International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

ASSESSMENT OF DECISION TREE ALGORITHMS ON STUDENT'S RECITAL

Arundthathi A[1], Ms. K. Glory Vijayaselvi[2], Dr. V. Savithri [3]

[1] *Post graduate Student, M.Sc. CST Women's Christian College, Tamil Nadu, India*
[2] *Assistant Professor, M.Sc. CST, Women's Christian College, Tamil Nadu, India*
[3] *Assistant Professor, M.Sc. CST, Women's Christian College, Tamil Nadu, India*

Abstract - *Data mining is a dominant progression to forecast upcoming behaviors. Data Mining is used in various domains and disciplines to solve an existing problem or to envisage comportments. The different Data mining techniques are Clustering, Association, Classification, Regression and Structured Prediction. Classification is a most important task of simplifying dataset. Decision tree method is commonly used in Classification technique. Decision tree model is represented by branch and nodes. There are several Decision tree algorithms to contrivance in data mining tools. The objective of this study is to compare the most frequently used Decision tree algorithms in various domains and holds the good in predicting the best decision tree algorithm. Educational dataset is implemented to find the accuracy of the Decision tree algorithms and to predict the student's performance level. This research provides an idea to educators on student progress level.*

Key Words: **Classification, Decision Stump, Decision Tree algorithms, J48, Hoeffding Tree, Random Forest, Random Tree, REPTree.**

1. INTRODUCTION

Data mining is a prevailing technology with prodigious prospective to emphasis the most essential information in data warehouses. Data mining tools are used to foretell future tendencies, making practical and knowledge-driven verdicts. Data mining tools can response queries that usually were complex to resolve. Data mining tasks is used to discover hidden patterns from existing information. Data mining is applicable for any kind of data repository. Data mining methods can be applied on existing software and hardware platforms to improve the significance of present information resources.

Classification is a data mining technique used to organize data objects according to given class labels. Classification process slants using training set of data in which all data objects are already classified with class labels. The classification algorithm absorbs the training set and constructs a model. The constructed model is used to classify unclassified large datasets. There are many classification algorithms.

Decision tree is a renowned classification technique commonly used in many researches. Decision tree model represents a flowchart-like structure where each node represents a test on data objects and the leaf node represents the class label. Decision tree are used in all domains to predict hidden patterns. Decision tree is well-known because it is simple and easy to interpret.

The aim of this research is to compare the efficiency of different decision tree algorithms. Education is one of the domains which is profited by Data mining. To compare the decision tree algorithms Educational dataset from a reputed college is implemented. Semester marks of college students is collected and analyzed by Data mining tool to classify students to Grade A, Grade B or Grade C and predict the next semester percentage of each student.

This study finds out the commonly used decision tree algorithms in various domains. The objective of this study is to list out the efficiency and accuracy of decision tree algorithms. College student's performance in exam is analyzed to rank them and predict their future performance. This research helps professors to predict achievement levels and identify a student or a group of students in need of further attention.

2. REVIEW OF LITERATURE

Data mining is used in many researches for various purposes in different field. Researchers have worked in educational field to predict loyal students and dropout students to improve educational quality. Medical field is widely used with data mining to diagnosis many diseases like Breast cancer, Diabetics, Typhoid. In organizational field data mining is supportive to make decision and set marketing goal. In weather domain data mining commonly used to predict weather. In environment domain data mining is implemented and analyzed with soil, iris flower and mushroom datasets.

Nilima Patil, Rekha Lathi, and Vidya Chitre [1] provided the way for decision making process of customers to recommend the membership card using classification

which is helpful to enterprise's development. This study concludes that the main factor that affects the customer ranks is income and found that C4.5 performs well than CART. Saeide kakavand, Taha Mokfi and Mohammad Jafar Tarokh [2] proposes a decision tree model to predict loyal student to increase educational quality. They conclude that data mining is a powerful technology which can be used to predict faithful students with 90% accuracy. They compared the three common decision tree algorithms and found CART performed well followed by C.5 and CHAID with lowest accuracy. Adeyemo and Adeyeye [3] equipped a comparative study of decision tree algorithms and multiple perception algorithms for prediction of typhoid. They analyzed, compared the algorithms on medical dataset and found MLP had greater accuracy and C4.5 had greater speed. Sweta Rai, Priyanka Saini and Ajit Kumar Jain [4] constructed a decision tree model with ID3 algorithm to make a decision whether first year students from undergraduates will continue their study or drop their study. The study concludes that a student dropout seems to be associated with the residence, stress, family type, stream in higher secondary, satisfaction level, enrolled for other institute, change in goal, infrastructure of university, participation in extra-curricular activity, adjustment problem in hostel, and family problem. The ID3 classifier gives accuracy of 98%. A. Sivasankari, Mrs. S. Sudarvizhi and S. Radhika Amirtha Bai [5] proposed a comparative study on different decision tree and clustering algorithms. Activities dataset is implemented to clustering algorithms and concluded that K-means has greater speed and SOM has greater accuracy. Tumor dataset is applied to decision tree algorithms and concluded that ID3 has greater speed and C4.5 has greater accuracy.

Badr HSSINA et al [6] equipped a comparative study on top two decision tree algorithms (ID3 and C4.5). Weather dataset is implemented to compare the efficiency of the algorithms and concludes that C4.5 is powerful decision tree algorithm.G. Sujatha and Dr. K. Usha Rani [7] used tumor datasets on top three frequently used decision tree algorithms (ID3, C4.5 and CART). ID3, CART and C4.5 algorithms are implemented on different types of tumor datasets and compared. This paper concludes that C4.5 classifier performs best and ID3 performs equally well with enhanced dataset. V.Shankar sowmien et al [8] proposes a prediction system for liver disease using C4.5 decision tree algorithm and results with good accuracy. Data mining method is used to recognize uncovering patterns from the warehoused data and Decision tree techniques is used to discover accurate and reliable results. This paper concludes the accuracy of C4.5 is 85.81% and can be applied on real time applications. Aman Kumar Sharma and Suruchi Sahni [9] construct a model to classify electronic mails as spam or non-spam. The four different decision tree algorithms ID3, J48, Simple CART and Alternating Decision Tree are used and compared. This paper concludes that J48 outperforms with 92.7624% accuracy and Simple CART also exhibited alike results that were only a little dissimilar from J48 algorithm. Sudheep Elayidom.M, Sumam Mary Idicula and Joseph Alexander [10] proposed a new decision tree algorithm which is implemented with different UCI datasets and compared with other decision tree algorithms. ADT tree, REPT tree, Random Tree, C4.5*stat , C 4.5 , Neural Network and Naïve Bayes algorithms are implemented in Iris, Segment, Diabetes, Breast cancer ,Glass and Labor datasets in data mining tool WEKA. The result illustrates neural networks showed a higher accuracy.

3. METHODOLOGY

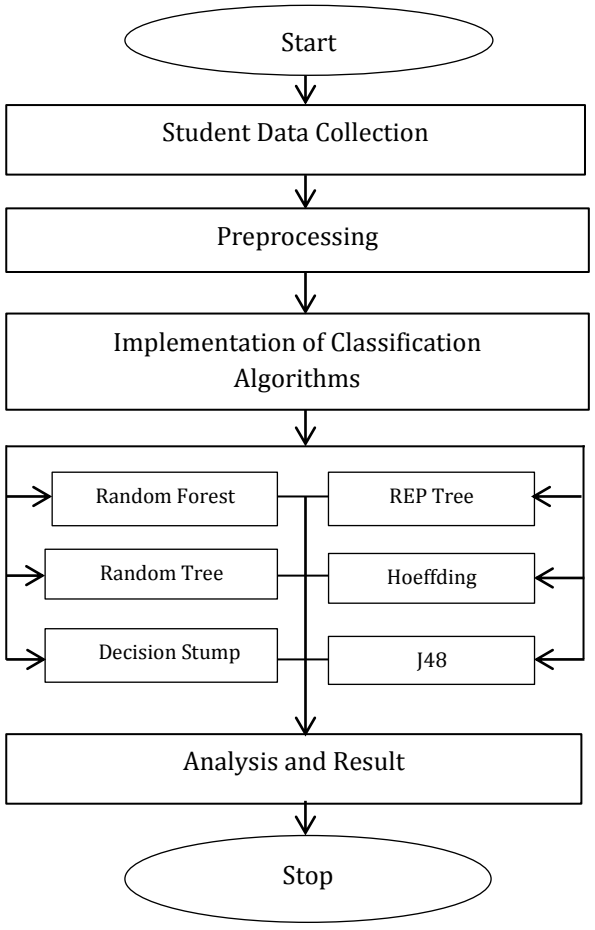

Fig - 1: Proposed Methodology

3.1 Data Collection

An educational dataset is used in this study. Student's semester percentage dataset from a reputed college is used. Dataset fields consist of student's name, register number, marks of various subjects and the percentage to classify students with Grade. Another Dataset fields consist of student's name, register number and the percentages of first five semesters to predict the percentage of sixth semester. The unnecessary attributes are removed and fields which influence the result are only selected.

Table -1: Students dataset's attributes and description for classification.

ATTRIBUTE	DESCRIPTION	DATA TYPE
Register Number	Unique register number for all students in the college	Numeric
Student's Name	Name of the registered student	String
Percentage	Percentage in the semester	Numeric

Table -2: Students dataset's attributes and description for prediction.

ATTRIBUTE	DESCRIPTION	DATA TYPE
Register Number	Unique register number for all students in the college	Numeric
Student's Name	Name of the registered student	String
Percentage of Semester I	Percentage in the first semester	Numeric
Percentage of Semester II	Percentage in the second semester	Numeric
Percentage of Semester III	Percentage in the third semester	Numeric
Percentage of Semester IV	Percentage in the fourth semester	Numeric
Percentage of Semester V	Percentage in the fifth semester	Numeric

3.2 Data Preprocessing

Data preprocessing is an important phase since real-world data are imperfect. In this process the missing values of attributes in the dataset are filled. The unnecessary attributes can be removed to improve performance. In the students dataset the only important field is the percentage. The final attributes selected are register number and percentage for classification process. In prediction process register number and five semesters percentage attributes are chosen.

3.3 Data Transformation

WEKA (Waikato Environment for Knowledge Analysis) data mining software is used in this study for implementation. Weka only accepts ARFF (Attribute Relation File Format) files. Hence the dataset should to be converted to ARFF. The dataset is stored in Microsoft Excel and saved as CSV (Comma Separated Value) file which is converted to ARFF file using Weka.

4. CLASSIFICATION ALGORITHM

4.1 Decision Tree Algorithms

Decision tree is a well-known classification technique. It is way to display algorithms in tree flowchart like structure. Decision tree algorithms are used commonly in researches due to simplicity and easy of understandability. Decision tree model uses flowchart symbol to represent the classification. The main advantage of decision tree is they can be combined with other decision-making techniques.

4.2 J48

This algorithm is an extension of well-known decision tree algorithm ID3. In WEKA which is the popular data mining tool J48 algorithm is the implementation of famous C4.5 algorithm. The C4.5 algorithm was developed by Ross Quinlan. It creates rules by which the dataset have to be classified. It creates decision tree centered on the given class labels.

4.3 Hoeffding Tree

The name of this decision tree is derived from the Hoeffding Bound which is used in decision tree induction process. The aim of Hoeffding bound is to give poise to the correct attribute to divide the tree to build the best model. Hoeffding tree is a streaming of decision tree induction technique. It was developed to overcome the previous streaming classification technique.

4.4 Random Forest

Random forest algorithm is famous in several competitions because of its powerful intensive calculation. It is similar to bootstrapping algorithm with CART (Classification and Regression Trees) model. It is programmed to build multiple CART model with diverse initial attributes. It is also capable of performing

regression technique in data mining. It builds multiple trees and each tree constructs a classification. The algorithm selects the best classification by "votes".

4.5 Random Tree

It is similar to Random Forest algorithm and they construct collaborative model with multiple decision trees by using distributed environment data. It tries to construct a huge number of models based on random subset of the input attributes. The strength of Random tree is they are robust and less prone to fitting because of bagging and field-sampling process.

4.6 REP Tree

Reduced Error Pruning Tree (REP Tree) is a fast decision tree algorithm. It can build both decision tree and regression tree by information gain and variance. It creates multiple trees in different iterations and can work only with numeric value attributes. Missing values in the dataset are handled by C4.5 algorithm's technique.

4.7 Decision Stump

It is a one-level decision tree model which has only one internal node connected to the leaves. This algorithm prediction process depends only on the value of single attribute. Many variations are conceivable depending on the type of selected attribute. They are mostly used as components in machine learning technique like boosting.

5. EXPERIMENTAL RESULTS

Waikato Environment for Knowledge Analysis is a popular data mining software which is written in Java and developed at the University of Waikato New Zealand. It is open-source software and licensed under the GNU General Public License. Weka has an easy understandable GUI (Graphical User Interface) and consists of several tools for data pre-processing and algorithms for decision making models. Weka comprises several data mining techniques like classification, clustering, visualization, association, feature selection and regression.

As Weka only accepts ARFF file format, first the student dataset is converted to ARFF file using Weka and imported. The classification tab in Weka Explorer is used for classification where several classification techniques or algorithms like bayes, trees, functions, rules and meta are present. All the decision tree algorithms are implemented to the imported educational dataset. The visualized tree, efficiency and the time consumed for each decision tree algorithm is noted.

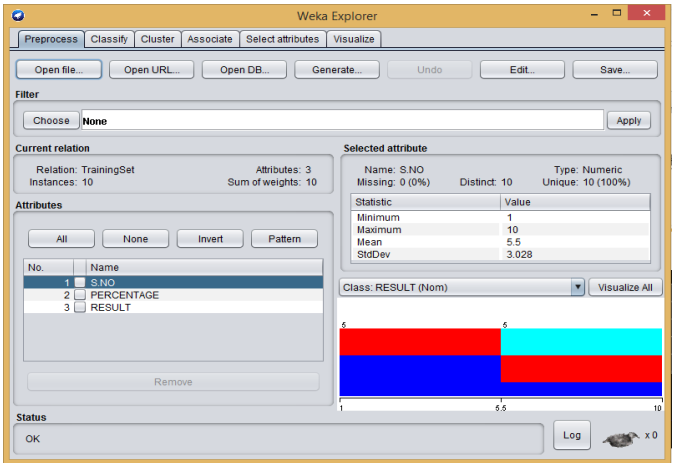

Fig - 2: Education dataset opened in Weka Explorer

5.1 Performance Evaluation

The decision tree algorithms are compared by efficiency. The number of currently classified instanced are known as the accuracy or efficiency of the algorithm. First the training dataset is implemented with the decision tree algorithm and then the test datasets are applied to classify the students with grade. In the training dataset each student is labeled with Grade A or B or C. The student is given Grade A if the percentage is above 75%, Grade B if the percentage is above 55% and Grade C if the percentage is less than 55%. Then with student's first five semesters percentage by Random Tree algorithm the student's sixth semester percentage is predicted. Datasets are separated by different years of students.

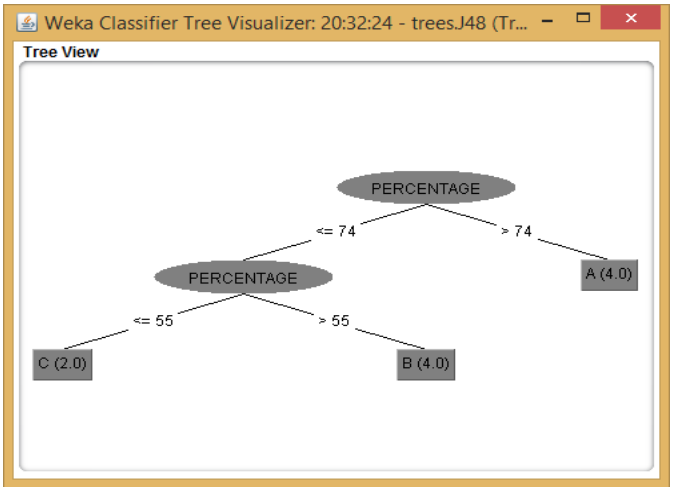

Fig - 3: J48 Tree model of Training dataset in classification.

Table -3: Comparison of Accuracy of Classifiers.

Classification Algorithm	Training Dataset	Dataset I	Dataset II	Dataset III
J48	100%	100%	100%	100%
Hoeffding	90%	100%	92.3077%	66.6666%
Random Forest	100%	100%	100%	100%
Random Tree	100%	100%	100%	100%
REP Tree	80%	96%	100%	83.3333%
Decision Stump	80%	100%	100%	100%

Table -4: Accuracy of Classifier in predicting student's percentage.

Classification Algorithm	Training Dataset I	Training Dataset II	Testing Dataset
Random Tree	99%	100%	85%

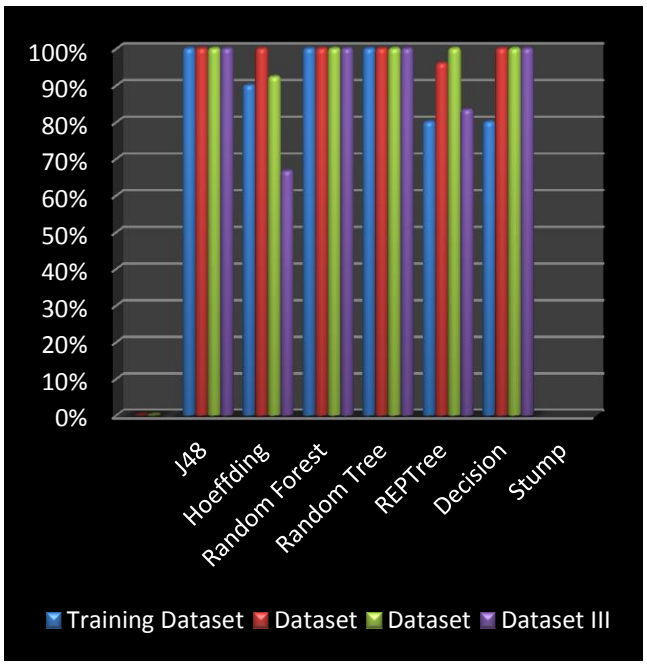

Chart -1: Graph on accuracy of decision tree Classifiers

Chart -2: Graph on accuracy of Random Tree to predict next semester percentage on different datasets.

6. CONCLUSION

In this study the different decision tree algorithms performance are compared by the model which predicts student's recital and classifies them according to their Grade. Six decision tree classifiers (Hoeffding , REP Tree, Decision Stump, Random Tree, Random Forest and J48) which work on numeric data are compared. Students of different year's datasets from a reputed college are used and the efficiency of the algorithms is analyzed. From the results it is clear that all the decision tree algorithms perform well with student's dataset to predict their recitals and it is proved that the efficiency of the algorithms differs with datasets. Among the six classifiers Random Tree, Random Forest and J48 algorithms show outstanding performance.

Thus, a comparative study on Classification Decision tree algorithms on educational datasets was done and the studies reveal that all the decision tree algorithms work with small datasets. It is also proved that data mining can be used thriving in educational domain.

The current study was on literature study and implementation of decision tree algorithms on small educational datasets. So future works will emphasis on comparison of decision tree algorithms with large datasets. This study proposed the prediction system on student's recital (Grade classification and Percentage prediction). It would be taken to the next level by predicting marks for each subject.

ACKNOWLEDGEMENT

The educational datasets used in this study were provided with consent by Ms. I.S Mary Ivy Deepa, HOD in the Department of Computer Science and Technology, Women's Christian College. I would like to acknowledge the exertions of all the teaching staff members especially Ms. Jerline Amutha, of department of M.Sc. CST for motivating to complete this study.

REFERENCES

[1] Prof. Nilima Patil, Prof. Rekha Lathi, Prof. Vidya Chitre "Customer Card Classification Based on C5.0 & CART Algorithms" International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 4, July-August 2012.

[2] Saeide kakavand, Taha Mokfi, Mohammad Jafar Tarokh "Prediction the Loyal Student Using Decision Tree Algorithms" International Journal of Information and Communication Technology Research, Volume 4 No. 1, January 2014.

[3] O..O.Adeyemo, T..O.Adeyeye , D.Ogunbiyi "Comparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever" African Journal of Computing & ICT, Vol 8. No. 1 – March, 2015.

[4] Sweta Rai, Priyanka Saini, Ajit Kumar Jain "Model for Prediction of Dropout Student Using ID3 Decision Tree Algorithm" International Journal of Advanced Research in Computer Science & Technology, Vol. 2 Issue 1 Jan-March 2014.

[5] Ms. A. Sivasankari, Mrs. S. Sudarvizhi, S. Radhika Amirtha Bai "Comparative Study of Different Clustering and Decision Tree for data mining algorithm" International Journal of Computer Science and Information Technology Research, Vol. 2, Issue 3, pp: (221-232), Month: July-September 2014.

[6] Badr Hssina, Abdelkarim Merbouha,Hanane Ezzikouri, Mohammed Erritali "A comparative study of decision tree ID3 and C4.5" International Journal of Advanced Computer Science and Applications , Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.

[7] G. Sujatha, Dr. K. Usha Rani "Evaluation of Decision Tree Classifiers on Tumor Datasets" International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 4, July – August 2013.

[8] V.Shankar sowmien, V.Sugumaran , C.P.Karthikeyan, T.R.Vijayaram "Diagnosis of Hepatitis using Decision tree algorithm" International Journal of Engineering and Technology (IJET), Vol 8 No 3 Jun-Jul 2016.

[9] Aman Kumar Sharma, Suruchi Sahni "A Comparative Study of Classification Algorithms for Spam Email Data Analysis" International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5 May 2011.

[10] Sudheep Elayidom.M, Sumam Mary Idicula, Joseph Alexander "A Novel Decision Tree Algorithm for Numeric Datasets - C 4.5*Stat" International Journal of Advanced Computing, ISSN: 2051-0845, Vol.36, Issue.1.

[11] Sunita B. Aher and Lobo L.M.R.J. "Comparative Study of Classification Algorithms" International Journal of Information Technology and Knowledge Management July-December 2012, Volume 5, No. 2, pp. 307-310.

[12] Irfan Ajmal Khan and Jin Tak Choi "An Application of Educational Data Mining (EDM) Technique for Scholarship

Prediction" International Journal of Software Engineering and Its Applications Vol. 8, No. 12 (2014), pp. 31-42.

[13] T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.Prasanna Venkatesan "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data" I.J. Modern Education and Computer Science, 2013, 5, 18-27.

[14] Harvinder Chauhan, Anu Chauhan "Implementation of decision tree algorithm c4.5" International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013.

[15] Jaimin N. Undaviaa, Dr. P.M.Doliab and Dr. AtulPatela "Comparison of Classification Algorithms to Predict Comparison of Decision Tree Classification Algorithm to Predict Student's Post Graduation Degree in Weka Environment " International Journal of Innovative and Emerging Research in Engineering Volume 1, Issue 2, 2014.

[16] Shikha Chourasia "Survey paper on improved methods of ID3 decision tree classification" International Journal of Scientific and Research Publications, Volume 3, Issue 12, December 2013.

[17] Masud Karim, Rashedur M. Rahman "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing" Journal of Software Engineering and Applications, 2013, 6, 196-206.

[18] Anuja Priyama, Abhijeeta, Rahul Guptaa, Anju Ratheeb, and Saurabh Srivastavab "Comparative Analysis of Decision Tree Classification Algorithms" International Journal of Current Engineering and Technology, Vol.3, No.2 June 2013.

[19] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, Rahul Pitale "Comparison of different Classification techniques using different datasets" International Journal of Advances in Engineering & Technology, Vol. 6, Issue 2, pp. 764-768, May 2013.

[20] Sreerama K.Murthy, Simon Kasif, Steven Salzberg "A System for Induction of Oblique Decision Trees" Journal of Artificial Intelligence Research [1994] 1-32.

[21] Jay Gholap "Performance tuning of J48 algorithm for prediction of Soil fertility".

[22] Kasra Madadipouya "A New Decision tree method for Data mining in Medicine" Advanced Computational Intelligence: An International Journal (ACII), Vol.2, No.3, July 2015.

[23] Amit Gupta, Ali Syed, Azeem Mohammad, Malka N. Halgamuge "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

[24] Abid hasan "Evaluation of Decision Tree Classifiers and Boosting Algorithm for Classifying High Dimensional Cancer Datasets" International Journal of Modeling and Optimization, Vol. 2, No. 2, April 2012.

[25] Pooja Sharma, Asst. Prof. Rupali Bhartiya "Implementation of Decision Tree Algorithm to Analysis the Performance" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 10, December 2012.

[26] Mai Shouman, Tim Turner, Rob Stocker "Using Decision Tree for Diagnosing Heart Disease Patients" CRPIT Volume 121 - Data Mining and Analytics 2011.

[27] Rachna Raghuwanshi "A Comparative Study of Classification Techniques for Fire Data Set" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 78-82.

[28] D.Lavanya and Dr.K.Usha Rani "Ensemble Decision tree Classifier for Breast Cancer data" International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.

[29] Md. Nurul Amin, Md. Ahsan Habib "Comparison of Different Classification Techniques Using WEKA for Hematological Data" American Journal of Engineering Research (AJER) 2015, Volume-4, Issue-3, pp-55-61.

[30] Pardeep Kumar, Nitin and Vivek Kumar Sehgal and Durg Singh Chauhan "A Benchmark to select Data mining based Classification algorithms for business intelligence and decision support systems" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.