
Proyecto Final Aprendizaje Automático

Pablo Rodríguez González
Universidad Pontificia Comillas
Estudiante N° 202314114

Abstract

El objetivo de este trabajo consiste en predecir la nota final de diferentes estudiantes en Matemáticas y/o lengua a partir de características demográficas, académicas y sociales. Para ello, se ha trabajado con un conjunto de datos con las características de los estudiantes en cuestión y se han aplicado distintos modelos de regresión: regresión lineal, regresión con regularización Lasso, Random Forest y Boosting. Con ello, pretendía evaluar cuál de ellos ofrece un mejor rendimiento predictivo sobre un conjunto de validación, utilizando métricas como el MAE, RMSE, R^2 y el porcentaje de aciertos. Tras analizar los resultados de los dos tipos de predicciones (con y sin las notas de los anteriores trimestres), he llegado a la conclusión de que Boosting ofrece, en términos de rendimiento predictivo puro, las mejores predicciones en ambos. Sin embargo, todas las predicciones no habrían sido posibles sin un preprocessamiento y limpieza de los datos, así como un análisis profundo de las variables y su importancia en la predicción final.

1 Introducción

En el contexto educativo, entender los factores que influyen en el rendimiento académico de los estudiantes es una prioridad tanto para los centros como los profesores. El análisis predictivo y exploratorio basado en datos permite identificar patrones y tomar decisiones informadas que mejoren los resultados de aprendizaje. Concretamente, predecir los resultados finales de los alumnos puede facilitar la detección temprana de estudiantes cuyo rendimiento académico puede ser inferior y tratar de mejorarlo.

Este trabajo se centra en la predicción de una variable objetivo ($T3$); es decir, la nota final del estudiante en el último trimestre, utilizando un conjunto de datos que contiene información detallada con características de cada alumno. A partir de este dataset, se plantea un enfoque que combina técnicas de aprendizaje supervisado (como regresión lineal, regularización y métodos de ensamblado) con métodos no supervisados (como clustering) para enriquecer el análisis y descubrir patrones ocultos.

Además del objetivo principal de la predicción de $T3$ para cada estudiante, se busca explorar también agrupaciones naturales entre los estudiantes que comparten características similares. Con este análisis conseguimos complementar la modelización supervisada gracias a conclusiones adicionales con las que resulta más sencillo entender los resultados y mejorar la toma de decisiones educativas basadas en evidencia.

2 Conjunto de Datos

El conjunto de datos utilizado proviene del archivo `rendimiento_estudiantes_train.csv`, que contiene información detallada sobre cada estudiante y su desempeño académico. Inicialmente, el dataset cuenta con 835 registros de estudiantes y 34 columnas, incluyendo la variable objetivo, cada una con una característica a tener en cuenta.

La variable objetivo que se busca predecir es de tipo numérica continua y representa el rendimiento académico del estudiante en el último trimestre con un valor entre 0 y 20. Entre las variables predictoras se incluyen características personales (como edad, género o situación familiar), académicas (asistencia a clase, horas de estudio, si va o no a academia), socioeconómicas (nivel educativo de los padres, acceso a internet etc.), entre otras. Existen variables tanto numéricas como categóricas.

Antes de comenzar se llevó a cabo una exploración inicial superficial en la que, de primeras, se identificaron valores faltantes (*NaN*) en varias columnas. Para tratarlos, se aplicaron estrategias de imputación como el uso de la moda en variables categóricas y la mediana en numéricas. También se detectaron variables categóricas con codificaciones heterogéneas, por lo que fue necesario estandarizar etiquetas para evitar duplicados (por ejemplo, “no” vs “No”). Asimismo, se podían ver con claridad outliers o valores sin sentido, para los cuales habría que buscar una solución. Por lo tanto, antes de comenzar las predicciones y el análisis, realicé un proceso de data cleaning con el objetivo de que el dataset no diera problemas y lograr mejores predicciones, con la amplia diversidad de características con las que se contaba.

	escuela	sexo	edad	entorno	TamFam	EstPadres	Medu	Pedu	Mtrab	P trab	...	TiempoLib	SalAm	AlcSem	AlcFin	salud	faltas	asignatura	T1	T2	T3
0	IC	M	19	U	>=4	J	2.0	1.0	casa	otros	...	4	3	1.0	3	5	210.910377	L	8	9	9
1	BG	F	18	U	>=4	J	4.0	4.0	sanidad	sanidad	...	4	4	1.0	1	4	15.000000	M	9	8	8
2	BG	F	16	R	>=4	J	4.0	4.0	sanidad	docencia	...	4	4	2.0	3	4	0.000000	L	17	16	16
3	BG	F	16	U	<4	J	4.0	3.0	docencia	servicios	...	4	3	1.0	2	1	2.000000	L	16	15	16
4	BG	M	18	U	<4	J	3.0	3.0	servicios	sanidad	...	2	4	2.0	4	4	13.000000	M	6	6	8
...
830	IC	F	19	U	>=4	J	2.0	NaN	casa	servicios	...	4	4	1.0	1	2	0.000000	L	9	9	10
831	IC	F	18	U	<4	J	1.0	1.0	casa	servicios	...	3	2	1.0	1	4	0.000000	L	19	17	18
832	IC	F	15	R	>=4	J	3.0	3.0	servicios	otros	...	5	4	NaN	1	1	4.000000	L	13	12	12
833	BG	F	17	U	>=4	J	NaN	NaN	servicios	otros	...	3	5	2.0	4	4	4.000000	L	12	16	16
834	BG	F	18	U	>=4	J	2.0	2.0	casa	servicios	...	3	3	1.0	1	3	0.000000	M	9	10	0

835 rows × 34 columns

Figure 1: Dataset Inicial.

3 Preprocesamiento

El preprocesamiento y data cleaning de los datos fue una etapa crítica para garantizar la calidad del análisis y la eficacia de los modelos predictivos. En él, se llevaron a cabo varias tareas fundamentales para trabajar con un buen dataset:

- **Eliminación de outliers:** A simple vista era muy fácil reconocer distintos outliers o valores sin sentido en determinadas variables por lo que, inicialmente opté por estudiar los que podía haber en cada variable. Así, observé que había gran cantidad de outliers en variables muy distintas pero todos ellos podían ser valores perfectamente reales salvo en las faltas de cada estudiante, en la que el boxplot ya resultaba, cuanto menos, extraño, por lo que este estudio se centró únicamente en dicha variable.

Tras un análisis profundo de la misma, fue sencillo percatarse de que algunos de los valores considerados outliers sí que podían ser correctos, pero otros resultaban excesivamente extremos por lo que decidí imponer un límite en 100. Así, se logró identificar 20 alumnos en los que, además de presentar valores extremos en las faltas, dicho valor era con decimales, lo cual era imposible. Por último, se comprobó que fueran esos los únicos valores con decimales de la variable y, como así era, decidí eliminar dichos estudiantes del dataset ya que, al ser este muy amplio, no iba a afectar significativamente.

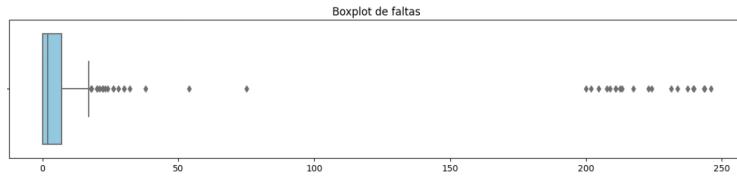


Figure 2: Boxplot Faltas.

- **Imputación de valores faltantes:** Se identificaron numerosos valores *NaN* en distintas variables, los cuales los imputé con la media de cada uno.

Medu	50
Pedu	100
TiempoEstudio	16
RefFam	16
AlcSem	19

Figure 3: Valores Faltantes.

- **Codificación de variables categóricas:** Para empezar, identifiqué dos tipos de variables categóricas, las cuales debía integrar de forma eficiente en los modelos lineales. Por un lado, contaba con variables categóricas binarias, (como sexo: M/F), las cuales se codificaron como 0 y 1. Por otra parte, también había variables categóricas nominales, las cuales fueron transformadas mediante *One-Hot Encoding*, tras observar los distintos tipos de valores que podían adquirir y evitar duplicados semánticos.

```
mapeos_binarios = {
    "sexo": {"M": 0, "F": 1},
    "escuela": {"IC": 0, "BG": 1},
    "entorno": {"U": 0, "R": 1},
    "TamFam": {"<4": 0, ">=4": 1},
    "EstPadres": {"J": 0, "S": 1},
    "apoyo": {"no": 0, "si": 1},
    "ApFam": {"no": 0, "si": 1},
    "academia": {"no": 0, "si": 1},
    "extras": {"no": 0, "si": 1},
    "enfermeria": {"no": 0, "si": 1}
    "EstSup": {"no": 0, "si": 1},
    "internet": {"no": 0, "si": 1},
    "pareja": {"no": 0, "si": 1},
    "asignatura": {"M": 0, "L": 1}
}
```

Figure 4: Variables Categóricas Binarias.

El resultado de este proceso fue un conjunto de datos limpio, numérico y adecuado para ser introducido en algoritmos con el fin de predecir la variable objetivo. Así, para comenzar a trabajar en los dos tipos de modelos (con y sin las variables T1 y T2) solo hacía falta eliminar de los datos de predicción la columna objetivo T3 y, si fuera necesario, también T1 y T2.

escuela	sesi	edad	entorno	fam_fam	EstPadre	Media	Pedi	TiempodeViaje	TiempodeEstudio	...	Mtrab_servicios	Prab_descencia	Prab_otros	Prab_viviad	Prab_servicos	ranci_optativas	ranci_otros	ranci_reputacion	notas_otros	notas_pade
1	1	1	16	0	1	0	4.8	3.0	3	...	0	0	1	0	0	0	0	1	0	0
2	1	1	16	0	1	0	4.8	3.0	3	...	0	0	1	0	0	0	0	1	0	0
3	1	1	16	0	0	0	4.8	3.0	3	...	0	0	0	0	1	1	0	0	0	0
4	1	0	18	0	0	0	3.0	3.0	1	...	1	0	0	1	0	0	0	0	0	1
5	1	0	18	0	1	0	3.0	2.0	2	...	0	0	0	1	0	0	0	0	1	0
...
830	0	1	19	0	1	0	2.0	2.0	1	...	0	0	0	0	1	1	0	0	1	0
831	0	1	18	0	0	0	1.0	1.0	2	...	0	0	0	0	1	1	0	0	0	1
832	0	1	19	0	1	0	2.0	2.0	1	...	0	0	0	0	1	0	0	0	0	0
833	0	1	17	0	1	0	3.0	2.0	1	...	1	0	1	0	0	0	0	1	0	0
834	1	1	18	0	1	0	2.0	2.0	1	...	0	0	0	0	0	0	0	0	0	0

Figure 5: Dataset Limpio.

4 Metodología y Modelos

El objetivo principal de esta sección fue estudiar distintos algoritmos predictivos y compararlos con el fin de conseguir el mejor para ser capaz de estimar con precisión el rendimiento final de los estudiantes en T3, de dos maneras distintas: usando o no las notas en los trimestres previos (T1 y T2).

4.1 Modelos Supervisados

Se entrenaron y evaluaron los siguientes modelos, cada uno con resultados diferentes y con importancias distintas para cada variable, las cuales se pueden observar en el código para cada algoritmo en cada modelo:

- **Regresión Lineal:** Sirvió como modelo base pero sus estadísticas dejaban que desear en ambos modelos al no capturar relaciones no lineales ni interacciones complejas entre variables
- **Regresión + Lasso:** Utiliza regularización L1 para realizar selección automática de variables, penalizando los coeficientes menos relevante, lo cual mejoró bastante el algoritmo previo.
- **Random Forest:** Algoritmo basado en ensambles de árboles de decisión. Es robusto ante variables no lineales y capture interacciones complejas entre variables, dando unos resultados mucho mejores que los previos.
- **Gradient Boosting:** Mostró gran capacidad predictiva, siendo uno de los modelos con mejor rendimiento, lo cual podremos ver en los resultados posteriores.

4.2 División de Datos y Evaluación

El dataset original se dividió en **train** y **validation**, empleando una partición del 80%–20% sobre la variable objetivo. Para la evaluación, se utilizaron las métricas estándar en regresión:

- **MAE (Mean Absolute Error):** Promedio del valor absoluto de los errores.
- **RMSE (Root Mean Squared Error):** Penaliza más los errores grandes.
- **R² (Coeficiente de Determinación):** Proporción de varianza explicada por el modelo.
- **Porcentaje de aciertos:** Porcentaje de predicciones correctas que ha realizado cada modelo sobre la variable objetivo

Así, todos los modelos fueron entrenados sobre los mismos datos preprocesados y comparados bajo las mismas condiciones para asegurar una evaluación justa.

5 Exploración Adicional

Con el objetivo de explorar patrones ocultos en los datos y agrupar a los estudiantes según características compartidas, se aplicó un análisis de clustering no supervisado utilizando el algoritmo K-Means. Para ello, se utilizó el dataset previamente limpiado, eliminando las variables directamente relacionadas con el rendimiento académico (notas T1, T2 y T3), con el fin de que el agrupamiento se basara únicamente en las características personales, familiares y escolares de los estudiantes.

Inicialmente, llevé a cabo dos técnicas de evaluación con el objetivo de conocer cuál era el número óptimo de clusters:

- **Método del codo:** Se graficó la inercia (suma de distancias cuadradas intra-cluster) para varios valores de k (de 2 a 10). El punto donde la reducción de inercia comienza a ser marginal indica el número ideal de clusters.
- **Silhouette score:** Se midió cómo de bien se separan los clusters generados. Un valor más alto indica mejor separación.

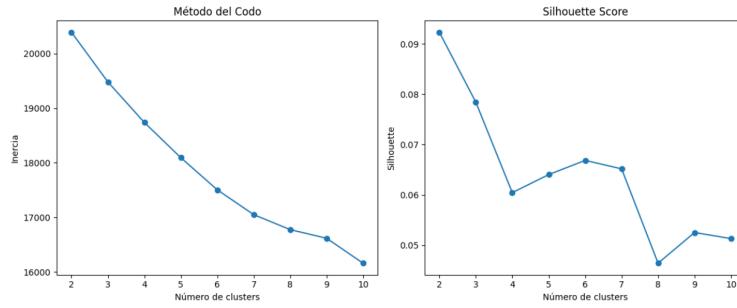


Figure 6: Resultados.

Tras los resultados de ambos métodos, decidí que agrupar los datos en tres clusters era la elección más razonable.

Para facilitar la interpretación visual del agrupamiento, se utilizó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y representar los estudiantes en un espacio bidimensional. El gráfico resultante mostró una separación razonablemente clara entre los tres grupos, lo que indica que el algoritmo captó patrones estructurados en los datos.

5.1 Clustering

Finalmente, se volvieron a incorporar las notas al dataset para analizar su distribución por grupo. Por un lado, se puede observar que el Cluster 0 contiene claramente a los estudiantes con mejor rendimiento académico promedio, asociados, con diferencia respecto a los demás, a factores como un mayor nivel educativo de los padres, más tiempo de estudio y mejor entorno familiar tal y como representan las estadísticas. El Cluster 1, por el contrario, agrupa a los estudiantes con peor rendimiento, quienes presentan menor apoyo familiar, menor tiempo de estudio y condiciones menos favorables. El Cluster 2 se caracteriza por un rendimiento más intermedio, con características variadas, lo cual se corresponde también con las estadísticas.

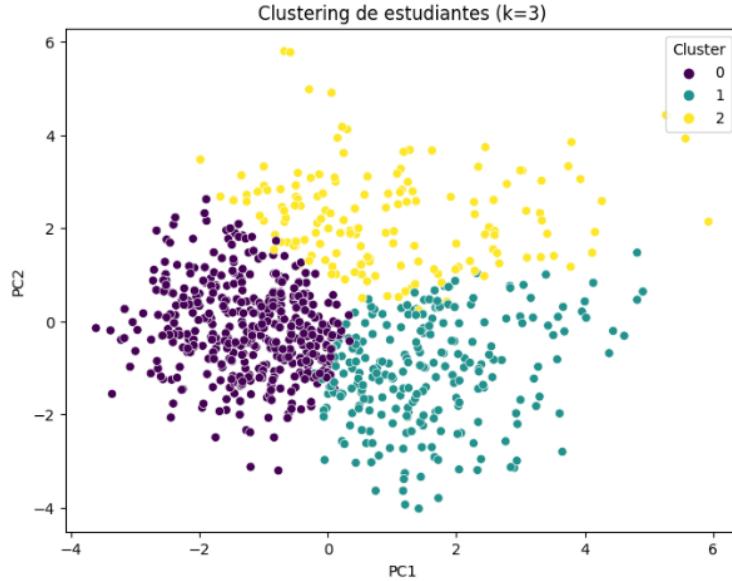
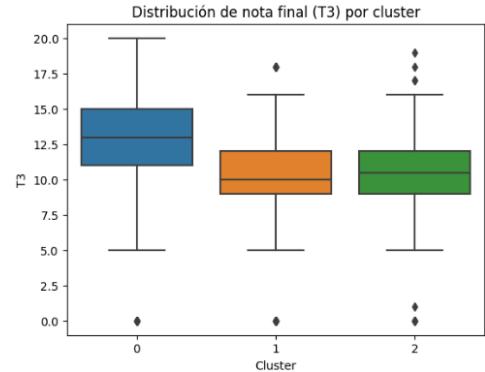


Figure 7: Clustering.

	escuela	sexo	edad	entorno	TamFam	EstPadres	\
Cluster	0.907363	0.648456	16.422880	0.130641	0.710214	0.114014	
0	0.400826	0.628099	17.140496	0.570248	0.714876	0.115702	
1	0.822368	0.236842	17.105263	0.283947	0.651316	0.118042	
2	3.092637	2.703088	1.315914	2.194774	...	0.064132	
0	1.756198	1.657025	1.851240	1.739669	...	0.004132	
1	2.895263	2.519737	1.572368	1.690789	...	0.026316	
3	0.261283	0.349169	0.083135	0.299287	...	0.004132	
0	0.260331	0.545455	0.132231	0.152893	...	0.004132	
1	0.348684	0.388158	0.118421	0.190789	...	0.004132	
4	tutor_others	tutor_padre	T1	T2	T3	...	
0	0.045131	0.249406	12.159145	12.332542	12.470309	...	
1	0.103306	0.190083	10.252066	10.041322	10.049587	...	
2	0.105263	0.223684	10.098684	10.171053	10.250579	...	

[3 rows x 43 columns]

(a) Estadísticas por cluster.



(b) Boxplot por cluster

Figure 8: Análisis Clustering

6 Resultados y Conclusiones

En esta sección se comparan los resultados obtenidos por los distintos algoritmos entrenados en los dos tipos de modelos. La evaluación se realizó, como se ha explicado antes, sobre el conjunto de validación y se presentan las métricas clave: MAE, RMSE, R² y el porcentaje de aciertos.

Table 1: Comparativa de rendimiento MODELO I.

Modelo	MAE	RMSE	R ²	% Aciertos
Regresión Lineal	0.93	1.40	0.86	38.7
Lasso Regression	0.83	1.37	0.87	44.2
Random Forest	0.78	1.14	0.91	41.7
Gradient Boosting	0.74	1.16	0.91	47.2

En ambos modelos el algoritmo **Gradient Boosting** fue el que alcanzó un mejor rendimiento puramente predictivo. En el Modelo I, al contar con las notas de T1 y T2, obviamente, los resultados

Table 2: Comparativa de rendimiento MODELO II.

Modelo	MAE	RMSE	R ²	% Aciertos
Regresión Lineal	2.47	3.21	0.27	16.0
Lasso Regression	2.41	3.21	0.27	18.4
Random Forest	2.38	3.17	0.29	14.1
Gradient Boosting	2.42	3.08	0.33	11.7

son bastante mejores que en el segundo en los cuatro algoritmos, pero Boosting destaca sobre todos ellos tanto en aciertos como en las estadísticas de rendimiento predictivo. Sin embargo, en el Modelo II, los 4 algoritmos caen mucho, debido a la importancia de las variables T1 y T2 para predecir el objetivo. Aunque el mayor porcentaje de aciertos lo consiguiera Regresión + Lasso y Boosting fuera el peor, he considerado valorar más las demás estadísticas que reflejan mejor la capacidad de predicción por lo que opté también por seleccionar Boosting como el mejor.