

PROYECTO 2 - CLASIFICACIÓN

CLASIFICACIÓN DE TRANSACCIONES FINANCIERAS
COMO FRAUDULENTES O LEGÍTIMAS

JULIA HERNÁNDEZ
ANA SOFIA HINOJOSA
SARA HERNÁNDEZ

DATA SET

- Data set de la Plataforma Kaggle:
- Contiene Variables tanto numéricas como categóricas:
 - Age
 - Merchant Group
 - Type of Card
 - Bank
 - Gender
 - Country of Transaction
 - Entry Mode
 - Type of Transaction
 - Day of Week
 - Fraud

Donde Cada registro representa una transacción individual.

DATASET

```
3 require File.expand_path('../config/environment', __FILE__)
4 # Prevent database truncation if the transaction fails
5 abort("The Rails environment is running in production mode!
6 require 'spec_helper'
7 require 'rspec/rails'
8
9 require 'capybara/rspec'
10 require 'capybara/rails'
11
12 Capybara.javascript_driver = :webkit
13 Category.delete_all; Category.create!(name: "Electronics")
14 Shoulda::Matchers.configure do |config|
15   config.integrate do |with|
16     with.test_framework :rspec
17     with.library :rails
18   end
19
20   # Add additional requires below this line
21
22   # Requires supporting files within the same directory as this file or,
23   # if further up the tree, look for the supporting files.
24   # spec/support/ and its subdirectories
25   # run as spec files by default. This means you can run specs
26   # in _spec.rb will both be required
27   # run twice. It is recommended that you do not name
28   # end with _spec.rb. You can run specs in
29   # nation on the command line (e.g.
30   # mongoid
31   # buffer
32
33 No results found for 'mongoid'
```

- **Preprocesamiento y transformaciones**

- Variables Categóricas a dummies (Merchant Group, Type of Card, Bank, Gender, Country of Transaction, Entry Mode, Type of Transaction)
- “Day of Week” se transforma a valores numéricos (0 = Monday, 6 = Sunday)
- Variables numéricas (Amount, Age) escaladas, para garantizar la compatibilidad con los modelos de aprendizaje supervisado, usando StandardScaler
- Eliminación de columnas irrelevantes (Transaction ID, Date, Shipping Address, Country of Residence).

GENERALES

Desarrollar un modelo predictivo que clasifique transacciones financieras como fraudulentas o legítimas, a través de métricas (ROC AUC), utilizando modelos de aprendizaje supervisado y optimización de hiperparámetros.

OBJETIVOS

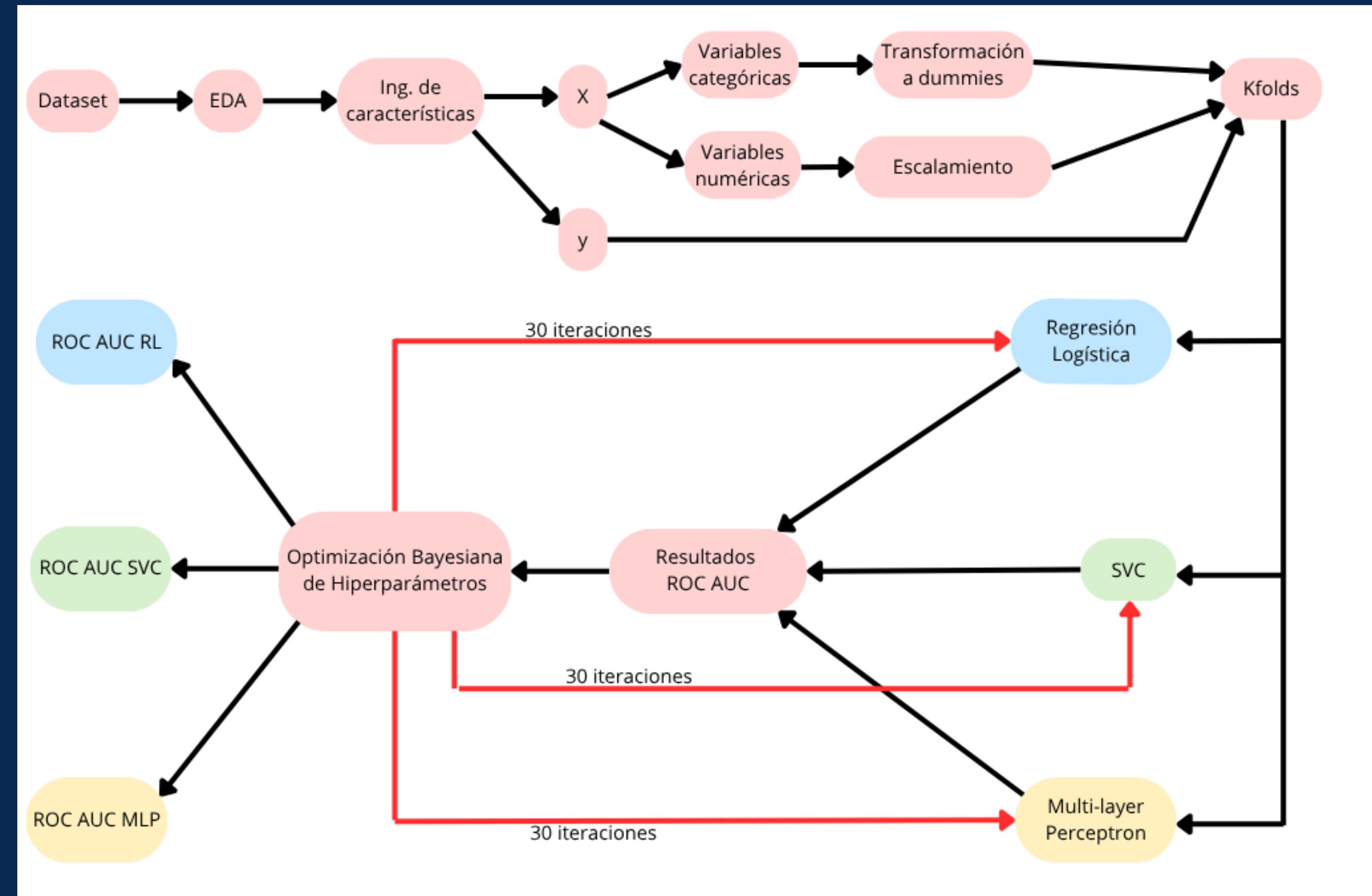


ESPECIFICOS

- Preprocesar los datos convirtiendo variables categóricas en dummies y estandarizar las variables numéricas para que sean compatibles.
- Entrenar un modelo de Regresión Logística, SVM y MLP para evaluar y analizar sus desempeños en distintas particiones del dataset, midiendo su capacidad de predicción de fraude.
- Evaluar métricas como accuracy y ROC AUC para cada modelo.
- Aplicar Optimización Bayesiana para encontrar la configuración óptima de los hiperparámetros de cada modelo (C para Regresión Logística y SVM, tamaño de capas ocultas para MLP) que maximice la ROC AUC.
- Comparar los modelos entrenados con las métricas, identificar cuál tiene mejor capacidad para detectar fraude.

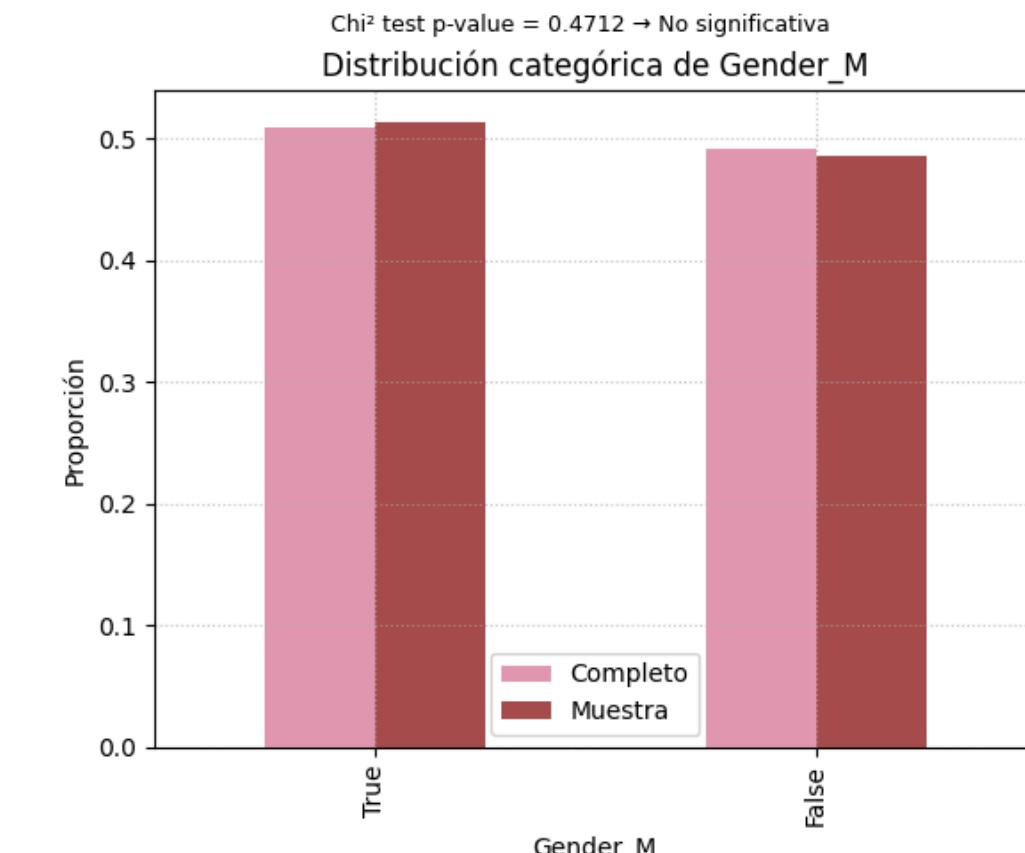
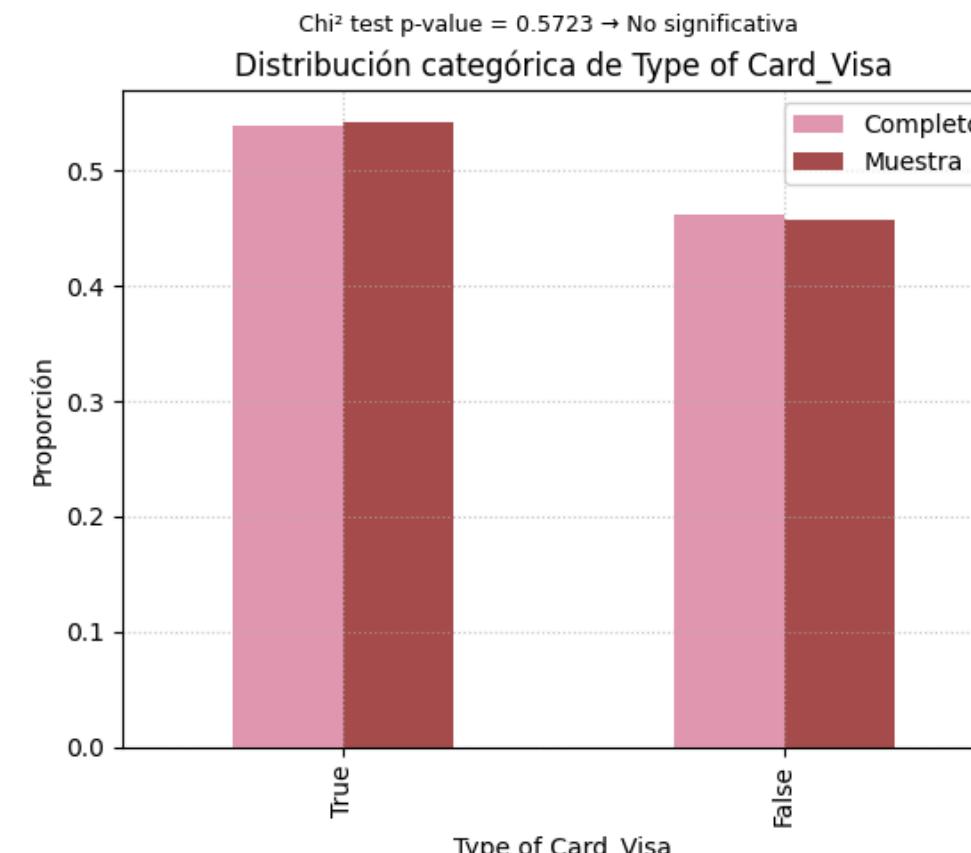
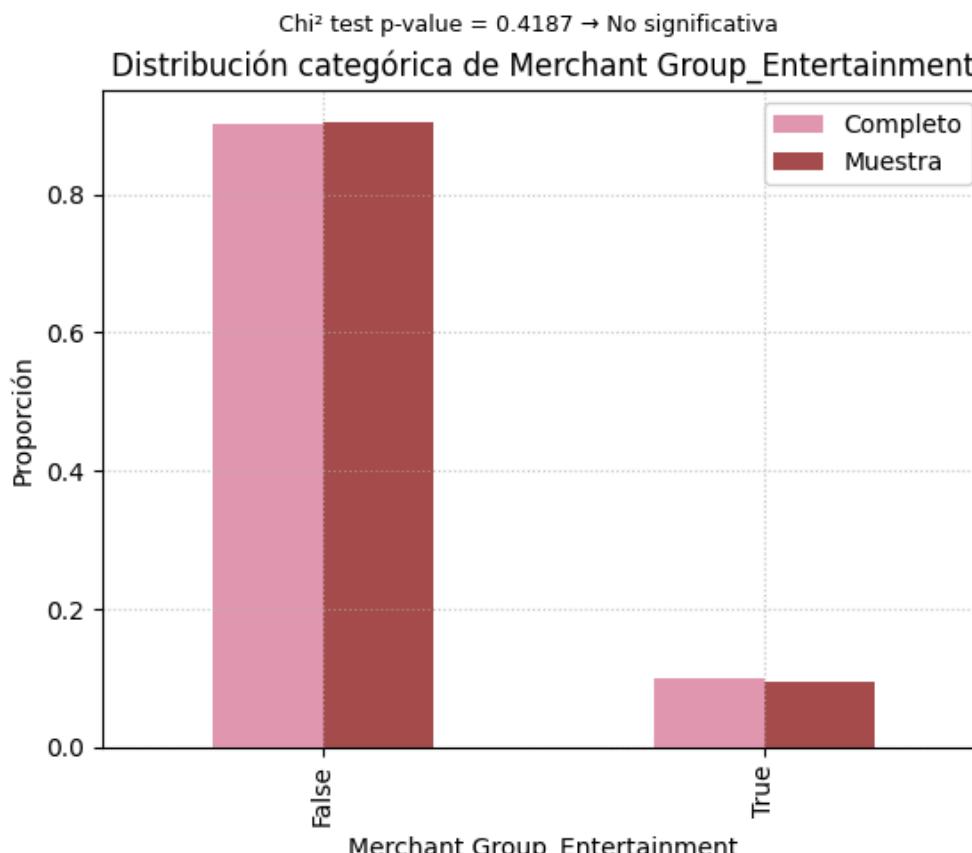
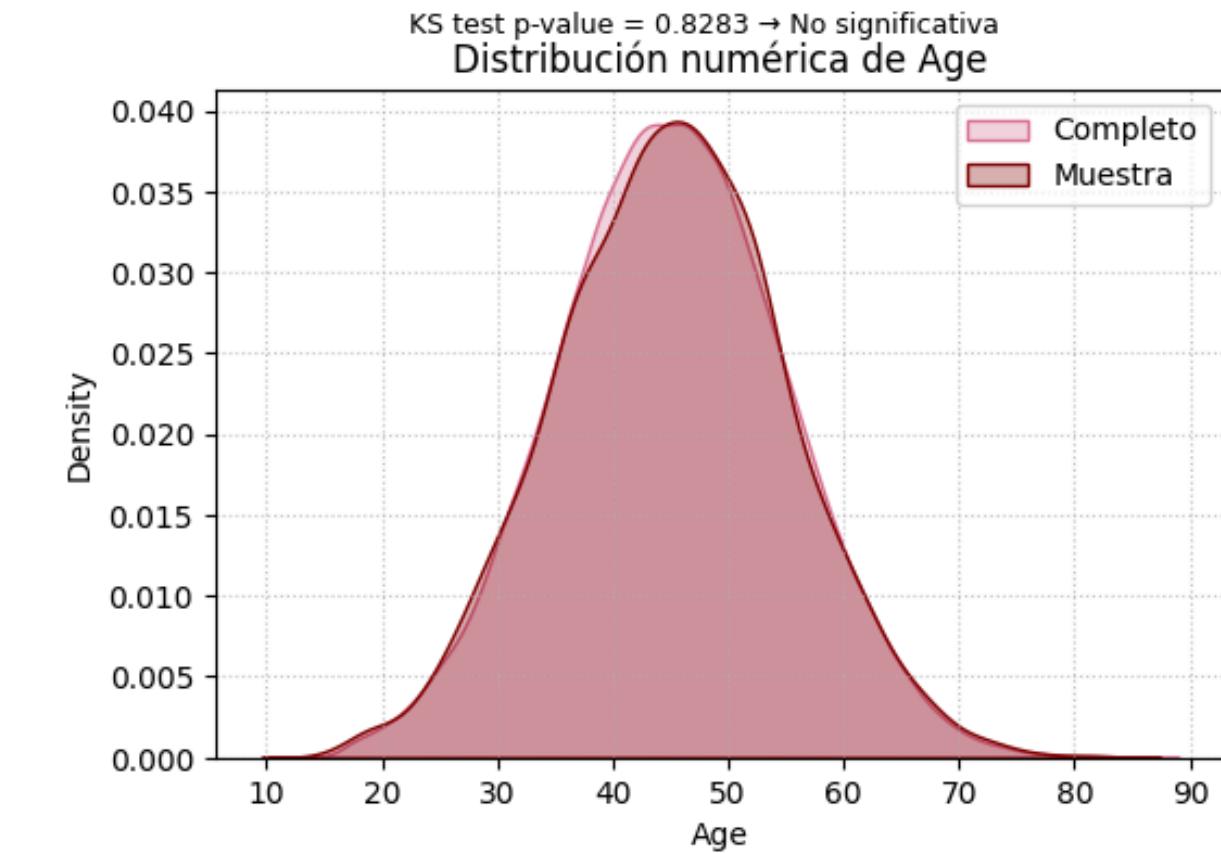
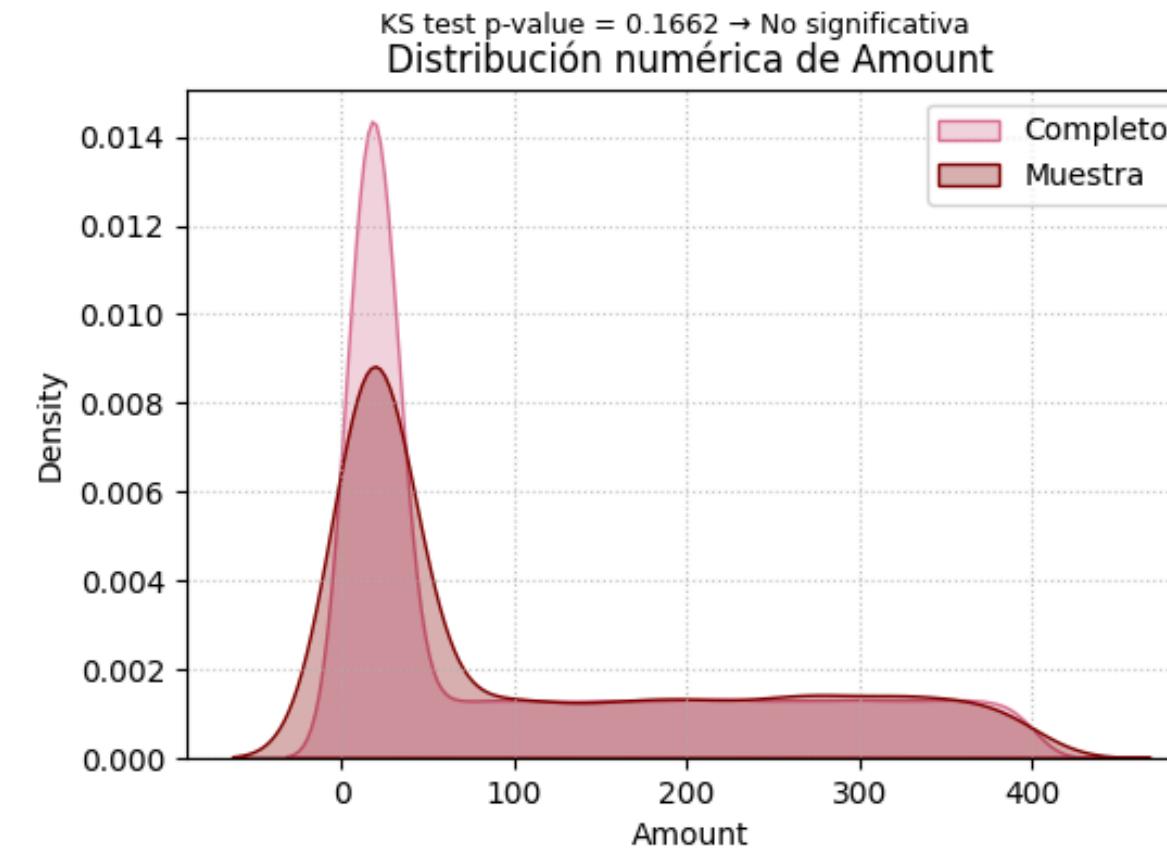
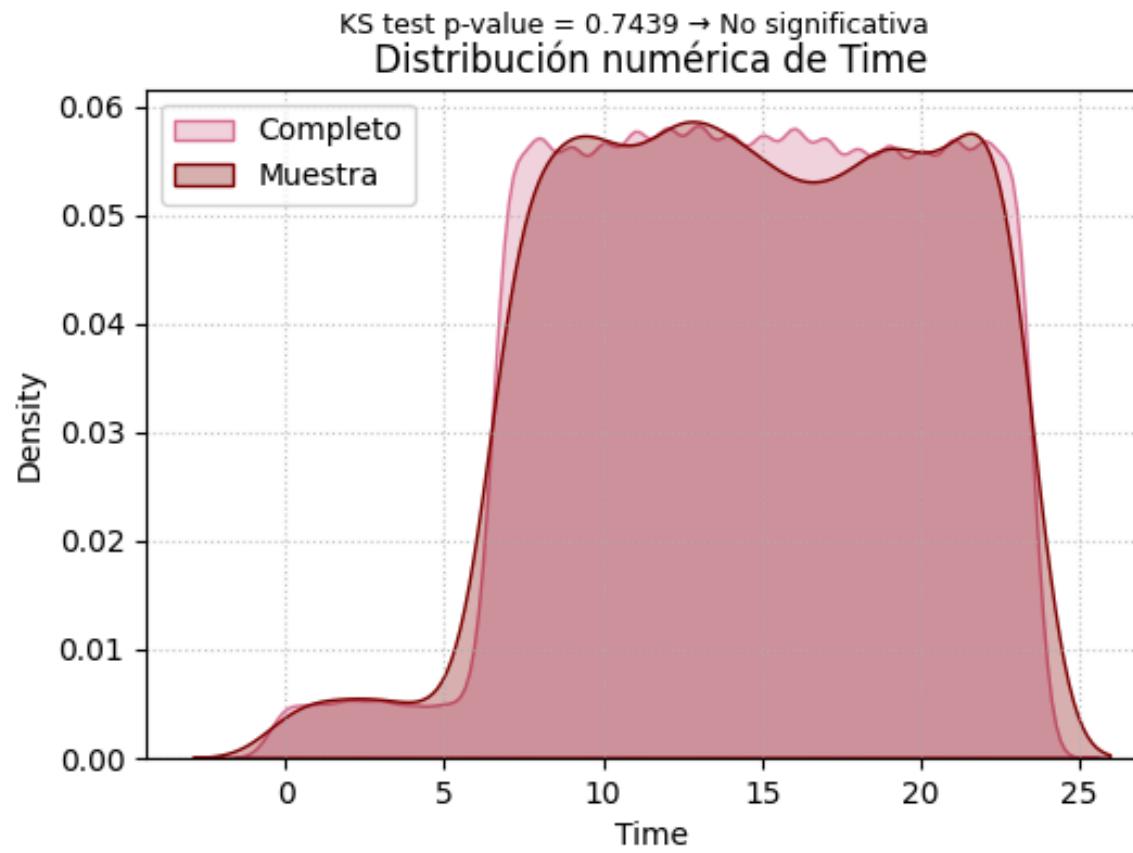
OBJETIVOS

PIPELINE

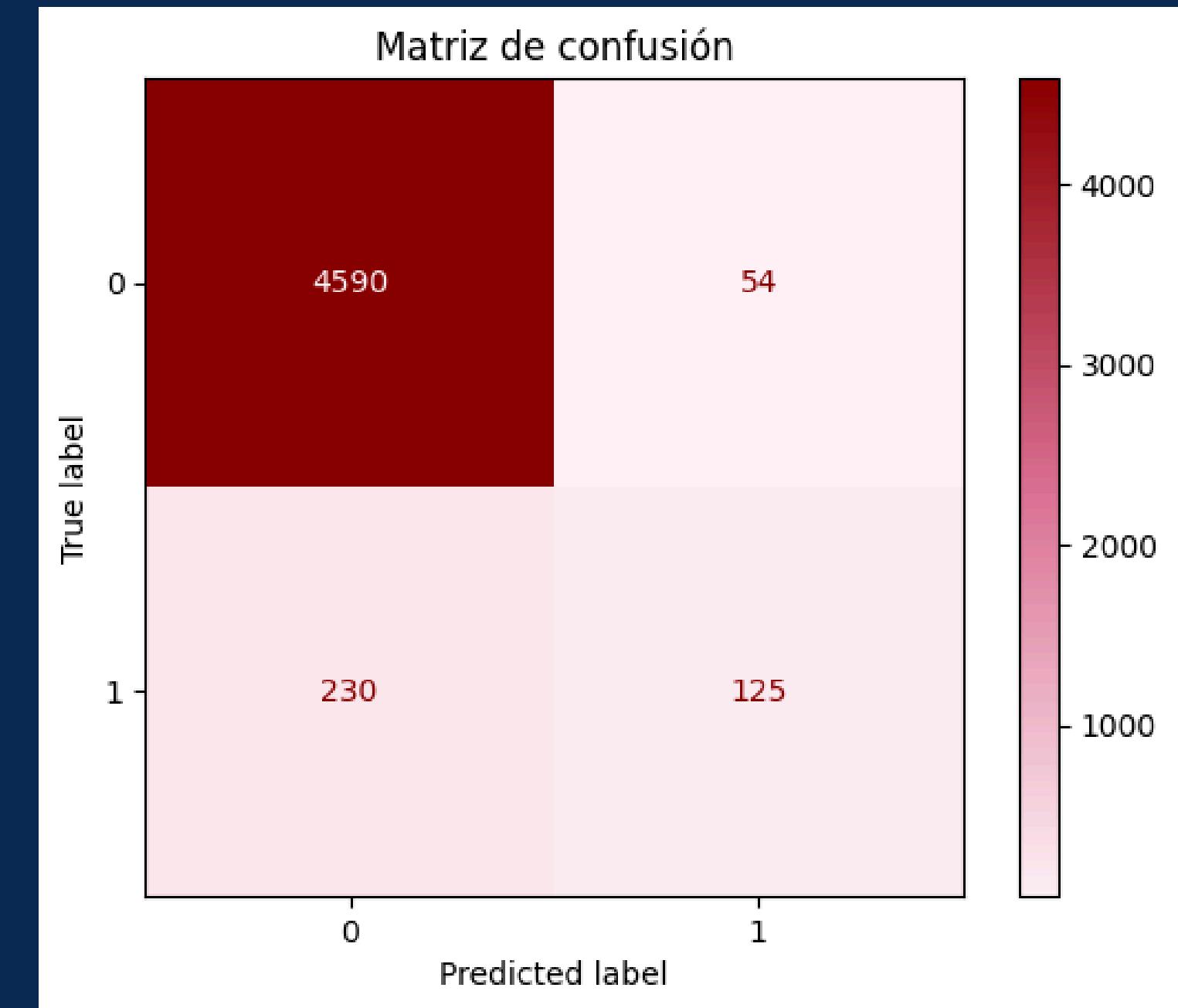
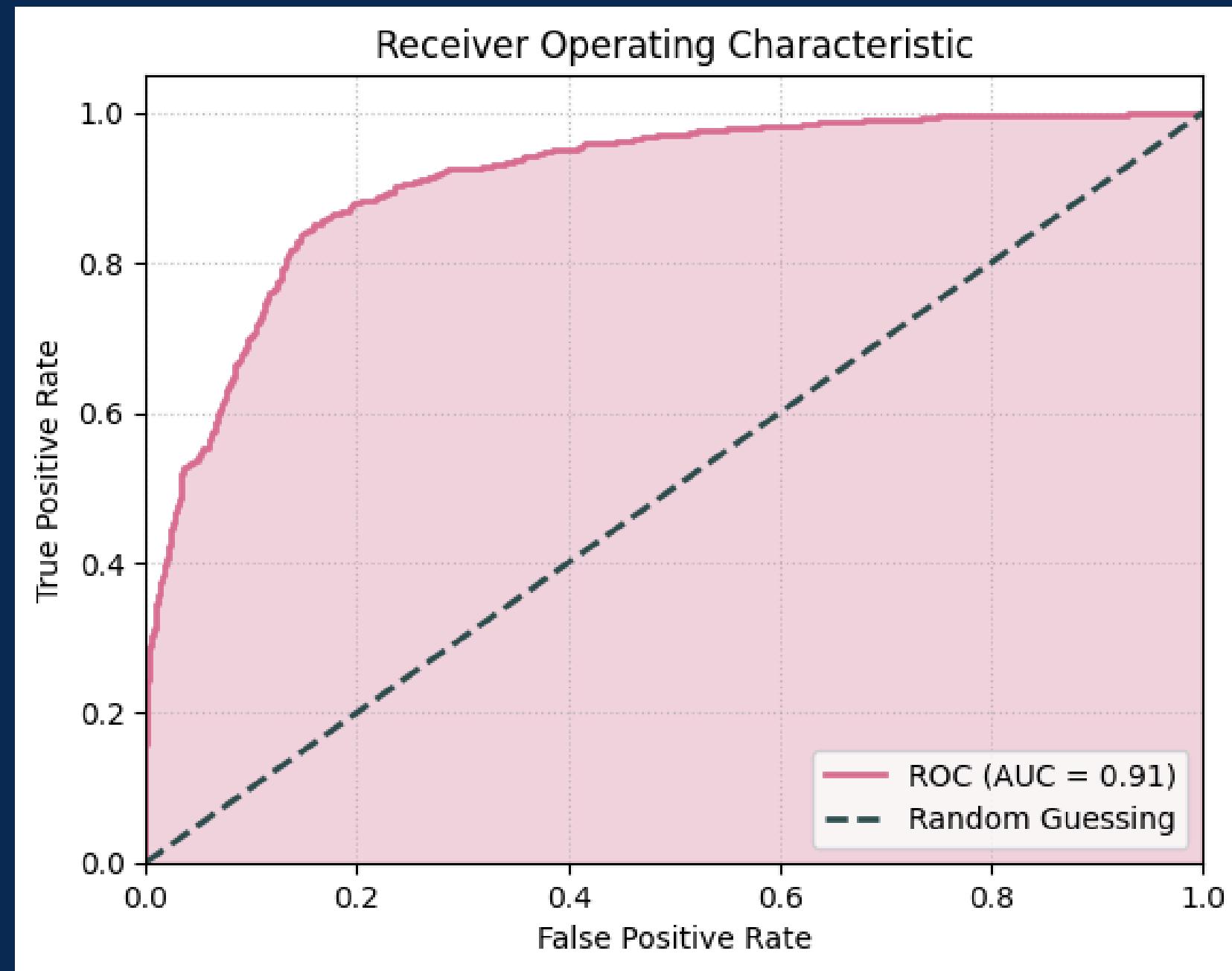


REDUCCIÓN DE DATASET USO DE PRUEBAS ESTADÍSTICAS

KS → comparación de distribuciones (data continua).
Chi-cuadrada → comparación de frecuencias/categorías (data categórica).



REGRESIÓN LOGÍSTICA



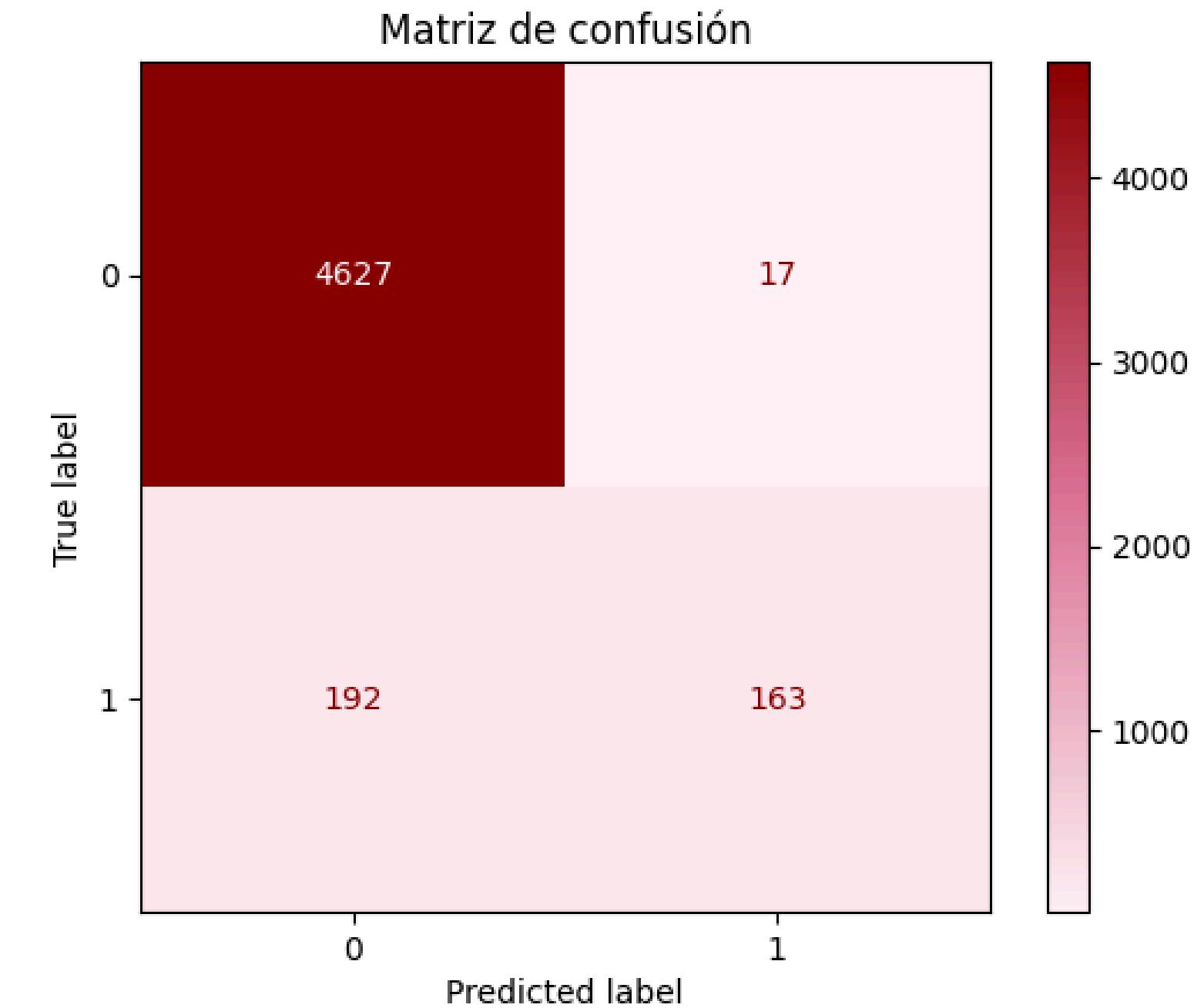
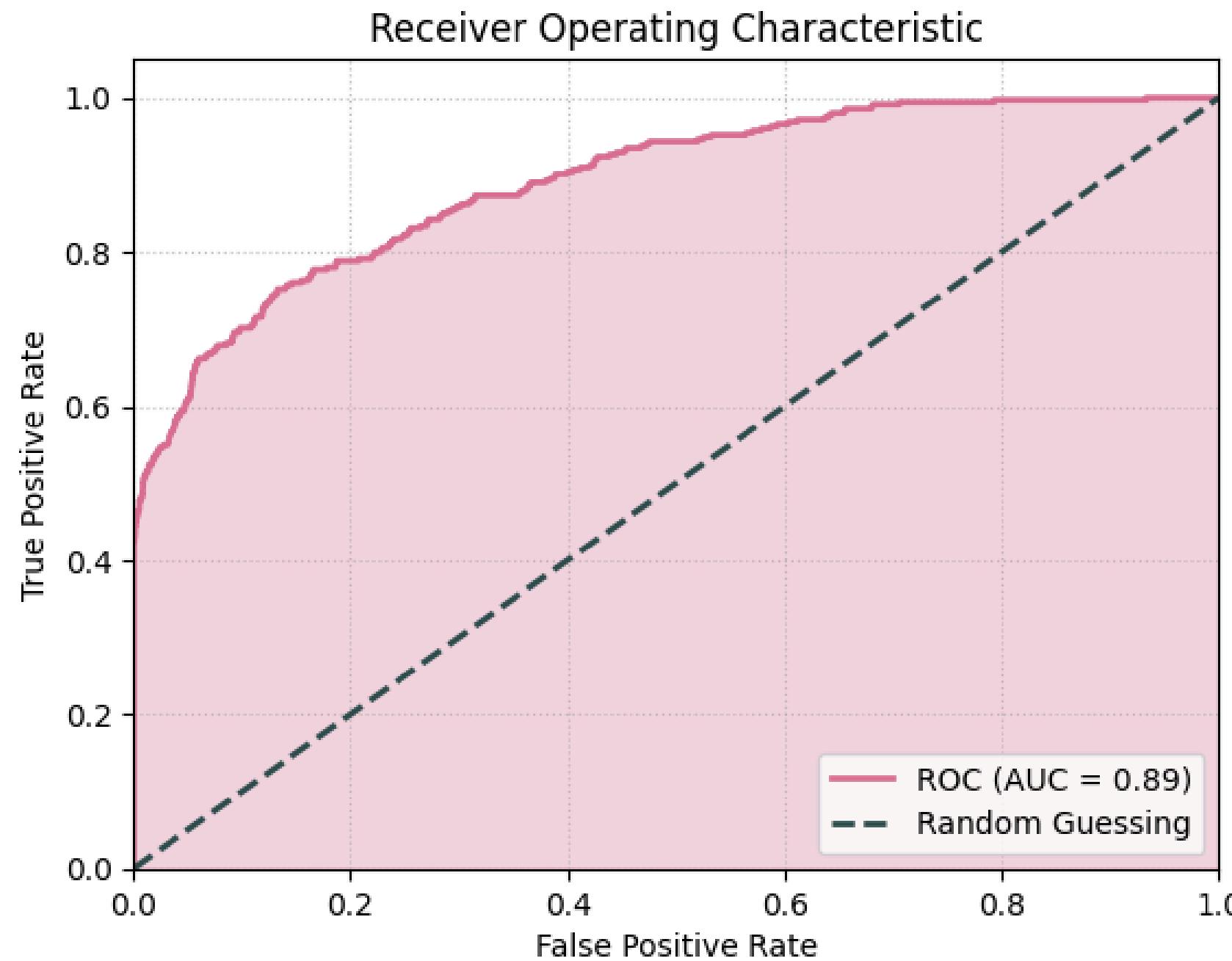
- 91% de capacidad para distinguir entre positivos y negativos.
- (4590 y 125) son los aciertos del modelo: 4590 casos negativos y 125 casos positivos.
- (54 y 230) son los errores: 54 falsos positivos y 230 falsos negativos.

APLICANDO PROCESO GAUSSIANO

Best C	ROC AUC
3.94133	0.910316

La mejora entre el modelo inicial y después de la optimización fue mínima (0.0002), indicando que el modelo ya estaba adecuadamente ajustado antes del proceso de optimización bayesiana.

MÁQUINA DE
SOPORTE
VECTORIAL CON
KERNEL RBF



- 89% de capacidad para distinguir entre positivos y negativos.

- (4627 y 163) son los aciertos del modelo: 4627 casos negativos y 163 casos positivos.
- (17 y 192) son los errores: 17 falsos positivos y 192 falsos negativos.

APLICANDO PROCESO GAUSSIANO

Best C

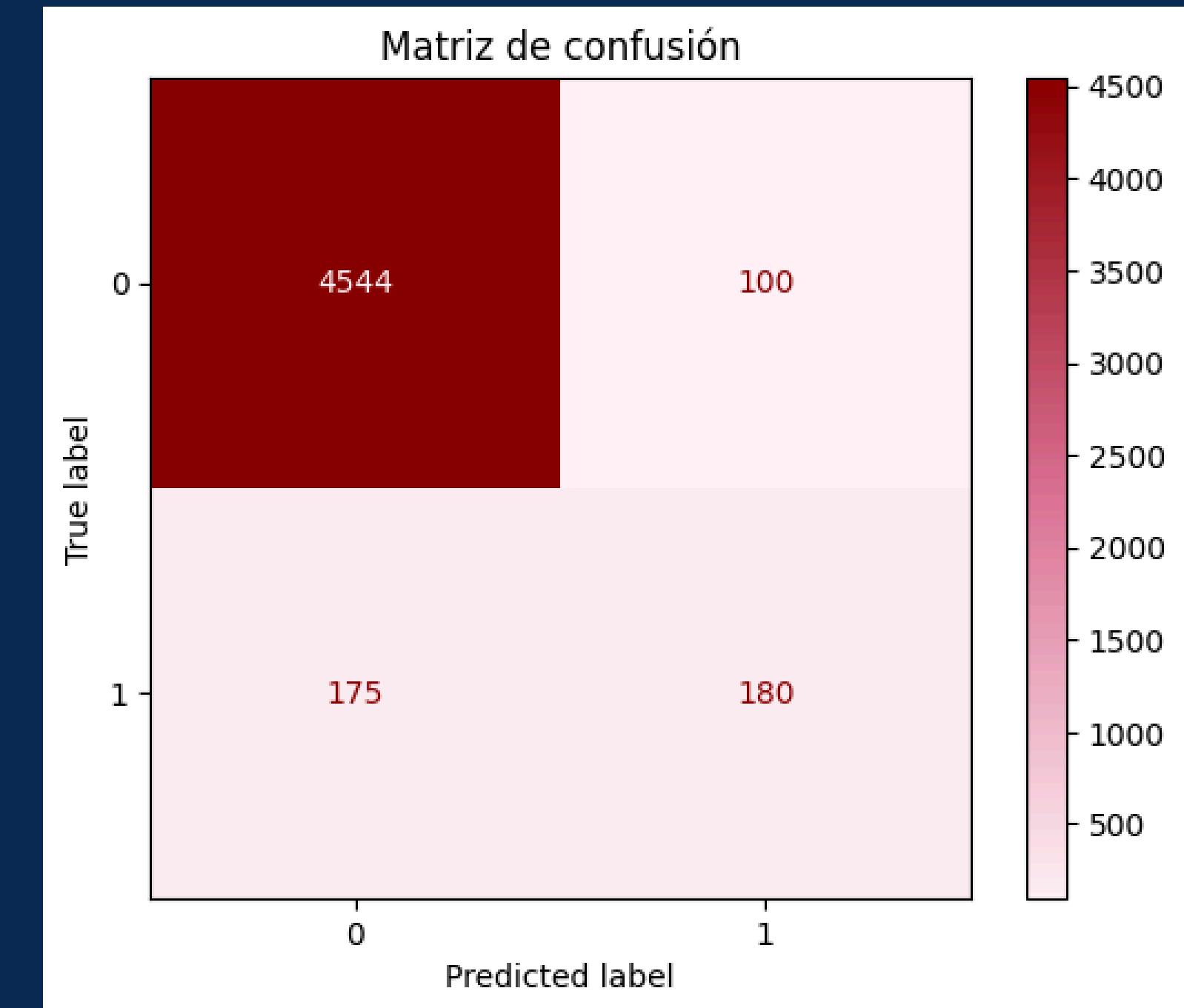
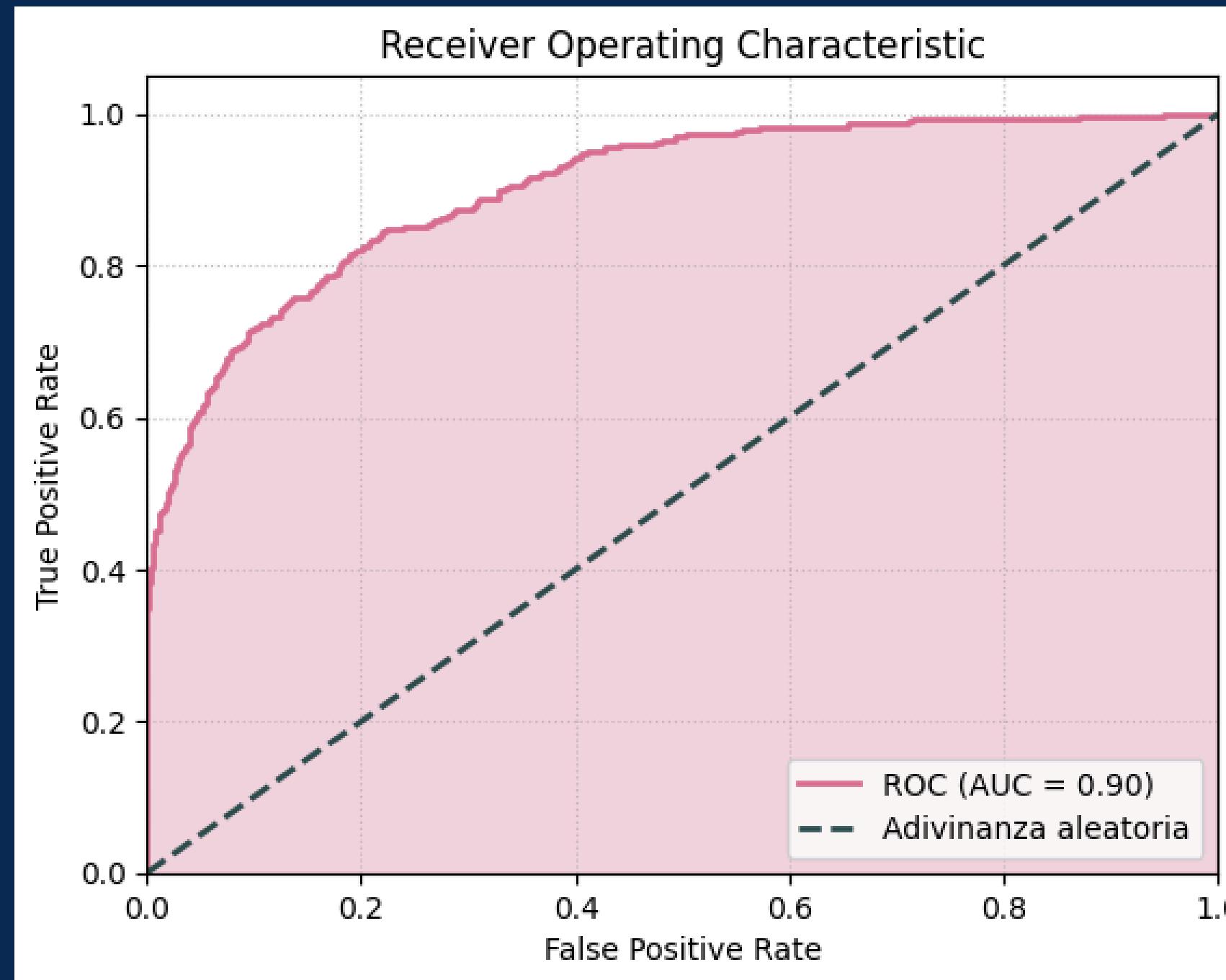
8.29228

ROC AUC

0.9047

Ligera mejora en su capacidad de discriminación respecto al desempeño obtenido antes de la optimización.

MULTI-LAYER PERCEPTRON



- 90% de capacidad para distinguir entre positivos y negativos.

- (4544 y 180) son los aciertos del modelo: 4544 casos negativos y 180 casos positivos.
- (100 y 175) son los errores: 100 falsos positivos y 175 falsos negativos.

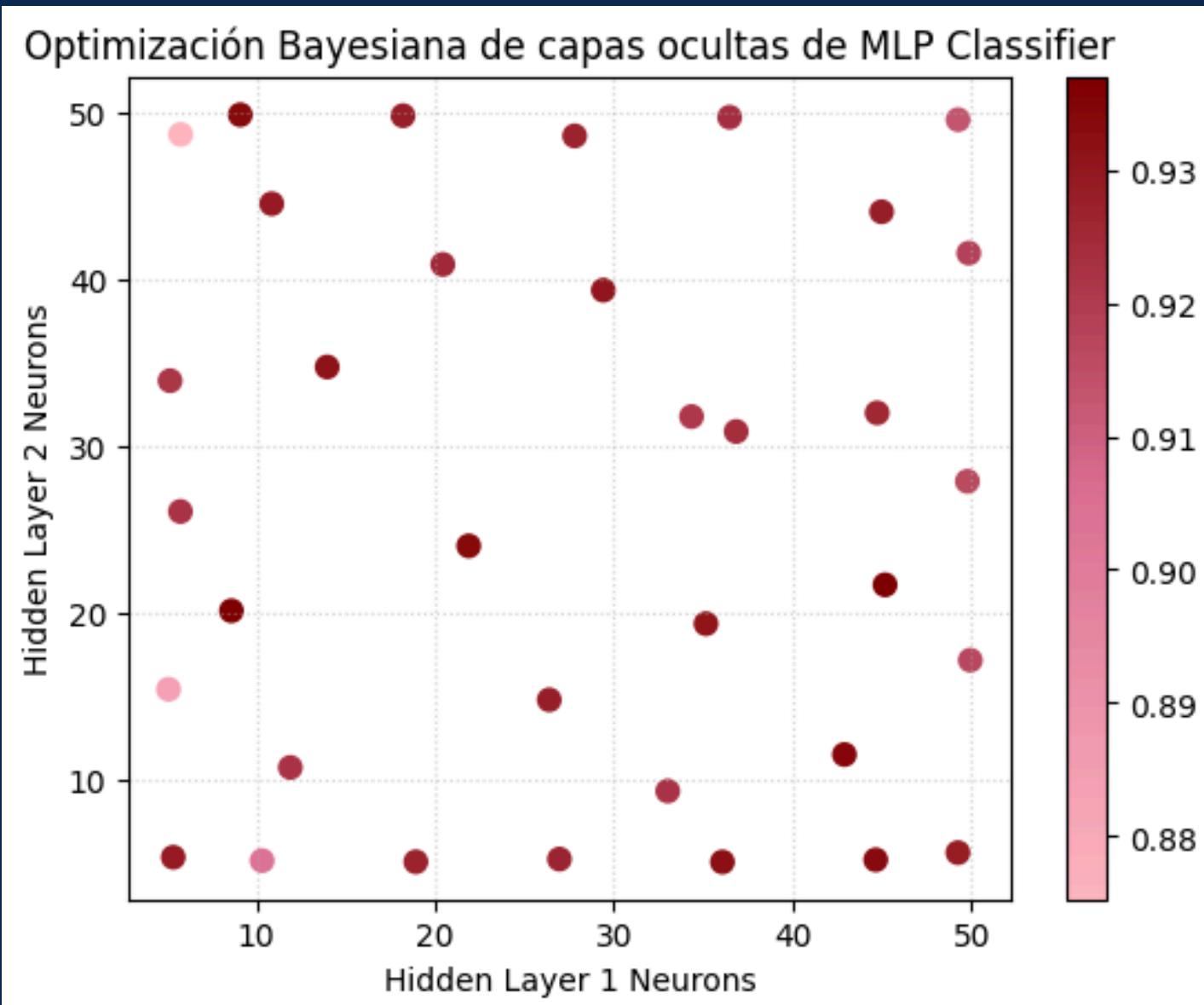
APLICANDO PROCESO GAUSSIANO

Best Hidden Layers

(45, 21)

ROC AUC

0.9369



(45, 21) significa que la red que mejor funcionó tiene dos capas ocultas:

- La primera con 45 neuronas.
- La segunda con 21 neuronas.

Cada punto es una combinación probada de neuronas durante las 30 iteraciones.

COMPARACIÓN DE MÉTRICOS

Model	accuracy_mean	accuracy_std	roc_auc_mean	roc_auc_std
Logistic Regression	0.943190	0.010083	0.910134	0.021114
Support Vector Machine	0.958191	0.008077	0.890746	0.031541
MLP Classifier	0.944989	0.010554	0.903499	0.024995

EL MEJOR MODELO INICIALMENTE FUE REGRESIÓN LOGÍSTICA, SIN EMBARGO, LA OPTIMIZACIÓN MOSTRÓ MEJORAS MÍNIMAS EN EL DESEMPEÑO...

MEJOR MODELO

Model	ROC AUC
Logistic Regression	0.910316
Support Vector Machine	0.904656
MLP Classifier	0.936939

EL MEJOR MODELO TRAS LA OPTIMIZACIÓN BAYESIANA ES EL **PERCEPTRÓN MULTICAPA**.

ESTE MODELO TIENE UNA MAYOR CAPACIDAD DE DISCRIMINACIÓN ENTRE TRANSACCIONES FRAUDULENTAS Y LEGÍTIMAS.

REGRESIÓN LOGÍSTICA

PRECISIÓN DEL 94.31% Y AUC DE 0.91, MOSTRANDO ALTA CAPACIDAD DE CLASIFICACIÓN. LA OPTIMIZACIÓN BAYESIANA ($C=3.9$) APENAS MEJORÓ EL AUC (0.91), EVIDENCIANDO QUE EL MODELO YA ESTABA BIEN AJUSTADO

SVM (KERNEL RBF)

PRECISIÓN DEL 95.82% Y AUC DE 0.89, MOSTRANDO BUEN DESEMPEÑO AL CLASIFICAR TRANSACCIONES. CON OPTIMIZACIÓN BAYESIANA, $C=8.29$, EL AUC AUMENTÓ LIGERAMENTE A 0.90, MEJORANDO SU CAPACIDAD DE DISCRIMINACIÓN

MLP CLASSIFIER

PRECISIÓN DEL 94.49% Y AUC DE 0.90, MOSTRANDO BUEN DESEMPEÑO EN LA CLASIFICACIÓN. TRAS LA OPTIMIZACIÓN BAYESIANA, CON UNA RED NEURONAL DE 45 Y 21 NEURONAS, EL AUC AUMENTÓ A 0.94, MEJORANDO SU RENDIMIENTO EN UN 3.35%.

CONCLUSIÓN

En este proyecto se desarrollaron y evaluaron tres modelos de clasificación, Regresión Logística, SVM y Perceptrón Multicapa, para detectar transacciones financieras fraudulentas, utilizando un conjunto de datos previamente limpiado, transformado y escalado. Mediante validación cruzada y Optimización Bayesiana, se ajustaron los hiperparámetros clave para maximizar su desempeño. Después de comparar sus resultados, el **MLP optimizado** mostró el mejor rendimiento, alcanzando un ROC AUC de 0.9369, superando a la Regresión Logística (0.9103) y al SVM (0.90), lo que demuestra su mayor capacidad de discriminación entre transacciones legítimas y fraudulentas.

MUCHAS GRACIAS

REPOSITORIO:

[HTTPS://GITHUB.COM/ANASOFIAHINOJOSA/PROYECTO2LABAPRESTADISTICO](https://github.com/ANASOFIAHINOJOSA/PROYECTO2LABAPRESTADISTICO)