www.kiet.edu
Delhi-NCR, Ghaziabad

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

Assessment Report

on

“DIABETES PREDICTION MODEL”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AI&ML)

By

Name :Harsh Vardhan Singh (202401100400093)
Nikhil Singh (202401100400128)
Naveen Kumar (202401100400122)
Nilesh Singh Yadav (202401100400129)
MaksudurRahman (202401100400117)

Section: B

Under the supervision of

“Mr.Abhishek Shukla sir”

KIET Group of Institutions, Ghaziabad

May, 2025

---

1. Introduction

What is Diabetes?

->>Diabetes is a chronic health condition that occurs either when the pancreas fails to produce enough insulin or when the body becomes resistant to insulin. It leads to elevated blood glucose levels, which can cause severe complications if not diagnosed and managed early.

- The objective of this project is to classify whether someone has diabetes or not.
- Dataset consists of several Medical Variables(Independent) and one Outcome Variable(Dependent)
- The independent variables in this data set are :-'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin','BMI', 'DiabetesPedigreeFunction', 'Age'
- The outcome variable value is either 1 or 0 indicating whether a person has diabetes(1) or not(0).

---

2. Problem Statement

Create a classification model to predict the likelihood of diabetes based on health data. Visualize model accuracy. Dataset Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

---

3. Objectives

- Preprocess the dataset for training a machine learning model.
- Train a Logistic Regression model to predict diabetes.
- Evaluate model performance using standard classification metrics.
- Visualize the confusion matrix using a heatmap for interpretability.

---

4. Methodology

- **Data Collection**: The user uploads a CSV file containing the dataset.
- **Data Preprocessing**:
  - Handling missing values using mean and mode imputation.
  - One-hot encoding of categorical variables.
  - Feature scaling using StandardScaler.
- **Model Building**:
  - Splitting the dataset into training and testing sets.
  - Training a Logistic Regression classifier.
- **Model Evaluation**:
  - Evaluating accuracy, precision, recall, and F1-score.
  - Generating a confusion matrix.

---

5. Data Preprocessing

The dataset is cleaned and prepared as follows:

- Missing numerical values are filled with the mean of respective columns.
- Categorical values are encoded using one-hot encoding.
- Data is scaled using StandardScaler to normalize feature values.
- The dataset is split into 80% training and 20% testing.

---

6. Model Implementation

Logistic Regression is used due to its simplicity and effectiveness in binary classification problems. The model is trained on the processed dataset and used to predict the diabetes risk on the test set.

---

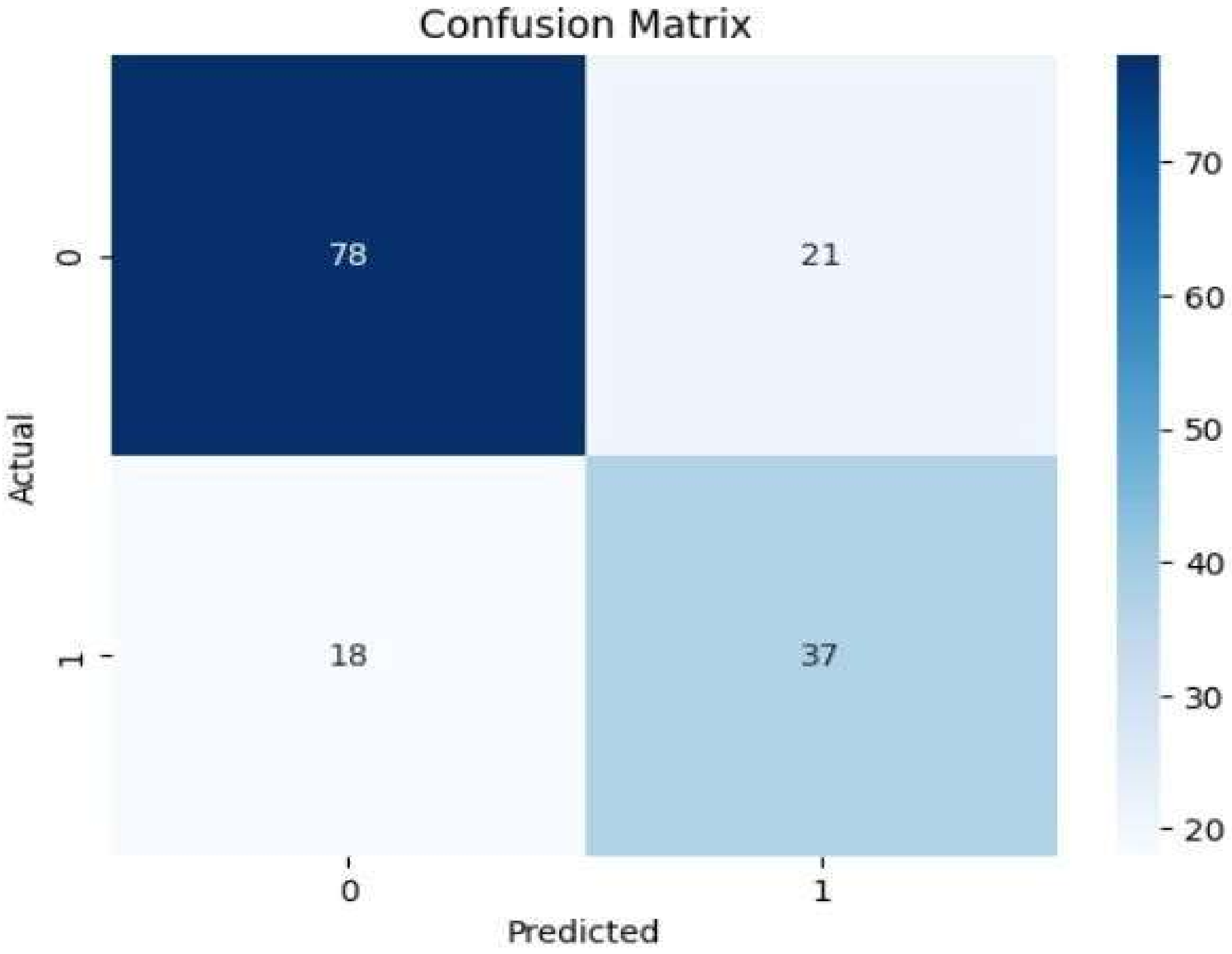
7. Evaluation Metrics

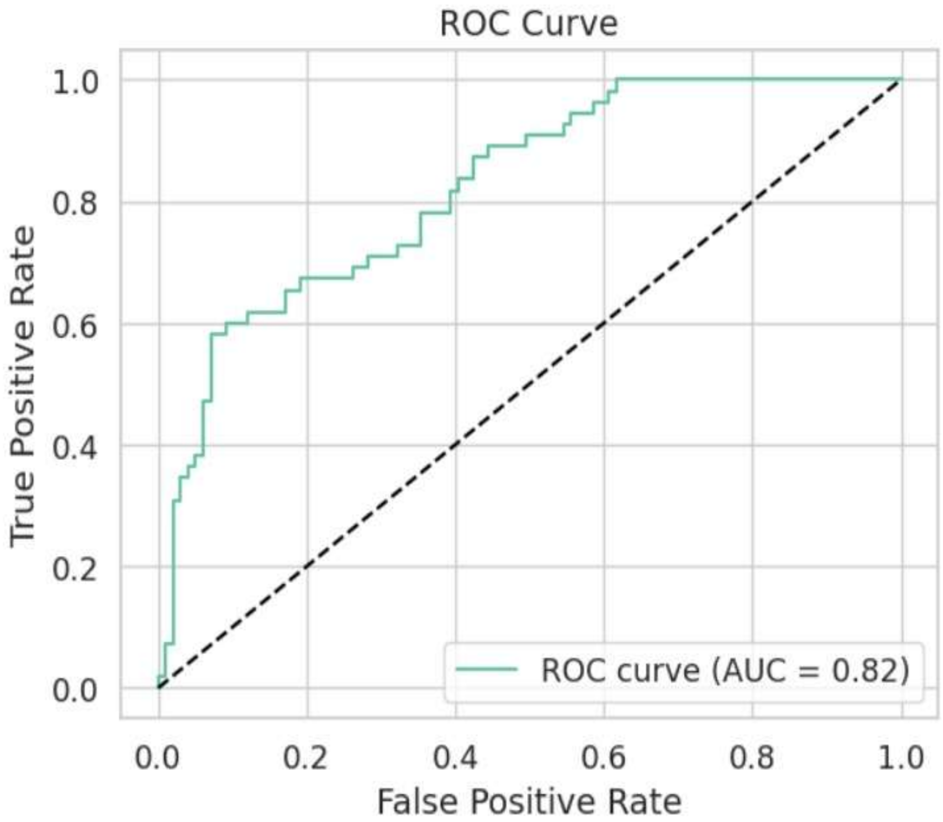
The following metrics are used to evaluate the model:

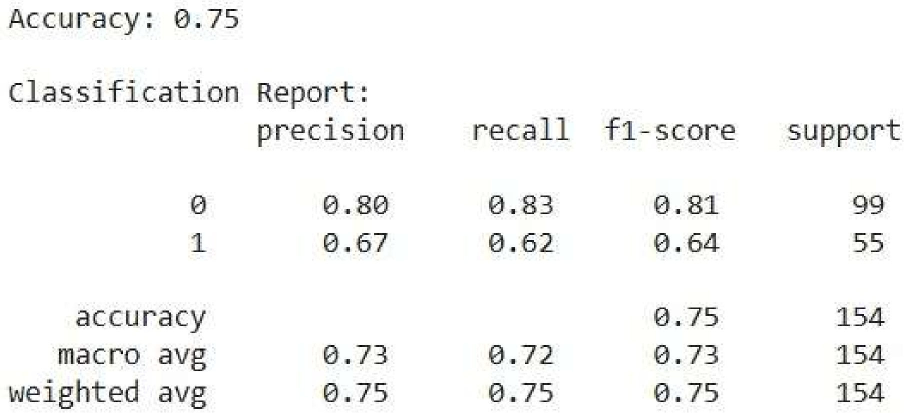
- **Accuracy**: Measures overall correctness.
- **Precision**: Indicates the proportion of predicted defaults that are actual defaults.
- **Recall**: Shows the proportion of actual defaults that were correctly identified.
- **F1 Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Visualized using Seaborn heatmap to understand prediction errors.

---

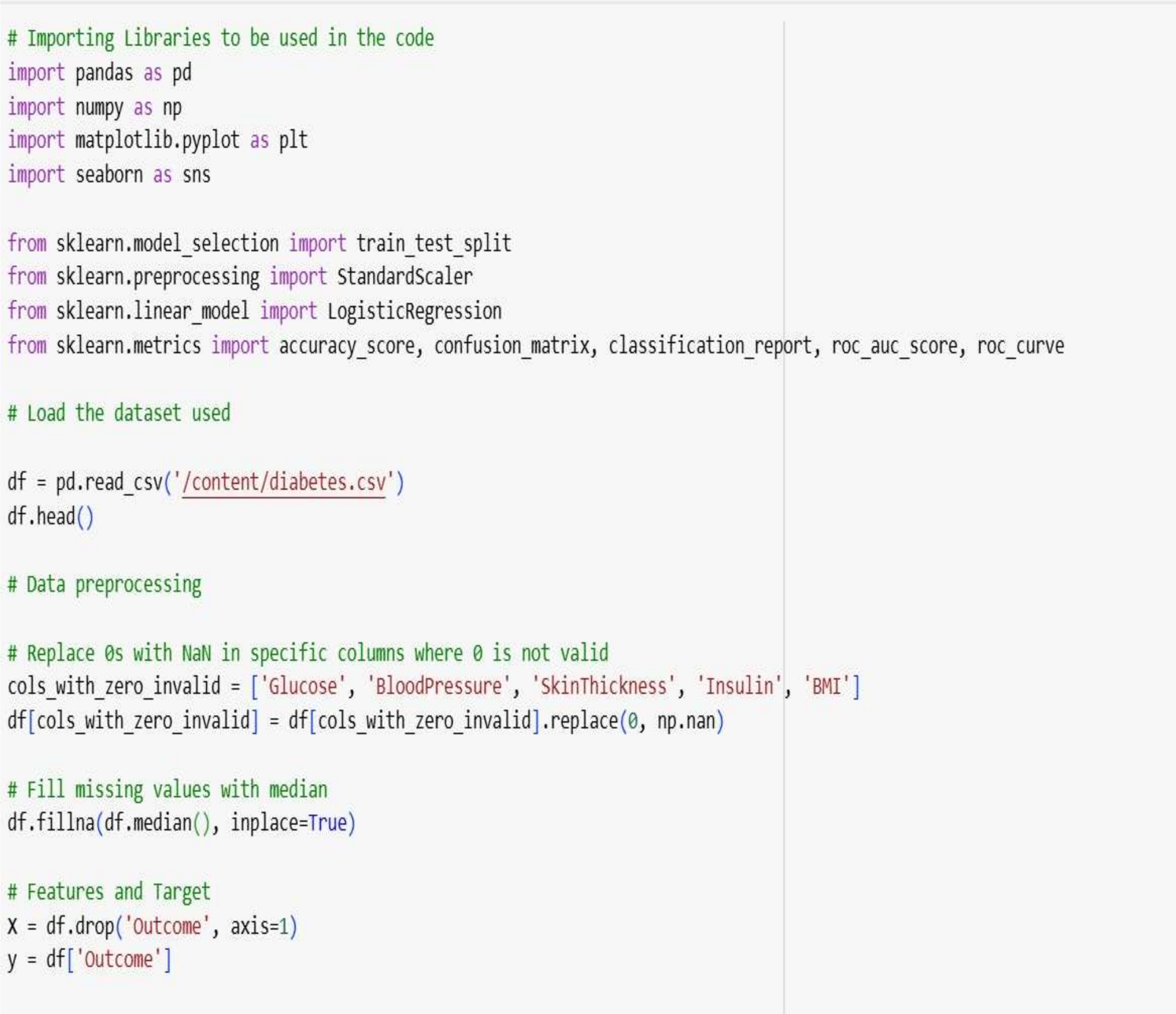
8. Results and Analysis



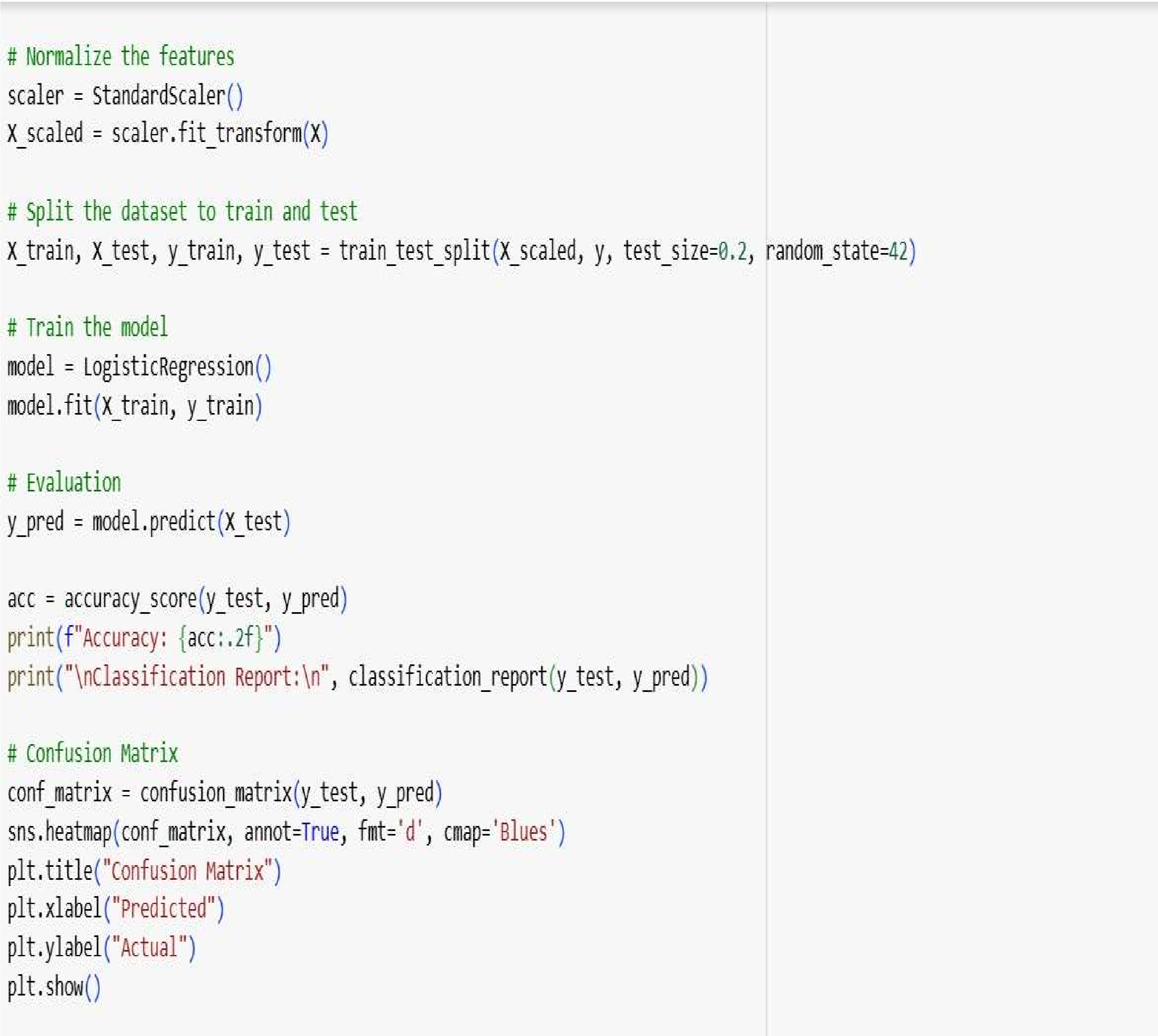




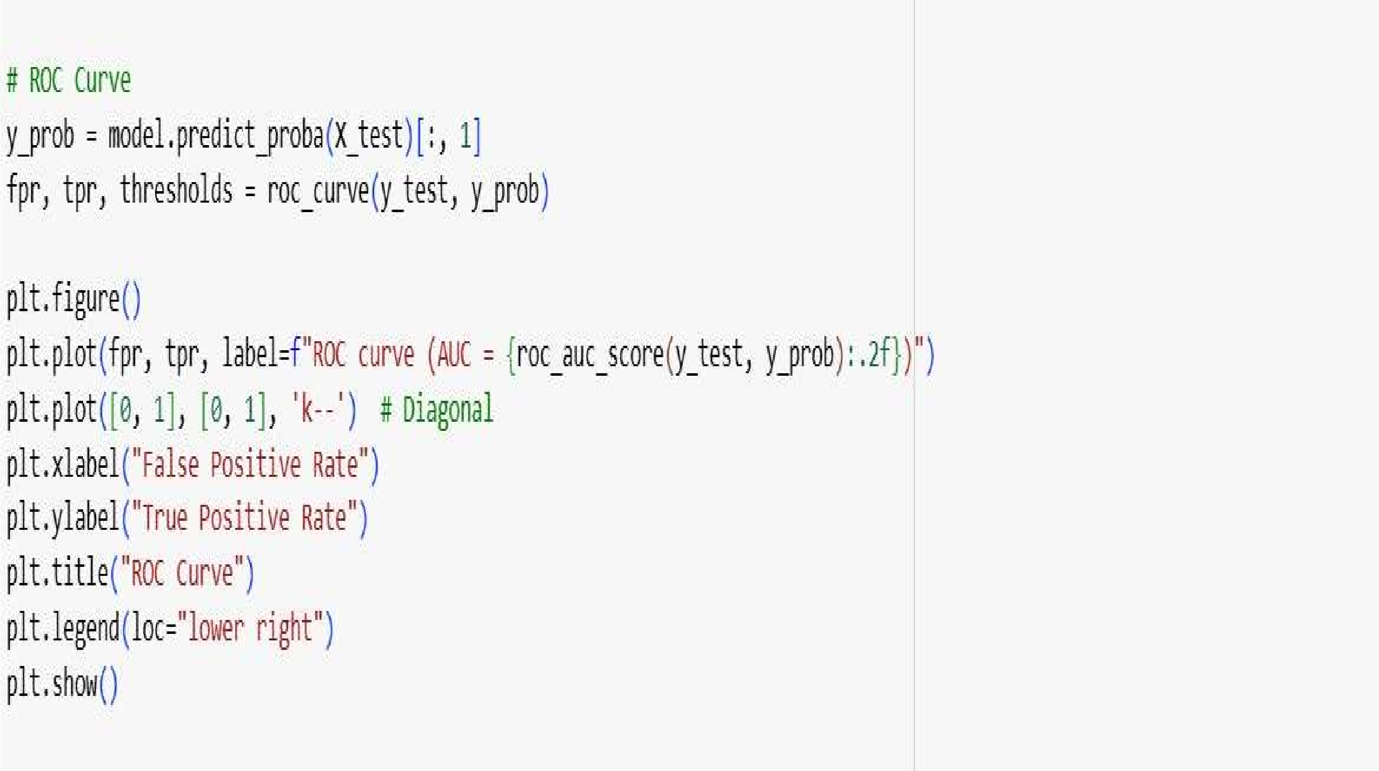
```
Accuracy: 0.75

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.83      0.81        99
           1       0.67      0.62      0.64        55

    accuracy                           0.75       154
   macro avg       0.73      0.72      0.73       154
weighted avg       0.75      0.75      0.75       154
```



```python
# Importing Libraries to be used in the code
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.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, roc_auc_score, roc_curve

# Load the dataset used

df = pd.read_csv('/content/diabetes.csv')
df.head()

# Data preprocessing

# Replace 0s with NaN in specific columns where 0 is not valid
cols_with_zero_invalid = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
df[cols_with_zero_invalid] = df[cols_with_zero_invalid].replace(0, np.nan)

# Fill missing values with median
df.fillna(df.median(), inplace=True)

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



```
# Normalize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split the dataset to train and test
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# Evaluation
y_pred = model.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print(f"Accuracy: {acc:.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred))

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



```
# ROC Curve
y_prob = model.predict_proba(X_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_prob)

plt.figure()
plt.plot(fpr, tpr, label=f"ROC curve (AUC = {roc_auc_score(y_test, y_prob):.2f})")
plt.plot([0, 1], [0, 1], 'k--')  # Diagonal
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.legend(loc="lower right")
plt.show()
```

9. Conclusion

Logistical regression is selected when the dependent variable is categorical, meaning they have binary outputs, such as "true" and "false" or "yes" and "no". The project demonstrates the potential of using machine learning for automating diabetes prediction and improving risk assessment. However, improvements can be made by exploring more advanced models and handling imbalanced data.

10. References

- scikit-learn documentation
- pandas documentation
- Seaborn visualization library
- Research articles on diabetes risk prediction