UD02_01 Práctica Predicción desempeño académico

Sistemas de Aprendizaje Automático.
Álvaro Martínez Lineros

Análisis del problema.

El objetivo de este estudio es desarrollar un modelo de clasificación que prediga el desempeño académico de los estudiantes utilizando un conjunto de datos que contiene información sobre más de 21,000 alumnos. La clasificación se realizará en tres categorías de rendimiento: "Bajo", "Medio" y "Alto". Los factores considerados incluyen aspectos académicos, socioeconómicos y de hábitos de estudio.

Este estudio permitirá identificar patrones que influyen en el rendimiento estudiantil, ayudando a las instituciones educativas a tomar decisiones informadas y a diseñar intervenciones personalizadas para mejorar el éxito académico.

Estudio preliminar de los datos.

Al hacer un recuento y exploración inicial de los datos se ha visto que hay un gran desbalance entre las clases proporcionadas en el dataset. Con grandes desbalances en las clases de los datos a predecir, los modelos pueden llegar a no generar reglas complejas para la predicción de los datos sino memorizar los datos de prueba.

Existen técnicas para combatir este problema, entre ellas:

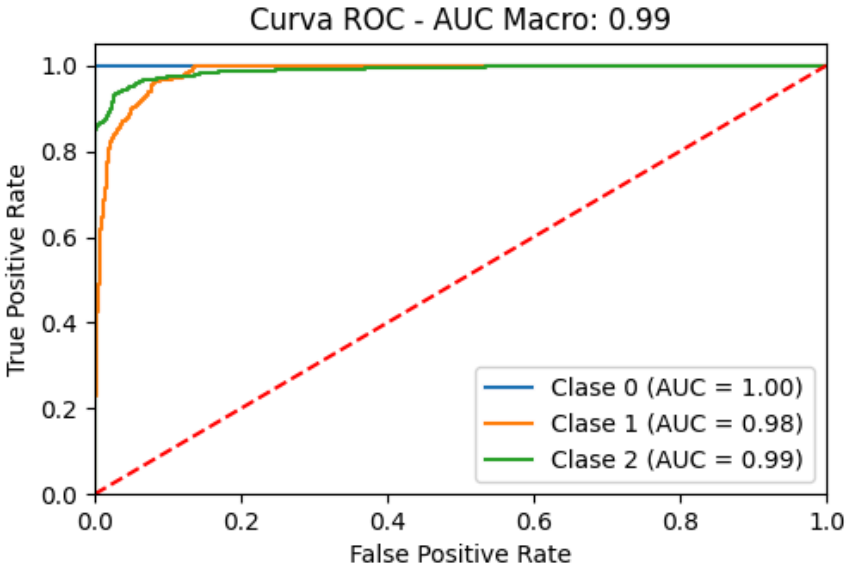
- SMOTE: genera muestras sintéticas de la clase minoritaria para balancear la cantidad de muestras por clase. Se utiliza la librería de imbalanced-learn.
- Undersampling: reducir el número de muestras de la clase mayoritaria.

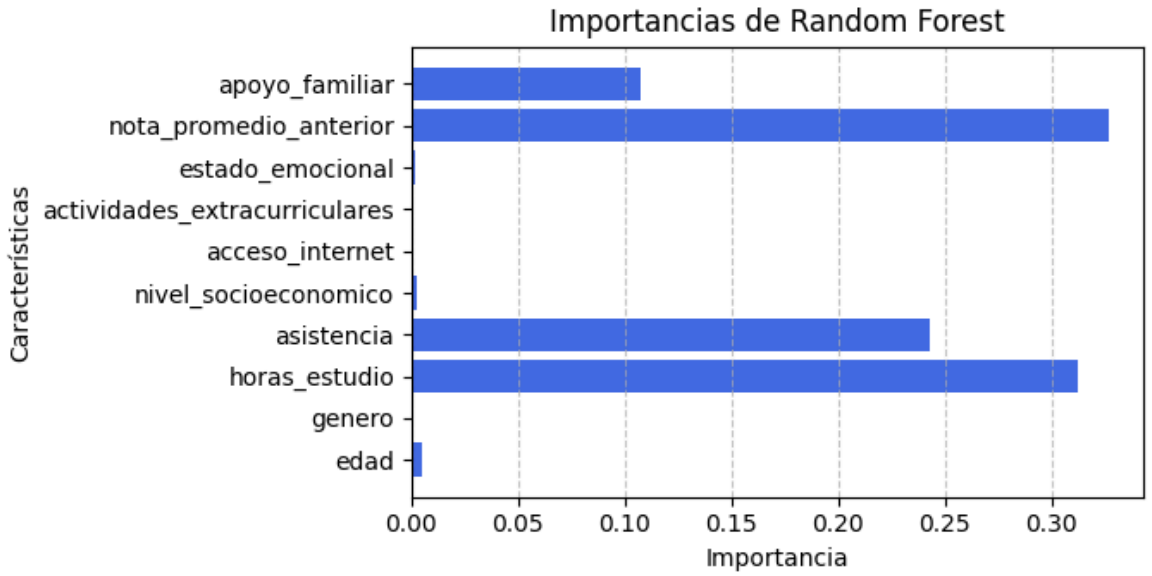
Estas técnicas pueden aplicarse individualmente o conjuntamente. SMOTE es especialmente útil cuando la clase minoritaria es mucho menor que la mayoritaria. Undersampling es más útil cuando hay un gran desbalance en la clase mayoritaria. En este caso, tanto la clase minoritaria (rendimiento académico alto) como la clase mayoritaria (rendimiento académico bajo) tienen valores demasiado extremos, por lo que se ha realizado una aplicación conjunta de SMOTE y undersampling.

Comando: pip install imblearn. Es necesario tener una versión de numpy igual o inferior a 2.1.0.

Random Forest.

Se entrenó un modelo utilizando el algoritmo de Random Forest con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:





Matriz de confusión:

	Predicción Alto	Predicción Bajo	Predicción Medio
Real Alto	156	0	0
Real Bajo	0	2417	66
Real Medio	32	127	1402

Reporte de clasificación:

	precision	recall	f1-score	support
Alto	0.83	1.00	0.91	156

Bajo	0.95	0.97	0.96	2483
Medio	0.95	0.90	0.93	1561
accuracy			0.95	4200
macro avg	0.91	0.96	0.93	4200
weighted avg	0.95	0.95	0.95	4200

Importancia de las características:

Característica	Importancia
edad	0.005109
genero	0.000055
horas_estudio	0.312800
asistencia	0.243173
nivel_socioeconomico	0.002600
acceso_internet	0.000082
actividades_extracurriculares	0.000574
estado_emocional	0.001300
nota_promedio_anterior	0.326819
apoyo_familiar	0.107489

Interpretación de los resultados.

Se obtuvo una precisión del 95% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 32 personas identificadas con rendimiento alto que no se previeron correctamente, 66 personas no identificadas con rendimiento bajo y 159 personas no identificadas con rendimiento medio.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

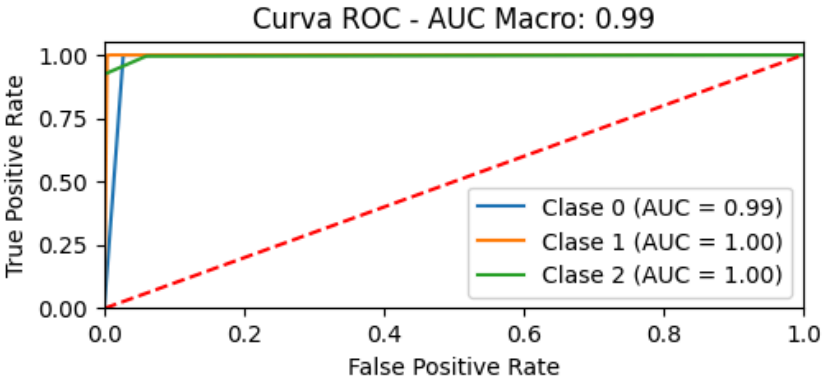
- Precisión:
    - Se acertó el 83% de predicciones para rendimiento alto.
    - Se acertó el 95% de predicciones para rendimiento bajo.
    - Se acertó el 96% de predicciones para rendimiento medio.
- Recall:
    - El 100% de los alumnos con rendimiento alto fueron identificados correctamente.

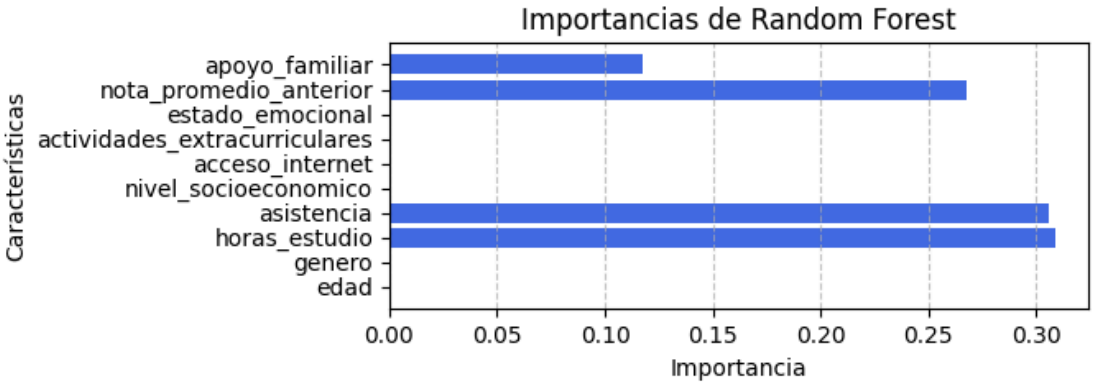
        - El 97% de los alumnos con rendimiento bajo fueron identificados correctamente.
        - El 90% de los alumnos con rendimiento medio fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para alumnos con rendimiento alto muestra muy buen rendimiento.
    - El balance entre precisión y recall para alumnos con rendimiento bajo muestra un rendimiento excepcional.
    - El balance entre precisión y recall para alumnos con rendimiento medio muestra muy buen rendimiento.
- Support:
    - La muestra cuenta con 156 alumnos con rendimiento alto.
    - La muestra cuenta con 2483 alumnos con rendimiento bajo.
    - La muestra cuenta con 1561 alumnos con rendimiento medio.
- Accuracy:
    - La precisión total del modelo es del 95%
- Weighted avg:
    - El promedio ponderado de las métricas es del 0.95

Finalmente, se puede observar la importancia de las características. La característica que más influye es la nota promedio anterior, seguido de las horas de estudio y la asistencia. Por otro lado, el género, las actividades extracurriculares y el acceso a internet tienen un impacto ínfimo o prácticamente nulo.

Árboles de decisión.

Se entrenó un modelo utilizando el algoritmo de árboles de decisión con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:





Matriz de confusión:

	Predicción Alto	Predicción Bajo	Predicción Medio
Real Alto	156	0	0
Real Bajo	0	2438	0
Real Medio	110	18	1433

Reporte de clasificación:

	precision	recall	f1-score	support
Alto	0.59	1.00	0.74	156
Bajo	0.99	1.00	1.00	2483

Medio	1.00	0.92	0.96	1561
accuracy			0.97	4200
macro avg	0.86	0.97	0.90	4200
weighted avg	0.98	0.97	0.97	4200

Importancia de las características:

Característica	Importancia
edad	0.000000
genero	0.000000
horas_estudio	0.308595
asistencia	0.305931
nivel_socioeconomico	0.000000
acceso_internet	0.000000
actividades_extracurriculares	0.000000
estado_emocional	0.000000
nota_promedio_anterior	0.267700
apoyo_familiar	0.117774

Interpretación de los resultados.

Se obtuvo una precisión del 97% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 110 personas identificadas con rendimiento alto que no se previeron correctamente y 128 personas no identificadas con rendimiento medio.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

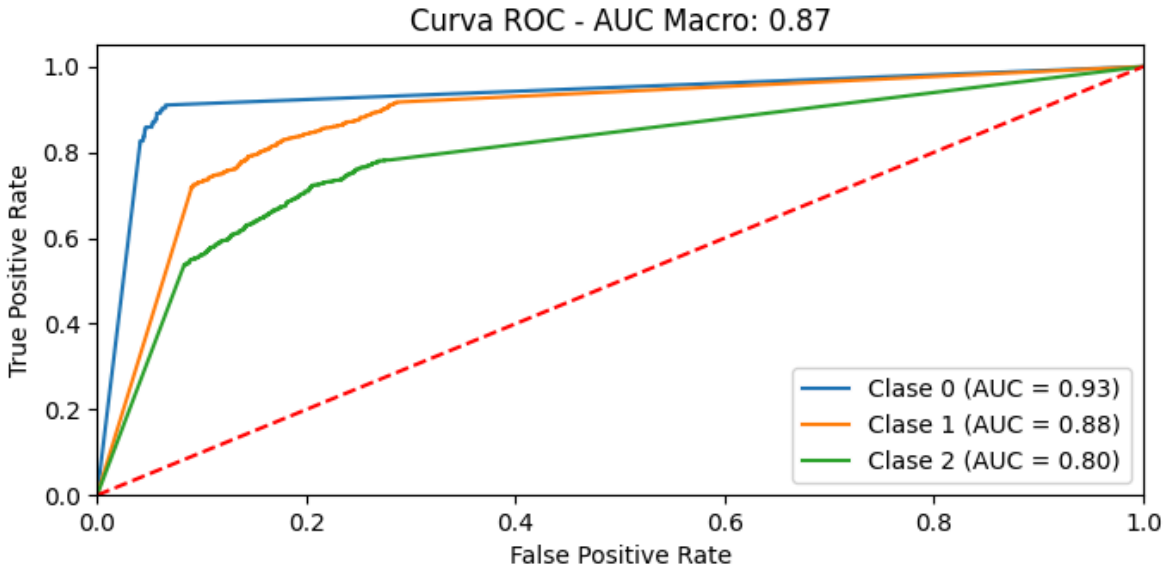
- Precisión:
  - Se acertó el 59% de predicciones para rendimiento alto.
  - Se acertó el 99% de predicciones para rendimiento bajo.
  - Se acertó el 100% de predicciones para rendimiento medio.
- Recall:
  - El 100% de los alumnos con rendimiento alto fueron identificados correctamente.
  - El 100% de los alumnos con rendimiento bajo fueron identificados correctamente.

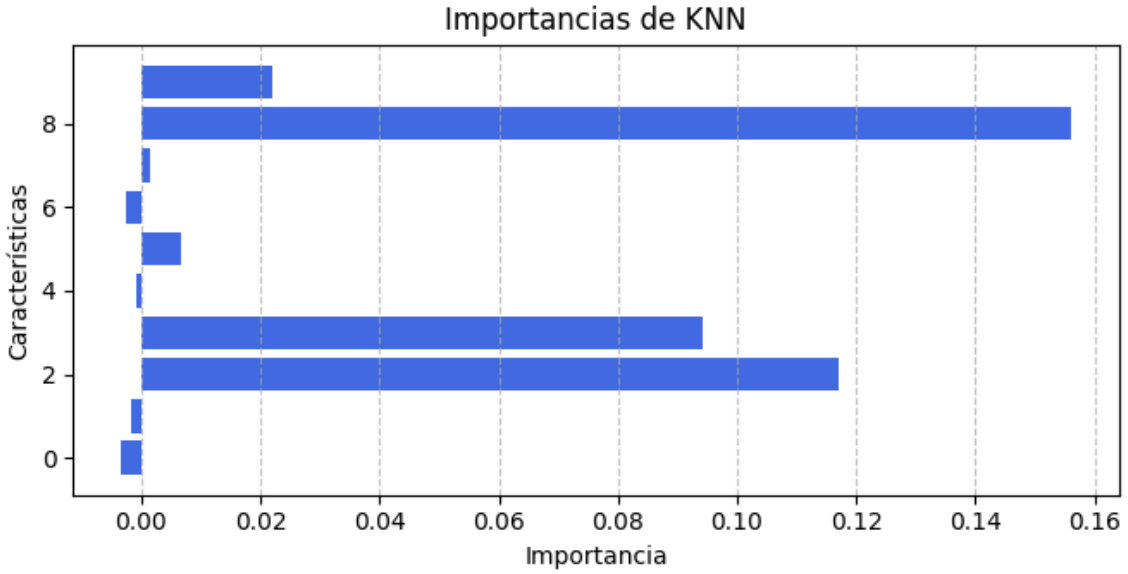
        - El 92% de los alumnos con rendimiento medio fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para alumnos con rendimiento alto muestra un rendimiento bajo.
    - El balance entre precisión y recall para alumnos con rendimiento bajo muestra un rendimiento excepcional.
    - El balance entre precisión y recall para alumnos con rendimiento medio muestra muy buen rendimiento.
- Support:
    - La muestra cuenta con 156 alumnos con rendimiento alto.
    - La muestra cuenta con 2483 alumnos con rendimiento bajo.
    - La muestra cuenta con 1561 alumnos con rendimiento medio.
- Accuracy:
    - La precisión total del modelo es del 97%
- Weighted avg:
    - El promedio ponderado de las métricas es del 0.97

Finalmente, se puede observar la importancia de las características. La característica que más influye son las horas de estudio, seguido de la asistencia y el apoyo familiar. Por otro lado, prácticamente el resto de características tienen un impacto nulo.

K-Nearest Neighbors (KNN).

Se entrenó un modelo utilizando el algoritmo de KNN con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:





Matriz de confusión:

	Predicción Alto	Predicción Bajo	Predicción Medio
Real Alto	135	0	21
Real Bajo	3	2058	422
Real Medio	211	312	1038

Reporte de clasificación:

	precision	recall	f1-score	support

Alto	0.39	0.87	0.53	156
Bajo	0.87	0.83	0.85	2483
Medio	0.70	0.66	0.68	1561
accuracy			0.77	4200
macro avg	0.65	0.79	0.69	4200
weighted avg	0.79	0.77	0.77	4200

Importancia de las características:

Característica	Importancia
edad	-0.003429
genero	-0.001762
horas_estudio	0.116929
asistencia	0.094190
nivel_socioeconomico	-0.000738
acceso_internet	0.006833
actividades_extracurriculares	-0.002548
estado_emocional	0.001381
nota_promedio_anterior	0.156143
apoyo_familiar	0.022095

Interpretación de los resultados.

Se obtuvo una precisión del 77% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 21 personas identificadas con rendimiento medio que no se previeron correctamente como rendimiento alto. También hubo 425 personas no identificadas con rendimiento bajo, y 523 personas no identificadas con rendimiento medio.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
    - Se acertó el 39% de predicciones para rendimiento alto.
    - Se acertó el 87% de predicciones para rendimiento bajo.
    - Se acertó el 70% de predicciones para rendimiento medio.
- Recall:

    - El 87% de los alumnos con rendimiento alto fueron identificados correctamente.
    - El 83% de los alumnos con rendimiento bajo fueron identificados correctamente.
    - El 66% de los alumnos con rendimiento medio fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para alumnos con rendimiento alto muestra un rendimiento extremadamente bajo.
    - El balance entre precisión y recall para alumnos con rendimiento bajo muestra un rendimiento aceptable.
    - El balance entre precisión y recall para alumnos con rendimiento medio muestra bajo rendimiento.
- Support:
    - La muestra cuenta con 156 alumnos con rendimiento alto.
    - La muestra cuenta con 2483 alumnos con rendimiento bajo.
    - La muestra cuenta con 1561 alumnos con rendimiento medio.
- Accuracy:
    - La precisión total del modelo es del 77%
- Weighted avg:
    - El promedio ponderado de las métricas es del 0.77

Finalmente, se puede observar la importancia de las características. La característica que más influye es la nota promedio anterior, seguido de la asistencia y el las horas de estudio. Por otro lado, prácticamente el resto de características tienen un impacto nulo.

Support Vector Machines (SVM).

Se entrenó un modelo utilizando el algoritmo de SVM con:

- kernel rbf: un kernel no lineal que mide la similitud entre los puntos de datos y crea una curvatura.
- C con valor de 1.0: valor por defecto para la tolerancia a errores
- Gamma de tipo scale: ajusta gamma según la fórmula $\gamma = 1 / n \cdot \text{varianza}(X)$. Gamma es el parámetro que controla la influencia de cada punto en la función de decisión.

Se entrenó el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:

Matriz de confusión:

	Predicción Alto	Predicción Bajo	Predicción Medio
Real Alto	153	0	1
Real Bajo	0	2348	108
Real Medio	81	153	1320

Reporte de clasificación:

	precision	recall	f1-score	support
Alto	0.65	0.99	0.79	154
Bajo	0.94	0.96	0.95	2492
Medio	0.92	0.85	0.89	1554
accuracy			0.92	4200
macro avg	0.84	0.93	0.87	4200
weighted avg	0.92	0.92	0.92	4200

Importancia de las características:

Característica	Importancia
edad	0.000452
genero	0.000262
horas_estudio	0.198143
asistencia	0.166905
nivel_socioeconomico	-0.001571
acceso_internet	0.002619
actividades_extracurriculares	-0.001333
estado_emocional	0.002452
nota_promedio_anterior	0.228714
apoyo_familiar	0.021190

Interpretación de los resultados.

Se obtuvo una precisión del 92% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 1 persona identificada con rendimiento medio que no se previó correctamente como rendimiento alto. También hubo 108 personas no identificadas con rendimiento bajo, y 234 personas no identificadas con rendimiento medio.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
    - Se acertó el 65% de predicciones para rendimiento alto.
    - Se acertó el 94% de predicciones para rendimiento bajo.
    - Se acertó el 92% de predicciones para rendimiento medio.
- Recall:
    - El 99% de los alumnos con rendimiento alto fueron identificados correctamente.
    - El 96% de los alumnos con rendimiento bajo fueron identificados correctamente.
    - El 85% de los alumnos con rendimiento medio fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para alumnos con rendimiento alto muestra un rendimiento mejorable pero aceptable.
    - El balance entre precisión y recall para alumnos con rendimiento bajo muestra un rendimiento excelente.
    - El balance entre precisión y recall para alumnos con rendimiento medio muestra un rendimiento adecuado.
- Support:
    - La muestra cuenta con 154 alumnos con rendimiento alto.
    - La muestra cuenta con 2492 alumnos con rendimiento bajo.
    - La muestra cuenta con 1554 alumnos con rendimiento medio.
- Accuracy:
    - La precisión total del modelo es del 92%
- Weighted avg:
    - El promedio ponderado de las métricas es de 0.92

Finalmente, se puede observar la importancia de las características. La característica que más influye es la nota promedio anterior, seguido de las horas de estudio y la asistencia. Por otro lado, prácticamente el resto de características tienen un impacto nulo.

Comparación de los resultados.

Casi todos los modelos entrenados presentan una predictibilidad excepcional, siendo el KNN el que menor precisión ha logrado. Basándonos en la matriz de confusión el modelo más acertado es el árbol de decisiones, con fallos ínfimos y una media ponderada de las métricas de 0.97.