International Journal of Research
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

Comparision Study Of Classification Techniques For Diabetes Prediction

M. Madhavi
Department of Computer Science Engineering
Anurag Group of Institutions, Venkatapur, Ghatkesar, Hyderabad, Telangana 500038.
Email- mmadhavicse@cvsr.ac.in
G. Vinoothna
Department of Computer Science Engineering
Anurag Group of Institutions, Venkatapur, Ghatkesar, Hyderabad, Telangana 500038.
Email- 16h61a05k3@cvsr.ac.in
D. Rushalika
Department of Computer Science Engineering
Anurag Group of Institutions, Venkatapur, Ghatkesar, Hyderabad, Telangana 500038.
Email- 16h61a05j7@cvsr.ac.in
B. Sai Kiran
Department of Computer Science Engineering
Anurag Group of Institutions, Venkatapur, Ghatkesar, Hyderabad, Telangana 500038.
Email- 17h65a0512@cvsr.ac.in

ABSTRACT: Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases. With advances in technology, machine learning helps to predict the occurrence of diabetes in a subject. But with new algorithms and methodologies being developed every day, it becomes difficult to choose one. We carry out a comparison study of classification algorithms that are widely used for prediction and establish which algorithm gives more accurate results. 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 – Decision Tree, ANN, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

Diabetes is a situation which causes deficiency due to less amount of insulin in the blood. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. Some of the symptoms are frequent urination, feeling thirsty, increased hunger. If it is not medicated, it will lead to deadly diseases. ML resolves the real-world difficulties by providing learning capability to workstation without supplementary program writing.

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 hypoglycaemic agent not responding properly to the hypoglycaemic agent created. Various information mining algorithms

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

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. EXPERIMENTAL

The proposed method is to predict the occurrence of diabetes in a person. It is done using multiple classification algorithms like Decision tree, Support Vector Machine, Naïve Bayes and Artificial Neural Network for accuracy authentication on a dataset consisting of 768 instances and 9 features. 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 them to predict the best accuracy to treat diabetes.
The architecture of the proposed system is divided into three modules which are shown below in Fig 1.

2.1 Dataset collection

The dataset is taken from PIMA, National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of 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. Out of 768 instances, 70% is used for testing and 30% is used for training [1].

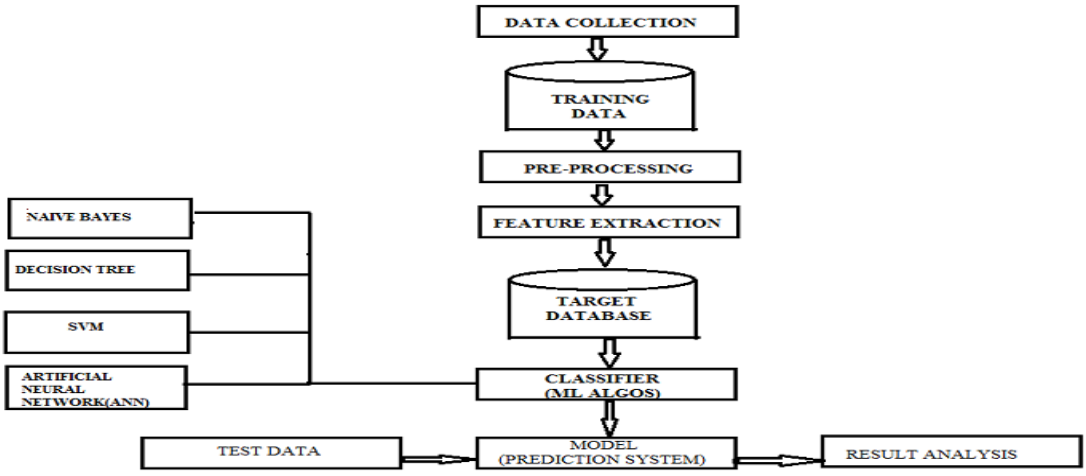

Fig 1. Architecture of the Proposed System

2.2 Pre-Processing

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

This 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. It is shown below in Fig 2.

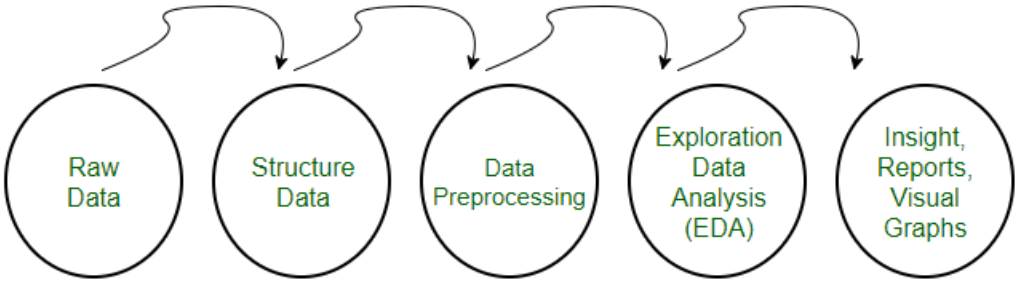

Fig 2. Data Pre-processing

2.3 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 attributes. Using the extracted feature instead of the complete initial data the chosen task can be achieved.

2.4 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 are Copy of Production.

2.5 Machine Learning Algorithms Used:

2.5.1 Decision Tree

It is a supervised learning method, which is used for solving classification problems. Decision tree is a technique which iteratively breaks the given dataset into two or more sample data. The goal of the method is to predict the class value of the target variable. The decision tree will help to segregate the data set and builds the decision model to predict the unknown class labels. A decision tree can be constructed to both binary and continuous variables. Decision tree optimally finds the root node based upon the highest entropy value. This gives decision tree an advantage of choosing the most consistent hypothesis among the training dataset. An input to the decision tree is a dataset, consisting of several attributes and instances values and output will be the decision model. Issues faced while building a decision model are selecting the splitting attribute, splits, stopping criteria, pruning, training sample, quality and quantity, the order of splits etc.

2.5.2 Support Vector Machine

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

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.

2.5.3 Naïve Bayes Classifier

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with unbalancing problems and missing values. Naive Bayes [24] is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P(C|X) can be calculated from P(C), P(X) and P(X|C).

Therefore, P(C|X) = (P(X|C) P(C))/P(X)

Where,

P(C|X) = target class's posterior probability.

P(X|C) = predictor class's probability.

P(C) = class C's probability being true.

P(X) = predictor's prior probability.

2.5.4 Artificial Neural Network

Artificial Neural network is typically organized in layers. Layers are being made up of many interconnected 'nodes' which contain an 'activation function'. A neural network may contain the following 3 layers: input, hidden, and output layer. The hidden layer consists of units that transform the input layer to the output layer. The output of one neuron works as the input for another layer. ANN detects complex patterns and learns on the basis of these patterns. The human brain contains billions of neurons. These cells are connected to other cells by axons and a single neuron is called as perceptron. Input is accepted by dendrites which is taken as stimuli. Similarly, the ANN is composed of multiple nodes that are connected with each other. The connection between units is represented by a weight. The objective of ANN is to convert input into significant output. Input is the combination of a set of input values that are associated with the weight vector, where the weight can be negative or positive. There is a function that sums the weight and maps the result to the output, such as y = The influence of a unit depends on the weighting; where the input signal of neurons meets is called the synapse. ANN works for both supervised and unsupervised learning techniques. Supervised learning was used in our study because the output is given to the model. In supervised learning, both input and output are known. After processing, the actual output with compared with required outputs. Errors are then back propagated to the system for adjustment. During training, the data is processed many times, so that the network can adjust the weights and refine them.

III. RESULTS AND DISCUSSIONS

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

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.

3.1 Machine Learning Metrics

3.1.1 Precision

Precision (also known as positive predictive value) is the fraction of relevant instances among the retrieved instances. 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.

$$\text{Precision} = TP/ (TP + FP)$$

3.1.2 Recall

Recall (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved. The recall can be defined as the number of true TP separated by the TP '+' FN.

$$\text{Recall} = TP/ (TP + FN)$$

3.1.3 F1-Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both 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 (a greater number of actual negatives).

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

3.1.4 Support

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

	Class	Precision	recall	f1-Score	support	accuracy
NaiveBayes	0	0.79	0.76	0.78	151	0.77
NaiveBayes	1	0.58	0.62	0.60	80	0.62
svm	0	0.75	0.85	0.80	151	0.81
svm	1	0.63	0.48	0.54	80	0.61
DecisionTree	0	0.79	0.72	0.75	151	0.82
DecisionTree	1	0.55	0.64	0.59	80	0.61
ANN	0	0.75	0.79	0.77	151	0.83
ANN	1	0.56	0.50	0.53	80	0.48

Fig 3. Output- Comparison between SVM, DTC, ANN and NBC algorithms.

IV. 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. Instability 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.

.

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. Uma Tejaswi, “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

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

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.
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.