Proyecto final - Aprendizaje automático

Pablo Berástegui
Grado en Ingeniería Matemática e Inteligencia Artificial
Universidad Pontificia Comillas
202311460@alu.comillas.edu

Abstract

Este proyecto aborda la predicción del rendimiento académico de estudiantes en una asignatura concreta, modelando la nota final del tercer trimestre a partir de variables sociofamiliares, escolares y personales recogidas durante el curso. Se entrenan y evalúan dos tipos de modelos: uno que incluye las calificaciones parciales como predictores, y otro que prescinde de ellas, con el objetivo de estimar el rendimiento final desde etapas más tempranas. Para ello, se implementan modelos de regresión lineal regularizada (*Elastic Net*) y métodos de ensamblado basados en árboles (*Random Forest* y *Gradient Boosting*), ambos optimizados mediante validación cruzada. El análisis incluye una exploración detallada del dataset, reducción dimensional selectiva y evaluación de la importancia de las variables. Se comparan los errores de predicción y se discute el equilibrio entre precisión y explicabilidad, así como el impacto de distintas categorías de factores en los resultados académicos. Finalmente, se proponen técnicas no supervisadas que enriquecen el análisis desde una perspectiva alternativa.

1 Análisis exploratorio de los datos (EDA)

1.1 Descripción general del dataset

El dataset proporcionado tiene un total de 835 muestras, 33 variables (31 variables para el segundo modelo) y 186 muestras en las que al menos un dato es faltante (NaN), carencia que precisa de una imputación de datos correcta.

De los 33 predictores que se dispone, 15 son numéricos y 18 son categóricos.

Muchas clases están desbalanceadas, pero estos desequilibrios son naturales y no producto de las concreciones de nuestra muestra de la población. Por ejemplo, el número de personas que acceden al servicio de enfermería es mucho mayor que las que no; eso concuerda con lo que uno podría esperar de cualquier centro educativo. Por lo tanto, se decide no hacer ningún tipo de *upsampling* o *downsampling*.

1.2 Tratamiento de las variables

Algunos de los numéricos son discretos, como la edad (entre 15 y 22), el tiempo de viaje hasta la escuela (niveles de 1 a 4) o la calidad de las relaciones familiares (niveles de 1 a 5). Esto implica que muchas de las variables numéricas se podrían tratar (idealmente) como categóricas, para evitar dinámicas estrictamente lineales en algunos casos. Por ejemplo, sería coherente que una mala relación familiar (1) afectase negativamente en mayor proporción a la nota final del estudiante que una relación decente (3). Tratando dichas variables como categóricas, cada número tendría un regresor, gracias al tratamiento *one-hot-encoding* que nos permite trabajar con variables categóricas.

No obstante, tratar como categóricas a dichos predictores numéricos, además de los que ya son categóricos, implicaría hacer *one-hot-encoding* con 29 variables, es decir, con la gran mayoría de

ellas. Esto nos daría algo más de 70 variables binarias con las que computar nuestros modelos, lo cual resultará muy poco eficiente. Por ello, dejaremos las variables numéricas como tal.

1.3 Posibles errores en el dataset

El dataset muestra posibles errores de entrada en diversas variables. Veremos cuáles son simplemente outliers y cuáles son datos erróneos.

1.3.1 Predictor `razón`

Al ver los valores únicos que toma la variable razon, podemos observar que aparece la categoría "otros" y "otras". Ambas parecen dar la misma información en el contexto de dicho predictor, que explica el motivo por el cual el estudiante pertenece a dicho centro educativo, por lo que computaremos ambos como "otros" para evitar inconsistencias.

1.3.2 Outliers en `faltas`

En la columna de faltas hay 20 valores mayores que 200, y todos estos valores son los únicos de tipo float en la columna. Ambas cosas son un claro indicio de datos introducidos erróneamente, ya que no es lógico que un estudiante tenga un número decimal de faltas, así como un número mayor que 200.

A la hora de imputar estos valores, se plantean distintas opciones, como caparlos a un valor coherente, sustituirlos por la media, o corregirlos manualmente. Sin embargo, no se encuentra ningún posible valor lógico para dichas muestras, por lo que las consideramos como **observaciones inválidas** y por tanto las eliminamos.

1.4 Relaciones entre variables

Conocer las correlaciones entre variables nos ayuda a entender las dinámicas establecidas en el rendimiento escolar. No obstante, se debe tener precaución para **no atribuir causalidad a lo que tan solo es correlación**.

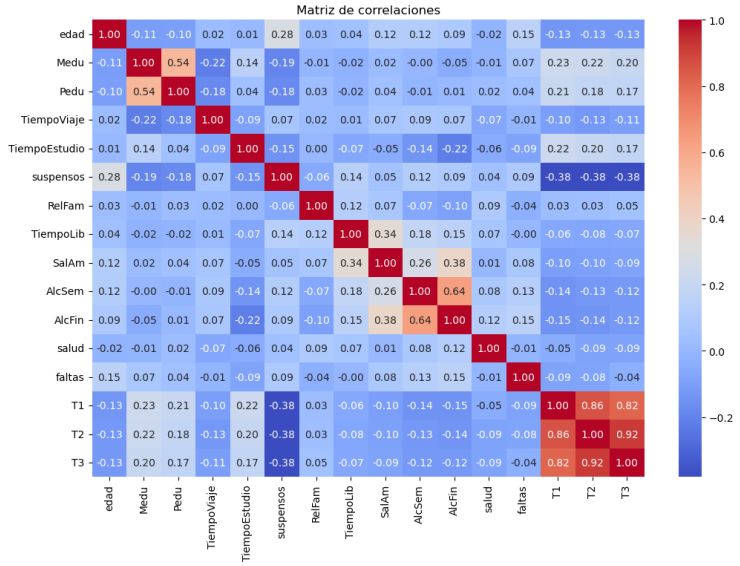

Figure 1: Matriz de correlaciones

Los tres bloques de correlación más acentuada son:

- `Medu`, `Pedu`: la educación de los padres suele ser similar, por lo que sería interesante estudiar la educación de ambos padres como una variable unificada, siendo esta una simple media

aritmética de la educación del padre y de la madre.

$$\texttt{PMEdu} = \frac{\texttt{Pedu} + \texttt{Medu}}{2}$$

- `AlcSem`, `AlcFin`: los estudiantes que consumen alcohol entre semana, con gran probabilidad lo harán en fines de semana. Ambos tienen una repercusión negativa sobre la nota final (correlación negativa). Resultará más útil unir ambas variables como el consumo de alcohol. Sin embargo, para tratar de dar más peso al consumo de alcohol entre semana, que puede afectar más gravemente en el rendimiento académico, calcularemos la nueva variable `Alc` como una ponderación entre las dos anteriores.

$$\texttt{Alc} = 0.7 * \texttt{AlcSem} + 0.3 * \texttt{AlcFin}$$

- `T1`, `T2`, `T3`: estas correlaciones tienen sentido, puesto que las notas del primer y segundo trimestre sirven como indicador de el nivel de conocimientos del alumno hasta el momento. Un estudiante con buena base de conceptos en los primeros trimestres obtendrá buenos resultados en la nota final con mayor facilidad que otro que no ha cultivado tales resultados antes.

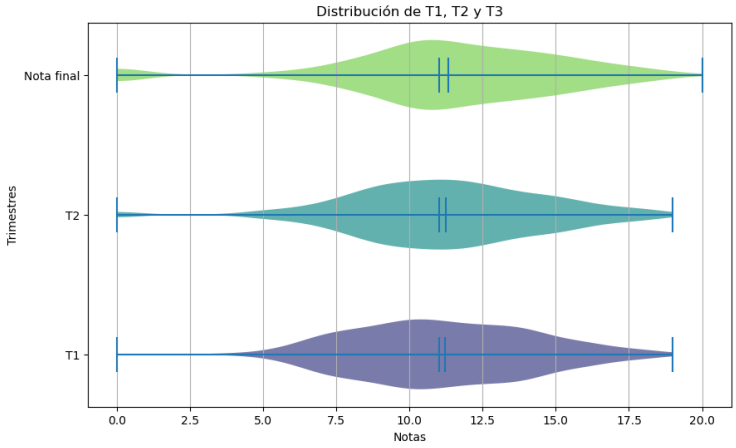

Figure 2: Distribución de T1, T2 y la nota final (T3)

Por este último motivo se estudian los modelos con y sin las variables `T1` y `T2`, puesto que, una vez conocidas éstas, es fácil predecir con alta precisión cuál será la nota final del alumno.

1.5 Preprocesamiento de las variables

1.5.1 Imputación de valores faltantes

Como mencionaba al comienzo del análisis, muchas de las observaciones cuentan con valores faltantes. Para entrenar modelos de Machine Learning es necesario lidiar con esos valores, ya que no se pueden computar. Para ello, la decisión por la que se opta es **imputar por moda** (variables categóricas) y **mediana** (variables numéricas). De esta manera nuestro dataset ya es enteramente numérico.

1.5.2 Estandarización de los datos

Para trabajar con modelos de regularización (además de otros como regresión logística, SVM, ...) es necesario tener un dataset estandarizado, con todas las variables con $\mu = 0$ y $\sigma = 1$, de tal modo que la relevancia de los coeficientes no dependa de la magnitud de las unidades de la variable.

$$z_i = \frac{x_i - \mu}{\sigma}$$

2 Implementación y comparación de los modelos

2.1 Un punto de partida

Para comenzar, es interesante observar el modelo más sencillo que puede haber: **la media**. Esta, ni más ni menos, nos dará un error y se procurará reducir implementando modelos más complejos.

Véase en la tabla 2 que la métrica R^2 para el predictor "media" es prácticamente cero. Esto se debe a que esta métrica nos indica "cuánto mejor predice nuestro modelo que la media". No obstante, puede no ser exactamente cero porque el conjunto de test no tiene exactamente la misma media que el conjunto completo o el conjunto de entrenamiento.

Una alternativa a este modelo tan sencillo es utilizar `T2` **como único predictor** para predecir `T3`. Esto tiene un coste ínfimo y tiene un rendimiento sorprendentemente alto. Cabe destacar que el gran rendimiento de este predictor tan "inocente" se debe a que hay muchas incidencias de estudiantes con $\texttt{T2} = 0$ y $\texttt{T3} = 0$. Encontrar una explicación clara a este fenómeno que nos permita tomar una decisión con los datos es complicado. Puede ser que el alumno se haya ido del centro escolar, o quizás ha sido expulsado. Sin embargo, la heterogeneidad del resto de variables de dichos estudiantes no permite sacar conclusiones precisas.

2.2 Conjunto de entrenamiento y validación

El planteamiento general para el entrenamiento de los distintos modelos será el siguiente: reservar un **20%** del dataset como test de cada modelo (*hold-out*), y con el **80%** restante entrenaremos cada modelo, buscando los hiperparámetros óptimos cuando sea necesario, empleando **validación cruzada** y técnicas de *grid-searching* y después se entrena de nuevo el 80% con los valores óptimos. En concreto, usamos k-fold CV con k=5 para todas las búsquedas de hiperparámetros.

2.3 Regresión lineal

Primero implementamos la versión más básica de este modelo, es decir, sin penalizar de modo alguno a las variables. Podemos anticipar los resultados, dado que en un dataset con tantas variables, es natural que *muchas de ellas no aporten valor ni explicabilidad*, por lo que la penalización en este sentido resultará beneficiosa.

Así, implementamos también las versiones regularizadas de la regresión: *Lasso*, *Ridge* y *ElasticNet*. Es interesante ver que las métricas de Lasso y ElasticNet son idénticas, y eso no es casualidad. Lo que ocurre es que, tras hacer una búsqueda exhaustiva con *grid-search* y validación cruzada, los parámetros óptimos hallados son un ratio de Lasso $r_{l1} = 1$ y una penalización $\alpha = 0.1$.

Lasso, y en este caso *ElasticNet* también, eliminan variables que no estima relevantes en la predicción. Por tanto, la obtención del parámetro $r_{l1} = 1$ es coherente con el análisis previo, puesto que observábamos que el gran número de variables iba a perjudicar el rendimiento de la regresión.

Veamos que los coeficientes que *Lasso* mantiene distintos a 0 son los siguientes:

Variable	Coeficiente
Suspensos	-0.056
Faltas	0.070
T1	0.385
T2	3.143
Madre trabaja (otros)	-0.004
Tiene pareja	-0.007
Asignatura M	-0.189

Table 1: Coeficientes del modelo para algunas variables seleccionadas

Los resultados concuerdan con la lógica que se podría esperar: variables como el número de suspensos, faltas o asignaturas más complicadas contribuyen negativamente a la nota, mientras que las notas del primer y segundo trimestre siguen una dinámica proporcional positiva con la nota final.

Otras son más difíciles de interpretar, como el trabajo de la madre o el hecho de que tenga o no pareja el estudiante. Sin embargo, estas mantienen coeficientes muy cercanos a cero, haciéndose prácticamente insignificantes en el cálculo de `T3`

2.4 Ensamblamiento de árboles

A *priori*, parece una idea muy intuitiva utilizar un árbol de decisión para determinar la nota del estudiante, por ejemplo: *"si tiene más de 2 suspensos en el curso anterior, va a tener una nota menor que 11, si tiene un consumo medio de alcohol de 5 tendrá una nota menor que 6, ..."*. Sin embargo, sabemos que el árbol de decisión por sí solo es un modelo débil, que tiende a sobreajustarse a los datos con gran facilidad. Como alternativa, podemos emplear múltiples árboles para ponderar o estimar en conjunto una mejor predicción.

2.4.1 *RandomForest*: óptimo para el Modelo i)

Tras una búsqueda exhaustiva de hiperparámetros óptimos, obtenemos las siguientes conclusiones:

- Los árboles simples (`max_depth` = 5) funcionan bien porque tenemos dos predictores (`T1` y `T2`) muy correlacionados a la variable objetivo y no hay necesidad de explorar estructuras complejas.
- Como `max_depth` = `None`, el modelo puede elegir a `T1` o `T2` en cada split, y dado su peso predictivo, es muy probable que lo haga.
- Hay una regularización ligera pero suficiente para evitar el potencial sobreajuste de los árboles.
- Con sólo 300 árboles, se obtuvo un rendimiento óptimo, lo que sugiere una convergencia rápida y eficiente

Nótese que los valores especificados pertenecen al ajuste del modelo i).

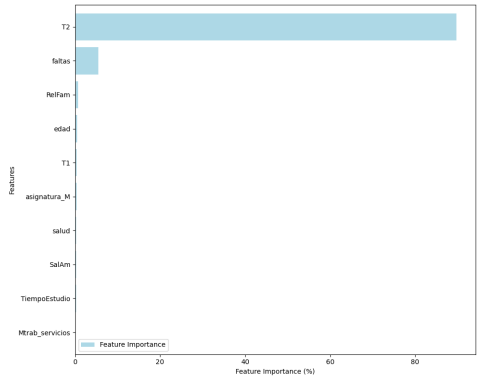

Figure 3: 10 variables con más importancia; modelo i) con RandomForest

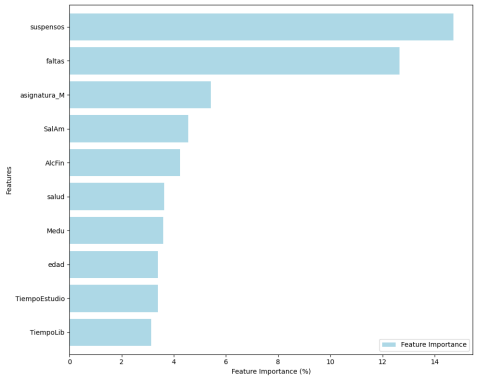

Figure 4: 10 variables con más importancia; modelo ii), RandomForest

2.4.2 XGBoosting: óptimo para el Modelo ii)

El mejor modelo excluyendo `T1` y `T2` se obtuvo mediante el adecuado ajuste de *XGBoost*. El modelo entrenó 762 árboles con profundidad 8, tasa de aprendizaje 0.05 y criterios de división exigentes ($\gamma = 0.49$, `min_child_weight` $= 7$), lo que sugiere una estructura compleja pero bien regulada. La aleatoriedad moderada (`colsample_bytree` $= 0.62$) y la regularización ℓ_2 ($\lambda = 0.38$) ayudaron a evitar el sobreajuste.

Para este modelo hemos hecho uso de validación cruzada con *Randomized-Search* para acortar los tiempos de ejecución y obtener resultados igualmente buenos.

Cabe mencionar que estos modelos son más lentos de entrenar, lo que nos puede hacer plantear si compensa el tiempo de entrenamiento y la pérdida de explicabilidad por unas ligeras mejoras del rendimiento. En general, dependerá de si el objetivo final del estudio es obtener resultados precisos o un análisis rico en información.

3 Los dos modelos. Resultados y observaciones

3.1 Modelo i): todas las variables disponibles

En este modelo anticipamos buenos rendimientos a partir de modelos sencillos, dado que tenemos variables muy explicativas (`T1`, `T2`). Veremos que modelos muy sofisticados, con largos tiempos de entrenamiento y menor explicatividad, lograrán una ligera mejora en nuestros resultados.

Modelo	**RMSE**	R^2	**MAE**
Media (μ)	4.12	-0.02	3.00
`T2`	1.63	0.83	0.85
Regresión lineal	1.60	0.85	0.99
Lasso (L1)	1.59	0.85	0.93
Ridge (L2)	1.61	0.84	0.99
ElasticNet	1.59	0.85	0.93
RandomForest	1.36	0.89	0.88
XGBoosting	1.39	0.88	0.93

Table 2: Rendimiento de modelos con `T1` y `T2`.

3.2 Modelo ii): exclusión de `T1` y `T2`

Lo interesante de este modelo es que se puede predecir (con un cierto error) la nota final del estudiante en el comienzo del curso. Esto ofrece al centro educativo la posibilidad de anticiparse a los malos resultados, proporcionando apoyo personalizado y más focalizado.

En este modelo vemos que no sólo `T1` y `T2` afectan directamente a la nota final (`T3`), sino que hay otras variables, como el número de suspensos, de faltas, la asignatura en cuestión o aspectos sociales del estudiante que afectan al rendimiento académico, como se puede ver en la figura 4

Modelo	**RMSE**	R^2	**MAE**
Media (μ)	4.12	-0.02	3.00
Regresión lineal	3.51	0.26	2.54
Lasso (L1)	3.44	0.29	2.46
Ridge (L2)	3.51	0.26	2.51
ElasticNet	3.44	0.29	2.46
RandomForest	3.34	0.33	2.43
XGBoosting	2.96	0.48	2.25

Table 3: Rendimiento de modelos sin `T1` y `T2`.

4 Técnicas adicionales: análisis no supervisado

Para comprender de forma visual y cuantitativa la estructura global del conjunto de datos con sus 33 variables, incorporamos un **Análisis de Componentes Principales (PCA)**. Esta técnica nos permite proyectar la información multivariante en un espacio de menor dimensión que retiene la mayor parte de la varianza original, facilitando:

- La visualización intuitiva de patrones y agrupamientos entre estudiantes,
- La detección de observaciones atípicas (outliers) que podrían influir de forma desproporcionada en los modelos supervisados,

Aunque inicialmente aplicamos *clustering* sobre las primeras componentes principales, con $k = 2$ los grupos resultantes coincidieron prácticamente con la división en $PC1 < 0$ / $PC1 > 0$, es decir, replicaron la separación "bajo vs. alto rendimiento" ya capturada por el análisis de componentes principales. Dado que el *clustering* no aportó perfiles adicionales significativos, nos limitamos a explorar lo que nos puede ofrecer PCA.

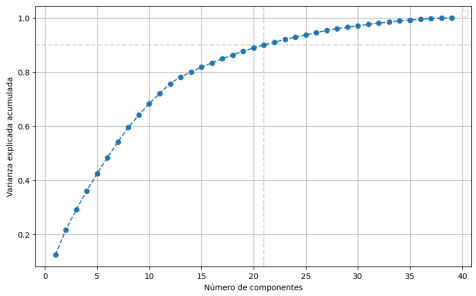

Figure 5: Varianza acumulada por los componentes

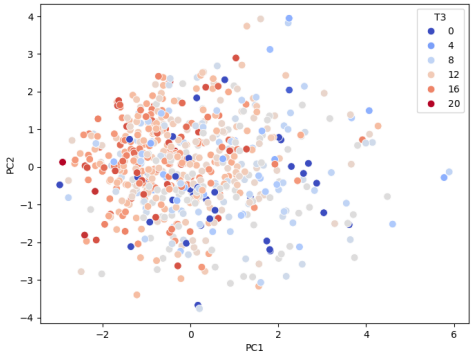

Figure 6: Relación de T3 con PC1 y PC2

Variable	**Peso**
Consumo de alcohol (Alc)	0.427
Suspensos	0.380
Salidas con amigos (SalAm)	0.370
Tiempo libre	0.345
Tiempo de estudio	−0.324
Edad	0.313
Nivel educativo parental	−0.265
Tiempo de viaje	0.246
Faltas	0.173
Sexo masculino	0.092

Table 4: Contribución de variables a PC1

Nótese (en la tabla 4)que en esta tabla, las variables con pesos positivos parecen ser aquellas que contribuyen negativamente a T3, mientras que las que tienen pesos negativos contribuyen positivamente. Esta relación inversa (y ciertamente contraintuitiva) cobra sentido viendo que PC1 sigue una dinámica inversamente proporcional a T3 (véase la figura 6), por lo que los pesos dentro de la primera componente principal también seguirán esta dinámica negativa. En otras palabras, PC1 está inversamente correlacionada con T3; por eso los loadings positivos indican menor rendimiento.

Como experimento adicional, entrenamos un modelo de regresión lineal sobre las 21 primeras componentes principales, que explican aproximadamente el 90% de la varianza total del dataset. El objetivo era reducir la dimensionalidad del problema y evaluar si las direcciones de mayor varianza contenían suficiente información para predecir `T3`. Aunque las PCs no tienen interpretación directa, el rendimiento obtenido ($RMSE = 3.41$, $R^2 = 0.30$) muestra que una parte sustancial de la variabilidad del rendimiento académico puede explicarse con un modelo más sencillo, obteniendo mejores resultados que muchos otros modelos más complejos y costosos.

5 Conclusiones

En conjunto, los resultados muestran que las calificaciones parciales (`T1` y `T2`) son el **mejor indicio temprano del rendimiento final**, pero un modelo construido únicamente con información sociodemográfica, familiar y de hábitos mantiene todavía un *RMSE* de 2.96 ($R^2 = 0.48$), que significa que la nota se predice con un error aproximado de 1.5 puntos sobre 10 (3 sobre 20). Esto permite discriminar con **notable precisión** a los alumnos que terminarán por debajo de la media.

En general, nunca se han visto reflejados con gran relevancia factores discriminantes de la persona como el género o el entorno (rural o urbano). Esto garantiza una evaluación justa e igualitaria.

Otro factor importante es la asignatura. Las notas de Matemáticas tienen una media más baja que las de Lengua. Esto se convierte en un factor determinante para la nota final también. Nos lleva a pensar si sería necesario ajustar la dificultad de las asignaturas, ofrecer apoyo concreto y personalizado para Matemáticas, y otras medidas, con el fin de que los resultados en ambas asignaturas no sean tan dispares.

En el segundo modelo (excluyendo `T1` y `T2`) encontramos la tendencia de *conocer mejor lo que perjudica al estudiante que lo que le beneficia*. Factores como el consumo de alcohol, el número de suspensos previos y el tiempo libre resultan determinantes en las calificaciones del alumno. A partir de estos hallazgos, sugerimos a los centros:

1. Vigilar a comienzos de curso a los estudiantes con más de un suspenso y absentismo elevado.
2. Reforzar programas de concienciación sobre el consumo de alcohol y su consecuente deterioro cognitivo.
3. Facilitar tutorías individualizadas cuando se detecta bajo nivel educativo familiar o pocos conocimientos cimentales de cursos anteriores.

El análisis no supervisado nos confirma una vez más que la tendencia de la nota final se ve fuertemente afectada por las mismas variables mencionadas anteriormente.

Aún así, el estudio se basa en una cohorte limitada (836 estudiantes, dos institutos, curso 2005) y no permite inferir causalidad. Futuros trabajos deberían replicar el modelo en centros adicionales, conocer el centro escolar más de cerca para conocer incidencias, costumbres y metodologías implantadas; incorporar variables psicosociales actuales y analizar la equidad del sistema de predicción.