International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 694-697 www.ijaem.net ISSN: 2395-5252

Prediction of Diabetes Using Machine Learning

Sanket Khape[1], Sumit Matode[2], Estyak Khan [3]
and Prof. Sangeetha Selvan
Department of Information Technology, PCE, Navi Mumbai, India - 410206

Submitted: 25-05-2021 Revised: 01-06-2021 Accepted: 05-06-2021

ABSTRACT: Diabetes is one of the common and growing diseases in countries and all of them are working to prevent this disease. In this project it predicts the symptoms of diabetes using algorithms. The main point of this is to compare all the algorithms. In this project we use 3 algorithms. Linear regression, Random Forest and Naive Bayes this 3 algorithms are used. To predict symptoms in medical data, various algorithms were used to achieve higher accuracy. But these three algorithms are the most effective algorithms because linear regression performs a regression task. Regression models a target prediction value based on independent variables. Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

KEYWORDS: diabetes mellitus, random forest, decision tree, neural network, machine learning, feature ranking

I. INTRODUCTION

Machine learning methods are widely used in predicting diabetes, and they get preferable results. Random Forest is one of popular machine learning methods in the medical field, which has great classification power. Random forest generates many decision trees. Naive Bayes are a most popular machine learning method, which has a better performance in many aspects. So in this study, we used a Linear Regression, random forest (RF) and Naive Bayes to predict diabetes.

II. LITERATURE SURVEY

2.1-Analysis of Diabetes using machine learning: Healthcare industry contains very large and sensitive data and needs to be handled very carefully. Diabetes Mellitus is one of the growing extremely fatal diseases all over the world. Medical professionals want a reliable prediction system to diagnose Diabetes. Different machine learning techniques are useful for examining the data from diverse perspectives and synthesizing it into valuable information. The accessibility and availability of huge amounts of data will be able to provide us useful knowledge if certain data mining techniques are applied on it. The main goal is to determine new patterns and then to interpret these patterns to deliver significant and useful information for the users. Diabetes contributes to heart disease, kidney disease, nerve damage and blindness. Mining the diabetes data in an efficient way is a crucial concern. The data mining techniques and methods will be discovered to find the appropriate approaches and techniques for efficient classification of Diabetes data set and in extracting valuable patterns. In this study a medical bioinformatics analysis has been accomplished to predict diabetes. The WEKA software was employed as a mining tool for diagnosing diabetes. The Pima Indian diabetes database was acquired from the UCI repository used for analysis. The data set was studied and analyzed to build an effective model that predicts and diagnoses diabetes disease. In this study we aim to apply the bootstrapping resampling technique to enhance the accuracy and then apply Naive Bayes, Decision Trees and (KNN) and compare their performance.

2.2-Classifier models to predict diabetes mellitus.

:Diabetes is one of the common and growing diseases in several countries and all of them are working to prevent this disease at an early stage by predicting the symptoms of diabetes using several methods. The main aim of this study is to compare the performance of algorithms those are used to predict diabetes using data mining techniques. In this paper we compare machine learning classifiers (J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines) to classify patients with diabetes mellitus. These approaches have been tested with data samples downloaded from UCI machine learning data repository. The performances of the algorithms have been measured in both the cases i.e data set with noisy data (before pre-processing) and data set set without noisy data (after pre-

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 694-697 www.ijaem.net ISSN: 2395-5252

processing) and compared in terms of Accuracy, Sensitivity, and Specificity.

2.3- Prediction of Diabetes using data mining approach: The main purpose of this paper is to predict how likely the people with different age groups are being affected by diabetes based on their lifestyle activities and to find out factors responsible for the individual to be diabetic. Hence it is interesting to implement statistical techniques in the medical field to understand which age group of people are being affected by diabetes.

Detection and Prediction of Diabetes Using Machine Learning Techniques Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a "key" to open our cells, to allow the glucose to enter -- and allow us to use the glucose for energy. But with diabetes, this system does not work. Several major things can go wrong – causing the onset of diabetes. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. This paper focuses on recent developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes.

Different Techniques Used For Predicting Diabetes Mellitus In today's world diabetes is the major health challenges in India. It is a group of a syndrome that results in too much sugar in the blood. It is a protracted condition that affects the way the body mechanizes the blood sugar. Prevention and prediction of diabetes mellitus is increasingly gaining interest in medical sciences. The aim of this paper is to conduct a survey on different techniques.

Literature Summary

A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. The summary is presented here.

Table 2.1 Summary of literature survey

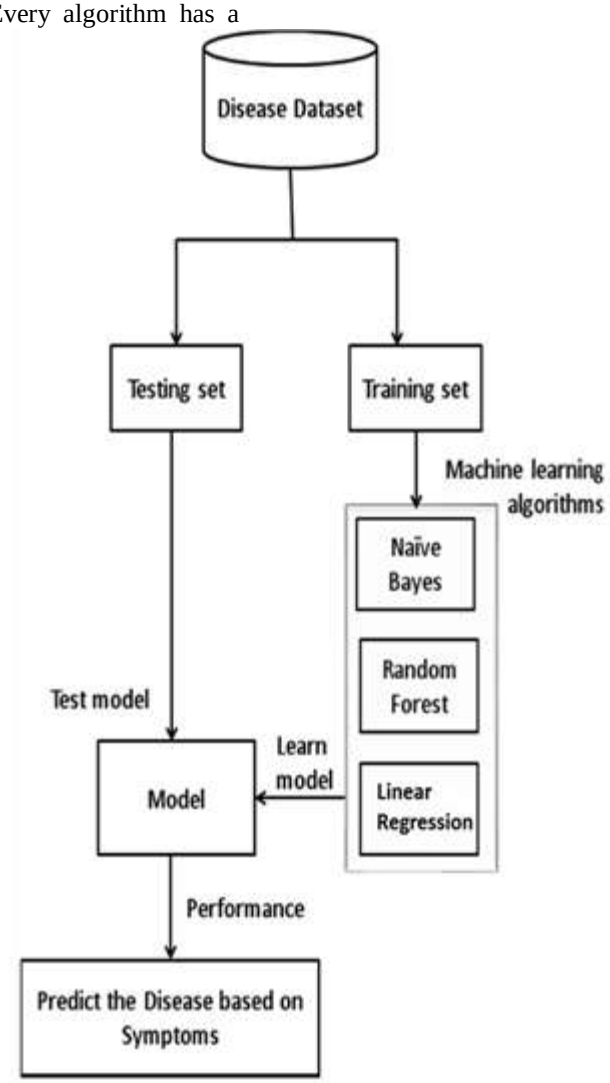
SN	Paper	Advantages and Disadvantages
1.	S.Saru, Subhashree	Advantages: High accuracy with accuracy value of 90.36% and decision Stump provided less accuracy then other by providing83.72 % accuracy Disadvantages:Only limited base classifier used.
2.	J.Pradeep Kandasamy, S.Balamurali	Advantages: 8 attributes under two different situation.one is before pre-processing the dataset .Accuracy of decision tree J48 is 73.82%.Removal of noisy dataset will provide good accuracy. Disadvantages: accuracy is low.
3.	Reshma R,Anjana S Chandran	Advantages:.Consists of double leveled algorithm i,e improved K-means and logic regression algorithms.Used to detect type 2 diabetes mellitus Disadvantage- It consumes more time during the part of preprocessing.
4	Priyanka Indoria, Yogesh Kumar Rathore,	Advantage: Bayes' theorem with strong independence assumptions between features and does not need a long computational time for training which is its major advantage. Disadvantage: According to compared results, the highest accuracy was achieved in Bayesian Network but also the smallest accuracy was shown in Bayesian Network.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 694-697 www.ijaem.net ISSN: 2395-5252

5	Saman hina, Anita Singh, Sohail Abdul Sattar	Advantages:Multilayer perceptron function is most effective hence it shows few errors.ZeroR is useful to determine baseline performance for other classification methods. Disadvantage:It takes too much processing time because it requires calculation of each node.

III. EXISTING SYSTEM

In the current system for prediction of diabetes using machine learning is done through using various different machine algorithms such as Linear Regression , Random Forest etc many such algorithms were used but the main reason is their accurate successful result. Every algorithm has a different success rate and has different ways of prediction. Different combinations of algorithms give different accurate rates. At present the accuracy rate is nearly 85 percent and the data sample they were using was also limited or small sample size and the input function was also limited.



IV. PROPOSED SYSTEM

In view of the problem statement described in the introduction section, we propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different classifiers like Linear Regression, Random Forest and Naive Bayes. The major focus is to increase the accuracy by using

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 694-697 www.ijaem.net ISSN: 2395-5252

resample technique on a benchmark well renowned diabetes dataset that was acquired from PIMA Indian Diabetes Dataset.

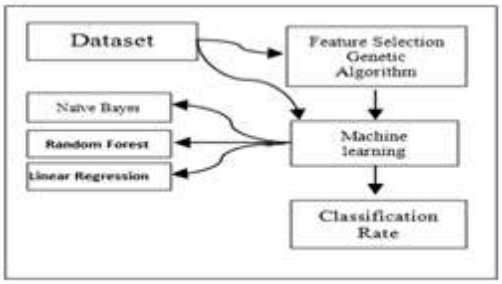


Figure 3.1.1

V. FUTURE SCOPE

In this study we concentrated only Diabetes disease for future it can be extended to apply this method in another diseases Small amount sample data used on this study.it can be apply in large amount of data for future extension .on this study also only a single data set used therefore for future multiple data set can be used for prediction .in this study only limited base classifier used .for future it is possible to use another base classifier like ANN, Naive Bayes, KNN, Random tree ,and other.

VI. CONCLUSIONS

Machine learning methods are widely used in predicting diabetes, and they get preferable results.Neural networks are a recently popular machine learning method, which has a better performance in many aspects. So in this study, we used a Linear Regression, random forest (RF) and Naive Bayes to predict diabetes. There are many challenges in the successful treatment of diabetes mellitus because of personal and economic costs incurred in diabetes therapy.

REFERENCES

[1]. Pamela Fry (Thompson Rivers University). Literature Review Template [Online]. Available FTP:
[2]. https://www.tru.ca/_shared/assets/Literature_Review_Template30564.pdf
[3]. J. K. Author, "Name of paper," Abbrev. Title of Periodical, vol. x, no. x, pp. xxx-xxx, Abbrev. Month, year.
[4]. R. E. Kalman, "New results in linear filtering and prediction theory," J. Basic Eng., ser. D, vol. 83, pp. 95-108, Mar. 1961.
[5]. A. B. Author. (year). Title (edition) [Type of medium]. Available FTP: Directory: File:
[6]. R. J. Vidmar. (1994). On the use of atmospheric plasmas as electromagnetic reflectors [Online]. Available FTP: atmnext.usc.edu Directory: pub/etext/1994 File: atmosplasma.txt.
[7]. Rahul Joshi, Minyechil Alehegn.2017.Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach from International Research Journal of Engineering and Technology, p-ISSN: 2395-0072
[8]. Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. 2016. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.
[9]. Song, Y., Liang, J., Lu, J., & Zhao, X.2017. An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.
[10]. Pradeep, K. R., & Naveen, N. C. 2016. Predictive analysis of diabetes using J48 algorithm of classification techniques In Contemporary Computing and Informatics (IC3I).
[11]. Dr. K. Thangadurai, N.Nandhini.2016. Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus from International Journal of Scientific & Engineering Research, Volume 7, Issue 5, ISSN 2229-5518