Urbina-Nájera, A.B., Camino-Hampshire, J.C., & Cruz-Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *RELIEVE*, *26*(1), art. 4. http://doi.org/10.7203/relieve.26.1.16061

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE 25 YEARS AÑOS

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa

University dropout: Patterns to prevent it by applying educational data mining

Urbina-Nájera, A.B. (1); **Camino-Hampshire, J.C.**(2), **& Cruz Barbosa, R.**(3)

(1) UPAEP-Universidad (México); (2) Accenture (México) & UPAEP-Universidad (México); (3) Universidad Tecnológica de la Mixteca (México)

Abstract

Recently, the use of educational data mining techniques has gained great relevance when applied to performance prediction, creation of predictive models of retention, behavior profiles, school failure, among others. In this paper we present the application of the attribute selection algorithm to identify the most important factors that influence the decision to drop out, decision trees are used to define patterns that can alert an imminent dropout. A tool was adapted and administered via web to 300 students from public HEIs and 200 students from private HEIs currently enrolled in higher level program. By means of the algorithm of attribute selection, 27 relevant factors were found. Within the three main factors, the lack of counseling, the lack of an adequate student environment and the lack of academic follow-up were recognized, while, by means of the decision tree, 7 patterns were found, where one of them includes factors such as: student environment, insufficient financial support, experience of an uncomfortable situation, place of career choice, among others. Finally, it has been seen that school drop-out does not depend on a single factor, but is multifactorial and that it is imperative to expand the sample to other cities so that various algorithms can be applied that provide greater information leading to the establishment of accurate mechanisms for reducing university drop-out rates according to the characteristics of the student population in each region.

Keywords: Student environment; Computer learning; Decision trees; Counseling; Feature selection

Resumen

Recientemente, el uso de técnicas de minería de datos educativa ha cobrado gran relevancia al aplicarlas en la predicción del desempeño, creación de modelos predictivos de retención, perfiles de comportamiento, fracaso escolar, entre otros. En este trabajo se presenta la aplicación del algoritmo selección de atributos para identificar los factores más importantes que inciden en la decisión de desertar; también, se utilizan árboles de decisión para definir patrones que pueden alertar una inminente deserción. Se adaptó un instrumento y se administró vía web a 300 estudiantes de IES pública y 200 estudiantes de IES privada actualmente inscritos en algún programa de nivel superior. Mediante el algoritmo selección de atributos se encontraron 27 factores relevantes, dentro de los tres factores principales se reconocen la falta de asesorías, la falta de un ambiente estudiantil adecuado y la falta de seguimiento académico, mientras que, por medio del árbol de decisión se encontraron 7 patrones, en donde uno de ellos incluye factores como: ambiente estudiantil, apoyos financieros insuficientes, experiencia de una situación incómoda, lugar que ocupa la elección de la carrera, entre otros. Finalmente, se ha visto que la deserción escolar no depende de un solo factor, sino que es multifactorial y que es imperativo ampliar la muestra a otras ciudades de manera que se puedan aplicar diversos algoritmos que proporcionen mayor información que conduzcan al establecimiento de mecanismos certeros para disminuir los índices de deserción universitaria en función de las características de la población estudiantil según la región.

Palabras clave: Ambiente estudiantil; Aprendizaje Computacional; Árboles de decisión; Asesoramiento; Selección de atributos

Received/*Recibido* 2019 November 04 | **Approved /*Aprobado*** 2020 April 20 | **Published/*Publicado*** 2020 June 30

Corresponding author / Autor de contacto: **Urbina-Nájera, A.B.** 17 sur #901. Barrio de Santiago. 72410. Puebla, Pue. (México). abunajera@gmail.com

La educación superior es el motor que impulsa el crecimiento de las economías modernas, pues dota a los estudiantes de competencias, habilidades técnicas, profesionales, humanas y disciplinares específicas que los cualifican para una pluralidad de funciones laborales (OECD, 2019). Es por ello, que la educación y el desarrollo de competencias son los pilares sobre los que las naciones deben construir su crecimiento, productividad y prosperidad futura (OECD, 2017). En este sentido, el sistema de educación superior de la República Mexicana ofrece una amplia gama de programas que han experimentado un desarrollo rápido en las últimas décadas. Particularmente, en los niveles que comprenden estudios de licenciatura o ingeniería, el 89% de los estudiantes en 2015 estaban matriculados en este tipo de programas, alcanzando un porcentaje mayor que el promedio de la OECD (por sus siglas en inglés *Organisation for Economic Co-operation and Development*) del 61%. Sin embargo, México tiene la proporción más baja entre los países de la OECD de adultos entre 25 y 64 años con un título de educación superior (17%) cifra inferior al promedio del 37% de la OECD (OECD, 2019).

A pesar de estas cifras, México ha conseguido avances notables respecto al aumento del logro educativo en los niveles de educación superior, pues en los últimos 16 años la proporción de adultos jóvenes que han finalizado este nivel pasó del 17% al 23% (OECD, 2019). Este lento crecimiento se debe a la deserción escolar, que es un aspecto educativo que ha tomado especial atención desde hace más de una década en todas las Naciones. De acuerdo con datos reportados por el Banco Mundial (2018) la mitad de los estudiantes entre 25 y 29 años termina sus estudios universitarios y el 50% de los abandonos ocurre en el primer año de estudios. Según antecedentes del Instituto Nacional de Estadística y Geografía (INEGI, 2018) de cada 100 estudiantes universitarios mexicanos solamente 8 concluyen sus estudios. Finalmente, la Secretaría de Educación Pública (SEP) reportó que en el ciclo escolar 2017-2018 se registró un 8.4% que aumentó considerablemente respecto al ciclo escolar 2016-2017 en donde se obtuvo un 7.2% (Secretaría de Educación Pública, 2019).

Entiéndase por deserción escolar aquel acto que conduce a desentenderse de los estudios o incorporación a otra institución y por abandono escolar cuando la prioridad no es estudiar sino atender obligaciones que se suplen con diversas necesidades del ser humano, en otras palabras, el abandono escolar está determinado por cifras estadísticas relacionadas al número de alumnos que dejan la escuela de un ciclo escolar a otro (Secretaría de Educación Pública, 2019). De esta manera, un alumno/a desertor/a se entiende por aquel individuo que ha abandonado los estudios y las obligaciones que le corresponden como estudiante, perdiendo su calidad de alumno y los derechos adquiridos en su inscripción en el centro educativo (Universidad Tecnológica de Tabasco, 2019).

Diversos estudios, muestran que las causas que originan una deserción escolar obedecen a razones multifactoriales que van desde aspectos personales, familiares, económicos, hasta políticos, culturales e institucionales. En este trabajo se usa el término deserción escolar para identificar a aquellos individuos que abandonan o deciden ya no retomar sus estudios en cualquier periodo de su formación académica.

Recientemente, el uso de técnicas de minería de datos educativa ha cobrado relevancia en el análisis de diversos aspectos educativos como la deserción escolar; su aplicación tiene como objetivo seguir la huella digital de los estudiantes y descubrir de manera oportuna un cambio en el comportamiento vinculado a aspectos académicos que puedan predecir, por ejemplo, una inminente deserción o abandono escolar. Esta técnica, también, se ha utilizado para predecir el comportamiento de los estudiantes a fin de realizar recomendaciones sobre el proceso de aprendizaje-enseñanza, desempeño, gestión de actividades, entre otros. De igual manera, se ha empleado para encontrar patrones ocultos en estudiantes en riesgo, retención, deudores y similares. La minería de datos educativa, busca crear métodos para explorar los tipos únicos de datos que provienen de entornos formativos con el objetivo de resolver y mejorar los procesos educativos de

manera automatizada (Romero y Ventura 2007).

De este modo, el objetivo de este estudio es hacer uso de esta técnica computacional aplicando el algoritmo selección de atributos y el algoritmo árboles de decisión, para identificar los principales factores que influyen en la deserción universitaria, a la vez, que se encuentran patrones para prevenirla, usando un conjunto de datos de 500 registros Recuperados a través de una encuesta administrada a estudiantes inscritos en Instituciones de Educación Superior (IES) públicas y privadas en la Ciudad de Puebla.

Para ello, se presenta una breve revisión de la literatura acerca de la deserción universitaria planteada desde dos aristas: metodologías tradicionales (cuantitativas y cualitativas) y aplicación de minería de datos educativa. Enseguida, se define a la minería de datos educativa, resaltando su capacidad para identificar patrones nuevos y no triviales para resolver y mejorar procesos educativos. Al mismo tiempo, se muestra la metodología implementada la cual se basa en la aplicación del proceso de descubrimiento de conocimiento en bases de datos. Por último, se presentan los resultados que permitieron identificar los patrones que inciden en una deserción escolar logrando concluir en acciones que favorecen una prevención oportuna que coadyuvarían en la disminución de los porcentajes de deserción universitaria locales y nacionales.

Minería de datos educativa

La minería de datos educativa ([EDM] *Educational Data Mining*) es una disciplina emergente, enfocada en crear métodos para explorar los tipos únicos de datos que provienen de entornos educativos con el objetivo de resolver y mejorar los procesos educativos de manera automatizada. Los métodos de la EDM se extraen de una variedad de áreas, incluyendo la minería de datos, aprendizaje computacional, psicometría, estadística, visualización de información y el modelado computacional (Romero & Ventura 2007). Hoy en día, hay muchos métodos (algoritmos) que han sido aplicados en varios problemas del mundo real con alta precisión, estos algoritmos incluyen árboles de decisión, máquinas de soporte vectorial, redes neuronales artificiales, aprendizaje bayesiano, métodos basados en instancias, métodos de kernel, entre otros. En este apartado, se da una breve descripción de los árboles de decisión, así como, de las métricas utilizadas para evaluar su desempeño.

Árboles de decisión (DT)

Los árboles de decisión se ubican dentro de una rama del aprendizaje computacional denominada aprendizaje simbólico, en la que también se encuentran los modelos de reglas de decisión, estrechamente relacionados con los árboles. El aprendizaje mediante árboles de decisión es una técnica que permite analizar decisiones secuenciales basadas en el uso de resultados y probabilidades asociadas. Mitchel (1997) lo define como "*Un método de aproximación de una función objetivo de valores discretos en el cual la función objetivo es representada mediante un árbol de decisión.*

Los árboles aprendidos también pueden representarse como un conjunto de reglas Si–entonces…" son uno de los métodos de aprendizaje inductivo más usado en los algoritmos de inferencia. Dicha representación se denota por un nodo de decisión, un nodo de probabilidad y una rama. El nodo de decisión representado por un cuadrado, indica que una decisión necesita tomarse en ese punto del proceso. El nodo de probabilidad representado por un rectángulo redondeado, indica que en ese punto del proceso ocurre un evento aleatorio. Finalmente, la rama muestra los distintos caminos que se pueden emprender cuando se toma una decisión o bien ocurre algún evento aleatorio representado por una línea (Frank, Hall, Mark & Witten, 2016).

La mayoría de los algoritmos que han sido desarrollados para el aprendizaje de los árboles de decisión son variaciones de un algoritmo que emplea un núcleo de arriba hacia abajo (*top-down*). Este enfoque, particularmente, incluye el algoritmo ID3 y su sucesor C4.5, ambos desarrollados por Quinlan en 1986 y 1993, respectivamente.

El algoritmo ID3 (*Induction Decision Trees*) es un sistema de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de ejemplos. Estos ejemplos o tuplas están constituidos por un conjunto de atributos y un clasificador o clase. Los dominios de los atributos y de las clases deben ser discretos. Además, las clases deben ser disjuntas. En general, es un algoritmo que genera descripciones que clasifican a cada uno de los ejemplos del conjunto de entrenamiento. Mientras que el algoritmo C4.5 o J48 (es una extensión de ID3), permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas. Los árboles que genera son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases; este algoritmo es popularmente utilizado en la minería de datos debido a su sencillez de interpretación y la forma visual de representar los resultados (Mitchell, 2000).

Otros algoritmos popularmente aplicados en esta área son *NBTree* y *Random Forest*. El algoritmo NBTree (*Naive Bayes Tree*) es considerado un algoritmo híbrido pues el árbol que genera, en las hojas contienen un clasificador Naive Bayes construido a partir de los ejemplos que llegan al nodo, es un eficiente y efectivo algoritmo de aprendizaje, de igual manera muestra datos de predicción tan eficientemente como el algoritmo C4.5, aunque con la limitante de únicamente representar cierto grado de separación entre las funciones binarias (Chen et al., 2017). Por otro lado, *Random Forest* o bosques aleatorios, se conforma de un gran número de árboles de decisión individuales que operan como un conjunto, en donde cada árbol individual del bosque aleatorio muestra una predicción de la clase y la clase con más votos se convierte en la predicción del modelo (Ustebay, Turgut y Ali, 2018).

De igual modo, hay otros algoritmos como CHAID o CART (ninguno disponible en Weka), el primero comúnmente utilizado para medir el grado de correlación entre las variables independientes y la clase (Cha, Kim, Moon & Hong, 2017) y el segundo utilizado para el análisis de regresiones lineales o múltiples (Sharma & Kumar, 2016).

Por otro lado, para construir un DT es necesario determinar qué atributos son los mejores, particularmente, cuál es el atributo que debe colocarse en el nodo raíz. Así, la entropía y la ganancia de información son utilizadas para dar respuesta a estas incógnitas. De acuerdo a (Mitchel, 1997) la entropía es una medida que permite calcular el grado de incertidumbre de una muestra. Si la muestra es completamente homogénea, su entropía = 0, al contrario de una muestra igualmente distribuida, cuya entropía = 1. En este sentido, se entiende por ganancia de información a la calidad de una variable, es decir, la ganancia de información verifica cuan homogénea es la distribución de la clase antes de instanciar cualquier variable; utilizada particularmente, en la creación de un árbol de decisión (Gupta, Rawat, Jain, Arora & Dhami, 2017).

Desempeño de los árboles de decisión

Un algoritmo debe ser analizado para determinar el uso de los recursos que utiliza y principalmente, el desempeño para realizar alguna tarea como clasificar, reconocer, identificar, agrupar, categorizar, entre otros. Existen medidas para estimar el desempeño de un algoritmo y este desempeño dependerá de qué medida se esté empleando como prioridad.

El criterio más obvio para estimar el rendimiento de un clasificador es su exactitud predictiva en instancias que no se ven. El número de casos que no se ven, algunas veces es potencialmente grande (si no es que infinita), por lo tanto, una estimación debe ser calculada en un conjunto de pruebas. A esto se le conoce comúnmente como validación cruzada. La validación cruzada (*cross-validation*) es una técnica empleada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición de datos de entrenamiento y prueba. Este método es muy preciso puesto que se evalúa a partir de *k* combinaciones de datos de entrenamiento y de prueba. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos. Lo más común es manipular la validación cruzada de 10 iteraciones (10-*fold*

cross-validation), si la muestra es muy grande (k>10) entonces k =3, y si la muestra es muy pequeña, entonces se toma el valor máximo de k (M.P. van der Aalst, 2011).

Por otro lado, una matriz de confusión, también llamada matriz de predicción o de clasificación, es una herramienta de visualización que se emplea para obtener información sobre las clasificaciones reales y predicciones realizadas por un sistema de clasificación (Bird, Klein & Loper, 2009). Entonces, la matriz de confusión es una tabla donde cada celda [*i*,*j*] indica cómo se clasificó alguna instancia con respecto a su clase pre-establecida (Tabla 1). Los casos bien clasificados se encuentran en las entradas de la diagonal (en este caso las celdas [*a*,*d*] de la Tabla 1) pues los grupos pronosticados y reales son los mismos; los elementos fuera de la diagonal se encuentran mal clasificados (Witten & Frank, 2005; Montero Lorenzo, 2007; Bird, Klein & Loper, 2009; Hamilton, 2009).

Tabla 1. Matriz de confusión cuando se tienen dos posibles resultados de clasificación: Negatigo y Positivo.

		Predicción	
		Negativo	**Positivo**
Actual	**Negativo**	a	b
	Positivo	c	d

Las métricas frecuentemente usadas que se obtienen de la matriz de confusión son: Exactitud, Precisión, Recuerdo, Medida F, cuyos conceptos se describen a continuación: Exactitud (*accuracy*) defininada como la proporción del número total de predicciones que son correctas; precisión (*precision*) (consistencia o confidencia) conocida como la proporción de la predicción de los casos positivos correctos; recuperación (*recall*) (verdadero positivo, completitud o sensibilidad) se interpreta como el número de casos que deberían haber sido recuperados en función de algunos criterios de búsqueda; finalmente, medida F (*F-Measure*) es una medida que combina la precisión con la recuperación para dar una puntuación única (Freitas, 2002; Witten & Frank, 2005; Bird, Klein & Loper, 2009 y M.P. van der Aalst, 2011).

Selección de atributos

Es frecuente que se tenga un gran número de atributos para cada instancia en un conjunto de datos, sin embargo, no todos pueden ser relevantes para caracterizar al objeto. De hecho, si se utilizan todos los atributos pueden, en muchos casos, causar un problema (Witten, Frank & Hall, 2011). En otras palabras, el gran número de atributos representa un espacio de alta dimensión, por lo que es necesario llevar a cabo una reducción de la dimensionalidad, seleccionando sólo unos pocos atributos. Este pequeño conjunto de atributos debe conservar la mayor cantidad de información posible que describa a los ejemplos (Bishop, 2007).

Por otro lado, la selección de atributos está compuesta de un evaluador y un método de búsqueda. Un evaluador (individual o subconjunto) define la forma en que los algoritmos evalúan atributos y se clasifican en *filter*, *wrapper y ranker*. Los primeros dos generan un subconjunto de atributos y el tercero, genera un ranking con todos los atributos.

Para poder ejecutar un evaluador, es requisito seleccionar un método de búsqueda, que para los evaluadores *filter y wrapper*, el método consiste en buscar un espacio entre los subconjuntos de datos utilizando métodos como: *CfsSubsetEval*: Evalúa un subconjunto de atributos considerando la habilidad predictiva individual de cada variable, así como el grado de redundancia entre ellas. *ConsistencySubsetEval*: Determina un subconjunto de atributos por el nivel de consistencia en los valores de la clase al proyectar las instancias de entrenamiento sobre el subconjunto de atributos. *ClassifierSubsetEval*: Estima los subconjuntos

de atributos en los datos de entrenamiento o en un conjunto de prueba independiente, utilizando un clasificador. *WrapperSubsetEval*: Calcula los subconjuntos de atributos utilizando un clasificador. Emplea validación cruzada para estimar la exactitud del esquema de aprendizaje en cada conjunto (Hall, 2011; Frank, Hall, Mark & Witten, 2016).

De igual manera, para el evaluador *ranker* (considerado como un evaluador individual), el método de búsqueda evalúa atributos individuales, bajo los métodos: *ChiSquaredAttributeEval*: calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo. *GainRatioAttributeEval*: evalúa cada atributo midiendo su razón de beneficio con respecto a la clase. *InfoGainAttributeEval*: estima los atributos midiendo la ganancia de información de cada uno con respecto a la clase. *OneRAttributeEval*: mide la calidad de cada atributo utilizando el clasificador OneR, el cual usa el atributo de mínimo error para predecir, discretizando los atributos numéricos (Frank, Hall, Mark y Witten, 2016).

Proceso para el descubrimiento del conocimiento

El descubrimiento de conocimiento en bases de datos (KDD, *Knowledge Discovery in Databases*), se concentra principalmente en las siguientes tareas: colección de datos, pre-procesamiento de los datos, selección de atributos y aplicación de algoritmos de aprendizaje computacional. Enseguida se describe cada una de estas tareas.

Colección de datos. Los métodos de recolección de datos incluyen la adquisición y almacenamiento de las nuevas observaciones, consultar bases de datos existentes de acuerdo con el problema, y si es necesario realizar cualquier combinación de datos (Han, Kamber & Pei, 2011). En el caso particular de este trabajo, los datos han sido recolectados en un formato digital, el proceso para ello se describe en las siguientes secciones.

Pre-procesamiento de los datos. El pre-procesamiento consiste en manipular, enriquecer, reducir o transformar los datos originales para ser accesibles, posteriormente, con mayor facilidad (Han, Kamber & Pei, 2011). Luego, la fase de transformación implica combinar datos que residen en diferentes fuentes para proporcionar una visión unificada de estos datos, de manera que estos datos se conviertan de un formato fuente a un formato de destino en función de la herramienta a utilizar y poder ser cargados en ella sin inconvenientes de lectura (Witten & Frank, 2005).

Selección de atributos. La razón por la cual es apropiado utilizar selección de atributos está en función de la mejora en la predicción, reducción del tiempo de entrenamiento del algoritmo y reducción en el espacio de almacenamiento.

Aplicación de algoritmos de aprendizaje computacional. El aprendizaje computacional (AC) es la rama de la Inteligencia Artificial que se dedica al estudio de los agentes/programas que aprenden o evolucionan basados en su experiencia, para realizar una tarea determinada cada vez mejor (Mitchell, 2000). La aplicación de algoritmos de aprendizaje computacional es la última fase del proceso para el descubrimiento del conocimiento, cuyo objetivo principal es utilizar la evidencia conocida para poder crear una hipótesis y poder dar una respuesta a nuevas situaciones desconocidas, es decir, partiendo de este hecho se deben seleccionar aquellos algoritmos que den respuesta a diversas situaciones.

Como se ha mencionado anteriormente, existen varios tipos de algoritmos de árboles de decisión, comúnmente utilizados en el área de minería de datos como ID3, C4.5, *NBTree*, *RandomForest* y otros como, CHAID y CART. Las razones por la cuales, en este trabajo se aplica el algoritmo C4.5 son las siguientes: facilita la interpretación de la decisión adoptada, proporciona un alto grado de comprensión del conocimiento, explica el comportamiento respecto a una determinada tarea de decisión, reduce el número de variables independientes y también permite desplegar visualmente un problema (Yukselturk, Ozekes & Kılıç, 2014; Jadhav y Channe, 2016), además

es el comúnmente utilizado para identificar las variables más importantes en un conjunto de datos con menos tasa de error y mayor precisión, entre otras ventajas asociadas a la toma de decisiones (Sharma & Kumar, 2016; Gupta, Rawat, Jain, Arora & Dhami, 2017). También, se emplea el algoritmo selección de atributos con el evaluador *ranker* y los métodos de búsqueda *GainRatioAttributeEval e InfoGainAttributeEval* debido a que, mediante ellos, es posible obtener una lista ordenada de los atributos más relevantes en el conjunto de datos analizado.

Trabajos relacionados

En este apartado, se presenta una breve revisión de la literatura acerca de la deserción universitaria planteada desde dos aristas: 1) Estudios que emplean metodologías convencionales (cuantitativa, cualitativa, métodos estadísticos o reflexiones teóricas) y 2) Estudios relacionados con la aplicación del aprendizaje computacional en diversos contextos educativos.

Estudios que emplean métodos estadísticos

Si bien la deserción escolar no obedece a una sola causa, sí hay una razón que origina la decisión de desertar. Para conocer sus causas, en este apartado se presentan 15 trabajos relacionados seleccionados bajo los siguientes criterios: a partir del año 2000 a la fecha, empleo de métodos estadísticos (enfoques mixtos, enfoques cualitativos, enfoques cuantitativos), centrados en la educación superior, estudios relacionados a nivel de maestría y doctorado se han excluido. La tabla 2 muestra la lista de factores que se han identificado por medio de métodos estadísticos ordenados alfabéticamente.

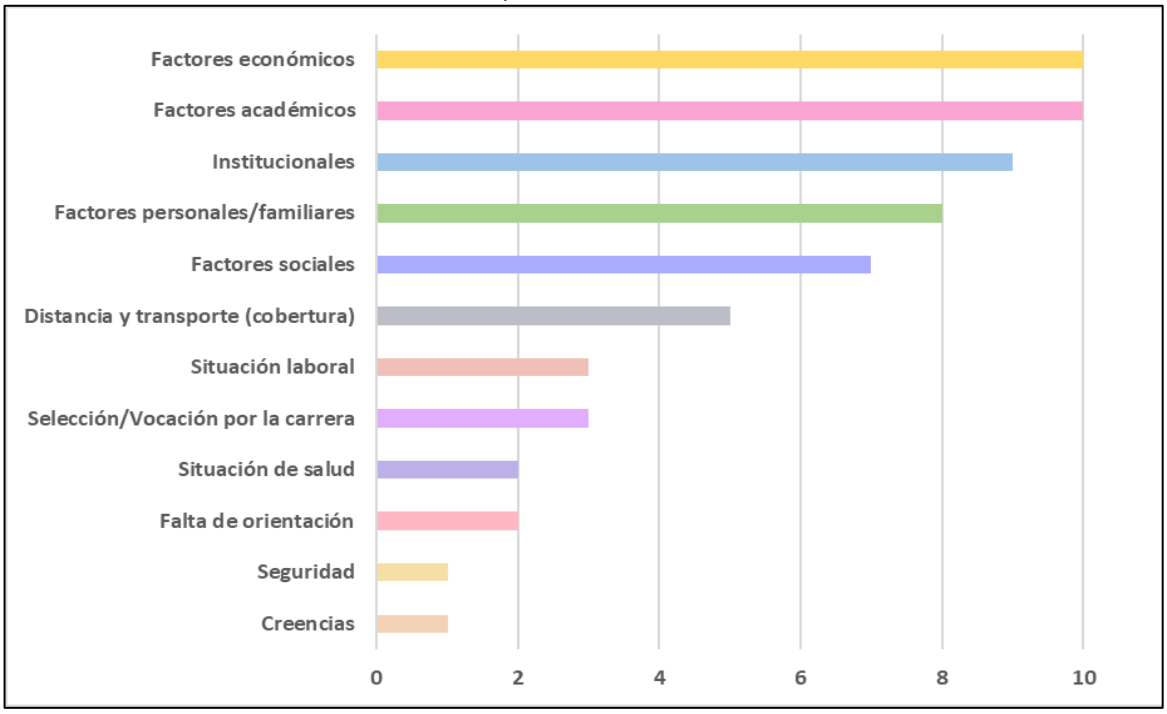
Tabla 2. Factores de la deserción universitaria identificados con métodos estadísticos de 2000-2016.

Factores	Autores
Creencias	Carvajal & Trejos (2016)
Distancia y transporte (cobertura)	Vélez & López (2004); Lavado & Gallegos (2005); Sandoval (2001); Abarca & Sánchez (2005); Lugo (2013)
Factores académicos	Ruíz (2009); Erazo, et. al. (2013); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Fozdar, Kumar y Kannan (2006); Vries, León Arenas, Romero & Hernández, (2011); Carvajal & Trejos (2016)
Factores económicos	Vélez & López (2004); Ruíz (2009); Lavado & Gallegos (2005); Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013) Carvajal y Trejos (2016)
Factores personales/familiares	Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez y González (2006); Londoño (2013); Rode, Bjornoy y Sogaard (2013); Lugo (2013); Fozdar, Kumar & Kannan (2006); Carvajal & Trejos (2016)
Factores sociales	Vélez y López (2004); Ruíz (2009); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Carvajal & Trejos (2016)
Falta de orientación	Abarca & Sánchez (2005); Rode, Bjornoy & Sogaard (2013)
Institucionales	De los Santos (2004); Ruíz (2009); Abarca & Sánchez (2005); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Fozdar, Kumar & Kannan (2006); Carvajal & Trejos (2016)
Seguridad	Vélez & López (2004)
Selección/Vocación por la carrera	Abarca & Sánchez (2005); Lugo (2013); Vries, León, Romero & Hernández (2011)
Situación de salud	Erazo, et. Al. (2013); Lugo (2013)
Situación laboral	Ruíz (2009); Rode, Bjornoy & Sogaard (2013); Vries, León, Romero & Hernández (2011)

Fuente: *Elaboración propia*

Como se detalla en la Figura 1 los factores con mayor frecuencia y en la misma proporción son los económicos y académicos; las cifras representan el número de autores que lo ha mencionado como principal causa en sus estudios (Tabla 2). De manera que, los factores económicos se han asociado, por mencionar algunos, a la falta de recursos para pago de colegiaturas, falta de apoyos financieros, pérdida de empleo del padre o tutor y subestimación de costos. En la agrupación de los factores académicos se ubican aquellos relacionados a la asesoría, tutoría, acompañamiento u orientación insuficiente o inadecuada, bajo rendimiento académico personal, incumplimiento de reglamento académico, nivel académico de la universidad demasiado alto/bajo, oferta de asignaturas u horarios insuficientes.

Figura 1. Factores que inciden en la deserción universitaria identificados por métodos estadísticos, revisión del 2000-2016.



Fuente: Elaboración propia

Estudios que aplican aprendizaje computacional

Después del año 2000 se han empleado otras técnicas científicas para identificar aquellas causas que originan la deserción escolar no solamente la universitaria, también la ocurrida en secundaria y preparatoria. Particularmente, el uso de técnicas de inteligencia artificial ha cobrado gran importancia, pues hay estudios donde se describe la aplicación de algoritmos de aprendizaje computacional para dar solución a alguna situación en el contexto educativo. En este sentido, se han identificado los algoritmos comúnmente utilizados.

- **Näive Bayes** (Kotsiantis, Pierrakeas Pintelas (2003); Dekker, Pechenizki & Vleeshouwers (2009); Pal (2012); Er (2012); Yukselturk, Ozekes & Kılıç (2014); Barbosa, Serra da Cruz & Zimbrão (2014); Sara, Halland, Igel & Alstrup (2015); Márquez-Vera, et al. (2016)). Siendo el porcentaje de predicción mayor reportado de 83% por Kotsiantis, Pierrakeas & Pintelas (2003).
- **Redes neuronales** (Kotsiantis, Pierrakeas y Pintelas (2003), Delen (2011); Yukselturk, Ozekes & Kılıç (2014); Alkhasawneh & Hargraves (2014); Barbosa, Serra da Cruz & Zimbrão (2014)). Siendo el porcentaje de predicción mayor

reportado de 81% en Delen (2011) y Alkhasawneh y Hargraves (2014)

- **Vecinos más cercanos** (K-NN) (Kotsiantis, Pierrakeas & Pintelas (2003); Yukselturk, Ozekes & Kılıç (2014); Márquez-Vera, et al. (2016); Aulck, Velagapudi, Blumenstock & West (2017)); Siendo el porcentaje de predicción mayor reportado de 87% por Yukselturk, Ozekes & Kılıç (2014).
- **Regresión (Lineal y logística)** (Kotsiantis, Pierrakeas & Pintelas (2003), Delen (2011); Aulck, Velagapudi, Blumenstock & West (2017); Yamao, Saavedra, Campos Pérez & Huancas (2018)). Siendo el porcentaje de predicción mayor reportado de 66.59% por Aulck, Velagapudi, Blumenstock & West (2017).
- **Máquinas de soporte vectorial** (Kotsiantis, Pierrakeas & Pintelas (2003); Barbosa Manhães, Serra da Cruz & Zimbrão (2014); Sara, Halland, Igel & Alstrup (2015); Márquez-Vera, Cano, Romero & Mohammad Noaman (2016); Yamao, Saavedra, Campos & Huancas (2018)). Siendo el porcentaje de predicción mayor reportado de 87.39% por Barbosa, Serra da Cruz & Zimbrão (2014).
- **Selección de atributos** (Márquez, Cano, Romero & Ventura (2012); Alkhasawneh & Hargraves (2014); Márquez-Vera, et al. (2016).

Por supuesto, que también se han empleado árboles de decisión para resolver diversas situaciones en el contexto educativo. En la Tabla 3 se presentan trabajos de 2003 a 2019 relacionados con la aplicación del algoritmo árboles de decisión en la solución de situaciones educativas en el nivel superior. Siendo el porcentaje de predicción mayor reportado de 82.87% por Yamao, Saavedra, Campos & Huancas (2018).

Tabla 3. Situaciones educativas resueltas aplicando árboles de decisión de 2003 a 2019

Situaciones Educativas	Autores
Predicción del desempeño	Kabra & Bichkar (2011); Vijayalakshmi & Kumar (2011); Márquez, Cano, Romero & Ventura (2012); Barbosa, Serra da Cruz & Zimbrão (2014); Al-Barrak & Al-Razgan (2016); Agaoglu (2016); Chiheb, Boumahdi, Bouarfa & Boukraa (2017); Yamao, Saavedra, Campos & Huancas (2018)
Captación de matrícula en IES particulares	Estrada-Danell, Zamarripa-Franco, Zúñiga-Garay & Martínez-Trejo (2016)
Éxito académico	Morales & Parraga-Alava (2018)
Predicción de la deserción escolar universitaria	Dekker, Pechenizki & Vleeshouwers (2009); Yukselturk, Ozekes & Kılıç (2014); Abu-Oda & El-Halees (2015); Márquez-Vera, et al. (2016); Sivakumar, Venkataraman & Selvaraj (2016);
Modelos predictivos de la deserción escolar universitaria	Aulck, Velagapudi, Blumenstock & West (2017); Rodríguez-Maya, Lara-Álvarez, May-Tzuc & Suárez-Carranza (2017);
Retención	Raju y Schumacker (2015); Delen (2011); Kumar, Bharadwaj & Pal (2012);
Perfiles de comportamiento	Guevara, et al. (2019)
Fracaso escolar	Márquez, Romero & Ventura (2012)
Estudiantes en riesgo	Er (2012)
Disminución de la tasa de abandono	Pal (2012)

Fuente: Elaboración propia

Después de la revisión presentada en la Tabla 3, se observa que solamente tres trabajos han aplicado el algoritmo selección de atributos para identificar aquellos factores que inciden en la deserción escolar en preparatoria. En el primero, utilizaron un conjunto de datos con 77 atributos (características) de 670 jóvenes entre 15-18 años; aplicaron 10 métodos para obtener el ranking de los factores predominantes, considerando solo aquellos que tuviesen una frecuencia mayor o igual a 2, seleccionando solamente 15 atributos de los 77 atributos totales entre los que destacan, según su importancia: calificación en diversas áreas del conocimiento (ocho atributos), nivel de motivación, puntaje Recuperado en educación

secundaria, edad, número de hermanos, grupo, hábitos de fumar y promedio del examen de admisión (Márquez, Cano, Romero & Ventura, 2013)

En el segundo, emplearon un conjunto de datos con 20 atributos de 1,966 jóvenes; aplicando selección de atributos obtuvieron los factores más importantes en la deserción: Género, aquellos relacionados con el total de créditos por periodo, créditos aprobados por periodo, promedio general y promedio en matemáticas (Alkhasawneh & Hargraves, 2014).

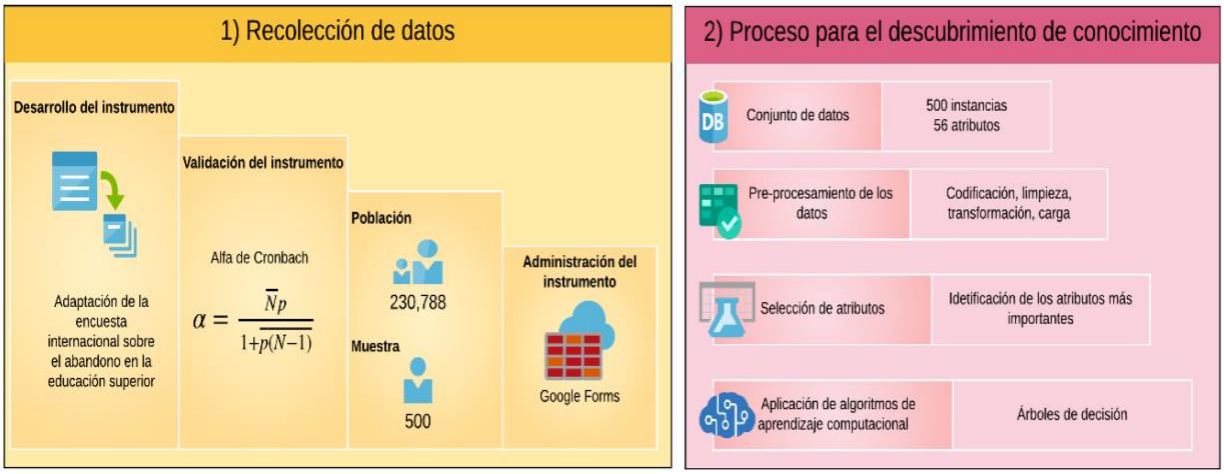
Y en el tercero, usaron un conjunto de datos con 60 atributos de 419 jóvenes de preparatoria y a partir del algoritmo selección de atributos obtuvieron los mejores atributos: promedio en la secundaria, grupo, número de estudiantes en el grupo, edad, asistencia, nivel educativo de la madre, distancia, consumo regular de alcohol, hábitos de fumar, sanciones administrativas, lugar utilizado para estudiar, nivel de motivación, calificación en matemáticas, ciencias sociales y humanidades (Márquez-Vera, et al., 2016).

Finalmente, con estos hallazgos, se puede determinar que el algoritmo árboles de decisión es el más utilizado para resolver situaciones en el contexto educativo cuyo porcentaje mayor de predicción reportado es de 82.87%; que el algoritmo con mejor desempeño al predecir algún evento educativo es máquinas de soporte vectorial con un porcentaje del 87.39% y el algoritmo con el menor desempeño para predecir una situación educativa es regresión lineal con un porcentaje del 66.59%. Estos aciertos, permiten en este trabajo, una oportunidad para identificar aquellos factores más importantes que intervienen en la deserción escolar universitaria a partir de un conjunto de datos de 56 atributos usando selección de atributos, y también, se vislumbra un reto de obtener un mejor desempeño en la aplicación del algoritmo árboles de decisión para contrastar los patrones identificados por los autores citados en la Tabla 3.

Método

La Figura 2 muestra la metodología aplicada para identificar aquellos factores que inciden en la deserción escolar universitaria. La descripción de cada proceso se detalla a continuación:

Figura 2. Metodología para determinar las características más relevantes en la deserción universitaria

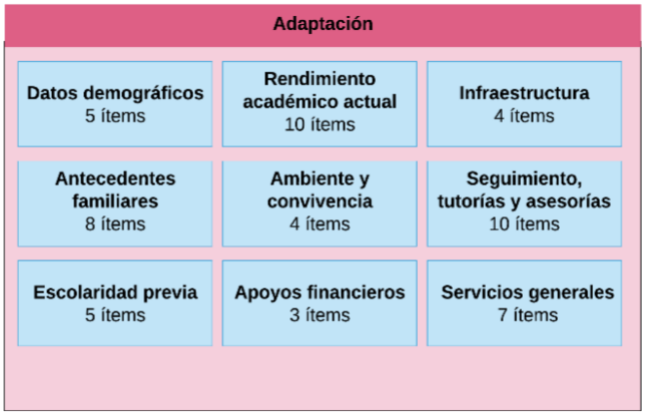


Fuente: Elaboración propia

El instrumento utilizado para la recolección de datos fue adaptado de la <Encuesta internacional sobre abandono en la educación superior> del proyecto Alfa Guía, 2014 (Valle, Eslava, Manzano & García, 2014). En la Figura 3 se presenta la descripción de dicha adaptación.

Figura 3. Adaptación del instrumento para la recolección de datos

Instrumento Proyecto Alfa Guía			
Bloque	**Población**	**No. preguntas**	**Categorías**
0	IES	5	Información sobre institución
		18	Datos estadísticos de institución
1	Encuestado	82	Individual
			Académico
			Social
			Económico
			Cultural
			Institucional
2	Encuestado	10	Posicionamiento
		15	Abandono
3	Encuestado	82	Perfil general



Fuente: Elaboración propia

Como se observa en la Figura 3, el instrumento adaptado tiene 56 preguntas divididas en 9 categorías, mismas que se describen a continuación:

1. Datos demográficos: Recolecta datos referentes a la edad, género, estado civil, número de hijos y empleo.
2. Antecedentes familiares: busca identificar el nivel de estudios del padre/madre/tutor y del hermano/a mayor, así como también la dependencia económica
3. Escolaridad previa: información sobre la preparatoria o bachillerato
4. Rendimiento académico actual: tipo de institución, área, promedio, entre otras.
5. Apoyos financieros: Beca, convenio, crédito educativo y dependencia que los otorga
6. Ambiente y convivencia: Recoge información relacionada al ambiente en la institución
7. Infraestructura: Estima el grado de satisfacción de los espacios universitarios
8. Seguimiento, tutorías y asesorías: Identifica el grado de satisfacción del seguimiento académico proporcionado por la parte académica y administrativa de la institución
9. Servicios: Mide el grado de satisfacción de los servicios generales prestados por la institución en atención al estudiante

A pesar que el instrumento fue adaptado, se consideró validar la fiabilidad de este mediante el coeficiente de Cronbach (Longest, 2019). De este modo, el instrumento tuvo un coeficiente de 0.8767 y cada ítem obtuvo un coeficiente entre 0.8 y 0.9, reflejando con esto una fiabilidad relevante.

Población y muestra

La población se determina considerando 230,788 estudiantes de educación superior del Estado de Puebla, México de acuerdo con datos del Sistema Nacional de Información Estadística Educativa [SNIE] (2019). El cálculo de la muestra fue realizado con un nivel de confianza del 97.5% y un error de muestreo admisible del 5%. Se ha utilizado la Fórmula 1 de muestreo aleatorio simple por proporciones de poblaciones finitas propuesta en Morillas (2014).

Fórmula 1

$$n = \frac{N\, Z^2_{1-\alpha/2}\, pq}{(N-1)\varepsilon^2 + Z^2_{1-\alpha/2}\, pq}$$

Donde: n = tamaño de la muestra; N = Tamaño de la Población; Z = Nivel de confianza; p = Probabilidad de éxito o proporción esperada, 0.5 cuando se desconoce el valor; q = Probabilidad de fracaso; ε = Error admisible. Por tanto, el valor de la muestra se observa en la Ecuación 1, de manera que se ha truncado a 500 encuestados.

Ecuación 1

$$n = \frac{(230{,}788)(2.24)^2(0.5)(0.5)}{(230{,}788-1)(0.05)^2 + (2.24)^2(0.5)(0.5)} = 500.7$$

Se utilizó un muestreo aleatorio simple en diversas universidades públicas y privadas de la Ciudad de Puebla para solicitar el llenado en línea del instrumento. Para poder obtener el total de muestras requeridas se solicitó permiso a los directivos de diversas universidades ubicadas en la Ciudad de Puebla y Ciudades circunvecinas. Se administró en aquellas universidades donde se tenía el permiso por escrito (sin importar si fuesen públicas o privadas) y a los grupos de estudiantes que se encontraban de manera presencial tomando clase en un laboratorio de cómputo (sin importar la carrera y semestre en que se encontraban inscritos). De esta manera, se obtuvo la participación de 300 estudiantes de IES públicas y 200 estudiantes de IES privadas. De los cuales, 311 son mujeres (176 de IES públicas y 135 IES privadas) y 188 son hombres (124 de IES públicas y 65 de IES privadas).

Administración del instrumento. Es preciso mencionar que se ha diseñado un formulario en *Google forms* (Google, 2019), para la administración del instrumento de forma tal que sea más sencillo recolectar la información para facilitar el proceso de descubrimiento del conocimiento (*véase* el formulario en el hipervínculo https://bit.ly/3dVKgJy).

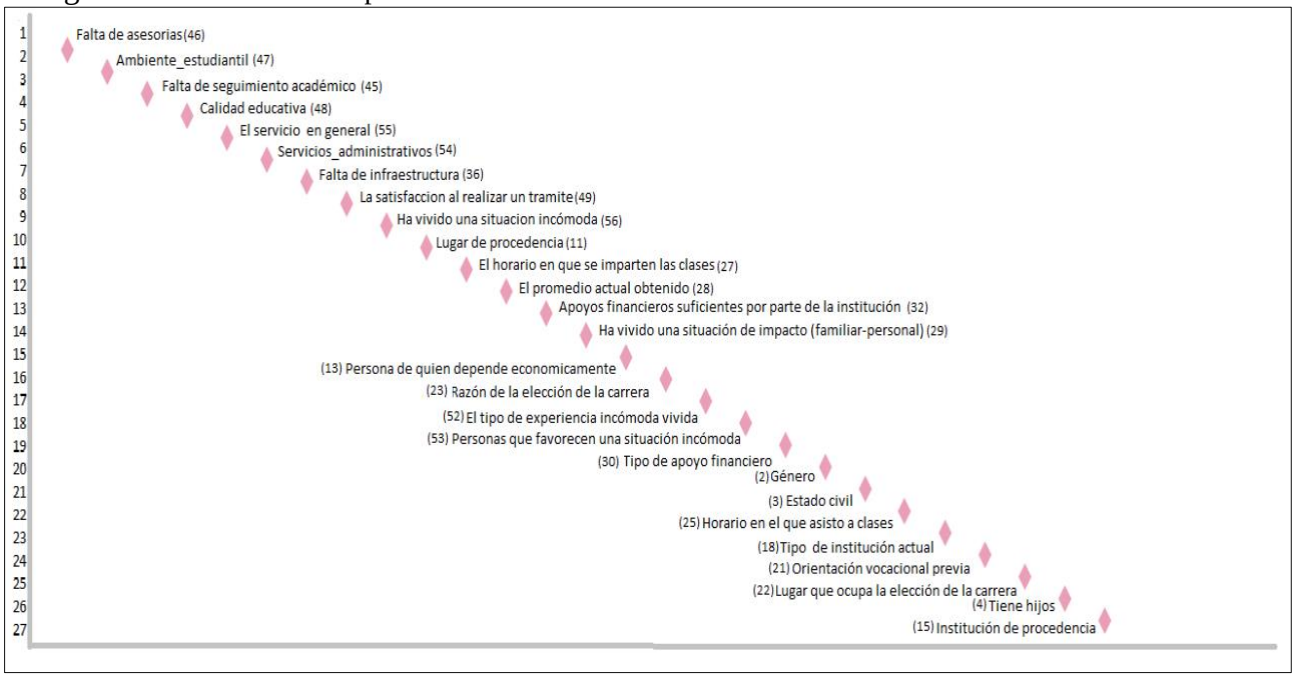
Resultados

En esta sección se presenta la aplicación del proceso para el descubrimiento de conocimiento en dos fases: 1) Identificación de los factores más relevantes aplicando evaluadores de selección de atributos y 2) Reconocimiento de patrones de la deserción universitaria con árboles de decisión.

Factores más relevantes de inciden en la deserción universitaria

Como se ha descrito con anterioridad, una de las fases del proceso para el descubrimiento del conocimiento es la selección de atributos. En la Figura 4 se presentan los resultados de haber aplicado los dos evaluadores (*GainRatioAttributeEval* e *InfoGainAttributeEval*), en la cual se denota una lista de los 27 principales atributos del conjunto de datos que contiene 56 atributos. Es notorio que el ranking de ambos evaluadores considera como importantes a los mismos atributos a pesar de otorgarles diferentes pesos. Dado que los resultados de ambos evaluadores coinciden, serán considerados para tipificar las reglas que pueden favorecer la identificación de una temprana deserción a partir de los árboles de decisión.

Figura 4. Atributos más importantes en la deserción escolar universitaria usando selección de atributos

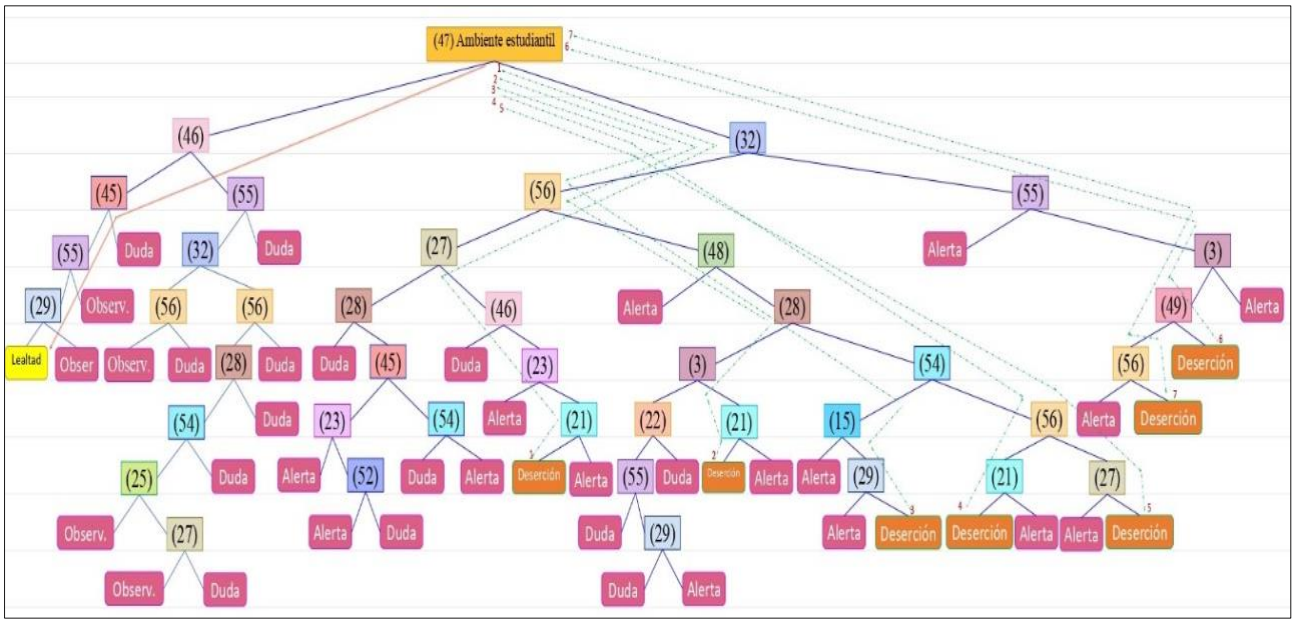


Fuente: Elaboración propia

Patrones de una inminente deserción

En esta fase se presentan los hallazgos encontrados una vez que se ha aplicado el algoritmo árboles de decisión C4.5 en sus dos fases; en la primera fase se ha utilizado el conjunto de datos completo, es decir, con los 56 atributos obteniendo un desempeño del 74.6% de exactitud (Tabla 6). En la segunda fase, se ha modificado el conjunto de datos basados en la aplicación de la selección de atributos descrita anteriormente. En otras palabras, se ha aplicado nuevamente el algoritmo árboles de decisión C4.5 al conjunto de datos solamente con los 27 atributos más relevantes de acuerdo con el algoritmo selección de atributos, obteniendo un desempeño de 92.6% de exactitud.

Figura 5. Árbol de decisión con la mejor clasificación y causas que predicen una inminente deserción universitaria.



Fuente: Elaboración propia

En la Figura 5 se muestra el árbol obtenido con los mejores atributos y las reglas que definen un patrón para aquellos jóvenes que tienen lealtad y convicción de permanecer estudiando. Dicho patrón está relacionado con la regla siguiente:

47_AMBIENTE ESTUDIANTIL <= 3
| 46_ FALTA DE ASESORÍAS<= 2
| | 45_ FALTA DE SEGUIMIENTO ACADÉMICO<= 2
| | | 55_SERVICIO GENERAL <= 1
| | | | 29_ HA VIVIDO UNA SITUACIÓN DE IMPACTO<= 1: LEALTAD

La cual se lee de la siguiente manera: SI el ambiente estudiantil es satisfactorio a neutral y la falta de asesoría y la falta de seguimiento académico casi no existe y el servicio en general es satisfactorio y no ha vivido una situación de impacto (muerte de un familiar, divorcio de padres, ruptura sentimental) ENTONCES permanezco estudiando. En otras palabras, un estudiante universitario decide continuar con sus estudios si existe un ambiente estudiantil que totalmente satisfactorio o satisfactorio, si las asesorías, seguimiento académico y servicios en general son totalmente satisfactorios o satisfactorios y

no ha vivido una situación de impacto. Por el contrario, para identificar una inminente deserción se han encontrado siete reglas (Figura 5). La regla 1 y la regla 2, por citar ejemplos, son descritas a continuación:

Regla 1

47_AMBIENTE ESTUDIANTIL > 3
| 32_ APOYOS FINANCIEROS INSUFICIENES<= 4
| | 56_ HA VIVIDO UNA SITUACIÓN INCÓMODA<= 3
| | | 27_HORARIO EN EL QUE SE IMPARTEN LAS CLASES >4
| | | | 46_ FALTA DE ASESORIAS >2
| | | | | 23_RAZÓN DE LA ELECCIÓN DE LA CARRERA>2
| | | | | | 21_RECIBIO ORIENTACIÓN PREVIA<=1:DESERCIÓN

Regla 2

47_AMBIENTE ESTUDIANTIL > 3
| 32_ APOYOS FINANCIEROS INSTITUCIONALES<= 4
| | 56_ HA VIVIDO UNA SITUACIÓN INCÓMODA> 3
| | | 48_SERVICIO GENERAL >2
| | | | 28_ PROMEDIO ACTUAL >4
| | | | | 54_SERVICIOS ADMINISTRATIVOS <=4
| | | | | | 15_INSTITUCIÓN DE PROCEDENCIA = PÚBLICA
| | | | | | | 29_HA VIDIDO UNA SITUACIÓN DE IMPACTO >2: DESERCIÓN

La regla 2 se puede explicar de la siguiente forma: SI el ambiente estudiantil es neutral a totalmente insatisfactorio y los apoyos financieros proporcionados por la institución son insuficientes, ha experimentado una situación incómoda (acoso, discriminación, maltrato), el servicio general y el promedio general son insatisfactorios y los servicios administrativos son totalmente satisfactorios o satisfactorios y procede de una preparatoria pública y ha vivido una situación de impacto ENTONCES decide desertar.

Para conocer el desempeño del algoritmo árboles de decisión, clasificando a la clase "pronóstico" identificada por 1: Lealtad, 2: Observancia, 3: Duda, 4: Alerta y 5: Deserción, se encuentran las dos matrices de clasificación obtenidas una vez que se ha probado con el conjunto de datos a) 56 atributos y b) 27 atributos más relevantes (Tabla 5).

Tabla 5. Matriz de confusión Recuperados con a) 56 atributos y b) 27 atributos más relevantes

a) Clasificación con 56 atributos					
a	**b**	**c**	**d**	**e**	**Clasificación**
65	0	24	0	4	a=Observancia
3	102	29	13	0	b=Alerta
21	30	115	3	0	c=Duda
1	19	0	42	0	d=Deserción
3	0	0	0	4	e=Lealtad

b) Clasificación con 27 atributos					
a	**b**	**c**	**d**	**e**	**Clasificación**
88	0	4	0	1	a=Observancia
1	135	10	1	0	b=Alerta
4	6	179	0	0	c=Duda
0	8	1	53	0	d=Deserción
1	0	0	0	8	e=Lealtad

Fuente: Elaboración propia

En la Tabla 5 se han seleccionado los valores en la diagonal de cada una de las matrices, la cual indica las instancias correctamente clasificadas en función de la clase pronóstico (clasificación). Por ejemplo, si se desea hacer un comparativo entre las matrices relacionadas al atributo <Deserción>, se observa que en la matriz *a*, 42 instancias fueron correctamente clasificadas como deserción, mientras que 1 fue incorrectamente clasificada como observancia y 19 instancias como alerta; en comparación con la matriz *b*, donde 53 instancias fueron clasificadas correctamente como deserción, 8 fueron clasificadas como alerta y 1 como duda. De esta manera, en la matriz *b*, el número de instancias correctamente clasificadas fue mayor que en la matriz *a*, situación que se nota para el resto de las clasificaciones. Del mismo modo, para conocer las métricas que estiman el desempeño del algoritmo, se presentan en la Tabla 6.

Tabla 6. Métricas de desempeño del árbol de decisión aplicado en dos fases.

Clasificación	Exactitud	Precisión	Recuperación	Medida-F	Instancias correctamente clasificadas
56 atributos	74.6	74.9	74.6	74.6	373
27 atributos	92.6	92.7	92.6	92.6	463

Se observa que utilizando el algoritmo selección de atributos previamente al algoritmo C4.5, la exactitud y precisión del algoritmo aumenta considerablemente; sin perder de vista que lo más importante es conocer cuáles son los factores más importantes que se deben considerar como alerta en la decisión de desertar, así como aquellas reglas o patrones que un estudiante universitario sigue para desertar de sus estudios.

Discusión y conclusiones

Como resultado de la aplicación de la minería de datos educativa, es posible concluir que las cinco principales causas de la deserción escolar universitaria se deben a la falta de asesorías, a un inadecuado ambiente estudiantil, a la falta de seguimiento académico, a la deficiente calidad educativa y al servicio en general. Estos hallazgos contrastan sensiblemente con los encontrados en la literatura en donde se indica que el primer factor es el disgusto por la carrera, seguido de la falta de recursos económicos, cambio de estado civil, distancia de la casa al centro de estudios, razones de tipo familiar y otras causas.

También, mediante la utilización de los evaluadores para la selección de atributos se pudo identificar que los aspectos relacionados con el disgusto por la carrera se encuentran en el lugar 13 y 19 de los 27 principales factores encontrados, así también, que los factores relacionados al aspecto económico ocupan los lugares 17 y 25 del ranking Recuperado a través del uso del algoritmo. Esto indica que los jóvenes universitarios han ampliado la gama de factores que satisfacen el deseo de permanecer estudiando, al menos para la muestra obtenida, y que discrepan con los estudios analizados en la Tabla 2.

Por otro lado, gracias a la aplicación del algoritmo árboles de decisión se ha logrado establecer una serie de 7 patrones que conducen a una deserción escolar; de la misma manera, se han logrado clasificar patrones que son dignos de atención mostrados en la Figura 5 como aquellos en donde se encuentre duda, observancia o alerta. Igualmente, se ha determinado el patrón que el estudiante universitario defiende para permanecer estudiando, el cual considera que exista un ambiente estudiantil adecuado, que le proporcionen asesorías y seguimiento académico, que los servicios en general sean adecuados y que no haya vivido una situación de impacto que le genere duda en la decisión de continuar con sus estudios.

Se sabe que la mayoría de IES han implementado acciones tutoriales y de asesorías, sin embargo, pareciera ser que el número asignado para cada profesor/tutor se ha visto rebasado en cuanto al número de estudiantes que deben atender y, por tanto, el servicio es deficiente o nulo. Los hallazgos obtenidos en este estudio, dan la pauta para que las IES apuesten por crear mecanismos que apoyen el servicio de asesorías mediante acciones que coadyuven a ofrecer mejores servicios, como la inclusión de estudiantes de últimos semestres que orienten a los de primeros semestres; del mismo modo, que profesores ofrezcan tutorías individualizadas para atender asuntos académicos y poder canalizar a instancias psicopedagógicas a estudiantes que se hayan detectado con alguna deficiencia o situación emocional de riesgo.

Asimismo, fomentar un clima educativo positivo a través del acercamiento a la comunidad universitaria, padres de familia, en donde la convivencia sea un factor que predomine en acrecentar la comunicación, el respeto, la inteligencia emocional, la solución de conflictos y el acercamiento de persona a persona para lograr sensibilidad, motivación y empatía. Al mismo tiempo, el fortalecimiento de la calidad educativa es crucial para garantizar que los estudiantes han adquirido las competencias y habilidades requeridas para incursionar con éxito en el ámbito laboral y evitar la deserción escolar en el primer año.

Finalmente, se ha visto que la deserción escolar no depende de un solo factor, sino que es causada por un conjunto de factores e interacción de estos, tal como se ha observado en los patrones Recuperados mediante el árbol de decisión C4.5; se logra deducir que dichos patrones pueden variar aún en contextos similares, debido a la región, lugar de procedencia, nivel socioeconómico o incluso creencias, según lo revisado en la literatura (Tabla 2).

De esta manera, este estudio proporciona una introducción a la minería de datos educativa para encontrar patrones que prevengan una inminente deserción y actuar antes que suceda. Por ello, para enriquecerlo es imperativo ampliar la muestra a otros estados o ciudades de manera que se puedan aplicar diversos algoritmos que proporcionen mayor información que conduzcan al establecimiento de mecanismos certeros para disminuir los índices de deserción universitaria que año con año se han estado reportando y que a pesar que los gobiernos federales o estatales implementan mecanismos para ello, no han sido suficientes para disminuir estos índices y que al contrario, han aumentado en los últimos ciclos escolares (2016-2019).

Referencias

Abarca R., A., & Sánchez V., M. A. (2005). La deserción estudiantil en la educación superior: el caso de la Universidad de Costa Rica. *Revista Electrónica "Actualidades Investigativas en Educación"*, *5*, 1-22. https://bit.ly/35TVeLE

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. International Journal of Data Mining y Knowledge Management Process (IJDKP), 5(1), 15-27. https://doi.org/10.5121/ijdkp.2015.5102

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

Al-Barrak, M. A., & 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. https://doi.org/10.7763/IJIET.2016.V6.745

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques. *Journal of STEM Education: Innovations and Research*, *5*(3), 35-42. ERIC. https://bit.ly/2Rd04hi

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting Student Dropout in

Higher Education. *Machine Learning in Social Good Applications*, 16-20. https://bit.ly/3aRtae6

Barbosa M. L. M., Serra da Cruz, S. M., & Zimbrão, G. (2014). The Impact of High Dropout Rates in a Large Public Brazilian University: A Quantitative Approach Using Educational Data Mining. *6th International Conference on Computer Supported Education* (págs. 124-129). Barcelona, Spain: INSTICC. https://bit.ly/2ZsYFbD

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. USA: O´Really Media, Inc.

Bishop, C. M. (2007). Pattern recognition and Machine Learning. Singapore: Springer.

Cabrera, L., Bethencour, J. T., Álvarez P. P., & González A. M. (2006). El problema del abandono de los estudios universitarios. *RELIEVE*, *12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Carvajal O. P., & Trejos C. Á. A. (2016). Revisión de estudios sobre deserción estudiantil en educación superior en Latinoamérica bajo la perspectiva de Pierre Bourdieu. *Congresos CLABES*. Quito, Ecuador: Escuela Politécnica Nacional. https://bit.ly/2UP9mlT

Cha, G.-W., Kim, Y.-C., Moon, H. J., & Hong, W.-H. (2017). New approach for forecasting demolition waste generation using chisquared automatic interaction detection (CHAID) method. Journal of Cleaner Production, 168, 375-385. https://doi.org/10.1016/j.jclepro.2017.09.025

Chen, W., Xie, X., Peng, J., Wang, J., Duan, Z., & Hong, H. (2017). GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. Geomatics, Natural Hazards and Risk, 8(2), 950-973. https://doi.org/10.1080/19475705.2017.1289250

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

Dekker, G. W., Pechenizki, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *2nd International Conference on Educational Data Mining* (págs. 41-50). Cordoba, Spain: International Educational Data Mining Society. https://bit.ly/2ZlH1a3

Delen, D. (2011). Predicting Student Attrition with Data Mining Methods. *Journal of College Student Retention: Research, Theory y Practice*, *13*(1), 17-35. https://doi.org/10.2190/CS.13.1.b

Del Pobil, A. P., Mira, J., & Ali, M. (1998). Tasks and Methods in Applied Artificial Intelligence. 11th International Conference on Industrial and Engineering Applications of Artificial In telligence and Expert Systems. 1416. Castellón, España: Springer.

Estrada-Danell, R. I., Zamarripa-Franco, R. A., Zúñiga-Garay, P. G., & 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. https://doi.org/10.15359/ree.20-3.11

Fozdar, B. I., Kumar, L. S., & Kannan, S. (2006). A Survey of a Study on the Reasons Responsible for Student Dropout from the Bachelor of Science Programme at Indira Gandhi National Open University. *International Review of Research in Open and Distance Learning*, *7*(3), 1-15. https://doi.org/10.19173/irrodl.v7i3.291

Frank, E., Hall, Mark A., & Witten I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Freitas, A. A. (2002). Data Mining and Knowledge Discovery with Evolutionary Algorithms. The Netherlands: Springer-Verlag. https://doi.org/10.1007/978-3-662-04923-5

Guevara, C., Sanchez-Gordon, S., Arias-Flores, H., Varela-Aldás, J., Castillo-Salazar, D., Borja, M., . . . Yandún-Velasteguí, M. (2019). Detection of Student Behavior Profiles Applying Neural Networks and Decision Trees. 1026, págs. 591-597. Munich, Germany: Springer, Cham. https://doi.org/10.1007/978-3-030-27928-8_9

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications, 163(8), 15-19. https://doi.org/10.5120/ijca2017913660

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Amsterdam: Morgan Kaufmann.

INEGI. (2018). Estadísticas a propósito del día mundial de la población (11 de julio). Ciudad de México: INEGI. https://bit.ly/2xbnZHd

Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research (IJSR), 5(1), 1842-1845. https://doi.org/10.21275/v5i1.NOV153131

Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Trees. *International Journal of Computer Applications, 36*(11), 9-12. https://bit.ly/2JdxckV

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. *Knowledge-Based Intelligent Information and Engineering Systems, 7th International Conference* (págs. 267-274). Oxford, UK.: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-45226-3_37

Kumar Y. S., Bharadwaj, B., & Pal, S. (2012). Mining Education Data to Predict Student's Retention: A comparative Study. *International Journal of Computer Science and Information Security, 10*(2), 113-117. https://bit.ly/2JJw9t1

Lavado, P., & Gallegos, J. (2005). La dinámica de la deserción escolar en el Perú: un enfoque usando modelos de duración. Lima, Perú: Universidad del Pacífico. https://bit.ly/39PH3rJ

Londoño A. L. F. (2013). Factores de riesgo presentes en la deserción estudiantil en la Corporación Universitaria Lasallista. *Revista Virtual Universidad Católica del Norte* (38), 183-194. https://bit.ly/1OnjEwM

Longest, K. C. (2019). Using Stata for Quantitative Analysis. California, USA: SAGE Publications.

M.P. van der Aalst, W. (2011). Process Mining: Discovery, Conformance and Enhancement of Business Processes (Google eBook). London, UK: Springer-Verlag. https://doi.org/10.1007/978-3-642-19345-3

Márquez-Vera C., Romero M. C., & Ventura S. S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *Revista Iberoamericana de Tecnologías del Aprendizaje, 7*(3), 109-117. https://bit.ly/2zoZKmo

Márquez-Vera, C., Cano, A., Romero, C., Mohammad N. A. Y., Fardoun, H. M., & 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. (1997). Machine Learning. Singapore: McGraw-Hill.

Mitchell, T. M. (2000). *Decision Tree Learning.* Washington State University. https://bit.ly/2N1AI32

Morales C. J., & Parraga-Alava, J. (2018). How Predicting The Academic Success of Students of the ESPAM MFL?: A Preliminary Decision Trees Based Study. 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM). Cuenca,

Ecuador: IEEE. https://doi.org/10.1109/ETCM.2018.8580296

Morillas, A. (2014). Muestreo en poblaciones finitas. Notas del curso. Málaga-España: Universidad de Málaga. https://bit.ly/2JLLA3K

OECD. (2017). *Skills Strategy Diagnostic Report: Mexico 2017, OECD Skills Studies*. París: OECD Publishing. https://doi.org/10.1787/9789264287679-en

OECD. (2019). *Higher Education in Mexico: Labour Market Relevance and Outcomes, Higher*. París: OECD Publishing. https://doi.org/10.1787/9789264309432-en

Pal, S. (2012). Mining Educational Data Using Classification to Decrease Dropout Rate of Students. *International Journal of Multidisciplinary Sciences and Engineering, 3*(5), 35-39. https://bit.ly/2xVhAjc

Raju, D. y Schumacker, R. (2015). Exploring Student Characteristics of Retention that Lead to Graduation in Higher Education Using Data Mining Models. Journal of college student retention: Research, Theory y Practice, 16(5), 563-591. https://doi.org/10.2190/CS.16.4.e

Rodríguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O., & Suárez-Carranza, B. A. (2017). Modeling Students' Dropout in Mexican Universities. *Research in Computing Science, 139*, 163-175. https://doi.org/10.13053/rcs-139-1-13

Ruíz C., L. (2009). Deserción en la educación superior recinto Las Minas. Período 2001-2007. *Ciencia e Interculturalidad, 4*(2), 30-46. https://doi.org/10.5377/rci.v4i1.288

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (págs. 319-324). Bruges, Belgium: i6doc.com. https://bit.ly/2MFgzkp

Secretaría de Educación Pública. (2019). Abandono escolar. Ciudad de México: SEP.

Secretaría de Educación Pública. (2019). Principales cifras del sistema educativo nacional 2018-2019. Ciudad de México: Dirección General de Planeación, Programación y Estadística. https://bit.ly/2yCwivX

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097. https://doi.org/10.21275/v5i4.NOV162954

Sivakumar, S., Venkataraman, S., & 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

Universidad Tecnológica de Tabasco. (2019). Glosario de Términos. Villermosa, Tabasco: Universidad Tecnológica de Tabasco. https://bit.ly/2xZ60DK

Ustebay, S., Turgut, Z., & Ali A. M. (2018). Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier. 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT) (págs. 71-76). Ankara, Turkey: IEEE. https://doi.org/10.1109/IBIGDELFT.2018.8625318

Vélez, A., & López, J. D. F. (2004). Estrategias para vencer la deserción universitaria. *Educación y Educadores* (7). 177-203. https://bit.ly/39MgeEJ

Valle G. R., Eslava G. G., Manzano P. A., & García M. M. (2014). Encuesta Internacional sobre el Abandono en la Educación Superior. Unión Europea. https://bit.ly/2p8k2Pk

Vijayalakshmi, M., & Kumar, A. S. (2011). Efficiency of decision trees in predicting student's academic performance. Computer Science y Information Technology, 335-343. https://doi.org/10.5121/csit.2011.1230

Vries, W., León A. P., Romero M. J. F., & Hernández S. I. (2011). ¿Desertores o

decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de la Educación Superior, 40*(160), 29-49. https://bit.ly/1TzOzru

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. San Francisco, CA: ELSEVIER.

Witten, I. H., Frank, E., & Hall, M.A. (2011). Data mining: Practical machine learning tools and techniques (3a. ed.). Morgan Kaufmann Publishers, Burlington. https://doi.org/10.1016/B978-0-12-374856-0.00001-8

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

Yang, S., Lu, O., Huang, A., Huang , J., Ogata, H., & Lin, A. (2017). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. Journal of Information Processing, 170-176. https://doi.org/10.2197/ipsjjip.26.170

Yukselturk, E., Ozekes, S., & Kılıç T. 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

Authors / Autores

Urbina-Nájera, A.B. (abunajera@gmail.com) 0000-0002-3700-7287

UPAEP-Universidad. Argelia B. Urbina Nájera, pertenece al Sistema Nacional de Investigadores Mexicano en el nivel Candidato. Sus líneas de investigación se enfocan en la aplicación de aprendizaje computacional, ciencia de datos e inteligencia de negocios en el ámbito educativo, salud y actividades comerciales y tecnología educativa. Obtuvo el grado de Doctora en Planeación Estratégica y Dirección de Tecnología por la Universidad Popular Autónoma del Estado de Puebla (UPAEP), tiene el grado de Maestra en Ciencias en Ingeniería de la Computación por la Universidad Autónoma de Tlaxcala; el grado de Maestra en Ciencias de la Educación por el IEU y la Licenciatura en Ciencias de la Computación por la Benemérita Universidad Autónoma de Puebla (BUAP). Actualmente es Profesora-Investigadora de Tiempo Completo adscrita al decanato de ingenierías en la UPAEP.

Camino-Hampshire, J.C. (josecarlos.camino@upaep.edu.mx) 0000-0002-4686-494

Accenture (México) & UPAEP-Universidad (México). José Carlos Camino Hampshire es Ingeniero Industrial Administrador por la (UPAEP), cuenta con una Maestría en Logística y Dirección de la Cadena de Suministro y la Maestría en Ciencia de Datos e Inteligencia de Negocios en la misma institución. Actualmente labora en la compañía de consultoría Accenture México como gerente en el área de Cadena de Suministro para empresas de la Industria de Productos (Empresas de consumo masivo, servicios, retail, hotelería, automotriz, entre otras).

Cruz Barbosa, R. (rcruz@mixteco.utm.mx) 0000-0002-5494-7027

Universidad Tecnológica de la Mixteca (México). Raúl Cruz-Barbosa tiene estudios de Licenciatura y Maestría por la Universidad Autónoma de Puebla, México. También cuenta con el doctorado en Inteligencia Artificial por la Universidad Politécnica de Cataluña, España. El Dr. Cruz-Barbosa es miembro del Sistema Nacional de Investigadores mexicano. Sus intereses de investigación están relacionados con aprendizaje computacional a gran escala, procesamiento digital de imágenes, minería de datos y reconocimiento de patrones así como su aplicación en Educación, Bioinformática y detección y diagnóstico asistido por computadora.

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]