Programming for AI – Lab Midterm

Exploratory Data Analysis and Classification Models

Tasks:

1. **Dataset Selection:**
    - Each student must choose a tabular dataset from Kaggle that is suitable for classification tasks. The dataset should contain a mix of categorical and numerical features.
2. **Data Preprocessing:**
    - Perform data cleaning and preprocessing tasks to prepare the dataset for modeling.
    - Handle missing values, outliers, and any other data quality issues.
    - Encode categorical variables appropriately.
3. **Exploratory Data Analysis (EDA):**
    - Conduct exploratory data analysis to gain insights into the dataset.
    - Generate descriptive statistics, visualizations, and correlations to understand the relationships between features.
4. **Model Building:**
    - Implement two classification models: Linear Classifier and Logistic Regression Classifier
    - You are also required to build them by scratch, without using sklearn library
    - Split the dataset into training and testing sets for model evaluation.
5. **Model Training and Evaluation:**
    - Train both models on the training set.
    - Evaluate the models using appropriate performance metrics (e.g., accuracy, precision, recall, F1-score).
    - Compare the performance of the two models.

Deliverables:

A singular report containing the following aspects:

1. **Dataset Selection Report:**
    - Provide details about the chosen dataset, including its source, features, and target variable.
2. **Data Preprocessing Report:**
    - Document the steps taken for data cleaning and preprocessing. Include any transformations applied to the data.

3. **Exploratory Data Analysis (EDA) Report:**
    - Present the findings of the EDA, including visualizations and insights gained from the analysis.
4. **Model Implementation and Evaluation Report:**
    - Describe the implementation of both the Linear and Logistic Classifiers.
    - Present the evaluation results, including performance metrics and any observations.
5. **Conclusion and Recommendations:**
    - Summarize the overall findings, compare the models, and provide recommendations for further improvements.
6. **Code Submission:**
    - Submit the Python code used for data preprocessing, EDA, and model building.

Submission Deadline: April 14, 2024 - Midnight

Grading Criteria:

- Data Preprocessing (10%)
- Exploratory Data Analysis (EDA) (30%)
- Model Implementation and Evaluation (40%)
- Conclusion and Recommendations (10%)
- Presentation and Clarity (10%)

Note:

- Perform this activity in a group of 2 students.
- Plagiarism and unauthorized collaboration with peers will result in penalties.
- Students are encouraged to seek clarification and assistance if needed.