UNIVERSIDADE FEDERAL DE SÃO CARLOS – CAMPUS SOROCABA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CURSO: INTELIGÊNCIA ARTIFICIAL

PROF KATTI FACELI

ATIVIDADE 4 / SAC: TRABALHO DE APRENDIZADO DE MÁQUINA

JESIMIEL EFRAIM DIAS - 726544

SOROCABA
2021

INTRODUÇÃO

Nesta atividade foram feitos a implementação dos algoritmos k-means e single-link em python utilizando as bibliotecas matprotlib, numpy e sklearn para o cálculo do Índice Rand Corrigido (IRA).

A comparação dos agrupamentos gerados pelos algoritmos implementados foram feitos através de gráficos e do IRA.

As saídas dos k clusters para cada conjunto de dados já se encontram organizadas junto com o arquivo .clu e sua respectiva imagem, porém, essa organização foi feita de modo manual e não automatizada.

IMPLEMENTAÇÕES

As implementações estão divididas em quatro algoritmos, divididos nos diretórios kMeans, singleLink e indiceRand, para respectivamente, k-means, single-link e índice rand.

No diretório do k-means, ao executar o main.py, será solicitado como entrada o nome do arquivo com a extensão que necessariamente tem de ser um dos arquivos que estão no diretório datasets, a quantidade de cluster desejado e a quantidade máxima de iterações.

No diretório do single-link, ao executar o main.py, será solicitado como entrada o nome do arquivo com a extensão que necessariamente tem de ser um dos arquivos que estão no diretório datasets, a quantidade de clusters mínima que será gerada e a quantidade máxima de clusters que será gerada, por exemplo, se digitar 2 e 5 para, respectivamente, o mínimo e o máximo de clusters, será gerado a clusterização para 2, 3, 4 e 5 clusters do conjunto de dados dado como entrada.

No diretório do indice-rand temos dois mains, um main que irá calcular o índice rand para o single-link que é o mainSingleLink.py e o mainKmeans.py para calcular o índice rand para o k-means. O mainSingleLink.py utiliza as saídas organizadas nos diretórios em saidaSingleLink no diretório do single-link e o mainKmeans.py utiliza as saídas organizadas em saidaKmeans no diretório do k-means.

CONJUNTO DE DADOS REAL

Os conjuntos de dados real se encontram no diretório dataset do diretório e as suas estruturas para cada conjunto de dados são:

SINGLE-LINK

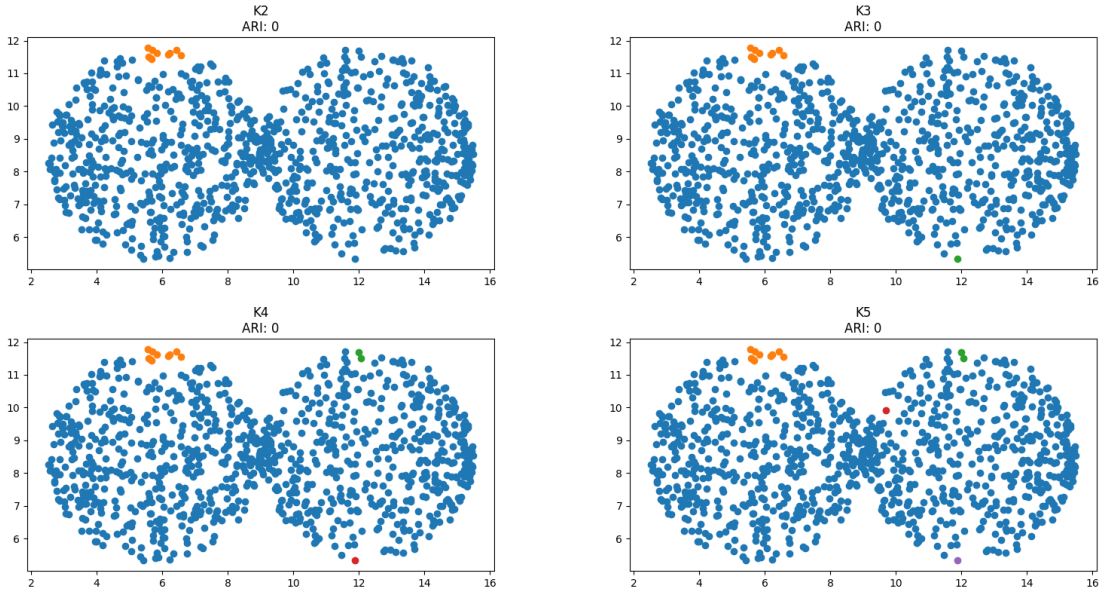
O single-link é um algoritmo de clusterização hierárquico determinístico onde cada ponto do conjunto de dados inicialmente é um cluster, esse algoritmo se apresentou efetivo para clusterização de pontos encadeados. A cada iteração do algoritmo, dois diferentes clusters são unidos baseados na menor distância e as distâncias dos demais clusters são atualizados pegando a menor distância entre os dois clusters.

Na implementação, para os n pontos, foi criado um vetor de tamanho n que representa os n cluster e em cada cluster possui um vetor de distância para cada ponto e a cada interação, é selecionada a menor distância entre dois pontos e são juntados em um único cluster atualizando as distâncias.

Logo, isso resulta em uma matriz de tamanho N x N, porém como os valores são espelhados em relação à diagonal principal, o vetor de distância de um cluster para todos os demais clusters em cada posição do vetor não precisa ser inteiramente calculado, logo, não temos uma matriz completa.

GLOBULARS

Com o ARI próximo ao zero, é possível concluir que o resultado não foi satisfatório em nenhuma das execuções, isso ocorre, pois entre os clusters, existem pouca separação espacial e como o single-link apresenta ser ideal para clusters que possuem pontos encadeados fica praticamente impossível fazer a separação entre os dois clusters, pois na fronteira entre os dois clusters muitos pontos ficam perto um dos outros, o que faz a clusterização encontrada pelo single-link ser bem diferente da clusterização real.

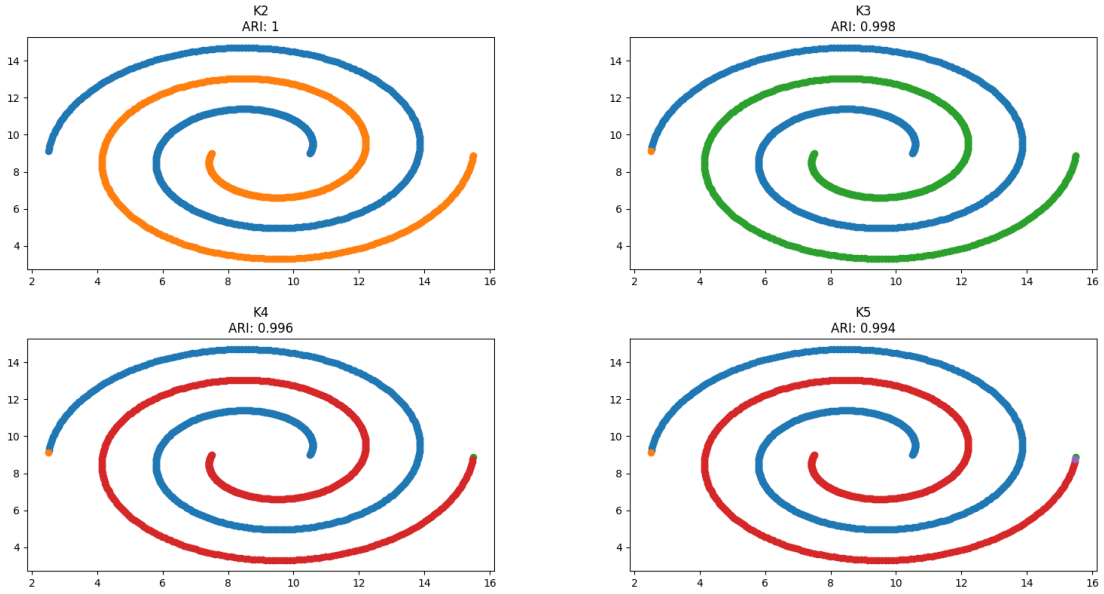


SPIRALS

Com o ARI igual a um ou próximo de um, é possível concluir que o resultado foi mais do que satisfatório, pois quando o algoritmo foi executado com dois clusters, o single-link foi capaz de encontrar a partição real dos dados.

O motivo do resultado mais do que satisfatório é a estrutura do conjunto de dados, pois os clusters estão bem separados e os pontos que fazem parte do mesmo cluster estão encadeados, ou seja, é o conjunto ideal para aplicar o algoritmo single-link.

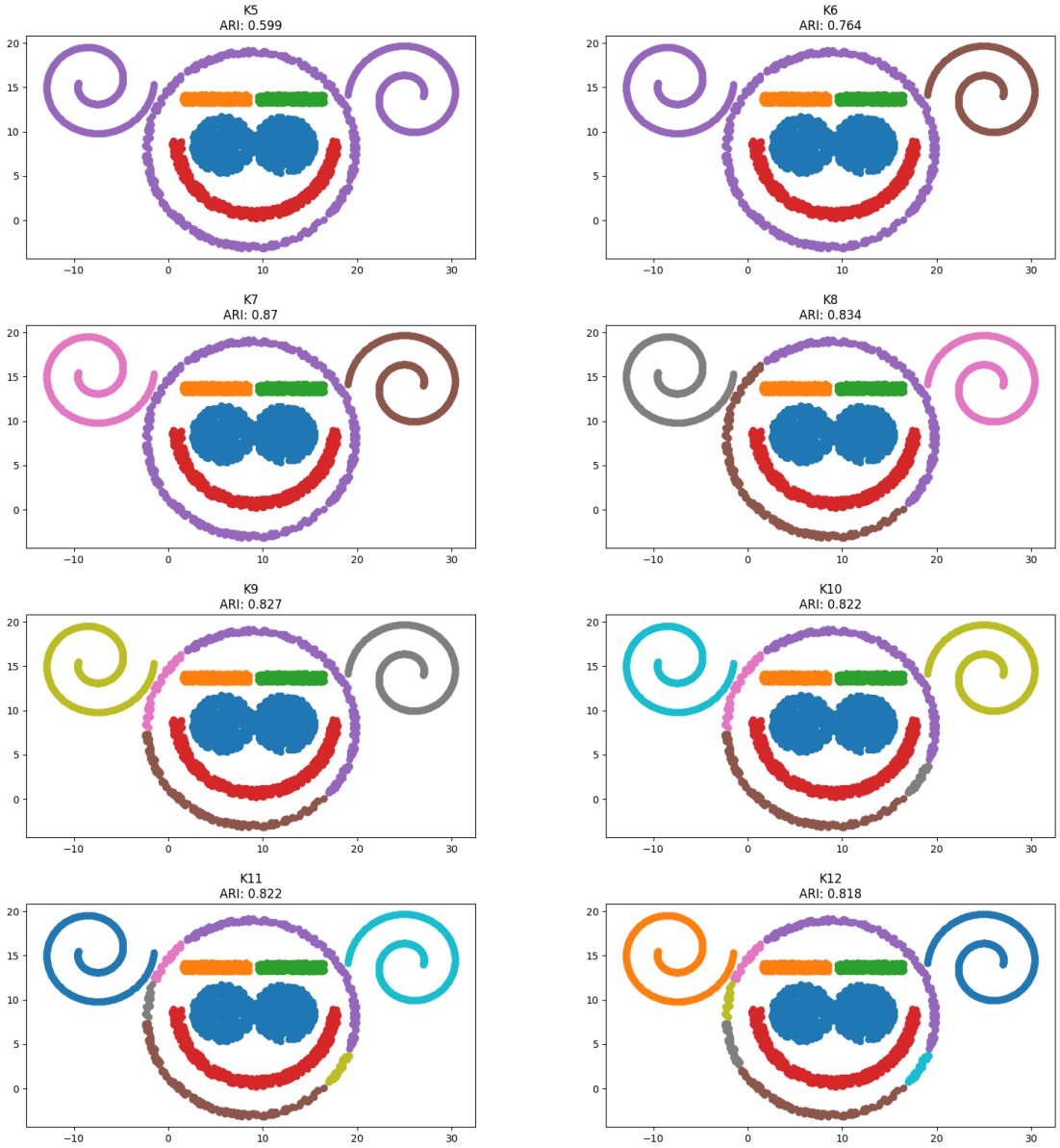
As demais execuções com mais de um cluster também tiveram resultados satisfatórios, pois o ARI chegou muito próximo a um.



MONKEY

O resultado obtido foi razoável, com oito clusters, o ideal seria a detecção de cada uma das partes do macaco como um clusters, porém, como os olhos do macaco apresentam uma estrutura parecida com o conjunto de dados globulars no qual já foi visto que o single-link não é o ideal, ocorre o mesmo problema de dois clusters diferentes resultarem em apenas um.

Com sete clusters, o ARI ficou com um valor mais próximo de um, porém, com o mesmo problema dos olhos do macaco, o motivo que fez a execução com sete clusters obter um resultado melhor foi, provavelmente, que o rosto não foi divido em dois clusters diferentes.



K-MEANS

O algoritmo k-means é um algoritmo interessante para clusterização de pontos baseado no centro, onde cada ponto de um cluster, é mais próximo do ponto central do cluster do qual o mesmo pertence do que dos demais pontos centrais dos outros clusters.

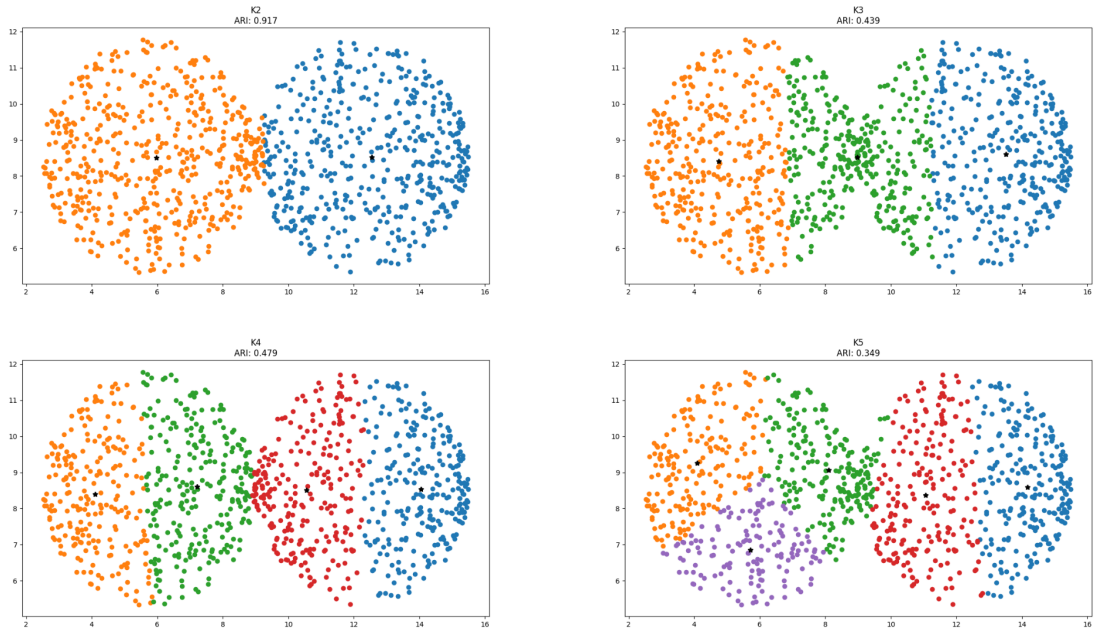
Na execução do algoritmo implementado, cada cluster, é inicializado com um dos pontos do conjunto de dados, essa inicialização é feita de modo randômico, porém, para se aproximar do databook, foi utilizado como padrão uma funcionalidade da biblioteca numpy que permite uma “semente” na inicialização randômica fazendo os valores sempre inicializarem com o mesmo valor, isso ocorreu, pois durante a leitura do data book, foi encontrado esse seed (semente) de valor 15.

Com os pontos iniciais dos clusters, começa então o processo de definir quais pontos pertencem a qual cluster baseado na menor distância euclidiana, após isso, os pontos centrais de cada cluster são recalculados utilizando a média dos pontos. Esse processo pode ocorrer após N iterações, onde N é um valor já definido ou quando os centróides convergirem, isso é, os centróides antigos e novos serem exatamente iguais.

Para melhorar a interatividade, nas imagens geradas pelo algoritmo é possível ver cada centróide de cada cluster representado por uma estrela.

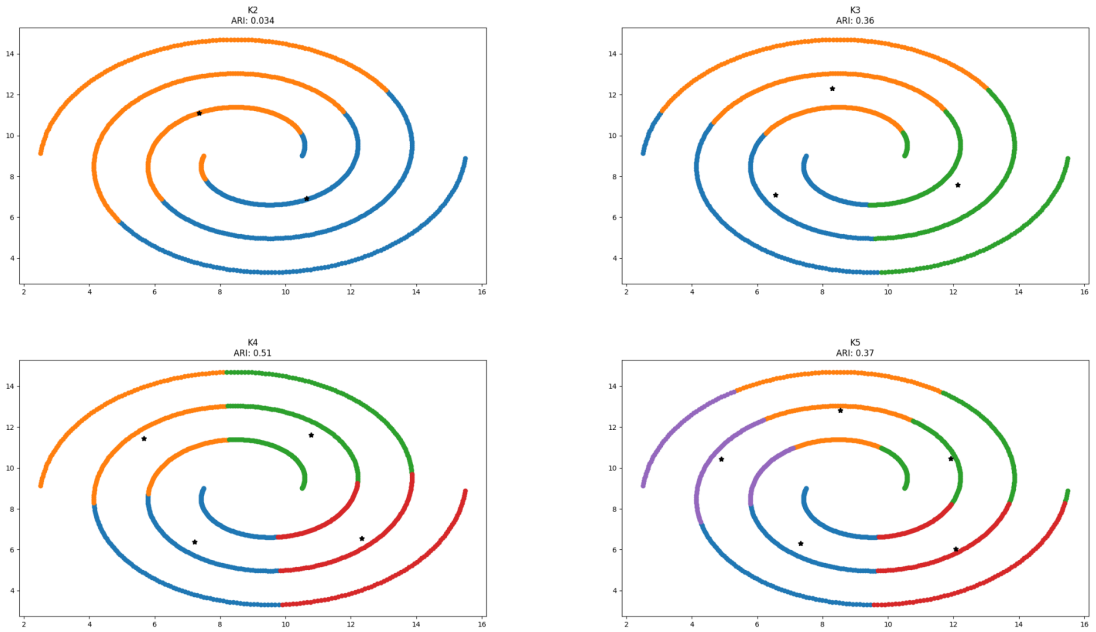
GLOBULARS

O modo em que os pontos estão dispostos no conjunto de dados faz o k-means perfeito para encontrar a clusterização, chegando a um ARI igual a 0,917 quando a quantidade de clusters é igual a dois. Porém, mesmo com um valor próximo de um, temos um problema, pois quando a quantidade de cluster é superior a dois o valor do ARI diminui drasticamente, ou seja, o número de clusters está relacionado com a proximidade da clusterização gerado pelo single-link com a clusterização real.



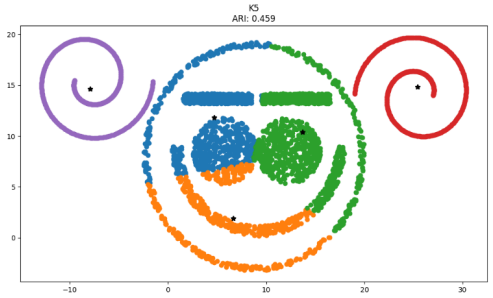
SPIRALS

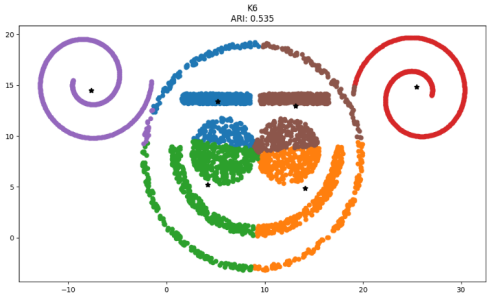
Com o ARI praticamente zero, fica evidente que a clusterização encontrada pelo k-means não é a ideal, isso ocorre porque o modo em que os pontos estão dispersos não é o ideal para o k-means. Se observar a imagem das clusterização geradas pelo k-means pode-se notar que os centróides ficam sempre no meio, ou seja, tentando agrupar os pontos pela proximidade do centro o que é bem diferente da clusterização real.

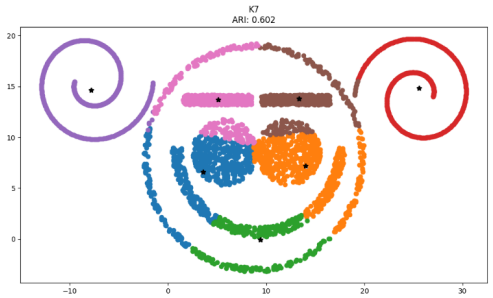


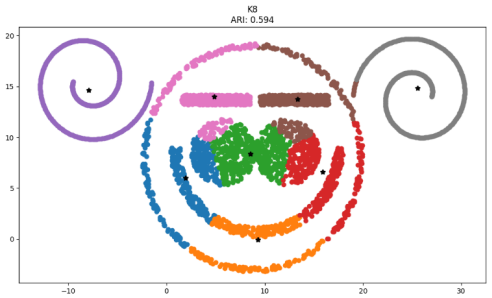
MONKEY

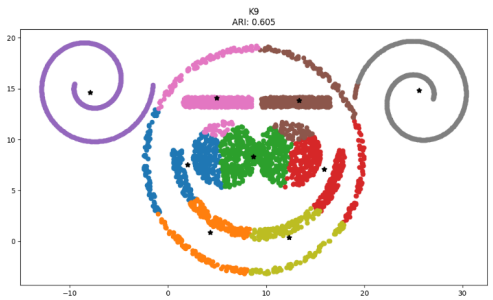
O resultado obtido foi regular e em alguns momento randômicos, pois em alguns casos foi possível separar as partes do macaco, como suas orelhas, partes da sobrancelha e partes dos olhos, porém, observando em um aspecto geral, as clusterização foram randômicas, com os ARI variando por causa da inicialização randômica do k-means, logo, apesar do ARI com sete clusters atingir 0.6 sendo o seed 15, pode-se dizer que o modo em que os pontos estão dispersos não é a melhor para o k-means e a sua performance não é confiável.

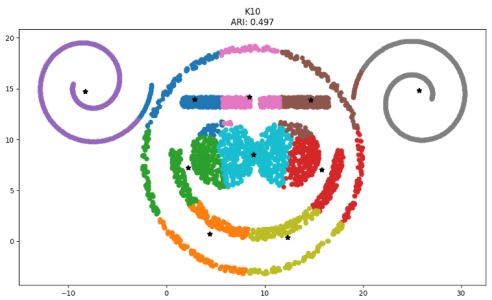


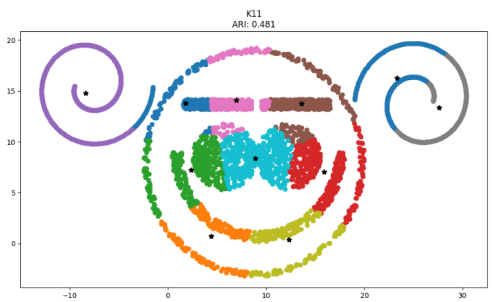


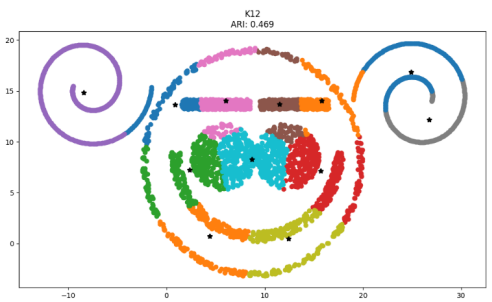












COMPARAÇÃO

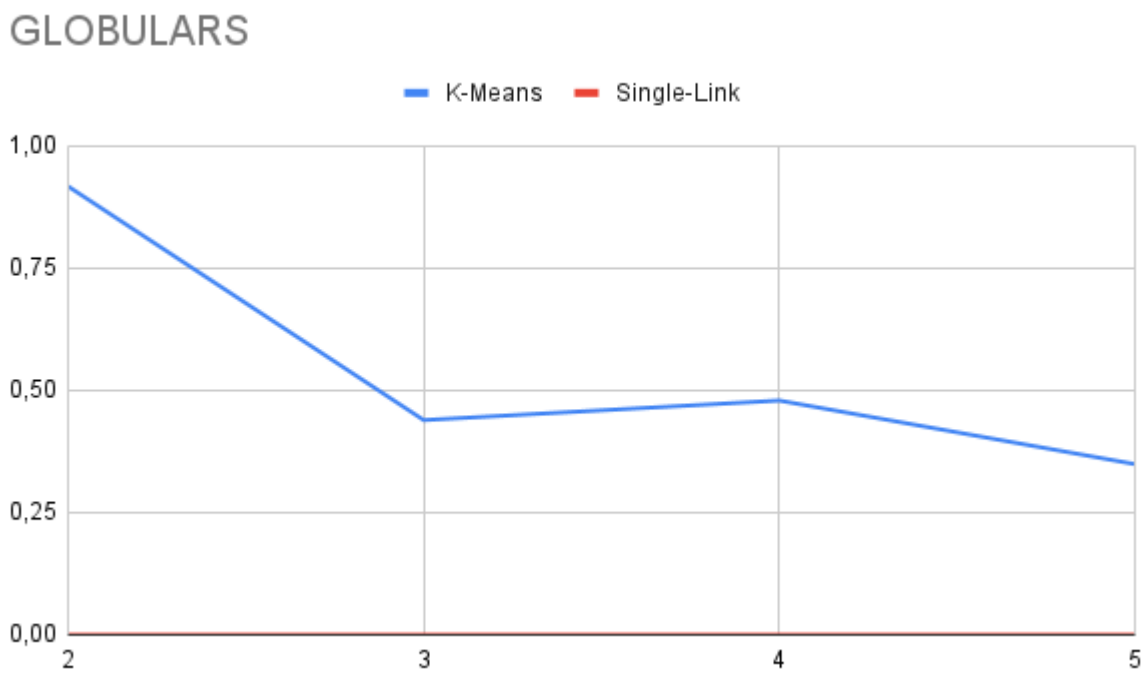
Agora segue a comparação dos algoritmos em relação aos conjuntos de dados, vale ressaltar que não existe algoritmo melhor, apenas diferentes conjuntos de dados em que cada algoritmo se adequa melhor.

GLOBULARS

Nesse conjunto de dados, o k-means foi superior ao single-link, porém, como já observado anteriormente, quando a quantidade de cluster é maior do que dois no k-means,

o rendimento decai muito, já o single-link, teve uma péssima performance independente da quantidade de clusters.

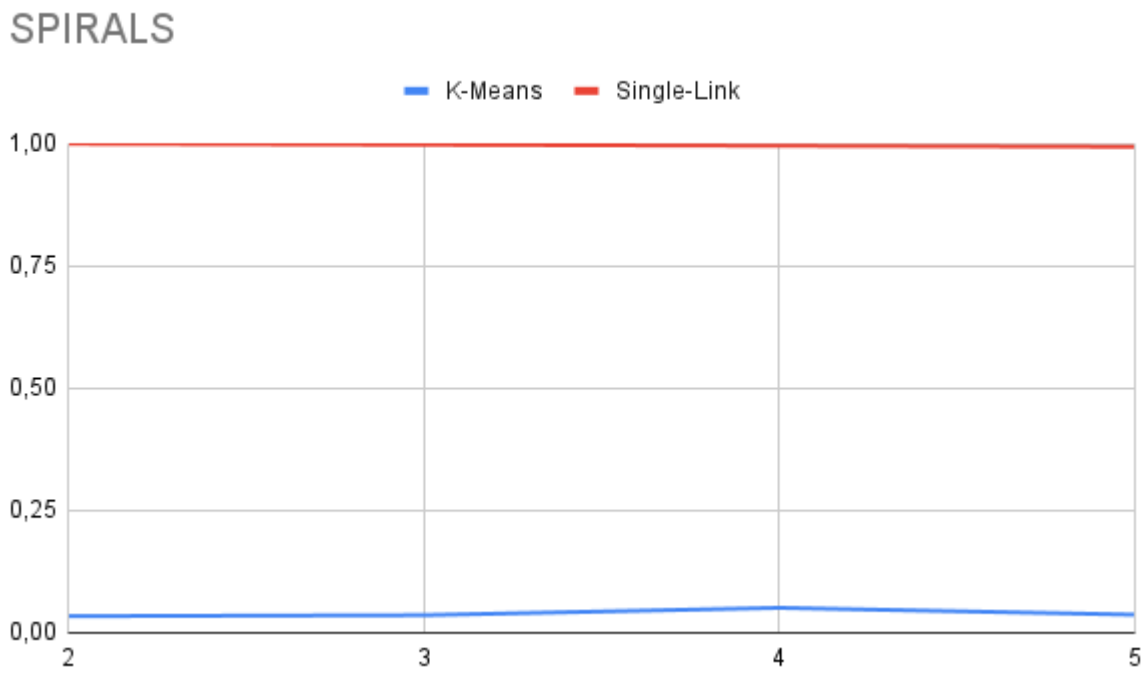
Os motivos que levaram o k-means a obter uma performance superior, foram as já debatidas anteriormente, por possuir em sua essência uma clusterização de pontos mais próximos de um ponto central do cluster é mais fácil encontrar a clusterização real de cluster que possuem pontos dispersos em formas elípticas.



SPIRALS

Nesse conjunto de dados, o single-link teve a melhor performance possível, pois como já visto, com dois clusters, foi possível conseguir a clusterização real do conjunto de dados, ainda assim, nas demais execuções com mais de dois clusters, o rendimento do algoritmo se manteve com valores bem próximos de um.

Isso ocorre pelo fato de que os pontos estão dispersos em uma forma que privilegia os algoritmos que conseguem trabalhar com pontos encadeados, que é o caso do single-link enquanto o k-means é restrito a formas elípticas.

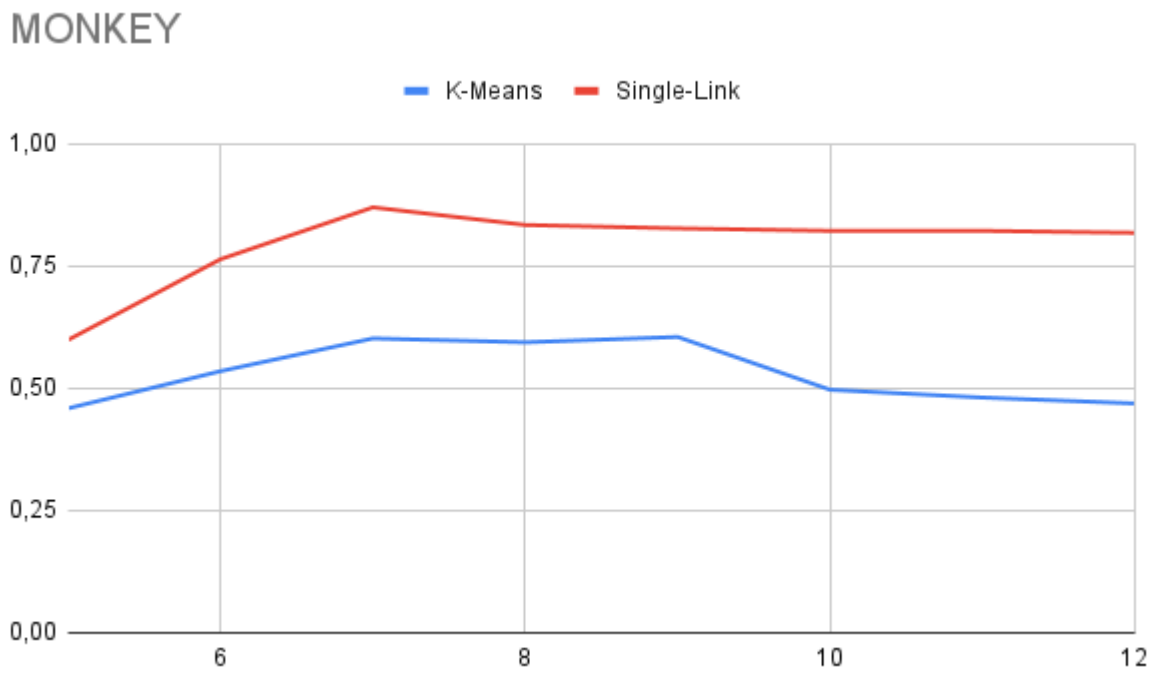


MONKEY

Nesse conjunto de dados, é interessante observar, pois nenhum dos dois algoritmos terão uma vantagem em relação ao outro, pois o modo em que os pontos estão dispersos é bem mais complexo que os demais visto até então.

Observando o gráfico, fica claro que o single-link teve uma performance superior, isso ocorre, pois em vários momentos, os cluster tem uma separação espacial, fazendo com que o single-link consiga diferenciá-los com exceção dos olhos do macaco em que ocorre o mesmo problema do conjunto de dados globulars.

O k-means até teve uma performance regular, porém, não confiável, pois não existe nenhuma garantia da onde o ponto central dos clusters irão estar devido a sua inicialização randômica.



NOTEBOOK

Em comparação com o notebook, o single-link obteve exatamente os mesmos resultados enquanto o k-means teve resultados parecidos, logo, pode-se considerar que pelo fato do single-link ser determinístico, os resultados irão sempre serem iguais enquanto a inicialização randômica do k-means possa resultar em clusterização diferentes, porém, em muitos dos casos, com valores de ARI bem semelhantes.

CONCLUSÃO

Nos conjunto de dados trabalhados, o k-means se saiu melhor no globulars enquanto o single-link se saiu melhor no spirals e monkey, sendo que no spirals foi possível até encontrar a clusterização real do conjunto de dados.

O k-means se saiu melhor no globulars devido a clusterização real ser próxima de uma elipse enquanto o single-link se saiu melhor nos demais por causa dos clusters terem estruturas que privilegiam pontos encadeados.

Uma das observações que é interessante ressaltar é o quão relacionado o k-means é a formas elípticas, sendo difícil utilizar o mesmo para outros conjunto de dados que não possuem estruturas elípticas enquanto o single-link até consegue realizar clusterização não apenas de pontos encadeados, mas também de outras formas desde que os demais clusters tenham uma distância espacial entre eles como o caso das sobrancelhas do

macaco, isso evidencia, que em estruturas complexas como o conjunto de dados monkey, a melhor aposta seria o single-link.