Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos

Predictive models of academic risk in computing careers with educational data mining

Enrique Ayala Franco
Universidad Autónoma de Yucatán. Mérida, México
enrique.ayala@correo.uady.mx

Rocío Edith López Martínez
Universidad Autónoma de Querétaro. Querétaro, México
rocio.edith.lopez@uaq.mx

Víctor Hugo Menéndez Domínguez
Universidad Autónoma de Yucatán. Mérida, México
mdoming@correo.uady.mx

Resumen
Los problemas de bajo rendimiento académico y rezago son recurrentes en instituciones educativas de nivel superior, especialmente al inicio de los estudios universitarios. En el contexto local, análisis diagnósticos han mostrado altos índices de reprobación y bajo rendimiento académico. En este trabajo, se utilizaron datos sociodemográficos y resultados de exámenes de admisión de 415 alumnos de las carreras del área de computación de la Universidad Autónoma de Yucatán (México), inscritos entre 2016 y 2019. El objetivo es generar modelos predictivos de riesgo académico, empleando métodos de la minería de datos educativa, que sirvan como herramientas de detección temprana de condiciones de riesgo académico y faciliten el despliegue de estrategias de intervención educativa. Se siguieron las etapas del Proceso de Extracción de Conocimiento en Bases de Datos, concretamente, se aplicaron técnicas de clasificación para el análisis, obtención y validación de los modelos. Los resultados muestran que el mejor modelo corresponde al algoritmo LMT, con un valor de precisión de 75.42% y un 0.805 para el área bajo la curva ROC. Se logró identificar a los mejores atributos predictores, particularmente las pruebas del examen de ingreso a licenciatura fueron muy significativas. Se propone el desarrollo de herramientas informáticas para la detección precoz de riesgo académico y estrategias de intervención educativa oportuna.
Palabras clave*:* minería de datos educativos, modelo predictivo, clasificación, riesgo académico, educación superior.

Abstract
The problems of poor academic performance and lag are recurrent in higher-level educational institutions, especially at the beginning of university studies. In the local context, diagnostic analyzes have shown high failure rates and low academic performance. In this work, sociodemographic data and admission exam results of 415 students of the computer science majors of the Autonomous University of Yucatán (Mexico), enrolled between 2016 and 2019, were used. The objective is to generate predictive models of academic risk, using educational data mining methods that serve as tools for early detection of academic risk conditions and facilitate the deployment of educational intervention strategies. The stages of the

Knowledge Extraction Process in Databases were followed, specifically, classification techniques were applied for the analysis, obtaining and validation of the models. The results show that the best model corresponds to the LMT algorithm, with a precision value of 75.42% and 0.805 for the area under the ROC curve. It was possible to identify the best predictive attributes, particularly the bachelor entrance exam tests were very significant. The development of computer tools for the early detection of academic risk and timely educational intervention strategies is proposed.
Keywords: educational data mining, predictive model, classification, academic risk, higher education.

1. Introducción

Las problemáticas derivadas del bajo rendimiento académico han sido temas comunes de preocupación y estudio en el contexto universitario (Padua, 2019). El rendimiento académico, también llamado desempeño académico o rendimiento escolar, es un constructo complejo que implica el logro de objetivos de aprendizaje establecidos en cursos o programas de estudio y puede ser entendido como un nivel de conocimientos demostrado, lo que involucra procesos de evaluación (Lamas, 2015).

Existe una gran variedad de factores personales, sociales e institucionales, en diferentes contextos geográficos o temporales, que explican los procesos de aprendizaje y sus resultados. El propósito de muchas de las investigaciones desarrolladas ha sido identificar el tipo de influencia y relaciones entre las distintas categorías o dimensiones de variables involucradas en el rendimiento académico (Muñoz, 2015).

El riesgo académico es un estado en el que confluyen diversos factores que colocan al estudiante en una situación de propensión de rezago escolar, bajo aprovechamiento escolar, bajo rendimiento académico o de fracaso escolar (Pacheco, Cruz y Serrano, 2019). Comúnmente, un resultado de bajo rendimiento académico se presenta cuando el estudiante no logra obtener las calificaciones establecidas por la institución educativa o cuando no aprueba asignaturas, lo cual es indicativo de riesgo académico que puede derivar en rezago o fracaso escolar, en el que el estudiante puede ver truncada su experiencia universitaria o continuarla, pero de una forma muy precaria.

Por su importancia, se han realizado numerosos estudios para entender los factores del rendimiento escolar desde diversas áreas del conocimiento como la psicología, la sociología o la pedagogía, y, recientemente, empleando técnicas de la Minería de Datos Educativos (MDE).

A partir de la conformación de nuevas disciplinas y comunidades para el análisis de datos educativos, se han incrementado los trabajos relacionados con los estudios de las características de los alumnos y sus ambientes escolares (Baker e Inventado, 2014). En particular, la disciplina emergente de la minería de datos educativos tiene el propósito de explorar datos provenientes de entornos educativos para encontrar patrones descriptivos y predicciones que caracterizan los comportamientos y logros de los alumnos, el contenido de conocimiento de dominio, evaluaciones, funcionalidades educativas y aplicaciones; empleando para ello diversos métodos, técnicas y algoritmos de descubrimiento de información (Peña-Ayala, 2014).

El uso de MDE en el contexto educativo tiene el potencial de trasformar los modelos existentes de enseñanza-aprendizaje al proporcionar nuevas herramientas de análisis,

interacción e intervención (Aldowah, Al-Samarraie y Fauzy, 2019). Los múltiples estudios sobre rendimiento académico han aportado información valiosa para el entendimiento y la planificación del proceso educativo, sin embargo, la problemática del bajo rendimiento escolar aún se sigue manifestando en muchas instituciones de educación superior, reflejándose en altos índices de reprobación y abandono, de una forma más notoria en los primeros años de estudio universitario (Silva, 2011).

En el contexto actual de la educación, se vuelve necesaria la incorporación de nuevas metodologías y tecnologías informáticas, como la minería de datos educativos y las analíticas del aprendizaje, para facilitar y acelerar el análisis de grandes volúmenes de información generados por los diversos sistemas de información escolares; aunque su empleo en el contexto de la educación superior aún es incipiente (Aldowah et al., 2019). Además, existe la necesidad y la oportunidad de emplear estas nuevas herramientas en el desarrollo de investigaciones aplicadas con el propósito de mejorar la calidad de la educación, o en el desarrollo de investigaciones puras que buscan mejorar el entendimiento del proceso de aprendizaje (Bakhshinategh, Zaiane, ElAtia e Ipperciel, 2018).

El uso de este tipo de tecnología educativa sigue siendo un reto en el contexto universitario para mejorar la calidad de la toma de decisiones de instructores y administrativos, así como para identificar y analizar muchos de los elementos del entorno académico de manera eficiente y rápida. La urgencia de su empleo también se justifica por el creciente uso de entornos virtuales de aprendizaje, redes sociales y diversos sistemas de información basados en Internet, los cuales generan grandes volúmenes de información digital, difícilmente manejados por los métodos tradicionales de análisis (Villanueva, Moreno y Salinas, 2018). En este sentido Gros (2015) sugiere utilizar las analíticas del aprendizaje y la minería de datos para interpretar la gran cantidad de datos de los estudiantes generados en los nuevos entornos digitales, también para registrar y analizar su progresión académica, predecir actuaciones futuras e identificar elementos problemáticos.

La problemática detectada en el contexto local, a partir de estudios diagnósticos, es referente a altos índices de reprobación y bajo rendimiento académico en los primeros semestres de estudio de las carreras del área de computación en la Facultad de Matemáticas de la Universidad Autónoma de Yucatán (UADY) (Ayala, López y Menéndez, 2020). Además, tales circunstancias ocasionan problemas de rezago y bajo aprovechamiento escolar, desencadenando situaciones que complican el resto de la trayectoria escolar de los estudiantes. Por ejemplo, dificultades para inscribir materias por traslapes en horarios y excesiva carga de trabajo por presentar exámenes o llevar cursos de regularización; en consecuencia, muchas de las veces abandonan sus estudios o son dados de baja por exceder las oportunidades para acreditar materias de acuerdo con la reglamentación institucional.

Adicionalmente, para el apoyo de los estudiantes, existe un programa de tutorías en el cual profesores-tutores actúan como guías en el medio académico para orientar en cuestiones de su vida escolar (UADY, 2012); sin embargo, en el primer semestre no reciben de manera oportuna y sistematizada información sobre los perfiles de los estudiantes, por lo cual es complicado hacer un diagnóstico preciso de sus problemáticas y atender de manera efectiva cada situación. A pesar de contar con sistemas de información institucionales, la carencia de mecanismos y procedimientos para extraer información relevante de la población estudiantil, principalmente en el

primer semestre de estudios, afecta la toma de decisiones de los directivos y docentes que, al no contar con una visión más precisa de los alumnos que reciben y sus problemáticas, dejan de implementar acciones para la mejora educativa.

Ante la problemática detectada en el contexto local, y los retos y tendencias en el uso y aplicación de la tecnología informática, particularmente los referentes a la minería de datos educativos, es que se propone el desarrollo de este trabajo de investigación; con el propósito de extraer conocimientos de los datos disponibles en el entorno institucional y generar modelos predictivos que ayuden a la identificación de estudiantes en riesgo.

El objetivo general es obtener modelos predictivos eficientes de riesgo académico de los alumnos en programas del área de computación de la Universidad Autónoma de Yucatán, generados mediante técnicas de la minería de datos educativos.

Los objetivos específicos consisten en:

1) Seleccionar atributos para ser empleados en la generación de los modelos predictivos, en función de su relevancia predictiva.
2) Determinar los modelos más efectivos para la predicción de riesgo académico de los alumnos de primer ingreso universitario.

Para la recuperación y análisis de los datos se siguen los pasos definidos en el Proceso de Extracción de Conocimiento en Bases de Datos (Márquez-Vera, Romero y Ventura, 2012; Peña-Ayala, 2014), y se emplean varias de las técnicas de la minería de datos educativos, particularmente de clasificación o predicción, para generar los modelos y evaluar su rendimiento.

Con los resultados obtenidos se pueden implementar los modelos en sistemas informáticos para ser utilizados en la detección de estudiantes con bajo rendimiento académico o en situación de riesgo escolar, al momento de su ingreso a la universidad. Así mismo, los modelos y la información adquirida ayudarán a la planificación educativa y la toma de decisiones oportunas por parte de tutores, docentes y directivos.

En los siguientes apartados se describen brevemente los fundamentos teóricos que aclaran algunas definiciones sobre rendimiento y riesgo académico, y se incluye un resumen introductorio a la minería de datos educativos, sus tareas principales y herramientas de software disponible; así como una descripción general de los métodos de clasificación y técnicas de evaluación de atributos y modelos. Seguidamente, se presenta la metodología utilizada, los resultados alcanzados, la discusión de los resultados con su interpretación y finalmente las conclusiones principales, en dónde se proponen acciones concretas de intervención educativa.

2. Fundamentos teóricos

En este apartado se describen algunos de los fundamentos teóricos y conceptuales que contribuyen al entendimiento del trabajo realizado. Se proporciona una explicación de los conceptos utilizados, así como una introducción a los principales elementos de la minería de datos educativos con el propósito de clarificar sus funciones y las herramientas disponibles para su aplicación en el entorno educativo.

2.1 Rendimiento y riesgo académico

El rendimiento académico es la expresión de las capacidades y las características psicológicas del alumno, desarrolladas y actualizadas a través del proceso de enseñanza aprendizaje que le posibilita obtener un nivel de funcionamiento y logros académicos a lo largo de un período o semestre, que se sintetiza en un calificativo final (cuantitativo en la mayoría de los casos) evaluador del nivel alcanzado (Chadwick, 1979, como se citó en Molina, 2015).

La variable rendimiento académico puede ser explicada por una multitud de causas tanto psicológicas como sociológicas, por ejemplo, inteligencia, sexo, aptitud, antecedentes académicos, antecedentes familiares, características de la escuela, resultados de evaluaciones previas, métodos de enseñanza, entre otros (Kerlinger y Lee, 2002).

Desde una perspectiva holística, el éxito académico en la universidad es determinado por la preparación de sus estudiantes, el tiempo de egreso, la inserción laboral, los programas de investigación, la proyección social, entre otros, por lo que la evaluación del rendimiento académico considera la suma de los diferentes y complejos factores que intervienen en el estudiante. Por ejemplo, factores personales, sociales y propios de las instituciones de educación superior que forman el conjunto de elementos determinantes del rendimiento escolar (López-Ramirez, 2015).

Para García (2015), existe una distinción entre el rendimiento inmediato, en el que se consideran las notas o calificaciones de las materias; y el rendimiento mediato, el cual incluye logros personales y profesionales. También destaca que es frecuente el uso de los siguientes indicadores para determinar el nivel de rendimiento académico: calificaciones, pruebas objetivas, cantidad de materias aprobadas y cantidad de créditos acumulados.

Para aclarar algunos conceptos que pudieran ser confusos, Carpio et al. (2018) proporcionan algunas definiciones relacionadas con el rendimiento y el riesgo académico:

El riesgo académico se define como un estado caracterizado por la concurrencia de factores que tornan al estudiante propenso a colocarse en alguna de las condiciones siguientes:

a) Rezago escolar, entendido como el desfase temporal entre los logros planeados en los tiempos curriculares y los alcanzados por los estudiantes en un momento determinado.

b) Bajo rendimiento escolar, referido a la mala calidad de los logros de los estudiantes.

c) Bajo nivel de aprovechamiento escolar, se refiere a un aprendizaje insuficiente para cubrir los objetivos programados en las asignaturas.

d) Fracaso escolar, es el incumplimiento irreversible de las metas escolares y de aprendizaje, imposibilita la continuación a otro nivel educativo.

Para estimar el bajo rendimiento escolar se emplean comúnmente los promedios de las calificaciones obtenidas en las distintas asignaturas cursadas, o mediante la razón *asignaturas cursadas* / *asignaturas aprobadas*. Algunos de los estudios para determinar predictores de riesgo académico, como el de Baker, Lindrum, Lindrum y Perkowski (2015), han considerado como alumnos en riesgo académico a aquellos con promedios de calificaciones iguales o menores a un 73%. En el mismo sentido, en la investigación

de Padua (2019), se trabajó con alumnos universitarios con bajo promedio de calificaciones y para ello se consideró un puntaje menor a 7.5 como indicador de tal estado académico. En ambos casos, una vez identificados los alumnos en riesgo, se les brindaba apoyo, tratando de indagar en factores familiares e individuales para atender las causas y reducir el riesgo académico.

Otros estudios como el de Dorio (2017), señalan que el primer año de universidad es el momento más significativo de la experiencia universitaria pues representa el inicio de un período de cambio clave en la vida del estudiante, tanto personal como académico. Se destaca que los abandonos de los estudiantes se concentran en mayor medida en este período inicial ante la dificultad para adaptarse a la nueva situación. Sin embargo, también se señala que las experiencias académicas y sociales positivas iniciales refuerzan y motivan la acción de continuar y persistir.

Por ello, una responsabilidad institucional es comprender las necesidades de los estudiantes que llegan a la universidad, de igual forma, dar el seguimiento adecuado para saber cómo transcurre su integración y adaptación al contexto universitario, y sobre todo identificar aquellos factores que facilitan o inhiben el proceso de transición. La implementación de mecanismos que recuperen dicha información contribuirá al diseño de programas de apoyo, a estudiantes de primer curso, que procuren establecer bases sólidas para el desarrollo de trayectorias escolares exitosas.

2.2 La minería de datos educativos

El desarrollo de herramientas y algoritmos computacionales para extraer información potencialmente útil y novedosa a partir de bases de datos generadas en contextos educativos ha dado lugar a una nueva disciplina emergente de las Ciencias Computacionales denominada Minería de Datos Educativos, la cual es un área de investigación interdisciplinaria que utiliza métodos para explorar los datos que surgen en un campo escolar (Baker y Yacef, 2009). Los enfoques computacionales que utiliza permiten examinar y estudiar datos escolares para responder a preguntas educativas. La MDE proporciona un conocimiento intrínseco del proceso de enseñanza y aprendizaje, que, mediante la generación de modelos, pueden responder a preguntas relacionadas con el rendimiento escolar y sus problemáticas asociadas y facilitar una planificación educativa efectiva (Anoopkumar y Rahman, 2016).

La MDE ofrece alternativas novedosas para analizar datos educativos y descubrir información relevante en las poblaciones estudiadas. A continuación, se describen algunas de sus principales aproximaciones o métodos de análisis de acuerdo con las técnicas que emplean (Peña-Ayala, 2014; Romero y Ventura, 2010):

- Predicción: desarrollo de modelos para inferir una variable a partir de la combinación de otros datos disponibles o variables predictoras. Algunos de los métodos usados para predecir son: *clasificación* (cuando la variable a predecir es un valor categórico), *regresión* (si la variable tiene valor continuo), o *estimación de densidad* (cuando la variable a predecir es una función de densidad de probabilidad). Se pueden aplicar estos métodos para predecir el éxito o el fracaso académico de los estudiantes.
- Agrupamiento: para encontrar conjuntos de datos que se agrupen naturalmente, y separarlos del conjunto completo en una serie de categorías. Se pueden crear grupos de estudiantes basados en sus patrones de aprendizaje o estrategias cognitivas, por ejemplo.

- Minería de relaciones: se usan para descubrir relaciones entre variables y codificarlas como reglas para aplicarlas posteriormente. Algunos de sus métodos son: *minería de reglas de asociación* (cualquier relación entre variables), *minería de patrones secuenciales* (asociaciones temporales entre variables), *minería de correlaciones* (correlaciones lineales), *minería de datos causales* (relaciones causales entre variables). Usadas para identificar relaciones, de las actividades de los estudiantes con sus resultados finales, y para modelar secuencias de actividades de aprendizaje.
- Descubrimiento mediante modelos: usado para validar el modelo de un fenómeno (mediante *predicción*, *agrupamiento* o *ingeniería del conocimiento*). Es usado para identificar la relación entre características y comportamiento de los estudiantes.
- Destilado de datos: para permitir a un humano identificar o clasificar rápidamente las propiedades de un conjunto de datos. Utiliza resúmenes, visualización e interfaces interactivas para destacar la información relevante y apoyar la toma de decisiones.
- Detección de valores atípicos: para descubrir datos significativamente diferentes al resto del conjunto. Se puede utilizar para detectar desviaciones en las acciones o comportamientos del alumno o educador, procesos de aprendizaje irregulares y para detectar estudiantes con dificultades de aprendizaje.
- Análisis de redes sociales (en inglés, Social Network Analysis, SNA): estudia las relaciones entre los individuos. El SNA considera las relaciones sociales en términos de la teoría de red, la cual considera nodos, que representan actores individuales dentro de la red, y conexiones o enlaces, que representan relaciones entre los individuos, como amistad, relaciones cooperativas, etc. Esto puede usarse para interpretar y analizar la estructura y las relaciones en tareas colaborativas e interacciones con herramientas de comunicación.

Además de las anteriores, se siguen sumando otras metodologías al repertorio de la MDE, por ejemplo, la Minería de Textos, la Minería de Procesos y el Trazado de Conocimientos, entre otras. La elección de una técnica determinada depende del entorno educativo, los objetivos de investigación y la disponibilidad de los datos.

2.3 Herramientas de la MDE

Hay disponibles varias herramientas que hacen factible la aplicación de las técnicas de la MDE, las cuales incluyen algoritmos para realizar distintos análisis de datos educativos. Algunas de las más destacadas, de acuerdo a Slater, Joksimović, Kovanovic, Baker y Gasevic (2017), son: *RapidMiner*, *DBminer*, *Waikato Environment for Knowledge Analysis (*WEKA*)*, *KEEL*, *KoNstanz Information MinEr (KNIME), Orange* y *Statistical Product and Service Solutions (*SPSS*)*; además se han desarrollado paquetes o módulos para ser utilizados con *Python* y *R*.

Las herramientas anteriores cuentan con un amplio repertorio de algoritmos y técnicas, que las hacen apropiadas para la mayoría de las tareas de la MDE, incluyen los principales métodos que han sido utilizados en el análisis del rendimiento académico. También, como lo indica Slater et al. (2017), hay herramientas diseñadas específicamente para otras tareas de la MDE, como la visualización o el análisis de

redes sociales, entre otras, por ejemplo el software *Graphical Interactive Student Monitoring (*GISMO*)*

En la revisión de la literatura, se señala el uso constante del software WEKA, el cual es un ambiente de simulación computacional que presenta un vasto soporte para la experimentación, con varios métodos estadísticos y de inteligencia artificial (Witten, Frank y Hall, 2011). El ser de libre distribución y de código abierto ha sido un factor importante para ganar popularidad en la comunidad de investigadores en el área de minería de datos. WEKA cuenta con un extenso repertorio de algoritmos de clasificación, predicción, agrupamiento y minería de asociaciones. También dispone de técnicas para aplicar varios criterios en la selección de atributos y comparar su desempeño, de tal forma que se pueda elegir el subconjunto de variables que mejores resultados obtengan para la clasificación.

Además, incluye varios métodos de validación y verificación de la precisión de los resultados. Puede usarse en línea de comandos, a través de su interfaz gráfica o mediante la interfaz de programación del lenguaje Java. Los modelos generados pueden ser almacenados en archivos para, posteriormente, ser utilizados con nuevos datos y realizar predicciones. Una característica importante es la posibilidad de desarrollar aplicaciones Java e incluir las clases WEKA. Esto permite recuperar datos, crear instancias de métodos de clasificación a partir de modelos previamente generados en WEKA, y utilizarlas para evaluar nuevos casos, entre otras capacidades. Por lo anterior, se decidió la utilización del software WEKA versión 3.8.3, para el desarrollo del presente trabajo.

2.4 Evaluación de atributos

Las técnicas de selección de atributos ayudan a identificar los más significativos, es decir aquellos que aporten el mayor poder predictivo en el modelo. Además, cuando existen muchas características, ayudan a reducir la complejidad de los modelos mejorando su comprensión, disminuyen la necesidad de más espacio de almacenamiento y acortan el tiempo para el entrenamiento y el procesamiento. De igual forma, la recopilación de nuevos casos se simplifica al ignorar aquellos atributos que no tienen potencial predictivo (Márquez-Vera, Romero y Ventura, 2012).

El objetivo de la selección de atributos es obtener un subconjunto más pequeño que al generar los modelos no se afecte significativamente el porcentaje de clasificación y que la distribución resultante sea similar al conjunto original. Este subconjunto de características es el resultado de un proceso de filtrado en función de su relevancia predictiva.

No existe una sola técnica de selección de atributos para obtener siempre los mejores resultados, muchas veces es necesario un especialista en el área de estudio para ayudar en la selección de variables de acuerdo con su experiencia, y, en muchos casos, es necesario hacer varias pruebas empíricas para encontrar los mejores subconjuntos.

Se describen a continuación algunas de las técnicas disponibles en WEKA para determinar los mejores subconjuntos de variables, que posteriormente serán empleadas por los algoritmos de predicción y clasificación (Bouckaert et al., 2018; García Gutiérrez, 2016).

CfsSubsetEval o CFS: evalúa el valor de un subconjunto de atributos al considerar la capacidad predictiva individual de cada característica junto con el grado de redundancia

entre ellas. Se prefieren los subconjuntos de características que están altamente correlacionados con la clase, mientras que tengan una baja intercorrelación entre atributos. La función de evaluación heurísticas es una función de correlación estadísticas. Los atributos con información redundante son penalizados puesto que tendrán una alta correlación con otras características, y los atributos irrelevantes son ignorados, puesto que tendrán una baja correlación con la clase.

CorrelationAttributeEval: evalúa el valor de un atributo midiendo la correlación (Pearson) entre él y la clase. Los atributos nominales se consideran valor por valor al tratar cada valor como un indicador. Se llega a una correlación general para un atributo nominal a través de un promedio ponderado.

WrapperSubsetEval: evalúa conjuntos de atributos utilizando un esquema de aprendizaje. Se trata de un método de envoltura en el cual, para evaluar y seleccionar los atributos, se necesita un método de inducción. Es decir, se prueban diversos parámetros y subconjuntos de atributos, aplicando un algoritmo de clasificación, y el criterio de parada es el nivel de rendimiento medido en cada caso. La validación cruzada se utiliza para estimar la precisión del esquema de aprendizaje para un conjunto de atributos.

2.4 Métodos y modelos de predicción

Dada la naturaleza del estudio sobre rendimiento académico, se han empleado principalmente métodos de Predicción, Regresión o Agrupamiento. Los algoritmos ubicados en estas categorías de tareas son de naturaleza inductiva, es decir, buscan derivar o descubrir inductivamente características o patrones a partir de los datos (datos de entrenamiento). El conocimiento obtenido se sintetiza en modelos predictivos los cuales pueden emplearse con nueva información (datos de prueba). Se definen dos grandes grupos de tipos de algoritmos: a) métodos supervisados, que realizan su proceso de aprendizaje con base en un conjunto de datos en donde los valores de las clases son conocidos, y b) métodos no supervisados, cuando no se conocen los valores de las clases, o no están definidas de antemano (Minguillón, Casas y Minguillón, 2017). Los principales métodos de minería de datos predictivos de tipo supervisado, de acuerdo con los mismos autores, son: k Vecinos más Cercanos (en inglés, *k-Nearest Neighbors*, k-NN), Redes Neuronales, Árboles de Decisión, Máquinas de Vector de Soporte (en inglés, *Support Vector Machines*, SVM) y métodos probabilísticos.

2.5 Validación de modelos

Para evaluar la precisión de los modelos obtenidos mediante las técnicas de clasificación, en primer lugar, se debe aplicar el algoritmo y obtener la clasificación o predicción de la clase, luego comparar con el valor real de las instancias y determinar la precisión y errores de la clasificación.

La precisión es el número de predicciones correctas, tanto para una clase positiva como para la clase negativa, sobre el número total de predicciones:

$$Precisión = [(VP + VN) / N] \quad (1)$$

En donde *VP* es verdadero positivo, *VN* es verdadero negativo y *N* la cantidad de instancias.

El error indica la tasa de mala clasificación y esta determinado por:

$$Error = [(FP + FN) / N] \quad (2)$$

En donde *FP* es falso positivo, *FN* es falso negativo y *N* la cantidad de instancias.

Una manera alternativa de evaluar el rendimiento de un modelo generado con algoritmos de aprendizaje máquina es mediante las curvas ROC (*Receiver Operating Characteristics*) (Fawcett, 2003), la cual es una técnica para visualizar su desempeño y ha sido adoptada por la comunidad de investigadores en minería de datos. Se calcula con base en los falsos positivos (FP) y los verdaderos positivos (VP). A partir de la curva ROC se determina el AUC (*Area Under the Curve*), que es un indicador de la calidad del modelo. Si el valor de AUC se acerca a la unidad, significa que el clasificador se comporta de manera óptima, aproximándose al clasificador perfecto.

De acuerdo a Minguillón et al. (2017), se pueden interpretar las curvas ROC utilizando los intervalos de valores AUC de la siguiente guía:

[0.5 – 0.6): Test malo

[0.6 – 0.75): Test regular

[0.75 – 0.9): Test bueno

[0.9 – 0.97): Test muy bueno

[0.97 – 1]: Test excelente

El software WEKA cuenta con las siguientes opciones para el proceso de evaluación y validación de modelos (Bouckaert et al., 2018):

- Utilizar el conjunto de entrenamiento: se evalúa el clasificador utilizando el mismo conjunto de datos utilizado para construir el modelo, por lo que el resultado suele ser demasiado optimista.
- Suministrar un conjunto de prueba: se consigue un conjunto de datos independiente para realizar las pruebas. Para cada instancia del nuevo conjunto se realiza la predicción y se estima la precisión del modelo.
- Validación cruzada (*k-fold cross validation*): se divide el conjunto de datos en *K* partes (*folds*) y se realizan *K* iteraciones en las que se va reservando una parte como datos de prueba y el resto como datos de entrenamiento. Al final se promedian los resultados de las precisiones y errores calculados en todas las iteraciones.
- Indicar un porcentaje de división: se dividen los datos en un grupo de entrenamiento y un grupo de prueba, de acuerdo con un porcentaje indicado.

La validación cruzada es una de las formas más consistentes de evaluar clasificadores, ya que los conjuntos se determinan de manera aleatoria y el error de la evaluación es muy bajo. Comparado con los demás métodos, es la validación más estricta para verificar la precisión de los modelos, esto asegura su capacidad de generalización.

2.6 Trabajos relacionados

En el trabajo de Kumar, Singh y Handa (2017) se hace una revisión de estudios realizados de 2007 a 2016, para identificar las principales técnicas de minería de datos utilizadas para predecir el progreso y desempeño de estudiantes e identificar los

atributos más importantes para las predicciones; en la mayoría de los trabajo se utilizaron los algoritmos de Decision Tree, Naive Bayes, Rule-Based, Artificial Neural Networks y K-Nearest Neighbor; y las mejores predicciones del rendimiento académico se han logrado con una combinación de atributos personales, socioeconómicos y de estudios previos.

Por su parte, Kumar y Singh (2017) evaluaron varias técnicas de la MDE para determinar las mejores. Utilizaron datos tanto académicos como familiares de alumnos de la Universidad Himachal Pradesh, India. El algoritmo de modelado predictivo Random Forest obtuvo los mejores resultados, comparado con los algoritmos Decision Tree, Naive Bayes, Bayes Network y árboles de clasificación, también emplearon técnicas de validación cruzada y conjuntos de entrenamiento.

Dada la creciente preocupación por el bajo rendimiento de los alumnos en los primeros cursos de estudios universitarios, en el trabajo llevado a cabo por Costa, Fonseca, Santana, de Araújo y Rego (2017), se analizó la efectividad de las técnicas de minería de datos (Decision Tree, Support Vector Machine, Neural Network and Naive Bayes), para la predicción oportuna del posible fracaso escolar en cursos de programación, en la Universidad Federal de Alagoas (UFAL), Brasil; los autores concluyen que las técnicas analizadas son suficientemente efectivas para el propósito planteado.

Martínez, Karanik, Giovannini y Pinto (2015) consideran al rendimiento académico como un factor crítico asociado a las altas tasas de deserción; analizaron resultados de bajo rendimiento académico en asignaturas del primer nivel en la Universidad Tecnológica Nacional, provincia del Chaco, Argentina; y mediante técnicas de la minería de datos determinaron perfiles de alumnos con bueno y bajo rendimiento académico, con base en antecedentes de estudios previos y datos familiares.

López, Guzmán y González (2015) generaron un modelo para predecir el bajo rendimiento académico en alumnos de la Universidad Nacional de Colombia, usan datos de los cuatro primeros períodos de estudio y aplicaron los algoritmos Naive Bayes y Decision Tree para la tarea de clasificación.

Por su parte, Merchan y Duarte (2016) construyeron un modelo predictivo de desempeño académico usando datos geográficos y académicos de estudiantes del programa de Ingeniería en Sistemas de la Universidad del Bosque en Bogotá, Colombia; para aplicar varias técnicas de clasificación de la MDE (J48, Part y Ridor); las precisiones y reglas obtenidas son diferentes para cada algoritmo, aunque el clasificador estrato social es común; observaron que los resultados obtenidos están altamente relacionados con las características de la institución y su contexto.

Menacho (2017) aplicó las técnicas de MDE de regresión logística, árboles de decisión, redes bayesianas y redes neuronales, usando datos escolares de estudiantes matriculados en el curso de Estadística General de la Universidad Nacional Agraria La Molina (Lima, Perú), para predecir la clasificación final (desaprobado o aprobado); el algoritmo de redes bayesianas obtuvo la mejor precisión; sugiere considerar datos socio económicos para mejorar la precisión del modelo.

El trabajo de Aziz, Hafieza y Ahmad (2014), utilizó la MDE para determinar el desempeño académico (pobre, promedio, bueno) de alumnos de primer semestre de una carrera en Ciencias de la Computación en Malasia. Los métodos para generar los modelos predictivos fueron Naive Bayes, Rule Based, y Decision Tree. Utilizaron 5 variables relacionadas con aspectos socioeconómicos y, como variable dependiente, el

promedio de calificaciones en primer semestre (en inglés, *Grade Point Average,* GPA). En las pruebas de validación emplearon variantes con diferentes porcentajes de valores de entrenamiento y validaciones cruzadas, con un 68.8% de exactitud como valor más alto de precisión.

En el caso de Miguéis, Freitas, Garcia y Silva (2018), se generaron modelos predictivos de clasificación temprana de estudiantes universitarios (Portugal), según su potencial rendimiento académico. Los resultados empíricos indicaron que el algoritmo Random Forest fue mejor al obtener un 95% de exactitud en las predicciones. Sin embargo, los algoritmos Decision Trees, Support Vector Machines, Bagged Trees y Boosted Trees, obtuvieron altos valores de exactitud, por lo que los modelos predictivos obtenidos también son efectivos para su aplicación en el contexto académico. Los atributos con mayor peso predictivo fueron el promedio de calificaciones de los exámenes de ingreso y el promedio de calificaciones al primer año. En menor medida tuvieron influencia la cantidad de créditos obtenidos en el primer año y el promedio de calificaciones en el nivel educativo previo, así como otros factores socioeconómicos. Se empleó este conjunto de variables para generar modelos confiables de predicción del rendimiento académico para el resto de la carrera.

Por su parte, Rico y Sánchez (2018) diseñaron y automatizaron un modelo predictivo del rendimiento académico de estudiantes del Instituto Politécnico Nacional (IPN), México. El modelo utilizó las calificaciones de cinco actividades académicas y la calificación final; también se construyó una plataforma que facilita la implementación del modelo para predecir automáticamente el desempeño académico de nuevos estudiantes, para ello se programó el algoritmo de Naive Bayes.

Para la creación de un sistema de detección temprana de alumnos en riesgo, Berens, Schneider, Görtz, Oster y Burghoff (2019), utilizaron el meta-algoritmo AdaBoost para combinar análisis de regresión, redes neuronales y árboles de decisión en la predicción de abandono escolar de alumnos de universidades de Alemania. La precisión de la predicción, usando datos demográficos y calificaciones disponibles al final del primer semestre, fue del 67%, y aumento a 80% en el cuarto semestre. Los resultados permiten identificar estudiantes en riesgo y ubicarlos en programas de apoyo en las universidades, además, de servir de punto de referencia para probar la efectividad de las intervenciones educativas.

La investigación de Imran, Latif, Mehmood y Shah (2019), desarrolló modelos predictivos mediante técnicas de clasificación supervisada con el objeto de determinar si los estudiantes aprueban o no un curso. Utilizaron calificaciones históricas de estudiantes, datos demográficos, sociales y atributos relacionados con la escuela. Se evaluaron los clasificadores J48, NNge y MLP, mediante técnicas de validación cruzada, el algoritmo J48 obtuvo la mejor precisión con un 90.2%. También emplearon el balanceo de la clase mediante la técnica conocida como *re-sampling* para mejorar el desempeño.

En el estudio de Buenaño-Fernández, Gil y Luján-Mora (2019), utilizaron técnicas de aprendizaje máquina para predecir la calificación final, basados en las calificaciones históricas de evaluaciones parciales, con diferentes pesos. Los sujetos de estudio fueron estudiantes de una carrera de ingeniería en sistemas computacionales en una universidad de Ecuador. Generaron árboles de decisión, mediante el algoritmo J48 implementado en el software WEKA. Los resultados muestran una precisión máxima de 96.5%, para predecir si un estudiante aprueba o no la asignatura.

De acuerdo con los trabajos revisados, las técnicas de MDE más utilizadas para la predicción del desempeño y riesgo académico, caracterización de los estudiantes, bajo desempeño y deserción, se ubican dentro de los métodos de regresión y clasificación; pero también se han utilizado métodos de agrupamiento y reglas de asociación.

El uso de los métodos de minería de datos ha permitido modelar la forma como los elementos del entorno y diversos factores académicos y no académicos tienen influencia en el rendimiento escolar y sus problemáticas asociadas. Sin embargo, sigue siendo necesario realizar estudios en contextos particulares dado que la mayoría de los resultados no son generalizables a todas las instituciones educativas.

3. Metodología

Considerando la naturaleza de los datos disponibles y su temporalidad, se determinó que para conducir este estudio, la investigación debía seguir un enfoque cuantitativo de corte no experimental o *ex post facto*, dado que la selección de los sujetos fue posterior a la realización del fenómeno y se buscó entender o reconstruir las posibles causas que lo ocasionaron, sin posibilidad de manipular las variables involucradas (Ballester, Nadal y Amer, 2017).

Este tipo de investigación se caracteriza por ser de carácter retrospectivo, partiendo del efecto a la determinación de las causas. El diseño y desarrollo del estudio se trata de recuperar y analizar información histórica, previamente registrada, con base en características relevantes de los datos obtenidos y utilizando variables bien identificadas (Valenzuela y Flores, 2012).

Específicamente, el desarrollo de la investigación se sustenta en el proceso de la minería de datos educativos para generar modelos predictivos de rendimiento académico y riesgo escolar, que posteriormente serán implementados en un sistema informático.

Se han formulado varios esquemas para aplicar la MDE basados en la metodología conocida como KDD (*Knowledge Discovery in Databases*) o Proceso de Extracción de Conocimiento en Bases de Datos (Márquez-Vera, Romero y Ventura, 2012; Peña-Ayala, 2014). Una propuesta del proceso de la minería de datos educativos se muestra en la Figura 1. Los pasos indicados permiten descubrir conocimiento que da respuesta a cuestionamientos concretos de los entornos educativos, lo que proporciona elementos para la toma de decisiones, además, en cada ciclo se van refinando los criterios para obtener mejores resultados.

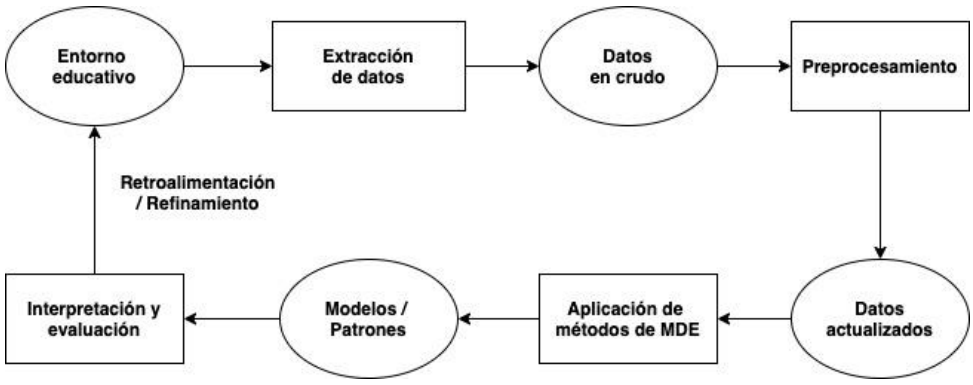

Figura 1. Proceso de la minería de datos educativa.

Fuente: Adaptación a partir de Alyahyan y Düştegör (2020) y Anoopkumar y Rahman (2016)

Las tareas por realizar en el proceso de la MDE se describen a continuación (Anoopkumar y Rahman, 2016; Merchan y Duarte, 2016; Peña-Ayala, 2014).

1) Extracción de datos: consiste en la selección de los datos de estudio del entorno educativo y su recuperación. Las fuentes de datos pueden ser heterogéneas: clases tradicionales, cursos *e-learning*, sistemas basados en web, redes sociales, entre otras. Se da preferencia a datos en formato digital provenientes de sistemas de información o entornos virtuales de aprendizaje, por conveniencia. Los datos pueden quedar agrupados en bases de datos en un formato crudo, es decir, sin ningún tratamiento aún. La información recuperada puede corresponder a la interacción y uso de los estudiantes en los sistemas, información de cursos, datos académicos, datos provenientes de encuestas socioeconómicas, antecedentes escolares, entre otros.

2) Preprocesamiento: una vez almacenados los datos en crudo, se limpian o eliminan los elementos con valores redundantes, incorrectos o faltantes, se fusionan datos provenientes de múltiples fuentes y se convierten valores de datos a formas más apropiadas para su manejo o análisis, es decir, se validan los datos para un manejo estandarizado y homogéneo. Esta etapa, por su complejidad, suele necesitar la aplicación de técnicas específicas para obtener datos correctamente estructurados. Es recomendable excluir variables o atributos irrelevantes, esto es, atributos que no aportan información útil para resolver el problema. Es posible que se requiera convertir los archivos a formatos específicos del sistema de minería de datos a utilizar. El resultado de esta etapa es un conjunto de datos actualizado y listo para su análisis.

3) Aplicación de métodos de MDE: consiste en la selección y aplicación de las técnicas más adecuadas acorde a las necesidades o hipótesis del estudio. Por los objetivos perseguidos en este trabajo, se emplearán métodos de clasificación/predicción para construir modelos que identifiquen casos de riesgo académico de los alumnos. Una vez cargados los datos, es posible seleccionar atributos para ser considerados en el análisis, según su relevancia. Para ello, pueden emplearse filtros que facilitan la elección de atributos concretos, además de la definición de los parámetros de funcionamiento de los algoritmos, con lo que se puede ir refinando la precisión de los modelos generados.

4) Interpretación de resultados: la salida resultante de los algoritmos se muestra para su interpretación, los modelos muestran los hallazgos de patrones, relaciones o tendencias en los datos. Se analizan los modelos con los mejores resultados para determinar el riesgo escolar de los estudiantes. También se pueden analizar los valores generados en la medición de la precisión de los resultados y utilizar métodos de validación. Dependiendo de la efectividad o validez de las reglas y modelos, pueden requerirse más iteraciones para refinar el funcionamiento de los algoritmos (incluir otros atributos, remover atributos "ruidosos" o que no tienen peso en el modelo, aplicar filtros, etc.).

5) Retroalimentación/Refinamiento: el conocimiento obtenido es informado a los actores del entorno educativo, a su vez, se pueden incluir recomendaciones; con lo cual es posible planificar acciones para modificar o mejorar procesos o condiciones del contexto escolar. También, en este proceso iterativo, se pueden ajustar diversos elementos para obtener modelos más confiables, al incorporar nuevas variables o datos

del entorno educativo, e incluir información de nuevos grupos de estudiantes, con lo cual se pueden mejorar los modelos predictivos.

A continuación, siguiendo las etapas previamente explicadas de la minería de datos educativos, se describen las actividades realizadas en un caso de estudio para determinar modelos predictivos de rendimiento escolar los cuales tienen el propósito de identificar alumnos en riesgo académico.

3.1 Extracción de datos

El estudio fue desarrollado en la Facultad de Matemáticas de la Universidad Autónoma de Yucatán. Esta institución educativa cuenta con seis programas de estudio a nivel licenciatura. Para propósitos de la investigación nos enfocamos exclusivamente en tres de los programas correspondientes al área de computación. Los programas de esta área son: Licenciatura en Ingeniería en Computación (LIC), Licenciatura en Ciencias de la Computación (LCC) y Licenciatura en Ingeniería de Software (LIS).

Para iniciar el proceso de generación de modelos predictivos, el primer paso consistió en ubicar los datos institucionales disponibles, los cuales pudieran aportar información relacionada con factores asociados al rendimiento académico de los alumnos de interés para el estudio. Fue seleccionada información académica, sociodemográfica y de resultados de exámenes de ingreso de las fuentes indicadas a continuación:

- Calificaciones de las asignaturas de los alumnos en su primer semestre de estudios, las cuales se recuperaron de los registros del Departamento de Control Escolar a partir de 2016, pues fue el año en que los programas LCC y LIS, comenzaron su operación con los nuevos planes de estudio basados en Modelo Educativo para la Formación Integral (MEFI) de la UADY; el cual considera una calificación de 70 como la mínima aprobatoria, en cambio en el modelo anterior la calificación mínima era de 60. Por tanto, los datos recabados corresponden al período de 2016 a 2019. Esta información fue proporcionada en 10 documentos de hojas de cálculo Excel, por lo que fue necesario homologar su estructura y consolidar los datos en un solo archivo.
- Los resultados del examen de ingreso Exani-II de los alumnos admitidos a los programas de estudios de las carreras del área de computación, entre 2016 y 2019. Esta información la proporcionó la Secretaría Académica mediante la entrega de 5 archivos en formato XLSX.
- Datos sociodemográficos de los alumnos de las carreras del área de computación. Esta información fue proporcionada por la Secretaría Académica a través de 4 archivos en formato XLSL.

Toda la información se recolectó en un momento único, cuando a la ocurrencia del fenómeno de estudio y los factores que influyeron en sus resultados ya habían sucedido. Al momento de consolidar la información, proveniente de todas las fuentes identificadas, se tuvo que hacer coincidir cada uno de los atributos o variables con el correspondiente alumno, mediante operaciones de ordenamiento, filtrado y copiado de datos. Finalmente, se obtuvo una sola base de datos almacenada en un archivo en formato XLSX. El total de instancias obtenidas corresponden a 415 alumnos, 90 de la Licenciatura en Ingeniería en Computación, 101 de la Licenciatura en Ciencias de la Computación y 224 de la Licenciatura en Ingeniería de Software. Cada uno de los registros de alumnos contaba con 65 atributos.

La confidencialidad de los datos de los estudiantes es un aspecto importante para el desarrollo del proyecto. En este sentido, los datos personales recuperados fueron manejados exclusivamente por el responsable del proyecto, quien también fue el único autorizado para mantener en resguardo las copias de los archivos de bases de datos. A partir de la obtención y consolidación de los datos en un solo archivo, también se eliminaron los atributos de nombres o matrículas escolares, con lo cual se evitó la identificación de los alumnos, de esta forma se garantiza el anonimato en la manipulación de la información en el proceso de análisis. Los modelos generados presentan información de forma genérica y sin mostrar datos de forma individual.

3.2 Preprocesamiento

Una vez unificados los datos, el preprocesamiento resultó complejo y tardado, pero se trató de una tarea fundamental para poder hacer un análisis más preciso de los datos. Se trabajó en la hoja de cálculo empleando las funciones disponibles para la edición de los datos. De los 65 atributos disponibles, mediante un proceso de limpieza y eliminación de atributos no relevantes para el estudio, fueron retirados manualmente de la base de datos los campos cuya información fue identificada fácilmente como irrelevante o poco útil. También, 5 variables fueron agregadas, derivadas a partir de otros atributos, por ejemplo, a partir de la fecha de nacimiento se obtuvo la edad del estudiante, a partir de las calificaciones de las materias se obtuvo el promedio y la cantidad de materias aprobadas, del nombre de la escuela de procedencia se obtuvo la variable Prepa, para indicar si estudió en preparatoria perteneciente a la UADY o no.

En este paso se destaca la importancia y necesidad de convertir, normalizar o discretizar las variables para obtener mejores resultados o para poder utilizar algoritmos que requieren ciertas características en el formato de las variables. Los atributos elegidos de los alumnos fueron aquellos que pudieran aportar información significativa para los propósitos del análisis. Un primer conjunto de variables se muestra en la Tabla 1.

Tabla 1.
Descripción del conjunto de variables inicial.

Nombre	Descripción
Carrera	Programa educativo elegido
Sexo	Sexo del estudiante
ICNE	Prueba competencias básicas (CB) del EXANI II
DIAG	Prueba de competencias disciplinares (CD) del EXANI II
Exani	Resultado global del examen de ingreso (CB 70% + CD 30%)
A_ing	Año de ingreso
N_insc	Número de materias inscritas
Edad	Edad al momento del ingreso
Mun	Municipio de procedencia
Edo	Estado de procedencia
Esc_Proc	Escuela de procedencia
Prepa	Preparatoria de procedencia (atributo derivado {UADY, No UADY})
E_Civil	Estado civil

N_Hijos	Número de hijos
S_Medico	Tipo de servicio médico
Resp	Responsable del alumno
Beca	Tipo de beca con la que cuenta
P_Aprob	Porcentaje de materias aprobadas
PROM	Promedio en primer semestre
RIESGO_A	Riesgo académico (P_Aprob < 0.8="SI", P_Aprob ≥ 0.8="NO")
RIESGO_P	Riesgo académico (PROM < 75="SI", PROM ≥ 75="NO")

Fuente: Elaboración propia.

Las variables dependientes RIESGO_A y RIESGO_P, se derivaron a partir de las variables PROM y P_Aprob. Fue necesario definirlas como variables nominales o categóricas (clases), ya que varios de los algoritmos utilizados la clasificación y generación de los modelos predictivos requieren este tipo de dato en la definición de las clases.

En el proceso de asignación de valores a la variable RIESGO_A, un valor de 0.8 o mayor del porcentaje de materias aprobadas (P_Aprob), se asignó como "NO" a la variable RIESGO_A, indicando que no hay riesgo académico, en caso contrario se asigna un "SI". Dado que en primer semestre regularmente se inscriben 5 materias, un valor P_Aprob menor a 0.8 indicaría que eventualmente reprobaría dos o más materias lo que se consideró como un riesgo académico alto para el alumno. De igual forma se procedió con la variable RIESGO_P, en cuyo caso se consideró un valor de PROM mayor o igual a 75 para indicar que" NO" hay riesgo académico, en caso contrario se asignaría un "SI" a la variable.

Por lo tanto, quedó definido que un promedio de calificación inferior a 75 o un porcentaje de materias aprobadas inferior a 0.8 son indicadores de un bajo rendimiento académico, y en consecuencia existe alta probabilidad de riesgo académico para el estudiante. Existen estudios en los que se adoptaron criterios similares para la determinación de perfiles de rendimiento académico (Baker et al., 2015; Martínez, Karanik, Giovannini y Pinto, 2015; Padua, 2019; Río-Jenaro, Calle, Martín y Robaina, 2018).

Para poder analizar los datos en el software WEKA, fue necesario convertir el archivo de base de datos del formato XLSX a un archivo de texto en formato CSV (valores separados por comas). Posteriormente, desde WEKA se leyó el archivo y se exportó al formato propio de esta herramienta, el cual utiliza archivos ARFF (formato de archivo de relación de atributos). De esta manera, todo quedó preparado para iniciar el proceso de análisis de la información.

3.3 Aplicación de métodos de MDE

Durante esta etapa se empleo el software de minería de datos WEKA para realizar tareas de selección de atributos, la aplicación de técnicas de minería de datos para la experimentación, refinamiento en los parámetros de ejecución y validación de los modelos predictivos generados.

3.3.1 Selección de variables significativas

A partir del conjunto inicial de variables, mostrado en la Tabla 1, se emplearon algunas de las técnicas integradas en el software WEKA para seleccionar un conjunto más reducido de variables significativas y que, posteriormente, se utilizan en el análisis y la generación de los modelos predictivos. Los resultados en la Tabla 2 corresponden a tres de los métodos de evaluación de atributos: *CfsSubsetEval*, *WrapperSubsetEval* con J48 y *WrapperSubsetEval* con NaiveBayes. La Tabla 3 contiene los resultados del método *CorrelationAttributeEval* para la selección de atributos.

Tabla 2.
Resultados de 3 métodos de selección de atributos

Atributo	CfsSubsetEval/ GreedyStepwise	WrapperSubsetEval/ J48/ BestFirst	WrapperSubsetEval/ NaiveBayes/ GreedyStepwise
Carrera *	0%	70%	80%
Sexo	0%	10%	50%
ICNE *	30%	0%	40%
DIAG *	100%	60%	0%
Exani *	100%	70%	80%
A_ing	0%	20%	50%
N_insc	0%	0%	20%
Edad *	0%	50%	40%
Mun *	80%	0%	10%
Edo	0%	20%	10%
Esc_Proc	20%	0%	10%
Prepa *	0%	30%	40%
E_Civil *	10%	40%	50%
N_Hijos	0%	10%	10%
S_Medico	0%	30%	10%
Resp *	0%	30%	60%
Beca	0%	10%	0%

Fuente: Elaboración propia.

El método *CfsSubsetEval* pondera de manera individual cada atributo, mientras que el método *WrapperSubsetEval* utiliza un algoritmo de clasificación para poder hacer la medición del rendimiento de los atributos que se están evaluando (ver Tabla 2). En el caso del método *CorrelationAttributeEval*, clasifica y ordena los atributos con base en la correlación que tienen con la clase (ver Tabla 3).

Tabla 3.
Resultados del cuarto método de selección de atributos

CorrelationAttributeEval / Ranker		
Mérito promedio	**Rango promedio**	**Atributo**
0.438 +- 0.014	1.2 +- 0.4	**5 Exani ***
0.43 +- 0.014	1.8 +- 0.4	**4 DIAG ***
0.407 +- 0.015	3 +- 0	**3 ICNE ***
0.247 +- 0.017	4 +- 0	**12 Prepa ***
0.172 +- 0.017	5 +- 0	**8 Edad ***
0.112 +- 0.009	6.5 +- 0.5	**1 Carrera ***
0.108 +- 0.015	6.9 +- 1.45	**16 Resp ***
0.088 +- 0.006	9 +- 0.63	11 Esc_Proc
0.088 +- 0.007	9.2 +- 1.54	7 N_insc

0.084 +- 0.01	9.8 +- 1.47	6 A_ing
0.082 +- 0.007	10 +- 0.89	9 Mun
0.058 +- 0.015	12.4 +- 1.28	15 S_Medico
0.06 +- 0.016	12.5 +- 0.92	10 Edo
0.032 +- 0.018	14.8 +- 1.4	13 E_Civil
0.023 +- 0.012	15.2 +- 0.98	17 Beca
0.02 +- 0.012	15.4 +- 1.11	2 Sexo
0.012 +- 0.008	16.3 +- 0.9	14 N_Hijos

Fuente: Elaboración propia.

Las variables seleccionadas en esta etapa son las indicadas con (*) en la Tabla 2 y Tabla 3. Los resultados de los diferentes métodos no coincidieron totalmente, ya que cada uno captura de forma diferente la información y la relevancia de las variables. Se optó por elegir un subconjunto considerando los atributos en los que había un valor de relevancia regular o alto para dos o más métodos, procurando de esta manera no eliminar aún atributos que pudieran ser relevantes en las predicciones.

El subconjunto de variables independientes o predictoras quedó definido como: *Carrera* (programa educativo elegido), *Exani* (Examen Nacional de Ingreso, que muestra el resultado global ponderado ICNE+DIAG), *ICNE* (Índice Ceneval, que refleja el resultado del examen Exani-II Admisión), *DIAG* (resultado del examen Exani-II Diagnóstico, para ingenierías y tecnología), *Edad* (edad al momento del ingreso), *Mun* (municipio de procedencia), *Prepa* (escuela preparatoria de procedencia), *E_Civil* (estado civil), *Resp* (responsable económico del estudiante). Las variables dependientes o clases son: *RIESGO_A* (riesgo académico determinado por la relación de materias aprobadas) y *RIESGO_P* (riesgo académico determinado por el promedio).

3.3.2 Aplicación de métodos de predicción/clasificación

Las técnicas de minería de datos educativas se emplearon considerando el subconjunto de variables determinado previamente. Por cada técnica utilizada y para cada variable dependiente se obtuvo un modelo predictivo y los resultados de la precisión y rendimiento del clasificador. En todos los casos se utilizó el conjunto de datos completo y el proceso para evaluar los resultados consistió en emplear validación cruzada con 10 iteraciones. Varios de los algoritmos de clasificación implementados en WEKA fueron utilizados. Después de varias pruebas se eligieron los que presentaron un mejor rendimiento y precisión, los cuales fueron: J48, RandomForest, LMT *(Logistic Model Trees*), Logistic y MultilayerPerceptron. Los modelos resultantes y la evaluación de su precisión se presentan en la sección de resultados.

4. Resultados

Los modelos obtenidos para cada algoritmo incluyen dos casos que corresponden a cada variable clasificadora: RIESGO_P, basada en el promedio de calificaciones y RIESGO_A, basada en el porcentaje de materias aprobadas. Ambas clases pueden tener dos estados finales del estudiante: 1) el estudiante esta en riesgo académico (valor de la clase = “SI”) y 2) el estudiante no esta en riesgo académico (valor de la clase = “NO”). A continuación, se describen los algoritmos empleados y los modelos generados con ellos.

4.1 Algoritmo J48

Un árbol de decisión es un método de aprendizaje supervisado que construye un árbol a partir de un conjunto de datos históricos (datos de entrenamiento), encontrando patrones predictivos inductivamente. Cada rama del árbol es una posible explicación al problema de clasificación o predicción planteado (reglas de decisión o clases). Cada regla se interpreta leyendo la secuencia de nodos y sus condiciones a partir de la raíz y hasta llegar al nodo hoja (Peña-Ayala, 2014).

El algoritmo consiste en calcular una razón de ganancia para cada variable, basado en un valor de entropía o incertidumbre; a menor incertidumbre más información aporta el atributo para explicar la clasificación. Se seleccionan los mejores atributos predictores para ir creando nodos de forma descendente y recursiva; los nodos más próximos a la raíz son los mejores predictores, y a partir de ellos el proceso se repite para generar ramas en el árbol con nuevos nodos a partir de los datos restantes. Si no es posible encontrar más predictores el árbol deja de crecer. Puede utilizar atributos nominales o continuos, ya que los valores continuos los convierte automáticamente a intervalos discretos. La variable dependiente debe ser de tipo nominal. Esta técnica permite procesar gran cantidad de variables, y muchas de las variantes de árboles son completamente no-paramétricas, es decir, no se requiere una distribución de población con características específicas (Menacho, 2017).

El método J48 de clasificación estadística es la implementación en WEKA del algoritmo C.45 (Quinlan, 1993). Ha sido uno de los más utilizados por su sencillez al momento de interpretar las reglas predictivas de sus modelos. Se puede usar tanto para clasificar o identificar factores determinantes en una población, como para predecir futuras instancias.

La Tabla 4 muestra las reglas de conocimiento obtenidas a partir del árbol J48 para la clase RIESGO_P y su interpretación en lenguaje natural. Por la dimensión del árbol y para el propósito de identificar condiciones de riesgo académico, se presentan las reglas que indican los casos en que la variable dependiente tomará un valor de “SI”, es decir, se predice la existencia de riesgo académico. Lo que significa que los alumnos podrían obtener un promedio de calificación inferior a 75 en el semestre.

Tabla 4.
Reglas de conocimiento derivadas del árbol de decisión J48 para la clase RIESGO_P

#	Regla de conocimiento	Interpretación en lenguaje natural	Ocurrencias (aciertos/errores)
1	DIAG <= 65 & DIAG <= 44: SI	Si el valor de DIAG es menor o igual que 44, entonces, SI hay riesgo académico.	(32.0)
2	DIAG <= 65 & DIAG > 44 & Carrera = LCC & Prepa = No UADY: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LCC y Prepa es igual a No UADY, entonces, SI hay riesgo académico.	(32.0/5.0)
3	DIAG <= 65 & DIAG > 44 & Carrera = LCC & Prepa = UADY & Resp = MADRE: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LCC y Prepa es igual a UADY y la responsable es la Madre, entonces SI hay riesgo académico.	(6.0/1.0)
4	DIAG <= 65 & DIAG > 44 & Carrera = LCC & Prepa = UADY & Resp = PADRE & ICNE <= 1180: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LCC y Prepa es igual a UADY y el responsable es el Padre y el ICNE es menor o igual a 1180, entonces, SI hay riesgo académico.	(9.0/2.0)
5	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = MADRE & DIAG <= 53: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y la responsable es la Madre y DIAG es menor o igual a 53, entonces, SI hay riesgo académico.	(7.0)
6	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = MADRE & DIAG > 53 & Prepa = UADY & ICNE <=1216: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y la responsable es la Madre y DIAG es mayor a 53 y Prepa es UADY y el ICNE es menor o igual a 1216, entonces, SI hay riesgo académico.	(4.0)
7	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = PADRE & ICNE <= 1150: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y el responsable es el Padre y el ICNE es menor o igual a 1150, entonces, SI hay riesgo académico.	(17.0/3.0)
8	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = PADRE & ICNE > 1150 & Prepa = No UADY & ICNE <=1198: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y el responsable es el Padre y el ICNE es mayor a 1150, pero menor o igual a 1198, y Prepa es No UADY, entonces, SI hay riesgo académico.	(27.0/9.0)
9	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = PADRE & ICNE > 1150 & Prepa = UADY & ICNE > 1174: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y el responsable es el Padre y el ICNE es mayor a 1174 y Prepa es UADY, entonces, SI hay riesgo académico.	(18.0/7.0)
10	DIAG <= 65 & DIAG > 44 & Carrera = LIS & Resp = UNO MISMO: SI	Si el valor de DIAG es mayor a 44 y menor o igual a 65 y Carrera es igual a LIS y el responsable es Uno mismo, entonces, SI hay riesgo académico.	SI (8.0/1.0)

Fuente: Elaboración propia.

El atributo DIAG es el más significativo, ya que a partir de él se derivaron todas las reglas de clasificación o predicción listadas en la Tabla 4. La columna de ocurrencias muestra la cantidad de casos que se presentaron para cada regla, tanto aciertos como errores en la clasificación, lo cual puede servir como indicador para detectar los atributos y condiciones más frecuentes que determinan situaciones de riesgo académico de los alumnos.

De forma similar, en la Tabla 5 se muestran las reglas de conocimiento obtenidas a partir del árbol J48 para la clase RIESGO_A y su interpretación en lenguaje natural. Para este caso, se observa que el atributo ICNE tiene el mayor peso predictivo, pues es el punto de partida para todas las reglas obtenidas.

Tabla 5.
Reglas de conocimiento derivadas del árbol de decisión J48 para la clase RIESGO_A

#	Regla de conocimiento	Interpretación en lenguaje natural	Ocurrencias
1	ICNE <= 1216 & DIAG <= 62 & Carrera = LIC & Prepa = No UADY: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es menor o igual que 62 y la Carrera es igual a LIC y Prepa es igual a No UADY, entonces, SI hay riesgo académico.	(40.0/15.0)
2	ICNE <= 1216 & DIAG <= 62 & Carrera = LIC & Prepa = UADY & Edad > 21: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es menor o igual que 62 y la Carrera es igual a LIC y Prepa es igual a UADY y Edad es mayor a 21, entonces, SI hay riesgo académico.	(5.0)
3	ICNE <= 1216 & DIAG <= 62 & Carrera = LCC & Prepa = No UADY: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es menor o igual que 62 y la Carrera es igual a LCC y Prepa es igual a No UADY, entonces, SI hay riesgo académico.	(49.0/6.0)
4	ICNE <= 1216 & DIAG <= 62 & Carrera = LCC & Prepa = UADY & Exani <= 781.25: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es menor o igual que 62 y la Carrera es igual a LCC y Prepa es igual a UADY y Exani es menor o igual a 781.25, entonces, SI hay riesgo académico.	(11.0/1.0)
5	ICNE <= 1216 & DIAG <= 62 & Carrera = LIS: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es menor o igual que 62 y la Carrera es igual a LIS, entonces, SI hay riesgo académico.	(88.0/25.0)
6	ICNE <= 1216 & DIAG > 62 & Prepa = No UADY & Edad <= 23 & Exani <= 803.25: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es mayor 62 y la Prepa es igual a No UADY y Edad es menor o igual a 23 y Exani es menor o igual a 803.25, entonces, SI hay riesgo académico.	(10.0/2.0)
7	ICNE <= 1216 & DIAG > 62 & Prepa = No UADY & Edad <= 23 & Exani > 803.25 & ICNE > 1192 & Edad <=20 & Edad > 18: SI	Si el valor de ICNE está entre 1193 y 1216 y DIAG es mayor 62 y la Prepa es igual a No UADY y Exani es mayor a 803.25 y Edad está entre 18 y 20 años, entonces, SI hay riesgo académico.	(6.0/1.0)
8	ICNE <= 1216 & DIAG > 62 & Prepa = No UADY & Edad > 23: SI	Si el valor de ICNE es menor o igual que 1216 y DIAG es mayor 62 y la Prepa es igual a No UADY y Edad mayor a 23, SI hay riesgo académico.	(4.0)
9	ICNE > 1216 & Edad > 25: SI	Si el valor de ICNE es mayor a 1216 y Edad es mayor a 25, SI hay riesgo académico.	(4.0)

Fuente: Elaboración propia.

Hay en total nueve reglas de decisión que devuelven un resultado igual a “SI”, indicando el riesgo de reprobar dos o más asignaturas en el semestre.

4.2 Algoritmo RandomForest

Los bosques aleatorios son una combinación de predictores de árboles de decisión. El método genera múltiples versiones del árbol tomando muestras aleatorias de los datos de entrenamiento y promediando estas para obtener un mejor clasificador (Breiman, 2001). Es posible que se genere una gran cantidad de árboles, incluso similares. Por ello, el algoritmo escoge un subconjunto de variables predictivas antes de dividir cada nodo, para obtener árboles diferentes no correlacionados del conjunto de datos. Para construir el modelo final, cada árbol individual en el bosque aleatorio proporciona una predicción de clase y la clase con más votos se convierte en la predicción del modelo.

El modelo resultante fue generado utilizando 100 iteraciones con parámetros de aprendizaje por defecto. La visualización del bosque de árboles no es posible debido a la gran cantidad de árboles obtenidos, sin embargo, WEKA tiene la opción de salvar en archivo los modelos, lo cual se realizó para los resultados conseguidos con las dos variables de clase RIESGO_A y RIESGO_P. Los archivos con los modelos pueden ser

cargados en WEKA o ser utilizados desde programas computacionales, que incluyan las bibliotecas apropiadas, para realizar la predicción de nuevos casos de alumnos.

4.3 Algoritmo LMT

Este algoritmo construye clasificadores de Árboles de Modelo Logístico (en inglés, *Logistic Model Trees*, LMT), que son árboles de clasificación con funciones de regresión logística en las hojas. El resultado de combinar modelos de estructuras de árboles y la regresión logística produce un árbol con estimaciones de probabilidad de clase explícitas en lugar de sólo clasificar (Landwehr, Hall y Frank, 2003).

LMT produce un solo árbol con divisiones binarias en atributos numéricos, divisiones de múltiples vías en variables nominales y modelos de regresión logística en las hojas. A pesar de no ser tan sencilla su interpretación, es más entendible que los resultados producidos por otros métodos, como en el caso de modelos con múltiples árboles. También el algoritmo se asegura de dejar sólo los atributos más relevantes en el modelo.

En la Figura 2 se muestra el árbol LMT. El árbol tiene sólo dos nodos hoja o terminales, correspondientes a los dos posibles valores de la clase RIESGO_A. Cada valor de la clase se modela como una función de variables explicativas que determinan la probabilidad de un posible resultado.

```
Class SI :
3.49 +
[Carrera=LIC] * -0.55 +
[DIAG] * -0.02 +
[Exani] * -0 +
[Edad] * 0.06 +
[Prepa=UADY] * -0.38 +
[Resp=PADRE] * -0.13

Class NO :
-3.49 +
[Carrera=LIC] * 0.55 +
[DIAG] * 0.02 +
[Exani] * 0    +
[Edad] * -0.06 +
[Prepa=UADY] * 0.38 +
[Resp=PADRE] * 0.13
```

Figura 2. Árbol de modelo logístico para la clase RIESGO_A.
Fuente: Elaboración propia.

La Figura 3 muestra el árbol LMT resultante para la clase RIESGO_P. Se observan similitudes en entre los dos árboles, al considerar los mismos atributos como factores relevantes en la predicción, variando en los coeficientes de las funciones logísticas de los nodos terminales.

```
Class SI :
3.99 +
[Carrera=LIC] * -0.77 +
[DIAG] * -0.03 +
[Exani] * -0 +
[Edad] * 0.07 +
[Prepa=UADY] * -0.22 +
[Resp=MADRE] * 0.08 +
[Resp=UNO MISMO] * 0.4  +
[Resp=HERMANO(A)] * -0.66 +
```

```
[Resp=OTRO] * 0.56

Class NO :
-3.99 +
[Carrera=LIC] * 0.77 +
[DIAG] * 0.03 +
[Exani] * 0    +
[Edad] * -0.07 +
[Prepa=UADY] * 0.22 +
[Resp=MADRE] * -0.08 +
[Resp=UNO MISMO] * -0.4 +
[Resp=HERMANO(A)] * 0.66 +
[Resp=OTRO] * -0.56
```

Figura 3. Árbol de modelo logístico para de la clase RIESGO_P.
Fuente: Elaboración propia.

El cálculo de la predicción se realiza sumando los elementos de la función. Cuando hay algún atributo nominal y se cumple la igualdad, por ejemplo [Carrera=LIC], entonces el valor de este término es igual a uno, o cero en caso contrario, y se multiplica por el valor indicado del peso de la variable. Si se trata de un atributo numérico, el valor de la variable se multiplica directamente por el peso del atributo. Cuando la suma de todos los términos de la función es cero la clase toma el valor "SI", que en nuestro modelo significa un resultado de riesgo académico verdadero.

4.4 Algoritmo Logistic

Los modelos de regresión logística lineal permiten predecir el resultado de una variable dependiente categórica. La regresión logística es un instrumento estadístico de análisis multivariado. Su propósito es predecir la probabilidad de que ocurra un evento (Le Cessie y Van Houwelingen, 1992).

En la Figura 4 se muestra el modelo logístico para la clase o variable dependiente RIESGO_A.

```
Coefficients...
                         Class
Variable                  SI
==============================
Carrera=LIC              -0.915
Carrera=LCC               0.2996
Carrera=LIS               0.4034
ICNE                     -0.0046
DIAG                     -0.0699
Exani                    -0.0044
Edad                      0.141
Prepa=UADY               -0.5613
Resp=MADRE                0.3755
Resp=PADRE                0.2016
Resp=UNO MISMO            0.8621
Resp=HERMANO(A)           0.4544
Resp=TIO(A)             -74.7232
Resp=OTRO                 1.3282
Intercept                10.1386

Odds Ratios...
                         Class
Variable                  SI
==============================
Carrera=LIC               0.4005
Carrera=LCC               1.3494
Carrera=LIS               1.497
ICNE                      0.9954
```

DIAG	0.9325
Exani	0.9956
Edad	1.1514
Prepa=UADY	0.5705
Resp=MADRE	1.4558
Resp=PADRE	1.2234
Resp=UNO MISMO	2.3682
Resp=HERMANO(A)	1.5752
Resp=TIO(A)	0
Resp=OTRO	3.7742

Figura 4. Coeficientes del modelo logístico para la clase RIESGO_A.
Fuente: Elaboración propia.

De igual forma, la Figura 5 muestra las ponderaciones de los atributos en el modelo logístico para la clase RIESGO_P.

Coefficients...

Variable	Class SI
Carrera=LIC	-1.1655
Carrera=LCC	0.4602
Carrera=LIS	0.4556
ICNE	-0.005
DIAG	-0.0777
Exani	-0.0049
Edad	0.1472
Prepa=UADY	-0.4098
Resp=MADRE	0.4287
Resp=PADRE	0.2634
Resp=UNO MISMO	1.1287
Resp=HERMANO(A)	-1.4086
Resp=TIO(A)	-78.285
Resp=OTRO	1.6384
Intercept	11.135

Odds Ratios...

Variable	Class SI
Carrera=LIC	0.3118
Carrera=LCC	1.5845
Carrera=LIS	1.5772
ICNE	0.995
DIAG	0.9253
Exani	0.9951
Edad	1.1586
Prepa=UADY	0.6638
Resp=MADRE	1.5353
Resp=PADRE	1.3014
Resp=UNO MISMO	3.0916
Resp=HERMANO(A)	0.2445
Resp=TIO(A)	0
Resp=OTRO	5.1468

Figura 5. Coeficientes del modelo logístico para la clase RIESGO_P.
Fuente: Elaboración propia.

Los coeficientes son las ponderaciones que se aplican a cada atributo antes de sumarlos. El resultado de la suma es la probabilidad de que la nueva instancia pertenezca a la clase. En los modelos presentados, un valor mayor 0.5 significa que las variables dependientes tendrán un valor igual a “SI”, lo que indica una predicción positiva de

riesgo académico. Las razones de probabilidades (*odds ratios*) indican qué tanta influencia o efecto tendrá un cambio de ese valor en la predicción.

4.5 Algoritmo MultilayerPerceptron

Se generó un modelo basado en red neuronal con propagación hacia atrás conocido como Perceptrón Multicapa (en inglés, *Multilayer Perceptron*, MLP). Este tipo de algoritmo emplea aprendizaje supervisado con el conjunto de entrenamiento y construye la red neuronal en una serie de niveles o capas, que determinan la clasificación de instancias. Deben existir al menos tres capas que son: capa de entrada, capa de procesamiento u oculta, y capa de salida. En cada capa puede existir un número variable de nodos o neuronas interconectadas que ejecutan funciones no lineales comúnmente de tipo sigmoidal o logística, en las que sea cual sea la entrada la salida estará comprendida entre 0 y 1. Cuando una neurona recibe una entrada, se activa y genera una salida que transmite a la siguiente capa, el valor transmitido dependerá del peso asignado a cada conexión (Mitra y Pal, 1995).

Los modelos para las clases RIESGO_A y RIESGO_P, tienen 10 nodos cada una, con los pesos de predicción que les fueron asignados en el proceso de aprendizaje. Un inconveniente de las redes neuronales es la dificultad de comprender los modelos que generan. En la Figura 6 se muestras 3 de los 10 nodos del modelo MLP obtenido.

```
Sigmoid Node 0
    Inputs    Weights
    Threshold    -1.3064491512227503
    Node 2    6.096383309821736
    Node 3    -6.219122630052431
    Node 4    -5.484905886307731
    Node 5    1.839999062521567
    Node 6    1.6192370537262368
    Node 7    3.8194015616049812
    Node 8    -2.921985664613799
    Node 9    1.8538491040214233
Sigmoid Node 1
    Inputs    Weights
    Threshold    1.3064565421745287
    Node 2    -6.096388571581784
    Node 3    6.219104407340171
    Node 4    5.484672734189203
    Node 5    -1.8400167841509871
    Node 6    -1.6198128315996037
    Node 7    -3.8189497476356724
    Node 8    2.9224180550225483
    Node 9    -1.8539339831271278
Sigmoid Node 2
    Inputs    Weights
    Threshold    0.31547768643364943
    Attrib Carrera=LIC    4.483306276354068
    Attrib Carrera=LCC    -2.7535846098268304
    Attrib Carrera=LIS    -2.0121033202057395
    Attrib ICNE    0.18408561410797677
    Attrib DIAG    -5.689743762265199
    Attrib Exani    -2.1842360451003153
    Attrib Edad    9.322224538137549
    Attrib Prepa=UADY    -7.526943203504304
    Attrib Resp=MADRE    -0.7600297549616539
    Attrib Resp=PADRE    0.8530293486362223
    Attrib Resp=UNO MISMO    -5.12567774369935
    Attrib Resp=HERMANO(A)    1.7827187040088694
    Attrib Resp=TIO(A)    -0.4131480848228954
    Attrib Resp=OTRO    2.3742931048318376
```

Figura 6. Vista parcial del modelo MultilayerPerceptron para la clase RIESGO_P.
Fuente: Elaboración propia.

4.6 Resumen de resultados

En la Tabla 6 se muestra el resumen de los resultados obtenidos al calcular la precisión de los modelos. Se incluyen los valores de precisión y el área bajo la curva ROC para determinar la calidad de los clasificadores. Se empleo el método de validación cruzada con 10 iteraciones, utilizando las 415 instancias disponibles en la base de datos.

Tabla 6.
Resultados de la evaluación de los modelos los predictivos.

Algoritmo	RIESGO_A		RIESGO_P	
	Precisión	AUC ROC	Precisión	AUC ROC
J48	69.8795 %	0.728	71.5663 %	0.736
RandomForest	68.9157 %	0.770	73.012 %	0.796
LMT	**71.0843 %**	**0.782**	**75.4217 %**	**0.805**
Logistic	**70.8434 %**	**0.780**	**74.9398 %**	**0.807**
MultilayerPerceptron	65.5422 %	0.721	74.6988 %	0.765

Fuente: Elaboración propia.

Al comparar los resultados generales entre la clase RIESGO_A y RIESGO_P se observa que la variable dependiente RIESGO_P, definida a partir del promedio de calificaciones, tiene mejores valores en el indicador AUC ROC y también en los niveles de precisión. En la comparación de los valores de precisión y calidad de cada uno de los algoritmos, los mejores resultados fueron obtenidos por los modelos LMT y Logistic; interpretando los valores AUC ROC, de ambos modelos, se puede decir que son buenos clasificadores.

4.7 Modelo propuesto

El modelo por desarrollar es el resultado de identificar los principales elementos que resultan fundamentales para la predicción. En ese sentido, fue necesario analizar qué atributos aportan el mayor poder predictivo y determinar cuáles algoritmos o combinaciones de algoritmos son los adecuados para establecer un buen modelo que sea implementado en una herramienta de software.

El subconjunto de variables predictoras se redujo a 9 atributos, según se indica en la sección 3.3.1, con lo cual se facilita y agiliza la recuperación de datos, ya sea que se soliciten a los alumnos o se recuperen desde algún repositorio digital.

Después de analizar los resultados de los algoritmos de clasificación, elegimos a los modelos LMT y Logistic como los más adecuados para implementarse en la aplicación de software; ya que, para la variable dependiente RIESGO_P, obtuvieron los valores de precisión más altos y sus valores de AUC ROC los ubican en el nivel de Test bueno, según Minguillón et al. (2017).

También, como lo explican Witten, Frank y Hall (2011), la herramienta WEKA tiene la capacidad de colocar en archivos los modelos generados por los algoritmos. Luego, mediante lenguajes de programación, se pueden recuperar los modelos y crear instancias de métodos de clasificación para realizar predicciones con los nuevos datos.

Con los elementos señalados, se presenta en la Figura 7 la propuesta del modelo predictivo para la detección de alumnos en riesgo académico, el cual será desarrollado e implementado en etapas posteriores.

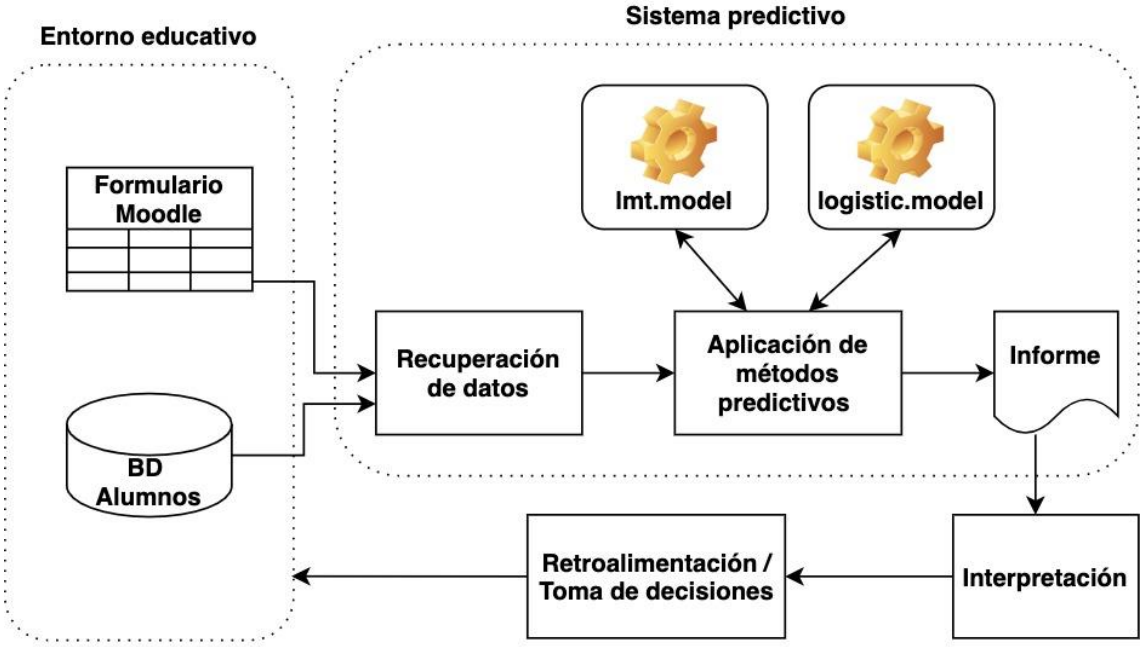

Figura 7. Propuesta de modelo predictivo de riesgo académico.

Fuente: Elaboración propia.

5. Discusión y conclusiones

5.1 Interpretación de resultados

En este apartado se comentan los resultados del proceso de selección de variables y algunas de las características de los modelos generados a partir de la aplicación de varias de las técnicas de clasificación/predicciones disponibles en el software WEKA.

De acuerdo con los valores de relevancia de los atributos disponibles (Tabla 2 y Tabla 3), el mejor método de selección de variables, para el conjunto de datos en estudio, es *CorrelationAttributeEval/Ranker*, ya que 7 de los 9 atributos seleccionados coinciden con sus resultados. Haciendo caso a los méritos calculados para los tributos, las mejores variables predictoras son: *ICNE, DIAG, Exani y Carrera*, las cuales tienen el mayor potencial predictivo, lo que se ve reflejado en los modelos obtenidos.

A pesar de no conseguir los mejores resultados, los modelos de árbol de decisión derivados del algoritmo J48, aportan mucha luz al entendimiento de los factores que influyen en el rendimiento académico, en especial para el propósito de este trabajo; ya permiten identificar con mucha claridad la serie de condiciones en las cuales los alumnos adquieren una condición de riesgo académico.

Considerando las reglas de conocimiento obtenidas por el algoritmo J48 para la clase *RIESGO_P*, como se observa en la Tabla 4, el atributo *DIAG* tiene el mayor peso como

variable predictora. En el siguiente nivel, considerando las reglas con más ocurrencias, vemos de forma general que, si el atributo *Carrera* es LCC o LIS, y la preparatoria de origen es igual a No UADY, entonces hay posible riesgo académico. También, hay condiciones en que el responsable económico o el *ICNE* pueden contribuir como factores de riesgo académico.

Las reglas de decisión presentadas en la Tabla 5**¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.**, fueron generadas por algoritmo J48, teniendo a la variable *RIESGO_A* como su clase. Se observa que los atributos *ICNE* y *DIAG* aparecen como los atributos de mayor relevancia como variables predictoras. En los casos con mayores ocurrencias, vemos que la variable *Carrera* es factor de riesgo cuando los alumnos son del programa LIS; o si son LIC o LCC y su preparatoria de origen es No UADY. Por otro lado, se detectan condiciones en los que a pesar de tener resultados aceptables en *DIAG*, *Exani* e *ICNE*, hay riesgo si *Prepa* es No UADY y la *Edad* del alumno es mayor a 18 años.

En general, en ambos árboles de decisión los resultados de las pruebas contenidas en el examen de ingreso Exani-II, tienen la mayor correlación respecto al rendimiento académico de los alumnos, ya sea que se considere la variable dependiente basada en el promedio de calificaciones o la basada en la cantidad de materias reprobadas. Hay coincidencia con los hallazgos de Miguéis et al. (2018), quienes determinaron que los atributos con mayor peso predictivo fueron promedios de calificaciones en exámenes de ingreso y calificaciones al primer año.

En menor medida se observan otros factores como la carrera elegida, lo que nos habla de diferencias en la conformación de los grupos pertenecientes a los diferentes programas; la preparatoria de origen, indicando una posible dificultad de los alumnos de otras escuelas para adaptarse al modelo empleado en la UADY; el responsable del alumno y su edad, que posiblemente estén relacionados con factores de su entorno familiar o problemáticas socioeconómicas.

Estos hallazgos son consistentes con el trabajo presentado por Montes y Lerner (2012), en donde se comenta que el rendimiento académico no sólo se explica por las calificaciones obtenidas, sino que existen otros aspectos relacionados con la dimensión académica, la económica, la familiar, la personal y la institucional, los cuales aportan a su comprensión; como también observaron Merchan y Duarte (2016).

Respecto a los modelos LMT mostrados en la Figura 2 y Figura 3, los atributos más relevantes son *Carrera, DIAG, Exani, Edad, Prepa* y *Resp*. Sin embargo, no es tan clara la interpretación del modelo como en el caso de los árboles de decisión, ya que el peso de los atributos en la función logística de los nodos terminales depende de si se trata de una variable nominal o numérica, y cuando es una variable numérica esta puede ser muy variada en el rango de sus valores, comparada contra otras variables numéricas del modelo. A pesar de lo anterior, estos modelos obtuvieron la mejor precisión de todos los algoritmos evaluados, por lo que se recomienda su empleo en la determinación del riesgo académico de los alumnos.

El modelo generado con el algoritmo Logistic, resulta muy similar al modelo LMT, ya que genera una función de regresión logística con la que se calcula la probabilidad de ocurrencia de un evento o valor de la clase. Los atributos relevantes son *Carrera, ICNE, DIAG, Exani, Edad, Prepa* y *Resp*. Este modelo es el segundo mejor en cuanto a

confiabilidad y resultados de predicción. Para este conjunto de variables las funciones logísticas parecen tener un buen desempeño.

El modelo RandomForest no se visualiza en los resultados que genera WEKA, debido a la gran cantidad de árboles obtenidos. Mientras que en el modelo MultilayerPerceptron (Figura 6) hay 10 nodos o neuronas, que contienen atributos y sus pesos, los cuales determinan el valor de la función sigmoidal. Se observa que las variables empleadas son las mismas que en el modelo Logistic. La interpretación del modelo es difícil pues habría que dar seguimiento a las conexiones entre capas y nodos de la red neuronal.

Al comparar los atributos y clases utilizados en otras investigaciones (Berens et al., 2019; Buenaño-Fernández et al., 2019; Imran et al., 2019; Rico y Sánchez, 2018) vemos que las variables utilizadas corresponden principalmente a calificaciones de cursos, y las predicciones van en el sentido de aprobar sólo un curso, además, los datos requeridos deben ser obtenidos cuando el curso ya ha sido iniciado, en ocasiones con poco margen para desplegar estrategias de intervención educativa. Por el contrario, en nuestra propuesta los datos requeridos son accesibles previo el comienzo del semestre y los resultados de la predicción resultan oportunos para ajustar el diseño de los cursos, informar a las instancias escolares responsables de la atención y apoyo estudiantil o desplegar estrategias adicionales de intervención.

Analizando la precisión de los mejores algoritmos e interpretando el indicador AUC ROC, podemos decir que los resultados alcanzados son suficientemente buenos para respaldar la pertinencia del empleo de los modelos en la predicción del riesgo académico de nuevas instancias de alumnos.

5.2 Conclusiones

En este trabajo se experimentó con varias técnicas de minería de datos educativas para generar modelos predictivos de rendimiento académico con un buen nivel de confiabilidad y exactitud, lo que permite identificar casos de alumnos en situación de riesgo académico, pertenecientes a grupos de nuevo ingreso de las carreras del área de computación de la Facultad de Matemáticas de la Universidad Autónoma de Yucatán.

Como se observó en la literatura consultada, la determinación del rendimiento académico es un problema multifactorial en donde confluyen una gran cantidad de variables, no sólo de orden académica o del entorno escolar, sino del contexto sociodemográfico o de aspectos cognitivos e interpersonales, entre otros.

La tarea de recopilación y procesamiento de datos resultó crucial para obtener información de calidad y atributos relevantes para ser utilizados en las tareas de la minería de datos educativas. Se analizaron diversas fuentes de datos y como resultado se obtuvo un archivo con 415 instancias de alumnos cada uno con 65 atributos. El método empleado para seleccionar variables permitió reducir a 9 la cantidad de atributos significativos asociados a los alumnos en estudio, con el consiguiente ahorro de espacio de almacenamiento, reducción de complejidad y la mejora en la calidad de los modelos generados.

La aplicación de los métodos de clasificación se realizó mediante el software WEKA y se utilizó la base de datos acotada al subconjunto de variables elegidas. También se consideraron dos casos de variables dependientes (clases) para determinar valores de predicción. La primera tomando el porcentaje de materias reprobadas, en el que la

reprobación de dos o más materias equivaldría a un valor afirmativo de riesgo académico (variable RIESGO_A). En la segunda opción, se determinó una predicción de riesgo cuando el promedio de calificaciones era inferior a 75 (variable RIESGO_P).

Se generaron modelos predictivos mediante 5 algoritmos de clasificación de la minería de datos educativa. Con cada uno de ellos se emplearon las dos opciones de variables de clase y se calcularon los valores de exactitud y predicción, para poder contrastar su rendimiento. De acuerdo con los resultados generales, la clase RIESGO_P obtuvo en todos los algoritmos los mejores valores de exactitud y confiabilidad. De igual forma el mejor algoritmo predictivo resultó LMT, con un nivel de instancias correctamente clasificadas de 71.08%, para la variable RIESGO_A y de 75.42% para la variable RIESGO_P y sus valores del área bajo la curva ROC mostraron un buen nivel de confiabilidad, con valores 0.782 y 0.805 respectivamente.

De los atributos seleccionados, en general observamos que el resultado de la prueba de competencias disciplinares *DIAG* y la prueba diagnóstica de competencias básicas *ICNE* tienen el mayor poder predictivo y determinan en buena medida el bajo rendimiento escolar o riesgo académico; le siguen en importancia: el resultado final del Exani-II, la carrera elegida, la preparatoria de origen, la edad y el responsable del alumno.

También se pudo ver la combinación de factores que delinean a aquellos alumnos con tendencia a reprobar más asignaturas u obtener promedios de calificaciones más bajos. En general, cuando los alumnos de nuevo ingreso obtienen un puntaje bajo en *DIAG* o *ICNE* y la preparatoria de origen no pertenece a la UADY presentan riesgo académico; o en algunos casos, aún cuando tienen puntajes regulares en las pruebas que conforman el Exani-II, pero provienen de preparatoria no UADY o dependen de alguien diferente a su padre o su edad es mayor a 18 años, también se configura un escenario de riesgo académico.

Se ha mostrado que las diversas técnicas de la MDE pueden emplearse de manera muy flexible y práctica para seleccionar subconjuntos de atributos significativos y con ellos aplicar algoritmos de clasificación. Los resultados en los niveles de precisión y confiabilidad de los algoritmos nos indican la posibilidad de emplear los modelos de manera efectiva, con el propósito de predecir el nivel de riesgo académicos de los estudiantes de nuevo ingreso los programas del área de computación.

El conocimiento adquirido con los modelos predictivos nos indica, por el lado académico, la necesidad de atender aspectos relacionados con los antecedentes académicos de los alumnos y tomar medidas para asegurar el entendimiento del modelo educativo de la UADY por todos los alumnos; pero, por otro lado, nos obliga a prestar atención en las condiciones socioeconómicas de los estudiantes, y considerar el despliegue de estrategias más efectivas de seguimiento y apoyo.

En trabajos futuros se continuará con la recopilación de datos para mejorar la precisión de los modelos predictivos, y se dará seguimiento a la implementación de los modelos mediante el desarrollo de un sistema informático, el cual será utilizado con las nuevas generaciones de estudiantes del área de computación para detectar de manera oportuna a estudiantes en situación de riesgo académico, lo cual dará la oportunidad de realizar intervenciones educativas que coadyuven a disminuir la problemática del bajo rendimiento académico.

Agradecimientos

Se agradece el apoyo recibido mediante el Programa para el Desarrollo Profesional Docente (PRODEP), a través de la beca número 511-6/2020-6858, de la Subsecretaria de Educación Superior, Secretaria de Educación Pública (SEP-SES, México). Un agradecimiento especial a la Secretaría Académica de la Facultad de Matemáticas de la UADY por las facilidades para acceder a los datos utilizados en el estudio.

Presentación del artículo: 15 de enero de 2021
Fecha de aprobación: 23 de marzo de 2021
Fecha de publicación: 30 de abril de 2021

Ayala Franco, E., López Martínez, R.E. y Menéndez Domínguez, V.H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. *RED. Revista de Educación a Distancia*, 21*(66).* https://doi.org/10.6018/red.463561

Financiación

Este trabajo no ha recibido ninguna subvención específica de los organismos de financiación en los sectores públicos, comerciales o sin fines de lucro.

Referencias

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37,* 13–49. https://doi.org/10.1016/j.tele.2019.01.007

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

Anoopkumar, M., & Rahman, A. M. J. (2016). A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, 122–133*. https://doi.org/10.1109/SAPIENCE.2016.7684113

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. Recuperado de: https://www.academiajournals.com/pubchetumal2020

Aziz, A. A., Hafieza, N., & Ahmad, I. (2014). First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms. *Proceeding of the International Conference on Artificial Intelligence and Computer Science(AICS 2014), (September), 100–109*.

Baker, R. S., & Inventado, P. S. (2014). Chapter 4 Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61–75). NewYork: Springer.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education e-Learning Courses. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015),* 150–155. Recuperado de: https://www.educationaldatamining.org/EDM2015/proceedings/full150-155.pdf

Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1(1),* 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23(1),* 537–553. https://doi.org/10.1007/s10639-017-9616-z

Ballester, L., Nadal, A., & Amer, J. (2017). Métodos y técnicas de investigación educativa (2 ed.). Palma: Ediciones UIB.

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining, 11(3)*, 1–41.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2018). WEKA Manual for Version 3-8-3. Hamilton, New Zealand: The University of Waikato. Recuperado de: https://user.eng.umd.edu/~austin/ence688p.d/handouts/WekaManual2018.pdf

Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32.

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability, 11(10),* 2833. https://doi.org/10.3390/su11102833

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256. https://doi.org/10.1016/J.CHB.2017.01.047

Dorio, I. (2017). La transición a la Universidad. El grado de maestro de Educación Infantil (Tesis Doctoral). Universitat de Barcelona, España. Recuperado de: http://diposit.ub.edu/dspace/handle/2445/109484

Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *HP Invent, 27*. https://doi.org/10.1.1.10.9777

García, D. (2015). Construcción de un Modelo para Determinar el Rendimiento Académico de los Estudiantes Basado en Learning Analytics (Análisis del Aprendizaje), mediante el Uso de Técnicas Multivariantes (Tesis Doctoral). Universidad de Sevilla, España. Recuperado de: https://idus.us.es/handle/11441/40436

García Gutiérrez, J. A. (2016). Comenzando con Weka : Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para la clase.

Madrid.

Gros, B. (2015). Retos y tendencias sobre el futuro de la investigación acerca del aprendizaje con tecnologías digitales. *Revista de Educación a Distancia (RED), (32)*. Recuperado de: https://revistas.um.es/red/article/view/233061

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (IJET), 14(14),* 92–104. https://doi.org/https://doi.org/10.3991/ijet.v14i14.10310

Kerlinger, F. N., & Lee, H. (2002). Investigación del comportamiento (4a ed.). México: McGraw-Hill.

Kumar, M., & Singh, A. J. (2017). Evaluation of Data Mining Techniques for Predicting Student's Performance. *International Journal of Modern Education and Computer Science, 8*, 25–31. Recuperado de:

http://www.mecs-press.org/ijmecs/ijmecs-v9-n8/IJMECS-V9-N8-4.pdf

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. *International Journal of Education and Management Engineering, 7(6),* 40–49. https://doi.org/10.5815/ijeme.2017.06.05

Lamas, H. (2015). Sobre el rendimiento escolar. *Propósitos y Representaciones, 3(1),* 351–386. https://doi.org/10.20511/pyr2015.v3n1.74

Landwehr, N., Hall, M., & Frank, E. (2006). Logistic model trees. *Machine Learning, 2837*, 241–252. https://doi.org/10.1007/978-3-540-39857-8_23

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics, 41(1),* 191–201.

López-Ramirez, V. M. (2015). Método sistémico para evaluar el rendimiento académico en instituciones de educación superior (Tesis Doctoral). Instituto Politécnico Nacional, México. Recuperado de: https://tesis.ipn.mx/handle/123456789/21401

López, C. E., Guzmán, E. L., & González, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Revista Iberoamericana de Tecnologías del Aprendizaje, 10(3),* 119–125. https://doi.org/10.1109/RITA.2015.2452632

Márquez-Vera, C., Romero, C., & Ventura, S. (2012). Predicción del Fracaso Escolar Mediante Técnicas de Minería de Datos. *IEEE-Rita, 7(3),* 109–117. Recuperado de: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

Martínez, D. L., Karanik, M., Giovannini, M., & Pinto, N. (2015). Perfiles de Rendimiento Académico: Un Modelo basado en Minería de datos. *Campus Virtuales, 6(1)*, 12–30. Recuperado de: http://uajournals.com/ojs/index.php/campusvirtuales/article/view/66

Menacho, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos, 78(1),* 26. https://doi.org/10.21704/ac.v78i1.811

Merchan, S. M., & Duarte, J. A. (2016). Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance. *IEEE Latin America Transactions, 14(6),* 2783–2788. https://doi.org/10.1109/TLA.2016.7555255

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of

students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Minguillón, J., Casas, J., & Minguillón, J. (2017). Minería de datos: modelos y algoritmos. Recuperado de: https://elibro.net/es/ereader/bibliouaq/58656?page=10

Mitra, S., & Pal, S. K. (1995). Fuzzy multi-layer perceptron, inferencing and rule generation. *IEEE Transactions on Neural Networks, 6(1),* 51–63.

Molina, M. (2015). Valoración de los criterios referentes al rendimiento académico y variables que lo puedan afectar. *Revista Médica Electrónica, 37(6),* 617–626. Recuperado de: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242015000600007

Montes, I. C., & Lerner, J. (2012). Rendimiento Académico de los estudiantes de pregrado de la Universidad EAFIT. *Perspectiva Cuantitativa, 158*. Recuperado de: https://publicaciones.eafit.edu.co/index.php/cuadernos-investigacion/issue/download/156/22

Muñoz, A. (2015). Modelos para la Mejora del Rendimiento Académico de Alumnos de la E.S.O. mediante Técnicas de Minería de Datos (Tesis Doctoral). Universidad de Murcia, España. Recuperado de: https://dialnet.unirioja.es/servlet/tesis?codigo=127044

Pacheco, V., Cruz, E., & Serrano, L. A. (2019). Rendimiento académico como factor de riesgo en estudiantes de licenciatura. *Revista Electrónica de Psicología Iztacala, 22(2)*, 2318–2336. Recuperado de: http://www.revistas.unam.mx/index.php/repi/article/view/70168

Padua, L. M. (2019). Factores individuales y familiares asociados al bajo rendimiento académico en estudiantes universitarios. *Revista Mexicana de Investigación Educativa, 24(80)*, 173–195. Recuperado de: http://www.scielo.org.mx/pdf/rmie/v24n80/1405-6666-rmie-24-80-173.pdf

Peña-Ayala, A. (2014). Educational Data Mining. *In Studies in Computational Intelligence (Vol. 524)*. https://doi.org/10.1007/978-3-319-02738-8

Quinlan, R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers.

Rico, A., & Sánchez, D. (2018). Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN / Design of a model to automate the prediction of academic performance in students of IPN. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 8(16),* 246–266. https://doi.org/10.23913/ride.v8i16.340

Río-Jenaro, C., Calle, R., Martín, E., & Robaina, N. (2018). Rendimiento académico en educación superior y su asociación con la participación activa en la plataforma Moodle. *Estudios Sobre Educación, 34*, 177–198. https://doi.org/10.15581/004.34.177-198

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 40(6),* 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Silva, M. (2011). El primer año universitario. Un tramo crítico para el éxito académico. *Perfiles Educativos, 33(Extra 0),* 102–114. Recuperado de: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-

26982011000500010

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics, 42(1),* 85–106. https://doi.org/10.3102/1076998616666808

UADY. (2012). Sistema de Atención integral al Estudiante. *Universidad Autónoma de Yucatán*. Recuperado de: https://www.saie.uady.mx/tutorias/

Valenzuela, J. R., & Flores, M. (2012). Fundamentos de investigación educativa (eBook, Vol. II). Monterrey, México: Editorial Digital del Tecnológico de Monterrey.

Villanueva, A., Moreno, L. G., & Salinas, M. J. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review, (33),* 235–266. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=6485868

Witten, I., Frank, E., & Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann.