Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES

Priyanka Sonar
Research Scholar
Mumbai University
Mumbai, India
prnksonar@gmail.com

Prof. K. JayaMalini
Research Scholar
Bharath University
Chennai, India
malini1301@gmail.com

***Abstract*—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 85%, for Naive Bayes 77% and 77.3% 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

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.

1.1 Types of Diabetes

1) 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.

2)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.

3) 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.

II. Literature Review

Veena Vijayan V. And Anjali C has discussed, the diabetes 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].

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].

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].

III Proposed System

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.

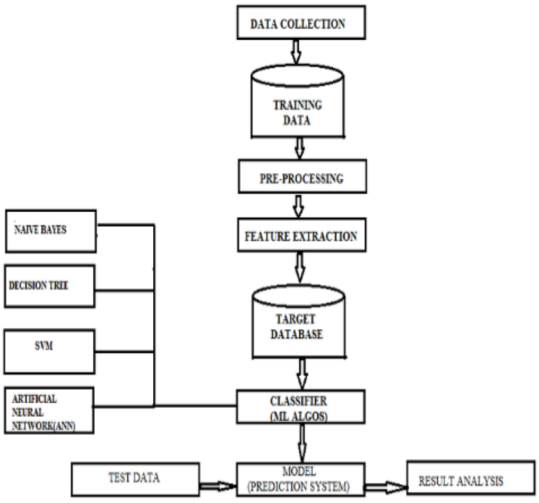


Fig 1.1 Block diagram of diabetes prediction system

3.1 Dataset Collection

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

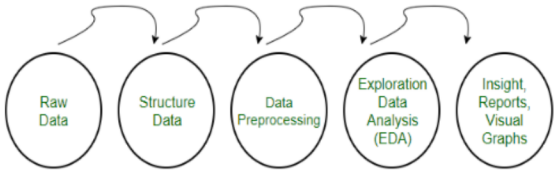
Percentage split option is provided for training and testing. Out of 768 instances 75 % is used for training and 25% is used for testing [1].

3.2 Training Data and Test Data

The training data set in Machine Learning is used to train the model for carrying out abundant actions. Detailed features are fetched from the training set to train the model. These structures are therefore combined into the prototype. In sentiment analysis, single words or sequences of consecutive words are taken from the tweets. Therefore, if the training set is labelled correctly, then the model will be able to acquire something from the features. So for testing the model such type of data is used to check whether it is responding correctly or not.

3.3 Pre-processing

Pre-processing refers to the transformations applied to our data before providing the data to the algorithm. Data Pre-processing technique is used to convert the raw data into an understandable data set. In other words, whenever the information is gathered from various sources it is collected in raw format that isn't possible for the analysis. Fig 1 Shown below data preprocessing.



3.4 Feature Extraction

Feature Extraction is used to transform the input information as the outcome of features. Attribute square measures are characteristic of input designs that facilitates in differentiating between the classes of input designs. In the algorithm if the input data is too huge for processing it will be suspected to be redundant as the repeat occurrence of images which are represented as pixels, which are changed into a condense set of attribute. Using the extracted feature instead of the complete initial data the chosen task can be achieved.

3.5 Target Database

The target database is the database to which the new changes are moved. For example, you install the certified Upgrade Source database, referred to as demo. Then you

produce a duplicate copy of your production database. You then copy the changed definitions from the Demo database into the Copy of Production. Here the Demo database is your source and the target is Copy of Production.

3.6 Machine Learning Algorithms Used:

3.6.1 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.

3.6.2 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.

3.6.3 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.

3.6.4 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.

3.7 Machine learning Matrix:

3.7.1 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.

$$\text{Precision} - \frac{TP}{TP + FP}$$

3.7.2 Recall

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

$$\text{Recall} - \frac{TP}{TP + FN}$$

3.7.3 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).

$$F1 - 2* \frac{\text{Precision*Recall}}{\text{Precision*Recall}}$$

3.8 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.

IV Implementation and Results

Table 1: Comparison between SVM, DTC, ANN and NBC algorithms.

	CLASSES	PRECISION	RECALL	F1 SCORE	SUPPORT	ACCURACY
Decision Tree Classifier	0	0.78	0.71	0.74	187	74
	1	0.45	0.54	0.49	82	49
Support Vector Classifier	0	0.70	1.00	0.82	187	82
	1	0.00	0.00	0.00	82	0
Gaussian Naive Bayes	0	0.82	0.79	0.80	187	80
	1	0.56	0.60	0.58	82	58
Artificial Neural Network	0	0.70	1.00	0.82	187	82
	1	0.00	0.00	0.00	82	0

Fig 4.1 Output of algorithms.

V Conclusion

SVM: Are 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. Unstability 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 training dataset. ANN: Gives good prediction and easy to implement. Difficult with dealing with big data with complex model. Require huge processing time.

VI References

1) Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum.
2) P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.
3) Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, , "Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.
4) Berina Alic, Lejla Gurbeta and Almir Badnjevic, "Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases", 2017 6th Mediterranean Conference On Embeded Computing (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO.
5) Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com
6) Rahul Joshi and Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017
7) Zhilbert Tafa and Nerxhivan Pervetica, "An Intelligent System for Diabetes Prediction", 4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro.
8) Sumi Alice Saji and Balachandran K, "Performance Analysis of Training Algorithms in Diabetes Prediction", International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.
9) Aakansha Rathore and Simran Chauhan, "Detecting and Predicting Diabetes Using Supervised Learning". International Journal of Advanced Research in Computer Science, Volume: 08, May-June 2017.
10) April Morton, Eman Marzban and Ayush Patel, "Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients,13th International Conference on Machine Learning and Applications",2014.
11) Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis". International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.

12) Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil "Diabetes Disease Prediction Using Data Mining". International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) 2016.