**Examen Data Science Part 1 :**

**Théorie et pratiques**

Aucune question ne pourra être posée durant l'examen.

En cas de doute concernant le sujet, vous poursuivrez votre réponse en expliquant vos hypothèses.

Durée : 1h30

Épreuve du 05/12/2024

# Modalités du travail

* Durée : 1h30 ;
* Aucun document autorisé, calculatrice non autorisée ;
* Ecrire vos réponses sur la copie, dans les cases réservées à cet effet ;
* Mettre vos noms et prénoms sur chaque feuille ;
* ***Toute réponse donnée sans explications sera considérée comme incorrecte***
* ***Tout texte indéchiffrable sera considéré comme une absence de réponse***

**Part 1: Intro to Data Science (5 points)**

1. **Qu'est-ce que la science des données, et quel est son rôle dans le domaine de l'informatique ? (0.25 pt)**

|  |
| --- |
|  |

1. **Décrivez les étapes principales du processus de la science des données, de la collecte des données à la prise de décision. (0.25 pt)**

|  |
| --- |
|  |

1. **On utilise les bibliothèques NumPy et Pandas pour quelles raisons ? (0.25 pt)**

|  |
| --- |
|  |

1. **Quel est le résultat de la sortie et expliquez comment il a été obtenu. (0.5 pt)**

**import numpy as np**

**arr = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])**

**print(arr[1:, ::2])**

|  |
| --- |
|  |

1. **Quel est le résultat de la sortie et expliquez comment il a été obtenu. (0.5 pt)  
    import pandas as pd   
    data = { 'A': [10, 20, 30, 40, 50], 'B': [1, 2, 3, 4, 5], 'C': ['x', 'y', 'z', 'w', 'v']}  
    df = pd.DataFrame(data)  
    df.loc[df['B'] > 2, ['A', 'C']]**

|  |
| --- |
|  |

1. **Quel est le résultat de la sortie et expliquez comment il a été obtenu (0.5 pt)**

**import pandas as pd**

**df = pd.DataFrame({**

**'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],**

**'Age': [25, 30, 35, 40, 45],**

**'Salary': [50000, 60000, 75000, 90000, 100000]**

**})**

**result = df.iloc[1:-1, [0, 2]]**

|  |
| --- |
|  |

1. **Que fait la méthode df.dropna() dans Pandas ? Quelle est la différence entre les paramètres axis=0 et axis=1 ? (0.5 p)**

|  |
| --- |
|  |

1. **Which Pandas method provides a statistical summary of numerical columns? (0.25)**

**a) head()**

**b) tail()**

**c) info()**

**d) describe()**

|  |
| --- |
|  |

1. **Etant donne le dataframe suivant :**

data = {'Product': ['A', 'B', 'C', 'D', 'E'],

'Price': [10, 20, 30, 40, 50],

'Stock': [100, 50, 75, 25, 60],

'Sales': [200, 150, 100, 250, 300]}

df = pd.DataFrame(data)

**Quel sera le résultat (0.5 pt):**

output1 = df.iloc[:3, -2:]

output2 = df.iloc[1:4, 1:]

|  |
| --- |
|  |

1. **Considérons le même data frame, comment sélectionner tous les produits avec le prix >= 30 ? (0.25)**

|  |
| --- |
|  |

1. **La différence entre df.loc[] et df.iloc[] dans Pandas? (0.25 p)**

|  |
| --- |
|  |

1. **La différence entre df.isnull().count() et df.isnull().sum()? (0.5 p)**

|  |
| --- |
|  |

1. **df.rename(columns = {‘old\_name’:’new\_name’, inplace = ?), que se passe-t-il si inplace = True et si inplace = False ? (0.5 p)**

|  |
| --- |
|  |

**Part 2: Data Visualization and EDA (5 points)**

1. **Pourquoi la visualisation des données est-elle une étape essentielle dans le processus d'analyse ? (0.5 pt)**

|  |
| --- |
|  |

1. **Vous disposez d'un ensemble de données sur les résultats des élèves aux tests dans quatre matières différentes : mathématiques, sciences, anglais et histoire. Quel type de graphique utiliseriez-vous pour comparer les résultats moyens de ces matières ? Expliquez votre choix. (0.5 pt)**

|  |
| --- |
|  |

1. **Décrivez deux types de graphiques adaptés pour comparer des variables continues et catégorielles. (0.75 pt)**

|  |
| --- |
|  |

1. **Quels graphiques utiliseriez-vous pour visualiser la répartition des âges dans une population ? Pourquoi ? (0.75 pt)**

|  |
| --- |
|  |

1. **Expliquez comment la méthode .info() de Pandas peut aider dans l’analyse d’un DataFrame. (1 pt)**

|  |
| --- |
|  |

1. **Pourquoi l’identification des valeurs aberrantes est-elle essentielle dans le cadre d’une EDA ? (0.5 pts)**

|  |
| --- |
|  |

1. **Expliquez comment détecter les valeurs manquantes dans un DataFrame Pandas. (1 pt)**

|  |
| --- |
|  |

**Part 3: Machine Learning Fundamentals (5 points)**

1. **Expliquez ce qu’est un modèle de classification et donnez deux exemples d’algorithmes couramment utilisés. (0.5 pt)**

|  |
| --- |
|  |

1. **Quel sera le résultat de X\_train, X\_test, y\_train et y\_test ? Expliquez le fonctionnement de train\_test\_split et l'effet du réglage random\_state=42. (0.25 pt)**

**from sklearn.model\_selection import train\_test\_split**

**X = [[1, 2], [3, 4], [5, 6], [7, 8]]**

**y = [0, 1, 0, 1]**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.25, random\_state=42)**

|  |
| --- |
|  |

1. **Quel est l'intérêt de diviser les données en ensembles d'entraînement et de test ? Quel sera le problème si nous utilisons les mêmes données pour l'entraînement et l'évaluation ? (0.5 pt)**

|  |
| --- |
|  |

1. **Complétez le code suivant pour créer un modèle de régression linéaire (0.25 pt)**

**from sklearn.linear\_model import LinearRegression**

**X = [[1], [2], [3], [4]]**

**y = [2, 3, 5, 7]**

**# Initializing an instance of the class LinearRegression**

**?**

**# training**

**?**

**# Prediction**

**?**

|  |
| --- |
|  |

1. **Écrivez le code pour générer la matrice de confusion (avec visualisation). Que représente chaque valeur de la matrice résultante ? (0.75 pt)**

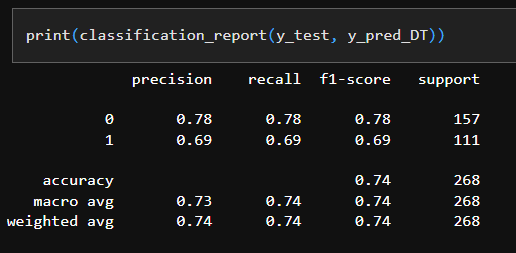
**from sklearn.metrics import confusion\_matrix**

**y\_true = [0, 1, 1, 0, 1, 0]**

**y\_pred = [0, 1, 0, 0, 1, 1]**

|  |
| --- |
|  |

1. **Expliquer les deux premières lignes de la sortie du rapport de classification (Classification report) ? Donnez les formules de Accuracy, Precision, Recall, and F1-Score (1.5 pt)**

****

|  |
| --- |
|  |

1. **Quels critères prendre en compte pour choisir un modèle d’apprentissage supervisé adapté à un problème donné ? (0.5 pt)**

|  |
| --- |
|  |

1. **Dans un ensemble de données avec un déséquilibre de classe, quelle mesure pourrait être trompeuse ? (0.75)**

|  |
| --- |
|  |

**Part 4 : Feature Engineering (5 points)**

1. **Pourquoi est-il nécessaire de normaliser ou standardiser les données avant d'utiliser certains algorithmes d'apprentissage automatique ? (0.75 pt)**

|  |
| --- |
|  |

1. **Expliquez le concept de data leakage avec un exemple. (0.75 pt)**

|  |
| --- |
|  |

1. **Quelle est la différence entre les techniques d'encodage one-hot et label encoding pour les variables catégorielles ? (0.5 p)**

|  |
| --- |
|  |

1. **Avec des features catégorielles à cardinalité élevée, celles qui ont un grand nombre de valeurs uniques, telles que les identifiants de produit, le code postal, quelle est la meilleure approche pour encoder ces features catégorielles ? et Expliquez pourquoi les variables catégorielles à haute cardinalité peuvent poser des problèmes dans les modèles si on utilise one-hot encoding. (0.5 p)**

|  |
| --- |
|  |

1. **Quelles sont les principales méthodes pour gérer les valeurs asymétriques (skewness) dans les données numériques? (0.5 p)**

|  |
| --- |
|  |

1. **Scénario : Vous travaillez sur un projet où vous devez prédire le prix d'une maison en fonction de plusieurs facteurs :**

* **Surface (en mètres carrés) : La superficie de la maison, généralement entre 50 et 300 m².**
* **Nombre de chambres : Un entier, généralement entre 1 et 5 chambres.**
* **Âge de la maison (en années) : Le nombre d'années écoulées depuis la construction de la maison, allant de 0 à 100 ans.**
* **Distance au centre-ville (en kilomètres) : La distance de la maison au centre-ville, généralement entre 5 km et 50 km.**

**Vous décidez d'utiliser la Régression Linéaire pour modéliser ce problème. Cependant, après des tests, vous remarquez que le modèle semble être fortement influencé par des caractéristiques comme la surface et la distance au centre-ville, tandis que le nombre de chambres et l'âge de la maison semblent avoir peu d'impact sur le prix prédit.**

**1.1 Identifiez le problème : Expliquez pourquoi le modèle pourrait privilégier des caractéristiques telles que la surface et la distance au centre-ville par rapport à d'autres comme le nombre de chambres et l'âge de la maison. (1 pt)**

**1.2 Proposez une solution : Suggérez une technique avec une équation pour résoudre ce problème et justifiez brièvement pourquoi elle serait utile dans ce scénario. (1 pt)**

|  |
| --- |
|  |