Fundamentos de Machine Learning e Modelos Supervisionados [24E2_3]

Regressão Linear, Regressão Logística,
Árvore de Decisão e Rede Neural

Grupo 3

Diego dos Santos

Filipe Rodrigues

Leandro Ribeiro

Rafael Diniz

Business Understanding

O Banco **Santander** está interessado em entender melhor o perfil de seus clientes para um determinado nicho através da Ciência de Dados.

Para isso, elaborou a seguinte tarefa a ser realizada:

- Identificar e reter clientes propensos a sair.

Referências Bibliográficas

Base de Dados Churn de Clientes:

https://www.kaggle.com/datasets/santoshd3/bank-customers/code?datasetId=35847&sortBy=voteCount

Objetivo dos Modelos

- Regressão Linear: prever a taxa de saída de clientes com base nas características dos mesmos.
- Regressão Logística: Prever a probabilidade de churn.
- Árvore de Decisão: Classificação de clientes em churn ou não churn.
- Rede Neural: Modelagem complexa para previsão de churn com maior precisão.

Data Understanding

A base de dados selecionada é o ***Churn.csv***, que contém dados sobre 10.000 clientes de um banco, com as informações abaixo.

Indicadores:

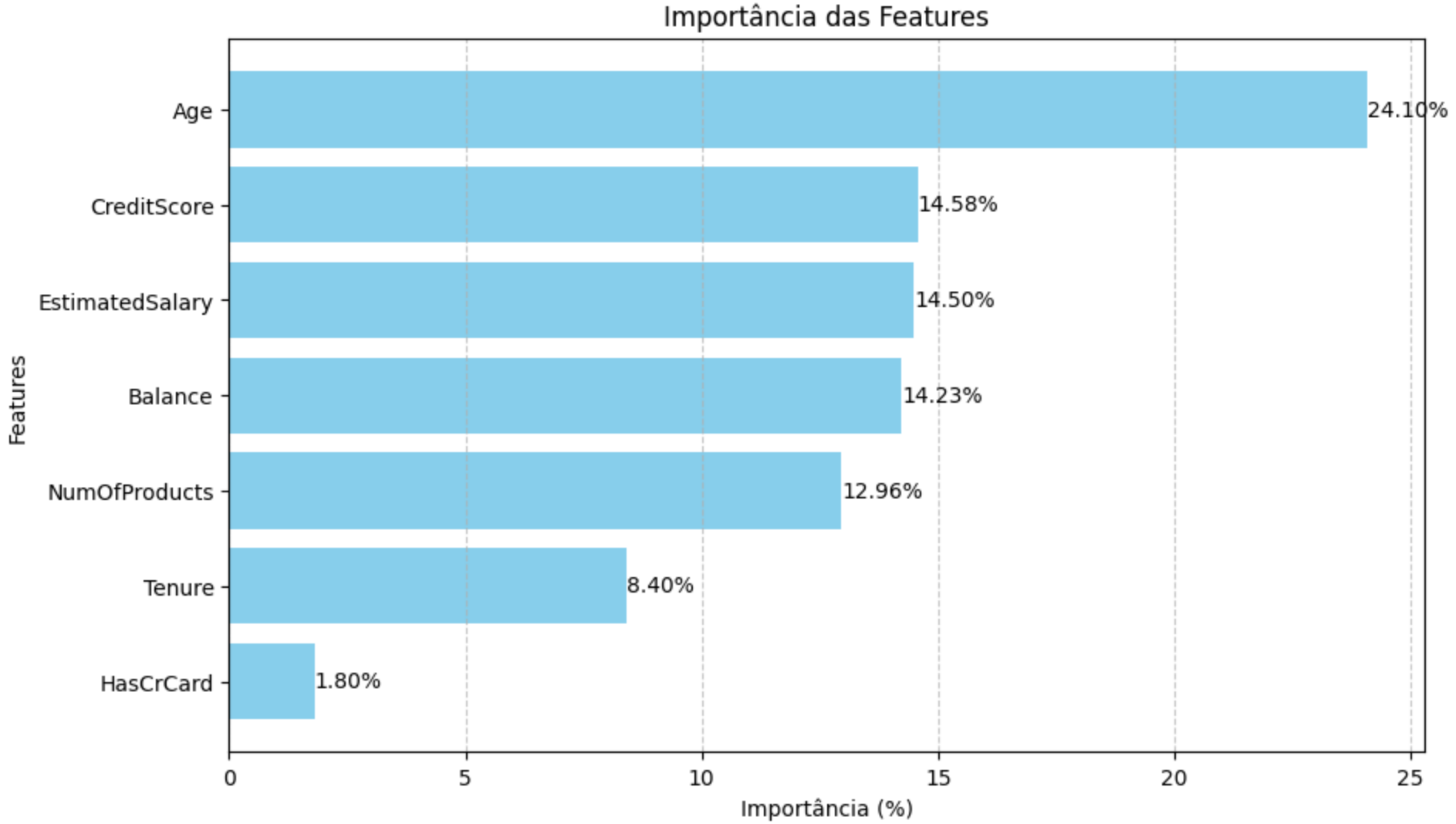
- País de residência, gênero, idade, tempo de conta, pontuação de crédito, saldo em conta, número de produtos, posse de cartão de crédito, membro ativo, salário estimado e encerramento de conta.
- Cada linha da base de dados representa um cliente com os indicadores acima.
- Total de linhas: 10.000

Variáveis Aleatórias

Colunas	Tipo	Descrição
RowNumber	Integer	Sequencial de linha
CustomerId	Integer	ID do cliente
Surname	String	Sobrenome
CreditScore	Integer	Pontuação de crédito
Geography	String	País de residência
Gender	String	Gênero
Age	Integer	Idade
Tenure	Integer	Tempo de conta
Balance	Float	Saldo em conta
NumOfProducts	Integer	Número de produtos
HasCrCard	SmallInt	Posse de cartão de crédito
IsActiveMember	SmallInt	Membro ativo
EstimatedSalary	Float	Salário estimado
Exited	SmallInt	Encerramento de conta

Data Preparation

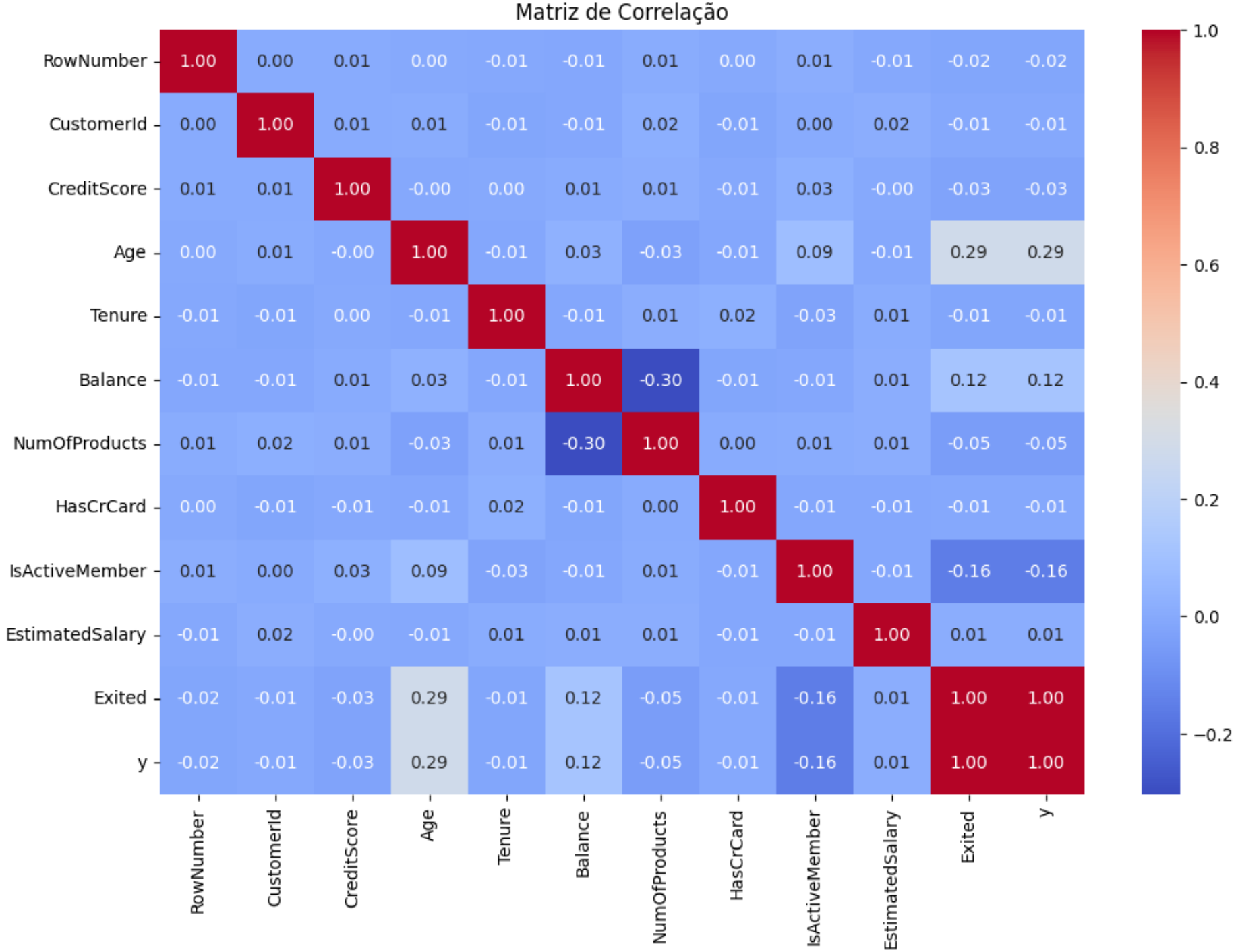
(VARIÁVEIS RELEVANTES)

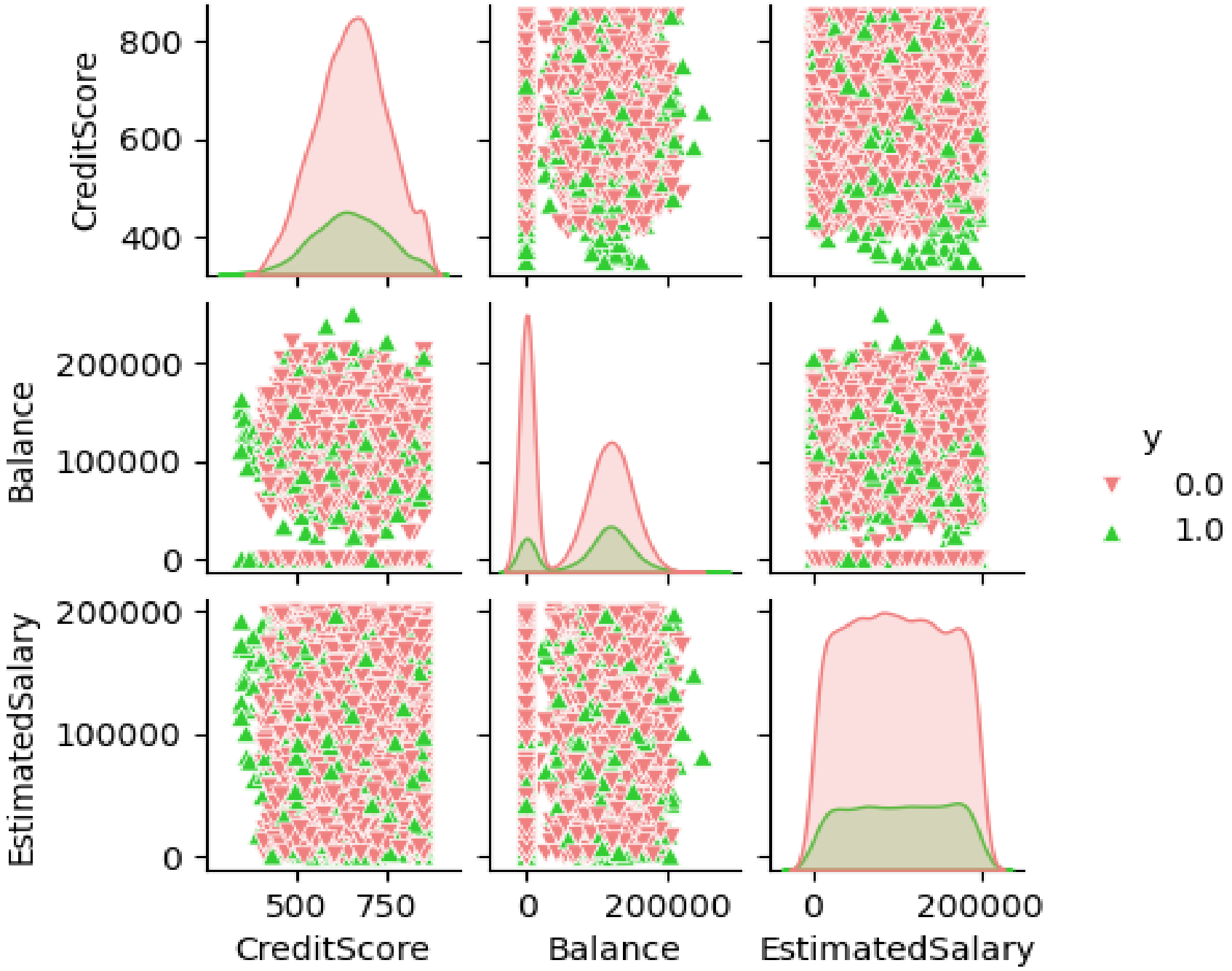


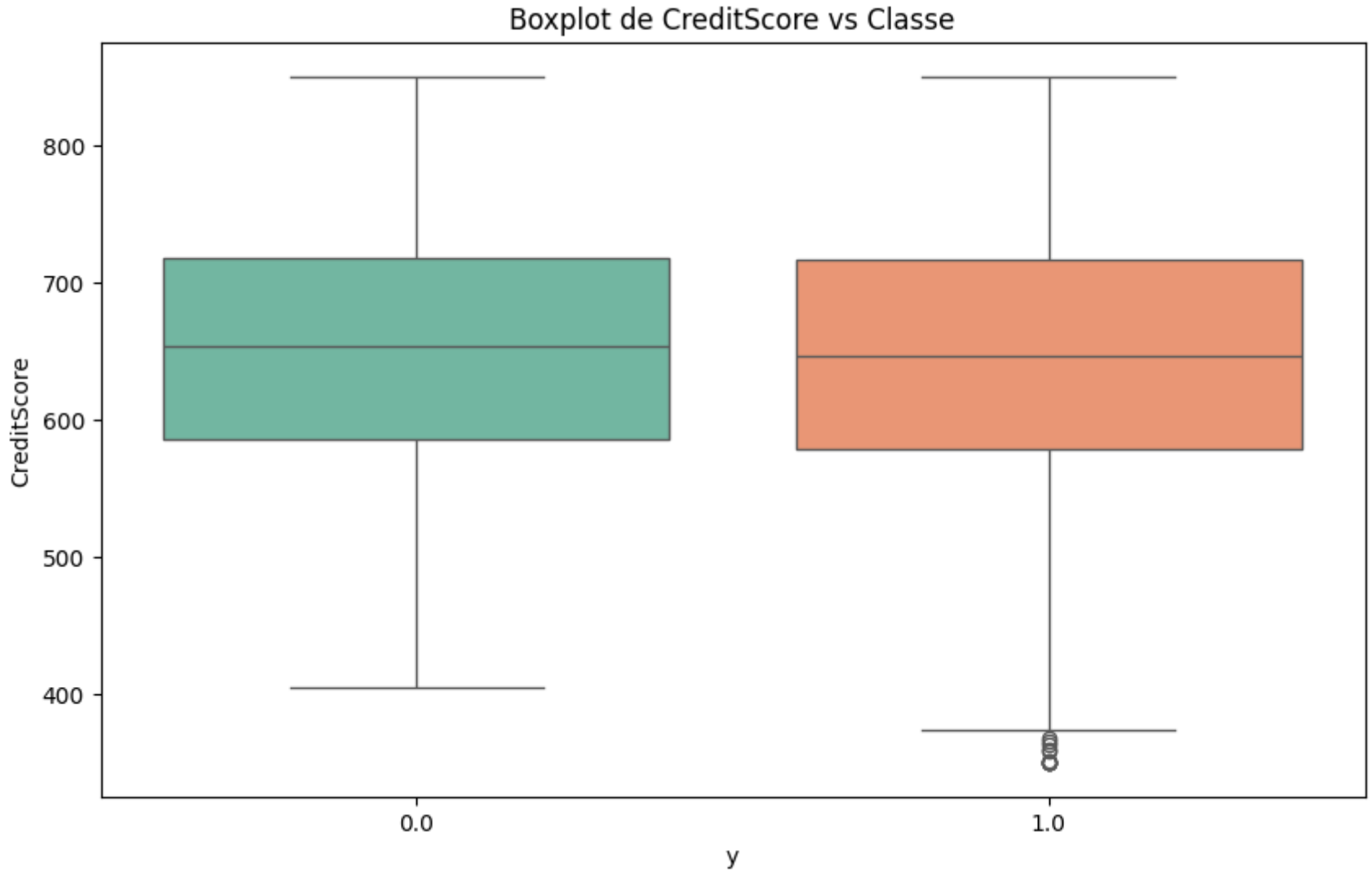
Data Preparation

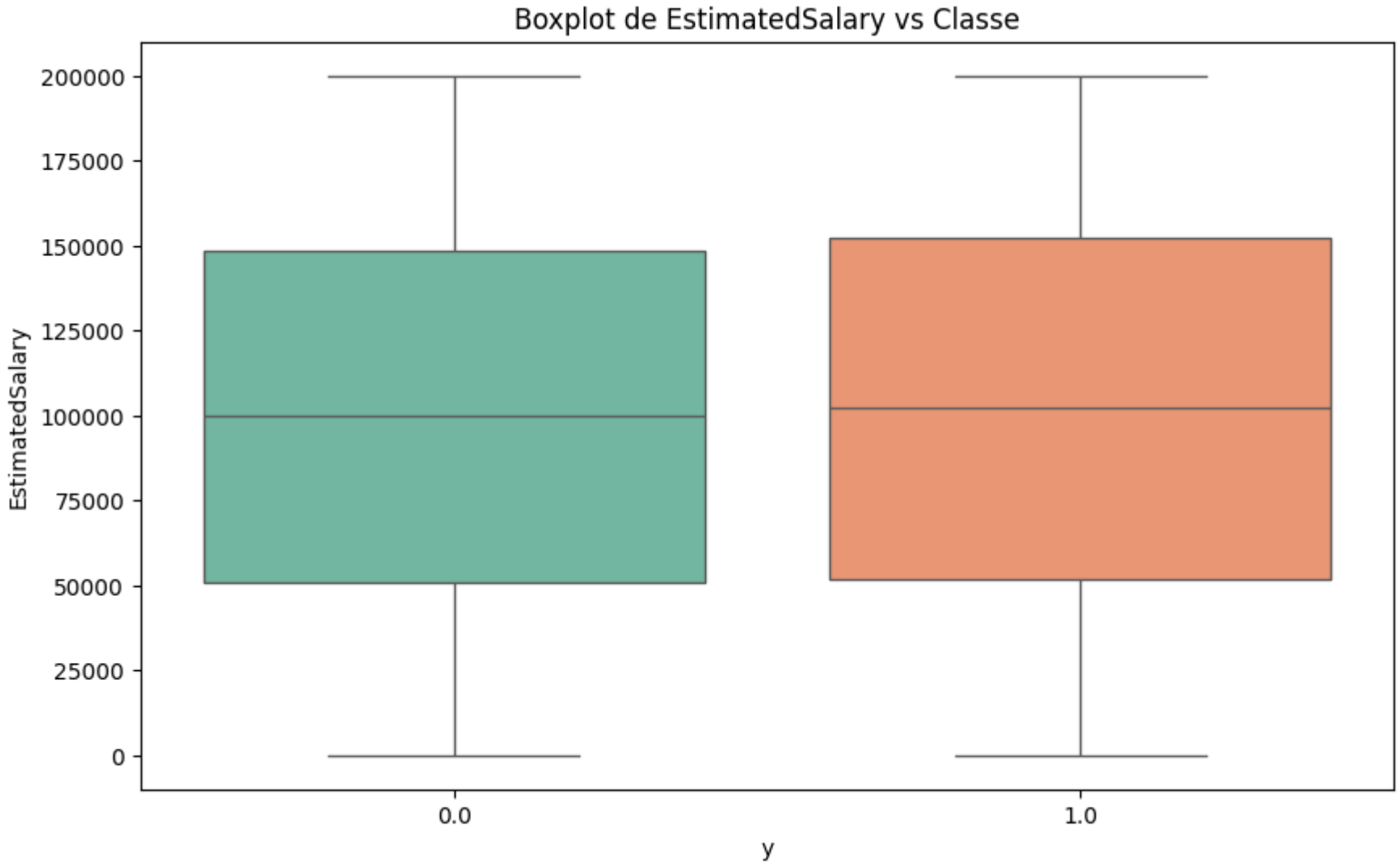
(TRANSFORMAÇÃO DOS DADOS)

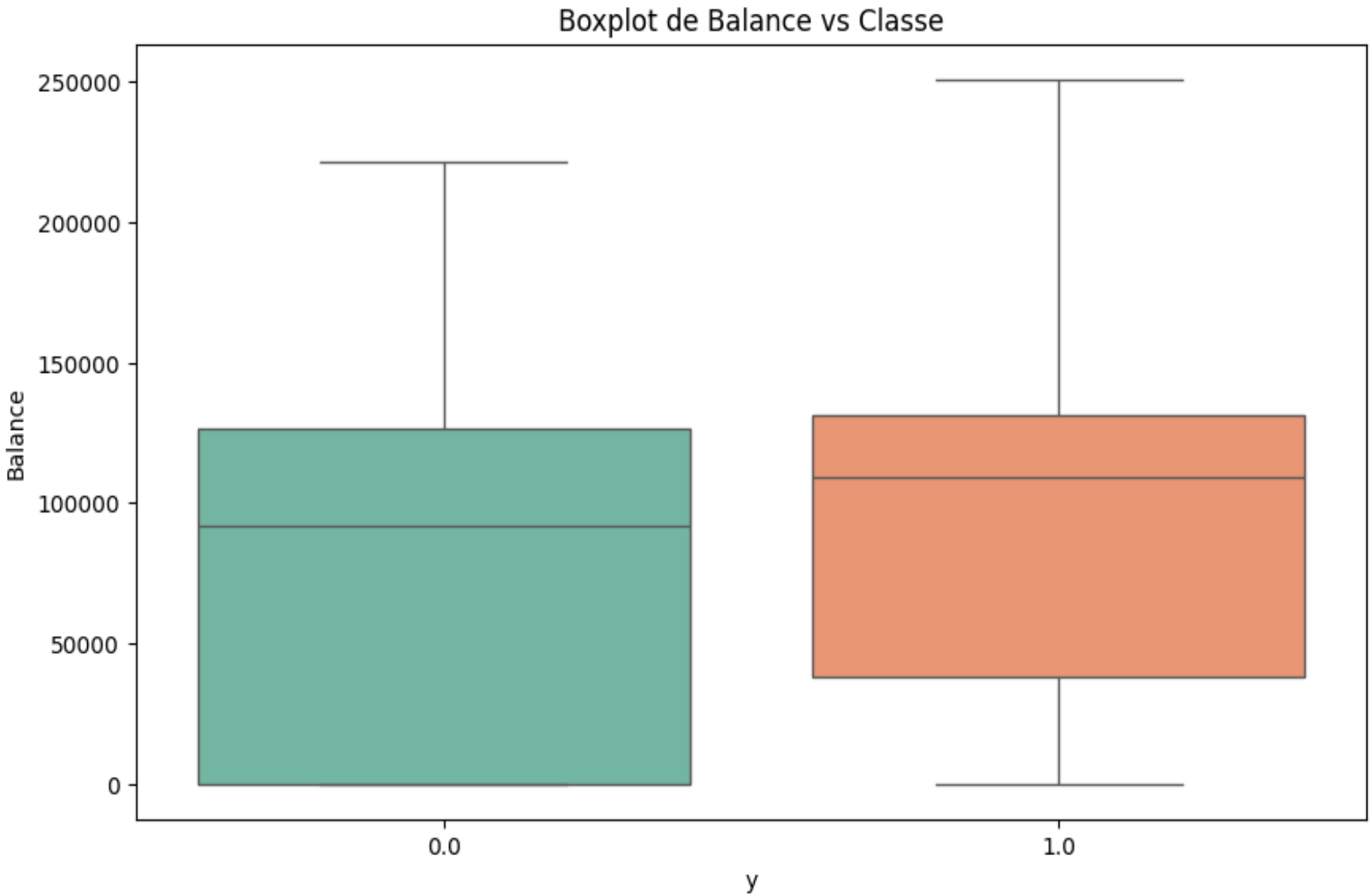
- Normalização dos dados.
- Conversão de variáveis categóricas em variáveis dummy.


Matriz de Correlação
RowNumber 1.00 0.00 0.01 0.00 -0.01 -0.01 0.01 0.00 0.01 -0.01 -0.02 -0.02
CustomerId 0.00 1.00 0.01 0.01 -0.01 -0.01 0.02 -0.01 0.00 0.02 -0.01 -0.01
CreditScore 0.01 0.01 1.00 -0.00 0.00 0.01 0.01 -0.01 0.03 -0.00 -0.03 -0.03
Age 0.00 0.01 -0.00 1.00 -0.01 0.03 -0.03 -0.01 0.09 -0.01 0.29 0.29
Tenure -0.01 -0.01 0.00 -0.01 1.00 -0.01 0.01 0.02 -0.03 0.01 -0.01 -0.01
Balance -0.01 -0.01 0.01 0.03 -0.01 1.00 -0.30 -0.01 -0.01 0.01 0.12 0.12
NumOfProducts 0.01 0.02 0.01 -0.03 0.01 -0.30 1.00 0.00 0.01 0.01 -0.05 -0.05
HasCrCard 0.00 -0.01 -0.01 -0.01 0.02 -0.01 0.00 1.00 -0.01 -0.01 -0.01 -0.01
IsActiveMember 0.01 0.00 0.03 0.09 -0.03 -0.01 0.01 -0.01 1.00 -0.01 -0.16 -0.16
EstimatedSalary -0.01 0.02 -0.00 -0.01 0.01 0.01 0.01 -0.01 -0.01 1.00 0.01 0.01
Exited -0.02 -0.01 -0.03 0.29 -0.01 0.12 -0.05 -0.01 -0.16 0.01 1.00 1.00
y -0.02 -0.01 -0.03 0.29 -0.01 0.12 -0.05 -0.01 -0.16 0.01 1.00 1.00
RowNumber
CustomerId
CreditScore
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
y
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


CreditScore
800
600
400
Balance
200000
100000
0
EstimatedSalary
200000
100000
0
500
750
CreditScore
0
200000
Balance
0
200000
EstimatedSalary
y
0.0
1.0


Boxplot de CreditScore vs Classe
CreditScore
800
700
600
500
400
0.0
1.0
y


Boxplot de EstimatedSalary vs Classe
EstimatedSalary
200000
175000
150000
125000
100000
75000
50000
25000
0
0.0
1.0
y


Boxplot de Balance vs Classe
250000
200000
150000
100000
50000
0
Balance
0.0
1.0
y

Procedimento de Validação Cruzada

- **K-Folds:** Utilizamos StratifiedKFold com n_splits=5 para realizar a validação cruzada estratificada, garantindo que a distribuição das classes fosse mantida em cada fold.

- **Semente Aleatória:** Utilizamos random_seed=42 para garantir a reprodutibilidade dos resultados.

Figura de Mérito

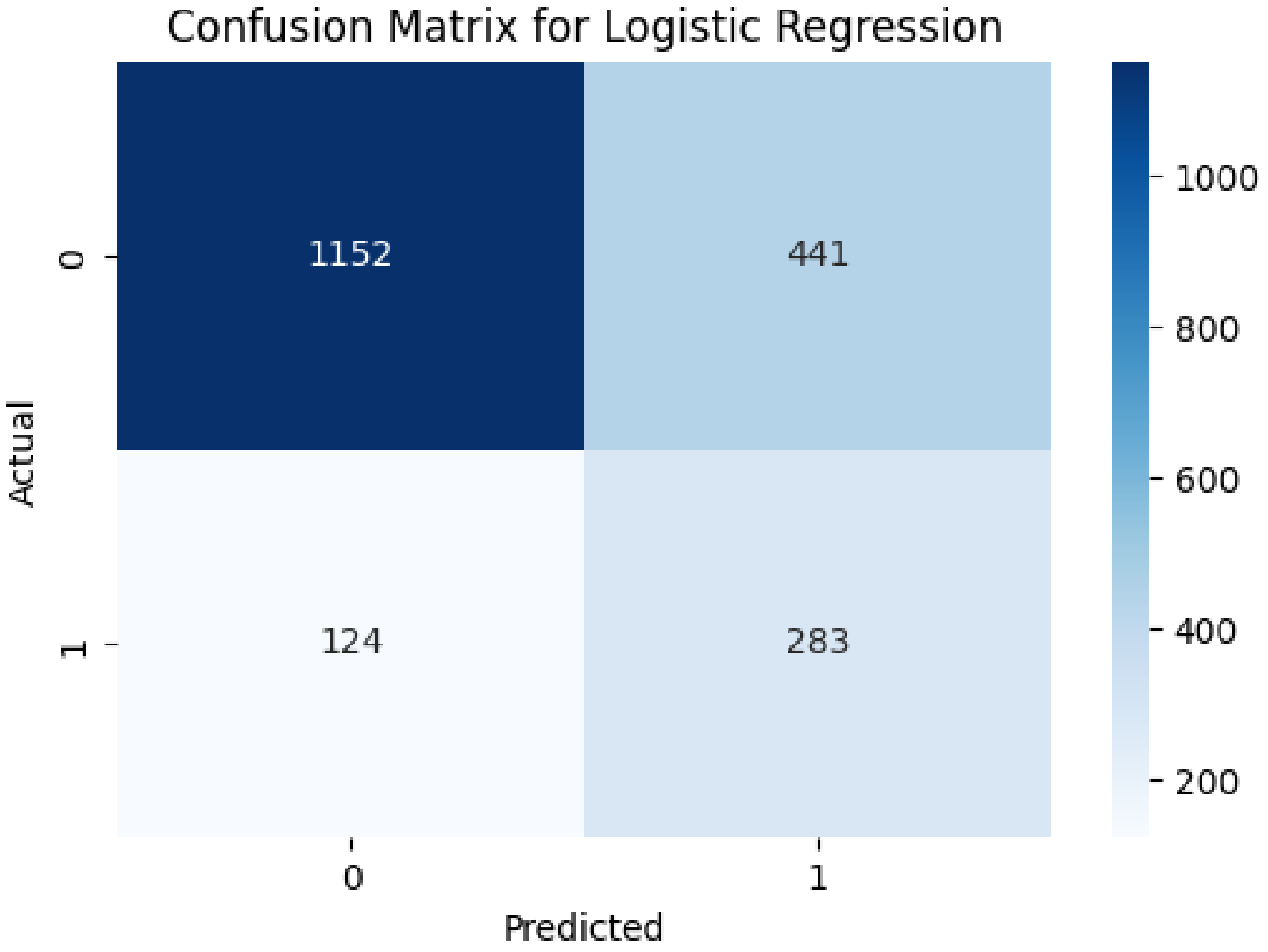
- Regressão Linear: Mean Squared Error (MSE)
- Regressão Logística: mean ROC AUC score
- Árvore de Decisão: mean ROC AUC score
- Rede Neural: mean ROC AUC score
- Random Forest: mean ROC AUC score

Generalização

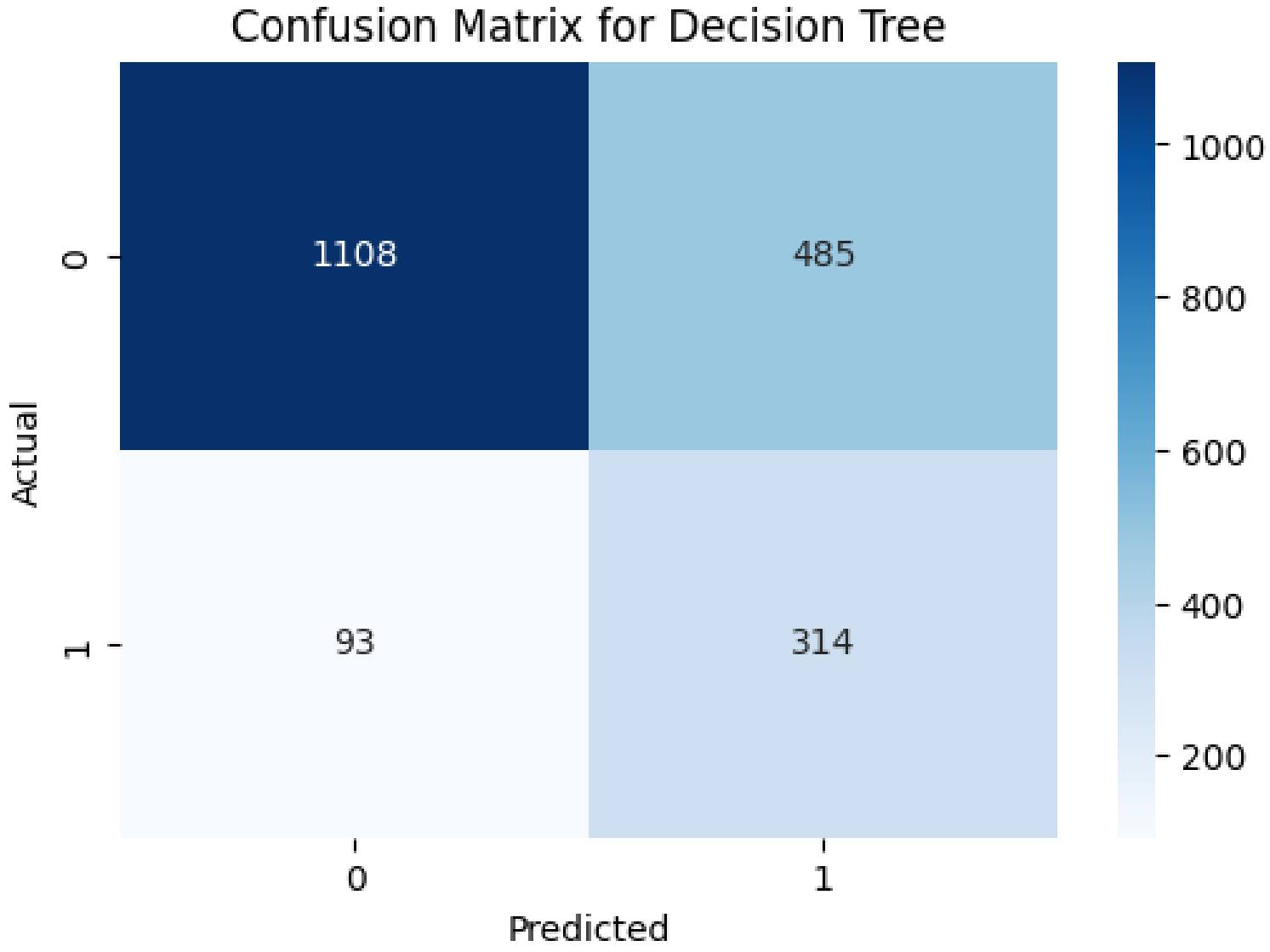
- **Validação Cruzada:** Garante que os modelos não estão sobreajustados aos dados de treinamento.

- **Evitar Overfitting:** Dividir os dados dessa forma ajuda a avaliar como o modelo se generaliza para novos dados não vistos durante o treinamento. Ao treinar o modelo em 80% dos dados e testá-lo nos 20% restantes, podemos ter uma estimativa mais precisa do desempenho do modelo em um cenário real, evitando o overfitting.

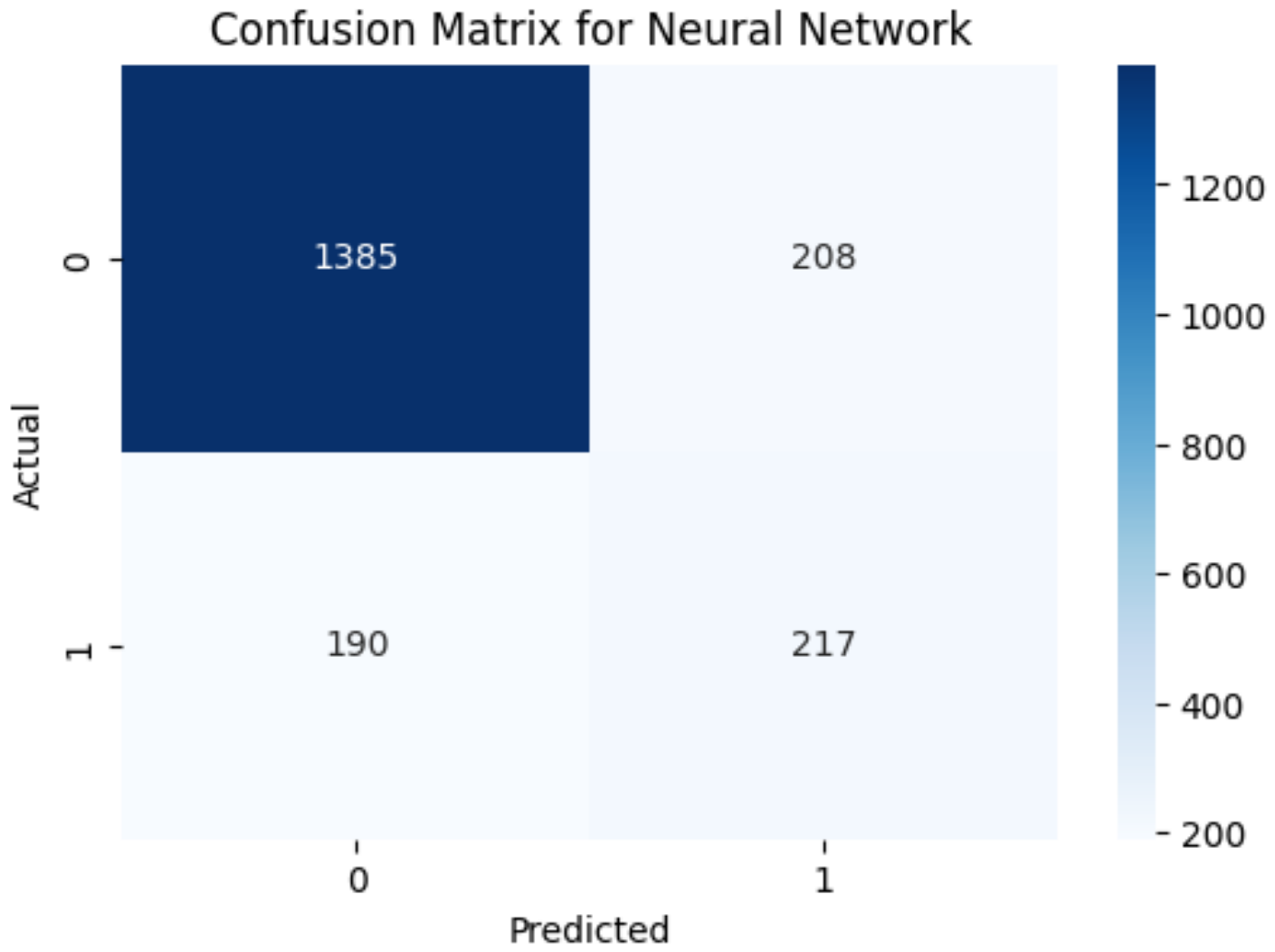
Logistic Regression Accuracy: 71,75%



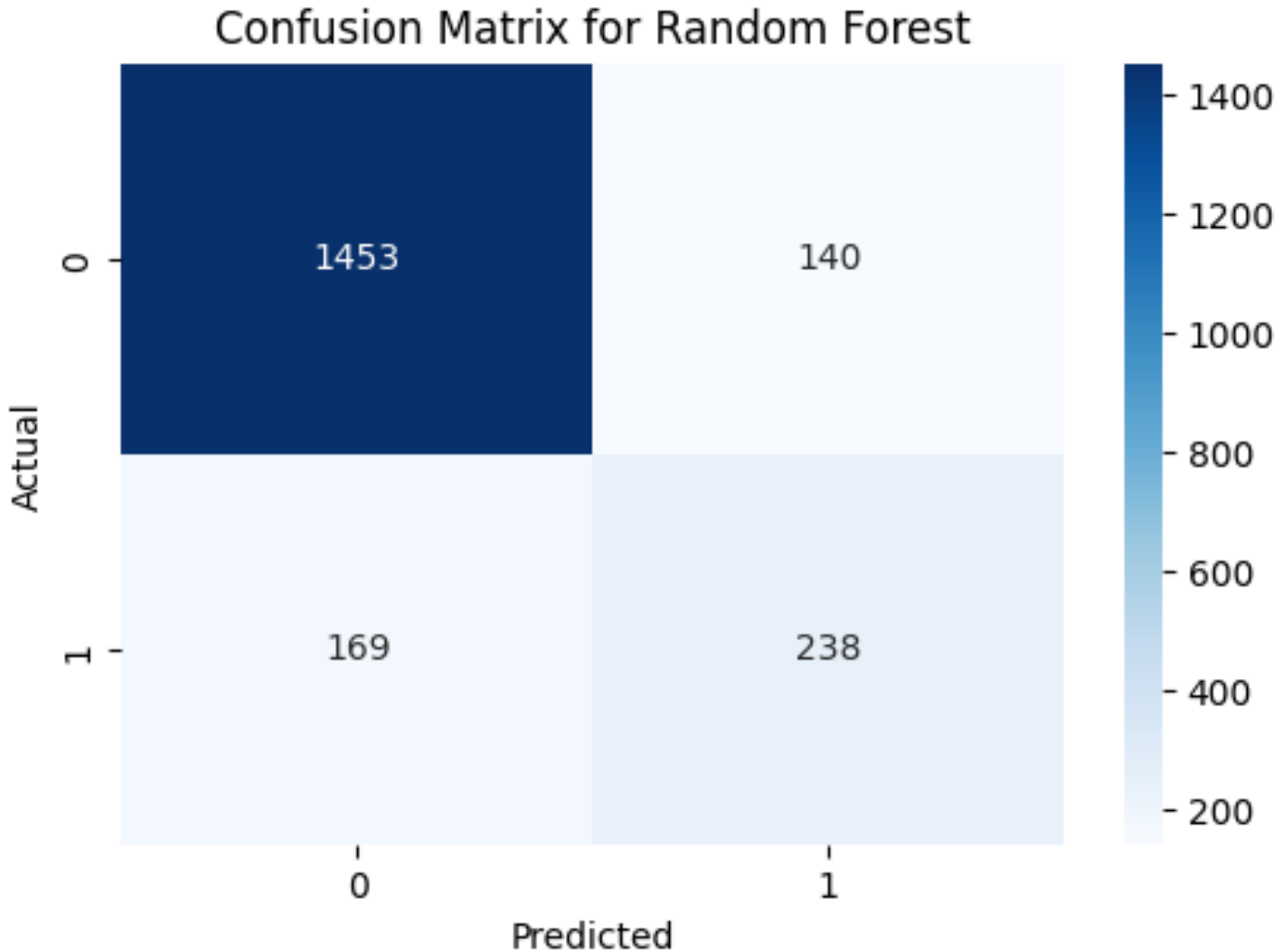
Decision Tree Accuracy: 71,10%



Neural Network Accuracy: 80,01%



Random Forest Accuracy: 84,55%



Resultado dos Modelos

ROC AUC (Validação Cruzada)

- Linear Regression:

  Mean score = 0.1962, Std = 0.0041

- Logistic Regression:

  Mean score = 0.7689, Std = 0.0145

- Decision Tree:

  Mean score = 0.8136, Std = 0.0081

- Neural Network:

  Mean score = 0.7873, Std = 0.0101

- Random Forest:

  Mean score = 0.8471, Std = 0.0054

Resultado dos Modelos

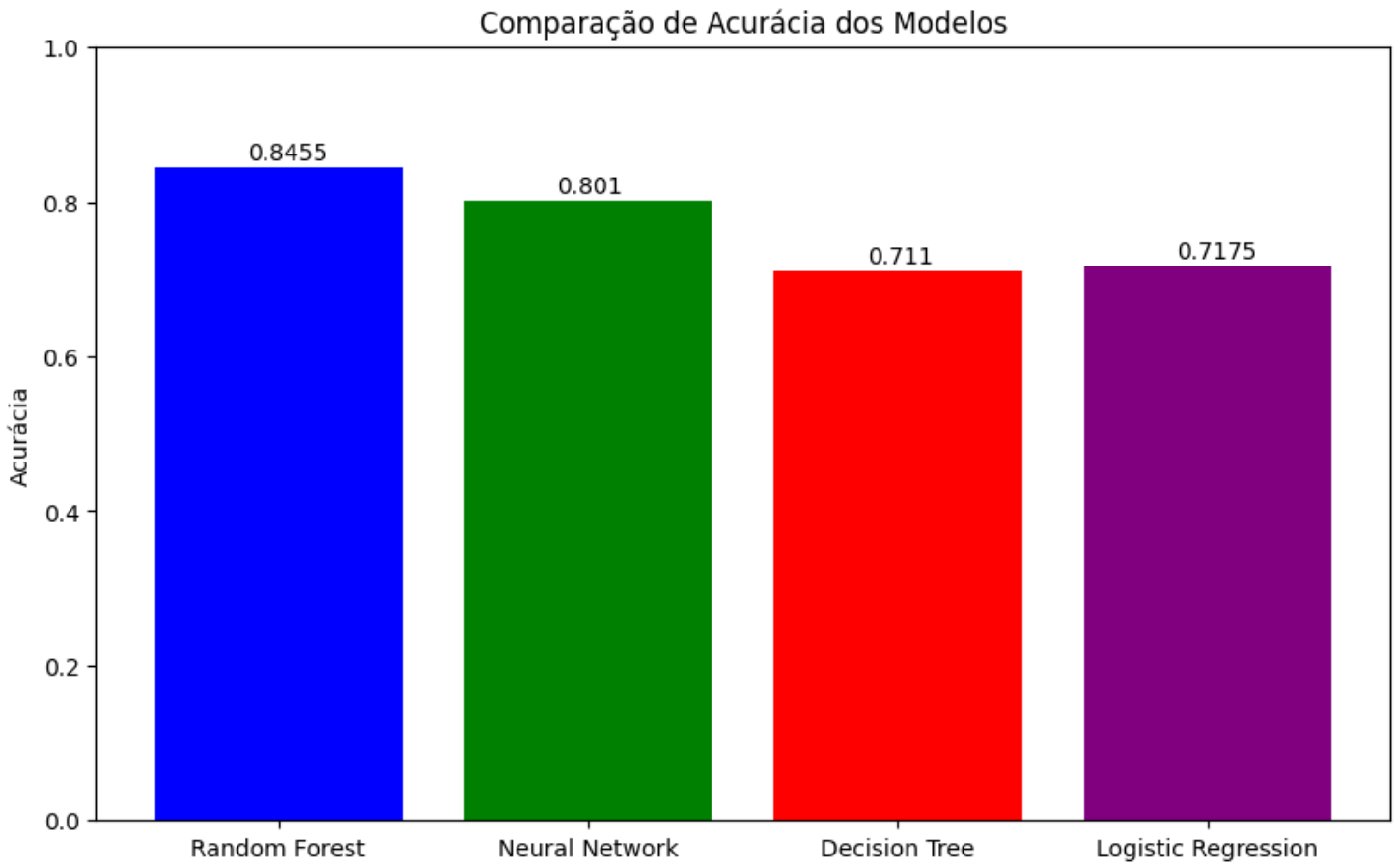
ROC AUC (Conjunto de Testes)

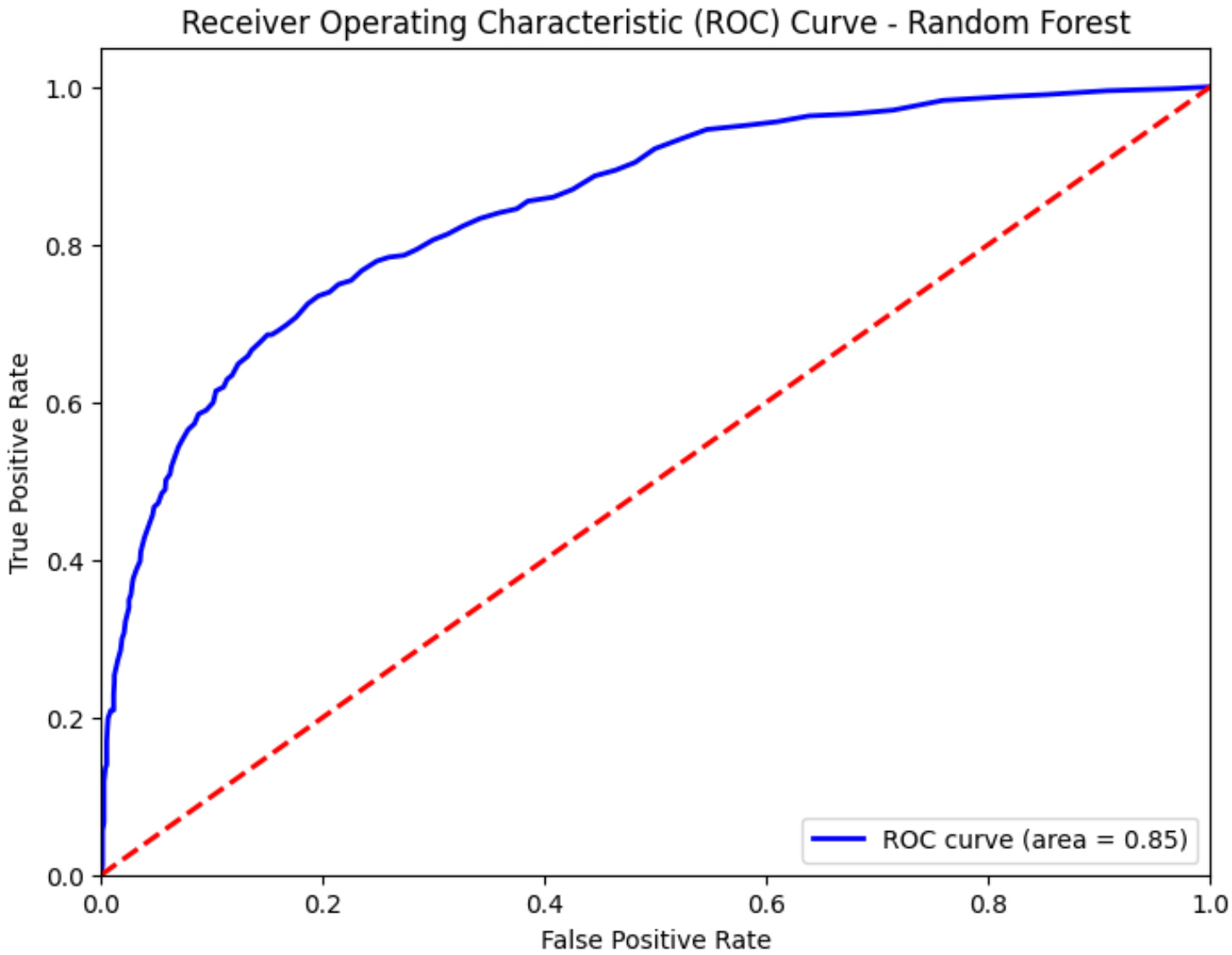
- Logistic Regression ROC AUC: 0.7754
- Decision Tree ROC AUC: 0.8107
- Neural Network ROC AUC: 0.7931
- Random Forest ROC AUC: 0.8479

Melhor Modelo

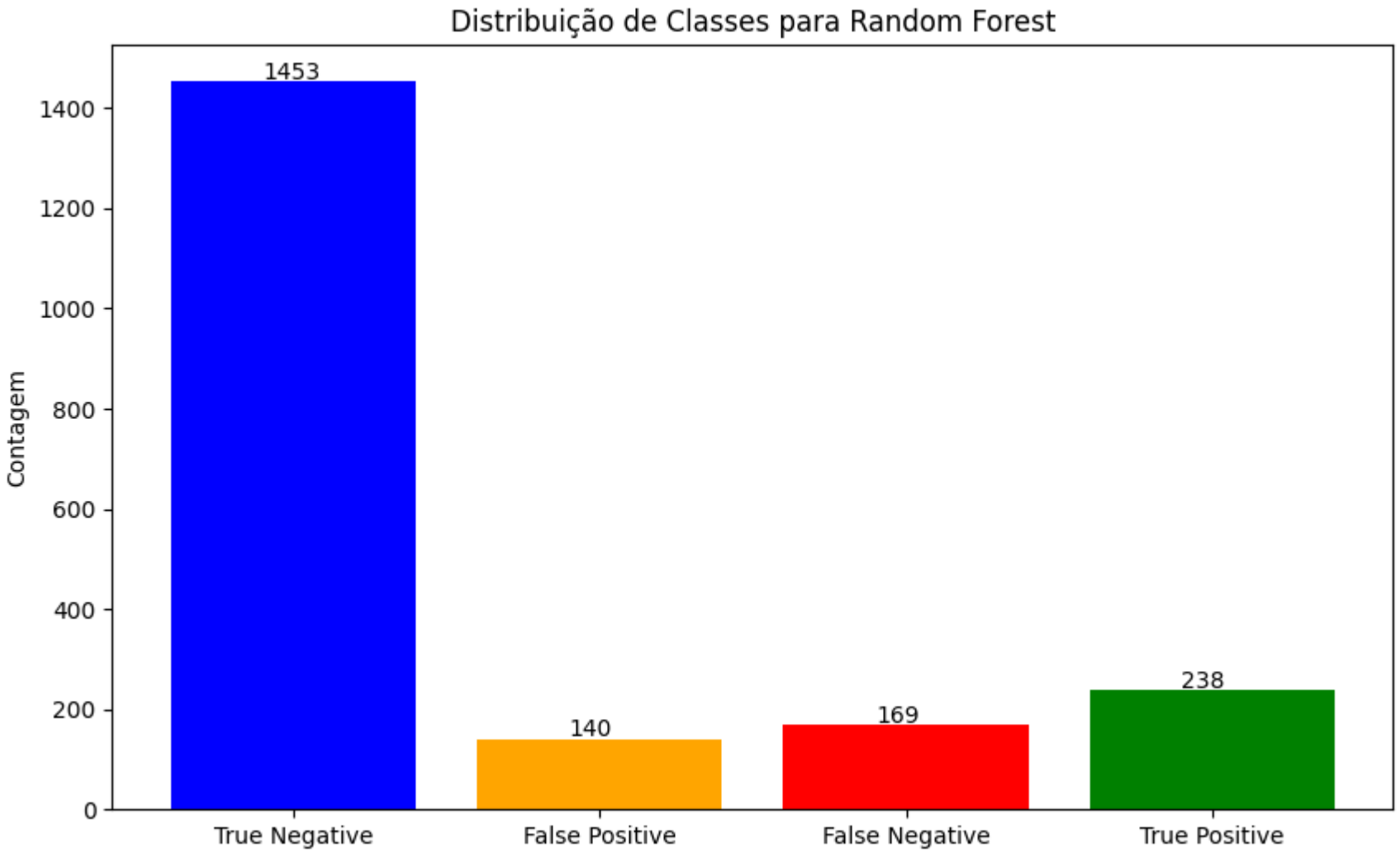
O melhor modelo é Random Forest com ROC AUC de 84,79%

Comparação entre os modelos

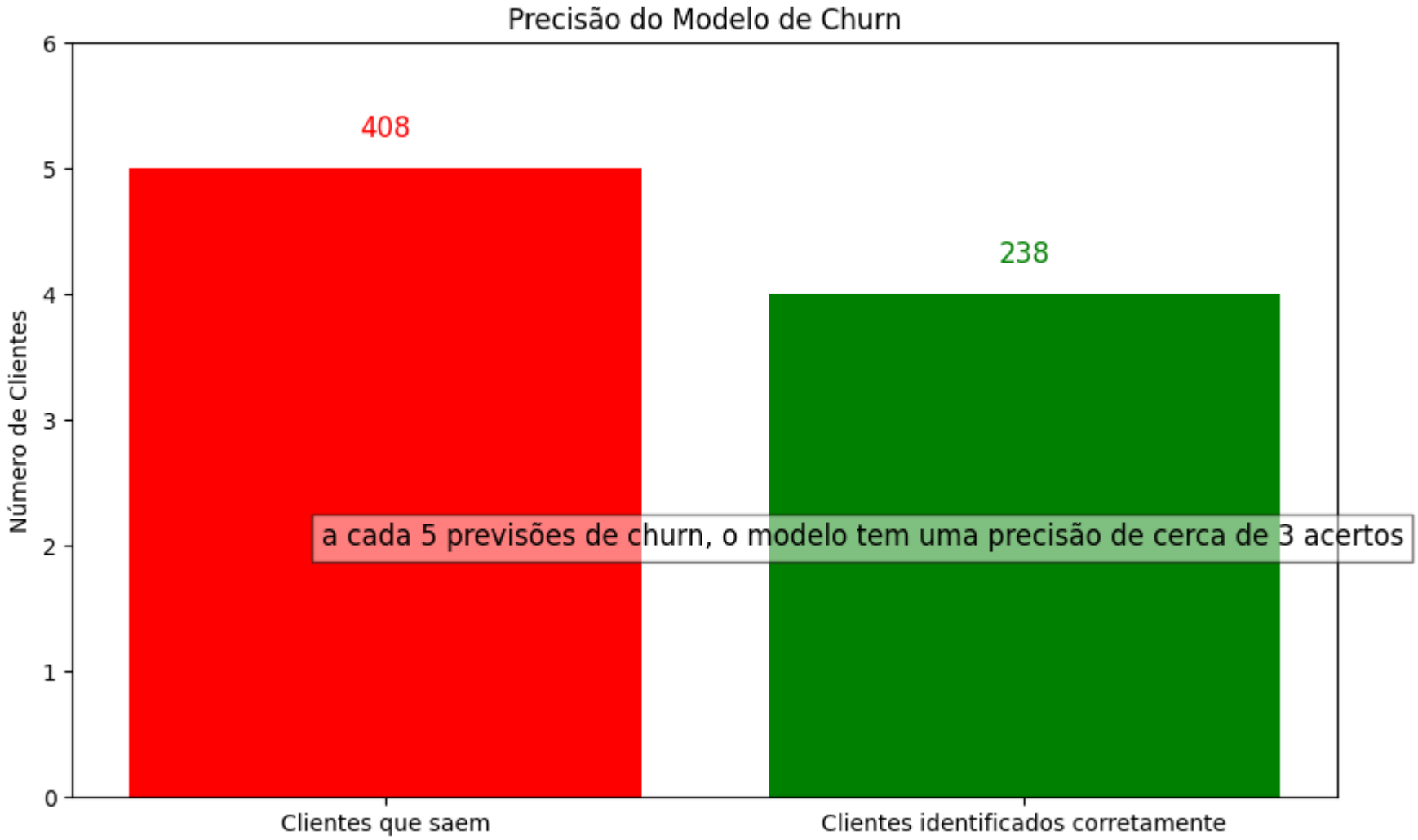



Receiver Operating Characteristic (ROC) Curve - Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.85)

Distribuição de Classes



Precisão do Random Forest



Próximos Passos

- Integrar o modelo ao CRM do banco.
- Segmentar clientes por risco de churn.
- Oferecer ofertas e atendimento personalizado para clientes de alto risco.
- Coletar novos dados para retraining do modelo.
- Ajustar o modelo com base no feedback das campanhas de retenção.

Conclusão

A comparação de resultados com base nos quatro modelos de algoritmos utilizados revela que o Random Forest foi o que apresentou o melhor valor de acurácia na assertiva de churn para os clientes do banco. O melhor modelo foi avaliado usando várias métricas.

A acurácia do modelo Random Forest foi de 84,55%, enquanto o ROC AUC foi de 84,79% no conjunto de testes e 84,71% na validação cruzada. Dado que o ROC AUC oferece uma visão mais completa da capacidade do modelo de distinguir entre clientes que permanecem e clientes que saem, especialmente em datasets desbalanceados como o de churn bancário, o modelo Random Forest pode ser considerado o melhor modelo com base no ROC AUC.

As estratégias propostas acima são fundamentais para manter uma linha de visão atualizada do quadro e avaliar melhor a redução de riscos.