Proyecto Final - Aprendizaje Automático

Jaime Pedrosa Comino
Grado en Ingeniería Matemática e Inteligencia Artificial
Universidad Pontificia Comillas - ICAI
jaimepedcom@hotmail.com
https://github.com/JaimePedrosaComino/ProyectoFinalMachineLearning.git

1 Introducción y Abstract

La predicción del rendimiento académico de los estudiantes es un área de investigación fundamental que busca comprender los factores que influyen en el desempeño educativo y anticipar posibles resultados para intervenir de forma temprana. El éxito académico no sólo depende de las capacidades cognitivas del alumnado, sino también de variables contextuales, sociales y motivacionales que, combinadas, generan una amplia diversidad de perfiles estudiantiles.

Este trabajo se centra en construir modelos de aprendizaje automático que permitan predecir la calificación final del tercer trimestre (`T3`) a partir de una combinación de variables académicas, familiares y personales de los estudiantes. La finalidad es doble: por un lado, identificar qué factores son determinantes en el rendimiento final, y por otro, disponer de un modelo predictivo que permita actuar a tiempo para mejorar los resultados del alumnado.

Se plantean dos escenarios de predicción: (i) un modelo que incluye las calificaciones previas (`T1` y `T2`) y (ii) otro que prescinde de estas para evaluar la capacidad predictiva del resto de variables. Ambos enfoques permiten analizar la influencia de las notas intermedias frente a otros factores en la estimación del desempeño final.

2 Metodología

2.1 Datos

El dataset utilizado proviene de datos reales anonimizados de estudiantes de secundaria, incluyendo variables demográficas, académicas y sociales. La variable objetivo es la calificación final del tercer trimestre (`T3`), expresada en una escala numérica.

Entre las variables se incluyen: género, edad, dirección familiar, nivel educativo de los padres, tiempo de estudio semanal, apoyo educativo adicional, actividades extracurriculares, entre otras. Esta diversidad permite capturar tanto dimensiones académicas como sociales del entorno estudiantil, aportando un contexto enriquecido al modelo predictivo.

2.1.1 Análisis Exploratorio de Datos (EDA)

El análisis inicial evidenció ciertos desequilibrios. Por ejemplo, la variable `sexo` está ligeramente desbalanceada hacia estudiantes de género femenino. Además, la mayoría de los estudiantes dispone de conexión a Internet en casa y apoyo escolar adicional. Se detectaron también variables numéricas con distribuciones sesgadas, como el tiempo de estudio semanal o la cantidad de ausencias.

La matriz de correlación (Figura 1) reveló relaciones clave entre las variables. Las calificaciones previas (`T1` y `T2`) mostraron una correlación muy alta con `T3` (0.82 y 0.92, respectivamente), confirmando su rol como predictores principales. También se observó una correlación negativa fuerte entre `suspensos` y `T3` (-0.38), indicando que un mayor número de suspensos está asociado con un peor

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

rendimiento. `TiempoEstudio` presentó una correlación positiva moderada con `T3` (0.18), mientras que `AlcFin` y `AlcSem` mostraron correlaciones negativas leves (-0.14 y -0.13), sugiriendo un impacto negativo del consumo de alcohol. Entre las variables predictoras, `Medu` y `Pedu` están altamente correlacionadas (0.61), y `AlcSem` y `AlcFin` también (0.64), lo que indica patrones consistentes en el nivel educativo de los padres y en el consumo de alcohol de los estudiantes. Este análisis resultó fundamental para orientar el preprocesamiento de datos y la selección de modelos.

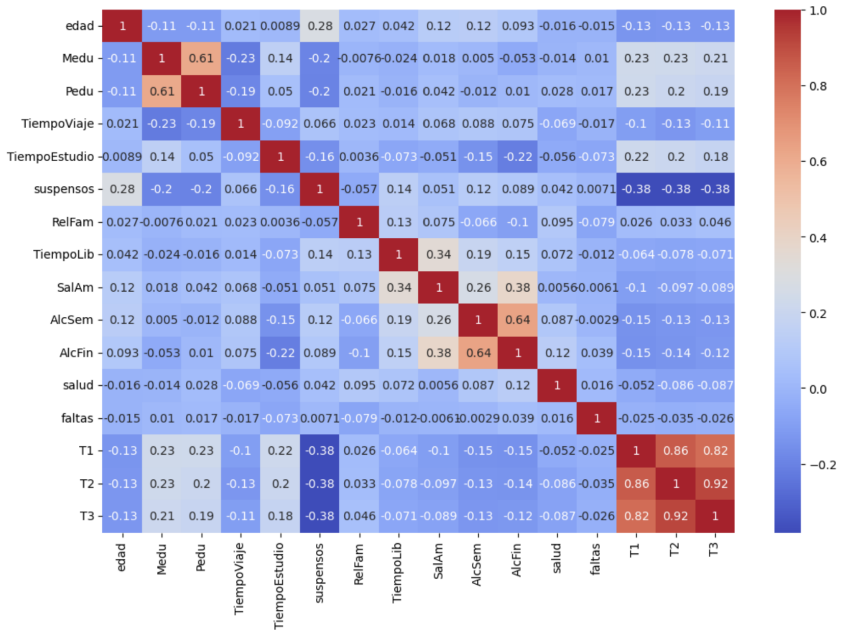


Figure 1: Matriz de correlación de las variables numéricas.

2.1.2 Preprocesamiento

El pipeline de preprocesamiento integró varias etapas cruciales:

- **Imputación de valores faltantes:** Se utilizaron estrategias de imputación basadas en la media para variables numéricas y en la moda para categóricas.
- **Estandarización de variables numéricas:** Mediante `StandardScaler` [**scikit-learn-standardscaler**], se aseguraron medias cero y desviaciones estándar unidad para mejorar la eficiencia del aprendizaje.
- **Codificación de variables categóricas:** Se aplicó `one-hot encoding`, transformando variables cualitativas en binarias para evitar sesgos de ordenamiento.
- **Transformación combinada:** Se utilizó `ColumnTransformer` [**scikit-learn-columntransformer**] para ejecutar estas transformaciones de forma simultánea, integrando datos numéricos y categóricos en una matriz final apta para el modelado.

2.2 Modelos Utilizados

Se seleccionaron tres enfoques complementarios para la modelización:

Regresión Lineal [**statsmodels-ols**]: Utilizando `statsmodels.OLS`, ofrece interpretabilidad directa mediante coeficientes que representan la relación lineal entre variables independientes y la nota final. Se aplicó selección de variables mediante la magnitud de los coeficientes, reteniendo aquellas con importancia absoluta superior a 0.01.

Random Forest [**sklearn-rf**]: El `RandomForestRegressor` permite capturar relaciones no lineales y manejar datos de alta dimensionalidad. Se calcularon importancias de variables para identificar predictores clave, replicando el proceso de filtrado del modelo lineal.

Red Neuronal Multicapa (MLP) [**pytorch-nn**]: Construida con PyTorch, incorpora dos capas ocultas, activación ReLU, Batch Normalization y Dropout para prevenir el sobreajuste. Se interpretaron las variables mediante *Permutation Importance*, proporcionando interpretabilidad incluso en modelos complejos.

Cada modelo se entrenó en dos versiones: (i) utilizando todas las variables, y (ii) empleando únicamente las seleccionadas tras el análisis de importancia.

3 Experimentos

La configuración experimental incluyó:

- **División del dataset:** 80% para entrenamiento y 20% para prueba.
- **Evaluación cruzada:** Validación interna durante el entrenamiento de la MLP.
- **Métricas evaluadas:** Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y Coeficiente de Determinación (R^2).

Estas métricas permiten evaluar tanto la precisión como la robustez de los modelos, considerando errores absolutos y su dispersión respecto a la varianza total.

4 Resultados

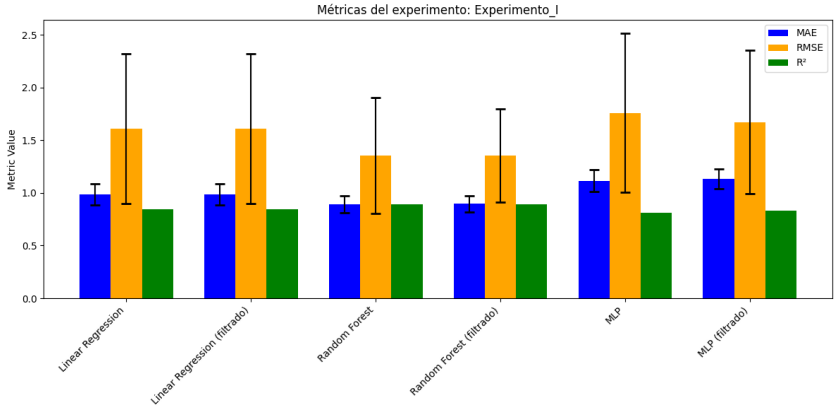
Los experimentos compararon el desempeño predictivo de los modelos en dos escenarios: Experimento I, que incluye las calificaciones previas (`T1` y `T2`), y Experimento II, que las excluye para evaluar la capacidad predictiva de otras variables. La Tabla 1 presenta las métricas de error (MAE, RMSE y R^2) para cada modelo en ambos experimentos, considerando las versiones con todas las variables y las filtradas (importancia absoluta > 0.01). Las métricas se reportan para el conjunto de prueba (Test).

Table 1: Comparación de métricas de error para los modelos en los Experimentos I y II. Las métricas se reportan para el conjunto de prueba (Test).

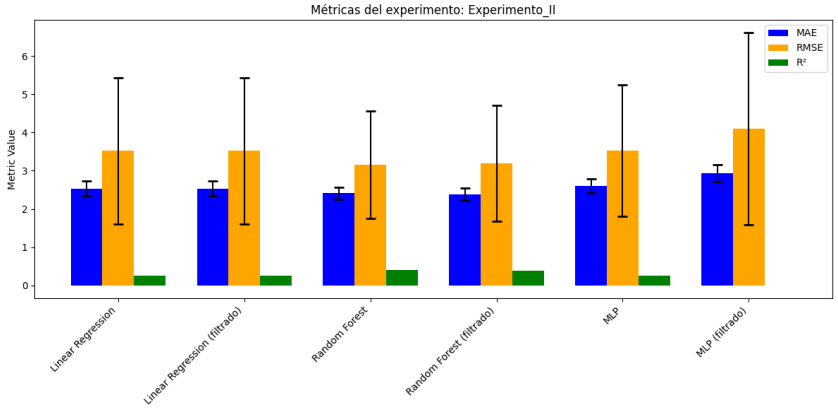
		Experimento I			**Experimento II**		
Modelo	**Versión**	**MAE**	**RMSE**	R^2	**MAE**	**RMSE**	R^2
Linear Regression	Todas las variables	0.985	1.611	0.844	2.534	3.522	0.255
	Filtrado	0.985	1.611	0.844	2.534	3.522	0.255
Random Forest	Todas las variables	0.891	1.355	0.890	2.407	3.159	0.401
	Filtrado	0.895	1.357	0.889	2.387	3.193	0.388
MLP	Todas las variables	1.115	1.760	0.814	2.600	3.527	0.253
	Filtrado	1.134	1.673	0.832	2.941	4.097	-0.008

En el Experimento I, los modelos aprovecharon la alta correlación de `T1` y `T2` con `T3` (0.82 y 0.92), logrando R^2 altos: Random Forest obtuvo el mejor resultado (0.890, MAE: 0.891, RMSE: 1.355), seguido por Regresión Lineal (0.844, MAE: 0.985, RMSE: 1.611) y MLP (0.814, MAE: 1.115, RMSE: 1.760). El filtrado no afectó a la Regresión Lineal, redujo ligeramente el R^2 de Random Forest a 0.889, pero mejoró el del MLP a 0.832 (MAE: 1.134, RMSE: 1.673), sugiriendo que la selección de variables puede beneficiar a modelos complejos al reducir ruido. En el Experimento II, al excluir `T1` y `T2`, el rendimiento cayó notablemente: Random Forest mostró mayor robustez (0.401, MAE: 2.407, RMSE: 3.159), seguido por Regresión Lineal (0.255, MAE: 2.534, RMSE: 3.522) y MLP (0.253, MAE: 2.600, RMSE: 3.527). El filtrado en el MLP resultó en un R^2 negativo (-0.008, MAE: 2.941, RMSE: 4.097), indicando que el modelo (con solo `TiempoEstudio` y `AlcSem`) fue

peor que predecir la media de T3. La diferencia en RMSE entre los experimentos (1.355–1.760 en Experimento I frente a 3.159–4.097 en Experimento II) subraya la importancia de T1 y T2, mientras que Random Forest destacó por su capacidad para capturar patrones con variables secundarias.



(a) Métricas Experimento I (con T1 y T2)



(b) Métricas Experimento II (sin T1 y T2)

Figure 2: Comparación gráfica de las métricas de los modelos en ambos experimentos. Las barras de error en MAE y RMSE indican la variabilidad de las predicciones.

Las figuras ilustran gráficamente las métricas de ambos experimentos. En el Experimento I, el gráfico muestra que Random Forest (con todas las variables) tiene las barras más bajas para MAE y RMSE (0.891 y 1.355), y la barra más alta para R^2 (0.890), confirmando su superioridad. Las barras de error para MAE y RMSE son pequeñas en todos los modelos, indicando una baja variabilidad en las predicciones, lo que refleja la robustez de los modelos cuando T1 y T2 están presentes. En el MLP, el filtrado reduce visiblemente la altura de la barra de RMSE (de 1.760 a 1.673) y aumenta la de R^2 (de 0.814 a 0.832), evidenciando el beneficio de la selección de variables. En el Experimento II, el gráfico revela un aumento drástico en las barras de MAE y RMSE para todos los modelos, especialmente en el MLP filtrado, donde RMSE alcanza 4.097 y R^2 cae a -0.008, reflejado por una barra negativa. Random Forest mantiene las barras más bajas para MAE y RMSE (2.407 y 3.159) y la más alta para R^2 (0.401), pero las barras de error son más amplias, indicando mayor variabilidad en las predicciones debido a la ausencia de T1 y T2. La comparación visual entre ambos gráficos subraya cómo la exclusión de estas variables clave afecta negativamente el desempeño de todos los modelos, especialmente del MLP filtrado.

5 Análisis e Importancia de Variables

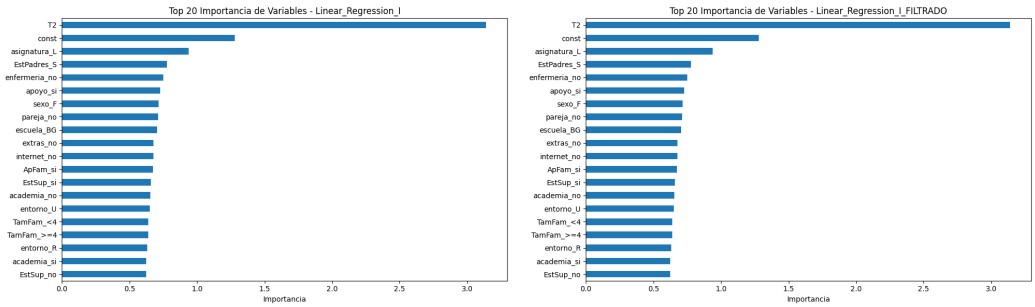
Se analizó la importancia de las variables en ambos escenarios, destacando diferencias significativas entre los modelos y los enfoques completo y filtrado.

5.1 Experimento I (con T1 y T2)

En todos los modelos, `T1` y `T2` emergieron como los predictores más relevantes, lo cual es coherente con su alta correlación con `T3` (0.82 y 0.92, respectivamente, según la matriz de correlación).

5.1.1 Regresión Lineal

`T2` presentó un coeficiente de 3.0, lo que indica que un aumento de 1 punto en `T2` incrementa `T3` en 3 puntos, manteniendo las demás variables constantes. `T1` mostró un coeficiente de 0.5, y otras variables como `asignatura_L` (0.8) también resultaron relevantes. El filtrado de variables (importancia absoluta > 0.01) mantuvo estas variables principales, eliminando aquellas con menor peso.

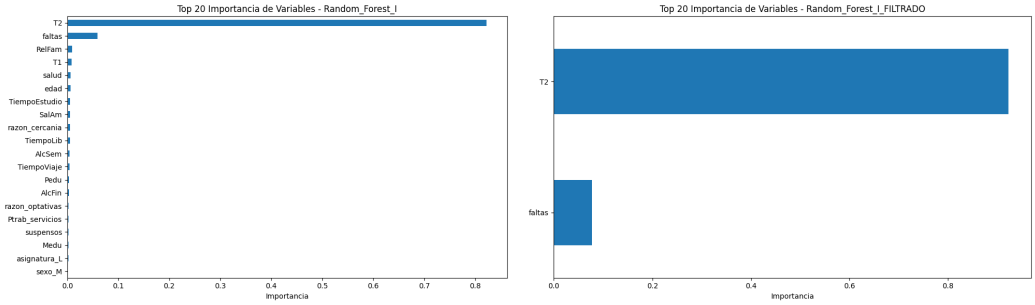


(a) Regresión Lineal completa (b) Regresión Lineal filtrado

Figure 3: Importancia de variables para Regresión Lineal en el Experimento I.

5.1.2 Random Forest

`T2` lideró con una importancia de 0.8, seguida por `faltas` (0.2). El filtrado redujo el análisis a estas dos variables, confirmando su dominancia en la predicción.

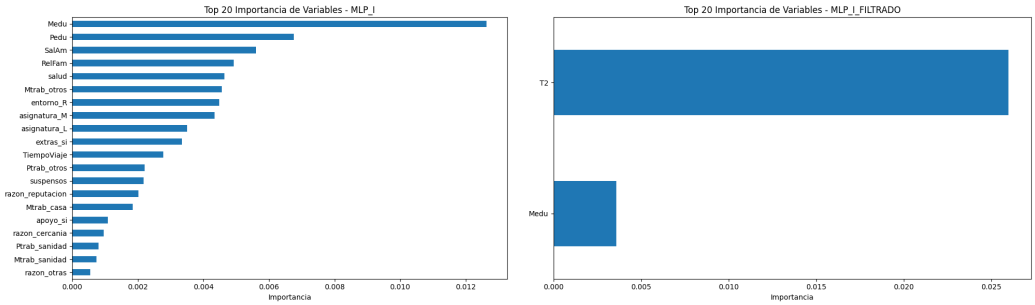


(a) Random Forest completo (b) Random Forest filtrado

Figure 4: Importancia de variables para Random Forest en el Experimento I.

5.1.3 MLP

`T1` (importancia 0.12) y `escuela_BG` (0.11) destacaron en el modelo completo. Tras el filtrado, `T1` (0.14) y `suspensos` (0.08) se consolidaron como las más influyentes, mostrando que el historial de suspensos también juega un papel relevante cuando se reduce el conjunto de variables.



(a) MLP completo (b) MLP filtrado

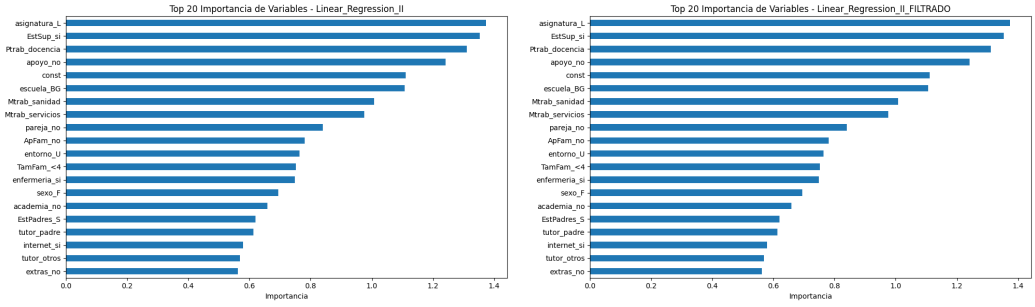
Figure 5: Importancia de variables para MLP en el Experimento I.

5.2 Experimento II (sin T1 y T2)

Al eliminar `T1` y `T2`, las variables contextuales y socioeducativas ganaron relevancia, reflejando su papel como predictores secundarios.

5.2.1 Regresión Lineal

`asignatura_L` (coeficiente 1.4) y `EstSup_si` (1.2) lideraron, indicando que el apoyo educativo adicional tiene un impacto positivo significativo en el rendimiento. El filtrado mantuvo estas variables, eliminando las de menor peso.

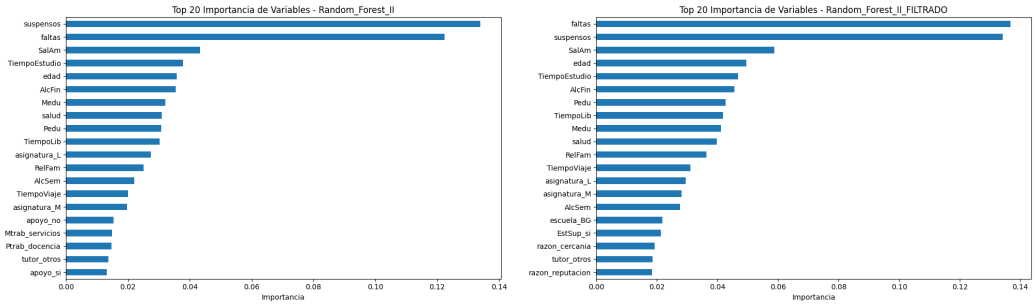


(a) Regresión Lineal completa (b) Regresión Lineal filtrado

Figure 6: Importancia de variables para Regresión Lineal en el Experimento II.

5.2.2 Random Forest

`suspensos` (importancia 0.14) y `faltas` (0.12) fueron las más influyentes, sugiriendo que la asistencia y el historial académico previo (en términos de suspensos) son proxies del compromiso estudiantil. El filtrado no alteró este ranking, manteniendo estas variables como las principales.

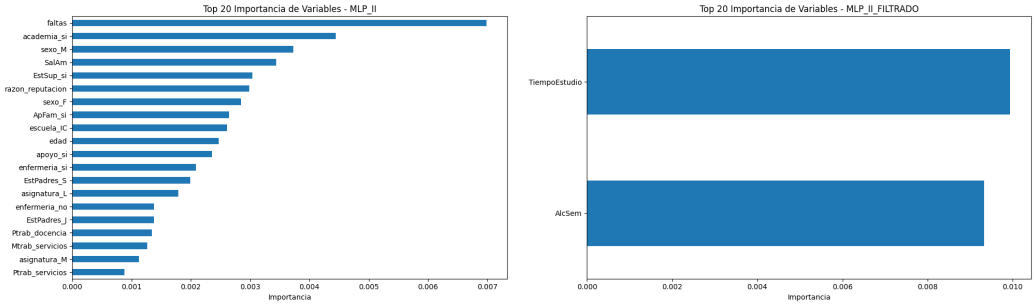


(a) Random Forest completo (b) Random Forest filtrado

Figure 7: Importancia de variables para Random Forest en el Experimento II.

5.2.3 MLP

`AlcFin` (importancia 0.12) y `TiempoLIB` (0.11) destacaron en el modelo completo, mostrando que el consumo de alcohol al final de semana y el tiempo libre impactan negativamente el rendimiento. Tras el filtrado, `TiempoLIB` (0.025) y `AlcFin` (0.015) se consolidaron como las más influyentes.



(a) MLP completo (b) MLP filtrado

Figure 8: Importancia de variables para MLP en el Experimento II.

5.3 Reflexión teórica

Los resultados confirman que las calificaciones previas (`T1` y `T2`) son predictores dominantes, como se esperaba por su alta correlación con `T3` (0.82 y 0.92, según la matriz de correlación). Sin embargo, en su ausencia, variables como `faltas`, `suspensos`, `TiempoLIB` y `EstSup_si` emergen como factores accionables. La matriz de correlación respalda la relevancia de `suspensos`, con una correlación negativa fuerte con `T3` (-0.38), lo que explica su importancia en el Experimento II. Reducir las ausencias (`faltas`) podría lograrse mediante políticas de asistencia más estrictas o programas de motivación, ya que `faltas` está relacionada con `suspensos` (correlación 0.0071). Por otro lado, `EstSup_si` mostró un impacto positivo, sugiriendo que las tutorías personalizadas podrían ser una intervención efectiva para mejorar el rendimiento. Variables como `Medu` (nivel educativo de la madre) o `AlcFin` (correlación -0.14 con `T3`) son menos accionables directamente, pero podrían abordarse mediante campañas de concienciación o apoyo familiar. Esto último es especialmente relevante, ya que `TiempoLib` está correlacionado con `AlcFin` (0.15), lo que indica que un exceso de tiempo libre puede derivar en hábitos menos productivos.

6 Conclusiones

Este trabajo demuestra que es posible predecir el rendimiento académico de los estudiantes mediante técnicas de aprendizaje automático, pero el desempeño depende en gran medida de la disponibilidad de predictores clave como las calificaciones previas. En el Experimento I, que incluye `T1` y `T2`, los modelos alcanzaron un R^2 de hasta 0.890 (Random Forest, con todas las variables), con un MAE de

0.891 y un RMSE de 1.355 en el conjunto de prueba, lo que refleja la alta correlación de `T1` y `T2` con `T3` (0.82 y 0.92, según la matriz de correlación). En contraste, en el Experimento II, al excluir estas variables, el mejor R^2 fue de 0.401 (Random Forest, con todas las variables), con un MAE de 2.407 y un RMSE de 3.159, lo que evidencia una caída significativa en el rendimiento predictivo.

La matriz de correlación respalda estos hallazgos: variables como `suspensos` (correlación -0.38 con `T3`) y `faltas` emergieron como predictores clave en el Experimento II, especialmente para Random Forest, que mostró mayor robustez frente a la ausencia de `T1` y `T2`. Por otro lado, el MLP fue más sensible a esta exclusión, con un R^2 negativo (-0.008) en su versión filtrada, indicando que el modelo filtrado (con solo `TiempoEstudio` y `AlcSem`) no logró capturar patrones útiles y fue peor que predecir la media de `T3`. Esto sugiere que las redes neuronales pueden requerir un conjunto más amplio de variables para generalizar bien en este contexto, mientras que Random Forest y Regresión Lineal logran aprovechar mejor las variables secundarias.

El impacto de factores socioeducativos también es notable. La matriz de correlación mostró un efecto negativo del consumo de alcohol (`AlcFin`, correlación -0.14 con `T3`, y `AlcSem`, correlación -0.13), y el análisis de importancia en el Experimento II destacó `AlcSem` como una variable relevante para el MLP filtrado. Además, `TiempoEstudio` (correlación 0.18 con `T3`) y `EstSup_si` (coeficiente 1.353 en Regresión Lineal) sugieren que fomentar hábitos de estudio y proporcionar apoyo educativo adicional pueden ser intervenciones efectivas para mejorar el rendimiento, especialmente cuando no se dispone de calificaciones previas. La correlación entre `TiempoLib` y `AlcFin` (0.15) indica que un exceso de tiempo libre puede derivar en hábitos menos productivos, lo que refuerza la necesidad de programas que promuevan un uso estructurado del tiempo.

La selección de variables mejoró la interpretabilidad y redujo la complejidad computacional, especialmente para el MLP en el Experimento I, donde el R^2 aumentó de 0.814 a 0.832 tras el filtrado. Sin embargo, en el Experimento II, el filtrado tuvo un impacto negativo en el MLP, lo que sugiere que un umbral de importancia más bajo o un enfoque menos agresivo podría ser más adecuado cuando las variables predictivas son menos informativas.

En conjunto, estos modelos pueden integrarse en sistemas de alerta temprana para identificar estudiantes en riesgo y aplicar intervenciones personalizadas. Por ejemplo, reducir las ausencias (`faltas`) mediante políticas de asistencia más estrictas y ofrecer tutorías (`EstSup_si`) podrían mitigar el impacto de factores como `suspensos` y `AlcFin`. Además, enfoques como el clustering propuesto podrían segmentar a los estudiantes en grupos de riesgo, permitiendo intervenciones más dirigidas. Futuras investigaciones podrían explorar la incorporación de datos longitudinales o variables dinámicas (como la evolución del compromiso estudiantil) para mejorar aún más la precisión predictiva y la aplicabilidad práctica de estos modelos.

References

[1] Scikit-learn documentation: *StandardScaler*. Disponible en `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

[2] Scikit-learn documentation: *ColumnTransformer*. Disponible en `https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html`

[3] Statsmodels documentation: *OLS (Ordinary Least Squares)*. Disponible en `https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html`

[4] Scikit-learn documentation: *RandomForestRegressor*. Disponible en `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`

[5] PyTorch documentation: *torch.nn*. Disponible en `https://pytorch.org/docs/stable/nn.html`