Factores que inciden en el desempeño escolar

Diana Cordero

July 2024

1 Introducción

El desempeño escolar es un fenómeno complejo y multifacético que ha sido objeto de estudio en diversas disciplinas como la psicología, la sociología y la educación. Comprender los factores que influyen en el rendimiento académico de los estudiantes es crucial para desarrollar estrategias efectivas que promuevan un aprendizaje significativo y equitativo.

El desempeño escolar no solo depende de las capacidades intelectuales innatas del estudiante, sino que está influenciado por una variedad de factores intrínsecos y extrínsecos, tales como el apoyo familiar, los intereses del estudiante, su edad, entre otros.

Este trabajo tiene como objetivo analizar y discutir los principales factores que inciden en el desempeño escolar de los estudiantes, a través de técnicas de aprendizaje supervisado debido a las características del conjunto de datos empleado en esta investigación.

2 Objetivo

Seleccionar el método de aprendizaje supervisado o no supervisado acorde a las características del conjunto de datos y emplearlo para determinar los factores que determinan el desempeño escolar de un estudiante.

3 Descripción de los datos

Se empleó un conjunto de datos de 2392 estudiantes cuyas edades oscilan entre los 15 y los 18 años. Las variables que se incluyen en este estudio, además, son: género, grupo étnico, apoyo parental, tiempo dedicado al estudio, ausencias en clase, tutoreo, nivel educativo de los padres; así como la participación en actividades extracurriculares, artísticas, deportivas o de voluntariado.

Se analiza el promedio en una escala de 2 a 4, y de acuerdo con ello se hace la clasificación de la calificación como se muestra a continuación:

Estos datos se obtuvieron de la plataforma kaggle.com

Calificación	Clasificación de la calificación
Mayor o igual que 3.5	0
Mayor o igual que 3 y menor que 3.5	1
Mayor o igual que 2.5 y menor que 3	2
Mayor o igual que 2 y menor que 2.5	3
Menor que 2	4

4 Marco teórico

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial (IA) que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y hacer predicciones o decisiones basadas en datos. Dentro de este campo, existen dos enfoques principales: el aprendizaje supervisado y el aprendizaje no supervisado. Estas técnicas se utilizan para diferentes tipos de problemas y tienen distintas aplicaciones.

4.1 Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica de aprendizaje automático que se utiliza con datos no etiquetados. En este caso, el modelo debe identificar patrones y estructuras en los datos por sí mismo sin la ayuda de etiquetas predefinidas. El objetivo es descubrir relaciones y agrupamientos dentro de los datos.

Entre los tipos más comunes de aprendizaje no supervisado se encuentran:

- Agrupación. Se utiliza para agrupar datos en grupos o clústeres basados en la similitud de las características. Entre las técnicas de agrupación más comunes se incluyen K-means, jerárquico y DBSCAN.

- Reducción de dimensionalidad. Es un proceso cuyo objetivo es reducir el número de variables aleatorias del conjunto de datos, asegurando que la información que proporciona en similar en ambos casos.

4.2 Aprendizaje supervisado

El aprendizaje supervisado es una técnica de aprendizaje automático donde el modelo se entrena utilizando un conjunto de datos etiquetados. Esto significa que cada instancia del conjunto de datos de entrenamiento incluye una entrada (características) y una salida (etiqueta). El objetivo del modelo es aprender una función que mapea entradas a salidas, de tal manera que pueda predecir correctamente las etiquetas de nuevas instancias no vistas.

5 Metodología

Dada la naturaleza del conjunto de datos, en este trabajo se empleó la técnica de aprendizaje supervisado "bosques aleatorios"

El método de bosques aleatorios es una técnica de aprendizaje automático que se utiliza tanto para problemas de clasificación como de regresión. Introducido por Leo Breiman en 2001, el algoritmo Random Forest combina múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste.

Un árbol de decisión es un modelo predictivo que utiliza un gráfico de decisiones en forma de árbol donde cada nodo representa una pregunta o condición sobre una característica, cada rama representa un resultado de esa condición, y cada hoja representa una etiqueta de clase o un valor de salida. Si bien los árboles de decisión son intuitivos y fáciles de interpretar, a menudo sufren de sobreajuste, especialmente cuando son profundos y complejos.

Random Forests mitiga este problema mediante la construcción de múltiples árboles de decisión y luego combinando sus resultados. La idea clave es que, aunque cada árbol individual puede ser un predictor débil, el conjunto de muchos árboles puede formar un predictor fuerte. Este enfoque se basa en los principios de aprendizaje conjunto (ensemble learning).

Se decidió emplear este método debido a la presencia de variables tanto numéricas como categóricas

6 Resultados

En la figura 1 se muestran la matriz de correlación de las variables presentes en el conjunto de datos.

Se observa una correlación fuerte entre el Id. del estudiante (StudentID) y la clasificación de la calificación (GradeClass). Debido a que Id. del estudiante es un identificador y no una variable, se elimina del conjunto de datos, ya que es posible que genere ruido en el análisis.

Una vez eliminada la columna Id. del estudiante, se utiliza el método de bosques aleatorios para obtener la importancia de las características, las cuales se muestran en la figura 2.

De acuerdo con lo anterior, se revisó la exactitud del modelo con un umbral de importancia de 0.1, y se obtuvo el siguiente mapa de calor

Luego de la reducción de características, se obtiene el mapa de calor mostrado en la Figura 3, en la que podemos observar que las características que se consideran más importantes en relación con el desempeño escolar son: inasistencias

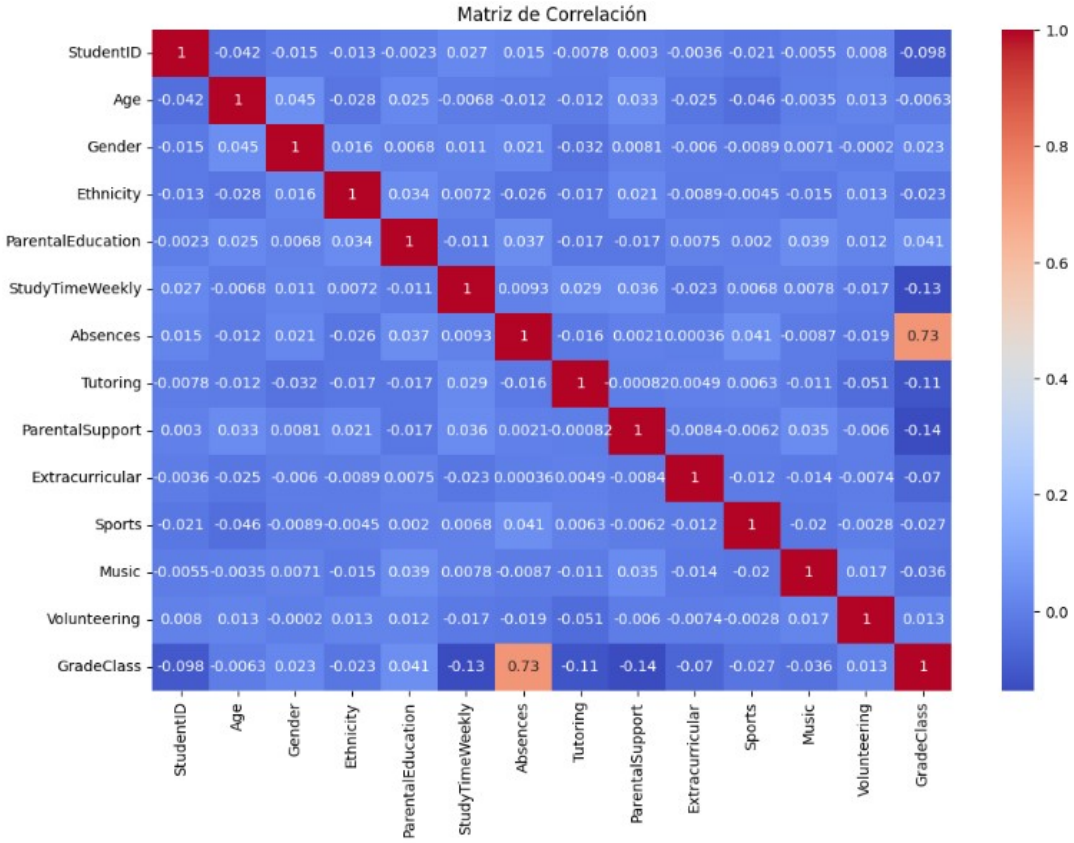

Figure 1: Matriz de correlación

Umbral	Exactitud del modelo
0.01	0.6805
0.02	0.6784
0.05	0.6638
0.1	0.5803

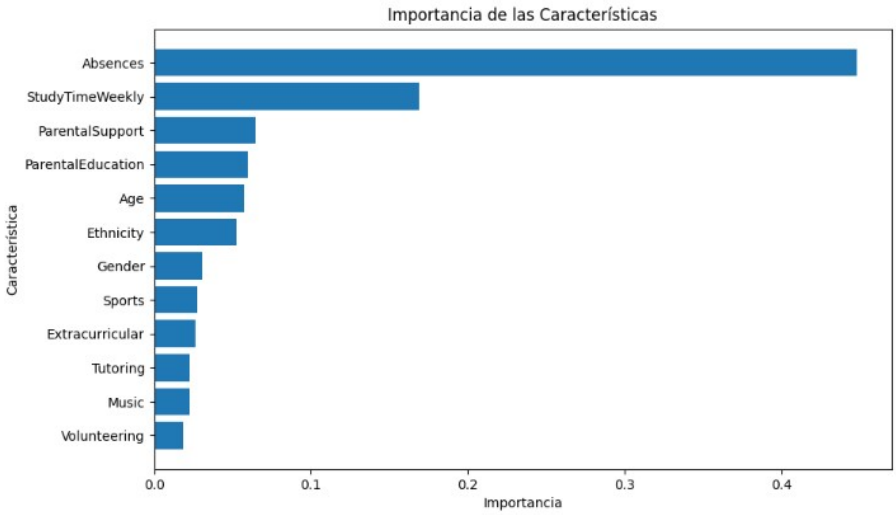

Figure 2: Importancia de las características

Clasificación	**Precisiónn**	**Recall**	**f1-score**	**Soporte**
0	0.21	0.18	0.20	22
1	0.34	0.37	0.35	49
2	0.39	0.44	0.41	85
3	0.34	0.30	0.32	86
4	0.82	0.81	0.82	237
Exactitud				0.5803
Macro Avg	0.42	0.0.42	0.42	479
Weighted Avg	0.58	0.58	0.58	479

(Absences) y tiempo de estudio por semana (StudyTimeWeekly). El resto de las características no resultan tan relevantes.

En la Tabla 3, se muestra el reporte de clasificación después de la reducción de características.

De acuerdo al reporte de clasificación se concluye que el modelo tiene dificultades para clasificar correctamente las clases menos representadas ('0.0', '1.0', '2.0', '3.0'), con precisiones y recalls más bajos en comparación con la clase mayoritaria ('4.0').

Del mismo modo, la matriz de confusión (Figura 3) refleja que las predicciones correctas están sesgadas hacia la clase mayoritaria, con confusiones significativas entre las clases menos representadas.

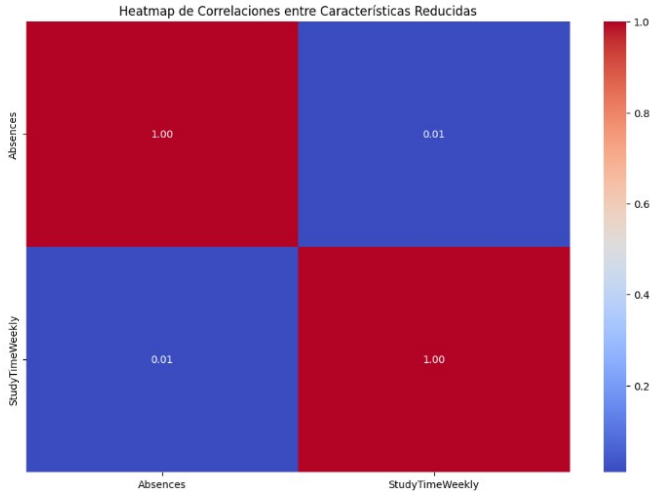

Figure 3: Mapa de calor de características reducidas

7 Conclusiones y siguientes pasos

Dado que el modelo cuenta con una exactitud baja, es importante realizar un análisis más detallado para determinar el umbral óptimo que maximice la precisión sin sacrificar la capacidad predictiva de las clases menos representadas.

Asimismo, se sugiere realizar un análisis más profundo de los errores del modelo para entender las razones detrás de las confusiones entre clases y cómo podrían abordarse.

Finalmente, se considera el uso posterior de técnicas como la validación cruzada para validar el rendimiento del modelo de manera más robusta y realizar ajustes iterativos en función de los resultados obtenidos.

8 Bibliografía

- Merino, R. F. M., Chacón, C. I. Ñ. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=6230447

- Jiménez-Cordero, M. A., Aplicada, A. M. E. E. I. o. y. M. (2022, 19 septiembre). Aprendizaje supervisado: métodos, propiedades y aplicaciones. Universidad de Málaga. https://riuma.uma.es/xmlui/handle/10630/25147