International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-05, Aug 2018

Classification of Breast Cancer Tissues using Decision Tree Algorithms

Nidhi[1] , Mukesh Kumar[2] , Shaveta Makkar[3]

Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India.

nidhi@chitkarauniversity.edu.in , mukesh.kumar@chitkarauniversity.edu.in

shaveta.makkar@chitkarauniversity.edu.in

Abstract - Nowadays, Healthcare sector data are enormous, composite and diverse because it contains a data of different types and getting knowledge from that data is essential. So for this purpose, data mining techniques may be utilized to mine knowledge by building models from healthcare dataset. At present, the classification of breast cancer patients has been a demanding research confront for many researchers. For building a classification model for the cancer patient, we used four different classification algorithms such as J48, REPTree, RandomForest, and RandomTree and tested on the dataset taken from UCI. The main aim of this paper is to classify the patient into benign (not cancer) or malignant (cancer), based on some diagnostic measurements integrated into the dataset.

Keywords - Data Mining, Classification, RandomForest, RandomTree, J48, REPTree

I. INTRODUCTION

One of the deadliest diseases among women across world is breast cancer. It is growing at very high rate. Predicting the type of cancer using data mining tools is a subject of research with good value. Today, it is the major cause of death in women. It is the topmost cancer in women worldwide and is expanding particularly in advancing countries where the majority of cases are detected in late stages [1]. The classification of Breast Cancer data can be useful to predict the genetic behaviour of tumours like benign (not cancer) or malignant (cancer).

Knowledge Discovery from Databases (KDD) in data mining includes different approaches such as classification, clustering, self-organising map (SOM) [2, 3, 4]. KDD has various data processing steps which includes selection, pre-processing, transformation, data mining and evaluation [5]. The relevant data selected for data mining is known as selection [5, 6]. Removal of inconsistent, noisy and incorrect data is called pre-processing [5, 6]. After pre-processing normalization and generalization of data is done which is known as data transformation [5, 6]? In order to get good patterns required for particular task different data mining methods are applied [5, 6]. Lastly, we represent or interpret the desired result known as evaluation [5, 6].

The main purpose of the paper is to build a data analytical model which can help doctors

i. To analyze and preprocess the available dataset with WEKA.
ii. To identify the type of cancer with respect to patient's attributes.

Previous researches using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [16] have shown significantly good result using machine learning algorithms [7], gated recurrent unit (GRU) with the support vector machine (SVM) [8].

In this paper, we will focus on different classification techniques and find out the best possible method for classification on type of tumour.

II. MATERIALS AND METHODS

The Dataset description: Classification algorithms are used to identify the type of breast cancer using Wisconsin Diagnostic Breast Cancer (WDBC) dataset. It consists of around 32 features as described in table 1.

Table 1: Dataset description used to implement classification algorithm

S.No	Attribute Name	Description of Attribute
1.	diagnosis	Diagnosis of breast tissues (M = malignant, B = benign)
2.	radius_mean	mean of distances from centre to points on the perimeter
3.	perimeter_mean	mean size of the core tumor
4.	area_mean	
5.	concavity_mean	mean of severity of concave portions of the contour
6.	concave points_mean	mean for number of concave portions of the contour
7.	area_se	
8.	radius_worst	"worst" or largest mean value for mean of distances from centre to points on the perimeter
9.	Perimeter_worst	

 DOI : 10.18231/2454-9150.2018.0636

10.	area_worst	
	concave points_worst	"worst" or largest mean value for number of concave portions of the contour

Classification Techniques: Classification techniques are widely used to analyse the given dataset and assign all its instances to a particular class so that error can be minimalized. It is two-step process. In the first step algorithms are applied on training dataset and then model is tested against the dataset in order to measure the accuracy. Different types of classification techniques have been proposed in literature that includes J48, Sequential Minimal Optimization (SMO), Naive-Bayesian methods, IBK, BF Tree etc.

J48 [9][10]: It is the modified version of c4.5 algorithm which produces decision tree as output. Decision tree consists of root node, in-between nodes and leaf node. On each node decision is labeled. Here, decision tree divides the input sample into different areas, each area having a label, a value or an action to describe its data points. On the basis of splitting, each attribute is decided to be placed at particular node.

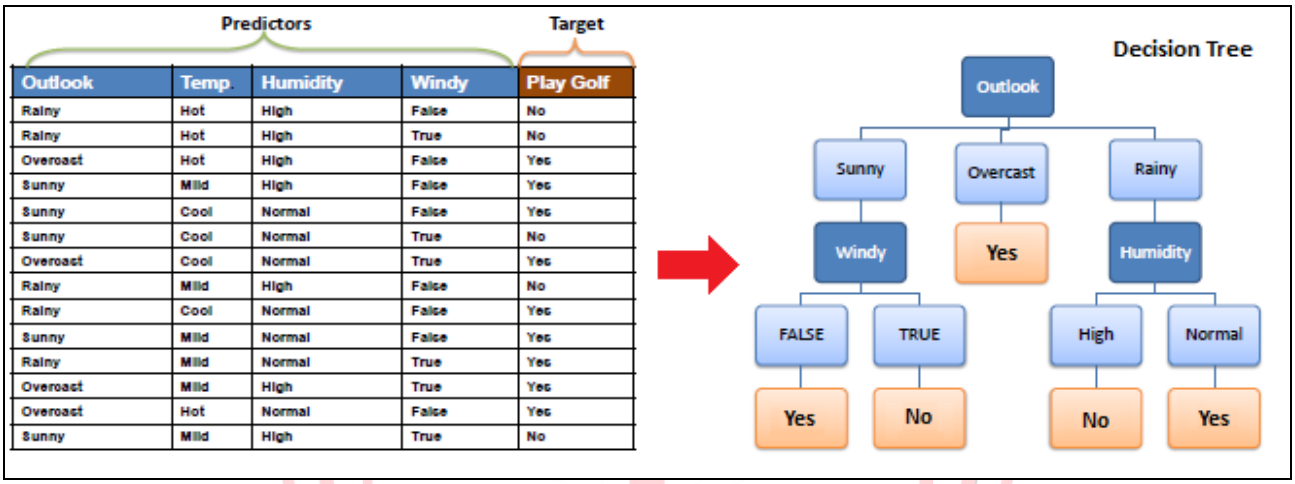


Figure 1: Decision Tree generated with given dataset

Random Forest [11]: This method is a supervised classification algorithm which consists of number of trees known as forests. As the name suggests forest (number of trees) is robust as it has many trees, in the same way if there are more number of trees in the classifier it would have high accuracy. Here, information gain or gini index is not used for modelling of decision tree. The main advantage of this model is that it can handle missing values too. Random forest selects m features from total n features. From those m features, it calculates the node d using best split method. It does this process until k numbers of nodes are selected.

BF Tree [12]: The divide-and conquer strategy is used each step in depth-first order in order to expand Best First Tree. Generally, left to right fixed order is used to expand the nodes. For best-first order boosting algorithm is used for the selection of best fit. In the growing phase, both gain and index is used for calculating the best node in the tree. The best node is used to reduce impurity among all nodes.

REPTree [14]: Reduced Error Pruning Tree generates multiple trees using different iterations and selects the best tree using sorted values of numeric attributes. This tree uses information gain method for splitting. Here, missing values are solved with C4.5 algorithm.

RandomTree [15]: Random tree constructs the tree which consists of K randomly chosen attributes at each node. This method do not use pruning and has no way to calculate target mean in the regression case on backfitting.

III. THEORY AND CALCULATION

All the experiments for classification of breast cancer dataset in this paper are performed on WEKA tool. WEKA has different tools for data pre-processing, classification, regression, clustering and visualization. Classification method used for this dataset are used with the slightly changes in the parameters in order to increase the accuracy of classification methods. The result in this paper clearly indicates that the attribute taken for classification easily classifies the type of cancer in breast. Pre-processing of dataset is done using ranker method by choosing attribute evaluator as InfoGainAttributeEval. InfoGainAttributeEval is used for feature selection for measuring how much feature contributes in decreasing the overall entropy [13]. After applying ranker method top 10 attributes were selected for classification in dataset.

Table 2:

Attribute Name	Minimum	Maximum	Mean	Standard Deviation
Radius_Mean	6.98	28.11	14.12	3.52
Perimetre_Mean	43.79	188.5	91.96	24.29
Area_Mean	143.5	2501	654.88	351.91
Concavity_Mean	0	0.427	0.089	0.08
Concave Points_Mean	0	0.201	0.049	0.039
Area_se	6.802	542.2	40.33	45.49
Radius_worst	7.93	36.04	16.29	4.83
Perimeter_worst	50.41	251.2	107.26	33.60
Area_worst	185.2	4254	880.58	569.35

 DOI : 10.18231/2454-9150.2018.0636

Concave points_worst	0	0.291	0.115	0.066

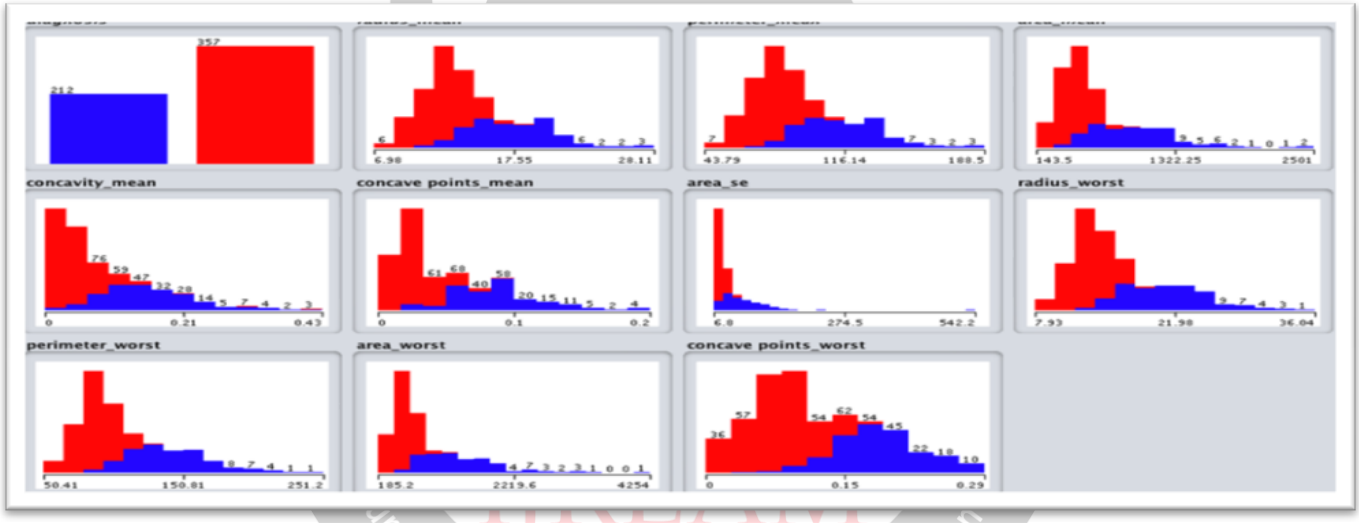
IV. ALGORITHM USED FOR CLASSIFICATION

In this study, we are using four different decision tree classification algorithms namely J48, RandomForest, REPTree, and RandomTree to build the model for classification of breast cancer patients. The dataset was further divided implemented on cross-validation and percentage split techniques. Here, we are using 10-fold cross-validation to prepare training and testing dataset. First of all we check our dataset for baseline accuracy with ZeroR classification algorithm. The baseline accuracy for our dataset is 62.74%, which implies that we need to do a lot of work to improve the accuracy of our dataset. After data pre-processing, the J48 algorithm is implemented on the dataset using WEKA tool kit which further divided data into "tested positive" or "tested-negative" as two classes. Table-3 Shows the experimental result of different classification algorithm like J48, RandomForest, REPTree, and RandomTree. We have conceded some implementation to evaluate the working and accuracy of classification algorithms for classifying diabetes patient's dataset.

Table 2: Performance of the Classifiers used for implementation

Classification Algorithm Implementation	Classification Algorithm used for Implementation			
	J48	REPTree	RandomForest	RandomTree
Timing to build the model	0 Sec	0.01 Sec	0.07 Sec	0.00 Sec
Correctly classified instances	533	532	541	536
Incorrectly classified instances	36	37	28	33
Accuracy (%)	93.67 %	93.49%	95.07%	94.20%

Figure 2 Attributes Selected after pre-processing with respect to diagnosis using WEKA.



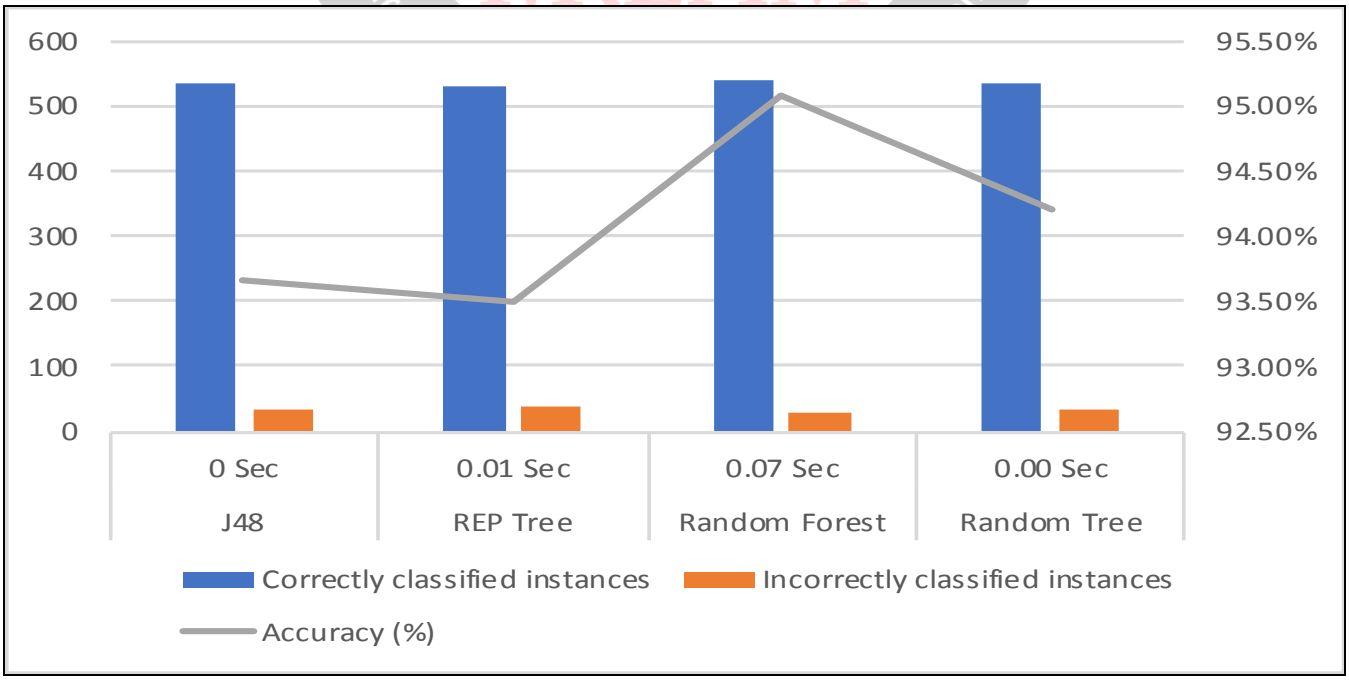


Figure 3: Classification algorithms with correctly or incorrectly classified instances

 DOI : 10.18231/2454-9150.2018.0636

Here we show that Random Forest decision tree classification algorithm has performed with very well with more accuracy as compared to other algorithm used. In WEKA tool, the percentage of correctly classified instances is frequently known as accuracy of the classifying model. We have other evaluation criteria for classification Algorithm implementation are Kappa Statistic (KS), Mean Absolute Error(MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) will be in numeric value only. In table 3 we illustrate the simulation result for different algorithm with their evaluation criteria in percentage for references and evaluation of algorithm.

Table 3: Training and Simulation error of each classifier used in the implementation

Evaluation Criteria for classification Algorithm implementation	Classification Algorithm used for Implementation			
	J48	REP Tree	Random Forest	Random Tree
Kappa statistic	0.8649	0.8616	0.8945	0.877
Mean absolute error	0.0904	0.0692	0.0774	0.058
Root mean squared error	0.232	0.2419	0.1967	0.2408
Relative absolute error	19.3262%	14.7888 %	16.5574%	12.4012%
Root relative squared error	47.9884 %	50.0238 %	40.6793%	49.8081 %

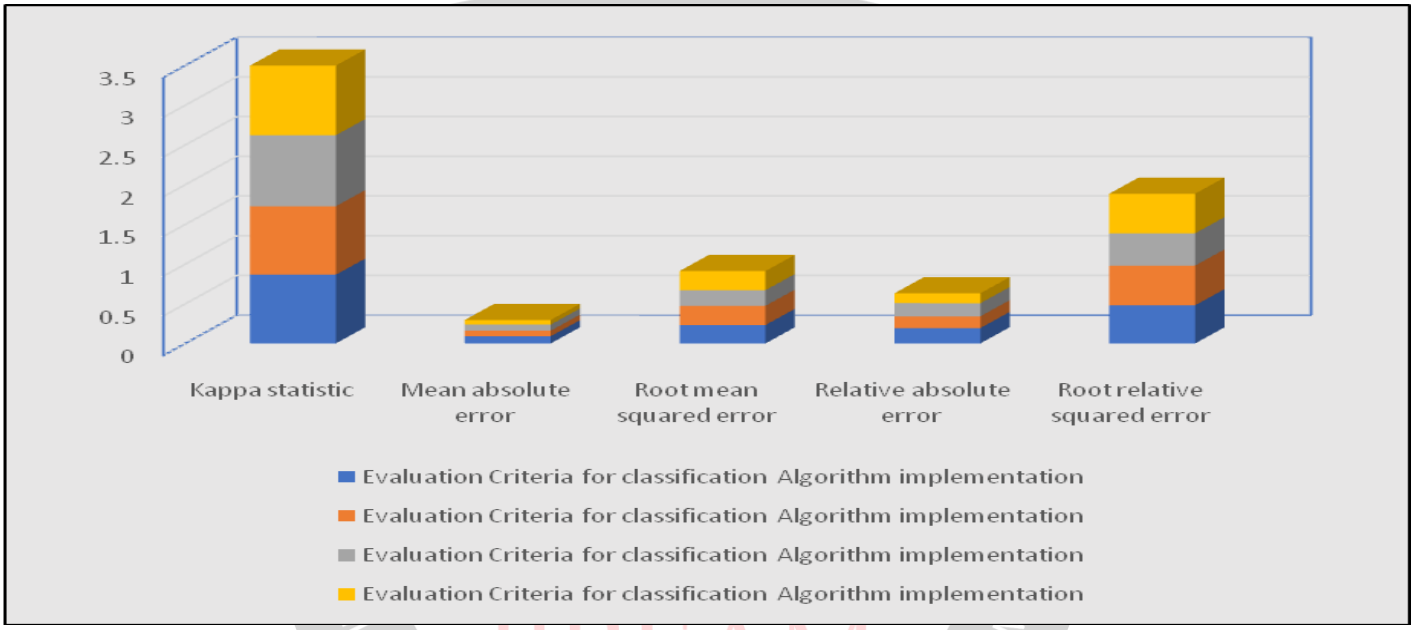


Figure 4: Comparison between different evaluation criteria of classification algorithm

Figures 4 show the graphical representations of the simulation result with are represented in table 3 above for proper visualization of the result. To choose the superlative algorithms for soaring performance, different algorithms are implemented and evaluated with respect to some evaluation criterion on selected dataset. The algorithm which achieves the utmost performance in terms of soaring recall and precision value is measured the finest algorithm. From table 4, it is clear that RandomForest classification algorithm achieves the maximum value.

The efficiency of the machine learning classifier can be assessed with numerous measures. The estimate of these measures basically depends on contingency table which is further obtained from the classification algorithm implemented. Table 4; contain the value of contingency table of a particular diabetes patient dataset.

Table 4: Comparison of accuracy measures of each classifier used in an implementation

Classification Algorithm implemented	True Positive	False Negative	Precision	Recall	Class
J48 Classification Algorithm	0.920	0.053	0.911	0.920	Malignant
	0.947	0.080	0.952	0.947	Begnin
REPTree Classification Algorithm	0.925	0.059	0.903	0.925	Malignant
	0.941	0.075	0.955	0.941	Begnin
RandomForest Classification Algorithm	0.929	0.036	0.938	0.929	Malignant
	0.964	0.071	0.958	0.964	Begnin
RandomTree Classification Algorithm	0.943	0.059	0.905	0.943	Malignant
	0.941	0.057	0.966	0.941	Begnin

The performance of any classification algorithm is extremely depending on the nature of training dataset used. In WEKA tool, confusion matrices which are generated

 DOI : 10.18231/2454-9150.2018.0636

after simulation of classification algorithm are very constructive for evaluating classifiers. The columns in confusion matrix represent the predicted classification classes, and the rows represent the actual class.

Table 5: Confusion Matrix of each classifier used in an implementation

Evaluation Criteria for Classification Algorithm implementation	Malignant	Beginn	Class
J48 Classification Algorithm	195	17	Malignant
	19	338	Beginn
REPTree Classification Algorithm	196	16	Malignant
	21	336	Beginn
RandomForest Classification Algorithm	197	15	Malignant
	13	334	Beginn
RandomTree Classification Algorithm	200	12	Malignant
	21	336	Beginn

Based on the above Figure 4, 5 and Table 2, we noticeably see that the maximum accuracy is 95.0791 % for RandomForest and the minimum accuracy is 93.4974 % for REP tree. By applying different classification algorithm, we found that approximately 534 instances out of 569 instances are found to be correctly classified with the highest score of 536 instances compared to 531 instances, which is the lowest score. The time taken to build the classification model is also an essential parameter. From Table 2, we say that RandomForest algorithm requires the minimum time which is around 0.07 sec and REPTree require maximum time which is around 0.01sec.

V. CONCLUSION

Different fields can be helped using various data mining algorithms for decision making. In this study, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is further collected from UCI machine learning repository having 569 records and ten different attributes. For better clinical decision in breast cancer patients, data mining model is constructed. However, this model can further be used for other diseases in order to protect patients.In the future, these results can be applied to create a proposal for breast cancer patients because breast cancer patients are normally not identified until a later stage of the disease or the development of complications.

REFERENCES

[1] http://www.who.int/cancer/detection/breastcancer/en/index1.html

[2] N. Abdelhamid, A. Ayesh, W. Hadi Multi-label rules algorithm based associative classification Parallel Process. Lett., 24 (01) (2014), pp. 1450001-14500021

[3] B. Ma, H. Zhang, G. Chen, Y. Zhao, B. Baesens Investigating associative classification for software fault prediction: an experimental perspective Int. J. Softw. Eng. Knowl. Eng., 24 (1) (2014), pp. 61-90

[4] S. Taware, C. Ghorpade, P. Shah, N. Lonkar, M. B k Phish detect: detection of phishing websites based on associative classification (AC) Int. J. Adv. Res. Comp. Sci. Eng. Inf. Technol., 4 (3) (2015), pp. 384-395

[5] Am J Cancer Res. 2017; 7(3): 610–627.Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges Rozita Jamili Oskouei,1 Nasroallah Moradi Kor,2,3 andSaeid Abbasi Maleki4

[6] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine. 1996;17:37–54.

[7] Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. 2012. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 32, 569 (2012), 2.

[8] Abien Fred M. Agarap. 2018. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. In ICMLSC 2018: ICMLSC 2018, The 2nd International Conference on Machine Learning and Soft Computing, February 2–4, 2018, Phu Quoc Island, Viet Nam. ACM, New York, NY, USA

[9] 9. http://www.saedsayad.com/decision_tree.htm accessed on 4/7/2018

[10] Volume 3, Issue 6, June 2013ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[11] http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/ accessed on 17/7/2018

[12] http://shodhganga.inflibnet.ac.in/bitstream/10603/72960/12/12_chapter%203.pdf accessed on 4/7/2018

[13] Anuj Sharma and Shubhamoy Dey. Article: Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications ACCTHPCA(3):15-20, July 2012.

[14] Sushil kumar Kalmegh ,IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015,ISSN 2348 – 7968

[15] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomTree.html accessed on 17/7/2018

 DOI : 10.18231/2454-9150.2018.0636