NAAN MUDHALVAN

ARTIFICIAL INTELLIGENCE

TITLE:

BASED DIABETES PREDICTION SYSTEM

Aanesht Raj G

REG NO:712221121001

PHASE 1

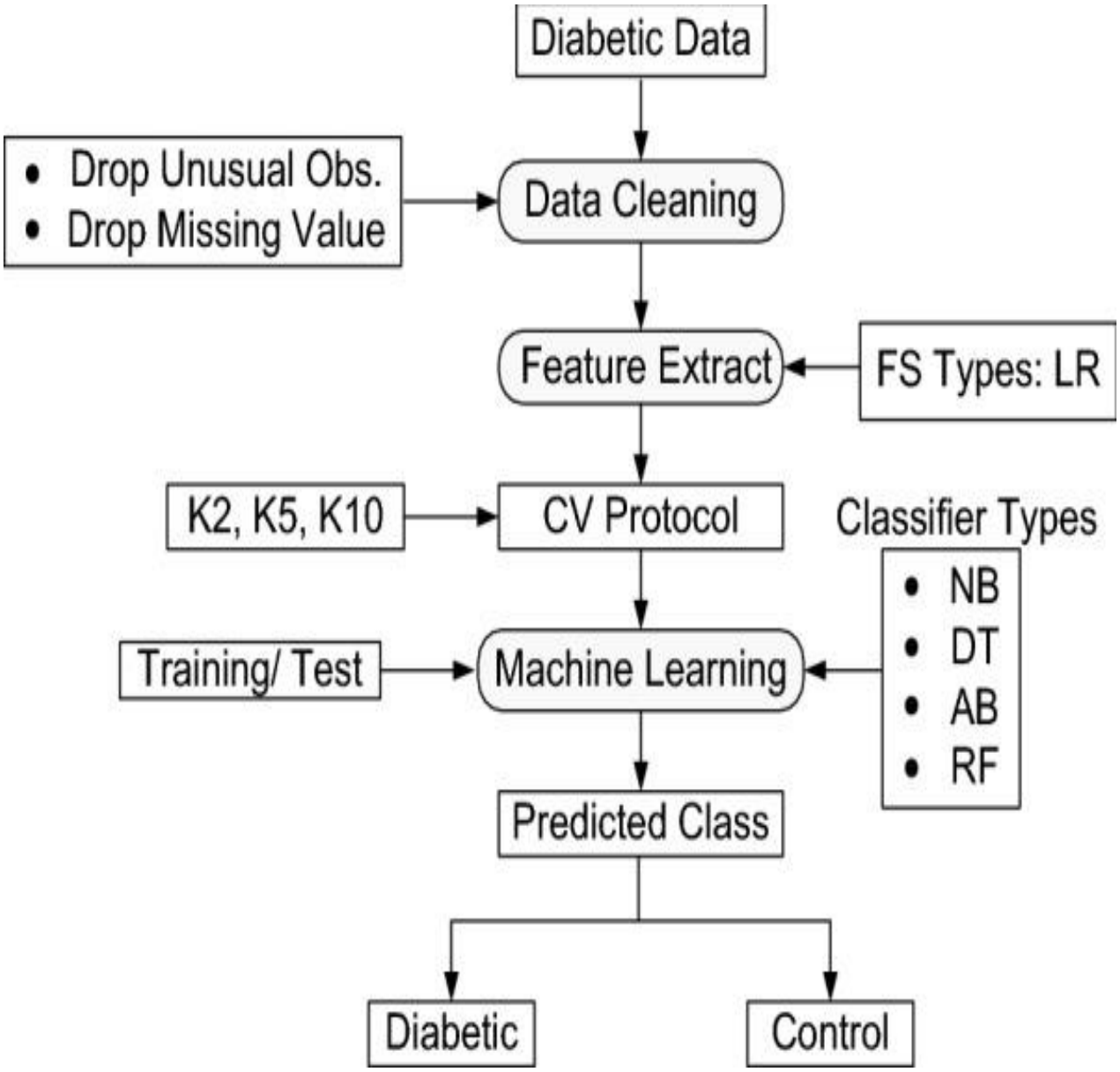
PROBLEM DEFINITION AND DESIGN THINKING

PROBLEM DEFINITION

Diabetes is a chronic disease characterized by high blood sugar. It may cause many complicated disease like stroke, kidney failure, heart attack, etc. About 422 million people were affected by diabetes disease in worldwide in 2014. The figure will be reached 642 million in 2040. The main objective of this study is to develop a machine learning (ML)-based system for predicting diabetic patients.

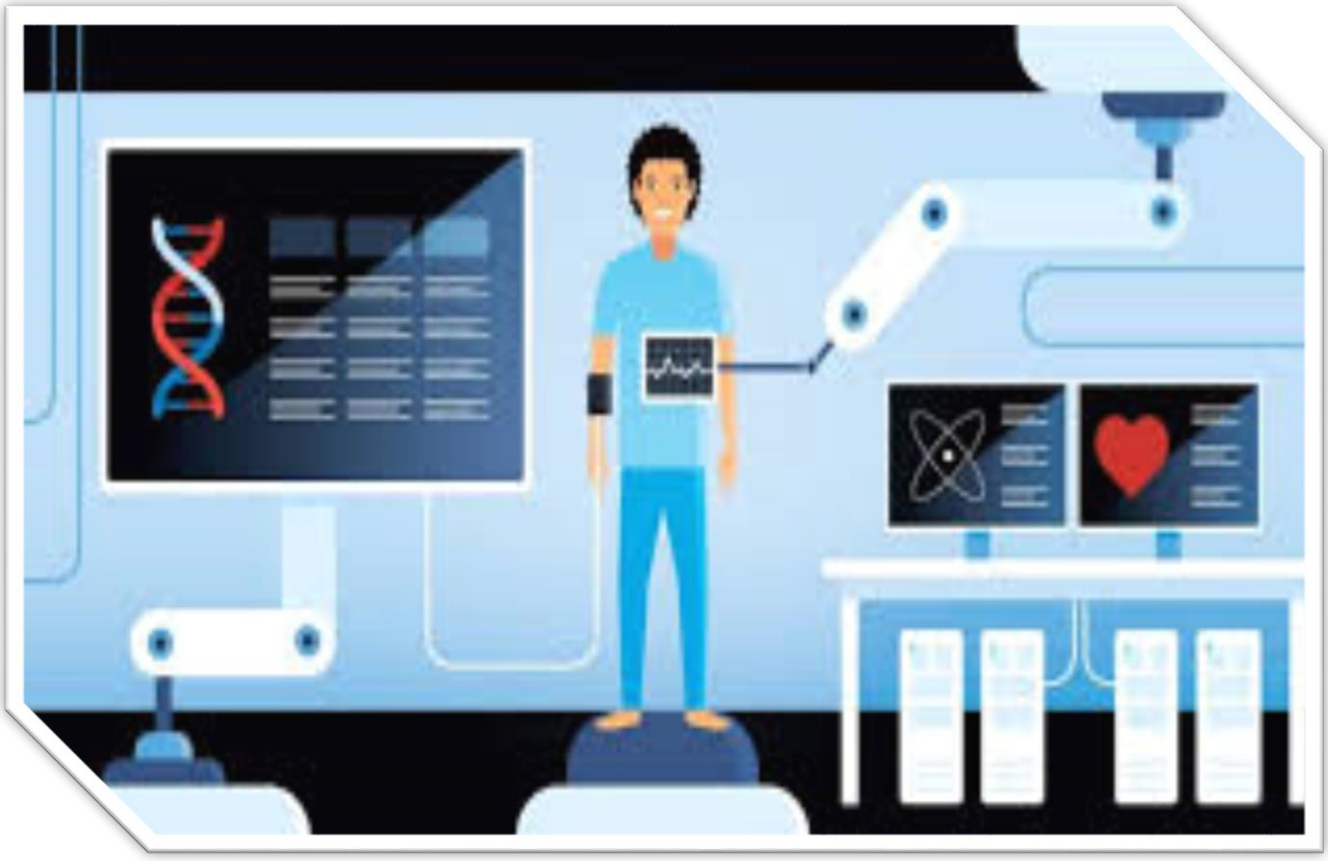
DESIGN THINKING

The combination of LR and RF-based classifier performs better. This combination will be very helpful for predicting diabetic patients.



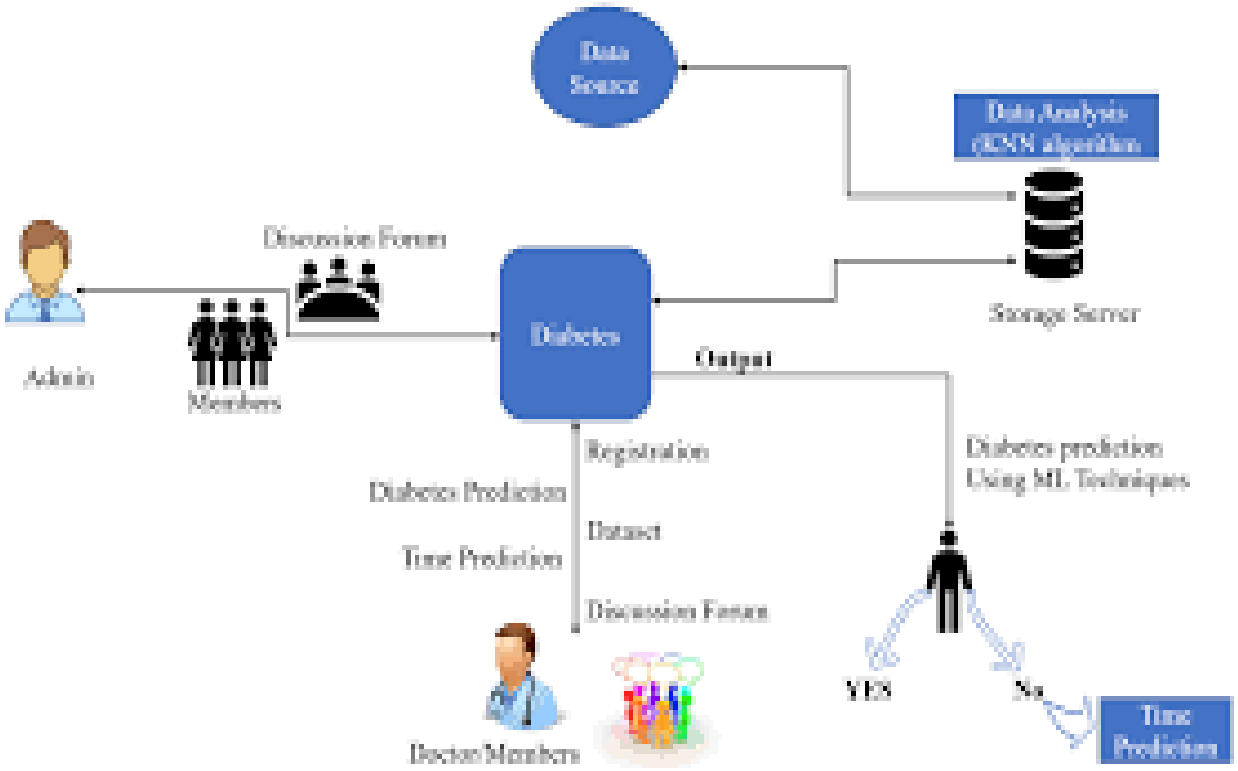
Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly.



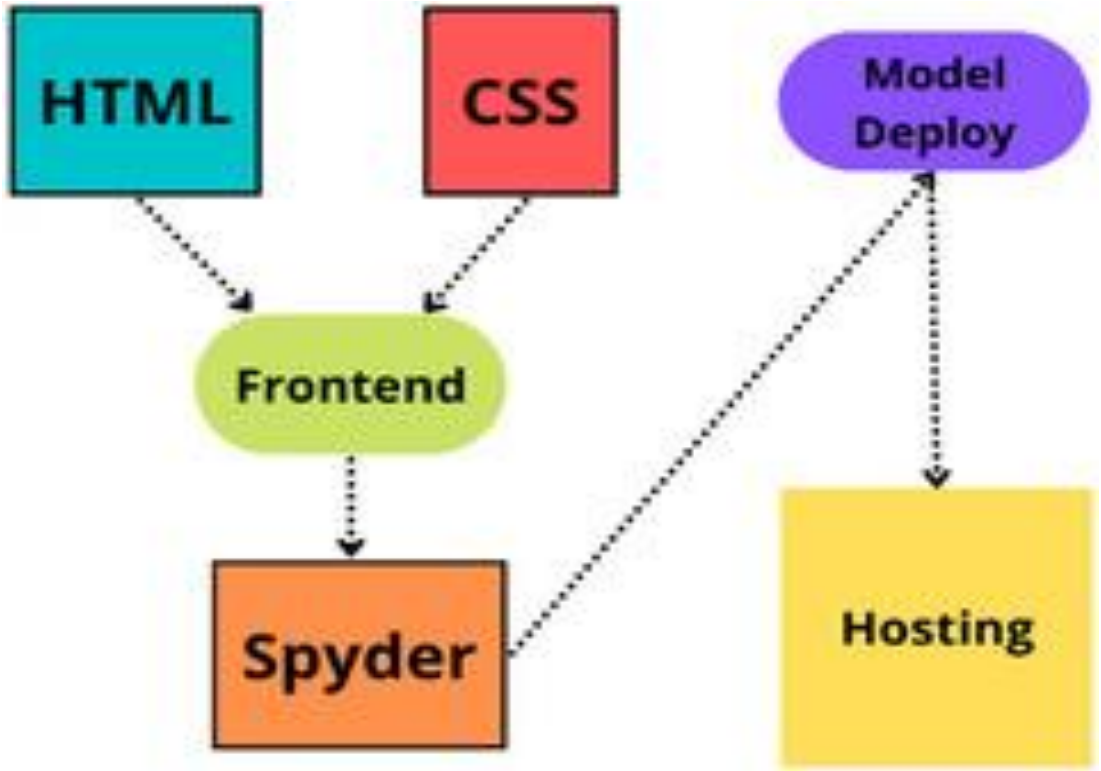
Materials and methods

Logistic regression (LR) is used to identify the risk factors for diabetes disease based on p value and odds ratio (OR). We have adopted four classifiers like naïve Bayes (NB), decision tree (DT), Adaboost (AB), and random forest (RF) to predict the diabetic patients. Three types of partition protocols (K2, K5, and K10) have also adopted and repeated these protocols into 20 trails. Performances of these classifiers are evaluated using accuracy (ACC) and area under the curve (AUC).



Deployment of the prediction system

The proposed machine learning-based diabetes prediction system has been deployed into a website and smartphone application framework to work instantaneously on real data.

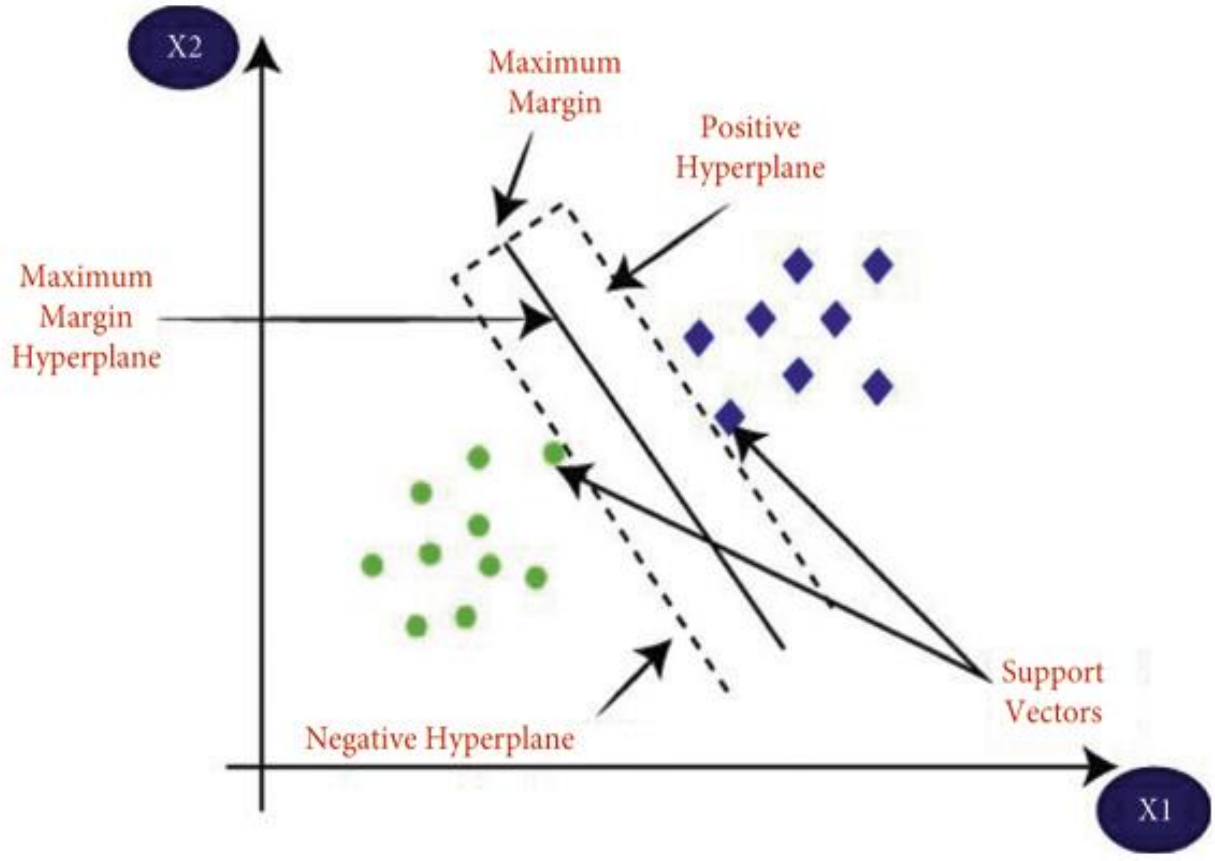


Android smartphone application: To demonstrate the automatic diabetes forecasting system in real time, we also designed an Android smartphone application to test its performance. Android Studio is used for the frontend part of this application. We employed Java as the necessary coding language. After that, the model has been implemented in Android Studio using the pickle package. While developing the API, we used Heroku to host our model on the corresponding hosting server.

Support Vector Machine

SVM is a nonprobabilistic classifier with a separating hyperplane as its formal definition. The technique creates an ideal hyperplane with the greatest distance from the support vectors based on the available training data (supervised learning). This hyperplane is a line that divides a plane into two classes in two-dimensional space. The epsilon ε, regularization, and kernel parameters are the SVM classifier's tuning parameters [6, 44].

The principle of SVM is



Experimental Results

In this section, we evaluate the performance of DNN algorithm by using the testing data to assess the effectiveness of our system based on several evaluation metrics. Besides, comparison between our proposed model and the machine learning algorithms described in section (2.3) has been conducted in order to demonstrate the superiority of our model. The used dataset was split into two subsets, the first one for training which contains 80% of the whole data (547 diabetics/1053 nondiabetics) and the other for testing which contains 20% of the whole data (137 diabetics/263 nondiabetics).

		predicted values	
		0	1
Actual values	0	True Negative (*TN*)	False Positive (FP)
	1	False Negative (FN)	True Positive (*TP*)

Conclusion

The combination of LR and RF-based classifier performs better. This combination will be very helpful for predicting diabetic patients.