BADJI MOKHTAR ANNABA UNIVERSITY
FACULTY OF TECHNOLOGY
Common Core Computer Engineer
3rd year Data science.

جامعة باجي مختار عنابة
كلية التكنولوجيا

Nom de l'enseignant : Bilal Dendani

Examen de Fondement des sciences des données

Durée : 1h : 30

Semestre 1 – Année Universitaire 2024/2025

Nom : Étudiant SD

Prénom : ING INF

Exercice 1 (6.5 points) : Sélectionnez toutes les bonnes réponses (une ou plusieurs réponses) :

1. Parmi les éléments suivants, lesquels constituent un exemple de données non structurées 0.25pt?

   A. Numéro de matricule de la voiture, marque, modèle
   B. Vidéos
   C. Fichiers audio
   D. Les deux B et C

2. Le machine learning ou apprentissage automatique est une sous-discipline …0.25pt

   A. De la science de données (DataScience)
   B. Du Deep Learning (apprentissage profond)
   C. De la robotique
   D. De l'intelligence artificielle

3. Quelles sont les étapes indispensables du processus de science des données 0.5pt ?

   A. Collecte et exploration des données.
   B. Visualisation et interprétation des résultats.
   C. Ajout de bruit dans les données pour tester la robustesse.
   D. Nettoyage et transformation des données.

4. Parmi les outils suivants, lesquels sont spécifiquement utilisés pour l'analyse des données 0.75pt ?

   A. Jupyter Notebook.
   B. TensorFlow.
   C. SQL.
   D. Tableau.

5. Le Deep learning ou apprentissage profonde, est une sous-discipline…0.25pt

A. Du Big Data
B. De l'apprentissage intelligent
C. De la science de données (Data-science)
D. Du machine learning (apprentissage automatique)

6. Qu'est-ce qu'une distribution symétrique de données ? 0.5pt
   A. Une distribution où les deux côtés du graphique sont des images miroir l'un de l'autre autour du centre.
   B. Une distribution qui est toujours de forme normale.
   C. Une distribution qui ne contient pas de valeurs extrêmes (outliers).
   D. Une distribution où les données sont uniformément réparties sur toute la plage des valeurs.
   E. Les données à gauche et à droite de la moyenne sont réparties de manière égale.

7. Lors d'une analyse exploratoire de données (EDA), quelles étapes sont cruciales pour détecter des valeurs aberrantes 0.75pt?
   A. Analyse de la matrice de corrélation.
   B. Standardisation des données.
   C. Comparaison des distributions avec des graphiques (histogrammes, KDE).
   D. Calcul des quartiles et de l'IQR (Interquartile Range).
   E. Utilisation de diagrammes en boîte (boxplot).

8. Quelle est l'importance de la matrice de corrélation dans une analyse exploratoire 0.5pt?
   A. Identifier les relations linéaires entre les variables.
   B. Détecter les variables catégorielles corrélées.
   C. Découvrir des relations non linéaires entre les variables.
   D. Évaluer la force des relations entre les variables numériques.

9. Quelles propriétés caractérisent un espace vectoriel 0.5pt?
   A. La somme de deux vecteurs peut ne pas appartenir à l'espace vectoriel.
   B. Il contient le vecteur nul.
   C. Tout sous-ensemble d'un espace vectoriel est nécessairement un espace vectoriel.
   D. Il doit satisfaire les propriétés associatives et distributives pour l'addition et la multiplication scalaire.

10. Un responsable des ventes dans une grande concession automobile souhaite déterminer les quatre modèles les plus vendus en se basant sur les données de ventes des deux dernières années. Quel est le type de graphique approprié pour cet objectif 0.25pt?

    A. Graphique en nuage de points
    B. Graphique en courbes
    C. Graphique circulaire
    D. Graphique en barres

11. Quelle hypothèse est faite dans la régression linéaire simple ? 0.25pt

A. Les variables sont indépendantes entre elles.
B. La relation entre les variables est linéaire.
C. La distribution des données est uniforme.
D. Toutes les valeurs manquantes sont remplies par des zéros.

12. Quel est l'objectif principal de la régression linéaire simple ? 0.25pt

A. Créer des histogrammes de données.
B. Prédire une variable cible en fonction d'une variable indépendante.
C. Classer les données en groupes distincts.
D. Stocker des données dans des bases relationnelles.

13. Quelle formule est utilisée pour la régression linéaire simple ? 0.5pt
A. $y = mx+b$
B. $y = ax2+bx+c$
C. $y = \beta 0+\beta 1x$
D. $y = x+b$

14. Dans les environnements Big Data, la vélocité désigne ? 0.5pt
A. Les données peuvent arriver à grande vitesse
B. Des ensembles de données énormes peuvent s'accumuler dans des périodes très courtes.
C. La vélocité des données se traduit par le temps qu'il faut pour que les données soient traitées.
D. Toutes les réponses ci-dessus

15. Quelle est la différence fondamentale entre un modèle supervisé et non supervisé ? 0.5pt

A. Un modèle supervisé nécessite des données étiquetées, tandis qu'un modèle non supervisé n'en nécessite pas.
B. Un modèle supervisé ne peut pas être utilisé pour la classification.
C. Un modèle non supervisé est utilisé pour prédire une variable dépendante.
D. Les deux modèles utilisent toujours des algorithmes linéaires

Exercice 2 : (6.5 points)

1. Décrivez les trois niveaux de structuration des données. Donner un exemple pour chaque niveau ?
2. Quelle est la différence entre big data et la data science ?

3. Vous travaillez sur un projet de prédiction des prix de l'immobilier. Expliquez comment vous structureriez chaque étape du processus de science des données pour ce projet. Quels défis spécifiques pourriez-vous rencontrer à chaque étape ?
4. Pourquoi la phase de nettoyage des données est -elle cruciale dans un projet de science des données ?
5. Expliquez pourquoi l'analyse exploratoire des données (EDA) est une étape essentielle du flux de travail de la science des données. Énumérez et décrivez trois informations que l'EDA peut révéler sur un ensemble de données avant de créer un modèle.
6. Expliquez l'importance des matrices dans le traitement des données. Donnez un exemple d'application dans la science des données**.**
7. Comparez les bibliothèques Pandas et NumPy en mentionnant leur rôle dans l'analyse de données**.**

……
……
……
……
……
……
……
……
……
……
……Voir dans le fichier Exercice 2……..
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

Exercice 03 : (07 points)

Une entreprise vous a fourni un jeu de données contenant des informations sur ses employés, avec les colonnes suivantes.

- `Employee_ID` : Identifiant unique de l'employé.
- `Department` : Département de l'employé.
- `Years_of_Experience` : Nombre d'années d'expérience de l'employé.
- `Work_Life_Balance` : Évaluation (1 à 5) de l'équilibre travail-vie personnelle.
- `Performance_Score` : Score de performance de l'employé (0 à 100).
- `Salary` : Salaire annuel de l'employé en dollars.

Le dataset contient des valeurs manquantes et des anomalies. Votre objectif est d'analyser ces données pour fournir des recommandations stratégiques.

1. Complétez le code suivant pour gérer les valeurs manquantes de manière appropriée. (1 pt)
2. Pourquoi est-il préférable d'utiliser la médiane pour imputer la colonne `Work_Life_Balance` plutôt que la moyenne ?

Réponse : Il est préférable d'utiliser la médiane pour `Work_Life_Balance`, car cette colonne est une évaluation discrète (échelle de 1 à 5). La médiane est robuste aux valeurs extrêmes et représente mieux la tendance centrale dans ce type de données catégoriques. (1 pt)

```
# Charger le dataset
import pandas as pd
df = pd.read_csv("employee_data.csv")

# Afficher les premières lignes du dataset
print(df.head())

# Identifier les valeurs manquantes
print("Valeurs manquantes par colonne :")
print(df.isnull().sum()) # indetification des valeurs manquantes (0.25pt)

# Compléter : Remplacer les valeurs manquantes de 'Performance_Score' par la moyenne
df['Performance_Score'].fillna(df['Performance_Score'].mean(), inplace=True)  (0.25pt)

# Compléter : Remplacer les valeurs manquantes de 'Work_Life_Balance' par la médiane
df['Work_Life_Balance'].fillna(df['Work_Life_Balance'].median(), inplace=True) (0.25pt)

# Compléter : Supprimer les lignes où 'Department' ou 'Salary' est manquant
df.dropna(subset=[ 'Department', 'Salary' ], inplace=True)  (0.25pt)

# Vérification finale
print("Données après traitement des valeurs manquantes :")
```

3. Complétez le code suivant pour détecter et gérer les anomalies dans les données ?

```
# Compléter : Détection des anomalies dans la colonne 'Salary' (salaires supérieurs à 200,000 ou inférieurs à 20,000)
anomalies = df[(df['Salary'] > 200000) | (df['Salary'] < 20000)]
                              (0.25pt)                (0.25pt)
# Afficher les anomalies détectées
print("Anomalies détectées :")
print(anomalies)

# Compléter : Remplacer les anomalies de 'Salary' par la médiane des salaires
median_salary = df['Salary'].median()
df.loc[(df['Salary'] > 200000) | (df['Salary'] < 20000), 'Salary'] = median_salary  (0.25pt)

# Vérification finale
print("Données après correction des anomalies :")
print(df['Salary'].describe())
```

4. Quels sont les risques pour les résultats d'analyse si les anomalies ne sont pas traitées correctement ?
   Si les anomalies ne sont pas traitées, elles peuvent biaiser les analyses statistiques ou les modèles de machine learning, conduisant à des résultats incorrects ou peu fiables. (1 pt)

5. Complétez le code suivant pour explorer la relation entre l'expérience et la performance ?

```
import matplotlib.pyplot as plt
plt.scatter (0.25pt) (df['Years_of_Experience'] (0.25pt) , df['Performance_Score'] (0.25pt) , alpha=0.7)

plt.title("Relation entre l'expérience et la performance")
plt.xlabel("Years of Experience")
plt.ylabel("Performance Score")
plt.show()
```

6. Quelle tendance pourrait être observée dans le graphique ?

   Une relation positive pourrait être observée, montrant que l'expérience améliore la performance. Cependant, des points de saturation pourraient apparaître (au-delà d'un certain nombre d'années). (1pt)

7. Si une relation positive est identifiée, comment cela pourrait-il influencer les politiques de recrutement de l'entreprise? (1.5pt)
   - Si une relation positive est, cela pourrait influencer les politiques de recrutement de l'entreprise par favoriser les employés expérimentés, mais
   - Investir également dans la formation continue pour améliorer les performances des employés.