Assignment No: - 2

Data Preparation

Problem Statement: -

Download advertising dataset from following link.

Advertising Dataset (kaggle.com)

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:

The objective of this task is to perform exploratory data analysis (EDA) on a dataset using Python.

Specifically, we aim to:

a) Compute and display summary statistics for each feature in the dataset to gain insights into the data distribution, central tendency, variability, and presence of outliers.

b) Create histograms for each feature in the dataset to visualize the distribution of values and understand their frequency and spread.

c) Perform data cleaning, integration, transformation, and model building to prepare the dataset for further analysis, such as classification or predictive modeling.

By achieving these objectives, we aim to gain a deeper understanding of the dataset, identify any data quality issues, visualize the distributions of features, and prepare the data for modeling or analysis tasks. This process will facilitate informed decision-making and help derive actionable insights from the data.

S/W Packages and H/W apparatus used: OS: Windows, Kernel: Python 3, Tools: Google Colab

Libraries and packages used: NumPy, Pandas, Matplotlib, Seaborn.

Theory: -

Methodology:

1. **Compute and Display Summary Statistics:**
   - **Python (using pandas):**
     - Use **describe ()** function to compute summary statistics.
2. **Data Visualization - Histogram Creation:**
   - **Python (using matplotlib or seaborn):**
     - Use **hist ()** function to create histograms for each feature.
3. **Data Cleaning, Integration, Transformation, Model Building:**
   - **Data Cleaning:**
     - Identify and handle missing values using techniques such as imputation or deletion.
   - **Data Integration:**
     - Merge or join multiple datasets based on common variables.
   - **Data Transformation:**
     - Normalize or scale features, encode categorical variables, and handle outliers.
   - **Model Building:**
     - Split data into training and testing sets.
     - Choose an appropriate machine learning algorithm (e.g., classification algorithm).
     - Train the model on the training data and evaluate its performance on the testing data.

Advantages and Disadvantages & Limitations/Example:

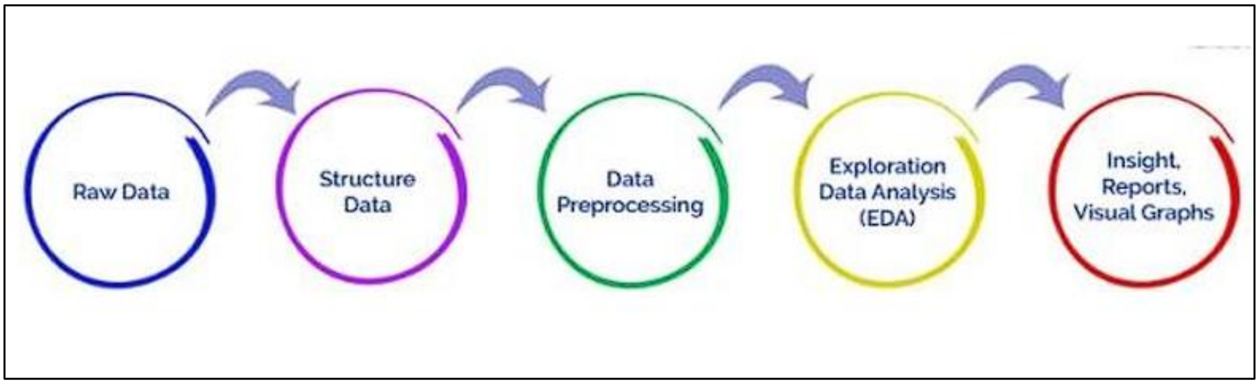
1. **Advantages:**
    - **Summary Statistics:**
        - Provides a quick overview of the dataset's characteristics.
        - Helps in identifying outliers and understanding the distribution of features.
    - **Data Visualization:**
        - Enables intuitive understanding of feature distributions.
        - 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.

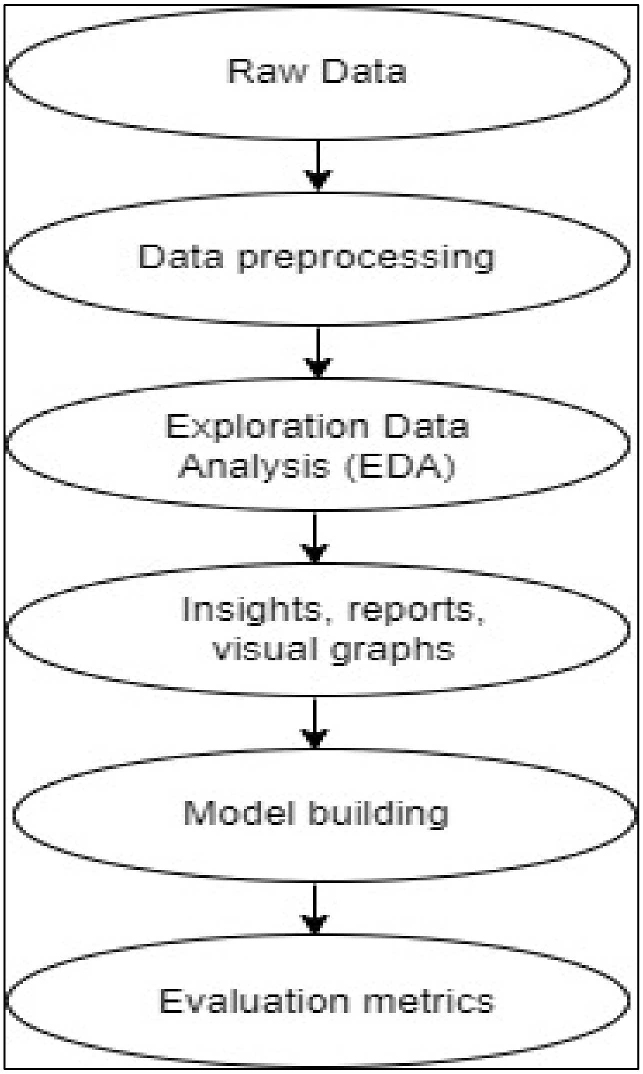
- Model performance heavily depends on data quality, feature selection, and algorithm choice.

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:


Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs


Raw Data
Data preprocessing
Exploration Data Analysis (EDA)
Insights, reports, visual graphs
Model building
Evaluation metrics

Conclusion

The methodology involves using Python for data analysis, including computing summary statistics, creating histograms, and cleaning and transforming data for classification modeling. Though versatile, these methods might oversimplify data and lack detailed insights. However, they can still provide valuable insights for decision-making when used carefully.