Predicción de desempeño académico basado en análisis de datos usando técnicas de clasificación: un estudio con datos de una universidad pública de Ecuador.

Luis Patiño Hernández, UPEC/UNIJUÍ,
luis.patinio@upec.edu.ec, https://orcid.org/0000-0001-5792-4122
Antonia María Reina Quintero, Universidad de Sevilla,
reinaqu@us.es, https://orcid.org/0000-0003-3698-6302
Sandro Sawicki, Rafael Z. Frantz, Fabricia Roos-Frantz, UNIJUÍ
sawicki@unijui.edu.br, https://orcid.org/0000-0002-7960-0775,
frfrantz@unijui.edu.br, https://orcid.org/0000-0001-9514-6560,
rzfrantz@unijui.edu.br, https://orcid.org/0000-0003-3740-7560

Abstract: *A pressing concern of universities is to improve the academic performance of their students, as poor performance can lead to student dropout. This article aims to analyze university academic data, to verify if it is possible to predict the performance of students before starting the academic year. To do this, educational data mining is used, analyzing not only performance data but also demographic data. The results show that it is possible to make early predictions of academic performance, finding an important percentage of prediction, 10.28 % students failed their academic level, the LightGBM and XGBoost algorithms based on the set of attributes obtained through the Correlation-Based Attribute Selection method obtained the best metrics.*

Keywords: data analysis, academic performance, early prediction.

Resumo: *Una preocupación constante de las universidades es mejorar el desempeño académico de sus estudiantes, ya que un bajo desempeño puede llevar a la deserción estudiantil. Este artículo tiene como objetivo analizar datos académicos universitarios, para verificar si es posible predecir el desempeño de los estudiantes antes de iniciar el ciclo académico. Para ello, se hace uso de la minería de datos educativos, analizando no solamente datos de desempeño sino también datos demográficos. Los resultados muestran que es posible hacer predicciones tempranas de desempeño académico. En este estudio hemos encontrado que el 10.28 % de estudiantes perdieron el nivel. Los algoritmos LightGBM y XGBoost con base en el conjunto de atributos obtenido mediante el método de Selección de Atributos Basado en Correlación obtuvieron las mejores métricas.*

Palabras clave: análisis de datos, desempeño académico, predicción temprana.

1. Introducción

Una preocupación constante de las universidades es mejorar el desempeño académico de sus estudiantes, entendiéndose por desempeño académico el grado de conocimientos obtenido a través del sistema educativo y se expresa por medio de una calificación asignada por el profesor (LOMELÍ, 2002). Un bajo desempeño puede llevar a la deserción estudiantil, que es considerada actualmente una de las principales preocupaciones de las instituciones universitarias porque trae consigo implicaciones de nivel social y económico (MARQUES *et al.*, 2019). Según el sitio web del Instituto Brasileño de Estudios e Investigación Educativa Anísio Teixeira (INEP), en el año 2020 la tasa de deserción en las universidades públicas brasileñas fue próxima al 40 %. En las universidades ecuatorianas, según la UNESCO, el

porcentaje de deserción también fue de un 40 % en el año 2021. Por lo que es importante que las instituciones educativas tomen acciones que permitan reducir estos porcentajes.

En el ámbito educativo, la predicción temprana del desempeño posibilita que profesores y gestores puedan tomar medidas preventivas para mejorar el rendimiento estudiantil reforzando el aprendizaje y, en consecuencia, aumentando las posibilidades de aprobar el nivel de estudios (HOFFAIT; SCHYNS, 2017). Para realizar actividades predictivas se requiere de un volumen considerable de datos de cualquier área, difícil de manejar a través de herramientas tradicionales. Además, muchos de estos datos necesitan de un tratamiento especial (corregir, formatear, eliminar). En su mayoría, los conjuntos de datos poseen muchos atributos ante lo cual es necesario verificar su relación con la variable objetivo con el fin de incluir solo aquellos que tengan relación directa con los resultados de predicción más exactos. La Minería de Datos Educativos (MDE) se presenta como una alternativa para crear modelos de predicción temprana de desempeño académico (COSTA *et al.*, 2017).

La Universidad Politécnica Estatal del Carchi (UPEC)[1], universidad ecuatoriana de tipo pública, presenta altas tasas de deserción, principalmente en las carreras de ingeniería. No existen herramientas para identificar desde el inicio del ciclo académico a los estudiantes con problemas de desempeño académico para posteriormente dar tutorías académicas o estrategias didácticas personalizadas que permitan mejorar su desempeño. Muchas de estas estrategias están orientadas a estudiantes que no lo necesitan. Los que tienen problemas, por desconocimiento, no tienen un seguimiento por parte del docente o gestores educativos.

Este artículo tiene como objetivo analizar datos de desempeño y demográficos para la identificación temprana de estudiantes con bajo desempeño en diferentes carreras. Para ello, generamos modelos predictivos evaluados por sus buenas tasas de acierto. Los datos que utilizamos son de estudiantes de primer nivel de estudios de las 9 carreras en modalidad presencial y son parte del Sistema Integrado de Información, módulo académico de la UPEC.

El aporte principal de este artículo es la definición de atributos que contribuyen de manera significativa en la predicción temprana de desempeño académico junto con algoritmos que alcanzaron mejores tasas de acierto utilizando datos académicos de la modalidad presencial de una universidad ecuatoriana. Hacemos un análisis para realizar predicciones tempranas, determinando cual de los dos conjuntos de datos, desempeño o demográfico, ofrece mejores resultados. Hemos encontrado que es posible realizar predicciones tempranas con base en el conjunto de datos demográficos, donde se obtiene un porcentaje importante de predicción, 10.28 % de estudiantes que perdieron el nivel. Los mejores modelos corresponden a los algoritmos LightGBM, XGBoost y fueron evaluados mediante las métricas: exactitud, especificidad, matriz de confusión, curva ROC y AUC ROC.

El resto del artículo se estructura de la siguiente manera: en la Sección 2 se describen los trabajos relacionados; en la Sección 3 se presenta la metodología usada para el desarrollo del trabajo; y, finalmente, la Sección 4 concluye el artículo.

2. Trabajos relacionados

En esta sección discutimos algunos de los trabajos relacionados.

Costa *et al.* (2017) desarrollan un estudio comparativo de las técnicas de MDE en la identificación temprana de estudiantes con probabilidades de reprobar el curso de introducción a la programación en una universidad pública brasileña en las modalidades presencial

[1] https://upec.edu.ec

y distancia. Nuestro estudio es más amplio: se desarrollan predicciones tempranas de desempeño para estudiantes de primer período utilizando notas de todas las materias de los primeros niveles.

Meghji *et al.* (2023) utilizando técnicas de clasificación y métodos de selección de atributos, predicen el rendimiento de fin de carrera de estudiantes de ingeniería de software en una etapa temprana durante su transcurso, para ello utilizan registros institucionales y datos de expedientes individuales del estudiante. Para evaluar los modelos utilizan las métricas, exactitud, F1-Score y Kappa. En el presente estudio se aplican técnicas de clasificación y métricas casi similares a diferencia del conjunto de datos que abarca diferentes carreras universitarias.

Hoffait and Schyns (2017) utilizando técnicas de clasificación identifican de manera temprana estudiantes de primer año que enfrentan dificultades para completar su año académico, utilizan características individuales, indicadores de desempeño pasados y algunos factores ambientales, datos correspondientes a tres años académicos. Cómo métricas utilizan la exactitud y la matriz de confusión. En nuestro trabajo se hace uso de datos individuales, demográficos y desempeño correspondientes a diez períodos académicos también del primer período de estudios.

Alhazmi and Sheneamer (2023) utilizan técnicas de agrupamiento y clasificación para predecir el desempeño de estudiantes en una etapa temprana, se utilizan datos individuales, notas de admisión, también notas de pruebas, rendimiento académico y aptitud general, datos correspondientes a estudiantes de una facultad de informática. Para validar el modelo utilizan las métricas, exactitud, precisión, recall, F1-score y matriz de confusión.

3. Metodología

En el desarrollo de este estudio usamos el proceso de descubrimiento de conocimiento en bases de datos (DCBD), donde la minería de datos se constituye en una de sus etapas.

3.1. Selección de datos

Se han seleccionado datos del módulo académico que guarda información de las nueve carreras de la UPEC, siendo ellas: Alimentos, Turismo, Agropecuaria, Computación, Enfermería, Administración de Empresas, Administración Pública, Comercio Exterior, Logística y Transporte. En este módulo se administran las calificaciones y los resultados de desempeño de los estudiantes en cada materia del nivel, informando cuales aprobaron, no aprobaron o tienen que rendir exámenes supletorios (recuperación).

El conjunto inicial de datos utilizado en este estudio tiene 145.444 registros y 18 atributos, que contienen la siguiente información: datos personales, demográficos y de desempeño de estudiantes matriculados en las nueve carreras, en los diez niveles. Estos datos corresponden a diez ciclos académicos comprendidos entre los años 2016 hasta 2022. De ese conjunto, fueron seleccionados los datos estudiantiles del primer nivel, ya que según los autores Delen (2010) y Herzog (2005) y por experiencia de la mayoría de carreras de la institución, de este nivel desertan más.

Con base al objetivo de este estudio, el conjunto de datos fue divido en dos: demográfico y desempeño.

- El conjunto de datos demográfico contiene los siguientes datos: código del ciclo académico, nombre de la carrera, cédula del estudiante, nombre del estudiante, lugar

de procedencia, número de materias cursadas, estado civil, nacionalidad, género, fecha de nacimiento, lugar de nacimiento y si ha aprobado el nivel. También se incluye, ingenieria, un código que distingue a las carreras de ingeniería de las licenciaturas. Según el reglamento, el estudiante que reprueba en más de dos materias, reprueba el nivel.

- El conjunto de datos de desempeño guarda información de las materias por cada estudiante: código del ciclo académico, nombre de la carrera, nombre de la materia, cédula del estudiante, nombre del estudiante, número de matrícula en la materia, nota1, nota2, nota3, nota final y si ha aprobado la materia o no.

3.2. Preprocesamiento

Para cada conjunto de datos realizamos la limpieza de datos porque existían valores nulos y registros duplicados, fue necesario eliminarlos o corregirlos utilizando Microsoft Excel. En el conjunto de datos demográfico realizamos ajustes a los datos del lugar de procedencia para estandarizarlos. Los nombres próximos a una ciudad capital o referente de provincia toman el mismo con el fin de reducir el número de nombres y formar menos grupos, por ejemplo: si tenemos estudiantes con lugares de procedencia El-Sagrario-Ibarra o San-Francisco-Ibarra, usamos el nombre estándar, Ibarra, que corresponde a una capital porque estos lugares pertenecen al área urbana de esta ciudad.

3.3. Transformación

Con los conjuntos de datos definidos procedimos a seleccionar sus atributos por medio de métodos de selección, ya que algunos atributos no contribuyen de forma significativa para la predicción de desempeño, pero si contribuyen al gasto de recursos desmesuradamente (FRANCO *et al.*, 2020). Los métodos utilizados, de acuerdo con (JANABI; KADHIM, 2018), disponibles en la herramienta Weka, fueron: Selección de Atributos Basado en Correlación(CFS), Chi-cuadrado (ChiSquared) y Ganancia de Información (IG).

3.3.1. Selección de Atributos Basada en Correlación

Este método calcula la correlación entre todos los atributos y la clase de salida y selecciona el mejor subconjunto de atributos utilizando la función de evaluación heurística basada en correlación. A continuación, se muestra la ecuación:

$$r_{zc} = \frac{k\bar{r_{zi}}}{\sqrt{k + k(k-1)\bar{r_{ii}}}} \tag{1}$$

Donde:

r_{zc} es la correlación (dependencia) entre los atributos y la variable de clase, k es el número de atributos, r_{zi} es el promedio de correlación entre el atributo clase y r_{ii} es el promedio de correlación entre atributos.

3.3.2. Selección de Atributos Basada en Chi-cuadrado

Este método es utilizado para atributos categóricos en un conjunto de datos. Se calcula el Chi-cuadrado entre cada atributo y la clase de salida, luego se seleccionan el número deseado de atributos con las mejores puntuaciones. A continuación, se muestra la ecuación:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

Donde, r es el número de intervalos que hay que comparar entre sí, c es el número de clases que hay, O_{ij} es la frecuencia observada y E_{ij} es la frecuencia esperada.

3.3.3. Selección de Atributos Basada en Ganancia de información

El método ganancia de información calcula la información recogida por la medida de entropía del atributo dado con relación a la clase.

$$GananciadeInfomacion(Clase, Atributo) = H(Clase) - H(Clase|Atributo) \quad (3)$$

Donde, H es el valor de la entropía (medida que cuantifica el desorden de un sistema).

3.3.4. Aplicación de los métodos de selección de atributos

Los tres métodos de selección descritos fueron aplicados a los dos conjuntos de datos utilizando el software Weka versión 3.6.2. El resultado de aplicar este proceso de selección de atributos se muestra en los Cuadros 1 y 2. Luego del proceso de selección de atributos, procedimos a transformar los datos de cada conjunto. Los de tipo nominal fueron cambiados a tipo numérico, por ejemplo, cada nombre de carrera fue sustituido por un número que la identifica. El motivo de este cambio se debe a que algunos algoritmos de clasificación seleccionados requieren datos numéricos. Una descripción de los atributos seleccionados por cada conjunto se muestra en los Cuadros 3 y 4.

Cuadro 1. Selección de atributos para el conjunto de datos de desempeño

Método de Búsqueda	Método de selección	Atributos seleccionados*	Atributos excluídos
GreedyStepwise/BestFirst	CFS	numero_matricula, nota1	ciclo_academico, cedula, carrera, materia, nivel_matricula, nota2, nota3
Ranker	ChiSquared	cedula, nota1, materia, carrera	ciclo_academico, numero_matricula, nivel_matricula, nota2, nota3,
Ranker	IG	cedula, nota1, materia, carrera	ciclo_academico, numero_matricula, nivel_matricula, nota2, nota3

* Del conjunto de atributos obtenido inicialmente, también fueron excluidos los atributos nota2 y nota3, ya que están relacionados directamente con la variable de resultado (Aprobo) y pueden causar un sesgo.

Cuadro 2. Selección de atributos para el conjunto de datos demográfico

Método de Búsqueda	Método de selección	Atributos seleccionados	Atributos excluídos
GreedyStepwise/BestFirst	CFS	cedula, numero_materias	carrera, lugar_procedencia, ingenieria, edad, genero, estado_civil, nacionalidad
Ranker	ChiSquared	cedula, numero_materias, carrera, lugar_procedencia, ingenieria	edad, genero, estado_civil, nacionalidad
Ranker	IG	cedula, numero_materias, carrera, lugar_procedencia, ingenieria	edad, genero, estado_civil, nacionalidad

Cuadro 3. Atributos del conjunto de datos de desempeño

Nombre	Descripción	Tipo de dato
cedula	Código de identificación del estudiante	Nominal
numero_matricula	Número de veces que el estudiante toma la misma materia	Numérico
nota1	Primera calificación del semestre registrada en el sistema	Numérico
materia	Nombre de la materia que estudia	Nominal
carrera	Nombre de la carrera que está matriculado	Nominal

Cuadro 4. Atributos del conjunto de datos demográfico

Nombre	Descripción	Tipo de dato
cedula	Código de identificación del estudiante	Nominal
numero_materias	Número de materias que el estudiante toma en el semestre	Numérico
carrera	Nombre de la carrera que está matriculado	Nominal
lugar_procedencia	Lugar donde ha vivido o vive el estudiante	Nominal
ingeniería	1 para carreras de ingeniería, 0 para carreras de licenciatura	Numérico

3.4. Minería de datos

En esta etapa utilizamos algoritmos de clasificación para generar los modelos de predicción del desempeño académico y posteriormente son evaluados para determinar el de mejor desempeño. Fueron seleccionados 4 algoritmos de la literatura (SORGATTO *et al.*, 2021; FRANCO *et al.*, 2020; NETO; VASCONCELOS; ZANCHETTIN, 2021; FONSECA *et al.*, 2019), con base en las mejores tasas de acierto siendo ellos: Random Forest, Naive Bayes, LightGBM y XGBoost. Con la herramienta Google Colaboratory[2] usamos el lenguaje Python para escribir el código de predicción.

Para generar modelos, los algoritmos para cada conjunto ya sea demográfico o desempeño necesitan dos conjuntos de datos distintos: entrenamiento y prueba, para estos conjuntos necesitamos definir el conjunto de atributos y sus datos. Por ello utilizamos los atributos definidos mediante los métodos selección de atributos, descrito en la sección anterior. El proceso de selección de atributos obtuvo dos tipos de conjuntos, uno mediante el método de selección CFS y el segundo mediante los métodos Chi-cuadrado e IG, porque obtuvieron el mismo resultado. Así, los Cuadros 1 y 2 muestran los conjuntos de atributos resultantes. Una vez definidos los atributos vino la conformación de datos.

El conjunto de prueba consta de datos de 5 carreras, 4 de ellas de ingeniería y una licenciatura, siendo ellas: Logística y Transporte, Computación, Agropecuaria, Alimentos y Administración de Empresas. Esta selección se justifica en base a la información de la tasa de retención de dirección académica UPEC para los años 2018 a 2020, en donde las carreras de ingeniería presentaron índices de retención escolar bajos con respecto a la base establecida por el organismo de evaluación de educación superior que es de 80 %. En las carreras de licenciatura la tasa de retención es más próxima a la base, por ello se selecciona una carrera. Estos datos son de estudiantes de primer nivel del ciclo académico Abril 2022 - Agosto 2022 . Se detalla el número de registros para cada carrera en el Cuadro 5 para los dos conjuntos de prueba.

El conjunto de entrenamiento es conformado con datos de las 9 carreras. Contiene 4140 registros de datos demográficos y 25648 registros para datos de desempeño, correspondientes a los ciclos académicos comprendidos entre los años 2016 hasta 2020 de estudiantes del primer nivel.

[2]https://research.google.com/colaboratory/faq.html

Cuadro 5. Número de registros por carrera del conjunto de prueba

Carrera	Tipo	Nro Reg (Demog)	Nro Reg (Desemp)
Logística y Transporte	Ingeniería	46	251
Computación	Ingeniería	72	337
Agropecuaria	Ingeniería	67	343
Alimentos	Ingeniería	60	320
Administración de Empresas	Licenciatura	66	377
Total		**311**	**1628**

Una vez generados los modelos, se obtuvo alrededor de 16 modelos de predicción que fueron evaluados utilizando las siguientes métricas: exactitud, especificidad, matriz de confusión, curva ROC y área bajo la curva (AUC ROC), las dos últimas fueron consideradas por el desbalanceamento de datos. Para el cálculo de las métricas exactitud y especificidad, se hizo uso de las matrices de confusión.

En la evaluación, por cada conjunto de datos se tomó como referencia los mejores promedios de cada métrica. El conjunto de datos demográfico obtuvo mejores resultados utilizando el conjunto de atributos obtenido por el método CFS. El Cuadro 6 muestra que en las métricas, exactitud, especificidad y AUC ROC, los algoritmos, LightGBM y XGBoost alcanzan las mayores tasas de acierto.

Cuadro 6. Métricas con datos demográficos, método de selección de atributos CFS

Exactitud				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	**XGBoost**
Computación	0,819	0,833	0,819	0,819
Alimentos	0,816	0,850	0,866	0,866
Agropecuaria	0,567	0,567	0,641	0,641
Logística y Transporte	0,804	0,760	0,782	0,782
Administración de Empresas	0,924	0,954	0,969	0,969
Promedio	**0,786**	**0,793**	**0,815**	**0,815**
Tasa de especificidad				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	**XGBoost**
Computación	0,416	0,444	0,375	0,375
Alimentos	0,642	0,800	1,000	1,000
Agropecuaria	0,640	0,640	0,928	0,928
Logística y Transporte	0,555	0,428	0,500	0,500
Administración de Empresas	0,600	0,750	0,857	0,857
Promedio	**0,571**	**0,612**	**0,732**	**0,732**
AUC ROC				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	**XGBoost**
Computación	0,673	0,614	0,613	0,613
Alimentos	0,665	0,740	0,708	0,708
Agropecuaria	0,524	0,620	0,620	0,620
Logística y Transporte	0,858	0,755	0,849	0,849
Administración de Empresas	0,886	0,893	0,893	0,893
Promedio	**0,721**	**0,724**	**0,737**	**0,737**

Los resultados obtenidos con el conjunto de desempeño se muestran en el Cuadro 7 y corresponden al método de selección CFS. En el análisis de las métricas exactitud y AUC ROC se verifica que el algoritmo que alcanza la mayor tasa de acierto es LightGBM. Las mejores tasas obtenidas para la métrica especificidad son del algoritmo XGBoost.

En el análisis de curvas ROC con datos demográficos, se verificó que en su mayoría los modelos presentan un acercamiento a la esquina superior izquierda del eje Y, están por encima de la línea imaginaria trazada entre el punto (0,0) y (1,1) lo que prueba que son

Cuadro 7. Métricas con datos de desempeño, método de selección de atributos CFS

Exactitud				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	XGBoost
Computación	0,836	0,848	0,857	0,851
Alimentos	0,800	0,850	0,846	0,840
Agropecuaria	0,749	0,749	0,766	0,758
Logística y Transporte	0,872	0,884	0,896	0,896
Administración de Empresas	0,880	0,885	0,883	0,885
Promedio	**0,827**	**0,843**	**0,850**	**0,846**
Tasa de especificidad				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	**XGBoost**
Computación	0,634	0,741	0,781	0,750
Alimentos	0,694	0,933	0,931	0,950
Agropecuaria	0,831	0,831	0,857	0,872
Logística y Transporte	0,700	0,787	0,823	0,866
Administración de Empresas	0,600	0,777	0,750	0,777
Promedio	**0,692**	**0,814**	**0,828**	**0,843**
AUC ROC				
Carrera	**Naïve Bayes**	**Random Forest**	**LightGBM**	**XGBoost**
Computación	0,787	0,782	0,816	0,813
Alimentos	0,856	0,873	0,885	0,883
Agropecuaria	0,856	0,872	0,893	0,890
Logística y Transporte	0,900	0,912	0,917	0,921
Administración de Empresas	0,762	0,760	0,747	0,750
Promedio	**0,832**	**0,840**	**0,852**	**0,851**

buenos modelos. Las curvas ROC de las carreras de Computación y Agropecuaria están más cerca de la línea imaginaria, indicando que los modelos tienen un desempeño aceptable.

En el análisis de curvas ROC con datos de desempeño la mayoría de los modelos presentan un acercamiento a la esquina superior izquierda del eje Y, están por encima de la línea imaginaria trazada entre el punto (0,0) y (1,1) lo que demuestra que son buenos modelos. Esto se verifica con los valores de la métrica AUC ROC en el Cuadro 7.

Con la finalidad de aclarar el uso de cada métrica se realiza una descripción de cada una de ellas de acuerdo con Ferreira (2022) y Junior *et al.* (2022).

Exactitud: evalúa con qué frecuencia las predicciones realizadas por el clasificador son correctas. Como ejemplo se calcula la exactitud para la carrera de Administración de Empresas correspondiente al algoritmo Naive Bayes con datos de desempeño, tenemos:

$$exactitud = \frac{(VP + VN)}{(VP + FP + FN + VN)} \tag{4}$$

Dónde :

VP = Corresponde a los valores, Verdadero positivo, VN = Verdadero negativo, FP = Falso positivo, FN = Falso negativo.

Reemplazando los valores tenemos:

$$exactitud = \frac{(323 + 9)}{(323 + 6 + 39 + 9)} = \frac{332}{377} = 0{,}880 \tag{5}$$

El valor de 0.880 representa la tasa de acierto del algoritmo con respecto a los estudiantes que aprobaron y reprobaron materias, la tasa obtenida en el ejemplo es muy buena.

Especificidad: mide la tasa de verdaderos negativos en relación al total de negativos. Esta métrica fue seleccionada para verificar cual algoritmo tiene mejor predicción de la clase negativa.

$$especificidad = \frac{(VN)}{(FP + VN)} \tag{6}$$

Como ejemplo se calcula la especificidad para la carrera de Agropecuaria correspondiente al algoritmo LightGBM con datos demográficos, tenemos:

$$especificidad = \frac{(13)}{(1 + 13)} = 0{,}929 \tag{7}$$

El valor de 0.929 representa la tasa de acierto del algoritmo respecto a los estudiantes que no aprobaron el nivel, la tasa obtenida en el ejemplo es muy buena.

Matriz de confusión: permite visualizar el desempeño del clasificador. Cada columna de la matriz representa el número de valores estimados de cada clase, mientras que cada fila representa a las instancias en la clase real. Ver el Cuadro 8.

Cuadro 8. Matriz de confusión

		Valor estimado	
		Si	**No**
Real	**Si**	Verdadero Positivo (VP)	Falso Negativo (FN)
	No	Falso Positivo(FP)	Verdadero Negativo (VN)

Curva ROC: es una representación gráfica que describe el desempeño de un sistema clasificador binario. Muestra la relación entre la tasa de verdaderos positivos (VP) y la tasa de falsos positivos (FP).

Área bajo la curva (AUC) ROC: cuánto más próximo a 1.0 es el área bajo la curva ROC, mejor es el modelo obtenido por el clasificador.

En la siguiente etapa se realiza la interpretación de los modelos encontrados con base en los mejores desempeños de los algoritmos aplicado a sus conjuntos de datos.

3.5. Interpretación de datos

En esta etapa se interpretan los resultados alcanzados con el uso de los conjuntos de datos, demográfico y desempeño. El objetivo es verificar si es posible hacer predicciones tempranas. Para esto usamos la métrica especificidad que guarda relación con el objetivo descrito anteriormente.

Según los resultados de la métrica especificidad del conjunto de datos demográfico, hemos encontrado que los algoritmos LighGBM y XGBoost alcanzaron las mejores tasas de acierto de verdaderos negativos (VN) que es de interés en este estudio y corresponden al valor acertado de estudiantes que no aprobaron el nivel, por ello se procedió a analizar los valores VN y FN de la matriz de confusión con el fin de determinar el comportamiento de los algoritmos frente a los valores negativos (estudiantes que no aprobaron el nivel). Se tiene:

VN = 3 + 7 + 13 + 3 + 6 = 32 (acertado - estudiantes que no aprobaron el nivel)
FN = 8 + 8 + 23 + 7 + 1 = 47 (no acertado - estudiantes que no aprobaron el nivel)

Total = 79 estudiantes que no aprobaron el nivel

Considerando el total de estudiantes, 311, matriculados en primer nivel de las cinco carreras, los 79 estudiantes corresponden al porcentaje de 25.40 %. Según los autores Hoffait and Schyns (2017) estos estudiantes en su mayoría son propensos a desertar. Según la tasa de deserción permitida por el organismo evaluador en Ecuador que es de hasta el 20 %, el porcentaje de 25.40 %, excede en 5.40 %, este valor muestra que tienen un problema. El porcentaje acertado de estudiantes corresponde al 10.28 % con relación al total, lo cual se considera un valor importante para la institución.

Con datos de desempeño y la métrica especificidad el algoritmo XGBoost en la mayoría de casos obtiene los mejores porcentajes, de la misma manera se procede a analizar los valores de VN y FN de la matriz de confusión.

VN = 24 + 38 + 96 + 26 + 7 = 191 (acertado - estudiantes que no aprobaron materias)
FN = 42 + 49 + 69 + 22 + 41 = 223 (no acertado - estudiantes que no aprobaron materias)
Total = 414 casos de estudiantes que no aprobaron alguna materia

Considerando el total de matrículas de estudiantes en las diferentes materias de las cinco carreras que es de 1628, se obtiene el porcentaje de 25.40 % que corresponde a los 414 estudiantes que no aprobaron las materias y que tienen alta probabilidad de desertar. Sin embargo no es posible establecer el número exacto de estudiantes que pueden perder el nivel.

4. Conclusiones

En los modelos de predicción generados en base al conjunto de datos demográfico y desempeño, los conjuntos de atributos obtenidos por el método CFS obtuvieron mejores resultados con base al promedio de las métricas de evaluación, siendo los algoritmos LightGBM y XGBoost los mejores. Comparando los promedios de las métricas de evaluación de los mejores modelos de predicción de desempeño académico, el modelo basado en datos de desempeño presenta tasas mayores en las 3 métricas que el modelo basado en datos demográficos, sin embargo el segundo permite obtener un porcentaje importante de predicción temprana (10.28 %) desde el inicio del ciclo académico, siendo posible determinar los estudiantes que no conseguirían aprobar el nivel. Este resultado se considera importante para la institución ya que si se toman las medidas correctivas adecuadas sería posible bajar la tasa de deserción estudiantil al nivel establecido por el organismo de evaluación de educación superior. En el análisis del conjunto de atributos obtenido por el método CFS con respecto a los datos demográficos se determina que el atributo, número_materias, es el más determinante en la predicción, restando importancia a los atributos de lugar_procedencia, estado_civil, nacionalidad que tienen más relación con la parte demográfica. Para trabajos futuros tenemos pendiente analizar las ventajas de utilizar los conjuntos de atributos obtenidos por los métodos CFS, ChiSquared e InfoGain en el desarrollo de modelos predictivos con datos demográficos ya que las diferencias encontradas entre sus métricas no fueron significativas.

Agradecimientos

Este trabajo ha sido financiado a través del convenio de colaboración científica y cultural entre la Universidad Politécnica Estatal del Carchi (UPEC) de Ecuador y la Universidad Regional del Noroeste del Estado de Río Grande del Sur (UNIJUÍ) de Brasil. Este trabajo

fue parcialmente apoyado por el Consejo Nacional de Desarrollo Científico y Tecnológico de Brasil (CNPq) en el marco de las siguientes subvenciones para proyectos 309425/2023-9, 402915/2023-2 y 311011/2022-5.

Referencias

ALHAZMI, E.; SHENEAMER, A. Early predicting of students performance in higher education. *IEEE Access*, IEEE, v. 11, p. 27579–27589, 2023.

COSTA, E. B. *et al.* Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior*, Elsevier, v. 73, p. 247–256, 2017.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, Elsevier, v. 49, n. 4, p. 498–506, 2010.

FERREIRA, G. G. *Medidas de avaliação de classificadores binários para classes desbalanceadas*. Tese (Doutorado) — Universidade de São Paulo, 2022.

FONSECA, S. C. *et al.* Adaptaçao de um método preditivo para inferir o desempenho de alunos de programaç ao. 2019.

FRANCO, J. J. *et al.* Usando mineraçao de dados para identificar fatores mais importantes do enem dos últimos 22 anos. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, 2020. p. 1112–1121.

HERZOG, S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in higher education*, Springer, v. 46, p. 883–928, 2005.

HOFFAIT, A.-S.; SCHYNS, M. Early detection of university students with potential difficulties. *Decision Support Systems*, Elsevier, v. 101, p. 1–11, 2017.

JANABI, K. A.; KADHIM, R. Data reduction techniques: a comparative study for attribute selection methods. *International Journal of Advanced Computer Science and Technology*, v. 8, n. 1, p. 1–13, 2018.

JUNIOR, G. B. V. *et al.* Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. *Revista CPAQV–Centro de Pesquisas Avançadas em Qualidade de Vida— Vol*, v. 14, n. 2, p. 2, 2022.

LOMELÍ, D. G. *El desempeño académico universitario: variables psicológicas asociadas*: Secretaría de Educación Pública, 2002.

MARQUES, L. T. *et al.* Mineraçao de dados auxiliando na descoberta das causas da evasao escolar: Um mapeamento sistemático da literatura. *RENOTE*, v. 17, n. 3, p. 194–203, 2019.

MEGHJI, A. F. *et al.* Early detection of student degree-level academic performance using educational data mining. *PeerJ Computer Science*, PeerJ Inc., v. 9, p. e1294, 2023.

NETO, M. V. G.; VASCONCELOS, G. C.; ZANCHETTIN, C. Mineração de dados aplicada à predição do desempenho de escolas e técnicas de interpretabilidade dos modelos. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 2021. p. 773–782.

SORGATTO, D. W. *et al.* Avaliação de classificadores para relacionar características escolares a indicadores educacionais. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 2021. p. 1232–1242.

Data de submissão: 31/04/2024 | Data de revisão: 22/06/2024 | Data de aceite: 28/06/2024 | Data de publicação: 31/07/2024