1 2 9 0
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

ANÁLISE DE CLUSTERING

APRENDIZAGEM COMPUTACIONAL

Ana Carolina Morais
Nº2021222056

INTRODUÇÃO

O objetivo deste relatório é avaliar diferentes técnicas de clustering aplicadas a vários conjuntos de dados, utilizando métodos como K-Means, Agglomerative Clustering, DBSCAN e Subtractive Clustering. A análise visa identificar a melhor técnica de Clustering e determinar o número mais adequado de 'clusters', considerando métricas de avaliação como o SSE (Sum of Squared Errors) e o Silhouette Score.

P2_CLUSTER1.csv

```
Melhor K para KMeans: 2 com Silhouette Score: 0.4290474635565197 com SSE: 3637734.5248080506
Melhor K para Agglomerative: 2 com Silhouette Score: 0.407540683014861 com SSE: 3637734.5248080506
Subtractive SSE: 0.0
Subtractive Silhouette Score: 0.029296875
DBSCAN: 47 clusters
DBSCAN Silhouette Score: -0.5490418760064064
```

Para o conjunto de dados em análise, o algoritmo KMeans com K=2 apresentou o melhor desempenho. O seu Silhouette Score de 0.429 foi o mais elevado entre todos os testes, indicando uma boa separação entre os 'clusters' e uma forte coesão interna. Embora o SSE (Soma dos Erros Quadrados) para K=2 não fosse o mais baixo, o equilíbrio entre coesão (SSE) e separação (Silhouette Score) faz deste o número ideal de 'clusters'.

Em comparação, o Agglomerative Clustering com K=2 apresentou um Silhouette Score ligeiramente inferior (0.407), e o desempenho deste método diminuiu ainda mais à medida que o número de 'clusters' aumentava. Métodos como Subtractive Clustering e DBSCAN não foram eficazes para este conjunto de dados.

Portanto, a escolha ideal é o KMeans com K=2, dado o seu melhor desempenho geral em termos de qualidade dos 'clusters'.

P2_CLUSTER2.csv

```
Melhor K para KMeans: 2 com Silhouette Score: 0.8733949416857487 com SSE: 137344.74658438523
Melhor K para Agglomerative: 2 com Silhouette Score: 0.8733949416857487 com SSE: 137344.74658438523
Subtractive SSE: 131.0
Subtractive Silhouette Score: 0.42413740014905
DBSCAN: 3 clusters
DBSCAN Silhouette Score: 0.7637116070931051
```

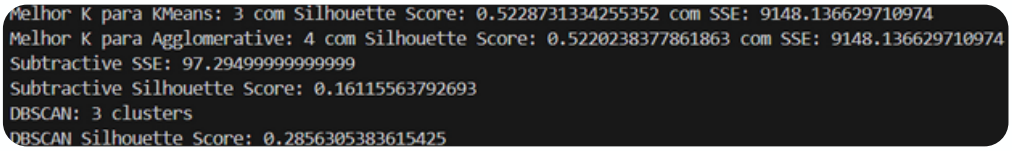
Para o conjunto de dados analisado, o KMeans com K=2 destacou-se com o melhor desempenho, apresentando um Silhouette Score de 0.873 e um SSE de 137344,75. Este Silhouette Score demonstra uma excelente separação entre os 'clusters', sugerindo que os dados estão bem agrupados e as observações dentro de cada 'cluster' são bastante coesas.

O Agglomerative Clustering com K=2 obteve resultados idênticos ao KMeans em termos de Silhouette Score (0.873) e SSE (137344.75), indicando que ambos os métodos são eficazes na segmentação dos dados. No entanto, o KMeans tende a ser preferido pela sua simplicidade e melhor escalabilidade para conjuntos de dados maiores.

O DBSCAN, que formou 3 'clusters', apresentou um Silhouette Score de 0.763, que é significativamente mais baixo do que os resultados obtidos com KMeans e Agglomerative. Embora DBSCAN seja adequado para encontrar 'clusters' de forma irregular e lidar com outliers, neste caso, o seu desempenho não superou os métodos baseados em K.

O método Subtractive Clustering, com um SSE de 131 e um Silhouette Score de 0.424, teve o desempenho mais fraco, o que sugere que não é a melhor abordagem para este conjunto de dados específico.
Portanto, a escolha ideal é o KMeans com K=2, dado que oferece a melhor qualidade de agrupamento, com o mais alto Silhouette Score e uma forte coesão interna entre os 'clusters'.

P2_CLUSTER3.csv



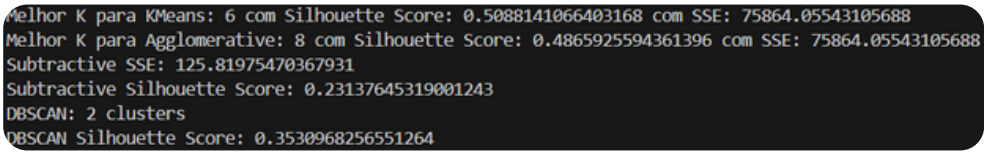
Para este conjunto de dados, o KMeans com K=3 foi o algoritmo que apresentou o melhor desempenho, com um Silhouette Score de 0.523 e um SSE de 9148.14. O valor do Silhouette Score sugere uma boa separação entre os clusters, indicando que os pontos de dados dentro de cada cluster estão bem agrupados, ainda que a separação entre os grupos seja moderada.

O Agglomerative Clustering com K=4 também demonstrou um desempenho semelhante ao KMeans, com um Silhouette Score de 0.522 e o mesmo SSE de 9148.14, sugerindo que ambos os métodos são comparáveis em termos de eficácia neste contexto. No entanto, o menor número de clusters encontrado pelo KMeans pode ser preferido, dado que simplifica a estrutura dos dados sem comprometer a qualidade do agrupamento.

O método DBSCAN identificou 3 ‘clusters’, mas apresentou um Silhouette Score de 0.286, que é consideravelmente mais baixo do que os obtidos pelos métodos anteriores. Isto indica que o DBSCAN teve dificuldades em segmentar bem os dados, o que pode sugerir a presença de ‘clusters’ sobrepostos ou outliers que o algoritmo não conseguiu lidar de forma eficaz.

Por outro lado, o Subtractive Clustering gerou um SSE de 97.29 e o Silhouette Score mais baixo de 0.161, o que reflete um desempenho inferior ao tentar agrupar estes dados, com ‘clusters’ menos definidos e menos coesos.
Assim, tanto o KMeans com K=3 quanto o Agglomerative Clustering com K=4 são as escolhas mais indicadas para este conjunto de dados, com o KMeans a ter uma ligeira vantagem pela sua simplicidade e menor número de ‘clusters’.

P2_CLUSTER4.csv



Para este conjunto de dados, o KMeans com K=6 foi o que teve melhor desempenho, com um Silhouette Score de 0.509 e um SSE de 75864.06, indicando uma boa separação entre os grupos e coesão interna.

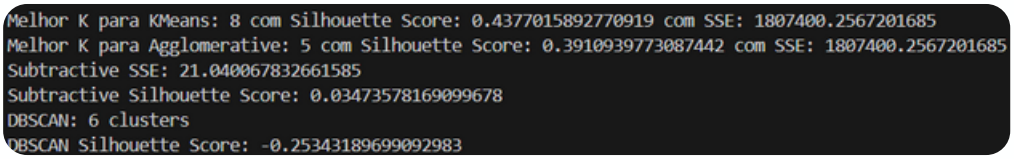
O Agglomerative Clustering com K=8 ficou logo atrás, com um Silhouette Score de 0.487, mas acabou por criar mais clusters sem melhorar muito o resultado.

O DBSCAN formou 2 clusters e teve um Silhouette Score de 0.353, mostrando que não conseguiu identificar bem os grupos.

Já o Subtractive Clustering foi o que teve pior desempenho, com um Silhouette Score de 0.231, não sendo muito eficaz a separar os dados.

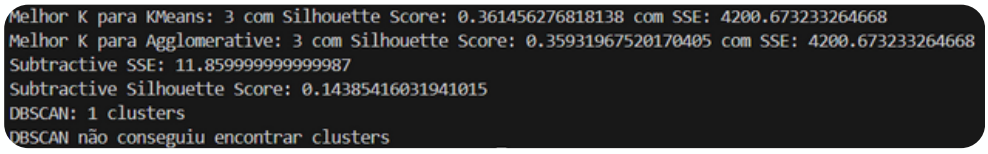
No geral, o KMeans com K=6 foi a melhor escolha para este caso.

P2_CLUSTER5.csv



Para este conjunto de dados, o KMeans com K=8 foi o que obteve o melhor desempenho, com um Silhouette Score de 0.438 e um SSE de 1807400,26, mostrando uma boa separação entre os grupos.
O Agglomerative Clustering com K=5 ficou um pouco atrás, com um Silhouette Score de 0.391, embora tenha o mesmo SSE do KMeans, o que significa que a qualidade dos 'clusters' foi inferior.
O DBSCAN gerou 6 'clusters', mas teve um Silhouette Score negativo (-0.253), sugerindo que não conseguiu separar bem os dados.
Por outro lado, o Subtractive Clustering teve um Silhouette Score muito baixo (0.035), mostrando-se ineficaz neste caso, apesar do SSE bastante reduzido (21.04).
No geral, o KMeans com K=8 foi a escolha mais adequada para este conjunto de dados.

P2_CLUSTER6.csv



Neste conjunto de dados, tanto o KMeans como o Agglomerative Clustering com K=3 apresentaram resultados bastante semelhantes. O KMeans teve um Silhouette Score de 0.361 e um SSE de 4200.67, enquanto o Agglomerative Clustering obteve um Silhouette Score ligeiramente inferior (0.359) com o mesmo SSE.
O Subtractive Clustering, embora tenha tido um SSE muito baixo (11.86), não apresentou um bom Silhouette Score (0.144), sugerindo que não conseguiu formar grupos coesos.
O DBSCAN não conseguiu formar clusters, indicando que não foi eficaz para este conjunto de dados.
Assim, KMeans com K=3 é a melhor opção, já que apresentou a melhor combinação entre separação e coesão dos clusters.