Faculté d'informatique
Departement IA & SD

DATA MINING

TP6: La Classification Supervisée (K-NN)

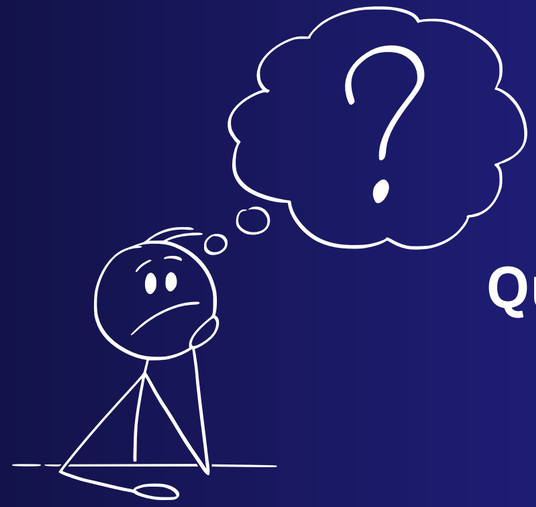
BENADJAL.N
Phd student in computer science
LAB: MOVEP
Email: nesrinebenadjel63@gmail.com

INGSOFT4

2025/2026

Introduction

Au cours des séances de TP précédentes, nous avons exploré différents algorithmes d'apprentissage non supervisé, notamment la segmentation (clustering), qui est une technique de regroupement de données.
Dans ce TP, nous allons aborder l'apprentissage supervisé et introduire un nouvel algorithme : le Nearest Neighbors (k-NN).
Ce modèle de machine learning nous permettra d'effectuer de la classification.



Qu'est-ce que la classification
l’apprentissage supervisé!!

La Classification (Supervisée)

La classification est une tâche fondamentale en apprentissage automatique supervisé.

Objectif :

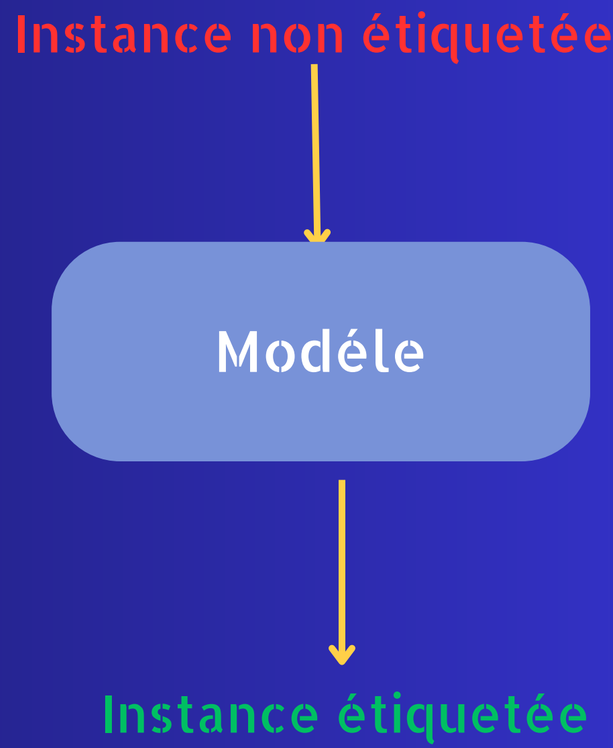
Prédire la classe ou la catégorie d'une nouvelle donnée(Instance) non étiquetée en se basant sur un ensemble de données d'entraînement déjà étiquetées .

Fonctionnement :

On entraîne un modèle sur les données étiquetées pour qu'il apprenne la relation entre les caractéristiques (features) des données et leurs étiquettes de classe (labels).

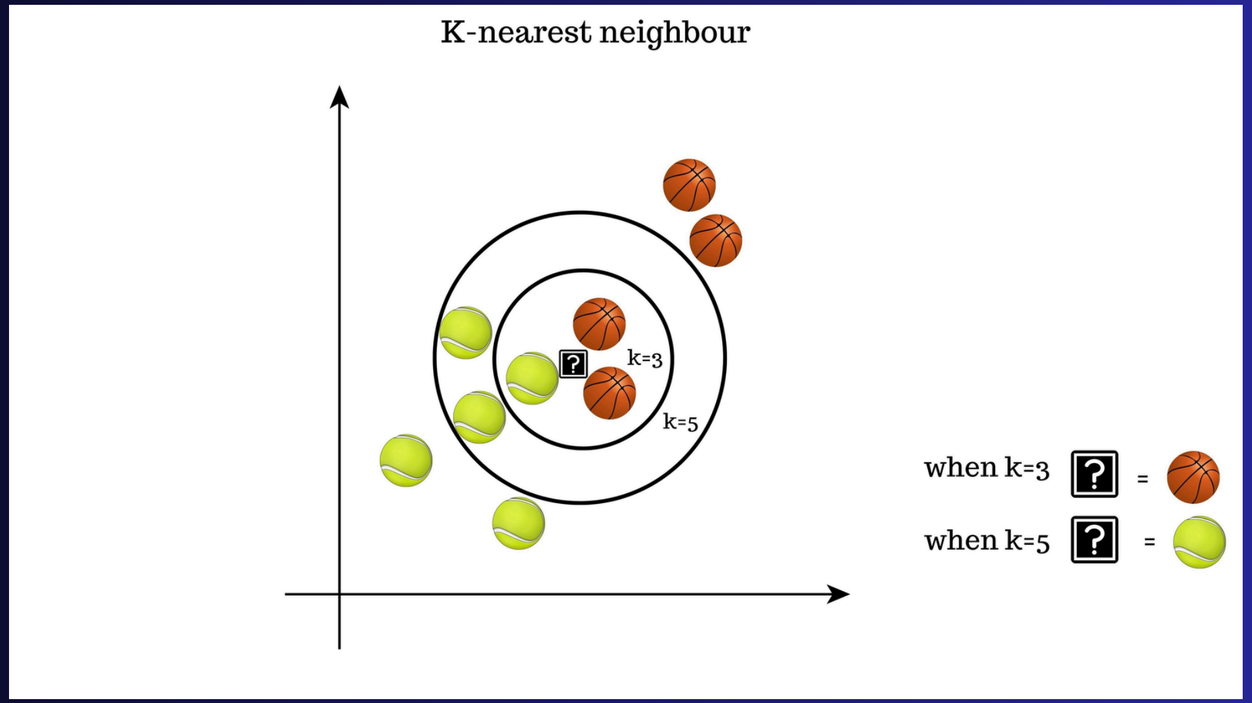
Exemples :

Déterminer si un email est spam ou non-spam (deux classes).
Reconnaître un chiffre manuscrit (classes 0 à 9).
Classifier un client comme acheteur ou non-acheteur.



K-NN

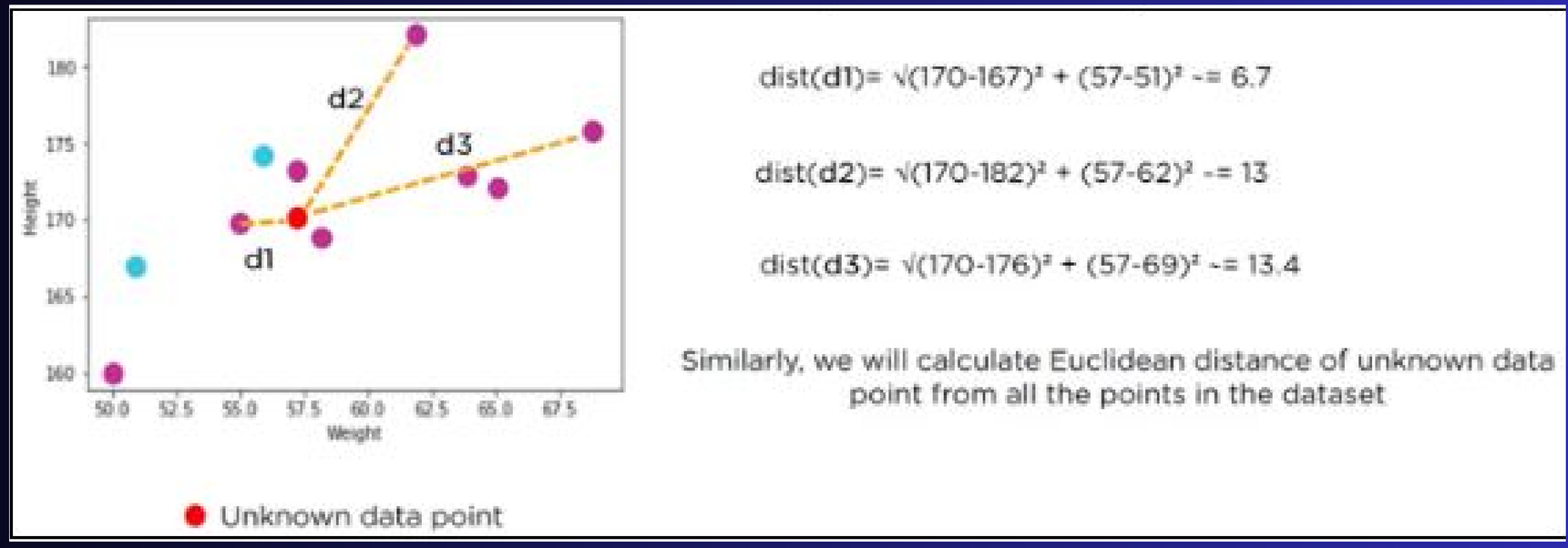
Le k-NN est un classifieur non-paramétrique, sont principe est de Classer une nouvelle donnée en lui attribuant la classe la plus fréquente parmi ses k plus proches voisins dans l'ensemble d'entraînement.



Les Étapes du k-NN

Le processus de classification d'une nouvelle donnée x_new par k-NN suit ces étapes :

- Choisir la valeur du paramètre k (le nombre de voisins à considérer).
- Calculer la distance entre la nouvelle donnée x_new et toutes les données de l'ensemble d'entraînement.(Distance Euclidienne)
- Sélectionner les k données d'entraînement qui ont la plus petite distance avec x_new .
- Identifier les classes de ces k voisins. La classe attribuée à x_new sera celle qui apparaît le plus fréquemment.
- Classer x_new dans la classe majoritaire.



Déroulement de l'algorithme KNN

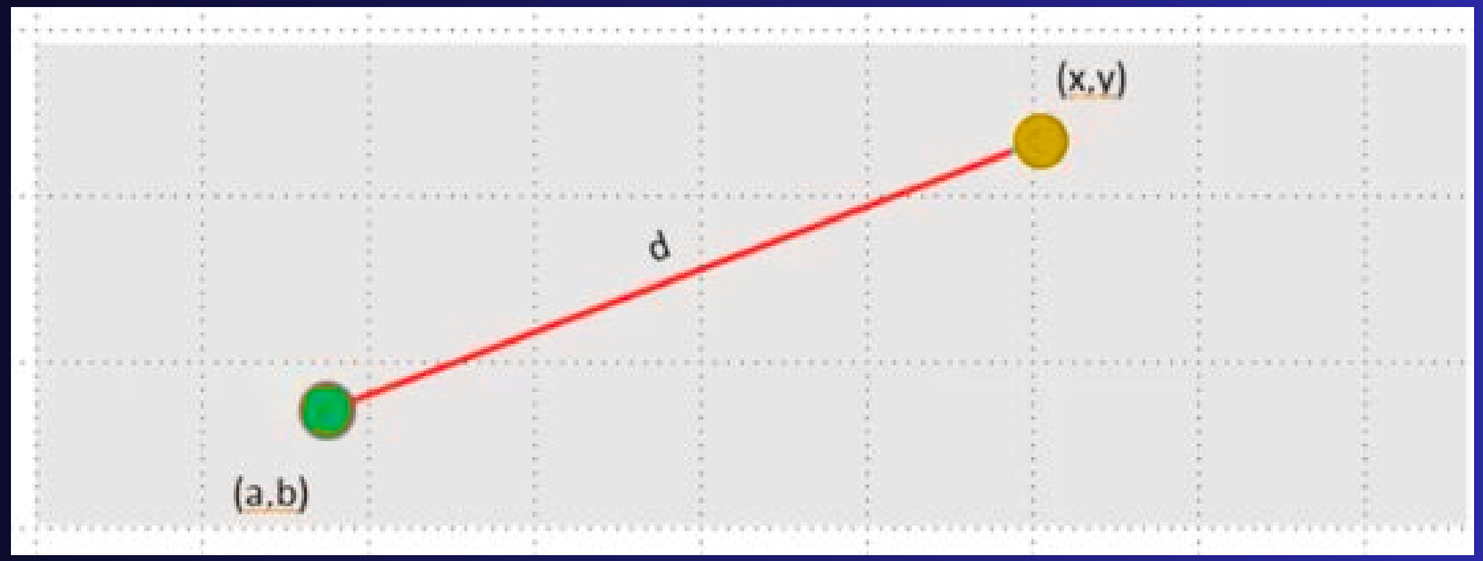
Considérons un ensemble de données contenant deux variables : la taille (cm) et le poids (kg). Chaque point est classé comme normal ou insuffisant en termes de poids.

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

Déroulement de l'algorithme KNN

Pour trouver les voisins les plus proches, nous calculerons la distance euclidienne. La distance euclidienne entre deux points du plan de coordonnées (x,y) et (a,b) est donnée par :

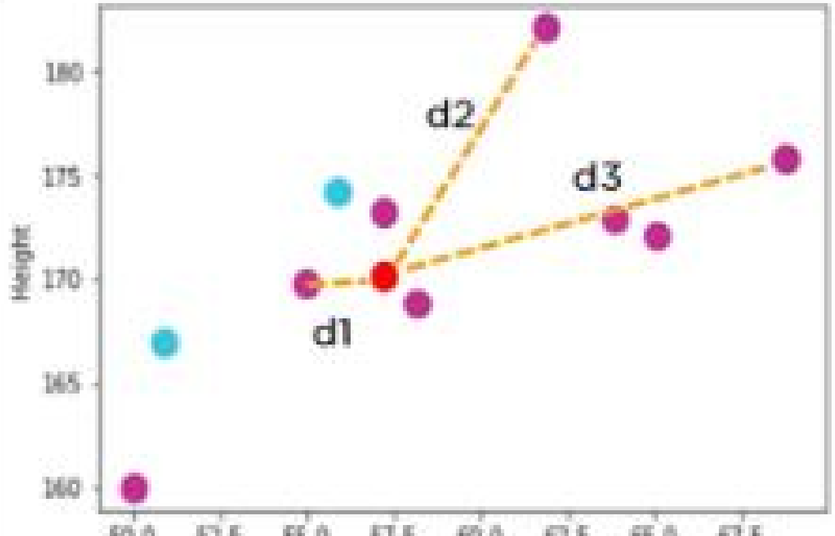
$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$



Déroulement de l'algorithme KNN

Maintenant, nous avons un nouveau point de données (x1, y1), et nous devons déterminer sa classe.

57 kg	170 cm	?



dist(d1)= √(170-167)² + (57-51)² ~= 6.7

dist(d2)= √(170-182)² + (57-62)² ~= 13

dist(d3)= √(170-176)² + (57-69)² ~= 13.4

Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

Unknown data point

Déroulement de l'algorithme KNN

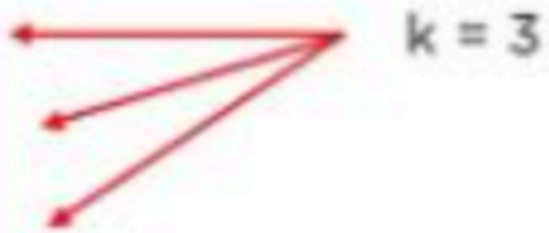
Le tableau suivant présente la distance euclidienne calculée des points de données inconnus par rapport à tous les autres points.

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

Déroulement de l'algorithme KNN

En regardant les nouvelles données, nous pouvons considérer les trois dernières lignes du tableau qui ont la plus petite distance avec x_new K=3.

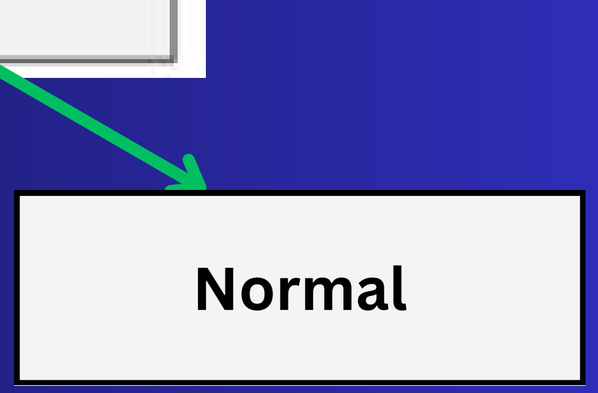
Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2



Déroulement de l'algorithme KNN

Étant donné que la majorité des voisins sont classés comme normaux selon l'algorithme KNN, le point de données (57, 170) devrait être normal.

57 kg	170 cm	?



La Matrice de Confusion

		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	
		Recall = TP / (TP + FN)		Accuracy = (TP + TN) / (TP + FP + TN + FN)