Comparación de dos técnicas de clasificación supervisada en la categorización de textos y evaluación en datos simulados: Árboles de clasificación y Regresión Logística

Celina Beltrán; Ivana Barbona
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
cbeltran2510@gmail.com

Abstract

This work compares the performance of two classification methods: Classification Trees (CT) and Logistic Regression (LR). This comparison is made 1) on a text categorization application and 2) an evaluation on simulated data under different scenarios.
1) For both methods, the functionality and performance in the texts classification is evaluated, describing how it is possible to use them to categorize and eventually characterize the texts. In this case, the classification criterion is the genre to which the text belongs (Scientific / Non-Scientific). The characterization of the texts is based on the frequency distribution of the morpho-syntactic categories. The texts were classified taking into account simultaneously the measurements made on them. The performance of the techniques was measured with the misclassification rate (MCR) calculated on a sample of texts not included in the estimation of the model and construction of the tree. The classification tree presented a MCR lower than that of the logistic model, managing to classify scientific texts more precisely. For CT the MCR was 4% for scientific texts and 28% for non-scientific texts. For the LR model, the MCR was 14% for scientific texts and 26% for non-scientific texts.
2) In the simulation study, it was observed as a main result, that in conditions where the predictor variables are highly correlated with the response, although the CT showed a significantly lower error rate in the classification, both methodologies work satisfactorily. However, when the conditions for obtaining a satisfactory classification are unfavorable (predictors with little correlation with the response), the CTs achieve a percentage of correct classification that is significantly higher than LR, with the disadvantage of obtaining a tree with numerous terminal nodes using the information from practically all the explanatory variables. In the unbalanced case, the majority class presents a higher correct classification percentage in logistic regression at the cost of worse performance in the minority class. This behavior is more marked in this technique than in the classification trees.

Keywords: Supervised classification techniques; logistic regression; classification trees; simulation; classification of texts

Resumen

En este trabajo se compara el desempeño de dos métodos de clasificación: Árboles de Clasificación (AC) y Regresión Logística (RL). Dicha comparación se realiza 1) sobre una aplicación en categorización de textos y 2) una evaluación sobre datos simulados bajo distintos escenarios.

1) Para ambos métodos se evalúa la funcionalidad y desempeño en la clasificación de textos describiendo cómo es posible utilizarlos para categorizar y eventualmente caracterizar los textos. En este caso, el criterio de clasificación es el género al que pertenece el texto (Científico / No Científico). La caracterización de los textos está basada en la distribución de frecuencias de las categorías morfo-sintácticas. Los textos se clasificaron teniendo en cuenta simultáneamente las mediciones realizadas sobre ellos. El desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC) calculada sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC resultó 4% para los textos científicos y 28% para los textos no científicos. Para el modelo de RL la TMC resultó 14% para los textos científicos y 26% para los textos no científicos.

2) En el estudio por simulación, 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. 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 RL que en los AC.

Palabras clave: Métodos de clasificación supervisada; regresión logística; árboles de clasificación; simulación; clasificación de textos

1. INTRODUCCION

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 [1]). Estas variables pueden ser cuantitativas, continuas o discretas, o cualitativas, nominales u ordinales (Pérez López, C. 2004 [2]). Uno de los objetivos de dichas técnicas es la clasificación de unidades u objetos en grupos (Cuadras, 2014 [3]). 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 (inferencia).

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 (RL) y Árboles de clasificación (AC) (Beltrán, 2012 [4]).

En este trabajo se propone el estudio de estas dos técnicas estadísticas multivariadas siendo de interés evaluar el desempeño de las mismas. Este desempeño será considerado en una primera instancia sobre datos reales: clasificación de textos. Para los dos métodos en consideración se presentan su funcionalidad y desempeño describiendo cómo es posible utilizarlos para clasificar y eventualmente caracterizar textos de distintos géneros o disciplinas. El criterio de clasificación es el género al que pertenece el texto (Científico / No Científico). La caracterización de los mismos está basada en la distribución de frecuencias de las categorías morfo-sintácticas. En una segunda instancia de evaluación de las dos técnicas consideradas será sobre 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. En ambas aplicaciones, se considera como medida para la comparación entre métodos la tasa de mala clasificación (TMC), o bien la tasa de clasificación correcta (TCC), calculada sobre una muestra de textos no incluidos en el proceso de construcción de la regla de clasificación.

2. MATERIAL Y METODOS

2.1. Datos reales

2.1.1. Diseño muestral

El marco muestral para la selección de la muestra de los textos científicos está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas científicas, extraídos de internet pertenecientes a distintas disciplinas. Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR. Este corpus se construyó con noticias extraídas de las páginas web de periódicos argentinos (noticias de tipo general, no especializadas en español). La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado.

Luego de obtener las muestras de los estratos, fueron evaluadas y comparadas respecto al número medio de palabras por texto. Se requiere esta evaluación para evitar que la comparación entre los géneros se vea afectada por el tamaño de los textos.

La muestra final para este trabajo quedó conformada por 90 textos científicos y 60 No científicos, con 162 y 135 palabras promedio por texto respectivamente.

2.1.2. Diseño y desarrollo de la base de datos

En Beltrán (2009) [5] se presenta la implementación de un etiquetador morfológico de textos basado en la información resultante de los análisis bajo el software Smorph (analizador y generador morfosintáctico desarrollado por Salah Aït-Mokhtar en el Groupe de Recherche dans les Industries de la Langue, Universidad Blaise-.Pascal, Clermont II) y el módulo post-smorph que permite resolver ambigüedades que resultan del análisis de Smorph. Esta herramienta se utiliza para el análisis morfológico de los textos. La información resultante de dicho análisis se dispuso en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto y tantas columnas como ocurrencia+lema+valores. Luego, a partir de esta base de datos por palabra, se confeccionó la base de datos por documento que es analizada estadísticamente.

2.2. Datos Simulados

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 desempeño 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 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 las técnicas multivariadas de clasificación anunciadas: ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN LOGÍSTICA. Por este motivo, para cada muestra, se simularon datos extras o suplementarios 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 versión 3.4.0 [6] (James, 2013 [7]).

2.3. Metodología estadística

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 enfoquen 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 (técnicas supervisadas). 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 (técnicas no supervisadas).
Las técnicas evaluadas en este trabajo tienen por objetivo construir un sistema que permita clasificar unidades en una de las categorías definidas y conocidas previamente en función de las variables relevadas.
Luego de la aplicación de cada una de las técnicas definidas en este apartado se debe evaluar la calidad de los resultados, es decir el desempeño para clasificar mediante la validación del mismo. Esto se realiza particionando el conjunto de unidades en dos grupos. Uno es utilizado para la estimación del mismo (entrenamiento del sistema) y el segundo conforma el grupo de validación para la fase de prueba. Se consideró como medida para la comparación entre métodos el error de mala clasificación calculada sobre una muestra de textos no incluidos en el proceso de construcción de la regla de clasificación.
A continuación se presentan dos de las técnicas multivariadas que tienen por objetivo clasificar unidades en categorías definidas a priori que fueron evaluadas en diferentes aplicaciones por los autores.

2.3.1. Análisis de regresión logística

Esta técnica es un caso particular de los modelos lineales generalizados, modela la probabilidad de que una unidad experimental pertenezca a un grupo en particular considerando información medida o registrada en dicha unidad (Agresti, 2002 [8]).
La regresión logística 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, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las p covariables que definen el vector **x** es:

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

donde $g(x)=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$

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:

$$\frac{P(y=1|X)}{1-P(y=1|X)}=e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

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

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=\log(e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p})=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$$

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 la función de discriminante 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 utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow (Hosmer, 1989 [9]) y, dado que el modelo es utilizado para clasificar

unidades, se utilizó también la tasa de mala clasificación calculada sobre la muestra independiente excluida en la etapa de estimación.

2.3.2. Árboles de clasificación

Los árboles de clasificación son una técnica de análisis discriminante no paramétrica que permite predecir la asignación de unidades u objetos a grupos predefinidos en función de un conjunto de variables predictoras. Esto es, dada una variable respuesta categórica, los árboles crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.
Es un algoritmo que genera un árbol de decisión en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten continuar la clasificación. Estas particiones recursivas logran formar grupos homogéneos respecto a la variable respuesta. El árbol determinado puede ser utilizado 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 es mínima. Generalmente esta búsqueda se realiza comparando árboles anidados mediante validación cruzada. La validación cruzada consiste, en líneas generales, en sacar de la muestra de aprendizaje o entrenamiento una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto excluido es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como “datos nuevos”.

3. RESULTADOS

3.1. Desempeño en datos reales: Clasificación de textos

En Beltrán 2013 [10] se realizó un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. En dicho estudio se evidenció que existen diferencias significativas entre los corpus respecto al tamaño de los textos (número de palabras por texto). Esta situación llevó a realizar las sucesivas comparaciones sobre los porcentajes o proporciones de las categorías gramaticales, hallando diferencias significativas ($p<0.05$) para todas las categorías gramáticas excepto la proporción de clíticos y de verbos en los documentos analizados. Asimismo, en un análisis de componentes principales, se dispusieron los textos en el plano de proyección demostrando que los textos procedentes del corpus No Científico presentan un mayor número de adverbios, respecto a las restantes categorías, que los textos Científicos.

3.1.1. Análisis de Regresión Logística

Se realizó un análisis de regresión logística para obtener una regla de clasificación que permita asignar los textos en estas dos poblaciones, definidas por el género al que pertenecen (Científico / No científico), en base a la frecuencia de cada categoría gramatical en el texto.
Para determinar cuáles categorías gramaticales son las responsables de la discriminación se utilizaron los tres algoritmos de selección de variables. En todos los casos se llega al mismo resultado: las variables que logran diferenciar a los dos corpus de textos analizados son el número de adjetivos, adverbios, conjunciones copulativas, determinantes, nombres y preposiciones.

Tabla 1: Coeficientes del modelo de regresión logística

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.6868	0.5507	1.5554	0.2123
adjetivos	1	0.1694	0.0562	9.0777	0.0026
adverbios	1	-0.3106	0.0800	15.0862	0.0001
Conj. Cop.	1	0.2769	0.1073	6.6566	0.0099
determinantes	1	0.1216	0.0464	6.8795	0.0087
nombres	1	-0.1995	0.0464	18.5044	<.0001
preposiciones	1	0.1575	0.0544	8.3925	0.0038

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus definidos por el género.

$$P(\in Cien/X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j+\ldots+\beta_p x_p}}$$

$$P(\in No\ Cien/X) = \pi(x) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j+\ldots+\beta_p x_p}}$$

Con este criterio un texto es asignado al corpus cuya probabilidad es máxima.

La bondad del ajuste se evaluó mediante el test de Hosmer-Lemeshow y la tasa de error de clasificación. Con el modelo de regresión logística obtenido durante la selección de variables se obtuvo una tasa de error global, estimada sobre el corpus de prueba, del 20% y la probabilidad asociada en el test de bondad de ajuste es p=0.9696 evidenciando lo adecuado del modelo. La tabla 2 presenta la tasa de error para cada género.

Tabla 2: Tasa de mala clasificación

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	14%	26%

Tabla 3: Razones de odds estimadas

Razón de odds			
Efecto	**Estimación puntual**	**IC 95%**	
adjetivos	1.185	1.061	1.323
adverbios	0.733	0.627	0.857
Conj. Cop.	1.319	1.069	1.628
determinantes	1.129	1.031	1.237
nombres	0.819	0.748	0.897
preposiciones	1.171	1.052	1.302

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. La razón de odds para el número de adjetivos es 1.19 lo cual indica que la chance de clasificar a un texto como Científico se incrementa en un 19% al aumentar en número de adjetivos en una unidad. Con respecto al número de adverbios la razón de odds es menor a la unidad por lo tanto si se interpreta el recíproco, 1/0.73=1.36, significa que la chance de clasificar un texto en el corpus No Científico aumenta un 36% al incrementarse en una unidad el número de adverbios. Si analizamos el efecto de las conjunciones copulativas, determinantes y preposiciones, al incrementar en una unidad cada una de estas categorías gramaticales, la chance de clasificar un texto como Científico se incrementa en un 32%, 13% y 17% respectivamente. Al igual que el efecto

del número de adverbios, la probabilidad de clasificar un texto como No Científico se incrementa en un 22% (1/0.82=0.22) al aumentar en una unidad la cantidad de nombres en el texto.

3.1.2. Árboles de Clasificación

Se aplicó la técnica de Árboles de Clasificación para obtener reglas de clasificación que permitan asignar los textos en dos poblaciones, definidas por el género al que pertenecen: CIENTÍFICO y NO CIENTÍFICO. De la misma manera que en el apartado previo, los predictores utilizados corresponden a la distribución de las distintas categorías morfológicas halladas en el análisis automático de los textos (proporción de cada categoría morfológica).
Las variables que mostraron una buena discriminación de los grupos son la proporción de adverbios, nombres, adjetivos, preposición y verbos. El árbol final presenta 10 nodos terminales.

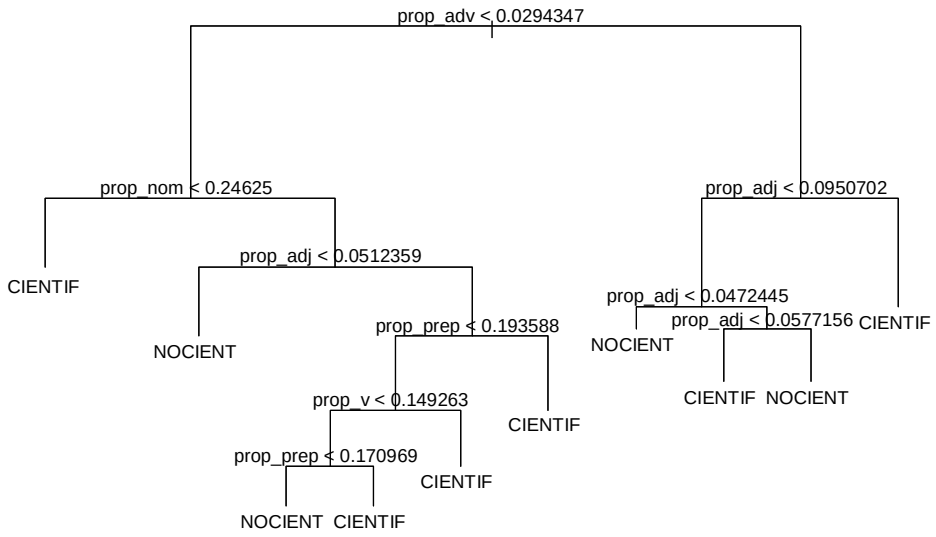

Gráfico 1: Árbol de clasificación

El gráfico 1 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el género es la proporción de adverbios en el texto, categorizada como superiores e inferiores a 0.029 (2.9%). El corpus general queda así dividido en dos grupos, los que presentan más del 2.9% de adverbios y los que no. Luego intervienen en las sucesivas subdivisiones el número de nombres, adjetivos, verbos y preposiciones. Interpretando el árbol resultante, se encuentran 10 perfiles de textos (que corresponden a los 10 nodos terminales) asociados con una de los dos géneros. Estos son:

- Textos con un porcentaje de adverbios inferior al 2.9% y un porcentaje de nombres menor a 25% son clasificados como textos CIENTÍFICOS.

- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25% y de adjetivos inferior al 5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición menor al 17% y de verbos menor al 15% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición entre 17% y 19%, y de verbos menor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición inferior al 19%, y de verbos mayor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición mayor al 19% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos inferior al 4.7% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 4.7% y 5.7% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 5.7% y 9.5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos superior al 9.5% son clasificados como textos CIENTÍFICOS.

El árbol final fue evaluado utilizando la muestra de prueba, no fue utilizada en la construcción del mismo, hallando una tasa de mala clasificación del 14%, siendo 4% para los textos científicos y 28% para los no científicos (Tabla 4).

Tabla 4: Tasa de mala clasificación

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	4%	28%

3.2. Desempeño en datos simulados

3.2.1. Análisis exploratorio de datos simulados

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. Los resultados de este análisis sugieren 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.

3.2.2. Aplicación de técnicas de clasificación

3.2.2.1.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 16% de mala clasificación, siendo el mínimo observado de 7% y el máximo de 25%.
Escenario 1b: En promedio, en las 500 muestras se observó en promedio un 11% de mala clasificación, siendo el mínimo observado de 3% y el máximo de 19%.

Escenario 2a: En promedio, en las 500 muestras se observó en promedio un 42% de mala clasificación, siendo el mínimo observado de 33% y el máximo de 51%.
Escenario 2b: En promedio, en las 500 muestras se observó en promedio un 25% de mala clasificación, siendo el mínimo observado de 15% y el máximo de 38%.

Escenario 3a: En promedio, en las 500 muestras se observó en promedio un 14% de mala clasificación, siendo el mínimo observado de 4% y el máximo de 23%.
Escenario 3b: En promedio, en las 500 muestras se observó en promedio un 10% de mala clasificación, siendo el mínimo observado de 3% y el máximo de 19%.

Escenario 4a: En promedio, en las 500 muestras se observó en promedio un 42% de mala clasificación, siendo el mínimo observado de 30% y el máximo de 54%.
Escenario 4b: En promedio, en las 500 muestras se observó en promedio un 25% de mala clasificación, siendo el mínimo observado de 15% y el máximo de 35%.

3.2.2.2.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 12% de mala clasificación, siendo el mínimo observado de 1% y el máximo de 21%.
Escenario 1b: En promedio, en las 500 muestras se observó en promedio un 10% de mala clasificación, siendo el mínimo observado de 5% y el máximo de 16%.

Escenario 2a: En promedio, en las 500 muestras se observó en promedio un 21% de mala clasificación, siendo el mínimo observado de 13% y el máximo de 29%.
Escenario 2b: En promedio, en las 500 muestras se observó en promedio un 17% de mala clasificación, siendo el mínimo observado de 8% y el máximo de 25%.

Escenario 3a: En promedio, en las 500 muestras se observó en promedio un 12% de mala clasificación, siendo el mínimo observado de 6% y el máximo de 18%.
Escenario 3b: En promedio, en las 500 muestras se observó en promedio un 9% de mala clasificación, siendo el mínimo observado de 4% y el máximo de 15%.

Escenario 4a: En promedio, en las 500 muestras se observó en promedio un 21% de mala clasificación, siendo el mínimo observado de 14% y el máximo de 28%.
Escenario 4b: En promedio, en las 500 muestras se observó en promedio un 17% de mala clasificación, siendo el mínimo observado de 10% y el máximo de 24%.

3.2.2.3.Comparación de los resultados hallados en datos simulados

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 5: 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 6: 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 6. 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.

4. CONCLUSIONES

Respecto a la aplicación en datos reales (clasificación de textos), el desempeño de las técnicas fue medido con la TMC calculada sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC resultó 4% para los textos científicos y 28% para los textos no científicos. Para el modelo de RL la TMC resultó 14% para los textos científicos y 26% para los textos no científicos.

La diferencia en la tasa de mala clasificación sólo se diferenció en el corpus de textos científicos para el cual con el árbol se obtuvo un 4% de mala clasificación versus un 14% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos están centradas principalmente en el porcentaje de adverbios, adjetivos, nombres y preposiciones presentes. Sin embargo, en el modelo de regresión logística han intervenido otras variables en la discriminación como los determinantes y conjunciones copulativas; mientras que el árbol de clasificación utiliza el porcentaje de verbos, categoría morfológica no utilizada en la regresión.

Una ventaja observada en el árbol de clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.

Dado que en estos datos las categorías de la variable respuesta están desbalanceadas, la clase mayoritaria (textos científicos) presenta un porcentaje de mala clasificación inferior a costa de un peor desempeño en la clase minoritaria (textos no científicos).

En la segunda instancia de evaluación, se ha comparado 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 a la TCC 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 una TCC 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 una TCC 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.

5. REFERENCIAS

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

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

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

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

[5] Beltrán C. Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario. Año 2009.

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

[7] James, G.;Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning with Applications in R. Springer, New York, 2013.

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

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

[10] Beltrán C. Estudio exploratorio para la comparación de distintos tipos de textos: Textos Científicos y Textos No Científicos. Revista de Epistemología y Ciencias Humanas. Nro. 5 Año 2013.