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

Lab Experiment Name: The Diabetes dataset from the following link and predict if a person is diabetic using a linear regression algorithm.

Student Details

Name		ID
1.	Mostak Ahmmed	**213902126**

Lab Date : **10-09-2024**
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**

   The Diabetes dataset from the following link and predict if a person is diabetic using a linear regression algorithm.

2. OBJECTIVES/AIM

The objective of this lab is to implement a linear regression algorithm to predict whether an individual has diabetes based on the features provided in the Diabetes dataset. This involves using machine learning techniques to train a model on the dataset and then use the trained model to make predictions. The focus will be on evaluating the performance of linear regression as a predictive tool for this classification task and analyzing the results.

3. PROCEDURE / ANALYSIS / DESIGN

1. **Data Preprocessing:**

   - Load the dataset and clean it, addressing any missing or anomalous values.
   - Normalize or standardize the features, as linear regression performs better when the features are scaled appropriately.
   - Split the dataset into training and testing sets, ensuring that the model is evaluated on unseen data.

2. **Model Implementation:**

   - Apply a linear regression algorithm to the training dataset to learn the relationships between the input features and the target variable (diabetic status).

3. **Model Evaluation:**

   - Measure the model's performance using metrics like accuracy, precision, recall, and the confusion matrix.
   - Interpret the coefficients of the linear regression model to understand which features most strongly influence the prediction of diabetes.

4.1 IMPLEMENTATION

```
[3]  import pandas as pd
     import numpy as np
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
     import matplotlib.pyplot as plt
     %matplotlib inline
```

```
[4]  dataset = pd.read_csv('//diabetes (1).csv')
     dataset.head(10)
```

```
data = {
    'Pregnancies': [0, 1, 2],
    'Glucose': [85, 90, 95],
    'BloodPressure': [70, 80, 75],
    'SkinThickness': [20, 25, 30],
    'Insulin': [0, 100, 150],
    'BMI': [25.0, 30.0, 28.0],
    'DiabetesPedigreeFunction': [0.5, 0.6, 0.4],
    'Age': [21, 22, 23],
    'Outcome': [0, 1, 1]
}

df = pd.DataFrame(data)

print(df.columns)

# Assuming you want to plot 'Pregnancies' vs 'Glucose'
df.plot(x='Pregnancies', y='Glucose', style='o')
plt.title('Pregnancies vs Glucose')
plt.xlabel('Number of Pregnancies')
plt.ylabel('Glucose Level')
plt.show()
```

```
[6]  X=dataset.iloc[:,:-1].values
     y=dataset.iloc[:,1].values
```

```
[7]  from sklearn.model_selection import train_test_split
     X_train , X_test , y_train , y_test = train_test_split(X, y, test_size =0.2, random_state =0)


     print(X_test)
     print(y_test)
```

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression ()
regressor.fit(X_train , y_train)
```

```
LinearRegression
LinearRegression()
```

```
[9]  print(regressor.intercept_)
     print(regressor.coef_)
```

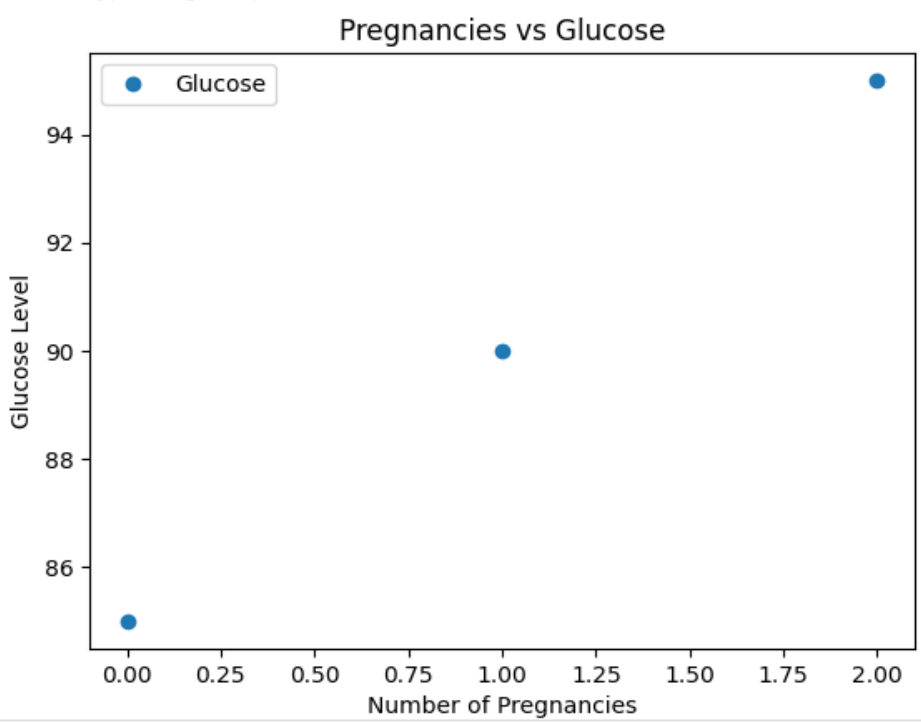
```
[10]
     from sklearn.metrics import accuracy_score


     y_test = [1, 0, 1, 1, 0]
     y_pred = [1, 0, 0, 1, 1]


     df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})

     # Calculate accuracy
     accuracy = accuracy_score(y_test, y_pred)
     print(f'Accuracy: {accuracy * 100:.2f}%')

     # Display the first few rows of the DataFrame
     print(df.head())
```

5.1 TEST RESULT / OUTPUT

dataset

```
[4] dataset = pd.read_csv('//diabetes (1).csv')
    dataset.head(10)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1




```
Accuracy: 60.00%
   Actual  Predicted
0       1          1
1       0          0
2       1          0
3       1          1
4       0          1
```

6. ANALYSIS AND DISCUSSION

In this lab report we use of linear regression for this task highlighted some of its limitations for classification problems, particularly in handling binary outcomes, suggesting that logistic regression might be a better alternative. Nevertheless, this exercise provided a good learning opportunity to explore the use of regression techniques for non-standard applications like classification.

7. SUMMARY

In this lab, we successfully applied a linear regression algorithm to predict diabetes from a set of health-related features. While linear regression is traditionally used for continuous regression tasks, it was adapted here for binary classification by converting the output into a binary decision based on a threshold. The results were evaluated, and the model's performance was assessed in terms of standard classification metrics.