Examen Data Science Part 2 :

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 08/03/2024

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 (4 points)

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

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

3. **Comment les entreprises utilisent-elles la science des données pour améliorer leur performance et prendre des décisions éclairées ? (0.25 p)**

4. **On utilise la bibliotheque NumPy pour quelle raison ? (0.25 p)**

5. **Prenons l'exemple d'un tableau 2D : arr_2d = np.array([[1,2,3], [4,5,6], [7,8,9]]) (0.5 p)**
    - Output de **print(arr_2d[1, 2])**
    - Output de **print(arr_2d[1])**
    - Output de **print(arr_2d[:, 1])**

6. **L'importance du bibliothèque Pandas ? (0.5 p)**

7. **Au cœur de Pandas se trouvent deux structures de données fondamentales : quelles sont-elles ? La différence entre eux ? (0.5 p)**

8. **La différence entre df.loc[] et df.iloc[] dans Pandas? (0.5 p)**

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

10. **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 (4 points)

11. **Expliquez l'importance de la visualisation des données dans le processus de prise de décision. (0.5 p)**

12. **Quels sont les types de données ? (0.5 p)**

13. **Quelles sont les différences entre les visualisations univariées, bivariées et multivariées ? Donnez un exemple de chaque type (et pour chaque type de données). (1 p)**

14. **Comment l'exploration des données (EDA) peut-elle aider à identifier des tendances et des anomalies dans un ensemble de données ? (0.5 p)**

15. **La différence entre Barplot et Histplot? (0.5 p)**

16. **Expliquez le concept de corrélation entre deux variables. Comment la corrélation peut-elle être interprétée dans le contexte de l'exploration des données (EDA) ? Donnez un exemple concret avec la formule. (0.5 p)**

17. **Quel est l'inconvénient de la corrélation de Pearson ? (0.5 p)**

Part 3 : Machine Learning Fundamentals (2 points)

18. **Donner une brève définition de l'apprentissage automatique ? (0.25 p)**

19. **La différence entre l'apprentissage automatique et la programmation traditionnelle (classique) ? (0.25 p)**

20. **La différence entre apprentissage supervisé, non supervisé et par renforcement ? (0.5 p)**

21. **Donner une définition d'apprentissage semi-supervisé ? (0.5 p)**

22. **C'est quoi XAI (eXplainable AI) ? Et l'importance de XAI ? (0.5 p)**

Part 4 : Feature Engineering (6 points)

23. **Qu'est-ce que l'ingénierie des caractéristiques (feature engineering) et pourquoi est-elle importante dans le processus de modélisation ? (1 pt)**

24. **Expliquer le concept GIGO (Garbage In Garbage Out) ? (0.5 p)**

25. **Lister les quatre types de données ? (0.5 p)**

26. **Pourquoi encoder les données catégorielles ? (0.5 p)**

27. **La différence entre one hot encoding et dummy encoding? (0.5 p)**

28. **C'est quoi le mean-encoding? (0.5 p)**

29. **C'est quoi le feature scaling? Et l'importance du feature scaling avec un exemple ? (0.5 p)**

30. **La différence entre normalization et standardization? Donnez une méthode pour chaque type avec la formule ? (1 p)**

31. **C'est quoi le data leakage (fuite de données)? Un exemple? (0.5 p)**

32. **C'est quoi skewed distribution ? Comment on le fixe ? (0.5 p)**

Part 5 : Intro to Deep Learning (4 points)

33. **Une petite définition du deep learning? (1 p)**

34. **Pourquoi utiliser le deep learning par rapport à l'apprentissage classique ? (1 p)**

35. **En 1957, Frank Rosenblatt proposait le premier réseau de neurones artificiels, quel était le nom de cet ANN ? (1 p)**

36. **Quelle était la limite (l'inconvénient) de l'ANN de Frank Rosenblatt ? (1 p)**