Análisis comparativo de modelos de clasificación en el estudio de la deserción universitaria

Emma V. Barreno Vereau

Resumen

El presente artículo tiene como finalidad mostrar una metodología para la comparación de los modelos de clasificación "regresión logística" y "árbol de clasificación". Este análisis comparativo se basó en el estudio de la deserción universitaria en una universidad particular. Se deseaba determinar si un alumno dado podría ser clasificado como un desertor potencial, teniendo como referencia determinadas variables explicativas. Para la aplicación de los modelos de clasificación se hizo uso del software comercial Minitab 16 y del software libre Weka 3-7-2, en los que se obtuvieron los modelos de regresión logística y el árbol de clasificación, respectivamente. En ambos casos se usaron los mismos datos de entrada y los datos de prueba para su evaluación. Entre las principales conclusiones se puede señalar que los dos modelos presentaron resultados similares, de acuerdo con las variables explicativas utilizadas, y que la elección del tipo de modelo a utilizar dependerá, además de la comparación de los resultados, de los requerimientos del estudio y de las facilidades de su implementación dentro del sistema de información de la institución que lo realice.

Palabras clave: deserción universitaria / predicción / árboles de clasificación / regresión logística.

Introducción

En diversas situaciones es necesario clasificar a un individuo en una categoría establecida, de acuerdo con ciertas características. Este es el caso de la deserción universitaria, en el cual se desea identificar a los posibles futuros desertores. Una de las interrogantes que se originan, entre otras, se encuentra relacionada con la elección del modelo de clasificación, la cual puede ser determinada a partir de la evaluación de diversos modelos. El artículo tiene como objetivo desarrollar una metodología para la comparación de modelos de clasificación, basado en el estudio de la deserción universitaria en una universidad particular mediante la aplicación de 2 modelos de clasificación: regresión logística y árbol de clasificación. En la metodología utilizada se señalan las ventajas y desventajas de la aplicación de los modelos de clasificación, lo que servirá de apoyo en el análisis de una realidad compleja como es la deserción universitaria, en la cual diversas características explicativas, cuantitativas o cualitativas influyen en la determinación de una característica respuesta de tipo cualitativa.

Inicialmente se expone la definición del problema; luego se explica el marco de referencia, en el cual se brindan los principales alcances relacionados con la deserción universitaria y los modelos de clasificación que serán evaluados: la regresión logística y los árboles de clasificación. Posteriormente se desarrollan los factores que determinan esta deserción, sección en la que se define la variable de interés deserción estudiantil en una institución universitaria, así como las variables explicativas que influyen en la deserción estudiantil. Luego se procede a desarrollar el análisis comparativo de los modelos de clasificación, para lo cual se presentarán los resultados obtenidos mediante la aplicación de los modelos de regresión logística y de árbol de clasificación a partir de una data de entrada compuesta por 1059 registros, como también de una data de prueba con 820 registros; después se presenta la comparación de los modelos de clasificación, señalando las bondades y dificultades de la aplicación de cada uno de los modelos analizados. Finalmente se brindan las conclusiones y recomendaciones derivadas del presente artículo.

1. Definición del problema

En análisis de la deserción estudiantil, cuando se desea aplicar un modelo de clasificación, el analista debe elegir entre una gran variedad de modelos, tales como: análisis discriminante, regresión logística, árboles de clasificación (también conocidos como árboles de decisión), redes neuronales artificiales, entre otros. Para esta elección, a veces realiza la aplicación de varios de estos modelos pero, a pesar de la obtención de los resultados correspondientes, no dispone de una forma o metodología de comparación que le permita definir el modelo que aplicará en su análisis,

así surge la siguiente pregunta: ¿Cuál es la metodología que se debe seguir para comparar los modelos de clasificación?

2. Marco de referencia

2.1 Deserción universitaria

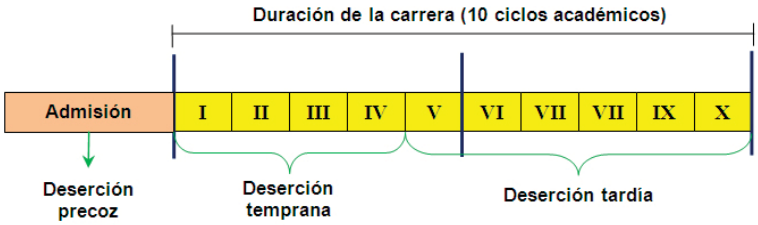
El análisis de la deserción estudiantil, y en forma particular la deserción universitaria, es un fenómeno que presenta una elevada complejidad en su análisis, que se ve incrementada si es que, además, se desean obtener modelos que permitan determinar si un alumno dado puede ser o no un futuro desertor. A continuación se brindan algunos alcances sobre los aspectos que se consideran para dicho análisis.

De acuerdo con el Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa (Sineace-Perú), la deserción, en términos generales, se encuentra relacionada con los alumnos que suspenden, cambian de carrera de estudios o la abandonan antes de finalizar los estudios y la correspondiente obtención del grado (2010: 25). Además, se indica que la deserción se puede definir, en la práctica, de diversas formas, según los requerimientos de cada institución, señalando que se puede evaluar en un horizonte de tiempo determinado por el número medio de años que se requieren para completar los estudios motivo de análisis. Asimismo, manifiesta que la deserción es un proceso de abandono de los estudios debido a la influencia de diversos factores; esto último se basa en lo señalado en el Glosario Internacional de la Red Iberoamericana para la Acreditación de la Calidad de la Educación Superior (Riaces 2004), tal como se señala en el mismo documento.

En relación con los tipos de deserción, el Ministerio de Educación Nacional de Colombia (Mineducación) señala que se pueden diferenciar 2 tipos de deserción en los estudiantes universitarios: deserción respecto al tiempo y deserción respecto al espacio (2009: 22). Para el presente artículo se tomará como base de análisis la primera de estas.

La deserción con respecto al tiempo se clasifica en precoz, temprana y tardía (ver figura 1), con las siguientes características:

a) Deserción precoz: sujeto que, luego de haber logrado una vacante en la universidad, no se matricula y, por lo tanto, nunca inicia los estudios en la institución.
b) Deserción temprana: alumno que abandona sus estudios durante los primeros ciclos de estudio de la carrera profesional a la cual ingresó.
c) Deserción tardía: alumno que abandona los estudios en ciclos de estudios avanzados de la carrera profesional (Mineducación 2009: 22).



La determinación de un estudio de deserción universitaria de acuerdo con el tiempo, es solamente uno de los muchos aspectos que se deben considerar en un análisis de este tipo, ya que existen diversos enfoques para el estudio de la deserción, entre los cuales se encuentra el análisis de la deserción intersemestral y el análisis de la deserción por cohortes (Guzmán 2009: 89). Además, otro de los aspectos que se deben tener en cuenta es si la deserción se analizará mediante una perspectiva macro (universidad) o micro (facultad), siendo posible tener una perspectiva más específica si se considera solamente una carrera profesional de las que ofrece la facultad (Guzmán 2009: 90).

Asimismo, Guzmán establece una diferenciación entre el desertor definitivo y el temporal, para lo cual es necesaria la determinación de un número de períodos consecutivos en los cuales el alumno puede dejar de presentar matrícula en su programa académico (Guzmán 2009: 87). Si el alumno se encuentra por debajo o a lo más en la cantidad de períodos antes determinados, sin presentar matrícula, entonces, es considerado como desertor temporal, en caso contrario, será considerado un desertor definitivo.

En líneas generales, estos aspectos deben ser considerados para el análisis de la deserción universitaria, además de otros, los cuales dependerán de los objetivos planteados, así como del alcance del estudio. Mención aparte merecen los factores que influyen en la deserción, sobre cuya base se extraerán algunas variables de análisis para el estudio correspondiente, que servirán de datos de entrada para los diversos modelos matemáticos propuestos para el estudio de la deserción universitaria.

1.2 Modelos de clasificación

Tal como señala Serna, "La clasificación es una actividad inherente al hombre, siempre existe la necesidad de ordenar o poner límites pues esto ayuda a entender fenómenos reales" (2009: 8). En efecto, por ejemplo, para el análisis de la deserción universitaria, más allá de simplemente analizar los factores que influyen directamente en el proceso de deserción, es de especial interés la determinación, llámese clasificación, de los posibles desertores.

Para la realización de la clasificación de individuos, con base en determinadas características, se recurre a los modelos de clasificación. Haciendo uso de la terminología de la "minería" de datos, existen 2 tipos de enfoques para esta: aprendizaje supervisado y aprendizaje no supervisado, aunque, tal como señalan Hernández et al., "La importancia de estos nombres actualmente es puramente terminológica" (2008: 144). A continuación se presenta una descripción de ambos enfoques, adaptado de lo señalado por Serna:

a) *Aprendizaje supervisado:* Cuando se conocen de antemano las categorías de clasificación y se desea clasificar a los individuos dentro de estas a partir de los valores de ciertas características. Las técnicas más utilizadas son la regresión logística, el análisis discriminante y los árboles de clasificación, entre otros.
b) *Aprendizaje no supervisado:* Cuando se desconocen las categorías de clasificación y lo que se desea es determinar las categorías o grupos de clasificación, a partir de los valores de ciertas características. La técnica por excelencia es el análisis de conglomerados (análisis clúster) (2009: 19).

El presente artículo analizará comparativamente 2 técnicas del denominado aprendizaje supervisado: la regresión logística y los árboles de clasificación. Se ha preferido la regresión logística por sobre el análisis discriminante pues los 3 procedimientos del análisis discriminante han sido comparados con la regresión logística en los estudios de Shelley & Donner (1987), Castrillón (1998), Usuga (2006) y Barajas (2007), obteniendo que, en general, la regresión logística produce mejores resultados (Serna 2009: 9).

2.1.1 Regresión logística

Según Cuadras, "El modelo de regresión logística permite estimar la probabilidad de un suceso que depende de los valores de ciertas covariables" (2012: 217). Este tipo de modelos también son considerados como modelos dicotómicos (Uriel y Aldás 2005: 324) con 2 alternativas o categorías, de tal forma que cada individuo tiene que pertenecer a una de ellas.

A manera de alcance, se brindarán algunos conceptos y nociones básicas de la regresión logística, en especial los relacionados con los planteamientos del modelo, la estimación e interpretación de parámetros y las pruebas de contraste o de evaluación del modelo.

A. Planteamiento del modelo

Supóngase que un evento A puede o no presentarse en cada uno de los individuos de una población en estudio compuesta por n individuos; por lo tanto, se considera una variable binaria (dicotómica) Y que adopta los siguientes valores:

$$Y = \begin{cases} 1, & \text{Si se presenta el evento A} \quad \text{Prob}(y_i = 1) = P_i \\ 0, & \text{Si no se presenta el evento A} \quad \text{Prob}(y_i = 0) = 1 - P_i \end{cases}$$

$$i = 1, \ldots, n$$

De la expresión señalada, tal como aclaran Uriel y Aldás, la esperanza de la variable ***Y*** es igual a la probabilidad de que la variable ***Y*** tome el valor de 1 (2005: 325):

$$E(Y) = 1*P_i + 0*(1 - P_i) = P_i$$

En la expresión de probabilidad señalada, la ocurrencia del evento ***A*** no se encuentra influenciada por otras características. Supóngase ahora que la probabilidad P_i depende de los valores de ciertas características o variables explicativas: X_1, … , X_m. Sea $x_i = (x_{i1}, \ldots, x_{im})^T$ un vector de variables explicativas con los valores correspondientes a las mencionadas características asociadas al ***i***-ésimo individuo, entonces, la probabilidad de que se presente el evento ***A***, dado $\boldsymbol{x}_i$, viene dada por la siguiente expresión:

$$\text{Prob}(y_i = 1|x_i) = P_i$$

La probabilidad de que el evento ***A*** no suceda, dado $\boldsymbol{x}_i$, viene dada por la siguiente expresión:

$$\text{Prob}(y_i = 0|x_i) = 1 - P_i$$

Para obtener un modelo que permita determinar la probabilidad de que en un individuo de análisis se presente el evento de interés (y = 1) o no se presente dicho evento (y = 0), en función de una serie de variables explicativas, se deberá tener en cuenta:

$$Pi = F(X), RF =_1 \langle 0,1 \rangle$$

Donde, F(X) es una función de las variables explicativas, cuyo rango debe encontrarse entre 0 y 1. Usualmente, la función F elegida para el análisis es la función de distribución logística, dada por:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im})}}$$

Función que puede expresarse de forma abreviada, haciendo uso del vector

$b = (\beta_1, \ldots, \beta_m)^T$, el cual representa los parámetros de regresión del modelo, así como del vector de variables explicativas $x_i = (x_{i1}, \ldots, x_{im})$, quedando la siguiente expresión reducida:

$$\text{Prob}(y_i = 1|x_i) = P_i = \frac{1}{1+e^{-(\beta_0+\beta_i^T x_i)}}$$

Y de igual forma:

$$\text{Prob}(y_i = 0|x_i) = 1 - P_i = \frac{1}{1+e^{(\beta_0+\beta_i^T x_i)}}$$

La función de probabilidad y de distribución logística posee la siguiente propiedad:

$$g_i = \log\left(\frac{\text{Prob}(y_i = 1|x_i)}{\text{Prob}(y_i = 0|x_i)}\right) = \log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im}$$

la cual es una expresión lineal denominada ***Logit***, que tal como señalan Ibarra y Michalus:

> Como se aprecia […] la Transformación Logit es lineal en los parámetros del modelo, de manera que permite realizar análisis muy similares a los de la Regresión Lineal; un coeficiente positivo aumenta la probabilidad de ocurrencia del evento, en tanto que uno con signo negativo la disminuye (2010: 51).

Cabe señalar que si la función F elegida hubiera sido la función de distribución normal estándar, se hubiera tenido el denominado modelo ***Probit***.

B. Estimación e interpretación de parámetros

Estimación

Para la estimación de los parámetros del modelo planteado, vector b, se hace uso del método de máxima verosimilitud, siguiendo a Uriel y Aldás (2005: 329) cuando dicen que "A diferencia de mínimos cuadrados ordinarios que tiene una solución analítica, los modelos Logit y Probit son modelos no lineales, que deben estimarse utilizando procedimientos iterativos".

Para el cálculo de estos parámetros, los procedimientos iterativos no se pre-

sentan en este artículo, pues escapan del alcance planteado. Pero se hará uso del software Minitab para realizar la estimación correspondiente.

Interpretación

La interpretación de un coeficiente de un modelo de regresión logística puede realizarse mediante el logaritmo neperiano de la ***razón de apuestas*** (RA) derivada respecto a la variable cuyo coeficiente desea interpretarse (Uriel y Aldás 2005: 326). A estos valores también se les denomina ***odds ratio***, lo cual indica "cuánto más probable es el éxito que el fracaso" (Giraldo 2009: 72), dado un cambio unitario en la variable asociada. La razón de apuestas, para la i-ésima observación, es el cociente entre la probabilidad de que suceda el evento de interés sobre la probabilidad de que no ocurra. Por lo tanto, el ***odds ratio*** de la k-ésima variable explicativa.

$$\textit{Odds ratio}\,(X_k) = \frac{\partial(\mathrm{LnRA_i})}{\partial X_k} = \frac{\partial\left(\mathrm{Ln}\left(\frac{P_i}{1-P_i}\right)\right)}{\partial X_k} = \frac{\partial\left(\mathrm{Ln}\left(e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}\right)\right)}{\partial X_k}$$

$$\textit{Odds ratio}\,(X_k) = \frac{\partial(\mathrm{LnRA_i})}{\partial X_k} = \frac{\partial\left(\beta_0+\sum_{j=1}^{m}\beta_j x_j\right)}{\partial X_k} = \beta_k$$

En conclusión, como el coeficiente de cada variable es igual al ***odds ratio*** asociado a la variable correspondiente, y este se encuentra asociado a la razón de apuestas, se pueden realizar las sigueintes interpretaciones:

- Un *odds ratio* $(X_k) < 1$ significa que la razón de apuestas, disminuirá por cada incremento unitario de la variable X_k, manteniéndose constantes las demás variables. Ejemplo: Si odds ratio $(X_k) = 0{,}5$, la razón de apuesta se reducirá a la mitad.
- Un *odds ratio* $(X_k) = 1$ significa que la razón de apuestas se mantendrá sin variación por cada incremento unitario de la variable X_k, manteniéndose constante las demás variables. Un odds ratio $(X_k) = 1$ significa que la razón de apuesta no se modifica.
- Un *odds ratio* $(X_k) > 1$ significa que la razón de apuestas aumentará por cada incremento unitario de la variable X_k, manteniéndose constantes las demás variables. Ejemplo: Si odds ratio $(X_k) = 2$, la razón de apuesta se duplicará.

Por lo tanto, ***odds ratios*** significativamente diferentes a la unidad son un indicio de que la variable es significativa para el modelo, mientras que un ***odds ratio*** unitario indica que la variable no es significativa para el modelo y podría ser retirada y así obtener un modelo más reducido que no considere dicha variable.

C. Pruebas de contraste o de evaluación del modelo

Las pruebas de contraste del modelo de regresión logística son variadas, entre las cuales se pueden mencionar a las siguientes:

a) Contraste de razón de verosimilitudes (modelo completo o subconjunto de variables explicativas).
b) Test de Wald (significancia de cada variable explicativa).
c) Pseudo R^2 (capacidad explicativa del modelo).
d) Contraste de bondad de ajuste de Hosmer y Lemeshow (discriminación del modelo).
e) Curva ROC (discriminación del modelo).
f) Matriz de confusión (discriminación del modelo).

Se procederá a una breve explicación de las 2 primeras pruebas; en relación con la matriz de confusión, esta será desarrollada para la evaluación de los árboles de clasificación.

- *Contraste de razón de verosimilitudes – significación del modelo:* Esta prueba puede aplicar para evaluar la significancia del modelo (significancia del conjunto total de variables explicativas) o para un subconjunto de dichas variables, incluso si solamente se desea evaluar una sola variable explicativa (Uriel y Aldás 2005: 330). Para la realización de esta prueba se construye el estadístico RV_0, el cual "[…] se distribuye asintóticamente […] como una Chi-cuadrado con k – 1 grados de libertad, siendo k el número de regresores del modelo incluido el término independiente (Uriel y Aldás 2005: 330).

$$\text{El estadístico } RV_0 \text{ se define como: } RV_0 = -2[\ln(\boldsymbol{L}_0) - \ln(\boldsymbol{L})]$$

Es decir el RV_0 se calcula a partir de la diferencia del logaritmo neperiano de la función de verosimilitud asociada al estimar el modelo completo, ln(L), y del logaritmo neperiano de la función de verosimilitud asociada al estimar el modelo a partir de únicamente el término independiente, $\ln(L_0)$, (Uriel y Aldás 2005: 329).Para la realización de este contraste se plantean las siguientes hipótesis y regla de decisión:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_i = 0$ (Modelo no es significativo).

H_1: Al menos un β_i 0 (Modelo sí es significativo).

Como en este caso, existen i + 1 regresores, incluyendo el término indepen-

diente, entonces, la distribución Chi-cuadrado poseerá *i* grados de libertad. Si $P(\chi_i > RV_0) \leq \alpha = 0{,}05$; así, se rechaza la hipótesis nula, es decir, que el modelo sí es significativo. En caso contrario, el modelo no es significativo.

- Test de Wald

  Sea el vector $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)^T$ el cual representa a los parámetros de regresión estimados, se tiene que la distribución del parámetro $\hat{\beta}_i$ es una normal $N(\hat{\beta}_i, Var(\hat{\beta}_i))$. El test de Wald que permite determinar la significancia del parámetro β_i utiliza el siguiente estadístico:

$$Z_0 = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}}$$

  Estadístico que se distribuye como una normal N(0,1) (Cuadras 2012: 221). Para la realización de este contraste se plantean las siguientes hipótesis y regla de decisión:

  H_0: $\beta_i = 0$ (Variable X_i no es significativa para el modelo)

  H_1: $\beta_i^1 0$ (Variable X_i sí es significativa para el modelo)

  Si $2*P(Z > |Z_0|) \leq \alpha = 0{,}05$, entonces se rechaza la hipótesis nula, es decir, la variable Xi sí es significativa para el modelo. En caso contrario, la variable no es significativa para el modelo. Cabe mencionar que valores de $\alpha = 0{,}10$ también son aceptables en la aplicación del presente test.

1.1.2 Árboles de clasificación

Este método realiza la clasificación más adecuada de cada instancia o registro (individuo o entidad en estudio) con base en los valores de ciertos atributos, cuantitativos o cualitativos, que miden diferentes características de la instancia. Los árboles de clasificación forman parte de los métodos denominados "aprendizaje supervisado", ya que el aprendizaje de la clasificación se realiza disponiendo del resultado real, denominado "clase de la instancia", atributo cualitativo que forma parte del conjunto de atributos de la instancia (Witten y Frank 2005: 42-43).

Los árboles de clasificación son uno de los métodos de la minería de datos que presentan mayor facilidad en su aplicación y comprensión (Hernández et al. 2004: 281), ya que se basa en un conjunto de reglas de decisión, las cuales se aplican a determinados atributos asociados a la instancia que se desea clasificar. Existen diversas descripciones relacionadas con los árboles de clasificación, una de ellas: "Un árbol de decisión es un conjunto de condiciones organizadas en una estruc-

tura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas" (Hernández et al. 2004: 281-282).

En relación con la organización mediante una estructura jerárquica, se refiere a que el árbol presenta niveles en los cuales se pueden encontrar nodos de decisión o nodos de clasificación (nodos terminales). El primer nivel del árbol se encuentra conformado únicamente por un nodo de decisión, al cual se le denomina "nodo raíz"; la decisión o evaluación efectuada en dicho nodo es un paso necesario para llegar a la clasificación final de la instancia, generalmente la evaluación se realiza basándose en solamente uno de los atributos asociados a dicha instancia.

A partir del nodo raíz, en el siguiente nivel, se originan otros nodos los cuales pueden también ser nodos de decisión o nodos de clasificación; a los nodos de clasificación, al ser nodos terminales, también se les denomina hojas. Los siguientes niveles surgen a partir de los nodos de decisión de los niveles previos, y se continúa con este proceso jerárquico hasta llegar al último de los niveles, el cual se encuentra compuesto solamente por nodos de clasificación.

Debido a esta ramificación de nodos, se va conformando un diagrama de decisiones y clasificaciones a la que se denomina "árbol de clasificación". Una representación esquemática de la descripción realizada se aprecia en la siguiente figura:

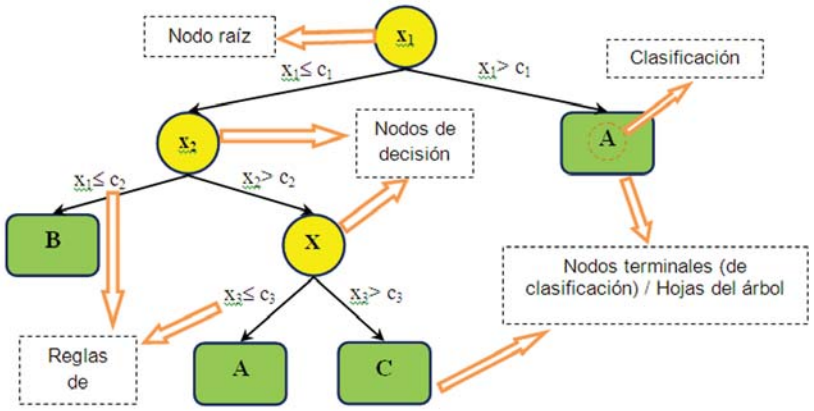


Figura 2. Estructura de un árbol de clasificación

Elaboración propia.

La figura muestra un árbol de clasificación hipotético, el que realiza la clasificación de las instancias de acuerdo a una clase con 3 categorías: A, B y C; los atributos que permiten realizar la clasificación son de tipo cuantitativo, y por lo tanto la regla de decisión asociada es una comparación con respecto a un valor constante; los atributos también pudieron ser cualitativos, y la regla de decisión sería comparar el

valor del atributo con cada una de las categorías que la conforman, en este caso se podrían originar tantas ramificaciones como categorías posea el atributo en evaluación. Además, un atributo cuantitativo utilizado en un nodo de decisión puede ser utilizado en ramas posteriores, más no así un atributo cualitativo.

Enseguida se describirán algunos aspectos básicos relacionados con los árboles de clasificación, como el modelamiento y las reglas de decisión, la obtención del árbol de clasificación e interpretación (nodos terminales) y la evaluación del modelo.

A. Modelamiento y reglas de decisión

Modelamiento

En un modelo de árbol de clasificación es necesario contar con un atributo de clasificación *Y*, atributo respuesta, al cual se le denomina clase, así como de un conjunto de atributos explicativos (X_1, X_2, …, X_p), los cuales pueden ser cuantitativos o cualitativos. El modelo del árbol consiste en particionar el espacio de los atributos explicativos en forma tal que los valores que toma el atributo respuesta sean cada vez más homogéneos (Juárez y Castells 2010: 109). Para una mejor comprensión de lo señalado se presenta la siguiente figura:

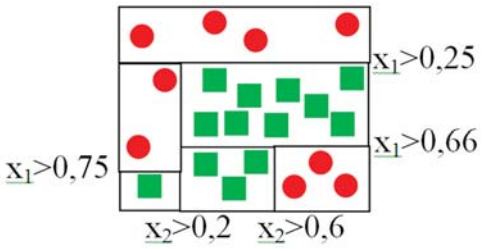


Figura 3. partición realizada con dos atributos x_1 y x_2 cuantitativos

Fuente: Adaptado de Hernández et al. (2004: 286).

Reglas de decisión

Las reglas de decisión, asociadas a una determinada partición, son un conjunto exhaustivo y concluyente (Hernández et al. 2004: 284). Los algoritmos de los árboles de decisión se diferencian, entre otros aspectos, por la forma como se determinan las reglas de decisión para la conformación de las particiones. En este artículo se describirá el algoritmo C4.5, el cual, según señala Hernández et al., es uno de los más usados. Este fue desarrollado por Quinlan, tomando como base otro algoritmo suyo, el ID3 (Hernández et al. 2004: 285).

Para Hernández et al. las particiones se realizan de la siguiente manera:

a) Particiones nominales: Sea x_i un atributo nominal (cualitativo), con las siguientes categorías: $\{v_1, v_2, \ldots, v_k\}$, únicamente se presenta un tipo de partición: $(x_i = v_1, x_i = v_2, \ldots, x_i = v_k)$, la cual es una condición de igualdad entre el atributo y cada posible categoría que pueda adoptar (2004: 285).
b) Particiones numéricas: Sea x_i un atributo numérico (cuantitativo), discreto o continuo, por lo tanto dicho atributo puede tomar mucho o infinitos valores. Entonces, se obtienen particiones a partir de intervalos: $(x_i \leq c_1, x_i > c_1)$, donde c_1 es un valor numérico constante, de tal forma que realice una discriminación adecuada de las instancias en evaluación. (2004: 285).

Tal como señalan Hernández et al. en relación con las reglas de decisión descritas:

> La expresividad resultante […] se conoce como expresividad proposicional cuadricular. El término proposicional se refiere a que son particiones que sólo afectan a un atributo de un ejemplo a la vez, es decir, ni relacionan 2 atributos del mismo ejemplo, ni 2 atributos de distintos ejemplos. El término cuadricular hace referencia al tipo de particiones que realizan, especialmente cuando atendemos a los atributos numéricos (2004: 285).

Este tipo de reglas de decisión son de gran simplicidad y permiten la obtención de árboles de clasificación bastante adecuados. El gran reto consiste en la elaboración de las reglas de decisión, de tal forma que realicen clasificaciones adecuadas y cuya determinación pueda ser realizad en forma sencilla. Un alcance de cómo se determinan estas reglas de decisión se presentan en el siguiente punto del artículo.

B. Obtención del árbol de clasificación e interpretación

Obtención del árbol de clasificación

La obtención de un árbol de clasificación se refiere a la obtención de las reglas de decisión que permitan realizar la clasificación más adecuada de las instancias en evaluación, utilizadas para la obtención del árbol, a la cual se le denomina "data de entrenamiento", como de otras instancias que no formen parte de la data de entrenamiento. Los aspectos relacionados con las reglas de decisión necesarias para no obtener árboles de clasificación demasiado específicos, es decir, que solamente se adecuen a los datos de entrenamiento, así como para no obtener árboles demasiado genéricos, no forman parte del alcance del presente artículo, pero se recomienda su profundización para una mayor comprensión del presente método de clasificación. Los criterios más utilizados para la determinación de las reglas de decisión se fundamentan en las frecuencias relativas de cada categoría de la clase en cada uno

de los nodos derivados, a los cuales se les denominará "nodos hijo", con respecto a las frecuencias relativas de las categorías de la clase correspondientes al nodo predecesor, denominado "nodo padre" (Hernández et al. 2004: 286).

Ejemplo aplicativo

Para una mejor comprensión, se presenta el siguiente caso: Sea un árbol de clasificación en el cual la clase posee solamente 2 categorías: A y B, y cuya data de entrenamiento se encuentra conformada por 100 instancias: 50 instancias A y 50 instancias B. Además, entre sus atributos explicativos se encuentran las variables cuantitativas x_1 y x_2. A continuación se presenta una figura que muestra cómo se utilizan las frecuencias relativas para determinar una mejor regla de decisión a partir del nodo raíz:

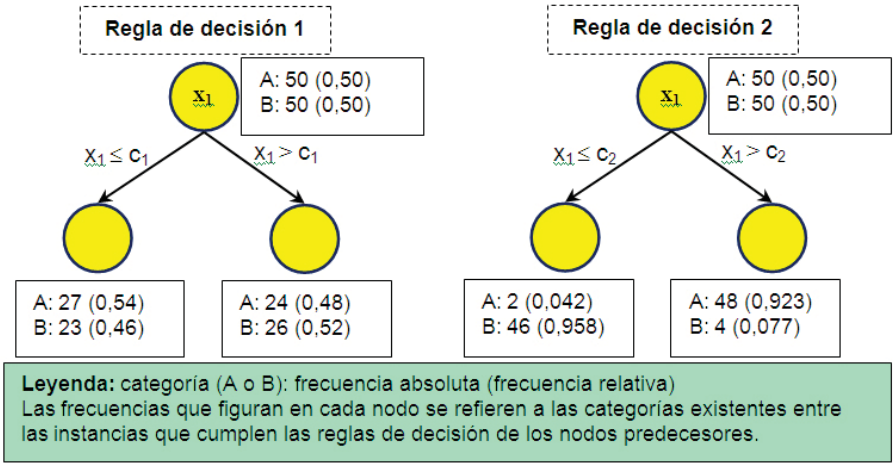


Figura 4. Comparación de reglas de decisión

Elaboración propia.

En la figura 4 se aprecia, a partir de las frecuencias relativas de las categorías de la clase, que la regla de decisión 2 ha discriminado de mejor forma a las categorías de la clase, mientras que en la regla de decisión 1 se aprecia que los nodos hijos presentan similares frecuencias relativas que el nodo padre y, por lo tanto, no realiza una adecuada discriminación de dichas categorías.

Interpretación

La clasificación que se realiza a una instancia se aprecia en las hojas del árbol (nodos terminales). Continuando con el ejemplo aplicativo, si la regla de decisión 1 hubiera sido la mejor de todas, se habría tenido que continuar con la elaboración de más

reglas, hasta llegar a obtener las clasificaciones correspondientes; mientras que en el caso de la regla de decisión 2, si bien se podría continuar obteniendo reglas de decisión adicionales, esto ya no sería conveniente, ya que se estaría obteniendo un árbol de clasificación demasiado específico. En este último caso, se podría considerar a dichos nodos como nodos terminales, es decir, como hojas, obteniéndose un árbol compuesto por 1 nodo de decisión y 2 hojas. Aunque es un ejemplo que no se ajusta a los casos reales de aplicación, sí permite una fácil interpretación de las clasificaciones realizadas.

a) Si una nueva instancia de evaluación es tal que su atributo x_2 es menor o igual a c_2, entonces se le clasificará como perteneciente a la categoría B de la clase de interés, ya que en la data de entrenamiento, de todas las instancias que cumplían dicha regla de decisión, el 95,8% eran de la categoría B, y el restante 4,2% eran de la categoría A.
b) Si una nueva instancia de evaluación, es tal que su atributo x_2 es mayor a c_2, entonces se le clasificará como perteneciente a la categoría A de la clase de interés, ya que en la data de entrenamiento, de todas las instancias que cumplían dicha regla de decisión, el 92,3% eran de la categoría A, y el restante 7,7% eran de la categoría B.

Es importante señalar que los diferentes programas informáticos que existen para la aplicación de árboles de clasificación, presentan de distinta forma los resultados de la clasificación; por lo tanto, es importante siempre recurrir a sus manuales de usuario, de tal forma que se puedan comprender de mejor manera los resultados que brinden dichos programas.

C. Evaluación del modelo

Para evaluar los modelos de clasificación de la minería de datos, y por tanto de los árboles de clasificación, se emplean diversos métodos, entre los cuales se pueden mencionar la matriz de confusión, la matriz de costos, el ***lift chart*** y la curva ROC.

Para el presente artículo se procederá a una breve explicación de la matriz de confusión, ya que es la de más fácil construcción e interpretación.

Matriz de confusión

Se procederá a la explicación de una matriz de confusión con base en los resultados de clasificación que se obtendrían a partir de una clase con 2 categorías: sí o no, compra, no compra, acepta o rechaza, etcétera. Además, una de las categorías de la clase debe ser considerada como la categoría de interés, la principal; la otra sería la categoría secundaria. Continuando con el ejemplo presentado, se determina que:

i. Categoría principal (+): A.
ii. Categoría secundaria (–): B.

Entonces, la clasificación realizada a una nueva instancia puede corresponder a uno de los siguientes 4 tipos de resultados: verdadero positivo, falso positivo, verdadero negativo y falso negativo (Witten y Frank 2005: 162). Estos tipos de resultados se muestran en la siguiente matriz de confusión:

		Clase predicha	
		A	**B**
Clase real	**A**	Verdadero positivo (VP)	Falso negativo (FN)
	B	Falso positivo (FP)	Verdadero negativo (VN)

Tabla 1. Matriz de confusión

Fuente: Adaptado de Witten y Frank (2005: 162).

Los verdaderos positivos (VP) y verdaderos negativos (VN) son las instancias clasificadas correctamente. Los falsos positivos (FP) son las instancias clasificadas como la categoría de interés (positivo) cuando es de la categoría secundaria (negativo). Los falsos negativos (FN) son las instancias clasificadas como la categoría secundaria (negativo) cuando es de la categoría principal (positivo) (Witten y Frank 2005: 162). Estos resultados, contextualizados al ejemplo aplicativo se presentan a continuación:

- Verdadero positivo (VP). Señala la cantidad de instancias clasificadas como la categoría A, perteneciendo realmente a dicha categoría.
- Verdadero negativo (VN). Señala la cantidad de instancias clasificadas como la categoría B, perteneciendo realmente a dicha categoría.
- Falso positivo (FP). Señala la cantidad de instancias clasificadas como la categoría A, perteneciendo realmente a la categoría B.
- Falso negativo (FN). Señala la cantidad de instancias clasificadas como la categoría B, perteneciendo realmente a la categoría A.

De acuerdo con estos tipos de resultados, usando la regla de decisión 2 del ejemplo aplicativo, se ha elaborado la siguiente matriz de confusión:

		Clase predicha	
		A	**B**
Clase real	**A**	48(VP)	2(FN)
	B	4(FP)	46(VN)

Tabla 2. Matriz de confusión – Ejemplo aplicativo (regla de decisión 2)

Elaboración propia.

De la matriz de confusión se pueden obtener algunas medidas, que permitirán medir el desempeño del árbol de clasificación obtenido (Ramírez 2010: 21-22)

Medida	**Forma de cálculo**		**Interpretación**
Éxito (Exactitud)	Clasificaciones acertadas con respecto al total de instancias.	$\text{éxito} = \frac{VP + VN}{\text{total}}$	Proporción de instancias bien clasificadas.
Error	Clasificaciones erradas con respecto al total de instancias.	$\text{error} = \frac{FP + FN}{\text{total}}$	Proporción de instancias mal clasificadas.
Sensibilidad	Clasificaciones acertadas con respecto al total de instancias de la categoría de interés.	$\text{sensibilidad} = \frac{VP}{VP + FN}$	Probabilidad de clasificar correctamente una instancia en la categoría de interés (+).
Especificidad	Clasificaciones acertadas con respecto al total de instancias de la categoría secundaria.	$\text{especificidad} = \frac{VN}{VN + FP}$	Probabilidad de clasificar correctamente una instancia en la categoría secundaria (–).

Tabla 3. Medidas de evaluación

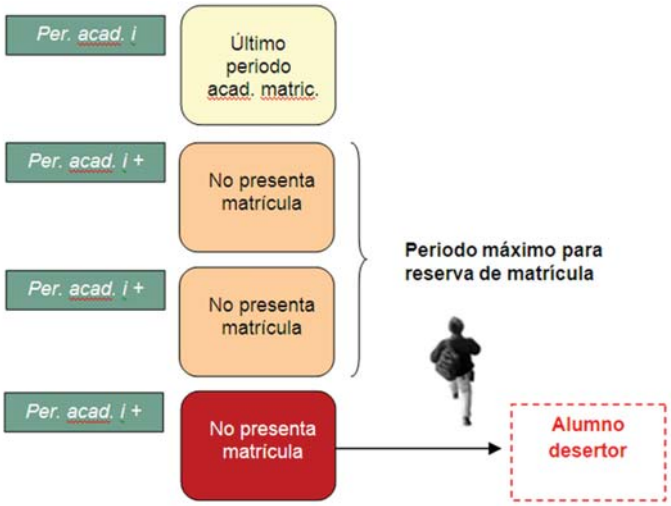
Elaboración propia.

3. Factores que determinan la deserción universitaria

Como se ha señalado anteriormente, la realidad de cada institución es muy disímil, debido a diversos motivos; por ejemplo, el análisis de la deserción en una universidad estatal presentará muchas diferencias con respecto a una universidad particu-

lar. Además, las diferencias también se pueden deber al alcance del estudio, entre otros múltiples motivos.

Como la finalidad primordial del presente artículo se centra en el análisis comparativo de 2 metodologías de clasificación, se hará uso de un caso específico del análisis de deserción en una entidad universitaria particular, trabajado para un curso de especialización realizado en la Universidad Nacional de Ingeniería. Barreno y Espíritu (2011) determinaron, de acuerdo con los requerimientos de su estudio, que un alumno es considerado desertor definitivo cuando presenta 3 periodos académicos consecutivos sin registrar matrícula (2011: 30). Asimismo, para dicho estudio no se consideraron como alumnos desertores a los alumnos expulsados.



Barreno y Espíritu analizaron la deserción universitaria en una determinada universidad particular, basándose en los factores que se detallan seguidamente.

3.1 Factores y variables que influyen en la deserción estudiantil

A continuación se detallan las variables consideradas para el análisis de la deserción estudiantil, variables respuesta y variables explicativas, que fueron clasificadas, estas últimas, de acuerdo con grupos de factores específicos.

- **Variable respuesta:**

Probabilidad de deserción:
- 1, Si el estudiante ha desertado,
- 0, En caso contrario

- **Variables explicativas:**

  En la tabla 4 se describen las variables explicativas utilizadas en el estudio, con su correspondiente clasificación de pertenencia (factores) y la fórmula asociada para su obtención, según sea necesaria, así como de los valores o intervalos de valores que pueden adoptar dichas variables:

Factor	Variable	Valores de la variable
Académico	Tipo colegio de procedencia (TCP).	1, Si proviene de colegio particular. 0, En caso contrario.

Factor	Variable	Valores de la variable
Académico	Efectividad Examen de Admisión (EEA). $EEA = \frac{\text{Puntaje obtenido Ex. Admisión}}{\text{Máximo puntaje posible}}$	[0;1]
Académico	Promedio Ponderado Acumulado (PPA). Media ponderada, según valor de créditos académicos de cada asignatura, de todas las notas obtenidas por el estudiante en las asignaturas que haya cursado hasta el periodo de análisis.	[0;20]
Académico	Proporción de créditos aprobados (PCA). $PCA = \frac{\text{Total créditos aprobados}}{\text{Total créditos matriculados}}$	[0;1]
Académico	Presenta acta de compromiso (AC). El acta de compromiso se extiende a aquellos alumnos con múltiples repitencias en una misma asignatura.	1, Si el alumno presenta acta de compromiso. 0, En caso contrario.

(continúa)

(continuación)

Factor	Variable	Valores de la variable
Económico	Solicitud de recategorización (SR). Se refiere a la solicitud que realiza el alumno para que se le cambie de categoría de pago de pensiones (cambio por una de menor importe económico).	1, Si el alumno ha solicitado recategorización. 0, En caso contrario.
Económico	Ingreso familiar (IF). Es una variable que se ha optado por colocar en categorías, ya que así fue recabada al momento de la inscripción para el proceso de admisión.	Menos de S/. 2000. Entre S/. 2000 y S/. 4000. Más de S/. 4000.

Tabla 4.Variables consideradas para el análisis de la deserción estudiantil

Fuente: Adaptado de Barreno y Espíritu (2011: 39).

4. Análisis comparativo de los modelos de clasificación

Para el análisis comparativo se ha adoptado la utilización de una similar metodología de trabajo en la aplicación de cada uno de los modelos de clasificación evaluados. Por lo tanto, se expondrán en forma resumida la aplicación de la siguiente metodología para cada uno de los modelos en análisis:

a) Adecuación de los datos de entrada.
b) Obtención del modelo preliminar.
c) Evaluación del modelo obtenido.
d) Reformulación y obtención del nuevo modelo.
e) Evaluación del modelo reformulado.
f) Interpretación del modelo.
g) Aplicación y prueba del modelo.

La última parte de la metodología es un estándar aplicado en estos tipos de análisis, tal como señalan Witten y Frank al referir que: "Filters are often applied to a training data set and then also applied to the test file"[1] (205: 393-394). Para cada modelo de clasificación evaluado se aplicará en su desarrollo la metodología señalada, donde cada punto de la metodología presentará sus propias peculiaridades según sea el modelo evaluado. Para el estudio realizado, y la correspondiente

1 Los filtros (algoritmos o modelos de clasificación) se aplican a menudo a un conjunto de datos de entrenamiento así como a datos de prueba entrenamiento, y luego también a datos de prueba.

comparación de los modelos de clasificación se utilizó una data de entrenamiento compuesta por 1059 registros, la cual presentaba 105 alumnos en condición de desertor definitivo; mientras que la data de prueba se encontraba compuesta por 820 registros con 98 alumnos desertores definitivos. Para la aplicación de ambos modelos se debe tener en cuenta la terminología que se especifica a continuación:

Modelo		Descripción
Regresión Logística	**Árbol de clasificación**	
Variable dependiente.	Clase.	Condición de desertor.
Categorías de la variable.	Categorías de la clase.	Desertor y No Desertor.
		Categoría de interés: Desertor. Categoría secundaria: No Desertor.

Tabla 5. Terminología utilizada en los modelos de clasificación en evaluación

Fuente: Elaboración propia.

4.1 Resultados obtenidos mediante la regresión logística

4.1.1 Regresión logística - Adecuación de los datos de entrada

El algoritmo de la regresión logística solamente acepta datos de entrada de tipo numérico, por ello, las variables cualitativas tienen que ser representadas de distinta forma para ser usadas por el algoritmo del modelo. Para las variables cualitativas que poseen solamente 2 categorías se hace uso de una variable binaria que toma valores 0 o 1, tal es el caso de las siguientes variables explicativas: TCP, AC, SR. Para el caso de variables cualitativas con 3 o más categorías se debe hacer uso de variables ***dummy*** para representar la variable "ingreso familiar". La representación utilizada, mediante variables ***dummy***, se presenta a continuación:

- ***Adecuación de la variable ingreso familiar***

  Para la adecuación de la variable cualitativa ingreso familiar se hace uso de las siguientes variables ***dummy***: IF1 y IF2, ambas binarias:

Categoría	IF1	IF2
Menos de S/. 2000	1	0
De S/. 2000 a S/. 4000	0	1
Más de S/. 4000	1	1

Tabla 6. Variables ***dummy*** para la variable ingreso familiar

Fuente: Barreno y Espíritu (2011: 39).

Cabe mencionar que la representación brindada no es la única, y que dependiendo de la representación utilizada, entonces, se deberá realizar su correspondiente interpretación.

4.1.2 Regresión logística - Obtención del modelo preliminar

Luego de procesar los datos en el software Minitab mediante la regresión logística, se obtuvieron los reportes correspondientes, de los cuales se presenta el siguiente extracto:

Tabla de regresión logística

Predictor	Coef	SE Coef	Z	P
Constant	31,6115	10,1069	3,13	0,002
TCP - T.Col.Proced.	0,721331	1,06006	0,68	0,496
EEA - Ef.Ex.Adm.	-3,11615	4,00140	-0,78	0,436
PPA - Prom.Pon.Ac.	-29,6558	15,2577	-1,94	0,052
PCA - Prop.Cred.Aprob.	-0,792092	0,797105	-0,99	0,320
AC - Acta.Compr.	2,47054	1,20539	2,05	0,040
SR - Sol.Recateg.	2,86075	1,24495	2,30	0,022
IF1 - Ing.Fam.1	-1,62847	1,35775	-1,20	0,230
IF2 - Ing.Fam.2	-2,81418	1,45181	-1,94	0,053

En ocasiones, no se obtienen buenos resultados, debido a la no convergencia de los resultados parciales en el proceso iterativo de solución, y se debe determinar que variable excluir del modelo, mediante otros criterios a fin de obtener un modelo que se pueda interpretar.

4.1.3 Regresión logística – Evaluación del modelo obtenido

Considerando las variables propuestas, el modelo obtenido no satisface la prueba individual en 4 de las 7 variables explicativas propuestas: TCP, EEA, PCA e ingreso

familiar (IF1 e IF2). Las variables mencionadas presentan un alto valor de P, razón por la cual se concluye que no son significativas para el modelo. A pesar de que la variable IF2 presenta un valor P aceptable (P value = 0,53 < 0,10), dicha variable tiene que ser evaluada en forma conjunta con la variable ***dummy*** (IF1), la cual presenta un alto valor P; por lo tanto, en conjunto dichas variables no son significativas para el modelo.

Las variables explicativas PPA, AC y SR, de acuerdo con sus valores P, se consideran significativas para el modelo. Estas conclusiones se ven reforzadas mediante la obtención de modelos donde se fueron retirando, de una a la vez, las variables explicativas con el peor valor P, es decir, se obtuvo un modelo utilizando las mismas variables, con excepción de TCP, la cual fue excluida en primera instancia al poseer el mayor valor P, y así se procedió en forma sucesiva.

4.1.4 Regresión logística – Reformulación y obtención del nuevo modelo

Luego de lo señalado anteriormente, se obtuvo un modelo considerando solo las variables PPA, SR y AC. A continuación se muestra un extracto del reporte obtenido:

Regresión logística binaria: Deserción vs. PPA, SR, AC

Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Prob.**
Constant	20,0187	5,01396	3,99	0,000	
PPA - Prom.Pon.Ac.	-2,31310	0,539404	-4,29	0,000	0,099
SR - Sol.Recateg.	2,63139	1,04469	2,52	0,012	13,893
AC - Acta.Compr.	2,44440	0,938926	2,60	0,009	11,524

Prueba que todas las pendientes son iguales a cero:
G = 188,180, DF = 3, P-Value = 0,000

4.1.5 Regresión logística – Evaluación del modelo reformulado

Todas las variables consideradas en el modelo final: PPA, SR y AC satisfacen la prueba individual, ya que presentan valores de P reducidos; adicionalmente, se apreció que dichas variables presentaban valores de ***odds ratio*** diferentes de la unidad. De acuerdo con la prueba global (prueba que todas las pendientes son iguales a cero), se concluye que el modelo en su conjunto es apropiado, ya que presentó un valor P igual a cero.

4.1.6 Regresión logística – Interpretación del modelo

El modelo matemático de regresión logística finalmente obtenido queda expresado de la siguiente forma:

$$\text{Prob(yi= desertor definitivo)} = \frac{1}{1+e^{-(20{,}0187-2{,}3131*\text{PPA}+2{,}63139*\text{SR}+2{,}4444*\text{AC})}}$$

Interpretaciones del modelo:

a) ***Promedio ponderado acumulado (PPA):*** El valor negativo (–) del coeficiente asociado a la variable explicativa PPA, significa que si aumenta el promedio ponderado acumulado, manteniéndose constante las demás variables, entonces, disminuye la probabilidad de que el alumno sea un desertor definitivo, es decir, disminuye P(Y = 1).

b) ***Solicitud de recategorización (SR):*** El valor (+) del coeficiente asociado a la variable explicativa SR, significa que si el alumno solicita recategorización en el pago de pensiones, manteniéndose constante las demás variables, entonces, aumenta la probabilidad de que sea un desertor definitivo, es decir, aumenta P(Y = 1).

c) Acta de compromiso (AC): El valor (+) del coeficiente asociado a la variable explicativa AC, significa que si el alumno posee acta de compromiso, manteniéndose constante las demás variables, entonces, aumenta la probabilidad de que sea un desertor definitivo, es decir, aumenta P(Y = 1).

Para utilizar el modelo obtenido, solamente es necesario reemplazar los valores que toman las variables correspondientes, asociadas a cualquier alumno de la facultad en estudio, en la expresión matemática señalada, lo cual se puede implementar rápidamente en los programas informáticos actuales. Enseguida, se aplicará el siguiente criterio para su interpretación:

i. Si la probabilidad de Prob($y_i = 1$) es menor a 0,50, entonces, el alumno es considerado como un no desertor.

ii. Si la probabilidad de Prob($y_i = 1$) es igual o mayor a 0,50; entonces, el alumno es considerado (clasificado) como un posible a futuro desertor definitivo.

4.1.7 Regresión logística – Aplicación y prueba del modelo

De acuerdo con la aplicación del modelo de regresión logística, finalmente obtenido en la data de prueba, y conforme a los criterios de interpretación de resultados, probabilidades menores o mayores iguales a 0,50; se obtuvo la matriz de confusión y la interpretación de los valores que aparecen en dicha matriz, tal como se presentan en las siguientes tablas:

		Clasificación	
		Desertor	**No desertor**
Valor Real	**Desertor**	87	11
	No desertor	36	686

Tabla 7. Matriz de confusión – Modelo de regresión logística aplicado a la data de prueba

Elaboración propia.

Cantidad de alumnos clasificados como desertores definitivos, perteneciendo realmente a dicha categoría.	VP	87
Cantidad de alumnos clasificados como no desertores, siendo realmente desertores definitivos.	FN	11
Cantidad de alumnos clasificados como desertores definitivos, siendo realmente no desertores.	FP	36
Cantidad alumnos clasificados no desertores, perteneciendo realmente a dicha categoría.	VN	686

Tabla 8. Interpretación de los valores de la matriz de confusión – Modelo de regresión logística aplicado a la data de prueba

Elaboración propia.

Asimismo, se obtuvieron las siguientes medidas de evaluación: error, éxito, sensibilidad y especificidad; las cuales fueron determinadas a partir de los valores presentados en la matriz de confusión. En la siguiente tabla se presentan dichas medidas, así como la correspondiente interpretación:

Medida de evaluación y forma de cálculo	Interpretación
$\text{error} = \frac{FP + FN}{total} * 100\% = \frac{36 + 11}{820} * 100\% \approx 5{,}732\%$	5,73% de alumnos erradamente clasificados.
$\text{éxito} = \frac{VP + VN}{total} * 100\% = \frac{87 + 686}{820} * 100\% \approx 94{,}286\%$	94,27% de alumnos exitosamente clasificados.
$\text{sensibilidad} = \frac{VP}{VP + FN} = \frac{87}{87 + 11} \approx 0{,}8878$	La probabilidad de clasificar correctamente a un alumno como desertor definitivo es de 0,888.
$\text{especificidad} = \frac{VN}{VN + FP} = \frac{686}{686 + 36} \approx 0{,}9501$	La probabilidad de clasificar correctamente a un alumno como no desertor es de 0,950.

Tabla 9. Medidas de evaluación – Modelo de regresión logística aplicado a la data de prueba

Elaboración propia.

4.2 Resultados obtenidos mediante el árbol de clasificación

4.2.1 Árbol de clasificación - Adecuación de los datos de entrada

El algoritmo del árbol de clasificación acepta datos de entrada de tipo numérico y de tipo categórico, así que las variables cualitativas pueden ser utilizadas directamente sin transformación alguna, lo cual no implica que no se puedan hacer transformaciones a la data de entrada; por ejemplo, alguna variable cuantitativa puede ser discretizada si es que así se considera conveniente. Esta discretización se realiza con fines de obtener un mejor modelo, de ser posible, y no es requisito para la aplicación del modelo. Para el presente caso de estudio no se discretizó ninguna de las variables explicativas, del tipo cuantitativo, que se propusieron.

4.2.2 Árbol de clasificación - Obtención del modelo preliminar

Se procesaron los datos en el software Weka, versión 3-7-2, mediante el árbol de clasificación J48, que, tal como señala Bouckaert et al., es una implementación del Weka para el algoritmo C4.5 desarrollado por Quinlan en 1993, el cual es uno de los

algoritmos más utilizados (2012: 19). Para el procesamiento respectivo, se usaron los parámetros por defecto del Weka con las siguientes modificaciones:

i. Obtención de un árbol no podado (Unpruned: True).

ii. Número mínimo de instancias por hoja: 3 (minNumObj: 3).

Las modificaciones señaladas se aplicaron con la finalidad de obtener un árbol más "frondoso" y poder apreciar así el poder discriminativo de la mayor cantidad de las variables. Luego de realizado el procesamiento se obtuvo el siguiente árbol de clasificación, además de algunos reportes que se detallan en el siguiente apartado (evaluación del modelo):

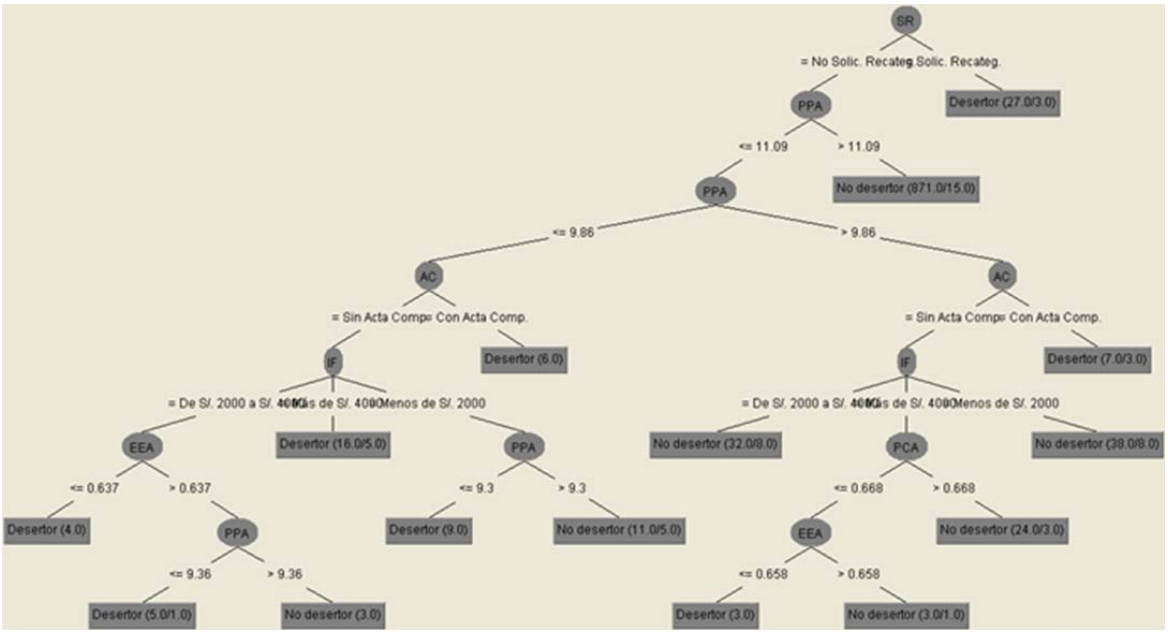


Figura 6. Árbol de clasificación (weka.classifiers.trees.J48 -U -M 3)

Elaboración propia. Reporte gráfico del Weka.

4.2.3 Árbol de clasificación – Evaluación del modelo obtenido

El árbol obtenido presenta 15 hojas, nodos donde se realiza la clasificación; y 12 nodos de decisión, donde se evalúa el valor de alguna de las variables.

```
=== Summary ===
Correctly Classified Instances        1007         95,0897 %
Incorrectly Classified Instances        52          4,9103 %
Kappa statistic                          0,6881
Mean absolute error                      0,0787
Root mean squared error                  0,1984
Relative absolute error                 43,9309 %
Root relative squared error             66,3925 %
Coverage of cases (0,95 level)          98,5836 %
Mean rel. region size (0,95 level)      57,6959 %
Total Number of Instances             1059
=== Detailed Accuracy By Class ===
     TP Rate FP Rate Precision Recall F-Measure ROCArea Class
       0,987   0,381    0,959   0,987    0,973    0,909 No desertor
       0,619   0,013    0,844   0,619    0,714    0,909 Desertor
W. A.  0,951   0,344    0,948   0,951    0,947    0,909
=== Confusion Matrix ===
   a   b   <-- classified as
942  12 |   a = No desertor
 40  65 |   b = Desertor
```

El árbol de clasificación obtenido presenta un 95,09% de clasificaciones correctas, siendo las variables con mayor poder discriminativo: SR, PPA, AC, IF, EEA y PCA. La variable TCP cuenta con un nulo poder discriminativo al no formar parte del árbol. El árbol obtenido utiliza 6 de las 7 variables explicativas en los nodos de decisión, que puede ser indicio de sobre ajuste, lo cual no es muy conveniente. Tal como señala Larose, "Also, retaining too many variables may lead to overfitting, in which the generality of the findings is hindered because the new data do not behave the same as the training data for all the variables" (2006: 1-2), es decir, afecta a la generalidad de su aplicación en datos diferentes a los de la data de entrenamiento.

4.2.4 Árbol de clasificación – Reformulación y obtención del nuevo modelo

Por lo expuesto, se optó por la búsqueda de un árbol que haga uso de un menor número de variables explicativas; luego de algunas pruebas, se obtuvo el árbol que se aprecia a continuación:

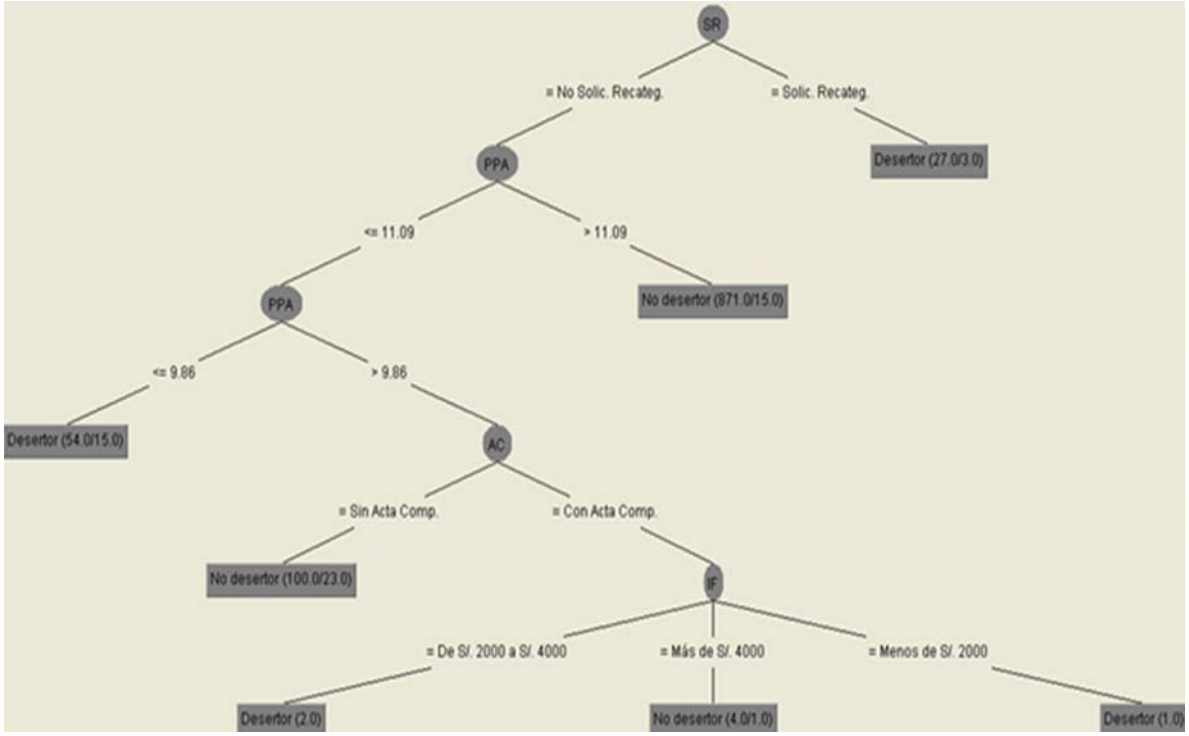


Figura 7. Árbol de clasificación (weka.classifiers.trees.J48 -C 0.25 –M

Elaboración propia. Reporte gráfico del Weka.

4.2.5 Árbol de clasificación – Evaluación del modelo reformulado

El árbol obtenido presenta 7 hojas, nodos donde se realiza la clasificación; y solamente 5 nodos de decisión, donde se evalúa el valor de alguna de las variables.

```
=== Summary ===
Correctly Classified Instances        1002          94,6176 %
Incorrectly Classified Instances        57           5,3824 %
Kappa statistic                          0,6693
Mean absolute error                      0,0882
Root mean squared error                  0,21
Relative absolute error                 49,2063 %
Root relative squared error             70,2658 %
Coverage of cases (0,95 level)          98,5836 %
Mean rel. region size (0,95 level)      58,7347 %
Total Number of Instances             1059
=== Detailed Accuracy By Class ===
     TP Rate FP Rate Precision Recall F-Measure ROCArea Class
       0,981   0,371    0,960   0,981    0,970    0,903 No desertor
       0,629   0,019    0,786   0,629    0,698    0,903 Desertor
W. A.  0,946   0,336    0,943   0,946    0,943    0,903
=== Confusion Matrix ===
   a   b   <-- classified as
936  18 |   a = No desertor
 39  66 |   b = Desertor
```

El nuevo árbol de clasificación presenta un 94,62% de clasificaciones correctas, casi 0,5% menos que el árbol de clasificación inicial; las variables con mayor poder de discriminación son: SR, PPA, AC e IF; por lo tanto, solamente se utilizan 4 de las 7 variables explicativas en los nodos de decisión.

4.2.6 Árbol de clasificación – Interpretación del modelo

El árbol de decisión obtenido, a diferencia del modelo de regresión logística, no se representa mediante una fórmula matemática, en su lugar hace uso de reglas de decisión, asociadas a las principales variables explicativas: SR, PPA, AC e IF. Estas reglas de decisión, las mismas que se observan en el gráfico del árbol de decisión, también son brindadas por el software, como se detalla a continuación:

```
SR = No Solic. Recateg.
|   PPA <= 11,09
|   |   PPA <= 9,86: Desertor (54,0/15,0)
|   |   PPA >9,86
|   |   |   AC = Sin Acta Comp.: No desertor (100,0/23,0)
|   |   |   AC = Con Acta Comp.
|   |   |   |   IF = De S/. 2000 a S/. 4000: Desertor (2,0)
|   |   |   |   IF = Más de S/. 4000 : No desertor (4,0/1,0)
|   |   |   |   IF = Menos de S/. 2000 : Desertor (1,0)
|   PPA > 11,09: No desertor (871,0/15,0)
SR = Solic. Recateg.: Desertor (27,0/3,0)
```

Para utilizar el modelo obtenido es necesario evaluar los valores que toman las variables explicativas, de acuerdo con las reglas de decisión señaladas; la evaluación se realiza según la prioridad de cada regla de decisión, es decir que solamente se realizarán tantas evaluaciones como sean necesarias hasta que se obtenga una clasificación para el alumno en evaluación. A modo de ejemplo, si se tuviera la siguiente información de un alumno, actualizada en un determinado período de análisis: TCP: colegio particular, EEA: 0,750; PPA: 11,012; PCA: 0,850; AC: sin acta de compromiso; SR: no solicitó recategorización e IF: más de S/. 4000, aplicando las reglas de decisión, a dicho alumno se le clasificaría como un "no desertor" a futuro.

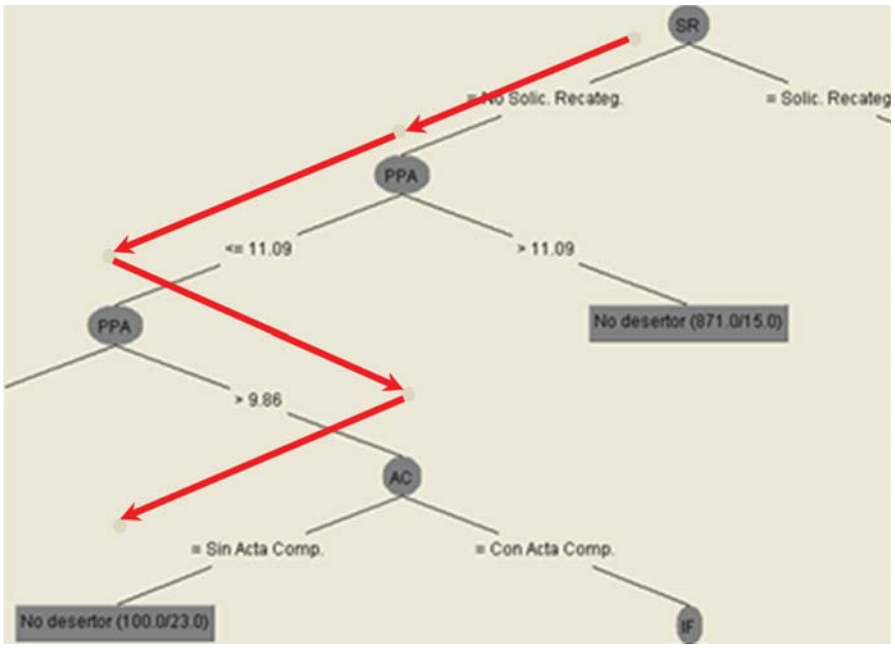


Figura 8. Uso del árbol de clasificación

Elaboración propia.

Esta aplicación a un determinado alumno se ha realizado en forma gráfica; si se deseara su aplicación en forma sistematizada, y si son pocas las reglas de decisión, entonces, estas pueden ser plasmadas sin dificultad, por ejemplo, en hojas de cálculo. Sin embargo, al incrementarse las reglas de decisión, lo más recomendable es utilizar un código de programación, donde se haga uso del seudocódigo brindado por el software Weka.

4.2.7 Árbol de clasificación – Aplicación y prueba del modelo

De acuerdo con la aplicación de las reglas de decisión determinadas por el árbol de clasificación, en la data de prueba, y conforme a las clasificaciones realizadas, se obtuvo la matriz de confusión, y la interpretación de los valores que aparecen en la matriz, así como las medidas de evaluación a partir de los datos de dicha matriz.

		Clasificación	
		Desertor	**No desertor**
Valor Real	**Desertor**	84	14
	No desertor	35	687

Tabla 10. Matriz de confusión – Árbol de clasificación aplicado a la data de prueba

Elaboración propia.

Cantidad de alumnos clasificados como desertores definitivos, perteneciendo realmente a dicha categoría.	VP	84
Cantidad de alumnos clasificados como no desertores, siendo realmente desertores definitivos.	FN	14
Cantidad de alumnos clasificados como desertores definitivos, siendo realmente no desertores.	FP	35
Cantidad alumnos clasificados no desertores, perteneciendo realmente a dicha categoría.	VN	687

Tabla 11. Interpretación de los valores de la matriz de confusión – Árbol de clasificación aplicado a la data de prueba

Elaboración propia.

Medida de evaluación y forma de cálculo	Interpretación
$\text{error} = \frac{FP + FN}{total} * 100\% = \frac{35 + 14}{820} * 100\% \approx 5{,}976\%$	5,98% de alumnos erradamente clasificados
$\text{éxito} = \frac{VP + VN}{total} * 100\% = \frac{84 + 687}{820} * 100\% \approx 94{,}024\%$	94,02% de alumnos exitosamente clasificados
$\text{sensibilidad} = \frac{VP}{VP + FN} = \frac{84}{84 + 14} \approx 0{,}8571$	La probabilidad de clasificar correctamente a un alumno como desertor definitivo es de 0,857.
$\text{especificidad} = \frac{VN}{VN + FP} = \frac{687}{687 + 35} \approx 0{,}9515$	La probabilidad de clasificar correctamente a un alumno como no desertor es de 0,952.

Tabla 12. Medidas de evaluación – Árbol de clasificación aplicado a la data de prueba

Fuente: Elaboración propia.

Como se puede apreciar en los resultados obtenidos, al aplicar el árbol de clasificación en la data de prueba, no difieren significativamente de los resultados obtenidos mediante la aplicación de la regresión logística.

4.3 Comparación de los modelos de clasificación

Dado que la metodología de análisis de un fenómeno como la deserción estudiantil es muy compleja, y la forma de analizarla es bastante diversa, no se puede indicar una única forma de análisis, pues esta dependerá de los objetivos planteados para dicho análisis; lo mismo ocurre con su alcance y con los recursos disponibles, en especial el tiempo y los recursos económicos. De igual forma, la determinación del método de clasificación más apropiado para dicho análisis o cualquier otro donde se tenga la necesidad de realizar clasificaciones de nuevos individuos, dependerá también de diversos factores, resultando a veces más conveniente un modelo de clasificación con respecto a otros modelos, que pueden variar de una realidad a otra.

Por eso, el análisis comparativo realizado servirá de apoyo para que los investigadores que así lo requieran, comparen diversos modelos de clasificación y no solamente los expuestos en el presente trabajo, y puedan seleccionar alguno de ellos para su aplicación práctica en su caso de estudio. Para el análisis comparativo de los modelos de clasificación en análisis: regresión logística y árbol de clasifica-

ción, se hará uso de la metodología de aplicación utilizada para ambos modelos; en cada paso de la metodología se brindará una apreciación comparativa de cómo se realizó la aplicación correspondiente.

	Regresión logística	**Árbol de clasificación**
Adecuación de los datos de entrada	Los datos cualitativos deben ser transformados previamente a ser usados por el algoritmo de clasificación. La transformación de las variables cualitativas de 2 categorías es de fácil interpretación, más no así cuando se hace uso de variables *dummy* en variables de 3 a más categorías.	No necesita, en forma obligatoria, de una adecuación previa de los datos, pero sí puede realizarse en beneficio de la calidad de los resultados. Por ejemplo: discretización de datos cuantitativos.
Obtención del modelo preliminar	En ocasiones no se obtienen buenos resultados (no convergencia del procedimiento).	Siempre se logra obtener un árbol de clasificación, con reglas de decisión claras. Se deben definir algunos parámetros para iniciar el algoritmo, para lo cual se requiere un mayor conocimiento del modelo.
Evaluación del modelo obtenido	Se evalúa en forma estadística, y se puede evaluar cada variable y al modelo en su conjunto, mediante una diversidad de pruebas. Además, se puede hacer uso de la matriz de confusión.	El modelo no puede ser evaluado en forma estadística de manera formal; principalmente se hace uso de la matriz de confusión.
Reformulación y obtención del nuevo modelo	Mediante una simple inspección de los reportes se puede realizar una rápida reformulación del modelo.	La reformulación del modelo depende en gran medida de la realización de varias pruebas, en las cuales se van modificando los diversos parámetros hasta obtener un árbol apropiado. Se debe evitar el sobre y el subaprendizaje.
Evaluación del modelo reformulado	Similar y más rápido que la evaluación del modelo original.	Similar y más rápido que la evaluación del modelo original.

(continúa)

(continuación)

	Regresión logística	Árbol de clasificación
Interpretación del modelo	Al brindar una fórmula matemática basada en la función exponencial, hace que su interpretación no sea tan intuitiva como la de la regresión lineal. Esta interpretación podría hacerse más complicada si en el modelo final se hace uso de variables ***dummy***. Para utilizar el modelo solamente se deben reemplazar los valores que toman las variables correspondientes en la fórmula matemática obtenida.	El árbol de clasificación hace uso de sencillas reglas de decisión, de fácil evaluación; evaluando un atributo a la vez (existen otros algoritmos que pueden crear reglas más elaboradas). Para utilizar el modelo obtenido se evalúan los valores que toman las variables correspondientes, de acuerdo con las reglas de decisión señaladas.
Aplicación y prueba del modelo	Al aplicar el modelo, con la data de prueba, se evalúan los resultados mediante la matriz de confusión. La aplicación del modelo en nueva data brindará una probabilidad de que un alumno sea un desertor potencial a futuro, lo cual permitiría una subclasificación de acuerdo con la probabilidad.	Al aplicar el modelo, con la data de prueba, se evalúan los resultados mediante la matriz de confusión. La aplicación del modelo en nueva data brindará solamente una clasificación, y no se puede determinar si un alumno es más o menos propenso a convertirse en desertor potencial.

Tabla 13. Análisis comparativo – Regresión logística vs. Árbol de clasificación

Elaboración propia.

5. Conclusiones y recomendaciones

De acuerdo con la aplicación de los modelos de clasificación de regresión logística y árbol de clasificación en un mismo caso de estudio: la reducción de la deserción estudiantil, así como de un análisis comparativo de los resultados y de la aplicabilidad de ambos modelos, se exponen las siguientes conclusiones y recomendaciones.

5.1 Conclusiones

a) La aplicación de los modelos de clasificación obtenidos (regresión logística y árboles de clasificación) brindaron similares resultados, en ambos casos con un porcentaje de clasificación exitosa de por lo menos 94% en el análisis de la deserción universitaria, considerando determinadas variables explicativas.

b) La determinación del tipo de modelo de clasificación más conveniente para el análisis de un determinado fenómeno deberá determinarse a partir de la aplicación y comparación de modelos, obtenidos mediante una data de entrenamiento, en una data de prueba, estableciendo previamente cuáles serán los tipos de modelos por evaluar.

c) La elección de los tipos de modelos de clasificación a comparar, además, dependerá de los requerimientos del estudio; por ejemplo, si se hubiera considerado necesaria la determinación de alumnos desertores potenciales de mayor riesgo (mayor probabilidad), esto no hubiera sido posible por medio de un árbol de clasificación. Para esta situación se hubieran tenido que seleccionar modelos de clasificación que brinden probabilidades, por ejemplo, una red neuronal.

d) El modelo de más fácil implementación es el obtenido en la regresión logística, ya que solamente se encuentra determinada por una fórmula matemática; mientras que las reglas de decisión, obtenidas de la aplicación de un árbol de clasificación, requerirán de un mayor esfuerzo, en especial si el árbol consta de varias reglas de decisión.

5.2 Recomendaciones

a) Para la obtención del modelo, generalmente se hará uso de software libre o comercial, ya que es mucho más práctico que implementar los algoritmos correspondientes dentro del mismo sistema de la institución que desea realizar el estudio. Lo que sí es recomendable que se realice en el sistema de la institución, es la implementación del modelo obtenido, para determinar rápidamente a los posibles estudiantes desertores.

b) Como todo modelo, sea cual fuere el modelo de clasificación elegido para el análisis de determinada problemática, siempre se recomienda una revisión periódica, ya que los comportamientos que intenta predecir el modelo pueden variar con el trascurso del tiempo.

Bibliografía

Barreno, Emma y Gustavo Espíritu (2011). "Reducción de la deserción estudiantil en una facultad de un centro de estudios universitarios". Trabajo de diplomatura. Lima: Universidad Nacional de Ingeniería

Bouckaert, Remco; Frank, Eibe; Hall, Mark; Kirkby, Richard; Reutemann, Peter; Seewald, Alex y David Scuse (2012). *Weka manual for version 3-7-6*. Universidad de Waikato, Hamilton. <http://ufpr.dl.sourceforge.net/project/weka/documentation/3.7.x/WekaManual-3-7-6.pdf >. [Consulta: 15 de marzo del 2012].

Cuadras, Carles (2012). *Nuevos métodos de análisis multivariante*. Barcelona: CMC Editions. http://www.ub.edu/stat/personal/cuadras/metodos.pdf. [Consulta: 24 de febrero del 2012].

Giraldo, Juan (2009). "Caracterización de algunas técnicas algorítmicas de la inteligencia artificial para el descubrimiento de asociaciones entre variables y su aplicación en un caso de investigación específico" Tesis para optar el grado de magíster. Facultad de Minas, Universidad Nacional de Colombia, Medellín. <http://www.bdigital.unal.edu.co/2272/1/71741491.2009.pdf>. [Consulta: 01 de marzo del 2012]

Guzmán, Sandra (2009). "Deserción y retención estudiantil en los programas de pregrado de la Pontificia Universidad Javeriana". Tesis para optar el grado de doctor. Pontificia Universidad Javeriana. <http://www.javeriana.edu.co/biblos/tesis/educacion/tesis81.pdf>. [Consulta: 30 de enero del 2012].

Hernández, José; Ramírez, M.ª José y César Ferri (2004). *Introducción a la minería de datos***.** Madrid: Pearson Educación, S.A.

Ibarra, María del Carmen y Juan Michalus (2010). "Análisis del rendimiento académico mediante un modelo logit". *Ingeniería Industrial* 9. Lima: Universidad de Lima, pp. 47-56. http://www.ici.ubiobio.cl/revista/index.php?option=com_docman&task=doc_download&gid=110&&Itemid=15. [Consulta: 5 de marzo del 2012].

Juárez, O. y Ernestina Castells (2010). "Modelos de árbol de regresión bayesiano: un estudio de caso". *Investigación Operacional* 31. Lima: Universidad de Lima, pp. 109-125. http://rev-inv-ope.univ-paris1.fr/files/31210/31210-02R.pdf. [Consulta: 10 de marzo del 2012].

Larose, Daniel (2006). *Data mining: Methods and models*. Nueva Jersey: Wiley-Interscience. <http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/Texts/%5B6%5D%202006%20Data%20Mining%20Methods%20and%20Models.pdf>. [Consulta: 12 de marzo del 2012].

Ministerio de Educación Nacional (2009). *Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención*. Bogotá: Imprenta Nacional de Colombia. <http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf>. [Consulta: 24 de enero del 2012].

Ramírez, Juliana (2010). "Regularización y métodos kernel para algoritmos de clasificación". Tesis para optar el grado de magíster. Universidad Nacional de Colombia, Manizales. <http://www.bdigital.unal.edu.co/1991/1/julianaramirezcandamil.2010.pdf>. [Consulta: 11 de marzo del 2012].

Serna, Sandra (2009). "Comparación de árboles de regresión y clasificación y regresión logística". Tesis para optar el grado de magíster. Universidad Nacional de Colombia, Medellín. <http://www.bdigital.unal.edu.co/671/1/42694070_2009.pdf>. [Consulta: 27 de enero del 2012].

Sineace (2010). Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa. "Propuesta del glosario de términos básicos de evaluación, acreditación y certificación del Sineace". Lima.

Uriel, Ezequiel y Joaquín Aldás (2005). *Análisis multivariante aplicado: aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo*. Madrid: Thomson-Paraninfo.

Witten, Ian y Eibe Frank (2005). *Data mining. Practical machine learning tools and techniques*. California: Elsevier Publishing. <http://177.101.20.73/docs/WittenFrank.pdf>. [Consulta: 14 de marzo del 2012].