Regresión Logística y Árboles de Clasificación. Un estudio de simulación para la comparación en el caso de grupos balanceados y desbalanceados.

Celina Beltrán; Ivana Barbona

Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina

beltranc@dat1.net.ar

Abstract

The goal of the present paper is to evaluate and compare two multivariate statistical techniques of classification, Logistic Regression and Classification Trees in order to evaluate the performance based data simulated under different conditions. Case 1 corresponds to data from a population in which the predictors are strongly correlated with the response but are not correlated between them. Case 2 proposes a simulation from a population with low correlation of the response with the predictor variables but these variable are correlated with each other. In case 3, the correlation present in the population is strong both, among the predictors and between them and the response. Finally, case 4 corresponds to a population in which there is no significant correlation between the variables, neither the predictors with the answer nor between them. These data presented two types of modalities for the dichotomous response variable: balanced case and unbalanced. For each sample, 30 extra data were simulated to be considered in the evaluation of the classification without having used them in the estimation processes. It was observed as a main result, that in conditions where the predictor variables are highly correlated with the response, although the CAs showed a significantly lower percentage of error in the classification, both methodologies work satisfactorily. However, when the conditions to obtain a satisfactory classification are unfavorable (predictors that are not correlated with the response), the AC achieved a correct classification percentage that is noticeably higher to the LR. In the unbalanced case, the majority class presented a higher correct classification percentage in the logistic regression at the expense of a worse performance in the minority class. This behavior was more marked in logistic regression than in the classification trees. In those cases where the percentages of correct classification for the two procedures are similar, the logistic regression model is better tan the trees, in the sense of the interpretation of the same parameters.

Keywords: Logistic regression; classification trees; simulation

Resumen

En esta investigación se propone el estudio, evaluación y comparación de dos técnicas estadísticas multivariadas de clasificación, Regresión Logística y Árboles de Clasificación, siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos simulados bajo distintas situaciones.
Se simularon datos bajo 4 condiciones diferentes que diferían en la estructura de correlaciones entre las variables. Asimismo, se combinaron estas cuatro situaciones con otras dos situaciones correspondientes a grupos balanceados y desbalanceados. El escenario 1 corresponde a datos provenientes de una población en la que los predictores están fuertemente correlacionados con la respuesta pero no entre ellos. El escenario 2 plantea una simulación a partir de una población con poca correlación de la respuesta con las variables predictoras pero éstas correlacionadas entre sí. En el escenario 3, la correlación presente en la población origen de la simulación es importante tanto entre las predictoras como entre éstas y la respuesta. Por último, el escenario 4 corresponde a una población original en la que no existe ningún tipo de correlación de magnitud importante entre las variables, ni de los predictores con la respuesta ni entre ellos. Asimismo, estos escenarios presentaban dos tipos de modalidades para la variable respuesta dicotómica: caso balanceado y desbalanceado. Para cada muestra, se simularon 30 datos extras o suplementarios para ser considerados en la evaluación de la clasificación sin haberlos utilizados en los procesos de estimación. Se observó como resultado principal, que en condiciones donde las variables predictoras están altamente correlacionadas con la respuesta, si bien los AC mostraron un porcentaje de error significativamente menor en la clasificación, ambas metodologías funcionan satisfactoriamente. Sin embargo, cuando las condiciones para obtener una clasificación satisfactoria son desfavorables (predictores poco correlacionados con la respuesta) los AC logran un porcentaje de clasificación correcta notablemente superior a la RL. En el caso desbalanceado, la clase mayoritaria presentó un porcentaje de clasificación correcta superior en la regresión logística a costa de un peor desempeño en la clase minoritaria. Este comportamiento estuvo más marcado en regresión logística que en los árboles de clasificación. En aquellos casos donde los porcentajes de clasificación correcta para los dos procedimientos son similares, el modelo de regresión logística tendría la ventaja con respecto a los árboles, en el sentido de la interpretación de los parámetros del mismo.

Palabras clave: regresión logística; árboles de clasificación; simulación

1. Introducción

El Análisis Multivariado se refiere al tipo de análisis que se realiza sobre n unidades experimentales sobre las cuales se han medido p variables y se pretende estudiar a todas las variables (o un gran número) en forma simultánea (Hair, J.F. 1999). Estas variables pueden ser cuantitativas, continuas o discretas, o cualitativas, nominales u ordinales (Pérez López, C. 2004). Uno de los objetivos de dichas técnicas es la clasificación de unidades u objetos en grupos. En la clasificación supervisada, tarea que concierne a este trabajo, se cuenta con un conocimiento a priori, es decir para la tarea de clasificar

un objeto dentro de una categoría o clase se cuenta con la información de p variables observadas en un conjunto de objetos cuya categoría o clase de pertenencia se conoce. Las técnicas de clasificación pueden diferenciarse en aquellos métodos clásicos estadísticos y los que provienen de la Minería de datos. En las técnicas clásicas se estima un modelo estadístico cuyos coeficientes permitirán caracterizar los grupos y construir la regla de clasificación para nuevas unidades. Las inferencias sobre las estimaciones realizadas permiten detectar aquellas características que aportan en el proceso de clasificación. Esto marca una diferencia con las provenientes de la Minería de datos ya que en estos casos generalmente los análisis son de tipo exploratorios y no se realiza una generalización sobre poblaciones de las cuales se extraen los datos.
Otra cuestión a tener en cuenta en esta tarea es la existencia de un desbalanceo de los grupos definidos por la variable respuesta binaria, es decir que existe una clase minoritaria y una mayoritaria, se presenta una dificultad en la clasificación de nuevas unidades. Esta dificultad o inconveniente se refleja en un deterioro del porcentaje de clasificación correcta en los grupos minoritarios, ya que en los grupos de mayor cantidad de observaciones las técnicas seguirán mostrando un buen desempeño. Esta situación es más problemática cuando justamente la clase o grupo de interés es el de menor tamaño.
Entre las técnicas de clasificación, correspondiente al enfoque clásico estadístico y el de minería de datos respectivamente, se pueden citar: Regresión Logística y Árboles de clasificación.
En este trabajo se propone el estudio de estas dos técnicas estadísticas multivariadas de clasificación siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos simulados bajo distintas situaciones que difieren en la estructura de correlaciones entre las variables intervinientes y en el desbalanceo o no de los grupos.

2. Metodología

2.1. Simulación de los datos

Se generaron mediante simulación 500 archivos de datos de 150 filas (unidades) y 6 columnas (variables) bajo distintas condiciones o escenarios. La simulación se realizó a partir de distribuciones normales estandarizadas multivariadas con matriz de correlaciones según cuatro estructuras diferentes. Se consideró la primer columna (X1) como la variable respuesta y las restantes variables (X2 a X6) como las variables predictoras o explicativas. Luego de la generación de los ficheros se transformó la variable respuesta (X1) para obtener una variable dicotómica utilizando:

a) La mediana de la distribución teórica (dos grupos balanceados)
b) el cuartil 3 de la distribución teórica (un grupo con el 25% de las observaciones y el otro con el 75% restante)

La variable respuesta siempre se la consideró transformada a categórica ya que el objetivo de este estudio es evaluar las técnicas encargadas de clasificación de unidades.

En este estudio, para evaluar el desempaño de las técnicas de clasificación en situaciones de tener dos grupos, se definieron las siguientes modalidades o condiciones sobre las cuales se evalúan los desempeños de las técnicas.

1- Escenario 1: Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$).
2- Escenario 2: Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las variables predictoras muy correlacionadas entre sí ($0.49<r<0.84$).
3- Escenario 3: Variable respuesta muy correlacionada con las predictoras ($0.36<r<0.83$) y las variables predictoras también muy correlacionadas entre sí ($0.43<r<0.87$).
4- Escenario 4: Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí ($r<0.06$).

De esta manera quedan definidas 4 bases de datos, correspondientes a cada una de estas situaciones. Cada una de estas bases exhibe 500 muestras de 150 unidades sobre las cuales se reconoce una variable respuesta dicotómica y 5 variables explicativas cuantitativas continuas con distribución Normal multivariada.

De esta manera se definen 8 bases de datos, correspondientes a cada una de estas situaciones, que presentan una variable respuesta dicotómica (para el caso a) y b) correspondiente a grupos balanceados y desbalanceados respectivamente) y 5 variables explicativas continuas. Resultan entonces, de la combinación de los escenarios 1, 2, 3 y 4 con las modalidades de la respuesta (a) y (b), las siguientes bases (BASE nº de escenario|modalidad de X1): BASE 1a - BASE 1b – BASE 2a – BASE 2b- BASE 3a – BASE 3b – BASE 4a- BASE 4b

Sobre las bases simuladas detalladas recientemente se comparan dos de las técnicas multivariadas de clasificación más utilizadas: ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN LOGÍSTICA. Por este motivo, para cada muestra, se simularon datos extras o suplementarios (30 filas para cada muestra) para ser considerados en la evaluación de la clasificación sin haberlos utilizados en los procesos de estimación (grupo de prueba).

El proceso de simulación de ficheros de datos como las aplicaciones estadísticas subsiguientes se lleva a cabo en el software R version 3.4.0.

2.2. Técnicas de clasificación

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoques agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.

Referido al primer enfoque, una de las técnicas más utilizadas es la Regresión Logística. La regresión logística estima la probabilidad de un suceso en función de un conjunto de

variables explicativas. Este modelo expresa matemáticamente la probabilidad de pertenencia a uno de los grupos, de manera que es posible calcularlas y asignar cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor. Otra técnica aplicada frecuentemente son los Árboles de Clasificación que crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.

2.2.1. Regresión logística

La Regresión Logística (RL) es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.

Sea **x** un vector de p variables independientes, esto es, **x´**= (x_1, x_2,…, x_p). La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y=1/X)=\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

donde $g(x)=\beta_0+\beta_1x_1+\ldots+\beta_px_p$

β_0 es la constante del modelo o término independiente

p el número de covariables

βi los coeficientes de las covariables

xi las covariables que forman parte del modelo.

Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 "variables de diseño" o "variables dummy". El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)}=e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1|\,X)}{1-P(y=1|\,X)}\right)=\log(e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p})=\beta_0+\beta_1x_1+\ldots+\beta_px_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.

Los coeficientes del modelo se estiman por el método de Máxima Verosimilitud y la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede verificar utilizando, entre otros, el test de Wald y el test de razón de verosimilitudes.

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en el modelo contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos. Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. En este trabajo se ajustó un modelo con todas los predictores sin realizar una selección. Sin embargo se evaluó la significación del aporte de cada una al modelo. El desempeño del modelo se valoró mediante el porcentaje de clasificación correcta calculado sobre un conjunto de datos (datos de prueba) no utilizado para la estimación del mismo.

2.2.2. Árboles de Clasificación

Los árboles de clasificación (AC) se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. En esta aplicación, los AC se emplean para asignar unidades a cada uno de los dos grupos definidos por la variable respuesta. Es un algoritmo que genera un árbol en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten utilizarlo para clasificar nuevas unidades.

Es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. Comienza el algoritmo con un nodo inicial o raíz a partir del cual se divide en dos sub-grupos o sub-nodos, esto es, se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. Luego se aplica el mismo procedimiento de partición a cada sub-nodo por separado. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes homogéneas.

Las divisiones sucesivas se determinan de modo que la heterogeneidad o impureza de los sub-nodos sea menor que la del nodo de la etapa anterior a partir del cual se forman. El objetivo es particionar la respuesta en grupos homogéneos manteniendo el árbol lo más pequeño posible.

La función de impureza es una medida que permite determinar la calidad de un nodo, esta se denota por i(t). Si bien existen varias medidas de impureza utilizadas como criterios de partición, una de las más utilizadas está relacionada con el concepto de entropía

$$i(t) = \sum_{j=1}^{K} p(j/t).\ln\ p(j/t)$$

donde j = 1, ..., k es el número de clases de la variable respuesta categórica y p(j|t) la probabilidad de clasificación correcta para la clase j en el nodo t. El objetivo es buscar la partición que maximiza

$$\Delta i(t) = -\sum_{j=1}^{K} p(j/t).\ln\ p(j/t).$$

De todos los árboles que se pueden definir es necesario elegir el óptimo. El árbol óptimo será aquel árbol de menor complejidad cuya tasa de mala clasificación (TMC) es mínima. En este trabajo se registraron cuáles variables fueron elegidas por el método de construcción del árbol final, dado que no todas fueron siempre necesarias.

El desempeño del árbol se comparó mediante el porcentaje de clasificación correcta calculado sobre un conjunto de observaciones no utilizado para la construcción del mismo (datos de prueba).

3. Resultados

3.1. Descripción de los conjuntos de datos simulados

Cada una de las bases de datos detalladas a continuación contiene 500 muestras de 150 filas cada una:

- Base 1a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.
- Base 1b: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta contienen el 25% y el 75% de las unidades respectivamente.

- Base 2a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las variables predictoras muy correlacionadas entre sí ($0.49<r<0.84$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.
- Base 2b: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las

variables predictoras muy correlacionadas entre sí (0.49<r<0.84). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta contienen el 25% y el 75% de las unidades respectivamente.

- Base 3a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta muy correlacionada con las predictoras (0.36<r<0.83) y las variables predictoras también muy correlacionadas entre sí (0.43<r<0.87). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.
- Base 3a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta muy correlacionada con las predictoras (0.36<r<0.83) y las variables predictoras también muy correlacionadas entre sí (0.43<r<0.87 Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta contienen el 25% y el 75% de las unidades respectivamente.

- Base 4a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí (r<0.06). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.
- Base 4a: Respuesta dicotómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí (r<0.06). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta contienen el 25% y el 75% de las unidades respectivamente.

Para explorar las bases de datos simuladas se aplica el test no paramétrico de Wilcoxon para muestras independientes, en cada conjunto de datos simulado, para llevar a cabo la comparación de los grupos definidos por la variable respuesta binaria respecto a cada variable explicativa. Esto se realiza para cada una de las muestras contenidas en cada escenario, hallando los siguientes resultados:

Base 1a: Estos datos fueron simulados a partir de una población donde la variable respuesta se encuentra asociada a las explicativas pero entre ellas no. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte la mediana de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 11.8% (295/2500) de las comparaciones no se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 43 casos (8.6%) para X3, 5 casos para X4 (1%), 141 (28%) para X5 y 30 (6%) casos para X6, mientras que no se encontraron casos no significativos para X2.

Base 1b: Estos datos fueron simulados a partir de una población donde la variable respuesta se encuentra asociada a las explicativas pero entre ellas no. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte el cuartil 3 de la variable X1. En la comparación de los grupos respecto a cada variable explicativa

se observa que en el 8.8% (219/2500) de las comparaciones no se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 55 casos (11%) para X3, 13 casos para X4 (3%), 173 (35%) para X5 y 54 (11%) casos para X6, mientras que no se encontraron casos no significativos para X2.

Base 2a: Estos datos surgen de simular muestras de poblaciones donde las explicativas no están relacionadas con la respuesta pero sí entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte la mediana de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 6% (150/2500) de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 19 casos (4%) para X2, 35 casos para X3 (7%), 24 casos para X4 (5%), 31 (6%) para X5 y 41 (8%) casos para X6.

Base 2b: Estos datos surgen de simular muestras de poblaciones donde las explicativas no están relacionadas con la respuesta pero sí entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte el cuartil 3 de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 6% (162/2500) de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 24 casos (5%) para X2, 39 casos para X3 (8%), 25 casos para X4 (5%), 42 (8%) para X5 y 32 (6%) casos para X6.

Base 3a: Estos datos surgen de simular muestras de poblaciones donde las explicativas están muy correlacionadas con la respuesta y también entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte la mediana de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 1% (36/2500) de las comparaciones no se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 6 casos (1%) para X3, 30 casos para X6 (6%) y para el resto de las variables las comparaciones resultaron significativas para todas las muestras.

Base 3b: Estos datos surgen de simular muestras de poblaciones donde las explicativas están muy correlacionadas con la respuesta y también entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte el cuartil 3 de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 3% (65/2500) de las comparaciones no se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 11 casos (2%) para X3, 1 caso (0.2%) para X4, 53 casos para X6 (11%) y para el resto de las variables las comparaciones resultaron significativas para todas las muestras.

Base 4a: Estos datos surgen de simular muestras de poblaciones donde las explicativas no están relacionadas con la respuesta y tampoco entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte la mediana de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 8% (193/2500) de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 35 casos (7%) para X2,

30 casos para X3 (6%), 33 casos para X4 (7%), 50 (10%) para X5 y 45 (9%) casos para X6.

Base 4b: Estos datos surgen de simular muestras de poblaciones donde las explicativas no están relacionadas con la respuesta y tampoco entre ellas. La variable respuesta define dos grupos creados a partir de utilizar como punto de corte el cuartil 3 de la variable X1. En la comparación de los grupos respecto a cada variable explicativa se observa que en el 7% (168/2500) de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 33 casos (7%) para X2, 30 casos para X3 (6%), 28 casos para X4 (6%), 42 (8%) para X5 y 35 (7%) casos para X6.

El análisis anterior sugiere que las técnicas de clasificación deberían mostrar sin inconvenientes un buen desempeño en las muestras correspondientes a las bases de los escenarios 1 y 3 ya que los grupos evidencian diferencias marcadas respectos a las variables utilizadas para la discriminación. No se esperaría lo mismo en datos provenientes de los escenarios 2 y 4 en los que los grupos parecen no mostrar discrepancias. Por este motivo en estos casos es interesante evaluar cómo se diferencian los resultados obtenidos en la clasificación con RL y AC, evaluando también el deterioro en la clasificación en el grupo minoritario en el caso desbalanceado.

3.2. Aplicación de técnicas de clasificación

3.2.1. Regresión Logística

Se ajustó un modelo de regresión logística para variable respuesta dicotómica y 5 variables explicativas continuas, para cada una de las 500 muestras en cada uno de los 4 escenarios evaluados y considerando los dos casos de variable respuesta: grupos aproximadamente balanceados (a) y no (b). Los resultados hallados en cada caso se presentan a continuación.

Escenario 1a: En promedio, en las 500 muestras se observó en promedio un 84% de clasificación correcta, siendo el mínimo observado de 75% y el máximo de 93%.

Escenario 1b: En promedio, en las 500 muestras se observó en promedio un 89% de clasificación correcta, siendo el mínimo observado de 81% y el máximo de 97%.

Escenario 2a: En promedio, en las 500 muestras se observó en promedio un 58% de clasificación correcta, siendo el mínimo observado de 49% y el máximo de 67%.

Escenario 2b: En promedio, en las 500 muestras se observó en promedio un 75% de clasificación correcta, siendo el mínimo observado de 62% y el máximo de 85%.

Escenario 3a: En promedio, en las 500 muestras se observó en promedio un 86% de clasificación correcta, siendo el mínimo observado de 77% y el máximo de 96%.

Escenario 3b: En promedio, en las 500 muestras se observó en promedio un 90% de clasificación correcta, siendo el mínimo observado de 81% y el máximo de 97%.

Escenario 4a: En promedio, en las 500 muestras se observó en promedio un 58% de clasificación correcta, siendo el mínimo observado de 46% y el máximo de 70%.

Escenario 4b: En promedio, en las 500 muestras se observó en promedio un 75% de clasificación correcta, siendo el mínimo observado de 65% y el máximo de 85%.

3.2.2. Árboles de Clasificación

Se aplicó la técnica de AC (Árboles de Clasificación) para variable respuesta dicotómica y 5 variables explicativas continuas, para cada una de las 500 muestras en cada uno de los 4 escenarios evaluados y considerando los dos casos de variable respuesta: grupos aproximadamente balanceados (a) y no (b). Los resultados se presentan a continuación para cada caso.

Escenario 1a: En promedio, en las 500 muestras se observó en promedio un 88% de clasificación correcta, siendo el mínimo observado de 79% y el máximo de 99%.

Escenario 1b: En promedio, en las 500 muestras se observó en promedio un 90% de clasificación correcta, siendo el mínimo observado de 84% y el máximo de 95%.

Escenario 2a: En promedio, en las 500 muestras se observó en promedio un 79% de clasificación correcta, siendo el mínimo observado de 71% y el máximo de 87%.

Escenario 2b: En promedio, en las 500 muestras se observó en promedio un 83% de clasificación correcta, siendo el mínimo observado de 75% y el máximo de 92%.

Escenario 3a: En promedio, en las 500 muestras se observó en promedio un 88% de clasificación correcta, siendo el mínimo observado de 82% y el máximo de 94%.

Escenario 3b: En promedio, en las 500 muestras se observó en promedio un 91% de clasificación correcta, siendo el mínimo observado de 85% y el máximo de 96%.

Escenario 4a: En promedio, en las 500 muestras se observó en promedio un 79% de clasificación correcta, siendo el mínimo observado de 72% y el máximo de 86%.

Escenario 4b: En promedio, en las 500 muestras se observó en promedio un 83% de clasificación correcta, siendo el mínimo observado de 76% y el máximo de 90%.

4. Comparación de los resultados hallados

Los porcentajes de clasificación correcta globales son superiores en los árboles de clasificación, para los 4 escenarios. La diferencia en el desempeño de las técnicas es más evidente en los escenarios en los que la respuesta se encuentra poco correlacionada con las variables explicativas. En los casos contrarios la diferencia es leve. Es decir, incluso en el escenario menos deseable los árboles se desempeñan mejor en la tarea de clasificar nuevas unidades.

Tabla 1: Porcentaje promedio de clasificación correcta según escenario, conformación de grupos y técnica estadística.

Escenario	**Caso balanceado**		**Caso desbalanceado**	
	Árboles de clasificación	**Regresión Logística**	**Árboles de clasificación**	**Regresión Logística**
1	88.0	84.5	89.5	88.5
2	79.2	58.2	83.1	75.3
3	88.5	86.5	90.6	89.6
4	79.3	58.2	82.9	75.3

Tabla 2: Porcentaje promedio de clasificación correcta según escenario, técnica estadística y clase, en el caso desbalanceado.

Escenario	**Caso desbalanceado**			
	Árboles de clasificación		**Regresión Logística**	
	Clase minoritaria	**Clase mayoritaria**	**Clase minoritaria**	**Clase mayoritaria**
1	76.8	93.5	71.2	94.0
2	61.4	90.0	2.9	99.2
3	79.7	94.1	74.6	94.4
4	60.9	89.8	2.6	99.2

Cuando se analiza el caso de grupos desbalanceados, el porcentaje de clasificación correcta dentro de cada clase se muestra en la tabla 2. Como era de esperar, la clase mayoritaria presenta un porcentaje alto mientras que la clase minoritaria muestra un mal desempeño, siendo mayor la diferencia principalmente en los escenarios en los que la respuesta no se correlaciona con las variables explicativas del modelo. Comparando las técnicas, en este caso el modelo de regresión logística presenta una clasificación más acertada sólo en la clase mayoritaria, mientras que los árboles clasifican con un mejor desempeño en la clase minoritaria, respecto a la otra metodología.

5. Discusión

En este trabajo se ha evaluado el desempeño de estas dos técnicas en datos simulados bajo distintas condiciones que diferían en:

- la estructura de correlaciones entre la variable respuesta y las predictoras y entre las predictoras mismas.
- Conformación de la respuesta: grupos balanceados y desbalanceados.

Entre las similitudes y diferencias halladas se puede enunciar:

- En condiciones en que las variables predictoras están altamente correlacionadas con la respuesta, ambas metodologías funcionan satisfactoriamente. Sin embargo, la superioridad de los AC respecto al porcentaje de clasificación correcta resultó significativa.
- Si bien en esta aplicación no se puede evidenciar, por no ser datos correspondientes a una problemática real, los modelos de RL presentan la ventaja de la interpretación de los coeficientes estimados que permiten reflejar información valiosa contenida en los datos.
- En condiciones desfavorables para obtener una clasificación satisfactoria, predictores poco correlacionados con la respuesta, los AC logran un porcentaje de clasificación correcta notablemente superior a la RL, con la desventaja de obtener un árbol con numerosos nodos terminales utilizando la información de prácticamente todas las variables explicativas.
- En el caso desbalanceado, la clase mayoritaria presenta un porcentaje de clasificación correcta superior en la regresión logística a costa de un peor desempeño en la clase minoritaria. Este comportamiento es más marcado en esta técnica que en los árboles de clasificación.

6. Bibliografía

Agresti, A. 2002. Categorical Data Analysis. Wiley & Sons. New Jersey.

Beltrán, C. 2012 Introducción al análisis estadístico en la investigación. Con aplicaciones en distintas disciplinas". Ediciones Juglaría. Rosario.

Beltrán C, 2012. Árboles de clasificación y su comparación con análisis de regresión logística aplicado a la clasificación de textos académicos. Revista INFOSUR. Número 6.

Barbona I, 2015. Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos. Revista INFOSUR. Número 7.

Beltrán C,; Barbona, I. 2017. Una revisión de las técnicas de clasificación supervisada en la clasificación automática de textos. Revista de Epistemología y Ciencias Humanas. Nro. Número 9.

Catena, A.; Ramos, M.M; Trujillo, H.M. 2003. Análisis multivariado. Un manual para investigadores. Biblioteca Nueva S.L. España.

Cuadras, C.M. 2014 Nuevos métodos de análisis multivariante. CMC Editions. Barcelona, España.

Hair, J.F., Anderson, R.L.,Tatham, R.L., Black, W.C. 1999. Análisis Multivariante. Prentice Hall Iberia, Madrid, España.

Hosmer, D.; Lemeshow, S. 1989. Applied Logistic Regression. Jhon Wiley & Sons. New York.

Johnson, D.E. 2000. Métodos multivariados aplicados al análisis de datos. Internacional Thomson Editores.

Flórez López, R.; Fernández Fernández, J.M. 2008. Las redes neuronales artificiales. Fundamentos teóricos y aplicaciones prácticas. Netbiblio S.L. España.

Johnson R.A. y Wichern D.W. 1992 Applied Multivariate Statistical Analysis. Prentice-Hall International Inc.

Khattree R., Naik D. (2000). Multivariate Data Reduction and Discrimination with SAS® Soft-ware. Cary, NC: SAS Institute Inc.

Pérez López, C. 2004. Técnicas de Análisis Multivariante de Datos. PEARSON EDUCACIÓN, S.A., Madrid, España.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.