Data Mining : K-nearest neighbors
TD N : 2

Exercice 1

Supposons que l'on a un problème de classification qui consiste à déterminer le poids d'une personne en se basant sur la taille et l'âge de cette personne. Le tableau suivant comprend la taille, l'âge et le poids (cible) pour 10 personnes

Travail à faire :

1. En se basant sur l'ensemble de données, identifier la nature de ce problème.
2. On souhaite maintenant utiliser l'algorithme K-NN pour prédire le poids de la personne ID11 en fonction de sa taille et de son âge.
    - Classer la nouvelle observation en appliquant l'algorithme K-NN. (Détailler les calculs).
    - Utiliser la distance euclidienne qui a la formule suivante : $D_E = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
    - Considérer la valeur : k=2 et par la suite k=5.

ID	Taille	Age	Poids
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	

Exercice 2

Supposons que l'on a un problème de classification qui consiste à déterminer la classe d'appartenance de nouvelles instances X_i. Le domaine de valeurs des classes possibles est 1,2,3. Le tableau suivant représente notre ensemble de données avec les 5 attributs : A_1, A_2, A_3, A_4 et A_5.

Travail à faire :

On souhaite maintenant utiliser l'algorithme K-NN pour déterminez à la main la classe de l'instance X_6, dont les valeurs pour les attributs numériques A1 à A5 sont <3,12,4,7,8>.

1. Classer la nouvelle observation en appliquant l'algorithme K-NN (Détailler les calculs).

2. Utiliser la distance de Manhattan qui a la formule suivante : $D_m(x,y) = \sum_{j=1}^{n} |x_j - y_j|$
3. Considérer la valeur : k=1 et par la suite k=4.

X_i	**A_1**	**A_2**	**A_3**	**A_4**	**A_5**	**Classe**
X_1	3	5	4	6	1	1
X_2	4	6	10	3	2	2
X_3	8	3	4	2	6	3
X_4	2	1	4	3	6	3
X_5	2	5	1	4	8	2