Early Detection of Diabetes Risk Through Machine Learning and Interactive Web Application SugerShield

Prepared by
Lakmini Thamodya

9/5/25

Machine Learning and Statistical Data Analysis

Abstract

Diabetes is one of the most prevalent chronic diseases worldwide, and its early detection is critical for effective management and prevention of complications. With the rapid growth of data-driven healthcare, machine learning has become a powerful tool to predict diabetes risk by analyzing both clinical and demographic factors. In this project, a Random Forest model was developed to predict the likelihood of diabetes using features such as HbA1c level, blood glucose level, age, BMI, and lifestyle indicators. The model achieved an accuracy of **97%**, demonstrating its strong predictive capability. To bridge the gap between technical analysis and practical use, the trained model was deployed as a web application, allowing users to input their health parameters and instantly obtain a prediction. This project highlights the potential of integrating machine learning into healthcare by providing an accessible, accurate, and scalable solution for early diabetes risk assessment, which can be applied in real-world industry and clinical settings.

Contents

Abstract 0

List of Figures 1

List of Tables 1

Introduction 2

Description of the Data set 2

Main results of the Descriptive Analysis 3

- Distribution of Numerical Variables 3
- Distribution of categorical variables 4
- Factor Analysis for Mixed Data 5

Results of Advanced Analysis 6

Model Deployment 7

- SugerSheild – Web Application 7

Conclusion 8

References 8

List of Figures

Figure 1: The distribution of Diabetes patients(Response) 3
Figure 2: Correlation Heatmap of Numerical Variables 5
Figure 3: Cramer's V Correlation matrix for categorical variables 5
Figure 4: The Score Plot (Factor analysis for mixed data) 5
Figure 5: Plot of Optimal Clusters (ELBOW) 6
Figure 6: Results of best model 7
Figure 7: Feature Importance Plot of Best model 7
Figure 8: Preview of Web app - SugarShield 8

List of Tables

Table 1: Description of Data 2

Introduction

Sri Lanka has a high and rising prevalence of diabetes, with estimates for 2024 showing 10.2% of adults (20-79) affected, totaling about 1.6 million people, according to the International Diabetes Federation (IDF). A 2019 national study found nearly a quarter (23%) of adults had diabetes and over a third (31%) had high blood sugar, with a significant number (38%) remaining undiagnosed, and another recent estimate suggests 4.2 million people have diabetes. The condition is exacerbated by increasing rates of obesity and a lack of physical activity, leading to premature deaths. The early prediction of having diabetes may help to a person to act accordingly.

This project focuses on predicting diabetes using patient medical and demographic data. The dataset includes key health indicators such as age, BMI, hypertension, heart disease, smoking history, HbA1c levels, and blood glucose levels. By applying data cleaning, exploratory analysis, and feature engineering, the dataset was prepared for machine learning. The Random Forest model showed the best performance and was saved for deployment. The project demonstrates how machine learning can be applied in healthcare to identify individuals at risk of diabetes and support early intervention strategies.

Description of the Data set

Diabetes Prediction dataset is sourced by Keggle which contains 100000 medical and demographic of patients There are nine attributes. The predictor variable is diabetes which indicates whether an individual having diabetes or not. The overview of the variables is given below.

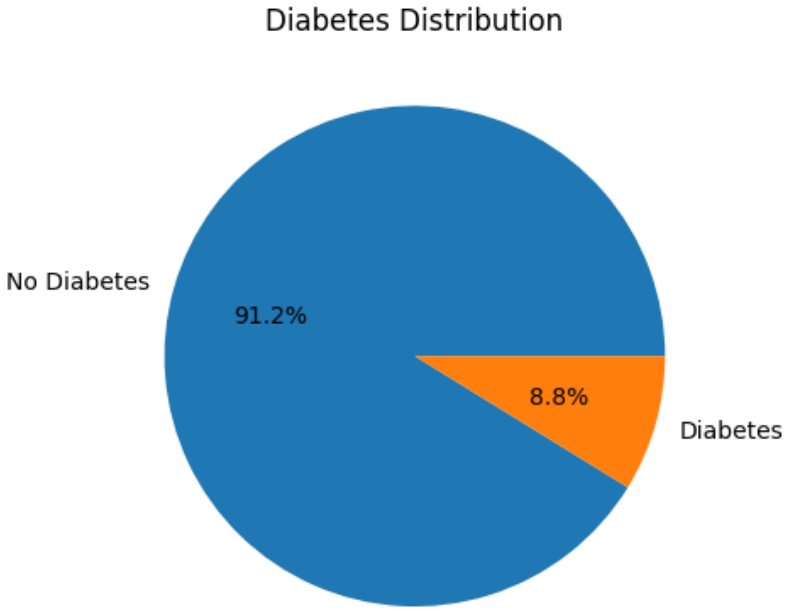
Table 1: Description of Data

Attribute	Description of the attribute	Type
Gender	Biological sex of the individual	Categorical
Age	Age	Numerical
Hypertension	a medical condition in which the blood pressure in the arteries is persistently elevated (No:0, Yes:1)	Categorical
Heart Disease	Having heart disease: 1, otherwise 0	Categorical
Smoking History	Values: Not current, Former, No Info, Current , Never	Categorical
BMI	Body Mass Index	Numerical
HbA1c level	a measure of a person's average blood sugar level over the past 2-3 months	Numerical
Blood glucose level	the amount of glucose in the bloodstream at a given time	Numerical
Diabetes	If presence: 1 , Not presence : 1	Categorical

Preprocessing

- ✓ Remove 3854 duplicate rows from the dataset
- ✓ No missing values have found
- ✓ Split the dataset into 80% training and 20% testing set.

Main results of the Descriptive Analysis



The distribution of the target variable or the distribution of the patients is given in pie chart. So that we can clearly see 91.2% of the patients are having diabetes.

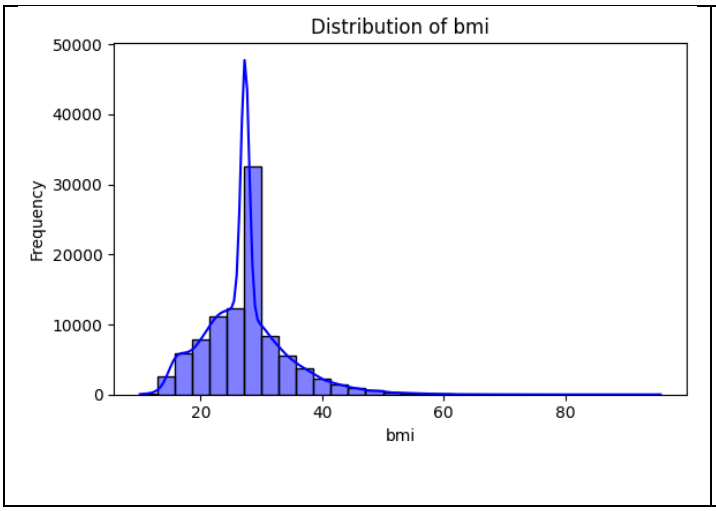
According to that the response variable is highly class imbalanced, I have taken the necessary techniques to solve that. (SMOTE technique)

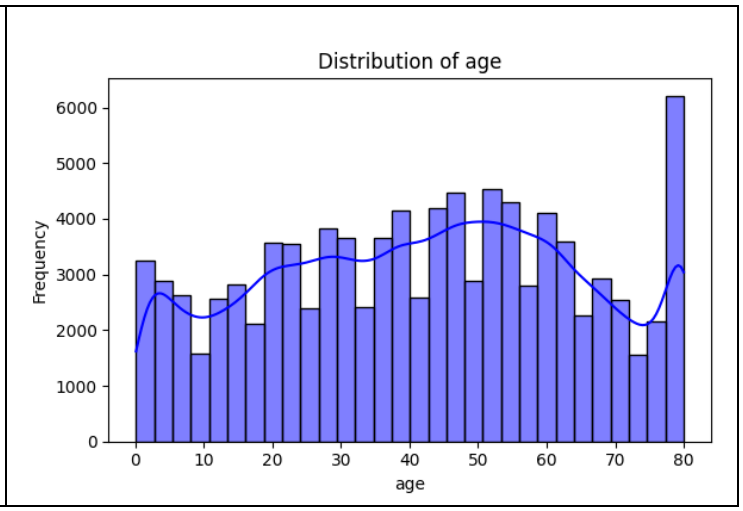
Figure 1: The distribution of Diabetes patients(Response)

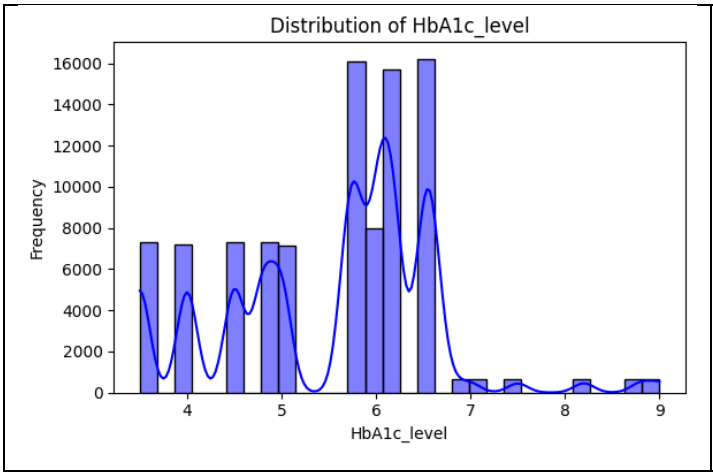
Distribution of Numerical Variables

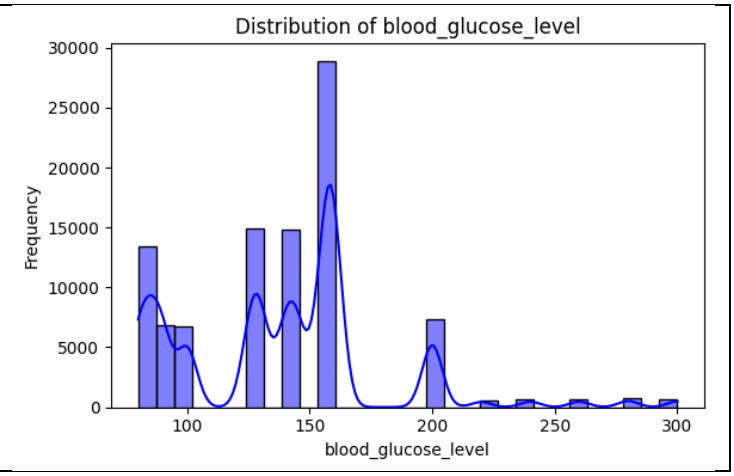
The distribution of BMI values are moderately right-skewed, with most of the BMI values are concatenated around the lower range 20 to 40. There are a few number of patients who has extreme values of BMI. So that most of the patients are having average BMI value.

The age variable is fairly evenly spread across its range, with no clear skewness. This indicates that the dataset covers a wide range of age groups, ensuring good representation of both younger and older individuals.









Among the patients most of them have been recorded 5 to 7 HbAC1 (glycated hemoglobin) level. Glycated hemoglobin measures a person's average blood sugar levels over the past three months by showing the percentage of hemoglobin in red blood cells that is coated with sugar. A normal A1c level is below 5.7%, while 5.7% to 6.4% indicates prediabetes, and 6.5% or higher suggests diabetes.

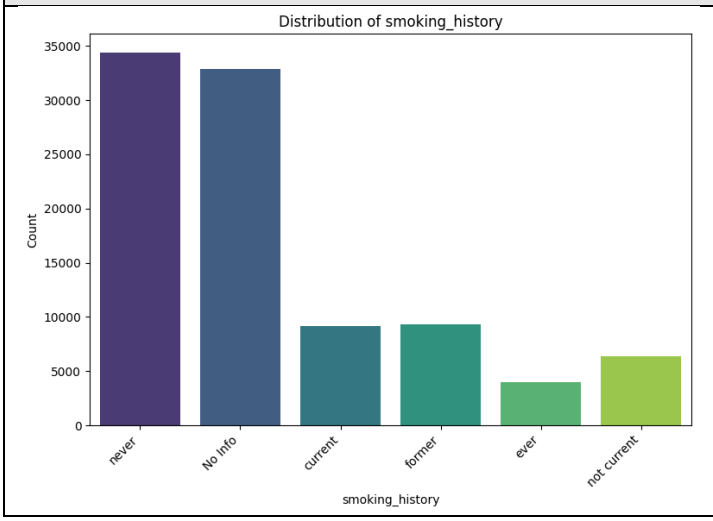
Similar to HbA1c, the blood glucose distribution is also multi-modal with a strong concentration around certain values. This pattern highlights the presence of distinct clusters in glucose measurements, which may correspond to normal, prediabetic, and diabetic ranges.

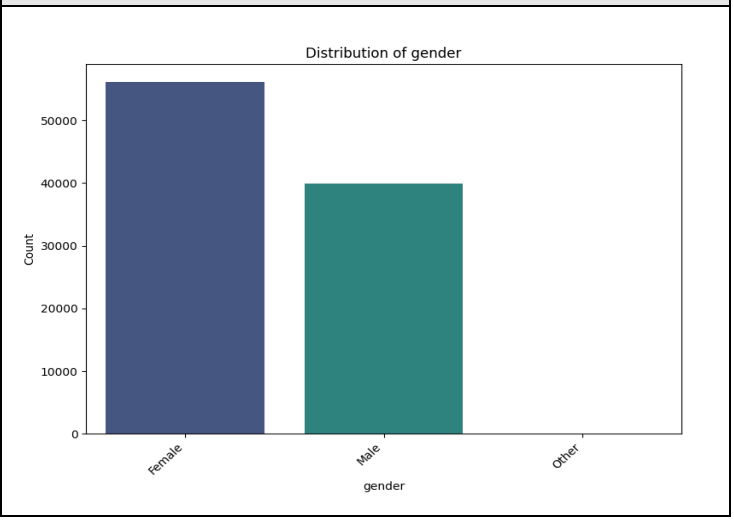
Distribution of categorical variables

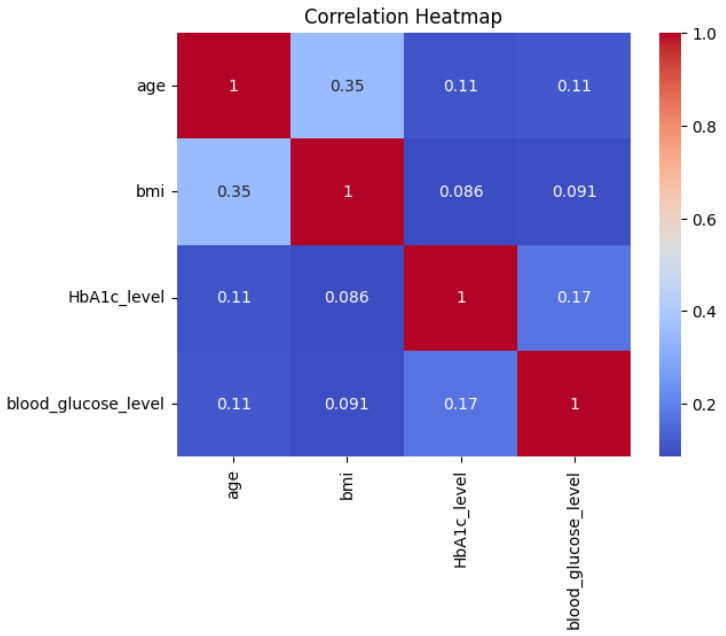
By the value counts precentage of hypertension and blood pressure , most of the diabetes patients have been suffering with high blood pressure and hypertension symptoms.

Similarly, 95.91% of patients do not have heart disease, while only 4.1% do. Though the prevalence is low, this factor is clinically significant since heart disease is a common complication associated with diabetes. Most of the patients have never smoked before at the moment of the data collection. And concatenating the sum of never and no info, half of the patients have not smoking habit before as smaller proportions are current , ever , not current, and former. Amoung patients larger proportions are female patients , other category is negligible.

Hypertension	Blood Pressure
0 88685 92.23% 1 7461 7.76%	0 92223 95.91% 1 3923 4.1%







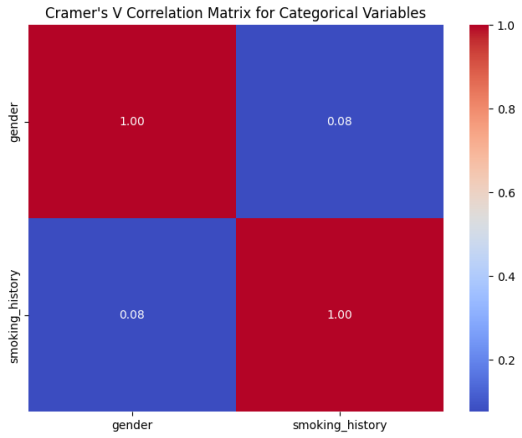


Figure 2: Cramer's V Correlation matrix for categorical variables

Figure 3: Correlation Heatmap of Numerical Variables

- The correlation heatmap does not show any severe multicollinearity between numerical variables.
- No need to perform dimension reduction technique due to miner correlations.
- Cramers'V correlation matrix is used to detect correlation between categorical variables , could not found correlation between gender and smoking history.

Factor Analysis for Mixed Data

Since the diabetes dataset contains both numerical categorical variables, Factor analysis for mixed data is performed. By the score plot of FAMD, could not detect any clusters among patients. But we can identify some outliers in this dataset. Approximately the proportion of outliers is very small, so I decided to keep them as it is ,to avoid biased results.

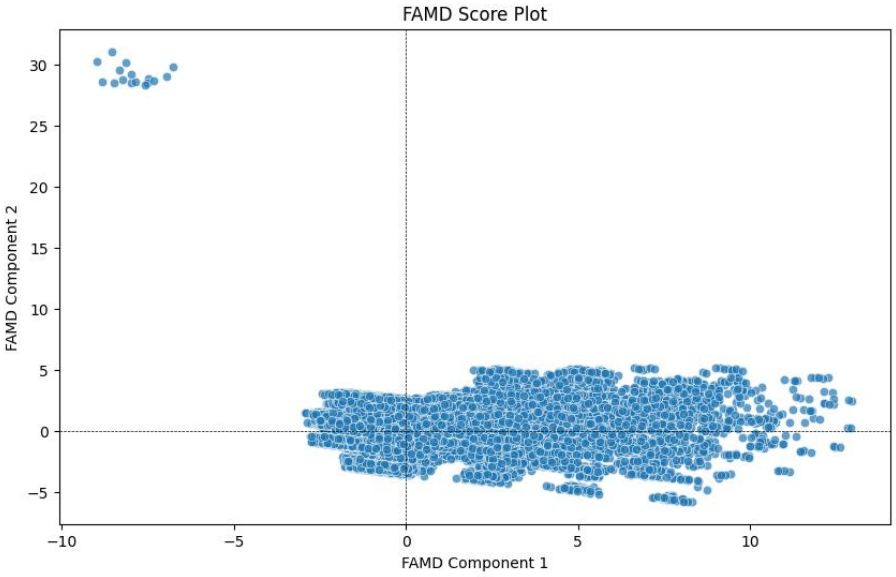
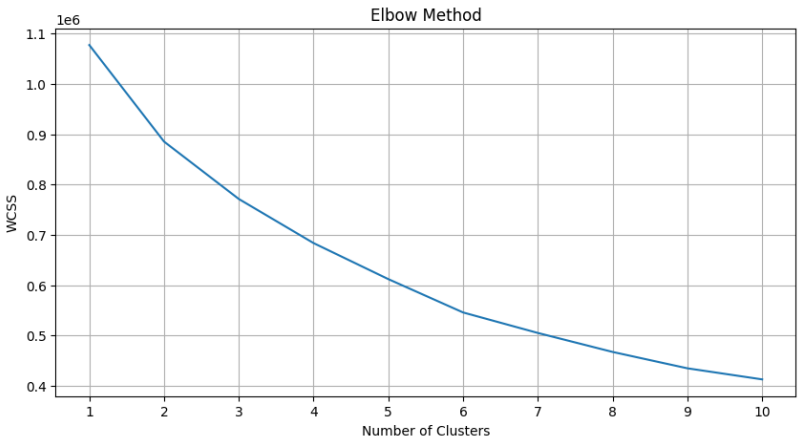


Figure 4: The Score Plot (Factor analysis for mixed data)



➢ Although clustering analysis was performed, no clear optimal number of clusters could be identified, as the elbow method did not reveal a distinct inflection point.

Figure 5: Plot of Optimal Clusters (ELBOW Method)

Results of Advanced Analysis

To build predictive models, categorical variables were first encoded and numerical variables were scaled to ensure consistency in the input features. Initially, Decision Tree and Random Forest classifiers were applied, achieving 99% training accuracy and 91% testing accuracy. This indicated strong performance but also suggested overfitting, as the models generalized less effectively on unseen data.

Since the dataset was imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the classes and improve model learning. After handling class imbalance, advanced models such as Decision Trees and Random Forest were reapplied. To further address the overfitting issue, 5-fold cross-validation was implemented. This approach provided more reliable evaluation and improved generalization, resulting in an average training accuracy of 97% and testing accuracy of 96.96%.

These results highlight that after addressing class imbalance and applying cross-validation, the models achieved high accuracy with reduced overfitting, making them robust for predicting diabetes.

Confusion Matrix (Train):

```
[[68319  1836]
 [ 2084 68071]]
```

Classification Report (Train):

	precision	recall	f1-score	support
0	0.97	0.97	0.97	70155
1	0.97	0.97	0.97	70155
accuracy			0.97	140310
macro avg	0.97	0.97	0.97	140310
weighted avg	0.97	0.97	0.97	140310

Confusion Matrix (Test):

```
[[17467    42]
 [  541  1180]]
```

Classification Report (Test):

	precision	recall	f1-score	support
0	0.97	1.00	0.98	17509
1	0.97	0.69	0.80	1721
accuracy			0.97	19230
macro avg	0.97	0.84	0.89	19230
weighted avg	0.97	0.97	0.97	19230

Figure 6: Results of best model

97% precision and 99% recall of predicting not diabetes were actually not positive. But recall of positive class is less in testing data.

By figure 7, The Random Forest model shows that HbA1c level and blood glucose level are by far the most influential predictors, together contributing more than 60% of the model's decision-making, highlighting their strong link to diabetes risk. Age and BMI also play important roles, while smoking history adds only a minor contribution.

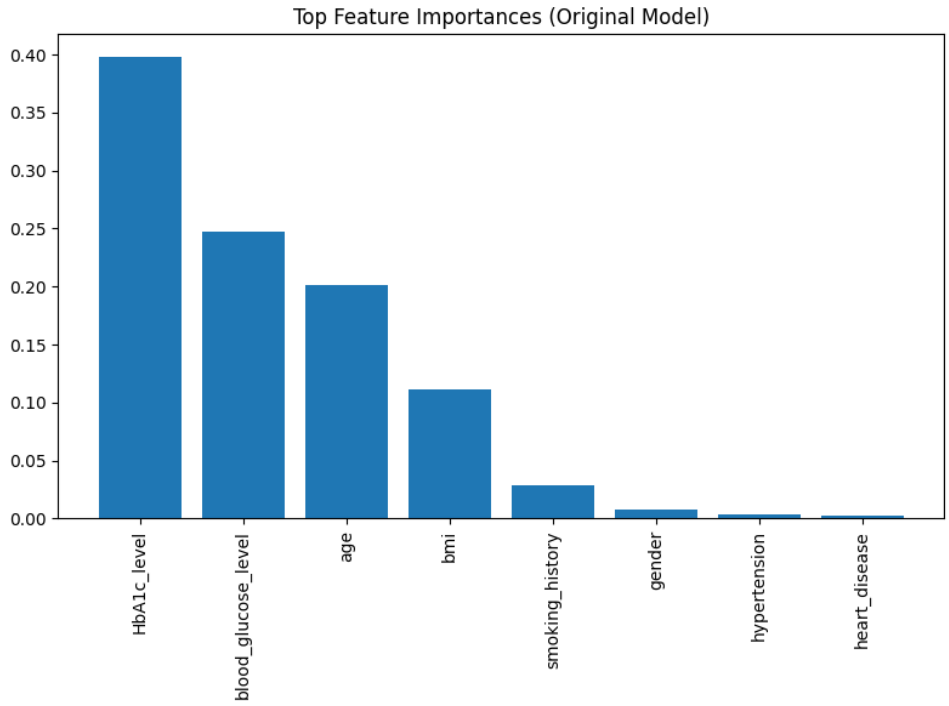


Figure 7: Feature Importance Plot of Best model

Model Deployment

The finalized model is random forest model with parameter tuning model which is deployed for build a web application to implement the outcome. To make the Random Forest model accessible for real-world use, it was deployed as a web application. The trained model was saved using **pickle** and integrated with a lightweight API built in Flask.The API exposes a /predict endpoint where users can input clinical parameters such as HbA1c level, blood glucose, age, BMI, and lifestyle factors. The system processes these inputs, applies the model, and returns the prediction along with a probability score. For usability, a simple web interface was developed, allowing users to enter values through a form and receive an immediate risk assessment. Below figures shows the overview of my product **SugerShield** Web app.

SugerSheild – Web Application

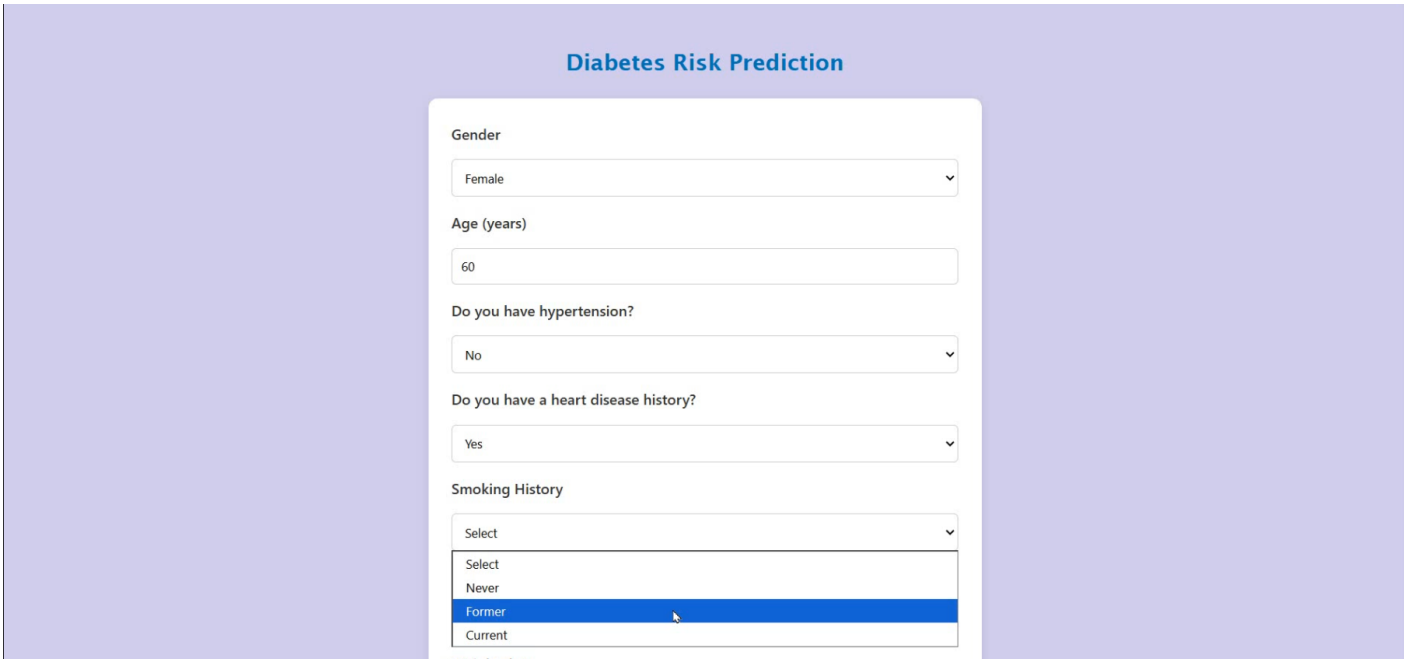


Figure 8: Preview of Web app - SugarShield

Conclusion

This project has effectively built and implemented a diabetes risk predictive model based on a Random Forest algorithm. The analysis demonstrates that HbA1c level and blood glucose level are the most dominant predictors, followed by age and BMI, with lifestyle and medical history factors having less weight in the model. By making the trained model available as a web application, the project demonstrates how advanced machine learning techniques can be translated into an accessible and usable tool. This deployment not only demonstrates the potential of data-driven healthcare solutions but also provides a platform that can be built upon with additional data, new features, and integration with clinical systems. Overall, the project illustrates the application of predictive modeling to support early risk identification and enhance informed health choices.

References

1. https://www.youtube.com/watch?v=dam0GPOAvVI
2. Dataset: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data
3. Git hub Repository: https://github.com/thamodya2001/Diabetes-Prediction/tree/main