esprit
Se former autrement
HONORIS UNITED UNIVERSITIES

Processus de préparation des données

UP: GL/BD
Réalisé par : Equipe ML Appliqué

EUR-ACE® Délivrée par la Commission des Titres d'Ingénieur
CONFERENCE DES GRANDES ÉCOLES
cdio

Plan

1. Objetifs du cours
2. Introduction
3. Problématique
4. Les activités de la preparation

Objetifs du cours

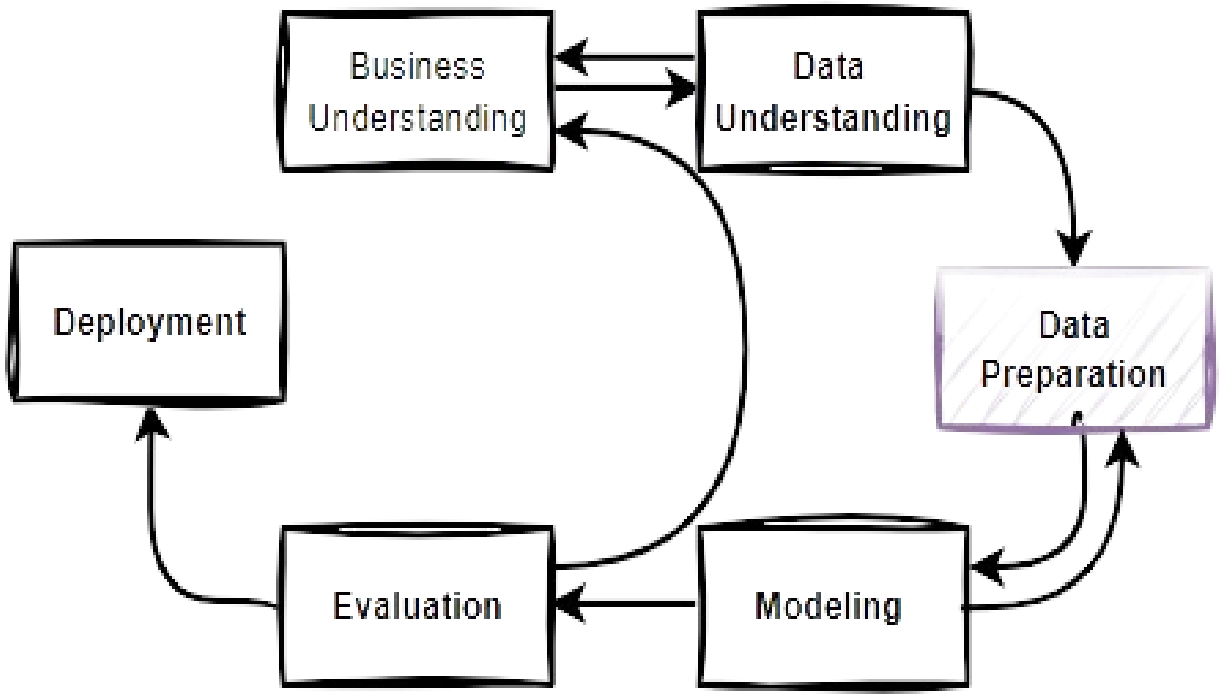
Comprendre l'importance de la préparation des données.

Comprendre les étapes essentielles de la préparation des données.

Appliquer des techniques pour la préparation de données

Introduction

La préparation des données est une étape essentielle pour un projet d'analyse de données ou d'apprentissage automatique.



Cette étape, consiste à transformer des données brutes en un format propre et utilisable par les modèles de machine learning.

Problématique

Avant d'utiliser les données, il est important de traiter les problèmes potentiels qui peuvent impacter la précision des modèles d'apprentissage. Parmi les problèmes les plus fréquents, on trouve :

Données manquantes :

- Certaines valeurs peuvent manquer dans le jeu de données.
- Ce qui peut perturber les algorithmes d'apprentissage.

Données bruitées ou incohérentes :

- Les données peuvent contenir du bruit.
- Des erreurs de saisie ou des valeurs incohérentes.

Valeurs aberrantes (outliers) :

- Certaines observations peuvent avoir des valeurs extrêmes.
- Elles se situent à l'extérieur de la plage normale ou attendue des données.

Problématique

Avant d'utiliser les données, il est important de traiter les problèmes potentiels qui peuvent impacter la précision des modèles d'apprentissage. Parmi les problèmes les plus fréquents, on trouve :

Données manquantes

- Certaines valeurs peuvent manquer dans le jeu de données.
- Ce qui peut perturber les algorithmes d'apprentissage.

Données bruitées ou incohérentes

- Les données peuvent contenir du bruit.
- Des erreurs de saisie ou des valeurs incohérentes.

Valeurs aberrantes (outliers)

- Certaines observations peuvent avoir des valeurs extrêmes.
- Elles se situent à l'extérieur de la plage normale ou attendue des données.

Problématique

Données non normalisées :

- Des algorithmes sensibles à l'échelle des variables.
- Des valeurs non normalisées peuvent entraîner des biais dans les résultats.

Trop grand nombre de dimensions :

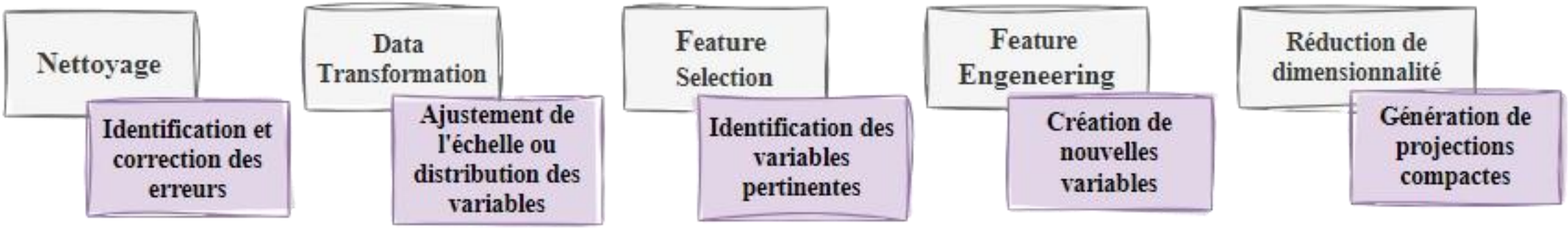
- Un grand nombre de variables peut rendre les modèles d’apprentissage inefficaces.
- introduire de la redondance et accroître le risque de surapprentissage (overfitting).

➡ ***Les données nécessitent souvent une préparation avant d'être exploitées efficacement.***

Intérêt de la preparation de données

- Les modèles apprennent mieux avec des données bien préparées.
- Réduction des erreurs et amélioration des performances.
- Une préparation rigoureuse des données augmente la précision des modèles.

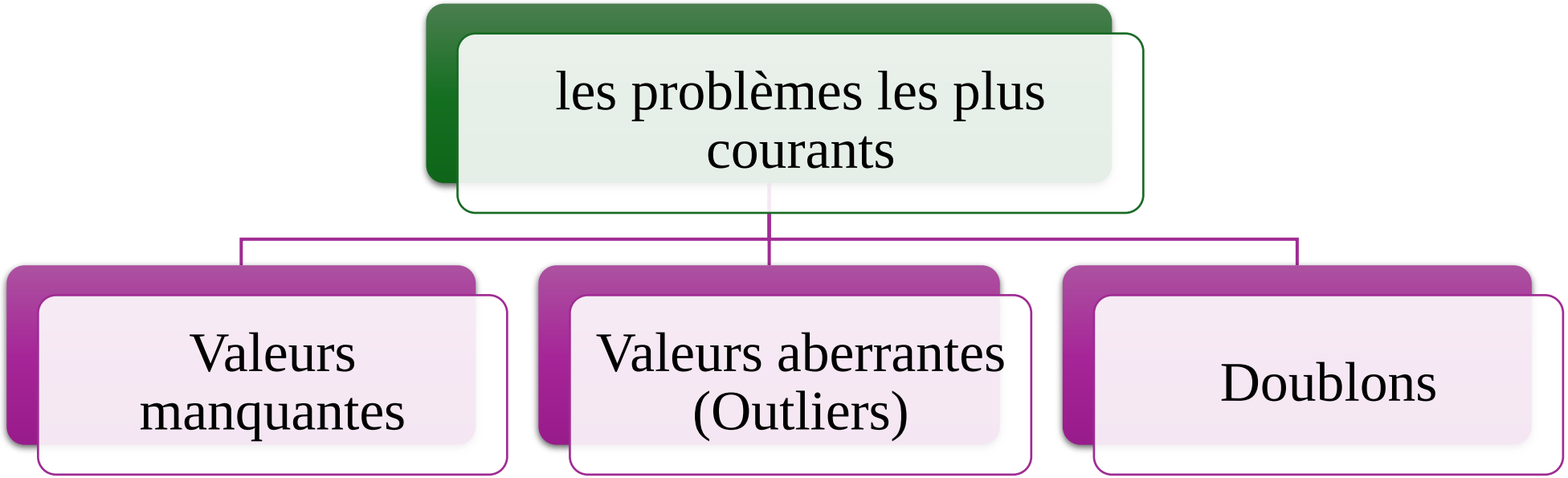
Les activités de la préparation de données



Nettoyage des données

Nettoyage des données
Problèmes courants

Lors du nettoyage des données, plusieurs problèmes **doivent être identifiés et traités**.



Nettoyage des données
Valeurs manquantes

Traitement des valeurs manquantes:

- **Suppression:**

Supprimer des individus avec des données manquantes ou nulles.

- **Imputation des données :**

Remplacer des valeurs manquantes par une valeur fixe ou des valeurs proches.

Nettoyage des données
Valeurs manquantes

Les méthodes d'imputation:

- ✓ **Mean**: Remplacer les valeurs manquantes en utilisant la moyenne le long de chaque colonne.
- ✓ **Median:** Remplacer les valeurs manquantes en utilisant la médiane le long de chaque colonne.

- ✓ **most_frequent:** Remplacer les valeurs manquantes par la valeur la plus fréquente de chaque colonne.
- ✓ **Constant:** remplacer les valeurs manquantes par fill_value.

Nettoyage des données
Valeurs aberrantes

- Les valeurs aberrantes, ou outliers, sont des observations dans un jeu de données qui diffèrent considérablement des autres.
- Elles peuvent résulter d'erreurs de mesure ou de saisie.
- Détecter et traiter les outliers est essentiel pour garantir la fiabilité des analyses et la performance des modèles de machine learning.

Exemple:

Voici un ensemble de données fictif représentant les notes des étudiants sur un test :

Notes : 12, 15, 14, 13, 17, 16, 14, 100

La note **"100"** se distingue des autres, qui se situent entre **12** et **17**.

Cette note pourrait être une **valeur aberrante**, car elle est extrêmement élevée par rapport aux autres.

Nettoyage des données
Valeurs aberrantes

Méthodes de traitement des valeurs aberrantes:

- **Suppression :**

Éliminer les valeurs aberrantes si elles sont dues à des erreurs ou si elles perturbent les analyses.
Remarque: La suppression à utiliser avec précaution pour éviter de perdre des informations importantes.

- **Transformation des données :**

Appliquer des transformations (logarithmique, racine carrée) pour réduire l'impact des outliers.

- **Imputation :**

Remplacer les valeurs aberrantes par des valeurs plus raisonnables (médiane, moyenne).

Remarque: Si les valeurs aberrantes sont pertinentes (ex. : anomalies significatives), elles doivent être étudiées séparément.

Nettoyage des données
Doublons

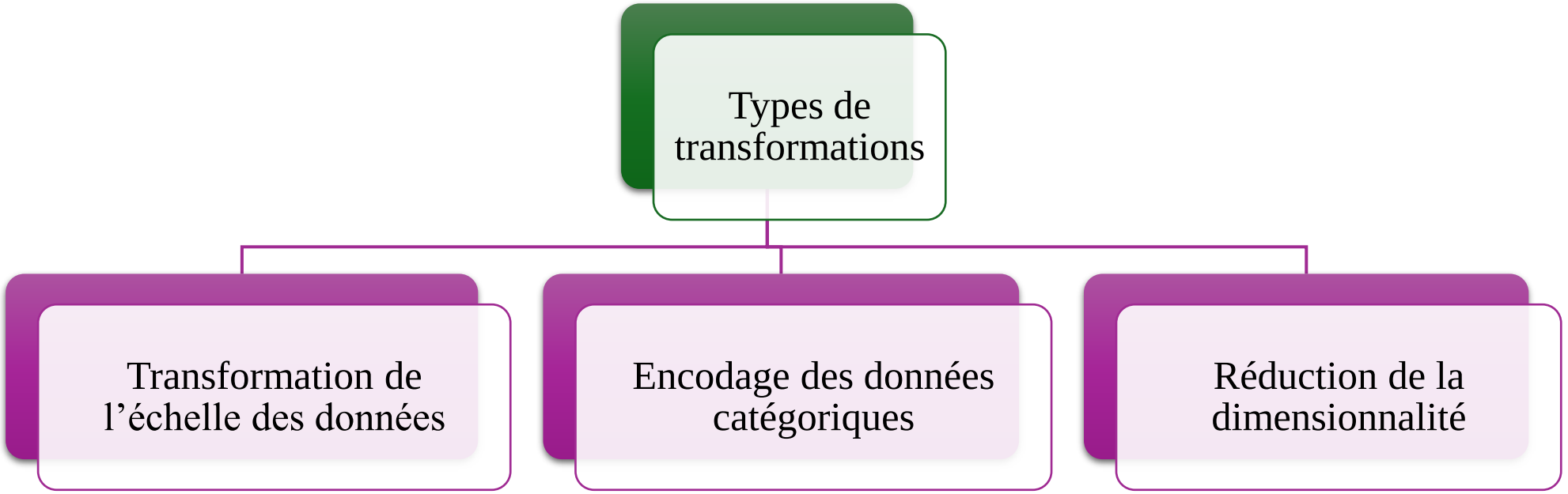
- **Identification des doublons :**
  - Il s'agit d'identifier les lignes identiques ou similaires dans le dataset.
  - Les doublons peuvent être sur l'ensemble des colonnes ou sur une partie spécifique des colonnes.
- **Suppression des doublons :**
  - Une fois identifiés, les doublons peuvent être supprimés en ne conservant qu'une occurrence unique de chaque ligne.
  - Cela permet de réduire la redondance dans les données.

Transformation de données

Transformation de données

- La **transformation des données** consiste à modifier les données brutes pour les rendre plus adaptées aux analyses ou aux algorithmes d'apprentissage.

- Ces transformations peuvent inclure des changements d'échelle, des normalisations, des encodages ou encore des manipulations mathématiques.



Transformation de données

Pourquoi transformer les données ?

Améliorer la performance des modèles

- Certains algorithmes **(ex. : régression linéaire, KNN)** nécessitent que les données soient normalisées pour donner des résultats optimaux.

Réduire la sensibilité aux échelles

- Les variables ayant des échelles différentes **(ex. : revenus en millions et âge en années)** peuvent biaiser les modèles basés sur les distances.

Rendre les données exploitables

- Les données catégoriques ou textuelles doivent être encodées pour être utilisées dans des modèles numériques.

Améliorer l'interprétabilité

- Certaines transformations, comme les logarithmes, permettent de mieux comprendre les relations entre les variables.

Transformation de données
Transformation de l'échelle des données

Standardisation:

- La standardisation est une technique de transformation des données utilisée pour mettre toutes les variables sur une même échelle. Elle consiste à centrer et réduire les caractéristiques, de sorte qu'elles aient :
  - **Une moyenne égale à 0.**
  - **Un écart-type égal à 1,** *avec l'écart-type indique à quel point les valeurs sont éloignées de la moyenne***.**

- Cela permet d'assurer que toutes les caractéristiques contribuent de manière équivalente aux calculs des modèles de machine learning, indépendamment de leur unité ou de leur amplitude initiale.

Formule :

Pour une caractéristique x, la standardisation est donnée par la formule : $$Z = \frac{x - \mu}{\sigma}$$

Où :

x : valeur originale de la caractéristique.

μ : moyenne des valeurs de la caractéristique.

σ : écart-type des valeurs de la caractéristique.

Transformation de données

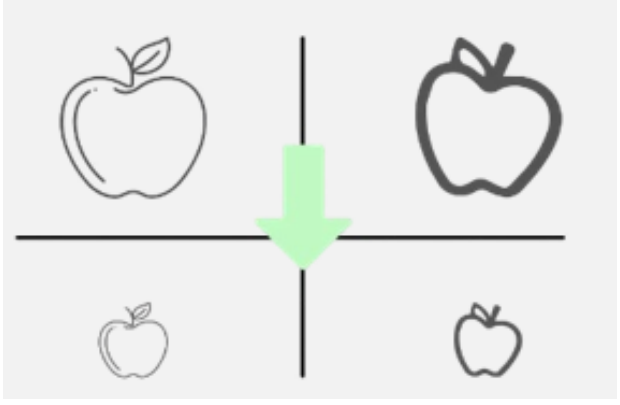
Transformation de l’échelle des données

Normalisation

- La **normalisation** est une technique de transformation des données qui consiste à redimensionner les valeurs d’une caractéristique pour les faire entrer dans une plage spécifique, souvent entre **0 et 1** ou entre **-1 et 1**.

- Elle est utilisée pour uniformiser les échelles des différentes variables tout en conservant les proportions relatives entre les valeurs.

- L'objectif est de rendre les variables comparables en ajustant leur échelle, ce qui est particulièrement important pour les algorithmes **de machine learning** qui sont sensibles à l'échelle des données.



Transformation de données

Transformation de l'échelle des données

Méthodes de normalisation:

Normalisation Min-Max

L'idée est de ramener toutes les valeurs de la variable dans l'intervalle [0;1], tout en conservant le rapport des distances entre les valeurs.

$$X_{\text{normalisé}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Où :
x : valeur d'une caractéristique.
min(x) : valeur minimale de la caractéristique.
max(x) : valeur maximale de la caractéristique.

Maximum Absolute Scaling

Cette méthode ramène chaque donnée dans la plage [−1,1][−1,1], tout en respectant les signes des données originales (les valeurs positives restent positives et les négatives restent négatives).

$$X_{\text{normalisé}} = \frac{X}{\max(|X|)}$$

Où :
X est la valeur d'origine de la donnée,
max(|X|) est la valeur absolue maximale parmi toutes les valeurs de la variable.s

Autres méthodes : Decimal scaling, Z-score normalization...

Transformation de données
Encodage

Variables quantitatives (numériques):
Représentent des mesures ou des valeurs numériques.
Peut être :

- **Continues** : Prennent des valeurs sur une plage infinie (ex. poids, taille, température).
- **Discrètes** : Prennent des valeurs entières uniquement (ex. nombre d'enfants, nombre de voitures).

Exemples :

- Âge : 25, 30.5
- Salaire : 45,000 €

Variables qualitatives (catégoriques):
Représentent des groupes ou des catégories.
Peut être :

- **Nominales** : Les catégories n'ont pas d'ordre (ex. couleur des yeux : bleu, vert, marron).
- **Ordinales** : Les catégories ont un ordre ou un rang (ex. niveau d'éducation : primaire, secondaire, supérieur).

Exemples :

- Ville : Paris, Marseille, Lyon
- Taille : Petit, Moyen, Grand

Transformation de données

Encodage

Les algorithmes d'apprentissage automatique fonctionnent principalement **avec des valeurs numériques**.

Les variables catégorielles, comme les noms ou les catégories **(par exemple, "Homme" ou "Femme", "Rouge" ou "Bleu")**, doivent donc être converties en un format numérique pour être utilisées par les modèles

Transformation de données
Encodage

Les méthodes d'encodage de données :

- **Label Encoding (Encodage par étiquettes) :**

Cette méthode attribue un entier unique à chaque catégorie.

Exemple:

pour une colonne avec les valeurs ["Faible", "Moyen", "Élevé"], le label encoding peut attribuer :

Faible → 0
Moyen → 1
Élevé → 2

Transformation de données
Encodage

- **One-Hot Encoding (Encodage binaire) :**

Cette méthode crée une colonne binaire pour chaque catégorie unique. Chaque catégorie est représentée par une seule colonne avec la valeur "1" pour indiquer la présence et "0" pour l'absence.

Exemple:

Pour le même exemple, avec les valeurs ["Faible", "Moyen", "Élevé"], cela produirait :

Faible	Moyen	Élevé
1	0	0
0	1	0
0	0	1

Transformation de données
Encodage

- **Binary Encoding (Encodage binaire sur les catégories) :**

Une alternative au one-hot encoding qui réduit le nombre de colonnes. Chaque catégorie est convertie en une représentation binaire, puis chaque bit est stocké dans une colonne séparée.

Exemple:

Pour le même exemple, avec les valeurs ["Faible", "Moyen", "Élevé"], cela produirait :

Faible → 001
Moyen → 010
Élevé → 001

- **Encodage fréquentiel :**

Remplace chaque catégorie par sa fréquence dans les données.

Transformation de données
Réduction de dimension

- La réduction de dimension est une technique cruciale en apprentissage automatique et en analyse de données, surtout lorsque le jeu de données contient un grand nombre de variables ou de caractéristiques.

- Cela permet de diminuer la complexité du modèle, d'améliorer la performance et d'éviter le surapprentissage (overfitting).

Exemple de techniques de réduction de dimension :

✓ PCA (Principal Component Analysis)

✓ LDA (Linear Discriminant Analysis)

Sélection de caractéristiques

Sélection de caractéristiques

Le fait d'avoir des caractéristiques **non pertinentes** dans vos données peut diminuer la **précision** de nombreux modèles.

La **sélection de caractéristiques** est une étape clé de la préparation des données en machine learning.

Elle consiste à identifier les caractéristiques les plus pertinentes pour entraîner un modèle, tout en éliminant celles qui sont inutiles ou redondantes.

Sélection de caractéristiques

Pourquoi faire de la sélection de caractéristiques ?

Améliorer les performances :

- Réduire le surapprentissage (overfitting) en limitant le bruit dans les données.
- Augmenter la précision des modèles en utilisant uniquement les données pertinentes.

Réduire la complexité :

- Simplifier les modèles en diminuant le nombre de variables.
- Réduire les coûts de calcul et de stockage.

Faciliter l'interprétation :

- Identifier les caractéristiques qui influencent le plus les prédictions.
- Aider à comprendre les relations sous-jacentes dans les données.

Sélection de caractéristiques

Types de caractéristiques dans les données:

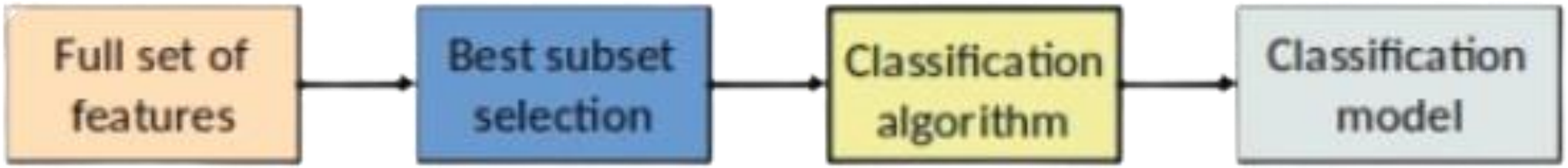
Caractéristiques pertinentes	Caractéristiques redondantes	Caractéristiques inutiles
• Directement corrélées avec la variable cible et utiles pour la prédiction.	• Correlées avec d'autres caractéristiques mais n'apportant pas d'information supplémentaire.	• Non liées à la variable cible et ajoutant du bruit au modèle.

Sélection de caractéristiques

- **Méthodes basées sur les filtres :**

Utilisent des critères statistiques pour évaluer l'importance des caractéristiques (features) de manière **indépendante du modèle** d'apprentissage machine.

Ces méthodes évalue chaque caractéristique indépendamment, en utilisant des critères statistiques ou des tests de corrélation pour déterminer son importance.
Exemples Filtrage basé sur la corrélation (ou corrélation de Pearson)



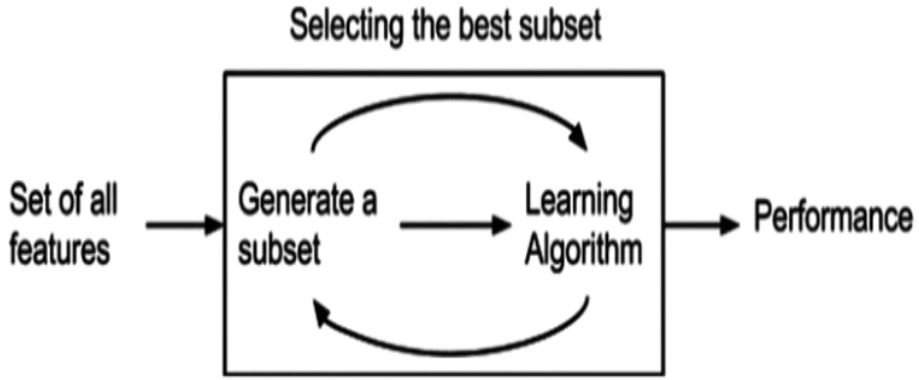
Sélection de caractéristiques

- **Méthodes basées sur les wrappers :** Les **méthodes basées sur les wrappers** sont des techniques de **sélection de caractéristiques** qui évaluent l'importance des caractéristiques en utilisant directement la performance d'un modèle de machine learning.

Contrairement aux méthodes basées sur les filtres, qui utilisent des critères statistiques indépendants du modèle, les méthodes basées sur les wrappers **considèrent la performance du modèle pour choisir les meilleures caractéristiques**.

Exemples:

- **Forward Selection :**
  Commence avec aucune caractéristique, puis ajoute progressivement celles qui améliorent le plus les performances.
- **Backward Elimination :**
  Commence avec toutes les caractéristiques, puis retire celles qui contribuent le moins.
- **Exhaustive Feature Selection :**
  Teste toutes les combinaisons possibles de caractéristiques (coûteux en temps de calcul).



Sélection de caractéristiques

- **Méthodes intégrées (Embedded Methods)**

Ces méthodes intègrent la sélection de caractéristiques dans le processus d'entraînement du modèle.

Exemple:

Régularisation :

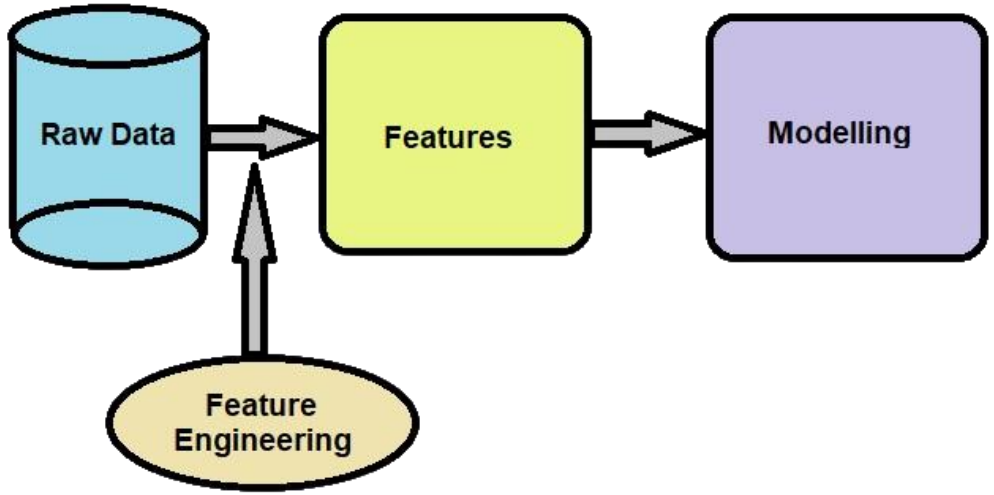
- **Lasso Regression (L1) :** Force certains coefficients de caractéristiques à zéro, éliminant ainsi les moins importantes.
- **Ridge Regression (L2) :** Réduit l'impact des coefficients mais ne les annule pas complètement.

Ingénierie des caractéristiques

Ingénierie des caractéristiques

- L’ingénierie des caractéristiques (ou Feature Engineering) est l'art et la science de transformer des données brutes en caractéristiques (features).

- Cela implique la création, la transformation et la sélection de caractéristiques pertinentes pour maximiser la précision et l’efficacité des modèles d’apprentissage.



Ingénierie des caractéristiques

Exemple de techniques de Feature Engineering :

- **Combinaison des features :** Combiner deux ou plusieurs variables existantes

  (ex. : Prix Total = Quantité × Prix Unitaire).

- **Transformations mathématiques :** Logarithmes, racines carrées, ou puissances pour ajuster les distributions.

- **Extraction des features à partir d'un texte:**

  Extraction de mots-clés, longueur des phrases, ou fréquence des termes.
  Topic extraction: extraire les sujets principaux à partir d'un texte

- **Extraction des features à partir d'une image:** Variance, Ecart type….