Proyecto Final de Machine Learning

Andrés Gil Vicente
202304626@alu.comillas.edu
Universidad Pontifica Comillas - ICAI
Grado en Ingeniería Matemática e Inteligencia Artificial
04/05/2025

Abstract

El objetivo principal de este proyecto es aplicar técnicas de aprendizaje automático supervisado para analizar y modelar datos reales, con el fin de identificar patrones relevantes y construir predicciones útiles en contextos educativos. En particular, se busca comprender qué factores influyen en el rendimiento académico de estudiantes de secundaria en Madrid, lo cual tiene un claro valor social y educativo.

1 Exploración y análisis descriptivo de los datos

El proceso de exploración y análisis de datos se ha desarrollado en el fichero `exploracion.ipynb`, con el objetivo de comprender la estructura interna del conjunto de datos y orientar adecuadamente el posterior proceso de modelado. El análisis exploratorio ha incluido un conjunto de herramientas estadísticas y gráficas que nos han permitido extraer conocimiento útil, detectar patrones de interés y observar relaciones relevantes entre variables de nuestro dataset. Destacan entre otras características:

- Variables directamente correlacionadas con variable objetivo: `T1`, `T2` y `faltas`.
- Variables inversamente correlacionadas con variable objetivo: `suspensos` y `TiempoViaje`.
- Numerosos valores atípicos en la variable `faltas`.
- Variables categóricas con errores de transcripción (ej. `razon`: “otras” en lugar de “otros”).
- Valores nulos o faltantes en variables: `Medu`, `Pedu`, `RelFam`, `TiempoEstudio` y `AlcSem`.

La información extraída de la exploración inicial ha sido fundamental para facilitar una correcta limpieza y comprensión del conjunto de datos, permitiendo sentar las bases para la posterior implementación de modelos más robustos y explicables.

Posteriormente, en el fichero `preprocesado.ipynb` se ha desarrollado el proceso de limpieza y procesado de los datos, preparándolos para poder utilizarlos en el entrenamiento de nuestros modelos predictivos. Algunas de las acciones más destacadas han sido:

- Corrección errores de transcripción (ej. “otras” y “otros” en la variable `razon`).
- **Imputación** de valores nulos (mediana para variables numéricas, moda para categóricas).
- Conversión de variables a un formato coherente (ej. `faltas`: a datos de tipo int).
- Codificación de variables categóricas empleando **dummy encoding**.
- Gestión de **outliers** y valores físicamente imposibles (ej. `faltas` con valores > 150).
- División del dataset (30% validación, 70% entrenamiento).
- **Estandarización** de las variables numéricas (usando los datos de train).

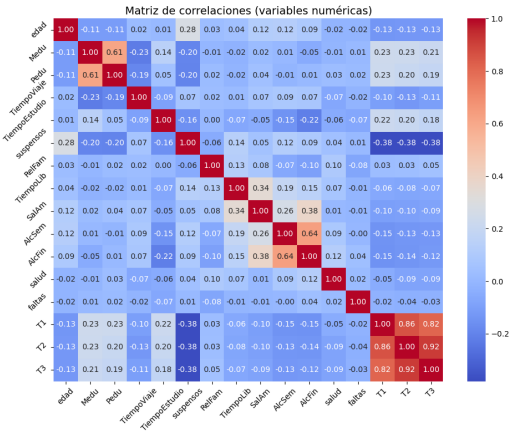

Figure 1: Matriz de correlaciones

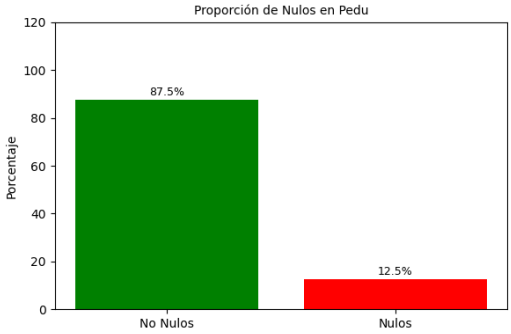

Figure 2: Valores nulos `Pedu`

2 Implementación y comparación de modelos predictivos

El siguiente paso del proyecto ha consistido en desarrollar modelos de aprendizaje automático que permitan abordar el problema planteado, de forma cuantitativa. En esta sección se documenta el proceso llevado a cabo en los archivos `modelo1.ipynb` y `modelo2.ipynb` desde la experimentación con distintos modelos, la evaluación y análisis crítico de su rendimiento, hasta la conclusión y elección del candidato óptimo. El objetivo es encontrar un modelo robusto que generalice bien sobre datos desconocidos, pero que a la vez no sea excesivamente complejo y permita una cierta interpretabilidad.

2.1 Modelo 1

En este caso, nuestro enfoque consiste en tratar de predecir la variable objetivo `T3` (nota final de los alumnos), utilizando toda la información disponible en el dataset de train. Hemos experimentado con diversos modelos supervisados, analizando las ventajas y limitaciones de cada uno. Todo este proceso ha sido desarrollado en `modelo1.ipynb`.

2.1.1 Regresión Logística

El modelo de regresión logística ha sido nuestro primer approach. Este modelo suele ser eficiente para problemas lineales de clasificación binaria. En nuestro caso, presentó un rendimiento aceptable (R2 = 0.69), pero se observaron ciertas limitaciones en cuanto a su capacidad para capturar relaciones no lineales en los datos. Es por ello que decidimos seguir explorando opciones.

2.1.2 KNN Regressor

A continuación, probamos a implementar un modelo de KNN pero aplicado a nuestro problema de regresión. Decidimos utilizar la media como estadístico para medir la contribución de cada vecino cercano al dato que queremos predecir. Tras emplear validación cruzada y elegir el valor óptimo para el hiperparámetro k (número de vecinos cercanos), evaluamos el modelo y obtuvimos un MSE ligeramente alto y un R2 inferior al de nuestro modelo de regresión logística (R2 = 0.66). De este modo, decidimos descartar este segundo approach.

2.1.3 Regresión Lineal

La implementación del modelo de regresión lineal tiene numerosas ventajas como que no hace falta entrenarlo, se conocen sus parámetros óptimos, dados por una fórmula cerrada, y que además es muy sencillo e interpretable. En nuestro caso nos proporcionó un rendimiento bastante bueno, con un MAE de 1.014 y un R2 de 0.83, lo cual lo convirtió en un buen candidato para nuestro modelo 1.

2.1.4 Decision Tree Regressor

Para este cuarto approach, optamos por implementar un modelo de árboles de decisión aplicados a regresión. Lo primero que hicimos fue utilizar validación cruzada para obtener el valor óptimo del hiperparámetro de la profundidad máxima del árbol, y a continuación, evaluamos la performance del modelo. Obtuvimos un MAE de 0.986, y un R2 de 0.834, mejorando el rendimiento de la regresión lineal, aunque añadiendo cierta complejidad y varianza al modelo, lo cual debemos tener en cuenta.

2.1.5 Random Forest

Finalmente, quisimos implementar un modelo de Random Forest, construido manualmente a partir de muchos árboles de regresión. Probando con distintos valores de hiperparámetros (número de árboles del bosque, máximo de features tomadas, máximo de muestras utilizadas), llegamos a conseguir una performance algo superior a las anteriores, con un R2 de hasta 0.862, MAE de 0.985 y MSE de 2.41.

2.1.6 Conclusión y elección del modelo

Aunque Random Forest y el árbol de decisión son opciones válidas y atractivas debido a su buen rendimiento, de hecho son los modelos que mejor rendimiento nos dan a nivel de métricas de performance, vamos a quedarnos con el modelo de Regresión Lineal como el elegido para el Modelo 1. Para justificar nuestra elección nos basamos en **la navaja de Ockham**, un principio filosófico y metodológico que defiende lo siguiente:

"En igualdad de condiciones, la explicación más simple suele ser la correcta."

Este principio establece que, entre varias explicaciones posibles para un fenómeno o relación, la explicación más simple suele ser la mejor, siempre y cuando esta explique adecuadamente los datos. Por ello, la **regresión lineal** es una opción adecuada, por su simplicidad, interpretabilidad y eficiencia computacional. Más adelante en `predicciones.ipynb`, aplicaremos este modelo.

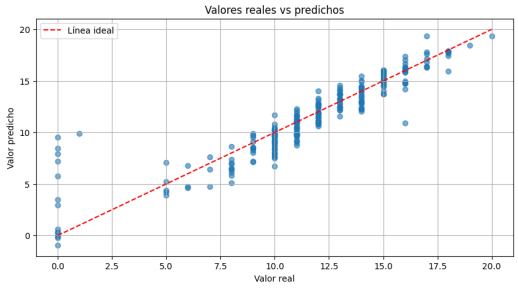

Figure 3: Rendimiento Regresión lineal

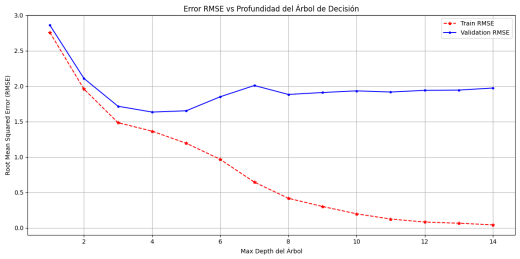

Figure 4: Cross Validation RegressionTree

2.2 Modelo 2

A diferencia del modelo anterior, ahora intentaremos predecir la variable objetivo `T3`, pero **sin contar con la información proporcionada por** `T1` **y** `T2`. Este proceso será muy interesante, puesto que nos ayudará a analizar qué otros factores clave son los que influyen en el rendimiento académico de los alumnos, permitiéndonos predecir qué tal va a ser su desempeño incluso antes de comenzar el curso. Todo el trabajo asociado a este enfoque se ha desarrollado en el `modelo2.ipynb`.

2.2.1 Random Forest

Como primer approach al problema, decidimos implementar Random Forest, ya que su rendimiento en el modelo 1 había sido bastante bueno y queríamos ver si lograba explicar correctamente la relación entre las features disponibles y la variable objetivo.

En este caso, desarrollamos un extenso proceso de **K-Fold Cross Validation** para encontrar la combinación óptima de los hiperparámetros. Tras sucesivas pruebas, logramos encontrar un rango interesante de valores, donde se conseguía una buena performance, además de un coste computacional no muy alto, y sin un número de árboles excesivo. Tengamos en cuenta que nuestro dataset no era muy extenso, por lo que generar muchísimos árboles acabaría generando árboles muy parecidos y redundantes. Finalmente, los hiperparámetros elegidos fueron:

- Número de árboles del bosque **(trees): 275**
- Máximo de variables tomadas para cada árbol **(features): 0.8**
- Máximo de muestras tomadas para cada árbol **(samples): 0.7**

Evaluando el rendimiento de este modelo, observamos que su R2 alcanzaba valores de 0.381, con un MAE de 2.2 y un RMSE de 3.156. Estas predicciones son obviamente muy limitadas, y el rendimiento del modelo no es el ideal; no obstante, teniendo en cuenta las características y dificultades de la tarea, se trata de una buena aproximación inicial.

2.2.2 Boosting

De igual modo que con Random Forest, desarrollamos un proceso de **K-Fold Cross Validation** para encontrar la combinación óptima de los hiperparámetros del modelo de Boosting. En este caso, nos costó menos encontrar una buena combinación de hiperparámetros, y como el modelo era algo más difícil de interpretar, tampoco quisimos añadir sofisticadas técnicas adicionales al modelo, ya que iban a dificultar su interpretabilidad y capacidad de generalización. De este modo, nos quedamos con la siguiente configuración:

- Número de **estimadores: 150**
- **Profundida máxima** de los árboles: **3**
- Proporción de muestras tomadas **(subsample): 0.7**

El rendimiento de este modelo fue algo peor que el de Random Forest, con un R2 de 0.342, un MAE de 2.342 y RMSE de 3.247 .

2.2.3 Conclusión y elección del modelo

Ambos modelos, tanto el de Boosting como el de Random Forest, nos brindan un **rendimiento limitado** y mejorable; sin embargo, hay que ser conscientes de que tratar de predecir el valor de `T3` a partir de variables que no incluyen directamente información sobre el rendimiento académico, como son `T1` y `T2`, es una tarea compleja.

Más adelante, en el apartado adicional del proyecto, seguiremos analizando más profundamente cuáles pueden ser los factores más influyentes en el rendimiento académico de los alumnos, así como las maneras de potenciarlo y remediar los problemas de los alumnos más necesitados. De momento, elegimos el modelo de **Random Forest** optimizado con validación cruzada, como nuestra opción para el Modelo 2, porque tiene mayor R2 que Boosting y es más sencillo de interpretar y de explicar.

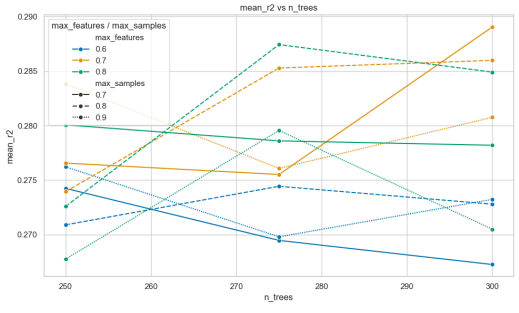

Figure 5: Cross Validation Random Forest

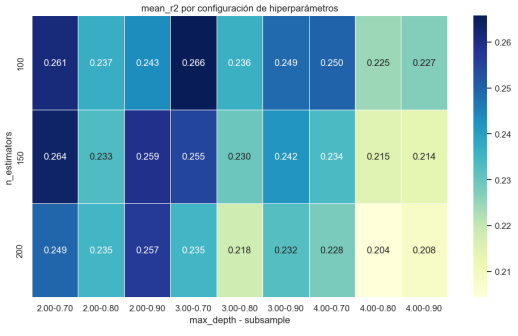

Figure 6: Cross Validation Boosting

2.3 Comparación de ambos modelos

Recordemos que nuestro modelo elegido para el modelo 1 era Regresión Lineal, mientras que para el modelo 2 era Random Forest optimizado con k-Fold Cross Validation.

Por un lado, el modelo de regresión lineal tiene una complejidad muy baja, no necesita ser entrenado ya que sus parámetros óptimos vienen dados por una fórmula cerrada, hace poco overfitting ya que solo es capaz de capturar relaciones lineales y puede ser bastante sensible a outliers.

En cambio, el modelo de Random Forest es mucho más complejo, es capaz de capturar relaciones no lineales, no es tan sensible a outliers, es muy flexible y puede tender a hacer overfitting. Además, sí que tiene ciertos hiperparámetros, por lo que requiere de alguna técnica de validación cruzada para hallar sus valores óptimos. A nivel computacional, es notablemente más costoso trabajar con este modelo y entrenarlo.

3 Análisis de los resultados

Una vez elegidos el Modelo 1 y el Modelo 2, los emplearemos para predecir datos que no han sido utilizados para su entrenamiento, los del dataset de test, del cual no tenemos la columna de `T3`. Este proceso se llevará a cabo en el fichero `predicciones.ipynb`.

Antes de nada, los datos de test deberán ser limpiados y tratados del mismo modo que se hizo con el conjunto de train en el fichero de `preprocesado.ipynb`, para que cuando dichos datos se introduzcan a los modelos como inputs, estos nos proporcionen predicciones coherentes sobre `T3`.

3.1 Comparación de resultados

Una vez hemos obtenido las predicciones de `T3` por parte de ambos modelos, podemos compararlas y analizar cuánto se parecen entre sí, si alguna predicción se sale del rango válido para las calificaciones, si las distribuciones son similares, etc.

Destacamos sobre todo las siguientes observaciones:

- La **máxima diferencia entre predicciones** de ambos modelos es menor que 3 puntos.
- Ambos conjuntos de predicciones tienen una **distribución parecida**, con simetría similar.
- Ambos conjuntos de predicciones tienen **la misma media de calificaciones.**
- Ambos conjuntos de predicciones tienen la **misma cantidad de suspensos y aprobados.**

4 Factores más influyentes en el rendimiento académico

En la misma línea del trabajo que veníamos haciendo en el fichero `modelo2.ipynb`, hemos seguido profundizando para explorar cuáles son los factores que mayor impacto tienen sobre los estudiantes, potenciando que estos consigan alcanzar mejores resultados académicos.

Obviamente `T1` y `T2` son las variables más influyentes en este aspecto, por lo que los sacaremos de la ecuación por un momento, con el objetivo de hallar otros factores también importantes que nos permitan hilar más fino.

4.1 Feature Importance

Para los dos candidatos con los que habíamos estado experimentando para el Modelo 2, analicemos qué variables son las que tienen más peso e importancia a la hora de hacer las predicciones. Hacemos el análisis tanto para Random Forest como para Boosting, y encontramos las siguientes similitudes en cuanto a variables más relevantes:

- `faltas`: Top 1 de importancia en Boosting y top 2 en RF
- `suspensos`: Top 2 de importancia en Boosting y top 1 en RF
- `asignatura_M`: Top 3 de importancia en Boosting y top 7 en RF
- `SalAm`: Top 5 de importancia en Boosting y top 3 en RF

En ambos modelos, destacan prácticamente las mismas variables a la hora de predecir las calificaciones de los alumnos. Podemos destacar, que la **asignatura de Matemáticas** está siendo un factor determinante en los alumnos, ya que o sacan bastantes buenas notas en dicha materia, o suspenden con notas preocupantemente bajas. Por otra parte, el **tiempo de ocio con amigos**, representa para algunos alumnos una liberación positiva y momento de descanso, mientras que para otros supone una peligrosa distracción. Son factores que debemos analizar en detalle.

4.2 Correlación con T3

Nuestro siguiente enfoque consiste en dividir el dataset en dos subconjuntos, formados por alumnos con calificaciones suspensas (< 10) o aprobadas (≥ 10). Estudiaremos en cada subconjunto las variables con mayor correlación con `T3`, para ver si hay factores que afectan más a un tipo de alumno o a otro. Los resultados que hemos obtenido son los siguientes:

Aprobados

Variable	Correlación
suspensos	-0.25
Medu	0.24
EstSup_si	0.23
TiempoEstudio	0.21

Suspensos

Variable	Correlación con T3
faltas	0.19
Ptrab_docencia	-0.17
TamFam_>=4	-0.17
asignatura_M	-0.17

Table 1: Top 4 variables más correlacionadas con T3 por subgrupos

En conclusión, entre los **aprobados**, el entorno familiar, las aspiraciones académicas y el acceso a tiempo y recursos para el estudio son claves para obtener buenos resultados. Estos elementos suelen estar vinculados a un perfil socioeconómico medio-alto.

Por otro lado, entre los alumnos **suspensos**, predominan factores que pueden reflejar desigualdad estructural o barreras socioeconómicas. Destacan familias numerosas , que pueden tener recursos limitados, asistencia irregular por parte de los alumnos y gran dificultad en asignaturas como Matemáticas, variable que ya hemos mencionado anteriormente.

A partir del análisis de los factores más influyentes en el rendimiento académico del alumnado, especialmente aquellos vinculados a perfiles con peores calificaciones, se podrían proponer las siguientes **políticas de mejora** en el ámbito educativo:

1. **Detección temprana de absentismo:** Establecer sistemas de alerta ante patrones de faltas reiteradas para actuar de forma preventiva, ofreciendo apoyo psicológico o social si es necesario.

2. **Adaptación metodológica de asignaturas críticas:** Revisar y flexibilizar la enseñanza en las asignaturas con mayores tasas de suspenso, incorporando nuevas metodologías.

3. **Reducción de ratios en aulas con vulnerabilidad:** Disminuir el número de alumnos por clase en grupos con mayor riesgo académico, favoreciendo atención más individualizada.

4. **Promoción de expectativas académicas positivas:** Trabajar con el alumnado para que visualicen la continuidad de sus estudios, mediante orientación académica y vocacional.

5. **Apoyo a familias numerosas o necesitadas:** Facilitar recursos específicos (becas, refuerzo escolar gratuito, comedor, transporte) a familias con dificultades económicas o logísticas.

5 Exploración adicional

5.1 Análisis de componentes principales (PCA)

El Análisis de Componentes Principales es una técnica de **reducción de dimensionalidad** que permite transformar un conjunto de variables posiblemente correlacionadas en un conjunto más pequeño de variables no correlacionadas. En este apartado, aplicaremos PCA a ambos subconjuntos de datos (suspensos y aprobados) utilizando el dataset de entrenamiento para identificar las componentes principales que explican la mayor parte de la varianza en cada caso.

Con este análisis, conseguimos ver que el grupo de suspensos necesita menos componentes principales para alcanzar una mayor varianza explicada que el grupo de aprobados. Esto sugiere que dicho grupo tiene menor complejidad estructural interna. Además, se ve que variables como `AlcFin` y `SalAm` tienen una considerable contribución a la primera componente principal del grupo de aprobados, del mismo modo que `suspensos` y `Medu` en el grupo de suspensos.

Así, pues, podemos deducir que el conjunto de alumnos aprobados, las variables con mayor peso están relacionadas con actividades sociales y tiempo libre, lo cual podría indicar un equilibrio positivo entre la vida personal y académica. En cambio, para el grupo de alumnos con peor rendimiento, las variables clave reflejan factores más estructurales, como el historial académico (suspensos) y el entorno familiar (nivel educativo de los padres), que parecen tener un impacto negativo en este subconjunto.

5.2 Clustering

El clustering es una **técnica de aprendizaje no supervisado** que nos permite agrupar observaciones en función de sus características, de manera que las observaciones dentro de un mismo grupo sean más similares entre sí que con las de otros grupos. En nuestro caso, vamos a aplicar clustering al dataset completo (excluyendo las variables T1 y T2) para encontrar grupos de estudiantes con características similares, evaluar qué tal es su rendimiento académico, y analizar las variables que diferencian a cada cluster del resto.

En este caso, hemos realizado **KMeans Clustering** sobre los datos proyectados en dos componentes principales, previamente obtenidas con PCA.

El **número de clústeres óptimo** para nuestro dataset es 3, y las características y distribuciones de cada clúster son las siguientes:

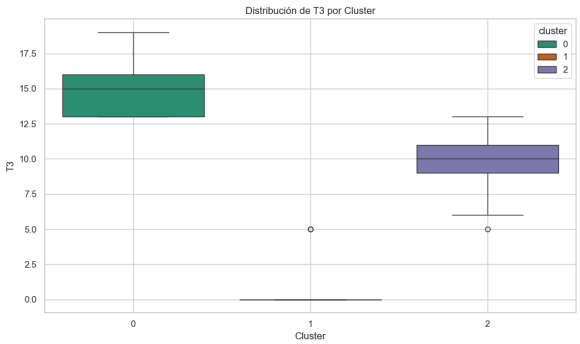

Figure 7: Distribución de `T3` según el clúster

Cluster	Tamaño	Media T3
0	227	14.92
1	34	0.29
2	323	9.89

Table 2: Resumen de estadísticas por cluster

El **cluster 0**, está formado por alumnos con un entorno familiar más favorable, madres con alto nivel educativo, buenos hábitos de estudio, bajo consumo de alcohol y pocos suspensos. Esto potencia claramente el rendimiento académico.

Por su parte, el **cluster 1** representa estudiantes en riesgo crítico de abandono escolar ya que sus calificaciones son prácticamente nulas, tienen una edad mayor que el resto (posiblemente han repetido algún curso), presentan notables dificultades en la asignatura de Matemáticas y viven un poco lejos de los centros escolares.

En el caso del **cluster 2**, se trata de alumnos con un rendimiento académico intermedio, los cuales podrían mejorar su rendimiento académico ajustando algunos hábitos de estudio y entorno. Este grupo presenta un consumo de alcohol algo mayor que el resto de clusters, y además dedican menos tiempo al estudio. Posiblemente, estos factores hacen que sus calificaciones no lleguen a ser excelentes aunque los estudiantes sí que tengan capacidad de conseguir mejores resultados.

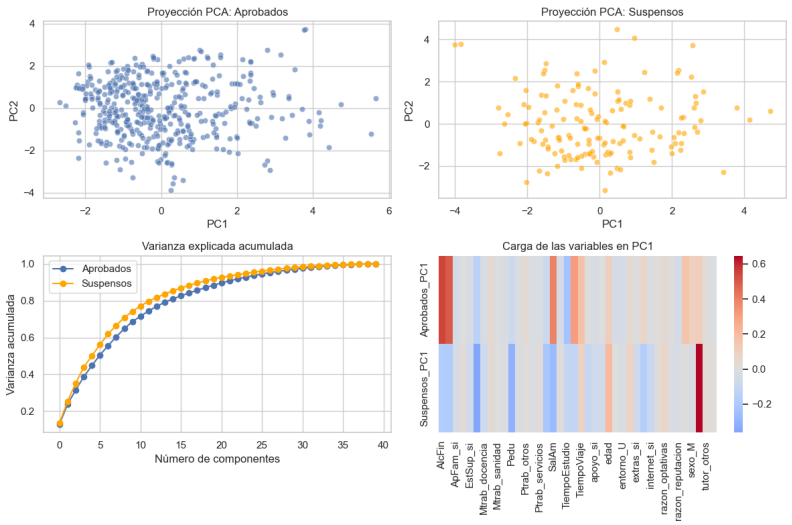

Figure 8: Resumen Estudio PCA

Figure 9: Proyección PC1 - PC2 tras Clustering

No se han añadido más imágenes en este documento para facilitar y amenizar su lectura; no obstante, se anima encarecidamente a **consultar los archivos adjuntos**, donde se recogen numerosos detalles sobre el análisis realizado, más gráficos ilustrativos y visualizaciones muy enriquecedoras, que sin duda facilitarán la comprensión del proceso que se ha desarrollado así como sus conclusiones.

6 Bibliografía

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (3rd ed.). Springer. Consultado en `https://www.statlearning.com/`

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, **12**, 2825–2830. Consultado en `https://scikit-learn.org/stable/`

[3] Castaño, S. (2018). *Comprende Principal Component Analysis (PCA) paso a paso*. Aprende Machine Learning. Consultado en `https://www.aprendemachinelearning.com/comprende-principal-component-analysis/`