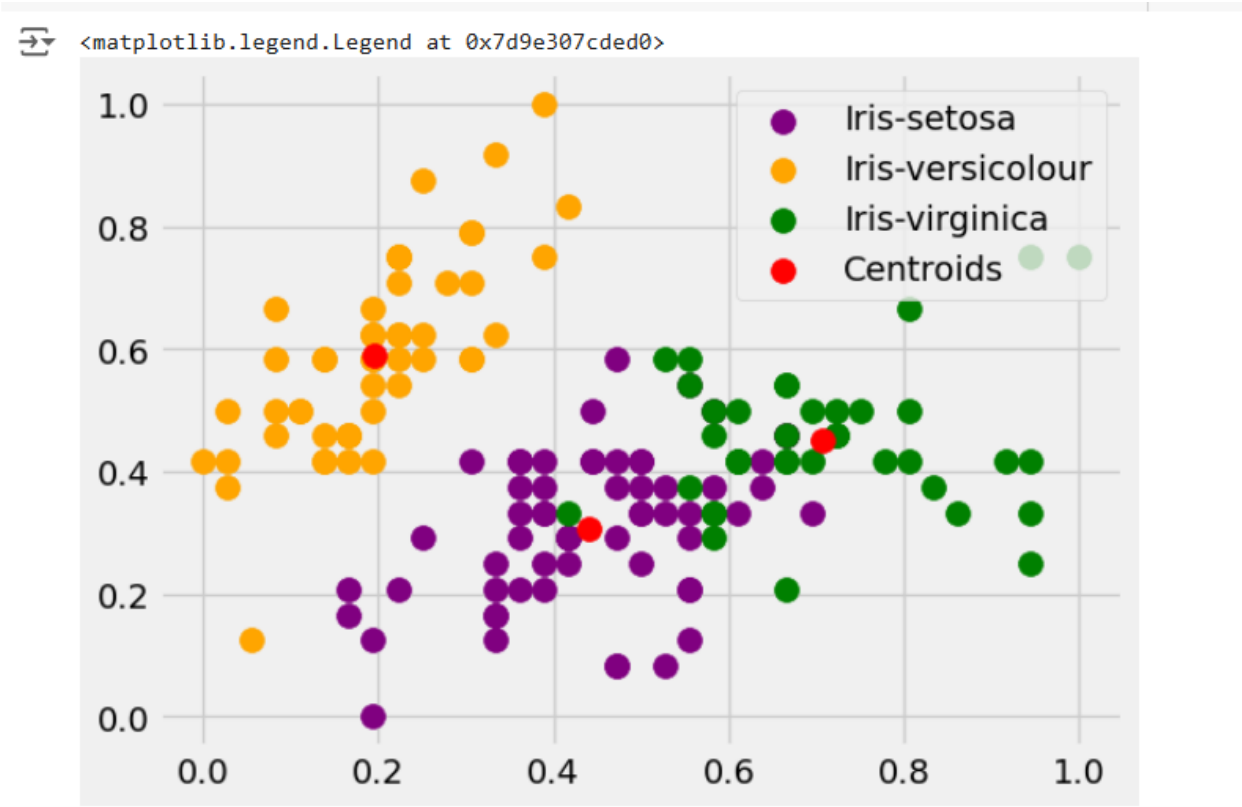
Lista 6 IA

Daniel Salgado Magalhães - 821429

1. Encontre os agrupamentos, discuta a qualidade destes agrupamentos (usando Silhouette e Elbow) e caracterize os agrupamentos obtidos.

É possível perceber que a base de dados foi dividida em 3 agrupamentos, que indicam as classes das flores, e os centroides de cada agrupamento.



Código implementado: https://colab.research.google.com/drive/1X9EYI8H8XqvacvIkeGYg8fYtiskEnbbE?usp=sharing

2. Explique como se obtém estas duas métricas, ou seja, explique as equações matemáticas.

Métrica Silhouette - é uma forma de medir a coesão e a separação dos clusters, ajudando a avaliar o quão bem os pontos estão agrupados. O índice de Silhouette é calculado para cada ponto e varia de -1 a 1.

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

- **a(i)** é a distância média entre i e todos os outros pontos do mesmo cluster.
- **b(i)** é a menor distância média entre i e os pontos do cluster mais próximo (ou seja, a distância média até o centróide do cluster mais próximo que não contém i).

Explicação:

- Se **s(i)**estiver próximo de 1, o ponto está bem agrupado.
- Se **s(i)**for próximo de 0, o ponto está no limite entre dois clusters.
- Se **s(i)** for negativo, o ponto provavelmente está no cluster errado.

Métrica Elbow - utilizada para ajudar a encontrar o número ideal de clusters em uma análise de agrupamento, com o objetivo de minimizar a soma das distâncias quadradas internas aos clusters**.**

```
WCSS = Σ(i=1)^n (i-1)! * (n-i)! * d(i,j)^2
```

COPIAR CÓDIGO

- `n` é o número de pontos na primeira sequência
- `i` é o índice de cada ponto na primeira sequência
- `d(i,j)` é a distância entre o ponto `i` na primeira sequência e o ponto `j` na segunda sequência

Explicação:

1. Para cada valor de K (número de clusters), você calcula o WCSS. Isso envolve somar as distâncias quadradas de todos os pontos em cada cluster até o seu respectivo centróide.
2. Em seguida, você plota o WCSS em função de K em um gráfico.
3. O ponto onde a redução do WCSS começa a diminuir de forma menos significativa (formando um "cotovelo") é o valor de K ideal.

3. Investigue, explique e implemente, pelo menos, mais 1 métrica de avaliação dos agrupamentos, diferentes das 2 anteriores

Índice de Dunn - serve como uma medida de separação e qualidade de compactação dos clusters formados por algoritmos de agrupamento, como K-means, DBSCAN e outros. Quanto maior o valor do Índice de Dunn, melhor é a separação entre os clusters, sugerindo que o agrupamento é de boa qualidade.

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}$$

- d(i,j) é a menor distância entre pontos de clusters diferentes i e j (distância intercluster).
- Delta(l) é o diâmetro do cluster l, ou seja, a maior distância entre qualquer par de pontos dentro do cluster l (distância intracluster).
- k é o número total de clusters.

Componentes do Índice de Dunn

1. **Distância Intercluster d(i,j)**:
    - Refere-se à distância mínima entre pontos de clusters diferentes. Isso mede o quão bem separados estão os clusters.
    - Pode ser calculada usando métricas como a distância Euclidiana, Manhattan, ou outras, dependendo do contexto.

2. **Distância Intracluster Delta(I)**:
    - Mede a dispersão dentro de um cluster, ou seja, o quão próximos estão os pontos de um mesmo grupo.
    - Geralmente, é a maior distância entre qualquer par de pontos no cluster.

Interpretação do Índice de Dunn

- Alta pontuação do Índice de Dunn: Indica que os clusters são bem separados entre si e compactos, significando que o agrupamento é de alta qualidade.
- Baixa pontuação do Índice de Dunn: Sinaliza que os clusters estão próximos (baixa separação) ou que há muita dispersão dentro de pelo menos um cluster, indicando um agrupamento de menor qualidade.

4. Utilizando mais dois algoritmos de agrupamento, por exemplo o DBSCAN e o SOM, verifique se estes métodos encontraram a mesma quantidade de grupos que o Kmeans. Faça uma análise dos grupos encontrados pelos 3 algoritmos

Os agrupamentos não encontraram a mesma quantidade de agrupamentos. O algoritmo que usei para o DBSCAN não conseguiu detectar agrupamentos de forma precisa, e indicou a presença de um único cluster.

O outro algoritmo utilizado foi o índice de Dunn, citado anteriormente, que mostra também 3 agrupamentos, e é representado no código abaixo.

Código implementado:
https://colab.research.google.com/drive/1bWfZlusRlvfOX8UvjBPW15UfRHBquay2?usp=sharing

5. Uma vez que a base é classificada (setosa, virgínica e versicolor), mostre visualmente que instâncias foram agrupadas incorretamente pelo kmeans. Discuta os resultados.

6. Faça um pequeno relatório explicando todas as etapas de pré-processamento realizadas e explicando todos os resultados obtidos.

1. A primeira etapa é a coleta de dados, que verifica a base de dados da Iris.
2. A segunda etapa conta com a remoção de valores nulos ou faltantes.
3. A terceira etapa conta com a conversão de dados categóricos em numéricos
4. A quarta etapa conta com a normalização ou Padronização dos Dados
5. A quinta etapa conta com a redução de Dimensionalidade (Opcional)
6. A sexta etapa conta com a detecção e Remoção de Outliers
7. A sétima etapa conta com a escolha de Features
8. A oitava etapa conta com a verificação de Dados Balanceados
9. A nona etapa conta com a escolha do Algoritmo de Agrupamento
10. A décima etapa conta com a avaliação dos Resultados