redie Revista Electrónica de Investigación Educativa
ISSN: 1607-4041
https://redie.uabc.mx

Vol. 23, 2021/e29

Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa

Patterns to Identify Dropout University Students with Educational Data Mining

Argelia Berenice Urbina-Nájera (1) https://orcid.org/0000-0002-3700-7287
Arturo Téllez-Velázquez (2) https://orcid.org/0000-0002-0787-7417
Raúl Cruz Barbosa (3) https://orcid.org/0000-0002-0677-3931

(1) Universidad Popular Autónoma del Estado de Puebla
(2) Universidad Tecnológica de Tlaxcala
(3) Universidad Tecnológica de la Mixteca
(Recibido: 2 de marzo de 2020; Aceptado para su publicación: 4 de diciembre de 2020)

Cómo citar: Urbina-Nájera, A. B., Téllez-Velázquez, A. y Cruz, R. (2021). Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa. *Revista Electrónica de Investigación Educativa, 23*, e29, 1-15. https://doi.org/10.24320/redie.2021.23.e29.3918

Resumen

En este trabajo se presenta un análisis de las características más relevantes de un potencial desertor universitario, mediante la aplicación de algoritmos de minería de datos educativa. Se utilizó un conjunto de datos de 10 635 instancias, adquiridas en el período 2014-2019, de 53 programas de licenciatura de una institución privada del estado de Puebla (México). Los resultados muestran que el modelo obtenido por los árboles de decisión ofrece mayor desempeño que otros algoritmos, así como una fácil interpretación de éste mediante reglas de decisión. Además, el rendimiento del modelo es mejor que otros modelos relacionados en la literatura aplicados al mismo problema. Los métodos de selección de características permitieron encontrar los atributos más importantes que identifican a un potencial desertor, tales como: el período, el último semestre cursado, créditos cursados, asistencia, materias reprobadas y programa. Utilizando los atributos y reglas de decisión encontradas se podrían crear mecanismos que favorezcan la prevención de la deserción.

Palabras clave: deserción escolar, características de la deserción, toma de decisiones

Abstract

This paper applies educational data mining algorithms to present an analysis of the most relevant characteristics of potential dropout students. The study used a dataset of 10,635 instances, acquired between 2014 and 2019 from 53 bachelor's degree programs at a private university in the state of Puebla (Mexico). The results show that the model obtained from the decision trees performs better than other algorithms and allows for easy interpretation through decision rules. Furthermore, the model performs better than other related models in the literature that have been applied to the same problem. The methods used to select characteristics yielded the most important attributes to identify potential dropouts, such as the period, last semester completed, credits completed, attendance, courses failed, and program. These attributes and decision rules can be used to create mechanisms that help prevent dropout.

Keywords: dropping out, dropouts characteristics, decision making

I. Introducción

La deserción escolar universitaria es una situación que ha tomado especial atención desde hace más de una década tanto en países desarrollados como en vías de desarrollo. De acuerdo con la Real Academia Española de la Lengua (2013) la deserción escolar es la acción de separarse o abandonar obligaciones en cuanto a compromisos escolares. Mientras que la Secretaría de Educación Pública [SEP], (2019a) lo expresa como el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo, originada por diversos factores intrínsecos (personales) y extrínsecos (familiares, personales, sociales, económicos). Para efectos de este estudio se utiliza el término apegado a la definición de la SEP.

Si bien la deserción escolar no obedece a una sola causa, sí hay una razón que origina la decisión de desertar; Aulck et al. (2017); Carvajal y Trejos (2016); López y Beltrán (2017); Márquez-Vera et al. (2016); Muñoz et al. (2018); Ramírez et al. (2016) y Sivakumar et al. (2016) afirman que las causas de este fenómeno se debe a factores religiosos, académicos, económicos, personales/familiares, sociales, institucionales, desempeño obtenido, si la madre no tiene estudios de posgrado, si el estudiante trabaja, si está presente alguna adicción (cigarro o alcohol), la distancia (hogar-escuela), el lugar de residencia, tipo de familia, satisfacción del curso, experiencia estresante familiar, infraestructura de la universidad, participación en actividad extracurricular, entorno del campus, cambio de objetivos, el bajo nivel de los estudiantes al ingresar a la institución, nulo interés vocacional, actitudes del estudiante, expectativas, la incorrecta elección de la carrera, o bien, el curso y tipo de período, entre otros.

De acuerdo con el reporte anual sobre las principales cifras del sistema educativo nacional, en el nivel superior los números estimados sobre deserción escolar en el ciclo escolar 2014-2015 alcanzaron un 12.6%, en el ciclo 2015-2016 un 12.1%, en el ciclo 2016-2017 un 7.2%; en el ciclo 2017-2018 aumentó a 8.4% y finalmente en el ciclo 2018-2019 se tuvo un ligero descenso al obtener un 8.3% sólo en la modalidad escolarizada (SEP, 2019b). Estas cifras indican que México tiene la proporción más baja de población con educación superior de todos los países de la OCDE (2019).

Para estudiar el fenómeno de la deserción escolar es pertinente identificar todos aquellos factores que intervienen en la decisión de abandonar los estudios universitarios. A partir de estudios recientes se ha determinado que las tecnologías que ayudan a analizar situaciones en este contexto es la minería de datos educativa (MDE o *Educational Data Mining*) y el aprendizaje computacional (*machine learning*) que han contribuido a la manera en que las instituciones de educación superior dan seguimiento y predicen el desempeño de los estudiantes (Clow, 2013).

La minería de datos educativa se define como un área interdisciplinaria emergente que trata con el desarrollo de métodos para explorar datos que se originan en el contexto educativo (Romero y Ventura, 2010). De esta forma, la MDE utiliza al aprendizaje computacional para analizar los datos relacionados con problemas en el contexto educativo. Dicho aprendizaje se define como una rama de la inteligencia artificial que permite programar a las computadoras para optimizar un criterio de rendimiento, utilizando datos de ejemplo o experiencia (Alpaydin, 2010). En otras palabras, esta área se dedica al estudio de los agentes o programas que aprenden o evolucionan basados en su experiencia para realizar una tarea determinada cada vez mejor (Mitchell, 2000).

Algunos métodos (algoritmos) de aprendizaje computacional se han aplicado para resolver problemas en este contexto, por ejemplo: se han utilizado árboles de decisión para predecir el desempeño de los estudiantes (Agaoglu, 2016; Al-Barrak y Al-Razgan, 2016; Chiheb et al., 2017; Yamao et al., 2018), medir el éxito académico (Morales y Parraga-Alava, 2018), captación de matrícula en Instituciones de Educación Superior (IES) particulares (Estrada et al., 2016) e identificación de perfiles de comportamiento (Guevara et al., 2019).

Para explicar las causas de la deserción escolar también se han aplicado algoritmos como Näive Bayes (Márquez-Vera et al., 2016), redes neuronales (Yukselturk et al., 2014), vecinos más cercanos (Aulck et al., 2017), regresión lineal o logística, máquinas de soporte vectorial (Agaoglu, 2016; Al-Barrak y Al-Razgan,

2016; Chihebet al., 2017; Yamao et al., 2018). Finalmente, el algoritmo selección de atributos ha sido empleado para identificar a los mejores atributos que pueden contribuir a una mejor clasificación (Márquez-Vera et al., 2016).

Este estudio tiene como objetivo presentar un análisis cuantitativo de los atributos que identifican a un estudiante desertor por medio de los algoritmos de minería de datos educativa.

1.1 Algoritmos de clasificación supervisada

La clasificación de instancias es aquella clase distinguida a partir de un conjunto de atributos. El aprendizaje computacional es utilizado para identificar patrones ocultos en grandes volúmenes de datos que pueden obtenerse a través del historial del estudiante, desde su ingreso a la universidad hasta obtener un título (Carvajal y Trejos, 2016; Witten et al., 2016).

Uno de los algoritmos aplicados es Bayes ingenuo (Theodoridis y Koutroumbas, 2008), el cual utiliza la regla de Bayes para predecir la pertenencia de una nueva muestra con alguna clase a partir de un conjunto de datos de entrenamiento, asumiendo que existe un modelo de probabilidad subyacente en los datos. Dicho modelo también supone que existe independencia entre los atributos. Otro algoritmo es el de árboles de decisión: "un modelo predictivo con el que se pueden representar clasificaciones y modelos de regresión; empleado para identificar la estrategia más probable para alcanzar el objetivo" (Rokach y Maimon, 2014, p. 5), con los cuales es posible construir un conjunto de reglas jerárquicas que están relacionadas directamente con los datos. Existen diversas variantes de árboles de decisión en la literatura; por ejemplo, el algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de expansión primero en profundidad (*depth-first*) (Mitchell, 2000). Mientras que el algoritmo REPTree, construye un árbol de decisión usando variación de información y lo simplifica usando una poda de error reducido (con ajuste posterior) (Witten et al., 2016).

Por otro lado, el algoritmo bosques aleatorios es una variante de los árboles de decisión, también conocido como un ensamble de árboles (Zhou, 2012), el cual genera una cantidad específica de árboles de decisión independientes, cuyos resultados son promediados. Dicha característica favorece en la generalización del modelo. Finalmente, las máquinas de soporte vectorial son un modelo que separa los datos de entrada transformándolos hacia un espacio de características de alta dimensión, en donde se construye un hiperplano de separación a razón de la distancia entre los vectores de soporte formados. El hiperplano de separación maximiza la distancia entre los vectores de soporte de las clases a separar.

Asimismo, la selección de atributos representa una opción que ayuda a mejorar el desempeño, tanto en la exactitud de clasificación como en la reducción de la complejidad del modelo. También, permite entender qué variables son relevantes o redundantes para el proceso de clasificación. Por esta razón, en este estudio se utilizaron tres métodos de selección de atributos para reducir el número de atributos en el proceso de clasificación y entender cuáles son las variables más importantes que describen a un estudiante desertor, a saber:

- Relief, que asigna una calificación a cada atributo mediante la técnica del k-ésimo vecino más cercano (Kira y Rendell, 1992; Kononenko et al., 1997). Relief es considerado como un método de *ranking*, porque entrega al usuario el orden de las variables con respecto a su importancia.

- La búsqueda secuencial hacia adelante, que representa un método empaquetador (*wrapper*), agrega un atributo a un subconjunto de atributos en cada iteración, de manera que se maximiza gradualmente el desempeño de algún clasificador hasta que se alcance la dimensionalidad requerida (Devijver y Kittler, 1982). Y la búsqueda secuencial hacia atrás también se trata de un método empaquetador, en donde la formación de subconjuntos de atributos se realiza empezando con todas las variables y en posteriores iteraciones se va eliminando de cada subconjunto una variable a la vez.

En este sentido, es importante evaluar el desempeño de cada uno de los algoritmos utilizados. Dentro de las técnicas para estimar el desempeño se encuentra la matriz de confusión también llamada matriz de error o tabla de contingencia, que es una técnica de visualización para obtener información sobre las clasificaciones reales y predicciones realizadas por un sistema de clasificación (Bird et al., 2009), en donde los casos bien clasificados se encuentran en la diagonal de la matriz, mientras que los elementos fuera de ella representan a las instancias mal clasificadas (Witten et al., 2016).

Incluso, la mayoría de las medidas de desempeño del clasificador se pueden enunciar a partir de dicha matriz; algunas de estas medidas frecuentemente usadas son: exactitud, precisión, recuento, medida F, entre otras. En este estudio, sólo se toman en cuenta las medidas de exactitud y la tasa de error balanceado (Witten et al., 2016), puesto que el conjunto de datos utilizado se encuentra desbalanceado, con respecto a sus dos clases, es decir, hay más elementos de una clase (no-desertor) que de la otra (desertor).

La exactitud representa la proporción de las instancias que fueron correctamente clasificadas con respecto al total de instancias que forman al conjunto de datos utilizado y la tasa de error balanceado (TEB) representa el porcentaje promedio de error por clase del clasificador, donde las clases están desbalanceadas, es decir, cuando existe un mayor número de instancias de una clase con respecto a la otra clase (Tabla 1).

Tabla 1. Algoritmos de aprendizaje automático utilizados en la predicción de la deserción escolar

Clasificador	Autores	Nivel de estudios	Tamaño del conjunto de datos
- Regresión (logística/lineal) - Selección de atributos	(Albán y Mauricio, 2019)	Universitario	3 777 registros 30 variables Recolectados en 2017
- Árboles de decisión - k-NN (k-nearest neighbors) - Regresión (logística/lineal)	(Aulck et al., 2017)	Universitario	32 538 registros 16 variables Recolectados en el verano de 2013
- Máquinas de Soporte Vectorial - Árboles de decisión - Näive Bayes - IBk (Instance-based lazy learning)	(Márquez-Vera et al., 2016)	Preparatoria	419 registros 60 variables recolectados de agosto a diciembre de 2012
- Näive Bayes - Árboles de decisión - Random forest - Ramdom Tree - Perceptron multicapa	(Rodríguez-Maya et al., 2017)	Universitario	274 registros 397 variables Obtenidos de las cohortes generacionales 2010-2015 y 2011-2016
- Näive Bayes	(Sara et al., 2015)	Preparatoria	72 598 registros 17 variables Datos de estudiantes matriculados después de 2009
- Árboles de decisión	(Sivakumar et al., 2016)	Universitario	240 registros 32 variables
- Árboles de decisión - Regresión (logística/lineal) - Máquinas de Soporte Vectorial	(Yamao et al., 2018)	Universitario	1 304 registros 8 variables Recolectados durante 2010-2015

1.2 Trabajos relacionados

Como se observa en la Tabla 1, los métodos frecuentemente usados en la predicción de algún evento son: árboles de decisión, máquinas de soporte vectorial y regresión lineal. La mejor precisión al clasificar aplicando árboles de decisión fue la obtenida por Yamao et al. (2018) con un 82.87%; usando Bayes ingenuo fue la obtenida por Sara et al. (2015) con un 76%; aplicando k-NN fue de 64.60% reportado por Aulck et al. (2017), quienes también aplicaron regresión lineal obteniendo una precisión de 66.59%; finalmente, aplicando máquinas de soporte vectorial, Márquez-Vera et al. (2016) obtuvieron un 83.22%. Empero, la aplicación de cualquier algoritmo tiene implicaciones respecto al desempeño del mismo, pues el porcentaje

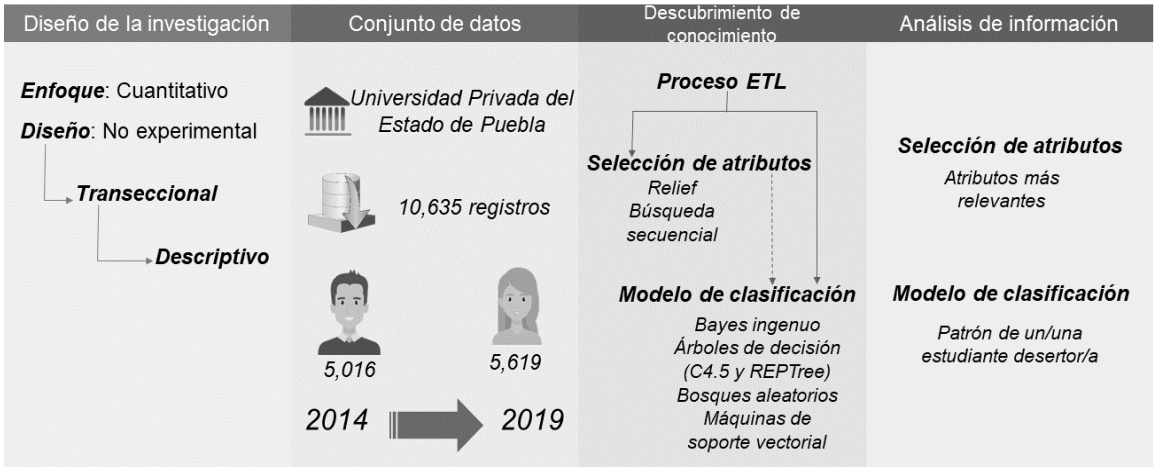
de exactitud está en función de las características del conjunto de datos utilizado.

Es importante resaltar que en este estudio se emplea un conjunto de datos de 10,635 registros con 12 variables, recolectados entre los años 2014-2019, que en comparación con los datos mostrados en la Tabla 1 son actuales y superan de forma considerable el conjunto de datos empleado en dichos estudios, aunque sólo se estudian 12 variables se tiene la ventaja de analizar a dos cohortes generacionales de 53 programas de estudio, en contraposición con los estudios presentados –que recolectaron datos de un solo programa.

II. Método

El presente trabajo proporciona un análisis cuantitativo del fenómeno de la deserción universitaria desde el punto de vista de la minería de datos educativa, con apoyo de la metodología mostrada en la Figura 1, cuyos procesos se detallan a continuación.

Figura 1. Metodología para determinar los atributos más relevantes en la deserción universitaria



Diseño de la investigación: Esta investigación tiene un enfoque cuantitativo y no experimental, de carácter exploratorio (Hernández-Sampieri y Mendoza, 2018; Rahi, 2017), ya que ninguna de las variables descritas en el conjunto de datos es manipulada, es decir, los atributos estudiados no se cambian de forma intencional sobre el efecto en la deserción escolar universitaria. Del mismo modo, el diseño es transeccional descriptivo (Hernández-Sampieri y Mendoza, 2018), al obtener los datos en un solo momento para indagar sobre la incidencia de las variables de estudio.

Conjunto de datos. Se ha considerado un conjunto de datos con 10 635 instancias recolectadas durante el período comprendido entre el otoño del 2014 y la primavera del 2019, provenientes de una institución privada de educación superior del estado de Puebla (México), de 53 programas educativos agrupados en 7 áreas del conocimiento. Los 12 atributos son descritos detalladamente en la Tabla 2. La base de datos está compuesta de información de estudiantes universitarios inscritos y desertores, de los cuales 5 016 son hombres y 5 619 son mujeres entre 17 y 63 años de edad.

Tabla 2. Descripción de los atributos utilizados en el análisis

Atributo	Intervalo de valores	Posibles valores
1. Período	[2014, 2019]	[otoño, verano, primavera]
2. Área	nominal	AH = Artes y Humanidades BIO = Ciencias Biológicas CEA = Ciencias Económico-Administrativas CS = Ciencias Sociales ELC = Estudios de Lengua y Cultura ING = Ingeniería SAL = Ciencias de la Salud
3. Programa	nominal	53 carreras agrupadas en cada una de las siete áreas
4. Género	nominal	Masculino y femenino
5. Edad	[17, 63]	Intervalo de edades
6. Estado de procedencia	nominal	Estados de procedencia de los estudiantes dentro de las 32 entidades federativas de la república mexicana
7. Promedio actual	[0, 10]	Promedio de calificaciones.
8. Créditos cursados	[0, 1]	Porcentaje de materias cursadas, con respecto a la totalidad de créditos de toda la carrera.
9. Último semestre	[0, 11]	Último semestre cursado.
10. Beca	[0, 1]	Sí o no.
11. Porcentaje de asistencia	[0, 1]	Porcentaje de asistencia general de los estudiantes
12. Porcentaje de materias reprobadas	[0, 1]	Porcentaje de materias reprobadas, con respecto al número de materias de la carrera.
Estatus	nominal	Estado actual del estudiante. De manera general, solamente hay dos estados en esta clase: desertor y no desertor. El atributo no desertor contempla a los estudiantes inscritos, pasantes o titulados.

Del total de muestras, 3 233 pertenecen a la clase desertor, mientras el resto (7 402) son de la clase no-desertor (Tabla 3).

Tabla 3. Distribución de clases en los conjuntos de datos de entrenamiento y prueba

Conjunto de datos	Clase "desertor"	Clase "no-desertor"	Total
Entrenamiento	2 587	5 922	8 509
Prueba	646	1480	2 126

Descubrimiento de conocimiento: En esta fase se realizó la selección y limpieza de los datos, que consiste en quitar inconsistencias como: datos duplicados, datos faltantes, datos erróneos, etc. Del mismo modo se han agrupado aquellos términos semejantes. Una vez que el conjunto de datos se ha limpiado, la base de datos está lista para iniciar la aplicación de los algoritmos descritos anteriormente.

Por otro lado, para obtener un modelo clasificador/predictor más amplio y que pueda generalizarse, es necesario separar el conjunto de datos en datos de entrenamiento y datos de prueba. Los primeros ayudan a enseñarle al algoritmo para que pueda predecir un valor o un conjunto de ellos; mientras que los segundos se utilizan para conocer el desempeño final de un modelo, es decir, para comprobar que el algoritmo ha aprendido a predecir un determinado valor (Witten et al., 2016). De esta manera, considerando que el conjunto de datos tiene 10 635 muestras, se ha separado el 80% de los datos para entrenamiento (8 509 muestras) y el resto para la realización de las pruebas. La distribución de clases se detalla en la Tabla 3.

Con la finalidad de validar los resultados obtenidos con muestras no conocidas (ver Tabla 3), se realiza validación cruzada de 10 particiones (es un método útil para calcular el porcentaje de aciertos esperado cuando se realiza una clasificación). Para descubrir las variables relevantes que intervienen en la deserción se utilizan los algoritmos C4.5 y el REPTree. También, para fines comparativos, se emplean los algoritmos

Bayes ingenuo, bosques aleatorios y máquinas de soporte vectorial, con la confianza de obtener rendimientos iguales o superiores a los presentados en la sección de trabajos relacionados. Finalmente, para la aplicación de los algoritmos descritos se utiliza el software para minería de datos Weka versión 3.8 (Frank et al., 2016).

Análisis de la información: Una vez descubiertas las primeras variables importantes, con el uso de árboles de decisión, se complementa el análisis usando algoritmos de selección de atributos, ya que favorecen la identificación de aquellos atributos más importantes en un conjunto de datos dado. Los algoritmos de clasificación empleados dentro de cada empaquetador (selector de características) son los mismos que fueron usados anteriormente. Otro factor que ayuda en el análisis y entendimiento de los resultados, obtenidos por el modelo de árboles de decisión es su fácil interpretación mediante reglas de decisión que son derivadas del árbol generado en el entrenamiento.

III. Resultados

Tal como se estableció en la metodología, antes de realizar el descubrimiento del conocimiento es necesario conocer cómo se comportan los modelos de aprendizaje computacional empleados para clasificar a los desertores, utilizando la totalidad de los atributos de la base de datos. Se observa en la Tabla 4 que, después de realizar validación cruzada usando el conjunto de datos de prueba, los clasificadores con la mejor exactitud son los árboles de decisión C4.5 y REPTree. De igual manera, los resultados obtenidos, tanto de los bosques aleatorios como de la máquina de soporte vectorial son competitivos, comparados con los dos árboles de decisión; no así para el clasificador Bayes ingenuo.

Como se ha mencionado antes, la base de datos utilizada está desbalanceada respecto a sus clases (desertor y no-desertor), por ello se ha utilizado la medida de tasa de error balanceado (TEB) obtenido por cada uno de los algoritmos (Tabla 4) para sopesar los resultados de la medida de exactitud, con el objetivo de poder elegir el mejor modelo. En este caso, se elige a ambos árboles de decisión, ya que proporcionan las máximas exactitudes y los mínimos TEB.

Tabla 4. Medidas de rendimiento de clasificación evaluando al conjunto de datos de prueba usando todos los atributos

Clasificador	Exactitud	TEB
Árbol de decisión C4.5	92.1919%	12.85%
Árbol de decisión REPTree	92.1919%	12.85%
Bosques aleatorios	90.3575%	15.87%
SVM (func. base radial)	90.2634%	15.54%
Bayes ingenuo	65.2399%	37.96%

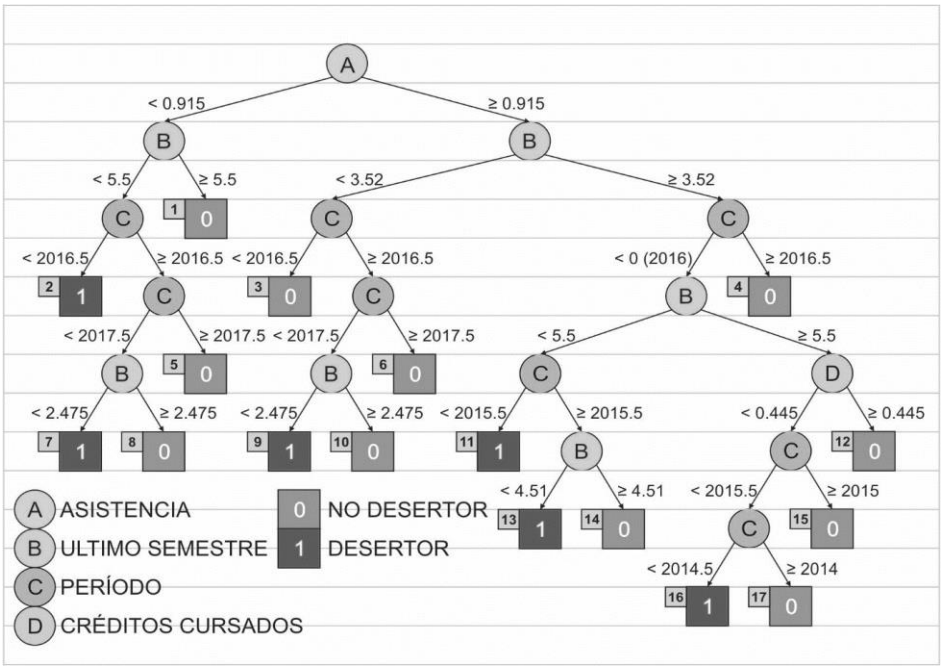
En la Tabla 5 se muestran las matrices de confusión obtenidas, donde se describe a detalle el desempeño por clase de cada clasificador. Asimismo, se advierte que todos los clasificadores (con excepción del Bayes ingenuo) se equivocan clasificando a los desertores en mayor proporción que los no desertores (nótese los resultados de la diagonal de cada matriz). Lo anterior se conoce como sesgo de clasificación hacia la clase mayoritaria. Aunque la mayoría de los clasificadores de la Tabla 4 tienen un desempeño similar al obtenido por los árboles de decisión, éstos no ofrecen la facilidad de interpretación que ofrecen los árboles de decisión. Por esta razón, y debido a la habilidad que tienen para realizar selección de características, se continúa con el análisis usando los algoritmos C4.5 y REPTree.

Aunque el desempeño de ambos árboles de decisión es exactamente el mismo (en términos de las medidas de exactitud y TEB), el árbol de decisión obtenido por REPTree es menos profundo y amplio que el árbol generado por C4.5. Esto significa que genera un árbol menos complejo, por lo que las reglas derivadas de éste (para caracterizar a un desertor) incluyen un número reducido de atributos, haciendo que la interpretación sea más fácil. Lo anterior es posible gracias al proceso de poda realizado por el árbol de decisión REPTree (ver Figura 2), el cual reduce el número de variables a cuatro (“asistencia”, “último semestre”, “período” y “créditos cursados”), las cuales pueden considerarse preliminarmente como las más representativas para este problema.

Tabla 5. Matrices de confusión resultantes de los clasificadores seleccionados

Clasificador	Predicción			Clasificador	Predicción	
	Verdadero	Falso			Verdadero	Falso
Bayes ingenuo	348	298	Verdadero	Árbol de decisión C4.5	480	166
	441	1,039	Falso		0	1,480
Árbol de decisión REPTree	480	166	Verdadero	Bosques aleatorios	441	205
	0	1,480	Falso		0	1,480
SVM (func. base radial)	450	196	Verdadero			
	11	1,469	Falso			

Figura 2. Árbol de decisión REPTree generado a partir del conjunto de datos de entrenamiento usando la totalidad de atributos



Los árboles de decisión, como el generado en la Figura 2, se pueden traducir en reglas de decisión, que ayudan a explicar qué factores intervienen en un problema de clasificación. Las reglas suelen ser de la forma "si premisas, entonces conclusión", donde las premisas que intervienen se conectan entre sí, mediante la operación lógica "y". Finalmente, la consecuencia o conclusión, en este caso representa a la clase en cuestión; es decir, si el estudiante es desertor o no-desertor.

Ahora, para traducir el árbol generado a reglas de decisión y conocer qué factores intervienen en el fenómeno de la deserción, basta con analizar al conjunto de premisas implicadas, exclusivamente de aquellas ramas cuyas hojas corresponden a la clase desertor. Por ejemplo, el árbol generado de la Figura 2, donde sólo se consideran aquellas hojas que corresponden a la clase desertor, es decir, cuya etiqueta es igual a "1", son útiles para generar las reglas de decisión del problema de deserción. Dichas reglas se muestran en la Tabla 4, donde la primera columna corresponde al número de la regla y la segunda columna corresponde a la regla de decisión correspondiente.

Por ejemplo, la regla número 7 pertenece al recorrido que se le hace al árbol desde su raíz hasta la hoja 7 de la Figura 2. Dicha regla establece que "los estudiantes desertaron, entre el período de 2016-2017, cuando la asistencia del estudiante fue inferior al 91.5% y el último semestre cursado del estudiante es inferior al quinto". Este tipo de reglas puede ayudar a planear una estrategia que ayude a prevenir la deserción enfocándose sólo en los estudiantes que cursan los primeros cinco semestres de la carrera.

Otro ejemplo es la regla número 16, la cual involucra a una mayor cantidad de premisas y establece que "repetidamente en algunos períodos de 2014 a 2016, un subconjunto de estudiantes desertó cuando cursaban algún semestre superior al quinto, contando con un porcentaje de créditos cursados inferior a 44.5%, y a pesar de que contaban con una asistencia superior al 91.5%".

Tabla 6. Reglas de decisión obtenidas a partir del entrenamiento del árbol de decisión REPTree usando la totalidad de los atributos

#	Reglas
2	(ASISTENCIA<91.5%) Y (ULTIMO_SEMESTRE<5.5) Y (PERIODO<2016.5) ENTONCES (ESTATUS=DESERTOR)
7	(ASISTENCIA<91.5%) Y (ULTIMO_SEMESTRE<5.5) Y (PERIODO>=2016.5) Y (PERIODO<2017.5) Y (ULTIMO_SEMESTRE<2.475) ENTONCES (ESTATUS=DESERTOR)
9	(ASISTENCIA>=91.5%) Y (ULTIMO_SEMESTRE<3.52) Y (PERIODO>=2016.5) Y (PERIODO<2017.5) Y (ULTIMO_SEMESTRE<2.475) ENTONCES (ESTATUS=DESERTOR)
11	(ASISTENCIA>=91.5%) Y (ULTIMO_SEMESTRE>=3.52) Y (PERIODO<2016.5) Y (ULTIMO_SEMESTRE<5.5) Y (PERIODO<2015.5) ENTONCES (ESTATUS=DESERTOR)
13	(ASISTENCIA>=91.5%) Y (ULTIMO_SEMESTRE>=3.52) Y (PERIODO<2016.5) Y (ULTIMO_SEMESTRE<5.5) Y (PERIODO>=2015.5) Y (ULTIMO_SEMESTRE<4.51) ENTONCES (ESTATUS=DESERTOR)
16	(ASISTENCIA>=91.5%) Y (ULTIMO_SEMESTRE>=3.52) Y (PERIODO<2016.5) Y (ULTIMO_SEMESTRE>=5.5) Y (CREDITOS_CURSADOS<44.5%) Y (PERIODO<2015.5) Y (PERIODO<2014.5) ENTONCES (ESTATUS=DESERTOR)

De este modo, el conjunto de reglas de la Tabla 6 indica que el problema de deserción se encuentra focalizado en los primeros cinco semestres de estudio. También, se puede interpretar que, para los primeros dos semestres de estudio, la "asistencia" es importante, mientras que, en los semestres de tercero a quinto, ésta ya no es tan relevante.

Al mismo tiempo, para identificar las principales variables que influyen en la decisión de desertar, se emplea el algoritmo de Relief para jerarquizar los atributos acorde con la relevancia para predecir a un desertor, tal como se muestra en la primera columna de la Tabla 7, donde se observa a las 12 variables en forma ordenada. Complementariamente, se han probado los algoritmos de selección secuencial hacia adelante y hacia atrás pero no se reportan los resultados debido a que las características seleccionadas no ayudan a mejorar el rendimiento de predicción.

Tabla 7. Máxima exactitud de clasificación, empleando el método de selección Relief y cada uno de los algoritmos de aprendizaje computacional seleccionados

Ranking de variables	Árbol de decisión C4.5	Árbol de decisión REPTree	Bosques aleatorios	SVM (func. base radial)	Bayes ingenuo
1. Período	69.6143%	69.6143%	69.6143%	69.6143%	69.6143%
2. Último semestre	92.1919%	92.1919%	92.1919%	92.1919%	84.7601%
3. Créditos cursados	88.7112%	92.1919%	90.2634%	89.6049%	71.4487%
4. Asistencia	92.1919%	92.1919%	90.1693%	87.3001%	80.0094%
5. Programa	92.1919%	92.1919%	90.5456%	87.5823%	80.0094%
6. Área	92.1919%	92.1919%	90.6397%	88.7582%	80.1035%
7. Estado de procedencia	92.1919%	92.1919%	90.7808%	89.2756%	79.9153%
8. Reprobadas	94.2145%	92.1919%	90.6867%	90.7338%	81.2794%
9. Promedio actual	94.2145%	92.1919%	90.7808%	90.9690%	65.3340%
10. Edad	92.1919%	92.1919%	90.4986%	90.8749%	65.2399%
11. Género	92.1919%	92.1919%	90.4986%	90.9219%	65.2399%
12. Beca	92.1919%	92.1919%	90.3575%	90.2634%	65.2399%

En el caso de Relief, se han colocado a los atributos "período" y "último semestre" como las variables más importantes que describen a la deserción universitaria. Como se ilustra en la Tabla 7, las columnas (de la segunda en adelante) representan los resultados de exactitud de clasificación obtenidos agregando un atributo a la vez al clasificador de manera progresiva y acumulativa, desde la variable más importante hasta la menos importante.

Por ejemplo, la intersección del tercer renglón con la tercera columna indica que REPTree alcanzó una exactitud de clasificación de 92.19%, utilizando como entrada solamente a los atributos “período” y “último semestre”.

En relación con los resultados analizados en la Tabla 7, casi todos los clasificadores utilizados ofrecen un resultado alto de exactitud de clasificación (92.19%), usando los atributos “período” y “último semestre”. No obstante, el árbol de decisión C4.5 obtiene una exactitud máxima de clasificación de 94.21%, usando ocho variables: “período”, “último semestre”, “créditos cursados”, “asistencia”, “programa”, “área”, “estado de procedencia” y “reprobadas”, de las cuales 4 coinciden con aquéllas que fueron relevantes para el algoritmo REPTree de la fase anterior. Si estas variables son utilizadas por el método C4.5, el árbol de decisión generado a partir de los datos de entrenamiento es el mostrado en la Figura 3. Con respecto al procedimiento de construcción de un árbol de decisión mencionado anteriormente, obsérvese que el árbol generado ha destacado sólo 6 de 8 variables para obtener una exactitud de clasificación de 94.21%, lo cual significa que ha descartado a los atributos “área” y “estado de procedencia”, por considerarlos irrelevantes, según la medida de importancia (es decir, usando la ganancia de información, medida en términos de entropía) incorporada en el algoritmo.

En consecuencia, en la Tabla 8 se detallan las reglas de decisión de la clase desertor referente al árbol de la Figura 3. A diferencia de la Tabla 7, la Tabla 8 incluye adicionalmente a los atributos “reprobadas” y “programa”, las cuales ayudan a elevar el desempeño del clasificador C4.5, por encima del desempeño del árbol REPTree. En términos de rendimiento, el modelo más apto para este problema de clasificación es el que corresponde a los árboles de decisión C4.5 (Figura 3), debido a que obtiene una exactitud de clasificación mayor que REPTree. Esto permite destacar la importancia de la selección de atributos para maximizar el rendimiento de los clasificadores, encontrando información adicional que finalmente es relevante para caracterizar de mejor manera a un desertor.

Figura 3. Árbol de decisión C4.5 generado a partir del conjunto de datos de entrenamiento usando las variables obtenidas por el método de selección Relief

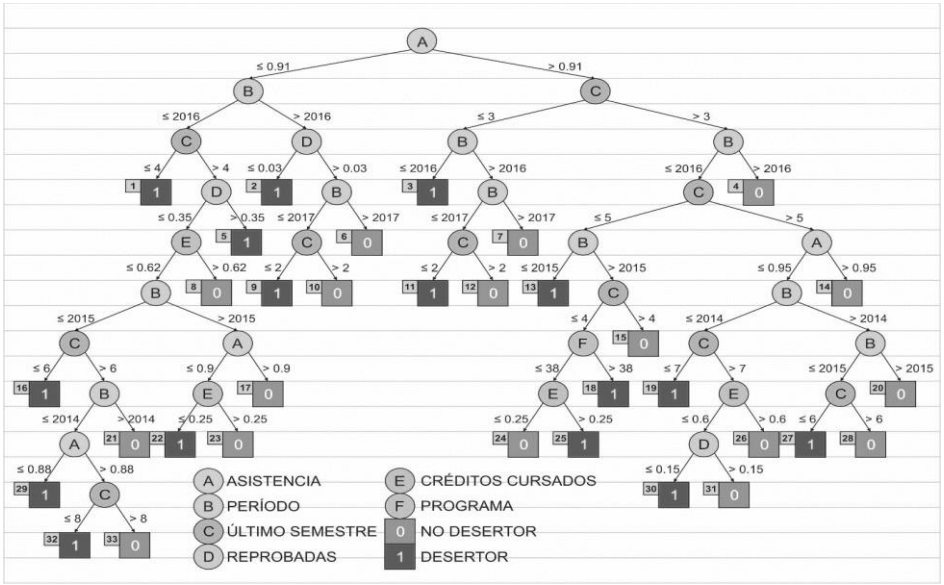


Tabla 8. Reglas de decisión construidas a partir del árbol de decisión C4.5 de la Figura 3 usando los atributos obtenidos por Relief

#	Reglas
9	(ASISTENCIA<=91%) Y (PERIODO>2016) Y (REPROBADAS>3%) Y (PERIODO<=2017) Y (ULTIMO_SEMESTRE<=2) ENTONCES (ESTATUS=DESERTOR)
3	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE<=3) Y (PERIODO<=2016) ENTONCES (ESTATUS=DESERTOR)
1	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE<=4) ENTONCES (ESTATUS=DESERTOR)
5	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>4) Y (REPROBADAS>35%) ENTONCES (ESTATUS=DESERTOR)
22	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>4) Y (REPROBADAS<=35%) Y (CREDITOS_CURSADOS<=62%) Y (PERIODO>2015) Y (ASISTENCIA<=90%) Y (CREDITOS_CURSADOS<=25%) ENTONCES (ESTATUS=DESERTOR)
16	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>4) Y (REPROBADAS<=35%) Y (CREDITOS_CURSADOS<=62%) Y (PERIODO<=2015) Y (ULTIMO_SEMESTRE<=6) ENTONCES (ESTATUS=DESERTOR)
29	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>4) Y (REPROBADAS<=35%) Y (CREDITOS_CURSADOS<=62%) Y (PERIODO<=2015) Y (ULTIMO_SEMESTRE>6) Y (PERIODO<=2014) Y (ASISTENCIA<=88%) ENTONCES (ESTATUS=DESERTOR)
32	(ASISTENCIA<=91%) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>4) Y (REPROBADAS<=35%) Y (CREDITOS_CURSADOS<=62%) Y (PERIODO<=2015) Y (ULTIMO_SEMESTRE>6) Y (PERIODO<=2014) Y (ASISTENCIA>88%) Y (ULTIMO_SEMESTRE<=8) ENTONCES (ESTATUS=DESERTOR)
2	(ASISTENCIA<=91%) Y (PERIODO>2016) Y (REPROBADAS<=3%) ENTONCES (ESTATUS=DESERTOR)
25	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE<=5) Y (PERIODO>2015) Y (ULTIMO_SEMESTRE<=4) Y (PROGRAMA<=38) Y (CREDITOS_CURSADOS>25%) ENTONCES (ESTATUS=DESERTOR)
18	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE<=5) Y (PERIODO>2015) Y (ULTIMO_SEMESTRE<=4) Y (PROGRAMA>38) ENTONCES (ESTATUS=DESERTOR)
19	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>5) Y (ASISTENCIA<=95%) Y (PERIODO<=2014) Y (ULTIMO_SEMESTRE<=7) ENTONCES (ESTATUS=DESERTOR)
30	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>5) Y (ASISTENCIA<=95%) Y (PERIODO<=2014) Y (ULTIMO_SEMESTRE>7) Y (CREDITOS_CURSADOS<=60%) Y (REPROBADAS<=15%) ENTONCES (ESTATUS=DESERTOR)
27	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE>5) Y (ASISTENCIA<=95%) Y (PERIODO>2014) Y (PERIODO<=2015) Y (ULTIMO_SEMESTRE<=6) ENTONCES (ESTATUS=DESERTOR)
13	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE>3) Y (PERIODO<=2016) Y (ULTIMO_SEMESTRE<=5) Y (PERIODO<=2015) ENTONCES (ESTATUS=DESERTOR)
11	(ASISTENCIA>91%) Y (ULTIMO_SEMESTRE<=3) Y (PERIODO>2016) Y (PERIODO<=2017) Y (ULTIMO_SEMESTRE<=2) ENTONCES (ESTATUS=DESERTOR)

De manera general, las reglas de la Tabla 8 ayudan a describir con detalle lo expresado gráficamente en el árbol de la Figura 3. Es decir, el atributo “asistencia” es crucial, sobre todo en los primeros tres semestres. Igual de importantes son los atributos “período” y “último semestre”. A partir del cuarto semestre, los atributos “reprobadas”, “créditos cursados” y “programa” cobran importancia, mientras que la “asistencia” disminuye su relevancia.

Como ejemplo de interpretación de las reglas de la Tabla 8, se presenta el caso de la regla número 32, la cual se podría descifrar de la siguiente manera: “En el período 2014 los estudiantes desertaban, si la asistencia era menor o igual al 91%, pero superior al 88%, y además se encontraban cursando entre el 6o. y el 8o. semestre, con un porcentaje de reprobadas inferior o igual al 35%, pero un avance de créditos cursados inferior o igual al 62%”. Las demás reglas pueden interpretarse de modo similar.

De acuerdo al conjunto de datos analizado, se encontró un grupo de reglas que describen las características de un estudiante desertor y que el porcentaje de incidencia ocurre en los primeros dos años de iniciados los estudios de educación superior. Los resultados obtenidos muestran que el algoritmo de árboles de decisión (método C4.5) ofrece un desempeño alto y mejor que los publicados en los trabajos relacionados (Tabla 1).

Es necesario considerar que los resultados publicados en la literatura relacionada no tienen punto de comparación con los resultados mostrados en las Tablas 4 a 8 de este estudio, debido a las diferencias con la base de datos utilizada. Otras diferencias que existen entre nuestra propuesta y las presentadas en el estado del arte son: la búsqueda de hiper-parámetros realizada y los algoritmos de minería de datos empleados, las cuales tienen como objetivo maximizar la exactitud de clasificación para asegurar que el clasificador prediga de la mejor manera a un desertor, utilizando muestras no conocidas.

IV. Discusión y conclusiones

Como se mencionó, el objetivo de este estudio es encontrar las variables más importantes que inciden en la deserción escolar universitaria con el fin de definir un patrón que identifique a un estudiante desertor. De este modo, usando algoritmos de selección de atributos se ha encontrado que los atributos más importantes son: asistencia, período, último semestre cursado, porcentaje de materias reprobadas, créditos cursados y programa. Estos hallazgos discrepan de los factores reportados en la literatura (Aulck et al., 2017; Carvajal y Trejos, 2016; López y Beltrán, 2017; Márquez-Vera et al., 2016; Ramírez et al., 2016; Sivakumar et al., 2016), en donde se afirma que las causas que originan la deserción escolar están relacionadas con factores académicos, económicos, sociales, creencias, desempeño y adicciones, entre otros. Dicha discrepancia se debe al enfoque que se ha dado al análisis, al método utilizado para ello y principalmente a las diferencias entre las bases de datos empleadas. En otras palabras, en este estudio se le dio una perspectiva diferente, dado que se han analizado variables como el 6, 7, 8, 9, 11 y 12 listados en la Tabla 2, que no han sido considerados por otros autores (Tabla 1) y que también son factores que inciden en la deserción.

En contraste, algunos resultados obtenidos en este estudio se asemejan a los obtenidos por Muñoz et al. (2018) al coincidir en que el período estudiado entre los años 2014-2019 sí es un factor significativo en la deserción. Por otro lado, se observa en los trabajos relacionados resumidos en la Tabla 1 que la exactitud más alta reportada fue de 83.22%, mientras que en este estudio se ha logrado una exactitud del 94.21%. Esto implica que un estudiante desertor puede caracterizarse mejor usando las reglas de decisión encontradas y listadas en la Tabla 8. De igual manera, partiendo de la literatura revisada es destacable que para el problema de la clasificación de deserción de estudiantes la selección de atributos es poco aplicada. Debido a esto, el presente trabajo logra no sólo identificar a un posible desertor, sino que permite identificar, de manera jerárquica, aquellos atributos que son más relevantes (Tabla 8).

Sin embargo, una de las limitaciones encontradas fue el bajo número de atributos analizados en comparación con los mostrados en la Tabla 1; el hecho de incluir un mayor número de atributos en el estudio agregaría la posibilidad de seleccionar uno o más variables importantes, no contempladas actualmente, que discriminen de mejor manera cada una de las clases. Otro inconveniente es haber encontrado registros incompletos, por lo que se tuvieron que descartar del estudio, los cuales podrían haber sido útiles para incrementar el rendimiento de los clasificadores empleados.

Finalmente, los resultados presentados sugieren la integración de una base de datos, construida a partir de estándares y protocolos, que incluya muestras adicionales de IES públicas y privadas. También, se debería incluir y analizar una mayor cantidad de atributos, y la aplicación de otras técnicas de minería de datos, así como evitar el desbalanceo de los datos en lo posible, usando algoritmos de balanceo de clases. De esta manera, se enriquecería aún más el modelo de clasificación obtenido asociado a un estudiante desertor. Es verdad que es valioso que estudios como éste, y otros descritos en los trabajos relacionados, puedan ayudar a las IES a prevenir el fenómeno de la deserción. Pero aún más valioso es generar un modelo, a partir de una base de datos de estas características, incluyendo a la mayoría de IES del país, con el cual sería posible ayudar a advertir la deserción oportunamente y con ello disminuir los índices de deserción a nivel local y nacional.

Referencias

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access, 4.* https://doi.org/10.1109/ACCESS.2016.2568756

Albán, M. y Mauricio, D. (2019). Factors that influence undergraduate university desertion according to students perspective. *International Journal of Engineering and Technology, 10*(6), 1585-1602. https://dx.doi.org/10.21817/ijet/2018/v10i6/181006017

Al-Barrak, M. A. y Al-Razgan, M. (2016). Predicting student's final GPA using decision trees: a case study. *International Journal of Information and Education Technology, 6*(7), 528-533. http://www.ijiet.org/vol6/745-IT205.pdf

Alpaydin, E. (2010). *Introduction to machine learning*. MIT Press.

Aulck, L., Velagapudi, N., Blumenstock, J. y West, J. (2017). Predicting Student Dropout in Higher Education. *Machine Learning in Social Good Applications, 16*(20). https://bit.ly/2R2XGdX

Bird, S., Klein, E. y Loper, E. (2009). *Natural language processing with python*. O'Reilly Media, Inc.

Carvajal, P. y Trejos, A. (2016). *Revisión de estudios sobre deserción estudiantil en educación superior en Latinoamerica bajo la perspectiva de Pierre Bourdieu*. Congreso CLABES, Quito, Ecuador. https://bit.ly/33YO5c2

Chiheb, F., Boumahdi, F., Bouarfa, H. y Boukraa, D. (2017). Predicting students' performance using decision trees: Case of an Algerian University. *International Conference on Mathematics and Information Technology* (ICMIT). IEEE, Adrar, Algeria. https://doi.org/10.1109/MATHIT.2017.8259704

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683-695.

https://doi.org/10.1080/13562517.2013.827653

Devijver, P. A. y Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice Hall.

Estrada, R. I., Zamarripa-Franco, R. A., Zúñiga-Garay, P. G. y Martínez-Trejo, I. (2016). Aportaciones desde la minería de datos al proceso de captación de matrícula de instituciones de educación superior particulares. *Revista Electrónica Educare, 20*(3), 1-21. http://dx.doi.org/10.15359/ree.20-3.11

Frank, E., Hall, M. A. y Witten I. H. (2016). *The WEKA workbench. Online appendix for data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Guevara, C., Sanchez, S., Arias, H., Varela, J., Castillo, D., Borja, M., Fierro, W., Rivera, R., Hidalgo, J. y Yandún, M. (2019). Detection of student behavior profiles applying neural networks and decision trees. En T. Ahram, W. Karwowski, S. Pickl y R. Taiar (Eds.), *Human systems engineering and design II. IHSED 2019. Advances in Intelligent Systems and Computing* (pp. 591-597). Springer. https://doi.org/10.1007/978-3-030-27928-8_90

Hernández-Sampieri, R. y Mendoza, C. P. (2018). *Metodología de la investigación: Rutas cuantitativa, cualitativa y mixta* (6a. ed.). McGraw Hill.

Kira, K. y Rendell, L. A. (1992). *A practical approach to feature selection. International Conference on Machine Learning* (pp. 249-256). Morgan Kaufmann Publishers.

Kononenko, I., Simec, E. y Robnik Sikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence, 7*(1), 39-55.

López, L. y Beltrán, A. (2017). La deserción en estudiantes de educación superior: tres percepciones en estudio, estudiantes, docentes y padres de familia. *Pistas Educativas,* (126), 143-159. https://bit.ly/2SXOab5

Márquez-Vera, C., Cano, A., Romero, C., Mohammad, A. Y., Fardoun, H. M. y Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33*(1), 107-125. https://doi.org/10.1111/exsy.12135

Mitchell, T. M. (2000). *Decision Tree Learning*. Washington State University.

Morales, J. y Parraga-Alava, J. (2018). How predicting the academic success of students of the ESPAM MFL?: A preliminary decision trees based study. Third Ecuador Technical Chapters Meeting (ETCM). Cuenca, Ecuador. https://bit.ly/2SHdDXu

Muñoz, S., Gallardo, T., Muñoz, M. y Muñoz, C. (2018). Probabilidad de deserción estudiantil en cursos de matemáticas básicas en programas profesionales de la Universidad de los Andes Venezuela. *Formación Universitaria, 11*(4), 33-42. https://bit.ly/38KFq8d

OECD. (2019). *OECD Skills Strategy 2019*. https://bit.ly/2P8GJwL

Rahi, S. (2017). Research design and methods: A systematic review of research paradigms, sampling issues and instruments development. *International Journal of Economics & Management Sciences, 6*(2).

Ramírez, E., Espinosa, D. y Millán, E. (2016). Estrategia para afrontar la deserción universitaria desde las tecnologías de la información y las comunicaciones. *Revista Científica, 24*, 52-62. https://doi.org/10.14483/udistrital.jour.RC.2016.24.a5

Rodríguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O. y Suárez-Carranza, B. A. (2017). Modeling Students' Dropout in Mexican Universities. *Research in Computing Science, 139*, 163-175. https://www.rcs.cic.ipn.mx/2017_139/Modeling%20Students_%20Dropout%20in%20Mexican%20Universities.pdf

Rokach, L. y Maimon, O. (2014). *Data mining with decision trees: Theory and applications*. World Scientific Publishing Co.

Romero, C. y Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, 40*(6), 601-618. https://ieeexplore.ieee.org/document/5524021

Sara, N. B., Halland, R., Igel, C. y Alstrup, S. (April, 2015). *High-school dropout prediction using machine learning: A danish large-scale study*. Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium.

Secretaría de Educación Pública. (2019a). *Glosario Educación Superior*. https://bit.ly/31PLNLu

Sivakumar, S., Venkataraman, S. y Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology, 9*(4), 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

Theodoridis, S. y Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press.

Witten, I. H., Frank, E., Hall, M. y Pal, C. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kauffman.

Yamao, E., Saavedra, L. C., Campos, R. y Huancas, V. D. (2018). Prediction of academic performance using data mining in first year students of peruvian university. *CAMPUS, 23*(26), 151-160. https//doi.org/10.24265/campus.2018.v23n26.05

Yukselturk, E., Ozekes, S. y Kılıç Türel, Y. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning, 17*(1), 119-133. https://doi.org/10.2478/eurodl-2014-0008

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman & Hall / CRC Hall / CRC Press.