Comparación del desempeño de técnicas multivariadas de clasificación en datos simulados bajo distintos escenarios: Regresión Logística y Árboles de Clasificación.

Celina Beltrán; Ivana Barbona

Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina

beltranc@dat1.net.ar

Abstract

This research proposes the study, evaluation and comparison of two multivariate statistical classification techniques, Logistic Regression and Classification Trees, being of interest to evaluate the performance based data simulated under different conditions that differed in the structure of correlations between the variables. 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. 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, with the disadvantage of obtaining a tree with numerous terminal nodes using the information from virtually all explanatory variables.

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. 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.
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, con la desventaja de obtener un árbol con numerosos nodos terminales utilizando la información de prácticamente todas las variables explicativas.

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.
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 o condiciones que difieren en la estructura de correlaciones entre las variables intervinientes.

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 la mediana de la distribución teórica. 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.

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 consideraron otras 150 filas para ser utilizadas en la evaluación de la clasificación sin haber intervenido en los procesos de estimación (grupo de prueba de igual tamaño al de entrenamiento).

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, y se asume lineal en las variables continuas incluidas en el modelo.

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 E1: 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 ($r<0.06$).
- Base E2: 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 ($0.49<r<0.84$).
- Base E3: 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 ($0.43<r<0.87$).
- Base E4: 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 ($r<0.06$).

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 E1: Estos datos fueron simulados bajo un escenario "ideal" para la tarea de clasificación en la que se tiene una población donde la variable respuesta se encuentra asociada a las explicativas pero entre ellas no. Se realizaron comparaciones univariadas de los grupos respecto a cada variable explicativa observándose que en el 91% de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró un 91% para X3, 99% para X4, 72% para X5 y 94% para X6, mientras que para X2 las comparaciones resultaron significativas en su totalidad. Esto concuerda con la matriz de correlaciones de la que fueron simulados los datos donde la variable X5 era la de menor magnitud.

Base E2: Estos datos surgen de simular muestras de una población donde las explicativas no están relacionadas con la respuesta pero sí entre ellas. En la comparación de los grupos respecto a cada variable explicativa, como se esperaba, se observa que sólo en el 6% de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró un 4% para X2, 7% para X3, 5% para X4, 6% para X5 y 8% para X6.

Base E3: Estas muestras simuladas provienen de una población donde las explicativas están muy correlacionadas con la respuesta y también entre ellas. En la comparación univariada de los grupos respecto a cada variable explicativa se observa que en el 99% de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró 99% de casos para X3, 94% para X6 y para el resto de las variables las comparaciones resultaron significativas para todas las muestras evaluadas. Esto está vinculado con el hecho que X3 y X6 son las variables con correlaciones de menor magnitud con la variable respuesta en la matriz poblacional utilizada para la simulación.

Base E4: Estos datos provienen de simular a partir de una población en la cual las variables explicativas no están relacionadas con la variable respuesta y tampoco entre ellas. En la comparación de los grupos respecto a cada variable explicativa se observa una situación similar al del escenario 2. En este caso sólo en el 8% de las comparaciones se observan diferencias significativas al 5%. Esto discriminado por variable se encontró un 7% de las muestras para X2, 6% para X3, 7% para X4, 10% para X5 y 9% de los 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 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.

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. El mismo se construyó con los efectos principales sin incluir interacciones entre los predictores, para cada una de las 500 muestras en cada uno de los 4 escenarios considerados. Para evaluar el supuesto de linealidad entre el logit y los predictores continuos se acudió a gráficos de dispersión en los cuales se enfrenta cada uno de los predictores del modelo y el logit. En todos los casos los gráficos no mostraron evidencia de un alejamiento de este supuesto. Los resultados se presentan por separado para cada escenario.

3.2.1.1.RL Escenario 1

En promedio, en las 500 muestras se observó un 85% de clasificación correcta, siendo el mínimo observado de 75% y el máximo de 92%. Respecto a la significación de las covariables en el modelo, la tabla 1 muestra los casos en los que resultó un efecto significativo para cada variable en los 500 modelos.

Tabla 1: Porcentaje de casos significativos para cada variable explicativa del modelo en las 500 muestras correspondientes al escenario 1.

Escenario	Variable	% de casos significativos en las 500 muestras
E1	X2	100
	X3	85
	X4	99
	X5	78
	X6	94

3.2.1.2.RL Escenario 2

En promedio, en las 500 muestras se observó un 61% de clasificación correcta, siendo el mínimo observado de 50% y el máximo de 69%. Respecto a la significación de las covariables en el modelo, la tabla 2 muestra los casos en los que resultó un efecto significativo para cada variable en los 500 modelos.

Tabla 2: Porcentaje de casos significativos para cada variable explicativa del modelo en las 500 muestras correspondientes al escenario 2.

Escenario	Variable	% de casos significativos en las 500 muestras
E2	X2	6
	X3	8
	X4	10
	X5	11
	X6	10

3.2.1.3.RL Escenario 3

En promedio, en las 500 muestras se observó un 86% de clasificación correcta, siendo el mínimo observado de 78% y el máximo de 96%. Respecto a la significación de las covariables en el modelo, la tabla 3 muestra los casos en los que resultó un efecto significativo para cada variable en los 500 modelos.

Tabla 3: Porcentaje de casos significativos para cada variable explicativa del modelo en las 500 muestras correspondientes al escenario 3.

Escenario	Variable	% de casos significativos en las 500 muestras
E3	X2	100
	X3	68
	X4	99
	X5	40
	X6	99

3.2.1.4.RL Escenario 4

En promedio, en las 500 muestras se observó un 59% de clasificación correcta, siendo el mínimo observado de 45% y el máximo de 70%. Respecto a la significación de las covariables en el modelo, la tabla 4 muestra los casos en los que resultó un efecto significativo para cada variable en los 500 modelos.

Tabla 4: Porcentaje de casos significativos para cada variable explicativa del modelo en las 500 muestras correspondientes al escenario 4.

Escenario	Variable	% de casos significativos en las 500 muestras
E4	X2	7
	X3	6
	X4	6
	X5	8
	X6	7

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. Los resultados se presentan para cada escenario.

3.2.2.1.AC Escenario 1

En promedio, en las 500 muestras se observó un 84% de clasificación correcta, siendo el mínimo observado de 77% y el máximo de 98%. Respecto a la significación de las covariables en el modelo, la tabla 5 muestra el porcentaje de casos en los que cada variable necesitó ser considerada por el árbol en los 500 ajustes.

Tabla 5: Porcentaje de casos en los que se requiere cada variable en la construcción del árbol en las 500 muestras correspondientes al escenario 1.

Escenario	Variable	% de casos que es usada la variable en los 500 árboles
E1	X2	100
	X3	88
	X4	98
	X5	71
	X6	90

3.2.2.2.AC Escenario 2

En promedio, en las 500 muestras se observó un 76% de clasificación correcta, siendo el mínimo observado de 70% y el máximo de 85%. Respecto a la significación de las covariables en el modelo, la tabla 6 muestra el porcentaje de casos en los que cada variable necesitó ser considerada por el árbol en los 500 ajustes.

Tabla 6: Porcentaje de casos en los que se requiere cada variable en la construcción del árbol en las 500 muestras correspondientes al escenario 2.

Escenario	Variable	% de casos que es usada la variable en los 500 árboles
E2	X2	96
	X3	93
	X4	91
	X5	92
	X6	93

3.2.2.3.AC Escenario 3

En promedio, en las 500 muestras se observó un 87% de clasificación correcta, siendo el mínimo observado de 79% y el máximo de 95%. Respecto a la significación de las covariables en el modelo, la tabla 7 muestra el porcentaje de casos en los que cada variable necesitó ser considerada por el árbol en los 500 ajustes.

Tabla 7: Porcentaje de casos en los que se requiere cada variable en la construcción del árbol en las 500 muestras correspondientes al escenario 3.

Escenario	Variable	% de casos que es usada la variable en los 500 árboles
E3	X2	100
	X3	49
	X4	71
	X5	61
	X6	86

3.2.2.4.AC Escenario 4

En promedio, en las 500 muestras se observó un 67% de clasificación correcta, siendo el mínimo observado de 59% y el máximo de 76%. Respecto a la significación de las covariables en el modelo, la tabla 8 muestra el porcentaje de casos en los que cada variable necesitó ser considerada por el árbol en los 500 ajustes.

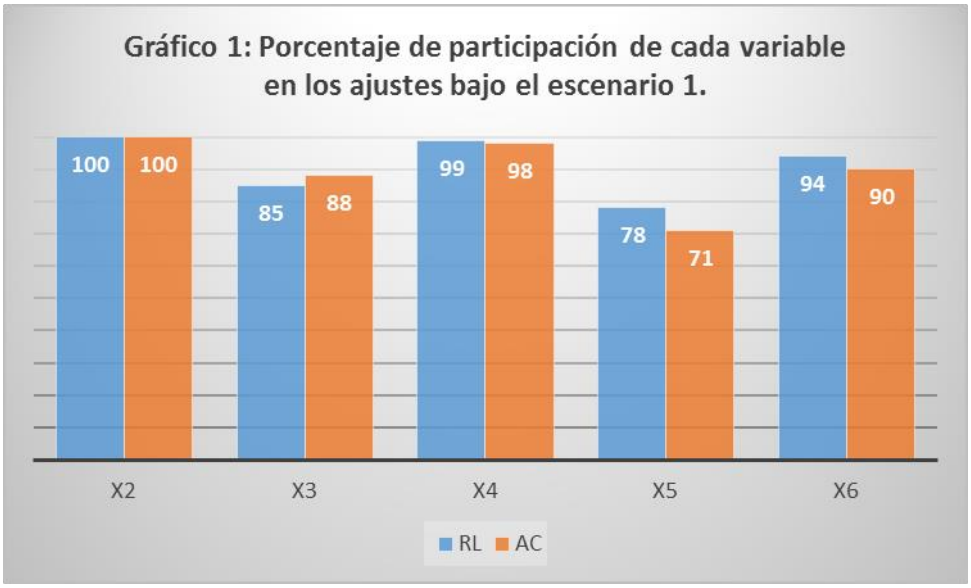
Tabla 8: Porcentaje de casos en los que se requiere cada variable en la construcción del árbol en las 500 muestras correspondientes al escenario 4.

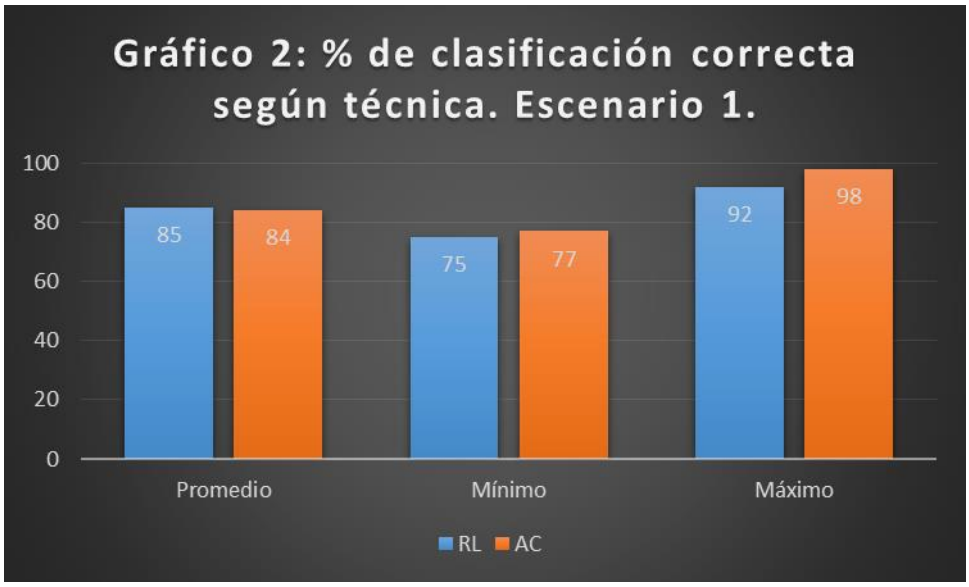
Escenario	Variable	% de casos que es usada la variable en los 500 árboles
E4	X2	95
	X3	94
	X4	92
	X5	91
	X6	92

4. Comparación de los resultados hallados

4.1. Escenario 1

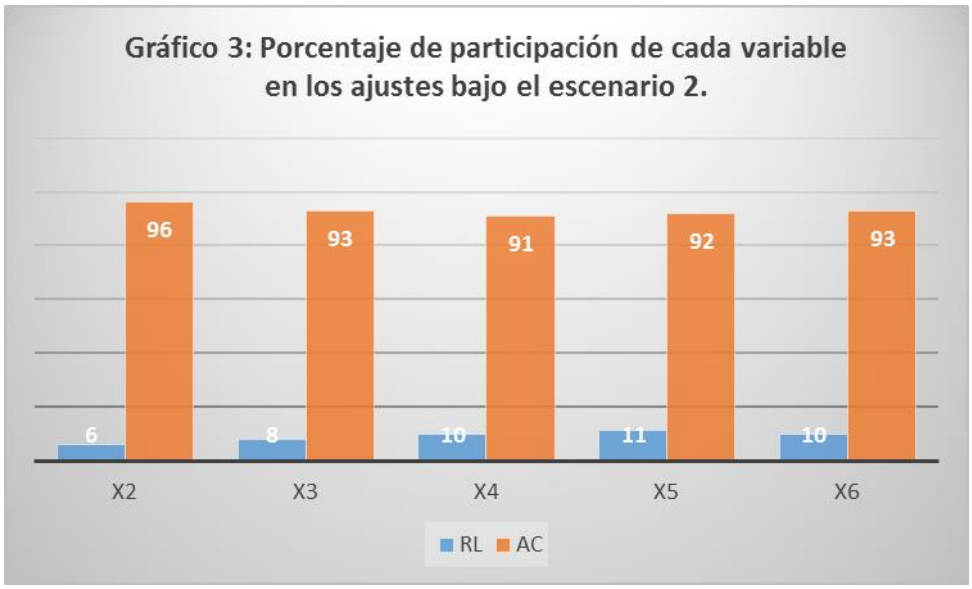
En este caso se observan similares participaciones de las variables en las dos técnicas evaluadas (Gráfico 1). Si nos enfocamos en el porcentaje de clasificación correcta se observa una superioridad a favor del RL (Gráfico 2) tanto para el promedio como los valores más extremos. Esta superioridad no parece ser de magnitud importante, el resultado del test de Wilcoxon para muestras apareadas utilizado para comparar los porcentajes promedios no revela diferencias significativas (p=0.171).

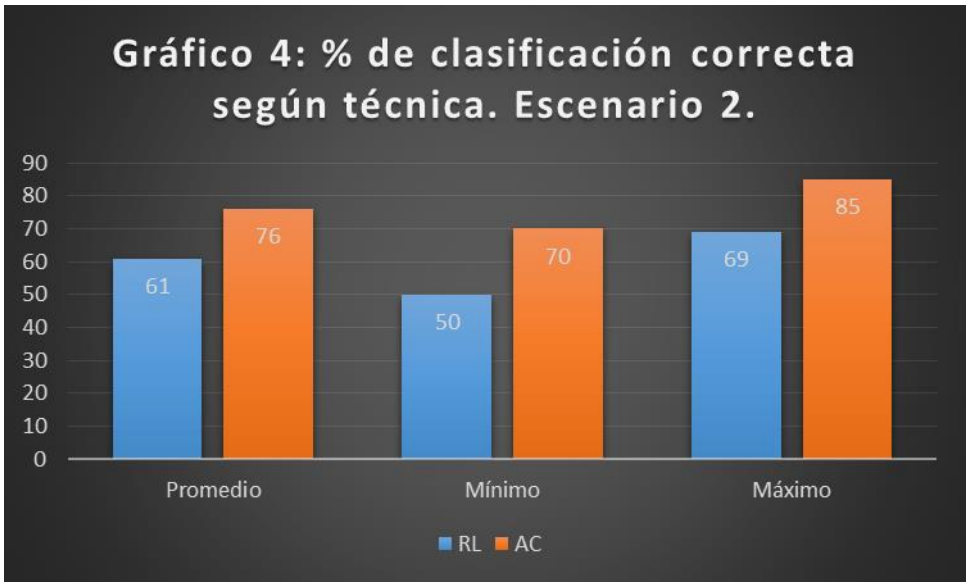




4.2. Escenario 2

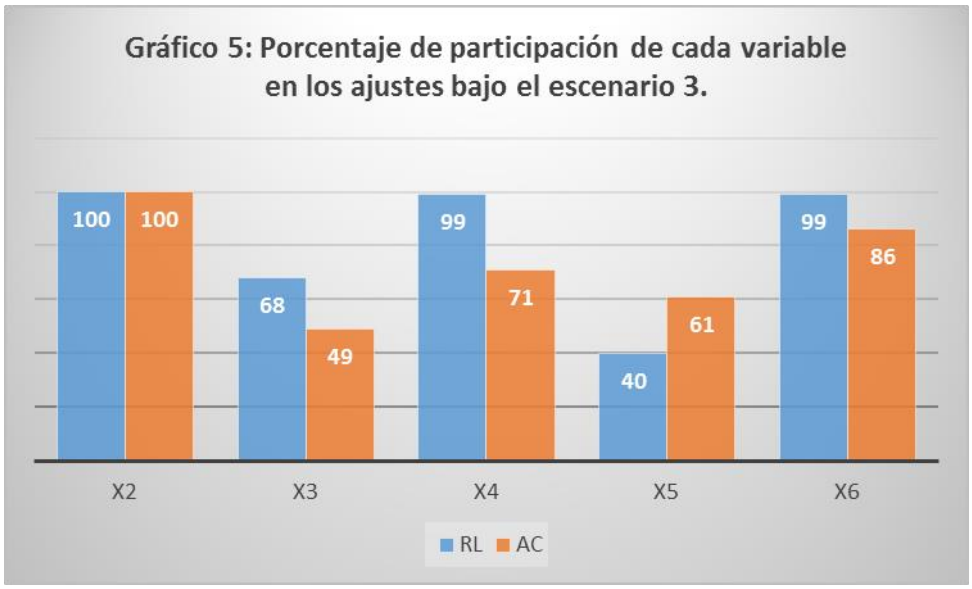
En las condiciones del escenario dos, donde las variables explicativas no presentan asociación con la respuesta pero sí entre ellas, menos del 10% de los modelos de RL presentan significación en las variables. Esto concuerda con los estudios descriptivos de los datos simulados. Esta situación se ve reflejada en los AC donde se observa que se requieren casi la totalidad de las variables para la construcción del árbol (Gráfico 3). El procedimiento prácticamente decide no excluir a ninguna variable dando a lugar árboles con muchos nodos terminales. Sin embargo, al analizar el porcentaje de clasificación correcta se evidencia una notable superioridad de los AC frente a los modelos de RL, tanto para el promedio como para todo el rango de variación (Gráfico 4). Esto se corresponde con el resultado del test de Wilcoxon para muestras apareadas utilizado para comparar los porcentajes promedios ($p<0.0001$) revelando diferencias significativas en el desempeño para clasificar correctamente.

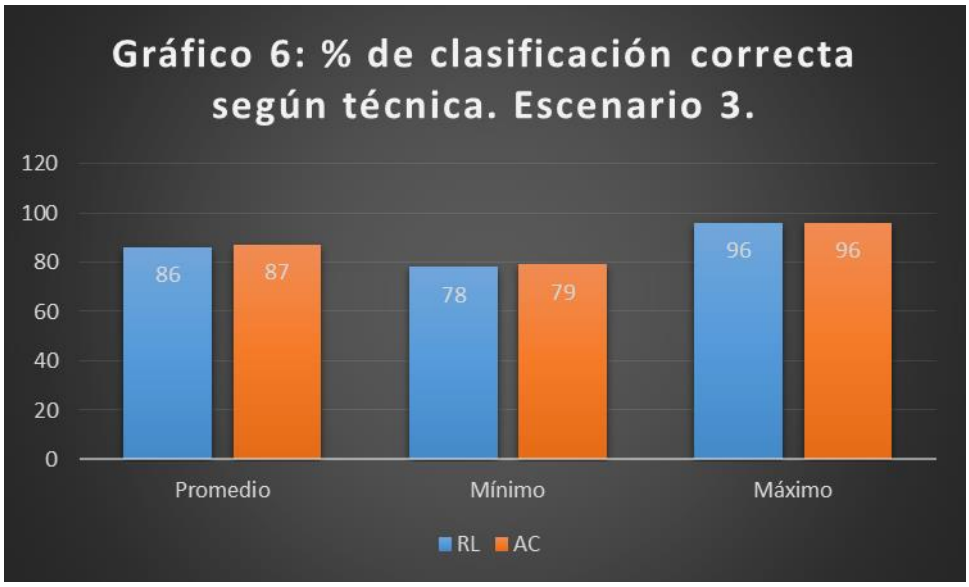




4.3. Escenario 3

En este caso se observan similares participaciones de las variables en las dos técnicas evaluadas (Gráfico 5). Si nos enfocamos en el porcentaje de clasificación correcta se observa una leve superioridad a favor del AC (Gráfico 6) para el promedio y el mínimo, siendo el valor máximo del porcentaje de clasificación correcta alcanzado con un modelo de RL. Esta superioridad no resulta significativa según el resultado del test de Wilcoxon para muestras apareadas utilizado para comparar los porcentajes promedios (p=0.139).

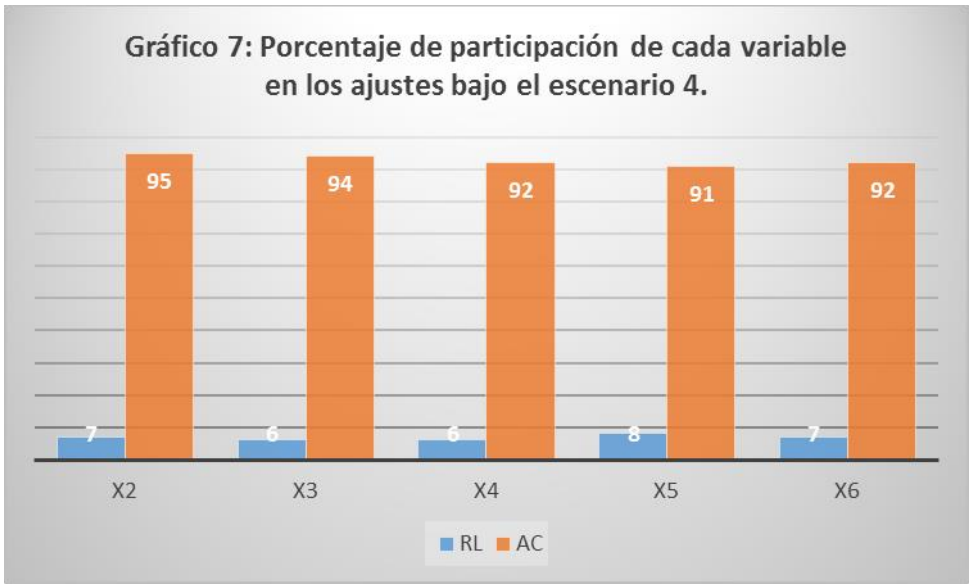


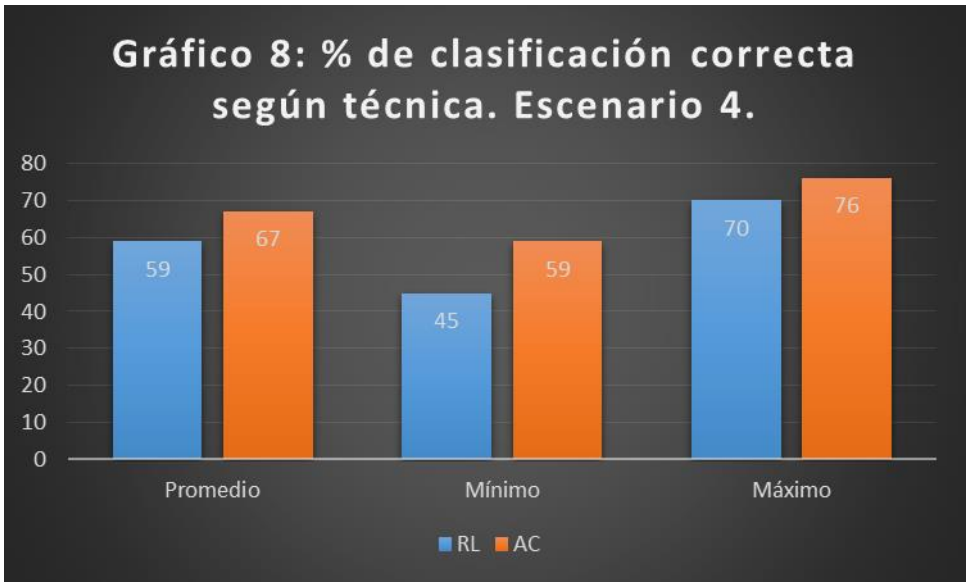


4.4. Escenario 4

En las condiciones del escenario 4, donde las variables explicativas no presentan asociación con la respuesta ni entre ellas, menos del 8% de los modelos de RL presentan significación en las variables. Esto concuerda con los estudios descriptivos de los datos simulados. Esta situación, al igual que en la aplicación sobre el escenario 2, se ve reflejada en los AC donde se observa que se requieren casi la totalidad de las variables para la construcción del árbol (Gráfico 7). Similarmente a lo presentado en el apartado 4.2., al analizar el porcentaje de clasificación correcta se evidencia una superioridad de los AC frente a los modelos de RL, tanto para el promedio como para todo el rango de variación (Gráfico 8). Esto se corresponde con el resultado del test de Wilcoxon para muestras apareadas utilizado para comparar los porcentajes promedios

($p<0.0001$) revelando diferencias significativas en el desempeño para clasificar correctamente.





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.

Entre las similitudes y diferencias halladas se puede enunciar algunas generales y otras particulares detectadas en este estudio:

- Ambas metodologías requieren para contar con datos históricos de casos en los que se conozca el grupo de pertenencia (variable respuesta) y los valores de cada una de los predictores.
- Los modelos de RL usan una expresión algebraica cuyos coeficientes son estimados con la información histórica, mientras que los AC usan dicha información para construir reglas secuencialmente.
- Con ambas metodologías puedo clasificar nuevos casos.
- Con el modelo de RL se calcula primero la probabilidad de pertenencia de la observación nueva a cada uno de los grupos y se asigna al grupo de mayor probabilidad; mientras que con los AC asignación se realiza siguiendo las reglas sucesivas directamente. Esta decisión de utilizar el punto de corte 0.50 para definir el grupo de pertenencia en RL se basó en la forma de generar la variable respuesta.
- En condiciones en que las variables predictoras están altamente correlacionadas con la respuesta, ambas metodologías funcionan satisfactoriamente. No se hallaron diferencias significativas respecto al porcentaje promedio de clasificación correcta con ambas técnicas.
- En condiciones desfavorables para obtener una clasificación satisfactoria, predictores poco correlacionados con la respuesta, los AC logran un porcentaje de clasificación correcta superior a la RL, con la desventaja de obtener un árbol con numerosos nodos terminales utilizando la información de prácticamente todas las variables explicativas.
- Considerar sólo efectos principales en el modelo de RL crea una clara desventaja frente a los árboles de clasificación quienes recogen el comportamiento no aditivo de las variables lo que genera una incorporación automática de interacciones.
- 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.

No se ha evaluado o comparado resultados en aplicaciones con distintos tamaños de muestras ni tampoco con presencia de datos faltantes, dejando planteado esta problemática para un futuro trabajo. Asimismo es importante destacar que, si bien en este trabajo se ha comparado la técnica de RL con AC, existen técnicas superadoras a los AC como por ejemplo bosques aleatorios (random forest). Estos algoritmos pueden obtener de 100 a 500 árboles y combinar sus resultados en un modelo. La decisión final en el conjunto de los árboles será la decisión que se tome en la mayor parte de los árboles constituyentes del grupo. Los bosques aleatorios, a diferencia de los AC, tienden a mostrarse resistentes al ruido y no sufren problemas de sobreajuste por aumentar el número de árboles creados en el proceso.

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/.