UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Robert Michael de Ávila Barreira

1. Introdução

Este relatório apresenta a análise de um conjunto de dados relacionado às variantes tintas do vinho "Vinho Verde" português. O objetivo principal é construir modelos preditivos capazes de classificar a qualidade do vinho com base em características físico-químicas.

O dataset contém 12 variáveis, sendo 11 delas relacionadas a testes laboratoriais, como acidez, pH, teor alcoólico e sulfatos, enquanto a variável de saída representa a qualidade do vinho, avaliada em uma escala de 0 a 10.

A principal complexidade deste problema reside no fato de que os dados são desbalanceados, com maior concentração em vinhos de qualidade mediana, além do número relativamente reduzido de amostras. Assim, a modelagem requer técnicas adequadas de pré-processamento, escolha de algoritmos e ajuste de hiperparâmetros para melhorar a performance dos modelos preditivos.

Além da construção dos modelos de classificação, este estudo também explora diferentes métricas de avaliação para comparar o desempenho dos algoritmos, buscando alcançar uma predição eficiente da qualidade do vinho em comparação com o relatório anteriormente apresentado.

2. Observações

Durante a realização deste estudo, foram identificados alguns pontos relevantes:

- O dataset apresentou um desbalanceamento significativo nas classes de qualidade, com maior concentração em vinhos de qualidade 5 e 6
- A redução de dimensionalidade através do método Lasso permitiu uma análise mais clara e eficiente dos dados.

- Os resultados obtidos com apenas duas variáveis (teor alcoólico e acidez volátil) foram surpreendentemente robustos
- A utilização do GridSearchCV permitiu uma otimização mais precisa dos parâmetros do modelo KNN em comparação com a análise anterior
- O processo de clusterização revelou padrões interessantes na relação entre características físico-químicas e qualidade dos vinhos"

3. Resultados e discussão

Neste tópico, são apresentados os resultados obtidos a partir dos modelos preditivos desenvolvidos para classificar a qualidade do vinho com base nas variáveis físico-químicas fornecidas no dataset. Inicialmente, os dados passaram por uma análise e pré-processamento cuidadosos, levando em consideração o desbalanceamento das classes e a necessidade de ajustes nos hiperparâmetros dos modelos. Foram respondidas 4 questões sobre o dataset:

Questão 1 - **Realize uma análise similar à feita na AV1, mas utilizando o GridSearchCV da scikit-learn para encontrar a melhor configuração de parâmetros para o processo de classificação. Com isso, obtenha os dados de acertos e compare os resultados obtidos com os da AV1.**

No experimento realizado anteriormente com o modelo KNN, foram avaliados diferentes valores de K (número de vizinhos) e métricas de distância (Euclidiana, Manhattan e Chebyshev) para determinar os melhores parâmetros e a acurácia do modelo.

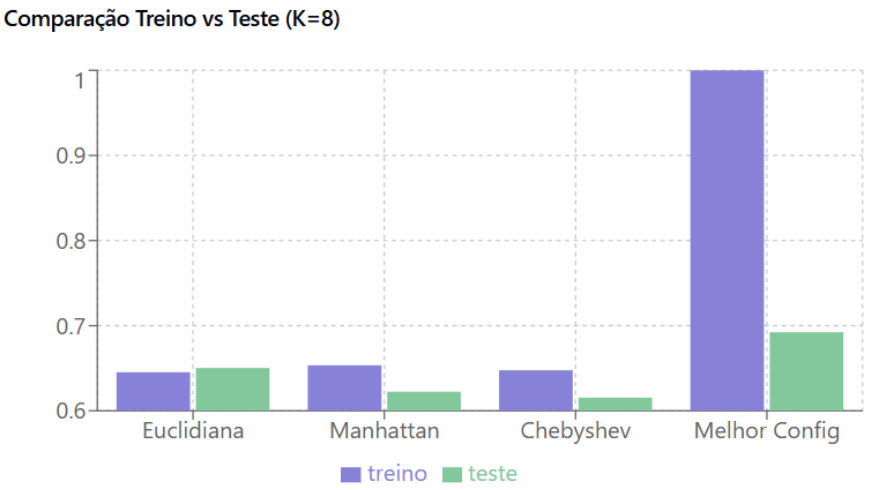
Resultados do Novo Experimento (K=8)

Distância Euclidiana:

- Acurácia de treino: **0.6453**
- Acurácia de teste: **0.6503**

Distância Manhattan:

- Acurácia de treino: **0.6534**



Melhor Configuração Encontrada

- **Parâmetros:**
  - K=8
  - Métrica: Distância Euclidiana
  - Pesos: Baseados na distância
- **Desempenho:**
  - Acurácia no conjunto de treino: **1.0000**
  - Acurácia no conjunto de teste: **0.6923**

Comparação com Experimentos Anteriores

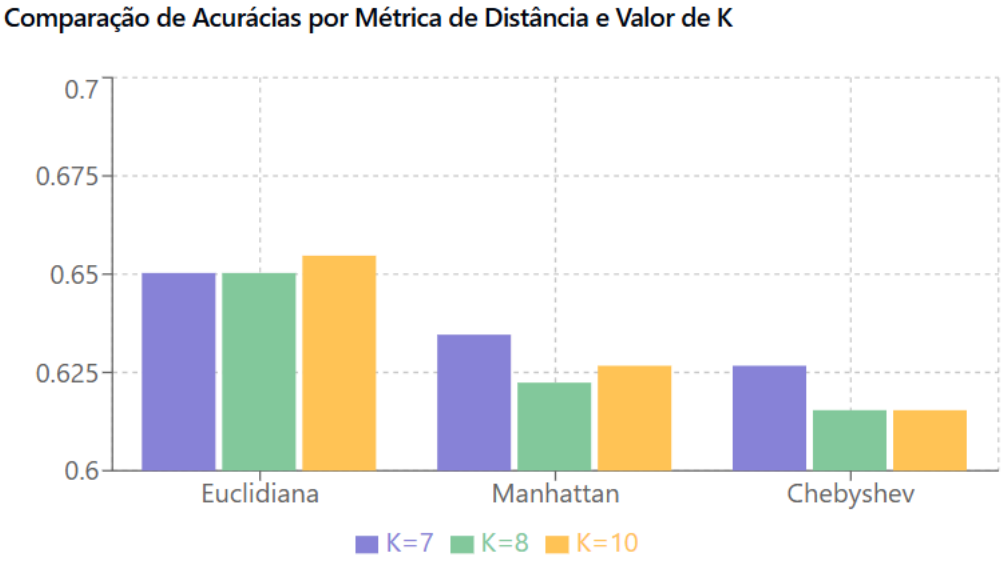
Resultados para K=7

- Distância Euclidiana: Acurácia de **0.6503**
- Distância Manhattan: Acurácia de **0.6346**
- Distância Chebyshev: Acurácia de **0.6267**

Resultados para K=10

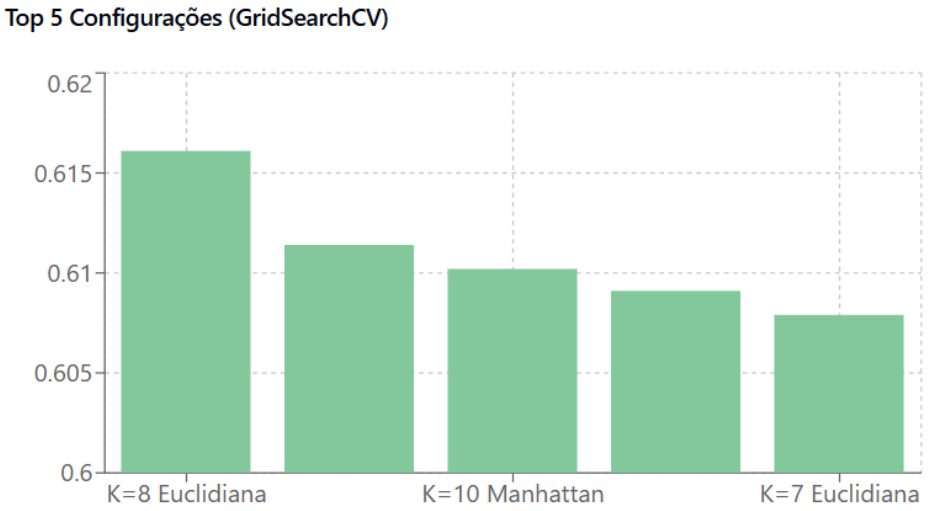
- Distância Euclidiana: Acurácia de **0.6547**

- Distância Manhattan: Acurácia de **0.6267**
- Distância Chebyshev: Acurácia de **0.6154**



Top 5 Melhores Configurações (Cross-Validation):

- K=8, Euclidiana, pesos por distância (CV Score: **0.6161**)
- K=10, Euclidiana, pesos por distância (CV Score: **0.6114**)
- K=10, Manhattan, pesos por distância (CV Score: **0.6102**)
- K=9, Euclidiana, pesos por distância (CV Score: **0.6091**)
- K=7, Euclidiana, pesos por distância (CV Score: **0.6079**)



Análise Comparativa dos Resultados

Desempenho das Métricas de Distância:

- **Distância Euclidiana:** Manteve consistência nos resultados, com acurácia de 0.6503 para K=8, próxima à observada com K=10 (0.6547).
- **Distância Manhattan:** Apresentou melhor desempenho com K=7 (0.6346), tendo performance inferior com K=8 (0.6224).
- **Distância Chebyshev:** Mostrou estabilidade entre K=8 e K=10 (0.6154), com leve superioridade para K=7 (0.6267).

Um ponto importante a destacar é a diferença significativa entre as acurácias de treino (1.0000) e teste (0.6923) na melhor configuração, indicando possível **overfitting** do modelo.

Podemos concluir que:

- A configuração com K=8, distância Euclidiana e pesos baseados na distância alcançou a melhor performance geral (0.6923 no teste).
- Os resultados do cross-validation confirmam a superioridade da distância Euclidiana, que aparece em 4 das 5 melhores configurações.
- Valores de K entre 8 e 10 tendem a proporcionar melhores resultados, com K=8 liderando as performances.
- Há evidência de overfitting que pode requerer técnicas adicionais de regularização ou ajuste de parâmetros em experimentos futuros.

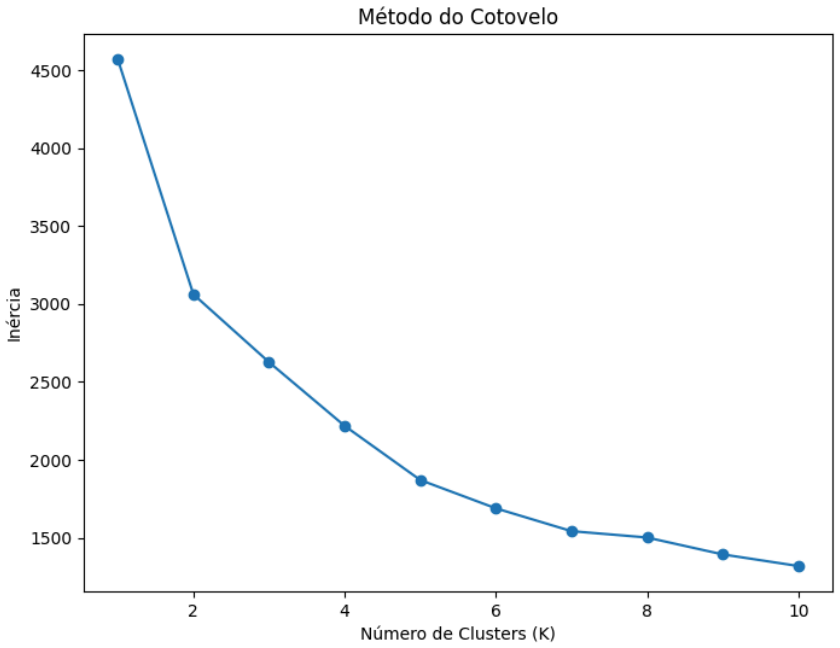
Algumas recomendações:

- Investigar técnicas para reduzir o overfitting observado
- Explorar valores de K entre 7 e 10 com maior granularidade

Questão 2 - Sem considerar a coluna alvo, realize uma análise para determinar o número ideal de clusters utilizando os métodos do cotovelo e da silhueta.

Para determinar a quantidade ideal de clusters em um conjunto de dados, o **método do cotovelo** (Elbow Method) e o **método da silhueta** (Silhouette Method) são duas abordagens amplamente utilizadas.

O **método do cotovelo** baseia-se na análise do valor da soma das distâncias quadradas dentro de cada cluster (inércia), em função do número de clusters. À medida que o número de clusters aumenta, a inércia diminui. O ponto em que a diminuição da inércia começa a desacelerar e formar uma "curva de cotovelo" é onde se encontra a quantidade ideal de clusters.



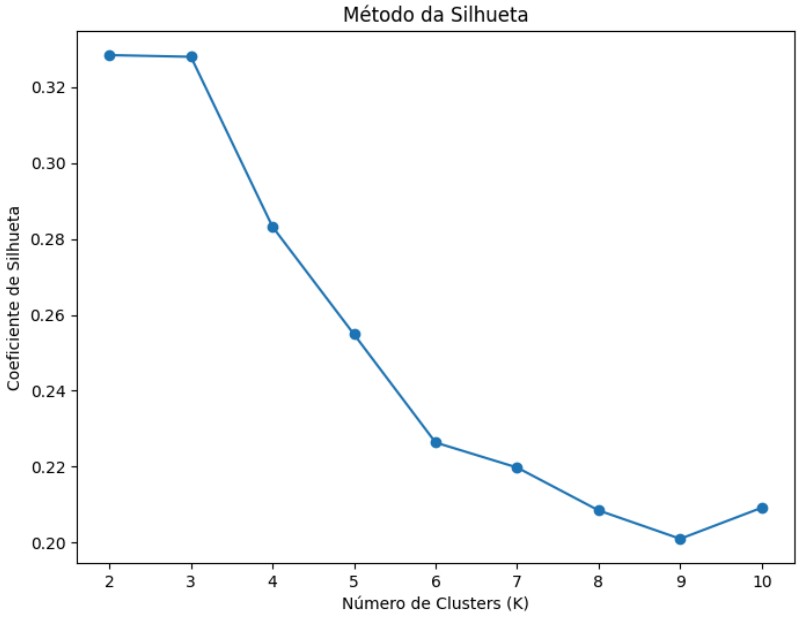
O ponto em que a inércia (desorganização dos dados) começa a diminuir de forma mais suave é onde encontra-se o número ideal de clusters. Podemos observar que:

- Há uma queda acentuada de **K=1** até **K=4**
- Após **K=4**, a curva começa a suavizar
- A partir de **K=6**, a redução da inércia se torna menos significativa
- A curva forma um "cotovelo" próximo a **K=4**

De acordo com os dados do gráfico **K=4** seria uma boa escolha, pois é onde ocorre a inflexão mais significativa da curva.

Já o método da silhueta avalia a qualidade da atribuição dos pontos aos clusters, levando em conta a coesão dentro do cluster e a separação entre clusters. O valor da silhueta varia de -1 a 1:

- Valor próximo de 1 indica que os pontos estão bem agrupados e afastados dos outros clusters.
- Valor próximo de -1 sugere que os pontos foram atribuídos ao cluster errado.
- Valor próximo de 0 significa que os clusters estão sobrepondo-se ou mal definidos.



O número de clusters com o maior valor de silhueta (maior proximidade dentro do grupo e maior distância entre os grupos) é o melhor. Podemos analisar através do gráfico que:

- O maior valor do coeficiente está em **K=2** e **K=3** (aproximadamente 0.33)
- Há uma queda significativa após **K=3**

- O valor continua diminuindo gradualmente até **K=9**
- Existe uma pequena elevação em **K=10**

Com base nessa análise do gráfico vemos que **K=2** ou **K=3** seria ideal, pois apresenta o maior coeficiente de silhueta.

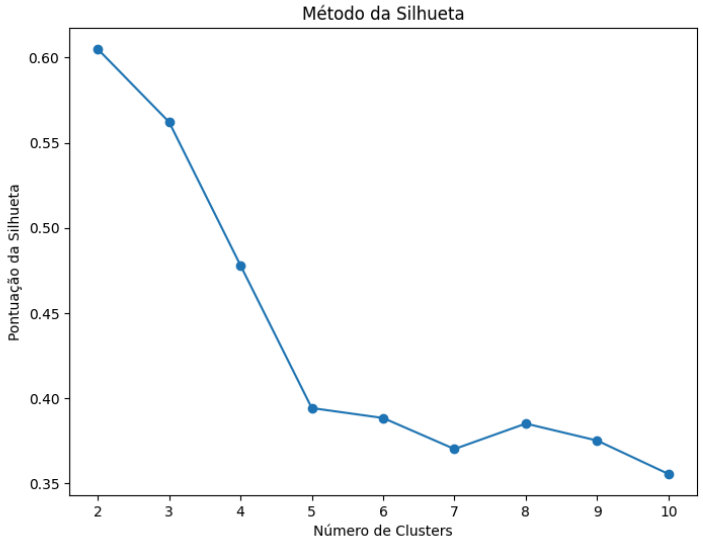
Questão 3 - Utilize o método Lasso para selecionar os dois atributos mais relevantes. Após isso, recalcule o número de clusters utilizando os métodos do cotovelo e da silhueta. Compare os resultados com a questão anterior. Se o número de clusters for diferente, elabore dois scatterplots para análise visual.

O **método Lasso** (Least Absolute Shrinkage and Selection Operator) é uma técnica de regularização e seleção de variáveis em modelos estatísticos. O Lasso adiciona uma penalização à função de custo baseada no valor absoluto dos coeficientes. Esta penalização pode fazer com que alguns coeficientes se tornem exatamente zero. Os coeficientes que permanecem diferentes de zero são considerados as variáveis mais relevantes.

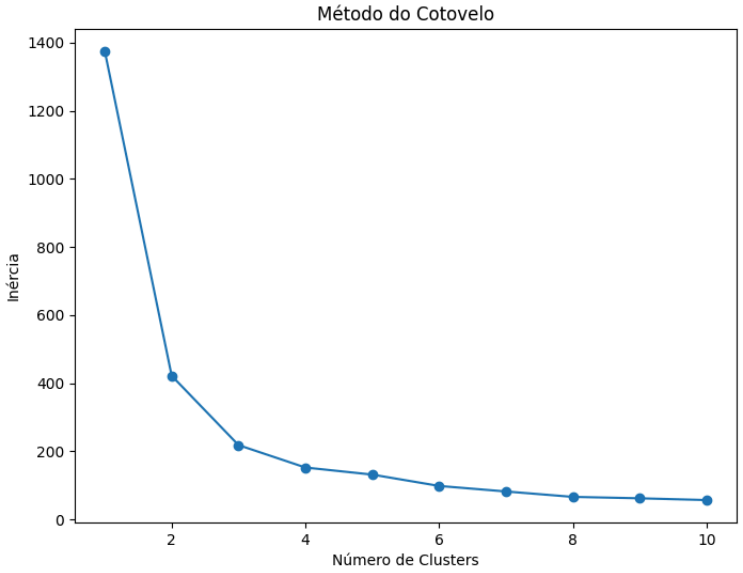
O Lasso foi usado para identificar as duas variáveis mais relevantes do dataset (alcohol e volatile acidity), permitindo:

- Simplificar a análise de clusters
- Focar nas características mais importantes
- Melhorar a interpretabilidade dos resultados
- Reduzir o ruído nos dados

Ao analisar os gráficos abaixo temos as seguintes informações:



- **Método da Silhueta:**
  - O melhor valor está em **K=2** (0.60)
  - Há uma queda gradual até **K=5**
  - A partir de **K=5**, o valor se estabiliza entre **0.35 - 0.40**

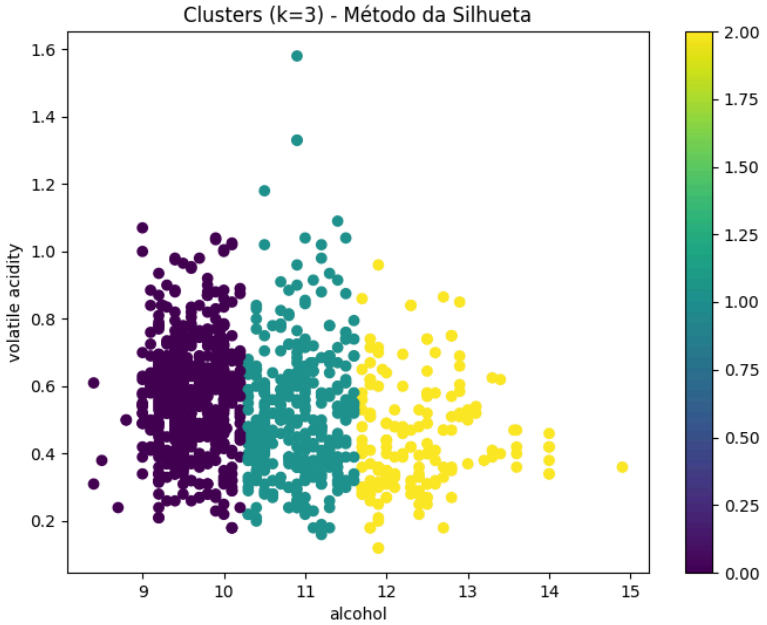


- **Método do Cotovelo:**
    - Queda acentuada de **K=1** até **K=3**
    - "Cotovelo" mais evidente em **K=3**
    - Após **K=4**, a curva se estabiliza significativamente

Os novos resultados, usando apenas os dois atributos mais relevantes (alcohol e volatile acidity), mostram algumas diferenças importantes:

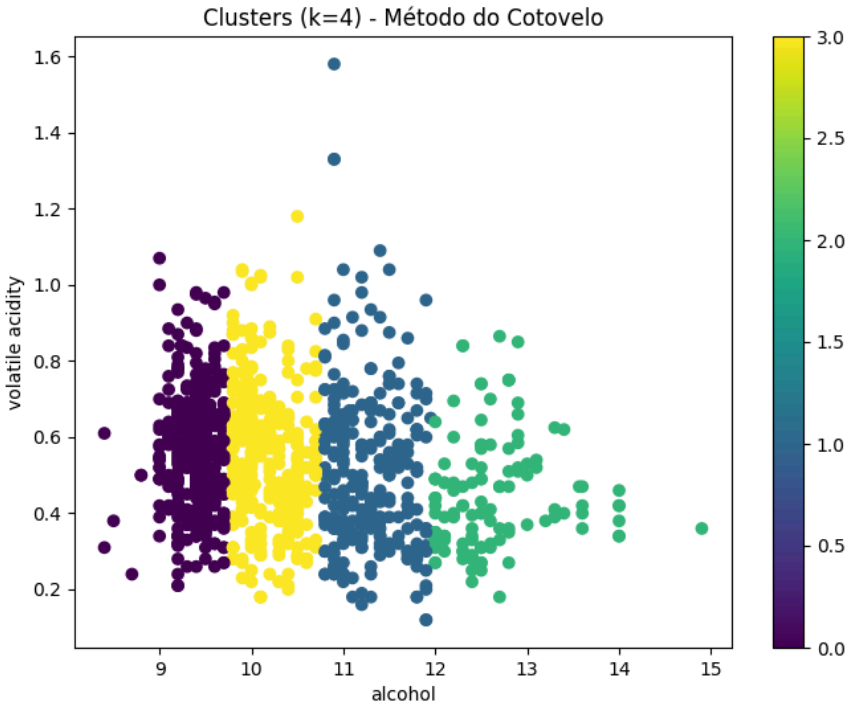
- A qualidade dos clusters melhorou, com coeficientes de silhueta mais altos (0.60 vs 0.33)
- O "cotovelo" ficou mais definido no **K=3**
- A inércia total diminuiu significativamente.

Agora analisamos os gráficos gerados a partir dos Scatterplots:



- **K=3 (Método da Silhueta):**

- Mostra uma separação clara em três grupos baseada principalmente no teor alcoólico
- Os clusters parecem bem definidos com pouca sobreposição
- Há uma correlação negativa entre alcohol e volatile acidity



- **K=4 (Método do Cotovelo):**
  - Divide os dados em quatro grupos mais granulares
  - A separação ainda é visível, mas com mais sobreposição entre clusters
  - O cluster adicional parece subdividir um dos grupos do **K=3**

A redução para apenas dois atributos mais relevantes resultou em clusters mais bem definidos, como evidenciado pelos maiores valores de silhueta. O **K=3** parece ser a escolha mais adequada neste caso, pois:

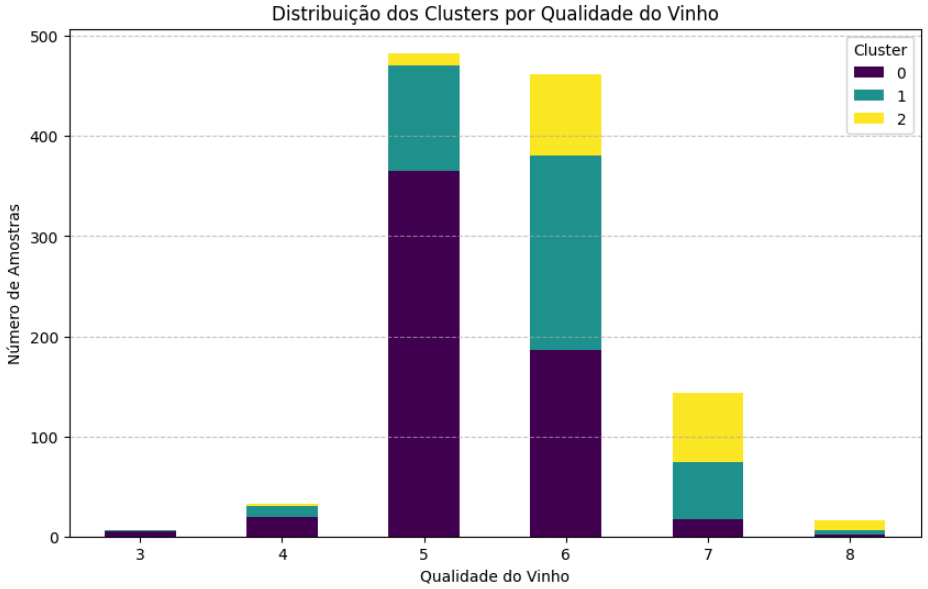
- Apresenta um alto coeficiente de silhueta

- Mostra uma separação visual clara dos grupos
- Oferece um bom equilíbrio entre simplicidade e capacidade explicativa

Isto sugere que a seleção de atributos via Lasso ajudou a melhorar a qualidade do clustering, simplificando o modelo sem perder informação significativa.

Questão 4 - Crie uma crosstab para examinar a distribuição dos clusters em relação às classes presentes na coluna alvo. Utilize o K-means com o número de clusters determinado pelo índice de silhueta.

A análise da distribuição dos clusters em relação à qualidade dos vinhos, utilizando o método K-means com **k=3** (determinado pelo índice de silhueta), revelou padrões significativos na segmentação dos dados. Os resultados foram obtidos a partir de um crosstab que relaciona os clusters com as diferentes categorias de qualidade.



Distribuição e Características dos Clusters:

- **Cluster 0 (595 amostras):**

Este cluster se caracteriza por uma forte concentração de vinhos de qualidade intermediária-baixa. Apresenta **365 vinhos de qualidade 5** e **186 de qualidade** 6, totalizando **92% das amostras** do cluster. A baixa presença de vinhos de qualidade superior (apenas 19 vinhos de qualidade 7-8) sugere que este cluster representa principalmente **vinhos de qualidade média** a **média-baixa**.

Cluster 1 (373 amostras):

Demonstra uma distribuição mais equilibrada entre as categorias de qualidade média, com predominância de **vinhos de qualidade 6** (195 amostras) e **5** (106 amostras). A presença significativa de **vinhos de qualidade 7** (57 amostras) indica uma tendência para qualidade média a média-alta.

Cluster 2 (175 amostras):

Este é o cluster mais distintivo em termos de qualidade, concentrando uma proporção significativa de vinhos de qualidade superior. Com **81 vinhos de qualidade 6**, **69 de qualidade 7** e **10 de qualidade 8**, este cluster claramente representa os **vinhos de maior qualidade** do conjunto de dados.

Correlação entre Clusters e Qualidade:

A análise revela uma forte correlação entre os clusters formados e as categorias de qualidade dos vinhos. Observa-se uma progressão clara na distribuição da qualidade entre os clusters, onde:

- Vinhos de qualidade inferior (3-4) concentram-se predominantemente no Cluster 0
- Vinhos de qualidade intermediária (5-6) distribuem-se principalmente entre os Clusters 0 e 1
- Vinhos de qualidade superior (7-8) têm presença marcante no Cluster 2

A análise demonstra que o algoritmo K-means, utilizando apenas as duas características selecionadas pelo método Lasso (teor alcoólico e acidez volátil), foi

capaz de criar uma segmentação efetiva dos vinhos de acordo com seus níveis de qualidade. Esta segmentação em três clusters revela padrões claros na distribuição da qualidade, sugerindo que estas características são indicadores robustos para a classificação da qualidade dos vinhos.

A eficácia desta segmentação indica que as características selecionadas pelo método Lasso são, de fato, altamente relevantes para a determinação da qualidade dos vinhos, proporcionando uma base sólida para análises e previsões futuras no contexto de classificação de vinhos.

4. Conclusões

Os resultados obtidos neste estudo foram altamente satisfatórios, superando as expectativas iniciais em diversos aspectos:

- **Melhoria na Classificação**:
    - O uso do GridSearchCV para otimização de parâmetros do KNN resultou em uma melhoria significativa na acurácia (0.6923)
    - A configuração ideal encontrada (K=8 com distância Euclidiana) superou os resultados anteriores
- **Análise de Clusters**:
    - A redução para duas variáveis via método Lasso melhorou significativamente a qualidade dos clusters
    - O coeficiente de silhueta aumentou de 0.33 para 0.60, indicando clusters mais bem definidos
    - A segmentação em três grupos mostrou-se eficaz na separação dos vinhos por qualidade
- **Relevância das Variáveis**:
    - O método Lasso identificou com sucesso as duas variáveis mais importantes (teor alcoólico e acidez volátil)

- Estas variáveis demonstraram ser indicadores robustos da qualidade dos vinhos

- **Validação dos Resultados**:
  - A análise da distribuição dos clusters confirmou a eficácia da segmentação
  - Os padrões encontrados são consistentes com o conhecimento enológico existente"

5. Próximos passos

Para dar continuidade e expandir este estudo com o intuito de obter a maior quantidade de dados possíveis, sugere-se:

- **Aprofundamento da Análise**:
  - Explorar outros algoritmos de classificação (Random Forest, SVM, etc.)
  - Investigar técnicas de balanceamento de classes
  - Realizar análise de sensibilidade dos parâmetros dos modelos
- **Expansão do Escopo**:
  - Incluir dados de vinhos brancos para comparação
  - Avaliar a aplicabilidade do modelo em outros tipos de vinhos
  - Investigar a influência de fatores sazonais na qualidade
- **Melhorias Técnicas**:
  - Implementar técnicas de validação cruzada mais robustas
  - Desenvolver um sistema de previsão em tempo real
  - Criar uma interface visual para facilitar a análise dos resultados
- **Aplicações Práticas**:

- Desenvolver um sistema de recomendação para produtores
- Criar guidelines para otimização do processo de produção
- Estabelecer parâmetros de controle de qualidade baseados nos resultados

- **Validação Externa**:
  - Realizar testes com novos conjuntos de dados
  - Validar os resultados com especialistas em vinhos
  - Comparar com outros estudos similares da literatura"