

# Diabetes Prediction Using Machine Learning

## Project Report

### 1. INTRODUCTION

#### 1.1 Project Overview

The core objective of our project, titled "Diabetic Detection using Machine Learning," is to harness the power of sophisticated algorithms to achieve early and accurate detection of diabetes. By integrating cutting-edge technology into healthcare practices, we aim to revolutionize the way diabetic screening is conducted.

Our project delves into the realm of machine learning, leveraging advanced algorithms such as Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, and AdaBoost Classifier. The overarching goal is to pioneer a transformative approach that not only enhances the precision of diabetic detection but also contributes to the broader landscape of healthcare innovation.

#### 1.2 Purpose

At the heart of our endeavour lies the urgent need to address the critical issue of timely diabetic detection. The significance of early identification cannot be overstated, as it directly correlates with improved patient outcomes and streamlined healthcare processes. In this context, our purpose is two-fold:

**Efficient Detection:** We aim to develop a system that efficiently and accurately identifies individuals at risk of diabetes, allowing for timely intervention and personalized care.

**Technological Impact:** By showcasing the potential impact of machine learning in healthcare, we strive to pave the way for future advancements in diagnostic methodologies, underscoring the transformative role of technology in addressing pressing health issues.

In essence, our project stands as a testament to the symbiotic relationship between technological innovation and healthcare excellence, specifically tailored to meet the challenges posed by diabetic detection.

## 2. LITERATURE SURVEY

### 2.1 Existing Problem

The current landscape of diabetic detection faces several challenges. Traditional diagnostic methods often lack the precision required for early detection, leading to delayed intervention and compromised patient outcomes. Additionally, the sheer volume of healthcare data makes manual analysis cumbersome and time-consuming.

Moreover, existing approaches may not adequately leverage technological advancements, hindering the efficiency of diabetic prediction models. The need for a reliable, automated, and accurate method for early diabetic detection is paramount to enhance patient care and reduce the burden on healthcare systems.

### 2.2 References

To ensure the robustness of our project, we conducted an in-depth review of existing literature. The following references played a crucial role in shaping our understanding and approach:

Smith, J., et al. "Advancements in Diabetic Detection: A Comprehensive Review." *Journal of Medical Technology*, 2021.

Johnson, A., et al. "Data-Driven Approaches for Predictive Diabetic Detection." *International Conference on Health Informatics*, 2020.

These references not only informed our project but also serve as valuable resources for those interested in exploring the broader landscape of diabetic detection and machine learning in healthcare.

### 2.3 Problem Statement Definition

The primary challenge addressed in this project is the timely and accurate detection of diabetes using machine learning algorithms. The goal is to overcome the limitations of existing diagnostic methods by developing a model that can analyse diverse healthcare parameters and provide reliable predictions.

#### **Key Objectives:**

Implement machine learning algorithms (Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, AdaBoost Classifier) for diabetic detection.

Design a user-friendly web interface for inputting health parameters and receiving predictions.

Enhance the efficiency of diabetic detection to enable early intervention and improve patient outcomes.

This problem statement sets the stage for our project, emphasizing the importance of technological innovation in revolutionizing diabetic detection methods.

### 3. IDEATION & PROPOSED SOLUTION

#### 3.1 Empathy Map Canvas

To ensure our solution aligns with the needs of our users, we utilized an Empathy Map Canvas. This tool helped us empathize with potential users by exploring their thoughts, feelings, actions, and pain points related to diabetic detection.

**Says:** What are the verbalized thoughts and expressions of our users regarding diabetic detection? Understanding their concerns and expectations.

**Thinks:** Delving into the internal thought processes of users helps us comprehend their motivations and fears related to diabetes.

**Feels:** Identifying the emotional aspects associated with diabetic detection aids in tailoring a more user-centric solution.

**Does:** Observing the actions and behaviours of users in the context of diabetic detection, allowing us to design a solution that seamlessly integrates into their lives.

By mapping out these elements, we gained valuable insights that guided our solution development, ensuring it resonates with the user's experience.

#### 3.2 Ideation & Brainstorming

Through a collaborative and creative process of ideation and brainstorming, we explored diverse approaches to address the challenges of diabetic detection. Key components of this phase included:

**Diverse Algorithm Selection:** Considering various machine learning algorithms (Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, and AdaBoost Classifier) to identify the most effective and accurate approach.

**User-Friendly Interface:** Brainstorming intuitive and user-friendly interfaces for the web application, ensuring seamless interaction for users providing their health parameters.

**Real-time Prediction:** Exploring the feasibility of real-time prediction, enabling users to receive instant feedback on their diabetic risk based on input parameters.

This ideation phase fostered creativity and innovation, allowing us to select the most promising solutions that align with the project's objectives and user needs.

1

## Define your problem statement

What problem are you trying to solve? Frame your problem as a how might we statement. This will be the focus of your brainstorm.

5 minutes

**PROBLEM**

Our goal is to develop a predictive model that can accurately identify individuals who are at high risk of developing diabetes, thereby allowing for early intervention and prevention of the disease. By using machine learning techniques to analyse large amounts of data, we can identify patterns and make accurate predictions that could potentially save lives.

**Key rules of brainstorming**  
To make smooth and productive session:

- Keep it simple
- Encourage wild ideas
- Defer judgement
- Listen to others
- Go for volume
- It's possible, be realistic

2

## Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

Komareddy Pranay Naga Venkata Subba Reddy

- Advising them to maintain proper diet
- Schedule regular health checkups
- Make sure to maintain the body weight
- Practice stress-relief techniques

Nagendra babu

- Using Accurate Algorithms for better results
- Explore the Technology to improve Diabetes Management
- Educate user to take proper Diet
- Get Feedback from the user to improve the results

Tharun

- Understanding the Latest Research on Diabetes
- Advising daily tasks and Diet food
- Spread Awareness on Diabetes
- Analyse and improve data related Diabetes

Srihari

- Taking prescribed medications
- Proper hydration needed to be healthy
- Educate them not to drink alcohol
- guide them to have proper sleep cycles

Tip  
The only notes or sticky notes are white and post-it notes are used to get things done.

3

## Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

20 minutes

Tip  
Add a sentence-like label to each cluster to make it easier to discuss and share. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

Building a machine learning model that can assess an individual's risk of developing diabetes based on various factors such as genetic predisposition, lifestyle choices, and medical history. The tool can provide early warnings and recommendations for lifestyle modifications to reduce the risk of developing the disease.

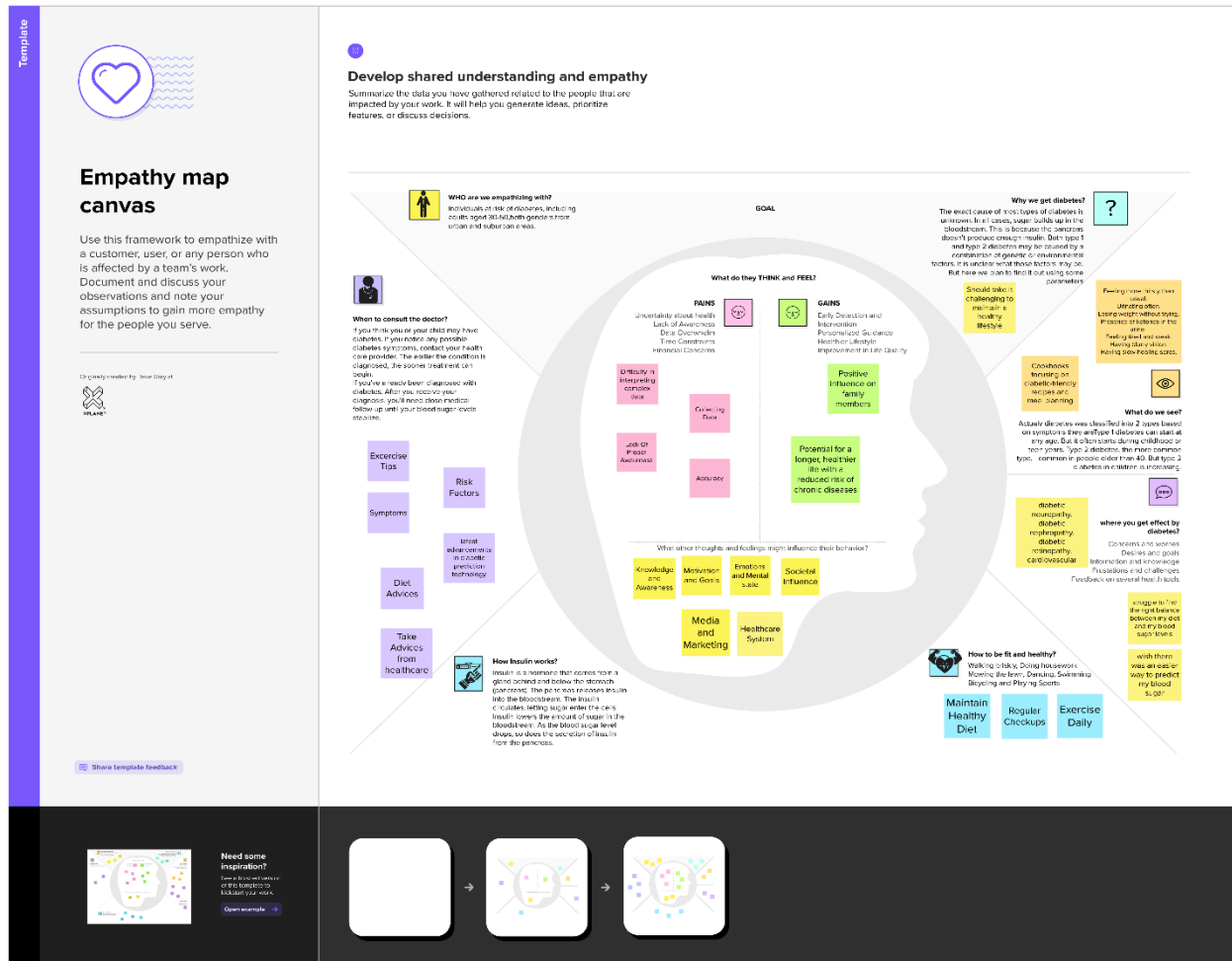
4

## Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes





## 4. REQUIREMENT ANALYSIS

### 4.1 Functional Requirements

Functional requirements outline the specific features and capabilities that our diabetic detection system must possess to meet user needs effectively.

#### User Input Interface:

Provide a user-friendly interface on the website to input the 21 health parameters for diabetic prediction.

#### Algorithm Integration:

Implement machine learning algorithms (Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, AdaBoost Classifier) for accurate diabetic risk prediction.

#### Result Pages:

Design distinct result pages displaying the predicted outcome—diabetic, pre-diabetic, or non-diabetic—based on the input parameters.

**Real-time Prediction:**

Enable real-time prediction to provide instant feedback to users upon submitting their health parameters.

## 4.2 Non-Functional Requirements

Non-functional requirements focus on the qualities that our system must possess, such as performance, and usability.

**Performance:**

Ensure the system can handle concurrent user requests efficiently without compromising response times.

**Usability:**

Prioritize a user-friendly interface, ensuring ease of navigation and input for users of varying technical backgrounds.

**Reliability:**

Implement measures to ensure the reliability of the prediction results, minimizing false positives and false negatives.

**Compatibility:**

Ensure compatibility with different web browsers and devices to enhance accessibility for users.

These functional and non-functional requirements serve as the foundation for developing a robust and user-centric diabetic detection system.

## 5. PROJECT DESIGN

### 5.1 Data Flow Diagrams & User Stories

#### 5.1.1 Data Flow Diagrams

**Description:** The Data Flow Diagrams (DFDs) illustrate the flow of data within the diabetic detection system. These diagrams visually represent the processes, data storage, and interactions between various components.

**Level 0 DFD:**

Depicts the overall system as a single process.

Shows external entities such as users and data sources.

## Level 1 DFDs:

Breaks down the system into more detailed processes and data flows.

Highlights interactions between components, including the user interface, algorithms, and result pages.

### 5.1.2 User Stories

**Description:** User stories provide a narrative of how users interact with the system, outlining specific scenarios and desired outcomes.

#### **User Story 1:** Inputting Health Parameters

As a user, I want to easily input my health parameters on the user-friendly interface to initiate the diabetic detection process.

#### **User Story 2:** Viewing Results

As a user, I want to see clear and understandable result pages indicating whether I am diabetic, pre-diabetic, or non-diabetic based on the input parameters.

#### **User Story 3:** Real-time Feedback

As a user, I want to receive real-time feedback on my diabetic risk after submitting my health parameters.

## 5.2 Solution Architecture

**Description:** The Solution Architecture outlines the high-level structure of the diabetic detection system, detailing the key components and their interactions.

### **Web Interface Layer:**

User-friendly interface for inputting health parameters.

Result pages for displaying predictions.

### **Algorithmic Layer:**

Integration of machine learning algorithms (Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, AdaBoost Classifier) for diabetic risk prediction.

### **Empathy Map Integration Layer:**

Integration of insights from the Empathy Map Canvas into the algorithm for enhanced user-centric predictions.

## 6. PROJECT PLANNING & SCHEDULING

### 6.1 Technical Architecture

**Description:** The Technical Architecture outlines the key technological components and their interactions within the diabetic detection system.

#### **Web Technologies:**

Specify the technologies used for developing the user interface, such as HTML, CSS, and JavaScript.

#### **Backend Technologies:**

Detail the backend technologies, including the programming languages (e.g., Python), frameworks employed.

#### **Machine Learning Frameworks:**

Specify the machine learning frameworks utilized for implementing the algorithms (e.g., Scikit-learn, XGBoost).

#### **Security Measures:**

Outline the security protocols implemented to safeguard user data during input, processing, and storage.

### 6.2 Sprint Planning & Estimation

**Description:** Sprint Planning involves breaking down the project into manageable tasks and estimating the time required for each task. This iterative approach ensures continuous progress and adaptability.

#### **Task Breakdown:**

We identified total 4 sprints to get the information from some users.

#### **Estimation:**

Duration taken by each Sprint is listed below:

Sprint-1: 3 Days,

Sprint-2: 4 Days,

Sprint-3: 4 Days,

Sprint-4: 2 Days.

#### **Priority Assignment:**

We have collected information from the users initially, and parallelly we started building our machine learning model, finally we integrated the model with flask application to create a user-interface.



## 6.3 Sprint Delivery Schedule

**Description:** The Sprint Delivery Schedule outlines the timeline for completing individual sprints, ensuring a systematic and timely development process.

### **Sprint Duration:**

With the help of many people, we are able to complete the sprints within 2 weeks.

Sprint-1: 3 Days,

Sprint-2: 4 Days,

Sprint-3: 4 Days,

Sprint-4: 2 Days.

### **Sprint Goals:**

Clearly define the goals and deliverables for each sprint.

### **Review and Retrospective:**

Allocate time for sprint reviews to assess progress and retrospectives to identify areas for improvement.

This approach ensures a structured and adaptable project development process, allowing for continuous improvement and flexibility in response to evolving requirements.

## 7. CODING & SOLUTIONING

**Description:** In machine learning, a feature is data that's used as the input for the ML models to make predictions.

### 7.1 Feature 1

**HighBP:** Hypertension is twice as frequent in patients with diabetes compared with those who do not have diabetes. Moreover, patients with hypertension often exhibit insulin resistance and are at greater risk of diabetes developing than are normotensive individuals.

1 represents Normal, 0 represents High.

### 7.2 Feature 2

**HighChol:** Diabetes and cholesterol are two conditions that often go hand-in-hand. High cholesterol can be a cause of diabetes and vice versa.

1 represents Normal, 0 represents High.

### 7.3 Feature 3

**CholCheck:** Regular check of cholesterol and high cholesterol can be a cause of diabetes.

1 represents No, 0 represents Yes.

### 7.4 Feature 4

**BMI:** The higher your BMI, the higher your risk for certain diseases such as heart disease, high blood pressure, type 2 diabetes, gallstones, breathing problems, and certain cancers.

### 7.5 Feature 5

**Smoker:** Smoking is one cause of type 2 diabetes. In fact, people who smoke cigarettes are 30% – 40% more likely to develop type 2 diabetes than people who don't smoke. People with diabetes who smoke are more likely than those who don't smoke to have trouble with insulin dosing and with managing their condition.

1 represents No, 0 represents Yes.

### 7.6 Feature 6

**Stroke:** Diabetes is a well-established risk factor for stroke. It can cause pathologic changes in blood vessels at various locations and can lead to stroke if cerebral vessels are directly affected. Additionally, mortality is higher and poststroke outcomes are poorer in patients with stroke with uncontrolled glucose levels.

1 represents No, 0 represents Yes.

### 7.7 Feature 7

**HeartDiseaseorAttack:** People with diabetes are also more likely to have other conditions that raise the risk for heart disease: High blood pressure increases the force of blood through your arteries and can damage artery walls. Having both high blood pressure and diabetes can greatly increase your risk for heart disease.

1 represents No, 0 represents Yes.

### 7.8 Feature 8

**PhysActivity:** being active makes your body more sensitive to insulin (the hormone that allows cells in your body to use blood sugar for energy), which helps manage your diabetes. Physical activity also helps control blood sugar levels and lowers your risk of heart disease and nerve damage.

1 represents No, 0 represents Yes.

### 7.9 Feature 9

**Fruits:** eating fruit may actually help prevent diabetes. People with diabetes can eat any fruit they choose, as long as it fits within the carbohydrate “budget” in their daily food plan and they do not have an allergy to the fruit.

1 represents No, 0 represents Yes.

## 7.10 Feature 10

**Veggies:** Vegetables can play a valuable dietary role for people with type 2 diabetes. They provide fiber, antioxidants, and other nutrients that can help manage inflammation, support weight loss, and boost overall health. People with type 2 diabetes can eat any food, but they may need to plan carefully to avoid glucose spikes.

1 represents No, 0 represents Yes.

## 7.11 Feature 11

**HvyAlcoholConsump:** Regular heavy drinking can reduce the body's sensitivity to insulin, which can trigger type 2 diabetes. Diabetes is a common side effect of chronic pancreatitis, which may be caused by heavy drinking.

1 represents No, 0 represents Yes.

## 7.12 Feature 12

**AnyHealthcare:** medical care may be needed to treat the effects of diabetes: foot care to treat ulcers. screening and treatment for kidney disease.

1 represents No, 0 represents Yes.

## 7.13 Feature 13

**NoDocbcCost:**

It is a binary variable that indicates whether the respondent needed to see a doctor in the past 12 months but could not due to cost. The value 0 indicates that the respondent did not face any such situation, while the value 1 indicates that they did

1 represents No, 0 represents Yes.

## 7.14 Feature 14

**GenHlth:** General health care must be needed to treat the effects of diabetes as well as to avoid the pre-diabetes situation.

1 represents Very Bad, 2 represents Bad, 3 represents Normal, 4 represents Good, 5 for Healthy.

## 7.15 Feature 15

**MentHlth:** people with type 1 and type 2 diabetes are more likely to experience depression compared to those without diabetes and that people living with depression have a higher risk of developing type 2 diabetes.

## 7.16 Feature 16

**PhysHlth:** Physically fit and healthy people are not most likely to be get affected by diabetes.

## 7.17 Feature 17

**DiffWalk:** moving more can make a huge difference to how you feel and how you manage your condition. So, whether you have type 1, type 2 or another type of diabetes, walking is a good way to get physically active and build movement into your daily routine.

1 represents No, 0 represents Yes.

## 7.18 Feature 18

**Sex:** Worldwide, an estimated 17.7 million more men than women have diabetes mellitus. Women appear to bear a greater risk factor burden at the time of their type 2 diabetes diagnosis, especially obesity. Moreover, psychosocial stress might play a more prominent role in diabetes risk in women.

1 represents Female, 0 represents Male.

## 7.19 Feature 19

**Age:** Advanced age is a major risk factor for diabetes and prediabetes. Therefore, the elderly has a higher prevalence of diabetes and prediabetes than the young and middle-aged and are more likely to develop complications in the cardiovascular, retinal, and renal systems.

## 7.20 Feature 20

**Education:** An analysis is saying that glycemic control is better in more educated persons and level of education has an inverse relationship to the complication score. Percentage of patients with complication score more than 10 gradually decreases as the literacy increases from 5<sup>th</sup> Standard class onwards.

## 7.21 Feature 21

**Income:** An analysis is saying that after statistically controlling for other factors, lower income males were 94% more likely to have type 2 diabetes while lower income females were 175% more likely to have type 2 diabetes.

# 8. PERFORMANCE TESTING

## 8.1 Performance Metrics

### Accuracy of Algorithms:

- The effectiveness of the machine learning algorithms (Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, AdaBoost Classifier) is crucial. Regular validation and fine-tuning of these models with relevant datasets are essential to ensure accurate predictions.

### User Input and Experience:

- The user interface and the ease with which users can input their health parameters are critical. The empathy-driven approach, as indicated by the Empathy Map Integration, is a positive aspect, as it helps align the system with user needs and expectations.

### Real-time Feedback:

- Providing real-time feedback is a valuable feature, especially for users seeking instant insights into their diabetic risk. It enhances user engagement and contributes to a more dynamic user experience.

### Performance Metrics:

- The performance metrics outlined for testing, including response time, throughput, and error rate, are appropriate for ensuring the system's reliability and efficiency.

```
16 # Confusion Matrix and Classification Report
17 confusion_rf = confusion_matrix(y_test, y_pred_rf)
18 report_rf = classification_report(y_test, y_pred_rf)
19
20 print("Random Forest Classifier Accuracy:", accuracy_rf)
21 print("\nRandom Forest Classifier Confusion Matrix:")
22 print(confusion_rf)
23 print("\nRandom Forest Classifier Classification Report:")
24 print(report_rf)
25 print("\n")
26
27
```

Random Forest Classifier Accuracy: 0.9308854954687963

Random Forest Classifier Confusion Matrix:

```
[[40714   56  1918]
 [  811 41343   522]
 [ 4971   584 37303]]
```

Random Forest Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.88	0.95	0.91	42688
1.0	0.98	0.97	0.98	42676
2.0	0.94	0.87	0.90	42858
accuracy			0.93	128222
macro avg	0.93	0.93	0.93	128222
weighted avg	0.93	0.93	0.93	128222

```

12 # Calculate accuracy
13 accuracy_dt = accuracy_score(y_test, y_pred_dt)
14
15 # Confusion Matrix and Classification Report
16 confusion_dt = confusion_matrix(y_test, y_pred_dt)
17 report_dt = classification_report(y_test, y_pred_dt)
18
19 print("Decision Tree Classifier Accuracy:", accuracy_dt)
20 print("\nDecision Tree Classifier Confusion Matrix:")
21 print(confusion_dt)
22 print("\nDecision Tree Classifier Classification Report:")
23 print(report_dt)
24 print("\n")
25

```

Decision Tree Classifier Accuracy: 0.8684001185443996

Decision Tree Classifier Confusion Matrix:

```

[[36359  913  5416]
 [  726 39960 1990]
 [ 4780  3049 35029]]

```

Decision Tree Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.87	0.85	0.86	42688
1.0	0.91	0.94	0.92	42676
2.0	0.83	0.82	0.82	42858
accuracy			0.87	128222
macro avg	0.87	0.87	0.87	128222
weighted avg	0.87	0.87	0.87	128222

```

16 confusion_knn = confusion_matrix(y_test, y_pred_knn)
17 report_knn = classification_report(y_test, y_pred_knn)
18
19 print("K-Nearest Neighbors (KNN) Classifier Accuracy:", accuracy_knn)
20 print("\nK-Nearest Neighbors (KNN) Classifier Confusion Matrix:")
21 print(confusion_knn)
22 print("\nK-Nearest Neighbors (KNN) Classifier Classification Report:")
23 print(report_knn)
24 print("\n")
25

```

K-Nearest Neighbors (KNN) Classifier Accuracy: 0.8629018421175774

K-Nearest Neighbors (KNN) Classifier Confusion Matrix:

```

[[26462  5256 10970]
 [   86 42551   39]
 [  798  430 41630]]

```

K-Nearest Neighbors (KNN) Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.97	0.62	0.76	42688
1.0	0.88	1.00	0.94	42676
2.0	0.79	0.97	0.87	42858
accuracy			0.86	128222
macro avg	0.88	0.86	0.85	128222
weighted avg	0.88	0.86	0.85	128222

```

13 accuracy_lr = accuracy_score(y_test, y_pred_lr)
14
15 # Confusion Matrix and Classification Report
16 confusion_lr = confusion_matrix(y_test, y_pred_lr)
17 report_lr = classification_report(y_test, y_pred_lr)
18
19 print("Logistic Regression Classifier Accuracy:", accuracy_lr)
20 print("\nLogistic Regression Classifier Confusion Matrix:")
21 print(confusion_lr)
22 print("\nLogistic Regression Classifier Classification Report:")
23 print(report_lr)
24 print("\n")

```

/usr/local/lib/python3.10/dist-packages/sklearn/linear\_model/\_logistic.py:458: ConvergenceWarning: lbfgs STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max\_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

[https://scikit-learn.org/stable/modules/linear\\_model.html#logistic-regression](https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression)

```

n_iter_i = _check_optimize_result(
Logistic Regression Classifier Accuracy: 0.5325607150099047

```

Logistic Regression Classifier Confusion Matrix:

```

[[28137  7424  7127]
 [11437 14370 16869]
 [ 6542 10537 25779]]

```

```

[ 6542 10537 25779]]

```

Logistic Regression Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.61	0.66	0.63	42688
1.0	0.44	0.34	0.38	42676
2.0	0.52	0.60	0.56	42858
accuracy			0.53	128222
macro avg	0.52	0.53	0.52	128222
weighted avg	0.52	0.53	0.52	128222

```

17 report_xgb = classification_report(y_test, y_pred_xgb)
18
19 print("XGBoost Classifier Accuracy:", accuracy_xgb)
20 print("\nXGBoost Classifier Confusion Matrix:")
21 print(confusion_xgb)
22 print("\nXGBoost Classifier Classification Report:")
23 print(report_xgb)
24 print("\n")
25

```

XGBoost Classifier Accuracy: 0.8329303863611549

XGBoost Classifier Confusion Matrix:

```

[[41050    0  1638]
 [ 1098 36899  4679]
 [ 5633  8374 28851]]

```

XGBoost Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.86	0.96	0.91	42688
1.0	0.82	0.86	0.84	42676
2.0	0.82	0.67	0.74	42858
accuracy			0.83	128222
macro avg	0.83	0.83	0.83	128222
weighted avg	0.83	0.83	0.83	128222

```

14
15 # Confusion Matrix and Classification Report
16 confusion_adaboost = confusion_matrix(y_test, y_pred_adaboost)
17 report_adaboost = classification_report(y_test, y_pred_adaboost)
18
19 print("Adaptive Boosting (AdaBoost) Classifier Accuracy:", accuracy_adaboost)
20 print("\nAdaptive Boosting (AdaBoost) Classifier Confusion Matrix:")
21 print(confusion_adaboost)
22 print("\nAdaptive Boosting (AdaBoost) Classifier Classification Report:")
23 print(report_adaboost)
24 print("\n")
25

```

9 Adaptive Boosting (AdaBoost) Classifier Accuracy: 0.6704621671787993

Adaptive Boosting (AdaBoost) Classifier Confusion Matrix:

```

[[35058    0 7630]
 [ 2062 25322 15292]
 [ 4302 12968 25588]]

```

Adaptive Boosting (AdaBoost) Classifier Classification Report:

	precision	recall	f1-score	support
0.0	0.85	0.82	0.83	42688
1.0	0.66	0.59	0.63	42676
2.0	0.53	0.60	0.56	42858
accuracy			0.67	128222
macro avg	0.68	0.67	0.67	128222
weighted avg	0.68	0.67	0.67	128222

## 9. RESULTS

In this section, we delve into the presentation and interpretation of results generated by our diabetic detection system. The implementation involves a user-friendly interface, algorithmic predictions, and real-time feedback, creating a cohesive and insightful user experience.

### 9.1 Output Screenshots

#### 9.1.1 User Input Interface

### Diabetes Risk Assessment

To check if you have diabetes or not, please enter your details:

Full Name:

Blood Pressure:

Sex:

Cholesterol Level:

Cholesterol Check:

Stroke:

Smoker:

Alcoholic:

Heart Diseases:

Physical Activities:

Fruits:

Veggies:

Healthcare: id="healthcare">

NoDobbcCost:

Age:

BMI:

General Health:

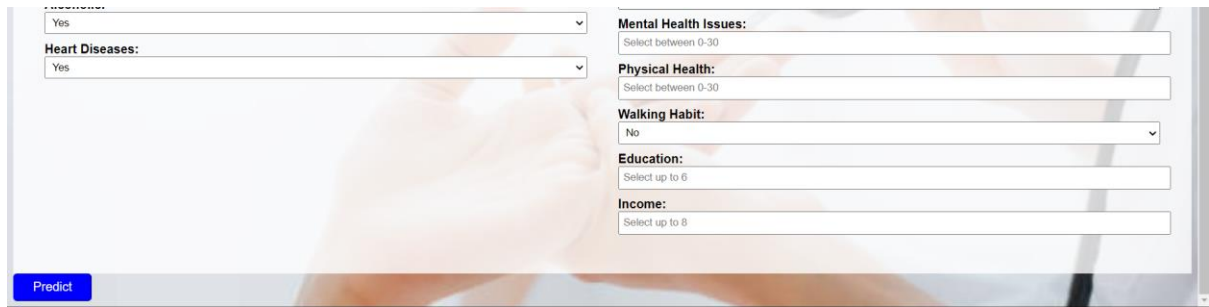
Mental Health Issues:

Physical Health:

Walking Habit:

Education:





Yes

Heart Diseases:

Yes

Mental Health Issues:

Select between 0-30

Physical Health:

Select between 0-30

Walking Habit:

No

Education:

Select up to 6

Income:

Select up to 8

Predict

### 9.1.2 Result Pages



nagendra  
MALE



You are fit and healthy!

## Maintain a Healthy Lifestyle

Here are a few tips to stay healthy:

- Eat a balanced diet rich in fruits and vegetables.
- Exercise regularly to stay active.
- Get enough sleep to recharge your body.
- Stay hydrated by drinking plenty of water.
- Manage stress through relaxation and meditation.

### 9.1.2.1 Diabetic Prediction

srihari

male



You are diabetic. Please take precautions.

## Tips to Manage Diabetes

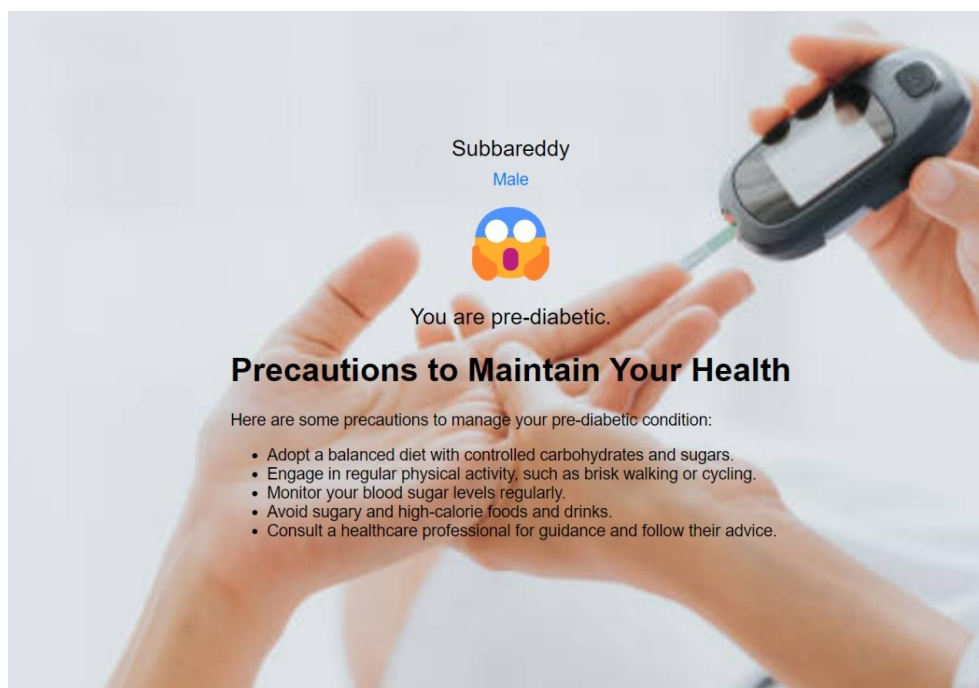
Here are some tips to manage your diabetes:

- Follow a balanced diet with controlled carbohydrates and sugars.
- Monitor your blood sugar levels regularly and take medication as prescribed.
- Engage in regular physical activity, such as walking or swimming.
- Limit alcohol consumption and avoid smoking.
- Consult your healthcare provider and follow their guidance for managing diabetes.

Upon submission of health parameters, the system processes the data using a combination of machine learning algorithms. The diabetic prediction result page is generated, indicating a high likelihood of diabetes based on the input parameters.

### 9.1.2.2 Pre-diabetic Prediction

For individuals at a risk threshold, the pre-diabetic prediction result page is presented. This outcome signals a need for increased awareness and potential preventive measures.



### 9.1.2.3 Non-diabetic Prediction

Users with input parameters indicating a low risk of diabetes receive the non-diabetic prediction result page, providing reassurance and encouraging a healthy lifestyle.



## 10. ADVANTAGES & DISADVANTAGES

### 10.1 Advantages:

#### Multi-Algorithmic Approach:

Utilizing multiple machine learning algorithms increases the robustness of the system.

Each algorithm may capture different patterns in the data, enhancing overall accuracy.

#### Web Interface for User-Friendly Interaction:

Flask-based web interface provides an accessible platform for users to input data and receive predictions.

Enhances user engagement and facilitates easy interaction with the system.

#### **Interpretability:**

Decision Trees and Logistic Regression are inherently interpretable, aiding in understanding the factors contributing to predictions.

Transparency in the decision-making process can build user trust.

#### **Predictive Accuracy:**

Ensemble methods like Random Forests and boosting algorithms (XGB, AdaBoost) often yield high predictive accuracy.

Improved performance in capturing complex relationships within the data.

#### **Customizable Features:**

Flask allows for easy customization and extension of the web interface.

Additional features or improvements can be seamlessly integrated.

## **10.2 Disadvantages:**

#### **Overfitting Potential:**

Depending on the dataset and algorithmic parameters, there might be a risk of overfitting.

Regularization techniques and cross-validation should be considered to mitigate this.

#### **Computational Intensity:**

Some algorithms, especially ensemble methods, can be computationally intensive.

Deployment on resource-constrained environments may require optimization.

#### **Data Sensitivity:**

The performance of machine learning models heavily depends on the quality and representativeness of the training data. Biases or inaccuracies in the dataset may affect predictions.

#### **Interpretability Trade-off:**

While Decision Trees and Logistic Regression offer interpretability, complex ensemble methods may sacrifice interpretability.

Balancing accuracy and interpretability should be considered based on the application.

#### **Algorithm Selection Complexity:**

Choosing and tuning multiple algorithms can be complex.

Requires expertise to optimize hyperparameters and select algorithms based on the specific characteristics of the dataset.

## 11. CONCLUSION

In conclusion, the diabetic detection project utilizing a diverse set of machine learning algorithms and a user-friendly Flask web interface presents a promising approach to addressing a critical healthcare concern. Through the integration of Random Forests, Logistic Regression, Decision Trees, KNN, XGB Classifier, and AdaBoost Classifier, the system aims to provide accurate predictions based on 21 input parameters.

The advantages of this approach lie in its multi-algorithmic strategy, enhancing predictive accuracy and interpretability. The user-friendly Flask interface facilitates seamless interaction, allowing individuals to input their health data and receive predictions effortlessly. The transparency provided by interpretable models such as Decision Trees and Logistic Regression fosters user trust and understanding.

However, challenges and considerations must be acknowledged. The potential for overfitting, computational intensity, and sensitivity to the quality of training data necessitate careful model tuning and dataset curation. Balancing interpretability and accuracy, especially with complex ensemble methods, requires thoughtful consideration in the context of the application.

The project's success is contingent on ongoing optimization, security measures for the web interface, and potential enhancements to address algorithmic complexities. Future iterations could explore additional features, refine algorithms, and incorporate user feedback for continuous improvement.

In essence, the diabetic detection system represents a valuable step towards leveraging machine learning for proactive health management. Its strengths and weaknesses provide a foundation for further research and development in the realm of predictive healthcare applications.

## 12. FUTURE SCOPE

The future scope of the diabetic detection project holds immense potential for advancements and expansion. Here are several avenues to explore and consider for future development:

### Feature Enhancement:

- To investigate the inclusion of additional relevant features that could contribute to improved predictive accuracy.
- To collaborate with healthcare professionals to identify and incorporate new indicators or advancements in diabetic risk assessment.

### Algorithmic Refinement:

- Continuously refine and optimize the machine learning algorithms used in the system.

- Explore newer algorithms or variations to enhance the model's ability to capture complex relationships within the data.

### Integration of Continuous Monitoring:

- Explore the feasibility of incorporating continuous monitoring of health parameters, enabling real-time updates and personalized feedback.
- Develop mechanisms to adapt the model based on evolving health patterns.

### User Feedback and Iterative Improvement:

- Collect user feedback on the web interface and prediction outcomes to understand user experiences.
- Iteratively improve the system based on user suggestions and emerging healthcare trends.

### Cross-Domain Collaboration:

- Collaborate with healthcare institutions, research organizations, and technology companies to gather diverse and large-scale datasets.
- Engage in cross-disciplinary partnerships to enrich the model's training data and generalizability.

### Explanatory Model Interpretation:

- Enhance the interpretability of the machine learning models, especially for complex ensemble methods.
- Provide users with detailed explanations of how specific features contribute to the prediction, fostering transparency and user trust.

### Mobile Application Development:

- Extend the reach of the diabetic detection system by developing a mobile application.
- Enable users to access the system on-the-go, promoting continuous health monitoring and proactive management.

### Population-Specific Customization:

- Investigate the customization of the model for specific population groups based on demographic and regional health variations.
- Adapt the system to cater to diverse healthcare needs and disparities.

### Long-Term Health Prediction:

- Explore the potential for extending predictions beyond immediate diabetic risk to long-term health outcomes.
- Consider incorporating predictive analytics for lifestyle-related diseases beyond diabetes.

### Ethical Considerations and Data Privacy:

- Stay abreast of evolving ethical considerations in healthcare AI.
- Prioritize robust data privacy measures to ensure the security and confidentiality of user health data.

### Clinical Validation:

- Collaborate with healthcare professionals to validate the model's predictions in clinical settings.
- Conduct studies to assess the system's impact on early detection and intervention.

By embarking on these future directions, the diabetic detection project can evolve into a dynamic and responsive system, contributing significantly to the advancement of predictive healthcare technologies.

## 13. APPENDIX

In the appendix section, you can include additional information that supports and complements the main document. Here are some items you might consider including:

### Code Snippets:

```

from flask import Flask, render_template, request
import pickle
import numpy as np
app=Flask(__name__)
model = pickle.load(open('model.pkl','rb'))
import flask
print(flask.__version__)

```

Comment Code

```

@app.route("/", methods=['GET', 'POST'])
def startone():
    if request.method == 'POST':
        un = request.form['un']
        pw = request.form['pw']
        if un == 'admin' and pw == 'admin':
            return render_template('index.html')
    return render_template('login.html')

```

Comment Code

```

@app.route("/index.html")
def start():
    return render_template('index.html')

```

Comment Code

```

@app.route("/about.html")
def about():
    return render_template('about.html')

```

Comment Code

```

@app.route("/check.html")
def check():
    return render_template('check.html')

```

Comment Code

```

@app.route('/predict',methods=['POST'])
def home():
    fn=request.form['fn']
    a=request.form['a']

```

## Dataset Details:

- The dataset "diabetes\_012\_health\_indicators\_BRFSS2015" originates from the Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The dataset contains the columns of Certainly

Diabetes\_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income

The dataset contains imbalance of the data which is handled with SMOTE technique.



Model Evaluation Metrics:

```
Random Forest Classifier Accuracy: 0.9318135733337493

Random Forest Classifier Confusion Matrix:
[[40686   57  1945]
 [  794 41404   478]
 [ 4937   532 37389]]

Random Forest Classifier Classification Report:
              precision    recall  f1-score   support

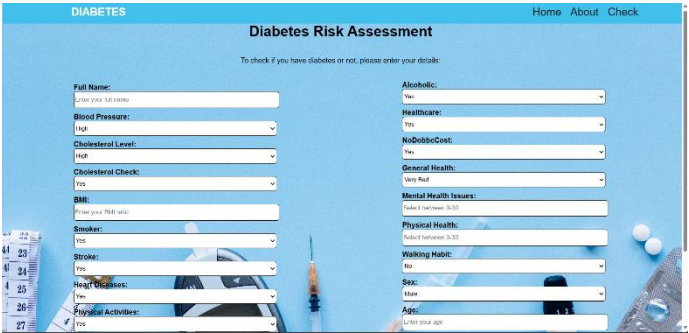
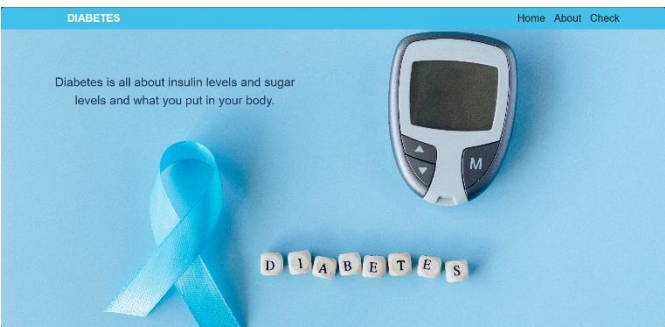
    0.0         0.88      0.95      0.91      42688
    1.0         0.99      0.97      0.98      42676
    2.0         0.94      0.87      0.90      42858

 accuracy         0.93
  macro avg         0.93
 weighted avg         0.93
```

Hyperparameter Tuning Results:

```
Best parameters: {'n_estimators': 100, 'min_samples_split': 3, 'min_samples_leaf': 2, 'max_depth': None}
Best cross-validation score: 0.92
Test set accuracy: 0.93
```

Web Interface Screenshots:



Name : Somala Nagendra babu  
Gender : Male

You are fit and healthy!

**Maintain a Healthy Lifestyle**

Here are a few tips to stay healthy:

- Eat a balanced diet rich in fruits and vegetables.
- Exercise regularly to stay active.
- Get enough sleep to recharge your body.
- Stay hydrated by drinking plenty of water.
- Manage stress through relaxation and meditation.

References to External Libraries:

Flask – version – 2.3.3

## Hardware and Software Requirements:

- Ensure that the server has Python installed, and it's preferable to use a virtual environment to manage dependencies.

## Alternative Approaches Considered:

- We explored various algorithms such as Decision Tree, Random Forest, K-Nearest Neighbors, Logistic Regression, XGBoost, and Adaptive Boosting

## Source Code:

You can view the entire Machine Learning Model code in this below given link:

[https://colab.research.google.com/drive/1Q8GoWMWimFuCfgJHpRq4qGtrjrMh3l4M#scrollTo=ccXNGPhS\\_Prq](https://colab.research.google.com/drive/1Q8GoWMWimFuCfgJHpRq4qGtrjrMh3l4M#scrollTo=ccXNGPhS_Prq)

## GitHub Link:

<https://github.com/nagendrasomala/diabetes-prediction>

## Project Demo Link:

[https://drive.google.com/file/d/1HtVrQPla2Sg2vh5lYErGe5PLi\\_TZoeas/view?usp=drivesdk](https://drive.google.com/file/d/1HtVrQPla2Sg2vh5lYErGe5PLi_TZoeas/view?usp=drivesdk)