**Assignment 2**

**Problem Statement:** Perform the following operations using R/Python on the data sets: a) Compute and display summary statistics for each feature available in the dataset. (e.g.minimum value, maximum value, mean, range, standard deviation, variance and percentiles)

b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g.

Classification)'''

# **Library:**

# Software used:Python 3.x, Google colab

Libraries and packages used**:** NumPy, pandas, matplotlib, sklearn

**Theory:**

**Methodology:**

**a) Compute and display summary statistics for each feature:**

1.Load the dataset into a Pandas DataFrame.

2.Use the describe () function to generate summary statistics for numerical features, including count, mean, standard deviation, minimum, maximum, and percentiles.

3.Calculate additional statistics such as range and variance if required.

4.Display the summary statistics for each feature.

**b) Data Visualization - Create a histogram for each feature:**

1.Import the Matplotlib or Seaborn library for data visualization.

2.Iterate over each numerical feature in the dataset.

3.Use the plt.hist() function to create a histogram for each feature.

4.Customize the histogram by setting labels, titles, and adjusting bins if needed.

5.Display the histograms for each feature.

**c) Data Cleaning, Data Integration, Data Transformation, Data Model Building (Classification):**

1.Data Cleaning:

* Handle missing values: Impute or remove missing data points.
* Remove duplicates if any.
* Address outliers if necessary.

2.Data Integration:

* Merge or concatenate multiple datasets if applicable.

3.Data Transformation:

* Encode categorical variables if required using techniques like one-hot encoding or label encoding.
* Scale numerical features to ensure they have similar scales.

4.Data Model Building (Classification):

* Split the dataset into training and testing sets.
* Choose an appropriate classification algorithm such as Logistic Regression, Decision Trees, Random Forests, etc.
* Train the classification model on the training data.
* Evaluate the model's performance using metrics like accuracy, precision, recall, and F1-score on the testing data.
* Fine-tune the model parameters using techniques like cross-validation or grid search if necessary.
* Deploy the trained model for making predictions on new data

1. Summary statistics: Computing summary statistics helps in understanding the basic properties of each feature in the dataset, such as mean, standard deviation, minimum and maximum values, percentiles, etc.
2. Data visualization: Creating histograms for each feature provides insights into the distribution of data, revealing patterns, skewness, and potential outliers.
3. Data cleaning, Integration, Transformation: These steps involve handling missing values, encoding categorical variables, scaling features, etc., to prepare the data for modeling.
4. Model Building: Building a classification model using machine learning algorithms such as Decision Trees, Random Forests, or Support Vector Machines.

**Advantages:**

1. EDA helps in understanding the structure and characteristics of the data, aiding in better decision making.
2. Data visualization facilitates the identification of trends, patterns, and outliers in the data.
3. Machine learning modeling enables predictive analysis, which can be used for various applications such as customer segmentation, fraud detection, medical diagnosis, etc.

**Disadvantages:**

1. EDA and modeling require domain knowledge and expertise to interpret the results accurately.
2. Over-reliance on machine learning models without proper understanding of the data can lead to biased or misleading conclusions.

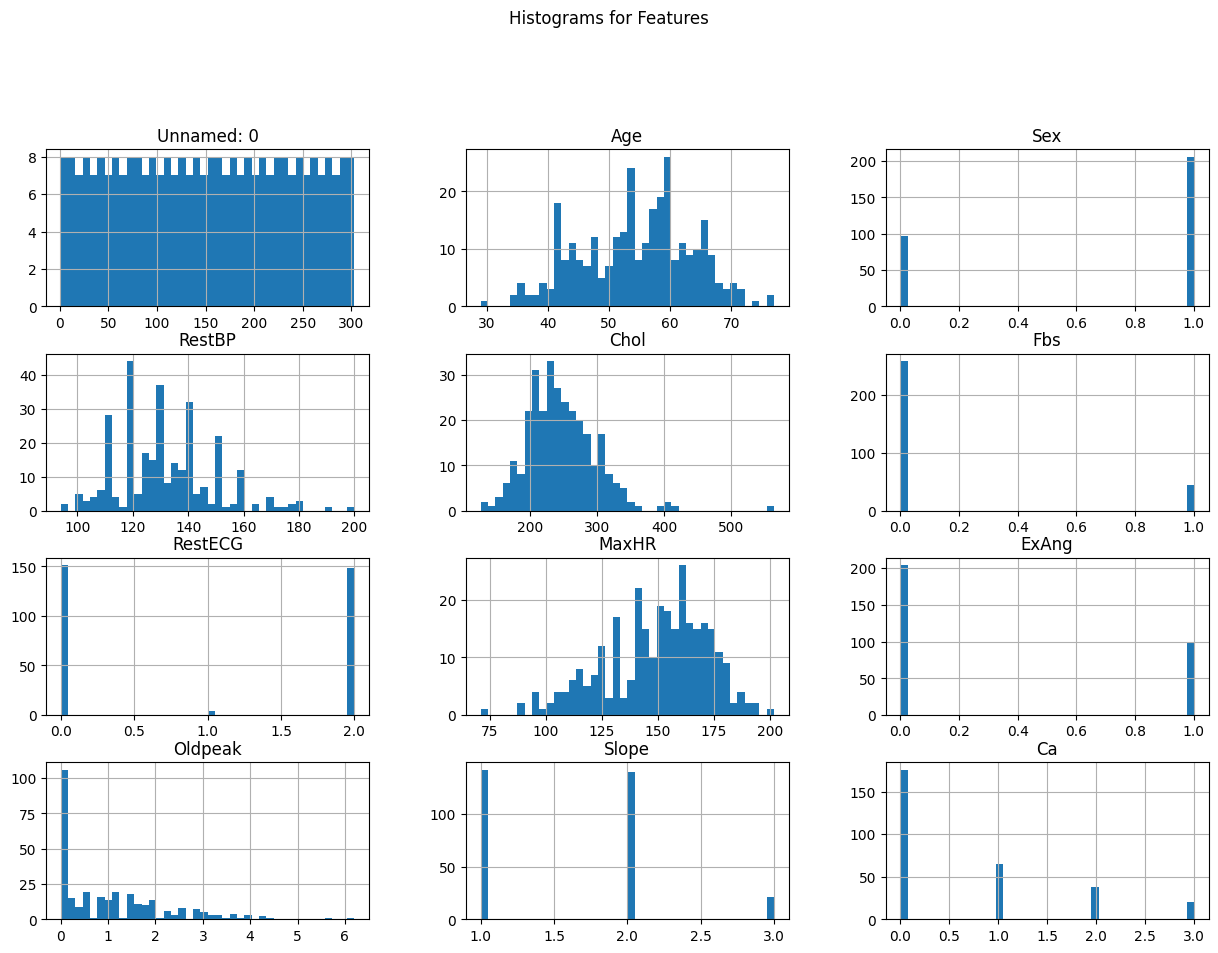
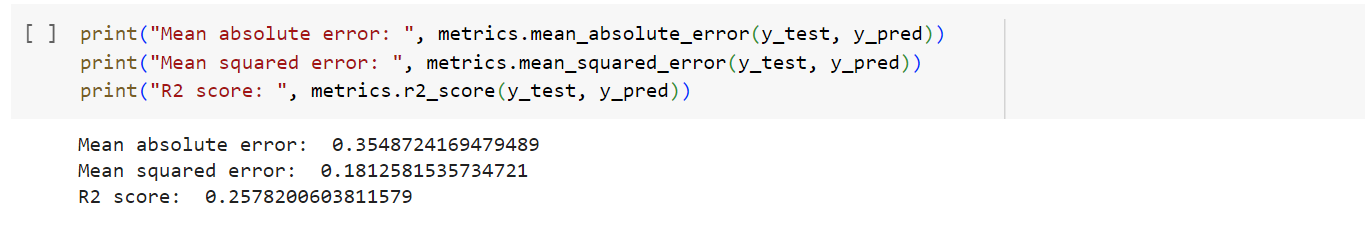
**Applications with example:**

1. EDA and machine learning modeling can be applied in various domains such as finance (credit risk analysis), healthcare (disease prediction), marketing (customer segmentation), etc.
2. Example: Predicting customer churn in a telecom company based on customer demographics, usage patterns, and service subscriptions.

**Working/ Algorithm:**

1. Load the dataset using Pandas.
2. Compute summary statistics using the **describe()** function.
3. Visualize data distributions using histograms with Matplotlib and Seaborn.
4. Perform data cleaning, integration, and transformation as necessary.
5. Build a machine learning classification model using Scikit-learn.
6. Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, etc.

**Diagram:**

****

**Conclusion:** In conclusion, this project demonstrates the importance of exploratory data analysis and machine learning modeling in understanding and extracting insights from data. By following a systematic approach, we can gain valuable insights into the data, identify patterns, and build predictive models that can be applied to real-world problems across various domains.