ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

ANALYSIS OF CLASSIFICATION ALGORITHMS USING HEART DISEASES DATA SET FOR PREDICTION ITS ACCURACIES

D. Meganathan[*1] & N. Marudachalam[2]

[*1]Research Scholar PG and Research Department of Computer Science, Dr.Ambedkar Government Arts College, Chennai - 600 039, India

[2]Associate Professor PG and Research Department of Computer Science, Dr.Ambedkar Government Arts College, Chennai - 600 039, India

Keywords: *Rapid Miner, Random Tree, Navie Bayes, Decision tree, Random forest, K-Means clustering.*

ABSTRACT

Heart disease is the very important role for human death and we predict it at earlier stage to save the human life. So many of classification algorithms available in the data mining, we selected as few classification algorithms for heart disease prediction and found the accuracies. Different algorithms give various levels of accuracies. In the paper comparing the accuracies of few classification algorithms are Random Tree,Naives Bayes,Decision Tree and Random Forest then used K-Means clustering. The hungarian_csv, cleveland.csv and switzerland.csv heart disease data set received from UCI repository with 1272 instance and 14 regular attributes age, sex,cp, restbps, chol,fbs,restecg, thalach, exang, oldpeak , slope, ca, thalm, num were used here for analysis. Rapid miner studio software is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining predicate analytics and business analysis. The different measures and result were tabulated and charted.

INTRODUCTION

Data mining this is discovery process in the raw data previously unknown, non-trivial, practically useful, the interpretation of the available knowledge necessary for decision-making in the various spheres of human activity[1].This search for relationship with existing large associated data that are hidden among large amounts of data and refers to the "mining" knowledge from large amounts of data. Existing systems are used to assist in decision-making, referred to as data mining. These systems represent an iterative sequence of pre-processing as cleaning, data integration, and data selection is correct the pattern identification of data mining and knowledge representation. I collect the diagnosis heart disease data set from various source. That data set are involved in preprocessing when resulting data set as good formatted.

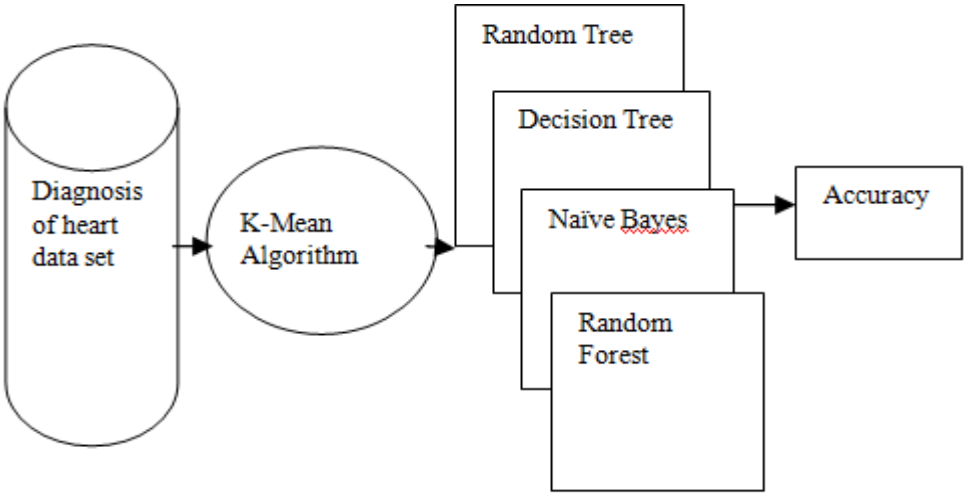

Fig 1 : Proposed Model

The propose problem of the papers is given above in the figure 1. The training heart disease data set and test data set is given as input of the K-Means clustering, classification algorithms and that accuracy compared for analysis. There are normally two types of data mining algorithms: one is supervised learning algorithms and

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

another one is unsupervised algorithms[6]. In supervised learning algorithms (like classification algorithms)is the data mining task of inferring a function from labeled training data[2]. A supervised learning algorithm analyzes the training data and produces an inferred function. In unsupervised learning algorithms (clustering algorithms)is that of trying to find hidden structure in unlabeled data.

The subsequent section of the paper present the brief literature review explaining the works of different researcher in this area. Section 3 describes the methodology followed in this research and in section 4 detailed analyses of results and significant extracted patterns from heart disease data is specified. Comparison and performance evaluation of Naïve Bayes with other algorithms is done is section 6. Conclusion and future work are explained in the last section 7.

LITERATURE REVIEW

Data mining uses its strong predicates models and algorithms which help in exploring, selecting and discovering the unknown/hidden information from a set of large data[4]. According to the literature reports that to predict heart diseases and to make heart disease decision support systems, developer or researchers use predictive models of data mining. To extract the significant patterns from coronary heart disease dataset and to predict heart attack, authors, have presented a dexterous approach. A rough set techniques associated to dynamic programming is suggested to abridge high interest features. In this research study authors have used Random Forest(RF), Decision Tree(DT)[5], Random Tree(RT) and Naïve Bayes(NB)[3] to classify the perilous heart disease cases. Three different approaches random forest are used in validation of results accuracy. It has been noted that forest-RI is the best among the different techniques. The relevant features are only takes into consideration which leads to reduce the complexity of the proposed model by focusing the study based on reduced features. Author in used Naives Bayes, Decision Tree and Neural Network to develop a prototype of Intelligent Heart Disease Prediction System(IHDPS)[7]. Beside it provides effective and cheap treatment and improves visualization and understanding. Among the three models the most efficient in prediction comes to be Naives Bayes followed by Neural Network and decision tree[8]. IHDPS is based on 15 medical attributes,909 record and only categorical data but it can be expanded to include more medical attributes[9], more techniques like Clustering, accuracy and Association Rules and continuous data as well.

FLOW OF WHOLE PROCESS

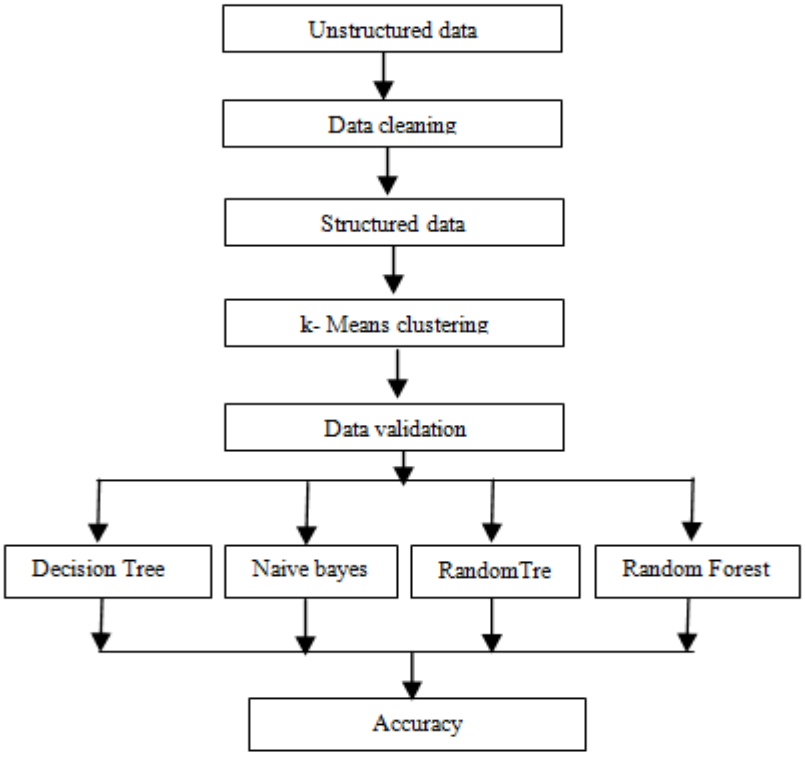

Fig 2 : Flow of Whole Process

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

METHODOLOGY

Rapid Miner
Rapid Miner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, validation and optimization.

Data Mining
Data Mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems[23]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data Mining is the analysis step of the "knowledge discovery in databases" process or KDD[11].

K-Means Clustering
k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem[10]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed apriori. The main idea is to define k centers, one for each cluster[24]. These centers should be placed in a cunning way because of different location causes different result. The heart disease data set clustering into two groups based on age.

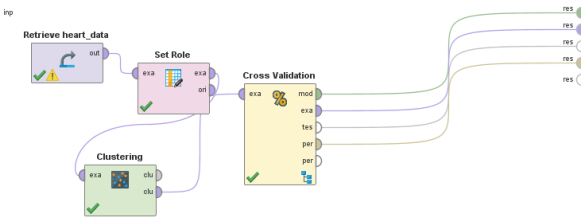

Fig 3 : K- Means clustering in rapid miner

CLASSIFICATION TECHNIQUES

Classification is a data mining function that assigns items in a collection to target category or classes[12]. The goal of classification is to accurately predict the target class for each case in the data. Classification models are tested by comparing the predicted values to known target values in a set of test data[13]. The historical data for a classification project is typically divided into two data sets: one for building the model, the other for testing the model. I have using few classification techniques only[22]. This are Naives Bayes, Random Tree, Random Forest and decision tree.

EXPLORATORY ANALYSIS AND RESULTS

Random Forests
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by[14] constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees[15]. Random decision forests correct for decision trees' habit of over fitting to their training set[16].

The heart disease data set with 1272 instance and 14 regular attributes is fed as input to the Random Forests classifier in Rapid Miner[25]. It gives the accuracy of 92.60%. The other related measures are Kappa static 0.1000, classification error 7.40% and weighted mean recall 55%.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

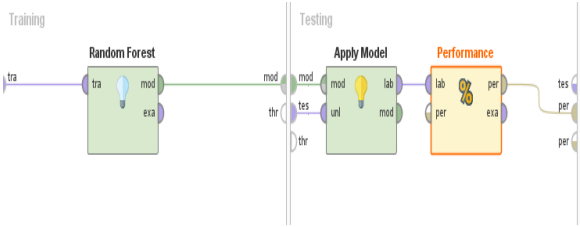

Fig 4 : Random forests in Rapid Miner

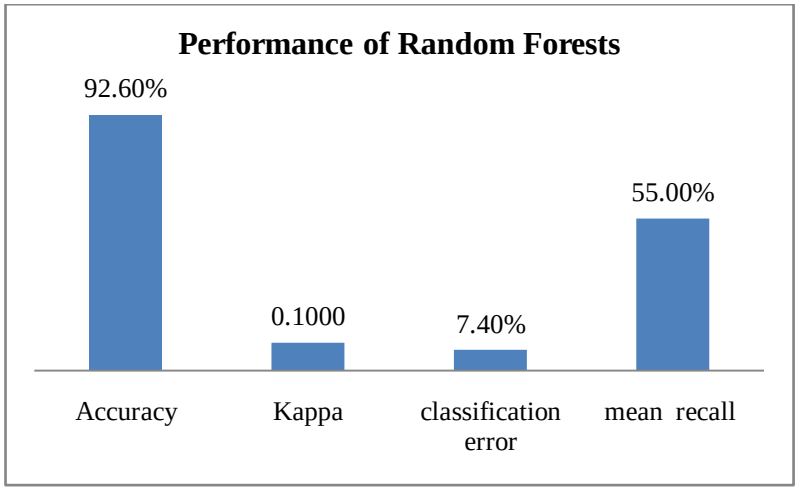

Fig 5 : Performance of Random Forests.

Table 1: Random Forest Confusion matrix

	True '<55_1'	True '>55'	Class precision
Pred '<55_1'	1167	94	92.55%
Pred '>55'	0	11	100.00%
Class recall	100.00%	10.48%	

Random Tree

The Random Tree operator works exactly like the Decision Tree operator with one exception: for each split only a random subset of attributes is available[17].

The heart disease data set with 1272 instance and 14 regular attributes is fed as input to the Random Tree classifier in Rapid Miner. It gives the accuracy of 95.5%. The other related measures are Kappa static 0.400, classification error 4.95% and weighted mean recall 70%.

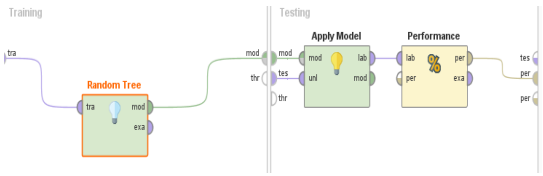

Fig 6 : Random Tree in Rapid Miner

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

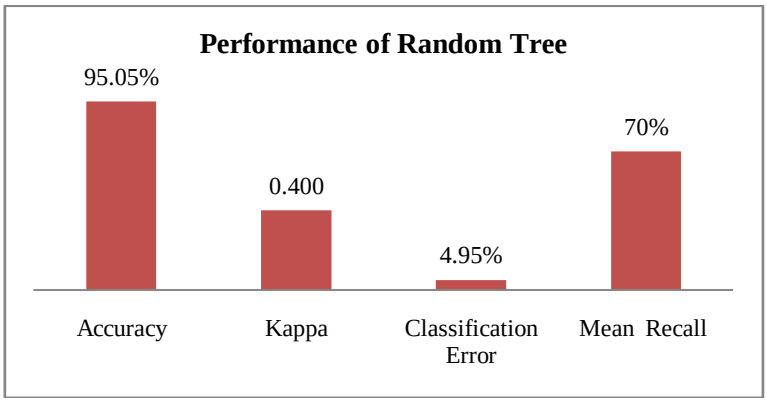

Fig 7 : Performance of Random Tree.

Table 2: Random Tree Confusion matrix

	True '<55_1'	True '>55'	Class precision
Pred '<55_1'	1127	63	94.88%
Pred '>55'	40	42	92.00%
Class recall	93.40%	40.00%	

NAIVE BAYES

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature[18].The heart disease data set with 1272 instance and 14 regular attributes is fed as input to the RandomTree classifier in Rapid Miner. It gives the accuracy of 98.51%. The other related measures are Kappa static 0.897micro, classification error 1.49% and weighted mean recall 93.06%.

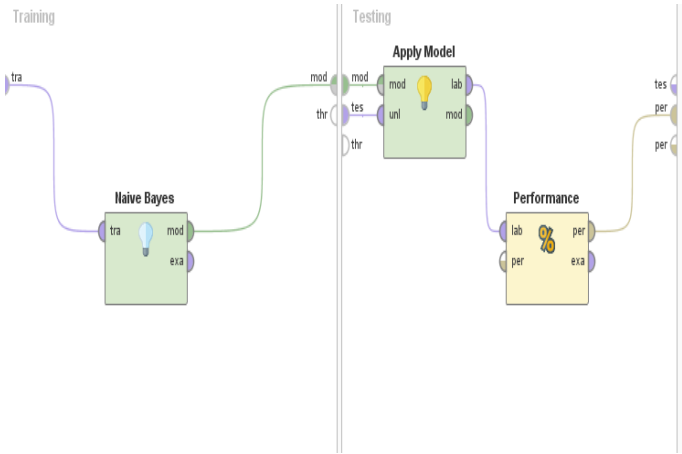

Fig 8 : Naive Bayes in Rapid Miner.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

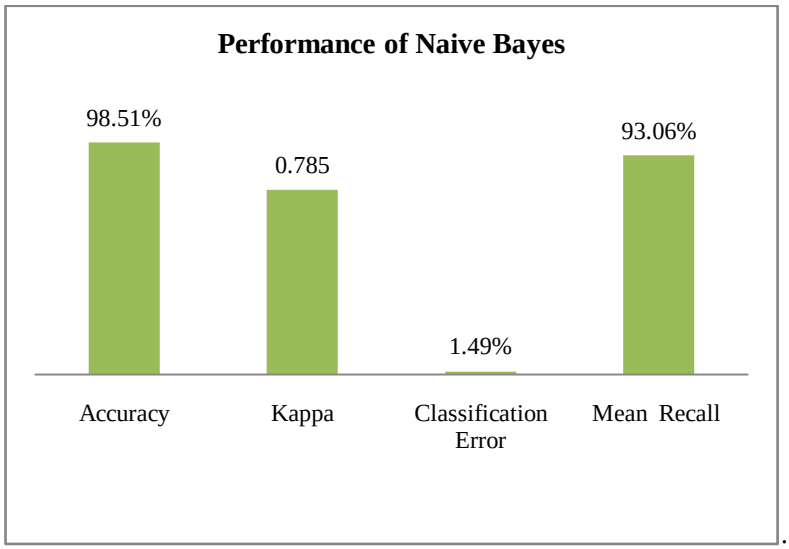

Fig 9: Performance of naïve bayes

Table 3: Naïve Bayes Confusion matrix

	True '<55_1'	True '>55'	Class precision
Pred '<55_1'	1162	14	98.81%
Pred '>55'	5	91	94.79%
Class recall	99.57%	86.67%	

DECISION TREE

Decision tree learning uses a decision tree as a predictive model observations about an item to conclusions about the item's target value It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees[19]; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels[20]. Decision trees where the target variable can take continuous values are called regression trees.

The heart disease data set with 1272 instance and 14 regular attributes is fed as input to the Random Tree classifier in Rapid Miner[21]. It gives the accuracy of 97.45%. The other related measures are Kappa static 0.785 micro, classification error 2.55% and weighted mean recall83.80%.

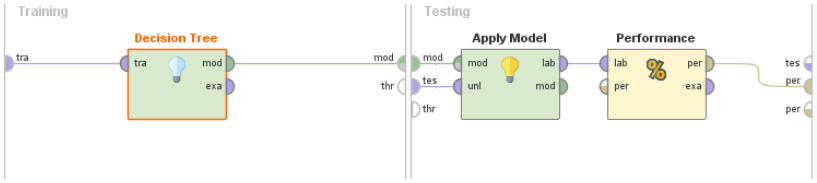

Fig 10 : Decision Tree in Rapid Miner

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

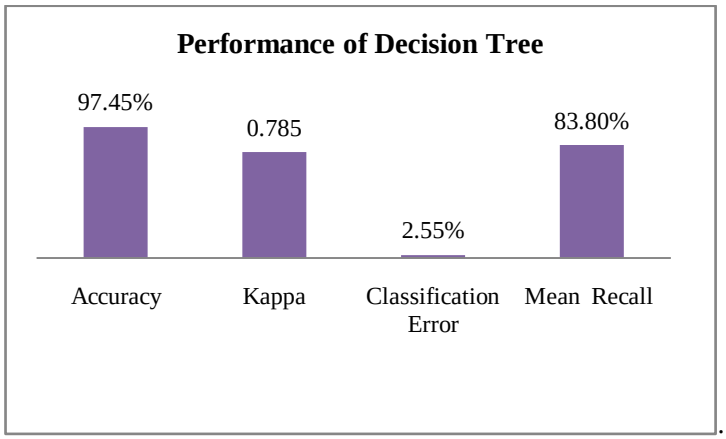

Fig 11:Performance of Decision Tree.

Table 4: Decision Tree Confusion matrix

	True '<55_1'	True '>55'	Class precision
Pred '<55_1'	1162	11	97.411%
Pred '>55'	3	93	93.39%
Class recall	97.57%	88.67%	

COMPARISON OF CLASSIFIERS

In this section to compared resultant of accuracy from above classifiers results.

Table 5: Comparison of accuracy

No	Classifier	True '>55'
1	Random Forest	92.60%
2	Random Tree	95.05%
3	Naïve Bayes	98.51%
4	Decision Tree	97.45%

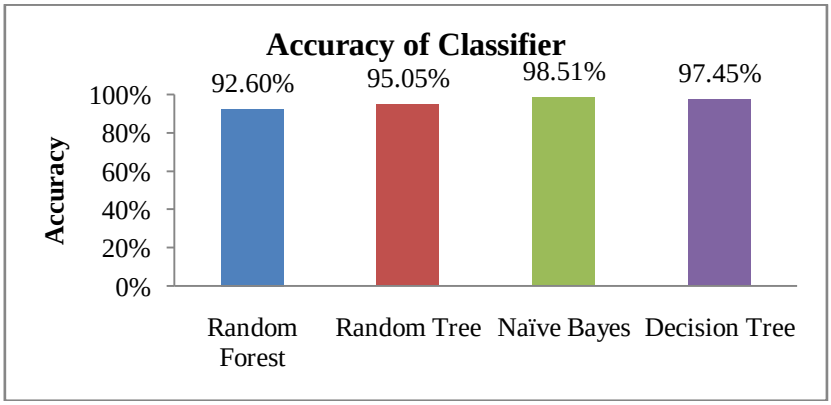

Fig 12 : Comparison of accuracy

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

CONCLUSION AND FUTURE ENHANCEMENTS

The accuracies of different classifiers with Heart disease data set is experimented with the support of Rapid Miner Software. The test and training datasets were passed as input to the Random tree, Naive Bayes, Random Forest and Decision Tree. We implement the data set and it found Naive Bayes better accuracy when compared to other classifiers is 98.51%. In future it can be implemented in artificial neural network in different large data set and improves efficiency and performance.

REFERENCES

1. *Jothikumar and Sivabalan.Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies.IDOSI Publication.2016.*
2. *JeroenEggermont and Joost N. Kok and Walter A. Kosters. Genetic Programming for data classification: partitioning the search space. SAC. 2004.*
3. *Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004.*
4. *Xiaoyong Chai and Li Deng and Qiang Yang and Charles X. Ling. Test-Cost Sensitive Naive Bayes Classification. ICDM. 2004.*
5. *Kaizhu Huang and Haiqin Yang and Irwin King and Michael R. Lyu and Laiwan Chan. Biased Minimax Probability Machine for Medical Diagnosis. AMAI. 2004.*
6. *David Page and Soumya Ray. Skewing: An Efficient Alternative to Lookahead for Decision Tree Induction. IJCAI. 2003.*
7. *Yuan Jiang Zhi and Hua Zhou and Zhaoqian Chen. Rule Learning based on Neural Network Ensemble. Proceedings of the International Joint Conference on Neural Networks. 2002.*
8. *Thomas Melluish and Craig Saunders and Ilia Nouretdinov and VolodyaVovk and Carol S. Saunders and I. Nouretdinov V.. The typicalness framework: a comparison with the Bayesian approach. Department of Computer Science. 2001.*
9. *Peter L. Hammer and Alexander Kogan and Bruno Simeone and SandorSzedm'ak. Rutcor Research Report. Rutgers Center for Operations Research Rutgers University. 2001.*
10. *Petri Kontkanen and Petri Myllym and TomiSilander and Henry Tirri and Peter Gr. On predictive distributions and Bayesian networks. Department of Computer Science, Stanford University. 2000.*
11. *Kristin P. Bennett and AyhanDemiriz and John Shawe-Taylor. A Column Generation Algorithm For Boosting. ICML. 2000.*
12. *Thomas G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning, 40. 2000.*
13. *Lorne Mason and Peter L. Bartlett and Jonathan Baxter. Improved Generalization Through Explicit Optimization of Margins. Machine Learning, 38. 2000.*
14. *EndreBoros and Peter Hammer and Toshihide Ibaraki and Alexander Kogan and Eddy Mayoraz and Ilya B. Muchnik. An Implementation of Logical Analysis of Data. IEEE Trans. Knowl. Data Eng, 12. 2000.*
15. *Kai Ming Ting and Ian H. Witten. Issues in Stacked Generalization. J. Artif. Intell. Res. (JAIR, 10. 1999.*
16. *IñakiInza and Pedro Larrañaga and Basilio Sierra and Ramon Etxeberria and Jose Antonio Lozano and Jos Manuel Peña. Representing the behaviour of supervised classification learning algorithms by Bayesian networks. Pattern Recognition Letters, 20. 1999.*
17. *Yoav Freund and Lorne Mason. The Alternating Decision Tree Learning Algorithm. ICML. 1999.*
18. *Jinyan Li and Xiuzhen Zhang and Guozhu Dong and KotagiriRamamohanarao and Qun Sun. Efficient Mining of High Confidience Association Rules without Support Thresholds. PKDD. 1999.*
19. *Floriana Esposito and DonatoMalerba and Giovanni Semeraro. A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Trans. Pattern Anal. Mach. Intell, 19. 1997.*
20. *Rudy Setiono and Huan Liu. NeuroLinear: From neural networks to oblique decision rules. Neurocomputing, 17. 1997.*
21. *Prototype Selection for Composite Nearest Neighbor Classifiers. Department of Computer Science University of Massachusetts. 1997.*
22. *Igor Kononenko and EdvardSimec and Marko Robnik-Sikonja. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Appl. Intell, 7. 1997.*
23. *D. Randall Wilson and Roel Martinez. Machine Learning: Proceedings of the Fourteenth International Conference, Morgan. In Fisher. 1997.*

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

24. *Kamal Ali and Michael J. Pazzani. Error Reduction through Learning Multiple Descriptions. Machine Learning, 24. 1996.*
25. *John G. Cleary and Leonard E. Trigg. Experiences with OB1, An Optimal Bayes Decision Tree Learner. Department of Computer Science University of Waikato.*