EDA e aplicação de Modelos supervisionados

na analise da dados de rede.

Matéria: 2025_2 - TÓPICOS COMPUTACIONAIS EM CIÊNCIA DE DADOS

Professora: Michele Nogueira Lima

Aluno: Lucas Albano Jansen

Curso: Especialização em ciência de dados com foco em ciência da computação

UFMG

UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Git: https://github.com/PlumbBobGaius/SG_atividade3

Resumo

Este projeto tem como objetivo aplicar técnicas de Machine Learning para detectar sites de phishing com base em características técnicas extraídas das URLs e páginas web.

- Exploração e limpeza dos dados
- Engenharia de features
- Treinamento de modelos de classificação
- Avaliação de desempenho
- Proposição de melhorias

Todo esse processo também pode ser acompanhado no Notebook disponível em: Github

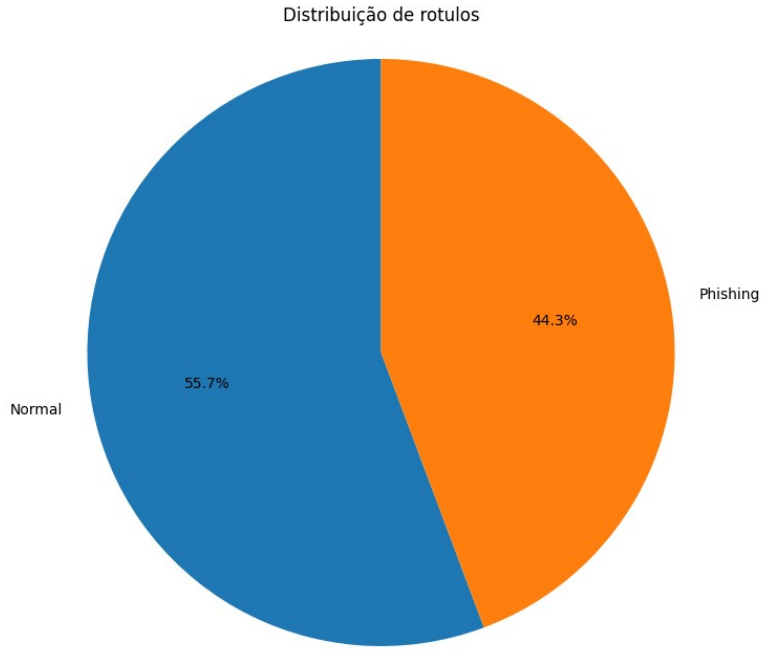
Distribuição dos Rótulos

O conjunto de dados utilizado apresenta duas categorias principais:

- Normal (1): 55.7%
- Phishing (-1): 44.3%

Essa distribuição é visualizada no gráfico de pizza ao lado, que mostra uma proporção relativamente equilibrada entre os dois tipos de instância.

Esse equilíbrio inicial permite que os algoritmos tenham uma base justa para aprender padrões de comportamento associados a ataques de phishing.

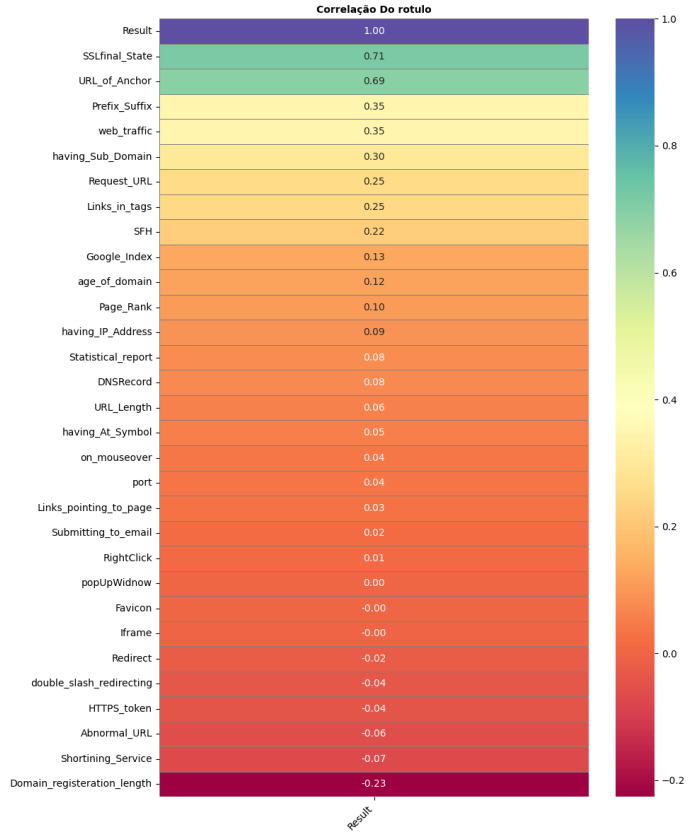


Correlação com o Rótulo

Objetivo da análise

Identificar quais variáveis têm maior influência sobre o rótulo de saída — ou seja, quais características estão mais associadas a sites phishing ou normais.
Destaques da correlação:

- SSLfinal_State: forte correlação positiva com sites normais. Presença de certificado SSL é um sinal de legitimidade.
- Prefix_Suffix: correlação negativa com o rótulo. Domínios com hífen são comuns em sites maliciosos.
- Domain_registration_length: correlação negativa de -0.23. Domínios com registro curto tendem a ser usados em campanhas de phishing temporárias.

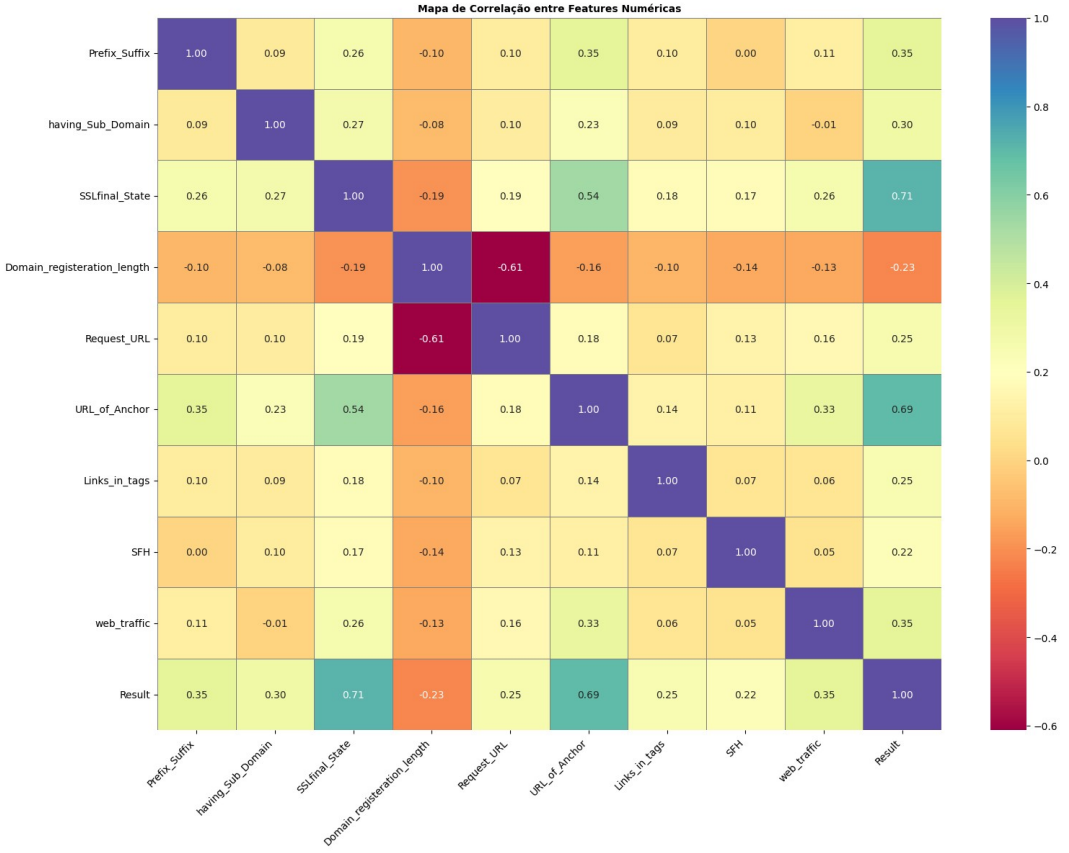


Seleção de Features Relevantes

Foram escolhidas as variáveis com correlação superior a 0.20 ou inferior a -0.20 em relação ao rótulo Result. Esse filtro permite focar nas features com maior poder discriminativo entre sites phishing e normais.
Principais variáveis selecionadas:

- SSLfinal_State
- Domain_registration_length
- Prefix_Suffix
- URL_of_Anchor
- Request_URL
- web_traffic

Essas variáveis apresentam padrões técnicos e estruturais que ajudam a identificar comportamentos maliciosos, como ausência de SSL, domínios temporários ou uso de hífen em URLs.

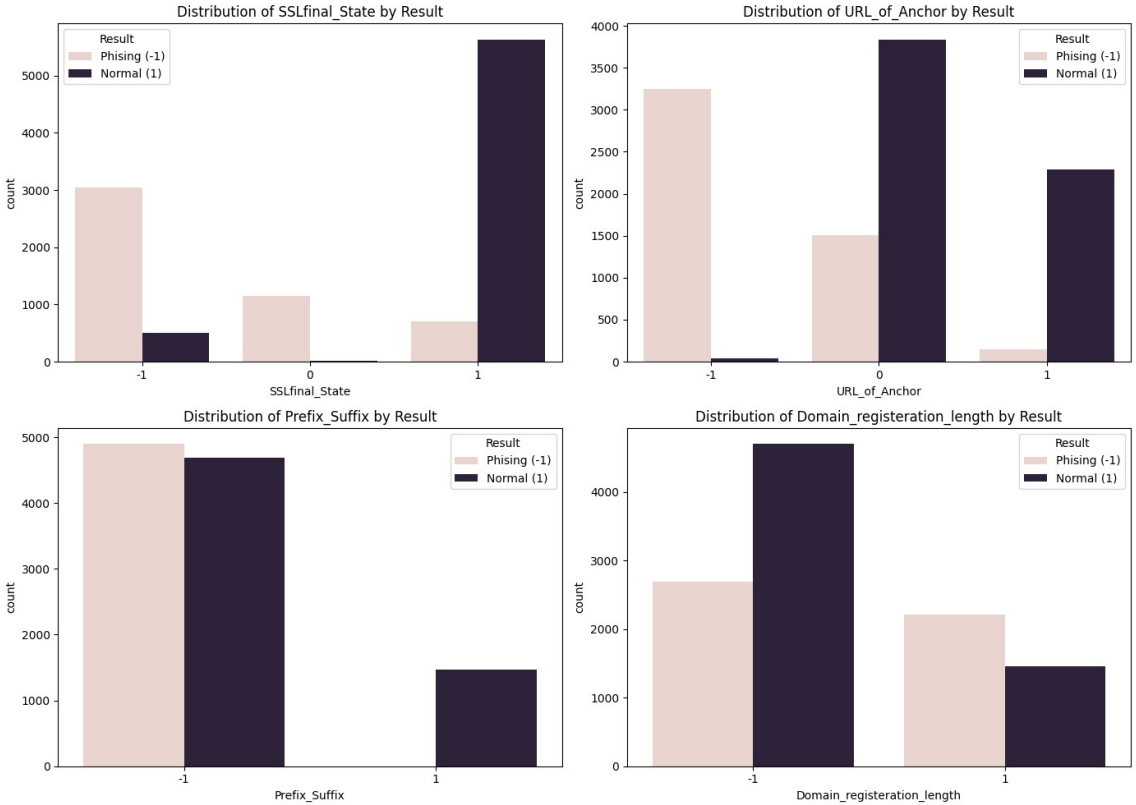


Distribuição de Features por Classe

Os gráficos ao lado mostram como algumas das principais features se distribuem entre sites phishing e normais. Essa visualização reforça padrões detectados na análise de correlação.
Destaques:

- SSLfinal_State Sites normais geralmente possuem certificado SSL válido (valor = 1), enquanto sites phishing tendem a não ter (valor = -1).
- URL_of_Anchor Sites phishing frequentemente usam âncoras quebradas ou vazias (valor = -1), enquanto sites legítimos mantêm links funcionais (valor = 1).
- Prefix_Suffix O uso de hífen no domínio (valor = -1) é comum em URLs fraudulentas. Sites legítimos evitam esse padrão (valor = 1).
- Domain_registration_length Domínios com registro curto (valor = -1) sã mais comuns em campanhas de phishing. Domínios duradouros (valor = 1) indicam maior confiabilidade

Essas variáveis capturam padrões estruturais e temporais que ajudam a distinguir comportamentos maliciosos, sendo fundamentais para a construção de modelos preditivos eficazes.



Modelos e Hiperparâmetros

Algoritmos Utilizados

Foram testados dois modelos de classificação supervisionada:

Logistic Regression Modelo linear que estima probabilidades com base em uma função logística. Ideal para problemas binários e interpretáveis.

Decision Tree Modelo baseado em regras de decisão. Segmenta os dados em nós com base nas features mais informativas.

Configurações Testadas

Para avaliar o impacto dos hiperparâmetros, foram criadas pipelines com diferentes ajustes:

Logistic Regression

C = 0.1: regularização mais forte

C = 1.0: regularização padrão

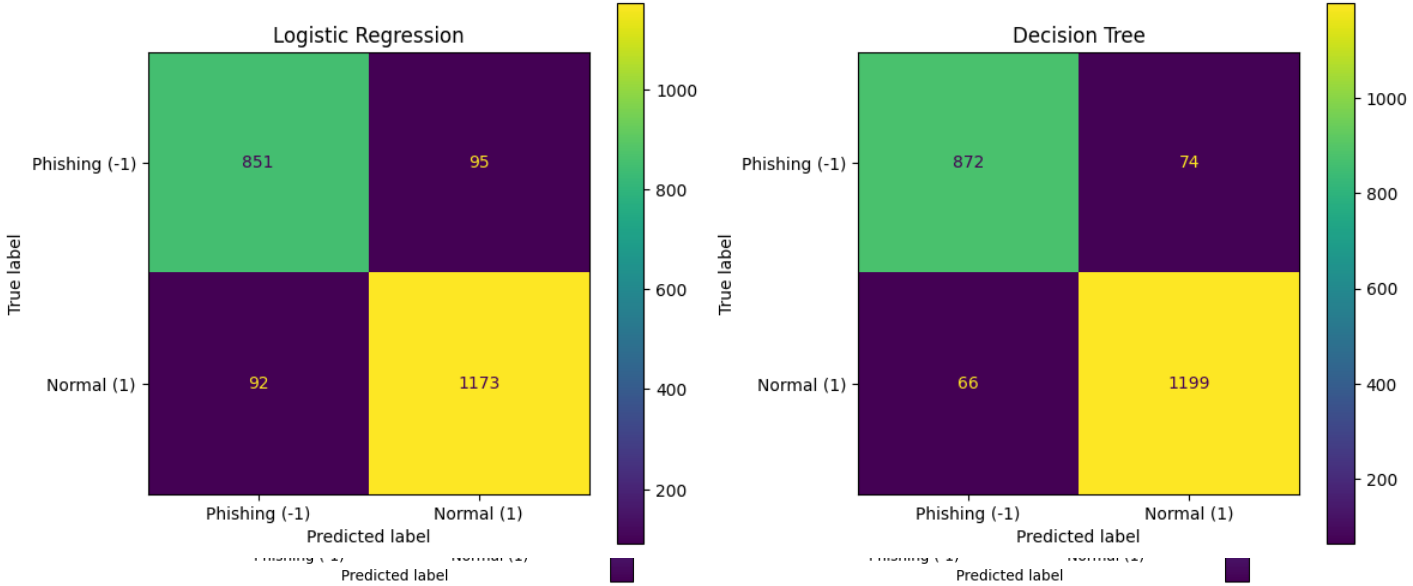
Decision Tree

max_depth = 3: árvore mais simples e generalista

max_depth = 5: árvore mais profunda e detalhada

Matrizes de Confusão — Comparação de Modelos (sem Hiperparâmetros)

As matrizes de confusão ao lado mostram o desempenho dos modelos Logistic Regression e Decision Tree na classificação de sites phishing e normais.



Logistic Regression

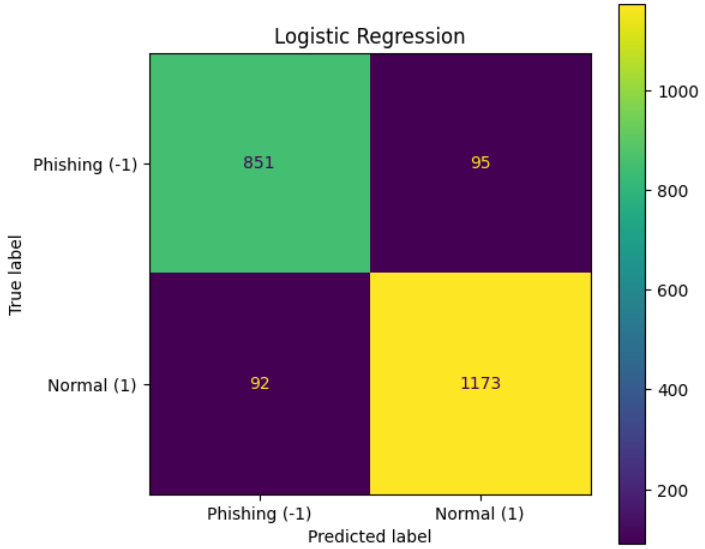
Verdadeiros Positivos (Phishing detectado corretamente): 851

Falsos Negativos (Phishing classificado como normal): 95

Falsos Positivos (Normal classificado como phishing): 92

Verdadeiros Negativos (Normal detectado corretamente): 1173

Acurácia: 92%



Decision Tree

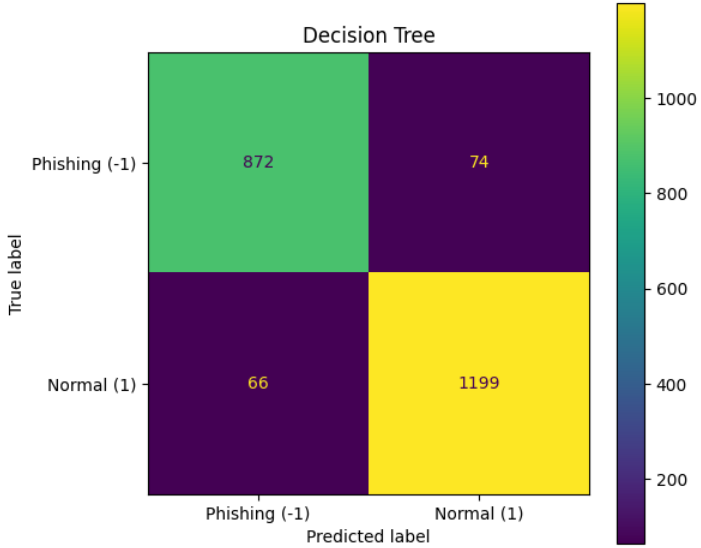
Verdadeiros Positivos: 872

Falsos Negativos: 74

Falsos Positivos: 66

Verdadeiros Negativos: 1199

Acurácia: 94%



O modelo de árvore de decisão apresenta melhor desempenho geral, com menos erros em ambas as classes.

Curva ROC — Comparação entre Modelo

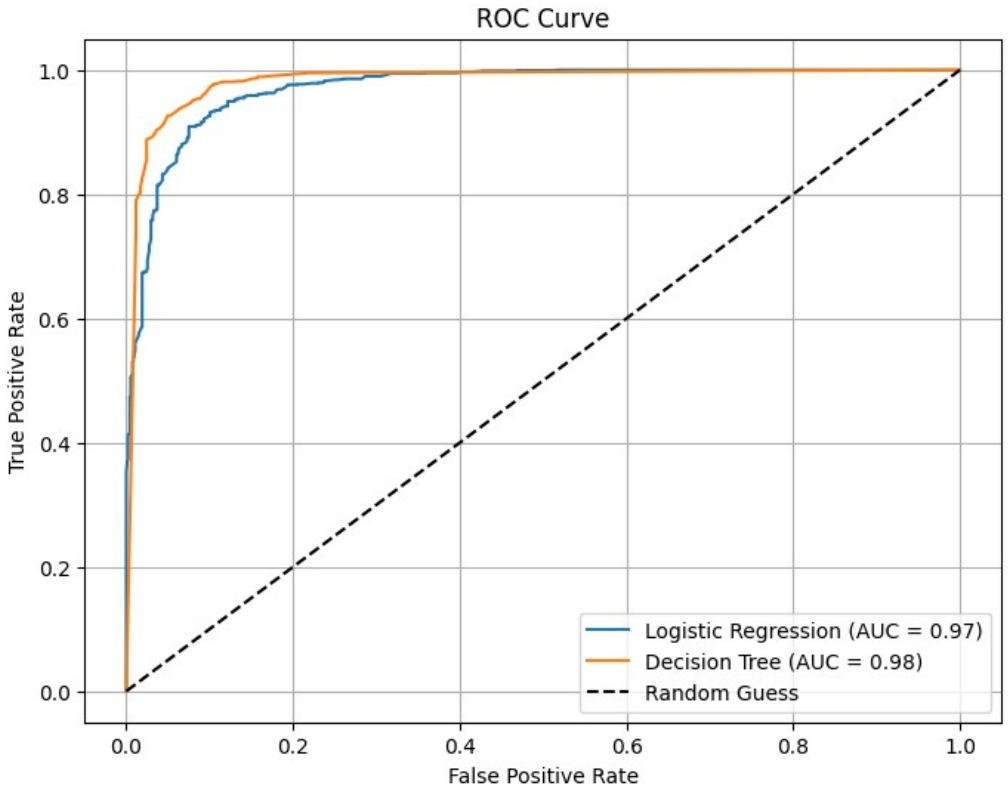
Logistic Regression

- AUC = 0.97
- Curva azul
- Excelente desempenho, com alta sensibilidade e especificidade

Decision Tree

- AUC = 0.98
- Curva laranja
- Leve superioridade em relação ao modelo logístico

Ambos os modelos reais apresentam curvas próximas ao canto superior esquerdo, indicando alto poder de separação entre sites phishing e normais. A árvore de decisão se destaca levemente, reforçando sua performance observada nas demais métricas.



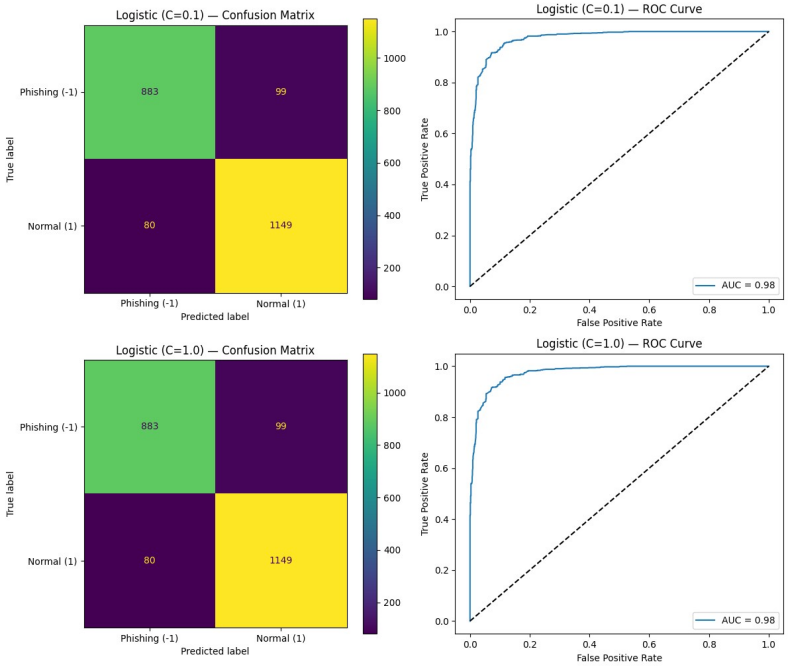
Avaliação com Hiperparâmetros Ajustados

Logistic Regression C = 0.1 e C = 1.0:

Acurácia: 92%
AUC: 0.98
Phishing: F1-score = 0.91
Normal: F1-score = 0.93
Acurácia geral: 92%

Desempenho estável e consistente

O modelo de Logistic Regression, testado com regularizações diferentes (C = 0.1 e C = 1.0), apresentou desempenho estável e consistente, com acurácia de 92% e AUC de 0.98 em ambos os casos. As métricas de F1-score para phishing (0.91) e normal (0.93) confirmam sua eficácia na separação entre classes, mesmo sob diferentes níveis de penalização.



Mapa de correlação completo

Decision Tree (depth=3):

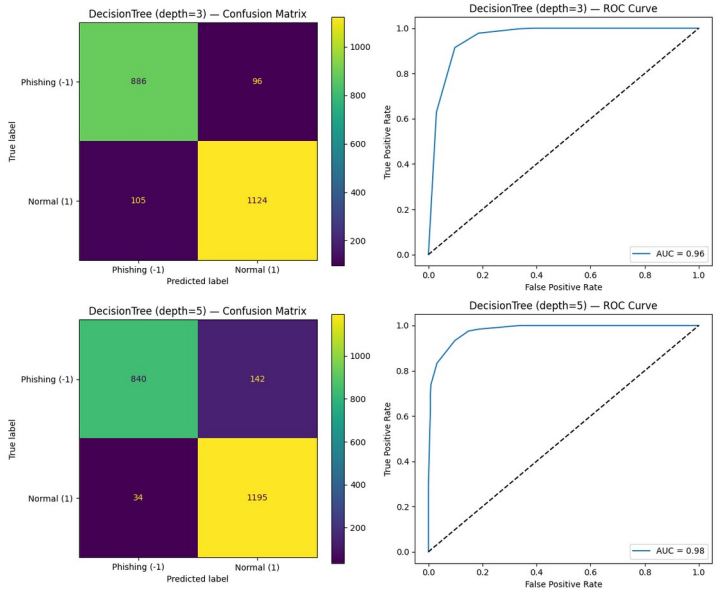
Acurácia: 91%
AUC: 0.91
Phishing: F1-score = 0.90
Normal: F1-score = 0.92
Acurácia geral: 91%

Modelo mais simples, mas com desempenho inferior

Decision Tree (depth=5):

Acurácia: 92%
AUC: 0.98
Phishing: F1-score = 0.91
Normal: F1-score = 0.93
Acurácia geral: 92%

Melhor separação entre classes, com destaque para revocação da classe normal (0.97)



O modelo Decision Tree com profundidade 5 apresenta o melhor equilíbrio entre precisão e revocação, especialmente na detecção de sites legítimos. Os modelos logísticos também se mostram consistentes e confiáveis, com desempenho estável em diferentes configurações.

Conclusões

O modelo Decision Tree com profundidade 5 apresentou o melhor desempenho geral, com alta acurácia (92%), AUC de 0.98 e excelente revocação para a classe normal (0.97).

Os modelos de Logistic Regression mostraram-se estáveis e confiáveis, com desempenho consistente mesmo sob diferentes níveis de regularização.

A análise exploratória revelou que features técnicas simples, como presença de SSL, uso de hífen no domínio e tempo de registro, são altamente eficazes na detecção de phishing.