Tarea 8. Métricas de desempeño

Diana Cordero

June 2024

1 Introducción

Durante esta investigación, se analizará un conjunto de datos acerca de las características y factores que influyen en el desempeño académico de los estudiantes. La finalidad es entrenar un modelo para clasificar la calificación de un estudiante, basado en los datos de entrada sugeridos a través de la selección de características.

2 Conjunto de datos

Este conjunto de datos contiene información integral sobre 2,392 estudiantes de secundaria, detallando su demografía, hábitos de estudio, participación de los padres, actividades extracurriculares y rendimiento académico. La variable objetivo clasifica las calificaciones de los estudiantes en categorías distintas.

Estos datos se obtuvieron a través de la plataforma kaggle.com. A continuación, se presenta el enlace a este registro de datos: students performance dataset

Las variables contenidas en este conjunto de datos son las siguientes:

- Identificador del estudiante (StudentId).
- Edad (Age)
- Género (Gender)
- Grupo étnico (Ethnicity)
- Nivel educativo de los padres (ParentalEducation)
- Tiempo semanal de estudio (StudyTimeWeekly)
- Inasistencias (Absences)
- Tutoría (Tutoring)
- Apoyo parental (ParentalSupport)
- Actividades extracurriculares (Extracurricular)
- Deportes (Sports)
- Música (Music)
- Voluntariado (Volunteering)
- Promedio (GPA)
- Clasificación de la calificación (GradeClass)

Para el estudio, se omitirá la variable "Identificador del estudiante" ya que, como su nombre lo indica, se trata de un identificador y no debería tener un impacto en el desempeño del estudiante. La variable "Promedio" está en una escala del 0 al 4, y, de acuerdo con el promedio, se realiza la clasificación de la calificación como se muestra en la tabla 1.

Table 1: Clasificación de la calificación

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

3 Antecedentes

Para el desarrollo de este proyecto se tuvo en cuenta trabajos previos internacionales como los que se describen a continuación:

- En "Prediction of Student's Performance Using Machine Learning", Chauhan (2019), utilizó técnicas de regresión como KNN, árbol de decisión, SVM, random forest y regresión lineal con el objetivo de crear una herramienta de aprendizaje automático que permita predecir el GPA (Promedio de calificaciones) del estudiante en base a datos pasados (2015-2019) en el curso de Ciencias de la Computación.
- En 2020, en el trabajo titulado "Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático", Contreras et.al., plantean la selección de variables que influyen en la predicción del rendimiento en estudiantes de ingeniería industrial de la Universidad Distrital (Colombia) por diferentes metodologías: filtro, envoltura e integrados. Se implementaron algoritmos de clasificación a través del lenguaje de programación Python como árbol de decisión, K vecinos más cercanos, perceptrón y otros, los cuales son comparados para conocer el mejor resultado de predicción.
- En 2022, Gutiérrez Villaverde et.al., emplearon las técnicas de árboles de decisión, bosques aleatorios, redes neuronales y máquinas de soporte vectorial con la finalidad de evaluar la factibilidad del uso de técnicas de Aprendizaje Automático para predecir el rendimiento académico de los alumnos, considerando resultados académicos anteriores, así como variables demográficas y sociales.

4 Estadística descriptiva

A continuación, se muestra la estadística descriptiva e histogramas de las variables numéricas, así como los gráficos de barras para las variables categóricas. Asimismo, se realizó la gráfica de correlación entre las variables numéricas.

Table 2: Estadística descriptiva para las variables numéricas

	Edad	Tiempo semanal de estudio	Inasistencias	Promedio
Total de datos	2392	2392	2392	2392
Media	16.4686	9.7719	14.5413	1.9061
Desviación estándar	1.1237	5.6527	8.4674	0.9151
Mínimo	15	0.0010	0	0
25%	15	5.0430	7	1.1748
50%	16	9.7053	15	1.8933
75%	17	14.4084	22	2.6222
Máximo	18	19.9780	29	4

5 Metodología

5.1 Análisis de componentes principales

La técnica de *análisis de componentes principales* propone la transformación del conjunto a un nuevo conjunto sintético de variables que no están correlacionados y se encuentran ordenados de mforma que

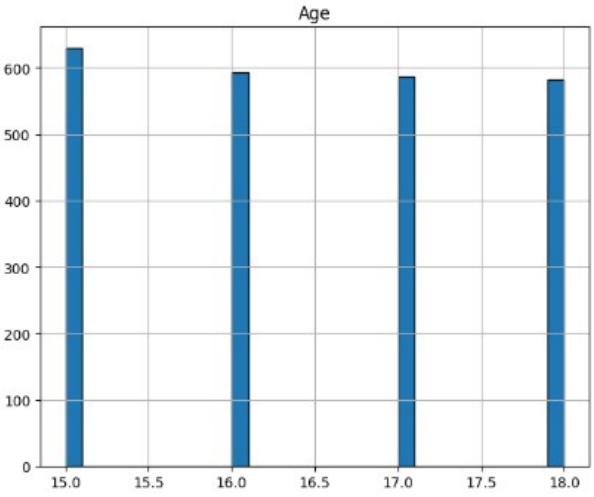

Figure 1: Histograma de Edad

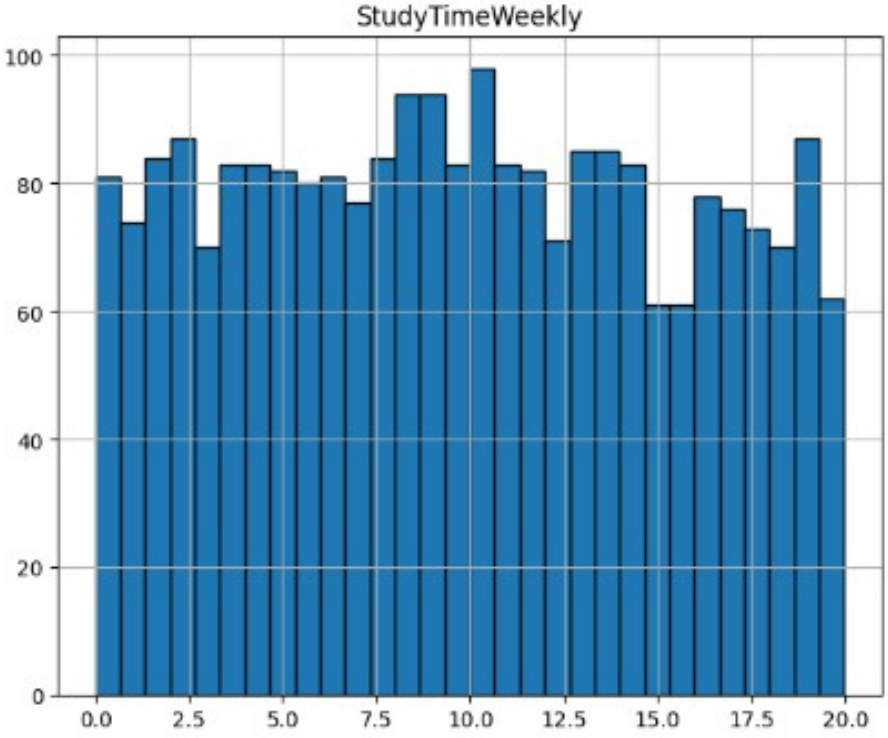

Figure 2: Histograma de Tiempo semanal de estudio

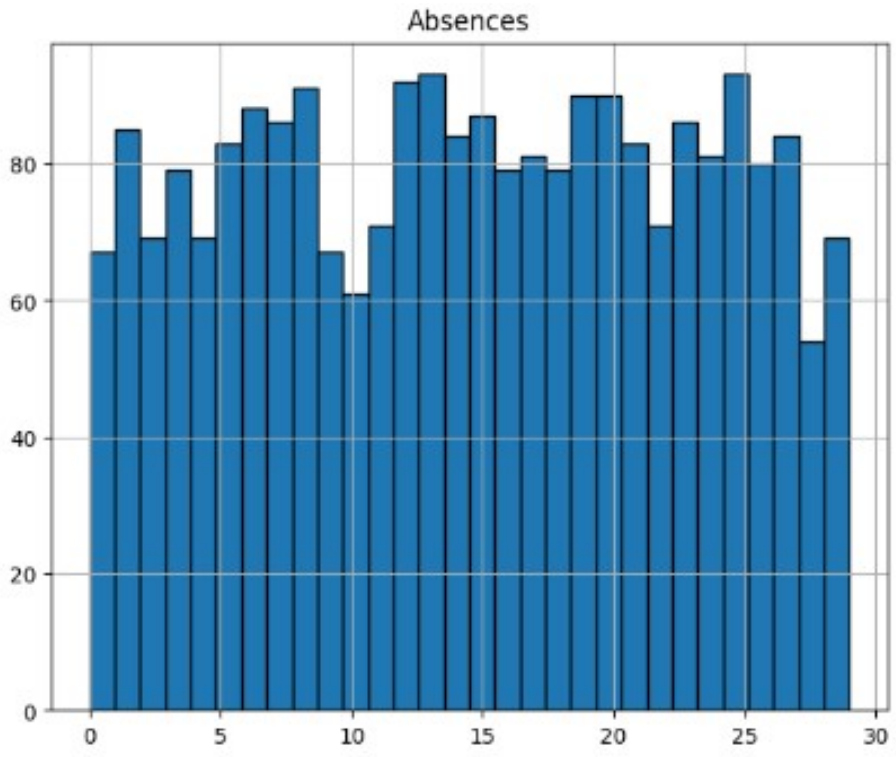

Figure 3: Histograma de Inasistencias

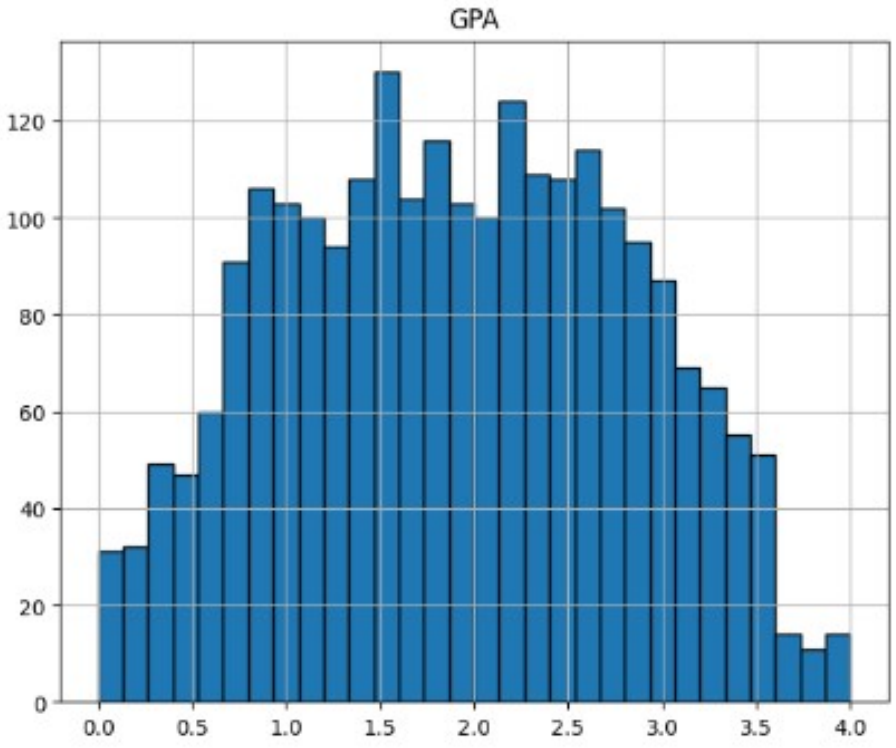

Figure 4: Histograma de Promedios

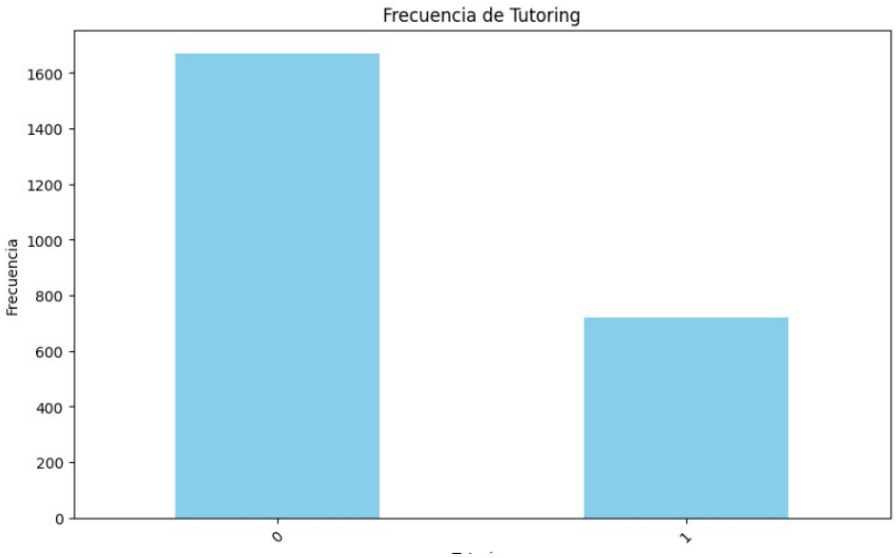

Figure 5: Frecuencia de Género

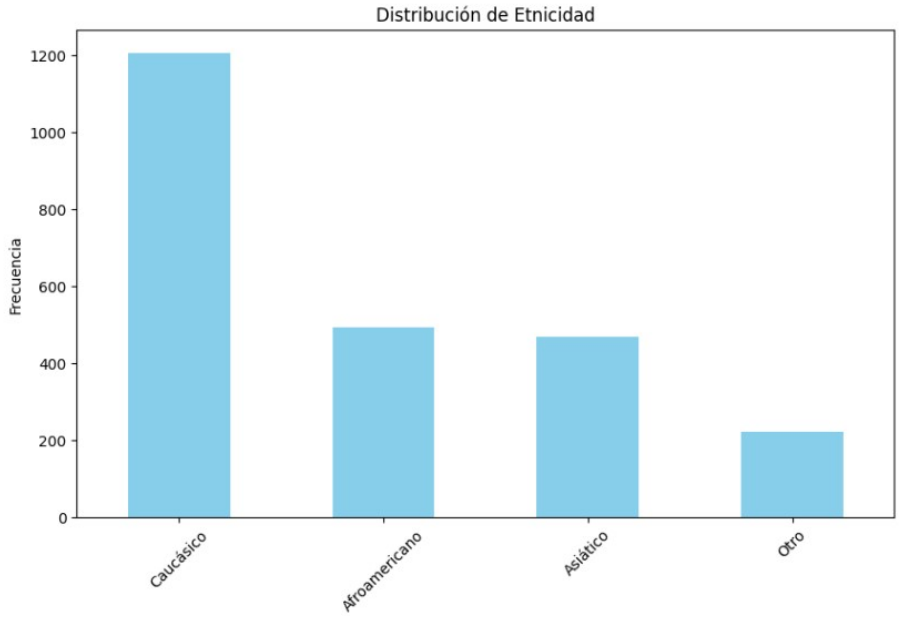

Figure 6: Frecuencia de Grupo étinco

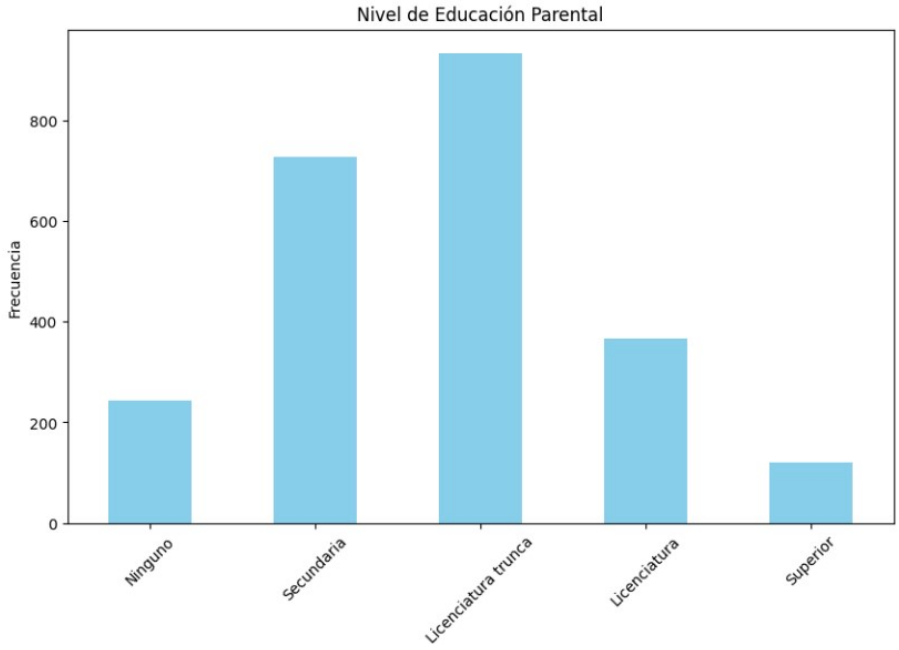

Figure 7: Frecuencia de Nivel educativo de los padres

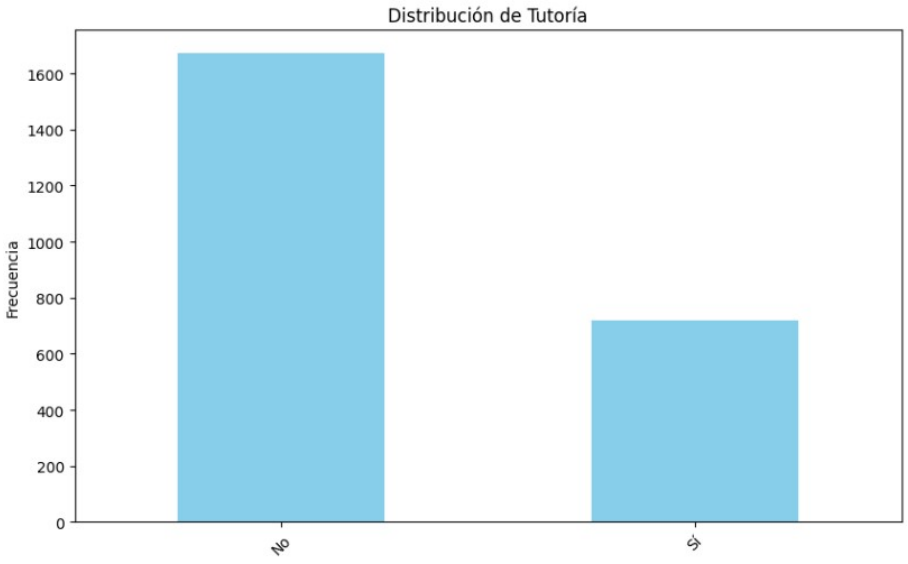

Figure 8: Frecuencia de Tutoreo

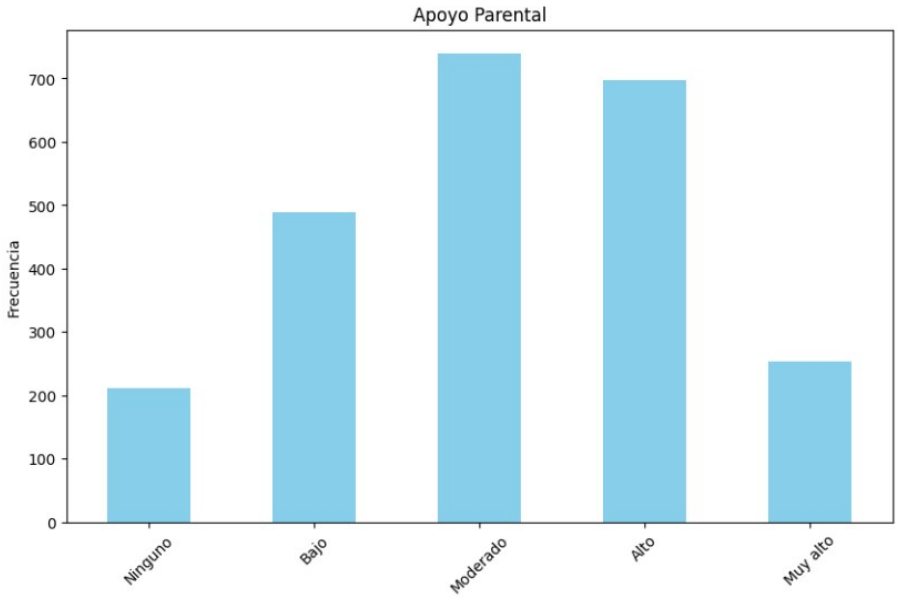

Figure 9: Frecuencia de Apoyo parental

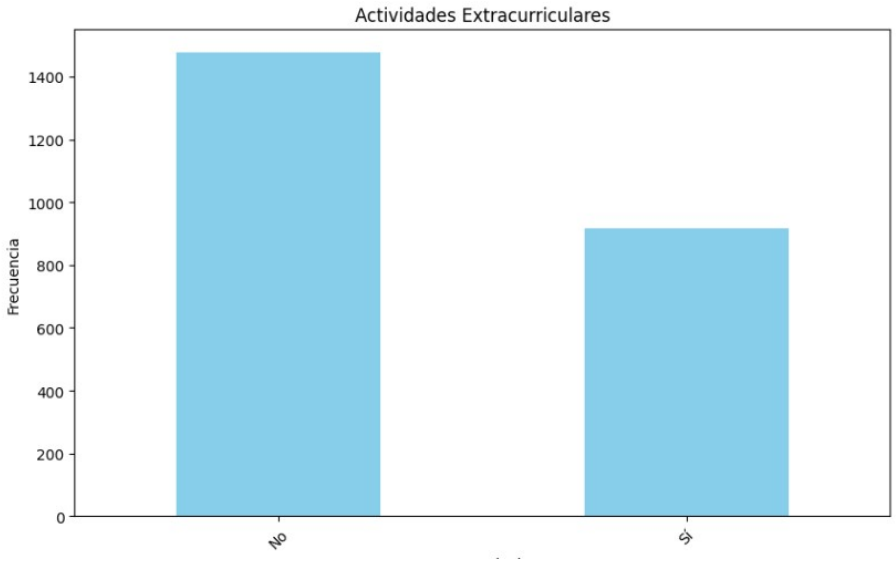

Figure 10: Frecuencia de Participación en act. extracurriculares

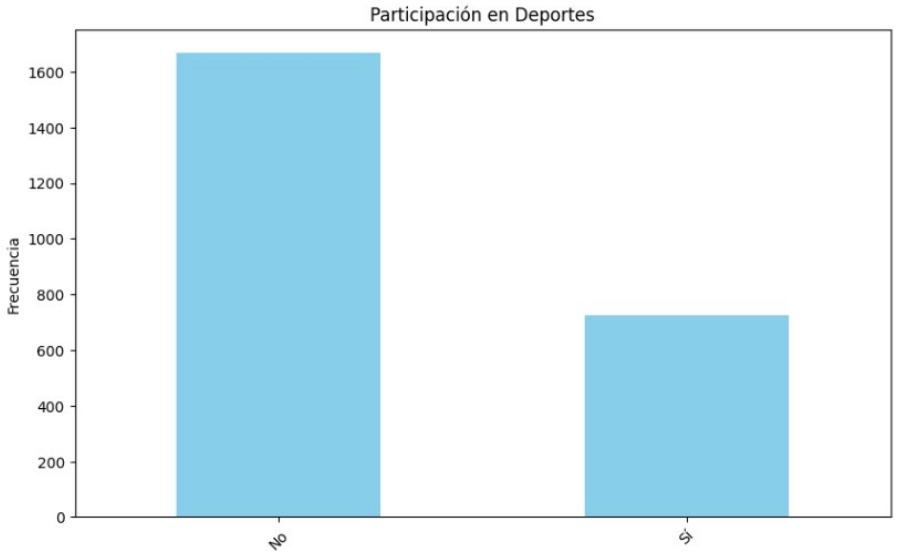

Figure 11: Frecuencia de Participación en deportes

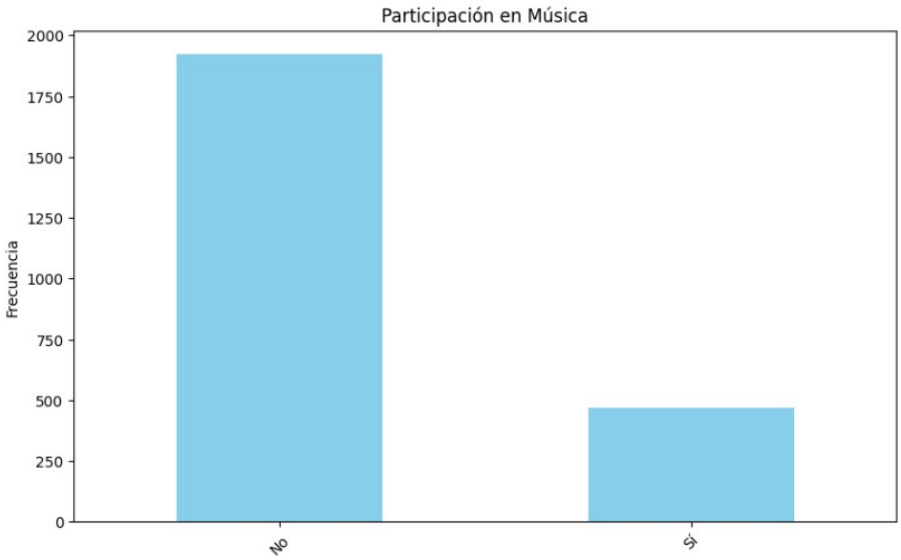

Figure 12: Frecuencia de Participación en música

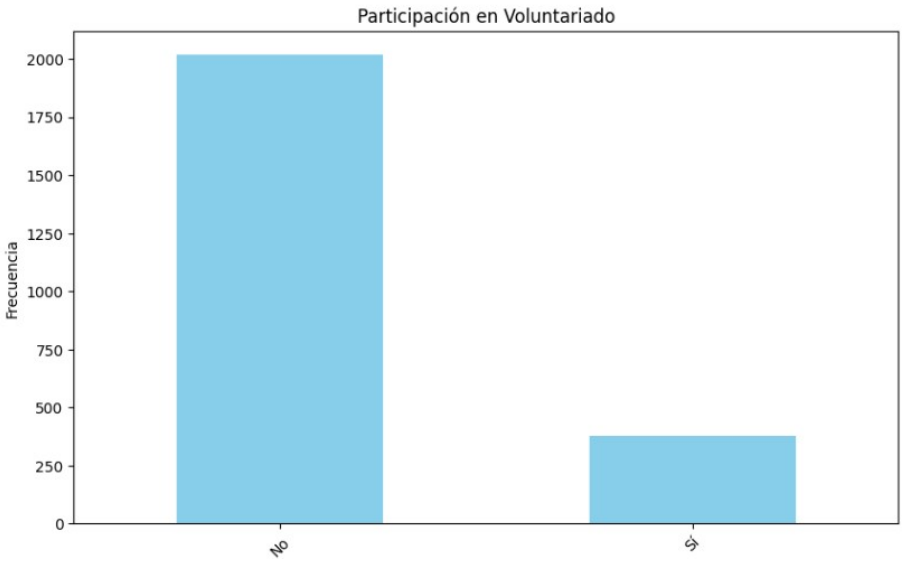

Figure 13: Frecuencia de Participación en voluntariado

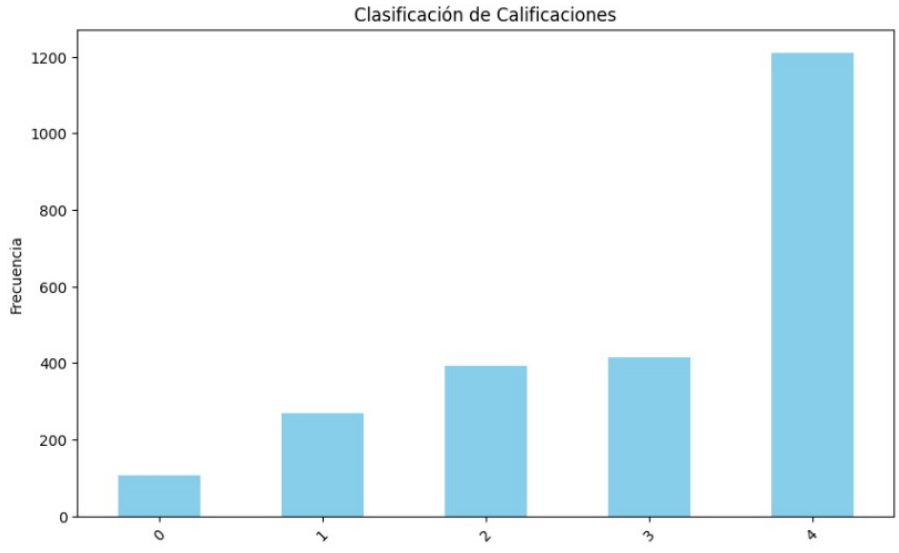

Figure 14: Frecuencia de Clasificación de la calificación

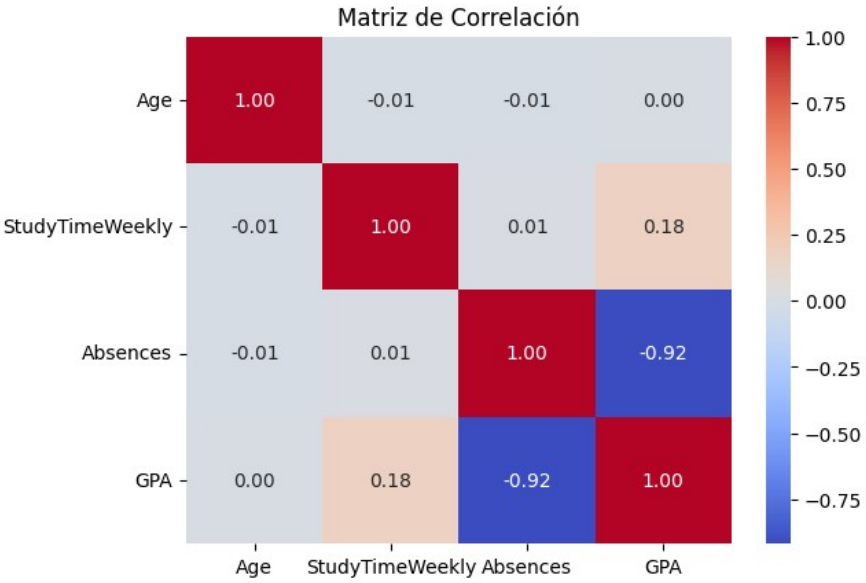

Figure 15: Matriz de correlación

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Figure 16: n = individuos, p = variables

los primeros conservan la mayor parte de la variación en todas las variables originales.

Se trata de una técnica estadística multivariada de simplificación. Las nuevas variables son combinaciones linealmente independientes de las variables originales.

Consideremos un grupo de variables $x_1, x_2, ..., x_p$ cada uno con el mismo número de muestras; cada variable registra información en un vector del mismo tamaño el cual, matricialmente da el 100% de la información y se puede representar matricialmente como se muestra en la figura 16.

La comparación de dos individuos i y j es evaluada con la distancia euclidiana clásica entre:

$$d^2(i,j) = \sum_{P=1}^{p} (x_{i_P} - x_{j_P})^2 \tag{1}$$

A partir de esta matriz calculamos un nuevo conjunto $(C_1, C_2, ..., C_p)$ no correlacionados entre si, cuyas varianzas van decreciendo progresivamente. Cada elemento C_j (j=1,. . . , p) representa una combinación lineal de las variables originales $(x_1, x_2, ..., x_p)$

$$C_j = a_{j,_1} x_1 + a_{j,_2} x_2 + ... + a_{j,_p} x_p = a'_j x \tag{2}$$

donde

$$a'_j = a_{1_j}.a_{2_j}, ..., a_{p_j} \tag{3}$$

$$X = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix},$$

a'_j es una matriz de constantes y representa el peso de las variables en cada componente; las componentes explican una parte de la varianza total, pretendiendo encontrar k componentes que representen casi toda la varianza de la nube de información.

El primer componente principal representa la mayor varianza observada sujeta a cumplir la condición de ortonormalidad; el segundo componente principal se obtiene calculando los valores de a_2 de tal manera que C_1 y C_2 sean no correlacionados (ortogonales). Igualmente se eligen los siguientes componentes cada uno con menor varianza que el anterior.

5.2 Gradient Boosting

Gradient Boosting es una técnica de aprendizaje automático que se utiliza para problemas tanto de regresión como de clasificación. Es un método basado en la idea de mejorar iterativamente modelos más débiles, comúnmente árboles de decisión, agregándolos de forma secuencial para crear un modelo final robusto y preciso. La clave de este enfoque es que cada nuevo modelo intenta corregir los errores cometidos por los modelos anteriores.

5.3 Bosques aleatorios

El algoritmo Random Forest (Breiman, 2001) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento. Los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado. Cada árbol se obtiene mediante un proceso de dos etapas:

Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que m ¡ M (donde M = total de predictores).

Cada árbol crece hasta su máxima extensión.

Cada árbol generado por el algoritmo Random Forest se construye a partir de un conjunto aleatorio de observaciones, seleccionadas mediante la técnica de bootstrap. Este método estadístico permite que una observación pueda aparecer en más de una muestra. Las observaciones que no son seleccionadas para un árbol específico (conocidas como "out of the bag") se utilizan para validar el modelo. Las predicciones de todos los árboles se combinan para producir una única salida final, conocida como ensamblado. Esta combinación se realiza generalmente mediante el promedio si las salidas de los árboles son numéricas, o mediante el conteo de votos si las salidas son categóricas.

5.4 Máquinas de vectores de soporte (SVM)

La teoría de las Máquinas de Soporte Vectorial (SVM por su nombre en inglés *Support Vector Machines*) es una técnica de clasificación. Una SVM primero mapea los puntos de entrada a un espacio de características de mayor dimensión y encuentra un hiperplano que los separe y maximice el margen m entre las clases en este espacio.

De acuerdo con Jara et. al. (2016), la SVM se fundamenta en un espacio de hipótesis de funciones, donde el cambio de dimensión es facilitado por un kernel. Este kernel eleva las características originales a una dimensión superior. Una de las formas más comunes en que las SVM aprenden es mediante el mapeo de las entradas X a un espacio de características en el que la clasificación resulta mucho más sencilla.

5.5 Ajuste de Hiperparámetros

Al entrenar modelos de machine learning, cada conjunto de datos y cada modelo requieren un conjunto específico de hiperparámetros, los cuales son variables que determinan el comportamiento del algoritmo de aprendizaje. La única manera de identificar estos hiperparámetros óptimos es mediante la realización de múltiples experimentos. En cada experimento, se selecciona un conjunto de hiperparámetros y se ejecuta el modelo con ellos. Este proceso se conoce como ajuste de hiperparámetros. En esencia, implica entrenar el modelo de manera secuencial utilizando diferentes combinaciones de hiperparámetros, con el objetivo de encontrar la configuración que maximice el rendimiento del modelo.

La búsqueda de rejilla, del inglés Grid Search (GS) funciona evaluando el producto cartesiano de un conjunto finito de valores definido por el usuario. Es fácil ejecutar GS en paralelo porque cada ensayo se ejecuta individualmente y el resultado es independientes de los de otros ensayos. Sin embargo, GS sufre la maldición de la dimensionalidad porque el consumo de recursos informáticos aumenta exponencialmente cuando hay hiperparámetros.

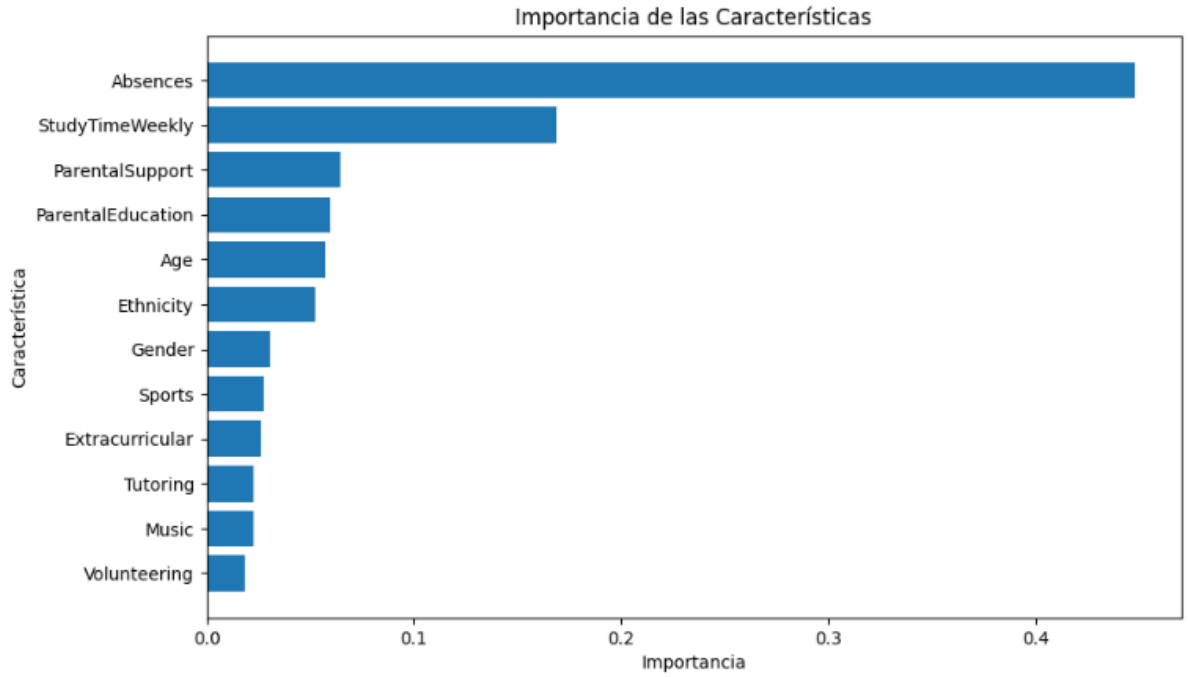

Figure 17: Importancia de las caraacterísticas

6 Principales Métricas de clasificación

La matriz de confusión es una tabla que describe el rendimiento de un modelo supervisado de ML en los datos de prueba, donde se desconocen los verdaderos valores. Se llama "matriz de confusión" porque hace que sea fácil detectar dónde el sistema está confundiendo dos clases.

A partir de los valores obtenidos de la matriz de confusión, se obtienen diferentes métricas que permitirán evaluar el modelo.

- Exactitud (Accuracy): Proporción de predicciones correctas sobre el total de predicciones.
- Precisión (Precision): Proporción de verdaderos positivos sobre el total de elementos clasificados como positivos.
- Recall (Sensibilidad o Tasa de Verdaderos Positivos): Proporción de verdaderos positivos sobre el total de elementos que realmente son positivos.
- F1 Score: Promedio armónico de la precisión y el recall, útil cuando necesitas un balance entre ambos.
- AUC-ROC: Área bajo la curva ROC (Receiver Operating Characteristic), que mide la capacidad del modelo para distinguir entre clases positivas y negativas.

7 Resultados y conclusiones

Los resultados obtenidos con el método de bosques aleatorios se muestran a continuación.

De acuerdo con lo anterior, se revisó la exactitud del modelo con un umbral de importancia de 0.1. Los resultados se muestran en la Tabla 3.

Luego de la reducción de características, se obtiene el mapa de calor mostrado en la Figura 18, en la que podemos observar que las características que se consideran más importantes en relación con el desempeño escolar son: inasistencias (Absences) y tiempo de estudio por semana (StudyTimeWeekly). El resto de las características no resultan tan relevantes.

En la Tabla 4, 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').

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

Table 3: Exctitud del modelo

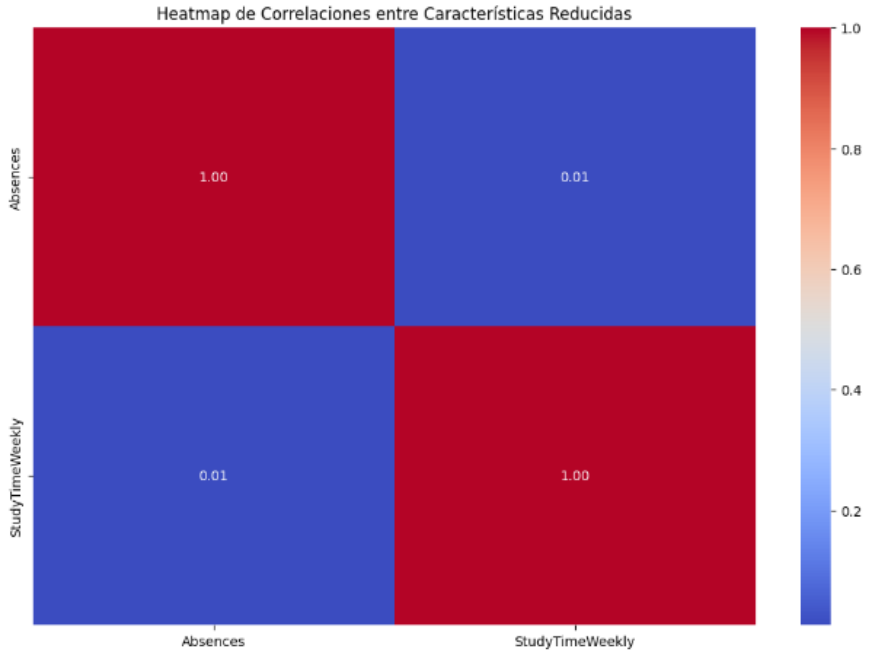

Figure 18: Mapa de calor de características reducidas

Clasificación	Precisión	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

Table 4: Reporte de clasificación

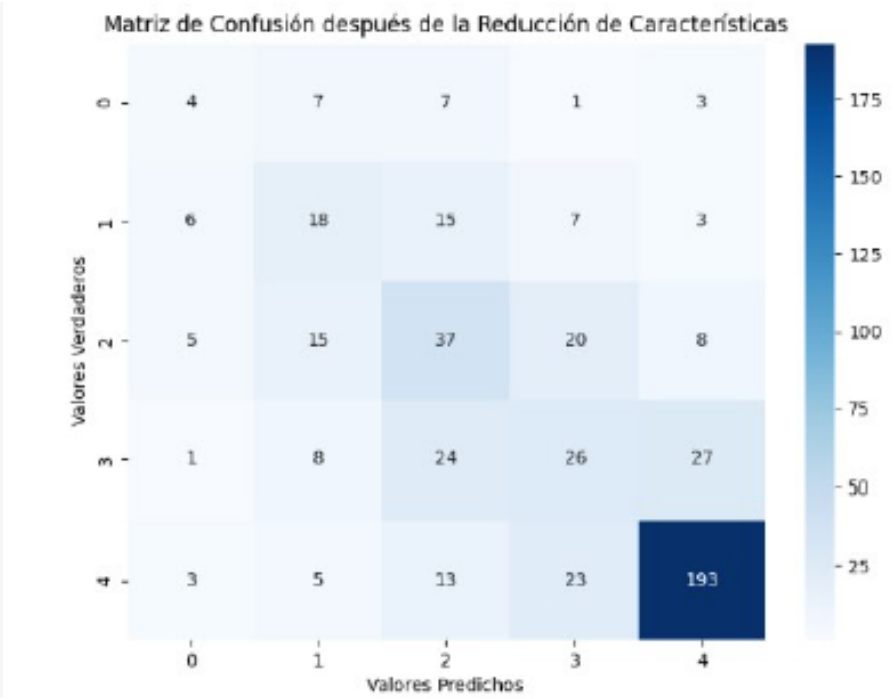

Figure 19: Matriz de confusión

Del mismo modo, la matriz de confusión (Figura 4) refleja que las predicciones correctas están sesgadas hacia la clase mayoritaria, con confusiones significativas entre las clases menos representadas.

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.

8 Bibliografía

- Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
- Contreras, Leonardo E., Fuentes, Héctor J., & Rodríguez, José I.. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. Formación universitaria, 13(5), 233-246. https://dx.doi.org/10.4067/S0718-50062020000500233 67/S0718-50062020000500233
- Chauhan, Nupur and Shah, Kimaya and Karn, Divya and Dalal, Jignasha, Prediction of Student's Performance Using Machine Learning (April 8, 2019). 2nd International Conference on Advances in Science & Technology (ICAST) 2019 on 8th, 9th April 2019 by K J Somaiya Institute of Engineering & Information Technology, Mumbai, India, Available at SSRN: https://ssrn.com/abstract=3370802 or http://dx.doi.org/10.2139/ssrn.3370802
- Betancourt, Gustavo. (2005). LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs). Scientia Et Technica.
- López-Sarmiento, D. A., Manta-Caro, H. C., & Vera-Parra, N. E. (2013). Clasificador Basado en una Máquina de Vectores de Soporte de Mínimos Cuadrados Frente a un Clasificador por Regresión Logística ante el Reconocimiento de Dígitos Numéricos. TecnoLógicas, (31), 37-51.
- Awad, Mariette & Khanna, Rahul. (2015). Support Vector Machines for Classification. 10.1007/978-1-4302-5990-9_3.
- Sánchez-Jiménez, Eduardo & Hernandez, Yasmin & Ortiz, Javier. (2022). Técnicas de Optimización de Hiperparámetros en Modelos de Aprendizaje Automático para Predicción de Enfermedades Cardiovasculares.