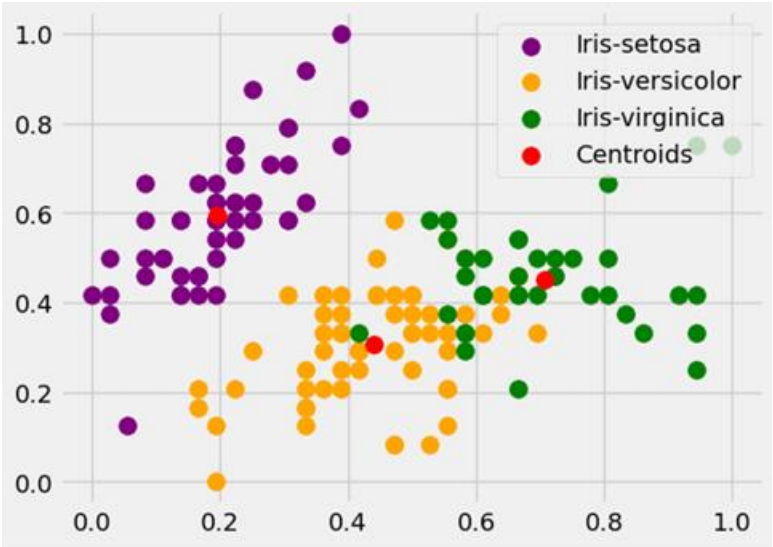
Lista 06 – Felipe Campolina Soares de Paula

Google colab:
https://colab.research.google.com/drive/1v3aRQzQWDtqTIk747VYSUk87O0840b1X?usp=sharing

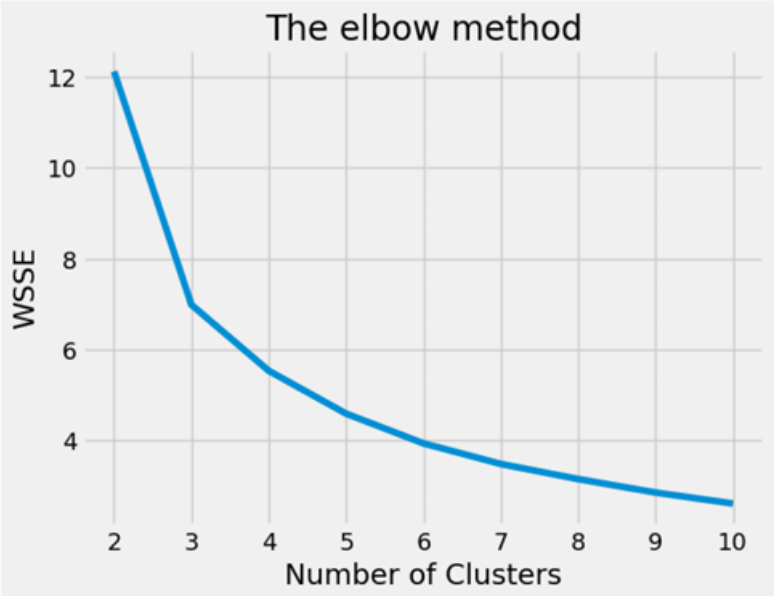
1) Os conjuntos resultantes podem ser encontrados na representação gráfica mostrada abaixo , enquanto o procedimento técnico está disponível no google colab. A Figura abaixo ilustra a distribuição dos dados após o processo de agrupamento com o método K-means. Ao analisar a imagem, é evidente que não todos os pontos foram classificados corretamente.



Para avaliar a qualidade do agrupamento, recorremos à métrica do índice de silhueta (SI), que foi calculada como 0.505. É importante ressaltar que os valores do índice de silhueta estão dentro do intervalo de [-1, 1]. Valores próximos de 1 indicam que os elementos estão relativamente próximos ao centroide de seu próprio cluster, enquanto valores mais próximos de -1 sugerem que os elementos estão mais próximos do centroide de outro cluster do que do seu próprio. Nesse caso, o valor positivo de SI, maior que 0.5, indica que os clusters se ajustam razoavelmente bem aos dados, apesar de não apresentarem agrupamentos totalmente distintos.

Por meio da abordagem do método elbow method, constatamos que o agrupamento com k = 3 clusters é razoável, conforme observado na representação abaixo. No código disponibilizado para essa análise, é importante notar que o algoritmo implementado

em Python indicou k = 4 inicialmente. No entanto, devido ao índice de silhueta desse valor ser inferior a 0.5, optou-se por utilizar o valor k = 3.



2) O Índice de Silhueta é uma métrica que avalia o quão bem os objetos estão agrupados, considerando tanto a coesão dentro de um cluster quanto a separação em relação aos outros clusters. Ele varia de -1 a 1, onde valores mais altos indicam um ajuste mais adequado do objeto ao seu próprio cluster e um ajuste menos adequado a outros clusters. Isso é calculado através da diferença entre a distância média do objeto para outros objetos no mesmo cluster e a distância média para objetos no cluster mais próximo, dividido pelo maior valor entre essas duas distâncias.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Por outro lado, o Método de "Cotovelo" (Elbow Method) é uma técnica usada para determinar o número ideal de clusters em um conjunto de dados. Ele envolve a avaliação da soma dos quadrados das distâncias de cada ponto de dados em relação ao centróide de seu cluster. O objetivo é identificar o ponto no gráfico onde a adição de mais clusters não resulta em uma melhora significativa na explicação da variância. Ao plotar a soma dos quadrados das distâncias em relação ao número de clusters, o ponto onde a curva começa a nivelar, formando um "cotovelo", indica o número ideal de clusters. Este ponto representa o equilíbrio entre a redução da variância intra-cluster e a minimização do número de clusters.

$$SSD = \sum_{i=1}^{n} \sum_{j=1}^{k} ||x_i - c_j||^2$$

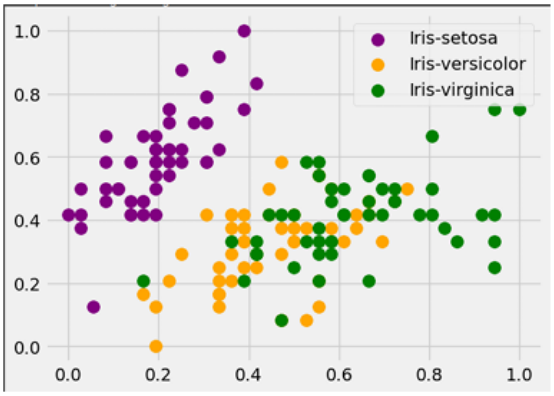
3) A métrica de Índice Dunn é uma medida adicional de avaliação de clusters que considera a relação entre a distância mínima entre os clusters e a distância máxima

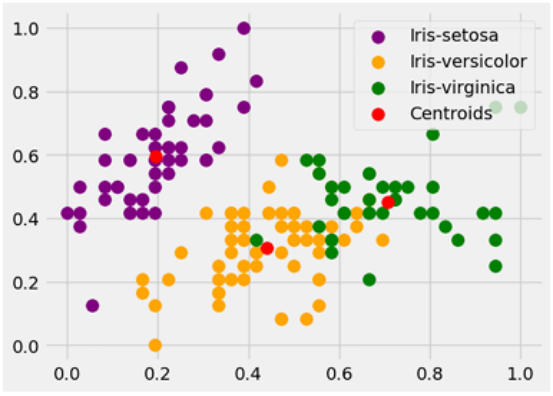
dentro dos clusters. É representada pela razão da menor distância intercluster pela maior distância intracluster, conforme a fórmula:

$$D = \frac{\min_{i \neq j} d(c_i, c_j)}{\max_{1 \leq i \leq n} \delta(c_i)}$$

Essa métrica é útil para avaliar simultaneamente a compacidade e a separação dos clusters. Um valor mais alto do Índice Dunn indica uma melhor separação entre os clusters, enquanto um valor mais baixo sugere uma maior sobreposição ou agrupamentos mais dispersos.

4) Na primeira figura, é possível observar a distribuição original dos dados em contraste com a distribuição resultante do algoritmo K-means. Além disso, a segunda figura destaca os elementos que foram incorretamente alocados pelo referido algoritmo. Para uma análise mais detalhada, consulte a documentação fornecida no Colab.





5)

Redução de Dimensionalidade: Técnica utilizada para reduzir a complexidade de conjuntos de dados com múltiplas variáveis, preservando informações relevantes. Ajuda na simplificação da análise por meio de métodos como PCA e t-SNE.

Remoção de Outliers: Processo crucial para eliminar dados discrepantes que não se encaixam no padrão geral dos dados, evitando distorções nos resultados. Estratégias como o uso de limites estatísticos e algoritmos de detecção de outliers, como Isolation Forest, são empregadas para identificar e remover tais pontos.

Preenchimento de Dados Ausentes: Abordagem para lidar com valores faltantes em conjuntos de dados, evitando impactos negativos na análise e modelagem. Estratégias comuns incluem preenchimento por média, mediana, moda e imputação por regressão.

Balanceamento dos Dados (Undersampling): Técnica para lidar com desequilíbrios em conjuntos de dados com classes minoritárias, reduzindo a quantidade de dados da classe majoritária para equiparar com a classe minoritária. Isso é feito para melhorar o desempenho de modelos de aprendizado de máquina, especialmente em tarefas de classificação.

Normalização dos Dados: Processo para garantir que diferentes características tenham uma escala comparável, essencial para algoritmos sensíveis à escala. Métodos como

normalização min-max, padronização (z-score) e normalização por valor máximo são comumente usados para transformar os dados em uma escala específica, facilitando a interpretação e o processamento.