



DIABETES PREDICTION SYSTEM

PREPARED BY:

BIKESH SITIKHU

RAM KATWAL

SABIN SUWAL

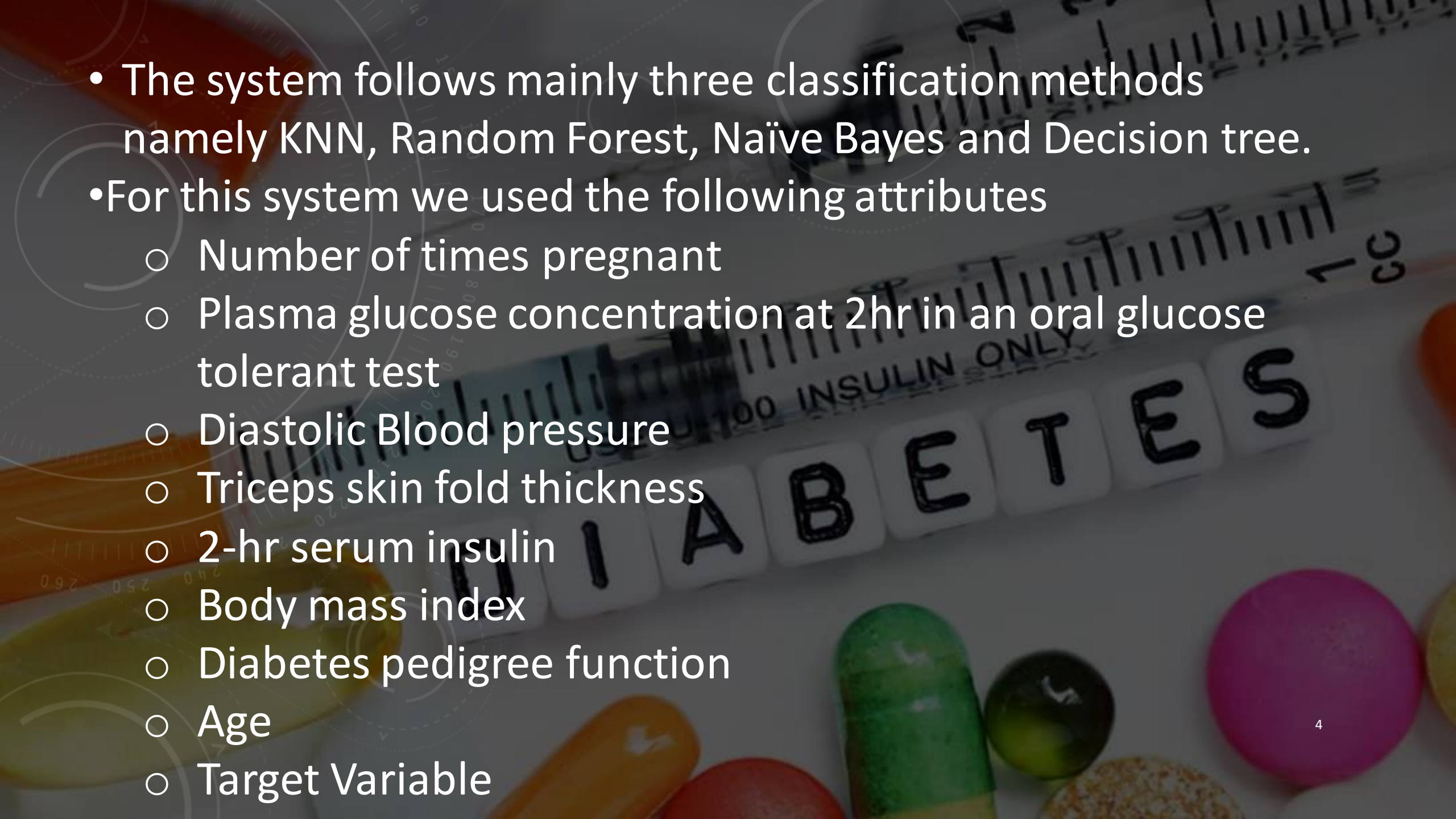
SACHIT KUMAR SHRESTHA₁

Contents

1. Introduction to Diabetes Mellitus
2. Objectives
3. Statement of Problems
4. Introduction to Machine Learning
5. Literature Review
6. Requirement Analysis
7. System Design and Architecture
8. Methodology
9. Project Implementation
10. Expected Outcomes
11. References

1. Introduction

- Diabetes Mellitus (DM) commonly known as diabetes is metabolic disorders characterized by high blood sugar level.
- About 422 million people worldwide have diabetes and the numbers are increasing.
- According to DRC early diagnosis of patients at risk can prevent 80 percent of complications.
- So we proposed a system Diabetes Prediction System that can be useful and helpful for doctors and practitioners

- 
- The system follows mainly three classification methods namely KNN, Random Forest, Naïve Bayes and Decision tree.
 - For this system we used the following attributes
 - Number of times pregnant
 - Plasma glucose concentration at 2hr in an oral glucose tolerant test
 - Diastolic Blood pressure
 - Triceps skin fold thickness
 - 2-hr serum insulin
 - Body mass index
 - Diabetes pedigree function
 - Age
 - Target Variable

2. Objective

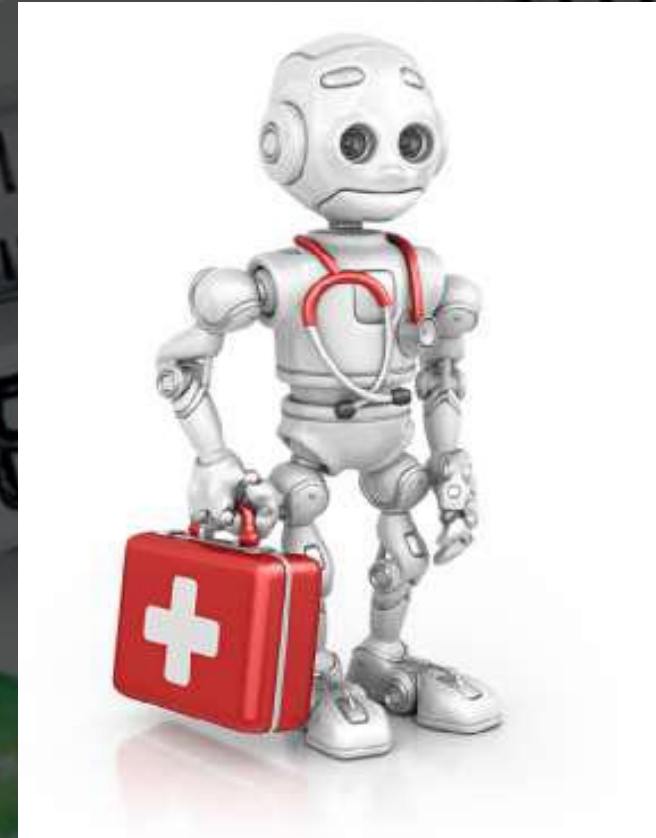
- To develop a system that can predict diabetes with a higher accuracy
- To evaluate the data set for the early prediction of the disease
- To provide an effective and easy system for all kinds of people
- To understand the concept of machine learning and data mining

3. Problem Statement

- So the normal identifying process is to visit a diagnostic center, consult their doctor and sit tight for a day or more to get their reports.
- Often this can bring many complications and more time and money.
- With the rise of Machine learning we have a solution to this issue.
- We are going to develop a system that has the ability to predict whether the patient has diabetes or not.

4. Introduction to Machine Learning

- In machine learning, computers apply statistical learning technique to automatically identify patterns in data.
- These techniques can be used to make highly accurate predictions.



Categories of Machine Learning

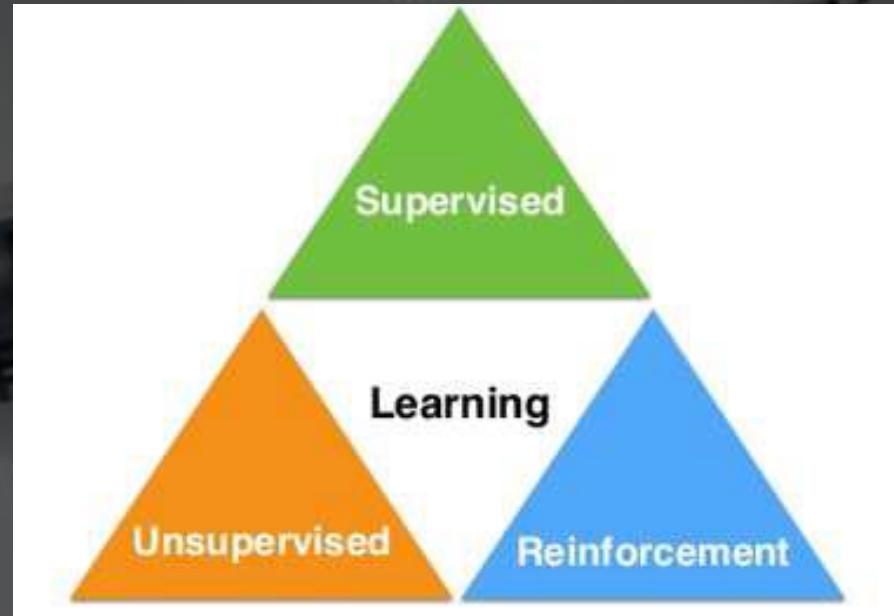
I. Supervised Learning

- i. Classification
- ii. Regression

II. Unsupervised Learning

- i. Association Rule Learning
- ii. Clustering

III. Reinforcement Learning



Types of Diabetes

Type 1 diabetes

10%

The body does not produce insulin

No

Type 2 diabetes

90%

The body produces insulin but is not used efficiently

CAN BE PREVENTED



HEALTHY & BALANCED EATING

APPROPRIATE BODY WEIGHT



MODERATE PHYSICAL EXERCISE

(E.G. WALKING 30 MINUTES A DAY)

HOW TO PREVENT?



YES

Gestational diabetes

High blood sugar levels in pregnancy

Usually, levels are stabilised after giving birth

-60% probability of suffering from type 2 diabetes



5. Literature Review

- Indian Researcher Sharmila et al in 2016 concentrate on gaining insight about the big data prediction of Indian diabetic dataset through Hadoop using K-means method.
- Another researcher Du Brava et al in 2017 conduct research with the objective to identify risk factor variables.
- In case of type II diabetes, it is correlated with diagnosis of neuropathy on electronic health records of diabetic patient using random forest modeling.
- Chinese Researcher Zheng et al (2018) proposed method using image from tongue color, texture and geometry features of diabetic patient.
- This research applied screening model using data mining techniques.

6. Requirement Analysis

6.1 Software Requirement

- Python
- Django
- PostgreSQL
- JavaScript
- Keras

Requirement Analysis contd...

6.2 Functional Requirement

- Transaction correction and cancellation
- Authorization to the user

6.3 Non-Functional Requirement

- Reliability
- Maintainability
- Performance
- Portability

7. System Design and Architecture

7.1 System Overview

- In this system we compared two machine learning task, supervised and unsupervised learning.
- Supervised learning is divided into classification and regression.
- The used supervised learning are Decision tree, Naive Bayes and Random Forest algorithms.
- Association and Clustering are used as unsupervised learning means to find hidden patterns of risk factor, being done by building separation of diabetes patient data into groups of data with similar characters.

System Design contd...

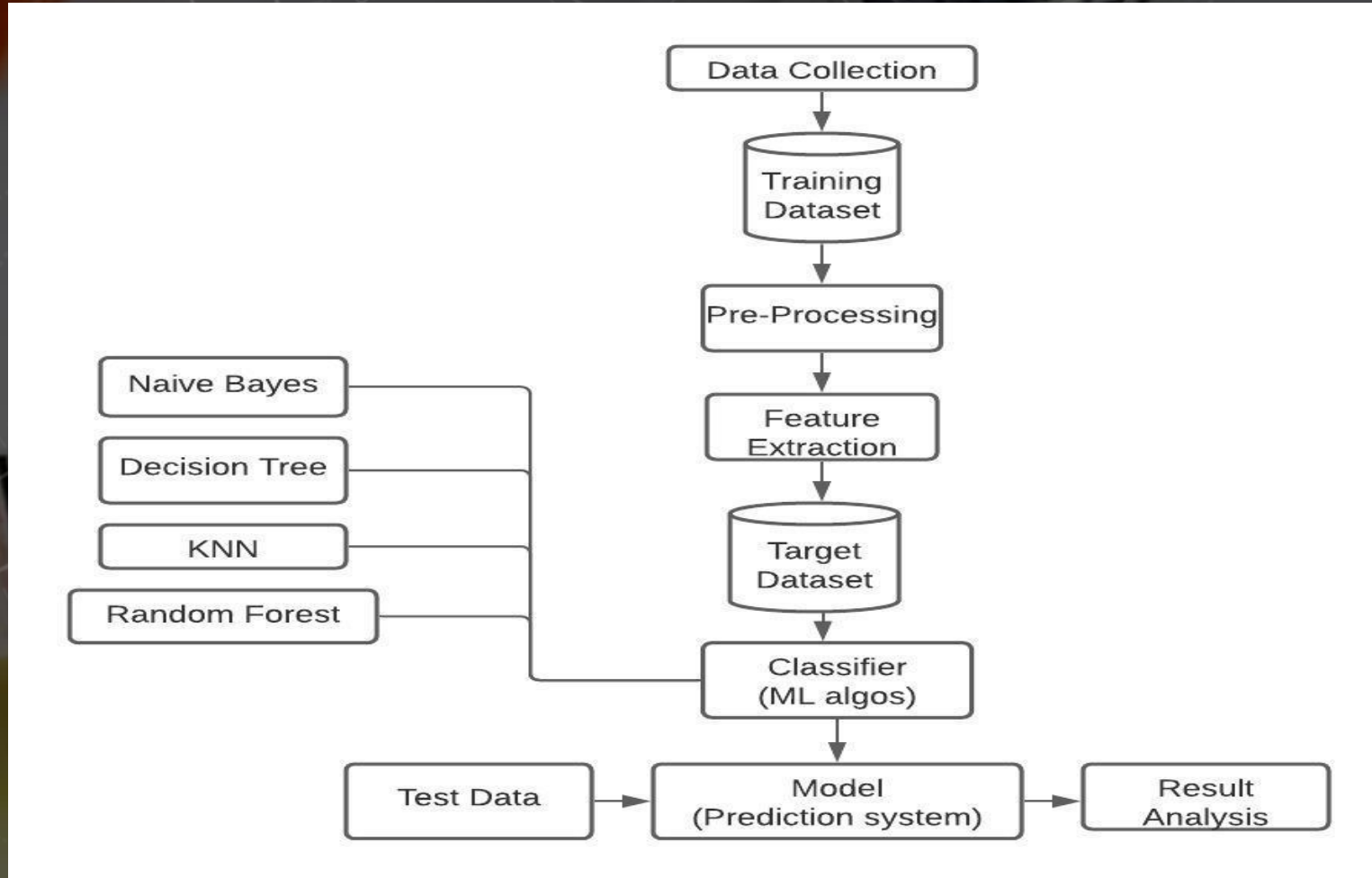


Figure: Block Diagram of Diabetes Prediction System

System Design contd...

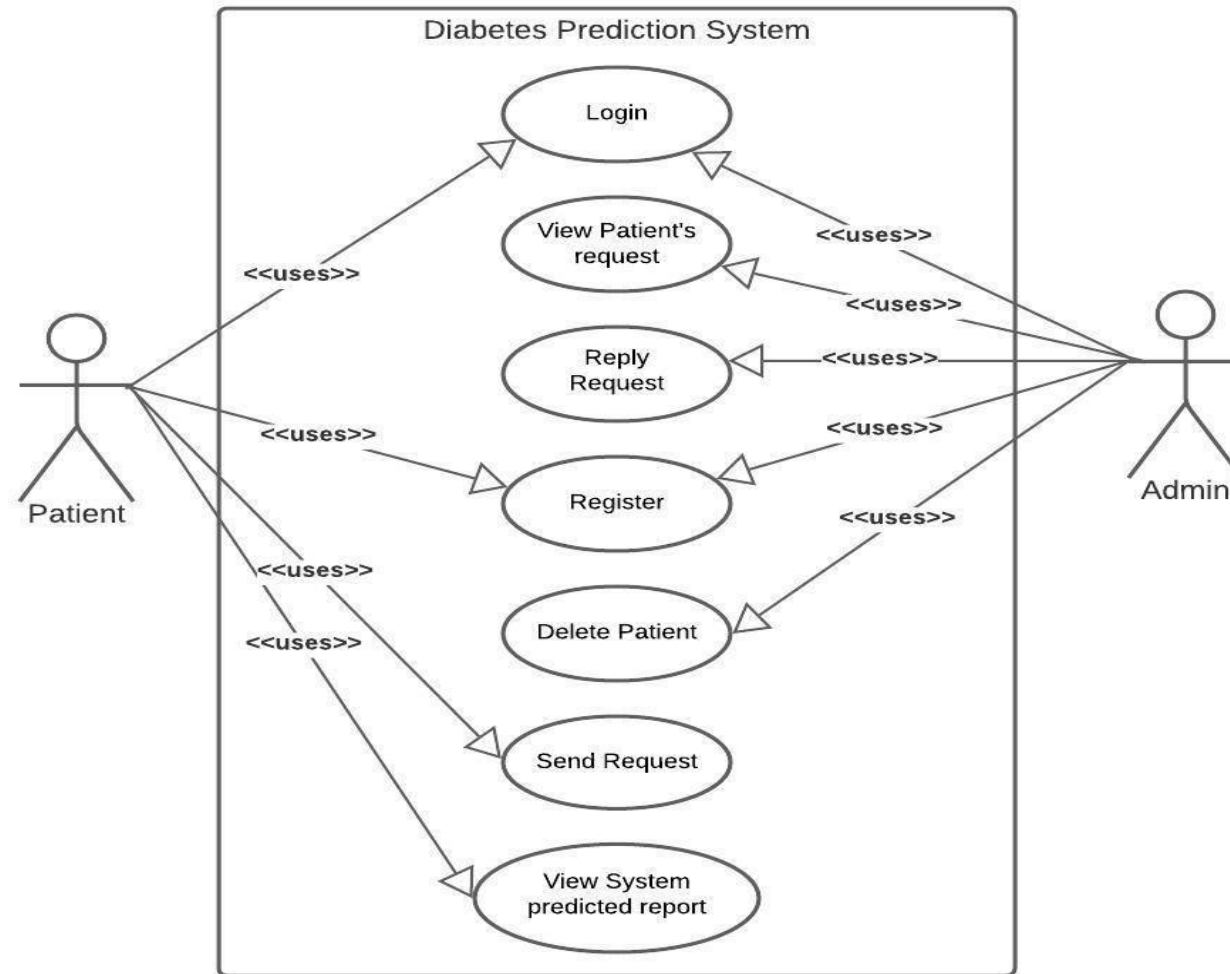
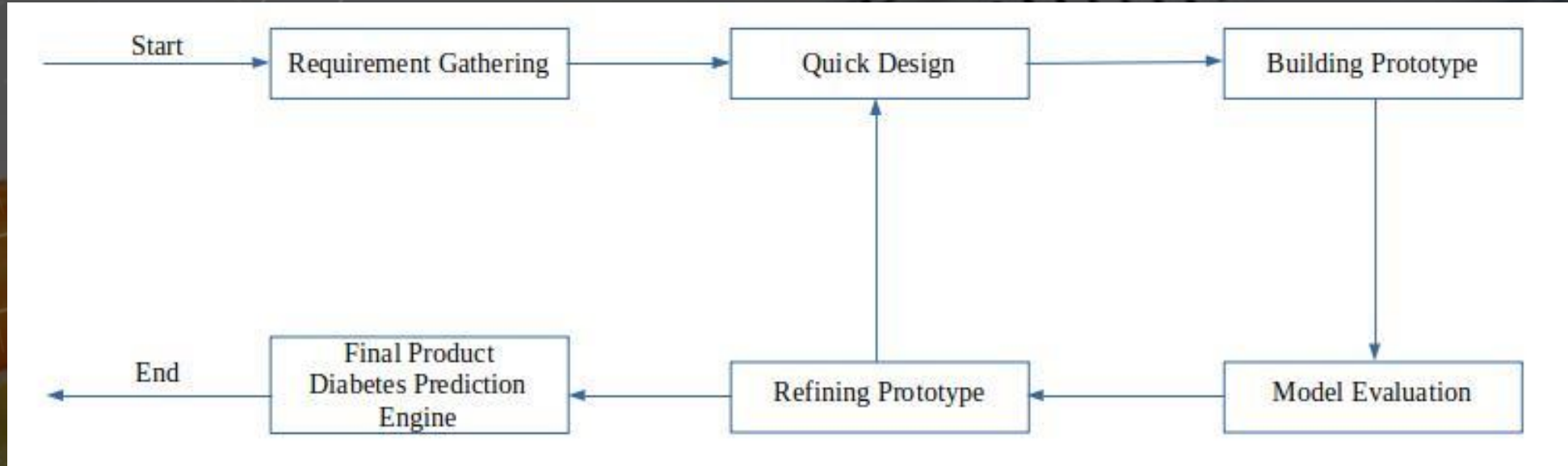


Figure: Use Case Diagram of Diabetes Prediction System

8. Methodology

8.1 Software Development Approach



8.2 Data Collection

- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- In particular, all patients here are females at least 21 years old of Pima Indian heritage.
- Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on .

8.3 Data Preparation

- Standard Scaler and label Encoder
- Label encoder
- X and Y
- Model performance
 1. Confusion matrix
 2. Metrics
 - Accuracy
 - Precision
 - Recall
- Scores Table

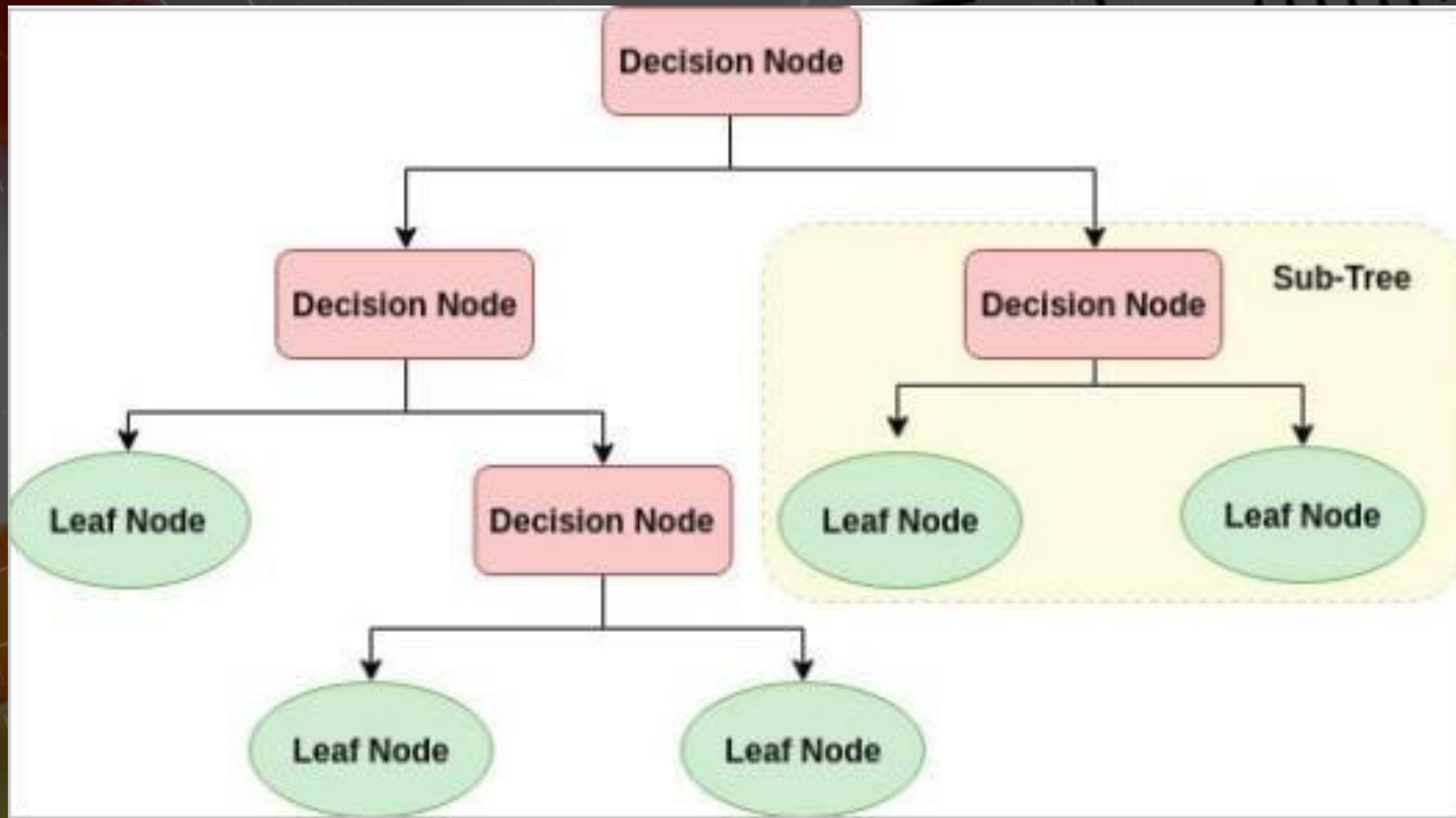
8.4 Training Data

- After data has been prepared the data is divided into two parts i.e. Train and test.
- Train data are used to create model during training process.



8.5 Decision Tree

- A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- Entropy Calculation
 - $\text{Entropy} = -(P/(P+N)) * \log(P/(P+N)) - (N/(P+N)) * \log(N/(P+N))$ where P = positive and N = negative
- Calculate average information:
 - $I(\text{attribute}) = \sum (P_i + N_i) / (P + N) \text{ entropy}(A)$
- Calculate information gain:
 - $\text{gain} = \text{entropy}(s) - I(\text{attribute})$

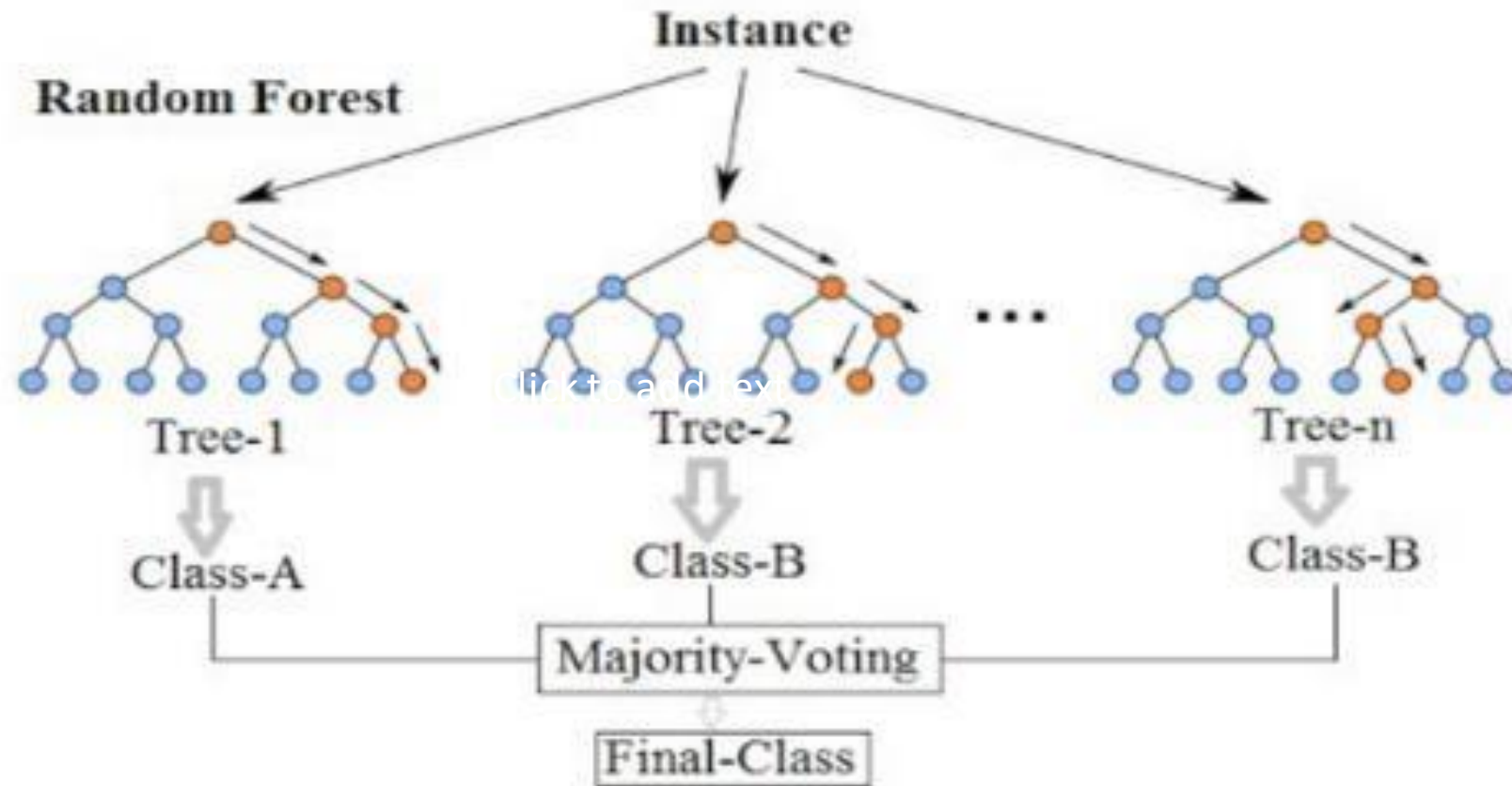


Decision Tree

8.6 Random Forest

- The random forest is a classification algorithm consisting of many decisions trees.
- Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.
- It also provides a pretty good indicator of the feature importance.

Random Forest Simplified



8.7 Naive Bayes

- Naive Bayes is a machine learning classifier which employs the Bayes Theorem.
- Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose.

$$P(C|X) = \frac{(P(X_1|C) * P(X_2|C) * P(X_3|C) * P(X_4|C) * P(C))}{(P(X_1) * P(X_2) * P(X_3) * P(X_4) * W)}$$

Finally, the value is predict by comparing value of probability of yes or no.

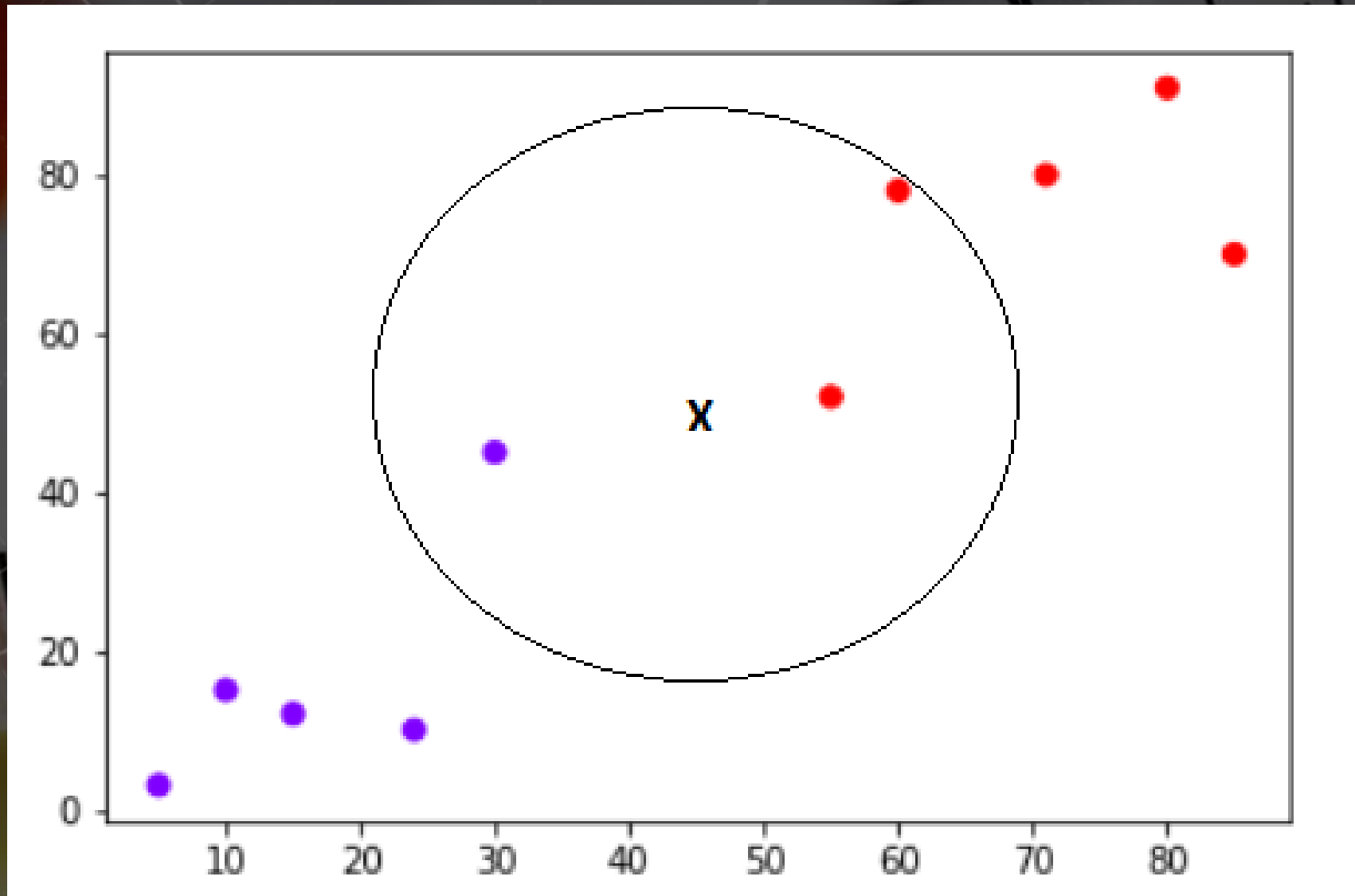
$$A = \frac{p(\text{yes}/X)}{p(\text{yes}/X) + p(\text{no}/X)}$$

$$B = \frac{p(\text{no}/X)}{p(\text{yes}/X) + p(\text{no}/X)}$$

If A is greater than B then ,the patiet

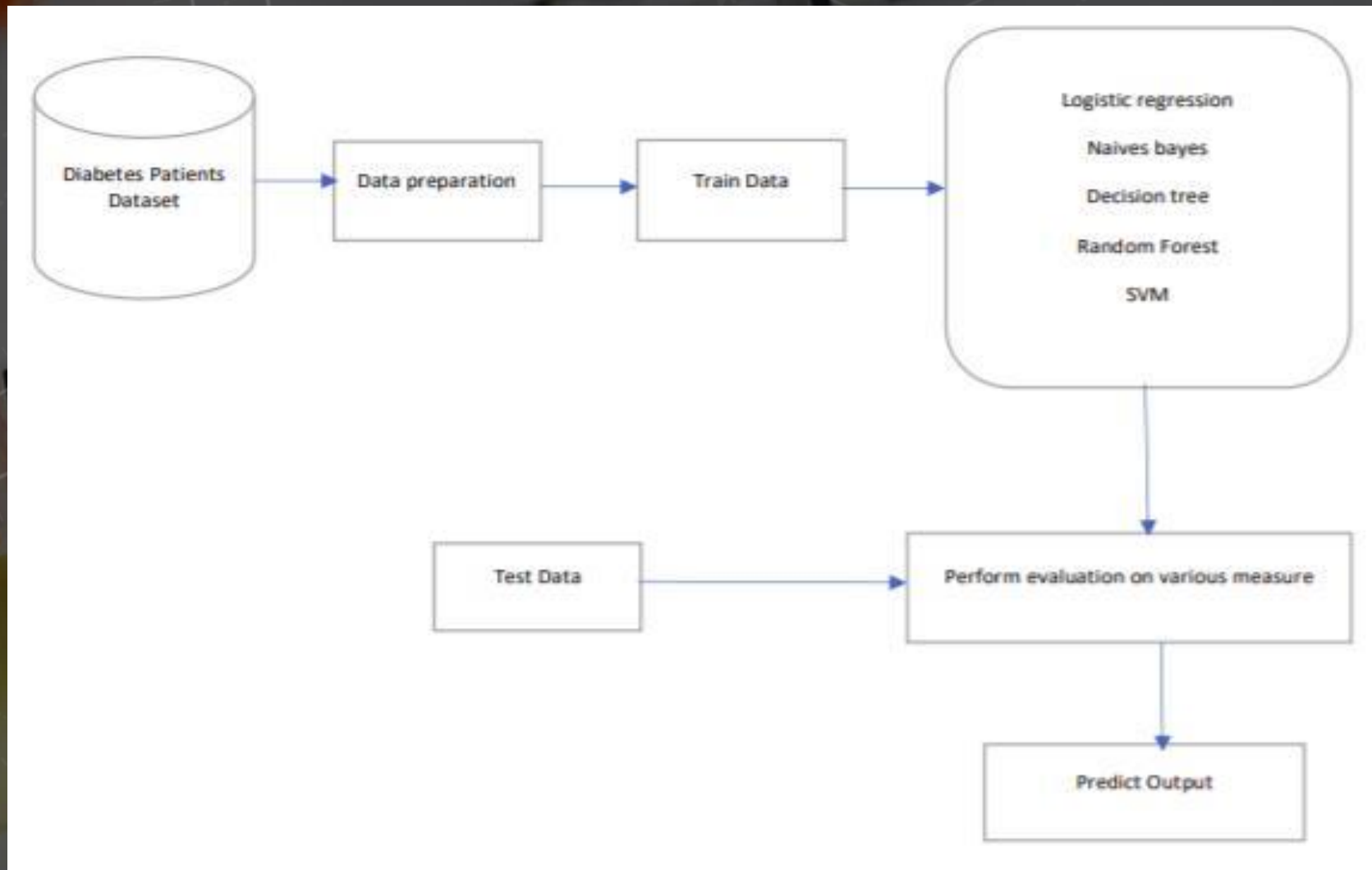
8.8 K-Nearest Neighbours(KNN) Algorithm

- It is one of the simplest supervised machine learning algorithm mostly used for classification.
- K in KNN is a parameter that refers to the number of nearest neighbors to include in the majority voting process i.e. diabetic or non-diabetic one.
- $D1 = (Pg - P1) + (Bg - B1) + (BMg - BM1) + \dots \dots \dots \text{outcome} = 1$
- $D2 = (Pg - P2) + (Bg - B2) + (BMg - BM2) + \dots \dots \dots \text{outcome} = 0$
- $\dots \dots \dots$
- Then we order the distance in ascending order and find it first K nearest neighborhood. the value is predicted on the basis of majority of Outcomes of neighborhood.



K-Nearest Neighbours(KNN)

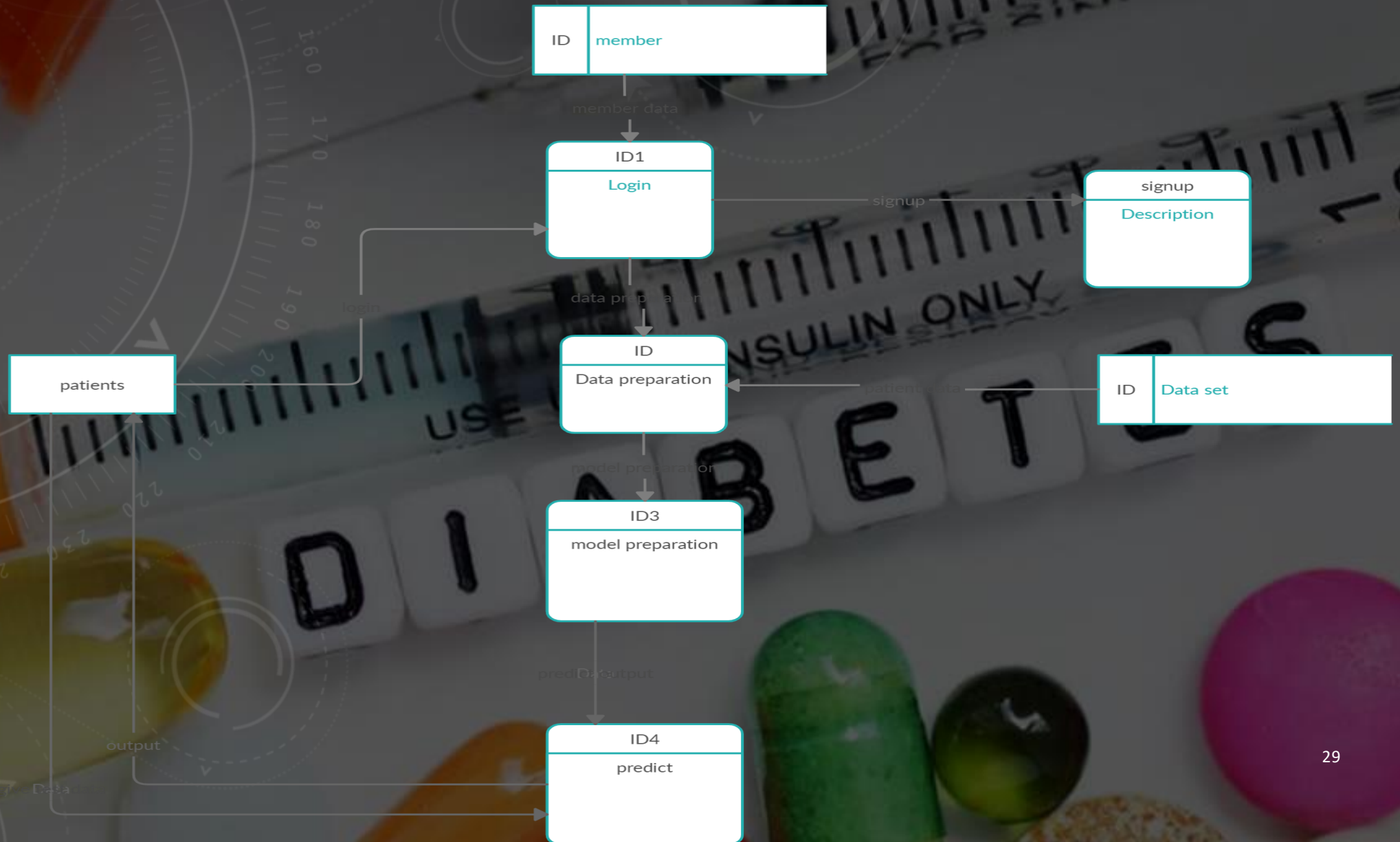
8.9 System Block Diagram



8.9.1 Tools Used

- GitHub
- Kanban Board
- Lucid Chart

8.9.2 Data Flow Diagram

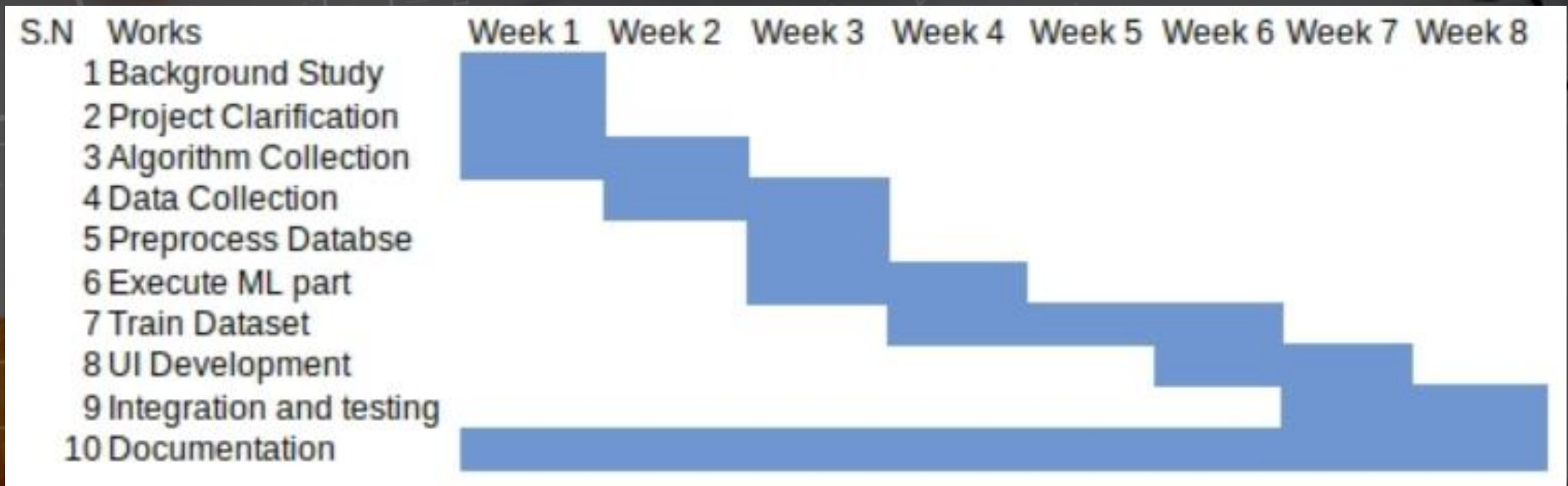


9. Project Implementation

The task needed to be done in order to complete this Diabetes Prediction System will be held with the equal participation of all the team members.

Project Name	Project Duration
Diabetes Prediction System	8 weeks

9.1 Gantt Chart



10. Expected Outcomes

- Proposed four machine learning algorithms
 - Naive Bayes
 - Decision Tree
 - Random Forest
 - KNN
- With the accuracy of 76.22%, 77.22%, 75.57%, 75.97% respectively

11. References

- [1] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. provisional report of a who consultation," Diabetic medicine, vol. 15, no. 7, pp. 539–553, 1998.
- [2] H. Temurtas, N. Yumuşak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks," Expert Syst. Appl., vol. 36, pp. 8610–8615, 05 2009.
- [3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia computer science, vol. 132, pp. 1578–1585, 2018.
- [4] J. Sun, "The study of pima indian diabetes," 10 2016.



DIABETES

THANK YOU!