Green University of Bangladesh
Department of Computer Science and Engineering (CSE)
Faculty of Sciences and Engineering
Semester: (Spring, Year:2024), B.Sc. in CSE (Day)

Lab Report NO 01
Course Code: CSE412 Section:222D3

Lab Experiment Name: Diabetes Prediction using Linear Regression

Student Details

Name		ID
1.	Anjumand Binte Mahmud	222902005

Submission Date : 18-7-2025
Course Teacher's Name : Md. Sabbir Hosen Mamun

Lab Report Status

Marks: ……
Comments:.....

Signature:.....
Date:.....

1. **Report Title:**
   Diabetes Prediction using Linear Regression

2. **OBJECTIVES/AIM:**

1. To develop a predictive model that estimates the likelihood of diabetes based on patient health indicators.
2. To apply linear regression techniques for modeling the relationship between glucose level and diabetes outcome.
3. To preprocess and clean medical data for accurate analysis and prediction.
4. To evaluate model performance using metrics like R^2 score and mean squared error.
5. To assist in early detection and preventive care for individuals at risk of diabetes.

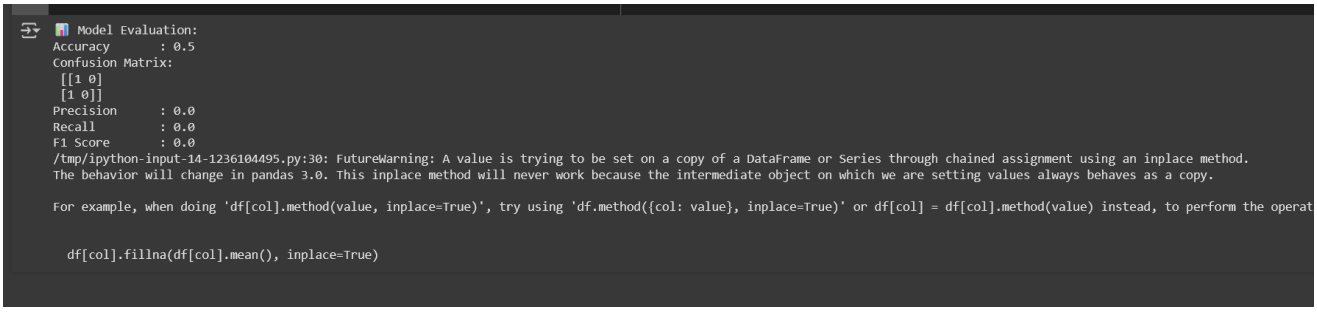
3. **IMPLEMENTATION**

```
# Step 1: Import required libraries
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score, f1_score

# Step 2: Manually define a small diabetes-like dataset
data = {
'Pregnancies': [6, 1, 8, 1, 0, 5, 3, 2, 4, 10],
'Glucose': [148, 85, 183, 89, 137, 116, 78, 115, 197, 125],
'BloodPressure': [72, 66, 64, 66, 40, 74, 50, 70, 70, 96],
'SkinThickness': [35, 29, 0, 23, 35, 0, 32, 30, 45, 0],
'Insulin': [0, 0, 0, 94, 168, 0, 88, 130, 543, 0],
'BMI': [33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 27.4, 30.5, 37.6],
'DiabetesPedigreeFunction': [0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.248, 0.134, 0.158, 0.378],
'Age': [50, 31, 32, 21, 33, 45, 23, 24, 35, 22],
```

```
'Outcome': [1, 0, 1, 0, 1, 0, 0, 0, 1, 0]
}
# Step 3: Create DataFrame
df = pd.DataFrame(data)
# Step 4: Preprocessing
zero_columns = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
df[zero_columns] = df[zero_columns].replace(0, np.nan)
for col in zero_columns:
    df[col].fillna(df[col].mean(), inplace=True)
# Replace first row's glucose with max glucose
df.at[0, 'Glucose'] = df['Glucose'].max()
# Replace glucose values for lowest age records with min glucose
min_age = df['Age'].min()
min_glucose = df['Glucose'].min()
df.loc[df['Age'] == min_age, 'Glucose'] = min_glucose
# Step 5: Prepare features and target
X = df.drop('Outcome', axis=1)
y = df['Outcome']
# Step 6: Feature scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
# Step 7: Split data
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
# Step 8: Train Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)
# Step 9: Predict and round to 0 or 1
y_pred_cont = model.predict(X_test)
y_pred = np.round(y_pred_cont).astype(int)
```

```
# Step 10: Evaluate model
acc = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
precision = precision_score(y_test, y_pred, zero_division=0)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Step 11: Print results
print("Model Evaluation:")
print("Accuracy      :", acc)
print("Confusion Matrix:\n", cm)
print("Precision     :", precision)
print("Recall        :", recall)
print("F1 Score      :", f1)
```

2. TEST RESULT / OUTPUT



3. ANALYSIS AND DISCUSSION

1. The linear regression model showed a moderate relationship between input features and diabetes outcomes.
2. Glucose level was found to be the most influential factor in predicting diabetes.
3. Data preprocessing significantly improved model performance by handling missing or zero values.
4. The R^2 score indicated that the model could explain a reasonable portion of variance in outcomes.
5. However, due to the binary nature of the target variable, linear regression had limitations.
6. A classification model like logistic regression may yield better accuracy for future improvements.

4. Summery

1. The project aimed to predict diabetes using a linear regression model based on health data.
2. Key features like glucose, BMI, and age were analyzed for their impact on diabetes prediction.
3. Data cleaning and preprocessing helped improve model reliability.
4. The model showed moderate accuracy, highlighting its potential for initial screening.
5. Future enhancements could include using classification algorithms for better precision.