Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Aprendizaje Automático

Artículo sobre predicción del rendimiento académico

Jessica Lizeth Hernández Bracho – 1842553

13 de noviembre de 2025

1. Introducción

El rendimiento académico de los estudiantes es un tema de interés constante en el ámbito educativo, ya que permite identificar los factores que influyen en el desempeño y, a partir de ello, diseñar estrategias que favorezcan la mejora del aprendizaje. En este contexto, el presente estudio se basa en el conjunto de datos Student Performance disponible en el repositorio UCI Machine Learning Repository, el cual recopila información académica, demográfica y social de estudiantes de secundaria de Portugal.

Particularmente, se trabajará con información correspondiente a la asignatura de Matemáticas, que contiene datos de aproximadamente 395 estudiantes y un total de 33 variables, las cuales abarcan aspectos personales, familiares, escolares y académicos.

Este estudio combina enfoques estadísticos y de aprendizaje automático para ofrecer una visión integral sobre los factores que afectan el rendimiento académico, contribuyendo al entendimiento de cómo ciertas variables pueden influir significativamente en los resultados educativos.

2. Descripción de los datos

"Student Performance Data Set"del UCI Machine Learning Repository, recopilado por Cortez 2008, contiene información sobre estudiantes de secundaria en Portugal, incluyendo variables personales, familiares, escolares y las calificaciones obtenidas en distintos periodos.

Para este trabajo, se seleccionaron únicamente las variables directamente asociadas al rendimiento académico con el fin de realizar análisis exploratorios, pruebas de hipótesis, selección de características y la construcción de modelos predictivos.

Datos

Datos Personales del Estudiante

Variable	Descripción	Tipo
sex	Género del estudiante (F = femenino, M = masculino)	Categórica
age	Edad del estudiante (15 a 22 años)	Numérica
address	Tipo de residencia (U = urbana, R = rural)	Categórica
famsize	Tamaño de la familia (LE3 = 3, GT3 = >3 miembros)	Categórica

Cuadro 1: Datos personales del estudiante

Contexto Familiar

Variable	Descripción	Tipo
Medu	Nivel educativo de la madre (0: ninguno, 4: universitario)	Numérica ordinal
Fedu	Nivel educativo del padre (0: ninguno, 4: universitario)	Numérica ordinal
Mjob	Trabajo de la madre (teacher, health, etc.)	Categórica
Fjob	Trabajo del padre	Categórica
guardian	Tutor principal del estudiante (mother, father, other)	Categórica

Cuadro 2: Contexto familiar

Información Escolar

Variable	Descripción	Tipo
studytime	Tiempo de estudio semanal (1: <2h, 4: >10h)	Numérica ordinal
failures	Número de materias reprobadas (0–3)	Numérica
schoolsup	Apoyo educativo adicional en la escuela (yes o no)	Categórica
famsup	Apoyo educativo familiar (yes o no)	Categórica
internet	Acceso a internet en casa (yes o no)	Categórica

Cuadro 3: Información escolar

Factores de Comportamiento y Hábitos

Variable	Descripción	Tipo
activities	Participa en actividades extracurriculares (yes o no)	Categórica
goout	Frecuencia con la que sale con amigos (1–5)	Numérica ordinal
freetime	Tiempo libre despúes de la escuela (1–5)	Numérica ordinal
Walc	Consumo de alcohol en fin de semana (1–5)	Numérica ordinal
health	Estado de salud autoevaluado (1–5)	Numérica ordinal

Cuadro 4: Factores de comportamiento y hábitos

Rendimiento Académico

Variable	Descripción	Tipo
G1	Nota obtenida en el primer periodo (0–20)	Numérica
G2	Nota obtenida en el segundo periodo (0–20)	Numérica
G3	Nota final (tercer periodo) (0–20)	Numérica

Cuadro 5: Rendimiento académico

3. Antecedentes

Los alumnos que obtienen un buen rendimiento académico son aquellos que alcanzan calificaciones óptimas para aprobar sus asignaturas, siendo una medida para valorar sus capacidades, que expresa lo aprendido durante su preparación. En este ámbito, existen varios factores relacionados con el rendimiento académico de los estudiantes y la calidad de la educación superior, entre los cuales se destacan: la infraestructura educativa, el sistema de evaluación docente, el entorno social, familiar y económico del estudiante, la malla curricular, entre otros que inciden en su propio rendimiento y la calidad de la educación.

Existen diversas investigaciones que abordan el análisis de este tipo de datos, una publicación particularmente relevante es la de Ramírez Lemus et al. [2], en donde se propone un enfoque similar al que se plantea en este trabajo, pues se evalúan distintas combinaciones y análisis al conjunto de datos con enfoque cuantitativo.

4. Metodología

El análisis se desarrollará a través de los siguientes pasos:

1. **Evaluación del tipo de variable:**

   Inicialmente se seleccionó las siguientes variables de interés para el análisis, por la importancia dentro del tema:

   - **Age:** Edad del estudiante.
   - **Studytime:** Tiempo dedicado al estudio.
   - **Failures:** Número de materias reprobadas anteriormente.
   - **Absences:** Número de faltas.
   - **G1, G2, G3:** Calificaciones de los tres periodos (G3 representa la calificación final).

   Se utilizó un gráfico Q-Q (Quantile-Quantile) para visualizar la distribución de cada variable y determinar si son o no paramétricas.

2. **Estadística descriptiva:** Se calcularon medidas descriptivas (media, mediana, desviación estándar, mínimos y máximos) para cada variable, con el objetivo de observar el comportamiento general de los datos.

3. **Matriz de correlación:** Se construyó una matriz de correlación entre las variables, lo que permitirá identificar relaciones lineales entre ellas.

   Se interpretarán acerca de las correlaciones más significativas, especialmente aquellas que puedan incidir en el rendimiento académico (G3).

4. **Pruebas de hipótesis:** A partir de las correlaciones observadas, se propone una hipótesis estadística, con el fin de determinar si cierta variable tiene un impacto significativo sobre la nota final del estudiante.

   Se realizó una prueba no paramétrica de Mann-Whitney U para comparar las calificaciones finales (G3) entre dos grupos de estudiantes:

   - Grupo con número de faltas bajas (menos que la media de faltas).
   - Grupo con número de faltas altas (mayor o igual que la media de faltas).

   Se elaboró un diagrama de caja comparando las calificaciones finales entre los dos grupos definidos anteriormente *(Fig. 7, pág. 11)*.

5. **Selección de características:** Se realiza una etapa de selección de variables para identificar las características más relevantes que permiten predecir la calificación final G3. Esta fase fue crucial para reducir la dimensión del problema y así, mejorar la precisión del modelo.

   Se utilizaron los siguientes métodos:

   i. **Transformación de variables categóricas**

      Se convirtieron variables con valores tipo 'yes'/'no' a formato binario (1/0) para ser utilizadas por los modelos. Por ejemplo: higher, internet, schoolsup, entre otras.

   ii. **Análisis de varianza (ANOVA - Valor F)**

      Se aplicó la prueba F de regresión (`f_regression`) para evaluar la relevancia estadística lineal de cada variable con respecto a G3. Las variables con p-valor inferiores a un umbral de significancia de $\alpha = 0.05$ fueron consideradas relevantes.

   iii. **Umbral de varianza**

      Se utilizó el método VarianceThreshold para descartar variables con muy baja dispersión, ya que estas no aportan suficiente información al modelo.

   iv. **Información mutua**

      Se aplicó la métrica de `'mutual_info_regression'` para capturar relaciones no lineales entre las variables predictoras y la variable objetivo.

   v. **Selección exhaustiva de características (EFS)**

      Se utilizó el método Exhaustive Feature Selector junto con regresión lineal como estimador, este método prueba todas las combinaciones posibles de variables dentro de un rango definido y selecciona el subconjunto que maximiza el rendimiento (en este caso, minimiza el MAE).

6. **Agrupamiento de datos:** Por lo aplicado en la sección anterior, se restablecen las variables a trabajar y las variables fueron estandarizadas utilizando StandardScaler, con el fin de evitar sesgos por diferencias en escalas numéricas, para utilizar el algoritmo no supervisado DBSCAN.

   DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)*, el cual agrupa observaciones en función de la densidad de los datos, sin necesidad de definir previamente el número de grupos (clústers) y permite identificar puntos atípicos (outliers).

Además, para evaluar la calidad de los agrupamientos, se utilizó la métrica **Silhouette Score**, que mide la relación interna de los grupos (clústers) frente a su separación entre sí.

Se utilizó un diagrama de caja para visualizar la distribución de la calificación final (G3) entre los distintos grupos formados por DBSCAN, lo que permitió observar las diferencias en rendimiento académico por clúster. *(Fig. 9, pág. 16)*.

7. **Pronóstico:** El conjunto de datos fue dividido en 80 % para entrenamiento y 20 % para prueba, esto se realizó utilizando la función `train_test_split()` de Scikit-Learn.

   Para la selección y aplicación de modelos supervisados, se aplicaron dos algoritmos de aprendizaje supervisado de tipo regresión:

   a) **Regresión Lineal Múltiple**, donde se ajustó una función lineal de la forma: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$
      Este modelo permite observar el peso de cada variable en la predicción del rendimiento final.
   b) **Bosque Aleatorio**, se aplicó como modelo complementario, capaz de capturar relaciones no lineales.
      Un método supervisado que construye múltiples árboles de decisión y promedia sus resultados para mejorar la precisión.

   Ambos modelos fueron evaluados mediante las siguientes métricas:

   - **MAE** *(Mean Absolute Error)*: Error promedio entre los valores reales y predichos.
   - **RMSE** *(Root Mean Squared Error)*: Representa la desviación estándar de los residuales (errores de predicción).

   Por último, se realizaron visualizaciones con gráficos de dispersión y de barras para analizar la calidad de las predicciones y comparar el rendimiento de los modelos.

8. **Diseño de Experimentos:** Se evaluaron tres tamaños de partición para el conjunto de prueba: 20 %, 30 % y 40 %, con el fin de analizar su impacto en el desempeño de las variables.

   Además, se utilizaron los dos modelos supervisados utilizados anteriormente *(Regresión Lineal Múltiple y Bosque Aleatorio)* agregando a esto las métricas junto con R^2 (Coeficiente de determinación).

5. Resultados

Tipo de Variables

Se sustituyeron las variables de interés en una prueba rápida de normalidad gráfico Q-QPlot.

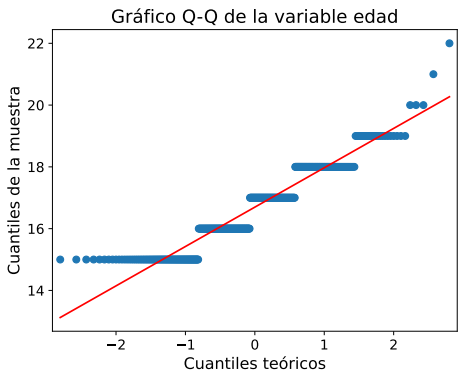


Figura 1: Gráfico QQ para la variable Edad

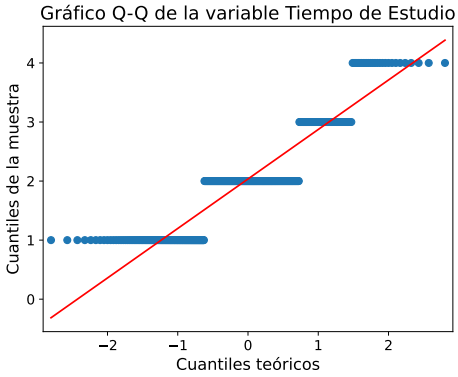


Figura 2: Gráfico QQ para la variable Tiempo de Estudio

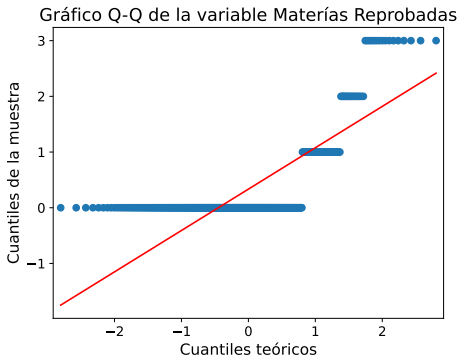


Figura 3: Gráfico QQ para la variable Materias Reprobadas

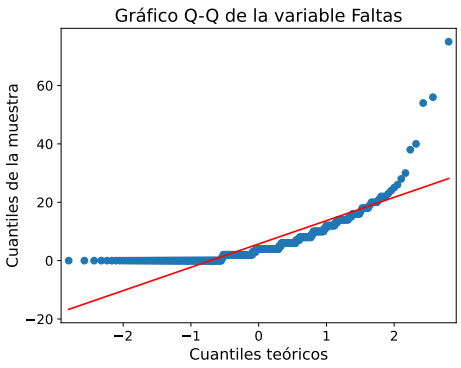
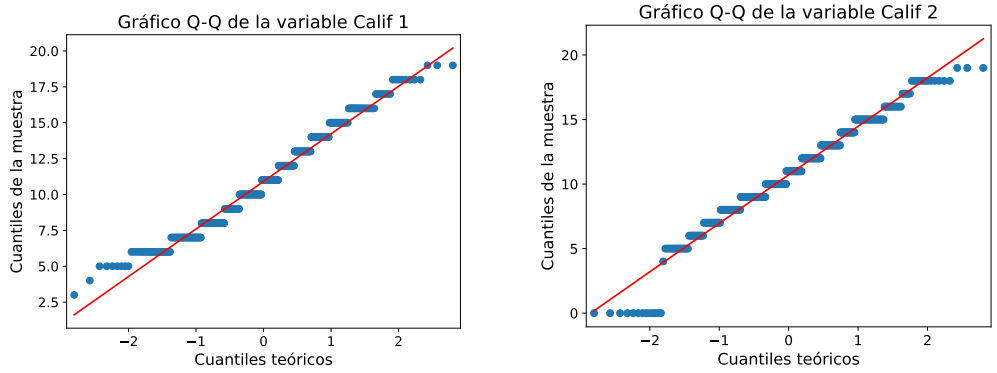
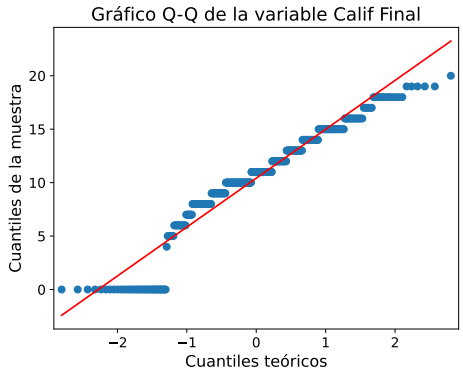


Figura 4: Gráfico QQ para la variable Faltas

Al ver los resultados de cada variable se concluye que son **datos no paramétricos**.



(a) (b)



(c)

Figura 5: Gráficas QQ variables: G1, G2 y G3 (calificaciones períodos)

Estadísticos Descriptivos

	age	studytime	failures	absences	G1	G2	G3
count	*395*	*395*	*395*	*395*	*395*	*395*	*395*
mean	16.6962	2.0354	0.3341	5.7088	10.9088	10.7139	10.4151
std	1.2760	0.8392	0.7436	8.0030	3.3191	3.7615	4.5814
min	15	1	0	0	3	0	0
25 %	16	1	0	0	8	9	8
50 %	17	2	0	4	11	11	11
75 %	18	2	0	8	13	13	14
max	22	4	3	75	19	19	20

Cuadro 6: Estadísticos Descriptivos

Correlación

Figura 6: Matriz de Correlación

Observaciones:

- La mayoría de las variables tienen una correlación positiva moderada.
- Las variables *G1, G2 y G3* (calificaciones de los periodos), tienen una correlación positiva y fuerte entre ellas.
- La variable *failures* (número de materias reprobadas anteriormente), es la que se muestra con una mayor correlación negativa en comparación con las demás.

Prueba de Hipótesis

Efecto del número de faltas en la calificación final

Con el objetivo de evaluar si el número de faltas escolares *(absences)* tiene un efecto significativo sobre la calificación final del estudiante *(G3)*, se realizó una prueba no paramétrica de Mann–Whitney U. Esta prueba se aplicó debido a que las distribuciones no cumplían con los supuestos de normalidad.

H_0: No hay diferencia significativa en $G3$ entre los grupos.
H_1: Hay diferencia significativa en $G3$ entre los grupos.

La muestra se dividió en dos grupos, tomando como punto de corte la media de faltas en el conjunto de datos:

- Grupo de faltas bajas: Estudiantes con un número de faltas inferior a la media.
- Grupo de faltas altas: Estudiantes con un número de faltas igual o superior a la media.

Se compararon las calificaciones finales *(G3)* de ambos grupos. Los resultados fueron los siguientes:

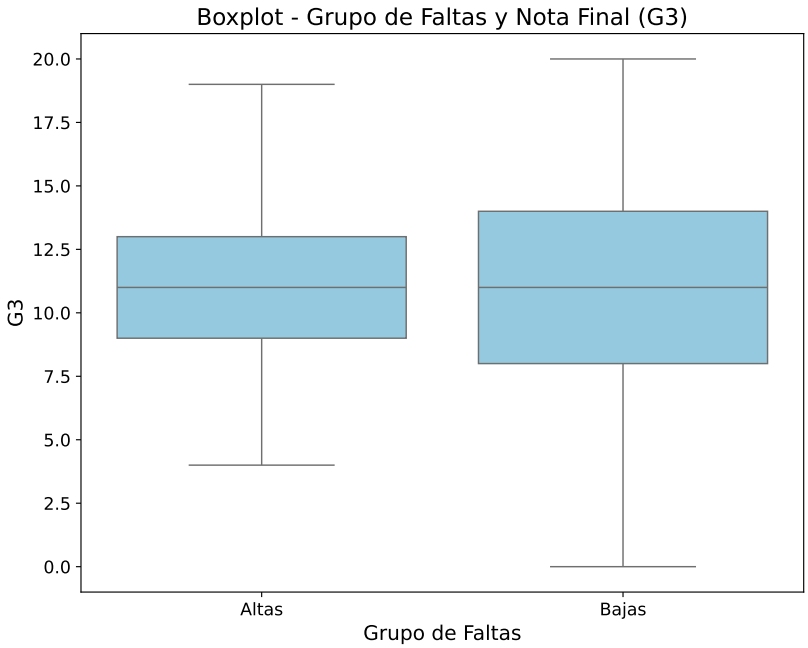
- Media de G3 en el grupo de faltas bajas: 10.17
- Media de G3 en el grupo de faltas altas: 10.84
- Estadístico U = 18,595
- Valor-p = 0.7019

Dado que el valor-p (0.7) es mayor a 0.05, **no se rechaza la hipótesis nula**.
Lo que indica que no existe una diferencia estadísticamente significativa en las calificaciones finales entre los estudiantes con menos faltas y aquellos con más faltas.

En la gráfica 7, pag. 11 se comparan las calificaciones finales (G3) entre los dos grupos de estudiantes con número de faltas bajas y altas.

Ambos grupos presentan una media muy similar, el número de faltas no parece afectar de forma clara la nota final, lo que concuerda con los resultados de la prueba estadística realizada, donde no se encontró una relación significativa entre las ausencias y el rendimiento final.

Figura 7: Diagrama de Caja - Comparación de Faltas y Nota Final (G3)



Selección de Características

A través de distintos métodos de selección de características, se identificaron las variables más relevantes para predecir la calificación final (G3) de los estudiantes.

- **ANOVA *(valor F)* -** Las variables con mayor relevancia estadística lineal para predecir G3 fueron:
  - **G2** (calificación del segundo periodo)
  - **G1** (calificación del primer periodo)
  - **Failures** (número de materias reprobadas)
- **Varianza -** Las variables G1 *(calificación del primer periodo)* y failures *(número de materias reprobadas)* presentaron una varianza considerable, lo cual respalda su utilidad al aportar información a los modelos. En cambio, variables como higher mostraron baja varianza, aunque se mantuvieron por su valor interpretativo.

– **Información mutua -** Este análisis mostró que variables como studytime, absences y higher resaltan, aunque no tan fuertes con el valor F, tienen una dependencia no lineal significativa con G3.

	valor_f	varianza	mir
G2	1775.7075	0.0331	0.0382
G1	705.8422	0.0780	0.0196
failures	58.6716	0.0612	0.0482
higher	13.5349	0.0113	0.1289
age	10.5354	0.0480	0
studytime	3.7968	0.0429	0.8105
absences	0.4614	0.0390	1.3948

Cuadro 7: Resultados: ANOVA-Varianza-MIR

– **Selección exhaustiva *(EFS)* -** El método EFS identificó que la mejor combinación de variables para minimizar el error fue:

  - **Failures** (número de materias reprobadas)
  - **G2** (calificación del segundo periodo)

  Esto coincide con los hallazgos previos, reforzando su importancia como predictores clave del rendimiento final. *(Fig. 8, pag. 13)*

– **Conclusión de la selección de variables**

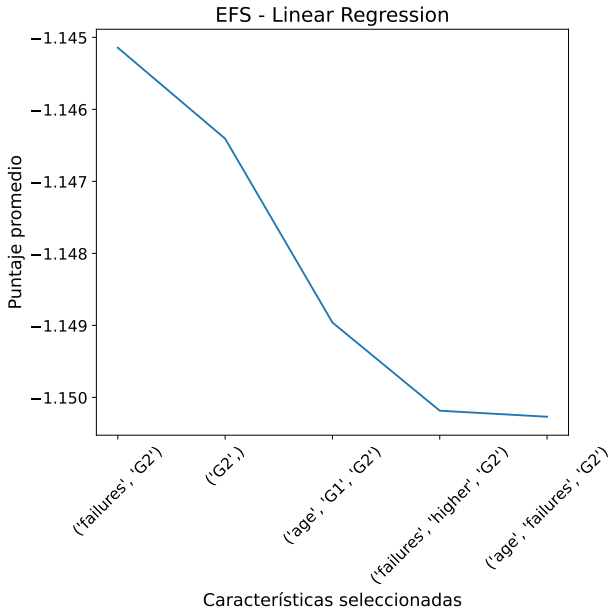
  Variables predictoras:

  - Studytime (tiempo de estudio).
  - Absences (faltas).
  - Failures (materías reprobadas).
  - G1 (calificación primer periodo).
  - G2 (calificación segundo periodo).

  Variable de respuesta:

  - G3 (nota final del curso).

Figura 8: Selección exhaustiva (EFS)



Agrupamiento de datos

Dado los resultados previos se restablecen las variables a trabajar, por lo que en este apartado se aplicará un algoritmo no supervisado para conocer mejor su densidad. Las variables fueron estandarizadas utilizando StandardScaler, con el fin de evitar sesgos por diferencias en escalas numéricas.

Algoritmo no supervisado: DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise), este es un método de agrupamiento basado en densidad porque encuentra un número de grupos (clústers) comenzando por una estimación de la distribución de densidad de los nodos correspondientes. [3]

- ***Modelo matemático de DBSCAN***

DBSCAN agrupa puntos en regiones de alta densidad y clasifica como valores atípicos (outliers) aquellos que se encuentran en zonas de baja densidad. Los parámetros principales son:

- *eps(ϵ):* radio de vecindad, es decir, la distancia máxima para que dos puntos se consideren vecinos.

- *min pts:* número mínimo de puntos requeridos para que una región se considere densa, sirve para formar un clúster (grupo).

El algoritmo comienza seleccionando un punto no visitado y determina si tiene al menos puntos (min pts) dentro de la distancia eps (ϵ). Si es así, se forma un clúster y se expanden los puntos vecinos densos, de lo contrario, el punto se clasifica como ruido (outlier).

Así, DBSCAN identifica automáticamente grupos de forma arbitraria y detecta valores atípicos.

- ***¿Por qué DBSCAN conviene para estos datos?***

En el conjunto de datos Student Performance, las características como tiempo de estudio, ausencias, calificaciones y número de materias reprobadas tienden a tener una alta diversidad agregando las variables de ausencias extremas o bajo rendimiento muy marcado.

-No requiere conocer a priori el número de clústers (grupos), lo cual es práctico dado que no se sabe cuántos perfiles de estudiantes podrían existir.

-Detecta automáticamente casos atípicos, ayudando a identificar estudiantes con comportamientos extremos (por ejemplo, altos niveles de ausencias).

- ***Silhouette Score***

Se utilizó la métrica Silhouette Score como estrategia para identificar la cantidad óptima de grupos en los datos, es adecuada para algoritmos como DBSCAN, ya que se refiere a un método de interpretación y validación de la coherencia dentro del análisis de grupos.

Esta métrica evalúa qué tan bien está asignado cada punto dentro de su grupo donde el valor oscila entre -1 a 1, donde un valor alto (cerca del 1) indica que el objeto está bien emparejado con su propio cúmulo. [4]

- ***Aplicación - Resultados***

Tras evaluar múltiples configuraciones del parámetro eps (ϵ), se identificó que la configuración óptima fue:

- eps (ϵ)= 2.0

- min pts (`min_samples`)= 5
- Número de clústers encontrados: 2
- Número de outliers detectados: 6
- Silhouette Score: Se obtuvo un valor aceptable (0.521), lo que sugiere una separación clara entre los grupos formados.

Descripción de los Clústers
El algoritmo identificó tres grupos:

(a) **Clúster 0**

- Número de estudiantes: 372
- Perfil: Estudiantes con buen rendimiento académico (promedio $G3 \approx 10.6$), pocas ausencias, pocas materias reprobadas y alta motivación para continuar estudios superiores. Representan el grupo más estable y exitoso.

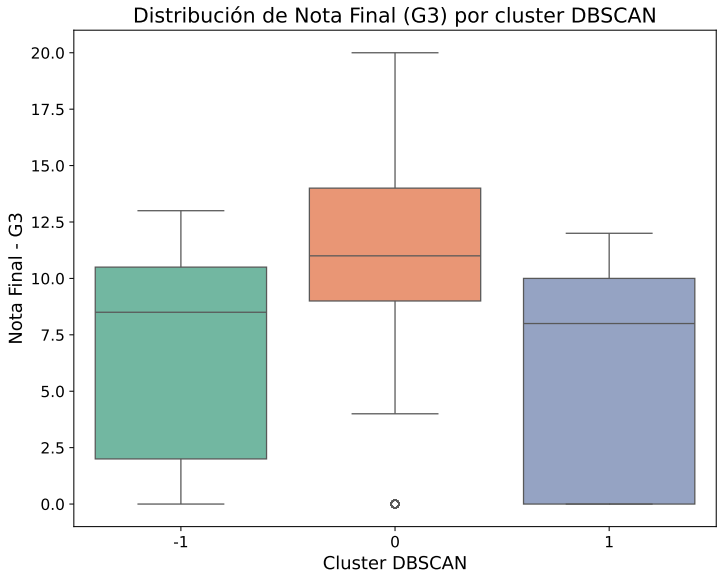
(b) **Clúster 1**

- Número de estudiantes: 17
- Perfil: Estudiantes con bajo rendimiento académico (promedio $G3 \approx 6.7$), mayor número de materias reprobadas y escaso interés en educación superior. Este grupo podría representar estudiantes en riesgo académico.

(c) **Clúster -1 (outliers)**

- Número de estudiantes: 6
- Perfil: Estudiantes con un número significativamente alto de faltas (≈ 34), comportamiento atípico frente al resto. Pueden estar influenciados por factores externos que afectan su rendimiento.

Figura 9: DBSCAN



Pronóstico

Algoritmo supervisado: Regresión Lineal

Para este proyecto se seleccionó un modelo de **regresión lineal múltiple**, una técnica supervisada que busca modelar la relación entre una variable dependiente (en este caso, la calificación final *G3*) y un conjunto de variables independientes *(tiempo de estudio, fallos previos, seguimiento de estudios, inasistencias, y las calificaciones previas (G1 y G2))*.

Este modelo asume que existe una relación lineal entre las variables predictoras y la variable de respuesta, y tiene la siguiente forma general [5]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Donde:

- *Variable dependiente* (y): Es la variable que se quiere predecir.
- *Variables independientes* $x_1, x_2, ..., x_n$*:* Son las variables que se utilizan para predecir la variable dependiente.

- *Coeficientes de regresión* $\beta_0, \beta_1, ..., \beta_n$*:*

    - β_0 (intercepto): Es el valor predicho de 'y' cuando todas las variables independientes son cero.
    - $\beta_i (i > 0)$: Miden el cambio en la variable dependiente 'y' por cada cambio de una unidad en la variable independiente correspondiente, manteniendo las demás variables constantes.

- *Término de error* ϵ*:* Representa la variabilidad en 'y' y captura las diferencias entre los valores predichos y los reales.

- ***Justificación del Modelo***

La regresión lineal múltiple fue elegida como modelo base por varias razones:

- *Relación entre variables:* Algunas variables como G1 y G2, tienen una relación evidente con G3, lo cual favorece un modelo lineal.

- *Simplicidad e interpretabilidad:* Es fácil de implementar y analizar, permite entender cómo cada variable influye en el resultado.

- *Línea base útil:* Sirve como punto de partida para comparar otros modelos más complejos. Si la regresión lineal da buenos resultados, puede ser suficiente en ciertos contextos.

- ***Métricas de Evaluación***

  Para medir el desempeño del modelo se utilizaron dos métricas comunes en problemas de regresión [6]:

- **MAE** *(Mean Absolute Error)*: Es el promedio de los errores absolutos entre las predicciones y los valores reales.
  Indica cuánto se desvía en promedio la predicción del valor real.
  Formula= $(1/n)\Sigma|y - \hat{y}_i|$

  Cuanto menor es el MAE, mejor es el modelo.

- **RMSE** *(Root Mean Squared Error)*: Es la raíz cuadrada del promedio del error cuadrático.
  Representa la desviación estándar de los residuales (errores de predicción).
  Formula = $\sqrt{(1/n)\Sigma(y - \hat{y}_i)^2}$

Un valor bajo de RMSE indica un buen ajuste.

- ***Aplicación - Resultados***
  El conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para prueba, utilizando la función `train_test_split()` de Scikit-Learn.

1. Modelo de Regresión Lineal Múltiple

La regresión asume una relación lineal entre las variables predictoras y la variable de respuesta.

El modelo ajustado obtuvo la siguiente ecuación:

$$\begin{aligned} y = & -1.872 - 0.0814\,(\text{studytime}) - 0.4309\,(\text{failures}) \\ & + 0.0397\,(\text{absences}) + 0.3027\,(\text{higher}) \\ & + 0.143\,(\text{G1}) + 0.9791\,(\text{G2}) \end{aligned}$$

Los resultados indican que las calificaciones previas *(G1 y G2)* son los predictores con mayor peso positivo sobre la nota final, mientras que el número de materias reprobadas *(failures)* tiene un efecto negativo.

En cuanto al desempeño, las métricas de evaluación fueron:

- MAE (Mean Absolute Error): 1.33
- RMSE (Root Mean Squared Error): 2.11

Estos valores reflejan un error promedio de aproximadamente 1.3 puntos en la predicción de la calificación final. Aunque el modelo logra capturar la tendencia general, se observa cierta dispersión entre los valores reales y los predichos.

La primera gráfica en la fig. 10, pág. 19 muestra la relación entre los valores reales y los predichos para el modelo lineal, donde la dispersión alrededor de la línea roja (predicción perfecta) evidencia la presencia de errores de estimación en varios puntos.

2. Bosque Aleatorio

Se aplicó el modelo Bosque Aleatorio *(Random Forest Regressor)*, un método supervisado que construye múltiples árboles de decisión y promedia sus resultados para mejorar la precisión.

En comparación con la regresión lineal, el modelo de Bosque Aleatorio presenta menores errores, lo que indica una mejor capacidad predictiva. La dispersión de

los valores reales y predichos se aproxima más a la línea ideal, reflejando una mayor precisión en la estimación de la calificación final.

- Regresión Lineal: MAE $\approx 1.33 - RMSE \approx 2.11$
- Bosque Aleatorio: MAE $\approx 1.06 - RMSE \approx 1.64$

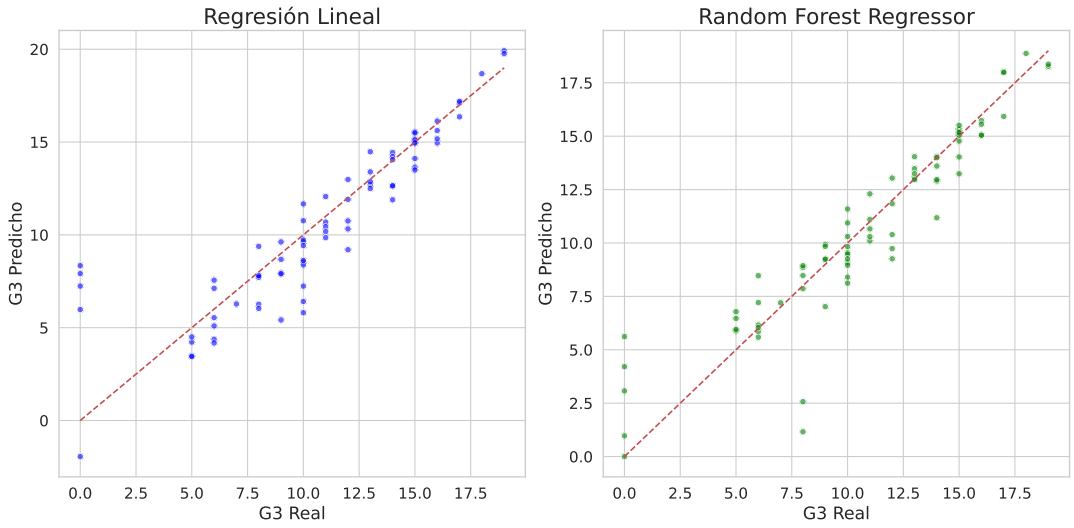


Figura 10: Regresión Lineal - Bosque Aleatorio

La Regresión Lineal Múltiple resulta útil como modelo base por su simplicidad e interpretabilidad, aunque su precisión es limitada ante comportamientos no lineales del conjunto de datos. Por el contrario, el modelo de Bosque Aleatorio demuestra un rendimiento superior en las métricas, gracias a su capacidad para modelar relaciones no lineales y manejar interacciones entre variables.

Ambos modelos permiten predecir la calificación final de los estudiantes con un error promedio menor a dos puntos, lo cual se considera un nivel de precisión aceptable para este tipo de datos educativos

Diseño de Experimentos

Se realizó un diseño factorial considerando tres elementos: modelo, normalización y tamaño del conjunto de prueba, evaluando su impacto sobre MAE, RMSE y R^2.

Factor	Descripción	Niveles
Modelo	Tipo de algoritmo utilizado	Regresión Lineal / Bosque Aleatorio
Normalización	Escalado de las variables numéricas	Sí / No
Tamaño del conjunto de prueba	Proporción de datos usados para prueba	0.2 / 0.3 / 0.4

Cuadro 8: Diseño de experimento

La normalización se aplicó únicamente a las variables numéricas continúas (studytime, failures, absences, G1, G2), excluyendo higher *(seguimiento de estudios)* por ser binaria.

Los resultados fueron los siguientes:

	Modelo	Normalización	Test Size	MAE	RMSE	R^2
0	Lineal	No	0.2	1.3290	2.1102	0.7828
1	RandomForest	No	0.2	1.0576	1.6442	0.8681
2	Lineal	Sí	0.2	1.3290	2.1102	0.7828
3	RandomForest	Sí	0.2	1.0911	1.6784	0.8626
4	Lineal	No	0.3	1.2754	2.0793	0.8033
5	RandomForest	No	0.3	0.9923	1.6152	0.8813
6	Lineal	Sí	0.3	1.2754	2.0793	0.8033
7	RandomForest	Sí	0.3	1.0138	1.6380	0.8779
8	Lineal	No	0.4	1.2541	2.0116	0.8156
9	RandomForest	No	0.4	1.0032	1.5993	0.8834
10	Lineal	Sí	0.4	1.2541	2.0116	0.8156
11	RandomForest	Sí	0.4	1.0267	1.6149	0.8812

Cuadro 9: Resultados - Diseño de Experimento

Observaciones:

- El modelo Bosque Aleatorio fue superior a la Regresión Lineal en todas las combinaciones de factores.
- Incrementar el tamaño del conjunto de prueba de 20 % a 40 % permitió que el modelo Bosque Aleatorio mostrará su mejor desempeño (R^2 = 0.8834).
- La normalización de variables numéricas no generó mejoras significativas en Bosque Aleatorio, aunque es más relevante para Regresión Lineal.

Comparando los resultados obtenidos para cada combinación de factores y recordando qué, cuanto menores sean MAE y RMSE, y mayor sea R^2, mejor será el modelo.

El modelo con mejor desempeño global fue el de Bosque Aleatorio sin normalización y con un tamaño de prueba del 40 %, alcanzando un $R^2 = 0.8834$, lo que indica aproximadamente el 88 % de la variabilidad en las calificaciones finales.

6. Conclusiones y discusión

El uso de técnicas estadísticas y métodos de aprendizaje automático permitió identificar y cuantificar los factores relacionados con el estilo de vida, comportamiento académico y bienestar personal que influyen de manera significativa en el rendimiento académico de los estudiantes, medido a través de la calificación final.

Como se vio en la matriz de correlación, las variables G1, G2 *(calificaciones de periodos)* tienen una correlacion positiva-fuerte entre ellas y 'Failures' *(número de materías reprobadas)* con una mayor correlación negativa en comparación con las demás, estás mismas fueron la mejor combinación de variables para minimizar el error en el método de Selección exhaustiva (EFS).

Si bien las variables se muestran de buena manera en las gráficas Q-QPlot, se determinó que los datos no son paramétricos, así al realizar la prueba de hipótesis se utilizó la prueba Mann–Whitney U debido a que no se cumplen los supuestos de normalidad y, según este análisis, el número de faltas escolares no parece tener un efecto determinante en el rendimiento final medido por G3.

El uso del algoritmo DBSCAN permitió segmentar a los estudiantes en grupos según su perfil académico y comportamiento escolar, estos resultados pueden ser útiles para:

- Confirmar patrones de éxito académico (Clúster 0, promedio calif. final ≈ 10.6), lo cual puede servir de referencia para prácticas educativas efectivas.
- Identificar estudiantes en riesgo (Clúster 1, promedio calif. final ≈ 6.7) y ofrecer apoyo focalizado.
- Analizar valores atípicos (Clúster -1, faltas ≈ 34), cuyas características podrían requerir intervención específica o estudio adicional.

La técnica de agrupamiento junto con el análisis de rendimiento ofrecen una herramienta útil para identificar perfiles de estudiantes, detectar casos en riesgo y orientar estrategias educativas.

Mediante la aplicación de un modelo de regresión lineal múltiple, se determinó que la ecuación del modelo es el siguiente:

$$\begin{aligned} y = & -1.872 - 0.0814\,(\text{studytime}) - 0.4309\,(\text{failures}) \\ & + 0.0397\,(\text{absences}) + 0.3027\,(\text{higher}) \\ & + 0.143\,(\text{G1}) + 0.9791\,(\text{G2}) \end{aligned}$$

El modelo obtenido presenta un coeficiente de determinación $R^2 = 0.783$, lo que indica que el 78 % de la variabilidad en las calificaciones puede explicarse por las variables seleccionadas.

Enlazado a esto, el diseño factorial permitió evaluar de manera sistemática el efecto del tipo de modelo, la normalización de variables y el tamaño del conjunto de prueba sobre el desempeño predictivo, medido mediante las métricas MAE, RMSE y R^2. Donde el modelo con mejor desempeño fue el de Bosque Aleatorio sin normalización y con un tamaño de prueba del 40 %, lo cual fue relevante como una opción más robusta y precisa para predecir las calificaciones finales dentro de las condiciones evaluadas con un 88 % de variabilidad en las calificaciones finales (R^2).

En conclusión, este estudio evidencia que la combinación de enfoques estadísticos y de aprendizaje automático puede ser una herramienta fuerte para mejorar la comprensión del proceso educativo, apoyar la toma de decisiones pedagógicas y diseñar intervenciones más efectivas que promuevan el desarrollo académico integral de los estudiantes.

Referencias

[1] Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2008. https://doi.org/10.24432/C5TG7T.

[2] Lidia Ramírez Lemus, Carlos Alberto Rodríguez Rodríguez, José Miguel Barrón-Adame, and Héctor Cuevas Vargas. Factores predominantes que influyen en el indicador de rendimiento académico en los universitarios *in situ*. *Acta universitaria*, 33(1):e3878, 2023. doi: 10.15174/au.2023.3878. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-62662023000100140.

[3] DataScientest. Machine learning & clustering: el algoritmo dbscan. https://datascientest.com/es/machine-learning-clustering-dbscan, 2023.

[4] Wikipedia. Silhouette (clustering), 2025. URL https://es.wikipedia.org/wiki/Silhouette_(clustering).

[5] GeeksforGeeks. Linear regression in machine learning. https://www.geeksforgeeks.org/ml-linear-regression/, 2025. (Última actualización: 14 Oct, 2025).

[6] Akshita Chugh. Mae, mse, rmse, coefficient of determination, adjusted r squared — which metric is better? https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e, 2020.

[7] Mishra et al. A multi-objective clustering approach based on different clustering measures combinations. *Computational and Applied Mathematics*, 44(59), 2024. doi: 10.1007/s40314-024-03004-x. URL https://www.redalyc.org/pdf/7600/760079749011.pdf.

[8] Elizabeth Guadalupe Chong González. Factores que inciden en el rendimiento académico de los estudiantes de la universidad politécnica del valle de toluca. *Revista Latinoamericana de Estudios Educativos (México)*, XLVII(1):91–108, 2017. URL https://www.redalyc.org/pdf/270/27050422005.pdf.

[9] Wikipedia. Prueba u de mann-whitney, 2025. URL https://es.wikipedia.org/wiki/Prueba_U_de_Mann-Whitney.

[10] Jason Brownlee. How to calculate nonparametric statistical hypothesis tests in python, 2019. URL https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/.