Análise dos Dados de Inadimplência: Perfil e Comportamento dos Devedores

ALEXANDER BANDEIRA LIRA & DIEGO WILSON PEREIRA DE ANDRADE

1. Introdução

Nesta análise, será realizado um estudo sobre inadimplência (Termo utilizado para descrever a situação em que uma pessoa ou empresa não cumpre com o pagamento de uma dívida ou obrigação financeira), tendo como objetivo identificar padrões e fatores que influenciam o comportamento dos clientes em relação ao pagamento de suas dívidas.

2. Fundamentos Teóricos e Metodológicos

Utilizando um conjunto de dados com variáveis socioeconômicas e comportamentais, vamos construir dois modelos de machine learning diferentes (Decision Tree & Random Forest) para prever a inadimplência. Os modelos serão comparados em termos de desempenho, permitindo avaliar qual deles oferece as melhores previsões e pode ser mais eficaz para tomada de decisões no contexto da análise de crédito.

3. Aplicação

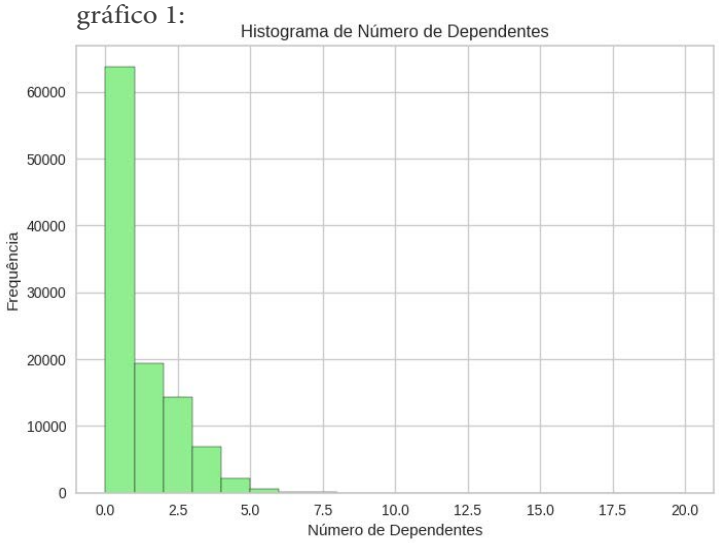
Inicialmente é feito o carregamento das bibliotecas necessárias, em seguida é feito o carregamento e o tratamento dos dados.

Apresentando as variáveis presentes no decorrer do Artigo:

- **inadimplente:** variável de resposta do modelo, indicando se o cliente está inadimplente (1 para inadimplente, 0 para adimplente);
- **util_linhas_inseguras:** percentual de utilização das linhas de crédito inseguras;
- **idade:** idade do cliente;
- **vezes_passou_de_30_59_dias:** número de vezes que o cliente passou de 30 a 59 dias em atraso no pagamento;
- **razao_debito:** razão entre o valor do débito e o total do crédito disponível;
- **salario_mensal:** salário mensal do cliente;
- **numero_linhas_crdto_aberto:** número de linhas de crédito abertas atualmente em nome do cliente;
- **numero_vezes_passou_90_dias:** número de vezes que o cliente passou 90 dias ou mais em atraso;
- **numero_emprestimos_imobiliarios:** número de empréstimos imobiliários que o cliente possui;
- **numero_de_vezes_que_passou_60_89_dias:** número de vezes que o cliente passou de 60 a 89 dias em atraso no pagamento;
- **numero_de_dependentes**: número de dependentes do cliente.

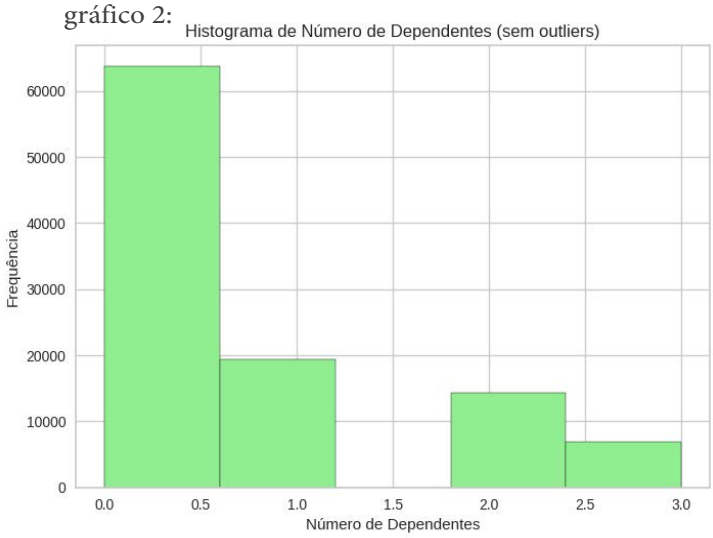
Gerando o histograma da variável numero_de_dependentes:

gráfico 1:



É feito a verificação dos percentis para limitar valores extremos.

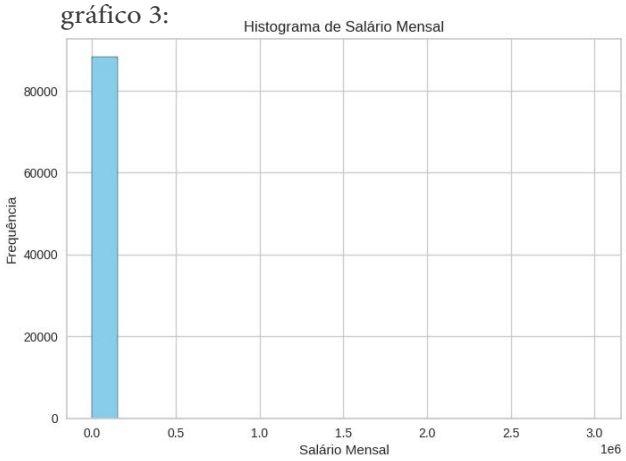
gráfico 2:



podemos observar que a grande maioria não possuí dependentes.

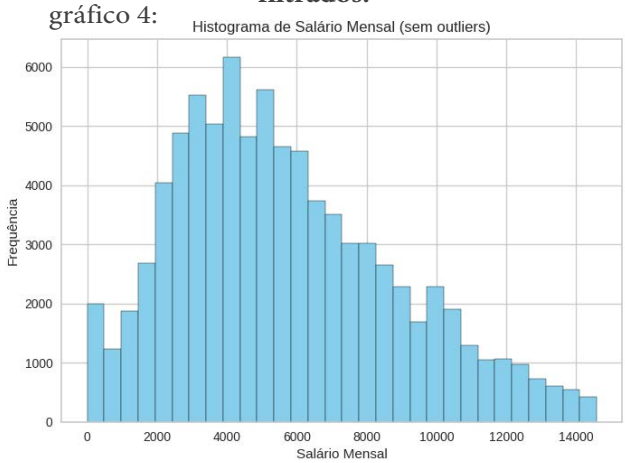
histograma da variável salario_mensal:

gráfico 3:



histograma da variável salario_mensal com valores filtrados:

gráfico 4:



A partir da distribuição dos dados é possível que na variável salario mensal possamos utilizar para substituir os valores nulos tanto a média como a mediana categorizado por idade por conta de termos uma distribuição aparentemente normal, mas como alcançamos isso apenas com a retirada dos outliers, verificaremos a performance primeiro com a mediana.
Já o número de dependentes mediana provavelmente será uma opção melhor por conta da assimetria encontrada.

	0
inadimplente	0
util_linhas_inseguras	0
idade	0
vezes_passou_de_30_59_dias	0
razao_debito	0
salario_mensal	3
numero_linhas_crdto_aberto	0
numero_vezes_passou_90_dias	0
numero_emprestimos_imobiliarios	0
numero_de_vezes_que_passou_60_89_dias	0
numero_de_dependentes	0

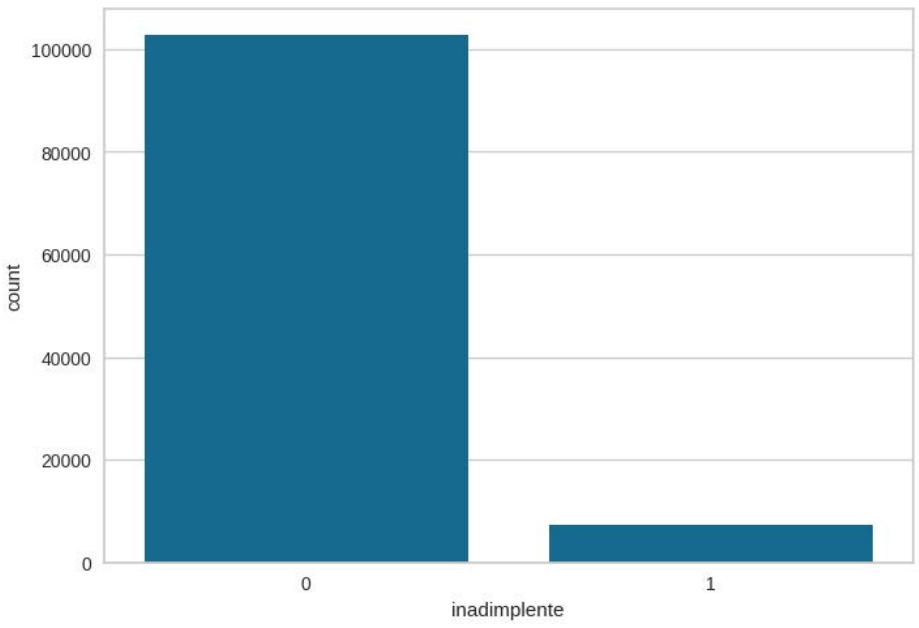
A coluna salario_mensal ainda possui valores nulos, indicando que o preenchimento por idade não foi totalmente eficaz.
As demais colunas, não possuem mais valores nulos, o que indica que o processo de preenchimento foi bem-sucedido.

em seguida é feito uma filtragem para a remoção dos valores nulos presentes em salario_mensal.

Quantidade dos dados inadimplentes:

inadimplente	
0	102666
1	7331

gráfico 5:



Logo após é feito uma descrição dos dados e obtido os seguintes resultados:

Desbalanceamento da Classe: A variável inadimplente apresenta um valor médio de 0.066, indicando que apenas cerca de 6,7% dos registros são de inadimplentes. Isso pode causar problemas de desbalanceamento na modelagem. Técnicas de balanceamento, como oversampling ou undersampling, podem ser necessárias.

Idade: A média de idade (52.25 anos) e a faixa (0 a 103 anos) sugerem uma distribuição possivelmente assimétrica. É possível que testemos alguma transformações, como logaritmo, tratar ou excluir os outliers.

Variáveis com Alto Desvio Padrão: Variáveis como salario_mensal e numero_linhas_crdto_aberto têm altos desvios padrão em relação à média. Isso pode indicar a presença de outliers que podem afetar a modelagem. É possível que testemos alguma transformações, como logaritmo, tratar ou excluir os outliers.

Dependentes: A média de dependentes é relativamente baixa (0.74), com um máximo de 20. Pode ser interessante analisar como a quantidade de dependentes influencia a inadimplência.

Subsequentemente é feito uma modelagem e um balanceamento dos dados.
Esse balanceamento é necessário pois como estamos lidando com inadimplência, nosso maior interesse de termos maior acerto é justamente aqueles que tem chance de inadimplência pois trazem prejuízo.

Decision Tree

Então é configurado o ambiente PyCaret com dados de treino e teste informados manualmente, é criado o
modelo 'dt': Decision Tree e em seguida é feito a predição do modelo.
Essa fase é essencial para verificar o desempenho do modelo em dados novos, ou seja, dados que ele não viu
durante o treinamento.
Através dessa predição, você pode avaliar o quão bem o modelo consegue fazer previsões com base nas novas
entradas e comparar os resultados com os valores reais (se existirem), analisando a precisão e a eficácia do
modelo.

Visualizando em forma de Gráfico

gráfico 6:

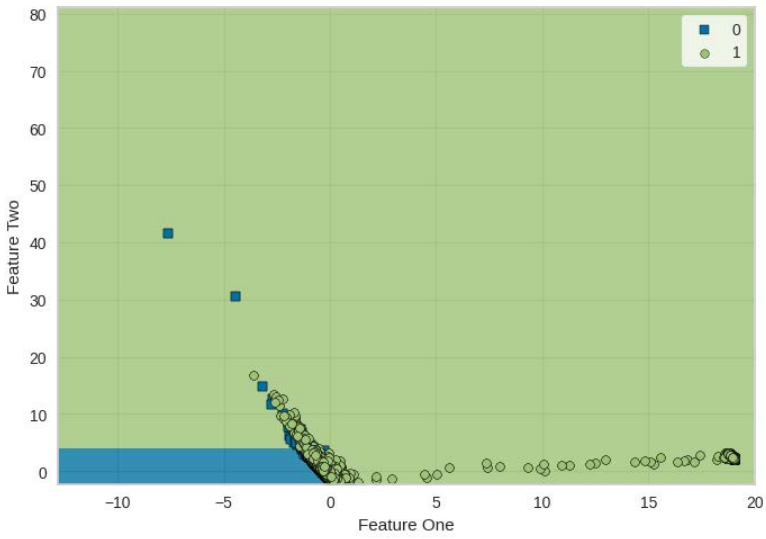
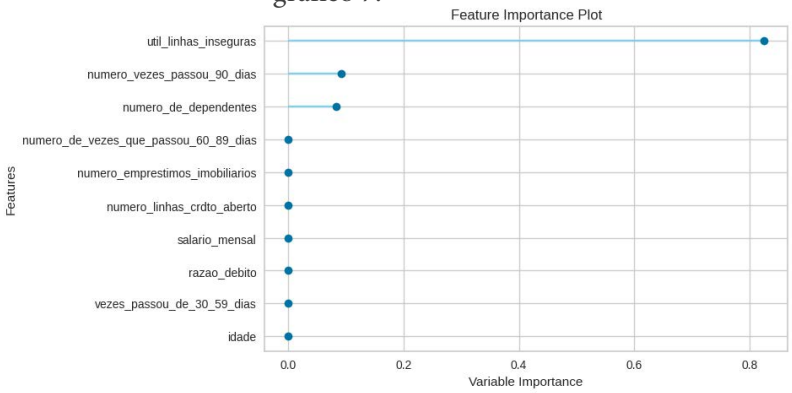


gráfico 7:



O gráfico acima observamos as variáveis mais importantes.

O gráfico ao lado é o grafico de aprendizado, útil para identificar se o modelo está overfitting (quando o modelo se ajusta muito bem aos dados de treinamento, mas falha em dados novos) ou underfitting (quando o modelo não captura a complexidade dos dados).

gráfico 8:

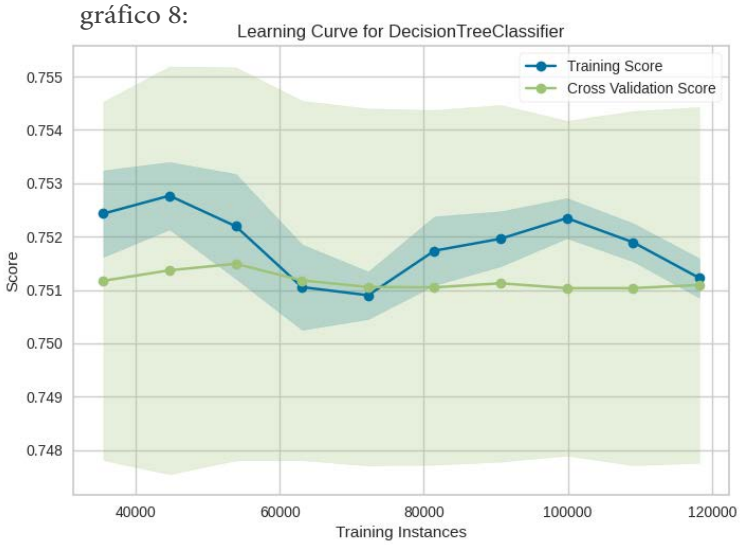
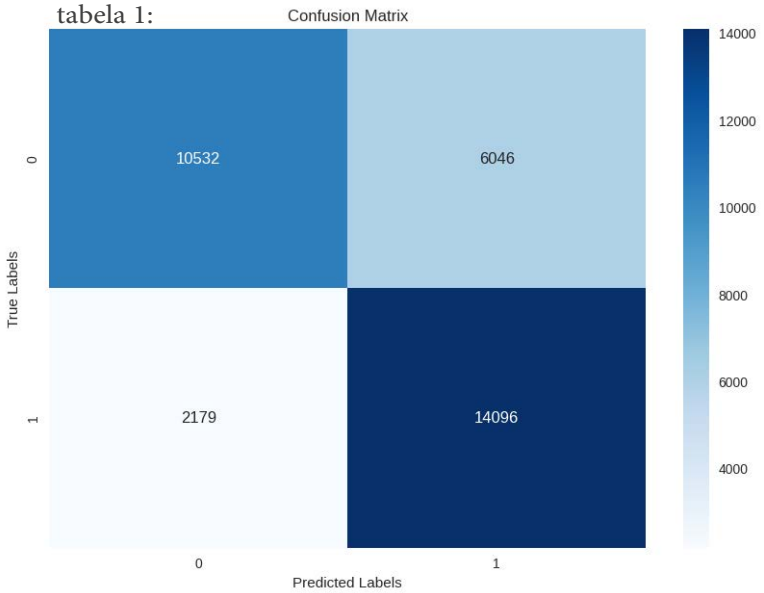


tabela 1:



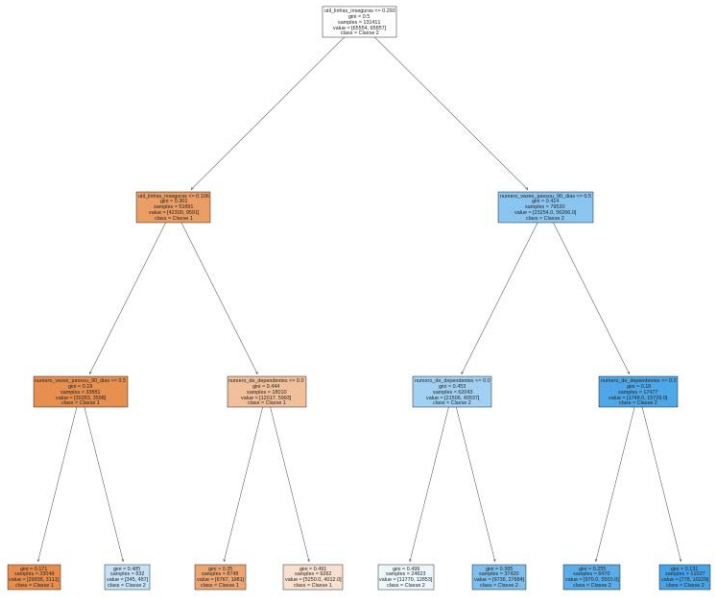
Accuracy: 0.7496423462088698
Kappa: 0.5003270713933081

É feito uma predição do modelo para descobrir alumas informações e em seguida obtemos o gráfico acima. O Qual avalia o desempenho do modelo de classificação usando a matriz de confusão.

Observamos que o modelo apresenta um desempenho razoável, especialmente em termos de identificar inadimplentes (classe 1). No entanto, a revocação da classe 0 indica que há uma necessidade de melhorar a capacidade do modelo de identificar não inadimplentes sem deixar muitos casos passarem.

Abaixo observa-se a árvore de decisão.

gráfico 9:



Random Forest

seguindo basicamente os mesmos passos da Árvore de Decisão, é configurado o ambiente para o treinamento e teste, é feito a predição do modelo e obtidos os seguintes gráficos.

gráfico 10:

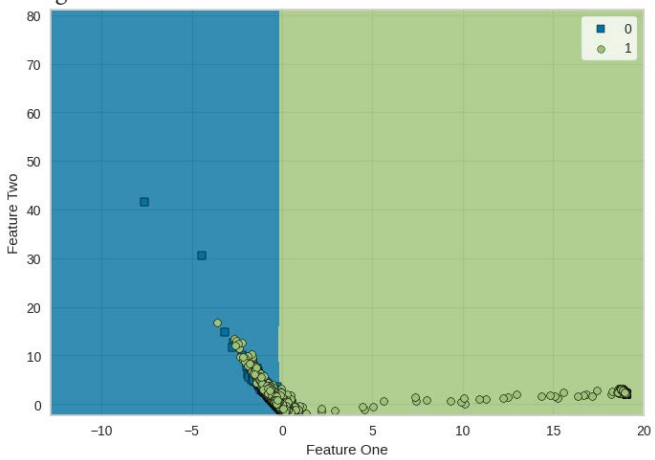


gráfico 11:

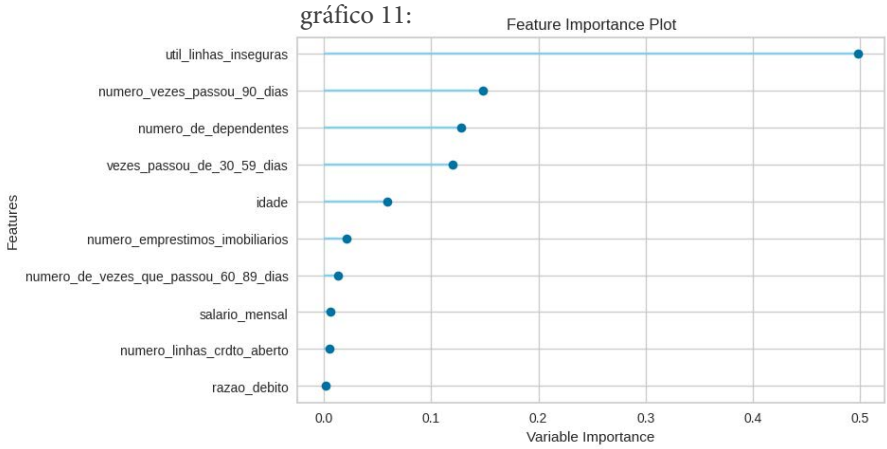
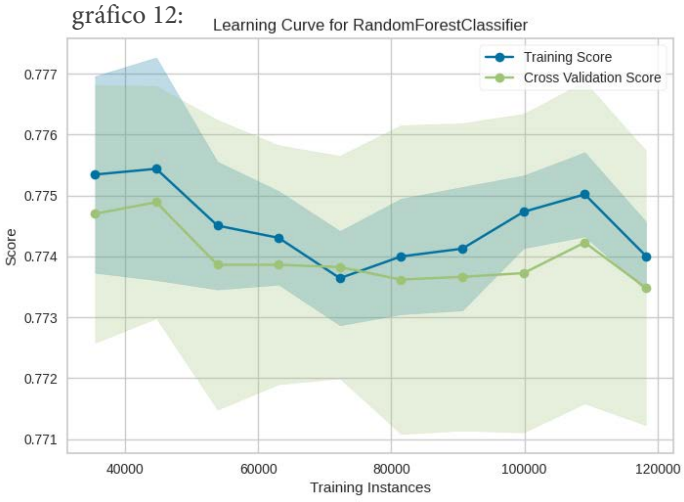
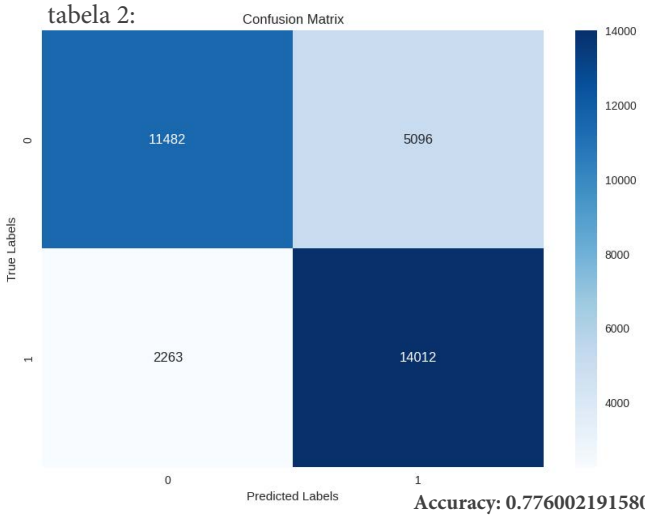


gráfico 12:



É feito a predição do modelo, obtido algumas informações como por exemplo, a acurácia e em seguida é gerado o gráfico.

tabela 2:



Accuracy: 0.7760021915806775
Kappa: 0.5526778570385158

CONCLUINDO

Observamos que a análise dos dados utilizando os modelos de Decision Tree e Random Forest revelou resultados semelhantes, com pequenas variações nas métricas de desempenho. O Random Forest apresentou uma leve vantagem em acurácia e concordância, indicando um desempenho um pouco superior na previsão da inadimplência. Embora ambos os modelos sejam eficazes, a robustez do Random Forest torna-o a escolha preferencial para essa tarefa. Assim, a decisão sobre qual modelo utilizar deve levar em conta o contexto específico e as necessidades do problema em questão, dado que ambos oferecem previsões confiáveis.

Links utéis:

https://www.kaggle.com
https://colab.research.google.com/drive/1VkpLFl_g66Gb9zvYarc69KN1fjGMZyhK?usp=sharing
https://github.com/Diegowil/INT_CIENCIA_DADOS/blob/main/EE3.ipynb
https://github.com/alexanderbandeiralira/EE3-Introducao_a_ciencia_de_dados/blob/main/EE3.ipynb

ALEXANDER BANDEIRA LIRA & DIEGO WILSON PEREIRA DE ANDRADE