I. (10%) **True or False Questions (Please answer T for True or F for False):**

1. In data analysis, using the Python Pandas package makes it convenient to handle structured data.

T

2. In data analysis, data cleaning only refers to removing missing values from raw data.

F

3. In machine learning, we typically divide data into training and testing data sets to evaluate the performance of the model.

T

4. Supervised learning is a machine learning approach where the model is trained based on labeled training data.

T

5. Decision tree is a supervised learning method used for both classification and regression tasks.

T

II. (10%) **Multiple Choice Questions (Please answer a, b, c, d):**

1. Which method can be used for statistical analysis in Python?

a) describe()

2. Which is the common algorithm used for classification tasks in machine learning?

b) K Nearest Neighbors

3. Which function is used for handling missing values in data analysis?

d) all of the above

4. Which method is used for grouping data in data analysis?

a) Groupby()

5. Which chart can be used to visualize the distribution of data?

d) all of the above

IV. (50%) Essay Questions:

1. (10%) Please explain the importance of data cleaning in the process of data analysis, and list three common data cleaning techniques, providing examples of their application scenarios.

資料清理在資料分析過程中非常重要，它會影響到分析結果的準確性和可靠性。未經處理的原始資料往往包含許多問題，如缺失值、異常值、重複記錄或不一致的資料格式，這些都可能導致分析結果產生偏差。

以下是三種常見的資料清理技術及其應用場景的例子：

1. 處理缺失值：處理方法包括刪除缺失值、填補缺失值、或者使用標記法標記缺失值。例如，在房價預測模型中，如果某些房屋的房間數量資料缺失，可以根據現有房屋的平均房間數來填補這些缺失值。

2. 識別和處理異常值：異常值可能由錯誤輸入、測量錯誤或是真實的偏離正常範圍的值引起。處理異常值的方法包括刪除、替換或使用轉換方法（如對數轉換）來減少異常值的影響。舉例來說，在醫療數據分析中，異常的患者體溫（如高於正常範圍太多）可能需要被視為測量錯誤並予以處理。

3. 資料格式標準化：不同來源的資料往往存在格式不一致的問題，資料格式

標準化是將這些資料轉換為一個共同的格式，以便於後續處理。這包括日期格式的統一、文本資料的大小寫轉換、數值資料的單位轉換等。例如，在處理一個國際公司的員工資料時，可能需要將不同國家的日期格式統一，以便進行統一的時間序列分析。

2. (10%) Please explain the significance of overfitting in machine learning and provide methods to avoid overfitting.

當一個模型過度學習訓練數據集中的細節和噪聲，以至於它在新的、未見過的數據上表現得不好時，就發生了過擬合。這意味著模型雖然能夠在訓練數據上達到很高的準確率，但是它對於新數據的泛化能力很差。

以下四個是避免模型過擬和的方法：

1. 使用更多的訓練數據
2. 減少模型複雜度
3. 使用正則化（如 L1 和 L2 正則化）
4. 交叉驗證

3. (10%) Please explain the working principles of the decision tree and the random forest models, compare their advantages and disadvantages, and provide examples of suitable application scenarios for each.

1. 決策樹和隨機森林的優缺點比較：

決策樹

優點：模型易於理解和解釋，可以直觀地看到決策過程；計算複雜度不高。

缺點：容易過擬合，對於有很多特徵的數據集，樹的結構可能會非常複雜；對於某些非線性問題表現不佳。

隨機森林

優點：具有很好的泛化能力，不易過擬合；既可以處理分類問題，也可以處理回歸問題；能夠處理高維度數據和大規模數據集。

缺點：模型的解釋性不如單一的決策樹；訓練和預測的時間可能比決策樹長。

2. 兩者適合應用場景：

決策樹：需要模型具有很好解釋性的場景，如醫療診斷、風險評估等；數據特徵與目標結果之間存在明顯的決策邏輯時。

隨機森林：對模型的泛化能力和預測準確率要求較高的場景，如金融詐騙檢

測、大規模影像分類等；數據集較大且特徵維度較高的問題。

4. (10%) Please discuss common data visualization tools in data analysis, such as Matplotlib and Seaborn, explaining their pros and cons and suitable usage methods.

1. Matplotlib

優點：

a) 靈活性：Matplotlib 是底層的庫，提供了大量的繪圖功能和細節控制，適合於需要精細定製圖表的情況。
b) 廣泛支持：作為 Python 數據視覺化的基石，許多其他繪圖庫都是建立在 Matplotlib 的基礎上，並且有大量的學習資源和社區支持。
c) 多樣的輸出格式：支持多種靜態、互動和動畫的輸出格式，方便在不同場景下使用。

缺點：

a) 學習曲線：由於其豐富的功能和靈活性，學習如何使用 Matplotlib 可能會有些困難。
b) 代碼冗長：為了達到精細的圖表定製，可能需要寫大量的代碼。

2. Seaborn

優點：

a) 易於使用：Seaborn 是基於 Matplotlib 的高級接口，提供了更加簡潔的繪圖 API，使得創建常見的統計圖表變得更加容易。
b) 美觀的默認風格：Seaborn 提供了多種預設的主題和顏色選擇，即使不進行復雜的定製也能創建美觀的圖表。
c) 更好的數據集整合：直接支持 Pandas DataFrame，方便於數據分析流程中的使用。

缺點：

a) 靈活性有限：相比於 Matplotlib，Seaborn 在定製性方面的選擇較少，對於一些非常特殊的圖表需求，可能需要回到 Matplotlib 進行實現。
b) 功能專注於統計圖表：Seaborn 主要專注於統計視覺化，對於一些非統計類型的高度定製圖表支持不足。

Matplotlib & Seaborn 兩者適合的場景：

Matplotlib 適合於需要高度定製圖表的場景，如科學論文、專業報告等。

Seaborn 則更適用於數據探索階段，快速獲得數據的直觀理解，以及準備初步的數據報告和展示。

5. (10%) Is it appropriate to use random forest regression and decision tree regression to detect anomalies? Please elaborate carefully.

隨機森林回歸和決策樹回歸本身主要是用於預測連續的目標變量值，而不是直接用於異常檢測。異常檢測（也稱為離群點檢測）的目的是識別那些與大部分其他數據明顯不同的數據點。儘管隨機森林和決策樹不是專門為異常檢測設計的，但在某些情況下，它們可以間接用於這一目的。

隨機森林回歸和決策樹回歸用於異常檢測的方法：

a) 基於預測誤差：一種可能的方法是，首先使用隨機森林或決策樹模型對數據進行回歸預測，然後計算每個數據點的實際值與預測值之間的誤差。那些具有異常高或異常低預測誤差的數據點可以被視為異常值。這種方法假設大多數數據點可以被模型準確預測，而異常點會導致預測誤差較大。

b) 基於模型的特徵重要性：隨機森林模型能夠提供特徵重要性的評估，這可以幫助識別哪些特徵對於模型預測來說是重要的。通過分析特徵的重要性，可以幫助找到可能導致異常的因素。

結論：

雖然使用隨機森林回歸和決策樹回歸進行異常檢測在技術上是可行的，但它們並不是為這一目的而特別設計的。在選擇異常檢測方法時，考慮使用專門針對異常檢測設計的算法可能會更為適合。然而，這並不意味著隨機森林和決策樹在某些特定情況下不能作為有價值的工具，特別是當結合其他方法或用於特定類型的數據分析時。