**Aula Prática 12**

**Tema: Comparação de Métodos de Agrupamento**

**Murielly Oliveira Nascimento – 11921BSI222**

## Ferramentas Usadas

Os pré-processamentos e agrupamentos foram feitos com a ferramenta Weka, enquanto as validações, com implementações na linguagem python usando as bibliotecas sckit-learn, numpy e pandas. Para a medida de pureza, em específico, foi usada a implementação de David Mugisha, disponibilizada no seguinte link:

https://gist.github.com/jhumigas/010473a456462106a3720ca953b2c4e2

As bases de dados escolhidas para testes foram Wine Data Set e Seeds Data Set, que podem ser encontradas nos seguintes links:

https://archive.ics.uci.edu/ml/datasets/seeds

https://archive.ics.uci.edu/ml/datasets/wine

## Algoritmos de Agrupamento

De acordo com Pang-Ning Tan o agrupamento objetiva encontrar grupos úteis de objetos, onde a utilidade seja definida pelos objetivos da análise de dados. Neste sentido, o algoritmo K-means é uma técnica Particional de agrupamento baseada em protótipos que tenta encontrar m número especificado pelo usuário de grupos (K), que são representados pelos seus centróides.

---

Algoritmo do K-means básico

---

1: Selecione K pontos como centroides iniciais.

2: **repita**

3: Forme K grupos atribuindo cada ponto ao seu centroide mais próximo

4: Recalcule o centroide de cada grupo/

5: **até que** os centroides não mudem.

---

A técnica Agrupamento Hierárquico possui duas abordagens: aglomerativa, comece com os pontos como grupos individuais e, em cada etapa, funda os pares mais próximos de grupos; e a divisiva, comece com um grupo inclusivo com tudo e, a cada etapa, divida um grupo até que restem apenas grupos únicos de pontos individuais.

Um agrupamento hierárquico é exibido frequentemente usando um diagrama do tipo árvore chamado dendrograma, que exibe tanto os

relacionamentos grupo-subgrupo quanto a ordem na qual os grupos são fundidos ou divididos.

Algoritmo de Agrupamento Hierárquico Aglomerativo

1: Calcule a matriz de proximidade, caso necessário.

2: **repita**

3: Funda os dois grupos mais próximos.

4: Atualize a matriz de proximidade para refletir a proximidade entre o novo grupo e os grupos originais.

5: **até que** reste apenas um grupo.

Quanto a proximidade entre os grupos há duas formas: Min (Single Link), define a proximidade dos grupos como a proximidade entre os dois elementos mais próximos que estão em diferentes grupos; e Max (Complete Link), a proximidade dos grupos é calculada como a maior distância entre dois elementos em grupos diferentes.

Para os agrupamentos feitos neste trabalho foram usados os parâmetros K igual a 3,2 e 4.

## Medidas de Avaliação

- Pureza: mede o quanto cada cluster contém exemplos de uma única classe.

  $p_i = \max p_{ij}$ pureza do i-ésimo *cluster*

  $pureza = \sum_{i=1}^{k} \frac{m_i}{m} p_i$ pureza total

- Coeficiente de Jaccard: medida orientada a similaridade.

$$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- Largura da Silhueta: mostra quais objetos estão bem situados dentro dos seus clusters e quais estão fora de um cluster apropriado.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad SWC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

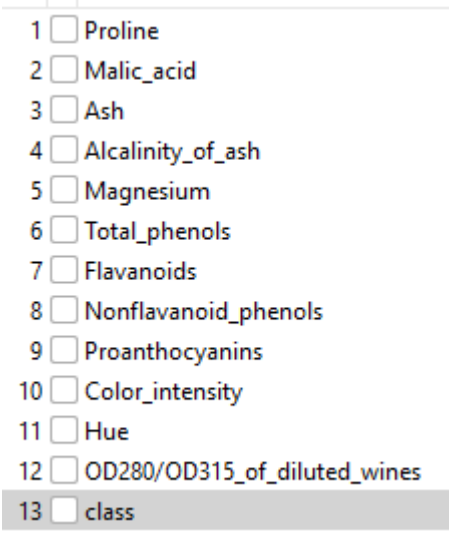
$$SWC \in [-1, +1]$$

# Análise da Base Wine

## Informações

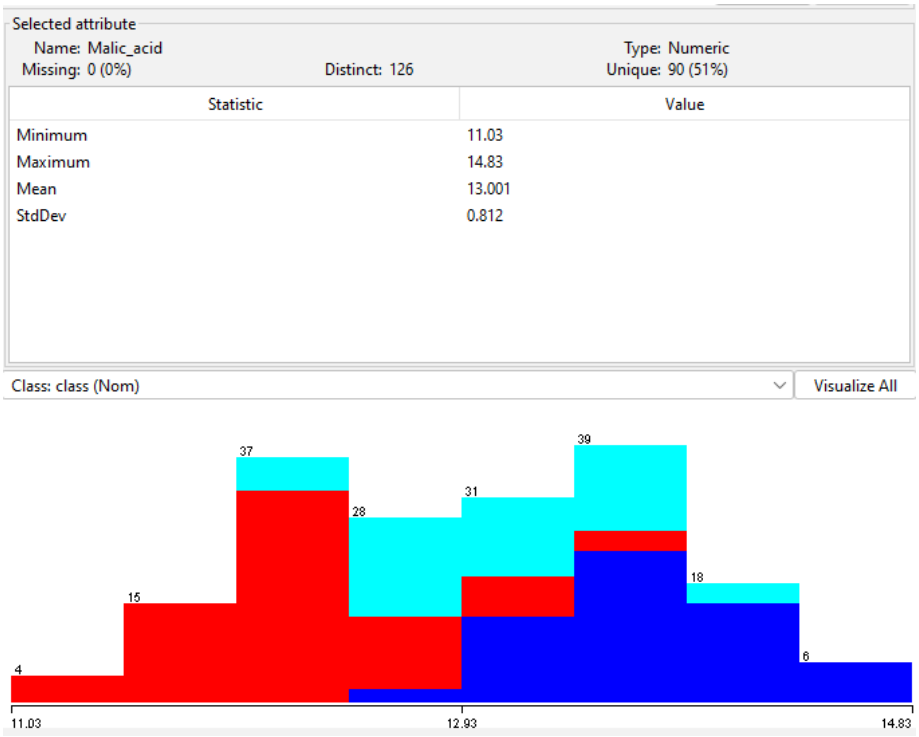
A base Wine é constituída por dados resultantes de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes. A análise determinou as quantidades de 13 constituintes encontradas em cada um dos três tipos de vinhos.

O Wine Data Set possui 178 instâncias com 13 atributos, os quais são inteiros ou reais. Não há valores ausentes e os atributos são descritos, de acordo com (MATOS, 2021), da seguinte forma:

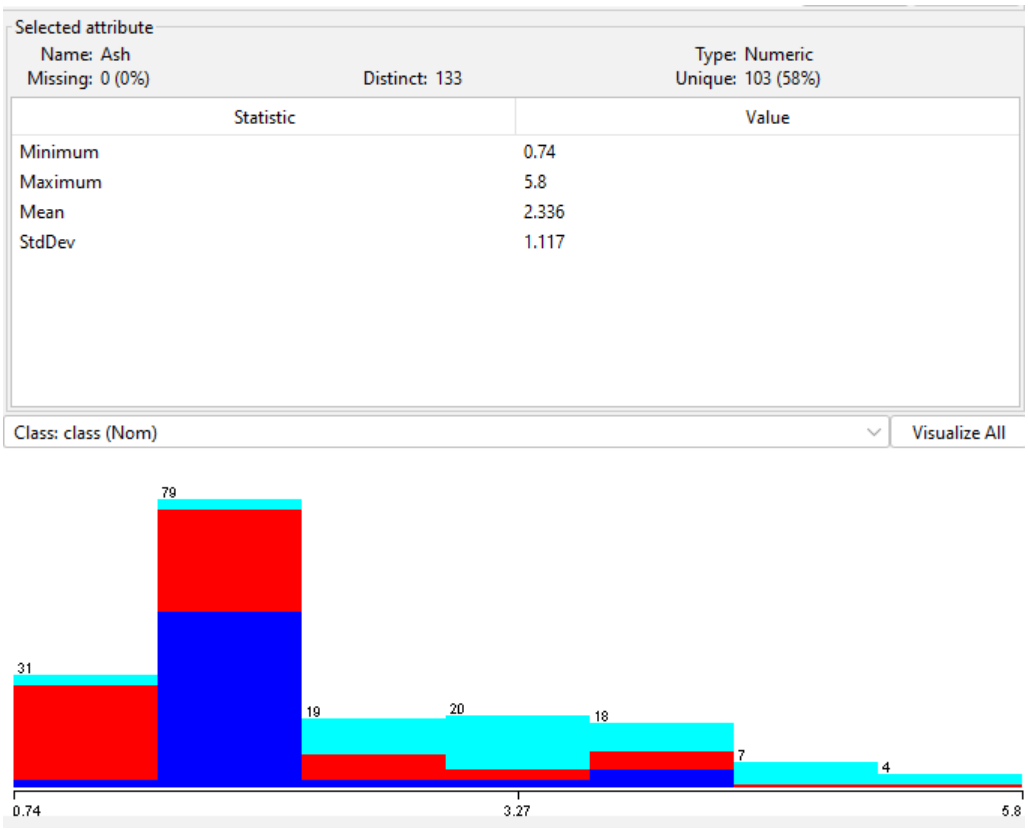


- **Classe:** dividida em 1,2,3 representando cada um dos vinhedos que foram analisados.

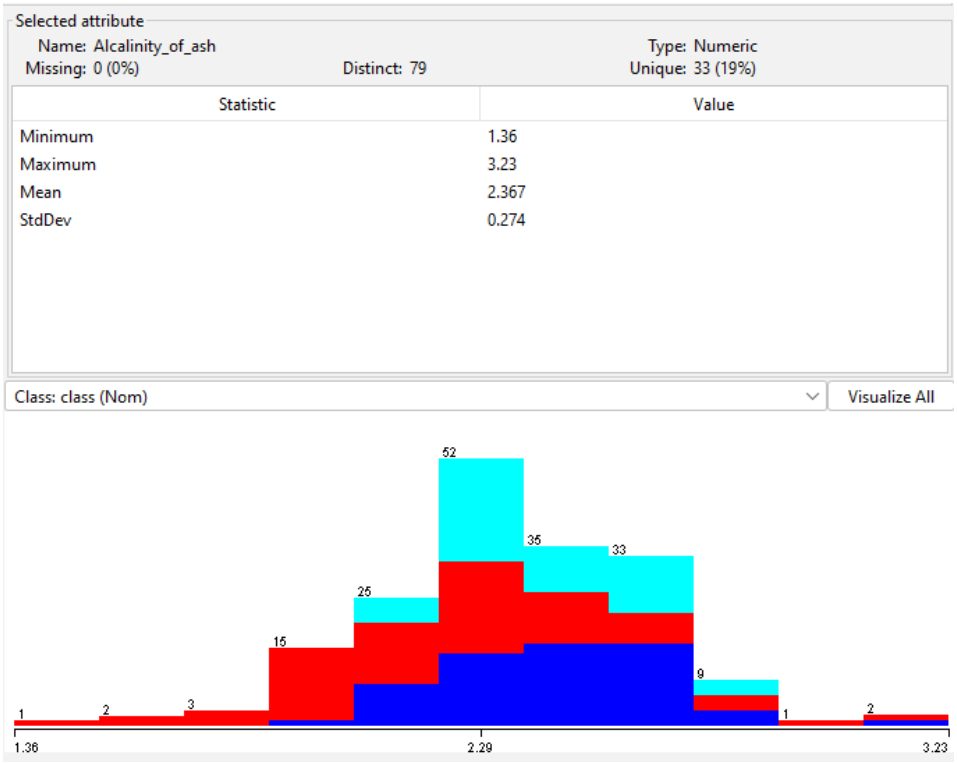

- **Ácido Málico:** o ácido málico encontra-se presente nas uvas, mais concretamente na casca, sementes e caules. Os valores de ácido málico fornecem informações sobre a fermentação malolática, pois quanto menor for o seu valor maior a percentagem que foi transformada com a fermentação malolática.



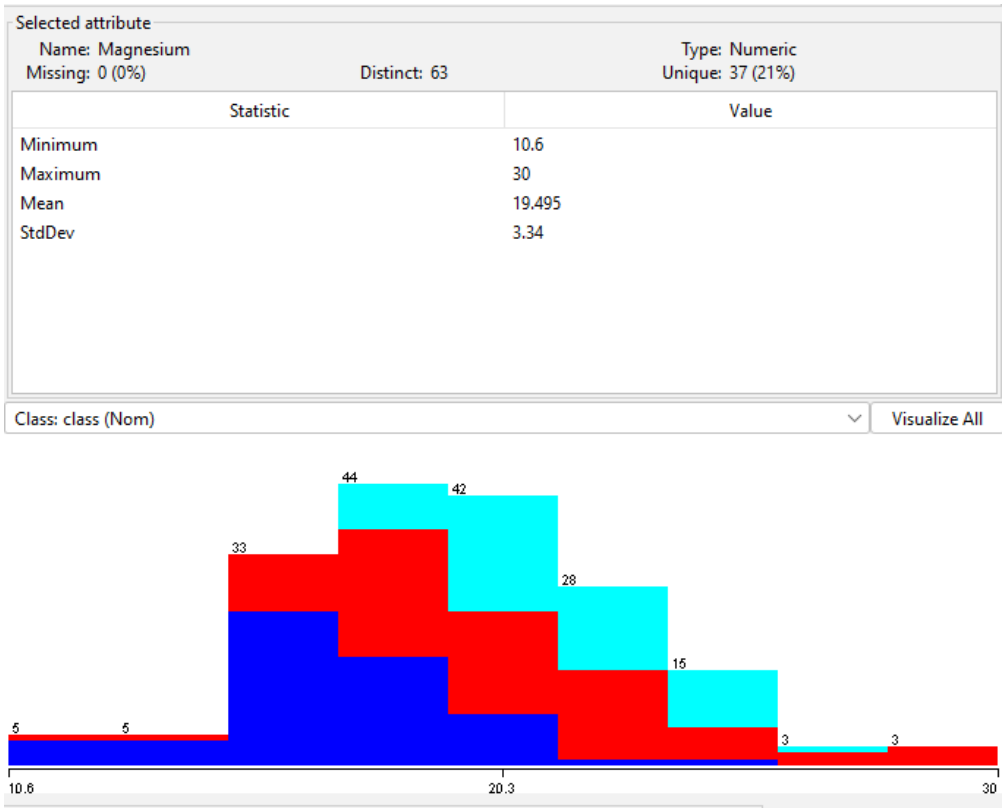
- **Cinzas**: a cinza, neste contexto, não é mais do que o conteúdo mineral presente nos mostos e respetivos vinhos, que resulta da combustão do extrato seco (resíduos de evaporação dos vinhos e mostos). Os valores da cinza permitem perceber se o vinho é ou não autêntico, dado que caso estes sejam muito baixos ou muito altos é sinónimo de que não se trata de um vinho adulterado. Os valores que a cinza apresenta, em média, no vinho são de 1.2 a 3 g/L, enquanto nos mostos esse valor é de 3 a 5 g/L. Relativamente ao tipo de vinho, existem valores mínimos tabelados pelo Instituto da Vinha e do Vinho, que para os vinhos brancos e rosados é de 1.6 g/L e para os tintos é de 1.8 g/L. Estes valores diferem entre si devido aos processos de clarificação



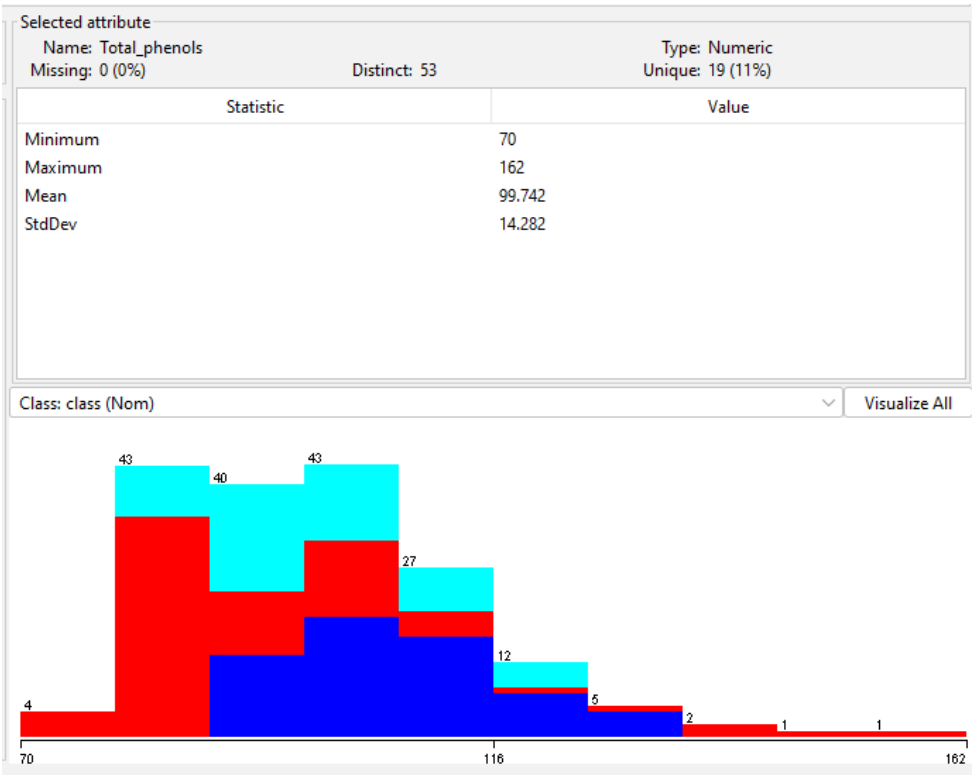
- **Alcalinidade das Cinzas:** A alcalinidade da cinza vai fornecer informação relativamente à quantidade de ácidos orgânicos presentes no vinho na forma de sal. Os valores da alcalinidade da cinza, permitem verificar a presença de ácidos na forma livre, mas também permitem verificar se existe ou não adulteração dos vinhos.



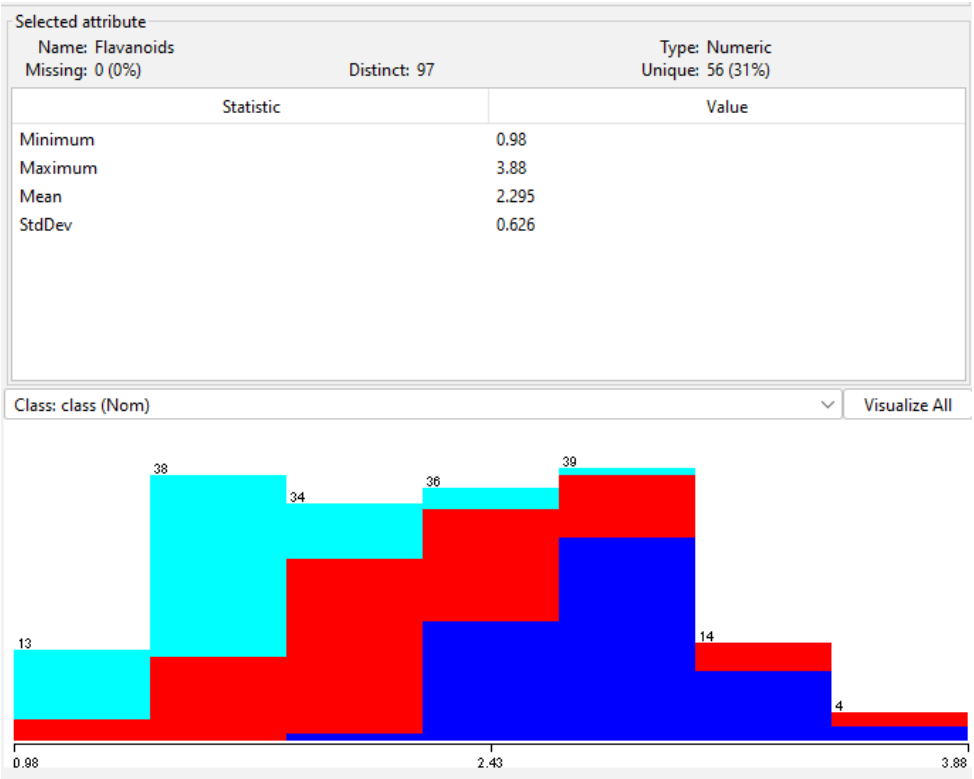

- **Magnésio**: O magnésio presente na uva está diretamente relacionado com o magnésio existente no solo de cultivo. Este composto vai estar mais presente em vinhos tintos, dado a sua produção uma vez que existe um maior contacto com a parte sólida da uva, local onde este se encontra. A presença deste nos vinhos tintos, vai ser fundamental para a fermentação malolática, visto que melhora o seu arranque e todo o processo.




- **Total de fenóis:** são responsáveis pela qualidade das uvas e vinhos. Este tipo de compostos varia de acordo com diversos fatores, tais como, o clima, o cultivo, a fermentação, a natureza do solo, entre outros. A distribuição dos compostos fenólicos nos vinhos brancos e tintos é diferente, sendo apresentados os vinhos brancos menores quantidades. A principal função destes compostos é serem antioxidantes, sendo que a sua atividade está diretamente relacionada com a sua estrutura química. Os fenóis podem se dividir em dois grandes grupos: os fenóis flavonoides e os fenóis não flavonoides.



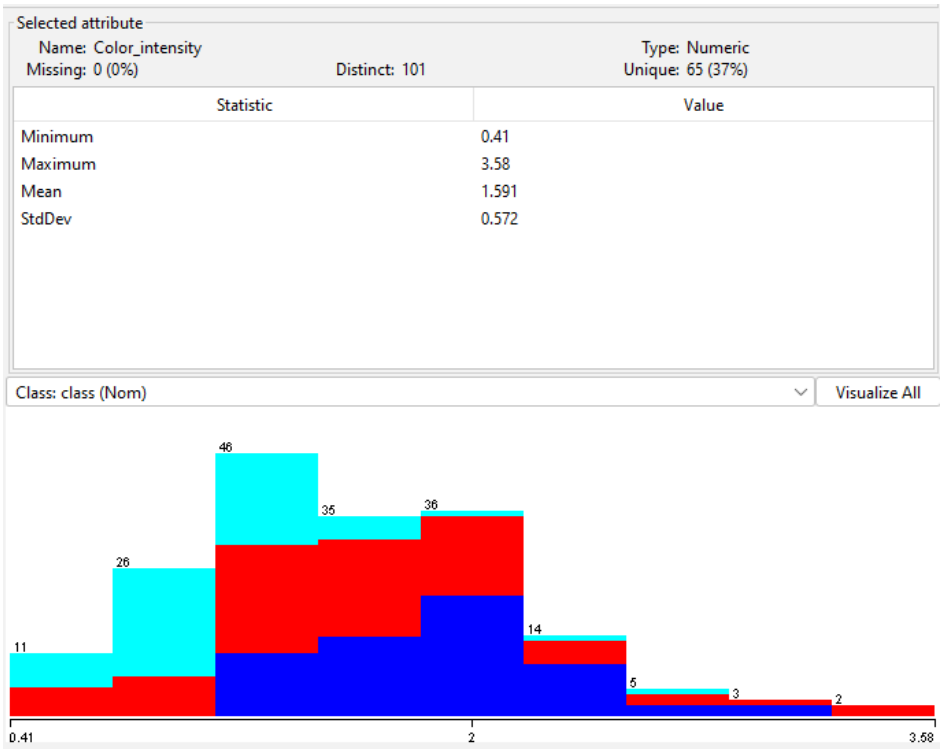

- **Flavonoides:** os flavonoides constituem o maior grupo de compostos poli fenólicos, estando estes diretamente associados às propriedades organoléticas do vinho, tais como o sabor, a cor e o aroma. Tal como a quantidade de fenóis, a quantidade de flavonoides varia de acordo com as condições ambientais e com o amadurecimento. A sua presença na uva ocorre, exclusivamente, nas partes sólidas (casca, grainhas e caule), sendo que a sua transferência para o mosto e vinho ocorre devido à maceração.



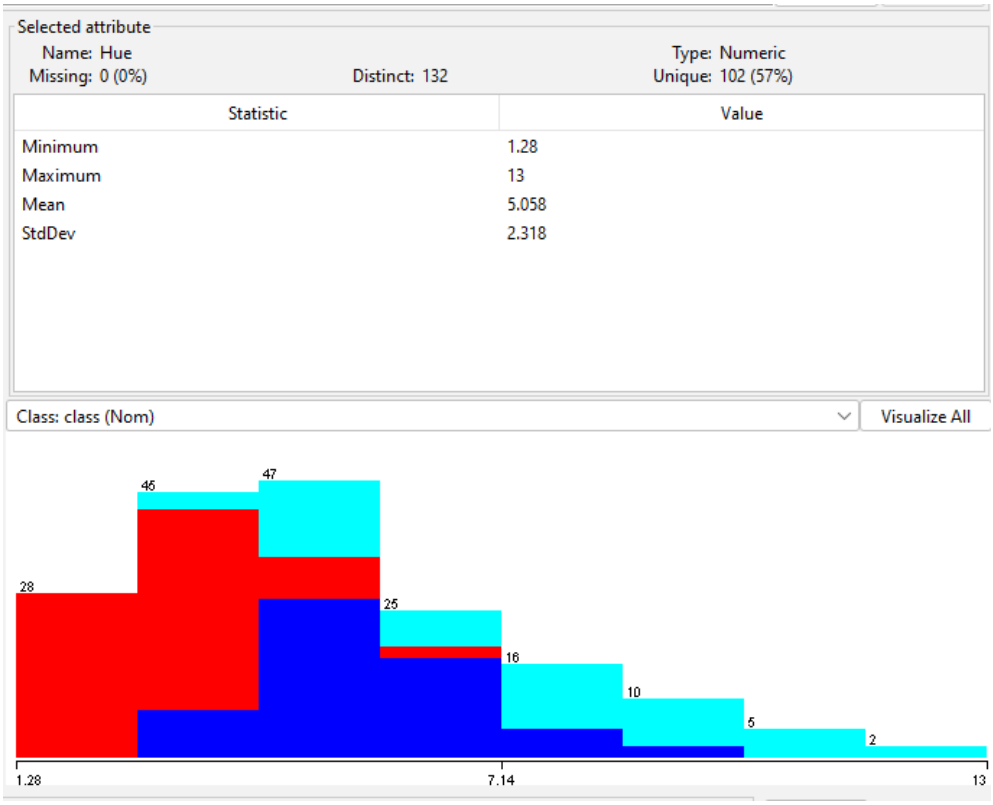

- **Não Flavonoides:** a este grupo de não flavonoides, vão pertencer todos os compostos poli fenólicos que não se encontrem no grupo anterior, como é o caso dos ácidos hidroxibenzóicos e dos estilbenos. Os não flavonoides encontram-se distribuídos pela casca e polpa das uvas, havendo uma diminuição das suas quantidades, há medida que existe o amadurecimento das uvas e na fermentação. Este tipo de constituintes é conhecido por realçar e estabilizar a cor dos vinhos tintos, sendo que contribuem também para o sabor do vinho e, no caso do resveratrol, contribui para as atividades biológicas.


- **Proantocianidinas:** as proantocianidinas, são vulgarmente conhecidas como taninos condensados. Estas são compostos poliméricos que dão origem às antocianinas, composto principal pela coloração do vinho. Quanto à quantidade destes compostos que é possível encontrar-se nas uvas, depende essencialmente da sua localização, sendo que a sua transferência para o vinho se dá devido aos processos usados na vinificação, como é o caso, do esmagamento, maceração e fermentação. Apesar da sua quantidade depender da sua localização, as proantocianidinas são responsáveis pelas características sensoriais. Desempenhando assim um papel importante no que diz respeito ao envelhecimento do vinho devido as suas capacidades de oxidar, condensar e polimerizar compostos.

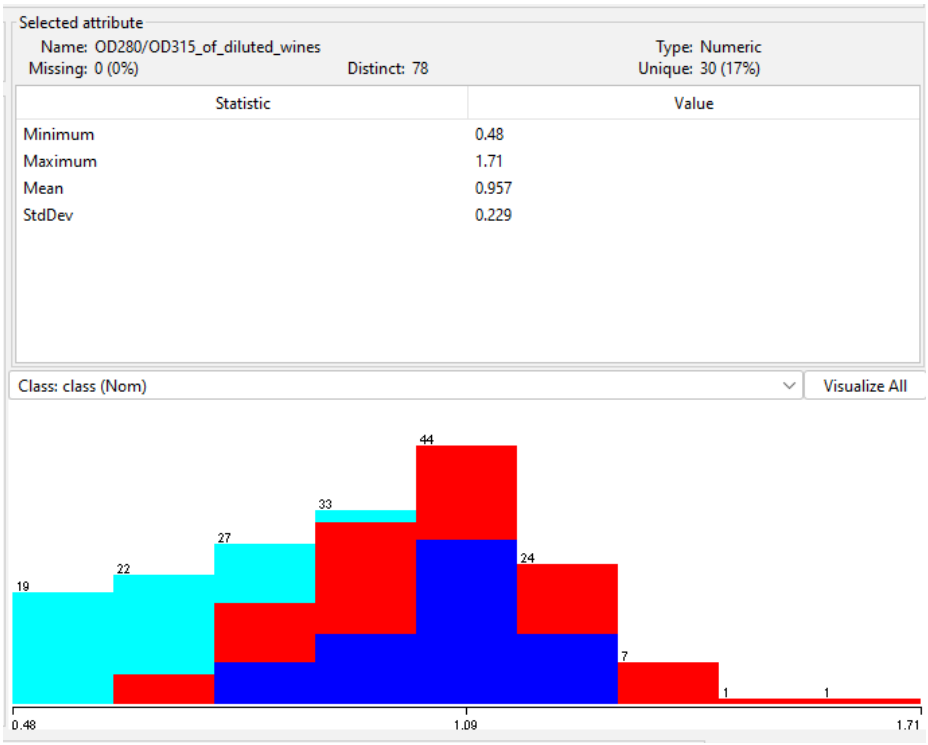

- **Intensidade da Cor:** ao observar-se a intensidade do vinho é possível obterem-se informações relativamente ao corpo do vinho, dado que quanto mais escuro é mais encorpado será. É ainda possível, verificar qual a idade e a acidez do vinho, uma vez que quanto mais brilhante mais jovem e mais ácido é.



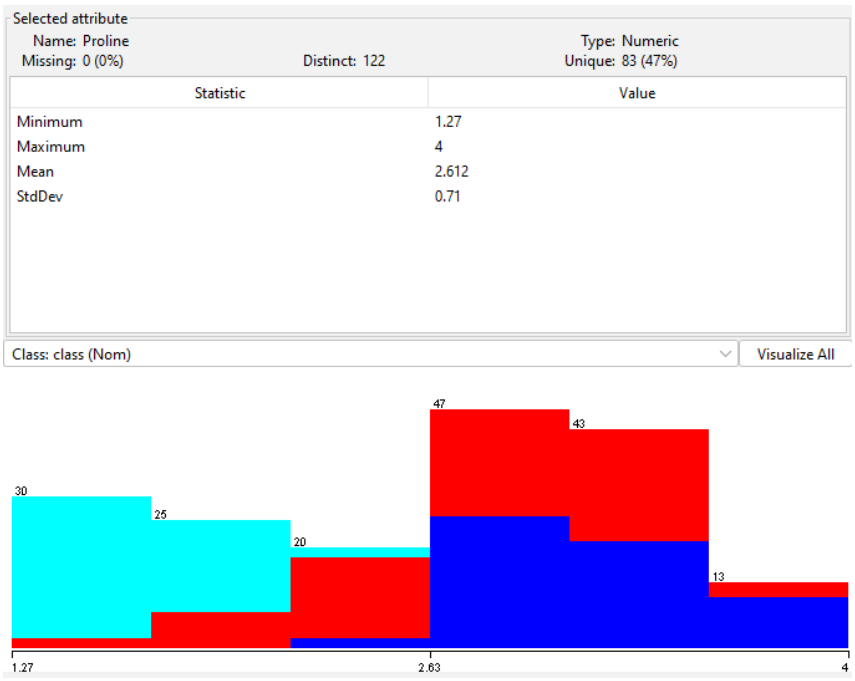
- **Tonalidade:** em relação à tonalidade do vinho, esta está relacionada com a cor das uvas, sendo que pode sofrer alterações durante os processos de vinificação e o envelhecimento. Pelo tom do vinho, pode ser possível perceber-se quais os métodos usados e as uvas, dado que as castas possuem cores diferentes entre si. Esta informação permite assim antever quais as características que cada vinho possui a nível de sabor.




- **Diluição de Vinhos pelo coeficiente OD280/OD315:** A diluição de vinhos usando o fator OD280/OD315, a densidade ótica das proteínas a 280 nm a dividir pela densidade ótica a 315 nm, permite que se determine a concentração de proteínas em vinhos diluídos. Isto vai ser útil para determinar o teor de proteína existente em cada vinho.




- **Prolina**: a prolina vai ser usada como indicador da maturação, dado que a quantidade desta aumenta brutalmente nesta fase devido a saturação de proteínas nas uvas, o que faz com que não haja a proteólise. No entanto, sendo um aminoácido, será uma fonte de nutrientes para as leveduras úteis para a fermentação. Esta é um dos constituintes do vinho que mais pode sofrer alteração de um ano para o outro, visto que esta está diretamente relacionada com as formas de cultivo usadas.



## Pré-processamento

Como o atributo classe ocupava a primeira coluna usei o filtro Reorder para colocá-lo na última. Assim o Weka mostra em cada atributo a quantidade de instâncias pertencentes a cada classe, o que resultou na imagem acima. Contudo, o agrupamento tem como premissa a descoberta de dados sem que estes estejam rotulados, portanto, removi o atributo classe para a execução dos algoritmos.

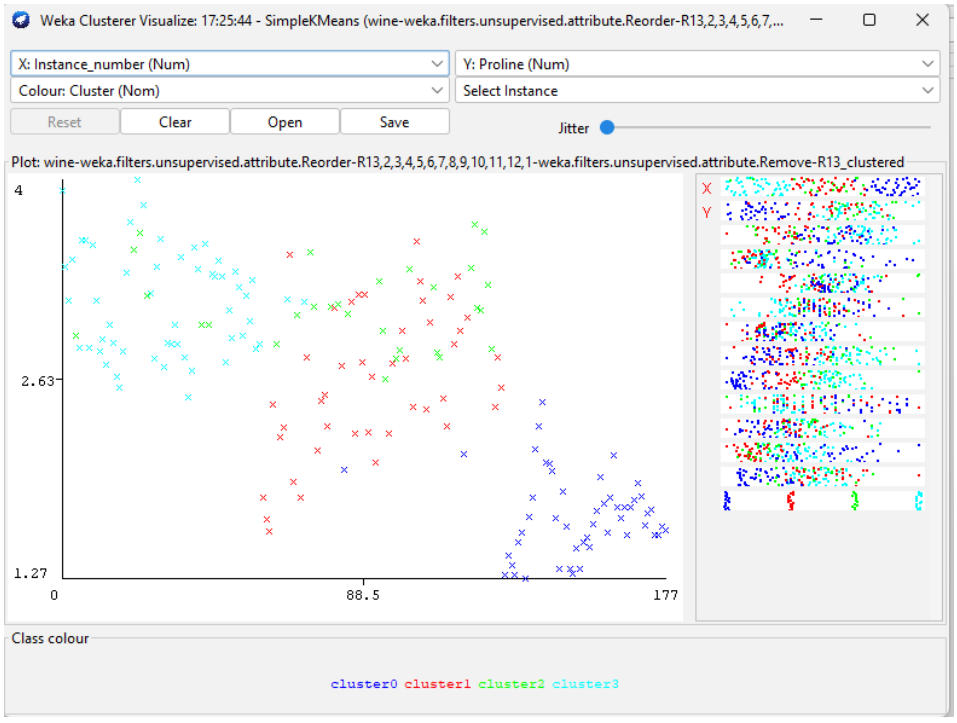
## Agrupamento

O algoritmo K-means apresentou os seguintes resultados:

- Com 3 centroides


- Com 2 centroides

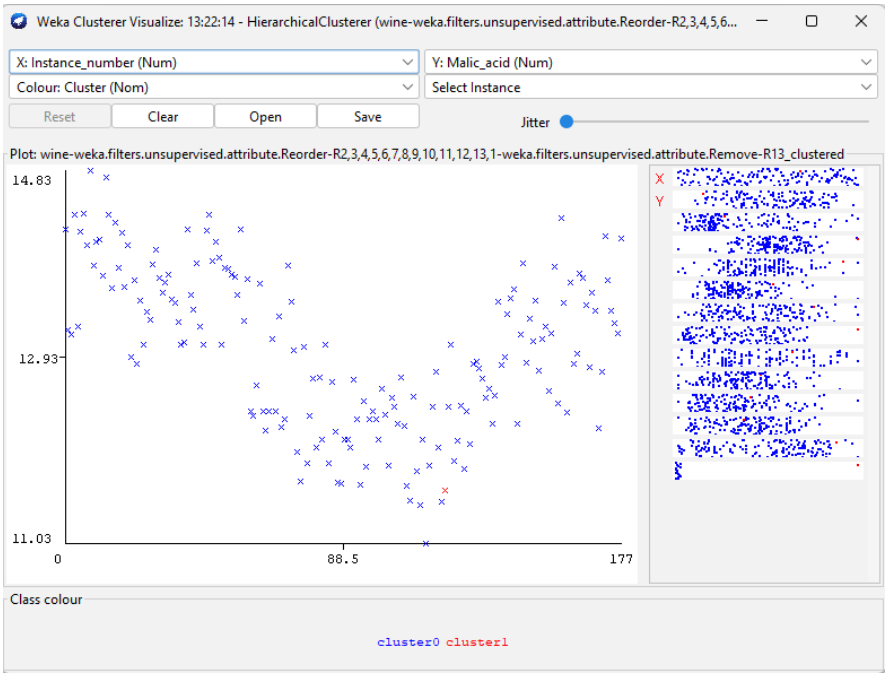

- Com 4 centroides

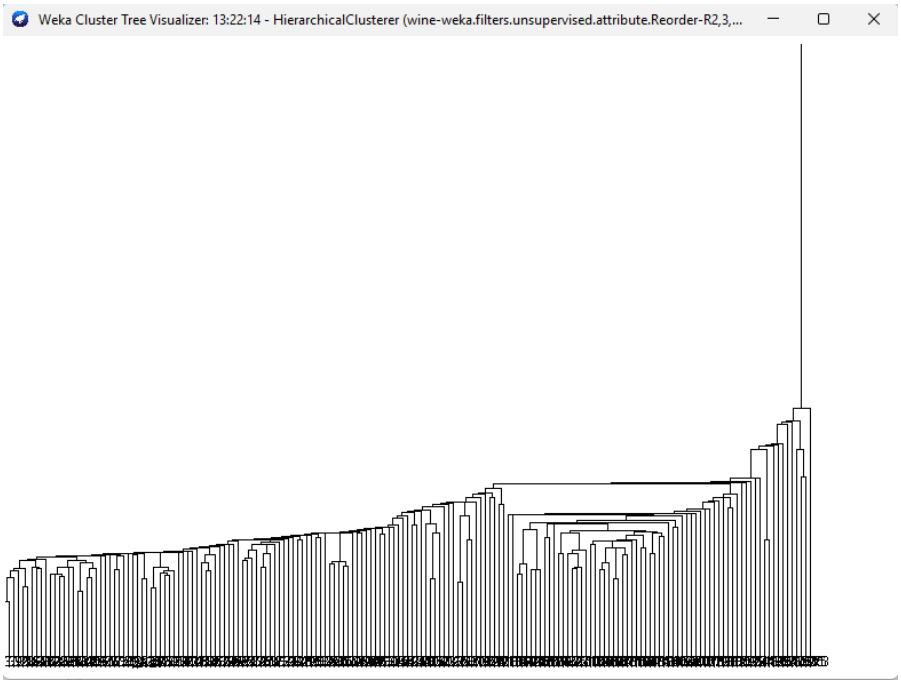



| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,70 | 0,61 | 0,62 |
| Pureza | 0,50 | 0,55 | 0,56 |
| Coeficiente de Jaccard | 0,12 | 0,11 | 0,06 |

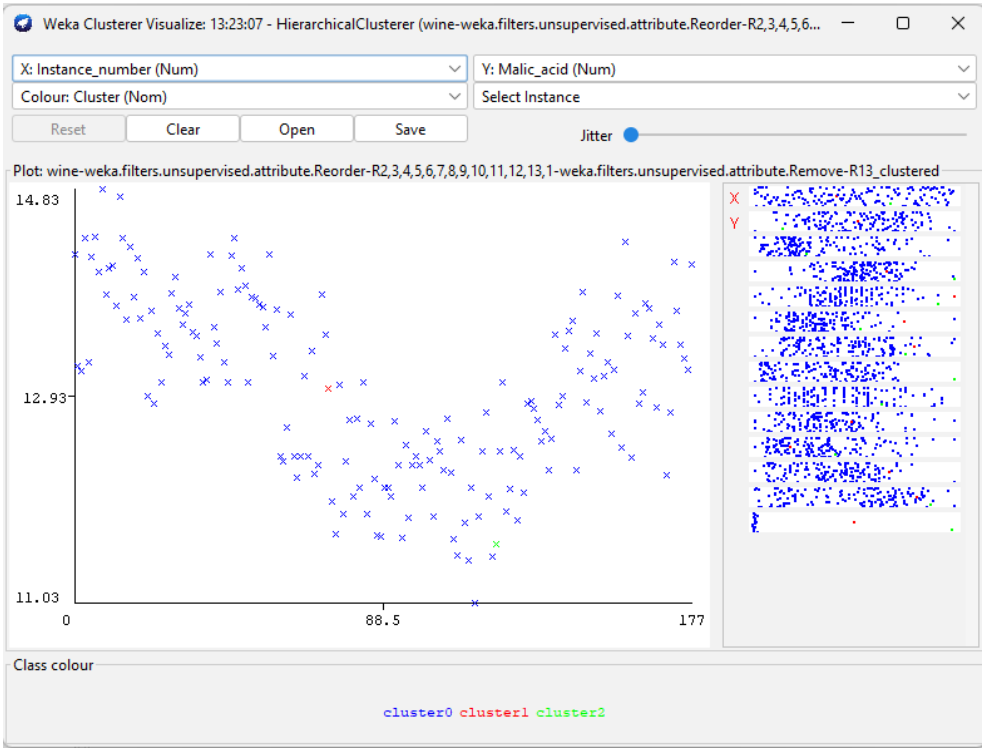
O algoritmo Hierárquico Single Link apresentou os seguintes resultados:

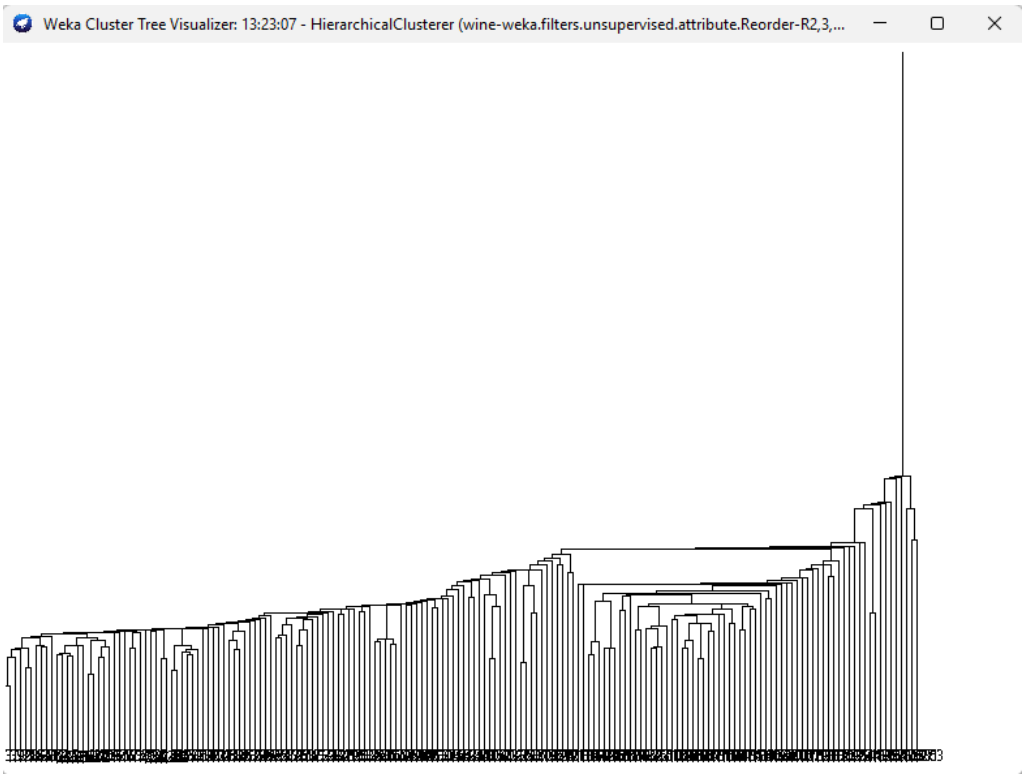
- Com 2 centroides



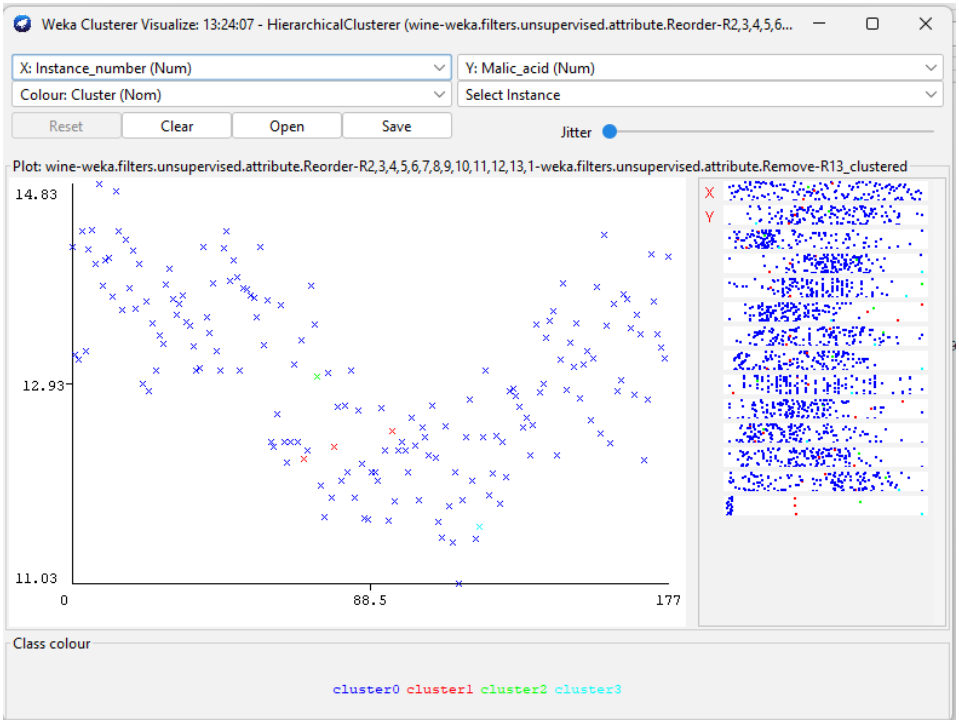



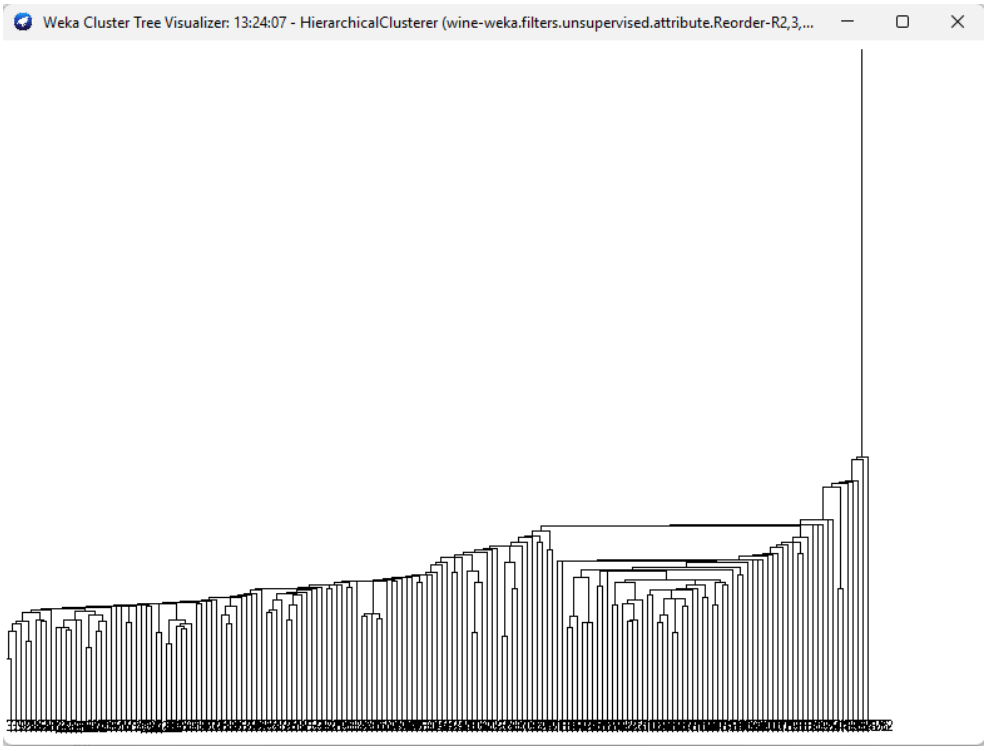
- Com 3 centroides



- Com 4 centroides

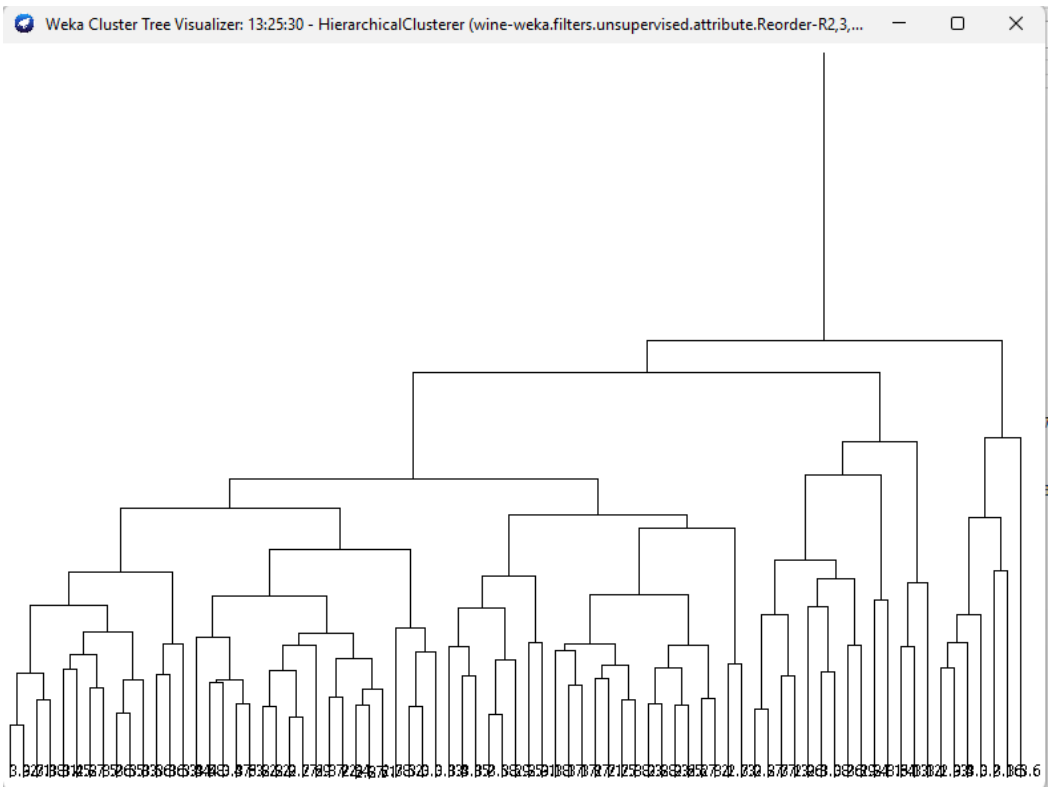


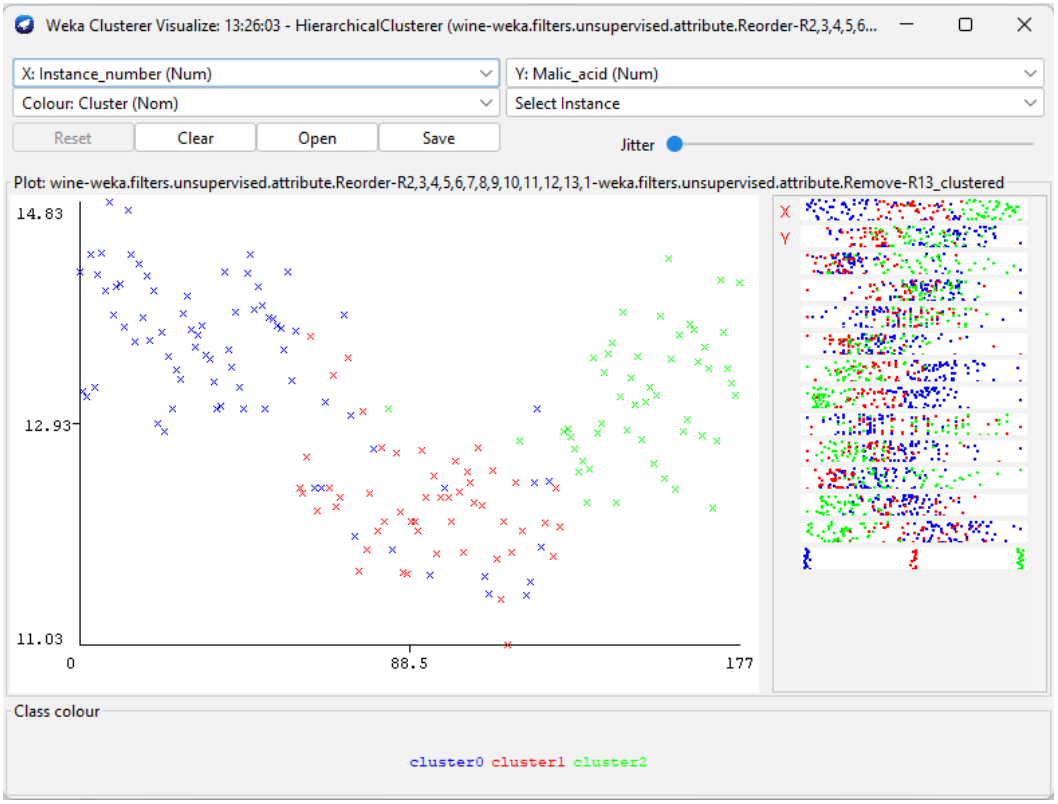

| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,83 | 0,78 | 0,68 |
| Pureza | 0,39 | 0,39 | 0,39 |
| Coeficiente de Jaccard | 0,24 | 0,20 | 0,20 |

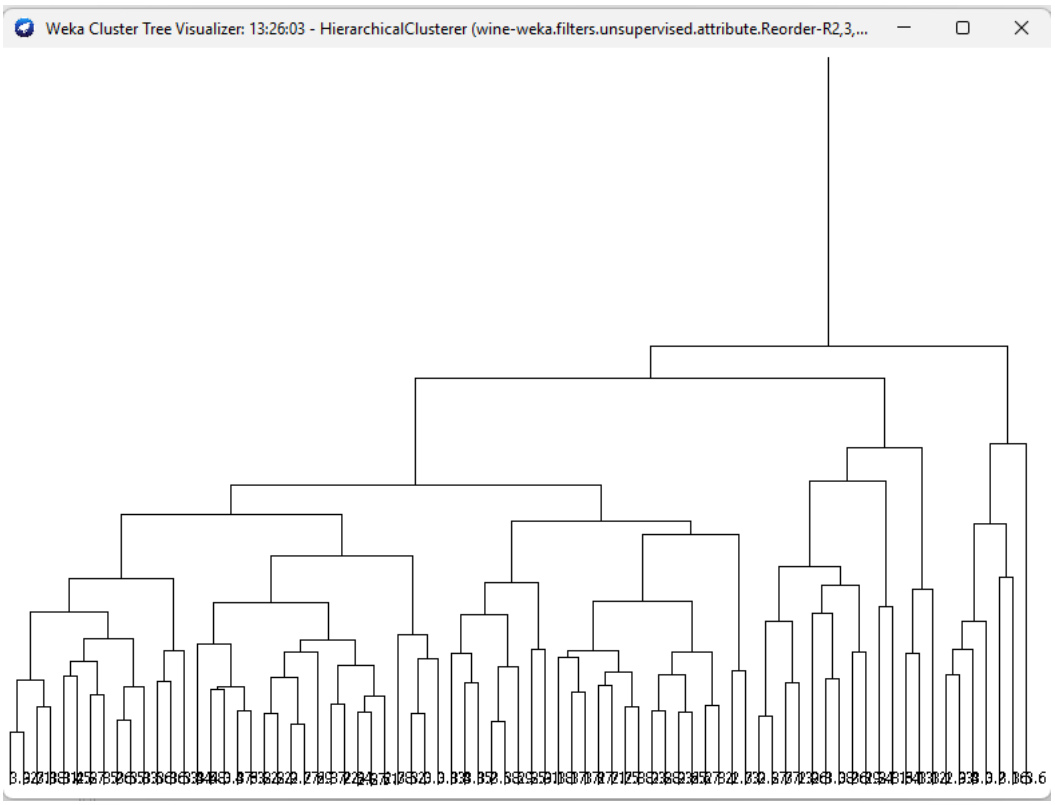
O algoritmo Hierárquico Complete Link apresentou os seguintes resultados:

- Com 2 centroides

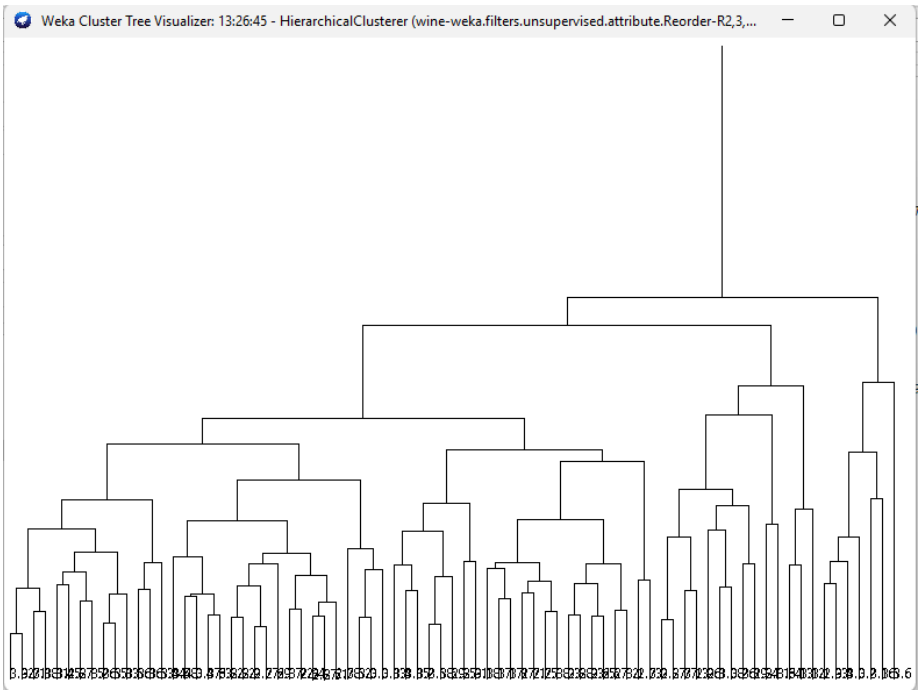



- Com 3 centroides



- Com 4 centroides

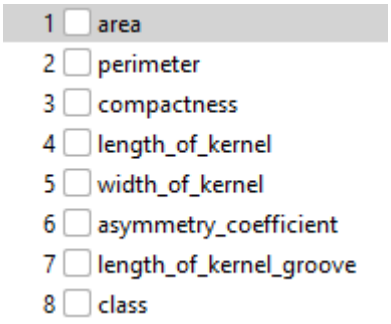

| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,72 | 0,68 | 0,40 |
| Pureza | 0,49 | 0,50 | 0,50 |
| Coeficiente de Jaccard | 0,32 | 0,12 | 0,28 |

# Análise da Base Seeds

## Informações

A base Seeds é constituída por dados de grãos pertencentes a três variedades de trigo: Kama, Rosa e Canadense, com 70 elementos cada, selecionados aleatoriamente para o experimento. A visualização de alta qualidade da estrutura interna do kernel foi detectada usando uma técnica de raios-X suave. É não destrutivo e consideravelmente mais barato do que outras técnicas de imagem mais sofisticadas, como microscopia de varredura ou tecnologia a laser. As imagens foram registradas em chapas de raios X 13x18 cm KODAK. Os estudos foram conduzidos usando grãos de trigo colhidos combinados provenientes de campos experimentais, explorados no Instituto de Agrofísica da Academia Polonesa de Ciências em Lublin.

O Seeds Data Set possui 210 instâncias com 7 atributos, os quais são inteiros ou reais. Não há valores ausentes e os atributos são descritos da seguinte forma:




- **Classe:** classes do grão Kama(0), Rosa(1) e Canadense(2)

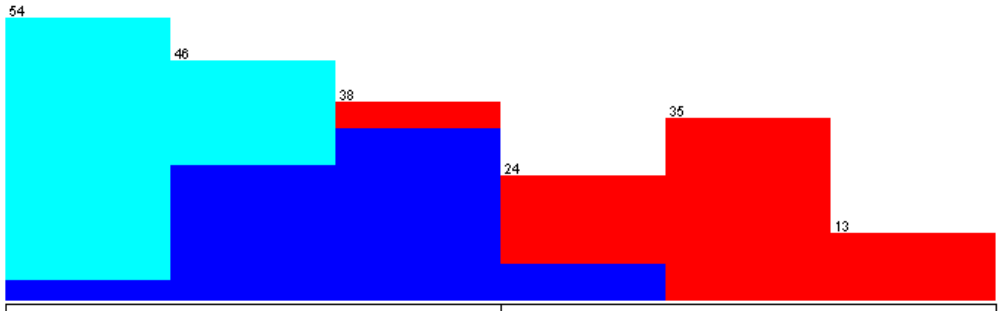
- **Área**

Selected attribute

Name: area
Missing: 0 (0%)
Distinct: 193
Type: Numeric
Unique: 179 (85%)

| Statistic | Value |
|---|---|
| Minimum | 10.59 |
| Maximum | 21.18 |
| Mean | 14.848 |
| StdDev | 2.91 |

Class: class (Nom) Visualize All



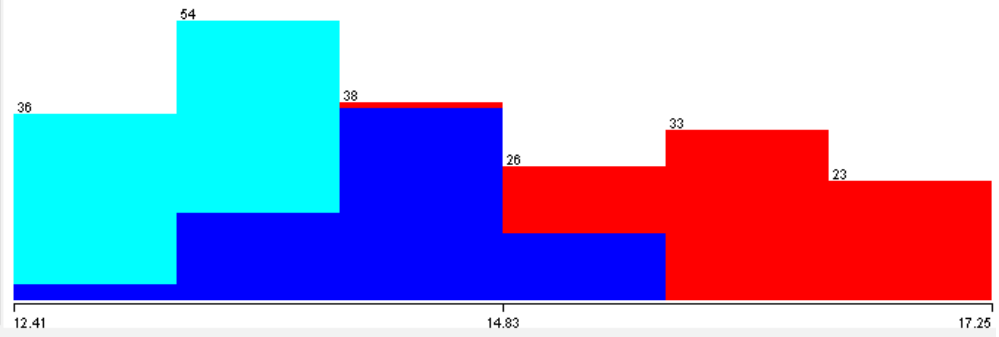

- **Perímetro**

Selected attribute

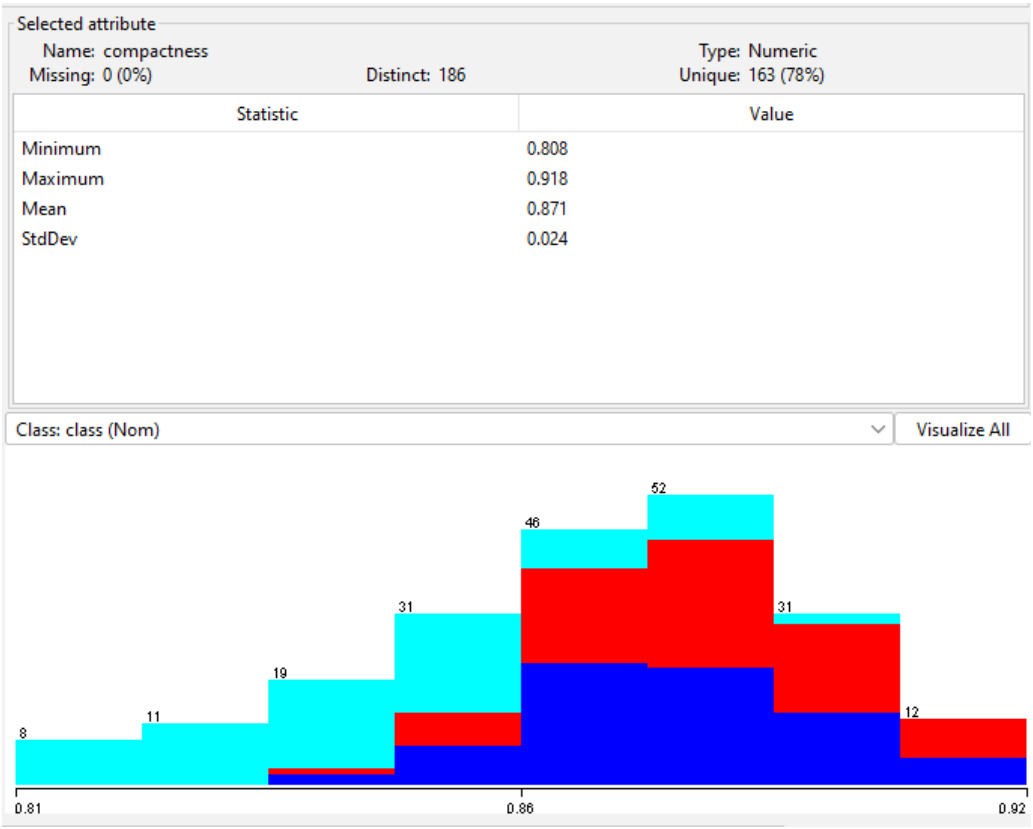
Name: perimeter
Missing: 0 (0%)
Distinct: 170
Type: Numeric
Unique: 137 (65%)

| Statistic | Value |
|---|---|
| Minimum | 12.41 |
| Maximum | 17.25 |
| Mean | 14.559 |
| StdDev | 1.306 |

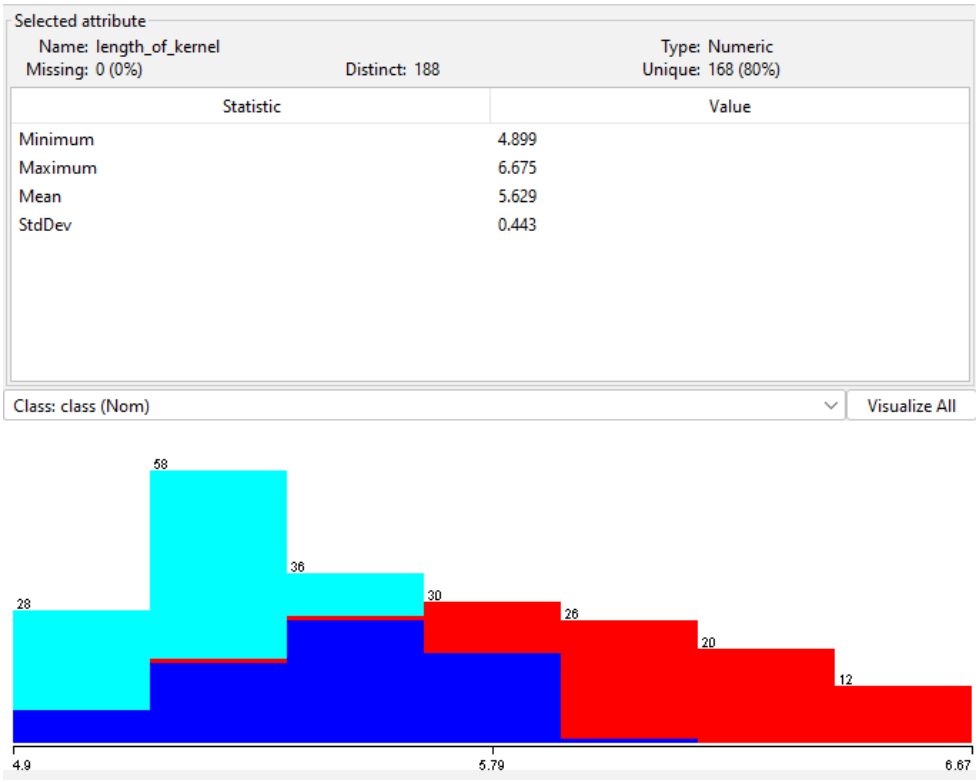
Class: class (Nom) Visualize All



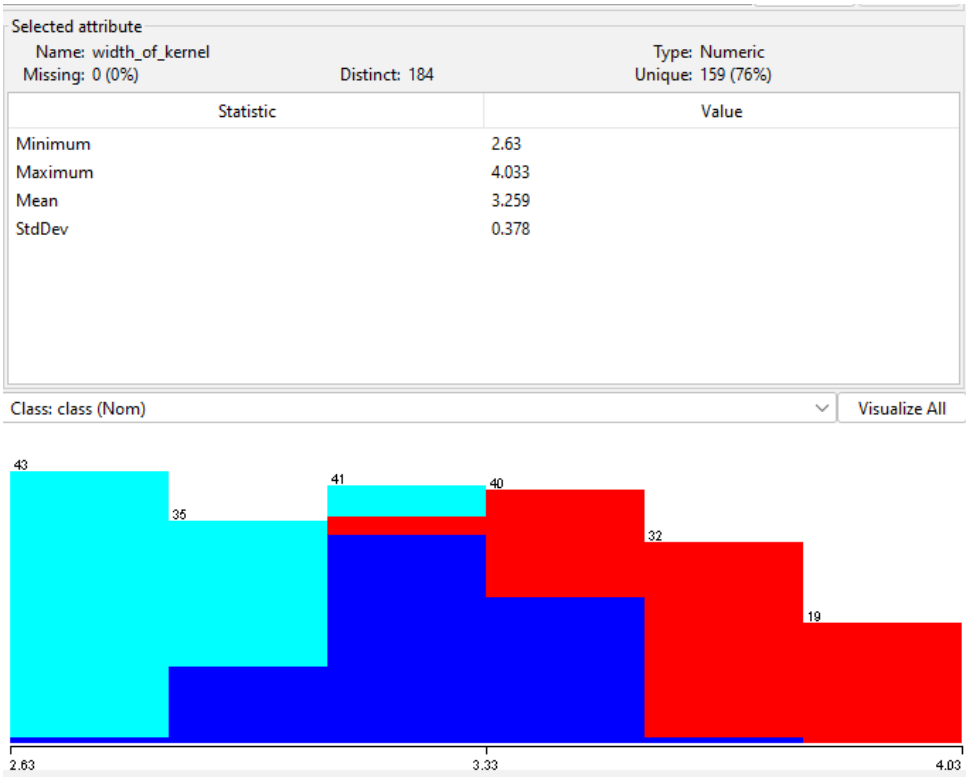
- **Compacidade**



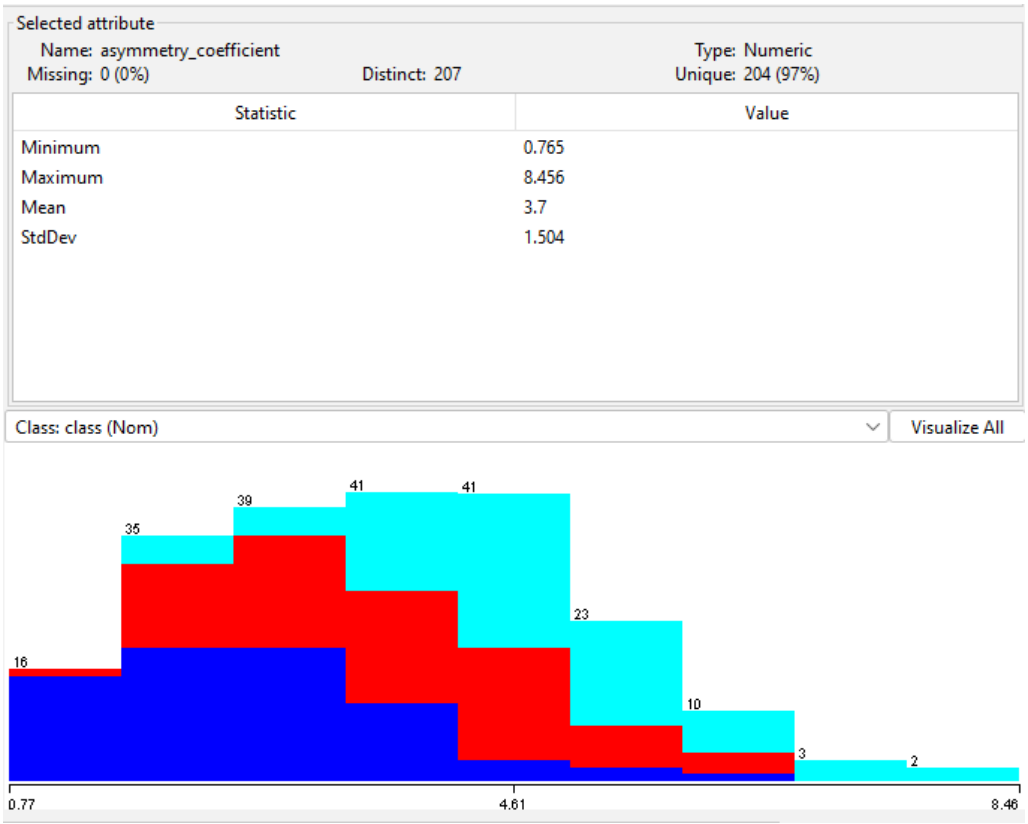

- **Comprimento do Grão**



- **Largura do Grão**




- **Assimetria**



- **Comprimento da Plantação de Grãos**

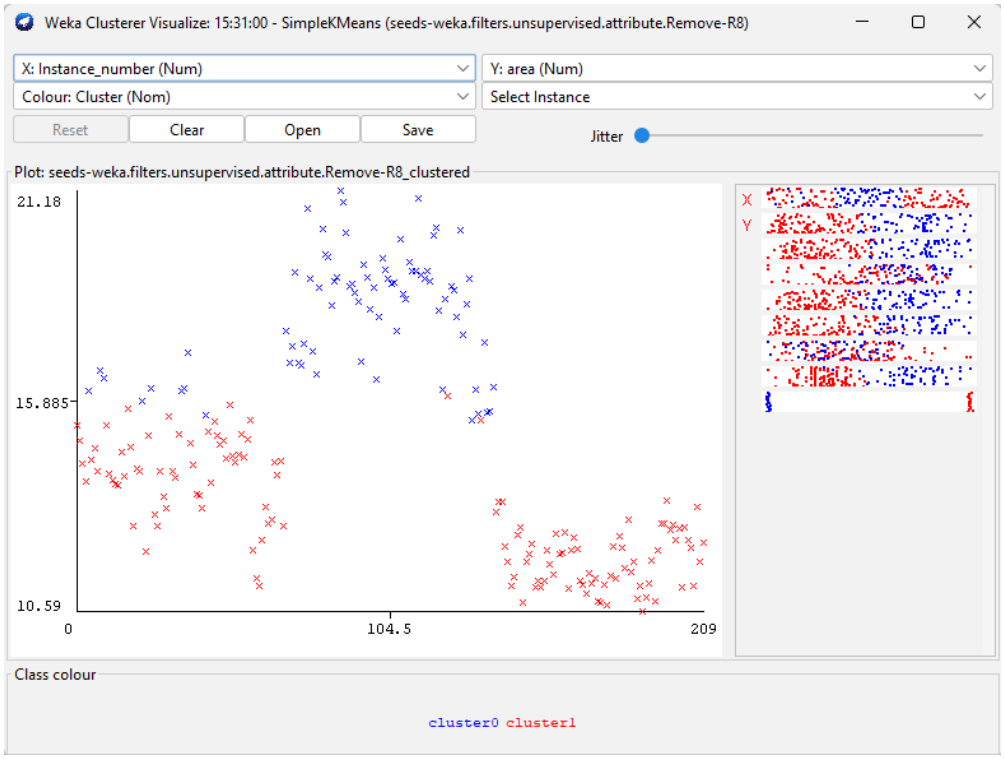

**Pré-processamento**

As instâncias da classe estavam descritas no intervalo de 1 a 3, para facilitar cálculos das medidas de avaliação modifiquei-as para ficarem no intervalo de 0 a 2. Para a execução dos algoritmos de agrupamento desconsiderei o atributo classe será usado na avaliação dos resultados.

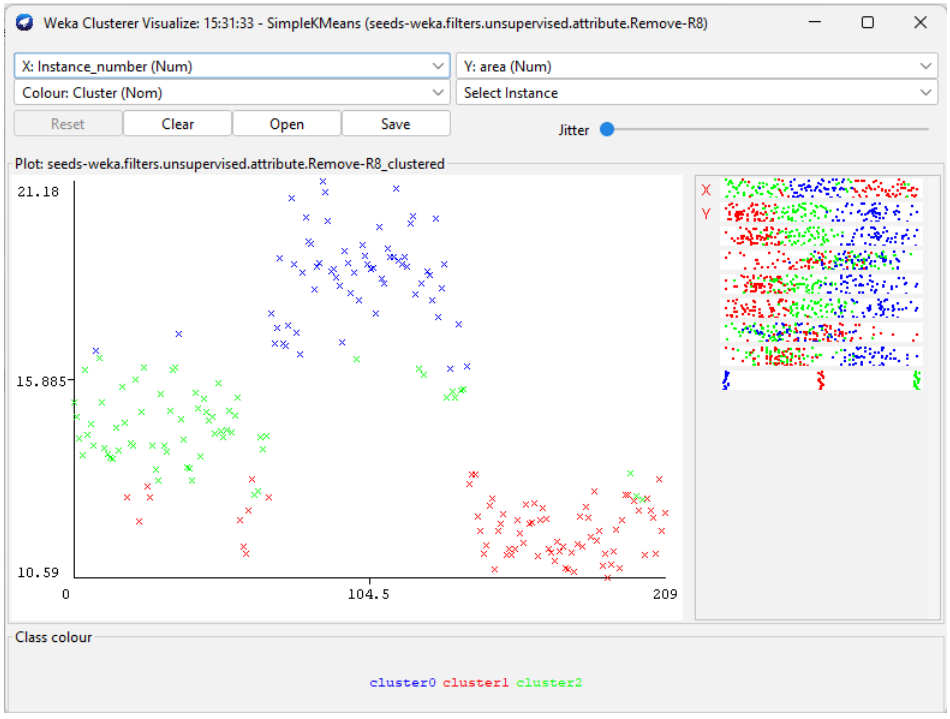
**Agrupamento**

O algoritmo K-means apresentou os seguintes resultados:

- Com 2 centroides




- Com 3 centroides

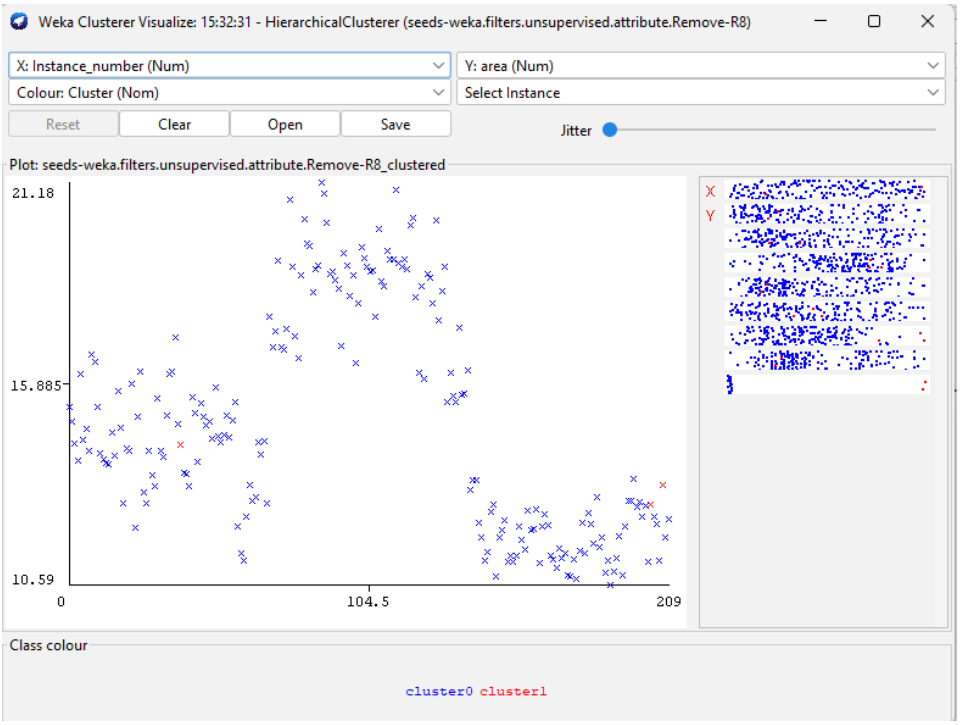


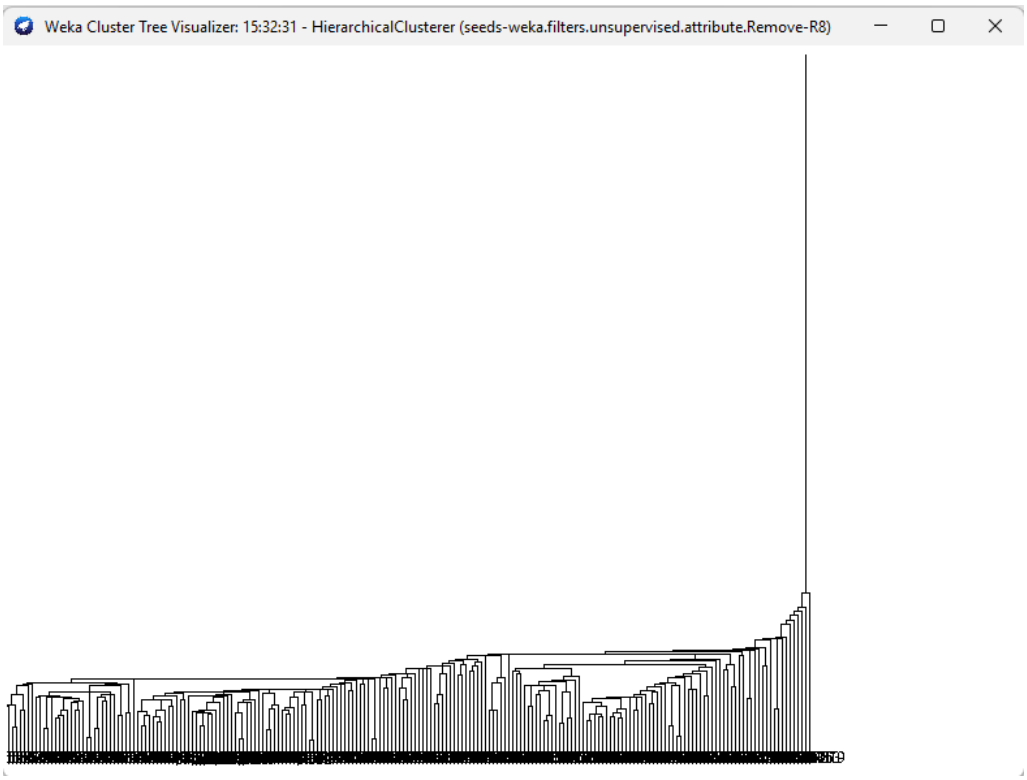
- Com 4 centroides


| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,71 | 0,66 | 0,57 |
| Pureza | 0,66 | 0,89 | 0,86 |
| Coeficiente de Jaccard | 0,42 | 0,16 | 0,28 |

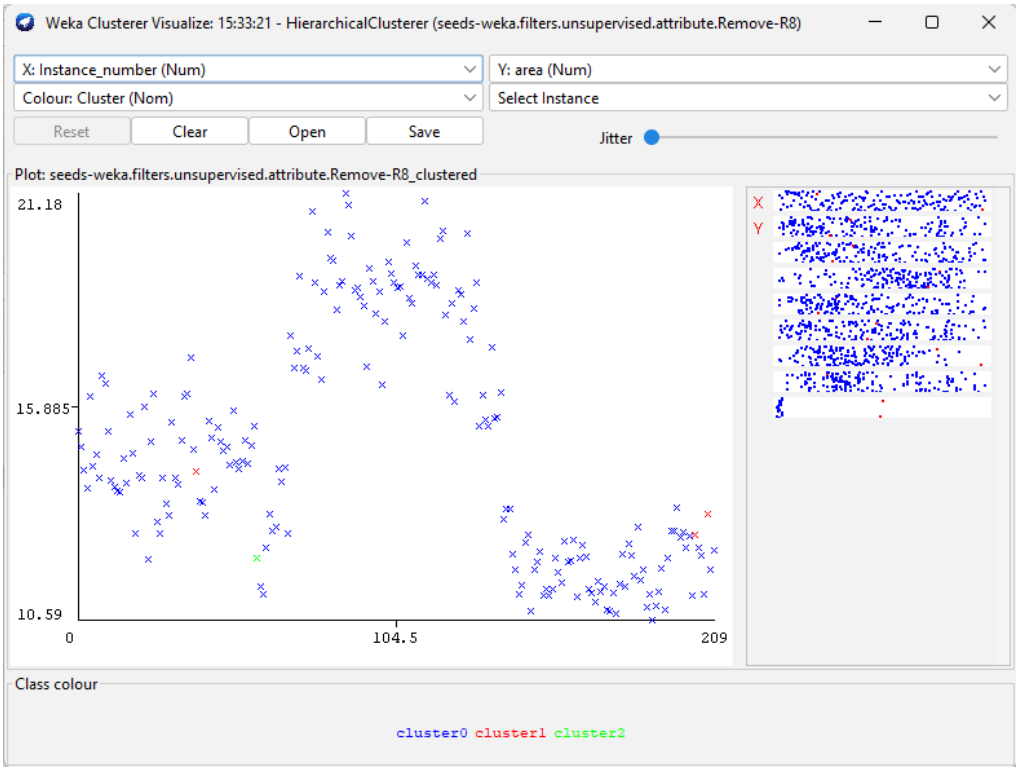
O algoritmo Hierárquico Single Link apresentou os seguintes resultados:

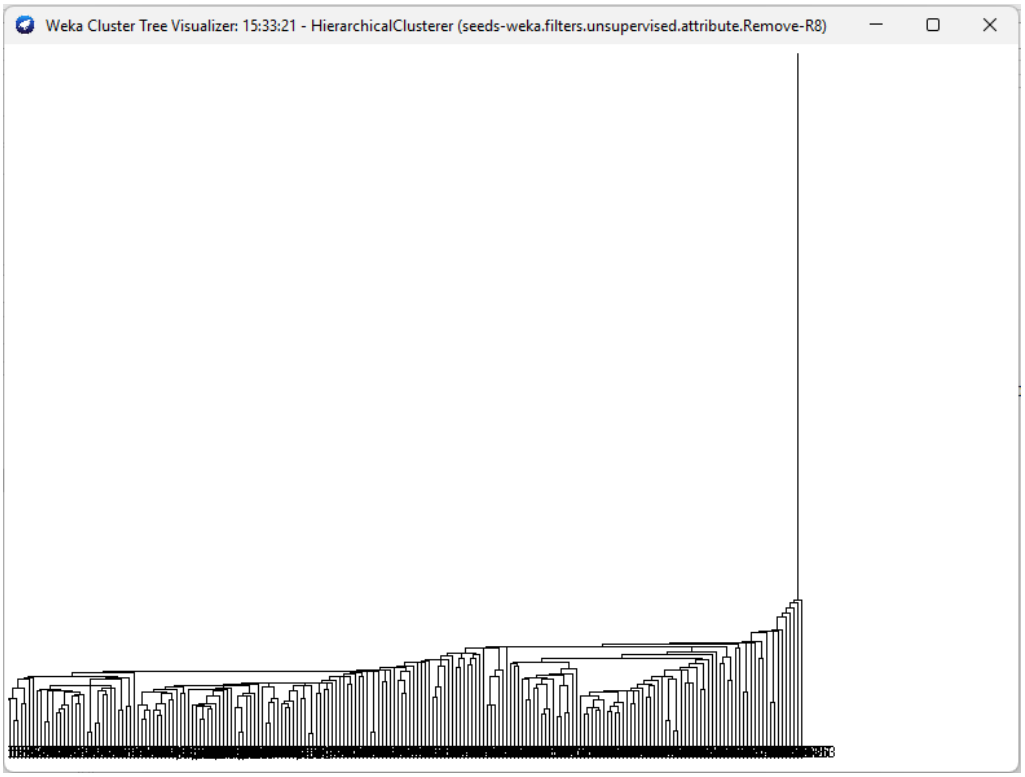
- Com 2 centroides


Weka Clusterer Visualize: 15:32:31 - HierarchicalClusterer (seeds-weka.filters.unsupervised.attribute.Remove-R8)
X: Instance_number (Num)
Y: area (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: seeds-weka.filters.unsupervised.attribute.Remove-R8_clustered
21.18
15.885
10.59
0
104.5
209
X
Y
Class colour
cluster0 cluster1

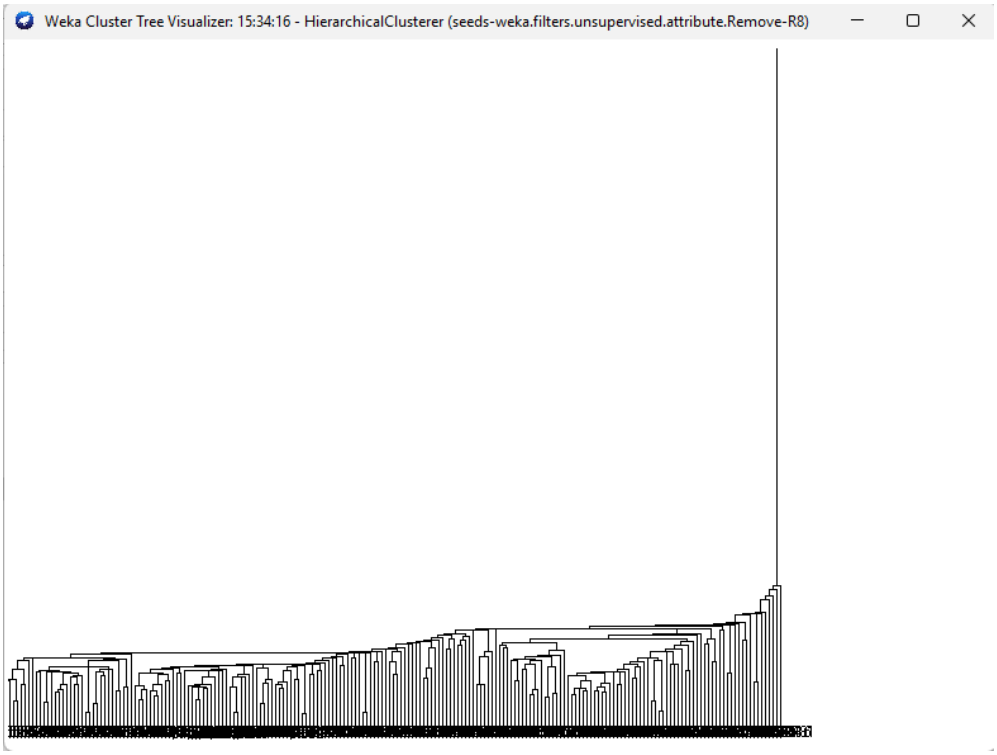

Weka Cluster Tree Visualizer: 15:32:31 - HierarchicalClusterer (seeds-weka.filters.unsupervised.attribute.Remove-R8)

- Com 3 centroides



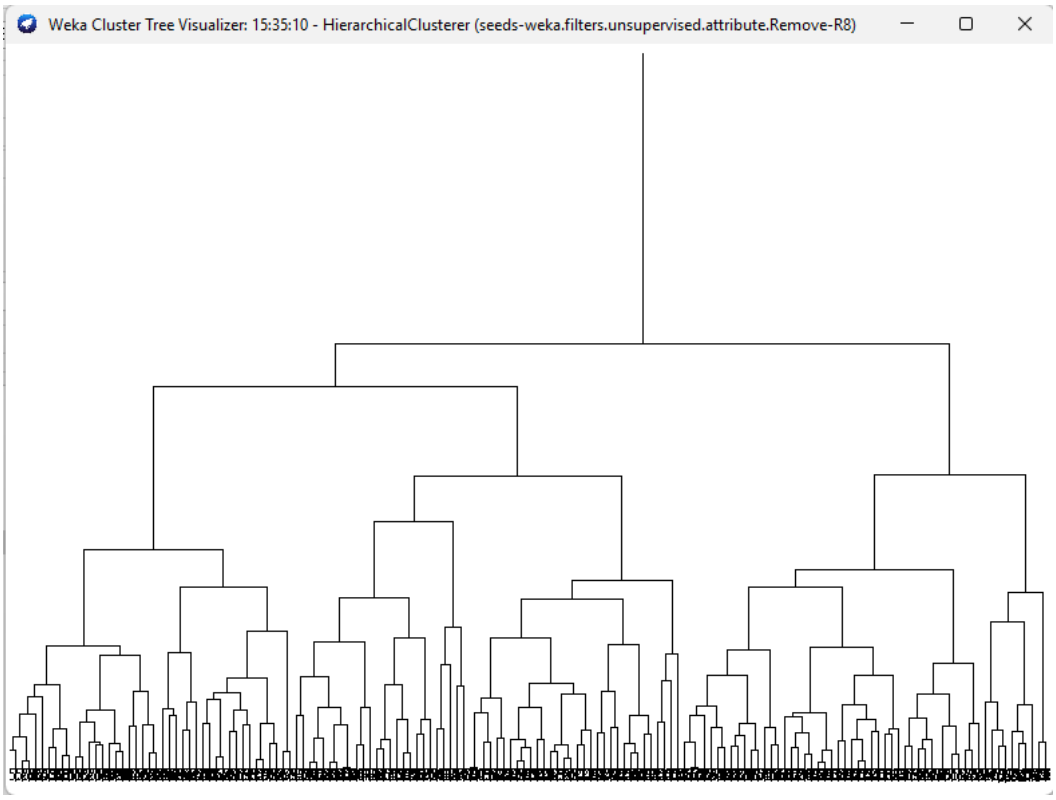
- Com 4 centroides

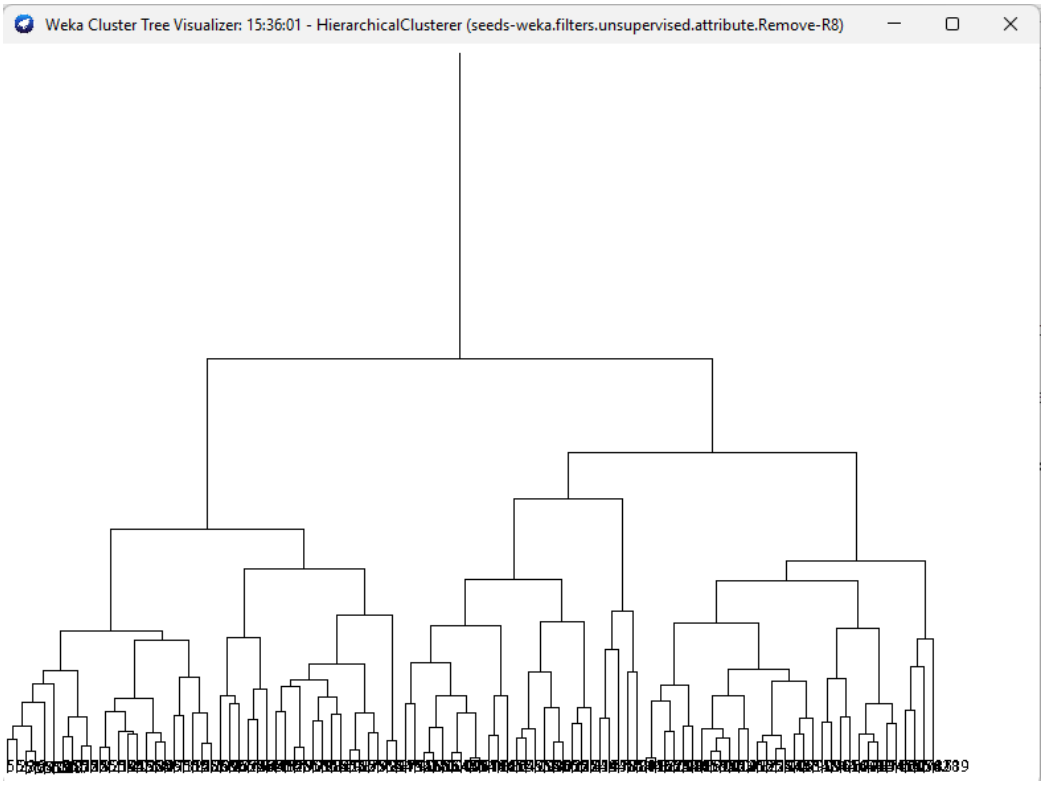

| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,22 | -0,02 | -0,25 |
| Pureza | 0,34 | 0,37 | 0,37 |
| Coeficiente de Jaccard | 0,2 | 0,2 | 0,19 |

O algoritmo Hierárquico Complete Link apresentou os seguintes resultados:

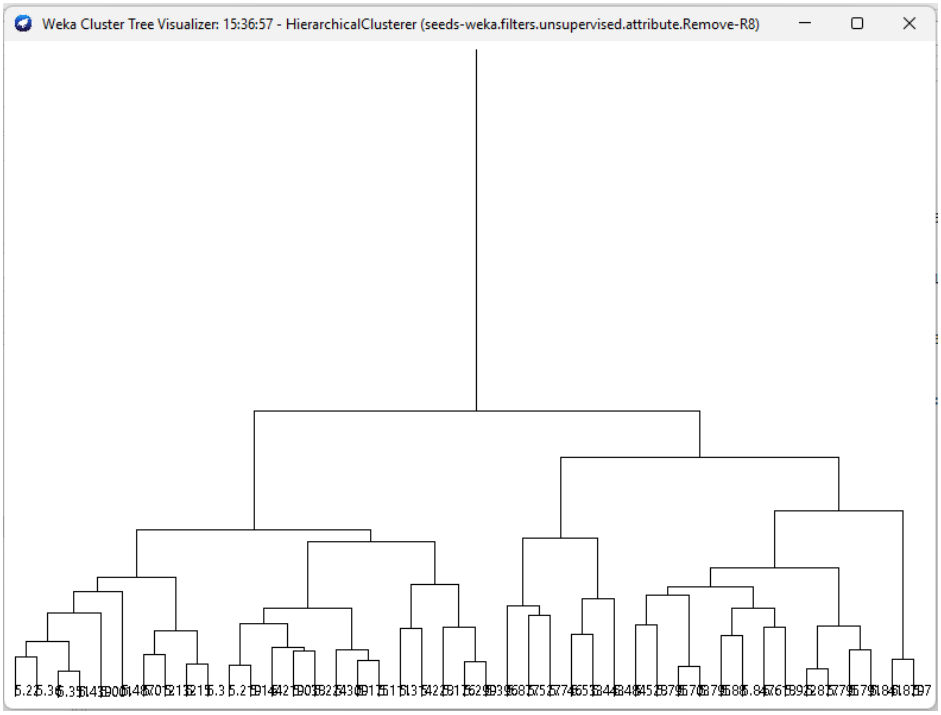
- Com 2 centroides

- Com 3 centroides

- Com 4 centroides


| | K=2 | K=3 | K=4 |
|---|---|---|---|
| Largura da Silhueta | 0,62 | 0,61 | 0,52 |
| Pureza | 0,55 | 0,80 | 0,80 |
| Coeficiente de Jaccard | 0,38 | 0,18 | 0,59 |

## Comparação dos Algoritmos

Neste trabalho foram usadas duas bases de dados de tamanhos similares. A primeira Wine possui 178 instâncias e a Seeds 210 instâncias. Para avaliar qual algoritmo teve o melhor desempenho analisamos os resultados que cada um teve nas duas bases de dados. Como três configurações dos algoritmos foram usadas estou considerando K= 3 para as análises finais.

- Largura da Silhueta

| | Wine | Seeds |
|---|---|---|
| K-Means | 0,61 | 0,66 |
| Hierárquico Single Link | 0,78 | -0,02 |
| Hierárquico Complete Link | 0,68 | 0,61 |

- Grau de Pureza

| | Wine | Seeds |
|---|---|---|
| K-Means | 0,55 | 0,89 |
| Hierárquico Single Link | 0,39 | 0,37 |
| Hierárquico Complete Link | 0,50 | 0,80 |

- Coeficiente de Jaccard

| | Wine | Seeds |
|---|---|---|
| K-Means | 0,11 | 0,16 |
| Hierárquico Single Link | 0,20 | 0,2 |
| Hierárquico Complete Link | 0,12 | 0,18 |

O Algoritmo com maior valor para largura da silhueta, grau de pureza e coeficiente de Jaccard, na base de dados Wine é o Hierárquico Single Link; e na Seeds é o K-Means. Podemos inferir que o conjunto de dados da base Wine produziu grupos mais homogêneos e com maior diferença entre os outros clusters. Enquanto os da base Seeds contribuíram para a formação de grupos aninhados.

## Referências Bibliográficas

MATOS, A. P. Cruz. “APLICAÇÃO DE FERRAMENTAS QUIMIOMÉTRICAS PARA CATEGORIZAÇÃO DE VINHOS ATRAVÉS DOS SEUS COMPONENTES QUÍMICOS”. Universidade de Coimbra, 2021.

Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson,2006.