Utilização de “K Means” para realizar a “clusterização” de vinhos tintos

Pedro Peixoto
Faculdade de Engenharia Elétrica
e de Computação - UNICAMP
p219818@dac.unicamp.br

Rafael Antonio Chinelatto
Faculdade de Engenharia Elétrica
e de Computação - UNICAMP
r187354@dac.unicamp.br

Ronald Gabriel Ferreira da Silva
Faculdade de Engenharia Elétrica
e de Computação - UNICAMP
r196815@dac.unicamp.br

***Resumo*—Esta atividade busca explorar conceitos de aprendizado de máquina relacionados a problemas de clusterização, aplicando técnicas de aprendizado não supervisionado para analisar e interpretar a qualidade dos vinhos tintos.**

I. Introdução

Muitos são os casos em que informações devem ser retiradas de um conjunto de dados, mas não possuem resultados pré-definidos para o modelo que será utilizado. Em um cenário onde a quantidade de dados cresce exponencialmente, a habilidade de extrair padrões de dados que não são supervisionados é valiosa. Portanto, o estudo de técnicas de aprendizado não supervisionado, como a clusterização e a redução de dimensionalidade, pode fornecer ferramentas cruciais para análise de dados em diversas áreas, desde a segmentação de clientes até a descoberta de novas tendências em pesquisas científicas. Sendo assim, conhecer tais metodologias e explorar tal conceito têm se tornado cada vez mais relevantes para compreender o mercado.

Utilizando a base de dados [Red Wine Quality], os dados serão tratados da devida maneira para que possam ser explorados os conceitos propostos. Neste trabalho, serão trabalhadas as seguintes abordagens:

- Tratamento de Dados: Realização de pré-processamento da base de dados, transformando os dados para garantir que estejam adequados para análise.
- Exploração de Dados: Utilizar métodos gráficos para realizar uma análise inicial, identificando características e padrões relevantes dentro do conjunto de dados.
- Clusterização: Aplicar o algoritmo de clusterização K-means para identificar grupos dentro dos dados que possam revelar diferentes categorias de vinhos com base em suas características.
- Redução de Dimensionalidade: Empregar técnicas como Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados, facilitando a visualização e a interpretação dos clusters formados.
- Avaliação e Interpretação dos Resultados: Avaliar a eficácia das técnicas aplicadas, interpretando os clusters obtidos.

II. Metodologia

A base de dados possui um conjunto de 1599 vinhos, cada vinho possui 12 atributos registrados. A partir disso, é possível analisar o conjunto e realizar modificações para que o algoritmo de clusterização seja implementado e que sejam obtidos melhores resultados.

Para avaliar os dados do dataset de forma satisfatória, é necessário realizar o pré-processamento adequado. Inicialmente a base de dados foi modificada de forma que a escalas dos atributos fossem equivalentes. Portanto, a normalização foi realizada de tal maneira que os valores fossem de todos os atributos estivessem entre zero e um.

Os atributos podem ser avaliados a partir de um mapa de calor, que avalia a correlação entre eles. Atributos com alta correlação incrementam pouca informação e, dessa forma, o mapa de calor muitas vezes é relevante para reduzir a dimensão da base.

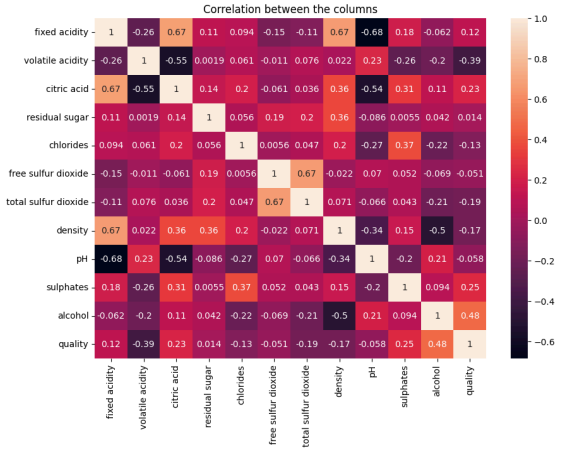

Figura 01- Mapa de calor do conjunto de dados

Observando o mapa de calor, existem alguns casos em que a correlação entre dois atributos é significativa. Tendo isso em vista, é pertinente supor que algumas classificações podem ser retiradas sem perder uma quantidade significativa de informação. Entretanto, a separação em componentes principais é tida como uma maneira mais adequada de descartar componentes do conjunto. Sendo assim, a atividade foi feita a partir de duas realizações. Inicialmente, o algoritmo de clusterização foi aplicado sobre o dado normalizado apenas, sem utilizar a mudança em componentes principais (PCA). Paralelamente, o mesmo conjunto de dados foi transformado em componentes principais e foi realizada a redução de dimensionalidade.

A. Clusterização sobre conjunto original

O algoritmo escolhido para realizar a separação dos "clusters" foi o "K-means". Ele divide a base de dados em uma quantidade "K" de subconjuntos. É utilizado o coeficiente Silhouette para realizar a interpretação e validação da consistência dentro dos clusters de dados. À medida que o número de subconjuntos aumenta, a variância obtida é reduzida(uma vez que para K igual ao número de dados, a variância é nula).

O método cotovelo foi utilizado justamente para verificar quando a relação de variação era pequena o suficiente. Assim, o algoritmo foi iterado 9 vezes, para valores de "K" entre 1 e 9, e verificado a curva de variância por número de conjuntos(Curva cotovelo). De forma similar, os valores de cada "Silhouette score" para cada "K" foi exibido e analisada a variação. Como observado abaixo, o valor adequado é para a divisão em 4 subconjuntos.

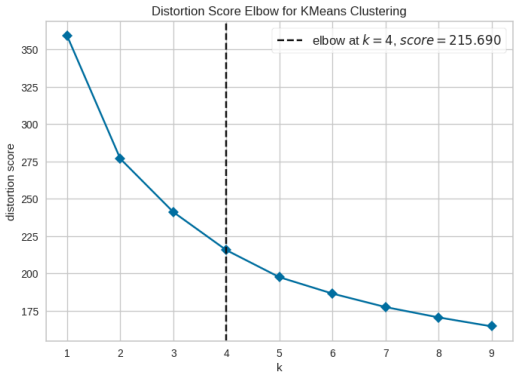

Figura 02- Curva cotovelo

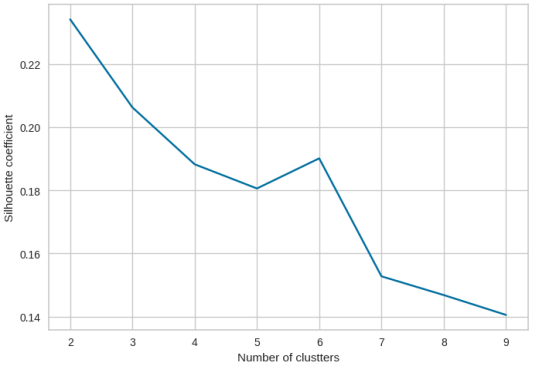

Figura 03- Silhouette Score para cada valor de K

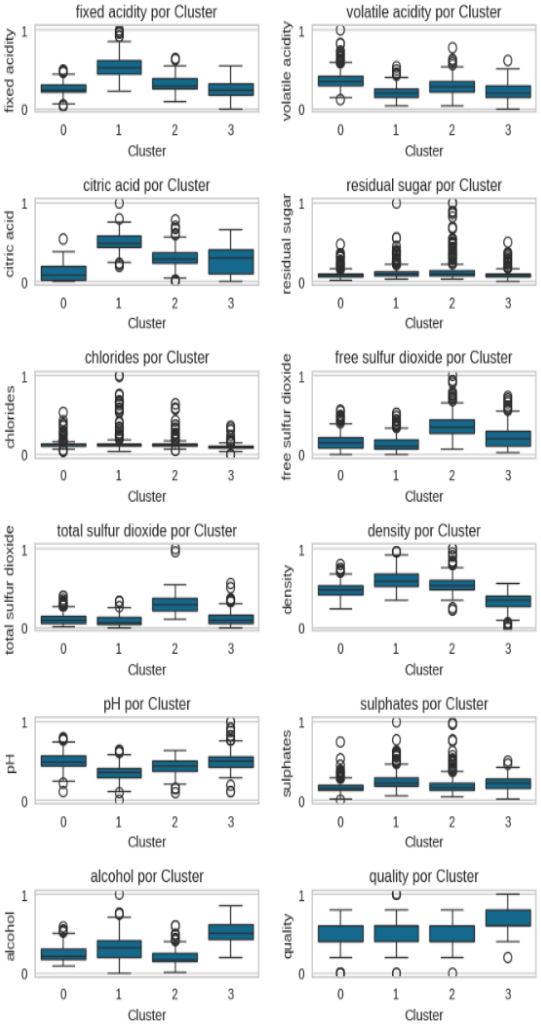

Figura 04- Box plot para cada valor de K

Pelo "box plot" de cada atributo em relação aos seus respectivos clusters, é perceptível que diferem em média e dispersão para cada cluster diferente. Sendo assim, é possível inferir que atributos que diferem entre si para clusters distintos armazenam maior informação em relação à classificação dos vinhos. Entretanto, atributos que pouco se

diferenciam ao se alterar o cluster não possui a capacidade de indicar diferentes tipos de vinhos.

B. Redução de dimensionalidade

A redução de dimensionalidade pode ser realizada através da utilização do algoritmo de PCA. A partir dele é feita a ortonormalização dos atributos, fazendo com que as novas componentes auxiliam na redução de redundância e podem melhorar o desempenho da clusterização em si. A magnitude dos autovalores de cada componente indicam a participação de cada uma em relação a variância.

Ao ser realizada a separação em componentes principais, combinações de atributos do conjunto de dados original são realizadas, de forma que novas componentes sejam formadas. Cada autovetor possui um autovalor associado, indicando a relevância daquela componente na variância. A tabela abaixo evidencia os três atributos com maiores pesos de cada autovetor, assim como seu autovalor.

Autovetor	1°	2°	3°	Autovalor
1	citric acid:0.66	fixed acidity:0.49	pH:-0.31	0.072
2	alcohol:0.64	quality:0.53	density:-0.39	0.049
3	free sulfur dioxide:0.77	total sulfur dioxide:0.52	residual sugar:0.14	0.030
4	quality:0.68	citric acid:-0.43	density:0.40	0.015
5	alcohol:0.60	volatile acidity: 0.41	residual sugar: 0.39	0.014
6	pH:0.54	sulphates:0.47	fixed acidity:-0.33	0.010
7	sulphates:0.47	volatile acidity:0.47	chlorides:0.44	0.010
8	residual sugar:0.60	fixed acidity:-0.36	pH:-0.35	0.007
9	volatile acidity:0.53	citric acid:0.43	residual sugar:-0.39	0.006
10	total sulfur dioxide: 0.64	chlorides:-0.46	free sulfur dioxide:-0.43	0.004
11	chlorides:0.67	sulphates:-0.34	total sulfur dioxide:0.34	0.002
12	density:-0.57	fixed acidity:0.56	pH:0.40	0.001

Tabela 01- Três maiores magnitudes de cada autovetor e autovalor

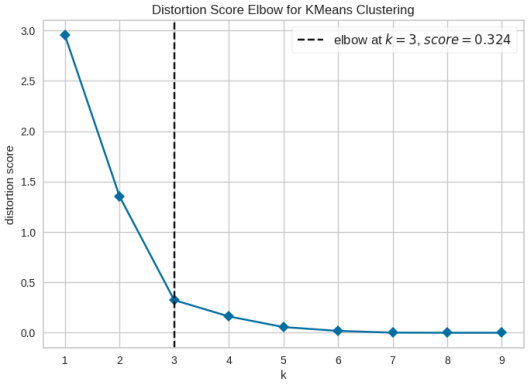

Figura 05- Curva cotovelo

É possível observar que os autovetores com maior autovalor associado possuem as maiores magnitudes dos pesos em atributos que possuem alta correlação entre si. Tal ocorrido era esperado, uma vez que o PCA reduz a redundância dos dados, e criam autovetores ortogonais entre si. Por conta disso, a correlação entre componentes é extremamente baixa.

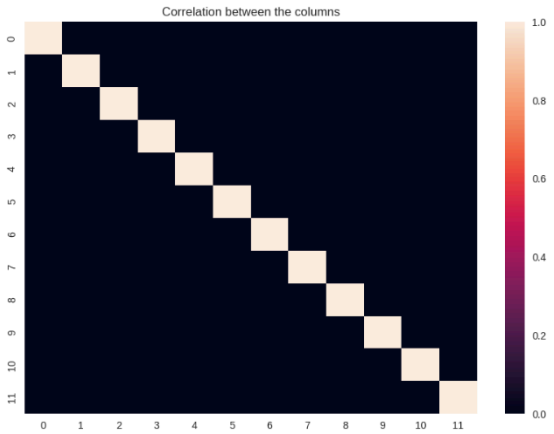

Figura 06- Mapa de calor após transformação

Com base no critério de manter os autovalores que correspondem a uma fração maior do que 95% do total, foram mantidos os nove primeiros autovalores, que correspondem a 96,06% da soma.

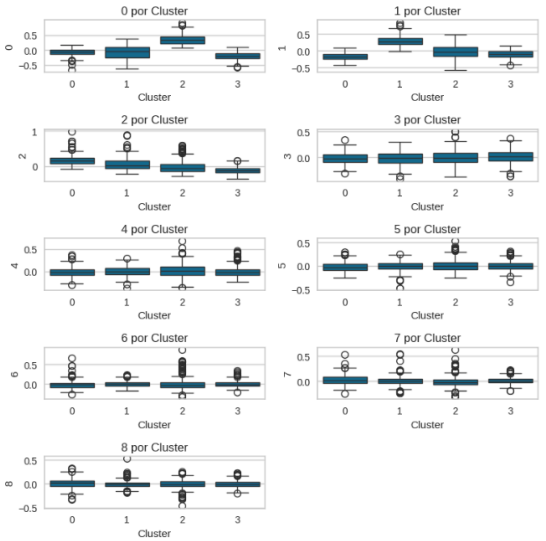

Figura 07- Box plot para cada valor de K

Pela imagem anterior, é possível perceber que os 3 primeiros atributos possuem valores médios distintos entre os clusters, indicando que possuem maior relevância para a separação em si. Além disso, os atributos na parte inferior possuem valores médios parecidos, indicando que as componentes retiradas realmente iriam ter pouca influência na separação entre clusters.

Abaixo é possível observar como as três primeiras componentes são separadas a partir dos respectivos clusters de cada amostra. A relação entre as demais componentes, em geral, apresentam apenas uma sobreposição, possuindo uma baixa quantidade de informação.

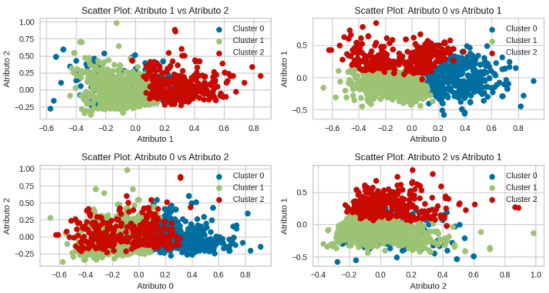

Figura 08- Componentes principais por cluster

III. Conclusão

A partir de todo desenvolvimento anterior, é possível obter algumas observações relevantes acerca do aprendizado não supervisionado, mais especificamente, a clusterização de um conjunto.

Inicialmente, observar a curva de correlação entre atributos pode indicar ao usuário que atributos possuem características em comum. Entretanto, a manipulação de atributos sem o auxílio de outra ferramenta tende a não produzir um desempenho ótimo. Nesse cenário, a aplicação de PCA para reduzir a dimensão do conjunto se torna relevante.

Comparando o "box plot" utilizando e sem utilizar a separação em componentes principais, percebe-se que atributos do conjunto original, em geral, possuem em si algum tipo de informação evidente, uma vez que as médias e dispersões se diferem para clusters distintos. Enquanto isso, após a transformação realizada, apenas as componentes principais apresentaram alguma diferença perceptível para clusters distintos, como mostrado na figura. Para componentes com menores autovalores, como visto na figura abaixo, apenas uma sobreposição de pontos, mesmo após serem retiradas as três componentes menos significativas.

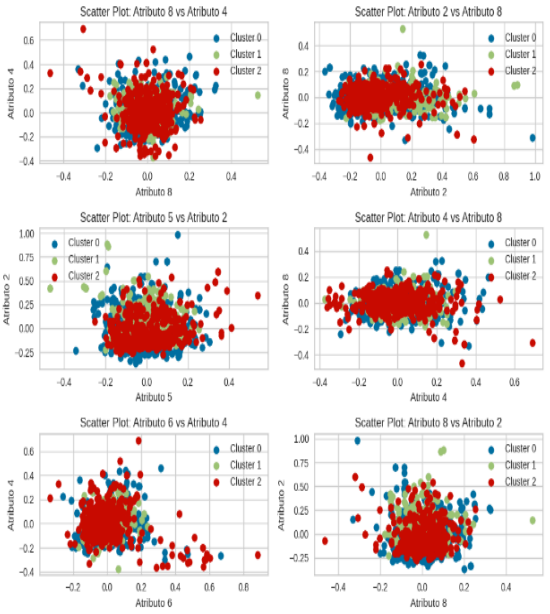

Figura 09- Demais componentes por cluster

Portanto, o PCA se torna uma ferramenta interessante para realizar a redução de dimensionalidade, mantendo as informações relevantes do conjunto de dados. Por sua vez, a redução da dimensão da base auxilia na redução de complexidade no tratamento dos dados.