

PREDICCIÓN DE RENDIMIENTO EN MATEMÁTICAS

PROBLEMÁTICA

El rendimiento en matemáticas es un indicador clave del desempeño académico y puede afectar el futuro educativo y laboral de los estudiantes.

En la actualidad, los docentes o coordinadores identifican estudiantes en riesgo principalmente con base en observación subjetiva o análisis tardío de calificaciones. Esto puede llevar a intervenciones poco oportunas o ineficaces en el proceso y formación académica de los estudiantes.

MÉTODO

- El conjunto de datos usado es “Students Performance in Exams” con 1.000 registros, que incluye variables demográficas y académicas de estudiantes.
- Se codificaron las variables categóricas y se escalaron los datos para preparar el dataset, haciendo el respectivo EDA previo.
- Se probó el modelo con modelos base como:
 - KNN
 - SVM
 - Regresión Logística
- Finalmente, se pasó al uso de modelos avanzados y optimización:
 - XGBoost
 - Stacking (técnica ensemble)
 - AutoML con FLAML

MÉTRICA DE ÉXITO

La métrica técnica clave fue alcanzar un $\text{Recall} \geq 0.85$ (85%). Esto significa que el modelo debe ser capaz de identificar correctamente al menos el 85% de los estudiantes que, de otra forma, suspenderían el examen de matemáticas (clase negativa).

ESCOGER EL MODELO

La escogencia de un modelo y otro dependerá del usuario final del modelo. ¿Qué priorizo?

- Alertarme de gente que no necesariamente va a reprobar.

Classification report:				
	precision	recall	f1-score	support
0	0.595	0.338	0.431	65
1	0.736	0.889	0.805	135
accuracy				
macro avg	0.665	0.614	0.618	200
weighted avg	0.690	0.710	0.684	200
Matriz de Confusión				
	Pred: No Reprobado	Pred: Reprobado		
Real: No Reprobado	22	43		
Real: Reprobado	15	120		

- Balancear los aciertos y que posiblemente se escapen algunos que reproben.

--- Reporte de Clasificación del Modelo Stacking ---				
Test Accuracy: 0.715				
Test F1 : 0.8106312292358804				
	precision	recall	f1-score	support
0	0.62	0.32	0.42	65
1	0.73	0.90	0.81	135
accuracy				
macro avg	0.68	0.61	0.62	200
weighted avg	0.70	0.71	0.69	200

RESULTADOS CLAVE

- El modelo óptimo encontrado fue XGBoost, el cual destacó frente a los demás candidatos tanto en desempeño como en estabilidad.
- La escasez de datos incrementó la sensibilidad de los modelos a pequeñas variaciones y probablemente contribuyó al sesgo observado.
- El mantenimiento de un conjunto amplio de características fue necesario para compensar la falta de volumen de datos.