Examen Intra

1. Information du dataset

Nom du dataset : student-mat1.csv

Contexte :

Les données proviennent d'une enquête menée auprès d'élèves de mathématiques du secondaire. Elles contiennent de nombreuses informations intéressantes sur les aspects sociaux, le genre et les études des élèves. Elles peuvent être utilisées pour évaluer l'AED ou **pour prédire la note finale des élèves.**

Description :

1. school - établissement scolaire de l'élève (binaire: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - sexe de l'élève (binaire: 'F' - féminin ou 'M' - masculin)
3. age - âge de l'élève (numérique: de 15 à 22)
4. Mjob - Emploi de la mère (nominal : « enseignante », « soins de santé », « services » civils (ex. : administratif ou policier), « à domicile » ou « autre »)
5. Fjob - Emploi du père (nominal : « enseignante », « soins de santé », « services » civils (ex. : administratif ou policier), « à domicile » ou « autre »)
6. studytime - temps d'étude hebdomadaire (numérique : 1 - < 2

heures, 2 - 2 à 5 heures, 3 - 5 à 10 heures, ou 4 - >10 heures)

7. échecs - nombre d'échecs passés (numérique : n si 1<=n<3, sinon 4)
8. absences - nombre d'absences scolaires (numérique : de 0 à 93)

Ces notes sont liées à la matière du cours mathématiques :

1. G1 - note de la première période (numérique : de 0 à 20)
2. G2 - note de la deuxième période (numérique : de 0 à 20)
3. G3 - note finale (numérique : de 0 à 20, **Variable cible**)

2. Importer les bibliothèques nécessaires

Importez les bibliothèques nécessaires pour le projet :
- **Pandas :** pour la manipulation des données.
- **Matplotlib** : pour créer des graphiques.

3. Charger et explorer le dataset

a. Informations générales sur le dataset : Chargez les données à partir du fichier CSV, affichez les 10 lignes et examinez les types de données.

b. Statistiques descriptives : Obtenez des statistiques descriptives pour les colonnes numériques (moyenne, écart-type,

valeurs minimales, maximales, etc.), ensuite sur les colonnes catégorielles

c. Grouper par une catégorie : Groupez les données par la colonne `school` (établissement scolaire de l'élève) et calculez des statistiques pour chaque groupe, telles que la moyenne des notes `G3` et du nombre d'absences.

d. Trier les groupes : Triez les groupes selon la variable `age` pour mieux comparer les statistiques des variables `G1`, `G2`, et `G3` entre les groupes.

4. Prétraiter les données

a. **Vérification des valeurs manquantes** : Vérifiez les valeurs manquantes et traitez-les en les remplissant par la moyenne, le mode ou en les supprimant si nécessaire.
b. **Vérification des valeurs aberrantes** : Identifiez les valeurs aberrantes dans les colonnes numériques, Traitez ces valeurs en les remplaçant par la moyenne, la médiane, ou en les supprimant selon le cas.
c. **Encodage des variables catégorielles** : Pour les variables catégorielles telles que `school`, `sex`, `Mjob`, et `Fjob`, appliquez un encodage approprié (par exemple, encodage One-Hot ou Label Encoding).

5. Créer des graphiques avec Matplotlib

a. **Histogramme** : Réalisez un histogramme pour visualiser la distribution de l'âge des élèves.

b. **Diagramme à barres** : Créez un diagramme à barres pour représenter le nombre d'élèves en fonction de leur sexe (`sex`).

c. **Diagramme circulaire (Pie Chart)** : Construisez un diagramme circulaire pour montrer la répartition des élèves selon leur lieu de résidence (`address`).

d. **Diagramme à moustaches (Boxplot)** : Tracez un boxplot pour observer la distribution des notes finales (`G3`).

6. Graphique en nuage de points (Scatter Plot)

Réalisez un nuage de points pour visualiser la relation entre l'âge des élèves (`age`) et leurs notes finales (`G3`).

7. Graphique en lignes

Créez un graphique en lignes pour visualiser la tendance d'une variable, par exemple l'évolution des notes moyennes (`G3`) en fonction de l'âge (`age`).