# Algorithme des k plus proches voisins - Exercices

## *Exercice 1*

On considère le fichier csv dont le contenu est représenté ci-contre. Il contient les données de 10 points de l'espace colorés selon une certaine logique spatiale.

```
x,y,z,couleur
3,7,5,noir
4,6,2,noir
3,7,8,blanc
0,1,2,noir
1,0,7,blanc
5,4,4,blanc
9,1,2,noir
5,3,3,noir
1,1,4,blanc
3,3,7,blanc
```

On considère un onzième point A de coordonnées x=4, y=4 et z=5 dont on cherche à prédire la couleur en appliquant l'algorithme des k plus proches voisins.

1. Combien de descripteurs vont être utilisés pour effectuer la prédiction ? Quels sont-ils ?

2. Combien y a-t-il d'étiquettes différentes ? Quelles sont-elles ?

3. Après un import de ce fichier dans Python sous forme de table, quel serait le dictionnaire associé à la ligne 3 du fichier ( 3,7,8,blanc) ?

4. On a obtenu le tableau incomplet des distances ci-contre qui donne les distances entre le point A(4;4;5) et les points de la table.
   Calculer les deux distances manquantes.

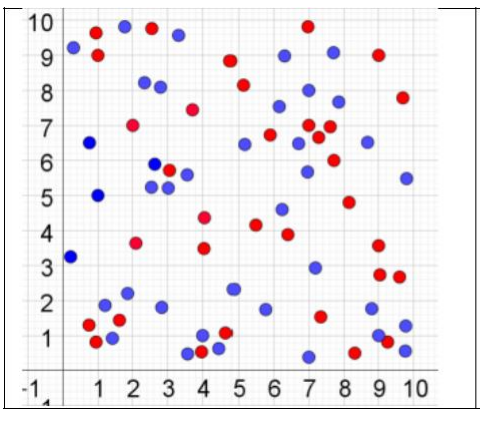
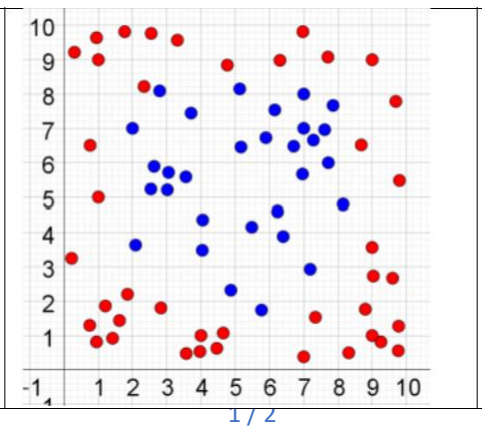
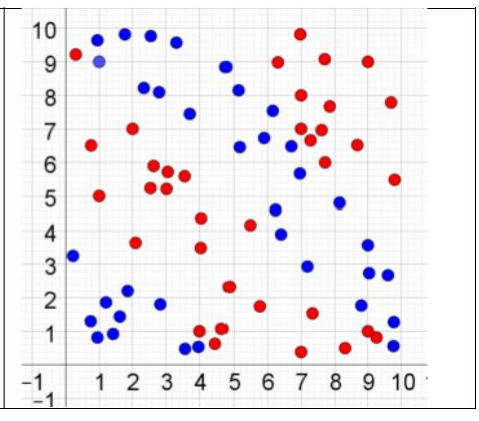
| Point de la table | Distance au point A |
|---|---|
| 3,7,5,noir | 3.16 |
| 4,6,2,noir | 3.60 |
| 3,7,8,blanc | 4.36 |
| 0,1,2,noir | |
| 1,0,7,blanc | 5.39 |
| 5,4,4,blanc | 1.41 |
| 9,1,2,noir | |
| 5,3,3,noir | 2.45 |
| 1,1,4,blanc | 4.36 |
| 3,3,7,blanc | 2.45 |

5. On applique l'algorithme des k plus proches voisins pour prédire la couleur au point A.
   - Si k = 1, quelle prédiction obtient-on ?
   - Si k = 3, quelle prédiction obtient-on ?
   - Si k = 5, quelle prédiction obtient-on ?

6. Pourquoi ne prend-on pas une valeur paire pour k ?

## *Exercice 2*

On considère trois jeux de données différents pour lesquels on a deux descripteurs sur lesquels baser la prédiction. La prédiction porte sur une couleur (étiquette rouge ou étiquette bleue). On a représenté les 3 jeux de données ci-dessous.
Lequel ou lesquels de ces 3 jeux de données ne vont pas permettre de réaliser des prédictions fiables ?

## *Exercice 3*

On applique l'algorithme des k-plus proches voisins sur une table table_donnees dont deux enregistrements sont par exemple :

```
#1 {'nb_clics': 0.489,  'duree': 0.517, 'panier': 0.854, 'satisfait': 'O'}
#2 {'nb_clics': 0.828,  'duree': 0.865, 'panier': 0.142, 'satisfait': 'N'}
```

Les valeurs associées aux différents descripteurs ont toutes été ramenées entre 0 et 1 afin de leur donner la même importance lors du calcul de distance.

1) D'après vous, de quel type d'étude provient cette table de données :
   - Étude de satisfaction sur les clients d'un fabricant de paniers en plastique,
   - Étude de satisfaction sur les clients d'un site web commercial.

2) Pour un nouveau client, on dispose de valeurs associées aux trois descripteurs dans un dictionnaire : `{'nb_clics' : 0.632, 'duree':0.321, 'panier': 0.242}`.
   On souhaite prédire la valeur d'un champ `'satisfait'` de ce client.

   a) En termes de vocabulaire, quelles sont les deux affirmations correctes ?
   - On cherche à effectuer une classification dans deux classes : 'O' et 'N'
   - On cherche à étiqueter le client en utilisant deux étiquettes : 'O' et 'N'
   - On cherche à distancer le client en utilisant trois descripteurs.

   b) Calculer la distance entre ce nouveau client et l'enregistrement `#1`.

   c) Écrire le code d'une fonction Python qui prend en paramètres :
   - Un enregistrement de la table client_table,
   - Un enregistrement client_nouveau, et renvoie la distance entre ces deux enregistrements (on utilisera `math.sqrt` du module `math` pour calculer la racine carrée).