Análise da Detecção de Fraudes em Transações Financeiras com Modelos de Aprendizado de Máquina

Frederico dos Santos, Gabriel de Souza, Gabriel Lima de Souza, Nikolas Louret

Sumário

01
introdução

Contexto

Crescimento das fraudes financeiras com o aumento das transações online

Impactos negativos na estabilidade econômica e confiança dos clientes

Desafio crescente devido à complexidade das operações financeiras

Métodos de Detecção de Fraudes Bancárias

Detecção de Uso Indevido:

- classifica transações com base em padrões conhecidos de fraude

Detecção de Anomalias:

- identifica desvios em relação ao comportamento normal com base em dados históricos.

Objetivo

Desenvolver um modelo de Machine Learning para detecção de fraudes com base em dados reais (Kaggle)

02
Metodologia

FERRAMENTAS

Configuração do Ambiente e Carregamento dos Dados

Configuração do Ambiente:

- Google Colab

Carregamento do Dataset:

- Fonte de dados: Kaggle – **Credit Card Fraud Detection**
- Utilização da biblioteca Pandas para importar os dados em um DataFrame

Análise Inicial do Dataset:

- Transações anonimizadas
- Dados reais de empresas europeias (Setembro de 2013)

Distribuição das Classes:

- **Não fraudulentas:** 99,828%
- **Fraudulentas:** 0,172%

Pré-processamento e Balanceamento dos Dados

Análise inicial:

- Verificação de valores ausentes ou duplicados.

Distribuição inicial

- Desequilíbrio entre classes identificada

Balanceamento

- Técnica aplicada: **Undersampling**
- **Motivação:** Equilibrar as classes e evitar vieses do modelo

Resultado(conjunto baseado em 984 transações)**:**

- **Não fraudulentas:** 50%
- **Fraudulentas:** 50%

Modelos de Aprendizado de Máquina

- **Random Forest**
- **Support Vector Machine (SVM)**
- **Logistic Regression**
- **Naive Bayes**
- **K-Nearest Neighbors (KNN)**

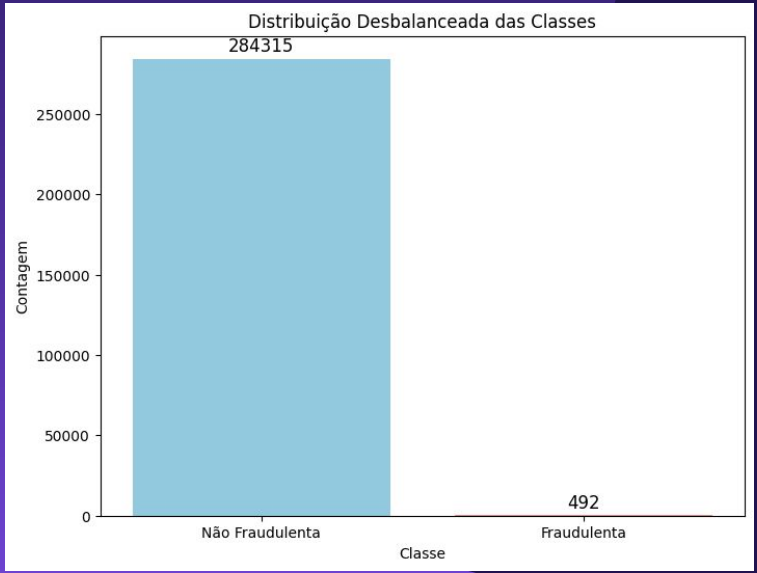
objetivo: comparar o desempenho na detecção de fraudes

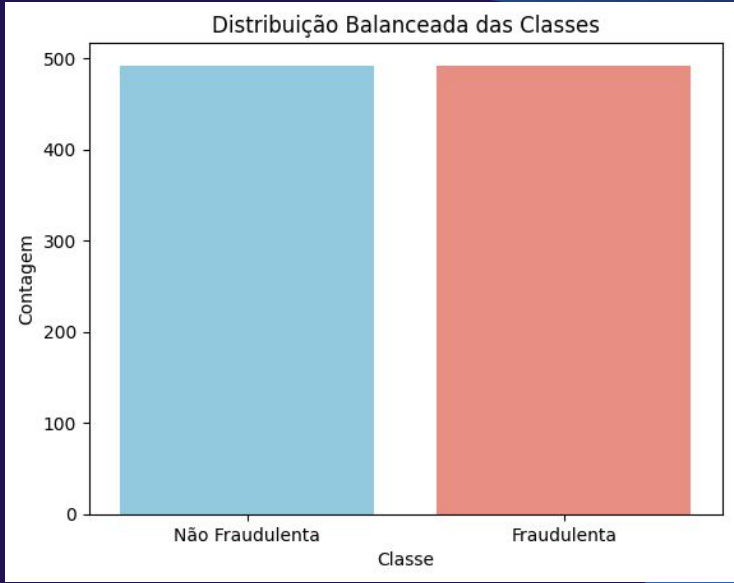
Avaliação dos Modelos

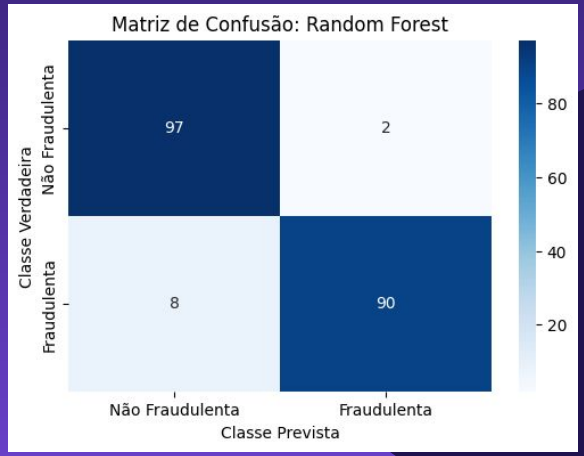
- **Acurácia**
- **Matriz de confusão**
- **Precisão, revocação e F1-score**
- **AUC da curva ROC**

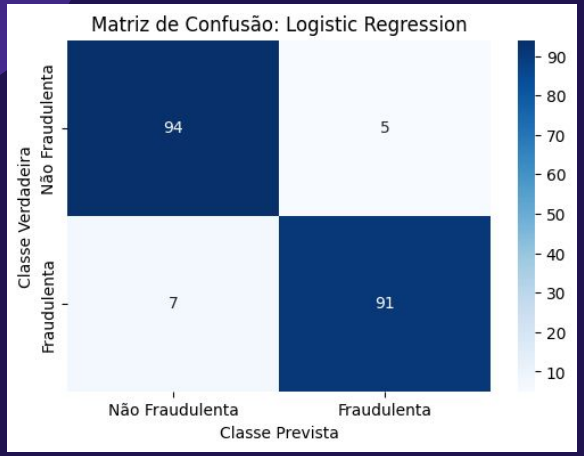
justificativa: análise ampla e detalhada do desempenho

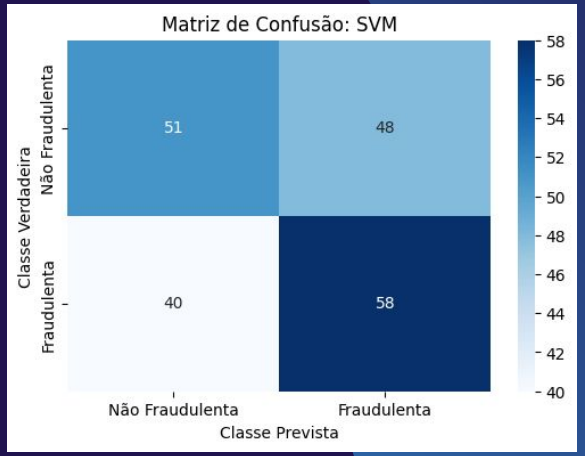
03
Resultados

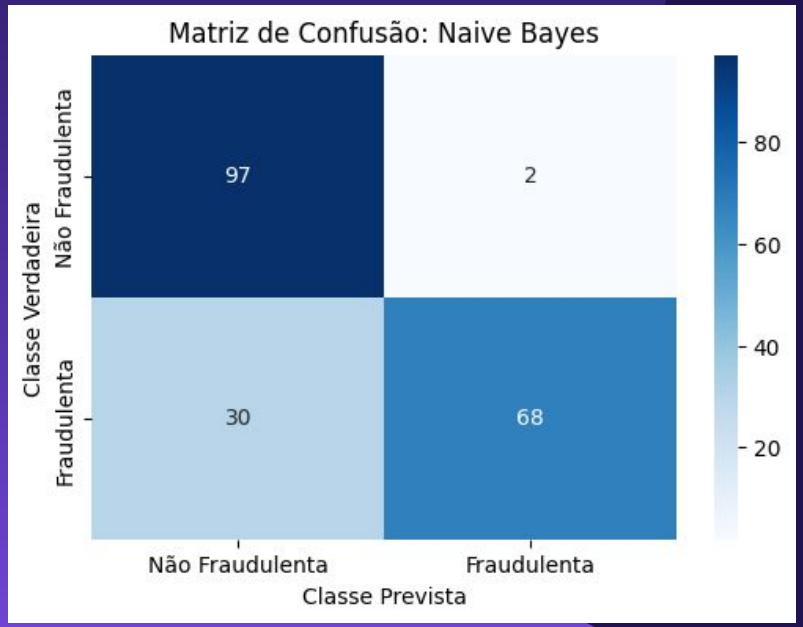

Distribuição Desbalanceada das Classes
284315
492
250000
200000
150000
100000
50000
0
Contagem
Não Fraudulenta
Fraudulenta
Classe

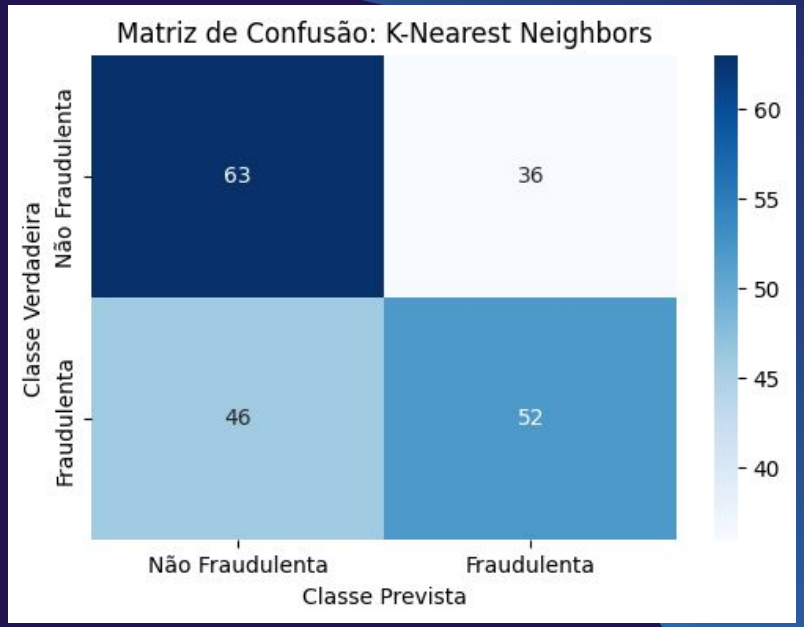

Distribuição Balanceada das Classes
500
400
300
200
100
0
Contagem
Não Fraudulenta
Fraudulenta
Classe

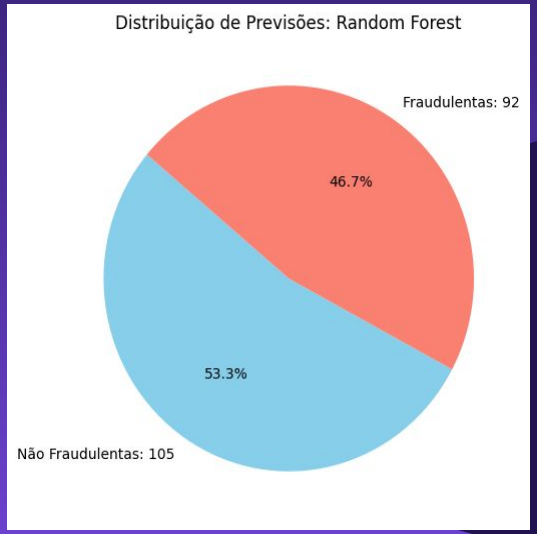

Matriz de Confusão: Random Forest
Classe Verdadeira
Não Fraudulenta
Fraudulenta
97
2
8
90
Não Fraudulenta
Fraudulenta
Classe Prevista
80
60
40
20

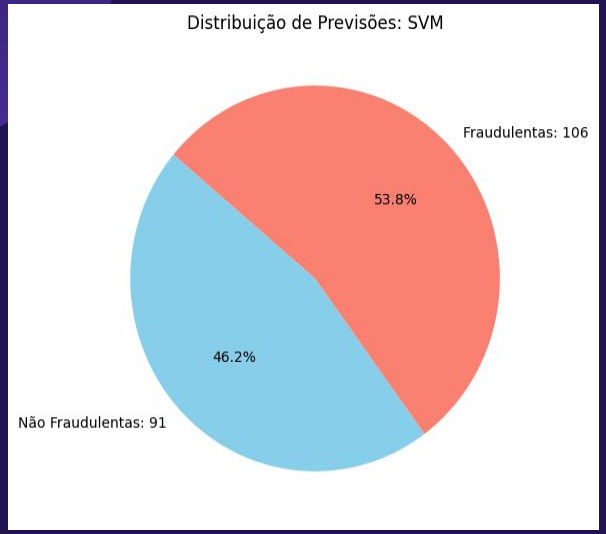

Matriz de Confusão: Logistic Regression
Classe Verdadeira
Não Fraudulenta
Fraudulenta
94
5
7
91
Não Fraudulenta
Fraudulenta
Classe Prevista
90
80
70
60
50
40
30
20
10

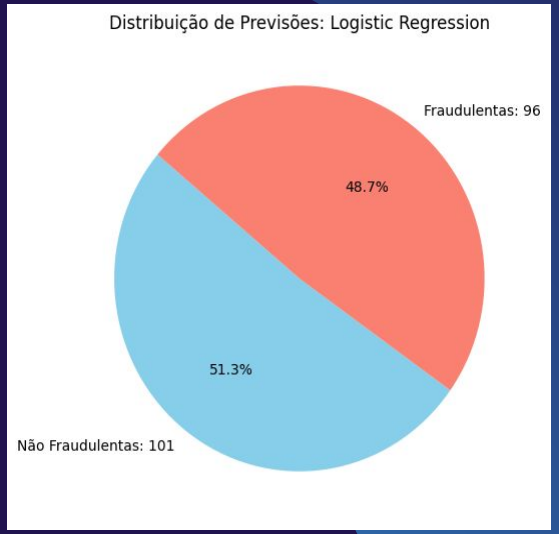

Matriz de Confusão: SVM
Classe Verdadeira
Não Fraudulenta
Fraudulenta
51
48
40
58
Não Fraudulenta
Fraudulenta
Classe Prevista
58
56
54
52
50
48
46
44
42
40

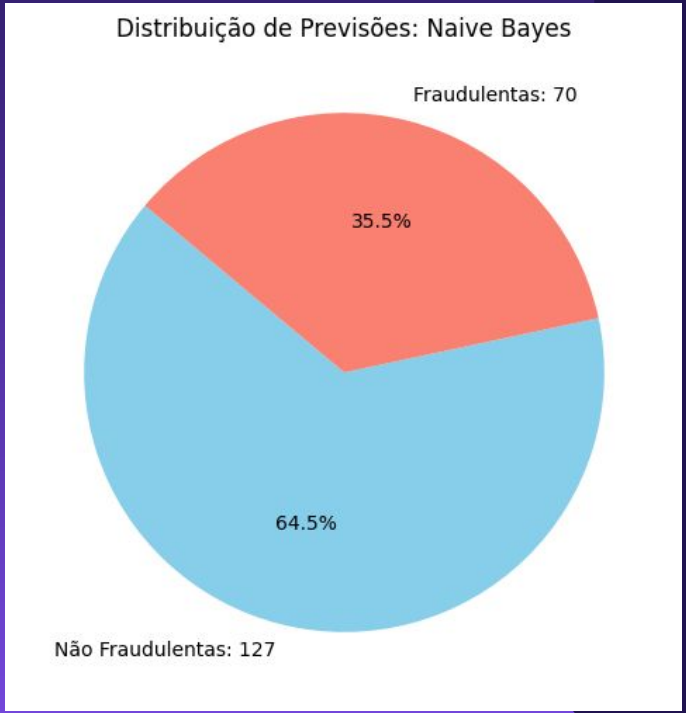

Matriz de Confusão: Naive Bayes
Classe Verdadeira
Não Fraudulenta
Fraudulenta
97
2
30
68
Não Fraudulenta
Fraudulenta
Classe Prevista
80
60
40
20

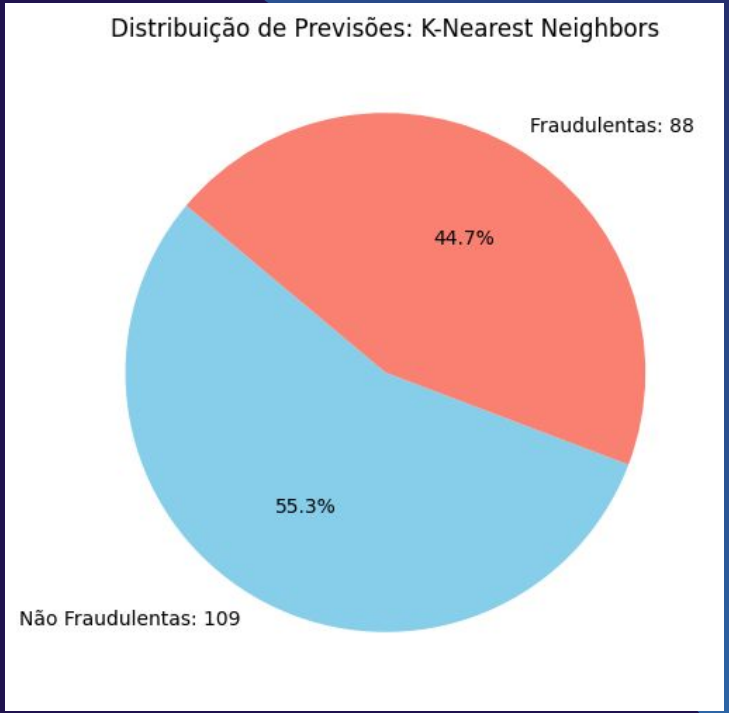

Matriz de Confusão: K-Nearest Neighbors
Classe Verdadeira
Não Fraudulenta
Fraudulenta
63
36
46
52
Não Fraudulenta
Fraudulenta
Classe Prevista
60
55
50
45
40

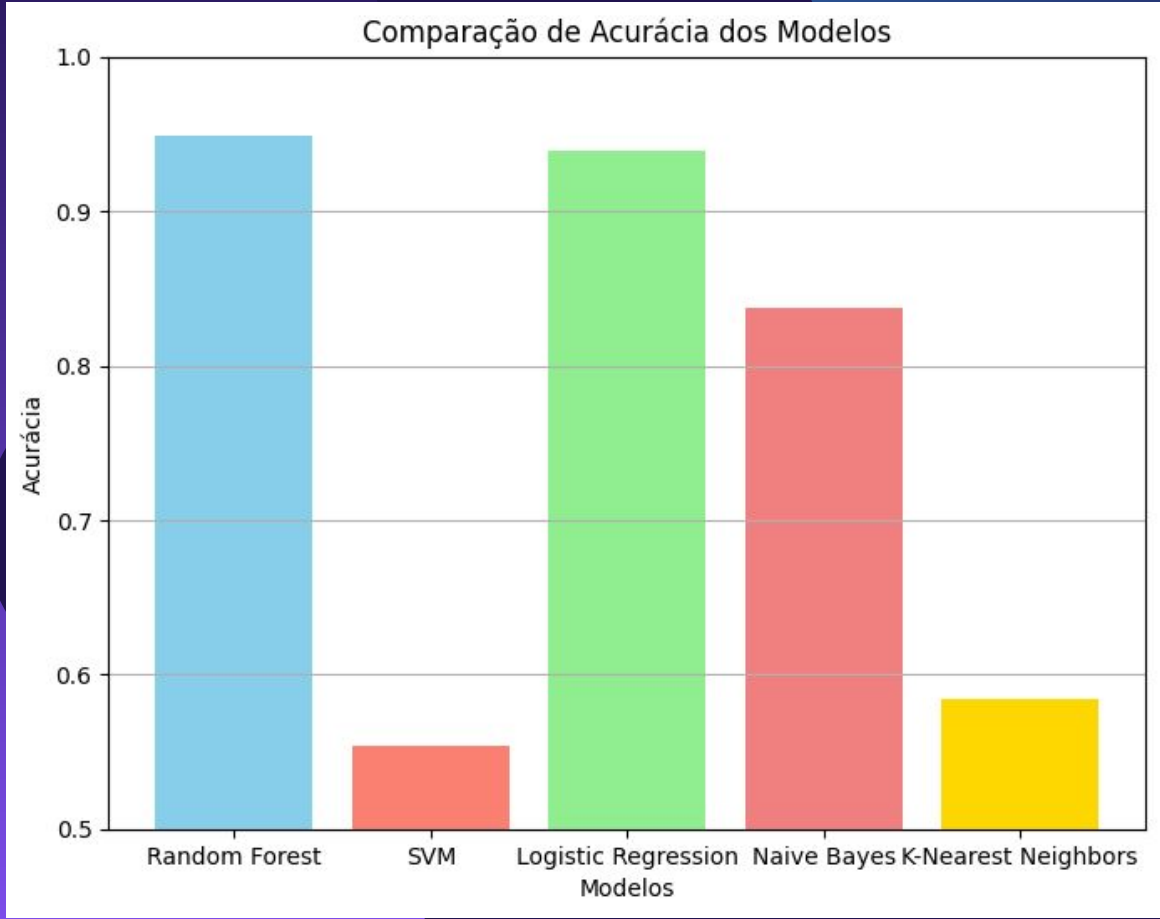

Distribuição de Previsões: Random Forest
Fraudulentas: 92
46.7%
53.3%
Não Fraudulentas: 105

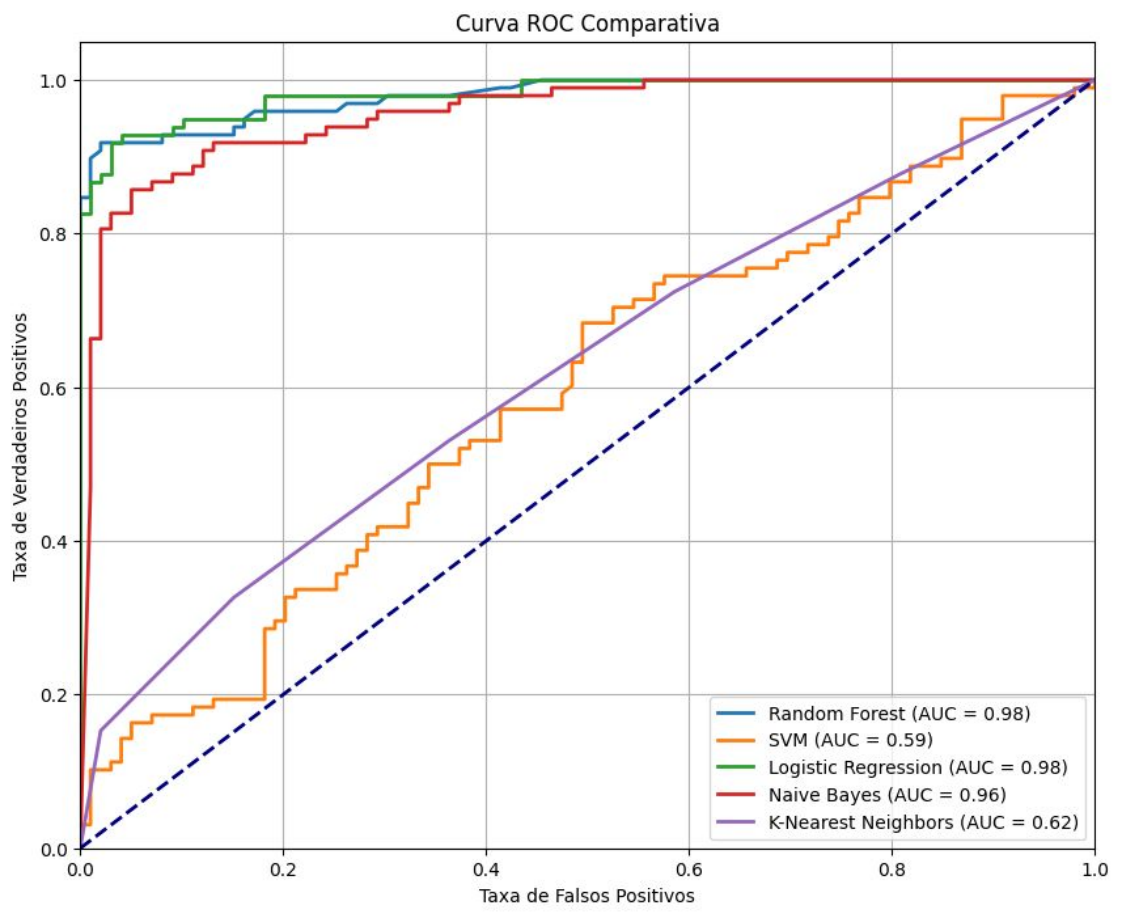

Distribuição de Previsões: SVM
Fraudulentas: 106
53.8%
46.2%
Não Fraudulentas: 91


Distribuição de Previsões: Logistic Regression
Fraudulentas: 96
48.7%
51.3%
Não Fraudulentas: 101


Distribuição de Previsões: Naive Bayes
Fraudulentas: 70
35.5%
64.5%
Não Fraudulentas: 127


Distribuição de Previsões: K-Nearest Neighbors
Fraudulentas: 88
44.7%
55.3%
Não Fraudulentas: 109


Comparação de Acurácia dos Modelos
1.0
0.9
0.8
0.7
0.6
0.5
Acurácia
Random Forest
SVM
Logistic Regression
Naive Bayes
K-Nearest Neighbors
Modelos


Curva ROC Comparativa
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 0.98)
SVM (AUC = 0.59)
Logistic Regression (AUC = 0.98)
Naive Bayes (AUC = 0.96)
K-Nearest Neighbors (AUC = 0.62)

04

Conclusão

Desempenho dos Modelos e Impacto do Balanceamento

Desempenho dos Modelos

- **Random Forest (RF)**: Melhor desempenho geral (AUC > 0,97, acurácia de 94,92%, F1-score de 93,69%).
- **Logistic Regression (LR)**: Alta acurácia e equilíbrio entre precisão e revocação.
- **SVM e KNN**: Desempenho insatisfatório (acurácia < 60%).
- **Naive Bayes (NB)**: Boa revocação, mas baixa precisão (muitos falsos positivos).

Impacto do Balanceamento (Undersampling)

- **RF e LR**: Menos sensíveis à redução da classe majoritária, mantendo boa generalização.
- **SVM e KNN**: Desempenho comprometido pela alteração na distribuição das classes.

Conclusões Gerais e Trabalhos Futuros

Conclusões Gerais

- RF e LR são adequados para detectar fraudes em bases desbalanceadas.
- Técnicas de balanceamento e escolha do modelo impactam diretamente os resultados.

Trabalhos Futuros

- Explorar técnicas avançadas de balanceamento (como o SMOTE).
- Avaliar modelos mais complexos, como redes neurais, para maior precisão e robustez.

Obrigado