International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 7, Issue 3
March 2018

Decision Support System for a Chronic Disease-Diabetes

Monisha. A
Loyola-ICAM College of Engineering and Technology
Nungambakkam, Chennai, Tamil Nadu, India
S.Shalin Christina
Loyola-ICAM College of Engineering and Technology
Nungambakkam, Chennai, Tamil Nadu, India
Nirmala Santiago
Loyola-ICAM College of Engineering and Technology
Nungambakkam, Chennai, Tamil Nadu, India

ABSTRACT

Diabetes is a chronic disease that occurs when the blood glucose level differs from the normal level. In recent years, the number of diabetic patients has increased drastically. Worldwide, it afflicts more than 422 million people. A number of computerized systems were designed using different classifiers for predicting and diagnosing diabetes.In this decision support system, the machine learning algorithms used are Naive Bayes statistical modeling, Logistic Regression and Extreme Gradient Boosting on the Pima Indian Diabetes dataset that consists of 768 patients.The accuracy of decision support system model using Naive Bayes Classifier, Logistic Regression and Extreme Gradient Boosting algorithm is found out. The accuracy obtained for Extreme Gradient Boosting is 81% which is greater compared to that of Naive Bayes Classifier and Logistic Regression.

Keywords : Decision Support System,Naive Bayes Classifier,Logistic Regression,Extreme Gradient Boosting

INTRODUCTION

The development of the computer-based methods would provide an efficient support to decision making in healthcare.A decision support system (DSS) is an information system that provides support indecision-making activities. DSSs help people make decisions about unstructured problems. Decision support systems can be either fully computerized or human-powered, or a combination of both. Machine learning is the area of artificial intelligence that uses statistical analyses.Based on the given dataset of diabetes,it can help in patient classification or probability prediction. The main strength is the ability of the algorithms to learn from data and to use that knowledge for later predictions and decisions.The idea of this decision support system is to predict the probability of a person having diabetes. The effectiveness of the decision support system is recognized by its accuracy. So the main aim of building a decision support system is to predict and diagnose a particular disease with maximum degree of accuracy.The Pima Indian Diabetes Dataset which is considered here consists of 768 instances and 8 attributes. All patients here are females at least 21 years old of Pima Indian heritage. Attributes considered are number of times pregnant,glucose,blood pressure (mm Hg),skin thickness (mm),insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function, age (years), class variable -outcome(0 or 1).

RELATED WORKS

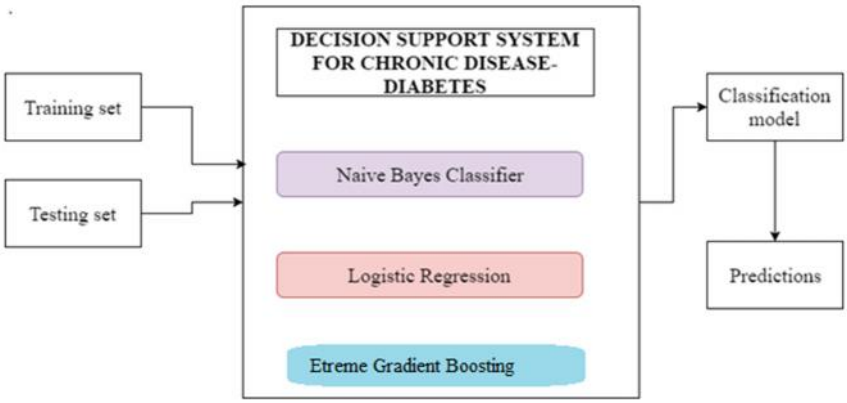
[1] proposes a decision support system that uses AdaBoost algorithm with Decision Stump as base classifier for classification. Additionally Support Vector Machine, Naive Bayes and Decision Tree are also implemented as base classifiers for AdaBoost algorithm for accuracy verification. [1] The accuracy obtained for AdaBoost algorithm with decision stump as base classifier is 80.72%.[2] review the benefits of different preprocessing techniques for decision support systems for predicting diabetes which are based on Support

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 7, Issue 3
March 2018

Vector Machine (SVM), Naive Bayes classifier and Decision Tree. The preprocessing methods focused on this study are Principal Component Analysis and Discretization. In [3] the algorithm works on the principle of maximum classifier rate and minimum rules. [4] Naive Bayes classifiers often work much better in many complex real-world situations. Here independent variables are considered for the purpose of prediction. In [5] Chi-Squared Test of independence was performed followed by application of the CART(Classification And Regression Trees) machine learning algorithm on the data and finally using Cross-Validation, the bias in the results was removed. In [6], the development of a decision support system based on ant colony optimized neural network has been done which is hybrid of feature selection with ant colony neural network.

PROPOSED SYSTEM

The proposed system focuses on to predict the probability of a person having diabetes using classification algorithms of machine learning approach. The effectiveness of the decision support system is recognized by its accuracy. So the main aim of building a decision support system is to predict and diagnose a particular disease with maximum degree of accuracy. The decision support system is modelled by machine learning approach that uses Naive Bayes classifier algorithm, Logistic Regression and Extreme Gradient boosting and the accuracy of these algorithms are predicted .Here no particular preprocessing is done for the dataset.80% of dataset has taken for training purpose and the remaining 20% of the dataset has taken for testing. The language used for the implementation is R programming.



DATASET-PIMA INDIAN DIABETES

The Pima Indians diabetes dataset is a publicly available dataset downloaded from UCI machine learning repository.The dataset comprises of 8 attributes, 768 instances and 1 binary class attribute.

1) Source of the dataset: UCI Machine learning repository

2) About the dataset: This dataset is a collection of healthinformation from Pima Indian women population of 21 yearsand above in the region of Arizona and Phoenix

3) Attributes: 9

a) Number of times pregnant

b) Glucose concentration

c) Diastolic blood pressure mm Hg

d) Skin thickness (mm)

e) Insulin (mu U/ml)

f) Body mass index Kgm-2

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 7, Issue 3
March 2018

g) Diabetes pedigree function

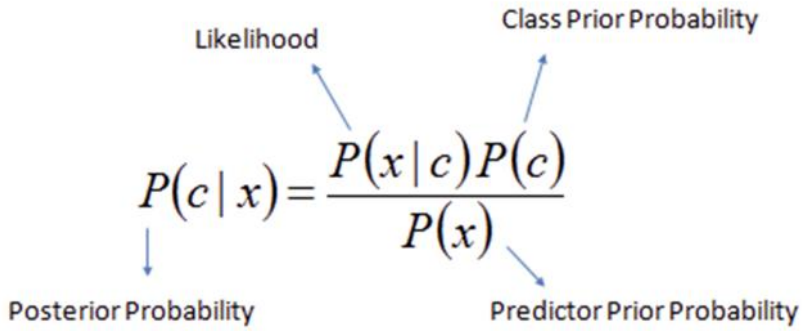
h) Age in years

i) Outcome- Binary Class variable

IMPLEMENTATION METHODS

1.NAIVE BAYES CLASSIFIER

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. i.e. every pair of features being classified is independent of each other. The assumptions made by Naive Bayes works well in practice.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

In the PIMA diabetes dataset ,80 % of the dataset has taken for training purpose and the remaining 20 % has taken for testing purpose.The naive bayes model is built with the training dataset and then predictions are done for the testing dataset. The accuracy is found out with the help of confusion matrix which is used to describe the performance of a classification model.

2.LOGISTIC REGRESSION

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a binary variable (in which there are only two possible outcomes).The outcome could be either 1(TRUE) or 0(FALSE).

The sigmoid function is guaranteed to produce an output between 0 and 1.

S(z)=1/1+e−z

S(z) = output between 0 and 1 (probability estimate)

z = input to the function (algorithm's prediction e.g. mx + b)

e = base of natural log.

This decision support system has taken 80% of PIMA dataset for training and the remaining 20% for testing. The attributes which are considered as less significant are ignored and the logistic model is built .The confusion matrix is determined by fixing the threshold rate as 0.5 .The threshold value is chosen in such a way that it produces the maximum true positive rate and minimum false positive rate.

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 7, Issue 3
March 2018

3.EXTREME GRADIENT BOOSTING

Extreme Gradient Boosting is a supervised learning algorithm that helps us to predict the outcome by combining the estimates of a set of simpler and weaker models. It has the capacity to compute parallely on a single machine. It is faster than the gradient boosting algorithm. It attempts to learn iteratively from the previously built weaker models and tries to minimize the error rate.

The PIMA diabetes dataset is splitted into training and testing dataset with the split ratio 0.8 and 0.2.The extreme gradient boosting model is built iteratively for 45 rounds with the learning rate of 0.4. The model is built iteratively in such a way it could find minimum error rate. The iteration which provides the minimum error rate of all the iterations is considered and the prediction is made. The accuracy of the extreme gradient boosting algorithm is found out.

RESULTS AND DISCUSSION

The accuracy of the machine learning algorithm is calculated with the following formula

Accuracy = (TN + TP)/(TN+TP+FN+FP)

(or)

= (Number of correct assessments)/(Number of all assessments)

TN-True Negatives

TP-True Positives

FN-False Negatives

FP-False Positives

The Naive Bayes Classifier gives 76% as accuracy. The accuracy of the logistic regression is 73% then after by removing the unnecessary attributes, and by fixing the threshold value to 0.4 which is determined using ROC curve, the accuracy has increased to 78%. The extreme gradient boosting algorithm which is an extension of gradient boosting algorithm provides 81% of accuracy.The accuracy obtained using extreme gradient boosting algorithm is greater compared to both naive bayes and logistic regression.

Table 1. Performance Metrics

Algorithm	Accuracy
Naive Bayes Classifier	76%
Logistic Regression	78%
Extreme Gradient Boosting	81%

Fig1 represents the plot for the prediction of probabilities of the test dataset using Naive Bayes classifier

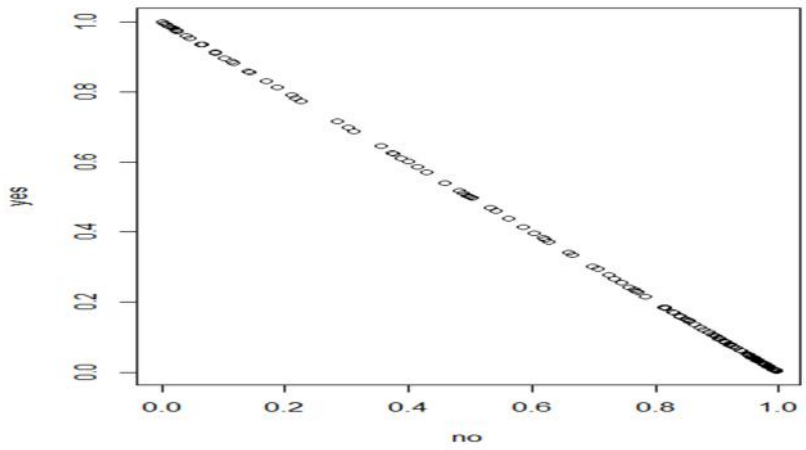


Fig 1:Naive Bayes plot

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 7, Issue 3
March 2018

Fig 2 Determination of the plot on true positive rate and false positive rate using ROC

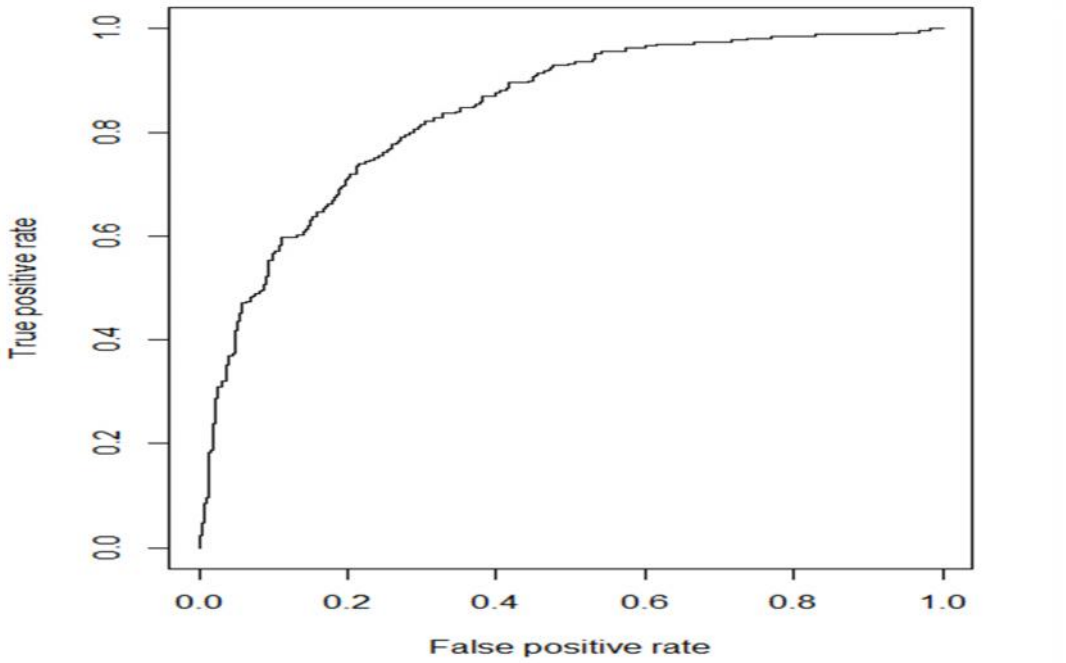


Fig 2: ROC curve

Fig 3 represents the error rate for the train dataset and test dataset for the specified number of iterations

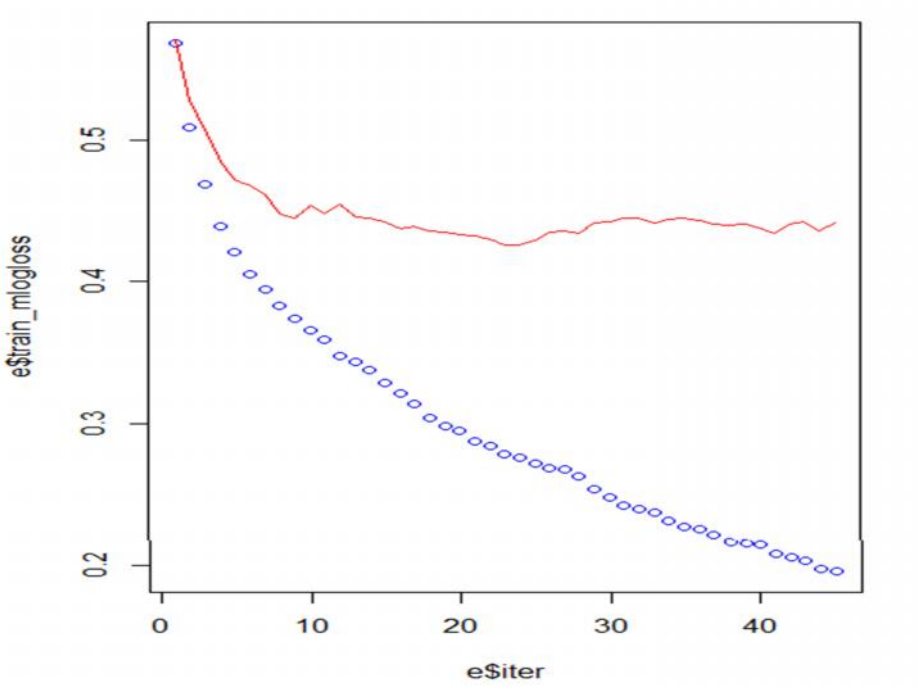


Fig 3: train error rate-blue color
test error rate-red color

Fig 4 represents the importance of attributes

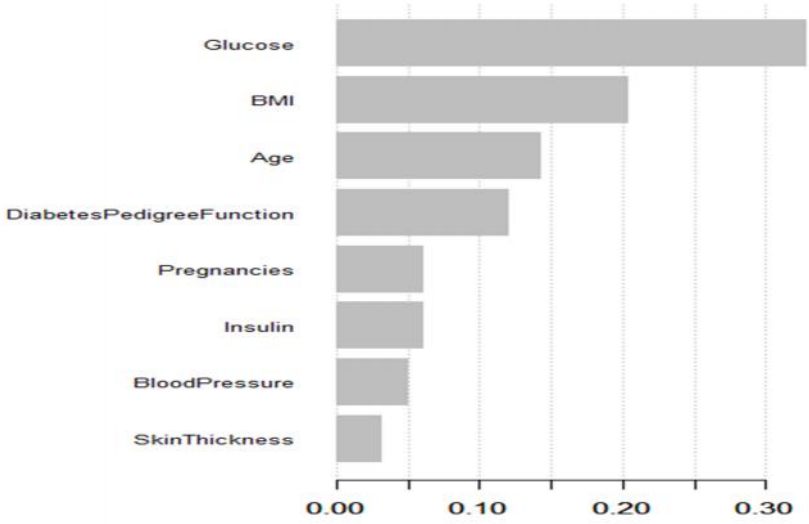


Fig 4:Importance of attributes

CONCLUSION AND FUTURE WORK

Machine learning approach is useful in disease diagnosis. The decision support system is implemented using machine learning algorithms such as Naive Bayes , Logistic regression and Extreme Gradient Boosting .The accuracy obtained for Extreme Gradient Boosting is 81% with the minimum error rate. This decision support system could be proposed to all other chronic diseases. This would result in the reliable prediction of diseases. The accuracy could be increased further with the implementation of other boosting algorithms.

REFERENCES

[1] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." *Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in*. IEEE, 2015.

[2] Vijayan, V. Veena, and C. Anjali. "Decision support systems for predicting diabetes mellitus—A Review." *Communication Technologies (GCCT), 2015 Global Conference on*. IEEE, 2015.

[3] Vaishali, R., et al. "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset." *Computing Networking and Informatics (ICCNI), 2017 International Conference on*. IEEE, 2017.

[4] Pattekari, Shadab Adam, and Asma Parveen. "Prediction system for heart disease using Naïve Bayes." *International Journal of Advanced Computer and Mathematical Sciences*3.3 (2012): 290-294.

[5] Anand, Ayush, and Divya Shakti. "Prediction of diabetes based on personal lifestyle indicators." *Next generation computing technologies (NGCT), 2015 1st international conference on*. IEEE, 2015.

[6] Kumar, Manoj, Anubha Sharma, and Sonali Agarwal. "Clinical decision support system for diabetes disease diagnosis using optimized neural network." *Engineering and Systems (SCES), 2014 Students Conference on*. IEEE, 2014.