Kaushal Galav D15C 11

AIM: Perform Regression Analysis using Scipy and Sci-kit learn.

THEORY:

1] **Regression Analysis:** Regression analysis is a statistical method used to understand the relationship between one dependent variable and one or more independent variables. It helps in predicting outcomes and identifying trends by analyzing past data. For example, a business can use regression to predict future sales based on factors like advertising spend and customer reviews

2] **Regression Model:** Regression Model for Prediction involves training a model on historical data to identify patterns and relationships between variables.
The trained model is then used to predict outcomes for new data. Depending on the dataset, different regression techniques (such as logistic or linear regression) are applied to achieve accurate predictions.

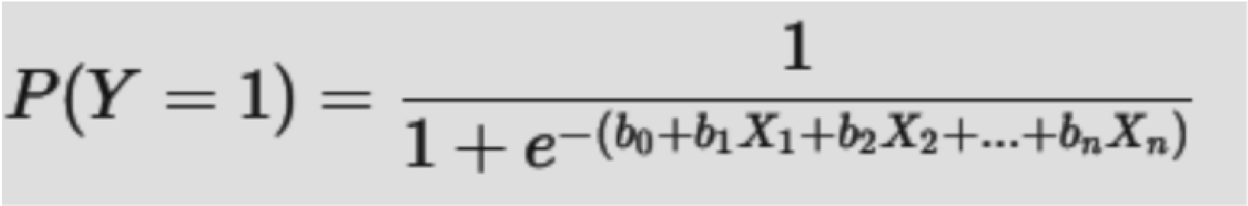
3] **Types of Regression Analysis:**

- **Linear Regression:**The simplest form, where a straight line shows the relationship between one dependent and one independent variable.
The formula for linear regression is :

$$Y_i = f(X_i, \beta) + e_i$$

- **Logistic Regression:** Logistic Regression is a statistical method used for binary classification problems, where the outcome is either 0 or 1 (e.g., success/failure, yes/no). It estimates the probability of an event occurring based on independent variables using the sigmoid function. Logistic Regression uses the sigmoid function (also called the logistic function) to model the relationship between the independent variables and the probability of a binary outcome.

The Formula for Logistic Regression is :



$$P(Y=1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + ... + b_n X_n)}}$$

Dataset:

The dataset, "Global Health Statistics," contains health-related statistics across different countries. It includes information on various diseases, treatment availability, mortality rates, and demographics. Key columns include:

- **Country**: The name of the country where the health data is recorded.
- **Disease Name**: The name of the disease being tracked in the dataset.
- **Disease Category**: The category under which the disease falls (e.g., infectious, non-infectious).
- **Age Group**: The age range affected by the disease.
- **Gender**: The gender of the individuals affected.
- **Treatment Type**: Type of treatment available for the disease.
- **Availability of Vaccines/Treatment**: A binary indicator of whether vaccines or treatments are available for the disease.
- **Mortality Rate (%)**: The percentage of deaths attributed to the disease in a given region.

The dataset is used for various analysis purposes, including predicting the availability of vaccines or treatment and understanding mortality trends

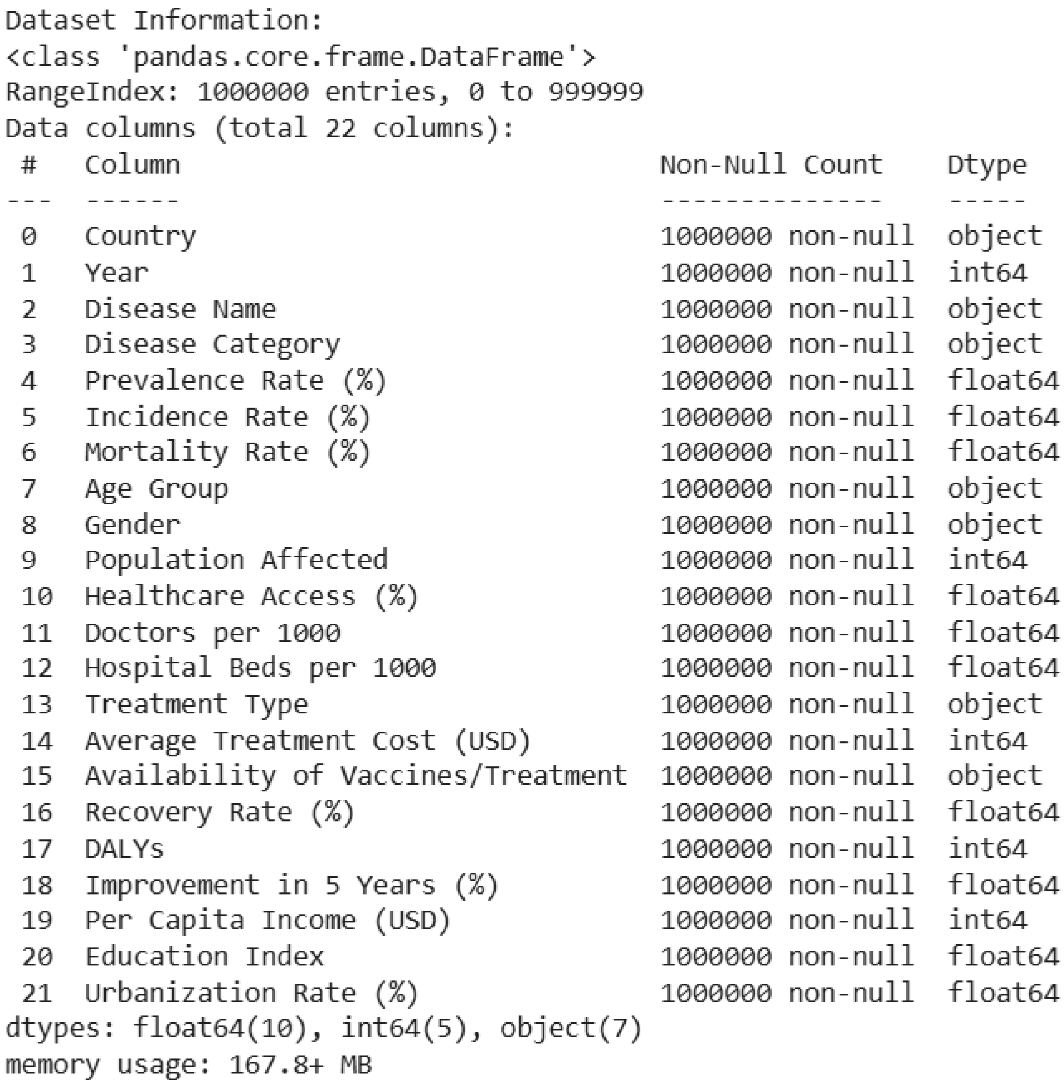
Steps:

1. **Load the Dataset**

   - The dataset is loaded into a pandas DataFrame using the `pd.read_csv()` function.
   - The file path to the dataset is specified.

2. **Display Basic Information**

   - The basic structure and data types of the dataset are displayed using `df.info()` to understand its contents.



```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 22 columns):
 #   Column                              Non-Null Count    Dtype
---  ------                              --------------    -----
 0   Country                             1000000 non-null  object
 1   Year                                1000000 non-null  int64
 2   Disease Name                        1000000 non-null  object
 3   Disease Category                    1000000 non-null  object
 4   Prevalence Rate (%)                 1000000 non-null  float64
 5   Incidence Rate (%)                  1000000 non-null  float64
 6   Mortality Rate (%)                  1000000 non-null  float64
 7   Age Group                           1000000 non-null  object
 8   Gender                              1000000 non-null  object
 9   Population Affected                 1000000 non-null  int64
 10  Healthcare Access (%)               1000000 non-null  float64
 11  Doctors per 1000                    1000000 non-null  float64
 12  Hospital Beds per 1000              1000000 non-null  float64
 13  Treatment Type                      1000000 non-null  object
 14  Average Treatment Cost (USD)        1000000 non-null  int64
 15  Availability of Vaccines/Treatment  1000000 non-null  object
 16  Recovery Rate (%)                   1000000 non-null  float64
 17  DALYs                               1000000 non-null  int64
 18  Improvement in 5 Years (%)          1000000 non-null  float64
 19  Per Capita Income (USD)             1000000 non-null  int64
 20  Education Index                     1000000 non-null  float64
 21  Urbanization Rate (%)               1000000 non-null  float64
dtypes: float64(10), int64(5), object(7)
memory usage: 167.8+ MB
```

3. **Encode Categorical Variables**

    - Categorical columns are encoded using `LabelEncoder` from `sklearn.preprocessing`.

    - This step converts textual categories into numerical values to be used in machine learning models.

4. **Logistic Regression (Binary Classification)**

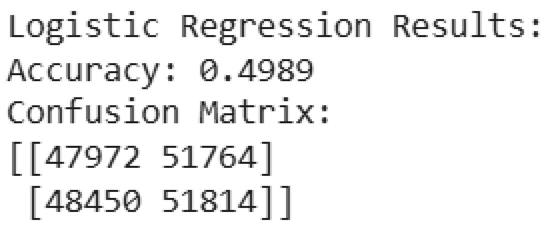
    - The target variable, **Availability of Vaccines/Treatment**, is selected for binary classification.

    - Features are separated from the target variable.

    - The data is split into training and testing sets using `train_test_split()`.

    - The features are standardized using `StandardScaler` to ensure better performance in the logistic regression model.

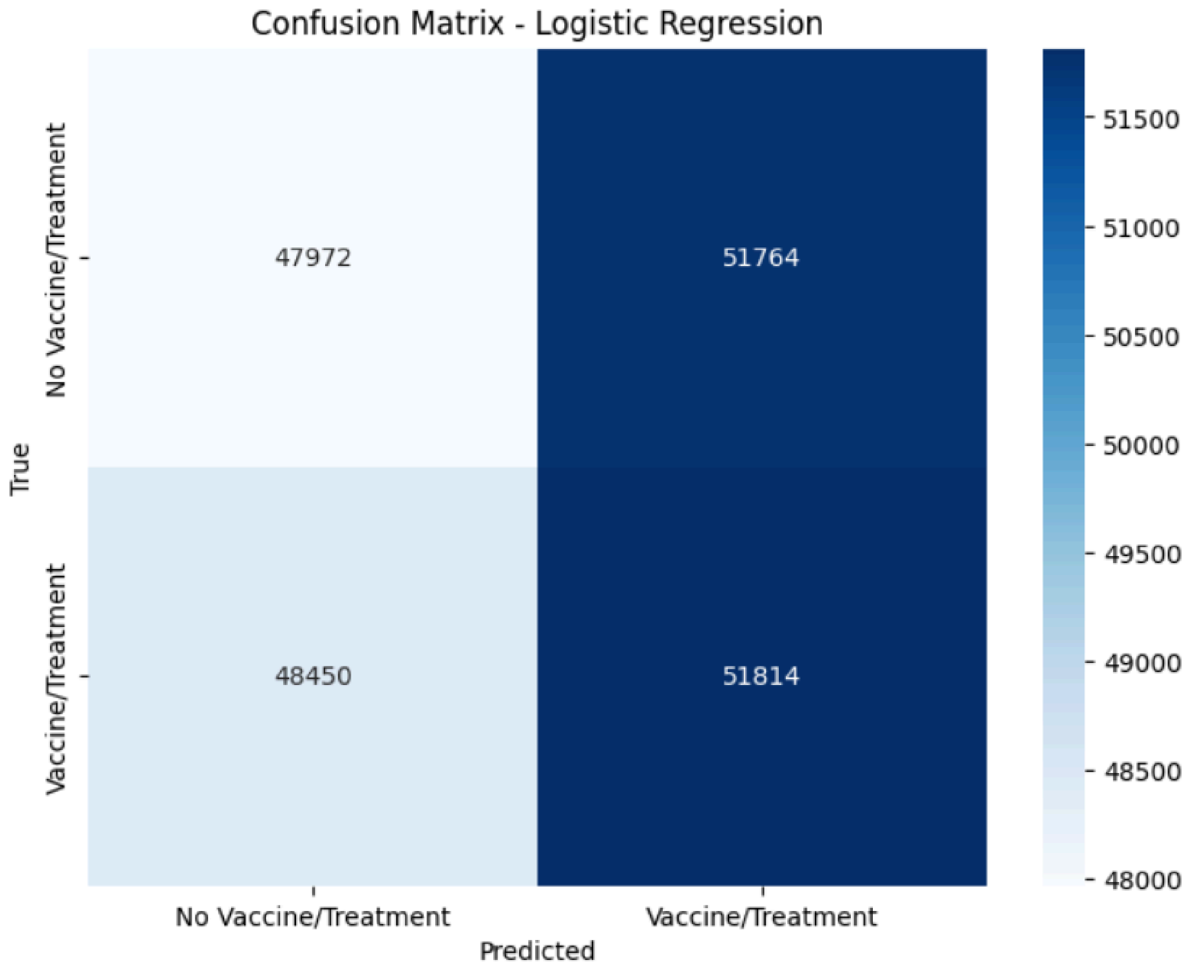
    - A logistic regression model is trained and evaluated using accuracy and confusion matrix.



```
Logistic Regression Results:
Accuracy: 0.4989
Confusion Matrix:
[[47972 51764]
 [48450 51814]]
```

5. **Confusion Matrix Heatmap**

    - The confusion matrix for logistic regression is visualized as a heatmap using `seaborn.heatmap()` to analyze model performance



6. **Linear Regression (Regression Analysis)**

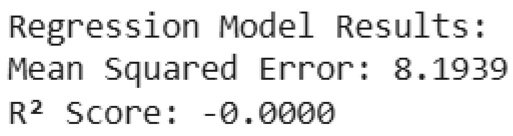
    ○ The target variable, **Mortality Rate (%)**, is selected for regression analysis.

    ○ Features are separated from the target variable.

    ○ The data is split into training and testing sets.

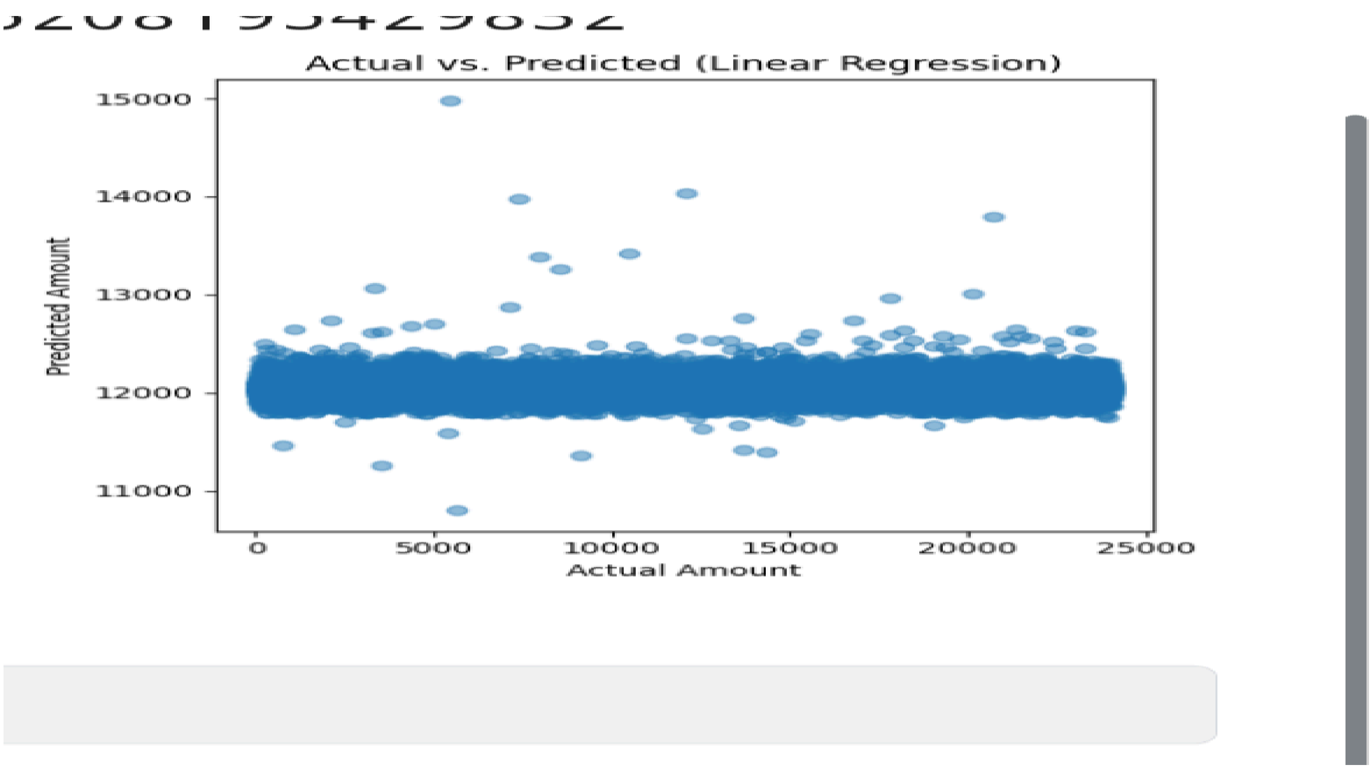
    ○ Features are standardized for the regression model.

    ○ A linear regression model is trained and evaluated using Mean Squared Error (MSE) and R² score.



```
Regression Model Results:
Mean Squared Error: 8.1939
R² Score: -0.0000
```

7. **Scatter Plot of Predicted vs. Actual Values**
   - A scatter plot is created comparing the predicted and actual mortality rates to visualize the model's prediction accuracy.



Conclusion:

In this analysis, we applied both **Logistic Regression** and **Linear Regression** models to a global health dataset. The logistic regression model aimed to predict the **Availability of Vaccines/Treatment**, and its performance was evaluated using accuracy and a confusion matrix, which highlighted the model's ability to classify binary outcomes. The linear regression model, on the other hand, predicted the **Mortality Rate (%)** and was assessed using Mean Squared Error (MSE) and R^2 score to measure the accuracy of predictions. Visualizations, such as the confusion matrix heatmap and the scatter plot of predicted vs. actual values, helped in interpreting the model outcomes. These models provide valuable insights for understanding health trends and the impact of available treatments and vaccines