A Comparative Analysis and Performance Evaluation of Classification Algorithm to Predict Diabetes Mellitus

[1]P.Pavithra, [2]Dr.P.B.Pankajavalli
[1]Research Scholar, [2]Assistant Professor
Department of Computer Science, Bharathiar University, Coimbatore, India.

Abstract: In this emerging world, diseases are increased besides treatment for those diseases also improved. At the same time, diagnosis of disease can lead more time consumption for experts because of its wide range of data. The data mining is the effective technology which can handle numerous amounts of data and extract hidden information from it. It is the emerging area in the field of medical diagnosis.Diabetes also referred as diabetes mellitus, defines group of metabolic diseases in which the person experiences high blood sugar level, because shorter insulin production or body cell's improper response to the insulin. Three major types of diabetes are type 1 diabetes, type 2diabetes and gestational diabetes. Among various data mining approaches, classification algorithms perform more efficient for the diagnosis of disease. In this paper, we have considered 10 classification algorithms like Naïve Bayes, Multilayer Perceptron, Adaboost, Logiboost, Decision Table, OneR, PART, J48, Random Forest, Random Tree, REPT and comparative analysis are performed. Finally, the best algorithm for diabetes diagnosis is keyed out based on its accuracy and error rate.

Keywords: Diabetes Mellitus, Naïve Bayes, Multilayer Perceptron, Adaboost, Logiboost, Decision Table, OneR, PART, J48, Random Forest, Random Tree, REPT

I. INTRODUCTION

Data mining is a knowledge discovery and analysis of large databases, providing unknown, hidden, meaningful, and useful patterns from major databases. Data mining algorithms such as neural networks, support vector machines, decision tree are used effectively in various medical fields. These algorithms can able to provide efficient solution for different disease diagnostic systems like as diabetes, heart disease, and breast cancer and so on. In healthcare society data mining plays a vital role for the detection of new trends which is helpful for the assembliesassociated with this field. Classification techniques have been widely used in the medical field for accurate classification of diseases (Mahmoud Heydari and Mehdi, 2016).

1.1 Data mining in healthcare

Electronic Health Records (EHR) is becoming more common among healthcare facilities. With increased access of large amount of patient data, healthcare providers can increase the efficiency and quality of their organizations by using data mining.In healthcare, data mining has implementedsuccessfully in the areas of analytical medicine, customer relationship management, fraud and abuse discovery, healthcare management. The purpose of data mining is,it mustidentifyuseful and understandable patterns by analyzing large sets of data. Data mining can be used to reduce costs through increasing efficiencies, save lives of more patients and improve patient quality of life (Mahmoud Heydari and Mehdi, 2016).

The future of healthcare may use data mining to reduce healthcare costs, identify treatment policy and best practices, calculate effectiveness, identify fraudulent insurance and medical claims, and toprogress the standard of patient care (EmreCelebi and Hassan, 2013). Healthcare providers bring into playwith data analysis and data mining to find best practices and the most effective treatments. These tools compare symptoms, causes, treatments and negative effects and then proceed to analyze the best action to prove most effective for a group of patients. Data mining tools provide support for healthcare groups to take better patient-related decisions (HasanTemurtas and NejatYumusak, 2009).

1.2 Diabetes Mellitus

1.2.1 Types of diabetes

Diabetes mellitus is a chronic, lifelong condition which affects body's ability to use the energy found in food. Type 1 diabetes, Type 2 diabetes and Gestational diabetes are the major types of diabetes. Type 1 diabetes also called as Insulin Dependent Diabetes Mellitus (IDDM), or juvenile-onset diabetes mellitus. In type 1diabetes, the pancreas suffers an autoimmune assault by the body, and reduces the capacity of making insulin. Irregular antibodies are found in majority of patients with type 1 diabetes. Antibodies are proteins in blood that are part of the body's immune system. The patient with type 1 diabetes must depend on insulin for survival. Type 2 diabetes is referred as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult-Onset Diabetes Mellitus (AODM). In type 2 diabetes, patients can still produce insulin, but do not support the patient body's needs, especially in the form of insulin resistance. In many cases, the pancreas produces larger than usual amounts of insulin. A major feature of type 2 diabetes is a lack of sensitivity to insulin by the body cells. Diabetes which arises temporarily during pregnancy, is referred as gestational diabetes (K. Saravananathan and T. Velmurugan, 2016).

II. METHODS

2.1 Data Source

The dataset is taken from UCI Machine learning repository. This larger database was obtained by the National Institutes of Diabetes Digestive and Kidney Diseases. It consists of two class variables which arerepresented by binary variable 0 and1. Here 1 represents the positive test diabetes and 0 represent the negative test for diabetes. The database contains 768 patients with 9 variables. There are 268 positive cases for class 1 and 500 cases inclass0. There are no missing values in the dataset.

2.2 Classification Algorithms

Classification is a machine learning technique, which is used to predict group relationship for data instances. It is used to analyze specified data set and takes each instance of it. To reduce the classification error and extract model, the instances are assigned to a particular class. The model defines important data classes from the given data set. Classification is a two-step process. In first step, the classification algorithm builds a model using class variable. Then the predefined test dataset is tested by the extracted model. It is to measure the performance and accuracy of the trained model. So classification assigns class label to the process from a data set whose class label is unknown (S.Neelamegam and Dr.E.Ramaraj, 2013).

Naive Bayes is a type of classifier which uses the Bayes Theorem. It predicts associative probabilities for each class such as the probability that gives evidence or data point belongs to a particular class. The class with the highest probability is measured as the most likely class. This is also known as Maximum A Posteriori (MAP) (S.Neelamegam and Dr.E.Ramaraj, 2013). Bayes theorem provides a way of calculating the posterior probability P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumesthe effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x)=P(x|c)P(c)/P(x) \qquad (2.1)$$

$$P(c|x)=P(x1|c)*P(x2|c)*\dots*P(xn|c)*P(c) \qquad (2.2)$$

Here, P(c|x) - posterior probability of class

P(c) - prior probability of class.

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

P(x) - prior probability of predictor (S.Neelamegam and Dr.E.Ramaraj, 2013).

Multilayer perceptrons are network of linear classifiers. It uses hidden layers to implement arbitrary decision boundaries. To create a network structure with number of perceptrons and connections, Weka contains a specified graphical interface. The feed-forward limitation creates a topological layering of the neurous in the network (S.Neelamegam and Dr.E.Ramaraj, 2013). **AdaBoost** is the short form of Adaptive Boosting and is a machine learning meta-algorithm. To improve the performance it can be applied in conjunction with other learning algorithms. Adaboost is sensitive to outliers and noisy data and derive the **LogitBoost** algorithm by considering the adaboost as generalized additive model and then applies the cost functional of logistic regression.

A **Decision Table** is used to represent conditional logic by creating list of tasks from the business level rules. Decision tables can be used when there is constant number of conditions that must be evaluated and assigned for the particular actions which can be used when the conditions are finally met (SamanHina and Anita Shaikh, 2017). **One Rule** is the expansion of OneR; it is a simple, accurate classification algorithm that generates one rule for each predictor in data, and then selects the rule with the smallest total error as its one rule. To create a rule, construct a frequency table for predictor (AiswaryaIyer and S. Jeyalatha, 2015).

PART stands for Projective Adaptive Resonance Theory. The algorithm produces set of rules called decision lists which are ordered. New information is compared with each rule in the list in turn, and the item is assign to the category of first matching rule; if no rule is successfully matches then a default is applied. PART builds a partial C4.5 decision tree during its each iteration and makes the best leaf into a rule. The algorithm is a combination of RIPPER and C4.5 rule learning.

The C4.5 algorithm for decision trees is implemented in Weka as a classifier called **J48**. To classify a new item, it needs to create a decision tree. The tree is based on the quality values of the training data. It identifies the attribute that differentiate the various instances most clearly. This feature can be able to classify the data instances and information gain (S.Neelamegam and Dr.E.Ramaraj, 2013).

Random Tree is a supervised classifier and it is an assembly learning algorithm that generates lots of individual learners. To construct a random set of data and make a decision tree, random tree employs bagging idea. **Random Forest** is a supervised learning algorithm and can be used for both regression and classification problems. It builds multiple decision trees and combines them together for more accurate results and it can develop a model in short period of time. In healthcare domain, random forest can be used to identify diseases by analyzing the combination of medicine and patient's history (SamanHinaand Anita Shaikh, 2017).

REPT is a Reducer Error Pruning Tree is a fast decision tree learning algorithm. It uses regression tree logic and generates multiple trees in altered iterations using information gain. It deals missing values by splitting the corresponding instances into pieces (AiswaryaIyer and S. Jeyalatha, 2015).

2.3 Performance Criteria

2.3.1 Statistical Analysis

Different classification algorithms are examined in this paper. Accuracy of each algorithm shows how the datasets are being classified. Recall and precision are the accuracy measures used for this work. Precision also known as predicted positive. Precision P uses True Positive (TP) and False Positive(FP) and the formula for precision is,

$$\text{Precision} = TP/\ (TP + FP) \qquad (2.3)$$

The proportion of positive cases were correctly identified which is known as True Positive Rate (TPR). It is calculated as,

Recall = TP/ (TP + FN).Here FN is False Negative Rate (2.4)

In this research work, three performance measures are used. Accuracy is calculated by an exact value. Equation (2.5) is used for the accuracy calculation and it uses TN= True Negative.

$$\text{Accuracy} = (TP + TN)\ /\ (TP + TN + FP + FN) \qquad (2.5)$$

$$\text{Sensitivity}=TP/(TP+FN) \qquad (2.6)$$

$$\text{Specificity}=TN/(TN+FP) \qquad (2.7)$$

The F-Measure can be computed as some average of the information retrieval precision and recall metrics.

$$F = 2^* \text{ Recall} * \text{Precision} / \text{Precision} + \text{Recall} \qquad (2.8)$$

TN –True Negative ; FN – False Negative (Sadri Sa'di and AmanjMaleki, 2015).

2.3.2 Terminologies of Test Statistics:

The following test measures are used to identify the error rate for employed classifiers. **Kappa Statistic** is a metric that compares Observed Accuracy with Expected Accuracy. **Mean Absolute Error** is the average of the absolute error between observed and forecasted value. **Root Mean Squared Error** measures differences between predicted model value and the actually observed value. **Relative Absolute Error** is the ratio of the absolute error of the measurement to the accepted measurement (Subhankar Mannaand Malathi, 2015).

III. RESULTS AND DISCUSSION

The classification accuracies were obtained from the diabetes dataset is present in Table 3.1.

Table 3.1 Performance Accuracy

Classifier/Instances	NB	MP(ANN)	AB M1	LB	DT	ONER	PART	J48	RF	RT	REPT
Correctly Classified	76.302	80.599	76.562	78.906	77.604	76.432	81.25	84.114	100	100	83.072
Incorrectly Classified	23.697	19.401	23.437	21.093	22.395	23.567	18.75	15.885	0	0	16.927

From Table 3.1, random forest, random tree and J48 approaches give a better classification results for the same dataset. The comparison between classification methods is done from the obtained results of correctly and incorrectly classified instances. Here correctly classified instance represents the percentage of records which are correctly identified for each classifier. The incorrectly classified instance is the metric to identify wrongly identified instances from the dataset. Table 3.2 gives the comparison of algorithms with respect to the error rate on employed classifiers.

Table 3.2 depicts the comparison of error rate for the algorithms used and random forest gives the better performance followed by random tree with least error rate. Decision table gives the least performance among other algorithms. Table 3.3 depicts detailed accuracy achieved by the algorithms. The different accuracy measures for each and every algorithm are obtained based on class variable.

Table 3.2 Error Reports

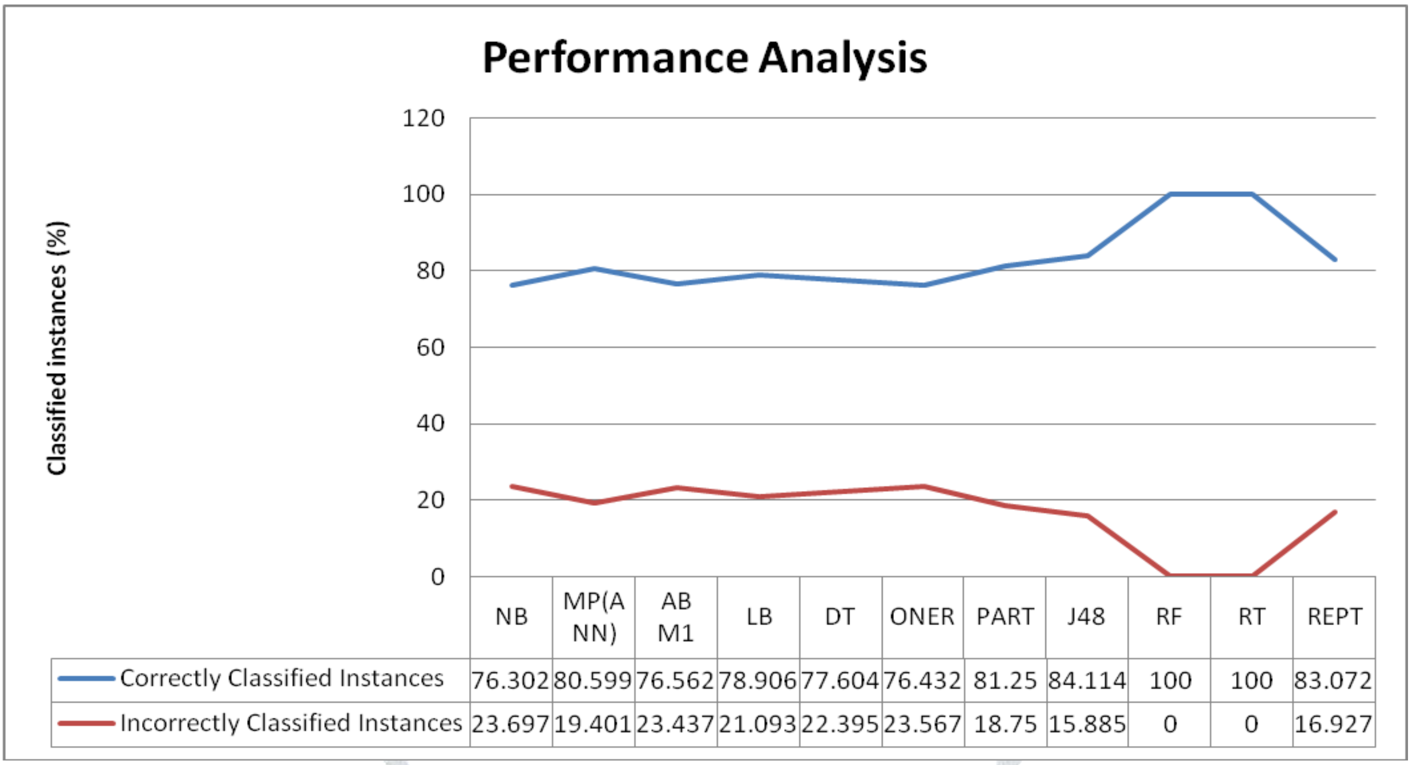
	NB	MP(ANN)	AB	LB	DT	ONER	PART	J48	RF	RT	REPT
KappaStatistic	0.4674	0.5904	0.4723	0.5067	0.4782	0.4484	0.6184	0.6319	1	1	0.6313
MAE	0.2811	0.2852	0.2956	0.2853	0.3223	0.2357	0.2466	0.2383	0.115	0	0.2498
RMSE	0.4133	0.3815	0.3922	0.3777	0.394	0.4855	0.3512	0.3452	0.1515	0	0.3534
RAE	61.8486%	62.75 %	65.0313%	62.7697 %	70.915 %	51.855 %	54.269%	52.433%	25.300 %	0 %	54.97 %
RRSE	86.7082%	80.049 %	82.2787%	79.251 %	82.6645%	101.851%	73.677%	72.420%	31.794 %	0 %	74.151%

Table 3.3 Detailed Accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
NB	0.763	0.305	0.759	0.763	0.760	0.469	0.825	0.826
MP(ANN)	0.806	0.191	0.819	0.806	0.809	0.596	0.872	0.874
AB M1	0.766	0.304	0.761	0.766	0.763	0.474	0.844	0.842
LB	0.789	0.312	0.786	0.789	0.780	0.518	0.863	0.862
DT	0.776	0.324	0.771	0.776	0.768	0.487	0.831	0.830
ONER	0.764	0.343	0.759	0.764	0.755	0.459	0.711	0.689
PART	0.813	0.144	0.845	0.813	0.817	0.639	0.888	0.858
J48	0.841	0.241	0.842	0.841	0.836	0.642	0.888	0.878
RF	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
RT	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
REPT	0.831	0.193	0.833	0.831	0.832	0.632	0.885	0.872

3.1 Performance Analysis

The performance of the specified algorithms are measured using classified error rate, precision, recall, true positive rate and false positive rate. In Figure 3.1.1 and Figure 3.1.2, the performance analyses of classified instances and error rate for specified algorithms are displayed.



	NB	MP(ANN)	AB M1	LB	DT	ONER	PART	J48	RF	RT	REPT
Correctly Classified Instances	76.302	80.599	76.562	78.906	77.604	76.432	81.25	84.114	100	100	83.072
Incorrectly Classified Instances	23.697	19.401	23.437	21.093	22.395	23.567	18.75	15.885	0	0	16.927

Figure 3.1.1 Performance Analysis for classified instances

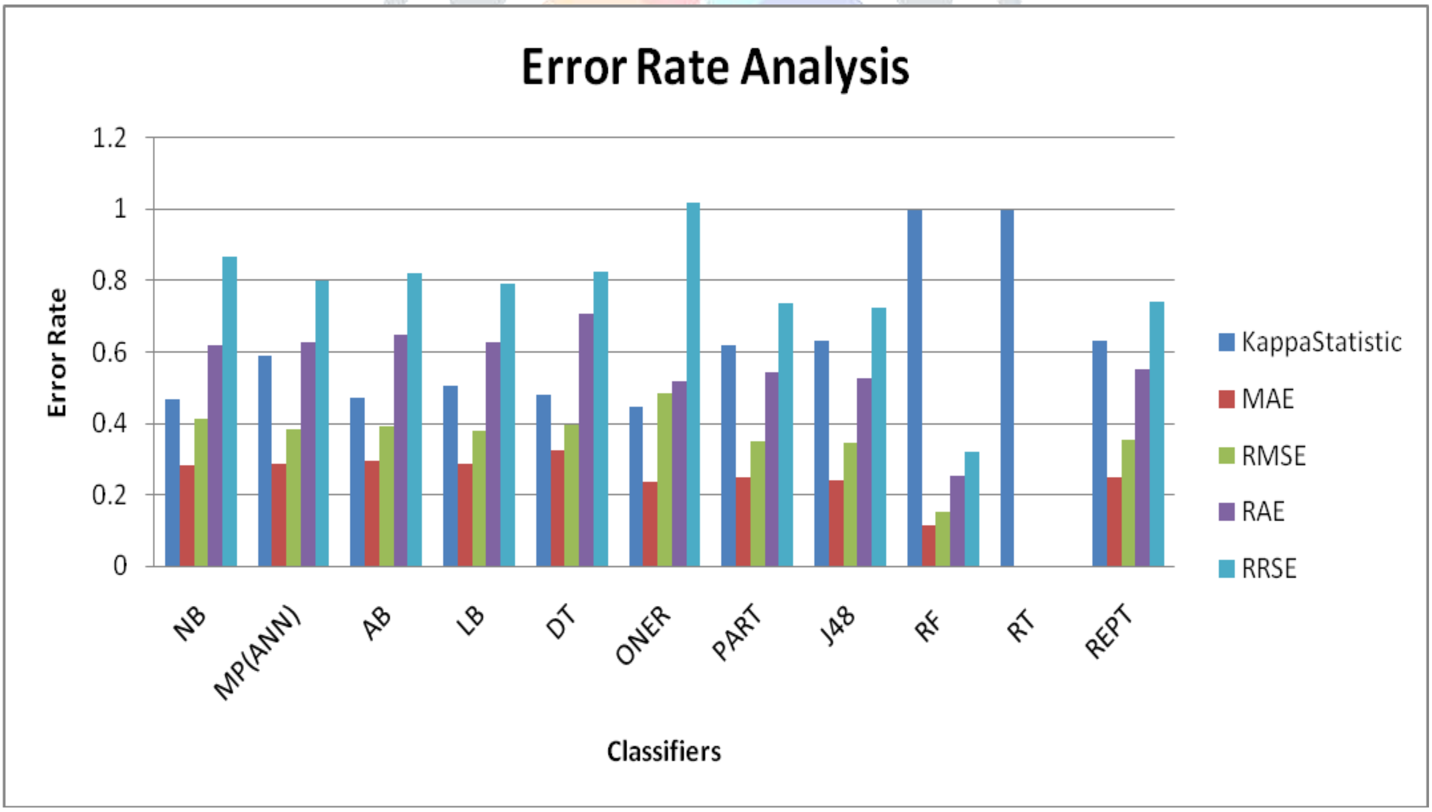


Figure 3.1.2 Error Rate Analysis

Figure 3.1.3 displays performance analysis of accuracy achieved by the algorithms.

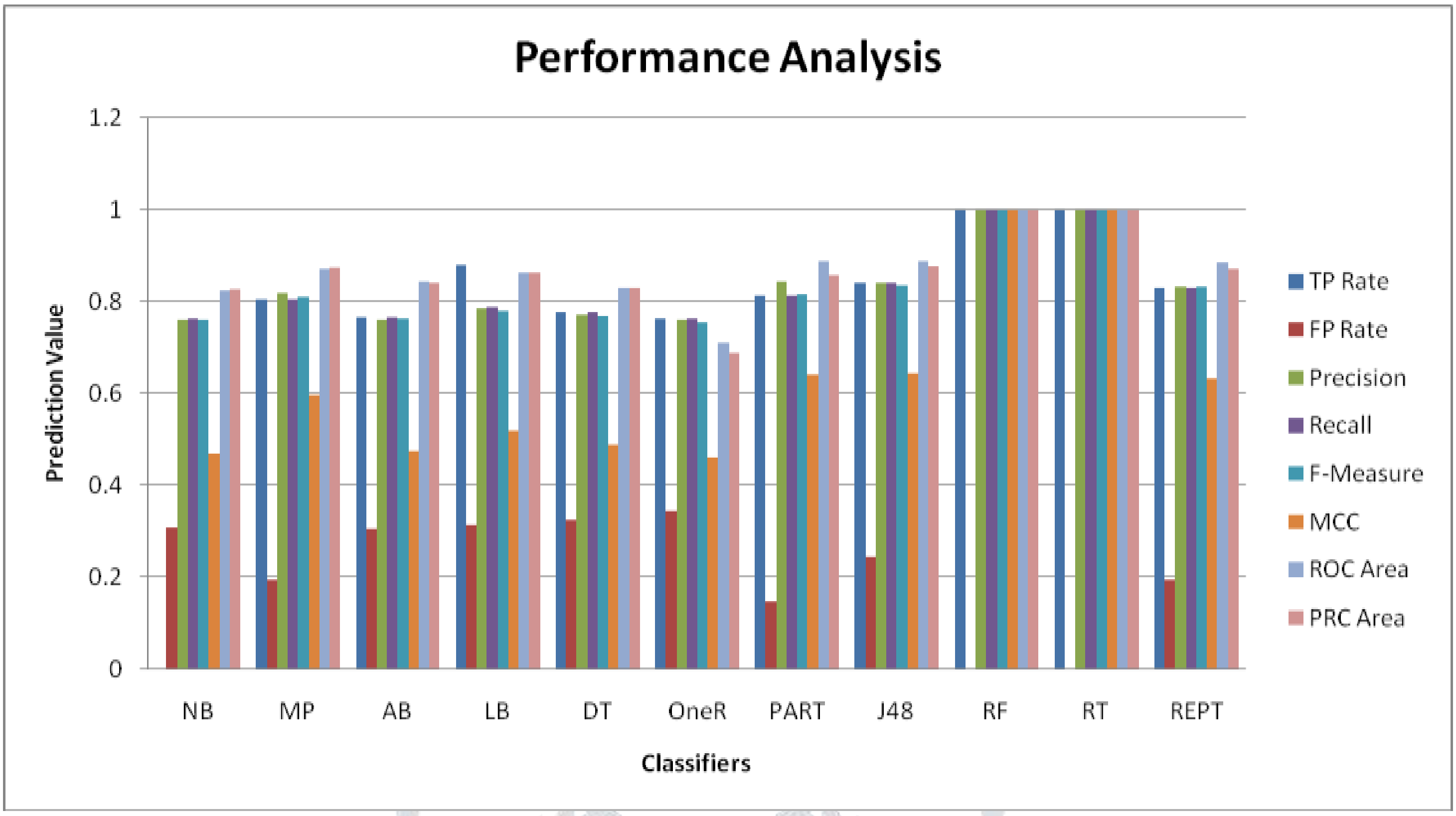


Figure 3.1.3 Performance Analysis based on Accuracy

IV. Conclusion

The primary destination of this study is to get better classification algorithms for the given dataset. This paper analysis various classification algorithms such as Naïve bayes, Multilayer Perceptron, Adaboost, Logiboost, Decision Table, OneR, PART, J48, Random Forest, Random Tree, REPT.From statistical analysis Random Forest and Random Tree gives least error rate and higher accuracy for diabetes dataset.Among different classification algorithms the random forest and random tree classifiers provides better results for diabetic diagnosis.

REFERENCES

[1] Mahmoud Heydari& Mehdi, Teimouri& Zainabolhoda, Heshmati&Seyed, Mohammad Alavinia, 2016, Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes DevCtries, Springer, 36(2):167–173.

[2] M. EmreCelebi, Hassan A. Kingravi, Patricio A. Vela, 2013, A comparative study of efficient initialization methods for the k-meansclustering algorithm. Expert Systems with Applications, Elsevier, 40, 200–210.

[3] HasanTemurtas, NejatYumusak, FeyzullahTemurtas, 2009, A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, Elsevier, 36, 8610–8615.

[4] K. Saravananathan1 and T. Velmurugan, 2016, Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874.

[5] S.Neelamegam, Dr.E.Ramaraj, 2013, Classification algorithm in data mining: An Overview. https://www.semanticscholar.org/paper/Classification-algorithm-in-Data-mining-%3A-An-Neelamegam-Ramaraj/2b19c569a03c5d2d232d13cd3eb1b56dc882d7db.

[6] SamanHina, Anita Shaikh and SohailAbulSattar, 2017, Analyzing Diabetes Datasets using Data Mining. Journal of Basic & Applied Sciences, Volume 13.

[7] AiswaryaIyer, S. Jeyalatha and RonakSumbaly, 2015, Diagnosis Of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.5, No.1.

[8] Sadri Sa'di, AmanjMaleki, RaminHashemi, Zahra Panbechi and KamalChalabi, 2015, Comparison Of Data Mining Algorithms Inthe Diagnosis Of Type Ii Diabetes. International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5.

[9] Thirumal P. C. and Nagarajan N., 2015, Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus - A Case Study. ARPN Journal of Engineering and Applied Sciences, Vol. 10, NO. 1, ISSN 1819-6608.

[10] Subhankar Manna, Malathi G., 2015, Performance Analysis Of Classification Algorithm On Diabetes Healthcare Dataset. International Journal of Research – Granthaalayah, Vol.5 (Iss.8), ISSN- 2350-0530(O), ISSN- 2394-3629(P) DOI: 10.5281/zenodo.890581.

[11] IoannisKavakiotis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, IoannaChouvarda, 2017, Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, Elsevier, 15, 104–116.

[12] JyotiKataria, BabitaKumari, 2017, Research Paper on Diabetic Data Analysis. International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Vol.8, No. 5.

[13] SamanHina, Anita Shaikh and SohailAbulSattar, 2017, Analyzing Diabetes Datasets using Data Mining. Journal of Basic & Applied Sciences, Vol.13, 466-471.

[14] Ashish Kumar Dogra, 2015, A Comparative Study of Selected Classification Algorithms of Data Mining. International Journal of Computer Science and Mobile Computing, Vol.4, Issue.6, pg. 220-229.

[15] MoloudAbdar, Sharareh R. NiakanKalhoriet.al , 2015, Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 6, pp. 1569~1576, 1569ρISSN: 2088-8708.

[16] ImranKurt, MevlutTure, A. TurhanKurum, 2008, Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease, Expert systems with applications, Vol. 34, 1, Pages 366-374.

[17] http://www.saedsayad.com/naive_bayesian.html

[18] https://data-flair.training/blogs/classification-algorithms/

[19] https://www.techopedia.com/definition/18829/decision-table-detab

[20] https://www.usfhealthonline.com/resources/healthcare/data-mining-in-healthcare/