INTRODUCTION ET ÉTHIQUE DE L'IA - MACHINE LEARNING

Expliquer et mobiliser les concepts de base en l'éthique de l'IA et des algorithmes. Choisir et à appliquer les algorithmes appropriés en fonction du problème et des données et explorer leurs applications.

1) Fondements, Applications et éthiques de l'IA
2) Apprentissage supervisé et non supervisé
3) Prétraitement et la répartition des données
4) Interprétation des résultats d'un modèle ML

AU: 2025/2026

PR. AZROUMAHLI CHAIMAE

AGENDA

INTRODUCTION ET ÉTHIQUE DE L'IA

Partie 1 : Introduction à l'Intelligence Artificielle

Partie 2 : Éthiques de l'IA et Réglementation générale sur la protection des données (RGPD)

Présentations : Applications de l'IA (Note inclus dans le CC)

CC + Examen

MACHINE LEARNING

Chapitre 1 : Prétraitement des données
Chapitre 2 : Apprentissage Supervisé
Chapitre 3 : Apprentissage non supervisé
Chapitre 4 : Sélection de modèles et validation croisée
Chapitre 5 : Les algorithmes ensemblistes
Chapitre 6 : Optimisation des hyperparamètres
Chapitre 7 : Interprétation des résultats d'un modèle ML

CC + Examen

2

AGENDA - MACHINE LEARNING

- **Chapitre 1: Prétraitement des données**
- Chapitre 2: Apprentissage Supervisé
- Chapitre 3 : Apprentissage non supervisé
- Chapitre 4 : Sélection de modèles et validation croisée
- Chapitre 5 : Les algorithmes ensemblistes
- Chapitre 6 : Optimisation des hyperparamètres
- Chapitre 7 : Interprétation des résultats d'un modèle ML

3

CHAPITRE 1: PRÉTRAITEMENT DES DONNÉES

SÉANCE 1

1) Les données
2) Prétraitement des données
3) Mesures d'évaluation

4

1) LES DONNÉES

1.1. Définition

Les données peuvent être vues comme une **collection d'objets** (enregistrements) et leurs **attributs**:

Attribut

- Un **attribut** est une propriété et ou une caractéristique de l'objet
- L'attribut est également appelé caractéristique, variable, champ
- Exemple: température, poids…

Objet

- Un **objet** est décrit par un ensemble d'attributs
- L'objet est également appelé enregistrement, observation, entité ou instance

5

1) LES DONNÉES

1.1. Définition

Attributs

Objets

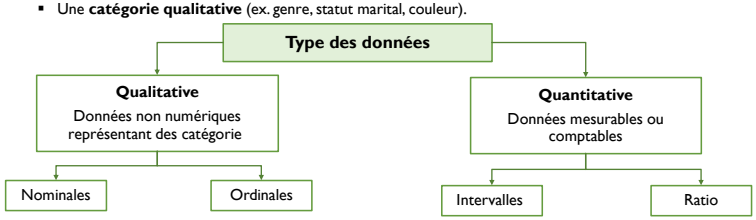
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

6

1) LES DONNÉES

1.2. Types des données

- En Machine Learning, les **données** représentent les **observations ou mesures** utilisées pour **entraîner**, **valider** et **tester** un modèle.
- Chaque **donnée** peut être :
  - Une **valeur numérique** (ex. revenu, âge, température),
  - Une **catégorie qualitative** (ex. genre, statut marital, couleur).



1) LES DONNÉES

1.2.1. Types des données Qualitatives

Nominales

- Ils présentent des catégories que l'on nomme avec un nom, comme par exemple le nom des journaux, le signe astrologique.
- Le seul calcul faisable sur les variables nominales est le nombre d'éléments ou pourcentage par catégorie.

Exemple:

Quel est votre sexe:
- Homme
- Femme

Votre couleur de cheveux:
- Noir
- Blond
- Châtain
- Autre

Ordinales

- Ils sont des catégories qui sont naturellement ordonnées.
- Ils désigne le rang : un peu, moyen, beaucoup, énormément
- Ça peut être le classement à une course, par exemple ou le résultat à questionnaire sur une échelle de **Likert**.

Exemple:

8

1) LES DONNÉES

1.2.2. Types des données Quantitatives

Données discrètes

- Ce sont des valeurs comptables et finies, souvent des entiers.
- Elles représentent des quantités précises.
- Exemples : Nombre d'étudiants dans une classe, Nombre de livres sur une étagère, Taille de chaussure (36, 37, 38…)
- Remarque : On ne peut pas avoir une demi-personne ou 2,5 livres !

Données continues

- Ce sont des valeurs mesurables, pouvant prendre toute valeur dans un intervalle, y compris des fractions ou décimales.
- Exemples : Âge (21,5 ans), Taille (1,73 m), Revenu (85 000,50 €), Température (36,7 °C)

Notion de "zéro absolu" (True Zero)

- Le zéro absolu indique l'absence totale de la quantité mesurée.
- Par exemple :
  - 0 kg → absence totale de poids,
  - 0 K (Kelvin) → absence totale de chaleur,
  - En revanche : 0°C ne signifie pas "pas de chaleur" → c'est un faux zéro (false zero).

1) LES DONNÉES

1.2.3. Sous-types de données Numériques

Interval
- Valeurs numériques avec intervalles égaux, mais sans vrai zéro
- Exemple: Température (°C ou °F), pH, …

Ratio
- Valeurs numériques avec intervalles égaux et vrai zéro
- Exemple: Poids, âge, revenu, vitesse

- Variables quantitatives contiennent des valeurs mesurables:
  - De **rapports**, exemple : distance, durée, valeur;
    - Discrètes exemple : âge, nombre d'habitants;
    - Continues exemple : distance.
  - D'**intervalles** exemple :date de naissance, heure d'arrivée;
    - Discrètes exemple : date en général;
    - Continues exemple : température.

1) LES DONNÉES

1.2.3. Sous-types de données Numériques

✓ Nous pouvons, non seulement, ordonner les items qui sont mesurés, mais également **mesurer** et **comparer** les tailles des différences entre elles.

Exemple :

- Nous pouvons dire qu'une température de 40 degrés Celsius est plus haute qu'une température de 30 degrés, et qu'une augmentation de 20 à 40 degrés est deux fois plus qu'une augmentation de 30 à 40 degrés.
- A 0°c, il y a toujours une température.

✓ Les **variables de ratios** sont très semblables à celles **d'intervalle** avec un point nul absolu identifiable

- Dans une donnée de ratio, le zéro signifie réellement l'absence de quelque chose.

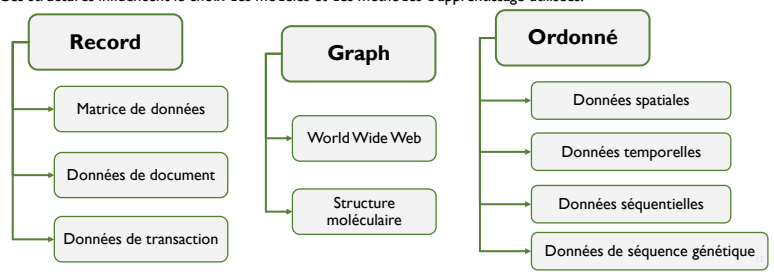
Exemple :

- Pour la durée d'un test, à 0, il n'y pas de temps.
- Si vous avez zéro produit laitier dans votre panier, c'est qu'il n'y a réellement aucun produit laitier.

1) LES DONNÉES

1.3. Types de structures de données en Machine Learning:

- Outre la distinction numérique / catégorielle, les données peuvent aussi être classées selon leur **structure**, c'est-à-dire la façon dont les observations sont **organisées** et **liées** entre elles.
- Ces structures influencent le choix des modèles et des méthodes d'apprentissage utilisées.



1) LES DONNÉES

1.3.1. Les Records: Données Tabulaires

- Les données de type ***record*** sont les plus courantes : chaque observation (ou ligne) représente un **enregistrement** avec plusieurs **attributs** (ou colonnes).

Exemple :

Refund	Marital Status	Taxable Income	Cheat
Yes	Single	125K	No
No	Married	100K	No

Domaines :
- Analyse financière
- Diagnostic médical
- Prédiction de churn (fidélité client)
- Crédit scoring

Modèles typiques :
- Régression linéaire / logistique
- Arbres de décision, Random Forest
- SVM, kNN, Réseaux de neurones fully-connected

1) LES DONNÉES

1.3.2. Matrice de donnée

- Les données sont représentées sous forme de **matrices** (2D ou multi-dimensionnelles), où :
  - les **lignes** représentent les observations,
  - les **colonnes** représentent les variables ou caractéristiques.

Exemple :

- En traitement d'image, une image est une **matrice de pixels** :

$$Image = \begin{bmatrix} 100 & 32 & 61 \\ 900 & 34 & 7038 \\ 12 & 19 & 986 \end{bmatrix}$$

- **Modèles typiques :**
  - PCA (réduction de dimension)
  - CNN (Convolutional Neural Network)
  - SVD, Autoencoders

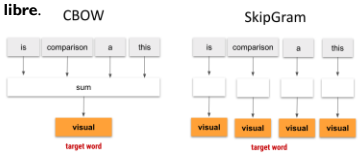
1) LES DONNÉES

1.3.3. Données de document (Textuelles)

- Ce sont des données **non structurées** sous forme de **texte libre**.

Exemples : Articles, commentaires, emails, tweets, etc.

- **Représentation :**
  - Bag of Words (BoW)
  - TF-IDF
  - Word Embeddings (Word2Vec, BERT…)
- **Modèles typiques :**
  - NLP (Natural Language Processing)
  - LSTM, Transformers



1.3.4. Données de transaction

- Chaque observation représente une **transaction composée d'un ensemble d'items**.

Exemple typique : données d'achat, logs d'activités.

- **Modèles typiques :**
  - Association Rules (Apriori, FP-Growth)
  - Recommandation de produits

1) LES DONNÉES

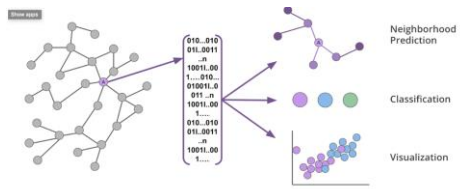
1.3.5. Données de graphe (Graph Data)

- Les données sont organisées sous forme de **nœuds (entités)** et **arêtes (relations)**.

Exemples :

- **Web graph** : pages web (nœuds) et liens hypertextes (arêtes)
- **Réseaux sociaux** : utilisateurs et leurs connexions
- **Structures moléculaires** : atomes et liaisons chimiques

- **Modèles typiques :**
  - Graph Neural Networks (GNN)
  - PageRank, Node2Vec



1) LES DONNÉES

1.3.6. Données ordonnées (spatio-temporelles et séquentielles)

Ces données ont une **dépendance d'ordre** (dans le temps, l'espace ou la séquence).

Données temporelles (Time Series)
- Exemples : température, ventes mensuelles, signaux ECG
- Modèles : RNN, LSTM, ARIMA

Données spatiales
- Exemples : localisation GPS, cartes géographiques, imagerie satellite
- Modèles : CNN, modèles géostatistiques

Données séquentielles
- Exemples : texte, séquences ADN, clics utilisateur
- Modèles : RNN, GRU, Transformers

Données de séquence génétique
- Exemples : chaînes de nucléotides (A, T, G, C)
- Modèles : Bioinformatics ML, LSTM, 1D CNN

17

CHAPITRE 1: PRÉTRAITEMENT DES DONNÉES

SÉANCE 1

1) Les données
2) Prétraitement des données
3) Mesures d'évaluation

21

2) PRÉTRAITEMENT DES DONNÉES

2.1. Exemple Introductif

Soit l'ensemble de données suivant auquel une technique du Machine Learning va être appliqué pour répondre à une question stratégique pour l'entreprise :

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airnair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

22

2) PRÉTRAITEMENT DES DONNÉES

2.1. Exemple Introductif

Etape 1: Corrections des doublons, des erreurs de saisie

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airnair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

23

2) PRÉTRAITEMENT DES DONNÉES

2.1. Exemple Introductif

Etape 2: Intégrité de domaine

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airnair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

24

2) PRÉTRAITEMENT DES DONNÉES

2.1. Exemple Introductif

Etape 3: Vérification des Information manquante:

- ✓ Cas où les champs ne contiennent aucune donnée.
- ✓ Parfois intéressant de conserver ces enregistrements car l'absence d'information peut être informative (e.g. fraude).

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	25/7/2005	BD
43342	Airnair	Rue de la source, Brest	30/05/20s05	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémol	Rue du moulin, Paris	NULL	Maison

25

2) PRÉTRAITEMENT DES DONNÉES

2.2. Pourquoi prétraiter les données ?

- Avant toute application d'un algorithme de Machine Learning, les données brutes doivent être préparées, nettoyées et transformées.
- En effet, les modèles d'apprentissage automatique ne peuvent produire des résultats fiables que si les données utilisées sont cohérentes, complètes et représentatives du phénomène étudié.

Dans la pratique, les données collectées à partir de différentes sources (bases de données, capteurs, formulaires, logs, etc.) sont souvent :

- **Incomplètes** (valeurs manquantes ou nulles),
- **Incohérentes** (erreurs de saisie, doublons, unités différentes),
- **Bruyantes** (valeurs aberrantes, erreurs de mesure),
- **Mal Structurées** (formats hétérogènes, typages erronés),
- **Non Normalisées** (échelles de valeurs très différentes).

Le prétraitement des données consiste donc à :

- ✓ **Nettoyer les données** (supprimer ou corriger les valeurs erronées),
- ✓ **Gérer** les valeurs **manquantes** et les **doublons**,
- ✓ **Convertir les données** dans un format utilisable (catégorisation, encodage, normalisation),
- ✓ et **Préparer** les **attributs pertinents** pour l'apprentissage (feature engineering).

2) PRÉTRAITEMENT DES DONNÉES

2.3. Principales étapes dans le prétraitement

- Le prétraitement des données regroupe l'**ensemble des techniques** qui permettent de rendre les données brutes exploitables par les algorithmes d'apprentissage automatique.

(1) Nettoyage des données	• Supprimer ou corriger les erreurs, valeurs manquantes, doublons ou incohérences présentes dans le jeu de données.
(2) Intégration des données	• Combiner plusieurs sources de données hétérogènes (bases de données, fichiers CSV, APIs, capteurs, etc.) pour former un jeu de données unifié et cohérent.
(3) Transformation des données	• Adapter les données au format, à l'échelle et au type attendus par les algorithmes de Machine Learning.
(4) Réduction des données	• Diminuer le volume ou la dimensionnalité des données tout en préservant leur information essentielle, afin d'améliorer : la vitesse d'entraînement, la précision des modèles, et la lisibilité des résultats.

27

2) PRÉTRAITEMENT DES DONNÉES

2.4. Nettoyage des données

- Le nettoyage des données est un processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes ➔ Cette étape fondamentale du **préparation des données**.

Problèmes courants	Techniques utilisées
• Données manquantes (NULL, NaN, champs vides) • Valeurs aberrantes ou extrêmes • Erreurs de saisie ou de format (ex. 11/11/1111 comme date) • Doublons d'enregistrements (même client, plusieurs lignes) • Incohérence d'unité (€, $, %, etc.)	• Suppression des enregistrements erronés ou inexploitables • Remplacement des valeurs manquantes (moyenne, médiane, mode, régression…) • Détection et traitement des valeurs aberrantes (z-score, IQR) • Uniformisation des formats (date, devise, casse des textes)

- **Objectif de cette étape :**
  - Garantir que seules des données propres et de haute qualité sont transférées vers les systèmes cibles.
  - Améliorer la cohérence, fiabilité et valeur des données.

28

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes

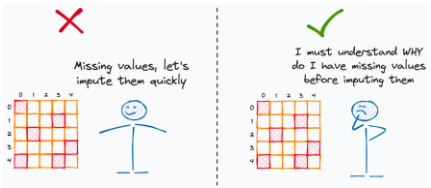
- Les **valeurs manquantes** (ou **Missing Values**) correspondent à des attributs sans valeur observée dans un jeu de données.
  - Autrement dit, certaines cases du tableau de données sont vides, NULL, ou contiennent des valeurs incohérentes (comme "?" ou "NaN").
- Les causes possibles des valeurs manquantes sont que les données peuvent être incomplètes pour plusieurs raisons :
  - **Problème technique:** Panne de capteur, perte de signal, erreur d'enregistrement.
  - **Erreur humaine ou omission:** Données non saisies par oubli, incompréhension ou manque de temps.
  - **Incohérences détectées:** Valeurs supprimées volontairement après détection d'anomalies.
  - **Pertinence jugée faible:** Certaines informations jugées non essentielles lors de la collecte.
  - **Fusion de données hétérogènes:** Lors de l'intégration de sources multiples, certaines variables ne sont pas présentes dans tous les ensembles.

29

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes: Pourquoi traiter les valeurs manquantes ?

- Les algorithmes de Machine Learning ne tolèrent pas toujours les données incomplètes. Les valeurs manquantes peuvent :
  - Perturber les calculs statistiques (moyenne, variance, corrélation),
  - Fausser l'entraînement du modèle (biais ou perte d'information),
  - Réduire la taille du jeu de données si elles sont nombreuses,
  - et donc Diminuer la précision et la robustesse du modèle.



30

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes: Comment remplir les trous ?

Suppression (Ignore/Delete)

Supprimer les lignes ou colonnes contenant des valeurs manquantes.

Avantages :
- Simple à mettre en œuvre.
- Évite d'introduire des biais dus à de fausses estimations.

Inconvénients :
- Risque de **perte d'information importante** si beaucoup de valeurs sont manquantes.
- Réduction de la taille du Dataset → moins représentatif.

Tolérance (Analyse avec données manquantes)

Certaines méthodes d'analyse ou algorithmes peuvent **travailler directement avec des valeurs manquantes**, sans ne les supprimer ni les remplir.

Avantages :
- Conservation de toutes les observations.
- Utile si les algorithmes intègrent une gestion interne (ex. arbres de décision).

Inconvénients :
- Non applicable à tous les modèles.
- Risque de sous-utilisation d'information.

Imputation

Remplacer les valeurs manquantes par des estimations cohérentes basées sur d'autres données.

Avantages :
- Préserve la taille du jeu de données.
- Permet une exploitation complète des informations disponibles.

Inconvénients :
- Risque d'introduire du **bruit artificiel** ou du **biais**.
- Complexité croissante avec des imputations avancées.

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes: Imputation par moyenne / médiane

- Calculer la **moyenne** / **médiane** des valeurs non manquantes dans une colonne,
- Remplacer les valeurs manquantes dans chaque colonne séparément et indépendamment des autres.
- Ne peut être utilisé qu'avec des données numériques.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes: Imputation par (le plus fréquent) ou (zéro / constante)

- **Plus fréquent:**
  - Remplacer les données manquantes par les valeurs les plus fréquentes dans chaque colonne.
  - Fonctionnel pour les données discrète.
- **Zéro/Constante:**
  - Remplace les valeurs manquantes par zéro ou une valeur constante.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

2) PRÉTRAITEMENT DES DONNÉES

2.4. Données manquantes: Imputation utilisant un algorithme

- Remplacer les données manquantes par la valeur la plus probable.
- Utiliser des algorithmes pour estimer la valeur des données manquantes.

Exemple:

- Imputation par le centre du groupe.
- Imputation à partir des k plus proches voisins.
- Imputation par une moyenne partielle.

Méthode	Description	Exemple
Constante	Remplacer par une valeur fixe	"Inconnu", 0, moyenne globale
Statistique	Utiliser la moyenne, médiane ou mode	df['Age'].fillna(df['Age'].mean())
Par similarité	Basée sur des enregistrements proches (k-NN)	Moyenne des k plus proches voisins
Régression	Prédire la valeur manquante via un modèle ML	Prédire le revenu selon âge, métier, etc.
Multivariée	Combinaison de plusieurs variables corrélées	Méthodes MICE (Multiple Imputation by Chained Equations)

2) PRÉTRAITEMENT DES DONNÉES

2.5. Données Bruitées: Correction des données bruitées

- Le **bruit** dans un jeu de données correspond à **une erreur aléatoire** ou **une variation imprévisible** affectant la valeur réelle d'une variable.
- Il ne reflète pas une tendance ou un comportement réel, mais plutôt une perturbation accidentelle dans la mesure ou la saisie des données.
- Causes principales du bruit: Les données bruitées peuvent apparaître pour plusieurs raisons :
  - ✓ Instrument de mesure défectueux (erreur de capteur, calibrage incorrect)
  - ✓ Erreur de saisie ou de transmission (typos, données mal enregistrées ou mal transférées)
  - ✓ Limitation technologique (résolution insuffisante, perte d'informations)
  - ✓ Incohérence dans les conventions de nommage (unités ou formats différents)
- Pourquoi corriger le bruit ?
  - ✓ Le bruit peut : Rendre les modèles moins précis,
  - ✓ Fausser les relations statistiques,
  - ✓ et Augmenter la variance de l'apprentissage.
  - ✓ Corriger ou réduire le bruit permet donc d'obtenir des modèles plus stables et plus généralisables.

2) PRÉTRAITEMENT DES DONNÉES

2. 5. Données Bruitées: Partitionnement simple ▸ lissage par moyenne

- Pour chaque partition, on calcule **la moyenne des valeurs** ▸ Les données sont ensuite remplacées par la moyenne de la partition.

Exemple : Température quotidienne d'une ville sur une semaine :

$$[30°C, 32°C, 35°C, 31°C, 100°C, 34°C, 33°C]$$

- Le pic de $100°C$ est probablement une erreur (une donnée bruitée).
- Pour atténuer cet effet, on divise les températures en groupes de 3 jours et calculer la moyenne de chaque groupe.
- Premier groupe (jours 1-3) : $Moyenne = \frac{(30+32+35)}{3} = 32.33°C$
- Deuxième groupe (jours 4-6) : $Moyenne = \frac{(31+100+34)}{3} = 55°C$
- Troisième groupe (jours 7) : $Moyenne = 33°C$ $(seul\ jour\ dans\ le\ groupe)$
- Résultat lissé :

$$[32.33°C, 32.33°C, 32.33°C, 55°C, 55°C, 55°C, 33°C]$$

37

2) PRÉTRAITEMENT DES DONNÉES

2.5. Données Bruitées: Partitionnement simple ▸ lissage par médiane

- On calcule **la médiane** pour chaque partition au lieu de la moyenne ▸ C'est utile quand les données contiennent des valeurs aberrantes, car la médiane est moins affectée par ces valeurs extrêmes.

Exemple : Prenons un ensemble de données représentant les revenus mensuels de 5 employés :

[2 000€, 2 100€, 2 500€, **50 000€**, 2 400€]

- Le revenu de 50 000€ est probablement une valeur aberrante.
- Ordonnons les données : [2 000€, 2 100€, 2 400€, 2 500€, 50 000€]
- Médiane des données est la valeur centrale : 2400€
- Résultat lissé : [2 000€, 2 100€, 2 500€, **2 400€**, 2 400€]

38

2) PRÉTRAITEMENT DES DONNÉES

2.5. Données Bruitées: Partitionnement simple ▸ Lissage par Frontières

- Chaque partition est représentée par ses **valeurs extrêmes** (minimum et maximum), et les données intermédiaires sont remplacées par ces frontières.

Exemple : Supposons que tu étudies les âges d'un groupe de personnes :

$$[18\ ans,\ 20\ ans,\ 22\ ans,\ 90\ ans,\ 23\ ans,\ 21\ ans]$$

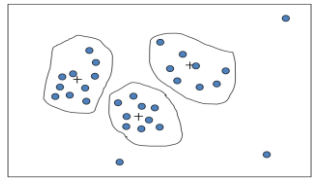
- L'âge de 90 ans est une valeur clairement aberrante.
- Partition : $[18, 20, 22, 90, 23, 21]$
- Valeurs frontières acceptables : $Minimum = 18\ ans, Maximum = 23\ ans$
- Résultat lissé : $[18\ ans, 18\ ans, 23\ ans, \mathbf{23\ ans}, 23\ ans, 23\ ans]$

39

2) PRÉTRAITEMENT DES DONNÉES

2.5. Données Bruitées: Clustering ▸ suppression des exceptions

- Le clustering consiste à **regrouper les données en clusters** (groupes) selon leurs similitudes.
- Cela peut aider à identifier les points de données qui ne correspondent à aucun cluster, ce qui signifie souvent qu'ils sont des valeurs bruitées.



Exemple :

- Les tailles des personnes dans un groupe :

$$[150\ cm, 160\ cm, 170\ cm, 180\ cm, 190\ cm, 250\ cm]$$

- En appliquant un algorithme de clustering (comme K-Means), la plupart des valeurs seront regroupées dans un ou deux clusters normaux.

40

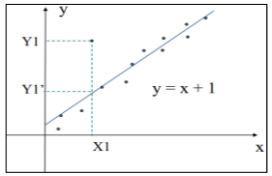
2) PRÉTRAITEMENT DES DONNÉES

2.5. Données Bruitées: Régression

- La régression est utilisée pour modéliser une relation entre des variables et prédire une valeur en se basant sur d'autres variables.
- Si certaines valeurs s'écartent fortement du modèle de régression, elles peuvent être considérées comme des données bruitées.

Exemple :
- On veut analyser l'âge et le poids de patients.
- On applique une régression linéaire pour prédire le poids en fonction de l'âge.

plus l'age augmente, plus le poids augmente, avec quelques variations normales.



- Si un point de donnée (par exemple, une personne de 30 ans pesant 300 kg) s'écarte fortement de la tendance, cela peut être identifié comme une valeur aberrante.

41

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes

- Outre les valeurs manquantes et les données bruitées, les jeux de données peuvent contenir plusieurs autres types d'imperfections qui compromettent la qualité de l'apprentissage. Les plus fréquents sont :

Enregistrements dupliqués	• Les doublons correspondent à des enregistrements identiques ou presque identiques répétés plusieurs fois dans le jeu de données. • Ils faussent les statistiques et biaisent le modèle en donnant un poids excessif à certaines observations.
Données incomplètes	• Les données incomplètes concernent les observations partiellement remplies — certaines colonnes ont des informations présentes, d'autres non.
Données incohérentes	• Les données incohérentes sont des valeurs contradictoires ou non plausibles par rapport à d'autres informations du jeu de données.

42

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Intégration des données et shémas

- **Intégration des données :** Combinaison de différentes sources en une seule
- **Intégration des schémas :** Intégrer les métadonnées de différentes sources
- **Problème de nommage :** identifier les différents noms des mêmes données réelles

Exemple : $numClient \Leftrightarrow clientId$
- Détecter et résoudre les conflits de valeurs
- Pour les mêmes entités réelles, les valeurs des attributs provenant de sources hétérogènes sont différentes
- Causes :
  - Représentation différentes,
  - Échelles différentes,

Exemple : cm et pouces, kilogramme et pounds

43

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Gestion de la redondance

- La redondance des données correspond à la présence d'informations répétées ou fortement corrélées dans un jeu de données.
  - ✓ Plusieurs variables apportent la même information, ce qui alourdit inutilement le modèle sans améliorer ses performances.

Méthode de détection → Le test de corrélation : La redondance peut être détectée à l'aide du coefficient de corrélation, qui mesure la force et la direction de la relation entre deux variables :

Méthode	Type de données	Description
Pearson (r)	Données continues	Mesure la corrélation linéaire entre deux variables
Spearman (ρ)	Données ordinales	Mesure la corrélation monotone (ordre croissant/décroissant)
Kendall (τ)	Données ordinales	Mesure la concordance entre paires de données

- ✓ Une corrélation proche de 1 → relation fortement positive
- ✓ Une corrélation proche de -1 → relation fortement négative
- ✓ Une corrélation proche de 0 → pas de relation significative

44

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Gestion de la redondance

Exemple :

- Dans un jeu de données sur des étudiants :

Variable	Description
Note_totale	Score final de l'étudiant
Heures_étude	Nombre d'heures passées à étudier

- Si le coefficient de corrélation de Pearson = 0,99, cela signifie que ces deux variables sont très liées : plus l'étudiant étudie, plus sa note augmente.
- Ces deux attributs sont donc redondants.

Solutions :

- Supprimer une des variables (ex. garder seulement Note_totale).
- Ou fusionner les deux dans une variable composite (score pondéré).

45

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Transformation (codage et normalisation)

- Cette étape est essentielle pour préparer les données en fonction de l'algorithme que l'on souhaite utiliser.
- **Regroupements:**
  - Lorsqu'un attribut contient de nombreuses valeurs discrètes, il peut être difficile de les traiter efficacement.
  - Si l'on souhaite simplifier l'analyse, on peut regrouper les adresses en catégories plus larges.
- **But du regroupement :**
  - Réduire le nombre de valeurs à traiter pour éviter la complexité excessive et améliorer l'efficacité des algorithmes.

46

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Transformation (codage et normalisation)

Exemple 1 : Calculs de distance ou de moyenne :

- Avant transformation :
  - Supposons le champ "Date de naissance" dans le jeu de données. Ce champ est souvent stocké sous forme de chaîne de caractères (par exemple, "15/03/2025").
- Après transformation
  - Pour calculer l'âge ou la distance en années entre deux personnes, il est nécessaire de convertir la "Date de naissance" en un type numérique (comme une différence en années). Cela permet de réaliser des calculs de distance ou de moyenne sur cet attribut.

Exemple 2 : Uniformisation des catégories :

- Si les données sont qualitatives comme "petit", "moyen", "grand", il peut être utile de convertir ces valeurs en nombres (1 pour petit, 2 pour moyen, 3 pour grand), surtout lors de l'utilisation des algorithmes qui nécessitent des entrées numériques.

Uniformisation d'échelle

Certains algorithmes sont basés sur des calculs de distance entre enregistrements : Variations d'échelle selon les attributs peuvent perturber ces algorithmes.

2) PRÉTRAITEMENT DES DONNÉES

2.5. Autres problèmes: Transformation (codage et normalisation)

Problème :

- Age et Revenu: Si l'âge varie entre 20 et 70 ans, mais que le revenu varie entre 10 000 € et 1 000 000 €, les algorithmes qui utilisent des distances pourraient accorder beaucoup plus d'importance au revenu qu'à l'âge, simplement à cause des valeurs plus grandes.

Solution : Une normalisation ou une standardisation pour uniformiser l'échelle :

Normalisation :

- Convertir les valeurs de chaque attribut dans une plage commune, par exemple, entre 0 et 1.
- Cela rend l'âge et le revenu comparables.

Standardisation :

- Transformer les données pour qu'elles aient une moyenne de 0 et un écart-type de 1.

48

2) PRÉTRAITEMENT DES DONNÉES

2.6. Normalisation

- Normaliser certains attributs numériques afin qu'ils varient dans une plage plus petite.
- Ces méthodes permettent de mettre les données sur une même echelle.

Exemple: Normaliser l'attribut Age pour qu'il varie entre 0 et 1.

- Méthode de normalisation:
  1) min-max
  2) z-score
  3) mise à l'échelle décimale

50

2) PRÉTRAITEMENT DES DONNÉES

2.6. Normalisation: Min-Max

- Les valeurs minimale et maximale des données sont extraites et chaque valeur est remplacée selon la formule suivante.

$$v' = \frac{v - min_A}{max_A - min_A}, \quad \begin{cases} A \text{ est les données d'attribut} \\ v', v : \text{La nouvelle et l'ancienne valeur de chaque entrée de données.} \\ max_A, min_A : \text{la valeur absolue minimale et maximale de A.} \end{cases}$$

Exemple: Nous avons un ensemble de données sur les salaires : $20000€, 50000€, 80000€20000€, 50000€, 80000€$

- Nomaliser ces salaires entre 0 et 1.
  - $x\_\min = 20\,000€$
  - $x\_\max = 80\,000€$
- Pour un salaire de $50\,000€ : X = \frac{50000-20000}{80000-20000} = \frac{30000}{60000} = 0.5$
- Le salaire de $50\,000€$ devient 0.5 après normalisation.

51

2) PRÉTRAITEMENT DES DONNÉES

2.6. Normalisation: Z-Score

- Dans cette technique, les valeurs sont normalisées en fonction de la moyenne et de l'écart type des données A.
- La formule utilisée est:

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}, \begin{cases} v', v : la\ nouvelle\ et\ l'ancienne\ valeur\ de\ chaque\ entrée\ de\ données. \\ \sigma_A, \bar{A} : l'écart\ type\ et\ la\ Moyenne\ de\ A\ respectivement. \end{cases}$$

Exemple: Les âges dans une classe : $20ans, 25ans, 30ans, 35ans$

- La moyenne (μ) est 27.5 ans, et l'écart-type (σ) est environ 5.59 ans.
- Pour un étudiant de 30 ans :

$$w = \frac{30 - 27{,}5}{5{,}59} \approx 0.45$$

- Cela signifie que l'âge de 30 ans est à environ 0.45 écart-type au-dessus de la moyenne.

52

2) PRÉTRAITEMENT DES DONNÉES

2.6. Normalisation: Mise à l'échelle décimale

- Cette technique normalise en déplaçant la virgule décimale des valeurs des données. Diviser chaque valeur des données par la valeur absolue maximale des données.
- La formule utilisée est:

$$v' = \frac{v}{10^j}, j : le\ plus\ petit\ entier\ telque \max(|v'|) < 1$$

Exemple: les populations de villes : $150\,000, 250\,0000, 1200000$

- La valeur maximale est $12\,000\,000$, qui a 7 chiffres. Pour normaliser, nous devons diviser chaque valeur par 10^7.
- Pour une ville de $2\,500\,000$ habitants :

$$x_{normalisé} = \frac{2500000}{10^7} = 0{,}25$$

- Cela réduit la valeur à 0.25, facilitant son traitement.

53

2) PRÉTRAITEMENT DES DONNÉES

2.7. Réduction de données

- Obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit les mêmes (ou presque) résultats analytiques.
- Stratégies :
  1) Réduction de dimension
  2) Réduction de numérosité
  3) Discrétisation

54

2) PRÉTRAITEMENT DES DONNÉES

2.7. Réduction de données: Echantillonnage

Réduction en ligne par échantillonnage :

- Pour des raisons de performance. Du fait de la complexité importante des algorithmes d'extraction.
- Plusieurs méthodes : échantillonnage aléatoire, échantillonnage par clustering.

Echantillonnage aléatoire :

- **Sélectionner aléatoirement** un sous-ensemble des données d'origine. L'idée est que si le jeu de données est suffisamment grand, un échantillon représentatif donnera des résultats similaires à ceux de l'ensemble complet.
- **Exemple :** Si nous avons un million de lignes de données sur des clients, on peut en prendre 10 000 aléatoirement pour faire nos analyses, car cela représente une proportion raisonnable des données d'origine.

Echantillonnage par clustering :

- **Regrouper les données similaires** en clusters, puis à sélectionner un ou plusieurs points représentatifs de chaque cluster. Cela garantit que chaque "type" de données est bien représenté dans l'échantillon.
- **Exemple :** Si nous avons des données sur des clients et on les regroupes en trois clusters basés sur des comportements d'achat (par exemple, gros acheteurs, acheteurs moyens, et petits acheteurs), on peut sélectionner des points de chaque cluster pour s'assurer que tous les types de clients sont représentés dans l'échantillon.

55

2) PRÉTRAITEMENT DES DONNÉES

2.7. Réduction de données: Réduction en colonne

- **Réduction en colonne** par suppression des attributs redondants:
  - Cas triviaux (âge et date de naissance).
  - Via une analyse des corrélation entre attributs

56

2) PRÉTRAITEMENT DES DONNÉESS

2.8. Discrétisation

- Diviser l'intervalle de valeurs possibles en sous intervalles.
  - Certains algorithmes acceptent seulement des attributs catégoriques.
  - Réduit le volume des données.
- Répartition des valeurs des attributs :
  - A chaque étape, on cherche à découper l'intervalle des données en K intervalles comportant le même nombre de valeurs.
  - On divise AGE= [0, 100] en A1 = [0, 20] et A2 = [20, 100] si 50 % des clients ont moins de 20 ans.

57

CHAPITRE I: PRÉTRAITEMENT DES DONNÉES

SÉANCE I

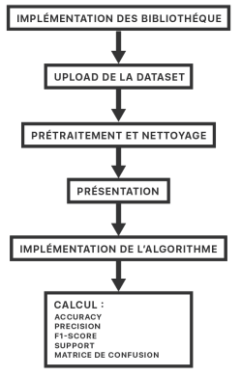
1) Les données
2) Prétraitement des données
3) **Mesures d'évaluation**

58

3) MESURE D'ÉVALUATION

3.1. Construction d'un modèle du Machine Learning

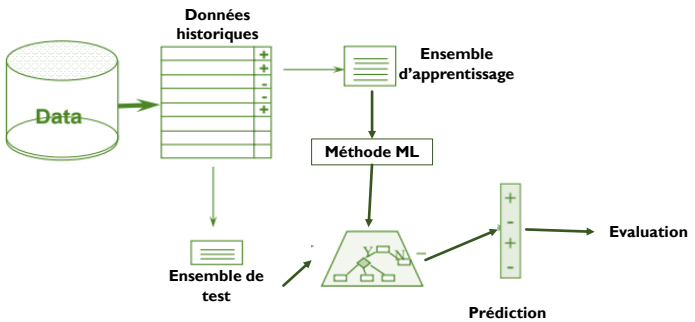
1. Diviser les données en ensemble d'apprentissage et ensembles de test
2. Construire le modèle ML en utilisant l'ensemble d'apprentissage
3. Evaluer le modèle en utilisant l'ensemble de test



59

3) MESURE D'ÉVALUATION

3.1. Construction d'un modèle du Machine Learning



60

3) MESURE D'ÉVALUATION

3.2. Echantillonnage de l'ensemble de données

- L'échantillonnage génère différents sous-ensembles de données à partir de l'ensemble initial D.

Hold-out	k-fold cross-validation	Leave-one-out
• Utiliser deux ensembles de données indépendants, par exemple, ensemble d'apprentissage (2/3), ensemble de test (1/3); avec un Échantillonnage aléatoire • Il est important que les données de test ne soient en aucun cas utilisées pour créer le modèle du ML. • Mieux adapté pour les données très volumineuses	• Il évite les ensembles de tests qui se chevauchent. • Première étape: les données sont divisées en k sous-ensembles de taille égale • Deuxième étape: utiliser k-1 sous ensemble comme données d'apprentissage et un sous ensemble comme données de test; répéter k fois. • Les estimations d'erreur sont moyennées pour produire une estimation d'erreur globale	• Une forme particulière de validation croisée, utilisé pour les données de petite taille. • Définir le nombre de sous ensemble en se basant sur le nombre d'instances d'apprentissage. • C'est-à-dire, pour n instances d'apprentissage, construire le modèle n fois mais à partir de n -1 exemples d'apprentissage … • N'implique aucun sous-échantillonnage aléatoire. • Assez cher en calcul!

61

3) MESURE D'ÉVALUATION

3.3. Qualités attendues d'un modèle DM

Précision	Le taux d'erreur, proportion d'individus mal classés doit être le plus bas possible.
Robustesse	Le modèle doit dépendre peu que possible de l'échantillon d'apprentissage et se généraliser à d'autres échantillons.
Concision	Les règles du modèles doivent être aussi simples et aussi peu nombreuses que possible.
Rapidité de calcul	Apprentissage rapide pour affinement du modèle.
Paramétrage	Pouvoir pondérer les erreurs de classement

62

3) MESURE D'ÉVALUATION

3.4. Méthodes d'évaluation

Classification	Association	Clustering
• Matrice de confusion • Taux d'erreur • Recall / precision • F-mesure • Courbe ROC	• Support • Confidence • Lift	• Modèle Machine Learning : longueur minimale de description

63

3) MESURE D'ÉVALUATION

3.4. Méthodes d'évaluation: TP, FN, FP, TN

	Prédit = Défaut (1)	Prédit = Pas de Défaut (0)
Réel = Défaut (1)	**TP** (vrai positif)	**FN** (faux négatif)
Réel = Pas de Défaut (0)	**FP** (faux positif)	**TN** (vrai négatif)

Interprétation (exemple bancaire)

- **TP (True Positive)** : le modèle a correctement identifié un client risqué. ✅
- **FP (False Positive)** : le modèle classe un bon client comme risqué (erreur → refus injustifié). ❌
- **FN (False Negative)** : le modèle n'a pas détecté un client qui fera défaut (erreur grave). ❌
- **TN (True Negative)** : le modèle a correctement identifié un client solvable. ✅

64

3) MESURE D'ÉVALUATION

3.4. Méthodes d'évaluation: TP, FN, FP, TN

TP (True Positives)	TN (True Negatives)	FP (False Positive)	FN (False Negative)
Les cas où la prédiction est positive, et où la valeur réelle est effectivement positive.	Les cas où la prédiction est négative, et où la valeur réelle est effectivement négative.	Les cas où la prédiction est positive, mais où la valeur réelle est négative.	Les cas où la prédiction est négative, mais où la valeur réelle est positive.
Exemple : le médecin vous annonce que vous êtes malade, et vous êtes vraiment malade.	Exemple : le médecin vous annonce que vous n'êtes pas malade, et vous n'êtes effectivement pas malade.	Exemple : le médecin vous annonce que vous êtes malade, mais vous n'êtes pas malade.	Exemple : le médecin vous annonce que vous n'êtes pas malade, mais vous êtes malade.

65

3) MESURE D'ÉVALUATION

3.5. Taux d'erreur: Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Interprétation:**
  - C'est la proportion totale de bonnes prédictions.
- **Exemple financier :**
  - Sur 1 000 clients, le modèle en classe correctement 900.
  - Accuracy = 900 / 1000 = 0.90 (90%)→ Le modèle semble bon si les classes sont équilibrées.
- **Limite :**
  - Si 95 % des clients sont solvables, un modèle qui dit toujours "solvable" a une accuracy de 95 %, mais il est inutile pour détecter les clients à risque.

66

3) MESURE D'ÉVALUATION

3.5. Taux d'erreur: Précision

$$Precision = \frac{TP}{TP + FP}$$

- **Interprétation:**
  - Parmi les clients que le modèle prédit comme "à risque", combien sont réellement à risque ?
- **Exemple financier :**
  - Le modèle prédit 200 clients à risque.
  - Parmi eux, 160 le sont réellement.
  - Precision = 160 / (160 + 40) = **0.80 (80%)**

  → Le modèle est fiable à 80 % lorsqu'il alerte sur un client à risque.
- **Limite :**
  - **Si la précision est faible** → le modèle "crie au loup" trop souvent (faux positifs).

67

3) MESURE D'ÉVALUATION

3.5. Taux d'erreur: Recall (Rappel ou Sensibilité)

$$Recall = \frac{TP}{TP + FN}$$

- **Interprétation:**
  - Parmi les vrais clients à risque, combien ont été détectés par le modèle ?
- **Exemple financier :**
  - Il y a 100 vrais clients à risque.
  - Le modèle en détecte 85.
  - Recall = 85 / (85 + 15) = 0.85 (85%)

  → Le modèle capture 85 % des mauvais payeurs.
- **Limite :**
  - **Si le rappel est faible** → le modèle laisse passer trop de clients risqués non détectés (FN).

68

3) MESURE D'ÉVALUATION

3.5. Taux d'erreur: F1 (Équilibre entre précision et rappel)

$$F1 = 2 \times \frac{Precision \times Recall}{Precsion + Recall}$$

- **Interprétation:**
  - C'est la moyenne harmonique entre précision et rappel.
  - Permet d'équilibrer les deux lorsqu'ils sont contradictoires.
- **Exemple financier :**
  - $Precision = 0.80, Recall = 0.85$
  - $F1 = 2 \times (0.8 \times 0.85) / (0.8 + 0.85) = 0.825$

  → Le modèle est équilibré : bon à la fois pour détecter les risques et éviter les erreurs.

69

3) MESURE D'ÉVALUATION

3.5. Taux d'erreur: Specificity (ou Taux de vrais négatifs)

$$Specificity = \frac{TN}{TN + FP}$$

- **Interprétation:**
  - Proportion de bons clients correctement reconnus comme tels.
- **Exemple financier :**
  - Sur 800 clients solvables, le modèle en classe 760 comme bons payeurs.
  - Specificity = 760 / (760 + 40) = **0.95 (95%)**

→ Le modèle évite de refuser des clients solvables.

70

3) MESURE D'ÉVALUATION

3.6. Courbe ROC

- Une courbe ROC (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification.
- Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.
- Le taux de vrais positifs (TPR) est l'équivalent du rappel:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

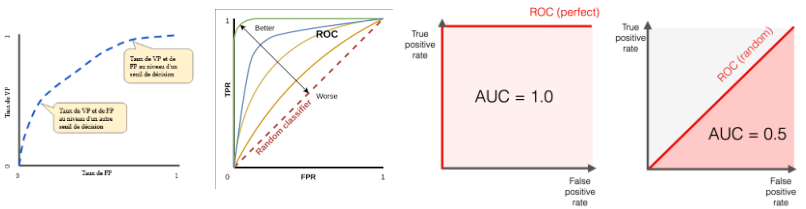
- Le taux de faux positifs (FPR) est défini comme suit

$$FPR = \frac{FP}{FP + TN}$$

71

3) MESURE D'ÉVALUATION

3.6. Courbe ROC



Graphique	Objectif	Interprétation
Courbe ROC / AUC	Mesure du compromis entre rappel et faux positifs	Plus proche du coin supérieur gauche = meilleur modèle

72

AGENDA - MACHINE LEARNING

- Chapitre 1: Prétraitement des données
- **Chapitre 2: Apprentissage Supervisé**
- Chapitre 3 : Apprentissage non supervisé
- Chapitre 4 : Sélection de modèles et validation croisée
- Chapitre 5 : Les algorithmes ensemblistes
- Chapitre 6 : Optimisation des hyperparamètres
- Chapitre 7 : Interprétation des résultats d'un modèle ML

73