
Análisis Predictivo del Rendimiento Académico de Estudiantes en Madrid - 2005

Jorge Carnicero Príncipe
ICAI, Universidad Pontificia Comillas
Madrid, España
jcarnicerop@alu.comillas.edu

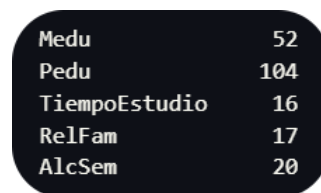
Abstract

Este informe muestra la implementación de varios modelos para predecir la nota final de estudiantes, empleando información demográfica, familiar y escolar. Además, se analizarán qué factores influyen realmente en ese rendimiento. Comparando los distintos modelos y que podemos hacer si queremos obtener mejores calificaciones o al menos esa es la finalidad.

1 Exploración inicial del conjunto de datos

Antes de comenzar con la fase de modelado y predicción hemos realizado una breve fase de exploración inicial, para conocer un poco más nuestros datos y corregirlos en caso de no encontrar algo en regla. Además de analizar comportamientos de ciertas variables con vistas al futuro y si pueden ser problemáticas.

He comenzado, analizando si encontramos algunos valores nulos en nuestro conjuntos de datos, ya que estos nos pueden causar problemas, y nos encontramos con que varias columnas los contienen.



Medu	52
Pedu	104
TiempoEstudio	16
RelFam	17
AlcSem	20

Figure 1: Columnas con valores nulos

Como vemos, tenemos 5 columnas con valores nulos, por tanto lo que he hecho es, dependiendo de si es una columna numérica o categórica, si fuera categórica los imputaría por la moda, y en caso de ser numérica lo imputaré por la mediana.

Después, nos damos cuenta de un error en la introducción de los datos, al encontrar en la columna de 'razon', algunos datos que son otros, cuando en el pdf del proyecto se nos indicaba que solo podían ser: cercanía, reputación, optativas u otras.

Por tanto, cambiaremos los valores en donde antes ponía otros por otras, para mantener un correcto formato.

optativas	348	→	optativas	348
cercania	204		cercania	204
reputacion	198		reputacion	198
otras	56		otras	85
otros	29			

Figure 2: Antes y después valores de la columna 'razon'

Luego, he decidido fijarme en ver si encontraba outliers en alguna de las columnas, y vemos lo siguiente:

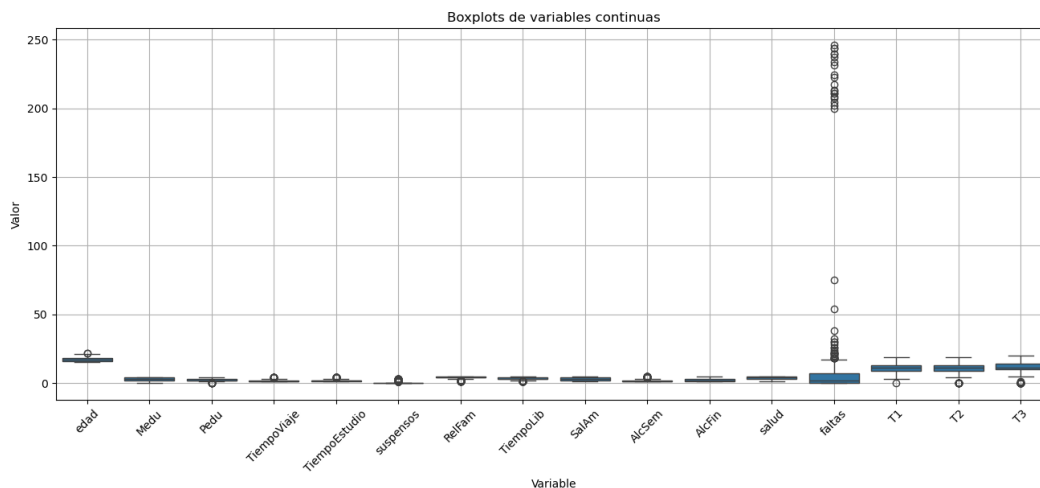


Figure 3: Boxplot columnas numéricas

Vemos como la más característica es la de faltas, viendo algunos valores que técnicamente son imposibles ya que no hay tantos días lectivos, rondando los 175 - 180 días al año. Y en el resto de columnas vemos algunos pero que son cosas que siempre pueden pasar.

Por tanto, lo que hemos hecho ha sido computar los valores nulos de la columna de faltas que sean mayores que 100 faltas, ya que menos puede ser posible (siempre hay algún alumno rebelde escondido).

Una vez visto estas cosas que pueden afectar a la hora de computar, me he querido enfocar y ver si encontramos desnivel en nuestros datos y como pueden llegar a afectar estos desniveles en nuestros datos.

Concretamente, me he enfocado en las columnas de Sexo, Asignatura y Escuela, entonces he querido ver que variedad de datos tenemos.

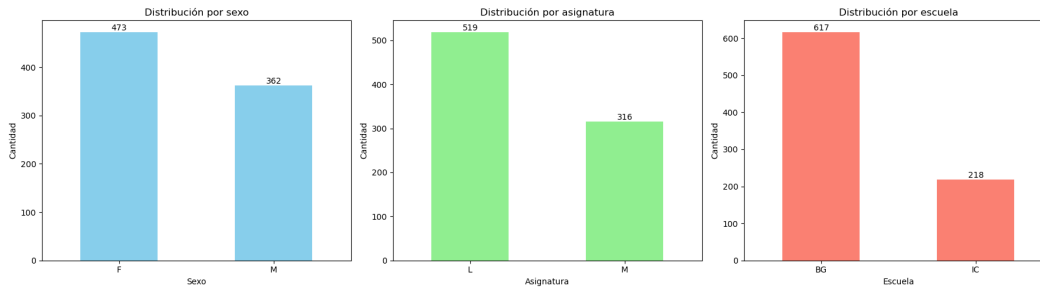


Figure 4: Boxplot columnas numéricas

Como vemos, encontramos gran disparidad de valores, entonces me ha dado por pensar si esto puede afectarnos a la hora de hacer predicciones, si afectará que sea mujer o hombre, que la asignatura sea matemáticas o lengua, o si va a un colegio o a otro.

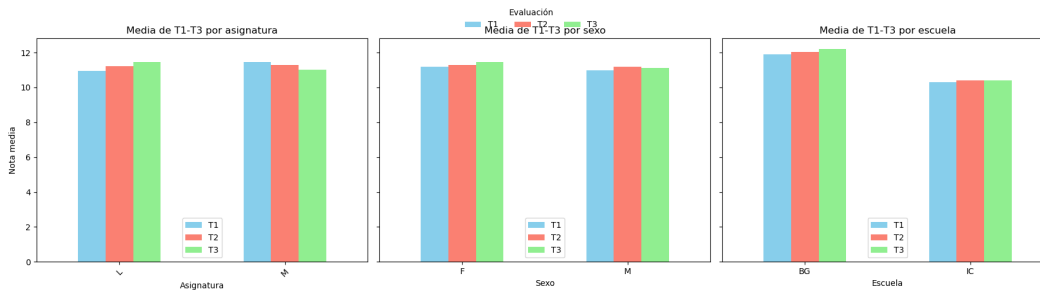


Figure 5: Media Notas Trimestres en función de valores de sexo, asignatura y colegio

Observamos como las medias de notas según sexo o asignatura apenas varían, mientras que entre ambos colegios existe una diferencia notable. Además, hice un análisis ANOVA y confirmó que ni el sexo ni la materia tienen un efecto significativo sobre las calificaciones, pero sí el centro educativo.

2 Modelos predictivos

2.1 Modelo I

Para este primer modelo he decidido usar dos modelos de predicción distintos:

- **Regresión lineal:** modelo creado desde cero en una de las prácticas de clase.
- **Random Forest:** implementado con `scikit-learn` siguiendo lo aprendido en clase y en la práctica.



Figure 6: Resultados Modelo I

Análisis de los resultados

Como podemos observar, ambos modelos ofrecen unos buenos resultados: con un R-squared por encima de $R^2 = 0.80$, y con un Mean Absolute Error (MAE) por debajo de 1, además de que el Root Mean Squared Error (RMSE) no supera 1.6. Estos resultados se explican, en parte, porque en el entrenamiento incluimos las dos variables más correlacionadas con la nota final T_3 , las notas de los trimestres anteriores T_1 y T_2 , lo que refuerza el modelo al aprovechar la continuidad habitual entre evaluaciones consecutivas. Pero finalmente, vemos como *Regresión Lineal* es ligeramente mejor que el *Random Forest*.

2.2 Modelo II

Para este segundo modelo he decidido volver a usar los dos modelos a ver que obtenemos:

- **Regresión lineal y Random Forest**

REGRESIÓN LINEAL	RANDOM FOREST
R2 : 0.2637037578781226	R2 : 0.4210932566135709
RMSE : 3.4525353003264136	RMSE : 3.0613722413323083
MAE : 2.416	MAE : 2.164

Figure 7: Resultados Modelo II

Análisis de los resultados

En este caso, el modelo de *Random Forest* ofrece predicciones notablemente mejores que la *Regresión Lineal*: el R-squared R^2 es 1.6 veces mayor, mientras que el error medio absoluto (MAE) se reduce en 0.3 puntos y la raíz del error cuadrático medio (RMSE) en 0.4. Estos valores más pequeños de MAE y RMSE también nos indican un ajuste más preciso y, por tanto, un rendimiento mejor del Random Forest.

2.3 Comparación de modelos

Comparando ambos modelos, nos encontramos como, el **Modelo I** destaca claramente por su mejor predicción, algo esperable dado que incluye las notas de los trimestres anteriores (T_1 y T_2), que están fuertemente correlacionadas con la nota final (T_3). Aportando mucha información sobre el rendimiento del estudiante, facilitando mucho más la predicción.

En cambio, en el **Modelo II** trabajamos sin estas notas de los cuatrimestres anteriores, por lo que hay que apoyarse únicamente en aspectos sociodemográficos, familiares y de comportamiento. Viendo cómo esto hace que sea menos preciso, pero también es una herramienta útil para identificar posibles dificultades desde el principio del curso, cuando aún no hay notas disponibles.

Un detalle interesante es que en el Modelo I, la *Regresión Lineal* funciona algo mejor que *Random Forest*, mientras que en el Modelo II sucede justo lo contrario. Esto se debe a que cuando las variables más importantes son numéricas y tienen una relación directa con el objetivo, como las notas previas, un modelo lineal puede capturar bastante bien el patrón. Sin embargo, cuando esa información desaparece y las variables son más variadas y menos lineales, Random Forest se adapta mejor al contexto al poder modelar interacciones complejas.

3 Interpretación de resultados

Los resultados obtenidos muestran que el rendimiento académico de los estudiantes podemos predecirlo con bastante precisión cuando tenemos información sobre los trimestres anteriores. Sin embargo, cuando no contamos con esta información, el modelo sigue siendo capaz de capturar ciertos patrones

relevantes a partir de variables contextuales y personales, aunque no con la misma exactitud, sino un poco menor.

Esto nos sugiere que los datos académicos previos determinan en gran medida las calificaciones siguientes, pero también que otros factores como el entorno familiar, el tiempo de estudio o otro tipo de actividades, tienen un papel importante en el rendimiento final.

Además, las diferencias de rendimiento que hemos observado entre regresión lineal y Random Forest confirman que la elección del modelo también depende del tipo de datos. Ya que, cuando disponemos de las calificaciones previas, un modelo lineal captura bien la tendencia; si solo contamos con variables contextuales y de hábitos, es preferible usar Random Forest para aprovechar sus interacciones complejas.

En la siguiente sección analizaremos que variables son influyentes en los modelos, buscando desde el punto de vista del Director cuales podrían solucionarse, y más adelante veremos otras variables que afectan en la nota final.

4 Factores clave que influyen en el rendimiento académico

Para ver los factores más clave, vamos a dejar de lado las notas de los trimestres anteriores, ya que estas obviamente determinan mucho. Y vamos a ponernos desde el punto de vista del Director, y prestar atención a aquellas en las que tal vez se pueda interferir.

1. **Apoyo al estudiante y suspensos.** Una de las variables que más ha llamado mi atención ha sido la del apoyo escolar, ya que recibía bastante peso tanto en el Modelo I como en el Modelo II. Por eso decidí analizar con más detalle cómo variaban las notas entre estudiantes que recibían apoyo y los que no.

En principio parecía que no había mucha diferencia, pero profundizando en el análisis para estudiantes que tienen entre 1 y 3 asignaturas suspensas, observamos resultados interesantes:

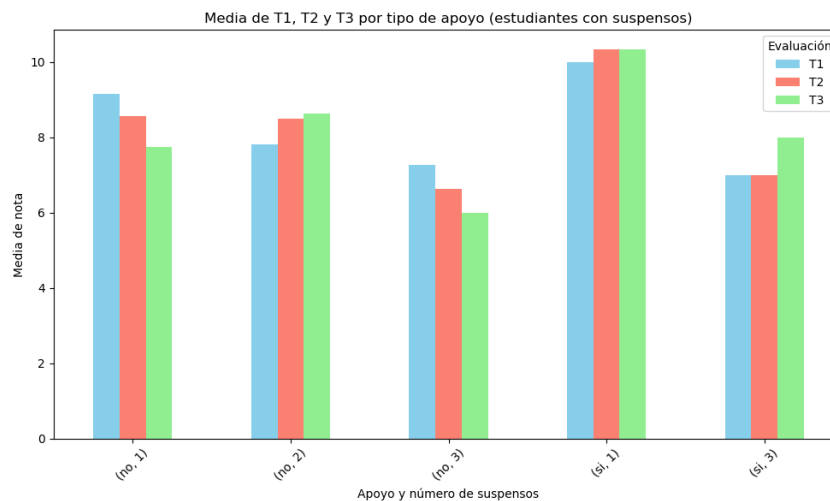


Figure 8: Relación entre apoyo escolar, suspensos y notas

Si nos centramos en aquellos estudiantes que tienen de 1 a 3 suspensos, las notas finales de quienes reciben apoyo son claramente superiores. Por tanto, sería recomendable ofrecer apoyo escolar especialmente a los estudiantes que presentan algunas asignaturas suspensas.

2. Otra cosa llamativa es si el estudiante tiene **internet en casa**. La media cambia casi 1.5 puntos entre quienes sí tienen y quienes no.

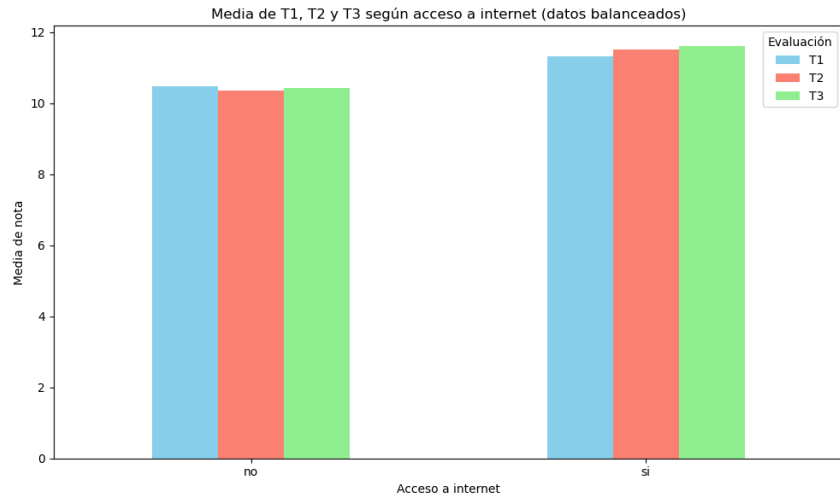


Figure 9: Relación entre acceso a internet y notas

Quizá se podría plantear un método de ayuda para aquellos que no tienen internet, y así poder instalar internet en casa y adaptarse.

3. Por último, es una variable que puede pasar desapercibida es el tema del **tiempo de viaje**.

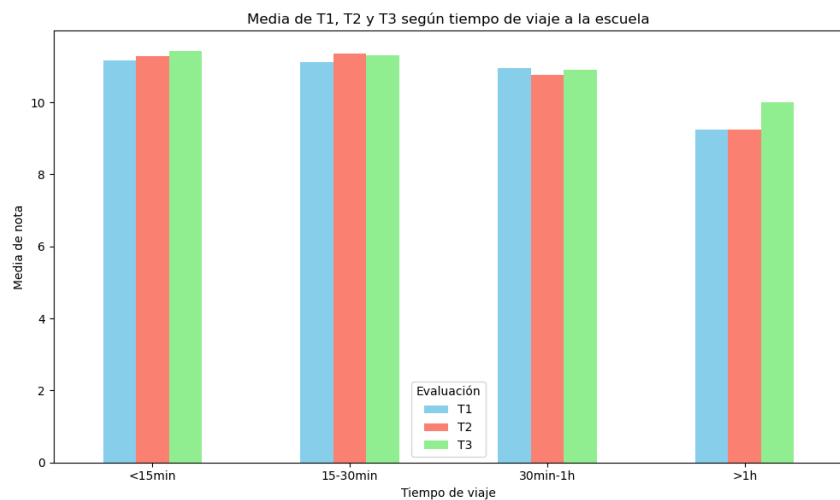


Figure 10: Relación entre tiempo de viaje y notas

Vemos como, cuanto más tarda el alumno en llegar a clase (independientemente el medio de transporte, el estudio es de 2005), peores son sus notas. Una idea podría ser montar rutas o autobuses para aquellas personas a las que les pueda beneficiar este transporte, y así acortarles el tiempo de viaje.

Estas son las variables que más me han llamado la atención y sobre las que, como director, puedo intervenir. Más adelante profundizaré con una exploración adicional de otros factores que también podrían influir en el rendimiento. El objetivo es detectar aspectos prácticos que podamos mejorar.

5 Exploración adicional y parte creativa

Para esta parte voy a hacer un poco más de exploración sobre el conjunto de datos para poder entenderlo mejor y que así me sirva para entender factores que afectan en las notas.

Me centraré en las variables de vida social y hábitos (relación familiar, tiempo libre, alcohol, etc.) para ver cuáles podría modificar o mejorar. Aunque los datos son de 2005 y las circunstancias puedan haber cambiado, sirven como referencia para entender mejor los patrones que influyen en las notas.

Una de las preguntas que más nos puede surgir es si el consumo de alcohol influye en el rendimiento académico. Disponíamos de datos tanto de fin de semana como de entre semana, pero me interesa más este último, ya que coincide con días de entre semana en donde se supone que debemos de estar concentrados en estudiar. Los resultados no sorprenden:

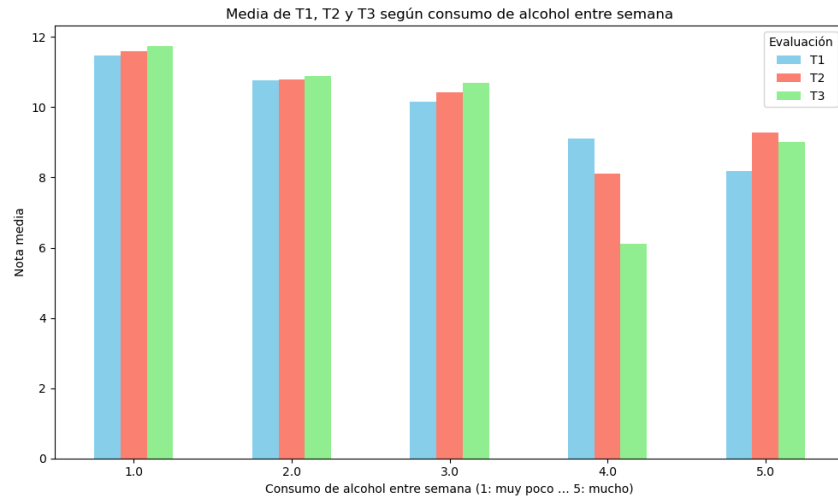


Figure 11: Relación entre consumo de alcohol entre semana (AlcSem) y notas

Como se aprecia, quienes consumen más alcohol entre semana obtienen peores notas. Por tanto, es algo que hay que tener en cuenta a la hora de plantearnos nuestro horario de estudio esta semana, ya que controlando esto, si antes lo hacíamos, y lo reducimos, podrían mejorar nuestras notas.

Por último, me ha parecido muy interesante explorar cómo se combinan el tiempo de estudio y las salidas con amigos, ya que podríamos pensar que estudiar mucho y salir poco es lo mejor para obtener buenas calificaciones. Por eso me ha gustado ver cómo se distribuye y qué patrones encontramos:

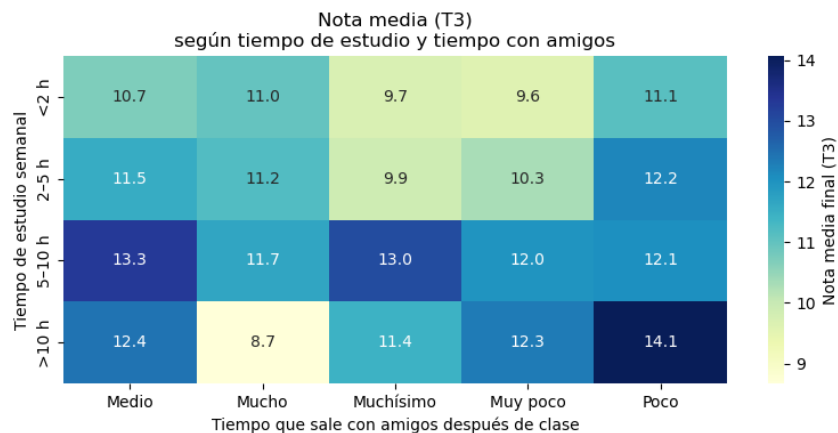


Figure 12: Heat-Map T3, Tiempo con amigos y Tiempo de estudio

Como podemos ver en el heat-map, en general quienes dedican el mínimo de horas al estudio obtienen notas más bajas, independientemente de su tiempo con amigos. Sin embargo, la mejor nota no se

corresponde con quienes estudian mucho y salen muy poco (lo que podríamos esperar), sino cuando se combina un alto tiempo de estudio con niveles bajo, medio o incluso alto de salidas con amigos.

Con esto podemos ver que no es necesario estar aislados todo el rato, sino que podemos tomarnos nuestros momentos de descanso y estar un rato con amigos y despejarnos, pero siempre y cuando tampoco nos olvidemos del estudio, viendo que esto puede llegar a ayudarnos a mejorar académicamente.

6 Conclusiones

En este trabajo hemos comparado dos enfoques de modelado supervisado para predecir la nota final (T_3) de estudiantes: el **Modelo I**, que incluye como variables predictoras las notas de los dos primeros trimestres (T_1 y T_2), y el **Modelo II**, que prescindir de esas notas y se basa únicamente en información sociodemográfica, familiar y de hábitos.

- *Modelo I*: obtuvo un R^2 superior a 0.85 y errores ($RMSE \approx 1.3$, $MAE \approx 0.9$), gracias a que contábamos con las notas anteriores. En este caso, la regresión lineal rindió ligeramente mejor que el Random Forest, pues la relación entre las variables y la nota final es prácticamente lineal.
- *Modelo II*: al eliminar T_1 y T_2 , el rendimiento global baja ($R^2 \approx 0.4$), pero aquí el Random Forest supera a la regresión lineal, al capturar interacciones y no linealidades entre las variables contextuales.

Respecto a los factores que más influyen cuando no disponemos de notas previas, destacan:

1. **Apoyo escolar y número de suspensos**: los alumnos con uno o varios suspensos mejoran claramente sus T_3 si estos reciben apoyo.
2. **Acceso a internet en casa**: quienes disponen de conexión en el casa obtienen, en promedio, 1.5 puntos más de que quienes no la tienen.
3. **Entorno familiar y hábitos de ocio**: un buen clima familiar y un uso moderado de salidas con amigos pueden llevar a mejores notas.
4. **Consumo de alcohol entre semana**: el consumo de alcohol en días lectivos influye en la nota media negativamente.
5. **Tiempo de viaje**: los estudiantes con desplazamientos excesivamente largos muestran de media, un rendimiento menor.

Por último, hay que recordar que estos resultados provienen de datos de 2005 y que correlación no implica causalidad. Sin embargo, permiten identificar variables de intervención (apoyo, recursos tecnológicos, organización del tiempo) que pueden ayudarnos a orientar nuestras acciones y mejorar nuestro rendimiento.