LL3. Utilizarea diferitor algoritmi de clusterizare pe diferite tipuri de date. Vizualizarea datelor.

Concepte de bază

- Cluster = grup de date care sunt similare.
- Clustering (grupare) = proces de identificare a clusterelor.
- Prototip (reprezentant)= "obiect" reprezentativ pentru datele dintr-un cluster.
  - *Centroid* = media datelor dintr-un cluster – centroidul nu este neapărat un element din setul de date;
  - Medoid = data din cluster care este cea mai apropiată de media clusterului – medoidul aparţine setului de date.
- Raza clusterului = abaterea standard a datelor din cluster (distanța medie față de centroid).
- Diametrul clusterului = distanţa (disimilaritatea) maximă dintre oricare două date ale clusterului.

Tipuri de clustering

Crisp vs fuzzy clustering

- Crisp clustering = fiecare dată aparţine unui singur cluster.
- Fuzzy clustering = o dată poate aparţine mai multor clustere (grad de apartenenţă pentru fiecare cluster).

Flat vs hierarchical clustering

- Flat (partitional) clustering = rezultatul este un set de clustere (o partiţie).
- Hierarchical clustering = rezultatul este o ierarhie de partiţii.

Variante de algoritmi

- Algoritmi partiţionali (ex: kMeans, Fuzzy cMeans).
- Algoritmi hierarhici (alg. aglomerativi, alg. divizivi).
- Algoritmi bazaţi pe densitate (ex: DBSCAN).
- Algoritmi bazați pe modele probabiliste (ex: EM = Expectation Maximization).

Măsuri de calitate

Nu există un indicator unic pentru evaluarea calităţii unei grupări. Cea mai comună abordare constă în estimarea:

- Compacității clusterelor (variabilitate intra-cluster – ar trebui să fie mică).
- Gradului de separare dintre datele aparținând unor clustere diferite (variabilitate inter-cluster – ar trebui să fie mare).

```
* from sklearn.metrics import silhouette_score
```

Sarcina individuală:

Alegeți spre implementare (în limbajul Python) câte o metodă din fiecare categorie de mai sus (cel puțin patru algoritmi). Explicați conceptul și pașii de bază ai fiecărui algoritm. Alegeți sau creați un set de date conform necesităților fiecărui algoritm elaborat. Evaluați rezultatele obținute.