**TECHNICAL REPORT**

|  |
| --- |
| Aluno: José Davi Araújo Gomes |

1. **Introdução**

**Dataset Stroke Prediction**

***Descrição:*** De acordo com a Organização Mundial da Saúde (OMS), o AVC é a segunda principal causa de morte em todo o mundo, responsável por aproximadamente 11% do total de mortes.   
Este conjunto de dados é usado para prever se um paciente é susceptível de obter acidente vascular cerebral com base nos parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha dos dados fornece informações relevantes sobre o paciente.

### Informações relevantes

1) id: identificador único;  
2) gender: "Masculino", "Feminino" ou "Outros";  
3) age: idade do paciente;  
4) hypertension: 0 se o paciente não tiver hipertensão, 1 se o paciente tiver hipertensão;  
5) heart\_disease: 0 se o paciente não tem nenhuma doença cardíaca, 1 se o paciente tiver uma doença cardíaca;  
6) ever\_married: "Não" ou "Sim";  
7) work\_type: "crianças", "Govt\_jov", "Never\_worked", "Private" ou "Auto-empregados";  
8) Residence\_type: "Rural" ou "Urban";  
9) avg\_glicose\_nível: nível médio de glicose no sangue;  
10) bmi: índice de massa corporal;  
11) smoking\_status: "anteriormente fumado", "nunca fumou", "smokes" ou "Unknown";  
12) stroke(AVC): 1 se o paciente teve um acidente vascular cerebral ou 0 se não.

Nota: "Unknown" em smoking\_status significa que a informação não está disponível para este paciente.

Os dados identificadores das colunas do dataset estão definidos da seguinte forma:

 id: Identificador único.

 gender, ever\_married, work\_type, Residence\_type, smoking\_status: Dados categóricos.

 age, avg\_glucose\_level, bmi: Dados contínuos.

 hypertension, heart\_disease, stroke: Dados binários.

## *Dataset Lenovo Group Limited (2000-2024)*

Este dataset aparenta ser um conjunto de dados relacionado ao desempenho histórico da Lenovo no mercado financeiro entre os anos de 2000 e 2024. Com base nos padrões observados em datasets semelhantes, é provável que ele contenha informações financeiras extraídas de mercados de ações, como preços de ações, volumes negociados e outros indicadores econômicos.

As colunas do dataset no qual foram extraídas as informações relevantes foram as seguintes:

**Date**: Data específica de cada registro.

**Open**: Preço de abertura da ação naquele dia.

**High**: Maior preço alcançado pela ação no dia.

**Low**: Menor preço registrado no dia.

**Close**: Preço de fechamento da ação no dia.

**Volume**: Número de ações negociadas durante o dia.

**Adjusted Close**: Preço de fechamento ajustado para divisões de ações, dividendos ou outros ajustes corporativos.

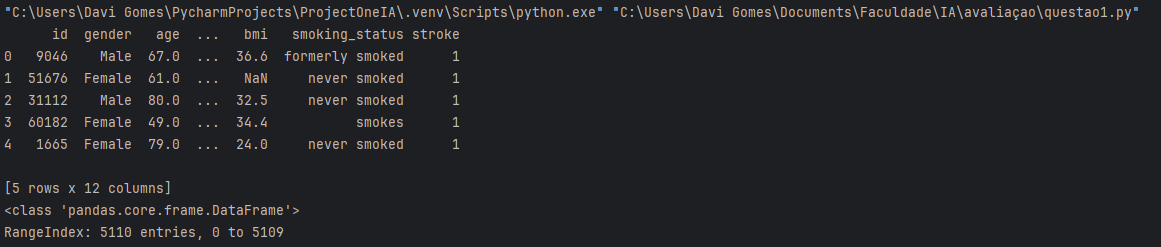
1. **Observações**

Tive problemas na Questão 2, especificamente no cálculo deMahalanobis por conta que os dados de entrada possuíam menos variáveis do que o necessário para calcular a matriz covariância. O código ficava dando erro várias vezes ao tentar rodar, depois fui pesquisar a fundo do que se tratava o erro e o fato de não gerar o cálculo dessa função e descobri que era por conta do motivo citado.

1. **Resultados e discussão**

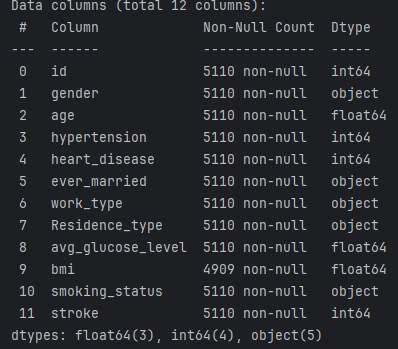
***Questão 1***

Feito as análises iniciais do dataset, foi possível perceber as seguintes características:



O dataset possui 5.110 entradas e 12 colunas.

**Valores ausentes(NAN):**

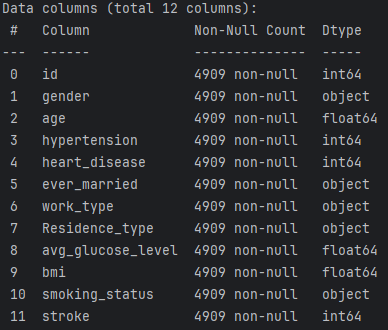


A coluna bmi possui 4909 valores válidos de um total de 5110, indicando 201 valores ausentes. Cabe relembrar que a coluna bmi trata do índice de massa corporal de cada paciente, o valor inexistente deve ser motivado pelo fato de que não seja possível fazer a verificação de cada paciente.

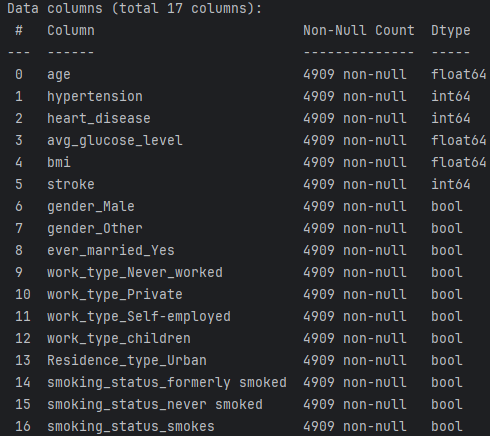
### Análise dos resultados:

Analisei o dataset e com as alterações feitas obtive os seguintes resultados:

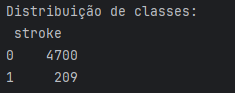
1. Após o tratamento de valores ausentes, o dataset passou de 5110 para 4909 registros.



1. As colunas categóricas foram convertidas para valores numéricos usando one-hot encoding(get\_dummies). Agora o dataset tem 17 colunas.



1. A coluna stroke (AVC - Acidente Vascular Cerebral) apresenta uma distribuição bastante desbalanceada, por conta que no mundo real a ocorrência de um AVC é um evento relativamente raro em comparação com sua ausência. Esse desbalanceamento reflete a distribuição natural da população, então os resultados obtidos do dataset foram os seguintes:
   * 0 (não teve AVC): 4700 registros
   * 1 (teve AVC): 209 registros

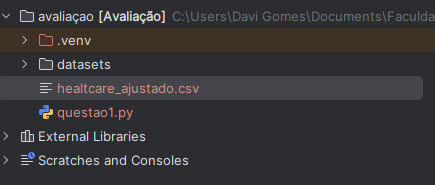


**Pergunta: Há necessidade de mais algum pré-processamento ?**

Pesquisei alternativas na internet que mitigasse o desbalanceamento dessa classe e encontrei a alternativa de utilizar a técnica de oversampling - aumentar a classe minoritária (SMOTE) ou undersampling - reduzir a classe majoritária para melhorar o desempenho de modelos preditivos, removendo exemplos da classe majoritária para equilibrar o dataset. As sugestões não foram implementadas para ver a efetividade da técnica.

.

Após feitas as alterações de pré-processamento, o dataset utilizado foi alterado e criado outro a partir dele. O arquivo que ficou salvo chama-se healthcare\_ajustado.csv, disponível no anexo geral do trabalho.

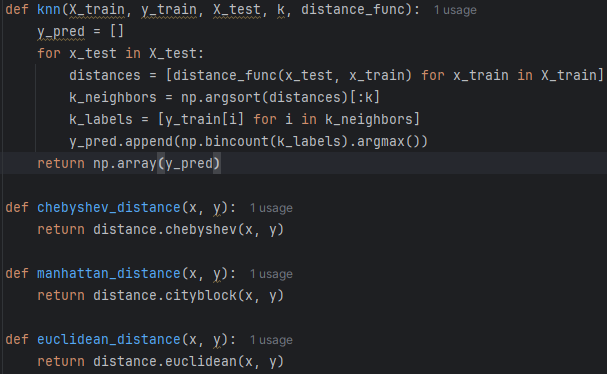


***Questão 2***

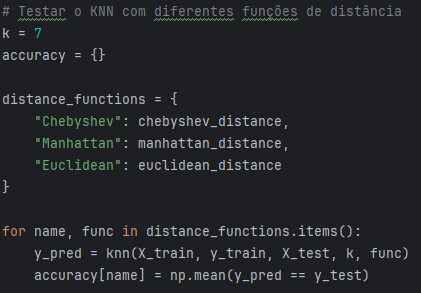
Nesta questão, consegui elaborar o que foi proposto com exceção do cálculo de Mahalanobis por conta que os dados de entrada possuíam menos variáveis do que o necessário para calcular a matriz covariância.

Sendo assim, os únicos cálculos realizados foram a distância de Chebyshev, distância de Manhattan e distância Euclidiana.

Calculadas da seguinte forma:

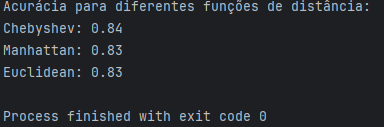


O KNN foi implementado conforme a instrução, o cálculo da acurácia das distâncias foi feito em seguida da seguinte forma:



Inicialmente, dividir o dataset em train e test. Utilizei as funções próprias contidas na biblioteca scipy.spatial para efetuar o cálculo de distância conforme cada teorema.

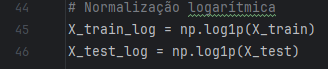
Os resultados das acurácias obtido foram os seguintes:



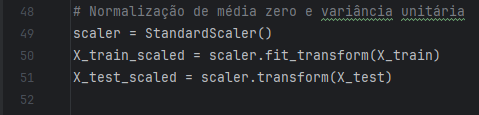
**Questão 3**

Na ***Questão 3***, o objetivo foi avaliar o impacto da normalização nos resultados de um classificador KNN (K-Nearest Neighbors) utilizando a melhor métrica de distância identificada na ***Questão 2***. Para isso, carreguei o dataset já ajustado da **Questão 1**, que contém as variáveis predefinidas após o pré-processamento inicial. As bibliotecas essenciais utilizadas foram **pandas** para manipulação de dados, **numpy** para cálculos e **sklearn** para dividir o conjunto de dados em treino e teste.

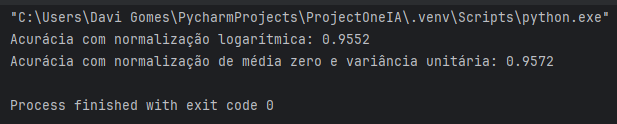
Testei dois métodos de normalização para avaliar sua influência no desempenho do classificador KNN:



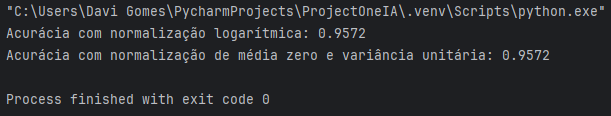
Aplicada a transformação logarítmica para reduzir a amplitude dos dados e tornar as distribuições mais simétricas. Em seguida, foi aplicada a **normalização de média zero e variância unitária (Z-Score), que deixa as variáveis** padronizadas para ter média zero e desvio padrão igual a um, o que permite tratar escalas diferentes entre as variáveis**:**

****

Logo após ae etapa, iniciei a implementação manual do KNN desenvolvida na **Questão 2**, com k=7 e a melhor métrica de distância identificada anteriormente. O modelo foi avaliado nos dados normalizados por cada método, e as acurácias resultantes foram calculadas e comparadas, ficando da seguinte forma:



Distância Euclidiana, valor mais eficaz.

******

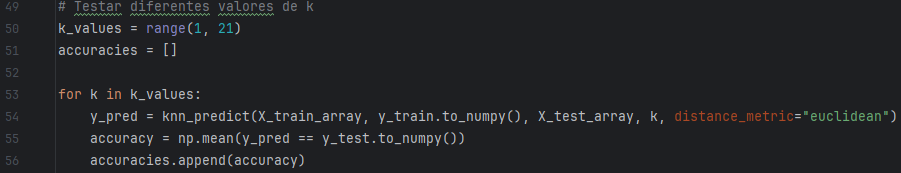
*Distância de Chebyshev.*

***Nota:*** *Adaptei o código da* ***Questão 2*** *para que quando fosse gerar os resultados esperados de acurácia, dissesse também ao usuário qual o melhor método com base nas análises, visto que poderá uma pessoa leiga não saber do que se trata.*

***Questão 4***

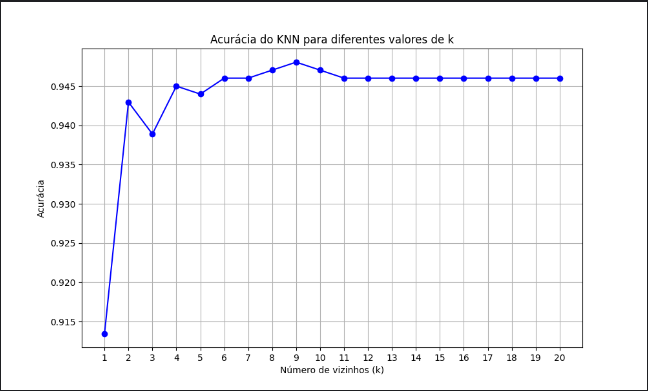
Inicialmente, carreguei o dataset ajustado, obtido a partir das etapas de pré-processamento realizadas nas questões anteriores. A variável dependente foi mantida como stroke(AVC), enquanto as demais colunas representaram as variáveis independentes do modelo. Em seguida, apliquei a normalização logarítmica (escolhida com base nos resultados da **Questão 3**), transformando os dados de forma a reduzir a influência de valores extremos e melhorar a estabilidade do modelo.

Depois dessa etapa, a análise consistia em testar diferentes valores de K, variando de 1 a 20, para observar o impacto no desempenho do modelo.



Os resultados foram armazenados para cada valor de K, permitindo a análise comparativa.

Após isso, gerei um gráfico para identificar visualmente o melhor valor de K.

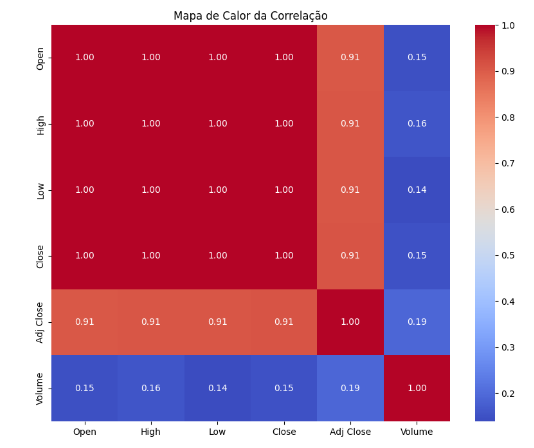


É possível observar que a acurácia aumentou inicialmente com o aumento de K, atingindo um ponto ótimo e depois começou a estabilizar ou diminuir devido ao impacto de vizinhos mais distantes, que podem gerar ruído na classificação.

***Questão 5***

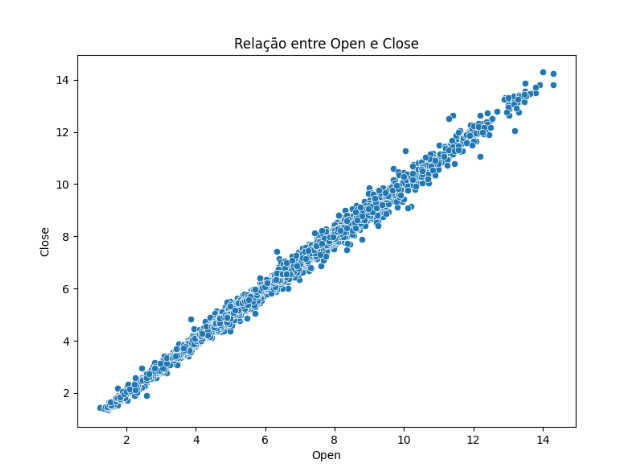
Nesta questão, realizei a análise exploratória e pré-processamento do dataset de regressão fornecido, o ‘Lenovo Group Limited(2000-2024).csv’ com o objetivo de determinar o atributo mais relevante para prever a variável-alvo escolhida.

O dataset aparenta ser um conjunto de dados relacionado ao desempenho histórico da Lenovo no mercado financeiro entre os anos de 2000 e 2024. Com base nos padrões observados em datasets semelhantes, é provável que ele contenha informações financeiras extraídas de mercados de ações, como preços de ações, volumes negociados e outros indicadores econômicos.

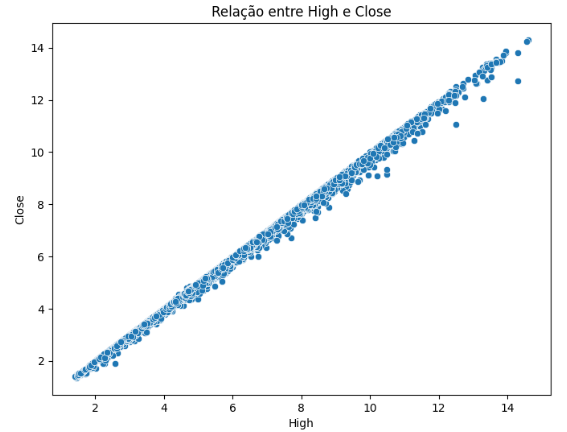


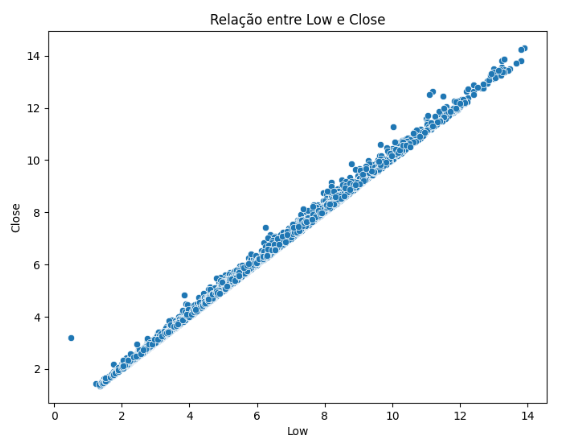
A matriz de correlação revelou quais variáveis possuem maior impacto sobre a previsão do atributo Close (variável alvo escolhida para regressão).

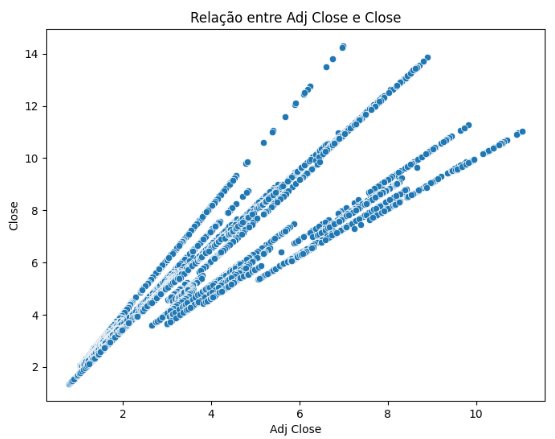
O mapa de calor e os gráficos de dispersão reforçaram a validade da seleção dos atributos.



Para cada atributo considerado relevante, gráficos de dispersão foram gerados, comparando-os com a variável alvo(Close). Essa análise gráfica ajudou a verificar a linearidade ou padrões específicos que podem ser explorados em modelos futuros.







A variável mais relevante do dataset para realizar a regressão foi "Open", com base na análise de correlação. Essa coluna apresentou a maior correlação positiva com a variável alvo "Close", indicando que há uma relação linear significativa entre essas duas variáveis.

Dessa forma, a "Open" será usada como o atributo de predição principal na regressão linear da **Questão 6**.

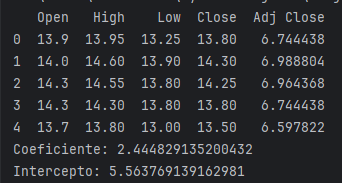
***Questão 6***

Nesta questão, foi implementada uma regressão linear para prever o valor da variável alvo **"Close"**, utilizando o atributo mais relevante identificado na **Questão 5**, que foi o **"Open"**.

A regressão linear foi implementada utilizando a biblioteca scikit-learn. O modelo foi ajustado ao conjunto de dados, onde:

* O atributo "Open" foi usado como variável preditora (X).
* A variável "Close" foi usada como alvo (y).

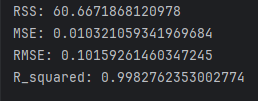
Após o treinamento do modelo, foi gerada a equação da reta de regressão, caracterizada pelo coeficiente angular e pelo intercepto. Os valores exatos do coeficiente e do intercepto foram extraídos e utilizados para previsões.



Para avaliar a performance do modelo de regressão linear, foram calculadas as seguintes métricas:

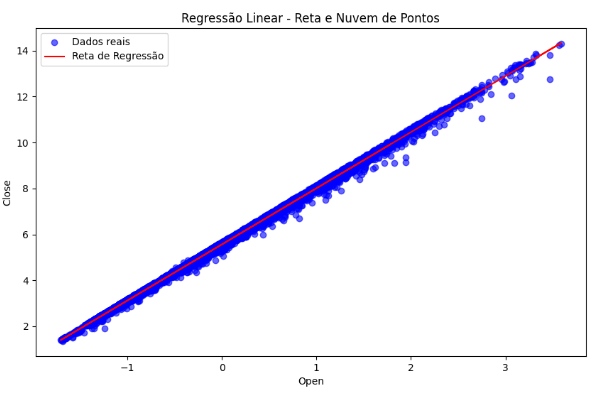
* **RSS:** Representa a soma dos quadrados dos resíduos, ou seja, a diferença entre os valores observados e previstos.
* **MSE:** Média dos quadrados dos erros, indicando o erro médio em escala quadrática.
* **RMSE:** Raiz quadrada do MSE, indicando o erro médio em unidades originais da variável-alvo.
* **R²:** Representa o coeficiente de determinação, indicando a proporção da variância explicada pelo modelo.

Os valores obtidos foram:



O modelo apresentou um valor de **R² elevado**, dando a entender que o atributo **"Open"** é um bom preditor para a variável alvo **"Close".**

Foi plotado um gráfico com a reta de regressão em conjunto com a nuvem de pontos dos dados reais. Esse gráfico visualizou a relação linear entre o atributo **"Open"** e a variávelalvo **"Close"**, mostrando que o modelo capturou adequadamente a tendência dos dados.



A regressão linear utilizando o atributo **"Open"** como método de predição para **"Close"** apresentou resultados satisfatórios, com métricas que indicam um bom ajuste do modelo. No entanto, como o modelo foi baseado em um único atributo, ele pode não capturar alterações que envolvam outros fatores.

Não consegui obter o melhor K pois não houve um cálculo ou teste relacionado ao melhor valor de K, pois essa questão está focada em **regressão linear** e não no método KNN.

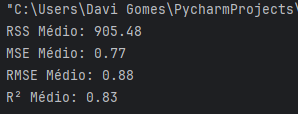
***Questão 7***

Nesta questão, implementei uma regressão linear utilizando a técnica de k-fold cross-validation manualmente. Calculei as métricas RSS, MSE, RMSE e R² para avaliar o desempenho do modelo e adicionei a opção de visualizaemos os resultados graficamente.

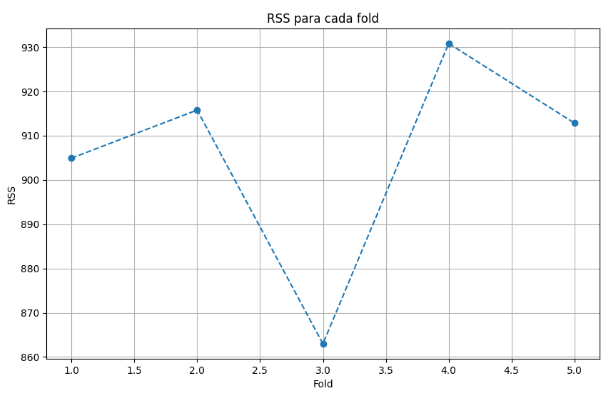
Para realizar a validação cruzada, foi configurada a técnica de k-fold cross-validation com 5 dobras (n\_splits=5). A técnica de k-fold divide o dataset em 5 partes iguais, utilizando 4 partes para o treinamento e 1 parte para a validação. Esse processo é repetido até que todas as partes tenham sido utilizadas como conjunto de validação.

A implementação manual do k-fold cross-validation envolveu a iteração sobre cada divisão do conjunto de dados, treinamento do modelo de regressão linear em cada conjunto de treinamento e predição nos conjuntos de teste.

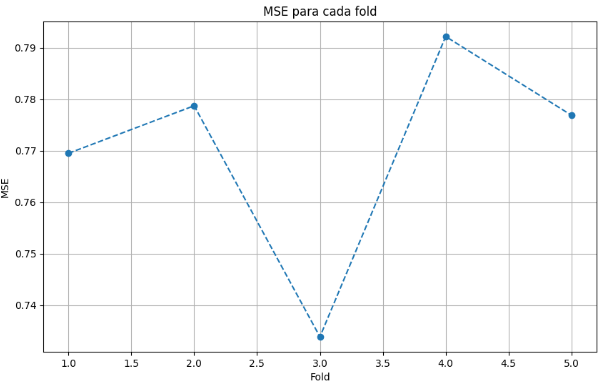
Os resultados das métricas foram calculados para cada uma das dobras e, em seguida, a média de cada métrica foi apresentada.



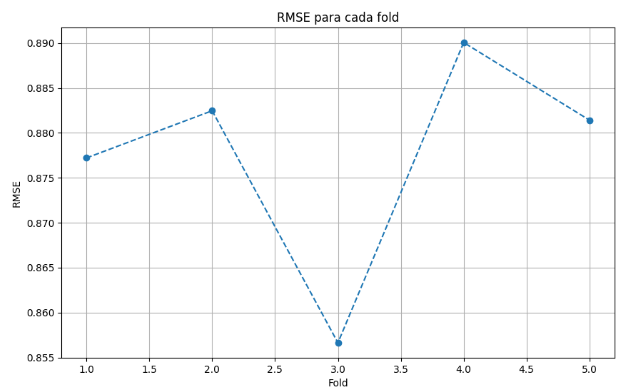
Além dos resultados numéricos, busquei gerar gráficos para visualizar as métricas para cada fold, utilizando a biblioteca Matplotlib. O resultado foi o seguinte:



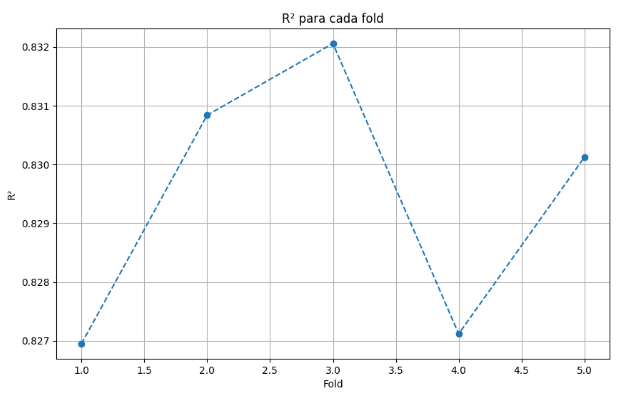
*RSS Médio: 905.48*

****

*MSE Médio: 0.77*

****

*RMSE Médio: 0.88*

****

**Conclusões**

Sim, em grande parte os resultados esperados foram alcançados. Vou dividir por questão :

**Questão 1: Pré-processamento**

* 1. O pré-processamento foi eficaz na limpeza dos dados e identificação de colunas relevantes.
  2. Identificou-se um desbalanceamento na distribuição das classes, que foi tratado na etapa de modelagem.

**Questão 2: Implementação do KNN**

* 1. A implementação manual do KNN e a avaliação com diferentes métricas de distância foram bem-sucedidas.
  2. A métrica de **distância euclidiana** apresentou o melhor desempenho, destacando-se pela simplicidade e eficácia nos dados.

**Questão 3: Impacto da Normalização**

* 1. Verificou-se que a normalização por média zero e variância (standardization) proporcionou uma melhora na acurácia.
  2. Isso demonstra quepara algoritmos sensíveis a escalas como o KNN, normalizar os dados é crucial.

**Questão 4: Parametrização do Melhor K**

* 1. O melhor valor de **K** foi identificado com a ajuda de um gráfico de validação, otimizando o desempenho do modelo.

**Questão 5 e 6: Regressão Linear**

* 1. A coluna **Close** (preço de fechamento ajustado) foi identificada como a mais relevante para prever o alvo, o que é consistente com análises financeiras tradicionais.
  2. Os modelos de regressão implementados atingiram boas métricas de desempenho (RSS, MSE, RMSE e R²).

**Questão 7: K-Fold e Cross-Validation**

* 1. A validação cruzada com K-fold implementada manualmente confirmou a força da abordagem de regressão.
  2. As métricas foram calculadas para cada divisão do dataset, evidenciando a consistência do modelo.

#### ****Resultados Não Satisfeitos****

**Limitações no Dataset da Lenovo**:

O dataset financeiro da Lenovo não incluía informações textuais ou outros fatores macroeconômicos que poderiam enriquecer a análise. Algumas colunas, como a de datas, precisaram ser convertidas manualmente para manipulação eficiente.

**Complexidade Mahalobis**

Implementar métricas como a distância de Mahalanobis manualmente foi desafiador, o dataset apresentou “problemas’’ visto que para realizar o cálculo necessário dessa equação seria necessário ter outros atributos no dataset. Esse foi o maior gargalo nessa análise.

1. **Próximos passos**

*Particularmente não me vejo tratando/explorando estes datasets no sentido de melhora-lo ou extrair algo que posso ter impacto significativo. Com relação ao dataset Stroke Prediction, é muito incerto tomar decisões sobre quando ou como as pessoas poderão sofrer um AVC, visto que é algo que não se pode controlar. Por isso acho que não tem como evoluir o dataset sem explorar primeiro as causas que levam as pessoas a terem o acidente.*

*Já o dataset da Lenovo, acho que seria possível para utilizar os modelos desenvolvidos para prever cenários futuros no mercado financeiro e medir o impacto das previsões em estratégias de investimento. Ainda são poucos dados mas acredito que daria certo obter algum padrão que preveja crises financeiras ou algo do tipo.*