EE 590 Basics of Data Analysis and Machine Learning
Assignment #4: Data Analysis and Modeling on Health Insurance Dataset

Due Date: Thursday, December 7, 2023

In this assignment, we will develop skills in data handling, analysis, and modeling using Python, specifically focusing on a health insurance dataset. You will learn about loading data, detecting, and handling outliers, dealing with missing values, exploring data distributions, calculating statistical measures, and building a linear regression model.

Dataset
The dataset contains 7 columns: Age (numeric), Gender (categorical: male/female), BMI (floating point < 100), Children (integer), Smoker (categorical: yes/no), Region (categorical: northwest/northwest/southeast/southwest), and Expenses (floating point).

Tasks

1. **Load Data into DataFrame:**
    - Import necessary libraries.
    - Load the dataset from a CSV file into Pandas DataFrame.
2. **Outlier Detection and Removal:**
    - Use boxplots to identify outliers in continuous columns (Age, BMI, Children).
    - Calculate Z-scores and remove data points where the absolute Z-score is greater than 3.
3. **Handling Missing Values:**
    - Count missing values in each column.
    - Apply appropriate strategies to handle missing data (removal or imputation). Explain your decision for each column.
4. **Data Distribution Visualization:**
    - Plot histograms for each numerical column.
    - Use bar charts or count plots for categorical columns.
5. **Descriptive Statistics and Variability Measures:**
    - Compute descriptive statistics (mean, median, mode, etc.) for numerical columns.
    - Calculate variability measures (standard deviation, variance, etc.).
6. **Compute Quartiles:**
    - Determine quartiles for numerical columns (Age, BMI, Children, Expenses).
7. **Probability Distribution Modeling:**

- Fit suitable probability distributions to selected columns and visualize the fits.

8. **Feature Scaling:**
    - Apply standardization or normalization to the numerical features.
9. **Linear Regression Modeling:**
    - Select 'Expenses' as the target variable.
    - Create and train a linear regression model using the remaining columns.
10. **Prediction and Model Evaluation:**
    - Make predictions using the model.
    - Evaluate model performance using R-squared, MSE, or other relevant metrics.

Submission Details

- A Jupyter Notebook or Python script containing all the code for the above tasks.
- A report (PDF/Word) summarizing your findings, including relevant visualizations.
- If attempting the bonus task, include your model performance metrics in the report.

Evaluation Criteria

1. Completeness and accuracy of the analysis.
2. Quality and relevance of visualizations.
3. Depth and clarity of insights drawn from the dataset.
4. Quality of the final report and recommendations provided.