المدرسة الوطنية للعلوم التطبيقية - بني ملال

ⵜⴰⵙⴷⴰⵡⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵙⵏⴰⵡⵉⵏ ⵜⵜⵓⵏⵖⵉⵏ - ⴰⵣⵔⴰⵡ

Ecole Nationale des Sciences Appliquées - Béni Mellal

# ANALYSE DES DONNÉES



Filière : Transformation Digitale Industrielle (TDI)

Pr. ESSWIDI AYOUB

A.U. 2024-2025



# PLAN PRÉVISIONNEL ET OBJECTIFS

- Analyse de données
  - Prétraitement et nettoyage des données
  - Visualisation des données (EDA).
  - Feature engineering, …
  - Analyse des données multimédia, Analyse du Web et des réseaux sociaux (exposé)
- Science de données et Data Mining
  - Introduction et types d'apprentissage
  - Algorithmes de ML
  - Algorithmes de Deep Learning
  - Exemple d'application en IA générative (exposés)
- Big Data
  - Introduction au Big Data: définition, caractéristiques, Historique, domaine d'application.
  - Architectures du Big Data
  - Ecosystèmes : Hadoop et Map-Reduce, Apache spark, kafka, HBASE, …

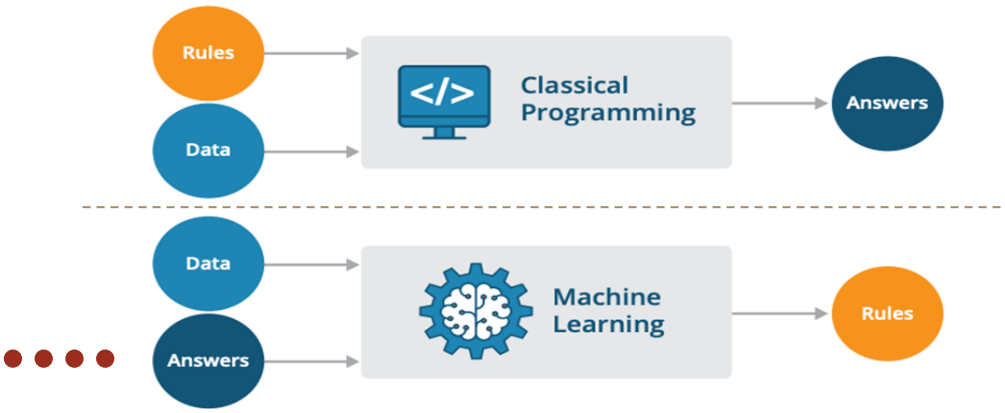
- Introduction
- Régression
  - Régression linéaire
- Classification
  - Régression logistique (pour la classification binaire)
  - KNN
  - Arbre de decision / random forest
  - Classification naïve bayesienne
  - SVM
- Clustering
  - K-means
- Deep Learning
  - ANN

3

3

# INTRODUCTION

- **Machine Learning (apprentissage automatique)** consiste à donner à une machine la capacité d'apprendre plutôt que de la programmer de façon explicite.




4

# DOMAINES D'APPLICATION

| Domaine | Ex. d'application | Objectif |
|---|---|---|
| Bancaire | Approbation ou non du crédit | modèle intelligent capables d'exploiter l'historique des données des clients passés, afin d'automatiser le processus de prise de décision et améliorer sa précision. |
| médical | Diagnostique médical (par exemple tumeur de cerveau) | Automatisation du processus de détection des tumeurs sans intervention humaine et augmenter la précision du système en utilisant (les images d'IRM) |
| Les voitures autonomes | Détecter des piétons pour les véhicules autonomes. | Détection et visualisation des piétons dans toutes les directions et dans n'importe quel environnement (nuit, brouillard, jour très ensoleillé…) |
| Chabots | ChatGPT | |
| … | | |

5

# CONDITIONS D'UTILISATION

**L'utilisation de ML exige la réalisation de trois conditions:**

- Existence des données: (condition suffisante +|-)

Il existe des données qui représentent le modèle.

- Existence d'un modèle à apprendre:

Il existe une corrélation entre les variables d'entrées et de sorties. On sait qu'un modèle existe même si on ne le connait pas.

- Modélisation mathématique est impossible:

On ne peut pas résoudre le modèle mathématiquement (pas de solution analytique).

**NB:** possible de construire des modèles basés sur des données **tabulaire**, **Texte**, **Image & vidéo**, **signale(audio)**

6

# TYPES D'ALGORITHMES DE ML

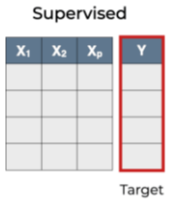
- Apprentissage supervisé (supervised learning)
- Apprentissage Non supervisé (unsupervised learning)
- Semi supervisé (semi-supervised learning)
- Par renforcement (reinforcement)

7

# TYPES D'ALGORITHMES DE ML

**Apprentissage supervisé (supervised learning)**

- Régression linéaire
- Algorithme des K-plus proches voisins (K-NN)
- Régression logistique
- Arbres de décision
- Forêts aléatoires (Random Forest)
- Machines à vecteurs de support (SVM)
- Réseaux neuronaux (appliqués en mode supervisé)
- …

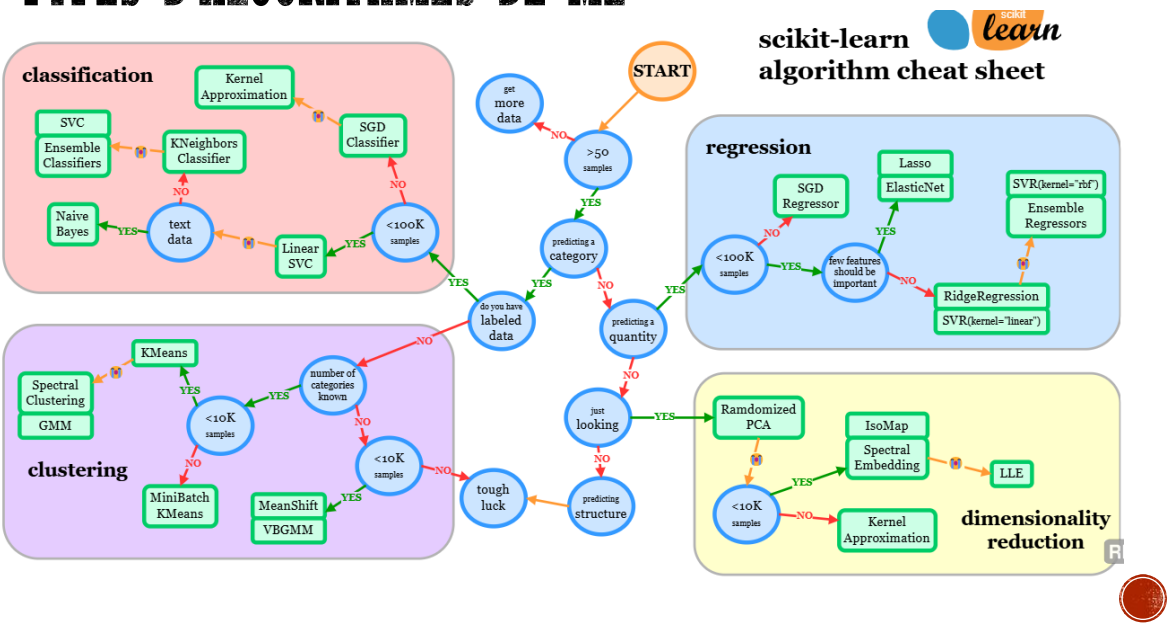



**Apprentissage Non supervisé (unsupervised learning)**

- K-means (clustering)
- Analyse en composantes principales (PCA)
- Réseaux auto-encodeurs (Autoencoders)
- …


8

# TYPES D'ALGORITHMES DE ML




9

# ALGORITHMES DE ML SUPERVISÉ

- Des données (entrées) annotées de leurs sorties pour entraîner le modèle,
- c'est-à-dire que à chaque entrée est associée à une classe cible (resp. valeur continue),une fois entraîné, le modèle (l'algorithme de ML) devient capable de classifier (resp. prédire) (éventuellement avec un pourcentage d'erreur) la cible sur de nouvelles données non annotées.
- On distingue deux types de problèmes

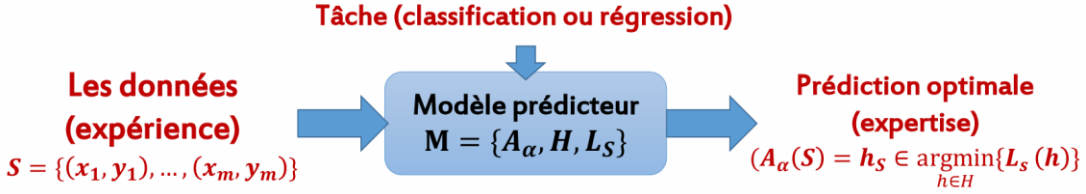
| Les régressions | Les classifications |
|---|---|
| Lorsqu'on cherche à prédire une quantité : | Lorsqu'on cherche à prédire une classe : |
|  |  |
| Exemples : prix d'une maison, durée d'un trajet, température, poids, ... | Exemples : un animal, un état de santé, une couleur, un type de véhicule, ... |



10

# ALGORITHMES DE ML SUPERVISÉ

❑ Le ML c'est un processus qui utilise l'expérience pour acquérir une expertise



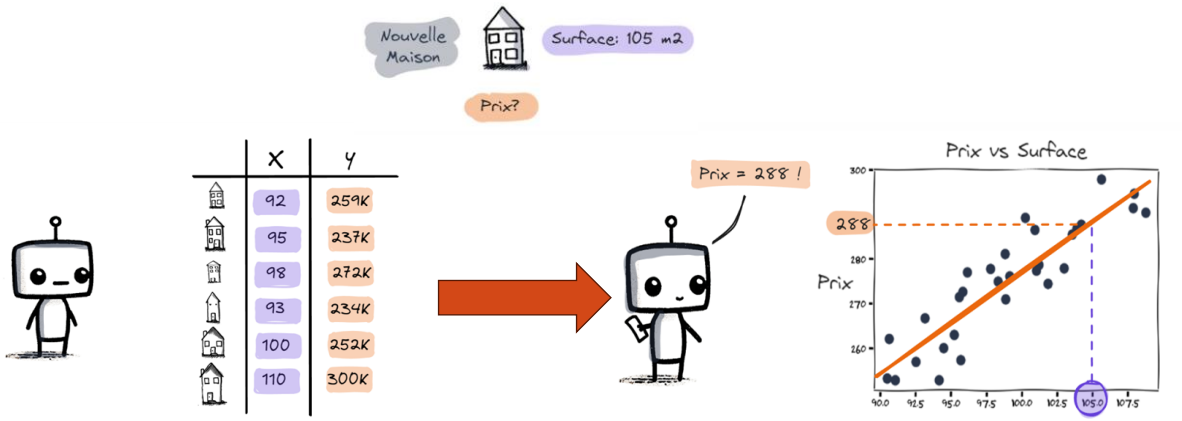

$x_i = (x_i^j, j = 1, \dots, d), x_i^j$ : caractéristique

$A_\alpha$ : l'algorithme d'apprentissage , $H$: l'ensemble des hypothèses ,
$L_S$: fonction d'erreur empirique et $\alpha$ est un vecteur de paramètres.
Il concerne l'utilisation des bonnes caractéristiques $x_i^j$, pour construire le bon modèle $h_S$ en minimisant $L_S$ permettant de réaliser les bonnes tâches.

11

# ML SUPERVISÉ (RÉGRESSION)

❑ Imaginez que vous souhaitiez prédire le prix d'une maison en fonction de sa surface habitable.




12

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

❑ Sélectionner **un estimateur** et préciser **ses hyperparamètres**:

Model = LinearRegression(.....)

objet — Constructeur — Hyperparamètres

```
# Entrainer le modèle sur les données X, y
model.fit( X,y )

# Évaluer le modèle
model.score (X,y)

# Utiliser le modèle
model.predict(X_test)
```

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X, y)
model.score(X,y)
model.predict(X)
```

```
from sklearn.naive_bayes import GaussianNB
model = GaussianNB()
model.fit(X, y)
model.score(X,y)
model.predict(X)
```

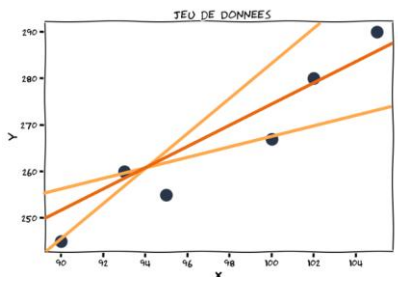
```
from sklearn.svm import SVR
model = SVR()
model.fit(X, y)
model.score(X,y)
model.predict(X)
```

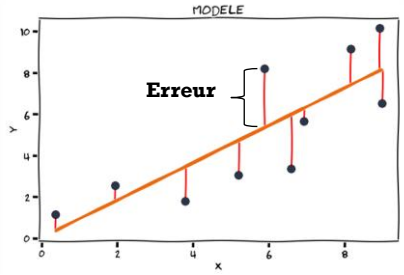
13

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

❑ Le modèle de régression linéaire peut être décrit avec la formule bien connue

$$f(x) = WX + b.$$




14

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

❑ Les fonctions d'Erreurs (Loss):

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y^i - f(x^i) |$$

- Mean Squared Error (MSE):

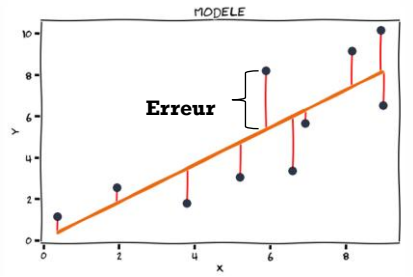
$$MSE = Loss = \frac{1}{m}\sum_{i=1}^{m} \left(y^i - f(x^i)\right)^2$$

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}$$

- R-squared ($R^2$):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



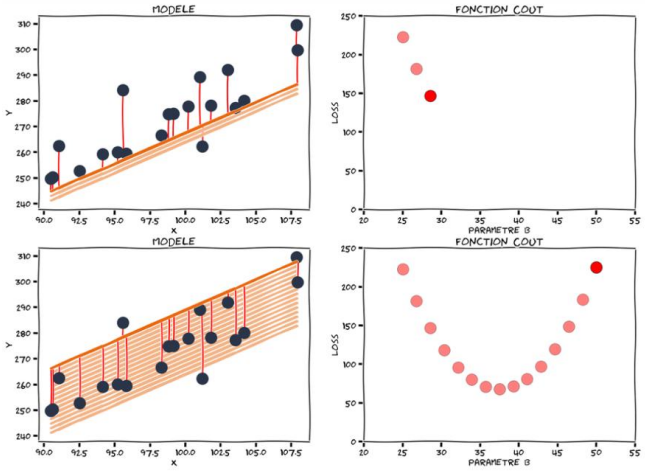

*Où*

- $y_i = actual\ values$
- $\hat{y}_i = f(x_i)$: *predicted* **values**
- $\bar{y}$: *the mean of* $y_i$

15

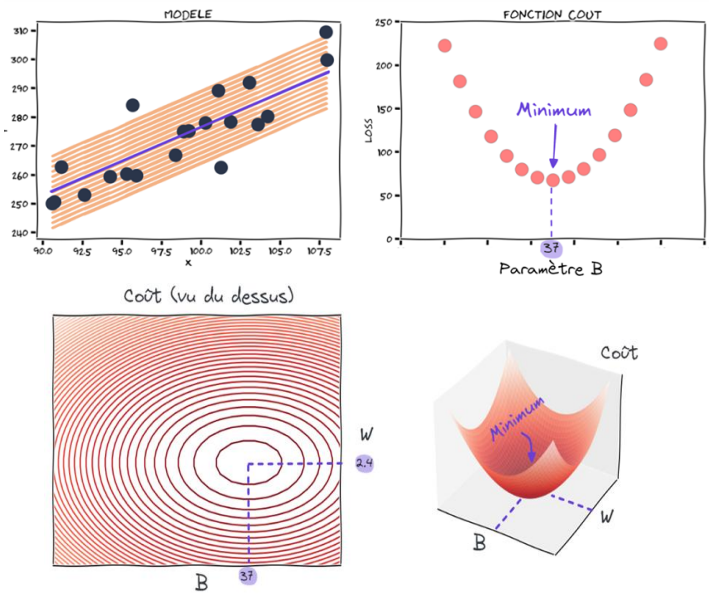
# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

❑ Allure de la fonction coût: $Loss = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - f(x^i)\right)^2$




16

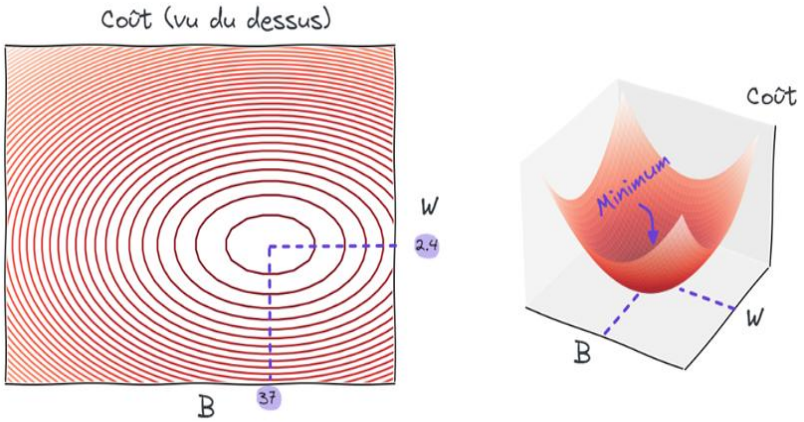
# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)




17

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

- Allure de la fonction coût: $L = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - f(x^i)\right)^2$




18

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)

**L'algorithme de la descente de gradient:**

- est l'un des algorithmes d'apprentissage les plus utilisés en Machine lLearning.
- Il consiste à calculer le gradient de la fonction coût,
- c'est-à-dire comment celle-ci évolue lorsque w et b varient légèrement, pour ensuite faire un pas dans la direction où la fonction coût diminue. D'où le nom de **descente de gradient**.

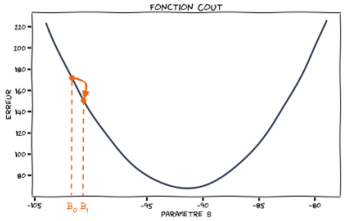
$$w_{t+1} = w_t - \eta \times \frac{\partial L}{\partial w_t}$$

$$b_{t+1} = b_t - \eta \times \frac{\partial L}{\partial b_t}$$

$\eta$ représente une vitesse d'apprentissage

19

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE SIMPLE)



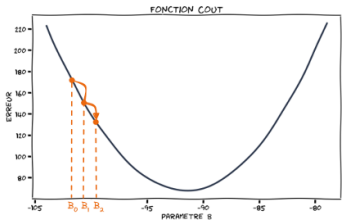

Ainsi, la formule de la descente de gradient donne : $supposons\ que\ (\eta = 1)$

$$b_1 = b_0 - \eta \times \frac{\partial L}{\partial b_0}$$

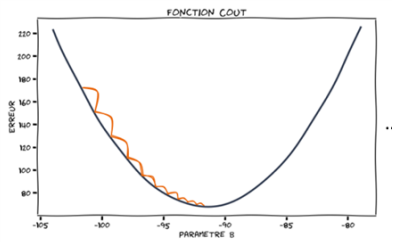
$$b_1 = b_0 - 1 \times (-2)$$

$$b_1 = b_0 + 2$$

$b_1$ est donc supérieur à $b_0$




On répète ainsi l'algorithme...




20

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE MULTIPLE)

**Régression linéaire simple**

Constante — Coefficient

$$y = b_0 + b_1 * x_1$$

La variable indépendante

La variable dépendante

**Régression linéaire multiple**

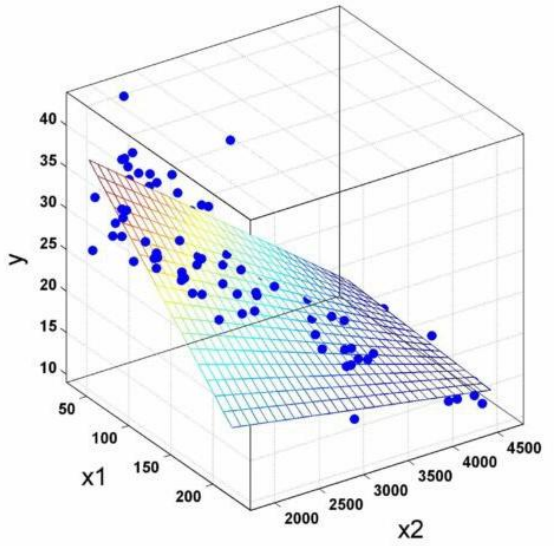
Les variables indépendantes

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Constante — Coefficients

21

# ML SUPERVISÉ (RÉGRESSION LINÉAIRE MULTIPLE)




22

```python
# Importation des bibliothèques nécessaires
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Chargement des données
data = pd.read_csv('house_price.csv')

# Sélection des caractéristiques (features) et de la variable cible (target)
# et que la cible (le prix de la maison) est dans la colonne 'price'
X = data[['feature1', 'feature2', 'feature3']]  # Remplacez par les noms de vos colonnes
y = data['price']

# Division des données en ensembles d'entraînement et de test (80% entraînement, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Création et entraînement du modèle de régression linéaire
model = LinearRegression()
model.fit(X_train, y_train)

# Prédiction des prix sur l'ensemble de test
y_pred = model.predict(X_test)

# Calcul des métriques d'évaluation
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
```

23

# ML SUPERVISÉ

**Autres algorithmes de Régressions:**

- Ridge regression
- Bayesian Regression
- Stochastic Gradient Descent – SGD
- Lasso
- Support Vector Machines (SVM)
- Perceptron
- Multi-layer Perceptron
- ….

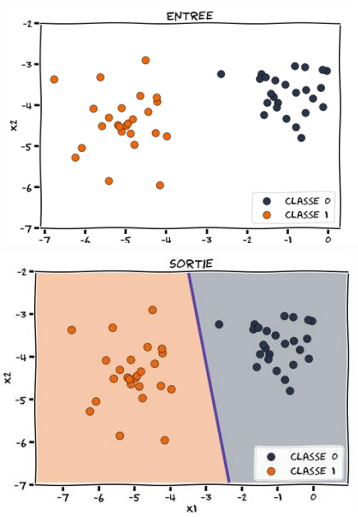
- Supervised learning — scikit-learn 1.5.2 documentation

24

# ML SUPERVISÉ (CLASSIFICATION)

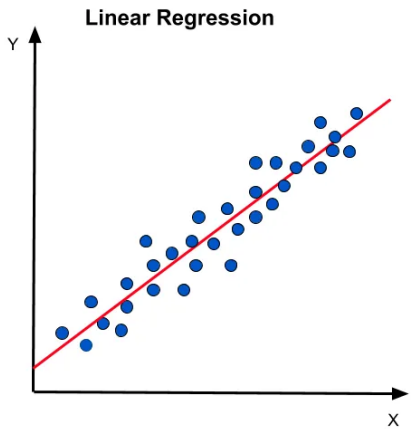
- Les modèles de classification sont utilisés pour prédire **la classe** ou **la catégorie** à laquelle appartient un objet

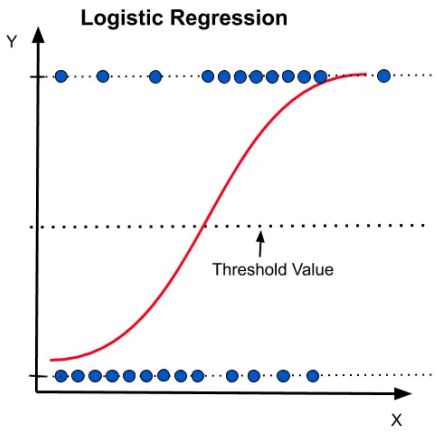
  - ✓ Prédire si un email est un spam ou non (2 classes possibles)
  - ✓ Détecter si une cellule est cancéreuse ou non (2 classes possibles)
  - ✓ Prédire le style de musique préféré d'une personne (environ 10 classes possibles)
  - ✓ Reconnaître une plante et donner son nom (des milliers de classes possibles)
  - ✓ …etc.




25

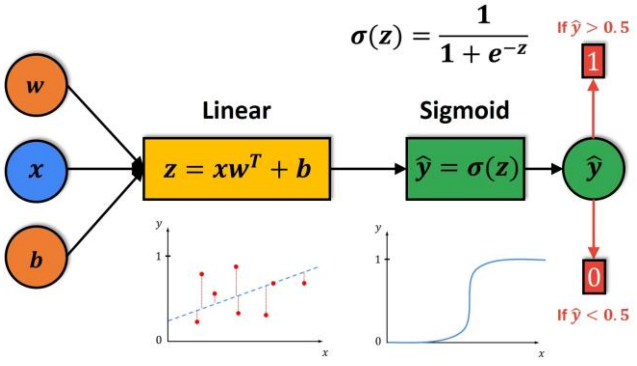
# ML SUPERVISÉ (RÉGRESSION LOGISTIQUE)




26

# ML SUPERVISÉ (RÉGRESSION LOGISTIQUE)




$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

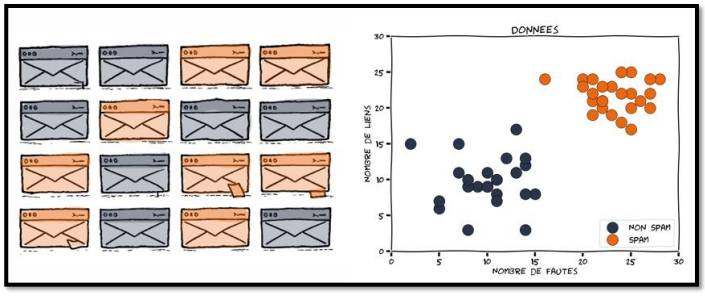
$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

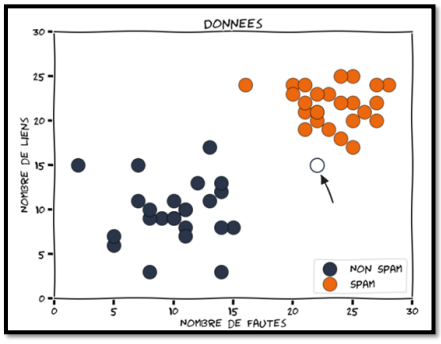
27

# ML SUPERVISÉ (CLASSIFICATION)

- **KNN k-nearest neighbors (k-plus proches voisins)**

- Prédire si un email est un spam ou non (2 classes possibles)




28