UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Matheus Feitosa de Oliveira Rabelo

1. Introdução

O estudo deste dataset tem como objetivo realizar uma análise e aplicar técnicas de aprendizado de máquina para identificar padrões no dataset "train_ajustado.csv", que contém informações sobre dispositivos móveis para a clusterização. Esses dados apresentam diversas características técnicas dos celulares, como capacidade de processamento, memória RAM, peso, tamanho da bateria, entre outras variáveis relevantes para segmentação e classificação. Suas colunas são:

- battery_power
- clock_speed
- fc
- four_g
- int_memory
- mobile_wt
- n_cores
- pc
- px_height
- px_width
- ram
- sc_h
- sc_w
- three_g
- touch_screen
- price_range

A análise proposta visa avaliar a estrutura dos dados e explorar técnicas de agrupamento para identificar relações entre as especificações dos dispositivos e suas faixas de preço. Para isso, utilizaremos métodos como *K-Nearest Neighbors (KNN), K-Means* e técnicas de validação cruzada, permitindo verificar quais características impactam diretamente na classificação dos aparelhos. Além disso, serão testadas diferentes métricas para definição da quantidade ideal de clusters, utilizando os *métodos do Cotovelo e do Coeficiente de Silhueta*. Por fim, será realizada uma análise comparativa para verificar se os grupos formados pelos algoritmos de clusterização

correspondem às classes reais do dataset, permitindo avaliar a eficácia da abordagem não supervisionada na identificação de padrões no mercado de dispositivos móveis.

2. Observações

Como o dataset utilizado continua sendo o Mobile Price, durante as análises do dataset, ainda é possível observar que em um dataset chamado test.csv, não possui a coluna mais importante, price_range, então foi utilizado somente o dataset train_ajustado.csv, já que ele contém essa característica principal sem valor nulo e sem colunas irrelevantes. Na questão 1 foi utilizado o dataset sem a normalização com StandardScaler(), pois como o dataset era bem estruturado, sem valores nulos e inconsistência, aplicar ele diminui a acurácia em quase 30%, na questão 2, foi necessário um estudo mais aprofundado na melhor escolha do K, que será abordado melhor em Resultados e discussão.

3. Resultados e discussão

Nesta seção deve-se descrever como foram as resoluções de cada questão. Crie sessões indicando a questão e discuta a implementação e resultados obtidos nesta. Explique o fluxograma do processo de cada questão, indicando quais processamentos são realizados nos dados. Sempre que possível, faça gráficos, mostre imagens, diagramas de blocos para que sua solução seja a mais completa possível. Discuta sempre sobre os números obtidos em busca de motivos de erros e acerto.

Questão 1

Objetivo:

- Determinar a melhor parametrização do K-Nearest Neighbors (KNN) utilizando GridSearchCV, variando o valor de K e a métrica de distância (Euclidiana e Manhattan).

Fluxograma:

Carregamento do dataset → Divisão dos dados em treino e teste → Aplicação do GridSearchCV → Comparação dos resultados com os da AV1

Discussão:

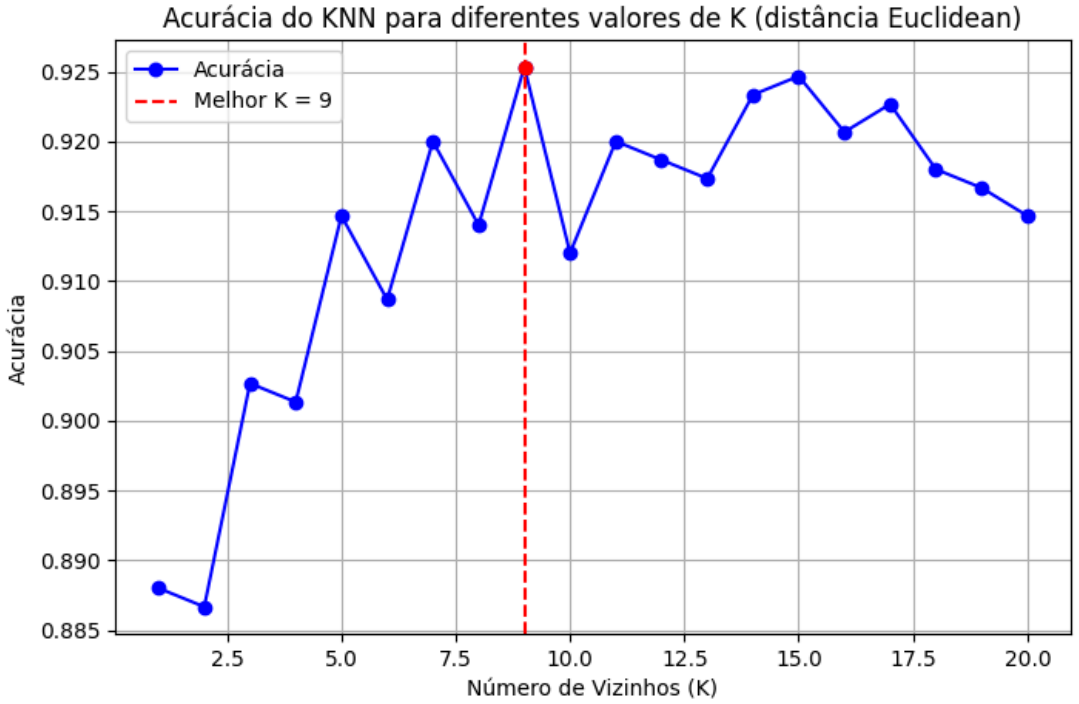
Os testes com o *GridSearchCV* indicaram que a melhor parametrização utilizou a métrica de distância *Euclidiana* e um valor de K = 9 . Diferentemente dos obtidos com a

AV1, onde a melhor distância foi a *Manhattan* e o melhor K foi 11. Apesar das diferenças, seus resultados de acurácias não foram tão diferentes, já que o *GridSearchCV* trouxe uma acurácia de 92.53 na validação cruzada e 95% no conjunto de teste quando comparada com a da AV1 que foi somente com o KNN uma acurácia de 95.4%, podendo afirmar que o *GridSearchCV* trouxe um diminuimento de acurácia de 0.4%. O que nos leva a tentar levar como estudo futuros, o porquê ao utilizar o *GridSearchCV* ele diminui a acurácia, sendo que ele é utilizado para encontrar o melhor hiperparâmetros de um modelo de aprendizado de máquina, testando várias combinações possíveis.

Resultados:

AV1

GridSearchCV



Acurácia no conjunto de validação cruzada: 92.53%

Acurácia no conjunto de teste: 95.00%

Questão 2

Objetivo:

- Determinar a quantidade ideal de clusters no dataset sem considerar a coluna alvo, utilizando os métodos do Cotovelo e Coeficiente de Silhueta.

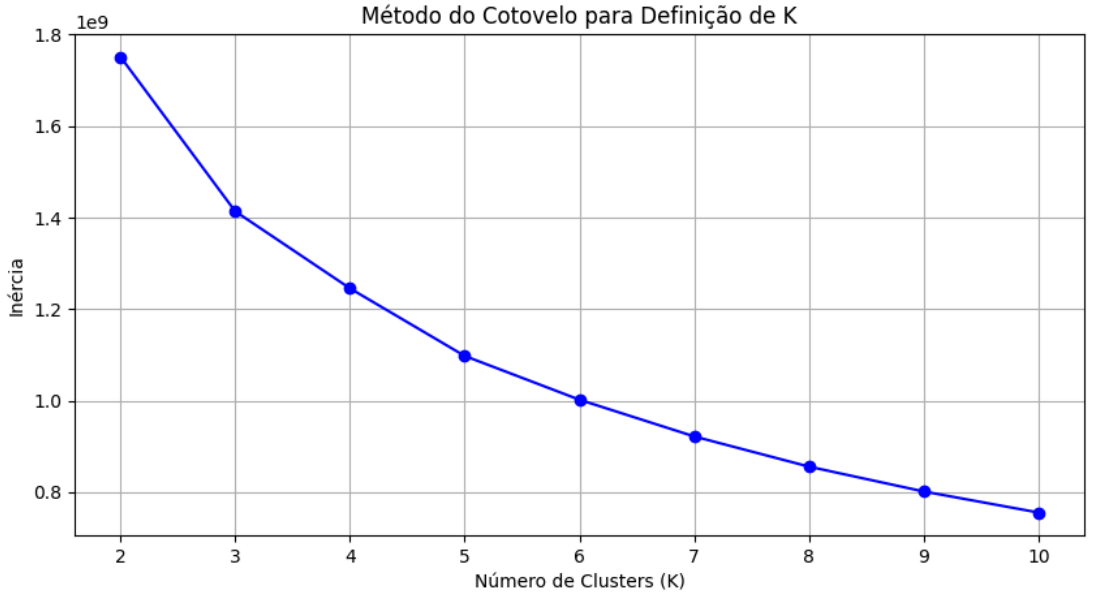
Fluxograma:

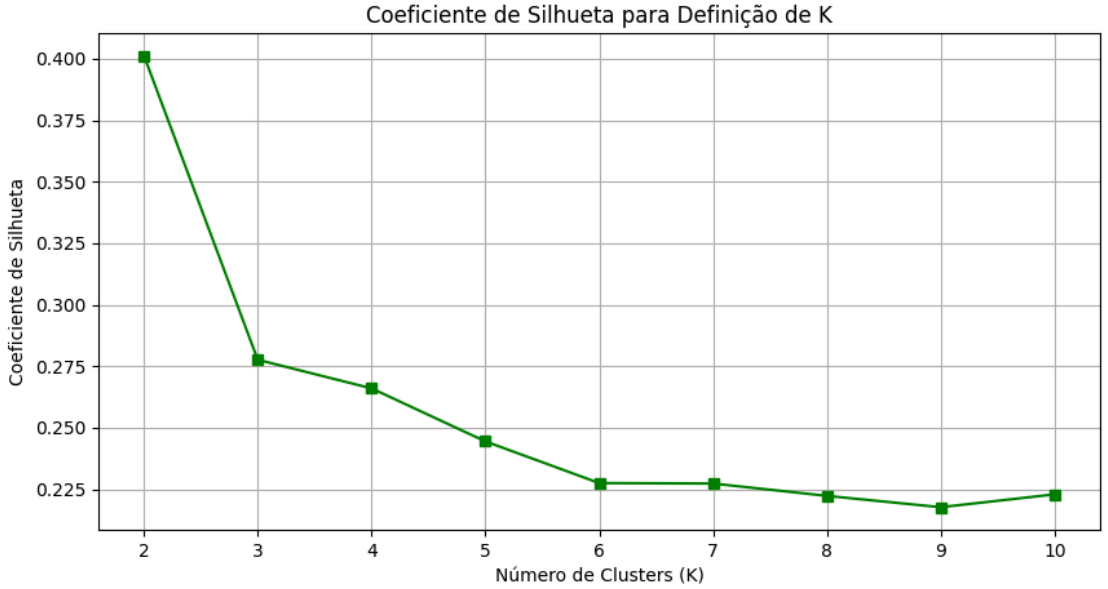
Remoção da variável alvo → Aplicação do Método do Cotovelo e Coeficiente de Silhueta → Comparação dos valores ideais de K entre os dois métodos

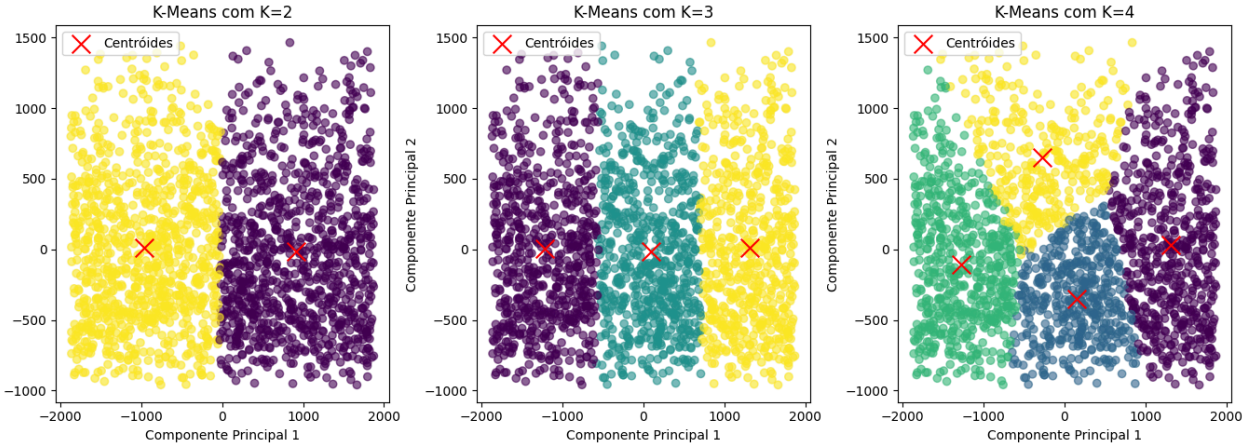
Discussão:

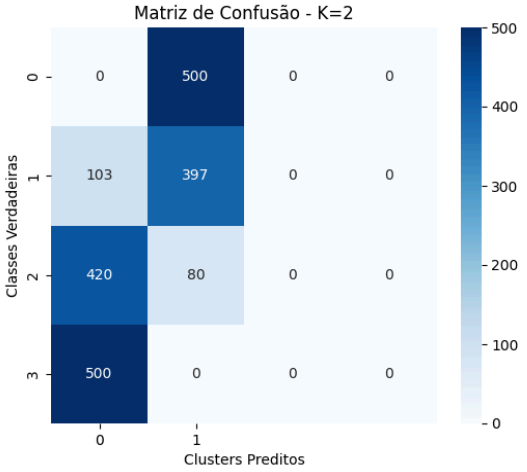
O *Método do Cotovelo* indicou que K = 3 ou 4 (imagem 1) seria ideal, pois a redução da inércia estabilizou mais ou menos nesses pontos. Já o *Coeficiente de Silhueta* apontou para K = 2 (imagem 2), onde a separação dos clusters foi melhor definida. Como a leitura do *método de cotovelo* estava difícil de se interpretar e o de *silhueta* deu valores diferente, foi optado a se aprofundar mais nessa análise ao invés de aplicar como estudo futuro, então foi analisado com o *K-Means* os cluster com a quantidade de K/centróides entre 2 e 4 para compreender melhor essa divisão (Imagem 3), através da geração dos gráficos, foi analisado que a divisão por 2 e 3 clusters era mais adequado por possuir uma boa separação, sendo clara e sem dados misturados, o que já podemos ver que está querendo surgir com 4 clusters, mas ainda temos que escolher entre 2 ou 3, então para aprofundar melhor a análise após isso, foi feito uma matriz de confusão para poder ter mais visualização sobre a definição do melhor K (Imagem 4, 5 e 6), e por fim um *classification report*. Após termos agora todas essas análises, e observar suas acurácias e as divisões da matriz de confusão, é possível observar que K = 3 é o mais ideal, porém ele ser 2 não seria uma péssima escolha em circunstâncias diferentes.

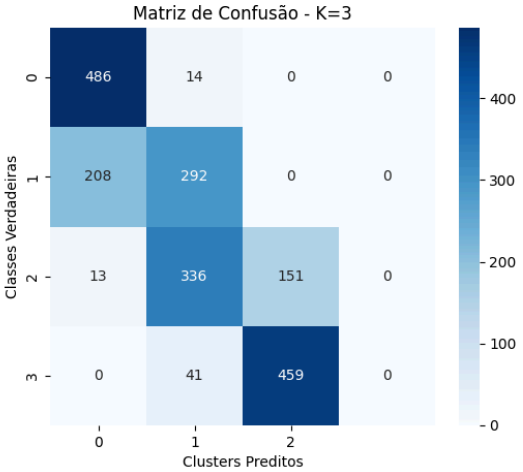
Resultados:

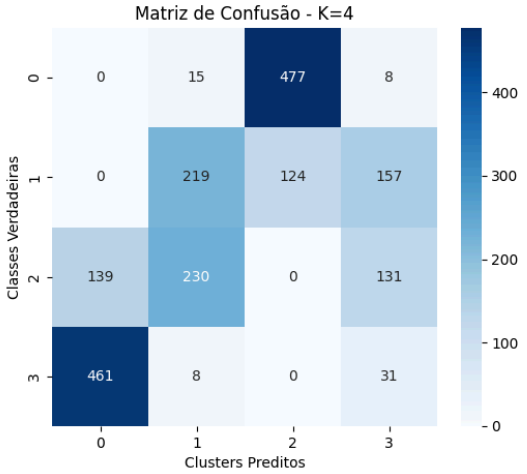



Coeficiente de Silhueta para Definição de K
Coeficiente de Silhueta
0.400
0.375
0.350
0.325
0.300
0.275
0.250
0.225
2
3
4
5
6
7
8
9
10
Número de Clusters (K)


K-Means com K=2
K-Means com K=3
K-Means com K=4
Centróides
Componente Principal 1
Componente Principal 2







Classification Report: K=2

	precision	recall	f1-score	Support
0	0.00	0.00	0.00	500
1	0.41	0.79	0.54	500
2	0.00	0.00	0.00	500
3	0.00	0.00	0.00	500
Acccuracy			0.20	2000
Macro avg	0.10	0.20	0.13	2000

weighted avg	0.10	0.20	0.13	2000

Classification Report: K=3

	precision	recall	f1-score	Support
0	0.69	0.97	0.81	500
1	0.43	0.58	0.49	500
2	0.25	0.30	0.27	500
3	0.00	0.00	0.00	500
Acccuracy			0.46	2000
Macro avg	0.34	0.46	0.39	2000
weighted avg	0.34	0.46	0.39	2000

Classification Report: K=4

	precision	recall	f1-score	Support
0	0.00	0.00	0.00	500
1	0.46	0.44	0.45	500
2	0.00	0.00	0.00	500
3	0.09	0.06	0.07	500
Acccuracy			0.12	2000
Macro avg	0.14	0.12	0.13	2000
weighted avg	0.14	0.12	0.13	2000

Questão 3

Objetivo:

- Selecionar os dois atributos mais relevantes do dataset utilizando *Lasso Regression* e refazer a análise de clusterização para verificar se a quantidade de clusters mudou.

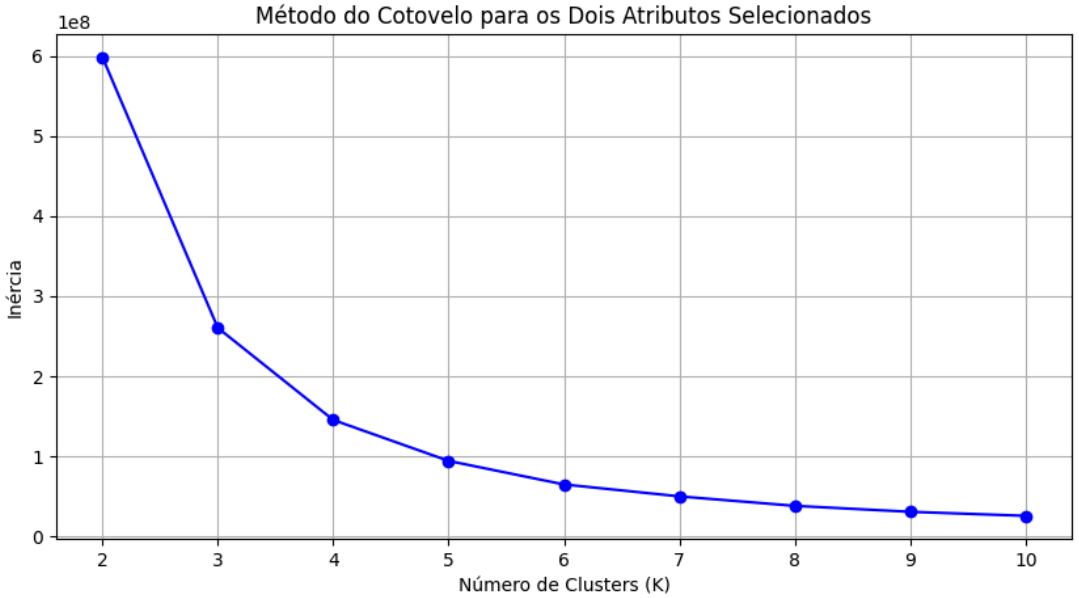
Fluxograma:

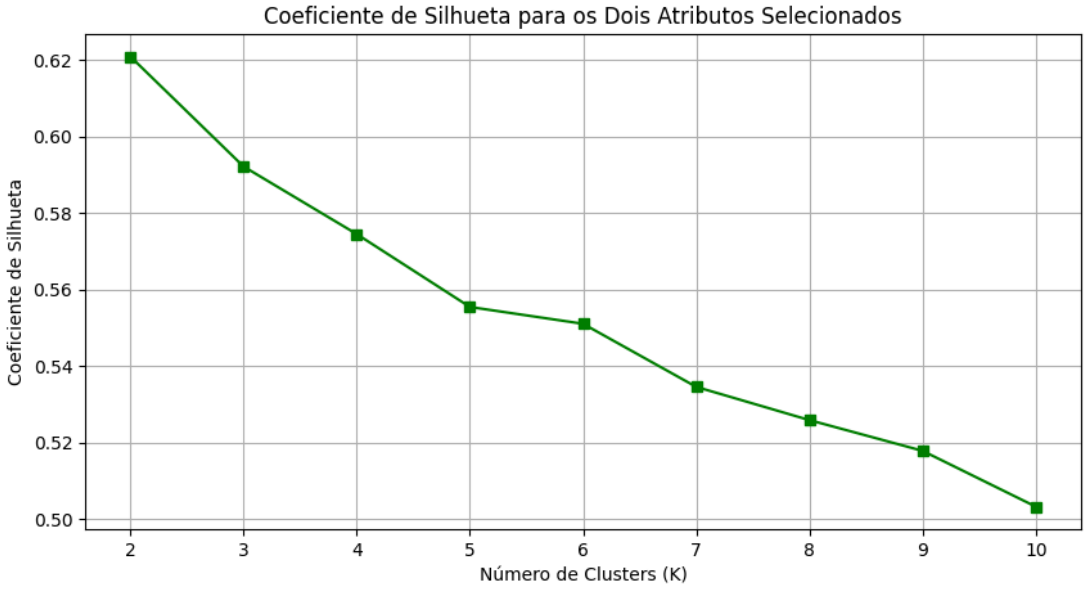
Aplicação do Lasso Regression para seleção de atributos → Redução do dataset para os dois atributos mais relevantes → Reaplicação dos métodos do Cotovelo e Silhueta → Comparação dos novos valores de K com os anteriores

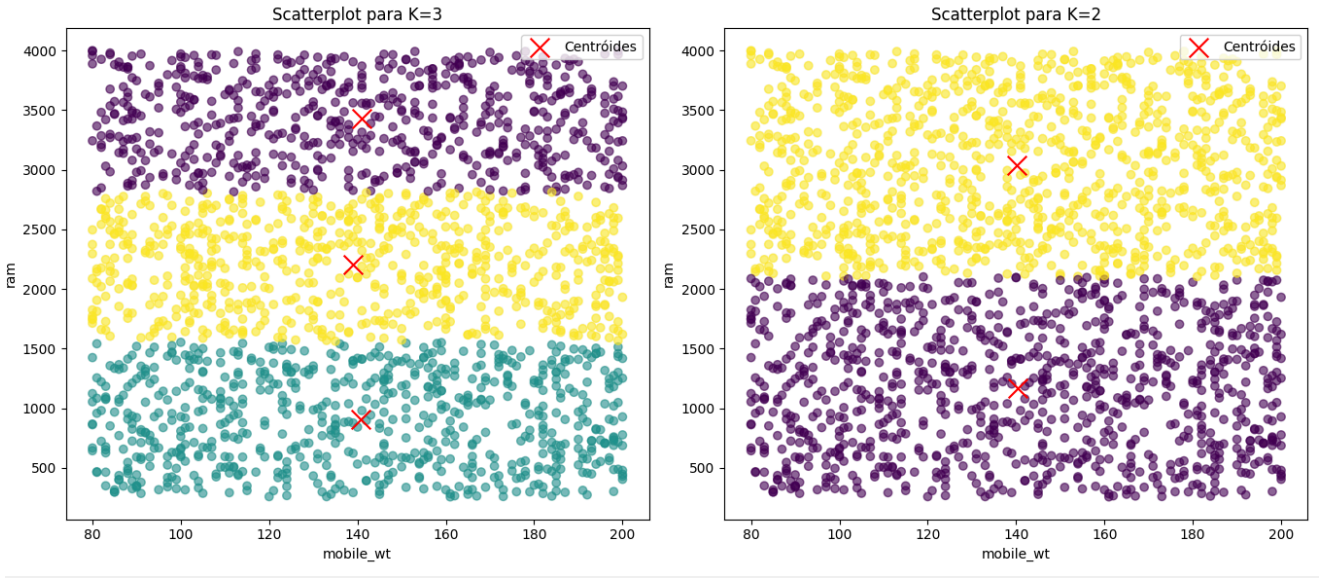
Discussão:

Após a seleção de atributos utilizando o *Lasso*, os dois mais relevantes foram *mobile_wt e ram*. A nova análise de clusterização indicou que o valor de K não mudou de certa forma para o *método do cotovelo*, mas podemos verificar que o 4 já não é mais uma opção, e sim somente o 3 agora, e no de *Silhueta* permanece sendo o K=2, sugerindo que a segmentação do dataset pode ser influenciada pelos atributos menos relevantes. Como dito no enunciado, por conter ainda N diferentes, foi feito um *Scatterplot* dos 2 K, para observar a divisão dos clusters, onde é possível observar que ambos os K (2 ou 3) tem boas divisões de clusters, assim como vimos na questão anterior, deixando aberto qual seja o melhor dependendo da situação, já que o K=2 tem a separação entre os clusters é mais clara e natural, porém pode ser simples demais, agrupando pontos que poderiam estar separados, e K=3 tem os centróides bem posicionados para representar cada cluster, porém pode estar segmentando um grupo extra desnecessário, já que vemos que o K=2 tem uma separação aceitável.

Resultados:


Método do Cotovelo para os Dois Atributos Selecionados
1e8
Inércia
Número de Clusters (K)


Coeficiente de Silhueta para os Dois Atributos Selecionados
Coeficiente de Silhueta
Número de Clusters (K)



Questão 4

Objetivo:

- Gerar uma crosstab para comparar a distribuição dos clusters formados pelo K-Means com as classes reais e avaliar se a clusterização corresponde às faixas de preço.

Fluxograma:

Aplicação do K-Means com o K definido pela Silhueta → Criação da crosstab entre clusters e classes reais → Análise da matriz de confusão para avaliar a correspondência entre clusters e classes reais

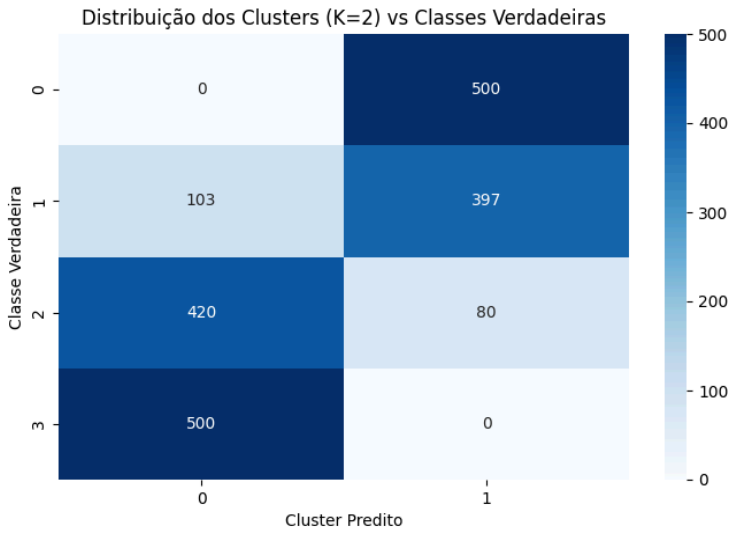
Discussão:

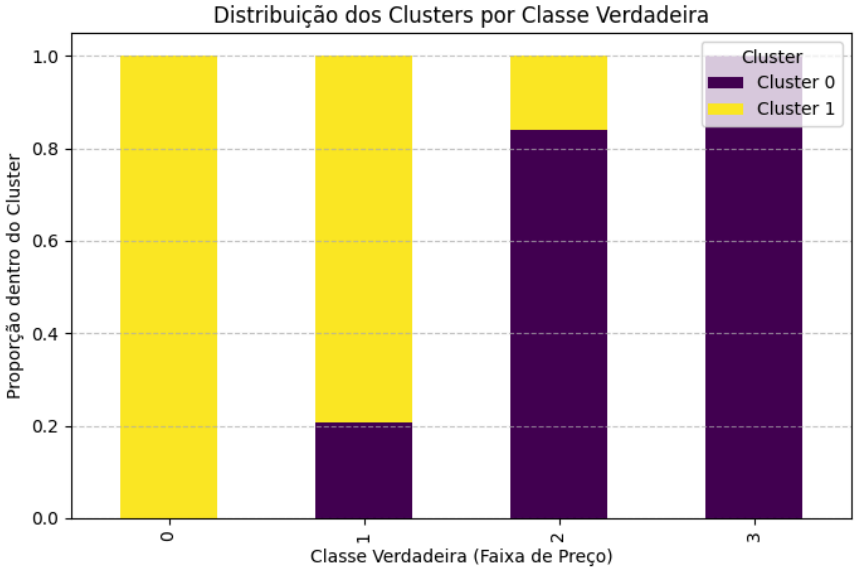
Como o enunciado pede que usemos o K obtido com o *Coeficiente de Silhueta,* utilizamos aqui K=2, podendo assim dizer que a matriz crosstab indicou que os clusters formados têm uma boa correspondência com as faixas de preço, pois ao observarmos a imagem 1 ou 2, podemos ver que apesar de na classe 1 ele possuir uma pequena parte do cluster 0, em consideração geral, ele conseguiu separar bem as classes 1 e 2 para o

cluster 0 e as classes 2 e 3 para o cluster 1. Isso sugere que o K-Means juntamente com o *Coeficiente de Silhueta* foi eficiente, e ajudou na decisão da quantidade de divisões/clusters.

Resultados:

	0	1
0 (Baixo custo)	0	500
1 (Médio custo)	103	397
2 (Alto custo)	420	80
3 (Premium)	500	0





4. Conclusões

Os estudos realizados demonstraram que as técnicas de aprendizado de máquina aplicadas ao dataset de dispositivos móveis permitiram uma organização coerente dos dados. A análise com o GridSearchCV mostrou que a escolha dos hiperparâmetros influencia a acurácia final, e a comparação entre os métodos de clusterização revelou que K=2 e K=3 foram as opções mais viáveis. Além disso, a abordagem com Lasso Regression indicou que apenas dois atributos já eram suficientes para manter a segmentação eficiente, reforçando a importância da seleção de variáveis no processo.

Embora o modelo tenha conseguido separar bem os grupos, especialmente ao comparar os clusters com as classes reais, alguns controversos foram observados, como a dificuldade na escolha do número ideal de clusters e a ligeira queda na acurácia ao utilizar GridSearchCV. No entanto, a metodologia aplicada permitiu uma análise aprofundada na validação dos resultados.

5. Próximos passos

Explorar técnicas alternativas como RandomizedSearchCV, para comparar a eficiência com o GridSearchCV. assim procurando entender também melhor o caso

da baixa acurácia com o uso de GridSearchCV e investigar a influência dos atributos selecionados pelo Lasso.