Project Report

Date	19 NOVEMBER 2023
Team ID	Team-591645
Project Name	Diabetes Prediction Using Machine Learning

1. INTRODUCTION

1.1 Project Overview

Objective: Develop a machine learning model to predict diabetes onset, enabling early intervention and prevention.

Dataset: Comprehensive health records including blood pressure, BMI, heart diseases, cholesterol levels, age, and family history.

Methodology:
1. Data Preprocessing: Clean and handle missing values.
2. Feature Selection: Identify key predictive features.
3. Model Selection: Experiment with logistic regression, decision trees, random forests, and support vector machines.
4. Training: Optimize model accuracy using a subset of the dataset.
5. Validation: Ensure model generalizability and assess performance.

Expected Outcomes:
1. Accurate Predictions: High accuracy in identifying individuals at risk.
2. Risk Factor Identification: Insights into key contributors to diabetes onset.
3. Early Intervention: Facilitate targeted preventive measures through early identification.

Impact: Improve public health outcomes by revolutionizing diabetes prevention through early detection and intervention.

1.2 Purpose

The Purpose of this project is to identify individuals who may develop diabetes at an early age by using computer programs and data analysis. Primary objectives are:

1. Early Identification: Develop a method to identify individuals who are at risk of developing diabetes before they exhibit any symptoms.
2. Preventive Action: Assist medical professionals and individuals in taking proactive measures to avoid or control diabetes.
3. Data analysis: Look for trends that point to a higher risk of diabetes by utilizing data on blood pressure, weight, and family history.
4. Healthcare Improvement: By figuring out how to anticipate and prevent diabetes, you may help make healthcare better and eventually help people live healthier lives.
5. Machine Learning Application: Investigate how machine learning methods can be used to evaluate big health data sets, producing precise forecasts that may even save lives.

To put it another way, the goal is to employ technology to identify diabetes early on so that people may take preventative measures to stay well. It's all about leveraging data to identify trends and support physicians in making better choices for their patients. The project's overall goal is to improve diabetes detection and prevention, which will benefit healthcare.

2. LITERATURE SURVEY

2.1 Existing problem

Background:

Diabetes poses a significant global health challenge, with its prevalence on the rise. Early detection and preventive measures are crucial for managing this chronic condition effectively. However, the existing healthcare landscape faces challenges in identifying individuals at risk of developing diabetes in a timely manner.

Challenges:

1. Late-stage Diagnosis: Nowadays, a large number of diabetes patients are identified when symptoms become noticeable. As a result, intervention is delayed and preventive measures' efficacy is reduced.
2. Limited Predictive Tools: Accurately forecasting diabetes risk is difficult with the incomplete tools available to traditional healthcare practices. This makes it more difficult to put preemptive initiatives for high-risk individuals into action.
3. Data Overload: Clinical measurements and lifestyle data are among the many types of patient data that healthcare systems produce in enormous quantities. Nevertheless, it is challenging to draw significant conclusions and patterns from this data due to a lack of advanced research techniques.
4. Resource Constraints: Healthcare workers frequently encounter time and technological resource limitations that impede their capacity to evaluate large datasets and generate early forecasts.

5. Ineffective Risk Communication: The current system struggles to effectively communicate the risk of diabetes to individuals, limiting their awareness and understanding of preventive actions.

Project Rationale:
The aforementioned challenges highlight the need for a more proactive and data-driven approach to diabetes prediction. This project seeks to address these issues by leveraging machine learning techniques to develop a predictive model. The aim is to overcome the limitations of current practices, enabling early identification of individuals at high risk and facilitating timely preventive interventions. Through the implementation of this project, we aspire to contribute to a paradigm shift in diabetes management, enhancing the overall effectiveness of healthcare strategies and improving health outcomes for individuals and communities.

2.2 References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/
2. https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012062/pdf
3. Health Organization, World. (Year). "Guidelines for Diabetes Prediction and Prevention." *World Health Organization Publications*, Publication Number.
4. https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/
5. Python Software Foundation. (Year). "Scikit-learn Documentation." *Scikit-learn Version*
6. Kaggle. (Year). "Diabetes Prediction Dataset." *Kaggle Datasets*

2.3 Problem Statement Definition

Problem:
The delayed recognition of individuals susceptible to diabetes is a significant concern in healthcare, resulting in missed opportunities for early intervention and prevention.

Challenges:
1. Conventional diagnostic methods often identify diabetes only when symptoms are apparent, limiting the effectiveness of preventive measures.
2. Analyzing a diverse array of health data, including blood pressure, BMI, and family history, poses a challenge in detecting subtle yet critical risk patterns.
3. The absence of advanced predictive tools in healthcare hinders the accuracy and efficiency of predicting the risk of developing diabetes.

Project Significance:
This project seeks to address these challenges by harnessing the power of machine learning to predict diabetes early. The ultimate goal is to transform healthcare practices, enabling timely and targeted interventions that have the potential to significantly improve health outcomes for individuals at risk of diabetes.

3. **IDEATION & PROPOSED SOLUTION**

3.1 Empathy Map Canvas

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user's behaviors and attitudes.

It is a useful tool to help teams better understand their users.

Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exercise of creating the map helps participants consider things from the user's perspective along with his or her goals and challenges.

Problem Statement:

Improving the early detection of diabetes using machine learning to enhance patient outcomes and simplify monitoring and management.

In this project, we aim to use machine learning algorithms to predict the onset of diabetes in individuals based on their health records and other relevant factors such as age, BMI, family history, and lifestyle habits. The dataset used in this project will include information on various clinical parameters such as blood pressure, BMI, Heart diseases and cholesterol levels.

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 analyze large amounts of data, we can identify patterns and make accurate predictions that could potentially save lives.

Overall, this project has the potential to contribute to the field of healthcare by improving early detection and prevention of diabetes, ultimately leading to better health outcomes for individuals and communities.

Reference:

https://app.mural.co/t/aiml3182/m/aiml3182/1698831903030/134f5459c44ae220956371faf91035f66ca5fb07?sender=u10d9bb41eecb4475d9751169

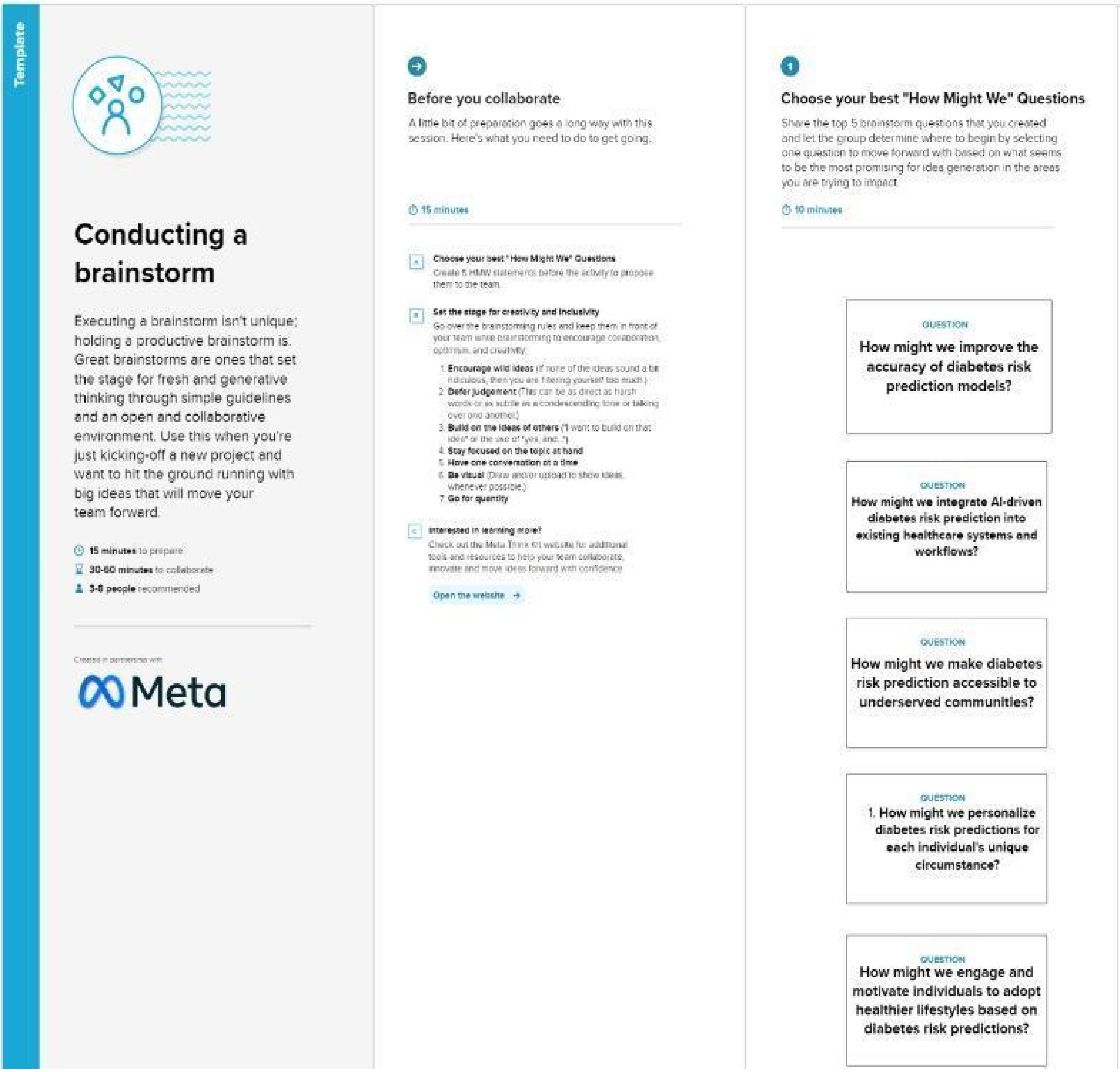
Empathy Mapping:

3.2 Ideation & Brainstorming

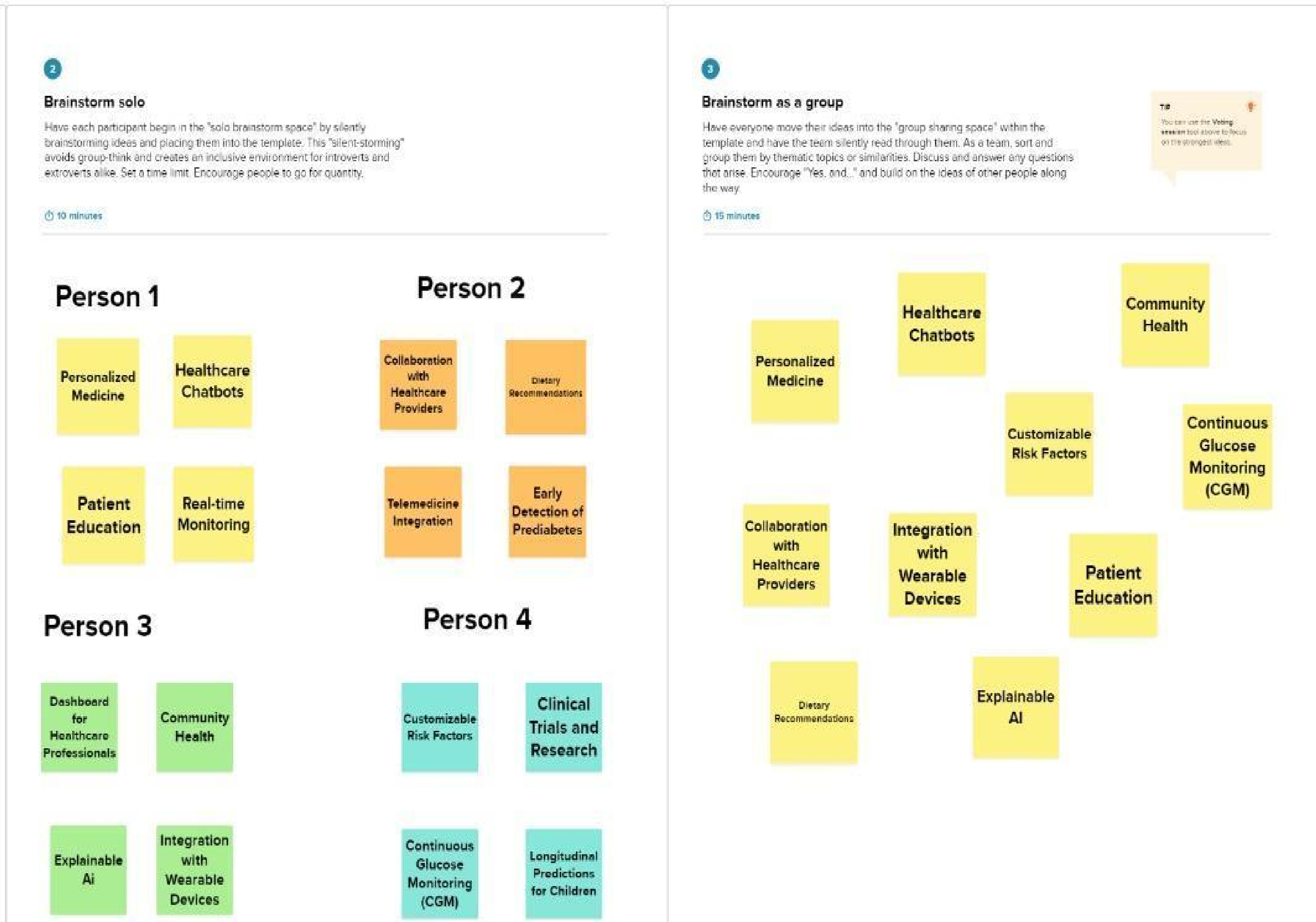
Brainstorm & Idea Prioritization Template:

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions.

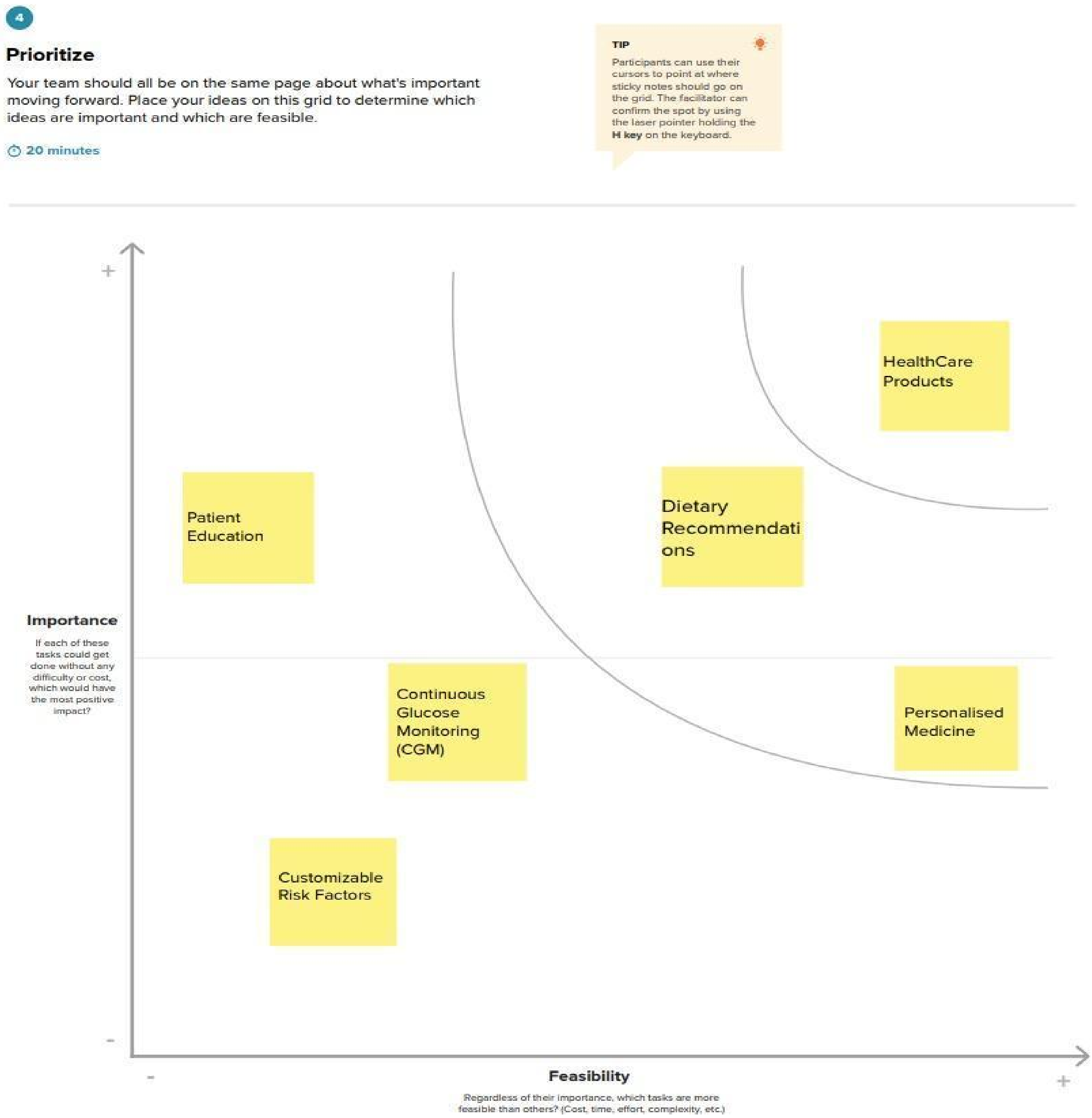
Step-1: Team Gathering, Collaboration and Select the Problem Statement



Step-2: Brainstorm, Idea Listing and Grouping



Step-3: Idea Prioritization Reference:



https://app.mural.co/t/vitapuniversity7296/m/vitapuniversity7296/1698841045421/1b15491df4a6a16542c7ad9ec9d0c1ff313d1385?sender=u623bf941642d1646117b6659
https://app.mural.co/t/project5088/m/project5088/1698845618841/be378332c291f8b51c0229a2160268a8b1e2e969?sender=u90a5584c59485c7f9fd98440

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

Functional requirements describe the specific features and capabilities that the system must have to meet its intended purpose. For the "Diabetes Prediction Using Machine Learning" project, the functional requirements include:

1. Data Input:

The system accepts diverse health data, including blood pressure, BMI, heart diseases, cholesterol levels, age, and family history.

2. Data Preprocessing:

Implementing data cleaning and preprocessing algorithms to handle missing values, outliers, and ensure data consistency.

Validating and standardizing input data to prepare it for analysis.

3. Feature Selection:

Identifying and selecting relevant features contributing to diabetes prediction.

Implement algorithms for automatic feature selection to enhance model efficiency.

4. Model Training:

Utilizing machine learning algorithms to train the predictive model.

Optimizing the model for accuracy using a subset of the dataset.

5. Prediction Output:

Generating accurate predictions for individuals at risk of developing diabetes.

Providing clear and interpretable results for healthcare practitioners.

6. Integration with Healthcare Systems:

Ensuring seamless integration with existing healthcare information systems.

Allowing for easy interoperability with electronic health records.

4.2 Non-Functional requirements

Non-functional requirements specify criteria that do not directly relate to the system's functionalities but are crucial for its overall effectiveness and performance. For this project, the non-functional requirements include:

1. Accuracy and Reliability:

The predictive model has a high level of accuracy in identifying individuals at risk of diabetes.

2. Scalability:

The system has a scalable to handle an increasing volume of health data as the dataset grows over time.

3. Interpretability:

Ensuring that the machine learning model's predictions are interpretable by healthcare practitioners.

Providing explanations for the factors influencing the model's predictions to build trust among users.

4. Security and Privacy:

Implements robust security measures to protect sensitive health data from unauthorized access.

Adheres to privacy regulations and standards, ensuring compliance with data protection requirements.

5.User Interface Design:

Designed an intuitive and user-friendly interface for healthcare practitioners to input and retrieve patient information easily.

Ensure that the interface facilitates effective communication of the model's predictions.

6. Performance:

The system processes data efficiently and delivers predictions in a timely manner.

Minimizes latency to support real-time or near-real-time decision-making.

7. Training:

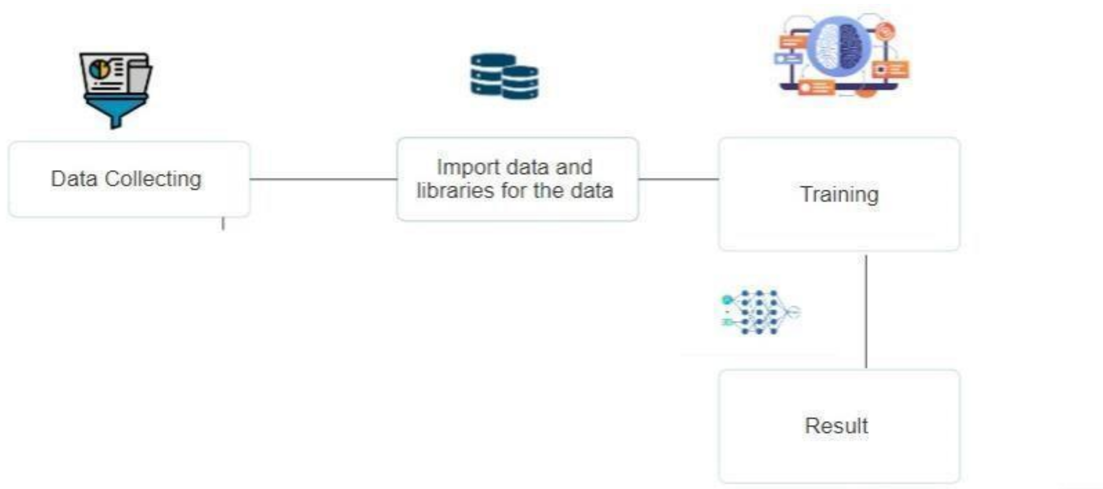
Provides training materials and support for healthcare practitioners to effectively use the system.

By defining these functional and non-functional requirements, the project team establishes a clear framework for developing a robust and effective Diabetes Prediction system using machine learning. These requirements guide the design, implementation, and evaluation of the system throughout the development lifecycle.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

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.



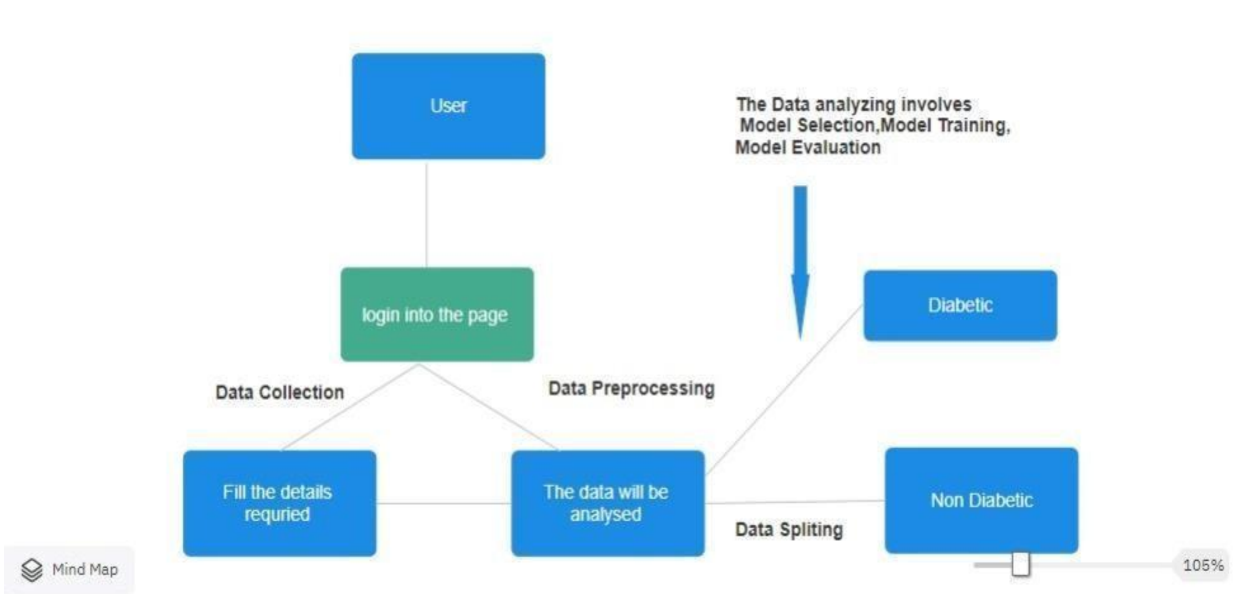
1.User figure out the symptoms available in the application

2.Gives the details of family history, food habits and common symptoms

3.The data is sent for processing to evaluation.

4.Finally the result is extracted and passed to application.

5.The Finalized data is visualized in the UI using the result.



User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Healthcare Provider	Data Input and Risk Assessment	USN-1	As a user, I can take Input from Patient Data for Diabetes Risk Assessment	~The system should provide a user-friendly interface for healthcare providers to input patient data, including age, gender, family history, BMI, blood pressure, glucose levels, and cholesterol. ~Upon data input, the system should accurately calculate and display the patient's risk of developing diabetes within a specified time frame. ~The system should make treatment recommendations based on the risk assessment.	High	Sprint-1
Patient	Self-Monitoring and Risk Assessment	USN-2	As a user, I will have Self-Monitoring and take Diabetes Risk Assessment	~Patients should be able to input their medical history and current health data into the system. ~The system should provide a diabetes risk assessment score based on the patient's input. ~Patients should receive personalized recommendations for lifestyle changes, diet, and exercise to reduce their diabetes risk.	Medium	Sprint-1
Data Scientist	Data Access and Model Training	USN-3	As a user, I can Access and Use Historical Data for Model Training	~Data scientists should have access to a diverse dataset with patient profiles and historical diabetes diagnosis records. ~The system should allow data scientists to extract, preprocess, and use this data for training and	High	Sprint-1

				validating machine learning models. ~The system should provide APIs or tools for data scientists to interact with the dataset and model training.		
Researcher	Result Analysis and Research	USN-4	As a user, I can Access and Analyze Prediction Results	~Researchers should have access to prediction results for further study and analysis. ~The system should provide relevant statistical information and visualization tools to help researchers understand factors influencing diabetes risk. ~The system should allow researchers to download prediction data for their studies.	High	Sprint-1
All Users (General)	Login	USN-5	As a user, I can log into the application by entering email & password	~Users who forget their passwords should be able to reset them securely through a password recovery process. ~The system should track login attempts and provide security measures against brute-force attacks.	Medium	Sprint-1
All Users (General)	Dashboard	USN-6	Create User Dashboards	~After successful login, each user should have access to a personalized dashboard. ~The dashboard should display relevant information based on the user's role and access rights, providing an overview of key features and data. Healthcare providers should see patient-related information and risk assessment tools. ~Data scientists should have access to data extraction and model training tools.	High	Sprint-1

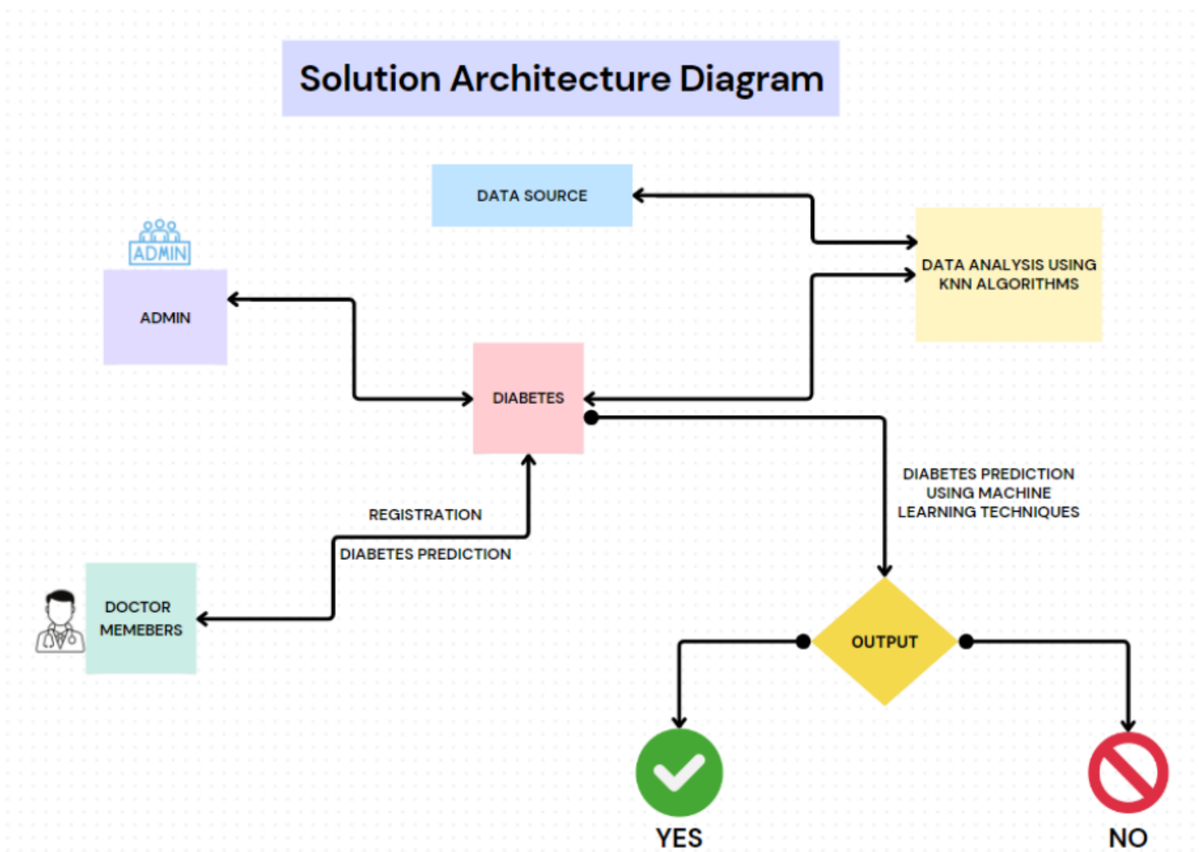
Customer (Web user)	User Registration and Profile	USN-7	Register and Create User Profile	~Customers should be able to register on the platform by providing necessary information, including name, email, and password. ~After registration, customers should have the ability to create and edit their user profiles, which may include personal information, medical history, and preferences ~Customers should receive a verification email to confirm their registration.	High	Sprint-1
Customer Care Executive	Customer Management	USN-8	Manage Customer Profiles	~Customer care executives should be able to search for and view customer profiles. ~Executives should have the ability to update customer information and medical data as provided by the customers. ~Customer care executives should be able to assist customers in using the system, including registration, risk assessment, and profile management.	High	Sprint-1
Administrator	Security and Compliance	USN-9	As a user, I can Ensure Data Security and Compliance	The system should implement robust security measures to protect patient data, including encryption, access controls, and regular security audits. The system should comply with data privacy regulations such as GDPR, ensuring that patient data is handled in a legally compliant manner.	High	Sprint-1

5.2 Solution Architecture

Our diabetes detection project leverages cutting-edge AI and machine learning techniques to revolutionize the way diabetes is diagnosed and managed. Our architecture combines data science, healthcare expertise, and technology to create a comprehensive solution that benefits patients, healthcare providers, and the environment.

Designing a solution architecture for a diabetes detection project using machine learning involves multiple components and considerations. Below is a high-level architecture for such a project:
Data collection, cleaning, and preprocessing.
Feature engineering and selection.
Model selection and training (e.g., logistic regression, decision trees, or deep learning).
Hyperparameter tuning for optimal performance.
Model evaluation using appropriate metrics.
Model deployment (e.g., API, containers, or cloud-based solutions).
Continuous monitoring, maintenance, and retraining.
User-friendly interface development.
Security and compliance with healthcare regulations.
Documentation and integration with healthcare systems.
Scalability and performance optimization.
Feedback loop for continuous improvement.

Solution Architecture Diagram:



6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture

S.No	Component	Description	Technology
1.	User Interface	Users can access the application through a web browser on their desktop or laptop.	HTML, CSS
2.	Application Logic-1	User Input Handling: When a user accesses the web application, display the homepage. Data Validation and Preprocessing: Preprocess the input data, such as normalizing and scaling numeric values, handling categorical data, and ensuring completeness	Python
3.	Application Logic-2	Machine Learning Model Integration: Load the trained machine learning model into the web application. Results Display: Display the prediction result to the user along with relevant information about the risk level.	HTML,CSS
4.	Application Logic-3	Built a web application using flask	Flask
5.	Database	A simple website is made using web development and any servers can be added based on requirement	MySQL or SQLite
6.	Cloud Database	We have the flexibility to choose for any cloud Databases	Amazon RDS (Relational Database Service etc
7.	File Storage	Dataset storage, Model Storage	Local Filesystem
8.	External API-1	Health data APIs act as bridges between applications and diverse health-related information sources, fostering the development of applications that support users in monitoring and improving their health and well-being.	Health Data API
9.	External API-2	Medical Data API provides access to relevant research findings, clinical trial data, or the latest information on diabetes treatments and medications. This additional data could enhance the project's accuracy and relevance, contributing to more informed predictions and recommendations.	Medical Data API
10.	Machine Learning Model	It uses machine learning models (K Nearest Neighbors, Support Vector Machine, and Naive Bayes) to analyze patterns in the data and make predictions. The goal is to identify people at high risk of diabetes early, allowing for timely intervention and prevention. The project contributes to healthcare by enhancing the early detection of diabetes, leading to better health outcomes for individuals and communities.	K Nearest Neighbors (KNN),Support Vector Machine (SVM) and Naive Bayes
11.	Infrastructure (Server)	Application Deployment on Local System Local Server Configuration: http://127.0.0.1:5000/	Local

6.2 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Form filling	USN-1	As a user, I can fill the form with details like age, insulin levels, BMI Index for the prediction.	2	High	Sadhia
Sprint-1		USN-2	As a user, I can give all the information for analysis of diabetes.	1	High	Purnima
Sprint-2		USN-3	As a user, I will click the predict button, so that the system proceeds.	2	Low	Meghana
Sprint-1		USN-4	As a user, I can get the results with in minutes.	2	Medium	Lalitha
Sprint-1	Results	USN-5	As a user, I will know whether I am having chances of diabetes or not.	1	High	Purnima

6.3 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	4 Days	31 Oct 2023	03 Nov 2023	20	19 Nov 2023
Sprint-2	20	6 Days	04 Nov 2023	09 Nov 2023		
Sprint-3	20	4 Days	10 Nov 2023	13 Nov 2023		
Sprint-4	20	6 Days	14 Nov 2023	19 Nov 2023		

Velocity:
Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

In the provided code, the features used for predicting diabetes are limited to the following five input fields in the HTML form:

1. Glucose Level:

Input field for the glucose level of the individual.

2. Insulin:

Input field for the insulin level of the individual.

3. MI (Body Mass Index):

Input field for the Body Mass Index of the individual.

4. Diabetes PF (Diabetes Pedigree Function):

Input field for the Diabetes Pedigree Function, which provides a measure of diabetes heredity.

5. Age:

Input field for the age of the individual.

7.2 Feature 2

1. Pregnancies

The number of pregnancies the individual has had. This can be an important factor in understanding the individual's health history.

2. Blood Pressure:

Input field for the blood pressure of the individual. Blood pressure is a vital health metric and can contribute to diabetes risk assessment.

3. Skin Thickness:

Input field for the skin thickness of the individual. Skin thickness may be correlated with certain health conditions related to diabetes.

The selection of features depends on the dataset, the characteristics of the target population, and the specific goals of the predictive model. Adding more relevant features can potentially improve the accuracy and reliability of the diabetes prediction.

Source Code:

Ditect.html

```
<!DOCTYPE html>

<html>

  <head>

  <title>DITECT</title>

  <meta name="viewport" content="width=device-width, initial-scale=1">

  <link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.min.css">


  <style>

  body {

  font-family: Arial, Helvetica, sans-serif;
```

```
}


input[type=text], select {
width: 1180px;
padding: 12px 20px;
margin-left: 15px ;
margin-right: 15px ;
margin-top: 10px ;
margin-bottom: 10px ;
display: inline-block;
border: 1px solid #ccc;
border-radius: 4px;
box-sizing: border-box;
padding:15px;
}
.registerbtn {
background-color: #e48912;
color: white;
padding: 16px 20px;
margin: 8px 0;
border: none;
cursor: pointer;
width: 100px;
opacity: 0.9;
text-align:center;
margin-left:580px;
margin-top:30px;
}
.homepage{

background-image:url('https://thumbs.dreamstime.com/b/digital-illustration-dna-structure-abstract-medical-background-technology-145822099.jpg') ;
    height:550px;
    width:1220px;
    margin:20px;
```

```
}

.registerbtn:hover {
opacity:1;
}
p{
    color: rgb(58, 139, 11);
    font-family:'Lucida Sans', 'Lucida Sans Regular', 'Lucida Grande', 'Lucida Sans Unicode',
Geneva, Verdana, sans-serif;
}


input[type=submit]:hover {
background-color: #17cce4;
}
.center {
margin: auto;
width: 60%;
border: 3px solid #ff9101c5;
padding: 10px;
}
.heading{
    color:#ffffff;
    text-align:center;
    marging-top :30px;
    font-size:25px;
    font-weight:900;
}

</style>
</head>
<body>
<div class="homepage">
```

```
    <h2> <p style="text-align:center;" class="heading" >Diabetes Prediction</p></h2>
    <form action="/predict"method="post">
   <!-- <input type="text" name="Pregnancies" placeholder="Pregnancies" required="required" /> -->
    <input type="text" name="Glucose Level" placeholder="Glucose Level" required="required" />
    <!--  <input type="text" name="Blood Pressure" placeholder="Blood Pressure" required="required"
/>
    <input type="text" name="Skin Thickness" placeholder="Skin Thickness" required="required"
/>-->
    <input type="text" name="Insulin" placeholder="Insulin" required="required" />
    <input type="text" name="BMI" placeholder="BMI" required="required" />
    <input type="text" name="Diabetes PF" placeholder="Diabetes PF" required="required" />
    <input type="text" name="Age" placeholder="Age" required="required" />
    <button type="submit" class="registerbtn">Predict</button>

    </form>
    </div>
</div>
    </body>
</html>
```

Result.html

```
<!DOCTYPE html>
<html>
<head>
  <title>DITECT</title>
<meta name="viewport" content="width=device-width, initial-scale=1">
<link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.min.css">

<style>
body {
  font-family: Georgia, serif;
}
h2{
```

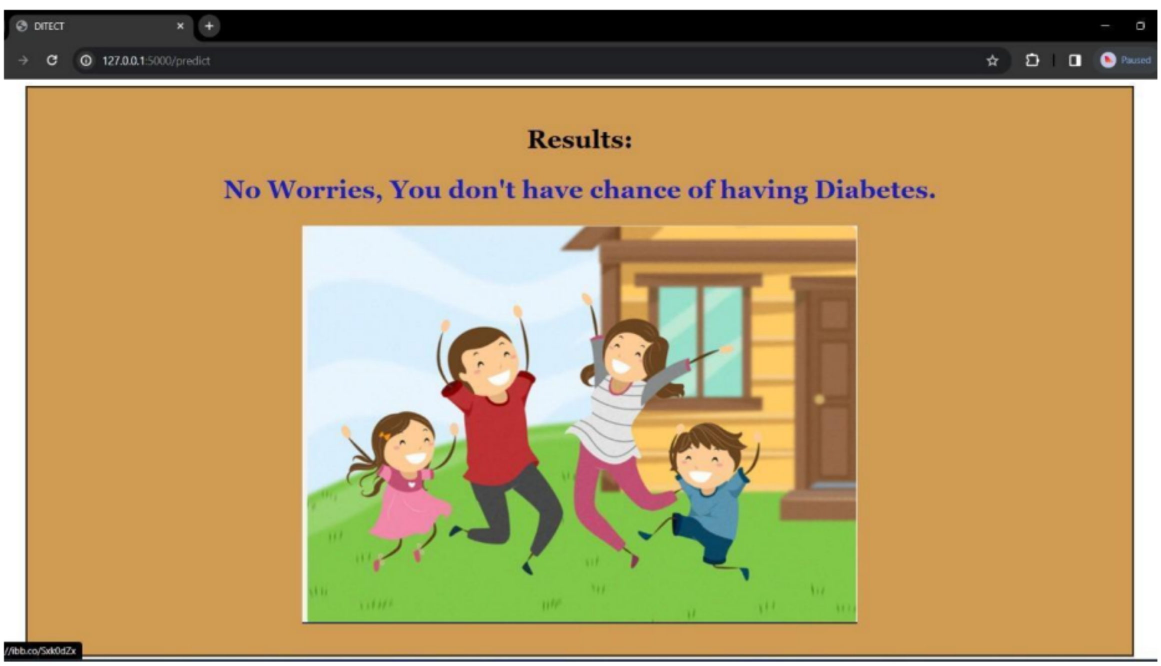
```
 color: rgb(0, 0, 0);
          font-family: Georgia, serif;
          align-content: center;
          font-size: 2rem;
}
.center {
 margin: auto;
 width: 60%;
 border: 3px solid #ccc;
 padding: 20px;
}

input[type=text], select {
 width: 100%;
 padding: 12px 20px;
 margin: 8px 0;
 display: inline-block;
 border: 1px solid #ccc;
 border-radius: 4px;
 box-sizing: border-box;
}

.registerbtn {
          font-family: "Open Sans", sans-serif,bold;
          font-size: 20px;
          letter-spacing: 3px;
          text-decoration: none;
          color: #000;
          cursor: pointer;
          border: 2px solid;
          padding: 0.75em 1em;
          position: relative;
          user-select: none;
          -webkit-user-select: none;
          touch-action: manipulation;
          width: 95%;
```

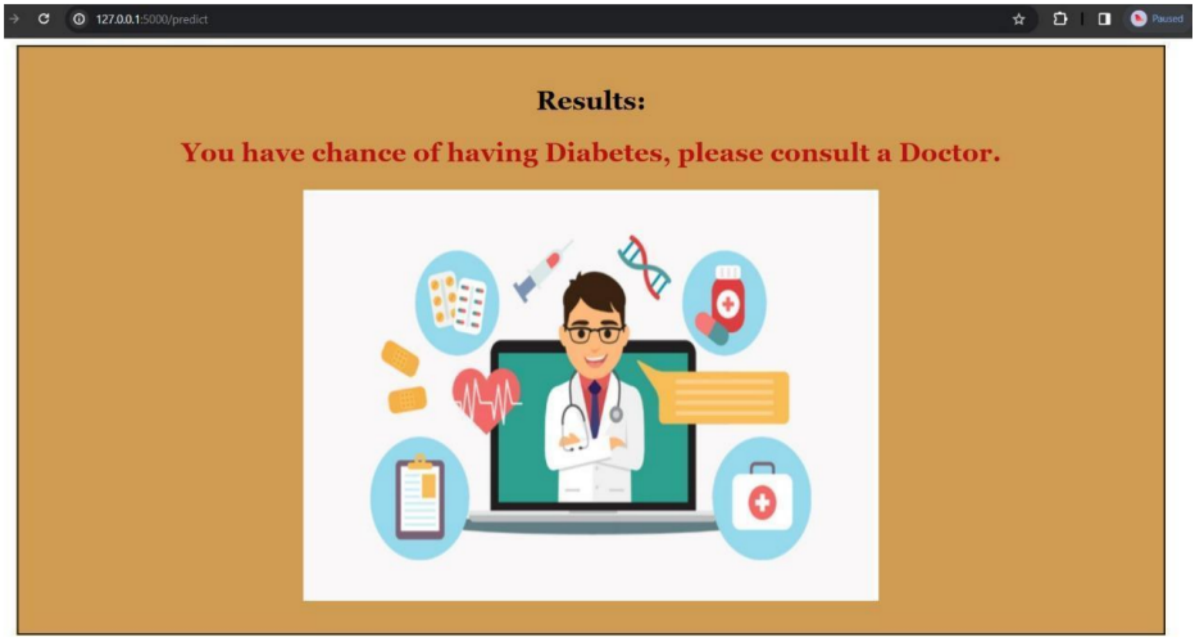
```
            box-sizing: border-box;
            margin-bottom: 15px;
            background-color: #254fc3
        }

        input[type=submit]:hover {
            background-color: #17cce4;
        }

        .center {
            margin: auto;
            width: 90%;
            border: 3px solid #000000c5;
            padding: 20px;
            text-align: center;
            background-color: rgb(208, 156, 83)
        }
</style>
</head>
<body>

<div class="center">
  <h2>Results: </h2>
  {% if prediction == 1%}
    <h2 style="color: rgb(189, 16, 16);">You have chance of having Diabetes, please consult a
Doctor.</h2>
    <p style="text-align:center;"> <a href="https://ibb.co/pyNgQJ6"><img
src="https://i.ibb.co/3zVwvB5/1.png" alt="HTML" width="700" height="500"></p>
  {% elif prediction == 0%}
    <h2 style="color: rgb(25, 31, 197)">No Worries, You don't have chance of having Diabetes.</h2>
    <p style="text-align: center;"></p> <a href="https://ibb.co/Sxk0dZx"><img
src="https://i.ibb.co/9vx2Vdv/2.png" alt="HTML" width="700" height="500"></p>
  {% endif %}
</div>

</body>
```

```
</html>
```

Output


Diabetes Prediction
Glucose Level
Insulin
BMI
Diabetes PF
Age
Predict

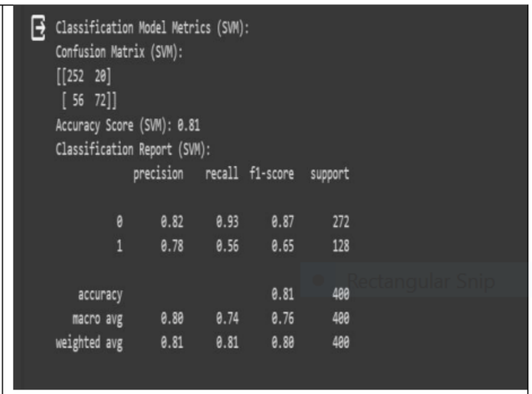
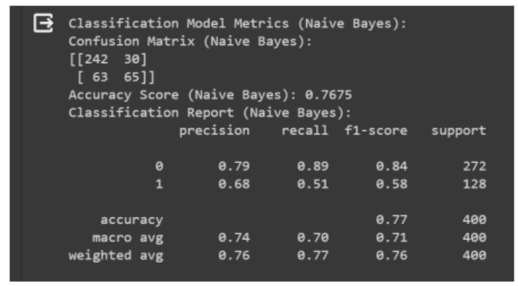

DITECT
127.0.0.1:5000/predict
Results:
No Worries, You don't have chance of having Diabetes.



8. PERFORMANCE TESTING

8.1 Performance Metrics

S.No.	Parameter	Values	Screenshot
1.	Metrics	**Regression Model:** MAE - 0.1725 , MSE - 0.1725 , RMSE - 0.4153311931459037, R2 score - 0.2072610294117646 **Classification Model:** Confusion Matrix - [[239 33] [36 92]], Accuray Score- & Classification Report - 0.8275 **precision recall f1-score support** **0 0.87 0.88 0.87 272** **1 0.74 0.72 0.73 128** **accuracy 0.83 400** **macro avg 0.80 0.80 0.80 400** **weighted avg 0.83 0.83 0.83 400**	

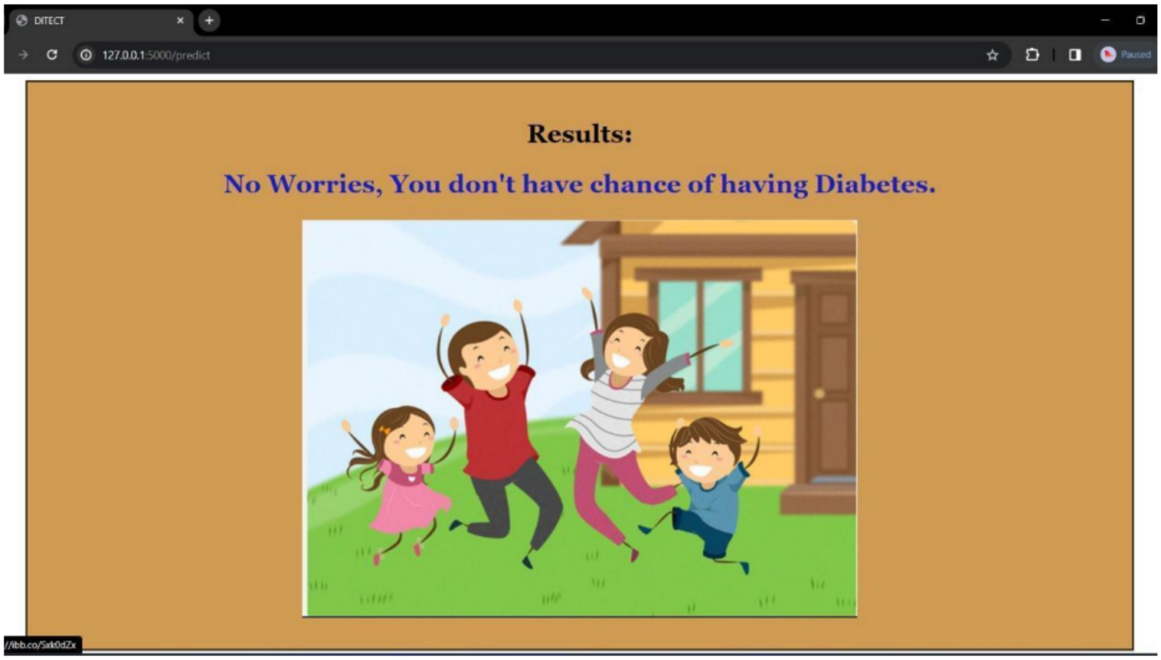
2.	Tune the Model	Hyperparameter Tuning - Validation Method -	 

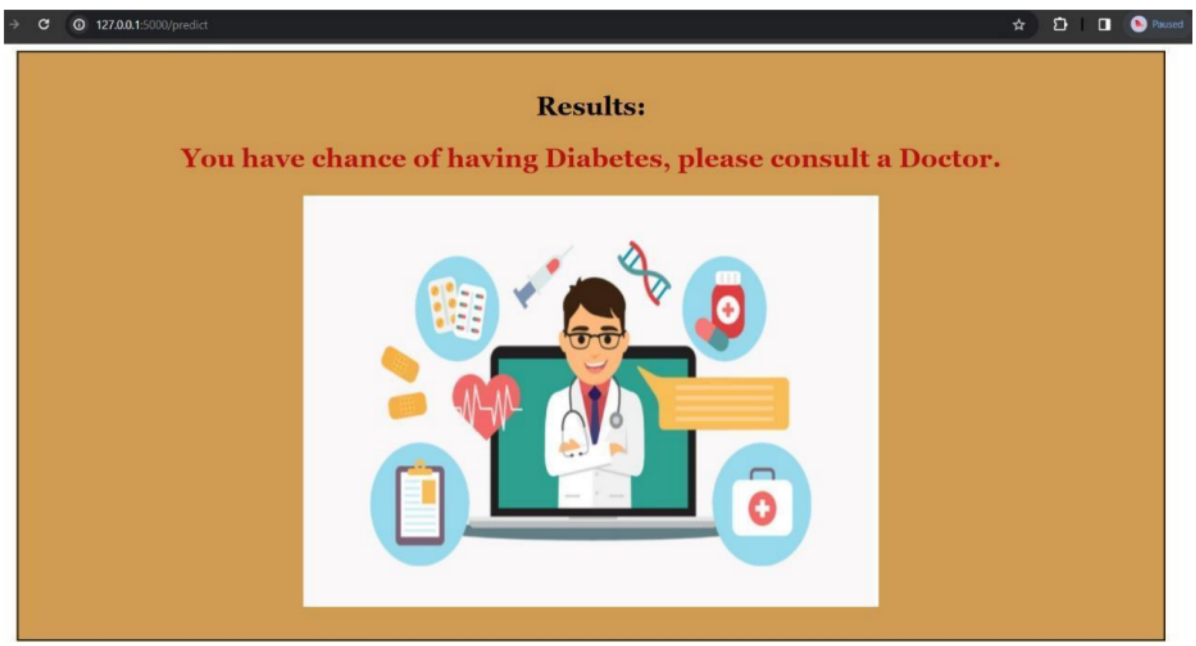
```
Classification Model Metrics (SVM):
Confusion Matrix (SVM):
[[252  20]
 [ 56  72]]
Accuracy Score (SVM): 0.81
Classification Report (SVM):
              precision    recall  f1-score   support

           0       0.82      0.93      0.87       272
           1       0.78      0.56      0.65       128

    accuracy                           0.81       400
   macro avg       0.80      0.74      0.76       400
weighted avg       0.81      0.81      0.80       400
```

```
Classification Model Metrics (Naive Bayes):
Confusion Matrix (Naive Bayes):
[[242  30]
 [ 63  65]]
Accuracy Score (Naive Bayes): 0.7675
Classification Report (Naive Bayes):
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       272
           1       0.68      0.51      0.58       128

    accuracy                           0.77       400
   macro avg       0.74      0.70      0.71       400
weighted avg       0.76      0.77      0.76       400
```

9. RESULTS

9.1 Output Screenshots





10. ADVANTAGES & DISADVANTAGES

Advantages :

- Early Detection:
  - Machine learning algorithms can analyze large datasets and identify patterns that might not be apparent to human observers. This can lead to the early detection of diabetes or prediabetes, allowing for timely intervention and management.
- Personalized Medicine:
  - ML models can take into account a wide range of variables, including genetics, lifestyle, and environmental factors, to create personalized predictions and treatment plans. This tailored approach can be more effective than one-size-fits-all solutions.
- Improved Accuracy:
  - Machine learning models can process vast amounts of data and learn complex relationships, potentially leading to more accurate predictions compared to traditional methods. This can help in identifying individuals at risk with higher precision.
- Cost-Effective:
  - Early detection and prevention of diabetes through machine learning can potentially reduce healthcare costs in the long run. By identifying high-risk individuals, resources can be directed towards targeted interventions, preventing complications and reducing the economic burden of diabetes.
- Continuous Monitoring:
  - Machine learning models can be used for continuous monitoring of individuals at risk. This real-time analysis can provide timely insights and alerts, enabling better disease management.

Disadvantages :

- Data Quality and Bias:
  - The accuracy of machine learning models heavily depends on the quality and representativeness of the data used for training. If the data is biased or incomplete, the model may not generalize well to diverse populations, leading to inaccurate predictions.
- Interpretability:
  - Many machine learning models, especially complex ones like deep neural networks, are often considered "black boxes" because their decision-making processes are not easily interpretable. This lack of transparency can be a barrier to gaining trust from healthcare professionals and patients.
- Privacy Concerns:
  - Predictive models often require access to sensitive health data. Protecting patient privacy while utilizing this information for training models is a significant challenge. Striking the right balance between data access and privacy is crucial.
- Overfitting:
  - Machine learning models may be prone to overfitting, where they perform well on the training data but fail to generalize to new, unseen data. Ensuring the model's robustness and generalizability is essential for its practical utility.
- Ethical Considerations:
  - There are ethical considerations regarding the use of predictive models in

healthcare. For example, the potential for discrimination based on predictions could lead to unequal access to resources or stigmatization of certain individuals.

11. CONCLUSION

In conclusion, the use of machine learning for diabetes prediction presents both promising opportunities and challenges in the field of healthcare. The potential for early detection, personalized medicine, improved accuracy, and cost-effective interventions highlights the positive aspects of integrating machine learning into diabetes management.Efforts to enhance the transparency and interpretability of machine learning models, ensuring data privacy and security, and addressing biases in training data are crucial for the successful implementation of these predictive tools. Additionally, ethical guidelines and regulatory frameworks must be established to govern the responsible development and deployment of machine learning applications in healthcare.Overall, while machine learning has the potential to significantly contribute to diabetes prediction and prevention, a thoughtful and balanced approach is essential to maximize its benefits while minimizing potential risks and ensuring equitable access to healthcare resources.

12. FUTURE SCOPE

The future scope of using machine learning for diabetes prediction and management is broad and holds the potential for transformative advancements in healthcare.

Advanced Predictive Models:

- Ongoing research may lead to the development of more sophisticated and accurate predictive models. Incorporating advanced machine learning techniques, such as deep learning and ensemble methods, could improve the ability to detect early signs of diabetes and enhance risk stratification.

Integration with Wearable Devices:

- The integration of machine learning models with wearable devices and continuous monitoring technology could enable real-time tracking of health parameters. This integration could provide timely feedback to individuals at risk and enhance personalized interventions.

Blockchain for Data Security:

- Given the sensitivity of health data, the integration of blockchain technology for secure and decentralized storage of patient information could address privacy concerns. Blockchain can provide a transparent and secure way to manage health records and ensure data integrity.

Patient Empowerment and Education:

- Future applications may emphasize not only prediction but also patient education and empowerment. Machine learning models could be used to develop user-friendly interfaces, providing individuals with actionable insights, lifestyle recommendations, and educational resources for diabetes prevention and management.

Global Health Impact:

- Machine learning models can be adapted for use in various global healthcare settings, particularly in regions with limited resources. The development of cost-effective and scalable solutions can contribute to the early detection and management of diabetes on a global scale.

13. APPENDIX

Source Code

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import pickle

# Extracting data
dataset = pd.read_csv('diabetes.csv')
dataset.head()

# Our dataset dimensions
dataset.shape

dataset.describe()

sns.countplot(x='Outcome', data=dataset)
```

```
dataset['Outcome'].value_counts()

dataset.groupby('Outcome').mean()

corr_mat = dataset.corr()
sns.heatmap(corr_mat, annot=True)

dataset.isna().sum()

# Feature matrix - Taking all our independent columns into a single array and dependent values into another array
x = dataset.iloc[:, :-1].values  # Independent matrix
y = dataset.iloc[:, -1].values

x.shape

y

#glucose for diabetic
fig = plt.figure(figsize =(16,6))

sns.distplot(dataset["Glucose"][dataset["Outcome"] == 1])
plt.xticks([i for i in range(0,201,15)],rotation = 45)
plt.ylabel("Glucose count")
plt.title("Glucose",fontsize = 20)

#insulin for diabetic

fig = plt.figure(figsize = (16,6))

sns.distplot(dataset["Insulin"][dataset["Outcome"]==1])
plt.xticks()
```

```
plt.title("Insulin",fontsize = 20)

#BMI for diabetic
fig = plt.figure(figsize =(16,6))

sns.distplot(dataset["BMI"][dataset["Outcome"]==1])
plt.xticks()
plt.title("BMI",fontsize = 20)

#diabeticpedigreefunction for diabetic
fig = plt.figure(figsize = (16,5))
sns.distplot(dataset["DiabetesPedigreeFunction"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("diabetespedigreefunction")

#Age for diabetic
fig = plt.figure(figsize = (16,6))

sns.distplot(dataset["Age"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("Age")

#Removing unnessary columns
x = dataset.drop(["Pregnancies","BloodPressure","SkinThickness","Outcome"],axis = 1)
y = dataset.iloc[:,-1]

# Splitting dataset into training set and test set
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

```
# K-Nearest Neighbors (KNN) Classifier with Hyperparameter Tuning
knn = KNeighborsClassifier()
knn_param_grid = {'n_neighbors': [5, 10, 15, 20, 25], 'metric': ['minkowski', 'euclidean',
'manhattan']}
knn_grid = GridSearchCV(knn, knn_param_grid, cv=5)
knn_grid.fit(x_train, y_train)
best_knn = knn_grid.best_estimator_

# Use the best model for predictions
knn_y_pred = best_knn.predict(x_test)

mae = mean_absolute_error(y_test, knn_y_pred)
mse = mean_squared_error(y_test, knn_y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, knn_y_pred)
print("Regression Model Metrics (KNN):")
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r2)

# Classification Model Evaluation Metrics for KNN
knn_cm = confusion_matrix(y_test, knn_y_pred)
print("Classification Model Metrics (KNN):")
print("Confusion Matrix (KNN):")
print(knn_cm)
accuracy_knn = accuracy_score(y_test, knn_y_pred)
print("Accuracy Score (KNN):", accuracy_knn)
classification_report_knn = classification_report(y_test, knn_y_pred)
print("Classification Report (KNN):")
print(classification_report_knn)
```

```
knn_cm = confusion_matrix(y_test, knn_y_pred)
sns.heatmap(knn_cm, annot=True)

# Support Vector Machine (SVM) Classifier with Hyperparameter Tuning
svc = SVC()
svc_param_grid = {'kernel': ['linear', 'poly', 'rbf', 'sigmoid']}
svc_grid = GridSearchCV(svc, svc_param_grid, cv=5)
svc_grid.fit(x_train, y_train)
best_svc = svc_grid.best_estimator_

# Use the best model for predictions
svc_y_pred = best_svc.predict(x_test)

svc_cm = confusion_matrix(y_test, svc_y_pred)
print("Classification Model Metrics (SVM):")
print("Confusion Matrix (SVM):")
print(svc_cm)
accuracy_svc = accuracy_score(y_test, svc_y_pred)
print("Accuracy Score (SVM):", accuracy_svc)
classification_report_svc = classification_report(y_test, svc_y_pred)
print("Classification Report (SVM):")
print(classification_report_svc)

# Naive Bayes (NB) Classifier
nb_classifier = GaussianNB()
nb_classifier.fit(x_train, y_train)
nb_y_pred = nb_classifier.predict(x_test)

# Classification Model Evaluation Metrics for Naive Bayes
nb_cm = confusion_matrix(y_test, nb_y_pred)
print("Classification Model Metrics (Naive Bayes):")
print("Confusion Matrix (Naive Bayes):")
```

```
print(nb_cm)
accuracy_nb = accuracy_score(y_test, nb_y_pred)
print("Accuracy Score (Naive Bayes):", accuracy_nb)
classification_report_nb = classification_report(y_test, nb_y_pred)
print("Classification Report (Naive Bayes):")
print(classification_report_nb)

# Save the best models and StandardScaler for future use
pickle.dump(best_svc, open('best_svc_classifier.pkl', 'wb'))
pickle.dump(best_knn, open('best_knn_classifier.pkl', 'wb'))
pickle.dump(sc, open('sc.pkl', 'wb'))
```

GitHub Link

https://github.com/sadhiasyed/Diabetes-Prediction-Using-Machine-Learning

Project Demo Link

https://www.youtube.com/watch?v=N9oOUm0XO8k