

=====

Esta tarefa consiste na execução do algoritmo de aprendizado não supervisionado para agrupamento (k-means) estudado na disciplina, para definição do melhor número de grupos usando o método do “cotovelo”, usando a biblioteca sklearn (<https://scikit-learn.org/stable/index.html>).

O conjunto de dados pode ser selecionado no repositório UCI (<http://archive.ics.uci.edu/ml/>) ou no site da ferramenta Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) e **deve ser diferente daquele usado no exemplo da aula**. Todas as funções necessárias para a implementação foram apresentadas na videoaula Avaliação de agrupamento.

As etapas que devem ser executadas e relatadas são:

- Selecionar conjunto de dados adequado para problemas de agrupamento. Se o conjunto selecionado tiver o atributo classe, ele pode ser retirado;
- Ler o conjunto de dados no formato .csv coma função read.csv() (biblioteca Pandas);
- Criar outra tabela de dados retirando a coluna referente às classes, se for necessário;
- Fazer as transformações que forem necessárias: transformar atributos nominais em binários com a função get\_dummies() e normalizar os atributos contínuos com a função fit() da classe MinMaxScaler (biblioteca sklearn.preprocessing);
- Definir um intervalo de valores para número de grupos (por exemplo, de 2 a 15, ou um intervalo maior caso seja necessário para visualizar a variação do índice no gráfico);
- Aplicar o algoritmo k-means no conjunto de dados (sem a classe) para todos os valores do intervalo definido usando a classe KMeans e a função fit() (módulo sklearn.cluster);
- Coletar o valor do índice soma quadrática das distâncias para todos os agrupamentos encontrados usando o atributo inertia\_ da classe KMeans (módulo sklearn.cluster);
- Plotar o gráfico dos índices encontrados para cada número de grupos e determinar qual é o melhor de acordo com o método do cotovelo explicado na aula.

Entrega:

- O trabalho deve ser entregue no formato IPython Notebook (.ipynb);
- **Todas as etapas descritas acima devem ser documentadas no notebook**, por meio de células de texto intercaladas com as células de código;
- O trabalho pode ser feito em duplas;
- Entregar (no ava – tarefa de arquivo único) com o arquivo no formato especificado acima;
- DATA DE ENTREGA: 30/04/2022.