Diabetes Prediction Model

DataDiagnostics: Aysha, Godswill, Kehlani & Sum

Project Overview

Why Diabetes?

Diabetes is a prevalent chronic disease, affecting millions of individuals and posing significant health risks if not managed effectively. Early detection and intervention are crucial in mitigating its adverse effects.

Our Goal

To develop a predictive model to identify individuals at risk of developing diabetes based on various health indicators.

Project Overview

Do you have diabetes?
How does a person know they have diabetes?

NHS Health Check 2022/2023 (England)

People invited:	2,925,325
People taking up:	1,136,770

Population in England:
approx. 56M

Project Overview

Our Focus: Type 2 Diabetes

In the UK, over 90% of all adults with diabetes have Type 2.

It's strongly associated with personal lifestyle factors like diet, exercise and family history.

Unlike Type 1 diabetes, Type 2 diabetes is largely preventable through lifestyle changes.

	Health Check	Questionnaire
How	measure blood glucose (sugar) levels	ask health and lifestyle related questions
+	highly accurate	extensive and cheap
-	expensive and limited in scope	less accurate

Can ML improve accuracy?

Data Sourcing and Preprocessing

Data Source

A dataset comprising of relevant health parameters such as BMI, age, blood pressure, family history of diabetes, and lifestyle factors

Preprocessing

Cleaned and pre-processed the collected data to handle missing values, converted categorical data to a numerical format and standardise features

Splitting

Defined the target and feature variables and split the data into subsets ready for the training and testing of models

Target and Feature Variables

Sample Code

```
# Split the data into features (X) and target variable (y)
X = df.drop('Diabetic', axis=1)
# Define target vector
y = df["Diabetic"].values.reshape(-1, 1)
y = y.ravel()
```

```
# Display the processed DataFrame
print("\nProcessed DataFrame:")
X.head()
```

```
Processed DataFrame:
```

	Age	Gender	FamilyDiabetes	PhysicallyActive	BMI	Smoking	Alcohol	Sleep	SoundSleep	RegularMedicine	JunkFood	Stress
0	2	0	0	3	39.0	0	0	8	6.0	0	0	1
1	2	0	0	1	28.0	0	0	8	6.0	1	2	1
2	1	0	0	3	24.0	0	0	6	6.0	0	0	1
3	2	0	0	3	23.0	0	0	8	6.0	0	0	1
4	1	0	0	1	27.0	0	0	8	8.0	0	0	1

```
print("\nTarget variable:")
y[:5]
```

```
Target variable:
 array([0, 0, 0, 0, 0])
```

Machine Learning Models

Logistic Regression

Trained a logistic regression model as a baseline model

Random Forest

Implemented a random forest classifier to capture non-linear relationships between features

Decision Tree

Built a decision tree classifier to understand the decision-making process

Support Vector Machine

Employed SVM with different kernels to find the best separating hyperplane. (Chosen Kernel: Linear)

K-Nearest Neighbors

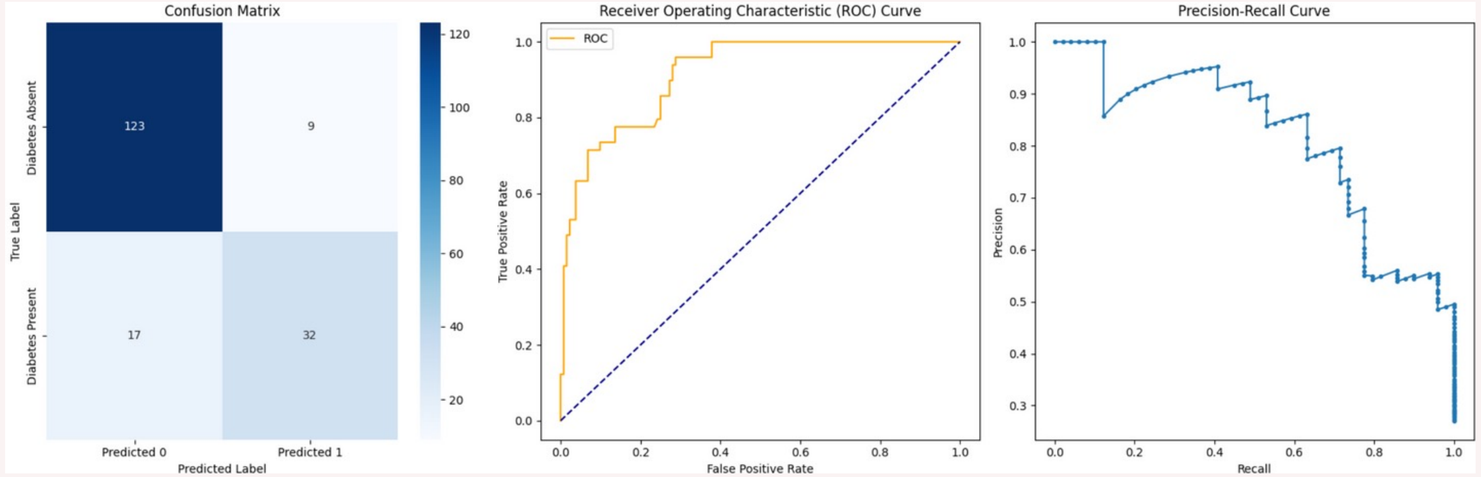
Implemented KNN to classify data points based on the majority class of their nearest neighbours

Neural Networks

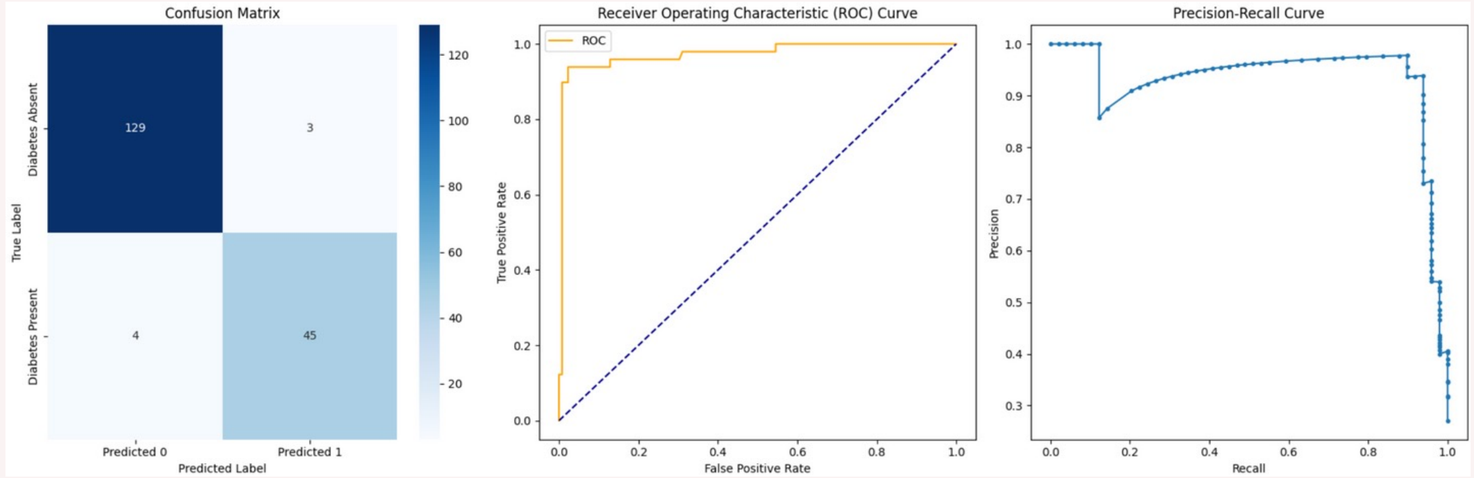
Constructed a Neural Network with multiple hidden layers using TensorFlow & optimised it using Keras-Tuner

Model Results & Evaluation

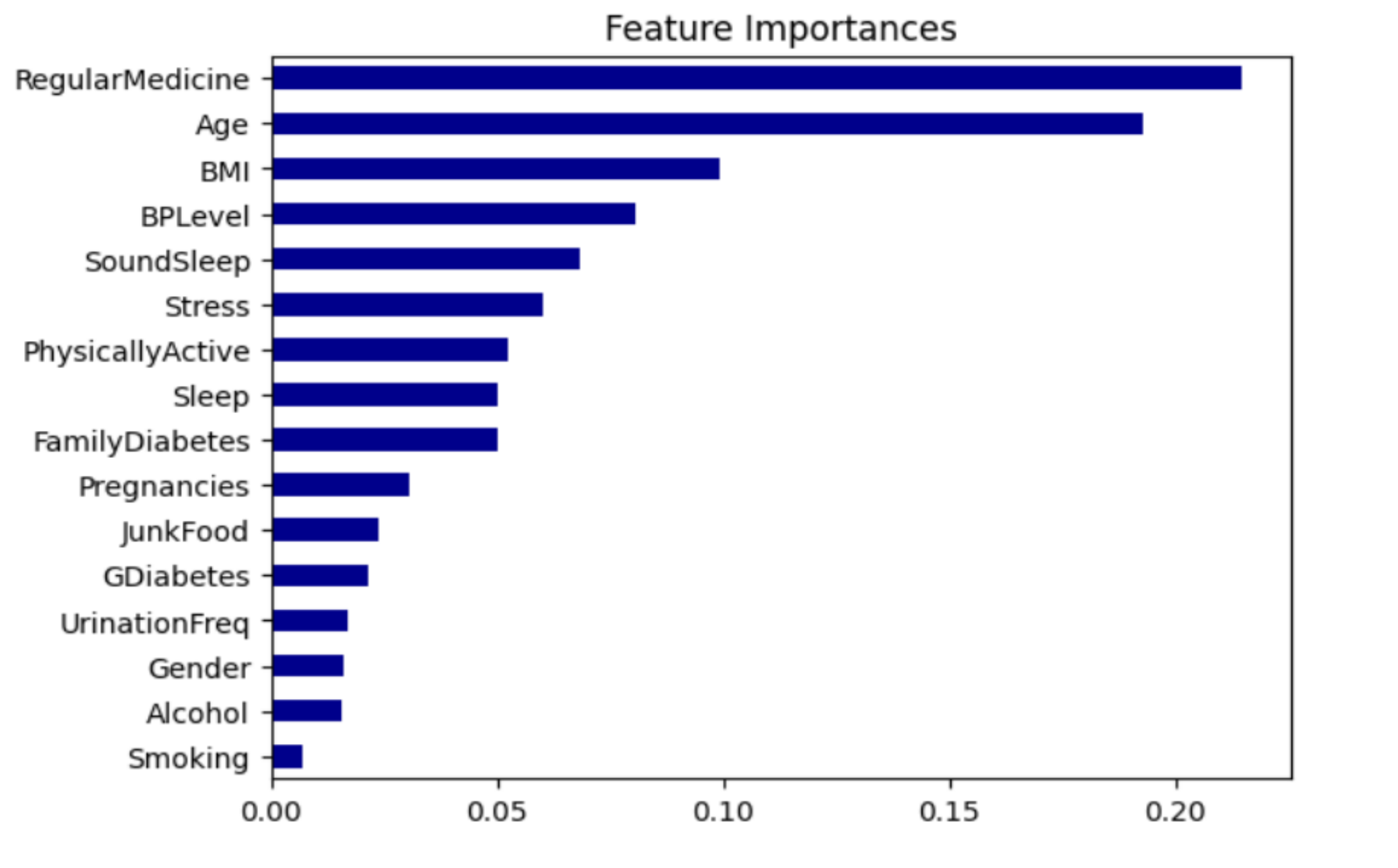
Logistic Regression



Random Forest

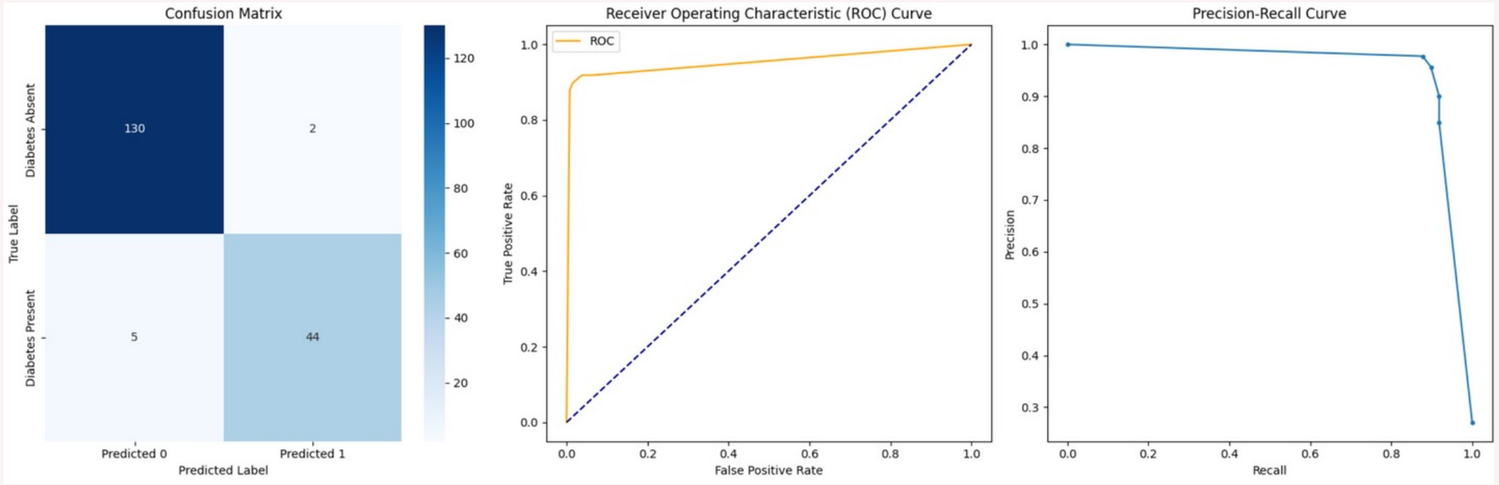


Model Feature Importances

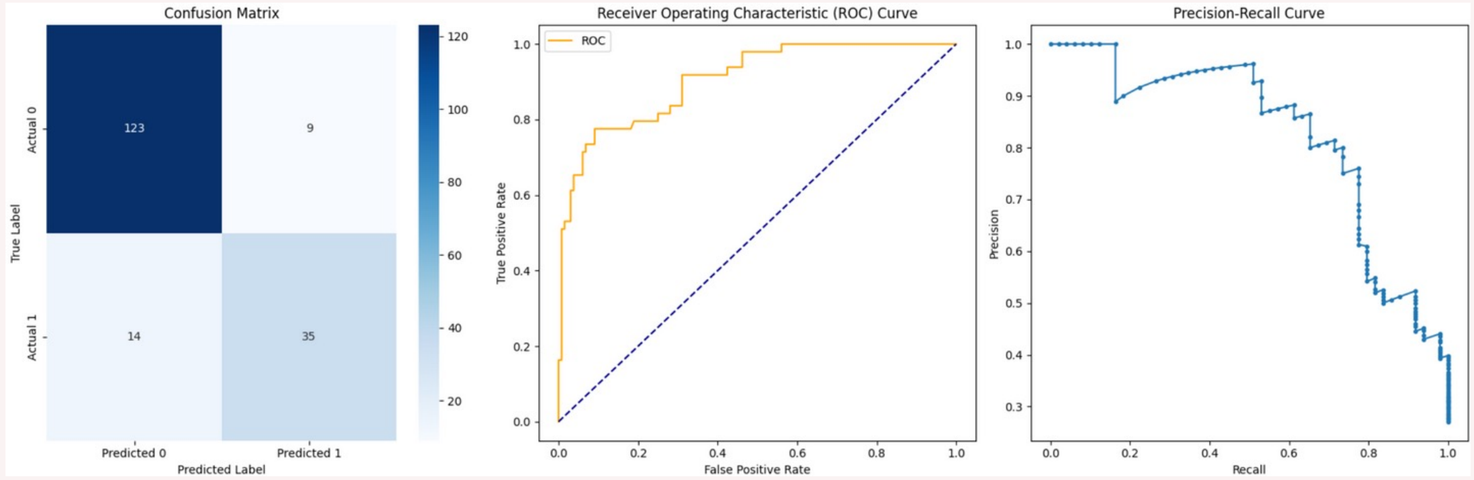


Model Results & Evaluation

Decision Tree

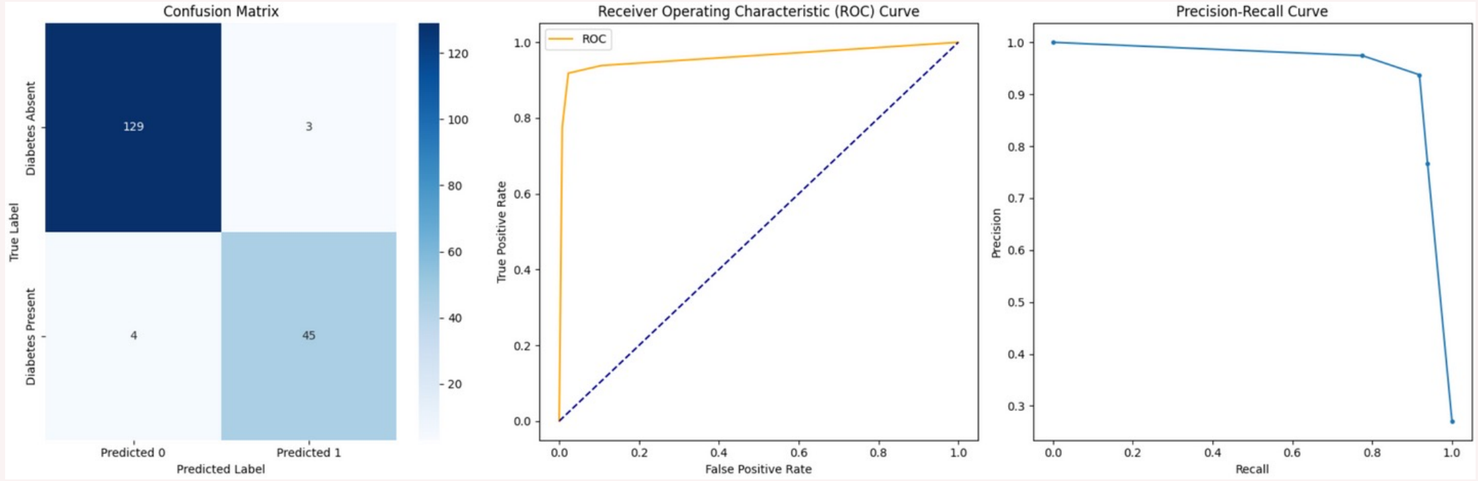


SVM

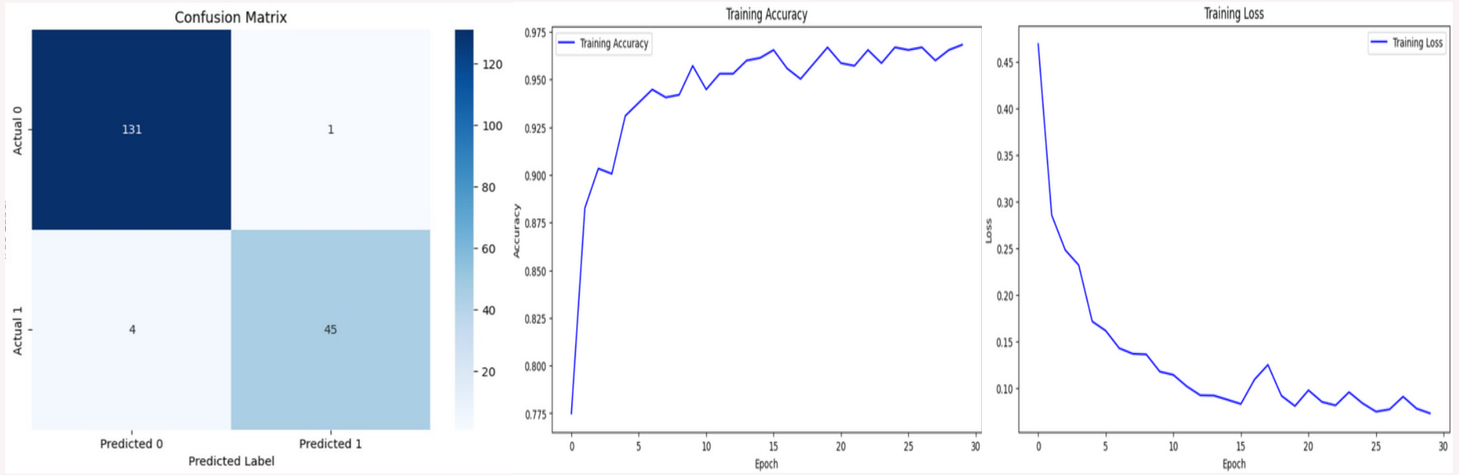


Model Results & Evaluation

KNN

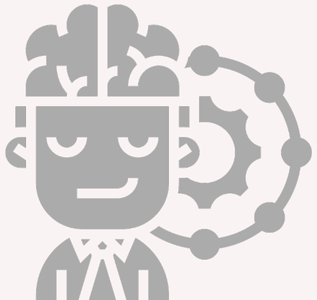


Neural Network



Model Results & Evaluation

Model	Accuracy	Precision (Diabetes Absent)	Precision (Diabetes Present)	Recall (Diabetes Absent)	Recall (Diabetes Present)	F1-score (Diabetes Absent)	F1-score (Diabetes Present)
Logistic Regression	85.64%	88%	78%	93%	65%	90%	71%
Random Forest	96.13%	97%	94%	98%	92%	97%	93%
Decision Tree	96.13%	96%	96%	98%	90%	97%	93%
SVM	87.29%	90%	80%	93%	71%	91%	75%
KNN	96.13%	97%	94%	98%	92%	97%	93%
Neural Network (optimised)	97.24%	97%	98%	99%	92%	98%	95%



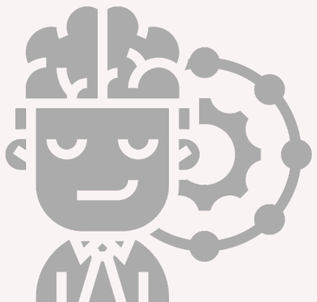
Front-end Visualisation Of Model Performance

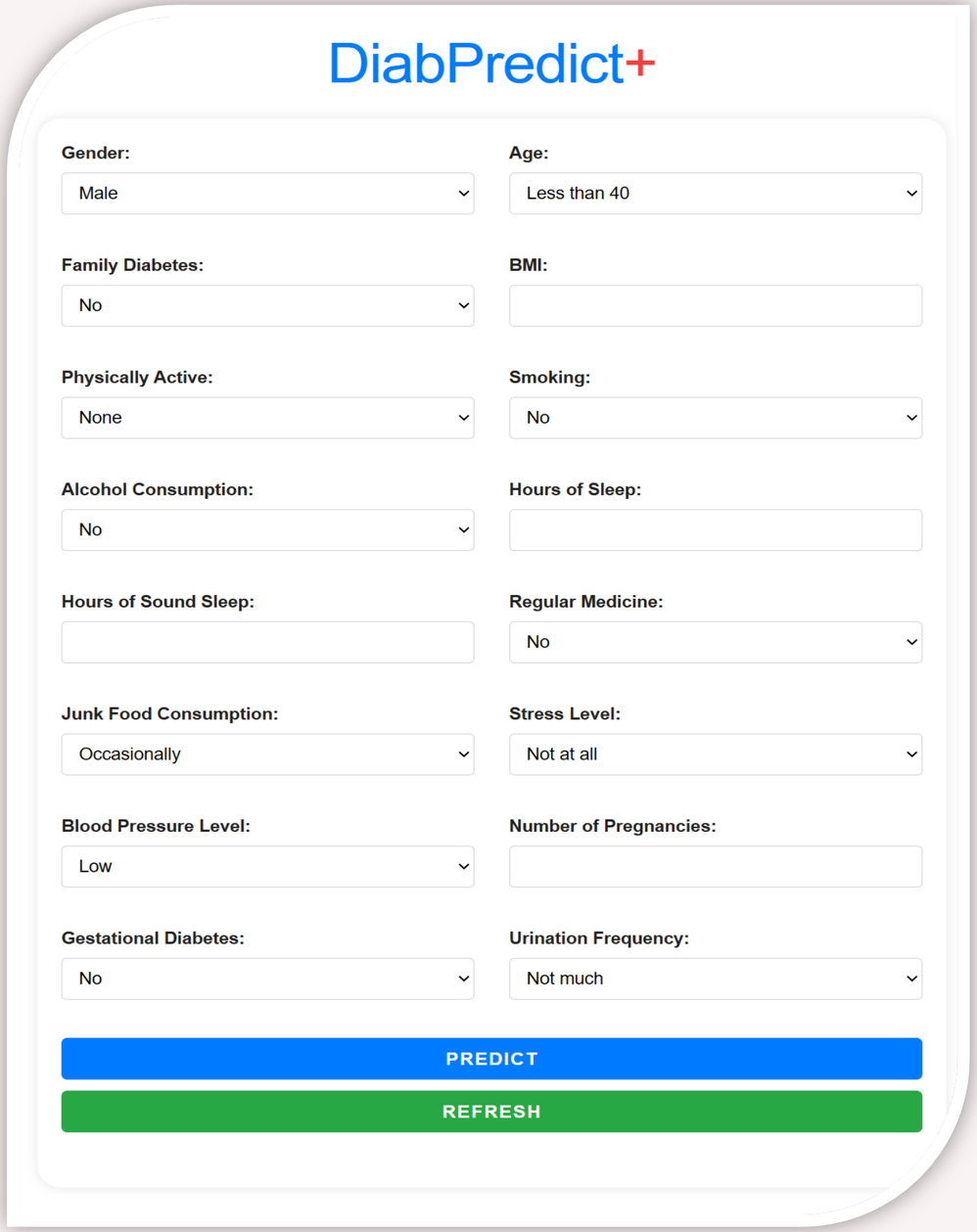
DiabPredict+

Web app predicts diabetes risk using machine learning.

With input on age, gender, lifestyle, and medical history, it instantly assesses diabetes status and probability.

Built on Flask and Keras, it offers accurate predictions and promotes health awareness.





Conclusion

Neural Network

Random Forest, Decision Tree, KNN

Support Vector Machine

Logistic Regression

Neural Network model shows the highest accuracy in predicting Type 2 Diabetes

Random Forest, Decision Tree, and KNN also exhibit promising results

SVM and Logistic Regression show comparatively lower accuracy and performance metrics

Top 3 features influencing the predictions:

- RegularMedicine
- Age
- BMI

Recommendation:

- **Accuracy:** Neural Network
- **Balanced approach:** Random Forest, Decision Tree, KNN

Thank you for your attention