Feature Engineering

Resolva as questões do módulo do Kaggle de Feature Engineering: https://www.kaggle.com/learn/feature-engineering (O módulo 6 é opcional).

Escolha de base de dados

Para as questões a seguir, usaremos uma base de dados e faremos a análise exploratória dos dados, antes da clusterização.

1. Baixe os dados disponibilizados na plataforma Kaggle sobre dados sócio-econômicos e de saúde que determinam o índice de desenvolvimento de um país. Esses dados estão disponibilizados através do link: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data
2. Quantos países existem no dataset?
3. Mostre através de gráficos a faixa dinâmica das variáveis que serão usadas nas tarefas de clusterização. Analise os resultados mostrados. O que deve ser feito com os dados antes da etapa de clusterização?
4. Realize o pré-processamento adequado dos dados.

Clusterização

Para os dados pré-processados da etapa anterior você irá:

1. Realizar o agrupamento dos países em 3 grupos distintos. Para tal, use:
    a. K-Médias
    b. Clusterização Hierárquica
2. Para os resultados, do K-Médias:
    a. Interprete cada um dos clusters obtidos citando:
        i. Qual a distribuição das dimensões em cada grupo;
        ii. O país, de acordo com o algoritmo, melhor representa o seu agrupamento. Justifique.
3. Para os resultados da Clusterização Hierárquica, apresente o dendograma e interprete os resultados.
4. Compare os dois resultados, aponte as semelhanças e diferenças e interprete.

Escolha de algoritmos

1. Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.

2. O algoritmo de K-médias converge até encontrar os centróides que melhor descrevem os clusters encontrados (até o deslocamento entre as interações dos centróides ser mínimo). Lembrando que o centróide é o baricentro do cluster em questão e não representa, em via de regra, um dado existente na base. Refaça o algoritmo apresentado na questão 1 a fim de garantir que o cluster seja representado pelo dado mais próximo ao seu baricentro em todas as iterações do algoritmo.
   Obs: nesse novo algoritmo, o dado escolhido será chamado medóide.

3. O algoritmo de K-médias é sensível a outliers nos dados. Explique.

4. Por que o algoritmo de DBScan é mais robusto à presença de outliers?

Assim que terminar, salve o seu arquivo PDF e poste no Moodle. Utilize o seu nome para nomear o arquivo, identificando também a disciplina no seguinte formato: “nomedoaluno_nomedadisciplina_pd.PDF”.