**Name:** Samruddhi Pandit

**Roll no:** 281014

**Batch:** A1

**Assignment 2**

**Statement:**

Q. 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) Illustrate the feature distributions using histograms.  
c) Data cleaning, Data integration, Data transformation, Data model building (e.g., Classification).

**Objective:**

1. This assignment focuses on analyzing and preprocessing a dataset using various statistical and visualization techniques.
2. Gain insights into computing and interpreting summary statistics for different features.
3. Utilize visualizations to explore data distributions and detect patterns.
4. Implement essential data cleaning, integration, and transformation processes.
5. Develop a classification model for predictive analysis.

**Resources Used:**

1. Software: Jupyter notebook
2. Libraries: Pandas, Scikit-learn, Matplotlib, Seaborn

**Introduction to Data Analysis and Classification:**

1. Data analysis involves summarizing, visualizing, and preparing datasets for further modeling.
2. Classification models are used to predict categorical outcomes based on input features.
3. The dataset includes maternal health attributes such as blood pressure, glucose levels, heart rate, and associated risk labels.

**Methodology:**

1. Computing Summary Statistics:
   * Calculate key statistical measures such as minimum, maximum, mean, range, standard deviation, variance, and percentiles for each feature.
2. Feature Distribution Visualization:
   * Use histograms and other visualization tools to analyze numerical feature distributions.
3. Data Cleaning and Preprocessing:
   * Identify and handle missing values while resolving inconsistencies in the dataset.
   * Normalize or scale numerical features when necessary.
4. Data Integration and Transformation:
   * Merge datasets if required and encode categorical variables for better processing.
   * Apply feature engineering techniques to enhance data quality.
5. Model Development (Classification):
   * Select an appropriate classification algorithm such as Logistic Regression, Decision Tree, Random Forest, or SVM.
   * Train and evaluate the model using suitable performance metrics.

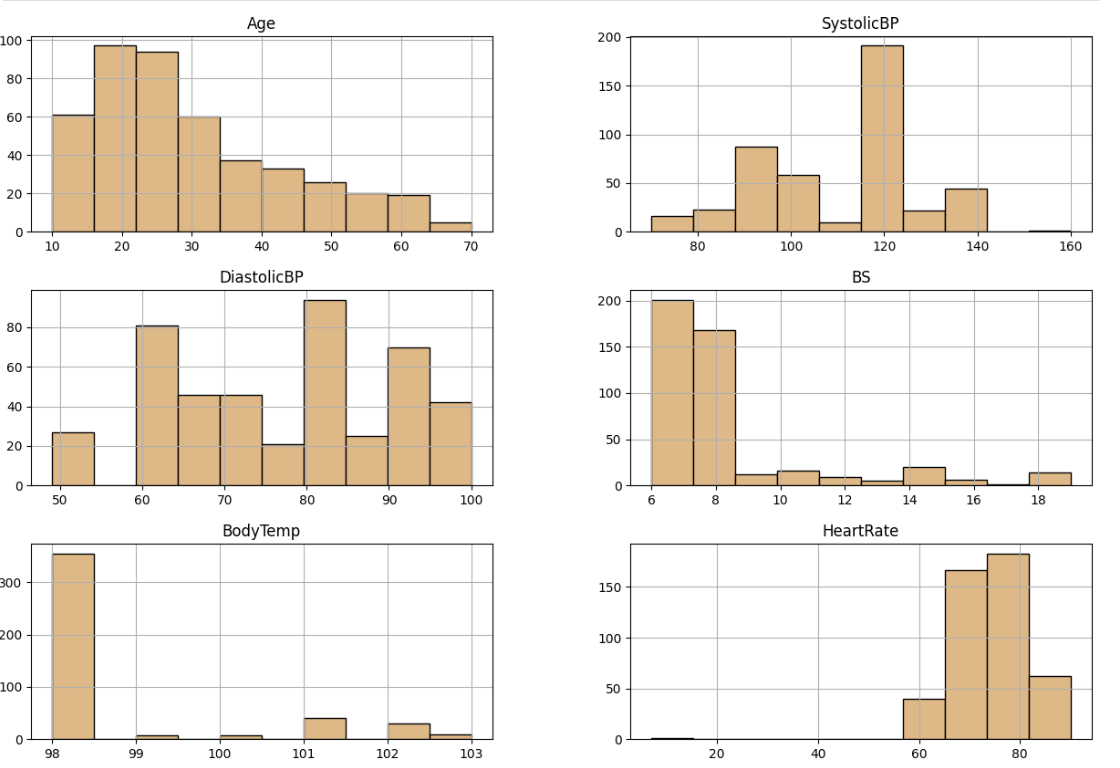
**Advantages:**

1. Enhances understanding of data characteristics and distributions.
2. Improves the accuracy of predictive models through robust preprocessing techniques.
3. Facilitates better decision-making, especially in healthcare applications.

**Disadvantages:**

1. Requires careful handling of missing and inconsistent data to avoid biases.
2. Model performance is highly dependent on dataset quality and preprocessing techniques applied.

**Results:**

****

**Conclusion:**

This assignment provided hands-on experience in dataset analysis by computing summary statistics and visualizing feature distributions. We applied essential preprocessing steps, including data cleaning, integration, and transformation. Finally, we built and evaluated a classification model to predict maternal health risks. These steps are fundamental for effective data-driven decision-making in healthcare analytics.