Sofía Hinojosa
Julia Hernández
Sara Hernández

PROYECTO 3

CLASIFICACIÓN DE SOLICITUDES CREDITICIAS

Objetivo general

Desarrollar un modelo de clasificación supervisada que prediga la aprobación de solicitudes crediticias, optimizando sus hiperparámetros mediante optimización bayesiana, para maximizar la discriminación entre aprobaciones y rechazos y obtener un alto desempeño general (alto AUC-ROC), de modo que el modelo pueda servir como una herramienta de apoyo a decisiones de crédito confiable y reproducible.

Objetivos específicos

01 Preprocesar los datos eliminando columnas irrelevantes, convirtiendo variables categóricas en dummies y estandarizando las variables numéricas para que sean comparables.

02 Entrenar y validar un modelo Random Forest mediante validación cruzada de K-Folds, evaluando su desempeño en la predicción de aprobación crediticia, y analizar su capacidad discriminativa utilizando métricas como ROC AUC y la matriz de confusión.

03 Desarrollar un modelo XGBoost para la clasificación de solicitudes crediticias, aplicando validación cruzada de K-Folds para medir su estabilidad y compararlo con Random Forest en términos de capacidad predictiva, ROC AUC y matriz de confusión.

04 Aplicar técnicas de optimización automatizada bayesiana mediante Optuna para identificar la configuración óptima de hiperparámetros en Random Forest y XGBoost, maximizando el desempeño del modelo según la métrica ROC AUC.

05 Comparar el desempeño de los modelos entrenados, Random Forest y XGBoost, tanto en su versión base como optimizada, utilizando métricas como ROC AUC, curvas ROC, matrices de confusión y medidas de precisión. Identificar el modelo con mayor capacidad para discriminar entre solicitudes aprobadas y no aprobadas.

Análisis del dataset

El dataset fue creado con criterios reales de bancos de Estados Unidos y Canadá. Basado en 3 años de experiencia práctica en la industria financiera, el dataset incorpora correlaciones realistas y lógica de negocio que reflejan cómo se toman decisiones de crédito en la vida real.

Variables

- Resume características del solicitante y del préstamo:
  - Edad, experiencia laboral y tipo de ocupación.
  - Nivel de ingresos, ahorros y deudas actuales.
  - Historial crediticio y registros de morosidad.
  - Tipo e intención del préstamo.
  - Resultado final del crédito (loan_status).

Modelos propuestos

Random Forest

Random Forest es un método de ensamble que combina muchos árboles de decisión entrenados con diferentes muestras. La predicción final se obtiene por votación, reduciendo la varianza y mejorando la generalización. Puede capturar relaciones no lineales, manejar datos ruidosos o de alta dimensionalidad y funciona bien sin necesidad de un ajuste complejo de hiperparámetros.

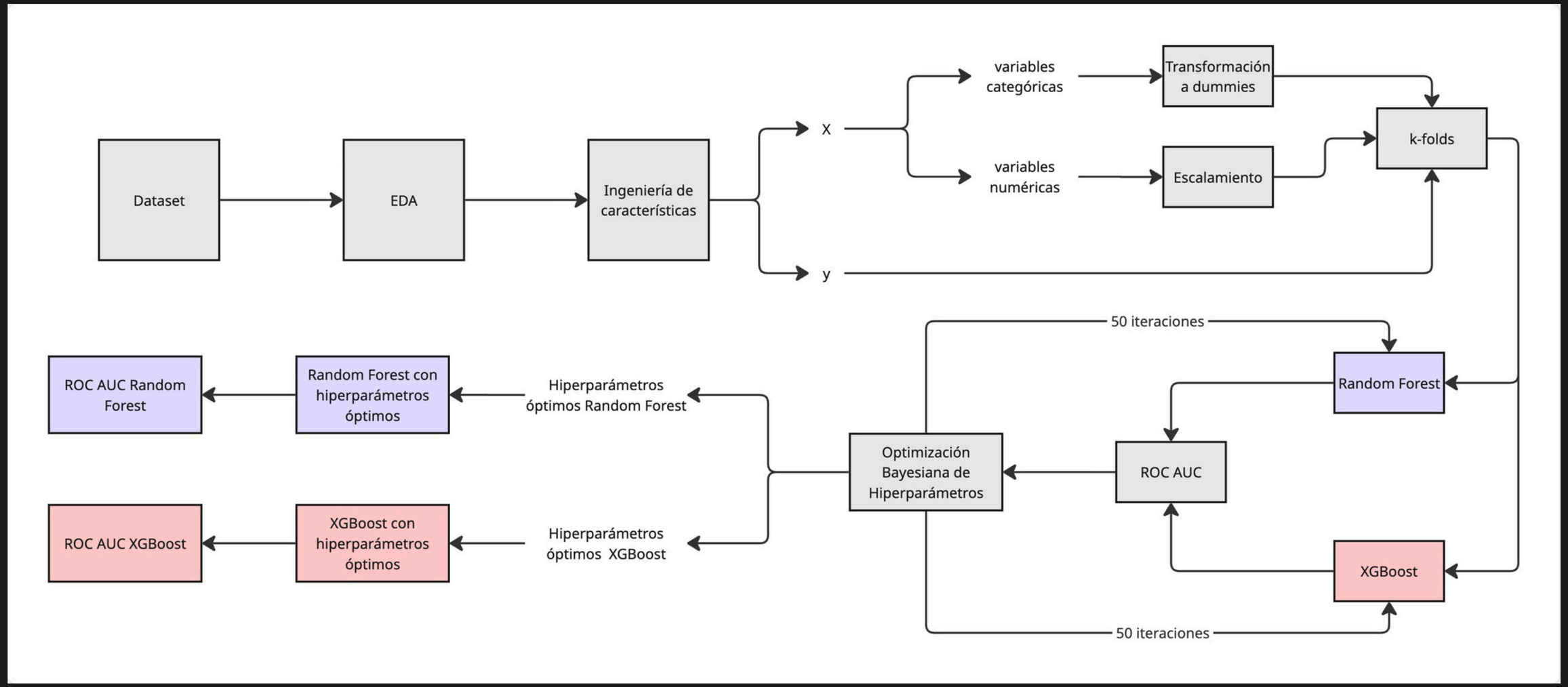
XGBoost

XGBoost es un algoritmo de gradient boosting que construye árboles de manera secuencial, corrigiendo los errores de los modelos anteriores. Este proceso le permite capturar patrones complejos con alta precisión, incorporando regularización y optimizaciones que reducen el sobreajuste. Destaca por su gran capacidad predictiva y su excelente rendimiento en datos estructurados.

Uso de ROC AUC como métrico principal

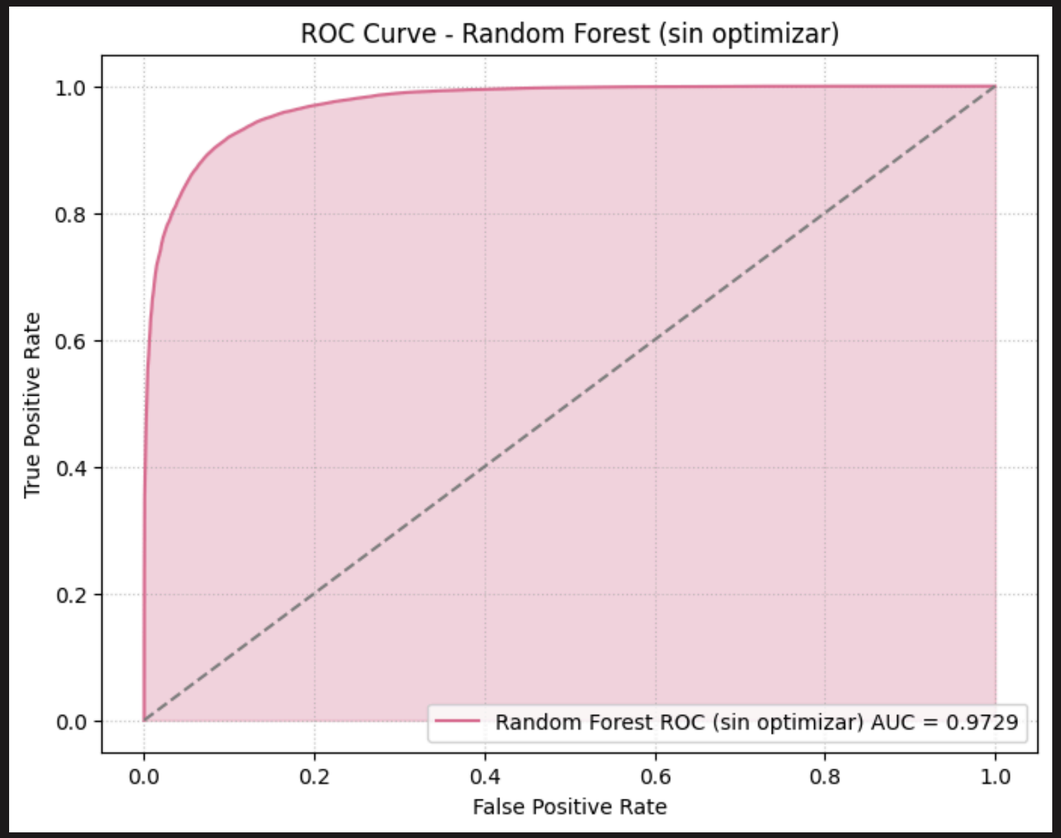
ROC AUC se eligió porque mide la capacidad del modelo para distinguir entre aprobaciones y rechazos sin depender de un umbral. Es robusto ante desbalance de clases, considera todas las combinaciones de sensibilidad y falsos positivos, y ofrece una evaluación estable y comparable entre modelos, ideal para contextos financieros y de riesgo.

Pipeline

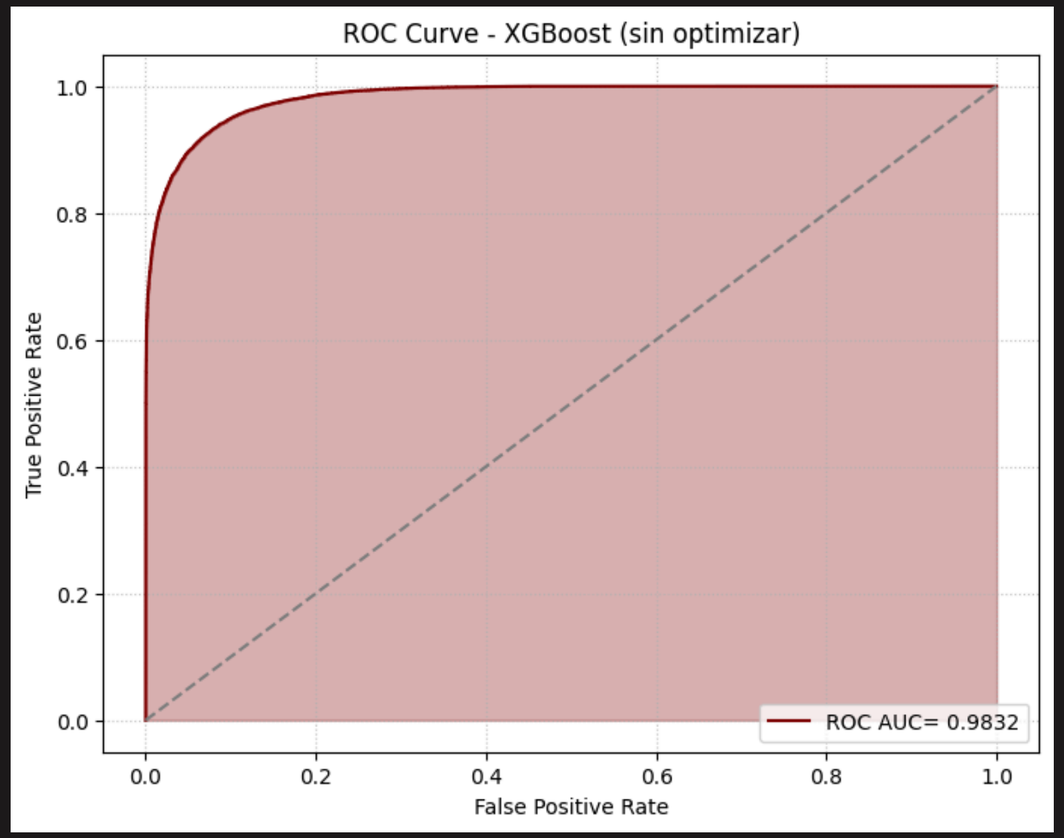

Dataset
EDA
Ingeniería de características
X
y
variables categóricas
variables numéricas
Transformación a dummies
Escalamiento
k-folds
50 iteraciones
Random Forest
ROC AUC
XGBoost
50 iteraciones
Optimización Bayesiana de Hiperparámetros
Hiperparámetros óptimos Random Forest
Hiperparámetros óptimos XGBoost
Random Forest con hiperparámetros óptimos
XGBoost con hiperparámetros óptimos
ROC AUC Random Forest
ROC AUC XGBoost

Resultados de los modelos antes de optimización de hiperparámetros

ROC AUC

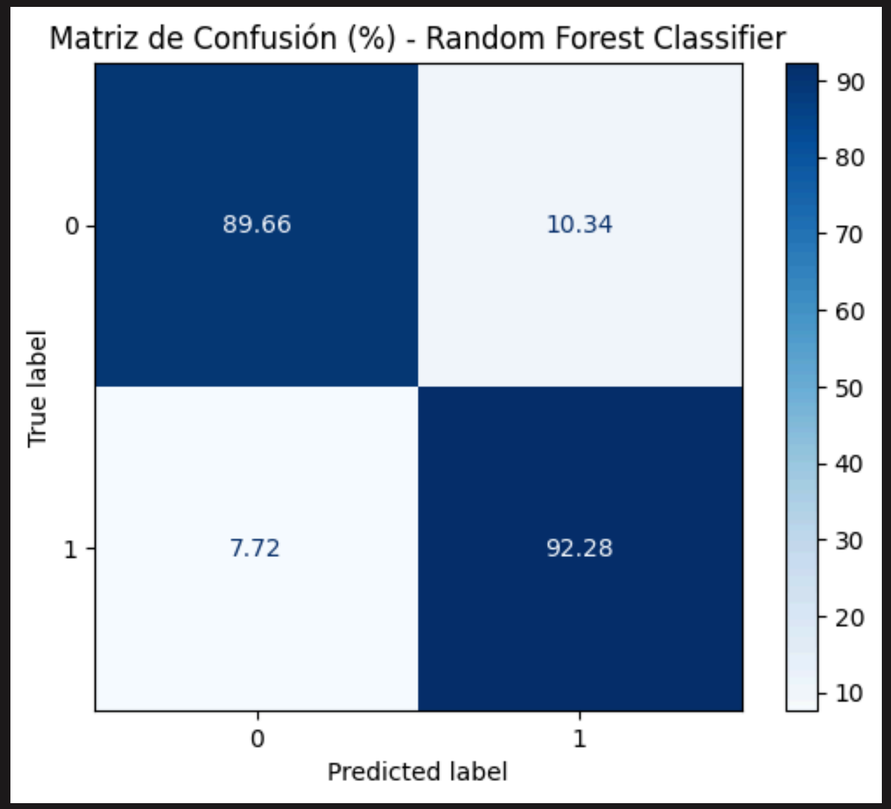


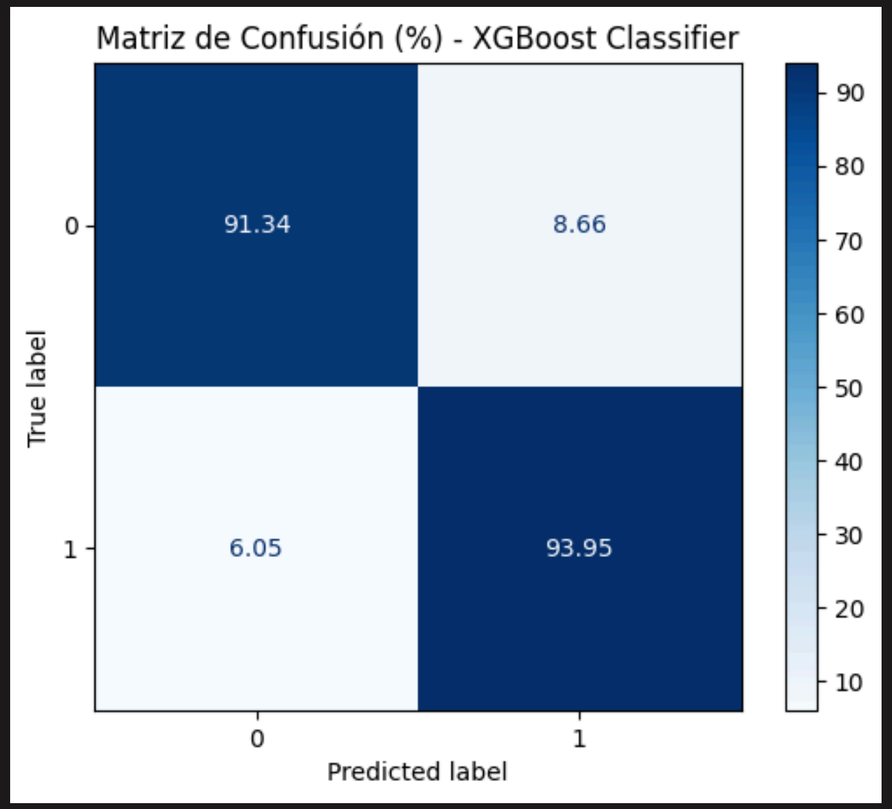
ROC AUC PROMEDIO: 0.9729
DESVIACIÓN ESTÁNDAR: 0.0011



ROC AUC PROMEDIO: 0.9832
DESVIACIÓN ESTÁNDAR: 0.0007

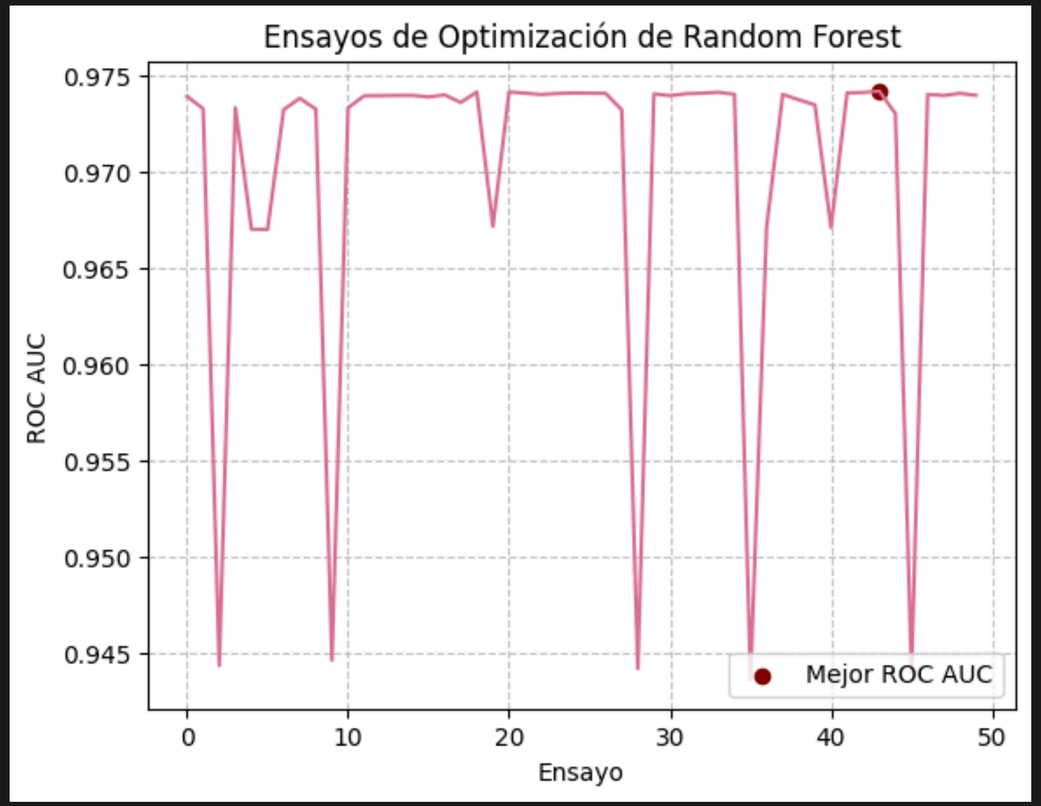
Matrices de confusión

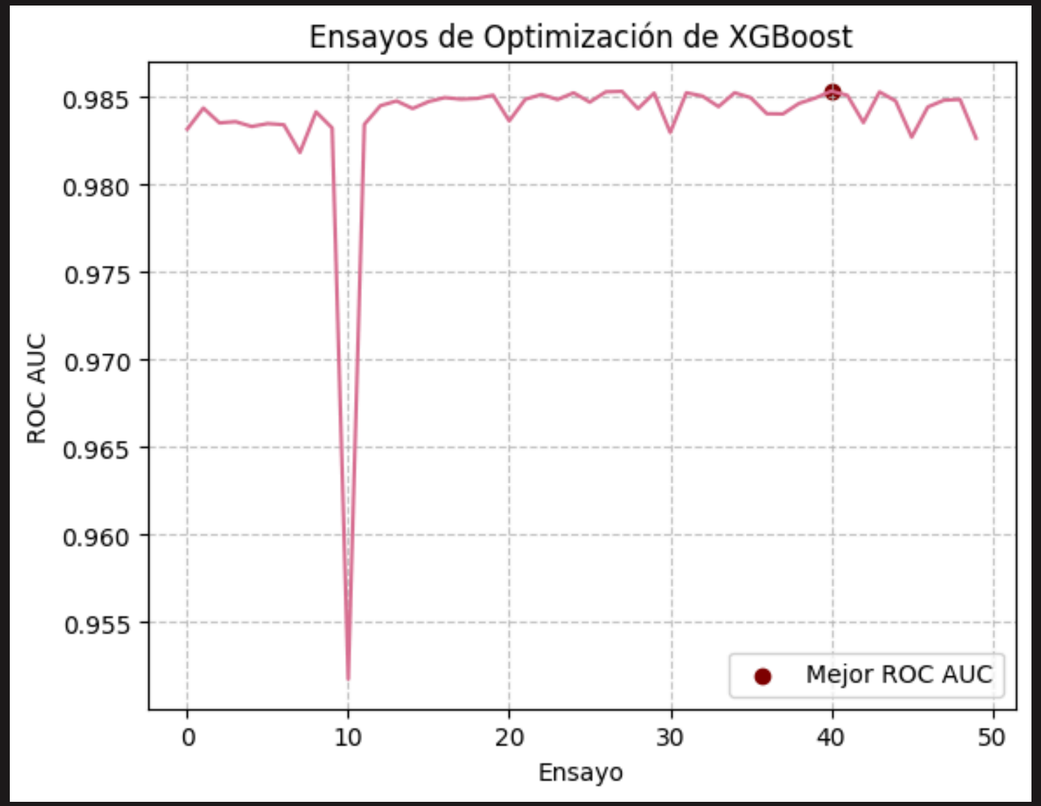




Resultados de los modelos después de optimización de hiperparámetros

Optimización Bayesiana

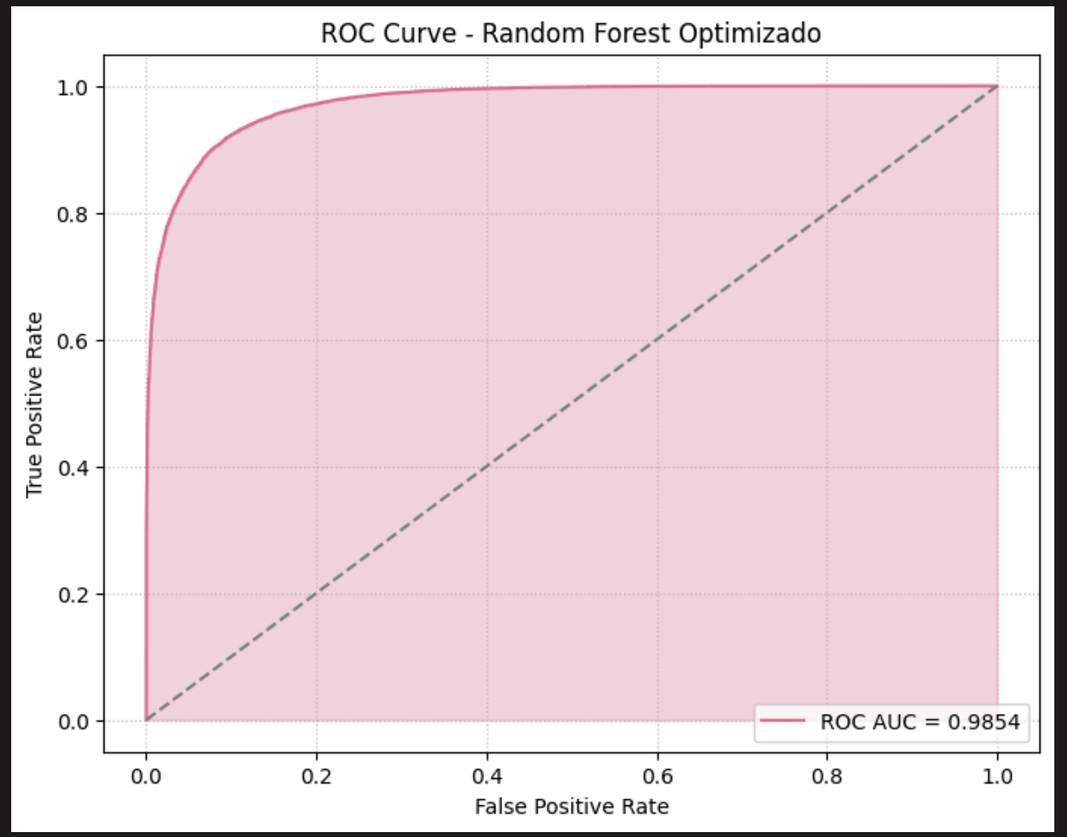




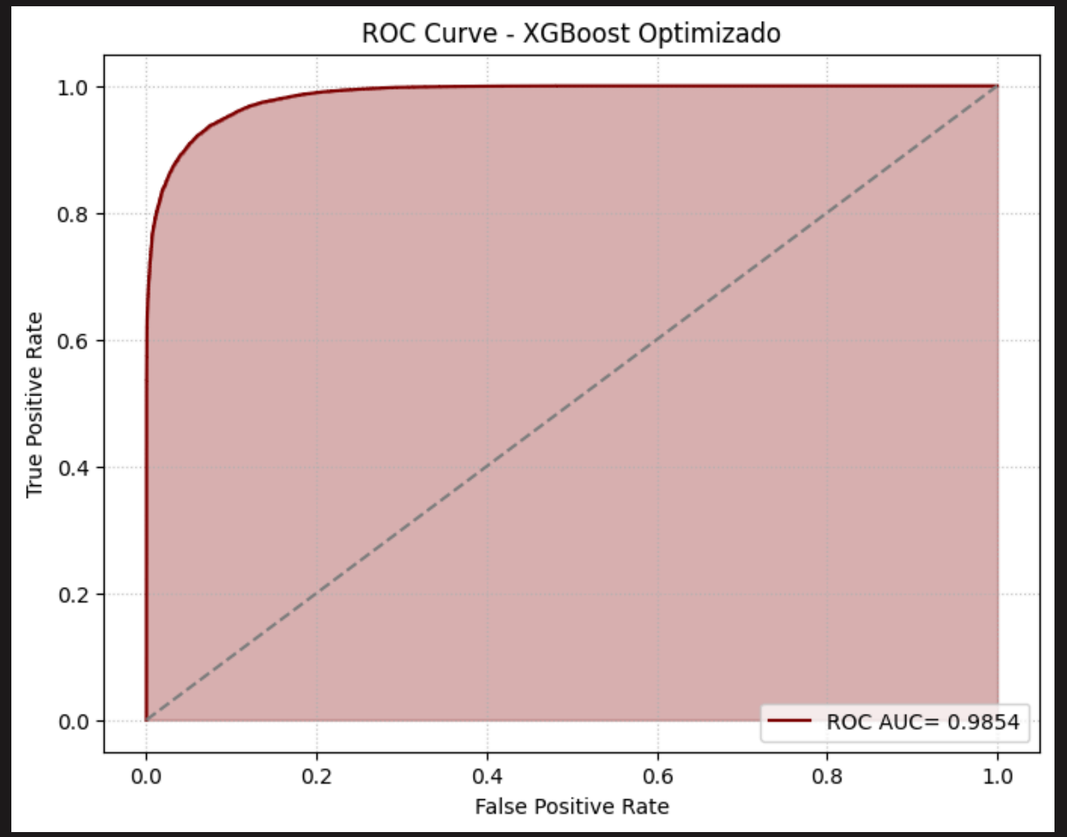
	Random Forest
n_estimators	669
max_depth	30
min_samples_split	3
min_samples_leaf	1

	XGBoost
n_estimators	269.000000
max_depth	3.000000
gamma	1.770670
learning_rate	0.186887

ROC AUC

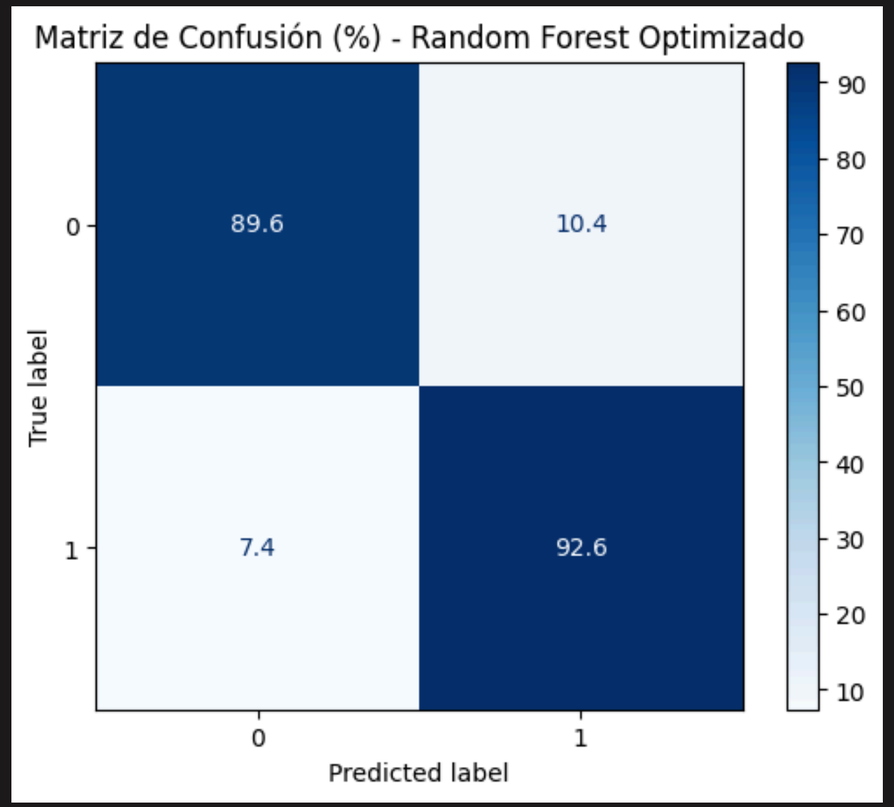


ROC AUC PROMEDIO: 0.9742
DESVIACIÓN ESTÁNDAR: 0.0014



ROC AUC PROMEDIO: 0.9853
DESVIACIÓN ESTÁNDAR: 0.0004

Matrices de confusión



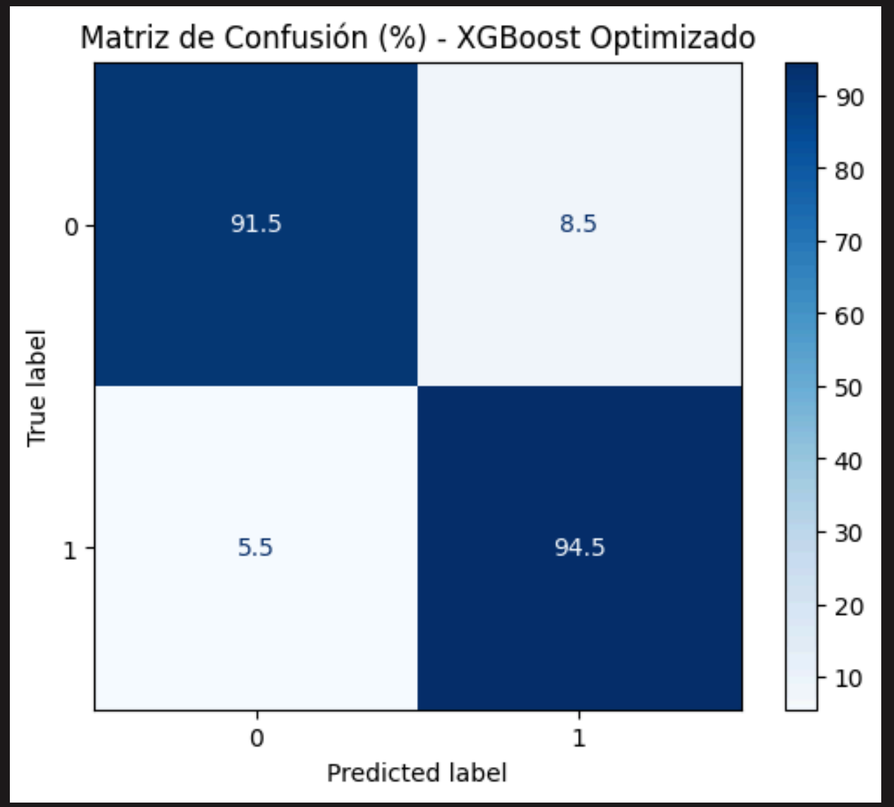


Tabla comparativa

Modelo utilizado	Promedio de ROC AUC	Desviación estándar de ROC AUC
Random Forest sin optimización	0.9729	0.0011
Random Forest optimizado	0.9742	0.0014
XGBoost sin optimización	0.9832	0.0007
XGBoost optimizado	0.9853	0.0004

Conclusiones

- El proyecto mostró que ambos modelos fueron efectivos para predecir aprobaciones crediticias, aunque XGBoost superó ligeramente a Random Forest en ambos casos. Tras el preprocesamiento y la optimización con Optuna, XGBoost optimizado obtuvo el mejor desempeño (mayor ROC AUC y mejor separación entre clases). Los resultados confirman que la combinación de buenos datos y ajuste de hiperparámetros mejora significativamente la capacidad predictiva.

- Se usó ROC AUC ya que mide qué tan bien el modelo diferencia aprobadas vs. rechazadas sin depender de un umbral y sin verse afectado por el desbalance. Además, en este proyecto ayuda a ver de forma clara y confiable qué tan bien está clasificando realmente el modelo.

GRACIAS

https://github.com/AnaSofiaHinojosa/Proyecto3LabAprEstadistico