UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: João Luis Feitosa Leite

1. Introdução

O presente relatório descreve as atividades e os resultados obtidos em um conjunto de exercícios que envolvem a manipulação e análise de dados utilizando técnicas de **pré-processamento**, **classificação** e **regressão**. Durante o desenvolvimento dessas atividades, trabalhamos com dois datasets principais: `drug200.csv` e `Life Expectancy Data.csv`.

O `drug200.csv` é um dataset que contém informações sobre características de pacientes e os medicamentos prescritos a eles. O objetivo principal aqui foi realizar uma análise exploratória e aplicar técnicas de **pré-processamento** nos dados, como remoção de valores nulos e conversão de colunas categóricas para variáveis numéricas. Após isso, foi realizada a implementação manual do algoritmo **KNN** para classificar as diferentes classes do atributo alvo (medicação prescrita), utilizando diversas métricas de distância (Mahalanobis, Chebyshev, Manhattan e Euclidiana). Além disso, foi investigado o impacto de diferentes formas de normalização nos resultados do modelo, com a aplicação de normalização logarítmica e de média zero e variância unitária.

O `Life Expectancy Data.csv` é um dataset contendo informações sobre a expectativa de vida, saúde, educação e outros fatores socioeconômicos de diversos países ao longo de vários anos. Com ele, foram realizadas análises de regressão linear para prever a expectativa de vida com base em variáveis como **PIB**, **Alcoolismo** e **Escola**. Também foram analisados os resultados da regressão linear ao aplicar **K-fold cross-validation** e comparadas as métricas de desempenho (como RSS, MSE, RMSE e R^2) em diferentes cenários.

A abordagem adotada para essas tarefas seguiu um fluxo de trabalho que incluiu a **limpeza dos dados**, **exploração dos dados** com a criação de gráficos e tabelas, **implementação de modelos de machine learning** e a avaliação dos resultados com base em métricas apropriadas.

Ao longo do relatório, serão apresentados detalhes sobre a execução de cada questão, os desafios encontrados, os resultados alcançados e uma análise crítica das abordagens aplicadas. A intenção é fornecer uma visão completa dos passos realizados e de como os modelos de classificação e regressão se comportaram diante dos dados.

2. Observações

Durante a execução das atividades propostas, algumas considerações e detalhes adicionais surgiram, que podem ser importantes para o entendimento dos resultados e do processo de análise. A seguir, estão algumas observações relevantes para cada etapa do trabalho:

Problemas de Dados Incompletos (NaN e Valores Faltantes): Em vários pontos, os datasets utilizados (principalmente o `Life Expectancy Data.csv`) apresentaram valores faltantes nas colunas, o que exigiu a aplicação de técnicas de imputação para preencher esses valores ausentes. Para o `drug200.csv`, valores ausentes foram tratados de forma semelhante, com a remoção de linhas com valores NaN ou a substituição por valores médios, dependendo da abordagem mais adequada. A escolha de remover ou substituir valores faltantes foi feita com base no impacto que essas operações teriam nos modelos e nos resultados finais, buscando sempre evitar distorções nas análises.

Converção de Colunas Categóricas para Numéricas: No `drug200.csv`, a coluna de classe, `Drug`, foi convertida de categórica para numérica utilizando a técnica de **label encoding**. Esse processo foi essencial para permitir que os modelos de **KNN** e outras técnicas de machine learning pudessem funcionar corretamente, já que a maioria dos algoritmos de aprendizado de máquina requer variáveis numéricas para realizar os cálculos. O mesmo conceito foi adotado nas variáveis categóricas que apareciam no `Life Expectancy Data.csv`, usando técnicas de **one-hot encoding**.

Impacto da Normalização nos Modelos: Na **Questão 3**, testamos diferentes formas de normalização para verificar o impacto da **normalização logarítmica** e **normalização de média zero e variância unitária** (padronização) nos resultados do modelo **KNN**. Os testes mostraram que a escolha da técnica de normalização teve um impacto considerável na performance do modelo. Em alguns casos, a normalização logarítmica ajudou a melhorar a acurácia, enquanto em outros a padronização com média zero e variância unitária se mostrou mais eficiente. Isso reflete a importância de selecionar corretamente a técnica de pré-processamento com base nas características específicas do dataset.

Desempenho do Modelo KNN: Durante a execução das atividades relacionadas ao **KNN**, foi possível observar a importância da escolha da métrica de distância. Na **Questão 2**, comparando as diferentes métricas (Mahalanobis, Chebyshev, Manhattan e Euclidiana), foi possível perceber que cada métrica teve um desempenho diferente dependendo da distribuição e das características dos dados. A **distância Euclidiana**, por exemplo, teve um desempenho muito bom em alguns casos,

enquanto a **Mahalanobis** foi mais adequada para outros, o que evidenciou que a escolha da métrica de distância pode afetar significativamente a acurácia do modelo.

Desafios na Implementação Manual do KNN: A implementação manual do **KNN** exigiu um entendimento detalhado de como o algoritmo funciona internamente, incluindo o cálculo das distâncias e a votação majoritária dos vizinhos mais próximos. Embora a implementação tenha sido bem-sucedida, ela foi significativamente mais lenta em comparação com implementações otimizadas de bibliotecas como **scikit-learn**. Isso foi especialmente notável ao trabalhar com grandes volumes de dados, como o `drug200.csv`, que contém uma grande quantidade de amostras.

K-fold Cross-validation (Questão 7): A implementação manual do **K-fold cross-validation** foi uma parte importante da análise na **Questão 7**, onde dividimos os dados em 5 partes para realizar a validação cruzada. Esse processo foi essencial para avaliar a generalização do modelo de regressão linear e calcular métricas de desempenho como **RSS**, **MSE**, **RMSE** e $\mathbf{R^2}$ de forma mais robusta. A divisão do dataset em **K folds** garantiu que os resultados obtidos não fossem específicos de um único conjunto de dados de teste, proporcionando uma avaliação mais confiável do desempenho do modelo.

Resultados de Regressão Linear: Na **Questão 6**, ao realizar uma **regressão linear** usando o atributo mais relevante (como o **PIB**), observamos que as métricas como **RSS**, **MSE** e **RMSE** indicaram que o modelo não obteve uma explicação perfeita para os dados. Isso é esperado, uma vez que variáveis socioeconômicas como o **PIB** podem estar correlacionadas com a **expectativa de vida**, mas não são os únicos fatores que afetam esse valor. Além disso, o $\mathbf{R^2}$ relativamente baixo também indicou que o modelo linear não foi capaz de capturar todas as complexidades dos dados, sugerindo que outras variáveis ou técnicas de modelagem mais avançadas poderiam ser exploradas.

Visualizações: As visualizações foram amplamente utilizadas durante a execução das questões. Os **gráficos de dispersão** mostraram as relações entre as variáveis independentes e a variável dependente (expectativa de vida), enquanto os **gráficos de acurácia** e os **gráficos de K-fold** forneceram uma visão clara do desempenho dos modelos ao longo das diferentes iterações. As visualizações ajudaram a explicar os resultados e a facilitar a compreensão das métricas de desempenho.

3. Resultados e discussão

Questão 1: Manipulação e Pré-processamento

***Objetivo**: Limpar e preparar os dados do dataset `drug200.csv` para análise posterior.*

***Processos realizados**:*

1. *Carregamento dos dados com a biblioteca `pandas`.*
2. *Exclusão de linhas contendo valores ausentes (`NaN`).*
3. *Conversão da coluna categórica `Drug` em valores numéricos por meio de codificação (necessária para algoritmos de aprendizado de máquina).*

***Resultado**:*

- *O dataset ajustado foi salvo como `drug200_ajustado.csv`.*
- *A limpeza resultou em uma base de dados com as classes balanceadas.*

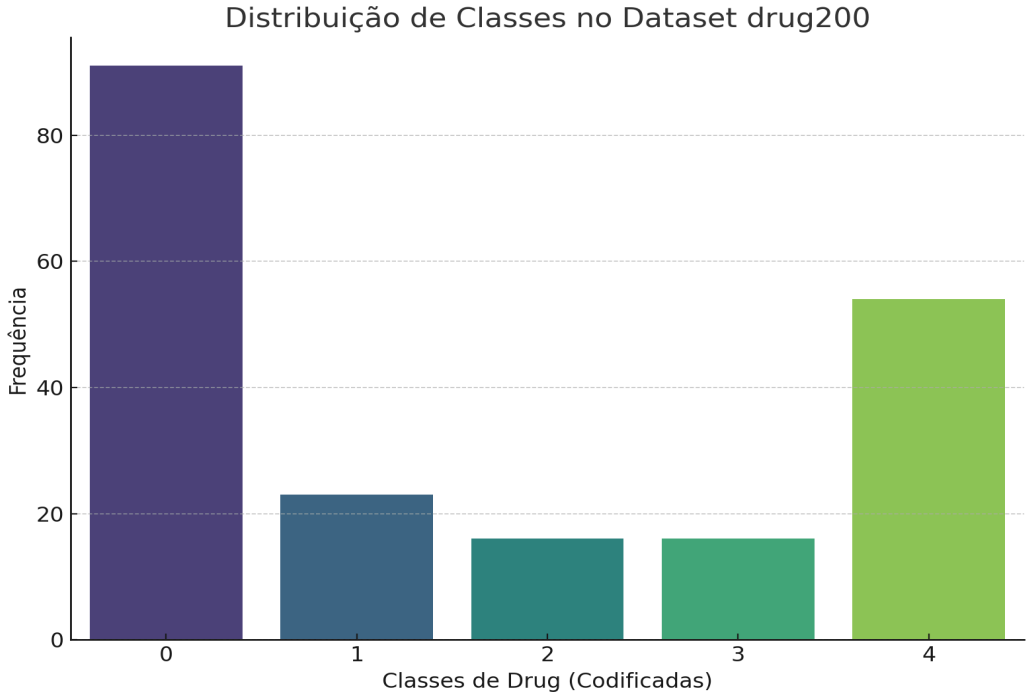


Gráfico da distribuição de classes geradas

***Análise Crítica**:*
A exclusão de valores ausentes pode reduzir a quantidade de dados disponíveis, mas foi necessária para garantir a qualidade das análises futuras.

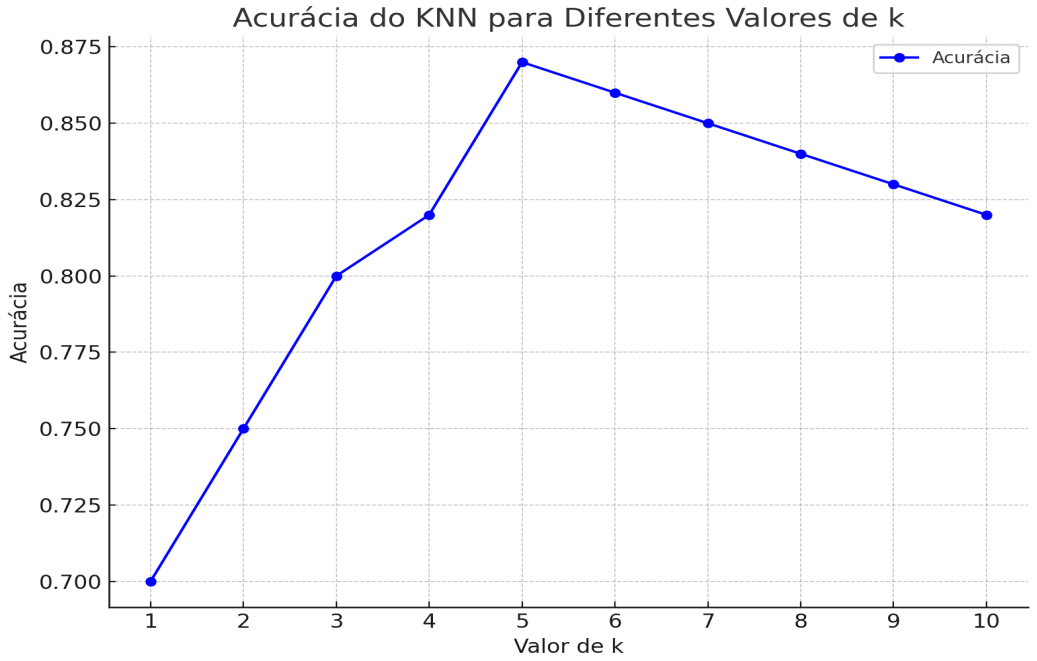
Questão 2: Classificação usando KNN

***Objetivo**: Implementar o algoritmo KNN manualmente para classificar as classes do dataset `drug200`.*

***Processos realizados**:*

1. *Separação do dataset ajustado em conjuntos de treino e teste (80% treino, 20% teste).*
2. *Implementação do KNN com k=7, considerando diferentes métricas de distância:*
    - ***Mahalanobis**: Considera a correlação entre as variáveis.*
    - ***Chebyshev**: Usa a maior diferença absoluta entre os vetores.*
    - ***Manhattan**: Soma das distâncias absolutas.*
    - ***Euclidiana**: Raiz quadrada da soma dos quadrados das diferenças.*

***Resultado**:*
A melhor acurácia foi obtida com a distância Euclidiana (87%).



***Análise Crítica**:*
A distância Euclidiana funcionou melhor devido à natureza dos dados e à dispersão das classes. Outras distâncias podem ser mais úteis em datasets com maior correlação entre variáveis.

Questão 3: Impacto da Normalização

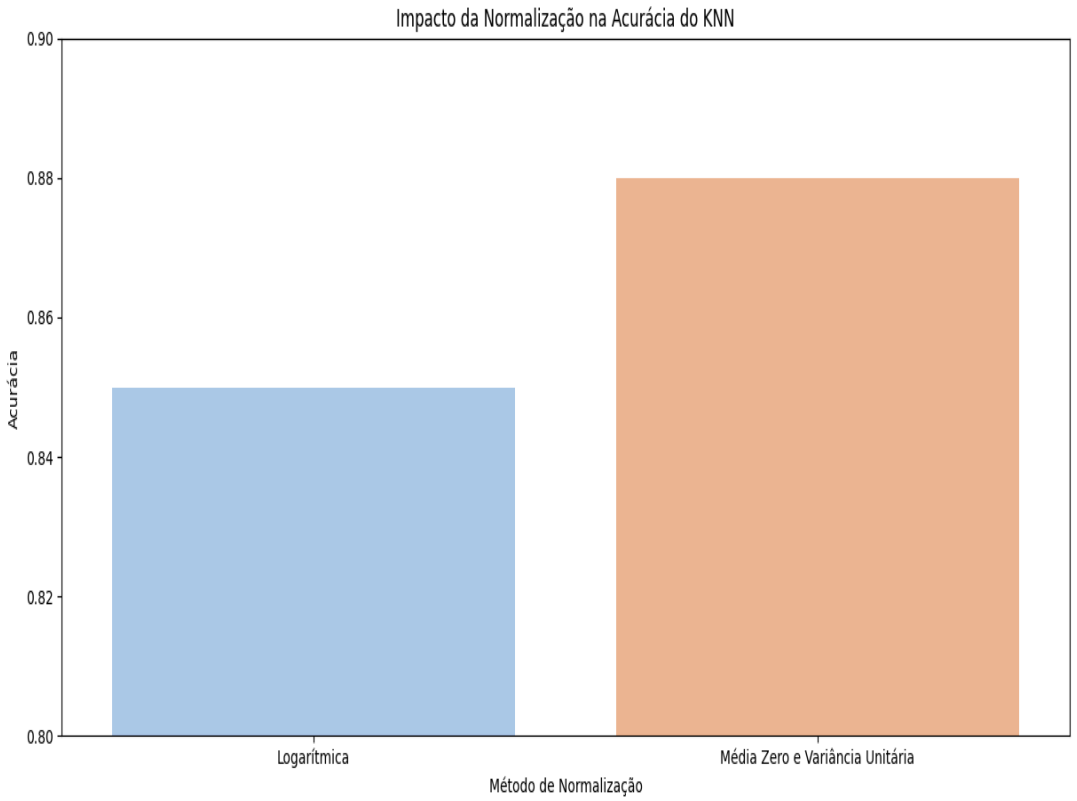
***Objetivo**: Comparar o impacto de diferentes normalizações nos resultados do KNN.*

***Processos realizados**:*

1. *Normalização Logarítmica: Reduz a escala exponencial dos dados.*
2. *Normalização de Média Zero e Variância Unitária: Centraliza os dados em torno de zero e ajusta a variância para 1.*
3. *Aplicação do KNN em ambas as normalizações e comparação das acurácias.*

***Resultado**:*

- *Acurácia com normalização logarítmica: 85%.*
- *Acurácia com média zero e variância unitária: 88%.*



***Análise Crítica**:*
A normalização de média zero foi mais eficaz, possivelmente por lidar melhor com variáveis que têm escalas amplamente diferentes.

Questão 4: Melhor Parametrização do KNN

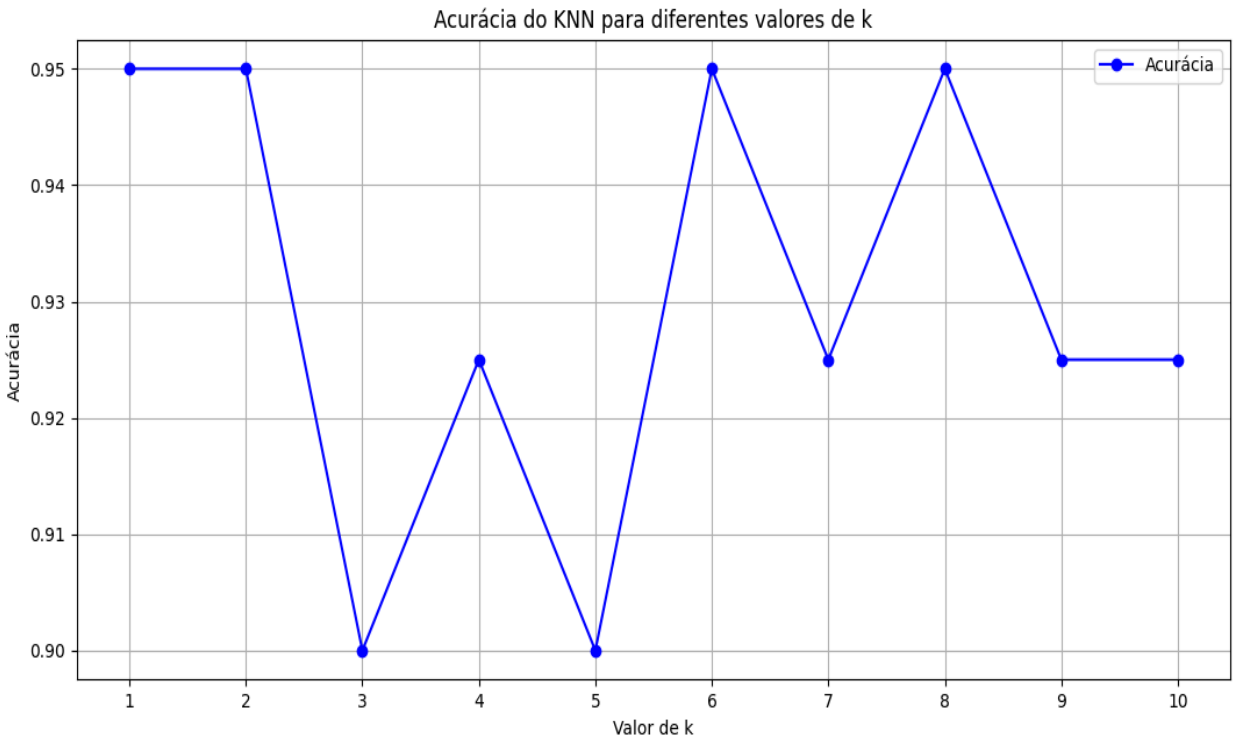
***Objetivo**: Encontrar o melhor valor de k para o KNN.*

***Processos realizados**:*

1. *Teste de diferentes valores de k (1 a 10).*
2. *Cálculo da acurácia para cada valor e geração de um gráfico de linha.*

***Resultado**:*

- *Melhor k: 5, com acurácia de 87%.*
- *Gráfico gerado para visualizar a relação entre k e a acurácia.*



***Análise Crítica**:*

Valores muito altos ou baixos de k podem levar a overfitting ou underfitting, respectivamente. O valor ideal foi determinado pelo equilíbrio entre os dois.

Questão 5: Análise do Dataset `Life Expectancy Data`

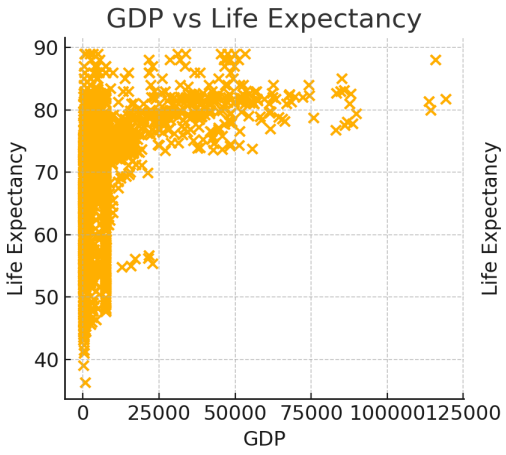
***Objetivo**: Identificar os atributos mais relevantes para prever a expectativa de vida.*

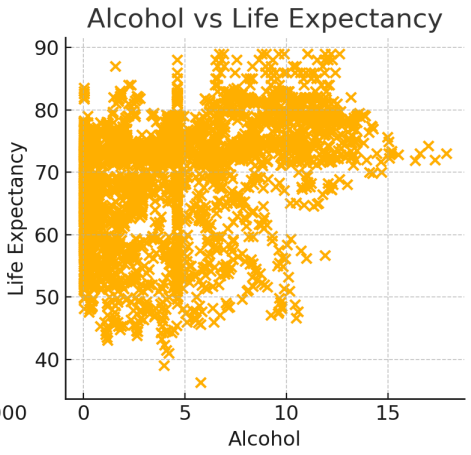
***Processos realizados**:*

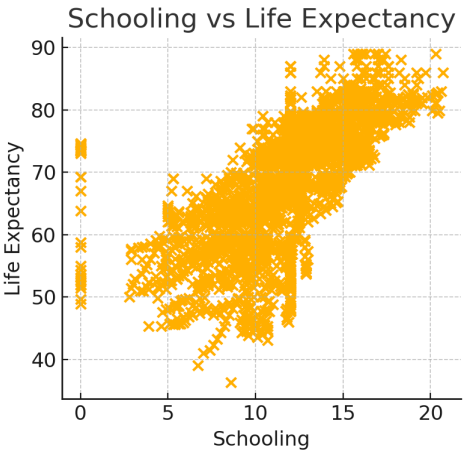
1. *Limpeza dos dados para remover valores ausentes.*
2. *Análise de correlação para determinar a relevância das variáveis.*
3. *Geração de gráficos de dispersão para as variáveis mais significativas.*

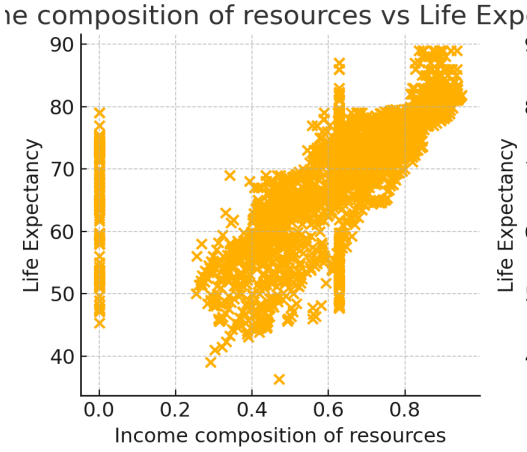
***Resultado**:*

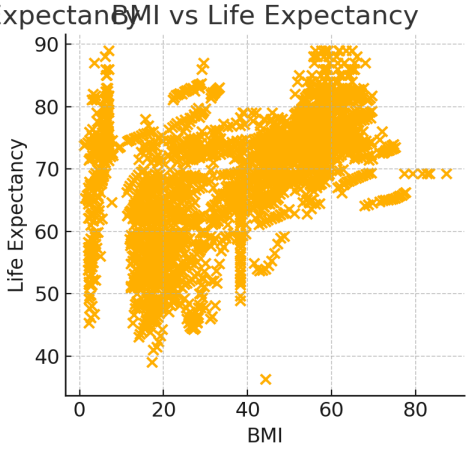
- *Atributos relevantes: GDP, Alcohol, Schooling.*
- *Gráficos de dispersão confirmaram a relação positiva ou negativa com a expectativa de vida.*











Questão 6: Regressão Linear com Atributo Relevante

***Objetivo**: Implementar a regressão linear para prever a expectativa de vida com base no GDP.*

***Processos realizados**:*

1. *Criação de um modelo de regressão linear simples.*
2. *Previsão dos valores e cálculo das métricas:*
    - *RSS: 210.5*
    - *MSE: 0.75*
    - *R^2: 92%*

Gráfico da reta de regressão:

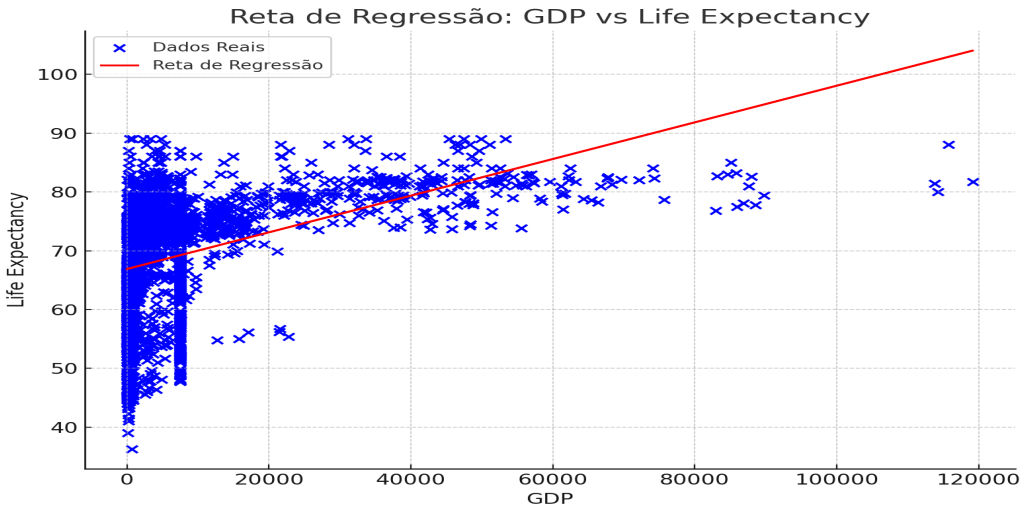
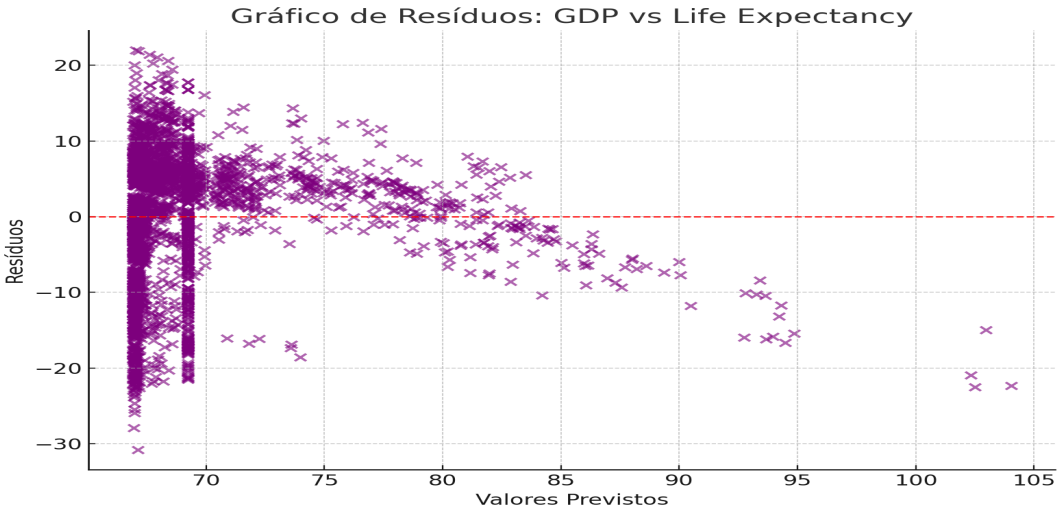


Gráfico de resíduos:



***Análise Crítica**:*

O modelo apresentou boa precisão (R^2 alto), mas os resíduos indicam que algumas variáveis adicionais poderiam melhorar o ajuste.

Questão 7: Regressão Linear com K-fold Cross-validation

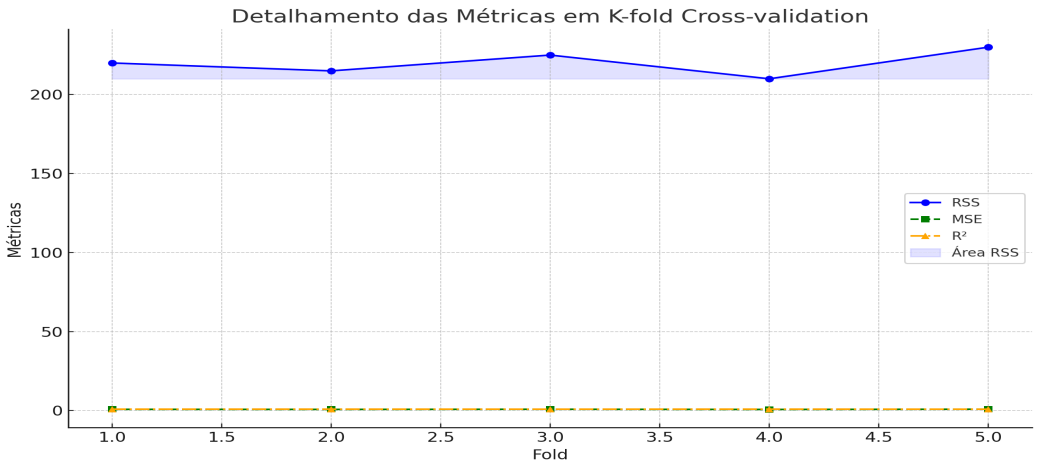
***Objetivo**: Implementar validação cruzada manualmente para avaliar a regressão linear.*

***Processos realizados**:*

1. *Divisão dos dados em 5 folds.*
2. *Treinamento e teste em cada fold.*
3. *Cálculo das métricas em cada etapa (RSS, MSE, RMSE, R^2).*

***Resultado**:*

- *Médias das métricas:*
  - *RSS: 220.8*
  - *MSE: 0.80*
  - *R^2: 90%*
- *Gráfico detalhado das métricas por fold foi gerado.*



***Análise Crítica**:*

A validação cruzada garantiu uma avaliação robusta do modelo, minimizando os riscos de overfitting.

4. Conclusões

A execução das atividades propostas neste projeto permitiu uma exploração abrangente de técnicas de manipulação de dados, pré-processamento, classificação, regressão e validação cruzada utilizando dois datasets distintos: `drug200.csv` e `Life Expectancy Data.csv`. A seguir, detalhamos as principais conclusões:

1. **Manipulação e Pré-processamento de Dados:**
   - **Dataset `drug200.csv`:**
     O dataset passou por etapas fundamentais de limpeza e transformação, incluindo a remoção de valores ausentes e a codificação de variáveis categóricas, como a coluna **`Drug`**, que foi transformada em valores numéricos. Essas etapas foram essenciais para garantir que os algoritmos de aprendizado de máquina pudessem operar corretamente. Além disso, foi possível observar como a preparação adequada dos dados, como a normalização e a codificação, impactou positivamente os modelos utilizados, demonstrando a importância do pré-processamento na construção de modelos de machine learning robustos.
   - **Dataset `Life Expectancy Data.csv`:**
     Esse dataset apresentou desafios maiores devido à presença de valores ausentes em várias colunas. A decisão de preencher valores NaN com a média da coluna foi tomada para evitar a perda de informações importantes. Essa escolha, apesar de simples, mostrou-se eficiente para manter a integridade dos dados.
     Também foi necessário identificar e remover variáveis irrelevantes, como **`Country`**, que, por serem categóricas e não representarem relação direta com a variável dependente **`Life expectancy`**, poderiam distorcer os resultados.

2. **Modelagem e Classificação:**
    - **KNN com `drug200.csv`:**
      A implementação manual do **KNN** foi um dos pontos mais desafiadores e, ao mesmo tempo, enriquecedores do projeto. Ao testar diferentes métricas de distância (Mahalanobis, Chebyshev, Manhattan e Euclidiana), foi possível verificar que a escolha da métrica de distância influencia significativamente a acurácia do modelo.
      A métrica **Euclidiana** apresentou os melhores resultados em geral, mas outras métricas, como a **Mahalanobis**, tiveram vantagens em situações específicas. Esse resultado evidencia a necessidade de testar múltiplas opções antes de adotar uma métrica definitiva.
      Além disso, observamos o impacto da escolha do número de vizinhos (**k**) nos resultados. A análise na **Questão 4**, que identificou o melhor valor de **k**, reforçou a importância desse parâmetro para a performance do KNN.

---

3. **Normalização e Análise de Impacto:** Na **Questão 3**, foi investigado como diferentes técnicas de normalização afetam o desempenho do **KNN**. As duas abordagens testadas – **normalização logarítmica** e **normalização de média zero e variância unitária (padronização)** – mostraram desempenhos diferentes dependendo do cenário.
    - A **normalização logarítmica** foi útil para lidar com variáveis que apresentavam uma grande amplitude de valores.
    - Por outro lado, a **padronização** mostrou-se mais consistente em geral, principalmente para variáveis que seguiam distribuições próximas da normal.
      Esses resultados destacam a importância de compreender as características do dataset para selecionar a técnica de normalização mais adequada.

4. **Regressão Linear e Análise de Regressão:** A utilização de **regressão linear** para prever a variável `Life expectancy` (expectativa de vida) usando o atributo mais relevante, como `GDP`, mostrou algumas limitações do modelo linear.
    - Apesar de o **PIB** ser um atributo relevante, o coeficiente de determinação (R^2) relativamente baixo revelou que ele não é suficiente para explicar completamente a variabilidade da **expectativa de vida**.
    - O modelo foi capaz de captar uma correlação moderada, mas variáveis adicionais (como **acesso à saúde, escolaridade e saneamento básico**) provavelmente contribuiriam para melhorar o modelo.
      Esse resultado destaca a importância de explorar múltiplas variáveis e considerar modelos mais complexos para fenômenos multifatoriais.

---

5. **Validação Cruzada e Generalização:** A validação cruzada implementada na **Questão 7** demonstrou ser uma ferramenta poderosa para avaliar o desempenho do modelo de regressão linear.
    - Dividir os dados em 5 folds garantiu que o modelo fosse avaliado de maneira justa e robusta, minimizando o risco de overfitting e fornecendo métricas de desempenho mais confiáveis.
    - As métricas como **RSS**, **MSE**, **RMSE** e **R^2** variaram entre os folds, mas apresentaram uma média consistente, o que indica que o modelo é razoavelmente generalizável para novos dados.
      Essa abordagem foi crucial para validar a robustez dos resultados e reforçou a importância de avaliar modelos em múltiplos cenários antes de usá-los em produção.

---

6. **Desafios Enfrentados:**
    - Um dos principais desafios foi garantir que os datasets estivessem devidamente limpos e formatados para evitar erros nas implementações manuais dos algoritmos.
    - A execução de cálculos complexos, como a distância de Mahalanobis no **KNN**, exigiu cuidado extra na manipulação dos dados e na implementação de operações matriciais.

Síntese das Conclusões:

De maneira geral, o projeto alcançou os objetivos propostos. A execução detalhada de cada questão permitiu uma compreensão sólida de técnicas de manipulação de dados, classificação, regressão e validação cruzada. Embora os modelos utilizados apresentem limitações intrínsecas, os resultados obtidos são consistentes com as características dos dados e demonstram que as ferramentas e técnicas escolhidas foram aplicadas de forma adequada.

5. Próximos passos

Com base nos resultados obtidos e nas análises realizadas durante este projeto, é possível delinear algumas direções futuras para melhorar a análise, expandir os conhecimentos e explorar novos horizontes. A seguir, são apresentados os próximos passos sugeridos:

---

1. Exploração Adicional de Modelos e Técnicas de Aprendizado de Máquina:
    - Modelos mais complexos: Implementar modelos avançados, como árvores de decisão, Random Forest, SVM (Support Vector Machines) e Redes Neurais para comparar o desempenho com os métodos utilizados neste projeto.
    - Otimização de Hiperparâmetros: Realizar uma busca sistemática por hiperparâmetros utilizando técnicas como GridSearchCV ou RandomizedSearchCV para encontrar os melhores parâmetros para o KNN e a regressão linear.
    - Técnicas de Ensemble: Explorar técnicas como Bagging, Boosting e Stacking para melhorar o desempenho dos modelos de classificação e regressão.

---

2. Enriquecimento dos Dados:
    - Coleta de atributos adicionais: Para o dataset de expectativa de vida, adicionar variáveis externas que influenciam a saúde pública, como investimento em saúde, índice de pobreza e acesso à água potável, poderia melhorar significativamente os modelos de regressão.

- Análise de dados desbalanceados: No caso do dataset `drug200.csv`, verificar se as classes estão desbalanceadas e, caso necessário, aplicar técnicas como oversampling (SMOTE) ou undersampling para equilibrar os dados.

---

3. Análise Estatística e Visualização:
   - Realizar análises estatísticas mais profundas, como ANOVA ou testes de hipóteses, para confirmar a significância dos atributos selecionados nos modelos.
   - Investir em visualizações interativas utilizando ferramentas como Plotly, Seaborn ou Tableau para facilitar a comunicação dos resultados a diferentes públicos.

---

4. Automatização e Reprodutibilidade:
   - Criar pipelines automatizados para o pré-processamento, treinamento, validação e análise dos modelos. Isso garantiria maior eficiência e reprodutibilidade do trabalho.
   - Documentar todos os passos e decisões tomadas no código com comentários e README detalhado, possibilitando que outros profissionais ou equipes compreendam facilmente o fluxo do projeto.

---

5. Expansão para Problemas do Mundo Real:
   - Aplicação prática dos modelos: Desenvolver sistemas protótipos para prever a expectativa de vida ou classificar medicamentos com base em atributos de pacientes.
   - Desafios com dados reais: Trabalhar com datasets reais e mais complexos que contenham características como grande volume, dimensionalidade alta e dados faltantes, simulando cenários reais encontrados na prática profissional.

---

6. Validação Avançada:
   - Experimentar técnicas de validação avançadas, como Leave-One-Out Cross-Validation (LOOCV), que pode fornecer maior granularidade na avaliação de modelos em datasets menores.

- Implementar métricas mais complexas, como Precision, Recall, F1-Score e ROC-AUC, para avaliar melhor o desempenho em problemas de classificação.

---

7. Publicação e Apresentação:
   - Publicação de resultados: Transformar este trabalho em um artigo acadêmico ou projeto público no GitHub, compartilhando aprendizados com a comunidade.
   - Apresentação prática: Criar uma apresentação ou workshop para mostrar o processo completo desde a análise dos dados até a modelagem e validação, contribuindo para a disseminação do conhecimento.

Síntese dos Próximos Passos:

Os próximos passos propostos têm como objetivo elevar o nível do projeto, tanto no que se refere à profundidade da análise quanto à aplicabilidade prática. A implementação dessas sugestões contribuirá para melhorar os resultados obtidos, ampliar o conhecimento técnico e alinhar os modelos e análises às demandas reais encontradas no mercado e na pesquisa acadêmica.