UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS AGRARIAS
Campo Experimental Villarino C.C. Nº 14
S2125ZAA | Zavalla | Santa Fe | Argentina
Tel/Fax: +54 0341 497-0080
www.fcagr.unr.edu.ar
agro@unr.edu.ar

"Doctorado en Ciencias Agrarias"

FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

ANÁLISIS Y APROVECHAMIENTO DE BASES DE DATOS AGRONÓMICAS RECURRIENDO AL PROCESO "KNOWLEDGE DISCOVERY IN DATABASES"(KDD) Y ALGORITMOS DE "DATA MINING"(DM).

Una Aplicación al Pronóstico de Producción de Frutas de Pepita en los Valles de Río Negro y Neuquén

Mg.Sc. Gustavo Néstor Giménez

TESIS PARA OPTAR AL TÍTULO DE DOCTOR EN CIENCIAS AGRARIAS

Director: Dr. Sergio Bramardi

Codirector: Dr. Guillermo Pratta

Codirectora: Dra. Celina Beltrán

Año: 2020

Declaración

ANÁLISIS Y APROVECHAMIENTO DE BASES DE DATOS AGRONÓMICAS RECURRIENDO AL PROCESO "KNOWLEDGE DISCOVERY IN DATABASES"(KDD) Y ALGORITMOS DE "DATA MINING"(DM).
Una Aplicación al Pronóstico de Producción de Frutas de Pepita en los Valles de Río Negro y Neuquén

Gustavo Néstor Giménez

Magister Scienteae-Ingeniero Agrónomo-Universidad Nacional del Comahue

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias Agrarias, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en ésta u otra Universidad. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en la Facultad de Economía y Administración, Universidad Nacional del Comahue, durante el período comprendido entre el año 2014 y el año 2019, bajo la dirección del Doctor Sergio Bramardi.

Nombre y firma del Doctorando: 

Nombre y firma del Director:

Nombre y firma del Co - Director

Nombre y firma del Co - Director

Defendida:

Agradecimientos

Este trabajo de tesis se realiza en virtud del apoyo de instituciones públicas como la Universidad Nacional del Comahue, la Facultad de Economía y Administración y la Universidad Nacional de Rosario a las que agradezco infinitamente.
Un agradecimiento especial al Dr. Sergio Bramardi quien no sólo me guió a lo largo del trabajo de investigación sino quien inculcó el estudio de la estadística.
El agradecimiento se hace extensivo al Dr. Guillermo Pratta y la Dra. Celina Beltran por su dedicación y tiempo. Los consejos del Dr. Guillermo Pratta siempre fueron más que oportunos.

Publicaciones y Presentaciones a Congresos

- Tassile, V.;Giménez, G.; Ibañez, A.; Rubio, N.;2014. Aproximación de Laplace en la Exponencial Completa para el Algoritmo EM en Modelos Mixtos.XIX Reunión del Grupo Argentino de Biometrıa.

- Rubio, N.; Giménez, G.; Lavalle, A.; Giménez, Ch.; Macchiavelli, R.;2015. Método de Laplace y Cuadratura de Gauss-Hermite como Técnicas de Aproximación de la Integral en la Función de Verosimilitud de un Modelo Generalizado Mixto. V Congreso de Matemática Aplicada Computacional e Industrial.

- Giménez, G.;Tassile, V.; Using Non-linear mixed models and artificial neural network in the fitting growth pattern in pears cv. 'Williams' tp predict final sizes at harvest. XV Conferencia Española y V Encuentro Iberoamericano de Biometrıa.Bilbao.España.Septiembre 2015.

- Giménez, G.; Rubio,N.; Haique, A.; Macchiavelli, R.;2016. Métodos alternativos para predecir categorias de frutos en base a curvas de crecimiento simuladas: modelo generalizado mixto multicategórico y support vector machine (SVM). XXI Reunión del Grupo Argentino de Biometrıa.

- Giménez, G.; Bramardi, S.; Díaz, S.;Montes, S.;2017. Calibrating the suport vector machine(SVM) from simulated growth curve by mixed Nonlinear models for the prediction of harvest in pears "Beurre D'Anjou".XVI Spanish Biometric conference. Oral presentation.

- Giménez, G.; Rubio, N.; Del Brío, D.; Bramardi, S.; 2017. Ponderación de la técnica "Support Vector Macchine" (SVM) para el mejoramiento en la predicción de la distribución de tamaños a cosecha de frutos de peras cultivar Beurre D'Anjou. XLV Coloquio de la Sociedad Argentina de Estadística. XXII Reunión Científica del Grupo Argentino de Biometría. Jornadas del Instituto Interamericano de estadística.

- Giménez, G.; Rubio, N.; Bramardi, S.; 2018. Comparison between Mixed Non-Linear Models and Support Vector Regresion in growth curves of pears. XXIX IBC Proceedings Book. Oral presentation.

- Reeb, P.;Bramardi, S.;Tassile, V.; Giménez, G.; Curetti, M.; Alvarez, H.2018. Tablas de Raleo. Crecimiento de Frutos de Pepita en el Alto Valle de Río Negro y Neuquén. Editorial UNCo-FaEA. Sec.Extensión FaEA.https://sites.google.com/site/bioestadisticafacauncoma/extensi%C3%B3n/pex2017?authuser=0
- Marticorena, M.; Giménez, G.; Gonzalez, C.; Bramardi, S.; 2018. tuckerR.mmgg: Three-Mode Principal Components Analysis. https://cran.r-project.org/package=tuckerR.mmgg

Conferencias brindadas

- Giménez, G.;Tassile, V.; Using Non-linear mixed models and artificial neural network in the fitting growth pattern in pears cv. 'Williams' to predict final sizes at harvest. XV Conferencia Española y V Encuentro Iberoamericano de Biometría.Bilbao.Spain. September 2015
- Giménez, G.; Bramardi, S.; Diaz, I.; Montes, S.; 2017. Calibrating the support vector machine (SVM) from simulated growth curves by mixed nonlinear models for the prediction of harvest in pears "Beurre D'Anjou". XVI Spanish Biometric Conference. Seville. Spain
- Giménez,G.;Rubio, N.; Bramardi, S.; 2018. Comparison between mixed non-linear models and support vector regresion in growth curves of pears. XXIX International Biometric Conference. Barcelona. Spain.

Premios logrados

- Using Non-linear mixed models and artificial neural network in the fitting growth pattern in pears cv. 'Williams' tp predict final sizes at harvest has been selected for the young statistician Showcase as the best research in Argentina and one of the best contributions presented by a young researcher in XVth. Spanish Biometric Conference and Vth Ibero-American Biometric Meeting with the work entitled "Using Non-liner Mixed Models and Artificial Neural Network in the Fitting Growth Pattern in Pears cv. Williams to Predict Final Sizes at Harvest". XVth. Spanish Biometric Conference and Vth Ibero-American Biometric Meeting.Bilbao. Spain. 2015.
- "International Biometrics Society Travel Awards" to assist at XVI Spanish Biometric Conference. Seville. Spain .2017.

Índice General

Índice de Figuras

Índice de Tablas

Lista de Abreviaturas

AFD Argentine Fruit Distribuitors

AIC Criterio de Información de Akaike por sus siglas en inglés Akaike Information Criteria

ANA Ácido Naftalén Acético

ANN Redes Neuronales Artificiales por su siglas en inglés Artificial Neural Network

BA Benciladenina

BIC Criterio de Información Bayesiano por sus siglas en inglés Bayesian Information Criteria

CART Algoritmo de Clasificación y Regresión Basado en Árboles por sus siglas en el inglés Clasification and Regresion Trees

ddplf días después de plena floración

DM Minería de datos o por su siglas en inglés Data Mining

FFNN Redes Neuronales Prealimentadas por sus siglas en inglés feed forward neural network

GA Giberelinas

KDD Descubrimiento de nuevo conocimiento en bases de datos o por sus siglas en inglés Knowledge Discovery in Data Bases

KNN K vecinos cercanos o por sus siglas en inglés K-nearest neighbor

LOESS Regresión Ponderada Localmente o por sus siglas en inglés Locally Estimated Scatterplot Smoothing

MLG Modelo Lineal Generalizado

MLP Perceptrones Multicapas por sus siglas en inglés Multi-layer Perceptron

MNL Método de los modelos no lineales

MOP Modelo de Odds Proporcionales

PMF Peso medio del fruto

RBF Función de Base Radial Gaussiano o por sus siglas en inglés Radia Base Function

SEFRN Secretaría de Estado de Fruticultura de la provincia de Río Negro

SGBD Sistema Gestor de Bases de Datos

SMN Servicio meteorológico Nacional

SQL Lenguaje estructurado de consultas ó por sus siglas en inglés Structured Query Language

SVM Máquina de soporte de vectores por sus siglas en ingles support vector machine

Resumen

Una de las principales actividades económicas en las provincias de Río Negro y Neuquén, es la producción de peras y manzanas. En dicha zona se ha llevado a cabo el pronóstico de producción desde el año 1992 durante 23 años. El pronóstico de producción de frutas de pepita ha sido una herramienta importante para planificar la cosecha y mejorar estrategias de mercado. El método de predicción de la producción, con antelación a la cosecha, de los principales cultivares se basó en curvas de crecimiento. La curvas de crecimiento no solo permitieron estimar la producción sino que, conjuntamente con la relación diámetro-peso de los frutos, los tamaños comerciales. Toda esta información ha generado un volumen de datos que resulta difícil procesar y aprovechar con los métodos estadísticos habituales. Una opción para estos casos es utilizar una técnica de extracción de conocimientos en bases de datos también llamado proceso KDD(Knowledge Discovery in Data Bases). El proceso KDD consta de tres etapas: preprocesamiento, análisis de datos aplicando técnica de minería de datos y extracción de conocimiento. El objetivo principal de esta tesis fue aplicar el proceso KDD y algoritmos de "Data Mining" como el SVM o máquina de soporte vectorial aplicados al pronóstico de producción. Otro objetivo de este trabajo fue diseñar una base de datos que pudiera preservar la información generada. Además, se aplicaron técnicas de preprocesamiento y visualización para detectar datos faltantes y con errores de registro; se buscaron relaciones entre variables como peso y diámetro. Para esto fue esencial programar nuevas funciones y algoritmos en R. Una vez sistematizados los datos de crecimiento se ajustó un modelo estadístico y se estimaron los efectos del mismo destacándose el efecto de la parcela. A partir de la estimación del modelo se simularon curvas de crecimiento para calibrar y entrenar el SVM. Aprovechando las curvas simuladas se verificó que el SVM mejoró el ajuste de las curvas de crecimiento observándo un error cuadrátivo medio menor que utilizando modelos estadísticos. La utilización del SVM como clasificador multiclase permitió predecir con antelación a la cosecha los tamaños comerciales de los frutos. La ventaja de aplicar el SVM residió principalmente en procesar mayor volumen de datos y lograr mayor precisión en el pronóstico. El alcance de las predicciones del SVM fue evaluado con una experiencia a campo donde se realizó una predicción 14 días posteriores a la cosecha comercial y se comparó con los tamaños de los frutos recolectados. La precisión expresada en tamaños comerciales correctamente clasificados fue de 30% pero al reagrupar las clases productivas en frutos pequeños, medianos y grandes se logró una precisión de 70%. Mediante esta tesis se logró sistematizar, procesar y analizar un gran volumen conformando una

base de datos de 17 tablas y 160.000 registros. La aplicación del proceso KDD y de algoritmos de DM permitió obtener predicciones de gran precisión.

Palabras claves: Pronóstico, Fruticultura, Modelos No Lineales, Máquina de Soporte de Vectores, Precisión.

Abstract

ANALYSIS AND USE OF AGRONOMIC DATABASES APPLYING THE “KNOWLEDGE DISCOVERY IN DATABASES” (KDD) PROCESS AND “DATA MINING” (DM) ALGORITHMS. An Application to the Forecast of Fruit Production in the Valleys of Río Negro and Neuquén.

In Río Negro and Neuquén provinces, pears and apples production forecast had been carried out since 1992 during 23 years. Forecasts are a valuable tool for planning harvesting and improving marketing conditions. The method to predict the entire production of the main cultivars before harvest are based on fruit growth curves, we use it not only to estimate the whole production, but also to know the relationship between weight and diameter required for ideal commercial size. All this information had generated a volume of data which is unable to process with an usual procedure. Therefore in these cases there are process of knowledge extractionn in databases called KDD which is the most recommended option, it consist of 3 stages: preprocessing, analysis using data mining(DM) techniques and obtaining knowledge. Based on the aforementioned, the overriding objective was to apply the KDD process and data mining algorithms such as Support Vector Machine(SVM) to the forecast data for its use. The aim of this research was design a data base which could be able to house the generated information. In addition, preprocessing and visualization techniques were applied to detect errors and missing data, and the relationship between variables such diameter and weight. It was essential to program new algorithms and functions in R. Once we had the data growth results, an statistical model was fitted, in the same way it could estimate the effect that standed out, which was the farm effect. Since the model, growth curves were simulated for calibration and tunning of the SVM. Using the simulated curves, it was verified that SVM improved the adjustment of the growth patterns with a smaller mean root error than the statistical models. Using the SVM as a multiclass classifier, it was possible to predict in advance the harvest the commercial size. A considerable advantage of SVM consist of primarily processing more data having a high accuracy of the forecast. The reach of the SVM predictions, was evaluated doing fieldwork, making predictions 14 days after the commercial harvest, comparing its result with the harvested fruits. The precision was $0,3$ but when regrouping the size in small, medium, and large fruits the precision remained at $0,7$. In this thesis. it has been possible to save, process and analyze a large volume of data(17 tables and 160.000 registers) from the KDD process, obtaining accurate predictions.

Keywords: Forecasting, Fruit Growing, Nonlinear Models, Support Vector Machine, Accuracy.

Capítulo 1

Presentación del Problema, Objetivos e Hipótesis

1.1 El problema de las grandes bases de datos

La rápida evolución de la computación en los últimos años ha incrementado sustancialmente nuestra capacidad tanto para generar datos, como para su registro en diversas fuentes. Enormes cantidades de datos fluyen de casi todos los aspectos de nuestras vidas, los cuales son almacenados en bases de datos (Han et al. (2012)). Las bases de datos y los sistemas de bases de datos son componentes esenciales de la vida en la sociedad moderna, la mayoría de las personas realizan sus actividades diariamente que involucran alguna interacción directa o indirecta con estas: transacciones bancarias, consultas en librerías, pago de impuestos, reservas de pasajes, reservas en hoteles y muchos otros ejemplos (Elmasri (2011)). Las ventajas de las bases de datos como forma de almacenamiento son indiscutibles, en primer lugar se eliminan los voluminosos archivos en papel, la velocidad en recuperar y actualizar datos es muy superior a la de cualquier otro medio, son menos laboriosas puesto que se elimina el trabajo de manipulación de archivos en papel y por último se puede disponer rápidamente de datos precisos y actualizados (Date (2001)). La pronta disponibilidad de los datos se debe en un grado no menos importante a internet, que ha permitido además una mayor accesibilidad a la información en variados formatos (Maimon and Rokach (2010)).

El método tradicional de convertir datos en conocimiento, es decir, el análisis e interpretación de los datos de forma manual es lento, costoso y altamente subjetivo (Efron (2018)). Por otro lado, los volúmenes de datos crecen exponencialmente, conforme avanza la tecnología de registro y digitalización de las observaciones. Como afirma (Romero (2009)) estos crecientes volúmenes de datos tienen asociados dos problemas fundamentales: el almacenamiento y gestión, y la obtención de información-conocimiento a partir de ellos. La gestión y almacenamiento es posible gracias a una colección de programas que permiten crear y mantener una base de datos, llamados sistemas gestores de bases de datos (SGBD) (Elmasri (2011)). El descubrimiento de información-conocimientos en bases de datos se ha desarrollado desde

hace un tiempo, utilizando análisis informático y análisis inteligente de datos conjuntamente con tecnologías de inteligencia artificial y existe una inmensa diversidad de técnicas en investigación sobre el proceso de descubrimiento de conocimiento en bases de datos o *Knowledge Discovery in Data Bases*(KDD) (Frawley et al. (1992)).

El avance de la computación ha producido incrementos tanto en los registros de las bases como en la cantidad de atributos de cada registro, dando lugar a los datos de alta dimensionalidad. Romero (Romero (2009)) distingue dos situaciones en alta dimensionalidad: mayor o igual cantidad de observaciones que de atributos y menor cantidad de observaciones que de atributos. Por otro lado, se observa no sólo un aumento en la cantidad de datos a registrar, sino también un incremento en los formatos en que estos se registran, ya no sólo es posible almacenar información en tablas, sino que la tecnología inalámbrica ha permitido el surgimiento del flujo constante de datos (o *data-stream*) que obliga a generar nuevas técnicas para su tratamiento (Maimon and Rokach (2010)).

1.2 Bases de datos en agronomía: pronósticos de producción

En el ámbito agronómico, con el advenimiento de la agricultura de precisión, una gran afluencia de información en forma de imágenes y datos constituyen enormes bases de datos. En el contexto de la producción de frutas, ha hecho su incipiente aparición la fruticultura de precisión, donde el incremento de datos se observa a partir de la digitalización de la información, el cual supera toda capacidad humana para procesarlos y analizarlos. La aplicación de nuevos instrumentos de recolección de datos, tales como sensores de temperatura, sensores de fertilidad, producción, etc. con la posibilidad de almacenar estas observaciones y estar disponibles para su uso posterior, es cada vez más frecuente y de mayor capacidad. Este fenómeno se observa por ejemplo en el registro y digitalización para la caracterización de suelos en amplias regiones agrícolas (Armstrong et al. (2007)), en el registro de variables climáticas para pronósticos meteorológicos, en la obtención de imágenes para la clasificación de frutos en base a su color, en la creación de bancos de germoplasma para la preservación del material vegetal y genético, en pronósticos de producción en cultivos y en muchas otras actividades agronómicas. En fruticultura la conformación de censos de productores, el relevamiento de información a nivel provincial y regional para la realización entre otros, de pronósticos de producción, requieren de herramientas para su organización, procesamiento y posterior análisis.

Los pronósticos de producción en fruticultura son de sumo interés ya que permiten predecir con anticipación de varias semanas la producción de fruta a cosecha y resulta una herramienta muy valiosa para llevar a cabo una adecuada logística al momento de la recolección de los frutos, particularmente en lo que respecta a la disponibilidad de materiales y envases (Lötze and Bergh (2004)). La estimación y el pronóstico de producción en diferentes cultivos son requeridos para muchas decisiones relacionadas con la conservación, distribución y logística, obtención de precios y comercialización. Además, contribuyen a mejorar la competencia en los mercados, en

la toma de decisiones al momento de la comercialización de los frutos y a orientar estrategias de manejo en el cultivo, entre otros. Cabe mencionar por ejemplo, el relevamiento llevado a cabo el año 2013 en el pronóstico de producción de peras y manzanas realizado en el Alto Valle de Río Negro y Neuquén donde, para implementar esta metodología de predicción, se relevaron 2865 árboles distribuidos en 327 parcelas productivas, contabilizando 5730 unidades muestrales, lo cual ha demandado el trabajo de 39 relevadores a lo largo de todas las localidades de la región (Tassile et al. (2013)). Considerando que cada una de la unidades muestrales constituye un registro, que cada registro posee un gran número de atributos o campos y que estos pronósticos se llevan a cabo desde el año 1992, se advierte que esta ingente cantidad de datos requiere un análisis y sistematización en su conjunto. Si bien la calidad de las estimaciones a lo largo de los años de pronóstico fue, en base a las variaciones interanuales de los últimos diez años, de un promedio de $3,5\%$, seguramente pueda ser mejorada con nuevas técnicas. El desafío es obtener conocimientos a partir de estas bases de datos crudos que permitan la implementación de nuevos métodos y técnicas para extender el conocimiento, en una de las tareas agronómicas mas importantes como es la estimación de producción de los cultivos (Raorane and Kulkarni (2013)).
Por otro lado, esta información generada puede resultar de gran utilidad para futuras predicciones que permitan alcanzar una mayor precisión. Por ello, se requiere la utilización de nuevas metodologías y procesos que contemplen esta gran masa de datos, y que permitan la sistematización y organización de los datos para usos futuros.

1.3 Distintas aproximaciones para el análisis de grandes bases de datos

Existen herramientas que toman la información, examinan los datos en busca de patrones que son relevantes para nuestro problema y devuelven una respuesta (o un resultado). El proceso de desarrollo de este tipo de herramientas ha evolucionado en gran número de campos como la física, la ciencia de la computación, la estadística y ha sido llamado “machine learning”, “artificial intelligence”, “data mining (DM)”, “predictive analitycs” y “ Knowledge Discovery in Data Bases (KDD)”. Mientras cada uno de los campos se aproxima al problema utilizando distintas perspectivas y conjunto de herramientas, el objetivo final es el mismo: realizar predicciones con la mayor precisión posible (Kuhn and Johnson (2013)). Una de las herramientas que resultaría muy útil para extraer conocimiento de bases de datos agronómicas es el proceso KDD.

El proceso KDD o de obtención de conocimientos a partir de bases de datos, integra distintos aspectos de disciplinas como la estadística, la informática, la inteligencia artificial y la minería de datos. Se define KDD al proceso de identificar patrones válidos, nuevos, potencialmente útiles y comprensibles en grandes volúmenes de información (Maimon and Rokach (2010)). Es un proceso compuesto de distintas etapas interactivas e iterativas, que involucran la selección, limpieza y reducción de las bases de datos intervinientes. El objetivo

final es poder utilizar posteriormente, distintas herramientas en los datos que son de interés para alcanzar la interpretación de los mismos (Fayyad et al. (1996)). Algunos autores (Maimon and Rokach (2010)), diferencian a lo largo del proceso un total de 9 etapas, las primeras 4 etapas están vinculadas a la comprensión del dominio y los objetivos de la aplicación del KDD, considerando también la selección de los datos y el preprocesamiento de los mismos. Las 3 etapas siguientes son netamente de utilización de minería de datos (DM), luego la evaluación e interpretación y finalmente la obtención del conocimiento. La DM conforma el epicentro matemático del proceso KDD, desarrollando modelos y otorgando los algoritmos que realizan la exploración para descubrir patrones significativos que son la esencia del conocimiento a obtener (Maimon and Rokach (2010)). Las aplicaciones de la minería de datos en el contexto del KDD han sido fundamentales: en medicina, para encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico, en mercadotécnia para identificar clientes susceptibles de responder a ofertas de productos, etc. No obstante, no es una metodología muy difundida en agronomía y en particular en la fruticultura donde podría brindar herramientas predictivas muy importantes y poderosas, en particular en pronósticos de producción.

Para algunos autores (Han et al. (2012)) la minería de datos es el proceso de descubrir conocimiento de interés de grandes volúmenes de datos, es un campo multidisciplinario con contribuciones de la estadística, aprendizaje de máquinas, reconocimiento de patrones y bioinformática, ampliamente utilizada en muchas áreas. Es un proceso apropiado para la detección de relaciones y patrones en grandes bases de datos, aunque también puede ser aplicada en bases pequeñas (Palmer et al. (2011)). La minería de datos aplicada al campo de la agricultura y particularmente al pronóstico de producción de los cultivos es una técnica muy novedosa para la investigación (Ramesh and Vishnu Vardhan (2013)). A menudo, suele utilizarse indistintamente el concepto de KDD y de DM no obstante, la DM es una etapa del proceso KDD, destacándose la necesidad de la colección e integración de los datos y las etapas de depuración y preparación de los datos, previas al DM (Palmer et al. (2011)).

1.4 Principales técnicas de minería de datos

Las principales técnicas para minería de datos incluyen clasificación, predicción, clustering o agrupamiento, detección de datos raros o outliers, reglas de asociación, análisis de secuencias, análisis de series de tiempo y procesamiento de texto y algunas nuevas técnicas como análisis de redes sociales (Zhao (2013)). Se debe resaltar que en el lenguaje computacional de minería de datos existen algunas diferencias terminológicas respecto de los términos de la estadística clásica. Las variables independientes o predictoras son denominadas variables de entrada, las variables de respuesta o dependientes corresponden a variables de salida (Hastie et al. (2008)).

Las técnicas de DM se pueden clasificar de acuerdo a si el entrenamiento requiere o no supervisión humana en: aprendizaje supervisado, aprendizaje no supervisado, entre otros de menor importancia. En el aprendizaje supervisado los datos que alimentan al algoritmo incluyen

las soluciones deseadas que son llamadas etiquetas. En los sistemas no supervisados los datos no proveen de etiquetas y los algoritmos deben aprender sin ellas (Géron (2019)).

Muchos son los algoritmos que se pueden aplicar en el contexto del DM.

En la tabla 1.1 se presentan distintas técnicas de DM de acuerdo al tipo de variable involucrada y al tipo de aprendizaje de los algoritmos:

Tabla 1.1: Técnicas de Data Mining de acuerdo a la naturaleza de las variables de entrada y salida (Var:Variables). Fuente: elaboración propia

	Var. Salida Contínua	**Var. Salida Categórica**	**Sin Variable de salida**
Variable de Entrada Contínua	Regresión Lineal Redes Neuronales Regresión no lineal K-NN Support Vector Machine	Regresión logística Redes Neuronales Regresión Logística Multinomial K-NN Support Vector Machine	Componentes Principales Análisis de Clúster Biclusters
Variable de Entrada Categórica	ANOVA Redes Neuronales Árboles de Regresión Support Vector Machine	Redes neuronales Árboles de clasificación Regresión Logística Support Vector Machine Naive Bayes Regresión Logística Multinomial	Reglas de asociación Análisis de coordenadas Análisis de correspondencias

Existe un gran debate respecto de si los algoritmos deben clasificarse como estadísticos o de aprendizaje de máquinas. Si bien algunos tienen su origen en la estadística otros han emergido desde la ciencia de la computación. No obstante, debe considerarse que la aplicación en una u otra disciplina difiere profundamente. Los métodos estadísticos se han centrado principalmente en la inferencia, que se logra mediante la creación y el ajuste de un modelo de probabilidad específico Bzdok (2018). El modelo nos permite calcular una medida cuantitativa de confianza de que una relación descubierta describe un efecto ”verdadero” que es poco probable que resulte del ruido. Además, si hay suficientes datos disponibles, podemos verificar explícitamente los supuestos (por ejemplo, igualdad de varianzas). Por el contrario, ML se concentra en la predicción mediante el uso de algoritmos de aprendizaje de propósito general para encontrar patrones en datos a menudo ricos y difíciles de manejar(Géron (2019)). ML hace suposiciones mínimas sobre los sistemas de generación de datos; pueden ser eficaces incluso cuando los datos se recopilan sin un diseño experimental cuidadosamente controlado y en presencia de interacciones no lineales complicadas Bzdok (2018). En definitiva tienen objetivos y marco de aplicación distintos.

Entre los algoritmos de DM de la tabla 1.1 se destaca las redes neuronales artificiales(ANN) como uno de los algoritmos más utilizados dentro de la minería de datos debido a su gran capacidad predictiva especialmente en el procesamiento de imágenes. Es un sistema de procesamiento de datos inspirado en las redes neuronales biológicas. Se basa fundamentalmente en el concepto de “neurona”, donde cada una de ellas aplica una función de activación a una entrada que puede provenir de otra neurona o de datos de entrada, cada neurona posee un peso que se otorga de manera iterativa, luego se obtiene un valor de salida que puede distribuirse al

resto de la red o como valor de salida final (Palmer et al. (2011)). Ha demostrado ser una buena aproximación a problemas donde la predicción de nuevos datos se hace impreciso o variante en el tiempo (Gironés Roig (2013)). No obstante, en agricultura no ha sido una técnica muy utilizada, sólo algunos casos de pronóstico de rendimientos de trigo utilizando como predictores sensores de fertilización (Ramesh and Vishnu Vardhan (2013)).

Otro algoritmo muy utilizado son los árboles de decisión realizan particiones secuenciales de un conjunto de datos que maximiza la diferencia respecto de la variable dependiente. Ofrecen una manera concisa de definir grupos que son consistentes en sus atributos pero varían en cuanto a los términos de la variable dependiente (Palmer et al. (2011)). Cabe aclarar, que se denomina árbol de decisión cuando la variable de salida es categórica y árbol de regresión cuando la variable de respuesta es contínua (Faraway (2006)). Se representan mediante nodos que corresponden a variables de entrada, ramas asociadas a grupos de las mismas variables y hojas que son valores de la variable de salida. Los árboles de decisión presentan una naturaleza descriptiva que permiten una fácil interpretación de las particiones realizadas por el modelo (Maimon and Rokach (2010)).

Otra de las técnicas muy utilizadas en minería de datos es K vecinos cercanos (KNN), está basada en el concepto de similaridad, el objetivo en este caso es identificar k-observaciones en los datos de entrenamiento similares a los datos que se desean clasificar, basado en los valores de las variables independientes, son empleadas k-observaciones similares para clasificarlas en una clase (Palmer et al. (2011)). Este tipo de técnicas ha sido utilizado para el pronóstico de precipitación y otras variables climáticas en meteorología (Ramesh and Vishnu Vardhan (2013)).

No obstante, las máquinas de soporte vectorial o *Support Vector Machine*(SVM) es uno de los métodos más eficientes para clasificación y regresión disponibles actualmente en minería de datos (Karatzoglou et al. (2004)). Su origen se remonta al año 1963 por Vladimir Vapnik y se basa en la teoría del perceptrón por lo que resulta en un clasificador binario sumamente eficiente.

Entre los algoritmos no supervisados, existen algunos basados en técnicas de reducción de dimensionalidad que se han utilizado para la agricultura, como son Componentes Principales otros son basadas en técnicas de agrupamiento. No obstante, en el contexto del aprendizaje no supervisado, los algoritmos pueden ser de gran utilidad pero no se han reportado muchas aplicaciones en agricultura (Ramesh and Vishnu Vardhan (2013)). Entre las técnicas estadísticas multivariadas que podrían revestir importancia al momento de implentarse como algoritmo de DM, se puede mencionar de forma general las técnicas de tres vías que se originan a partir de registros de tres modos, cuya estructura contiene individuos y variables en distintos ambientes o condiciones, muy utilizada principalmente en la caracterización de recursos genéticos (Marticorena et al. (2013)).

La implementación es un aspecto importante, para ello el software R (R Core Team (2019)) ofrece una gran gama de herramientas que facilitan la aplicación de estás técnicas. El ambiente[1]

[1]se denomina a un sistema coherente y plenamente planificado, con herramientas específicas y flexibles

R es una suite integrada de facilidades de software para la manipulación, cálculo y manejo de datos con potentes herramientas gráficas. Entre otras cualidades, posee un efectivo manejo de datos y facilidades en su almacenamiento, un conjunto de operadores para cálculos en arreglos y matrices, una colección integrada de herramientas para el análisis de datos, facilidades gráficas para el análisis y visualización de datos, y provee un desarrollado, simple y efectivo lenguaje de programación llamado S (Venables et al. (2013)). El ambiente R permite interactuar directamente con las bases de datos y con los SGBD, posibilitando crear, diseñar y gestionar bases de datos con lenguaje SQL mediante librerías que ofician de interfaz entre R y las bases de datos (James and DebRoy (2012)). Por otro lado, R posee un gran repositorio con una importante cantidad de librerías aplicables a las principales técnicas de minería de datos (Zhao (2013)). Por lo anteriormente expuesto, R es un ambiente completo, actualizado y muy versátil para aplicar el proceso KDD en su totalidad.

Los procesos que se proponen para abordar las nuevas tecnologías de relevamiento de datos van a permitir no sólo una organización y sistemación de los registros, sino una herramienta de análisis que va a ser plausible la interpretación en grandes masas de datos agronómicos.

1.5 Objetivos e hipótesis

1.5.1 Objetivo General

Estudiar la aplicabilidad y evaluar los resultados del proceso KDD y algoritmos de DATA MINING en bases de datos agronómicas de alta dimensionalidad.

1.5.2 Objetivos específicos

1. Organizar y sistematizar, en bases de datos, información proveniente del relevamiento frutícola del pronóstico de producción que se realiza desde el año 1992 en la zona de Río Negro y Neuquén.

2. Realizar un preprocesamiento de datos dentro del contexto del KDD adaptado a las curvas de crecimiento que permita explorar las características de la base de datos y las variables involucradas.

3. Detectar patrones de comportamiento en los datos mediante el proceso KDD para las diferentes temporadas de pronóstico de producción.

4. Comparar técnicas de minería de datos y los modelos de predicción de cosecha implementados en el pronóstico de producción.

5. Desarrollar algoritmos de minería de datos para mejorar la detección de patrones, en el pronóstico de producción frutícola considerando, las localidades, las distintas variables o atributos y las temporadas, es decir la estructura de 3 modos, de los datos.

1.5.3 Hipótesis

La utilización del proceso KDD permite ordenar, clasificar y encontrar patrones de manera más eficiente que los métodos estadísticos utilizados actualmente en bases de datos agronómicos de alta dimensionalidad. Las predicciones del pronóstico de producción de los principales cultivares de peras y manzanas de los Valles de Río Negro y Neuquén, podrán ser más precisos mediante este proceso respecto de los modelos actualmente implementados. Haciéndose esto extensivo a otros pronósticos frutícolas.

1.6 Importancia de esta tesis

Actualmente, en las ciencias agrarias la afluencia de ingentes volúmenes de datos requiere en primera medida, del diseño y creación de sistemas de bases de datos capaces de contener y gestionar la disponibilidad de los mismos. Los pronósticos de producción no son la excepción, y utilizan grandes masas de datos de distintas fuentes para obtener predicciones de mayor precisión. Al respecto, la mayor cantidad de trabajos se observan en dominios como meteorológia, imágenes, mapas, geografía, registros agropecuarios y de catastro, entre otros, pero no en bases de pronóstico de producción que además tiene una metodología propia.

La importancia de esta tesis radica en utilizar un proceso de análisis que permita organizar bases de datos agronómicas de pronóstico de producción para extraer información y obtener conocimientos para futuros pronósticos y, de esta forma, mejorar los pronósticos de cosecha aprovechando los datos de bases de datos tan particulares como la que se emplea en esta tesis.

Por otro lado, esta tesis provee métodos para sistematizar datos con la particularidad de estar basados en curvas de crecimiento de gran utilidad en variado tipo de técnicas predictivas.

Además de aprovechar las bases de datos se proponen y testean algoritmos que permiten mejorar la exploración de datos y los modelos estadísticos de crecimiento del fruto.

En definitiva se propone una metodología de trabajo para pronósticos de producción abordando el problema de alta dimensionalidad, para aprovechamiento y extracción de conocimiento de grandes bases de datos.

Por otro lado, existe una demanda de los sectores productivos para implementar sistemas de evaluación y mejora de la calidad de frutos frescos que permitan automatizar el proceso, acelerarlo y reducir los problemas de clasificación manual. Al respecto, este trabajo de tesis pretende realizar un aporte de mejora en le precisión de la predicción de los tamaños comerciales componente fundamental de la calidad de los frutos.

Por último, este trabajo brinda algoritmos entrenados y testeados para la creación de futuras aplicaciones informáticas que permitan al productor conocer de acuerdo a los calibres que presenten los frutos de su parcela, el tamaño y calidad comercial de los frutos.

Con la firme convicción de que la información es una fuente de riqueza inconmensurable en los tiempos que corren y esta tesis realiza un pequeño aporte para su mejor aprovechamiento.

Capítulo 2

Revisión de antecedentes

Dado que el presente trabajo de tesis aborda la problemática desde distintas perspectivas: frutícola, estadística y computacional. Se considera necesario una introducción que equipare los conocimientos de los distintos puntos de vista. Por este motivo, al lector con amplios conocimientos en producción frutícola se sugiere leer directamente el punto 2.3 “Descripción del proceso de KDD”, en tanto que al experto en Data Mining los puntos 2.1 y 2.2.

2.1 Pronósticos de producción

Los pronósticos de producción son un claro ejemplo de aplicación de grandes bases de datos agronómicas. Existen muchos ejemplos de pronósticos de producción cuya información provee una herramienta para la logística, comercialización y producción de frutas. En general, dicha práctica requiere de distintas fuentes de información que no puede ser manipulada y recuperada eficientemente por lo cual, las bases de datos son una solución factible. Dada la gran importancia que revisten los pronósticos de producción de frutas en particular de peras y manzanas existe una asociación internacional denominada “WAPA”(The World Apple and Pear Association en http://www.wapa-association.org) cuya función es registrar y elaborar informes de la producción anual de peras y manzanas en el mundo. También es de su competencia generar y presentar anualmente un evento denominado “Prognosfruit” ó pronósticos de producción de distintas regiones del mundo donde se presenta la estimación de producción de la región europea, del hemisferio sur y de Estados Unidos.

Asimismo, el NASS (National Agricultural Statistics Service) es una agencia dependiente del Departamento de Agricultura de Estados Unidos (USDA) cuya función es registrar y difundir información actualizada y completa sobre la producción en general a todos los productores de ese país. Básicamente, provee estadísticas actualizadas, precisas y útiles a la agricultura de Estados Unidos. Recolecta, analiza, procesa y difunde datos sobre todos los aspectos de la agricultura basados en encuestas, e información administrativa y satelital. Dicha información se vuelca en un informe anual que publica de forma accesible y donde se observan los datos registrados para la mayoría de los cultivos de interés para ese país entre los que se encuentran

peras y manzanas (RoSS (2018)) .

El centro de servicios Hortofrutícola de Italia(CSO o Centro Servizi Ortofrutticoli Italy https://www.csoservizi.com/servizi/statistiche-di-produzione-e-mercato/) realiza previsiones e informes de cosecha para distintos cultivos como kiwi, manzana, peral, duraznos y nectarines, damasco, etc., tanto para Italia como para países productores de la Unión Europea.

En Argentina, se registra en la provincia de Mendoza una metodología desarrollada para el pronóstico en manzanos cv. "Fuji", que tiene en cuenta las formas del fruto y la época y la intensidad del raleo (Gil (1996)). Posteriormente, el Instituto de Desarrollo Rural (IDR) de Mendoza, implementó prónosticos de producción en distintas variedades de durazno, ciruela y peral.

Por su parte, en la provincia de Corrientes se ha desarrollado la metodología necesaria de construcción de curvas de crecimiento, con clasificación por tamaños comerciales, para la realización del pronóstico de producción en cítricos, en particular de naranjos dulces (Avanza (2010)). En la provincia de Entre Ríos se están desarrollando las curvas de crecimiento y las distintas metodologías de pronóstico de producción en mandarinas.

La Facultad de Ciencias Agrarias de la Universidad Nacional del Comahue ha realizado pronósticos de producción de los principales cultivares de peras y manzanas desde la temporada 1992/1993 hasta la temporada 2013/2014 en el Alto Valle de Río Negro y Neuquén siendo uno de los programas de pronóstico de producción de frutas de pepita más extensos y de mayor trayectoria del país, cuyos resultados y metodologías son motivo de análisis de la presente tesis. La zona de incumbencia y las principales características productivas serán descriptas a continuación.

2.1.1 Descripción de la zona productiva del Alto Valle de Río Negro y Neuquén

Las provincias de Río Negro y Neuquén comprenden regiones productoras de pepita muy importantes (Figura 2.1), se pueden distinguir cuatros zonas: Alto Valle, Valle Medio, General Conesa y Río Colorado (Sanchez and Villareal (2015)).

El Valle Medio se localiza en el departamento Avellaneda de la provincia de Río Negro. Se extiende sobre el Río Negro desde la localidad de Chelforó hasta el paraje Fortín Castre. Posee un gran potencial de producción bajo riego y desde hace algunos años, varias empresas frutícolas han comenzado su desarrollo con importantes proyectos productivos tanto de peras y manzanas como de cerezas. El clima es de tipo continental, templado árido, con una precipitación que alcanza los $280\ mm$ concentradas en otoño y primavera (Godagnone and Bran (2009)). Los suelos son complejos de origen aluvial, es decir, surgen a partir de los materiales acumulados en sucesivos aportes del río. El $50\ \%$ de los suelos tienen problemas de sodicidad y salinidad lo cual representa severas limitaciones para el uso frutícola (Sanchez and Villareal (2015)).

El Valle inferior del Río Negro(Figura 2.1) se extiende desde la localidad de Viedma, precisamente la desembocadura del río en el océano hasta el Valle de Conesa. Se caracteriza

por una incipiente actividad agrícola con mayor desarrollo de la ganadería y una importante actividad minera, pesquera y turística, destacándose los servicios portuarios (Sanchez and Villareal (2015)).

El valle de Río Colorado es una zona que se encuentra en el largo y angosto valle al noroeste de la Provincia de Río Negro, en el departamento Pichi Mahuida. Limita con la provincia de La Pampa y al sur con la meseta patagónica. Posee un clima continental con una precipitación anual que alcanza los $390\ mm$ (Godagnone and Bran (2009)), que propicia el desarrollo de enfermedades fúngicas. Los suelos son aluvionales, de depósito reciente y escaso desarrollo por lo que el material original está prácticamente sin modificar (Sanchez and Villareal (2015)).

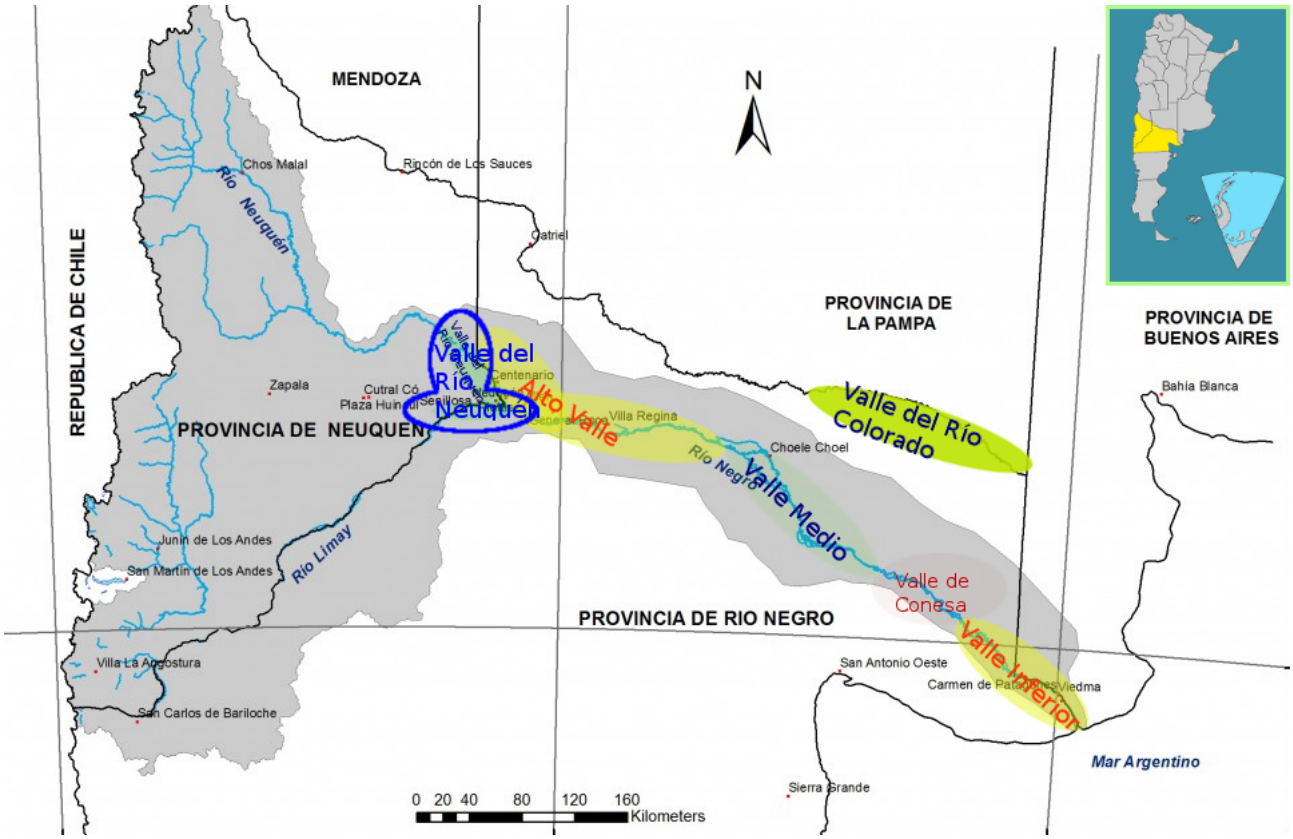

Figura 2.1: Zonas productoras de pepita en las cuencas de los ríos Negro y Colorado (Readaptado de Sanchez and Villareal (2015))

No obstante, es el Alto Valle de Río Negro y Neuquén la región de producción de peras y manzanas por excelencia en Argentina (Casamiquela and von Wagner (1999)). Dicha región está ubicada en los departamentos General Roca de la provincia de Río Negro y Confluencia de la provincia de Neuquén. Se extiende por ambas márgenes de los ríos Limay, Neuquén hasta la confluencia y luego primera parte del recorrido del Río Negro. El Alto Valle se puede situar entre las coordenadas geográficos $38°40'$ y $39°20'$ de latitud sur y $66°50'$ a $68°20$ de longitud oeste a una altura promedio respecto al nivel del mar de 243 mts.

El área productiva de la región está estimada en unas $100.000\ has$, aproximadamente el sesenta por ciento se encuentra bajo riego y de las cuales se cultivan alrededor de $40.000\ has$.

El área es destinada principalmente a la actividad frutícola, en particular al cultivo de peras y manzanas para la exportación en fresco y la industria de jugos concentrados. También se destaca el cultivo de la vid para la elaboración de vinos y en menor medida frutales de carozo como ciruelas, duraznos y pelones y finalmente el cultivo del tomate y la alfalfa (CPIARN (2015)).

Los suelos del Alto Valle son típicos de desierto, suelos grises que no tienen suficiente meteorización excepto aquellos suelos cercanos a las inundaciones del río o bajo cultivo (Casamiquela and von Wagner (1999)). Geomorfológicamente se puede describir como mesetas en avanzado proceso de desgastes, vertientes, pendientes de distintos gradientes, cauces, planicies aluviales lagunas y salitrales. Gran parte de los materiales son de naturaleza arcillosa están vinculadas con formaciones geológicas aflorantes. Los órdenes de suelos más comunes son los Aridisoles y los Entisoles (Godagnone and Bran (2009)). De acuerdo a Apcarian et al. (2006), los suelos del orden Entisoles corresponden al $28,4\%$ de los suelos del valle, los suelos Aridisoles constituyen el $64,4\ \%$ totalizando $64.000\ has$ (Apcarian et al. (2006)).

Por otro lado, teniendo en cuenta la textura de los suelos se han confeccionado mapas de distribución textural para la zona del Alto Valle:

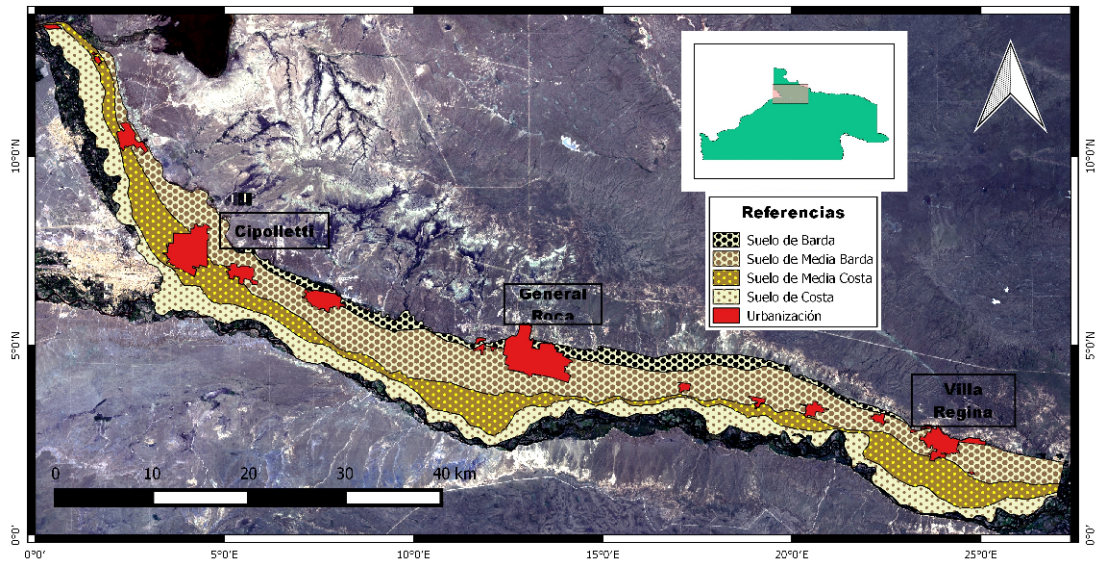

Figura 2.2: Mapa de distribución de las texturas de los suelos del Alto Valle (Extraído de Bestvater and Casamiquela (1983))

De acuerdo a la distribución de textura, los suelos de meseta, son los que se encuentran bordeando el norte del valle al pie de elevaciones llamadas 'Bardas' están formados por una gama de arenas en ocasiones muy gruesas y gravillosas (Figura 2.2). Próximos a los suelos de barda se encuentran los suelos de media barda, caracterizados por texturas que van de franco-limoso a franco. Suelen tener problemas de salinidad y alcalinidad, su productividad aparece directamente vinculada a su bajo drenaje. Más al sur de las bardas (Figura 2.2) encontramos los suelos de media costa, se definen con texturas que van de franco-arenosa a arenosa-franca, siendo más difusa su ubicación que los dos grupos anteriores. Finalmente, los suelos que

se disponen al sur del valle y próximos al río son los suelos de costa: que poseen texturas arenosas, en ocasiones interrumpidas por causes vinculados transitoria o permanentemente con el curso principal del río (Bestvater and Casamiquela (1983)).

En cuanto al clima, al igual que las demás zonas, es de tipo continental, templado árido aunque su régimen térmico es más moderado que aquellas regiones de la misma latitud por el efecto de la oceanidad, es decir, la baja proporción de tierra en relación al mar (Godagnone and Bran (2009)).

El Alto Valle se caracteriza por una baja humedad relativa del aire, lo cual hace que la región sea poco propensa a enfermedades fúngicas que usualmente afectan en otras zonas típicas de producción de frutales. También resalta el importante déficit hídrico durante el ciclo productivo dado que la región se encuentra entre las isoyetas de $150\ mm$ y $250\ mm$, con una precipitación media anual de $188\ mm$. De acuerdo a Cordon et al.(2000), el Alto Valle se caracteriza por una alta radiación: con una radiación teórica astronómica, para la latitud $39°00'$ máxima en el mes de diciembre de $1044\ cal.gr/cm^2.dia$ registrándose los valores más bajos en el mes de Junio con $313,4\ cal.gr/cm^2.dia$. Respecto a la radiación global, es máxima en el mes de enero alcanzando valores de $607,9\ cal.gr/cm^2.dia$.

En cuanto a la temperatura media de la región en enero, de acuerdo a algunos trabajos (Godagnone and Bran (2009)) se establece en los $22°\ C$, para el mes de julio, la zona se halla ubicada entre las isotermas $6°\ C$ y los $7°\ C$. La temperatura máxima media en Cinco Saltos es de $29,6°\ C$ en enero y $12,6°\ C$ en el mes de julio (Cordon et al. (2000)). La temperatura mínima media alcanza su mayor expresión en el mes de enero $12,8°\ C$ y los valores menores en julio con $-0,8°\ C$. Es quizás, la ocurrencia de heladas primaverales la adversidad climática más grave en cuanto a los efectos que produce en la actividad frutícola de esta zona, reportando año tras año importantes pérdidas no sólo en volumen sino en cuanto a calidad de los frutos (FAO (2015)). Dicho fenómeno es el que más explica las oscilaciones de producción anual de fruta en esta actividad. Si bien existen distintos métodos de control activo de heladas, sólo un 20% de las chacras de la provincia de Río Negro posee métodos activos de control y el 56% utiliza algún combustible (Godagnone and Bran (2009), Cordon et al. (2000)). La fecha media de heladas tardías es el 11 de octubre y la fecha extrema de helada tardía se registró el 19 de diciembre. La frecuencia de días con temperaturas por debajo de cero grado es de $69,9$ días .

Un aspecto importante para el desarrollo de los árboles frutales es el total medio acumulado de horas de frío, para la región es de $1559\ h$, el total anual acumulado mínimo de $1233\ h$, mientras que el total anual acumulado máximo es de $1948\ h$ (Cordon et al. (2000)).

Una característica propia de la zona es la presencia de vientos intensos dominantes del sector oeste sudoeste, con mayor frecuencia en primavera. La velocidad media del viento se corresponde con los sectores de mayor frecuencia y alcanza velocidades de entre $18,8\ km/h$ y $15,9\ km/h$. La presencia casi constante de vientos y la baja precipitación generan un clima seco y de baja humedad poco propicio para el desarrollo de enfermedades fúngicas.

Los fuertes vientos y la intensa radiación solar son la razón de los típicos sistemas de conducción observados en la región con estructuras de sostén en el diseño de las plantaciones

frutales.

2.1.2 Producción frutícola del Alto Valle y el contexto internacional

Argentina tiene una gran importancia como productor de peras y manzanas, en el decenio 2003-2013 ocupaba el puesto cuatro como productor mundial de peras con un volumen de entre $600.000\ tn$ y $700.000\ tn$ levemente inferior a Italia, país por excelencia en el cultivo de esta especie.La producción del Alto Valle, su conformación y estructura productiva puede explicarse conociendo la historia de su desarrollo y expansión como valle productivo que se presenta en la sección 8.1 del Anexo del presente trabajo de tesis.

No obstante, China es el país que ha liderado la producción de peras durante muchos años alcanzando ampliamante $9.000.000\ tn$. En cuanto a la producción de manzanas, Argentina se encuentra en el puesto 11 (FAOSTAT (2019)), por debajo de Rusia. En virtud que la manzana es un fruto cultivable en condiciones menos restrictivas que la pera, son muchos más los países intervinientes en el mercado. Nuevamente, el primer productor mundial es China con más de $20.000.000\ tn$ con tendencia en aumento en el decenio mencionado. A pesar de que el país asiático lidera ampliamente la producción de peras y de manzanas, no sucede lo mismo dentro del comercio internacional de las frutas (Nations (2019)), seguramente debido a su alto consumo interno. Segun la división de comercio y estadística de las Naciones Unidas (Nations (2019)), durante los años $2003 - 2014$ Argentina se posicionó como el primer exportador mundial de peras, logrando colocar un volumen promedio para el período de $414.000\ tn$ por encima de China y Holanda, los principales exportadores. Pero en cambio sólo logró el puesto 11 en la exportación de manzanas con paulatino descenso con apenas $208.000\ tn$.

Según los datos relevados por el SENASA (Senasa (2017)), de su sede del Centro Regional Patagonia Norte, en el año 2017 se registra una superficie implantada en peras de $20.879\ ha$, de las cuales $18.498\ ha$ corresponden a la superficie en la Provincia de Río Negro y $2.381\ ha$ a la provincia de Neuquén. Las principales variedades implantadas son *"Williams"* , *"Packham's Triumph"* y *"Beurre D'Anjou"* las cuales reúnen una superficie de $17.388\ ha$ que representa un $83,3\%$ de la superficie total implantada con peras (ver tabla 2.1). Sólo el cultivar *"Williams"* , está implantado con el $40,4\%$ y es históricamente el cultivar de peras más implantado y en general más demandado tanto para su comercio exterior como el mercado interno. Tanto el cultivar *"Williams"* , como *"Packhams Triumph"* , *"Beurre D'anjou"* y *"Abate Fetel"* han experimentado un importante retroceso en cuanto a la superficie implantada de los últimos años. El más afectado es el cultivar *"Williams"* cuya superficie se redujo un $20,5\%$ desde el año 2009 al 2017, en tanto que, los cultivares *"Beurre D'Anjou"* y *"Packham's Triumph"* se mantuvieron constantes con una leve reducción en la implantación de árboles frutales. Respecto a la comercialización, el organismo(Senasa (2017)) informa que para el año 2017 la región exportó un volumen total de $254.464\ tn$ de peras, donde $227.999\ tn$ son aportados por la provincia de Río Negro y el restante $26.465\ tn$ por la provincia de Neuquén. El destino principal para las peras regionales es el vecino país de Brasil absorbiendo $88.200\ tn$, en tanto que Rusia demandó $60.127\ tn$ y quedó

como segundo lugar mientras en tercer lugar Estados Unidos importó de la región $31.430\ tn$.

En cuanto a las manzanas la realidad es muy distinta. En primer lugar, de acuerdo a la tabla 2.1, la variedad con mayor superficie implantada es *"Red Delicious"* y sus clones mejorados: "Red Chief", "Angius"(llamada regionalmente "Chañar 28") y "Atwood"("Chañar 34"), alcanzando más de un 60% de la superficie total implantada. Esto se debió en parte al poco recambio varietal que existió en la región en cuanto a las manzanas y puede explicar la posición relegada en la que se encuentra la exportación de manzanas desde hace algunos años. En segundo lugar se observa el cultivar *"Granny Smith"* con sólo el $13,2\%$ del área y en tercer lugar el cultivar *"Gala"* y sus clones mejorados: *"Royal Gala"* , *"Mondial Gala"* , *"Galaxy"* , etc.

Tabla 2.1: Superficie (Sup) destinada a los cultivares de peras y manzanas en Río Negro y Neuquén, en hectáreas(ha) y respecto de la superficie total($\%$) en el año 2017.

Peras			Manzanas		
Variedad	$Sup(ha)$	$\%$	Variedad	$Sup(ha)$	$\%$
Williams	8.434	$40,4\%$	Red Delicious	12.516	$64,2\%$
Packham's Triumph	6.007	$28,8\%$	Granny Smith	2.578	$13,2\%$
Beurre D'anjou	2.947	$14,1\%$	Gala	2.565	$13,2\%$
Abate Fetel	1.112	$5,3\%$	Cripps Pink	888	$4,6\%$
Red Bartlett	1.002	$4,8\%$	Rosy Glow	227	1.4%
Beurre Bosc	535	$2,6\%$	Starkrimson	161	$0,8\%$
Beurre Giffard	340	$1,7\%$	Golden Delicious	106	$0,5\%$
Otras	480	$2,3\%$	Otras	409	$2,1\%$
Total	20.879	100%	Total	19.496	100%

Cabe resaltar el cultivar *"Cripps Pink"* , un cultivar novedoso que ha ganado superficie en los últimos años alcanzando el cuarto lugar en cuanto a importancia. Siguiendo el informe del organismo oficial SENASA (Senasa (2017)) la implantación del cultivar *"Red Delicious"* se retrajo respecto al 2009 un $18,5\%$, mientras que los cultivares *"Granny Smith"* y *"Gala"* se mantuvieron constantes en el mismo lapso. La exportación de manzanas regionales alcanzó $71.500\ tn$, los destinos fueron muy variados destacándose Paraguay y Brasil como los más representativos con un volumen aproximado de $12.000\ tn$, luego le siguen en orden de importancia Estados Unidos y Brasil colocando aproximadamente $9000\ tn$.

Claramente, dada la importancia en la producción de peras y manzanas de la región y la contribución al comercio internacional de frutas en especial de peras, el pronóstico de producción es una herramienta que pueda brindar una ventaja competitiva en todo el ciclo comercial y productivo de los frutos.

2.1.3 Historia del pronóstico de producción del Alto Valle de Río Negro y Neuquén

En el Alto Valle de Río Negro y Neuquén se ha llevado a cabo el pronóstico de producción en peras y manzanas desde el año 1992, durante 23 años se aplicaron técnicas que han incorporado metodologías desarrolladas durante mucho tiempo de investigación en la Facultad

de Ciencias Agrarias de la Universidad Nacional del Comahue. En el año 1991 se comienza con las primeras reuniones en las cuales las Secretaría de Fruticultura de la provincia de Río Negro y de producción de la provincia de Neuquén plantean la necesidad de un pronóstico de cosecha o de Producción. A partir de la necesidad explícita de un pronóstico se crea una comisión regional de pronóstico de producción que involucra a distintas instituciones, como la Dirección de Fruticultura y Producción Agraria del Ministerio de Economía de la Pcia, Río Negro, el Ministerio de la Producción de la Pcia. Neuquén, el Instituto Nacional de Tecnología Agropecuaria (INTA), la Facultad Ciencias Agrarias de la U.N.Comahue, la Cámara Argentina de Fruticultores Integrados (CAFI), la Cámara de la Industria y Exportación de Jugo de Peras, Manzanas y Afines (CINEX) y Federación de Productores de frutas de Río Negro y Neuquén.

El objetivo del pronóstico es conocer en forma anticipada la producción de frutos de peras y manzanas con el firme propósito de mejorar la competencia en los mercados, tomar decisiones en la comercialización, orientar las estrategias hacia los cultivos, proyectar ingresos y claramente establecer la logística de recolección (Tassile and Giménez (2013)). La estimación de la producción en toneladas depende de la carga que poseen las plantas, en cantidad de frutos por planta o expresado en frutos por hectárea, el tamaño del fruto a cosecha y de la cantidad de plantas o superficie implantada. De manera que para estimar la producción de una variedad A (P_A) se utiliza la siguiente ecuación:

$$\boxed{P_A = \frac{\sum_i^n (NF_i * PMF_i * PL_i)}{k}} \tag{2.1}$$

Donde:

n = número de estratos, definido por rangos de edad de las plantaciones en combinación con sistemas de conducción libre y espaldera

NF = número medio de frutos por árbol

PMF = peso estimado del fruto medio a cosecha

PL = número de árboles del estrato

k = factor de corrección que considera el error que se produce durante el conteo por frutos ocultos

La estimación de la carga depende de métodos de muestreo, es decir, requiere de un diseño de la muestra. El tamaño medio de los frutos a cosecha precisa de modelos estocásticos o de curvas de crecimiento, mientras que la cantidad de plantas o superficie se proporciona a través de censos productivos o del conocimiento de estructuras del sistema productivo.

Método de la estimación del peso medio del fruto por carga (Método estocástico)

En el año 1992 en convenio con el Centro Operativo Hortofrutícula de Ferrara se realiza el primer pronóstico de producción para el Alto Valle de Río Negro y Neuquén. Posteriormente, en el año 1993 se realiza el censo provincial de la agricultura bajo riego, cuyo procesamiento estuvo a cargo de la Facultad de Ciencias Agrarias de la Universidad Nacional del Comahue (Damico (1993)). Este censo, permitió conocer características de las parcelas productivas del Alto Valle como variedades implantadas, distanciamientos, número de plantas, etc. El censo fue un aporte muy importante para mejorar la precisión de los pronósticos, al proveer de información más certera de las parcelas del Alto Valle y su estructura agraria. Tassile and Giménez (2013).

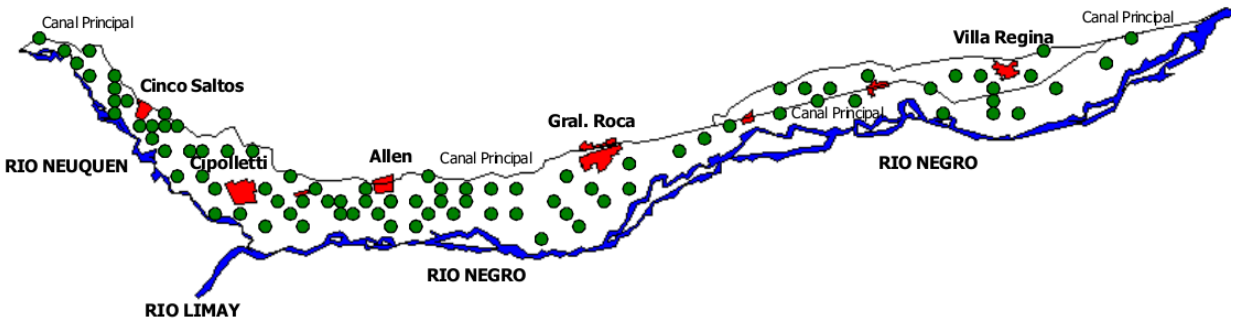

Figura 2.3: Distribución de las chacras de muestreo a lo largo de la región productiva para elaboración del pronóstico

Con la información del censo se establecieron 86 parcelas (o chacras) para el muestreo y 10 chacras experimentales a lo largo de todo el Alto Valle de Río Negro y Neuquén como se visualiza en la figura 2.3. Las parcelas experimentales fueron instaladas con el objetivo de calcular el parámetro k del modelo de la ecuación 2.1 y obtener estimaciones históricas del peso mínimo y del peso máximo del fruto. Asimismo, el conocimiento de la estructura parcelaria mediante el censo y del comportamiento de las plantaciones a lo largo de las temporadas en base al seguimiento de las chacras muestreadas (Figura 2.3) productivas permitió definir estratos en función de los rangos de edad de la plantación: estrato I de 10 a 19 años, estrato II de 20 a 29 años, estrato III de 30 a 39 años y estrato IV mayor a 40 años; y el sistema de conducción en libre y espaldera para cada variedad.

De acuerdo al modelo estocástico (Tassile et al. (2013)) implementado en ese momento el

peso del fruto se estimaba a partir de la carga de las plantas según la ecuación:

$$\boxed{P = \gamma \cdot e^{-\beta Q}} \qquad (2.2)$$

En el caso de la ecuación 2.2 el peso depende de los parámetro γ y de β y de la carga frutal

Q. Para la temporada 1994 se había logrado encontrar un modelo exponencial que describía como el aumento de la carga de frutos en la planta afectaba inversamente el peso y por ende el tamaño del fruto.

El comportamiento del peso del fruto y la carga se observa en la figura 2.4, en este caso se analizaba la relación del peso del fruto en gramos y la carga expresada en cientos de frutos por planta, a medida que la carga aumentaba se detectaba una disminución exponencial en el peso del fruto.

Del gráfico presentado en la figura 2.4 se destaca que el ajuste del modelo exponencial no describía satisfactoriamente la relación entre peso de los frutos y carga frutal especialmente debido a una gran dispersión de los datos expresado en un bajo valor de coeficiente de determinación (R^2) no superaba el valor de $0,7$.

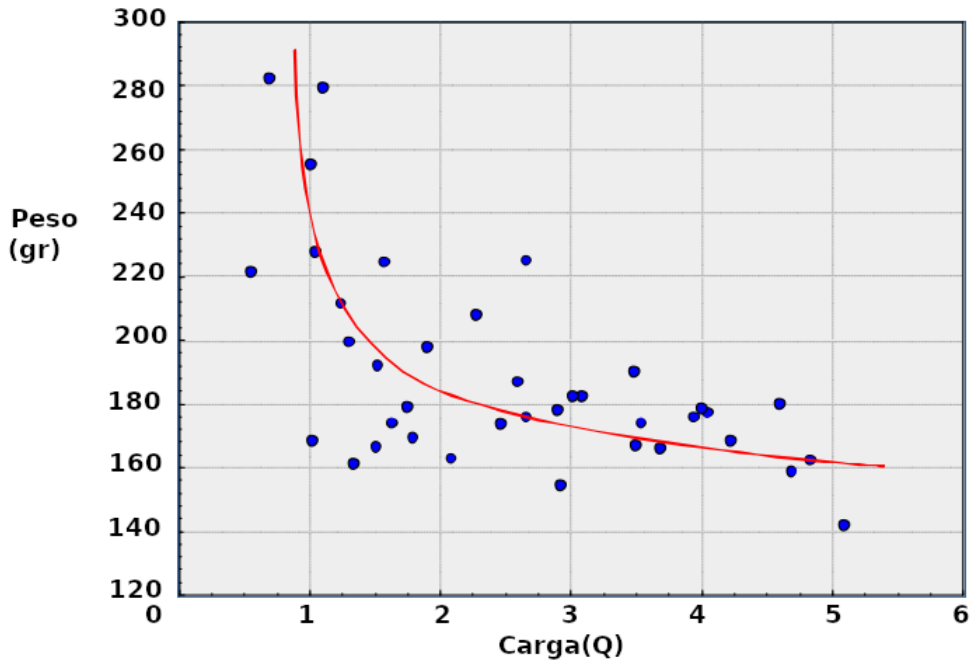

Figura 2.4: Relación entre el peso promedio en gramos de los frutos(P) y la carga frutal promedio(en cientos por planta) de las plantas(Q)

La baja relación presentaba la principal causa de los problemas de estimación en el pronóstico de producción.

Continuando con la descripción del método, para obtener el peso del fruto a cosecha, fue necesario conocer la distribución de la carga frutal y la distribución de los frutos. Por ello que, siguiendo con el método estocástico, se estudiaba la distribución de la carga frutal de las plantas en la temporada, la cual no se ajustaba a una distribución normal sino a una distribución gamma y respondía al patrón que se muestra en la figura 2.5. En la misma se observa un histograma de acuerdo a la carga frutal y una función de densidad aproximada al histograma.

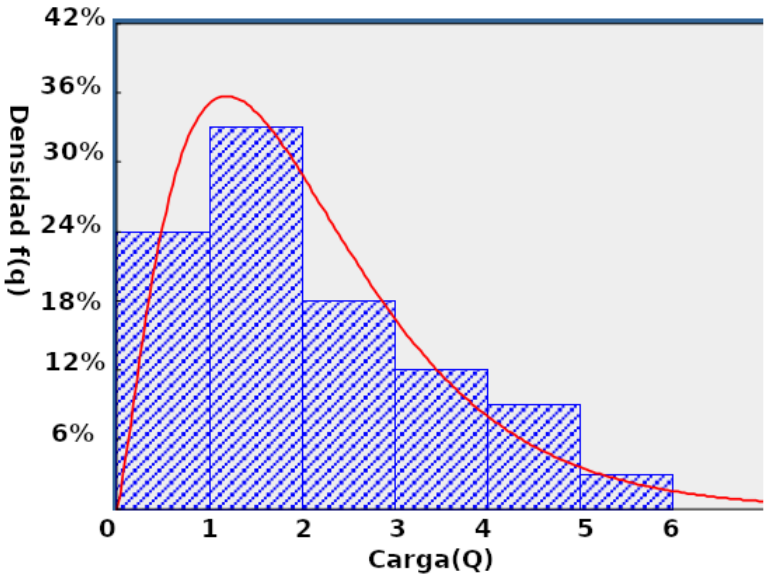

Figura 2.5: Histograma y Función de densidad de la carga(q) histórica de frutos de pepita.

La distribución de la carga es una función de la distribución Γ con el parámetro α y donde q hace referencia a la carga (Bramardi et al. (2005)). Claramente, la distribución de la carga de frutos no posee una distribución simétrica sino una distribución de tipo asimétrica y continua, particularmente una distribución Γ.

$$\boxed{f_q(q) = \frac{\alpha}{\Gamma(r)}(\alpha q)^{r-1}e^{-\alpha q}} \tag{2.3}$$

De esta manera, a partir de la relación histórica del peso y la carga de los frutos y la distribución de la carga de la temporada, se obtiene la distribución de los pesos de la temporada integrando las funciones anteriores. De manera que:

$$\boxed{f_p(p) = p^{-1}(\Gamma(r))^{-1}\frac{\alpha}{\beta}\left[\frac{\alpha}{\beta}(ln\gamma - lnp)\right]^{r-1}e^{-\frac{\alpha}{\beta}(ln\gamma - lnp)}} \tag{2.4}$$

Finalmente, para lograr la estimación del peso medio del fruto componente de la ecuación 2.1 se calcula mediante la esperanza de la función 2.4, entonces la $E(P)$ se define:

$$\boxed{PMF = \left(\frac{\alpha}{\alpha+\beta}\right)^{r} \cdot \gamma} \tag{2.5}$$

A diferencia de lo sucedido en Italia, los modelos estocásticos no resultaron precisos para nuestra región. Una de las posibles causas sea la gran heterogeneidad de los montes frutales de los valles de Río Negro y Neuquén. De manera que la relación peso carga hallada en la

región estaba pobremente descripta por el modelo 2.2. Por otro lado, tampoco era posible realizar predicciones de la distribución de los tamaños comerciales, esto último de gran interés para los productores frutícolas y las empresas empacadoras en general.

Método de la estimación del PMF por mediciones sucesivas o curvas de crecimiento

Dado que el método estocástico no brindaba resultados satisfactorios, se comienza el desarrollo de modelos de predicción de tamaño de frutos a cosecha en base a mediciones sucesivas durante el período de crecimiento de los frutos. Se inicia ajustando los patrones de crecimiento, en un principio de manzanas a partir de los primeros registros del crecimiento en los años 1968/1969 (Hector and Tiscornia (1968)) y luego en peras (Bramardi (1989)) pudiendo estimarse los tamaños de los frutos a cosecha de manera anticipada en cualquier momento del ciclo de crecimiento (Bramardi et al. (2006)). Los patrones de crecimiento permitían estimar con gran precisión los tamaños medios a cosecha (PMF en la ecuación 2.1) y por ende también de la producción. Además, se simulaban las estrategias de cosecha de la región, logrando la predicción de la distribución de los tamaños de los frutos de los cultivares más importantes en peras y manzanas (Tassile and Giménez (2013)).

A partir del año 1995 se comienza a implementar el nuevo método en la predicción de tamaño de los frutos, basados en los modelos de predicción de tamaños de frutos a cosecha a partir de mediciones sucesivas. A diferencia del método estocástico, el método de curvas de crecimiento requería un importante desarrollo y puesta a punto, construyendo curvas de crecimiento desde septiembre a marzo para lograr el conocimiento del crecimiento completo de los frutos en los cultivares más importantes de peras y manzanas. El desarrollo de de los nuevos modelos implicaba mayores costos en la puesta a punto del método de curvas de crecimiento. Además, requería para su procesamiento el relevamiento de la carga frutal y los diámetros de los frutos en un determinado momento del crecimiento del fruto, en general a los 80 días de crecimiento del fruto. El método de curvas de crecimiento se adecuaba de mejor forma a la heterogeneidad de las plantaciones de peras y manzanas característica del Valle de Río Negro y Neuquén.

En el año 1996 se retoma el estudio de curvas de crecimiento en las variedades de manzanas. Es decir, se desarrollan las curvas de crecimiento de las restantes variedades para lograr un pronóstico integral de los frutales de pepita a nivel regional. Al igual que en peras, para manzanos *"Red Delicious"* y *"Granny Smith"* se ajustan modelos no lineales de la familia logística. Se suma el crecimiento de frutos de carozo y bayas. Para el caso de los frutos de nectarines y ciruelas se ajustaron modelos doble sigmoideos (Alvarez et al. (1996)).

En el 2002 se comienza el estudio de las características de las parcelas productivas del Alto Valle mediante el empleo de las técnicas multivariadas con el fin de estratificar y mejorar el muestreo. Al mismo tiempo, se desarrollaron modelos no lineales mixtos con la estimación de incertidumbre en los párametros y se probaron modelos de crecimiento incluyendo covariables del ciclo de crecimiento y temperaturas, en particular índices de crecimiento basados en

umbrales biológicos (Alvarez et al. (2002)).

En al año 2005 se implementa definitivamente la estimación del PMF por el método de las curvas de crecimiento y se deja de lado la metodología de la estimación por cargas, métodología estocástica. Ese mismo año se realiza el censo agrícola rionegrino llamado ”CAR2005“, dicho censo mejora notablemente la base de expansión del pronóstico. La implementación de la nueva metodología permitió hacer una estimación de la distribución de los frutos en tamaños comerciales a cosecha tal cual se había previsto(Tassile and Giménez (2013)).

En el año 2006 se publican las tablas de raleo, producto del estudio y desarrollo de las curvas de crecimiento de peras y manzanas de la región, ajustadas por la cátedra de Bioestadística y Matemática de la Facultad de Ciencias Agrarias de la Universidad Nacional del Comahue (Bramardi et al. (2006)). Las tablas de raleo poseen tabulados los diámetros de los frutos desde momentos tempranos del desarrollo hasta días posteriores a la cosecha comercial, referenciados a tamaños comerciales. Las mismas permiten conocer a partir del diámetro del fruto referenciados a un tamaño comercial, en cualquier del crecimiento del mismo, el tamaño a cosecha de los frutos. De esta manera se pueden extraer con mayor antelación, los frutos pequeños que no alcancen el tamaño comercial deseado a cosecha mediante una tarea cultural denominada raleo. En el mismo momento, se desarrollaron las tablas de raleo para los principales cultivares de peras *“William’s”* , *“Packham’s Triumph”* y manzanas *“Red Delicious”* y *“Granny Smith”* .

La investigación de curvas de crecimiento continuó y su desarrollo se extendió a otros cultivares, por ello en 2008 se incluye la variedad de manzana Gala y sus clones; una variedad que fue incrementando su área cultivada y su relevancia, por eso la necesidad de incorporarla en la estimación del volumen final. Este cultivar presenta una serie de particularidades: en primer lugar gran diversidad de clones mejorados como *“Gala”* , *“Royal Gala”* , *“Mondial Gala”* y *“Galaxy”* y además suelen estar implantados en una variada gama de portainjertos. La combinación portainjerto-variedad podía lograr una entrada en producción más temprana y por eso su diversidad. El crecimiento de los frutos está influenciado por la interacción clon-portainjerto, particularmente en la distribución de tamaños comerciales a cosecha. La variedad *“Gala”* tiene un ciclo de crecimiento corto si la comparamos con otras variedades de manzana, y es muy afectada por las temperaturas primaverales. Es un cultivar cuyos tamaños comerciales dependen mucho de las prácticas culturales como poda, riegos y principalmente raleo(Tassile et al. (2013)).

En el año 2009 se comenzó a cuantificar la calidad de las estimaciones, cotejando el volumen estimado por el pronóstico con el total de producción que registraban los distintos canales de comercialización. Para comparar las distintas temporadas presentadas en la tabla 2.2 se consideraron los registros de volúmenes exportados, el volumen comercializado dentro del país, la fruta destinada a procesamiento por distintas industrias, principalmente la industria juguera y una estimación aproximada del consumo interno realizada por el equipo de la SEFRN (Secretaría de Estado de Fruticultura de la provincia de Río Negro).

Tabla 2.2: Tabla comparativa de la producción estimada por el pronóstico respecto a la producción total registrada por SEFRN (en miles de toneladas)

	2001		2002		2003		2004		2006		2007		2008	
	Prono	SEFRN	Prono	SEFRN	Prono	SEFRN	Prono	SEFRN	Prono	SEFRN	Prono	SEFRN	Prono	SEFRN
Total	1.844	1.523	1.500	1.185	1.648	1.339	1.493	1.283	1.585	1.521	1.567	1.531	1.525	1.436
Manzana	1.366	976	1.045	687	1.139	816	1.087	796	949	899	946	905	923	836
Pera	478	547	456	498	509	523	407	487	637	622	622	626	603	600

Como se observa en la tabla 2.2, para el caso de las peras los volúmenes estimados por el método del pronóstico de producción fueron, en las temporadas analizadas, muy próximos a los volúmenes registrados por la SEFRN. En tanto que, en manzanas los volúmenes estimados por el pronóstico de producción fueron siempre superiores a los registrados por SEFRN. En este punto, la diferencia se puede asumir por una sobreestimación del método o como volumen de producción que, o bien no se retira de las parcelas debida a daños en la producción o bien no se comercializa por canales de comercialización que sean registrados por la SEFRN y el SENASA. En general, la información de interés tanto para los productores como para las distintas empresas se expresa en términos de variación anual de la producción total y se calculan las variaciones interanuales como diferencia de una temporada respecto a la anterior en relación al volumen de la temporada precedente (Tassile et al. (2013)).

En la tabla 2.3 se presentan las variaciones anuales de volumen de producción estimado por el pronóstico y las variaciones porcentuales a partir de los volúmenes registrados por la SEFRN. Por ejemplo, para el año 2002 (ver tabla 2.3) la variación anual del total de la producción según el pronóstico fue de $-18,6\ \%$,es decir, que la producción del año 2002 sufrió una merma de $-18,6\ \%$ respecto al año 2001. Este valor se calcula a partir de los valores de producción total para el año 2002 que fue de $1.500.000\ tn$ menos el volumen total del año 2001 de $1.844.000\ tn$ respecto el volumen total del año 2001. Para el ciclo 2002 la variación del pronóstico y la variacion del SEFRN ambas resultaron negativas con un error estimado, suponiendo correcto el registro y estimación del SEFRN, de $3,6\%$. En los ciclos productivos sucesivos las tendencias arrojadas por el pronósticos siempre estuvieron acorde a las obtenidas por el SEFRN de manera que el error promedio desde el ciclo 2002 al 2008 fue del $3,46\%$.

Tabla 2.3: Variaciones porcentuales anuales respecto al año anterior y error promedio de estimación según el pronóstico de producción (celda en verde) y SEFRN

	2002		2003		2004		2007		2008	
Total	$-18,6$	$-22,2$	$9,8$	$13,0$	$-9,4$	$-4,2$	$-1,12$	$-0,66$	$-2,7$	$-6,2$
Error	$3,6$		$3,2$		5.2		$1,78$		$3,5$	
	$3,46$									

Si bien en este período los errores de estimación en las variaciones interanuales fueron aceptables, posteriormente se observaba que las diferencias se hacían cada vez mayor, donde

además de la falta de censos agropecuarios actualizados, insumo fundamental del pronóstico de producción, los modelos de crecimiento no contemplaban el efecto climático sobre las curvas (Tassile and Giménez (2013)).

2.2 Crecimiento de los frutos

2.2.1 Descripción y caracterización botánica de los cultivos

La implementación del método de mediciones sucesivas(ver sección 2.1.3) para estimar el peso medio del fruto en el pronóstico de producción, requiere el estudio de los patrones de crecimiento de los diferentes cultivares, el patrón de crecimiento se determina a partir de diversos aspectos fisiológicos y agronómicos que es necesario conocer.

Los manzanos y perales son especies leñosas de hojas caducas originarias de climas templados, se caracterizan por ser plantas que tienen una fase inicial en vivero de dos a tres años y luego son extraídas e implantadas definitivamente en la parcela comercial, alcanzando la plena producción aproximadamente a los cinco o seis años dependiendo especie, cultivar y portainjerto. Las árboles frutales de las parcelas productivas son individuos bimembres conformadas por dos secciones: el “pie” o “portainjerto” correspondiente a la parte radicular, a partir del cual se busca control vegetativo y acelerar el comienzo de la producción, conjuntamente con resistencia a plagas y enfermedades del suelo; y la sección del injerto que corresponde a la parte área también denominado “cultivar” o “clon” y es la parte productiva y comercial. A lo largo de esta tesis, se refiere como variedad o cultivar indistintamente, no obstante, es menester definir el cultivar como conjunto de individuos que por sus caracterısticas de cultivo y mercado, época de cosecha, forma, color y tamaño de la fruta, crecimiento de la planta, destino de la producción, etc. concuerdan entre sı y con sus descendientes. En tanto, el clon se designa a los individuos que dentro de un cultivar presenten pequeñas diferencias en sus caracterısticas de cultivo o mercado como por ejemplo fruta de mejor color o plantas de menor porte y precocidad. Esto ocurre comúnmente en cultivares de manzano y en general estas pequeñas diferencias se dan por efecto de mutaciones somáticas.

Botánicamente, peras y manzanos corresponden a los géneros *Pyrus* y *Malus* respectivamente, son miembros de la familia Rosacea, pertenecen a la subfamilia Amygdaloideae, tribu Pyreae y subtribu Pyrinae. La familia Rosacea incluye especies que son productiva y económicamente muy importantes como duraznos, pelones, cerezas, ciruelas, almendras, damascos, etc. y las ya mencionadas peras y manzanas (Yamamoto and Terakami (2016)). Existen al menos 28 especies del género Pyrus, entre las cuales cuatro especies son utilizadas por sus frutos comestibles: pera japonesa *P. pyrifolia Nakai*, pera europea *P. communis L.* y pera china *P. bretschneideri Rehd. y P. ussuriensis Maxim.*. El género Malus comprende un variado número de especies, se estima entre 25 a 50, sin lugar a dudas, la especie más representativa es *Malus X domestica Borkh*.

En los últimos años se ha logrado secuenciar tanto el genoma de las peras como el de

las manzanas: mientras que las peras poseen un genoma de $512\ Mb$ en manzanas se han secuenciado $603,9\ Mb$. Se estima una cantidad de 42.800 genes putativos en peras y 57.386 en manzanas (Jung, Sook et al. (2019)), dispuestos en $x = 17$ cromosomas, en general las peras y las manzanas son funcionalmente diploides aunque en estas últimas depende de la especie. La diferencia en el tamaño del genoma entre peras y manzanas se explica principalmente por la presencia de secuencias repetitivas, aunque la región génica y los genes que codifican proteínas son similares en ambas especies. Se han encontrado 1219 genes en pera europea, que son únicos de la especie y no se han repetido en el genoma de ninguna otra especie estudiada. Las especies de la tribu Pyreae se caracterizan por una flor de ovario ínfero, gineceo de dos a cinco carpelos y un fruto indehiscente denominado pomo, donde el ovario es expandido y vulgarmente llamado “corazón” y por otro lado, el cortex o hipanto que es la parte comestible del fruto (Yamamoto and Terakami (2016)).

2.2.2 Aspectos fisiológicos del crecimiento de los frutos

El proceso de desarrollo del fruto, tiene características comúnes en muchas especies y resulta en la expansión del tejido próximo a la semilla de manera coordinada con el desarrollo de la semilla. En etapas tempranas del desarrollo, el tejido del fruto experimenta varios ciclos de división celular, seguido por la expansión celular durante la cual los frutos almacenan metabolitos y energía en forma de almidón o azúcares (Gillaspy et al. (1993)). Luego de que las semillas maduran, la fruta experimenta una serie de cambios bioquímicos que convierten el almidón en compuestos más disponibles, como los azúcares; además de producir metabolitos secundarios volátiles que podrían actuar como atrayentes para animales o insectos que dispersan la semilla.

La regulación del crecimiento de los frutos está a cargo de sustancias llamadas biorreguladores o reguladores de crecimiento vegetal. Los reguladores de crecimiento como las auxinas, las citoquininas y las giberelinas(GA), son sustancias orgánicas no nutritivas que, en bajas concentraciones, promueven, inhiben o modifican los procesos de desarrollo de la planta. Diversos procesos de síntesis y regulación ocurren frecuentemente en el fruto y especialmente en el embrión y las semillas. En el desarrollo normal del fruto, el embrión o la semilla controla la división celular sostenida, esto se ve reflejado en que el tamaño y el peso final del fruto están frecuentemente correlacionados por el número total de semillas, además la presencia de múltiples óvulos fertilizados induce rápidamente el crecimiento del ovario para convertirse en fruto (Gillaspy et al. (1993)). Este proceso se atribuye a reguladores como las citoquininas puesto que a menudo se registran altos niveles en las semillas en desarrollo al momento de gran actividad de división celular en los tejidos que las rodea. Otro grupo de fitohormonas como las auxinas son responsables del incremento en la expansión celular en el tejido, siendo en la mayoría de los frutos la concentración de auxinas mayor en las semillas que en los tejidos. Las auxinas causan un incremento en la extensibilidad de las paredes celulares e inducen la absorción y retención de agua y soluto en la células. Los niveles de auxinas registran dos picos a lo largo del desarrollo del fruto: un primer pico en la postantesis y coincide con la iniciación de

la expansión celular, y en algunos frutos, un segundo pico a final del desarrollo del fruto que coincide con la fase final de desarrollo del embrión.

En el caso de las GAs, participan en los procesos de la división y expansión celular además se encuentran en las semillas y desencadenan procesos de germinación. Se ha observado especialmente en el desarrollo del tomate que se registran dos picos de acumulación de la hormona, uno en la división celular y otro en la plena expansión celular (Gillaspy et al. (1993)). En manzanos, la acumulación de GA se lleva a cabo en las semillas alcanzando su concentración máxima de seis a diez semanas posteriores a la plena floración y decrece rápidamente hasta desaparecer conforme madura la semilla (Manabu et al. (2008)). En frutos partenocárpicos, la acumulación de giberelinas es mucho mayor en las primeras fases de la división celular pero mucho menor en la fase de expansión, esto puede explicar porque en general los frutos partenocárpicos tienen tamaños más pequeños. En manzanos del cultivar “Fuji” y “Ohrin” la aplicación exógena de giberelinas antes de la floración indujo frutos partenocárpicos (Manabu et al. (2008)). En el mismo ensayo la combinación de giberelinas y citoquininas produjo además un estimulación en el crecimiento de los frutos, indicando que los frutos partenocárpicos tienen, además una deficiencia en citoquininas que lleva a un fruto de menor tamaño. Asimismo, su aplicación exógena retrasa la caída natural de los frutos aumentando la retención de los mismos.

Entre otras funciones que cumplen las giberelinas dentro del complejo proceso de desarrollo del fruto, juega un rol esencial en el continuo fuente-ruta-destino. La partición de asimilados a un órgano destino incrementa la llamada fuerza del destino. La “fuerza del destino” es definida como la habilidad competitiva de un órgano para recibir o atraer asimilados, y este proceso está regulado por fitohormonas. La distribución sistemática de los fotosintatos es lo que se conoce como partición de asimilados y es el mayor determinante del crecimiento y la productividad de la planta.

La GA estimula el transporte de nutrientes aumentando la descarga del floema o actuando sobre el metabolismo y compartimentación de la sacarosa y el sorbitol (Iqbal et al. (2011)). Algunos autores mostraron que en peras cv “Kousoi” con aplicación exógena de GA aumentaron considerablemente el área seccional del pedúnculo de los frutos con un marcado incremento en el área seccional del floema y asimismo del xilema, aunque este más tardío en el ciclo de crecimiento (Zhang et al. (2005)). Como se mencionó anteriormente las giberelinas estimulan tanto la división celular como la expansión celular lo cual resulta en mayores sitios para la colocación de asimilados e incrementa la acumulación de materia seca (Gillaspy et al. (1993)). En ensayos realizados en el cultivar Kosui se observó que la aplicación exógena de GA a 40 Días después de antesis producía, efectivamente frutos de mayor peso fresco. Este aumento se debía a un incremento en el contenido de agua ya que la materia seca sólo representaba un $5\ \%$ del aumento total del peso fresco del fruto (Zhang et al. (2005)).

El desarrollo y el crecimiento del fruto es dependiente de la fijación de dióxido de carbono en las hojas y la traslocación a las células del fruto. Durante la etapa temprana del crecimiento del fruto, debido a su alta actividad metabólica y división celular resultan destinos activos o

de utilización. Mientras que, durante la fase de expansión celular desarrollo y maduración de la semilla los frutos acumulan altos niveles de carbohidratos y los frutos funcionan como destinos de almacenamiento. En definitiva, las GAs aumentan el potencial de la fuente y destino estimulando las enzimas fotosintéticas, aumentando el área foliar para una mayor intercepción lumínica de la radiación fotosintéticamente activa y mejorando la eficiencia de uso de los nutrientes (Iqbal et al. (2011)).

En ensayos realizados con ^{13}C sobre peras *"Kousoi"* se demostró que los frutos tratados con GA aumentaban significativamente la tasa específica de fijación de ^{13}C en el fruto, es decir, que las GA utilizadas ejercían un efecto en la fuerza del destino (Zhang et al. (2005)). De acuerdo a estos autores, la acción principal de la GA es el efecto ejercido sobre los frutos como fuerza de destino más que la vascularización de los tejidos del pedúnculo o la tasa fotosintética. La mayoría de los genes biosintéticos de esta hormona han sido identificado y la expresión de ellos depende de condiciones ambientales y del desarrollo. Por otro lado, la giberelinas interactúan con otras fitohormonas como el ácido abscísico, ácido salicílico y el ácido jasmónico. El ácido abscísico tiene un pico de acumulación en la expansión celular y luego comienza a descender hasta el final del desarrollo del fruto, se considera que su presencia regula la desecación de la semilla e induce a la dormancia para prevenir una germinación precoz.

A nivel molecular, en estudios de expresión génica de manzanos cultivar *"Royal Gala"* aplicando "microarrays" muestreando en ocho momentos del desarrollo desde la polinización hasta la madurez del fruto, se observaron cambios significativos de la expresión génica en 1955 genes (Janssen et al. (2008)). En dicho estudio se obtuvieron cuatro patrones mayores con expresión génica coordinada, se identificó un grupo de genes con expresión en estadío de brotes florales pero inhibidos a lo largo del desarrollo del fruto; un segundo grupo de genes expresados durante estadíos tempranos del desarrollo e inhibidos posteriormente y dos grupos de genes que se expresaron en el estadío medio y en la maduración. En un análisis funcional posterior de los cuatro patrones encontrados, se destaca que la proporción de genes de función metabólica y energética es alta en las yemas florales pero disminuye en el desarrollo para luego volver a incrementarse en la madurez. El incremento tardío podía observarse en una mayor expresión de genes de metabolitos secundarios como compuestos aromáticos que se presentan en la madurez del fruto. Tambien se encontraron perfiles de expresión de enzimas metabólicas del almidón involucradas tanto en el almacenamiento como en su degradación. No obstante, las enzimas del metabolismo del almidón se regulan transcripcionalmente en el fruto, esta regulación según otros autores (Smith,Steven M et al. (2004)), podría darse tanto en tejidos del frutos como en tejidos de hojas, es decir, tanto en órganos fuente como en órganos destino proceso fundamental para coordinar la partición de carbohidratos a través de la planta.

2.2.3 Aspectos agronómicos del crecimiento de los frutos

En el desarrollo de frutos de pepita se distinguen 3 fases(ver Figura 2.6), la fase 1 o período de crecimiento exponencial asociado a la multiplicación celular; la fase 2 o período de crecimiento

lineal en el que predomina el agrandamiento y expansión celular y la fase 3 o etapa final correspondiente al período de maduración de los frutos. Un fruto pequeño que posee un menor número de células en la primera fase, será pequeño al final del ciclo aún si se dan todas las condiciones favorables para su desarrollo. El tamaño final del fruto depende del número de células al momento del cuaje, del número posterior de divisiones celulares y de la expansión celular.

Los primeros estudios en morfología y citología en el desarrollo de frutos de manzano, mostraron que la división celular es completada dentro de los 35 a 45 días posteriores a la antesis, registrando al menos cuatro ciclos de divisiones celulares incrementando diez veces del número de células. Luego del cese de la división de células meristemáticas, el crecimiento del fruto es caracterizado en una primera instancia por un proceso de vacuolización de las células, posteriormente por un rápido incremento en el tamaño de células individuales y finalmente por un acelerado proceso de desarrollo de los espacios intercelulares (Warrington et al. (1999)).

El proceso de división celular predomina temprano en el desarrollo del fruto por ello se considera que los primeros días de la fase I del crecimiento (ver figura 2.6) son cruciales para el tamaño final a cosecha, este condiciona el patrón a lo largo de todo el ciclo.

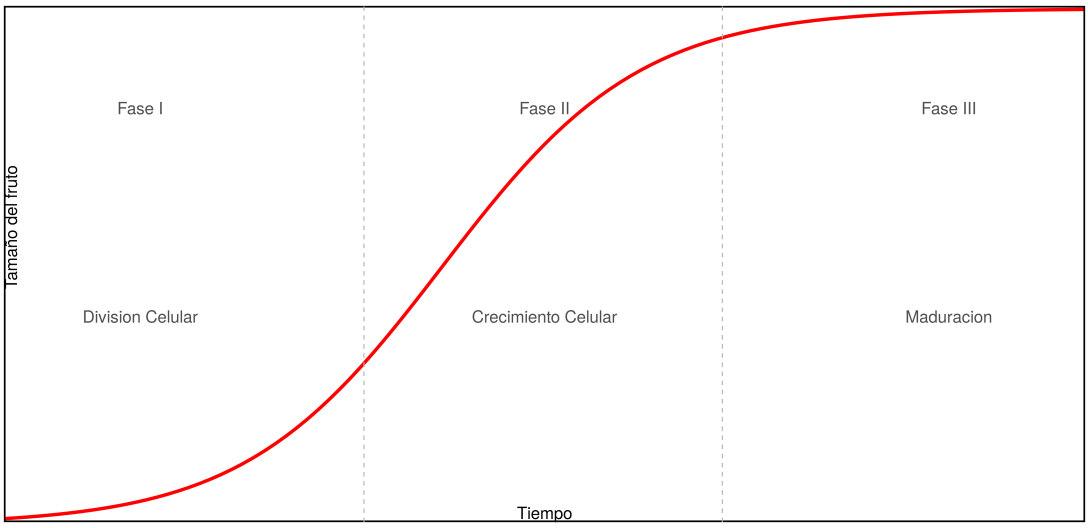

Figura 2.6: Fases del crecimiento típico de frutos de pepita desde el cuaje hasta la madurez

En manzanos del cultivar “Royal Gala” se ha observado la división celular entre la plena floración y 35 ddplf (días después de plena floración) , a partir del cual se registra un incremento en la expansión celular y el inicio en la acumulación de almidón en el fruto. La expansión celular entre los 20 ddplf y la maduración del fruto, con un pico de máxima expansión en los 60 ddplf y de máxima acumulación de almidón, cabe resaltar que entre 25 y 35 ddplf se observan tanto procesos de división como de expansión celular. Algunos autores (Zhang et al. (2006)), estudiaron la división celular sobre 46 cultivares comerciales y tres cultivares naturales de *Pyrus pyrifolia* con variadas características en tamaño del fruto y largo del ciclo. Observaron que el número de células en el mesocarpo al momento de polinización no mostró correlación con

el peso final del fruto, mientras que se calculó una correlación de $r = 0.76$ entre el número de células al momento de cosecha y el tamaño en fresco de los frutos al final del ciclo. Esto evidencia por un lado, que el número de células es determinante en momentos posteriores a la floración y el cuaje y por otro lado que, a mayor número de células mayor tamaño de fruto al final del desarrollo.

Otro aspecto del crecimiento de los frutos es el alargamiento o elongación celular asociado a la fase II del crecimiento en frutos de pepita (Figura 2.6). Si bien está estipulado que el alargamiento celular, luego del período de división es una característica determinante del tamaño de los frutos a cosecha, trabajos sobre *Pyrus pyrifolia* (Zhang et al. (2006)) no encontraron relación entre el largo de la célula, como medida de la elongación celular, y el tamaño de los frutos a cosecha. Es decir, frutos de gran tamaño tenían una longitud de células similar que los frutos de pequeño tamaño, esto implica que el tamaño de los frutos está condicionado directamente al número de células y no al tamaño de las mismas. En la región del Alto Valle, estudios realizados sobre frutos de la especie *Pyrus communis* cv. 'Williams', definieron que el área de las células al comienzo de la temporada era de $10,44\ \mu m^2$ alcanzando los $120\ ddplf$ $210\ \mu m^2$, este incremento se logró describiendo un patrón exponencial (Zon et al. (2011)) a lo largo del tiempo de estudio.

El patrón de crecimiento y el tamaño del fruto al final del ciclo también está influenciado por la edad o duración del ciclo de crecimiento del fruto y este ligado estrechamente a la variedad y el cultivar. En general, los cultivares de maduración tardía tienden a tener frutos de mayor tamaño que aquellos de madurez temprana (Figura 2.7).

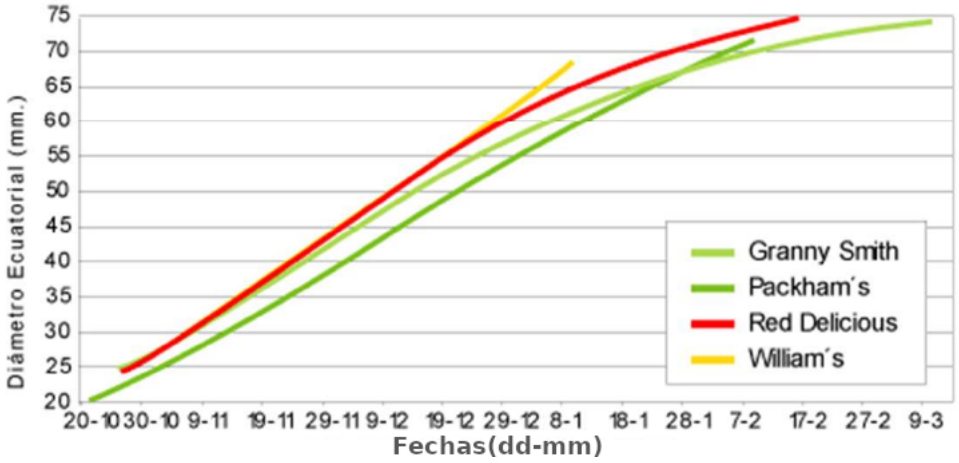

Figura 2.7: Patrones de crecimiento de las variedades de peras "Williams" y "Packham's Triumph" y de manzanas "Red Delicious" y "Granny Smith" del Alto Valle

En peras de la especie *Pyrus pyrifolia*, se obtuvo una correlación de $r = 0,84$ entre los días del período de maduración y el peso fresco del fruto al final del mismo. Esto implica que los cultivares de ciclo de crecimiento más largo poseen mayor tamaño final (Zhang et al. (2006)).

En el Alto Valle el cultivar de pera *"Williams"* cuyas edad histórica promedio del fruto es

de 105 días (Bramardi et al. (2006)) tienden a ser de tamaños pequeños y experimentar a cosecha severos inconvenientes en alcanzar el calibre deseado para su recolección, mientras que el cultivar de peras *"Packham's Triumph"* de ciclo más largo, alrededor de los 135 días (ver figura 2.7), posee frutos que alcanzan tamaños medianos a grandes presentando en algunas temporadas frutos con exceso de crecimiento. En manzanas se observa un comportamiento similar, en cultivares de ciclo corto como *"Royal Gala"* los frutos se presentan de tamaños pequeños mientras que en cultivares de ciclo largo como *"Granny Smith"* los frutos alcanzan tamaños más grandes. En la figura 2.7 se ve reflejado el patrón característico de los cuatro cultivares más destacados del Alto Valle, en peras los cultivares *"Williams"* y *"Packham's Triumph"* y en manzanas los cultivares *"Red Delicious"* y *"Granny Smith"* , en estos gráficos se toma como inicio del ciclo la fecha de plena floración histórica y marcando el final del ciclo la fecha histórica de cosecha comercial.

Desde un punto de vista agronómico el crecimiento de los frutos y el tamaño final a cosecha es definido por la interacción de factores genéticos, ambientales y por las prácticas culturales realizadas en el monte frutal (Bound (2005)). El ambiente en el cual el fruto crece, atenúa este potencial, y factores como la luz y la temperatura son esenciales para entender el crecimiento del fruto. En ensayos realizados en manzanas del cultivar *"Braeburn"* , *"Delicious"* , *"Golden Delicious"* y *"Fuji"* bajo condiciones controladas y carga frutal regulada, se estimó que la tasa de expansión de los frutos medida en $mm.d^{-1}$, de los $10\ a\ 40\ ddplf$ tenían una relación lineal significativa con la temperatura diaria. Se encontró, para un rango de temperaturas de $6\ a\ 20°C$ la tasa promedio expansión del diámetro del fruto se incrementaba de $0.062\ a\ 0.075\ mm.d^{-1}$ por cada grado de temperatura (Warrington et al. (1999)). Del mismo estudio se desprende que, mientras las temperaturas en los días próximos a la floración, afectan principalmente la división celular y la tasa de crecimiento, las temperaturas próximas a cosecha afectan directamente la madurez de los frutos. No obstane, cuando las temperturas fueron inferiores con régimenes de máximas y mínimas de $9\ y\ 3°C$ y $13\ y\ 3°C$ hasta los $40\ ddplf$ se observó una extensión del período de división celular, y luego de colocarlo en condiciones ambientales normales un crecimiento compensatorio. A pesar del crecimiento compensatorio, los frutos de "Delicious" sometidos a régimenes de temperatura superiores de $25\ y\ 15°C$ durante los $10\ a\ 40\ ddplf$ manifestaron el doble del tamaño del fruto en peso que los sometidos a régimenes más frescos.

En base a registros de crecimiento de los frutos y de las temperaturas ambientales se establecieron como umbral para el desarrollo en manzanas $14°C$, es decir, que en las temporadas de mayor cantidad de horas por encima del umbral se registraban frutos de mayor tamaño (Bergh (1990)). La importancia de las temperaturas al comienzo del ciclo de crecimiento se basa en que en primavera, las plantas frutales poseen reservas de carbohidratos y nutrientes, los frutos ven restringido su tamaño por la temperatura, durante el verano las reservas de carbohidratos son limitadas y las temperaturas superan ampliamente los umbrales de desarrollo del fruto, es decir, son las temperaturas un factor limitante en la primera fase de crecimiento del fruto (Zhang et al. (2005)).

Otro aspecto agronómico fundamental en las parcelas productivas es la regulación de

la carga frutal. Si bien las plantas frutales experimentan una caída fisológica de los frutos denominada “June Drop” como mecanismos de autorregulación, esta no resulta suficiente a los fines de obtener frutos comercialmente aceptables (Costa et al. (2006)). Se considera que es necesario sólo de un 3% a un 5% frutos cuajados del total de flores, para alcanzar calibres comerciales. En cultivares de manzana *“Starkrim”* , *“Starkinmson”* , *“Golden Delicious”* y *“Granny Smith”* se determinó que un alto número de frutos por centímetro de circunferencia del tronco de la planta estaba estrechamente relacionado a tamaños pequeños de los frutos (Bergh (1990)). Si bien en este estudio se mostró que la temperatura hasta los 42 ddplf era un factor importante, la carga de frutos tenía un efecto dominante en el tamaño final a cosecha.

La carga frutal puede regularse mediante prácticas agronómicas como el raleo, que consiste en extraer el exceso de frutos de las plantas para mejorar y garantizar un equilibrio entre la producción y el tamaño de frutos (Costa et al. (2018)). Cuanto más temprano, en el ciclo de crecimiento se interviene mejores son los resultados obtenidos ya que mayor es el área foliar por fruto lograda, mayor la disponibilidad de fotoasimilados al fruto y por ende el tamaño alcanzado a cosecha. La práctica de forma temprana se fundamenta en que por un lado permite a los frutos que quedan en la planta incrementar su tamaño ya que estimula la división celular, por otro lado, evita la vecería o “añerismo”. Fenómeno que es muy frecuente cuando no se realizan raleos adecuados y oportunos. La vecería es el fenómeno que presentan los árboles frutales por el cual si no se interviene presentan años con alternancia de cargas, es decir, un año poseen fuertes carga de frutos, también llamado año “on” y el siguiente una cantidad escasa de frutos o prácticamente nula o año “off” (Costa et al. (2018)). Dicho fenómeno está fuertemente ligado a regulaciones a nivel de fitohormonas donde están involucradas principalmente las giberelinas (Gariglio et al. (2007)). La fuerte presencia de giberelinas inhiben la inducción de yemas florales para el año próximo. Como se señaló en la sección 2.2.2, la acumulación de GA en manzanos se realiza en la semilla entre seis a diez semanas después de plena flor, por lo tanto, el raleo debe realizarse con anterioridad a este período para lograr una carga equilibrada el año siguiente. En peras el fenómeno de vecería es menos frecuente debido a que generalmente la acumulación de GA es más tardía a los 60 ddplf por lo tanto, se dispone de mayor tiempo para la intervención en la regulación de la carga.

En general, la estrategia de raleo en las parcelas comerciales es obtener más flores de las necesarias y alcanzar un cuaje óptimo de los frutos, para compensar posibles pérdidas por condiciones ambientales adversas como por ejemplo las heladas primaverales. Los métodos más utilizados para llevar a cabo el raleo de frutos son el raleo químico y el raleo manual. El raleo manual posee la ventaja de ser una práctica precisa y de bajo riesgo en la producción dado que su realización es posterior a la ocurrencia de heladas pero el raleo manual no es práctico ni económicamente viable (Bound (2005)). No obstante, es una práctica que aún se utiliza en las parcelas comerciales con el fin de mejorar la distribución y el espaciamiento de los frutos además de ser amigable con el ambiente y muy difundida en la producción orgánica.

Dentro de las prácticas del raleo químico, suele aplicarse en post floración puesto que supone menos riesgo de pérdidas que el raleo en floración por esa razón en general es

preferido por los productores. La mayoría de las sustancias químicas utilizadas para el raleo químico corresponden a reguladores de crecimiento vegetal. Algunos productos como el ácido-naftalén-acético (ANA), son auxinas aplicadas tanto a perales como a manzanos que han tenido efectos más erráticos y poco eficientes en el control de la carga frutal. Existen raleadores como la benciladenina(BA) que han demostrado ser eficientes en la disminución de la carga frutal en manzanos, mayores tamaños a cosecha y un mayor retorno a floración. Este regulador a una dosis de $100\ ppm$, ha demostrado reducir la carga frutal eficientemente y obteniendo frutos medios de mayor peso a cosecha, aunque con una merma en los rendimientos por planta.

La BA es un regulador de crecimiento perteneciente al grupo de las citoquininas, que posee un efecto raleador y además incrementa el tamaño de los frutos de manera independiente al efecto de raleo, aumentando la tasa de división celular en los tejidos (Wismer et al. (1995)). En ensayos realizados en la región del Alto Valle, sobre peras cultivar "Williams" han demostrado que la BA es efectiva para la reducción de la carga expresada tanto en frutos por ramilletes como en frecuencia de estructuras fructíferas con un fruto y sin frutos, al mismo tiempo se destaca un aumento importante en la frecuencia de unidades de mayor tamaño comercial (Giménez et al. (2010b)). Estudiando el crecimiento de frutos de pera *"William's"* con aplicación de BA y sin aplicación se registró un aumento significativo en la tasa promedio de crecimiento en los frutos tratados respecto de los frutos sin tratamiento, traducido en mayores tamaños a cosecha (Giménez et al. (2010a)). No obstante, la eficiencia en el raleo de la mayoría de los raleadores químicos es altamente variable a lo largo de los años y en la misma temporada, esto hace difícil predecir con exactitud el mejor momento de aplicación como así también anticipar los resultados finales de la aplicación (Robinson and Lakso (2004)). Por ello, es que una estrategia frecuentemente utilizada es la aplicación de raleadores químicos en post floración y una vez evaluado el efecto del raleador químico un ajuste de la carga mediante el raleo manual.

El raleo manual es una práctica corriente en las parcelas productivas tanto del Alto Valle de Río Negro y Neuquén como de la mayoría de las zonas productivas de peras y manzanas. Para realizar un raleo manual que permita seleccionar los frutos por sus tamaños comerciales a cosecha es necesario la construcción de tablas de raleo a partir de las curvas de crecimiento de los frutos (Bramardi et al. (2006)). Las confección de las tablas dependen del cultivar de interés, la zona y otras características agronómicas por lo que no pueden ser extrapolables a otras zonas o regiones agrometeorológicas. Las tablas de raleo construidas en el Alto Valle representan el crecimiento medio de frutos referenciados por distintos tamaños comerciales de acuerdo a los envases más utilizados en el mercado. Cada uno de los tamaños comerciales hace referencia a la cantidad de unidades de un determinado peso que puede contener un envase, de manera que un tamaño comercial 100 para una envase de $18\ kg$ indica que mencionada caja requiere de 100 frutos de un peso medio de $180\ gr$. Es por ello que, las tablas de raleo requieren el ajuste de modelos de peso y diámetro para poder referenciar los frutos de acuerdo a su crecimiento medio en diámetro a los tamaños comerciales registrados en peso.

2.2.4 Modelos de crecimiento de los frutos

En base a lo expuesto en la sección anterior el desarrollo de los frutos de pepita describe un patrón sigmoideo a lo largo del ciclo, es decir, desde el cuaje hasta la madurez. Dicho patrón en forma de "S" se observa al realizar mediciones sucesivas sobre el diámetro ecuatorial de un fruto a lo largo de su ciclo de vida, definido desde la plena floración, que se establece entre el 50% y el 80% de las flores abiertas hasta la madurez del mismo. Conocer el patrón de crecimiento permite por un lado describir el desarrollo ontogénico del fruto mediante un modelo matemático, por otro lado, predecir los tamaños de los frutos a cosecha a partir de los diámetros de los frutos en cualquier momento del ciclo. Es por eso que, se ha puesto un gran esfuerzo en encontrar el modelo que mejor ajuste a cada variedad, lo cual, se ve reflejado en gran cantidad de trabajos.

En manzanos *"Empire"* y *"Golden Delicious"* en dos regiones distintas de producción(EEUU e Italia), bajo condiciones ideales de cultivo el crecimiento del fruto evaluado en peso, describe un modelo expolineal con sólo dos etapas de crecimiento (Lakso et al. (1995)). El modelo ajustado respondió a la ecuación de Goudrian-Montheit, donde la fase exponencial coincidió con el período de división celular en tanto que la fase lineal con la expansión celular.

En la región del Alto Valle, realizando muestreos destructivos en parcelas productivas de peras *"William's"* el peso del fruto describió un modelo exponencial desde el período de cuaje hasta la cosecha comercial (Zon et al. (2011)), no obstante, al considerar el diámetro del fruto coincidió con un patrón sigmoideo. En la misma región se ha utilizado un modelo logístico de tercer parametrización para describir el crecimiento de peras del cultivar *"William's"* y *"Packham's Triumph"* (Bramardi et al. (1998)), desde los $20\ ddplf$ hasta pasada la cosecha comercial. También en cultivares de manzanas *"Red Delicious"* y *"Granny Smith"* de ciclo medio y largo se arribó a patrones sigmoideos de crecimiento realizando mediciones sucesivas (Stangaferro et al. (2001), Alvarez et al. (1996)).

En peras del cultivar *"Kosui"* , algunos autores encontraron patrones de crecimiento sigmoideo a partir del peso fresco de los frutos tanto naturalmente como aplicando giberelinas (Zhang et al. (2005)). En cítricos, específicamente naranjos de variedades tardías como *"Valencia Late"* se describe asimismo un crecimiento sigmoideal con 3 fases de crecimiento similar al observado en frutos de pepita, destacándose que el naranjo tiene una longitud del ciclo mayor. En este caso utilizando criterios de no linealidad y de varianza residual se ajustó la quinta parametrización del modelo logístico (Avanza et al. (2008)).

Los modelos de crecimiento tipo sigmoideo son curvas que parten de algún punto fijo e incrementan su pendiente en forma monótona hasta alcanzar un punto de inflexión, momento en que la aceleración del proceso cambia de una velocidad creciente a decreciente, luego la función comienza a aproximarse en forma asintótica al valor definitivo (Alvarez (1999)).

El modelo más común para describir fenómenos de crecimiento es el modelo de crecimiento

logístico (Davidian (1995)), cuya derivada se puede expresar de la siguiente forma

$$\boxed{\frac{dx}{dY}\Big/ Y = k(1 - \frac{Y}{a})} \tag{2.6}$$

donde la parte derecha de la ecuación es una función lineal de Y y $k > 0, a > 0$. En la ecuación 2.6 se indica que la tasa de crecimiento en relación al tamaño actual declina linealmente, con el tamaño creciente. Integrando la función 2.6 resulta:

$$\boxed{Y = \frac{\beta_0}{1 + \beta_1\ exp(-\beta_2\ x)}} \tag{2.7}$$

En el caso del modelo expresado en 2.7, $\beta_0 = a$, $\beta_1 = k$ y β_2 es el valor en que $\beta_0/(1 + \beta_1)$ representa el tamaño cuando $x = 0$. Se debe considerar que cuando el tiempo es mayor $x \to \infty$ entonces la función se aproxima a β_0. Es decir, que β_0 caracteriza la asíntota superior de la función y conjuntamente con β_1 caracterizan el inicio del crecimiento al momento de $x = 0$. Mientras que β_2 describe el cambio de crecimiento en el tiempo.

Para este modelo, en términos de las curvas de crecimiento, los parámetros del modelo (2.7) pueden ser interpretados de manera que β_0 representa la asíntota superior asociado con el tamaño final del fruto, $\beta_1/(1 + \beta_2)$ representa la asíntota inferior y se relaciona al tamaño inicial del fruto, en tanto que, β_2 se asocia con la tasa de crecimiento promedio desde β_0 hasta β_1. La variable de respuesta y es usualmente el diámetro del fruto en tanto que la covariable x el tiempo t, medido en días. Las distintas parametrizaciones de la ecuación 2.7 genera una familia de modelos logísticos que de acuerdo a sus propiedades puede ser más adecuada para una u otra especie a describir, las más destacadas son:

$$\begin{array}{lll} y = \dfrac{1}{\beta_0 + \beta_1.e^{(-\beta_2.x)}}(L2) & y = \dfrac{1}{\beta_0 + \beta_1.\beta_2^x}(L3) & \\ y = \dfrac{1}{\beta_0 + e_1^{\beta}.\beta_2^x}(L5) & y = \dfrac{1}{1 + \beta_0.e^{(-\beta_2.x)}}(L6) & y = \dfrac{\beta_0}{1 + e^{\beta_1}.\beta_2^x}(L4) \end{array} \tag{2.8}$$

Dentro de la familia logística de parametrizaciones de la ecuación 2.8 se muestran las más frecuentemente reportadas en curvas de crecimiento donde las parametrizaciones se señalan con L en este caso. Existen otros modelos no lineales frecuentemente utilizados como el modelo de Gompertz, Morgan-Mercer-Flodin, Richards y Weibull entre otros.

La estimación de los parámetros de los distintos modelos se logra a partir de la técnica de regresión. El análisis de regresión es una técnica estadística que establece una relación funcional entre dos o más variables y explica una de ellas a través de las demás. El modelo de regresión básico es el modelo de regresión lineal que se expresa de la siguiente manera:

$$\boxed{y_i = \beta_0 + \beta_1.x_i + \varepsilon_i} \tag{2.9}$$

Un modelo de regresión es lineal cuando lo es respecto de sus parámetros y simple cuando existe sólo una variable explicatoria x. Para ajustar el modelo y encontrar las estimaciones de los parámetros se suele utilizar el método de mínimos cuadrados, el método consiste en minimizar la función respecto de los valores observados, que se deriva respecto a los parámetros y se iguala a cero obteniendo las ecuaciones normales que son lineales. El método tiene una forma cerrada y fácil de resolver. El modelo de regresión puede incluir dos o más variables explicatorias extendiendo a un modelo de regresión lineal múltiple como así también polinomios de distintos grados y otros modelos en los que se incluyan términos trigonométricos.

Claramente en un modelo lineal como el representado en la ecuación (2.9), la variable y_i se expresa como combinación lineal de los β y no así respecto de β_0, β_1 y β_2 de la ecuación (2.7). Un modelo no lineal general se expresa matemáticamente de la siguiente manera:

$$\boxed{y_i = f(x_i, \theta) + \varepsilon_i} \tag{2.10}$$

En este caso, θ corresponde a un vector de parámetros $p \times 1$ desconocidos, $f(x_i, \theta)$ cualquier función matemática no lineal, ε_i es un error aleatorio no correlacionado que se asume con ditribución normal, con $E(\varepsilon_i) = 0$ y de $Var(\varepsilon_i) = \sigma^2$. Los métodos de regresión no lineal son adecuados para analizar datos para los cuales hay un relacion funcional $f()$ empírica o teórica establecida entre la respuesta y el predictor. Cada medida se distorsiona por algún error relacionado al proceso de medición entonces la observación y_i difiere de la media esperada $E(y_i|x_i)$ por algún valor que se denota con ε_i(Ritz and Streibig (2008)).

Por lo tanto, los modelos utilizados para el ajuste de curvas de crecimiento como el indicado en la ecuación (2.7) son modelos no lineales cuya estimación no es simple de resolver, puesto que las ecuaciones normales resultantes son no lineales. Para la estimación de los parámetros se utilizan métodos iterativos que minimizan el cuadrado medio del error residual. El método frecuentemente utilizado es el de linealización de funciones no lineales en un método iterativo de Gauss-Newton. La linealización se aplica con un desarrollo en serie de Taylor de $f(x, \theta)$, respecto al punto $\theta_0' = [\theta_{10}, \theta_{20} \ldots \theta_{p0}]$ reteniendo sólo los términos lineales (Montgomery et al. (2007)).

El método para encontrar los estimadores de los parámetros requiere de valores iniciales θ_0 para lograr una solución satisfactoria, existen distintos métodos para encontrar los mismos, requiere de conocer el fenómeno y estudiar previamente el comportamiento de los datos.

En un modelo de regresión lineal simple y asumiendo que los errores tienen distribución normal e independiente con varianza constante se puede aplicar el método de máxima verosimilitud donde se comprueba que maximizar la verosimilitud es equivalente a minimizar la suma de cuadrados (Ritz and Streibig (2008)). De manera que a partir de la distribución normal se tiene:

$$L(\theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)} exp - \frac{1}{2\sigma^2} \sum_{1=1}^{n} [y_- f(x, \theta)]^2 \tag{2.11}$$

Entre los criterios de bondad del ajuste se suelen utilizar los indicadores clásicos de regresión lineal como coeficiente de determinación (R^2), significación de los parámetros mediante pruebas t análisis de la distribución de los residuales y de la varianza residual (σ^2), y medidas propias de modelos no lineales como medidas de no linealidad intrínseca (IN) y no linealidad del efecto de los parámetros (PE) (Avanza et al. (2008)). No obstante, la inferencia en regresión no lineal aún con distribución de los errores normal y cumpliendo los supuestos anteriormente mencionados, las pruebas estadísticas y los intervalos de los coeficientes no son exactos y dependen de la muestra grande o asintóticos, es decir, que las pruebas sólo son aproximadas en forma asintótica (Montgomery et al. (2007)).

La construcción de curvas de crecimiento no destructivas requiere, por cuestiones operativas, que los frutos sean seleccionados y demarcados para realizar mediciones periódicas a lo largo del ciclo de crecimiento. Un aspecto importante es determinar el número de frutos representativos para caracterizar la curva de crecimiento en un cultivar. En el Alto Valle se determinó que la selección de cinco árboles, 15 frutos por árbol distribuidos en cinco frutos pequeños, cinco medianos y cinco grandes eran los suficientemente representativos para describir el crecimiento tanto en peras como manzanas en parcela productiva típica (Bramardi (1989)). Esto debido a que la variabilidad entre frutos en un árbol era superior a la variabilidad entre árboles. Resultados similares fueron encontrados en naranjos donde el 82% de la variabilidad en el diámetro de los frutos corresponde a la variabilidad entre frutos y el 18% a la variabilidad entre árboles, de ahí es que se determinó para este caso utilizar siete árboles y 30 frutos por árbol (Avanza (2010)).

La característica distintiva es que en estos estudios se cuenta con un conjunto de observaciones tomadas secuencialmente sobre la misma unidad (en el caso de las curvas de crecimiento, el fruto). Las mediciones registradas de esta forma no son independientes, de manera que el supuesto de independencia no se cumple y los errores presentan correlaciones que deben ser contempladas en la estructura del modelo. Este tipo de mediciones brinda algunas ventajas como la eficacia en el estudio de cambios, es decir, los estudios longitudinales tienen la capacidad de separar, en el contexto de la población estudiada, los efectos dentro de cada individuo de los efectos de cohorte o de diferencias entre individuos; principalmente un investigador puede separar el efecto del tiempo (cambios de los individuos a través del tiempo) de los efectos de cohorte (diferencias entre los individuos al inicio del estudio) (Rubio (2016)).

Davidian y Giltinan(Davidian (2003)) destacan en este tipo de estudios tres particularidades:

1. observaciones repetidas de una respuesta continua sobre el tiempo,

2. variabilidad en la relación entre respuesta y tiempo u otra condición entre individuos,

3. disponibilidad de un modelo científicamente relevante que caracterice el comportamiento individual en términos de parámetros significativos que varían de un individuo a otro y que determinan la variación en los patrones de respuesta temporal.

Claramente para el caso de las curvas de crecimiento la medición de los diámetros semanales de los frutos es una medida repetida, la evolución del crecimiento a lo largo del ciclo, es decir en el tiempo, presenta un incremento en la variabilidad a medida que el fruto crece y cumple la tercera opción dado que posee un modelo ya ampliamente estudiado donde los parámetros ya explicados tienen significancia fisiológica del fruto.

Las curvas de pepita, son estudiadas sobre distintas parcelas y temporadas dada la variabilidad que existe en el manejo y el material vegetal presente en las parcelas productivas y por otro lado las temporadas de crecimiento imprimen una enorme variabilidad debido a factores climáticos. Por lo tanto, se puede concebir el estudio de las curvas de crecimiento en un modelo jerárquico donde las mediciones de los frutos integran el último eslabón de dicha cadena jerárquica. Es por esta razón que el enfoque de los modelos mixtos no lineales son apropiados para analizar datos jerárquicos, correlacionados y son utilizados ampliamente no sólo en curvas de crecimiento sino en un variado campo de investigación (Davidian (2003)). De acuerdo a los mismos autores, la aplicación de modelos no lineales mixtos permite comprender el comportamiento “típico” de los fenómenos representados por los parámetros, asimismo la medida en que los parámetros varían de un individuo a otro y conocer si parte de la variación está sistemáticamente asociada con atributos individuales. La predicción a nivel individual también puede ser de interés.

En los modelos no lineales mixtos es posible de estudiar un fenómeno desde dos niveles: nivel individual y a nivel poblacional, es decir, sujeto específico y promedio poblacional o marginal. Si bien dicha distinción no es tan importante en los modelos lineales es crítico en los modelos no lineales. El modelo mixto lineal y no lineal supone que los individuos provienen de una población y, por lo tanto, comparten características comunes.

En general los datos referidos al crecimiento de un individuo u órgano, donde la trayectoria del cambio es no lineal, son casos típicos para el ajuste de modelos no lineales mixtos donde se trata al individuo u órgano de medición como un grupo o clúster y las medidas realizadas sobre él como observaciones individuales anidadas dentro del grupo o clúster. En esta circunstancia los efectos fijos describen la trayectoria de la población de individuos en tanto que los efectos aleatorios reflejan la variabilidad entre individuos (Stegman et al. (2017)).

Las curvas de crecimiento se construyen con los datos de los frutos individuales, no obstante el objetivo es caracterizar el crecimiento del fruto para cada uno de los cultivares, es decir, el patrón de crecimiento de los frutos y los parámetros característicos, o la población de curvas de crecimiento de los frutos. Entender el patrón de los frutos en la población significa comprender cómo las curvas individuales de los diámetros y el tiempo y los parámetros que los caracterizan varían entre la población de individuos. El modelo logístico de la ecuación 2.7 y el modelo derivado del crecimiento relativo en la ecuación 2.6 pertenecen al comportamiento del fruto,

donde el modelo es una descripción teórica de los procesos biológicos que tienen lugar a lo largo del tiempo dentro de un sujeto dado.

Si se desea estudiar el comportamiento del individual del fruto entonces se denota la $j - esima$ medición en el tiempo t_j. Y el modelo de regresión 2.7 entonces queda especificado como en la ecuación 2.12.

Como en general el interés de estos estudios se centra en la población de frutos, entonces en la ecuacion 2.12 se indiza los frutos como $i = 1 \ldots m$, entonces Y_{ij} representa el diámetro del $i - esimo$ fruto al $j - esimo$ tiempo de medición donde el tiempo de medición se puede especificar como t_{ij}, $j = 1, \ldots, n_i$ donde el momento de medición puede ser distinto para cada uno de los frutos.

$$\boxed{E(y|x_j) = \frac{\beta_0}{1 + \beta_1 exp(-\beta_2 x_j)}} \tag{2.12}$$

Del razonamiento anterior, se puede afirmar que cada fruto tiene una relación de regresión de la forma 2.12; sin embargo, como el crecimiento del fruto es un proceso individual, sería de esperar que cada sujeto tenga sus propios parámetros de crecimiento $\boldsymbol{\beta}$ que gobiernen su comportamiento individual. Entonces se podría pensar en un modelo para el sujeto i dependiente de su conjunto de parámetros específico individuales del crecimiento $\boldsymbol{\beta_i} = \beta_{0i}, \beta_{1i}, \beta_{2i}$.

Los modelos mixtos pueden ofrecer para este caso la mejor alternativa, ya que proveen una flexible y potente herramienta para el análisis de datos agrupados. Donde los datos agrupados pueden incluir problemas de datos longitudinales y de medidas repetidas, diseños en bloque y datos multinivel. Otra de las grandes ventajas de un modelo mixto es modelar las correlaciones intra grupales presentes en los datos agrupados y también para conjunto de datos desbalanceados. Además de presentar mejores propiedades estadísticas que los modelos de efectos fijos, en los modelos mixtos se puede interpretar una variable observada para un individuo bajo un tiempo o covariable dada como un valor que tendría una cierta probabilidad de pertenecer a un grupo de individuos con características ambientales particulares, representado por una curva promedio (Carrero et al. (2008)).

Los modelos mixtos son denominados de esa forma ya que incorporan efectos fijos y aleatorios en su formulación. Mientras que los efectos fijos son parámetros asociados a una población entera o con determinados niveles repetibles de un factor experimental, los efectos aleatorios están asociados a unidades experimentales extraídas aleatoriamente de una población. Como refiere Stegman et al., 2017 (Stegman et al. (2017)), los modelos mixtos son utilizados cuando los datos consisten en grupos de observaciones, donde los parámetros fijos describen el patrón general de los datos mientras los parámetros de los efectos aleatorios describen los grupos específicos. Un modelo lineal de efectos aleatorios de una sóla vía de

clasificación se puede expresar de la siguiente forma:

$$\boxed{y_{ij} = \beta + b_i + \varepsilon_{ij}} \tag{2.13}$$

Donde β representa al efecto fijo y se interpreta como la media de la población y b_i es la variable aleatoria que indica la desviación del individuo i de la media de la población, en tanto que, ε es la variable aleatoria que marca la desviación de la observación j del individuo i de la media del individuo i. Es decir que la ecuación 2.13 se puede interpretar como :

$$\boxed{y_{ij} = \overline{\beta} + (\beta_i - \overline{\beta}) + \varepsilon_{ij}} \tag{2.14}$$

Como ya se mencionó β corresponde a la media poblacional $\overline{\beta}$ y las desviaciones respecto de la media que se traducen como $\beta_i - \overline{\beta}$. Se le llaman "efectos" aleatorios porque representan una desviación de la media general. Asumiendo en una primera instancia con datos independientes y varianza constante entonces:

$$\boxed{b_i \sim \mathcal{N}(0, \sigma_b^2) \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)} \tag{2.15}$$

Un aspecto interesante de los modelos mixtos es que, las observaciones de un individuo i comparten el mismo efecto b_i por lo tanto, inducen la correlación. Entonces, la covarianza entre observaciones de un mismo individuo σ_b^2 corresponden a la correlación de

$$\boxed{\sigma_b^2/(\sigma_b^2 + \sigma^2)} \tag{2.16}$$

La ventaja es que los parámetros del modelo a estimar son β, σ_b^2, σ^2, donde independientemente de los individuos o grupos de i la cantidad de parámetros a estimar siempre es tres. Se debe tener en cuenta que los efectos aleatorios $b_i,\ i = 1, \ldots, M$ para los M grupos o individuos, se comportan como parámetros pero son en realidad otro nivel de variación aleatoria del modelo por lo que no se estiman como parámetros. Aunque se pueden realizar predicciones $\hat{b_i}$ de los valores de los efectos aleatorios a partir de los datos observados (Pinheiro and Bates (2000)). Si bien la notación anterior permite dar una descripción de los modelos mixtos, la notación utilizada frecuentemente es la notación matricial de manera que el modelo completo es:

$$\begin{aligned} y_i &= \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \varepsilon_i \quad i = 1, \ldots, M \\ b_i &\sim \mathcal{N}(0, \sigma_b^2) \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \end{aligned} \tag{2.17}$$

En este caso, $\mathbf{X}$ indica la matriz de diseño de los efectos fijos con vectores fijos β, en tanto que, $\mathbf{Z}_i$ indica la matriz de diseño de los efectos aleatorios con b_i efectos aleatorios. Donde esta última tiene su propia distribución $b_i \sim \mathcal{N}(0, \sigma_b^2)$ independiente asimismo de $\varepsilon_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

Cuando se considera que los niveles de los efectos aleatorios son significativos para todos los individuos i, es posible modelar la estructura de la matriz de varianzas covarianzas de los efectos aleatorios. Entonces, en ese caso la especificación del modelo sería similar a 2.17 pero $b_i \sim \mathcal{N}(0, \mathbf{\Psi}) \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ y la matriz $\mathbf{\Psi}$ es definida positiva con correlaciones entre los distintos efectos aleatorios. Un aspecto importante es la predicción de la respuesta y los efectos aleatorios bajo un modelo mixto. La estimación de los valores predichos que contemple la respuesta y los efectos aleatorios es lo que denomina "BLUP" o *Best linear unbiased predictions*.

A menudo, los datos adquieren una estructura jerárquica ya sea porque surgen naturalmente de ese modo o por el diseño experimental al cual son sometidos. Como se mencionó, las curvas de crecimiento no son la excepción y requieren contemplar múltiples efectos aleatorios anidados. Para considerar dicha estructura en los datos algunos autores (Goldstein (1998)) proponen estudiar los modelos mixtos desde la óptica de los modelos multinivel. Los modelos multinivel con efectos aleatorios en dos niveles pueden ser expresados de la siguiente manera:

$$\begin{aligned} \mathbf{y}_{ij} &= \mathbf{X}_{ij}\beta + \mathbf{Z}_{i,j}b_i + \mathbf{Z}_{ij}b_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, M \quad j = 1, \ldots, M_j \\ b_i &\sim N(0, \mathbf{\Psi}_1) \quad b_{ij} \sim N(0, \mathbf{\Psi_2}) \quad \varepsilon_{ij} \sim N(0, \sigma^2\mathbf{I}) \end{aligned} \tag{2.18}$$

Para este caso el vector de la variable de respuesta desde el nivel mas interno de agrupamiento se escribe como $\mathbf{y}_{ij}\ i = 1, \ldots, M \quad j = 1, \ldots, M_j$ donde M es el número de grupos del primer nivel, en tanto que M_i es el número de grupos del segundo nivel dentro del primer nivel i y la longitud del vector $\mathbf{y}_{ij}$ es n_{ij}. La matriz de efectos fijos $\mathbf{X}_{ij}$ tiene una dimensión de $n_{ij} \times p$, con efectos aleatorios de primer nivel b_i con dimensión q_i y de efectos aleatorios de segundo nivel b_{ij} con dimensión q_2 con una matriz de diseño de efectos aleatorios $\mathbf{Z}_{i,j}$ con dimensión $n_i \times p_1$ y $\mathbf{Z}_{ij}$ con dimensión $n_i \times p_2$ donde p representa los parámetros del modelo. Se asume que el nivel de efectos aleatorios 1, b_i es independiente del nivel de efectos aleatorios 2, b_{ij} y de los errores ε_{ij} dentro de los grupos. Y la formulación expuesta en la ecuación 2.18 puede extenderse a Q niveles. Las matrices de varianza-covarianza, para una mejor solución matricial (Pinheiro et al. (2019)), suele expresarse en términos de factor de precisión relativa $\mathbf{\Delta}$ que puede ser expresado como:

$$\boxed{\frac{\mathbf{\Psi}^{-1}}{1/\sigma^2} = \Delta^{-1}\Delta} \tag{2.19}$$

En el caso de los modelos mixtos las estimaciones no pueden realizarse por medio del método de mínimos cuadrados como en el caso de los modelos de regresión clásicos sino que una estrategia es recurrir a la estimación por el método de máxima verosimilitud.

De acuerdo a algunos autores (Pinheiro and Bates (2000)), para el caso del modelo expuesto en la ecuación 2.13 la función de verosimilitud es la función de densidad de probabilidad de los datos dados los parámetros pero expresada como los datos fijos dado los parámetros, entonces

se escribe como:

$$\boxed{L(\beta, \theta, \sigma^2|y) = p(y|\beta, \theta, \sigma^2)} \tag{2.20}$$

Donde L es la verosimilitud, θ un parámetro de efectos aleatorios que contiene a Δ y p es la función de densidad de probabilidad. Ya que los efectos aleatorios b_i no son en realidad observaciones directas y se debe considerar en el modelo, entonces se debe integrar la densidad condicional de los datos dado los efectos aleatorios con respecto a la densidad marginal de los efectos aleatorios para obtener la densidad marginal de los datos. Resultando la expresión de la siguiente manera:

$$\boxed{L(\beta, \theta, \sigma^2|y) = \prod_{i=1}^{M} \int p(Y|b_i, \beta, \sigma^2) p(b_i|\theta, \sigma^2) db_i} \tag{2.21}$$

Teniendo en cuenta que la densidad condicional de y es una normal multivariada entonces:

$$\boxed{p(Y|b_i, \beta, \sigma^2) = \frac{exp(- \parallel Y_i - \mathbf{X_i}\beta + \mathbf{Z}b_1) \parallel^2 / 2\sigma^2}{(2\pi\sigma^2)^{(n_i)/2}}} \tag{2.22}$$

Para el caso de la probabilidad marginal de b_i también es normal multivariada y en este caso:

$$\boxed{p(b_i|\theta, \sigma^2) = \frac{exp(-b_i^T \mathbf{\Psi}^{-1} b_i)}{(2\pi)^{q/2}\sqrt{|\mathbf{\Psi}|}}} \tag{2.23}$$

La integración y la posterior solución deriva en una función de log verosimilitud que requiere de métodos iterativos para su estimación. Uno de los algoritmos más utilizados es el algoritmo EM cuyo primer paso calcula la esperanza y luego el paso M consiste en maximizar dicha esperanza.

El algoritmo de optimización por excelencia es el algoritmo de Newton-Raphson, no obstante, es un algoritmo computacionalmente costoso y muy inestable cuando se encuentra lejos de los valores finales. Es por eso que algunas librerías como **nlme** del software **R** aplican un híbrido de dicho algoritmo, utilizando el método EM para estimar valores iniciales y luego buscando los valores finales con el algoritmo de optimización de Newton-Raphson.

Alcanzado este punto, una vez logradas las estimaciones de los parámetros del modelo y obtenida la verosimilitud de las funciones anteriormente mencionadas mediante los métodos de optimización, se puede establecer un método de comparación de modelos: “la prueba de razón de verosimilitud”. El test permite probar términos de modelos anidados, se dice que un modelo estadístico está anidado en otro modelo cuando este representa un caso especial del otro modelo (Schabenberger and Pierce (2002)). De manera que si $L2$ es la verosimilitud de un modelo más general y $L1$ es la verosimilitud de un modelo reducido entonces $L2 > L1$ y el test

de razón de verosimilitud(LRT) se calcula de la siguiente forma:

$$\boxed{LRT = -2log(L2/L1) = -2[log(L2) - log(L1)]} \tag{2.24}$$

La distribución del estadístico de LRT bajo hipótesis nula tiende asintóticamente a una distribución χ^2 con $p_2 - p_1$ grados de libertad. El valor calculado de LRT se compara con una $\chi^2_{pL2-pL1,\alpha}$ donde $pL2$ es el número de parámetros del modelo general y $pL1$ número de parámetros del modelo reducido y α el nivel de significación. Si $LRT > \chi^2_{pL2-pL1,\alpha}$ entonces el modelo correcto es el general, en caso contrario debe seleccionarse el reducido.

En la comparación de los efectos o los términos aleatorios de los modelos mixtos suelen utilizarse también criterios como el "criterio de información de Akaike(AIC)" y el "criterio de información bayesiano(BIC)". Ambos ampliamente utilizados en la comparación de modelos se obtienen de la siguiente manera:

$$\boxed{AIC = -2logLik + 2n_{par}} \tag{2.25}$$
$$\boxed{BIC = -2logLik + 2n_{par}log(N)}$$

Si bien ambos criterios dependen del valor de verosimilitud del modelo, el AIC penaliza por el número de parámetros del modelo mientras que el BIC penaliza no sólo por la cantidad de parámetros sino por el logaritmo de la cantidad de datos de manera que este afecta más a los modelos con menor cantidad de datos. Utilizando este criterio se prefieren modelos con menor valor de AIC o BIC. En tanto, en la comparación de los efectos fijos suelen utilizarse directamente pruebas F (F de Snedecor) entre modelos anidados.

Antes de realizar la inferencia de un modelo mixto se deben verificar supuestos acerca de si la distribución que se asume es válida a los datos analizados. En estos modelos hay dos supuesto básicos que se deben chequear, el primero es acerca de los errores dentro de los grupos, que se suponen normales, identicamente distribuidos con esperanza 0 y varianza σ^2 e independiente de los efectos aleatorios. Y además, se verifica que los efectos aleatorios sean normalmente distribuidos con media 0 y matriz de covarianzas $\mathbf{\Psi}$ e independiente para los distintos grupos.

Es frecuente que al estudiar curvas de crecimiento y medidas repetidas en general se observen dos problemas en el momento de verificación de supuestos: la falta de independencia de los datos y presencia de heteroscedasticidad, es decir, un aumento de la varianza de los errores a medida que el crecimiento del individuo va alcanzando su valor final. El hecho de realizar mediciones repetidas sobre el mismo sujeto implica que no es posible aleatorizar el factor tiempo, por lo que las medidas tomadas sobre un mismo individuo están autocorrelacionadas y, por tanto, no se cumple el supuesto de independencia de los errores (Schabenberger and Pierce (2002)). Es por eso que es posible contemplar la modelación de la correlación y la heteroscedasticidad de los errores dentro de los grupos de manera que la expresión 2.17 queda

de la siguiente manera:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \varepsilon_i \quad i = 1, \ldots, M \\ b_i &\sim N(0, \mathbf{\Psi}) \quad \varepsilon_i \sim N(0, \sigma^2 \mathbf{\Delta_i}) \quad i = 1, \ldots, M \end{aligned} \tag{2.26}$$

En este caso $\mathbf{\Delta_i}$ es una matriz definida positiva de los efectos fijos o marginales y puede ser generalizada para casos de modelos multinivel. De manera que la estructura de varianza covarianza de la variabilidad entre grupos se puede descomponer en (Davidian (1995)):

$$\boxed{\mathbf{\Delta_i} = \mathbf{V_i C_i V_i}} \tag{2.27}$$

Donde $\mathbf{V}_i$ es una matriz diagonal correspondiente a la varianza y $\mathbf{C}_i$ la correlación entre los errores, de manera que podemos expresarlo como:

$$\boxed{Var(\varepsilon_{ij}) = \sigma^2 [\mathbf{V}_i]^2_{jj}, \qquad cor(\varepsilon_{ij}, \varepsilon_{jk}) = [\mathbf{C}_i]_{jk}} \tag{2.28}$$

La matriz $\mathbf{\Delta_i}$ puede descomponerse en una estructura de correlación que contiene la heteroscedasticidad y la correlación entre errores, dos aspectos importantes en la modelación de curvas de crecimiento. Algunos software muy difundidos como R utilizan dichos métodos donde, además, se han desarrollado funciones para modelar la estructura de los errores dentro de grupos utilizando covariables. Entre las funciones especificados contemplan casos de varianza fija, cuando la variabilidad de los grupos aumenta de forma proporcional las diferentes varianzas por estrato o grupo. También considera el ajuste de varianzas independientes por grupos expresadas como la razón entre la varianza del primer estrato y de l-ésimo estrato. La variabilidad se puede modelar asimismo como funciones exponenciales y potenciales entre los estratos de los grupos que se modelan.

Como se observa en la ecuación 2.28 la descomposición de la matriz $\mathbf{\Delta}_i$ contempla estructuras de correlación para modelar la dependencia entre observaciones. Generalmente, la correlación entre observaciones se deben a dos tipos de datos: datos asociados al tiempo o datos asociados al espacio. En general las estructuras de correlación dentro de cada grupo se expresan de la siguiente forma:

$$\boxed{cor(\varepsilon_{ij}, \varepsilon ij') = h[d(\mathbf{p}_{ij}, \mathbf{p}'_{ij}), \rho]} \tag{2.29}$$

Entonces la correlación de los errores, en un modelo para el i-ésimo grupo del j-ésima medición puede obtenerse a partir de una función de correlación $h()$ aplicada sobre vectores de posición $\mathbf{p}$ y donde ρ es un vector de parámetros de correlación Pinheiro et al. (2019). Este tipo de funciones está especialmente implementados en librerías de R como el caso de la librería **nlme**.

La ecuación 2.29 indica que la correlación de los errores intra grupos, depende de la posición del vector de posición a través de una distancia entre ellos que se modela mediante una función h donde ρ es un vector de parámetros de correlación, h toma valores de -1 a 1. De manera que

$h(0, \rho) = 1$ dado que dos observaciones tienen la misma posición, son la misma observación y por lo tanto tienen correlación 1.

Existen implementadas varias funciones que permiten contemplar distintas estructuras de correlación de los errores, en general, la mayoría son utilizadas para modelar la dependencia en una serie de tiempo como en el caso de las curvas de crecimiento de los frutos donde un dato es observado regularmente en el tiempo y los errores se asocian a un vector de posición uni dimensional ($\mathbf{p}$). En el contexto de los datos de series de tiempo la función $h()$ de correlación es llamada "autocorrelación" y se expresa de la siguiente forma:

$$\boxed{cor(\varepsilon_{ij}, \varepsilon'_{ij}) = h(|\mathbf{p}_{ij} - \mathbf{p}'_{ij}|, \rho)} \tag{2.30}$$

Las estructuras de correlación más frecuentemente utilizadas son: simetría compuesta, autoregresiva de media móvil, autoregresiva de media móvil con tiempo continuo, entre otras. La estructura denominada simetría compuesta, es la estructura de correlación más simple ya que admite un único valor de correlación de los errores dentro del grupos. Dicha estructura es la que se asume cuando se utiliza un modelo mixto como fue descripto en la ecuación 2.16. Una de las estructuras que más se ajusta a la problemática de las curvas de crecimiento, es la autoregresiva de media móvil. Este tipo de estructuras asumen que las observaciones se realizan en un tiempo entero determinado y donde la distancia entre observaciones es equis-espaciada llamado también *lag*.

Los modelos autoregresivos expresan la observación actual como una función lineal de las observaciones previas más un término de ruido. La función de autocorrelación autoregresiva de orden 1 es la más aplicada, se expresa de manera que la correlación decrece en valor absoluto exponencialmente con el lag como se observa en la siguiente ecuación:

$$\boxed{h(k, \phi) = \phi^k, \quad k = 0, 1, \ldots} \tag{2.31}$$

Esta función de correlación puede ser generalizada a una medida continua del tiempo dando origen a la estructura de correlación autoregresiva contínua.

En el caso de los modelos no lineales mixtos, en general se cumple los descripto hasta el momento para los modelos lineales mixtos, sólo que los modelos dejan de ser lineales y la resolución de la maximización de la función de verosimilitud resulta mucho más compleja. Los modelos no lineales mixtos se escriben de manera similar a lo visto en la ecuación 2.10 pero de una forma más general contemplando los efectos aleatorios y su correspondiente matriz de diseño en un modelo jerárquico utilizando el criterio de Laird and Ware Laird and Ware (1982):

$$\begin{aligned} y_{ij} &= f(\phi_{ij}, x_{ij}) + \varepsilon_{ij} \quad i = 1, \ldots, M; j = 1, \ldots, n_i \\ \phi_{ij} &= \mathbf{A}_{ij}\beta + \mathbf{B}_{ij}b_i, \\ b_i &\sim \mathcal{N}(0, \mathbf{\Psi}) \quad \varepsilon_{ij} \sim \mathcal{N}(0, \mathbf{\Delta_i}) \end{aligned} \tag{2.32}$$

En el caso de la ecuación 2.32 f representa una función no lineal, ϕ_{ij} un vector de parámetros dependiente del modelo, x_{ij} un vector de predictores y ε_{ij}, el vector de errores cuya distribución es normal con esperanza 0 y admite una estructura particular que relaja el supuesto de homocedasticidad y correlación de los errores. En la ecuacion 2.32 ϕ contiene a las matrices de diseño para los efectos fijos representados con A y aleatorios B, mientras que β es un vector de parámetros poblacionales y b_i es el vector de efectos aleatorios asociado al grupo i, y ψ constante en j en una matriz de varianzas covarianzas. La estimación de los parámetros del modelo se realiza más comunmente, o al menos, en los paquetes estadísticos más frecuentemente utilizados, maximizando la función de verosimilitud, que para este caso quedaría de la siguiente manera:

$$L(\beta, \sigma^2, \Psi|y) = \int p(Y|b_i, \beta, \sigma^2)p(b_i|\Psi)db_i \tag{2.33}$$

La log verosimilitud que se obtiene a partir de la ecuación 2.33 puede aproximarse al menos, mediante 4 algoritmos distintos:

- Una aproximación laplaciana modificada.
- La cuadratura Gaussiana.
- Muestreo de importancia.
- Método alternativo de Lindstrom y Bates.

El método alternativo de Lindstrom y Bates, utiliza un algoritmo que alterna entre dos pasos, un paso que aplica mínimos cuadrados penalizados y un paso que utiliza modelos lineales mixtos(LME) considerando que ambos son iterativos. Básicamente en el primer paso se estima la matriz de varianzas covarianzas de los efectos aleatorios la cual se fija y luego se estiman los efectos fijos y los aleatorios minimizando el algoritmo PNLS, cuya función objetivo a minimizar es:

$$\sum_{1=1}^{M} \parallel y_i - f_i(\beta, b_i) \parallel^2 + \parallel \mathbf{\Delta} b_i \parallel^2 \tag{2.34}$$

Luego el segundo paso, a partir de los efectos ya estimados actualiza la estimación de la matriz de varianzas covarianzas, en este caso expresado como $\mathbf{\Delta}$ basado en una serie de expansión de Taylor. Como se mencionó anteriormente el paso LME involucra el algoritmo de EM para encontrar los valores iniciales y el algoritmo de Newton-Raphson que converge más rápidamente hasta alcanzar el valor final. Pinhero y Bates, encontraron que tanto el método alternativo de Lindstrom y Bates como la aproximación Laplaciana, eran muy precisos estadísticamente y eficientes computacionalmente. Esta conclusión la obtuvieron los autores luego de ajustar dos modelos no lineales mixtos, un modelo logístico con efecto aleatorio sobre la asíntota superior

y un modelo compartimentarizado de primer orden con efectos aleatorios sobre tres de sus parámetros (Pinheiro and Bates (1995)). Otros autores han destacado asimismo la rapidez con la cual se arriba a la convergencia, no obstante, han observado que los resultados obtenidos son de menor precisión que la aproximación laplaciana (Stegman et al. (2017)).

El ajuste de modelos no lineales mixtos ha sido ampliamente difundido y permite resolver distintos problemas que surgen tanto al momento de modelar crecimiento de órganos e individuos como a nivel espacial, correlaciones generadas por efecto del suelo. Entre los ejemplos de estos modelos recientemente detallados se puede mencionar el de Carrero et al. (2008), quienes aplicaron modelos no lineales mixtos en el ajuste de índices de sitios, es decir, en la medición y seguimiento de alturas de árboles representativos de un rodal por unidad de área en el tiempo. Los mismos especificaron la asíntota superior como efecto aleatorio y al observar una fuerte correlación en las mediciones sucesivas, una matriz autoregresiva de primer orden(AR1).

En la modelación del crecimiento del diámetro de los troncos de naranjos a lo largo de 1200 días, problema clásico que gracias a la aplicación de un modelo logístico y efectos aleatorios en la asíntota superior permitieron un avance y una modelación completa del fenómeno en cuestión (Pinheiro and Bates (1995)). En el estudio del crecimiento en especial la variable altura de niños, en sus primeros tres años de vida, a partir del cual, se aplica un modelo no lineal de Jenss-Bayley asumiendo efectos aleatorios en los parámetros b_1, b_2 y b_3, permitiendo estimar el efecto del individuo en cada parámetro (Stegman et al. (2017)). Y existen un gran número de ejemplos que se puede mencionar en particular en la rama de la farmacocinética donde tuvieron estos modelos un pleno auge.

Los modelos no lineales mixtos están implementados en softwares como SAS, JULIA y R entre otros. Particularmente, en R está programado en funciones de distintos paquetes como `nlme()` del paquete **nlme**, en este caso se implementó el algoritmo de Bates et al., (Pinheiro et al. (2019)), anteriormente descripto. La función `nlmer()` del paquete **lme4** que utiliza dos alternativas para la aproximación a la función de verosimilitud, la primera una aproximación de la cuadratura Gaussiana adaptativa, cuando se especifica un efecto aleatorio y un sólo punto de cuadratura, entonces la cuadratura Gaussiana es equivalente a la aproximación de Laplace. La segunda alternativa de aproximación no integra los efectos aleatorios, de modo que estos sólo influyen en las estimaciones de los efectos fijos a través de su modos condicionales estimados. Hay tres pasos para esta optimización: mínimos cuadrados penalizado iterativamente re ponderados (PIRLS) para estimar el condicional de los efectos aleatorios, integrar los efectos aleatorios sobre sus modos condicionales, y optimización no lineal de la función objetivo (es decir, los resultados de la integración).

Por otro lado, la función `saemixModel()` del paquete **saemix**, utiliza los denominados algoritmos de aproximación estocástico EM para estimar los parámetros de los modelos de efectos mixtos utilizando una aproximación estocástica de la verosimilitud a partir de una modificación del algoritmo EM. Como se describió anteriormente el algoritmo EM es utilizado para calcular la logverosimilitud en un modelo lineal y en caso de modelos no lineales no se

logra una forma cerrada de la integral. A partir de la modificación del algoritmo EM se logra por aproximación estocástica de la verosimilitud sortear el problema de las funciones no lineales.

Existen otras librerías que implementan modelos Bayesianos como es el caso de librería **brms**, pensado especialmente para modelos multinivel (Bürkner (2017)) y contempla el caso de modelos mixtos no lineales, que permiten además tener estimaciones e intervalos de confianza de los parámetros de los distintos modelos, resultando en convergencias más rápidas cuando los modelos son muy complejos. La ventaja de utilizar un marco bayesiano en modelos no lineales, es que permite incorporar conocimientos previos en la estimación de los parámetros, como es la distribución *a priori* de los parámetros β_0, β_1 y β_2 y asimismo de la matriz de varianzas covarianzas de los efectos aleatorios. Esto se logra luego de haber ajustado la suficiente cantidad de curvas de crecimiento y estudiar el comportamiento de dichos parámetros en el conjunto de curvas.

La función `nlme` tiene mayor rapidez y facilidad de convergencia que las demás funciones de otros paquetes, permite modelar estructuras de heteroscedasticidad y correlación de los errores con mayor facilidad y, ofrece una interface más simple al momento de especificar el modelo y su estructura (Stegman et al. (2017)). Por esta razón, se ha optado su aplicación en esta tesis, en el ajuste de modelos no lineales en curvas de crecimiento.

Dado el gran volumen de datos generado a partir del estudio del crecimiento de frutos de diferentes cultivares y del pronóstico de cosecha de las regiones del Alto Valle de Neuquén y Río Negro, los métodos estudiados anteriormente, para estimar la log verosimilitud y encontrar las estimaciones a los parámetros se tornan computacionalmente difíciles. Una opción es aplicar herramientas de la teoría de aprendizaje de máquinas concebidas para este fin y en un contexto de grandes bases de datos como es el caso del proceso KDD, es por eso que se cree pertinente comenzar el desarrollo del mismo a partir de ahora.

2.3 Descripción del proceso KDD

Como se menciona en la sección 1.1, la sociedad actual está experimentando un período en el cual el avance de la electrónica, la informática y la computación, con un gran aporte de la telemática han llevado a una revolución en las comunicaciones que se traduce en la generación de enormes masas de datos. Esto se evidencia en dos efectos particulares: la tecnología basada en la computación ha permitido a investigadores recolectar enormes conjuntos de datos en órdenes de magnitud mucho mayores de lo que las teorías estadísticas han sido concebidas para analizar; estos enormes volúmenes de datos demandan nuevas metodologías y requieren innovadores algoritmos estadísticos aplicados en informática (Efron (2018)). El proceso KDD es una de las metodologías surgidas en respuesta a dicha demanda.

El proceso KDD por sus siglas en inglés *Knowledge Discovery in Data Bases* permite encontrar patrones válidos, potencialmente útiles en grandes volúmenes de información. El término “descubrimiento de conocimiento en bases de datos” fue acuñada en 1989 para enfatizar

que el conocimiento es el producto final de un descubrimiento basado en datos (Fayyad et al. (1996)). El concepto de encontrar patrones útiles en los datos ha recibido un gran número de denominaciones como minería de datos, extracción de conocimiento, descubrimiento de información, cosecha de información, etc. En tanto que, minería de datos o *"Data Mining"* DM ha sido utilizado ampliamente por estadísticos, analistas de datos y profesionales del manejo de información extendiéndose hasta el campo de las bases de datos. La tarea de aprender a partir de los datos ha sido objeto de dos disciplinas independientes pero al fin convergentes: desde la disciplina del aprendizaje de máquinas o machine learning, con énfasis en la aproximación algorítmica y los modelos estadísticos que hacen hincapié en la elección de un modelo de distribución de probabilidades de los datos observados (Ciampi (2007)).

El proceso KDD es una metodología que evoluciona de manera interdisciplinaría a partir de distintos campos de investigación como machine learning, bases de datos, estadística, inteligencia artificial, visualización de datos, etc. cuyo objetivo común es extraer información de alta calidad en el contexto de grandes conjuntos de datos. Uno de los componentes fundacionales del KDD es el DM que recibe un importante aporte de disciplinas como aprendizaje de máquinas, reconocimiento de patrones y especialmente de la estadística. Básicamente el proceso KDD hace hincapié en encontrar patrones en los datos que puedan resultar en conocimiento útil o interesante. En este proceso la estadística como disciplina cumple un rol central ya que provee un marco para cuantificar la incertidumbre de los algoritmos de aprendizaje y muy especialmente cuando se trata de inferir a partir de patrones de una muestra, patrones más generales aplicados a la población.

Por otro lado, las bases de datos cobran especial importancia dentro del proceso KDD, dado que uno de los principales problemas es que en muchos casos el gran volumen de datos no puede ser alojado en la memoria de las computadoras (al menos en memoria RAM). Por esta razón, se requiere de sistemas de gestión de datos y bases de datos que permitan acceder a los mismos de manera eficiente, ordenando y agrupando operaciones, y al mismo tiempo permitiendo optimizar consultas.

En el contexto de DM se considera dato a todo hecho conocido que puede ser registrado y que tiene un significado implícito. De acuerdo a Elmasri (2011) las bases de datos, son una colección de datos relacionados que poseen propiedades implícitas como:

- Representan algún aspecto del mundo real, algunas veces llamado **minimundo** o **universo de discurso**. Donde cambios en el minimundo son reflejados en la base de datos
- Es una colección coherente lógicamente de datos con algún significado inherente.
- Es diseñada, construida y poblada con datos con un propósito específico.

Cabe destacar que la mayoría de las bases de datos actuales son de tipo relacionales. Esto implica, el usuario percibe la información como tablas, sobre las que se aplican operadores que derivan en otras tablas (Date (2001)). Para interactuar con este tipo de bases se utiliza una serie de instrucciones que constituyen un lenguaje llamado SQL

En el mismo contexto, se refiere a patrón como un subconjunto de datos o un modelo aplicado a dicho subconjunto. Básicamente encontrar un patrón es hallar la estructura de los datos, aplicar un modelo a los datos o realizar una descripción de alto nivel sobre los mismos (Fayyad et al. (1996)). Dentro del proceso KDD la búsqueda de patrones en los datos requiere de una serie de fases repetibles, iterables e interactivas cuyas decisiones son evaluadas por el investigador. Las fases y acciones del proceso KDD se presentan esquemáticamente en la figura 2.8. Donde se observa que la primera fase consiste en comprender el dominio de los datos, constituir la base de datos y plantear los objetivos de la aplicación del proceso KDD. Luego, se establecen los datos o variables a analizar, es decir, se establece un conjunto de datos objetivo a partir de los cuales se planifica el análisis. Otro punto no menos importante, señalado en la figura 2.8, es la fase de preprocesamiento y "limpieza" de los datos. Se basa en remover el ruido de los datos, recolectando información útil y que permita poner en evidencia los mismos. En este punto se identifican los datos faltantes y se deciden estrategias para su procesamiento.

La siguiente fase involucra la reducción de los datos y la proyección. En esta fase se utilizan las variables más destacadas dado el interés y el objetivo de la aplicación del proceso o se procede a la reducción de la dimensionalidad y a la transformación de los mismos en caso que así lo requiera. Aquí se desea encontrar un número efectivo de variables para la obtención de los patrones. Estos primeros pasos suelen resumirse como preprocesamiento de los datos y en ámbitos netamente estadísticos se refiere como análisis exploratorio de datos o EDA[1] (Behrens (1997)).

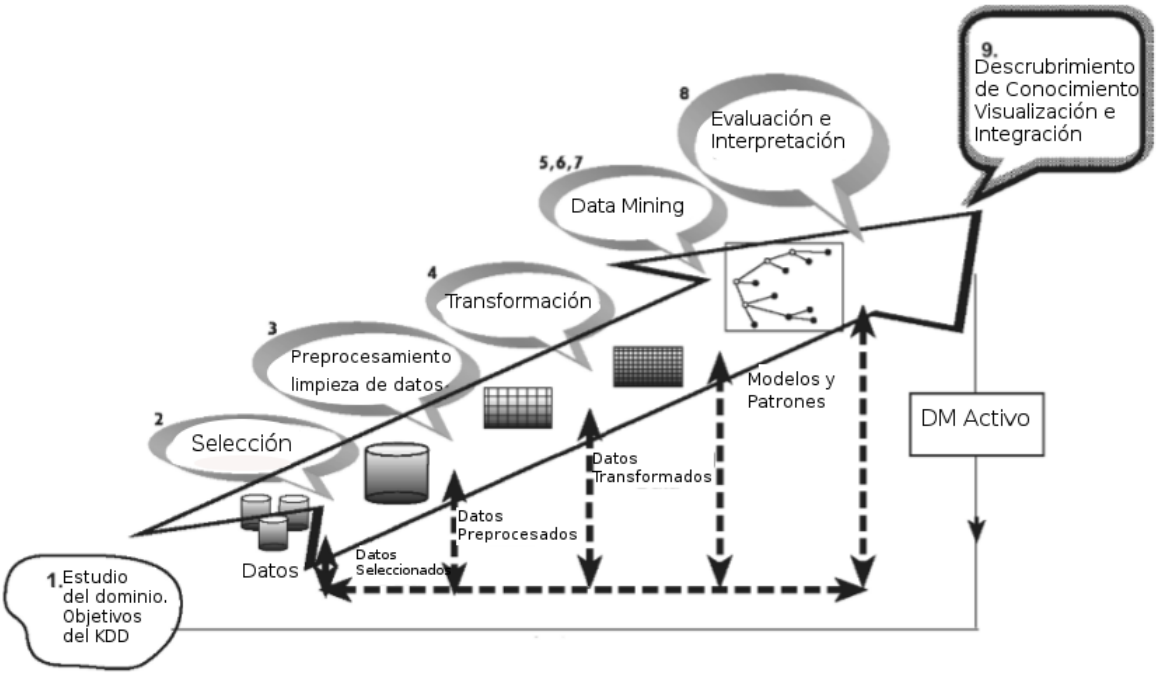

Figura 2.8: Fases y acciones del proceso KDD a partir de *Maimon et Rokach, (2010)*

[1]por sus siglas en inglés Exploratory Data Analysis

Continuando con la descripción del esquema de la figura 2.8, se procede a la aplicación de algoritmos del campo de la minería de datos que responda a los objetivos del KDD, entre los cuales pueden ser algoritmos de clasificación, regresión, agrupación, reducción de dimensionalidad, detección de datos raros y datos faltantes. Este punto involucra un análisis exploratorio con la selección del modelo apropiado y de la hipótesis, es decir, seleccionar el algoritmo más acorde a los datos y posibles patrones. También contempla buscar los parámetros e hiperparámetros de los algoritmos seleccionados. Los mismos dependen del tipo de variable de salida, de modo que si la variable de salida es cuantitativa se torna un problema de regresión y si es categórica un problema de clasificación. Por ejemplo, al aplicar el SVM los hiperparámetros que se deben calibrar no son los mismos si es un problema de regresión o de clasificación (como se describirá en la sección 2.3.4).

A partir de este punto se arriba a la esencia del proceso KDD: la utilización de algoritmos de DM, donde se aboca a la búsqueda de patrones de interés en alguna forma de representación particular o grupo de representaciones como pueden ser reglas de clasificación, árboles o regresión. En un paso ulterior se procede a interpretar los patrones encontrados. El proceso KDD admite, en caso que así lo requiera volver a alguno de los pasos anteriores del que se está analizando y por esta razón se dice que es iterativo e interactivo, permite retornar a alguna de las fases anteriores cuando se necesita buscar alguna explicación al patrón encontrado.

El punto 8 de la figura 2.8 correspondiente a la evaluación e interpretación se pueden utilizar herramientas para representaciones gráficas de los patrones extraídos o hacer visualización de los datos dados los algoritmos utilizados. Estas fases pueden ser sintetizadas como procesamiento o análisis propiamente dicho, lo cual en esta tesis se procederá a simplificar como aplicación de algoritmos.

Finalmente, se supone que ya se ha arribado a los patrones definitivos y, por lo tanto, se procede a documentar y reportar el nuevo conocimiento. También puede ser aplicado a nuevos sistemas para acciones adicionales o para otros análisis más complejos. Una vez obtenidos los patrones, estos deben ser válidos en nuevos conjuntos de datos con cierto grado de certeza, los patrones encontrados deberían ser novedosos, es decir, brinden un nuevo aporte al investigador y potencialmente útiles en el sentido de brindar algún beneficio en la tarea. Dicho esto, se pueden definir medidas que cuantitativamente logren evaluar los patrones extraídos. En general estas medidas se refieren a medidas de certeza como la precisión en la predicción de nuevos conjuntos de datos.

El proceso KDD persigue dos finalidades el descubrimiento de nuevos patrones y la verificación. Con el término verificación se limita meramente a la verificación de una hipótesis planteada por el investigador, en tanto que, descubrimiento es cuando el producto final es la obtención de nuevos patrones. En este contexto el término descubrimiento hace referencia por un lado a que los patrones hallados permiten realizar estimaciones futuras de ciertas entidades; y por otro lado a la descripción donde el nuevo patrón encontrado se utiliza para presentar de una forma entendible y aprovechable como nuevo conocimiento obtenido de las bases de datos(Fayyad et al. (1996)).

Cabe resaltar que al referirse a “Data Mining”, resulta en una fase particular del proceso KDD como se puede observar en el esquema de la figura 2.8, donde se dispone de una batería de algoritmos específicos para la búsqueda de patrones en los datos y provee las herramientas estadístico-computacionales del proceso traducidas en algoritmos. Otros autores (Bramer (2007)) prefieren mencionar directamente al término data mining como el proceso en sí y consideran que data mining es la extracción automática o conveniente de patrones que representan conocimiento almacenado o capturado implícitamente en grandes bases de datos, almacenes de datos, la internet, otros repositorios de información masiva o flujos de datos. Dentro del proceso KDD cobra importancia puesto que se encarga de realizar el análisis de los datos y descubrir algoritmos que bajo supuestos computacionales, genera una enumeración de patrones sobre los datos. En la presente tesis se va a considerar como parte central del proceso KDD donde existen pasos previos que son esenciales para alcanzar el objetivo final que es la obtención del conocimiento a partir de bases de datos.

Como se mencionó párrafos arriba los objetivos más importantes de la aplicación del KDD es la predicción y la descripción. Desde el punto de vista del Data Mining la predicción involucra la utilización de algunas variables o campos en la base de datos para encontrar valores desconocidos o futuros valores e incluso variables de interés. Estadísticamente la aplicación de los modelos permite predecir en tanto que estimar valores del rango de estudio que son desconocidos. Mientras que la estimación de valores fuera del rango de estudio y en predicciones futuras le corresponde el término de pronosticar o pronóstico.

Para lograr los objetivos de predecir y clasificar el DM como disciplina brinda una serie de métodos particulares que es preciso describir con detalle.

2.3.1 Algoritmos de data mining (DM)

Es importante comprender aspectos fundamentales de los algoritmos en el proceso KDD y en particular en el contexto de DM. Todo algoritmo[2] se puede descomponer en tres partes: la representación, el criterio de evaluación del modelo y la búsqueda del algoritmo. La representación del algoritmo hace referencia a como la máquina propuesta describe los datos problema, si la representación del algoritmo es limitada entonces ni aún la disponibilidad de muchos datos ni el tiempo de entrenamiento podrán reproducir un algoritmo preciso para los datos.

En el extremo opuesto, cuando el poder de representación es muy alto entonces se corre el riesgo de sobreajustar un algoritmo en el entrenamiento del mismo y generar predicciones pobres en datos nuevos. El sobreajuste es un problema frecuente al momento de buscar los parámetros de la máquina de aprendizaje y se alcanza no sólo cuando se modela el patrón general de los datos sino al tratar de ajustar el ruido de los mismos por lo que las predicciones en nuevos datos son muy poco precisas. La evaluación del modelo en general es cuantitativa, un algoritmo suele juzgarse por su capacidad predictiva o su precisión en la predicción de datos

[2] en este contexto también máquina de aprendizaje

futuros. Este aspecto se centra en cuán bien un modelo representa a los datos y si el patrón hallado responde a los objetivos del KDD (Fayyad et al. (1996)).

La última componente del modelo se refiere a la búsqueda del método, una vez que la representación del modelo y el criterio de evaluación del modelo se fijan, el proceso de DM se reduce a un problema de optimización, es decir, encontrar los parámetros del algoritmo que optimicen el criterio de evaluación.

Para definir la representación que posee un algoritmo o máquina de aprendizaje resulta interesante ver los conceptos desde el contexto del aprendizaje estadístico. Dado que el proceso KDD tiene una gran afluencia de conceptos estadísticos, es interesante exponer desde dicha concepción las ideas de aprendizaje de algoritmos. Partiendo de una función general, donde se supone que una variable Y está relacionada a una o varias X:

$$\boxed{Y = f(X) + \varepsilon} \tag{2.35}$$

En la ecuación 2.35 Y es una variable de salida o variable de respuesta, en tanto que X corresponde a la o las variables de entrada, variables predictoras o también denominadas variables explicatorias, en algunos contextos como en bioinformática suelen denominarse *features*. Por último, $f()$ es una función desconocida que representa la parte sistemática de la ecuación 2.35 y ε el término de error aleatorio. El objetivo del aprendizaje de los algoritmos es estimar una $\hat{f}()$ a partir de los datos que permita realizar una descripción de los datos y fundamentalmente predicciones valederas, ya que este es además el fundamento de los algoritmos en DM. El aprendizaje consiste justamente en aproximar $\hat{f}()$ a la función "real" que describa el conjunto de datos que relaciona a Y con X (Gareth et al. (2013)). Es decir, encontrar un algoritmo no sólo que ajuste bien a los datos ya conocidos sino también que logre predecir observaciones futuras de manera precisa. Aproximar una función implica aprovechar los datos en un procedimiento de aprendizaje y validación, preprocesando las variables y las variables objetivo, calibrando los parámetros e hiperparámetros y evaluando la performance del algoritmo (Boehmke and Greenwell (2019)).

Para lograr una estimación de $\hat{f}$ a partir de los datos se debe disponer de un conjunto o subconjunto de las observaciones recolectadas para "enseñar" a estimar la función, es decir, aplicar algún método de aprendizaje que permita aproximar la función f. En otras palabras encontrar una función para el conjunto de datos (X, Y), donde :

$$\boxed{Y \approx \hat{f}(X)} \tag{2.36}$$

A dicho subconjunto de observaciones se les llama "datos de entrenamiento". Este subconjunto de datos se utiliza para someter a una función a un entrenamiento necesario no sólo para encontrar los parámetros sino también comparar y encontrar el mejor algoritmo. Una vez que el algoritmo ha sido seleccionado, se utiliza un subconjunto de datos complementario al anterior llamado "datos de testeo" que permite realizar una evaluación insesgada de la

performance de la máquina de aprendizaje. No obstante, dado que el entrenamiento se realiza sobre un mismo subconjunto de datos al igual que el testeo, los errores de predicción suelen ser altos y por otro lado las posibilidades de sobreajustar el algoritmo se incrementan. Una opción es considerar una parte de los datos de entrenamiento para encontrar el mejor algoritmo y sus hiperparámetros, a este último conjunto de datos se le da el nombre de datos de validación (Géron (2019)) .

Para definir el sesgo de un algoritmo implementado, a partir de una $\hat{f}$ determinada y un X fijo se obtiene un conjunto de predicciónes $\hat{Y}$ de manera que la diferencia entre los datos observados y predichos permite descomponer el error como lo muestra la ecuación 2.37. En la ecuación 2.37 $E(Y - \hat{Y})^2$ corresponde al cuadrado de la esperanza entre el dato observado y predicho que puede descomponerse en sesgo y varianza (Hastie et al. (2008)).

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{“sesgo”}} + \underbrace{Var(\varepsilon)}_{\text{varianza}} \end{aligned} \tag{2.37}$$

Mientras el sesgo es un error reducible porque depende de la función $\hat{f}$, es decir, que se puede encontrar una función que reduzca esta diferencia, la varianza es irreducible porque independientemente del $\hat{f}$ que se aplique siembre existe una variabilidad intrínseca de los datos. La varianza indica los cambios producidos en las estimaciones de $f(\hat{x})$ cuando se utilizan diferentes conjuntos de datos. Uno de los desafíos en el contexto del aprendizaje de máquinas es encontrar un método que alcance un sesgo y una varianza mínimo. Aunque los métodos más complejos suelen tener valores altos de varianza y bajos sesgo por el contrario métodos más simples suelen tener una varianza reducida y sesgos altos. Por ello se debe buscar un balance entre la complejidad del método el sesgo y la varianza (Géron (2019)) .

La aplicación de estos conceptos es posible a través del proceso de aprendizaje de los algoritmos, donde es fundamental cuánto se destina de los datos al entrenamiento y cuánto al testeo. En general, algunos autores recomiendan asignar el 60% de los datos a entrenamiento y el restante 40% a testeo (Boehmke and Greenwell (2019)). Cuando el porcentaje de los datos destinados al entrenamiento es mayor al 80% es decir, se desperdicia muchos datos en entrenamiento no se obtiene una buena evaluación de la performance predictiva del algoritmo y se cae en el riesgo de un sobreajuste de los datos. Por el contrario, cuando los datos destinados a testeo son mayores al 40% los parámetros no son estables y no se alcanza una buena performance de los algoritmos.

Existe un gran y variado número de algoritmos disponibles para aplicar en el contexto del DM, no obstante, no hay una única opción que pueda contemplar todas las bases de datos, cada uno ajusta a un conjunto de datos mejor que otros y de ello dependen algunas cuestiones importantes como los tipos de datos, el problema a abordar y el objetivo final del proyecto. También cobra importancia el balance entre sesgo y varianza, buscando algoritmos que preserven el mínimo sesgo sin afectar la varianza.

Dado la gran variedad de métodos existentes y de criterios para categorizarlos, se presenta en la tabla 2.4, en función del tipo de aprendizaje como supervisado o clasificación supervisada y los de tipo no supervisada o clasificación no supervisada.

Tabla 2.4: Clasificación de los algoritmos en supervisados y no supervisados más frecuentemente utilizados en DM (en paréntesis se señalan los autores consultados.)

Entrenamiento Supervisado	**Entrenamiento no Supervisado**
Regresión Lineal (Géron (2019))	Componentes Principales (Kuhn and Johnson (2013))
Regresión logística (Géron (2019))	Detección de anomalías (Kuhn and Johnson (2013))
Regresión logística multinomial (Géron (2019))	Reglas de asociación (Géron (2019))
Regresión Logística $> 2\ clases$ (Gareth et al. (2013))	Análisis de Clúster (Gareth et al. (2013))
Redes Neuronales (Géron (2019))	Análisis de cluster jerárquicos (Géron (2019))
Árboles de Regresión y clasificación (Géron (2019))	k-means (Géron (2019))
Support Vector Machine (Géron (2019))	
K-NN (Géron (2019))	
Clasificación utilizando Bayes (Gareth et al. (2013))	

En la tabla 1.1 de la sección 1.1 se clasificaron de acuerdo al tipo de variable de salida y de entrada.

Los de tipo supervisados son considerados predictivos y se utilizan para obtener un modelo válido que prediga casos futuros a partir del aprendizaje de casos conocidos. Esto implica que, dado un conjunto de casos objetos, descriptos por un vector de características y del cual conocemos la clase a la que pertenece cada objeto, se construye un conjunto de datos de entrenamiento o aprendizaje donde, este tipo de algoritmos hace factible clasificar objetos nuevos de los que no conocemos la clase a la que pertenece Gironés Roig (2013). Una de las ventajas que poseen las técnicas supervisadas es la posibilidad de evaluar la bondad del modelo mediante la tasa de error o por medio de matrices de confusión.

Los algoritmos no supervisados o de clasificación no supervisada son esencialmente descriptivos, también llamados de descubrimientos de patrones y se aplican para obtener modelos válidos que clasifican objetos a partir de la similitud de sus características. Donde, a partir de un conjunto de objetos descriptos por un vector de características, contemplando una métrica que defina el concepto de similitud entre objetos, se construye un modelo o regla general que clasifica todos los objetos. Como ejemplo de esta clasificación se puede considerar al agrupamiento(*clustering*) y segmentación. Algunos autores mencionan además, los de tipo semisupervisados en el caso que se dispone para un conjunto de datos una determinada cantida de observaciones en las cuales existe tanto datos en la variable de salida como en las variables de entrada, pero se presentan datos similares con variables de entrada pero sin variables de salida y en este caso se desea adoptar una técnica que incorpore tanto la información de los datos que disponen de la variable de salida como a aquellos que no (Gareth et al. (2013)) la poseen.

2.3.2 Evaluación de la calidad predictiva de los algoritmos

Para evaluar la performance de los algoritmos es necesario realizar un adecuado entrenamiento de los mismos como ya se vio anteriormente, pero además se deben definir medidas de precisión especialmente al evaluar la predicción que realiza cada uno de los métodos presentados en la tabla 2.4. Las medidas de precisión van a depender de la naturaleza de la variable de respuesta y por lo tanto asociado a un problema de regresión o de clasificación. Cuando la variable de respuesta es continua, el problema presentado es de regresión y en este caso es de interés cuantificar la distancia entre la variable de predicción y el dato observado, una de las medidas más utilizadas es el error cuadrático medio (MSE), que se calcula de la siguiente manera.

$$\boxed{MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2} \tag{2.38}$$

Como se observa en la ecuación 2.38 cuanto menos difiere el valor predicho del valor observado menor será entonces el MSE y ello depende directamente de $\hat{f}$, al referirse al valor que se obtiene del entrenamiento se denomina error cuadrático medio de entrenamiento (Hastie et al. (2008)). Pero lógicamente el MSE de mayor interés es el MSE sobre las nuevas predicciones o sea sobre el MSE obtenido durante el testeo. En caso que se desee expresar el error promedio en las mismas unidades que los datos originales entonces la medida a utilizar es:

$$\boxed{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}} \tag{2.39}$$

Claramente,la ecuación 2.39 es la misma que la anterior pero aplicada la raíz cuadrada. Se pueden mencionar muchas otras medidas de precisión para el caso de regresión en ML como por ejemplo: error absoluto medio, entropía, deviance,etc. No obstante, tanto el MSE como el RMSE son los más frecuentemente aplicados.

En el caso de los problemas de clasificación donde la variable de respuesta es de tipo categórica, la medición de la calidad predictiva es un tanto más compleja. Además, para determinar si una observación corresponde a una u otra clase se utilizan valores umbrales que establecen los límites a partir del cual se considera una u otra categoría. Para evaluar la calidad de las predicciones de los algoritmos se implementa la llamada matriz de confusión. La matriz de confusión es una tabla de contingencia donde se presentan las clases o categorías observadas versus las clases predichas por el método utilizado. Generalmente acompañando la matriz de confusión suelen presentarse distintas medidas estadísticas que evalúan la performance o la predicción del algoritmo (Kuhn (2008)). Para una matriz de confusión de dos clases la constitución de la misma se presenta de la siguiente forma:

Tabla 2.5: Matriz de confusión general. A la izquierda se utiliza la notación alfabética tradicional(A: éxitos correctamente clasificados, B: no éxitos incorrectamente clasificados como éxitos, C: éxito incorrectamente clasificados como no éxitos y D: no éxitos correctamente clasificados) A la derecha la interpretación directa (V: verdaderos, F: falsos, +: positivos y -:negativos)

	Observadas	
Predichos	Éxito	No Exito
Éxito	A	B
No Éxito	C	D

$\Leftrightarrow$

	Observadas	
Predichos	Éxito	No Éxito
Éxito	V+	F+
No Éxito	F-	V-

Podemos interpretar la matriz de confusión mostrada en la tabla 2.5 como un experimento Bernulli definiendo éxito cuando la clase está presente y no éxito cuando la misma no se encuentra. La diagonal en verde representa las predicciones correctas, y la diagonal en rojo las clasificadas incorrectamente. A partir de estas se definen los siguientes estadísticos:

$$Sensibilidad = \frac{A}{A+C} \tag{2.40}$$

La sensibilidad, o más conocido como *recall*, corresponde a la proporción de clases correctamente clasificadas con éxito respecto de las totales observadas como éxitos. Donde A es también denominada verdaderos positivos y C falsos negativos o error de tipo II. En el área de la teoría de señales se suele definir verdaderos o falsos positivos a las predicciones positivas que son correctas e incorrectas respectivamente; por el contrario verdaderos y falsos negativos a las predicciones negativas que son correcta e incorrectamente clasificadas (Powers (2012)). Esta medida permite detectar la habilidad del algoritmo en identificar las clases con éxito, es decir, los verdaderos positivos.

Otra medida de evaluación de las matriz de confusión es la especificidad que se calcula de la siguiente forma:

$$Especificidad = \frac{D}{B+D} \tag{2.41}$$

Representa la proporción de no éxitos correctamente clasificados por el algoritmo,respecto de los totales conjuntamente con las incorrectamente clasificadas como éxito. En este caso D corresponde a los verdaderos negativos y B a los falsos negativos o error de tipo I. Esta medida evalúa la habilidad del algoritmo de “rechazar” las clases que no se clasifican como éxito.

Sin embargo, uno de los más importantes es el estadístico de exactitud o Accuracy[3], que se calcula como:

$$Exactitud = \frac{A+D}{A+B+C+D} \tag{2.42}$$

[3]en esta tesis se va a utilizar el término exactitud puesto que existe una medida denominada *precision* que se refiere a los valores verdaderos predichos

La exactitud se refiere a todos las clases correctamente clasificadas, ya sea verdaderos positivos como verdaderos negativos respecto del total de la población. Es decir, tanto los que se clasifican correctamente cómo éxito como los que se clasifican correctamente como no éxito respecto del total de la tabla. Es una media ponderada del recall y la especificidad.

Algunos softwares, en particular el paquete **caret** de la suite R suelen presentar intervalos de confianza para la exactitud a partir de una distribución binomial (Wing et al. (2019)). Cuando el conjunto de datos se encuentra muy desbalanceado por ejemplo en el caso de que los datos posean 95 no éxito y cinco éxitos clasificando todos los valores como de no éxito tendríamos una exactitud de 95% es por eso que se aplica un estadístico llamado balanceo de exactitud y resulta el promedio ponderado entre la sensibilidad y la especificidad.

La tasa de no información es una medida de interés especialmente cuando las clases se encuentran muy desbalanceadas y se calcula como la proporción de clases de mayor frecuencia respecto de la totalidad. Suelen presentarse pruebas a una cola donde se testea si la proporción de la exactitud es estadísticamente mayor respecto de la proporción de tasa de no información, no rechazar la hipótesis implica que la exactitud evaluada no es representativa (Kuhn (2008)).

Otras medidas que suelen presentarse son los valores positivos predichos y los valores negativos predichos, el primero se obtiene como:

$$\boxed{Val.Pos.Pred = \frac{A}{A+B}} \tag{2.43}$$

La medida expresada en la ecuación 2.43 indica la proporción de clases clasificadas correctamente por el algoritmo respecto de la totalidad de predicciones realizadas en la categoría éxitos. Desde la perspectiva del DM se la conoce como como “Precisión” (Géron (2019)), y es una medida acompañada frecuentemente con el recall. Dependiendo del objetivo del proyecto se puede priorizar la precisión o el recall. Aunque se debe tener en cuenta que el aumento de la precisión repercute en una disminución del recall. El balance de estas dos medidas se encuentra mediado por el valor umbral que se haya establecido en la clasificación. Una estrategia es buscar valores umbrales o también denominados en este contexto “valores de corte” que permitan maximizar tanto el recall como la precisión.

Otra medida importante corresponde a la proporción de valores negativos predichos se representa de la siguiente forma:

$$\boxed{Val.Neg.Pred = \frac{C}{C+D}} \tag{2.44}$$

En este caso, representa la proporción de no éxitos correctamente clasificados por el algoritmo respecto del total de predichos sobre la categoría de no éxitos.

El estadístico Kappa (Powers (2012)), compara la exactitud del sistema o del algoritmo respecto de la exactitud de un sistema aleatorio, es decir, compara la exactitud observada con una esperada bajo aleatoridad. La exactitud aleatoria se calcula suponiendo independencia entre los valores observados y predichos, donde se calcula los valores esperados mediante el

producto de sus correspondientes marginales filas y columnas. Luego, el estadístico Kappa se calcula de la siguiente manera:

$$\boxed{kappa = \frac{(Exactitud_{observada} - Exactitud_{esperada})}{1 - Exactitud_{esperada}}} \tag{2.45}$$

En esencia, el estadístico kappa evalúa qué tan cerca las categorías clasificadas por el algoritmo coinciden con los datos observados. No se establece una interpretación exacta del estadístico Kappa, se considera que $0 - 0,20$ como leve, $0,21 - 0,40$ como regular, $0,41 - 0,60$ como moderado, $0,61 - 0,80$ como sustancial y $0,81 - 1$como casi perfecto. Algunos autores (Fleiss (1981)) consideran $kappa > 0,75$ como excelente, $0,40 - 0,75$ como regular a bueno y $< 0,40$ como pobre.

Otros estadísticos como la denominada prevalencia suele utilizarse como estadístico de clasificación aunque dicho concepto es más epidemiológico y se define como la proporción de éxitos (o casos positivos) en la población (o clases).

Se han definido hasta 17 estadísticos para evaluar la performance de un algoritmo de clasificación (Kuhn and Johnson (2013)), en esta tesis se presentan los más destacados, serán luego mencionados en caso que resulte de interés en alguno de los algoritmos aplicados.

Cuando las matrices de confusión se aplican sobre datos multicategóricos, uno de los criterios que se suelen implementar es el de uno versus el resto(*one-against-all*) donde se reconstruye una tabla de contingencia por cada clase comparándola con el resto de las clases como si fuera una única.

Anteriormente, se había mencionado que el proceso de aprendizaje de los algoritmos se realizaba a partir de la partición de los datos en datos de entrenamiento y datos de testeo. No obstante, uno de los inconvenientes que se suelen suscitar es que al dividir los datos tan sólo una única vez, la evaluación de la performance del algoritmo puede estar sesgado si dicha asignación no se establece con métodos debido a que los datos con alguna característica particular podrían caer en una u otra partición. Es por eso que se suelen utilizar criterios de validación llamados validación cruzada, uno de los más utilizados es el método k-fold cross validation (k-fold-cv).

La validación cruzada por k-fold-cv es un método de remuestreo que consiste en dividir el total de los dato en k grupos iguales, el modelo se ajusta sobre los $k - 1$ grupos y se evalúa en el restante grupo como se indica en el esquema 2.9. Este proceso se repite k veces, de manera que en cada repetición un grupo distinto es utilizado como validación. Esto permite que se tengan k estimaciones del error y luego se obtiene un error final de la validación cruzada promediando los errores de los k validaciones. Se llega así a la última instancia que es elegir el método de aproximación para estimar los parámetros del modelo que se va aplicar.

Figura 2.9: Esquema del proceso de validación cruzada por k-fold cross validation según Boehmke and Greenwell (2019)

En este punto es importante identificar y distinguir entre los parámetros y los hiperparámetros de los modelos o técnicas.

Los parámetros corresponden a las variables de configuración interna de un modelo cuyos valores pueden ser encontrados a partir de los mismos datos. Estas variables internas son requeridas al momento de realizar predicciones y, dependiendo del método, pueden ser interpretados en términos del problema. Un ejemplo estadístico clásico de un parámetro es cuando se analiza una variable aleatoria cualquiera asumiendo que la misma posea una distribución gaussiana, es decir, un modelo de campana de Gauss, en ese caso los parámetros que caracterizan el modelo son μ y σ y son estimados a partir de los datos por medio de los respectivos estimadores $\overline{X}$ y S. Otro ejemplo claro de los parámetros de un modelo se puede proporcionar mediante el modelo de la ecuación 2.7. En este caso $\beta_0, \beta_1, \beta_2$ son parámetros del modelo logístico que se estiman a partir de los datos y donde se pueden interpretar en términos del problema. Desde la perspectiva del DM los parámetros no pueden ser proporcionados por la práctica o el usuario sino que son obtenidos (o calculados) como parte del entrenamiento del algortimo. En este sentido se puede concebir al modelo como la hipótesis a verificar y los parámetros la adaptación de esa hipótesis a un conjunto específico de datos.

En tanto que se refiere a hiperparámetros de un modelo como una configuración que es externa al modelo y cuyos valores no pueden o no deberían ser estimados directamente de los datos. Es decir, variables cuyos valores deben estár fijadas previas al aprendizaje del algoritmo. Son utilizados usualmente en procesos que ayudan a estimar parámetros de los modelos. Son especificados en la práctica o por el usuario y se encuentran de manera heurística, y lo más importante es que son sometidos a un proceso de "calibración" para un problema determinado. Es importante destacar que no se puede conocer de antemano los hiperparámetros para un problema dado, se pueden utilizar valores de otros problemas similares o buscar los mejores valores por ensayo y error. En general, se encuentran de manera heurística utilizando una grilla de valores candidatos para los hiperparámetros que intervienen en el modelo, que se prueban y mediante una función de pérdida se verifica que hiperparámetro o combinación de

hiperparámetros tiene mejor performance. Algunos ejemplos de hiperparámetros son: tasa de aprendizaje(ν) en el entrenamiento de una red neuronal artificial, los parámetros C (costo) y Sigma de un SVM, el parámetro k en una técnica de k-nearest neighbors Kuhn and Johnson (2013).

Dado que el interés en este trabajo es realizar predicciones futuras de frutos de diferentes categorías, el desarrollo posterior se basará en los algoritmos de aprendizaje supervisado más ampliamente utilizados como son redes neuronales, support vector machine y árboles de regresión. Se comenzará con la descripción de las redes neuronales puesto que otorgan el contexto teórico para desarrollar las demás técnicas.

2.3.3 Redes Neuronales

Las redes neuronales artificiales (ANN) son, en el marco del Data Mining (DM) técnicas inspiradas en el sistema de neuronas biológicas del cerebro (Van Gerven and Bohte (2018)). Están basadas en una colección de unidades o nodos llamadas "neuronas artificiales" que se conectan entre sí transmitiendo la señal o sinapsis de una neurona a otra. Cada neurona es una "unidad computacional" que posee entrada(input) y salida(output) de señales de información. La primera generación de redes neuronales, fueron los llamados "perceptrones", como se muestra en el esquema 2.10 y se concebieron entre los años 1950 y 1960 por Frank Rosenblatt a partir del modelo de neurona de McCulloch-Pitt.

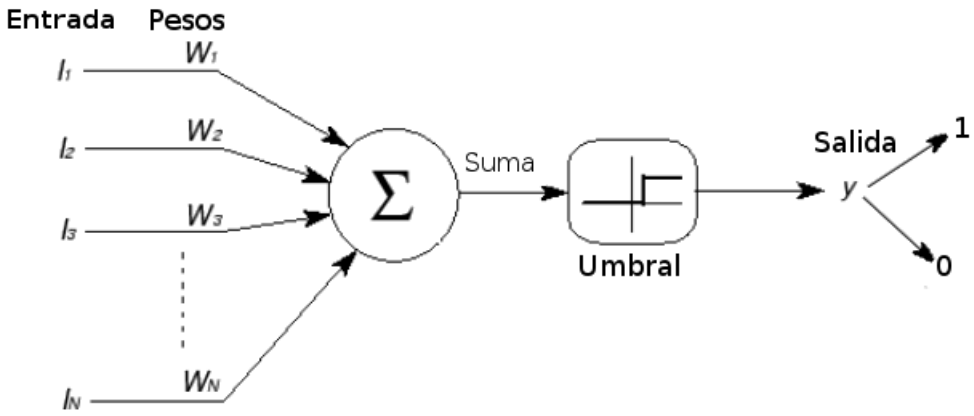

Figura 2.10: Esquema de un perceptron

El concepto de perceptrón, es también, la base de algunos algoritmos de DM como por ejemplo la máquina de soporte vectorial. El perceptrón se basa en varias entradas binarias $x_1, x_2, x_3 \ldots,$ (figura 2.10) a cada una de las entradas se la asigna un peso $w_1, w_2, w_3, \ldots,$, es decir, números reales que representan la importancia que tiene cada una de las entradas respecto a la salida y produce una sóla respuesta (output) binaria $0, 1$ que depende de la suma de las entradas por el peso $\sum_{j=1}^{N} w_j x_j$ comparado a un valor umbral (Nielsen (2015)). Si la sumatoria es menor al valor umbral asignado entonces el valor de salida es 0 y en caso que lo supere el valor es 1.

Las redes neuronales que se utilizan actualmente son a partir de neuronas de la segunda generación, pueden ser aplicadas para aproximar casi cualquier relación funcional compleja entre entradas y salidas. Una neurona típica posee datos de entrada y salida. A cada entrada de la neurona se le asocia un peso o ponderación que es un valor dentro del conjunto de los reales (representado como **W_i** en la figura 2.11).

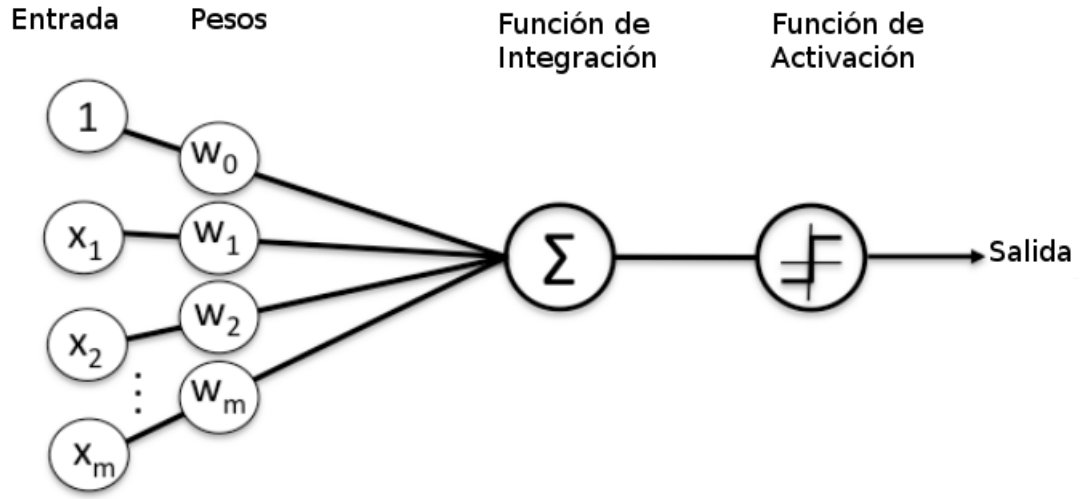

Figura 2.11: Esquema básico de una Red Neuronal Artificial

La neurona multiplica cada una de las entradas por los pesos asociados y suma, mediante una función de integración, todos los resultados que luego son aplicados en una nueva función no lineal llamada "función de activación" (simbolizada como f), obteniendo la salida (output) de la neurona (Nielsen (2015)). Las entradas de cada capa de una red neuronal se pueden expresar como un vector y así simplificar los cálculos matemáticos, y representar los pesos de las conexiones entre las diferentes capas por medio de matrices. Una capa conectada es algebraicamente la multiplicación de vector y matrices donde el vector de entrada X está multiplicado por el peso W, de la siguiente manera:

$$o(x) = f(w_0 + \sum_{i=1}^{n} w_i x_i) = f(w_0 + W^T X) \tag{2.46}$$

En la ecuación 2.46 w_0 se denomina sesgo[4], $W = (w_1, \ldots, w_n)$ corresponde al vector de pesos de las sinapsis y $X = (x_1, \ldots, x_n)$ el vector de variables de entrada. La función es matemáticamente equivalente a un MLG (Modelo Lineal Generalizado) con una función de enlace f^{-1}, donde los pesos son equivalentes a los coeficientes de una regresión. Prácticamente cualquier modelo estadístico para clasificación supervisado o no supervisado puede ser representado como una red neuronal (Ciampi (2007)).

La función de activación es una característica que aparece en la segunda generación de

[4]En machine learning se llama sesgo al intercepto y no debe confundirse con el sesgo de la ecuación 2.37

ANN, algunas funciones de activación son la función sigmoidal, la función tangente, etc.

La neurona artificial, como la presentada esquemáticamente en la figura 2.11, puede estar conectada a otras neuronas conformando una red donde la salida de una puede estar integrada a la entrada de otra u otras neuronas armando una red de neuronas interconectadas. Esta red conectada, puede estar organizadas en capas, donde la salida de las neuronas de una capa provee la entrada a las neuronas de la capa siguiente, las redes neuronales multicapas son también denominadas “perceptrones multicapas (MLP)”. Las típicas redes neuronales multicapas, están compuestas de varias capas de neuronas que definen la transferencia de información entre las capas de entrada y salida.

Un ejemplo típico de las redes neuronales de segunda generación son las “feed forward neural network”(FFNN), que tiene una estructura como la indicada en el figura 2.12.

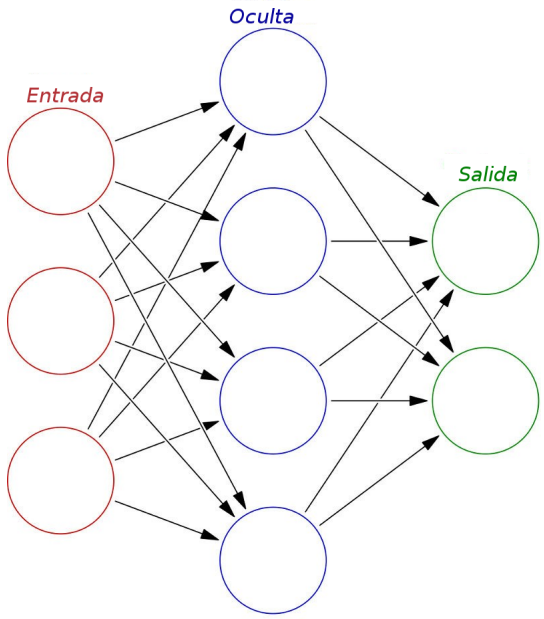

Figura 2.12: Capas de una red neuronal simple

En las “FFNN” la información viaja en un sentido, donde un conjunto de valores de las variables en las capas de entrada, se propaga de una o más capas ocultas a la última capa de variable de respuesta.Las capas ocultas juegan un rol fundamental dentro de las ANN, así como las capas de entrada y de salida están conformadas por variables o nodos, las capas ocultas están compuestas con nodos ponderados que definen la fuerza del flujo de la información (Beck (2018)).

Las redes en las cuales, la salida de cada neurona está conectada con todas las otras neuronas de la siguiente capa son denominadas redes neuronales completamente conectadas. Las redes neuronales típicas están constituídas por una capa de entrada asociada a las variables de entrada, una capa de salida que consiste en las variables de salida (Günther and Fritsch (2010)) y una o más capas intermedias denominadas generalmente como capas ocultas. La inclusión de capas “ocultas” incrementa la flexibilidad de la modelación, en general, se

asume que una capa oculta es suficiente para ajustar cualquier función continua. Las redes neuronales más simples (ver figura 2.12), están conformadas por una capa de neuronas de entrada, una capa oculta y una capa de neuronas de salida, las capas de entrada y las capas intermedias incluyen una neurona constante que no se ve influenciada por las variables de entrada. Una red neuronal multicapa con j neuronas ocultas se rige bajo la siguiente función (Günther and Fritsch (2010)):

$$\boxed{o(x) = f\left(w_0 + \sum_{i=1}^{J} \cdot f\left(w_{0j} + \sum_{i=1}^{n} w_{ij}x_i\right)\right) = f\left(w_0 + \sum_{j=1}^{J} wj \cdot f(w_{0j} + W_j^T X)\right)} \tag{2.47}$$

Donde w_0 denota el sesgo de la neurona de salida y w_{0j} el sesgo de la j neurona oculta, w_j denota el peso de la sinapsis que comienza en la j-ésima neurona oculta, mientras que $W_j = (w_{1j}, \ldots, w_{nj})$ es el vector de todos los pesos de las sinapsis que llevan a la j-ésima neurona oculta y $X = (x_1, \ldots, x_n)$ el vector de las covariables. Tanto las neuronas ocultas y las neuronas de salida obtienen los resultados a partir de:

$$\boxed{f(g(z_0, z_1, \ldots, z_k)) = f(g(Z))} \tag{2.48}$$

La salida o el resultado de todas las neuronas precedentes $z_0, z_1, \ldots, z_k$ es integrada por una función g que es la función de integración y luego da una respuesta a partir de la función f ó función de activación. La función de integración se define como:

$$\boxed{g(Z) = w_0 z_0 + \sum_{i=1}^{k} w_i z_i = w_0 + W^t Z} \tag{2.49}$$

La función de activación, es una función no lineal, derivable que limita la salida de los valores, como puede ser un ejemplo clásico la función logística. La misma acota la salida de las neuronas a los valores $[0, 1]$.

$$\boxed{f(u) = \frac{1}{1 + e^{-u}}} \tag{2.50}$$

La función de activación le otorga a la red neuronal mayor estabilidad en los resultados que el perceptrón. Mientras que en el perceptrón pequeñas variaciones de los pesos producen cambios rotundos de 0 a 1 la función de activación produce cambios graduales que no implican necesariamente un cambio drástico de la respuesta. Esto es debido a que la función de activación, es una función suavizada del perceptrón.De manera que cuando $e^{-z} \approx 0$ entonces $f(u) \approx 1$, es decir, $u = w \cdot x + w_0$ es un número positivo grande, la salida de la neurona es un valor de salida aproximadamente 1.Por otro lado, cuando $u = w \cdot x + w_0$ es un valor muy negativo $e^{-z} \to 0$ y $f(u) \approx 0$, es decir, el valor de salida de la neurona es 0 (Nielsen (2015)).

Desde el paradigma del aprendizaje supervisado, las redes neuronales ajustan a los datos a través de algoritmos en un proceso de entrenamiento por adaptación de sus parámetros,

es decir, los pesos. El aprendizaje consiste en el entrenamiento de las redes neuronales para encontrar los pesos (W_i) que definen las conexiones de las capas del modelo, los pesos óptimos son aquellos que minimizan el error de predicción de un conjunto de datos de prueba. El entrenamiento de las ANN se logra a partir de un algoritmo denominado "gradiente descendiente estocástico", el cual es un algoritmo de optimización que estima el gradiente del error para el estado actual del modelo utilizando los datos y luego actualiza los pesos del modelo a partir del algoritmo "Backpropagation"(Goodfellow et al. (2016)). Más precisamente, los pesos son hallados a partir de los datos maximizando una función de costo $c(Y, out(X|\mathbf{w})$, recurriendo algoritmos como los llamados "Backpropagation" y "Resilient Backpropagation". Estos algoritmos consisten en, calcular los valores de salida y comparar con los valores predichos por el algoritmo, adaptando los parámetros o pesos de acuerdo a la comparación, donde los pesos son inicializados con valores aleatorios usualmente a partir de una distribución normal (Günther and Fritsch (2010)). En forma muy general, el algoritmo entrena durante un proceso iterativo de la siguiente manera:

1. Las redes neuronales calculan una salida $o(x)$ para una entrada x y los pesos actuales (en la primera iteración los pesos se obtienen a partir de valores aleatorios de una distribución normal). Si el proceso no está completo, los valores predichos de o diferirán de los observados y.

2. La diferencia entre los valores predichos y los valores observados se mide con alguna funcion de error E o función de costo, como por ejemplo la suma de cuadrados del error o la entropía.

3. Los pesos son recalculados de acuerdo a alguna medida de aprendizaje como por ejemplo calculando un ajuste del error y la derivada de la función de activación, este es un aspecto muy importante que se rige de acuerdo a la siguiente ecuación:

$$\boxed{w_k^{t+1} = w_k^t - \eta \cdot \frac{\partial C^{(t)}}{\partial w_k^t}} \tag{2.51}$$

   La ecuación 2.51 indica que el k-esimo peso en el tiempo $t+1$ se obtiene a partir del valor de k-esimo peso en el tiempo t menos la tasa de aprendizaje η por la derivada de la función de costo respecto a los k pesos en el tiempo t, donde η es la tasa de aprendizaje. En el caso del algoritmo Resilient Backpropagation η está multiplicado por el signo de la derivada que mejora y acelera la convergencia.

La tasa de aprendizaje η refleja la cantidad o tamaño del paso, que debe actualizar los pesos durante el entrenamiento. Específicamente, es un hiperparámetro configurable utilizado en las redes neuronales que controla la velocidad a la cual las ANN aprenden. Un valor muy grande de tasa de aprendizaje permite a las ANN aprender más rápidamente pero a riesgo de alcanzar un conjunto de pesos subóptimo, en tanto, que un valor muy pequeño permite alcanzar pesos óptimos pero con una mayor demanda de entrenamiento(Goodfellow et al. (2016)).

El aprendizaje finaliza de acuerdo al cumplimiento de algún criterio pre especificado como la cantidad de iteraciones o a valores de tolerancia arbitrarios impuesto sobre las derivadas parciales de la función de costo.

Desde el punto de vista estadístico, la función de costo puede identificarse con la deviance de los modelos generalizados expresado como $-2logL$ (menos dos veces la log-verosimilitud de los datos), es decir, una medida de bondad de ajuste obtenida por máxima verosimilitud, visto de esta forma el entrenamiento de las ANN es una aproximación a la maximización de la función de verosimilitud Ciampi (2007). Es más, la elección de la función de activación es similar a la elección de la función de enlace en los MLG y la elección de la función de densidad se asemeja a la elección en la función de costo. Se puede profundizar aún mas y asociar una red de redes neuronales con un modelo generalizado aditivo.

Desde el punto de vista de los MLG, la utilización de las redes neuronales lleva a un modelo hiperparametrizado, además un número arbitrario de parámetros en las redes neuronales otorga una ventaja predictiva, pero complica la extracción de información del modelo. Asimismo, resulta complejo evaluar la importancia de las variables y la sensibilidad de la técnica, aspectos importantes al momento del análisis de datos Beck (2018). Es por esta razón que (Günther and Fritsch (2010)) introdujeron el concepto de peso generalizado, una medida relativa que permite conocer la contribución de cada covariable en la ANN y permite interpretar a los pesos como los parámetros de un modelo de regresión lineal. Dichos pesos o ponderaciones pueden ser utilizados para obtener diagramas de interpretación de redes neuronales, evaluar la importancia de las variables y la sensibilidad en el modelo (Beck (2018)).

Algunos autores afirman que los pesos de las redes neuronales tienen una distribución normal multivariada si la red es identificable, es decir, no posee neuronas irrelevantes, tanto en las capas de entrada como en las capas ocultas (Ciampi (2007)). Para lograr una red identificable se han propuesto técnicas de poda ó “pruning” de pesos no relevantes en las redes neuronales (Bergmeir and Benítez (2012)). En base a ello, algunos softwares incluyen intervalos de confianza a los pesos de las neuronas Günther and Fritsch (2010).

En la actualidad está muy vigente el uso de técnicas como “deep learning” para el reconocimiento de imágenes. Aquellas redes neuronales con dos o más capas ocultas son llamadas “redes neuronales profundas” o “deep neural network”, donde la profundidad está dada por la cantidad de capas ocultas que se asignen Nielsen (2015). Otros ejemplos incluyen las “redes neuronales recurrentes” y “redes neuronales recursivas”.En el caso de las redes neuronales recurrentes, la salida de una capa de neuronas puede retroalimentar la red siendo la entrada de una capa anterior. Básicamente, implica que una red neuronal puede poseer un bucle o un ciclo en la cual permite que el resultado de una capa i al tiempo t llegue a una capa anterior en el momento $t+1$. Esta condición le provee a las redes neuronales una cierta memoria a corto plazo. Este tipo de redes son utilizadas para clasificación de secuencias y de series temporales.

Existe una tercer generación de redes neuronales que emplean neuronas espinosas o “spiking neuron”. Este tipo de neuronas se acerca aún más al modelo de actividad de neurona biológica Van Gerven and Bohte (2018).

Las ANN se utilizan para un amplia variedad de aplicaciones a lo largo de muchas disciplinas, en el ámbito de salud, financiero, de marketing y minería de datos, negocios, logística e ingeniería. En medicina la comparación de las ANN, los árboles de decisión y regresión respecto de los modelos de regresión clásico, mostraron que las ANN y los árboles de regresión tenían mejor capacidad predictiva que el modelo clásico de regresión múltiple en datos de fumadores Razi and Athappilly (2005). Se ha observado que las ANN son más versátiles y tienen mayor facilidad de configuración que, por ejemplo los modelos no lineales o no lineales mixtos. Además, las ANN no requieren validar supuesto ni distribución alguna, aunque no provee de parámetros interpretables de los datos ajustadosGaleano-Vasco (2013).

En el dominio agronómico, en estudio de medidas longitudinaes, modelando el crecimiento de pollitas Lohman, se destaca un gran desempeño de las ANN, mejorando las predicciones respecto de un modelo no lineal, pero no logrando modelar la variabilidad y las correlaciones entre las observaciones como si pudo observarse en el ajuste de modelos no lineales mixtos Galeano-Vasco (2013). En en este estudio la mejor confección de las redes se logró con una capa oculta de 6 neuronas y una entrada y una salida. También se utilizaron para modelar el crecimiendo de patos, en este ejemplo, se utilizó una FFNN con 3 capas y una capa oculta y 4 neuronas en su capa oculta Kaewtapee et al. (2011). En este último, se utilizó una FFNN con función de activación Gaussiana, que mejoró el entrenamiento y se sostuvo que las ANN son una herramienta apropiada para resolver complejas predicciones de crecimiento, en este caso la función de activación Gaussiana logró predecir el crecimiento con mayor precisión que los modelos no lineales. Las ANN también fueron aplicadas para predecir el crecimiento en peso de la bayas de uva “tempranillo” contemplando las variables climáticas, de temperatura, humedad y viento Fernandez-Martinez et al. (2011). No obstante, no fueron en términos de error cuadrático medio y de tiempo de procesamiento la técnica que mejor ajustó al crecimiento de las bayas sino que la mejor performance se observó en técnicas no paramétricas como las llamadas de “Proceso Gaussiano”. Aunque no se estableció claramente cómo fue confeccionada la estructura de la red neuronal.

Para la implementación de las ANN se dispone en el ambiente R de una serie de librerías frecuentemente empleadas como **nnet**, **neuralnet**, **RSNNS** , otras librerías como **caret** y **NeuralNetTools** suelen utilizarse como herramientas suplementarias, confeccionar las redes y extraer información de la misma. Otros paquetes menos frecuentemente utilizados son **AMORE**, **FCNN4R**, **monmlp** y **qrnn**. Para la implementación específica de “deep learning” existe entre otros **MXNet** y **deepnet**(Rong (2014)). Cabe aclarar que el paquete **FCNN4R** ha sido removido del CRAN y el **MXNet** sólo se encuentra disponible desde repositorios alternativos como github. En el caso específico de la técnica de deep learning es recomendable otro lenguaje como Python que dispone de la librería **scikit learn** que implementa fundamentalmente las redes neuronales multicapa y tienen mucho mayor desarrollo computacional que en R.

Si bien las redes neuronales son excelentes algoritmos y de gran potencia son mucho menos comprensibles e interpretables que los árboles de regresión y mucho más complejas de implementar que las máquinas de soporte de vectores(SVM).

2.3.4 Máquinas de Soporte Vectorial(SVM)

El SVM (máquinas de soporte vectorial) es una de las técnicas más eficientes para clasificación y regresión disponibles actualmente en minería de datos Karatzoglou et al. (2004). Además es un algoritmo de fácil implementación si se compara por ejemplo con redes neuronales. Surge a partir de la teoría de aprendizaje estadístico propuesta por Vladimir Vapnik en el año 1998 (Vapnik (1998)). Es, dentro de los algoritmos de minería de datos, un algoritmo de tipo supervisado. Es decir, requiere de una variable de salida para cotejar el entrenamiento y el aprendizaje del mismo. Cuando no existe una variable de salida definida, no es posible aplicar el método supervisado por lo que se recurre a una aproximación no supervisado que intenta encontrar la agrupación natural de los datos, en este caso se conoce como Support Vector Clustering.

El SVM, inicialmente se concibió como un clasificador binario aplicado a problemas de separación de datos dicotómicos, inspirado en el perceptrón ya descripto en la sección 2.3.3. Posteriormente, se extendieron a problemas de clasificación multiclase, regresión y detección de datos atípicos. El SVM básicamente, aplica un método lineal en un espacio hiperdimensional de características no lineales respecto de la variable de entrada.

La idea principal del SVM consiste en encontrar un hiperplano que logre una separación lineal entre dos clases de datos maximizando la brecha entre los conjuntos de clases a separar. Básicamente, a partir de un conjunto de datos construye un modelo que permite predecir nuevos casos. Es decir, un algoritmo que encuentra patrones y luego los replica frente a nuevos datos.

Para ilustrar lo descripto se puede ver la figura 2.13 corresponde a un esquema donde se observan dos clases de datos representados por las manzanas y las peras.

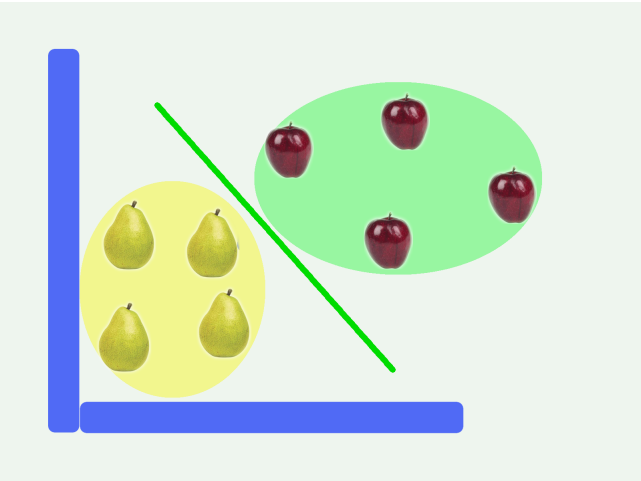

Figura 2.13: Esquema de implementación del SVM en dos clases de datos: peras y manzanas

Si se deseara separar las manzanas de las peras, a partir de dos atributos por ejemplo peso y diámetro, el objetivo sería encontrar una recta que permitiera, clasificar o separar ambos grupos procurando no cometer error, es decir, clasificar una especie por otra. En la figura 2.13,

las peras y las manzanas están esquematizadas en un plano confeccionado por dos variables de interés y son separables a partir de la recta. Si fuera separable pero en n dimensiones se requeriría de un hiperplano para su separación. Un hiperplano, geométricamente es un subespacio de una dimensión menor al del ambiente espacial (un punto en una dimensión, una recta en 2 dimensiones, plano en 3 dimensiones, etc) de dimensión d es un conjunto de datos $x \in \mathbb{R}^d$ que satisface la siguiente ecuación:

$$\boxed{W^T X + b = 0} \tag{2.52}$$

Donde W es un vector (de pesos o coeficientes) y X vector de variables de entrada, entonces $W^T X$ es el producto interno entre ambos y b el sesgo. Como se observa guarda una estrecha relación con la ecuación 2.46, en este caso llamaremos b al sesgo (intercepto con el eje vertical) y no w_0 para separar la notación de los métodos. [5] Esta notación se utiliza puesto que resulta práctico expresar en más de dos dimensiones.

De acuerdo al hiperplano definido podemos asociar cada clase a un valor $+1$ o -1 (manzanas y peras respectivamente en la figura 2.13), y se define una función h:

$$h(x_i) = \begin{cases} +1 \quad \text{si} \quad W \cdot X_i + b \geq 0 \\ \\ -1 \quad \text{si} \quad W \cdot X_i + b < 0 \end{cases}$$

El cual es equivalente a:

$$\boxed{h(X_i) = sgn(W \cdot X_i + b)} \tag{2.53}$$

A cada uno de los datos se le asigna una etiqueta, a los frutos que están por encima del hiperplano se le otorga seguramente un valor $+1$, en tanto que, los frutos por debajo un valor -1. Recordando que la función sgn es una función matemática definida a trozos y devuelve $-1, 0, 1$. Dado que se utiliza la ecuación del hiperplano 2.52 que genera una combinación lineal de los valores, la función h se llama **clasificador lineal**. Donde $h(x)$ es una función hipótesis, es decir, una función que se busca y se prueba hipotéticamente ya que pueden ser infinitas las posibles funciones que puedan utilizarse para clasificar el conjunto de datos Bennett and Campbell (2000).

Se puede reformular la ecuación 2.52 de manera de simplicar aún más, removiendo la constante b, primero agregando la componente $x_0 = 1$ al vector $X_i = (x_1, x_2, \ldots, x_n)$ entonces es $\hat{X}_i = (x_0, x_1, \ldots, x_n)$ y de la misma forma $w_0 = b$ al vector $w = (w_1, w_2, \ldots, w_n)$ y obteniendo

[5] Guarda una estrecha relación con una recta ya que $y = ax + b$ si renombramos $y = x_2$ vemos que $x_2 = ax_1 + b$ entonces podemos despejar $ax_1 - x_2 + b = 0$ definiendo un vector bidimensional $X = (x_1, x_2)$ y $W = (a, -1)$ se obtiene otra forma de expresar una recta donde $W \cdot X$ es el producto escalar de W por X

el vector $\hat{w}_i = (w_0, w_1, \ldots, w_n)$ Luego, la ecuación 2.53 se redefine como:

$$\boxed{h(X_i) = sgn(W \cdot X_i)} \tag{2.54}$$

El único término de la ecuación que influye en la forma del hiperplano es W cambiando el vector podemos hallar infinitos números de hiperplanos que asimismo separen las clases de la imagen 2.13. En este punto, el objetivo es encontrar el hiperplano óptimo para la clasificación de los datos.

El método SVM, para un conjunto de datos linealmente separables busca dos hiperplanos paralelos que separen las dos clases de los datos, donde la distancia entre las líneas es máxima. La región entre estos dos hiperplanos es conocida como margen, de manera que se selecciona el hiperplano que maximiza desde éste hasta el punto de datos más cercano en cada lado. El mismo se denomina “hiperplano de máximo margen” y el clasificador lineal que se define se denomina “clasificador de máximo margen”.

Geométricamente encontrar el máximo margen equivale a calcular la distancia entre el hiperplano y el dato más cercano, al duplicar dicha distancia se define un margen priorizando la correcta clasificación de las clases. Es decir, para clasificar adecuadamente dos grupos de datos el método encuentra dos hiperplanos que separan los datos sin dejar puntos entre estos y maximizando la distancia hasta alcanzar el primer caso. La distancia maximizada es la que corresponde al margen como lo muestra el esquema de la figura 2.14. Para encontrar el máximo margen es preciso encontrar el margen geométrico (ver en Anexo 8.2 el fundamento geométrico y matemático). Encontrar el hiperplano óptimo es encontrar los valores de W y b que hacen máximo el margen geométrico.

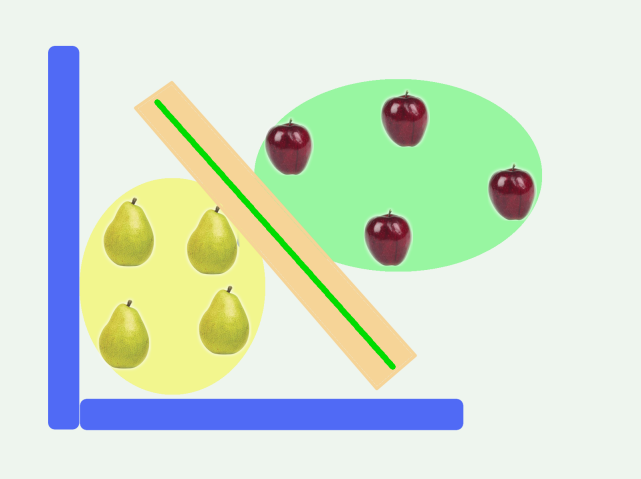

Figura 2.14: Esquema del margen en el SVM para separar las categorías representadas por peras y manzanas

Lo cual resulta en un problema de optimización. Entonces, el “hiperplano de separación óptima” es el hiperplano definido por el vector normal W y el sesgo b para el cual el margen

geométrico M es mayor. Esto es equivalente a minimizar Bennett and Campbell (2000):

$$\begin{aligned} \min_{\mathtt{w,b}} \quad & \frac{1}{2}\|W\|^2 \\ s.a \quad & y_i(W \cdot X_i) + b - 1 \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{2.55}$$

Para encontrar un mínimo o un máximo local de una función sujeto a una restricción(en la ecuación 2.55 $s.a.$) se puede utilizar el método de multiplicadores de Lagrange. La ecuación 2.55 en término de multiplicadores de Lagrange (Ver anexo 8.2.1), queda expresada de la siguiente forma:

$$\boxed{\mathcal{L}(W, b, \alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(W \cdot x_i + b) - 1]} \tag{2.56}$$

Donde $\mathcal{L}$ denota el método de Lagrange para encontrar máximos o mínimos en W, b y α son los multiplicadores de Lagrange. Se aplica una constante α por cada restricción introducida. La solución de la ecuación 2.56 puede ser $\mathcal{L}(W, b, \alpha) = 0$ no obstante, esta solución es meramente analítica y sólo aplicable a un conjunto pequeño de datos. Entonces se debe reescribir el problema en términos del principio dual(Kowalcxyk (2017)). Entonces, la ecuación 2.56 se puede expresar:

$$\boxed{\mathcal{L}(W, b, \alpha) = \frac{1}{2} W \cdot W - \sum_{i=1}^{m} \alpha_i [y_i(W \cdot x_i + b) - 1]} \tag{2.57}$$

A partir de la ecuación 2.57 y el teorema de Slater el problema primordial a resolver es:

$$\begin{aligned} min_{w,b} \quad max_{\alpha} \quad & \mathcal{L}(W, b, \alpha) \\ s.a \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{2.58}$$

Es decir, minimizar la ecuación de Lagrange en función de w y b maximizando α .

Para resolver el problema de minimización hay que obtener las derivadas parciales de $\mathcal{L}$ respecto a w y respecto a b (ver anexo 8.2.1). Sustituyendo, distribuyendo y reagrupando se obtiene:

$$\begin{aligned} max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j \\ s.a. \quad & \alpha_{\geq} 0, \quad \mathtt{para} \quad \mathtt{cualquier} \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{2.59}$$

El problema de optimización es ahora un problema dual y no depende de W ni de b y la función objetivo $W()$ no sólo depende de los multiplicadores de Lagrange. Una vez resuelto el problema dual, se obtiene un vector α que contiene todos los multiplicadores de Lagrange. Entonces, a partir del problema primordial el objetivo es encontrar w y b a partir de los valores de los multiplicadores de Lagrange.

Lo mas destacable, que se desprende es que los vectores soporte son los datos que tienen un valor positivo en los multiplicadores de Lagrange (Kowalcxyk (2017)). Es decir, son aquellos valores para los cuales la restricción $y_i(W \cdot x_i + b) \geq 0$ es activa (Bennett and Campbell (2000)). Recordando que los vectores soportes son aquellos puntos o datos de entrenamiento que son utilizados para definir o maximizar el margen y corresponden a los puntos cercanos a los límites de decisión.

Lo descripto hasta ahora corresponde al cálculo del margen rígido que puede ser aplicable solamente a conjuntos de datos linealmente separables y bien definidos. Existen, al menos, dos casos para los cuales los datos en presencia de valores raros presentan un problema para el margen tal cual está presentado:

- Cuando un dato raro está muy próximo al grupo de clases contrario, de esta forma el margen se reduce al mínimo.
- El dato de una clase se encuentra entre los datos del grupo para lo cual no es linealmente separable y el margen no tiene solución.

Para resolver este problema, en 1995 Vapnik y Cortes introducen una versión del SVM que permite al algoritmo cometer algunos errores en la clasificación de los datos, originando lo que se denomina margen flexible. El objetivo es ahora que el clasificador cometa la menor cantidad posible de errores para poder flexibilizar la clasificación en caso de que existan datos extremos o raros.

Para alcanzar un margen que sea más permisivo en la clasificación se modifica la restricción introduciendo una variable ζ de manera que la restricción en 2.55 es ahora:

$$\boxed{y_i(W \cdot x_i + b) \geq 1 - \zeta_i} \tag{2.60}$$

Se debe tener en cuenta que se podría elegir un valor ζ para cada dato y de esta manera todas las restricciones serían satisfechas. Para evitar esto es que se modifica la función objetivo principal descripta en la ecuación 2.55 que penaliza la elección de ζ:

$$\begin{aligned} \underset{\mathtt{w,b},\zeta}{min} \quad & \frac{1}{2}\|W\|^2 + \sum_{i=1}^{m}\zeta \\ s.a \quad & y_i(W \cdot X_i) + b \geq 1 - \zeta, \quad i = 1, \ldots, m \end{aligned} \tag{2.61}$$

A partir de la ecuación 2.61 el término agregado suma los valores individuales de ζ_i a la función objetivo y se denomina regularización (Karatzoglou et al. (2006)). Esto permite

maximizar el margen permitiendo un pequeño error. A partir de la regularización surgen dos inconvenientes, el primero que el nuevo término admite valores negativos y en segundo lugar se necesitaría tener algún control sobre el término agregado y por lo tanto sobre el "margen flexible" (Chang and Lin (2011)). Para el primer caso se resuelve agregando una nueva restricción donde $\zeta \geq 0$ y para el segundo utilizando el parámetro C de manera que la ecuación 2.61 queda definida:

$$\begin{aligned} \underset{\mathtt{w,b},\zeta}{min} \quad & \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\zeta \\ s.a \quad & y_i(W \cdot X_i) + b \geq 1 - \zeta, \quad i = 1, \ldots, m \\ & \zeta \geq 0 \quad \mathtt{para} \quad \mathtt{todo} \quad i = 1, \ldots, m \end{aligned} \tag{2.62}$$

La restricción de la ecuación 2.62 es denominada también restricción en caja puesto que α se encuentra entre C y 0 .

El margen flexible es una mejora importante respecto del margen rígido, permitiendo clasificar los datos correctamente aún cuando hay datos raros o datos que rompen la linealidad de la separación. No obstante, el costo de incorporar el margen es el hiperparámetro C que se debe encontrar o calibrar por medio de prueba y error. El hiperparámetro C o "costo" controla la penalidad por las clasificaciones erróneas del SVM en datos de entrenamiento, cuanto mayor el valor de C mayor la reducción de los errores de clasificación incrementado la rigidez del algoritmo y aumentando la posibilidad de sobreajustar los datos, es decir, al margen flexible se asemeja a un margen rígido Karatzoglou et al. (2006). Desde otro punto de vista, valores grandes del hiperparámetro C le otorga un costo por error de clasificación muy alto y le brinda menor sesgo a la clasificación pero mayor varianza y fuerza al algoritmo a explicar los datos de entrada muy estrictamente y potencialmente sobreajustar. En tanto que, valores bajos otorga un alto sesgo pero baja su varianza. En el mismo sentido, valore bajos en el hiperparámetro C hace bajo el error de clasificación haciendo más difícil el aprendizaje. El objetivo es calibrar el hiperparámetro de costo por medio de la validación cruzada y remuestreo utilizando una grilla de valores tentativos de búsqueda.

Existen otros tipos de márgenes flexibles, por ejemplo el "Margen flexible 2 normal o regularización L2" en el cual se minimiza:

$$\boxed{\underset{\mathtt{w,b},\zeta}{min} \quad \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\zeta^2} \tag{2.63}$$

La ventaja de este nuevo tipo de margen es que, al plantear el problema dual no se requiere de la restricción de caja mencionada en la ecuación 2.62.

Se ha concebido además otro tipo de margen como es el $\nu-$SVM (Chang and Lin (2011)). Dado que C se ve afectado por el espacio de las variables se propone otro formulación del problema que es el νSVM. El parámetro ν tiene interesantes propiedades ya que acota los

valores de C entre 0 y 1, es decir, establece un límite superior en el error de entrenamiento y un límite inferior en la fracción de vectores soportes encontrados, controlando la complejidad de la función de clasificación (Karatzoglou et al. (2006)).

Entonces el problema dual a resolver ahora es:

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i x_j \\ s.a. \quad & 0 \leq \alpha \leq \frac{1}{m}, \\ & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & \sum_{i=1}^{m} \geq \nu \quad \texttt{para} \quad \texttt{cualquier} \quad i = 1, \ldots, m \end{aligned} \tag{2.64}$$

Muchas veces ocurre que las clases de interés de un conjunto de datos no son linealmente separables como se ilustra en la figura 2.15 se observa que las peras y las manzanas no se pueden separar por medio de una recta, sino que los mismos se disponen de una forma no lineal y teniendo en cuenta que el SVM es un separador lineal no sería factible su separación y clasificación, al menos en $\mathbb{R}^2$.

Figura 2.15: Esquema de dos clases representados por peras y manzanas no linealmente separables

Una opción es transformar cada vector de dos dimensiones (x_1, x_2) en un vector tridimensional para ver si en el nuevo espacio, los datos pueden ser separables linealmente, en este caso por un plano.

Una posible transformación podría ser aplicando una función polinomial donde $\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ definida de la siguiente manera:

$$\boxed{\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)} \tag{2.65}$$

Luego de la transformación, esperaríamos encontrar una situación como la del esquema 2.16. En la misma se observa la proyección realizada por la transformación de los mismos datos pero un sistema tridimensional.

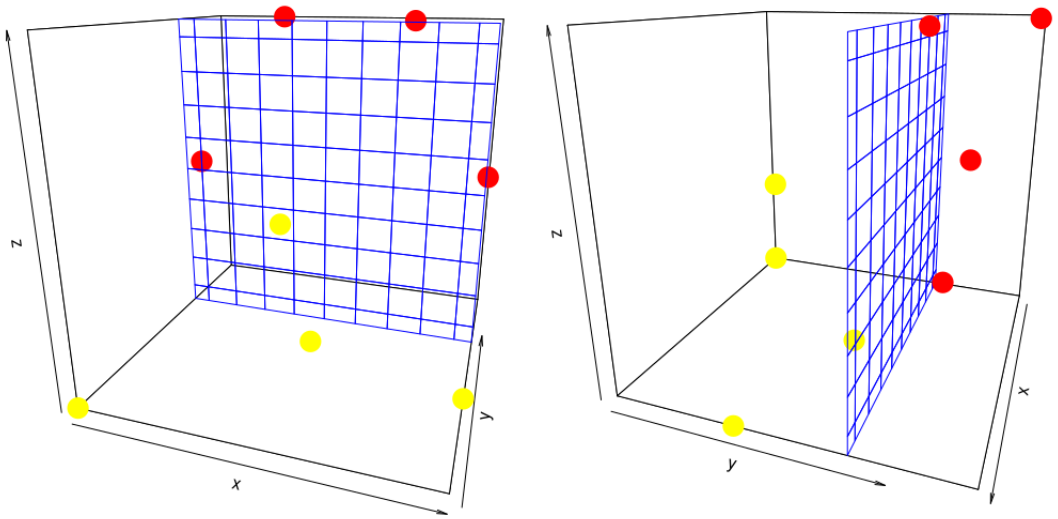

Figura 2.16: Separación, realizando una proyección a una dimensión superior de los datos

En una primera aproximación los datos no pueden ser separados en el plano pero al rotar el sistema claramente los datos rojos y amarillos permiten la separación por un plano. Como se observa en un plano bidimensional los datos nos son separables linealmente, sin embargo, al utilizar una transformación en un espacio de mayores dimensiones, se puede esperar claramente la posibilidad de separarlos linealmente por un plano. Es decir, que se realiza una proyección a una dimensión superior sin agregar variable alguna.

Los datos podrían transformarse a una espacio hiperdimensional de mayores dimensiones hasta lograr una mejor separación de las clases. El problema de transformar cada uno de los datos surge cuando se debe realizar una transformación compleja en un gran conjunto de datos, lo cual demanda una gran cantidad de tiempo de procesamiento. Por dicho motivo se concibió una forma de obtener una proyección en un espacio dimensional mayor sin transformar los datos, utilizando los llamados Kernels Karatzoglou et al. (2004).

El kernel es una función que devuelve el resultado de un producto interno realizado en un espacio distinto (ver definición matemática en Anexo 8.2.2). Si el Kernel lo podemos definir como $K(x_i, x_j) = x_i \cdot x_j$ entonces podemos escribir el margen flexible de la siguiente forma Bennett and Campbell (2000):

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i x_j) \\ s.a. \quad & 0 \leq \alpha_i \leq C, \quad \texttt{para} \quad \texttt{cualquier} \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m}\alpha_i y_i = 0 \end{aligned} \tag{2.66}$$

(2.67)

Existen diferentes kernel o funciones de transformación entre los más importantes se pueden mencionar los siguientes:

- Kernel “lineal”, es otra forma de representar el SVM original permite hacer una separación lineal de los datos.

$$\boxed{k(x, x') = x \cdot x'} \tag{2.68}$$

  Este kernel se ha reportado como un buen clasificador de texto entre otras aplicaciones.

- la forma genérica del kernel “polinomial”

$$\boxed{k(x, x') = (x \cdot x' \ + c)^d} \tag{2.69}$$

  En este caso c es una constante y d corresponde al grado del polinomio. A medida que se incrementa el grado del polinomio la banda de decisión se vuelve más compleja y se ve mas influenciada por datos individuales, altos grados del polinomio le otorgan una mejor performance en el proceso de clasificación es más factible de alcanzar un sobreajuste de los datos.

- el kernel “función de base radial Gaussiano o RBF”

$$\boxed{k(x, x') = exp(-\gamma \|x - x'\|^2)} \tag{2.70}$$

  Es uno de los Kernels más utilizado y que mejor performance provee en clasificación, básicamente es una función cuyos valores dependen de la distancia del origen a los datos. Como se mencionó anteriormente, la función kernel devuelve el resultado del producto escalar realizado en un espacio de dimensiones distinta. En el caso del RBF, devuelve el resultado del producto escalar realizado en $\mathbb{R}^{\infty}$

- el kernel “tangente hiperbólica”

$$\boxed{k(x, x') = tanh(\gamma \cdot \langle x, x' \rangle + coef)} \tag{2.71}$$

Existen muchos otros y variados kernels pero en general los más utilizados y los implementados en distintos softwares son el kernel lineal, kernel RBF y el kernel polinomial. Si bien el kernel más utilizado es e RBF, su aplicación demanda mayor tiempo y esfuerzo de procesamiento que el kernel lineal.

La implementación de uno u otro Kernel requiere la calibración de distintos hiperparámetros. En el caso del Kernel lineal el único hiperparámetro que se debe calibrar es el hiperparámetro C (referido al costo) que ya se describió anteriormente. En el caso del kernel radial, se debe proceder calibrando el hiperparámetro costo y el γ (ver ecuación 2.70, algunos autores afirman que γ representa cuan lejos está de la influencia de un sólo conjunto de datos de entrenamiento, es decir, un parámetro de complejidad (Kowalcxyk (2017)). Cuando el valor de γ es muy pequeño, el modelo está muy restringido y en general no puede capturar la complejidad o la forma de los datos. El kernel polinomial, posee tres parámetros para calibrar: el grado del polinomio, el parámetro escala y el costo.

Se debe destacar que el SVM ha demostrado una gran performance en caso de ser datos separables no linealmente a partir de utilizar un kernel adecuado. Asimismo, realiza un gran trabajo en casos de alta dimensionalidad, es decir, cuando el problema posee un gran número de predictores Bennett and Campbell (2000). Y por otro lado, no tiene problemas de multicolinealidad como pueden experimentar otros algoritmos de aprendizaje.

Si bien el método SVM fue concebido para realizar clasificaciones binarias, también se extiende al caso de datos con mayor número de clases es decir, para la clasificación multiclase, a partir de varias aproximaciones. La primera aproximación es la denominada *"one-against-all"* o *"one-against-rest"* (uno-contra-todos). Esta aproximación consiste en crear, para clasificar K clases, k diferentes clasificadores binarios. Para una determinada clase, se clasifican como positivos los datos de la clase y negativos los que no son correspondientes a la misma. Esta aproximación puede devolver resultados inconsistentes dado que las etiquetas pueden estar asignadas a múltiples clases o a ninguna. Aunque presenta algunas inconsistencias, tiene grandes ventajas en cuanto a que resulta fácil su implementación y asimismo fácil entender su aplicación por lo tanto, suele ser la aproximación más utilizada en los softwares.

Otra aproximación dentro de la clasificación multiclase que compensa esta desventaja, es la aproximación *"One-Against-One"* ("Uno-contra-uno"). A diferencia de la aproximación "uno-contra-todos" se distingue una clase de otra y no una clase de las restantes. Es decir, frente a k niveles, $k > 2$ se entrenan $k(k-1)/2$ clasificadores binarios. Cada clasificador es entrenado en un subconjunto de datos que produce su propia banda de decisión. Las predicciones son realizadas utilizando una estrategia de "voto". Es decir, en cada dato que se va a predecir, cada clasificador predice una clase que es registrada, luego la clase que posee mayor votos asignados al dato es la clase que finalmente se otorga. Esta aproximación presenta aún una dificultad, que surge cuando dos clases reciben la misma cantidad de votos de los clasificadores. Para resolver este problema se proponen algunas estrategias por ejemplo, seleccionar la clase con menor valor de índice (Hsu et al. (2002)).

Otras formas de aplicar la clasificación involucran problemas de resolución de optimización

única en lugar de resolver varios problemas binarios de optimización. Uno de los métodos es el de Vapnik, Weston y Watkins, este método es una generalización del problema de optimización directa (Kowalcxyk (2017)). Consiste, en agregar una restricción por cada una de las clases en el problema de optimización. Como resultado el problema es proporcional al número de clases y puede ralentizar el proceso de entrenamiento si el número de clases es muy grande.

Dentro del marco de la clasificación multiclase por optimización, Crammer y Singer propusieron una alternativa a la de Weston y Singer. De la misma forma resuelve un problema de optimización pero reduciendo las variables poco explicativas. Si bien dicha propuesta mejora considerablemente el tiempo de entrenamiento de los datos, Hsu y Lin (Hsu et al. (2002)) encontraron que este método es especialmente lento cuando se utilizan valores grandes del parámetro de regularizacion C.

De las aproximaciones, para clasificación multiclase someramente descriptas, algunos autores (Hsu et al. (2002)) sostienen que el método "uno-contra-uno" es el método más apropiado por la perfomance, presentando resultados robustos en la predicción y además está presente en la gran mayoría de los paquete de minería de datos.

Se debe tener en cuenta que el SVM realiza la clasificación multiclase considerando que las clases poseen similar frecuencia. En el caso de que las clases no poseen la misma frecuencia, y se encuentren desbalancedas se puede ponderar los vectores soporte por las proporciones de las frecuencias de las categorías de los datos de entrenamiento. Las ponderaciones permiten sesgar el modelo afectando el parámetro C para cada categoría de clasificación y compensar las categorías que están menos representadas en los datos (Chang and Lin (2011)). Estas pueden ser realizadas según distintos criterios: $1/w$, $w - 1/2$,$1500 * w - 1/2$, donde w es la proporción de cada categoría. Es interesante resaltar que las SVM pueden calcular las probabilidades de ocurrencia de cada una de las clases de los datos, a partir de un función de probabilidad a posteriori de Platt (Lin et al. (2007)), que corresponde a una función sigmoidea de la siguiente expresión:

$$\boxed{P(y = 1|f) = \frac{1}{1 + e^{Af+B}}} \tag{2.72}$$

La ecuación 2.72 es ajustada a los valores de decisión f de los clasificadores binarios del SVM, donde A y B son valores estimados por minimizar la función negativa de la log-verosimilitud. Esto es equivalente a ajustar modelos de regresión logística a los valores de decisión estimados. El mismo, se extiende al caso de la clasificación multiclase donde las probabilidades de los clasificadores binarios son combinados.

El SVM no sólo es aplicable a clasificadores binarios o multiclase sino también puede ser aplicado a problemas de regresión. En problemas de regresión el svm también puede ser aplicado con misma base teórica sólo que el método en este caso es el de "regresión-epsilon"(SVM-ϵ), en cuyo kernel debe tenerse en cuenta el parámetro de costo y ϵ. Una de las ventajas de ajustar un SVM-ϵ es que se ven minimizados los valores raros o extremos en la regresión. Por esta razón es que algunos autores le llaman la técnica de "regresión ϵ no

sensible" (Kuhn and Johnson (2013)).

Pero debe considerarse una función de pérdida distinta denominada "$\epsilon-$insensitive" que consiste en resolver $\|y - f(x)\|_\epsilon = max\{0, \|y - f(x)\| - \epsilon\}$, Karatzoglou et al. (2006) donde la variable dependiente o variable de salida es continua. El objetivo es al igual que en los problemas de clasificación maximizar el margen, que en este caso significa minimizar el error. Se establece un margen de tolerancia (ϵ) en aproximación al método SVM. Entonces la formulación general para el caso del SVM aplicado a la regresión es la que se muestra en la ecuación 2.73 (Chang and Lin (2011)). Si la desviación entre el valor actual y el valor predicho es menor a ϵ, entonces la función de regresión no se considera un error, es decir, el hiperparámetro ϵ define el error tolerable. Desearíamos entonces $-\epsilon \leq W\dot{X}_i - b - y_i \leq \epsilon$.

$$\begin{aligned} \underset{\mathtt{w,b,\zeta}}{min} \quad & \frac{1}{2}\|W\|^2 + \frac{C}{m}\sum_{i=1}^{m}(\zeta_i - \zeta_i*) \\ s.a \quad & ((\phi(X_i)\cdot W) + b) - y_i \geq \epsilon - \zeta_i, \quad i = 1, \ldots, m \\ & (y_i - (\phi(X_i)\cdot W) + b) \geq \epsilon - \zeta_i*, \quad i = 1, \ldots, m \\ & \zeta_i* \geq 0 \quad \mathtt{para} \quad \mathtt{todo} \quad i = 1, \ldots, m \end{aligned} \tag{2.73}$$

Geométricamente, se puede visualizar como una banda o un "tubo" alrededor de la hipótesis de la función y cualquier valor fuera del "tubo" puede ser considerado un error de entrenamiento Bennett and Campbell (2000).

El SVM es aplicado en un gran número de campos desde la bioinformática, visión computacional, procesamiento de lenguaje, neuroimágenes, identificación de partículas, identificación de rostros, categorización de texto, etc (Bennett and Campbell (2000)). En el ámbito agronómico se registran ensayos realizados para clasificar sobre imágenes, frutos de tomate sanos de frutos enfermos con distintas afecciones sobre distintos ángulos del fruto obteniendo una precisión del 92%. En este caso, se probaron distintos kernels: lineal, polinomial, y función de base radial, la mejor performance se observó a partir del kernel lineal y el RBF (Semary et al. (2014)). Asimismo, otros autores (Padol and Yadav (2016)), utilizando procesamiento de imágenes aplicaron SVM para detectar automáticamente enfermedades en vid. En este caso se utilizaron imágenes de oídio y mildiu en hojas, donde se entrenó un SVM lineal para clasificar las enfermedades de hoja tanto en estadíos iniciales como en estadíos finales de la enfermedad. Se logró detectar y clasificar con un 89% de precisión las enfermedades combinadas. En manzanas (Suresha et al. (2012)), lograron clasificar en manzanas verdes y rojas utilizando un procesamiento de imágenes con una precisión cercana al 100%. En este caso se utilizaron un centenar de imágenes de baja resolución y se entrenó en función del color utilizando una segmentación previa.

Existe una gran y variada gamma de softwares que implementa SVM, se destacan las librerías **libsvm** y **bsvm** ambas programadas en ***C++***. La primera provee una robusta y rápida implementación del método tanto para clasificación utilizando C-SVM, $\nu-$SVM y $\epsilon-$ SVM en regresión, es la librería más utilizada e implementada (Chang and Lin (2011)). En tanto que la

librería **bsvm**, está programada para la solución de problemas en regresión y clasificación en grandes bases de datos e incluyen entre otros la clasificación multiclase utilizando la formulación de Crammer y Singer Hsu et al. (2002).

Dentro de la suite R se puede mencionar el paquete **e1071** (Meyer et al. (2019a)), fue la primera implementación de la librería **libsvm**, el mismo está diseñado para ser una interface amigable desde R. Implementa funciones no sólo para analizar datos mediante el SVM sino también funciones para la calibración de los hiperparámetros, la graficación de las grillas de validación cruzada y la separación final de los datos. Existen dentro de la suite de R una variada cantidad de librerías que aplican distintas versiones del SVM. Por ejemplo la librería **kernlab** (Karatzoglou et al. (2004)) provee un variado conjunto de métodos basados en distintos kernel implementando las librerías **libsvm** y **bsvm**, extendiendo y flexibilizando las herramientas del SVM. La librería **klaR** (Weihs et al. (2005))incluye una interface a SVMlight que es una popular aplicación de SVM que ofrece adicionalmente herramientas de clasificación y otras como Análisis Discriminante Regularizado.

2.3.5 Árboles de regresión y clasificación

Los árboles de clasificación y regresión son métodos no paramétricos surgidos del campo del DM, que permite la construcción de modelos de predicción a partir de los datos. Es quizás, uno de los métodos más versátiles y al mismo tiempo interpretables, al menos de los algoritmos mencionados en la tabla 2.4. Se basa en la partición recursiva del espacio predictor, y en el ajuste de un modelo simple en cada una de las particiones (Loh (2011)). Una de las grandes ventajas del método abordado es que los resultados se representan gráficamente mediante árboles binarios invertidos que resultan sumamente interpretables. Los primeros algoritmos de los árboles de regresión fueron concebidos hace más de 50 años con la publicación de Morgan y Sonquist en 1963 y se le dio el nombre de algoritmo THAID (Loh (2014)).

En el caso de los árboles de regresión, se denominan de esta forma cuando la variable de respuesta es de tipo continua o en su defecto un valor discreto u ordinal. Los árboles de regresión pueden ajustar prácticamente la gran mayoría de los modelos utilizados en estadística como los modelos lineales, logísticos, no lineales, con distribución Poisson y otros, incluyendo casos con datos longitudinales y de efectos multirrespuesta. En el caso de los árboles de clasificación estos utilizan la clase de mayor representación en lugar del valor de respuesta media y para realizar la predicción en probabilidades, utiliza las proporciones de cada clase de los subgrupos. En una primera instancia, los árboles de decisión (así como también los de regresión), pueden ser descriptos de acuerdo a su topoforma.

En la figura 2.17 se describen las principales partes y topoformas de un árbol teórico. En la parte superior o nodo raíz (o principal) se disponen la totalidad de los datos, los puntos a lo largo del árbol dónde se divide el espacio predictor son denominados nodos internos.

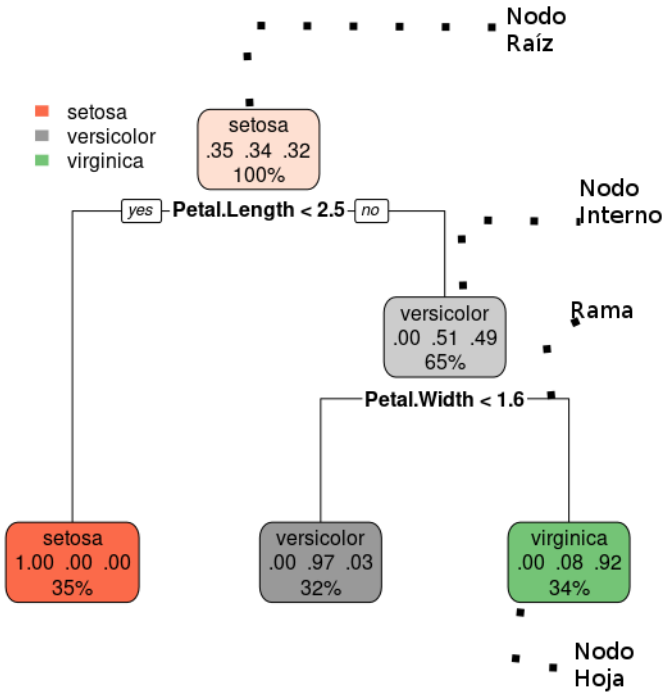

Figura 2.17: Descripción de las partes en un árbol de decisión binario implementado para la clasificación de las distintas especies de Iris

Cada una de las decisiones que se especifican a partir de los nodos son llamadas ramas y los nodos terminales corresponden a las hojas del árbol. Para la construcción de los árboles de regresión y particularmente la obtención de modelos predictivos existen muchos algoritmos y cada uno ha sufrido modificaciones dejando distintas versiones. Uno de los algoritmos más utilizados es algoritmo de clasificación y regresión basado en árboles o más comunmente llamado CART. Básicamente el algoritmo CART procede de manera que los árboles se dividen recursivamente en particiones binarias y conforman en espacio de variables, subconjuntos rectangulares utilizando una variable predictoria a la vez. El término recursivo hace referencia a que la división de un nodo depende del nodo superior y así sucesivamente. La división del espacio de predicción se realiza en rectangulos hiper dimensionales que se les suele dar el nombre de caja. El objetivo es encontrar las cajas que logren minimizar:

$$ESS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_J}) \tag{2.74}$$

Es decir, la mejor disposición de las cajas es aquella que minimice las suma de cuadrados del error de la ecuación 2.74, donde $\hat{y}_{R_J}$ corresponde a la respuesta media para los datos de entrenamiento de la R_j, dado que es inviable considerar todas las particiones posibles, es que se realiza una aproximación descendente que se denomina división recursiva binaria (Gareth et al. (2013)). Suelen referirse como descendentes o de arriba hacia abajo porque comienza en el extremo del árbol donde todas las observaciones corresponden a una única región y luego

se divide sucesivamente el espacio predictivo donde cada división es indicada por dos nuevos brazos del árbol.

El algoritmo CART procede de la siguiente forma, dado un problema de regresión en el cual existe una variable de respuesta continua Y y variables predictoras X_1 y X_2, se generan particiones del espacio de las variables, simplificadas en intervalo unidad y particionadas en líneas rectas como muestra la figura 2.18. Primero se divide el espacio predictivo en dos regiones, y se modela la respuesa Y en cada una de las regiones, luego se selecciona la variable que mayor reducción genere en la suma de cuadrados del error de la ecuación 2.74 y el punto de división que logre el mejor ajuste. Las divisiones de las regiones corresponden a los valores medios de la variable considerada para la región o subregión que se está dividiendo. Luego, una o ambas de estas regiones se dividen en dos regiones más, y este proceso continúa, hasta alcanzar alguna regla de convergencia o regla de detención.

Observando el esquema 2.18 la primera división se realiza en $X_1 = t_1$ luego, en la región $X_1 < t_1$, en el siguiente paso se divide en $x_2 = t_2$ y en la región $X_1 > t_1$ es nuevamente dividido en $x_1 = t_3$. Por último, la región $x_1 > t_3$ es dividido en $x_4 = t_4$ (Hastie et al. (2008). El resultado de este proceso es la partición en 5 regiones. El mismo resultado se presenta en el esquema 2.18 a partir de árboles binarios, donde básicamente los datos completos se sitúan en la parte superior del árbol, las observaciones que satisfacen la condición en cada unión se asignan a la izquierda y las demás a la derecha. Mientras que los nodos terminales corresponden a las regiones $R_1, R_2 \ldots R_5$. Como se mencionó anteriormente la ventaja de los árboles binarios recursivos es su interpretabilidad ya que también el espacio de las variables es plenamente descripto en un árbol.

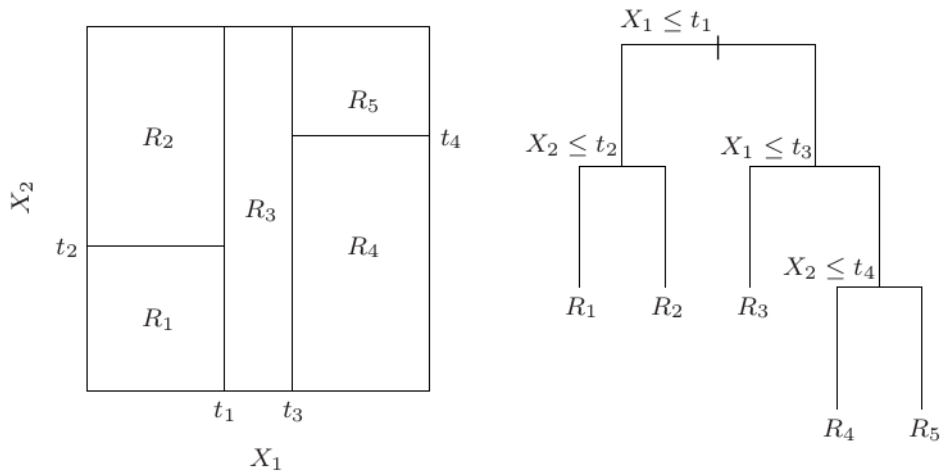

Figura 2.18: División del espacio predictivo (esquema a izquierda) y representación de los árboles de regresión(esquema a derecha) a partir del algoritmo CART

En el esquema 2.18, la predicción de la variable Y para un caso en R_1 está dado por el valor de la respuesta media donde se cumple la condición $X_1 < t_1$ y el valor de la respuesta

media que también cumple la condición $X_2 < t_2$, es decir, donde $R_1 = \{X|X_1 < t_1, X_2 < t_2\}$ y de similar forma para definir las restantes regiones.

El proceso de construcción de los árboles de regresión se basan en la estratificación del espacio de las variables, que presenta en forma resumida dos pasos:

1. dividir el espacio predictivo, es decir, el conjunto de posibles variables $X_1, X_2 \ldots X_p$ en J regiones no superpuestas $R_1, R_2 \ldots R_J$.

2. para cada observación que cae en la misma región R_j se realiza la misma predicción que corresponde al valor medio de los datos de entrenamiento para la respuesta de esa region.

Llegado a este punto y de acuerdo al conjunto de datos a analizar, se debe tener en cuenta cuán complejo o profundo debe ser el árbol a construir. Es decir, cuantas divisiones recursivas debe realizar y cuál es el criterio a utilizar para finalizar las divisiones. Si el desarrollo del árbol es demasiado profundo para el conjunto de datos se corre el riesgo de sobreajustar los datos con el consecuente problema de tener predicciones de datos futuros muy poco precisas. Si la complejidad del árbol es inferior a la requerida para los datos, entonces también se tendrán precisiones poco exactas que podrían ser mejoradas. Por este motivo, se debe encontrar un balance en la complejidad de los árboles, para lo cual es necesario someter el algoritmo a un proceso de calibración donde el hiperparámetro que se debe optimizar es justamente la profundidad del árbol. Este proceso aplicado a los árboles de regresión se denomina *pruning* (Boehmke and Greenwell (2019)). El *pruning* consiste en generar árboles de la mayor complejidad y profundidad posible para luego "podar" hasta encontrar un árbol óptimo. Dicho subárbol óptimo se obtiene aplicando un parámetro de costo o complejidad (ζ) en la función a minimizar (ecuación 2.74) para el número de nodos terminales, de manera que la penalización resulta:

$$\boxed{min\{SSE + \zeta|T|\}} \tag{2.75}$$

Para valores de ζ pequeños mayor será la complejidad del árbol, en tanto, que para valores mayores menor será el desarrollo del árbol. De manera que para un valor determinado de ζ se busca la configuración del árbol expresado por el número de T que minimice el cuadrado del error. El valor de ζ es un hiperparámetro que se debe calibrar y una de las técnicas más utilizadas es la validación cruzada.

Si bien existen un gran número de algoritmos disponibles en árboles de decisión como son los algoritmos THAID, CHAID, C4.5, el más utilizado y que se encuentra muy desarrollado en softwares es el ya descripto CART. En la suite R, se encuentra implementado en al paquete **rpart**, las funciones desarrolladas incluyen la calibración o *pruning* y la validación cruzada para optimizar el número de nodos (Therneau and Atkinson (2019)). También cabe mencionar la

librería **rpart.plot**, una interface con gran número de opciones para graficar árboles de acuerdo al tipo de variable analizada (Milborrow (2019)).

No obstante, los árboles de regresión y clasificación suelen no alcanzar la capacidad predictiva de otros métodos de aprendizaje como las redes neuronales, tampoco la performance y la exactitud del SVM pero brinda una presentación gráfica de los resultados que pocos métodos poseen.

Capítulo 3

Materiales

3.1 Descripción del dominio de los datos

El pronóstico de producción de peras y manzanas fue realizado en las provincias de Río Negro y Neuquén desde el año 1992 hasta el año 2015. Consistió fundamentalmente en la recolección de información sobre los principales cultivares de pepita de la región. El relevamiento de datos abarcó las zonas del Alto Valle, Valle Medio, Río Colorado correspondientes a la provincia de Río Negro y las zonas productoras de Neuquén como se señala en la Figura 3.1.

Figura 3.1: Zonas de relevamiento para pronóstico de producción de frutas (gentileza del Dr. Sergio Bramardi, esquema para el pronóstico de producción)

Los trabajos a campo se iniciaban, todos los años el 30 de noviembre y se extendían hasta el día 14 de diciembre, se basó en el recuento de frutos por planta y en el registro de diámetros por fruto. La metodología implementada se describe en la sección 2.1.3 y los datos registrados utilizados para esta tesis se describen a continuación.

Los registros de información se realizaron sobre los cultivares de manzana *"Red Delicious"* , *"Royal Gala"* (con sus clones mejorados) y *"Granny Smith"* , y las variedades de peras *"William's"* , *"Packham's Triumph"* y *"Beurre D'Anjou"* . Para cada variedad el relevamiento de información se realizó por estratos de edad (10 a 19 años, 20 a 29 años, 30 a 39 años y más de 40 años) y por sistemas de conducción ("espaldera" y "monte libre"). En el caso del cultivar *"Gala"*, se contemplaron los clones más frecuentemente implantados como *"Gala"*, *"Royal Gala"*, *"Mondial Gala"* y *"Galaxy"*.

Además se relevó, el distanciamiento entre plantas y filas, apreciación subjetiva de la carga y el tamaño del fruto, y observaciones sobre características del monte frutal como las prácticas culturales realizadas sobre la parcela, distribución de los frutos, porte del árbol, problemas sanitarios y del suelo, daños por granizo y heladas y estado sanitario en general.

Para el presente trabajo de tesis se logró unificar y sistematizar un total de 174.540 registros, y es el material de partida disponible para su procesamiento.

3.2 Datos para el pronóstico de volúmenes de cosecha

Para aplicar la ecuación 2.1 fue necesario conocer para cada estrato tres fuentes de información:

i) Carga de los árboles al momento del pronóstico, expresadas en número de frutos por planta o por hectárea.

ii) Tamaño esperado a cosecha del fruto medio.

iii) Cantidad de plantas por estrato o alternativamente la superficie efectiva del estrato.

La primera fuente de información, carga de frutos, se realizó mediante el conteo total de dos árboles por unidad muestral que en general se corresponde con un cuadro de la parcela productiva o chacra.

Cabe aclarar que en la región del Alto Valle cada parcela productiva se denominaba "chacra" y poseía en promedio entre 10 y 20 hectáreas generalmente implantado con dos o más variedades de peras y manzanas. Los cuadros eran definidos por el productor generalmente con una superficie de una hectárea donde el marco de plantación, sistema de conducción y fecha de implantación eran comunes.

El número de unidades muestrales dentro de cada estrato se determinó en función de la variabilidad de las parcelas de acuerdo a trabajos anteriores realizados para tal fin (Reeb et al. (2003)). Esto permitió conocer la carga media por árbol, y complementariamente con los datos de distanciamiento entre árboles y entre filas, realizar la expansión a número de frutos por hectárea.

Al mismo tiempo, para obtener los tamaños de frutos medios a cosecha en cada temporada, al momento del relevamiento anual del pronóstico, es decir, unos 40 días previos a la cosecha dependiendo el cultivar (Bramardi et al. (2006)). Se requirió de la medición de cuatro frutos por árbol, en cada uno de las unidades muestrales consideradas en la primer fuente de datos.

Esta tabla conjuntamente con las curvas de crecimiento permiten obtener el tamaño en milímetros de los frutos que se espera al momento de cosecha.

Las fuentes de datos se consideran tablas de dimensión $F \times C$ donde, para este caso las filas F corresponden al registro de cada árbol y C las columnas a las variables de ese árbol. Las C columnas están relacionadas a la fecha de relevamiento, localidad, número de cuadro, cultivar, sistema de conducción, distancia entre filas y plantas y cantidad de frutos por planta.

Esta fuente de información se reunió en una única tabla que se le dio el nombre de "Carga" pudiéndose albergar unos 10.485 registros. Dichos registros son los correspondientes a los pronósticos de producción de los años 1993, 1994, 1995, 1996, 1998, 1999, 2000 y 2009, que alcanzaron las distintas localidades de las zonas productoras de Alto Valle, Valle Medio, Río Colorado y Neuquén.

La segunda fuente de información es referida al tamaño del fruto a la fecha de cosecha y se requiere de 3 tablas de datos.

3.2.1 Diámetros correspondientes a curvas de crecimiento

Los datos referidos a las curvas de crecimiento, fueron una de las tablas más importantes del pronóstico de producción puesto que permitió conocer el patrón de crecimiento de los frutos y la predicción a cosecha del tamaño comercial. Las curvas de crecimiento se han construido y desarrollado para las condiciones climáticas y de producción de la región desde el año 2005 en manzanas, y desde el año 1975 en peras. Es la base de la metodología de las mediciones sucesivas (que se describe en la sección 4.3 de métodos) y por lo tanto, el fundamento del pronóstico de producción desde que se ha implementado dicha metodología que permitió además realizar predicciones con distribución de tamaños comerciales.

Para esta tabla de datos, se considera que por cultivar se obtiene el diámetro medio(en milímetros) de 75 frutos durante aproximadamente 12 semanas de seguimiento de acuerdo a la metodología ya implementada en el pronóstico y ampliamente estudiada (Bramardi (1989), Bramardi et al. (1998)). Se dispuso de un conjunto de curvas de crecimiento importantes que en total alcanzó los 117.902 registros. En los cuales se tuvo en cuenta el tipo de plantación, es decir, la variedad, el portainjerto y el sistema de conducción, la fecha en la cual se llevó a cabo la medición, las personas que realizaron la tarea, el número de la planta, el tamaño del fruto, el número de fruto, los diámetros menor, mayor y el diámetro medio resultante, el nombre de la chacra, su ubicación y la localidad, también se calculó y se registró los días después de plena floración correspondientes para el momento de realizada la medición. La cantidad de frutos seguidos y registrados dependen del cultivar y es por ello que es preciso una descripción más pormenorizada en función de los cultivares presentada en la sección de resultados.

Comenzando por el cultivar de peras *"Williams"* el total de registros disponibles fue de 15.748 con las condiciones ya mencionadas anteriormente. Se destaca asimismo que se han evaluado frutos en la temporada 1975, 1976, 1977, 1986, 1991, 1992, 1993, 1996, 1998, 2009, 2010, 2011, 2012, 2013, 2014 y 2015 realizadas en distintas parcelas productivas de las localidades de Allen, Cipolletti, Fernandez Oro, Cinco Saltos, Neuquén, SP. Del Chañar, Contralmirante MM Guerrico y JJ.Gomez.

Respecto del cultivar de peras *"Packham's Triumph"* se tuvieron 22.293 registros, que corresponde al seguimiento de estos frutos durante las temporadas 1975, 1976, 1977, 1978, 1986, 1991, 1992, 1996, 1997, 1998, 2009, 2010, 2011, 2012, 2013, 2014 y 2015, en las localidades de Cipolletti, Centenario, Contralmirante M. Guerrico, Cinco Saltos, Neuquén, San Patricio del Chañar, Fernandez Oro y J.J.Gomez.

El cultivar de peras *"Beurre D'Anjou"* el total de registros sistematizados fue de 13.084 datos. Las curvas fueron relevadas en los años 2.005 y 2.006, 2.012,2.013, 2.014 y 2.015, fue una de las últimas variedades incorporadas al estudio de las curvas de crecimiento.

En cuanto a las manzanas, uno de los cultivares más sobresaliente sin lugar a dudas fue y continúa siendo el cultivar *"Red Delicious"* es por eso que el total de registros de las curvas fue de 20.649. Los registros incluyen tanto el cultivar tradicional de *"Red Delicious"* como los mejorados, es decir, los denominados regionalmente "Chañar 28" y "Chañar 34". Los datos se han registrado durante las temporadas 1987, 1996, 1997, 1998, 1999, 2000, 2009, 2010, 2011, 2012 y 2013 en las localidades de Cipolletti, Cinco Saltos, San Patricio del Chañar, J.J. Gomez y Neuquén.

Otro cultivar históricamente importante es el cultivar *"Granny Smith"* , cultivar de ciclo largo que asimismo fue ampliamente estudiado como lo demuestra los 18.255 registros cargados. Las parcelas establecidas para su estudio se encontraron en las localidades fundamentalmente de Cipolletti, Cinco Saltos, San Patricio del Chañar y Neuquén. Las fechas de medición registradas son 2009, 2010, 2011, 2012, 2013 y 2014, si bien para este cultivar existieron mediciones anteriores al pronóstico de producción, al provenir de bases de datos distintas las mismas se perdieron y sólo se recuperaron las mediciones y los días a partir de plena flor.

También los cultivares de manzana *"Gala"* fueron considerados en los últimos programas de pronóstico de producción, como se había mencionado anteriormente, su implantación fue en aumento paulatino hasta cobrar importancia en la producción regional. En el caso particular de este cultivar existe un gran número de clones y de combinaciones clon portainjerto que afectan el crecimiento del fruto. Es por eso que es muy importante considerar dicha combinación en las curvas de crecimiento de los frutos. El total de registros sistematizados para las diversas combinaciones de portainjerto y clon asciende a 20.817 registros. No obstante, la combinación más representativa dentro del cultivar *"Gala"* , es *"Royal Gala"* en portainjerto "M4". Las localidades en las cuales se realizaron las mediciones corresponden a Cipolletti y Fernandez Oro. Las mediciones fueron llevadas a cabo los años 2005, 2011, 2012, 2013, 2014 y 2015.

Otro de los cultivares que fue incorporado recientemente a la base de datos del pronóstico es el cultivar *"Cripps Pink"* , el cual se caracteriza por tener un ciclo muy largo, aún más largo

que el cultivar *"Granny Smith"* y por poseer un fruto marcadamente bicoloreado. Dada la importancia creciente en su implantación es que se lo ha incorporado a la base de datos y futuros pronósticos de producción. Hasta el momento se dispone de 7.056 registros que incluyen distinta combinación de portainjertos, los cuales fueron llevados a cabo durante los años 2017 y 2018. El total de los registros se ha evaluado en la localidad de Contralmirante M. Guerrico en una parcela comercial.

3.2.2 Registro de los pesos y diámetros para hallar la relación peso y diámetro de los frutos

Las estimaciones del pronóstico son basadas en el peso total en kilogramos por cultivar y en los tamaños según envases comerciales, mientras que las tablas de datos descriptas anteriormente en la sección 3.2.1, sólo proveen el tamaño del fruto a cosecha en milímetros. Para realizar la transformación de milímetros a pesos en gramos, se estudió la relación peso y diámetro a partir de frutos que son recolectados de dos plantas a los cuales se les midió el diámetro ecuatorial y el peso en gramos. Los registros de peso y diámetro de los frutos se realizaron para todas las variedades de peras y manzanas en el pronóstico al momento de cosecha comercial. Para obtener estos datos, se registró para cada fruto el peso mediante una balanza de precisión y se midió el calibre a la altura ecuatorial en dos mediciones ortogonales, dado que el fruto no es una esfera perfecta y en general presenta irregularidades. Las dos mediciones a la altura ecuatorial del fruto pretenden captar la irregularidad de la superficie del fruto. De dichas mediciones se dispone de 5.488 registros de los principales cultivares de peras y manzanas que son respectivamente *"Williams"* , *"Beurre D'Anjou"* , *"Packhams Triumph"* y *"Red Delicious"* , *"Galas"* en general y *"Granny Smith"* . En esta tabla se ha registrado el número de fruto, fecha de evaluación, promedio de los diámetros ecuatoriales, peso y la parcela de la cual fue relevado.

Además, dada la importancia que poseen en el ciclo productivo se han registrado los eventos fenológicos para cada año y temporada de crecimiento de los frutos, las fechas de plena floración (ddplf) y fechas de cosecha comercial, marcando la duración del ciclo de cada cultivar. Estos registros se incluyen para las distintas regiones productivas: Alto Valle, Valle Medio y Valle Inferior, el total de registros asciende a 434 datos.

La tercera fuente de información corresponde al número de plantas reales de cada estrato de edad o superficie efectiva ocupada por el estrato, es suministrada por organismos fiscalizadores como el SENASA y la FUNBAPA, por tal motivo, no es una fuente construida por el pronóstico de manera directa. No obstante, resultan de suma importancia para conocer la cantidad de plantas y la superficie implantada por variedad, estrato de edad y sistema de conducción, que posteriormente permite realizar las estimaciones del volumen total. Esta información se reunió en una tabla denominada "Superficie_parcelas" y se obtuvo un total de 15.362 datos, en este caso los datos reunidos pertenecían a todas las plantaciones declaradas por el productor del año 2012 brindadas por el organismo de la FUNBAPA.

Con la información precedente la estimación del volumen de producción para cultivar “A” contemplando cada estrato de edad, región, y sistema de conducción estará dada por la ecuación 2.1. La ecuación 2.1 permite calcular para la variedad A la producción total en peso considerando el error por el efecto de “fruto oculto”, que se genera a partir del conteo por defecto dada la dificultad de contabilizar frutos no expuestos por el follaje y sobretodo en los sectores altos de los árboles.

3.3 Datos climáticos

Un objetivo específico de esta tesis es **utilizar la información generada durante estos años para predicciones futuras** como se mencionó en la sección 1.5.2 y por otro lado, contemplar en dichas predicciones los efectos de las variables climáticas que pueden afectar el crecimiento de los frutos (Warrington et al. (1999)). Se recolectaron los datos de las distintas estaciones meteorológicas dispuestas por distintas entidades públicas para la descripción de las condiciones climáticas en las temporadas de crecimientos estudiadas. En este contexto, si bien la región dispone de estaciones meteorológicas ubicadas en distintas localidades, se dispone de los datos de tres de ellas: la estación meteorológica Neuquén Aero, dependiente del Servicio Meteorológico Nacional, ubicada en el aeropuerto de la ciudad de Neuquén, la estación meteorológica automática de INTA Alto Valle centrada en Contralmirante Guerrico y la estación meteorológica de la Autoridad Interjurisdiccional de Cuencas (AIC) en Cipolletti. Cabe destacar que no todas las estaciones disponían de todas las variables mencionadas anteriormente, y solamente la estación meteorológica Neuquén Aero proporcionaba datos a partir del año 1975. Otras estaciones como la AIC realiza tareas de mantenimiento durante el mes de febrero, por lo que en general no se disponen de registros climáticos completos para ese momento crucial del desarrollo de los frutos.

Esta información se estructura en una tabla de datos de dimensiones definidas por los días de registros de temperaturas y las variables climáticas de interés y asimismo otras variables que podrían cobrar importancia al momento del pronóstico como son: temperatura máxima, temperatura media, temperatura mínima, temperatura máxima sin abrigo, temperatura mínima sin abrigo, precipitación, heliofanía efectiva, humedad relativa máxima, humedad relativa media, humedad relativa mínima, viento medio, velocidad máxima del viento, presión atmosférica, temperatura media sin abrigo, viento medio a 10 metros de altura, viento máximo a 10 metros de altura, dirección predominante del viento a 10 metros de altura, viento medio a 2 metros de altura, viento máximo a 2 metros de altura, dirección predominante del viento a 2 metros de altura, radiación global, radiación global máxima, índice ultravioleta, dosis de radiación ultra violeta, índice ultravioleta máximo y evapotranspiración.

El total de registros de información de los datos climáticos de las tres estaciones meteorológicas alcanza los 23.840 registros. Disponer de los datos de las distintas estaciones meteorológicas permite poseer datos de al menos tres puntos claves de la región en estudio, en especial

para estudiar curvas de crecimiento, dada la gran importancia que la temperatura posee en el desarrollo del fruto en particular a comienzos del ciclo. El mayor registro fue brindado por la estación meteorológica del Servicio Meteorológico Nacional en el aeropuerto de la ciudad de Neuquén, totalizando unos 16.801 registros, mientras que la estación meteorológica de la AIC registró 2.715 y por último en INTA Alto Valle Guerrico 4.324. Cabe destacar que los datos del SMN fueron los más completos y que abarcó momentos desde el 1 de enero del año 1970 hasta la actualidad cubriendo las primeras curvas de crecimiento.

Capítulo 4

Métodos

En primer lugar, es importante mencionar que todo el software aplicado en esta tesis tanto para la escritura, análisis, procesamiento y presentación de los resultados, como así también su comunicación, es software libre sin utilizar software propietario o privativo de ningún tipo. Particularmente, el procesamiento y análisis de datos se realizó mayoritariamente a partir de la suite informática R (R Core Team (2019)) versión 3.6.2 en una plataforma bajo el sistema operativo GNU/LINUX distribución "Debian Stretch 9".

4.1 Creación y Gestión de bases de datos

Una de las primeras actividades fue la creación de la base de datos que físicamente se situó en un servidor remoto bajo la siguiente dirección web `https://miredlibre.ddns.net/`. Para acceder se debe identificar y solicitar un usuario y contraseña. El objetivo fue que los datos se encuentren en un sitio virtual de mayor acceso, seguro y a futuro ser consultados desde otras computadoras por distintos usuarios interesados. El servidor constaba de una computadora, conectada en forma permanente a internet que oficia de servidor bajo GNU/Linux distribución RedHat-Fedora 29 (edición para servidores). Esto permitió que todo usuario del grupo de investigación pudiese acceder a los datos para realizar consultas u obtener datos para su análisis.

Para la creación de la base de datos se utilizó un sistema de bases de datos de tipo relacional denominado "MariaDB", la cual es un desarrollo en paralelo o *folks* del sistema "MySQL", que es sin lugar a dudas el más utilizado para la creación y administración de bases de datos cuya versión del servidor fue la 10.3.18-MariaDB - MariaDB Server.

No obstante, para realizar el diseño de la base de datos, la creación de las tablas y la generación de las claves primarias y foráneas se utilizó una interface denominada "PHPmyAdmin", cuya versión fue la 4.9.1. Este software permitió interactuar con las tablas y el diseño en sí de la base de una forma más amigable, rápida y brindaba otras herramientas gráficas de visualización de la estructura de la base de datos. Una vez creadas las tablas y definidas los tipos de datos para cada campo, la importación de los datos y las tablas de datos se realizó directamente

desde el software R. La importación de datos directamente desde “PHPmyAdmin” no resultaba práctico especialmente cuando existían gran número de registros en el archivo.

Esto se logró utilizando la librería **RMariaDB** (Müller et al. (2019)), la cual provee los controladores(drivers) y la interface entre el sistema de base de datos y el software R. Para llevar a cabo las operaciones de bases de datos desde R como exportación e importación de datos, filtrado, actualización y consulta, también se requirió de la librería **DBI** ((R-SIG-DB)), que asimismo es una interfaz que complementa la librería anterior y permitió la comunicación entre R y los sistemas de gestión de bases de datos. Por una cuestión de seguridad informática, el llenado o poblado de las tablas se realizó accediendo remotamente al servidor remoto haciendo uso de una máquina virtual, en el mismo servidor, tanto con R como con las librerías mencionadas. El acceso al servidor y a la máquina virtual se hizo mediante el protocolo SSH ó *secure shell*, que permitió accesos a servidores remotos de forma segura, ya que la información se transfiere de manera encriptada. Para evitar la interrupción de los procesos iniciados en la máquina remota se ejecutó conjuntamente con el protocolo SSH una aplicación SCREEN, es decir, un programa informático que puede usarse para multiplexación de terminales y que entre otras cosas permitío mantener activa la sesión y los procesos iniciados en la máquina virtual aún habiéndose desconectado de ella.

Desde la máquina virtual se realizó la conexión entre R y MariaDB utilizando funciones de la librería **DBI** ((R-SIG-DB)), de manera que la conexión quedó establecida de la siguiente manera.

```
conex <-  dbConnect(RMariaDB::MariaDB(),
  host = "localhost",
  user = "gustavo",
  password = "*****",
  dbname= "gustavo")
```

En la primera línea se definió el driver a utilizar, en tanto que, al utilizar la máquina virtual desde el mismo servidor el host correspondió al mismo por lo que se especificaba como “localhost”, luego el usuario y la contraseña y por último el nombre de la base de datos. En este caso el nombre de la base de datos fue el mismo que el del usuario y para resguardo de los datos se necesitó de una contraseña de acceso. La conexión quedó establecida y guardada en un objeto de R que en este caso se le indicó como `conex`.

Una vez que las tablas fueron creadas y las relaciones entre las mismas establecidas, se procedió al llenado de las mismas exportando los datos desde R hasta la base de datos, utilizando la función `dbWriteTable()` de la librería **DBI**. La función requiere de la conexión que se lo provee mediante el objeto `conex`, un marco de datos de R con los datos a exportar y el nombre de la tabla de la base de datos. Un ejemplo de exportación de datos es el que se presenta a continuación

```
dbWriteTable(conex, "Floracion", datosflor, append = T)
```

En este caso la conexión se otorgó mediante el objeto `conex`, el nombre de la tabla de la base de datos fue “Floracion” y el marco de datos con la totalidad de registros de fenología de los cultivares en el objeto `datosflor`, el argumento `append=T` permitió agregar filas a la tabla ya creada, caso contrario la función crearía una nueva tabla.

Para que la exportación de los datos desde el software de procesamiento R a la base de datos fuese exitosa, las variables del marco de datos debieron tener exactamente el mismo nombre que aquellas definidas en la base de datos. De la misma manera debía existir una correspondencia entre el tipo de datos de la base y del marco de datos del software R, de manera que las columnas definidas como números enteros en la base de datos debían especificarse como enteros en el marco de datos y lo mismo con los datos tipo caracter o de punto flotante. Asimismo, las columnas con fechas que fueron determinadas en la base de datos como tipo “date” debieron poseer el mismo tipo de datos en el marco de datos de R. En el caso de los campos autoincrementales de las bases de datos, en R se especifica como NA o NULL para que fuese el sistema gestor el encargado de otorgarle el número correspondiente a la fila agregada de acuerdo a la numeración de la tabla.

Es importante destacar que la preparación de los datos para su exportación se realizó en R utilizando el ecosistema **tidyverse**(`https://www.tidyverse.org/`), en especial las librerías **dplyr**(Wickham et al. (2019)), **purrr**(Henry and Wickham (2018)) y **tibble** (Müller and Wickham (2019)). Para la preparación de los datos tipo fechas o “dates” se utilizó la librería **lubridate**(Grolemund and Wickham (2011)) conjuntamente con el ecosistema **tidyverse**. Este conjunto de paquetes mencionados, además de simplificar la tarea previa a la exportación a bases de datos hizo mucho más compatible su interacción con la misma, puesto que la semántica de las funciones es semejante a la utilizada en el lenguaje SQL. A manera de ejemplo:

```
grannydb <- grannys %>% select(registro, id_pareja, id_plantas, date,
                        treen, fruitn,size, diametro_mayor, diametro_menor,
                        diameter, id_chacra, dafb)  %>%
                        rename(pareja_de_relevadores=id_pareja,
                        plantacion = id_plantas, fecha_registro = date,
                        planta_n = treen, fruto_n = fruitn,
                        tamanio = size, diametro_medio = diameter,
                          chacra = id_chacra, ddplf = dafb)
```

En este caso al marco de datos `grannys` que contenía los registros de la variedad de manzanas *“Granny Smith”* se le aplicó la función de selección no sólo para filtrar los datos sino también para ordenar las columnas y posteriormente con la función `rename` se cambiaron los nombres de las variables o columnas para que las mismas sean compatibles con los nombres de la base de datos. De manera similar se trabajó sobre el conjunto de datos para los distintos cultivares y las restantes tablas de esta tesis.

Luego de que los datos fueron exportados a las tablas, una operación frecuentemente realizada fue la consulta de los datos, requeridos para el análisis. Dado que la base de datos reunía la totalidad de los datos en todas sus tablas, las consultas sobre la base de datos se

acompaña con el filtrado de datos. Esta acción se realizó mediante la función `dbGetQuery()`, la cual necesitó de los datos de la conexión alojados en el objeto ya descripto y luego una sentencia en lenguaje SQL encomillada. Por ejemplo para obtener todos los datos del cultivar de peras *"Packhams Triumph"* se realizó la siguiente consulta:

```
datos_packhams <- dbGetQuery(conex, "SELECT*FROM ‘Crecimiento_Frutos‘
                  WHERE plantacion IN (16,17);")
```

Para seleccionar los datos de la tabla "Crecimiento_Frutos" que posee todos los registros de las curvas de crecimiento, se utilizó la sentencia de SQL "SELECT*FROM", la misma permite ejecutar una consulta a partir de una selección de los datos de la tabla que para el ejemplo anterior corresponde a la tabla "Crecimiento_Frutos", donde la sentencia "WHERE" permitió filtrar los datos identificados con la plantación 16 y 17 cuyos códigos corresponden a *"Packhams Triumph"* en sendos sistemas de conducción "Espaldera" y "Monte Tradicional". Se pueden mencionar otras operaciones sobre la base de datos mediante R como por ejemplo la corrección o modificación de datos ya exportados a la base. A menudo se suelen detectar errores que son necesarios corregir como por ejemplo errores en el tipeo de los ciclos productivos en los datos de floración donde se procedió de la siguiente manera:

```
dbExecute(conex, "UPDATE ‘Floracion‘ SET ‘Ciclo_crecimiento‘ = ’2019/20’
                 WHERE ‘Floracion‘.‘id_flor‘ IN (427,428);")
```

En este caso se utilizó la función `dbExecute()` la librería **DBI**, como en los casos mostrados anteriormente el primer argumento fue la conexión a la base, luego la sentencia UPDATE correspondiente al lenguaje SQL sobre la tabla Floración modificando el ciclo de crecimiento (que anteriormente estaba escrito como 2019/2020) a 2019/20 para los casos cuyas filas de id eran 427 y 428.

4.2 Métodos aplicados en el preprocesamiento de los datos

Posteriormente a la creación y administración general de la base de datos se procedió al preprocesamiento de los datos para cada una de las tablas de la base. El primer paso del preprocesamiento de los datos fue identificar los datos faltantes de las distintas tablas y ver a que variable o columna correspondían mayoritariamente, esto se realizó con funciones base de R y funciones como `glimpse()` y `tibble()` de la librería **tibble**.

El preprocesamiento de los datos también involucra la representación gráfica de los datos, es decir, la visualización de las registros de las distintas tablas para detectar posibles errores en la carga de datos y ver los patrones y tendencias de los mismos. Para este caso, se utilizó una librería que también es parte del ecosistema tidyverse como es **ggplot2**(Wickham (2016)). Esta librería provee una amplia gama de herramientas gráficas para la representación de diversos tipos de datos además la misma está concebida para la exploración de grandes volúmenes de datos. En esta tesis se utilizó para la presentación y graficación tanto en el preprocesamiento como en la presentación de resultados.

Continuando con los métodos implementados en el preprocesamiento de los datos, cabe aclarar que, en el caso de las curvas de crecimiento, dada las características particulares de dichos datos, en cuanto a que son datos de crecimiento en diámetro de los frutos a lo largo del ciclo, expresado en días después de plena floración, se programaron funciones ex novo que contemplaron esta característica. En primer lugar, se programó una función en R que permitió individualizar cada fruto para el cultivar ingresado y otorgarle una identificación unívoca como número de fruto y contabilizar la cantidad de mediciones para cada fruto. Esta función (ver función 8.1 del Anexo) identifica el fruto de acuerdo a todas las variables de clasificación ingresadas como año, chacra, planta, tamaño,etc.

Dado la gran frecuencia con la que suelen cometerse errores de tipeo o errores que provienen de campo ya sea en la medición como en la sustitución de frutos caídos fue necesario crear una función que permitió verificar que todas las mediciones fueran crecientes con el transcurso de los ddplf, en caso que esto no se cumpliese el registro se suprimía de futuros análisis (ver función 8.2 en Anexo). Luego, se incorporó una nueva función que permitió chequear la cantidad de registros por fruto y en caso de que la cantidad de registros fuera inferior a cuatro el mismo se descartaba (ver función 8.3 en Anexo). Una curva de crecimiento con un número inferior a cuatro registros no es posible de ajustar adecuadamente por medio del método de Gauss-Newton teniendo problemas en la convergencia y en la estimación de los parámetros.

Además fue necesario programar una función para identificar aquellos frutos que, ya sea por errores en la medición o en el tipeo, no ajustaban correctamente al modelo propuesto. Para ello se construyó un algoritmo para ajustar el modelo logístico a cada uno de los frutos para cada variedad, dicha función permitió identificar frutos cuyo modelo tuviera problemas de ajuste tanto por la cantidad de iteraciones como por el error estándar de estimación, ambos criterios fueron utilizados para separar frutos (ver función 8.5). El ajuste del modelo para los frutos individuales se realizó utilizando la función `nls()` de la librería **stats** donde el modelo se especificó por una expresión en R. Dado que esta función requería de valores iniciales para alcanzar estimaciones adecuadas y que el algoritmo converga se programó una función denominada *selfstarting* (Ritz and Streibig (2008)) que permitió hallar los valores iniciales mediante la linealización de la ecuación 2.7 y obteniendo el valor inicial para la estimación de β_0 el máximo del diámetro. Sobre dicha linealización aplicó una regresión lineal para obtener los valores iniciales correspondientes a β_1 y β_2 (Ver función 8.4 an Anexo).

En todos los casos se programó en el lenguaje de R bajo el estilo de programación orientada a objetos S3 (Wickham (2019)), puesto que en primer lugar no se encontraron funciones en las librerías que realizaran las operaciones mencionadas y por otro lado, la implementación resultaba acorde a la necesidad y practicidad.

Dentro del preprocesamiento de los datos, dada las características particulares del pronóstico fue necesario estudiar gráfica y analíticamente la relación a cosecha del peso y el diámetro de los frutos para todos los cultivares. Si bien ya habían sido estudiados los principales cultivares (Bramardi et al. (1998), Bramardi et al. (2006)), se extendió a los cultivares *"Beurre D'Anjou"* ,

Galas y algunos clones mejorados. En este caso se utilizó el modelo potencial definido como se expresa en el ecuación 4.1. Encontrar los parámetros A y B es fundamental para cada variedad ya que esta relación es la que permitió transformar el diámetro de los frutos a peso al momento de cosecha y el peso de los mismos a la clasificación en tamaños comerciales de acuerdo al envase que se considere.

$$\boxed{Peso(gr) = A.Diametro(mm)^B} \tag{4.1}$$

Para realizar esta operación se programó una función para ajustar el modelo potencial transformando por el logaritmo la variable peso y diámetro, obteniendo los parámetros y retornando los valores en la escala original (como se muestra en la función 8.6 del Anexo).

Estudiar la relación entre el peso y el diámetro permitió convertir el diámetro del fruto en peso y a partir del mismo lograr la clasificación de los frutos en un tamaño comercial. Para eso fue necesario seleccionar los envases más importante para peras y manzanas. Los envases comerciales considerados para el análisis fueron distintos para peras y para manzanas. En el caso de peras se utilizó el cajón estándar $4/5$ de 19 kilogramos y el cajón de $18,2$ kilogramos, los dos envases mayormente utilizados para mercados externos e internos. Con el mismo criterio, se utilizaron los envases caja Mark IV de 18 kilogramos y la caja de 19 kilogramos para manzana. Los tamaños comerciales de acuerdo a los pesos son los que se muestran en la tabla 4.1.

Tabla 4.1: Peso del fruto de acuerdo a los envases comerciales para peras y manzanas(FT_p fuera de la clasificación comercial: tamaño muy pequeño, FT_g fuera de la clasificación comercial: tamaño muy grande)

Envases para Manzana				**Envases para Pera**			
Caja 19 kg		Caja "MarkIV" 18 kg		Cajón estándar 4/5 de 19 kg		Cajón de $18,2$ kilogramos	
Pesos(gr)	Tamaño	Pesos(gr)	Tamaño	Pesos(gr)	Tamaño	Pesos(gr)	Tamaño
112	FT_p	$114,5$	FT_p	120	FT_p	105	FT_p
122	163	127	150	131	150	116	165
132	150	142	135	141	140	128	150
145	138	157	120	152	130	143	135
160	125	172	110	$165,5$	120	$158,5$	120
179	113	190	100	181	110	174	110
202	100	$212,5$	90	201	100	192	100
226	88	241	80	224	90	215	90
250	80	$278,5$	70	251	80	244	80
280	72	321	60	280	72	282	70
318	64	800	FT_g	800	FT_g	333	60
800	FT_g					800	FT_g

El tamaño comercial indica la cantidad de frutos que puede contener un envase de acuerdo a los kilos estipulados del envase. Entonces para una caja de 19 kg de manzana, el tamaño 125 indica que se pueden ubicar 125 frutos de un peso entre 160 y 145 grs, si tomamos el valor medio $152,5$ y asumimos 125 unidades alcanzamos el peso total de 19 kg nominal de la caja. Los pesos de los tamaños de las distintas cajas están guardadas en un marco de datos (estructura de datos más común en R ver código 8.7 del Anexo); mediante una sentencia

programada para tal fin (ver código 8.8 del Anexo), al ingresar un valor de peso busca en las tablas recorriendo cada uno de los pesos y al encontrar el correspondiente devuelve el tamaño comercial.

4.3 Métodos en el procesamiento de los datos: SVM

Luego del preprocesamiento de los datos, cabe destacar que los resultados de la aplicación de los algoritmos de data mining se muestran exclusivamente sobre el cultivar de peras *"Beurre D'Anjou"* . Se seleccionó este cultivar en particular puesto que, en primer lugar, es un cultivar de ciclo medio que permite analizar adecuadamente el tipo de patrón que desarrolla a lo largo del crecimiento del fruto, a diferencia de los ciclos de frutos cortos cuyos patrones de crecimiento son más contraídos y no desarrollan plenamente la curva del fruto. En ese sentido *"Beurre D'Anjou"* es más representativa del ciclo de crecimiento. Además, es un cultivar destinado netamente a exportación donde el calibre de los frutos es una variable limitante por lo que su estudio y predicción a cosecha son fundamentales. En contrapartida, los cultivares de ciclo largo en general poseen tamaños de frutos grandes con un patrón plenamente desarrollado que no representan inconvenientes en el tamaño final.

Ante todo es de suma importancia conocer el patrón de crecimiento de los frutos a partir de los datos completos, es decir, aquellas curvas con toda la información referida al ciclo productivo. Dichos patrones se estudiaron utilizando el método de mediciones sucesivas, el cual consistió en la selección de una parcela productiva que cumpliera las condiciones fitosanitarias y de manejo de manera que los frutos alcanzaran un desarrollo comercial satisfactorio. En dichas parcelas para cada variedad de interés se seleccionaron cinco árboles y en cada árbol se identificaron cinco frutos pequeños, cinco frutos medianos y cinco frutos grandes, los cuales fueron medidos con calibre en el sector ecuatorial del fruto. Las mediciones se efectuaron desde días posteriores al cuaje hasta días posteriores a la cosecha comercial. En caso que alguno de los frutos cayera era sustituído por uno de similares características.

A partir de los datos obtenidos de la metología de mediciones sucesivas se ajustó un modelo no lineal mixto que permitió estimar los parámetros del modelo donde la parte aleatoria del modelo correspondió a la variabilidad debido a las distintas fuentes como chacras, temporada, parcela, conducción, árbol, tipo de fruto y fruto individual. Para llevar a cabo este ajuste se utilizó la función `nlme` del paquete **nlme** en el cual se especificó el modelo a ajustar y asimismo los distintos efectos a estimar. La estimación de los parámetros de los efectos fijos se evaluó con pruebas t y F en tanto que las varianzas de los efectos aleatorios se evaluaron realizando pruebas de razón de verosimilutud así también como las funciones de correlación de los errores (Schabenberger and Pierce (2002)).

Dada el gran número de efectos aleatorios a ser testeados en el modelo no lineal mixto inicial se comenzó con los efectos cuyas estimaciones resultaban valores bajos para el parámetro considerado (Pinheiro and Bates (2000)). Las pruebas de razón de verosimilitud consistieron en

ajustar un modelo con todos los efectos llamado modelo completo y luego actualizar el ajuste sin el efecto a verificar o modelo incompleto, obteniéndose los valores de verosimilutud que luego se contrastaban con una distribución χ^2 (chi^2), si la prueba no resultaba significativa se continuaba con el modelo "incompleto", sin el efecto testeado. Después de los test ya descriptos, una vez logrado el modelo definitivo, se procedió a verificar los supuestos básicos de forma gráfica. Se obtuvieron los gráficos referidos a la normalidad como el gráfico de qqplot y los gráficos de residuos estandarizados versus valores predichos tanto marginales como condicionales a los distintos efectos significativos.

El siguiente paso en el procesamiento de los datos y la implementación de algoritmos de Data Mining(DM) fue comparar el desempeño de los algoritmos de aprendizaje con métodos de base estadística para evaluar la capacidad predictiva y sopesar ventajas y desventajas de uno y otro método. El algoritmo utilizado es el SVM o máquina de soporte vectorial, dado que resulta en uno de los más simples, eficientes y precisos particularmente en datos con patrones lineales y no lineales (Karatzoglou et al. (2004)).

La estimación de los parámetros, tanto de los efectos fijos como aleatorios, permitió luego realizar simulaciones utilizando el modelo con los valores encontrados en el ajuste y de esta manera evaluar las distintas técnicas propuestas en esta tesis. La simulación de curvas de crecimiento brinda datos suficientes para encontrar los hiperparámetros del algortimo de aprendizaje implementado en esta tesis, como se mencionó párrafo arriba, correspondiente al SVM. La importancia de realizar este paso previo residió en que la utilización directa de los datos para la determinación de los hiperparámetros aumenta las posibilidades de sobreajuste del algoritmo de manera que se propone simular curvas de crecimiento para encontrar los mejores hiperparámetros, reduciendo al mínimo las posiblidades de sobreajuste. Para la aplicación del SVM se utilizó un paquete clásico como es el **e1071** (Meyer et al. (2019b)) que implementa el algoritmo **LIVSVM** ampliamente difundido.

A partir de los datos simulados de las curvas de crecimiento (ver función 8.8 creada en R de la sección Anexo) se procedió a la calibración del o los hiperparámetros, dependiendo el caso, a partir de una grilla de valores posibles. La calibración consistió en evaluar para cada valor de la grilla mediante el método de validación cruzada 10-kfold-cv el algoritmo obteniendo alguna medida de error y una vez evaluadas todas las posibilidades de la grilla elegir el o los parámetros que menor error de entrenamiento hayan alcanzado. El proceso de calibración se realizó en dos etapas consecutivas, una primera fase donde los valores de la grilla se seleccionan utilizando una progresión geométrica con el objetivo de tener un dominio más amplios, dado que, se desconoce a priori el valor mas adecuado para los hiperparámetros. Una vez que se realiza la primera fase de calibración a la que se podría llamar "calibración gruesa", se procede a tomar valores para la grilla próximos a los encontrados como óptimos en la primera calibración esto se realiza en primer lugar para encontrar el valor más preciso y en segundo lugar porque el proceso se realiza por validación cruzada donde la partición de los datos es aleatoria, y de esta forma encontrar hiperparámetros consistentes. Este procedimiento se realizó con ayuda del paquete **caret** (Wing et al. (2019)).

La librería **caret** provee un gran número de herramientas para una ingente cantidad de algoritmos de aprendizaje facilitando tareas como la calibración, validación y predicción de los algoritmos. Las funciones utilizadas fueron principalmente `createDataPartition()` que permite particionar el conjunto de datos en datos de entrenamiento y datos de testeo, la función `train()`, corresponde a la función más importante para los algoritmos de aprendizaje, a los efectos de la calibración, requiere también de la función `trainControl()` que le confiere a la función `train` una serie de parámetros como el tipo de validación cruzada que se requiere, el número de particiones de la validación y cantidad de repeticiones, para este caso se utilizó 10-kfold-cv.

Para comparar el SVM y un modelo no lineal mixto en el crecimiento de los frutos particularmente aplicando la función `train()` se debió ingresar las variables de los datos que en el caso de regresión como variable de salida se especificó el diámetro medio de los frutos y como variable de entrada los días después de plena floración (ddplf). Luego se procedió a calibrar los hiperparámetros costo y gamma con los datos simulados probando además tanto SVM lineal como SVM radial. En este caso se simularon 300 curvas de crecimiento de frutos pequeños, medianos y grandes respectivamente. Como ya fue descripto se particionaron las curvas en datos de entrenamiento y datos de testeo utilizando los mismos datos de entrenamiento para ajustar el modelo y entrenar el algoritmo y los mismos datos de testeo para predecir el crecimiento de los frutos con ambas técnicas. En ambos casos, se utilizó el error cuadrático medio para evaluar la capacidad predictiva de ambos métodos y poder así compararlos.

Posteriormente, a partir de los datos de tamaños comerciales,se evaluó la performance del algoritmo SVM para clasificación multiclase. En este caso, se simuló el diámetro de 1200 curvas de crecimiento correspondientes a 400 frutos pequeños, 400 medianos y 400 grandes de peras del cultivar *“Beurre D´anjou”* en un ciclo de 140 días. Al momento histórico de cosecha que para el caso del cultivar es de 125 ddplf se calculó el peso del fruto a partir del diámetro simulado para dicho momento, utilizando la relación peso díametro y posteriormente lograr clasificarlos en las distintas categorías comerciales que se presentaron en la tabla 4.1.

En esta experiencia se pretende comparar entre dos posibles técnicas para realizar esta clasificación. La primera es, teniendo en cuenta que la variable de respuesta de tipo multicategórica se ajustó un modelo generalizados multinomial para realizar las estimaciones de los parámetros y predecir las clases de los tamaños comerciales. Se probaron modelos como el modelo logit multinomial con categoría de referencia aunque dadas las características de los tamaños comerciales no era de gran utilidad la comparación de las categorías con una de base ya que no es ese el objetivo. Luego el modelo logit de odds proporcionales(MOP) para el ajuste del mismo se asumió y se corroboró que las pendientes de las variables predictoras son las mismas y finalmente el modelo de logit acumulado que estima una pendiente por variable y umbral con el consiguiente aumento de las estimaciones y los grados de libertad del modelo. Para el ajuste del modelo lineal generalizado multicategórico logit de categorías de referencia se necesitó la función `multinom` de la librería **nnet** y para el modelo de odds proporcionales se requirió de la librería **ordinal** (Christensen (2019)) y la función `clm()`. Para verificar el supuesto

de proporcionalidad de los odds se realizó el test de proporcionalidad utilizando la función disponible en éste última. Para el caso del modelo logit acumulado se aplicó la función `vglm` de la librería **vgam**. Para la comparación de los distintos modelos se utilizó el criterio AIC, la logverosimilitud y el test de proporcionalidad.

La segunda técnica para clasificación es aplicar la máquina de soporte vectorial (SVM) para realizar predicciones de los tamaños comerciales a cosecha directamente desde el diámetro de los frutos en los momentos históricos del pronóstico ,es decir, a 74, 81, 88 y 95 ddplf. Ambos métodos se compararon utilizando el mismo criterio, particionando el conjunto de datos en datos de entrenamiento y datos de testeo. Con los datos de entrenamiento se entrenó el SVM para encontrar los vectores soportes y con los mismos datos se ajustó el modelo logit multicategórico. Con los datos de testeo se realizaron las correspondientes predicciones tanto bajo SVM como con el modelo logístico. En ambas técnicas se tuvo en cuenta al momento del ajuste, las variables diámetro del fruto y los ddplf y como variable de respuesta la categoría comercial de fruto para la caja de peras de 4/5 de 19 kg. El criterio de comparación consistió en obtener las matrices de confusión para ambas técnicas y evaluar los estadísticos asociados. Para ello el paquete **caret** dispone de la función `confusiónMatrix()` que requiere de los datos predichos y observados para construir la matriz y calcular los estadísticos predictivos. Entre los estadísticos predictivos o de clasificación general se presentan la exactitud ó *accuracy* y su intervalo de confianza a partir de una distribución binomial, la tasa de no información, la prueba de contraste de precisión y tasa de no información y el estadístico Kappa. Otros estadísticos calculados son la prevalencia, tasa de detección, detección de prevalencia y balanceo de precisión. Estos estadísticos son los que ofrece el paquete **caret** para la evaluación de los algoritmos frente a la clasificación. En la introducción de la presente tesis como parte de la evaluación de las predicciones de los algoritmos en la sección 2.3.2 se definieron los estadísticos anteriormente mencionados.

4.3.1 Métodos utilizados para comparar la metodología del pronóstico de producción y el SVM

En este apartado se describen los métodos aplicados para implementar el SVM en el pronóstico de producción y comparar la performance con el método de mediciones sucesivas utilizado hasta ahora. El pronóstico de producción se basa en las curvas de crecimiento para realizar las predicciones a cosecha. Tanto para construir las curvas como para entrenar el SVM se utilizaron un 70% de las curvas para ajustar y entrenar y un 30% para testear. Para referenciar las predicciones a cosecha con los tamaños comerciales fue necesario ajustar la ecuación 2.7 del modelo no lineal explicado en la sección 2.2.4, a cada uno de los frutos relevados a partir de las mediciones del diámetro a lo largo de los ddplf y realizar una predicción al momento histórico de cosecha que depende de cada cultivar. Una vez encontrado el diámetro del fruto para el momento histórico de cosecha se aplica el modelo potencial que relaciona el diámetro con el peso y que también es dependiente del cultivar, esto se realiza para estimar el peso del

fruto en ese momento. Luego de calcular el peso del fruto, se clasifica de acuerdo a los envases comerciales de interés, quedando de esta manera cada una de las curvas de crecimiento referenciadas a un tamaño comercial de los precisados en la tabla 4.1. Posteriormente, se agrupan todos los fruto de un mismo tamaño comercial y se ajustó un modelo no lineal por tamaño comercial obteniendo un patrón de crecimiento para cada uno de los tamaños, denominandolo método de los modelos no lineales (MNL).

El estudio de las curvas de crecimiento permitió también construir las denominadas tablas de raleo. Las tablas de raleo no estaban disponibles hasta ahora para el cultivar *"Beurre D'Anjou"* es por eso que se utilizaron los modelos ajustados por tamaños comerciales para realizar predicciones del diámetro de los frutos, cada 7 días a partir de los 40 ddplf hasta los 130 ddplf. Estos registros son tabulados ordenándolos por tamaños comerciales y tiempo de forma semanal.

Una vez construidas las curvas por tamaños comerciales se procede con el pronóstico propiamente dicho, que consistió en evaluar entre los 70 a 80 ddplf el diámetro de los frutos y luego a partir de las curvas de tamaños comerciales predecir el tamaño a cosecha. Dado que es muy improbable que la medición del fruto para un momento dado coincida exactamente con alguna de las curvas de los tamaños, se realizó una interpolación entre el diámetro del fruto al momento de pronóstico y al momento de cosecha. Es decir, a partir del diámetro del fruto que se desea pronosticar se detecta entre cuáles curvas se encuentra y se calcula el valor medio entre ambas curvas para un ddplf determinado, en caso que el diámetro del fruto sea mayor al diámetro promedio de las curvas se clasifica según la curva superior y en caso que el diámetro del fruto fuera menor al promedio se clasifica según la curva inferior. Este procedimiento, correspondiente al pronóstico de producción se tradujo en un algoritmo *Ex novo* en R (ver función 8.10 en Anexo), puesto que el algoritmo original se había programado en un sistema estadístico distinto. Este procedimiento se comparó con la aplicación del SVM multiclase cuyas particularidades ya han sido descriptas.

4.3.2 Métodos para evaluar el alcance de las predicciones mediante una experiencia a campo

Finalmente, para evaluar el alcance de las predicciones del SVM se realizó una experiencia en la estación experimental INTA del Alto Valle en la localidad de Contraalmirante Guerrico. La experiencia se llevó a cabo en un cuadro del cultivar de peras *"Beurre D' Anjou"* en plena producción conducido en espaldera. Consistió en demarcar e identificar 100 frutos al azar al momento habitual de relevamiento de pronóstico, específicamente el 21 de diciembre de 2018 que corresponde a los 93 días después de plena floración, momento en el cual se procedió a realizar las mediciones de diámetros ecuatoriales mayor y menor. Posteriormente, el 7 de febrero de 2019 luego de 20 días de establecida la cosecha comercial, a los 141 ddplf, se retornó y se recolectaron los frutos que habían sido demarcados. Dichos frutos fueron medidos en su sección ecuatorial utilizando un calibre digital y al mismo tiempo pesados mediante

balanza de precisión granulométricas con hasta 2 decimales. Es importante destacar que en dicha parcela no se habían realizado recolecciones escalonadas como suelen hacerse en las parcelas comerciales.

El análisis consistió en seleccionar las curvas de crecimiento de la base de datos, que poseían mediciones de hasta 141 ddplf o posteriores, para reducir el tiempo de entrenamiento y concentrar el proceso sólo en las curvas más largas que alcanzaran el tiempo requerido. A partir del diámetro medio al momento precisado, se transformaron los diámetros a peso mediante la función potencial y estos a tamaños comerciales del envase estándar de 4/5 para peras. Se sometió el SVM al entrenamiento de las curvas de crecimiento pero con tamaños comerciales referenciados a 141 ddplf considerando el valor del hiperparámetro encontrado en la calibración y la ponderación correspondiente. De la misma forma, con los datos relevados de la experiencia, a partir de los pesos, se expresó la variable en tamaños comerciales del mismo envase comercial. Luego, utilizando el SVM ya entrenado sobre los datos relevados en la experiencia de INTA se realizó una predicción de las clases comerciales a 141 ddplf a partir de los 93 ddplf. Sobre los resultados obtenidos del SVM se calcularon los estadísticos de las predicciones.

Capítulo 5

Resultados

5.1 Bases de datos

La primera fase del proceso KDD consiste en comprender el dominio de los datos, constituir la base de datos y plantear los objetivos de la aplicación del proceso KDD, es por eso que el primer resultado que se presenta en este trabajo de tesis es haber creado una base de datos que permitiera recuperar la mayor cantidad de información generada a lo largo del pronóstico. Hasta el momento los datos obtenidos en más de 20 años, no se habían logrado reunir en un único sistema, disponiéndose en diversos formatos. Los formatos de archivos trabajados originalmente eran de lo más variado donde intervenían otros SGBD(Sistemas Gestores de Bases de Datos), hojas de cálculo, archivos planos, datos que aún estaban en papel y fueron digitalizados y por último datos que estaban en formatos específicos de programas de procesamiento y análisis de datos. Una vez digitalizados todos los archivos se trabajó con formatos de extensión .csv dado que estos formatos de archivos son los de mayor compatibilidad entre los distintos softwares utilizados.

Una vez dispuestos los datos en un único formato se procedió al diseño y la creación de las tablas de datos. Para ello, se lograron sistematizar en total 117.902 registros de los distintos cultivares discriminados de la siguiente forma. Comenzando por el cultivar de peras *"Williams"* el total de registros que se han logrado sistematizar fue de 15.748 con las condiciones ya mencionadas en la sección 3.2.1 de materiales. Dado que un fruto requiere varios registros y depende del cultivar y que los cultivares como *"Williams"* son de ciclo corto, la cantidad de registros por fruto son menores que en un cultivar de ciclo largo. En el caso de dicho cultivar la cantidad total de frutos relevados fue de 1.846 frutos, logrando un promedio de nueve mediciones por fruto. Respecto del cultivar de peras *"Packham's Triumph"* , cultivar de ciclo más largo, se han sistematizado 22.293 registros, que corresponden al seguimiento de 2.025 frutos; se obtiene un promedio de 11 mediciones por fruto. Continuando con el cultivar de peras *"Beurre D'Anjou"* el total de registros sistematizados es de 13.084 datos obtenido a partir de la medición de 1.975 frutos, es decir, con un promedio de siete mediciones por fruto. En cuanto a los cultivares de manzanas, uno de los cultivars más sobresaliente es sin lugar a dudas el cultivar *"Red*

Delicious" es por eso que el total de registros obtenidos de las curvas fue de 20.649. El total de registros corresponde al seguimiento y medición de 1.389 frutos que hace un promedio de 15 mediciones por fruto. Otro cultivar históricamente importante es el cultivar *"Granny Smith"* cultivar de ciclo largo, que asimismo fue ampliamente investigado como lo demuestra los 18.255 registros cargados. El total de frutos estudiados en las distintas temporadas fue de 1059, es decir, que la cantidad de registros promedio asciende a 17 registros por fruto. Los cultivares de *"Gala"* y sus clones mejorados fueron también sistematizados y agregados a la base de datos. El total de registros sistematizados para las diversas combinaciones de portainjerto y clon asciende a 20.817 registros, como es el caso del cultivar *"Royal Gala"* en portainjerto "M4" la cantidad de registros es de 7.690 sobre 1.366 frutos, con un promedio de seis mediciones por fruto. Recientemente se ha alcanzado a sistematizar el cultivar *"Cripps Pink"* ; hasta el momento se dispone de 7.056 registros que incluyen distinta combinación de portainjertos, los cuales fueron llevados a cabo durante los años 2.017 y 2.018. El total de los registros se ha evaluado en la localidad de Contralmirante M. Guerrico y el total de frutos medidos fue de 645, es decir, que en promedio se realizaron 11 mediciones por fruto en la temporada.

Las bases de datos relacionales, como la utilizada en la presente tesis deben su nombre justamente por el manejo de los datos sobre distintas tablas que luego pueden ser vinculadas entre sí y de esta forma optimizar el espacio en memoria, facilitar el acceso a los datos y permitir el acceso de usuarios múltiples. Dentro de la base de datos se crearon tablas que alojan los datos centrales para el pronóstico de producción y tablas accesorias que eran vinculadas. Las primeras tablas creadas y llenadas eran las tablas accesorias para luego poder aprovechar dicha información en las restantes.

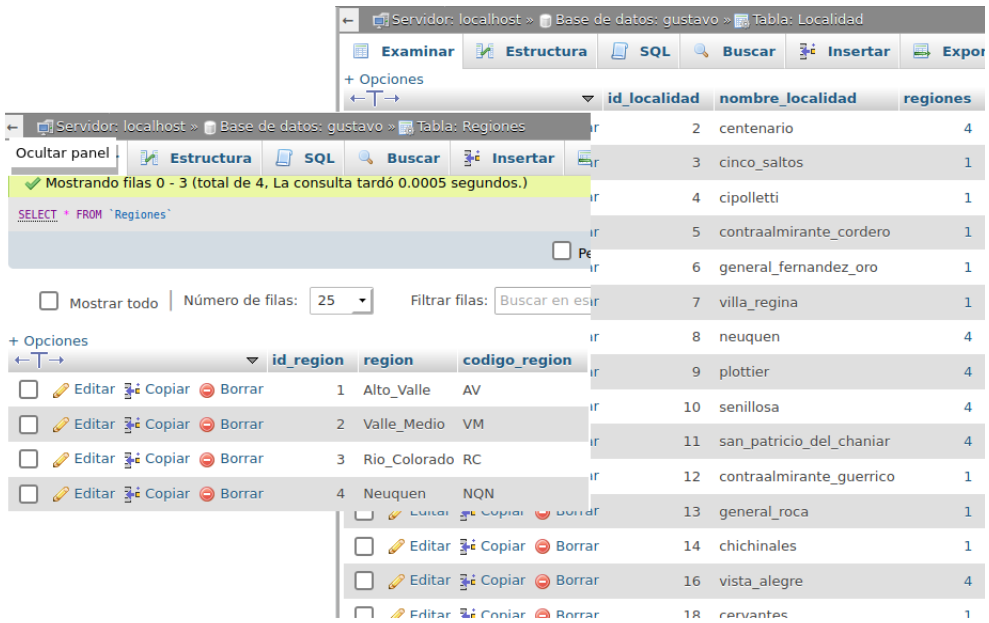

Figura 5.1: Tablas superpuestas correspondientes a las regiones y las localidades

Una de las primeras tablas accesorias que se crea es la denominada "Regiones", dicha

tabla posee las cuatro regiones que integran el pronóstico como son Alto Valle, Valle Medio, Río Colorado y Neuquén. Como se observa en la figura 5.1, los datos son identificados unívocamente con una columna que posee un número 1,2,3 y 4 respectivamente que se le otorga a cada región, la columna identificatoria se la designó asimismo como clave primaria. También se agregó un código para reducir la escritura del nombre en caso que sea necesario y asimismo hacer compatible esta tabla con la anterior base de datos. Posteriormente, se creó la tabla "Localidad" que posee todas localidades de las regiones ya mencionadas en la tabla anterior, en total son 40 localidades y a cada una se asoció a la región correspondiente con el número de la tabla regiones. Dichas tablas fueron relacionadas mediante una clave foránea entre el campo de región de la tabla "Localidad" y la clave primaria de la tabla regiones. Acción que se realiza para implementar una restricción de integridad sobre la tabla de localidades ya que al momento del llenado las regiones están limitadas a las correspondientes a la tabla "Regiones". Este tipo de restricciones que se implementaron en distintas tablas ofreció múltiples ventajas como el ahorro de espacio evitando copiar innumerables veces el nombre de la región. Otra ventaja es que en caso de cometer un error al momento del tipeo del nombre del registro sólo es necesario modificar una única vez el mismo. Otra condición que se observa en la figura 5.1 es la condición de cardinalidad "uno a muchos", donde un elemento de la tabla "Regiones" se relaciona con muchos elementos de la tabla "Localidad".

Continuando con la descripción de la base de datos, de manera similar a las tablas anteriores, se generaron tablas para contener los datos correspondientes a las variedades del pronóstico con el nombre de "Variedades", en dicha tabla además de poseer todos los cultivares de peras y manzanas se agregaron también los clones mejorados de los distintos cultivares como por ejemplo *"Galaxy"* y *"Royal Gala"* que son clones mejorados del cultivar *"Gala"* . Se creó una tabla para guardar las especies, con el mismo criterio se utilizó la tabla regiones para las localidades. Como se mencionó en la sección 2.2.1 los árboles frutales son individuos bimembres conformados por variedad y portainjerto. Por esta razón, se creó la tabla "Portainjertos" que registra todos los portainjertos implantados en las regiones estudiadas tanto para peras como para manzanas. Para describir completamente una plantación frutal se generó una tabla con el nombre de "Plantación" donde se especifica la variedad y clon, que se vincula a la tabla "Variedades", el portainjerto también relacionado a los registros de la tabla de igual manera, se agregó el sistema de conducción y el nombre resumido a manera de acrónimo para compatibilizar esta tabla con la base anterior del pronóstico.

La tabla "Chacra" fue creada con el objetivo de guardar los registros de las parcelas productivas donde fueron evaluados los frutos de alguna instancia del pronóstico. Los datos guardados fueron el nombre ficticio de la parcela, la razón social o empresa, la dirección y la localidad donde se encontraba la parcela esta columna vinculada a la tabla "Localidades" y por último se especificaba el tipo de producción de la parcela si correspondía a producción convencional, orgánico, biodinámico y en transición.

Se tuvo en cuenta además la creación de una tabla para los relevadores que participaron de los trabajos del pronóstico que permitió contemplar qué relevamiento había realizado cada

persona y teniendo en cuenta que algunas apreciaciones fueron subjetivas conocer éste aspecto para, en caso que fuese necesario, agregarlo al modelo estadístico o de aprendizaje correspondiente. Dicha estructura se la llamó “Relevadores” incluyó campos como los nombres completos, localidad de residencia y documento de identidad.

Teniendo en cuenta que los trabajos se realizaban, en su gran mayoría en parejas de relevadores y que a lo largo de los años dichos relevadores no eran las mismas personas o cambiaban de grupo de trabajo, se confeccionó una tabla “Pareja_relevadores” donde se identificaba a cada pareja de relevadores y las personas que la conformaban. También se agregó una columna con el nombre resumido para compatibilizarla con la tabla de la base de datos antigua del pronóstico. Es importante aclarar de la tabla “Pareja_relevadores” que los nombres de los integrantes se vinculaban a la tabla de “Relevadores”, es decir, que no existía necesidad de reescribir dato alguno. Una particularidad de ésta tabla, respecto de la cardinalidad, es que es una tabla muchos a muchos, ya que muchos registros de “Relevadores” pueden generar distintos registros de la tabla “Pareja_relevadores” y asimismo ésta última múltiples registros de otras tablas que van a ser descriptas a continuación. Las características de esta última son similares en cuanto a relaciones y estructura de datos a la tabla “Plantación”.

Una de las fuentes de datos más destacadas en el pronóstico es la correspondiente a la carga de los montes frutales al momento del pronóstico, es decir, el relevamiento realizado en las distintas parcelas de las regiones involucradas donde se obtenían los datos de carga frutal, diámetro de los frutos y distintas percepciones subjetivas de los montes frutales. Toda esta información fue volcada a la tabla que se le dio el nombre de “Carga”. En la tabla, además del identificador del registro, se consideraron variables como le fecha de registro, la chacra de la parcela que se relevó, el cuadro, el tipo de plantación vinculado a la tabla ya descripta párrafos arriba, la edad de las plantas que permitió discriminar el estrato productivo, la distancia entre plantas y entre filas, el número de frutos contabilizados de las plantas, la percepción que tenían los relevadores respecto de la carga frutal del monte frutal y del tamaño de los frutos. El diámetro de los cuatro frutos que debían medir por planta, es decir, ocho columnas en las cuales figura el diámetro mencionado, tres columnas más destinadas a las notas que se relevaban de los montes, la pareja de relevadores relacionadas con la tabla y finalmente una columna destinada a las notas generales de la parcela. Esta tabla de grandes dimensiones y que posee además muchas vinculaciones a partir de claves foráneas permitió compatibilizar dos sistemas de pronóstico de producción, el implementado hasta el año 2000 utilizando el método estocástico y el método de mediciones sucesivas.

También fue creada una tabla denominada “Superficie_Parcelas”, donde se volcó toda la información provista por la FUNBAPA para la expansión de las estimaciones del pronóstico de producción. La tabla en cuestión registra un primer campo de identificación unívoca, la variedad, el sistema de conducción, el portainjerto, la edad de la plantación, el número de plantas del cultivar correspondiente, la superficie ocupada por la plantación y la localidad donde se encontraba la parcela. Tanto la variedad, como el sistema de conducción, el portainjerto como la localidad están vinculadas con las respectivas tablas.

Una de las tablas más destacables de la base, considerando el método de mediciones sucesivas es la tabla que reúne los datos de curvas de crecimiento de todos los cultivares, ya descripto en la sección materiales. A la tabla en cuestión se le asignó el nombre de "Crecimiento_Frutos" y además de tener el número de registro posee las columnas de la pareja de relevadores que llevó a cabo las mediciones, el tipo de plantación que define el cultivar implantado y el sistema de conducción, la fecha de registro, los diámetros del fruto para ese momento,la planta que es identificada a campo con un número, el número de fruto y el tamaño del fruto, la chacra correspondiente y los ddplf (ver figura 5.2). Dado que no siempre las mediciones de curvas de crecimiento se hicieron en parejas en ese caso la clave foránea señala a un único relevador, dejando el segundo campo como NULL.

registro	pareja_de_relevadores	plantacion 2 variedad-pie-conduccion	fecha_registro 1	planta_n 3 planta numero	fruto_n 4 fruto numero	tamanio	diametro_mayor	diametro_menor	diametro_medio	chacra
110717	3	25	2015-01-14	1	3	M	62.38	61.15	61.765	1
110718	3	25	2015-01-14	1	4	G	71.32	70.41	70.865	1
110719	3	25	2015-01-14	1	5	P	47.2	47.64	47.42	1
110720	3	25	2015-01-14	1	6	P	61.79	62.08	61.935	1
110721	3	25	2015-01-14	1	7	P	53.13	52.89	53.01	1
110722	3	25	2015-01-14	1	8	P	59.5	60.02	59.76	1
110723	3	25	2015-01-14	1	9	M	57.77	57.81	57.79	1
110724	3	25	2015-01-14	1	10	M	66.01	63.48	64.745	1
110725	3	25	2015-01-14	1	11	P	61.61	59.71	60.66	1
110726	3	25	2015-01-14	1	12	G	62.96	62.33	62.645	1
110727	3	25	2015-01-14	1	13	M	62.52	60.19	61.355	1

Figura 5.2: Extracto de los datos de curvas de crecimiento en vista tabla

En la extracto que se muestra en la figura 5.2, la columna de pareja de relevadores, como el de plantación y chacra están vinculados mediante una clave foránea a las respectivas tablas que ya fueron mencionadas. Un aspecto importante de dicha tabla son las distintas restricciones impuestas con el propósito de evitar errores en la carga e importación de los datos. Teniendo en cuenta que no en todos las curvas de crecimiento se disponía del total de la información, es que se han permitido agregar valores faltantes (en SQL corresponde el valor NULL) a las distintas columnas de la tabla. Las columnas de diámetro mayor y menor son las que mayores valores faltantes poseían puesto que muchos relevadores registraban directamente el resultado promedio de ambas mediciones. Tanto los campos de diámetro medio como de ddplf no se admitieron datos faltantes puesto que son los datos mínimos para construir una curva de crecimiento sin los cuales no tendría sentido almacenar el registro. Dichos campos fueron configurados para admitir valores decimales, específicamente tipos de datos denominados "double", como puede observarse en el extracto de la figura 5.3. En los campos referidos a la fecha de registro, sólo se admiten datos tipo fecha con el formato "YYYY%MM%DD", es decir, año escrito completo con las 4 dígitos, mes y día con dos dígitos respectivamente. Tanto el número de planta como el de fruto y los ddplf sólo admite un número entero positivo menor a 1000. El caso de tamaño de frutos sólo se admiten los caracteres "P", "M" y "G"

correspondiente a los frutos pequeños, medianos y grandes que el relevador debía seleccionar para el seguimiento de la curva de crecimiento. Para este campo también se admiten datos faltantes puesto que las primeras curvas no utilizaban ésta notación sino que directamente se ponía el número de frutos de 1 a 75.

#	Nombre	Tipo	Cotejamiento	Atributos	Nulo	Predeterminado	Comentarios	Extra
1	**registro**	int(10)			No	*Ninguna*		AUTO_INCREMENT
2	**pareja_de_relevadores**	int(3)			No	*Ninguna*		
3	**plantacion**	int(3)			Sí	*NULL*	variedad-pie-conduccion	
4	**fecha_registro**	date			No	*Ninguna*		
5	**planta_n**	int(4)			Sí	*NULL*	planta numero	
6	**fruto_n**	int(4)			No	*Ninguna*	fruto numero	
7	**tamanio**	enum('G', 'P', 'M', '')	latin1_swedish_ci		Sí	*NULL*		
8	**diametro_mayor**	double			Sí	*NULL*		
9	**diametro_menor**	double			Sí	*NULL*		
10	**diametro_medio**	double			No	*Ninguna*		
11	**chacra**	int(5)			No	*Ninguna*		
12	**ddplf**	int(4)			No	*Ninguna*		

Figura 5.3: Extracto de los datos de curvas de crecimiento en vista diseño y las configuraciones de las columnas

En la misma tabla que se muestra en el extracto de la figura 5.3 utilizando una vista de diseño, se puede observar las distintas columnas o atributos así como los tipos de datos asignados para cada uno de ellos; se distingue con una llave amarilla las claves primarias y en llaves grises las columnas que poseen una clave foránea con otras tablas.

Otra de las tablas importantes dentro del pronóstico de producción es la que corresponde al peso y diámetro de los frutos de los distintos cultivares. La misma se definió como “Peso_Diametro” y registró además de la identificación unívoca del dato, la variedad, la chacra en la cual se extrajeron los frutos, la fecha de evaluación, el peso y el diámetro. De la misma forma que en las tablas anteriores se vinculó a las columnas “Variedad” y “Chacra” con las tablas correspondientes y se establecieron restricciones de acuerdo al tipo de dato en cada caso.

Se destinó una tabla para guardar los datos fenológicos más importantes de los diversos cultivares de peras y manzanas cuyo nombre fue “Floracion”. Se crearon columnas con el ciclo productivo, fecha del momento de plena floración y de cosecha comercial para ese ciclo productivo los días de vida o de duración del fruto, la variedad y finalmente la región en la cual se determinó la fenología. Tanto los campos de variedad y región fueron vinculadas a las correspondientes tablas y se establecieron restricciones respecto de las fechas y los ddplf.

Por último, los datos climáticos fueron guardados en la tabla de nombre “Datos_Climaticos” con todas las variables descriptas en la sección de materiales, donde se agregó el campo de la localidad indicando la ciudad donde se encontraba la estación meteorológica vinculada a la tabla de las localidades. En este caso se permitió que todas las variables pudieran tener datos faltantes excepto la columna de las fechas de registro, puesto que sin el momento de registro

sería inviable el dato climático.

Para lograr compatibilizar los datos, en una misma base de datos, de las distintas fases del pronóstico se debieron crear un total de 17 tablas. Las mismas se muestran en la siguiente imagen.

Las tablas de la figura 5.4 son las que se describieron en la presente sección y en la sección de materiales.

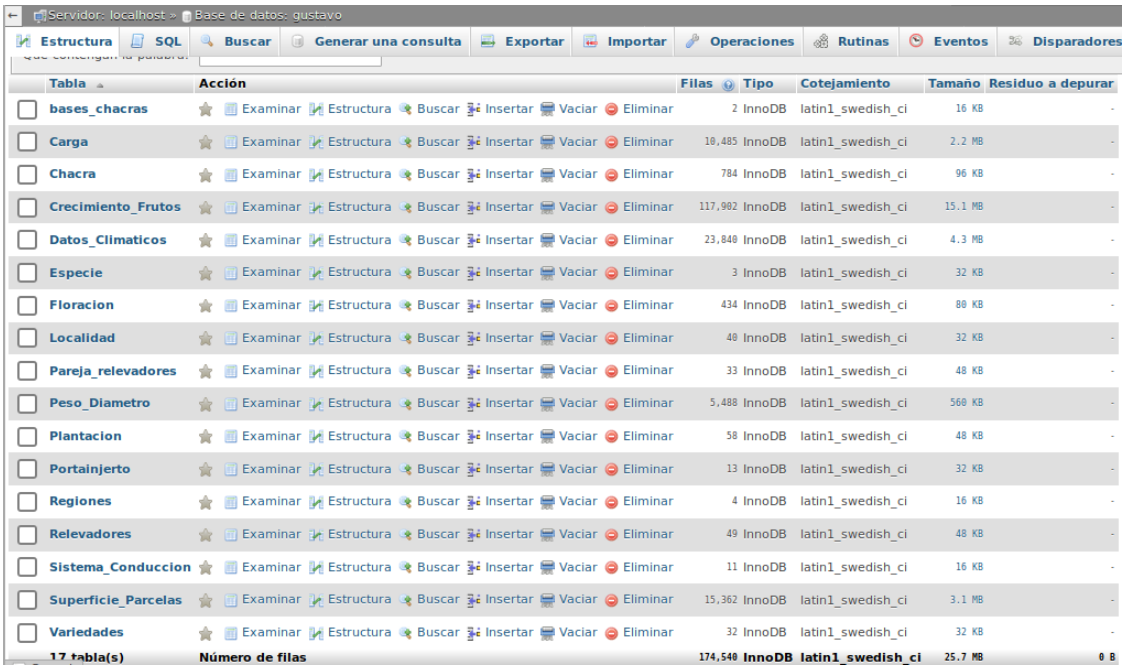

Figura 5.4: Base de datos del pronóstico con todas las tablas correspondientes

Cabe destacar que las distintas tablas se encuentran vinculadas entre sí mediante las claves foráneas con distinta cardinalidad tanto de uno a muchos como de muchos a muchos.

La vinculación entre las tablas son las que se generan o se planifican en el diseño de la base de datos, por tal motivo es que se muestra en la figura 5.5 un esquema de relaciones de todas las tablas de la base de esta tesis.

En la figura 5.5, se pueden observar todas las tablas de la base de datos, en los distintos recuadros, sin considerar los atributos de las tablas. Las líneas indican cómo se interrelacionan cada una de ellas, por ejemplo, en la parte central de la figura se encuentra la tabla “Floración” que se encuentra vinculada por las claves foráneas con las tablas “Localidad” y “Variedades”. Asimismo, la tabla Crecimiento_Frutos se encuentra vinculada con las tablas Plantación, Chacra y Pareja_relevadores. Esta vista de las relaciones, es muy útil al momento de diseñar la base y ver cuáles son las tablas relacionadas y cómo son éstas relaciones. A pesar de que son 17 las tablas que intervienen claramente se muestra cómo se encuentran relacionadas las mismas aunque no con qué columnas o atributos se relacionan.

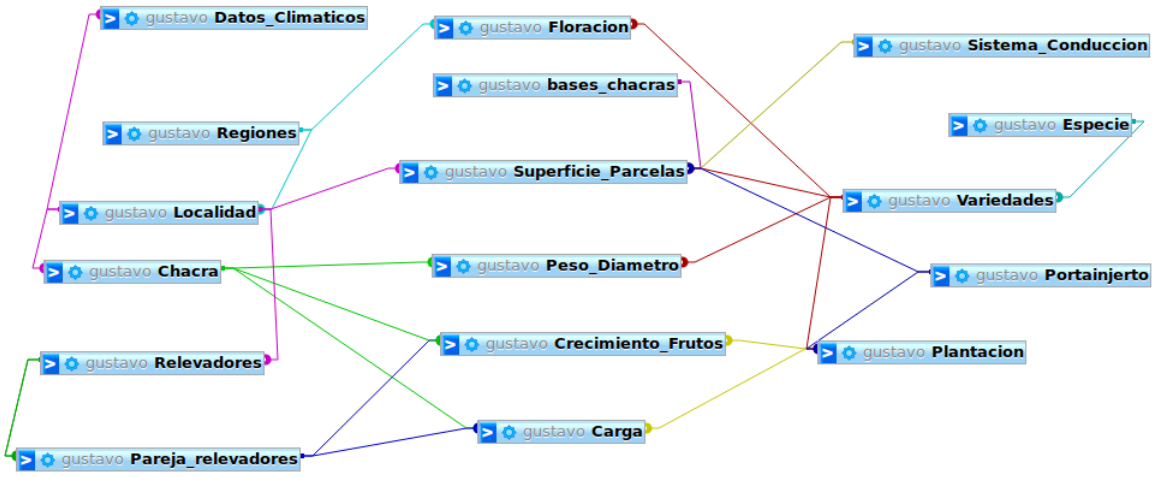

Figura 5.5: Esquema del diagrama relacional de la base de datos

5.2 Preprocesamiento de los datos

Una de las fases esenciales en la implementación del proceso KDD es el preprocesamiento de la información, como se describe en la figura 2.8 de la sección 2.3, consiste en la selección, limpieza, transformación y normalización de los datos. Es la fase del proceso que mayor demanda en cuanto tiempo y recursos computacionales se refiere.

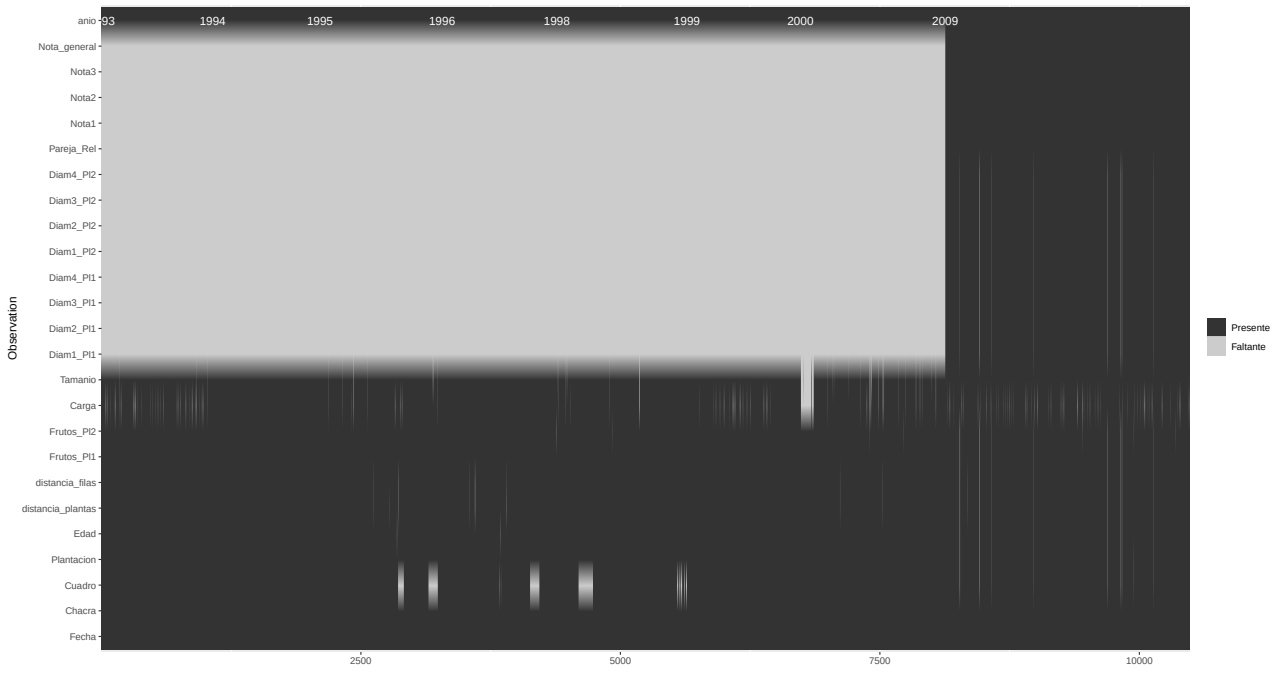

Figura 5.6: Gráfico de datos faltantes para la tabla de “cargas”, los datos faltantes se indican con color gris y los datos presentes con color negro. Las variables en el eje Y y el número de registro en X.

El primer paso es la identificación y análisis de los datos faltantes, detectar las variables que mayor cantidad de datos faltantes posee y buscar un patrón para poder arribar a la causa de los mismos, así también como determinar el impacto de los mismos sobre el análisis que se va a desarrollar.

La figura 5.6 muestra en un gráfico los datos presentes en color negro y ausentes en color gris para la totalidad de los datos de la tabla “Cargas”.

Donde se denotan en el eje “X” los registros ordenados numéricamente desde el 1 hasta el 10000 para cada una de las variables presentadas en el eje “Y”, también se observa el año del pronóstico en las etiquetas superiores. Claramente, el gráfico marca un importante contraste de gris y negro entre los primeros 2500 datos de los restantes, tendiendo en cuenta que los últimos registros corresponden a los datos de los últimos años y los registros anteriores a los diez primeros años, la falta de datos corresponde a las variables no evaluadas en los primeros pronósticos donde se prescindía de los mismos. En esta tabla el mayor número de datos faltantes se detecta en las columnas de los diámetros de los frutos, puesto que se alojan dos tipos de bases de datos correspondientes a las distintas metodologías de relevamiento de pronóstico de producción, en el primero no se registraban los diámetros por no estar basado en curvas de crecimiento sino que sólo se requería la carga de los frutos de las plantas. En tanto que los diámetros son la base del método de las curvas de crecimiento.

En esta fase la visualización representa una herramienta útil y viable a los fines de entender el comportamiento de los datos y asimismo errores en los mismos. Es por eso que para explorar las curvas de crecimiento de los frutos, representar las mediciones en gráficos resulta una alternativa muy valiosa. Para esto se graficó, como se observa en la figura 5.7, el seguimiento de las mediciones de tres cultivares de manzanas (*“Granny Smith”* , *“Royal Gala”* y *“Galaxy”* -clon mejorado de *“Royal Gala”* -) y dos de peras(*“Beurre D’Anjou”* y *“William’s”*), todos del ciclo 2.013 y 2.014.

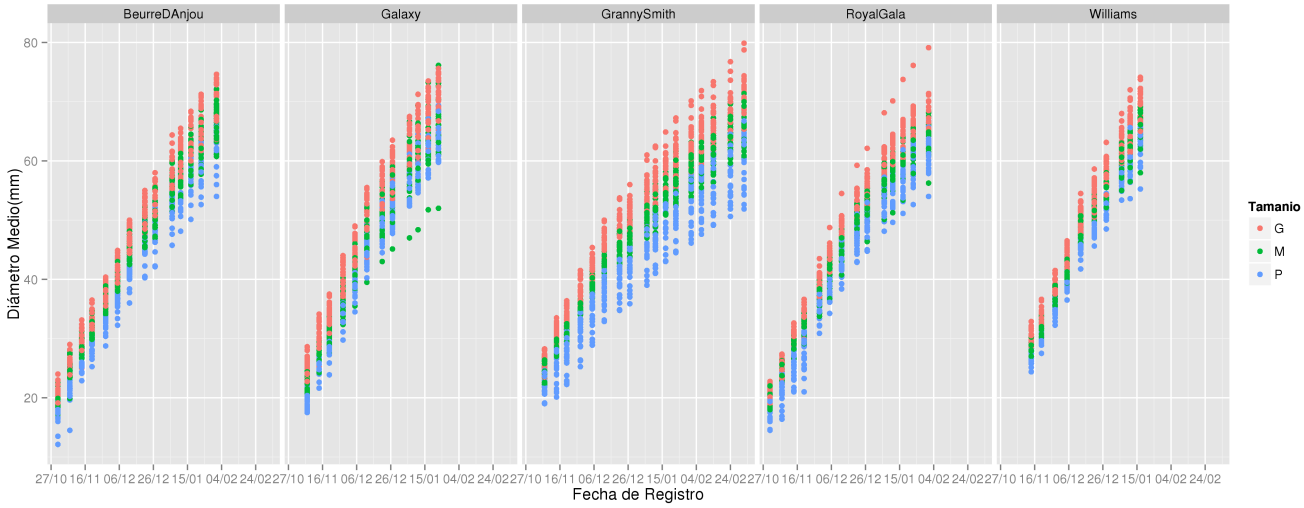

Figura 5.7: Seguimiento de frutos para 3 cultivares de manzanas y 2 cultivares de peras del ciclo 2013-2014.G:frutos grandes,M.frutos mediando P:frutos pequeños

Cada punto representa la medida promedio del diámetro ecuatorial de un fruto en una fecha determinada, el cual fue seguido en toda la temporada, hasta días posteriores al momento de cosecha. El seguimiento de los frutos en los distintos cultivares se realizó mediante la metodología descripta en la sección 3.2.1.

A partir de la figura 5.7 se puede observar el comportamiento de cada uno de los cultivares, donde las curvas presentan un patrón sigmoideo más marcado en los cultivares de ciclo largo y menos marcado en aquellos cultivares de ciclo corto. El caso de *"William's"* es un cultivar de peras de ciclo corto que a diferencia de otros cultivares de ciclo más largo, posee menor variabilidad entre los frutos y a lo largo del ciclo de crecimiento. También se observa claramente que en el cultivar *"Granny Smith"* , un cultivar de ciclo largo, los frutos presentan una gran variabilidad, desde frutos pequeños a frutos muy grandes y con un marcado patrón sigmoideo.

Por otro lado, la representación gráfica permitió identificar algunos problemas de medición, por ejemplo, en el cultivar *"Galaxy"* se observa que uno de los frutos medianos identificados con el color verde verdes en la figura 5.7, comienza a decaer hasta obtener un crecimiento menor a los frutos pequeños, no condice con el comportamiento esperado por el tamaño del fruto ni tampoco con el patrón sigmoideo esperado. Esto pudo deberse a una sustitución del fruto por otro de menor tasa de crecimiento. El mismo análisis se llevó a cabo para todos los cultivares en la mayoría de los ciclos productivos.

Con el objetivo de detectar estos frutos cuyos registros presentan alguna anomalía en la medición o en la sustitución de las unidades del árbol, se preprocesaron los datos en una primera instancia detectando registros decrecientes y posteriormente ajustando modelos no lineales a cada uno de los frutos para todos los cultivares de la base de datos. Como se describió en la sección 4 de métodos se programó un algoritmo para ajustar a partir de los datos ya depurados de registros espúreos, el modelo de la ecuación 2.12 ampliamente estudiado por Bramardi et al. (1998), Stangaferro et al. (2001) y Alvarez et al. (1996) y minuciosamente descripto en la sección 2.2.4. Teniendo en cuenta esta metodología se presenta la salida del algoritmo programado de los diez primeros frutos individuales ajustados para el cultivar de manzanas "Granny Smith".

```
##  frutind  Alpha     Beta      Gamma   predic    niter  std_err
##        1 67.07387 1.390287 0.03188456 65.82691     6 0.4801630
##        2 66.14714 1.848709 0.04250924 65.81625     5 0.2756200
##        3 59.76673 1.998131 0.04466878 59.52497     5 0.3917596
##        4 67.89647 1.668860 0.03543745 66.97366     7 0.5356242
##        5 65.41793 1.751223 0.04194307 65.09146     5 0.3724302
##        6 54.52772 1.692502 0.04394422 54.34409     5 0.2028385
##        7 55.52270 1.673307 0.04146793 55.24508     5 0.3715848
##        8 64.93210 1.583849 0.03367749 63.84610     6 0.6456934
##        9 59.26082 1.578423 0.03570046 58.55563     7 0.5931030
##       10 59.14153 1.398046 0.03628281 58.60713     5 0.3429231
```

En la salida presentada, `frutind` indica el número de fruto, `Alpha` la estimación del parámetro β_0, `Beta` la estimación del parámetro β_1 y la estimación de `Gamma`, la columna

`predic` hace referencia a una acción del algoritmo que permite predecir el calibre del fruto al momento de cosecha comercial, `niter` corresponde al número de iteraciones necesarias por el algoritmo Gauss-Newton para alcanzar la convergencia y finalmente `std_err` el error estándar de estimación. En los primeros diez frutos extraídos el número de iteraciones es bajo y el valor de error de estimación no supera $0.65\ mm$. Esto indica que los ajustes son satisfactorios y las mediciones de los frutos no acusan un problema grave al menos en cuanto a los valores hallados. A partir del postulado que un gran número de iteraciones, acompañado de un valor alto del error estándar son indicadores de un problema de ajuste y seguramente a errores en los datos, se aplicó una condición de retención de frutos con iteraciones mayores a 50 ó errores estándar mayores a 1 obteniendo para el caso del cultivar *"Granny Smith"* los siguientes resultados:

```
##   frutind      Alpha      Beta       Gamma   predic niter  std_err
##        72 1054.50784 4.1196094 0.009856257 82.71374   150  1.3213542
##       633  136.99658 0.6528773 0.003152165 64.28025    52  0.3442218
##       641  312.07146 2.1761860 0.006335161 77.24416   150  1.8255992
##       662  226.50446 1.9404314 0.007196120 73.58435    91  3.4733518
##       671  133.05578 1.1032186 0.009036458 80.13467   100  1.9908295
##      1022   87.52408 1.1727495 0.012494648 62.69651    54  2.9999405
```

En la salida se observa que los frutos 72, 633, 641, 662, 671 y 1022 superan las 50 iteraciones y en casi todos los casos tienen valores altos en sus errores estándares.

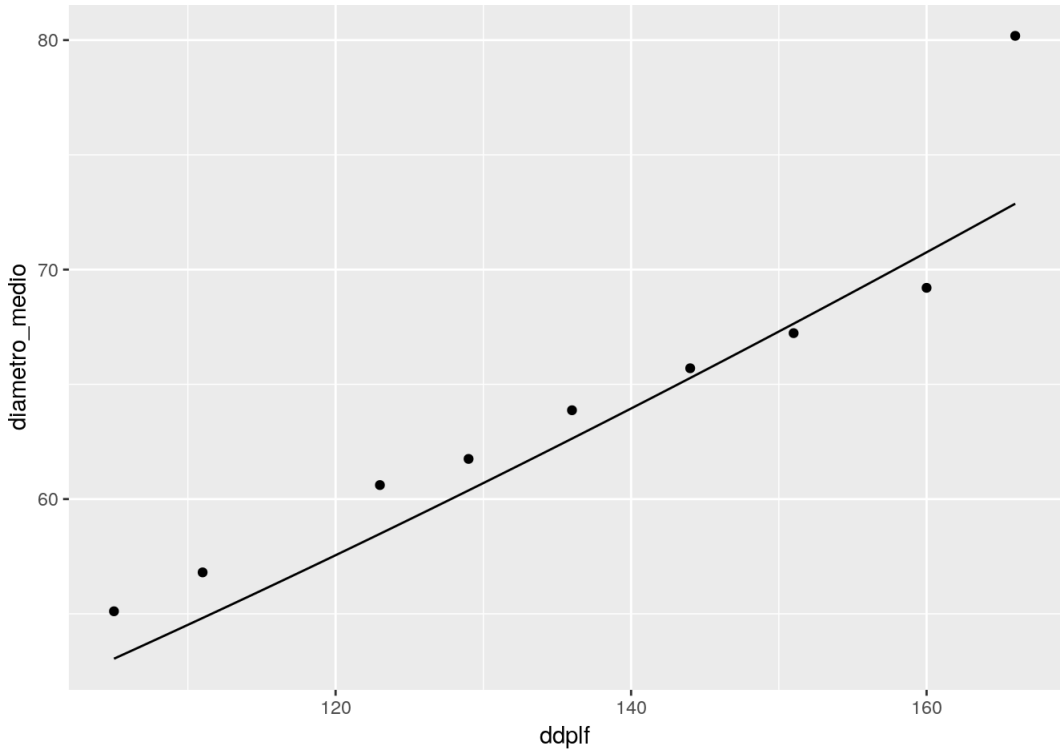

Figura 5.8: Gráfico de curva de crecimiento del fruto 72 donde se observa un alto número de iteraciones y error estándar mayor a 1 mm, detectado por el algoritmo programado (ver anexo 8.5)

Toda esta información indica problemas en el ajuste debidos seguramente a errores en los datos. Para evidenciarlo, se graficó el fruto 72 de la siguiente forma: Por otro lado, las estimaciones de `Alpha` no son valores coherentes en el sentido de que ningún fruto tendrá potencialmente diámetros de 1054 mm.

Se observa en la figura 5.8 que existe al menos un error en la medición o escritura del dato registrado luego de los 160 ddplf, por otro lado, los puntos no describen un patrón sigmoideo. Es importante la detección de los errores en mediciones de frutos puesto que pueden generar ruido al momento de predecir tamaños de frutos a cosecha y muy especialmente en el entrenamiento de los algoritmos de aprendizaje.

Este procedimiento y el chequeo de las mediciones logradas a través del algoritmo programado ha identificado frutos con problemas en los registros de crecimiento con la consecuente pérdida de esas curvas. La tabla 5.1 que se muestra a continuación contabiliza la cantidad de registros perdidos por problemas identificados en la presente fase de preprocesamiento.

Tabla 5.1: Tabla de registros procesados y eliminados luego de aplicar los algoritmos de preprocesamiento

Cultivar	Registros	Datos retenidos	$\%$ Registros Eliminados	Curvas Procesadas
"Royal Gala"	7690	6126	$20,33$	1021
"Red Delicious"	20649	20205	$2,15$	1347
"Granny Smith"	18555	16711	$9,93$	983
"Williams"	15748	14202	$9,82$	1578
"Beurre D'Anjou"	13084	9702	$25,85$	1386
"Packhams Triumph"	22293	18293	$17,94$	1663

De acuerdo a la misma tabla el cultivar que mayor número de registros tuvo suprimidos por los algoritmos debido a errores fue *"Beurre D'Anjou"* con un 24%. No obstante, la cantidad de curvas de crecimiento procesadas asciende a 1.386, más que suficiente para la aplicación de modelos estadísticos y algoritmos de DM. Por otro lado, el cultivar *"Red Delicious"* tuvo la menor proporción de datos con errores representando apenas un 2.15% de los registros. Las causas de la eliminación de datos fueron errores en las mediciones observándose diámetros de frutos que no correspondían con la medición anterior o con la posterior al momento de registro. También problemas de ajuste de los frutos o frutos que ajustaron correctamente pero que poseían un valor fuera de escala en la estimación de los parámetros. Es importante insistir en la importancia de detectar errores en los registros y eliminarlos para evitar que, especialmente, que sean fuente de datos de los modelos estadísticos ya que son los que mayor dificultad poseen en la convergencia de los algoritmos. Se desconoce la causa por la cual en algunas variedades de la tabla 5.1 se observan mayor pérdida de datos que en otras. Podría sospecharse que algunos cultivares como *"Royal Gala"* los frutos son raleados más fuertemente que otros cultivares y genere mayor pérdida de frutos en la medición.

El estudio de las curvas de crecimiento está íntimamente relacionado, especialmente en el

pronóstico de producción, al peso del fruto para un momento y un diámetro determinado. Es por eso que, la visualización de los datos también cobra interés cuando se busca relacionar el peso del fruto con el diámetro a cosecha, es decir, transformar una medición realizada en milímetros a gramos. Como se describió ampliamente en la sección 4, encontrar la relación peso-diámetro permite determinar a cosecha el tamaño comercial del fruto. Es por eso que se debe conocer la relación entre el diámetro de los frutos en milímetros y el peso en gramos para los cultivares estudiados en el pronóstico de producción. A continuación, en el gráfico 5.9 se presenta la relación de los diámetros de los frutos y sus correspondientes pesos, cada color representa un cultivar distinto y cada uno de los puntos representa un fruto con un diámetro definido en "X" y con un peso definido en el eje "Y". A partir del gráfico se puede deducir una relación potencial entre las variables diámetro y peso para todos los cultivares, regido por la ecuación 4.1. En el mismo gráfico se destaca que cada cultivar tiene un comportamiento distinto, regido por el mismo patrón potencial pero que difieren en los valores de los parámetros.

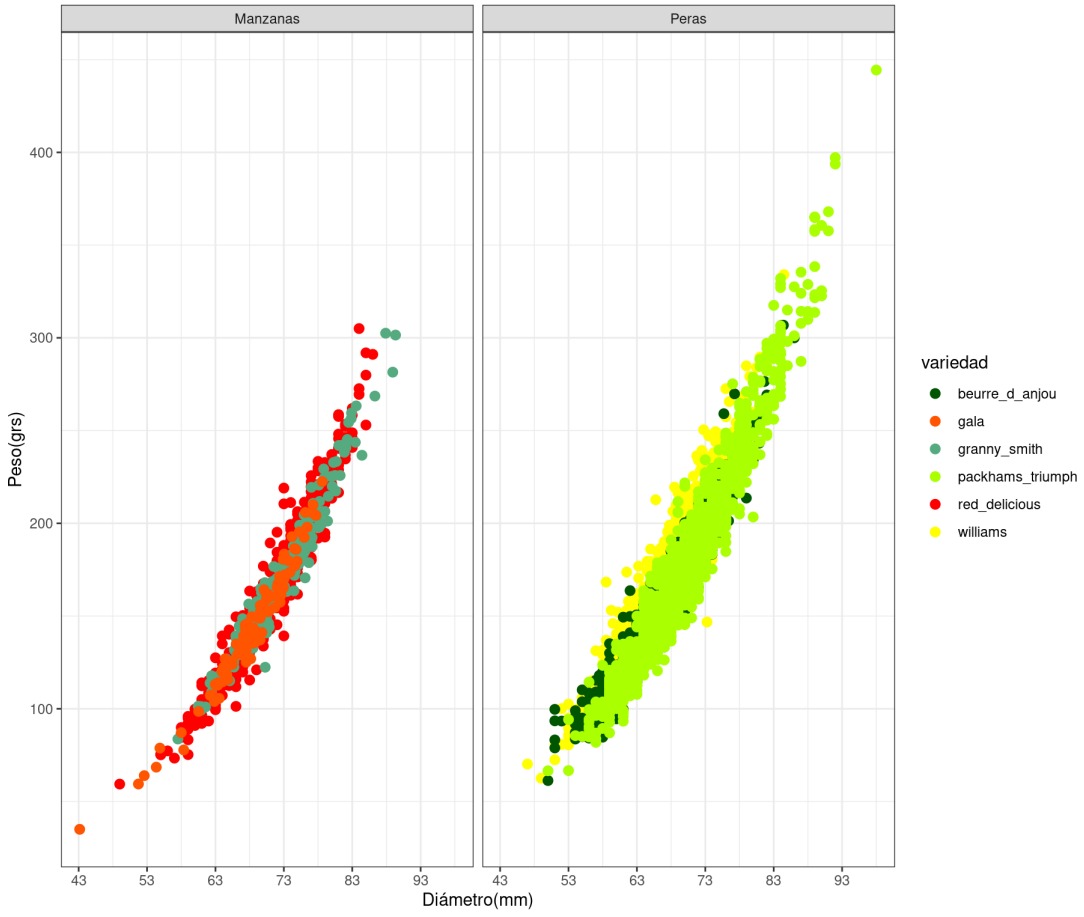

Figura 5.9: Gráfico de la relación entre el peso en gramos y el diámetro en milímetros de los principales cultivares de peras y manzanas.

Se puede resaltar que los patrones de los cultivares de peras son similares destacándose que el cultivar *"Packham's Triumph"* posee frutos de mayor diámetro y peso, es decir, posee un rango de datos más amplio que los otros cultivares de peras. Mientras que en manzanas, el

cultivar *"Gala"* se asemeja al cultivar *"Granny Smith"* , aunque este último tiene mayor rango de datos. Las diferencias más importantes se observan entre las peras y las manzanas donde las peras son de mayor peso para un mismo diámetro que las manzanas en particular *"Gala"* y *"Granny Smith"* . Las diferencias entre peras y manzanas se evidencian al observar los patrones del cultivar *"William's"* que se denotan con puntos de color amarillo y cotejarlo con el patrón de *"Granny Smith"* demarcado con puntos verde-oliváceo. Para un fruto de un mismo diámetro por ejemplo de 60 milímetros, *"William's"* alcanza un peso aproximado de 140 grs en tanto que *"Granny Smith"* de 100 grs, poniendo de manifiesto que el cultivar de pera posee un fruto de mayor peso que este cultivar de manzanas. Por dicha razón, se hace necesario el ajuste y la estimación de los parámetros A y B del modelo 4.1 para cada uno de los cultivares analizados.

En el caso del modelo potencial, el parámetro B es un factor de forma, a medida que los frutos se asemejan más a una esfera el factor tiende a tres en tanto que cuando predomina la altura del fruto por el diámetro el valor tiende a alejarse de tres. Por este motivo, es que se observan en la tabla 5.2 diferencias entre los valores de B para peras y manzanas, tendiendo a ser mayores esta última especie a excepción del cultivar *"Granny Smith"* que tiende a ser una manzana de mayor altura respecto a su diámetro.

Tabla 5.2: Estimación de los parámetros A y B y del coeficiente de determinación ajustando el modelo potencial para cada cultivar

Especie	Cultivar	A	B	R^2
Peras	William's	0.001652	2.7398178	0.91
	Beurre D'Anjou	0.002489	2.6272006	0.94
	Packham'S Triumph	0.000995	2.8401246	0.94
Manzanas	Gala	0.000379	3.0365703	0.97
	Red Delicious	0.000442	3.0017283	0.93
	Granny Smith	0.001077	2.7917068	0.96

El comportamiento del peso y del diámetro se describe satisfactoriamente a partir de un modelo potencial, en todos los cultivares tanto de peras como de manzanas, los valores de R^2 son mayores a $0,9$ y los residuales (no mostrado) no indican problemas de ajuste ni de heteroscedasticidad.

En base a los parámetros estimados de la tabla 5.2 se presenta en la figura 5.10, a manera de ejemplo, la relación peso en función del diámetro ecuatorial para el cultivar *"Granny Smith"* . El peso se expresó en gramos y el diámetro correspondiente a la sección ecuatorial de los frutos, en milímetros. En base al gráfico y con las estimaciones del modelo potencial podemos predecir el peso de los frutos en un rango de 55 milímetros a 90 milímetros.

La estimación del modelo potencial permite realizar, teniendo en cuenta el R^2, predicciones del peso de los frutos por medio del diámetro, con gran precisión, para las variedades estudiadas permitiendo además la conversión al tamaño comercial correspondiente. La aplicación del modelo resulta de la siguiente manera: para un fruto de la variedad *"Granny Smith"* que posee

$70\ mm$ se estima a partir de la ecuación 4.1 y los valores de A y B, un peso medio de $152,47\ gr$ que también se puede proyectar en el gráfico 5.10.

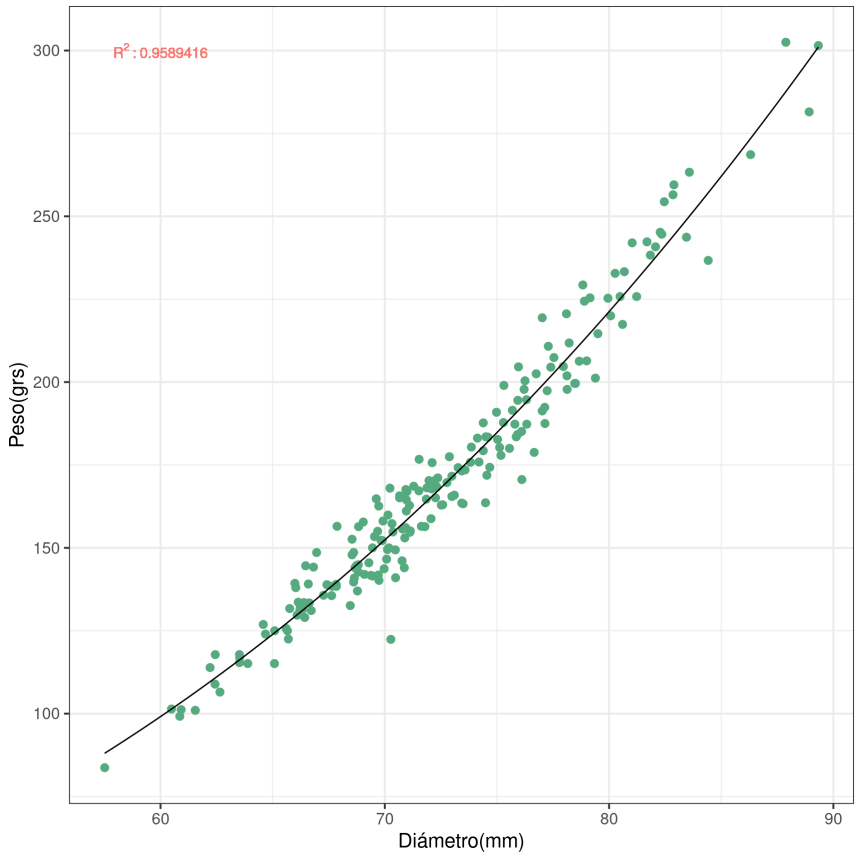

Figura 5.10: Ajuste de un modelo potencial del peso del fruto(en gramos) vs el diámetro(en milímetros) en fecha de cosecha del cultivar “Granny Smith”

El ciclo de crecimiento de los frutos, y por lo tanto la edad del fruto, está definido por la fenología de los cultivares en particular dos eventos como son la plena floración y la cosecha comercial. Además, la transformación de diámetros a peso y la clasificación comercial está referenciado al momento de cosecha comercial y la edad histórica del fruto. Por ello, un aspecto importante en el preprocesamiento de los datos fue la exploración de los registros fenológicos de los principales cultivares. Dentro de la fenología del fruto cobra particular importancia el momento de plena floración, el momento definido como cosecha comercial y la duración del ciclo de crecimiento.

El momento que se determina como inicio del ciclo de crecimiento es la plena floración y se define como la fecha en la cual se produce el 50% de apertura de las flores de un monte frutal. La cosecha comercial, en tanto, se define en función de las características organolépticas del fruto y a partir de muestreos sobre los que se evalúa variables como acidez, azúcares totales, firmeza de la pulpa, degradación de almidón, etc. y por lo tanto varía de temporada a temporada. Por último, la duración del ciclo está calculado en días y corresponde a la diferencia entre la fecha de cosecha comercial y la plena floración. En la figura 5.11 se presentan las fechas de

plena floración, con puntos verdes y cosecha comercial con puntos rojos para el cultivar de peras *"William's"* , el segmento que une ambos puntos es el largo del ciclo del fruto.

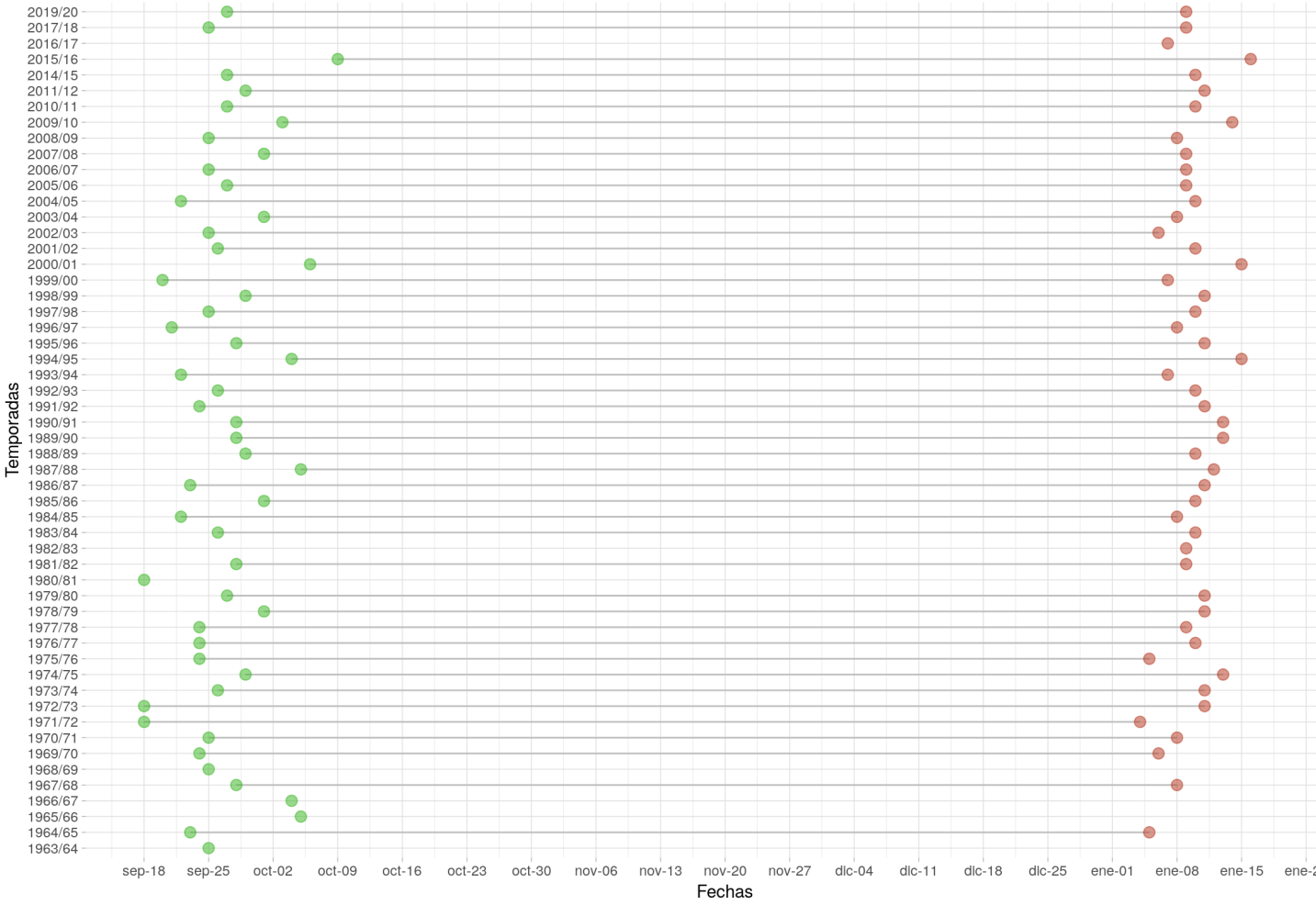

Figura 5.11: Fechas de plena floración(puntos verdes) y cosecha comercial(puntos rojos) del cv. William's para todas las temporadas registradas desde el año 1964 hasta el año 2020

Los primeros registros fenológicos del cultivar datan de la temporada 1963/1964. Para el caso de peras *"William's"* por ejemplo la duración histórica de la vida del fruto es de 107 días. Cabe destacar que algunos ciclos productivos por distintos motivos no presentan registro ya sea de cosecha comercial o de plena floración, que se representa en la figura 5.11 como un único punto. Para todos los cultivares estudiados se realiza un análisis similar, puesto que en la predicción de los tamaños a cosecha se referencian las curvas de crecimiento a la duración histórica del ciclo del fruto.

Es muy importante considerar las variables climáticas, puesto que las mismas tienen un efecto sobre el crecimiento de los frutos e indirectamente sobre la predicción de la producción a cosecha. Dichas variables pueden explicar los errores de predicción y algunos desfasajes que surgieron en años anteriores, en el pronóstico de producción, y mejorar sustancialmente las futuras predicciones. Es por ello, que se registran los datos climáticos de estaciones

meteorológicas que son representativas no sólo por el sitio dentro del valle, sino por sus mediciones históricas. Por tal motivo, se realizó un análisis exploratorio gráfico para conocer el comportamiento de dos variables climáticas, la precipitación y la temperatura media. Tanto los registros de temperaturas como de precipitación que se presentan en la figura 5.12 fueron datos del servicio meteorológico nacional cuyo parque meteorológico se sitúa en el aeropuerto de la ciudad de Neuquén y presenta la serie de datos más completos de la región.

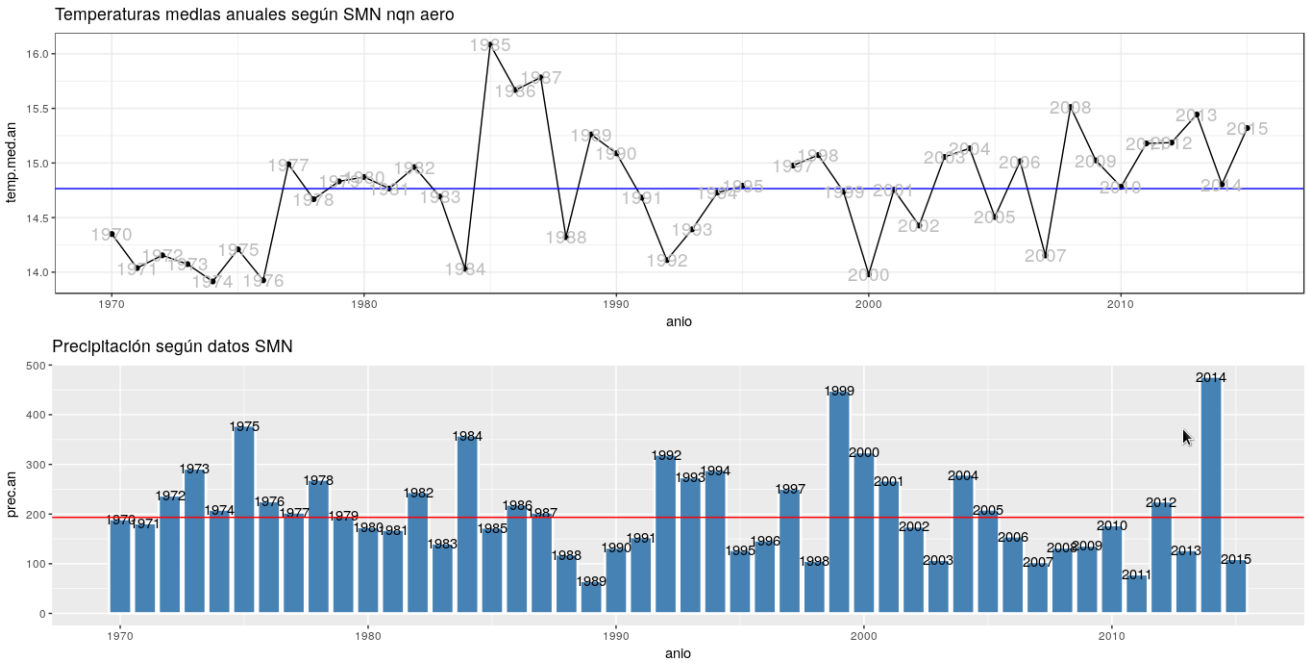

Figura 5.12: Gráfico de temperaturas medias y precipitación anual del período comprendido entre 1970 y 2015 (En el panel superior la línea azul representa la temperatura media histórica del período, en el panel inferior la precipitación media histórica)

En primer lugar, cabe destacar que el gráfico se dispone en dos paneles, donde el panel superior está destinado a las temperaturas medias anuales y el panel inferior a las precipitaciones anuales con una serie de tiempo que comprende los años 1970 a 2015. En el panel superior se denota con una línea de color azul continua la temperatura media de la serie estudiada, que ronda los $14,5°C$ y en base a ésta se puede definir períodos y años más cálidos o más fríos. En el caso de los años 1970 a 1976 fue un período de temperaturas medias inferiores a la media histórica, luego se destaca los años 1984, 1992 y los años 2000 y 2007 también con temperaturas inferiores a la media. En tanto que las temperaturas que superan ampliamente la media corresponde al ciclo 1985 a 1987 alcanzando temperaturas medias de $16°C$. Luego, los años 2008 y 2013 se presentan como ciclos cuyas temperaturas merezca resaltar llegando a los $15,5°C$. En los años 2007 y 2008 se sucedieron dos períodos en los que existieron contrastes climáticos dado que el año 2007 fue un año de temperaturas medias frías mientras que el año 2008 fue un año de temperaturas más cálidas.

En el caso del panel inferior de la figura 5.12, la línea roja contínua indica la precipitación media histórica que se encuentra en torno a los $190\ mm$ tal cual lo descripto en la sección 2.1.1

del presente trabajo de tesis. Si bien la precipitación en la región no es una variable climática que sea sobresaliente, es interesante resaltar algunos aspectos como por ejemplo el hecho de que hasta el año 1990 la precipitación fue bastante homogénea. En tanto que a partir de década señalada se observan períodos de escasa precipitación como los años 2002 y 2003 y el período 2006 a 2011 donde algunos años la precipitación cayó a hasta $100\ mm$, alternándose con años como 1999,2000 y 2014 donde la precipitación se duplicó y en el caso particular del año 2014 alcanzó casi los $500\ mm$, con una importante ocurrencia durante los meses de septiembre y octubre (datos no mostrados). Algunos años como como por ejemplo 1992, 1993 y 2000 se destacan por ser años de mayor precipitación a la media y temperaturas medias anuales por debajo de la media histórica, estos años se esperaría que el crecimiento de los frutos fuera inferior a los diámetros normales. En contraste, los años 1989,1990,2008,2013 y 2015 se caracterizan por temperaturas medias superiores y precipitaciones bajas en relación a los valores históricos.

Dado la importancia que poseen las temperaturas en especial en el período de desarrollo de yemas, floración y crecimiento de los frutos es válido presentar los gráficos correspondientes a las temperaturas medias de esos meses para el período en estudio. Los meses de agosto a diciembre, resultan críticos para el desarrollo del frutos en virtud de que las temperaturas resultan limitantes para el crecimiento de los órganos vegetales (Warrington et al. (1999)). Por ello, se realizó un preprocesamiento para dichos meses utilizando la variable de temperaturas medias y los años de análisis aplicando una regresión LOESS, es decir, una regresión ponderada localmente.

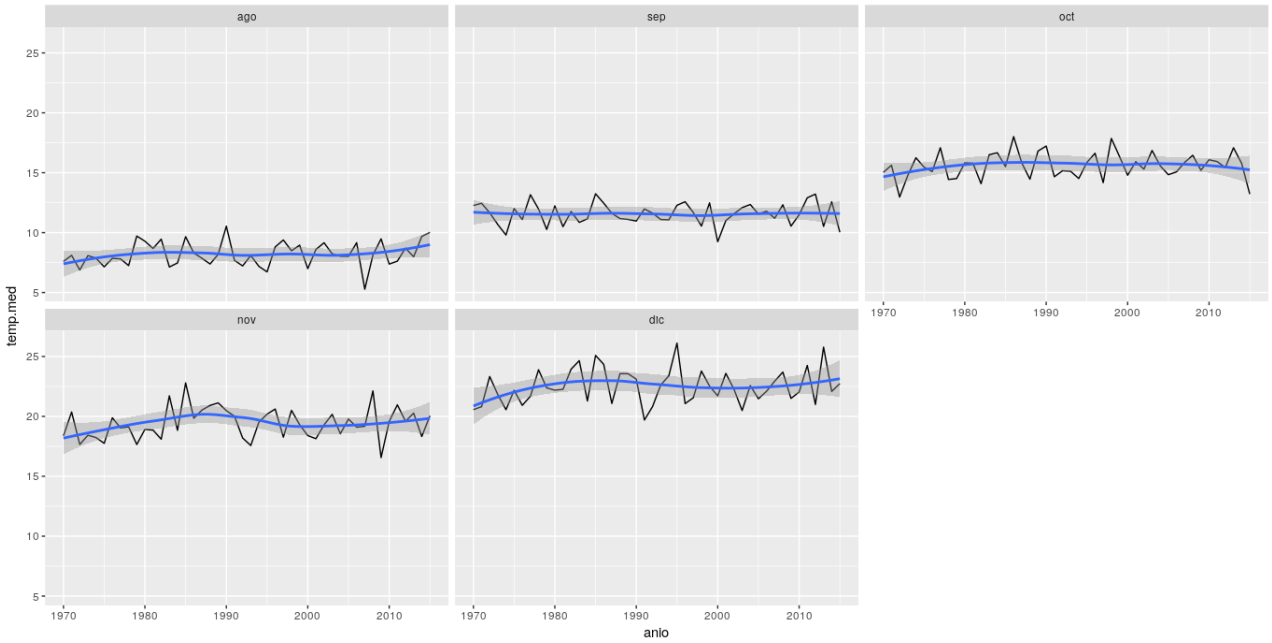

Figura 5.13: Gráfico de temperaturas medias modeladas mediante LOESS para los meses críticos de crecimiento del fruto en la serie de años de 1970 a 2015

En el gráfico 5.13 se presentan las temperaturas medias mensuales y los años del ciclo

estudiado separados en paneles, donde cada uno de ellos corresponde a los meses de agosto, septiembre, octubre, noviembre y diciembre por separado.

La línea negra representa la evolución de las temperaturas medias mensuales a lo largo de los años, la línea azul el ajuste de la regresión LOESS y el sombreado la franja de confianza. A partir del gráfico de la figura 5.13 se puede observar el aumento escalonado de las temperaturas medias a lo largo de los meses, destacándose el mes de octubre como el mayor incremento de temperaturas medias. Por otro lado, se puede observar que los meses de agosto, octubre, noviembre y diciembre presentan desde los años 1970 a 1980 un aumento, que es leve en agosto pero más marcado en noviembre y más aún en diciembre. Mientras que para el mes de septiembre no se observa tendencia alguna sino que las temperaturas se encuentran estables en el ciclo.

5.3 Aplicación de algoritmos de Data Mining

5.3.1 Aplicación de algoritmos en el patrón de crecimiento

La presente sección se basa en la aplicación de técnicas de minería de datos, como primera aproximación, a curvas de crecimiento y a la predicción de tamaños comerciales de frutos a cosecha. La comprensión del fenómeno de crecimiento de los frutos ha llevado a la aplicación de diversos modelos estadísticos y matemáticos que describen con mayor precisión el desarrollo de los frutos. Se pueden mencionar distintas aproximaciones para describir el fenómeno: el ajuste de modelos no lineales, con excelentes propiedades de no linealidad y de curvatura interna (Bramardi et al. (1998)), ajustes que contemplen la correlación de las mediciones, modelos no lineales mixtos y la aplicación de modelos no lineales mixtos con inferencia bayesiana. La comprensión del fenómeno ha permitido lograr una mejor predicción del calibre de los mismos y por ende de los tamaños comerciales al momento de cosecha.

Uno de los mayores desafíos que se presentan en los últimos años es la modelación de las curvas de crecimiento y de la predicción de los tamaños comerciales a cosecha, contemplando distintas variables climáticas. La importancia que revisten las variables climáticas en el crecimiento de los frutos es cada vez mayor. Y esto se debe seguramente,a que año a año las estaciones meteorológicas dan cuenta de mayores variaciones climáticas en muchas zonas del país.

Si bien los modelos ajustados han dado excelentes predicciones y además poseen cualidades al momento de interpretar los parámetros hallados, como así también la de permitir realizar inferencias sobre los mismos, la gran cantidad de curvas de crecimiento construidas y la enorme cantidad de datos registrados hace que la aplicación de los complejos algoritmos de máxima verosimilitud y de optimización no lineal haga muy difícil la tarea computacional de estimación. Sumado a la necesidad de modelar efectos fijos y efectos aleatorios donde intervienen: parcela, planta, tamaño del fruto, temporada, etc.; en cada cultivar y a lo largo de todos los ciclos productivos hace que sea prácticamente imposible la utilización de toda esta información en

futuras predicciones.

En los últimos años han surgido desde las ciencias de la computación y en particular del aprendizaje de máquinas una serie de técnicas computacionales que poseen gran capacidad predictiva y de clasificación. Una de las técnicas de aprendizaje supervisado, es la llamada máquina de soporte vectorial (SVM), basado en ideas simples a partir de la teoría del aprendizaje estadístico, surgida en la década de los '90 fue ampliamente difundida debido a su versatilidad y excelentes predicciones en gran variedad de situaciones y de conjuntos de datos (Karatzoglou et al. (2006)). La simplicidad de esta técnica proviene de aplicar un clasificador linear a los datos en un espacio hiperdimensional. Esta técnica requiere entre otras cosas la calibración de los hiperparámetros del algoritmo. Los hiperparámetros dependen del tipo de método a aplicar y del objetivo de la aplicación del algoritmo,es decir, del tipo de kernel asociado a los datos que se desea analizar. Por ejemplo, si se desea aplicar un método de clasificación cuyos datos son separables linealmente sólo se requiere calibrar el parámetro costo, para el caso de regresión los parámetros a calibrar son γ y $costo$ como se indica en la ecuación 2.70.

En esta primera experiencia de aplicación del SVM se comenzó con la calibración de los hiperparámetros de los algoritmos, utilizando el método "regresión-epsilon". Para lograr la calibración se propone la simulación de curvas de crecimiento como datos de validación para evitar problemas de sobreajuste y lograr predicciones de la mayor exactitud posible. Las curvas simuladas van a ser una fuente de datos de validación necesaria para encontrar los hiperparámetros de los algoritmos de aprendizaje tanto para regresión como para clasificación. Los modelos no lineales mixtos, como se describió en la sección 2.2.4 brindan grandes ventajas al momento de ajustar los datos referidos a crecimiento de frutos y en este caso permite estimar no sólo los efectos fijos sino las fuentes de variabilidad que intervienen en la generación de los datos. A partir de la estimación de los parámetros poblacionales del modelo y conociendo la variabilidad que generan los distintos efectos introducidos en los datos, se puede proceder a generar datos mediante un proceso de simulación.

Para realizar el ajuste del modelo se utilizaron datos correspondientes a registros de crecimiento de peras cv. *"Beurre D'Anjou"* de las temporadas 11/12, 12/13, 13/14 y 14/15 de tres parcelas representativas del Alto Valle donde se seleccionó una parcela en la localidad de Cipolletti, una segunda en la localidad de General Fernandez Oro y otra parcela en Centenario provincia de Neuquén. En la parcela de Fernández Oro se consideraron además dos sistemas de conducción libre y espaldera, el sistema de conducción libre sólo se hallaba en esta unidad productiva. En el análisis se consideraron cuatro temporadas en tres parcelas y dos sistemas de conducción, en cada parcela y sistema de conducción los cinco árboles frutales, para cada árbol frutal se identificaron tres grupos de frutos: pequeños, medianos y grandes. Finalmente, de cada tamaño se tuvieron en cuenta cinco frutos, es decir, que por árbol se midieron 15 frutos y por parcela y sistema de conducción 75 frutos a lo largo del ciclo de crecimiento. En base a dicha estructura de los datos se propone ajustar el modelo no lineal mixto expuesto en la ecuación 5.1. El modelo propuesto no había sido ajustado hasta el momento en ninguno de los cultivares y no se ha encontrado registro de un modelo similar para el cultivar en cuestión. Se

plantea la ecuación que describe la curva no lineal ya mencionada en la sección 2.2.4 y que mejor describe el desarrollo ontogénico de los frutos. Los parámetros Φ_1,Φ_2 yΦ_3 se encuentran expresados como un modelo multinivel.

$$Diametro = \frac{\Phi_{1_{ijklmn}}}{1 + exp(\Phi_{2_{ijklmn}} - \Phi_{3_{ijklmn}} * ddplf)} + \varepsilon_{ijklmn} \tag{5.1}$$

donde :

$$i : 1 \dots S, j : 1 \dots S_i, k : 1 \dots S_{ij}, l : 1 \dots S_{ijk}, m : 1 \dots S_{ijkl}, n : 1 \dots S_{ijklm}$$

Donde i corresponde el i-ésimo nivel de temporada (cuatro), j la j-ésima parcela productiva (tres) aunque no evaluadas en todas las temporadas, k el k-ésimo sistema de conducción (dos), l la l-ésima planta (cinco), m el m-ésimo tamaño de los frutos (Grandes, Medianos y Pequeños) y finalmente n el n-ésimo fruto individual(cinco) para cada tamaño.

Efectos aleatorios (continua modelo no lineal mixto ecuación 6.1) :

$$\begin{aligned}
\Phi_{1ijkl} &= \beta_0 + b_{0i} + b_{0i,j} + b_{0ij,k} + b_{0ijk,l} + b_{0ijkl,m} + b_{0ijklm,n} \\
\Phi_{2ijkl} &= \beta_1 + b_{1i} + b_{1i,j} + b_{1ij,k} + b_{1ijk,l} + b_{1ijkl,m} + b_{1ijklm,n} \\
\Phi_{3ijkl} &= \beta_2 + b_{2i} + b_{2i,j} + b_{2ij,k} + b_{2ijk,l} + b_{2ijkl,m} + b_{2ijklm,n}
\end{aligned}$$

Matrices de varianza covarianza (continua modelo no lineal mixto ecuación 5.1) :

$$b_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_1}), b_{i,j} = \begin{bmatrix} b_{0i,j} \\ b_{1i,j} \\ b_{2i,j} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_2}), b_{ij,k} = \begin{bmatrix} b_{0ij,k} \\ b_{1ij,k} \\ b_{2ij,k} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_3}),$$

$$b_{ijk,l} = \begin{bmatrix} b_{0ijk,l} \\ b_{1ijk,l} \\ b_{2ijk,l} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_4}), b_{ijkl,m} = \begin{bmatrix} b_{0ijkl,m} \\ b_{1ijkl,m} \\ b_{2ijkl,m} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_5}), b_{ijklm,n} = \begin{bmatrix} b_{0ijklm,n} \\ b_{1ijklm,n} \\ b_{2ijklm,n} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Psi_6})$$

$$\varepsilon_{ijklmn} \sim \mathcal{N}(0, \sigma^2)$$

donde:$\Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, \Psi_6$; son matrices diagonales

Función de Correlación Autorregresiva Continua (CAR1) :

$$h(S, \phi) = \phi^S \quad S \geq 0, \phi \geq 0$$

Los efectos fijos y aleatorios del modelo están expresados de acuerdo al criterio de Laird and Ware (Laird and Ware (1982)), donde β_0 representa el efecto fijo de la asíntota superior en tanto que $b_{0i} \dots bo_{ijklmn}$ representan los efectos aleatorios para cada uno de los niveles señalados en los subíndices. De la misma manera, para β_1 indica el efecto fijo que caracteriza la asíntota

inferior y $b_{1i} \dots b_{1ijklmn}$ el efecto aleatorio de los niveles respectivos a los subíndices. El último parámetro, β_2 como efecto fijo del parámetro que caracteriza la tasa de crecimiento del fruto y los respectivos efectos aleatorios $b_{2i} \dots b_{2ijklmn}$.

En el modelo propuesto, los efectos aleatorios conforman una matriz de varianza covarianza que expresan la variabilidad respecto de los parámetros para cada nivel jerárquico. Es por eso que en el modelo de la ecuación 5.1 se expresan las matrices de varianza covarianza referida a los parámetros para los seis niveles considerados. Se modeló para todos los casos una misma estructura de matriz correspondiente a la matriz diagonal que sólo considera la estimación de los efectos aleatorios b_0, b_1, b_3 y no la correlación entre ellos asumiendo una distribución normal multivariada independiente cada una de ellas e independiente del error final. Esta estructura de las matrices de varianza-covarianza permitía una mejor convergencia de los algoritmos y fue en definitiva la estructura que mejor ajustó a los datos. Como las mediciones se realizaron sobre los mismos frutos se considera modelar la correlación de los errores en la matriz marginal, se testearon diferentes funciones de correlación disponibles en el software donde resultó que la autoregresiva contínua de orden 1, mejor describió la correlación de las mediciones y al mismo tiempo la de menor valor de AIC y de verosimilitud. El modelo presentado en 5.1 es el propuesto inicialmente y durante el análisis se procedió secuencialmente a encontrar aquél que resultara más parsimonioso, haciendo uso de las pruebas de razón de verosimilitud como fue descripto en la sección 4.

De manera que el modelo final de los efectos significativos se presenta junto con las estimaciones de la tabla 5.3. En dicha tabla se encuentran los valores estimados para el modelo no lineal mixto correspondiente a la ecuación 5.1. Los parámetros del modelo no lineal típicos del patrón de crecimiento para los efectos fijos son también promedio poblacionales que caracterizan el cultivar, donde $\hat{\beta}_0 = 88.37$, $\hat{\beta}_1 = 1.929$ y finalmente $\hat{\beta}_2 = 0.0233$. La importancia de los modelos no lineales radica en que los parámetros pueden ser interpretables en el contexto del problema, en este caso el parámetro β_0 caracteriza el crecimiento máximo potencial del fruto y es propio de cada cultivar, β_1 caracteriza la asíntota inferior y β_2 está relacionado a la tasa promedio de crecimiento entre ambas asíntotas.

Por otra parte, la estimación de los efectos aleatorios indica que la mayor variabilidad se debe al efecto de la parcela ya que es donde se detecta el mayor valor de desvío estándar alcanzando los $7.78\ mm$, en tanto que el árbol es el efecto de menor variabilidad porque registra un valor de $1.275763e-07$. El efecto de las temporadas es prácticamente despreciable en relación a la estimación de la variabilidad de la parcela. Si bien se esperaba que debido a la influencia del clima el efecto en los parámetros del modelo fueran mayores, no fueron significativos para el parámetro β_1 y β_2 y apenas significativo el efecto sobre la asíntota.

Dado que el objetivo final del ajuste de este modelo es estimar la variabilidad de los efectos considerados es que también se asumió efecto aleatorio al sistema de conducción y el tamaño de los frutos. Además, se debe considerar que el tamaño de los frutos, discriminados en pequeños, medianos y grandes dependían de la planta.

Al observar detenidamente las estimaciones de los parámetros en tabla 5.3 se desprende

que no todos los efectos aleatorios fueron significativos en todos los casos.

Tabla 5.3: Estimación de los parámetros del modelo no lineal mixto del cultivar *"Beurre D'Anjou"* para los efectos fijos y aleatorios (ns: no significativo) los últimos expresados en desvío estándar

PARÁMETROS		
β_0	β_1	β_2
$88,372$	$1,92866$	$0,02332$
EFECTOS ALEATORIOS		
Temporada		
$0,01819495$	ns	ns
Parcela		
$7,778209$	$0,08515845$	0.00351
Sistema de conducción		
$0,007356164$	$0,08463421$	$0,0008014243$
Árbol		
$1,275763e-07$	ns	ns
Tamaño		
$4,502842$	$0,1267108$	$0,0008752909$
Fruto		
$2,342049$	$0,0006291349$	ns
Residual		
$1,355485$		
Correlación		
$\phi =$	$0,6807813$	

Si bien el efecto aleatorio fue significativo para el parámetro b_0, no resultaron estadísticamente significativos b_1 y b_2 para los niveles temporada y árbol, tampoco fue significativa la estimación de la variabilidad para el parámetro b_2 a nivel de fruto. En cuanto a la variabilidad de los árboles, a partir de los efectos considerados en este modelo, resulta prácticamente despreciable y este resultado es concordante con los resultados hallados en otros trabajos donde el efecto árbol explica mucho menor variabilidad respecto del efecto fruto y parcela (Bramardi et al. (1998), Avanza (2010)).

La curva característica del crecimiento de los frutos, para el cultivar *"Beurre D'Anjou"* describe un patrón sigmoideo como se observa en la línea de color negra en la figura 5.14. La curva se construyó a partir de los valores estimados del modelo encontrado en la tabla 5.3 considerando las predicciones desde 30 ddplf hasta los 128 ddplf, es decir, un ciclo de duración de crecimiento histórico. El efecto de la parcela señalado en la figura 5.14 mediante el sombreado de color amarillo, es claramente de mayor amplitud para el crecimiento del fruto representando un desvío de $7,78\ mm$ resultando el efecto más sobresaliente por su variabilidad. La selección de los tamaños en el árbol también marcan un factor importante de variabilidad mostrado en la figura 5.14 por el color amarillo intenso. La variabilidad debida al fruto resulta en tercer lugar de importancia en cuanto al efecto a considerar en la modelación, apenas perceptible en su importancia denotado con el color rosado. Cabe resaltar, que tanto el efecto sistema de conducción y parcela son similares en cuanto al valor b_1 encontrado por el modelo,

expresado en desvío estandar y afectan particularmente la asíntota inferior (ver tabla 5.3).

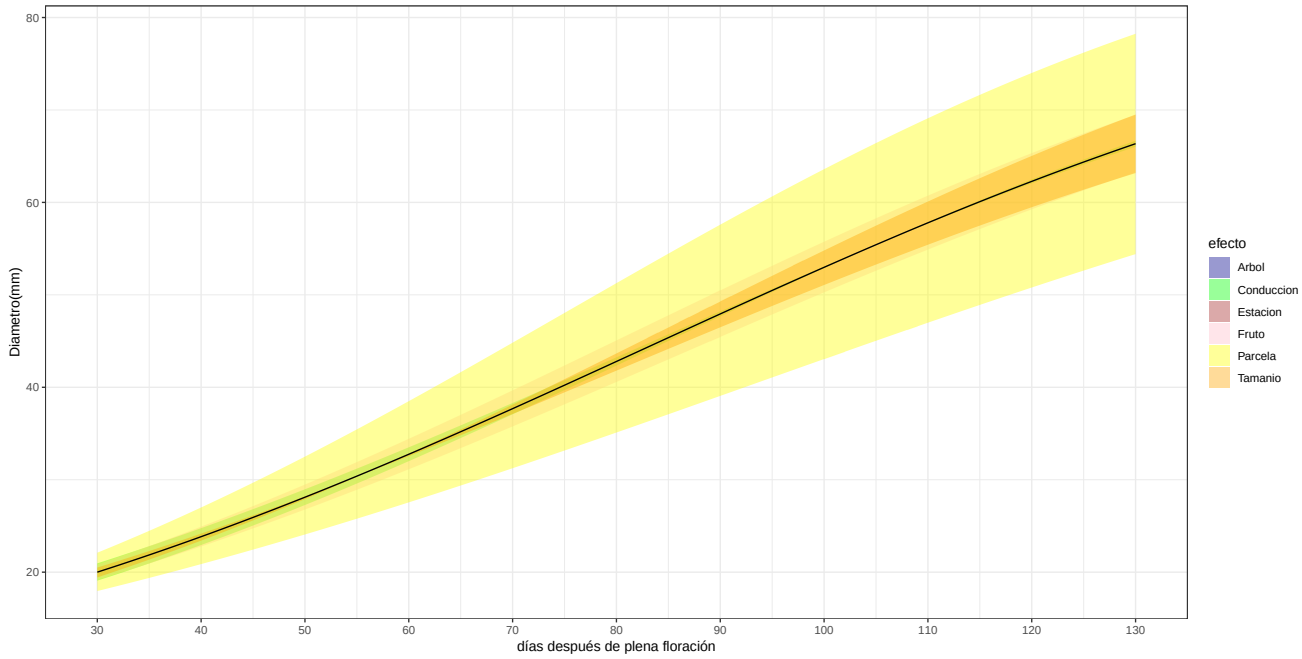

Figura 5.14: Curva característica del cv. Beurre D'Anjou y sombreados que denotan los desvíos estándar de los distintos efectos aleatorios involucrados

El efecto sistema de conducción se observa en la figura 5.14 como un sombreado de color verde que sobresale en el sector de los primeros días de crecimiento del fruto prácticamente hasta el punto de inflexión, dicho efecto es similar al efecto tamaño.

Otro aspecto importante del modelo, es que si bien se contemplaron varios niveles de efectos aleatorios, la estimación del parámetro de correlación entre los residuos de los datos fue significativo a nivel de fruto en el test de verosimilitud. El valor obtenido fue de $\phi = \hat{0.68}$, al tratarse de medidas repetidas de un mismo individuo la modelación de este aspecto debía ser contemplada. Respecto de los supuestos del modelo, no se observaron mayores desviaciones de los residuos en el ajuste marginal como tampoco en los efectos condicionales. El supuesto de normalidad, se chequeó observando los residuos en gráficos denominados "quantiles-quantiles plot" sin presentar en ningún caso problemas importantes de los mismos. Por el análisis anterior se puede asumir que el modelo se ajustó adecuadamente al patrón de los frutos, afirmación que se puede verificar en el ajuste del frutos individuales como se puede apreciar en la figura 5.15.

La figura 5.15 muestra un extracto de los ajustes por el modelo mixto estimado a nivel de fruto en 25 unidades del total de los frutos. Los frutos están representados en cada uno de los paneles y las mediciones realizadas por los círculos de color celeste. En todas las curvas presentadas se observa que la curva de color azul se ajusta y describe satisfactoriamente el comportamiento del crecimiento del fruto pudiéndose asegurar que el modelo es acorde al patrón de crecimiento de los frutos para todos los casos.

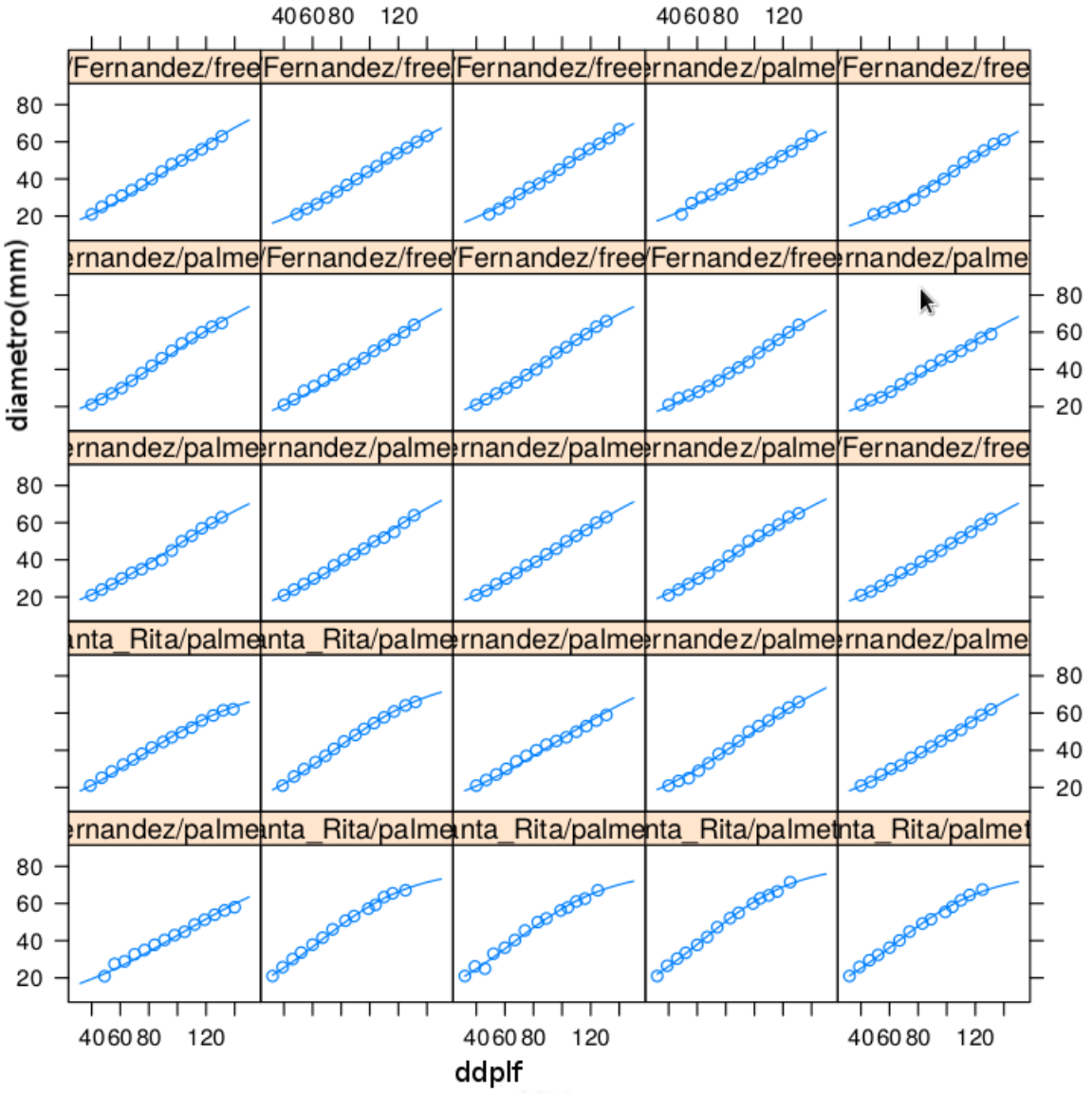

Figura 5.15: Gráfico de 25 frutos con el ajuste sujeto-específico del modelo estimado sobre el cv. *"Beurre D'Anjou"*

A partir de la ecuación 5.1 y teniendo en cuenta las matrices de varianzas y covarianzas, recordando que éstas últimas eran diagonales, se procedió mediante una distribución normal multivariada (considerando los tres parámetros) a la simulación del crecimiento de los frutos, tal cual se detalló en la sección 4 de métodos. Se realizaron 300 simulaciones para frutos pequeños, medianos y grandes, de los 30 ddplf a 135 ddplf.

Los datos fueron simulados y luego ajustados nuevamente para verificar que corresponden aproximadamente a los valores inicialmente introducidos como se verifica en la función 8.9 del Anexo. Posteriormente, se procedió a graficar los resultados que se vuelcan en la figura 5.16. Las curvas simuladas ofrecen la posibilidad no sólo de calibrar los hiperparámetros del SVM, sino de tener un conjunto de datos completo en todo el período de crecimiento del fruto reproduciendo toda la variabilidad captada por el modelo. La simulación describe un

modelo sigmoideo que posee además un comportamiento entre los datos muy similar, visto gráficamente, a los datos originales. Precisamente, la variabilidad de las curvas de crecimiento tienen la particularidad de incrementarse a medida que los valores del órgano de crecimiento aumentan, ésto se ve perfectamente en la simulación de los datos.

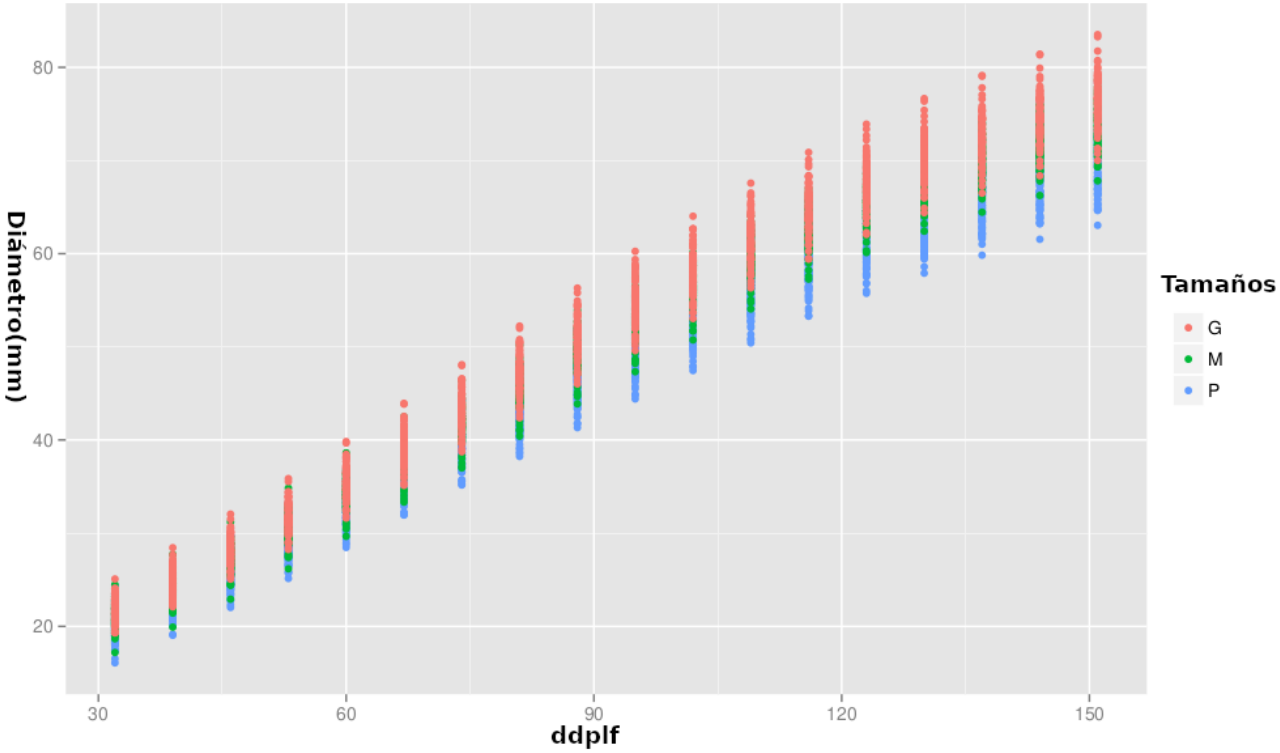

Figura 5.16: Curvas Simuladas a partir de los parámetros encontrados de las curvas del cv. *"Beurre D'Anjou"*

En la simulación se tuvo en cuenta además la correlación entre las mediciones donde se utilizó una función tipo "arima 1" con el valor encontrado en el ajuste. Cabe aclarar respecto al gráfico 5.16 que los frutos grandes en color rojo se superponen a los frutos medianos de color verde y no es que los mismos predominen.

Una vez simulados los frutos de acuerdo a las características mencionadas se procedió a buscar los hiperparámetros del algoritmo máquina de soporte vectorial, que mejor describieran el crecimiento de los frutos, esto es, probar un conjunto de valores candidatos para encontrar, en base al error de predicción, el que menor error produjese. La búsqueda de los valores óptimos para los hiperparámetros se realiza a partir de una grilla de valores combinada para los hiperparámetros costo y gamma, de manera que el entrenamiento del SVM para cada combinación obtiene una predicción y un error cuadrático medio. En esta instancia, la determinación de los mejores hipérparámetros se realizó a partir de una metodología"k-fold-cv" donde k=10, es decir, una validación cruzada donde los datos fueron divididos en diez partes de testeo y entrenamiento que obtienen un valor de error promedio. Como se mencionó en la sección 4 de métodos, la calibración se realizó en dos pasos, un primer paso con un mayor

dominio de los datos utilizando una progresión geométrica y un segundo paso tomando valores más próximos a los encontrados en la primera calibración. Se podría decir que el primer paso se logra una calibración “gruesa”, en tanto la segunda, una calibración “fina”. La representación de la figura 5.17 muestra en curvas de color azul intenso y curvas de nivel las zonas donde las combinaciones poseen valores de error cuadrático medio menor y por lo tanto indicando el rango óptimo de los hiperparámetros. Para este caso los valores óptimos se encuentran para un gamma menor a $0,5$ y un costo entre 4 y $0,5$ seleccionando un valor de costo=2, que es donde aproximadamente comienza a aumentar el error cuadrático medio de acuerdo a la tonalidad de azul. Por lo tanto, se puede asumir valores óptimos para los hiperparámetros de este conjunto de datos en $costo = 2$ y $gamma = 0,25$.

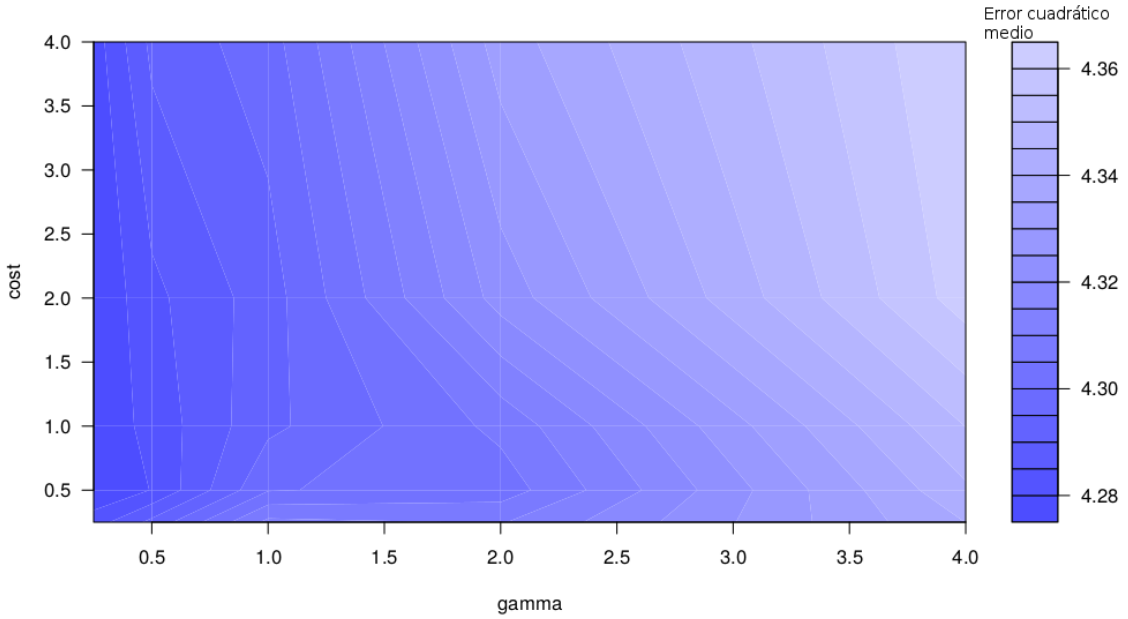

Figura 5.17: Primer paso de la calibración de los hiperparámetros Costo y Gamma del SVM

Una vez seleccionados los valores para los hiperparámetros, se procede a aplicar el algoritmo SVM sobre los datos simulados para corroborar la performance sobre las curvas de crecimiento. Para lo cual, como se describió en la sección de materiales y métodos se procede utilizando dos terceras partes del conjunto de datos para el entrenamiento de la máquina de soporte de vectores y una tercera parte, para el testeo del mismo. Una vez realizado este procedimiento se realizan las predicciones por grupo de frutos pequeños, medianos y grandes donde el resultado se observa sobre el gráfico presentado en la figura 5.18.

Se puede observar a partir del gráfico 5.18 que las curvas predichas por el método, contemplando los tamaños de frutos, describen el patrón de crecimiento desde que los frutos comienzan su desarrollon hasta el final del ciclo, contemplando además el incremento de la variabilidad de los datos, es decir, no son predicciones de curvas fijas sino que las mismas se adaptan a las características de los datos.

Respecto del cálculo del error cuadrático medio, para las curvas construidas a partir del SVM,

se obtuvo un valor de $4,086\ mm^2$, o expresado en error estándar de estimación $2,021\ mm$. No obstante, si tenemos en cuenta que la predicción del SVM no contempla el efecto individual de los frutos sino que realiza la predicción sobre los grupos de tamaños de frutos se puede pensar en recalcular el error cuadrático medio teniendo en cuenta sus valores medios. Es decir, se recalcula el error cuadrático medio considerando, para cada fecha y tamaño de fruto (Pequeño, Mediano y Grande) el valor medio y se comparó con el valor predicho por el método de SVM. Entonces, recalculando el error cuadrático medio con este criterio se obtiene $0,01325\ mm^2$ ó expresado en error estándar $0,1151\ mm$. Recordando que los valores hallados corresponden a las curvas de frutos simuladas.

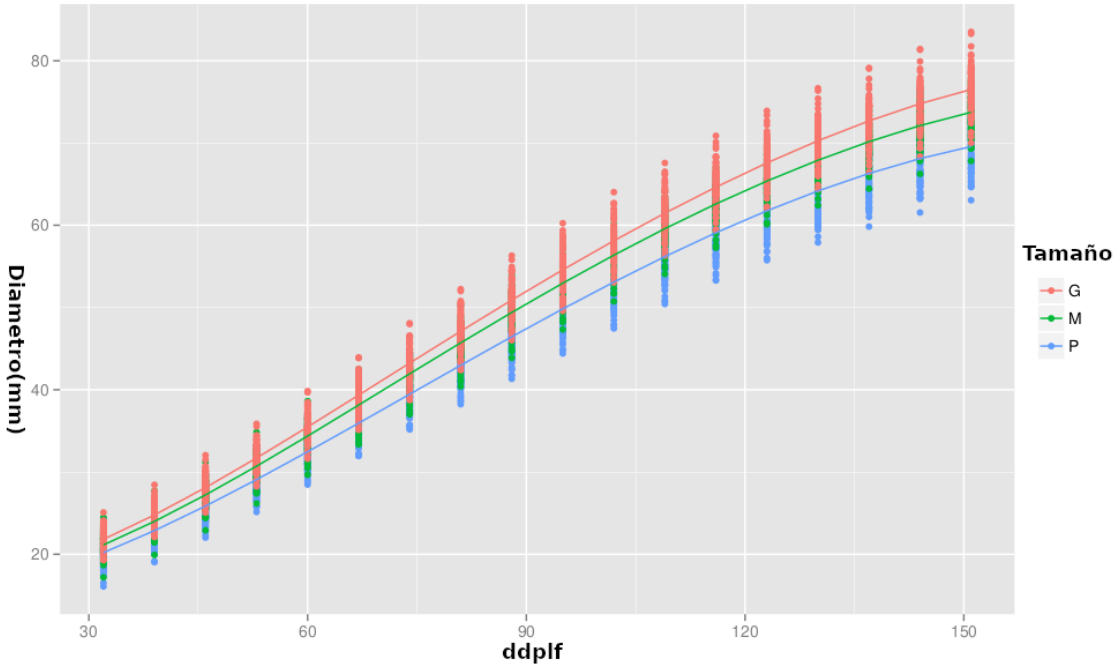

Figura 5.18: Gráfico de ajuste del SVM a las curvas simuladas por tamaño de fruto

Una vez que se obtuvo la calibración definitiva del SVM y se corroboró que los mismos son adecuados y que además ajustan un patrón con excelentes cualidades predictivas sobre los frutos, se procede a aplicar el SVM sobre las curvas de crecimiento originales de los frutos.

Para el caso de las curvas de crecimiento originales, realizado el procedimiento que se explicó más arriba, se obtuvo que el error cuadrático medio es de $7,015\ mm^2$ ó en términos del error estándar de estimación $2,648\,mm$. En tanto que, si se recalcula el error cuadrático medio por el valor medio de los frutos se obtiene $2,877\ mm^2$ expresado como error estándar $1,696\ mm$. Como se observa en el gráfico 5.19 las predicciones para las mediciones originales, siguen en cada uno de los grupos de tamaños de frutos, un comportamiento muy cercano al patrón que describen sus mediciones. Para poder comparar los resultados obtenidos con está técnica, se ajustó asimismo sobre los datos de curvas originales un modelo no lineal mixto. Se utilizó, el mismo conjunto de datos de entrenamiento que el utilizado para SVM pero en este caso para la estimación de los parámetros del modelo y de los efectos aleatorios. Luego,

sobre el mismo conjunto de datos de testeo se realizó la predicción con los parámetros del modelo estimado, técnica muy utilizada dentro del campo del aprendizaje estadístico (“Statistical Learning”).

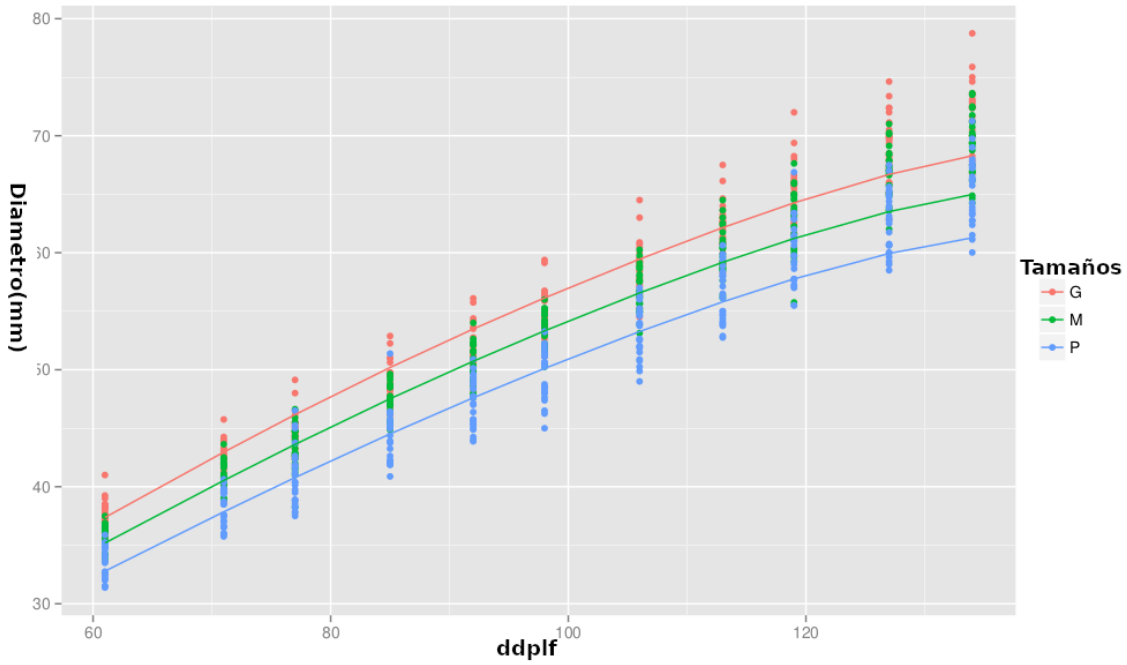

Figura 5.19: Curvas predichas a partir del SVM en el cultivar *“ Beurre D’Anjou”*

El error cuadrático medio calculado para este caso es de $8,008\ mm^2$, ó expresado como error estándar de estimación $2,823\ mm$. Teniendo en cuenta que el error cuadrático medio y el error estándar calculado mediante el SVM era de $7,015\ mm^2$ y $2,648\, mm$ los errores de estimación son evidentemente menores utilizando ésta última técnica o al menos se puede considerar que el SVM puede ser aplicado para la predicción de los patrones de crecimiento de la misma forma que los modelos estadísticos. Como se mencionó en la introducción, los modelos estadísticos poseen bondades muy interesantes respecto de la estimación de los parámetros y de la interpretación de los mismos además de la teoría estadística que lo avala pero asimismo muchas limitaciones cuando la cantidad de datos es muy grande. Por tal motivo, el SVM puede ser una alternativa eficiente y que tiene la capacidad de realizar excelentes predicciones y donde los volúmenes de datos no representan una dificultad sino que mejoran su entrenamiento y predicción.

En la presente sección, se ha avanzado en la implementación de técnicas de minería de datos especialmente en herramientas como las máquinas de soporte vectorial, observando que la técnica describe los patrones de crecimiento de los frutos considerando distintas características de los mismos. Si bien, la técnica de SVM y los modelos no lineales mixtos se originan en distintos campos el primero desde la informática y el segundo desde la modelación estadística y que ambos poseen distintas bondades no cotejables, al momento de efectuar predicciones y sometidos a igual condición, el SVM ha mostrado mejores valores predictivos,

detectado a partir de un menor valor de error cuadrático medio.

5.3.2 Clasificación multiclase de tamaños comerciales de frutos del cv Beurre D'Anjou a partir del diámetro aplicando, algoritmos de DM

Comparación entre el SVM y el modelo logístico en la predicción de los calibres

Tanto, el ajuste de modelos no lineales mixtos aplicados a las curvas de crecimiento, como el entrenamiento y aprendizaje del SVM apropiado a los patrones de los frutos ha sido un avance para este trabajo. Un punto central en los pronósticos de producción y en el presente trabajo es lograr la predicción de los frutos no sólo a partir del diámetro de los frutos, dado que luego requieren del modelo que predigan el peso para convertirlos a tamaños comerciales. Entonces, sería necesario lograr que el algoritmo tenga la capacidad de encontrar el patrón que, a partir de los diámetros al momento de pronóstico, permita predecir a cosecha directamente los tamaños comerciales correspondientes. Para ello, se debe tener en cuenta la importancia de la puesta a punto de la técnica que se utilizará posteriormente en la expansión de todos los pronósticos. Para lograrlo se compara en esta sección la aplicación de dos técnicas de clasificación que poseen distinto principios pero en condiciones similares en cuanto a la predicción: la técnica de SVM y una técnica de origen estadístico como es el ajuste de modelos logit multicategóricos proporcionales. En estas técnicas se busca una mayor precisión en la predicción de los frutos, anticipándose el mayor tiempo a la cosecha.

Los datos a utilizar en este análisis se muestran en la figura 5.20, donde se representa en eje de abscisas el diámetro de los frutos en milímetros y en ordenadas el peso en gramos que le corresponde a cosecha. En el mismo gráficos se observa el patrón entre ambas variables y el tamaño comercial correspondiente.

Asimismo está representado en formas (círculo, cuadrado, triángulos y cruz) las 4 fechas más recomendadas para realizar el pronóstico de cosecha, es decir, a los 74, 81, 88 y 95 ddplf.

Es importante aclarar que los tamaños comerciales denotados con diversos colores, son referenciados al momento histórico de cosecha o 128 ddplf. Se debe mencionar que los datos presentados corresponden a curvas simuladas, agrupadas luego por tamaños comerciales y referenciadas al momento de cosecha, para cada uno de los momentos de pronóstico.

En la figura 5.20 los momentos de pronóstico denotados por distintas formas de puntos, definieron 4 franjas de datos, donde la primera de formas circulares coincide con los 74 ddplf, los triángulos 81 ddplf, los cuadrados 88 ddplf y las cruces 95 ddplf, los cuales son los momentos más frecuentes del pronóstico de producción. En el mismo gráfico, los colores expresan los tamaños comerciales de los frutos para un diámetro dado en un momento determinado, referenciados a cosecha comercial. De manera que un fruto que posee $60\ mm$ al momento $95\ ddplf$ (marcador tipo cruz) se espera que tenga a cosecha alrededor de $220\ gramos$ y un tamaño comercial de 80, o sea un fruto de gran tamaño.

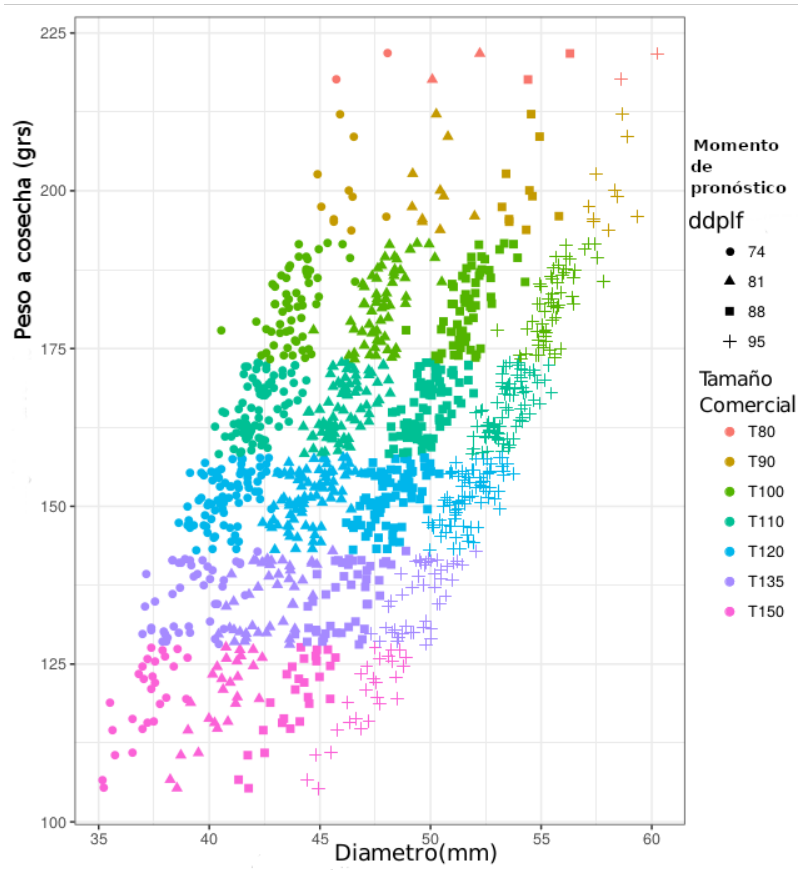

Figura 5.20: Diámetro de los frutos en distintos momentos de pronóstico y su peso correspondiente a cosecha con la clasificación comercial correspondiente en el cv. *" Beurre D'Anjou"*

Por el contrario, un fruto que posee 35 milímetros al momento de 74 ddplf (marcador tipo círculo en el gráfico) se espera tenga alrededor de 105 gramos a cosecha y un tamaño comercial de 150, es decir, resulta un fruto muy pequeño.

En el gráfico 5.20 se evidencia que la separación de los tamaños comerciales se da en forma lineal, esto se detecta viendo cómo los colores se concentran, en forma horizontal y paralela para cualquiera sea el momento de pronóstico. Al seleccionar, por ejemplo, el tamaño 150 con un color púrpura se separa linealmente del tamaño 135 de color violeta, independientemente del diámetro que tome el fruto. Los tamaños comerciales definen franjas horizontales a lo largo del diámetro en los distintos momentos considerados, existiendo franjas de datos más estrechas como el caso del tamaño 135 y otras franjas más amplias o anchas. Esto implica que no sería necesaria la implementación de un Kernel radial sino que sería suficiente la implementación de un kernel lineal con la consecuente disminución en el tiempo de calibración.

No obstante, es necesario de igual modo calibrar el hiperparámetro costo para mejorar las predicciones, testeando una grilla de valores candidatos como ya se describió en la sección 4 de métodos.

Los resultados del método de calibración se muestran en la figura 5.21 donde se observa

el comportamiento del error en función de los valores de costo, para los datos de tamaños comerciales. El error disminuye a medida que aumenta el valor del hiperparámetro costo hasta un valor de cuatro donde el error se estabiliza sin observarse mayor reducción. Por lo cual, el valor cuatro del hiperparámetro se considera el mejor candidato dado que al aumentar se podría sobreajustar el algoritmo.

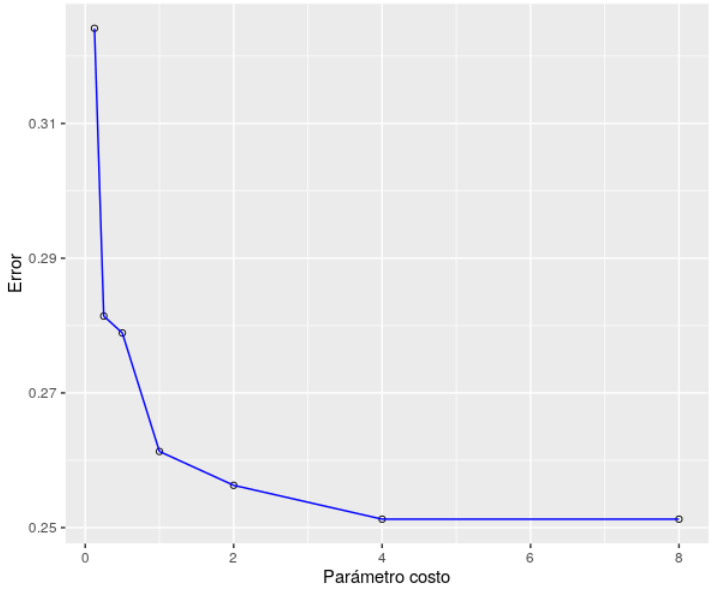

Figura 5.21: Error estándar para distintos valores del parámetro costo durante la calibración de los datos simulados de peras para clasificación multiclase

La evaluación comparativa entre el modelo de regresión ordinal y el SVM que se pretende realizar respecto de las predicciones sólo es a los fines de validación del SVM. En esta experiencia se evalúa el algoritmo de aprendizaje como clasificador y por otro lado se desea encontrar ventajas en su implementación como así también posibles problemas en las predicciones. Como se mencionó en la sección 4 de métodos de esta tesis distintos modelos multicategóricos fueron ajustados para la modelación de los tamaños comerciales. En primer lugar se ajustó un modelo multicategórico nominal con base de referencia, luego un modelo logit acumulado y por último un modelo logit multicategórico ordinal de odds proporcionales. Dado que este último cumplió el supuesto de proporcionalidad y resultó de acuerdo a los criterios clásicos de AIC y BIC el mejor modeló se optó por aplicarlo para compararlo con el SVM.

El modelo logit multicategórico definitivo resulta el siguiente:

$$\begin{aligned} Y_i &\sim M(n, \Pi_1, \ldots, \Pi_{J-1}) \\ g(\Pi_j) &= logit(\Pi_j) \\ \eta &= x'\beta_j \end{aligned} \tag{5.2}$$

Donde la variable de respuesta Y_i se define con distribución multinomial con $J-1$ categorías, en este caso las categorías corresponden a los tamaños comerciales, $g(\Pi_j)$ es la función

de enlace que para estos datos se utiliza el logit, el enlace canónico del modelo multinomial. Por último η es el predictor lineal y con él quedan expresadas las partes del modelo lineal generalizado en el contexto de los datos. Dado que el modelo utilizado fue el modelo logit de odds proporcionales(MOP) queda definido como se escribe en la ecuación 5.2 donde se considera que la pendiente β es la misma para todas las $J-1$ categorías y donde cada logit acumulado tiene su propio intercepto o umbral $\{\alpha_j\}$ (Agresti (2013)).

$$log\frac{P(Y \leq j|x)}{1-P(Y \leq j|x)} = \alpha_j + \beta' x, \text{para} \quad j = 1, \ldots, J \tag{5.3}$$

En el modelo logit, se puede considerar que las $J-1$ categorías que en este caso corresponden a los tamaños comerciales tienen una variable subyacente con una distribución particular que para los tamaños comerciales es claramente simétrica y similar a una distribución Gaussiana. Los umbrales hacen referencia a la estimación de las ordenadas de los α_j del modelo, recordando que la escala de la función de enlace corresponde al logit como lo indica la ecuación 5.2. Por último, el vector $'x_i$ hace referencia a las variables predictoras ddplf y diámetro para el momento de pronóstico.

Posteriormente, se estimaron las ordenadas y se probaron los distintos criterios de umbrales que disponía el software, en particular la librería ordinal **ordinal** de la suite R. Este aspecto resulta muy importante dado la gran cantidad de tamaños comerciales y por lo tanto de valores umbrales que debían estimarse, es decir, el logit de la ordenada para las clases 90/100, 100/110, y así sucesivamente con las restantes tamaños comerciales. Se testeó el umbral flexible que asume las ordenadas de las regresiones logísticas no poseen ningún patrón y por lo tanto, deben ser estimadas para cada uno de los odds de los tamaños comerciales (Christensen (2019)). Otros umbrales como los equidistantes restringen, entre las ordenadas su distancia para que sean iguales reduciendo la cantidad de estimaciones de los valores de α_j. Se obtuvo el umbral flexible como la mejor opción a pesar de realizar mayores estimaciones seguramente debido a que las ordenadas de las clases productivas no eran equidistantes ni respondían a restricción alguna.

En el caso del MOP, los datos de entrenamiento fueron utilizados para la estimación de los parámetros del modelo, en tanto que los datos de prueba fueron utilizados para predecir los tamaños comerciales y evaluar su eficiencia en la clasificación. La variable de respuesta utilizada tanto para el modelo ajustado como para el SVM fueron los tamaños comerciales y la variable de entrada o covariables fueron los diámetros, y el tipo de fruto (mediano, pequeño y grande) así también como la fecha de predicción en ddplf del cultivar de peras *"Beurre D'Anjou"* . En las tablas(tabla 5.4 y tabla 5.5) que se presentan podemos observar los estadísticos de predicción que permiten evaluar la performance en las clasificaciones de los tamaños comerciales realizados por ambos métodos.

Es necesario interpretar los estadísticos más importantes en la tabla 5.4: la exactitud, como la proporción total de categorías clasificadas correctamente.

Tabla 5.4: Estadísticos de predicción para los distintos tamaños comerciales en los datos de testeo para el modelo odds proporcionales

Estadísticas de clase	Tamaños comerciales predichos					
	T90	T100	T110	T120	T135	T150
Sensibilidad	$0,8$	$0,73$	$0,77$	$0,7$	$0,61$	$0,77$
Especificidad	$0,98$	$0,96$	$0,89$	$0,91$	$0,94$	$0,96$
Val.Pos.Pred	$0,53$	$0,83$	$0,72$	$0,74$	$0,67$	$0,65$
Val.Neg.Pred	$0,99$	$0,94$	$0,91$	$0,89$	$0,93$	$0,98$
Prevalencia	$0,03$	$0,19$	$0,27$	$0,27$	$0,16$	$0,08$
Tasa de detección	$0,02$	$0,15$	$0,21$	$0,19$	$0,09$	$0,06$
Detección de prevalencia	$0,04$	$0,18$	$0,29$	$0,26$	$0,14$	$0,09$
Balanceo de precisión	$0,89$	$0,84$	$0,82$	$0,80$	$0,78$	$0,87$
Estadísticos generales						
Exactitud = $0,7204$						
95% CI:$(0,67; 0,76)$						
Tasa de no información:$0,272$						
P-Valor[Prec>NIR]:$2,2e-16$]						
Kappa:$0,6441$						

En este caso el 72% de las categorías fueron clasificadas adecuadamente utilizando el método de MOP, en tanto que utilizando el método de SVM se alcanzó una precisión de 73% ligeramente superior. Analizando la sensibilidad en la tabla de resultados del MOP posee valores altos excepto al momento de clasificar el tamaño 135 donde se observa una menor performance disminuyendo hasta $0,61$. En tanto la especificidad alcanza valores altos en todas las clases comerciales predichas.

Mediante un método de remuestreo, se calcularon intervalos de confianza para los valores de precisión, donde no existieron diferencias entre ambos métodos desde el punto de vista estadístico (Ver tablas 5.4 y 5.5). La tasa de no información (NIR) corresponde a la proporción de categorías observadas más frecuentes, en el caso del método MOP el valor es del $0,272$ y mientras que el SVM es $0,292$. El software implementado realiza un test de la precisión versus NIR, si la tasa de no información es alta la precisión no cobra sentido por ello es importante dicho test, en ambos casos se rechazan ampliamente.

Otro estadístico importante es el estadístico Kappa. El mencionado estadístico es utilizado cuando dos variables binarias son medidas por dos “individuos” y se evalúa el grado de concordancia entre ambos individuos. En este caso los “individuos” corresponden a los valores observados y predichos por el método. Como se describió en la sección 2.3.2 de la introducción, los valores que toma el estadístico son siempre menores a uno, cuanto más próximo a uno más concordancia existe entre el método y los valores observados. Tanto el método de MOP como el SVM tienen valores de $0,6$, el cual indica una buena concordancia entre los valores predichos y los valores observados de acuerdo al criterio de Fleiss (Fleiss (1981)). Dentro de las clases comerciales predichas, los estadísticos más importantes son: Sensibilidad y Especificidad. Para ello, se debe tener en cuenta que el software utilizado, cuando se trata de una clasificación multiclase, mostrará una matriz de confusión con el método “one-versus-all”, es decir, el cálculo

Tabla 5.5: Estadísticos de predicción para las distintos tamaños comerciales predichos en los datos de teste aplicando el SVM

Estadísticas de clase	Tamaños comerciales predichos					
	T90	T100	T110	T120	T135	T150
Sensibilidad	$0,6$	$0,83$	$0,67$	$0,75$	$0,68$	$0,78$
Especificidad	$0,99$	$0,93$	$0,92$	$0,89$	$0,94$	$0,98$
Val.Pred.Pos	$0,82$	$0,71$	$0,77$	$0,71$	$0,67$	$0,81$
Val.Pred.Neg	$0,98$	$0,96$	$0,87$	$0,91$	$0,95$	$0,98$
Prevalencia	$0,04$	$0,18$	$0,29$	$0,26$	$0,14$	$0,09$
Tasa de detección	$0,02$	$0,15$	$0,19$	$0,19$	$0,09$	$0,07$
Detección de prevalencia	$0,03$	$0,21$	$0,25$	$0,27$	$0,15$	$0,09$
Balanceo de precisión	$0,80$	$0,88$	$0,80$	$0,82$	$0,81$	$0,88$
Estadísticos generales						
Exactitud = $0,7305$						
95% CI:$(0,684; 0.7735)$						
Tasa de no información:$0,292$						
P-Valor[Prec>NIR]:$2,2e-16$]						
Kappa:$0,6581$						

se basa en la clase de interés en relación a las restantes clases, de esta forma la sensibilidad se calcula teniendo en cuenta la primer clase en relación a todas los datos de las demas clases. Se debe remarcar que en el método de MOP, la sensibilidad posee un valor homogéneo en general en todas las clases. En tanto, en el SVM el valor de sensibilidad con las clases menos representadas son menores en particular con el tamaño comercial 90 (T90),mostrando valores de sensibilidad mayores en las clases comerciales más frecuentes. Indicando posibles dificultades en la clasificación multiclase que se encuentren desbalanceado respecto de las frecuencias de las clases. No obstante, los valores en general de precisión son aceptables para considerar la técnica como una alternativa al momento de realizar una predicción de cosecha considerando el diámetro como variable de entrada y las clases comerciales como variables de salida. Otra medida estadística que resulta importante es el balanceo de precisión que resulta en un promedio entre los valores de sensibilidad y especificidad (Kuhn (2008)). El SVM tiene asimismo un valor menor respecto del MOP para el tamaño comercial 90.

Al comparar la técnica de SVM con la aplicación de un modelo lineal generalizado se detectó mayores errores en la clasificación de los tamaños comerciales menos representadas, es decir, en los tamaños más pequeños y más grandes razón por la cual llevó a estudiar nuevamente el método para encontrar una solución. Teniendo en cuenta que el modelo lineal generalizado posee una variable subyacente que considera el desbalance entre las frecuencias de las clases. El SVM en cambio, asume que las proporciones de las categorías son iguales en todos los casos es por ello que una solución es ponderar el SVM por las frecuencias de los tamaños comerciales.

5.3.3 Ponderación en el método de SVM para mejorar las predicciones en datos multiclase desbalanceados

Teniendo en cuenta que la técnica SVM fue concebida como un clasificador binario y que luego se extendió a clasificadores multiclase, la técnica multiclase asume previamente que las categorías para los datos analizados, se encuentran balanceadas en similares proporciones (Karatzoglou et al. (2006)).

Por lo tanto, frente a datos multiclase que no poseen la misma proporción el SVM puede realizar clasificaciones erróneas dado que otorgaría la misma posibilidad a una clase que a otra al momento de realizar la predicción. Una forma de contemplar este sesgo en las predicciones multiclases es "ponderar" a cada clase por su proporción. Las ponderaciones permiten sesgar el modelo afectando el parámetro costo para cada categoría de clasificación y compensar las categorías que están menos representadas en los datos.

Las categorías pueden ser compensadas de acuerdos a distintos criterios que son expresados en funciones como por ejemplo:$1/w$, $\sqrt{(1/w)}$, $1500 \times \sqrt{(1/w)}$, donde w es el vector de proporciones de cada categoría. Es por eso que el paso siguiente fue a partir de datos del cultivar de *"Beurre D'Anjou"* , tomando la calibración anterior que provenía de los datos simulados, a partir de la misma metodología se procedió a utilizar distintas ponderaciones y se presenta aquella que obtuvo mejores resultados desde el punto de vista de las matrices de confusión y de la precisión.

Comenzando por la predicción de frutos a partir de la aplicación del SVM sin ponderar se construye la matriz de confusión de la tabla 5.6, contrastando los valores predichos por el método y los valores observados a campo.

Tabla 5.6: Matriz de confusión para los datos sin ponderar

		Datos Observados						
		T90	T100	T110	T120	T130	T140	T150
Predichos	T90	0	0	0	0	0	0	0
	T100	8	26	0	0	0	0	0
	T110	0	39	87	15	0	0	0
	T120	0	0	41	116	49	5	1
	T130	0	0	1	28	44	29	6
	T140	0	0	0	0	0	5	0
	T150	0	0	0	0	1	14	85
Estadísticos generales								
Exactitud = $0,605$								
95% CI:$(0,56; 0,64)$								
Tasa de no información:$0,26$								
P-Valor[Prec>NIR]:$2,2e-16$]								
Kappa:$0,5$								

A partir de la aplicación del SVM sin ponderar se logró una exactitud ("accuracy") de $0,60$, una tasa de no información de 0.26 y el test de no información vs precisión un $p-valor < 0,0001$. En los datos sin ponderar se observa claramente que las clases menos representadas poseen

una pobre predicción, las clases menos representadas sobretodo los tamaños de frutos grandes son de gran importancia en particular para el objetivo del pronóstico y para el cultivar que se trabaja. El tamaño 90 fue erróneamente clasificado en todos sus frutos como un tamaño inferior, lo cual podría indicar la presencia de un sesgo en el entrenamiento del método. También se destaca que los tamaños de frutos medianos que poseen una mayor frecuencia también posee una mayor proporción de categorías clasificadas correctamente. Luego de haber realizado las predicciones con distintas criterios de ponderación sobre los datos de peras cultivar D'Anjou, el criterio de ponderación que obtuvo mayor precisión fue $\sqrt{(1/W)}$ alcanzando el $0,63$ como se puede comprobar en la tabla 5.7. Aunque el aspecto más importante es que se mejoró la precisión en los tamaños pequeños y grandes a valores de $0,93$ y $0,81$ respectivamente, muy superiores a los valores sin ponderación. Al aplicar la ponderación con la función $\sqrt{(1/w)}$ sobre el SVM las predicciones en las clases menos frecuentes mejoran ostensiblemente. En el caso de los tamaños comerciales 90 existe una mayor proporción correctamente clasificado. En general, se observa una mayor proporción de clases correctamente clasificadas en prácticamente todas las clases menos representadas y unas pocas clases mal clasificadas en aquellas más representadas.

Tabla 5.7: Matriz de confusión aplicando SVM ponderado con la función $\sqrt{(1/w)}$

		Datos Observados						
		T90	T100	T110	T120	T130	T140	T150
Predichos	T90	5	8	0	0	0	0	0
	T100	3	33	1	0	0	0	0
	T110	0	24	88	16	1	0	0
	T120	0	0	39	110	40	5	1
	T130	0	0	1	31	42	20	4
	T140	0	0	0	2	10	15	6
	T150	0	0	0	0	1	13	82
Estadísticos generales								
Exactitud = $0,625$								
95% CI:$(0,58; 0,66)$								
Tasa de no información:$0,265$								
P-Valor[Prec>NIR]:$2,2e-16$]								
Kappa:$0,524$								

Como estadísticos globales se observa una mejoría en la precisión del método ponderado incrementando el valor de $0,6$ a $0,65$. No obstante, lo más destacado es que las clasificaciones menos representadas fueron mejoradas notablemente. De esta forma el algoritmo logra compensar las proporciones de las distintas categorías logrando una precisión satisfactoria dada la gran cantidad de categorías que se están logrando predecir con esta técnica.

5.3.4 Implementación del SVM en el pronóstico de producción

Uno de los logros más valorados del pronóstico de producción aplicando el método de mediciones sucesivas fue, sin lugar a dudas, la posibilidad de predecir con gran antelación a

partir del diámetro de los frutos, los tamaños comerciales a cosecha. Esto permitió brindarle al productor no sólo la posibilidad de conocer el volumen de producción mucho antes de la cosecha comercial sino además la distribución de los calibres de los frutos a ser recolectados. Esto se alcanzó en virtud al estudio de muchos años y de personal evaluando, durante las distintas temporadas, el crecimiento de los frutos a lo largo de su ciclo como ya se describió en varias oportunidades en esta tesis.

Si bien los resultados de los pronósticos realizados hasta el año 2015 han sido satisfactorios los datos han ido creciendo de manera rápida y en gran volumen lo cual requiere de variados modelos que deben ser ajustados cada vez que se realiza alguna predicción. La gran acumulación de datos y de información en general dificulta cada vez más ajustar los modelos y encontrar los parámetros para la gran cantidad de curvas registradas. Es por eso que se propone en este apartado evaluar la performance del método aplicado en el pronóstico de producción y el algoritmo estudiado SVM ya calibrado y ponderado y compararlo con el método aplicado en el pronóstico de producción.

Para llevarlo a cabo se consideraron todas las curvas de crecimiento del cultivar *"Beurre D'Anjou"* ya preprocesadas y seleccionadas y las cuales se someten como se especifica en la sección 4.3.1 de métodos, en una fase de entrenamiento o ajuste y otra de testeo. En el método de mediciones sucesivas los datos de entrenamiento son utilizados para construir las curvas de crecimiento de referencia por tamaños comerciales. Dichas curvas se construyeron, como se especifica en la sección de métodos, ajustando a cada fruto el modelo no lineal y luego del ajuste mediante una predicción al momento de cosecha comercial convirtiendo el diámetro a peso y este a tamaño comercial. Una vez identificada cada curva con su tamaño comercial se ajusta un único modelo no lineal al grupo de frutos y por ende de curvas que comparten un mismo tamaño comercial.

El resultado del ajuste de los modelos por grupos de curvas asociados a los tamaños comerciales se muestra en la figura 5.22. Es importante aclarar que el modelo utilizado en todos los casos corresponde al descripto en la ecuación 5.2, cuyas diferencias radican en los distintos parámetros estimados para cada clase comercial.

En la figura 5.22 se identifican además con tonalidades azules a los frutos grandes, las de tonalidades verdes frutos medianos, en tanto que las tonalidades anaranjadas son frutos pequeños. Claramente, las clases de tamaños más pequeños tienen curvas que se diferenciadan entre sí en casi todo el ciclo siendo más marcada hacia el final del ciclo. Los frutos de tamaños grandes, medianos y pequeños, pueden ser distinguibles entre los grupos como se puede apreciar observando las distintas tonalidades de la figura 5.22. Pero no así dentro de cada uno de los grupos donde en varios casos se pueden confundir las curvas, en especial al comienzo del ciclo de crecimiento.

Los frutos de tamaños grandes como los tamaños 80 y 90 no presentan visualmente diferencias al inicio del ciclo. Esto también puede deberse a la menor cantidad de frutos con esas características que se logran ajustar.

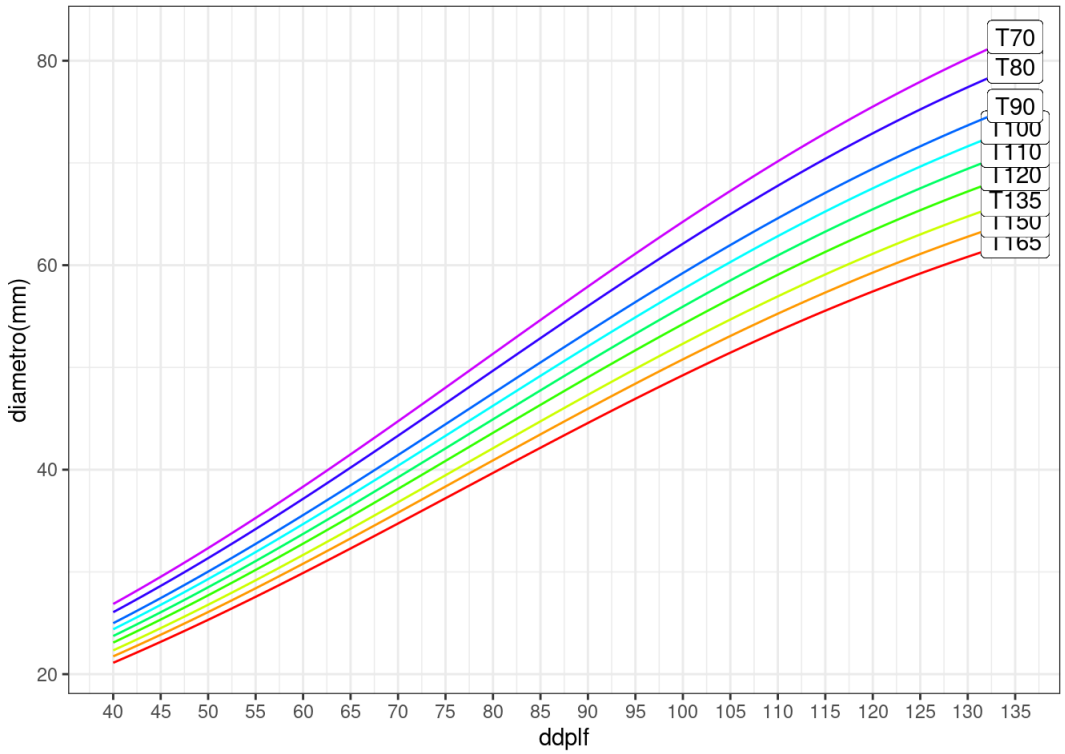

Figura 5.22: Curvas de crecimiento ajustadas a cada tamaño comercial del envase de $18,2\ kg$ en el cultivar de peras *"Beurre D'Anjou"*

Las diámetros por debajo del tamaño comercial 150 se consideran fuera de tamaño comercial (Ft_p), en el caso de los frutos de calibres grandes mayores a los tamaños 80 se clasifican como fuera de comercialización (Ft_g) aunque son muy pocos a improbables los frutos de tamaños tan grandes.

En la figura 5.22 se observa la evolución del ciclo completo, desde 40 ddplf hasta diez días posteriores al momento histórico de cosecha, para los frutos agrupados por tamaños comercial. No obstante, el momento en el que se realiza el pronóstico de producción se sitúa entre los 70 y 90 ddplf y es a partir de las curvas de crecimiento registradas el momento que se propone clasificar. La estrategia de análisis es situarse en el momento clave del pronóstico para luego realizar las predicciones de acuerdo al método de MNL y el SVM y cotejar los resultados entre uno y otro método respecto del tamaño real del fruto.

El método MNL se basa en, como se mencionó en la sección 4.3.1, a partir del diámetro de los frutos al momento de predicción seleccionar las dos curvas más próximas y calcular el valor promedio de las respectivas predicciones, si el diámetro del fruto supera la media se clasificaba con el tamaño superior caso contrario en el tamaño inferior. En el caso de que un fruto tuviese un diámetro inferior al predicho por el modelo de tamaño comercial 150, se clasifica directamente como tamaño pequeño o fuera de tamaños que en la práctica implica un fruto descartado. La performance del SVM se comparó con las predicciones de curvas de crecimiento donde se seleccionó el 70% de los registros aleatoriamente para entrenar el SVM

y ajustar los MNL, el 30% se utilizó para testear ambas técnicas. El algoritmo se entrenó con kernel lineal y el parámetro costo = 4 los cuales fueron hallados en la sección anterior.

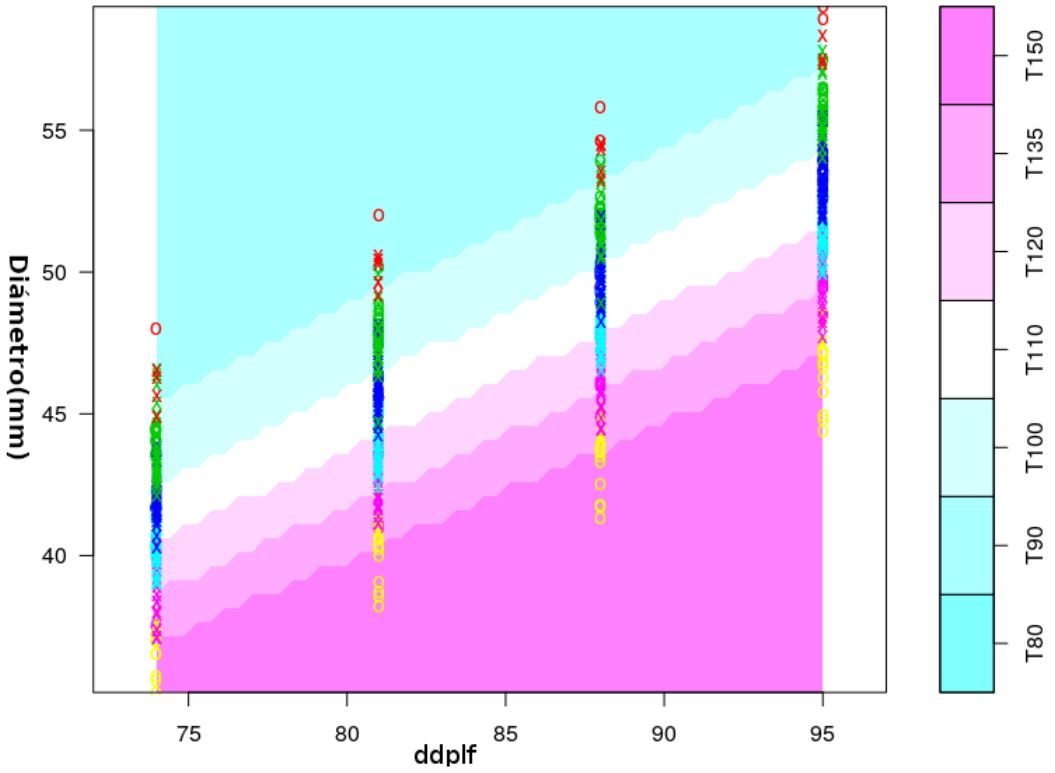

Figura 5.23: Regiones de clasificación comercial de tamaños comerciales del cv *"Beurre D'Anjou"* determinadas a partir de los vectores soporte. Los puntos son los datos observados cuyos colores denotan tamaños comerciales en cuatro momentos de pronóstico.

El testeo se realizó tomando como respuesta el tamaño comercial de cada fruto y como variables predictoras los ddplf y el diámetro del fruto entre los 30 y 40 días antes de cosecha.

De la misma manera que el método de mediciones sucesivas implementado en el pronóstico mediante los modelos no lineales define las curvas para la clasificación de los frutos, el SVM define las regiones de clasificación de acuerdo a los vectores soporte encontrados. Como se observa en la figura 5.23, las distintas áreas coloreadas representan las clasificaciones de los distintos tamaños comerciales definidas por los vectores soporte.

La región de color rosado intenso representa la zona definida para los tamaños pequeños de los frutos que coincide con los frutos de menor diámetro con una tendencia creciente a medida que transcurren los ddplf.

En la figura 5.23, los tamaños pequeños a medianos como el 120 y 110 son franjas mucho más estrechas de rosados menos intensos que corresponden con frutos de diámetros mayores y crecientes a medida que transcurren los ddplf.

Otra región de clasificación que merece mencionarse es la de color turquesa sobre la parte superior del gráfico y que representa la zona de clasificación para los frutos de mayor tamaño, específicamente del tamaño 90. No se encuentra representada la región para el tamaño 80.

Sobre las distintas regiones se pueden observar cuatro sucesiones de puntos que representan los diámetros de los frutos en cuatro momentos distintos de predicción. Los distintos tipos de puntos se asocian a los tamaños comerciales observados superpuestos a las distintas zonas de clasificación definidas por el SVM. En la figura 5.23 se destaca que los círculos amarillos se sitúan bastante acordes al tamaño 150, no obstante se observa en las clases intermedias que los puntos verdes, azules y rojos se entremezclan sin resultar bien definidos como sucede con las clases de frutos más pequeños. Esto puede deberse a que en general las clases de frutos más pequeños tienen mayor frecuencia y están mejor representadas que las clases de frutos medianas a grandes con lo cual la zona de clasificación se define con menor número de datos y de vectores soporte.

La performance se puede evaluar obteniendo la sensibilidad, especificidad y exactitud (*accuracy*) en la clasificación de los frutos de acuerdo a ambos métodos como se presentan mediante barras en la figura 5.24.

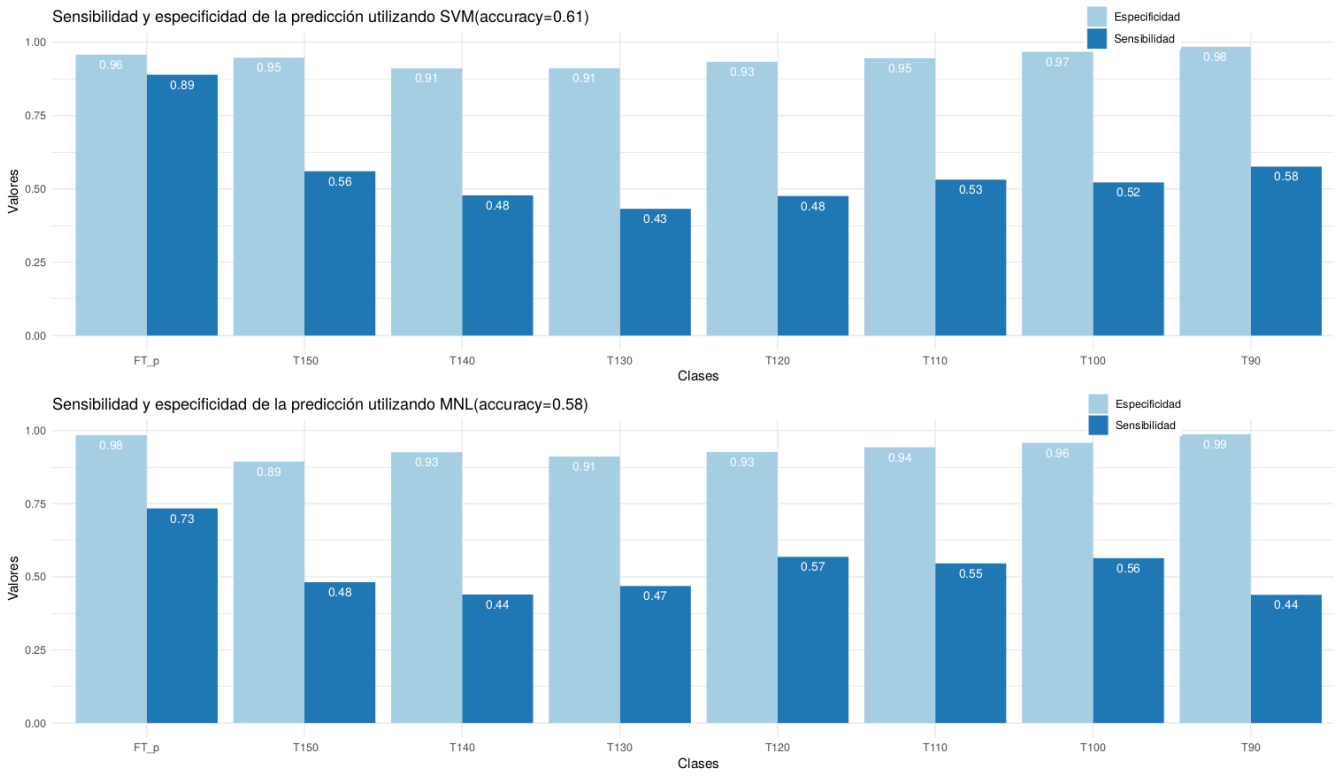

Figura 5.24: Sensibilidad y especificidad de las predicciones para el método no lineal(MNL) y el SVM

Las barras azules se destinan a los valores de sensibilidad y las barras celestes a la especificidad. Los resultados mostraron una precisión del SVM del $0,61$, levemente superior a los MNL que presentaron una precisión del $0,58$. Si analizamos la especificidad, es decir, las clases correctamente clasificadas que no pertenecen a la clase analizada, para ambos métodos es alta y en ambos casos similares, solamente se destaca el tamaño 150 donde el SVM supera en seis puntos al MNL. Mientras que la sensibilidad, es decir, aquellas clases

correctamente clasificadas a la clase analizada, en ambas técnicas es mucho menor a los valores de especificidad se destaca en promedio una mayor sensibilidad en el método de SVM. En los frutos clasificados como pequeños o fuera de tamaño la sensibilidad alcanzada por el algoritmo SVM es de $0,89$ mientras que resulta de $0,73$ para la técnica de MNL. La sensibilidad decae en ambos métodos en la clasificación de frutos medianos aunque dicha caída resulta aún mayor en el método de MNL. La especificidad promedio para el método de SVM es de $0,945$ un valor similar al valor del MNL que alcanza los $0,94$. Para el caso de la sensibilidad es de $0,56$ para el SVM y de $0,53$ para el MNL. El tamaño 90 es la clase que menor sensibilidad registra de acuerdo el método de los modelos no lineales pero aún mayor para el caso de la clasificación del SVM. Estos resultados pueden explicarse puesto que el ajuste de los modelos para el caso de los frutos grandes posee pocos frutos y no se logró obtener, dada la poca cantidad de datos, una buena estimación de los parámetros del modelo.

En cambio el SVM es un algoritmo que en general suele mantener su capacidad predictiva aún con pocos datos, sumado al hecho de haber realizado previamente el testeo de la ponderación del hiperparámetro costo en la clasificación.

Las comparaciones realizadas entre la técnica de modelos no lineales y el SVM no tiene como objetivo determinar cuál de las técnicas resulta mejor que la otra. Ambas técnicas tienen una base teórica muy distintas y aportan diferente información y dependiendo el contexto puede ser más conveniente una u la otra. En un contexto de descripción del fenómeno donde los datos a modelar sean limitados pero suficientes entonces la aplicación del método MNL resulte óptimo, mientras que, si el objetivo es predictivo pero además existe una gran cantidad de datos disponibles con la posibilidad de mayor registro entonces el SVM es en ese contexto el recomendable.

A partir de lo visto, fue factible evaluar la capacidad predictiva de ambos métodos para validar el uso de una nueva herramienta como es el SVM. En el mismo sentido las predicciones del SVM son apenas más precisas que los modelos no lineales de manera que nos permite considerar al SVM como un método alternativo.

5.3.5 Alcances del SVM en el pronóstico de producción

En este instancia, una de las preguntas que surge es acerca del alcance que posee la capacidad del SVM en la predicción de cosecha. En otras palabras, hasta que momento después de la cosecha el algoritmo continúa siendo preciso en la predicción de los tamaños comerciales. Responder dicha pregunta es importante puesto que, a menudo, la recolección de los frutos suele retardarse ya sea porque la mayoría de los frutos no alanzan un calibre requerido para la comercialización o porque la recolección se hace en al menos dos oportunidades o “pasadas”, como una estrategia de cosecha. Donde la primera pasada se realiza temprano pocos días después de la autorización oficial de cosecha, recolectando los frutos mayores de un diámetro determinado que usualmente debe superar los 66 o 67 mm y una segunda pasada una semana o más días posterior recolectando y seleccionando el resto. En otras oportunidades simplemente

un retraso de la cosecha por superposición en la recolección de diversos cultivares.

Un dato importante a tener en cuenta es que, para el cultivar de peras *"Beurre D'Anjou"* la edad promedio histórica del fruto, es decir, el tiempo desde que la floración de los frutos se encuentra en su estadío fenológico de plena floración hasta la cosecha comercial es de 128 días. Es por eso que, como se mencionó en la sección métodos se llevó a cabo una experiencia en la cual se decidió demorar la recolección hasta al menos los 141 días y evaluar el alcance del algoritmo para contemplar esta situación.

Como se describe en la sección 4.3.2 de métodos, se entrenó el algoritmo sólo con los frutos de la base datos cuyas curvas de crecimiento alcanzaron los 141 ddplf o posteriores, que fueron un total de 225 frutos y 3300 registros. En este caso se utiliza el mismo hiperparámetro y la misma función de ponderación que en la hallada en las secciones anteriores.

La predicción se realizó sobre los diámetros de los 100 frutos marcados a los 93 ddplf y se cotejó con el peso, transformado a tamaño comercial de los mismos que fueron recolectados a los 141 ddplf, es decir, 14 días posteriores al momento histórico y 20 días posteriores al momento de cosecha comercial de ésa temporada. Las predicciones del algoritmo sobre los 100 frutos marcados a campo se muestran en el gráfico de matriz de confusión de la figura 5.25.

Figura 5.25: Gráfico de matriz de confusión en la predicción de los tamaños comerciales de frutos a campo. Se representan frecuencias en valores y color representa la clasificación por el método:rojo erróneamente clasificados y verde correctamente clasificados

En dicho gráfico el eje "x" corresponde a las categorías observadas, en tanto que, el eje "y" las categorías predichas por el SVM. En el centro de las celdas se denotan las frecuencias para la combinación predichos versus observados.

En el gráfico 5.25, el color verde de las celdas representa las categorías correctamente clasificadas por el SVM, en tanto que, el color rojo las categorías erróneamente predichas por el

mismo algoritmo. Por otro lado, la intensidad del color verde representa la proporción de tamaños comerciales correctamente clasificados respecto del total de categorías observadas para ese tamaño comercial. De manera que el tamaño 150 posee 5 frutos correctamente clasificados de un total de seis frutos razón por la cual el color verde es más intenso. La intensidad del color rojo, por el contrario, representa la proporción de categorías erróneamente clasificadas del total de las observadas. Por ejemplo, el tamaño 80 posee tres frutos erróneamente clasificados de un total de cuatro, razón por la cual el color rojo es más intenso.

De la figura 5.25 se puede destacar que los mayores problemas en la clasificación se observaron en los tamaños comerciales centrales como son los 110, 120 y 130, que tendieron a ser clasificados como un tamaño superior al observado. Mientras que los tamaños, 72, 90,100 y 150 fueron los que mayor proporción de frutos clasificados correctamente poseen. Del gráfico 5.25 se deduce que la precisión en la predicción de los frutos identificados a campo resulta baja, dado la baja frecuencia de frutos clasificados adecuadamente, en particular, en los frutos de tamaños centrales.

Es por eso que como se advierte en la tabla de estadísticos de precisión 5.8, la precisión sólo alcanzó un $0,3939$ con un intervalo de confianza del 95% obtenido por bootstraping de entre $0,2972$ y $0,4972$, indicando una baja precisión del método en la clasificación de los tamaños comerciales para los 141 ddplf.

Tabla 5.8: Estadísticos para la predicción a 141 ddplf

Estadístico	Valor
Precisión	$0,3939$
Intervalo de confianza95%	$0,2972$ - $0,4972$
Tasa de no información:	$0,2626$
P-valor [$Acc > NIR$] :	$0,002943$
Estadístico Kappa :	$0,2846$

No obstante, este valor de precisión es significativamente distinto y superior a la tasa de no información. Aunque, el estadístico Kappa indica valores bajos de correlación en los valores predichos. Con todo lo descripto hasta el momento podemos asumir que las predicciones a 141 ddplf tiene una baja exactitud y bajos valores de los estadísticos de clasificación más importantes indicando, para los 141 ddplf una pobre predicción.

Por otro lado, se debe considerar que los errores de clasificación por un tamaño comercial que es la mayoría de los errores observados no representan en realidad errores graves en el pronóstico, debido a que en general se encuentran en el mismo grupo de tamaños pequeños, medianos y grandes. Es por esa razón que, una alternativa para la clasificación a 141 ddplf utilizando ésta metodología es reducir la cantidad de categorías a tres: tamaños grandes, medianos y pequeños. Donde se consideró tamaño grande a los calibres comerciales inferiores o iguales a 90, a los tamaños medianos los calibres 100, 110 y 120 y los tamaños comerciales pequeños fueron los considerados mayores e iguales a 130. De ésta manera la clasificación del algoritmo genera las predicciones que se muestran en la tabla 5.9.

Tabla 5.9: Matriz de confusión para las categorías reducidas(Pred.: categorías predichas)

		Observados		
		Grande	Mediano	Pequeño
Pred.	Grande	19	3	0
	Mediano	9	45	1
	Pequeño	0	9	13

Claramente, en éste caso predominan los frutos de tamaños medianos y luego los frutos de tamaños grandes donde los tamaños comerciales pequeños tienen una menor representación en la muestra.

Respecto de los frutos medianos se observa en la tabla 5.9 una altísima proporción de frutos correctamente clasificados donde 45 frutos de un total de 57 se clasificaron como medianos, es decir, una proporción de $0,79$. En cuanto a los frutos grandes, 19 de 28 frutos fueron predichos correctamente o una proporción de $0,68$ y finalmente los frutos pequeños 13 de 14 frutos se predijeron correctamente, lo que representa una proporción de $0,93$.

Con éste nuevo criterio de reagrupación de categorías la exactitud del algoritmo en la clasificación se incrementó a $0,778$, con un intervalo de confianza bootstrap de $0,68$ a $0,86$. Si tenemos en cuenta a partir de la tabla 5.8 el valor de exactitud era de tan solo $0,3939$ el incremento fue de $1,97$ veces. Bajo la nueva reagrupación la prueba de precisión versus la tasa de no información se rechaza con un p-valor menor y los restantes estadísticos como el estadístico kappa se incrementaron indicando una relación entre los valores predichos y los correctamente predichos.

Tabla 5.10: Estadísticos para la predicción a 141 ddplf

Estadístico	Valor
Precisión	$0,7778$
Intervalo de confianza95%	$0,6831 - 0,8552$
Tasa de no información:	$0,5758$
P-valor [$Acc > NIR$] :	$2,015e-05$
Estadístico Kappa :	$0,6207$

Claramente, la reagrupación de las categorías mejoró considerablemente las clasificaciones respecto de la predicción con diez clases comerciales, indicando qué esta metodología es factible de utilizar para el caso de predicciones de lapsos mayores a los referidos a cosecha comercial. El aumento en la precisión del algoritmo se logra no sin perder valiosa información desde el punto de vista comercial. Aunque, continúa siendo un importante aporte al sector productivo más preocupado por que sus frutos vean alcanzado el tamaño deseado y en general en concentrar la producción en los tamaños comerciales medios.

Hasta ahora se ha logrado demostrar que el SVM es una herramienta adaptable a las necesidades del pronóstico de producción en particular en la predicción de tamaños comerciales. El gran volumen de datos que genera el pronóstico de producción y las curvas de crecimiento requieren la utilización de herramientas que no dependan de las limitaciones computacionales

o de algoritmos que poseen los modelos no lineales o no lineales mixtos. El SVM es un gran aporte que permite además realimentarse de los nuevos datos que van surgiendo en la construcción de curvas de crecimiento y del pronóstico de producción.

Si retomamos la ecuación 2.1 aplicada en el pronóstico de producción del Alto Valle de Río Negro y Neuquén para la estimación de los volúmenes totales a cosecha, podemos observar que la producción total por variedad depende del número medio de frutos por árbol, el número de plantas y del peso estimado medio del fruto a cosecha. Los esfuerzos de éstas últimas secciones se han puesto en mejorar las técnicas para la estimación del peso medio del fruto, es decir, en obtener mejores estimaciones de los frutos a cosecha.

Capítulo 6

Discusión

Como se ha señalado desde la introducción, la creciente acumulación de datos dada por la digitalización de documentos, la aparición de la telemetría, la internet, acompañada del desarrollo de la informática hacen imposible el procesamiento y análisis utilizando los métodos conocidos. Mientras que la disponibilidad de datos se incrementa exponencialmente, la capacidad humana para su procesamiento se mantienen constantes (Efron (2018)).

En el dominio del pronóstico de producción, se pueden mencionar distintos ejemplos como las distintas fuentes de datos que lo conforman: los registros climáticos a través de centrales meteorológicas automáticas con sensores de gran precisión y mayor registros en tiempo, los registros de fenología agrícola de los frutos, que proveen una fuente valiosa de información, etc. Otros ejemplos relacionados son las bases de datos de instituciones oficiales que fiscalizan las parcelas y almacenan la información de la estructura productiva de todas las explotaciones comerciales de las regiones frutícola de la norpatagonia, los datos de los censos productivos relevados por los estados provinciales y nacionales, son todos ejemplos de grandes masas de datos agronómicas.

Esta enorme masa de datos requirió la implementación urgente de uno o varios métodos para su procesamiento, con el objetivo de aprovechar y mejorar los pronósticos de producción. Si bien existe una variada gama de métodos actuales para aprovechar este conjunto de datos, como el uso de ciencia de datos, "*Big Data*", es el proceso KDD con algoritmos de data mining, la metodología que más se adapta a las necesidades de las bases de datos comprometidas en este caso. En primer lugar, involucra una serie de pasos desde la confección de la base de datos hasta la obtención de patrones y que una vez interpretados generan un nuevo conocimiento. Otra metodologías factible de aplicar que se encuentra vigente en la actualidad, aplicable al dominio de las base de datos del pronóstico, es la ciencia de datos(*data science*). En primer lugar, cabe resaltar que es una disciplina que se la designa como ciencia, cuyo principio es extraer información importante de los datos, una ciencia que depende de la estadística, la matemática y fuertemente de la informática. Es decir, a diferencia del procesos KDD no se identifican distintas fases donde aplicar técnicas definidas sino que consiste en el desarrollo de habilidades para la extracción de información del conjunto de datos (Baumer et al. (2017)). La

ciencia de datos podría aplicarse en esta tesis como una fase previa a la fase de aplicación de algoritmos de minería de datos. Entre otras ventajas el proceso KDD contempla la necesidad de explorar y procesar los datos previamente de aplicar algoritmo alguno.

Dado que transcurrimos una era denominada "*Big Data*" (Efron (2018)) merece detenerse brevemente para sentar las diferencias respecto del proceso KDD aplicado en el presente trabajo de tesis. La metodología de *"Big Data"* responde a la necesidad de analizar enormes volúmenes de datos que superan las capacidades de procesar las bases de datos convencionales. Otra característica del Big Data es dar respuesta a la gran velocidad con que los datos son intercambiados para poder analizarlos u automatizar su análisis en un tiempo definido, como es el claro ejemplo de las redes sociales o las transacciones comerciales. Asimismo, esta metodología brinda el marco para procesar datos en distintos formatos como audio, texto e imágenes, en definitiva, el *"Big Data"* responde a las necesidades de los grandes volúmenes de datos, la velocidad de intercambio de los mismos y por lo tanto, la celeridad en su análisis y la variedad de los formatos que intervienen, por esta razón se lo define como la metodología de las tres "V": volumen, velocidad y variedad (Dumbill (2012)). Dadas las características de los datos que se disponen en el pronóstico no responden a la competencia del *"Big Data"*.

El proceso KDD abarca un amplio estudio desde la generación de los datos hasta la obtención de conocimientos, destina la primera fase del método para el estudio del dominio y de base de datos en general. En el proceso KDD se pueden identificar nueve fases interactivas e iterativas que a los fines prácticos se van a resumir en preprocesamiento, procesamiento y aplicación de algoritmos de minería de datos y finalmente de extracción de conocimiento (Maimon and Rokach (2010)). Dicho esquema fue el que se implementó en la presente tesis, donde el primer paso fue el entendimiento y comprensión del dominio de los datos para luego proponer objetivos del KDD con esos datos, el preprocesamiento en ésta tesis requirió el diseño, creación, llenado(o población) y administración de la base de datos. En la sección 3.1 de materiales, se dejó plasmado el dominio de los datos comprendidos dentro de la ciencias agronómicas específicamente fruticultura y referida al pronóstico de producción. Una vez definido el dominio surgen los interrogantes y los objetivos que se centran en mejorar el pronóstico de producción en base a los datos y la experiencia adquirida durante los años que se ha realizado el mismo.

Por otro lado, el proceso KDD se nutre de ciencias de la computación y en particular de teorías de bases de datos, es por eso la importancia de tratarse la base de datos como un objeto de estudio en sí misma (Fayyad et al. (1996)). Las bases de datos son ampliamente difundidas y utilizadas en todos los aspectos de la vida cotidiana actual, en las más variadas actividades diarias interviene algún tipo de sistema de gestión de bases de datos, por ejemplo: cuando accedemos a la banca virtual, buscamos números de claves tributarias, pagamos impuestos o reservamos pasajes de avión entre otras muchas actividades. Otros ejemplos más agronómicos se refieren a la implementación de bases de datos para la caracterización de suelos en cuanto a la descripción química y física de los perfiles que conjuntamente con las imágenes satelitales y otros datos importantes requieren un sistema gestión (Vamanan and Ramar (2011)). No

obstante, en el ámbito experimental no es tan utilizada como herramienta de almacenamiento de datos y tampoco en lo que respecta a pronósticos de producción, en el sentido de que en general no se piensa o diseña una base de datos para un sólo fin.

Por las características del pronóstico de producción en cuanto a volumen de datos, cantidad de variables e información a registrar, compatibilidad con otras bases de datos, es que se ha invertido un enorme esfuerzo para lograr obtener una base de datos con la mayor cantidad de información posible. La creación y población de una base de datos ha permitido, recuperar grandes cantidades de datos en variados formatos incluyendo formatos en papel que fueron digitalizados y sistematizados en esta base. La utilización de bases de datos relacionales no sólo permite almacenar un conjunto de registros relacionados sino que además pretende representar un aspecto de la realidad o minimundo(Elmasri (2011)) en este caso el pronóstico de producción. La creación de base de datos tuvo como meta no solamente la recuperación de la información, sino hacer la misma disponible para el acceso de otros usuarios de manera que pueda ser aprovechada también para otros análisis a futuro. Por medio de este trabajo de tesis, se creó y diseñó una base de datos con el propósito específico de almacenar datos de pronóstico de producción y de las demás fuentes de información que la alimentan por lo cual se dispone de una gran variedad y riqueza de datos y con el firme propósito que pueda seguir siendo poblada con datos propios de la fenología y especialmente de curvas de crecimiento de los frutos, fuente esencial para todo tipo de predicciones de cosecha.

A menudo la creación, el diseño de bases de datos y sistemas de bases de datos para estos fines suele ser cuestionado, por el tiempo y la complejidad que demanda su uso, no obstante la misma proporciona enormes ventajas. Algunos autores(Silberschatza et al. (2002)) establecen que la creación y diseño de bases de datos no se justifica al menos que en ella intervengan varios usuarios en especial de forma remota. No obstante, las bases de datos relacionales como la aplicada en esta tesis, ofrecen grandes ventajas respecto del manejo clásico en archivos. Se puede incluir numerosos tipos de datos que se pueden interrelacionar entre sí, por ejemplo: relacionar las variedades de peras y manzanas con los datos fenológicos de la tabla FLORACION, como se mostró en el esquema 5.5 de la sección resultados se puede representar complejas relaciones entre los datos, lo cual no sería factible almacenando datos en archivos. De la misma manera, a partir de las relaciones mencionadas este tipo de bases de datos es muy eficiente en la actualización y renovación de dichas relaciones (Elmasri (2011)).

Por otro lado, los sistemas gestores proveen de herramientas para la recuperación y generación de copias de seguridad que aseguren la información en caso de problemas en el hardware de los equipos en particular ante fallas en los discos de almacenamiento, fenómeno bastante frecuente. Un aspecto muy importante que brinda la creación de bases de datos en sistemas tipo SQL es la ejecución de restricciónes para reducir los errores en la carga de datos (Silberschatza et al. (2002)). En el mismo sentido, la restricción por integridad evita cargar datos cuyos registros no se encuentren en una tabla, de manera que el sistema rechaza los datos que no dispongan los registros de las tablas vinculados previamente. Un correcto diseño y gestión de la base de datos permite evitar la carga de registros redundantes, puesto

que no admite tuplas que posean exactamente los mismos individuos. La redundancia, es decir, el almacenamiento de exactamente los mismos datos lleva distintos problemas, duplica el esfuerzo de copiado y recuperación de los datos, desperdicia espacio en memoria de los discos de almacenamiento y fundamentalmente los datos repetidos pueden generar inconsistencias (Elmasri (2011)).

El diseño de la base y la administración de la misma debe almacenar cada dato lógico en únicamente un lugar de la base de datos y es lo que se denomina normalización de los datos. Las bases de datos proporcionan estructuras de almacenamiento y técnicas de búsqueda que permiten procesar consultas de forma más eficiente que cualquier otro sistema (Date (2001)). Una vez poblada la base de datos, los sistemas de gestión poseen herramientas altamente eficientes para recuperar rápidamente los datos de la misma. A menudo para realizar un análisis determinado no se requiere de la totalidad de los datos sino un subconjunto de los registros, para lo cual, la manera más eficiente de obtenerlos es mediante las consultas a la base de datos y los mecanismos de filtrado a la base de datos.

La implementación de las bases de datos y los sistemas de bases de datos para el uso de los pronósticos de producción se justifica, dada la gran variedad de tablas y datos que se han recabado y las prestaciones para las cuales se han planificado dichos datos. Otra ventaja de la utilización de bases de datos relacionales, es la posibilidad de guardar información por medio de los metadatos de las tablas . Los metadatos de las distintas tablas requieren los tipos de datos y restricciones además de comentarios que permiten agregar información a la base acerca de los datos almacenados que en otros formatos, no es factible (Elmasri (2011)). Siguiendo el proceso KDD, una vez poblada la base de datos, estableciendo los objetivos de su implementación se continúa con la etapa del procesamiento.

El procesamiento de los datos es una de las fases tanto del proceso KDD como de otras metodologías, que más tiempo y recursos informáticos demanda y no es la excepción en los datos constituídos en la base de la presente tesis. En general, uno de los primeros pasos es el preprocesamiento y está relacionado con la detección de datos faltantes y la identificación de errores en los mismos(Maimon and Rokach (2010)). La detección de errores y de problemas en los datos exige la utilización y aplicación de algoritmos adecuadamente programados para las tablas y el dominio de los datos que se estén trabajando, puesto que el proceso manual de limpieza de los datos es laborioso y en la mayoría de los casos imposible, demanda de excesivo tiempo y propenso a generar errores (Maletic and Marcus (2010)).

En la búsqueda de datos faltantes se procura analizar si los mismos responden a algún patrón particular o alguna distribución de probabilidades que pueda explicar el problema de la pérdida de estos datos. En este último caso, permitiría generar la imputación de los mismos si requiriese la metodología o el algoritmo a utilizar. La presencia de datos faltantes se puede atribuir a dos causas a la no existencia de información o a la pérdida aleatoria (Boehmke and Greenwell (2019)). La pérdida de información implica una causa estructural de la falta de datos y puede indicar deficiencias en la recopilación de datos o anormalidades en el entorno de observación. La falta al azar implica que es independiente del proceso de recopilación de los

datos.

Para los datos de esta tesis, los datos faltantes se pueden atribuir a los cambios en la metodología de relevamiento, por ejemplo en la tabla denominada “Cargas” los datos relevados en la implementación del pronóstico de acuerdo al modelo estocástico posee diferencias respecto del método de mediciones sucesivas, en el primer caso no se registraban diámetros de los frutos en el momento de realizar el pronóstico de producción, variables esenciales en el caso del pronóstico por mediciones sucesivas (ver figura 5.6).

La visualización es otro aspecto relevante en la fase de preprocesamiento de los datos, donde se pretende encontrar patrones de los datos mediante su representación gráfica (Kuhn (2008)). Los gráficos son una herramienta esencial en la búsqueda tanto de problemas u errores en los datos como en el estudio previo de los patrones que presentan los mismos. Por ejemplo, en el caso de los registros de curvas de crecimiento para los distintos cultivares, como se presenta en la figura 5.7 es una herramienta utilizada frecuentemente que permite ver tanto el comportamiento de los frutos para los distintos cultivares como para la detección de errores en las mediciones. Pudiendo identificar con relativa facilidad aquellos puntos que no siguen el patrón esperado y ser sujetos a verificación.

Además de la visualización en la detección de errores suelen aplicarse, asimismo, modelos o algoritmos (Maletic and Marcus (2010)) que permiten detectar datos erróneos u otros problemas en los datos. Para el caso de los datos de crecimiento de esta tesis y en el ajuste de modelos no lineales, se ha aplicado el criterio de que el número alto de iteraciones en el algoritmo Gauss-Newton es un posible indicador de problemas en la curva ajusta y por ende en alguna de las mediciones (Ritz and Streibig (2008)), para automatizar funciones que detecten problemas en las curvas de crecimiento. Este criterio fue implementado en todas las curvas de crecimiento como el caso presentado en la figura 5.8, donde se observó un número elevado de iteraciones para llegar a la convergencia del modelo y el error estándar superior a la unidad. Cuyo error fue detectado conjuntamente con otros no visualmente observables sino por medio del algoritmo programado especialmente para las curvas de crecimiento y sus características particulares (ver algoritmo 8.5 del anexo). Algunos autores han aplicado diversos algoritmos de preprocesamiento, como reglas de asociación, algoritmos basados en patrones, clusterización y algoritmos estadísticos utilizando el teorema de Chebyshev sobre el sistema de infomración de personal de oficiales de la armada de EEUU (Maletic and Marcus (2010)). Utilizando un subconjunto de 5.000 registros con el objetivo de demostrar que estos métodos se pueden utilizar con éxito para identificar valores atípicos que constituyen errores potenciales. Dentro del estudio se utilizaron reglas del teorema de Chebyshev y entre los 5.000 registros del conjunto de datos experimentales, 164 tuvieron valores atípicos detectados con este método, posteriores métodos visuales permitieron detectar errores en los datos.

Además de los datos faltantes y la detección de errores cabe mencionar la búsqueda de variables con varianza cero, la imputación de datos, la búsqueda de dependencias lineales entre variables y la transformación de la variable predictora entre otros muchos aspectos que se pueden abordar durante el análisis previo (Boehmke and Greenwell (2019)). Se pueden

enumerar distintas transformaciones que dependen de los datos para mejorar su características o su distribución. Por ejemplo, en caso que los datos tengan una marcada asimetría a izquierda en su distribución empírica suele sugerirse una transformación logarítmica para que los datos sean más simétricos y poder aplicar algoritmos estadísticos y aprovechar sus propiedades.

En el caso de los datos que se abordan desde la presente tesis, en primer lugar se procede al análisis de la relación entre variables comenzando por el estudio de la relación entre el peso y el diámetro de los frutos. Si bien mencionada relación ya había sido estudiada (Bramardi et al. (1998)) para peras de los cultivares *"Williams"* y *"Packhams Triumph"* , se extiende a los restantes cultivares de peras " *"Beurre D'Anjou"* " y de manzana *"Royal Gala"* y *"Pink Lady"* . Dado que la relación de los frutos, observado a partir de la aplicación de técnicas de visualización, era de tipo marcadamente potencial. Por ello se recurrió a transformaciones potenciales para su linealización. La linealización tiene como objetivo la transformación de las medidas de los frutos que se encuentran expresadas en milímetros a gramos al momento de cosecha pudiendo clasificar en tamaño comercial.

Es decir, se aplicó una transformación que permitió convertir una variable de entrada cuantitativa en una variable cualitativa pero que representa mayor interés a los productores, especialmente en el aspecto comercial de los frutos. En trabajos realizados en naranja se encontraron asimismo una similar relación potencial entre el peso y el diámetro de los cítricos aplicando el mismo criterio de transformación del diámetro en milímetros a peso en gramos y posteriormente en tamaños comerciales (Avanza (2010)).

La fase de preprocesamiento es una instancia clave del proceso KDD para lograr datos de calidad y posteriormente poder aplicar técnicas de Machine Learning que resulten efectivas, precisas y puedan alcanzar resultados confiables. Las técnicas de ML son el eje central del proceso y son las encargadas de brindar los algoritmos de aprendizaje que en éste caso tienen una función predictiva. Algunos autores han mencionado que el DM es una metodología cada vez más implementada en agricultura dado el gran volumen de datos que se registra y cumple un rol importante en la predicción de los rendimientos de los cultivos (Raorane and Kulkarni (2013)). Es una herramienta clave, por ejemplo, en la predicción de rendimientos de cereales a partir de gran cantidad de datos climáticos, de suelos e imágenes satelitales que se dispone en ese estudio. Otra aplicación importante de algoritmos de minería de datos en agricultura es para la clasificación de suelos a partir de grandes bases de datos. En este caso se testearon algoritmos de aprendizaje supervisado como distintas versiones de Naive Bayes, árboles de clasificación y Random Forest, donde se determinó que la mejor clasificación de suelos se generaba con el clasificador Naive Bayes y Random Forest, no obstante, en este trabajo se utilizó un número reducido de variables (Vamanan and Ramar (2011)). En Nueva Zelanda en cultivos de Kiwi los algoritmos de ML fueron probados para la toma de decisiones en el control de insectos y la aplicación de insecticidas sobre larvas laminadoras de hojas (Hill et al. (2014)). En este caso, para la toma de decisiones se registraron 80 variables y se probaron 6 algoritmos de ML incluyendo un modelo de regresión logística. Si bien no se alcanzó una precisión aceptable para definir el momento de aplicación de los insecticidas se lograron buenos

resultados en la decisión de no aplicación del insecticida. Aunque en esta experiencia escaso número de temporadas se destinaron al entrenamiento de los algoritmos, quizás con mayor tiempo de experimentación el entrenamiento de los algoritmos podría mejorar las predicciones de los momentos de intervención con agroquímicos.

A pesar de lo descripto, la implementación de DM no se encuentra todavía muy difundido en el análisis y la solución de problemas en las ciencias agrarias, si lo comparamos con la importancia que posee en otras áreas como medicina, astronomía o mercadotecnia, resulta todavía un método muy novedoso en este campo, donde se destacan algunos trabajos en la predicción de rendimiento de cultivos (Patel and Kathiriya (2017)), detección de enfermedades en los cultivos, detección de malezas, calidad de cultivos, reconocimiento de especies (Liakos et al. (2018)) y otros trabajos de determinación de calidad en naranjas (Kavitakomal (2019)).

Dentro de las técnicas de DM más utilizadas en problemas relacionados a la agricultura se citan técnicas tales como k-means, redes neuronales artificiales, bicluster, Naive Bayes, Suport Vector Machine, k-vecinos cercanos (Kavitakomal (2019)). Para los mismos autores es el algoritmo k-means el más frecuentemente aplicado, quizás porque prescinde de una calibración previa y sólo requiere de un proceso de entrenamiento. Otros autores (Ghaiwat and Arora (2014)) reconocen que el algoritmo k-vecinos cercanos es uno de los algoritmos más simples en lo que respecta a problemas de clasificación. KNN es un algoritmo simple que puede aplicarse a un conjunto de datos pequeño pero no es recomendable en grandes bases de datos, principalmente porque es un algoritmo de aprendizaje muy lento además que es muy influenciable por el ruido de los datos y fundamentalmente se deben seleccionar con mucha precaución las variables que intervienen en el análisis dado que es muy sensible a outliers y variables redundantes. Otros ejemplos que pueden mencionarse son la utilización de SVM en la predicción de rendimiento y de los distintos estadíos del arroz y la aplicación de ANN para pronóstico meteorológicos, ambientales, económicos y de cosecha (Liakos et al. (2018)). En trabajos de agricultura de precisión para el manejo de plagas la determinación del momento óptimo de aplicación de insecticidas se han probaron técnicas de aprendizaje ensamblado como Random Forest y Adaboost que resultaron las técnicas de mayor capacidad predictiva.

En cuanto al procesamiento de los datos de la presente tesis, se dispone de curvas de crecimiento cuya característica es la medición repetida de los frutos a lo largo del ciclo de crecimiento. Por eso el ajuste de curvas de crecimiento, fue realizado y tratado especialmente, ajustando el modelo logístico 2.7 utilizando técnicas de modelos no lineales, a partir de los estudios realizados en peras *"Williams"* y *"Packhams Triumph"* donde se determinó que la tercera parametrización de la familia de modelos logístico en 2.8, presentaba además de un buen ajuste, y excelente capacidad predictiva, buenas cualidades respecto de la curvatura interna y de la no linealidad del modelo (Bramardi et al. (1998)). Posterioremte, se lograron ajustar modelos logísticos a los cultivares de *"Red Delicious"* y *"Granny Smith"* obteniendo ajustes de gran capacidad predictiva y cuya bondad de ajuste fue igualmente satisfactoria(Stangaferro et al. (2001), Alvarez et al. (1996)).

Otros autores(Lakso et al. (1995)), realizando seguimientos en peso de los cultivares de

manzana *"Golden Delicious"* y *"Empire"* , bajo condiciones ideales, es decir, sin competencia de frutos y en estado nutricional óptimo encontraron un modelo expolineal. Donde el ajuste fue evaluado mediante pruebas de falta de ajuste indicando que el modelo se adecuó a los datos. Por otro lado, en otros cultivos, en frutos de naranjo dulce "Valencia Late" a partir de un estudio de modelos de distintas parametrizaciones del modelo logístico y de otras familias de modelos como Weibull, Gompertz y Richards, se arribó a la conclusión que el modelo más acorde al desarrollo de los frutos fue el modelo logístico de quinta parametrización evaluado a partir de medidas de no linealidad y varianza residual (Avanza et al. (2008)). Otros trabajos(Godoy et al. (2008)) de curvas de crecimiento realizados sobre arándanos observaron que las variedades tardías exhibían un patrón doble sigmoideo que permitió ajustar un modelo no linel mixto correspondiente al modelo de Gomertz II, sin embargo, en el caso de las variedades tempranas no se observaba un patrón doble sigmoideo bien definido.

Los modelos de crecimiento no son exclusivos de los frutos, algunos trabajos fueron realizados por ejemplo curvas de índice de sitio para rodales de "Eucalypto" ajustaron asimismo modelos no lineales mixtos tipo Chapmand Richars, a partir del estudio de la altura de los rodales a lo largo de los años, considerando efectos aleatorios en la asíntota del modelo. La valoración de los ajustes fue realizado a partir de los criterios de información AIC, BIC, etc. obteniendo mejores ajustes que los modelos lineales fijos (Carrero et al. (2008)).

En los trabajos llevados a cabo en la presente tesis se logró ajustar las curvas de crecimiento en peras *"Beurre D'Anjou"* mediante un modelo no lineal mixto contemplando los distintos efectos del crecimiento de los frutos (ver ecuación 5.1). El ajuste de modelos mixtos resulta novedoso dado que no había sido descripto el patrón de crecimiento a este cultivar y permitió estimar los efectos fijos y aleatorios de los parámetros del modelo, debido a la variabilidad del año, el manejo de las parcelas, los sistemas de conducción, la planta, la selección de los frutos por parte de los relevadores y finalmente el fruto. En el mismo se determinó que la mayor variabilidad, teniendo en cuenta el parámetro β_0, se encontraba en el factor debido a la parcela (ver tabla 5.3). Si consideramos que las parcelas correspondían a distintos productores que poseían distinto manejo de sus chacras sería coherente pensar en que dicha variabilidad podría ser la mayor de todas las estimadas. Luego, la variabilidad es importante en el tamaño de los frutos, es decir, en como se distribuyen los tamaños grandes, medianos y pequeños en la planta que depende de como son los frutos en la planta y del criterio de los relevadores para su marcación y posteriormente la variabilidad del fruto y prácticamente despreciable la variabilidad debida al árbol. La variabilidad debida al año, donde entre otros efectos podemos detectar la incidencia de las variables climáticas, no presentó en las temporadas estudiadas un efecto importante, resultando no significativo para los parámetros β_1 y β_2 y prácticamente despreciable para β_0. El análisis del efecto climático en las curvas de crecimiento, seguramente requiera del estudio de un ciclo de más años y en los distintos cultivares con la implementación y testeo de otras herramientas dado el volumen de datos que implica.

La estimación de la varianza de los distintos efectos es importante puesto que las simulaciones posteriores dependen de ello, de ahí la importancia de analizar la variabilidad de los efectos

estimados por el modelo. Algunos autores (Avanza et al. (2010)), analizando la partición de variabilidad en el diámetro de frutos de naranja, con el fin de encontrar el tamaño de muestra, aplicando un modelo anidado con los factores huerto y año fijos, en tanto que los factores de árbol y fruto aleatorios, encontraron un efecto significativo del huerto y el año y atribuyeron la mayor variabilidad debida al efecto del fruto, similar a lo hallado en esta tesis.

En el ajuste del modelo no lineal mixto de peras en esta tesis se observó una importante y significativa correlación, dadas las medidas repetidas de los frutos que logró modelarse mediante una función autoregresiva de primer orden contínua en la matriz de varianza covarianza residual. Teniendo en cuenta las medidas repetidas otras autores encontraron en las mediciones de las curvas de índices de sitio autocorrelación en los residuos alcanzando un coeficiente de 0.63, similar al encontrado en las curvas de crecimiento de peras Danjou, y modelados mediante una autoregresiva de primer orden (Carrero et al. (2008)).

Lograr un modelo que describa el patrón de crecimiento de los frutos y además una estimación de la variabilidad de los parámetros debido a los distintos efectos ya mencionados tiene una importancia fundamental. Por un lado, describir el comportamiento promedio poblacional (Davidian (2003)) del fruto para el cultivar *"Beurre D'Anjou"* y a partir de ello utilizar esta información para análisis posteriores como realizar simulaciones de curvas de crecimiento.

Definitivamente la técnica de modelos no lineales mixtos es una herramienta que puede describir adecuadamente el crecimiento de distintos órganos y organismos vivos. En este caso, describir el patrón de los frutos con excelentes propiedades estadísticas, estimando y analizando la variabilidad de los factores del pronóstico y con una excelente capacidad predictiva, no obstante, el ajuste de los modelos dada las dificultades de los algoritmos de estimación en particular de máxima verosimilitud se ve limitado por los niveles de anidamiento y por otro lado por la cantidad de datos (Pinheiro and Bates (2000)), que redunda en tiempo computacional y dificultad de convergencia de los algoritmos de la función utilizada. Posiblemente, la utilización de otros algoritmos heurísticos y de técnicas bayesianas permitirían afrontar ésta dificultad (Bürkner (2017)) aunque no se podría aplicar al caso de las curvas de crecimiento que no se realizaron con la misma metodología, es decir, la selección de frutos en distintos árboles como fue el caso de las primeras mediciones sucesivas.

En esta tesis se comparó el ajuste del modelo no lineal mixto con un algoritmo de minería de datos como las máquinas de soporte vectorial, para describir el crecimiento de los frutos encontrando una mejor descripción y predicción sobre el modelo no lineal mixto donde el error cuadrático medio para el modelo fue de $8,008$ y en tanto del SVM fue de $7,015$ (ver gráfico 5.16). Otros trabajos en que también evaluaron crecimiento pero de pollitas Lohman los algoritmos como redes neuronales por ejemplo, no mejoraron la capacidad predictiva del MNLM aunque sí respecto de los modelos no lineales fijos(Galeano-Vasco (2013)). En trabajos previos de esta tesis se comparó la capacidad predictiva de los modelos no lineales y las redes neuronales artificiales en curvas de crecimiento del cultivar *"Williams"* , donde éstas últimas no lograron alcanzar la capacidad predictiva de los modelos no lineales(Giménez (2015)), además de

destacar que la calibración y el entrenamiento es más complejo que el SVM. En las ANN se debe encontrar de manera heurística la mejor combinación de neuronas, considerar la arquitectura de las mismas, las funciones de activación y también las tasas de aprendizaje. Ésta observación se contradice con la experiencia de otros autores quienes consideraron que las ANN presentaron mayor facilidad de programación y ejecución puesto que sólo requiere de la variación heurística de las neuronas de la capa oculta, para la modelación de curvas de crecimiento en pollitas Lohman (Galeano-Vasco (2013)).

La implementación del SVM en regresión y en el ajuste de curvas de crecimiento con un patrón marcadamente no lineal obtuvo, desde el punto de vista predictivo resultados satisfactorios. No obstante, el interés se centra en la capacidad del algoritmo como clasificador multiclase a partir de los ddplf y del calibre de los frutos al momento previo a la cosecha. Dado que en el proceso de evaluación de la producción agrícola, el cumplimiento del estándar de calidad y el precio de mercado, la clasificación de la fruta es esencial (Kavitakomal (2019)). Realizar la clasificación de los productos o frutos manualmente resulta lento, engorroso y sujeto a errores.

Se ha reportado que el algoritmo SVM es uno de los más competitivos, precisos para problemas de clasificación y con uno de los mejores algoritmos de aprendizaje programados para el análisis de datos de alta dimensionalidad (Ghaiwat and Arora (2014)). Otros autores(Karatzoglou et al. (2006)) han sostenido que el SVM es uno de los métodos más eficientes para clasificación y regresión destacándose por su simplicidad. Respecto de la clasificación multiclase en casos agronómicos, existe evidencia donde demuestra la eficiencia en la determinación de calidad del fruto en mango, donde se comparan distintos algoritmos como SVM, y distintas versiones de ANN, entre otros y se encuentra que la mayor precisión se logra con la aplicación de SVM multiclase (Agilandeeswari et al. (2017)) aunque las diferencias en la precisión fueron de 97% y 96.7% para SVM y ANN respectivamente.

Un aspecto fundamental en la aplicación de SVM, previo a realizar el entrenamiento y la predicción de las categorías fue la calibración y la determinación de los valores adecuados para los hiperparámetros del algoritmo. Una forma de encontrar los valores más adecuados es mediante la simulación de datos y evitar de ésta forma el sobreajuste del método en datos de entrenamiento y evitar problemas en la predicción de nuevos valores. Se entiende por simulación de datos a la generación de números aleatorios a partir de un proceso estocástico que se describe mediante una serie de características y propiedades de un modelo o distribución (Kéry and Royle (2016)). Las simulaciones brindan una técnica poderosa para responder un amplio conjunto de problemas metodológicos y teórico y provee un marco para responder preguntas específicas de la investigación(Hallgren Kevin A (2013)) . En particuar, la simulación de datos es una técnica de gran valor puesto que puede aplicarse para comprobar un gran número de efectos, como comprender un modelo y los alcances del mismo; en algunos casos(Morgan Kain et al. (2015)) utilizan la técnica de simulación para realizar análisis de potencia de pruebas para detectar diferencias de tratamientos en modelos generalizados mixtos donde la determinación análitica de las pruebas resultaría mucho más compleja. En otras experiencias

(Moineddin Rahim et al. (2007)), se aplican simulaciones en modelos generalizados multinivel para determinar la cantidad de inviduos a muestrear en cada nivel en base a la variabilidad de los datos. Dicho esto la simulación de datos para generar un conjunto validación resulta novedoso y un nuevo aporte de este trabajo. Otros autores((Kéry and Royle (2016)) sugieren que las simulaciones son una herramienta para calibrar los parámetros de un modelo derivado, es decir, permite ajustar los parámetros para obtener conjuntos de datos significativos.

La simulación de datos, más precisamente de curvas de crecimiento, luego de ser transformadas a tamaños comerciales, fueron las utilizadas para la calibración de los hiperparámetros como se describió en la sección 5.3.1. Respecto de la predicción de tamaños comerciales a cosecha se determinó el tipo de kernel a utilizar donde se evaluó tanto el kernel lineal como el radial, gráficamente se determinó que las categorías podrían ser separables linealmente como se mostró en la figura 5.20. Desde un punto de vista analítico, a partir de validación cruzadas, no existieron diferencias entre uno y otro kernel, utilizando un criterio estadístico de parsimonia se aplicó el kernel de menos hiperparámetros, es decir, el kernel lineal. En el mismo proceso, se determinó que el valor del hiperparámetro costo que mejor performance obtenía se encontraba entre cuatro y ocho. Recordando que el parámetro que se está calibrando controla la penalidad por las clasificaciones erróneas (Kowalcxyk (2017)), donde valores altos del hiperparámetro *costo* hacen al algoritmo más rígido y valores bajos lo convierten en un algoritmo muy permisivo. Algunos autores señalan que con valores de *costo* igual 100 el margen se hace muy estrecho y rígido, mientras que valores de uno corresponden a márgenes muy amplios que admiten mayor proporción de categorías erróneamente clasificados(Yang and Song (2007)). Por ello la importancia de la calibración del hiperparámetro para determinar un valor de costo del margen acorde a los datos.

Los resultados obtenidos en esta tesis indican que el SVM como clasificador es un algoritmo apropiado para la predicción de los frutos en el pronóstico de producción; a partir de un kernel lineal se ha logrado una precisión de 0.73 que a pesar de no estar ponderado tiene mejor performance que el modelo logístico aplicado como algoritmo de aprendizaje. Los estadísticos muestran altos valores de especificidad y medios a bajos valores de sensibilidad, dichos valores se traducen en altos valores de predicciones negativas y valores menores en predicciones positivas (ver tabla 5.5). Otros trabajos han realizado una clasificación de manzanas en verdes y rojas utilizando 90 imágenes para dicha clasificación donde se utilizó un kernel lineal y arrojó una precisión cercana al 100% (Suresha et al. (2012)). No obstante, no está claro cómo se realizó el proceso de entrenamiento ni tampoco la calibración, pudiendo indicar que ha existido un sobreajuste de los datos de entrenamiento y que dichas predicciones no son extrapolables a nuevos datos. Otros trabajos realizados en tomate han utilizado el algoritmo SVM para realizar clasificación de tomates sanos y tomates dañados donde se alcanzó una precisión de 90% (Semary et al. (2014)). En dicho trabajo se utilizaron imágenes de tomates en distintas condiciones de sanidad cuyas imágenes fueron preprocesadas y segmentadas para mejorar la clasificación de las mismas, utilizando el 70% para entrenar el algoritmo y el 30% para testear. En el mismo ejemplo se observó que el kernel lineal y el kernel radial lograban la

mayor precisión. No obstante, no se aclara cuáles fueron las condiciones de calibración. En otros ensayos realizados en el contexto de agricultura de precisión se testeó la predicción de aplicaciones de insecticidas para el control de larvas minadoras en Kiwi donde el SVM alcanzó una sensibilidad del 80% y una precisión del 93%, no obstante, la implementación de ADAboost presentó superiores valores de sensibilidad de la predicción (Hill et al. (2014)).

Para el problema de clasificación de frutos en tamaños comerciales, si tenemos en cuenta el gran número de categorías (ver tabla 4.1) la clasificación resulta mucho más compleja que una clasificación binaria o de tres categorías. Es por eso que se considera que valores de precisión de 0.7 o 0.6 indican una muy buena performance del algoritmo, desde el punto de vista productivo si el error de clasificación es sólamente por una categoría comercial dentro del mismo grupo de tamaños no resultaría estrictamente en un error grave. Por ejemplo, si el tamaño real fuera 113 para una caja de 19 kilogramos pero el clasificador indicara un tamaño 100 ambos tamaños comerciales corresponden a un tamaño intermedio y comercialmente aceptable. En otros trabajos de clasificación multiclase, se ha reportado al SVM en la clasificación de enfermedades nutricionales de nitrógeno, potasio y magnesio a través de imágenes de árboles para aceite de palma. En este caso se aplicó un kernel polinomial con un polinimio de grado 3 y un valor de costo =0.1, y se logró obtener una precisión de 87%, 100% y 98% para la clasificación respectiva de deficiencias en potasio, magnesio y nitrógeno. También se encuentran resultados satisfactorios con la aplicación de SVM multiclases, en la clasificación por imágenes de frutos de mangos como frutos de muy buena calidad, buena calidad y mala calidad. En dicha experiencia se utilizó el 80% de datos para entrenar el algoritmo y el 20% restante para evaluar, logrando un 97% de precisión en la clasificación de testeo (Agilandeeswari et al. (2017)).

Como se mencionó tanto en la introducción como en los resultados de la clasificación multiclase a partir del SVM, es de vital importancia considerar las proporciones de las categorías a clasificar. Los resultados hallados en el presente trabajo que se muestran en la tabla 5.7 indican pequeñas diferencias en la precisión final a partir de los datos de curvas de crecimiento entre la clasificación de los datos con la ponderación de las categorías y la precisión alcanzada en los datos sin la ponderación del algoritmo. No obstante, si consideramos los tamaños extremos y los tamaños menos representados vemos importantes diferencias en la predicción de los mismos donde se observa una mejora con el algoritmo ponderado. Una diferencia importante entre los trabajos citados y el llevado a cabo en esta tesis en cuanto a la precisión reportada se fundamenta en que la clasificación por imágenes suelen utilizarse un grupo importante de variables de entrada, en tanto, en el caso de esta tesis sólo se disponde de la variable ddplf y el diámetro al momento de predicción. En los datos simulados, al agregar el tamaño de frutos, la precisión supera el 0.7. Algunos autores, utilizando el algoritmo de SVM para la evaluación de la calidad en naranjas determinando defectos y enfermedades a partir de un tratamiento de imágenes previo, lograron una precisión entre 60% y 90%, mejorando la precisión de otros clasificadores como Naive Bayes.

Las curvas de crecimiento, de la figura 5.22 ajustada por tamaños para un envase comercial de $18,2$ kg son además la base para la construcción de las tablas de raleo (Bramardi et al.

(2006)). Las tablas de raleo no se encontraban con anterioridad disponibles para el cultivar *"Beurre D'Anjou"* tampoco para el cultivar de manzanas *"Royal Gala"* , por tal motivo, se considera un nuevo aporte de ésta tesis. Las tablas de raleo son una herramienta para el productor que permite a partir de los tamaños comerciales de interés descartar o eliminar durante el ciclo de crecimiento, mediante una práctica denominada raleo manual, los frutos que no alcancen los diámetros definidos por dichas tablas. Como se mencionó en la introducción, cuanto mayor sea la anticipación de dicha práctica mejores serán los resultados obtenidos en cosecha. Las tablas de raleo no sólo fueron construidas en manzanas y peras para el Alto Valle, otros autores han implementado la metodología para el caso de tablas de raleo en naranjo tardío "Valencia Late" con resultados similares, es decir, como una herramienta de raleo al productor. La importancia de las tablas de raleo es que son la metodología esencial de predicción de pronóstico de cosecha (Avanza (2010)).

Las curvas encontradas para los distintos tamaños comerciales permiten realizar una clasificación multiclase y es la metodología hasta ahora vigente en el pronóstico de producción de peras y manzanas de la región del Alto Valle y Valle Medio. Los resultados de la figura 5.24 indican que la precisión en la estimación de los tamaños comerciales a cosecha son sensiblemente superiores aplicando el SVM que las curvas de crecimiento. El método utilizado en el pronóstico de producción de clasificación, mediante modelos no lineales, para las distintas categorías comerciales a partir de agrupar curvas de crecimiento de los frutos que corresponden a cada tamaño comercial posee ventajas y desventajas a tener en cuenta. En primer lugar, al ser construidos a partir de modelos no lineales podemos utilizar los parámetros para interpretar y caracterizar los tamaños comerciales a partir de las curvas. Pero entre las desventajas observadas, se destaca que requiere de un importante conjunto de datos para la construcción de las curvas por lo tanto, las curvas de crecimiento para los frutos muy grandes suelen ser escasas y el modelo que describe el tamaño comercial de frutos grandes y muy grandes es deficiente. Al respecto, el SVM requiere en general menor número de datos para calcular los vectores soportes y lograr de ésta manera mejores predicciones en frutos poco representados. Al respecto el algoritmo de SVM presenta la desventaja de no poseer parámetros que puedan describir o interpretar el crecimiento de los frutos. No obstante, en el contexto de la busqueda del mejor modelo predictivo para tener mejoras en los pronósticos de producción, dicha desventaja no se sostiene. Y por los resultados hallados el SVM resulta una metodología que mejora las predicciones además de resultar apropiada al momento del procesamiento de grandes volúmenes de datos.

Un requerimiento en el pronóstico de producción es determinar el alcance de las predicciones que realiza el método implementado. Es decir, estudiar mediante una experiencia a campo si la eficiencia en la predicción continúa siendo elevada aún después de 14 días del momento histórico de cosecha comercial. Los resultados indicaron que la precisión en la predicción de diez categorías comerciales disminuía la exactitud y en particular la sensibilidad a valores muy bajos. Mucho tiene que ver el incremento de la variabilidad a medida que transcurren los días de vida del órgano, como señalan algunos autores (Davidian (2003),Carrero et al.

(2008)), sumado al hecho de que en general todos los datos recabados fueron referenciados al momento definido como cosecha comercial y no momentos posteriores. También cabe mencionar en que la cantidad de frutos que alcanzan este momento para la construcción de las curvas de crecimiento son mucho menores y por lo tanto, el entrenamiento del algoritmo se ve acotado a ese número de curvas. Por otro lado, se puede observar en el gráfico 5.25 que a 141 ddplf cambia la distribución de los tamaños predominando los tamaños medianos y grandes. En definitiva el alcance predictivo es bajo en esta experiencia con 10 categorías pero podría recomendarse una predicción con sólo tres categorías definidas como “tamaños pequeños”, “medianos” y “grandes”, que aún puede resultar de gran utilidad para el sector productor.

Si bien en trabajo de la presente tesis se implementó SVM no se puede dejar de mencionar técnicas como los árboles de clasificación. Aunque también existe su versión para variables de respuesta continua denominado árboles de regresión. Algunos autores(Wu et al. (2009)) han testeado algoritmos en árboles de regresión para la predicción de la edad de los moluscos, involucrando importante cantidad de variables de clasificación como así también de categorías en la variable predictora. En éste ejemplo, se utilizó en la predicción de la edad de moluscos (*“Haliotis sp.”*). Obteniendo una precisión en la predicción del 35% en la edad de las ostras. El método, el de árboles de regresión y clasificación, también ha sido implementado en problemas agrononómicos. Se ha ajustado árboles de clasificación en la detección de distintas enfermedades en soja (precisamente en 15 enfermedades) a partir de 35 atributos que describen la fecha, la temperatura ambiente, humedad y características del cultivo, donde la precisión en la detección de dichas enfermedades es del 32.5%. Tambien se ha aplicado el algoritmo CART en la predicción de rendimientos de trigo, en Polonia, a partir de 19 variables entre las cuales se encuentran aspectos del suelo, manejo general del cultivo, características de la semilla y climáticas (Iwańska et al. (2018)). En éste trabajo el análisis fue realizado entre los años 1992 y 2003 en dos regiones de dicho país, donde la precisión alcanzada fue de 20% a 22% dependiendo la zona, atribuyendo su baja precisión a la variabilidad de las regiones.

Las técnicas de árboles de clasificación y regresión se destacan por tener una representación gráfica de los resultados que no poseen otros algoritmos. También es una característica de éstos algoritmos la aplicación en problemas donde existen un gran número de atributos o variables de entrada de manera que, el algoritmo permite ponderar por su importancia en la predicción, las distintas variables. Es decir, resulta en una herramienta de gran utilidad al momento de buscar las principales variables que afectan la predicción de un fenómeno, lo cual puede ser muy útil en la búsqueda de las causas en distintos ejemplos agronómicos de predicción de rendimientos. El principal inconveniente radica en que en general no posee la potencia predictiva de algoritmos como SVM o ANN.

Estamos atravesando una era denominada *“Big Data”* en que la implementación de algoritmos de minería de datos es un requirimiento para la gran mayoría de los campos de investigación. En el área agronómica, existe una demanda de los sectores productivos de frutos frescos para la implementación de sistemas con técnicas de evaluación de calidad que permitan automatizar el proceso, acelerarlo y reducir los problemas de clasificación manual que resultan imprecisos,

inconsistentes y requieren de recursos humanos y tiempo (Agilandeeswari et al. (2017)). Es por ello que las técnicas implementadas y discutidas en este trabajo pretenden hacer un aporte hacia esa dirección y dando lugar a aplicaciones y desarrollos más complejos que aporten al pronóstico de frutas y se conviertan en herramientas del sector productivo y comercial.

En esta época particular se requiere y demanda la aplicación de softwares más complejos y el desarrollo de nuevos algoritmos. En este trabajo de tesis las distintas etapas de implementación del proceso KDD requirió manejo de datos o generación de nuevos datos a partir de simulaciones. Esto resultaría imposible sin la creación de nuevos algoritmos y el consiguiente desarrollo de nuevo código. Este desarrollo intervino en distintas etapas de la tesis: para *parsear* o convertir los diferentes formatos de archivos a .csv que permita la importación desde .RData a la base definitiva. En la etapa de preprocesamiento, para la identificación de datos y curvas de crecimiento que posean errores de carga, en la visualización y detección de datos faltantes, etc. La creación de la base de datos, el manejo y consulta de los datos necesita la interacción de un sistema gestor y un lenguaje de programación. El desafío actual del análisis de datos requiere el uso de lenguajes de programación, la utilización de softwares como R y la creación de programas pueden intervenir en prácticamente todas las etapas del análisis y procesamiento desde la importación, preparación de los datos, transformación visualización, aplicación de modelos y hasta la comunicación de los resultados (Wickham (2019)). En definitiva esta nueva era es una era de computadoras, algoritmos y procesamiento de datos donde el objetivo es obtener nuevo conocimiento a partir de grandes bases de datos.

Capítulo 7

Conclusiones

- El proceso KDD fue una metodología esencial para recuperar y sistematizar una masa de datos generada a partir del pronóstico de producción de peras y manzanas del Alto Valle de Río Negro y Neuquén que se constituyó en una base de datos resultando una herramienta primordial para la consulta de los datos, disponible para su ulterior uso.

- Se implementó exitosamente un preprocesamiento de datos específico para datos del pronóstico de producción pudiendo identificar problemas de datos faltantes y su origen como así también errores de medición en curvas de crecimiento. Se adaptaron gráficos para la visualización y exploración de los datos acordes a esta etapa del proceso KDD.

- Se lograron identificar patrones para encontrar la relación entre peso y diámetro de los cultivares para todas las temporadas de pronóstico donde se relevaron datos. También se identificaron patrones en la predicción de tamaños comerciales a cosecha respecto a distintos momentos de pronóstico donde se aprovecharon las curvas de crecimiento para mejorar las predicciones.

- Se evaluaron distintas técnicas para describir curvas de crecimiento y estimar el peso medio del fruto a cosecha donde se seleccionó el SVM, el cual mediante técnicas de simulación fue calibrado para curvas de crecimiento. La implementación del SVM mejoró el ajuste de las curvas de crecimiento y el pronóstico de cosecha mediante la predicción multiclase de los tamaños comerciales superando la precisión del métodos de modelos no lineales. Se logró evaluar el alcance del SVM en el tiempo de predicción después de cosecha destacando que su efectividad depende del número de categorías a predecir con tres categorías el pronóstico es factible en términos de precisión aún 14 días posteriores a la cosecha comercial.

- Se desarrollaron nuevos algoritmos para la implementación en el proceso KDD: en la fase de preprocesamiento permitieron detectar errores en los registros de los diámetros de los frutos y en el ajuste de curvas de crecimiento. Se programó un nuevo código para la simulación de nuevas curvas de crecimiento. Fue creado un algoritmo que permitió realizar

la predicción de los tamaños comerciales a cosecha de los frutos utilizando modelos no lineales.

Basado en los trabajos realizados, el proceso KDD y las técnicas de minería de datos implementadas se considera que, fueron apropiadas para aprovechar la masa de datos del pronóstico de producción, para la extracción de conocimientos y son una fuente de información para futuros trabajos de predicción a cosecha. Entre los algoritmos de minería de datos estudiados y evaluados, el SVM es un algoritmo apropiado para la clasificación multiclase aplicados a la predicción de tamaños comerciales en el pronóstico de producción obteniendo una gran precisión a cosecha.

Capítulo 8

Anexo

8.1 Reseña histórica del Alto Valle

El origen de la región como valle irrigado se remonta al año 1910, cuando se comienza con la construcción del dique “Ballester” sobre el río Neuquén y una importante red de canales de riego desde la localidad de Contraalmirante Cordero hasta Chichinales con una extensión de más de 100 kilómetros. El crecimiento económico de la región se inició sin lugar a dudas, primero por el desarrollo de las obras de riego y luego con la llegada del ferrocarril permitiendo la valorización de las tierras y la puesta en producción de las mismas.

Históricamente, en el sistema productivo, se pueden distinguir tres etapas bien diferenciadas (CPIARN (2015)): la primera, relacionada con las empresas de capital inglés como Ferrocarril del Sud, Compañía de Tierras del Sud y Argentine Fruit Distribuitors (AFD), que abarcó desde 1911 hasta 1948 (fecha de nacionalización de los ferrocarriles y empresas conexas). La segunda, está vinculada a las empresas nacionales dedicadas a la comercialización interna y externa de frutas frescas que ocuparon el espacio vacío dejado por la desaparición de la AFD. Esta etapa comenzó en 1948 y su fin puede fijarse, tentativamente, en la década de los años 70. La tercera, aún vigente, está caracterizada por la entrada de empresas trasnacionales y el fortalecimiento de empresas locales ligadas a los mercados externos, que decidieron invertir en la producción, empaque y frío para asegurarse producción y abastecimiento propio de calidad para exportación.

Efectivamente, en el año 1913 la “Compañía Tierras del Sud” subsidiaria del “Ferrocarril Sud” establece la colonia “La Picasa” y comienza la venta y entrega de chacras de fracciones de entre 2 a 50 hectáreas. Las mismas fueron adquiridas por inmigrantes españoles e italianos y en menor medida ingleses, trabajadores muchos de ellos de las obras de irrigación y ferroviarias. Los primeros cultivos fueron netamente agrícolas como alfalfa, cereales, papas y vid para luego dar paso a la producción frutícola, en esa época en pequeñas parcelas de explotación familiar. Este ejemplo fue seguido por otros dueños de grandes extensiones iniciándose un proceso de subdivisión de tierras de oeste a este del valle, en el sentido del canal de riego Blanco (1999).

A mitad de la década de 1920 el cultivo de la alfalfa era desplazado por la producción de

vides, perales y manzanos. Posteriormente, la alfalfa era producida solamente en las regiones que se incorporaban al sistema de riego, es decir, a la parte este del valle, hacia la zona de Chichinales. Por entonces, se veía un crecimiento sostenido respecto al área cultivada con frutales de peras y manzanas, en tanto que el cultivo de vid experimentaba una franca retracción. Se destacaban el cultivo de peras de la variedad "Williams" que encontró en la región excelentes condiciones para su producción y en manzanas, la variedad "Deliciosa", que con el tiempo se convirtió en el cultivo por excelencia de la región.

Hacia fines de la década de 1920 principios de 1930 el volumen producido era considerable y la calidad apta para ser comercializable, es por eso que a instancias del mismo "Ferrocarril del Sud" se comenzó a mejorar el rudimentario acondicionamiento de la fruta para ser exportada (Bandieri (2014)). Ésto llevo a la instalación de los primeros galpones de empaque, de manera que la fruta llegaba desde las chacras a los galpones de empaque donde era clasificada por tamaño y calidad y luego enviada por vagones ventilados al puerto de Buenos Aires.

En la década de 1930 la producción de frutas estaba en pleno auge y ya el 70% de las peras y manzanas argentinas eran producidas en el Alto Valle que por ese entonces alcanzaba el primer puesto en producción nacional. El cultivo más importante era sin lugar a dudas la pera "Williams", sin embargo debido al incremento de precios internos y externos en poco tiempo la manzana, especialmente el cultivar "Red Delicious" fue incrementando su participación hasta lograr el primer puesto en producción. Como se observa en la tabla 8.1 desde el año 1944 en adelante la producción comienza a revertirse y las manzanas obtienen mayor producción que las peras; para la temporada 1944/1945 se registraron en manzanas 82.342 toneladas representando el 48% de la producción nacional.

Tiempo después comienza la industrialización de la fruta con la aparición de las primeras plantas productoras de caldo de sidra o también llamadas "sidreras" y posteriormente la aparición de fábricas deshidratadoras y productoras de pulpa de manzana. Éste hecho modifica la matriz comercial de la fruta para la región donde ya su destino se diferenciaba en exportación en fresco, mercado interno e industrialización. Los aumentos de la producción, como se observa en la tabla 8.1, el deterioro de la calidad y otros factores hicieron que fuera cada vez mayor el volumen que se destinaba a las industrias(Bandieri (2014)).

Sus singulares condiciones climáticas y particularmente la estacionalidad de la región, la colocaban en un lugar excepcional para la época y es por eso que, los mercados europeos de Alemania, Suiza, Inglaterra, Holanda, etc. fueron el destino mayoritario, recibiendo hasta el 60% de las exportaciones argentinas en esos años. El desenlace de la Segunda Guerra Mundial provocó un drástico cambio comercial, gran parte de los mercados cerraron el ingreso de fruta y otros redujeron considerablemente la demanda, sumado al hecho de que otros competidores del hemisferio sur como Chile, Sudáfrica y Nueva Zelanda hacían su aparición en el mercado de frutas. La situación obligó a que los destinos de fruta fueran redirigidos a otros mercados como Brasil y EEUU, este hecho fue el que se destacó en las décadas subsiguientes.

La finalización de la guerra y los nuevos mercados de Brasil y EE.UU. marcan el inicio de la segunda etapa, vinculada a las empresas nacionales dedicadas a la comercialización

interna y externa de frutas frescas que ocuparon el espacio vacío dejado por la desaparición de la AFD (CPIARN (2015)). La demanda, alta rentabilidad y menores exigencias de calidad del mercado brasileño, ofrecía amplias posibilidades que se tradujo en un masivo proceso de plantación de manzanas entre los años 1969 y 1972, redundando en años posteriores (como se observa claramente en la tabla 8.1) en una casi duplicación del volumen de producción. En esas décadas, entre el 50% y el 60% de la fruta valletana fue exportada al vecino país. Luego de algunos períodos en los cuales la demanda brasilera disminuyó notablemente, este mercado se volvió a convertir en uno de los principales destinos de la fruta de la región(Bandieri (2014)).

En la década de 1960, la producción del Alto Valle alcanzó y se mantuvo en las $300.000\ tn$ en manzanas y en $60.000\ tn$ para peras un volúmen considerable para la región. Al mismo tiempo, se experimentaron varios cambios que mejoraron distintos aspectos en la comercialización de la fruta. En primer lugar, en el sistema de transporte hubo un paulatina sustitución del camión por el tren lo cual permitió reducir considerablemente los costos de transporte sumado al hecho de volcar la producción por el puerto de Bahía Blanca en lugar del puerto de Buenos Aires. Por aquellas épocas, se logra integrar el empaque con los frigoríficos logrando una unidad funcional que hizo más eficiente la conservación de los frutos (Blanco (1999)).

La integración de distintos aspectos de la conservación y la comercialización,características de la tercera etapa(CPIARN (2015)), propició un proceso de concentración económica a través de la integración vertical de la actividad que permitió la consolidación y capitalización de los sectores relacionados al acondicionamiento y comercialización en detrimento del productor independiente. Con el transcurso de los años éstas empresas lograron posicionarse en la comercialización controlando no sólo las cadenas de frío sino los precios otorgados por su producto a productores independientes.

Tabla 8.1: Producción histórica de peras y manzanas de Río Negro y Neuquén en toneladas(tn.) y porcentaje ($\%$) participación en la producción nacional

Año Cosecha	Manzana		Peras		Año Cosecha	Manzana		Peras	
	tn.	$\%$	tn.	$\%$		tn.	$\%$	tn.	$\%$
1943/44	49.290	39	69.803	58	1959/60	308.400	72	77.750	69
1944/45	82.342	48	61.570	59	1960/61	278.500	65	35.800	47
1945/46	77.413	48	74.842	62	1961/62	283.500	68	77.700	68
1946/47	78.371	55	54.248	58	1962/63	360.300	76	62.700	64
1947/48	31.060	32	31.920	44	...	...	...	...	...
1948/49	102.960	60	79.050	68	1972/73	162.900	70	22.850	53
1949/50	94.900	48	46.900	51	1973/74	641.000	82	74.100	67
1950/51	180.800	64	82.300	68	1974/75	464.000	76	63.950	66
1951/52	129.150	58	62.800	59	1975/76	425.700	74	88.250	72
1952/53	154.432	58	52.570	57	1976/77	691.000	84	130.700	82
1953/54	165.372	68	36.938	59	1977/78	703.000	87	128.000	85
1954/55	223.950	62	48.000	49	1978/79	804.300	83	126.500	78
1955/56	152.600	60	49.960	52	1979/80	779.000	80	117.500	76
1956/57	293.700	70	54.850	54	1980/81	761.400	84	92.400	71
1957/58	201.250	71	65.610	59	1981/82	690.000	86	106.300	77
1958/59	327.900	71	52.200	56	1981/82	640.500	78	148.600	84

De esta forma el productor debió ceder poco a poco beneficios que fueron apropiados por el empacador y, sobre todo, cuando perdieron la capacidad de negociar la cosecha en el momento óptimo de madurez. Allí finalizaba para muchos una etapa de oro en la actividad frutícola y se iniciaba una crisis que, afectando en primer lugar al pequeño productor, se haría luego extensiva al conjunto de la actividad. La integración entre la mayor parte de la cadena de comercialización, comercialización-empaque-frío había otorgado un decisivo poder de negociación a los sectores intermediarios, éstos dependían en última instancia del productor para la entrega de la fruta en la calidad y cantidad que le permitiera cumplir los compromisos asumidos (Bandieri (2014)).

Pero a partir de los años 1970 las empresas integradas pusieron en práctica una nueva estrategia: incorporar la producción a sus funciones. La importancia que alcanzó a lo largo de la década, hizo que las empresas integradas adquirieran un poder absoluto en la fijación del precio, comenzando a demandar menor cantidad de fruta y a requerir determinada calidad por parte de los productores independientes, al tiempo que fijaban el precio en niveles inferiores. A fines de los años 70 debido a la crisis que presentaba el sector muchos productores optaron por créditos para poder sobrellevar la situación y dado que las políticas de estado fueron la desregulación de las tasas de interés generaron un endeudamiento creciente en el sector.

La crisis se hizo extensiva al conjunto de los sectores involucrados en la actividad frutícola, con la sola excepción de algunas empresas fuertemente concentradas que lograron crecer ejerciendo un monopolio, sobre las etapas de transformación y comercialización. En cambio, muchas otras empresas iban quedando en el camino hasta desaparecer. En la actualidad, todos los sujetos sociales vinculados a la fruticultura regional-con escasas excepciones- se encuentran inmersos en una crisis que lleva veinte años (Blanco (1999)).

En estos momentos, sumido en una crónica y profunda crisis económica los valles continúan con la actividad productiva de peras y manzanas, aportando la mayor proporción en la producción Argentina.

8.2 Fundamentos del SVM

Para encontrar el máximo margen es preciso encontrar el margen geométrico (Kowalcxyk (2017)).

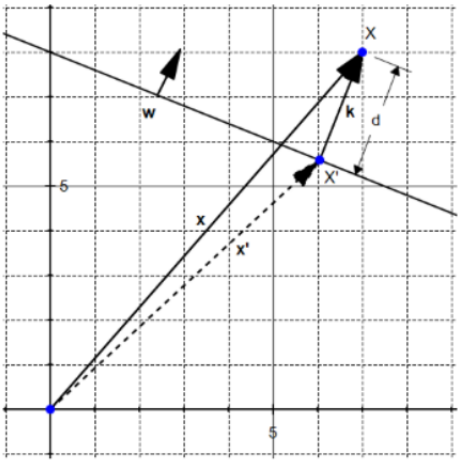

Figura 8.1: Esquema geométrico para los márgenes del SVM

En la figura 8.1 el margen geométrico del vector k resulta en calcular la distancia d. Dado que el vector k tiene el mismo sentido que W o sea $\frac{X}{\|W\|}$ entonces el vector $d(k)$ es la norma o módulo del vector y puede ser definido como:

$$k = d\frac{W}{\|W\|} \tag{8.1}$$

Se puede expresar a k en término de $x\prime$ y dado que $x\prime$ pertenece al hiperplano, entonces:

$$d = \frac{W}{\|W\|} \cdot x + \frac{b}{\|w\|} \tag{8.2}$$

8.2.1 Multiplicadores de Lagrange

El método de los multiplicadores de Lagrange es un clásico método de optimización que permite determinar extremos locales de una función sujeto a restricciones. Dada una función objetiva $f(x, y)$ a ser maximizada o minimizada sujeto a $g(x, y)$ entonces la función auxiliar de Lagrange es:

$$F(x, y, \theta) = f(x, y) - \theta g(x, y) \tag{8.3}$$

Donde θ es una variable desconocida llamada Multiplicadores de Lagrange.

La solución de la ecuación 2.56 puede ser $\mathcal{L}(W, b, \alpha) = 0$ no obstante, dicha solución es meramente analítica y sólo aplicable a un conjunto pequeño de datos. Entonces se debe reescribir el problema en términos del principio dual. El principio dual hace referencia a que el problema de optimización puede ser visto desde dos perspectivas, desde la perspectiva del

problema primordial, que en este caso es el problema de minimización y la perpectiva dual que en este caso es un problema de maximización. El mismo principio asegura que la búsqueda del máximo del problema dual sera menor o igual al mínimo del problema primordial, es decir, provee un límite inferior a la solución del problema primordial. Dado que se está tratando de resolver un problema de optimización convexa, se puede aplicar el teorema de Slater que sostiene para dualidades fuertes. Asegura que "resolver el problema dual es lo mismo que resolver el problema primordial". Entonces, la ecuación 2.56 se puede expresar:

$$\mathcal{L}(W, b, \alpha) = \frac{1}{2} W \cdot W - \sum_{i=1}^{m} \alpha_i [y_i (W \cdot x_i + b) - 1]$$

La resolución a la ecuación 2.57 se encuentra en base al teorema de Slater que establece en una dualidad fuerte el valor máximo del problema dual es igual al mínimo del problema primordial.

$$\begin{aligned} min_{w,b} \quad max_{\alpha} \quad & \mathcal{L}(W, b, \alpha) \\ s.a \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Para resolver el problema de minimización hay que obtener las derivadas parciales de $\mathcal{L}$ respecto a w y respecto a b:

$$\begin{aligned} \nabla_w c = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \\ \frac{\partial \mathcal{L}}{b} = - \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{8.4}$$

De la primera ecuación se obtiene:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{8.5}$$

Luego sustituímos el resultado de le ecuación 8.2.1 en la ecuación 2.57 y entonces se obtiene:

$$\begin{aligned} w(\alpha, b) &= \frac{1}{2} \left(\sum_{i=1}^{m} \alpha_i y_i x_i \right) \cdot \left(\sum_{j=1}^{m} \alpha_j y_j x_j \right) - \sum_{i=1}^{m} \alpha_i \left[y_i \left(\sum_{i=1}^{m} \alpha_i y_i x_i \cdot x_i + b \right) - 1 \right] \\ &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{i=1}^{m} \alpha_i y_i \left(\sum_{i=1}^{m} \alpha_i y_i x_i \cdot x_i + b \right) + \sum_{i=1}^{m} \alpha_i \\ &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j - b \sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j - b \sum_{i=1}^{m} \alpha_i y_i \end{aligned} \tag{8.6}$$

Teniendo en cuenta la ecuación 8.2.1, sabemos que el término $\sum_{i=1}^{m} \alpha_i y_i = 0$ entonces finalmente:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j \tag{8.7}$$

El problema de optimización es ahora un problema dual.

$$\begin{aligned} max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j \\ s.a. \quad & \alpha_{\geq} 0, \quad \texttt{para} \quad \texttt{cualquier} \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{8.8}$$

Hay que tener en cuenta que, la aplicación de multiplicadores de Lagrange requiere que las restricciones sean igualdades en tanto que, en éste caso se emplean restricciones que son desigualdades, por lo tanto, existe una condición que debe cumplirse y es la condición Karush-Kuhn-Tucker (KKT). La primer condición que debe cumplir es la Condición de Slater ya descripto anterioriormente para el SVM. Dado que el problema primordial que se trata de resolver es un problema de optimización convexa, las condiciones KKT son suficientes para el punto primordial y dual óptimo. Si la solución satisface las condiciones KKT se garantiza siempre una óptima solución. Para obtener W se calcula a partir del gradiente $\nabla_w \mathcal{L}$:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{8.9}$$

En tanto que para calcular b se obtiene a partir de la restricción del problema primordial, a partir de:

$$y_i(W \cdot x_i - b) - 1 \geq o \tag{8.10}$$

Teniendo en cuenta que los puntos próximo al margen son igual a 1 (ya que w está estandarizado).

$$y_i(W \cdot x_i - b) = 1 \tag{8.11}$$

Despejando:

$$b = y_{-} w \cdot x_i \tag{8.12}$$

Algunos autores aseguran que calculando el promedio de los vectores soporte es una forma numéricamente más estable que tomando todos los valores por separdo:

$$b = \frac{1}{S} \sum_{i=1}^{S} (y_i - W \cdot x_i) \tag{8.13}$$

Donde en éste caso S son el número de vectores soporte.

8.2.2 La función del kernel en el svm

Matemáticamente, dados dos vectores x_1 y x_2 en un espacio $\Re^2$ la función kernel calcula su producto escalar como si los vectores hubieran sido transformados a un $\Re^3$ pero sin realizar dicha transformación y sin calcular el producto escalar.

Entonces, el kernel es una función que devuelve el resultado de un producto interno realizado en un espacio distinto. Formalmente, dada una función $\phi : \chi \longrightarrow \upsilon$ se llama a la función el kernel $K : \chi \longrightarrow \Re$ definido por $K(X, X') = \langle \phi(X), \phi(X') \rangle$.

Lo que implica que la aplicación del Kernel se debe reemplazar el producto escalar por una función kernel.

8.2.3 Funciones creadas en R

A continuación se presentan las funciones desarrolladas en R.

La siguiente función permite simplificar el proceso de filtrado de la base de datos a partir de la selección de una variedad de interés. Genera una tabla de transición e individualiza y enumera los frutos.

```
trans <- function(variedad){ #Consignar una variedad
    transi <-variedad %>% select(pareja_de_relevadores,
    chacra, plantacion, anio, planta_n, tamanio, fruto_n,
    diametro_medio, ddplf) %>%
    group_by(pareja_de_relevadores, chacra, plantacion, anio,
    planta_n, tamanio,fruto_n) %>% summarize( n_obs= n())
transi$frutind <- 1:nrow(transi)
return(transi)
}
```

Código R 8.1: Función para generar una tabla de transición o resumen a partir de la base de datos

Función para chequear que los diámetros de los frutos son crecientes, elimina del registro cuyo dato es decreciente.

```
#óFuncin para sacar las mediciones que se decrementan
check_frut_crec <- function(datanl){
  torden <- datanl[order(datanl$ddplf),]
  p <- 1
  mint <- -1
  while (mint < 0) {
    difer <- diff(torden$diametro_medio)
    mint <- difer[which.min(difer)]
    torden$difer <- c(0,difer)
    tfals <-  torden$difer < 0
    torden <- torden[!tfals,]
    p <- p + 1
  }
  return(torden)
}
```

Código R 8.2: Función para verificar que los datos son crecientes

Función para preprocesar datos, verifica que los datos de curvas de crecimiento sean crecientes y descarta las curvas que poseen menos de cuatro registros. Guarda las curvas de crecimiento ya verificadas en una estructura de datos tipo lista.

```r
preproces <- function(vari){
  antrans <- trans(vari) #Crea tabla de ótransicin
  dft_an <- as.data.frame(antrans)
  menos3a <- dft_an$n_obs > 3
  dft_an <- dft_an[menos3a, ]
   # Inicializa  objeto para llenarlo luego con el los datos finales
  dfan = NULL
  dif_obss <- c()
  ide_frut <- c()
  #El algoritmo itera por fruto sobre la tabla de ótransicin
  for (i in dft_an$frutind ) {
    frut     <- dft_an[dft_an$frutind == i,]
    datanl <- selbasfrutb(vari,frut)#datos originales
    datana <- nrow(datanl)
    datanl <- check_frut_crec(datanl)
    datanb <- nrow(datanl)
    dif_obs <- datana-datanb
    dfan <- rbind(dfan,
                  data.frame(pareja_de_relevadores = datanl$pareja_de_
                     relevadores,
                  chacra=datanl$chacra, plantacion=datanl$plantacion,
                     anio=datanl$anio,
                  planta_n=datanl$planta_n, tamanio=datanl$tamanio,
                     fruto_n=datanl$fruto_n,
                  diametro_medio=datanl$diametro_medio, ddplf=datanl$
                     ddplf, stringsAsFactors=FALSE))
    ides_frut <- paste("Fruto", i)
    dif_obss <- c(dif_obss,dif_obs)
    ide_frut <- c(ide_frut,ides_frut)
  }
  trans2 <- trans(dfan)
  menos3 <- trans2$n_obs > 3
  trans2 <- trans2[menos3, ]
  lista <- list(trans_ult = trans2, res_frut =
                  data.frame(iden_frut=ide_frut,
                             dife_obs = dif_obss, mas3=menos3),
                dfan=dfan,tot_frut=nrow(antrans), tot_elim = sum(!
                   menos3)+sum(!menos3a)
  )
}
```

Código R 8.3: Función para preprocesamiento de los datos y verificación de curvas de crecimiento

Función para encontrar valores iniciales en un modelo no lineal permite ser utilizado dentro de la función `nls` de R.

```
#Modelo no lineal
#Diametro_Medio ~ alpha/(1 + exp(beta - gamma * DDPF))
#Modelo no lineal escrito en terminos de funcion
frutcurv1Model <- function(predictor, crecmax,mincrec,tasa){
    crecmax/(1 + exp(mincrec - tasa * predictor))
}
##Definir los  valores iniciales
frutcurv1ModelInit <- function(mCall,LHS,data){
   #ordena los datos por el predictor
   xy <- sortedXyData(mCall[["predictor"]], LHS, data)
    maximo <- max(xy[,"y"])
    crecmax <- maximo +maximo * 0.01
    lmFit <- lm(log(crecmax/xy[,"y"]-1) ~ xy[,"x"])
    coefs <- coef(lmFit)
    mincrec <- coefs[1]
    tasa <-abs(coefs[2])
    value <- c(crecmax,mincrec,tasa)
    names(value) <- mCall[c("crecmax","mincrec","tasa")]
    value
}
#Asigna el modelo y los valores iniiciales a selfStart
SSfrutcurv1 <- selfStart(frutcurv1Model, frutcurv1ModelInit,
    c("crecmax","mincrec","tasa"))
```

Código R 8.4: Función para encontrar los valores iniciales en un modelo no lineal(selfStart)

Función para ajustar el modelo no lineal a cada uno de los frutos.

```
## Funcion que ajusta modelos no lineales a los frutos individuales
## Ademas permite realizar predicciones para un ddplf en cada modelo ajustado
ajusteNLSfrut <- function(datanl,dafb,prediccion = NULL){
        frutcurva <- nls(diametro_medio ~ SSfrutcurv1(dafb,
                        crecmax,mincrec,tasa), data = datanl,
                        control=list(maxiter = 150, tol = 1e-05,
                        minFactor = 1/1024, printEval = FALSE,
                        warnOnly = T))
        resume  <- summary(frutcurva)
        salida <- list()
        iterar <- frutcurva$convInfo$finIter
        veccoef <- coef(frutcurva)
        Alpha   <- veccoef[1]
        Beta    <-  veccoef[2]
        Gamma   <- veccoef[3]
        std_err <- resume$sigma
        paramdf <- data.frame(Alpha = Alpha, Beta = Beta,
                              Gamma = Gamma,
                              niter = iterar,std_err=std_err)
        salida <- paramdf
        if(!is.null(prediccion)) {
            prediccionval <- predict(frutcurva,
            data.frame(dafb = prediccion))
           salida <- list(parametors = paramdf,
           prediccion = prediccionval) }
        return(salida)
}
nitergreat <- dfan[dfan$niter > 50 | dfan$std_err >2.5,] #Cambio 50
```

Código R 8.5: Función para ajustar un modelo no lineal a cada fruto (utiliza la función selfStart)

Función para ajustar un modelo potencial a los datos de peso y diámetro. Extrae los coeficientes y los linealiza disponiéndolos para su posterior uso en la escala original.

```
#Ajustar el peso y el diametro para encontrar la órelacin potencial
fitpesodiam <- function(formulate, data, imprimir=FALSE){
    datapesodiam <- as.data.frame(data)
    form <- formula(formulate)
    Yresp <- datapesodiam[,all.vars(form)[1]]
    Xpred <-  datapesodiam[,all.vars(form)[2]]
pesoddanjou.lm <- lm(log(Yresp) ~ log(Xpred))
    summarypes <- summary(pesoddanjou.lm)
    #Guardar los coeficientes
coeflin <- coef(pesoddanjou.lm)
    #Como es un modelo potencial transformo con el antilog
    a_pesodim <- exp(coeflin)[1]
    b_pesodim <- coeflin[2]
    r2 <- summarypes$r.squared
    sterr <- summarypes$sigma
    salida <- list(resumen=summarypes,a_lin =  a_pesodim, b_lin = b_
        pesodim,
                   R2 = r2, StanErr = sterr)
    if (imprimir){
    cat("Resumen_____________________________________________", "\n",
        "coef.lin.a=", salida$a_lin,
        "|", "coef.lin.b=", salida$b_lin, "\n" )
    }
}
```

Código R 8.6: Función para ajustar el modelo potencial que relaciona el peso y el diámetro de los frutos

Marcos de datos para los distintos envases trabajados para manzanas y peras con los pesos para tamaños comerciales de cajas de 19 kilogramos y tipo Mark IV para manzana y cajón estándar 4/5 de 19 kilogramos y $18,2$ kilogramos para pera. Los tamaños comerciales son guardados en vectores de tipo factor porque permite que los caracteres guardados mantengan el orden correspondiente al tamaño comercial.

```
1
2 #Caja de 19 kilogramos para manzana
3 caja19mza <- data.frame(peso = c(112, 122, 132, 145, 160,179,
4                 202, 226, 250, 280, 318, 800),
5                 tamanios = factor(c("FT_p","T163","T150",
6                 "T138","T125","T113","T100","T88","T80",
7                 "T72","T64","FT_g"), levels =c("FT_p",
8                 "T163","T150","T138","T125","T113","T100",
9                 "T88","T80","T72","T64","FT_g")))
10 #Caja Mark IV de 18 kgs para manzana
11 mark4mza <- data.frame(peso = c(114.5, 127, 142, 157, 172,
12                 190, 212.5, 241, 278.5,321, 800),
13                 tamanios = factor(c("FT_p", "T150", "T135",
14                 "T120","T110", "T100", "T90", "T80","T70",
15                 "T60","FT_g"),
16                 levels = c("FT_p", "T150", "T135", "T120",
17                 "T110", "T100", "T90", "T80","T70", "T60",
18                 "FT_g")))
19 #óCajn áestndar 4/5 de 19 kilogramos para pera
20 std4_5pera <- data.frame(peso = c(120, 131, 141, 152, 165.5,
21                 181, 201, 224, 251, 280, 800),
22                 tamanios = factor(c("FT_p","T150","T140",
23                 "T130","T120","T110","T100","T90","T80",
24                 "T72", "FT_g"),
25                 levels = c("FT_p","T150","T140","T130",
26                 "T120","T110","T100","T90","T80","T72",
27                 "FT_g")))
28
29 #óCajn de 18,2 kilogramos para pera
30 std182pera <- data.frame(peso = c(105, 116, 128, 143,
31                 158.5, 174, 192, 215, 244, 282, 333, 800),
32                 tamanios = factor(c("FT_p","T165","T150",
33                 "T135","T120","T110", "T100","T90","T80",
34                 "T70","T60", "FT_g"),
35                 levels = c("FT_p","T165","T150","T135",
36                 "T120", "T110", "T100","T90","T80", "T70",
37                 "T60","FT_g")))
```

Código R 8.7: Marco de datos para los envases de peras y manzanas más utilizados en su comercialización

Función para clasificar frutos para cualquier caja definida como data frame y tamaño comercial a partir del registro de peso de los frutos.

```
selec_todastablas <- function(data.tabla,sss){
    x <- 1
    while (data.tabla[x,1] < sss) {
        x <- x + 1
    }
    tamasel <- data.tabla[x,2]
    return(tamasel)
}
## Funcion 'clasifica' permite clasificar un vector de pesos utilizando
## la ófuncin anterior
clasifica <- function(p,data.tabla){
    clasificacion <- sapply(p,function(p) selec_todastablas(data.tabla,p))
    return(clasificacion)
    }
```

Código R 8.8: Clasificación del peso de los frutos a tamaño comercial a partir de un envase comercial

Algoritmo de simulación de curvas de crecimiento

```
library(nlme) # Ajuste de modelos y ósimulacin AR1
library(MASS) # óFuncin de ósimulacin normal multivariada
alpha <- 88.37197;beta  <- 1.92866;gamma <- 0.02332
temporada <- gl(4,1, label = paste("Temp", 1:4, sep = ""))
ntemporada <- length(temporada)
parcela <- gl(3,1, label = c("Parc1", "Parc2", "Parc3"))
nparcela <- length(parcela)
train <- c("Cond1", "Cond2")
ntrain <- length(train)
planta <- gl(5,1, label = paste("Pl",1:5, sep=""))
nplanta <- length(planta)
size <- gl(3, 1, label = c("P", "M", "G"))
nsize <- length(size)
fruto <- 1:5
nfruto <- length(fruto)

dafb <- seq(30, 135, 7)
ndafb <- length(dafb)

datos <- expand.grid(dafb, fruto, size, planta,train,parcela,
   temporada)
colnames(datos) <- c("dafb", "Fruto", "Size", "Planta", "Train",
                     "Parcela", "Temporada")

#Efecto temporada
ef_temporada <- rnorm(ntemporada, 0,  0.01819495)
ef_temporadas <- rep(ef_temporada,
                     each = ndafb*nfruto*nsize*nplanta*ntrain*
                        nparcela)

datos <- cbind(datos, ef_temporadas)

                                        #Parcela
## En la ósimulacin se ingresa la matriz de varianzas convarianzas
ef_parc <- mvrnorm(ntemporada*nparcela,
                   mu = rep(0,3),
                   Sigma = diag(c(7.778209^2, 0.08515845^2,
                      0.00351^2)))
ef_parcs <-  rep(as.vector(ef_parc), each = ndafb*nfruto*nsize*
   nplanta*ntrain)
ef_parcsm <- matrix(ef_parcs, ncol = 3)
colnames(ef_parcsm) <- c("parc_alfa", "parc_beta", "parc_gamma")
datos <- cbind(datos, ef_parcsm)



                                        #train(sistema de conduccion)
ef_trainn <- mvrnorm(ntemporada*nparcela*ntrain, mu = rep(0,3),
    diag(c(0.007356164^2, 0.08463421^2, 0.0008014243^2)))
ef_trains <-  rep(as.vector(ef_trainn), each = ndafb*nfruto*nsize*
```

```
    nplanta)
47 ef_trainm <- matrix(ef_trains, ncol = 3)
48 colnames(ef_trainm) <- c("train_alfa", "train_beta", "train_gamma")
49 datos <- cbind(datos, ef_trainm)
50
51
52 #planta
53 ef_planta <- rnorm(nplanta*ntrain*nparcela*ntemporada, 0,
    0.0000001275763)
54 ef_plantas <- rep(ef_planta, each = ndafb*nfruto*nsize)
55
56 datos <- cbind(datos, ef_plantas)
57                                          #Size
58
59 ef_sizen <- mvrnorm(nsize*nplanta*ntrain*nparcela*ntemporada, mu =
    rep(0,3),
60     diag(c(4.502842^2, 0.1267108^2, 0.0008752909^2)), empirical = T)
61 ef_sizens <-  rep(as.vector(ef_sizen), each = ndafb*nfruto)
62 ef_sizem <- matrix(ef_sizens, ncol = 3)
63 colnames(ef_sizem) <- c("size_alfa", "size_beta", "size_gamma")
64 datos <- cbind(datos, ef_sizem)
65
66
67 ## Fruto
68
69 ef_fruit  <- mvrnorm(nfruto*nsize*nplanta*ntrain*nparcela*ntemporada,
70                      mu = rep(0,2), Sigma= diag(c(2.342049^2,
                         0.0006291349^2)))
71 ef_fruits <-  rep(as.vector(ef_fruit), each = ndafb)
72 ef_fruitm <- matrix(ef_fruits, ncol = 2)
73 colnames(ef_fruitm) <- c("fruit_alfa", "fruit_beta")
74 datos <- cbind(datos, ef_fruitm)
75
76 ##Factores anidados
77 datos$PlanTemp <- paste(datos$Fruto, datos$Size, datos$Planta, datos$
    Train,
78                         datos$Parcela, datos$Temporada, sep="/")
79
80 diam <- NULL
81 for (j in unique(datos$PlanTemp)){
82 valor <- NULL
83 valor <- datos[datos$PlanTemp== j, ]
84 diam0 <- frutcurv1Model(valor$dafb,
85                         alpha  + valor$ef_temporadas[1] + valor$parc_
                            alfa[1] +
86                         valor$train_alfa[1] + valor$ef_plantas[1] +
87                         valor$size_alfa[1] + valor$fruit_alfa[1],
88                         beta + valor$parc_beta[1] + valor$train_beta
                            [1] +
89                         valor$size_beta[1] + valor$fruit_beta[1],
90                         gamma + valor$parc_gamma[1] + valor$train_
```

```
                                gamma[1] +
91                            valor$size_gamma[1])
92 diam <- c(diam, diam0)
93 }
94
95 datos <- cbind(datos, diam)
96
97 ##Simulacion de la correlacion de los datos
98 curAR1 <- corAR1(0.66, form = ~ 1 | Temporada/Parcela/Train/Planta/
       Size/Fruto)
99 cualquiera_group <- groupedData(diam ~
100                                    dafb|Temporada/Parcela/Train/
                                          Planta/Size/Fruto,
101                                data = datos)
102      cursAR1 <- Initialize(curAR1, data = cualquiera_group)
103 Sigma4 <- matrix(unlist(corMatrix(cursAR1)[1]),ncol = ndafb)
104 datos4 <- mvrnorm(nfruto*nsize*nplanta*ntrain*nparcela*ntemporada,
105                   rep(0, ndafb),
106                   Sigma4, empirical = TRUE)
107
108 err_cor <- as.vector(t(datos4))
109
110 datos <- cbind(datos, err_cor)
111
112 datosnew <- within(datos, {
113     Fruto <- as.factor(Fruto)
114     diametro_final_cor<- diam + err_cor*1.32
115 }
116 )
117
118 ##Verificacion de la simulacion
119 curvas_group <- groupedData(diametro_final_cor ~
120                                  dafb|Temporada/Parcela/Train/Planta/
                                        Size/Fruto,
121                              data = datosnew)
122
123 curvas_nlme <- nlme(diametro_final_cor ~ frutcurv1Model(dafb, alpha,
       beta, gamma),
124                     data = curvas_group, fixed = alpha+beta+gamma~1,
125                     random = list(Temporada = pdDiag(alpha ~ 1),
126                                   Parcela   = pdDiag(alpha + beta +
                                        gamma ~ 1),
127                                   Train   = pdDiag(alpha + beta +
                                        gamma ~ 1),
128                                   Planta = pdDiag(alpha ~ 1),
129                                   Size   = pdDiag(alpha + beta + gamma
                                        ~ 1),
130                                   Fruto =pdDiag(alpha + beta  ~ 1)),
131                     start=c(alpha=70,beta=1.9,gamma=0.03),
132                     cor = corAR1()
133                     )
```

Código R 8.9: Algoritmo de simulación para los efectos considerados en el modelo no lineal mixto ajustado

Algoritmo de predicción de tamaño a cosecha a partir de los parámetros de los modelos de crecimiento para cualquier caja comercial y los ddplf.

```
pronostico <- function(mat_param, dafb, valor_diam,
imp_list = FALSE){
 #Input: matriz de parametros alpha, beta y gamma para los
 #tamanios comerciales,ddplf en que se realiza el pronostico
  mat_param <- as.matrix(mat_param)
 #Function modelo Curvas
  frutcurv1Model <- function(predictor, crecmax, mincrec,
  tasa){crecmax/(1 + exp(mincrec - tasa * predictor))
  }
 #Diametros para los ñ tamaos comerciales al momento ddplf
  predic_mat <- apply(mat_param, 1, function(x) frutcurv1Model(dafb
      ,
                  x[1],x[2], x[3]))
  ultfil <- length(predic_mat)
 #Si los valores son mayores o menores a los calibres
 #comerciales directamente los clasifica como ft_g y ft_p,
 #si los valores estan dentro de los calibres calcula el
 #promedio entre dos clases comerciales y se compara el
 #diametro con ese promedio,si es mayor se clasifica con
 #la clase mayor si es menor con  la clase menor
  if (valor_diam < predic_mat[1] | valor_diam > predic_mat[ultfil])
     {
      decis_mat <- ifelse(valor_diam < as.numeric(predic_mat[1]), "
         ft_p", "ft_g")
  } else {
      i <- 2
      while (predic_mat[i] <  valor_diam ) {
          i <- i + 1
      }
      media_mat <- mean(c(predic_mat[i], predic_mat[i-1]))
      decis_mat <- ifelse(valor_diam > media_mat, names(predic_mat)
         [i],
                          names(predic_mat)[i-1])
  }
  lista_prono <- list(matriz_de_prediccion = predic_mat,
                      calibre_predicho = decis_mat)
  if (imp_list == TRUE )  return(lista_prono)
  if (imp_list == FALSE) return(decis_mat)
}
```

Código R 8.10: Algoritmo de prediccion de cosecha a partir de las curvas de crecimiento

Bibliografía

Agilandeeswari, L., Prabukumar, M., and Goel, S. (2017). Automatic grading system for Mangoes using multiclass SVM classifier. *International Journal of Pure and Applied Mathematics*, 116(23):515–523.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., third edition edition.

Alvarez, O., Bramardi, S., Hervatin, J., Reeb, P., Tassile, V., Carranza, P., Bogado, D., and De Bernardin, D. (2002). Proyecto de Investigación 04/A065 'Técnicas multivariadas y modelos no-lineales mixtos aplicados al mejoramiento del pronóstico de producción de fruta de pepita' (2002-2005). Technical report, Universidad Nacional del Comahue.

Alvarez, O., Bramardi, S., Stangaferro, S., Hervatin, J., Boche, S., Lavalle, A., and Reeb, P. (1996). Proyecto de Investigación 04/A031 'Análisis y generación de modelos de crecimiento en frutos de pepita, carozo y baya' (1996-1998). Technical report, Universidad Nacional del Comahue.

Alvarez, Omar & Boche, S. (1999). Modelos Matemáticos para Describir Crecimientos Doble Sigmoideos en Frutos de un Nectarín Tardío (c.v. Sun Grand). *Agro Sur*, 27(1):21–28.

Apcarian, A., Echenique, M. d. C., Aruani, C., and Reeb, P. (2006). Efecto de capas endurecidas de suelos sobre el potencial productivo de viñedos, Alto Valle de Río Negro, Patagonia Argentina. *Agricultura Técnica*, 66(1):70–79.

Armstrong, L., Diepeveen, D., and Maddern, R. (2007). The application of data mining techniques to characterize agricultural soil profiles. In Christen, P., Kennedy, P., Li, J., Kolyshkina, I., and Williams, G., editors, *Proceedings of the sixth Australasian conference ond Data Mining and analytics*, volume 70, pages 85–100. Australian Computer Society.

Avanza, M. (2010). *Desarrollo de una propuesta metodológica para la predicción de cosecha de naranjo dulce (Citrus sinensis L. Osbeck) var. Valencia late en la provincia de Corrientes, Argentina* . PhD thesis, Universidad Nacional del Nordeste.

Avanza, M., Bramardi, S., and Mazza, S. (2008). Statistical Model to Describe Growth Pattern in Sweet Orange "Valencia Late". *Spanish Journal of Agricultural Research*, 6(4):577–585.

Avanza, M. M., Bramardi, S. J., and Mazza, S. (2010). Optimal sample size for evaluate the growth pattern of 'Valencia Late' orange fruit. *Revista Brasileira de Fruticultura*, 32(4):1154–1163.

Bandieri, Susana & Blanco, G. (2014). La fruticultura en el Alto Valle de río Negro. Auge y crisis de una actividad capitalista intensiva. *Revista de Historia*, 0(2):127–141.

Baumer, B., Kaplan, D., and Horton, N. (2017). *Modern Data Science with R*. Text in Statistical Science. CRC press.

Beck, M. W. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, 85(11):1–20.

Behrens, J. (1997). Principles and Procedures of Exploratory Data Analysis. *Psychological Methods*, 2(2):131–160.

Bennett, K. P. and Campbell, C. (2000). Support Vector Machines: Hype or Hallelujah? *SIGKDD Explorations*, 2:1–13.

Bergh, O. (1990). Effect of Temperature during the first 42 days following full bloom on apple fruit growth and size at harvest. *South African Journal of Plant and Soil*, 70(1):11–18.

Bergmeir, C. and Benítez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software*, 46(7):1–26.

Bestvater, C. and Casamiquela, C. (1983). Distribución textural de los suelos del Alto Valle del Río Negro. Technical report, Alto Valle del INTA.

Blanco, G. (1999). El Alto Valle del río Negro y la fruticultura La historia de un origen pionero, un pasado de gloria y un presente difícil. In *Fruticultura Moderna – Tecnología, transferencia, capacitación, organización – 9 Años de cooperación técnica. 1990/1999.*, pages 1–10. Inta/GtZ.

Boehmke, B. and Greenwell, B. (2019). *Hand-On Machine Learning with R*. The R Series. Chapman & Hall/CRC, first edition.

Bound, S. A. (2005). *The Impact of Selected Orchard management practices on apple (Malus domestica L.) fruit quality*. PhD thesis, University Of Tasmania.

Bramardi, S. (1989). Modelos de predicción de tamaño de frutos a la cosecha en base a mediciones sucesivas durante el período de crecimiento. Peras cv. Williams y Packham's Triumph en la región del Alto Valle de Río Negro. Master's thesis, Universidad Nacional del Comahue-Universidad de Buenos Aires-Instituto Nacional de Tecnología Agropecuaria.

Bramardi, S., Castro, H., and Zanelli, M. (1998). Fruit Growth Pattern of Pears cv. 'Williams' and 'Packhams Triumph'to Improve Hand Thinning. *Acta Hort.*, 475:283–287.

Bramardi, S., Tassile, V., and Reeb, P. (2006). Crecimiento de Frutos de Pepita en el Alto Valle. Tablas de Raleo. Technical report, Universidad Nacional del Comahue.

Bramardi, S. J., Tassile, V., Reeb, P., and De Bernardin, F. (2005). Comparación de métodos para la predicción anticipada del peso del fruto medio a cosecha: curvas de crecimiento vs. modelos estocásticos. *X Reunión Científica del Grupo Argentino de Biometría.*

Bramer, M. (2007). *Principles of Data Mining*. UTICS. Springer.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28.

Bzdok, Danilo ; Altman, N. K. M. (2018). Statistics versus Machine Learning. *Nature Methods*, 15(4):21–28.

Carrero, O., Jerez, M., Macchiavelli, R., Orlandoni, O., and Stock, J. (2008). Ajuste de Curvas de Índice de Sitio Mediante Modelos Mixtos para Plantaciones de Eucalyptus urophylla en Venezuela. *Interciencia*, 33(4):265–272.

Casamiquela, C. and von Wagner, A. (1999). *La fruticultura en las Provincias de Río Negro y Neuquén*. LyM S.R.L.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM : a library for support vector machines. *AcM Transactions on Intelligent Systems and Technology*, 2(27):1–27. ISBN 3-900051-07-0.

Christensen, R. H. B. (2019). ordinal—Regression Models for Ordinal Data. R package version 2019.4-25. http://www.cran.r-project.org/package=ordinal/.

Ciampi, Antonio & Lechevallier, Y. (2007). *Statistical Models and Artificial Neural Networks: Supervised Classification and Prediction Via Soft Trees*, pages 239–261. Birkhäuser Boston, Boston, MA.

Cordon, V. H., Forquera, J. C., Gastiazoro, J., Lässig, J., Bastanski, M., and Nordenstrom, G. (2000). Caracterización Climática del Alto Valle del Río Negro, Neuquén y Limay Inferior. Technical Report 5722, Universidad Nacional del Comahue.

Costa, G., Botton, A., Vizzotto, G., et al. (2018). Chapter 4:Fruit Thinning: Advances and Trends. In Warrington, I., editor, *Horticultural Review*, volume 46, page 185. John Wiley & Sons.

Costa, G., Dal Cin, V., and Ramina, A. (2006). Physiological, Molecular and Practical Aspects of Fruit Abscission. *Acta Hort.*, 727:301–309.

CPIARN (2015). Aportes para la Reconstrucción de una Fruticultura Sustentable. Technical report, Consejo Profesional de Ingeniería Agronómica Río Negro.

Damico, J. (1993). Censar '93. Censo Agrícola Rionegrino. Technical report, Subsecretaría de Fruticultura.

Date, C. (2001). *Introducción a los Sistemas de Bases de Datos*. Pearson Prentice Hall, séptima edition.

Davidian, M & Giltinan, D. (2003). Nonlinear models for repeated measurements: an overview and update. *Journal of Agricultural Biological and Environmental Statistics*, 8(4):387–419.

Davidian, Marie & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, London.

Dumbill, E. (2012). What is Big Data? In *Big Data Now*, chapter 2, page 123. O'Really.

Efron, Bradley & Hastie, T. (2018). *Computer Age Statistical Inference. Algorithms, Evidence, and data science.* Cambridge University Press.

Elmasri, Ramez & Navathe, S. (2011). *Fundamentals of Database Systems*. Prentice Hall International, 6 edition.

FAO, U. A. (2015). Informe de Diagnóstico de los princiaples valle y áreas con potencial agrícola de la provincia de Río Negro.

FAOSTAT (2019). Datasets Argentina.

Faraway, J. (2006). *Extending the Linear Model With R*. Chapman & Hall/CRC Taylor & Francis Group.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, 39(11):27–34.

Fernandez-Martinez, R., Ascacibar, F. M.-d.-P., Espinoza, A. V. P., and Lorza, R. L. (2011). Predictive modelling in grape berry weight during maturation process: comparison of data mining, statistical and artificial intelligence techniques. *Spanish Journal of Agricultural Research*, 9(4):1156–1167.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. John Wiley, 2nd edition.

Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, 138(3):57–70.

Galeano-Vasco, Luis & Cerón-Muñoz, M. (2013). Modelación del crecimiento de pollitas Lohmann LSL con redes neuronales y modelos de regresión no lineal. *Rev.MVZ Córdoba*, 18(3):3861–3867.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning. With applications in R.* Springer.

Gariglio, N., Pilatti, R. A., Fonfría, M. A., et al. (2007). Capítulo 2: requerimiento ecofisiológicos de los árboles frutales. In Sozzi, G., editor, *Árboles frutales: ecofisiología, cultivo y aprovechamiento.* Universidad de Buenos Aires. Facultad de Agronomía.

Ghaiwat, S. and Arora, P. (2014). Detection and Classification of Plant Leaf Diseases Using Image processing Techniques: A Review . *International Journal of Recent Advances in Engineering and Technology*, 2(3):2347–2812.

Gil, F. (1996). Bases para elaborar un pronóstico de cosecha en manzano (Malus pumila Mill) cv. Fuji. Factores que modifican el crecimiento del fruto. Master's thesis, Universidad de Chile.

Gillaspy, G., Ben-David, H., Gruissem, W., et al. (1993). Fruits: A Developmental Perspective. *The Plant Cell*, 5.

Giménez, G., Dussi, M. C., Reeb, P., Zon, K., Nyeki, J., Szabó, Z., and Racsko, J. (2010a). Fruit Growth and Abscission Pattern of 'Williams' Pear Treated With Benzyladenine. *Acta Horticulturae*, 884:481–490.

Giménez, G., Reeb, P., Dussi, M. C., Elosegui, F., FAntaguzzi, S., and Sugar, D. (2010b). Optimizing Benzyladenine Application Timing in 'Williams' Pear. *Acta Horticulturae*, 888:265–272.

Giménez, Gustavo & Tassile, V. (2015). Using Non-linear mixed models and artificial neural network in the fitting growth pattern in pears cv. 'Williams' tp predict final sizes at harvest. In *Proceedings of the XV Conferencia Española y V Encuentro Iberoamericano de Biometría.*

Gironés Roig, J. (2013). *Algoritmos*. openlibra, business analytics edition.

Godagnone, R. E. and Bran, D. E., editors (2009). *Inventario integrado de los recursos naturales de la provincia de Río Negro. Geología, Hidrogeología, Geomorfología, Suelos, Clima, Vegetación y Fauna*. Ediciones INTA.

Godoy, C., Monterubbianesi, G., and Tognetti, J. (2008). Analysis of Highbush blueberry (Vaccinum corymbosum L.) fruit growth with exponencial mixed models. *Scientia Horticulturae*, 115:368–376.

Goldstein, H. (1998). Multilevel Models. In Armitage, P. and Colton, T., editors, *Encyclopaedia of Biostatistics*. Wiley.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Grolemund, G. and Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Günther, F. and Fritsch, S. (2010). neuralnet: Training of Neural Networks. *The R Journal*, 2(1):30–38.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. Grupo Editorial Patria, second edition edition.

Hallgren Kevin A (2013). Conducting Simulation Studies in the R Programming Environment. *Tutorials in quantitative methods for psychology*, 9(2):43–60.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining.Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, third edition edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning.Data Mining, Inference and Prediction*. Springer, second edition.

Hector, C. and Tiscornia, J. (1968). Predicción del Tamaño que Alcanzan los Frutos de Manzano en la Cosecha. Technical report, INTA.

Henry, L. and Wickham, H. (2018). *purrr: Functional Programming Tools*. R package version 0.2.5.

Hill, M., Connolly, P., Reutemann, P., and Fletcher, D. (2014). The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. *Computers and Electronics in Agriculture*, 108:250–257.

Hsu, C.-W., Lin, C.-J., et al. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 138:415–425.

Iqbal, N., Nazar, R., Khan, M. I. R., Masood, A., Khan, N. A., et al. (2011). Role of gibberellins in regulation of source– sink relations under optimal and limiting environmental conditions. *Current Science*, 100(7).

Iwańska, M., Oleksy, A., Dacko, M., Skowera, B., Oleksiak, T., and Wójcik-Gron, E. (2018). Use of classification and regression trees (CART) for analyzing determinants of winter wheat yield variation among fields in Poland. *Biometrical Letters*, 55(2):197–214.

James, D. and DebRoy, S. (2012). RMySQL: R interface to the MySQL database. http://biostat.mc.vanderbilt.edu/RMySQL.

Janssen, B., Thodey, K., Schaffer, R., Alba, R., Balakrishnan, L., Bishop, R., Bowen, J., Crowhurst, R., Gleave, A., Ledger, S., McArtney, S., Pichler, F., Snowden, K., Ward, S., et al. (2008). Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. *BMC Plant Biology*, 8(16).

Jung, Sook, Lee, Taein, Cheng, Chun-Huai, Buble, Katheryn, Zheng, Ping, Yu, Jing, Humann, Jodi, Ficklin, Stephen P, Gasic Ksenija, Scott Kristin, Frank Morgan, Ru Sushan, Hough Heidi, Evans Kate, Peace Cameron, Olmstead Mercy, DeVetter Lisa W, McFerson James, Coe Michael, Wegrzyn Jill L, Staton Margaret E, Abbott Albert G, Main Dorrie, et al. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic acids research*, 47(D1):D1137–D1145.

Kaewtapee, C., Khetchaturat, C., and Bunchasak, C. (2011). Comparison of growth models between artificial neural networks and nonlinear regression analysis in Cherry Valley ducks. *The Journal of Applied Poultry Research*, 20(4):421–428.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, 15(9):1–28.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software, Articles*, 11(9):1–20.

Kavitakomal, S. (2019). Quality Assessment Of Orange Fruit Using Svm Classifier And Gray Level Co-Occurrence Matrix Algorithm. . *International Journal of Scientific & Technology Research*, 8(11).

Kéry, M. and Royle, J. A. (2016). Chapter 4 - Introduction to Data Simulation. In Kéry, M. and Royle, J. A., editors, *Applied Hierarchical Modeling in Ecology*, pages 123–143. Academic Press, Boston.

Kowalcxyk, A. (2017). *Support Vector Machines. Succintly.* Syncfusion.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5):1–26.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *BIOMETRICS*, 38(4):963–974.

Lakso, A., Corelli Grapadelli, L., Barnard, J., and Goffinet, M. (1995). An Expolinearmodel of the growth pattern of the apple fruit. *Journal of Horticultural Science*, 70(4):389–394.

Liakos, K., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Review Machine Learning in Agriculture: A Review . *Sensors*, 18:2674.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276.

Loh, W.-Y. (2011). Classification and Regresion Trees. *WIREsComputational Statistics*, 1:14–23.

Loh, W.-Y. (2014). Fifty years of Classification and Regression Trees. *International Statistical Review*, 82(3):329–248.

Lötze, E. and Bergh, O. (2004). Early Prediction of Harvest Fruit Size Distribution of an Apple and Pear Cultivar. *Scientea Horticulturae*, 101:281–290.

Maimon, O. and Rokach, L. (2010). *Data Mining and knowledge Discovery Handbook*. Springer.

Maletic, J. and Marcus, A. (2010). Data Cleansing: A prelude to Knowledge Discovery. In *Data Mining and knowledge Discovery Handbook*, chapter 2, page 1306. Springer, second edition.

Manabu, W., Hideyuki, S., Masanobu, M., Satoru, S., Sadao, K., et al. (2008). Effects of Plant Growth Regulators on Fruit Set and Fruit Shape of Parthenocarpic Apple Fruits. *Journal of the Japanese Society for Horticultural Science*, 77(4):350–357.

Marticorena, M., Bramardi, S., and Defacio, R. (2013). Aplicación de Análisis Multibiplot a la evaluación de Recursos Fitogenéticos. In *Libro de Resúmenes X Clatse.*, pages 240–242. Grupo Argentin de Biometría.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019a). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-0.1.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019b). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-1.

Milborrow, S. (2019). *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. R package version 3.0.8.

Moineddin Rahim, Matheson Flora I, and Glazier Richard H (2007). A simulation study of sample size for multilevel logistic regression models. *BMC medical research methodology*, 7:34–34.

Montgomery, D., Peck, E., and Vining, G. (2007). *Introducción al Análisis de Regresión Lineal*. Grupo Editorial Patria, 3a.edición edition.

Morgan Kain , Ben Bolker , and Michael McCoy (2015). A practical guide and power analysis for GLMMs: detecting among treatment variation in random effects. *PeerJ*, 3:e1226–e1226.

Müller, K., Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2019). *RMariaDB: Database Interface and 'MariaDB' Driver*. R package version 1.0.8.

Müller, K. and Wickham, H. (2019). *tibble: Simple Data Frames*. R package version 2.1.3.

Nations, U. (2019). United Nations Statistics Division.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press.

Padol, P. B. and Yadav, A. A. (2016). SVM Classifier Based Grape Leaf Disease Detection. *Conference on Advances in Signal Processing*, pages 175–179.

Palmer, A., Jimenez, R., and Gervilla, E. (2011). Data Mining: Machine Learning and Statistical Techniques. In Funatsu, K., editor, *Knowledge-Oriented Applications in Data Mining*.

Patel, A. and Kathiriya, D. (2017). Data Mining Trends in Agriculture: A Review. *An International e Journal*, 6(4):637–645.

Pinheiro, J. and Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2019). *nlme: Linear and Nonlinear Mixed Effects Models*.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Powers, D. M. W. (2012). The Problem with Kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355, Avignon, France. Association for Computational Linguistics.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

(R-SIG-DB), R. S. I. G. o. D., Wickham, H., and Müller, K. (2019). *DBI: R Database Interface*. R package version 1.1.0.

Ramesh, D. and Vishnu Vardhan, B. (2013). Data Mining Techniques and Applications to Agricultural Yield Data. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(9):3477–3480.

Raorane, A. and Kulkarni, R. (2013). Review- Role of Data Mining in Agriculture. *International Journal of Computer Science and Information Technologies*, 4(2):270–272.

Razi, M. A. and Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models . *Expert Systems with Applications*, 29:65–74.

Reeb, P., Bramardi, S. J., and Alvarez, O. (2003). Estudio de la variabilidad de la producción de manzanas Red Delicious en los montes frutales del Alto Valle de Río Negro. *Agrosur*, 31:21–26.

Ritz, C. and Streibig, J. C. (2008). *NonLinear Regression With R*. Springer.

Robinson, T. and Lakso, A. (2004). Between Year and Within Year Variation in Chemical Fruit Thinning Efficacy Of Apple During Cool Springs. *Acta Horticulturae*, 636:283–294.

Romero, M. d. C. (2009). Selección de atributos en contextos de alta dimensionalidad. Master's thesis, Universidad Nacional de Córdoba.

Rong, X. (2014). *deepnet: deep learning toolkit in R*. R package version 0.2.

RoSS, J. (2018). Agricultural Statistics 2018. Technical report, USDA.

Rubio, N. (2016). Modelación de Respuestas Ordinales Longitudinales Mediante Modelos Lineales Generalizados Mixtos. Master's thesis, Universidad Nacional del Comahue.

Sanchez, E. E. and Villareal, P. (2015). Cadena Frutales de Pepita. Technical report, INTA.

Schabenberger, O. and Pierce, F. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences.* CRC press.

Semary, N. A., Tharwat, A., Elhariri, E., and Hassanien, A. E. (2014). Fruit-Based Tomato Grading System Using Features Fusion and Support Vector Machine . *Advances in Intelligent Systems*, 323.

Senasa (2017). Anuario Estadístico 2017. Technical report, Servicio Nacional de Sanidad y Calidad Agroalimentaria-Centro Regional Patagonia Norte.

Silberschatza, A., Korth, H., and Sudarshan, S. (2002). *Fundamentos de Bases de Datos*. McGrowHill, cuarta edition.

Smith,Steven M, Fulton, Daniel C, Chia, Tansy, Thorneycroft, David, Chapple, Andrew, Dunstan, Hannah, Hylton, Christopher, Zeeman, Samuel C, and Smith, Alison M (2004). Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in Arabidopsis leaves. *Plant physiology*, 136(1):2687–2699.

Stangaferro, S., Alvarez, O., and Bramardi, S. (2001). Modelos de predicción anticipada del peso medio de los frutos en producción de manzanas. In - VELAEM, A., editor, *X Congreso Latinoamericano de Biomatemática*. ALAB - VELAEM.

Stegman, G., Jacobucci, R., Harring, J., and Grimm, K. (2017). Nonlinear Mixed-Effects Modeling Programs in R. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–6.

Suresha, M., Shilpa, N., and Soumya, B. (2012). Apples Grading based on SVM Classifier. *International Journal of Computer Applications*, pages 27–30.

Tassile, V. and Giménez, G. (2013). Pronóstico de producción en frutales de pepita en la región del Alto Valle. In *Presentación en el primer encuentro de pronosticadores.* IDR.

Tassile, V., Giménez, G., Bramardi, S., Sepúlveda, M., García, A., et al. (2013). Resultados del Pronóstico de Producción 2013-2014. Technical report, Secretaría de Estado de Fruticultura de Río Negro-Ministerio de Desarrollo Territorial de la Provincia de Neuquén-Facultad de Ciencias Agrarias de la Universidad Nacional del Comahue.

Therneau, T. and Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Vamanan, R. and Ramar, K. (2011). Classification of agricultural land soils a data mining approach. *International Journal on Computer Science and Engineering(IJCSE9*, 3(1).

Van Gerven, M. and Bohte, S. (2018). *Artificial Neural Networks as Models of Neural Information Processing*. Frontiers in Computational Neuroscience, marcel van gerven and sander bohte edition.

Vapnik, V. (1998). *Statistical Learning Theory*. Jhon Wiley and Sons.

Venables, W. N., Smith, D., and Team, R. D. C. (2013). *An Introduction to R*.

Warrington, I., Fulton, T., Halligan, E., and De Silva, H. (1999). Apple fruit growth and maturity are affected by early season temperatures. *Journal of American Society for Horticultural Science*, 124:468–477.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343. Springer-Verlag.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2019). *Advanced R*. Chapman & Hall/CRC The R Series. Chapman and Hall/CRC, second edition.

Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.1.

Wing, M. K. C. f. J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, t. R. C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2019). *caret: Classification and Regression Training*. R package version 6.0-75.

Wismer, P., Proctor, J., and Elvfing, D. (1995). Benzyladenine Affects Cell Division and Cell Size During Apple Fruit Thinning. *Journal of American Society for Horticultural Science*, 120(5):802–807.

Wu, J., olesnikova, A., Song, C.-H., and Lee, W. D. (2009). The Development and Application of Decision Tree for Agriculture Data. In *Second International Symposium on Intelligent Information Technology and Security Informatics*, pages 1–6. IEEE.

Yamamoto, T. and Terakami, S. (2016). Review:Genomics of pear and other Rosaceae fruit trees. *Breeding Science*, 66:148–159.

Yang, X. and Song, Q. (2007). Weighted support vector machine for data classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(5):961–976.

Zhang, C., Tanabe, K., Tamura, F., Matsumoto, K., Yoshida, A., et al. (2005). 13 C-photosynthate accumulation in Japanese pear fruit during the period of rapid fruit growth is limited by the sink strength of fruit rather than by the transport capacity of the pedicel. *Journal of Experimental Botany*, 56(420):2713–2719.

Zhang, C., Tanabe, K., Wang, S., Tamura, F., Yoshida, A., Matsumoto, K., et al. (2006). The Impact of Cell Division and Cell Enlargement on the Evolution of Fruit Size in Pyrus pyrifolia. *Annals of Botany*, 98:537–543.

Zhao, Y. (2013). R and Data Mining: Examples and Case Studies 1. In *Data Mining Application with R*. Elsevier.

Zon, K., Sepúlveda, G., Giménez, G., Dussi, M. C., et al. (2011). Fruit Growth and Cell Evolution in Williams pears. *Acta Horticulturae*, 909.