DIABETES PREDICTION USING MACHINE LEARNING

Kishan Patel
Department of IT Engineering
PHCET
Rasayani Panvel

Manu Nair
Department of IT Engineering
PHCET
Rasayani , Panvel

Shubham Phansekar
Department of IT Engineering
PHCET
Rasayani , Panvel

Abstract — Diabetes Mellitus is a chronic disease characterized by hyperglycemia. It is a common disease for human body caused by metabolic disorder when the sugar level is high. It can cause many complications. According to growing morbidity by the year 2050, the world's diabetic patients will reach 740 millions, which means that one of ten adults or children may suffer diabetes. Early prediction of such disease can save human life. To achieve this goal researchers are mainly working on this risk factor related to diabetes using machine learning techniques. With rapid development of machine learning, machine learning has been applied in many aspects of medical health. In this study, we are using some popular machine learning algorithms namely, Random Forest, K-Nearest Neighbor (KNN), Decision Tree (DT) and Logistic Regression to predict diabetes mellitus. In our experimental results it shows that Logistic Regression have achieved the highest accuracy compared to other machine learning techniques.

Keywords— Diabetes Mellitus, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, Machine Learning, Prediction
.

I. Introduction

Diabetes is a common chronic disease which can pose great threat to human health. Diabetes can be identified when blood glucose is higher than normal level, which is caused by high secretion of insulin or biological effects. Diabetes can cause various damage to our body and can disfunction tissues, kidneys, eyes and blood vessels. Diabetes can be divided into two categories, type 1 diabetes and type 2 diabetes. Patients with type 1 diabetes are normally younger with an age less then 30 years old. The clinical symptoms are increase thirst and frequent urination this type of diabetes cannot be cleared by medications as it requires therapy. Type 2 diabetes occurs more commonly on middle-aged and old people, which can show hypertension, obesity and other diseases. with our living standards diabetes has increased commonly in people's daily life. So how to analyse diabetes is worth studying. As we get the diagnosis earlier we can control it. Machine learning can make preliminary judgement on diabetes mellitus according to physical examination data, and by reference with doctors. Recently, many algorithms are used to predict diabetes, including machine learning methods like Random Forest, (KNN) K-Nearest Neighbor, Decision Tree and so on. With this machine learning techniques we are able to predict diabetes by constructing predicting models which are obtained by medical datasets. By extracting such knowledge we are able to predict diabetic patient. We use the best technique to predict based on our attributes of the given datasets in order to get the perfect accuracy to predict diabetes mellitus

II. Literature Survey

This section shows our existing recent literature work and provide us the understanding the challenges of our given approaches.

Various computing techniques were used in this healthcare domain. The focus on this literature survey is the use of different machine learning algorithms used for predicting diabetes mellitus. In order to get the perfect accuracy we extract the knowledge from the given medical data. Faisal [1] developed a predictive analysis model using random forest algorithm. The Asaduzzaman [2] used 10 fold cross validation aa an evaluation method for three different algorithms which included decision tree, naive bayes and SVM where naïve bayes have shown the accuracy of 75% than other given algorithms. Chun li [3] used random forest, KNN, naïve bayes, SVM, decision tree to predict diabetes mellitus early stage. Currently in the healthcare domain we are implementing machine learning algorithms and statistical data to understand the diseased data which was discovered. Since the machine learning domain consists of various techniques and researches to make a comparison based on which algorithm is faster in giving the results of prediction. The classification of algorithm was not evaluated by cross validation method. To predict and analyze diabetes mellitus different data mining techniques were used. As we use three data mining techniques we used real word data sets by collecting information from the given datasets.

In this work we have analyzed real diagnostic medical data based on various risk factors for the classification

of machine learning techniques and for predicting diabetes mellitus.

III. METHODOLOGY

In order to achieve our goal, our methodology comprises if few steps from which we accumulate datasets of the given attributes for the patients and we will do the pre-processing of our given attribute to apply on the given machine learning techniques tp find out the predictive analysis of the data.

A. DATASET AND ATTRIBUTES

In this work, we collect diabetes data from Medipath Diagnostic Center (MDC), from Mumbai, Maharashtra, India. The dataset consists of various attributes for diabetes mellitus for 700 patients. the attributes are given in the below table.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0

B. DATA PREPROCESSING

To achieve the goal some data pre-processing is done on the given diabetes dataset. As it converts raw data in numerical form from which we are able to get the values of the attribute to predict diabetes. Here for example we can say that the age of the patient can be divided inti three categories, such as young (10-23 years), adult (24-49 years), old (50 and above). Similarly a patients weight can also be classified into three categories as less (below 40 kg), normal(40-60kg) and overweight (above 60kg). blood pressure is classified as normal (120/80 mmdl), low (less than 80mmdl), high (more than 120 mmdl).

C. APPLYING MACHINE LEARNING TECHNIQUES

Once the data has been created for modelling we employ our four machine learning classification algorithm which we are going to implement to predict diabetes mellitus. Some overview of these techniques.

1. *RANDOM FOREST:*

   Are an ensemble learning method for classification and regression and other task that operates by constructing a multitude of decision tree at training time and outputting the class that is the mode of the classes or mean prediction of individual trees. The first algorithm for random decision forests was created by Tin Kam Ho using random subspace method. Ho established that to gain the accuracy it should over train where it can randomly restrict sensitive selected features of the given data.

2. *K-NEAREST NEIGHBOR:*

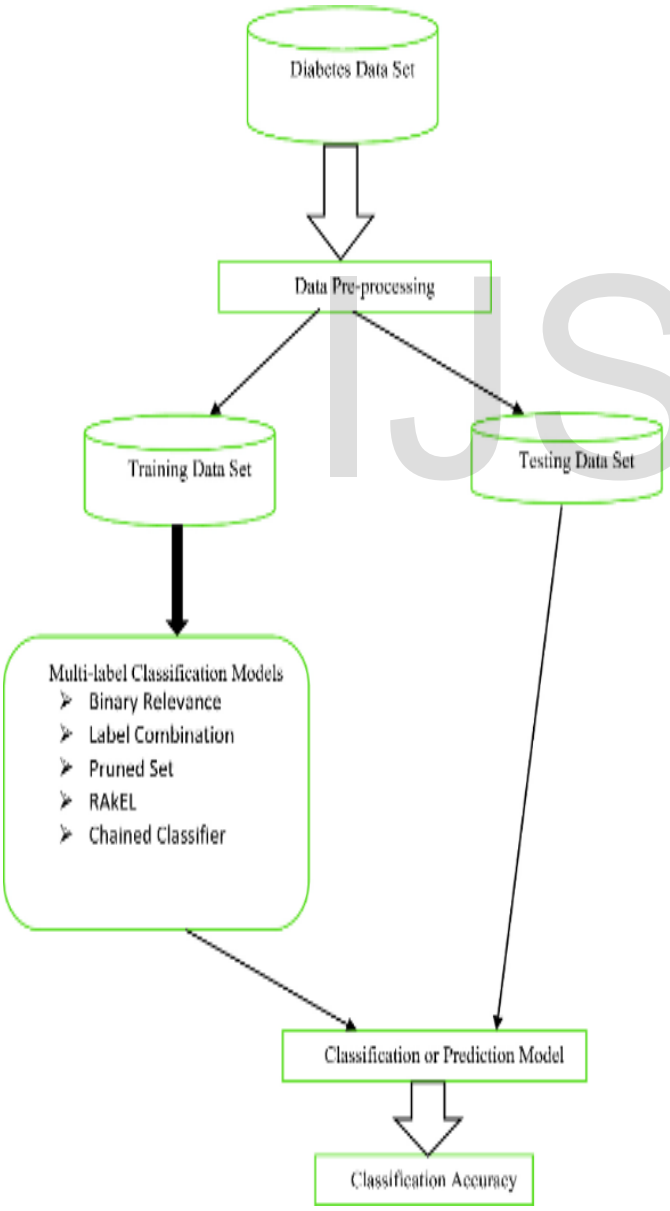
   The k-nearest neighbor is a non-parametric method used for classification and regression. The input consist of k-closest training example in the feature space. To determine the distance from point of interest to point of training data set it uses. In classification technique, the value of k is always a positive integer of the nearest neighbor. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

3. *DECISION TREE:*

   Decision tree are a type of supervised machine learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decision or the final outcome and the decision nodes are where the data is split. The leaf node is labeled by attribute and each attribute is assigned by a target value. The highest information gain of all attribute id calculated first. It is a method commonly used for data mining.

4. LOGISTIC REGRESSION:

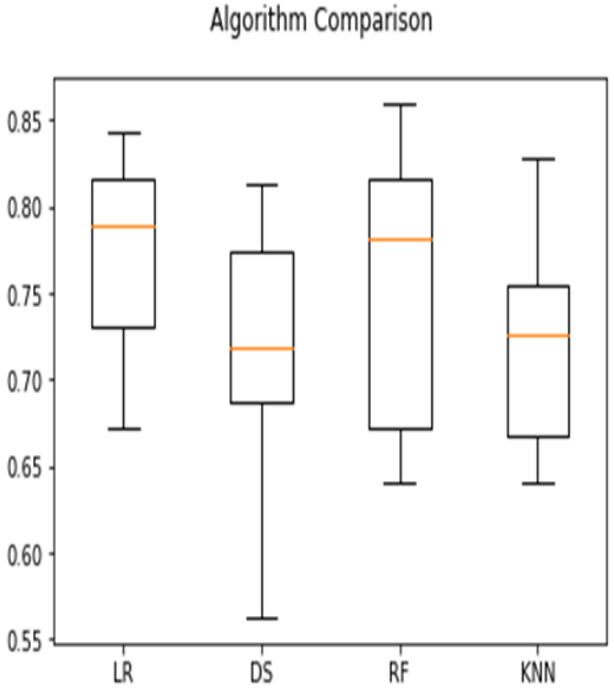
Logistic regression is a machine learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and it is based in the concept of probability.
We can call a logistic regression a linear regression model but the logistic regression uses a more complex cost function, it is also known as logistic function. The hypothesis of logistic regression tends to limit the cost function between 0 and 1. Therefore linear functions fail to show it, as it can make an outcome for the given function to the possible hypothesis if it is great then 1 and if it is less then it is 0.



The figure shows the overall process of the work which is been implemented. After defining the given problem we processed the data for constructing a prediction model. After applying the preprocessed data we get a training dataset and testing dataset. Now after applying the given machine learning techniques we get the test data which is been formed will now give us the performance of the techniques, thus by providing us with the best classifier ofr predicting diabetes mellitus.

D. COMPARISON RESUTS:

The performance of the given machine learning techniques shows us the prediction result where we recall the algorithms used as Random Forest, K-Nearest Neighbor, Decision Tree, Logistic Regression. Here we see that logistic regression classifier shows the better results than other classifier to predict diabetes.
According to the figure logistic regression shows 78% accuracy on this dataset, which is greater than all other machine learning techniques. This it shows that logistic regression performs well on this given medical dataset for predicting diabetes mellitus.



In this work, we have chosen the best machine learning technique to predict diabetes mellitus to that we can achieve high performance based on our

evaluation as shown in the above box plot from which we can estimate the testing dataset to show high performance of precision for our given dataset As logistic regression has considered the best accuracy choice for this dataset as it shows that it can achieve upto 78% of accuracy to predict diabetes mellitus for the medical dataset.

IV. CONCLUSION

In this experiment we have analysed the early prediction of diabetes by taking all the related factors in its tests and implementing using machine learning techniques by extracting knowledge from our real health care medical dataset to predict diabetic patients thus we have done our experiment using some various machine learning algorithms namely Random Forest, K-Nearest Neighbour, Decision Tree, Logistic Regression on Indian datasets to predict diabetes.