A

Assesment Report

on

“Diagnose Diabetes”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Utkarsh Kumar Singh(202401100400202)

Under the supervision of

“Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

INTRODUCTION

Diabetes is a global health challenge affecting millions of individuals worldwide. It is a metabolic disorder characterized by elevated blood sugar levels over a prolonged period.
Undiagnosed or poorly managed diabetes can lead to serious complications including heart disease, kidney failure, vision loss, and nerve damage.

Given the increasing prevalence of diabetes worldwide, especially in low and middle-income countries, early diagnosis and prevention strategies have become more crucial than ever. Machine learning, a subset of artificial intelligence (AI), provides advanced tools to analyze large datasets and uncover hidden patterns that traditional statistical techniques might miss.

By employing machine learning techniques, healthcare providers can not only predict diabetes but also identify the key factors influencing disease progression. These insights are vital for designing personalized treatment plans, optimizing healthcare delivery, and improving patient outcomes.

The role of data science in healthcare is expanding rapidly. Predictive modeling using electronic health records (EHRs), wearable devices, and diagnostic reports is helping bridge the gap between reactive and proactive healthcare. Early intervention, enabled through accurate predictions, can save lives and reduce the long-term costs of chronic disease management.

This project demonstrates how even basic machine learning models like Logistic Regression can significantly impact early detection and decision-making in medical diagnostics.

Methodology

The project workflow includes the following steps:

1. Data Upload: The user uploads a medical record dataset in Excel format.

2. Data Preprocessing: We check for missing values and separate the features from the target label ('Outcome').

3. Data Splitting: We split the data into training and testing sets (80%-20%).

4. Feature Scaling: We standardize the feature values using StandardScaler.

5. Model Training: A Logistic Regression classifier is trained on the training data.

6. Model Prediction: The trained model is used to predict the labels of the test set.

7. Evaluation: We use a confusion matrix, and calculate accuracy, precision, and recall to assess model performance.

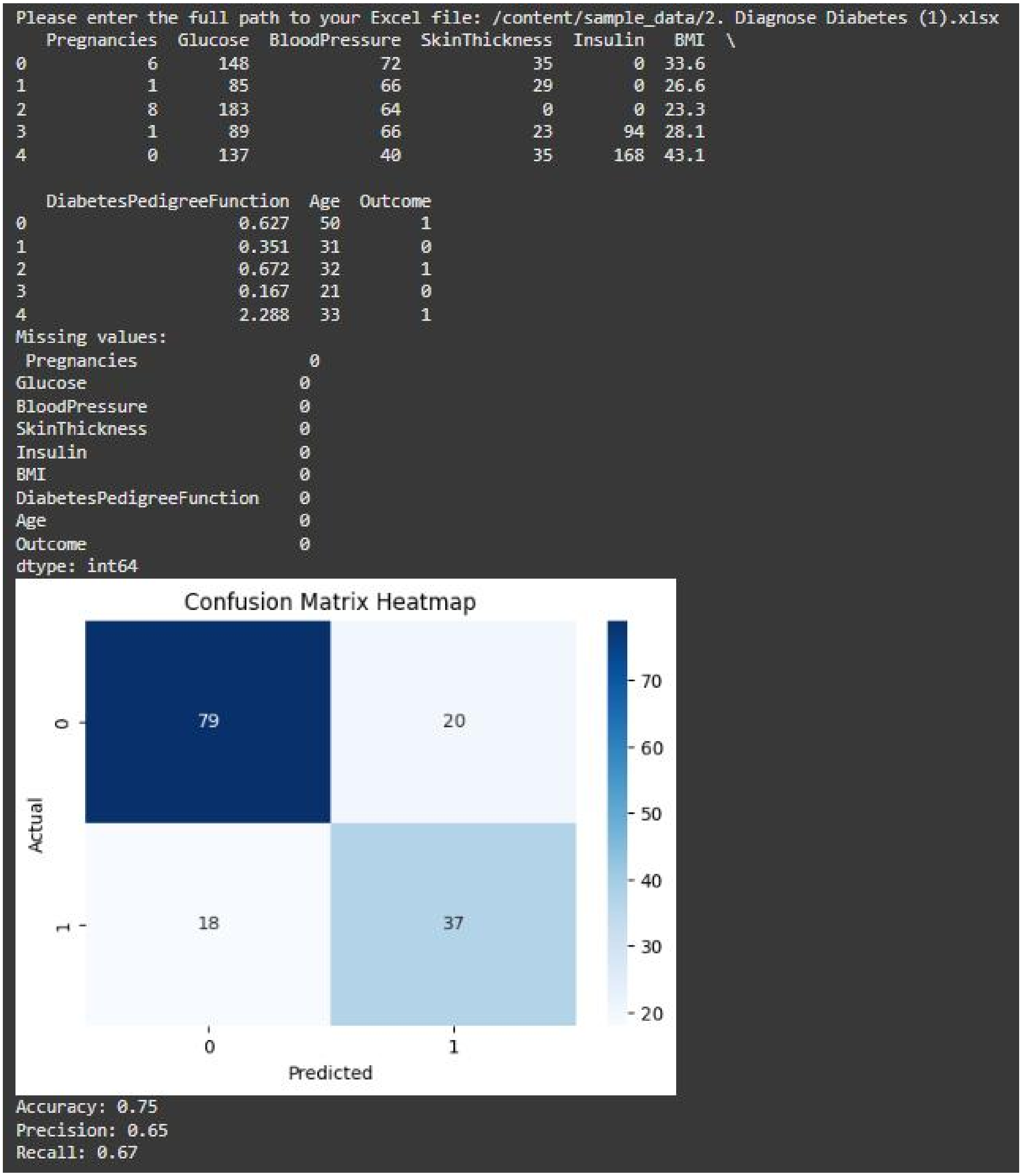
8. Visualization: A heatmap of the confusion matrix is generated to visualize model performance.

CODE

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score

# Ask user to input the file path
file_path = input("Please enter the full path to your Excel file: ")
try:
    # Load the dataset
    df = pd.read_excel(file_path)
    # Display the first few rows
    print(df.head())
    # Check for missing values
    print("Missing values:\n", df.isnull().sum())
    # Let's assume the target column is named 'Outcome'
    # Separate features and target
    X = df.drop(columns=['Outcome'])  # Feature columns
    y = df['Outcome']                # Target column
    # Train/test split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
    # Feature scaling
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)
    # Train a Logistic Regression model
    model = LogisticRegression()
    model.fit(X_train, y_train)
    # Predictions
    y_pred = model.predict(X_test)
    # Confusion matrix
    cm = confusion_matrix(y_test, y_pred)
    # Plot heatmap
    plt.figure(figsize=(6,4))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
    plt.title('Confusion Matrix Heatmap')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
    # Evaluation metrics
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    print(f"Accuracy: {accuracy:.2f}")
    print(f"Precision: {precision:.2f}")
    print(f"Recall: {recall:.2f}")
except Exception as e:
    print(f"An error occurred: {e}")
```

OUTPUT/RESULT



```
Please enter the full path to your Excel file: /content/sample_data/2. Diagnose Diabetes (1).xlsx
   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
Missing values:
 Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

```
Accuracy: 0.75
Precision: 0.65
Recall: 0.67
```

REFERENCES/CREDIT

-Dataset: The CSV file is uploaded from Kaggle

- Scikit-learn documentation: https://scikit-learn.org/stable/

- Pandas documentation: https://pandas.pydata.org/docs/

- Matplotlib documentation: https://matplotlib.org/stable/index.html

- Seaborn documentation: https://seaborn.pydata.org/

- WHO - Diabetes Facts: https://www.who.int/news-room/fact-sheets/detail/diabetes