

Tarea 6: Aprendizaje no supervisado

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				
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	0.042	0.42	479
Weighted Avg	0.58	0.58	0.58	479

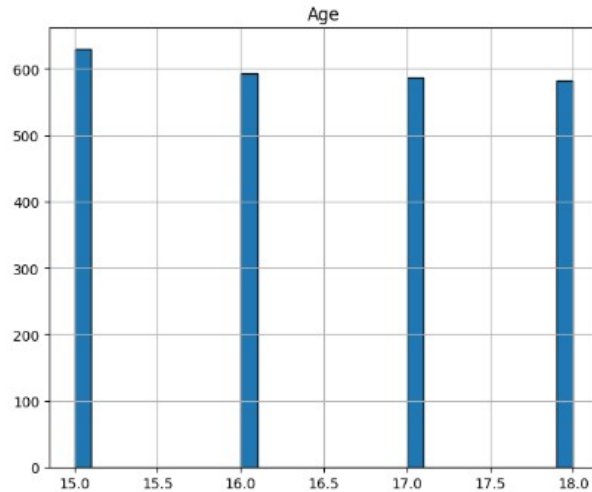


Figure 1: Histograma de Edad

3 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

4 Metodología

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

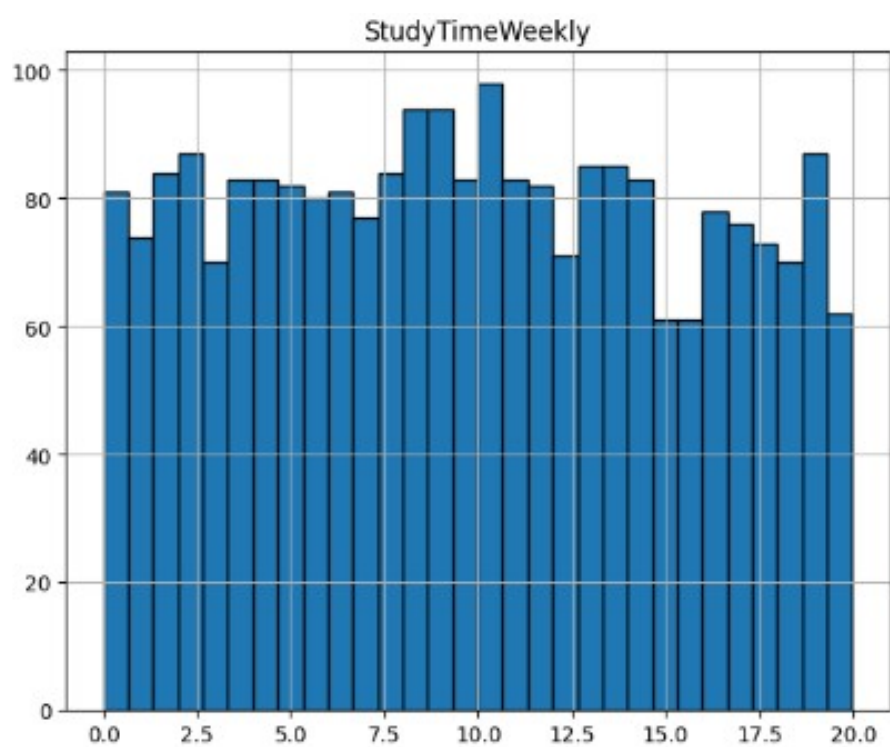


Figure 2: Histograma de Tiempo semanal de estudio

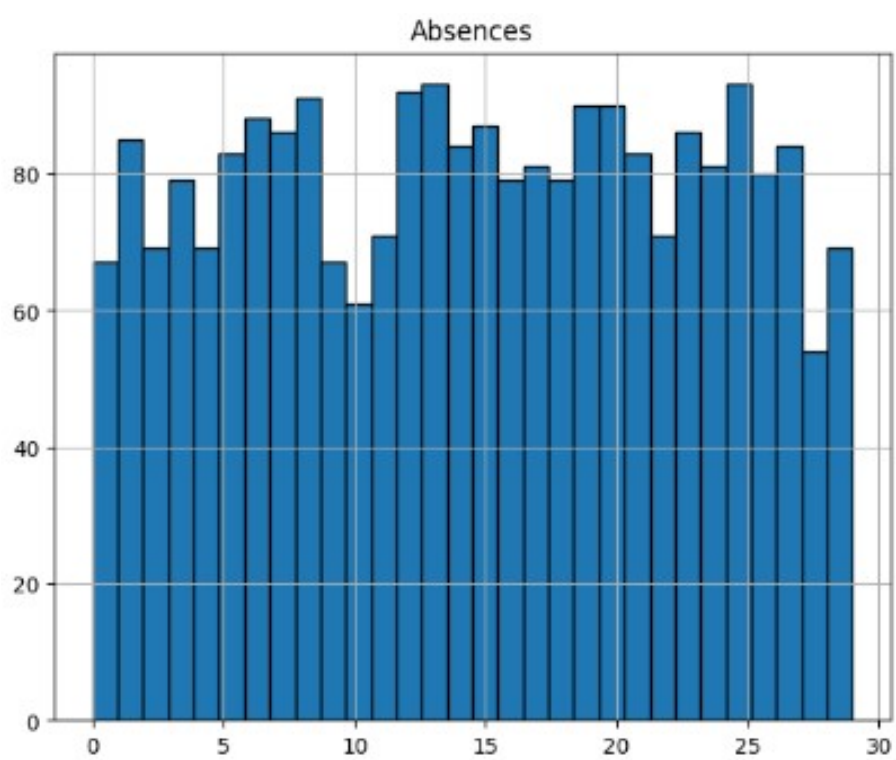


Figure 3: Histograma de Inasistencias

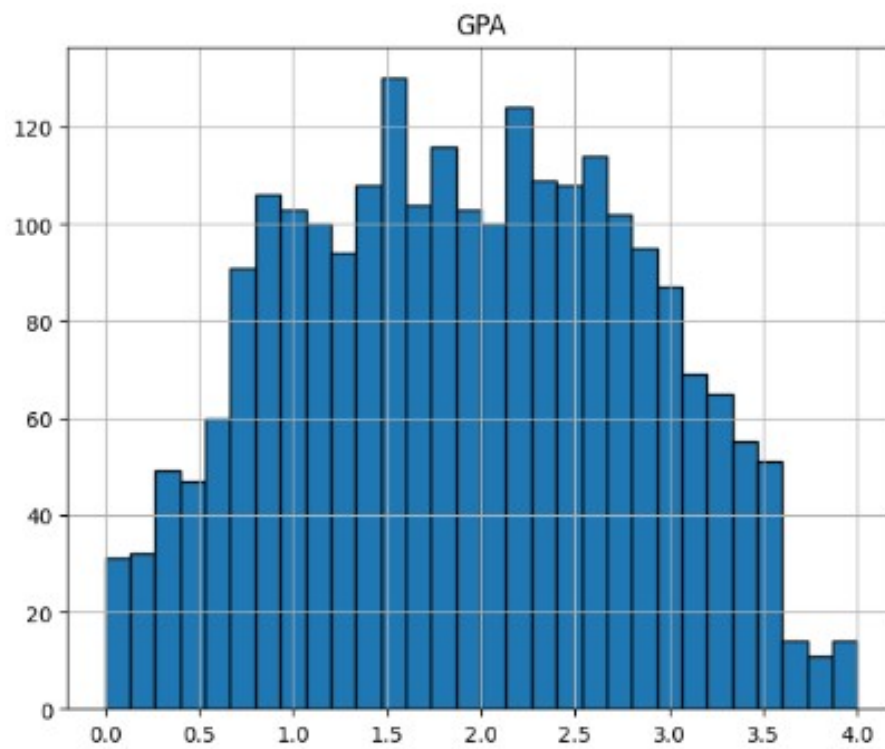


Figure 4: Histograma de Promedios

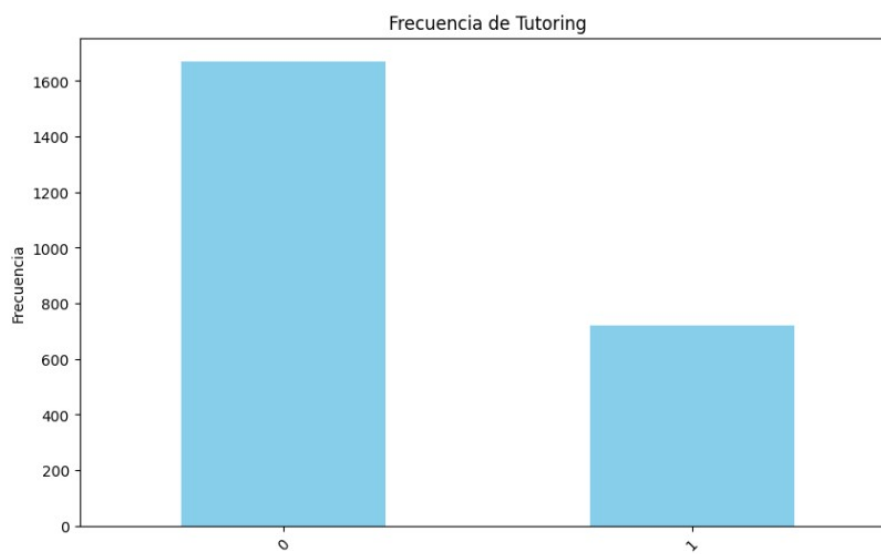


Figure 5: Frecuencia de Género

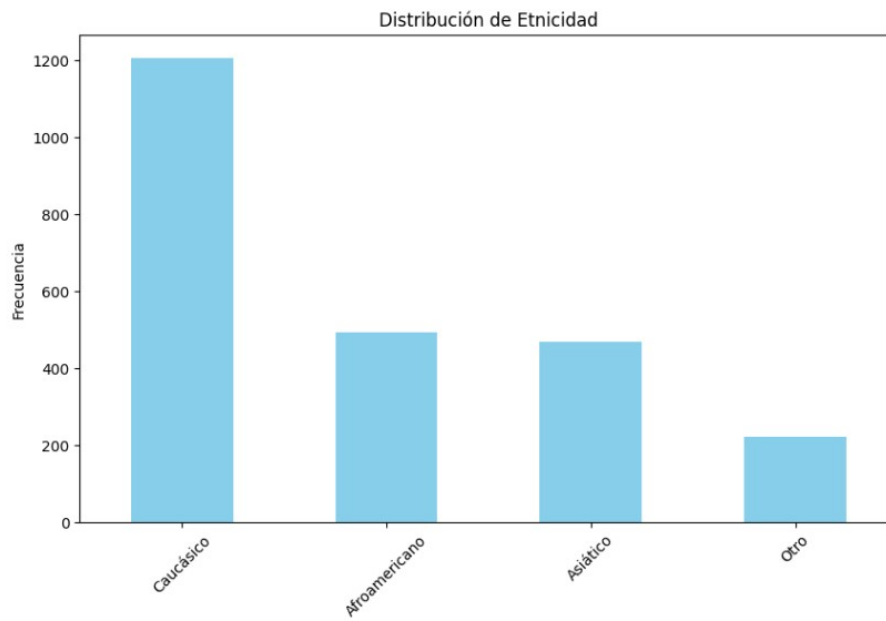


Figure 6: Frecuencia de Grupo étnico

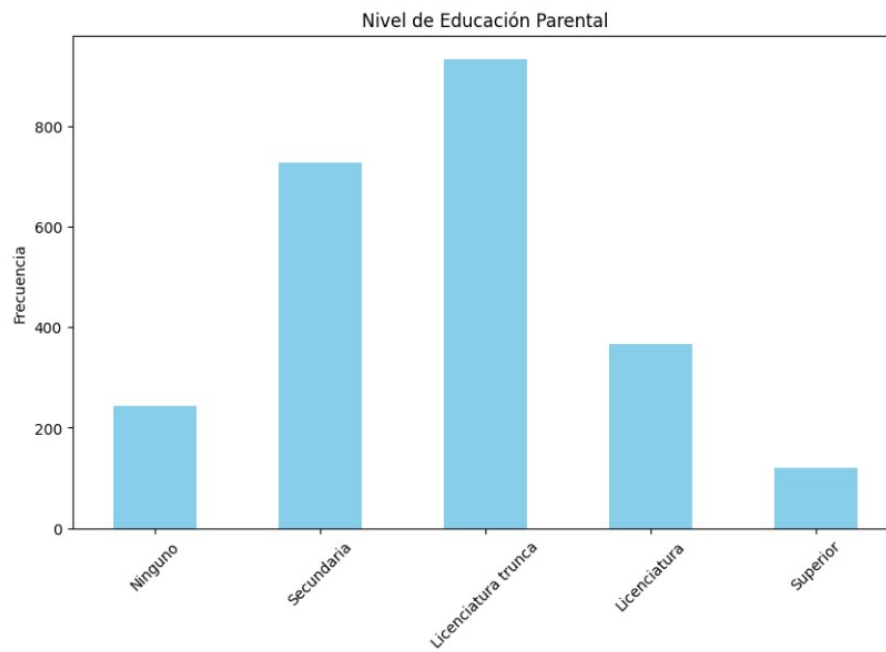


Figure 7: Frecuencia de Nivel educativo de los padres

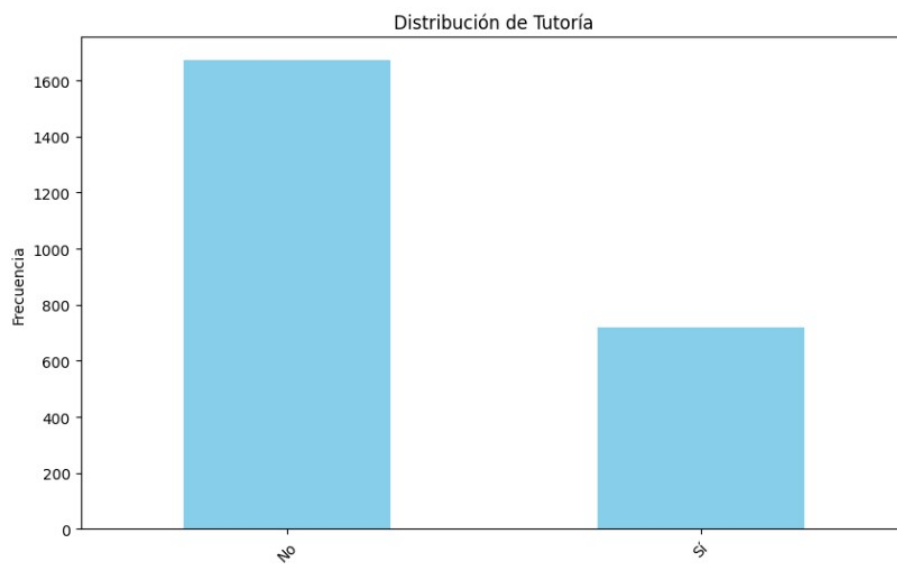


Figure 8: Frecuencia de Tutorio

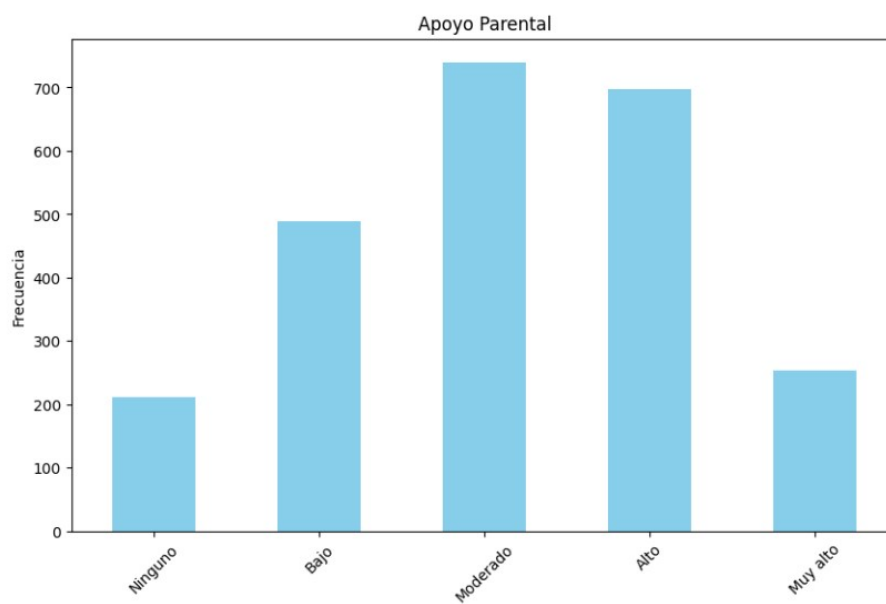


Figure 9: Frecuencia de Apoyo parental

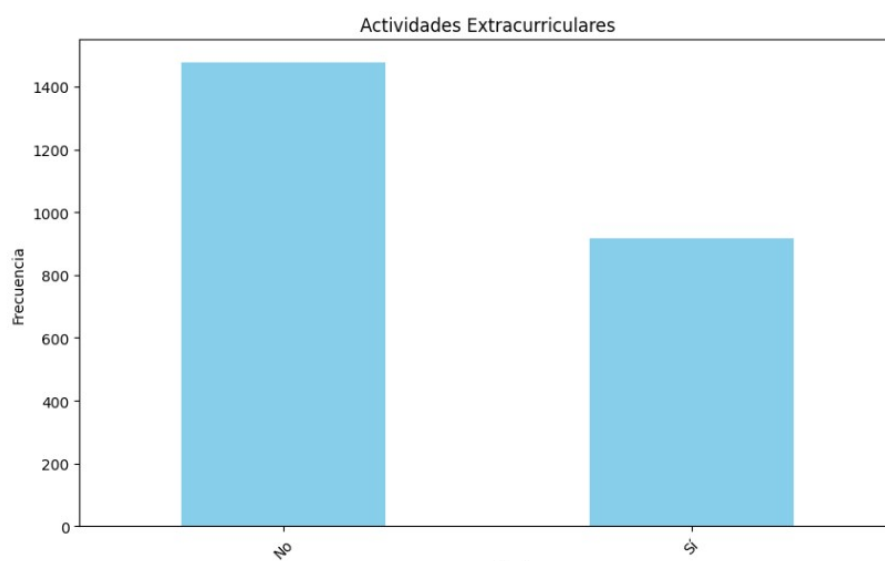


Figure 10: Frecuencia de Participación en act. extracurriculares

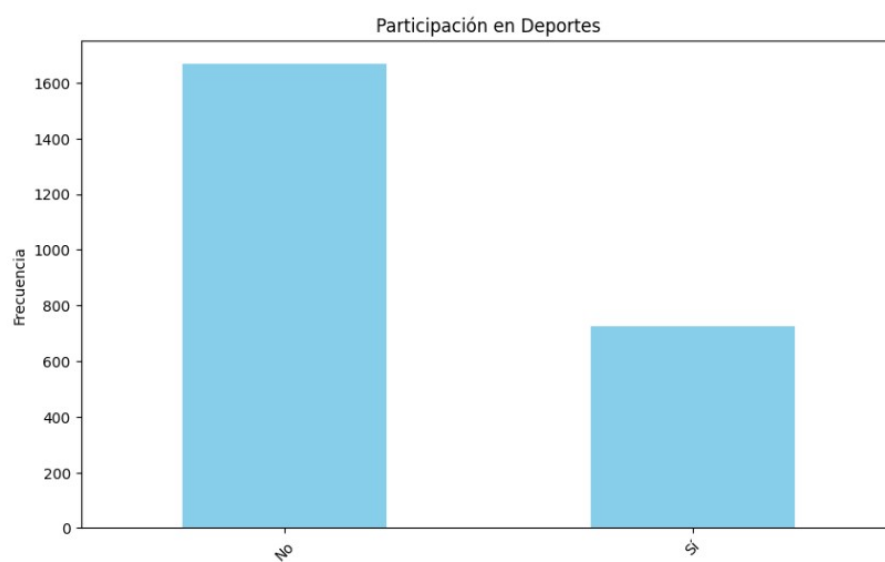


Figure 11: Frecuencia de Participación en deportes

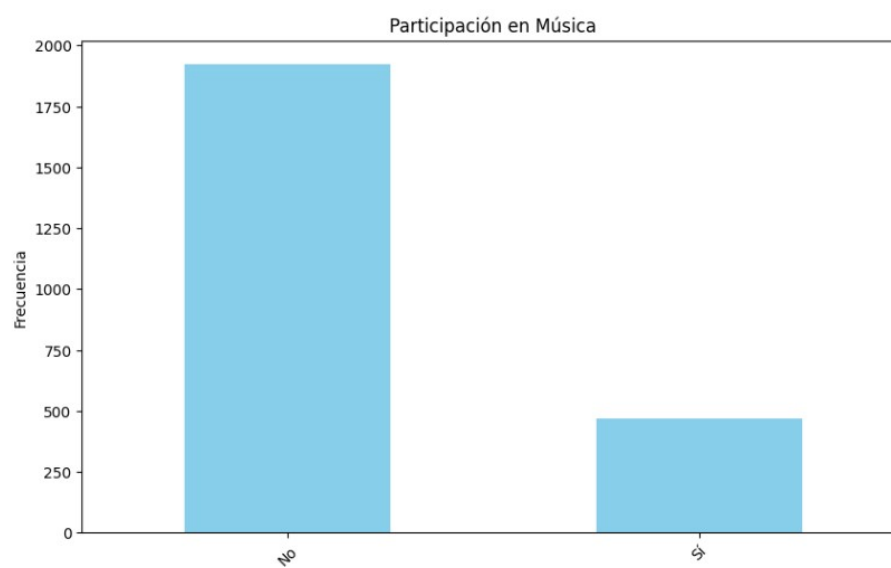


Figure 12: Frecuencia de Participación en música

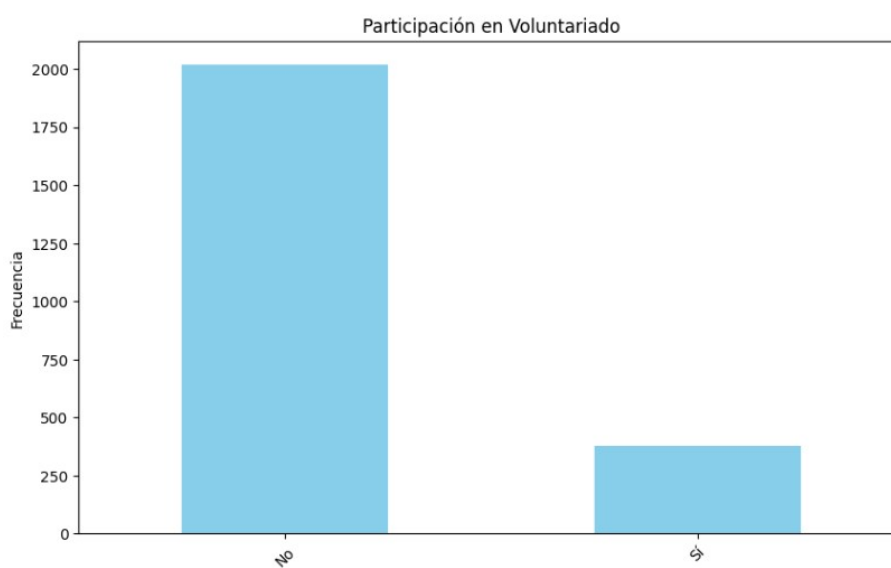


Figure 13: Frecuencia de Participación en voluntariado

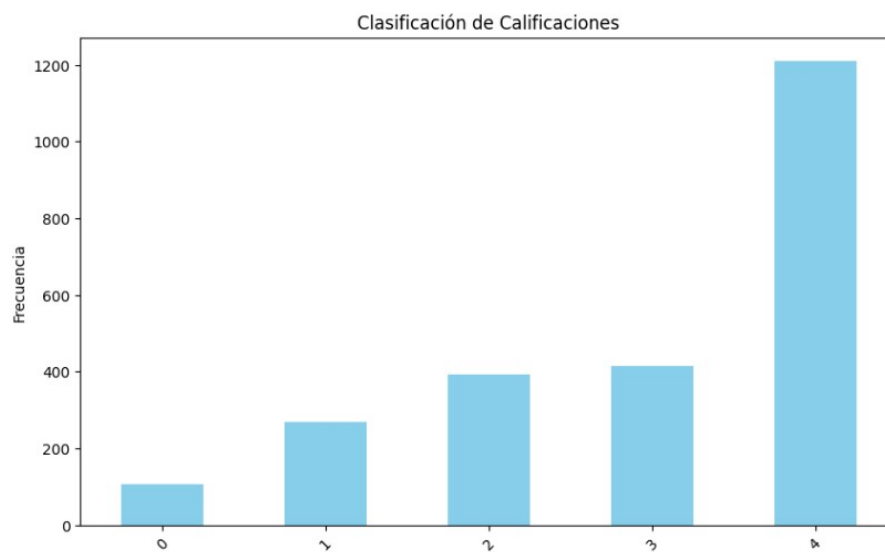


Figure 14: Frecuencia de Clasificación de la calificación

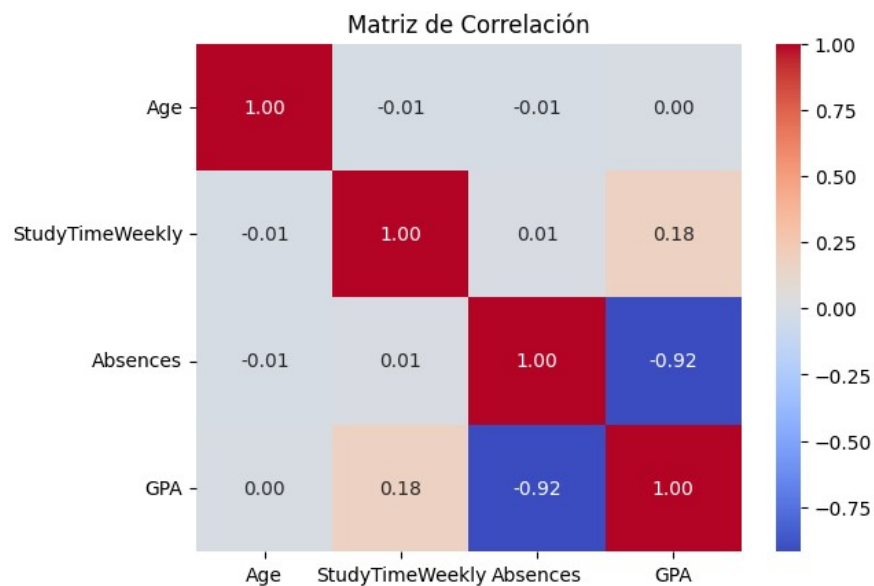


Figure 15: Matriz de correlación

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

Figure 16: n = individuos, p = variables

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, \dots, 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:

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

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

$$C_j = a_{j,1}x_1 + a_{j,2}x_2 + \dots + a_{j,p}x_p = a'_j x \quad (2)$$

donde

$$a'_j = a_{1,j}, a_{2,j}, \dots, a_{p,j} \quad (3)$$

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

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.

4.2 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.

4.3 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 $m \ll M$ (donde M = total de predictores).

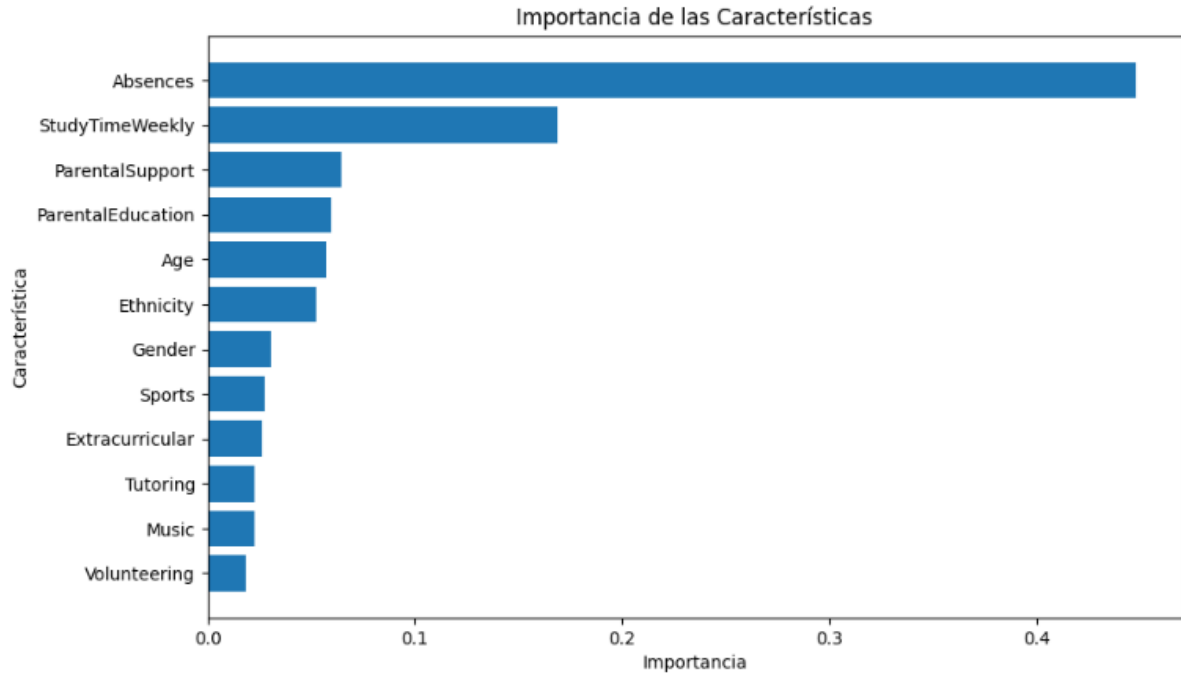


Figure 17: Importancia de las características

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

Table 3: Exactitud del modelo

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 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').

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.

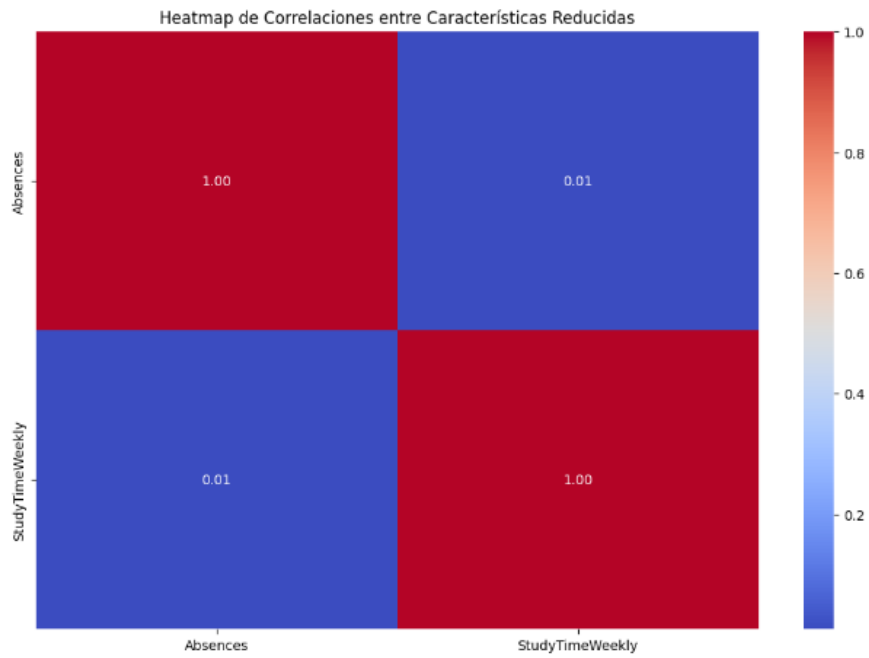


Figure 18: Mapa de calor de características reducidas

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	0.42	0.42	479
Weighted Avg	0.58	0.58	0.58	479

Table 4: Reporte de clasificación

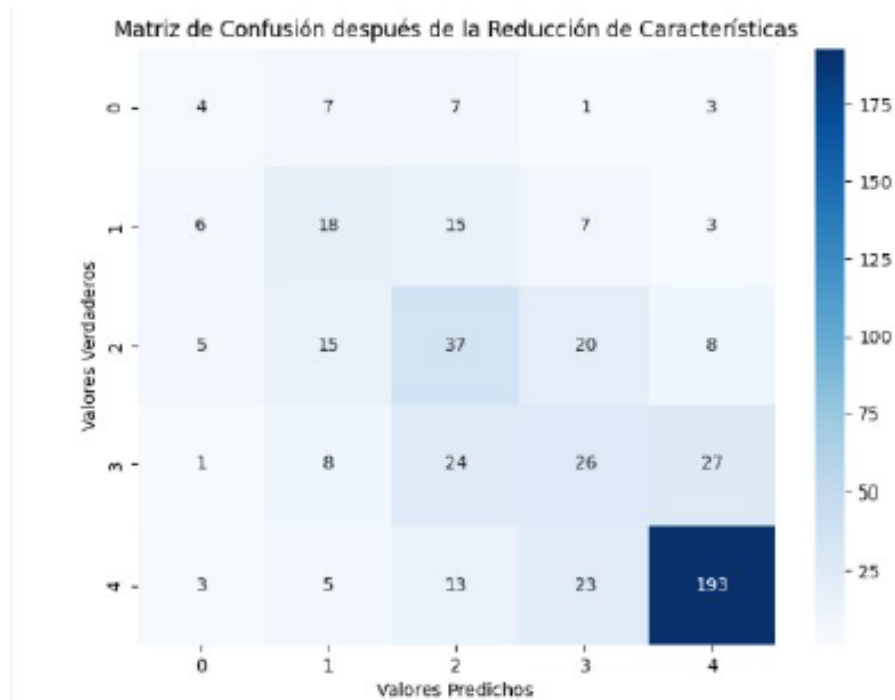


Figure 19: Matriz de confusión

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.

6 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/S0718-5006202000050023367/S0718-50062020000500233>