Ciencia Latina Revista Científica Multidisciplinar
Septiembre-Octubre, 2023, Volumen 7, Número 5
https://doi.org/10.37811/cl_rcm.v7i5.8363

Modelo Predictivo de la Deserción Escolar en Educación Superior: una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM

M.C. Simón Guadalupe Cornejo Sifuentes[1]
simon.cs@mochis.tecnm.mx
https://orcid.org/0009-0006-3408-1592
Tecnológico Nacional de México/ Instituto Tecnológico de Los Mochis
Los Mochis – México

Dra. Ligia Gabriela Vega Pérez
ligia.vp@mochis.tecnm.mx
https://orcid.org/0000-0001-7673-1430
Tecnológico Nacional de México/ Instituto Tecnológico de Los Mochis
Los Mochis - México

MDC. María Guadalupe Naranjo Cantabrana
maria.nc@mochis.tecnm.mx
https://orcid.org/0009-0005-0623-9543
Tecnológico Nacional de México/ Instituto Tecnológico de Los Mochis
Los Mochis – México

Ing. Isabel Francisca Osúa Acosta
isabel.oa@vyaqui.tecnm.mx
https://orcid.org/0009-0009-8531-6170
Tecnológico Nacional de México/ Instituto Tecnológico del Valle del Yaqui
Valle del Yaqui – México

Ing. Fausto Alberto Ávila Santana
fausto.as@mochis.tecnm.mx
https://orcid.org/0009-0000-7434-4728
Tecnológico Nacional de México/ Instituto Tecnológico de Los Mochis
Los Mochis - México

M.C. María de los Ángeles Sotomayor Fierro
maria.sf@mochis.tecnm.mx
https://orcid.org/0009-0005-5037-7409
Tecnológico Nacional de México/ Instituto Tecnológico de Los Mochis
Los Mochis – México

RESUMEN

Este artículo presenta el desarrollo de un modelo para predecir, de manera temprana y oportuna, casos de estudiantes que muestren un potencial riesgo de deserción escolar, mediante el uso de técnicas de minería de datos. La deserción escolar a nivel superior es un problema multifactorial y complejo de analizar por la intervención de elementos de diversa índole, como factores familiares, académicos, educacionales, la situación económica familiar, las habilidades intelectuales de los estudiantes o la didáctica de los profesores. Este gran volumen de información a analizar no es fácilmente manejable con técnicas estadísticas tradicionales, sino que se precisa buscar estrategias que permitan operar con los bancos de datos de modo más eficiente y rápido. En el desarrollo de la propuesta se aplicó de una manera novedosa la minería de datos, para explorar los cambios en los comportamientos de los estudiantes, vinculados a diferentes causas de abandono escolar, utilizando la metodología CRISP-DM, con datos de 1,374 estudiantes de una institución de educación superior. Los resultados muestran las técnicas utilizadas para identificar y seleccionar factores asociados a la deserción estudiantil y los algoritmos para generar los modelos predictivos, de los cuáles pudo seleccionarse el más preciso, con mayor puntuación y facilidad de interpretación.

Palabras clave: *deserción escolar; educación superior; minería de datos; modelos predictivos; metodología CRISP-DM*

[1] Autor principal
Correspondencia: simon.cs@mochis.tecnm.mx

Predictive Model of School Dropouts in Higher Education: An Approach From Data Mining Using the CRISP-DM Methodology

ABSTRACT

This article presents the development of a model to predict, in an early and timely manner, cases of students who show a potential risk of dropping out of school, through the use of data mining techniques. High school dropout is a multifactorial and complex problem to analyze due to the intervention of diverse elements, such as family, academic, educational factors, the family economic situation, the intellectual abilities of the students or the didactics of the teachers. This large volume of information to be analyzed is not easily manageable with traditional statistical techniques, but it is necessary to find strategies that allow operating with data banks more efficiently and quickly. In the development of the proposal, data mining was applied in a novel way to explore changes in student behaviors, linked to different causes of school dropout, using the CRISP-DM methodology, with data from 1,374 students from an institution of higher education. The results show the techniques used to identify and select factors associated with student dropout, and the algorithms to generate predictive models, from which the most precise one could be selected, with the highest score and ease of interpretation.

Keywords: *school dropout; higher education; data mining; predictive models; CRISP-DM methodology*

Artículo recibido 14 setiembre 2023
Aceptado para publicación: 25 octubre 2023

INTRODUCCIÓN

Uno de los pilares del desarrollo económico y social de un país es la educación; la población que tiene la oportunidad de estudiar una carrera profesional, por lo general, logra un mejor empleo y mayores ingresos económicos que las personas que no lo hicieron, o que nomás cursaron educación básica (Spring, 1998). En algunos países, ciertos sectores de la población no cuentan con oportunidades educativas, por lo que se ven limitados en su desarrollo como personas y como sociedad. El acceso a la educación es, sin lugar a duda, trascendental (Camarena, Saavedra, & Saldívar, 2015), sin embargo, el hecho de que los estudiantes culminen su formación profesional y no fracasen en el intento por conseguirlo, tiene igual trascendencia, pues les permite contar con las herramientas y competencias necesarias para hacer frente a los retos y a las oportunidades futuras.

En lo concerniente a México, en el periodo escolar 2021- 2022, las instituciones educativas de sostenimiento público y particular, a nivel licenciatura, atendieron a 5’068,493 estudiantes, según la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES, 2021), distribuidos en 3,057 instituciones educativas. Según el último censo del Instituto Nacional de Estadística Geografía e Informática (INEGI) en el año 2020, la población de jóvenes entre 15-19 años era de 10’806,690, mientras que la de jóvenes entre 20-24 era de 10’806,690 (INEGI, 2021). Por lo anterior, se puede inferir que nomás el 30% de los jóvenes entre 18-23 años estudia una licenciatura, encima, es muy lamentable que algunos de ellos no logren finalizarla.

De acuerdo con Himmel (2002), la deserción escolar es el abandono del curso del plan de estudios de una institución educativa, antes de lograr obtener un certificado o título, sin que el estudiante tenga alguna posibilidad de volverse a integrar a la institución. Las causas de la deserción escolar de un estudiante pueden ser por múltiples factores y de diversa índole como la académica, económica, social, psicológica o de interacción. Sin importar cuales sean las causas que intervengan o confluyan para que un alumno decida no concluir sus estudios universitarios, la institución pierde un estudiante, y se ve afectada en sus indicadores institucionales, porque disminuye el número de alumnos egresados, lo cual se reflejará en los índices de su Eficiencia Terminal (ET).

Cabe destacar que algunos investigadores han diseñado modelos conceptuales, desde perspectivas diferentes, según los factores que influyan para que los estudiantes tomen la decisión de desertar o

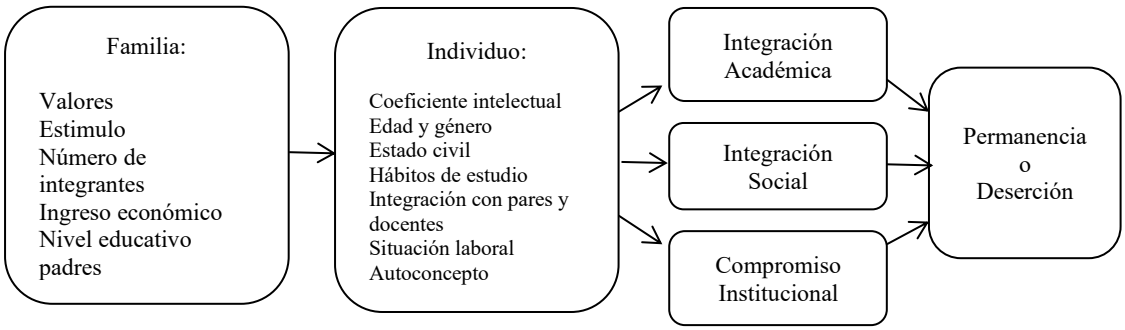
culminar sus estudios. Spady (1970), mantiene un enfoque social en donde la familia, la interacción del alumno con sus compañeros de clases y su desarrollo intelectual, tienen gran influencia en las tasas de graduación estudiantil. Ethington (1990) sostiene un enfoque psicológico en la culminación de los estudios profesionales de los estudiantes, el cual incluye los antecedentes familiares, el antecedente académico, el estímulo y apoyo familiar, el concepto académico de sí mismo y la percepción de la dificultad de los estudios.

A propósito de lo anterior, Cabrera, Nora y Asker (1999) mencionan que los factores económicos podrían determinar la permanencia en los estudios superiores, cuando el estudiante percibe mayores beneficios económicos y sociales que los que obtendría en otras actividades, como un trabajo. Vincent Tinto (1975) es uno de los autores más citados, es quien amplió el modelo creado por William G. Spady. Según Tinto, los estudiantes actúan de acuerdo con la teoría del intercambio en la construcción de su integración social y académica. Este autor considera que mientras el alumno va realizando sus estudios, diversos factores influyen para que se adapte a la institución educativa. Estos factores comprenden antecedes familiares como el nivel socioeconómico y cultural, atributos personales y experiencia preuniversitaria.

En suma, los modelos estudiados exponen que la base familiar es fundamental para garantizar la continuidad de los estudios de los jóvenes y que las características de la familia del alumno guardan relación con los factores a considerar en la deserción, según se muestra en la Figura 1.

Figura 1

Factores de los diferentes enfoques para el análisis de la deserción escolar



Por otro lado, la Minería de Datos (DM, por sus siglas en inglés Data Mining) es un área multidisciplinaria utilizada exitosamente para resolver problemas en muchos campos de aplicación como los negocios, la ciencia, la salud, la educación, entre otros. DM se define como un proceso de

extracción de patrones válidos, no triviales, implícitos, previamente desconocidos y potencialmente útiles. En otras palabras, la DM hace referencia al descubrimiento de conocimiento a partir de los datos (Han, 2007). En la DM confluyen múltiples disciplinas tales como Máquinas de Aprendizaje, Reconocimiento de Patrones, Métodos Estadísticos, Algoritmos, Visualización, Cómputo de Alto Rendimiento, Tecnología de Bases de Datos, entre otros, y desde hace relativamente poco tiempo también se han utilizado algoritmos de cómputo suave tales como los algoritmos evolutivos.

Adicionalmente, entre las diferentes tareas que pueden realizarse con DM están la descripción, la predicción, la segmentación y la asociación. Particularmente la predicción puede realizarse por medio de la clasificación, la cual es una de las actividades que más frecuentemente realiza el ser humano en su vida cotidiana. La clasificación ocurre cuando es necesario asignar un objeto a un grupo predefinido o clase, lo cual se decide basándose en un determinado número de atributos que se observan en el objeto (Zhang, Oussen, Clark, & Kim, 2010). El objetivo de la clasificación es inducir un modelo para predecir a qué clase pertenece un objeto.

Profundizando en el tema, un caso concreto de aplicación de DM es en el campo de la educación, donde se le denomina Minería de Datos Educativa (EDM, por sus siglas en ingles Educational Data Mining) (Romero & Ventura, 2010). Actualmente, la EDM se está aplicando para tratar problemas en los sistemas tradicionales de educación como el predecir el rendimiento académico de los estudiantes, en los cursos basados en la web, en los sistemas de gestión para el aprendizaje de contenidos, en los sistemas inteligentes de aprendizaje, entre otros usos.

Agregando a lo anterior, la red Internet y las plataformas didácticas permiten que las instituciones educativas tengan suficiente información de sus estudiantes en repositorios electrónicos de datos (Kovacic, 2010). Generalmente, las instituciones pueden disponer fácilmente de esta información, pero no la procesan a la par que se genera u obtiene, y es seguro que se puede encontrar, en esos datos, un conocimiento sustancialmente útil.

Ahora bien, en México, las instituciones educativas de nivel superior presentan índices de ET del 39%, considerando la titulación del alumno, según la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES, 2021). Los índices de ET son indicadores muy importantes para una

institución educativa ya que reflejan el trabajo de su personal docente y directivo, aun cuando el rendimiento de los alumnos se vea afectado por algunos factores externos.

Relacionado con ello, existen políticas y programas de gobierno (Cervantes, Warschauer, Nardi, & Sambasivan, 2011), creados con la finalidad de disminuir los índices de reprobación y deserción, no sólo por la preocupación de la imagen de las instituciones educativas, sino por la conveniencia de que los estudiantes culminen una carrera profesional y que exista un mayor número de personas preparadas. Desde luego, las instituciones educativas de nivel superior están comprometidas y buscan opciones para lograr disminuir estos índices.

Concerniente a este argumento, en el Instituto Tecnológico de Los Mochis (ITLM), no se cuenta con ningún padrón de deserción de alumnos; únicamente se registran índices semestrales de conformidad con el aprendizaje, mismos que están compuestos por el índice de reprobación y el de aprobación. La institución tiene registros de índices de ET, por cohorte generacional, que indican de manera porcentual, cuántos alumnos terminan su programa de estudios y se titulan (por ejemplo, 51% sin titulación, 28% con titulación, según datos del cohorte 2011); estos índices son actualizados al finalizar cada semestre. Indiscutiblemente, el fenómeno de la deserción escolar en el ITLM afecta a la institución porque disminuye el número de alumnos egresados (Antunes, 2011); en este escenario, no existen indicadores que señalen de manera temprana a los alumnos que presenten un riesgo potencial de fracaso por reprobación o deserción.

En definitiva, resulta de gran importancia que las instituciones educativas tengan la capacidad de analizar toda la información acumulada de sus alumnos y obtengan el conocimiento que les facilite predecir e identificar a los estudiantes que estén en riesgo de desertar. Esto les permitirá tomar acciones y evitar, hasta donde sea posible, que los alumnos pasen a formar parte de la estadística de los índices de deserción escolar.

En este documento se presenta un estudio cuyo objetivo fue definir un modelo para predecir, de manera temprana, los casos de estudiantes que presenten un potencial riesgo de deserción escolar, a través de técnicas de minería de datos, utilizando factores académicos, psicológicos, sociales, económicos y de interacción. La aplicación de este modelo brindaría al personal directivo y docente de la institución

educativa, la posibilidad de tomar las medidas preventivas adecuadas para apoyar oportunamente a los alumnos en riesgo.

REVISIÓN DE LITERATURA

Z. Kovacic (2010).

Realizó un estudio sobre predicción temprana del éxito de los alumnos, en el Open Polytechnic, con estudiantes de un curso, de 2006 a 2009, analizó datos de más de 450 estudiantes que se inscribieron en el curso Information Systems y los utilizó para realizar un análisis cuantitativo. Con base en las técnicas de minería de datos (como selección de características y árboles de clasificación), identificó los factores más importantes para el éxito de los estudiantes y un perfil de los típicos estudiantes exitosos y no exitosos.

Para la identificación de los factores utilizó nomás un algoritmo de selección de atributos llamado chi-square para conocer el ranking de los atributos, tenía solo nueve atributos.

Para la predicción utilizó árboles de clasificación CHAID, QUEST, y CART. Los resultados mostraron que el árbol de clasificación y regresión (CART) fue el más exitoso en el crecimiento del árbol con un porcentaje general de clasificación correcta del 60.5%; y tanto el riesgo estimado por la validación cruzada como el diagrama de ganancia sugiere que todos los árboles, basados solo en los datos de inscripción, no son muy buenos para separar a los estudiantes exitosos de los que no lo lograron.

Márquez-Vera, Romero y Ventura (2013).

Realizaron un estudio de predicción de deserción escolar, en una escuela de bachillerato en Zacatecas, México, con técnicas de minería de datos, empleando los 10 algoritmos de selección de atributos siguientes: CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, ReliefFAttributeEval y SymmetricalUncertAttributeEval. Estos autores, tenían 77 atributos y seleccionaron los mejores 15; posteriormente, balancearon los datos con un filtro de datos supervisado, Synthetic Minority Over- sampling Technique, y obtuvieron 10 archivos de entrenamiento balanceados y 10 archivos de prueba no balanceados.

Para obtener los modelos predictivos Márquez-Vera, Romero y Ventura, aplicaron también 10 algoritmos de clasificación, como son: JRip, NNge, OneR, Prism, Ridor, ADTree, J48, RandomTree,

REPTree, SimpleCart; de los cuales, mostraron mejores resultados JRip, Prism, ADTree, Simplecart. Como trabajo futuro, los autores proponen desarrollar su propio algoritmo basado en programación genética, y así predecir de la manera más temprana posible a los estudiantes en riesgo de deserción.

METODOLOGÍA

En este estudio, se realizaron pruebas con una base de datos relacional MySQL, con un total de 1,374 registros de estudiantes del Instituto Tecnológico de Los Mochis (ITLM), de la carrera de Ingeniería Industrial. El método para predecir a los estudiantes que pudieran presentar un potencial riesgo de reprobar o abandonar sus estudios profesionales, se fundamentó en la metodología CRISP-DM (Wirth, 2000) y abarcó las siguientes seis etapas:

1) Comprensión del problema.

Se buscó el alcance de los objetivos y los requisitos del proyecto desde una perspectiva institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. En este caso se determinaron los diversos factores que influyen en la deserción escolar, tanto académicos como psicológicos, sociales, económicos, y de interacción. También, se identificaron las posibles fuentes de datos que se utilizarían para el proyecto.

2) Comprensión de los datos.

Esta etapa consistió en recolectar toda la información disponible de los estudiantes. Para ello, se debieron detectar todos los factores que podrían afectar en el rendimiento académico de los estudiantes y recopilar la información de las fuentes disponibles. Finalmente, toda la información fue integrada en un solo almacén de datos.

3) Preparación de los datos.

En esta etapa, el conjunto de datos se preparó para la posterior aplicación de las distintas técnicas de DM. Se realizaron tareas típicas de pre-procesado como la limpieza de datos y la transformación de variables. Otras técnicas, más específicas, como la selección de atributos y el balanceo de datos pudieron ser aplicadas para resolver los problemas de alta dimensionalidad y desbalanceo, que suelen presentarse en este tipo de conjunto de datos.

4) Modelado.

En esta etapa, ya con el conjunto de datos pre-procesado, se aplicaron diversas técnicas de DM para predecir el fracaso escolar de los estudiantes. Para este problema se utilizaron las diferentes técnicas de clasificación que existen. Se propuso un algoritmo de clasificación basado en programación genética y se compararon sus resultados con otros algoritmos de clasificación clásicos, basados en reglas de clasificación y en árboles de decisión. Además, se aplicó la clasificación sensible a costos, la cual considera diferentes costos por error de cada clase, con el fin de resolver el problema del desbalanceo del conjunto de datos.

5) Evaluación.

En esta última etapa, fueron revisados los modelos de clasificación descubiertos en la etapa anterior. Se analizaron las salidas de los distintos algoritmos y se verificaron cuáles son los factores que aparecían en las reglas y en los árboles de decisión, y cómo se relacionaban. A partir de estos análisis se pudo realizar una interpretación del problema y su magnitud, para la futura toma de decisiones que pudiese reducir el problema del fracaso y el abandono.

6) Implementación.

Una vez que el modelo fue construido, está en proceso de ser evaluado por las autoridades educativas de la institución, en un programa piloto, a fin de llevar a cabo otra etapa de validación.

RESULTADOS

En síntesis, en esta investigación, el método utilizado para predecir a los estudiantes que presentan un potencial riesgo de reprobar o abandonar sus estudios profesionales está basado en la metodología CRISP-DM (Wirth, 2000).

Para el estudio, se utilizó una base de datos relacional con la información académica de 1,374 estudiantes, de cohorte generacional 2014, 2015, 2016, 2017 y 2018, de la carrera de Ingeniería Industrial en el ITLM, como muestra de población; asimismo, archivos de Excel con los resultados de los exámenes de ingreso EXANI II, y también archivos DBF con los datos socioeconómicos de los alumnos capturados a través de la encuesta de ingreso, aplicada por CENEVAL.

Con los tratados de los modelos conceptuales ya expuestos, de tipo social, psicológico, económico y de interacción, se comprendieron cuáles son los factores que influyen en los alumnos para tomar la decisión

de continuar sus estudios o desertar. Esto fue de utilidad al momento de analizar las fuentes de datos de la institución educativa donde se realizó la investigación, ya que sirvió de base para saber identificar cuáles datos se debían tomar como factores.

Por cierto, en estas fuentes de datos se detectaron 42 factores, como variables independientes, que afectan la decisión del alumno de perseverancia en sus estudios. Entonces, se integraron las diferentes fuentes en una sola base de datos en MySQL para facilitar su análisis y la aplicación de los algoritmos para el modelado.

Dicho sea de paso, una de las tareas fundamentales cuando se preparan los datos para un adecuado análisis es la limpieza, la cual consiste en eliminar registros incompletos o con datos corrompidos; parece algo sencillo, pero si no se realiza de manera minuciosa puede llegar a alterar los resultados de los experimentos y, por lo tanto, arrojar resultados sesgados o falsos.

Seguidamente, para reducir la alta dimensionalidad de factores, se utilizaron algoritmos de selección de atributos y se eligieron los 10 atributos con mayor peso. Esta elección se realizó debido a que es muy complejo generar modelos predictivos con 42 factores.

A continuación, se utilizó el software libre llamado Orange Canvas y se aplicaron seis algoritmos que fueron útiles para la selección de los atributos más relevantes. Estos algoritmos son ANOVA, ChiSquared, GiniDecrease, GainRatio, InformationGain, ReliefF. La Figura 2 muestra, en la parte superior, los algoritmos aplicados y en el lado izquierdo, los 10 atributos o factores.

Figura 2

Los 10 factores más relevantes

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF
REG_PROC		0.502	0.519	0.208	167.797	112.138	0.764
ANO_NAC		0.263	0.139	0.072	2.111	12.536	0.005
SER_AUTO		0.125	0.074	0.038	3.258	5.200	0.079
PROM_BAC		0.124	0.062	0.029	1.970	3.301	0.059
POS_SUS		0.089	0.044	0.016	0.250	3.292	-0.006
EST_ALCA		0.072	0.075	0.034	6.174	2.089	0.161
EXA_EXT		0.070	0.049	0.028	2.458	7.755	0.042
MES_NAC		0.065	0.033	0.017	1.092	2.230	0.118
MOD_BAC		0.061	0.044	0.014	1.872	8.273	-0.001
PCNE		0.059	0.029	0.013	2.655	1.660	0.038

Ahora bien, debido a que se encontró una diferencia significativa entre el número de instancias que están activas y las que desertaron, como se puede observar en la Figura 3, se empleó software libre

llamado WEKA para aplicar el algoritmo SMOTE y así equilibrar el número de instancias de estas dos clases. Asimismo, en la Figura 4 se muestran los datos después de ser balanceados con SMOTE.

Figura 3

Datos antes de balancear

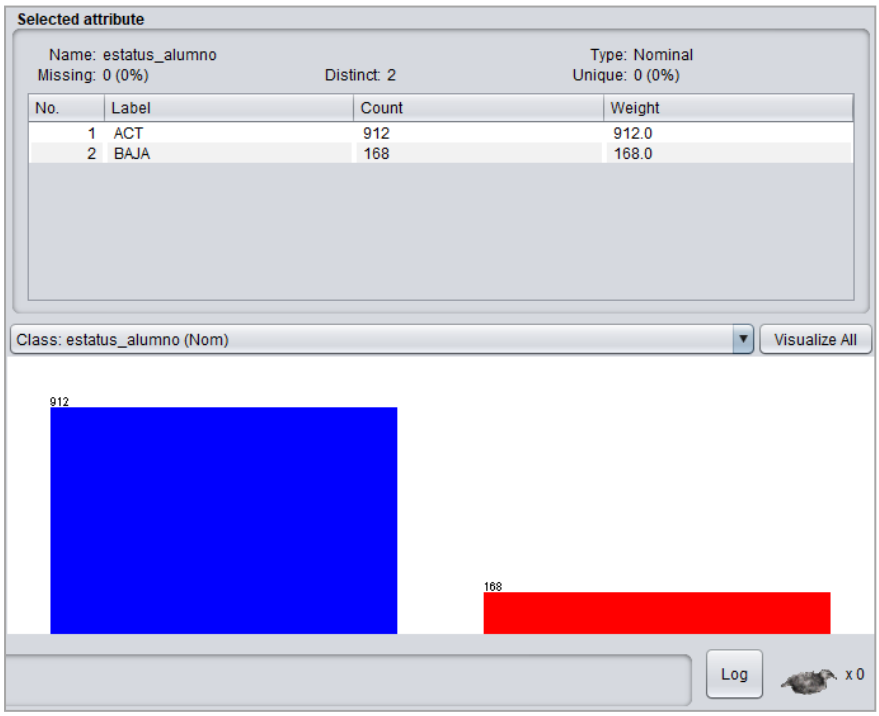
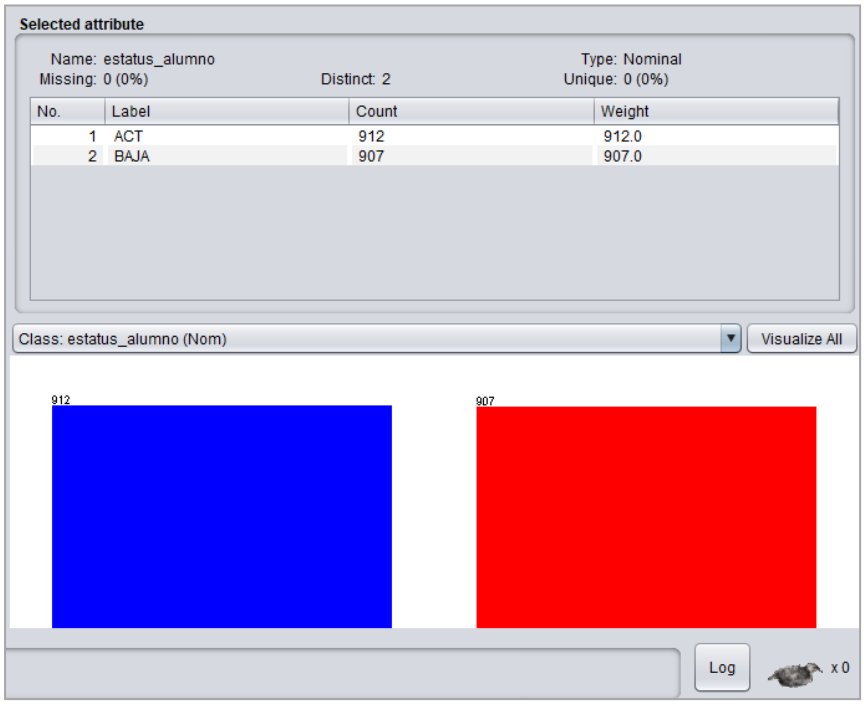


Figura 4

Datos balanceados con SMOTE



Así pues, una vez obtenido el conjunto de datos pre-procesado, se aplicaron diversas técnicas de Minería de Datos para predecir el fracaso escolar de los estudiantes. Para este problema se emplearon diferentes técnicas de clasificación y para realizar los experimentos se utilizaron algoritmos como DecisionTable, JRip, OneR, PART, DecisionStump, HoeffdingTree, J48, LMT, RandomForest, RandomTree, REPTree.

En la Tabla 1, se puede observar la comparación de los resultados de los algoritmos de clasificación basados en reglas de clasificación y en árboles de decisión.

Tabla 1
Resultados de la clasificación utilizando solo los mejores atributos sin balancear los datos

Algoritmo	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
DecisionTable	0.849	0.818	0.788	0.849	0.793	0.089	0.533	0.766
JRip	0.836	0.855	0.724	0.836	0.776	-0.052	0.491	0.746
OneR	0.840	0.820	0.765	0.840	0.788	0.046	0.510	0.750
PART	0.852	0.766	0.809	0.852	0.808	0.177	0.526	0.770
DecisionStump	0.852	0.852	0.726	0.852	0.784	0.000	0.446	0.736
HoeffdingTree	0.852	0.852	0.726	0.852	0.784	0.000	0.446	0.748
J48	0.852	0.852	0.726	0.852	0.784	0.000	0.446	0.748
LMT	0.852	0.852	0.726	0.852	0.784	0.000	0.500	0.748
RandomForest	0.830	0.804	0.763	0.830	0.787	0.048	0.581	0.783
RandomTree	0.756	0.782	0.740	0.756	0.748	-0.028	0.484	0.744
REPTree	0.852	0.852	0.726	0.852	0.784	0.000	0.500	0.748

NOTA. Elaborado por los autores.

Precisión: La precisión se define como la fracción de elementos clasificados correctamente como positivos de todos los elementos que el algoritmo clasifica como positivos, se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos positivos (Precision = TP / (TP + FP).

Recall: El Recall se define como la fracción de elementos clasificados correctamente como positivos de todos los elementos positivos, se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos (Precisión = TP / (TP + FN).

F-Measure: Es la media armónica de la Precision y el Recall (F-Measure = 2 * Precision * Recall /(Precision + Recall)). Este indicador es el que se tomó como referencia para evaluar el mejor modelo. En definitiva, el algoritmo que proporcionó la mejor precisión fue RandomForest, pero desafortunadamente este algoritmo no explica cómo va generando el modelo, solo arroja los resultados, por lo cual no es factible utilizarlo para crear un modelo predictivo.

Ahora bien, en la Tabla 2, se puede observar que los algoritmos de árboles de clasificación LMT y RandomTree obtuvieron una precisión muy similar; de los algoritmos de reglas de clasificación, PART aparece con la mejor precisión. A diferencia de la Tabla1, en esta segunda tabla se utilizaron los datos balanceados, es decir se tuvieron la misma cantidad de registros de alumnos que desertaron y de alumnos que continuaron sus estudios, lo cual aumento la precisión.

Tabla 2
Resultados de la clasificación utilizando solo los mejores atributos con los datos balanceados

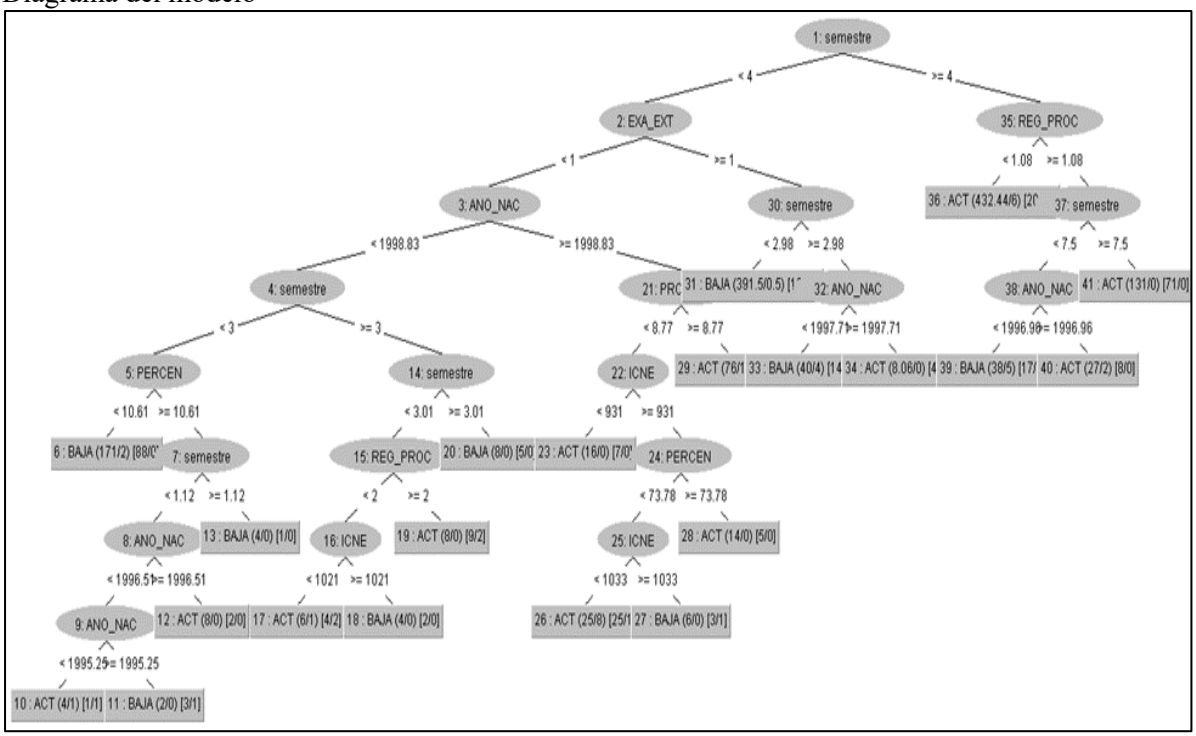
Algoritmo	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
DecisionTable	0.701	0.296	0.712	0.701	0.698	0.414	0.755	0.735
JRip	0.709	0.292	0.711	0.709	0.708	0.419	0.742	0.700
OneR	0.608	0.387	0.636	0.608	0.589	0.244	0.610	0.570
PART	0.828	0.171	0.830	0.828	**0.828**	0.658	0.858	0.825
DecisionStump	0.603	0.393	0.623	0.603	0.588	0.226	0.605	0.565
HoeffdingTree	0.637	0.362	0.638	0.637	0.637	0.275	0.724	0.726
J48	0.808	0.191	0.815	0.808	0.807	0.623	0.855	0.824
LMT	0.839	0.160	0.845	0.839	**0.838**	0.684	0.857	0.825
RandomForest	0.866	0.133	0.868	0.866	**0.866**	0.735	0.933	0.922
RandomTree	0.839	0.160	0.844	0.839	**0.838**	0.683	0.870	0.828
REPTree	0.786	0.214	0.787	0.786	0.786	0.572	0.834	0.811

NOTA. Elaborado por los autores.

Para terminar, a manera de representación visual, en la Figura 5 se presentan los factores que afectan el rendimiento escolar de los alumnos y pueden propiciar el abandono escolar. En primer lugar se encuentra el semestre que cursan los alumnos; en segundo lugar interviene si los estudiantes presentaron exámenes extraordinarios o no y su región de procedencia; en tercer lugar están la edad y nuevamente el semestre; en cuarto lugar, el semestre, la edad y el promedio de bachillerato. Además, el quinto lugar lo ocupan el porcentaje obtenido en su examen de admisión, el semestre y su índice CENEVAL; en

sexto lugar, concurren el semestre, la región de procedencia y el porcentaje obtenido en su examen de admisión; en el séptimo lugar quedan la edad del alumno y su índice CENEVAL; por último, en el octavo lugar, se encuentra la edad.

Figura 5
Diagrama del modelo



CONCLUSIONES

Para concluir, se puede afirmar que es de suma importancia sumergirse en la literatura de los modelos conceptuales, sin importar cuán antiguos sean, debido a que, conocer los puntos de vista de diferentes autores, permite al investigador comprender las múltiples causas que se conjugan y conducen al estudiante por diferentes caminos en su trayecto de formación profesional. Sin lugar a duda, la revisión de la literatura fue de gran utilidad al momento de identificar los factores a analizar, según la información de los alumnos con la que contaba la institución educativa, habiendo seleccionado factores académicos, psicológicos, sociales, económicos y de interacción.

Un punto importante a mencionar es que, año tras año, CENEVAL ha modificado el instrumento para recabar los aspectos de carácter general de los estudiantes, por lo que, al integrar la información de las cohortes generacionales en estudio, se encontraron factores diferentes. Esta situación dificultó concentrar en el archivo histórico los factores seleccionados, lo cual era de suma importancia ya que

los algoritmos utilizados para generar los modelos predictivos se deben entrenar con datos históricos. Por esta razón, se sugiere a la institución educativa generar un instrumento propio de registro de información de aspirantes, para tener control sobre los datos que se consideren objeto de estudio y generar, simultáneamente, un archivo histórico.

Vale la pena resaltar que en el estudio se comprobó que, al utilizar los datos balanceados, los modelos obtienen una mayor precisión que los modelos sin balancear. Esto se puede observar en los resultados que se exhiben en la Tabla 1 y en la Tabla 2.

Como resultado del estudio, después de haber comparado los resultados de los diferentes algoritmos de clasificación, se propone implementar el modelo predictivo que se generó con el algoritmo RandomTree, por haber obtenido la puntuación más cercana al uno, lo cual demuestra que es el algoritmo más preciso y su modelo es fácil de interpretar.

En resumidas cuentas, la predicción del abandono de estudios es esencial, con miras a que los centros de educación tomen medidas para controlar esta problemática; en virtud de lo cual, los resultados de este trabajo, constituyen una herramienta para que en el Instituto Tecnológico de Los Mochis se puedan tomar las acciones preventivas convenientes, a fin de retener a los estudiantes e incrementar los índices institucionales de eficiencia terminal.

REFERENCIAS BIBLIOGRAFICAS

Antunes, C. (2011). Anticipating student's failure as soon as possible. In Handbook of Educational Data Mining (pp. 353–363). https://doi.org/10.1201/b10274-28

ANUIES. (2021). ANUARIO_EDUCACION_SUPERIOR-LICENCIATURA_2021-2022. www.anuies.mx/

Cabrera, A. E., Nora, A., & Asker, E. H. (1999). Economic Influences on Persistence Reconsidered.

Cervantes, R., Warschauer, M., Nardi, B., & Sambasivan, N. (2011). Infrastructures for low-cost laptop use in Mexican schools. Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11, 945. https://doi.org/10.1145/1978942.1979082

Ethington, C. A. (1990). A Psychological Model of Student Persistence, 31(3), 279–293.

Han, J. (2007). Data Mining : Concepts and Techniques.

Himmel K., E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior -

Retención y movilidad estudiantil. Revista Calidad En La Educación, 91–108. http://www.alfaguia.org/alfaguia/files/1318955602Modelo de analisis de la desercion estudiantil en la educacion superior.pdf

INEGI. (2021). Anuario estadístico y geográfico de los Estados Unidos Mexicanos 2021, 45–81. www.inegi.org.mx/

Kovacic, Z. (2010). Early Prediction of Student Success : Mining Students Enrolment Data.

Marquez-Vera, C., Romero Morales, C., & Ventura Soto, S. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje, 8(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

Romero, C., & Ventura, S. (2010). Educational Data Mining : A Review of the State of the Art, 40(6), 601–618.

Spady, W. G. (1970). Dropouts from Higher Education : An Interdisciplinary Review and Synthesis 1.

Tinto, V. (1975). Review of Educational. https://doi.org/10.3102/00346543045001089

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, (24959), 29–39. https://doi.org/10.1.1.198.5133

Zhang, Y., Oussen, S., Clark, T., & Kim, H. (2010). Middlesex University Research Repository.