CS4101– Applied Machine Learning

Assignment 01

Due Date: 19th September 2023 Total Marks: 100

Description: Carry out all the following eight (08) tasks on the dataset of your choice. Additionally, furnish a brief **conclusion or summary,** consisting of 2-3 lines, for each activity within these tasks.

Task 1. **Data Summarization**:

- Calculate basic summary statistics (mean, median, standard deviation, etc.) for each numerical variable.
- Count the frequency of unique values for categorical variables.
- Calculate the number of missing values for each variable.

Task 2. **Data Visualization**:

- Create histograms or density plots to visualize the distribution of numerical variables.
- Generate bar plots or pie charts to visualize the distribution of categorical variables.
- Create box plots to identify outliers and understand the spread of data.
- Construct scatter plots to explore relationships between pairs of variables.
- Use heatmaps to visualize correlations between variables.

Task 3. **Handling Missing Data**:

- Explore the patterns of missing data across variables.
- Decide on an appropriate strategy for handling missing values (imputation, removal, etc.).

Task 4. **Outlier Detection and Treatment**:

- Identify and visualize outliers in numerical variables.
- Decide whether to remove, transform, or treat outliers based on domain knowledge and analysis goals.

Task 5. **Data Distribution Analysis:**

- Visualize the data distribution and assess skewness and kurtosis.

Task 6. **Bivariate Analysis**:

- Analyze relationships between pairs of variables through scatter plots.

Task 7. **Grouping and Aggregation**:

- Group data by categorical variables and calculate summary statistics within each group.
- Explore differences or patterns between different groups.

Task 8. **Data Transformation**:

- Apply mathematical transformations (e.g., logarithmic or exponential transformations) to normalize data.
- Convert categorical variables to numerical format using encoding techniques.