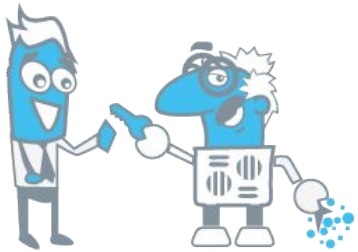
TRAITEMENT DES VALEURS MANQUANTES

Data preprocessing



HUPI

VALORISER ANALYTIQUEMENT VOTRE DIFFÉRENCE

INDEX

1 — 2 — 3

Introduction

Conséquences

Impact des données manquantes sur les analyses

Techniques

Techniques de traitement des valeurs manquantes

01 INTRODUCTION

1

« Quelles peuvent être les raisons de valeurs manquantes ? »

1 Définition

Une valeur manquante représente une absence d'observation pour une variable donnée dans un jeu de données.

Elles peuvent apparaître pour plusieurs raisons :

- Erreurs de collecte ou de saisie de données.
- Problèmes techniques lors des mesures.
- Refus de réponse dans des enquêtes.

2 Types de données manquantes

1. **MCAR (Missing Completely At Random)** :
   - Les données manquent de manière totalement aléatoire.
   - *Exemple* : Une panne aléatoire d'un appareil de mesure.

1. **MAR (Missing At Random) :**
   - Les données manquantes dépendent d'autres variables observées, mais pas de la variable elle-même.
   - *Exemple* : Les personnes plus âgées répondent moins souvent à certaines questions dans un sondage.

1. **MNAR (Missing Not At Random) :**
   - Les données manquantes dépendent de la variable elle-même.
   - *Exemple* : Des revenus élevés non déclarés dans un questionnaire.

02 CONSÉQUENCES

1 Impact sur l'analyse

Visualisation biaisée :

- La distribution des données peut sembler différente lorsque des valeurs manquent.
- *Exemple* : En étudiant la taille d'une population, l'absence de valeurs extrêmes (très grands ou petits) peut fausser la visualisation.

Corrélation biaisée avec suppression des données :

- Supprimer les lignes contenant des valeurs manquantes peut fausser les corrélations si les données ne sont pas manquantes aléatoirement.
- Cela réduit la taille de l'échantillon et diminue la puissance statistique.

Réduction de la puissance de l'analyse :

- Moins de données signifie des estimations moins précises et des intervalles de confiance plus larges.
- Si les données ne sont pas remplacées correctement, le modèle risque de perdre en robustesse.

03 TECHNIQUES COURANTES

1

« Quelles techniques connaissez-vous ? »

1 Suppression des valeurs manquantes **(dropna *Python*)**

Méthode :

suppression des lignes

Avantages/Inconvénients :

- Simple,
- mais peut entraîner une perte d'informations importante.

Quand l'utiliser ?

Lorsque le nombre de valeurs manquantes est faible.

2 Imputation des valeurs manquantes **(fillna *Python*)**

Méthode : Imputation/remplacement

- **par une constante,**
- **par la moyenne,**
- **par la médiane,**
- **par la donnée la plus représentée, ...**

Avantages/Inconvénients :

- Simple à appliquer
- mais, ne prend pas en compte la variance réelle des données.

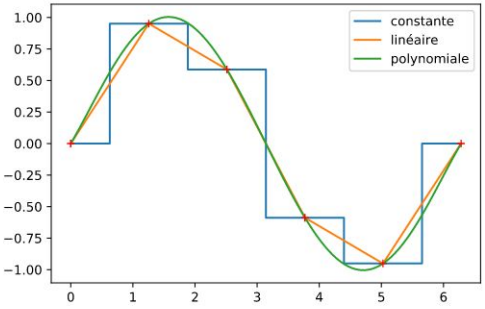
Quand l'utiliser ?

- Lorsque les données manquantes ne sont pas nombreuses :
  - Si le pourcentage de valeurs manquantes est faible (par exemple, inférieur à 5-10%), l'imputation est souvent une meilleure alternative que la suppression.

3 Imputation par méthode avancée : Interpolation **(interpolate *Python*)**

Méthode : à partir d'un nombre fini de points reconstruire une fonction :

- **linéaire,**
- **polynomiale,**
- **spline (une fonction polynomiale par morceaux),**
- **cubic (cas particulier de polynôme d'ordre 3), …**



Utilise les tendances pour combler les lacunes (par exemple, dans des séries temporelles).

Avantages/Inconvénients :

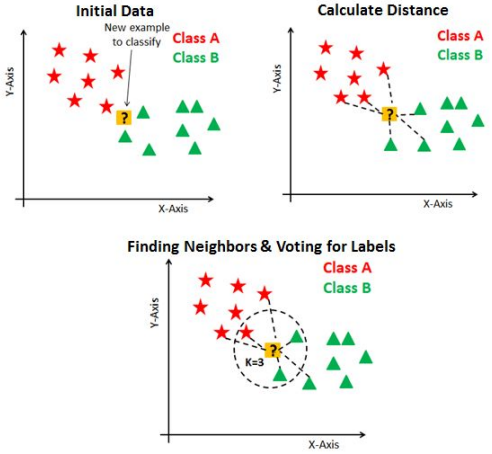
- Préserve la tendance des données et permet de gagner en précision
- mais, ne fonctionne pas bien avec de grandes plages de données manquantes.

Quand l'utiliser ? L'interpolation est idéale pour les séries temporelles ou toute donnée qui possède une structure ordonnée.

4 Imputation par méthode avancée : Imputation par modèle (KNN, régresseurs, etc.) **(KNNImputer *Python*)**

Méthode : utilise des algorithmes pour prédire les valeurs manquantes

- **K-nearest neighbors (KNN)**
- **modèle prédictif (régression, arbre de décision, …)**



Avantages/Inconvénients :

- Prend en compte les relations entre les variables.
- mais, plus complexe et nécessite davantage de calcul.

Quand l'utiliser ?

- Lorsque les relations entre les variables sont complexes et que les données ne suivent pas des tendances simples.
- Lorsque les données contiennent des relations non linéaires.
- Lorsque les données sont modérées en taille (car KNN est coûteux en termes de calcul pour les grands jeux de données).

.

CONCLUSIONS

Bonnes pratiques :

- Identifier et comprendre les valeurs manquantes avant de les traiter.
- Imputer ou supprimer les valeurs manquantes selon leur impact sur l’analyse.
- Choisir la méthode d’imputation la plus appropriée (moyenne, médiane, modèles prédictifs, etc.).
- Vérifier l'impact du traitement des valeurs manquantes sur la qualité des données et des modèles.
- Toujours tester les performances du modèle avant et après le traitement des valeurs manquantes pour éviter tout biais.