**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 : 2h

Épreuve du 08/11/2024

# Modalités du travail

* Durée : 2h ;
* 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**

**array = np.array([[10, 15, 20], [25, 30, 35], [40, 45, 50]])**

**output = array[1:, :2]**

|  |
| --- |
|  |

1. **Quelle est la sortie de la sortie et que fait l’opération .T sur le tableau ? (0.25 pt)**

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

**output = array.T**

|  |
| --- |
|  |

1. **Quelle est la valeur de sortie et expliquez la logique derrière le calcul. (0.5 pt)**

**array = np.array([10, 20, 30, 40, 50])**

**output = np.mean(array[array > 25])**

|  |
| --- |
|  |

1. **Au cœur de Pandas se trouvent deux structures de données fondamentales : quelles sont-elles ? La différence entre eux ? (0.25 p)**

|  |
| --- |
|  |

1. **Décrivez ces méthodes d'exploration : head(), tail(), info(), describe() ? (0.5 p)**

|  |
| --- |
|  |

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 ? (0.5)**

|  |
| --- |
|  |

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. **Quels sont les types de données ? (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. **Vous disposez d'un ensemble de données contenant des variables catégorielles et continues et vous souhaitez analyser les différentes variables et analyser aussi les relations entre elles. Décrivez trois types de visualisations adaptées à l'exploration de ces relations et expliquez les scénarios dans lesquels chaque type serait le plus efficace. (1 pt)**

|  |
| --- |
|  |

1. **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. (0.5 pt)**

|  |
| --- |
|  |

1. **Les valeurs aberrantes peuvent avoir une incidence significative sur les résultats d'un modèle. Qu'est-ce qu'une valeur aberrante ? Expliquez deux techniques de visualisation qui permettent d'identifier les valeurs aberrantes dans un ensemble de données et décrivez comment vous les interpréteriez dans chaque graphique. (0.5 pt)**

|  |
| --- |
|  |

1. **On vous donne deux variables numériques, « Âge du client » et « Revenu annuel », et vous souhaitez déterminer s'il existe une relation entre elles. Décrivez la visualisation que vous utiliseriez, comment elle pourrait être utilisée à cette fin et quelles informations supplémentaires pourraient être obtenues en calculant le coefficient de corrélation. (Donnez le nom du coefficient de corrélation + formule) (1.5 pts)**

|  |
| --- |
|  |

1. **Quel est l’inconvénient de la corrélation de Pearson ? (0.5 p)**

|  |
| --- |
|  |

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

1. **La tâche la plus importante d'un data scientist est de sélectionner un modèle approprié pour résoudre un problème. Quelles sont les considérations à prendre en compte lors du choix d'un modèle d'apprentissage supervisé ? (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. (0.75 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 ? (0.75 pt)**

|  |
| --- |
|  |

1. **Complétez le code suivant pour créer un modèle de régression linéaire (0.5 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 ? (0.5 pt)**

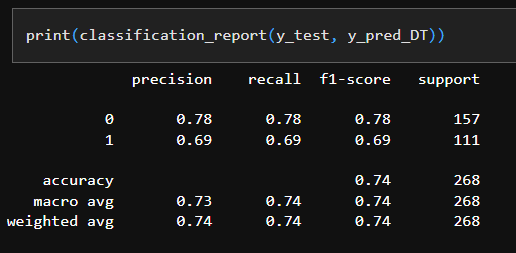
**from sklearn.metrics import confusion\_matrix**

**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) ? (1.5 pt)**

****

|  |
| --- |
|  |

1. **Quelle était la limitation du perceptron de Rosenblatt ? (0.5 pt)**

|  |
| --- |
|  |

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

1. **Comment remplir les valeurs manquantes pour différents types de données ? (0.75 pt)**

|  |
| --- |
|  |

1. **Qu'est-ce qu'une distribution asymétrique (skewed distribution) ? et pourquoi c'est un problème ? (0.75 pt)**

|  |
| --- |
|  |

1. **C’est quoi l’importance d’encoder 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 ? (0.5 p)**

|  |
| --- |
|  |

1. **C’est quoi le data leakage ? (0.5 p)**

|  |
| --- |
|  |

1. **Vous travaillez sur un projet de classification pour une application de covoiturage, où vous cherchez à prédire si un client acceptera une suggestion de course en fonction de divers facteurs, notamment :**

* **Distance (en kilomètres) : la distance entre le point de prise en charge et l'emplacement actuel du client, généralement comprise entre 0 et 50 km.**
* **Estimation du tarif (en dollars) : le tarif estimé de la course, qui peut varier de 5 $ à 100 $.**
* **Note du conducteur (de 1 à 5) : la note moyenne du conducteur, souvent comprise entre 3,0 et 5,0.**
* **Nombre de courses précédentes (nombre entier) : le nombre de courses effectuées par le client sur la plateforme, compris entre 0 et 200.**

**Vous décidez d'utiliser les k-Nearest Neighbors (k-NN) comme modèle pour cette tâche, car il fonctionne bien dans les scénarios où les distances et les similitudes entre les points de données sont cruciales. Cependant, après les premiers tests, vous remarquez que le modèle semble être biaisé en faveur des fonctionnalités avec des plages numériques plus larges, comme la distance et l'estimation du tarif, par rapport à l'évaluation du conducteur et au nombre de trajets précédents.**

* 1. **Identifiez le problème : expliquez pourquoi le modèle pourrait privilégier des fonctionnalités telles que la distance et l'estimation du tarif par rapport à d'autres dans ce cas. (1 pt)**
  2. **Proposez une solution : suggérez une technique avec l'équation pour résoudre ce problème et justifiez brièvement pourquoi elle serait utile dans ce scénario. (1 pt)**

|  |
| --- |
|  |