CREDIT CARD FRAUD DETECTION

Deteção de Fraude 2024

Matilde Isabel da Silva Simões 202108782

COMPREENSÃO DO TEMA

Este problema tem como objetivo prever se uma transação de cartão de crédito é **legítima** ou **fraudulenta**, utilizando um *dataset*.

O objetivo do modelo machine learning, especializado em classificação binária, consiste em classificar cada transação num dos dois grupos mutuamente exclusivos:

Legítima: Transação autorizada, realizada pelo titular do cartão.

Fraudulenta: Transação não autorizada, efetuada por terceiros.

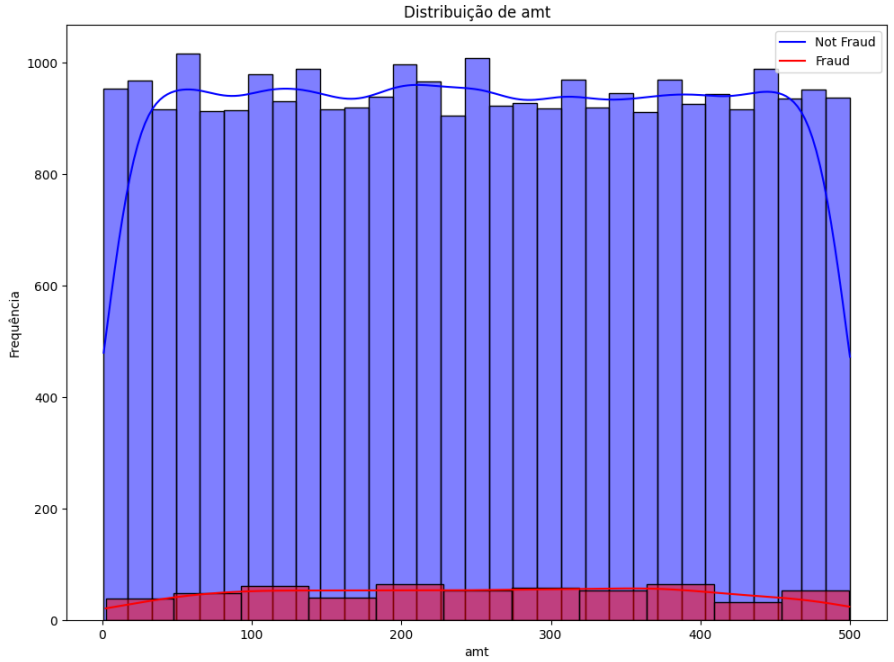
Será treinado utilizando uma abordagem de *supervised learning.*

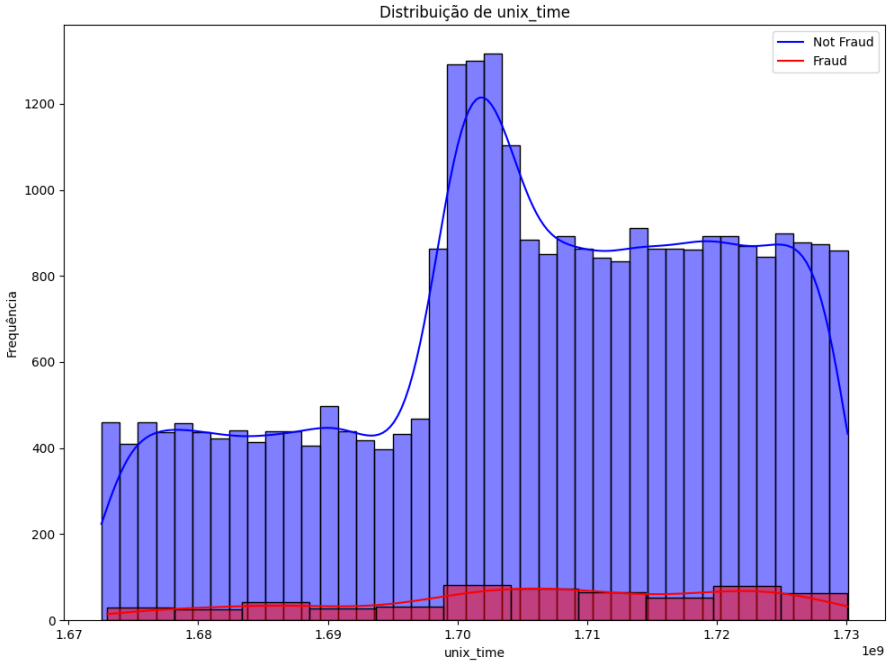
COMPREENSÃO DOS DADOS

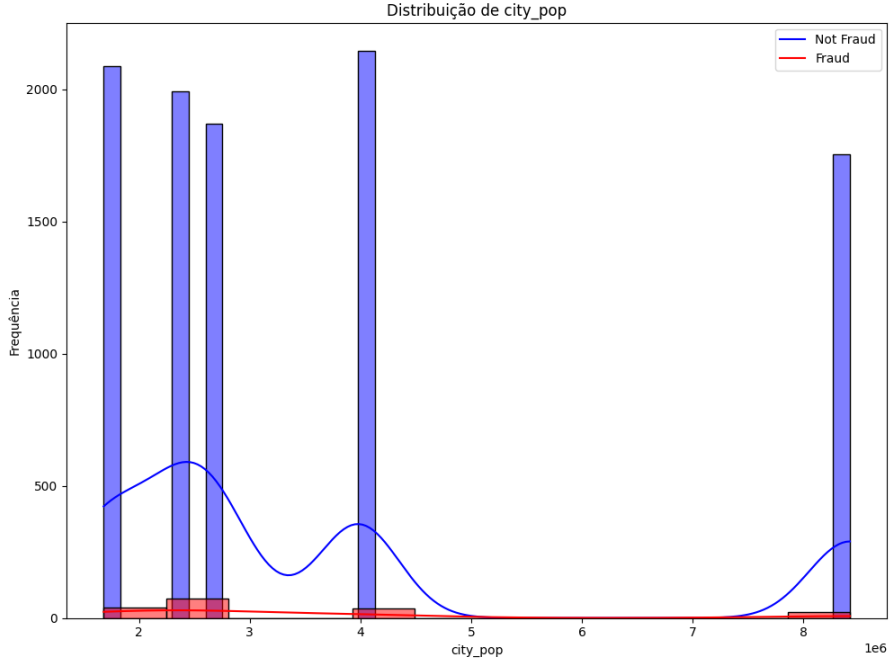
index	int64
trans_date_trans_time	object
cc_num	int64
device_os	object
merchant	object
amt	float64
trans_num	object
unix_time	int64
is_fraud	int64
first	object
last	object
gender	object
street	object
city	object
zip	float64
job	object
dob	object
lat	float64
long	float64
city_pop	float64
state	object
category	object
merch_lat	float64
merch_long	float64
merchant_id	float64

~~index~~	~~int64~~
trans_date_trans_time	object
~~cc_num~~	~~int64~~
device_os	object
merchant	object
amt	float64
trans_num	object
unix_time	int64
~~is_fraud~~	~~int64~~
~~first~~	object
~~last~~	object
gender	object
street	object
city	object
~~zip~~	~~float64~~
job	object
dob	object
lat	float64
long	float64
city_pop	float64
state	object
category	object
merch_lat	float64
merch_long	float64
~~merchant_id~~	~~float64~~

Remover
Ligadas a localização
Mudança de tipo

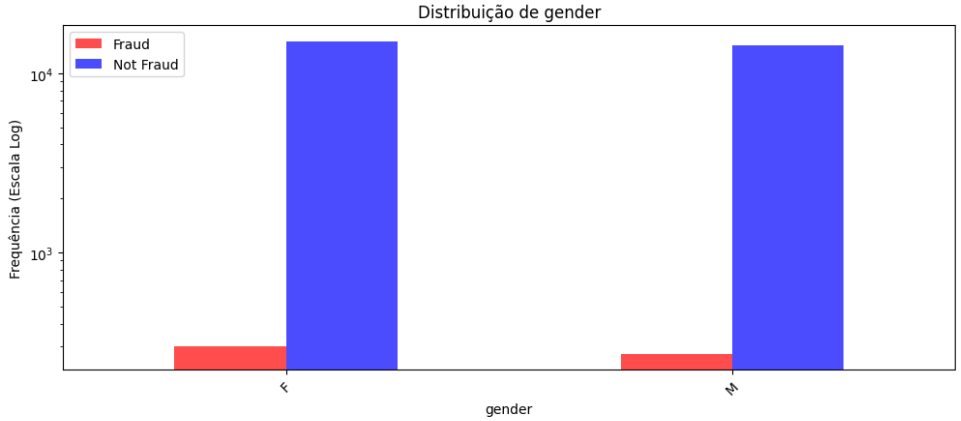


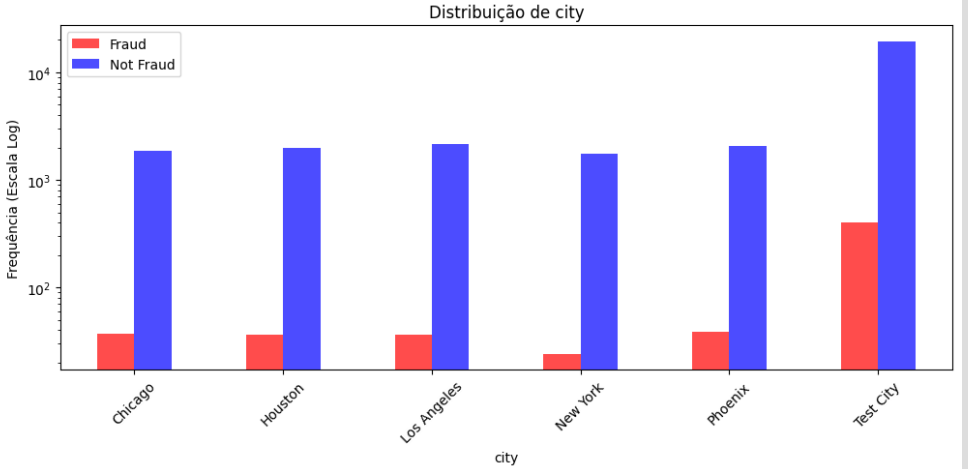


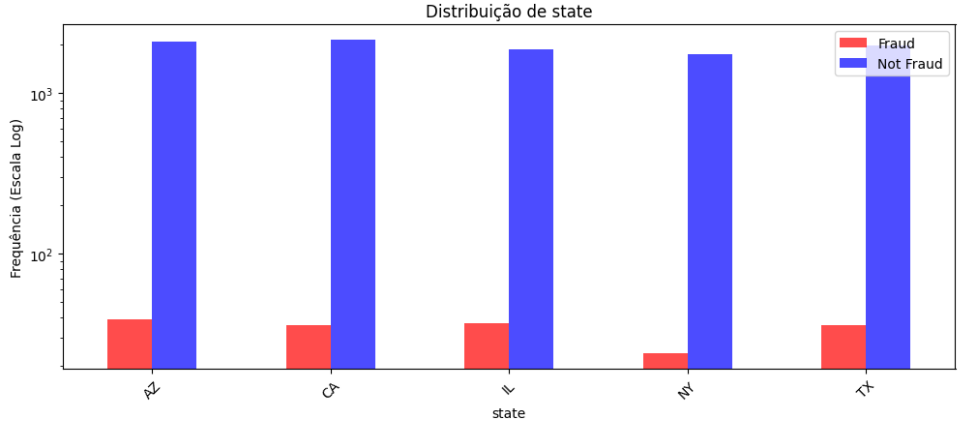


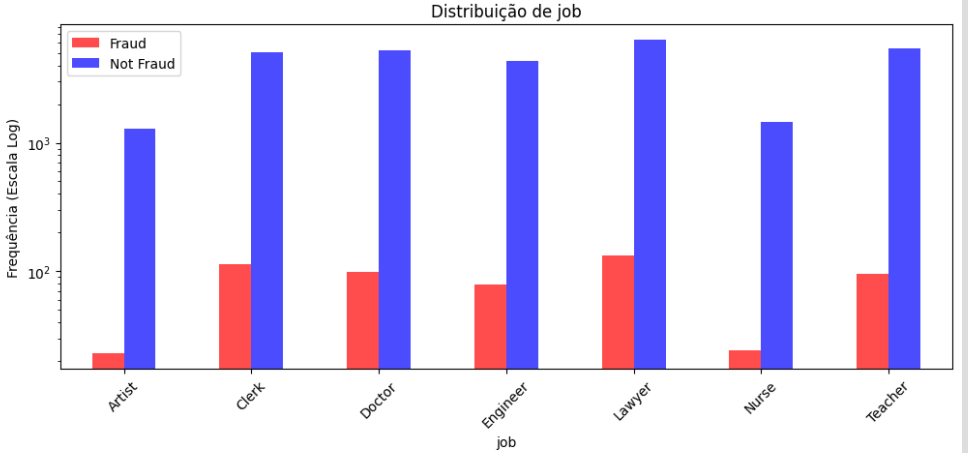
day_of_week	Não Fraude	Fraude
Monday	4315	93
Tuesday	4169	92
Friday	4162	92
Thursday	4251	86
Wednesday	4081	77
Saturday	4242	71
Sunday	4209	60

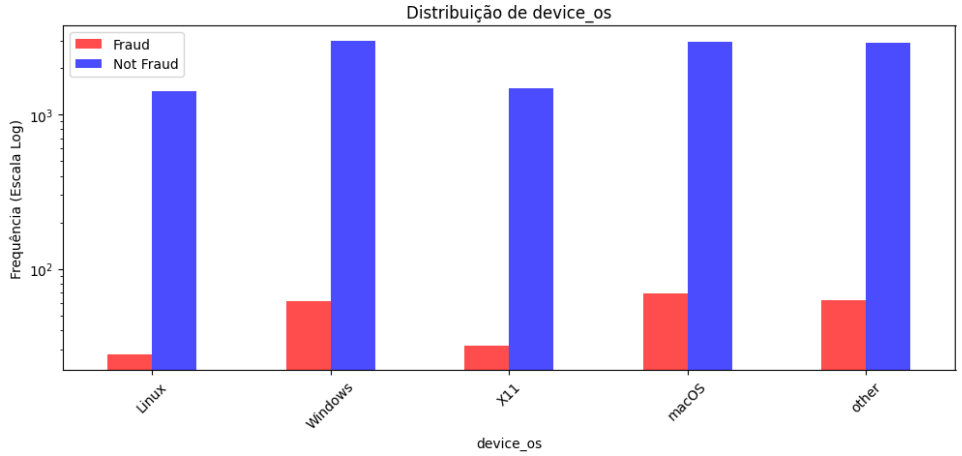
```
Period of Day: Madrugada
  Total Frauds: 186
  Total Transactions: 10074
  Fraud Rate: 1.85%
--------------------------------
Period of Day: Manhã
  Total Frauds: 128
  Total Transactions: 6259
  Fraud Rate: 2.05%
--------------------------------
Period of Day: Noite
  Total Frauds: 149
  Total Transactions: 7500
  Fraud Rate: 1.99%
--------------------------------
Period of Day: Tarde
  Total Frauds: 108
  Total Transactions: 6167
  Fraud Rate: 1.75%
```

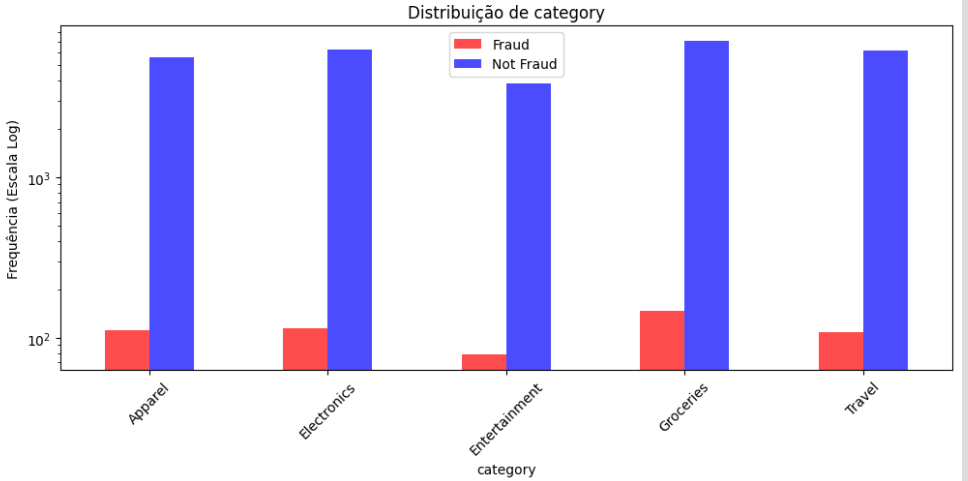




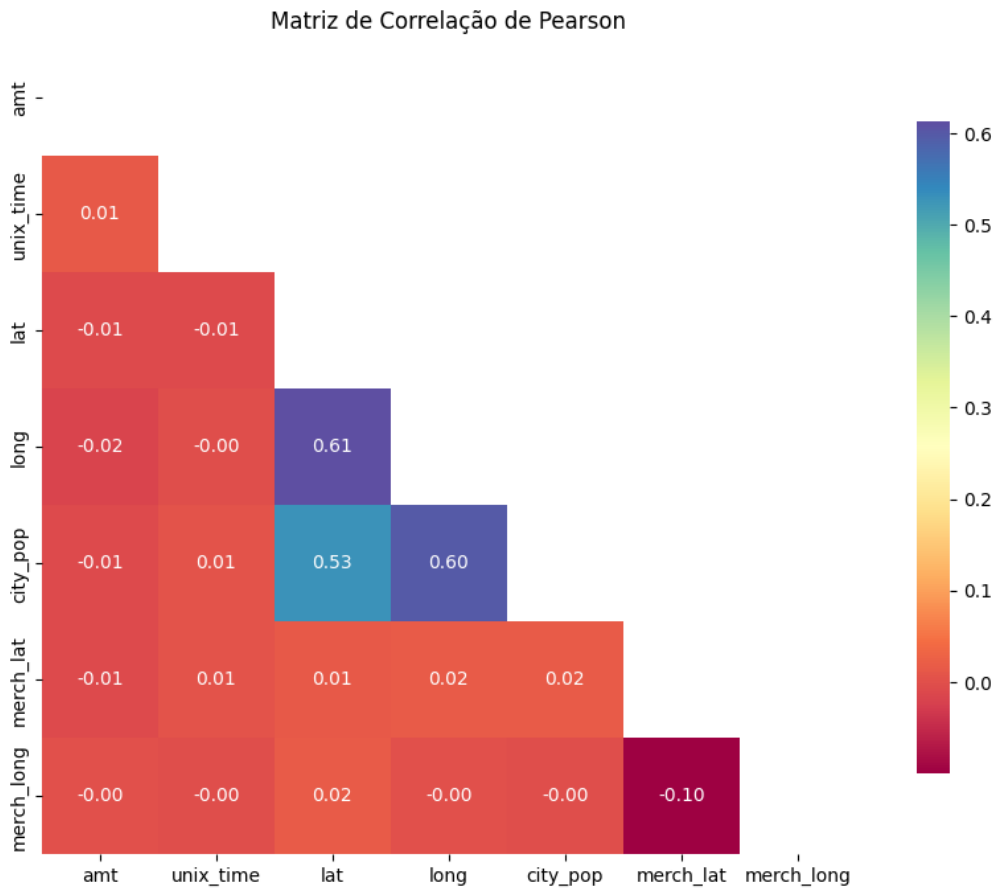


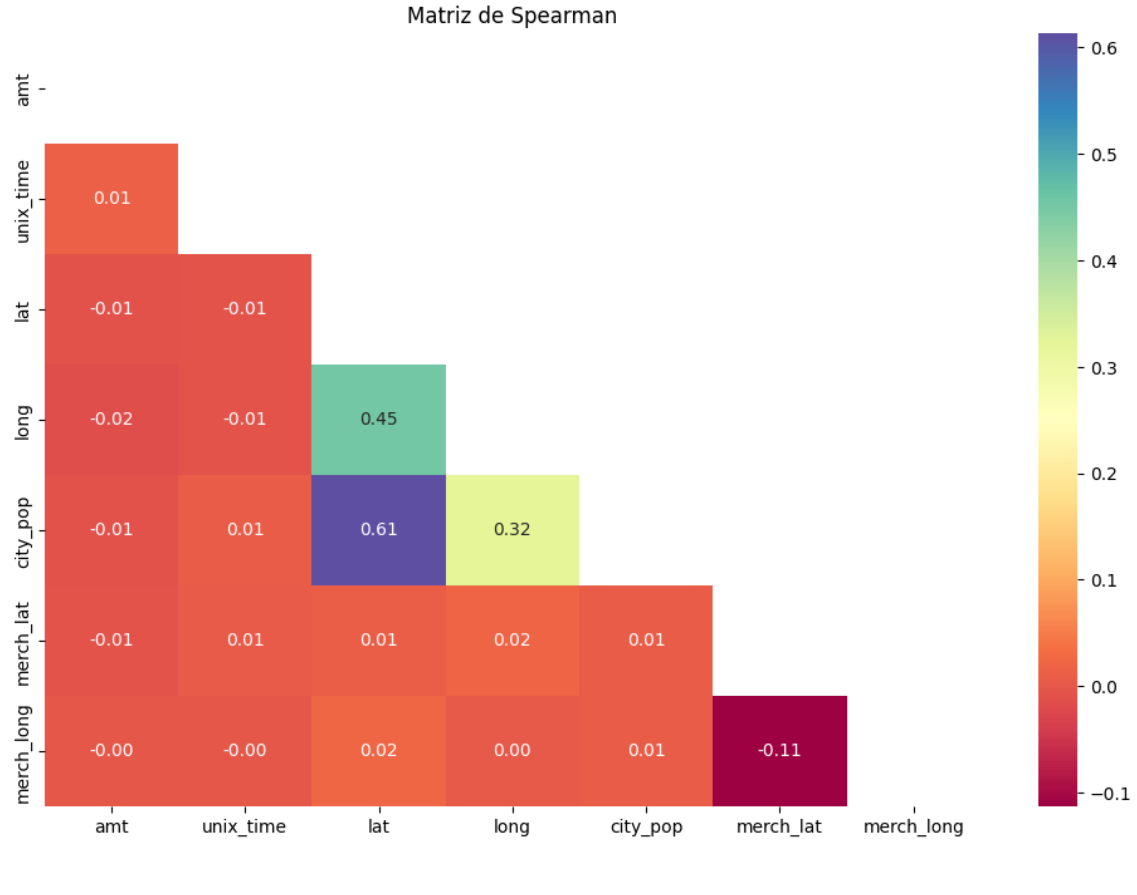


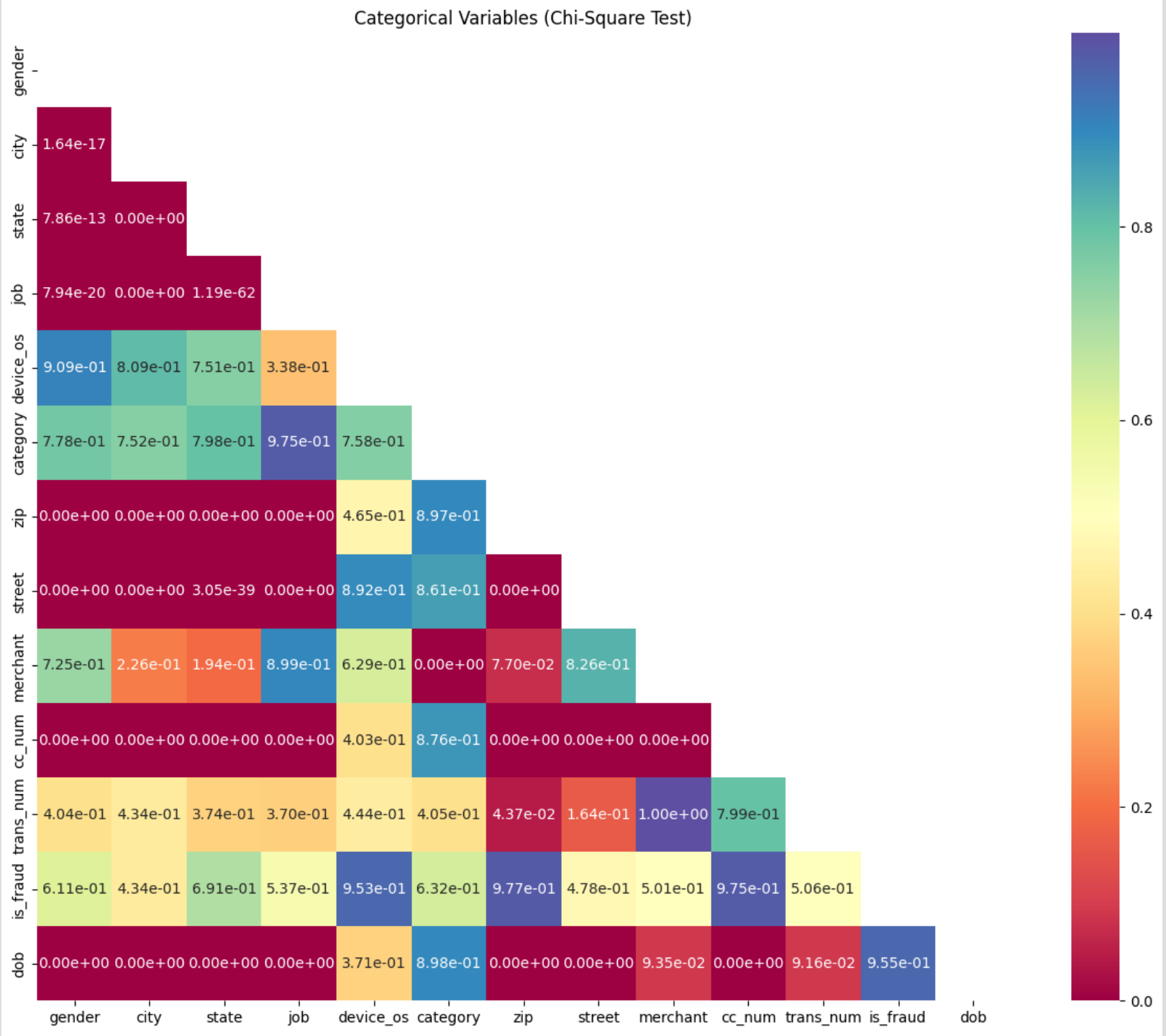





```
Age Category: Adult
  Total Frauds: 192
  Total Transactions: 9875
  Fraud Rate: 1.94%
  Min Age: 37.0
  Max Age: 55.0
-----------------------------
Age Category: Middle Age
  Total Frauds: 209
  Total Transactions: 11849
  Fraud Rate: 1.76%
  Min Age: 56.0
  Max Age: 74.0
-----------------------------
Age Category: Senior
  Total Frauds: 0
  Total Transactions: 10
  Fraud Rate: 0.00%
  Min Age: nan
  Max Age: nan
-----------------------------
Age Category: Young Adult
  Total Frauds: 170
  Total Transactions: 8266
  Fraud Rate: 2.06%
  Min Age: 18.0
  Max Age: 36.0
```


Matriz de Correlação de Pearson
amt
unix_time
lat
long
city_pop
merch_lat
merch_long
0.01
-0.01 -0.01
-0.02 -0.00 0.61
-0.01 0.01 0.53 0.60
-0.01 0.01 0.01 0.02 0.02
-0.00 -0.00 0.02 -0.00 -0.00 -0.10
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Matriz de Spearman
amt
unix_time
lat
long
city_pop
merch_lat
merch_long
0.01
-0.01 -0.01
-0.02 -0.01 0.45
-0.01 0.01 0.61 0.32
-0.01 0.01 0.01 0.02 0.01
-0.00 -0.00 0.02 0.00 0.01 -0.11
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1


Categorical Variables (Chi-Square Test)
gender
city
state
job
device_os
category
zip
street
merchant
cc_num
trans_num
is_fraud
dob
1.64e-17
7.86e-13 0.00e+00
7.94e-20 0.00e+00 1.19e-62
9.09e-01 8.09e-01 7.51e-01 3.38e-01
7.78e-01 7.52e-01 7.98e-01 9.75e-01 7.58e-01
0.00e+00 0.00e+00 0.00e+00 0.00e+00 4.65e-01 8.97e-01
0.00e+00 0.00e+00 3.05e-39 0.00e+00 8.92e-01 8.61e-01 0.00e+00
7.25e-01 2.26e-01 1.94e-01 8.99e-01 6.29e-01 0.00e+00 7.70e-02 8.26e-01
0.00e+00 0.00e+00 0.00e+00 0.00e+00 4.03e-01 8.76e-01 0.00e+00 0.00e+00 0.00e+00
4.04e-01 4.34e-01 3.74e-01 3.70e-01 4.44e-01 4.05e-01 4.37e-02 1.64e-01 1.00e+00 7.99e-01
6.11e-01 4.34e-01 6.91e-01 5.37e-01 9.53e-01 6.32e-01 9.77e-01 4.78e-01 5.01e-01 9.75e-01 5.06e-01
0.00e+00 0.00e+00 0.00e+00 0.00e+00 3.71e-01 8.98e-01 0.00e+00 0.00e+00 9.35e-02 0.00e+00 9.16e-02 9.55e-01
gender city state job device_os category zip street merchant cc_num trans_num is_fraud dob
0.8
0.6
0.4
0.2
0.0

index	29970
trans_date_trans_time	29868
cc_num	1101
device_os	5
merchant	101
amt	22615
trans_num	29470
unix_time	29959
is_fraud	2
first	108
last	108
gender	2
street	102
city	6
zip	1077
job	7
dob	1062
lat	5
long	5
city_pop	5
state	5
category	5
merch_lat	98
merch_long	100
merchant_id	100

index	0
trans_date_trans_time	100
cc_num	0
device_os	17964
merchant	0
amt	100
trans_num	0
unix_time	0
is_fraud	0
first	10
last	10
gender	10
street	10
city	10
zip	216
job	216
dob	10
lat	19980
long	19980
city_pop	19980
state	19980
category	599
merch_lat	599
merch_long	10
merchant_id	10

~~index~~
~~trans_date_trans_time~~
~~cc_num~~
device_os
~~merchant~~
amt
~~trans_num~~
unix_time
is_fraud
~~first~~
~~last~~
gender
~~street~~
city
~~zip~~
job
dob
~~lat~~
~~long~~
~~city_pop~~
~~state~~
category
~~merch_lat~~
~~merch_long~~
~~merchant_id~~

Valores Duplicados!

PREPARAÇÃO DOS DADOS

ETAPAS

- Tirar duplicados
- **Feature Engineering**
- Correr o teste de **Shapiro-Wilk** para ver a distribuição das variáveis - **MinMaxScale**
- Dividir em **treino e teste** (com *stratify* e *shuffle*) e **normalização**
- Usar o **kNN de R** para **imputar** os valores no **treino**
- Usar a **média** e a **moda** para imputar no **teste**
- Transformar para **variáveis numéricas**

Faixas etárias

Transações feitas até ao momento

Com quantos comerciantes fez transação até ao momento

Tempo que passou entre a última transação

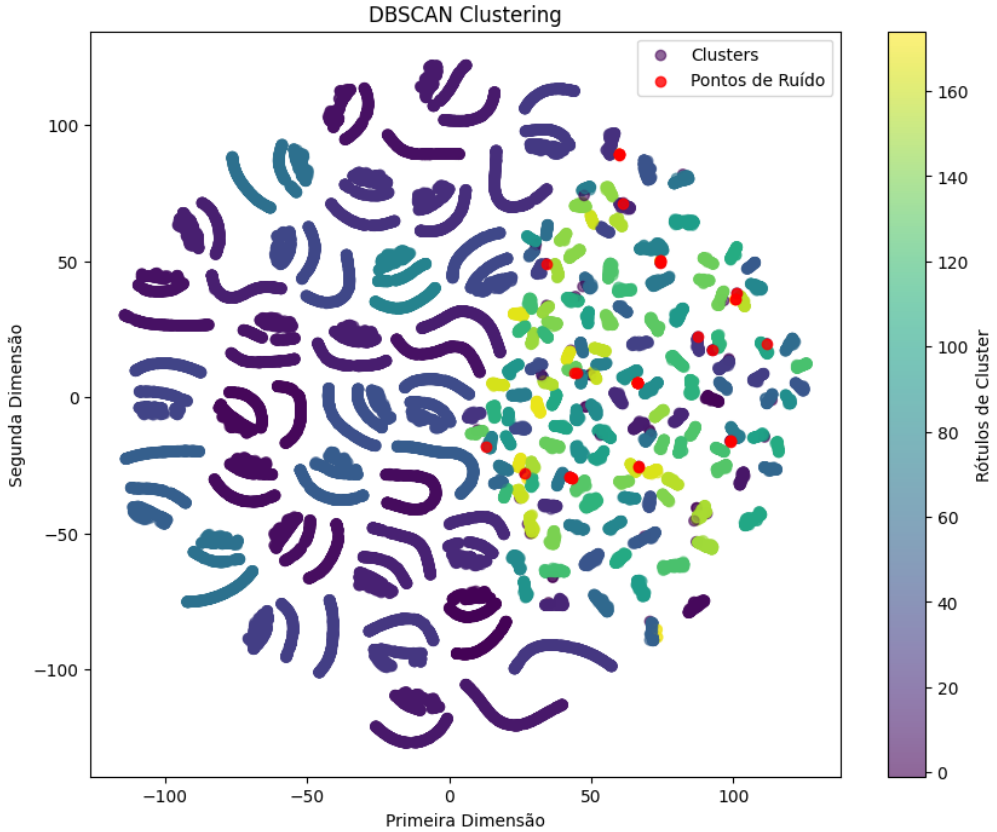
device_os	amt	gender	city	job	category	age_category	transactions_count	unique_merchants_count	time_since_last	day_of_week_sin
other	0.503056	M	Test City	Doctor	Travel	Adult	0.285714	1.0	0.010520	0.277479
macOS	0.743241	M	Los Angeles	Clerk	Travel	Adult	0.000000	1.0	0.265077	0.277479
Linux	0.633570	M	Test City	Clerk	Apparel	Middle Age	0.000000	1.0	0.019855	1.000000
macOS	0.567772	F	Los Angeles	Doctor	Travel	Adult	0.000000	1.0	0.000000	1.000000
macOS	0.422648	M	Test City	Engineer	Groceries	Adult	0.428571	1.0	0.003864	1.000000

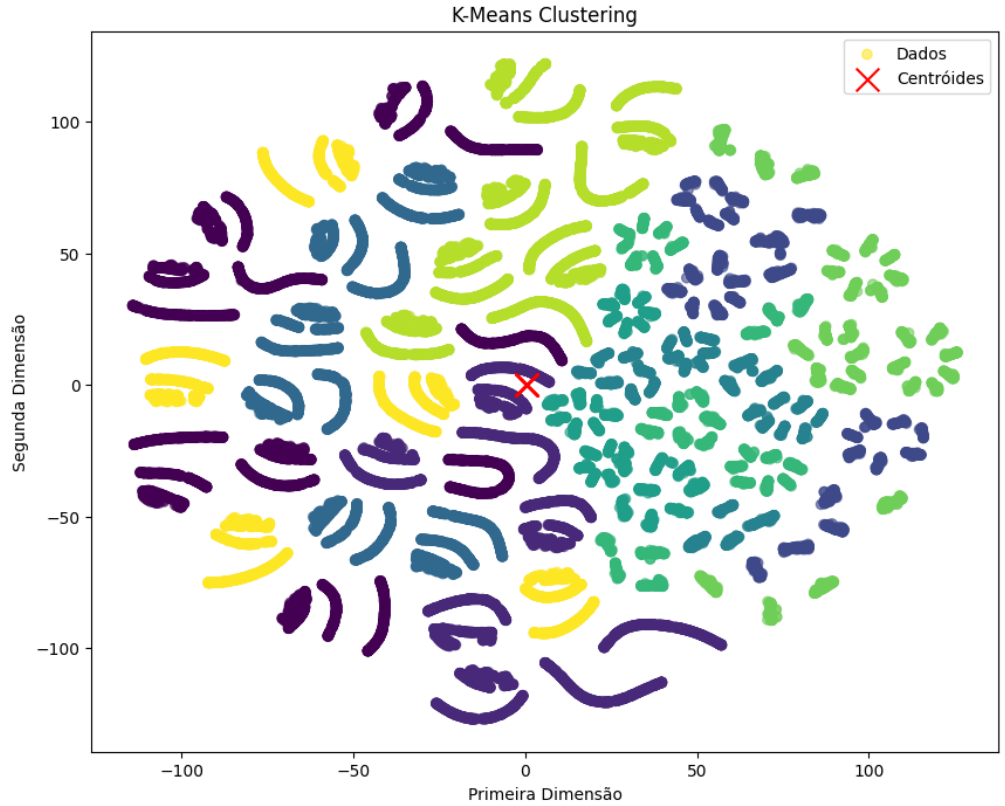
Dias da semana

day_of_week_cos	period_of_day_sin	period_of_day_cos	city_has_info
0.000000	1.0	0.5	0
0.000000	0.0	0.5	1
0.356896	0.0	0.5	0
0.356896	0.5	1.0	1
0.356896	0.0	0.5	0

Período do dia

Se existe informação sobre a cidade


DBSCAN Clustering
Clusters
Pontos de Ruído
Segunda Dimensão
Primeira Dimensão
Rótulos de Cluster
100
50
0
−50
−100
−100
−50
0
50
100
160
140
120
100
80
60
40
20
0


K-Means Clustering
Dados
Centróides
Segunda Dimensão
Primeira Dimensão
100
50
0
−50
−100
−100
−50
0
50
100

MODELAÇÃO

ETAPAS

- Usado ***pipeline***
- ***Cross validation***
- Usado ***grid search*** nalguns modelos
- ***Smote*** para fazer ***oversampling*** com ***Tomek links*** para ***undersampling*** informado
- Retiradas métricas como ***AUC***, ***f1***, **precisão** e ***recall***

MODELOS IMPLEMENTADOS

DECISION TREE (SEM GRIDSEARCH)

DECISION TREE (COM GRIDSEARCH)

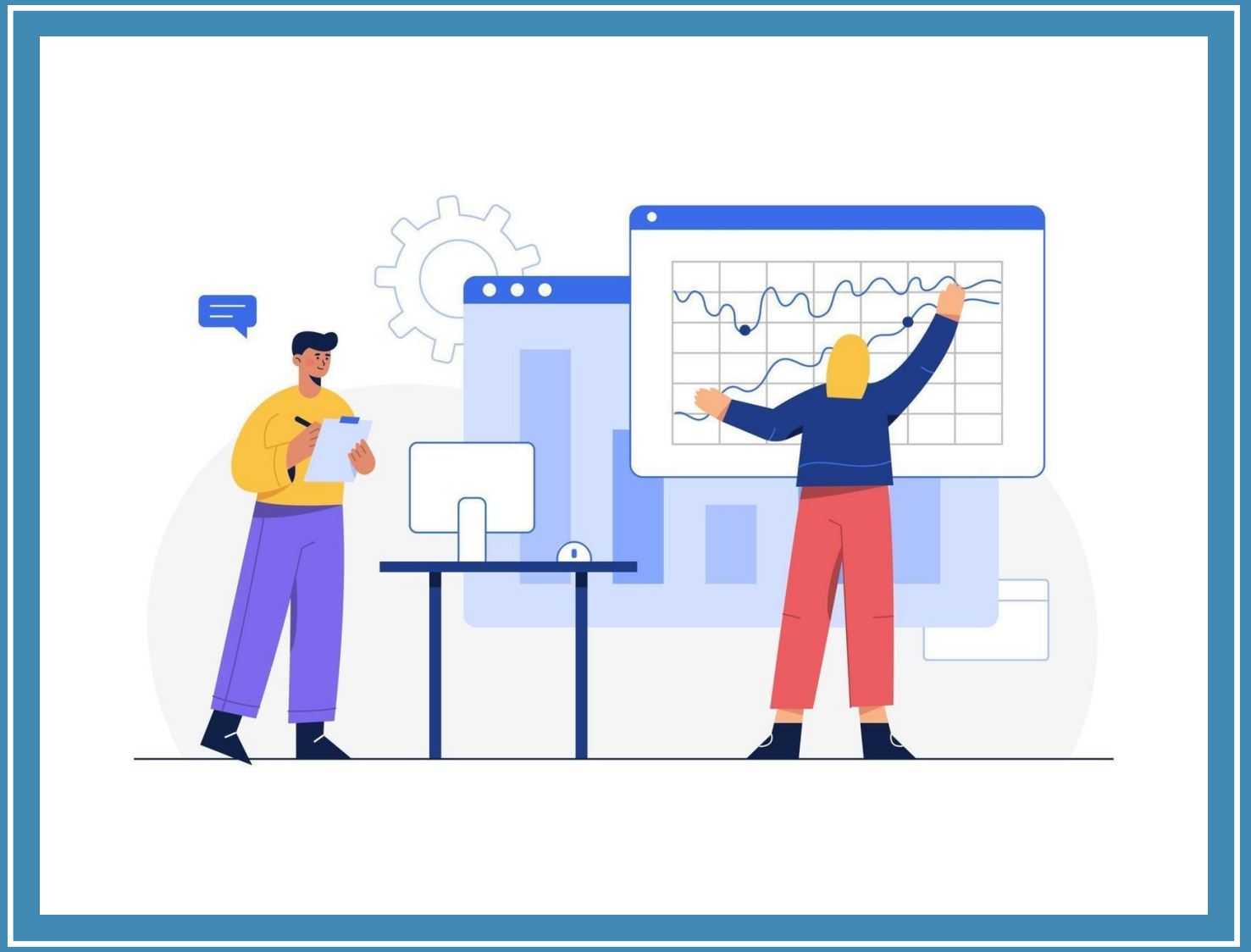
NEURAL NETWORK

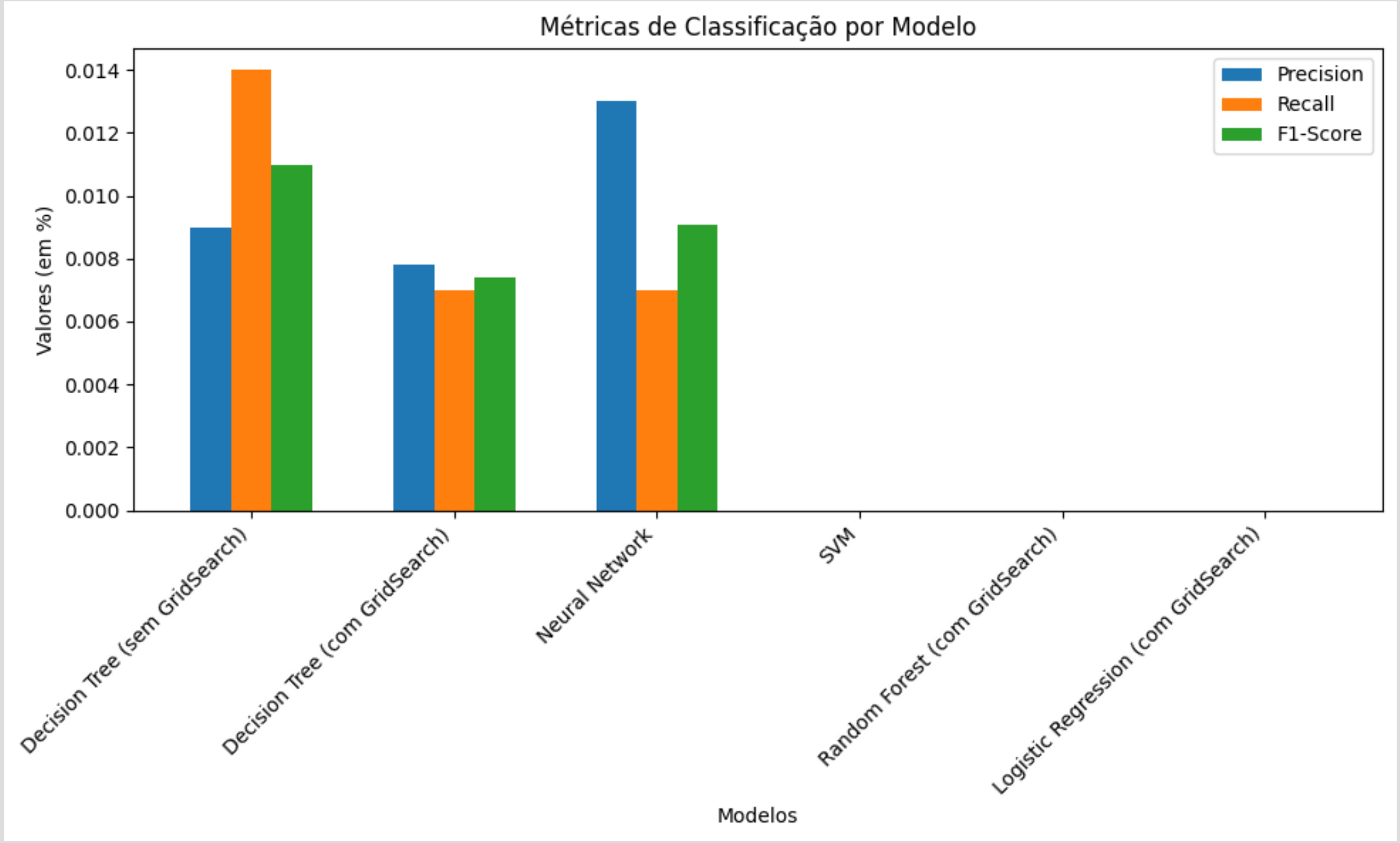
SVM

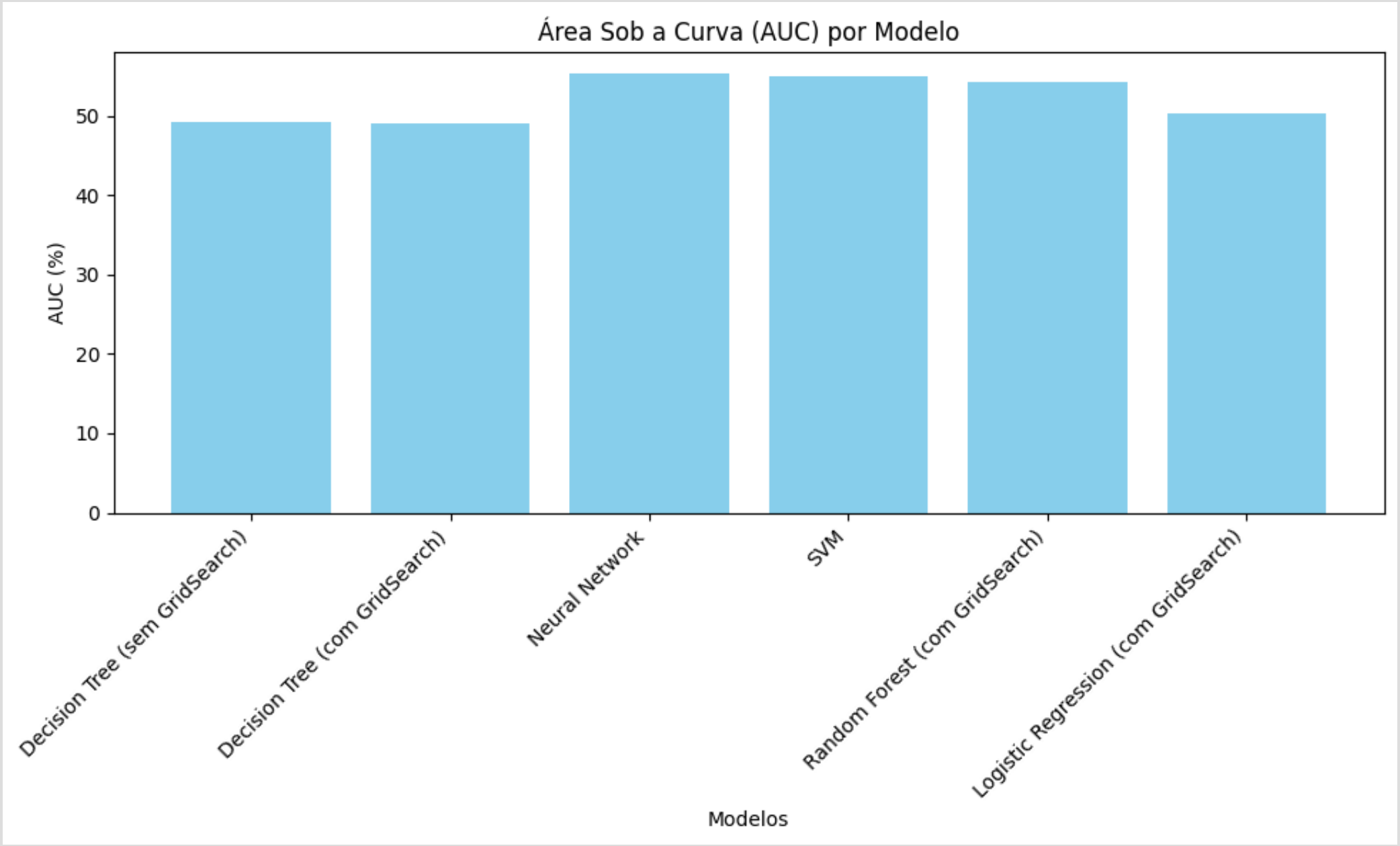
RANDOM FOREST (COM GRIDSEARCH)

LOGISTIC REGRESSION (COM GRIDSEARCH)

AVALIAÇÃO




Métricas de Classificação por Modelo
Precision
Recall
F1-Score
Valores (em %)
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Decision Tree (sem GridSearch)
Decision Tree (com GridSearch)
Neural Network
SVM
Random Forest (com GridSearch)
Logistic Regression (com GridSearch)
Modelos


Área Sob a Curva (AUC) por Modelo
AUC (%)
0
10
20
30
40
50
Decision Tree (sem GridSearch)
Decision Tree (com GridSearch)
Neural Network
SVM
Random Forest (com GridSearch)
Logistic Regression (com GridSearch)
Modelos

Local

Modelo	Precision	Recall	F1-Score	AUC
Decision Tree (sem GridSearch)	0.90%	1.40%	1.10%	49.20%
Decision Tree (com GridSearch)	0.78%	0.70%	0.74%	49.09%
Neural Network	1.30%	0.70%	0.91%	55.32%
SVM	0.00%	0.00%	0.00%	54.97%
Random Forest (com GridSearch)	0.00%	0.00%	0.00%	54.28%
Logistic Regression (com GridSearch)	0.00%	0.00%	0.00%	50.26%

Kaggle

Nome do Teste	Modelo Utilizado	Valor Obtido
teste1.csv	Árvore de Decisão Sem GridSearch	0.50345
teste2.csv	SVM Sem GridSearch	0.54154
teste3.csv	Logistic Regression Com GridSearch	0.48420
teste4.csv	Neural Network Sem GridSearch	0.63801
teste5.csv	Random Forest Com GridSearch	0.59903

CONCLUSÕES, LIMITAÇÕES E TRABALHO FUTURO

A **Rede Neural** foi o modelo com melhor AUC, apesar do desempenho limitado nas outras métricas.

O desequilíbrio extremo do *dataset* afetou a capacidade de classificação de vários modelos, mesmo com ***SMOTE-Tomek Links***.

Os resultados do **clustering** podem ser incorporados como novas variáveis nos modelos para melhorar a capacidade preditiva e identificar grupos com maior propensão a fraudes.