Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios

Classification models to recognize patterns of desertion in university students

Joshua Zárate-Valderrama[1] Norka Bedregal-Alpaca[2]* Víctor Cornejo-Aparicio[2]

Recibido 05 de agosto de 2020, aceptado 16 de octubre de 2020
Received: August 05, 2020 Accepted: October 16, 2020

RESUMEN

La deserción universitaria es un problema relacionado con el estudiante, como responsable directo, y con la institución universitaria, conocer las posibilidades de deserción es relevante para la institución. En este trabajo se propone utilizar modelos de clasificación para encontrar patrones y predecir posibles casos de deserción en estudiantes universitarios.
Se ha implementado una aplicación que utiliza información proporcionada por la universidad y que genera modelos de clasificación a partir de diferentes algoritmos (redes neuronales, ID3, C4.5), y utiliza los atributos más significativos dentro de la información disponible. Se comparó el rendimiento de estos modelos para definir aquel que aportaba mejores resultados y que servirá para realizar la clasificación de los estudiantes.
Los resultados muestran que el algoritmo de C4.5 presentó mejoras medidas de rendimiento con respecto a la red neuronal y al ID3 y que la relación de créditos aprobados por un estudiante con respecto a los créditos que debería haber llevado es la variable más significativa en la construcción del modelo, seguida por las calificaciones, mientras que, la modalidad de examen de admisión mediante la cual el estudiante ingresó a la universidad no resultó ser significativa, pues no aparece en el árbol de decisión generado.

Palabras clave: Minería de datos educativos, algoritmo ID3, algoritmo C4.5, red neuronal artificial, algoritmos de clasificación, deserción estudiantil.

ABSTRACT

University dropout is a problem related to the student, as a direct responsible, and with the university institution, knowing the possibilities of attrition is relevant for the institution. In this paper, it is proposed to use classification models to find patterns and predict possible dropouts in university students.
An application has been implemented that uses information provided by the university and that generates classification models from different algorithms (neural networks, ID3, C4.5), and uses the most significant attributes within the available information. The performance of these models was compared to define the one that provided the best results and that will be used to classify the students.
The results show that the algorithm of C4.5 presented improvements in performance with respect to the neural network and the ID3 and that the relation of credits approved by a student related to the credits that he should have taken is the most significant variable in the construction of the model, followed by the

[1] Universidad Nacional de San Agustín de Arequipa. Escuela Profesional de Ingeniería de Sistemas. Arequipa, Perú. E-mail: ashleyjzaratev@gmail.com
[2] Universidad Nacional de San Agustín de Arequipa. Departamento Académico de Ingeniería de Sistemas e Informática. Arequipa, Perú. E-mail: nbedregal@unsa.edu.pe; vcornejo@unsa.edu.pe
* Autor de Correspondencia: nbedregal@unsa.edu.pe

qualifications, while the modality of the admission exam through which the student entered the university did not turn out to be significant, since it does not appear in the generated decision tree.

Keywords: Educational data mining, ID3 algorithm, C4.5 algorithm, artificial neural network, classification algorithms, student desertion.

INTRODUCCIÓN

La educación es fundamental para el desarrollo y el bienestar de una sociedad, por tanto, los estudiantes son el activo fundamental de cualquier institución educativa. El desarrollo social y económico de un país está directamente relacionado con el rendimiento académico de sus estudiantes [1].

En este contexto y por las implicaciones que tiene, la deserción universitaria es un problema que genera preocupación en los directivos de las instituciones de educación superior; por un lado, se afectan las finanzas de la universidad y por otro, se cuestiona la eficiencia del sistema de educación superior, pues, solo un número reducido de jóvenes que inician estudios universitarios logra culminarlos.

Una forma de atacar este problema, es contar con información oportuna sobre la posibilidad de deserción de un estudiante, de allí el interés por analizar los posibles factores que pueden llevar a un estudiante a entrar en condiciones de abandono de los estudios.

Hoy en día, son múltiples las aplicaciones de la inteligencia artificial al ámbito educativo, se usan técnicas de minería de datos para descubrir patrones importantes y obtener información útil de sistemas de información de base académica [2].

La minería de datos utiliza diferentes técnicas: clasificación, agrupación, predicación, entre otras. La literatura existente muestra que las técnicas de minería de datos son eficaces para predecir el rendimiento académico de los estudiantes. Pimpa [3], aplica algoritmos de árboles de decisión y redes bayesianas utilizando Weka y validación cruzada 10 folds. Gart y Sharma [4] comparan varias técnicas C4.5, ID3, CART y J48, Naive de Bayes, Redes Neuronales, k-medias y k-vecino más cercano para predecir el rendimiento académico. Dole y Rajurkar [5] aplican el algoritmo Naive de Bayes y árbol de decisión para predecir la graduación y la condición final de los estudiantes: aprobado y desaprobado. Thai [6] compara la precisión de los algoritmos de árboles de decisión y redes Bayesianas, aplica Weka en los algoritmos árboles de decisión J48 y M5P y red Bayesiana para predecir el rendimiento académico de los estudiantes no graduados y graduados sobre su calificación final.

Este trabajo se centra en técnicas de clasificación, técnicas que emplean un conjunto de ejemplos preclasificados para desarrollar un modelo que puede clasificar una población de registros análogos [7].

Se usan diferentes algoritmos de clasificación: redes neuronales, ID3 y C4.5. El proceso de clasificación involucra un conjunto de entrenamiento consistente de datos con etiquetas de clase conocidas, conjunto que se usa para construir un modelo de clasificación que posteriormente se aplica a un conjunto de prueba que posee datos con etiquetas de clase desconocidas.

En este proceso se tiene como objetivo identificar las mejores variables que caracterizan al estudiante y que pueden servir para generar un modelo de clasificación confiable y que pueda ser utilizado por los directivos de la institución universitaria para detectar aquellos estudiantes que se puedan encontrar en riesgo de abandonar la carrera. En base a esa información, se podrán tomar medidas remediales para mejorar el rendimiento académico estudiantil y por tanto disminuir las tasas de deserción.

Si bien, en este trabajo el interés está en conocer si un estudiante abandonará o no los estudios (2 tipos de salida), el aplicativo desarrollado puede modificarse fácilmente para generar más clases de salida, por ejemplo, detectar el nivel de riesgo de deserción.

MARCO TEÓRICO

Clasificación

La clasificación es la técnica de extracción de datos más comúnmente aplicada. Utiliza un conjunto de

ejemplos preclasificados para desarrollar un modelo que puede clasificar la población de registros en general [7]. Esta técnica emplea un árbol de decisiones o los algoritmos de clasificación basados en redes neuronales.

El proceso de clasificación de datos implica dos fases, aprendizaje y clasificación. En la fase de aprendizaje, el algoritmo de clasificación analiza los datos de entrenamiento. En la fase de clasificación, se utilizan los datos de prueba para estimar la precisión de las reglas de clasificación. Si la precisión es aceptable, las reglas se pueden aplicar a las nuevas tuplas de datos.

El algoritmo de entrenamiento del clasificador utiliza los ejemplos preclasificados para determinar el conjunto de parámetros necesarios para una discriminación adecuada, luego codifica esos parámetros en un modelo llamado clasificador.

Redes neuronales artificiales (RNA)
Las Redes Neuronales Artificiales (RNA), son algoritmos de computación que pueden resolver problemas complejos imitando los procesos del cerebro animal de manera simplificada [8].

Se basan en una estructura de grafo dirigido, compuesto por un conjunto de neuronas que se interconectan a través de arcos dirigidos con un peso asociado que determina la fuerza y el signo de la conexión. Las neuronas se organizan por niveles o capas. Las RNA, tienen dos funciones definidas: activación y salida.

La RNA más conocida es la Perceptrón Multicapa (PM), que comprende una capa de entrada, una oculta y una de salida, este tipo de redes se están usando para realizar procesos de minería de datos [9].

Las redes PM están constituidas por neuronas dispuestas en capas e interconectadas por pesos sinápticos y pueden filtrar y transmitir información, de manera supervisada, para construir un modelo predictivo que clasifique los datos almacenados en la memoria.

Árboles de Decisión
Son estructuras en forma de árbol cuyos nodos representan una elección entre varias alternativas y cada nodo hoja representa una decisión [10-11].

Utilizan algoritmos de extracción de datos reales para ayudar con la clasificación. Se utilizan como apoyo para la elección entre varias líneas de acción, permitiendo explorar los posibles resultados para varias opciones, y evaluar el riesgo y las recompensas para cada posible curso de acción. Estas decisiones generan reglas, que luego se usan para clasificar los datos. Entre los algoritmos de aprendizaje de árboles de decisión destacan ID3, C4.5 y ASSISTANT.

Algoritmo ID3 - Iterative Dichotomizer 3
Algoritmo inventado por Ross Quinlan, iterativamente divide los atributos en dos grupos: el más dominante y los otros para construir el árbol. Calcula la entropía y la ganancia de información de cada atributo para determinar el atributo más dominante, que se coloca en el árbol como un nodo de decisión. Se repite el proceso con ellos atributos restantes, de modo que se encuentra el siguiente atributo más dominante. El procedimiento continúa hasta alcanzar una decisión para esa rama, es por ello que se denomina Dicotomizador Iterativo.

Construido el árbol, se le aplica a una tupla de la base de datos, dando como resultado la clasificación de esa tupla. El conjunto de datos ejemplo debe estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido).

Algoritmo C4.5
Se basa en el algoritmo C4.5 de Hunt, maneja atributos categóricos y continuos para construir un árbol de decisión. Para manejar los atributos continuos, C4.5 divide los valores de los atributos en dos particiones según el umbral seleccionado, de modo que todos los valores por encima del umbral se asignan a un hijo y el resto al otro hijo. También maneja valores de atributo faltantes. C4.5 utiliza la relación de ganancia como una medida de selección de atributo para construir un árbol de decisión. Elimina el sesgo de la ganancia de información cuando hay muchos valores de resultado de un atributo.

MATERIALES Y TÉCNICAS

Materiales
Los datos corresponden a los registros académicos de 970 estudiantes pertenecientes a la Escuela Profesional de Ingeniería de Sistemas.

Las variables consideradas para el estudio se describen en la Tabla 1.

La aplicación utilizada para implementar, ejecutar y comparar los algoritmos y sus resultados se desarrolló utilizando el lenguaje de programación Java, el motor de base de datos MySQL, además se uso de JPA, que es un framework de Java útil para manejar datos relacionales.

Técnicas de Minería de Datos

En esta investigación, las técnicas de minería de datos que se proponen para la clasificación de los estudiantes y predecir el riesgo de deserción (Abandona, No abandona) son: redes neuronales artificiales, algoritmo de C4.5 y algoritmo ID3.

Técnicas para evaluar clasificadores

La evaluación de las técnicas de clasificación es importante, pues permite validar la bondad de ajuste del modelo en relación al conjunto de entrenamiento. Así mismo, la evaluación permite comparar entre varias técnicas de clasificación y seleccionar aquella que proporcione mayor precisión.

Para la evaluación de las técnicas de minería de datos utilizadas en este trabajo se calcularon 3 medidas de rendimiento: exactitud, precisión y sensibilidad.

LA PROPUESTA

Para el desarrollo de esta propuesta se siguió el proceso clásico de KDD (Figura 1). El proceso KDD (*Knowledge Discovery in Databases*) consiste en extraer patrones en forma de reglas o funciones, obtenidas a partir de los datos. El proceso implica pre-procesar los datos, hacer minería de datos y presentar resultados [12, 13]; se trata de un proceso interactivo e iterativo que involucra la participación del usuario en la toma de decisiones.

Tabla 1. Variables y sus valores.

Variable	Valor
Género	0 – Masculino 1 – Femenino
Promedio de Notas de Asignaturas de Carrera	0 a 1 (Normalizado con min = 0 y max = 20)
Promedio de Notas de Asignaturas Generales	0 a 1 (Normalizado con min = 0 y max = 20)
Colegio de Procedencia	1 – Estatal 2 – Particular 3 – Parroquial
Relación de Asignaturas de Carrera y sus correspondientes créditos	0 a 1
Relación de Asignaturas Generales y sus correspondientes créditos	0 a 1
Calificación de examen de admisión	0 a 1 (Normalizado con min = 0 y max = 100)
Tipo de Admisión	1 – Ordinaria 2 – Extraordinaria
Relación de créditos que posee con respecto a los que debería poseer	0 – 1
Relación de número de asignaturas desaprobadas con respecto al total de asignaturas que se lleva en un semestre	0 – 1
Relación de número de asignaturas abandonadas con respecto al total de asignaturas que se lleva en un semestre	0 – 1

Fuente: Baradwaj [7].

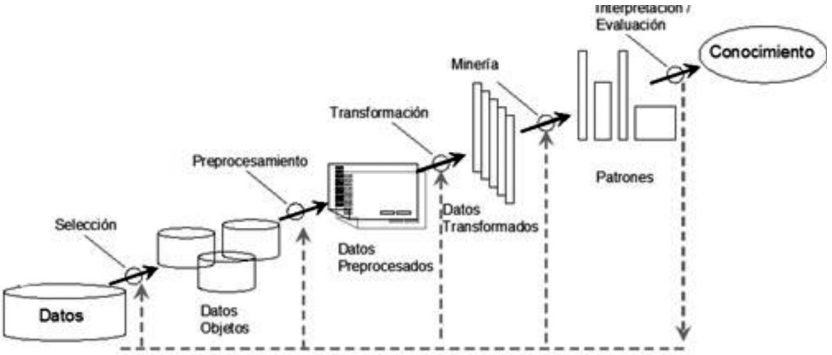

Figura 1. Pasos para extraer conocimiento de los datos.

Datos y selección

Se tomó como entrada los datos de estudiantes de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín (UNSA) de Arequipa, Perú.

La información la proporcionó el Instituto de Informática de la UNSA, mediante 3 archivos con extensión .xlsx. Esta información se migró a una base de datos cuya estructura de tablas y campos almacenados permite trabajar de mejor manera con los clasificadores.

El primer archivo contiene información básica de los estudiantes como son: nombres y apellidos, C.U.I. (código único de identificación), fecha de nacimiento, género, y tipo y lugar del colegio de procedencia. El segundo archivo contiene el registro de calificaciones finales de los estudiantes en las asignaturas consideradas en el plan de estudios. Se especifica en este archivo el nombre y código de la asignatura que cursó un estudiante (identificado por su C.U.I.), el grupo, el número de matrícula, la nota, la condición al finalizar el curso (aprobó, desaprobó o abandonó la asignatura), el año y ciclo académico en que el estudiante llevó la asignatura. El tercer archivo contiene el puntaje obtenido por el estudiante en la prueba de admisión, la modalidad de ingreso y el puesto en que ingresó.

Preprocesamiento y transformación

Identificadas las fuentes de datos, fue necesario preprocesarlos y transformarlos con el fin de tenerlos en forma de archivo plano.

Una vez transformados los datos a un formato más adecuado, se identificaron los registros incompletos, incorrectos e irrelevantes y se eliminaron los errores de tipeo. Posteriormente, se filtraron los datos para eliminar la información no relevante.

La data contenía 970 registros, sin embargo, luego de la limpieza y depuración, el número de registros con información útil y completa se redujo a 300.

El promedio de las calificaciones obtenidas por un estudiante es una de las variables más utilizadas en otras investigaciones en torno a rendimiento académico, en este trabajo se han incluido diferentes formas de calcular ese promedio: promedio aritmético simple, promedio ponderado por el número de créditos de la asignatura, promedio que considera sólo las calificaciones aprobatorias. Otros tipos de promedio utilizados son: promedio de notas en un semestre, promedio desde el primer semestre hasta un semestre específico, promedio en asignaturas de especialidad.

Se consideraron también otras variables: número de créditos que debería poseer el estudiante al finalizar un semestre, número de asignaturas reprobadas o abandonadas durante el semestre.

Por tanto, los modelos implementados utilizaron no solo variables obtenidas directamente de la base de datos inicial, sino que utilizaron también variables calculadas.

Para trabajar con redes neuronales es necesario normalizar los datos, por lo que variables como las notas que se encuentran en un rango 0 a 20, se escalaron a valores entre 0 y 1.

Completada la etapa de preprocesamiento, tenía que dividirse la base en subconjuntos, se tomó un 80% de la base para el conjunto de datos de entrenamiento y un 20% para el de prueba.

Minería de datos

Fase de Entrenamiento

El conjunto de entrenamiento se utilizó para entrenar los modelos de clasificación mencionados. También se evaluó durante este proceso el tiempo de entrenamiento que toma ejecutar cada uno de los algoritmos, esto se hizo en razón de que este puede ser un indicador significativo, pues si se desea cambiar de base de datos, el tiempo puede variar significativamente debido al número de variables de entrada y al tamaño del conjunto de entrenamiento.

Para realizar el entrenamiento de los modelos se le proporcionó al sistema un archivo con las variables de entrada (género, promedio de notas, relación de créditos, etc.) y un campo clasificado como salida (abandona o no) del modelo. También se le proporcionó un archivo con la lista de los códigos de estudiantes que componían el conjunto de entrenamiento junto con la etiqueta de clase a la que pertenecían, clasificación que se hizo previamente en base a información real (Figura 2). Esta acción se justifica en que el entrenamiento se realiza usando datos conocidos y cuando se realiza la clasificación

real, se hace sobre datos no trabajados ni en la fase de entrenamiento ni en la fase de prueba, por lo que los resultados así obtenidos son una predicción.

El sistema entonces recibe una matriz (Estudiante x Variables) y construye, según la elección del usuario, un modelo de clasificación (Tabla 2).

Si la opción elegida es una red neuronal, se construye una red perceptrón multicapa entrenada con el algoritmo backpropagation. Para detener el entrenamiento, se establecieron dos criterios diferentes: llegar a una tasa de error del 0,01% o llegar a un máximo de 100000 iteraciones.

Al terminar el entrenamiento, se guardaron los pesos de las conexiones de las neuronas en un archivo, lo que permitiría armar nuevamente la red en caso se quisiera utilizar nuevamente el modelo.

Luego de la fase de entrenamiento de los algoritmos ID3 y C4.5 (algoritmos que construyen un árbol de decisión utilizando fórmulas de entropía y ganancia de información), se construyeron los árboles de decisión respectivos y se guardaron en la base de datos del sistema.

En el caso del algoritmo C4.5 se hizo uso de las fórmulas de la tasa de ganancia (Gain Ratio) en lugar la ganancia; y la fórmula de información de división.

Tabla 2. Matriz Estudiante x Variable.

Estudiante	**Variable**			
	V1	**V2**	**…**	**Vm**
E1	E1V1	E1V2	…	E1Vm
E2	E2V1	E2V2	…	E2Vm
E3	E3V1	E3V2	…	E3Vm
…	…	…	…	…
En	EnV1	EnV2	…	EnVm

Fuente: Aplicación implementada específicamente para esta investigación.

LISTA DE VARIABLES

VARIABLE	CLASE	VALOR
GÉNERO	MASCULINO	1.0
GÉNERO	FEMENINO	-1.0
PROMEDIO NOTAS	MUY BUENO	16.0 - 20.0
PROMEDIO NOTAS	BUENO	14.0 - 16.0
PROMEDIO NOTAS	REGULAR	11.0 - 13.0
PROMEDIO NOTAS	MALO	7.0 - 10.0
PROMEDIO NOTAS	MUY MALO	0.0 - 6.0
RELACIÓN CRÉDITOS	BUENO	75.0 - 100.0
RELACIÓN CRÉDITOS	REGULAR	50.0 - 75.0

CONJUNTO DE ENTRENAMIENTO

GÉNERO	PROMEDIO NO...	RELACIÓN CR...	COLEGIO PRO...	PUNTAJE EXA...	CLASIFICACIÓN
MASCULINO	11.0	73.0	PARROQUIAL	53.0	NO ABANDONA
MASCULINO	3.0	10.0	PARROQUIAL	63.0	ABANDONA
MASCULINO	6.0	10.0	ESTATAL	60.0	ABANDONA
MASCULINO	10.0	38.0	PARROQUIAL	58.0	ABANDONA
FEMENINO	9.0	47.0	PARTICULAR	48.0	NO ABANDONA
FEMENINO	10.0	53.0	PARROQUIAL	56.0	ABANDONA
MASCULINO	13.0	86.0	PARROQUIAL	68.0	NO ABANDONA
FEMENINO	10.0	62.0	PARTICULAR	48.0	NO ABANDONA

CONJUNTO DE PRUEBA

GÉNERO	PROMEDIO ...	RELACIÓN ...	COLEGIO P...	PUNTAJE E...	CLASIFICACI...	RESULTADO
FEMENINO	13.0	89.0	PARTICULAR	65.0	NO ABANDO...	NO ABANDO...
MASCULINO	11.0	64.0	PARTICULAR	58.0	NO ABANDO...	ABANDONA
FEMENINO	12.0	84.0	PARTICULAR	58.0	NO ABANDO...	NO ABANDO...
MASCULINO	1.0	0.0	ESTATAL	60.0	NO ABANDO...	ABANDONA
MASCULINO	14.0	100.0	ESTATAL	65.0	NO ABANDO...	NO ABANDO...
MASCULINO	12.0	81.0	PARROQUIAL	58.0	NO ABANDO...	NO ABANDO...
MASCULINO	6.0	35.0	PARROQUIAL	58.0	ABANDONA	ABANDONA
FEMENINO	13.0	91.0	ESTATAL	58.0	NO ABANDO...	NO ABANDO...
MASCULINO	12.0	86.0	PARROQUIAL	68.0	NO ABANDO...	NO ABANDO...

Figura 2. Tabla de variables y conjunto de entrenamiento y de prueba generados.

En cualquiera de los tres modelos explicados, las variables de salida clasifican los registros en NO ABANDONA y ABANDONA.

Fase de Prueba

Posteriormente a la fase de entrenamiento se deben realizar las pruebas respectivas con el conjunto de prueba. Para cada modelo entrenado se calcularon 3 medidas de rendimiento: exactitud (ecuación (1)), precisión (ecuación (2)) y sensibilidad (ecuación (3)).

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precisión = \frac{TP}{TP + FP} \quad (2)$$

$$Sensibilidad = \frac{TP}{TP + FN} \quad (3)$$

Donde, TP (True Positive) es la proporción de casos que se clasificaron como la clase real e indica qué parte de la clase se capturó correctamente, TN (*True Negative*), FP (*False Positive*) es la proporción de casos que se clasificaron como una clase en particular, pero pertenecen a una clase diferente y FN (*False Negative*).

Para realizar la fase de entrenamiento, se proporcionó al sistema el conjunto de registros que pertenecían al conjunto de pruebas. Análogamente a la fase de entrenamiento, se especificó la clase a la que pertenece el estudiante (ABANDONA y NO ABANDONA), el objetivo era aplicar el modelo de clasificación a cada registro y comparar la salida con la clase real a la que pertenece el registro y aplicar las medidas de rendimiento.

Considerando los valores obtenidos para las medidas de rendimiento establecidas, se puede determinar qué modelo es el más adecuado al momento de clasificar, pues presenta menos errores al determinar si un estudiante se encuentra o no en uno de los estados (ABANDONA O NO ABANDONA).

RESULTADOS

La Figura 3 muestra el árbol de decisión construido por el clasificador C.45.

Como un árbol de decisión puede representarse linealmente por medio de reglas de decisión, donde el resultado es el contenido del nodo hoja, y las condiciones a lo largo del camino forman una conjunción en la cláusula "if", entonces el árbol obtenido se podría interpretar el árbol de la siguiente manera:

- Si la RELACIÓN DE CRÉDITOS (créditos actuales/créditos que debería tener) es menor

Fuente: Aplicación implementada específicamente para esta investigación.

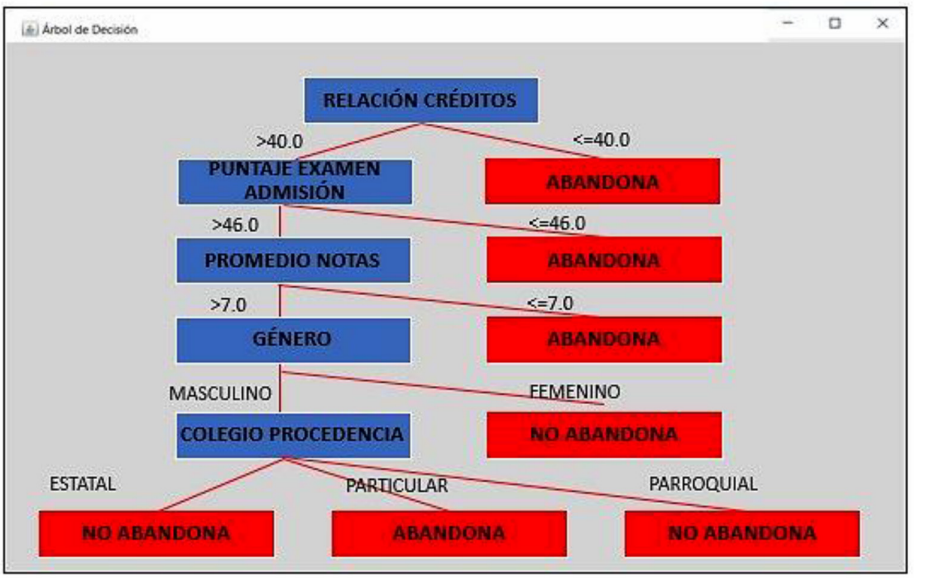

Figura 3. Árbol de Decisión Generado por el Aplicativo.

o igual al 40% entonces el estudiante corre el riesgo de abandonar.
- Caso contrario, si su PUNTAJE EN EL EXAMEN DE ADMISIÓN fue menor o igual a 46 entonces corre el riesgo de abandonar.
- Caso contrario, si el PROMEDIO DE NOTAS es menor a 7 entonces corre el riesgo de abandonar.
- Caso contrario, si el GÉNERO del estudiante es mujer no abandona la carrera. Si es hombre entonces abandonará la carrera si ha estudiado en un colegio particular y no abandona si ha estudiado en un colegio estatal o parroquial.

Después de generar el modelo se clasificó cada estudiante dentro del conjunto de prueba para aplicar métricas de rendimiento tales como exactitud (qué tan cerca está un valor medido respecto al valor verdadero), precisión (proporción de casos que realmente tiene la clase real) y sensibilidad (Figura 4). Los modelos generados utilizando un modelo de red neuronal artificial, dieron resultados muy similares al obtenido por el C4.5; mientras que el clasificador ID3 produjo resultados algo diferentes (Tabla 3).

Luego de generar los modelos utilizando cada uno de los algoritmos de clasificación se escogió el mejor considerando las medidas de rendimiento calculadas para cada uno (Figura 5).

Tabla 3. Métricas obtenidas para los diferentes algoritmos empleados.

Algoritmo	Exactitud	Precisión	Sensibilidad
Redes neuronales	66%	62%	85%
C4.5	68%	65%	92%
ID3	60%	58%	75%

DISCUSIÓN DE LOS RESULTADOS

El algoritmo de C4.5 es el que produjo métricas de rendimiento con mejores resultados, mientras que la red neuronal tuvo resultados muy cercanos. Se podía asumir que ID3 iba tener resultados inferiores al menos al C4.5 ya que éste último es una versión mejorada del ID3 que permite trabajar tanto con valores continuos como categóricos mientras que el ID3 solo permite categóricos.

El modelo podría mejorarse si se utilizase un mayor número de datos y variables, por ejemplo, en un futuro recolectar información acerca de variables relacionadas a factores sociales. Esto podría ayudar en generar un modelo con medidas de rendimiento más confiables.

El árbol generado por el algoritmo C4.5 incluye todas las variables ingresadas en el entrenamiento

Fuente: Aplicación implementada específicamente para esta investigación.

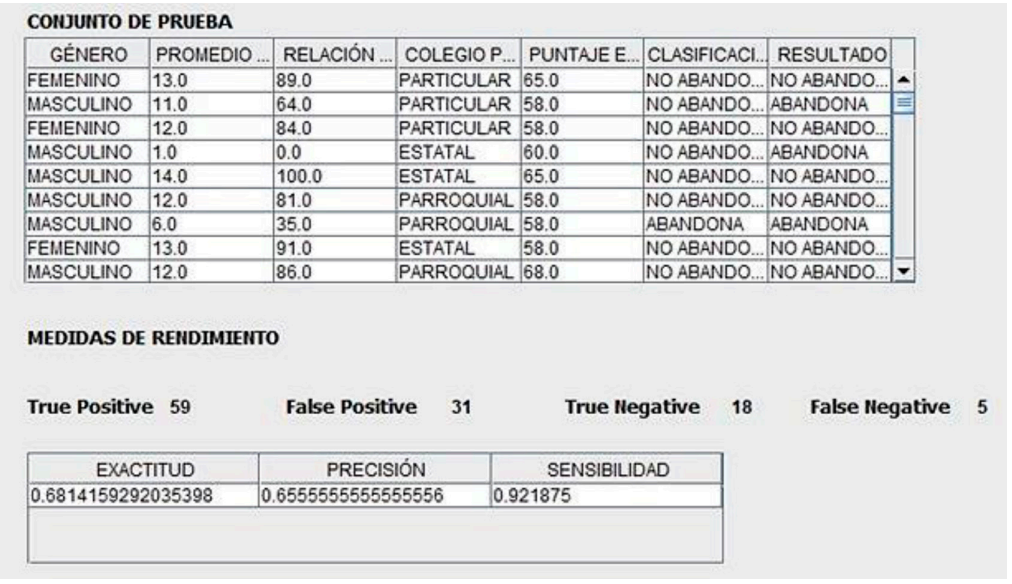

Figura 4. Clasificación realizada en conjunto de prueba y métricas de rendimiento por el Aplicativo.

Fuente: Aplicación implementada específicamente para esta investigación.

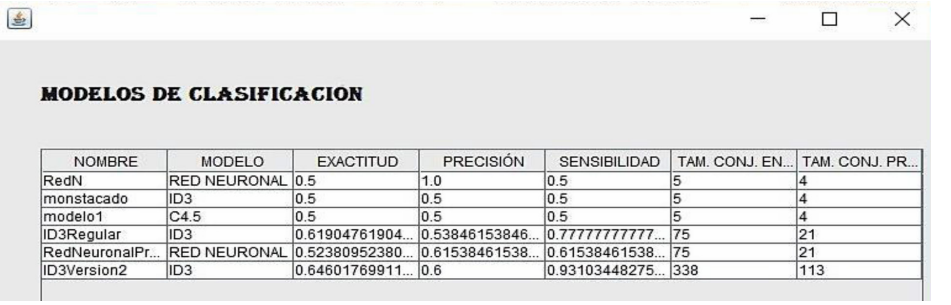

Figura 5. Medidas de rendimiento de los modelos de clasificación.

a excepción de la modalidad del examen de admisión, lo cual se debería a que no es una variable significativa dentro del modelo, mientras que la relación de créditos que tiene el estudiante sería la variable más significativa al encontrarse como raíz del árbol de decisión.

Debido a que el modelo generado por el C4.5 fue el más exitoso se utilizará para que se realicen las nuevas clasificaciones cuando sean requeridas por los directores de escuela.

CONCLUSIONES

Las Técnicas de Minería de Datos han mostrado ser herramientas eficaces para obtener modelos que permitan predecir la permanencia de los estudiantes matriculados en una carrera de ingeniería.

Con el desarrollo de este trabajo se ha podido determinar aquellos factores que han incidido en la tasa de deserción, para ello se utilizaron distintas variables del tipo personal y académicas.

El algoritmo de C4.5 presentó mejoras medidas de rendimiento con respecto a la red neuronal y sobre todo con respecto al algoritmo ID3 (debido a que el ID3 no puede trabajar con datos continuos).

La relación de créditos actuales de un estudiante con respecto a los créditos que debería poseer resultó ser la variable más significativa en la construcción del modelo, seguido por las notas, mientras que el tipo o modalidad de examen de admisión con el que el estudiante ingresó a la universidad no resultó ser significante ya que no aparece en el árbol de decisión generado.

Es así que los resultados mostraron que la deserción está relacionada con los créditos en los que se matriculó el estudiante y que la proporción de créditos aprobados sobre los créditos matriculados es una buena medida de su rendimiento académico. En este resultado se coincide con otros estudios en los que se observó que el promedio de créditos en los que se matriculó el estudiante es mayor en aquellos que no desertan comparado con aquellos que sí lo hacen.

Las medidas de rendimiento obtenidas podrían mejorarse, así como la tasa de aciertos de clasificación de los modelos al aumentar nuevas variables en un futuro, como variables sociales, económicas, etc. Por ello para ampliar los resultados de la investigación se propone considerar variables institucionales y socio-económicas.

El sistema elaborado puede adaptarse para crear modelos con diferentes variables de entrada y salida (clases) y almacenar los modelos con mejor rendimiento para ser utilizados en varias escuelas profesionales o que cada escuela utilice un modelo de clasificación diferente.

RECONOCIMIENTOS

Este trabajo se realizó con el apoyo de nuestra casa de estudios, la Universidad Nacional de San Agustín, en la que el Vicerrectorado de Investigación canaliza los recursos provenientes del canon minero y convoca a un conjunto de esquemas financieros concursables.

Es mediante uno de ellos que se financió el proyecto “Modelo de evaluación de desempeño académico para

la detección de estudiantes destacados y estudiantes en riesgo académico", en el que el primer autor participa como tesista de pregrado, el segundo autor es la investigadora principal del proyecto y el tercer autor es co-investigador en el mismo.

REFERENCIAS

[1] I. Mushtaq and S. Nawaz. "Factors Affecting Students' Academic Performance". Mohammad Ali Jinnah University Islamabad. Pakistan. 2012.

[2] S. Oloruntoba and J. Akinode. "Student Academic Performance Prediction using Support Vector Machine". Computer Science Department, the Federal Polytechnic, ILARO, Ogun State, Nigeria. 2017.

[3] C. Pimpa. "Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program". Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. I. 2013.

[4] S. Gart and A. Sharma. "Comparative Analysis of Data Mining Techniques on Educational Dataset". International Journal of Computer Applications. Vol. 74 Issue 5, pp. 1-5. 2013.

[5] L. Dole and J. Rajurkar. "A Decision Support System for Predicting Student Performance". International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2. 2014.

[6] N-Thai. "A comparative analysis of techniques for predicting academic performance". 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge. 2007.

[7] B. Baradwaj and S. Pal. "Mining Educational Data to Analize Students' Performance". International Journal of Advanced Computer Science and Applications. Vol. 2 Nº 6. 2011.

[8] N. Zacharis. "Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks". International Journal of Artificial Intelligence and Applications (IJAIA). Vol. 7 Nº 5. 2016.

[9] K. Manchandia and N. Khare. "Implementation of Student Performance Evaluation through Supervised Learning Using Neural Network". 2017.

[10] I. Ganiyu. "Data Mining: A Prediction for Academic Performance Inprovement of Science Student using Classification". 2016.

[11] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules". vldb Conference, Santiago, Chile. 1994.

[12] G. Piatetsky-Shapiro, R. Brachman and T. Khabaza. "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications". Association for the Advancement of Artificial Intelligence. 1996.

[13] J. Han and M. Kamber. "Data Mining Concepts and Techniques". San Francisco, Morgan Kaufmann Publishers. 2001.