CREDIT CARD FRAUD DETECTION

1.COMPREENSÃO DO ESTUDO

O objetivodo trabalho é aplicar técnicas de aprendizado supervisionado para detectar fraudes em transações financeiras com cartão de crédito. Inicialmente, são construídos e comparados diversos modelos de classificação, como Árvores de Decisão, Regressão Logística, KNN, Floresta Aleatória, SVC, SGDClassifier e Naive Bayes, utilizando métricas como recall, precisão e F1-score para selecionar os modelos mais promissores a serem estudados ao longo do trabalho. **Devido ao alto desbalanceamento dos dados, a métrica recall é priorizada para maximizar a detecção de fraudes reais**, adotando técnicas de balanceamento como undersampling e oversampling para melhorar o desempenho dos modelos.

Além da comparação inicial de modelos, **o foco principal do trabalho está na análise detalhada e no desempenho comparativo dos classificadores de Regressão Logística e Floresta Aleatória.** Esses modelos são treinados e avaliados com diferentes estratégias de balanceamento de classes, empregando métricas como acurácia, precisão, recall e F1-score para medir sua eficácia. A análise de curvas ROC e a métrica AUC também são exploradas para compreender melhor o comportamento desses classificadores.

Adicionalmente, **o trabalho realiza um processo de seleção de variáveis (feature selection)** para identificar os atributos mais relevantes na detecção de fraudes, aprimorando a eficiência e a precisão dos modelos. **Também são utilizadas validação cruzada e otimização de hiperparâmetros para ajustar os parâmetros dos classificadores**, visando maximizar a performance preditiva. Esse refinamento resulta em uma abordagem mais robusta e confiável para a detecção de fraudes, garantindo que os modelos estejam configurados para alcançar o melhor equilíbrio entre precisão e generalização.

1.1 MÉTRICAS

Para o estudo e análise comparativa dos resultados que serão adquiridos, serão consideradas as métricas recall e precisão. Durante a apresentação dos resultados, também será exibida a acurácia obtida; no entanto, essa não será a principal métrica, **visto que, conforme analisado abaixo, o conjunto de dados apresenta um alto desbalanceamento dos dados da variável alvo. Logo, a acurácia não é a melhor métrica nesses casos.** Como o estudo tem como finalidade identificar todos os casos realmente fraudulentos, a métrica de recall será a mais importante no critério de escolha, sendo sempre balanceada e analisada em conjunto com um valor coerente de precisão.

2.COMPREENSÃO DO ESTUDO

O presente conjunto de dados que será trabalhado nesse projeto, está disponível no seguinte **Credit Card Fraud Detection**

O conjunto de dados contém transações feitas por cartões de crédito em setembro de 2013 por titulares de cartão europeus. [1]

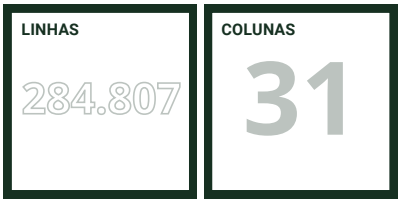
Este conjunto de dados apresenta **transações que ocorreram em dois dias, onde temos 492 fraudes de 284.807 transações. O conjunto de dados é altamente desbalanceado, a classe positiva (fraudes) é responsável por 0,172% de todas as transações.** [1]

INFORMAÇÕES DA BASE:

- 'Time'(tempo): contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados;
- 'Amount'(valor): é o valor da transação;
- 'Class' (classe): é a variável de resposta e assume valor 1 em caso de fraude e 0 caso contrário;
- **Devido a questões de confidencialidade, não podemos fornecer os recursos originais e mais informações básores sobre os dados. Características V1, V2 ... V28 são os princípais componentes obtidos com "PCA".**

3.ANÁLISE INICIAL

Nesta etapa, será realizada uma análise preliminar do conjunto de dados para compreender suas principais características e identificar possíveis desafios que possam influenciar o desempenho dos modelos de aprendizado supervisionado. O foco inicial é verificar a qualidade e a estrutura dos dados, avaliar a distribuição da variável alvo e definir estratégias para lidar com problemas, como o desbalanceamento de classes.

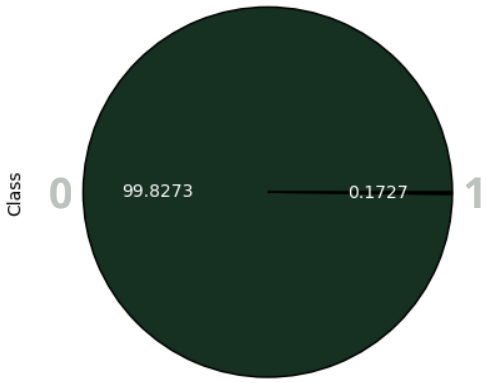


O conjunto de dados utilizado neste projeto contém **284.807 registros e 31 colunas.**

```
<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
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

	Time	Amount	Class
count	284807.00	284807.00	284807.00
mean	94813.86	88.35	0.00173
std	47488.15	250.12	0.042
min	0.00	0.00	0.00
25%	54201.50	5.60	0.00
50%	84692.00	22.00	0.00
75%	139320.50	77.17	0.00
max	172792.00	25691.16	1.00

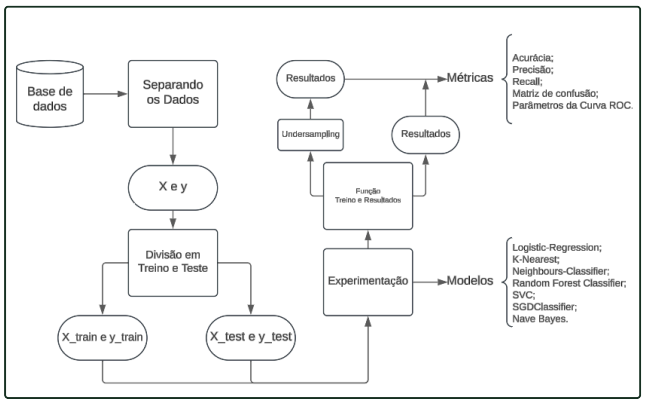
O resumo estatístico inicial sugere uma ampla variação nas variáveis, especialmente em Amount e Time, reforçando a necessidade de normalização ou padronização em etapas posteriores.



Esse desbalanceamento exige o uso de técnicas específicas, como o balanceamento das classes, para garantir a eficácia dos modelos de classificação. **Caso contrário, o modelo pode apresentar alta acurácia geral, mas com baixa eficiência na identificação de fraudes.**

4.ESCOLHA DO MODELO DE ESTUDO

Nesta seção, serão analisados diferentes algoritmos de classificação disponíveis na biblioteca scikit-learn do Python, voltados para o aprendizado supervisionado. O principal objetivo é avaliar o desempenho desses modelos quando aplicados ao conjunto de dados, com ênfase nas métricas curva ROC, recall e precisão. Com base nesses indicadores, serão selecionados os dois modelos com o melhor desempenho geral para a aplicação em questão. Além disso, ao longo das seções do trabalho, será explorado o comportamento desses modelos utilizando diferentes técnicas e abordagens, examinando como as métricas de avaliação variam conforme as estratégias adotadas.



Esta primeira tentativa é com o conjunto de dados original, porém, **vamos perceber que, por questões de tempo e pela quantidade de dados, é mais vantajoso aplicar o undersampling para reduzir a quantidade de dados. Veremos que o resultado obtido é semelhante.**

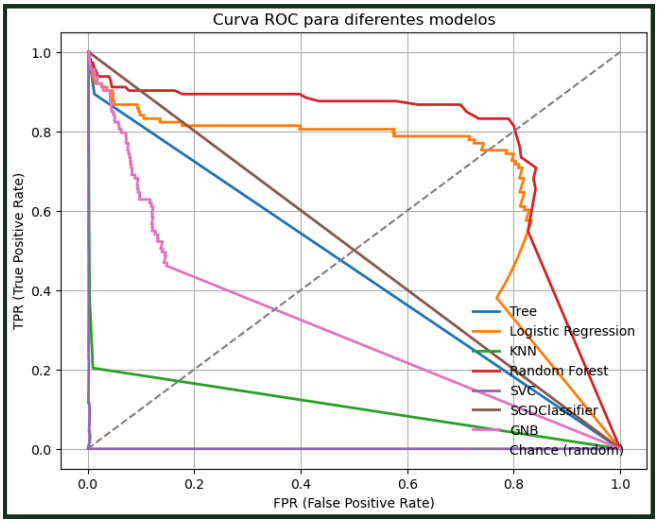
Treinamento dos Modelos

43,7 min

com undersampling

31.2 s

PRIMEIROS RESULTADOS



Modelos	Modelos	Acurácia	Precisão	Recall	F1-Score
1	SGDClassifier	0.0016	0.0016	1.0000	0.0032
2	Logistic Regression	0.9458	0.0263	0.9204	0.0511
3	Random Forest	0.9720	0.0494	0.9115	0.0937
4	Tree	0.8884	0.0125	0.8938	0.0248
5	GNB	0.9862	0.0805	0.7345	0.1452
6	KNN	0.6548	0.0026	0.5752	0.0053
7	SVC	0.4827	0.0017	0.5487	0.0034