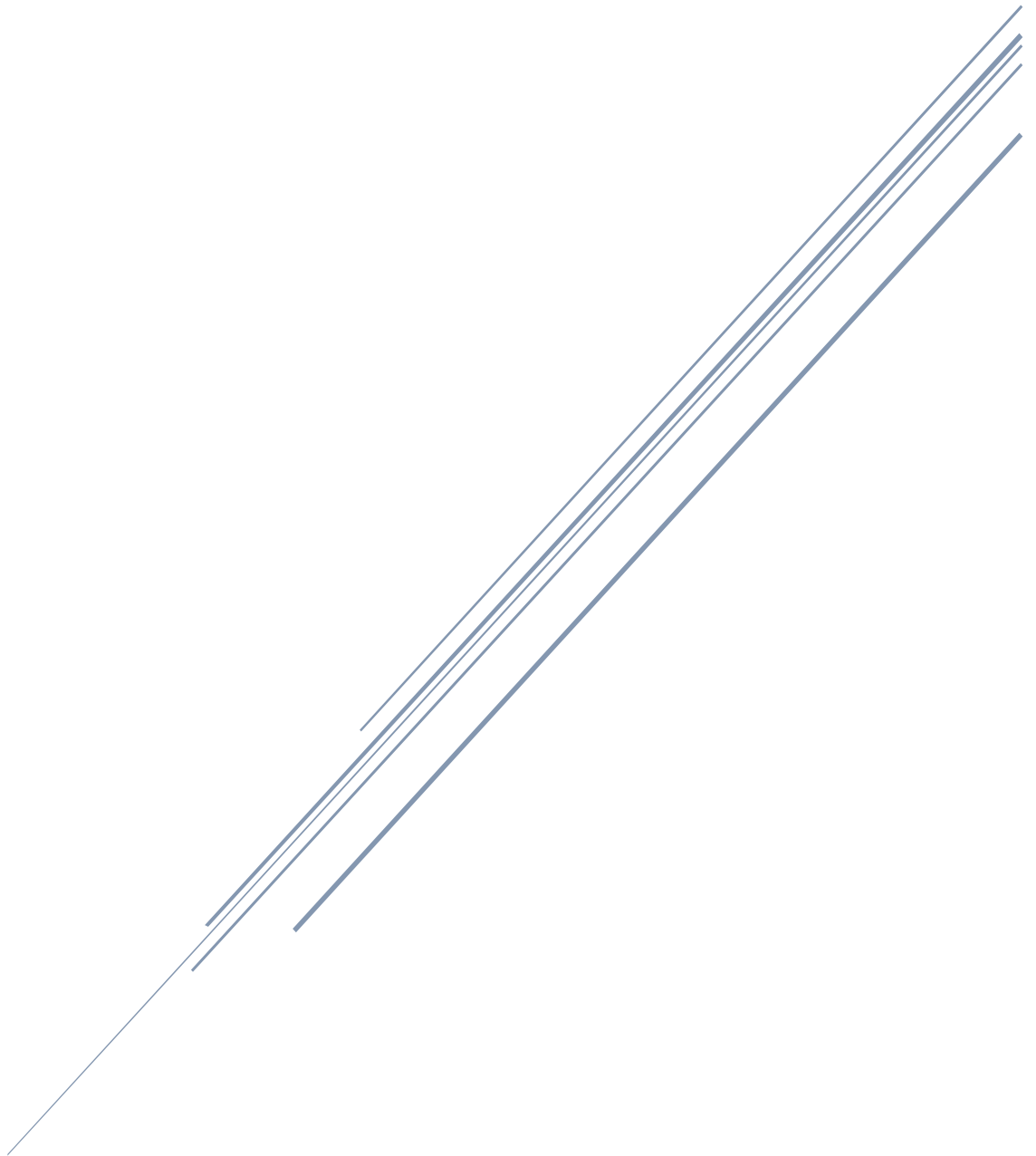


DIABETES PREDICTOR

USING SCIKIT-LEARN AND STREAMLIT



ANIRUDH RAMESH

Project Report: Diabetes Predictor

By: Anirudh Ramesh

Introduction:

The Diabetes Predictor is a machine learning-based application developed using the Streamlit framework. It aims to predict the likelihood of an individual developing diabetes based on various input parameters such as gender, age, hypertension, heart disease, BMI, HbA1c level, and blood glucose.

Objective:

The main objective of the Diabetes Predictor is to provide users with an estimation of their risk of developing diabetes, allowing them to take necessary precautions and seek medical advice if needed.

Methodology:

The application utilizes two machine learning algorithms for prediction: Logistic Regression and Random Forest. The steps involved in the methodology are as follows:

a. Data Preparation:

The application imports the required libraries, including pandas, scikit-learn, and Streamlit.

It reads the dataset from a CSV file (attached to the repository), which contains historical data related to diabetes.

The dataset is preprocessed by dropping irrelevant columns and performing one-hot encoding for the 'gender' feature.

b. Model Training and Evaluation:

The dataset is divided into training and testing sets using the `train_test_split` function.

The Logistic Regression model is trained on the training data using the 'newton-cg' solver.

The Random Forest model is trained on the training data.

The accuracy of both models is evaluated using the score function on the testing data.

c. Prediction:

The user provides input through the Streamlit interface, including gender, age, hypertension, heart disease, BMI, HbA1c level, and blood glucose.

The Logistic Regression model predicts the likelihood of diabetes using the provided input.

The Random Forest model also predicts the likelihood of diabetes.

The prediction values are taken and a color-coded result is printed in the UI. The result could be Low, Medium and High

The predicted probabilities of both models are displayed to the user in the application sidebar.

Results:

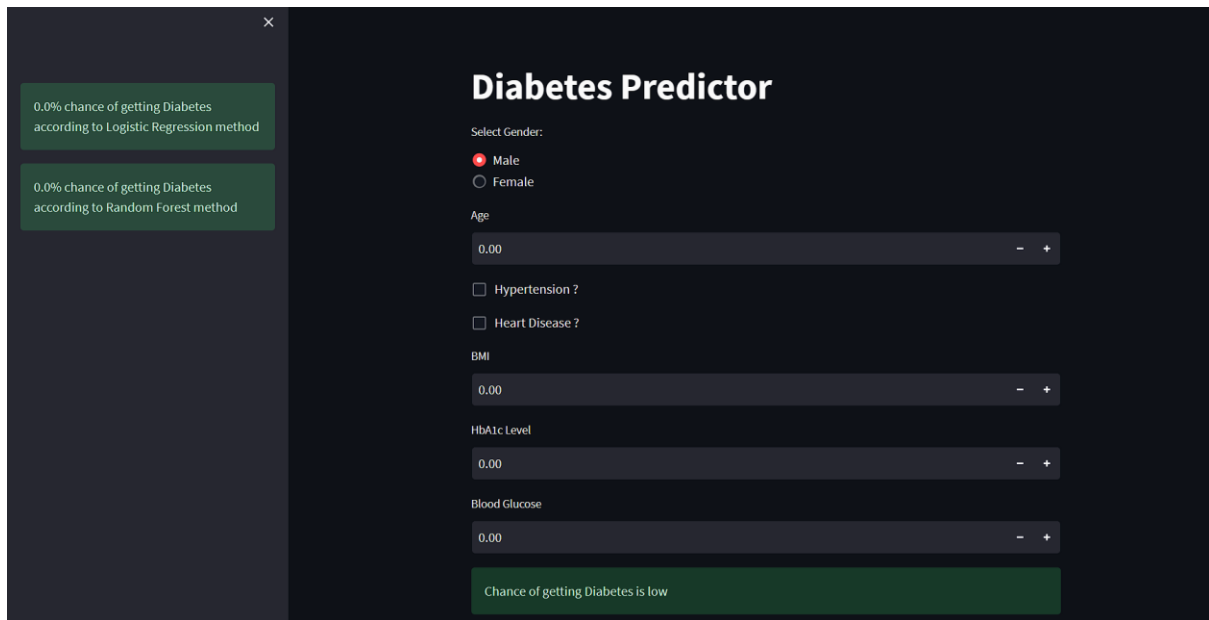
The accuracy scores of the trained models have been found out as 0.958 for the Logistic Regression model and 0.969 for the Random Forest Model. Also, predicted percentages displayed in the Streamlit interface for a more user-friendly experience.

Conclusion:

The Diabetes Predictor application provides an easy and convenient way for individuals to assess their risk of developing diabetes. By inputting relevant information, users can instantly receive predictions

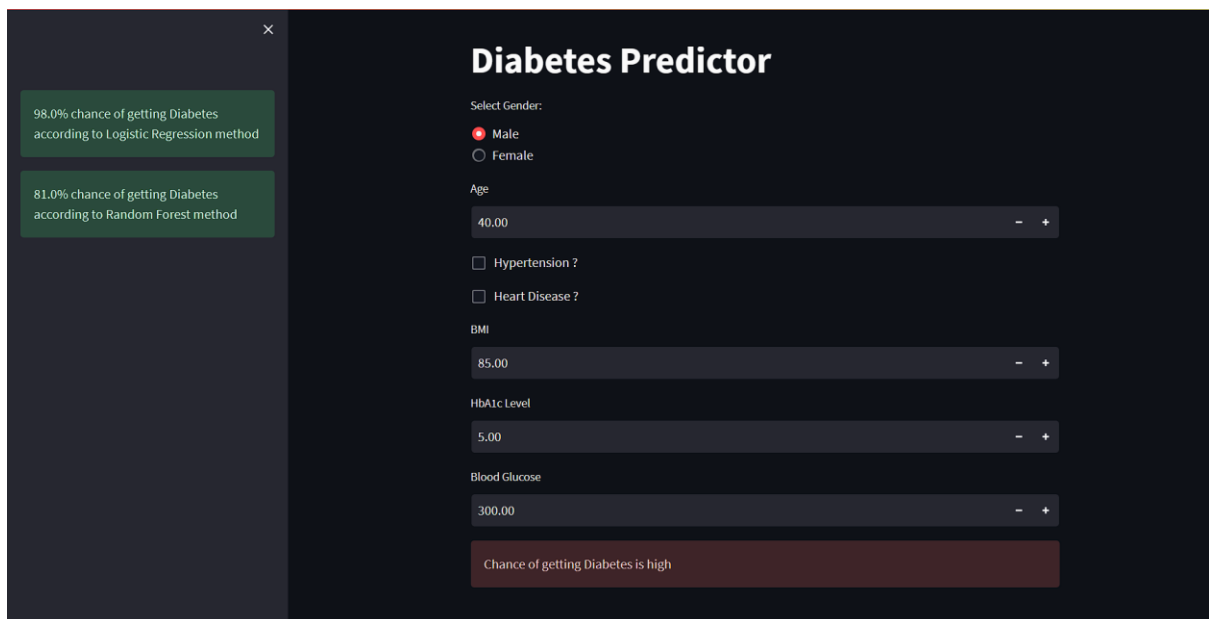
based on two different machine learning algorithms. However, the accuracy, though appreciable, is not perfect and therefore, can be improved. The application could also be further enhanced by including additional features, improving the user interface, and conducting more extensive evaluation of the models' performance.

Screenshots:



The screenshot displays the 'Diabetes Predictor' interface. On the left, a sidebar shows two green boxes: '0.0% chance of getting Diabetes according to Logistic Regression method' and '0.0% chance of getting Diabetes according to Random Forest method'. The main area contains the following inputs: 'Select Gender' with 'Male' selected, 'Age' set to 0.00, 'Hypertension?' and 'Heart Disease?' both unchecked, 'BMI' set to 0.00, 'HbA1c Level' set to 0.00, and 'Blood Glucose' set to 0.00. A green button at the bottom indicates 'Chance of getting Diabetes is low'.

Figure 1 - Basic UI



The screenshot displays the 'Diabetes Predictor' interface with sample data. The sidebar shows two green boxes: '98.0% chance of getting Diabetes according to Logistic Regression method' and '81.0% chance of getting Diabetes according to Random Forest method'. The main area contains the following inputs: 'Select Gender' with 'Male' selected, 'Age' set to 40.00, 'Hypertension?' and 'Heart Disease?' both unchecked, 'BMI' set to 85.00, 'HbA1c Level' set to 5.00, and 'Blood Glucose' set to 300.00. A red button at the bottom indicates 'Chance of getting Diabetes is high'.

Figure 2 - Prediction of High Chance (Random Sample Data)

×

63.800000000000004% chance of getting Diabetes according to Logistic Regression method

7.000000000000001% chance of getting Diabetes according to Random Forest method

Diabetes Predictor

Select Gender:

☒ Male
☐ Female

Age

40.00 - +

☐ Hypertension ?
☐ Heart Disease ?

BMI

85.00 - +

HbA1c Level

5.00 - +

Blood Glucose

200.00 - +

Chance of getting Diabetes

Figure 3 - When Prediction Percentages vary heavily, Warning Given