Proceedings of International Conference on Advances in Computer Technology and Management (ICACTM)
In Association with Novateur Publications IJRPET-ISSN No: 2454-7875
ISBN No. 978-81-921768-9- 5
February, 23rd and 24th, 2018

KIDNEY DISEASE DETECTION IN INDIAN PATIENTS IN AN EARLY STAGE USING WEKA TOOL

HARSHALI PATIL
Associate Professor, MET ICS, Bandra(W) & Research Student, UDCS, University of Mumbai, Mumbai, India
[1]harshalip_ics@met.edu

MANISHA DIVATE
Department of Computer Science,University of Mumbai, Mumbai, India,
divate.manisha.79@gmail.com

ABSTRACT
Systems like health care, educational organization, financial institution etc., and produce huge amount of data. For decision making, there is a need of processing a raw data generated by these systems automatically. For example there is need of decision support system in health care to decrease the hospitalization rate. The quality of service provided will be low if the rate of hospitalization is more [1].
Machine learning (ML) is involved in the automatic identification of hidden patterns if there exists any in a given raw data. ML helps discovering knowledge from within a raw data that data mining aims at. Such knowledge plays a vital role in decision making. Mathematical and statistical techniques, namely, regression, classification, clustering, and association rule mining are generally employed [2]. Several software are available. We here discuss features of one such, WEKA, a collection of ML tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The WEKA library can be used directly or one can embed the functions in Java code.
The case presented here is from the health care sector: Diagnosis of chronic kidney disease based upon the patient's history. Record containing blood pressure, sugar, red blood cells count etc., of 400 patients has been used to build a classification model [3]. Results are compared with LMT(Logistic model tree), Random Tree, REP Tree (Reduced Error Pruning) and the accuracy is more in J48. In LMT method the results are 98%, in Random tree it is 96.5% , REP Tree method results are 96.75%whereas with J48 algorithm the result is 99%. Classification result of Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithm shows the accuracy of 76.32% and 87.70% respectively [4]. The paper concludes that J48 classification algorithm comparatively generates the best result.
KEYWORDS — Data Mining, Kidney Disease, Classification, decision tree, confusion matrix

I. INTRODUCTION

Good health discusses on a person or groups' freedom from illness. Health is therefore best understood as a vital basis for defining a person's sense of well being. In India as per survey it has been identified that 60% death rate is due to chronic dieses [5]. As per prediction, new health challenges are likely to come forward in India in next few decades. As per a possible scenario of the burden of disease (BOD) for India in the year 2020, Murry and Lopez <WorId Bank B 2000> have provided, a statistical model calculating the change in disability-adjusted life year (DALY). As per survey results of 201, 17% of urban Indians have kidney dieses [6]. The Global Burden of Disease (GBD) study 2015, identifies that chronic kidney disease is ranked 17th among the causes of deaths globally (age-standardised annual death rate of 19·2 deaths per 100 000 population) [7], and as per world health ranking in 2016 India is on 24th rank and having high death rate as 21.56% [8]. Hence the proactive stand for early stage disease detection and proper treatment has become a need. In this research paper the kidney dieses prediction is done using classification algorithms, applied on WEKA tool. The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning tool written in java. Weka was developed at the University of Waikato in New Zealand.

II. LITERATURE SURVEY

Dr. S. Vijayarani et al [9] research work is to predict kidney diseases by using Support Vector Machine (SVM) and Navie Bays. The research mainly focused finding best classification algorithm, which is done by comparing the performance of these two algorithms on the basis of its accuracy and execution time. As per the research results it is observed that the SVM has the maximum classification accuracy and Naïve Bayes requires minimum execution time.
Andrew Kusiak et al [10] research work is based on data mining approach to obtain information about interaction between the patient survival and parameters such as demographic as well as clinical parameters, medications, medical interventions, and the dialysis treatment prescription. Two different data mining algorithms, Decision tree and rough set were used knowledge extraction in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. The research results are based on data collected from four dialysis sites. The approach presented in this paper provides patients selection for clinical studies with reduced cost and efforts. Predictive algorithms are used for Patients selection.
Tommaso Di Noia et.al [11] research is a developed software tool that takes advantage of the power of artificial neural networks (ANN) to classify patients' health status which is leading to End Stage of Kidney Disease (ESKD). The tool which has been refined and made derivable both as an online web application and as an android mobile app. The developed tool is important for clinical use and it is based on the largest group worldwide

III. METHODOLOGY

Dataset: The kidney disease dataset has been used for analysis of kidney disease. This dataset contains four hundred instances and twenty five attributes are used in this comparative analysis. Out of 25 attributes 11 attributes are numeric and 14 attributes are nominal type. The attributes in the dataset are Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, Packed cell volume, White Blood Cell Count, Red Blood Cell Count, hypertension, Diabetes Mellitus , Coronary Artery Disease , appetite, Pedal Edema, Anemia, and Class. The class distribution is in two classes' chronic kidney dieses and not chronic kidney dieses; these are interpreted as "ckd" and "notckd" respectively [3].

Classification: Classification is also known as classification trees or decision trees. The decision tree it creates is a tree where each node in the tree represents a spot where a decision must be made based on the input data. It is also referred as node split point. One can move from root node to the next node and the next until leaf node is found; which tells the predicted output. Classification is a data mining algorithm which finds out the output of a new data instance. In this research paper the experimental study is conducted on various classification algorithms and best algorithm is identified for chronic kidney dieses.

Confusion matrix: A confusion matrix is a table that is usually used to describe the performance of a classification model on a set of test data for which the true values are known. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa) [16]. Sensitivity and specificity are statistical measures of the performance of a binary classification test. Sensitivity (TPR) measures the proportion of positives that are correctly identified. Specificity (TNR) measures the proportion of negatives that are correctly identified.
In general, Positive = identified and negative = rejected [17]. Therefore:

True positive (TP) = correctly identified
False positive (FP) = incorrectly identified
True negative (TN) = correctly rejected
False negative (FN) = incorrectly rejected

The sensitivity and specificity is calculated by following formulas (1) & (2)

The number of real positive cases in the data is denoted by P.
The number of real negative cases in the data is denoted by N.

Sensitivity (TPR) = TP/P = TP/(TP+FN) ----------(1)

Specificity (TNR) = TN/N =TN/(FP+TN) ----------(2)

Objectives of the paper are

1. To extract hidden patterns with classification accuracy for chronic kidney disease detection.
2. To compare the results with different Classification algorithms on the given datasets.
3. To identify the best algorithm based on the correct classification of instances for detecting the chronic kidney disease out of J48, LMT, RandomTree and REPTree algorithms used in this paper.

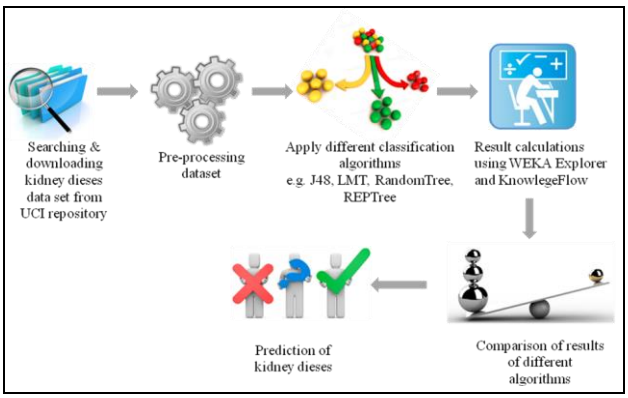

Fig.1: Flow diagram of system

The above diagram is showing the flow of methodology used in this paper. In this research work, first web search for dataset of kidney disease is done. The suitable dataset was found in UCI repository. The dataset was having missing values, so the pre-processing using unsupervised filter ReplaceMissingvalues, discretize and normalization on the attributes is applied on the dataset. Then one by one the classfification algorithms J48, LMT, RandomTree and REPTree are applied on the filtered dataset. After the comparisons of results is completed and the above mentioned objectives of this research paper are achieved.

IV. RESEARCH FINDINGS

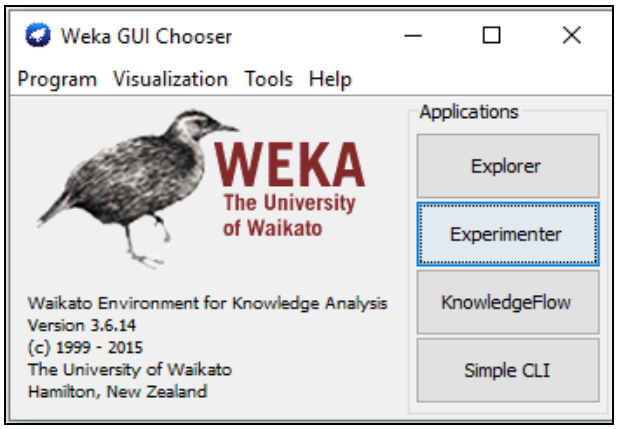

Fig.2: WEKA tool snapshot

J48

Decision tree J48 is the extension of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. C4.5 builds decision trees from a set of training data as like ID3, using the concept of information entropy. In WEKA tool, J48 is an open source Java implementation of the C4.5. the following figure depicts the KnowledgeFlow environment.

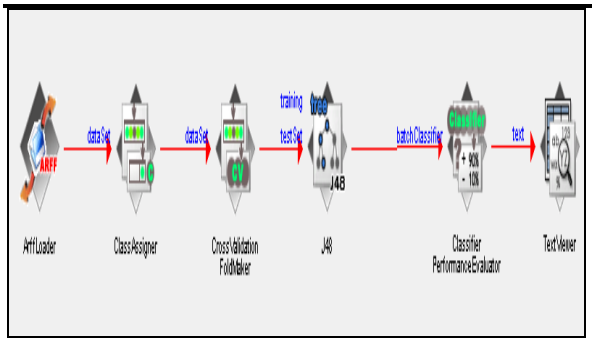

Fig.3: Classification with KnowledgeFlow environment (algorithm J48)

The decision tree generated is as follows.

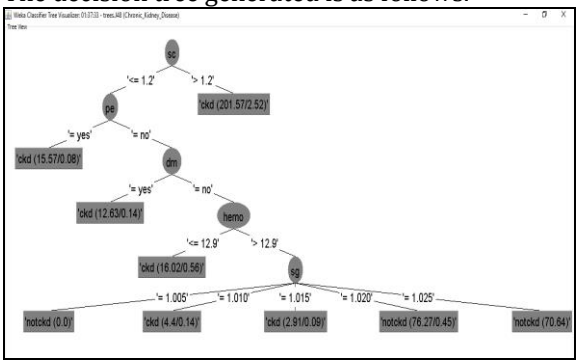

Fig.4: Decision tree using J48

Some rules generated using decision tree (J48 algoritm) are as follows

Rule 1: If serum creatinine <= 1.2 and pedal edema= 'yes' then class='ckd'

Rule 2: If serum creatinine <= 1.2 and pedal edema= 'no' and diabetes mellitus = 'yes' then class='ckd'

Rule 3: If serum creatinine <= 1.2 and pedal edema= 'no' and diabetes mellitus = 'no' and haemoglobin <=12.9 then class='ckd'

Rule 4: If serum creatinine <= 1.2 and pedal edema= 'no' and diabetes mellitus = 'no' and haemoglobin >12.9 and specific gravity = 1.005 then class='nonckd'

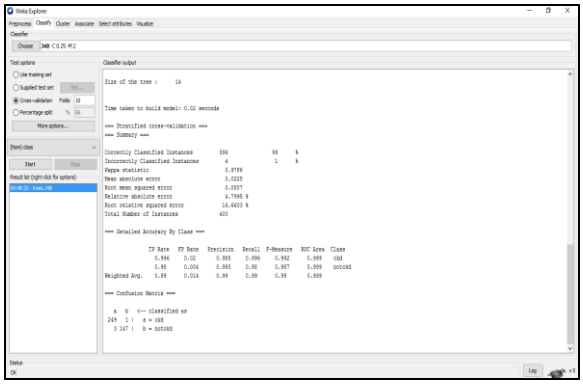

Fig.5: J48 Classifier output

After selecting Visualize threshold curve it gives a plot with FP Rate on the x-axis and TP Rate on the y-axis. Depending on the classifier used, this plot can be quite smooth or it can be fairly irregular.

The Receiver operating characteristic (ROC) curve and Area under curve (AUC) for J48 algorithm the threshold curve for chronic kidney dieses obtained as follows.

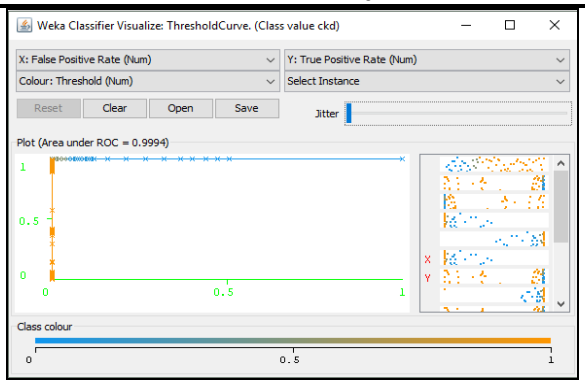

Fig.6: ROC curve and AUC

LMT

For supervised learning the most popular techniques are Tree induction methods and linear models. Both the techniques are for the prediction of nominal classes and numeric values. 'Model trees' is a approach used for predicting numeric quantities; where these two techniques are combined. Model tree is that contain linear regression functions at the leaves [12].

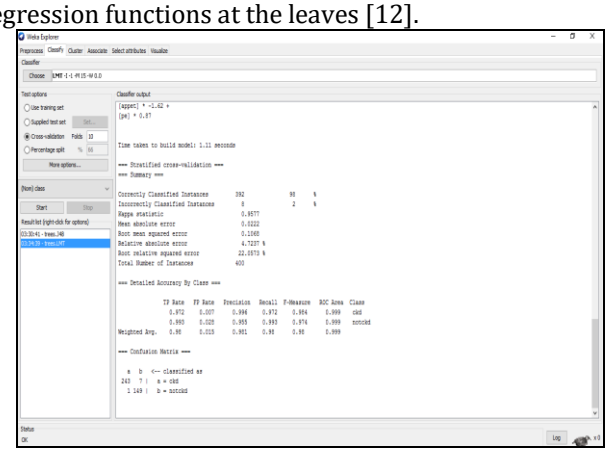

Fig.7: LMT classifier output

RandomTree

A random tree is a collection of tree predictors that is called forest. A decision tree is constructed by using random set of data. Standard tree each node is split using the best split among all variables. In RandomTree node split, a random subset of all attributes is considered at every node, and the best split for that subset is computed. The random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label as ckd [13].

The following figures shows the decision tree using RandomTree algorithm and classification output.

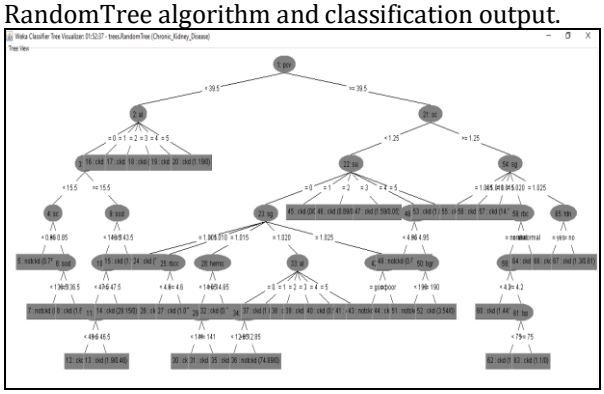

Fig.8: Decision tree using RandomTree

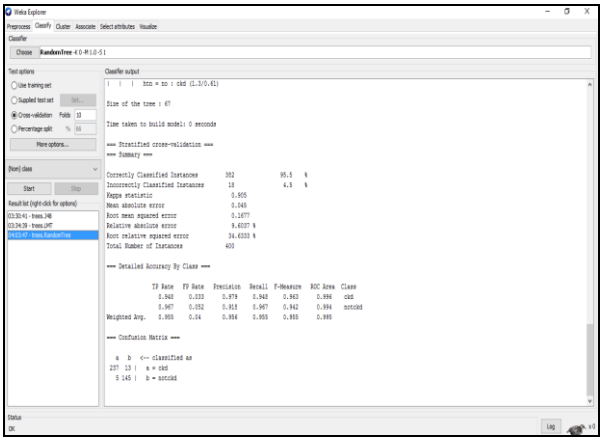

Fig.9: RandomTree classifier output

REPTree

Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree uses information gain as splitting criterion, and prunes it using reduced error pruning. It sorts the values for numeric attributes once. REPTree uses the regression tree logic and generates several trees in different iterations. Best tree is selected from all generated trees and it is considered as the representative. Mean square error is used as a measure on predictions for tree pruning [14].

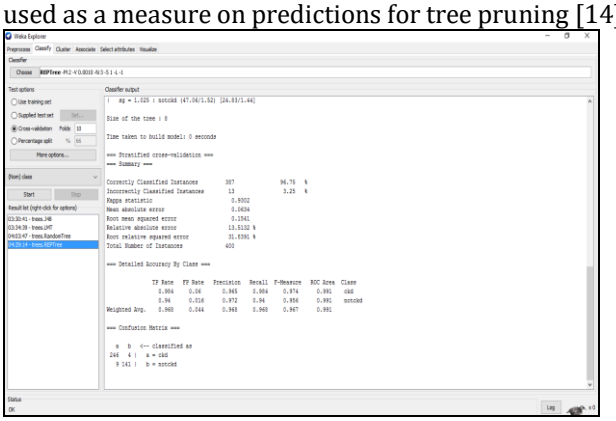

Fig.10: REPTree classifier output

Using the WEKA experimenter the result of algorithms is as shown in following figure. The average result of J48 is better than remaining three algorithms.

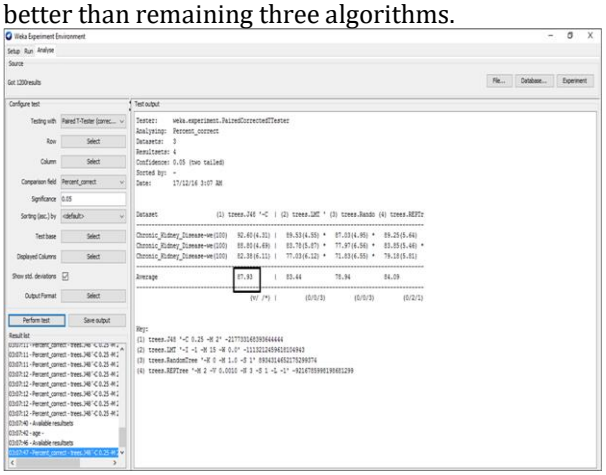

Fig.11: WEKA experimenter result

The KnowledgeFlow provides an alternative to the Explorer as a GUI to WEKA's core algorithms. The KnowledgeFlow can handle data either incrementally or in batches. The Explorer handles batch data only. KnowledgeFlow process multiple batches or streams in parallel. It has feature to view models produced by classifiers for each fold in a cross validation. KnowledgeFlow has a facility that one can easily add new components in it [15]. The KnowledgeFlow can draw multiple ROC curves in the same plot window, whereas the Explorer cannot have that feature In this example we use J48, LMT, RandomTree and REPTree as classifiers. The multiple algorithm model snapshot is as follows.

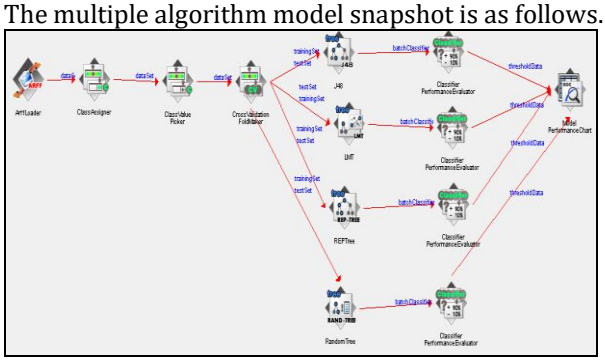

Fig.11: Model evaluation

The following model performance plot shows multiple ROC curves. The model performance chart results the threshold curve for chronic kidney dieses and we get better true positive rare for J48 algorithm.

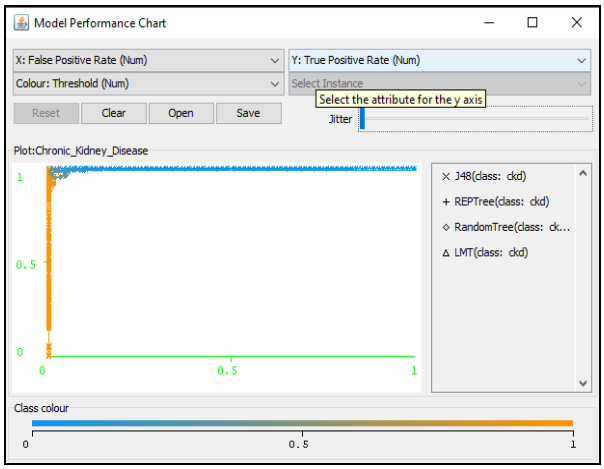

Fig.12: Multiple ROC curves - Model Performance chart

TABLE I- COMPARISON OF CLASSIFICATION ALGORITHM OUTCOME

Algorithm	**Total instances (400)**		**Mean absolute error**	**%**
	Correct	Incorrect		
J48	396	4	0.0225	99%
LMT	392	8	0.0222	98%
RandomTree	382	18	0.045	95.5%
REPTree	387	13	0.0634	96.75%

Table1 depicts that smallest error rate is in LMT but the percentage of correct classification is comparatively better in J48.

TABLE II - SENSITIVITY & SPECIFICITY ANALYSIS

Algorithm	Sensitivity(TPR)	Specificity (TNR)
J48	0.996	0.98
LMT	0.972	0.99333333
RandomTree	0.948	0.96666667
REPTree	0.984	0.94

Table 2 information states that true positive rate of J48 algorithm are better than the other three algorithms.

V. CONCLUSION

In this paper, we have proposed the comparative analysis of four different types of classification algorithms with the help of WEKA data mining tool. In the experimental study we have used the chronic kidney dataset where individual algorithm results were compared and best algorithm is selected on the basis of accuracy and time required for model evaluation. Comparative study of classification algorithm identifies that J48 is giving better results. In the experimenter we have used dataset with noise component as 5%, 10% and 15% respectively and then we performed test using paired corrected T-test on the pre-processed data, it has been observed that average of J48 test is 87.93%, for LMT it is 83.44%, for RndomTree it is 78.94% and for REPTree it is 84.09%. This experimental study will be useful for accurate prediction of chronic kidney dieses at an early stage and will be useful for the health care sector for taking proactive initiatives/decisions to cure the patient and reduce the death rate.

ACKNOWLEDGMENT

We thank our family members, friends and our mentors from computer science and distance learning department of University of Mumbai, Department of Mathematics Kirti College and MET ICS who were the constant source of inspiration. Their guidance and expertise greatly assisted the research.

REFERENCES

1) Yeh, Jinn-Yi, Tai-Hsi Wu, and Chuan-Wei Tsao. "Using data mining techniques to predict hospitalization of hemodialysis patients." Decision Support Systems 50.2 (2011): 439-448.
2) Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
3) L.Jerlin Rubini, P.Eswaran ,2015, https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
4) S. Vijayarani, Mr.S.Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS " International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015
5) Ashok Kumar, "Chronic diseases account for 60% of deaths in India: study", article in OneWorld South Asia, April 2015.
6) Durgesh Nanda Jha, "17% of urban Indians have kidney disease: Study", Article in TOI, June 2013
7) Vivekanand Jha, Gopesh Modi "Uncovering the rising kidney failure deaths in India", The Lancet global Health Journal, Volume 5, No. 1, e14–e15, January 2017
8) Kidney dieses death rate by country, http://www.worldlifeexpectancy.com/cause-of-death/kidney-disease/by-country/
9) Dr. S. Vijayarani1, Mr.S.Dhayanand," DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION", International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.
10) AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page no 311–327
11) Tommaso Di Noia, Vito Claudio Ostuni, Francesco Pesce, Giulio Binetti, David Naso, Francesco Paolo Schena, Eugenio Di Sciascio,(2013) An end stage kidney disease predictor based on an artificial neural networks ensemble, Elsevier Publication, Expert Systems with Applications 40, page no 4438–4445
12) Niels Landweh, Mark Hall and Eibe Frank,"Logistic Model Trees",extended version of a paper that appeared in the Proceedings of the 14th European Conference on Machine Learning (Landwehr et al., 2003).
13) Sushilkumar Rameshpant Kalmegh, "Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 5, Issue 1, January 2015
14) Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp-155-160
15) Mark Hall, Peter Reutemann, "WEKA KnowledgeFlow Tutorial", University of Waikato, Jun 2008.
16) Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63.
17) "Detector Performance Analysis Using ROC Curves – MATLAB & Simulink Example". www.mathworks.com. Retrieved 11 August 2016.