International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

A Novel Classification Model to Predict Diabetes Using Machine Learning and Data Mining Techniques

Dr. Gangolu Yedukondalu[1], Prof. M. Srinivasa Rao[2], Prof. J. Anand Chandulal[3]
[1]CSE Dept, Vignan Institute of Technology and Science, Hyderabad, India
[2] Retd. Professor, School of Information Technology, JNTU Hyderabad, India
[3]SASI Institute of Technology and Engineering, Andhra Pradesh, India

Article Info
Volume 9, Issue 1
Page Number : 287-293

Publication Issue
January-February-2022

Article History
Accepted : 05 Jan 2022
Published : 20 Jan 2022

ABSTRACT

Now-a-days, people face various diseases due to the environmental condition and their living habits. So, the prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. The diabetes is one of lethal diseases in the world. It is additional a inventor of various varieties of disorders foe example: coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation. Due to every time they have to invest their time and currency. But with the growth of Machine Learning methods, we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital. Information withdrawal has the flexibility to remove unseen data from a large quantity of diabetes associated information. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms. For Decision Tree, the models give precisions of 88%, for Naive Bayes 79% and 78.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.

Keywords- Machine Learning, Support vector machine, Artificial Neural Network, Decision Tree, Naive Bayes, Data Mining.

I. INTRODUCTION

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience. The machine learning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep the track of patient's health. ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

Diabetes is a situation which causes deficiency due to less amount of insulin in the blood. Warning sign of high blood sugar results in frequent urination, feeling thirsty, increased hunger. If it is not medicated, it will lead to many difficulties. This difficulty lead to death. Severe difficulties lead to
cardiovascular disease foot sores, and eye blurriness. When there is a rise within the sugar level within the blood, it is referred to as prior diabetes. The prior diabetes isn't therefore great than the traditional worth. Diabetes is appreciations to either the exocrine gland not manufacturing plentiful
hypoglycemic agent not responding properly to the hypoglycemic agent created. Various information mining algorithms presents different decision support systems for assisting health specialists. The effectiveness of the decision support system is recognized by its accuracy. Therefore, the objective is to build a decision support system to predict and diagnose a certain disease with extreme amount of precision. The AI consist of ML which is its subfield that resolves the real world difficulties by "providing learning capability to workstation without supplementary program writing.

II. LITERATURE SURVEY

Dr. M. Renuka Devi and J. Maria Shyla has discussed about the analysis of various skills of mining to guess diabetes using Naive Bayes, Random forest, Decision Tree and J48 algorithms [5].

Rahul Joshi and Minyechil Alehegn has discussed the ML techniques which are used to guess the datasets at an initial phase to save the life. Using KNN and Naive Bayes algorithm [6].

Zhilbert Tafa and Nerxhivane Pervetica has discussed the result of algorithms that are implemented in order to progress the diagnosis reliability [7].

Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. has discussed the study of Machine Learning Algorithms such as Support Vector Machine, Naïve Bayes, Decision Tree, PCA for Special Disease Prediction using Principal of Component Analysis [11].

Ridam Pal , Dr.Jayanta Poray and Mainak Sen has presented the Diabatic Retinopathy (DR) which is one of the leading cause of sight inefficiency for diabetic patients. In which they reviewed the performance of a set of machine learning algorithms and verify their performance for a particular data set [3].

Veena Vijayan V. And Anjali C has discussed, the abetes disease produced by rise of sugar level in the plasma. Various computerized information systems were outlined utilizing classifiers for anticipating and

diagnosing diabetes using decision tree, SVM, Naive Bayes and ANN algorithms
[1].

P. Suresh Kumar and V. Umatejaswi has presented the algorithms like Decision Tree, SVM, Naive Bayes for identifying diabetes using data mining techniques [2].

The scope of this research is primarily on the performance analysis of disease prediction approaches using different variants of supervised machine learning algorithms. Disease prediction and in a broader context, medical informatics, have recently gained significant attention from the data science research community in recent years. This is primarily due to the wide adaptation of computer-based technology into the health sector in different forms (e.g., electronic health records and administrative data) and subsequent availability of large health databases for researchers. These electronic data are being utilized in a wide range of healthcare research areas such as the analysis of healthcare utilization [10], measuring performance of a hospital care network, exploring patterns and cost of care, developing disease risk prediction model, chronic disease surveillance, and comparing disease prevalence and drug outcomes. Our research focuses on the disease risk prediction models involving machine learning algorithms (e.g., support vector machine, logistic regression and artificial neural network), specifically - supervised learning algorithms. Models based on these algorithms use labelled training data of patients for training. For the test set, patients are classified into several groups such as low risk and high risk.

III. SUPERVISED MACHINE LEARNING ALGORITHM

At its most basic sense, machine learning uses programmed algorithms that learn and optimize their operations by analyzing input data to make predictions within an acceptable range. With the feeding of new data, these algorithms tend to make more accurate predictions. Although there are some variations of how to group machine learning algorithms, they can be divided into three broad categories according to their purposes and the way the underlying machine is being taught. These three categories are: supervised, unsupervised and semi-supervised.

In supervised machine learning algorithms, a labelled training dataset is used first to train the underlying algorithm. This trained algorithm is then fed on the unlabeled test dataset to categories them into similar groups. Using an abstract dataset for three diabetic patients, Fig. 1 shows an illustration about how supervised machine learning algorithms work to categories diabetic and non-diabetic patients. Supervised learning algorithms suit well with two types of problems: classification problems; and regression problems. In classification problems, the underlying output variable is discrete. This variable is categorized into different groups or categories, such as ‘red’ or ‘black’, or it could be ‘diabetic’ and ‘non-diabetic’. The corresponding output variable is a real value in regression problems, such as the risk of developing cardiovascular disease for an individual. In the following subsections, we briefly describe the commonly used supervised machine learning algorithms for disease prediction.

IV. TYPES OF DIABETES

Type one diabetes outcomes due to the failure of pancreas to supply enough hypoglycemic agent. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in children beneath twenty years old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended by doctors.

The type two diabetes starts with hypoglycemic agent resistance, a situation inside which cells fail to response the hypoglycemic agents efficiently. The sickness develops due to the absence of hypoglycemic agent that additionally built.

This type was spoken as "non-insulin-dependent polygenic disease mellitus". The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existences of diabetes mellitus are condensed by 3% in rural zone as compared to urban zone. The pre hyper tension is joined with bulkiness, fatness and diabetes mellitus. The study found that an individual United Nations agency has traditional vital sign.

Type 3 Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women in pregnancy have diabetes. So, in the older age there is a risk of emerging the gestational diabetes in pregnancy.

The obesity is one of the main reasons for type-2 diabetes. The type-2 polygenic disease are under control by proper workout and taking appropriate regime. When the aldohexose level isn't reduced by the higher strategies then medications are often recommended. The polygenic disease static report says that 29.1 million people of the United States inhabitants has diabetes.

V. PROPOSED SYSTEM

Most of the chronic diseases are predicted by our system. It accepts the structured type of data as input to the machine learning model. This system is used by end-users i.e. patients/any user. In this system, the user will enter all the symptoms from which he or she is suffering. These symptoms then will be given to the machine learning model to predict the disease. Algorithms are then applied to which gives the best accuracy. Then System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. Naïve Bayes algorithm is used for predicting the disease by using symptoms, for classification KNN algorithm is used, Logistic regression is used for extracting features which are having most impact value, the Decision tree is used to divide the big dataset into smaller parts. The final output of this system will be the disease predicted by the model.

VI. Methodology

To calculate performance evaluation in the experiment, first, we denote TP, TN, Fp and FNias true positive(the number of results correctly predicted as required), true negative (the number of results not required), false positive (the number of results incorrectly predicted as required), false negative(the number of results incorrectly predicted as not required)respectively. We can obtain four measurements: recall, precision, accuracy, and F1 measures as follows:

The proposed system focuses using algorithms combinations shown above in the block diagram. The base classification algorithms are: Decision tree, Support Vector Machine, Naive Bayes and ANN for accuracy authentication.

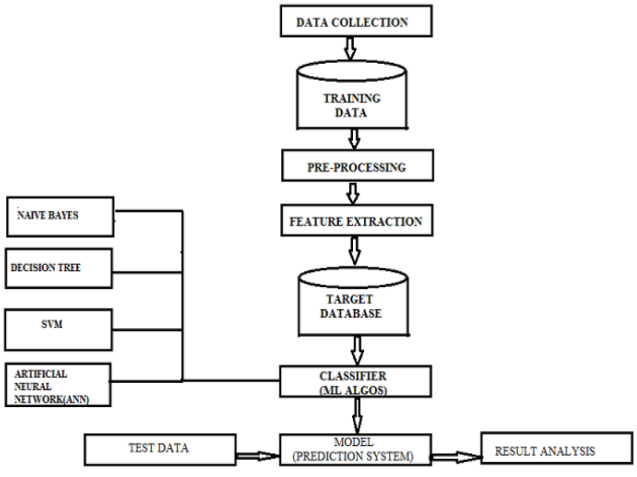


Fig 1.1 Block diagram of diabetes prediction system

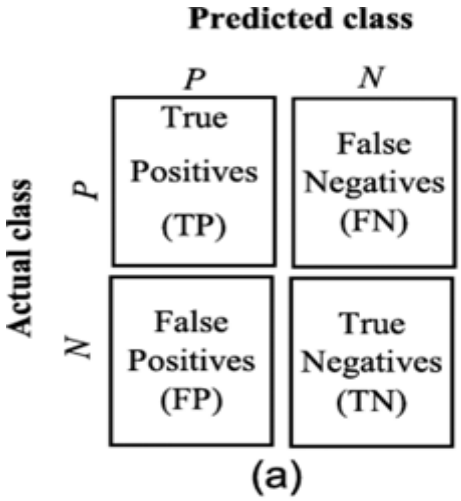
VII. DATA SET DESCRIPTION

Global dataset:

The training phase is completed. The dataset contains seven sixty eight instances and nine features. The dataset features are:

- Total number of times pregnant
- Glucose/sugar level
- Diastolic Blood Pressure
- Body Mass Index (BMI)
- Skin fold thickness in mm
- Insulin value in 2 hour
- Hereditary factor- Pedigree function
- Age of patient in years

Percentage split option is provided for training and testing. Out of 768 instances 75 % is used for training and 25% is used for testing [1].



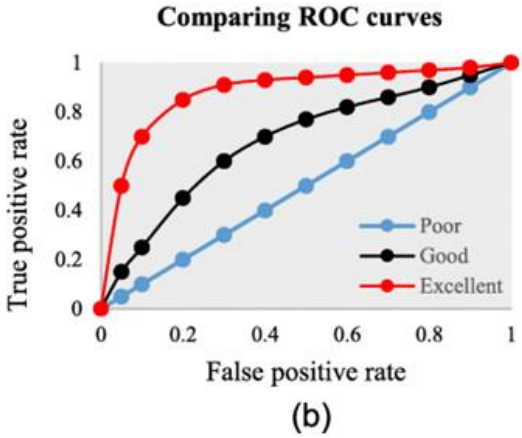


Fig 2. Depicts F1 Scores

Decision Tree

It is the extensive, forecast modelling tool that has applications crossing a number of diverse zones. In general, decision trees are constructed as an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used method for supervised learning. The aim is to build a prototype that predicts the worth of a target variable by learning straightforward decision tree instructions and it does not require any parameter setting, and therefore it is appropriate for discovery of the knowledge. The rules that decision tree follows are generally in the form of if-then-else statements. Decision trees performs classification without requiring much computation. Decision trees is capable to handle continuous as well as categorical variables.

Support Vector Machine Classifier

The occurrences of points in area is denoted by the SVM algorithm that are then plotted so that the classes are separated by strong gap. The goal is to determine the maximum-margin hyperplane which provides the greatest parting between the classes. The occurrences which is closest to the maximum-margin hyperplane are called support vectors. The vectors are chosen which are based on the part of the dataset that signifies the training set. Support vectors of two classes enable

the creation of two parallel hyperplanes. Therefore, larger the periphery between the two hyperplanes, better will be the generalization error of the classifier. SVMs are implemented in a unique way as compared with other machine learning algorithms.

Naive Bayes Classifier

The probability of an event occurring is rest on prior knowledge of circumstances that might be related to the event, focused by Naive Bayes. Naive Bayes is the most up-front and rapid classification algorithm, which is suitable for an enormous block of data. There are varied applications such as sentiment analysis, text categorization, spam filtering and recommender systems, where NB classifier is being used. Bayes theorem of probability is used for predicting the unknown classes. Naive Bayes is straightforward and easy to implement algorithm. Because of which, when the quantity of data is sparse it might out perform more complex models.

$P(H|E) = (P(E|H) * P(H)) / P(E)$

Where,

• P(H|E) the probability of hypothesis in which H gives the event E, a posterior probability.

P(E|H) given that the hypothesis H is true, when the probability of event is E.

• P(H) the probability of hypothesis where the H is true, a preceding probability.
• P(E) states the probability of the event that is been occurring.

Artificial Neural Network

The supervised learning is used by Artificial neural network which classifies the input information into the desired product. The artificial neurons consist with weighted interconnections that regulate the effect of the corresponding input signals, therefore neural network make use of supervised learning to categorize the load parameters of diabetes. Firstly, in classification of diabetes neural network gathers and identifies the data as an input to the network. With defined training dataset the network is trained and choose the training algorithm. ANN is tested after the training process to acquire the reaction of the network which states whether the disease is classified magnificently or not.

VIII. MACHINE LEARNING MATRIX

Precision:

The precision can be defined as the number of TP upon the number of TP '+' number of FP. False positives are cases where the model is incorrectly tagged as positive that are actually negative.

Precision = TP-----→TP + FP

Recall

The recall can be defined as the number of true TP separated by the TP '+' FN.
Recall = TP.
TP + FN

F1-Score

F1 is a function of Precision and Recall. F1 Score is needed when you want to seek a balance between Precision and Recall and there is an uneven class distribution (more number of actual negatives).

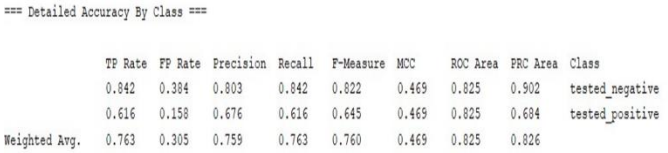
F1=2* Precision*Recall
Precision*Recall

IX. EXPERIEMNTAL SETUP

Result Analysis

After taking the input dataset the model will predict the data by applying the ML algorithms and provide

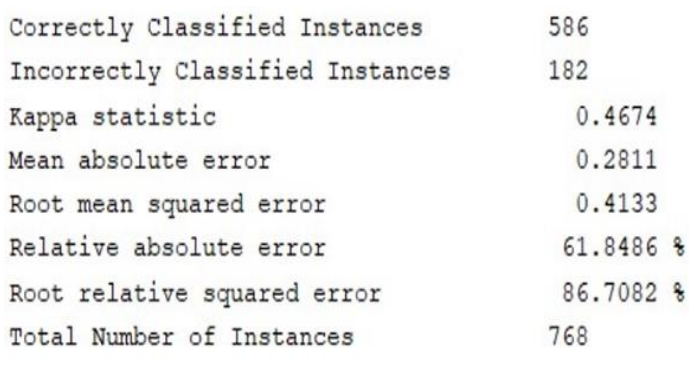
the best result in the form of comparison between to predict the best accuracy to treat diabetes.



=== Detailed Accuracy By Class ===

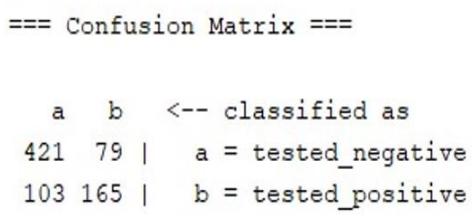
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.842	0.384	0.803	0.842	0.822	0.469	0.825	0.902	tested_negative
	0.616	0.158	0.676	0.616	0.645	0.469	0.825	0.684	tested_positive
Weighted Avg.	0.763	0.305	0.759	0.763	0.760	0.469	0.825	0.826	

Fig 3. Detailed accuracy by class



Correctly Classified Instances	586
Incorrectly Classified Instances	182
Kappa statistic	0.4674
Mean absolute error	0.2811
Root mean squared error	0.4133
Relative absolute error	61.8486 %
Root relative squared error	86.7082 %
Total Number of Instances	768

Fig 4. Evaluation on Training set



=== Confusion Matrix ===

a	b	<-- classified as
421	79	a = tested_negative
103	165	b = tested_positive

Fig 5. Confusion matrix

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor was successfully implemented using the grails framework. This system gives a user-friendly environment and easy to use. As the system is based on the web application, the user can use this system from anywhere and at any time. In conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the hospital data.

By using Naive Bayes Classifier

=== Summary ===

Correlation coefficient	0.4015
Mean absolute error	11.5919
Root mean squared error	17.7161
Relative absolute error	91.7122 %
Root relative squared error	91.588 %
Total Number of Instances	768

By using Support Vector

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient	0.2511
Mean absolute error	12.3643
Root mean squared error	18.7236
Relative absolute error	97.8235 %
Root relative squared error	96.797 %
Total Number of Instances	768

X. CONCLUSION

SVM is very good when we have no idea on the data. Even with unstructured and semi structured data like text, images and trees SVM algorithm works well. The drawback of the SVM algorithm is that to achieve the best classification results for any given problem, several key parameters are needed to be set correctly. Decision tree: It is easy to understand and rule decision tree. Un stability is there in decision tree, that is bulky change can be seen by minor modification in the data structure of the optimal decision tree. They are often relatively inaccurate. Naive Bayes: It is robust, handles the missing values by ignoring probability estimation calculation. Sensitive to how inputs are prepared. Prone bias when increase the number of trainings dataset. ANN: Gives good prediction and easy to implement. Difficult with dealing with big data with complex model. Require huge processing time.

XI. REFERENCES

[1]. Agusti Solanas, Fran Casino, Edgar Batista and Robert Rallo," Trends and Challenges in Smart Healthcare Research: A Journey from Data to Wisdom", IEEE 2017.

[2]. Fuad Rahman, "Application of Big-Data in Healthcare Analytics –Prospects and Challenges", IEEE 2017.

[3]. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime and Thomas Noel, "Big Data in healthcare: Challenges and Opportunities", IEEE 2015.

[4]. Pranjul Yadav, Michael Steinbach, Vipin Kumar and Gyorgy Simon, "Mining Electronic Health Records (EHRs): A Survey", ACM Computing Surveys, Vol. 50, No. 6, Article 85. Publication date: January 2018.

[5]. Weider D. Yu, Jaspal Singh Gill, Maulin Dalal, Piyush Jha and Sajan Shah, "Big Data Approach in Healthcare used for Intelligent Design", 2016 IEEE International Conference on Big Data (Big Data)

[6]. Rohan Bhardwaj, Ankita R. Nambiar and Debojyoti Dutta, "A Study of Machine Learning in Healthcare", 2017 IEEE 41st Annual Computer Software and Applications Conference.

[7]. Niharika G. Maity, Dr. Sreerupa Das, "Machine Learning for Improved Diagnosis and Prognosis in Healthcare", IEEE 2017.

[8]. Emrana Kabir Hashi, Md. Shahid Uz Zaman , Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, IEEE.

[9]. Md. Golam Rabiul Alam, Rim Haw, Sung Soo Kim, Md. Abul Kalam Azad, Sarder Fakhrul Abedin, Choong Seon Hong, "EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO.6, DECEMBER 2016.

[10]. Parisa Naraei, Abdolreza Abhari, Alireza Sadeghian, "Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data", FTC 2016 Future Technologies Conference 2016, 6-7 December 2016 — San Francisco, United States, IEEE 2016.

[11]. Dr.Neeraj Bhargava, Sonia Dayma, Abishek Kumar, Pramod Singh, "An Approach for Classification using Simple CART Algorithm in Weka", 2017 11 th International Conference on Intelligent Systems and Control (ISCO)" , IEEE 2017.

[12]. Dhomse Kanchan B., Mr. Mahale Kishor M," Study of Machine Learning Algorithms for Special Disease Prediction using Principal Component Analysis", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE 2016.

Cite this Article

Dr. Gangolu Yedukondalu, Prof. M. Srinivasa Rao, Prof. J. Anand Chandulal, "A Novel Classification Model to Predict Diabetes Using Machine Learning and Data Mining Techniques", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 1, pp. 287-293, January-February 2021.

Journal URL : https://ijsrst.com/IJSRST2294103