Problèmatique du Prétraitement des Données et de la Discrétisation

I. Le prétraitement

Les données brutes sont rarement prêtes à être utilisées telles quelles.
Il est necessaire de réaliser des étapes de prétraitement pour :
- Nettoyer les données (valeurs manquantes, doublons, erreurs).
- Standardiser les formats (nombres, catégories, dates).
- Créer ou modifier des variables pour mieux capturer les relations utiles.
- Adapter les données aux exigences de certains algorithmes.

II. Nettoyage des données

a) Suppression de colonnes inutiles :
drop(colomns=)
Objectif : supprimer des colonnes peu informatives ou trop vides.
Risque : perte d'information utile si suppression non justifiée.

b) Suppression des lignes avec valeurs manquantes :
dropna(inplace=True)
Objectif : garantir l'intégrité des données.
Limite : peut introduire un biais si les valeurs manquantes ne sont pas aléatoires.

III. Imputation des valeurs manquantes

a) Imputation numérique avec la moyenne :
numeric_imputer = SimpleImputer(strategy='mean')

b) Imputation catégorielle :
categorical_imputer = SimpleImputer(strategy='most_frequent')

c) Imputation robuste avec la médiane :
imputer = SimpleImputer(strategy='median')

IV. Feature Engineering

Définition : processus de création ou transformation de variables pour améliorer la performance du modèle.
Exemples :
iris_df['sepal_area_cm2'] = iris_df['sepal_length'] * iris_df['sepal_width']
titanic_df['family_size'] = titanic_df['SibSp'] + titanic_df['Parch'] + 1
Objectif : ajouter des relations ou du sens métier non visibles dans les variables d'origine.

V. Analyse statistique exploratoire

Permet de comprendre la distribution des variables et détecter des valeurs aberrantes.
Exemples de fonctions utilisées : describe(), histplot(), value_counts()

VI. Discrétisation des données

Définition : transformer une variable continue en variable discrète (catégorielle).
Objectifs :
- Faciliter l'apprentissage
- Réduire le bruit
- Capturer des effets de seuil
- Rendre les données interprétables

VII. Méthodes de discrétisation

Avec KBinsDiscretizer :
discretizer = KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')
Paramètres :
- n_bins=3 : nombre de classes
- encode='ordinal' : codage en 0, 1, 2
- strategy='uniform' : intervalles de taille égale

VIII. Limites et précautions

Ne pas discrétiser systématiquement. Évaluer les effets sur la performance et l'interprétabilité.

Risque de perte d’information ou de sursimplification.