International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Improved Diabetic Data Analytic Model for Complication Prediction

K.Vidhya, R.Shanmugalakshmi

ABSTRACT--- Data Analytic model examines large datasets and reveals the hidden information like useful patterns and their correlations in it. Especially in the healthcare analytics accurate analysis and correct prediction would be much more important for prevention of further complications. The prediction here is based on the prior treatment details and readmission possibility based on the health condition from the Diabetes dataset. Upon aiming to analyze and predict the possibility of diabetes complication, the diabetic data is preprocessed and analyzed using Decision Tree Algorithm. As per execution the accuracy of the algorithm is only 55% only. We improved the accuracy value to 84% by the application improved AdaBoost based ID3 algorithm. This enhanced system shows the improved result for accuracy precision, recall and F-measure.

Keywords - BigData, Healthcare, Diabetes, Decision Tree, Classifier, AdaBoost, Accura

1. INTRODUCTION

Diabetes Mellitus is a very common disease in the society. Nowadays the prevalence of diabetes keep on increasing due to the changes in the life style of human. In early days the type 2 diabetes due to improper usage of insulin in our body would be at the middle age and old age. But nowadays even the child gets diabetes [1].The cost for medical care changes an economic burden for the patients and it also affects their employment in many ways. There are a large number of complications in Diabetics that can affect heart, kidney, nerve disorder, eyes and feet. Its risk can be reduced only upon early detection and right management, so there is a need to detect the presence of diabetes in an early stage [2]. The patient's health factors are collected and preserved as big data [3]. The Big Data Analytic model is the most important and widely spreading tool that helps in analyzing and predicting diabetes. The various parameters are taken to test every individual's symptoms and the disease that can be caused due to those symptoms. The dataset which is downloaded from Kaggle has over 1 lakh tuples and 28 attributes that is imported into the algorithm and trained for providing a better accuracy.

Big Data is the term for a collection of datasets that is so large and complex as it became difficult by using traditional Database tools to store and process, its challenges also include capture, curation, search, transfer, storage, sharing, analysis and visualization [4]. It is an advanced form of analytics which involves, elements that are predictive and applications that are complex and also performs what-if analysis and statistical algorithms that are powered by high-performance analytics systems. It enables predictive modelers, data scientists, statisticians and other analytics professionals to analyze the growing volumes of data in many forms. Raw data are taken from both internal and external sources and it may be structured or semi-structured data. Classification strategies are widely used in the medical field for classifying data into different classes according to some constraints to compare every individual classifier. Diabetes is not only affected by height, weight, gene and insulin but the major reason is considered to be the sugar concentration level in Blood [5]. If diabetes remains untreated many complications, will arise and the early identification is the only way to stay away from those complications [6]. Data Mining and Machine learning algorithms gain strength due to the capability of managing a large amount of data and to combine data from several different sources and also integrating the background information in the study. Several other classification techniques and ensemble classifier are proposed in the field of data mining for diabetes diagnosis [7]. As compared to single classifiers, ensemble classifier accuracy is considered to be more precise in performance and also in prediction [8]. Their structure is also more flexible and different alternatives are chosen to provide the best solution in predicting high performance and greater accuracy to determine the chances for a person to get diabetes.

2. LITERATURE REVIEW

The accuracy of prediction is highly important in medical field. The effective prediction technique must overcome the problem of erroneous data and missing value.Edgar Acuna and Caroline Rodriguez proposed the different techniques for the treatment of missing values and its effect in the classifier accuracy by calculating cross validation errors, median imputation, mean imputation, case deletion and KNN imputation [9]. K.G.Li, Mohd Ibrahim Shapiai, Asrul Adam presented the feature scaling model for EEG human concentration using particle swarm optimization, digital filters, wavelet transforms [10].Atul Kumar Pandey, Prabhat Pandey, K.L.Jaiswal represents the model intelligent heart disease prediction system built with the help of data mining techniques like Decision tree, Naive Bayes and Neural Network [11].

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
K.Vidhya, Assistant Professor(Sr.G), Department of CSE, KPR Institute of Engineering and Technology, Coimbatore. T.N, India. (E-mail: vidhya.k@kpriet.ac.in), Phone: +919865511224.
Dr.R.Shanmugalakshmi, Professor & Head, Department of Electrical and Electronics Engineering, Government College of Technology, Coimbatore. T.N, India. (E-mail: drshanmi@gct.ac.in)

KaiqiZou, Hongzhi Yu, Wenming Sun, Feng xin Liu established the fraud detection model with ID3 decision tree and also described the application fields of fraud model [12].Zhang H& Zhou, R shows that the classification accuracy of ID3 decision tree with higher when compared to other classification algorithms and it also achieves the structure of decision tree for improving the efficiency of the algorithm [13]. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha proposes the information gain for splitting the attribute with the highest information gain [14].Chen Jin, Luo De-lin, &Mu Fen-xiang shows that the decision tree is the important for both inductive method and data mining, as it takes a sample training dataset as example for deriving the tree[15]. Rashmi Saini and S.K. Ghosh reviews the different ensemble especially useful for remote sensing and satellite image processing [16].

SabaBashir, Usman Qamar, Farhan Hassan Khan "An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles“ proposed the decision trees were trained and performance is increased to classify unlabeled diabetes instance using Adaboost [17]**.** Yoav Freund Robert, E. Schapire proposed the boosting algorithm called Adaboost which extends the communication between the boosting algorithm and the weak learner[18].Saba Bashir, Usman Qamar, Farhan Hassan Khan, M.Younus Javed shows ensemble classifiers like C4.5,CART,ID3 and the decision tress is trained and the result is improved using Adaboost algorithm[19].Yusuf Engin Tetik,Bülent Bolat discussed diabetes using the pedestrian from the images by calculating the weight,height and age of the person [20].Von Kirby P. German, Bobby D. Gerardo and Ruji P.Medina discussed about enhancing the Adaboost algorithm when there is numerous amount of data and to improve the accuracy of the data [21].

Wen Zhu, Nancy Zeng, Ning Wang computes the sensitivity, specificity, accuracy and the confidence interval and ROC for the improved data sets [22].Dr. Achuthsankar S.Nair, Aswathi B.L discussed about the sensitivity, specificity and the relation between them and accuracy is the dependable on both specificity and sensitivity [23].Chunxue Wu, Chong Luo, Naixue Xiong, Wei Zhang and Tai-Hoon Kim discussed about greedy deep learning method for medical disease analysis the highest accuracy level is achieved and the ROC curves are defined [24].Francesco Mercaldoa, Vittoria Nardoneb, Antonella Santone deals with diabetes affected persons and their classification using various algorithms and the f-measure is calculated [25].

3. PROPOSED WORK AND METHODOLOGY

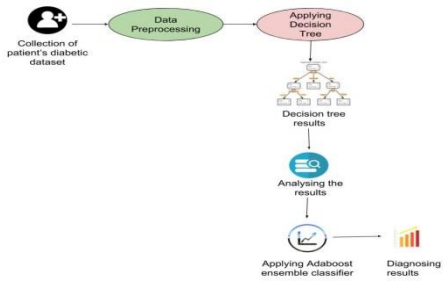

Figure 1. Proposed System Methodology

The Decision tree based data analytic system analyses the data from Kaggle diabetes dataset .The data set involves various parameters like patient number, race, gender, age, weight, Admission type, admission_id, insulin level etc. From these the proposed ID3 algorithm will analyze and identify the possibility of the patients getting readmission to the hospital based on the level of their insulin. The above Figure 1 represents the methodology of the proposed system. The proposed system consists of Data collection, Data preprocessing, implementation of Decision tree, Adaboost classifier to enhance the accuracy of outcome of diagnosis .Each process is expressed as follows.

3.1 Data Preprocessing

Real world data are generally incomplete due to the lack of attribute values or values that contain only aggregate data, inconsistent or some certain behavior lacking or too many errors. A method for resolving such issues is Data preprocessing . The process involves the following steps :

1. Importing the necessary libraries and the Dataset.
2. Checking out the missing values.
3. Categorical values are viewed.
4. The Dataset is split into training and testing sets.
5. Perform Feature Scaling

Once the library files and dataset are included , the missing values are checked .Because the inaccurate data will provide an inaccurate result.There are many ways to handle the missing values .

1. One way is to handles null values in the data. It will either delete a particular column having more than 75% null values or row if it is having null for a particular feature .This is only applicable to estimate when there are enough samples in the dataset but removing of data will lead to inaccurate results.
2. Another way calculates the mean, median, mode values of the feature and replace it with those missing values .This is merely an approximation and it adds variance to the dataset.Data loss can be negotiated as it yields the better results than the removal of rows and columns using the above mentioned method .

Machine learning models are based on some mathematical equations and if we do not remove the categorical data it would cause some problems as our equations need only numbers so they are converted into numerical values . Dummy variables can be used as that takes the value 0 or 1 to indicate the presence or absence of some categorical effect that may affect the outcome or sometimes, KNN imputation method can also be used .In this method the missing values are imputed by the given number of attributes similar to those that of missing and the similarities are determined using a distance function . Next, the data sets are split into two separate sets one is Training set and other is Testing set . As the algorithms we use generally learn from the data to make predictions the data test is split into 70:30 or 80:20 ratio where 70% data is trained and tested against 30% of data .It varies accordingly to the shape and size of the dataset .

The final step is Feature Scaling .One can normalize or standardize the data as it is the method of limiting the range of the variables so that they can be compared on the common grounds.

3.2 Decision Tree(ID3)

Decision tree algorithms are used for both classification and regression problem. It is often mimic to the human level of thinking that makes understanding and data interpretation easier. A Decision tree is a tree where each node represents an attribute, link represents the rule and each leaf represents an outcome that will either be categorical or continuous value. The tree is created for the entire data and there will be a single outcome at every leaf (or minimises the error in every leaf).The proposed system uses ID3 algorithm. Iterative Dichotomiser 3(ID3) algorithm iteratively divides attributes into two groups consisting of most dominant attribute on one side and others construct a tree . The most dominant attribute can be found by calculating Entropy and Information Gain and it is put as a decision node in the tree. The next most dominant attribute is also found in the same manner and this continues until reaching a decision for that branch .The best attribute is the one with highest Information gain in ID3. In order to define information gain we start by defining a measure used in information theory called Entropy.

Entropy

Let S is the training set contains positive and negative examples, then the entropy of S relative to this classification is:

Entropy H(S) [13] is the amount of uncertainty in the set S.

$$H(S) = \sum_{c \in C} -p(c)log_2 p(c)$$

p(c) is the proportion of S belonging to class I.
c is the set of classes.
∑ is over c.
H(S) is a measure of the amount of uncertainty in the set S.

From the random of 20 samples taken in our dataset, the readmission variable has 12 No classes and 8 Yes classes and the entropy is calculated for the readmission variable is given by,

Entropy= (12/20, 8/20)

=-12/20log (12/20)-8/20log (8/20)

=-3/5log (3/5)-2/5log (2/5)

=0.2922

Information Gain

Information Gain IG(A) [14]is the measure of difference between in entropy from before to after the set S splits on an attribute A.

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,{\displaystyle \mathrm {H} (S)}H(S) - Entropy of

set S {\displaystyle S}

Target Readmission			
Gender	**No**	**Yes**	**Total**
Female	6	3	9
Male	4	7	11
Total	10	10	20

- T-The subsets created from splitting set S by attribute A
- p(t)– The proportion of the number of elements in t to the number of elements in set S
- H(t)– Entropy of subset

Sample entropy and information gain calculation for the outcome variable (readmission) for our diabetic dataset is as shown in table 1 below.

Table 1: Details for Entropy calculation

1).Entropy (Parent)

P0 = p(target=No) = Proportion of Target values No P0 = 0.5
P1 = p(target=Yes) = Proportion of Target values Yes P1 = 0.5
Entropy (Parent) = -P0 log(P0) -P1 log (P1) = 0. 5 + 0.5 = 1

2).Entropy (Gender=Male)

Count of Target value Yes:3 Count of Target value No: 6

P0 = p(target=Yes) = 0.33 P1 = p(target=No) = 0.67
Entropy (Gender=Male)
=-0.33log(0.33) + -0.67log(0.67)
= 0.1589+ 0.1165== 0.2754

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

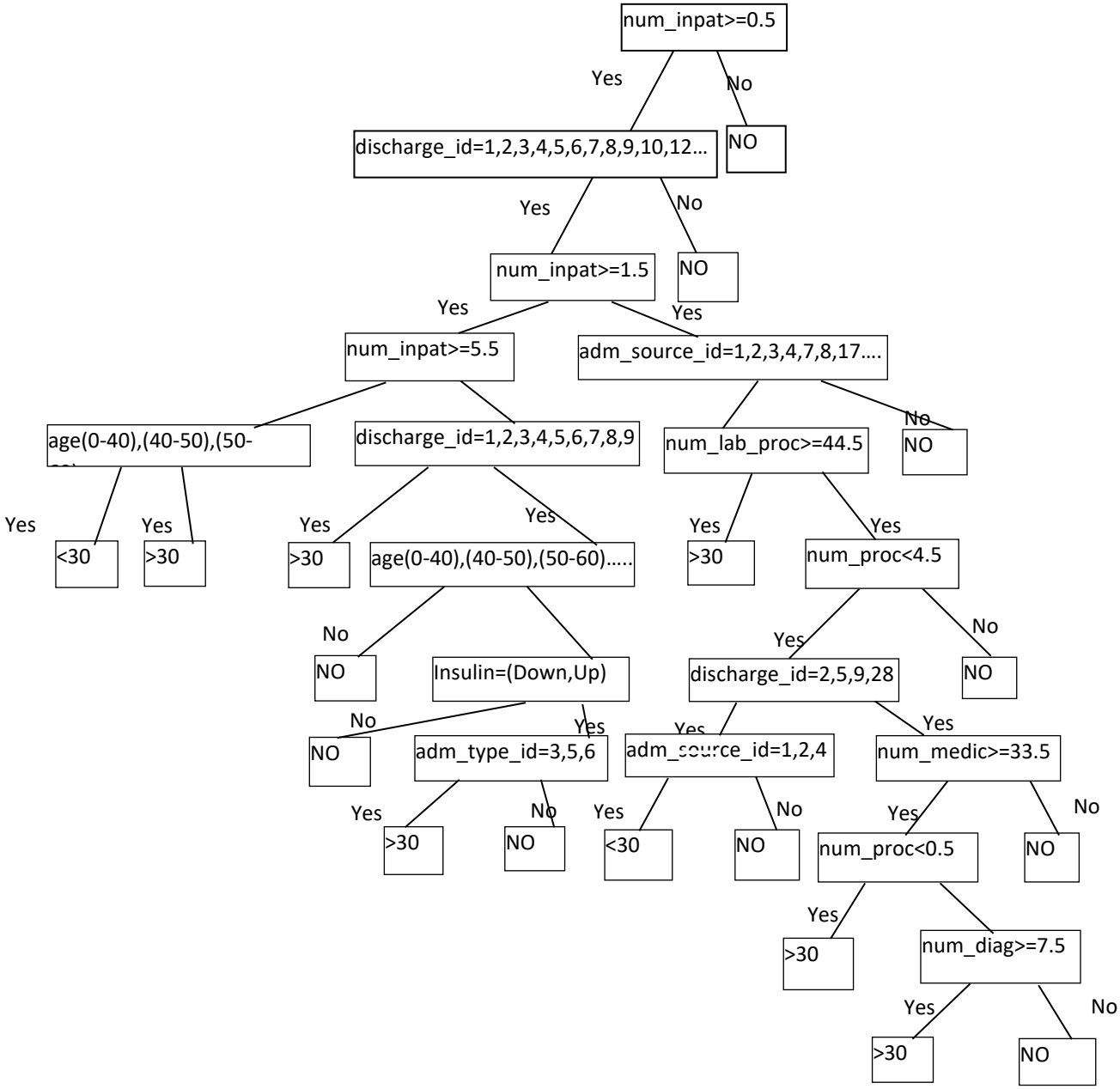

Figure.2 Decision Tree Result

3).Entropy (Gender=Female)

Count of Target value Yes:4 Count of Target value No: 7

P0 = p(target=Yes) = 0.364 P1 = p(target=No) = 0.636

Entropy (Gender=Female)

= -0.364log (0.364) + -0.636log (0.636)

= 0.1597+0.1250

=0.2847

The attributes used for the construction of decision tree in Figure.2 includes the number of inpatients, admin source id, age, discharge id, number of lab procedures, admission type id, admission source id, number of medications, number of diagnosis. Based on the information gain values calculated and with the help of rule induction operator, the following rules are extracted, If Number of inpatients is < 0.5 the patient is healthy. Else If number of inpatients is >=0.5 then discharge id is checked.If discharge id =1,2,3,4,5,6,7,8,9,10,12,17,27 and if number of inpatients>=1.5, if inpatient >=5.5 and the age lies within (0-40), (40-50), (50-60) the patient will be readmitted within 30 days otherwise, they will be readmitted after 30 days. If number of inpatients is <5.5 then patients discharge id is checked, if the value lies between 1-8, they will be readmitted after 30 days. Otherwise if age range is between (0-60) then they will not be readmitted else if insulin value is down or up then the patient will not be readmitted, else if based on admission type id has values like 3,5 or 6 then they will be readmitted after 30 days otherwise they will not be readmitted. Similarly, information gain is calculated for every other feature and decision tree is constructed from the values obtained. From the above the accuracy level is calculated based on the results. Here True Negative, True Positive, FalsePositive,

False Negative are used to calculate the percentage of accuracy, sensitivity and specificity. The most common associated terms of a Binary classification test are specificity, sensitivity, accuracy and they statistically measure the performance. In a binary classification, a given data set is divided into two categories on the basis of whether they have common properties or not by identifying their significance and in a binary classification test, as the name itself conveys, it needs two datastes. In general, of these two categories, how well the test predicts one category indicates Sensitivity and how well the test predicts the other category measures the Specificity. Whereas the accuracy is expected to measure how well the test predicts both categories. Accuracy is derived from Sensitivity and Specificity.

$$\text{Accuracy} = \frac{\text{Sensitivity}+\text{Specificity}}{2} \quad [24]$$

1) Accuracy- It is defined as the ratio of correctly identified cases to the total number of test cases which is represented by

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad [24]$$

2) Recall- The ability of a classification model to find all relevant cases within a dataset

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad [24]$$

3) Precision- The ability of a classification model to identify only the relevant data points

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad [24]$$

4) The F1-Measureis a measure of a test's accuracy. This measure can be interpreted as a weighted average of the precision and recall: [24,25]

$$\text{F1-Measure} = \frac{2*\text{Sensitivity}}{\text{Sensitivity}+\text{Specificity}}$$

Based on the above parameters the accuracy value obtained is 55% only. In medical diagnosis the accuracy of prediction plays a vital role in the life of patients. So such an improved accuracy model is essential.

4. IMPLEMENTING ADABOOST FOR ACCURACY IMPROVEMENT

In Data mining there are well known classification algorithms, where each classifier performs predictions based on learning. An ensemble classifier is learned from more prediction models.The basic idea behind ensemble classifiers is to weigh several individual classifiers and then combine them to obtain the result which outperforms every individual classifier. The performance of classification and prediction accuracy of ensemble classifier is higher than single classifiers. The ensemble classifiers are Bagging, Adaboost, Majority Voting, Stacking, Bayesian Boosting. The proposed project focused on Adaboost ensemble classifiers for improving accuracy of the predicted results.

Adaboost is the popular ensemble algorithm and it performs boosting by iterative processing. It mainly focuses on the instances that are difficult to classify using other classification techniques. The level of focus depends upon the weight that is associated with instances during each iteration. Initially, all instances are assigned equal weight. In every iteration, the weight of misclassified instances is increased whereas weight is reduced for the instances which are correctly classified. The measure of overall accuracy of that individual classifier is calculated by an associated weight. By considering their weight and prediction class, the classifiers are then combined.Now the decision trees are trained and performance is boosted using Adaboost ensemble classifier to classify unlabeled diabetes instances. Mathematically, the Adaboost ensemble method can be written as ,

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t \,.\, M_t\,(x)\right) \quad [19]$$

where - $M_t$denotes the classification based on voting of all classifiersfor a particular instance a_t is weight.

The weighted average of the weak classifiers is used to make predictions . For an each new input instance, each weak learner calculates a predicted value as either +1.0 or -1.0. The predicted values are weighted and compared by each weak learner's stage value. A sum of the weighted predictions is taken as the resultant prediction of an ensemble model. If the sum of weights is positive, then the first class is predicted, if it is negative then the second class is predicted.

5. RESULT AND EVALUATION

The implementation involves confusion matrix to appraise the performance of the model for incidence of diabetes and three evaluated indices for accuracy, sensitivity and specificity. The classification results of Decision tree without and with Adaboost algorithm is depicted by a confusion matrix of M(2×2) for comparative analysis of different parameters. In this matrix, element M_{ij} shows the total cases of class i which is actually classified as class j. In the binary classification problem, confusion matrix provides the details of the following values:

Table 2 : Comparative Result Performance

Metrics	Decision Tree Prediction	Decision Tree with Adaboost
Accuracy	0.55	0.84
Precision	0.51	0.84
Recall	0.50	0.82
F-measure	0.51	0.83

Accuracy, Sensitivity, Specificity and FMeasure (F-M) are used for the performance evaluation of the ensemble technique.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

It has the best performance results as compared to the other similar prediction approaches. It has achieved the highest accuracy levels for the diabetes datasets whereas Sensitivity, Specificity and F-measure values are also very randomly chosen is classified.The ROC curve of decision tree without Adaboost algorithm implementation is as shown in Figure.3 below.

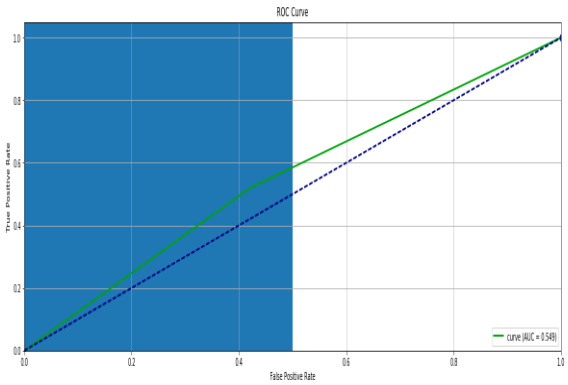

Figure 3. Decision Tree Result

Here the level is nearly 0.55 only.The improves ROC with Adaboost algorithm is as shown below Figure 4.

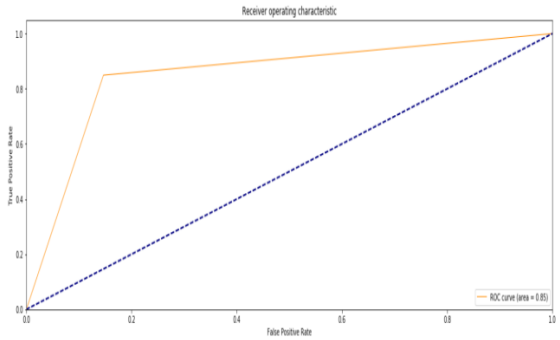

Figure 4. Improved Accuracy using Adaboost

In Figure 3,4 and in Figure 5, the X axis indicates the False Positive rate and the Y axis indicates the True Positive rate. The blue dotted line is the centre of the axis. The yellow line denotes the ROC curve which indicates the accuracy after using Adaboost. The result varies from 0.549(0.55) to 0.85 after using Adaboost.

Figure 5 shows the confusion matrix for the two classifiers proposed. The four basic parameters are used to predict that the possibility of readmission due to diabetes (insulin level based) is present or not. True Positive (TP) is the positive values that indicate that they have diabetes. True Negative (TN) is the negative values that indicate that they don't have diabetes. False Positive (FP) is the negative value but they are incorrectly described as positive. False Negative (FN) is the positive value that are mislabeled as negative.

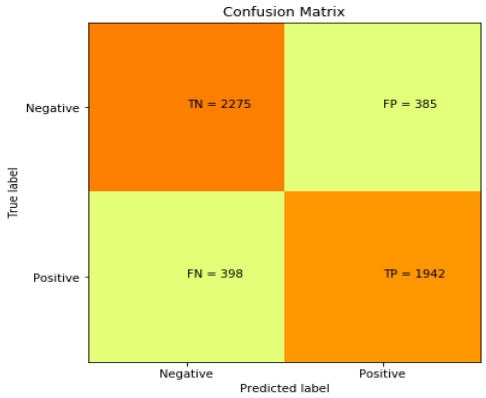

Figure 5. Confusion Matrix

The comparative prediction performance of Decision tree with and without Adaboost algorithm is as shown in Figure 6 below.It shows that the accuracy,precision,recall and F-measure of Decision Tree algorithm with Adaboost is nearly 85% which is the 30% improvement of early algorithm.

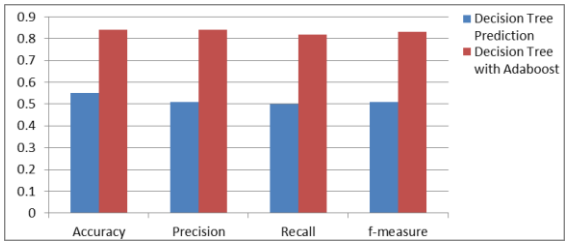

Figure 6:Comparative Performance with Adaboost

6. CONCLUSION AND FUTURE WORK

Thus the detection of diabetes related complication at its early stage is must for the society of diabetic .If we were able to detect diabetes even at the earlier stage then it is possible to avoid damage of vital organs of body like kidney, Heart ,Eyes and Nerves and especially readmission frequency of diabetic patients in the hospital.If the accuracy of prediction rate is high, the medical guidelines and respective treatment also accurate.So that the complexity of treatment also minimized early. In future similar ensemble approaches can be incorporated on other disease datasets such as Hypertension, Coronary heart disease and Dementia for early prediction and prevention. Furthermore, diverse individual techniques like Naïve Bayes, SVM and neural networks etc can be combined as base learners in ensemble framework.

REFERENCES

1. Harleen Kaur, Vinita Kumari," Predictive modelling and analytics for diabetes using a machine learning approach ",2018, Journal of Applied Computing and Informatics
2. Venkatesan, C., Karthigaikumar, P., & Satheeskumaran, S. (2018). Mobile cloud computing for ECG telemonitoring and real-time coronary heart disease risk detection. Biomedical Signal Processing and Control, 44, 138-145.

3. Vijayan, V.V., Anjali, C.,"Prediction and diagnosis of diabetes mellitus A machine learning approach." 2015, IEEE Recent Advances in Intelligent Computational Systems.
4. E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects", Nature, vol. 489, pp. 49-51, 2012.
5. Michael E. J. Lean and Lisa Te Morenga"Sugar and Type 2 diabetes"British Medical Bulletin, 2016
6. Satheeskumaran, S., and M. Sabrigiriraj. "A new LMS based noise removal and DWT based R-peak detection in ECG signal for biotelemetry applications." National Academy Science Letters 37, no. 4 (2014): 341-349.
7. K Prasanna Jyothi, Dr R SivaRanjani, Dr Tusar Kanti Mishra, S Ranjan Mishra"A Study of Classification Techniques of Data Mining Techniques in Health Related Research".
8. Rashmi Saini and S.K. Ghosh"Ensemble classifiers in remote sensing: A review"2017 International Conference on Computing, Communication and Automation (ICCCA)
9. Edgar Acu˜na1 and Caroline Rodriguez"The treatment of missing values and its effect in the classifier accuracy"
10. K. G. Li,Mohd Ibrahim Shapiai,Asrul Adam “Feature scaling for EEG human concentration using particle swarm optimization”
11. Atul Kumar Pandey1 ,Prabhat Pandey2 ,K.L. Jaiswal3 ,Ashish Kumar Sen4”A Heart Disease Prediction Model using Decision Tree”
12. Venkatesan, C., P. Karthigaikumar, Anand Paul, S. Satheeskumaran, and R. Kumar. "ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications." IEEE Access 6 (2018): 9767-9773.
13. Zhang, H., & Zhou, R. (2017)” The analysis and optimization of decision tree based on ID3 algorithm2017 9th International Conference on Modelling, Identification and Control (ICMIC).
14. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao “predicting students’ performance using id3 and c4.5 classification algorithms”
15. Chen Jin, Luo De-lin, & Mu Fen-xiang. (2009). Uopeng Meng “BigData visualization: Parallel coordinates using density approach ” Systems and Informatics (ICSAI), 2ND international conference,Shanghai 2014 , pp. 1056-1058.
16. CHUNXUE WU1 , (Member, IEEE), CHONG LUO 1 , (Member, IEEE), NAIXUE XIONG2 , (Senior Member, IEEE), WEI ZHANG3 , AND TAI-HOON KIM4 , (Member, IEEE)" A Greedy Deep Learning Method for Medical Disease Analysis " (2018)
17. Francesco Mercaldoa,∗ , Vittoria Nardoneb, Antonella Santone"Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques"International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France