Comparing the Performance of Data Mining Tools: WEKA and DTREG

Neha Sharma, Hari Om

Abstract— The objective of the paper is to compare two data mining tools on the basis of various estimation criteria. The data mining tools which are evaluated are WEKA and DTREG. These tools are used to build multilayer perceptron which is a data mining model to predict the survivability of the oral cancer patients. Oral cancer database is considered as it is estimated to be 8th most common cancer worldwide and extremely grave problem in India as well. Early detection is the only way to prevent the disease and reduce this burden. Dtreg is a proprietory data mining tool whereas weka is an open source. Classification accuracy of multilayer perceptron model developed using dtreg is 70.05% and using weka is 59.70%. 10-fold cross-validation method is used for validation by dtreg and stratified cross validation is used by weka. The data mining tool dtreg has demonstrated better results in terms of true negative, false negative, specificity, recall and area under ROC curve. However, weka displays better results in terms of true positive, false positive, precision and f-measure. Analysis run time of dtreg is less than weka and the report generated by dtreg is also more expressive and descriptive in comparison to weka, which makes dtreg a better data mining tool for multilayer perceptron models.

Index Terms—Data Mining, Model, Multilayer Perceptron, Oral Cancer, Weka, DTREG, Data Mining Tool

1 INTRODUCTION

Data mining is an effective new innovation with incredible potential to help organizations understand the most critical data in their data warehouses [1,2]. Computer software programs or packages that enable the extraction and identification of patterns from stored data are popularly known as data mining tools. Data mining tools predict future trends and behaviors allowing businesses to make proactive knowledge driven decisions [3,4,5]. They scour database for hidden patterns finding predictive information that experts may miss because it lies outside their expectations [6]. There are various data mining tool available that typically serve as a software interface which interacts with a large database containing customer or other important data. Data mining is extensively used by companies and public bodies for marketing, detection of fraudulent activity, and scientific research [1]. Gradually, the medical community has also understood the importance of data mining and intends to extract meaningful pattern and knowledge from healthcare data collected over the long period of time [7]. The results of these methods may possibly offer clinical medicine structure and build the model for medical issues that can provide extraordinarily benefit to healthcare industry

In this paper, we attempt to apply data mining to oral cancer database as it was estimated to be 8th most common cancer worldwide in 2000, with approximate 267,000 new cases and 128,000 deaths. Oral cancer has the greatest burden mainly in developing countries [8]. Moreover oral cancer data is of significant public health importance in India, as the public health officials, private hospitals, and academic medical centers within country have recognized oral cancer as a grave problem [9]. Efforts to increase the body of literature on the knowledge of the disease etiology and regional distribution of risk factors have begun gaining momentum. However, early detection is the only way by which we can prevent the disease and reduce this burden. In light of this, we construct a data mining model using multilayer perceptron to predict the survivability of oral cancer patients. Multilayer perceptron model is implemented with the help of two different tools and subsequently the performance of the tools is compared.

The first tool used is DTREG (pronounced D-T-Reg) which is a predictive modeling software that builds classification and regression decision trees, neural networks, support vector machine (SVM), GMDH polynomial networks, gene expression programs, K-Means clustering, discriminant analysis and logistic regression models that can describe data relationships [10]. The DTREG can be used to predict values for future observations and also has full support for time series analysis. It accepts a dataset in the form of table containing number of rows, whose columns represent attributes/variables. One of the variables is the "target variable" whose value is to be modeled and predicted as a function of the "predictor variables". The DTREG analyses the data and generates a model showing how best it predicts the values of target variable based on the values of predictor variables [10]. The second tool used is WEKA3.7.9 which is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering and association rule extraction. It is a Java based open source tool created by researchers at the University of Waikato in New Zealand [11].

- *Neha Sharma is currently pursuing PhD in computer science and engineering in Indian School of Mines, Dhanbad, India, PH-09923602490. E-mail: nvsharma@rediffmail.com*
- *Dr. Hari Om is currently working as Assistant Professor in computer science and engineering in Indian School of Mines, Dhanbad, India, PH-09430768169. E-mail: hariom4india@gmail.com*

The rest of the paper is organized as follows: Section 2 reviews related literature and section 3 covers the information about oral cancer. In section 4, data mining model is discussed briefly. Section 5 presents the experimental results of both the tools and section 6 compares the performance of the tools. Section 7 concludes the paper.

2 Literature Review

Literature survey is done to comprehend the research work carried out by various authors in the field of oral cancer to understand the gravity of the disease and also to review various application of data mining adopted by eminent researchers to extend the benefits to medical fraternity. Yeole et al.[12] have studied the data on survival of oral cancer patients registered by the Bombay population-based cancer registry in India, during 1992-1994. They found that the overall 5-year observed and relative survival rates were 30.5% and 39.7%, which declines further with advancing age and advanced clinical stages. Five-year observed survival was 59.1% for localised cancer, 15.7% for cancers with regional extension and 1.6% for those with distant metastasis. Those with tongue, buccal mucosa and retromolar trigone cancers had poor survival rates. Their study clearly shows that detecting oral cancer in early stages, when these are amenable to single modality therapies, offers the best chance of long-term survival. Misra et al. [13] performed a prospective clinic-histological study of premalignant and malignant lesions of the oral cavity, and compared it with a 10-year retrospective data, especially in terms of incidence, age distribution, personal habits, site and type of lesion. The study shows that the histology along with a detailed clinical workup is found to be a useful, reliable and accurate diagnostic technique for lesions of the oral cavity. An increase in premalignant lesions in the prospective study, associated with increased pan masala intake is alarming and needs to be taken care, as suggested by authors.

Anurag Upadhayay et al. [14] present a snapshot of various forces driving the e-health applications and challenges for their widespread adoption. They also attempted to provide a conceptual framework for successful deliverance of e-health services using two Decision Tree technique (C4.5 and C5.0). Singh et al. [15] have applied the apriori algorithm with transaction reduction on the data of cancer symptoms by considering five different types of cancer to find the symptoms that help the cancer to spread and also the cancer type that spreads faster. Srikant et al. [16] have considered the problem of integrating constraints in the form of boolean expression that appoint the presence or absence of items in rules. Swami et al. [17] discuss the multidimensional association rules and the model for smoking habits in order to take some preventive measures to reduce various habits of smoking in youths. Milovic et al. [18] discuss the applicability of data mining in healthcare and explain how the patterns can be used by physicians to determine diagnoses, prognoses, and apply for patients in healthcare organizations. Nahar et al. [19] discuss the significant prevention factors for a particular type of cancer. Prevention factor dataset was constructed and then three association rule mining algorithms: Apriori, Predictive apriori, and Tertius algorithms have been applied. Experimental results illustrate that the Apriori is the most useful association rule-mining algorithm for discovery of the prevention factors. Kaladhar et al. [20] predict oral cancer survivability using the CART, Random Forest, LMT and Naïve Bayesian classification algorithms, which classify the cancer survival using 10 fold cross validation and training dataset. Among these algorithms, the Random Forest technique classifies dataset of cancer survival more accurately as compared to other methods.

3 Oral Cancer

Oral malignancy is a heterogeneous assembly of tumors rolling out from diverse parts of the oral cavity, with distinctive predisposing factors, prevalence, and treatment outcomes. Oral tumor is one of the ten most incessant diseases worldwide with a yearly occurrence of over 300,000 cases, of which 62% arise in advancing nations [21]. There is a huge contrast in the rate of oral tumor in diverse regions of the worlds. The age-adjusted rates of oral tumor differ from over 20 for every 100,000 population in India, to 10 for every 100,000 in the U.S., and less than 2 for every 100,000 in the Middle East [22]. In contrast with the U.S. population, where oral cavity malignancy represents only about 3% of malignancies, it accounts for over 30% of all growths in India. The variation in incidence and pattern of oral cancer is due to regional differences in the prevalence of risk factors. But as oral cancer has well-defined risk factors, these may be modified – giving real hope for primary prevention.

The main clinician's issue is to separate malignant lesions from a nearly infinite amount of other poorly characterized, questionable, and crudely comprehended sores that additionally occur in the oral cavity. Most oral sores are benign, yet many have a manifestation that may be effectively befuddled with threatening lesions and some are considered premalignant because they have been statistically correlated with subsequently cancerous changes [23]. On the other hand, some malignant lesions seen in an early stage may be mistaken for a benign. Early carcinomas are presumably asymptotic and ensuing signs are regularly misjudged in light of the fact that they imitate numerous benevolent lesions and the distress is negligible. Professional consultation is thus often delayed, increasing the chance for local spread and regional metastases. Stress must be placed on gaining access to high risk individuals for periodic oral examinations and efforts to increase the educational skill of primary health care providers in recognizing this problem. Squamous cell carcinoma accounts for 90% of the total number of malignant oral lesions. Therefore, the problem of oral cancer is primarily that of pathogenesis, diagnosis and management of squamous cell carcinoma originating from oral muscular surface [24]. Oral tumor presenting with nodal metastases would appear to have a less favorable prognosis [25].

4 Data Mining Models

Data mining, an analytic process, has been designed to explore

data in search for consistent patterns and/or systematic relationships between variables and then to validate the findings by applying the detected patterns to new subsets of data. Its ultimate goal is prediction. Predictive data mining is the most common type of data mining and one that has the most direct business applications. In this paper, multilayer perceptron predictive model is built for predicting the survivability of oral cancer patients and is implemented using two data mining tools ie. DTREG and WEKA.

4.1 Multilayer Perceptron Model (MLP)

The Artificial Neural Network is one of the most commonly used models based on human cognitive structure. Some different types of the Artificial Neural Network (multi-layer perception, Radial Basis Function Neural Network and Kohonen's self-organizing map) are proposed to solve non-linear problem by learning. When used without qualification, the terms ―Neural Network (NN) and ―Artificial Neural Network (ANN) usually refer to a Multilayer Perceptron Network (MLP). The diagram shown in figure 1 illustrates a perceptron network with three layers. This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons. There is one neuron in the input layer for each predictor variable (x1…xp). In the case of categorical variables, N-1 neurons are used to represent the N categories of the variable.

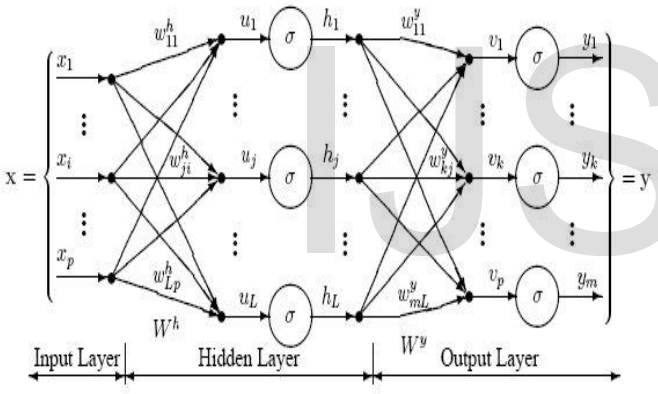

Fig 1. Three Layered Multilayer Perceptron Model

The network diagram shown above is a full-connected, three layered, feed forward, perceptron neural network. Fully connected network means that the output from each input and hidden neuron is distributed to all of the neurons in the following layer. Feed forward means that the values only move from input to hidden to output layers; no values are fed back to earlier layers. When there is more than one hidden layer, the output from one hidden layer is fed into the next hidden layer and separate weights are applied to the sum going into each layer.

5 Experimental Results

The database for this case study has been created by collecting the data related to oral cancer through a retrospective chart review in non-randomized or non-probabilistic method. The complete process of data preparation, data integration and data cleaning was strictly adhered to create the database of oral cancer patients [26]. The database has 1025 oral cancer patients' record which is described with the help of 35 attributes. The oral cancer data is initially stored in MS Excel sheet, which is converted into comma separated values (.csv) file format which is the desirable format for DTREG tool and subsequently saved it as attribute relation file format (.arff) which is the format accepted by the WEKA tool. Performance estimation of multilayer perceptron model built using two different data mining tools is presented in this section.

5.1 Building MLP Model using DTREG Tool

The attribute 'survival' is considered as a target variable. Classification technique is used for analysis, category weights are distributed over entire data file, misclassification costs are equal and variable weights are also equal. Number of layers is 3(input, hidden and output). Hidden layer and Output layer activation function used in this model is Logistic. Cross validation method with 10 folds is used for validation whereas network size evaluation is performed using 4-fold cross-validation. To build MLP model, the prior probability for the category survival =D is 0.4029268 and for the category survival = A is 0.5970732. The architecture of multi-layer perceptron network and training Statistics of the network is presented in Table [1] and Tables [2] respectively.

TABLE 1
ARCHITECTURE OF MLP MODEL USED BY DTREG TOOL

Layer	Neurons	Activation	Min. Weight	Max. Weight
Input	48	Passthru	-	-
Hidden 1	3	Logistic	-1.277e+000	1.468e+00
Output	2	Logistic	-9.397e-001	1.157e+00

TABLE 2
TRAINING STATISTICS OF MLP MODEL USED BY DTREG TOOL

Process	Time	Evaluations	Error
Conjugate gradient	00:00:00.2	142,065	1.1888e-001

The performance of data mining tool DTREG is evaluated on the basis model estimation criteria which are presented as follows:

1. Misclassification Table

If the target variable is categorical and a classification tree is build, then a misclassification summary table presents the number of rows with a particular category that were misclassified by the tree, for both training as well as validation dataset. Misclassification table for the model is presented in Table [3] for training data and in Table [4] for validation data.

TABLE 3
MISCLASSIFICATION TABLE FOR TRAINING DATA BY DTREG TOOL

	Actual		Misclassified			
Category	Count	Weight	Count	Weight	%	Cost
A	612	612	212	212	34.641	0.346
D	413	413	95	95	23.002	0.230
Total	1025	1025	307	307	29.951	0.300
Overall accuracy = 70.05%						

TABLE 4
MISCLASSIFICATION TABLE FOR VALIDATION DATA BY DTREG TOOL

	Actual		Misclassified			
Category	Count	Weight	Count	Weight	%	Cost
A	612	612	236	236	38.562	0.386
D	413	413	74	74	17.918	0.179
Total	1025	1025	310	310	30.244	0.302
Overall accuracy = 69.76%						

2. Confusion matrix

Confusion Matrix provides detailed information about how data rows are classified by the model. The numbers in the diagonal cells are the weights for the correctly classified cases where the actual category matches the predicted category. The off-diagonal cells have misclassified row weights. Confusion Matrix for both training and validation data is shown in Table [5].

TABLE 5
CONFUSION MATRIX FOR MLP MODEL BY DTREG TOOL

	Testing Data		Validation Data	
Actual Category	Predicted Category		Predicted Category	
	A	D	A	D
A	400	212	376	236
D	95	318	74	339

3. Sensitivity and Specificity

Sensitivity means probability that the algorithms can correctly predict non malignancy and specificity means probability to correctly predict malignancy. Survival = D is considered as a positive and Survival = A is negative for the developed model. The patients who are predicted as malignant among malignant patients are True Positive (TP) cases. The patients who are predicted as non malignant among non malignant patients are True Negative (TN) cases. The patients who are predicted as non malignant among malignant patients are False Positive (FP) cases. The patients who are predicted as malignant among non malignant patients are False Negative (FN) cases.

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

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

The detail regarding sensitivity and specificity of the model along with positive/negative ratio, true positive (TP), true negative (TN), false positive (FP), false negative (FN), geometric mean of sensitivity and specificity, positive predictive value (PPV), negative predictive value (NPV), geometric mean of ppv and npv, precision, recall, F-measure and area under Receiver Operating Characteristics (ROC) curve for training and validation data for the models is shown in Table [6].

4. Probability Calibration

The probability calibration report generated by the tool shows how the predicted probability of a target category is distributed and provides a means for gauging the accuracy of predicted probabilities. Probability calibration for Survival = D and Survival = A is same. Average weighted probability error for training data = 0.040631. Average weighted squared probability error for training data = 0.050201. Average weighted probability error for validation data = 0.058048. Average weighted squared probability error for validation data = 0.062428.

5. Probability Threshold

The probability threshold report generated by the tool provides information about how different probability thresholds would affect target category assignments. Area under ROC curve (AUC) for training data = 0.751660. Threshold to minimize misclassification for training data = 0.459764. Threshold to minimize weighted misclassification for training data = 0.459764. Threshold to balance misclassifications for training data = 0.516274. Area under ROC curve (AUC) for test data = 0.733567. Threshold to minimize misclassification for test data = 0.493657. Threshold to minimize weighted misclassification for test data = 0.493657. Threshold to balance misclassifications for test data = 0.544517.

TABLE 6
SENSITIVITY AND SPECIFICITY FOR MLP MODEL BY DTREG TOOL

	Training Data	Validation Data
Positive/ Negative ratio	0.6748	0.6748
True positive (TP)	318 (31.02%)	339 (33.07%)
True negative (TN)	400 (39.02%)	376 (36.68%)
False positive (FP)	212 (20.68%)	236 (23.02%)
False negative (FN)	95 (9.27%)	74 (7.22%)
Sensitivity	77.00%	82.08%
Specificity	65.36%	61.44%
Geometric mean of sensitivity-specificity	70.94%	71.01%
Positive predictive value (PPV)	60.00%	58.96%
Negative predictive value (NPV)	80.81%	83.56%
Geometric mean of PPV and NPV	69.63%	70.19%
Precision	60.00%	58.96%
Recall	77.00%	82.08%
F-Measure	0.6744	0.6862
Area under ROC curve	0.769	0.739

6. Lift and Gain

The lift and gain table is a useful tool for measuring the value of a predictive model. Lift and gain values are especially useful when a model is being used to target (prioritize) marketing efforts. The basic idea of lift and gain is to sort the predicted target values in decreasing order of purity on some target category and then compare the proportion of cases with the category in each bin with the overall proportion. The lift and gain for the MLP model for training and validation data is presented in Table [7].

TABLE 7
LIFT AND GAIN CHART FOR MLP MODEL BY DTREG TOOL

Lift and Gain	Training Data		Validation Data	
	Survival			
	A	D	A	D
Average gain	1.278	1.362	1.270	1.308
Percent of cases with Survival	59.71%	40.29%	59.71%	40.29%

7. Analysis run time

The time taken by the tool to build the model is 00:05.37 sec.

5.2 Building MLP Model using WEKA Tool

The scheme adopted by the tool to build the MLP model is weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a. The Table [8] shows that model has correctly classified 677 instances whereas 348 instances were incorrectly classified.

TABLE 8
INSTANCE CLASSIFICATION FOR MLP MODEL BY WEKA TOOL

	No. of Instances	Percentage
Correctly Classified Instances	677	66.05%
Incorrectly Classified Instances	348	33.95%

Stratified cross-validation method is used for confirmation of model. In stratified k-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. This is a case of a dichotomous classification, which means that each fold contains roughly the same proportions of the two types of class labels. The other measures used by the tool to evaluate the model are as follows:

1. Kappa coefficient

It is a statistical measure for qualitative (categorical) items regarding inter-rater agreement or inter-annotator agreement [27]. It is a robust measure in comparison to simple percent agreement calculation since kappa coefficient takes into account the agreement occurring by chance [28][29].

2. Mean absolute

It is a risk function corresponding to the expected value of the squared error loss.

3. Root mean squared error

It is a frequently used measure of the differences between values predicted by a model and the values actually observed.

4. Relative absolute error

It takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

5. Root Relative Squared Error

By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

Table [9] presents the summary of the performance estimation carried out by WEKA.

TABLE 9
PERFORMANCE ESTIMATION FOR MLP MODEL BY WEKA TOOL

Measure	Value
Kappa Statistics	0.2966
Mean Absolute Error	0.3478
Root Mean Squared Error	0.5504
Relative Absolute Error	72.2712 %
Root Relative Squared Error	112.2057 %

Detailed accuracy of the model by class regarding true positive rate (TP), false positive rate (FP), precision, recall, f-measure, receiver operating character area (ROC Area) and precision recall curve area (PRC Area) is presented in Table [10] and confusion matrix is presented in Table [11].

TABLE 10
DETAILED ACCURACY BY CLASS FOR MLP MODEL BY WEKA TOOL

	Alive	Dead	Weighted Average
True Positive Rate	0.709	0.588	0.660
False Positive Rate	0.412	0.291	0.363
Precision	0.719	0.577	0.662
Recall	0.709	0.588	0.660
F-Measure	0.714	0.583	0.661
ROC Area	0.702	0.702	0.702
PRC Area	0.784	0.555	0.692

TABLE 11
CONFUSION MATRIX FOR MLP MODEL BY WEKA TOOL

Classified as	Alive	Dead
Alive	434	178
Dead	170	243

6. Analysis run time

The time taken by tool to build the model is 00:284.55 sec.

6 COMPARISON OF DATA MINING TOOLS

In this section, both the tools (DTREG and WEKA) used for developing multi-layer perceptron model, are compared. The first criteria on which both the tools are evaluated are receiver operating characteristic (ROC). ROC is a graph that demonstrates the performance of a binary classifier model. It is created by plotting true positives out of the total actual positives (TPR = true positive rate) and false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. TPR is sensitivity and FPR is specificity. The ROC is also known as a relative operating characteristic curve as it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes [30]. ROC analysis allows tools to choose perhaps optimal models and to discard others independently from the cost context or the class distribution. ROC analysis has been used profusely in data mining research in a direct and natural way to cost/benefit analysis of diagnostic decision making. Figure 2 and Figure 3 present the graphs generated to represent ROC Area by DTREG and WEKA respectively.

The second estimation criteria is probability threshold report, which is a graphical depiction of the different probability thresholds affecting target category assignments. The report presents the tradeoff between impurity and loss as the probability threshold is varied. The report is generated only for a classification analysis is performed with two target categories. Usually the category with the highest probability is selected as the predicted category. Figure 4 and Figure 5 present the

Threshold Chart generated by DTREG and WEKA respectively.

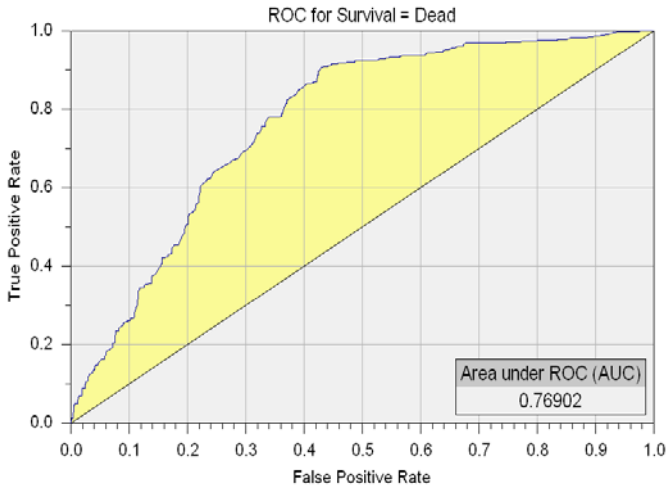

Fig 2. ROC Curve for MLP Model build by DTREG

Fig 3. ROC Curve for MLP Model build by WEKA

Fig 4. Prodability Threshold for MLP Model build by DTREG

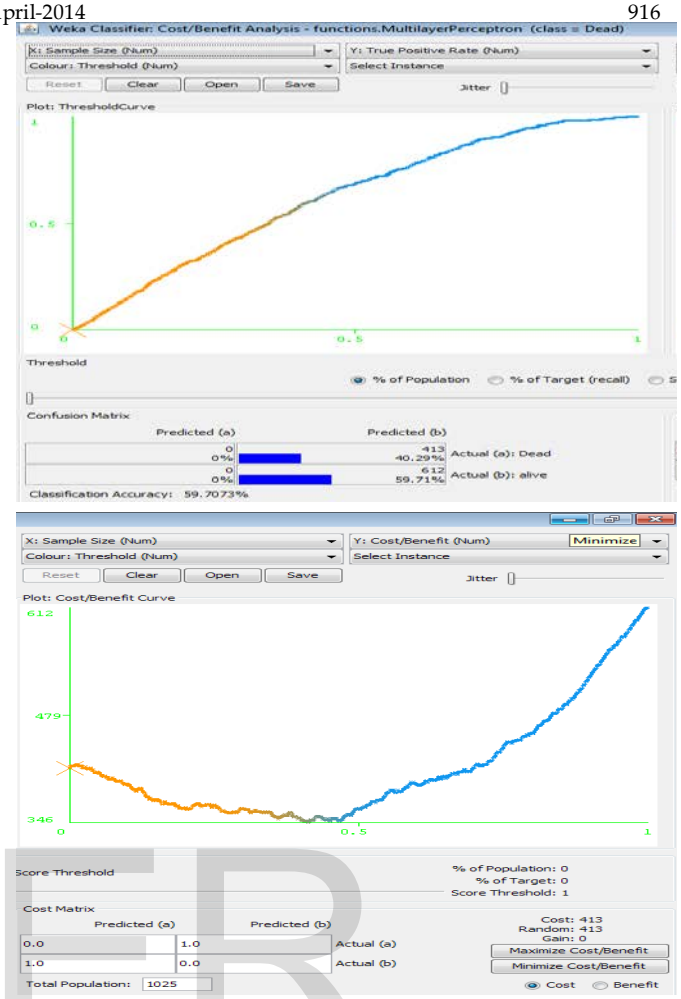

Fig 5. Prodability Threshold for MLP Model build by WEKA

DTREG and WEKA are also compared on various model estimation criteria and is presented in Table [12] as well as figure 6.

TABLE 12
COMAPRISON OF DATA MINING TOOLS: DTREG AND WEKA

	DTREG	WEKA
Accuracy	70.05%	59.70%
True positive (TP)	31.02%	23.70%
True negative (TN)	39.02%	42.35%
False positive (FP)	20.68%	17.36%
False negative (FN)	9.27%)	16.58%
Sensitivity	77.00%	58.85%
Specificity	65.36%	70.91%
Precision	60.00%	66.20%
Recall	77.00%	70.90%
F-Measure	0.6744	0.714
Area under ROC curve	0.769	0.702

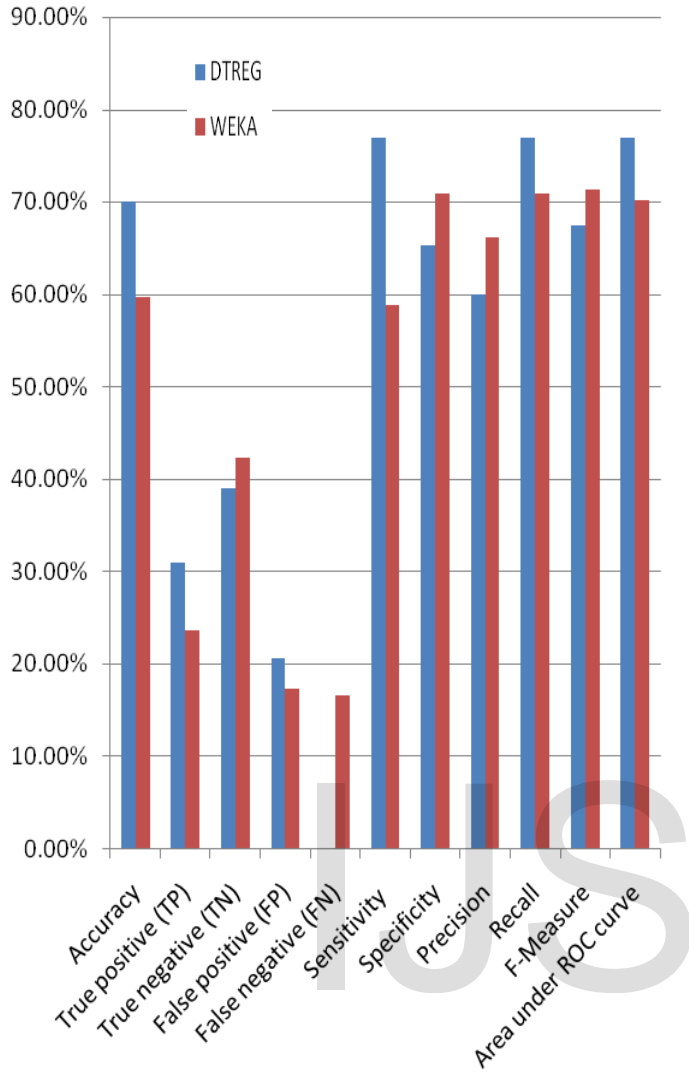

Fig 6 Comparison of DTREG and WEKA

DTREG took 00:05.37 sec and WEKA took 00:284.55 sec to build the MLP model.

7 Conclusions

The data mining tool DTREG has demostrated slightly better results in terms of accuracy, true negative, false negative, specificity, recall and area under ROC curve. However, WEKA displays better results in terms of true positive, false positive, precision and f-measure. Also, WEKA took more time in comparison to DTREG for building multilayer perceptron model and generating the analysis report. WEKA is a open source tool whereas DTREG is a licensed tool. After comparing on the basis of various estimation criteria, it is observed that DTREG is a better tool. Our future work shall include using more tools and compare their performance with DTREG and WEKA.

References

[1] J. Han and M. Kamber, "Data Mining , Concepts and Techniques", Morgan Kaufmann , Third Edition, 2012.

[2] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence (AAAI-AI Magazine), 1996, pp. 37–54 .

[3] Data Mining Curriculum, ACM SIGKDD, 2006-04-30.

[4] C. Clifton, Encyclopædia Britannica: Definition of Data Mining, 2010.

[5] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2009.

[6] M. E. Yahia and M. E. El-taher, "A New Approach for Evaluation of Data Mining Techniques", International Journal of Computer Science Issues, September 2010, Vol. 7, Issue 5.

[7] C. S. Shital, K. Andrew, A. Michael and O. Donnell, "Patient recognition data mining model for BCG-plus inerferon immunotherapy bladder cancer treatment", Computers in Biology and Medicine, 2006, Vol. 36. pp.634–655.

[8] K.P. Exarchos, G.Rigas, Y.Goletsis and D.I. Fotiadis, "Modelling of Oral Cancer Progression Using Dynamic Bayesian Networks", Data Mining for Biomarker Discovery, Springer Optimization and its Applications, 2012, pp. 199-212.

[9] K. R. Coelho, "Challenges in Oral Cancer Burden in India", Journal of Cancer Epidemiology, 2012, Vol 2012, Article ID 701932, 17 pages.

[10] www.dtreg.com

[11] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tool and Techniques, Second Edition, Elsevier.

[12] B. Yeole, A. V. Ramanakumar and R Sankaranarayanan, "Survival from oral cancer in Mumbai (Bombay), India.", December 2003;Vol 14, issue 10, pp.945-52.

[13] V. Misra, P.A. Singh, N. Lal, P. Agarwal and M. Singh, "Changing pattern of oral cavity lesions and personal habits over a decade: hospital based record analysis from allahabad", Indian J Community Med., October 2009, Vol 34, Issue 4, pp. 321-5, doi: 10.4103/0970-0218.58391.

[14] A. Upadhayay, S. Shukla and S. Kumar, "Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set" , International Journal of Computer Science & Communication Networks, 2012, Vol 3, isuue 1, pp.64-68

[15] S. Singh, M. Yadav and H. Gupta, "Finding the Chances and Prediction of Cancer through Apriori Algorithm with Transaction Reduction", International Journal of Advanced Computer Research. (ISSN (print): 2249-7277 ISSN (online): 2277-7970, June 2012, Vol 2, issue 2, pp. 23-28.

[16] R. Srikant, Q. Vu and R. Agrawal, "Mining association rules with item constraints", Proceeding KDD97, 1997, pp. 67-73.

[17] S. Swami, R. S. Thakur and R.S. Chandel, "Multi- dimensional Association Rules Extraction in smoking Habits Database", Int. J. Advanced Networking and Applications, 2011, Vol 03, issue 03, pp. 1176-1179.

[18] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining" International Journal of Public Health Science, December 2012, Vol 01, issue 02, pp. 69-78.,

[19] J. Nahar, S. T. Kevin, A. B. M. S. Ali and Y. P. Chen, "Significant cancer prevention factor extraction: An association rule discovery approach", J Med Syst, Springer, October 2009, DOI 10.1007/s10916-009-9372-8.

[20] D.S.V.G.K Kaladhar, B. Chandana and P. B. Kumar, "Predicting cancer survivability using Classification algorithms", International Journal of Research and Reviews in Computer Science, April 2011, Vol 02, issue 02, pp.340–343.

[21] D.M. Perkin and E. Lara, "Estimates of the World wide frequency of sixteen major cancers", 1980, vol 41, pp 184-197.

[22] R. Sankaranarayan, E. Masuyer, R. Swaminathan, J. Ferley and S. Whelan, Head and neck cancer: a global perspective on epidemiology and prognosis. Anticancer Res 18:4779-86, 1998

[23] American Cancer Society, Cancer Facts and figures, Atlanta (GA), the society, 1996.

[24] H.P. Sobin, International union Against Cancer TNM classification of malignant tumors, 4th Ed, 2nd revision" LH Editors, Berlin, Springer-Verlag; 1992.

[25] Cancer Research Capign, Oral cancer, Fact sheet vol 14, no 1, 1990.

[26] N. Sharma and Hari Om, "Framework for early detection and prevention of oral cancer using data mining", International Journal of Advances in Engneering and Technology, September 2012, Vol 4, Issue 2, pp. 302-310.

[27] J. Carletta, 1996 Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), pp. 249–254.

[28] J. Strijbos, R. Martens, F. Prins, W. Jochems, 2006. "Content analysis: What are they talking about?". Computers & Education 46: 29–48. doi:10.1016/j.compedu.2005.04.002.

[29] J.S. Uebersax, 1987. "Diversity of decision-making models and the measurement of interrater agreement" (PDF).Psychological Bulletin 101: 140–146. doi:10.1037/0033-2909.101.1.140.

[30] S.A. John 1996. Signal detection theory and ROC analysis in psychology and diagnostics : collected papers, Lawrence Erlbaum Associates, Mahwah, NJ.