

# PROYECTO 2 - CLASIFICACIÓN

---

CLASIFICACIÓN DE TRANSACCIONES FINANCIERAS  
COMO FRAUDULENTES O LEGÍTIMAS

JULIA HERNÁNDEZ  
ANA SOFIA HINOJOSA  
SARA HERNÁNDEZ

## 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

# 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.

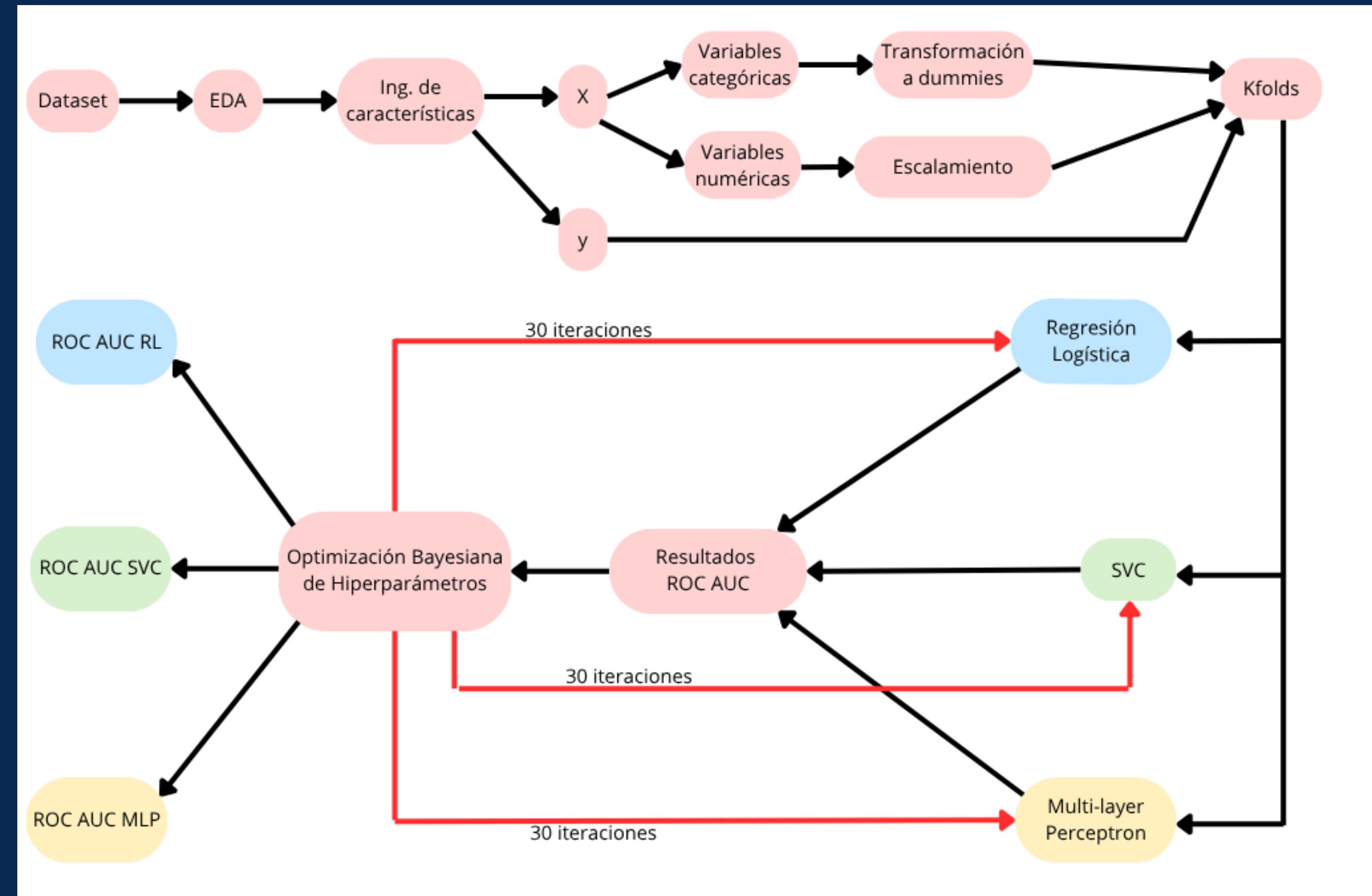
# DATASET

```
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**

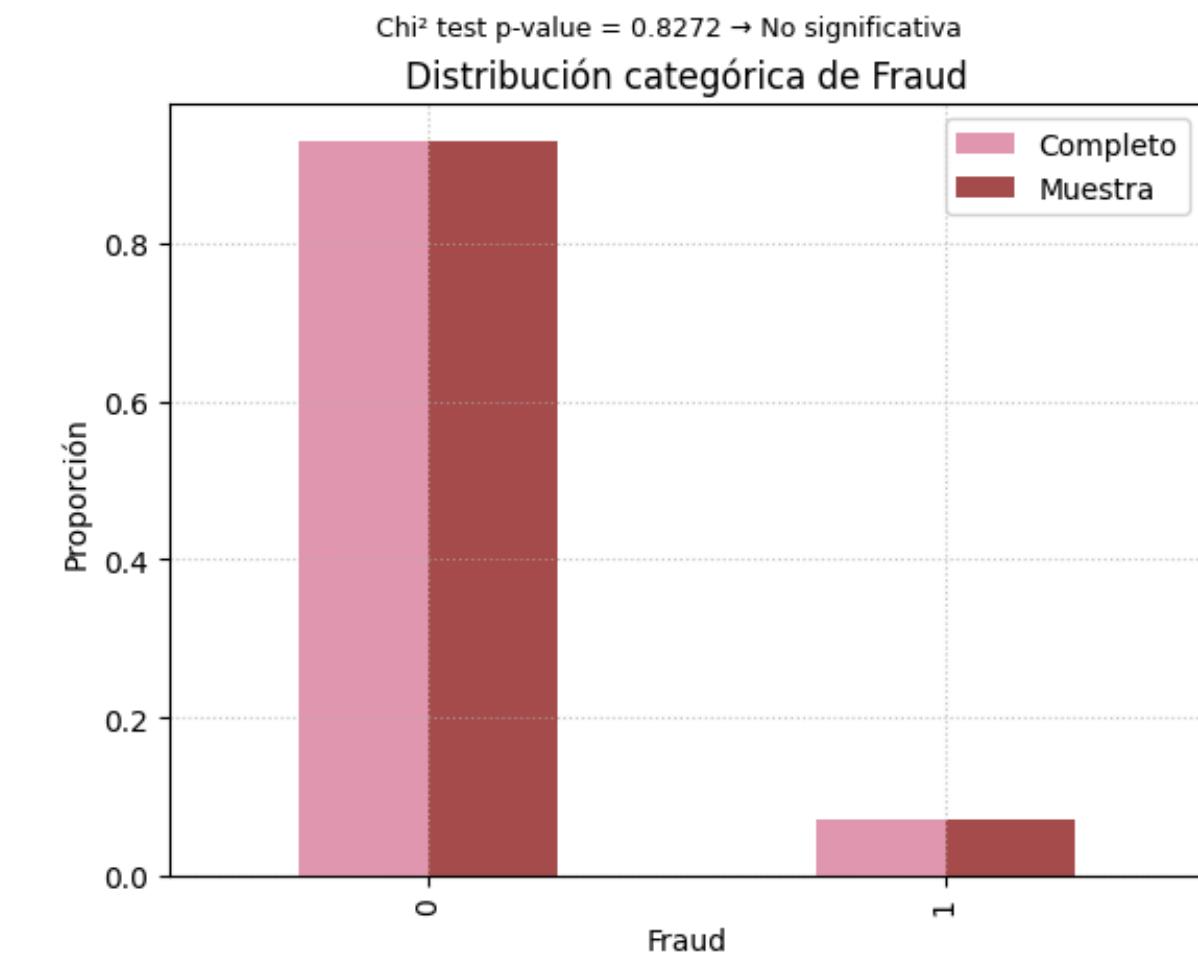
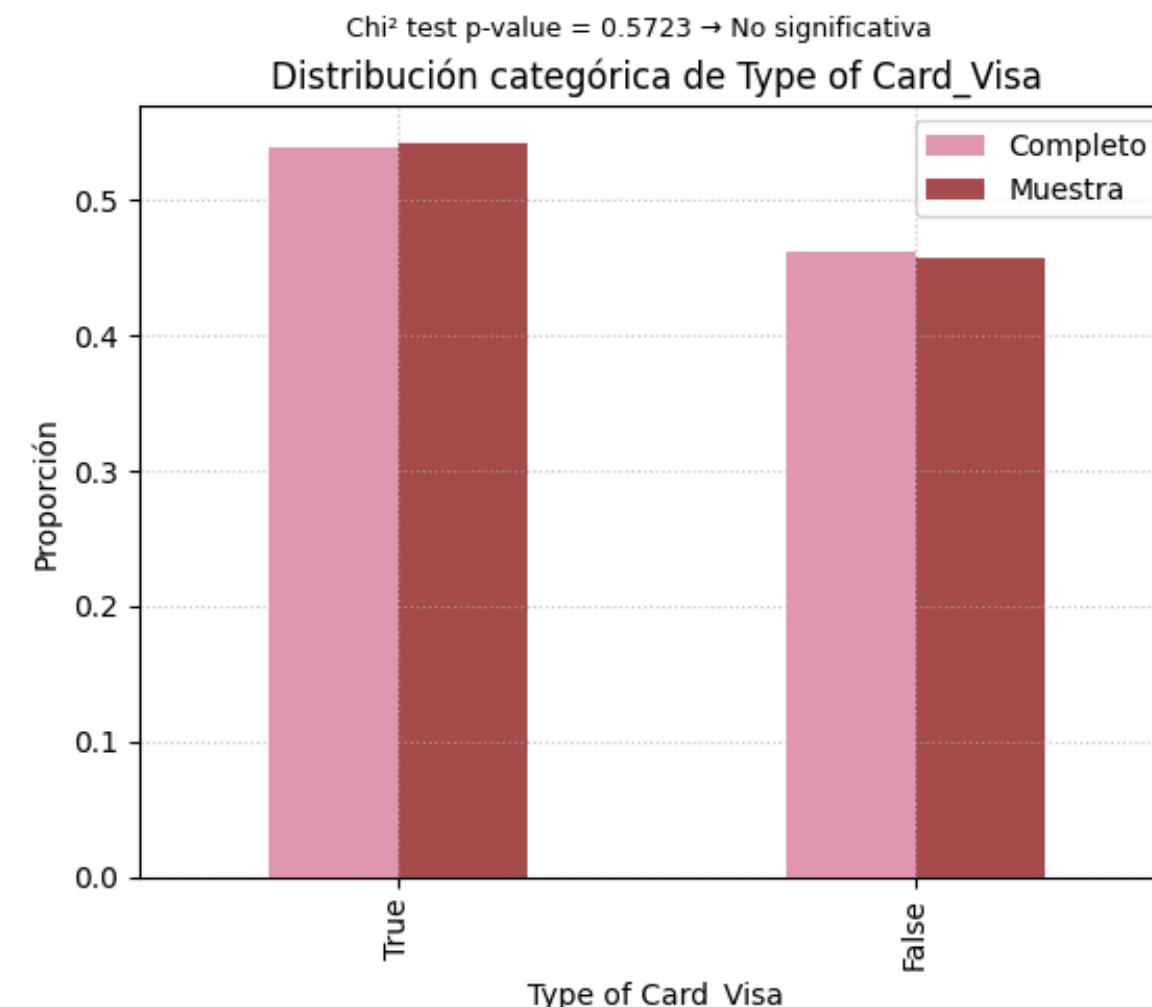
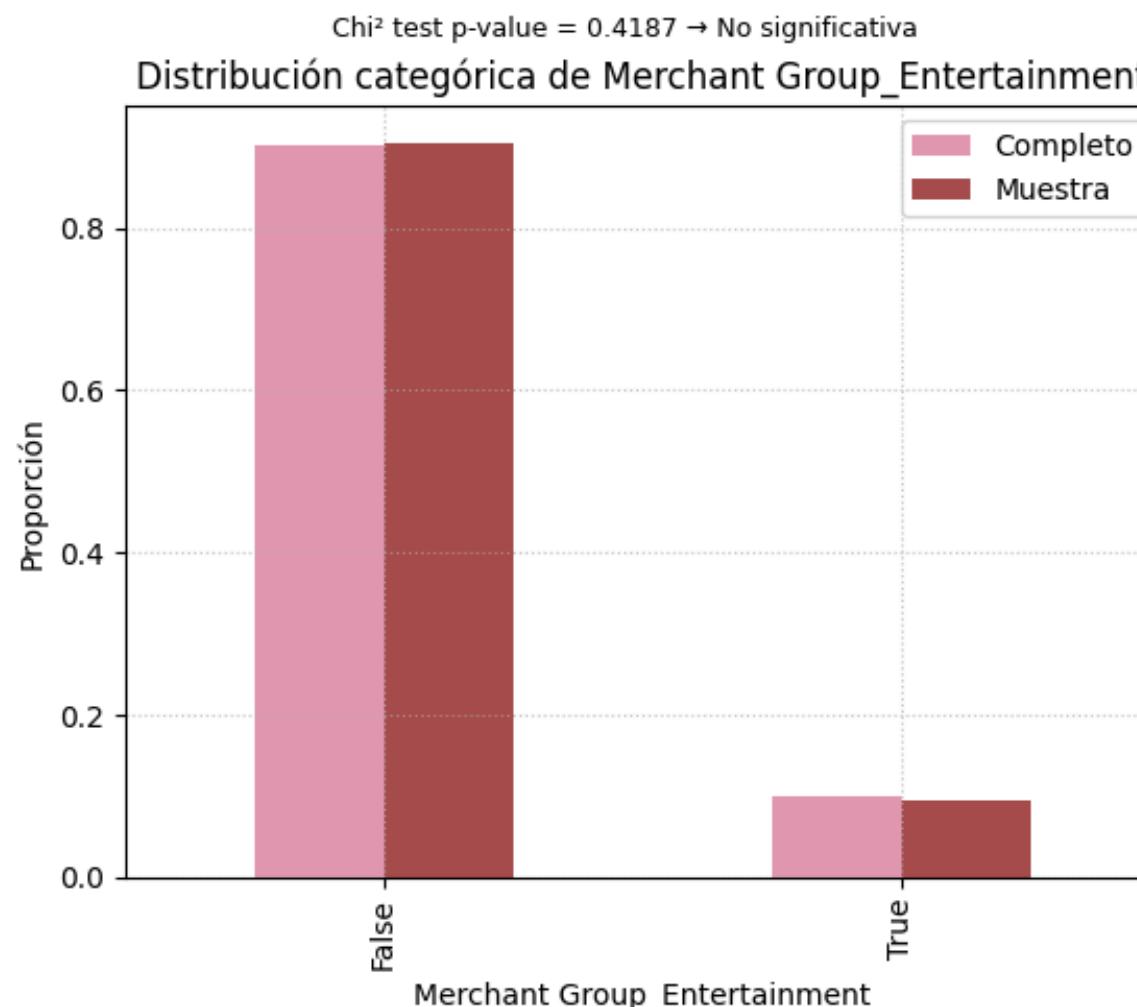
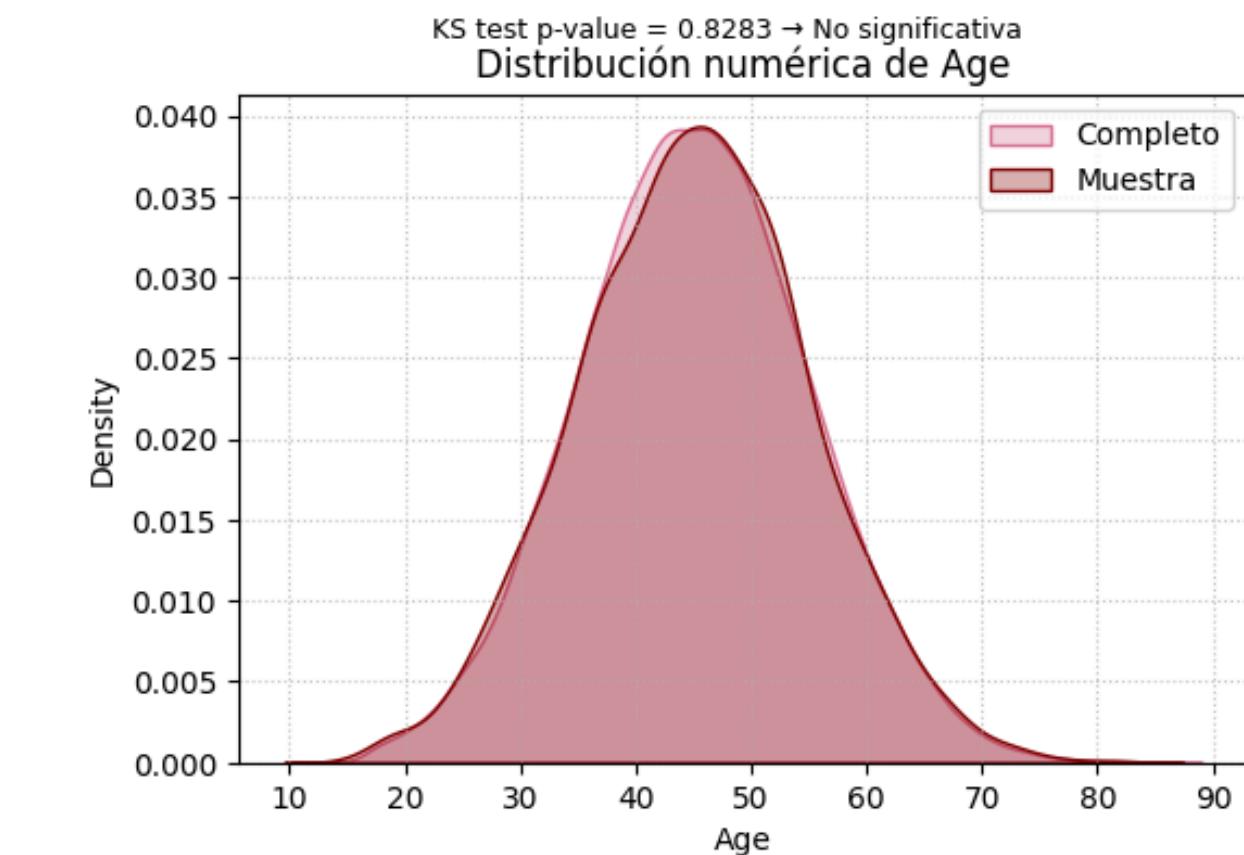
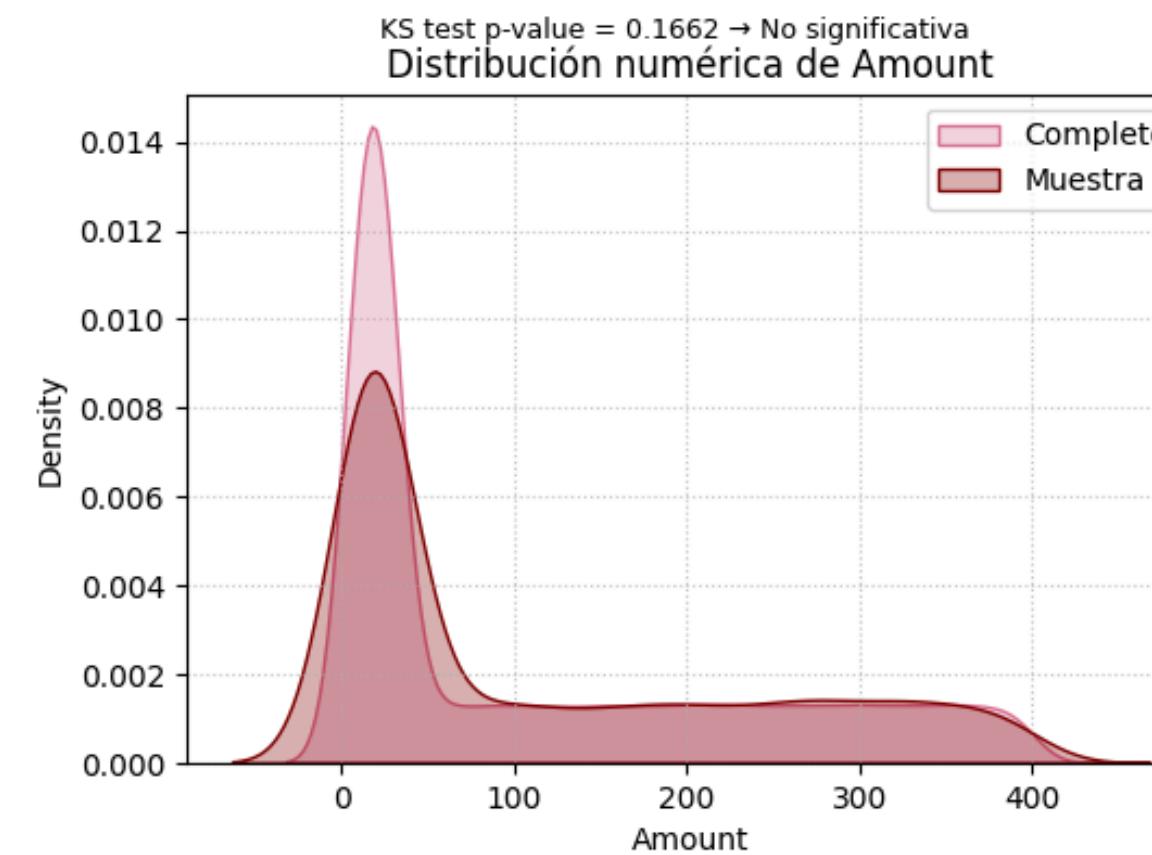
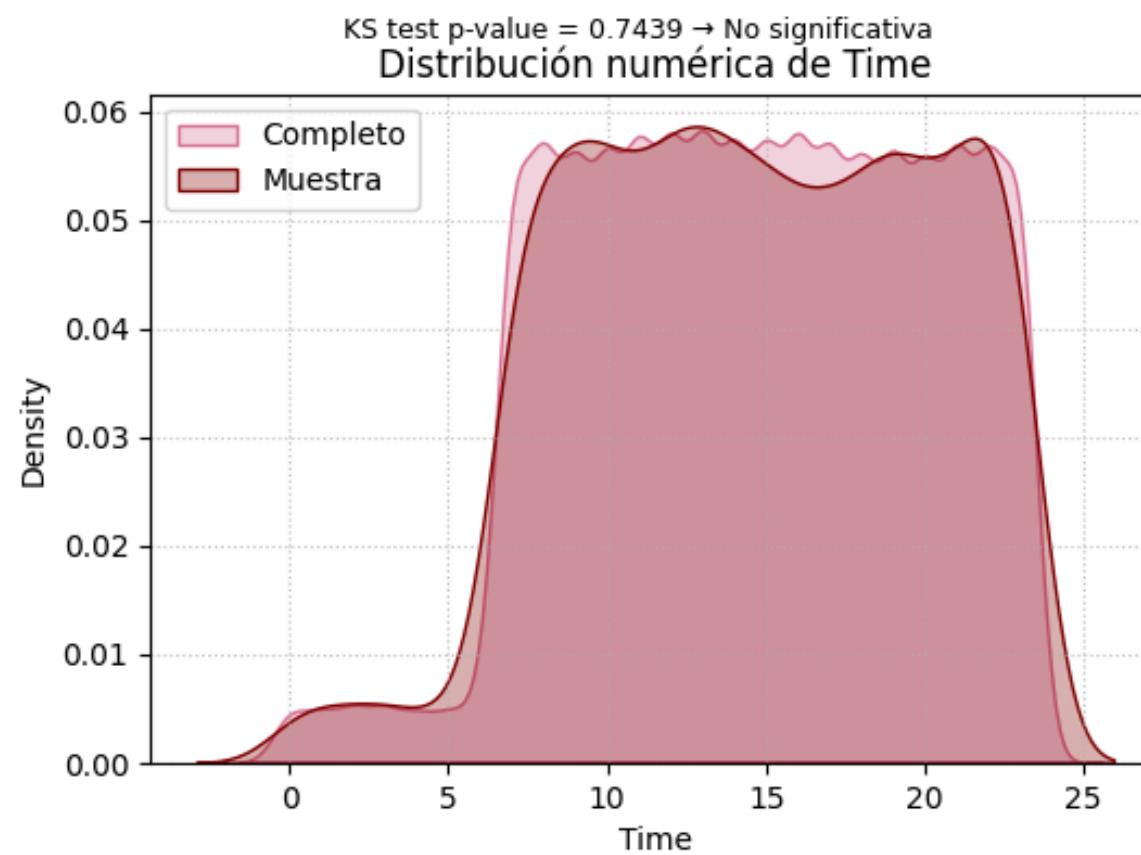
- Variables categóricas a dummies (Merchant Group, Type of Card, Bank, Gender, Country of Transaction, Entry Mode, Type of Transaction)
- “Day of Week” se transforma a valores numéricos (0 = Monday, 6 = Sunday)
- Variables numéricas (Amount, Age) escaladas, para garantizar la compatibilidad con los modelos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

# PIPELINE

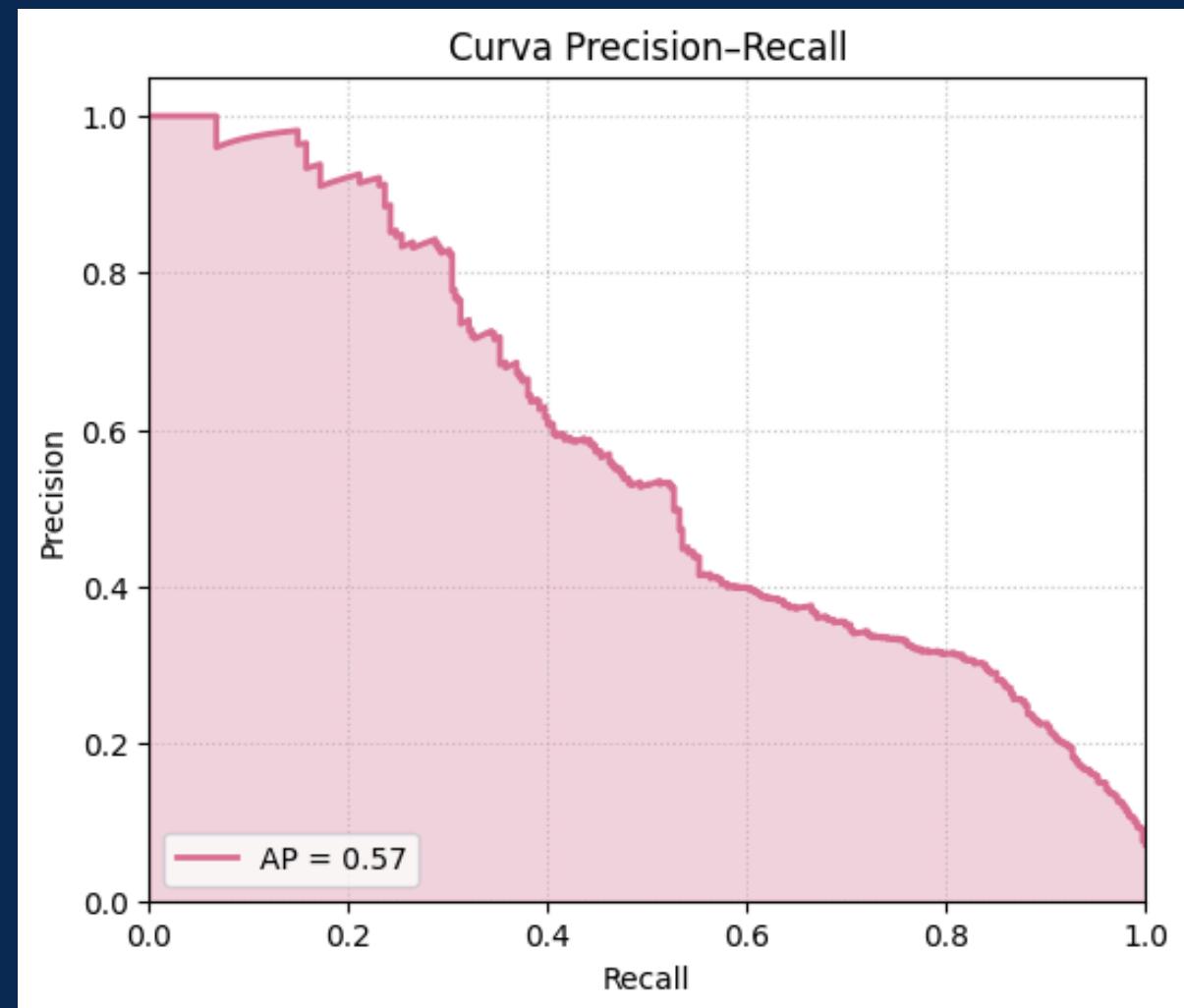
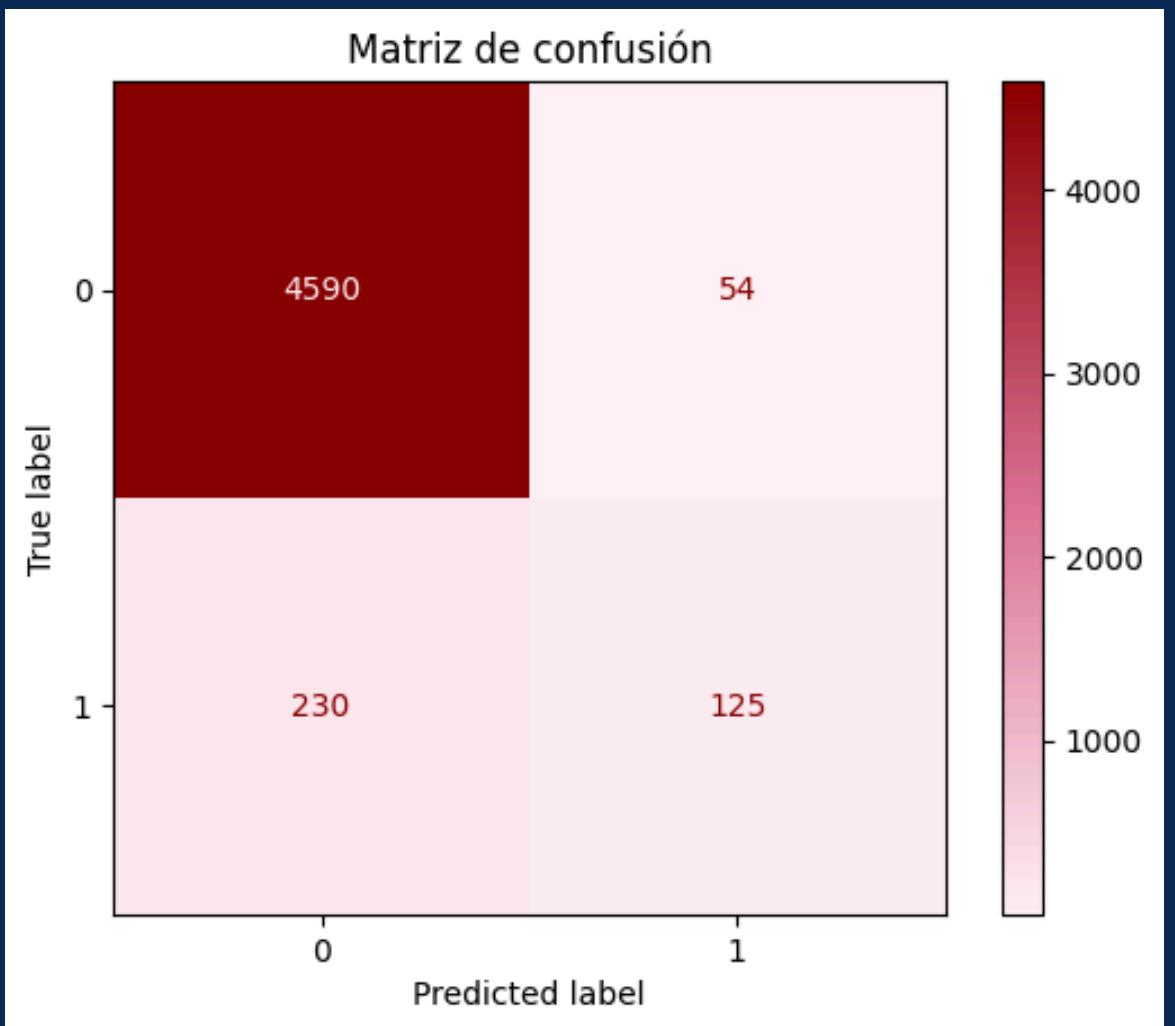
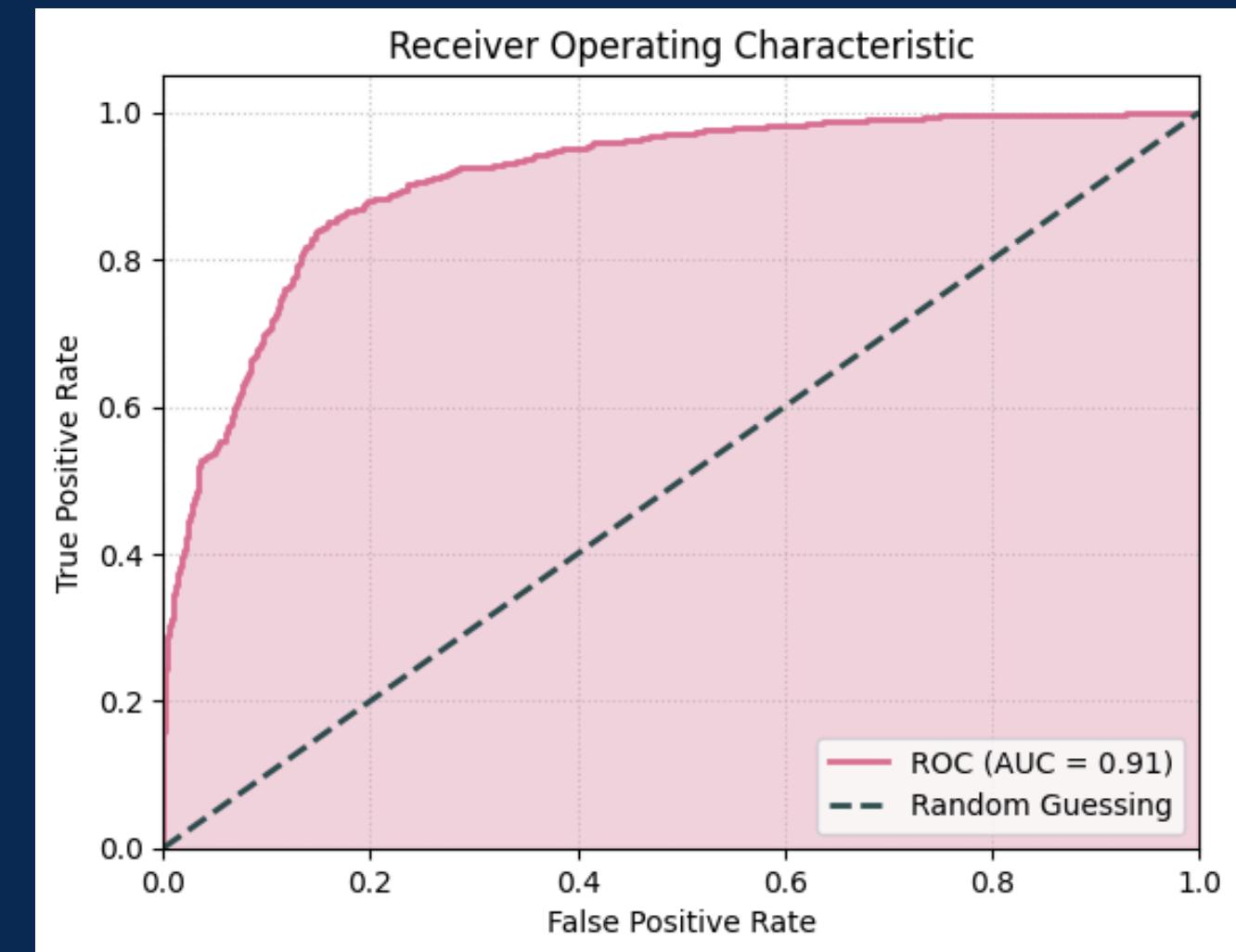


# REDUCCIÓN DE DATASET USO DE PRUEBAS ESTADÍSTICAS

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



# REGRESIÓN LOGÍSTICA



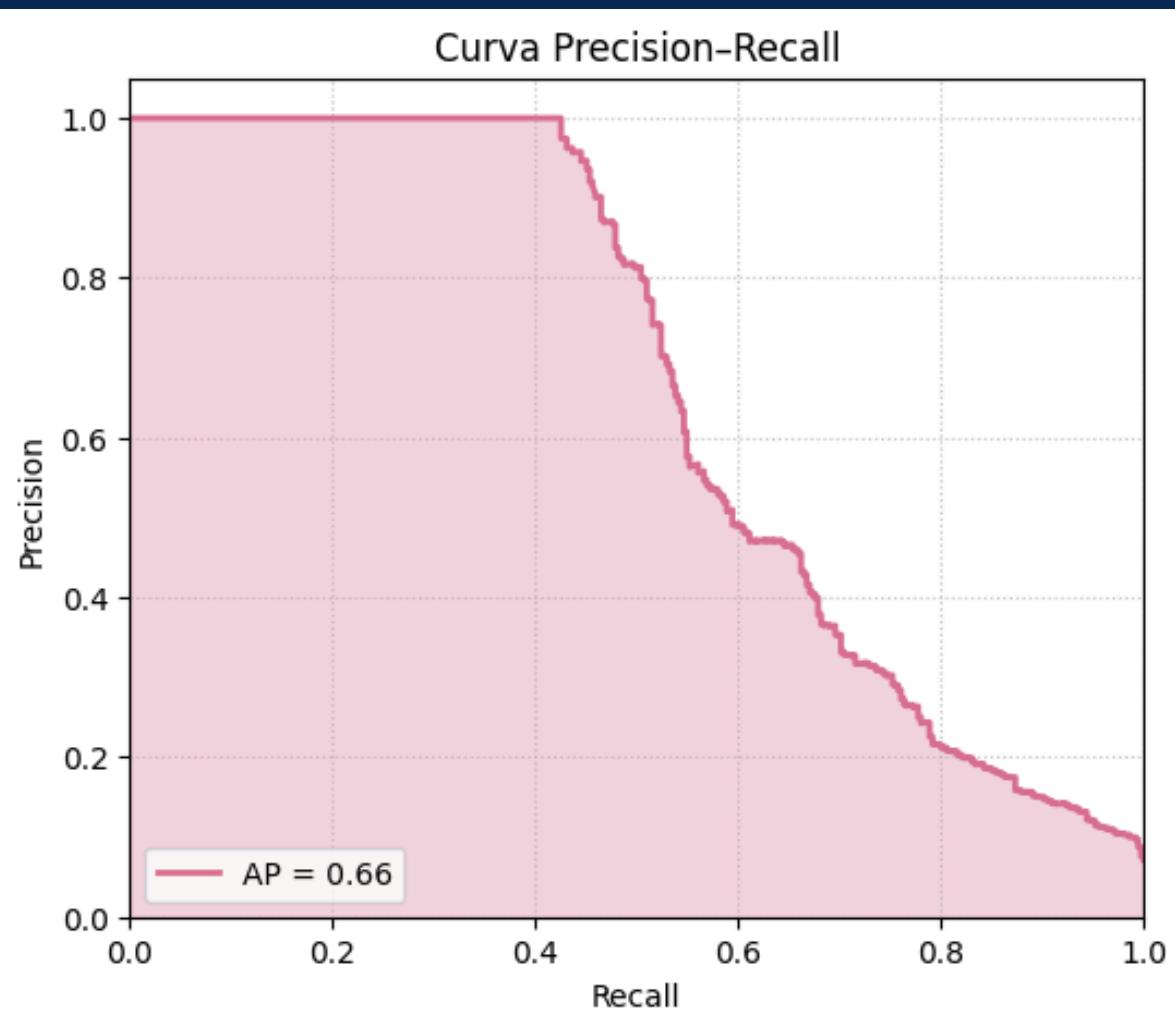
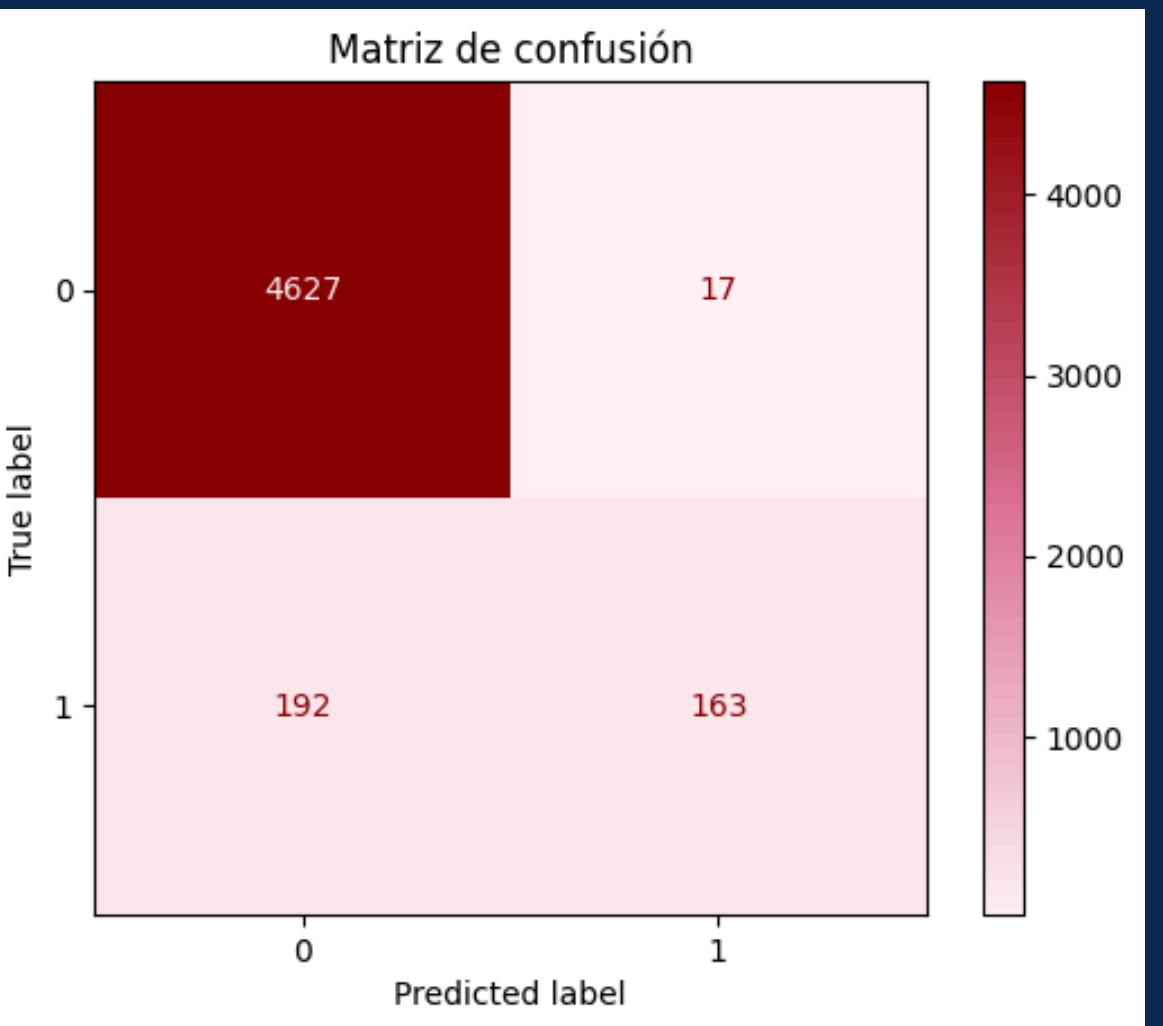
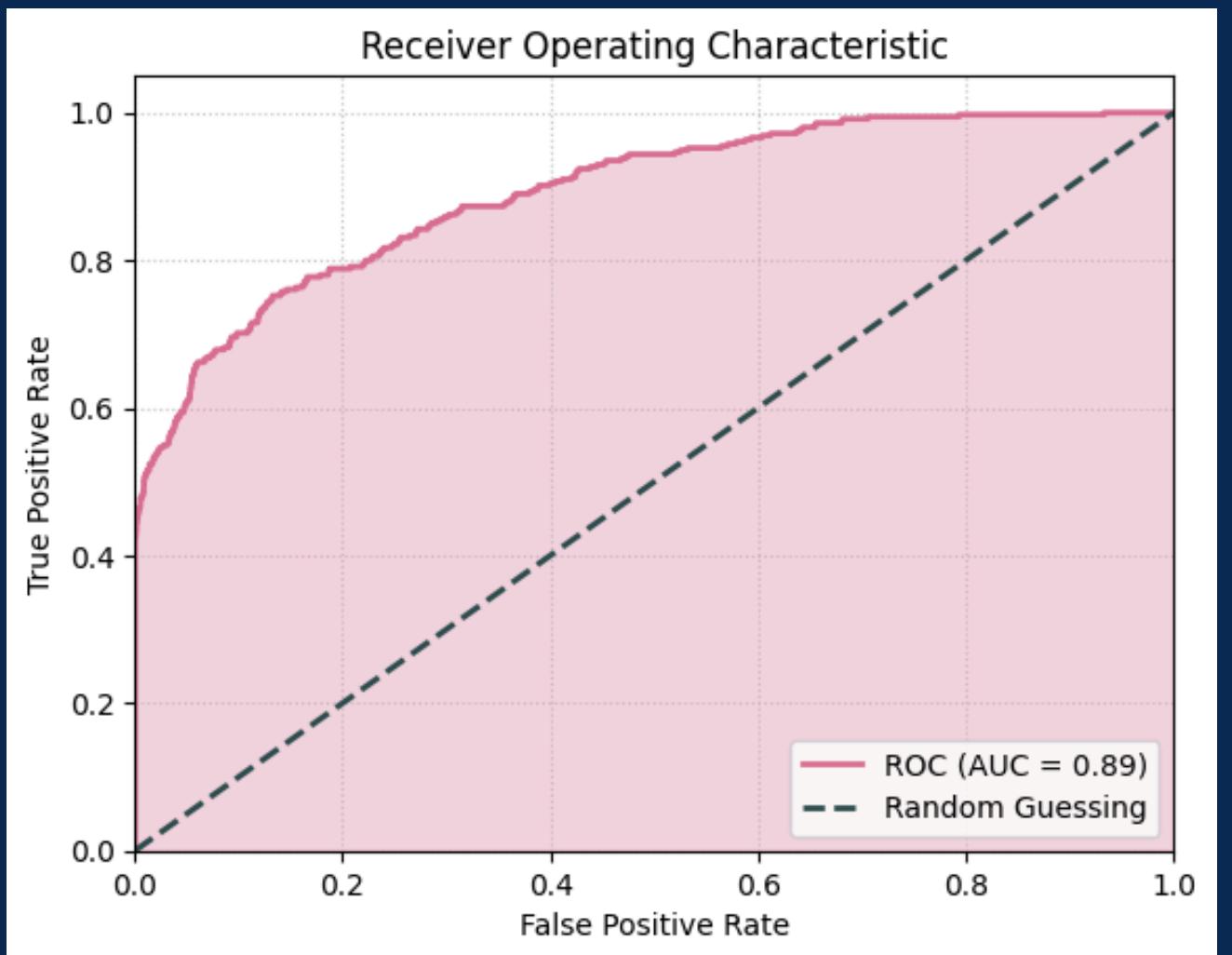
- 91% de capacidad para distinguir entre positivos y negativos.
- (4590 y 125) son los aciertos del modelo: 4590 casos negativos y 125 casos positivos.
- (54 y 230) son los errores: 54 falsos positivos y 230 falsos negativos.
- APM = 0.5697, más de la mitad de las alertas emitidas corresponden efectivamente a fraudes reales

## APLICANDO PROCESO GAUSSIANO

Best C	Average Precision Mean	ROC AUC
0.00333	0.5965	0.9103

El modelo ya había alcanzado un nivel de desempeño cercano a su límite con respecto a esta métrica, notando que hubo una mejora de únicamente 2.7 puntos porcentuales con la optimización

MÁQUINA DE  
SOPORTE  
VECTORIAL CON  
KERNEL RBF



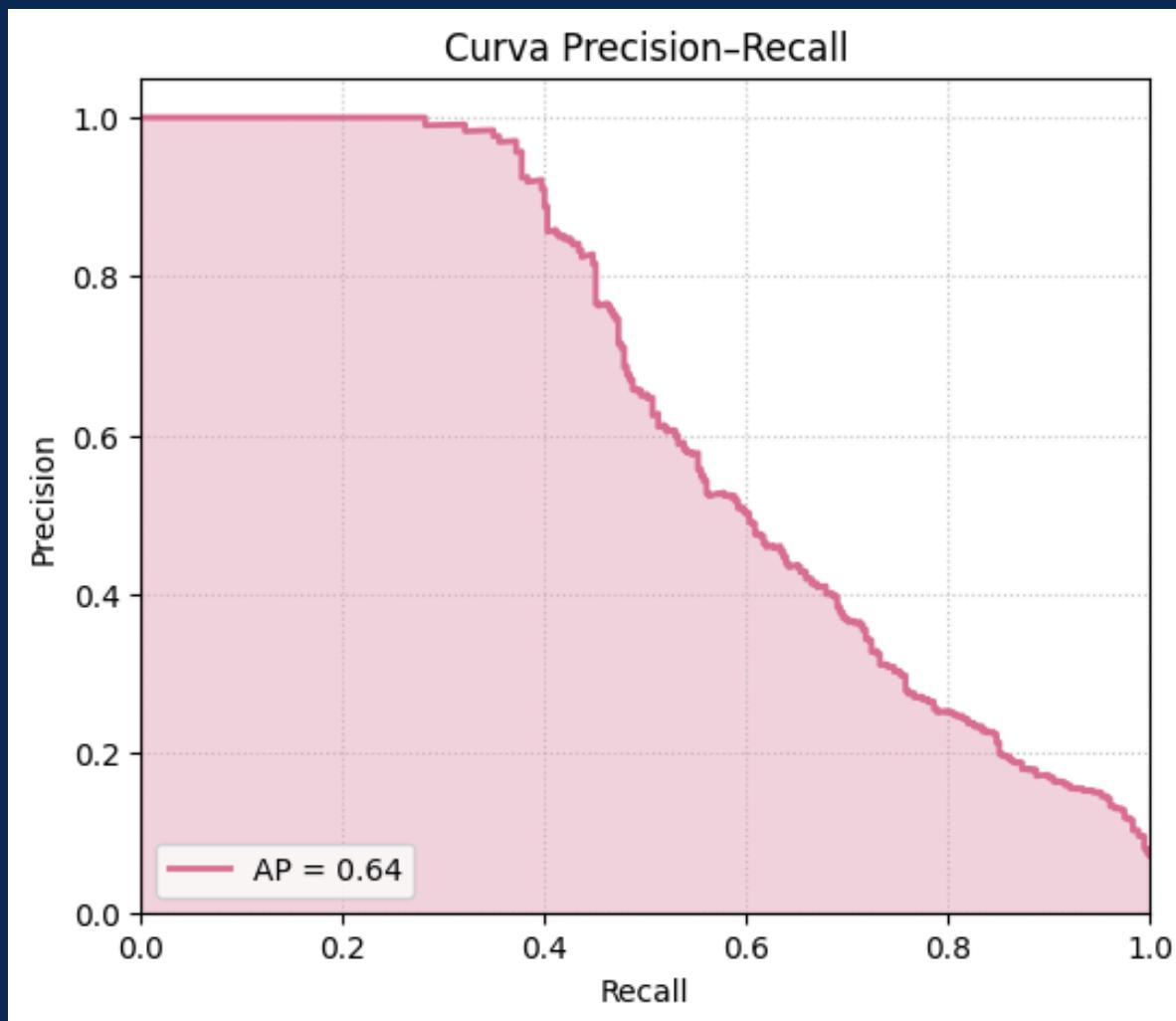
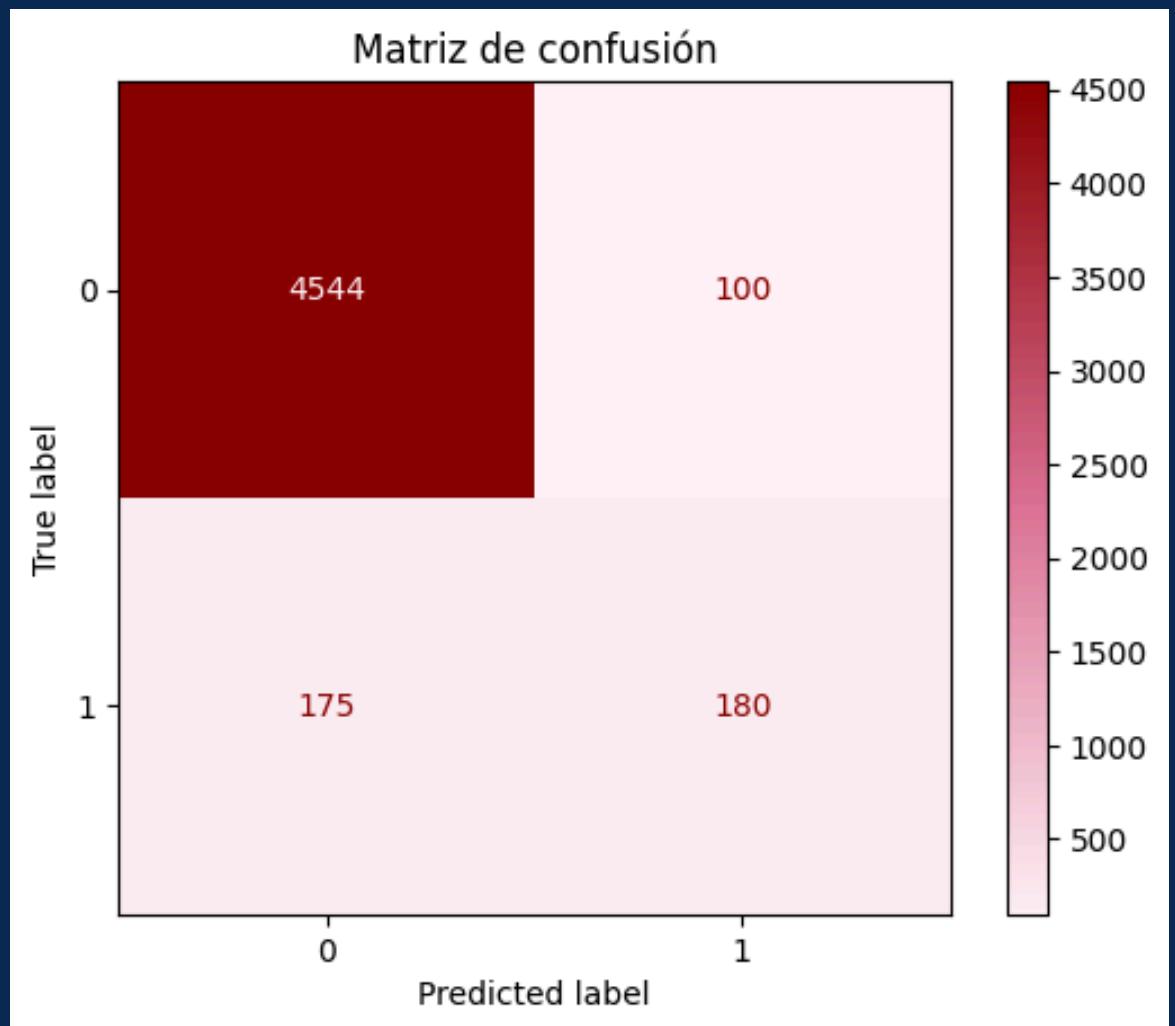
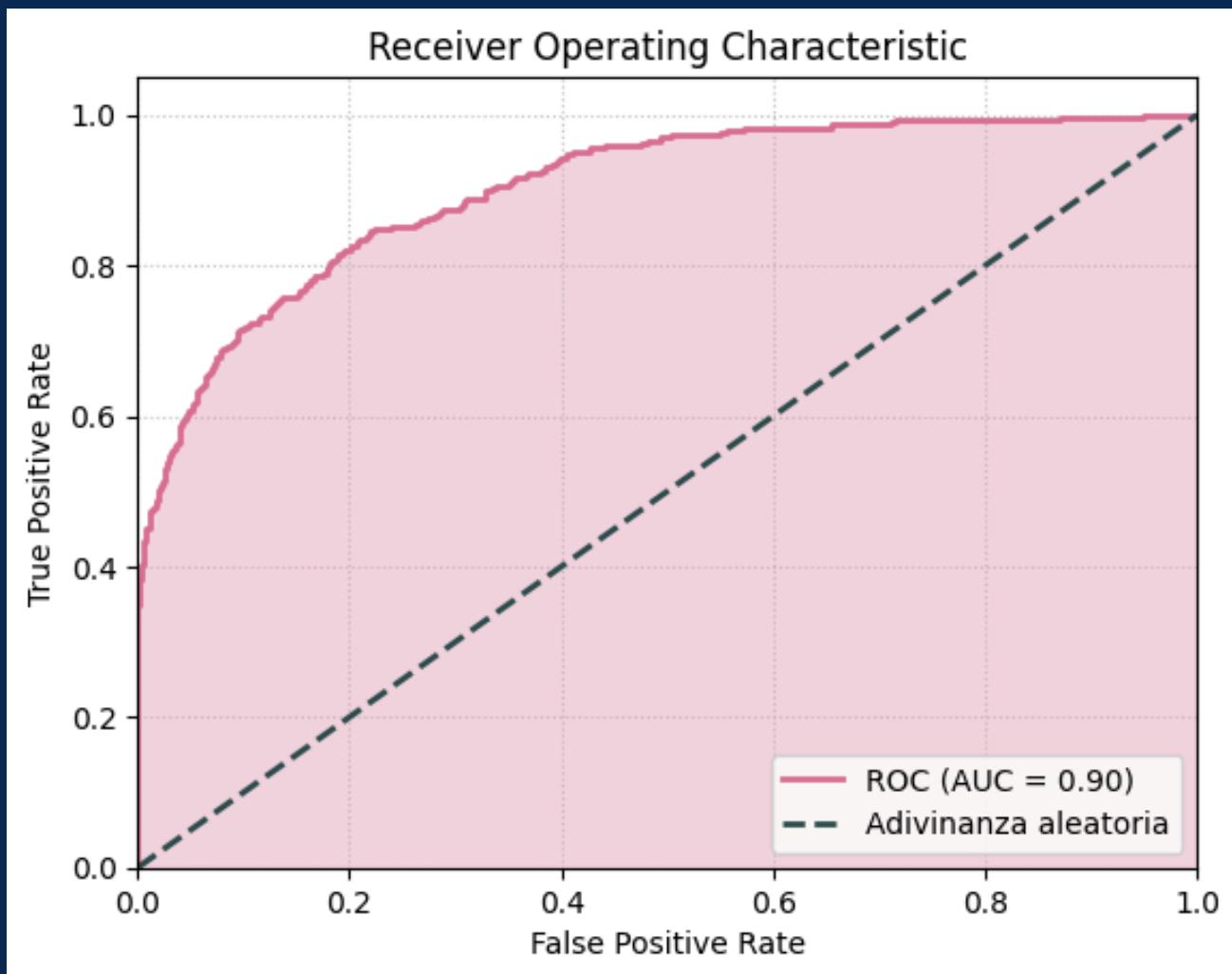
- 89% de capacidad para distinguir entre positivos y negativos.
- (4627 y 163) son los aciertos del modelo: 4627 casos negativos y 163 casos positivos.
- (17 y 192) son los errores: 17 falsos positivos y 192 falsos negativos.
- APM = 0.6583, con una desviación estándar de 0.06, reflejando un desempeño estable, además de una mayor capacidad para generar alertas precisas en comparación con RL.

## APLICANDO PROCESO GAUSSIANO

Best C	Average Precision Mean	ROC AUC
1.2961	0.6602	0.8858

Con este ajuste en el APM, se observa que aproximadamente dos tercios de las alertas emitidas por el modelo corresponden efectivamente a fraudes reales, representando una ligera mejora respecto al modelo base

# MULTI-LAYER PERCEPTRON



- 90% de capacidad para distinguir entre positivos y negativos

El modelo original tiene 2 capas ocultas:

- La primera con 25 neuronas.
- La segunda con 17 neuronas.

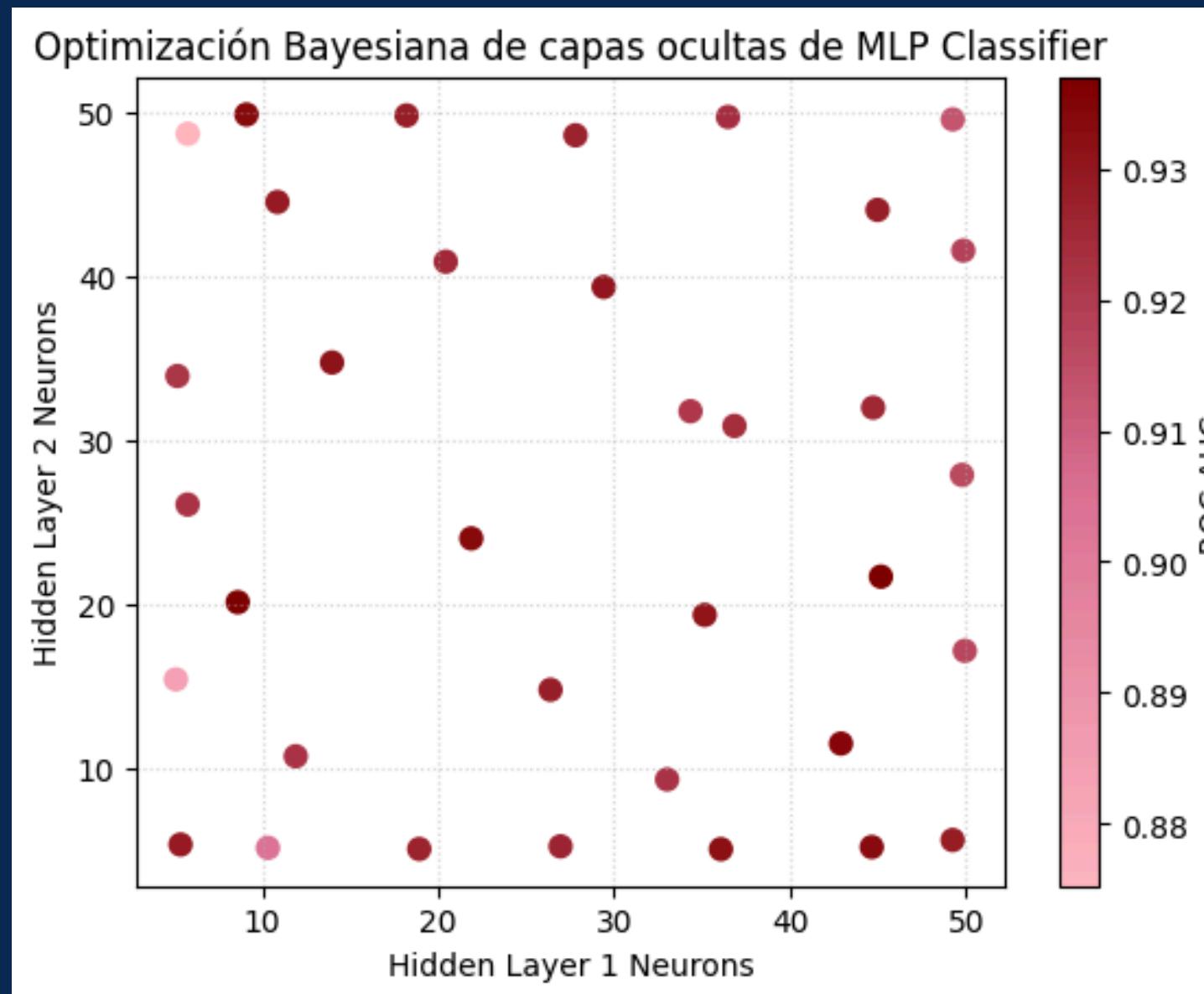
- (4544 y 180) son los aciertos del modelo: 4544 casos negativos y 180 casos positivos.

- (100 y 175) son los errores: 100 falsos positivos y 175 falsos negativos.

- APM = 0.6398, indicando que el modelo logra un equilibrio sólido entre precisión y capacidad discriminante

# APLICANDO PROCESO GAUSSIANO

Best Hidden Layers	Average Precision Mean	ROC AUC
(5, 47)	0.7043	0.9286



((5, 47) significa que la red que mejor funcionó tiene dos capas ocultas:

- La primera con 5 neuronas.
- La segunda con 47 neuronas.

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

# COMPARACIÓN DE MÉTRICOS SIN OPTIMIZACIÓN

Model	Average Precision Mean	Average Precision std	ROC AUC mean	ROC AUC std
Logistic Regression	0.5696	0.0972	0.9101	0.0211
Support Vector Machine	0.6583	0.0654	0.8907	0.0315
MLP Classifier	0.6398	0.0737	0.9035	0.0249

TOMANDO COMO PARÁMETRO EL APM, EL MEJOR MODELO INICIALMENTE  
FUE LA MÁQUINA DE SOPORTE VECTORIAL (SVM).  
SIN EMBARGO, LA OPTIMIZACIÓN MOSTRÓ MEJORAS MÍNIMAS EN EL  
DESEMPEÑO...

# MEJOR MODELO

Model	Average Precision Mean	ROC AUC
Logistic Regression	0.5966	0.9103
Support Vector Machine	0.6603	0.8858
MLP Classifier	0.7044	0.9286

EL MEJOR MODELO TRAS LA OPTIMIZACIÓN BAYESIANA ES EL  
**PERCEPTRÓN MULTICAPA.**  
ESTE MODELO ES CAPAZ DE DISTINGUIR CON MAYOR PRECISIÓN LAS  
TRANSACCIONES FRAUDULENTAS DE LAS LEGÍTIMAS. ADEMÁS, LAS  
ALERTAS EMITIDAS POR EL MODELO TIENEN UNA PROBABILIDAD  
CONSIDERABLEMENTE MAYOR DE CORRESPONDER A FRAUDES REALES.

# REGRESIÓN LOGÍSTICA

APM DEL 0.56 Y AUC DE 0.91, MOSTRANDO MODERADA CAPACIDAD DE CLASIFICACIÓN. LA OPTIMIZACIÓN BAYESIANA ( $C=0.003$ ) APENAS MEJORÓ EL APM (0.59), EVIDENCIANDO QUE EL MODELO YA ESTABA BIEN AJUSTADO

## SVM (KERNEL RBF)

APM DEL 0.65 Y AUC DE 0.89, MOSTRANDO MEJOR DESEMPEÑO AL CLASIFICAR TRANSACCIONES QUE LA REGRESIÓN LOGÍSTICA. CON OPTIMIZACIÓN BAYESIANA,  $C=1.29$ , EL APM AUMENTÓ LIGERAMENTE A 0.66, PRESENTANDO UNA LIGERA MEJORA RESPECTO AL MODELO BASE

## MLP CLASSIFIER

APM DEL 0.63 Y AUC DE 0.90, MOSTRANDO BUEN DESEMPEÑO EN LA CLASIFICACIÓN. TRAS LA OPTIMIZACIÓN BAYESIANA, CON UNA RED NEURONAL DE 5 Y 47 NEURONAS, EL APM AUMENTÓ A 0.70 Y ROC A 0.92, MOSTRANDO UNA NOTABLE MEJORA EN AMBAS MÉTRICAS

## 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 Average Precision Mean de 0.7043 y un ROC AUC de 0.9286, superando a la Regresión Logística y al SVM en ambas métricas, 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)