📝 Assignment Task 1

⚠️ Academic Integrity Notice

Strict Prohibition on AI-Generated Code

The use of ChatGPT, Copilot, or any other AI tools to generate code for this assignment is strictly prohibited. Such actions constitute academic misconduct. Any submission found to contain AI-generated code will receive a grade of 0. We employ advanced detection methods to identify AI-generated content. Additionally, instructors may assess your understanding by asking you to explain your code and reasoning. So be ready to answer any questions based on your homework.

Please note that AI-generated code often lacks the unique stylistic elements and logical progression characteristic of human-written code. Such discrepancies are easily detectable by experienced instructors and automated tools.

Homework: choose at least two of the listed below problems and solve (two correctly solved problems account for 100 points)

Note: you need to use Python3 for this Homework

1. Linear Regression (50 points)

Dataset: https://archive.ics.uci.edu/dataset/186/wine+quality
Description: This dataset contains physicochemical properties of red and white wine samples, with the goal of predicting wine quality.

1.1 Mathematical Formulation (6 points)

In a simple linear regression model, the relationship between the dependent variable yyy and the independent variable xxx is expressed as:

$$y=\beta 0+\beta 1x+\epsilon$$

a) Explain the significance of each term in the equation.

b) Discuss the assumptions underlying linear regression models.

c) What are the potential consequences if these assumptions are violated?

1.2 Python Implementation with Scikit-Learn(44 points)

1. **Load and Inspect Data**
    - Import the dataset and display the first few rows.
    - Check for missing values and handle them appropriately.
    - Identify and remove any duplicate entries.
2. **Descriptive Statistics**
    - Calculate summary statistics (mean, median, standard deviation) for each feature.
    - Identify outliers using box plots and handle them accordingly.
3. **Feature Engineering**
    - Create new features that might be relevant (e.g., alcohol-to-density ratio).
    - Assess the impact of these new features on model performance.
4. **Univariate Analysis**
    a. Plot histograms for each feature to understand their distributions.
    b. Use box plots to visualize the spread and detect outliers.
5. **Bivariate Analysis**
    a. Create scatter plots to examine relationships between pairs of features.
    b. Compute and visualize the correlation matrix using a heatmap.
6. **Dimensionality Reduction**
    a. Apply Principal Component Analysis (PCA) to reduce data to two dimensions.
    b. Visualize the PCA results and interpret the variance explained.
7. **Data Normalization**
    a. Normalize the dataset using Min-Max scaling or Standardization.
    b. Compare model performance before and after normalization.
8. **Model Training**
    a. Train multiple regression models (e.g., Linear Regression, Ridge, Lasso) to predict wine quality.
    b. Evaluate models using Mean Squared Error (MSE) and R^2 score.
9. **Model Evaluation**
    a. Generate and interpret residual plots to assess model fit.
    b. Calculate and visualize the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
10. **Hyperparameter Tuning**
    a. Use GridSearchCV or RandomizedSearchCV to find optimal hyperparameters for models.
    b. Compare the performance of tuned models with default ones.

2. Logistic Regression (50 points)
https://archive.ics.uci.edu/dataset/2/adult
Description: This dataset contains demographic information about adults and aims to predict whether a person earns more than $50K/year based on attributes like age, education, and occupation.
Use Case: Ideal for binary classification tasks, as the target variable is categorical (income >50K or <=50K).

2.1 Mathematical Formulation (6 points)

a) Explain the concept of a decision threshold in classification models.

b) How does varying the decision threshold affect the model's performance metrics such as accuracy, precision, and recall?

c) What is the trade-off between precision and recall, and how is it quantified?

2.2 Python Implementation with Scikit-Learn (44 points)

1. **Load and Inspect Data**
    a. Import the dataset and display the first few rows.
    b. Check for missing values and handle them appropriately.
    c. Identify and remove any duplicate entries.
2. **Descriptive Statistics**
    a. Calculate summary statistics (mean, median, standard deviation) for each feature.
    b. Identify outliers using box plots and handle them accordingly.
3. **Feature Engineering**
    a. Create new features that might be relevant (e.g., age-to-hours-per-week ratio).
    b. Assess the impact of these new features on model performance.
4. **Univariate Analysis**
    a. Plot histograms for each feature to understand their distributions.
    b. Use box plots to visualize the spread and detect outliers.
5. **Bivariate Analysis**
    a. Create scatter plots to examine relationships between pairs of features.
    b. Compute and visualize the correlation matrix using a heatmap.
6. **Categorical Feature Analysis**
    a. Visualize the distribution of categorical variables using bar plots.
    b. Examine the relationship between categorical features and the target variable.
7. **Data Normalization**
    a. Normalize the dataset using Min-Max scaling or Standardization.

    b. Compare model performance before and after normalization.
8. **Model Training**
    a. Train multiple classification models (e.g., Logistic Regression, Decision Trees, Random Forests, Support Vector Machines) to predict income.
    b. Evaluate models using accuracy, precision, recall, and F1-score.
9. **Model Evaluation**
    a. Generate and interpret confusion matrices for each model.
    b. Plot ROC curves and calculate the AUC.
10. **Hyperparameter Tuning**
    a. Use GridSearchCV or RandomizedSearchCV to find optimal hyperparameters for models.
    b. Compare the performance of tuned models with default ones.

3. K-Means Clustering (50 points)

Dataset: https://archive.ics.uci.edu/dataset/45/heart+disease
Description: This dataset contains features related to heart disease, including attributes like age, sex, chest pain type, and maximum heart rate achieved.

Use Case: Ideal for clustering tasks, as it allows for the identification of patterns and groupings among patients based on their medical attributes.

- **3.1 Mathematical Formulation (5 points)**
    - Explain the K-Means clustering algorithm, including the objective function.
    - Discuss the convergence criteria and potential limitations of the algorithm.
- **3.2 Implementation (45 points)**

1. Data Preprocessing:
    a. Load the dataset and inspect the first few rows.
    b. Handle missing values appropriately.
    c. Encode categorical variables using techniques like one-hot encoding.
    d. Normalize or standardize numerical features to ensure uniform scale.
2. Exploratory Data Analysis (EDA):
    a. Visualize the distribution of key features using histograms and box plots.
    b. Examine correlations between numerical features using a heatmap.
    c. Identify potential relationships between features and the target variable.
3. Determine Optimal Number of Clusters:
    a. Use the Elbow Method to find the optimal number of clusters:

    b. Compute the within-cluster sum of squares (WCSS) for a range of cluster numbers.
    c. Plot WCSS against the number of clusters and identify the "elbow" point.
4. Apply K-Means Clustering:
    a. Implement the K-Means algorithm with the determined number of clusters.
    b. Assign cluster labels to each data point.
    c. Visualize the clusters using dimensionality reduction techniques like PCA.
5. Interpretation and Analysis:
    a. Analyze the characteristics of each cluster.
    b. Compare cluster centroids to understand distinguishing features.
6. Discuss potential medical insights gained from the clustering results. Reporting:
    a. Prepare a report summarizing the methodology, findings, and interpretations.
    b. Include visualizations and statistical analyses to support conclusions.