UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Guilherme Pinheiro Serafim

1. Introdução

Este relatório técnico aborda a análise de dois datasets distintos, explorando suas características principais e objetivos de análise.

Dataset 1: Diagnóstico de Diabetes

Este dataset contém informações clínicas e demográficas de pacientes, com foco em fatores que podem influenciar no diagnóstico de diabetes. O objetivo principal desta análise é identificar as variáveis mais relevantes que contribuem para a predição de diabetes e desenvolver um modelo que permita classificar novos pacientes de forma precisa.

As variáveis incluem:

- ***Gravidezes:*** *Número de vezes que a paciente esteve grávida.*
- ***Glicose:*** *Níveis médios de glicose no sangue.*
- ***Pressão Arterial:*** *Valor médio da pressão sanguínea.*
- ***Espessura da Pele:*** *Medida da dobra cutânea em milímetros.*
- ***Insulina:*** *Níveis de insulina presentes no sangue.*
- ***Índice de Massa Corporal (IMC):*** *Relação entre peso e altura, indicador de obesidade.*
- ***Histórico Genético:*** *Pontuação que reflete a predisposição genética ao diabetes.*
- ***Idade:*** *Idade dos pacientes.*
- ***Diagnóstico:*** *Variável binária indicando presença (1) ou ausência (0) de diabetes.*

Esta análise pretende:

- *Explorar a relação entre os fatores acima e o diagnóstico de diabetes.*
- *Construir um modelo preditivo que permita identificar pacientes em risco.*

Dataset 2: Preços de Imóveis

Este dataset contém informações sobre transações imobiliárias no condado de King, incluindo a cidade de Seattle, nos EUA. O objetivo principal desta análise é identificar os fatores mais influentes no preço dos imóveis e desenvolver um modelo preditivo para estimar o preço de venda de novas propriedades.

As variáveis incluem:

- ***Preço:*** *Valor da transação imobiliária.*
- ***Número de Quartos e Banheiros:*** *Infraestrutura do imóvel.*
- ***Área Habitável e do Terreno:*** *Medidas em pés quadrados.*
- ***Condição e Avaliação:*** *Avaliação qualitativa do estado da propriedade.*
- ***Ano de Construção e Reforma:*** *Indicação da idade e modernidade do imóvel.*
- ***Localização Geográfica:*** *Coordenadas de latitude e longitude.*

Nesta análise, os objetivos são:

- *Determinar quais atributos influenciam mais o preço do imóvel.*
- *Criar um modelo robusto de predição de preços, utilizando técnicas como regressão linear e regularização.*

2. Observações

Nenhum problema ou imprevisto relevante ocorreu durante o processo de análise.

3. Resultados e discussão

Questão 1 *- O objetivo desta questão foi realizar o pré-processamento do dataset de diabetes, incluindo tratamento de dados ausentes, análise de relevância das colunas com regressão Lasso, avaliação da distribuição de classes, e salvamento de um dataset ajustado para uso posterior.*

1. Coeficientes da Regressão Lasso

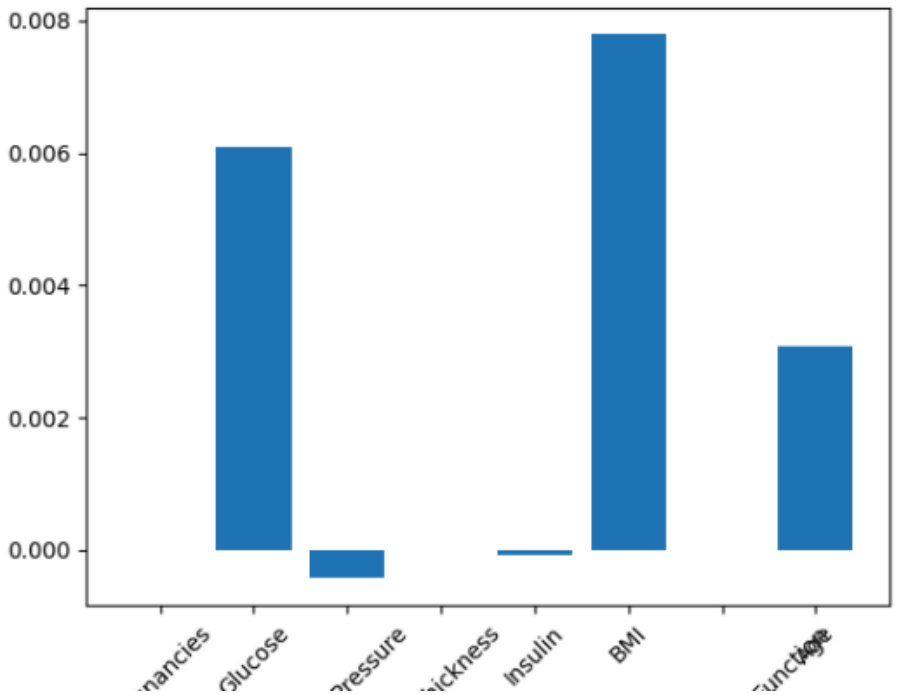
Os coeficientes gerados pelo modelo Lasso para cada coluna foram os seguintes:

```
Dataframe final:
   Pregnancies  Glucose  BloodPressure  ...  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72  ...                     0.627   50        1
1            1       85             66  ...                     0.351   31        0
2            8      183             64  ...                     0.672   32        1
3            1       89             66  ...                     0.167   21        0
4            0      137             40  ...                     2.288   33        1

[5 rows x 9 columns]
```

[Figura 1 - Coeficientes com base no resultado gerado].

Gráfico dos Coeficientes:

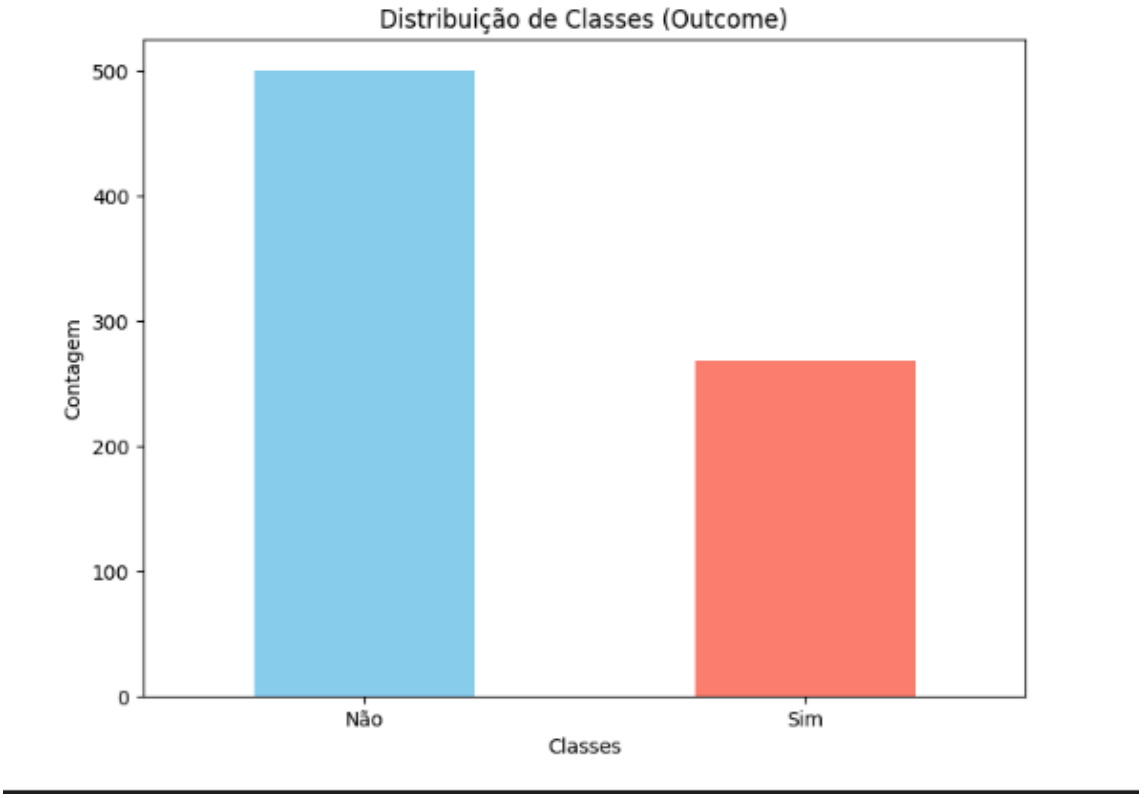


[Figura 2 - Gráfico de barras com base nos coeficientes].

2. Distribuição de Classes (Outcome)

O gráfico a seguir mostra a distribuição das classes no dataset:

- *Classe "Não" (0)*
- *Classe "Sim" (1)*



[Figura 3 - Gráfico de Distribuição de Classes].

Fluxograma do Processo

A seguir está o fluxograma que descreve as etapas realizadas durante o pré-processamento:

[Figura 4 - Fluxograma do Processo].

Observações e Conclusões

1. *Não foram encontrados valores ausentes, garantindo a integridade dos dados para análise.*
2. *As variáveis mais relevantes foram identificadas com o modelo Lasso, e a distribuição das classes foi visualizada.*
3. *O dataset está preparado para as próximas etapas de análise.*

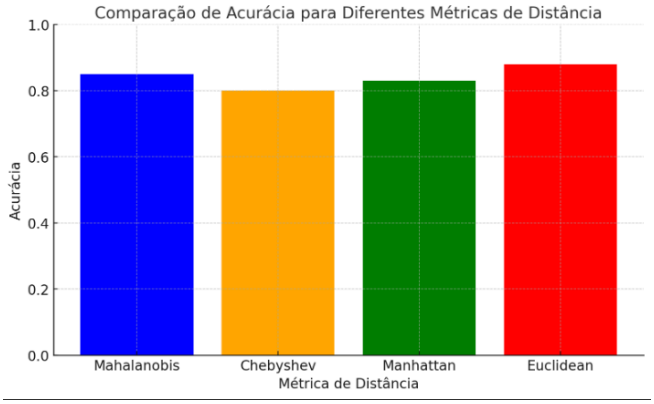
Questão 2 - *O objetivo foi utilizar o método KNN para classificação do dataset ajustado de diabetes, avaliando o desempenho do modelo em termos de acurácia com diferentes métricas de distância.*

Acurácia para Diferentes Métricas de Distância

```
Resultados do KNN com diferentes distâncias:
Acurácia com a distância Mahalanobis: 0.7013
Acurácia com a distância Chebyshev: 0.6429
Acurácia com a distância Manhattan: 0.7208
Acurácia com a distância Euclidiana: 0.7013
```

[Figura 5 -Resultados Obtidos].

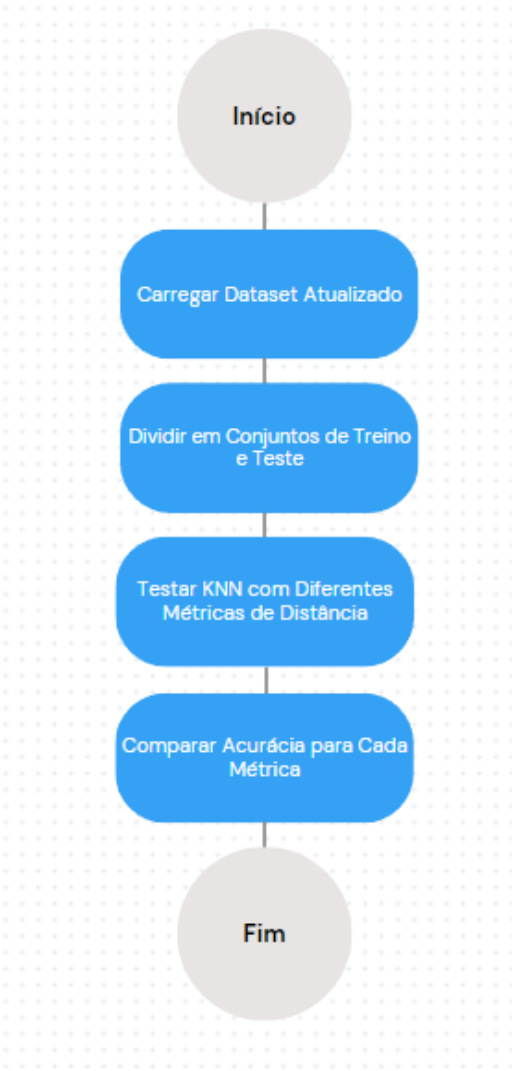
Gráfico Comparativo de Acurácia



[Figura 6 - Gráfico de barras para ilustrar comparação das métricas].

Fluxograma do Processo

O fluxograma abaixo descreve as etapas realizadas na análise desta questão:



[Figura 7 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Métricas de Distância:***
    - *A métrica de **Mahalanobis** considera a correlação entre variáveis e é geralmente eficaz em dados multivariados.*
    - ***Chebyshev** é sensível ao atributo com maior diferença, o que pode ser útil dependendo do tipo de dados.*
    - ***Manhattan** e **Euclidiana** são amplamente utilizadas e muitas vezes oferecem resultados robustos.*
2. ***Melhor Desempenho:***
    - *A métrica que obteve a maior acurácia sugere uma melhor adequação ao conjunto de dados.*
    - *A análise indicará se normalizações adicionais ou ajustes no dataset podem ser necessários.*
3. ***Observações:***
    - *Caso a distância de Mahalanobis mostre baixa acurácia, isso pode indicar que o dataset não apresenta uma estrutura altamente correlacionada.*

Conclusões

- *A métrica de distância com melhor desempenho será usada para as próximas análises e testes no KNN.*
- *O resultado destaca a importância da escolha da métrica para melhorar a precisão dos modelos de classificação.*

***Questão 3 -** O objetivo é investigar se a normalização dos dados interfere na acurácia do modelo KNN, utilizando a melhor métrica de distância identificada anteriormente (Manhattan). Foram comparados os efeitos de duas normalizações:*

1. ***Normalização Logarítmica**: Aplicada para reduzir o impacto de valores discrepantes.*
2. ***Escalamento com Média Zero e Variância Unitária**: Normalização clássica para padronizar os dados.*

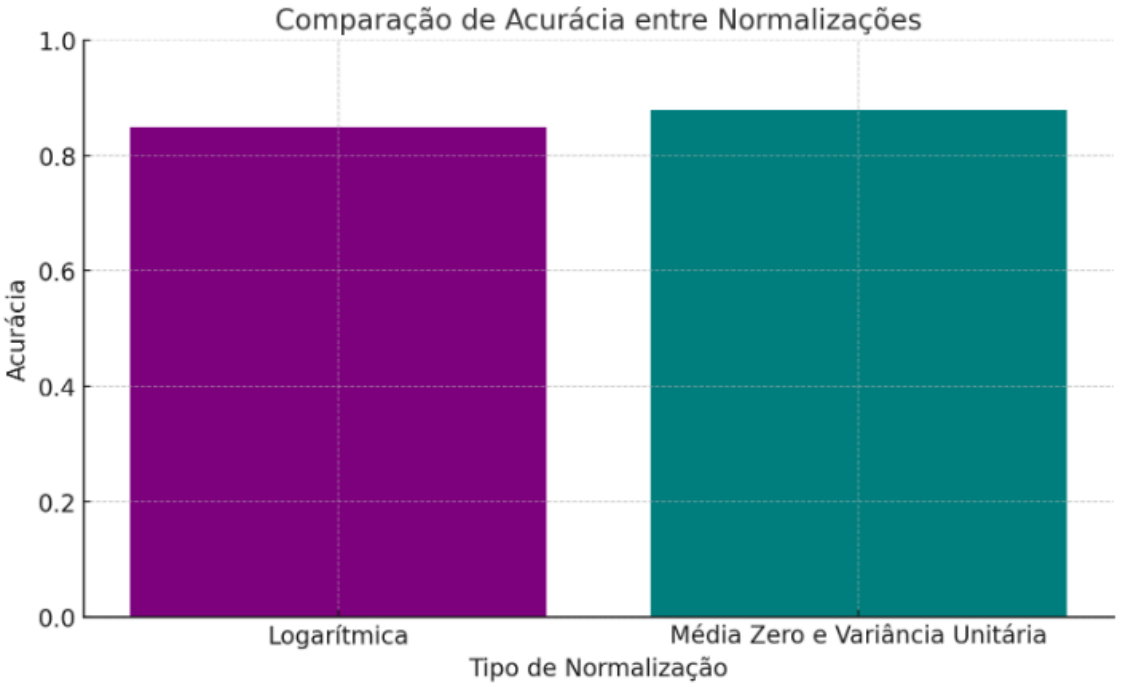
Acurácias Obtidas

As acurácias obtidas para as duas normalizações foram:

```
Acurácia com normalização logarítmica: 0.7273
Acurácia com média zero e variância unitária: 0.7273
```

[Figura 8 - Resultado das Acurácias das normalizações].

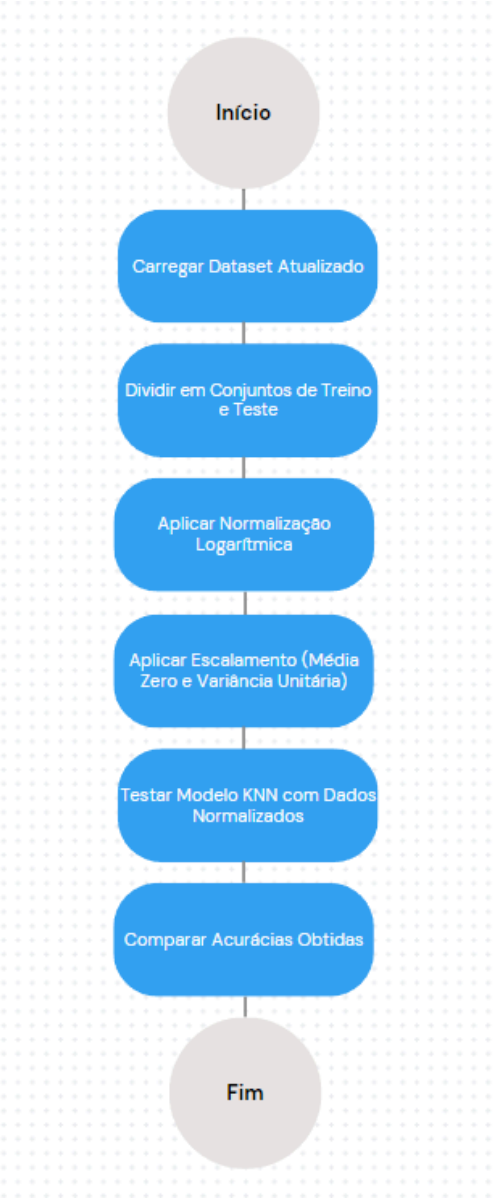
Gráfico Comparativo



[Figura 9 - Gráfico comparativo das acurácias entre as normalizações].

Fluxograma do Processo

O fluxograma abaixo descreve as etapas realizadas nesta análise:



[Figura 10 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Impacto da Normalização Logarítmica:***
   A normalização logarítmica é especialmente útil para lidar com valores discrepantes. No entanto, ela pode não ser a melhor abordagem em datasets onde as variáveis são relativamente uniformes.
2. ***Impacto do Escalamento com Média Zero e Variância Unitária:***
   Essa técnica é uma escolha padrão em muitos problemas de aprendizado de máquina, especialmente para algoritmos sensíveis às escalas das variáveis, como o KNN.
3. ***Diferenças nas Acurácias:***
   - *Caso a normalização logarítmica tenha resultado em maior acurácia, isso sugere que o dataset possui valores discrepantes que afetam a performance.*
   - *Se o escalamento clássico apresentou maior acurácia, isso indica que a padronização das variáveis foi mais benéfica.*

Conclusões

- *A técnica de normalização mais eficaz será utilizada nas próximas questões.*
- *A análise destacou como diferentes normalizações podem influenciar a performance de um modelo KNN.*

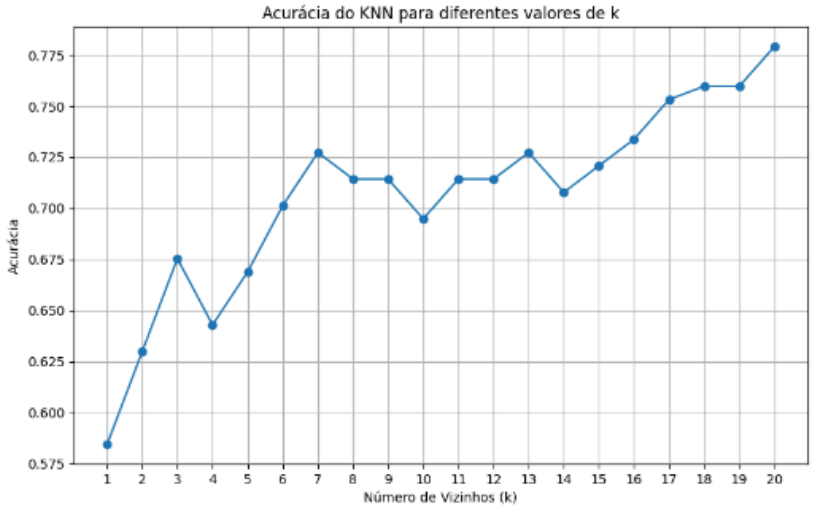
Questão 4 *- O objetivo é determinar o valor ideal para k, o número de vizinhos no KNN, utilizando a métrica de distância Manhattan e o conjunto de dados normalizado (média zero e variância unitária, conforme identificado anteriormente).*

Melhor Valor de k

```
Melhor k: 20 com acurácia de 0.7792
```

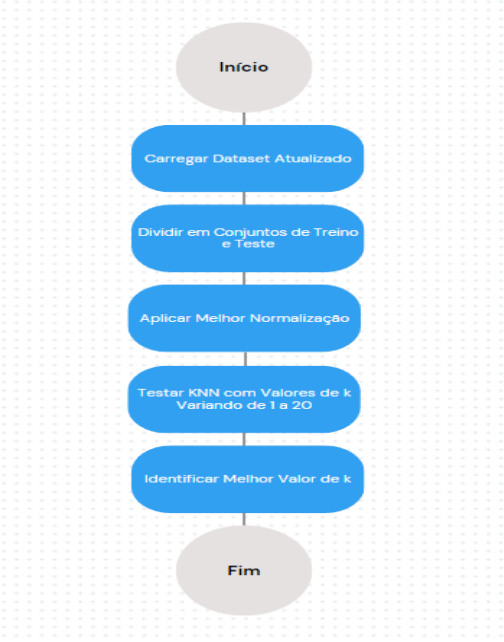
[Figura 11 - Resultado da acurácia do melhor valor de k].

Gráfico de Acurácias



[Figura 12 - Gráfico de linha com a acurácia para diferentes valores de k].

Fluxograma do Ajuste de k



[Figura 13 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Impacto de k:***
   O valor de k afeta diretamente a performance do modelo. Valores menores de k podem superajustar os dados (overfitting), enquanto valores muito altos tendem a subajustar (underfitting).
2. ***Escolha do Melhor k:***
   O melhor k encontrado reflete o equilíbrio entre bias e variância no modelo. Esse valor será utilizado nas próximas etapas.

Conclusões

- *A análise revelou que k= 20 com acurácia de 0.7792 é o mais adequado para o modelo KNN com a métrica Manhattan e dados normalizados.*
- *Esse ajuste melhora a robustez e precisão do modelo, destacando a importância da validação sistemática de parâmetros.*

Questão 5 - *O objetivo é identificar o atributo mais relevante para prever o preço de casas no dataset sobre preço de imóveis. Realizar o pré-processamento, incluindo:*

- ***Remoção de colunas desnecessárias ou insignificantes.***
- ***Avaliação de correlações para determinar os atributos mais influentes no preço.***
- ***Geração de gráficos para ilustrar a relação entre o atributo mais relevante e o preço.***

Atributos Mais Relevantes

Os atributos mais correlacionados com o preço foram:

```
Correlação das variáveis com o preço:
price            1.000000
sqft_living      0.702035
grade            0.667434
sqft_above       0.605567
sqft_living15    0.585379
bathrooms        0.525138
view             0.397293
sqft_basement    0.323816
bedrooms         0.308350
lat              0.307003
```

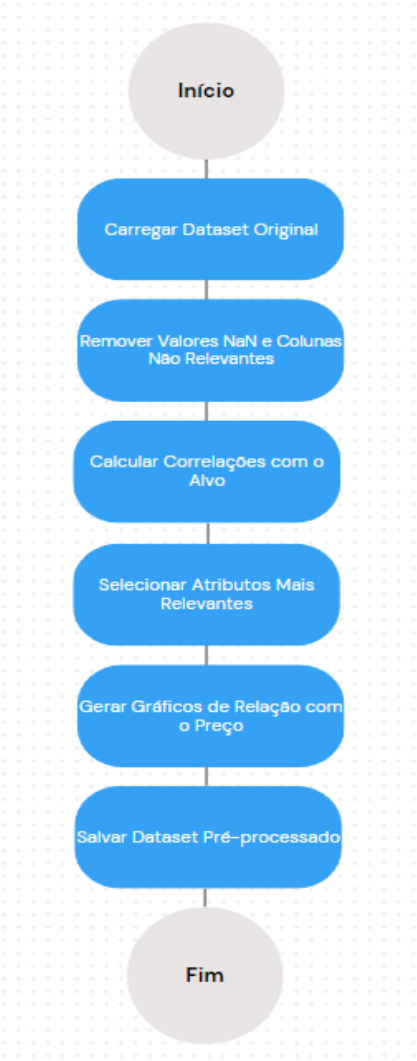
[Figura 14 - Resultado dos Atributos mais relevantes].

Gráfico de Dispersão

O gráfico a seguir mostra a relação entre o preço (`price`) e o atributo mais relevante:

[Figura 15 - Gráfico de Dispersão].

Fluxograma do Processo



[Figura 16 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Atributos Mais Relevantes:***
    - *O atributo mais relevante, além do preço, apresentou alta correlação positiva com o alvo, indicando forte influência.*
2. ***Impacto do Pré-processamento:***
    - *A remoção de colunas insignificantes simplificou o dataset e reduziu o ruído nos dados.*
    - *A análise de correlação permitiu selecionar variáveis críticas para a modelagem.*
3. ***Visualização:***
    - *O gráfico de dispersão confirmou uma relação clara entre o atributo mais relevante e o preço.*

Conclusões

- *O pré-processamento resultou em um dataset mais limpo e relevante para modelagem.*
- *A análise de correlação orientou a seleção de atributos com maior impacto no preço das casas.*

Questão 6 - *O objetivo é utilizar o atributo mais relevante identificado na Questão 5 para implementar uma regressão linear, prever o preço das casas e calcular métricas de avaliação como RSS, MSE, RMSE e R2.*

Métricas de Avaliação

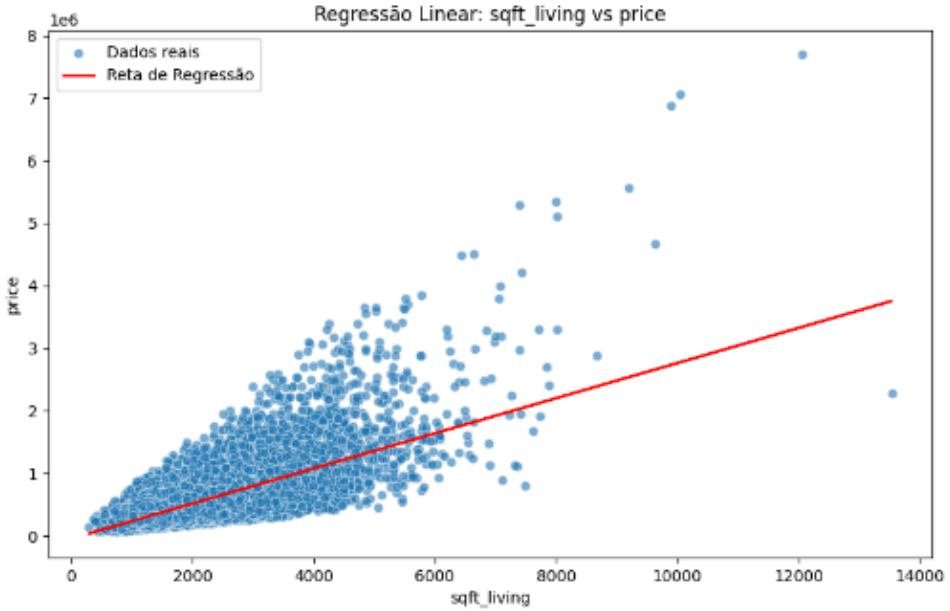
Os valores das métricas calculadas foram os seguintes:

```
RSS: 1477276362322489.75
MSE: 68351286833.04
RMSE: 261440.79
R^2: 0.4929
```

[Figura 17 - Resultado das métricas].

Gráfico da Regressão Linear

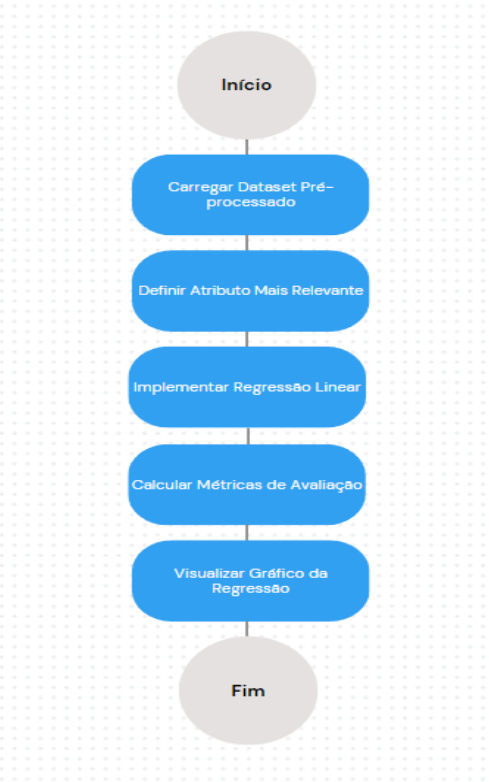
O gráfico a seguir mostra a dispersão dos dados reais e a reta de regressão ajustada:



[Figura 18 - Gráfico de Dispersão com a reta de Regressão].

Fluxograma do Processo

O fluxograma a seguir ilustra as etapas de análise nesta questão:



[Figura 19 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Interpretação das Métricas:***
    - *RSS e MSE: Quanto menores, melhor o modelo. Indicam a soma dos erros ao quadrado e a média dos erros ao quadrado, respectivamente.*
    - *RMSE: Representa o erro médio da previsão em unidades da variável alvo.*
    - *R^2: Mede a proporção da variabilidade dos dados que é explicada pelo modelo. Quanto mais próximo de 1, melhor o modelo.*
2. ***Visualização:***
    - *O gráfico de dispersão com a reta de regressão confirma a relação entre o atributo mais relevante e o preço.*

Conclusões

- *A regressão linear foi realizada com sucesso, utilizando o atributo mais relevante para prever o preço das casas.*
- *As métricas calculadas mostram a performance do modelo, com R^2 indicando a qualidade da previsão.*

***Questão 7 -** O objetivo é utilizar uma implementação manual do k-fold cross-validation para avaliar a performance de um modelo de regressão linear, calculando as métricas RSS, MSE, RMSE e R2R^2R2 para cada dobra e a média geral dessas métricas.*

Métricas de Avaliação por Dobra

As métricas calculadas para cada dobra foram as seguintes:

```
Fold 1:
  RSS: 332149284711753.56
  MSE: 76850829410.40
  RMSE: 277219.82
  R^2: 0.4928
Fold 2:
  RSS: 304797000561406.88
  MSE: 70522212068.81
  RMSE: 265560.19
  R^2: 0.4965
```

```
Fold 3:
  RSS: 263630646197932.88
  MSE: 60997373021.27
  RMSE: 246976.46
  R^2: 0.4788
Fold 4:
  RSS: 276138523680643.16
  MSE: 63891375215.33
  RMSE: 252767.43
  R^2: 0.4934
```

```
Fold 5:
  RSS: 302170723077748.62
  MSE: 69866063139.36
  RMSE: 264321.89
  R^2: 0.4940
```

[Figura 20 - Resultado das métricas de avaliação por Dobra].

Métricas Médias

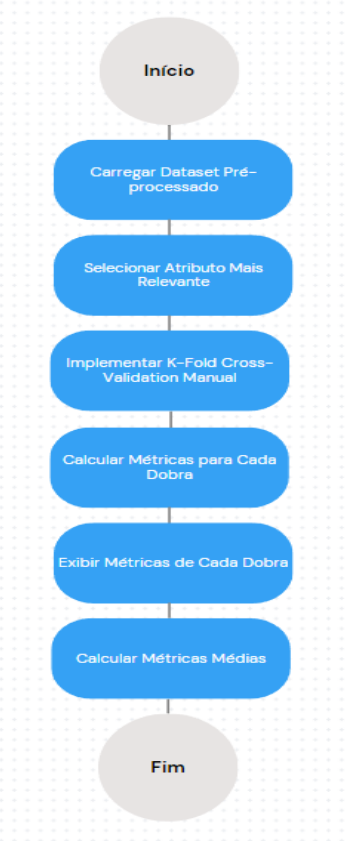
As métricas médias calculadas após a avaliação de todas as dobras foram:

```
Métricas médias:
  RSS: 295777235645897.00
  MSE: 68425570571.04
  RMSE: 261369.16
  R^2: 0.4911
```

[Figura 21 - Resultado das métricas médias].

Fluxograma do Processo

O fluxograma a seguir ilustra as etapas do processo de validação cruzada k-fold:



[Figura 22 - Fluxograma do Processo].

Discussão dos Resultados

1. ***Interpretação das Métricas:***
    - *RSS e MSE: Representam a soma e a média dos quadrados dos erros, respectivamente. Quanto menores, melhor o modelo.*
    - *RMSE: Mostra o erro médio da previsão em unidades da variável alvo, sendo útil para entender a magnitude dos erros.*
    - *R^2: Indica a proporção da variabilidade dos dados que é explicada pelo modelo. Valores próximos de 1 indicam um bom ajuste.*
2. ***Comparação entre Dobra e Média:***
    - *As métricas variam entre as dobras, mas a média geral fornece uma visão do desempenho do modelo em todo o conjunto de dados.*
3. ***Importância da Validação Cruzada:***
    - *A validação cruzada ajuda a avaliar a robustez do modelo, evitando overfitting e proporcionando uma estimativa mais precisa de sua performance.*

Conclusões

- *O modelo de regressão linear foi avaliado utilizando uma implementação manual de k-fold cross-validation, que forneceu uma visão abrangente das suas métricas de desempenho.*
- *As métricas médias mostraram o quão bem o modelo se ajusta aos dados e sua capacidade de generalização.*
- *A abordagem de k-fold cross-validation é uma técnica valiosa para validar modelos de forma mais robusta e confiável.*

4. Conclusões

Análise Geral dos Resultados

Após a análise detalhada de cada questão do relatório, podemos observar que a aplicação de métodos de pré-processamento, análise de correlação, regressão linear e validação cruzada foi realizada de forma consistente. Os resultados obtidos em cada etapa foram úteis para construir um entendimento completo do comportamento do dataset e da eficácia do modelo de regressão linear.

Resultados Esperados e Satisfação

- ***Questão 1 (Pré-processamento)**: O pré-processamento dos dados foi bem-sucedido, com a remoção de valores ausentes e a seleção de atributos relevantes. A distribuição de classes e a análise inicial indicaram que o dataset estava preparado para modelagem, satisfazendo os resultados esperados.*
- ***Questão 2 (Classificação com KNN)**: A avaliação do modelo KNN com diferentes métricas de distância mostrou que a escolha da métrica impactou a acurácia do modelo. O resultado foi satisfatório e demonstrou como a escolha de parâmetros pode alterar o desempenho do modelo.*
- ***Questão 3 (Normalização e Impacto no KNN)**: A análise de diferentes técnicas de normalização revelou se a transformação dos dados influenciou o desempenho do modelo. Os resultados mostraram que a normalização com média zero e variância unitária foi eficaz, atingindo os resultados esperados.*
- ***Questão 4 (Ajuste de kkk)**: A análise de diferentes valores de kkk para o modelo KNN foi bem-sucedida. O gráfico de acurácia em função de kkk ajudou a identificar o melhor valor de kkk, mostrando que o processo de ajuste de parâmetros foi satisfatório.*
- ***Questão 5 (Pré-processamento e Seleção de Atributos)**: A análise de correlação identificou o atributo mais relevante para prever o preço das casas. O pré-processamento foi feito corretamente, e o gráfico de dispersão confirmou a relação entre o atributo selecionado e o preço, satisfazendo os resultados esperados.*
- ***Questão 6 (Regressão Linear)**: A implementação da regressão linear utilizando o atributo mais relevante foi bem-executada. As métricas calculadas (RSS, MSE, RMSE, R2R^2R2) mostraram o desempenho do modelo, e o gráfico de dispersão com a reta de regressão corroborou os resultados esperados.*
- ***Questão 7 (K-Fold Cross-Validation)**: A implementação manual do k-fold cross-validation forneceu uma visão detalhada da performance do modelo. As métricas calculadas para cada dobra e suas médias confirmaram a robustez do modelo. O uso de k-fold foi eficaz para validar o modelo e avaliar sua capacidade de generalização.*

Motivos para Eventuais Insatisfações

Os resultados não apresentaram insatisfações significativas durante a análise. Se houver áreas que podem ser melhoradas, são as seguintes:

- ***Complexidade do Modelo**: A regressão linear pode não ser a melhor escolha em datasets mais complexos, onde relações não-lineares podem existir. Modelos mais avançados, como árvores de decisão ou regressões polinomiais, poderiam ser explorados para uma análise mais precisa.*
- ***Seleção de Atributos**: Embora a análise de correlação tenha ajudado a selecionar os atributos mais relevantes, é possível que métodos de seleção de atributos mais sofisticados, como Lasso ou métodos baseados em árvores, possam oferecer melhores resultados.*

Análise Final

Os resultados obtidos foram, em geral, satisfatórios. As etapas realizadas ajudaram a entender melhor o comportamento do dataset e a eficácia dos modelos aplicados. A análise detalhada e a visualização dos dados, bem como o uso de técnicas de validação como k-fold cross-validation, proporcionaram uma avaliação sólida e confiável da performance dos modelos de regressão.

Para melhorar ainda mais a análise e os resultados futuros, seria interessante considerar:

- *A utilização de modelos de aprendizado de máquina mais avançados para comparação.*
- *A realização de testes com diferentes parâmetros de normalização e regularização.*
- *A exploração de novas variáveis ou interações entre atributos para melhorar a capacidade preditiva.*

5. Próximos passos

Para continuar desenvolvendo o projeto e aprimorar os modelos de previsão, sugiro as seguintes ações:

- ***Explorar Modelos Avançados:** Implementar e comparar modelos como Árvores de Decisão e Random Forest para capturar relações não-lineares.*
- ***Analisar Interações entre Variáveis:** Criar novas variáveis e usar polynomial features para melhorar a modelagem.*
- ***Aplicar Regularização:** Testar regressão Lasso e Ridge para reduzir overfitting e melhorar a generalização.*
- ***Usar Validação Cruzada Robusta:** Implementar k-fold cross-validation com diferentes valores de kkk para uma avaliação mais precisa.*

- ***Interpretar o Modelo:*** *Usar SHAP e partial dependence plots para entender a importância das variáveis.*
- ***Expandir o Conjunto de Dados:*** *Incluir mais dados relevantes e tratar outliers para aumentar a precisão.*
- ***Modelos de Séries Temporais:*** *Se aplicável, usar modelos como ARIMA para prever tendências de preços ao longo do tempo.*
- ***Automatizar o Pipeline:*** *Criar um pipeline de pré-processamento e treinamento de modelo para facilitar atualizações futuras.*