**Assignment:2**

**Problem Statement:**

Get CC General dataset from following link:

[CC General dataset](https://drive.google.com/file/d/1ImYbuVM9NvtQSN9L8RzeRhfhrTl-AUTc/view?usp=classroom_web&authuser=0)

Perform the following operations using 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:**Perform data analysis operations on a given dataset using Python.

1. Compute and display summary statistics for each feature available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Perform data cleaning, data integration, data transformation, and data model building (e.g., Classification).

**Software and Hardware Requirements:**

* Operating System: Windows, Linux, macOS
* Python Environment: Jupyter Notebook, GoogleColab
* Libraries and Packages: Pandas, NumPy, Matplotlib, Seaborn

**Libraries and Packages Used:**

* Pandas: For data manipulation and analysis
* NumPy: For numerical computing
* Matplotlib: For data visualization
* Seaborn: For advanced data visualization

**Methodology:**

1. Data Loading: Load the dataset using Pandas.
2. Summary Statistics: Use Pandas to compute summary statistics for each feature.
3. Data Visualization - Histograms: Use Seaborn to create histograms for each feature.
4. Data Cleaning: Check for and handle missing values and duplicates in the dataset.
5. Data Integration: Combine data from different sources into a single dataset (if necessary).
6. Data Transformation: Convert categorical variables to numerical using encoding techniques (e.g., label encoding, one-hot encoding).
7. Data Model Building (Classification): Split the dataset into training and test sets, build a classification model using Scikit-learn (e.g., Logistic Regression), and evaluate the model's performance.

**Advantages and Disadvantages & Limitations/Example:**

1. **Advantages:**
   * **Summary Statistics:**
     + Provides a quick overview of the dataset's characteristics.
     + Helps in identifying outliers and understanding the distribution of features.
   * **Data Visualization:**
     + Enables intuitive understanding of feature distributions.
     + Facilitates identification of patterns and trends in the data.
   * **Data Cleaning, Integration, Transformation, Model Building:**
     + Enhances data quality and prepares it for analysis.
     + Facilitates the development of predictive models for classification tasks.
2. **Disadvantages & Limitations/Example:**
   * **Summary Statistics:**
     + May not capture all nuances of the data distribution, especially in complex datasets.
     + Outliers can skew summary statistics, affecting their interpretability.
   * **Data Visualization:**
     + Histograms may not provide sufficient detail for understanding complex relationships.
     + Interpretation of histograms can be subjective and influenced by binning choices.
   * **Data Cleaning, Integration, Transformation, Model Building:**
     + Data cleaning and transformation can be time-consuming, especially for large datasets.
     + Model performance heavily depends on data quality, feature selection, and algorithm choice.

**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:**Performing these operations on the dataset allows for a comprehensive analysis, providing insights into the data distribution, relationships between variables, and the potential for building predictive models.