Inteligência Artificial

Lista 9 - Etapas de Pré-Processamento e Agrupamento

Aluno: Otávio Augusto de Assis Ferreira Monteiro
Matrícula: 851568
Link para o código:
https://github.com/otavioaugustoafm/Faculdade/blob/main/IA/Listas/Lista%209/pp.ipynb
Link para a base:
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download&select=creditcard.csv

Belo Horizonte, 2025

Questão 1

Visualização de dados:

Foi realizada a inspeção inicial dos dados:

```
--- Primeiras 5 linhas do DataFrame ---
   Time        V1        V2        V3        V4        V5        V6        V7  \
0   0.0 -1.359807 -0.072781  2.536347  1.378155 -0.338321  0.462388  0.239599
1   0.0  1.191857  0.266151  0.166480  0.448154  0.060018 -0.082361 -0.078803
2   1.0 -1.358354 -1.340163  1.773209  0.379780 -0.503198  1.800499  0.791461
3   1.0 -0.966272 -0.185226  1.792993 -0.863291 -0.010309  1.247203  0.237609
4   2.0 -1.158233  0.877737  1.548718  0.403034 -0.407193  0.095921  0.592941

         V8        V9  ...       V21       V22       V23       V24       V25  \
0  0.098698  0.363787  ... -0.018307  0.277838 -0.110474  0.066928  0.128539
1  0.085102 -0.255425  ... -0.225775 -0.638672  0.101288 -0.339846  0.167170
2  0.247676 -1.514654  ...  0.247998  0.771679  0.909412 -0.689281 -0.327642
3  0.377436 -1.387024  ... -0.108300  0.005274 -0.190321 -1.175575  0.647376
4 -0.270533  0.817739  ... -0.009431  0.798278 -0.137458  0.141267 -0.206010

        V26       V27       V28  Amount  Class
0 -0.189115  0.133558 -0.021053  149.62      0
1  0.125895 -0.008983  0.014724    2.69      0
2 -0.139097 -0.055353 -0.059752  378.66      0
3 -0.221929  0.062723  0.061458  123.50      0
4  0.502292  0.219422  0.215153   69.99      0

[5 rows x 31 columns]

--- Dimensões (Linhas, Colunas) ---
(284807, 31)
```

```
--- Tipos de Dados e Valores Não-Nulos ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
```

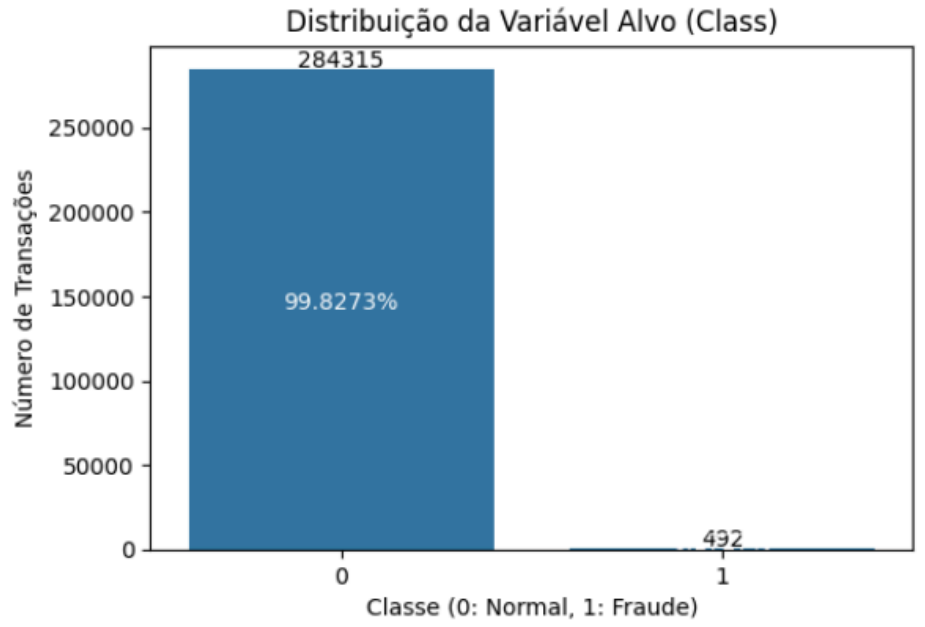
A base de dados tem uma estrutura de 284.807 registros e 31 atributos. A maioria das colunas está anonimizada e já passou por uma transformação,

significando que já está em uma escala padronizada e sem problemas de codificação. A base não apresenta valores ausentes também.

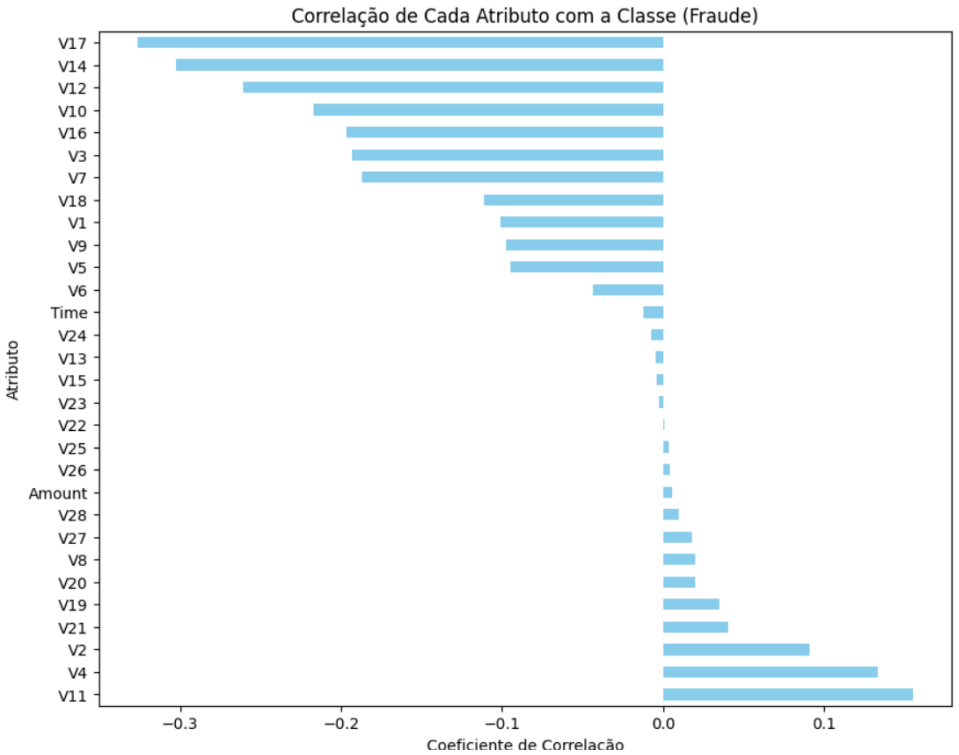
```
--- Contagem de Classes ---
Class
0    284315
1       492
Name: count, dtype: int64

--- Proporção de Classes (%) ---
Class
0    99.827251
1     0.172749
Name: count, dtype: float64
```

Analisando o balanceamento das classes, podemos concluir que existe um desbalanceamento extremo entre elas: apenas 0.17% das transações são fraudes. O gráfico a seguir demonstra esse desbalanceamento, confirmando que a classe 1 é praticamente invisível quando comparada a classe 0.



Seleção de atributos:

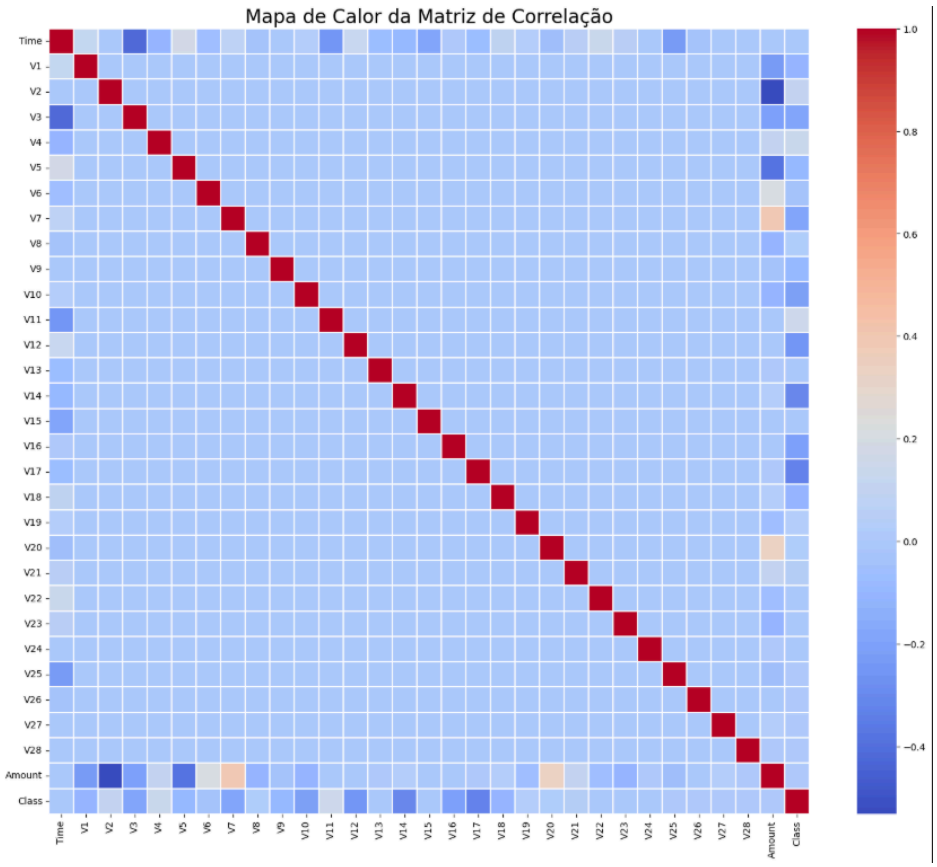


O gráfico acima busca compreender a correlação entre os atributos. É importante realizar esse passo, pois podemos identificar quais são os atributos que não precisam ser levados em consideração, pois não importam muito para a classificação da instância.

Relevância Forte: As variáveis que apresentam a maior correlação negativa são V17, V14, V12 E V10. Um valor baixo nessas variáveis está fortemente associado a uma transação de fraude. Devem ser mantidas.

Relevância Moderada: As variáveis com a maior correlação positiva são a V11 e V4. Um valor alto nessas variáveis está associado a uma transação de fraude. Devem ser mantidas.

Irrelevância: As variáveis com correlação próxima a zero são consideradas de baixa relevância. São elas: V28, V27, V23, V22, V25, V13 e V15. Podemos ignorar esses atributos, filtrando, por exemplo, valores absolutos abaixo de 0.03.



Por sua vez, o mapa de calor acima permite identificar redundâncias. Podemos analisar, então, que o mapa não apresenta correlações extremas, que seriam pontos no mapa onde um tom de vermelho escuro ou azul escuro se destacam muito (exceto na diagonal principal).

Desse modo, concluindo essa seção, devemos focar na remoção dos atributos irrelevantes, como explicado anteriormente. Não é necessário fazer nenhuma mudança quanto às redundâncias, afinal não temos casos extremos que poderiam prejudicar o desempenho do processo.

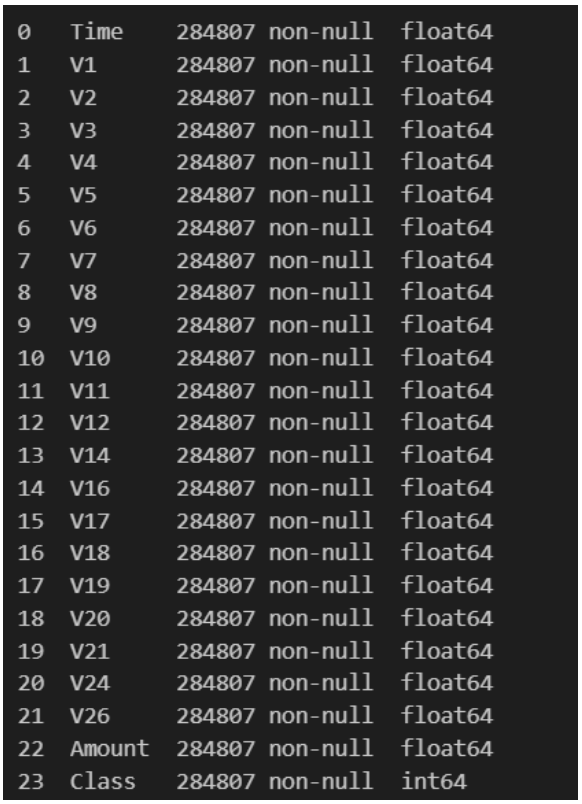
Vamos remover as seguintes colunas: V28, V27, V23, V22, V25, V13 e V15. O amount tem uma correlação baixa, mas será mantido, pois o valor de uma transação pode significar algo - análise empírica.

```
Número de colunas originais: 31
Número de colunas após a remoção: 24

Primeiras 5 linhas do novo DataFrame:
   Time        V1        V2        V3        V4        V5        V6        V7  \
0   0.0 -1.359807 -0.072781  2.536347  1.378155 -0.338321  0.462388  0.239599
1   0.0  1.191857  0.266151  0.166480  0.448154  0.060018 -0.082361 -0.078803
2   1.0 -1.358354 -1.340163  1.773209  0.379780 -0.503198  1.800499  0.791461
3   1.0 -0.966272 -0.185226  1.792993 -0.863291 -0.010309  1.247203  0.237609
4   2.0 -1.158233  0.877737  1.548718  0.403034 -0.407193  0.095921  0.592941
```

Codificação:

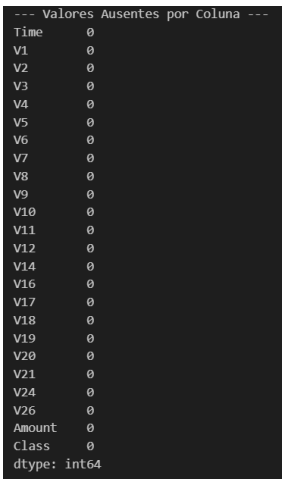
Como a base trabalhada não tem atributos nominais ou ordinais, não precisamos utilizar o One-Hot ou Label encoding. Desse modo, essa seção será apenas para a verificação dos atributos que restaram após a remoção dos não essenciais.



0	Time	284807 non-null	float64
1	V1	284807 non-null	float64
2	V2	284807 non-null	float64
3	V3	284807 non-null	float64
4	V4	284807 non-null	float64
5	V5	284807 non-null	float64
6	V6	284807 non-null	float64
7	V7	284807 non-null	float64
8	V8	284807 non-null	float64
9	V9	284807 non-null	float64
10	V10	284807 non-null	float64
11	V11	284807 non-null	float64
12	V12	284807 non-null	float64
13	V14	284807 non-null	float64
14	V16	284807 non-null	float64
15	V17	284807 non-null	float64
16	V18	284807 non-null	float64
17	V19	284807 non-null	float64
18	V20	284807 non-null	float64
19	V21	284807 non-null	float64
20	V24	284807 non-null	float64
21	V26	284807 non-null	float64
22	Amount	284807 non-null	float64
23	Class	284807 non-null	int64

Eliminação das Inconsistências e Redundâncias:

Já tínhamos visto anteriormente que não temos valores ausentes, mas podemos verificar novamente isso por meio de um somatório, visto na imagem a seguir:



```
--- Valores Ausentes por Coluna ---
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V14       0
V16       0
V17       0
V18       0
V19       0
V20       0
V21       0
V24       0
V26       0
Amount    0
Class     0
dtype: int64
```

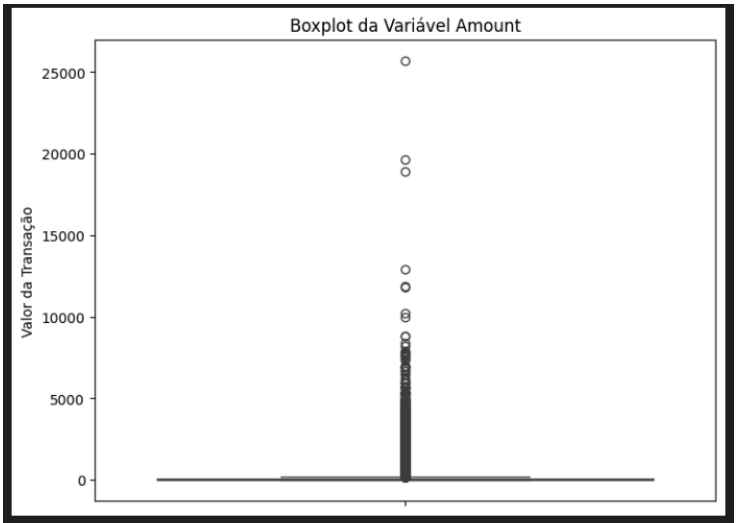
Agora, em relação a registros duplicados, tivemos 1081 instâncias iguais, as quais foram removidas:

```
Número de registros duplicados encontrados: 1081
Registros duplicados removidos. Novo número de linhas: 283726
```

Observações: Como a base de dados já passou por transformações, não acredito que terão inconsistências onde os mesmos conjuntos de atributos resultam em classes diferentes em duas instâncias. Por isso, vamos seguir ao próximo passo.

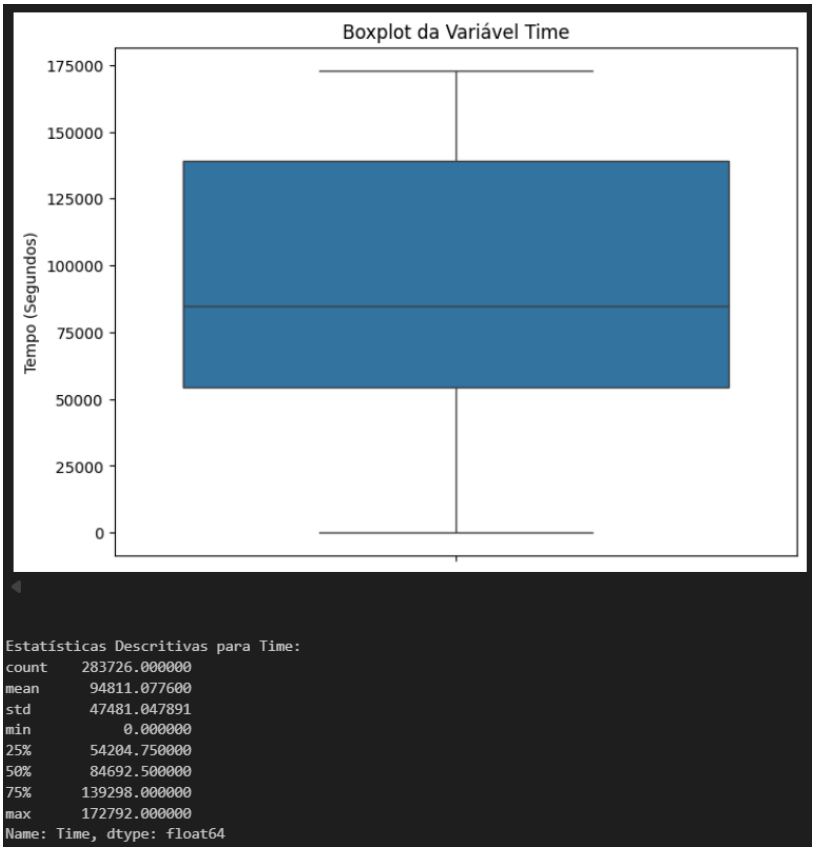
Eliminação de outliers:

Nesta seção serão utilizados dois boxplots para visualizar a distribuição das variáveis e identificar valores que estendem muito além dos limites interquartil (outliers).




```
Estatísticas Descritivas para Amount:
count    283726.000000
mean         88.472687
std         250.399437
min           0.000000
25%           5.600000
50%          22.000000
75%          77.510000
max       25691.160000
Name: Amount, dtype: float64
```

A partir deste gráfico, podemos observar que a maioria dos dados da variável Amount está concentrada em valores muito baixos. Alguns outliers de valores altos estão visíveis, mas é importante mantê-los, pois valores extremos podem sim estar relacionados a transações fraudulentas.



A partir desse gráfico a respeito da variável Time, podemos perceber que não existem outliers extremos, tornando ele muito simétrico e uniforme.

É importante ressaltar a necessidade de escalonar as variáveis, para que elas estejam em escalas que se respeitem e gerem qualidade para a base.

Divisão de treino e teste:

Agora a divisão entre o conjunto de treino e o de teste será feita. É importante notar que a divisão foi feita de forma que o conjunto de teste tenha 30% dos dados e o de treino 70%.

```
--- Distribuição Após Divisão Estratificada ---
Total de amostras de treino: 198608
Total de amostras de teste: 85118

Proporção de Classes no Treino:
Class
0    99.83334
1     0.16666
Name: proportion, dtype: float64

Proporção de Classes no Teste:
Class
0    99.833173
1     0.166827
Name: proportion, dtype: float64
```

Normalização/Padronização e Balanceamento:

Finalizando as duas últimas etapas do pré-processamento, vamos realizar a normalização e balanceamento dos conjuntos.

Para ajustar as escalas de Time e Amount, vamos utilizar o StandardScaler para isso:

```
--- Aplicando Standard Scaler ---
Colunas Time e Amount padronizadas em X_train.
Colunas Time e Amount padronizadas em X_test.

Primeiras linhas de X_train após padronização:
            Time        V1        V2        V3        V4        V5        V6  \
257334  1.330896  2.139297 -2.379957 -0.588094 -2.116887 -0.151810  4.930636
118498 -0.417262 -2.150992  0.447565  0.584405  1.181052  0.180772 -1.200828
120679 -0.400015 -6.000510 -5.868708 -1.210423  2.234168 -0.925608  0.019870
163000  0.434726  2.021325 -0.535640 -0.416146 -0.124310 -0.178832  0.918878
171948  0.546950  0.149536  0.807852 -0.590906 -0.509802  1.145394 -1.116329

             V7        V8        V9  ...       V14       V16       V17  \
257334 -3.133495  1.408309  0.431508  ... -1.698677 -0.888080  0.895687
118498 -0.106578  0.771604 -1.441208  ...  1.305198  0.017997  0.076337
120679  0.012067  0.910109 -0.905567  ...  0.811206 -1.228141  0.642475
163000 -1.042599  0.318182  1.072180  ... -0.347979  1.092741 -1.270309
171948  0.935687 -0.234237  0.014499  ... -0.829211  0.439958  0.064027

            V18       V19       V20       V21       V24       V26    Amount
257334  0.094946  0.003088 -0.379746 -0.073778  0.738886  0.044060 -0.308400
118498  0.076609  0.944236  0.028937 -0.066225  0.501689 -0.732049 -0.363961
120679  1.794655 -0.141484  0.545803 -0.132715 -0.747164 -0.205593  1.876398
163000  0.711504  0.305460 -0.078728  0.020950 -0.311931  0.495448 -0.359922
171948  0.898992 -0.346549 -0.122272  0.220651 -0.734172 -0.114120 -0.344877

[5 rows x 23 columns]
```

Para o balanceamento, apenas o conjunto de treino será balanceado, de forma que tenha dados suficientes da classe minoritária para aprender a fraude. Usaremos o smote:

```
--- Contagem de Classes Antes do SMOTE (Treino) ---
Counter({0: 198277, 1: 331})

--- Contagem de Classes Após o SMOTE (Treino) ---
Counter({0: 198277, 1: 198277})
```

A conclusão desta seção é: O resultado do StandardScaler foi que as colunas Time e Amount agora tem seus valores centrados em zero e na mesma ordem de magnitude, o que ajudará ao nosso modelo de machine learning a não ser distorcido por diferenças de escalas; O Smote, por sua vez, criou 197.946 instâncias sintéticas para a classe minoritária, o que permite o modelo realmente verificar as características da fraude e aprendê-las, em vez de ignorá-la.

Agrupamento:

K-Means:

```
--- K-Means (k=2) ---
Número de amostras por cluster: 0    149168
1    134558
Name: count, dtype: int64
Valor do Coeficiente de Silhueta: 0.1190
```

Um valor de Silhueta próximo de zero (como **0.1190**) indica que os grupos (clusters) encontrados pelo K-Means **não estão bem separados**. Isso sugere que a separação dos dados em apenas dois grupos não é clara ou bem definida usando as características atuais, ou que a estrutura interna dos dados é mais complexa do que o K-Means consegue capturar com K = 2.

DBScan:

```
--- DBSCAN ---
Número de clusters encontrados: 103
Número de amostras por cluster:  1      178953
-1       32730
 3       20786
 2       14775
 5       14343
         ...
 102         9
 57          8
 90          7
 80          6
 91          4
Name: count, Length: 104, dtype: int64
Valor do Coeficiente de Silhueta (Excluindo Ruído): -0.0572
```

Um valor de Silhueta **negativo** (como **-0.0572**) é um **resultado ruim** para o agrupamento. Ele indica que, em média, as amostras dentro de um *cluster* estão **mais próximas de amostras de outros *clusters*** do que de suas próprias. Isso confirma que a separação não é coesa. O DBSCAN não conseguiu encontrar dois grupos bem definidos; em vez disso, ele interpretou a estrutura dos dados como muitos pequenos aglomerados de alta densidade e uma grande quantidade de ruído.