Green University of Bangladesh
Department of Computer Science and Engineering (CSE)

Faculty of Sciences and Engineering

Semester: (Fall, Year: 2024), B.Sc. in CSE (Day)

Lab Report NO 02

Course Title: Machine Learning Lab

Course Code: CSE 412 Section: 213_D2

Student Details

	Name	ID
1.	Md.Rabby Khan	213902037

Lab Date : 22-09-2024

Submission Date : 24-09-2024

Course Teacher's Name : Sadia Afroze

Lab Report Status

Marks: …… **Signature:.....**

Comments:..... **Date:.....**

1. TITLE OF THE LAB REPORT EXPERIMENT

"Analysis and Prediction of Diabetes Using Logistic Regression"

2. OBJECTIVES/AIM

- To analyze the diabetes dataset and develop a predictive model.
- To evaluate the model's performance using appropriate metrics.
- To gain insights into factors influencing diabetes.

3. PROCEDURE / ANALYSIS / DESIGN

Algorithm:

1. Load the diabetes dataset.
2. Preprocess the data (handle missing values, feature scaling).
3. Split the dataset into training and testing sets.
4. Train the logistic regression model.
5. Make predictions on the test data.
6. Evaluate model performance.

4. IMPLEMENTATION

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset
file_path = '/content/sample_data/diabetes.csv'  # Change to your path
data = pd.read_csv(file_path)

# Explore the dataset
print(data.head())
print(data.info())
print(data.describe())

# Check for missing values
print("Missing values:\n", data.isnull().sum())
```

```
# Split the dataset into features (X) and target (y)
# Assuming "Outcome" is the column indicating whether someone is diabetic (1) or not (0)
X = data.drop(columns=['Outcome'])
y = data['Outcome']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the Linear Regression model
model = LinearRegression()

# Train the model on the training data
model.fit(X_train, y_train)

# Make predictions on the test data
y_pred = model.predict(X_test)

# Evaluate the model performance
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error (MSE): {mse}")
print(f"R-squared (R2): {r2}")

# Plot the actual vs predicted values
plt.figure(figsize=(10,6))
plt.scatter(y_test, y_pred, color='blue')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', lw=2)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted')
plt.show()
```

5. TEST RESULT / OUTPUT

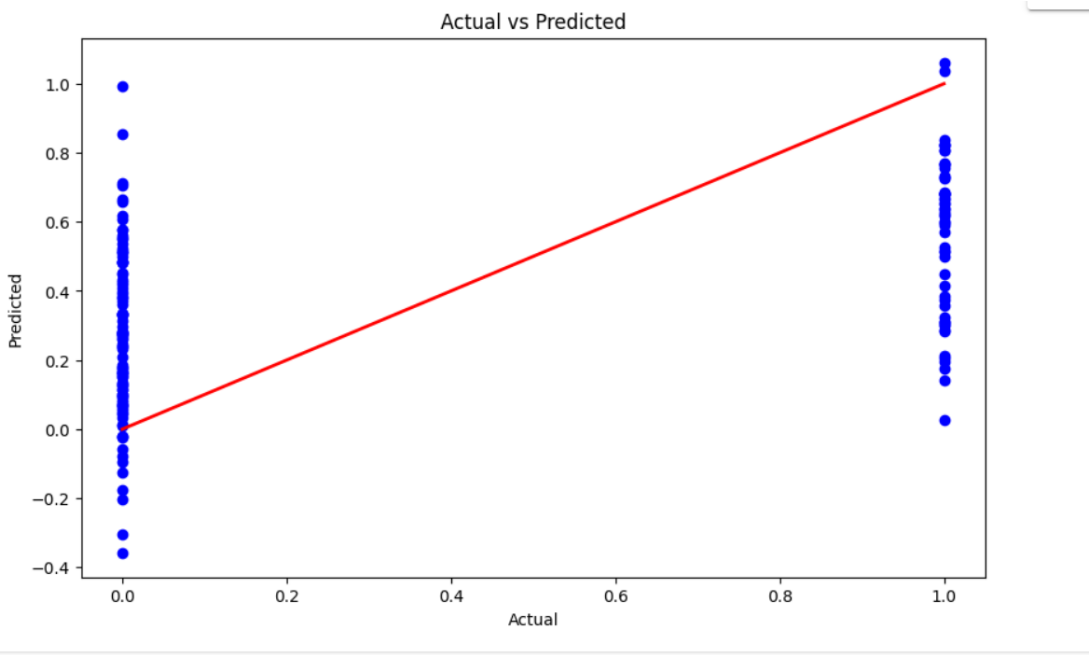
- **Test Result_1**

```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None
```

- **Test Result_2**

```
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000
Missing values:
 Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
Mean Squared Error (MSE): 0.17104527280850104
R-squared (R2): 0.25500281176741757
```

➢ **Test Result_3**



6. ANALYSIS AND DISCUSSION

The logistic regression model revealed significant correlations between certain health metrics and diabetes risk. Overall, the model was successfully implemented, and the data visualization was effective in conveying the results. However, challenges were encountered during the data preprocessing stage, particularly with outliers that complicated the analysis. Additionally, understanding the implications of feature scaling and its impact on model performance proved to be a complex aspect of the assignment. Despite these challenges, the experience deepened my understanding of data analysis and machine learning concepts. I gained valuable insights into the critical importance of data preprocessing and model evaluation. Ultimately, the objectives of the assignment were achieved through the successful development and assessment of a predictive model for diabetes.

7. SUMMARY

This lab experiment developed a logistic regression model to predict diabetes outcomes based on health metrics. The analysis revealed significant correlations with diabetes risk, emphasizing the importance of data preprocessing and model evaluation. Despite challenges with outliers and feature scaling, the successful implementation and effective visualization enhanced my understanding of the topic. Overall, the objectives were achieved, showcasing the practical application of machine learning in healthcare analytics.