UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Antonio Lucas Melo de Sousa

1. Introdução

Neste relatório, será analisado o dataset ***Cancer_Data.csv***, obtido da plataforma Kaggle, com o objetivo de explorar suas características e realizar diferentes etapas de processamento e agrupamento dos dados. Diferente da análise realizada na **AV1**, onde o foco estava na **classificação preditiva**, esta avaliação concentra-se em técnicas de **aprendizado não supervisionado**, como a segmentação dos dados utilizando **métodos de clusterização**.

Dataset de Classificação:Cancer_Data.csv

O dataset contém **569 instâncias e 32 colunas**, abrangendo medições estatísticas de tumores mamários, como ***radius_mean*****,** ***texture_mean*****,** ***perimeter_mean*****,** e ***area_mean***, além da variável alvo ***diagnosis***, que classifica os tumores como **M (maligno)** ou **B (benigno)**. O objetivo desta análise, será explorar a segmentação natural dos dados utilizando **métodos de agrupamento**, avaliando a coerência dos clusters formados e sua relação com a variável de diagnóstico real.

O objetivo principal é verificar se os padrões identificados nos dados permitem separar amostras de tumores malignos e benignos de forma não supervisionada, além de avaliar métricas como o **índice de silhueta** para definir o número ideal de clusters. Para isso serão aplicadas técnicas de **normalização dos dados**, remoção de colunas desnecessárias e visualização dos agrupamentos obtidos.

2. Observações

Durante a execução das atividades tive alguns problemas com NaN, quando realizava a normalização mas permanecia com NaN devido a coluna ***Unnamed*** que estava vazia, e a coluna ***Id*** que era irrelevante para a clusterização.

Na questão 4, os resultados não estavam satisfatórios na exibição do gráfico dos ***clusters***, pois havia intersecção entre os 2 centróides. Então foi aplicado o PCA para reduzir as dimensões para 2 e visualizar os ***clusters***.

3. Resultados e discussão

Nesta seção será apresentado os resultados de cada questão e uma discussão sobre a metodologia e fluxo de resolução.

Q1 - Na AV1 foi realizado uma análise do KNN para descobrir o melhor valor para **k** e também a forma mais adequada de determinar a métrica de distância. Na AV2 utilizando o mesmo ***dataset*** foi utilizado o ***GridSerachCV***, além do método de "Cotovelo" e "Silhueta" para determinar o número ideal de ***clusters*** e comparar os resultados com a variável alvo (***Diagnosis***) por meio de uma tabela cruzada (**crosstab**).

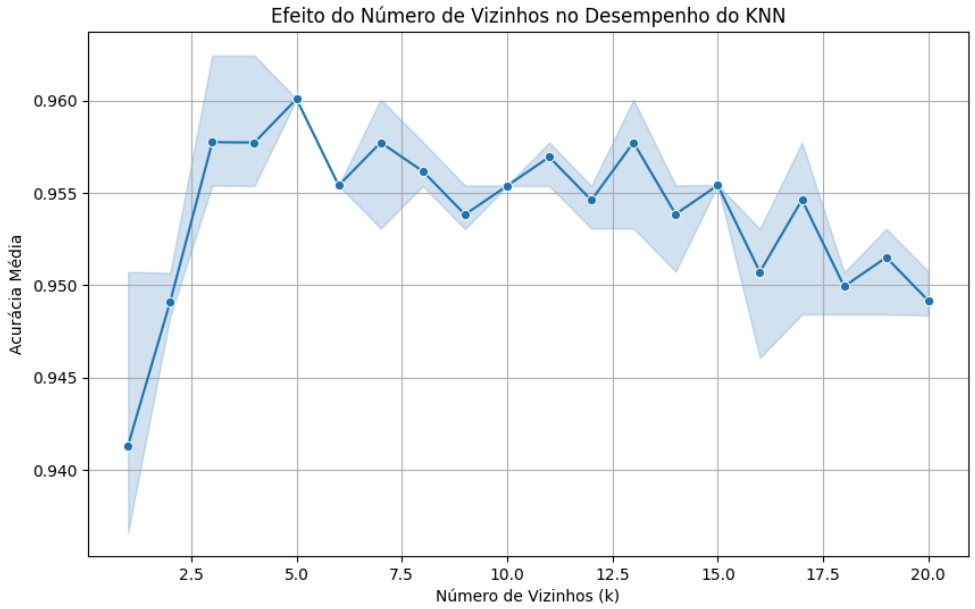
O fluxo de Resolução para esta questão começa no pré processamento dos dados, carregando o ***dataset*** e removendo as colunas irrelevantes ***Unnamed*** e ***Id***, além de converter a variável ***Diagnosis*** para valores números, 1 para malignos, e 0 para benignos. Os valores ausentes encontrados foram preenchidos com a média da respectiva coluna, além de ter sido realizada uma padronização dos dados com o ***StandardScaler.***

Em seguida por meio do **K-Means,** variando **k** de 2 a 10, é calculada o índice de "Silhueta" para avaliar a qualidade dos agrupamentos, e logo em seguida é utilizado o método de "Cotovelo" para observar a inércia e identificar um ponto de estabilização.

O resultado da aplicação do **K-Means** utilizando PCA para a visualização foram:

O número ideal de ***clusters*** foi **k = 3,** conforme indicado pelo índice "Silhueta", a tabela cruzada entre os ***clusters*** e os diagnósticos reais resultou em acurácia de 97.20%, o tamanho do ***test size*** está padronizado em 0.25, o recomendado para ***datasets*** nessa faixa de amostras é de 0.2 a 0.5 , então testei as melhores tamanhos um por um:

Cluster	precision	recall
0	0.98	0.98
1	0.96	0.96



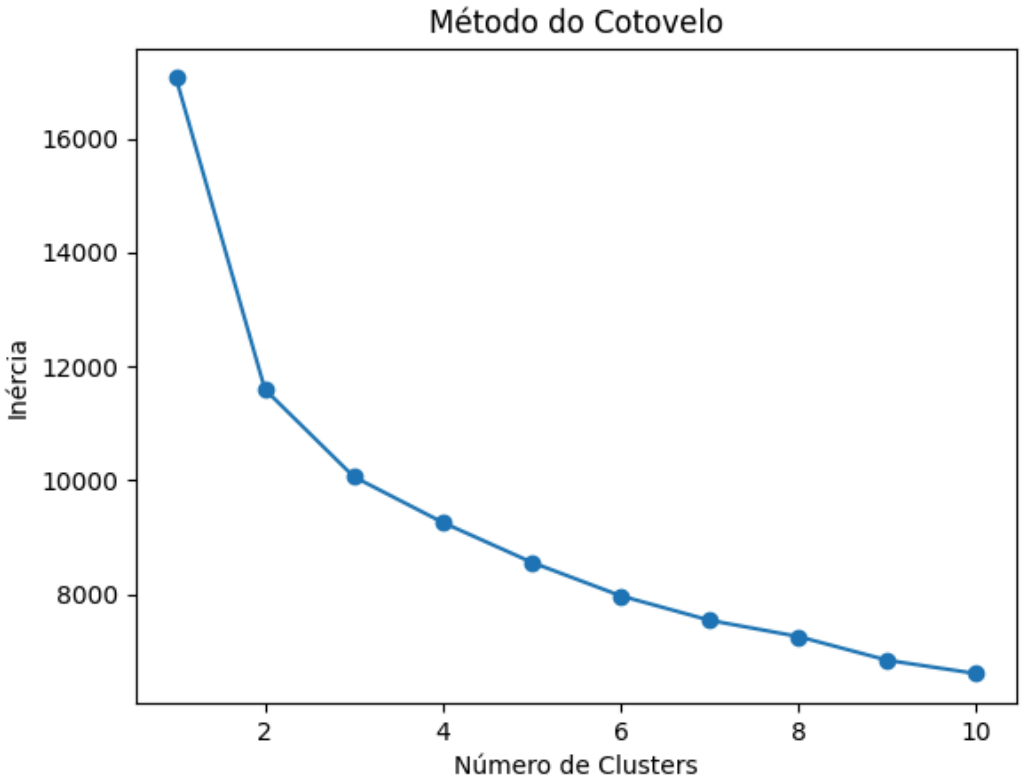
Q2 - Nesta questão não foi considerada a coluna alvo do ***dataset,*** a ***Diagnosis*** (a variável no qual se deseja explicar ou prever), então é feita uma análise para descobrir qual a quantidade ideal de ***clusters,*** usando o método cotovelo e da silhueta.

O fluxo da resolução começa no carregamento e processamento dos dados, as colunas ***Unnamed*** e ***Id***, que não possuíam relevância para o agrupamento foram removidas. Em seguida as colunas constantes foram descartadas para garantir que apenas variáveis com variação relevante para a clusterização fossem mantidas.

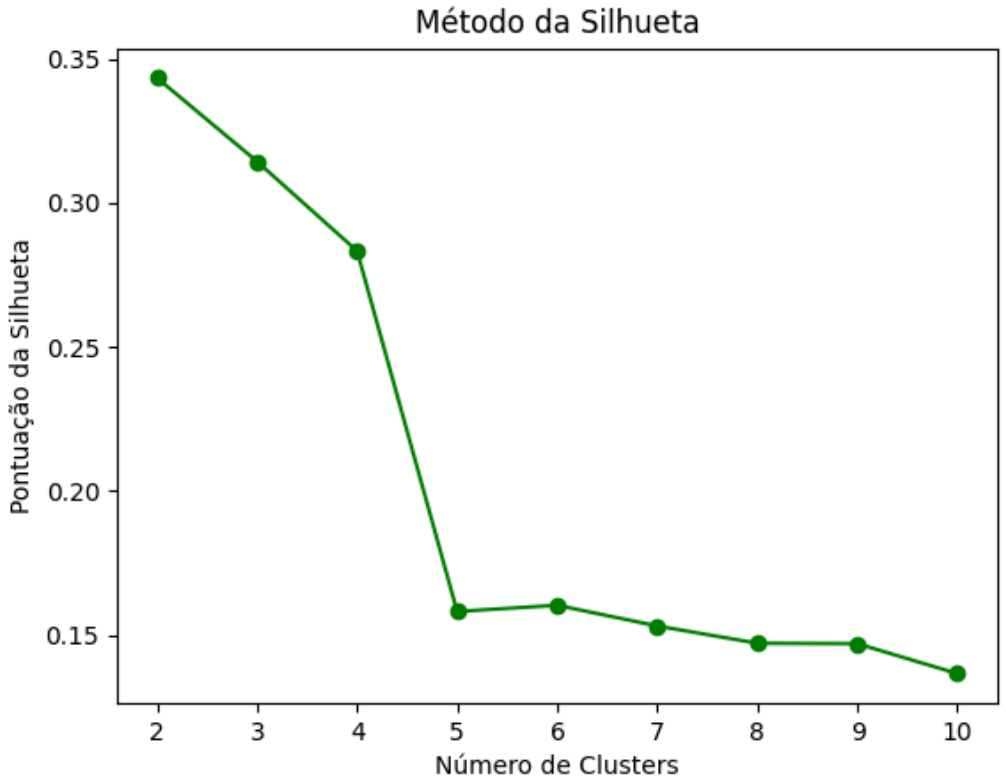
Em seguida é feita a normalização dos dados, garantindo que as variáveis estivessem na mesma escala, os dados foram normalizados utilizando ***StandardScaler***, uma vez que o ***K-Means*** é sensível à escala das variáveis, sendo importante para a clusterização.

Por último foi feito a aplicação dos métodos de Determinação de ***Clusters***, utilizando o método de "Cotovelo" e "Silhueta", foi realizada a análise do número ideal de ***Clusters***, sendo utilizado uma variação de 1 a 10 (valor utilizado em aula).

No Gráfico de "Cotovelo" mostra a inércia para valores de **k** entre **1** e **11.** A partir do **k=2**, a inércia começa a apresentar uma diminuição mais suave, indicando a escolha do número de ***Clusters.***



No Gráfico de “Silhueta” o gráfico apresenta as pontuações para diferentes valores de **k.** A maior pontuação foi obtida com 2 ***Clusters,*** reforçando a escolha do número ideal.



Com a análise realizada, concluímos que os dados podem ser segmentados em dois clusters, o que está alinhado com a ideia de que existem dois tipos principais de tumores no ***dataset: malignos*** *e* ***benignos.*** *As métricas assim indicam que a escolha do número de* ***Clusters*** foi adequada, mas podem haver variações em diferentes implementações, ou com métodos adicionais de validação.

Q3 - Nesta questão, o método de regressão ***Lasso*** foi utilizado para selecionar os dois atributos mais relevantes do ***dataset.*** Com esses atributos, a quantidade de ***Clusters*** foi recalculada utilizando os métodos do "Cotovelo" e "Silhueta", em busca de valores diferentes.

Primeiramente foi feito o carregamento e limpeza dos dados, o ***dataset*** foi carregado e as colunas irrelevantes ***Unnamed*** e ***Id*** foram removidas.

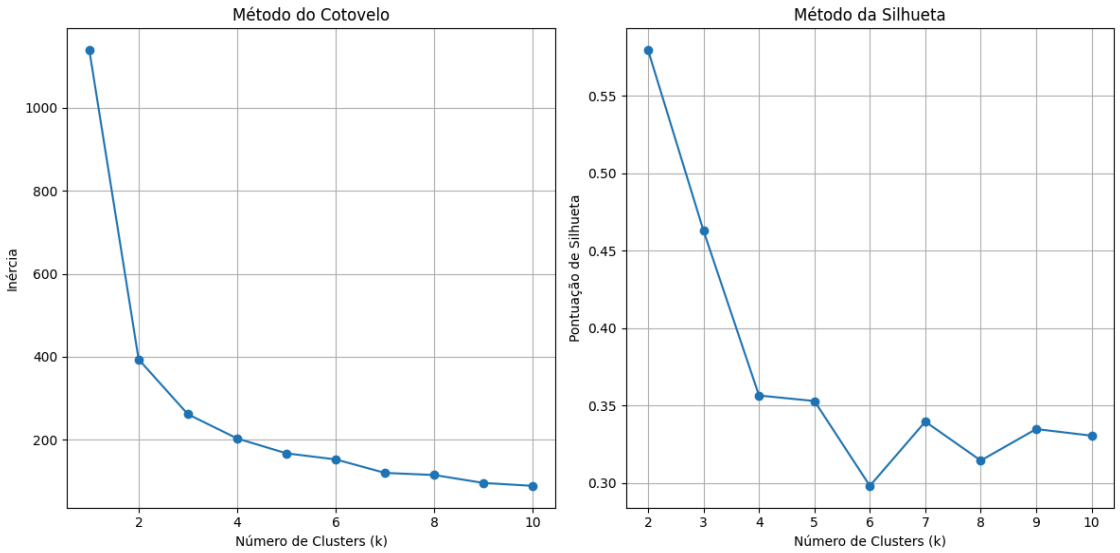
Em seguida foi feita a separação da variável alvo, neste caso a ***Diagnosis*** foi transformada em uma variável binária (1 para maligno e 0 para benigno) para ser usada no modelo de seleção de atributos, nas outras questões o mesmo foi feito em relação a essa coluna.

Após realizar a separação da variável alvo, é feito a normalização dos dados utilizando o ***StandardScaler***, para padronizar os dados e garantir que todas as variáveis tivessem a mesma escala.

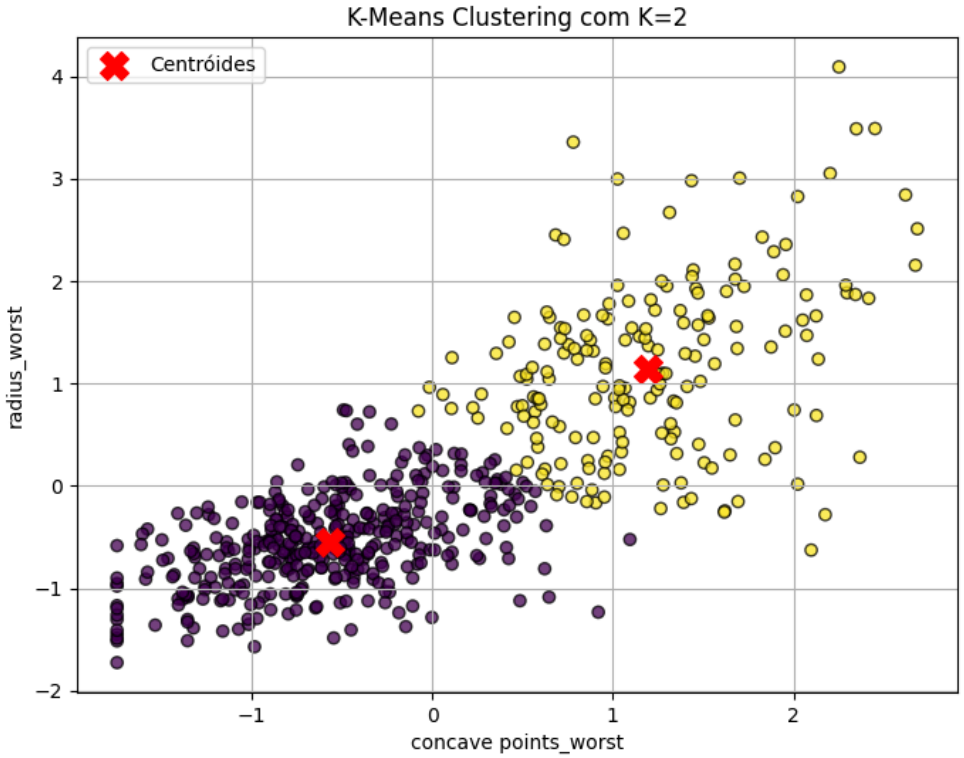
Logo em seguida é feita a aplicação do ***Lasso,*** o modelo foi treinado com um valor de regularização (alpha = 0.01), permitindo a seleção automática dos atributos mais relevantes.

Por último, os atributos com maiores coeficientes absolutos foram escolhidos.

Quando finalizado o processo de seleção dos atributos mais relevantes, é iniciado a determinação do número ideal de ***Clusters,*** primeiramente com o método "Cotovelo" que busca identificar o ponto em que a inércia começa a desacelerar. Tanto esse método quanto o de "Silhueta" seguem a lógica e propósito da questão anterior, veja o resultado de ambos no gráfico abaixo.



Após a execução dos métodos percebe-se que os valores para **k** permaneceram 2, o que indica que os melhores atributos já haviam sido utilizados, ou eram próximos, possuindo variações mínimas nos gráficos, com exceção do gráfico de "Silhueta" que mudou o comportamento, mas manteve o valor de **k**, verifique o gráfico anterior da questão 2. Abaixo é possível observar uma visualização melhor do agrupamento dos ***Clusters,*** com **k = 2.**



Q4 - Nessa questão foi utilizado o ***K-Means,*** para agrupar os dados do conjunto ***Cancer_Data.csv***, para determinar o número ideal de ***clusters*** aplicando o índice de "Silhueta", que mede a qualidade dos agrupamentos formados. Posteriormente, é comparado os ***clusters*** com a distribuição da variável alvo ***Diagnosis*** através de uma tabela cruzada (***crostab***).

Primeiramente é realizado o carregamento do ***dataset*** removendo logo em seguida as colunas irrelevantes como feito anteriormente, assim como o da coluna ***Diagnosis*** para valores numéricos.

Em seguida é preenchido os valores ausentes com a média das colunas correspondentes, além da normalização usando o ***StandardScaler*** para evitar distorções devido às diferenças de escala das variáveis.

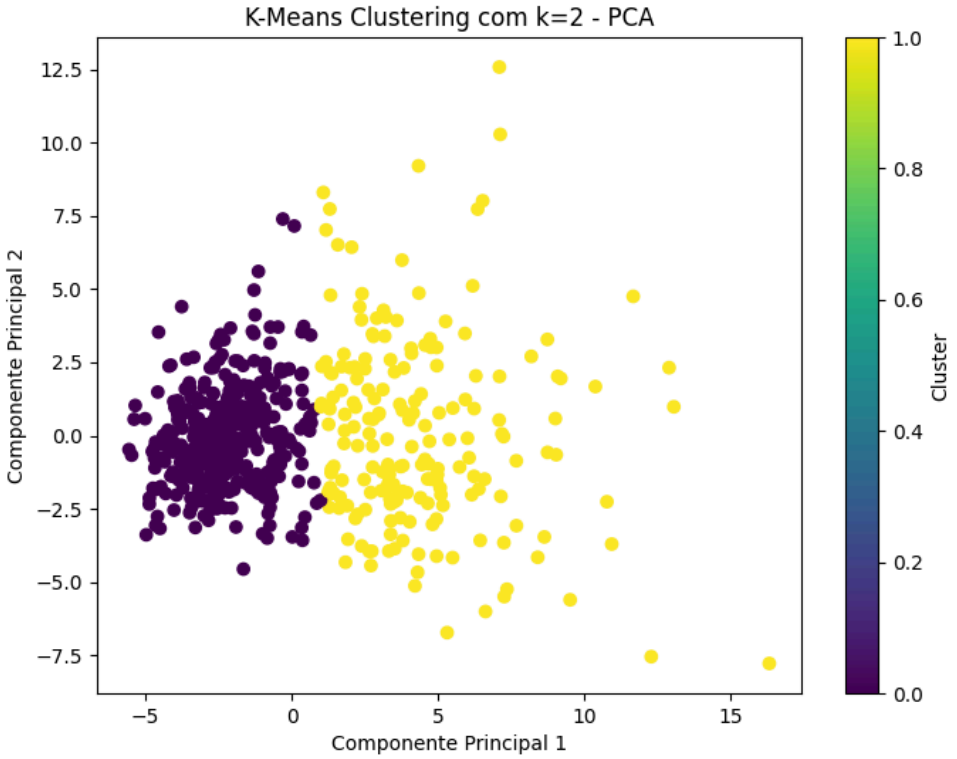
Em seguida é determinado o melhor **k** utilizando o **K-Means** variando o número de ***clusters*** entre 2 e 10, em seguida é avaliado tanto com o método "Cotovelo" quanto

"Silhueta". Com os melhores **k** obtidos é comparado os rótulos dos ***clusters*** com a variável ***Diagnosis*** por meio de uma **crosstab.**

O melhor valor para **k** foi 2, e a Matriz de confusão retornou:

Cluster	0	1
0	339	18
1	36	176

Vale ressaltar que foi necessário a utilização do PCA para ajustar a separação visual dos grupos, uma vez que estavam se sobrepondo, no final foi obtido uma separação relativamente boa, mas a distância do agrupamento poderia ser melhor, no gráfico abaixo é possível visualizar o resultado.



4. Conclusões

Ao final boa parte dos resultados foram satisfatórios, atingindo pontuações melhores que na AV1, houve uma melhora significativa, tanto na organização do código, desafios encontrados, quanto nos resultados. Apesar dos bons resultados uma questão é aberta, as questões foram resolvidas das melhores formas, ou corretamente? na grande maioria sim, mas na questão 1 por exemplo é possível abrir uma discussão, uma vez que a comparação com a AV1 se torna complicada devido a má implementação da mesma, o código e erros apresentados eram frequentes no processo de resolução. Mas no final foi possível analisar métodos diversos e compreendê-los através dos gráficos de visualização.

5. Próximos passos

Após o término do relatório e resolução das questões, seria interessante revisar os códigos com o intuito de otimizar o seu tamanho, e analisar melhor algumas saídas, e realizar o mesmo com a AV1.