UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Antonio Lucas Melo de Sousa

1. Introdução

Neste relatório, serão analisados dois datasets distintos: um de classificação, denominado *Cancer_Data*, e outro de regressão, denominado *possum*. Ambos os datasets foram obtidos da plataforma *Kaggle*, sendo utilizados para diferentes tipos de análise preditiva.

Dataset de Classificação: *Cancer_Data.csv*

Este dataset contém informações sobre características físicas de tumores em exames de mamografia, com o objetivo de prever se o tumor é maligno ou benigno. Com 569 instâncias e 33 colunas, a variável alvo, denominada *diagnosis*, classifica os tumores como *M* (maligno) ou *B* (benigno). As características presentes no dataset, como *radius_mean*, *texture_mean*, *perimeter_mean*, e *area_mean*, são medições relacionadas ao tamanho, forma e textura dos tumores. O objetivo da análise é explorar a distribuição das classes, realizar o pré-processamento dos dados, incluindo a limpeza de dados ausentes e a seleção das colunas mais relevantes, e, finalmente, treinar modelos preditivos para identificar tumores malignos e benignos.

Dataset de Regressão: *possum.csv*

O dataset *possum.csv* contém dados relacionados às características físicas de 365 indivíduos de gambás(marsupiais), sendo utilizado para modelagem preditiva em um problema de regressão. As variáveis incluem medidas como *age* (idade), *hdlngth* (comprimento da cabeça), *skullw* (largura do crânio), e *footlgth* (comprimento do pé), entre outras. O objetivo da análise é investigar como as características físicas dos gambás, podem ser usadas para prever variáveis como idade ou comprimento da cabeça, além de explorar possíveis relações entre essas medições e o sexo ou a localização do animal.

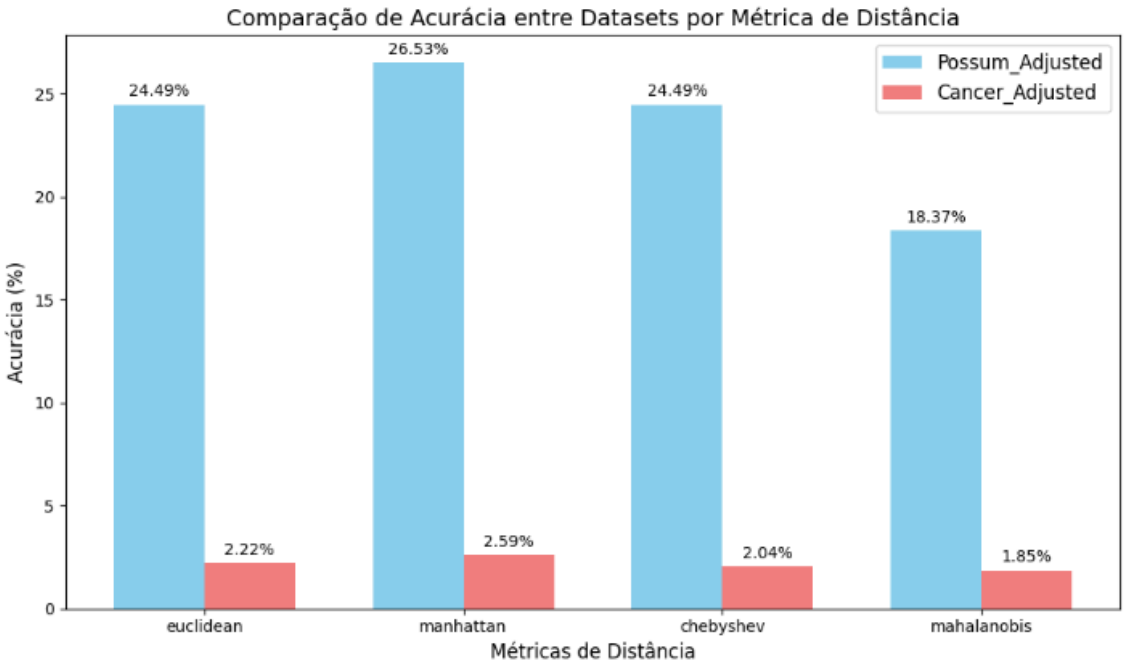
Observações

1º e 2º Questão, Baixa Acurácia

Na primeira questão deveria ser feito a normalização dos datasets, mas ao iniciar a 2º questão no qual pede para realizar os cálculos de distância com o KNN, como

euclidiana e manhattan, percebi que o dataset apresentava erros, pois estava utilizando um não ajustado, o que ocorreu foi um erro no ajuste no dataset de classificação, no qual ao remover dados vazios e inúteis resultou na remoção de todas as colunas que tinham ao menos 1 dado vazio, assim deletando o dataset praticamente por completo, após corrigir o código para não eliminar as colunas com o dropna, especificando os dados a serem deletados e não permitindo deletar a coluna toda, consegui ajustar o dataset.

Em seguido ocorreu outro problema, a acurácia estava muito baixo, o que resultou em uma dúvida quando ao sucesso do ajuste, então utilizei o *StandarScaler* do *sklearn* para normalizar as escalas, mas mesmo assim manteve a acurácia baixa, quando alterando a quantidade de exemplos esse valor caiu mais ainda, então se eu possuísse um dataset maior em questão de dados corretos, a acurácia aumentaria, no atual momento está em cerca de 24% a depender da distância utilizada, então outro fator pode ser o ajuste incorreto, pois alguns dados foram convertidos para valores numéricos, e outros valores que não são utilizados foram ignorados ou deletados.
Resultado final do *Possum_Data_Adjusted.csv vs Cancer_Data_Adjsuted:*



2. Resultados e discussão

1-

Nesta 1º questão o objetivo é realizar a manipulação de um dataset e realizar o pré-processamento.

O 1º erro encontrado foi a presença de uma linha vazia no final do *Cancer_Data.* O fluxo segue os propostos no readme, começando pela importação das bibliotecas necessárias, nesse caso *pandas* e *scikit*. Em seguida foi feito o carregamento de ambos os datasets, o de classificação *Cancer_Data* e o de regressão p*ossum.* Em seguida fiz uma pré-visualização das primeiras linhas, além disso foi verificado células vazias e feito a remoção de colunas irrelevantes, como "Unnamed: 32" no dataset de classificação e "case" no dataset de regressão. O próximo passo foi selecionar as colunas mais relevantes a serem utilizadas. Por último foi feito a renomeação de classes no dataset de classificação, pois possuía dados como strings importantes, então convertendo M para 1 e B para 0, sendo B tumores classificados como benignos e M como maligno. Em seguida utilizei o *StandardScaler* para realizar uma normalização nos dados do *possum*, o dataset de regressão assim finalizando o código com os datasets ajustados.

2-

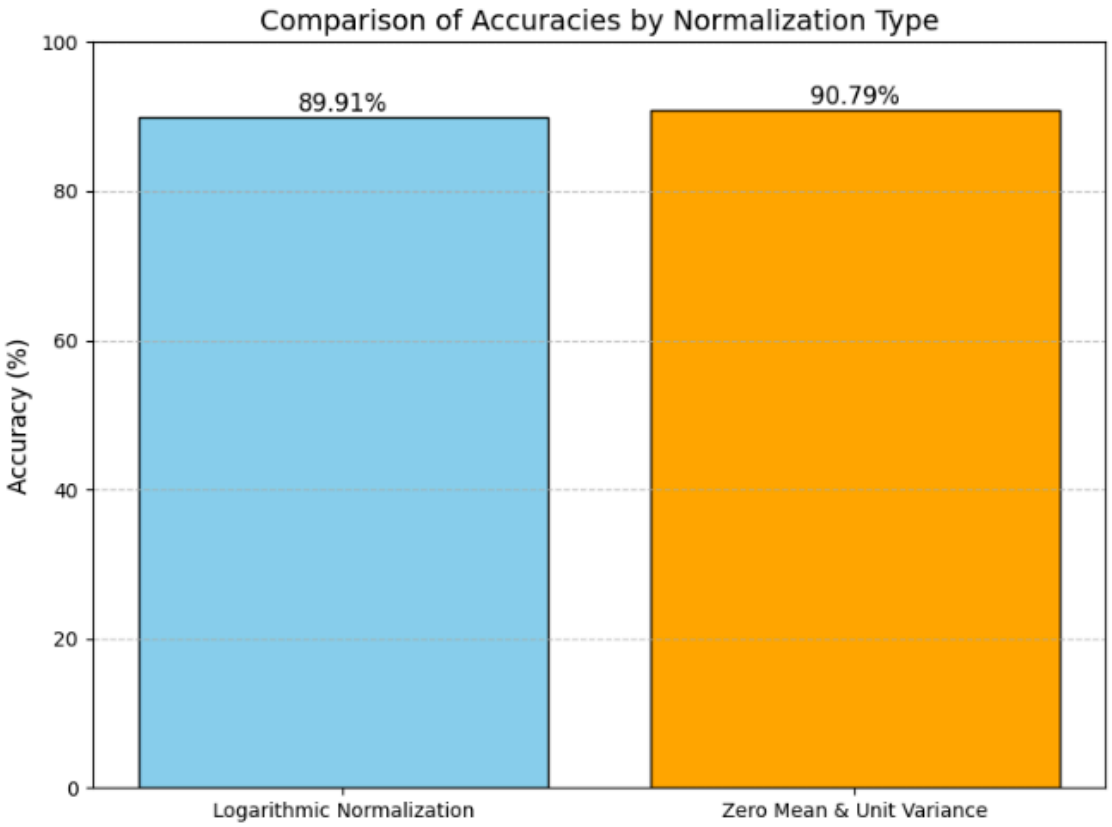
Na segunda questão foi realizada uma classificação utilizando o KNN implementando de forma manual. Neste exemplo eu utilizei os datasets ajustados, *Cancer_Data_Adjusted* e *Possum_Adjusted.* Em seguida utilizei o *StandardScaler* para realizar uma normalização nos dados do *possum*, o dataset de regressão assim finalizando o código com os datasets ajustados. Sem normalizar o conjunto como pedido, os dados foram divididos em treino e tese. Em seguida é feito a implementação do KNN exibindo a acurácia dos dados de teste, utilizando um valor de k = 7. Para realizar as comparações entre as 4 distâncias, *manhalanobis, chebyshev, manhattan, euclidiana*. No tópico de observações há um gráfico que representa esses resultados.

3-

Nesta questão foi escolhida a distância com melhor acurácia, no caso a manhattan. O objetivo é verificar se a normalização impacta no resultado.

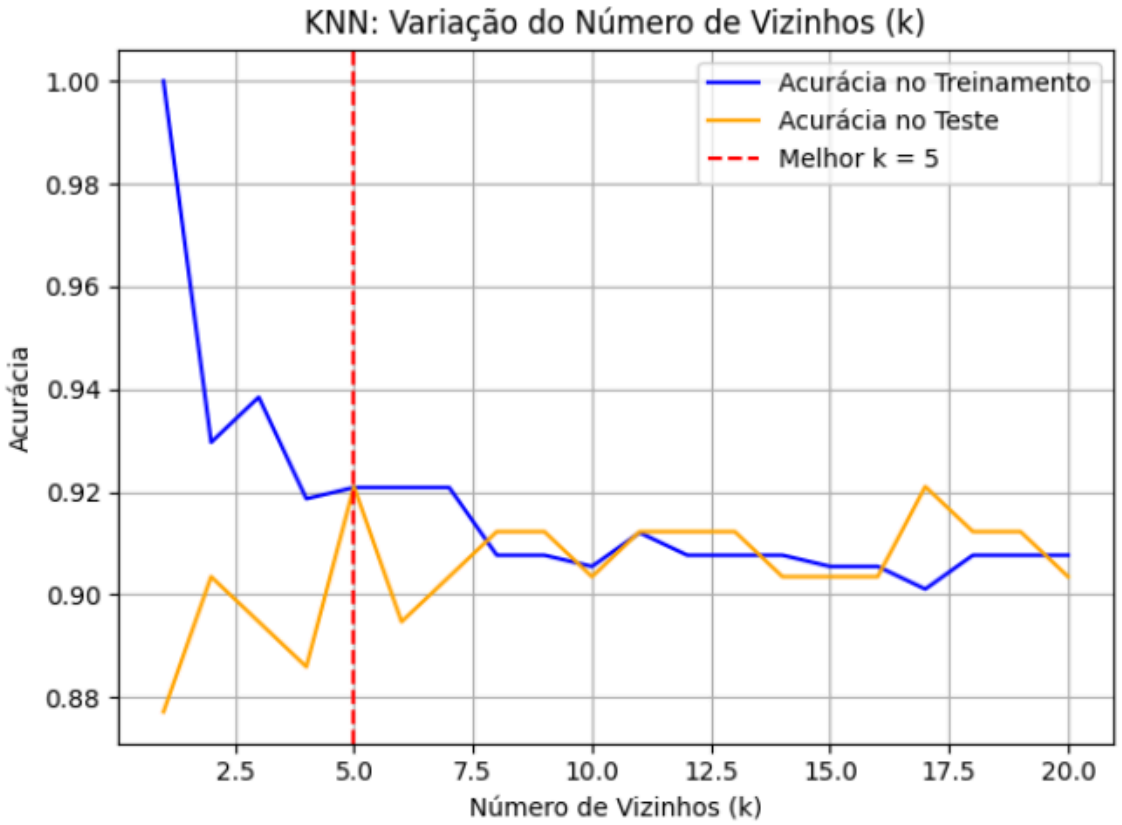
Começando pelo carregamento do dataset ajustado, *Cancer_Data_Adjusted*, em seguida dividi as características, e acrescentei um trecho de código para remover classes com menos de 2 instâncias, em seguida foi feita a separação dos conjuntos de teste e treino. Em seguida foi feita a normalização logarítmica e a normalização de média zero e variância unitária. Encerrando com a avaliação da normalização,

nesse caso observou um grande aumento na acurácia, o que antes era 24% agora chegou a cerca de 90%, abaixo é mostrado ambos os resultados deste código.



4-

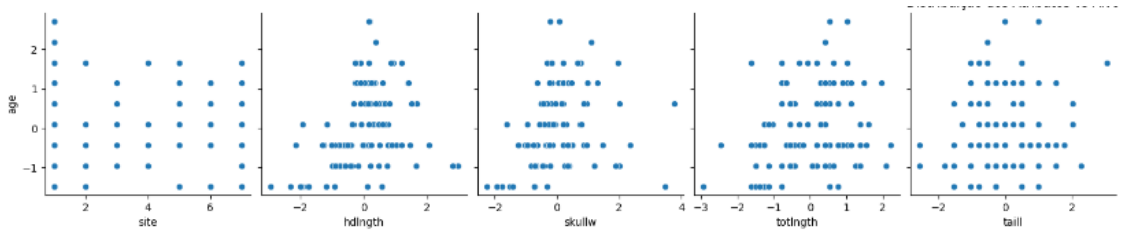
Nesta questão o objetivo é buscar a melhor parametrização do knn da questão anterior, novamente utilizando o datasset ajustado, Cancer_Data_Adjusted, dividindo os dados de treino e teste, e em seguida realizando a normalização StandarScaler(), seguindo das mesmas normatização do exemplo anterior. Logo em seguida é testado diferentes valores de K, no total foram testados 20 valores para k, do 1 ao 21, após realizar uma análise de cada acurácia exibindo o melhor K, que nesse caso foi 5, atingindo uma acurácia de 92.11%, veja o gráfico abaixo.



5-

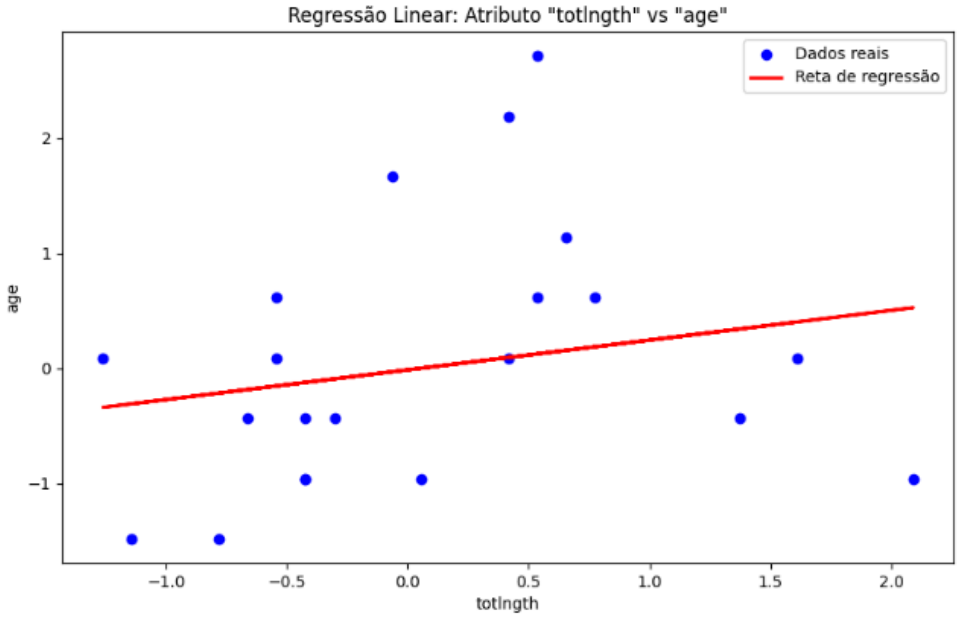
Nesta questão é realizado um pré-processamento e uma análise para identificar o atributo alvo e para realizar a regressão e o atributo mais relevante

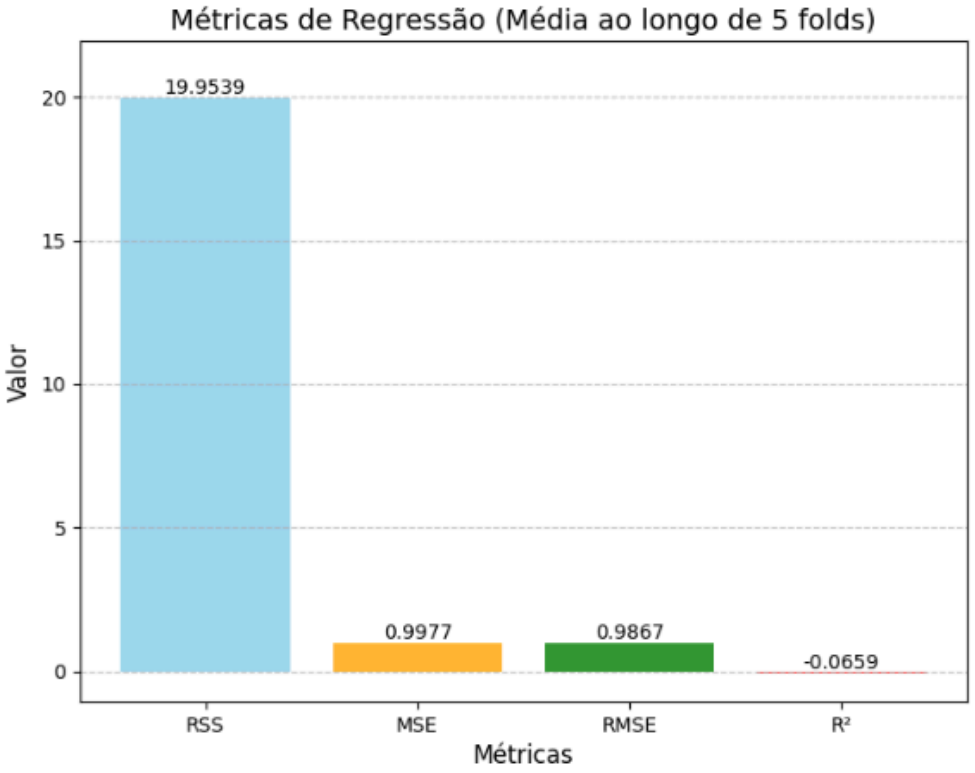
Os passos seguem os mesmos de melhor normalização, e o resultado para esse exemplo pode ser visto logo abaixo:



6-

Nesta questão são utilizados os atributos mais importantes da questão anterior, agora implementando uma regressão linear, abaixo temos o resultado, e um segundo gráfico mostrando os valores de cada valor pedido na questão: RSS, MSE, RMSE, e R^2





7-

Esta é a última questão, nela é utilizado o kfold e cross-validation para fazer uma regressão linear. Foi utilizado uma implementação manual do kfold e cross-validation. As métricas calculadas foram, RSS, MSE, RMSE e R_squared, os resultados foam os mesmo do gráfico anterior.

3. Conclusões

A conclusão final foi que algumas questões podem não terem sido resolvidas corretamente, seja por domínio nos conteúdos, ou pela amostragem de dados utilizados no qual foi comentado nas observações, mas foi um trabalho realmente interessante específico em muitos pontos, fazendo com que o aluno revise constantemente.

4. Próximos passos

Não tenho recomendações para o projeto, o fato dos datasets serem escolhidos pelo professor é interessante, pois não limitado os alunos a fazerem apenas o que dominam, trazendo uma análise dos próprio conjunto de dados.