Clustering

Objetivo: agrupar observações em grupos homogêneos internamente e heterogêneos entre si

Técnica exploratória não supervisionada: não há caráter preditivo

Método Hierárquico Aglomerativo:
quantidade de clusters é definida ao longo da análise

1) Análise das unidades de medida
Caso estejam em unidades de medidas diferentes, aplica-se o Z-score

$$ZX_{ji} = \frac{X_{ji} - \bar{X}_j}{s_j}$$

2) Medida de dissimilaridade (distância)
Distância de Manhattan, Chebychev, Canberra e Correlação de Pearson (também pode ser utilizada já que é uma medida de SEMELHANÇA)

3) Métodos de encadeamento: esquemas hierárquicos aglomerativos

Nearest Neighbor (single linkage): privilegia menores distâncias, recomendável em casos de observações distintas

Furthest Neighbor (complete linkage): privilegia maiores distâncias, recomendável no caso de observações parecidas

Between groups (average linkage): junção de grupos pela distância média entre todos os pares de observações

4) Quantos agrupamentos?

Como critério para o número final de clusters, adota-se o tamanho dos saltos para incorporação do cluster seguinte. Saltos muito altos indicam o agrupamento de observações com características mais distintas.

Para isso, utilizam-se os dendogramas.

5) Comparação entre variabilidades

Após a análise, é importante comparar se a variabilidade dentro do grupo criado é menor que a variabilidade entre os grupos. Para isso, aplica-se o **teste F** para análise de variância:

$$F = \frac{Variabilidade entre grupos}{Variabilidade dentro dos grupos}$$

Graus de liberdade no numerados: K-1
Graus de liberdade no denominador: n-K

Sendo **k o número de clusters** e **n o tamanho da amostra**

Método Não Hierárquico K-means:
quantidade de clusters definida à priori

1) Análise das unidades de medida
Caso estejam em unidades de medidas diferentes, aplica-se o Z-score

$$ZX_{ji} = \frac{X_{ji} - \bar{X}_j}{s_j}$$

2) Esquemas não-hierárquicos

A quantidade K de clusters é escolhida à priori. Para isso, utiliza-se de base os centros de aglomeração, em que as observações são arbitrariamente alocadas aos K clusters para o cálculo dos centroides iniciais.

3) Comparação da proximidade dos centroides

As observações serão comparadas pela proximidade aos centroides de outros clusters. Se houver realocação a outro cluster por estar mais próxima, os centroides devem ser recalculados (em ambos os clusters)

4) Processo iterativo

O procedimento K means encerra-se quando não for possível realocar qualquer observaçãopor estar mais próxima do centroide de outro cluster. Ou seja, indica que a soma dos quadrados de cada ponto até o centro do cluster alocado foi minimizada.

$$SS = \sum_{k=1}^{k} \sum_{x_i \epsilon c_k} (x_i - \mu_k)^2$$