PONTIFÍCIA UNIVERSIDADE CATÓLICA DE CAMPINAS
CURSO DE ENGENHARIA DE SOFTWARE
PROFESSOR FERNANDO H. C. SILVA

Grupo 1

22000354 - Bruno Tasso Savoia

23010918 - Luigi Mazzoni Targa

23013238 - Marcela Franco

23009486 – Nicole Silvestrini Garrio

20018040 – Vitor Hugo Amaro Aristides

ATIVIDADE 03: AGRUPAMENTO

CAMPINAS - SP

2025

SUMÁRIO

1. Introdução

O objetivo desta atividade é aplicar e analisar as técnicas de agrupamento em duas bases de dados, sendo uma delas a base Iris, e a outra, que escolhemos para este trabalho, a base Wine, também da mesma biblioteca. O foco principal é utilizar métodos de agrupamento hierárquico e particional, como o K-Means, Bi-secting K-Means e Linkage (single, average, complete e ward), para segmentar as amostras dessas bases de dados, estudando e selecionado parâmetros e técnicas que podem ser úteis nesse processo.

2. Descrição do problema

Especificamente, usando os critérios de avaliação e técnicas de parametrização discutidas em aula:

1. Aplicar um algoritmo hierárquico e um particional nas duas bases.
2. Selecionar melhores parâmetros utilizando técnica de joelho/elbow ou alguma das métricas de avaliação (interna ou externa).
3. Avaliar grupos gerados.
4. Descrever e discutir informações obtidas ao aplicar os algoritmos.

3. Implementação e algoritmos utilizados

Para a implementação, escolheu-se Python devido à sua facilidade de uso e à disponibilidade de bibliotecas especializadas que permitem uma visualização clara e intuitiva dos gráficos e dendrogramas gerados.

Neste projeto, organizamos o código em dois arquivos distintos para modularizar e separar os efeitos de cada dataset: iris.py e wine.py. Cada arquivo contém funções específicas para exibir os dados, aplicar o algoritmo de agrupamento hierárquico e realizar o agrupamento particional, permitindo uma análise mais clara e independente de cada conjunto de dados.

Não foi necessário realizar o upload dos datasets, pois ambos, o *Iris* e o *Wine*, são disponibilizados diretamente pela biblioteca sklearn. Além disso, a biblioteca matplotlib também foi essencial para o desenvolvimento, pois com ela foi possível exibir os gráficos para estudo.

Algoritmos utilizados:

- Algoritmo Particional K-Means

  Um algoritmo de aprendizado não supervisionado que tenta dividir os dados em k grupos minimizando a soma das distâncias dos pontos ao centróide do cluster.

  Etapas seguidas:

  1. Definir um intervalo de valores para k.
  2. Utilizar o método do cotovelo (Elbow Method) para encontrar o número ideal de clusters.
  3. Executar os algoritmos K-Means e Bi-Secting K-Means com o k escolhido (k = 3 para ambos os datasets).
  4. Avaliar os resultados visualmente e por métricas como Silhouette Score.

- Algoritmo Hierárquico Linkage

  Os métodos hierárquicos criam uma estrutura de dendrograma para representar a junção de grupos de dados.

  Etapas seguidas:

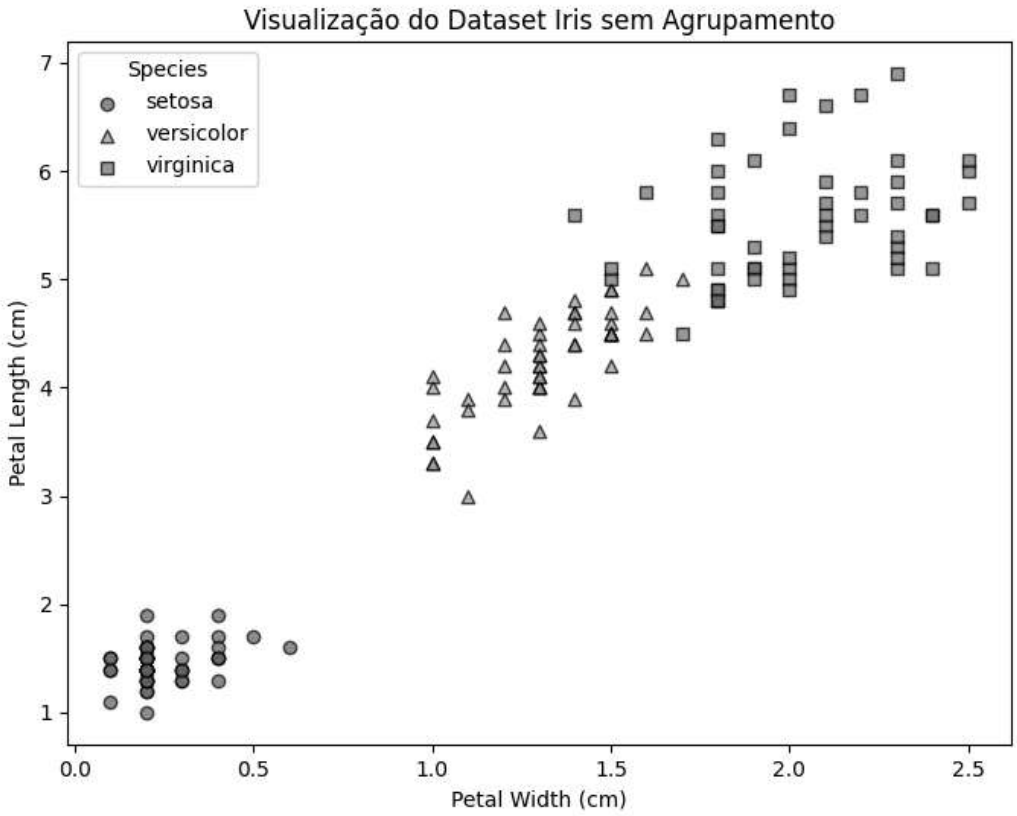
  1. Aplicação dos métodos linkage (single, average, complete, ward)
  2. Definição de k (corte) no dendrograma.
  3. Geração dos gráficos de dispersão.

Além da implementação dos algoritmos de agrupamento, a classe "main.py" foi continuamente ajustada ao longo do desenvolvimento para integrar as funções criadas, permitindo a execução dos algoritmos em um único “menu”, facilitando a execução.

4. Bases de dados

Para esta atividade, foram selecionadas as bases de dados ‘Iris’ e ‘Wine’, ambas estão disponíveis na biblioteca sklearn.datasets e são famosas no estudo e teste de algoritmos de Machine Learning.

4.1 Iris



- Contém medições de 150 flores de três espécies (Setosa, Versicolor e Virginica).
- Cada amostra possui quatro atributos numéricos relacionados às características das sépalas e pétalas (comprimento e largura).

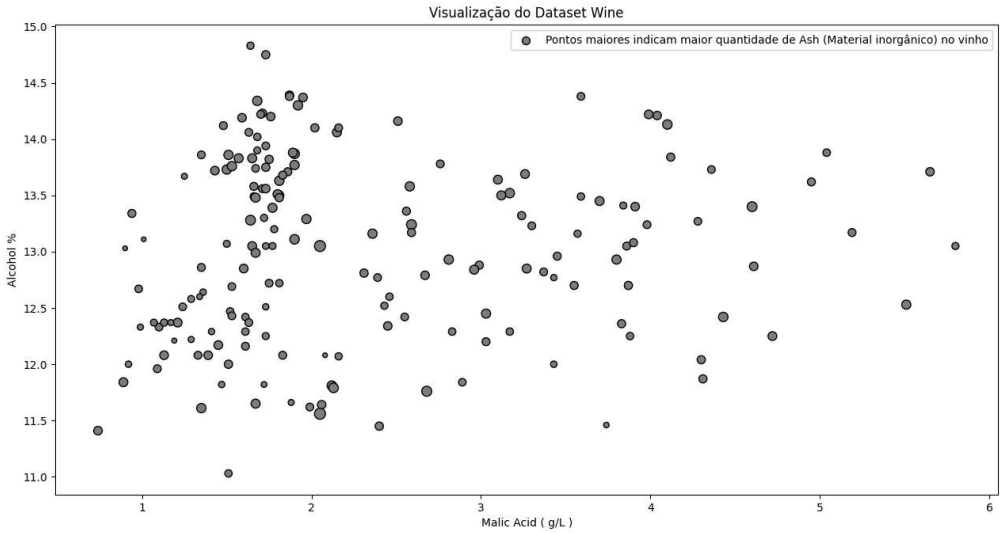
4.1.1 Atributos selecionados de Iris Dataset

- species: Usada para diferenciar visualmente os clusters.

- petalLength: Utilizada na coordenada Y do gráfico.
- petalWidth: Utilizada na coordenada X do gráfico.

Para o dataset íris, escolhemos apenas 2 atributos (Além das espécies, porém, como é um rótulo, não entra no cálculo dos clusters, e sim apenas para visualização gráfica), sendo esses atributos o comprimento e a largura da pétala da flor, pois dessa maneira é fácil criar uma imagem mental da largura e comprimento das pétalas da flor por conta de ser duas informações da mesma parte da planta.

4.2 Wine



- Consiste em 178 amostras de três tipos diferentes de vinho
- Caracterizada por 13 atributos químicos, como teor alcoólico, acidez e concentração de minerais.

4.2.1 Atributos selecionados de Wine Dataset

- Alcohol: Coordenada Y no gráfico.
- malic acid: Coordenada X no gráfico.
- ash: Determina o tamanho dos pontos no gráfico.

Já no dataset wine, foram escolhidos 3 atributos, o percentual de álcool porque o teor alcoólico é um dos principais fatores que diferenciam os tipos de vinho, ácido málico porque está relacionado com a acidez e o sabor do vinho, e ash (material inorgânico), que mesmo não tendo um impacto muito grande no vinho, adiciona uma dimensão a mais ao gráfico.

5. Normalização dos dados

Como serão aplicados algoritmos hierárquicos, é importante considerar que variáveis com escalas diferentes podem influenciar o agrupamento de forma desigual, pois as variáveis de maiores magnitudes podem dominar a análise dos dados.

Por exemplo, na base de dados ‘Wine’ temos a seguinte situação:
- A variável "magnesium" pode ter valores na faixa de 50 a 200.
- A variável "flavanoids" pode variar de 0 a 3.

Sem a normalização, as variáveis com valores numéricos maiores poderiam ter um impacto desproporcional nas distâncias calculadas, afetando diretamente a qualidade do agrupamento. Portanto, para resolver esse problema, aplicamos o processo de normalização nas bases de dados, a partir do uso da ferramenta MinMaxScaler, da biblioteca sklearn. Essa técnica garante que todas as variáveis terão uma faixa de valores comum, o que pode melhorar a performance dos algoritmos que são sensíveis à magnitude das variáveis.

6. Aplicação de Algoritmo Hierárquico

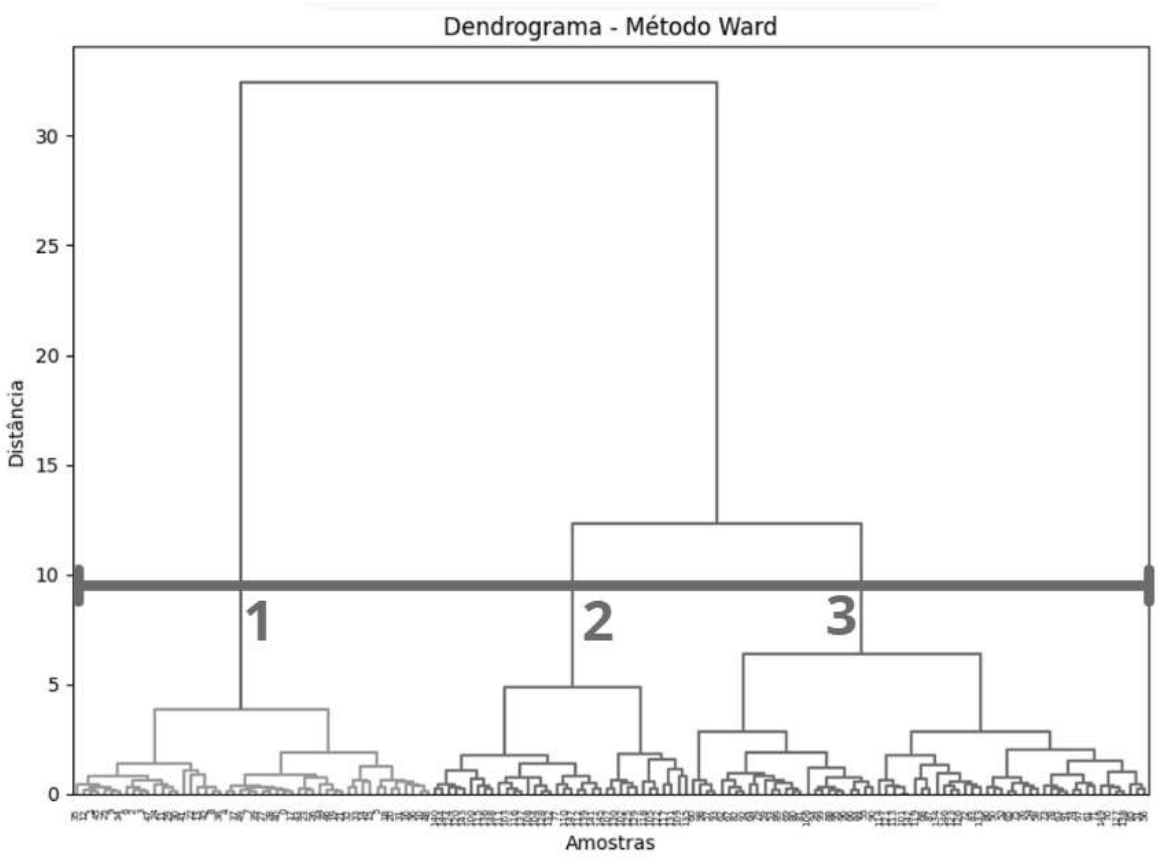
Nos testes, serão aplicados nas bases de dados quatro algoritmos hierárquicos, de forma aglomerativa, são eles: Single Linkage, Average Linkage, Complete Linkage e Ward Linkage. Temos como objetivo determinar qual algoritmo oferece melhores resultados em termos de qualidade de agrupamento. Para esse estudo, usaremos como objeto de análise os gráficos de dispersão e os

dendrogramas de duas dimensões (para melhor visualização) gerados.

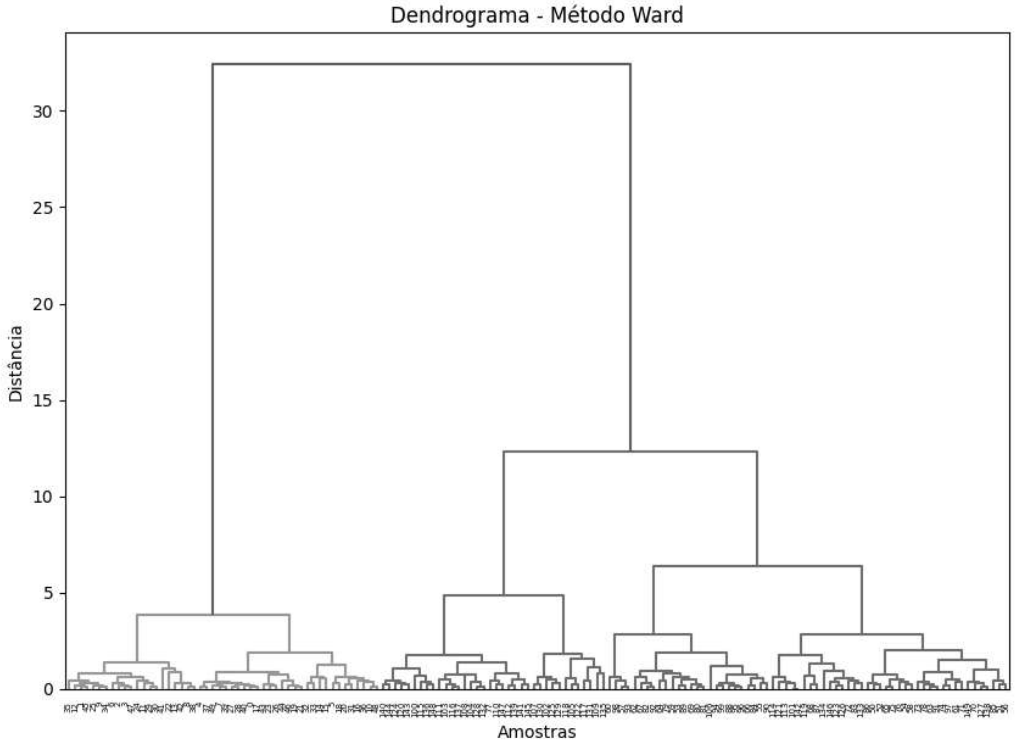
6.1 Iris hierárquico

Ao analisar as visualizações geradas pelos algoritmos, o ‘Ward Linkage’ apresentou-se como a melhor escolha para aplicação na base de dados Iris.

Nesta base de dados, a escolha do número de clusters foi feita com base na função determinar_n_clusters. Essa é uma abordagem automática e objetiva para definir o número de clusters. Ela calcula as distâncias entre as fusões no processo de aglomeração hierárquica e identifica o maior salto entre essas distâncias. Esse maior salto indica o ponto onde a fusão entre clusters se torna menos significativa, sugerindo o número ideal de clusters.



No caso do dataset Iris, a função sugeriu 3 clusters, alinhando-se com o número real de espécies de flores na base.



Dendrograma da base Iris com aplicação do método Ward Linkage.

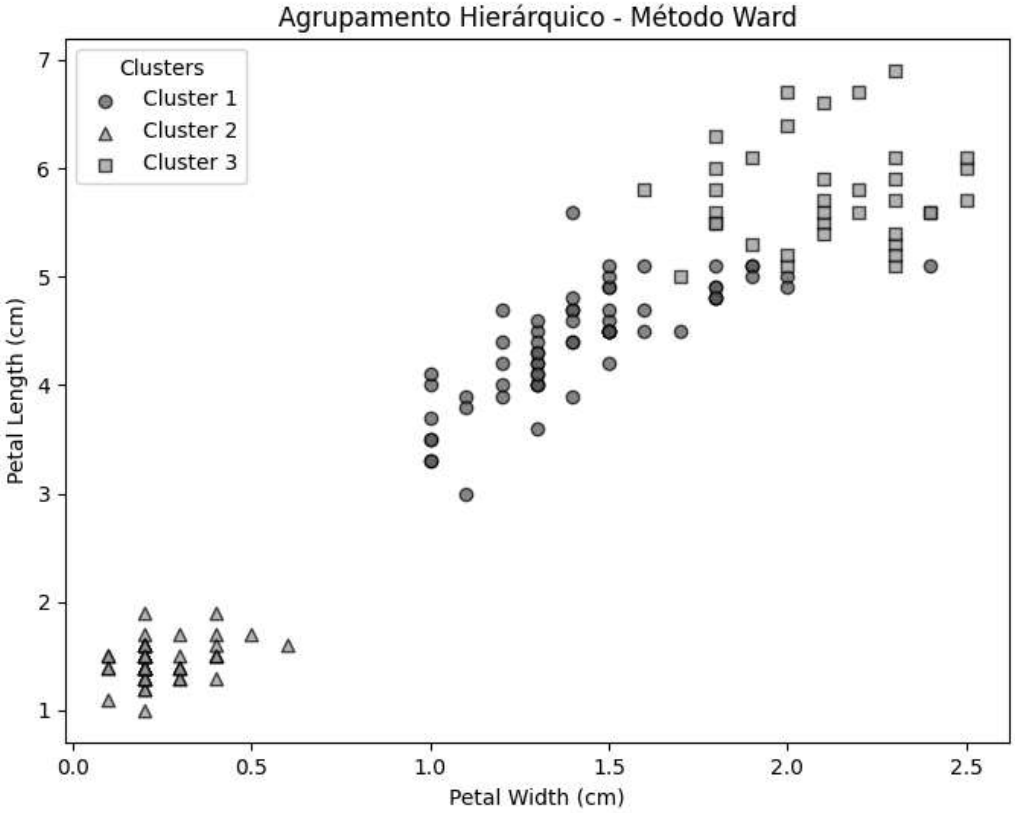
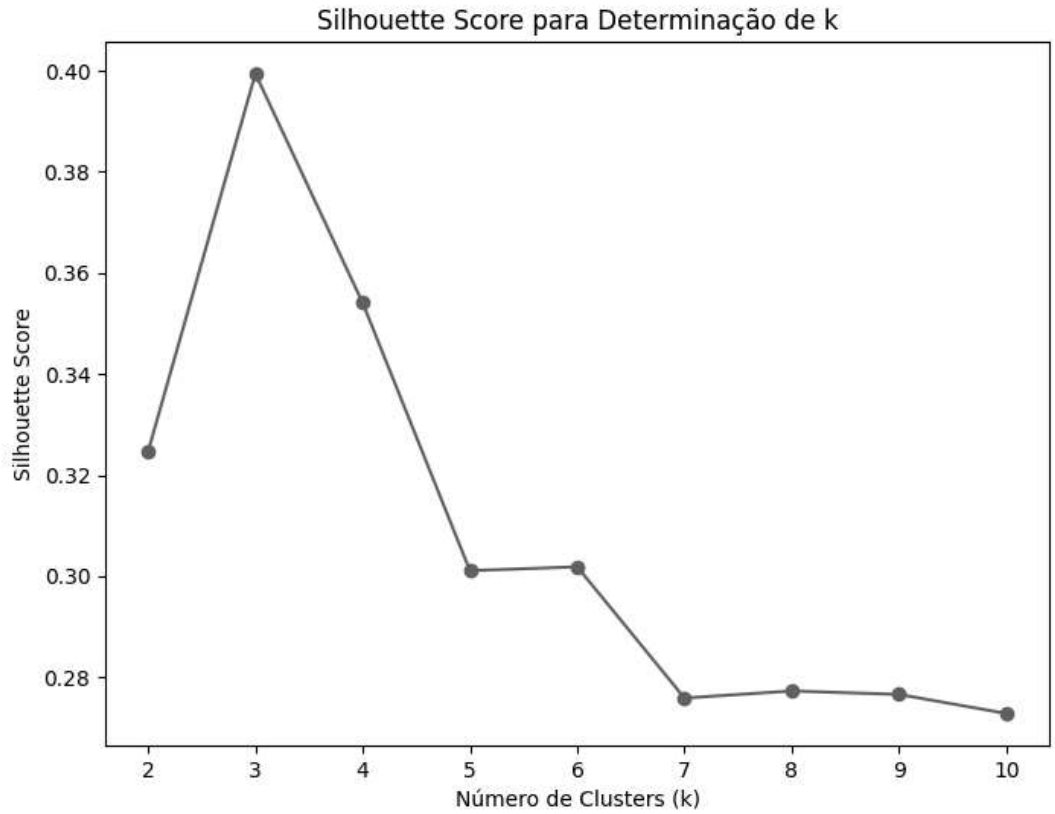


Gráfico de Dispersão da base Iris com aplicação do método Ward Linkage.

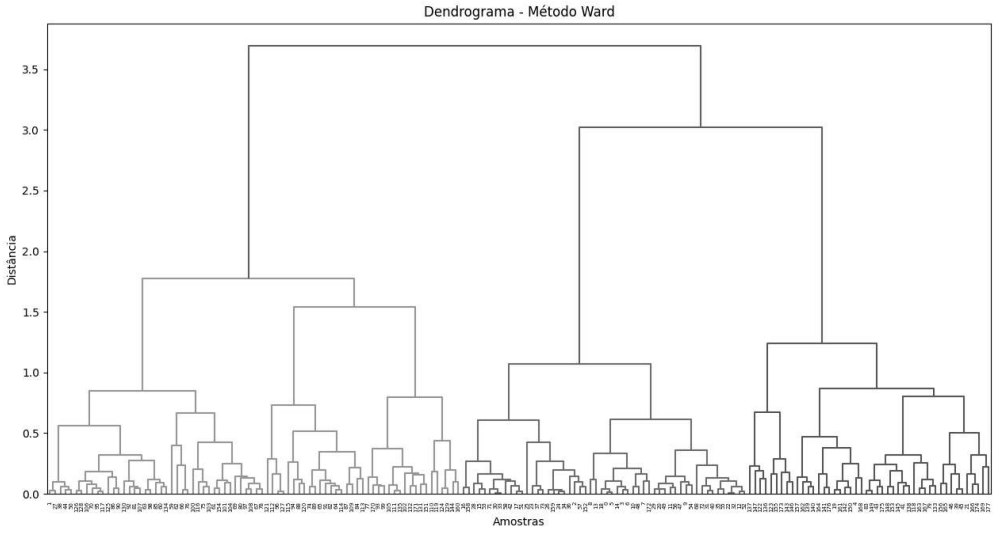
6.2 Wine hierárquico

Ao analisar as visualizações geradas pelos algoritmos, o 'Ward Linkage' apresentou-se como a melhor escolha para aplicação na base de dados Wine.

Nesta base de dados, o Silhouette Score foi escolhido para determinar o número de clusters, pois é uma métrica interna que avalia a qualidade do agrupamento. Ele mede o quão bem cada ponto está dentro de seu próprio cluster em relação aos outros clusters. O número de clusters com o maior valor de Silhouette Score é selecionado como o melhor.



Neste caso, o melhor número de clusters (k) baseado no Silhouette Score é 3.



Dendrograma da base Wine com aplicação do método Ward Linkage.

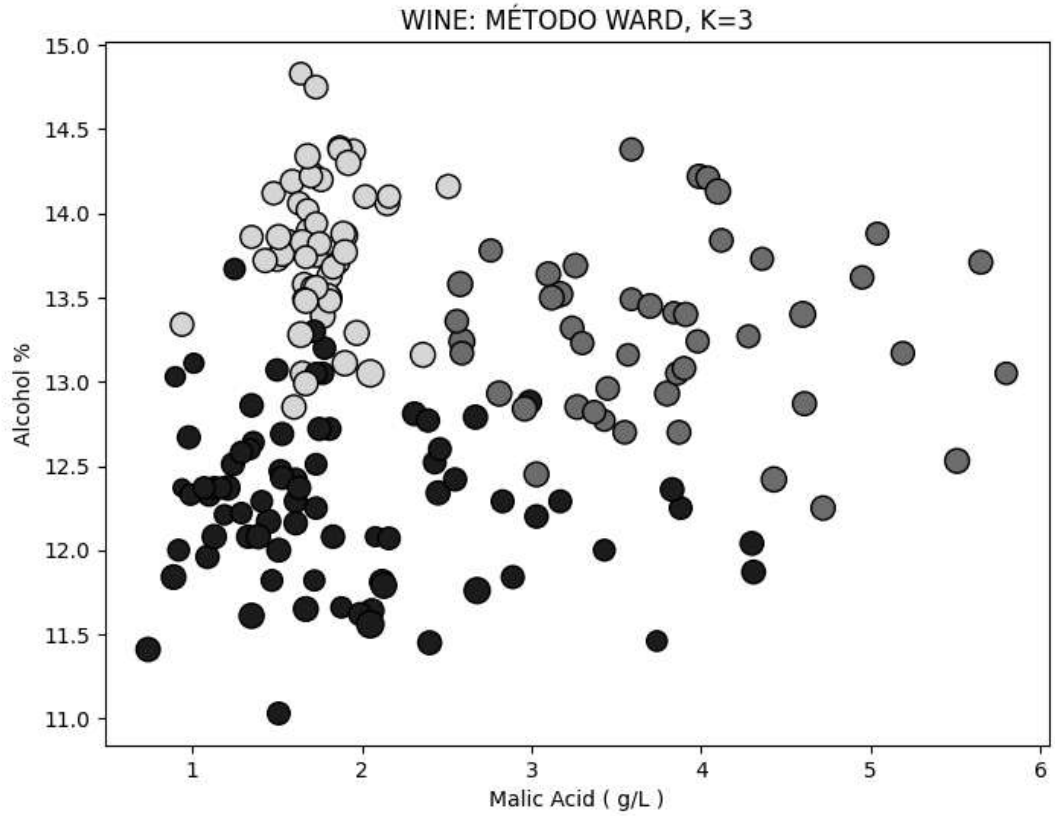


Gráfico de Dispersão da base Wine com aplicação do método Ward Linkage.

6.3 Algoritmo de melhor desempenho

Sabendo que a escolha do melhor algoritmo hierárquico de agrupamento (entre Single, Average, Complete e Ward) depende de vários fatores, incluindo a distribuição dos dados e a qualidade dos clusters desejada, realizamos testes aplicando os quatro algoritmos e, nas duas bases de dados, o que desempenhou melhor foi o Ward Linkage.

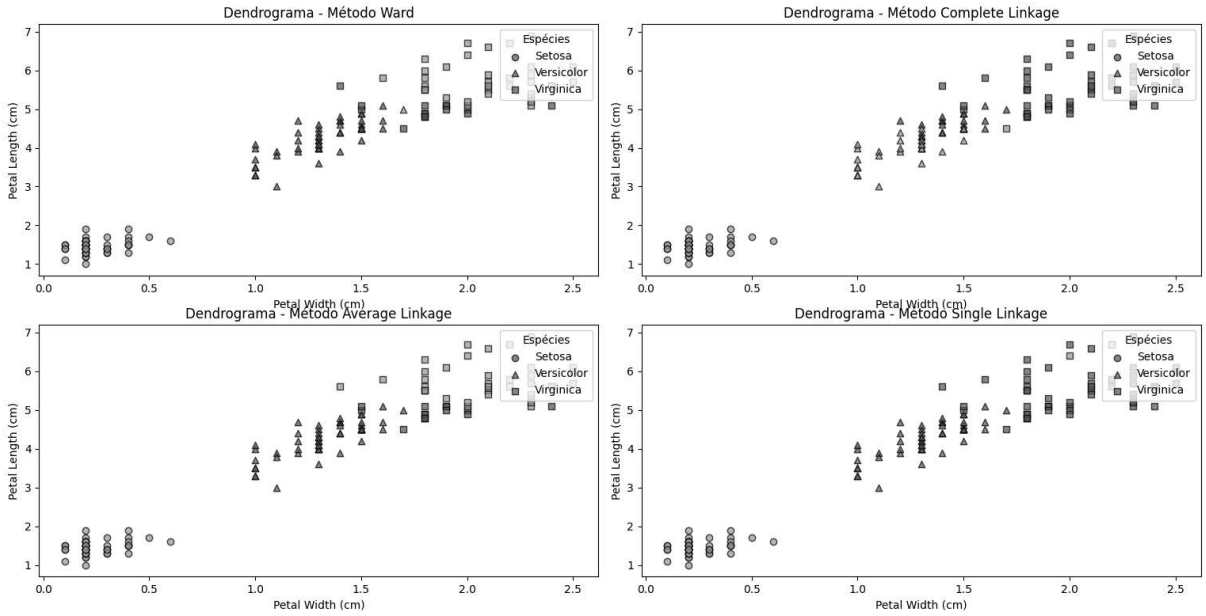
Esse algoritmo tende a gerar clusters compactos e bem definidos, que são adequados para dados com classes separadas, como é o caso do Wine e da Iris. Ele funciona de forma que haja a minimização da variância intracluster a

cada fusão, o que garante que, internamente, os clusters sejam mais compactos e homogêneos.

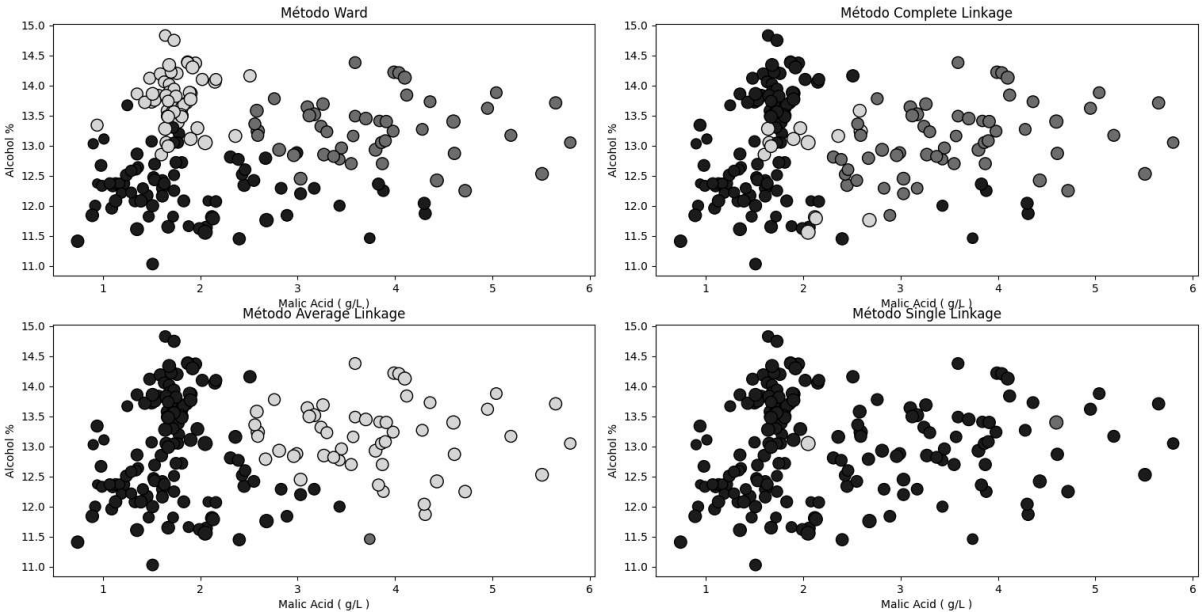
Para comparação:

- O método ‘Single Linkage’ foi o pior, pois é extremamente sensível a outliers e sofre de chainning effect.
- O método ‘Average Linkage’, que utiliza da média das distâncias entre todos os pares, desempenhou bem, porém este algoritmo não utiliza a variância interna.
- O método ‘Complete Linkage’ nesse caso foi suscetível a formar agrupamentos menos compactos ou mais dispersos.

Métodos Linkage aplicados na base Iris



Métodos Linkage aplicados na base Wine



7. Aplicação de Algoritmo Particional

- Para os algoritmos particionais, tanto no Wine como no Iris, o algoritmo de agrupamento usado foi o “K-Means”

- Para a validação dos clusters, o Silhouette Score foi escolhido em vez do Xie-Beni Index porque ele oferece uma interpretação mais intuitiva e é amplamente utilizado para avaliar a qualidade dos clusters de forma independente da forma e densidade dos grupos.

- Durante a execução do código, os algoritmos são executados n vezes, com os centróides iniciando em lugares aleatórios todas as vezes e armazenando o melhor Silhouette Score. Após isso, só é mostrado ao usuário o melhor resultado.

8. Aplicação do PairPlot

O PairPlot é uma ferramenta útil para visualizar a distribuição de atributos e a relação entre eles em um conjunto de dados. Ele gera gráficos de dispersão para cada par de variáveis numéricas e histogramas para visualizar a distribuição univariada de cada atributo. Essa abordagem permite uma inspeção visual da separação dos grupos e possíveis padrões de agrupamento antes da aplicação dos algoritmos de clustering.

8.1 Iris PairPlot

Para a base de dados Iris, aplicamos o PairPlot para observar como as diferentes espécies de flores se distribuem em relação aos seus atributos. Os pares de variáveis escolhidos foram:

- **Petal Length vs. Petal Width**
- **Sepal Length vs. Sepal Width**

Os gráficos revelam que as espécies apresentam padrões distintos em algumas combinações de atributos, facilitando a visualização dos agrupamentos naturais.

Resultados:

- A espécie Setosa é bem separada das demais usando apenas **Petal Length e Petal Width.**
- Versicolor e Virginica possuem alguma sobreposição, tornando mais difícil a distinção sem um modelo de clustering adequado.

8.2 Wine PairPlot

Para a base de dados Wine, o PairPlot foi utilizado para visualizar as relações entre os atributos químicos dos vinhos. As variáveis selecionadas foram:

- **Alcohol vs. Malic Acid**
- **Alcohol vs. Ash**

- **Malic Acid vs. Ash**

Os gráficos mostram que há uma separação razoável entre os tipos de vinho em certas combinações de variáveis, mas em algumas dimensões há sobreposição significativa. A normalização dos dados foi essencial para garantir que nenhuma variável dominasse a análise.

Resultados:

- Atributos como **Alcohol e Malic Acid** já sugerem separações visíveis entre alguns tipos de vinho.
- A variável **Ash**, apesar de apresentar pouca variação, adiciona uma dimensão extra ao agrupamento e pode contribuir para uma melhor segmentação.

O uso do PairPlot auxiliou na escolha das variáveis mais relevantes para os algoritmos de clustering, proporcionando uma visualização clara dos agrupamentos e permitindo ajustes antes da aplicação dos métodos hierárquicos e particionais.

9. Conclusão

Neste estudo, foram aplicados e analisados algoritmos de agrupamento em duas bases de dados conhecidas: Iris e Wine. Através da implementação de técnicas de clusterização hierárquica e particional, foi possível segmentar os dados e avaliar a qualidade dos agrupamentos utilizando métricas como Silhouette Score e o método do cotovelo (Elbow Method).

Os resultados mostraram que o algoritmo Ward Linkage se destacou nos agrupamentos hierárquicos, formando clusters mais compactos e homogêneos em ambas as bases. No agrupamento particional, o K-Means demonstrou um

bom desempenho ao encontrar padrões nos dados, especialmente após a normalização.

Além disso, a aplicação do PairPlot contribuiu para a visualização e validação dos clusters gerados, permitindo uma interpretação mais intuitiva da separação das classes.

Com base nos experimentos realizados, conclui-se que a escolha adequada dos algoritmos e a seleção criteriosa dos parâmetros são essenciais para a obtenção de agrupamentos consistentes e interpretáveis. A combinação de diferentes técnicas de validação ajudou a garantir que os clusters formados fossem representativos das estruturas naturais dos dados. Assim, este trabalho reforça a importância da análise exploratória e da experimentação com múltiplos métodos para alcançar melhores resultados em problemas de agrupamento de dados.