IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115170

Diabetes Disease Prediction using Machine Learning Technique

Dr. G RAJIV SURESH KUMAR[1], Shubham Kumar Mishra[2], Merwin Prabhu[3], Vishnu Priya MK[4], Sruthi S[5]

[1]PROFESSOR & HOD (CSE), JCT College of Engineering and Technology

[2-5]BE Computer Science Engineering, JCT College of Engineering and Technology

Abstract: we aim to develop a prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Ensemble Decision Tree Algorithms for high feature selection. A significant attention has been made to the accurate detection of diabetes which is a big challenge for the research community to develop a diagnosis system to detect diabetes in a successful way in the e-healthcare environment. The existing diagnosis systems have some drawbacks, such as high computation time, and low prediction accuracy. To handle these issues, we have proposed diagnosis system using machine learning methods, such as preprocessing of data, feature selection, and classification for the detection of diabetes disease in e- healthcare environment. Model validation and performance evaluation metrics have been used to check the validity of the proposed system. We have proposed a filter method based on the Decision Tree algorithm for highly important feature selection. Two ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest are also used for feature selection and compared the classifier performance with Wrapper based feature selection algorithms also. Machine learning classifier Decision Tree has been used for the classification of healthy and diabetic subjects. The experimental results show that the Decision Tree algorithm based on selected features improves the classification performance of the predictive model and achieved optimal accuracy. Additionally, the proposed system performance is high as compared to
the previous state-of-the-art methods. High performance of the proposed method is due to the different combinations of selected features set. Furthermore, the experimental results statistical analysis demonstrated that the proposed method would be effectively detected diabetes disease.

Index Terms: Machine Learning , Random Forest, PIMA Dataset, IDT-3, ADA Boost, e-Health Care.

I.INTRODUCTION

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.
The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers. A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

A. Problem Definition

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.

II.LITERATURE REVIEW

Following is some of the search which has been reviewed for the proposed system: -
1) Intelligible support vector machines for diagnosis of diabetes mellitus

N. H. Barakat, et al [1]Diabetes mellitus is a chronic disease and a major public health challenge worldwide. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. Furthermore, 3.8 million deaths are attributable to diabetes complications each year. It has been shown that 80% of type 2 diabetes complications can be prevented or delayed by early identification of people at risk. In this context, several data mining and machine learning methods have been used for the diagnosis, prognosis, and management of diabetes. In this paper, we propose utilizing support vector machines (SVMs) for the diagnosis of diabetes. In particular, we use an additional explanation module, which turns the “black box” model of an SVM into an intelligible representation of the SVM's diagnostic (classification) decision. Results on a real-life diabetes dataset show that intelligible SVMs provide a promising tool for the prediction of diabetes, where a comprehensible ruleset have been generated, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Furthermore, the extracted rules are medically sound and agree with the outcome of relevant medical studies.

2) Medical diagnosis on Pima Indian diabetes using general regression neural networks

K. Kayaer and T. Yıldırım [2] The performance of recently developed neural network structure, general regression neural network (GRNN), is examined on the medical data. Pima Indian Diabetes (PID) data set is chosen to study on that had been examined by more complex neural network structures in the past. The results of early studies and of the GRNN structure presented in this paper is compared. Close classification accuracy to the reference work using ARTMAP-IC structured model, which is the best result obtained since now, is achieved by using GRNN, which has a simpler structure. The performance of the standard multilayer perceptron (MLP) and radial basis function (RBF) feed forward neural networks are also examined for the comparison as they are the most general and commonly used neural network structures. The performance of the MLP was tested for different types of backpropagation training algorithms.

I. PROPOSED SYSTEM

In this system we design e-health care diagnosis system for diabetes detection.

In this system we propose Filter based DT-(ID3) algorithm for features selection and the proposed algorithm select more appropriate features from the dataset. Also, two DT ensembles algorithms, such as Ada Boost and Random Forest are used for feature selection and compared the performance of DT on the proposed feature selection algorithm with these two FS algorithms and Wrapper based feature selection methods.

In this system we use the classifier DT and the performance have been checked on original features set and on selected features set along with cross validation methods, such as Training/testing set, K-fold, and LOSO. The LOSO is more suitable than train/test and k-folds validations.

In this system we recommend that the proposed method can be used to effectively detect the diabetes disease and the system can be easily incorporated in healthcare.

2. ALGORITHMS

Random Forest:

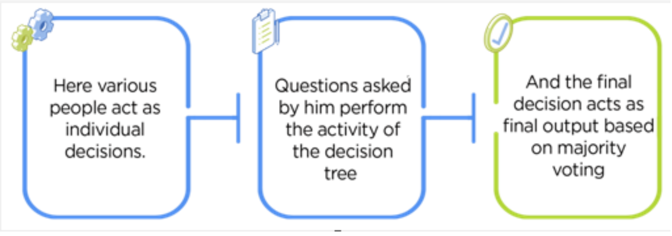
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems.

It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.

It performs better results for classification problems.

TABLE I. Random Forest EXAMPLE


Here various people act as individual decisions.
Questions asked by him perform the activity of the decision tree
And the final decision acts as final output based on majority voting

Problem: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

TABLE II. FREQUENCY TABLE

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	1	2
Sunny	2	3
Grand total	5	9

TABLE III. LIKELIHOOD TABLE

Likelihood Table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	= 5/14	= 9/14		
	0.36	0.64		

P (Yes in Sunny) = P (of Sunny in Yes) * P (Yes) / P (Sunny)

Now that we possess P (Sunny in Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Also, there is P (Yes in Sunny) = 0.33 * 0.64 / 0.36 = 0.60, so this has probability.

AdaBoost:
AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method.
The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split.

These trees are also called Decision Stumps.

A. Working of Ada Boost:

Steps For Ada Boost: -

The example is in below Fig. 2 to acknowledge this algorithm. Following is a wide spread of red circles (RC) and green squares (GS):

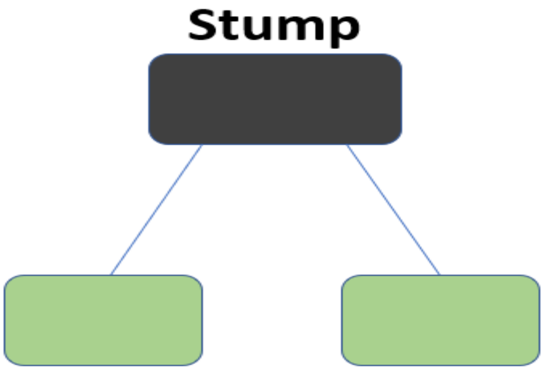


Fig 2. Ada Boost Example

DOI: 10.17148/IJARCCE.2022.115170

The formula to calculate the sample weights is:

$$w(x_i, y_i) = \frac{1}{N}, \quad i = 1, 2, \ldots . n$$

Where N is the total number of datapoints Here since we have 5 data points so the sample weights assigned will be 1/5.

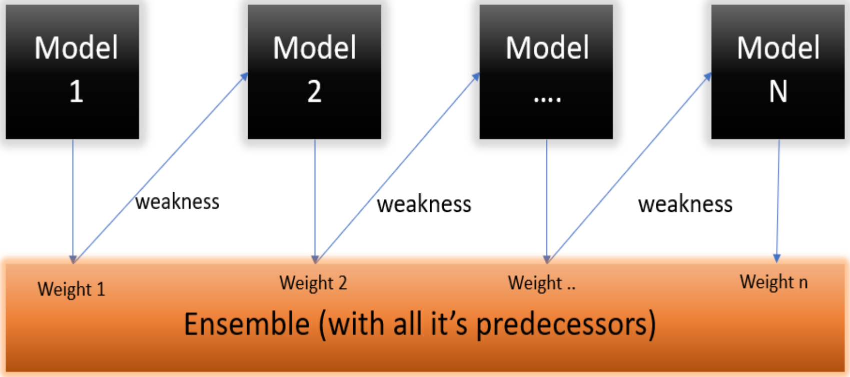


Fig 3. K-NN Example

$$Performance\ of\ the\ stump = \frac{1}{2}\log_e\left(\frac{1 - Total\ Error}{Total\ Error}\right)$$

$$\alpha = \frac{1}{2}\log_e\left(\frac{1-\frac{1}{5}}{\frac{1}{5}}\right)$$

$$\alpha = \frac{1}{2}\log_e\left(\frac{0.8}{0.2}\right)$$

$$\alpha = \frac{1}{2}\log_e(4) = \frac{1}{2}*(1.38)$$

$$\alpha = 0.69$$

1. Assign **equal weights** to all the datapoints
2. Find the stump that does the **best job classifying** the new collection of samples by finding their Gini Index and selecting the one with the lowest Gini index
3. Calculate the **"Amount of Say"** and **"Total error"** to update the previous sample weights.
4. Normalize the new sample weights.

III. EXPECTED RESULT

The goal of our project is to know whether patient is diabetic or not, patient will be diagnosed and it will be depending on the attributes that we are going to take, such as age, pregnancy, pg concentration, tri fold thick, serum ins, body mass index (bmi), dp function, diastolic bp i.e. the factors which are majorly responsible for diabetes.

So, to reduce the correctly know whether the patient is diabetic or not, we are developing a system which will be a prediction system for the diabetes patients. Another best thing about the system is it is will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that we are going to use added the recommendations we are going to provide based on the diabetic levels of the patients. Also, the prediction of the disease will be done with the help of Random Forest Algorithm algorithm and Ada Boost algorithm.

CONCLUSIONS

By our in-depth analysis of literature survey, we acknowledged that the prediction done earlier did not use a large dataset [12]. A large dataset ensures better prediction. Also what it lacks is recommendation system. When we predict we will give some recommendation to the patient on how to control or prevent diabetes in case of minor signs of diabetes.

The recommendations would be such, that when followed it will help the patient. Thus we will build up a system which will anticipate diabetic patient with the assistance of the Knowledge base which we have of dataset of around 2000 diabetes

DOI: 10.17148/IJARCCE.2022.115170

patients and furthermore to give suggestions on the premise of the nearness of levels of diabetes patients. Prediction will be done with the help of two algorithms Random Forest and Ada Boost Neighbor and also we will compare which algorithm gives better accuracy on the basis of their performance factors. This system which will be developed can be used in HealthCare Industry for Medical Check of diabetes patients.

FUTURE SCOPE

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes:

1. Increase the accuracy of the algorithms.
2. Improvising the algorithms to add more efficiency of the system and enhance its working.
3. Working on some more attributes so to tackle diabetes even more.
4. To make it as a complete healthcare diagnosis system to be used in hospitals.

REFERENCES

[1] Y. Cai, D. Ji,D. Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor”, Proceedings of NTCIR-8 Workshop Meeting, 2010.

[2] I. Rish, “An empirical study of the naive Bayes classifier”, T.J. Watson Research Center, 2001.

[3] M.Elkourdi, A.Bensaid, T.Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm”, Alakhawayn University, 2001.

[4] L.Wang, L.Khan and B.Thuraisingham,“An Effective Evidence Theory based on nearest Neighbor (KNN) classification”, IEEE International Conference, 2008.

[5] M.Muja, David G.Lowe, “Fast Approximate Nearest Neighbors

[6] B. Gallwitz, ‘‘Implications of postprandial glucose and weight control in people with type 2 diabetes: Understanding and implementing the inter- national diabetes federation guidelines,’’ Nov. 2009

[7] Ramezani, Rohollah, Mansoureh Maadi, and Seyedeh Malihe Khatami. "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis." Alexandria engineering journal 57.3 (2018): 1883-1891

[8] Amin Ul Haq et.al, Comparative Analysis of the Classification Performance of Machine Learning Classifiers and Deep Neural Network Classifier for Prediction of Parkinson Disease, 2018 15th International computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE,14- 16 Dec 2018.

[9] Y. Liu et al., "Detecting Diseases by Human-Physiological-Parameter-Based Deep Learning," in IEEE Access, vol. 7, pp. 22002-22010, 2019. doi: 10.1109/ACCESS.2019.2893877

[10] A. Tsanas, et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 59, pp. 1264-1271, January 2020.