Predicción del rendimiento académico en la UNADECA por medio de sistemas de clasificación

Predict Academic Performance at UNADECA through Classification Systems

Dodanim Castillo Aráuz
Universidad Adventista de Centroamérica
Costa Rica, dcastillo@unadeca.ac.cr

Jairo Jonathán Martínez
Universidad Autónoma de Honduras
Honduras, jairo.martinez@unah.edu.hn

Cómo citar / How to cite

Castillo Aráuz, D., & Martínez, J. J. (2023). Predicción del rendimiento académico en la UNADECA por medio de sistemas de clasificación. *Unaciencia Revista De Estudios E Investigaciones*, *16* (31), 17–35. https://doi.org/10.35997/unaciencia.v16i31.738

Fecha de recepción: 15 de agosto de 2023
Fecha de aprobación: 16 de noviembre de 2023

Resumen

Predecir el rendimiento académico de los estudiantes no es solo una tarea que atrae a los investigadores sino también al personal administrativo de la facultad universitaria. Es posible crear modelos efectivos mediante algoritmos específicos para minería de datos educativos supervisados y no supervisados. Al conjunto de datos se le aplicaron técnicas de limpieza y codificación. La ejecución de los algoritmos y la comparación de sus métricas permitieron determinar los cursos a los que se debe dar asistencia con mayor atención en la búsqueda de mejorar el rendimiento académico de los estudiantes. Los datos fueron fraccionados en dos grupos, uno para aprendizaje y otro para predicción. Se usaron algoritmos en el lenguaje Python y una herramienta gráfica, *RapidMiner Studio*. No se trabajaron agrupamientos por falta de información consistente en los

datos originales. El algoritmo de clasificación que tuvo las mejores métricas fue el *Random Forest* superando en los distintos casos el 90% de *accurracy*. En cambio, *RapidMiner* el algoritmo con mejores resultados *fue Gradient Boosted Trees* con un *accuracy* del 93.6%, con la predicción específica del resultado final de aprobado o reprobado. Se hizo una comparativa por escuelas, con resultados muy similares para Enfermería, Psicología y Teología, con una precisión aproximada de 93%.

Palabras clave: Minería de datos educativos, aprendizaje automático, random forest, métricas.

Abstract

Predicting the academic performance of students is not only a task that attracts researchers but also the administrative staff of university faculty. Effective models can be created using specific algorithms for supervised and unsupervised educational data mining. Cleaning and coding techniques were applied to the data set. The execution of the algorithms and the comparison of their metrics made it possible to determine the courses that should be assisted with greater attention in the quest to improve students' academic performance. The data were divided into two groups, one for learning and the other for prediction. Algorithms in the Python language and a graphical tool, RapidMiner Studio, were used. No clustering was performed due to lack of consistent information in the original data. The classification algorithm that had the best metrics was Random Forest, exceeding 90% accurracy in the different cases. RapidMiner, on the other hand, the algorithm with the best results was Gradient Boosted Trees with an accuracy of 93.6%, with the specific prediction of the result of pass or fail. A comparison was made by schools, with very similar results for Nursing, Psychology and Theology, with an accuracy of approximately 93%.

Key Words: Educational Data Mining, Machine Learning, random forest, metrics.

1. INTRODUCCIÓN

Las instituciones educativas terciarias trabajan en alcanzar altos estándares y ser ubicadas en los primeros lugares de las clasificaciones mundiales. Una herramienta que se utiliza para el mejoramiento continuo es la predicción del rendimiento académico. Con estos métodos se puede anticipar el nivel de conocimiento logrado por los discentes. La información que se recopila se puede crear programas y políticas educativas más efectivas y personalizadas. Además, al identificar los factores que influyen en el éxito estudiantil, las universidades pueden implementar programas de apoyo dirigidos a quienes necesiten. El impacto que tiene en reducir las tasas de deserción y mejora de la retención de personas es impresionante. Así mismo, al anticipar las necesidades y desafíos, las entidades lograr desempeñar un rol más efectivo en su formación.

Sin embargo, predecir el rendimiento académico, es una tarea desafiante. El manejo adecuado de los grandes volúmenes de información que emanan del quehacer educativo requiere de técnicas de minería de datos. Esta disciplina descubre patrones, relaciones y tendencias ocultas en los datos. Por eso, cada vez es más frecuente usar técnicas de minería aplicados a la educación

y algoritmos de aprendizaje automático que permitan monitorear y predecir el rendimiento académico estudiantil, según una serie de características que cada institución decide especificar.

La Universidad Adventista de Centroamérica (UNADECA), actualmente, no cuenta con ningún estudio que mida el rendimiento académico estudiantil histórico. Se tienen datos académicos y otras variables desde el primer cuatrimestre del año 2002 hasta el primer cuatrimestre del año 2022 en el sistema de información actual. Sin embargo, no han sido estudiados. Así, ¿es viable para la UNADECA comparar algoritmos de agrupamiento y predicción en un sistema que monitorea y predice el rendimiento académico de sus estudiantes? Por ello se comparan algoritmos supervisados y no supervisados al conjunto de datos para explorar, medir e identificar cuáles puedan predecir con las mejores métricas el rendimiento académico.

En primera instancia, se aplicaron técnicas de limpieza y agrupamiento a los datos. Seguidamente se corrieron en distintos algoritmos usando Python y RapidMiner (Mierswa, Igno, 2022) en la búsqueda de los parámetros con mejores rendimientos. Se trabajó por género, por países y finalmente por escuelas, para detectar los cursos con mejor posibilidad en la predicción.

Al final del artículo se hace una breve discusión de los resultados y conclusiones de las pruebas. Este estudio ha generado una gran motivación a trabajar con el departamento académico de registro en afinar las características de los datos y poder realizar clasificaciones o agrupamientos más específicos. Estos últimos no se completaron en este estudio por la falta de datos consistentes -se trabajará como una recomendación de mejora.

2. METODOLOGÍA

APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es considerado una rama de la inteligencia artificial, ya que aplica diversas técnicas en aprendizaje del pasado (Alférez et al., 2022), (Domingos, 2012). Se relaciona con las ciencias computacionales, las ingenierías, la estadística, la minería de datos y con reconocimiento de patrones. Los componentes clave para obtener resultados ideales de acuerdo con el problema en cuestión, son tres, a saber: la representación, la evaluación y la optimización (Domingos, 2012). A su vez, se ramifica en aprendizaje supervisado y no supervisado.

Aprendizaje automático supervisado

Se conoce como aprendizaje supervisado aquel en que los datos son etiquetados con el objetivo de entrenar un modelo (Almasri et al., 2020); esto conduce a una categorización o clase. El algoritmo aprende de las características de los datos para realizar predicciones. Muchos algoritmos son agrupados en esta categoría, como por ejemplo los siguientes:

K-Nearest Neighbor: todos los casos posibles son almacenados; las nuevas muestras serán clasificadas en función de la etiqueta más frecuente de sus k vecinos más cercanos (Hasterok et al., 2019)

Logistic Regression (regresión logística): el resultado es modelado basado en la probabilidad como función de las características individuales –considerando las variables independientes. Las características se multiplican por un peso y luego se suman. Posteriormente, esta suma se pone en una función lógica y devuelve el resultado que puede tomarse como una estimación de probabilidad.

Decision Trees (árboles de decisión): Este modelo ordena la información en un conjunto de datos que se mueven a través de una estructura de consulta desde la raíz hasta que llega a la hoja, que representa una clase. Los pasos a seguir se resumen así: coloca todos los ejemplos de entrenamiento en función de los atributos seleccionados, selecciona los atributos mediante el uso de algunas medidas estadísticas y continúa recursivamente el proceso de partición hasta que el árbol está completamente formado (Rimadana et al., 2019). Cada bloque tiene una medida que se conoce como entropía, mide el desorden o la incertidumbre en un grupo de muestras. Cuando mayor es la entropía, más confusos son los datos. Este algoritmo intenta disminuir la entropía a medida que avanza cada bloque. Cuando se llega a una conclusión la entropía es cero (Alférez et al., 2022).

Support Vector Machine (máquinas de vectores de soporte): para este algoritmo los datos son trazados en un espacio *n* dimensional (basado en el número de características) y un límite de decisión (o hiperplano) que divide los datos en clases. Cuanto más lejos estén los puntos de datos graficados del límite de decisión, más seguro estará el algoritmo sobre la predicción. Los puntos de datos más cercanos al hiperplano se denominan vectores de soporte (Alférez et al., 2022)

Multilayer Perceptron (perceptrón multicapa): es la red neuronal artificial (RNA) más común. Tiene dos capas conectadas directamente con el entorno, conocidas como entrada y salida; entre las dos están las intermedias que se denominan ocultas. Cada una contiene neuronas conectadas con la siguiente. La señal que transmiten las neuronas sigue una única dirección (de la entrada hasta la salida) sin formar bucles. Esta estructura es llamada *feed fordward*. También utiliza un algoritmo llamado reprogramación para minimizar los errores entre los resultados del modelo y los resultados esperados (Marius-Constantin et al., 2009).

Gaussian Naive Bayes (NB): se considera un clasificador probabilístico fácil y simple que depende de la aplicación del teorema de Bayes. Cada variable de atributo se constituye en una variable independiente, requiere pocos datos para ser entrenado y generar resultados (Kamel et al., 2019).

Aprendizaje automático no supervisado

Por otro lado, en el aprendizaje no supervisado, los datos no son etiquetados. Lo que el algoritmo realiza es un agrupamiento para descubrir patrones en ellos. Algunos métodos son:

Principal Component Analysis (PCA): es una técnica utilizada para reducir la dimensionalidad de los datos buscando que la pérdida de información sea la menor posible. El conjunto de datos se transforma en un sistema de coordenadas. Se elige un primer eje en la dirección de mayor variación en los datos. Se escoge un segundo eje ortogonal al primer eje y con la mayor varianza posible. Este proceso se repite hasta que todos los datos estén todos incluidos en los componentes principales generados; presenta una mejora en el rendimiento de los algoritmos y reduce el tiempo de entrenamiento (Hasterok et al., 2019).

K-Means: es un algoritmo para formar clústeres con características similares. El centro de cada conglomerado se llama centroide y es la distancia media de los valores en cada conglomerado. El algoritmo encuentra k conglomerados únicos y cada muestra se asigna al conglomerado con el centro más cercano. Es un algoritmo iterativo, es decir, ejecuta ciclos o iteraciones cada vez que se actualizan los centros de los clústeres (Alférez et al., 2022).

En el presente artículo se muestran los resultados de comparar distintos algoritmos y sus métricas, con el fin evaluar los que mejor predicen el rendimiento académico con los datos en estudio.

3. RESULTADOS

En la búsqueda de los algoritmos que den mejores resultados en la predicción del rendimiento académico, se decidió primero hacer un tratamiento exhaustivo de los datos. Luego, se trabaja con algunos algoritmos en Python y en *RapidMiner Studio* como herramientas principales. Finalmente se hizo una valoración de los resultados.

Preparación de los datos

Una tarea importante para realizar con los datos es prepararlos de forma adecuada para usarse en algoritmos de aprendizaje automático. La metodología ETL (*Extract, Transform and Loading*) se aplicó para el tratamiento de estos. Se solicitaron reportes históricos de los estudiantes a partir del primer cuatrimestre del año 2002 hasta el primer cuatrimestre del año 2022. Se obtuvieron 59,253 registros agrupados en 11 atributos.

Se procedió a la limpieza de datos. En primera instancia, se eliminaron características que permitían el tratamiento privado y anónimo de los datos. Así, se omitieron los campos como nombre del estudiante, fecha de nacimiento y fecha de ingreso. En la tabla 1 se presenta una descripción de los datos en estudio y sus atributos. Los datos quedaron listos para aplicar la codificación necesaria.

Tabla 1.

Descripción de los datos.

Atributo	Descripción
Carrera	Tipo nominal, basada en las 6 escuelas de la Universidad.
Grado	Reúne los 22 énfasis en las distintas escuelas.
Materia	El campo nominal más amplio, pues cuenta con 600 cursos en total que se ofrecen.
Num_créditos	Campo numérico que indica el número de créditos que tiene como peso la materia.
Cuatrimestre	Campo numérico invertido en orden (1 para segundo cuatrimestre 2022) hasta completar 61 períodos lectivos (primer cuatrimestre 2002).
Nota	Campo numérico con la nota obtenida en el curso (será una de las variables a predecir).
Status	Campo nominal, que muestra el resultado final obtenido en el curso A aprobado o R Reprobado. Se utiliza como predicción en algoritmos binarios.
País	Campo nominal con los 24 países de donde se tienen estudiantes históricos.
Filiación	Campo nominal donde se registra la filiación religiosa del estudiante resumido en cinco opciones.
Género	Campo nominal, con dos opciones para indicar el género o sexo del estudiante.
Edad	Campo numérico que se calculó la edad para dejarla como campo fijo.

Fuente: elaboración propia.

Ya con los datos limpios, se obtuvieron estadísticas básicas. En la tabla 2 se observa que la nota promedio fue de 81.43 con una desviación típica de 20.63. Además, se observa una tendencia no central de los datos, pues en su mayoría está agrupado a la derecha, como se observa en la figura 1. El acercamiento que tienen los cuartiles bajo esa tendencia se puede observar en la figura 2.

Tabla 2.

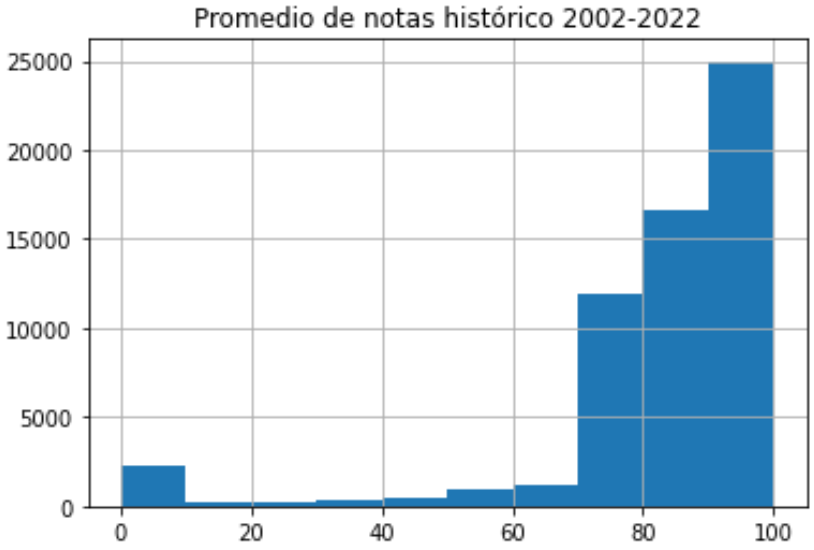
Estadística básica.

Medida	Valor
Cantidad de datos	59253
Media aritmética	81.432249
Desviación estándar	20.638517
Valor mínimo	0.000000
Cuartil 1 - 25%	77.00
Cuartil 2 - 50%	87.00
Cuartil 3 - 75%	94.00
Máximo	100.00

Fuente: elaboración propia.

Gráfico 1.

Análisis del Rendimiento Académico en la Universidad.

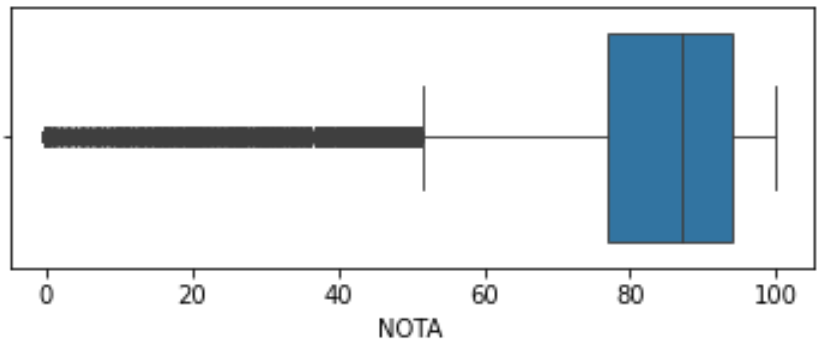


Fuente: elaboración propia.

Nota: El rendimiento promedio de los estudiantes de la Universidad refleja una tendencia alta a superar la nota de aprobación de 70. Aun así, existen bastantes registros con nota tendientes a cero, que deberán ser evaluadas las razones conducentes a esas calificaciones.

Gráfico 2.

Gráfico de Caja: Media y Mediana.



Fuente: elaboración propia.

Nota: En este gráfico de caja se observa una clara tendencia de la media aritmética y la mediana, superando ambas la nota de aprobación de 70%.

Aplicación de algoritmos

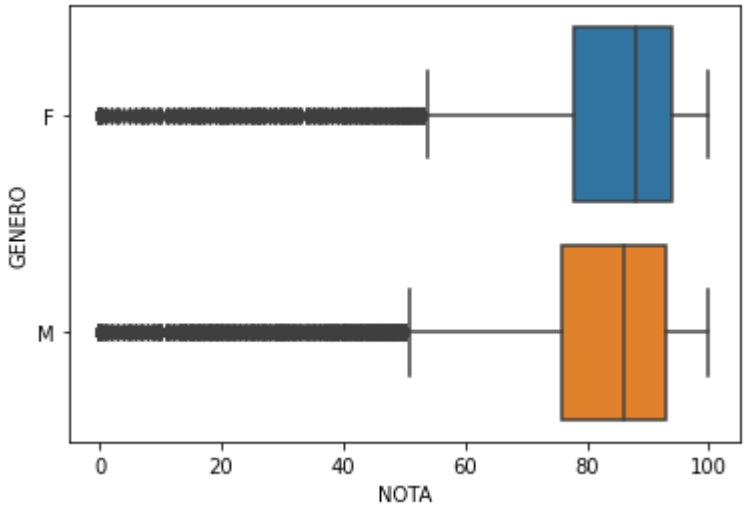
La predicción se trabajó con el campo STATUS, donde se define si un estudiante aprobó o no un curso. Se corrieron los algoritmos *Dummy Classifier* (DC), *Logistic Regression* (LR), *Random Forest* (RF) y *XGBoost* (XGB). A los campos nominales se aplicó la técnica de codificación *one-hot encoding* (Rodríguez et al. 2018): dividir los campos nominales y asignarles un 1 en los que presenta la opción y 0 donde no está presente la clasificación. Los resultados obtenidos en distintas ejecuciones son los siguientes:

Análisis por género

La primera característica que se analizó de los datos fue el género. Los de género femenino acusan mejor rendimiento académico con un promedio de 82.54 respecto al masculino que ponderó 80.4776. En la figura 3 se puede apreciar la ligera tendencia a un rendimiento académico más alto por parte del género femenino. Las desviaciones estándar son de 20.05 y 21.086 para femenino y masculino respectivamente. Según los resultados que se muestran en la tabla 3 el algoritmo con mejor *accuracy*, F1 *Score* y *Precision* no se define claramente pues presentan igual medida tanto LR, RF y XGB con 0.9201.

Gráfico 3.

Comparativa de Rendimiento por Género.



Fuente: elaboración propia.

Nota: La nota promedio según género muestra un sustancial mejor rendimiento en el género femenino respecto al masculino. Además, en ambos casos el rango intercuartil es bastante cercano. Por lo que acusan rendimientos muy similares.

Tabla 3.

Resultados de algoritmos por género.

Modelo	Accuracy	F1 Score	Precision	Recall
LR	0.9201	0.9584	0.9201	1.00
RF	0.9201	0.9584	0.9201	1.00
XGB	0.9201	0.9584	0.9201	1.00
DC	0.8513	0.9191	0.9199	0.9183

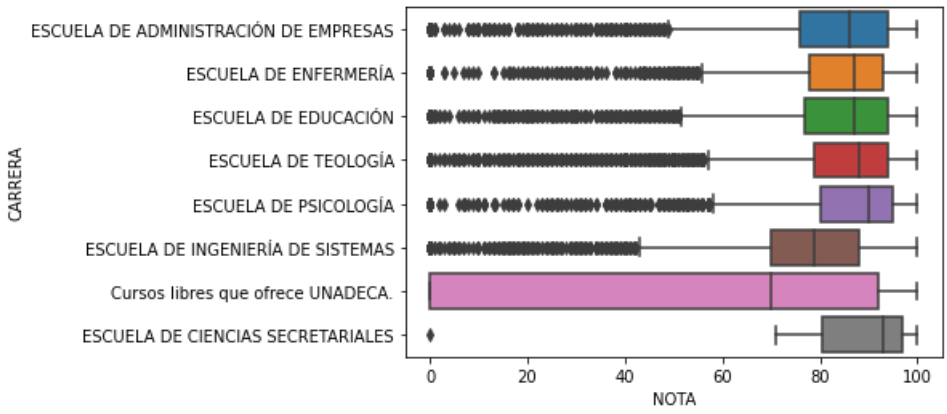
Fuente: elaboración propia.

Análisis por Escuelas

La segunda característica que se trabajó fue por escuelas. Aquí se aplicó one-hot encoding en dicha característica (resultando 8 campos adicionales). La que obtuvo el rendimiento académico promedio más alto fue la de Ciencias Secretariales con un promedio de 87.20 y que además tiene la menor desviación estándar. La figura 4 muestra los resultados del promedio por escuelas. Los resultados de los algoritmos (ver tabla 3) reflejan un Accuracy, F1 Score y Precision iguales tanto para LR, RF y XGB con 0.9194, 0.9580 y 0.9194 respectivamente. Por último, se midió cuál tenía una mejor predicción (ver figura 5) y fue la de Ciencias Secretariales.

Gráfico 4.

Promedio por escuelas.



Fuente: elaboración propia.

Nota: La nota promedio por escuelas tiene comportamientos similares.

Tabla 4.

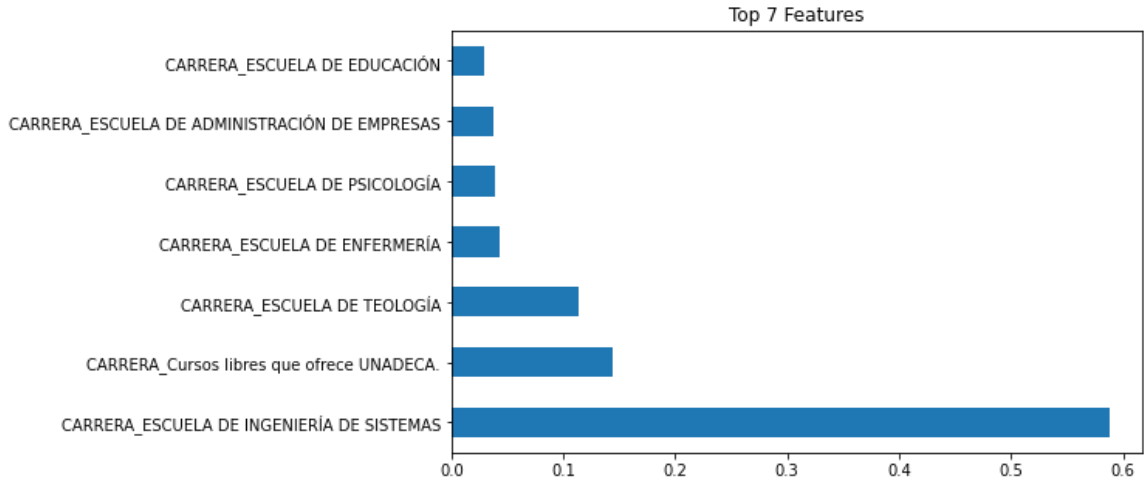
Resultados de algoritmos por escuelas.

Modelo	Accuracy	F1 Score	Precision	Recall
LR	0.9194	0.9580	0.9194	1.00
RF	0.9194	0.9580	0.9194	1.00
XGB	0.9194	0.9580	0.9194	1.00
DC	0.8520	0.9580	0.9194	0.9197

Fuente: elaboración propia.

Gráfico 5.

Características y su predicción probable.



Fuente: elaboración propia.

Análisis por países

A los distintos países de procedencia se aplicaron los algoritmos para medir su rendimiento. La universidad cuenta con una familia de 24 nacionalidades distintas (algunas con poca representatividad). Fue interesante notar que la nacionalidad con un mejor rendimiento académico fue la española con un promedio de 94.53 y una desviación estándar de 18.78. Los algoritmos LR, RF y XGB tuvieron resultados iguales tanto para *Accuracy*, F1 Score como *Precision*, con valores de 0.9226, 0.9597 y 0.9226 respectivamente (según se detalla en la tabla 5).

Tabla 5.

Resultados de algoritmos por países.

Modelo	Accuracy	F1 Score	Precision	Recall
LR	0.9201	0.9584	0.9201	1.00
RF	0.9201	0.9584	0.9201	1.00
XGB	0.9201	0.9584	0.9201	1.00
DC	0.8513	0.9191	0.9199	0.9183

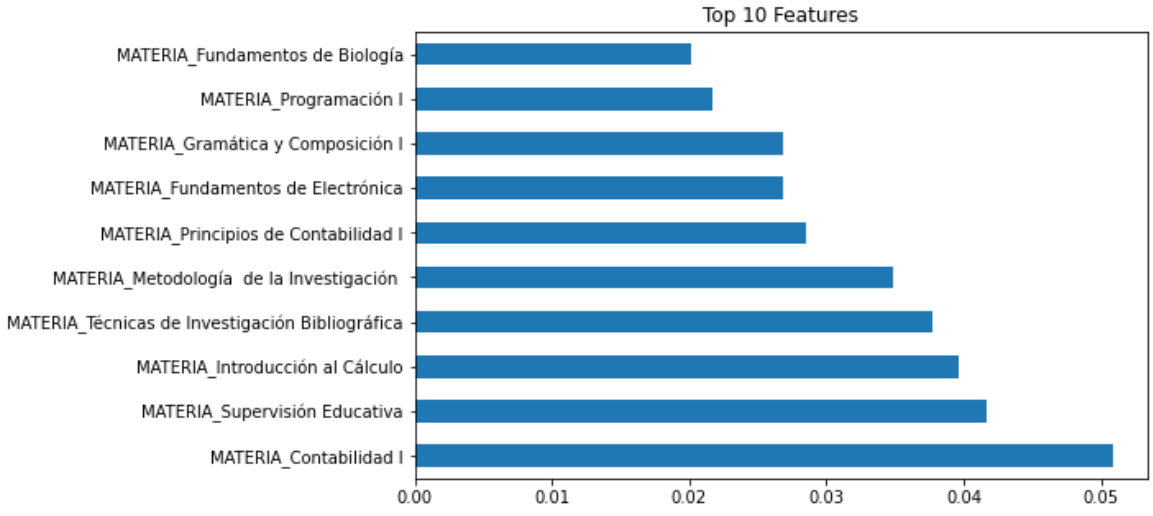
Fuente: elaboración propia.

Análisis por materias

Para probar los algoritmos por cursos, se aplicó el mismo proceso de codificación con los 600 disponibles. La Figura 6 muestra los primeros diez que presentaron una mejor predicción. Esto permitirá definir estrategias internas para trabajar en ellos. El resultado del área bajo la curva ROC (AUC) para esta característica fue de 0.69 para el algoritmo RF como el de mejor puntuación. Los resultados particulares de las distintas métricas se muestran en la tabla 6.

Gráfico 6.

Los cursos con mejor predicción.



Fuente: elaboración propia.

Tabla 6.

Resultados de algoritmos por escuelas.

Modelo	Accuracy	F1 Score	Precision	Recall
LR	0.9170	0.9567	0.9170	1.00
RF	0.9170	0.9566	0.9172	0.9996
XGB	0.9170	0.9567	0.9170	1.00
DC	0.8562	0.9219	0.9185	0.9253

Fuente: elaboración propia.

Análisis individuales por Escuelas

Con el objetivo de tener criterios más específicos por Escuelas, se corrió el algoritmo *Random Forest*, para determinar los cursos con mejores opciones de predicción y que, a la vez, debe ponerse atención especial en su rendimiento académico. Los resultados se muestran en la tabla 7. La escuela de enfermería, con un AUC de 0.76 es la que presenta un mejor posicionamiento para predicción del rendimiento en sus estudiantes.

Tabla 7.

Resultados de algoritmos por métrica y cursos con mejor predicción por escuelas.

Escuela	Resultados	Materias
Ingeniería	Data = 5 645 AC=0.8501 F1 = 0.9190 *Precision* = 0.8501 AUC = 0.72	Contabilidad 1 Introducción al cálculo Metodología de la investigación Fundamentos de Electrónica Programación 1
Enfermería	Data = 6 549 AC = 0.9351 F1 = 0.9665 *Precision* =0.9351 AUC = 0.76	Fundamentos de biología Microbiología y Parasitología Fundamentos de Enfermería 1
Psicología	Data = 5 619 AC = 0.9265 F1 = 0.9618 *Precission* = 0.9265 AUC = 0.62	Introducción a la Estadística Descriptiva Psicología de la Educación Especial Métodos y Diseños de Investigación en Psicología Clínica

Educación	Data = 12 422 AC = 0.9139 F1 = 0.9549 Precision =0.9164 AUC = 0.65	Supervisión Educativa Técnicas de Investigación Bibliográfica Solfeo y Entrenamiento Auditivo I Instrumento Complementario Gramática y Composición 1
Administración de Empresas	Data = 9 111 AC = 0.9151 F1 = 0.9557 Precision = 0.9151 AUC = 0.72	Principios de Contabilidad 1 Informática para Administradores 1 Precálculo Principios de contabilidad 2 Matemática Universitaria Básica
Teología	Data = 19 832 AC = 0.9366 F1 = 0.9673 Precision = 0.9366 AUC = 0.68	Gramática y Composición 1 Griego 1 Historia Eclesiástica 1 Técnicas de Investigación Bibliográfica

Fuente: elaboración propia.

Pruebas con RapidMiner

Se procedió a realizar pruebas en RapidMiner Studio, con el conjunto de datos, tomando como parámetro de prueba el Status (Aprobado o Reprobado), con los algoritmos de clasificación y se obtuvieron los resultados mostrados en la tabla 8. Una muestra de predicciones se ve en la Fig.7.

Tabla 8.

Pruebas de algoritmos en RapidMiner.

Algoritmo	**Métricas**
Naive Bayes,	Accuracy: 93.6% Classification Error: 6.4% AUC = 0.599 Precision = 93.6% F Measure = 96.7%
Generalized Linear Model,	Accuracy: 93.6% Classification Error: 6.4% AUC =0.614 Precision = 93.6%
Logistic Regression,	Accuracy: 93.6% Classification Error: 6.4% AUC =0.611 Precision = 93.6%
Fast Large Margin	Accuracy: 93.6% Classification Error: 6.4% AUC =0.556 Precision = 93.6%
Deep Learning	Accuracy: 93.6% Classification Error: 6.4% AUC =0.616 Precision = 93.7%
Decision Tree	Accuracy: 93.6% Classification Error: 6.4% AUC =0.5 Precision = 93.6%
Random Forest	Accuracy: 93.6% Classification Error: 6.4% AUC =0.645 Precision = 93.6% F Measure = 96.7%
Gradient Boosted Trees	Accuracy: 93.6% Classification Error: 6.4% AUC =0.648 Precision = 93.6%

Support Vector Machine	*Accuracy*: 93.6% *Classification Error:* 6.4% AUC =0.483 *Precision* = 93.6% *F Measure* = 96.7%

Fuente: elaboración propia.

Gráfico 7.

Muestra de predicciones en RapidMinner.

Fuente: elaboración propia.

4. DISCUSIÓN

Poder medir o cuantificar el avance educativo mediante el rendimiento académico es un indicador usado con frecuencia en investigaciones realizadas en latitudes diversas. Se compara qué características influyen y, en general, se asocian con el género, la edad, el personal docente y el aprendizaje de los estudiantes. En la educación y su administración, tener la capacidad de predecir el éxito académico acapara el interés. Pero ¿a qué se refiere la medida del rendimiento académico? A medir numéricamente el cumplimiento o no de los objetivos de aprendizaje inmediato y a largo plazo. De aquí se deriva la clasificación de las universidades. Se ve mejorada según tenga un historial sólido y sus logros académicos. Desde la perspectiva estudiantil, el poder mantener un rendimiento académico sobresaliente aumenta sus posibilidades de empleabilidad, al ser uno de los principales aspectos que los empleadores evalúan. (Alsariera et al., 2022) (Hasan et al., 2019) (Babu & Varghese, 2020)

Diversas técnicas se utilizaban para mantener actualizado ese registro histórico. Una de ellas, el *Data Mining*, se define como la práctica de examinar una base de datos grande

preexistente con el fin de generar información nueva (Osman, 2019). La técnica de mayor interés para el sector académico se conoce como *Educational Data Mining* (EDM), es considerada una disciplina emergente (Jawad, 2021). El objetivo es visualizar una variable predictiva a partir de un conjunto de valores conocidos como predictores de datos y así buscar resolver inquietudes netamente académicas (Vasani & Gawali, 2013). Una de sus ventajas consiste en identificar estudiantes con un lento aprendizaje, para que, desde una etapa temprana, se tomen medidas que ayuden a mejorar su rendimiento académico (Burman & Som, 2019). EDM tiene un conjunto de algoritmos que extraen patrones ocultos de los registros históricos del rendimiento académico estudiantil (Almasri et al., 2020).

En la búsqueda de otras técnicas que permitan trabajar con estas predicciones, se pueden utilizar métodos de clasificación en forma integrada como *Artificial Neural Network* (ANN), Naïve Bayes (NB), *Logistic Regression y Decision Trees* (Altabrawee et al., 2019). Otros tratados han incursionado en obtener distintas métricas para distintos grupos de datos, utilizando algoritmos *Decission Trees* (DT), Naïve Bayes (NB) y *rule-based* (RB) (Jawad, 2021). Para encontrar mejores métricas se han estudiado las redes, distintas características, determinando la correlación con el rendimiento académico (Priya et al., 2021).

Por ejemplo, (Rimadana et al., 2019) recopiló los resultados de varios estudios que mostraron características, algoritmos y resultados de rendimiento, como: factores académicos, para los algoritmos *Logistic Regression* (LR), *Multiple Regression* (MR) y *Neural Network* (NN), con un 83% de rendimiento en regresión; información personal, información de educación y geográfica, con algoritmos como DT y NN, que mostraron resultados del 97% con cuatro niveles de clasificación.

El estudio presentado por (Gull et al., 2020) mostró que el algoritmo con mejor rendimiento con 0.81 fue el *Linear Discriminant Analysis* (LDA), y el menor fue el de *Logistic Regression* (LR) con un 0.65. Este análisis se realizó con las principales características en evaluaciones (quiz1, quiz2, *midterm exam and Project*) de 250 estudiantes, además de utilizar la nota como la característica de prueba final. Otro aspecto relevante a estudiar es la decisión de los atributos estudiantiles importantes a usarse en los algoritmos de predicción (Dhilipan et al., 2021). El estudio realizado por (Rimadana et al., 2019) encontró que el resultado del Coursework es la característica más importante para la clasificación con un 93% de *accuracy* utilizando clasificación binaria.

Las estadísticas básicas del conjunto de datos muestran información relevante, por ejemplo: la nota promedio global es de 81.43; el género femenino acusa un rendimiento que supera al masculino en un poco más de dos puntos (82.54 a 80.47); la escuela con un mejor rendimiento académico es Psicología con un 83.44; de los países de América Central que son más representativos por la cantidad de datos que tienen, es Guatemala con un promedio de 84.15.

Por otro lado, apoyados en los resultados de los distintos algoritmos con la base de datos de origen, el que presentó mejores parámetros fue el RF con un *accuracy* en las distintas clasificaciones: por género un 92.01%, por escuelas un 91.70%, por países un 92.01%. Estos resultados superan los realizados por (Rimadana et al., 2019) que obtuvo un 83% para los algoritmos de clasificación. Además, se aplicaron algoritmos por escuelas específicas para determinar los cursos con mayor sensibilidad a la predicción (ver tabla 7). La escuela con un AUC más alto fue la de enfermería con un valor de 0.76 que se cataloga como una prueba buena.

5. CONCLUSIONES

Utilizando Python se pudieron comparar los algoritmos DM, LR, RF y XGBoost. Fue RF a nivel de clasificación el que tuvo los mejores resultados tanto por género, por países como por escuelas y materias. Al comparar con RapidMiner Studio, el algoritmo con una mejor predicción fue *Gradient Boosted Tree*s seguido de RF.

Los resultados de los algoritmos permitieron establecer el nivel de predicción tanto por género, por países, por escuelas y por asignaturas. Esta clasificación presenta oportunidades claras para establecer estrategias de trabajo en pro de mejorar el rendimiento académico. Así se puede focalizar por país, por género y las materias que requieren una mejor atención.

Las pruebas realizadas ayudarán a mejorar la toma de decisiones para optimizar el rendimiento académico estudiantil en UNADECA. Se identificaron cursos sensibles para cada escuela donde focalizar esos esfuerzos. En este artículo no se discuten aspectos de metodología de enseñanza ni cómo perfeccionarlos, pero sería una motivación para trabajos futuros, así como trabajar algoritmos de agrupamiento para predecir tanto rendimiento académico como la deserción estudiantil.

CONFLICTO DE INTERÉSES

Los autores declaramos no tener ningún conflicto de interés relacionado con la investigación presentada.

FINANCIACIÓN

Para completar la presente investigación, no se requirió ninguna fuente de financiación.

REFERENCIAS BIBLIOGRÁFICAS

A Alférez, G. H., Esteban, O. A., Clausen, B. L., & Ardila, A. M. M. (2022). Automated machine learning pipeline for geochemical analysis. *Earth Science Informatics, 15*(3), 1683-1698. https://doi.org/10.1007/s12145-022-00821-8

Almasri, A., Alkhawaldeh, R. S., & Çelebi, E. (2020). Clustering-Based EMT Model for Predicting Student Performance. *Arabian Journal for Science and Engineering*, *45*(12), 10067–10078. https://doi.org/10.1007/s13369-020-04578-4

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/4151487

Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting Students' Performance Using Machine Learning Techniques. *JOURNAL OF UNIVERSITY OF BABYLON for Pure and Applied Sciences*, *27*(1), 194–205. https://doi.org/10.29196/jubpas.v27i1.2108

Babu, C., & Varghese, R. (2020). *Predicting Student' s Performance Using Educational Data Mining. 9*(1).

Burman, I., & Som, S. (2019). Predicting Students Academic Performance Using Support Vector Machine. *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, 756–759. https://doi.org/10.1109/AICAI.2019.8701260

Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. *IOP Conference Series: Materials Science and Engineering*, *1055*(1), 012122. https://doi.org/10.1088/1757-899x/1055/1/012122

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Gull, H., Saqib, M., Iqbal, S. Z., & Saeed, S. (2020). Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning. *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, 1–4. https://doi.org/10.1109/INOCON50539.2020.9298266

Hasan, H. M. R., Shahariar, A., & Rabby, A. (n.d.). Hasan, H. M. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019). Machine Learning Algorithm for Student's Performance Prediction. *In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). Kanpur, India. https://doi.org/10.1109/ICCCNT45670.2019.8944629

Hasterok, D., Gard, M., Bishop, C. M. B., & Kelsey, D. (2019). Chemical identification of metamorphic protoliths using machine learning methods. *Computers and Geosciences*, *132* (March), 56–68. https://doi.org/10.1016/j.cageo.2019.07.004

Jawad, S. M. D. (2021). Student Performance Analysis System Using Machine Learning. *Journal of Resource Management and Technology, 12*(84), 84–90.

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). Cancer Classification Using Gaussian Naive Bayes Algorithm. *Proceedings of the 5th International Engineering Conference, IEC 2019*, 165–170. https://doi.org/10.1109/IEC47844.2019.8950650

Marius-Constantin, P., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), 579–588.

Mierswa, Igno, and K. (2022). *RapidMiner*. RapidMiner Studio. https://rapidminer.com/

Osman, A. S. (2019). Data mining techniques: Review. *International Journal of Data Science Research*, *2*(1), 1–4.

Priya, S., Ankit, T., & Divyansh, D. (2021). Student performance prediction using machine learning. *Advances in Parallel Computing*, *39*(03), 167–174. https://doi.org/10.3233/APC210137

Rimadana, M. R., Kusumawardani, S. S., Santosa, P. I., & Erwianda, M. S. F. (2019). Predicting Student Academic Performance using Machine Learning and Time Management Skill Data.

2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019, 511–515. https://doi.org/10.1109/ISRITI48646.2019.9034585

Vasani, V. P., & Gawali, R. D. (2013). Classification Performance Evaluation. *Encyclopedia of Systems Biology*, *3*(3), 411–411. https://doi.org/10.1007/978-1-4419-9863-7_100213