**ASSIGNMENT HELP**

**MANUAL**



SUBMITTED

TO

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

FOR THE SKILL AND COMPETENCY EVALUATION OF

**DATA SCIENCE & MACHINE LEARNING**

IN

**CSE AI DEPARTMENT**

BY

**Divyansh Mohta [22210998]**

**Class: S.Y. BTech Division: A Batch: A3**

**Batch Teacher**

**Dr. ANURADHA YENKIKAR.**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SR. NO.** | **CONTENTS** | **PAGE NO.** |
| **1** | **PROBLEM STATEMENT** | **4-5** |
| **2** | **LIBRARY USED** | **5-6** |
| **3** | **THEORY** | **6-7** |
| **4** | **METHODOLOGY** | **7-14** |
| **5** | **ADVANTAGES & DISADVANTAGES** | **14-15** |
| **6** | **WORKING** | **16-17** |
| **7** | **DIAGRAM** | **17-18** |
| **8** | **CONCLUSION** | **18-19** |

**1.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)

**2. LIBRARY USED:**

For Python:

* Pandas: For data manipulation and analysis
* Matplotlib/Seaborn: For data visualization
* Scikit-learn: For building classification models

**3. THEORY:**

**1)Computing Summary Statistics:**

Summary statistics provide a concise summary of the dataset's characteristics, including measures of central tendency, dispersion, and distribution.

Descriptive statistics functions like describe() in pandas or summary() in R are commonly used to compute summary statistics such as minimum, maximum, mean, standard deviation, variance, and percentiles.

**2)Data Visualization:**

Data visualization is a powerful technique for visually representing data to gain insights and identify patterns.

Histograms are graphical representations that display the frequency distribution of continuous variables by dividing the data into bins and plotting the number of observations in each bin.

Histogram plotting functions like hist() in matplotlib/seaborn for Python or geom\_histogram() in ggplot2 for R are used to create histograms.

**3)Data Cleaning, Integration, Transformation, and Model Building:**

Data cleaning involves identifying and handling missing values, outliers, or inconsistencies in the dataset to ensure data quality.

Data integration merges data from multiple sources into a single dataset, ensuring seamless integration of relevant information.

Data transformation techniques like normalization, scaling, or encoding categorical variables prepare the data for modeling by standardizing the features or converting categorical variables into numerical format.

Model building involves selecting suitable machine learning algorithms based on the problem type and dataset characteristics and training the model to make predictions.

**4. METHODS:**

**1)Compute Summary Statistics:**

Use descriptive statistics functions like describe() in pandas or summary() in R to compute summary statistics for each feature.

**2)Data Visualization:**

Utilize histogram plotting functions like hist() in matplotlib/seaborn for Python or geom\_histogram() in ggplot2 for R to create histograms for each feature.

**3)Data Cleaning, Integration, Transformation, and Model Building:**

Employ techniques like dropna() or fillna() in pandas or na.omit() in R for handling missing values.

For data integration, use functions like merge() in pandas or merge() in dplyr for R.

Implement data transformation methods such as normalization using MinMaxScaler or encoding categorical variables using OneHotEncoder in Python's scikit-learn, or corresponding functions in R.

Build classification models using algorithms like Logistic Regression, Decision Trees, or Random Forests using fit() and predict() functions from scikit-learn in Python, or train() and predict() functions from caret in R.

**5. ADVANTAGES AND DISADVANTAGES:**

**Advantages:**

* Computing summary statistics provides a quick overview of the dataset's characteristics, aiding in initial data exploration.
* Data visualization through histograms offers intuitive insights into the distribution of features, helping identify potential patterns or anomalies.
* Data cleaning, integration, transformation, and model building collectively contribute to preparing the dataset for analysis, ensuring data quality and enhancing predictive performance.

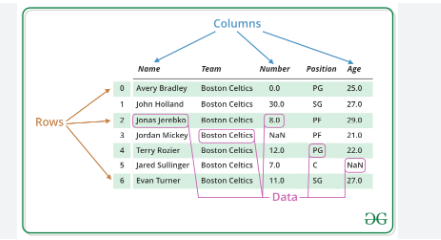
**Disadvantages:**

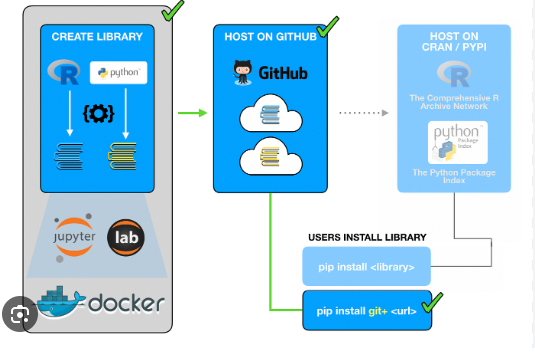
* Depending solely on summary statistics may overlook nuances in the data distribution, necessitating complementary visualizations for comprehensive understanding.
* Histograms may not effectively represent the distribution of features in complex datasets with high dimensionality.
* Data cleaning, integration, and transformation processes require careful handling and domain expertise to avoid unintended consequences or bias in the analysis

**6. WORKING:**

The workflow involves sequentially executing each step outlined in the problem statement using suitable functions and libraries in R or Python. Starting with computing summary statistics, followed by creating histograms for visualization, and then proceeding with data cleaning, integration, transformation, and model building. Each step contributes to the overall process of data analysis and model development, culminating in actionable insights or predictive models.

**7. DIAGRAM :-**





DIAGRAMS FOR THE ABOVE PROGRAM

**8. CONCLUSION:**

By systematically performing summary statistics computation, data visualization, data cleaning, integration, transformation, and model building, analysts can gain comprehensive insights into the dataset and develop reliable predictive models for decision-making. This approach ensures data quality, enhances interpretability, and fosters informed decision-making processes based on empirical evidence extracted from the data.