**Assignment 2**

**Exploratory Data Analysis and Machine Learning Modelling**

**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**).**

**Objective:** The objective of this project is to perform exploratory data analysis (EDA) on a given dataset, including computing summary statistics, visualizing data distributions, and then building a machine learning classification model. The dataset contains multiple features, and the goal is to understand the characteristics of the data and create a model that can classify instances accurately.

**Software used:**

1. Python 3.x
2. Google colab

**Libraries and packages used:**

NumPy, pandas, matplotlib, sklearn

**Theory:**

Methodology:

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:

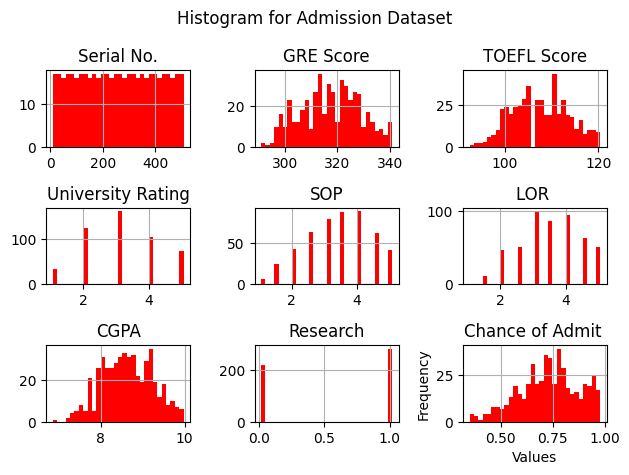
1. EDA and machine learning modeling can be applied in various domains such as finance (credit risk analysis), healthcare (disease prediction), marketing (customer segmentation), etc.
2. Example: Predicting customer churn in a telecom company based on customer demographics, usage patterns, and service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