MC106: MATLAB Programming (P7)

Diabetes Prediction

Submitted by-

Rishabh Ranjan Singh

24/B06/019

Submitted to-

Prof. C.P. Singh

Mr. Jamkhongam Touthang

Dr. Nitika Sharma

Mr. Lokesh Chander

Date: 07/04/2025

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY, SHAHBAD DAULATPUR (FORMERLY DELHI COLLEGE OF ENGINEERING) MAIN BAWANA ROAD, Rohini, Delhi 110042

Diabetes Prediction using Machine Learning and Data Science

1. Abstract

This MATLAB-based application is a Diabetes Predictor that integrates machine learning with a clean, interactive Graphical User Interface (GUI) for efficient health assessments. Built on the Pima Indians Diabetes dataset, it uses a Bagged Ensemble Classifier to predict the likelihood of diabetes based on user-provided inputs. Key details such as age, gender, glucose, blood pressure, insulin, weight, and height are collected, with an additional field for pregnancies in female users. The system auto-calculates BMI and normalizes all inputs before making a prediction. Results are displayed with clear, color-coded messages for easy interpretation. If the outcome is uncertain, the app prompts for family history to enhance assessment. With smooth navigation, error checks, and a refresh option, the tool offers a simple yet powerful user experience.

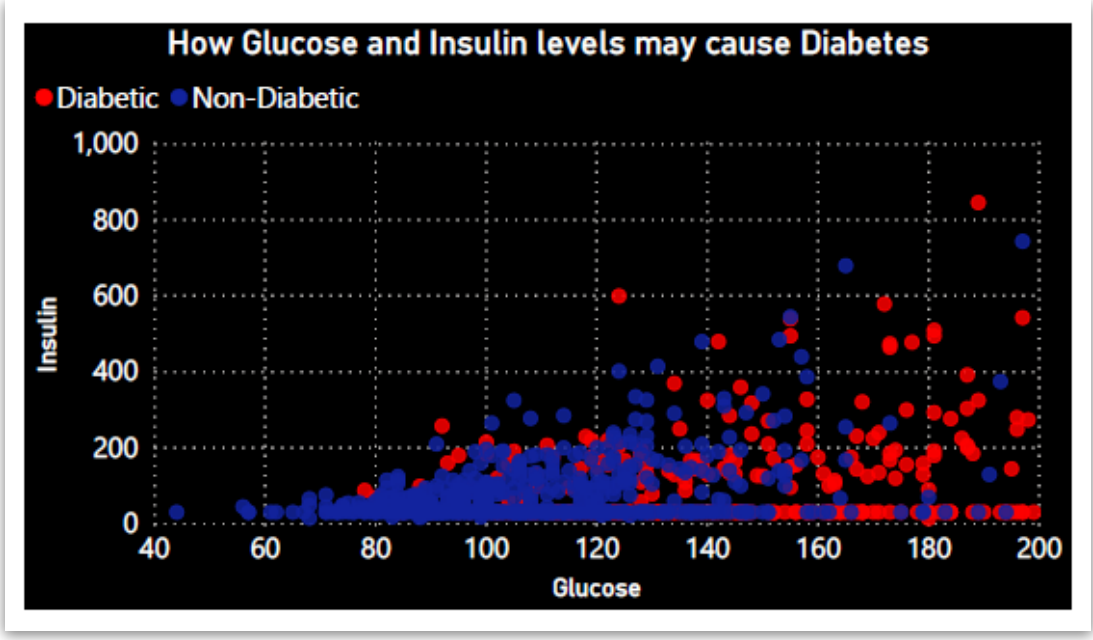
2. Introduction

Diabetes is a chronic health condition that affects how the body processes blood sugar (glucose), and if left undiagnosed, it can lead to serious complications. Early detection is crucial for effective management and prevention of long-term damage to vital organs. However, identifying diabetes at an early stage can be challenging without proper medical testing and expert analysis. With the rise of data-driven solutions, machine learning offers powerful tools to assist in early diagnosis using readily available health indicators. This project addresses the need for a quick, accessible, and reliable method to predict diabetes risk based on basic patient information.

3. Problem Statement

Diabetes is a chronic metabolic disorder that affects millions globally and is becoming increasingly common due to sedentary lifestyles and poor dietary habits. One of the major challenges is that diabetes often remains undiagnosed in its early stages, as symptoms can be mild or go unnoticed. Delayed diagnosis can lead to severe complications such as heart disease, kidney failure, nerve damage, and vision problems.

- **Early Risk Identification:** A diabetes predictor can flag individuals at high risk before symptoms appear, encouraging early lifestyle changes or medical consultation.
- **Cost-Effective Screening:** Predictive models minimise the need for frequent lab testing and can act as a first-line screening tool.
- **Data-Driven Insights:** Using machine learning or statistical analysis, predictors can analyse various factors like age, BMI, blood pressure, and glucose levels to assess risk with high accuracy.
- **Scalable & Accessible:** A digital diabetes predictor can be integrated into mobile apps or health platforms, making it easy to use anywhere, especially in areas with limited medical infrastructure.
- **Support for Clinicians:** Helps doctors prioritise patients for further testing or monitoring, improving efficiency in healthcare delivery.



4. Dataset Description

The model uses the **PIMA Indian Diabetes Dataset** available publicly. It consists of 768 observations and 9 variables:

- Pregnancies
- Glucose
- Blood Pressure (Diastolic)
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- Outcome (0 = Non-diabetic, 1 = Diabetic)

diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0

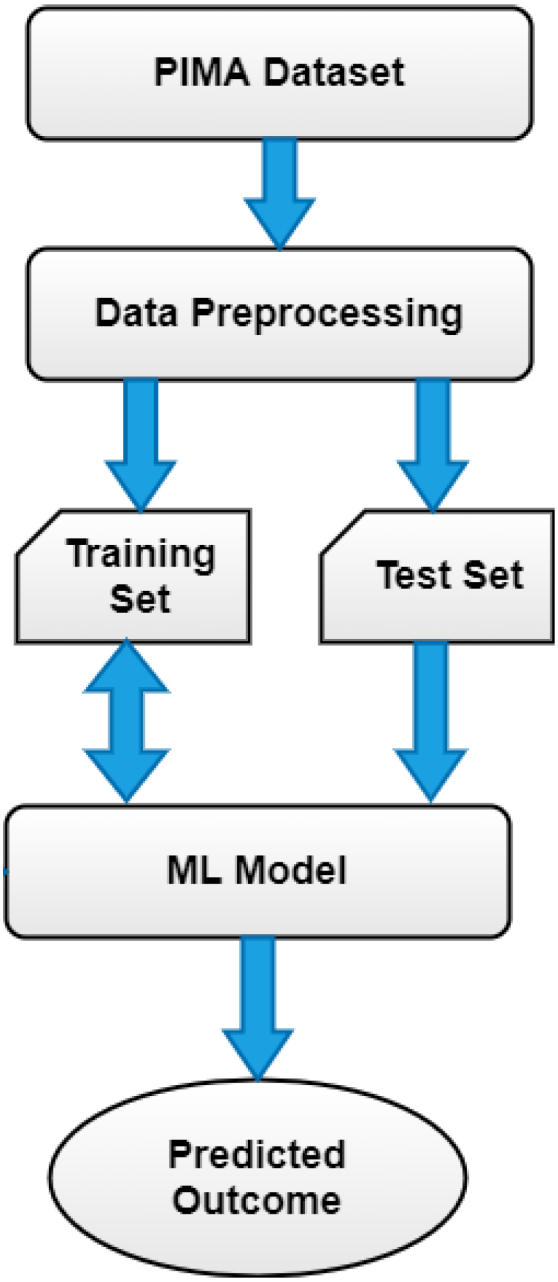
5. Methodology

5.1 Data Preprocessing

Missing values in fields such as Insulin and Skin Thickness are replaced with average values. Feature normalisation(mean subtraction and division by standard deviation) is applied to training data.

5.2 Model Selection

A Random Forest classifier (Bagged Trees) was chosen for its robustness and performance on tabular data. It handles overfitting better and gives good accuracy. The model was trained using 70% of the dataset.



5.3 Feature Engineering

BMI is computed from user-provided weight and height in the GUI. Skin Thickness, Diabetes Pedigree Function, and in some cases, Insulin and Pregnancies are substituted with average values for simplicity.

5.4 GUI Design

The GUI is divided into three pages:

- Page 1: Name, Age, Gender Input
- Page 2: Glucose, Blood Pressure, Weight, Height, Insulin, Pregnancies (if Female)
- Page 3: Optional Family History dropdown

Predicted result and calculated BMI are then displayed.

6. Implementation

The core of the project is implemented in a MATLAB script (`DiabetesPredictorUI`). It loads the dataset, preprocesses the data, trains a Random Forest model, and builds the GUI.

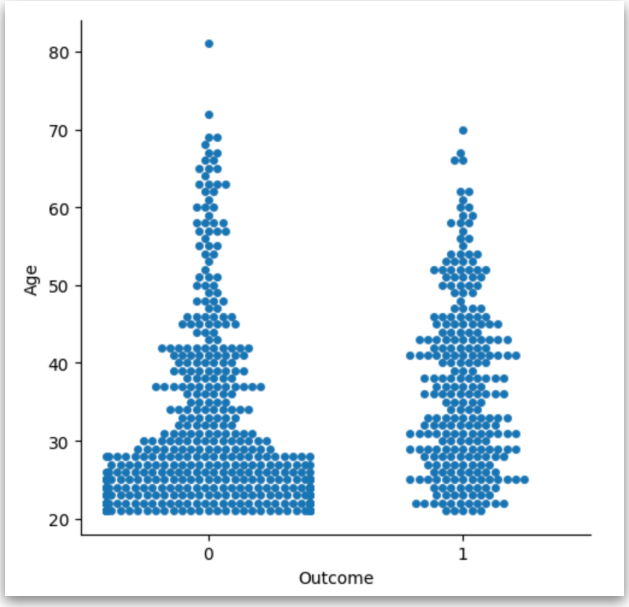
Each user input is processed, validated, and fed into the model for prediction. The prediction result is displayed with risk levels (Low/High) depending on glucose and model confidence.

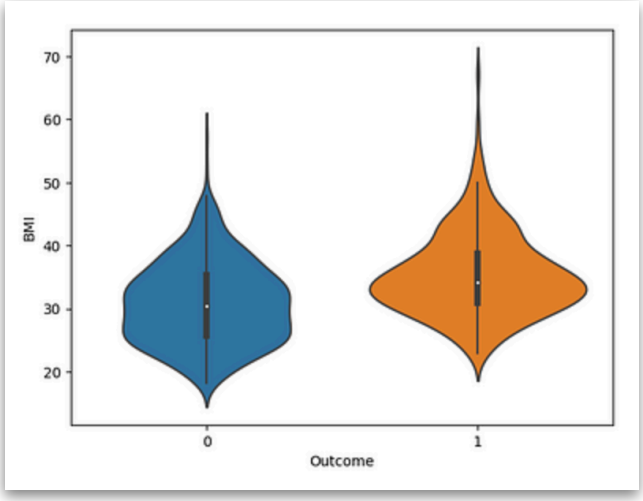
6.1 Softwares & Tools

- **MATLAB:** Used as the core platform for coding, model training, and GUI development.
- **Statistics and Machine Learning Toolbox:** For implementing machine learning algorithms like Random Forest (fitcensemble).
- **MATLAB GUI Builder (uicontrol):** To design the user interface for input collection and prediction display.

6.2 Algorithm Flow Diagram

- **Flowchart:** A detailed flow diagram, various Catplots, Violinplots outlines the sequential steps from data acquisition and preprocessing through segmentation, feature extraction, and Normalisation.





7. Working of Model

7.1 Data Loading & Preprocessing

- The model starts by loading the Pima Indians Diabetes dataset (diabetes.csv), which contains medical data of patients, including features like glucose level, blood pressure, insulin, BMI, age, and whether they have diabetes (Outcome column).
- The data is split into features (inputs) and labels (outcomes). It uses a 70-30 train-test split with cvpartition, although only the training data is used for model building.

7.2 Normalisation

- To ensure that all features contribute equally, the training features are normalized using Z-score normalization: [(X - μ) / σ], where mu is the mean and sigma is the standard deviation of each feature.

```
X_train = (X_train - mu) ./ sigma;
X_test = (X_test - mu) ./ sigma;
```

7.3 Model Training

Powered by a forest, not just a tree.

This bagged ensemble model builds 100 decision trees on random data slices and feature splits, combining their votes for stable, accurate diabetes predictions. Input normalisation keeps performance sharp and overfitting in check.

```
% Train model
model = fitcensemble(X_train, y_train, 'Method', 'Bag', 'NumLearningCycles', 100);
```

7.4 Designing of GUI

The GUI for this diabetes predictor is built using MATLAB's App Designer-style approach with figure, uipanel, and uicontrol elements to create an interactive, multi-page interface. It consists of three main panels that represent different stages of user input and result display.

- Created using figure, with fixed size and a custom title.
- Acts as the container for all GUI components.

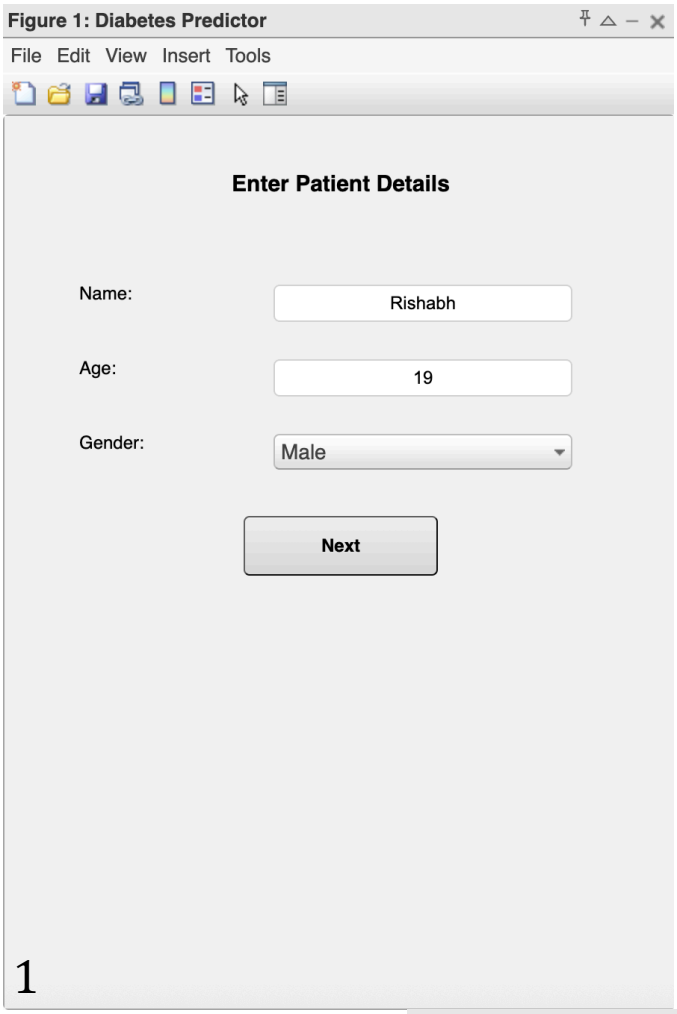


Figure 1: Diabetes Predictor
File Edit View Insert Tools
Enter Patient Details
Name:
Rishabh
Age:
19
Gender:
Male
Next
1

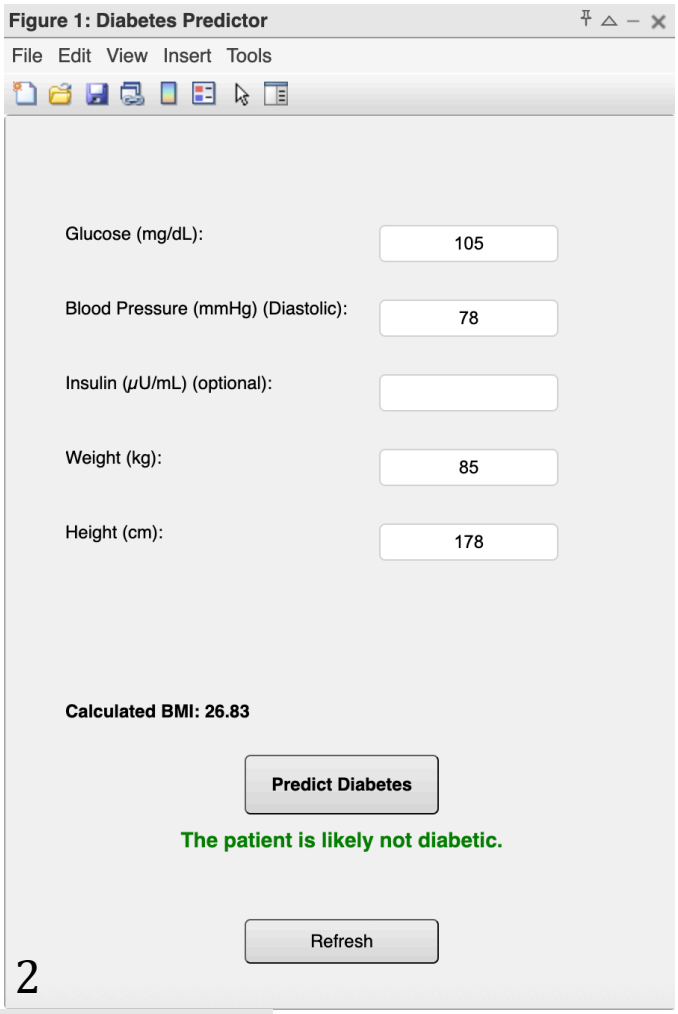


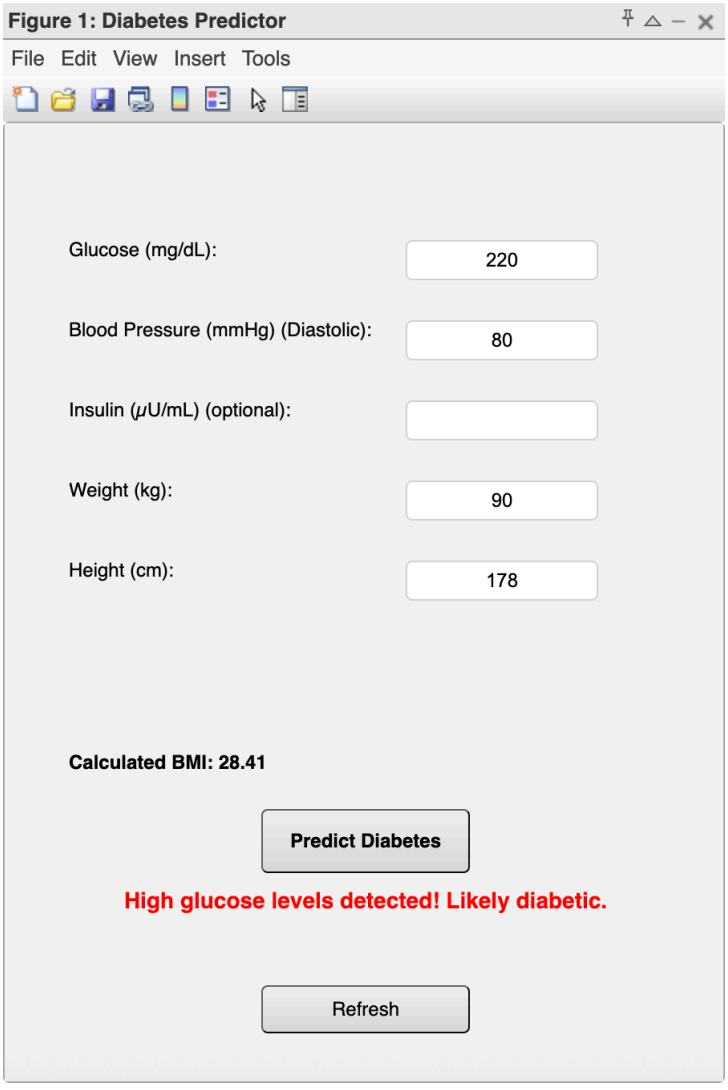
Figure 1: Diabetes Predictor
File Edit View Insert Tools
Glucose (mg/dL):
105
Blood Pressure (mmHg) (Diastolic):
78
Insulin (μU/mL) (optional):
Weight (kg):
85
Height (cm):
178
Calculated BMI: 26.83
Predict Diabetes
The patient is likely not diabetic.
Refresh
2


Figure 1: Diabetes Predictor
File Edit View Insert Tools
Any Family History of Diabetes?
Yes
You are not diabetic yet, but a family history indicates potential risk. Please stay cautious.
Enter
Start Over
3

8. Conclusion

This project highlights how machine learning, combined with a simple GUI, can make early diabetes risk detection quick and accessible. It offers a handy first alert system—especially useful for those in remote areas—encouraging timely awareness and action before clinical care is even sought.

This smart diabetes prediction model is more than just technology—it's a lifesaver, especially for those in villages or remote areas with limited medical facilities. With just a few easy inputs like glucose, weight, and age, it can detect early signs of diabetes and issue timely warnings. For example, if someone only knows their glucose level is 220—boom, the model flags the risk instantly. It helps bridge the gap between awareness and medical attention, turning simple data into potentially life-saving alerts. A small step that can make a big difference—because prevention truly is better than cure.



9. References

S.No.	Website / Source	Purpose / Information Used
1	Kaggle - Diabetes Dataset https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database	Dataset used to train and test the diabetes prediction model
2	MathWorks - fitcensemble https://www.mathworks.com/help/stats/fitcensemble.html	Documentation used to implement the Bagged Ensemble (Random Forest) classifier in MATLAB
3	MathWorks - uicontrol https://www.mathworks.com/help/matlab/ref/uicontrol.html	Used for designing the GUI interface elements in MATLAB
4	Wikipedia - Random Forest https://en.wikipedia.org/wiki/Random_forest	Understanding Random Forest algorithm and ensemble methods
5	Stack Overflow https://stackoverflow.com	Used for solving code errors and UI/logic implementation tips
6	ScienceDirect Article https://www.sciencedirect.com/science/article/pii/S1877050920300557	Research article on machine learning models for diabetes prediction in healthcare
7	IEEE Xplore - ML for Diabetes https://ieeexplore.ieee.org/document/7938895	Reference for ML-based approaches in early-stage diabetes detection
8	Comet Blog - Pima Diabetes https://www.comet.com/site/blog/pima-indian-diabetes-prediction/	Practical tutorial for building ML models with the Pima Indian dataset
9	Data Science MathWorks Video Series https://in.mathworks.com/videos/series/data-science-tutorial.html	Guided tutorials for applying data science workflows in MATLAB
10	YouTube - Data Science Basics https://www.youtube.com/watch?v=b4sq1dIdBS8	Introductory video used for understanding basic data science and preprocessing concepts