IMSP

TD : Premiers pas

Les datasets publics sont disponible sur **Kaggle** ou via **sklearn.datasets**

Exercice1 :

Téléchargez un dataset au format CSV et effectuez une première exploration.

Tâches :

1. Chargez le dataset avec pandas.
2. Affichez le nombre de lignes et de colonnes.
3. Affichez un échantillon de 10 lignes du dataset.
4. Listez les noms des colonnes et leurs types de données.
5. Affichez un résumé statistique des colonnes numériques.
6. Identifiez les colonnes qui contiennent des valeurs manquantes.
7. Calculez le pourcentage de valeurs manquantes par colonne.
8. Supprimez les colonnes ayant plus de 50% de valeurs manquantes.
9. Remplacez les valeurs manquantes des colonnes numériques par leur médiane.
10. Remplacez les valeurs manquantes des colonnes catégorielles par la modalité la plus fréquente.
11. Détectez le nombre de lignes dupliquées.
12. Supprimez les doublons et affichez la nouvelle dimension du dataset.
13. Affichez un boxplot pour visualiser les outliers d'une colonne numérique.
14. Utilisez la méthode de l'IQR pour identifier les outliers.
15. Supprimez les valeurs aberrantes détectées.
16. Convertissez une colonne date en type datetime et extrayez l'année, le mois et le jour.
17. Effectuez un one-hot encoding sur une colonne catégorielle.
18. Normalisez une colonne numérique entre 0 et 1.
19. Créez une nouvelle colonne indiquant si une personne est majeure ou non (basé sur une colonne âge).
20. Catégorisez une colonne continue en intervalles (ex : âge en jeunes/adultes/seniors).
21. Créez une variable binaire indiquant si une personne a un salaire supérieur à la médiane.

Exercice 2 Fusion et jointure de deux datasets

1. Chargez deux datasets(csv) et affichez leur taille respective.
2. Effectuez une jointure sur une colonne commune.
3. Ajoutez de nouvelles colonnes à un dataset en fusionnant les informations d'un second fichier.