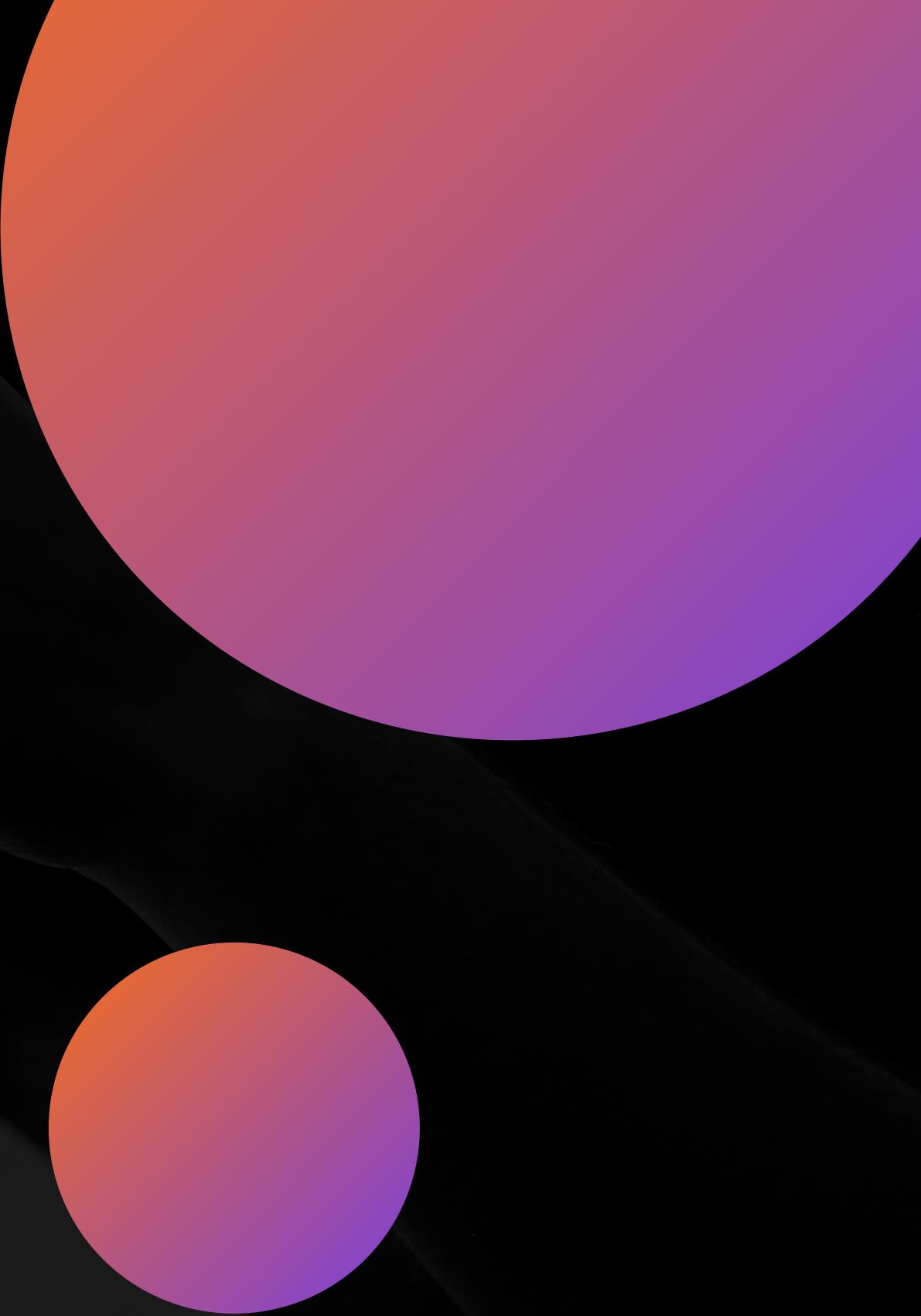


# Big Data

BANK CUSTOMER CHURN PREDICTION



# Introducción al Dataset

¿ DE QUÉ SE TRATA ?

# Bank Customer Churn

Es un dataset de la industria bancaria. Tiene información de clientes que se quedaron en el banco y clientes que abandonaron (churn).

01

## Audiencia

Este conjunto de datos está **dirigido principalmente a profesionales y analistas de la industria bancaria** que están interesados en comprender y predecir el comportamiento de sus clientes

02

## Temática

El caso de negocio requiere que clasifiquemos y predigamos si un cliente va a abandonar o no el banco

# Bank Customer Churn

Es un dataset de la industria bancaria. Tiene información de clientes que se quedaron en el banco y clientes que abandonaron (churn).

03

## Participantes

Los participantes en este conjunto de datos son los ***clientes bancarios que han sido seleccionados como sujetos de estudio.***

04

## Metodología

Extracción de información de las bases de datos internas del banco, donde se almacenan los detalles de los clientes y sus transacciones



# Metadata de variables

- 01 Customer ID**  
Una identificación única para cada cliente
- 02 Surname**  
El apellido del cliente
- 03 Credit Score**  
Un valor numérico que representa el puntaje crediticio del cliente
- 04 Geography**  
El país de residencia del cliente (Francia, España o Alemania)
- 05 Gender**  
El género del cliente (masculino o femenino)
- 06 Tenure**  
El número de años que el cliente lleva en el banco



# Metadata de variables

- 07 Balance**  
El balance de la cuenta del cliente
- 08 NumOfProducts**  
La cantidad de productos bancarios que el cliente utiliza
- 09 HasCrCard**  
Si el cliente tiene o no tarjeta de crédito (1 = sí, 0 = no)
- 10 IsActiveMember**  
Si el cliente es activo o no (1 = sí, 0 = no)
- 11 EstimatedSalary**  
El salario estimado del cliente
- 12 Exited**  
Si el cliente abandonó o no (1 = sí, 0 = no)



# Plan de Trabajo

¿ CUÁL ES EL PASO A PASO ?

# Plan de Trabajo

- 
- A dark, semi-transparent background image shows two people laughing heartily. One person is wearing a baseball cap and a hoodie, while the other is in a dark long-sleeved shirt. They are positioned in front of a large bookshelf filled with books.
- 01 Caso de negocio
  - 02 Preparación de los datos
  - 03 Manipulación de datos
  - 04 Análisis exploratorio
  - 05 Modelado

# Definición del Problema

¿ A Q U É N O S E N F R E N T A M O S ?

# Caso de negocio

El banco necesita saber que cantidad de sus clientes se van a ir y le interesa captar a todos sin importar si entra alguno que se iba a quedar. Quieren hacer una campaña de marketing para fidelizarlos



## Costo del falso positivo

El costo del falso positivo, es decir, de ***predecir que el cliente se va a ir cuando en realidad se queda***, es el ***costo de hacer una campaña de marketing que no era necesaria***. Ese costo es de 50 USD por persona.



## Costo del falso negativo

El costo del falso negativo, es decir, de ***predecir que alguien se queda cuando en realidad se va***, es la ***pérdida de la mitad del valor medio de un cliente***. El valor medio de un cliente es de 500 USD.



## Ganancia del verdadero positivo

La ganancia por el verdadero positivo, es decir, de ***predecir que el cliente se va a ir y efectivamente se va***, es la ***mitad del valor medio del cliente***. El valor medio de un cliente es de 500 USD.



# Preparación y tratamiento

¿ CÓMO MODIFICAMOS LOS DATOS?



# Preparación de datos

## 01 Importación de libreria

Importamos Pandas, Scikit, Numpy y otras

## 02 Cargar los datos la notebook

Caragamos el dataframe con pandas desde github

## 03 Forma de los datos

Inspeccionamos la forma de los datos

## 04 Información general de los datos

Analizamos cada variable y si tiene sentido



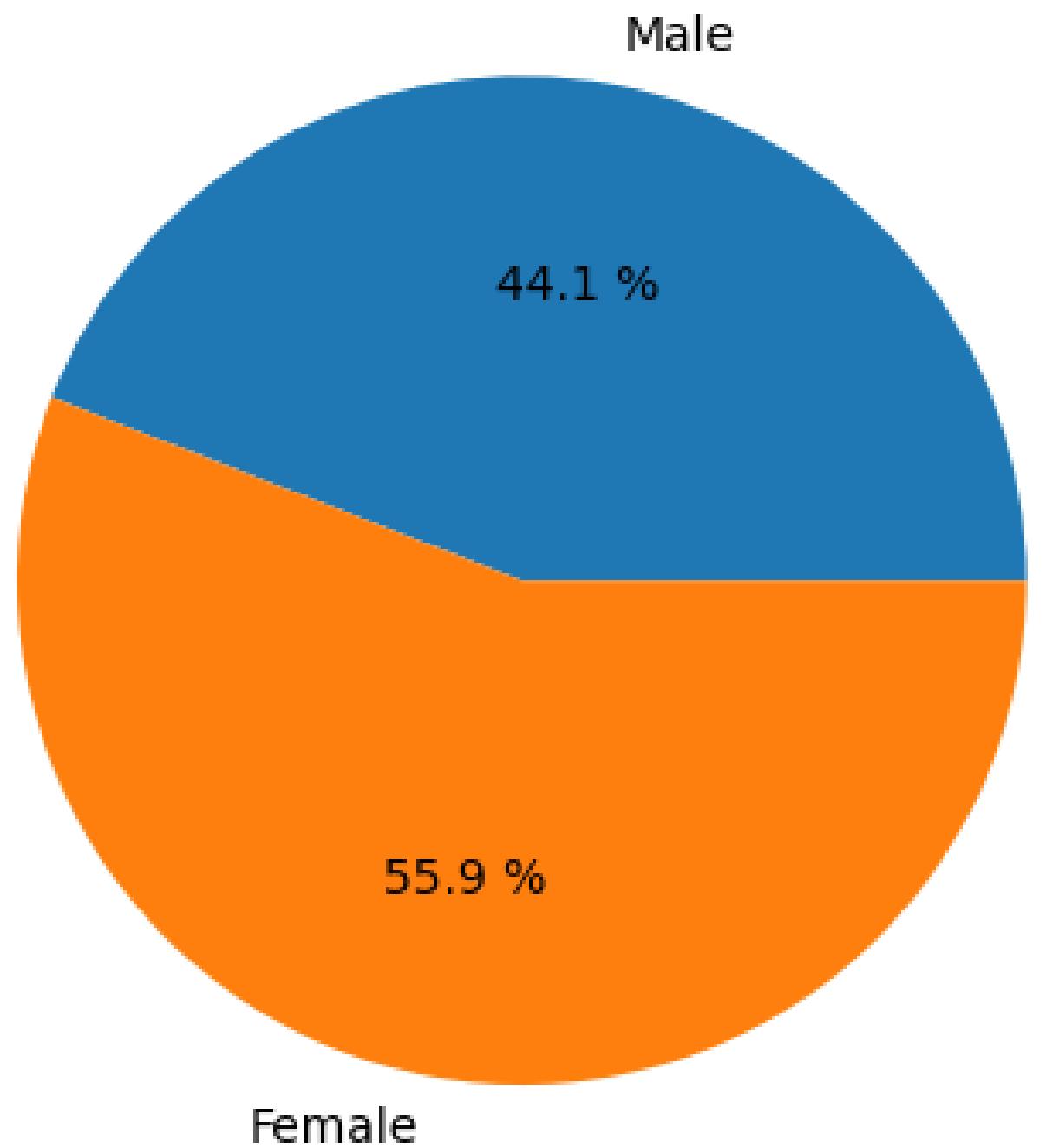
# Manipulación de datos

- 01 Tratando valores nulos**  
Chequeamos si existen y los eliminamos
- 02 Tratando valores NAN**  
Chequeamos si existen y los eliminamos
- 03 Tratando duplicados**  
Chequeamos si existen y los eliminamos
- 04 Creando dummies**  
Creamos dummies para Geography y Gender
- 05 Clasificar edad**  
Agrupamos la edad por categorías

# Análisis exploratorio

¿ CÓMO SE RELACIONAN LOS DATOS ?

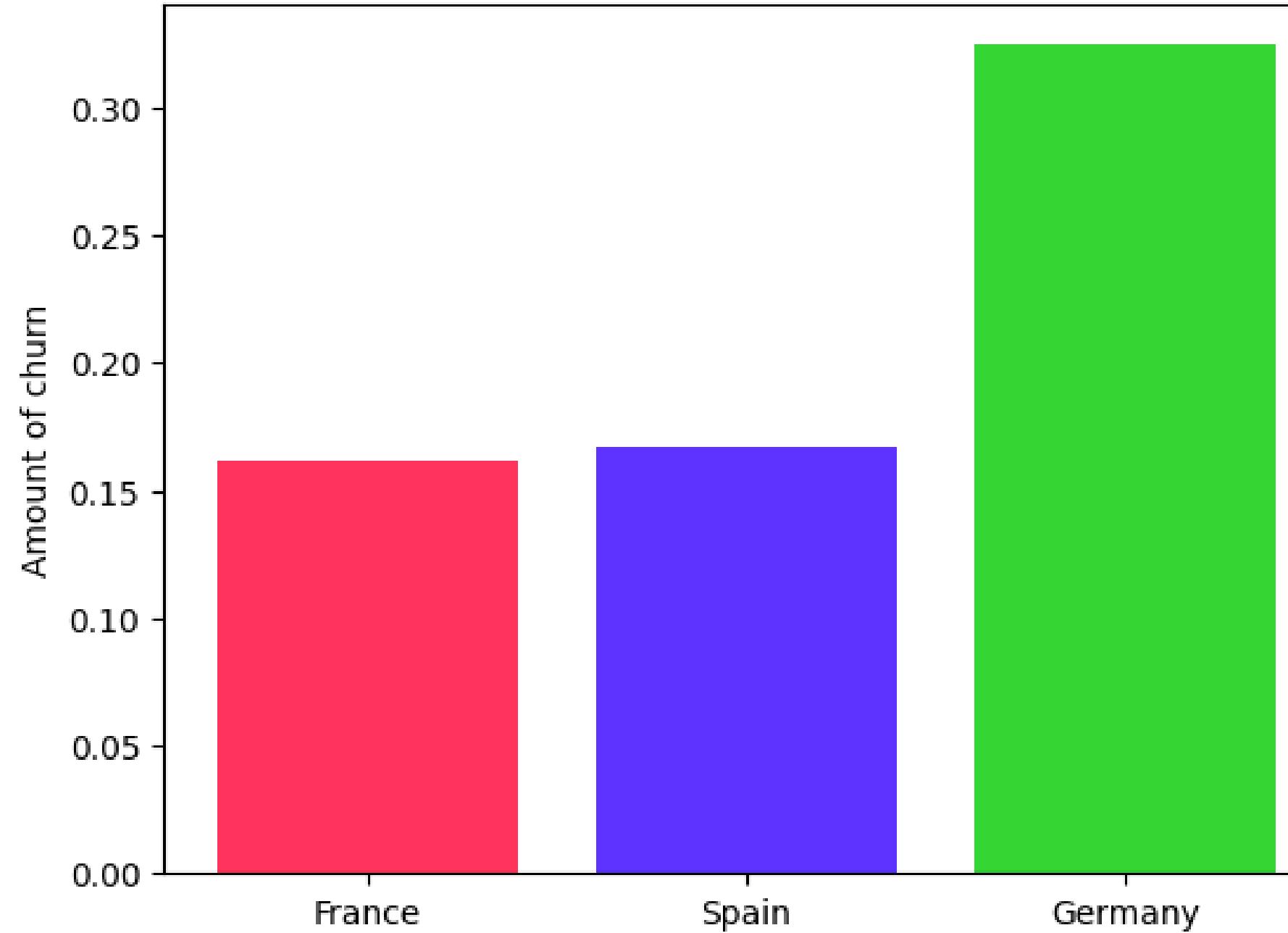
**Churn by gender Pie Chart**



## **Abandono por sexo**

Este gráfico de torta nos muestra el porcentaje de los clientes que han abandonado, según su género.

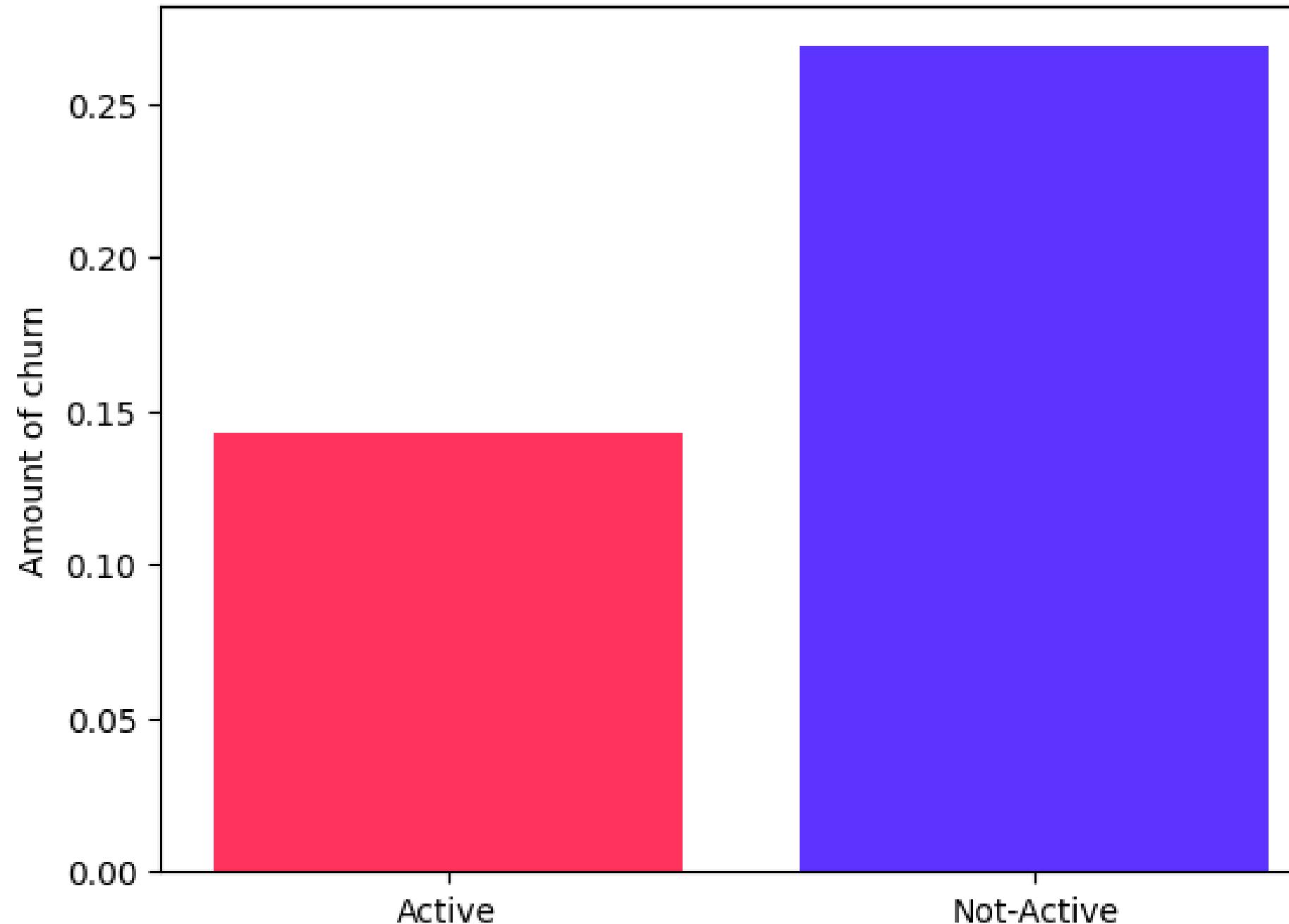
### Churn by country Bar Chart



## Abandono por país

Este gráfico de barras nos muestra la propensión a abandonar por país de residencia.

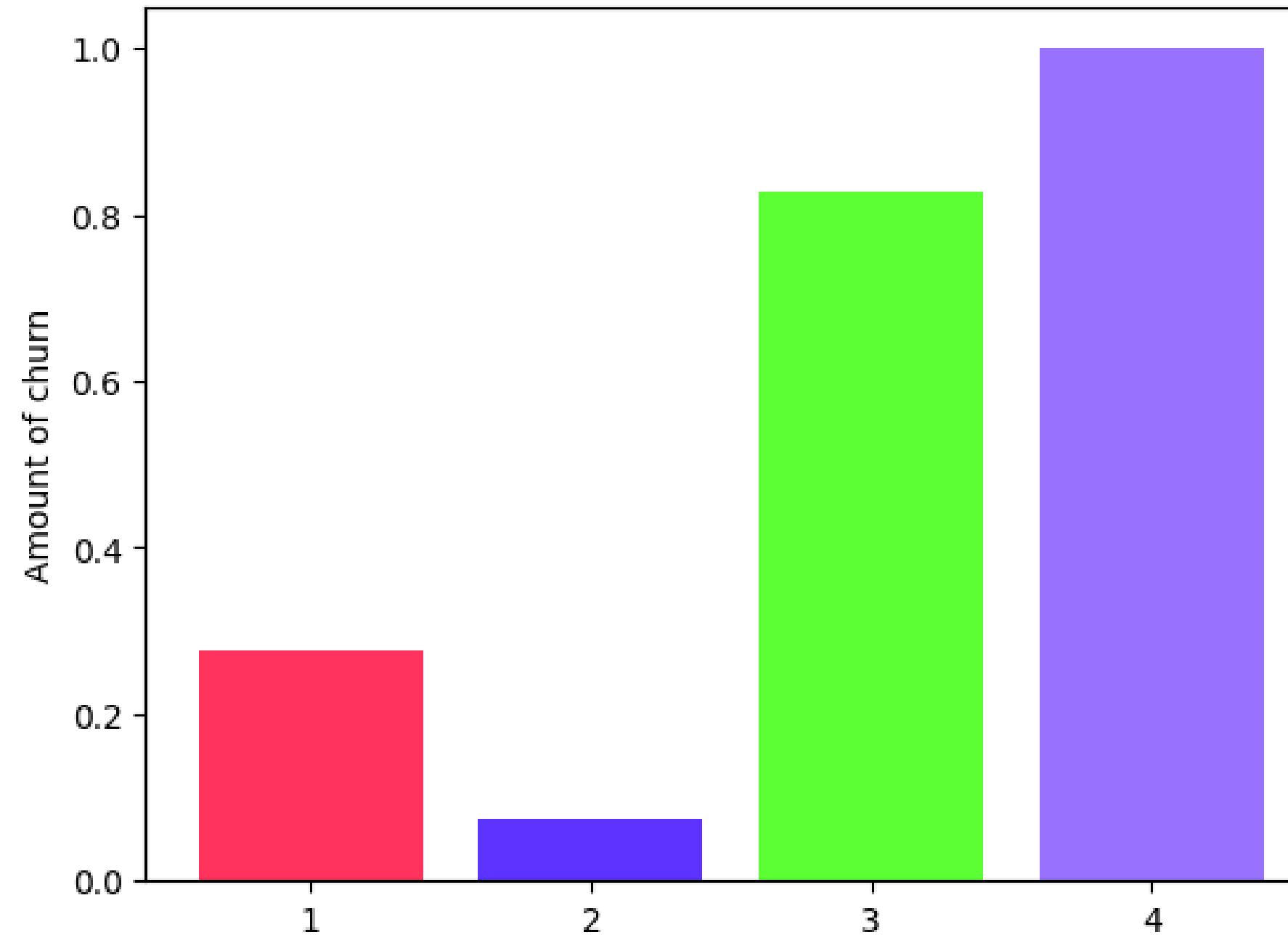
**Churn by active member Bar chart**



# Abandono por actividad

Este gráfico de barras nos muestra la propensión a abandonar según si son miembros activos o no.

**Churn by number of products Bar chart**

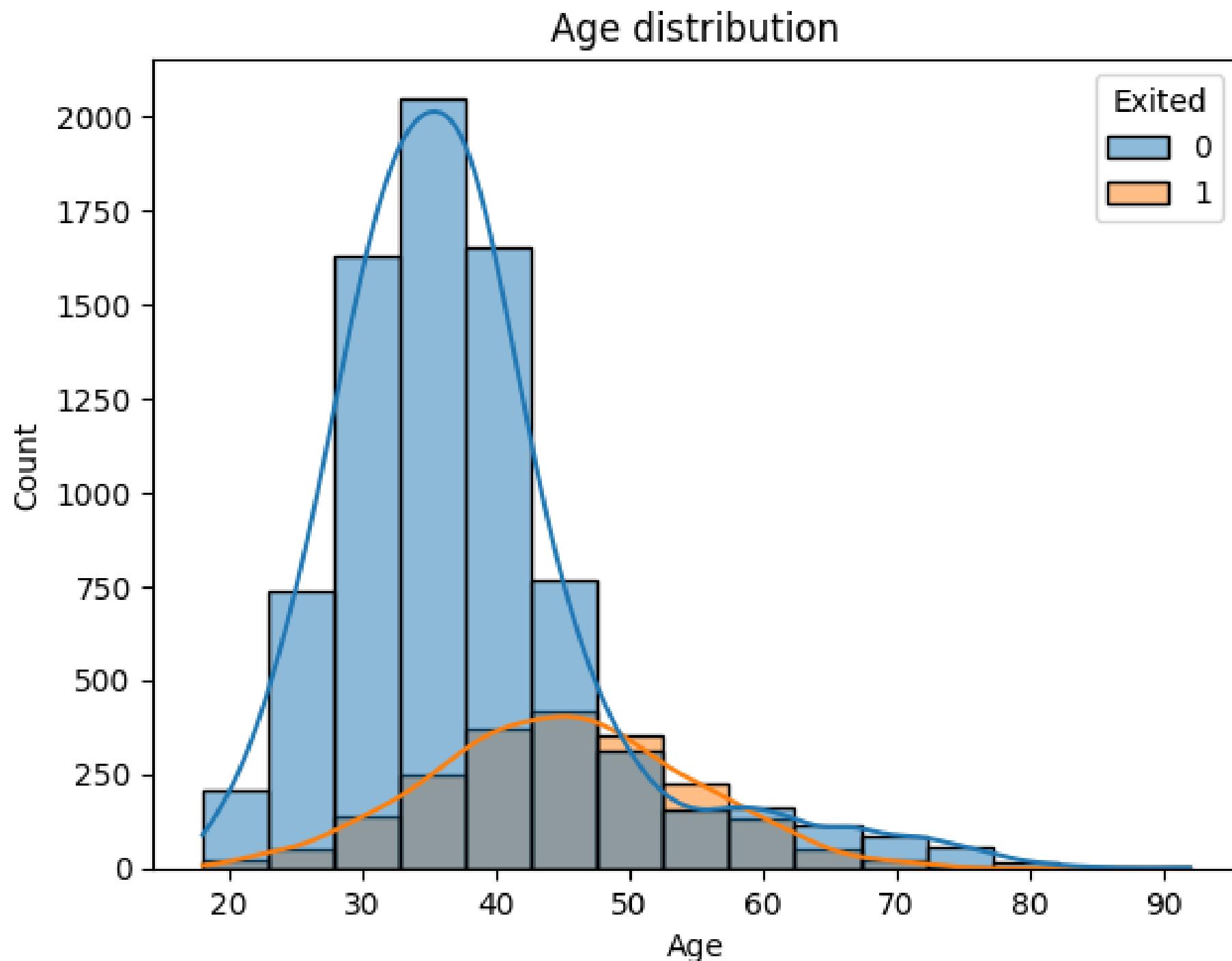


# Abandono por cant productos

Este gráfico de barras nos muestra la propensión a abandonar por número de productos.

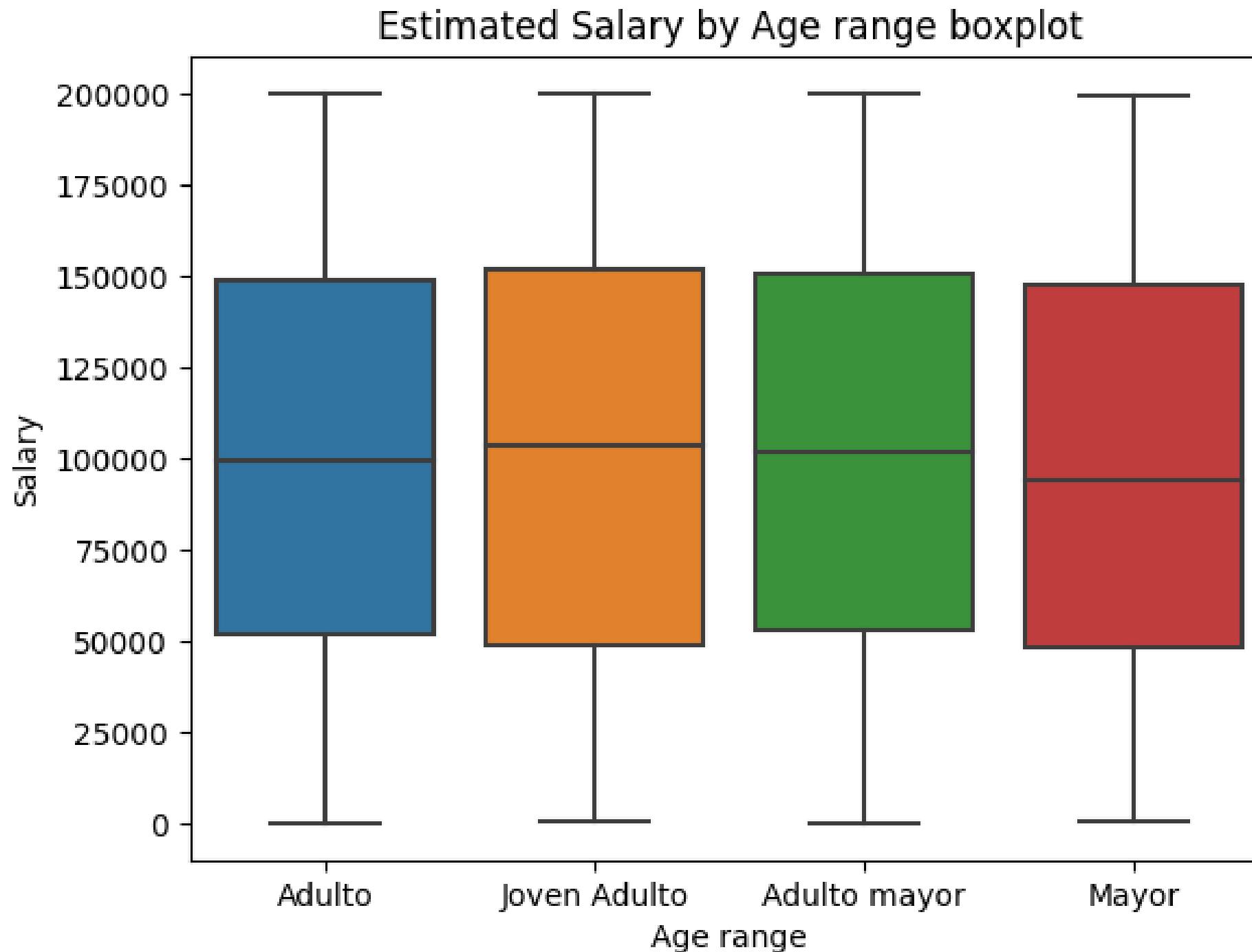
# Histograma de edad

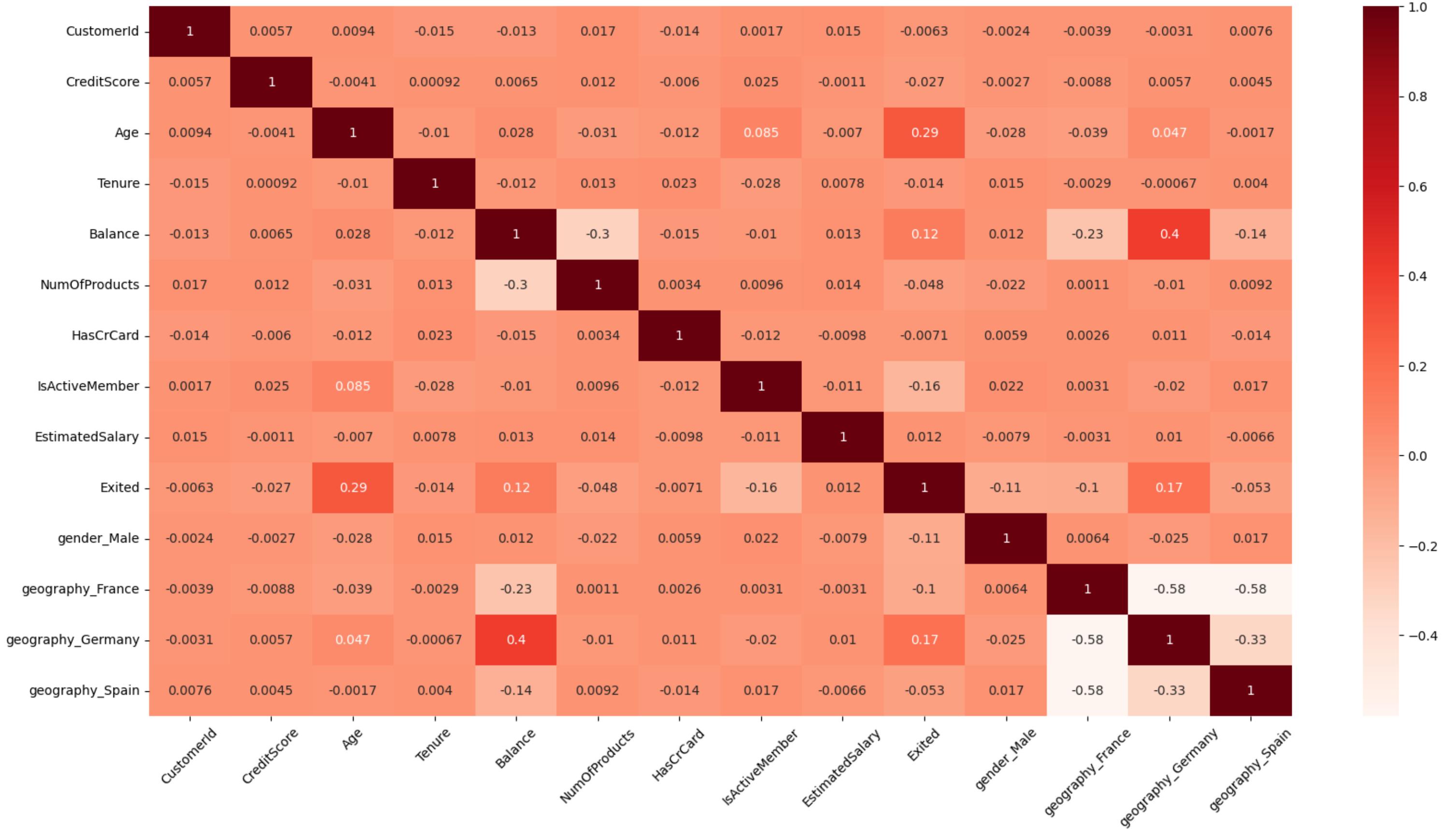
En el histograma de edad podemos visualizar la distribución, las frecuencias de edades y ver qué edades son más recurrentes en los clientes que otras.



# Salario por rango etario

En el boxplot de salario estimado podemos visualizar la comparación entre los diferentes salarios según el rango de edad.





# Modelado y evaluación

¿ CÓMO CLASIFICAMOS ?

# Modelos de clasificación

En nuestro caso vamos a utilizar dos modelos para comparar su desempeño

01

## Decision Tree

Modelo que toma decisiones basadas en reglas lógicas en forma de árbol.

02

## Random Forest

Conjunto de árboles de decisión que trabajan juntos para mejorar la precisión y evitar el sobreajuste.

# Preparación del modelo

Para preparar los datos para los modelos vamos a hacer un pretratamiento

01

## Definir variables

Definimos "Exited" como variable de interés (y), y todas las demás como variables independientes (x).

02

## Train y test

Separamos una porción de nuestro dataset que usaremos luego para testear, del resto del dataset que usaremos para entrenar. En nuestro caso, 30:70.

# Decision tree

El árbol se construye dividiendo los datos de entrenamiento en función de las reglas lógicas, lo que permite tomar decisiones basadas en las características de entrada.



## Ajuste y evaluación con random seed

Entrenamiento y evaluación del árbol de decisión con una semilla aleatoria establecida para resultados reproducibles.

**Recall 0.54**



## Ajuste y evaluación con cross validation

Entrenamiento y evaluación del árbol de decisión mediante validación cruzada para una evaluación más robusta.

**Recall 0.50**



## Ajuste de hiperparámetros y evaluación

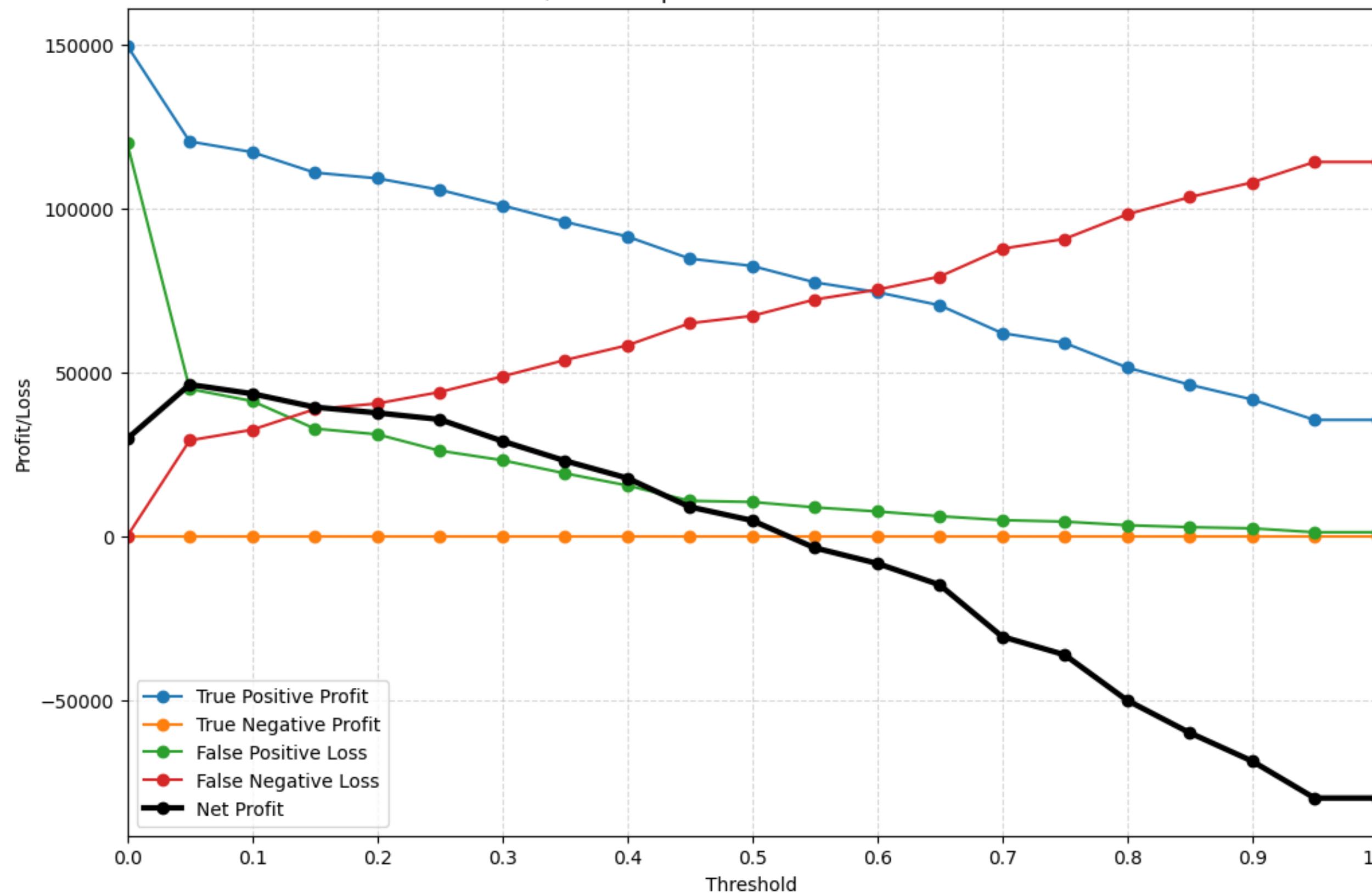
Ajuste de los parámetros configurables del árbol de decisión para mejorar su rendimiento mediante técnicas como la búsqueda en cuadrícula.

**Recall 0.51**



<b>threshold</b>	<b>precision</b>	<b>recall</b>	<b>Net Profit</b>
0.00	0.199733	1.000000	29750.0
0.05	0.349022	0.804674	46300.0
0.10	0.362442	0.782972	43500.0
0.15	0.403270	0.741235	39400.0
0.20	0.412653	0.729549	37650.0
0.25	0.447619	0.706177	35650.0
0.30	0.465438	0.674457	29050.0
0.35	0.500000	0.641068	23050.0
0.40	0.541420	0.611018	17750.0
0.45	0.610811	0.565943	8950.0
0.50	0.612245	0.550918	4800.0

Profit/Loss Comparison for Different Thresholds



# Random forest

El Random Forest es un conjunto de árboles de decisión. El Random Forest utiliza múltiples árboles y combina sus resultados para obtener una predicción más precisa y robusta.



## Ajuste y evaluación con random seed

Entrenamiento y evaluación del Random forest con una semilla aleatoria establecida para resultados reproducibles.

**Recall 0.49**



## Ajuste y evaluación con cross validation

Entrenamiento y evaluación del Random Forest mediante validación cruzada para una evaluación más robusta.

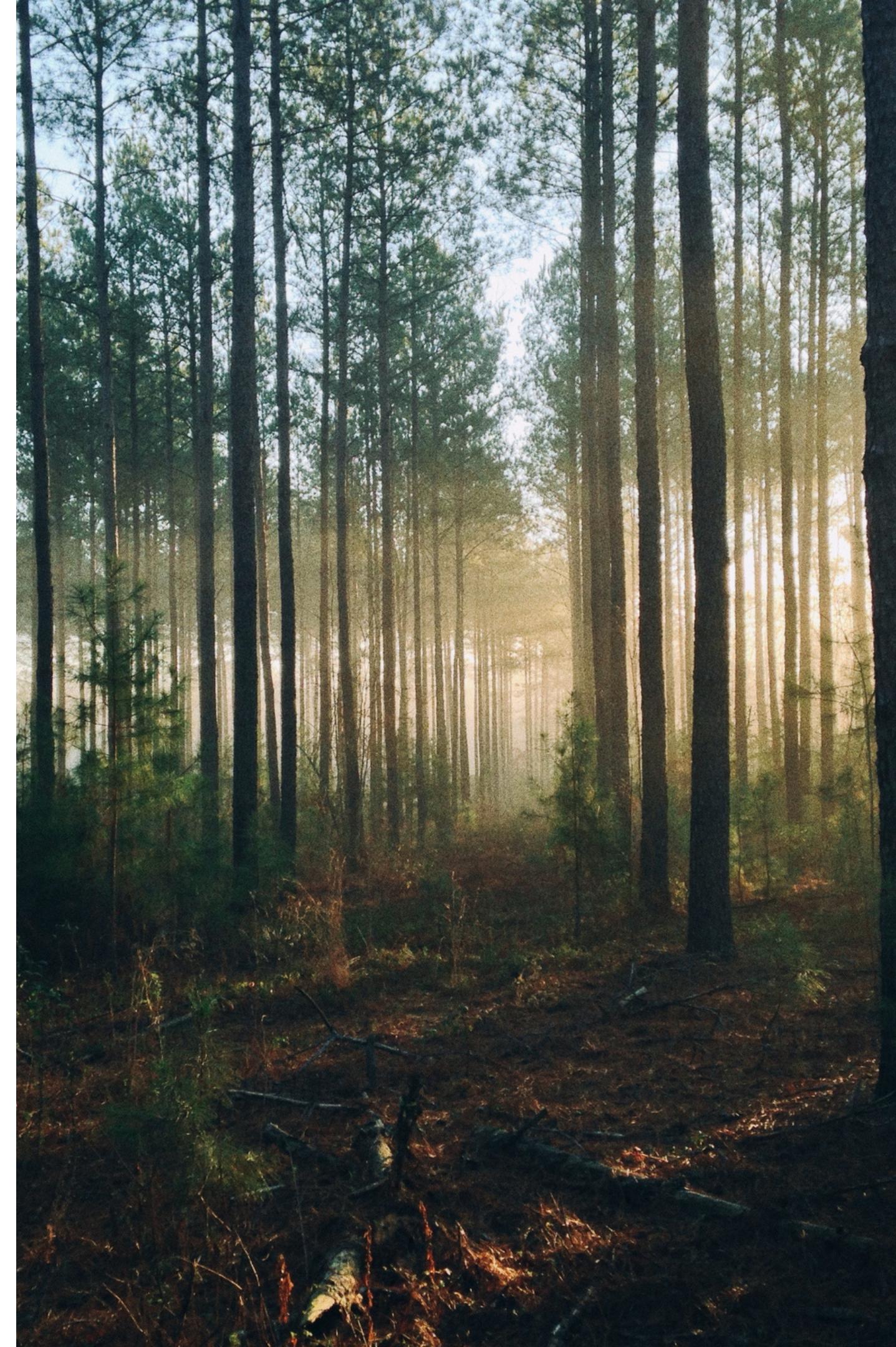
**Recall 0.46**

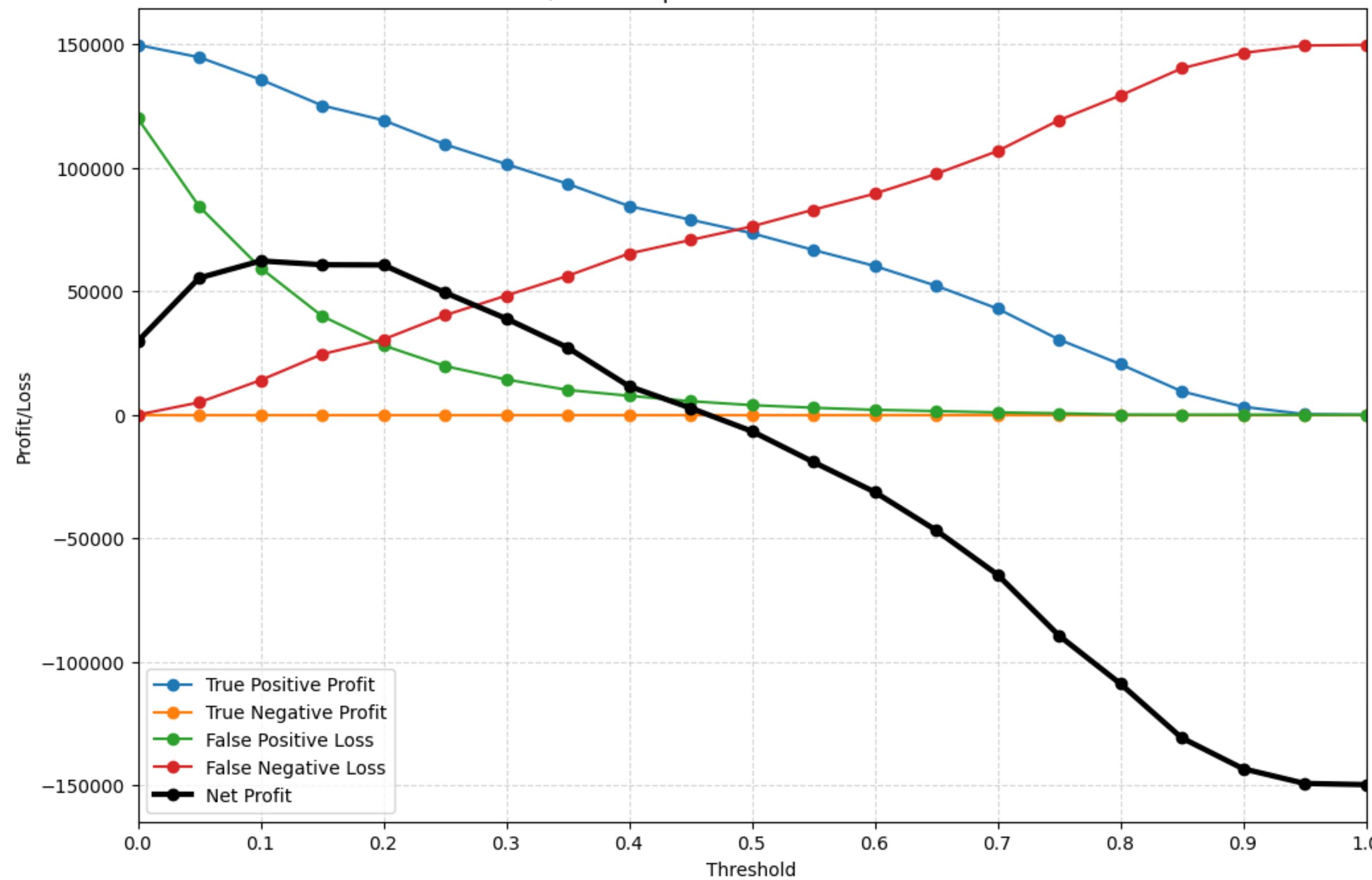


## Ajuste de hiperparámetros y evaluación

Ajuste de los parámetros configurables del Random forest para mejorar su rendimiento mediante técnicas como la búsqueda en cuadrícula.

**Recall 0.46**

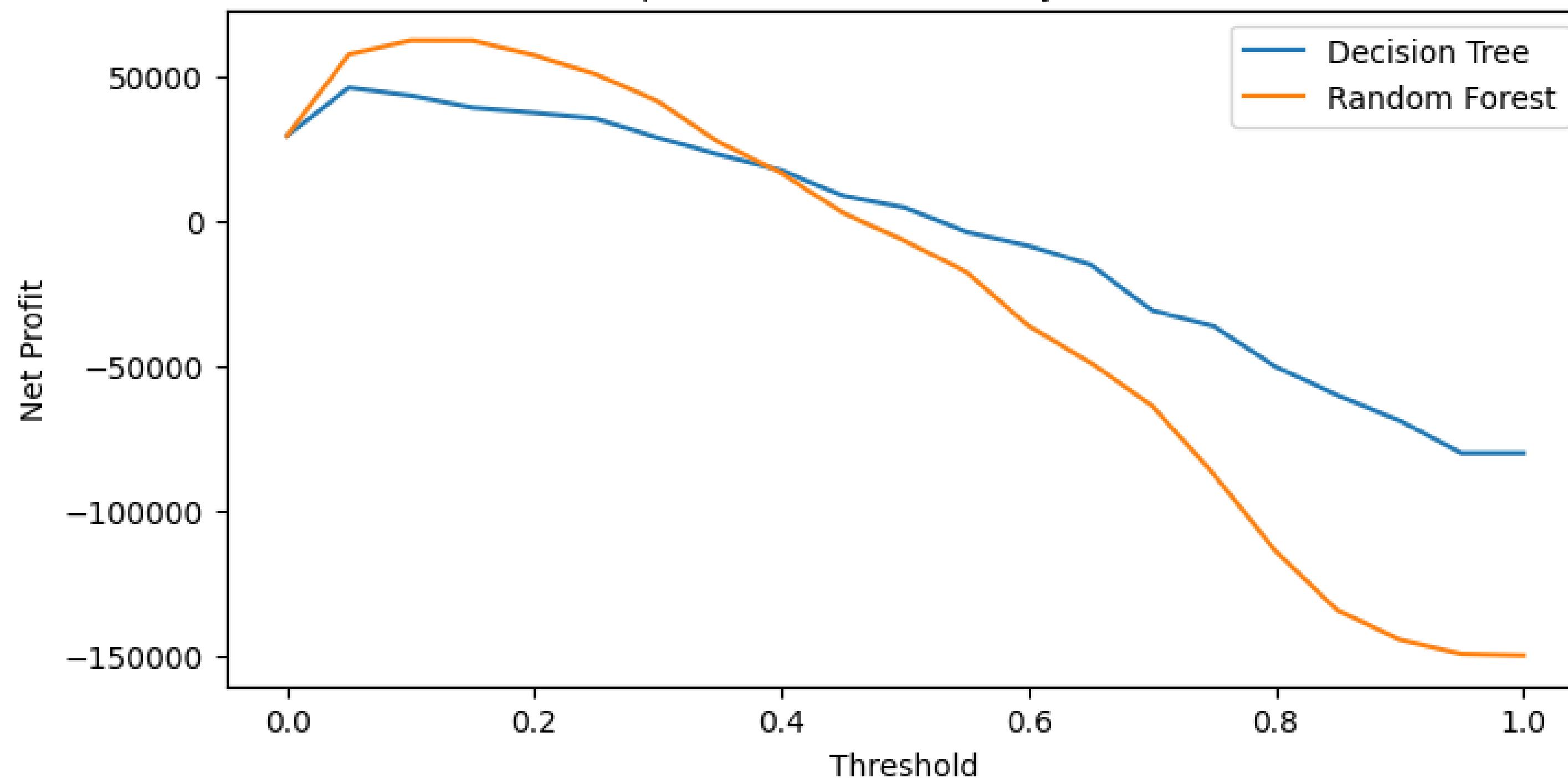


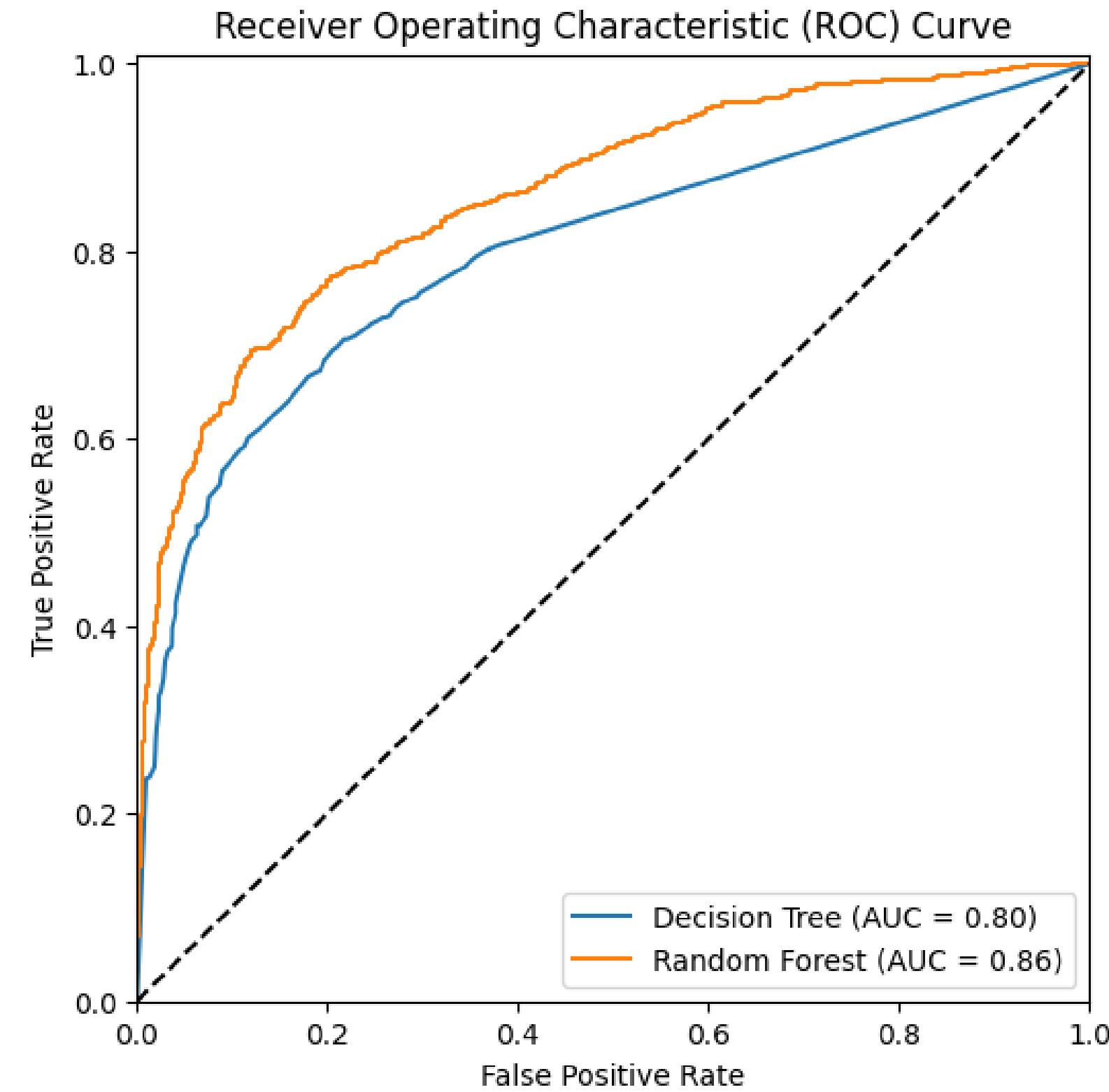


# Comparación de modelos

¿ CUÁL ES MEJOR ?

### Comparison of Net Profit by Threshold





### Comparación de Importancia de Características

