HE2: IA

Parcial 1

Problema a resolver:

Predecir la probabilidad de que un estudiante apruebe o repruebe la materia teniendo en cuenta que la nota mínima aprobatoria es 10. Este tipo de predicciones resulta útil para las instituciones educativas, ya que permite identificar tempranamente a los estudiantes en riesgo de reprobar y diseñar estrategias de acompañamiento. Asimismo, tiene aplicaciones en políticas públicas orientadas a reducir la deserción escolar y mejorar los niveles de rendimiento académico, lo cual lo hace un problema de interés social y económico.

Base de datos:

La base de datos se obtuvo mediante Hugging Face, esta contiene 1044 registros y 33 características acerca de diferentes características y situaciones de estudiantes de primaria y secundaria. Tales como la edad, género, dirección, actividades, tiempo de estudio, entre otras. Además, se escogió 10 como la nota mínima aprobatoria, ya que, en Portugal (de donde viene este dataset de *student alcohol consumption*), el sistema escolar suele usar: 10/20 como nota mínima aprobatoria. Es decir, cualquier calificación 10 se considera reprobada.

Escala de Classificação Portuguesa | DGES

Modelos:

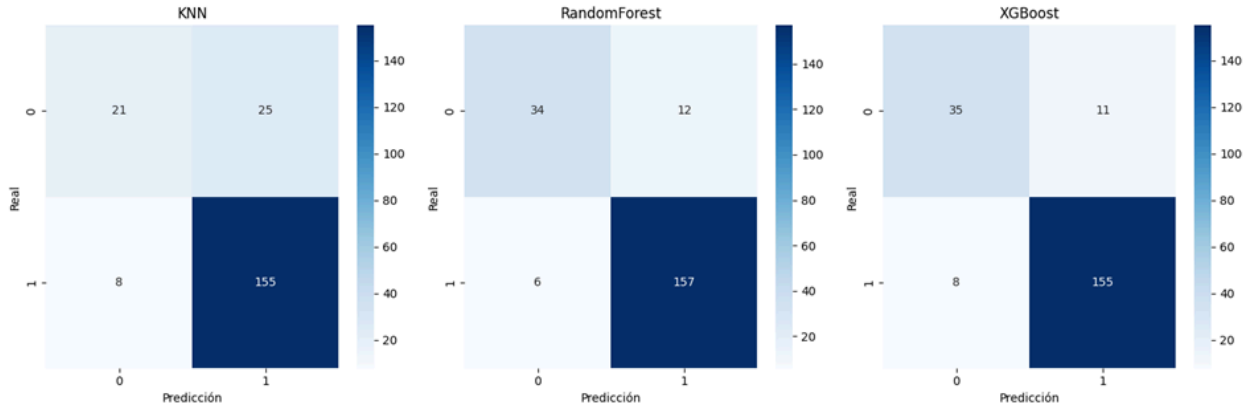
- **KNN**: dado que este modelo nos ayuda a poder predecir si un estudiante aprobará la materia si la mayoría de los estudiantes parecidos a él aprueban (en factores como hábitos, condiciones socioeconómicas, tiempo de estudio, consumo de alcohol). Además de identificar perfiles de estudiantes similares
- **Random Forest:** Permite identificar la importancia de las variables, como saber qué factores (consumo de alcohol, nivel socioeconómico, apoyo familiar) impactan más en que un estudiante apruebe o repruebe una materia. Esto es posible dado que este modelo puede trabajar con datos heterogéneos.
- **XGboost**: Ayuda a identificar factores que influyen pero con mejor detalle, es más preciso en las predicciones ya que corrige los errores de los árboles anteriores. Entonces es excelente para predecir con mayor precisión la probabilidad de aprobar o reprobar. Permite simular escenarios (ej. si mejora el tiempo de estudio, ¿cómo cambia la probabilidad de aprobar?).

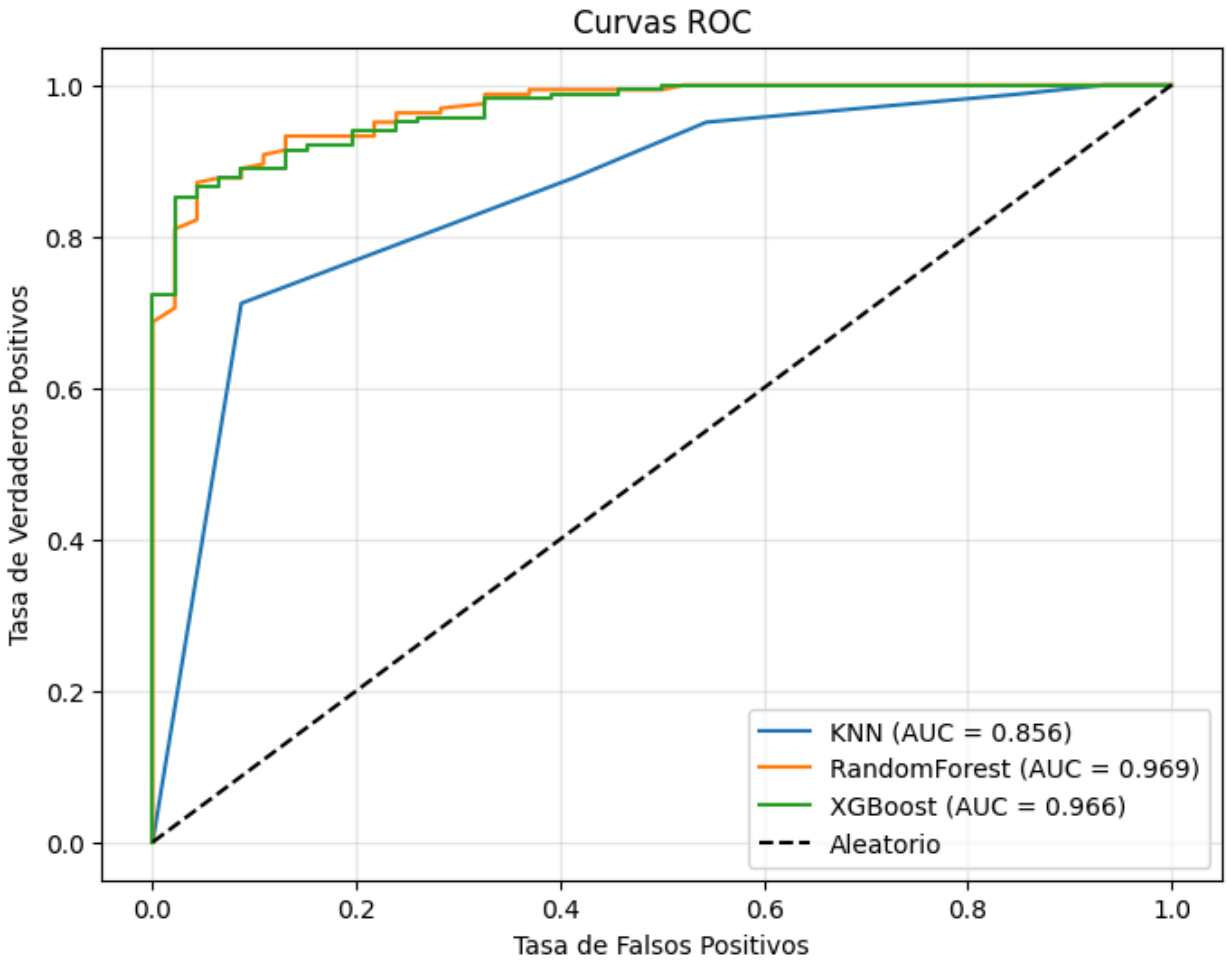
Resultados obtenidos:

Para encontrar los mejores hiperparámetros de cada modelo se aplicó un proceso de búsqueda con validación cruzada. En este procedimiento, se probaron diferentes combinaciones de valores posibles para cada hiperparámetro y se evaluó el desempeño de los modelos utilizando la métrica F1. De esta manera, fue posible identificar la configuración que maximizaba el rendimiento de cada algoritmo y evitar el riesgo de sobreajuste asociado a una sola partición de datos.

Resultados de la búsqueda de hiperparámetros:

- **KNN**: n_neighbors=7, weights=distance, metric=euclidean → F1 = 0.908
- **Random Forest**: n_estimators=200, max_depth=None, min_samples_leaf=1, max_features=sqrt → F1 = 0.952
- **XGBoost**: n_estimators=200, learning_rate=0.1, max_depth=10, subsample=0.8, colsample_bytree=0.8 → F1 = 0.954





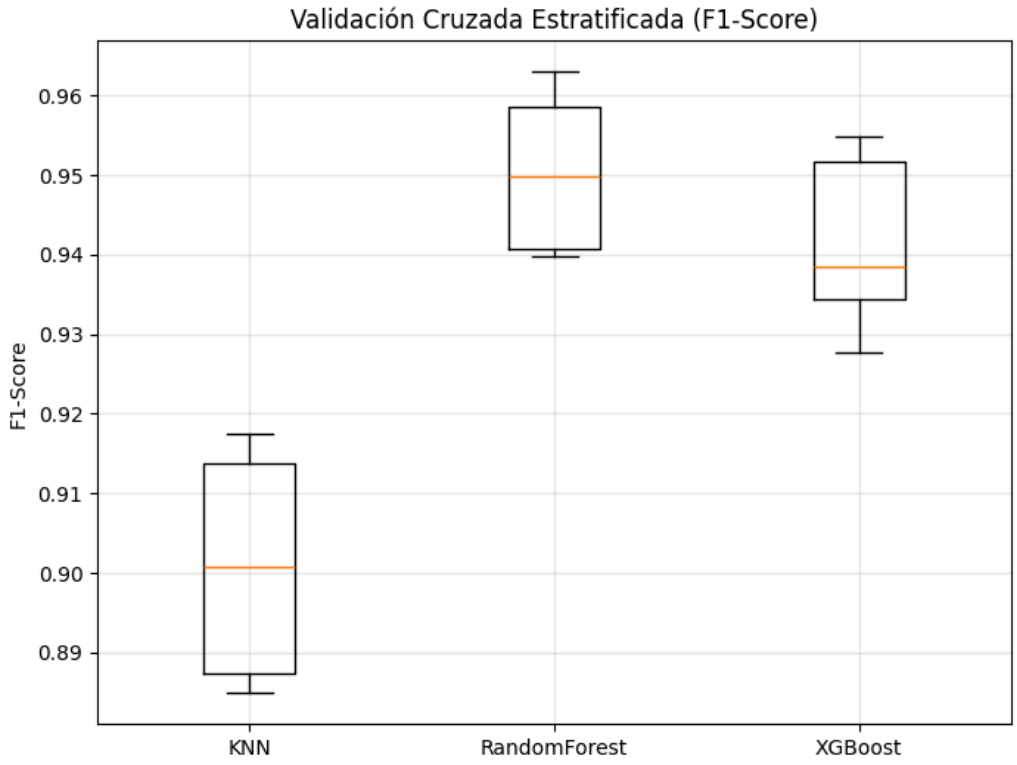
- Se evidencia que todos los modelos tienen un buen rendimiento, sin embargo, KNN con tiene el peor rendimiento de los tres modelos, ya que, detecta bien a los estudiantes que aprueban/reprueban, pero comete más falsos positivos. En cambio, sobresale Randomforest, siendo este modelo el de mejor rendimiento de los tres, lo que implica que distingue con mucha precisión entre aprobar y reprobar.

Modelo	Accuracy	Precision	Recall	F1	AUC
KNN	0.842	0.861	0.951	0.904	0.856
RandomForest	0.914	0.929	0.963	0.946	0.969
XGBoost	0.904	0.928	0.951	0.939	0.966

Los tres modelos muestran buenos resultados : las accuracies son superiores a 0.84.

- **KNN**: El recall es muy alto lo que significa que el modelo detecta casi todos los casos positivos pero tiene menor precisión y AUC, lo que indica que genera más FP

- **XGBoost**: El modelo es confiable y robusto porque su AUC es muy alto lo que demuestra una excelente capacidad de distinguir entre clases. El recall y la precisión presentan un buen equilibrio.
- **RandomForest**: El modelo minimiza los FP y discrimina con mayor eficiencia entre las clases. Tiene la mejor precisión y el mejor AUC



Random Forest alcanza los valores de F1 más altos en la validación cruzada y muestra una gran consistencia entre los distintos folds. KNN, en cambio, presenta el peor desempeño: sus F1-scores son claramente más bajos y con mayor variabilidad, lo que indica que depende más de la partición de datos que le toque. Por su parte, XGBoost también exhibe un rendimiento muy competitivo, con resultados cercanos a los de Random Forest, aunque con una ligera mayor dispersión en los folds.

Una limitación importante es que el dataset corresponde a estudiantes de Portugal, por lo cual los resultados pueden no ser totalmente generalizables a otros contextos educativos. Además, aunque Random Forest y XGBoost ofrecen gran capacidad predictiva, también pueden ser más susceptibles al sobreajuste si no se regulan adecuadamente los hiperparámetros.

En términos prácticos, la identificación de variables relevantes como el tiempo de estudio o el apoyo familiar sugiere que las instituciones educativas podrían enfocar esfuerzos en reforzar hábitos de estudio y acompañamiento parental para aumentar la probabilidad de éxito académico.

Conclusión:
Aunque RandomForest obtuvo el mejor rendimiento, en un escenario real la elección entre XGBoost y Random Forest dependería de los recursos disponibles: XGBoost ofrece un buen rendimiento pero con mayor costo computacional, mientras que Random Forest es más estable y menos demandante, logrando resultados casi igual de sobresalientes.