Rappels des définitions

Titre du cours : Analyse exploratoire des données

Code officiel : 420-A55-BB

Professeur : Dr Komi SODOKE

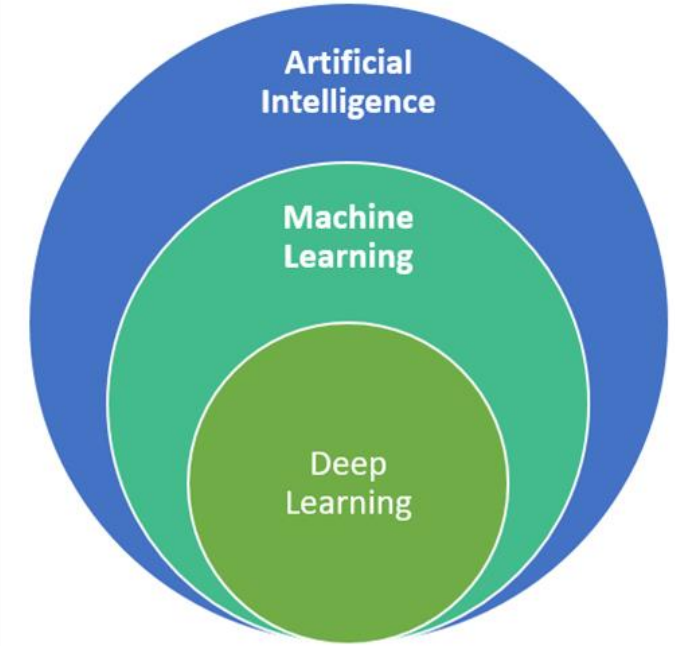
Plan

I- Introduction

Le Machine Learning (ML) est un domaine en pleine expansion qui révolutionne divers secteurs, allant de la santé à la finance, en passant par le marketing et l'industrie. Il repose sur des techniques permettant aux ordinateurs d'apprendre à partir des données et d'effectuer des tâches sans être explicitement programmés. L'objectif de ce cours est d'introduire les concepts fondamentaux du Machine Learning, d'explorer l'analyse exploratoire des données (EDA) et d'examiner les approches statistiques utilisées pour construire et évaluer des modèles prédictifs.

Qu’est-ce que le Machine Learning ?

Le Machine Learning (ML) est une branche de l’intelligence artificielle (IA) qui vise à permettre aux ordinateurs d’apprendre à partir des données et à prendre des décisions sans être explicitement programmés. Il repose sur des algorithmes capables d’identifier des modèles, de faire des prédictions et d’améliorer leurs performances avec l’expérience.



Exemples d'Applications du Machine Learning

- Reconnaissance d'images : Détection d'objets, reconnaissance faciale (exemple : déverrouillage des smartphones).
- Systèmes de recommandation : Netflix, Amazon, YouTube utilisent le ML pour proposer du contenu personnalisé.
- Détection de fraudes : Analyse des transactions bancaires pour repérer des comportements suspects.
- Voitures autonomes : Analyse en temps réel des conditions de conduite pour la navigation autonome.
- Santé : Diagnostic de maladies à partir d'images médicales et personnalisation des traitements.
- ...

II- Approches d'Apprentissage Machine (IA)

A. Apprentissage Supervisé

L'apprentissage supervisé en machine learning consiste à enseigner à un ordinateur à reconnaître des schémas dans les données en lui fournissant des exemples étiquetés. Ces exemples sont des paires de données, où chaque entrée est associée à une sortie attendue.

L'ordinateur utilise ces exemples pour apprendre à faire des prédictions précises sur de nouvelles données. En d'autres termes, il apprend à partir de "bons exemples" où la bonne réponse est déjà connue, afin de pouvoir généraliser et faire des prédictions sur des données non vues auparavant.

Exemples des algorithmes :

- Régression linéaire et logistique : Utilisées pour les prédictions numériques et les classifications binaires.
- Arbres de décision et forêts aléatoires : Construisent des modèles robustes en divisant les données en sous-groupes successifs.
- Support Vector Machines (SVM) : Trouve l'hyperplan optimal séparant différentes classes.
- Réseaux de neurones artificiels : Inspirés du cerveau humain, très performants pour la vision et le traitement du langage naturel.

Cas d'usage :

- Prédiction du prix de l'immobilier (régression linéaire).
- Détection de spams dans les e-mails (classification binaire).
- Reconnaissance faciale sur les smartphones (réseaux de neurones).

B. Apprentissage non supervisé

L'apprentissage non-supervisé en machine learning consiste à analyser des données sans étiquettes ni réponses préétablies. Contrairement à l'apprentissage supervisé, où le modèle est entraîné sur des exemples étiquetés pour prédire des résultats, l'apprentissage non-supervisé cherche à découvrir des structures ou des modèles intrinsèques dans les données elles-mêmes.

Cela peut inclure des tâches telles que :

- le regroupement de données similaires en clusters,
- la réduction de la dimensionnalité pour simplifier la représentation des données
- la détection d'anomalies pour identifier des comportements inhabituels,
- la découverte de motifs fréquents ou d'associations entre les variables

B. Apprentissage non supervisé

- **Exemples d'algorithmes :**
  - K-Means : Regroupe les données en clusters selon leur similarité.
  - Analyse en Composantes Principales (PCA) : Réduit la dimensionnalité des données tout en conservant l'essentiel des informations.
  - Clustering hiérarchique : Créer une arborescence des relations entre les observations.
- **Cas d'usage :**
  - Segmentation de clients dans le marketing.
  - Détection de fraudes bancaires en trouvant des transactions anormales.
  - Regroupement d'articles d'actualité similaires.

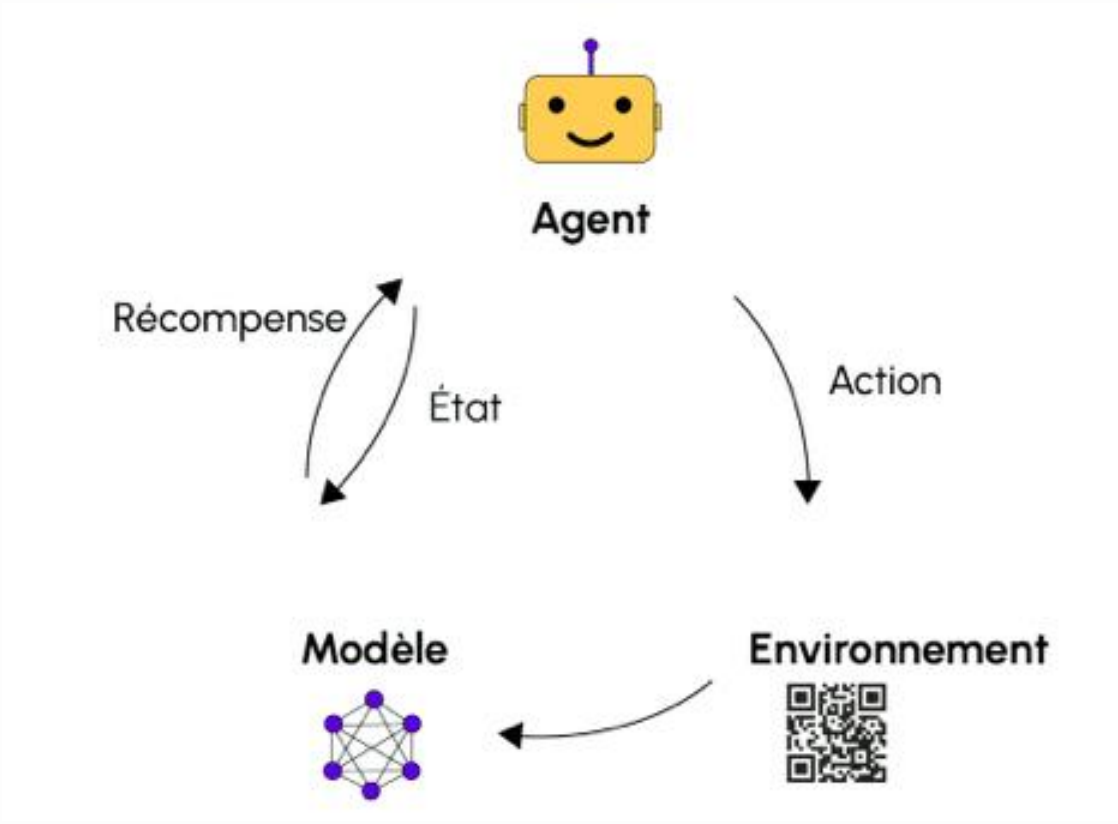
C. Apprentissage par Renforcement

L’apprentissage par renforcement désigne l’ensemble des méthodes qui permettent à un agent d’apprendre à choisir quelle action prendre, et ceci de manière autonome.

Plongé dans un environnement donné, il apprend en recevant des récompenses ou des pénalités en fonction de ses actions. Au travers de son expérience, l’agent cherche à trouver la stratégie décisionnelle optimale qui puisse lui permettre de maximiser les récompenses accumulées au cours du temps.

Concepts clés :

- **Agent :** L'entité qui prend les décisions.
- **Environnement :** L’univers dans lequel évolue l’agent.
- **Politique :** Stratégie de prise de décision de l’agent.
- **Récompense :** Retour mesurant la qualité d’une action.


Agent
Récompense
État
Action
Modèle
Environnement

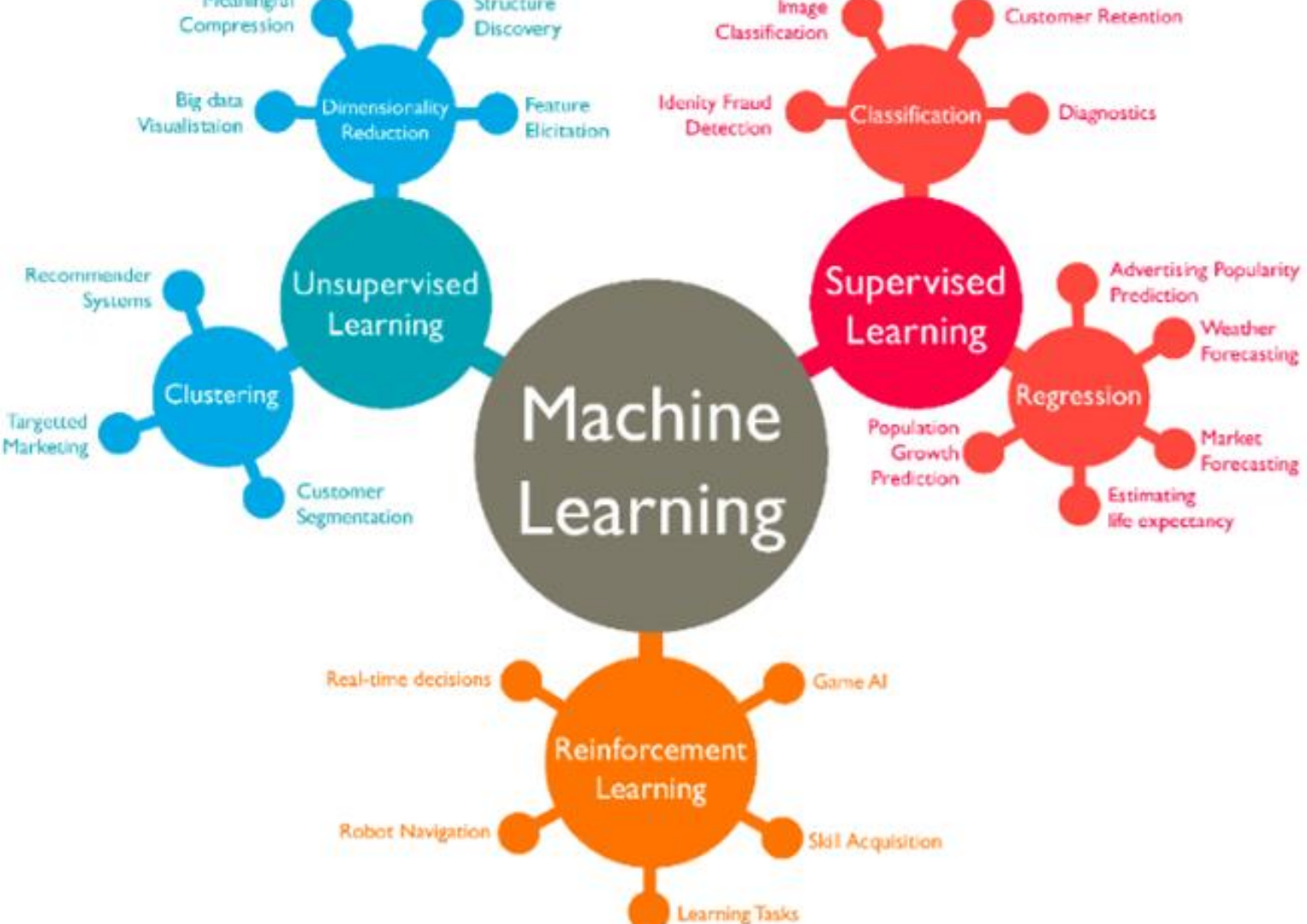
Exemples d'algorithmes :

- Q-Learning : Apprend une politique optimisée par essais et erreurs.

- Deep Q-Networks (DQN) : Utilise les réseaux de neurones pour améliorer l’apprentissage par renforcement.

Cas d’usage :

- Automatisation des jeux vidéo (exemple : AlphaGo battant les champions d’échecs).
- Contrôle de robots autonomes.
- Optimisation des recommandations publicitaires.


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualisation
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

III- Analyse exploratoire des données

- **Définition :**
  L'Analyse Exploratoire des Données (EDA) est une étape essentielle du Machine Learning permettant de comprendre la structure des données **avant** de construire un modèle. Elle aide à détecter les valeurs aberrantes, à identifier les tendances et à choisir les bonnes transformations des données.
- **Importance :**
  - Comprendre la distribution des données
  - Détecter les valeurs aberrantes
  - Choisir les transformations nécessaires pour améliorer les modèles ML

A. Techniques d’ analyse exploratoire des données (EDA)

1. **Résumé statistique des données** : Moyenne, médiane, écart-type, quartiles.
2. **Visualisations** :
    - Histogrammes pour comprendre la distribution des variables.
    - Diagrammes en boîte (boxplots) pour détecter les valeurs aberrantes.
    - Matrices de corrélation pour identifier les relations entre variables.
3. **Détection et gestion des valeurs manquantes** : Suppression, imputation avec la médiane ou interpolation.
4. **Transformation des variables** : Normalisation, standardisation, encodage des variables catégorielles.
5. **Analyse des distributions** : Vérification de la normalité des données à l’aide de tests statistiques (ex : test de Shapiro-Wilk).
6. **Détection des valeurs extrêmes** : Utilisation de méthodes comme l’IQR (Interquartile Range) pour identifier les outliers.

Exemples Concrets

- **Finance** : Histogrammes de la distribution des revenus des clients pour créer un modèle de scoring de crédit.
- **Santé** : Boxplots comparant les âges des patients en fonction des diagnostics pour détecter des schémas cachés.
- **Marketing** : Matrices de corrélation pour observer les liens entre habitudes d'achat et montants dépensés.

L'EDA est une étape cruciale qui garantit la qualité des données avant toute modélisation en Machine Learning, réduisant ainsi le risque d'erreurs et améliorant la performance des modèles.

1. Résumé statistique des données

Le **résumé statistique** donne une vue d'ensemble des variables numériques.

Mesure	Définition	Utilité
Moyenne (Mean)	Somme des valeurs / nombre d'observations	Donne une idée du centre des données, mais sensible aux valeurs extrêmes.
Médiane (Median)	Valeur centrale des données triées	Moins affectée par les outliers que la moyenne.
Écart-type (Standard deviation, σ)	Dispersion des valeurs autour de la moyenne	Indique si les données sont concentrées ou dispersées.
Quartiles (Q1, Q2, Q3)	Valeurs qui divisent les données en 4 parties égales	Utilisés pour détecter les valeurs aberrantes avec la méthode IQR.

Pourquoi c’est important ?

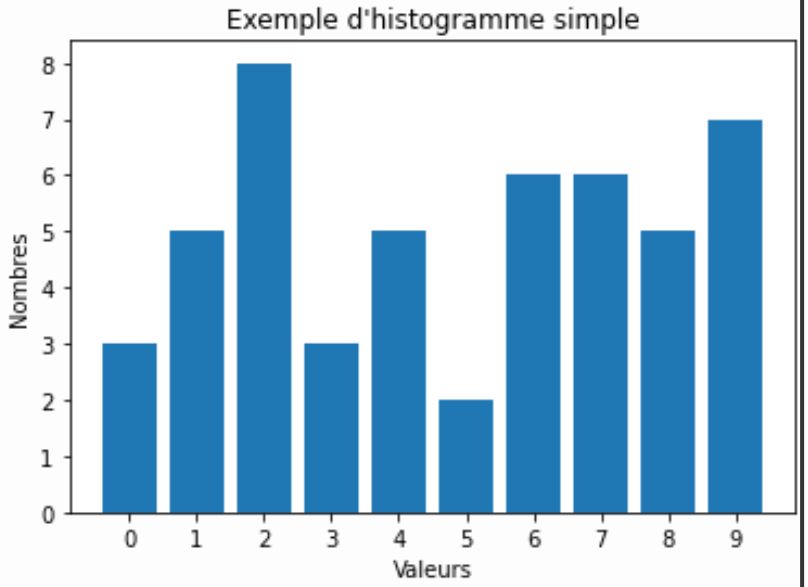
- **La moyenne** et la **médiane** permettent de comprendre la tendance centrale.
- **L’écart-type** et les quartiles montrent si les valeurs sont concentrées ou dispersées.
- Une variable avec une forte dispersion peut nécessiter une transformation avant d’être utilisée dans un modèle.

2. Visualisations

Les graphiques sont essentiels pour repérer les patterns cachés que les statistiques seules ne montrent pas.

a- Histogrammes

- Un histogramme montre la **distribution des valeurs** d'une variable numérique.
- Permet de voir si une variable suit une **distribution normale** ou **asymétrique.**
- Aide à détecter les valeurs inhabituelles et les écarts.

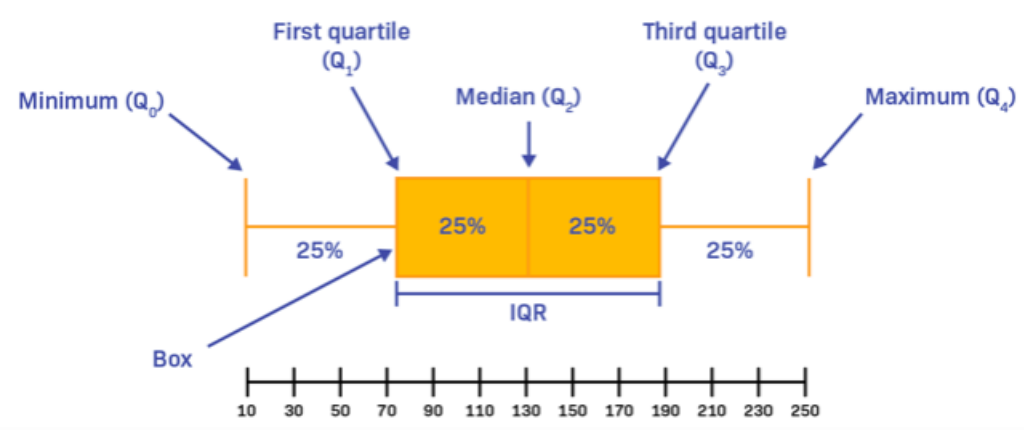


Exemple : Un histogramme des salaires peut montrer si la majorité des employés gagnent un revenu moyen ou s'il y a des écarts importants.

b. Boxplots (diagrammes en boîte)

Un boxplot permet de:

- Visualiser les **quartiles** et d'identifier les valeurs aberrantes.
- Affiche la médiane, Q1, Q3 et les **valeurs extrêmes.**
- Aide à comparer plusieurs distributions (ex : salaires selon le secteur d'activité).



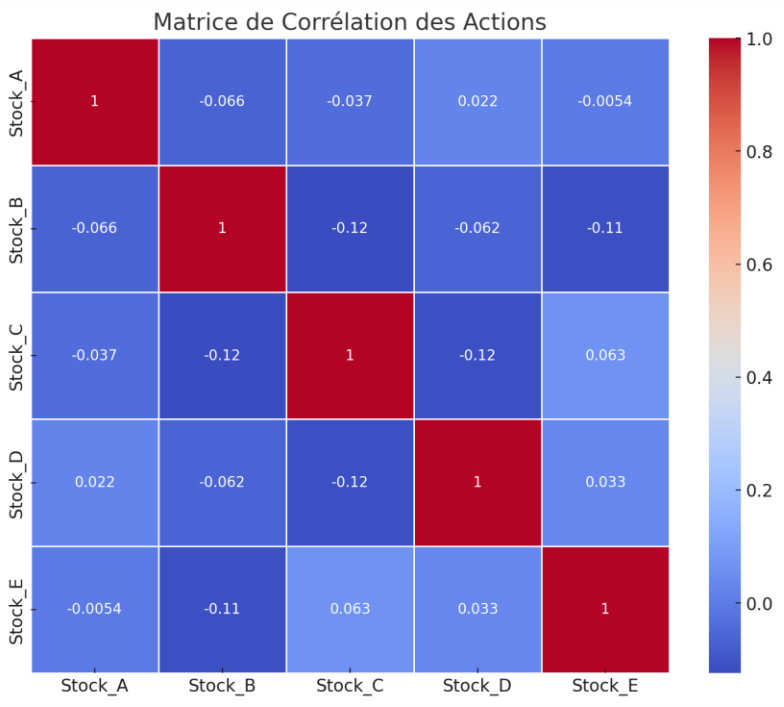
Exemple : Un boxplot des notes d'examen permet de voir si certains étudiants ont des résultats extrêmement bas ou élevés.

c- Matrice de corrélation

Une matrice de corrélation, souvent représentée sous forme de **heatmap**, permet d'analyser les relations entre différentes variables.

- Plus le coefficient de corrélation est proche de 1 ou -1, plus la relation entre les variables est forte.
- Un coefficient proche de 0 signifie qu'il n'y a pas de lien significatif entre elles.

Exemple : Si la température est fortement corrélée aux ventes de glaces, on peut exploiter cette relation pour construire un modèle de prévision des ventes.



3- Détection et gestion des valeurs manquantes

Les **valeurs manquantes** peuvent **fausser l'analyse, réduire la qualité des modèles** et **entraîner des erreurs** si elles ne sont pas bien gérées.

❖ **Pourquoi sont-elles problématiques ?**

- **Biais des résultats :** Si les valeurs manquent de façon non aléatoire, l'analyse peut être faussée.
- **Perte d'information :** Supprimer ces données peut nuire à la précision du modèle.
- **Incompatibilité avec certains algorithmes :** Certains modèles ne fonctionnent pas avec des données manquantes.

Exemple : Dans un dataset médical, si la glycémie est manquante surtout chez les patients au mode de vie sain, supprimer ces données pourrait biaiser l'étude en sous-estimant la gravité du diabète.

Stratégies pour gérer les valeurs manquantes

Méthode	Description	Quand l'utiliser ?
Suppression des lignes	Enlever les lignes contenant des valeurs manquantes	Si le nombre de valeurs manquantes est très faible (ex : < 5 %).
Imputation avec la médiane/moyenne	Remplacer les valeurs manquantes par la médiane ou la moyenne	Si les valeurs manquantes sont aléatoires et peu nombreuses.
Imputation par interpolation	Estimer les valeurs manquantes à partir des données existantes (ex : interpolation linéaire)	Pour des séries temporelles ou des données continues.

Exemple : Dans un dataset médical, si 10 % des patients n'ont pas renseigné leur âge, on peut **remplacer par la médiane plutôt** que de supprimer ces patients.

4. Transformation des variables

Certaines transformations sont nécessaires pour que les modèles fonctionnent mieux.

❖ Normalisation et Standardisation

- **Normalisation (Min-Max Scaling) :** Met les valeurs entre 0 et 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Utilisée pour **KNN, réseaux de neurones**.

- **Standardisation (Z-score normalization) :** Met les valeurs sur une distribution de moyenne **0** et écart-type **1**.

$$X' = \frac{X - \mu}{\sigma}$$

Utilisée pour **régressions linéaires, SVM, PCA.**

❖ Encodage des variables catégorielles

Les modèles de machine learning ne peuvent pas traiter directement les **variables textuelles** (ex : "Femme", "Homme").
- **Encodage One-Hot :** Convertit chaque catégorie en une colonne binaire (1 ou 0).
- **Encodage Ordinal :** Attribue un numéro à chaque catégorie si elles ont un ordre (ex : "Faible"=1, "Moyen"=2, "Élevé"=3).

Exemple : Une variable "Type de voiture" ("SUV", "Berline", "Compacte") peut être encodée avec One-Hot comme ceci :

SUV	Berline	Compacte
1	0	0
0	1	0
0	0	1

5. Analyse des distributions et détection des valeurs extrêmes

❖ **Détection des valeurs extrêmes (Outliers)**

Les valeurs aberrantes peuvent fausser l'entraînement des modèles.

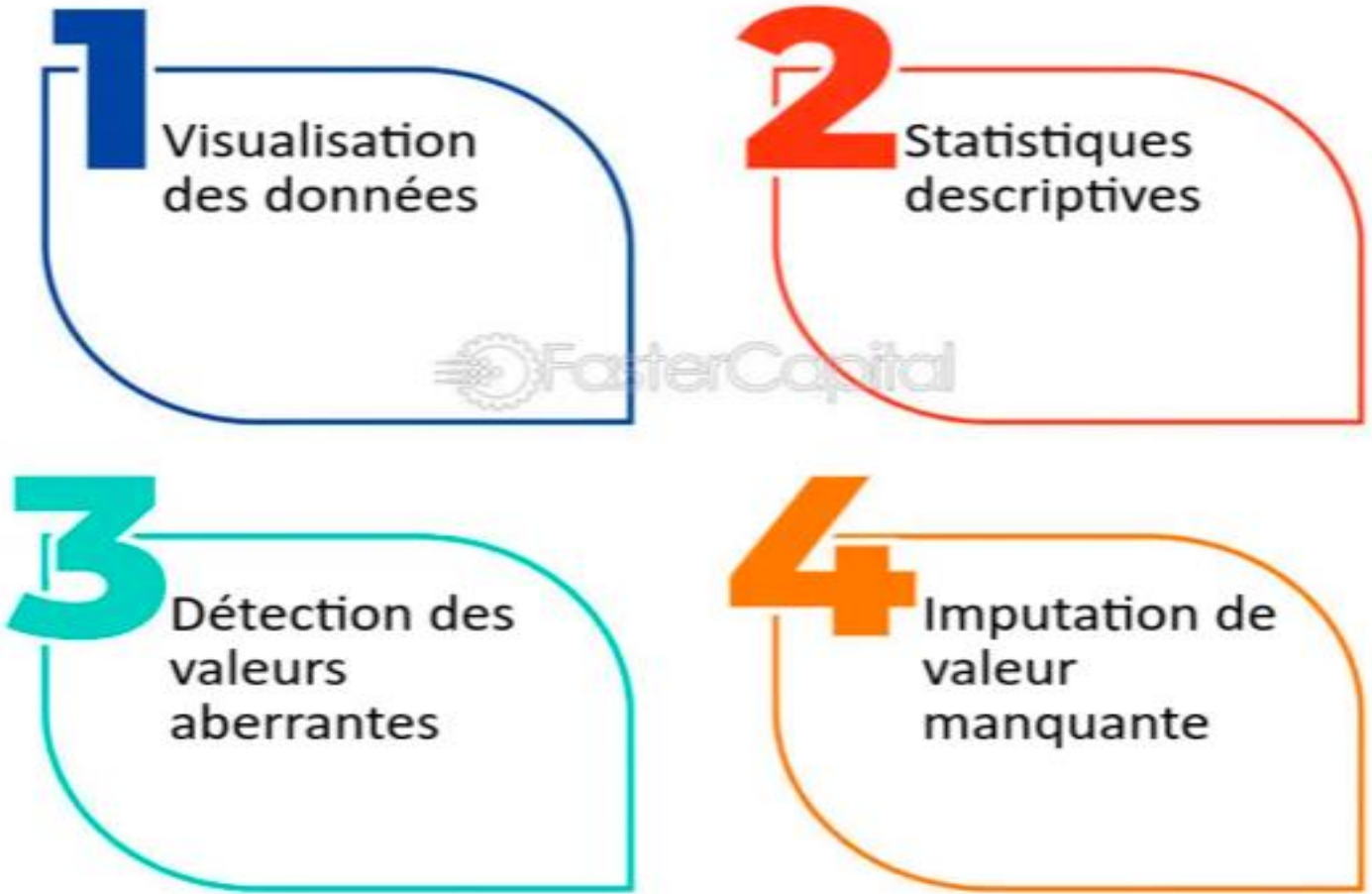
- Méthode de l'écart interquartile (IQR)

On définit : $$IQR = Q3 - Q1$$

Une valeur est considérée comme aberrante si elle est en dehors de :

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

Exemple : Si le revenu médian est de 50 000 $, un salaire de 1 000 000 $ sera détecté comme outlier.


1
Visualisation des données
2
Statistiques descriptives
3
Détection des valeurs aberrantes
4
Imputation de valeur manquante

Etape EDA

B. Outils pour l'analyse exploratoire des données

L'Analyse Exploratoire des Données (EDA) repose sur des outils puissants permettant de visualiser et d'explorer les données avant de les utiliser pour l'entraînement d'un modèle de Machine Learning.

Langages et bibliothèques :

- **Python :**
  - **Pandas** : Manipulation et nettoyage des données (ex : df.describe(), df.info()).
  - **Matplotlib** : Création de graphiques de base (histogrammes, nuages de points, etc.).
  - **Seaborn** : Visualisation avancée pour identifier des tendances et relations entre variables.
  - **Scipy** : Outils statistiques pour tests et distributions de données.

B. Outils pour l'analyse exploratoire des données

R :

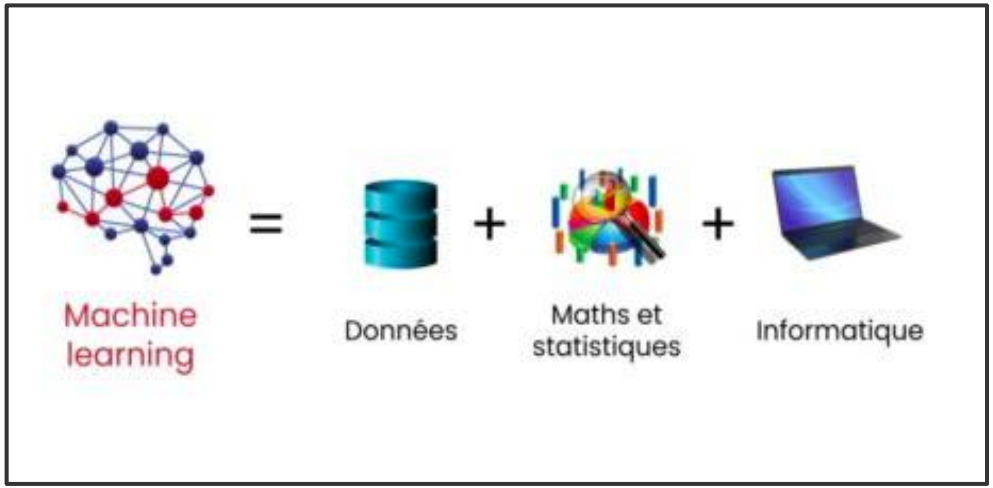
- ggplot2 : Visualisation avancée avec des graphiques personnalisables.
- dplyr : Manipulation des données (filtrage, regroupement, agrégation).
- tidyverse : Collection d'outils pour l'analyse et la visualisation des données.

Excel :

- Outils intégrés pour des graphiques simples et des calculs statistiques.
- Analyse rapide des tendances via les tableaux croisés dynamiques.

IV- Approches statistiques

Les statistiques jouent un rôle fondamental en Machine Learning, car elles permettent de comprendre, modéliser et interpréter les données. Elles sont utilisées à plusieurs étapes du processus d'apprentissage, notamment pour analyser les tendances, vérifier les hypothèses et évaluer la performance des modèles prédictifs.



On distingue deux grandes catégories d'approches statistiques en Machine Learning :

- Statistiques descriptives : Elles permettent d'analyser et de résumer les données à travers des indicateurs numériques et des visualisations.
    - Exemples : Moyenne, médiane, mode, écart-type, histogrammes.
- Statistiques inférentielles : Elles servent à tirer des conclusions à partir d'un échantillon de données et à généraliser les résultats à une population plus large.
    - **Exemples** : Tests d'hypothèses, intervalles de confiance, régressions.

A- Statistique descriptive

- **Mesures de tendance centrale :**
  - Moyenne : Somme des valeurs divisée par le nombre total d'observations.
  - Médiane : Valeur centrale d'un ensemble de données triées.
  - Mode : Valeur la plus fréquente dans un ensemble de données.

  Exemple : Pour les notes d'étudiants [10, 12, 14, 14, 16], la moyenne est 13.2, la médiane est 14, et le mode est 14.
- **Mesures de dispersion :**
  - Écart-type : Indique la variation ou la dispersion des données par rapport à la moyenne.
  - Variance : Moyenne des carrés des écarts à la moyenne.

  Exemple : Un faible écart-type signifie que les valeurs sont proches de la moyenne, tandis qu'un écart-type élevé indique une grande variabilité.

A- Statistique descriptive

- **Distribution des données :**
  - **Loi normale :** Distribution en cloche où la majorité des valeurs sont proches de la moyenne.
  - **Loi de Poisson :** Modélise les événements rares dans un intervalle de temps ou d'espace donné.

Exemple : La distribution de la taille des adultes suit souvent une loi normale, tandis que le nombre d'appels reçus dans un centre d'appels suit une loi de Poisson.

B- Tests Statistiques en ML

Les tests statistiques permettent de vérifier des hypothèses et d'évaluer des relations entre différentes variables avant d'appliquer des modèles de Machine Learning.

- **Test de Student (t-test) :**
  - Utilisé pour comparer les moyennes de deux groupes et déterminer si une différence est statistiquement significative.
  - ***Exemple*** : Comparer les performances de deux algorithmes de classification sur des ensembles de données différents.
  - ***Cas d'usage en ML* :** Vérifier si une augmentation des données d'entraînement améliore réellement la performance d'un modèle.
- **Test du khi-deux (χ^2) :**
  - Évalue l'indépendance entre deux variables catégoriques en comparant les fréquences observées avec les fréquences attendues.

B- Tests Statistiques en ML

- ***Exemple* :** Vérifier s'il existe une relation entre le niveau d'éducation et la préférence pour un type de produit.
- ***Cas d'usage en ML* :** Sélection de variables pertinentes en évaluant la dépendance entre les features et la variable cible.

- **ANOVA (Analyse de la variance)** :
  - Teste s'il existe une différence significative entre les moyennes de plusieurs groupes.
  - ***Exemple* :** Comparer la précision de trois modèles de classification entraînés avec différentes configurations d'hyperparamètres.
  - ***Cas d'usage en ML* :** Comparer l'effet de différentes techniques de prétraitement sur la performance d'un modèle.

C- Importance des statistiques dans le ML

Les statistiques sont essentielles en machine learning (ML) pour plusieurs raisons :

1. **Compréhension des données :** Elles permettent d'analyser les données, d'identifier des tendances et des anomalies.
2. **Prétraitement des données :** Les statistiques aident à traiter les données manquantes, à identifier les valeurs aberrantes et à normaliser les données.
3. **Modélisation :** Elles sont à la base de nombreuses méthodes de ML, comme la régression, où des concepts comme l'estimation des paramètres et les distributions de probabilité sont utilisés.
4. **Évaluation des modèles :** Les statistiques servent à mesurer la performance des modèles à l'aide de métriques (précision, rappel, etc.) et à vérifier leur significativité.
5. **Optimisation des modèles :** Les techniques comme la validation croisée reposent sur des principes statistiques pour ajuster les hyperparamètres et sélectionner les meilleures caractéristiques.

C- Importance des statistiques dans le ML

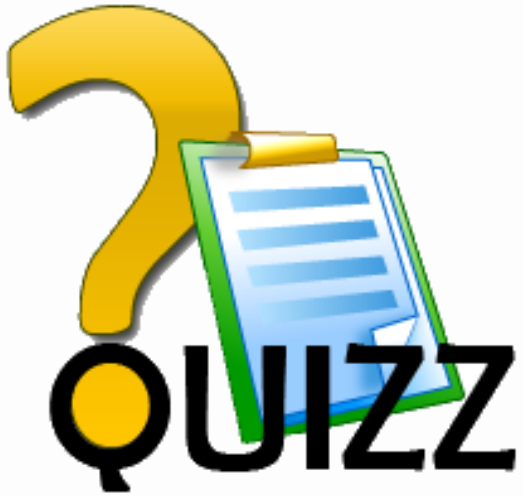
6. Interprétation des résultats : Elles aident à interpréter les prédictions des modèles avec des outils comme les intervalles de confiance et les tests de significativité.

7. Gestion de l'incertitude : Le ML traite souvent des données bruitées, et les statistiques permettent de quantifier cette incertitude, rendant les modèles plus robustes.

8. Sélection des features et réduction de dimension :

- *Exemple* : Utilisation de l'ACP (Analyse en Composantes Principales) pour réduire la dimensionnalité d'un dataset et améliorer les performances des modèles.
- *Impact* : Permet de réduire le surapprentissage et d'améliorer la vitesse de calcul des modèles.

V- Quiz et Travaux Pratique



VI- Conclusion

En somme, nous avons exploré les bases du Machine Learning et de l’Analyse Exploratoire des Données (EDA), mettant en évidence leur rôle essentiel dans la création de modèles performants. L'EDA permet de mieux comprendre les données, tandis que les statistiques jouent un rôle clé dans la validation des hypothèses et la sélection des variables pertinentes. Les différentes approches d’apprentissage machine, qu’elles soient supervisées, non supervisées ou par renforcement, offrent des solutions adaptées à divers types de données et problématiques.

En définitive, le Machine Learning est un outil puissant pour résoudre des problèmes complexes et effectuer des prédictions précises.