Ciencias Económicas 33-No.1: 2015 / 63-73 / ISSN: 0252-9521

DOI: http://dx.doi.org/10.15517/rce.v33i1.19971

UNA METODOLOGÍA PARA ENCONTRAR EL MEJOR CLASIFICADOR EN DECISIÓN EMPRESARIAL

José C. Vega Vilca[1]
David A. Torres Núñez[2]

Recibido: 30/01/2015 Aprobado: 16/06/2015

RESUMEN

En la investigación, se presenta una metodología para mejorar las estrategias de análisis en situaciones donde la clasificación supervisada se convierte en la herramienta fundamental de decisión empresarial. La necesidad de catalogar a los nuevos clientes en uno de varios grupos, definidos de acuerdo a las características del sujeto, es analizada mediante el cálculo de la tasa de error. Con este propósito, se elaboraron programas en lenguaje R para calcular la tasa de error de cada uno de los nueve clasificadores, usando el método de validación cruzada 10 (Stone, 1974), en 50 permutaciones de los datos en estudio. Para cada conjunto de datos analizados se demostró, mediante ANOVA, que efectivamente existen diferencias significativas en el promedio de tasas de error de los clasificadores (p=0.00); por lo tanto, se concluye que el mejor clasificador es aquel con la mínima tasa de error.

PALABRAS CLAVE: CLASIFICACIÓN SUPERVISADA, VALIDACIÓN CRUZADA, TASA DE ERROR, CLIENTE, DECISIÓN ESTADÍSTICA, ANÁLISIS MULTIVARIABLE.

ABSTRACT

In this research, a methodology is presented to improve strategies of analysis in situations where supervised classification becomes the fundamental tool for business decision. The need to categorize the new customers into one of several groups, according to the characteristics of the subject, is analyzed through the calculation of the error rate. Programs were written using the statistical software package R, to calculate the error rate of each of nine classifiers, using cross-validation method 10 (Stone, 1974), in the 50 permutations of the data under consideration. For each of the analyzed data sets it was demonstrated, through ANOVA, that there are indeed significant differences in the average error rates of classifiers (p=0.00); therefore, it is concluded that the best classifier is the one with the lowest error rate.

KEYWORDS: SUPERVISED CLASSIFICATION, CROSS VALIDATION, ERROR RATE, CUSTOMER, STATISTICAL DECISION, MULTIVARIATE ANALYSIS.

1 Universidad de Puerto Rico, P.0 BOX 23332, Código Postal 00931, San Juan, Puerto Rico; jose.vega23@upr.edu

2 Universidad de Puerto Rico, P.0 BOX 23332, Código Postal 00931, San Juan, Puerto Rico; david.torres9@upr.edu

I. INTRODUCCIÓN

Un análisis adecuado tanto de las características o dinámicas de comportamiento de los clientes actuales o potenciales resulta fundamental, así como de los datos sobre insumos, mercados, distribuidores, etc., los cuales son básicos para el diseño de estrategias empresariales. En este artículo, se plantea una metodología para mejorar las estrategias de análisis de datos en situaciones donde los clientes, al igual que los insumos o cualquier otro sujeto de estudio (personas, animales o cosas), deben ser catalogados correctamente en grupos definidos de acuerdo a sus características, para encontrar patrones favorables o negativos. Así, por ejemplo, los solicitantes de préstamo ante una entidad bancaria brindan información personal como ingresos, edad, sexo, situación familiar, antigüedad en su puesto de trabajo, gastos, número de dependientes, etc. Estas características están registradas en la base de datos del banco. A partir de los datos obtenidos en casos antiguos, se identifican los rasgos particulares de los clientes cumplidores, con el fin de determinar si se conceden o no los préstamos a los solicitantes.

La Clasificación Supervisada (Witten, Frank y Hall, 2011) es una herramienta estadística, cuyo propósito es construir un clasificador con mínima tasa de error de clasificación, con la finalidad de ubicar nuevos sujetos en uno de los grupos posibles, de acuerdo a las características del sujeto y del grupo donde será ubicado. Para la construcción del clasificador se necesita una matriz de datos **X** de orden n x p , donde n indica el número de sujetos y p el número de variables en estudio. Cada fila de la matriz **X** contiene las respuestas de cada sujeto a las p-variables. Además es necesario un vector de grupos o clases $\boldsymbol{Y}$ de orden $n\times1$ que contiene un indicador del grupo al que pertenecen cada uno de los n sujetos.

En el proceso de clasificación de un nuevo sujeto, este puede ser ubicado en un grupo que realmente no le corresponde; en ese momento, se considera que el clasificador ha cometido error de clasificación. Resulta indispensable conocer la tasa de error de clasificación, definida como la probabilidad que tiene el clasificador de ubicar un nuevo sujeto en una categoría que no le corresponde. Además. se debe experimentar con muchos clasificadores para encontrar un clasificador con mínima tasa de error. En este trabajo se calcula la tasa de error de nueve clasificadores sobre cada una de dos diferentes bases de datos en estudio; el mejor clasificador queda definido por la tasa de error mínima, que no necesariamente es el mismo para ambos conjunto de datos.

La investigación está enfocada en las siguientes hipótesis:

- Existen diferencias significativas entre los promedios de la tasa de error de los nueve clasificadores aplicados sobre un mismo conjunto de datos.
- No existe un clasificador que logre, en cualquier conjunto de datos, detectar la mínima tasa de error.

II. METODOLOGÍA

En este trabajo se analizan las características de los nuevos clientes de una empresa, con la finalidad de predecir su comportamiento. La clientela nueva será catalogada en categorías establecidas según el comportamiento de los clientes actuales de una empresa, –con la finalidad de elaborar estrategias empresariales particulares de acuerdo a las características del grupo de clientes. En general, el problema de predecir el comportamiento de un nuevo cliente, desde la óptica de clasificación supervisada, se resume de la siguiente manera:

> *Problema:* Se tiene un nuevo sujeto caracterizado por las p-variables estudiadas. ¿En cuál de los G grupos debe ser clasificado?
>
> *Respuesta:* El nuevo sujeto debe ser clasificado en el grupo, donde la proba-

bilidad de pertenecer a dicho grupo es mayor que la probabilidad de pertenecer a otros grupos.

Estrategia: Con base en la matriz de datos **X** de orden $n \times p$ y el vector de clases **Y**, se debe construir un clasificador con una mínima tasa de error.

A continuación, se presentan nueve clasificadores utilizados habitualmente en clasificación supervisada: Regresión Logística y Análisis Discriminante: Lineal y Cuadrático, los K-Vecinos Más Cercanos con K=1,3 y 5, Naive Bayes y Arboles de Clasificación: *Recursive Partitioning and Regression Trees (rpart) y Conditional Inference Trees (ctree).*

- **Regresión Logística**

Es un modelo de regresión ampliamente empleado para analizar datos, donde la variable respuesta es binaria, dicótoma y, en algunos casos, polítoma; mientras, las variables predictivas pueden ser continuas o categóricas. La regresión logística es un caso del Modelo Lineal Generalizado (GLM, por sus siglas en inglés); donde la estimación de parámetros y, consecuentemente la estimación de probabilidades, se realiza por el método de máxima verosimilitud, (Dobson, 2002)

- **Análisis Discriminante**

Es una técnica del análisis multivariante que construye una función clasificadora basada en datos multivariantes, los cuales pertenecen a clases o grupos bien definidos, con la finalidad de asignar nuevos sujetos a uno de estos grupos. La función clasificadora se construye como una combinación lineal de un grupo de variables independientes o predictivas; en caso de existir homogeneidad de las matrices de covarianza de los grupos en estudio, se aplicará Análisis Discriminante Lineal y en caso contrario, se empleará Análisis Discriminante Cuadrático. (Venables y Ripley, 2002)

- **K-vecinos más cercano**

El clasificador *K-Nearest Neighbor* (KNN), es un clasificador sencillo, basado en distancias. Un nuevo individuo será clasificado en la clase más frecuente a la que pertenecen sus K-vecinos más cercanos. Para cada uno de los valores más usados de K=1,3 y 5 , existe un clasificador diferente. (Ripley, 1996)

- **Naive Bayes**

Es un algoritmo de clasificación sencillo pero muy eficiente, basado en el teorema de Bayes para el modelado de predicción de la clase de un nuevo individuo. La palabra naive (ingenuo en inglés) se emplea porque el algoritmo utiliza técnicas bayesianas, pero no tiene en cuenta las dependencias entre variables predictoras, que realmente puedan existir. (Manning, Raghavan y Schutze, 2008)

- **Árboles de clasificación**

Es un clasificador que divide recursivamente el intervalo de los valores posibles de las variables predictivas, con la finalidad de construir redes lógicas y establecer reglas que representen el conocimiento del problema, mediante una estructura de árbol. Se presentan dos clasificadores de este tipo: *Recursive Partitioning and Regression Trees (rpart)*, según establecen Breiman, Friedman, Olshen y Stone (1984) y *Conditional Inference Trees (ctree)*, según establece Hothorn, Hornik, van de Wiel y Zeileis (2006).

- **Tasa de error de clasificación**

Aunque muchos autores evalúan al clasificador mediante la tasa de error aparente (Smith, 1947), esta no fue considerada porque generalmente es optimista y tiene un sesgo grande. La tasa de error de clasificación fue estimada mediante el método de validación cruzada 10 (Stone, 1974), que consiste en dividir la muestra en 10 partes. Con nueve partes de la

muestra, se construyó el clasificador utilizado para predecir y comparar la clase de la décima parte de la muestra que no intervino en la construcción del clasificador. Entonces, la tasa de error por validación cruzada 10 fue calculada como la suma de los errores cometidos en cada décima parte, dividido por el tamaño de la muestra (Witten et al., 2011).

La tasa de error por validación cruzada tiene poco sesgo, pero una alta variabilidad. Para reducir dicha variabilidad, se repite la estimación varias veces; por tal motivo, se incluyó 50 permutaciones de la muestra, en cada una de ellas se procedió al cálculo de la tasa de error por validación cruzada 10. Finalmente, la tasa de error fue calculada por el promedio de las 50 tasas de error por validación cruzada 10.

- **Bases de datos**

Como herramientas para trabajar con los sistemas de clasificación descritos en el apartado anterior, se utilizaron dos bases de datos: ***credit*** y *churn*, las cuales se describen a continuación.

- **Datos credit**

En esta investigación se usó los datos *credit*, la cual es una base de datos de IBM SPSS, con información demográfica y el historial de créditos bancarios, obtenida desde el subdirectorio *Samples* del directorio de instalación del software. La variable dependiente es "valoración de crédito". Esta recoge la evaluación de la institución bancaria sobre el cliente (crédito bueno o malo); las demás variables son llamadas variables independientes o predictoras. La base de datos consta de la información de 2464 clientes que solicitan crédito a un banco. El cuadro 1 presenta una descripción de los datos utilizados en este estudio. Solo la variable edad es cuantitativa, las demás variables son categóricas.

CUADRO 1
DESCRIPCIÓN DE LOS DATOS CREDIT

Variables	Descripción
X1: Edad	cuantitativa
X2: Nivel de ingresos	niveles: Bajo, Medio, Alto
X3: Número de tarjetas de crédito	niveles: Menos de 5, 5 o más
X4: Nivel de educación	niveles: Pre-Universidad, Universidad
X5: Número de préstamos de carro	niveles: Uno o ninguno, Dos o más
Y: Valoración de crédito	niveles: Malo, Bueno

Fuente: Elaboración propia, con base en los datos *credit*.

Datos churn

La base de datos *churn* (Blake y Merz, 1998) contiene el historial de las características de 3333 clientes de una compañía de telecomunicaciones. La variable dependiente es "descontinuación de suscripción", la cual indica si el cliente descontinuó la suscripción o no.

Esta base de datos fue empleada por Antipov y Pokryshevskaya (2010), para calcular la tasa de error de dos modelos de clasificación; para ello, los investigadores dividen la muestra total en muestra de entrenamiento (60%) y muestra de prueba (40%); sobre la primera, se construye un primer modelo de clasificación, el cual consiste en una Regresión Logística que, al ser aplicada a la muestra de prueba, se obtiene una tasa de error de 15%. Un segundo modelo de clasificación se obtuvo al segmentar la muestra de entrenamiento en 4 grupos de clientes, mediante la técnica del árbol de decisión CHAID; seguidamente los investigadores

construyen un modelo de regresión logística en cada grupo, para clasificar los datos de la muestra de prueba. De esta forma, lograron bajar la tasa de error a 7,9%. El cuadro 2 presenta una descripción de los datos utilizados en este estudio.

CUADRO 2
DESCRIPCIÓN DE LOS DATOS CHURN

variables	Descripción
X1: Número de meses en la cuenta	cuantitativo
X2: Plan internacional	niveles: Sí, No
X3: Plan de mensajes de voz	niveles: Sí, No
X4: Número de mensajes de voz	cuantitativo
X5: Total de minutos en el día	cuantitativo
X6: Total de minutos en la tarde	cuantitativo
X7: Total de minutos en la noche	cuantitativo
X8: Total de minutos internacional	cuantitativo
X9: Número de llamadas a Servicio al Cliente	cuantitativo
Y: Descontinuaron la suscripción	niveles: Sí, No

Fuente: Elaboración propia, con base en los datos *churn*.

III. RESULTADOS

A continuación se detallan los resultados obtenidos en cada uno de los modelos de clasificación supervisada propuestos en esta investigación, mostrando su aplicación en las bases de datos antes descritas. Los datos *credit* y *churn*, cuyas variables se describen en los cuadros 1 y 2respectivamente, fueron analizados por los nueve clasificadores y evaluados mediante el cálculo de la tasa de error de clasificación, a través de una aplicación desarrollada en el programa R.

Un resumen de la estimación de la tasa de error por validación cruzada 10, de cada una de las 50 permutaciones de la muestra se expone en los cuadros 3 y 4, para los datos *credit* y *churn*, respectivamente.

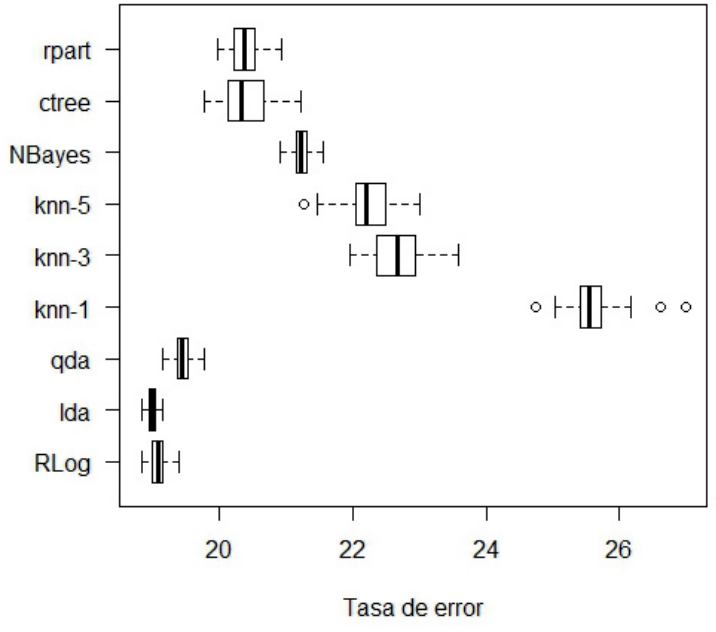
CUADRO 3
TASA DE ERROR DE CLASIFICACIÓN PARA DATOS CREDIT

MÉTODO	mínimo	máximo	Media	S	cv (%)
Reg. Logística	18,83	19,40	19,09	0,1092	0,57
Lda	18,83	19,16	19,00	0,0752	0,40
Qda	19,16	19,76	19,44	0,1147	0,59
knn-1	24,76	26,99	25,59	0,3862	1,51
knn-3	21,96	23,58	22,66	0,3723	1,64
knn-5	21,27	23,01	22,22	0,3936	1,77
Naïve Bayes	20,90	21,55	21,23	0,1546	0,73
Ctree	19,76	21,23	20,41	0,3376	1,65
Rpart	19,97	20,94	20,40	0,2273	1,11

Fuente: Elaboración propia.

En el cuadro 3, se presentan los valores mínimo, máximo, media, desviación estándar (S) y coeficiente de variabilidad (cv= $S/\bar{X}$) de la tasas de error de clasificación para los datos *credit*. Los clasificadores Regresión Logística y Discriminante Lineal alcanzan una tasa de error mínima (18, 83%) y el clasificador knn-1 (1-Vecino Más Cercano), calcula la tasa de error más alta (26, 99%). Respecto al promedio de las tasas de errores de las 50 permutaciones, se observa que el clasificador Discriminante Lineal proporciona la tasa de error mínima ($\bar{X}$=19.00,S=0.0752), con lo que este clasificador queda identificado como el mejor para el conjunto de datos *credit*. La figura 1 muestra la dispersión del cálculo de las tasas de errores por validación cruzada 10, en las 50 permutaciones de la muestra; se observa poca variabilidad de la tasa de error con los clasificadores Regresión Logística, Discriminante Lineal, Discriminante Cuadrático y Naive Bayes.

FIGURA 1
DISPERSIÓN DE LAS TASAS DE ERRORES DE CLASIFICACIÓN. DATOS CREDIT



Fuente: Elaboración propia.

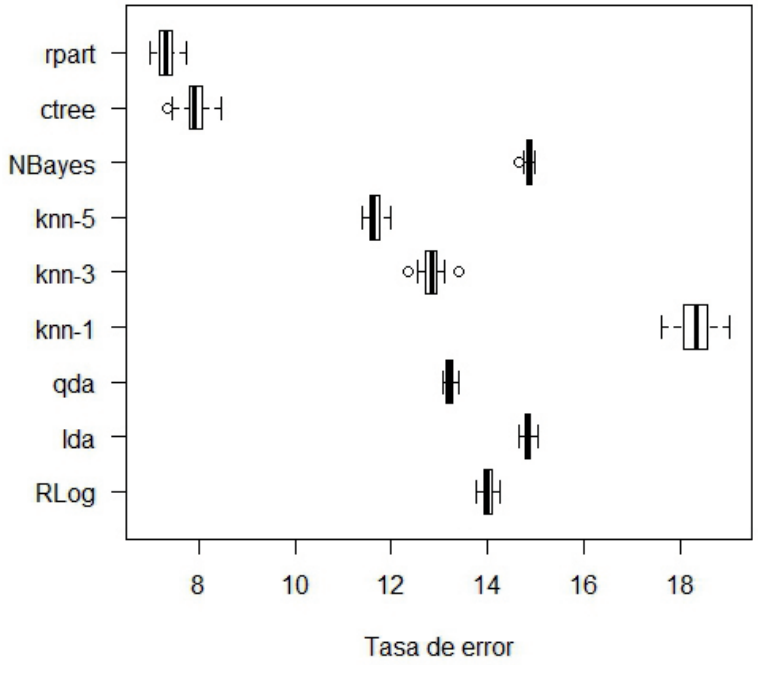
En el cuadro 4, se presentan los valores mínimo, máximo, media, desviación estándar (S) y coeficiente de variabilidad (cv= $S/\bar{X}$) de las tasas de error de clasificación para los datos de *churn*. El clasificador *rpart* alcanza la tasa de error mínima (6, 99%) y el clasificador knn-1 (1-Vecino Más Cercano), calcula la tasa de error máxima (18, 99%). Respecto al promedio de las tasas de errores de las 50 permutaciones, se observa que el clasificador *rpart* proporciona la mínima tasa de error ($\bar{X}$ =7.32,S=0.1831), con lo que este clasificador queda identificado como el mejor para el conjunto de datos *churn*. La figura 2 muestra la dispersión del cálculo de las tasas de errores por validación cruzada 10, en las 50 permutaciones de la muestra. Como en el caso anterior, se observa muy poca variabilidad de la tasa de error con los clasificadores Regresión Logística, Discriminante Lineal, Discriminante Cuadrático y Naive Bayes.

CUADRO 4
TASA DE ERROR DE CLASIFICACIÓN PARA DATOS CHURN

	mínimo	máximo	media	S	cv (%)
Reg, Logística	13,74	14,25	13,99	0,1104	0,79
Lda	14,64	15,03	14,83	0,0893	0,60
Qda	13,05	13,41	13,21	0,0821	0,62
knn-1	17,61	18,99	18,30	0,2937	1,60
knn-3	12,33	13,38	12,82	0,1794	1,40
knn-5	11,40	11,97	11,65	0,1344	1,15
Naïve Bayes	14,64	14,97	14,86	0,0651	0,44
Ctree	7,35	8,46	7,92	0,2144	2,71
Rpart	6,99	7,74	7,32	0,1831	2,50

Fuente: Elaboración propia.

FIGURA 2
DISPERSIÓN DE LAS TASAS DE ERRORES DE CLASIFICACIÓN. DATOS CHURN



Fuente: Elaboración propia.

- **Justificación de los Modelos de Clasificación propuestos**

Debido a que en cada permutación de la muestra la tasa de error por validación cruzada 10 fue calculada por cada uno de los nueve clasificadores, entonces cada clasificador calculó la tasa de error 50 veces. Un análisis de varianza (ANOVA) de dos factores se utilizó para probar si existen diferencias en los promedios de tasas de error calculado por cada clasificador.

Para los datos *credit*, el ANOVA de dos factores: clasificadores (tratamientos) y permutación de la muestra (bloques) concluyó que existen diferencias significativas en los promedios de tasas de error de los clasificadores ($F(8,392)=3329,14, p=0,00$). La prueba de comparación múltiple de Tukey fue utilizada para detectar diferencias significativas entre todos los pares de promedios de tasas de error de los clasificadores. No se encontraron diferencias entre Regresión Logística y Discriminante Lineal ($p=0,76$), con lo cual ambos se identificaron como los mejores clasificadores para los datos *credit*, porque el promedio de la tasa de error fue mínima. Tampoco se encontraron diferencias significativas ($p=0.99$) en los promedios de tasas de error de los clasificadores *ctree* y *rpart* .

Para los datos *churn*, el ANOVA de dos factores: clasificadores (tratamientos) y permutación de la muestra (bloques) concluyó que existen diferencias significativas en los promedios de tasas de error de los clasificadores ($F(8,392)=22684,52, p=0,00$). La prueba de comparación múltiple de Tukey fue usada para detectar diferencias significativas entre todos los pares de promedios de tasas de error de los clasificadores. No se encontró diferencias entre Discriminante Lineal y Naive Bayes ($p=0.99$), los demás promedios son todos diferentes entre sí, por lo tanto el clasificador *rpart* es el mejor para los datos *churn*, porque el promedio de la tasa de error es mínima (7, 32%).

- **Clasificadores**

El clasificador para el conjunto de datos *credit* debe ser el Discriminante Lineal (*DL*); aunque no hay diferencias en los promedios de tasas de error con los clasificadores Regresión Logística y Discriminante Lineal, se eligió el segundo, por tener un menor valor en el promedio de tasa de error (19, 00% frente a 19, 09%). El cuadro 5 presenta los coeficientes del clasificador, el cual es una combinación lineal de las variables contenidas en el conjunto de datos, $DL(X)=a_0+\sum^5_{i=1} a_i \times X_i$.

CUADRO 5
COEFICIENTE DEL CLASIFICADOR DISCRIMINANTE LINEAL. DATOS CREDIT

Variables	Coeficientes
X_1 : Edad	a_1=-0.10838505
X_2 : Nivel de Ingreso	a_2=-2.00548509
X_3 : Número de tarjetas de crédito	a_3= 2.15926201
X_4 : Nivel de educación	a_4=-0.06911501
X_5 : Número de préstamo de carros	a_5= 0.19916936
constante	a_0= 3.39986700

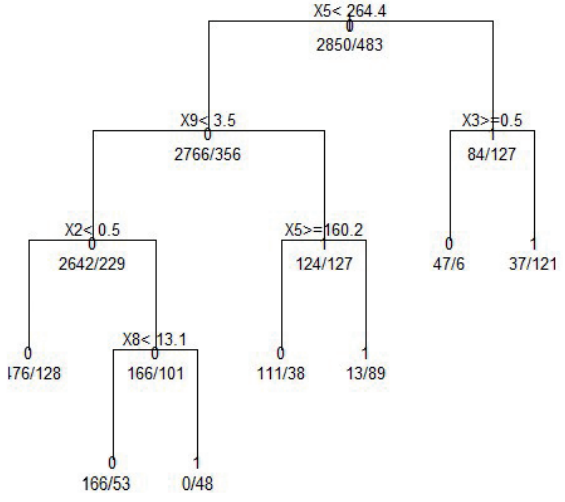
Fuente: Elaboración propia.

Para un nuevo cliente representado por $\mathbf{X}=(X_1, X_2,\ldots,X_5)$, la regla de clasificación es: si $DL(\mathbf{X})\geq 0$ entonces será clasificado como "mal cliente"; en caso contrario, será clasificado como "buen cliente". Así por ejemplo, un cliente de 34 años de edad (X_1=34), con nivel de

ingreso medio (X_2=2), con menos de 5 tarjetas de crédito (X_3=1), con educación de Escuela Superior (X_4=1) y con más de un préstamos de auto (X_5=2)., al aplicar el clasificador se tiene DL(X)=–1,807709; el cliente será clasificado como "buen cliente", con probabilidad de error de 19,00%, es decir con 81,00% de posibilidades de haber sido clasificado bien.

El clasificador para el conjunto de datos *churn*, debe ser *rpart*, ya que el promedio de la tasa de error de clasificación fue mínima (7, 32%). La figura 3 presenta un modelo simplificado de este clasificador (árbol de clasificación) para los datos *churn*. Las variables de estos datos X_1,X_2,…, X_9; están identificadas en el cuadro 2.

FIGURA 3
ÁRBOL DE CLASIFICACIÓN. DATOS CHURN



Fuente: Elaboración propia.

Para un nuevo cliente representado por X=(X_1,X_2,…, X_9), la figura 3 representa la regla de clasificación, la cual consiste de preguntas sobre el valor de una variable donde las respuestas pueden ser "sí" o "no". Si la respuesta fue "sí", el camino a seguir es por la izquierda, en caso contrario se sigue por la derecha. Ambos caminos conducen a la siguiente pregunta o al final del camino. Se puede llegar al valor "0" (no descontinuó la suscripción) o se puede llegar al valor "1" (se descontinuó la suscripción).

La figura 4 es un diagrama complementario, para una mejor lectura del árbol de clasificación. Así, por ejemplo, la clasificación de un cliente con 137 meses de antigüedad (X_1=137), sin plan internacional (X_2=0), sin plan de mensajes de voz (X_3=0), con ningún mensaje de voz (X_4=0), con total de 243,4 minutos en el día (X_5=243,4), 121,2 minutos en la tarde (X_6=121,2), 162,6 minutos en la noche (X_7=162,6), con 12,2 minutos internacionales (X_8=12,2) y ninguna llamada a servicio al cliente (X_9=0); es representado por el vector (X_1=137, X_2=0, X_3=0, X_4=0, X_5=243.4, X_6=121,2, X_7=162,6, X_8=12,2, X_9=0).

Al aplicar el clasificador se sigue la siguiente lógica, dada por el árbol de la figura 3: ¿Es X_5<264,4? cuando la respuesta es "sí", se continúa a la izquierda por la siguiente pregunta, la cual es ¿Es X_9<3,5?, cuando la respuesta es "sí", se prosigue hacia la misma dirección, donde se encuentra la siguiente interrogante: ¿Es X_2<0,5?, en caso de que la respuesta sea "sí", se sigue a la izquierda; en este caso, se llega al final de las preguntas, al valor "0"; entonces, el nuevo cliente será clasificado como que "no

descontinuó la suscripción", con probabilidad 0,95084485. Este valor se obtiene desde la figura 4, cuarta fila.

El árbol de clasificación representado en la figura 3 se complementa con el diagrama de lectura de la figura 4, donde lo más representativo se observa en las filas con asterisco: 4, 6, 7, 9, 10, 12 y 13. Estas filas representan los nodos terminales del árbol. Por ejemplo, si la secuencia lógica de los datos de un nuevos clientes conducen a la fila 7 (en la figura 3, tercer nodo terminal desde la izquierda), serán clasificados en "1" (descontinuará la suscripción) con probabilidad 1.

FIGURA 4
DIAGRAMA PARA LA LECTURA DEL ÁRBOL DE CLASIFICACIÓN, DATOS CHURN

```
      1) root 3333 483 0 (0.85508551 0.14491449)
        2) X5< 264.45 3122 356 0 (0.88597053 0.11402947)
          4) X9< 3.5 2871 229 0 (0.92023685 0.07976315)
            8) X2< 0.5 2604 128 0 (0.95084485 0.04915515) *
            9) X2>=0.5 267 101 0 (0.62172285 0.37827715)
             18) X8< 13.1 219  53 0 (0.75799087 0.24200913) *
             19) X8>=13.1 48   0 1 (0.00000000 1.00000000) *
      5) X9>=3.5 251 124 1 (0.49402390 0.50597610)
          10) X5>=160.2 149  38 0 (0.74496644 0.25503356) *
          11) X5< 160.2 102  13 1 (0.12745098 0.87254902) *
        3) X5>=264.45 211  84 1 (0.39810427 0.60189573)
          6) X3>=0.5 53   6 0 (0.88679245 0.11320755) *
          7) X3< 0.5 158  37 1 (0.23417722 0.76582278) *
```

Fuente: Elaboración propia con base en los datos *churn*.

IV. CONCLUSIONES

Al trabajar la metodología propuesta con dos bases de datos, se validaron las hipótesis que originaron este trabajo de investigación: se demostró que efectivamente existen diferencias significativas en el cálculo de la tasa de error de cada uno de los nueve clasificadores, cuando son aplicados sobre un conjunto de datos; y también que no existe un clasificador óptimo que logre una mínima tasa de error en dos conjuntos de datos. Para cada nuevo conjunto de datos se debe estudiar todos los clasificadores, uno de ellos será el más eficiente (tasa de error mínima).

Para lograr la mínima tasa de error en la utilización de clasificadores, se desarrolló un Modelo de Clasificación Supervisada basado en el cálculo de tasas de error de diferentes clasificadores por validación cruzada y en la utilización de permutaciones de la muestra. La tasa de error de clasificación por validación cruzada es la mejor medida de fiabilidad para la evaluación y uso de un clasificador. La inclusión de permutaciones de la muestra mejora aún más el cálculo de esta tasa de error, esta afirmación se hace evidente por la variabilidad observada en su cálculo. En cada una de las 50 permutaciones de la muestra, se calculó una tasa de error por validación cruzada, que va desde un valor mínimo hasta un máximo. La tasa de error que mejor representa al clasificador fue calculada como promedio de esos 50 valores.

La tasa de error de un clasificador no es un valor constante, varía de acuerdo al conjunto de datos; no se encontró un único clasificador óptimo que logre la mínima tasa de error para los dos cualquier conjuntos de datos en estudio. Por ese motivo, se propone un modelo basado en la búsqueda de un clasificador con mínima tasa de error, a partir de la experimentación con muchos clasificadores, hasta identificar uno de ellos que alcance el mínimo esperado y, por lo tanto, sea el clasificador óptimo que se está buscando.

Este método, desarrollado como aplicación en lenguaje R, podría convertirse en una herramienta fundamental en decisiones empresariales debido a sus múltiples aplicaciones.

V. REFERENCIAS

Antipov, E., & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement and Analysis for Marketing*, 18 (2), 109-117.

Blake, C. L., & Merz, C. J. (1998). *Churn Data Set*. University of California. Department of Information and Computer Science, Irvin, CA. Recuperado de: http://www.sgi.com/tech/mlc/db/churn.data

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press LLC.

Dobson, A. (2002). *An Introduction to Generalized Linear Models*. Boca Raton, FL: CRC Press LLC. doi: 10.1002/sim.1493

Hothorn, T., Hornik, K., van de Wiel, M., & Zeileis, A (2006). A Lego System for Conditional Inference. *The American Statistician*, 60 (3), 257–263. doi: 10.1198/000313006X118430

Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. London: Cambridge University Pres

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. London: Cambridge University Press.

Smith, C. (1947). Some examples of discrimination. *Ann. Eugenic* 18, 272–282.

Stone, M. (1974). Cross-validatory choice and the assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, B 36, 111-133.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer-Verlag. doi: 10.1007/978-0-387-21706-2

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.