*EDU REVIEW* | Vol. 11, No. 2, 2023 | ISSN 2695-9917
*The International Education and Learning Review / Revista Internacional de Educación y Aprendizaje*
https://doi.org/10.37467/revedu.v11.4784


# APLICACIÓN MÓVIL PARA ESTIMAR EL PORCENTAJE DE DESERCIÓN ESCOLAR USANDO MINERÍA DE DATOS

Mobile Application to Estimate the Percentage of School Dropouts Using Data Mining

ELENA FABIOLA RUIZ LEDESMA, LORENA CHAVARRÍA BÁEZ, JUAN CARLOS VELIZ MARTÍNEZ

Instituto Politécnico Nacional, Mexico

**KEYWORDS**

*CRISP-DM*
*Terminal efficiency*
*Data Mining*
*School dropout*
*Application*
*Students*
*University*

**ABSTRACT**

*The paper reports the development of a mobile application, using data mining techniques, to determine a model of decision rules and include them in the application using JSON. The methodology used for the development was CRISP-DM. For the data mining process, information from 949 students from a public institution in Mexico was used, considering 7 variables. The application generated allows estimating the probability that a student has of not completing their studies, in order to act accordingly, looking for strategies to support the improvement of the students' conditions, according to the detected factors.*

**PALABRAS CLAVE**

*CRISP-DM*
*Eficiencia terminal*
*Procesamiento de datos*
*Deserción escolar*
*Aplicación*
*Estudiantes*
*Universidad*

**RESUMEN**

*Este artículo reporta el desarrollo de una aplicación móvil, haciendo uso de técnicas de minería de datos, para determinar un modelo de reglas de decisión e incluirlas en la aplicación usando JSON. La metodología utilizada para el desarrollo fue CRISP-DM. Para el proceso de minería de datos se utilizó la información de 949 estudiantes de una institución pública de México, considerando 7 variables. La aplicación generada permite estimar la probabilidad que tiene un estudiante de no culminar sus estudios, para poder actuar en consecuencia, buscando estrategias que apoyen la mejora de las condiciones de los estudiantes, según los factores detectados.*

Recibido: 27/ 01 / 2023
Aceptado: 19/ 04 / 2023

## 1. Introducción

La deserción estudiantil en los distintos niveles educativos es una problemática que se ha dado desde tiempo atrás, pero actualmente se ha acentuado en el nivel superior, esto es, ha habido un incremento en la cantidad de estudiantes que no terminan sus estudios de nivel universitario alrededor del mundo, así como también en México. Dentro de los indicadores que son reportados por las instituciones tanto públicas como privadas (Ayala, López, & Menéndez, 2020; Toscano de la Torre, 2016), se encuentra la elevada cantidad de alumnos que reprueban sus cursos, la necesidad de tener que trabajar para apoyar a sus familias, entre otros factores, lo que los conduce a abandonar la escuela y por ende no culminar sus estudios. La finalidad de este artículo es dar a conocer el proceso llevado a cabo para realizar una aplicación móvil que estima el porcentaje que un alumno tiene de abandonar sus estudios, para que las autoridades de las instituciones empleen estrategias que permitan apoyarlos para que continúen con sus estudios. Para construir el modelo que fue programado en la aplicación móvil, fue necesaria la detección de patrones dentro de una base de datos, para ello, se empleó la metodología de desarrollo CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es utilizada para la creación de aplicaciones que involucran minería de datos. Inicialmente se realizó una revisión de investigaciones que se han llevado a cabo, considerando los estudios que convergen en factores que propician el abandono escolar a nivel universitario (Alyahyan & Düştegör, 2020; Torre et al., 2015; Estrada-Danell et al., 2016) y aquellos que los involucran utilizando técnicas de minería de datos (Oviedo & Jiménez, 2019; Marcano & Rodríguez, 2014); posteriormente, se mencionan los factores que se consideraron, tomando en cuenta aquéllos que no han sido poco empleados en otros estudios y finalmente se plantea la forma en que fue creada la aplicación móvil.

De acuerdo con información destacada por Toscano de la Torre (2016), en las instituciones de nivel superior en México, para el año 2013 hubo un ingreso a primer semestre de 568,669 alumnos y para el cierre generacional en el 2018 existieron 424,018 egresados, obteniéndose una eficiencia terminal (ET) del 74.56%. De la misma forma, la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), que se encarga de promover la mejora continua de los servicios que facilitan las instituciones de educación superior (IES) en México, menciona que de cada 100 estudiantes que inician sus estudios en una IES, solo el 60% de ellos egresan.

A su vez, de los 34 países considerados desarrollados que pertenecen a la Organización para la Cooperación y el Desarrollo Económicos (OCDE), México es uno de los países que se ubica en los últimos lugares en lo que se refiere a cobertura en educación, además de presentar una alta tasa de deserción escolar ya que los estudiantes abandonan de forma prematura la escuela (Torre, et al, 2015). De forma global, el conocimiento es uno de los pilares que fomentan el desarrollo social y hace crecer la economía, gracias a esto muchos países con niveles educativos mayores en toda su población son los que cuentan con la capacidad de enfrentar nuevas problemáticas de índole global y logran mantener una mejor equidad y participación social (Márquez, 2012; Fadhl & Sofian, 2015).

Con base a lo señalado en la Declaración de los Derechos Humanos (Naciones Unidas, 2015), la educación es una prioridad a nivel mundial, de tal forma que toda persona tiene derecho a la educación con el objetivo de tener un desarrollo pleno en la personalidad humana, así mismo se favorece la comprensión y la tolerancia. Por lo consiguiente es indispensable que una persona pueda completar sus estudios, sin importar las múltiples adversidades a las cuales se pudiera enfrentar. También, la Comisión Nacional de Derechos Humanos (CNDH), expresa que toda persona tiene derecho a recibir educación y asume que los niveles básicos deben de ser gratuitos con la finalidad de fomentar el desarrollo de mejores personas. Además, en el Plan Nacional de Desarrollo (2019-2024) se enfatiza como primordial el contar con educación de calidad, teniendo como una de sus estrategias el disminuir el abandono escolar, aumentar la eficiencia terminal (ET) en distintos niveles educativos y a la vez incrementar el ingreso de un nivel a otro. Sin embargo, en algunas Instituciones de Educación Superior (IES), de acuerdo con su informe de rendición de cuentas (Instituto Tecnológico de Iztapalapa, 2018), en el apartado de eficiencia terminal se ilustra que la ET del 2017 fue de solo el 25%.

Teniendo en consideración que la educación es uno de los ejes que son contemplados como fundamentales a nivel mundial y que a su vez cada gobierno está comprometido con realizar los cambios que sean necesarios, permite justificar la realización de la presente investigación, que culminó con el desarrollo de una herramienta informática en aras de colaborar a combatir los bajos índices de ET.

El objetivo que se tiene de forma general es el de desarrollar un sistema que se encuentre apoyado en la ciencia de datos, que sea capaz de determinar el porcentaje que un alumno tiene de culminar su carrera, lo anterior basado en las experiencias aprendidas a partir de otros alumnos. Para poder lograr esto es necesario identificar qué factores, tanto socioeconómicos como académicos, influyen en la ET. Seguido de lo mencionado, es necesario desarrollar un modelo de predicción, basado en algoritmos de minería de datos, que identifique los patrones de comportamiento de una población y que proporcione reglas de decisión, basadas en los patrones detectados, posteriormente se describe el desarrollo de una aplicación móvil que contiene las reglas obtenidas y que permite a sus usuarios conocer la probabilidad que tienen de terminar o desertar de sus estudios.

## 2. Estado del Arte

Según un estudio realizado en el sector educativo (Fayyad et al., 1996), es indispensable disponer de sistemas de gestión que permita tomar decisiones académicas y a partir de este conocimiento oportuno ser capaces de elaborar estrategias en beneficio de los estudiantes. Según los reportes presentados por Peinado & Jaramillo (2018), la eficiencia terminal del Centro de Investigación e Innovación Tecnológica en el periodo de egreso del 2018 fue del 22.17%, lo cual indica la baja cantidad de alumnos que terminan sus estudios universitarios. En una investigación reportada en (Pérez et al., 2016), se menciona que en muchas ocasiones la mala planeación que se tiene de los horarios de clases se convierte en una complicación para el avance de los estudiantes, esto porque no cumple con sus necesidades académicas. Así mismo, señala que cuando un alumno reprueba una materia no conoce los programas de apoyo que existen de reforzamiento para dichas asignaturas.

En un estudio realizado por la Universidad Autónoma Metropolitana (UAM) de Azcapotzalco (Villalobos, 2017), se encontró que entre las causas con mayor índice de afectación hacia los estudiantes y que desembocan en que estos no terminen sus estudios son: las responsabilidades laborales (estudiar y trabajar) y la falta de espacio en las materias que debían cursar. En una investigación realizada en la Universidad de Zulia en Venezuela (Marcano & Rodríguez, 2014), se determinó como principal factor de la baja ET los conocimientos previos al nivel en que los estudiantes se encuentran, así como las pocas horas dedicadas a sus estudios fuera de la escuela.

La minería de datos (MD) en la educación no es un concepto nuevo (Eckert, & Suénaga, 2015), la utilización de técnicas de MD puede permitir determinar la probabilidad que un estudiante tiene de desertar o continuar en la escuela, así como también ir prediciendo su desempeño durante el tiempo que dure sus estudios. De acuerdo con Román, Sánchez & García (2017), la Minería de Datos Educativa es considerada como un área multidisciplinaria que permite construir un modelo ajustado a un contexto educativo, con el fin de predecir comportamientos futuros.

La MD orientada a la educación permite predecir factores, características, fenómenos o situaciones tales como la probabilidad de abandono de estudios de un alumno (Oviedo & Jiménez, 2019). También la MD puede ayudar a tomar mejores decisiones a las personas relacionadas con la gestión eficiente de los recursos, retención escolar, eficiencia terminal, así como también métodos de evaluación, aspectos que son importantes para que una institución educativa tenga éxito (Estrada-Danell, 2016).

## 3. Aspectos Teóricos

### *3.1. Minería de Datos*

La minería de datos se puede considerar como un proceso innovador que permite de forma fácil llevar a cabo métodos y técnicas cuantitativas para analizar una gran cantidad de datos de forma automática o semiautomática, con la finalidad de obtener patrones que no son fácilmente detectables. La MD utiliza algoritmos computacionales que permiten extraer nuevos conocimientos que no se encuentran a simple vista, este conocimiento puede dar como resultado anomalías que no se esperaban y la oportunidad de tomar mejores decisiones sobre nuevas situaciones. La MD se puede definir como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al analizar grandes cantidades de datos; estos procesos son actualmente utilizados en múltiples áreas, por ejemplo análisis de mercado, análisis financieros, educación, entre muchos más (Reyes et al., 2017; Dicovskiy & Pedroza, 2018).

La MD es un campo en el que se mezclan varias disciplinas como la estadística y las ciencias de la computación, tratando de detectar patrones o registros poco usuales en repositorios de datos enormes, además también se apoya de la inteligencia artificial, los sistemas de bases de datos y el aprendizaje automático (Román, Sánchez & García, 2017). La MD en educación se utiliza para explorar datos que involucran a los estudiantes, es decir, que provienen de un entorno educativo, y su uso se da con la intención de comprender mejor a los estudiantes, así como el ecosistema en el que se encuentran aprendiendo (Panizzi, 2019).

En el desarrollo de esta investigación se hace uso de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual está orientada para proyectos de minería de datos. La MD se utiliza para obtener patrones con los cuales se pueda predecir el porcentaje estimado que tiene un estudiante de abandonar sus estudios, esto lográndose gracias a la utilización del algoritmo C4.5 de aprendizaje supervisado, el cual está basado en aprender de los comportamientos y características previamente observados y poder así inferir la estimación mencionada. Se hace uso del software Weka (2019) para la ejecución del algoritmo C4.5, utilizando la implementación denominada J48 que provee el software.

### *3.2. Factores de deserción escolar*

Se han detectado distintos factores que son atribuidos a la deserción de alumnos de sus estudios, en el aspecto socioeconómico se han determinado que las malas condiciones de vida, los bajos ingresos familiares, una estructura familiar disfuncional, así como el bajo acceso a medios de esparcimiento son algunos de los principales alicientes para que un alumno abandone la escuela. Sumado a lo anterior, el hecho de que las instituciones no conozcan la situación socioeconómica de sus alumnos favorece la deserción, esto al no poder brindarle una atención personalizada derivada del entorno en el que se encuentra (Marcano & Rodríguez, 2014; Eckert & Suénaga, 2015; Ayala, López, & Menéndez, 2021).

El factor académico también suele ser recurrente, este a su vez tiene distintas vertientes, en algunas ocasiones el docente puede estar bastante preparado y con una trayectoria amplia, sin embargo, con nulos conocimientos de docencia, lo cual complica la transferencia de conocimientos hacia sus alumnos. Otro de los aspectos que son relevantes, es el hecho de la baja preparación que un alumno tiene del nivel anterior, ocasionando un bajo desempeño, así mismo la forma en que algunas materias son evaluadas siendo los exámenes el único método utilizado y dejando de lado otros instrumentos que pudieran ser utilizados para apreciar desde otros puntos el aprovechamiento del alumno (Marcano & Rodríguez, 2014; Ayala, López, & Menéndez, 2021).

Los factores sociales juegan también un papel importante en la deserción de los estudiantes, (Hernández et al., 2015). El hecho de compartir un espacio continuamente puede resultar en problemas de adaptación, quedando expuesto a problemas con sus compañeros y profesores, el prepararse para una exposición frente al grupo puede generar tensión, miedo, ansiedad, generando con esto desinterés que puede convertirse en un factor por el cual abandonan sus estudios (Romero, Ultrilla & Utrilla, 2014). Según Álvarez, Gómez & Morfín (2012), en algunos estudios se ha encontrado que la eficiencia terminal no tiene un incremento considerable cuando el estudiante percibe un apoyo económico o algún tipo de beca; sin embargo, el factor económico invariablemente afecta a las personas de escasos recursos y que les es difícil acceder a la educación.

## 4. Métodos y materiales

Como se mencionó en el marco teórico conceptual, la metodología de desarrollo utilizada fue CRISP-DM (Schröer, Kruse y Gómez, 2021), esta metodología se divide en 6 fases: compresión del negocio, análisis de los datos, preparación de los datos, modelamiento, evaluación y despliegue. A continuación, se describe de forma general cómo fueron utilizadas cada una de las fases de la metodología mencionada.

### *4.1. Comprensión del negocio*

El primer paso de todo proyecto es entender qué es lo que se quiere lograr, para lo cual primero se comprendió la problemática, así como las formas en que se ha abordado por otros investigadores y los resultados que se han obtenido, es por ello que en el estado del arte se incluyeron algunos de los estudios revisados. Una vez que la problemática fue comprendida en un mayor grado, se modificó el cuestionario

de la Gestión Universitaria Integral del Abandono (Proyecto ALFA GUIA DCI-ALA/2010/94), cuya autoría es de la Comisión GUIA y Grupo de Análisis creados con ayuda de la Unión Europea. Este cuestionario se adaptó al entorno educativo de México. El cuestionario se aplicó a dos bloques de 200 alumnos cada uno, un bloque estuvo conformado por alumnos de primer semestre y otro de último semestre, ambos bloques correspondieron a estudiantes que cursaban su carrera en el Instituto Tecnológico de Iztapalapa de la Ciudad de México, con la finalidad de conocer las diferencias que existen entre alumnos que acaban de ingresar y de aquellos que están terminando sus estudios. Se utilizó este cuestionario ya que se encuentra diseñado para detectar factores individuales, académicos, socioculturales, económicos e institucionales. Una vez aplicado el cuestionario, se obtuvo la siguiente información:

1. El estado civil del 80% de la muestra es soltero, lo cual es de suponer que, al no contar con responsabilidades adicionales, se centran únicamente en sus estudios.
2. Un poco más del 80% viven con sus padres, esto significa que dependen directamente de ellos.
3. Son estudiantes que cuentan con 1 o 2 hermanos, lo que se puede interpretar en que las familias que son pequeñas cuentan con mayores oportunidades para estudiar el nivel superior.
4. Casi el 50% de los alumnos tiene hermanos que cuentan con título universitario.
5. El 70% de los padres de los alumnos únicamente terminó el nivel básico o medio superior, lo cual no implica que un estudiante solo llegue a concluir estos niveles.
6. En contraste con el punto anterior, solo en promedio el 15% de los padres cuentan con estudios universitarios.
7. Un poco más del 90% de los estudiantes proviene del mismo contexto geográfico, lo que indica que no se han estado desplazando en diversos estados.
8. Casi el 70% considera que en su casa se favorecen los hábitos de estudio.
9. El 75% dependen económicamente de sus padres, el 20% de sí mismos y el resto de otros familiares.
10. Solo el 45% considera que durante sus estudios ha contado con suficientes recursos económicos para su sostenimiento.
11. Menos del 25% ha contado con becas o subsidios económicos para sus estudios.
12. Casi el 100% de la muestra de estudiantes, cursó el nivel medio superior en instituciones de educación públicas.
13. El 50% nunca ha interrumpido sus estudios.
14. El 25% durante su trayectoria escolar ha experimentado situaciones como acoso, discriminación, maltrato o indiferencias.
15. Solo el 2% ha requerido acondicionamientos especiales por cuestiones de capacidades diferentes.
16. El 4% considera que las relaciones académicas que ha tenido con los profesores son malas o muy malas.
17. También el 4% considera que las relaciones que ha tenido con sus compañeros han sido malas o muy malas.
18. El 40% de los estudiantes señaló que no han tenido éxito en sus materias de Matemáticas, por lo que han presentado exámenes a título de suficiencia o han recursado la asignatura, lo que los ha retrasado en la culminación de su carrera.

## *4.2. Análisis de los datos*

De los factores revisados en la literatura encontrada, así como con la información detectada en la aplicación del cuestionario, se decidió dirigir el trabajo hacia el camino de encontrar si existe una relación con el rendimiento de las matemáticas en las materias de tronco común y el hecho de contar con beca en la trayectoria escolar con la Eficiencia Terminal. Se optó por estos factores como indicadores que puedan generar patrones que permitan apoyar en la toma de decisiones, dado que son factores que no han sido trabajados con técnicas de minería de datos.

Se cuenta con una base de datos relacional con la información de 949 alumnos de dos carreras, Ingeniería en Sistemas Computacionales (ISC) e Ingeniería en Gestión Empresarial (IGE). Se optó por

estas dos carreras ya que, a pesar de que ambas son ingenierías, tienen un enfoque distinto (ciencias exactas y ciencias administrativas) y, como se desea observar si existe relación con las matemáticas y la ET, se podría llegar a detectar factores que se sean de utilidad. Entre los datos que se tienen son el turno en el que cada estudiante inicia, además de poder determinar si ha contado con distintos turnos, se cuenta con las calificaciones que forman parte del historial académico de los estudiantes. Por otra parte, se conoce la edad de los alumnos al momento de ingresar y así mismo es posible calcular la edad que tiene en cada semestre cursado. Es posible también conocer de dónde proviene el alumno (nivel anterior) y el origen del alumno. Como conocimiento preliminar se tiene que los datos con los que se trabajaron corresponden a 7 generaciones, el 41% es de IGE y el 59% de ISC, el 38% son mujeres y el 62% hombres, 57% de los estudiantes tuvieron un turno mixto, 25% se mantuvo en el turno matutino, 18% se mantuvo en el turno vespertino y el 44% de la muestra concluyó sus estudios.

### *4.3. Preparación de los datos*

En esta etapa se debe realizar la selección de las tablas, registros y atributos que serán utilizados para construir el conjunto final de datos, así como la transformación y limpieza de datos para las herramientas que se encargarán del modelado. De las tablas que se encuentran en el modelo relacional con el que se cuenta, se encuentran por mencionar las más importantes:

1. Alumnos: contiene los datos generales de los alumnos, como periodo en el que ingresaron, fecha de nacimiento, lugar de origen, bachillerato de procedencia, carrera que estudia.
2. Historia_alumnos: contiene todas las materias que los alumnos han cursado, calificaciones que han obtenido, semestre en el que cursaron las materias.
3. Horario_materias: mantiene los datos de los horarios en que se han impartido las materias, así como el profesor que atendió cada horario.
4. Alumnos_generales: se encuentran los datos para conocer si el alumno está becado o no, fecha de ingreso a la carrera, entre otros.

Dado que existen otras tablas en el modelo relacional como info_servicio_social, mensajes, etc., que no son de importancia, no son mencionadas y tampoco su contenido, ya que carecen de valor para el desarrollo del trabajo. De las tablas que contienen datos valiosos, se han tomado los campos que contienen información de las calificaciones de los alumnos, con las cuales se puede obtener el promedio general, el promedio de las materias de tronco común, el promedio de las materias que no son de tronco común, el promedio incluyendo materias que son cursadas en más de una ocasión (recursamientos). Con el horario de las materias, se puede asociar a cada alumno a un turno fijo (matutino o vespertino) o un turno mixto. Con la fecha de ingreso y la fecha de nacimiento se puede obtener la edad de ingreso a la universidad y con la fecha en la que cursó la última materia se puede calcular la edad en la que egresó. Con la información de lugar donde estudió el nivel medio superior se puede saber cuántos provienen del mismo entorno demográfico y que porcentaje no, de esta muestra de estudiantes.

En la universidad de donde se están evaluando los datos, se encontró las calificaciones aprobatorias, las cuales se expresan en porcentajes, de tal forma que el 70% corresponde a la calificación más baja aprobatoria y la más alta 100%, considerando no aprobado a los estudiantes que obtengan un porcentaje entre 0% y 69%. Debido a lo anterior, al ser muchos los valores que se pudieran obtener se decidió discretizar los valores de las calificaciones y promedios en 4 rangos: bajo, que toma valores del 0 al 69, regular del 70 al 79, bueno del 80 al 89 y excelente del 90 al 100%. Dado que los alumnos pueden tomar materias en múltiples horarios, se decidió que si el 80% o más de sus estudios son por la mañana pertenecen al turno matutino, si el 80% o más fue por la tarde pertenecen al turno vespertino, de lo contrario se les asigna un turno mixto; sin embargo, también se considera el turno en el que iniciaron sus estudios. Con la finalidad de que el modelo generado presente información confiable, se descartaron a aquellos alumnos de los cuales no se cuenta con su información completa, por lo que se presentan datos de los alumnos que se encuentran bien clasificados respecto a la conclusión de sus estudios o al abandono de la escuela.

### *4.4. Modelamiento*

Al trabajar en esta etapa se seleccionaron y aplicaron los algoritmos de MD que fueron pertinentes al problema, además de que se calibraron con los valores que se determinaron como óptimos. En la mayoría de las ocasiones, para poder utilizar diversos algoritmos de MD, es necesario regresar a la fase de preparación de los datos, ya que algunos algoritmos requieren de formatos especiales en los datos de entrada.

De los algoritmos que se encuentran a disposición para la fase de modelamiento, están los que generan árboles de decisión, estos algoritmos consisten en forma general, en organizar los datos en elecciones que forman ramas de influencia después de una decisión inicial. El nodo raíz del árbol representa la decisión inicial y utiliza decisiones condicionales o multicondicionales, y en cada elección se desciende en el árbol en ramas divergentes hasta llegar a un final. Se utiliza por lo general para predecir el porcentaje de certeza que se tendrá de una situación con respecto al camino tomado (Hernández, Ramírez & Ferri, 2004). Por ejemplo, saber si una persona es soltera o casada, tomando en cuenta datos como el rango de edad, ocupación, estudios, entre otros.

Tomando en consideración que lo que se pretende en este trabajo es conocer el porcentaje estimado que tiene un alumno para completar o no sus estudios de nivel superior, se definió a los árboles de decisión como el algoritmo que se debe utilizar, considerando que, usándose de la forma adecuada tiene la capacidad de predecir en qué situación se puede encontrar un alumno, tomando como entrada lo que se obtiene en la fase de preparación de los datos. Existen varios algoritmos para la creación de árboles de decisión, tales como Hunt's Algorithm, CART, ID3, C4.5, SLIQ, SPRINT, sin embargo, actualmente el algoritmo más utilizado para crear árboles de decisión por sus diversas implementaciones es el C4.5 el cual es una extensión del algoritmo ID3 y es el utilizado en el desarrollo de este trabajo. Su uso se da en la aplicación WEKA (Waikato Environment for Knowledge Analysis), la cual contiene una incrustación de dicho algoritmo en el lenguaje de programación Java y que se conoce como J48 a la implementación que además es *open source*.

El algoritmo C4.5 contiene los pasos necesarios para construir un árbol de decisión a partir de un grupo de datos que sirven de entrenamiento, esto se realiza tomando en cuenta el concepto de entropía de información. La entropía, también conocida como *entropía de Shannon*, es la encargada de medir la incertidumbre de una fuente de información. Cuando el algoritmo C4.5 se está procesando, elige un atributo de los datos, tomando en consideración a aquel que sea más eficaz en cuanto al conjunto de muestras, tomándolo como el subconjunto de datos que es mayormente enriquecido por la clase clasificadora. Para poder llevar a cabo esta elección, el algoritmo toma en cuenta el criterio normalizado de ganancia de información, el cual hace referencia a la diferencia de entropía, que es el resultado de la elección como atributo para dividir los datos. En la configuración del programa Weka, se convirtieron todos los datos a nominales y a enteros en el caso de las edades, se verificó que no existieran valores perdidos, en el caso de las edades se encontraban en el rango de los 17 hasta los 53 años, en el caso de los municipios de los que asisten a la institución se detectaron 23 distintos, los cuales pertenecen a la Ciudad de México y al Estado de México. Se cuenta con 949 instancias correctamente clasificadas, más del 50% de los alumnos se asignaron al turno mixto.

En los distintos modelos que se obtuvieron, el promedio general de los alumnos siempre era el que contaba con mayor ganancia de información, por lo cual era siempre el inicio del árbol generado. Los alumnos que contaban con una beca se iban descartando cuando se ejecutaba el algoritmo, ya que la ganancia de información que representaba era muy baja. En la mayoría de los modelos obtenidos, los alumnos que ingresaban sin haber cumplido la mayoría de edad eran clasificados como que no terminarían sus estudios.

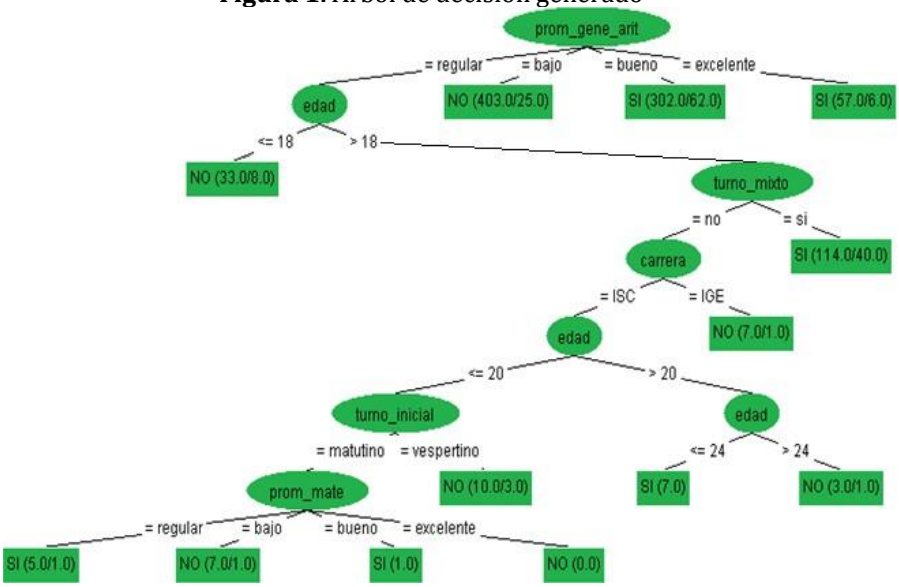
### *4.4. Evaluación*

Cuando se llega a la etapa de evaluación de un proyecto de MD es porque ya se han construido uno o varios modelos, que alcanzan una calidad suficiente desde una perspectiva de análisis de los datos. Una vez que se encuentra en esta etapa es importante evaluarlo a fondo, además de considerar si con los datos que se cuentan se ha omitido algo que pudiera ser de importancia.

En la Fig. 1 se muestra el árbol generado con la herramienta WEKA, dicho árbol alcanzó más de un 90% de certeza en las predicciones realizadas, los modelos generados que no alcanzaron un mínimo de

90% de certeza se fueron descartando, regresando a la fase de preparación de los datos para modificar la forma en que se estaban integrando los datos.

**Figura 1.** Árbol de decisión generado




Fuente: Elaborado por los autores, 2022.

Una vez que se obtuvo un árbol de decisión que cumplía con un mínimo de 90%, se puso a prueba con un 20% de datos no utilizados, obteniendo nuevamente un valor por encima del 90% de certeza en las predicciones realizadas, por lo tanto, el porcentaje del 90% de certeza se obtuvo tanto con el conjunto de entrenamiento como con el conjunto de prueba. Esta información se explica más a detalle en el siguiente párrafo:

Para poder validar el funcionamiento de los modelos desarrollados en MD se realizaron revisiones con datos reales previamente bien clasificados. Siempre es importante validar los modelos antes de implementarlos en un entorno real. Para la validación se utilizó el software Weka, y entre las funciones del algoritmo J48 se encuentra la opción de test. Una de las formas de realizar las pruebas es mediante la partición de los datos en 80% para entrenamiento y 20% para observar el comportamiento del clasificador. Otra de las formas de revisar el modelo es con una validación cruzada, que consiste en utilizar el 100% de los datos para generar un modelo, y el mismo 100% para obtener qué porcentaje de certeza se tiene.

En el caso de la validación con 80% para entrenamiento y 20% para pruebas, se dividió en 759 registros para entrenamiento y 190 para pruebas, teniendo como resultado a 170 estudiantes correctamente clasificados, lo que es equivalente a que el modelo tenga una certeza del 89.9%.

En el caso de la validación cruzada en donde se incluyen los 949 estudiantes para generar el modelo y los mismos 949 para realizar pruebas, se obtuvo 863 alumnos correctamente clasificados, lo que es equivalente a que el modelo tenga una certeza del 90.93%.

### *4.4. Despliegue*

Cuando se termina con el desarrollo de un modelo de MD no significa que sea el final del proyecto, el conocimiento que se pudo obtener en la etapa de modelamiento y evaluación, es parte de lo que se presentará al cliente para que éste pueda utilizarlo. Obedeciendo a lo que se plantea obtener, el proceso puede ser tan sencillo, así como muy complicado hasta la puesta en producción, la cual puede requerir procesos automatizados.

En este trabajo se llevó la solución a una aplicación móvil, para que fuera útil a los estudiantes y a las autoridades de una institución educativa, para conocer de primera mano una aproximación a la probabilidad que tiene un estudiante de desertar o de concluir con sus estudios. Para poder llevar esto a cabo se generó un archivo que contiene la Notación de Objetos de JavaScript (JSON), que es un formato ligero de intercambio de datos, el cual contiene la estructura necesaria en formato de pregunta-respuesta para poder utilizar el árbol generado, se creó de tal forma que pueda ser recorrido de forma

recursiva y que en caso de que los datos crezcan y se puedan volver a analizar y el árbol sufra modificaciones, sea necesario únicamente modificar este archivo y la aplicación funcione sin necesidad de realizar modificaciones adicionales, un fragmento del archivo JSON generado se muestra en la Fig. 2.

**Figura 2.** Fragmento de archivo JSON


```
}, {
    "respuesta": "Edad mayor a 18",
    "siguientePregunta": {
        "pregunta": "Turno Mixto?",
        "posiblesRespuestas": [{
            "respuesta": "sí",
            "mensaje": "Sigue asi!!! 65% de probabilidad que termi
        }, {
            "respuesta": "no",
            "siguientePregunta": {
                "pregunta": "Carrera?",
                "posiblesRespuestas": [{
                    "respuesta": "IGE",
                    "mensaje": "Sigue asi!!! 86% de probabilidad q
                }, {
                    "respuesta": "ISC",
                    "siguientePregunta": {
                        "pregunta": "Edad ingreso?",
```


Fuente: Elaborado por los autores, 2022.

El lenguaje de programación utilizado para dispositivos móviles fue Kotlin. En dicho lenguaje se programó una aplicación con una interfaz muy intuitiva, con la que solo basta ir pulsando entre las respuestas que ofrece para obtener un resultado. Se cuenta con un botón en caso de que se desee reiniciar el proceso, además de contar con el formato mencionado de pregunta-respuesta, con el cual basta con leer lo que solicita la aplicación y pulsar en la opción de nuestra preferencia hasta que nos dé como resultado la estimación expresada en porcentaje de que un estudiante pueda terminar la carrera.

## 5. Resultados y análisis

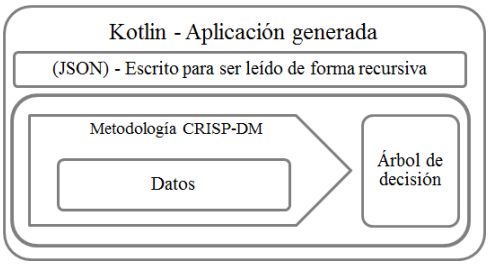
A partir del modelo obtenido, se puede observar que los alumnos que fueron contemplados en la muestra no necesariamente deben de contar con una beca para poder terminar o no sus estudios, de hecho, la ganancia de información que el algoritmo C4.5 debió obtener fue demasiado baja, ya que fue completamente descartada del árbol resultante.

En el caso de las materias de matemáticas, el árbol generado indicó que solamente en el caso de los alumnos que son de la carrera de Sistemas Computacionales influye en su formación profesional, ya que contar con calificaciones muy bajas en esta área, dan como resultado la deserción.

También se detectó que los alumnos que ingresan sin haber cumplido la mayoría de edad, en alrededor de un 75% no concluyen sus estudios.

En la Figura 3 se presenta el esquema general de cómo se desarrolló el proyecto completo.

**Figura 3.** Esquema general del desarrollo del proyecto




Fuente: Elaborado por los autores, 2022.

En el esquema general del proyecto presentado en la Figura 3, se observa el trabajo en su totalidad, partiendo de los datos, se utiliza una metodología orientada a proyectos de MD, dando como resultado un árbol de decisión, con el cual se genera un archivo con estructura JSON y creado para que pueda ser recorrido de forma recursiva. En el nivel más alto se encuentra el lenguaje de programación Kotlin con el cual se desarrolló la aplicación final.

A continuación, se expone un ejemplo de la forma de trabajo de la aplicación, describiendo parte del recorrido del árbol de decisión implementado. Para el caso presentado se supone un estudiante que ingresó a la carrera de ISC siendo mayor de edad de 19 años, con promedio general regular, que no ha tenido turno mixto, del turno matutino y con un promedio regular en matemáticas.

En la Figura 4 se muestra la interfaz de la aplicación y que es la que aparece cuando se abre la aplicación, en esta pantalla se puede observar el botón de iniciar con el cual se comienza a interactuar con la aplicación, además en la parte superior derecha se encuentra el botón de ayuda, que permite conocer algunos aspectos acerca de la aplicación.

**Figura 4.** Pantalla de inicio


Fuente: Elaborado por los autores, 2022.

En la Figura 5 se muestra la pantalla de ayuda, si se tiene alguna duda de lo que hace la aplicación o lo que predice, en esta pantalla se aclaran algunas situaciones, también se muestran los rangos considerados para las clasificaciones de los promedios de los estudiantes, del mismo modo, en caso de que se quiera conocer más acerca de la aplicación o de la investigación con la que se desarrolló, se proporciona el correo del autor. La pantalla de ayuda es indispensable en todas las aplicaciones, ya que brinda información que es de utilidad para los usuarios. En esta pantalla se aprecia el texto alineado a la izquierda, sin embargo, dependerá de la versión del sistema operativo Android en el cual se encuentre instalado, ya que en versiones más recientes se puede encontrar el texto justificado.

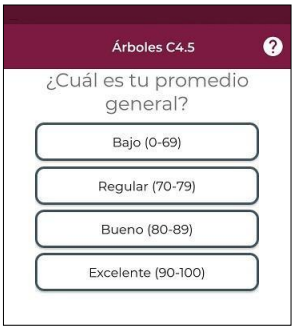
**Figura 5.** Pantalla de inicio


Fuente: Elaborado por los autores, 2022.

En la Figura 6 aparece la primera pantalla que se da al pulsar el botón de 'iniciar', en ella se observa la primera interacción con el usuario realizándose una pregunta acerca del promedio general, este promedio sirve para conocer cuál es el rendimiento general de un estudiante y a partir de éste se inicia el recorrido del árbol de decisión.

En la segunda interfaz se pregunta por la edad a la que ingresó el estudiante al nivel superior, esta pantalla es resultado de haber seleccionado un promedio general regular, y lo que trata de discernir es si un estudiante ingresó siendo menor de edad o no.

**Figura 6.** Pantalla solicitud de promedio

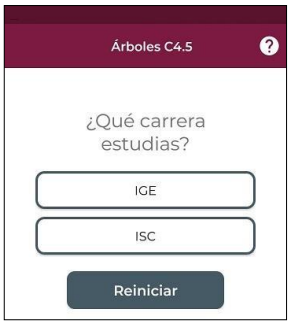



Fuente: Elaborado por los autores, 2022.

Enseguida se pregunta si el alumno ha tenido turno mixto, esta pantalla es resultado de haber seleccionado que cuando ingresó era mayor de edad, esta pregunta trata de identificar si un estudiante se ha mantenido en un turno o no.

En caso de que un estudiante no haya estado en un turno mixto se muestra la pantalla que aparece en la Figura 7, en ella se aborda el tema de la carrera en que el estudiante se encuentra para poder decidir hacia qué ramificación del árbol clasificador moverse.

**Figura 7**. Pantalla solicitud de Carrera




Fuente: Elaborado por los autores, 2022.

Si como respuesta se seleccionó que el estudiante pertenece a la carrera de Ingeniería en Sistemas Computacionales, nuevamente el algoritmo C4.5 determinó para el árbol el utilizar la edad de ingreso, es por esto que en la Figura 8 se muestra como pregunta la edad a la que ingresó un estudiante al nivel superior.

**Figura 8.** Pantalla solicitud de edad de ingreso


Fuente: Elaborado por los autores, 2022.

Cuando un estudiante tiene 19 ó 20 años, la siguiente pregunta que se le realizará es para conocer el turno en el que ingreso al nivel superior.

Finalmente, en el recorrido que se está mostrando como ejemplo, en caso de haber contestado que el estudiante ingresó en el turno matutino, se mostrará la pantalla que se muestra en la Figura 9, en esta pantalla se pregunta por el rendimiento que el alumno ha tenido en las materias de matemáticas exclusivamente.

**Figura 9.** Pantalla solicitud del promedio de matemáticas


Fuente: Elaborado por los autores, 2022.

Si el alumno selecciona que cuenta con un promedio regular en las materias de matemáticas, aparecerá la pantalla que se muestra en la Figura 10, en esta pantalla se concluye el recorrido de una de las ramificaciones del árbol de decisión, y se termina con un mensaje en el que se indica la probabilidad que tiene un estudiante de concluir sus estudios con unas palabras de aliento, en cuanto a las respuestas ofrecidas por el sistema es posible que los mensajes finales cambien tanto para las autoridades como para los estudiantes, con la finalidad de presentar a las autoridades un dato crudo y que al estudiante no se le vaya a afectar con una predicción negativa.

**Figura 10.** Pantalla respuesta del sistema


Fuente: Elaborado por los autores, 2022.

Como se pudo apreciar a lo largo de las distintas capturas de pantalla del sistema, interactuar con la aplicación es muy sencillo, basta con ir respondiendo lo que se pregunta para ir avanzando al siguiente cuestionamiento. En estas figuras se mostró el recorrido a lo largo de una de las ramificaciones del árbol de clasificación, siendo la intención conocer parte del sistema, ya que mostrar todas las pantallas de todos los casos posibles se tornaría repetitivo.

## 6. Conclusión

El modelo obtenido a través de la metodología CRISP-DM muestra que la mayoría de los alumnos que se mantienen en un solo turno tienen mayores probabilidades de completar sus estudios, lo cual sugiere que el estar en distintos turnos puede generar un proceso en el que no se logran adaptar. También se puede observar que llevando un promedio bajo es casi una tendencia a que el estudiante decida abandonar los estudios. Otro de los puntos que se puede observar en el modelo generado es el hecho de que, a pesar de que se contaba con información de los alumnos que tenían beca, al no aparecer en el modelo, significa que la ganancia de información fue tan baja que fue descartada, lo cual sugiere al igual

que otros trabajos revisados, que contar con una beca no es garantía de completar los estudios (Marcano & Rodríguez, 2014; Eckert & Suénaga, 2015; Ayala, López, & Menéndez, 2021).

Es importante seguir apoyando con herramientas que permitan otras lecturas, apoyados de diversas técnicas computacionales, con la intención de mantener más informados a los alumnos y al personal que pueda tomar decisiones que coadyuven en que crezca el número de estudiantes que terminan su formación universitaria, lo que coincide con lo planteado por Fayyad et al. (1996).

Para terminar, se propone para mejorar el proceso de discretización en la fase de preparación de los datos, se utilice Lógica Difusa, esto llevaría a tener mayor sentido en las clasificaciones realizadas a las calificaciones de los alumnos y que fueron interpretadas como bajas, regulares, buenas y excelentes; ya que mediante este mecanismo se podría decidir, por ejemplo, qué tanto pertenece un alumno de 79 a los alumnos regulares, o si ya debería de ser evaluado como un alumno bueno.

## Referencias


Álvarez, M., Gómez, E. &. Morfín, M. (2012). Efecto de la beca CONACYT en la eficiencia terminal en el posgrado. *Rev. Electrónica Investig. Educ., vol. 14*, núm. 1, 153–163.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1). https://doi.org/10.1186/s41239-020-0177-7

Ayala, E., López, R. E., & Menéndez, V. H. (2020). Factores asociados al bajo rendimiento académico de estudiantes de primer semestre en carreras de computación. *Congreso Internacional de Investigación Academia Journals Chetumal 2020, 12*(2), 38–43. https://www.academiajournals.com/pubchetumal2020

Comisión Nacional de los Derechos Humanos (CNDH). (2018). *Derecho a la educación* | Comisión Nacional de los Derechos Humanos México. http://www.cndh.org.mx/Derecho_Educacion.

Dicovskiy L. M. & Pedroza, M. E. (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico universitario, *Revista Científica FAREM- Estelí, 24*, 143, https://doi.org/10.5377/farem.v0i24.5557

Eckert, K. B. & Suénaga, R. (2015). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos, *Formación Universitaria, 8*(5), 3–12. https://doi.org/10.4067/S0718-50062015000500002

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 en Instituciones de Educación Superior particulares. *Revista Electrónica Educare*, *20*(3), 1-21. https://doi.org/10.15359/ree.20-3.11

Fadhl, F. y Sofian, S. (2015). A review of balanced scorecard framework in higher education institution (HEIs). *International Review of Management and Marketing, 5*(1), 26-35.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. Aaai Press.

Hernández, J., Ramírez, M. J., Ferri, C. (2004). *Introducción a la minería de datos*. Pearson.

Hernández, D., Vargas, A., Almuiñas, J. y García, J. (2015). Los indicadores actuales de la eficiencia académica: necesidad de su perfeccionamiento. *Pedagogía Universitaria, 20*(3), 53-62. http://cvi.mes.edu.cu/peduniv/index.php/peduniv/article/view/690

Instituto Tecnológico de Iztapalapa, (2018). *Rendición de cuentas*.

Marcano, Y. J. & Rodríguez, R. (2014). Minería de datos aplicada a la deserción estudiantil. Caso: Licenciatura en Computación de la Universidad del Zulia, NPF, *Revista EDUCARE - UPEL-IPB - Segunda Nueva Etapa 2.0, 18*(2), 31–51. https://doi.org/10.46498/reduipb.v18i2.131

Márquez, A. (2012). El financiamiento de la educación en México: Problemas y alternativas. *Perfiles Educativos 34*(1), 107– 117.

Naciones Unidas (2015). *La Declaración Universal de Derechos Humanos*, http://www.un.org/es/universal-declaration- human-rights/.

Oviedo, A. I., Jiménez, J. (2019). Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas saber-pro. *Revista Politécnica, 15*(29), 128-140. https://doi.org/10.33571/rpolitec.v15n29a10

Secretaría de Gobernación (SEGOB). (2023). *Plan Nacional de Desarrollo 2019-2024.* México. https://www.dof.gob.mx/nota_detalle.php?codigo=5565599&fecha=12/07/2019#gsc.tab=0

Panizzi, M. (2019). Establecimiento del estado del arte sobre la Minería de Datos Educacional en el Nivel Superior: Un Estudio de Mapeo Sistemático. *Revista de Investigaciones Científicas de la Universidad de Morón, 4*(2), 51-60.

Peinado, J. y Jaramillo, D. (2018). La eficiencia terminal del Centro de Investigación e Innovación Tecnológica. *Revista Electrónica de Investigación Educativa, 20*(3), 126-134. https://doi.org/10.24320/redie.2018.20.3.1797

Proyecto ALFA GUIA DCI-ALA/2010/94. Encuesta Internacional sobre el Abandono en la Educación Superior, 2014.

Reyes, A., Flores, A., Alejo, F. & Rendón, E. (2017). Minería de datos aplicada para la identificación de factores de riesgo en alumnos, *Research in Computing Science 139*, 177-189.

Román, A. B., Sánchez, D. & García, R. (2017). Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo, *Latin-American Journal of Physics Education 4*(1), 7.

Romero, A., Utrilla, A. & Utrilla V. M. (2014). Las actitudes positivas y negativas de los estudiantes en el aprendizaje de las matemáticas, su impacto en la reprobación y la eficiencia terminal. *Ra Ximhai, vol. 10*, núm. 5.

Toscano de la Torre, B. (2016). La Eficiencia Terminal como un Indicador de la Calidad en la Educación Superior en México. En R. Enciso (Ed.). *Las Universidades y sus Estrategias de vinculación* (pp. 6-9). UTP.

Schröer, Ch., Kruse, F., Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science 181*, 526–534

Torre, B. A., Gallegos, J. C., Meza, J. & Granados, S. A. (2015). *Análisis de la Eficiencia Terminal en un Programa Educativo de Tecnologías de Información. Caso: Universidad Autónoma de Nayarit*. https://doi.org/10.13140/rg.2.1.3993.4568

Vera, J. A., Ramos, D. Y., Sotelo, M. A., Echeverría, S., Serrano D. M. & Vale, J. J. (2012). Factores asociados al rezago en estudiantes de una institución de educación superior en México. *Rev. Iberoam. Educ. Super., vol. 3*, núm. 7, 41–56, 2012.

Villalobos, L. O. (2017). La eficiencia terminal en tres generaciones de alumnos de la división de ciencias básicas e ingeniería de la UAM Azcapotzalco, *XIV Congreso Nacional de Investigación Educativa COMIE*, p.13.

Weka. (2019, February). University of Waikato. *Machine Learning Group*. www.cs.waikato.ac.nz/ml/weka/downloading.html.