Name: Rugved Sawant
Roll No: 281065
Batch: A3

Assignment 2

Statement

Q. Perform the following operations using R/Python on the given datasets:

a) Compute and display summary statistics for each feature (e.g., minimum, maximum, mean, range, standard deviation, variance, and percentiles).
b) Illustrate feature distributions using histograms.
c) Perform data cleaning, integration, transformation, and build a classification model.

Objective

1. To analyze and preprocess the dataset using statistical and visualization techniques.
2. To compute and interpret summary statistics for each feature.
3. To visualize data distributions for better insight into feature behavior.
4. To perform key data preparation steps including cleaning, integration, and transformation.
5. To implement a **classification model** for predictive analysis.

Resources Used

- **Software**: Google Colab
- **Libraries**: Pandas, Scikit-learn, Matplotlib, Seaborn

Introduction to Data Analysis and Classification

Data analysis is the process of inspecting, cleansing, transforming, and modeling data to discover useful information and support decision-making. Classification, a supervised machine learning technique, is used to predict categorical outcomes based on input variables.

In this assignment, we worked with a dataset containing **maternal health attributes** such as blood pressure, glucose level, heart rate, and corresponding **risk labels**. The goal was to understand the data and build a model to predict health risks effectively.

Methodology

1. **Summary Statistics Computation**
   - Calculated **minimum**, **maximum**, **mean**, **range**, **standard deviation**, **variance**, and **percentiles** for each numerical feature using built-in Pandas functions.
2. **Feature Distribution Visualization**
   - Plotted **histograms** for each numerical column using **Matplotlib** and **Seaborn** to analyze feature distributions and detect skewness or outliers.

3. **Data Cleaning and Preprocessing**
   - Identified and handled **missing values**.
   - Addressed inconsistencies and ensured data quality.
   - Applied **normalization or scaling** where required to standardize features.
4. **Data Integration and Transformation**
   - Merged datasets if needed.
   - Encoded categorical variables using **label encoding** or **one-hot encoding**.
   - Performed feature engineering for model enhancement.
5. **Model Building (Classification)**
   - Selected an appropriate classification algorithm (e.g., **Logistic Regression**, **Decision Tree**, **Random Forest**, or **SVM**).
   - Split data into training and testing sets.
   - Trained and evaluated the model using **accuracy**, **confusion matrix**, and other performance metrics.

Advantages

1. Provides deep insight into data behavior through statistical analysis.
2. Improves model accuracy via robust preprocessing.
3. Supports effective decision-making, especially in healthcare-related applications.

Disadvantages

1. Sensitive to data quality – missing or inconsistent values can affect results.
2. Model performance varies based on preprocessing and feature selection.

Results

- Generated detailed summary statistics for all features.
- Visualized feature distributions, identifying patterns and anomalies.
- Preprocessed the data for model training.
- Built and evaluated a classification model for maternal health risk prediction, achieving a reasonable accuracy based on the chosen algorithm and dataset characteristics.

Conclusion

This assignment provided hands-on experience in **data analysis, preprocessing, and classification modeling**. By computing descriptive statistics and visualizing feature distributions, we gained valuable insights into the dataset. The cleaning, transformation, and classification steps further highlighted the importance of proper data preparation for building effective predictive models—especially in **healthcare analytics**, where accurate predictions can aid in timely and informed decisions.