www.kiet.edu
Delhi-NCR, Ghaziabad

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

Assesment Report

on

“Diagnose Diabetes”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

Name of discipline

SUDEEKSHA SINGH

202401100300252

Under the supervision of

“ABHIKESH SHUKLA”

KIET Group of Institutions, Ghaziabad

Affiliated to

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

(Formerly UPTU)

May, 2025

INTRODUCTION

This project aims to use machine learning to diagnose whether a patient has diabetes based on various medical features. We use the popular Pima Indians Diabetes dataset, which includes parameters like glucose levels, BMI, age, and blood pressure. Our goal is to build a classification model and evaluate its performance.

Diabetes is a chronic health condition that affects how the body turns food into energy. If left undiagnosed or unmanaged, it can lead to serious health complications including heart disease, kidney failure, and nerve damage. Early detection is crucial to managing diabetes and improving patient outcomes.

This project focuses on leveraging machine learning techniques to classify whether an individual has diabetes based on a set of medical attributes. The dataset used for this analysis is the **Pima Indians Diabetes Dataset**, which contains diagnostic measurements for female patients of at least 21 years of age, including features such as glucose level, blood pressure, insulin levels, BMI, and age.

By applying a classification model—specifically a Random Forest Classifier—this project aims to accurately predict the presence of diabetes. The performance of the model is assessed using key evaluation metrics such as accuracy, precision, and recall. The results are visualized using a confusion matrix heatmap, providing insight into the effectiveness of the prediction model.

The goal is to demonstrate how machine learning can assist in medical diagnoses and provide decision support to healthcare professionals.

METHODOLOGY

The methodology of this project involves several key steps to develop a reliable and interpretable machine learning model for predicting diabetes diagnoses based on medical features. The process is outlined as follows:

3.1 Data Collection

The dataset used is the **Pima Indians Diabetes Dataset**, which consists of 768 records and 9 columns. The features include:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI (Body Mass Index)
- Diabetes Pedigree Function
- Age
- Outcome (target variable: 1 = diabetes, 0 = no diabetes)

3.2 Data Preprocessing

- The dataset was loaded using **Pandas**.
- No missing values were found, so minimal cleaning was required.
- The data was divided into **features (X)** and the **target (y)**.
- It was then split into **training and testing sets** using an 80/20 ratio to ensure fair evaluation.

3.3 Model Selection

- A **Random Forest Classifier** from the **scikit-learn** library was selected for training. This algorithm was chosen due to its robustness, ability to handle feature interactions, and its effectiveness in binary classification tasks.

3.4 Model Training and Prediction

- The model was trained using the training data.
- Predictions were made on the test data to evaluate model performance.

3.5 Evaluation Metrics

To assess the quality of the model's predictions, the following evaluation metrics were computed:

- **Accuracy**: The proportion of total correct predictions.
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positives correctly identified by the model.

A **confusion matrix** was also generated and visualized using a **heatmap** to provide a more detailed view of model performance.

3.6 Visualization

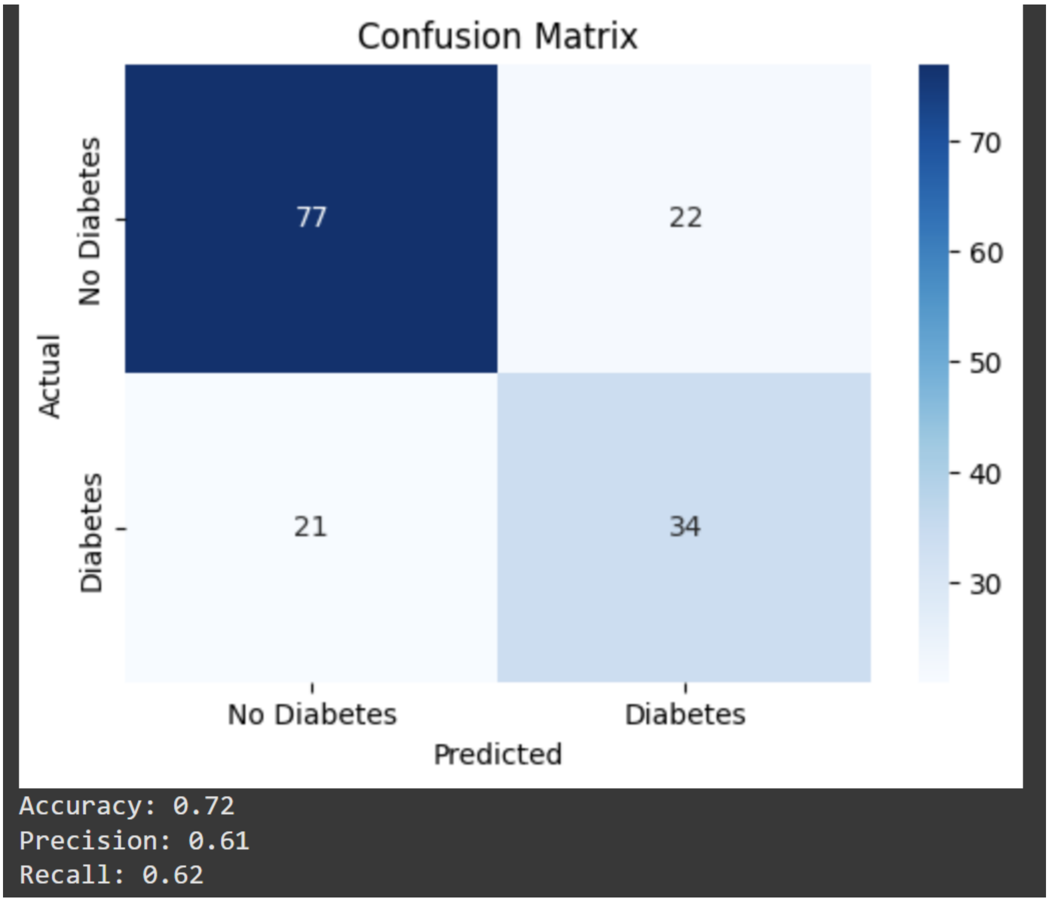
- The confusion matrix was visualized using the **Seaborn** and **Matplotlib** libraries to help interpret the prediction results and assess where the model performs well or needs improvement.

CODE

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("/content/2. Diagnose Diabetes.csv")

# Split into 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)

# Train Random Forest Classifier
clf = RandomForestClassifier(random_state=42)
clf.fit(X_train, y_train)
```

```
# Predict on test set
y_pred = clf.predict(X_test)

# Confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Heatmap
plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Diabetes', 'Diabetes'], yticklabels=['No Diabetes', 'Diabetes'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
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}")
```

OUTPUT



```
Accuracy: 0.72
Precision: 0.61
Recall: 0.62
```