PREDICTING DIABETES WITH ADVANCED-ML

Sakshi Gangwar[1], Er. Harshit Gupta[2], Dr. Ruchin Jain[3]

[1]Student, Department of Computer Science and Engineering, Rajshree Institute of Management and Technology, Bareilly, U.P., India.

[2]Assistant Professor, Department of Computer Science and Engineering, Rajshree Institute of Management and Technology, Bareilly, U.P., India.

[3]Head, Department of Computer Science and Engineering, Rajshree Institute of Management and Technology, Bareilly, U.P., India.

DOI: https://www.doi.org/10.58257/IJPREMS40329

ABSTRACT

There are several machines gaining knowledge of strategies which can be used to carry out predictive analytics over massive information in diverse fields. Predictive analytics in healthcare is a hard mission however ultimately can assist practitioners make big facts-knowledgeable timely choices about patient's health and remedy. This paper discusses the predictive analytics in healthcare; six special system mastering algorithms are used in this studies work. For experiment motive, a dataset of patient's medical report is acquired and 6 different gadgets getting to know algorithms are applied on the dataset. Overall performance and accuracy of the carried out algorithms is discussed and in comparison. Assessment of the special device getting to know strategies used in this look at reveals which set of rules is great suitable for prediction of diabetes. This paper objective to assist docs and practitioners in early prediction of diabetes the usage of machine learning techniques.

Diabetes is one of the commonplace and developing illnesses in nations and all of them are working to prevent this disease. On this assignment it predicts the signs and symptoms of diabetes the usage of algorithms. The principle point of this is to evaluate all of the algorithms. On this venture we use 3 algorithms. Linear regression, Random forest and Naive Bayes those 3 algorithms are used. To predict signs and symptoms in scientific records, various algorithms were used to attain higher accuracy. However those 3 algorithms are the handiest algorithms due to the fact linear regression plays a regression challenge. Regression models a target prediction price based totally on independent variables. Naive Bayes set of rules is a supervised learning set of rules, that's primarily based on Bayes theorem and used for solving category issues. Random forest builds multiple decision timber and merges them collectively to get an extra correct and strong prediction.

Keywords: Diabetes mellitus, Random forest, Decision tree, neural network, Machine learning, Feature ranking.

1. INTRODUCTION

In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source – everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose.

The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas.

Insulin works like a key to a door. Insulin attaches itself to ‘doors’ on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycemia) and diabetes develops. Diabetes Mellitus

means high levels of sugar (glucose) in the blood stream and in the urine. 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.

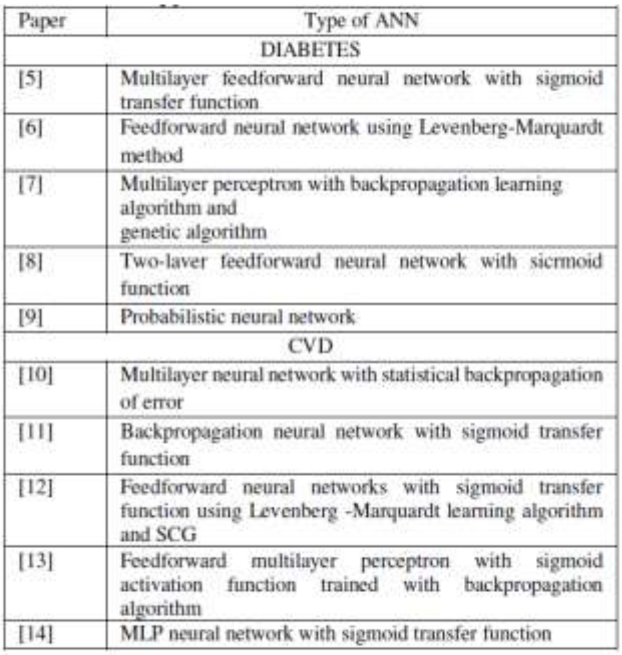
2. LITERATURE SURVEY

2.1 Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases

In this paper authors designed to perform a review of Artificial Neural Network and Bayesian Network and their application in classification of diabetes and CVD diseases. The purpose is to show the comparison of machine learning techniques and to discover the best option for achieving the highest output accuracy of the classification. This paper represents the comparison of application of two machine learning techniques, Artificial Neural Network and Bayesian Network in classification of diabetes and cardiovascular diseases.

but in different field of studies, the literature review was done using 20 published papers in order to obtain the relevant results about diabetes and CVD classification in the period from 2008 to 2017.

Table 1 Ann Types for Classification of Diabetes and Cvd



Paper	Type of ANN
DIABETES	
[5]	Multilayer feedforward neural network with sigmoid transfer function
[6]	Feedforward neural network using Levenberg-Marquardt method
[7]	Multilayer perceptron with backpropagation learning algorithm and genetic algorithm
[8]	Two-laver feedforward neural network with sicrmoid function
[9]	Probabilistic neural network
CVD	
[10]	Multilayer neural network with statistical backpropagation of error
[11]	Backpropagation neural network with sigmoid transfer function
[12]	Feedforward neural networks with sigmoid transfer function using Levenberg -Marquardt learning algorithm and SCG
[13]	Feedforward multilayer perceptron with sigmoid activation function trained with backpropagation algorithm
[14]	MLP neural network with sigmoid transfer function

2.2 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 re-sampling technique to enhance the accuracy and then apply Naive Bayes, Decision Trees and (KNN) and compare their performance.

2.3 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 without noisy data (after pre-processing) and compared in terms of Accuracy, Sensitivity and Specificity.

2.4 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 groups of people are being affected by diabetes.

2.5 Detection and Prediction of Diabetes Using Machine Learning Techniques

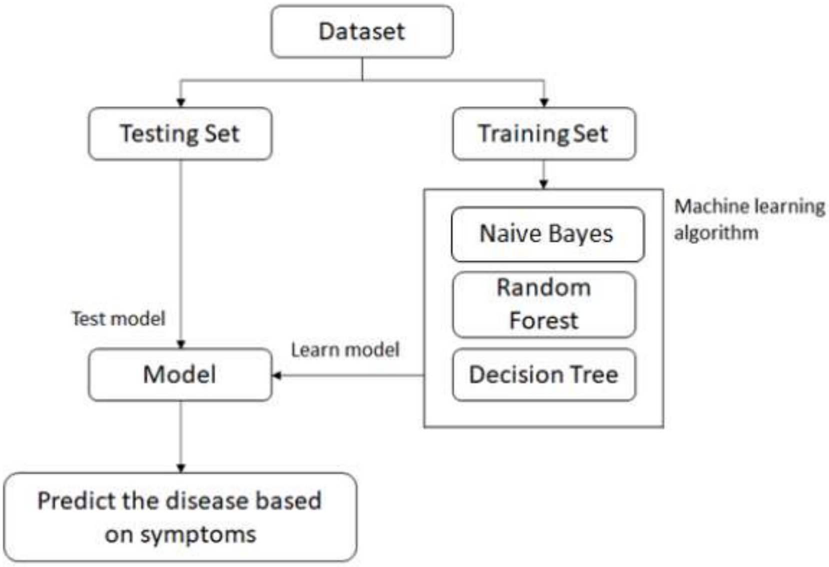
Signs or Symptoms of Diabetes: Frequent Urination, Increased thirst, increased hunger, Tired/Sleepiness, Weight loss, Blurred vision. Mood swings, Confusion and difficulty concentrating, frequent infections / poor healing. Type 1 diabetes: In Type 1 diabetes the beta cells of the pancreas have been injured or attacked by the body’s own immune system (auto immunity). As a result of this attack, the beta cells die and are therefore unable to make the required amount of insulin to move glucose into the cells, causing high blood sugar (hyperglycemia). Type 1 diabetes occurs in about 5 -10% of those with diabetes and usually in people less than 30 years of age, but can occur at any age. The signs and symptoms have a rapid onset and are usually intense in nature. As Type 1 diabetes is caused by a lack of insulin, people need to replace what the body cannot produce itself. According to the latest **American Heart Association's** Heart Disease and Stroke Statistics, about 8 million people 18 years and older in the United States have type 2 diabetes and do not know it. Often type 1 diabetes remains undiagnosed until symptoms become severe and hospitalization is required. Left untreated, diabetes can cause a number of health complications. That's why it's so important to both know what warning signs to look for and to see a health care provider regularly for routine wellness screenings. 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.

2.6 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.

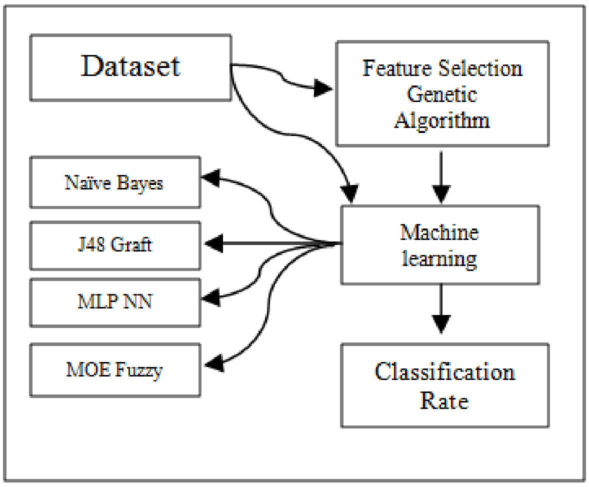
3. EXISTING SYSTEM

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.

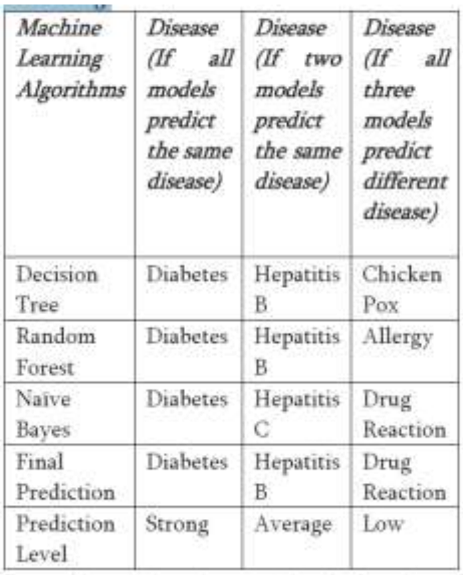


4. PROPOSED SYSTEM

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



Therefore, we have used three different ways to obtain more accurate predictions. If all classifiers are predicting the different diseases, then final prediction is considered on the basis of Naïve Bayes classifier. Because, a Naïve based classifier gives more accuracy and also it doesn't have the problem of over fitting.



Machine Learning Algorithms	*Disease (If all models predict the same disease)*	*Disease (If two models predict the same disease)*	*Disease (If all three models predict different disease)*
Decision Tree	Diabetes	Hepatitis B	Chicken Pox
Random Forest	Diabetes	Hepatitis B	Allergy
Naïve Bayes	Diabetes	Hepatitis C	Drug Reaction
Final Prediction	Diabetes	Hepatitis B	Drug Reaction
Prediction Level	Strong	Average	Low

5. RESULT

The model trained on 132 symptoms and 42 diseases and its respective medicines. From the below table, we can infer that all the three algorithms show excellent results but Naive Bayes performs the best and achieves the highest accuracy of 98.12 percent. The training accuracy has been described in the table below. As we can see that the efficiency of training is higher in Naïve Based classifiers. It is because it overcomes the problem of over fitting, which is common in the case of Decision Tree and Random Forest classifiers. We have seen that there are many diseases that share common medicines for treatment, if symptoms are common between the diseases. So, the algorithm finds the most common disease then suggests the medicine to the user.

Table of algorithms and accuracy

ALGORITHM	**ACCURACY**
DECISION TREE	**0.9763**
RANDOM FOREST	**0.9793**
NAIVES BAYES	**0.9887**

6. 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.

7. 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. Moreover, our system also recommends the suitable medicine for the predicted diseases. Now we set out to create a system which can predict disease and its medicine on the basis of symptoms given to it. On an average we achieved accuracy of ~98%. System has an easy-to-use interface so anyone can use it very easily.

8. REFERENCES

[1] www.diabetesresearch.org/document.doc?id=284

[2] D. Yu, and L. Deng, 2011, “Deep learning and its applications to signal and information processing,” IEEE Signal Process. Mag., vol. 28, no. 1, pp. 145-154.

[3] Habibi, N., Hashim, S. Z. M., Norouzi, A., & Samian, M.R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. BMC bioinformatics, 15(1), 134.

[4] Langarizadeh, M., & Moghbeli, F. (2016). Applying Naïve Bayesian Networks to Disease Prediction: a Systematic Review. Acta Informatica Medica, 24(5), 364.

[5] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using artificial neural network. International Journal of Scientific and Engineering Research, 5(10).

[6] Jayalakshmi, T., & Santhakumaran, A (2010, February). A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. OSDE, 159-163.(2010)

[7] Pradhan, M., & Sahu, R. K. (2011). Predict the onset of diabetes disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies

[8] (E-ISSN: 2044-6004).

[9] Sejdinovic, Dijana, et al. "Classification of Prediabetes and Type 2 Diabetes using Artificial Neural Network." Springer. CMBEBIH 2017.

[10] Soltani, Z., & Jafarian, A (2016). A New Artificial Neural Networks Approach for diagnosing Diabetes Disease Type II. International Journal of Advanced Computer Science & Applications, 1(7), 89-94.

[11] Atkov, O. Y., Gorokhova, S. G., Sboev, A G., Generozov, E. Y., Muraseyeva, E. v., Moroshkina, S. Y, & Cherniy, N. N. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of cardiology, 59(2), 190-194.

[12] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseases diagnosis using neural networks arbitration. International Journal of Intelligent Systems and Applications, 7(12), 72.

[13] Colak, M. C. et. al., Predicting coronary artery disease using different artificial neural network modelslkoroner arter hastaliginin degisik yapay sinir agi modelleri lie tahmini. The Anatolian Journal of Cardiology (Anadolu Kardiyoloji Dergisi), 8(4), 249-255, (2008).

[14] Can, M. (2013). Diagnosis of cardiovascular diseases by boosted neural networks.

[15] Sayad, A T., & Halkarnikar, P. P. Diagnosis of heart disease using neural network approach. In Proceedings of IRF International Conference, 13th April-2014, Pune, India, ISBN (pp. 978-93).

[16] Pamela Fry (Thompson Rivers University). Literature Review Template [Online]. Available FTP: https://www.tru.ca/shared/assets/Literature_Review_Template30564.pdf

[17] R.E. Kalman, “New results in linear filtering and prediction theory,” J. Basic Eng., ser. D, vol. 83, pp. 95-108, Mar. 1961.

[18] 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.

[19] 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.

[20] 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.

[21] Song, Y., Liang, J., Lu, J., & Zhao, X.2017. An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26-34.