Artículo de investigación

Optimizando el aprendizaje de los lenguajes de programación. Un enfoque basado en la analítica de datos para los estudiantes de Ingeniería de Sistemas en la Fundación Universitaria Los Libertadores

Optimizing programming language learning: A data analytics-based approach for Systems Engineering students at Los Libertadores University Foundation

PERSPECTIVAS

Javier Daza Piragauta
Fundación Universitaria Los Libertadores, Colombia
jdazap@libertadores.edu.co

Jhonn Edgar Castro Montaña
Fundación Universitaria Los Libertadores, Colombia
jecastromo@libertadores.edu.co

Hernán Ávila Puentes
Fundación Universitaria Los Libertadores., Colombia
jdazap@libertadores.edu.co

Revista Perspectivas
vol. 9, núm. 24, p. 234 - 256, 2024
Corporación Universitaria Minuto de Dios, Colombia
ISSN: 2145-6321

Resumen: La investigación propone una metodología sólida para construir modelos predictivos de rendimiento académico en estudiantes de Ingeniería de Sistemas, basada en el análisis detallado de datos académicos, especialmente en programación y programación intermedia. Busca mejorar la analítica educativa y la predicción del rendimiento en áreas especializadas utilizando diversas métricas. Se analizan estudiantes del primer y segundo semestre de los últimos cinco años, recopilando datos meticulosamente y empleando tres técnicas de aprendizaje automático para construir modelos predictivos con una precisión cercana al 65 %.

Así, el modelo basado en Naïve Bayes se destaca por identificar estudiantes propensos a repetir y abandonar, examinando las características más relevantes para la predicción del rendimiento académico. Este enfoque no solo busca mejorar la capacidad predictiva en Ingeniería de Sistemas, sino también ser replicable en otros cursos y programas.

En consecuencia, la metodología contribuirá a mejorar los procesos de aprendizaje y a generar estrategias de intervención para estudiantes en riesgo de repitencia y deserción, alineándose con la misión de la investigación de catalizar un

ISSN-E: 2619-1687
Periodicidad: Frecuencia continua
perspectivas@uniminuto.edu

Recepción: 29 Noviembre 2023
Aprobación: 21 Mayo 2024
Publicación: 04 Septiembre 2024

DOI: https://doi.org/10.26620/uniminuto.perspectivas.9.24.2024.234-256

URL: http://portal.amelica.org/ameli/journal/638/6384708015/

impacto positivo en la calidad educativa y el bienestar estudiantil, actuando como un agente de transformación interdisciplinaria.

Palabras clave: aprendizaje, lenguajes de programación, analítica de datos, ingeniería, sistemas de información, educación superior, evaluación del rendimiento.

Abstract: The research proposes a solid methodology for constructing predictive models of academic performance in Systems Engineering students, based on detailed analysis of academic data, especially in programming and intermediate programming. By utilizing various metrics, it seeks to enhance educational analytics and the prediction of performance in specialized areas. Students from the first and second semesters of the last five years will be analyzed, meticulously collecting data and employing three machine learning techniques to construct predictive models with an accuracy close to 65 %.

Thus, the Naïve Bayes-based model stands out for identifying students prone to repeating and dropping out, examining the most relevant characteristics for predicting academic performance. This approach aims not only to improve predictive capacity in Systems Engineering but also to be replicable in other courses and programs.

Consequently, the methodology will contribute to improving learning processes and generating intervention strategies for students at risk of repetition and dropout, aligning with the research mission of catalyzing a positive impact on educational quality and student well-being, acting as an agent of interdisciplinary transformation.

Keywords: Learning, programming languages, data analysis, engineer, information systems, higher education, performance evaluation.

Introducción

En la era digital actual, el aprendizaje de los lenguajes de programación se ha convertido en una habilidad fundamental para los estudiantes de Ingeniería de Sistemas. Con el fin de garantizar un proceso educativo efectivo, es esencial contar con herramientas que permitan monitorear el progreso y desempeño de los estudiantes en esta área. En este contexto, la analítica de datos emerge como una solución prometedora para el seguimiento y la mejora continua del aprendizaje.

Sin duda, existe un avance tecnológico sin precedentes en años recientes en diversas áreas (Sánchez, 2009), en particular, en el área educativa. En la literatura se ha observado que el uso de herramientas tecnológicas puede fomentar con prácticas adecuadas la enseñanza y aprendizaje (Torres y Cobo, 2017).

Las tecnologías de la información y la comunicación (TIC) impulsan las actividades cognitivas de los usuarios, específicamente, las herramientas tecnológicas que permiten almacenar, analizar e interpretar diversas características de estudiantes para predecir comportamientos académicos futuros y poder realizar intervenciones de manera oportuna en lugar de esperar hasta que el alumno repruebe alguna actividad y sea necesaria una recuperación académica que lleva más tiempo y es más costosa tanto para el alumno como para la institución educativa.

Uno de los objetivos del análisis de datos académicos es encontrar patrones y predicciones que permitan caracterizar el desarrollo académico de estudiantes, no obstante, se requiere la recopilación de datos de las características de los estudiantes teniendo en cuenta el contexto, para así conseguir una mayor comprensión de los resultados obtenidos. Algunas de estas características son factores socioeconómicos, datos familiares y escolares del estudiante.

Por lo tanto, para el análisis de datos se emplean diferentes métodos, técnicas y algoritmos (Peña, 2014). La predicción del rendimiento académico se realiza con diversos propósitos, tales como detectar el riesgo de abandono o la posibilidad de deserción por parte de los estudiantes.

Así, el rendimiento académico puede medirse en diferentes fases del proceso de formación académica del estudiante, así como también se pueden recopilar diversas variables o características del estudiante asociadas al rendimiento. Esta información puede almacenarse para ser analizada posteriormente para predecir el rendimiento académico del estudiante y tomar decisiones adecuadas de manera oportuna para mejorar los resultados del aprendizaje (Gutiérrez *et al.*, 2021).

Resulta pertinente mencionar que, para realizar la predicción, se han analizado datos para construir modelos predictivos a partir de técnicas de aprendizaje automático debido a que las herramientas estadísticas clásicas pueden no funcionar adecuadamente con grandes cantidades de datos y con varias características de estudiantes (Romero y Ventura, 2010). Estas técnicas de aprendizaje automático se centran en el uso y manejo de datos para obtener resultados representados como decisiones y son útiles para desarrollar modelos de predicción. Típicamente, estas técnicas se utilizan en áreas de tipo comercial o empresarial (Han, 2012), sin embargo, se han comenzado a emplear en el diseño de modelos predictivos del rendimiento académico a partir de factores o características de estudiantes (Romero y Ventura, 2010).

En esta investigación, se plantea la siguiente pregunta: ¿cómo desarrollar una metodología para construir modelos predictivos del rendimiento académico a partir de los datos académicos de los estudiantes de Ingeniería de Sistemas de la FULL? El objetivo es desarrollar una estrategia de analítica de datos para monitorear el aprendizaje de los lenguajes de programación de los estudiantes del Programa de Ingeniería de Sistemas de la FULL.

Marco teórico

El presente artículo se fundamenta en un marco teórico que abarca diversas áreas de estudio clave, proporcionando una base sólida para la optimización del aprendizaje de los lenguajes de programación en estudiantes de Ingeniería de Sistemas mediante un enfoque basado en la analítica de datos.

Aprendizaje de programación

Es de resaltar el hecho de que los procesos de enseñanza y aprendizaje de programación han sido objeto de atención en la literatura académica. Autores como Sebesta (2015) y Lister *et al.* (2004) ofrecen perspectivas valiosas sobre estrategias efectivas y desafíos comunes en la enseñanza de la programación.

Analítica de datos educativos

La aplicación de la analítica de datos en el ámbito educativo ha cobrado relevancia. Autores como Siemens y Long (2011) y Clow (2013) exploran el potencial de la analítica de datos para mejorar la toma de decisiones en educación, destacando su papel en la identificación de patrones y el diseño de intervenciones personalizadas.

Tecnologías de la información y comunicación (TIC)

Sin duda el uso y apropiación de TIC en la educación es un tema ampliamente discutido. Anderson y Dron (2011) ofrecen una visión integral de cómo las TIC pueden transformar la enseñanza y el aprendizaje, destacando su papel en la personalización de la educación y el acceso a recursos digitales.

Rendimiento académico

En efecto la medición y predicción del rendimiento académico es crucial. Autores como Wang y Woo (2007) proporcionan enfoques para evaluar el rendimiento académico, mientras que Adelman (2006) analiza factores que influyen en el éxito académico, proporcionando insights para el diseño de estrategias efectivas.

Modelos predictivos - tipos y beneficios

Cabe reslatar que la construcción de modelos predictivos en educación ha sido explorada por diversos investigadores. Baker y Yacef (2009) presentan un enfoque detallado sobre la construcción de modelos predictivos en entornos educativos, abordando cuestiones relacionadas con la predicción del rendimiento estudiantil.

Los modelos predictivos son una técnica estadística comúnmente utilizada para predecir comportamientos y resultados probables en el futuro según los datos utilizados.

De otra parte, el modelado predictivo es una forma de minería de datos que analiza datos internos (históricos y actuales, para nuestro caso se utilizó la información de la materia programación únicamente teniendo en cuenta la nota final de los estudiantes). Estos, en conjunto con datos oficiales y data alternativa, generan un modelo que capaz estimar resultados en un tiempo determinado

Modelos predictivos más utilizados

Con los avances en inteligencia artificial, *big data, data mining* y ánálisis de datos, los modelos predictivos se han convertido en una herramienta esencial para todos aquellos que desean tomar decisiones.

Todos estos modelos se basan en algoritmos de aprendizaje y técnicas de Machine Learning. Es así que con la ayuda de los programadores e ingenieros se logra obtener información valiosa a partir de grandes conjuntos de datos. El *machine learning* es una de las ramas de la inteligencia artificial (IA), funciona a través de algoritmos y el objetivo principal es que las máquinas tengan la capacidad de identificar patrones para predecir datos que son de utilidad.

Dependiendo del campo de trabajo y de los tipos de datos que se utilicen, los modelos predictivos se pueden clasificar diferentes grupos: modelos de previsión, modelos de clasificación, modelos de regresión, modelos de datos atípicos, modelos de series de tiempo, árboles de decisión y, por supuesto, el tema de moda en el mundo, las redes neuronales.

Ingeniería de sistemas

En el contexto de la ingeniería de sistemas, autores como Sommerville (2011) ofrecen una perspectiva general sobre la ingeniería de *software*, proporcionando contexto y principios fundamentales que son esenciales para comprender la aplicación de la analítica de datos en este contexto.

En sintesis, en virtud de las ideas anteriores se integran conocimientos clave en programación, analítica de datos educativos, tecnologías de la información, rendimiento académico, modelos predictivos y la disciplina específica de Ingeniería de Sistemas.

Metodología

Después del análisis de diversas fuentes de información relacionadas con la analítica de datos y los modelos predictivos, es evidente resaltar que en Colombia existen pocos trabajos que aplican técnicas de aprendizaje automático para el diseño de modelos predictivos del rendimiento académico, a pesar del potencial beneficio que pueden tener en el desempeño académico de estudiantes. En particular, la predicción del rendimiento académico ofrece la oportunidad de elaborar planes de prevención de repitencia y deserción académica mediante la realización de estrategias de intervención en lugar de estrategias de recuperación académica.

Es decir, estos modelos permiten a los profesores e instituciones educativas realizar intervenciones desde el principio del curso y no al final cuando es demasiado tarde para realizar alguna acción para evitar la repitencia del estudiante y por ende la deserción.

Es pertinente destacar que la recopilación de datos se efectuó con base en las notas finales de los estudiantes que cursaron los espacios académicos de programación y programación intermedia, espacios académicos correspondientes al primer y segundo semestre del programa de Ingeniería de Sistemas, datos correspondientes a una ventana de tiempo de cinco años, es decir de 2017-2 a 2022-2. Esto permitió anticipar la probabilidad de aprobación del estudiante antes de que finalice cada semestre lectivo. Esta anticipación brinda a los profesores y a las instituciones educativas la posibilidad de disponer de un tiempo razonable para planificar y ejecutar intervenciones pertinentes con el objetivo de mitigar los índices de reprobación.

Es de resaltar que en el presente estudio, la construcción y evaluación de los modelos predictivos se llevó a cabo con el respaldo del *software* de licencia libre Weka (Waikato Environment for Knowledge Analysis). Este *software*, conforme a la obra de Witten *et al.* (2005), alberga diversas técnicas de aprendizaje automático y proporciona una interfaz que facilita la visualización de datos en múltiples formatos. Weka posibilita la inserción de datos contenidos en archivos que poseen varios registros, cada uno con un conjunto definido de atributos, como se ha detallado en la investigación de Díaz *et al.* (2021).

Finalmente, se evalúan los modelos y se comparan con base en métricas representativas, así, la metodología propuesta se resume como sigue:

Recopilación de datos de estudiantes de Ingeniería de Sistemas primer semestre => Construcción de los modelos predictivos => Evaluación de los modelos predictivos.

Análisis de datos con Python

Aunque el análisis de datos se ha convertido en una ciencia arraigada en nuestra cultura, las palabras por sí solas son insuficientes para captar su alcance. En este contexto, surge una pregunta crucial: ¿cómo podemos llevar a cabo el análisis de datos de manera efectiva? ¿Existen herramientas tecnológicas que puedan asistirnos en este proceso? La respuesta es afirmativa. Mientras que algunos optan por utilizar aplicaciones de oficina para filtrar, buscar y generar figuras a partir de datos, otros prefieren explorar lenguajes de programación. Aunque estos lenguajes pueden no ser tan intuitivos para el usuario final, el lenguaje de programación Python se destaca como una solución directa y eficaz para el análisis de datos. Python ofrece una amplia gama de bibliotecas y herramientas que facilitan la manipulación, el análisis y la visualización de datos, convirtiéndolo en una opción preferida entre los profesionales del campo.

Ventajas de Python para el análisis de datos

Para abordar este tema en profundidad, es fundamental destacar y recomendar Python como una herramienta clave en el análisis de datos. Python es un lenguaje de programación altamente flexible, multipropósito y multiplataforma, además de ser de código abierto y gratuito. Su popularidad en la ciencia de la analítica de datos ha aumentado significativamente con el tiempo. A diferencia de otros lenguajes de programación que requieren compilación o instalaciones extensas y susceptibles a errores, Python permite realizar prácticamente cualquier tarea de manera eficiente. Esta versatilidad y facilidad de uso hacen de Python una elección preferida para los profesionales que buscan una solución robusta y fiable en el análisis de datos.

Librerías

Pandas

Es una biblioteca de código abierto que opera de manera eficiente y distintiva, facilitando la manipulación de diversos formatos de datos, como archivos .CSV y bases de datos SQL. La biblioteca crea objetos en Python denominados "*dataframes*", los cuales son esencialmente tablas estructuradas que permiten una visualización amigable tanto para desarrolladores como para usuarios finales. Pandas permite trabajar con estructuras de datos de alto nivel y proporciona herramientas avanzadas para la manipulación, análisis y visualización de datos, contribuyendo significativamente a la toma de decisiones informadas.

Numpy

Otra biblioteca ampliamente utilizada en el ámbito del análisis de datos con Python es NumPy. El nombre de esta biblioteca, abreviatura de "Python Numérico", refleja su propósito: ofrecer una extensa colección de funciones precompiladas para la ejecución de rutinas numéricas, la gestión de estructuras de datos y el trabajo con matrices multidimensionales. Estas capacidades permiten realizar cálculos complejos de manera eficiente. En resumen, NumPy es la biblioteca de referencia para la informática científica, que proporciona una base sólida para la manipulación y el análisis de datos en entornos de investigación y desarrollo.

Numba

Para concluir con las opciones disponibles para la analítica de datos, es pertinente mencionar Numba. Esta biblioteca tiene la capacidad de traducir funciones escritas en Python a código máquina optimizado durante la ejecución. Los algoritmos numéricos compilados con esta notable biblioteca pueden alcanzar velocidades de ejecución comparables a las de lenguajes de programación como Fortran o C. Numba permite a los desarrolladores aprovechar las ventajas de Python para la prototipación rápida, mientras que ofrece el rendimiento de lenguajes de bajo nivel en aplicaciones intensivas en cálculos.

Matplotlib

En el ámbito de la visualización de datos, aunque es posible representar información a través de tablas u otros mecanismos que muestran filas y columnas, la utilización de figuras es una opción más efectiva para comunicar análisis complejos de manera clara y concisa. En este contexto, una de las bibliotecas más utilizadas en el análisis de datos con Python es Matplotlib. Esta herramienta permite generar figuras de alta calidad que están listos para su publicación, facilitando así la interpretación y presentación de los datos.

Ciencia de datos con Google Colab

En este aparte se destaca el uso y aprehensión de Colab con toda la potencia de las bibliotecas más populares de Python, para analizar y visualizar datos.

Carga y descripción de datos

Cabe destacar que, los datos proporcionados consisten en notas de estudiantes en dos cursos: Programación y Programación Intermedia, a lo largo de varios semestres. Se utilizó Pandas para la carga y manipulación de los datos, y se realizó un análisis descriptivo para identificar tendencias y patrones, es como sigue:

Formato de muestra

Formato de muestra

El formato de muestra de datos seleccionado es el siguiente:Gru_Semestre	Est_Código	Mat_Código	Nombrematería	Nota
2017-2	201720003601	IS0205	Programación	3,3
2017-2	201720003601	IS0205	Programación	2,5
2017-2	201720003601	IS0205	Programación	3,9
2017-2	201720003601	IS0205	Programación	3,6
2017-2	201720003601	IS0205	Programación	3,9
2017-2	201720003601	IS0205	Programación	3,4

Python Code

El código de Python generado para cargar los datos es como sigue:

```
from google.colab import files
import pandas as pd
import io
# Cargar los datos
uploaded = files.upload()
df = pd.read_csv(io.BytesIO(uploaded['datos.csv']))
print(df)
```

Agrupación y conteo de notas por semestre

Los datos muestran el número de notas registradas por semestre para ambos cursos. Aquí están las cifras para el curso de Programación y Programación Intermedia:

Curso de Programación

Semestre	Recuento de notas
2017-2	11
2018-1	25
2018-2	19
2019-1	24
2019-2	19
2020-1	25
2020-2	12
2021-1	28
2021-2	38
2022-1	40
2022-2	48

Curso de Programación Intermedia

Semestre	Recuento de notas
2018-1	7
2018-2	19
2019-1	13
2019-2	22
2020-1	14
2020-2	20
2021-1	16
2021-2	28
2022-1	29
2022-2	35

Total general de notas: 492

Análisis de tendencias

• **Incremento de matriculados:**en ambos cursos, hay un notable incremento en el número de notas registradas a lo largo de los semestres, lo que podría indicar un aumento en la matrícula de estudiantes o una mayor retención en estos cursos.

• **Variación semestral:** el curso de Programación muestra una variación significativa en el número de notas registradas entre semestres, particularmente notable entre los semestres de 2017-2 y 2022-2. Esto podría deberse a cambios en el plan de estudios, en la metodología de enseñanza, o en la inscripción de estudiantes.

• Similarmente, el curso de Programación Intermedia también muestra un incremento constante en el número de notas registradas desde 2018-1 hasta 2022-2.

• **Pico de registros:** ambos cursos alcanzan picos de registros en los semestres más recientes (2022-1 y 2022-2), sugiriendo un mayor interés o una mayor necesidad de los cursos en estos periodos.

T3 Análisis comparativo

Al comparar ambos cursos, se observa que el curso de Programación tiene consistentemente más notas registradas por semestre en comparación con Programación Intermedia. Esto es esperable, ya que Programación puede ser un curso de nivel más básico y con mayor número de estudiantes inscritos inicialmente.

La diferencia en el número de notas entre los cursos se reduce en los semestres más recientes, lo que podría indicar un aumento en la progresión de los estudiantes desde Programación hacia Programación Intermedia.

Implicaciones educativas

• **Ajustes en la enseñanza:** la variabilidad en el número de notas y el incremento constante sugieren que podría ser necesario ajustar los

recursos educativos y las estrategias de enseñanza para manejar el creciente número de estudiantes y mejorar la calidad del aprendizaje.

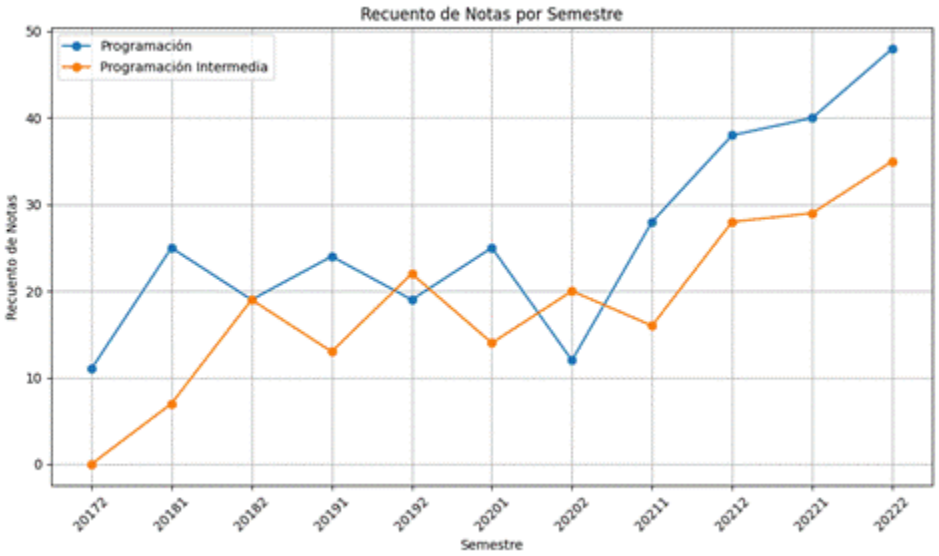
- **Intervenciones específicas:** la identificación de semestres con menor número de notas registradas podría ayudar a focalizar intervenciones educativas para mejorar la retención y el éxito académico en esos periodos específicos.
- **Seguimiento continuo:** un seguimiento continuo de estas tendencias permitirá realizar ajustes oportunos en la planificación de los cursos y en el apoyo académico proporcionado a los estudiantes.

Como se muestra en la figura 1, el análisis de los datos de rendimiento académico en los cursos de Programación y Programación Intermedia presenta un incremento general en la matrícula y la retención de estudiantes. Este patrón sugiere la necesidad de ajustes en la estrategia educativa para acomodar el crecimiento y asegurar el éxito académico continuo. Una metodología basada en el análisis de datos permitirá realizar intervenciones más efectivas y mejorar la experiencia educativa de los estudiantes.

Figura 1. Total de notas

Figura 1

Total de notas



elaboración propia.

La figura 1 muestra el recuento de notas por semestre para los cursos de Programación y Programación Intermedia. Esta gráfica presenta cómo ambos cursos han experimentado variaciones en el número de notas registradas a lo largo de los semestres, destacando un incremento significativo en los semestres más recientes.

- **Eje X (horizontal):** semestres desde 2017-2 hasta 2022-2.
- **Eje Y (vertical):** recuento de notas.

• **Línea azul con marcadores:** recuento de notas en el curso de Programación.

• **Línea naranja con marcadores:** recuento de notas en el curso de Programación Intermedia.

Esta visualización permite identificar las tendencias en la matrícula y la retención de estudiantes en estos cursos, así como a planificar posibles intervenciones educativas.

Análisis del incremento de estudiantes

La figura 2 muestra la suma de notas por semestre, reflejando el incremento en la cantidad de estudiantes a lo largo del tiempo.

Figura 2

Suma de notas por semestre

elaboración propia.

Con base en la figura 2, a continuación, se presentan algunos puntos destacados del análisis.

• **Tendencia ascendente:** la suma de notas muestra una clara tendencia ascendente desde el semestre 2017-2 hasta el semestre 2022-2, indicando un aumento significativo en el número de estudiantes.

• **Picos significativos:** se observan picos notables en los semestres 2021-2 y 2022-2, lo que sugiere un incremento considerable en la inscripción o en el rendimiento académico de los estudiantes durante estos periodos.

• **Variabilidad:** aunque la tendencia general es de crecimiento, hay semestres como 2020-2 que muestran una ligera disminución en la suma de notas, lo cual podría deberse a factores específicos de ese periodo, como cambios en el plan de estudios, la metodología de enseñanza, o circunstancias externas.

• **Impacto en la planificación educativa:** este incremento constante en la suma de notas subraya la necesidad de ajustar los recursos educativos y las estrategias de enseñanza para manejar el creciente número de estudiantes y asegurar la calidad del aprendizaje.

En síntesis, el análisis sugiere un crecimiento sostenido en el número de estudiantes, con algunos picos de incremento que podrían requerir atención adicional para mantener la calidad educativa y gestionar adecuadamente los recursos disponibles.

Análisis del histograma

La figura 3 muestra dos componentes principales: la frecuencia de las clases y el porcentaje acumulado.

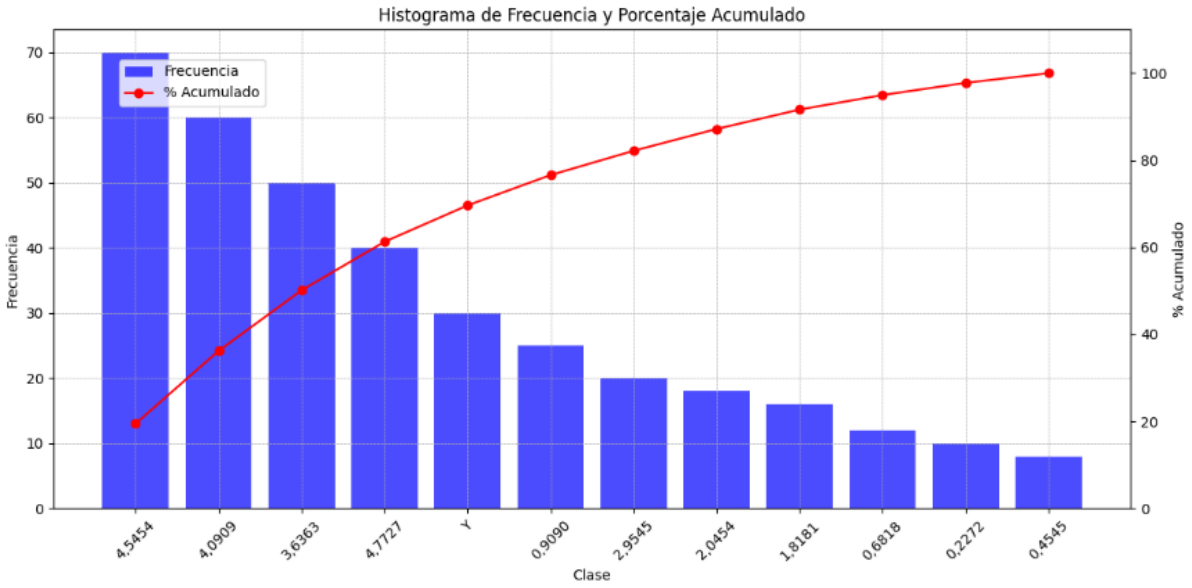


Figura 3

Histograma

elaboración propia.

A partir de la figura 3 se presenta el siguiente análisis detallado:

• **Frecuencia de clases**

- **El eje X** representa las clases (rango de valores).

- **El eje Y** de la izquierda muestra la frecuencia de observaciones en cada clase.

- **La barra más alta** en el histograma representa la clase con la mayor frecuencia de observaciones, lo que indica que esa clase es la más común en el conjunto de datos.

• **Porcentaje acumulado:**

- **El eje Y** de la derecha representa el porcentaje acumulado.

- **La línea roja** con marcadores muestra cómo se acumula el porcentaje a medida que se suman las frecuencias de las clases.

- **El porcentaje acumulado** alcanza el 100 % al final del histograma, lo que indica que se han contabilizado todas las observaciones.

• **Distribución de los datos:**

- **El histograma parece mostrar** una distribución de frecuencias que decrece a medida que se mueven hacia la derecha en el eje X, indicando que las observaciones más altas son menos comunes.

- **La figura sugiere** una posible distribución asimétrica, donde la mayoría de las observaciones se encuentran en las primeras clases a la izquierda.

• **Interpretación del porcentaje acumulado:**

- **El porcentaje acumulado** sube rápidamente al principio, lo que indica que una gran proporción de los datos se encuentra en las primeras clases.

- **Este rápido incremento inicial** del porcentaje acumulado sugiere que las primeras clases contienen la mayor parte de los datos, mientras que las clases posteriores contienen menos observaciones.

• **Conclusión:**

- **El histograma y la figura** de porcentaje acumulado proporcionan una visión clara de la distribución de los datos y cómo se acumulan las observaciones a lo largo de las clases.

- **La visualización** es útil para identificar la clase más frecuente y para entender la distribución acumulada de los datos.

En síntesis, el histograma ilustra una distribución de datos con una mayor concentración en las primeras clases, con una disminución gradual en la frecuencia hacia las clases posteriores. El porcentaje acumulado proporciona una visión clara de cómo se distribuyen las observaciones a lo largo de las clases.

Resultados

La construcción de modelos predictivos conlleva la necesidad imperativa de evaluar meticulosamente su desempeño mediante la aplicación de métricas de evaluación establecidas. En el presente estudio, hemos optado por emplear medidas tales como la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos, siguiendo la propuesta metodológica de Durairaj y Vijitha (2014). La exactitud, como se ha definido en la sección anterior, se refiere a la proporción de predicciones acertadas en relación con el total de predicciones realizadas..

La tasa de verdaderos positivos se determina dividiendo el número de registros predichos como positivos entre el total de registros que realmente son positivos. En este estudio, esta métrica denota las predicciones correctas de estudiantes aprobados en comparación con el total de estudiantes aprobados. Por otro lado, la tasa de verdaderos negativos se refiere al número de registros predichos como reprobados en relación con el total de registros de estudiantes que reprobaron.

En ese orden, para calcular la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos de los modelos predictivos basados en las técnicas de Naïve Bayes, k vecinos más cercanos y árbol de decisión C4.5, hemos aplicado la validación cruzada con diez particiones, conforme a la metodología propuesta por Mueen *et al.* (2016).

Este procedimiento implica la división aleatoria de los datos de entrenamiento en diez particiones, utilizando nueve de ellas para desarrollar el modelo predictivo y prever la aprobación en la partición restante, lo que facilita el cálculo de las métricas de evaluación mencionadas. Este proceso se repite, reservando una partición diferente para realizar las predicciones en cada iteración, y se calcula el promedio de las métricas de evaluación. Este enfoque se replica para cada una de las tres técnicas de aprendizaje automático empleadas.

En efecto, las figuras 4, 5 y 6 de este documento exhiben la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos, respectivamente, de los modelos desarrollados mediante las técnicas de aprendizaje automático mencionadas.

Por ende, la representación gráfica proporcionada en la figura 4 revela que la exactitud muestra una similitud notable entre los tres modelos analizados. Sin embargo, cabe destacar que el modelo generado mediante la técnica de k vecinos más cercanos exhibe la mayor exactitud entre las predicciones realizadas.

Cabe destacar que, la precisión de los modelos predictivos del rendimiento académico se refiere a la medida de la proporción de predicciones correctas realizadas por el modelo con respecto al total de predicciones. Sin duda representa una métrica para evaluar la capacidad general del modelo para predecir los casos positivos y los casos negativos. Una mayor exactitud indica una mayor capacidad del modelo para realizar predicciones correctas en general, proporcionando una evaluación global de su rendimiento.

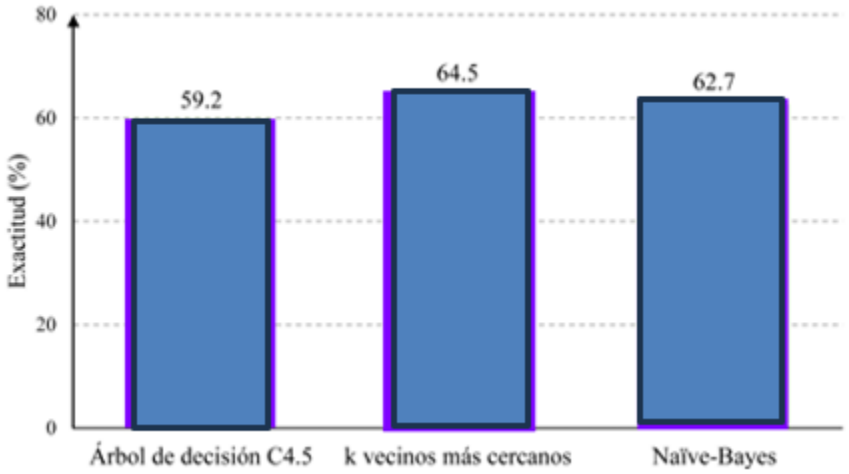


Figura 4

Precisión de los modelos predictivos del desempeño académico

elaboración propia

La revisión de la figura 5 pone de manifiesto que el valor más alto de la tasa de verdaderos positivos se obtiene al aplicar la técnica de k vecinos más cercanos. Además, se observa que las técnicas de árbol de decisión y Naïve Bayes muestran valores comparativamente cercanos en esta métrica.

Por consiguiente, la tasa de verdaderos positivos en los modelos predictivos del rendimiento académico (MPRA) refleja la proporción de casos correctamente identificados como positivos (por ejemplo, estudiantes que aprobarán) entre todos los casos reales positivos. En términos simples, indica la capacidad del modelo para predecir de manera precisa los resultados positivos, lo que la convierte en una métrica esencial para evaluar la efectividad y la capacidad predictiva de la herramienta utilizada en el contexto académico.

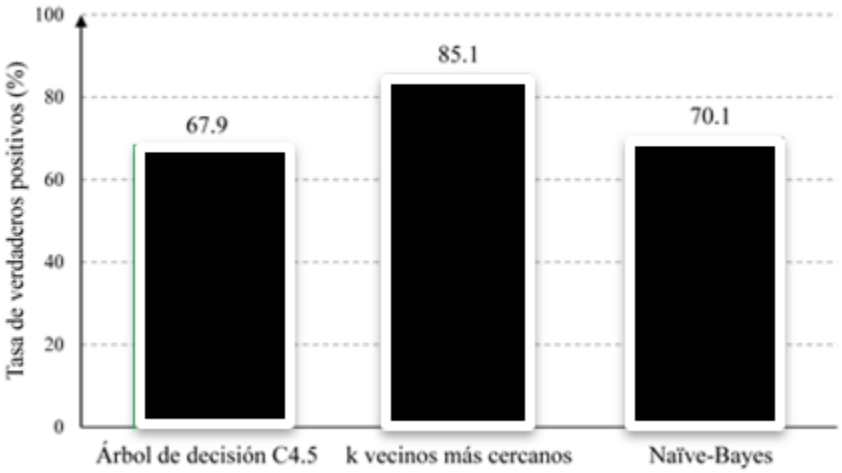


Figura 5

Porcentaje de aciertos en las predicciones positivas de los MPRA

elaboración propia.

La inspección de la figura 6 resalta que la tasa de verdaderos negativos alcanza su punto máximo en el modelo predictivo aplicado mediante la técnica de Naïve Bayes, mientras que el valor más bajo se registra en el modelo basado en la técnica de k vecinos más cercanos. Este patrón contrasta con los hallazgos de las métricas de evaluación previamente examinadas.

En el orden de las ideas anteriores, la tasa de verdaderos negativos en los MPRA denota la proporción de casos correctamente identificados como negativos (por ejemplo, estudiantes que no aprobarán) entre todos los casos reales negativos. En términos más concretos, refleja la precisión del modelo al prever con exactitud los resultados negativos, lo que contribuye significativamente a evaluar la efectividad y la capacidad predictiva de la herramienta utilizada.

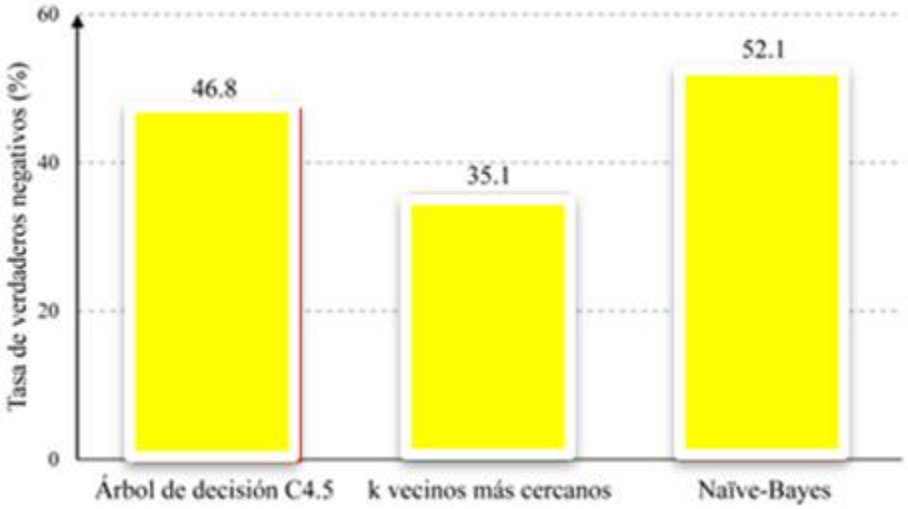


Figura 6

Porcentaje de aciertos en las predicciones negativas de los MPRA

elaboración propia

Conclusiones

La investigación ha cumplido plenamente su objetivo de desarrollar una metodología para predecir el desempeño académico de estudiantes de programación y programación intermedia en el programa de Ingeniería de Sistemas de la Fundación Universitaria Los Libertadores. Esta metodología, basada en la analítica de datos aplicada al aprendizaje de lenguajes de programación, demuestra ser una herramienta efectiva para mejorar el proceso de enseñanza y aprendizaje en este ámbito específico, corroborando así la alineación con los objetivos establecidos en la investigación. En consecuencia, los aspectos significativos a partir de la investigación son como sigue:

- **Efectividad del enfoque analítico de datos:** la analítica de datos mejora el aprendizaje de los estudiantes al optimizar los métodos de enseñanza de programación.
- **Relevancia para la ingeniería de sistemas:** el método se alinea bien con las necesidades específicas de los estudiantes de ingeniería de sistemas.
- **Personalización del aprendizaje:** la analítica de datos permite adaptar la enseñanza a las necesidades individuales de los estudiantes.
- **Potencial de replicación:** la metodología es replicable en otras instituciones y campos relacionados con la programación e informática.
- **Integración tecnológica y educativa:** el éxito de la metodología requiere herramientas tecnológicas adecuadas y formación docente.

Este estudio demostró que es viable desarrollar modelos predictivos de rendimiento académico al inicio de cursos de programación en

ingeniería de sistemas. La evaluación utilizó métricas de exactitud y tasas de verdaderos positivos y negativos, comparando diversas técnicas de aprendizaje automático. El promedio actual del estudiante fue clave, con diferentes umbrales de influencia en Naïve Bayes y árboles de decisión, respaldando hallazgos previos.

La evaluación mostró que la técnica de k vecinos más cercanos superó a Naïve Bayes en un 2 % y a árboles de decisión en un 5 % en términos de exactitud. La mayor disparidad se observó en la tasa de verdaderos positivos, con k vecinos superando a Naïve Bayes en un 15 % y a árboles de decisión en un 17 %, sugiriendo su superioridad para predecir estudiantes aprobados.

Comparado con estudios previos, como Juárez *et al.* (2014) y Salal *et al.* (2019), que alcanzaron mayor exactitud, este trabajo utilizó solo nueve características y evaluó tasas de verdaderos positivos y negativos, ofreciendo información adicional sobre los modelos predictivos y facilitando la recolección y análisis de datos.

Castrillón *et al.* (2020) usaron exclusivamente la técnica de árbol de decisión con 22 atributos y 460 registros, logrando una precisión del 91 %. Sin embargo, la evaluación se realizó con métodos de poca aleatoriedad, como la validación cruzada con dos particiones, lo que podría sesgar el modelo hacia los datos de entrenamiento.

A mayor precisión en las predicciones, mayor será la utilidad del modelo en entornos prácticos para predecir el rendimiento académico.

La analítica de datos en el seguimiento del aprendizaje de lenguajes de programación es fundamental. Permite la recolección, análisis y visualización de datos académicos, identificando patrones, tendencias y áreas de mejora, y ofreciendo intervenciones tempranas para optimizar el proceso educativo.

La analítica de datos aplicada al monitoreo del aprendizaje en ingeniería de sistemas permite anticipar el rendimiento estudiantil y tomar medidas proactivas. Facilita la identificación de estudiantes en riesgo y el diseño de estrategias personalizadas de intervención, contribuyendo a un entorno educativo más eficiente y orientado al éxito de cada estudiante.

Finalmente, las particularidades empleadas en este estudio pueden replicarse en la construcción de modelos predictivos para otros cursos, lo que permite aplicar la metodología desarrollada en diversos contextos. Esto facilita la recolección de atributos antes o al inicio del curso, lo cual es esencial para la planificación de intervenciones educativas.

Discusión

La investigación ha cumplido su objetivo de desarrollar una metodología para predecir el desempeño académico en cursos de programación del programa de Ingeniería de Sistemas en la

Fundación Universitaria Los Libertadores, utilizando la analítica de datos. Esta metodología ha demostrado ser eficaz para mejorar el proceso de enseñanza y aprendizaje en este contexto específico, como lo muestran los siguientes puntos específicos.

- **Efectividad de la metodología de analítica de datos:** la metodología basada en la analítica de datos ha mostrado un impacto positivo en el aprendizaje, permitiendo una detallada recopilación y análisis de datos académicos. Esto facilita la identificación de patrones y áreas de mejora en el rendimiento estudiantil, subrayando la importancia de la analítica de datos en la educación.
- **Relevancia de la metodología para la ingeniería de sistemas:** esta metodología resulta especialmente efectiva para estudiantes de ingeniería de sistemas, alineándose con las necesidades de este campo. La capacidad de anticipar desafíos académicos y ofrecer intervenciones tempranas es crucial para optimizar el proceso educativo y garantizar el éxito en la programación.
- **Importancia de la personalización del aprendizaje:** la analítica de datos permite una personalización efectiva del aprendizaje, adaptando los métodos de enseñanza a las necesidades individuales de los estudiantes. Esto es particularmente relevante en la programación, donde los niveles de habilidad varían ampliamente.
- **Potencial para replicación y aplicación en otros contextos:** la metodología puede ser replicada en otras instituciones educativas y campos relacionados con la programación, utilizando sus características para construir modelos predictivos en diversos cursos y planificar intervenciones educativas.
- **Necesidad de integración tecnológica y educativa:** la integración de herramientas tecnológicas y educativas es esencial para el éxito de esta metodología. La infraestructura tecnológica y la formación de los profesores son cruciales para implementar métodos de enseñanza basados en datos. Este estudio ha demostrado la viabilidad de desarrollar modelos predictivos de rendimiento académico utilizando diversas técnicas de aprendizaje automático.
- **Evaluación de las técnicas de aprendizaje automático:** se evaluaron métricas como la exactitud y las tasas de verdaderos positivos y negativos, empleando técnicas como Naïve Bayes, k vecinos más cercanos y árbol de decisión C4.5. El modelo basado en k vecinos más cercanos superó a los otros modelos en términos de exactitud y tasa de verdaderos positivos, sugiriendo su mayor idoneidad para predecir aprobaciones.
- **Comparación con estudios previos:** este trabajo proporcionó una evaluación más detallada que estudios anteriores, como los de Juárez *et al.* (2014) y Salal *et al.* (2019), al incluir tasas de verdaderos positivos y negativos. Esto ofrece una comprensión más profunda de la efectividad de los modelos predictivos.

- **Implicaciones y aplicaciones futuras:** la analítica de datos aplicada al monitoreo del aprendizaje de lenguajes de programación es fundamental para identificar patrones y anticipar desafíos académicos. Esta metodología predictiva facilita la creación de estrategias personalizadas de intervención, contribuyendo a un entorno educativo más eficiente y orientado al éxito estudiantil. La metodología desarrollada puede ser aplicada en diversos contextos, permitiendo una planificación proactiva y efectiva de las intervenciones educativas.

Referencias

Adelman, C. (2006). T*he toolbox revisited: Paths to degree completion from high school through college*. U.S. Department of Education.

Anderson, T., & Dron, J. (2011). Three generations of distance education pedagogy. *The International Review of Research in Open and Distributed Learning*, 12(3), 80-97. https://doi.org/10.19173/irrodl.v12i3.890

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17. https://doi.org/10.5281/zenodo.3554657

Castrillón, O. D., Sarache, W. y Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación Universitaria*, 13(1), 93-102 (2020) http://dx.doi.org/10.4067/S0718-50062020000100093

Clow, D. (2013). An overview of learning analytics. *Journal of Learning Analytics*, 1(1), 4-19. https://doi.org/10.18608/jla.2014.11.3

Díaz, B., Meleán, R. y Marín, W. (2021). Rendimiento académico de estudiantes en educación superior: predicciones de factores influyentes a partir de árboles de decisión. Telos: *Revista de Estudios Interdisciplinarios en Ciencias Sociales*, 23(3), 616-639. https://doi.org/10.36390/telos233.08

Durairaj, M. y Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, 5(4), 5987-5991. https://www.semanticscholar.org/paper/Educational-Data-mining-for-Prediction-of-Student-Durairaj-Vijitha/892b0182c44c34a2ae68daec819eaec301c3bd9c

Gutiérrez, J. A., Garzón, J. y Segura, A. M. (2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación Universitaria*, 14(1), 13-24. http://dx.doi.org/10.4067/S0718-50062021000100013

Han, J. (2012). *Data mining: Concepts and techniques.* Morgan Kaufmann Publishers.

Juárez, P., Morales, M., Sánchez, J., & López, D. (2014). *Innovative approaches to software development: Challenges and practices*. Springer.

Lister, R., Adams, E. S., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., ... & Thomas, L. (2004). A multi-national study of reading and tracing

skills in novice programmers. *ACM SIGCSE Bulletin*, 36(4), 119-150. https://doi.org/10.1145/1041624.1041673

Mueen, A., Zafar, B., y Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science* (IJMECS), 8(11), 36-42.

Peña, A. (2014). *Métodos y técnicas de análisis de datos en la investigación científica*. Síntesis.

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

Salal, Y., Ullah, S., Khan, M. A., & Shah, M. A. (2019). Analyzing the impact of feature selection on classification techniques: A case study of email spam filtering. *Journal of King Saud University - Computer and Information Sciences*, 31(4), 412-423. https://doi.org/10.1016/j.jksuci.2018.03.011

Sánchez, J. A. (2009). La revolución informática y la educación. *Educación XX1*, 12(1), 17-38.

Sebesta, R. W. (2015). *Concepts of Programming Languages* (11th ed.). Pearson.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 30-32.

Sommerville, I. (2011). *Software engineering* (9th ed.). Addison-Wesley.

Torres, P. C. y Cobo, J. K. (2017). Tecnología educativa y su papel en el logro de los fines de la educación. *Educere*, 21(68), 31-40.

Wang, Q., & Woo, H. L. (2007). Systematic planning for ICT integration in topic learning. *Educational Technology & Society*, 10(1), 148-156.

Witten, I. H. Frank, E. y Hall, M. A. (2005). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Información adicional

Artículo derivado de proyecto de investigación: "Analítica de datos para monitorear el aprendizaje de los lenguajes de programación en los estudiantes de Ingeniería de Sistemas de la Fundación Universitaria Los Libertadores".

Javier Daza Piragauta, Jhonn Edgar Castro Montaña, Hernán Ávila Puentes

Optimizando el aprendizaje de los lenguajes de programación. Un enfoque basado en la analítica de datos para los estudiantes de Ingeniería de Sistemas en la Fundación Universitaria Los Libertadores

Optimizing programming language learning: A data analytics-based approach for Systems Engineering students at Los Libertadores University Foundation

Revista Perspectivas
vol. 9, núm. 24, p. 234 - 256, 2024
Corporación Universitaria Minuto de Dios, Colombia
perspectivas@uniminuto.edu

ISSN: 2145-6321
ISSN-E: 2619-1687

DOI: https://doi.org/10.26620/uniminuto.perspectivas.9.24.2024.234-256