I J C T A, 10(8), 2017, pp. 59-67 ISSN: 0974-5572

Comparative Analysis of Machine Learning Algorithms for Chronic Kidney Disease Detection using Weka

Milandeep Arora* and Ajay Sharma**

ABSTRACT

Chronic kidney disease, a dangerous and life threatening disease which is very fatal and common now a days. We have worked to control and detect this disease very minutely in this paper by comparing the results of various outcomes of different algorithms used here. Three algorithms have been used in this paper i.e. Naïve bayes, AD tree and LWL, all are of different classifier groups. The comparisons have been made by testing the results of these three algorithms in Explorer and Experimenter interfaces of WEKA data mining tool based on four parameters i.e. number of instances either correctly or incorrectly classified, ROC area, mean absolute error and classified accuracy. In the end after all the comparisons and analysis, it has been found that AD tree is the best analysis classifier algorithm for detecting chronic kidney disease(CKD).

Index Terms: Chronic kidney disease(CKD), WEKA, classification algorithms, etc.

1. INTRODUCTION

In this paper, we have used medical datasets of chronic kidney disease readily available on UCI (university of California) repository [1] and made them introduce to WEKA data mining tool [2] with different algorithms as mentioned above. We have used two different interfaces in this paper to compare the results. Various symptoms of chronic kidney have been used in this paper to study the comparison of different algorithms. The main aim of this paper is to make acute comparative analysis of chronic kidney disease and to know which algorithm turns out to be the best in analyzing the disease. And if we want to number the objectives of this paper, it can be as follows

1. To analyze the results of chronic kidney disease medical datasets in WEKA.
2. Compare the results in Explorer and Experimenter interface with various parameters.

After that the paper follows this procedure i.e. section two tells about the details of symptoms of chronic kidney disease, section three tells about the medical datasets used in this paper, section four lets you know about the literature survey that is being used to design this paper, section five tells us about the methodology used in this paper, results are displayed and compared in section six and conclusion and future scope is given in the section seven. In the end, the references are given from whom the thought and concept of this paper is carried out and without their support and help, this research wouldn't have been the same or as effective.

2. SYMPTOMS OF CHRONIC KIDNEY DISEASE

We have used 24 symptoms of chronic kidney disease which are considered while detecting this disease and they are as follows

* M.Tech.(Research Scholar (CSE)), *Email: Milandeep9@gmail.com*

** Associate Professor ((CSE)-ACET, Amritsar)), *Email: Ajaysharma@acetamritsar.org*

Table 1
List of Symptoms Used to Detect Chronic Kidney Disease

S. No.	*Symptom*	*Short Form*
1.	Age	Age
2.	Blood pressure	Bp
3.	Specific gravity	Sg
4.	Albumin	Al
5.	Sugar	Su
6.	Red blood cells	Rbc
7.	Pus cell	Pc
8.	Pus cell clumps	pcc
9.	Bacteria	ba
10.	Blood glucose random	bgr
11.	Blood urea	bu
12.	Serum creatinine	sc
13.	Sodium	sod
14.	Potassium	pot
15.	Hemoglobin	hemo
16.	Packed cell volume	pcv
17.	White blood cell count	wc
18.	Red blood cell count	rc
19.	Hypertension	htn
20.	Diabetes mellitus	dm
21.	Coronary artery disease	cad
22.	Appetite	appet
23.	Pedal edema	pe
24.	Anemia	ane
25.	Class	class

The symptoms used in this paper for detecting chronic kidney disease are given above in Table 1 with their names and short forms which are used in this paper [3].

3. MEDICAL DATASETS

Dataset is a collection of data or a single statistical data where every attribute of data represents variable and each instance has its own description. For the prediction of Chronic kidney disease, we have used medical datasets [4] in order to compare their accuracy using wekas Explorer and Experimenter interface. The datasets used by us contains 25 attributes and 400 instances out of which 250 are suffering from the disease and 150 are not suffering from the disease. We have applied different algorithms using WEKA data mining tool for our analysis purpose.

4. LITERATURE SURVEY

Naganna chetty et al [6] has built classification models with different classification algorithms i.e. wrapper subset attribute evaluator and best first search method to predict and classify the CKD and non CKD patients. The models have been applied to medical datasets and it has been concluded that classifiers has performed better on reduced datasets than the original ones.

Lambodar jena et al [7] has suggested the use of six classifiers present in weka data mining tool and then studied their performance based on various parameters.

Milandeep et al [8] in his paper tells about the complete details of chronic kidney disease, its symptoms and the datasets that were helpful in predicting this disease effectively.

S.Ramya et al [9] in his paper has performed experiments on chronic kidney disease datasets with various algorithms such as back propagation neural network, radial basis function and random forest and it has been found that radial basis function algorithm performs the best out of three.

Parul sinha et al [10] has compared the performance of results performed on CKD datasets based on SVM and KNN classifiers and it has been found that KNN is better than SVM.

Dhamodran et al [11] has done prediction of liver disease using naïve bayes and functional tree algorithms and concluded that naïve bayes algorithm is best predicting this disease.

N k kameswara rao et al [12] has tried to discover the fast, easy and efficient data mining algorithm in prediction of epidemic disease with minimum errors, having large datasets and show reasonable patterns with dependent variables.

5. METHODOLOGY

The above Figure 1 shows the methodology used in this paper or the flow of work done accordingly, first the data is been searched from various sources available and then it is integrated to one suitable form i.e. ARFF and .CSV formats. After that it is being introduced to WEKA data mining tool on two interfaces i.e. Explorer and Experimenter and in the end the results are carried out and compared using suitable tables. We have used three different algorithms for this purpose i.e. naïve bayes, AD (alternating decision) tree and

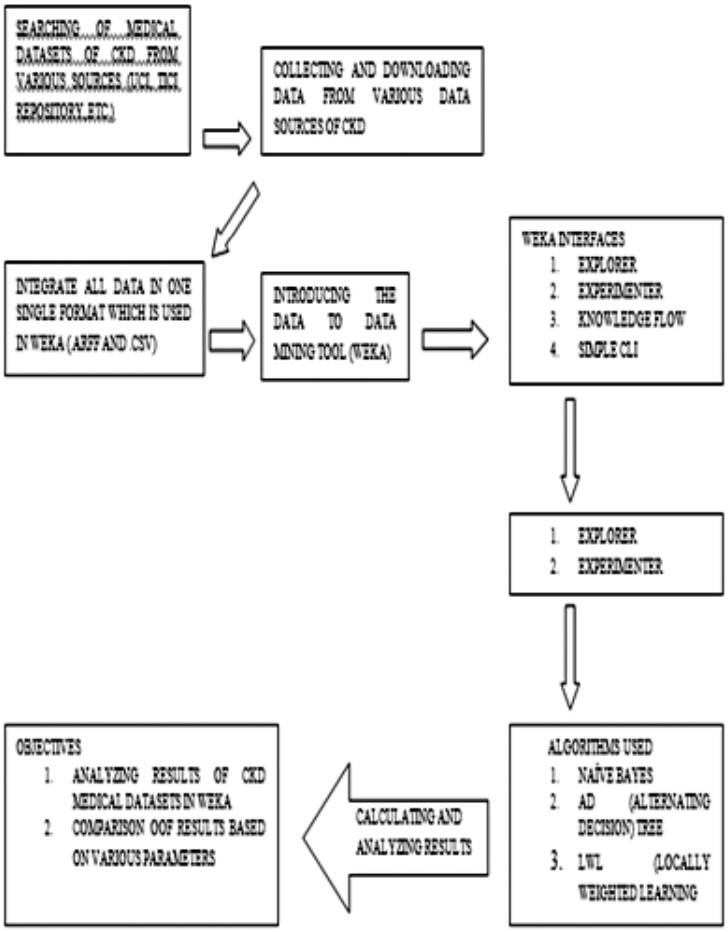

Figure 1: Showing Flow of Methodology

Table 2
Showing data mining techniques

Software	*WEKA Interface*	*Classification algorithms*	*Purpose*
WEKA	Explorer	Naïve bayes, AD tree, LWL	Analyzing and Comparison
	Experimenter	Naïve bayes, AD tree, LWL	Analyzing and Comparison

LWL (locally weighted learning). In order to carry out experimentations and implementations Weka was used as the data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java developed at Waikato. WEKA is an excellent data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics [13]. In data mining tools classification deals with identifying the problem by observing characteristics of diseases amongst patients and diagnose or predict which algorithm shows best performance on the basis of WEKA's statistical output [14] Table 2 shows the WEKA data mining techniques that have been used in this paper along with other prerequisites like data set format etc. by using different algorithms.

The interfaces we have used in this paper are Explorer and Experimenter. In this study we classified the accuracy of different algorithms Naïve bayes, AD tree and LWL on different datasets and compared the results to know which algorithm shows best performance. All the algorithms used by us were applied to a chronic kidney disease are from different types i.e. naive bayes is from bayes classifier, LWL is from lazy classifier and AD tree is from tree classifier. In order to obtain better accuracy 10 fold cross validation was performed. For each classification we selected training and testing sample randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The parameters used by us are:

1. Number of instances i.e. 400
   Either correctly classified or incorrectly classified dependent on algorithm used
2. ROC (receiver operating characteristic) area
3. Mean absolute error
4. Classified accuracy

6. RESULTS AND COMPARISONS

6.1. Explorer Interface

6.1.1. Naïve Bayes

In Figure 2 classification accuracy achieved is 95% out of total 400 instances in which there are 380 correctly classified instances and 20 are not correctly classified, mean absolute error is 0.0479 and ROC area is 1.

6.1.2. AD(Alternating Decision) Tree

In Figure 3 classification accuracy achieved is 99.75% out of total 400 instances in which there are 399 correctly classified instances and 1 is not correctly classified, mean absolute error is 0.0203 and ROC area is 1.

6.1.3. LWL(Locally Weighted Learning)

In Figure 4 classification accuracy achieved is 92.25% out of total 400 instances in which there are 369 correctly classified instances and 31 are not correctly classified, mean absolute error is 0.1132 and ROC area is 0.994.

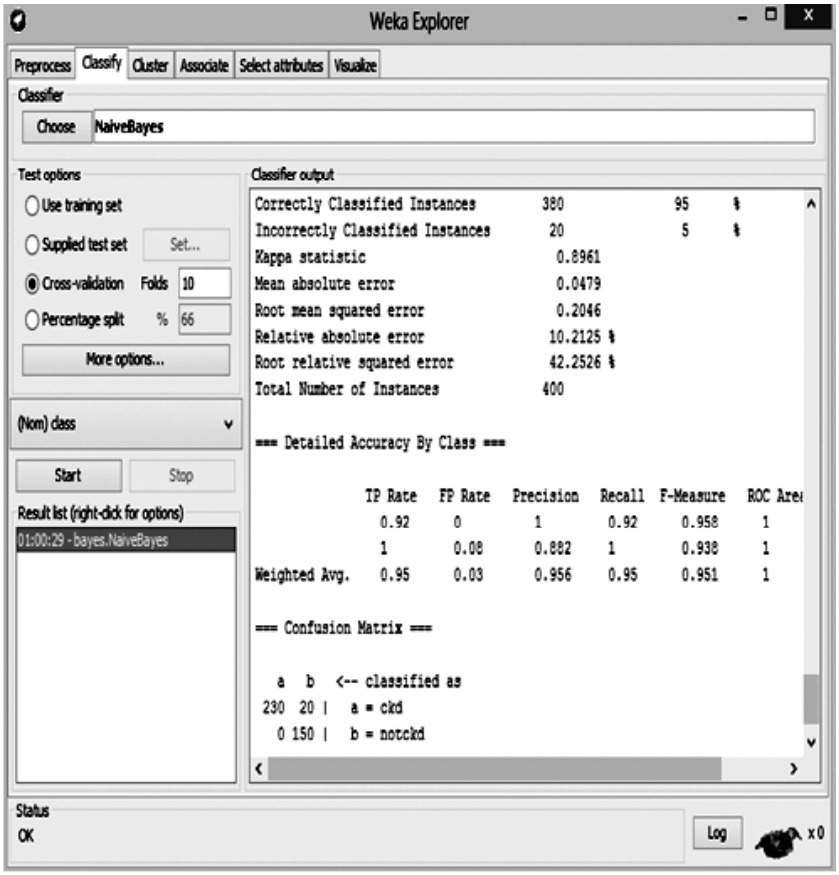

Figure 2: Results of Naive Bayes Algorithm

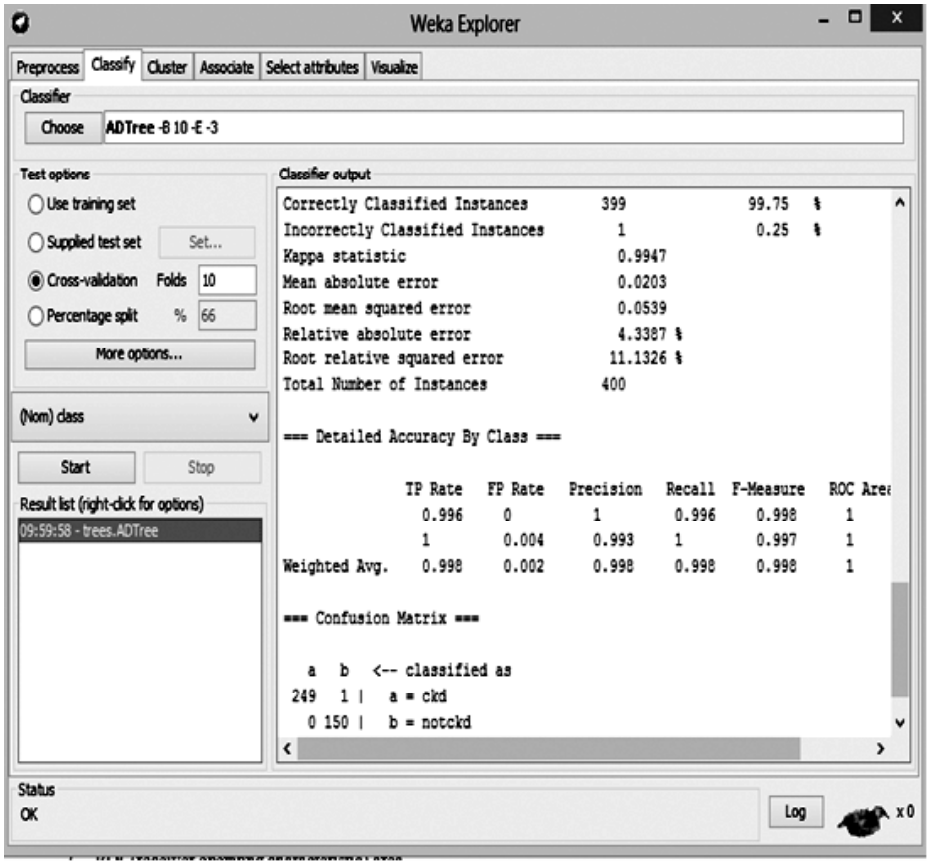

Figure 3: Results of AD Tree Algorithm

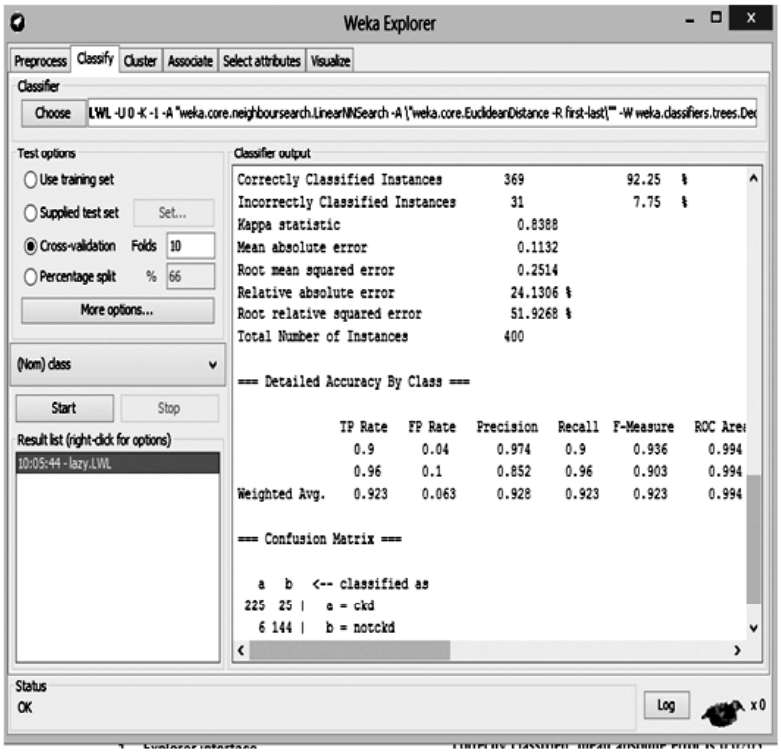

Figure 4: Results of LWL Algorithm

6.2. Experimenter Interface

6.2.1. Naive Bayes

In Figure 5 classification accuracy achieved is 95.20% out of total 400 instances in which there are 381 correctly classified instances and 19 are not correctly classified, mean absolute error is 0.05 and ROC area is 1.

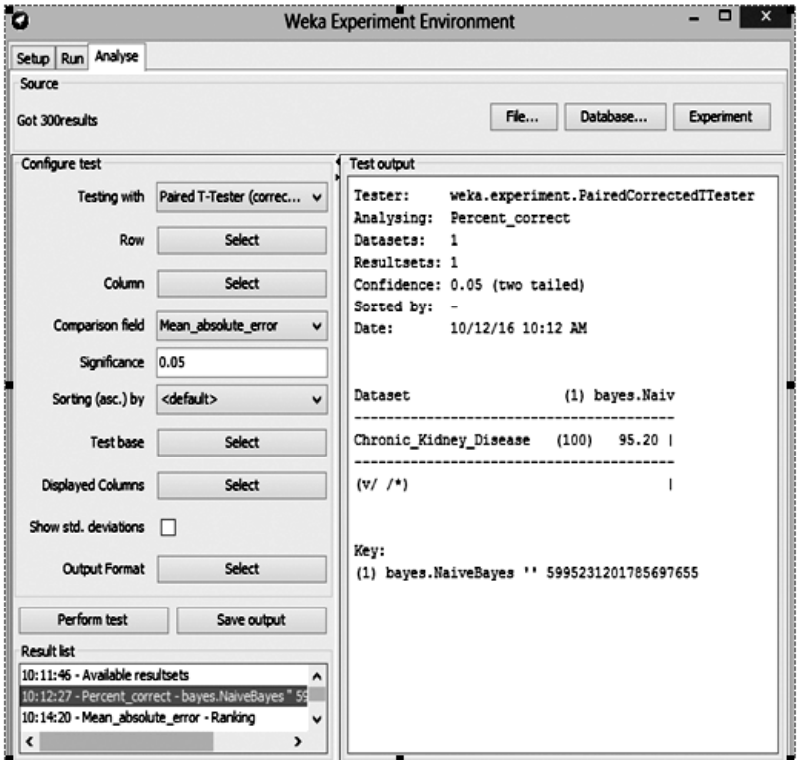

Figure 5: Results of Naive Bayes Algorithm

6.2.2. AD(Alternating Decision) Tree

In Figure 6 classification accuracy achieved is 99.58% out of total 400 instances in which there are 398 correctly classified instances and 2 are not correctly classified, mean absolute error is 0.02 and ROC area is 1.

6.2.3. LWL(locally weighted learning)

In Figure 7 classification accuracy achieved is 92.28% out of total 400 instances in which there are 369 correctly classified instances and 31 are not correctly classified, mean absolute error is 0.11 and ROC area is 1.

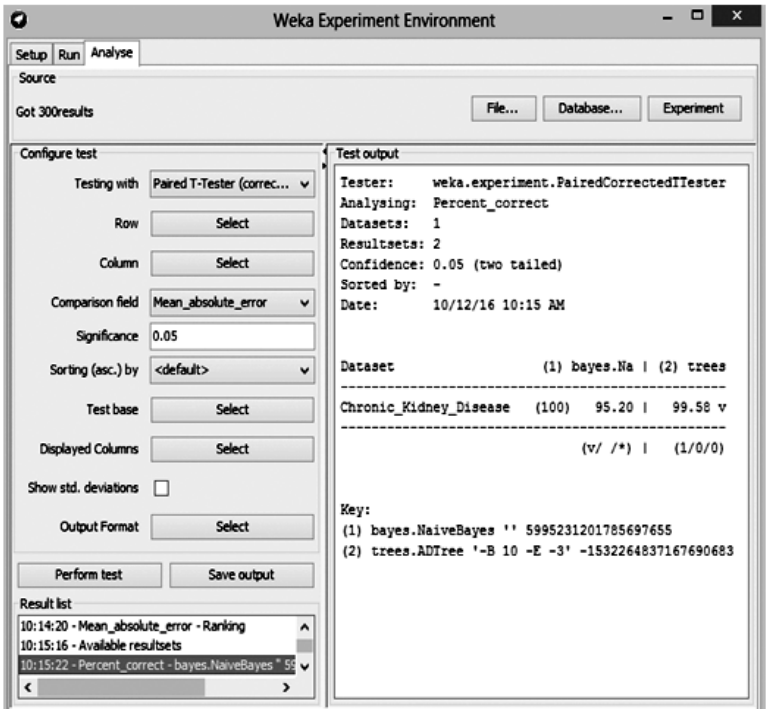

Figure 6: Results Of AD Tree Algorithm

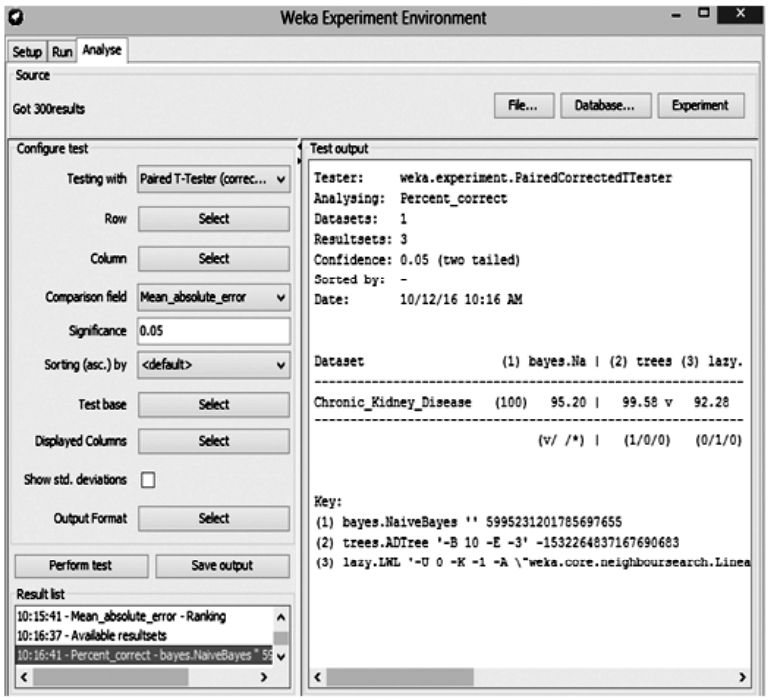

Figure 7: Results Of LWL Algorithm

Table 3
Result Of Explorer Interface

			Explorer		
	No. of Instances		*ROC*	*MAE*	*Classified Accuracy*
	400				
Naïve Bayes	C	IC	1	0.0479	95%
	380	20			
	400				
AD TREE	C	IC	1	0.0203	99.75%
	399	1			
	400				
LWL	C	IC	0.994	0.1132	92.25%
	369	31			

Table 4
Result Of Experimenter Interface

			Explorer		
	No. of Instances		*ROC*	*MAE*	*Classified Accuracy*
	400				
NAÏVE BAYES	C	IC	1	0.05	95.20%
	381	19			
	400				
AD TREE	C	IC	1	0.02	99.58%
	398	2			
	400				
LWL	C	IC	1	0.11	92.28%
	369	31			

The above Table 3 and Table 4 shows the comparison of results of two interfaces i.e. Explorer and Experimenter of weka data mining tool between three algorithms i.e. naive bayes, AD tree and LWL. The parameters used in this comparison are Number of instances either correctly classified(C) and incorrectly classified(IC), ROC(receiver operating characteristic) area, MAE(mean absolute error) and classified accuracy. The above comparison shows that there is very minute difference between the results of Explorer and Experimenter interface of weka data mining tool and from readings from both the interfaces, it is clearly visible that, LWL i.e. locally weighted learning algorithm outperforms other algorithms and hence is the best in analysing and detecting chronic kidney disease. We have used both these interfaces because there is a slight difference between there results and we haven't used the third interface that is Knowledge flow because of two reasons i.e. first it is an alternate method to Explorer interface and secondly we have used it in our earlier paper i.e. Chronic kidney disease detection by analysing medical datasets in weka [15] which was published in International journal of computer applications in august 2016 edition.

7. CONCLUSION AND FUTURE SCOPE

The main aim of this paper is to compare the results of three different algorithms of different class and it is being justified by using naïve bayes algorithm which belongs to bayes class, AD tree algorithm which belongs to tree class and LWL algorithm which belongs to lazy class. After performing all the experiments,