DIABETES PREDICTION USING MACHINE LEARNING

PROJECT DOCUMENTATION

Team ID – 592321

1. INTRODUCTION

This research initiative focuses on harnessing machine learning algorithms to forecast the likelihood of diabetes onset in individuals based on their medical records and pertinent factors, including age, body mass index (BMI), familial medical history, and lifestyle choices. The dataset under scrutiny encompasses comprehensive clinical parameters, spanning blood pressure, BMI, heart conditions, and cholesterol levels.

The primary aim is to construct a predictive model capable of accurately identifying individuals with a high susceptibility to diabetes, thereby facilitating early intervention and preemptive measures against the disease. Leveraging machine learning methodologies to scrutinize substantial data sets allows for the recognition of discernible patterns and the formulation of precise predictions, potentially translating into life-saving implications.

In essence, this undertaking holds promise for the healthcare domain, given its potential to enhance the early identification and prevention of diabetes, thus fostering improved health outcomes for both individuals and communities.

1.1 Project Overview

The project aims to develop a Diabetes Prediction System using advanced machine learning techniques for early detection and proactive management of diabetes.

1.2 Purpose

The purpose of this system is to assist healthcare professionals in accurately predicting the likelihood of an individual developing diabetes, enabling timely intervention and personalized patient care.

2. LITERATURE SURVEY

2.1 Existing Problem

The existing problem lies in the lack of efficient tools for early diabetes detection, leading to delayed interventions and compromised patient outcomes.

2.2 References

- ML Concepts

- Supervised learning: https://www.javatpoint.com/supervised-machine-learning
- Unsupervised learning: https://www.javatpoint.com/unsupervised-machine-learning
- Decision tree: https://www.javatpoint.com/machine-learning-decision-tree-classificationalgorithm
- Random forest: https://www.javatpoint.com/machine-learning-random-forest-algorithm
- KNN: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
- Xgboost: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-tounderstand-the-math-behind-xgboost/
- Evaluation metrics: https://www.analyticsvidhya.com/blog/2019/08/11-important-modelevaluation-error-metrics/

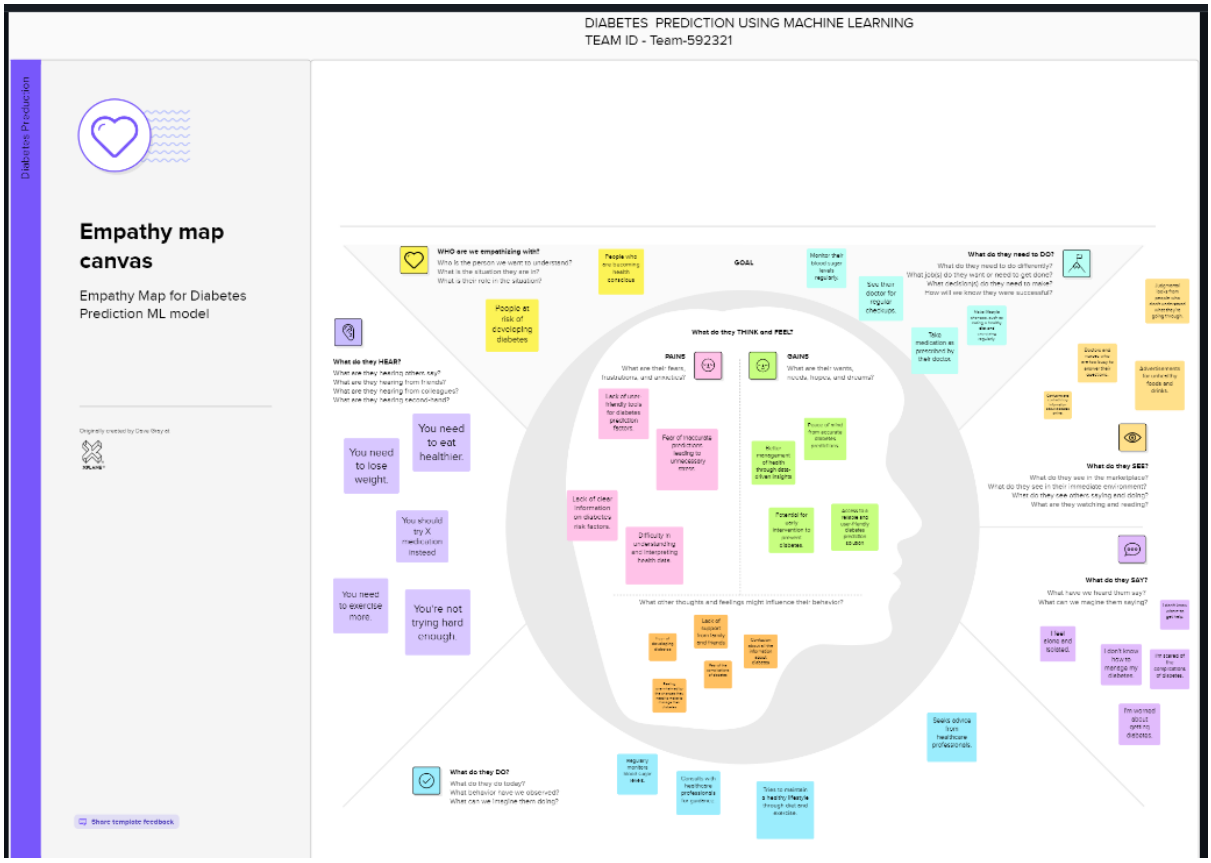
- NLP:- https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_python.htm
- Flask Basics: https://www.youtube.com/watch?v=lj4I_CvBnt0

2.3 Problem Statement Definition

The project addresses the need for a reliable and accurate diabetes prediction system to improve preventive healthcare and enhance patient well-being.

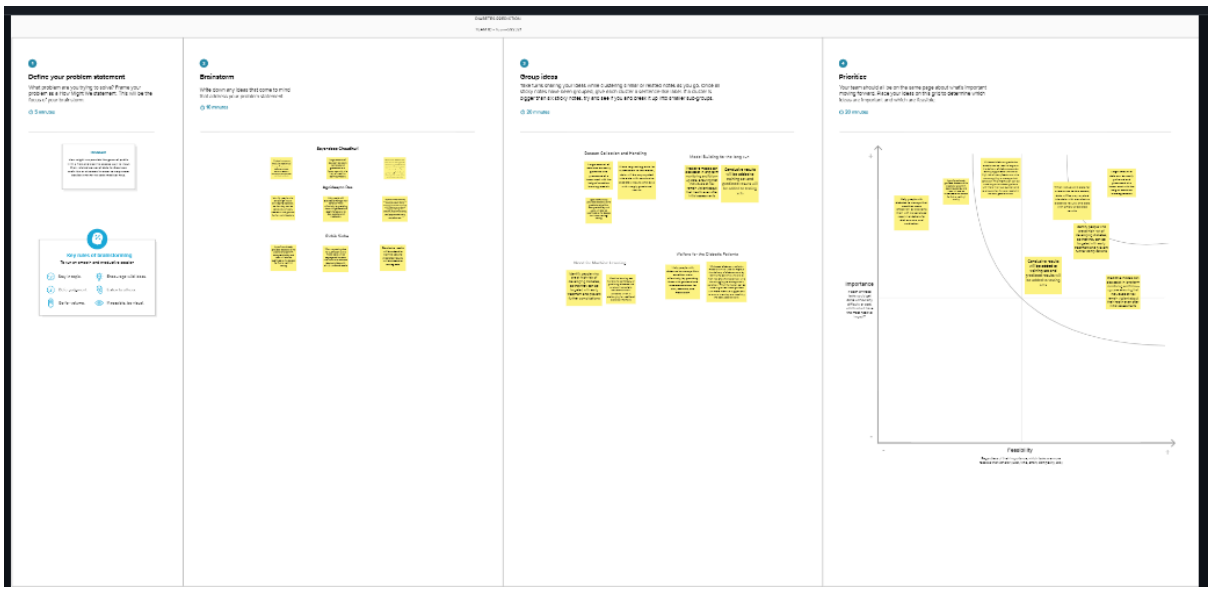
3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming

The proposed solution involves leveraging machine learning algorithms to analyze health indicators and risk factors for accurate diabetes prediction.



4. REQUIREMENT ANALYSIS

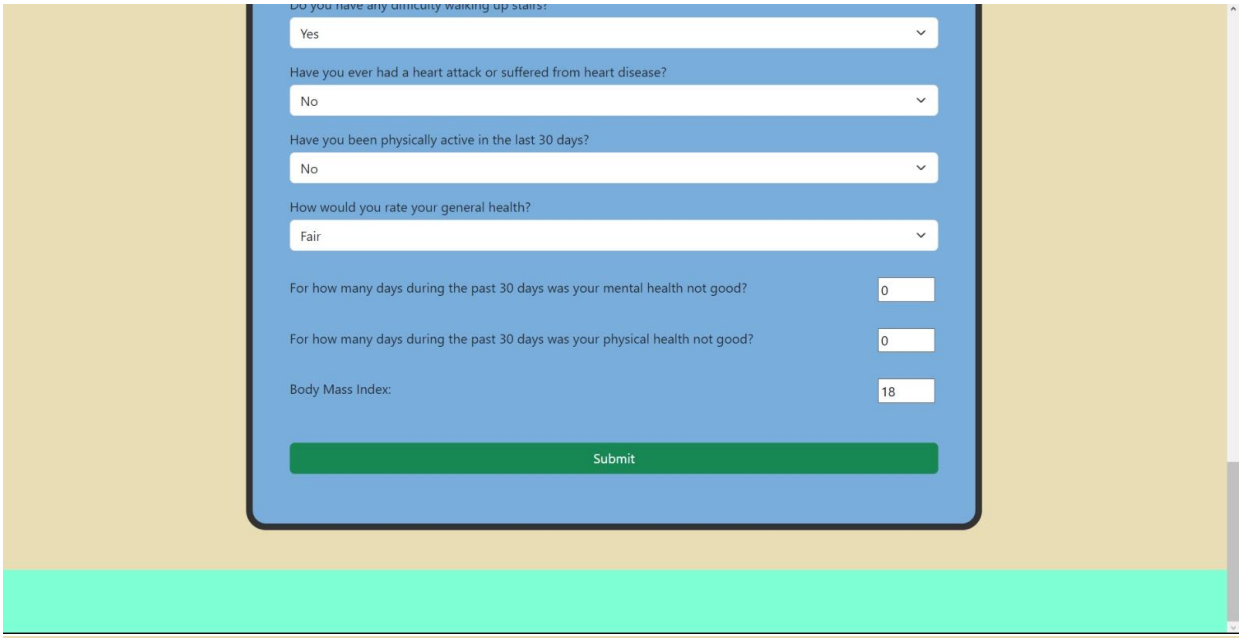
4.1 Functional Requirements

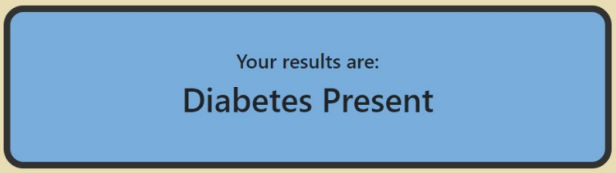
- Data input interface
- Machine learning model integration
- Prediction result display

4.2 Non-Functional Requirements

- User-friendly interface
- High accuracy and reliability
- Scalability for future enhancements


Enter Your Details
Sex:
Female
Age:
70-74
How many years of education have you had?
Grades 1 through 7
What is your estimated total household income?
$30000 - $40000
Do you have any health care coverage?
Yes
Was there a time in the past year when you needed to see a doctor but could not because of costs?
No
Do you have high blood pressure?
Yes
Do you have high cholestrol?
Yes
When was the last time you got your cholestrol checked?
Last 5 Years
What is your daily fruit intake?
Almost never
What is your daily vegetable intake?
Almost never
How many cigarettes have you smoked in your entire life?
None
Have you ever suffered a stroke?
No
Are you a heavy drinker?
No
Do you have any difficulty walking up stairs?
Yes
Have you ever had a heart attack or suffered from heart disease?
No
Have you been physically active in the last 30 days?
No
How would you rate your general health?


Do you have any difficulty walking up stairs?
Yes
Have you ever had a heart attack or suffered from heart disease?
No
Have you been physically active in the last 30 days?
No
How would you rate your general health?
Fair
For how many days during the past 30 days was your mental health not good?
0
For how many days during the past 30 days was your physical health not good?
0
Body Mass Index:
18
Submit

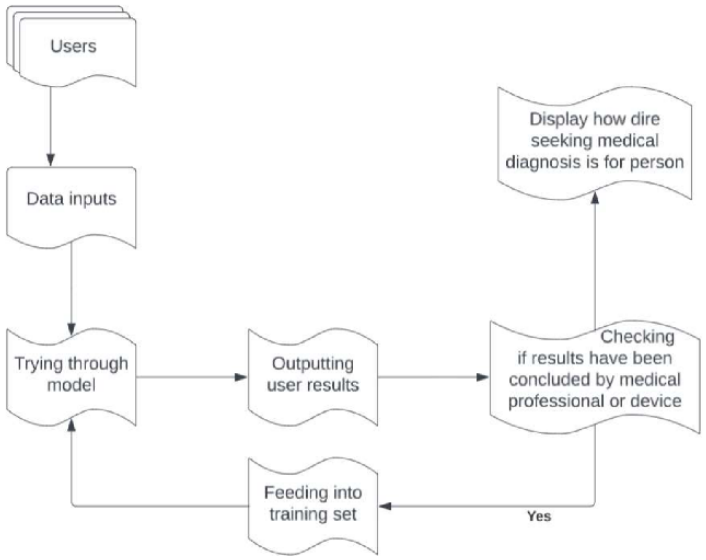

Your results are:
Diabetes Present

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

Data Flow Diagrams:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

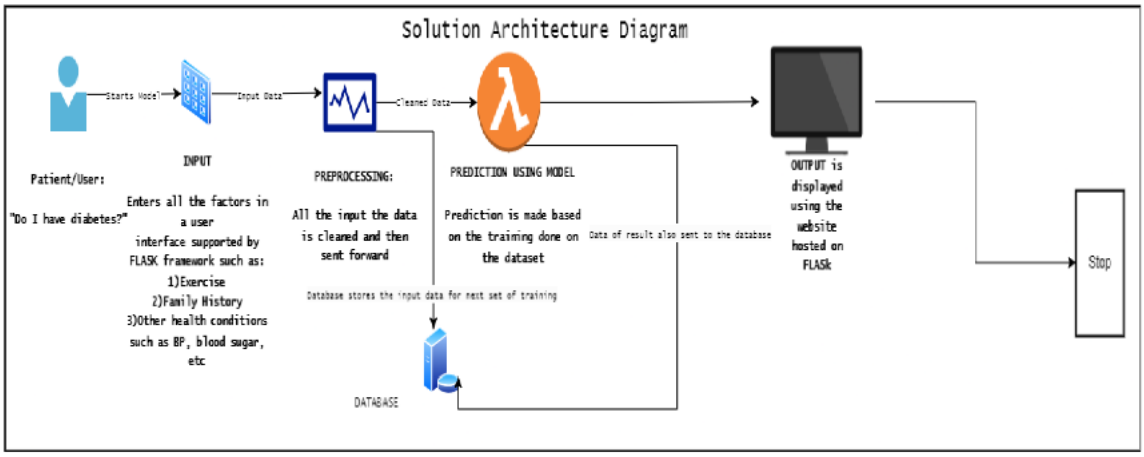


User Stories:

User type	Functional Requirement	User Story No.	User Story	Acceptance Criteria	Priority
Customer	Data Inputs	USN-1	As a user, I need to input my details to get a diagnosis but I do not have my family history	I will input my personal data to my best possible knowledge	Medium
		USN-2	As a user, I can input all my details to get a diagnosis	I will input all required data to the best of my knowledge	High
		USN-3	As a user, I am unaware of my personal medical data but I can input an estimation of the same	I will input values in accordance to what I believe are correct	Low
	Medically diagnosed	USN-4	As a user, I have been medically diagnosed and would like to input my details to feed the model better for more accurate diagnoses in the future for me and my loved ones who might need it	I will input my medical data and input my diagnosis as a professionally backed one	High
		USN-5	As a user, I have not been medically diagnosed yet and would like to input my details to get an estimate of if I should seek out the same	I will input my most latest accurate medical data to get the model's results	Medium

5.2 Solution Architecture

Solution Architecture Diagram



6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture

Table-1 : Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	User gives input in the website created using Flask along with HTML templates	HTML, CSS, Flask
2.	Database	Local filesystem is use to hold the datasets	Local Filesystem
3.	File Storage	Local filesystem is use to hold the datasets	Local Filesystem
4.	Machine Learning Model	A CNN model comprised of Sigmoid/ReLU neurons in order to make a binary classifier	Binary Prediction Model, etc.
5.	Infrastructure (Server / Cloud)	Application Deployment on Local System	Localhost

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	TensorFlow APIs with keras, Flask framework, NumPy and Pandas frameworks, Scikit-learn framework, Seaborn framework	NumPy and Pandas are used for data manipulation and preprocessing. TensorFlow is used to make the CNN model. Scikit-learn is used to scale the dataset and also to evaluate metrics. Seaborn is used to do various variate analysis.
2.	Scalable Architecture	The machine learning model can be deployed on a larger scale using AWS, where datasets can be held in a server,	AWS for model deployment and hosting servers, Django for making the website for a larger scalable

		hence allowing cloud computing. Flask mircoframework can be scaled largely by changing to Django.	architecture.

S.No	Characteristics	Description	Technology
3.	Availability	Using AWS servers help in balancing the load and can also be used for cloud computing to further fasten the computations. Also, the source code will be available on GitHub for open source availability.	AWS services will be used for scalability while source code is available publicly on GitHub

6.2 Sprint Planning & Estimation

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement	User Story No.	User Story	Acceptance Criteria	Story Points
Sprint-1	Data Inputs	USN-1	As a user, I need to input my details to get a diagnosis but I do not have my family history	I will input my personal data to my best possible knowledge	10
Sprint-1		USN-2	As a user, I can input all my details to get a diagnosis	I will input all required data to the best of my knowledge	15
Sprint-4		USN-3	As a user, I am unaware of my personal medical data but I can input an estimation of the same	I will input values in accordance to what I believe are correct	5
Sprint-3	Medically diagnosed	USN-4	As a user, I have been medically diagnosed and would like to input my details to feed the model better for more accurate diagnoses in the future for me and my loved ones who might need it	I will input my medical data and input my diagnosis as a professionally backed one	15
Sprint-1		USN-5	As a user, I have not been medically diagnosed yet and would like to input my details to get an estimate of if I should seek out the same	I will input my most latest accurate medical data to get the model's results	10

6.3 Sprint Delivery Schedule

Project Tracker, Velocity & Burndown Chart:

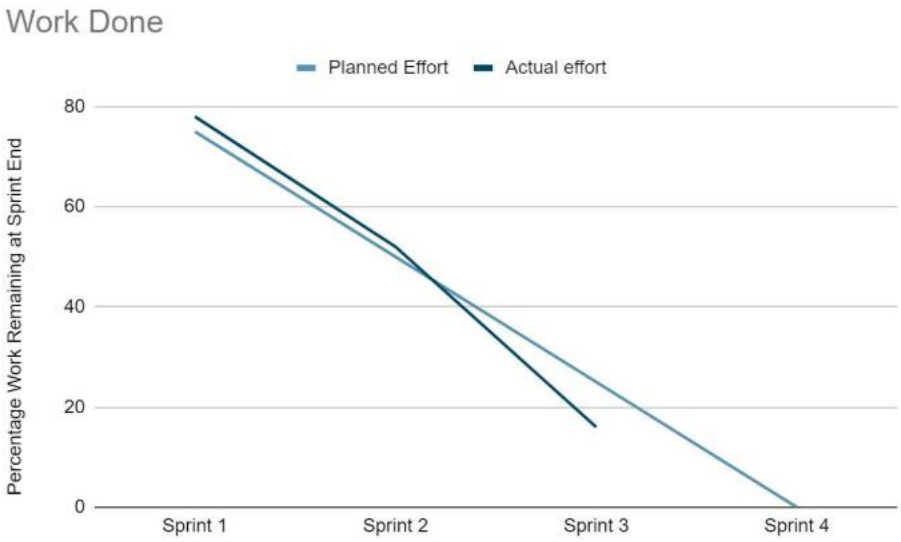
Sprint	Story Points	Duration	Sprint Start Date	Sprint End Date
Sprint-1	15	5 days	20-10-2023	25-10-2023
Sprint-2	15	5 days	25-10-2023	30-10-2023
Sprint-3	15	5 days	30-10-2023	04-11-2023
Sprint-4	15	4 days	04-11-2023	08-11-2023

Average Velocity:

The average velocity of our work for 5-day sprints with 15 story points for each sprint is:

$$AV = \frac{Points\ per\ sprint}{Days\ per\ sprint} = \frac{15}{5} = 3$$

Burndown Chart:



7. CODING & SOLUTIONING

7.1 TensorFlow Model

```
model=Sequential()
model.add(Dense(30,input_dim=21,activation='relu'))
model.add(Dense(10,activation='relu'))
model.add(Dense(1,activation='sigmoid'))
```

```
[ ] model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
[ ] model.fit(x_train, y_train, epochs=50, batch_size=10)
```

7.2 Random Forest Classifiers

```
[ ] rf=RandomForestClassifier(max_depth=100,n_estimators=50,random_state=0)
    rf.fit(x_train,y_train)
```

```
RandomForestClassifier
RandomForestClassifier(max_depth=100, n_estimators=50, random_state=0)
```

7.3 Decision Tree Classifiers

Decision Tree Classifier

```
[ ] dt=DecisionTreeClassifier(max_depth=12)
    dt.fit(x_train,y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=12)
```

7.4 Logistic Regression

Logistic Regression

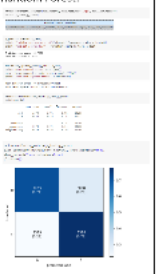
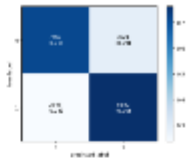
```
[ ] lg=LogisticRegression(max_iter=1500)
    lg.fit(x_train,y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=1500)
```

8. PERFORMANCE TESTING

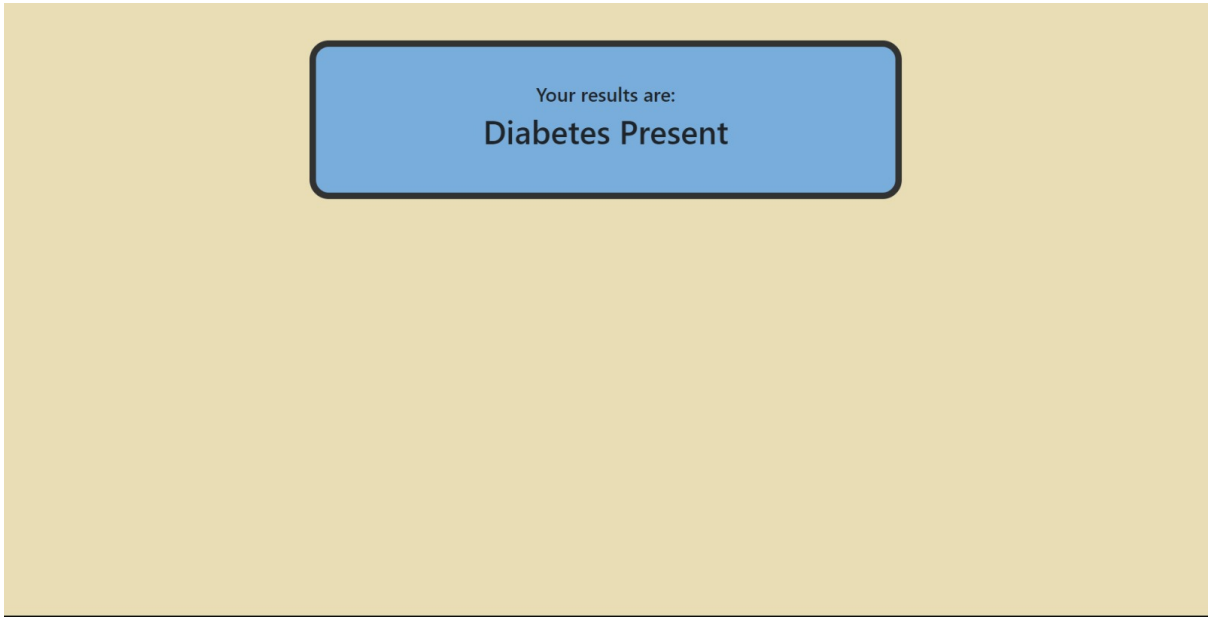
8.1 Performance Metrics

Model Performance Testing:

S.No.	Parameter	Values	Screenshot
1.	Metrics	**Regression Model:** MAE - , MSE - , RMSE - , R2 score - **Classification Model:** Confusion Matrix - , Accuracy Score- & Classification Report	Random Forest:  Decision Tree: 
			 Logistic Regression: 
2.	Tune the Model	Hyperparameter Tuning - Validation Method	Random Forest:  Decision Tree:  Logistic Regression:

9. RESULTS

9.1 Output Screenshots



10. ADVANTAGES & DISADVANTAGES

Advantages

1. Early Detection: Enables early identification of individuals at risk of developing diabetes, facilitating timely intervention and preventive measures.

2. Personalized Care: Facilitates personalized healthcare plans based on individual risk assessments, leading to improved patient outcomes.

3. Efficient Healthcare Management: Helps healthcare providers allocate resources effectively by focusing on high-risk individuals and optimizing treatment strategies.

Disadvantages

1. Dependency on Data Quality: Relies heavily on the quality and quantity of the input data, making the system susceptible to inaccuracies and biases if the data is incomplete or biased.

2. Ethical Concerns: Raises concerns about data privacy and confidentiality, necessitating robust data security measures to safeguard sensitive patient information.

3. Technological Limitations: May face limitations in accurately predicting diabetes in certain populations or individuals with complex health conditions, necessitating continuous model refinement and validation.

11. CONCLUSION

The development of the Diabetes Prediction System represents a significant step towards proactive diabetes management and improved healthcare outcomes.

12. FUTURE SCOPE

The system can be further enhanced to incorporate additional health parameters and to support predictive analytics for other health conditions.

13. APPENDIX

https://colab.research.google.com/drive/1Q_TsgQX6wZMhQwaeyT7Gaq3Ax1m8Lr2a?usp=sharing-
https://drive.google.com/file/d/1aXkLtu8iF2-6RK2C5x_P-RgO0CQrLnIx/view?usp=drive_link