Diabetes Disease Prediction using Machine Learning

Lomani Nayak[1], Dr. Gayatri S Pandi[2]

[1]Department of Computer Engineering, L.J Institute of Engineering and Technology, Ahmedabad, India.
[2]Head of Department of Post- Graduation, L.J Institute of Engineering and Technology, Ahmedabad, India.

Abstract: *Diabetes caused due to increase in amount of sugar or glucose which is condensed into the blood Identifying process of Diabetes is the glucose and sugar levels needs to be checked before and after meal, there are fluctuations before and after meal, this whole process of patient visiting a doctor is tiresome. The leap in Machine Learning approaches and algorithms helps us to solve this issue. The motive of this study and research is to make use of features and try to predict the likelihood of the disease, Decision Tree, Random Forest and Support Vector Machine are the algorithm or approaches that have been applied to detect and predict diabetes at an early stage. We have also compared the support vector machine algorithm with k-nearest neighbor and the decision tree.*

Key Words: *Diabetes, Machine Learning, SVM, Decision Tree, K nearest neighbor, Accuracy.*

1. INTRODUCTION:

Diabetes is a condition in which your body is unable to produce the required amount of insulin which is needed to regulate the amount of sugar in the body.

Basically it is found out that there are two general reasons for diabetes:

(1) The pancreas does not produce enough amount of insulin or the body is not able to produce adequate insulin. This kind of problem come under Type-1 diabetes problem.

(2) Cells do not respond to the insulin that is produced is come under the Type-2 diabetes problem machine learning algorithms are used to classify and diagnosis the diseases, in order to eliminate the problem and reduce the required cost. Besides that, using the machine learning algorithm lead to meaningful and accurate decisions.

1.1 Support Vector Machine

A support vector machine (SVM) is a machine learning algorithm that constructs a hyper plane or set of hyper planes in a high dimensional space, which can be used for classification.

1.2 Decision Tree

The main motive using Decision Tree in this research work. For the purpose of classification and prediction nodes and internodes are used. Features or attributes that can be easily differentiated between classes are considered as the root node. Leaf nodes represent classification, while root nodes can have two or more branches where the features are further segregated.

1.3 K NEAREST NEIGHBORS ALGORITHM

KNN is a supervised machine learning algorithm, or known as the lazy learning because it doesn't use the training data at the time of training, it directly make use of testing data The k-NN algorithm is the simplest among all machine learning algorithms .When the size of the dataset is very less so it can easily classify .It does not perform well with the large dataset.

1.4 Naive Bayes (NB)

Naive Bayes classifiers assume attributes have independent distributions. It is considered to be fast and space efficient. It also provides simple approach. It is known as Naive because it relies on two important simplifying assumptions. The predictive attributes are conditionally independent and secondly it assumes that no hidden attributes bias the prediction process. It is very fast to train and fast to classify.

1.5Logistic Regression (LR)

Logistic regression is a type of probabilistic statistical classification model for analyzing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary or dichotomous, that means it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.). Logistic regression generates the coefficients of a formula to predict a log it transformation of the probability of presence of the characteristic of interest.

2. REVIEW WORK:

In this paper, the review and comparison of the various machine learning algorithm.

Rohini Patil, Leena Majumder, Manisha Jain, Vedanti Patil are uses three algorithm SVM, RF and DT.

Prediction of Diabetes Mellitus is made via Data Mining. The goal of Data Mining Methodology is to extract, transform and load data from a dataset and change it into reasonable structure for further use.

J. Krishnendhu, G. Arnesh, B. Harish, K. Vidhya are use of the SVM and Logistic Regression algortithm to get batter prediction.

Md. Aminul Islam, Nusrat Jahan are use the Naïve bayes, Logistic Regression, Multilayer Perceptron, SVM, IBK

Geetha Guttikonda, Madhavi Katamaneni, MadhaviLatha Pandala are use the SVM, Decision Tree, K nearest neighbor proposed a system for diabetes disease classification using Support Vector Machine (SVM) A fast and accurate diabetes prediction system is proposed in this paper. The proposed system focused on the features analysis and classification parts. The effects of using the algorithms of the proposed system through achieving a higher classification rate that the other systems.

In this paper, from different machine learning algorithms Random Forest provide us highest accuracy with ROC Method on Indian Pima Dataset.

3. Problem state:

The current systems working on diabetes disease prediction works on a small dataset. The aim of our system is to work on a larger dataset to increase the efficiency of the overall system. The number of medical tests also affects the performance of the system; thus, our aim is to reduce the number of medical tests to increase the efficiency of the system.

4. Proposed System:

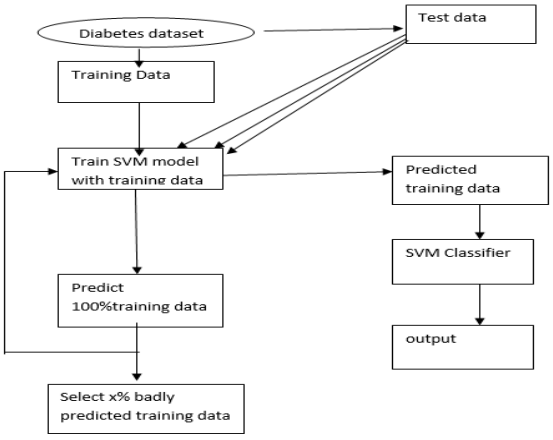

Fig -1: Architecture of the proposed model

The diabetes dataset is divided into two parts such that some randomly samples are chosen from the Diabetes dataset, known as training data and they are trained using the SVM Model. And then the remaining other samples are known as the testing data, and here the actual prediction is performed using this testing data. The entire code is performed in python and machine learning concepts are also used.

4.1 Support Vector Machine Algorithm

SVM is very popular and widely used supervised learning classification algorithm. An advantage of using this algorithm is that it can operate in even infinite dimension. SVM finds a hyper plane that leads to a homogeneous partition of data. A good separation is achieved by the hyper plane that has largest distance to the nearest training data points of any class .so we have to maximize the margin.

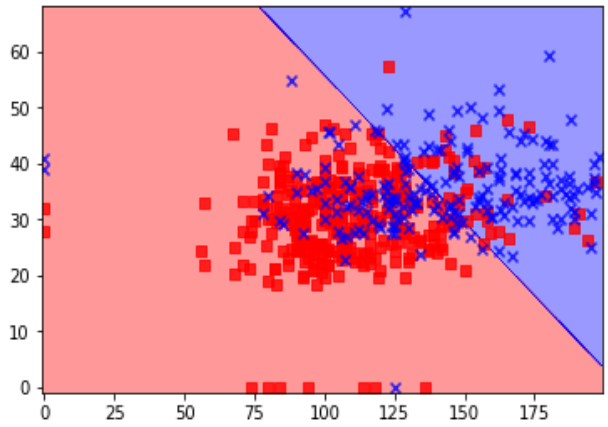

Fig2: when kernel = Linear

When kernel value='LINEAR' is chosen then always the straight line will be drawn and that hyper plane separates the two different classes. This method is found inefficient to classify when data is in 3-D form or more-higher level. It can only be applied when data in 2-D form.

4.2 K NEAREST NEIGHBORS ALGORITHM

KNN is a supervised machine learning algorithm, or also popular and known as the lazy learning, because it doesn't use the training data at the time of training, it directly make use of testing data .The k-NN algorithm is the simplest among all machine learning algorithms .When the size of the dataset is very less so it can easily classify .It does not perform well with the large dataset.

4.3 Decision Tree

The main motive using Decision Tree in this research work. For the purpose of classification and prediction nodes and internodes are used. Features or attributes that can be easily differentiated between classes are considered as the root node. Leaf nodes represent classification, while root nodes can have two or more branches where the features are further segregated.

5. CONCLUSIONS:

We have applied 1. KNN Algorithm 2. SVM Algorithm 3.DecisionTree on the Pima Indian Diabetes Dataset. We compare the three different algorithm base on there accuracy the support vector machine learning got the 73.95% , k Nearest Neighbors algorithm got the 71.35% and the decision tree got the 72% accuracy. We got The best Accuracy of 73.95% using support vector machine algorithm. We were able to perform a lot of data analysis and came to a conclusion that SVM is a good and practical choice to classify a medical data.

Table 1: Accuracy of Algorithm

Algorithm	Accuracy
SVM	73.95%
KNN	71.35%
DT	72%

6. References:

[1].Rohini Patil, Leena Majumder, ”Manisha Jain, Vedanti Patil Diabetes Disease Prediction Using Machine Learning” IJRESM(2020).

[2] J. Krishnendhu, G. Arnesh, B. Harish, K. Vidhya “Diabetes Prediction Using SVM and Logistic Regression ” IJRESM(2020).

[3] KanMhaMustafa S. Kadhm Ikhlas Watan Ghindawi Duaa Enteesha“Diabetes Prediction System Based on K-means Clustering” IJRESM(2018).

[4] Md. Aminul Islam, Nusrat Jahan “Prediction of Onset Diabetes using Machine Learning Techniques “IJRESM(2018)

[5] Geetha Guttikonda, Madhavi Katamaneni, MadhaviLatha Pandala “Diabetes prediction using machine learning “IJRESM(2019)

[6] Mimoh Ojha and Kirti Mathur, "Proposed application of big data analytics in healthcare at Maharaja Yeshwantrao Hospital", 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1-7, 2019

[7] K.I. Nkuma-Udah and G.A. Chukwudebe, "Medical Diagnosis Expert System for Malaria and Related Diseases for Developing Countries", 2017 IEEE 3rd International Conference on Electro-Technology for National Development (NIGERCON)

[8] Faiza Nouh, Mariam Omar and Manal Younis, "Prevalence of Hypertension among Diabetic Patients in Benghazi: A Study of Associated Factors", Asian Journal of Medicine and Health

[9]Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, “Diabetes Disease Prediction Using Data Mining,” 2017.

[10] Vrushali Balpande, Rakhi Wajgi, “Prediction and Severity Estimation of Diabetes Using Data Mining Technique,” 2017.

[11]Vrushali B., and Rakhi W., “Review on Prediction of Diabetes using Data Mining Technique”, International Journal of Research and Scientific Innovation (IJRSI), Volume IV, Issue IA, pp. 43-46, January 2017

[12]G. Krishnaveni*, T. Sudha,” A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques” in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT-2017), vol. 3, Issue 1, pp. 5-11, 2017