KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning
www.kiet.edu
Delhi-NCR, Ghaziabad

An Assessment Report

on

“Diagnose Diabetes”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY
DEGREE

SESSION 2024-25

in

CSE(AI&ML)

By

Name:- Ram Ji

University Roll no:- 202401100400152

Under the supervision of

“Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

Diabetes Classification using Machine Learning

1. Introduction

Diabetes is a chronic medical condition affecting millions of people worldwide. Early diagnosis is essential to managing the disease and preventing complications. In this project, we use machine learning techniques to classify whether a patient has diabetes based on medical data. We use the Pima Indians Diabetes Dataset, which is a well-known dataset from the UCI Machine Learning Repository containing diagnostic measurements for female patients of Pima Indian heritage aged 21 and above.

2. Objective

To develop a machine learning model that can accurately classify individuals as diabetic or non-diabetic using the Pima Indians Diabetes Dataset.

3. Methodology

- Data Preprocessing:
  - Load the dataset.
  - Replace zeroes in certain columns with the mean of those columns.
  - Split the data into training and test sets.
  - Normalize feature values using StandardScaler.
- Model Selection:
  - We use the Random Forest Classifier for classification.
- Model Evaluation:
  - Evaluate the model using metrics like accuracy, precision, recall, and F1-score.
  - Visualize the confusion matrix.

4. Code

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv("diabetes.csv")

# Check for missing values (0 in some columns is considered missing)
cols_with_zero = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
for col in cols_with_zero:
    df[col] = df[col].replace(0, np.nan)
    df[col].fillna(df[col].mean(), inplace=True)  # Fill with mean

# Features and Target
X = df.drop("Outcome", axis=1)
y = df["Outcome"]

# 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)

# Model: Random Forest Classifier
```

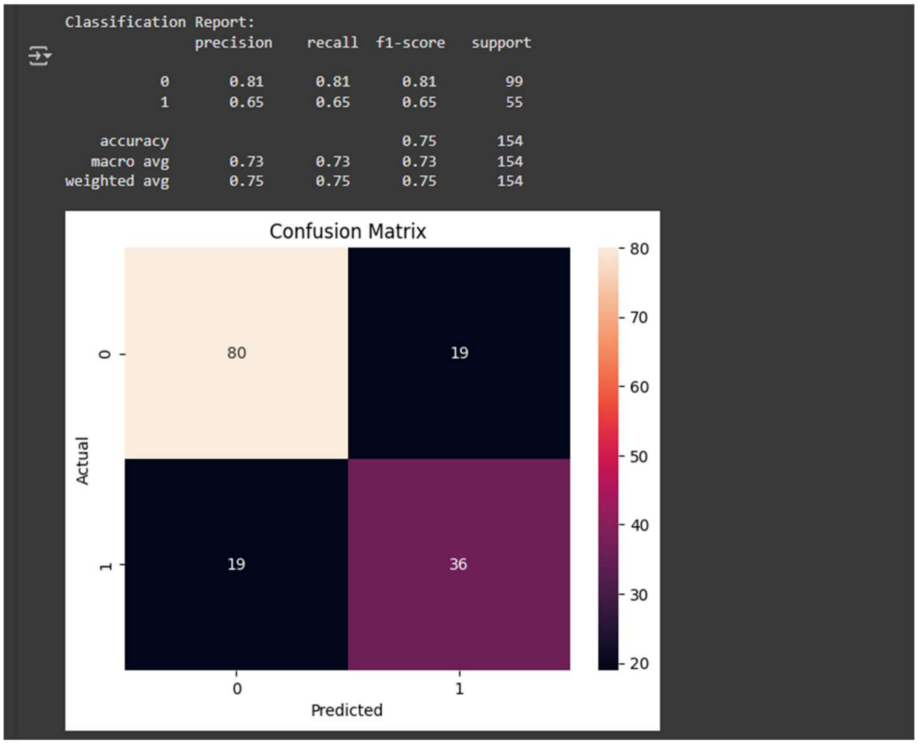
```
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluation
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))

# Confusion Matrix
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d')
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

5. Results



```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.81      0.81        99
           1       0.65      0.65      0.65        55

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.75      0.75      0.75       154
```

6. References / Credits

- Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

- Python Programming Language