Diabetics Prediction using Gradient Boosted Classifier

J. Beschi Raja, R. Anitha, R. Sujatha, V. Roopa, S. Sam Peter

Abstract : Diabetes is one of the most common disease for both adults and children. Machine Learning Techniques helps to identify the disease in earlier stage to prevent it. This work presents an effectiveness of Gradient Boosted Classifier which is unfocused in earlier existing works. It is compared with two machine learning algorithms such as Neural Networks, Radom Forest employed on benchmark Standard UCI Pima Indian Dataset. The models created are evaluated by standard measures such as AUC, Recall and Accuracy. As expected, Gradient boosted classifier outperforms other two classifiers in all performance aspects.

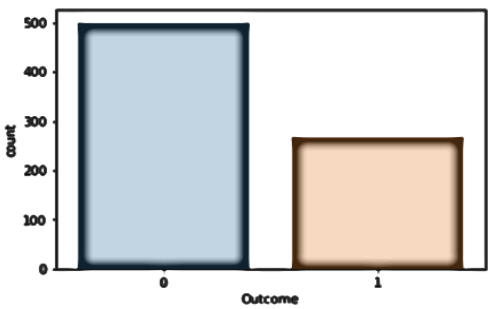
Keywords: Gradient Boosted Classifier, Pima Indian dataset, Diabetes, Evaluation measures.

I. INTRODUCTION

Diabetes disease is mainly caused due to lack of insulin content in human blood [1]. Some symptoms of diabetes are frequent urination, thirst and hunger. The seriousness will increase if the disease is untreated at initial stage which further leads to stroke, disorder of all parts [2]. The main duty of pancreas is to secrete insulin to the human body. If it fails to secrete enough amount of insulin, Diabetes are occurred.The three types of diabetics are Type I, Type II and gestational diabetes. The cause of Type I is inadequate amount of insulin creation by pancreas. Type II is caused by malfunction of body cells due to less insulin secretion and Gestational diabetes are caused to pregnant women due to high sugar level. The recent study depicts that more than 18% of women are affected by Type III diabetes in their pregnancy times [3]. Data analysis helps medical field researches for knowledge extraction from dataset leads to take appropriate decisions that makes a good progress for health care industry [4]. Diabetes prediction using Machine Learning methods has been a striking research area due to its massive importance. Data mining methods used to predict the decisions by pattern recognitions, cluster analysis and classification techniques [14]. In existing works, the researches focus on supervised methods than unsupervised methods for Diabetes prediction. we discussed some prominent recent and various existing research works related to diabetics' and other medical disease prediction using ML Techniques. A study was done using SOM and NN, PCA for prediction of diabetic analysis [5]. In another work, a multi model was designed by kumar et al. employing SVM method for attribute selection and eliminating unwanted attributes in data sample [6]. Similarly, hybrid work designed using NB and K-Means for clustering the data samples into groups [7]. Similar study was examined [8] diabetic analysis employing MLP, NB, Decision Tree. They proved that NB has good efficient performance compared to other methods on pima dataset. similarly, a model is developed for Heart disease using NN machine learning techniques [9]. The foundation of this article constructed as follows. We elucidated pre-processing process and dataset explanation in section 2. Next, ecosystem and framework implemented in this paper are highlighted in section 3. Section 4 carries various machine learning methods employed in this work. Finally, section 5 and 6 gives results and conclusion of the paper.

A. Pre-processing

The first step in Pre-Processing process is analysing for missing values. It is observed that nearly five features such as blood pressure, skin thickness, glucose, insulin and BMI have value as 0, which indicates the missing values in dataset. the missing values should be treated to improve the efficiency of the model. So, we replaced missing values into average values of each and every column. Next the dataset is introduced to Spearman method to evaluate the correlation between the values. We noticed that majority of the data belongs to healthy people and less amount belongs to people who suffers from diabetics. After applying Spearman method for attribute correlation, it is observed that age, pregnancy, insulin, glucose, skin thickness and BMI are correlated to target variable [15]. In that glucose and insulin are highly correlated to the target.



Revised Manuscript Received on October 05, 2019.
***** Correspondence Author
J.Beschi Raja, Assistant Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore, TamilNadu, India,
R.Anitha,*, Assistant Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore, TamilNadu, India,
R.Sujatha,, Assistant Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore, TamilNadu, India,
V.Roopa,, Assistant Professor, Department of IT, Sri Krishna College of Technology, Coimbatore, TamilNadu, India
S.Sam Peter, Assistant Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore, TamilNadu, India

Fig.1.2. Count Plot for Diabetes Vs Healthy ones

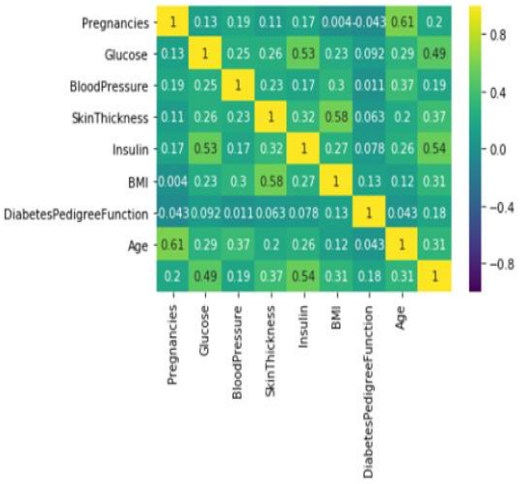


Fig.1.3. Correlation between the Attributes

a. Dataset

Pima Indian data sample for this analysis were collected from UCI database. It is one the standard 768 instances and eight various features. We observed 500 records were non diabetics and only 268 were diabetics patients [13]. The records were collected from women staying near phoenix location. The various features of pima Indian sample are depicted in fig.2.1.

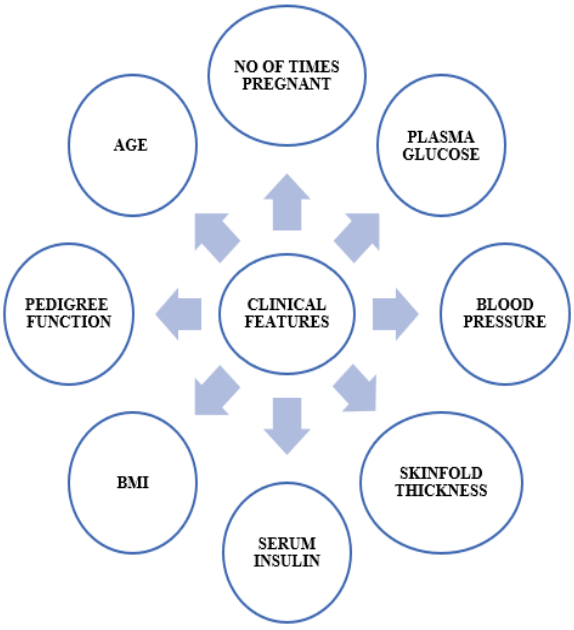


Fig.2.1. various features in pima Indian data sample

II. FRAMEWORK PROPOSED

The framework in this work consists of three phases. The first phase employs pre-processing stage which deals with missing value analysis. We used Spearman Technique for feature selection and correlation. Next stage consists of classification process. Here we employed three benchmark algorithms such as Random Forest, Neural Networks and Gradient Boosting. The last phase takes the prediction phase. We examined the dataset with these three methods and predicted Gradient Boosting algorithm performs well among all. For evaluating the model, three standard metrics were used such as ROC, Recall and Accuracy.

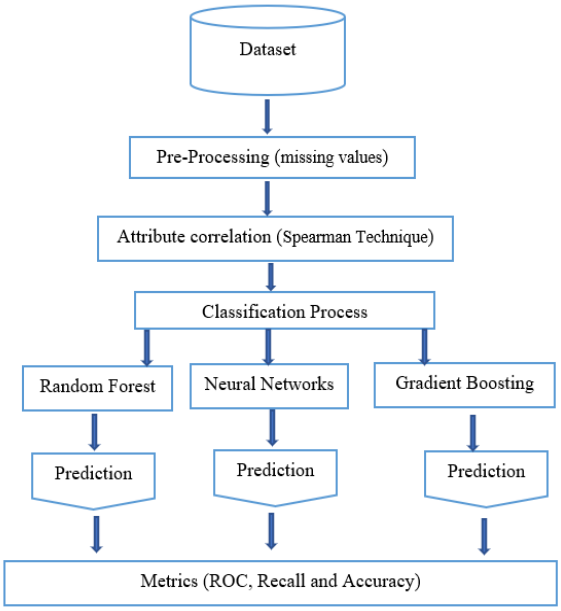


Fig.3.1. Systematic Task Flow of Work Employed

III. MACHINE LEARNING METHODS

a. Neural Networks

Multilayer perceptron is a method for designing feed-forward ANN. It integrates various perceptron to design a non-linear boundary [12]. The main role of perceptron holds on activation function, weights, processor and bias. The three main layers in Multilayer perceptron is final layer, middle layer and initial layer. The input data are feed into initial layer and perceptron parameters are passed along with layer. By modifying the parameters, the higher accuracy is obtained. The activation function is accelerated in middle intermediate layer to predict the target.

The steps for Neural network algorithm

1. For each and every layer in NN model, nodes are generated automatically,
2. Initially, target value is set as zero
3. Update the target value with weights from previous layer estimation
4. Target value+=weight* value obtained from other nodes in layer
5. Target value rate = sig (Target value rate)

b. Random Forest

Random forest method creates a forest with various number of trees. It is a supervised algorithm that is robust and produce high accuracy [11]. This algorithm is more useful for both regression and classification work. Handling missing values as well as overfitting are well performed using random forest. It creates a lot of subsets with a random value to make decisions. Random forest algorithm works as a large collection of decorrelated decision trees.

With all decision trees to create a ranking of

classifiers and make the class prediction.

The steps for random forest algorithm

1. Randomly pick "U" attributes from overall "Q" features
2. Consider U<<Q
3. Evaluate the node "X" by best fragment position.
4. Daughter nodes are obtained by dividing by best fragmentation.
5. Iterate the process till desired number of nodes are obtained.
6. Finally, forest by designed by "N" tress

c. Gradient booting

Gradient boosting method creates step-wise process and increments the algorithm on basis of loss function [10]. The errors are detected and rectified to improve the accuracy. Generally, boosting checks models which decrease the loss function obtained from trained samples. From these calculations the errors are measured and analysed for optimal prediction of results. Loss function calculates the range of detected rate which compares with desired target. Onward stepwise process is most popular method for updating different with various attributes. The accuracy is optimized by reducing loss function and adding base learners at all stages.

Steps for gradient Boosting method

1. Consider a sample of target values as P
2. Estimate the error in target values
3. Update and adjust the weights to reduce error M
4. P[x]=p[x]+alpha M[x]
5. Model Learners are analysed and calculated by loss function F
6. Repeat steps till desired & target result P

IV. EXPERIMENTAL ANALYSIS & RESULTS

The first model was created using Neural Network and cross validation was employed to optimize the regulation parameter. They number of hidden layers are fixed by trial and error process to increase the accuracy of model. Multi-Layer perceptron is used for designing Neural Network classifier. We fixed solver as 'lbfgs', activation function as relu, learning rate as 'adaptive', hidden layers as 10,10,10 and random state as 9. It is observed that this model works good only for healthy patients but it struggles for diabetes patients. Next, we designed random forest model using "n_estimators". We see that, the same classification results happened in Neural Networks i.e providing good results for healthy samples alone. Finally, we created a Gradient Boosting Classifier with Max_depth parameter. It is observed that gradient boosting models outperforms other two models by producing high scores in all criteria.

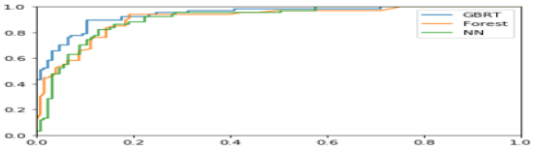


Fig.5.1. ROC Comparison

Table.1.2. Experimental Results of Machine Learning methods used in this work

	Neural Network	Random Forest	Gradient Boosted Classifier
Recall	0.701	0.656	0.761
AUC	0.907	0.907	0.942
Accuracy	0.838	0.822	0.897

V. CONCLUSION

Diabetes is one of the most common disease for humans today. Data mining methods are very helpful for detecting it in early stage. This work presents machine learning based diabetes prediction using classifier methods. The comparative experimental analysis reveals that Gradient Boosting Classifier outperforms Random forest and Neural networks. The future scope can be extended to figure out the features impacting by hybrid of feature selection methods with classifiers for real life large dataset.

REFERENCES

1. S. Siddiqui, Depression in type 2 diabetes mellitus—a brief review. Diabetes Metab. Synd.Clin. Res. Rev. 8(1), 62–65 (2014)
2. K. Rajesh, V. Sangeetha, Application of data mining methods and techniques for diabetes diagnosis. Int. J. Eng. Innov. Technol. (IJEIT) 2(3) (2012)
3. S. Sarma Kattamuri, Predictive modeling with SAS enterprise miner: practical solutions for business applications (SAS Institute, 2013)
4. I. Yoo et al., Data mining in healthcare and biomedicine: a survey of the literature. J. Med. Syst. 36(4), 2431–2448 (2012)
5. MehrbakhshNilashia Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset", Fuzzy Information and Engineering, Volume 9, Issue 3, September 2017, Pages 345-357.
6. Kumar, Binit, et al. "Retinal neuroprotective effects of quercetin in streptozotocin-induced diabetic rats." *Experimental Eye Research* 125 (2014): 193-202.
7. Pandeeswari, L., Rajeswari, K.: K-means clustering and Naïve Bayes classifier for categorization of diabetes patients. Int. J. Innov. Sci. Eng. Technol. (IJISET) **2**(1) (2015)
8. Koklu, M., Unal, Y.: Analysis of a population of diabetic patients databases with classifiers. World Acad. Sci. Eng. Technol. **7**(8) (2013)
9. S. Palaniappan, R. Awang, BIntelligent Heart Disease Prediction System Using Data Mining Techniques^. IJCSNS.2008;Vol. 8, No. 8.
10. Friedman, Jerome H. "Stochastic gradient boosting." *Computational statistics & data analysis* 38.4 (2002): 367-378.
11. Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.
12. Peter Salamon. "Neural network ensembles." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 10 (1990): 993-1001.
13. Hayashi, Y. and Yukita, S., 2016. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. *Informatics in Medicine Unlocked*, *2*, pp.92-104.
14. Schultz, Matthew G., et al. "Data mining methods for detection of new malicious executables." *Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001*. IEEE, 2000.
15. Hamilton, Martin A., Rosemarie C. Russo, and Robert V. Thurston. "Trimmed Spearman-Karber method for estimating median lethal concentrations in toxicity bioassays." *Environmental Science & Technology* 11.7 (1977): 714-719.