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?
   R: **Existem 167 países 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?
   R: **Deve ser feito a normalização dos dados.**
   Análise Exploratória:

   As colunas gdpp(PIB per capita) e income(Renda), imports(Importações) e exports(Exportações), child_mort(Óbito) e total_fert(Taxa de natalidade) são altamente correlacionados positivamente considerando que life_expec(Expectativa de vida) e child_mort(Óbito) são altamente correlacionados positivamente.

4. Realize o pré-processamento adequado dos dados.
   Feito o pré-processamento pelo jupyter notebook.

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

        R: **Países subdesenvolvidos que representam o cluster 0, possuem uma alta mortalidade infantil, baixo PIB per capita e baixa inflação.**

        Países em desenvolvimento que representam o cluster 1, possuem baixa mortalidade infantil, alto PIB per capita e alta inflação.

        Países desenvolvidos que representam o cluster 2, possuem alta mortalidade infantil, baixa PIB per capita e alta inflação.

        ii. O país, de acordo com o algoritmo, melhor representa o seu agrupamento.

        Jamaica

3. Para os resultados da Clusterização Hierárquica, apresente o dendograma e interprete os resultados

    R: **Países subdesenvolvidos que representam o cluster 0, possuem uma baixa mortalidade infantil, alto PIB per capita e baixa inflação.**

    Países em desenvolvimento que representam o cluster 1 possuem alta mortalidade infantil, baixo PIB per capita e alta inflação.

Países desenvolvidos que representam o cluster 2, possuem alta mortalidade infantil, baixa PIB per capita e alta inflação.

4. Compare os dois resultados, aponte as semelhanças e diferenças e interprete.
   R: **O resultado da clusterização hierárquica podemos cortar a árvore em inúmeros pontos para definir o número de cluster, já o K-means é definido antes da inicialização. No k-means os pontos ficam mais sensíveis a outliers, diferente do que na clusterização hierárquica que é mais eficiente na detecção de outliers. K-means pode usar mediana ou média como centro de cluster para representar cada cluster, os métodos aglomerativos começam com n clusters e combinam sequencialmente clusters semelhantes até que apenas um cluster seja obtido.**

Escolha de algoritmos

1. Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.
   R: **1º passo: Dado o parâmetro de K clusters, inicializa K centróides em pontos aleatórios;**
   2º passo: Para cada ponto, encontrar qual o centróide mais próximo.
   3º passo: Para cada ponto, calcular o centróide de menor distância. Cada ponto pertencerá ao centróide mais próximo;
   4º passo: Reposicionar o centróide. A nova posição do centróide deve ser a média da posição de todos os pontos do cluster;
   5º passo: Repetir novamente o segundo passo até o número de interações pré-especificado for atingido ou até a posição dos centróides não mudar mais.

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.
   R: **O algoritmo de K-medoids comporta-se de uma maneira similar ao K-means, mas ao invés do centróide mover-se com base na média da distância dos cluster, o centróide passa a ser o cluster que está mais próximo do centro.**

   Obs: nesse novo algoritmo, o dado escolhido será chamado medóide.
3. O algoritmo de K-médias é sensível a outliers nos dados. Explique.
   R: **Por ser baseado em médias acaba alterando muito a média dos clusters. O centróide é o centro de gravidade e o outlier vai exercer tanta gravidade sobre o centróide que ele vai deslocar completamente esse centróide de um lugar razoável, então ele vai ficar longe dos pontos que ele deveria efetivamente estar ligado por isso que se não tratar o outliers é um problema para o K-means.**

4. Por que o algoritmo de DBScan é mais robusto à presença de outliers?
   R: **Porque a utilização do método de densidade, o algoritmo DBScan realiza a clusterização procurando por regiões densas no espaço dos dados, permitindo que sejam encontrados grupos com formatos excessivos e sejam detectados outliers.**