



INTRODUCTION DATA INTELLIGENCE & DATA SCIENCE

MODULE 2 – MACHINE LEARNING

Informatique – orientation IA – 1DA/IA

PLAN

- Introduction & définitions
- Méthodes supervisées
- Méthodes non supervisées

INTRODUCTION & DÉFINITIONS



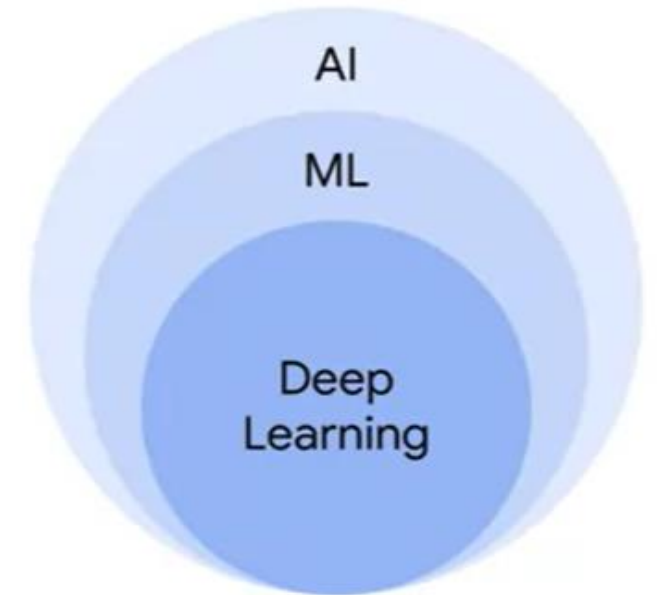
INTRODUCTION & DÉFINITIONS

INTELLIGENCE ARTIFICIELLE

“L’Intelligence Artificielle est l’ensemble des théories et des techniques développant des programmes informatiques complexes- capables de simuler certains traits de l’intelligence humaine (raisonnement, apprentissage, ...)”

Le petit Robert

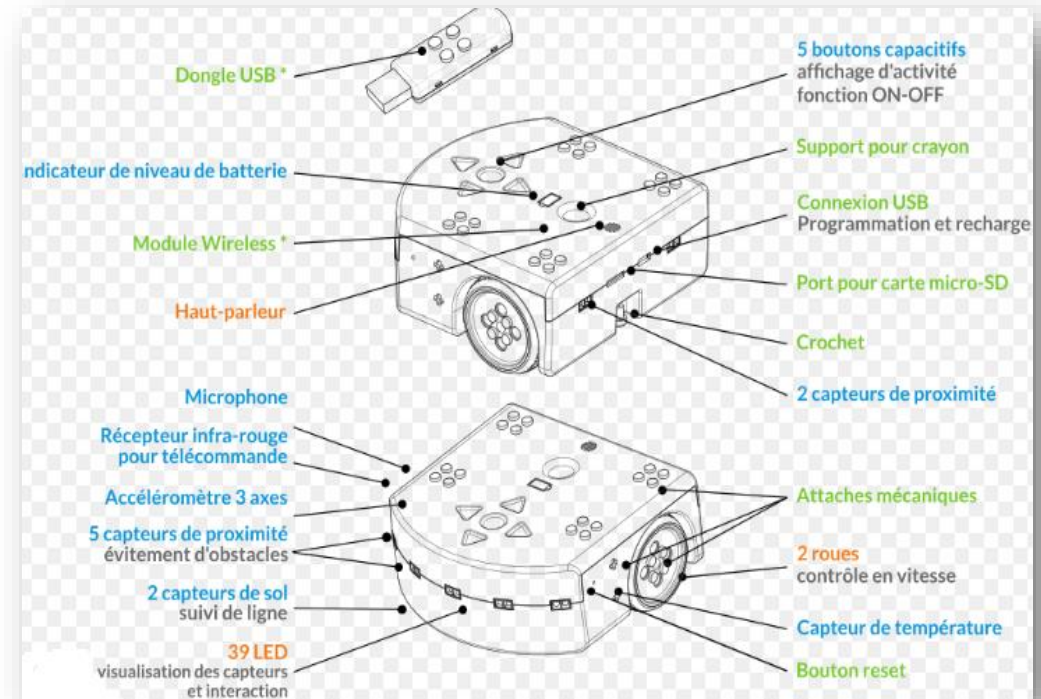
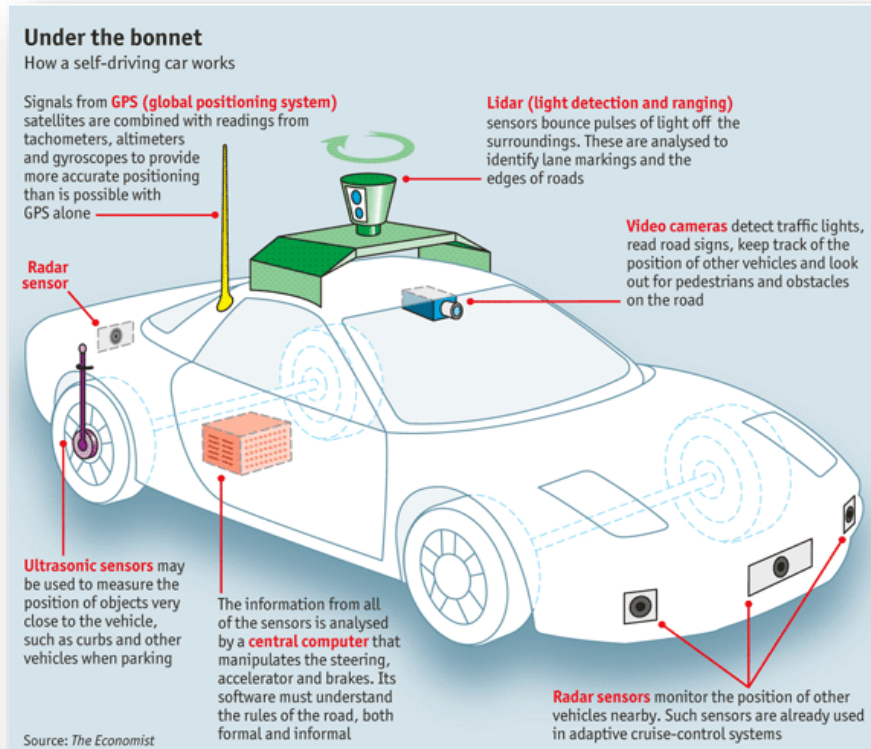
- Pas de définition universelle
- « Intelligence » en anglais signifie plus renseignements qu’intelligence
- Encore loin d’imiter l’intelligence
- Technologies spécifiques



INTRODUCTION & DÉFINITIONS

INTELLIGENCE ARTIFICIELLE - APPLICATIONS

- Voiture autonome : GPS , analyse d'images, détection d'obstacles, ...



INTRODUCTION & DÉFINITIONS

INTELLIGENCE ARTIFICIELLE - APPLICATIONS

- Utilisation courante et potentielle



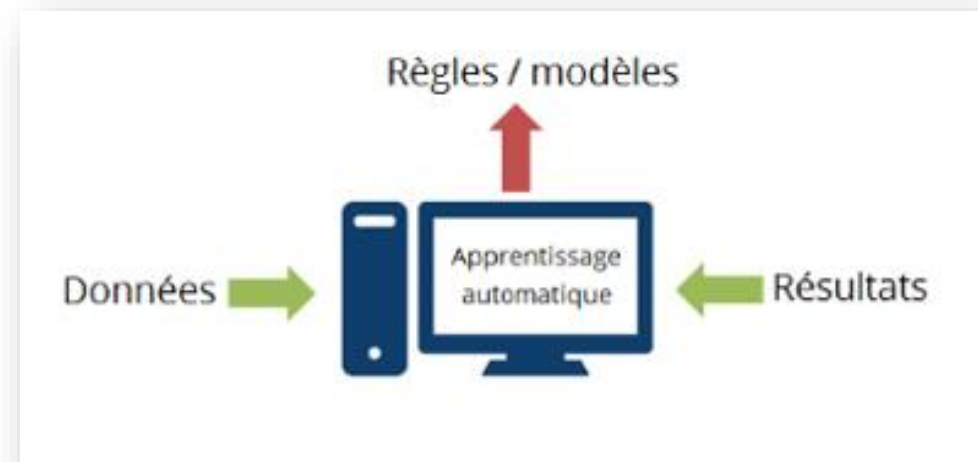
INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – MACHINE LEARNING

L'apprentissage automatique consiste à « **donner à une machine la capacité d'apprendre sans la programmer de façon explicite** »

Arthur Samuel (pionnier de l'IA)

- « apprendre à apprendre » à l'ordinateur
- Pas de programmation explicite
- Ajustements successifs



INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- **Contexte**

Vous travaillez pour une entreprise alimentaire renommée qui produit une gamme de céréales pour petits-déjeuners.

Vous voudriez proposer un nouveau produit de très haute qualité nutritionnelle. Vous souhaitez donc savoir quels facteurs influencent (expliquent) ce taux nutritionnel.

Pour ce faire, vous avez récemment mené des analyses nutritionnelles approfondies sur plusieurs produits de la gamme afin de mieux comprendre leur composition. Les données recueillies comprennent la quantité de sucre (en grammes par portion), la quantité de fibres et beaucoup d'autres paramètres ainsi que le taux nutritionnel global.

Dans un premier temps, votre supérieur vous a demandé d'explorer la relation entre un des paramètres à disposition : le sucre et la qualité nutritionnelle des céréales. Il souhaite modéliser le lien existe entre ce paramètre et la qualité nutritionnelle des produits.

INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- Etape 1 : Les données (le « dataset »)

Features /
variables

Target / cible

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	100%_Bran	N	C	70	4	1	130	10.0	5.0	6	280	25	3	1.0	0.33	68.402973
1	100%_Natural_Bran	Q	C	120	3	5	15	2.0	8.0	8	135	0	3	1.0	1.00	33.983679
2	All-Bran	K	C	70	4	1	260	9.0	7.0	5	320	25	3	1.0	0.33	59.425505
3	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14.0	8.0	0	330	25	3	1.0	0.50	93.704912
4	Almond_Delight	R	C	110	2	2	200	1.0	14.0	8	-1	25	3	1.0	0.75	34.384843
...
72	Triples	G	C	110	2	1	250	0.0	21.0	3	60	25	3	1.0	0.75	39.106174
73	Trix	G	C	110	1	1	140	0.0	13.0	12	25	25	2	1.0	1.00	27.753301
74	Wheat_Chex	R	C	100	3	1	230	3.0	17.0	3	115	25	1	1.0	0.67	49.787445
75	Wheaties	G	C	100	3	1	200	3.0	17.0	3	110	25	1	1.0	1.00	51.592193
76	Wheaties_Honey_Gold	G	C	110	2	1	200	1.0	16.0	8	60	25	1	1.0	0.75	36.187559

Observation

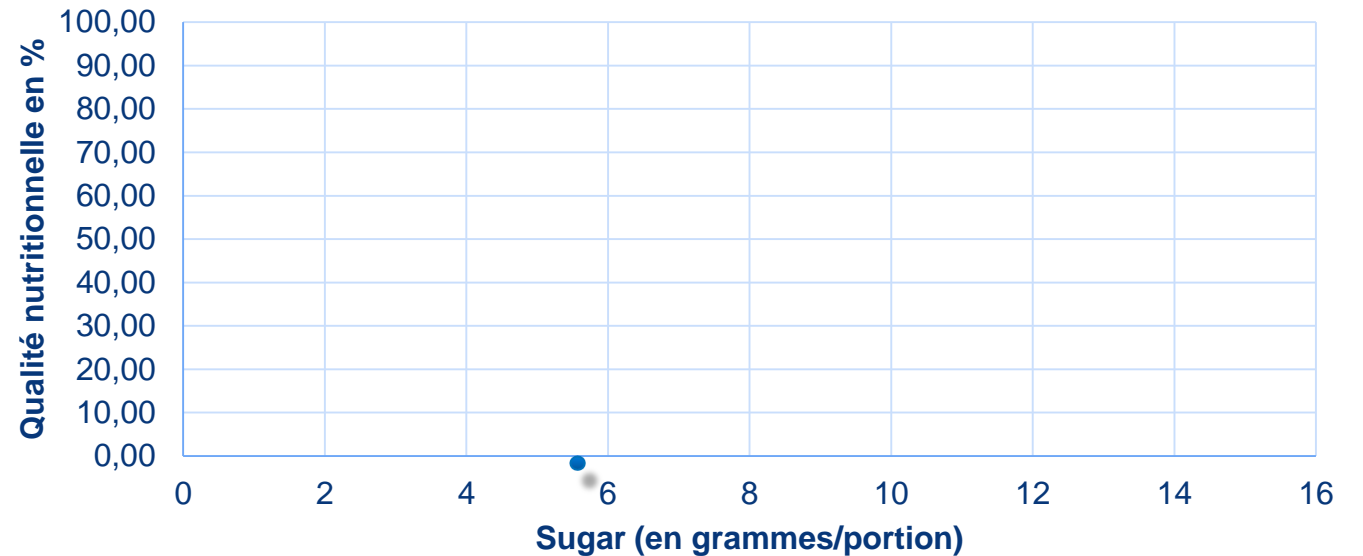
céréales.xlsx

INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- **Etape 2 : « features engineering »**
 - Etape de préparation des données
 - Choix de variables
- Ici, choix porté sur le paramètre « sugar »

Nuage de points sugar - rating
Source : ... - data ...



INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- **Etape 3 : Choix de l'algorithme**

- Dépend du problème à résoudre, du type et du volume de données ;
- Recherche de la valeur optimale → réduction du coût ;

- **Qu'est ce qu'un algorithme ?**

Ensemble de règles opératoires dont l'application permet de résoudre un problème énoncé au moyen d'un nombre fini d'opérations. Un algorithme peut être traduit, grâce à un langage de programmation, en un programme exécutable par un ordinateur.

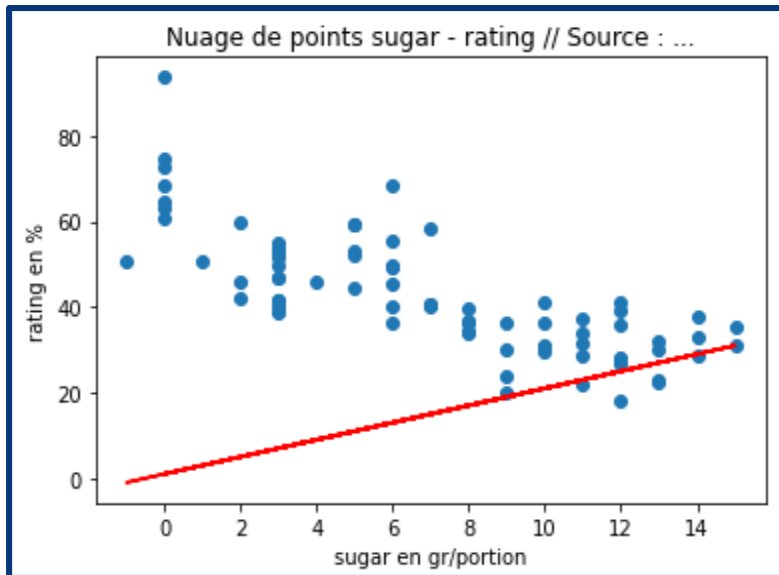
- Dans notre exemple :

Choix de l'algorithme : **Régression linéaire simple** ($y = ax+b$)

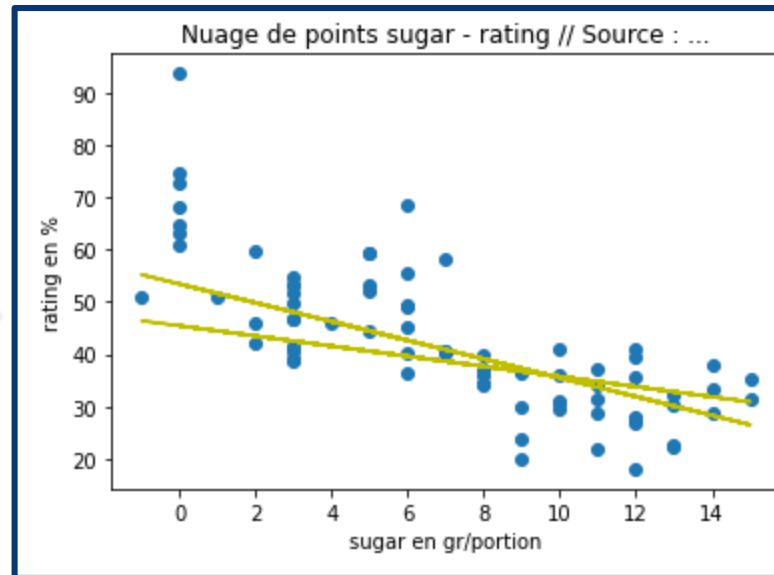
INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

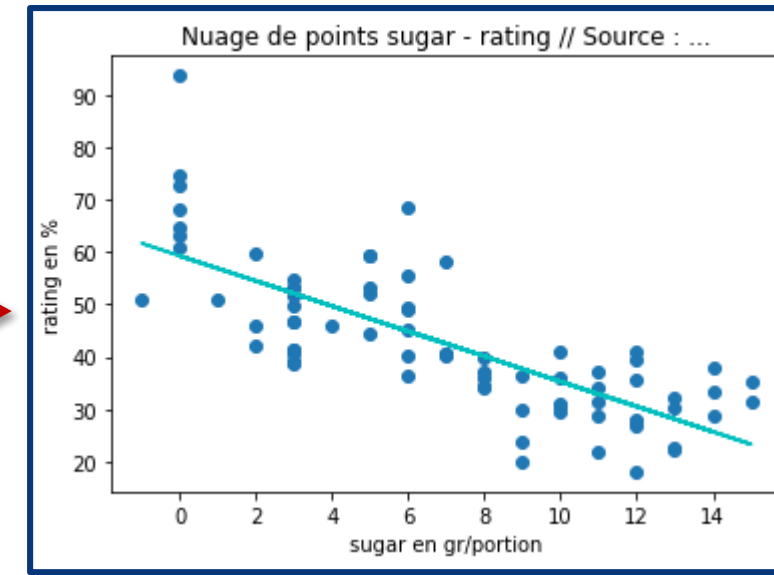
- Etape 4 : Entrainement et itérations



Premier essai



Intermédiaire



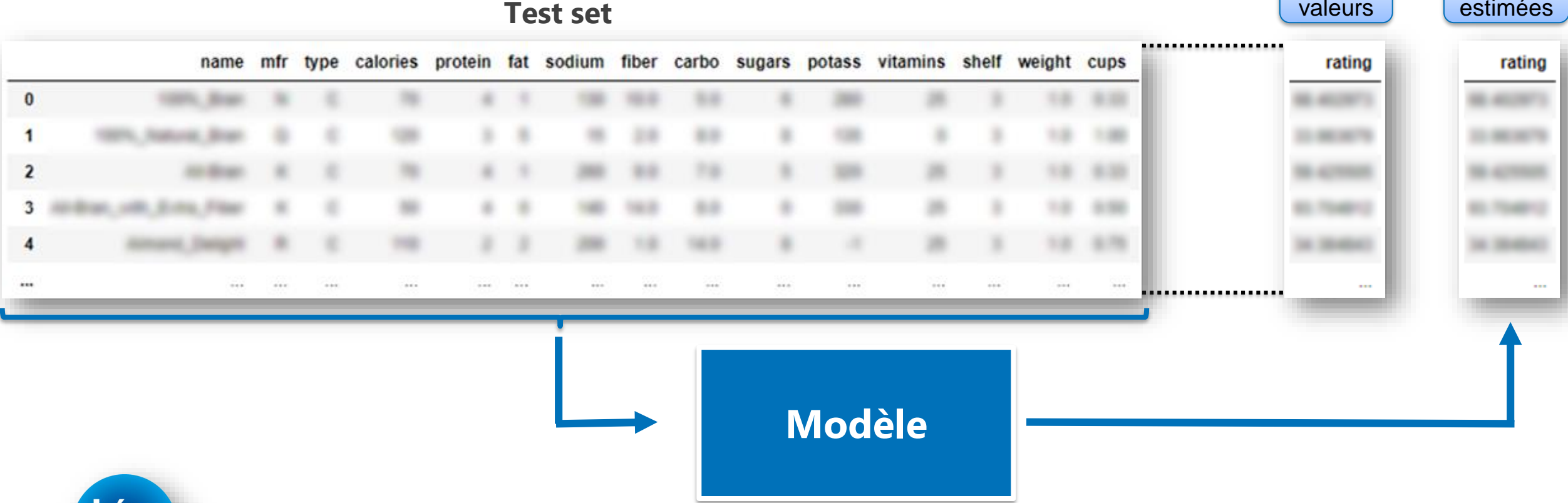
Après convergence

$$\rightarrow y = -2,45x + 59,3$$

INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- Etape 5 : Evaluation du modèle



INTRODUCTION & DÉFINITIONS

APPRENTISSAGE AUTOMATIQUE – EXEMPLE

- **Etape 6 : Utilisation du modèle**
 - Quid de la qualité nutritionnelle d'une céréale si on augmente la quantité de sucre de 1gr ?
 - Quelle sera la valeur de la qualité nutritionnelle d'une céréale si on y incorpore 5gr de sucre ?
 - Quid du lien entre le sucre et qualité nutritionnelle ?

INTRODUCTION & DÉFINITIONS

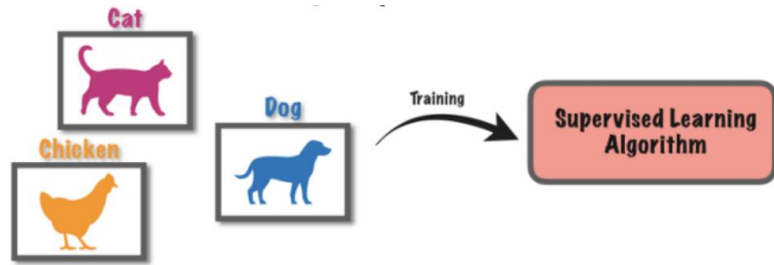
RÉSUMÉ

- **Une observation** : correspond à une ligne du dataset. Chaque observation représente un enregistrement spécifique dans le dataset. Dans l'exemple ci-dessus, chaque ligne du dataset contient les caractéristiques d'une céréale.
- **Un dataset (ou jeu de données)** : regroupe un ensemble de données cohérentes qui peuvent se présenter sous différents formats. Il est composé d'observations.
- **Train set** : sous-ensemble des observations sur lequel l'algorithme entraîne le modèle.
- **Test set** : sous-ensemble des observations sur lequel le modèle sera évalué. Il doit contenir des observations qui n'ont pas été utilisées pour l'entraînement.
- **Target ou cible** : est la variable dont on veut que le modèle trouve la valeur
- **Indicateur de performance** : est une mesure utilisée pour évaluer la qualité d'un modèle d'apprentissage automatique. Cette mesure dépend de l'algorithme utilisé ainsi que du contexte.
- **Modèle** : ce qui est donné pour servir de référence. Dans l'exemple ci-dessus, $y = - 2.4 * x + 59.3$

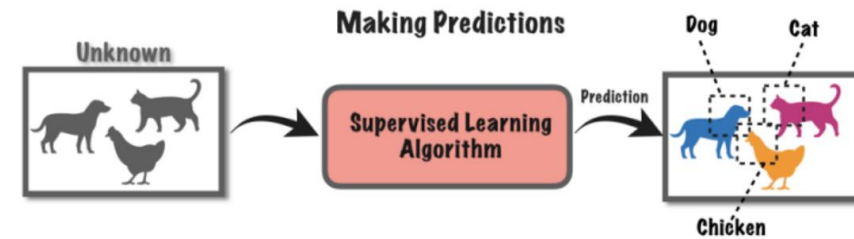


INTRODUCTION & DÉFINITIONS

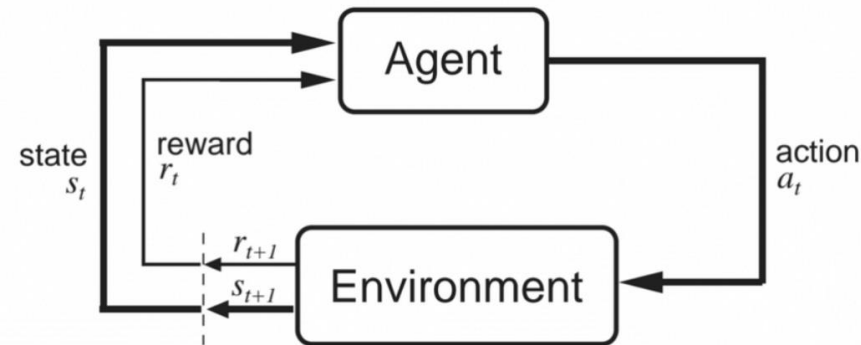
LE MACHINE LEARNING



Apprentissage supervisé



Apprentissage non supervisé



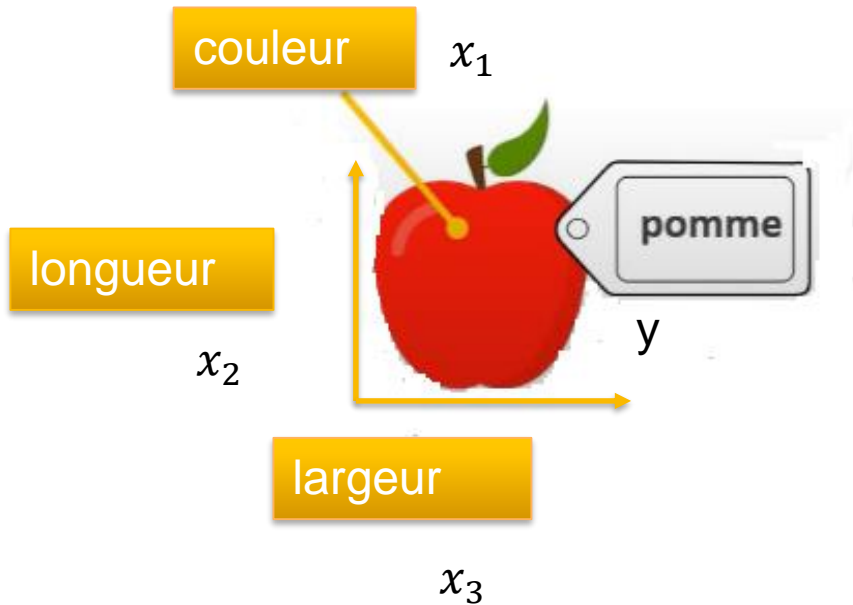
Apprentissage par renforcement

MÉTHODES SUPERVISÉES



APPRENTISSAGE SUPERVISÉ

TYPES D'ALGORITHMES



CLASSIFICATION

A diagram illustrating supervised learning for regression. It shows a table with house data. A vertical double-headed arrow labeled m indicates the number of rows (samples). A horizontal double-headed arrow labeled n indicates the number of columns (features plus target).

Target y	Features		
	x_1	x_2	x_3
Prix	Surface	Qualité	Adresse postale
313,000	90	3	95000
720,000	110	5	93000
250,000	40	4	44500
290,000	60	3	67000
190,000	50	3	59300
...

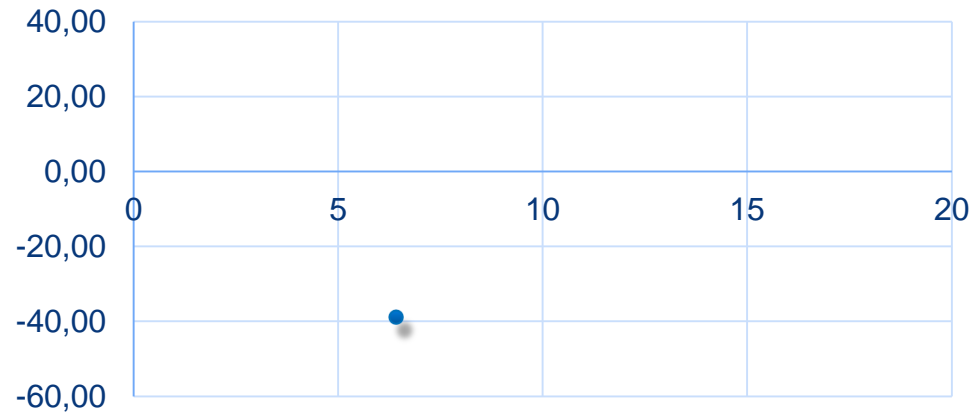
RÉGRESSION

APPRENTISSAGE SUPERVISÉ

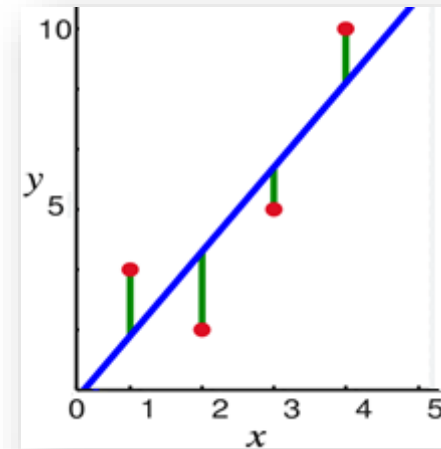
LA RÉGRESSION LINÉAIRE SIMPLE

- Simplifions le modèle : $y = ax$;

rating-mod



Train set



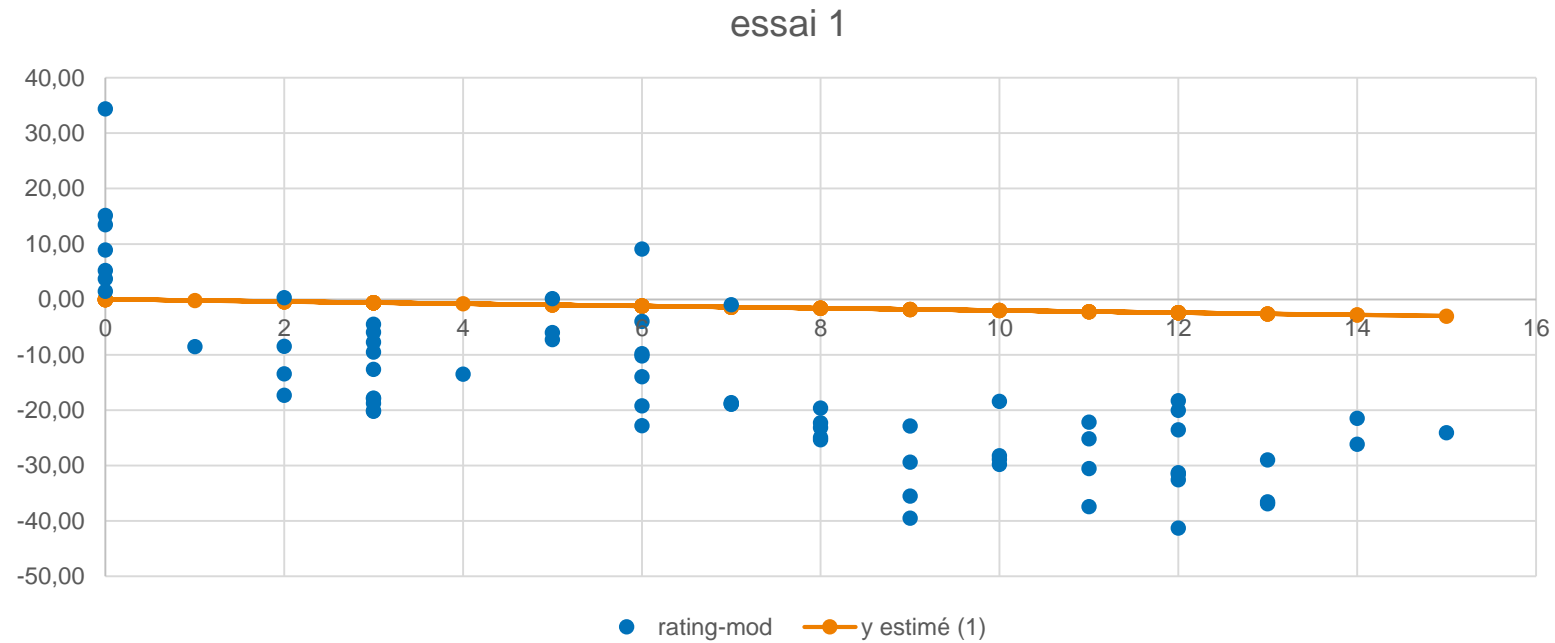
$y = ax$



$$J(a) = \frac{1}{2m} * \sum_{i=1}^m (y_i - ax_i)^2 \text{ où } m \text{ est le nombre d'observations.}$$

APPRENTISSAGE SUPERVISÉ

LA RÉGRESSION LINÉAIRE SIMPLE (EXEMPLE)

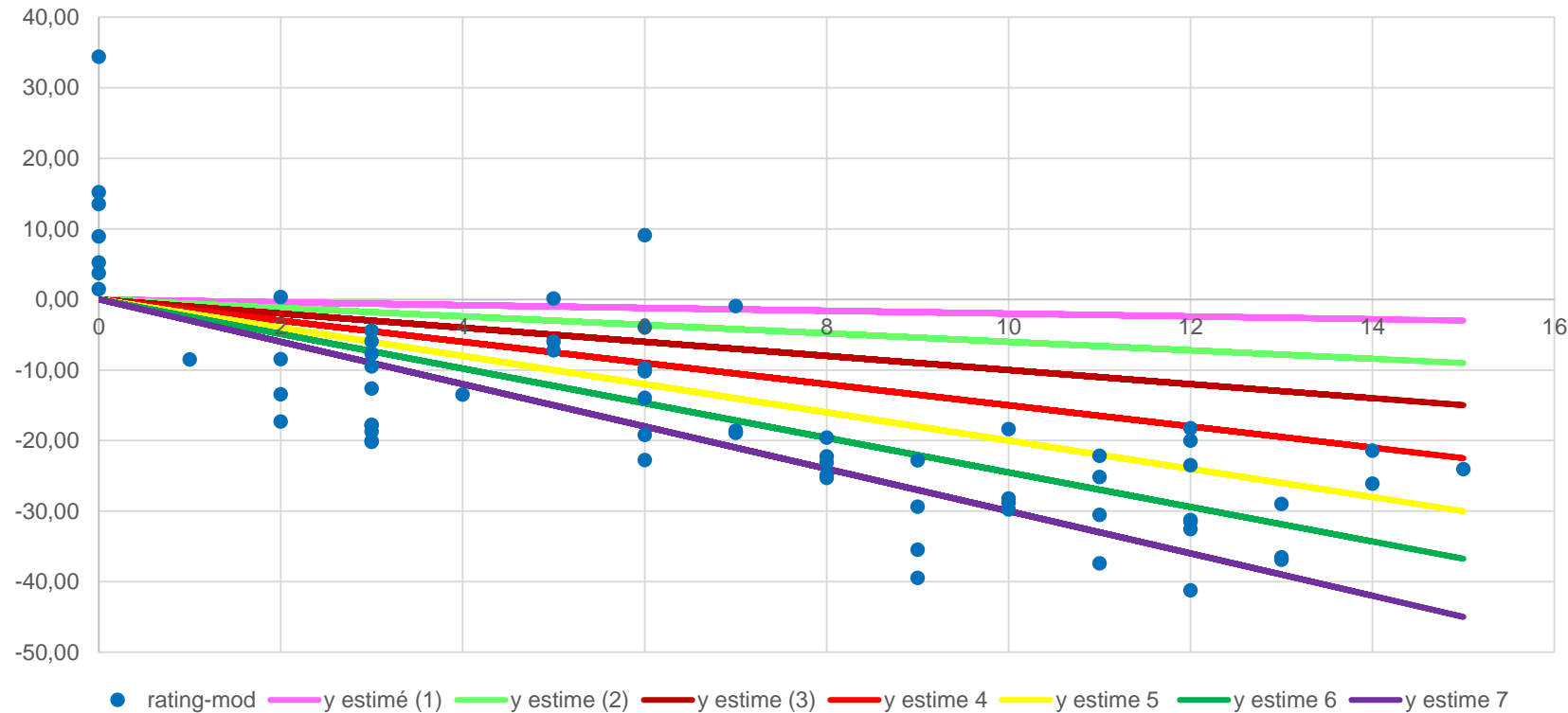


Avec $y = -0,2x$

APPRENTISSAGE SUPERVISÉ

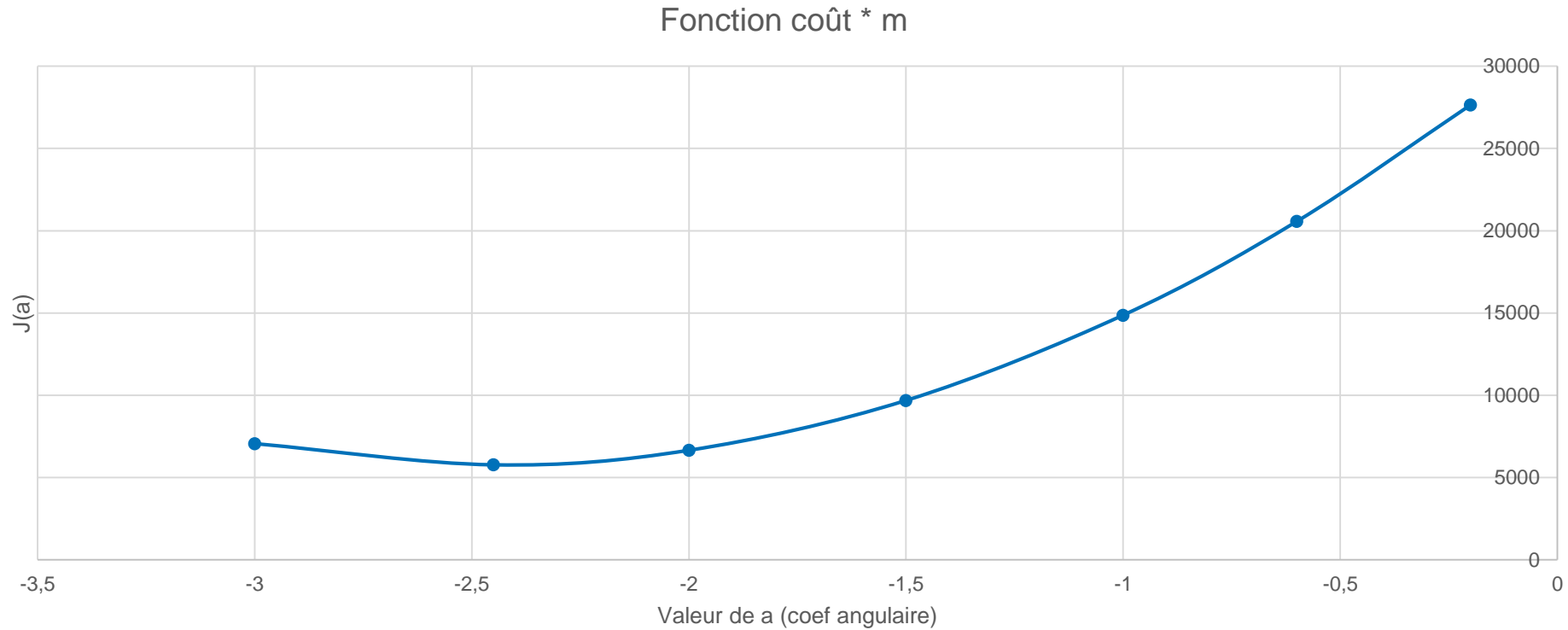
LA RÉGRESSION LINÉAIRE SIMPLE (EXEMPLE)

essais 1 à 7



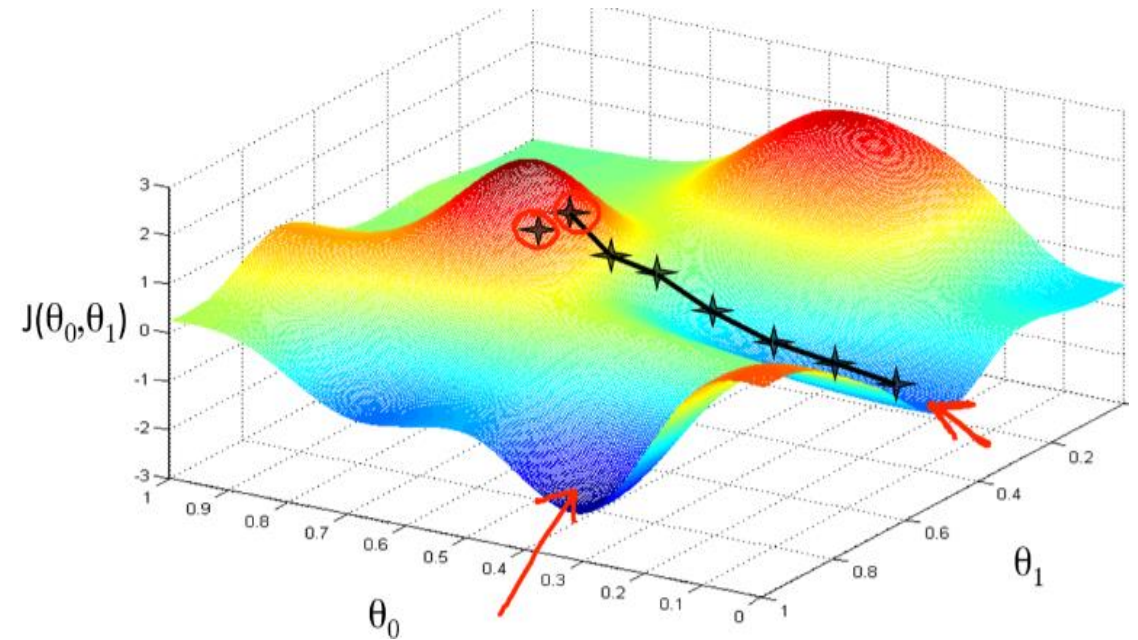
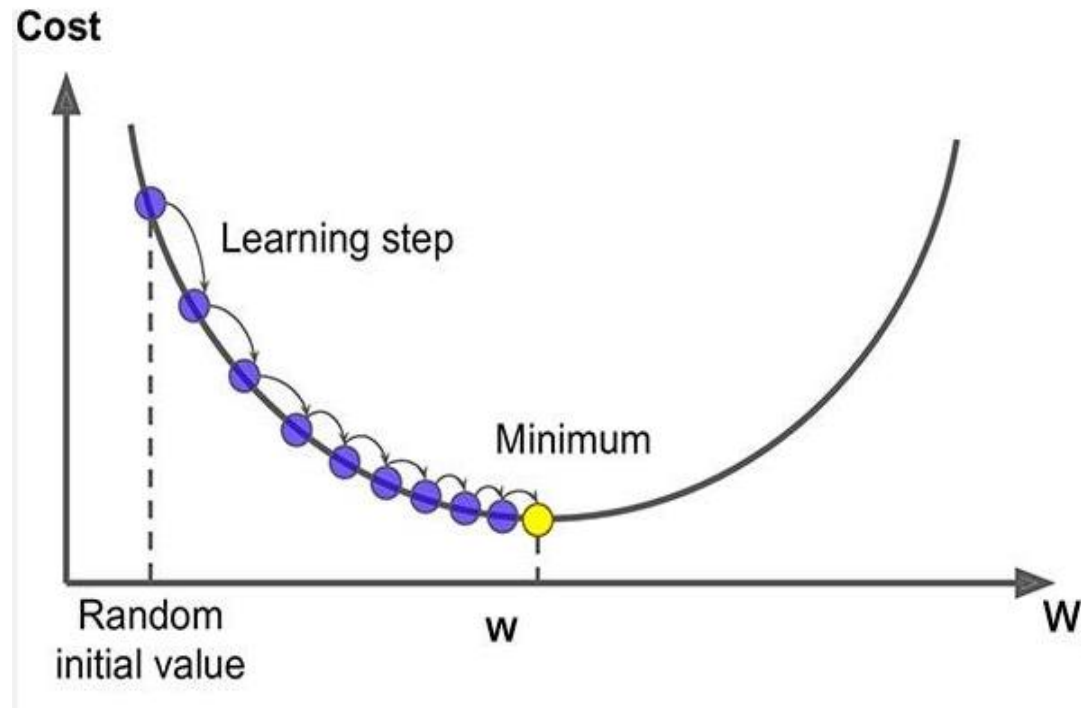
APPRENTISSAGE SUPERVISÉ

LA RÉGRESSION LINÉAIRE SIMPLE (EXEMPLE)



APPRENTISSAGE SUPERVISÉ

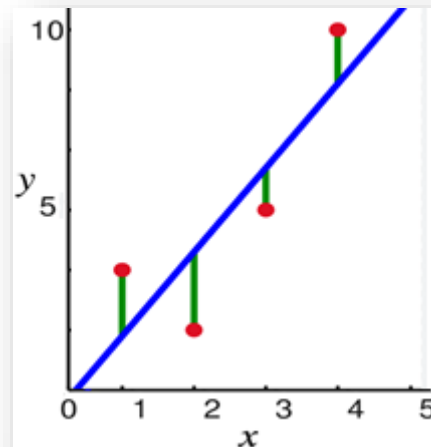
LA RÉGRESSION LINÉAIRE SIMPLE – LE GRADIENT DESCENT



APPRENTISSAGE SUPERVISÉ

LA RÉGRESSION LINÉAIRE SIMPLE (EXEMPLE)

- **Evaluation du modèle :** $RMSE = \sqrt{\frac{1}{n} * (y_i - 2.45 * x_i)^2}$



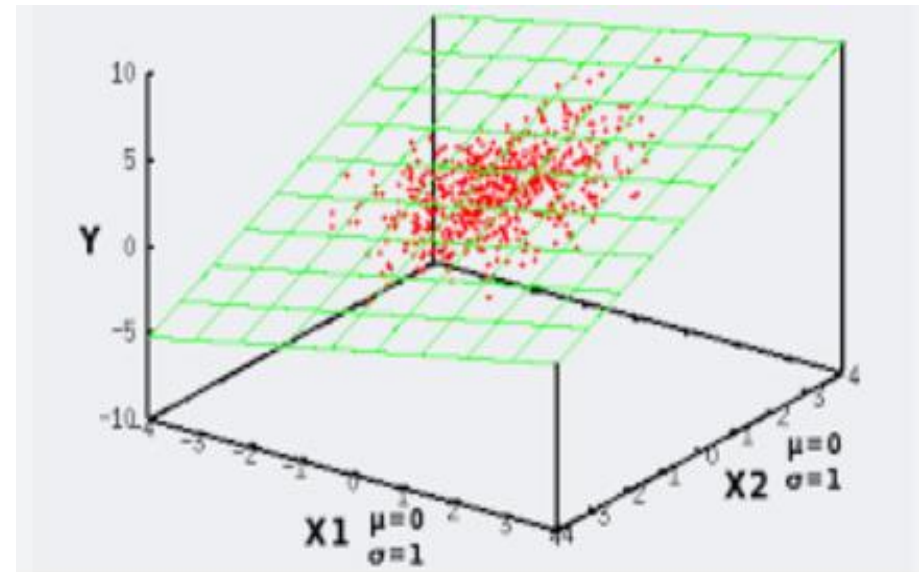
APPRENTISSAGE SUPERVISÉ

LA RÉGRESSION LINÉAIRE MULTIPLE

- Fonction de la régression linéaire multiple : $y = a_0 + a_1x_1 + \dots + a_nx_n$
 - Avec
 - y : la cible
 - a_i : les coefficients recherchés
 - x_i : les variables explicatives (les features)



Les variables x_i doivent être standardisées !



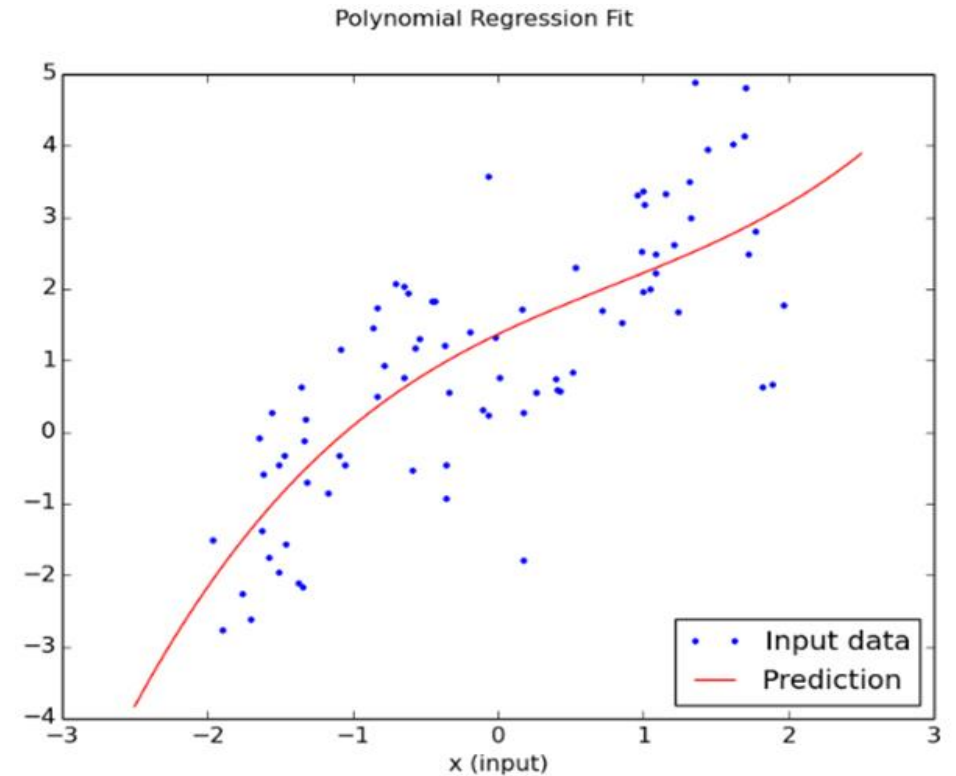
APPRENTISSAGE SUPERVISÉ

LA RÉGRESSION POLYNOMIALE

- **Fonction de la régression polynomiale :** $y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$
 - Avec
 - **y** : la cible
 - **a_i** : les coefficients recherchés
 - **x** : la variable explicative

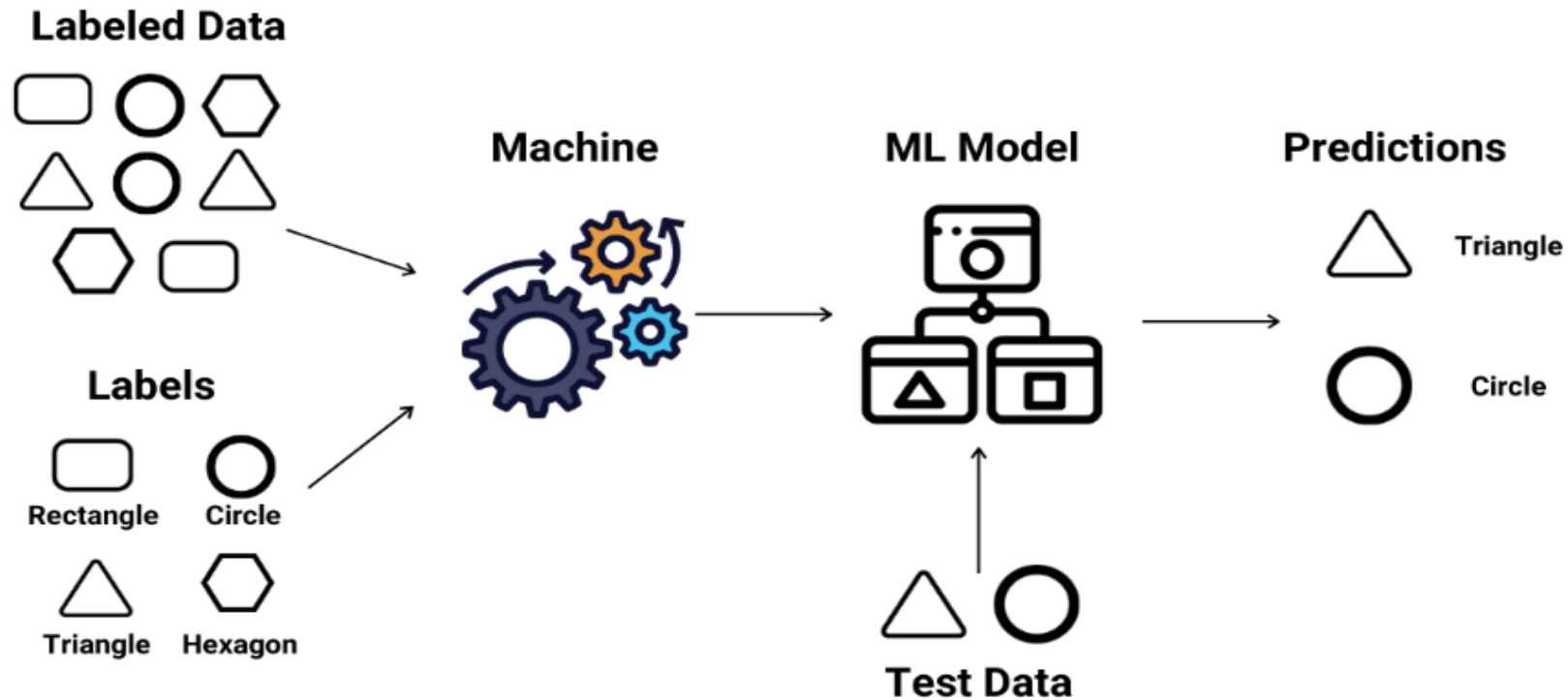


La variable x doit être standardisée !



APPRENTISSAGE SUPERVISÉ

LA CLASSIFICATION



APPRENTISSAGE SUPERVISÉ

LA CLASSIFICATION - KNN

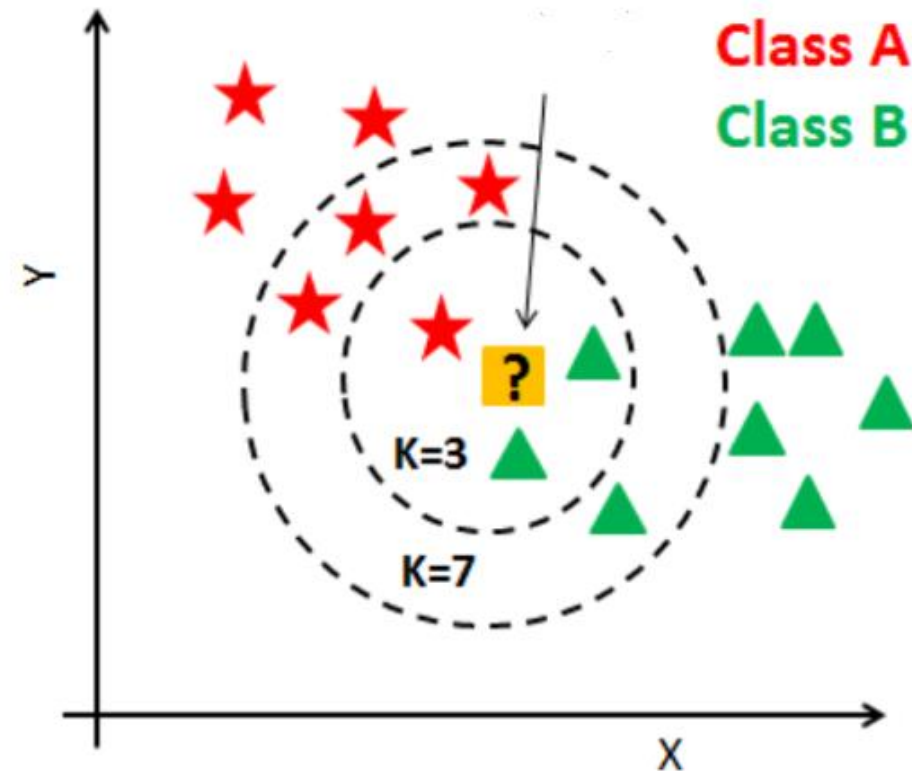
- KNN : K Nearest Neighbors

- Distance calculée par :

$$\text{dist}((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

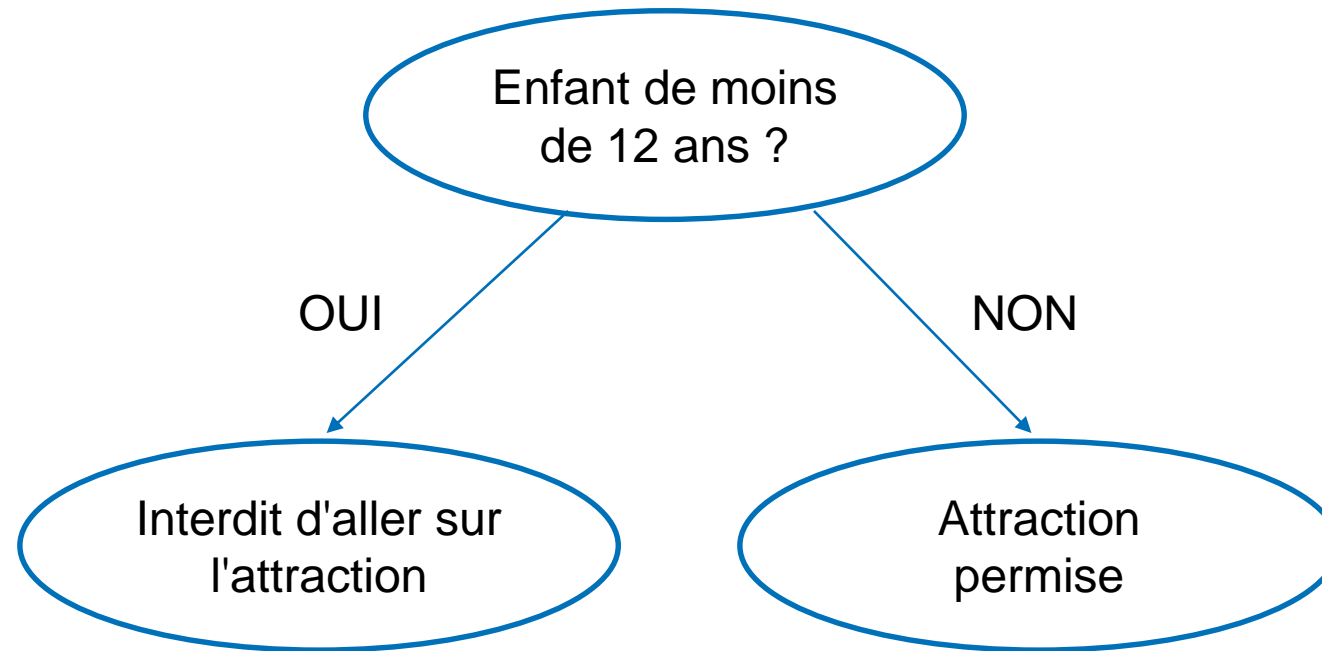
- Indicateur de performance :

$$\text{accuracy} = \frac{\text{nbre bien classés}}{\text{nbre total}} * 100$$



ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)

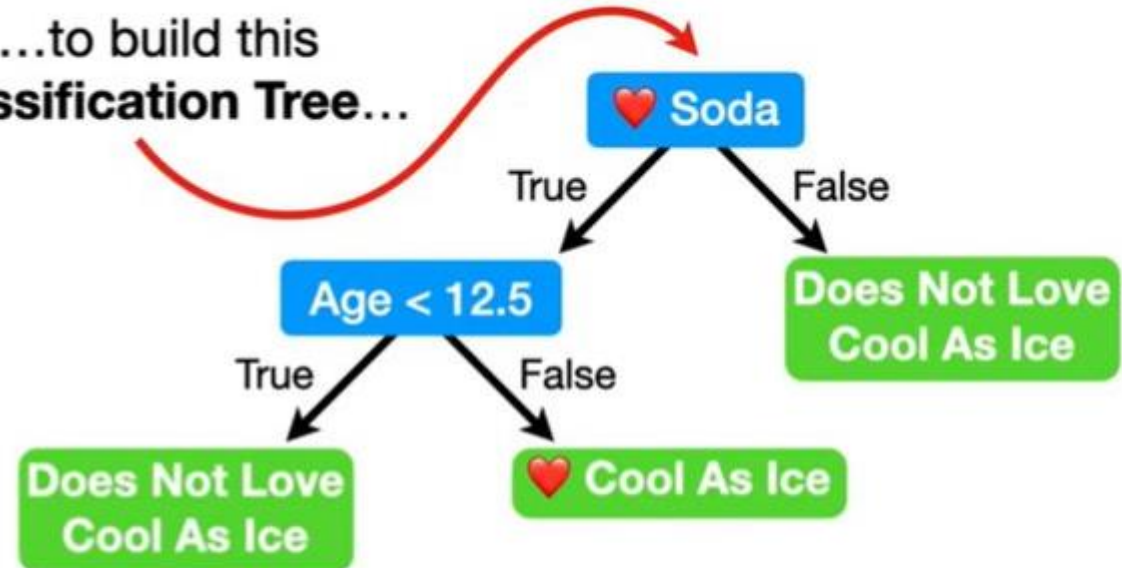


ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

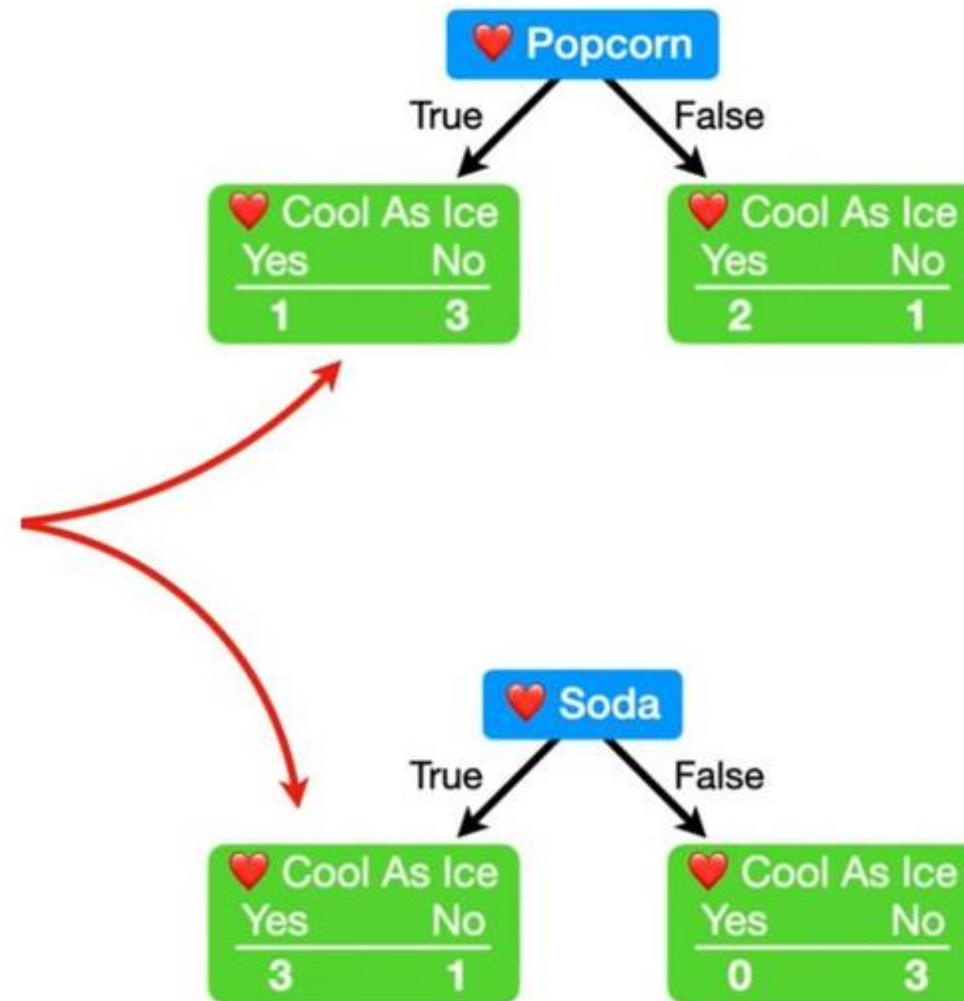
...to build this
Classification Tree...



ALGORITHMES DE CLASSIFICATION

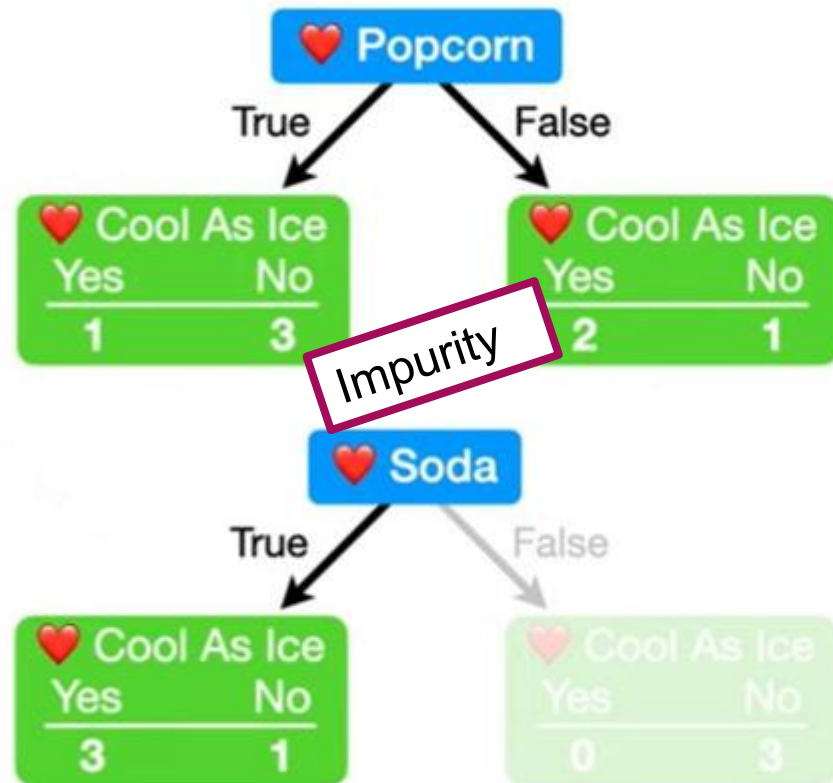
ARBRE DE DÉCISION (DECISION TREE)

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

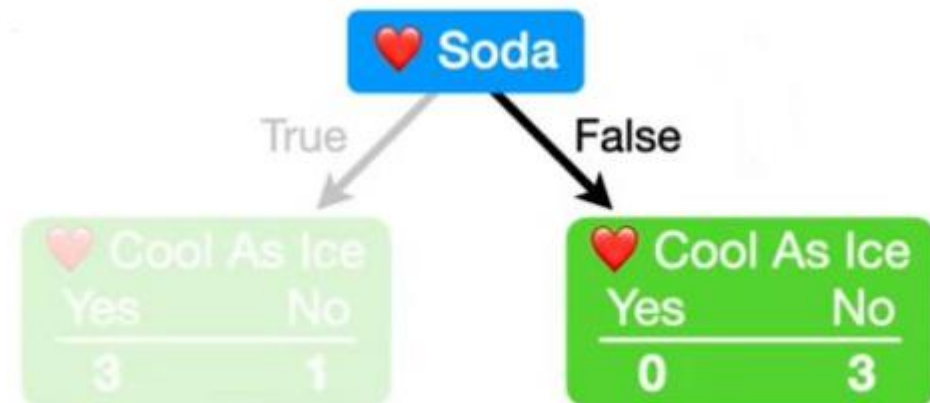


ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)



Comment quantifier la différence entre « aimer les pop corns » et « aimer les sodas »?

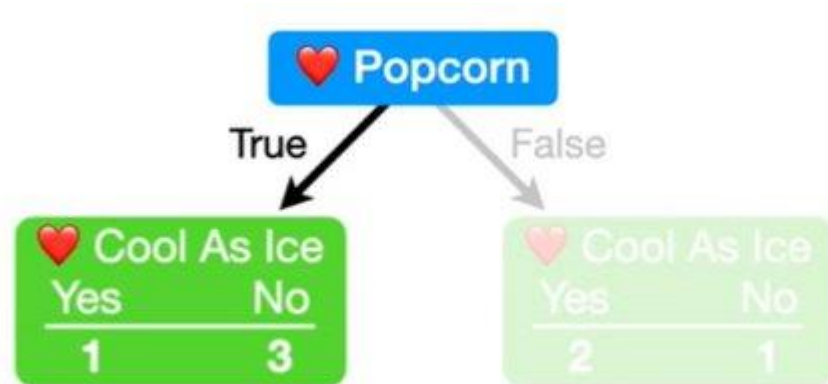


Plusieurs méthodes

- Giny Impurity
- Entropy and information gain

ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)



Gini Impurity for a Leaf = $1 - (\text{the probability of "Yes"})^2 - (\text{the probability of "No"})^2$

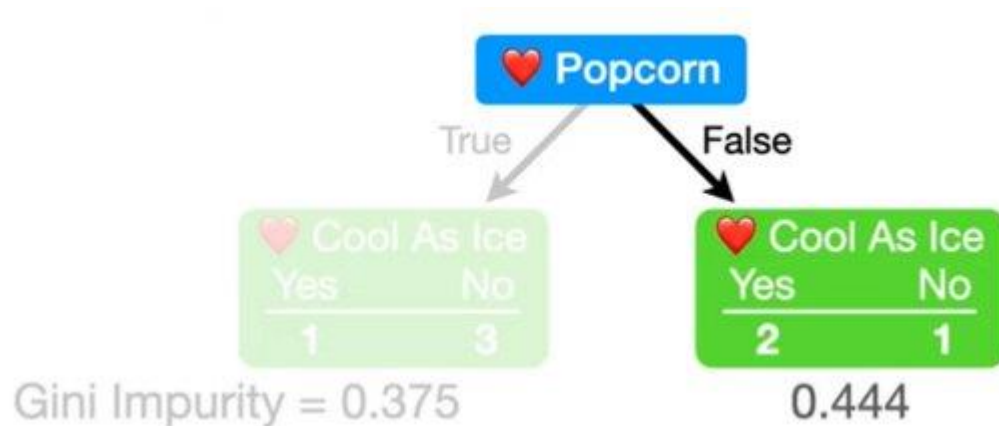
$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$$= 0.375$$

And when we do the math, we get **0.375**.

ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)



Gini Impurity for a Leaf = $1 - (\text{the probability of "Yes"})^2 - (\text{the probability of "No"})^2$

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2$$

$$= 0.444$$

And when we do the math we get **0.444**.

ALGORITHMES DE CLASSIFICATION

ARBRE DE DÉCISION (DECISION TREE)

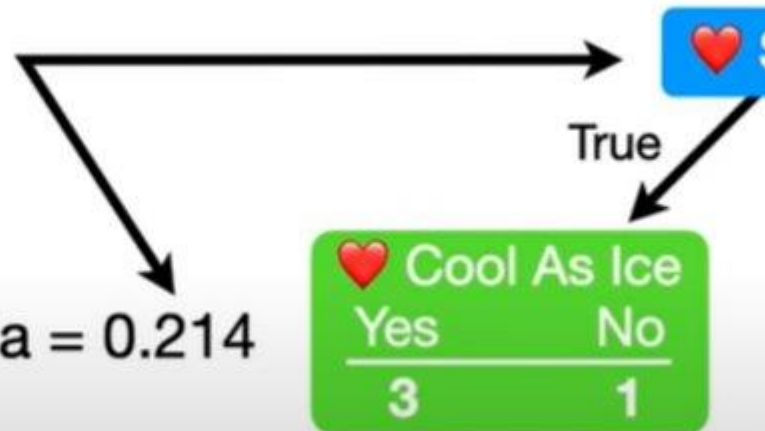
Total **Gini Impurity** = weighted average of **Gini Impurities** for the

$$= \left(\frac{4}{4+3} \right) 0.375 + \left(\frac{3}{4+3} \right) 0.444$$

$$= 0.405$$

Likewise, the **Gini Impurity** for **Loves Soda** is **0.214**.

Gini Impurity for Loves Soda = 0.214



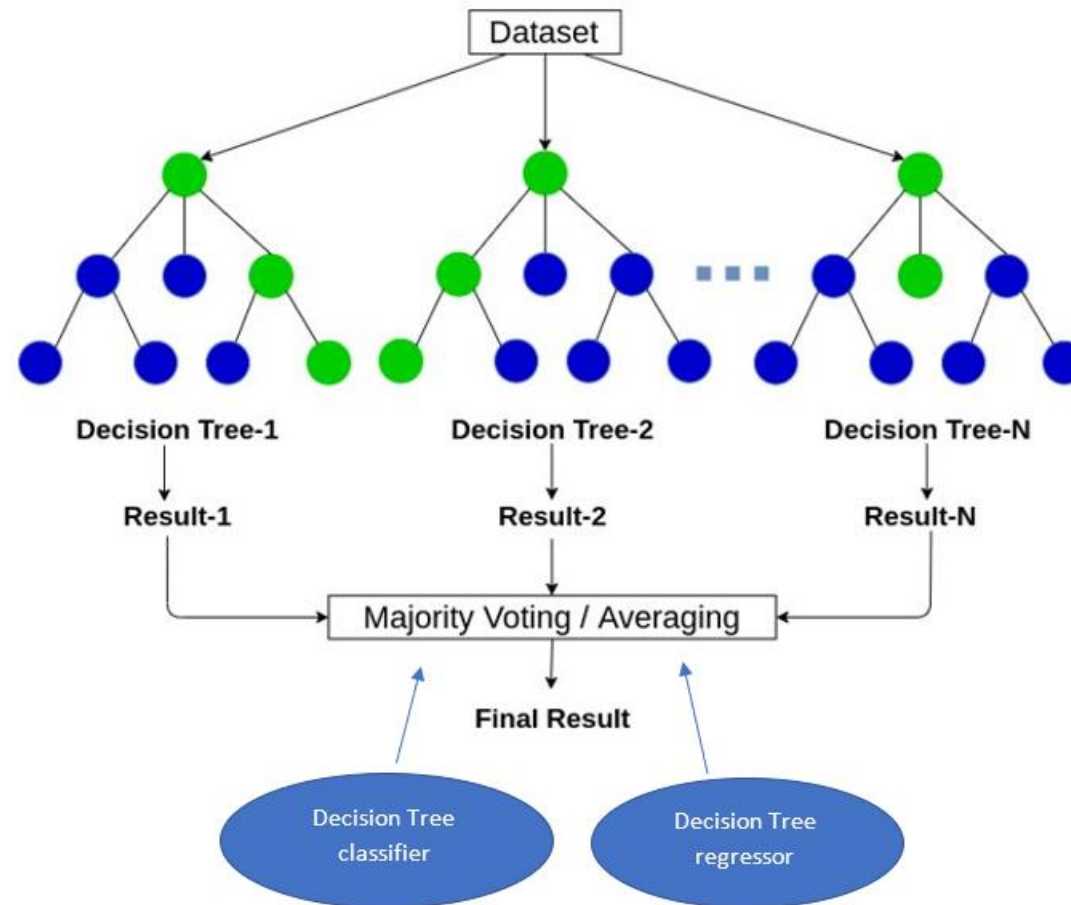
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

A green leaf node from the decision tree, containing the following data:

Cool As Ice	
Yes	No
0	3

APPRENTISSAGE SUPERVISÉ

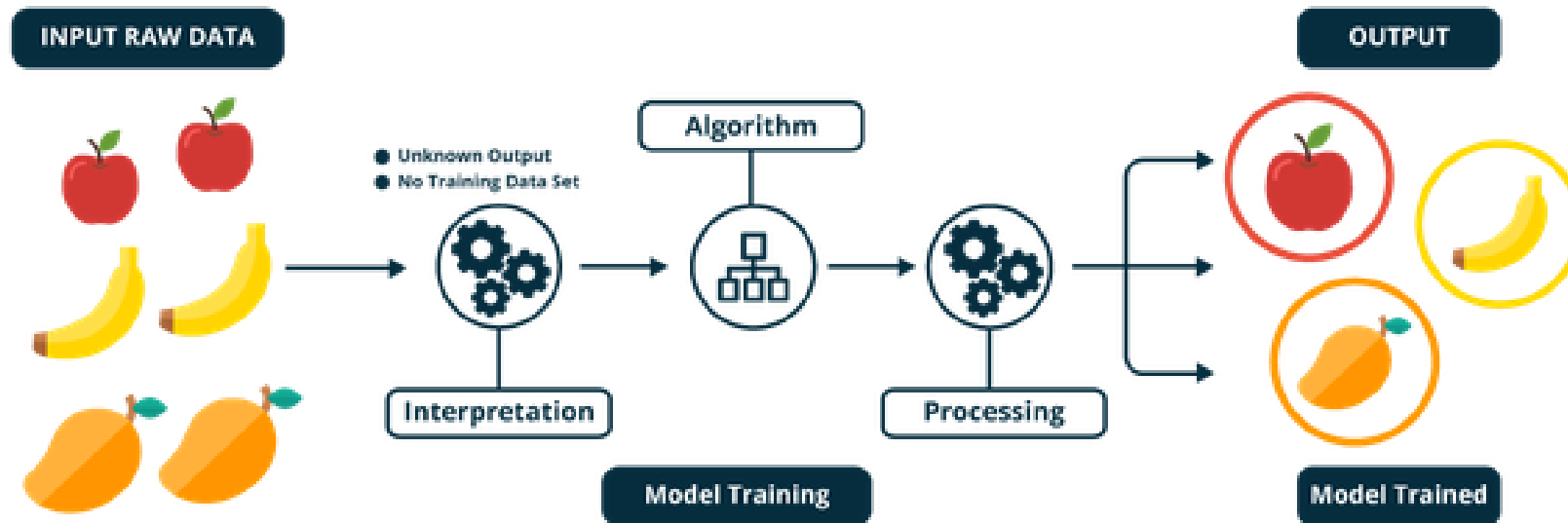
RANDOM FOREST



MÉTHODES NON SUPERVISÉES

APPRENTISSAGE NON SUPERVISÉ

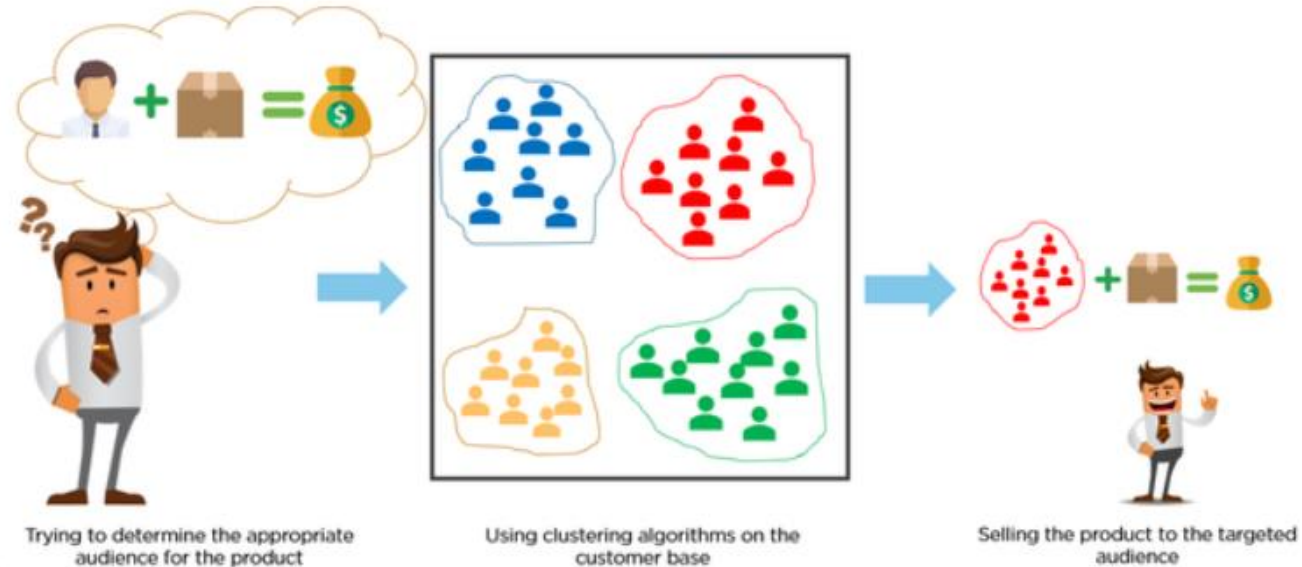
CLUSTERING – SEGMENTATION



APPRENTISSAGE NON SUPERVISÉ

CLUSTERING – APPLICATIONS

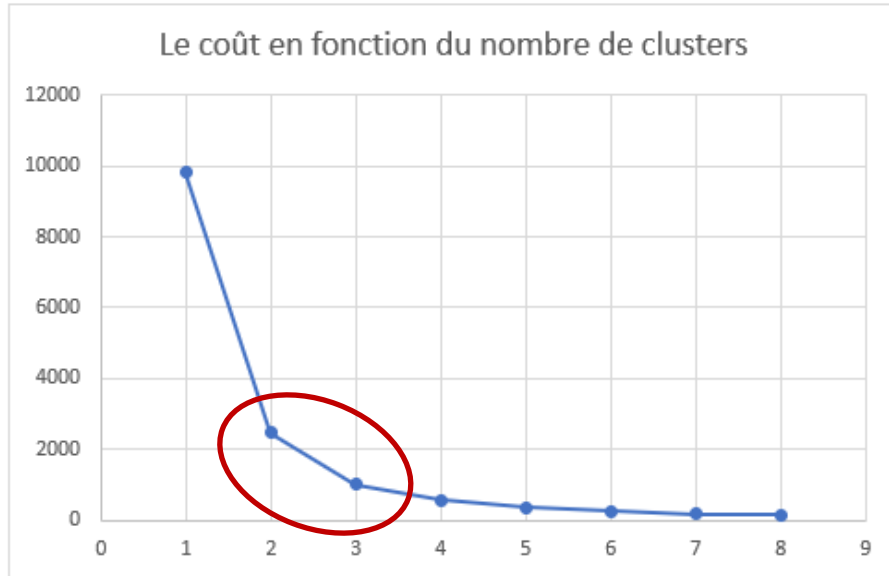
- Segmentation de la clientèle



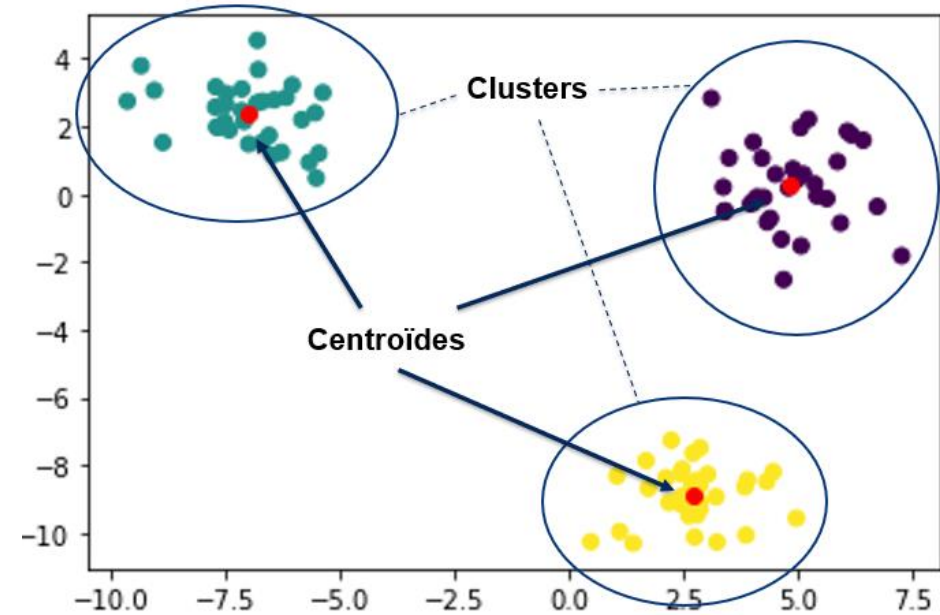
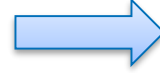
- Déceler des individus similaires
- Segmentation de documents

APPRENTISSAGE NON SUPERVISÉ

CLUSTERING – K MEANS



méthode Elbow

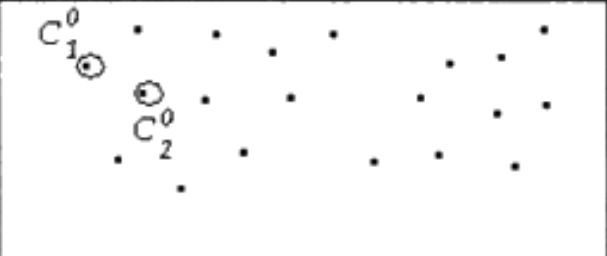
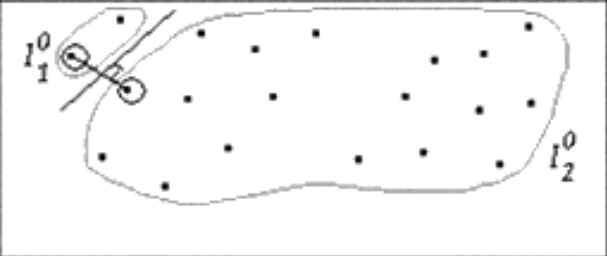
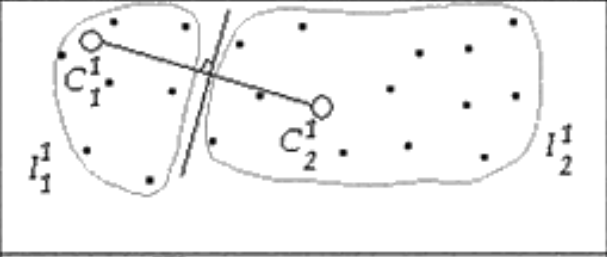



KMeans

- **Fonction coût :** $V = \sum_j \sum_{x_i \rightarrow c_j} D(c_j, x_i)^2$
 - c_j : Le centre du cluster (le centroïd)
 - x_i : la i ème observation dans le cluster ayant pour centroïd c_j
 - $D(c_j, x_i)$: La distance (euclidienne ou autre) entre le centre du cluster et le point x_i

APPRENTISSAGE NON SUPERVISÉ

CLUSTERING – K MEANS – PRINCIPE

	Tirage au hasard des centres C_1^0 et C_2^0
	Constitution des classes I_1^0 et I_2^0
	Nouveaux centres C_1^1 et C_2^1 et nouvelles classes I_1^1 et I_2^1
	Nouveaux centres C_1^2 et C_2^2 et nouvelles classes I_1^2 et I_2^2

APPRENTISSAGE NON SUPERVISÉ

RÉDUCTION DES DIMENSIONS - EXEMPLE

	Math	Physique	Franc	Anglais
Jean	6	6	5	5,5
Alain	8	8	8	8
Annie	6	7	11	9,5
Monique	14,5	14,5	15,5	15
Didier	14	14	12	12,5
André	11	10	5,5	7
Pierre	5,5	7	14	11,5
Brice	13	12,5	8,5	9,5
Evelyne	9	9,5	12,5	12

Communalities

	Initial	Extraction
Math	1,000	,999
Physique	1,000	,999
Franc	1,000	,999
Anglais	1,000	,998

Extraction Method: Principal Component Analysis.

Component Matrix^a

	Component	
	1	2
Math	,811	-,584
Physique	,902	-,430
Franc	,753	,657
Anglais	,915	,401

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Total Variance Explained

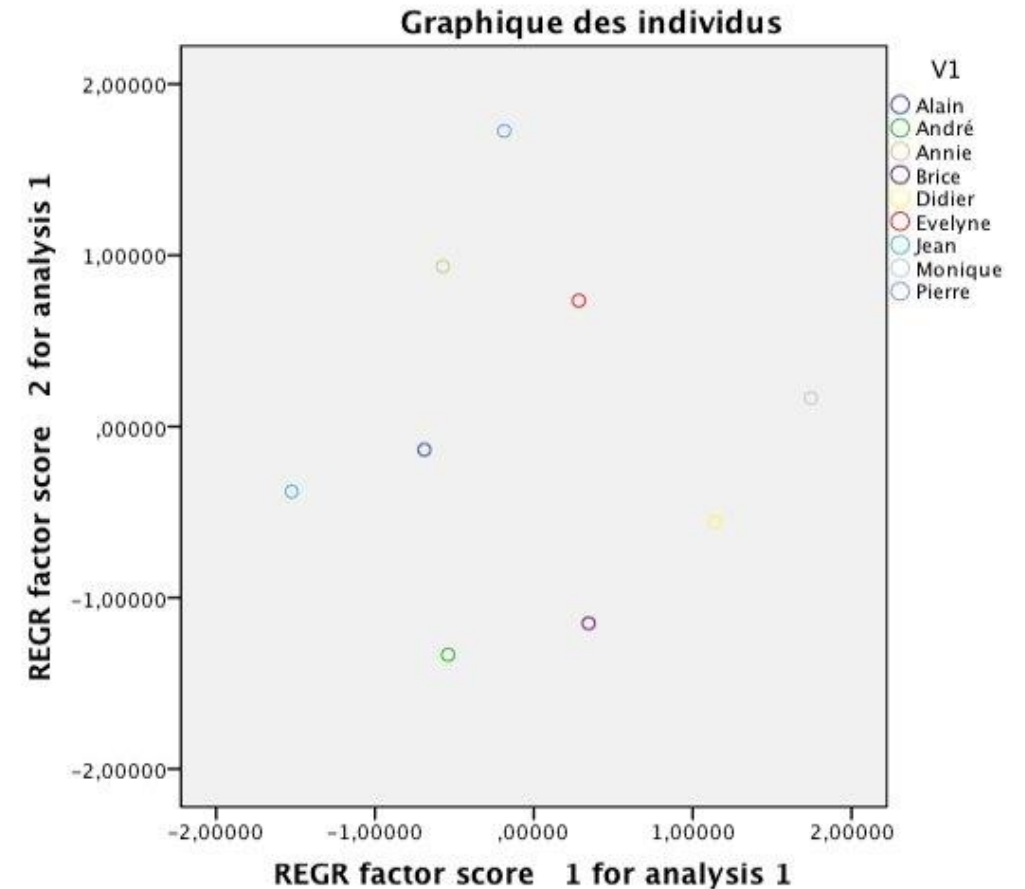
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,876	71,892	71,892	2,876	71,892	71,892
2	1,120	27,992	99,884	1,120	27,992	99,884
3	,004	,089	99,974			
4	,001	,026	100,000			

Extraction Method: Principal Component Analysis.

APPRENTISSAGE NON SUPERVISÉ

RÉDUCTION DES DIMENSIONS - EXEMPLE

	Math	Physique	Franc	Anglais
Jean	6	6	5	5,5
Alain	8	8	8	8
Annie	6	7	11	9,5
Monique	14,5	14,5	15,5	15
Didier	14	14	12	12,5
André	11	10	5,5	7
Pierre	5,5	7	14	11,5
Brice	13	12,5	8,5	9,5
Evelyne	9	9,5	12,5	12



APPRENTISSAGE NON SUPERVISÉ

APPRENTISSAGE PAR RENFORCEMENT

