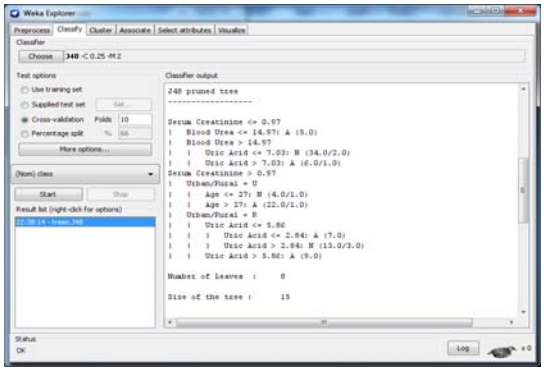

Figure 4: Generated Decision Tree

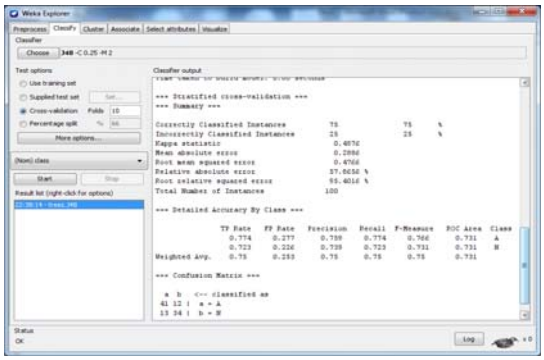

Figure 5: Interpreting Classifier Performance

- In Table 1, we can interpret the performance of Weka using different test options. The performance is interpreted in terms of the accuracy and error rate. It is found that the performance of Weka is best when tested with "Use Training Set" followed by "Cross validation" with 10 folds than with "Percentage Split" option.

Table 1: Performance of Weka Under Different Test Options

Test Options	*Accuracy*	*Error Rate*	*Kappa Statistic*	*Mean Absolute Error*
Use Training Set	92 %	8%	0.840	0.134
Cross Validation (10 folds)	75%	25%	0.497	0.288
Percentage Split (66%)	58.8%	41.1%	0.167	0.423

3.3 Classification in Tanagra

- Open Tanagra and then load the dataset in txt format. The dataset appears in Tanagra as shown in screenshot in Fig. 6. Tanagra detects the variable types automatically. It can be seen that there are 100 examples (records) and 12 attributes out of which there are 4 discrete attributes and 8 continuous attributes.

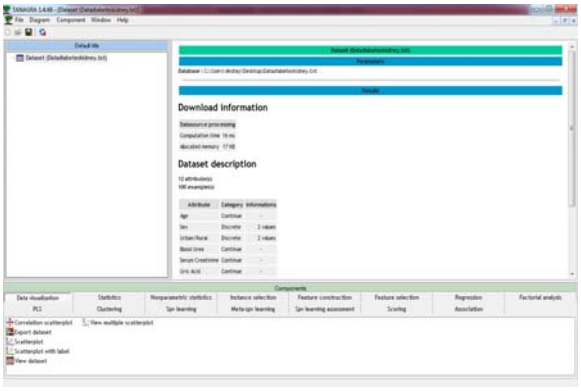

Figure 6: Opening Page

- Then, from the "View Dataset" component present inside the Data Visualization Tab, a pop-up menu appears. On choosing view menu, the data set would be displayed from Tanagra.

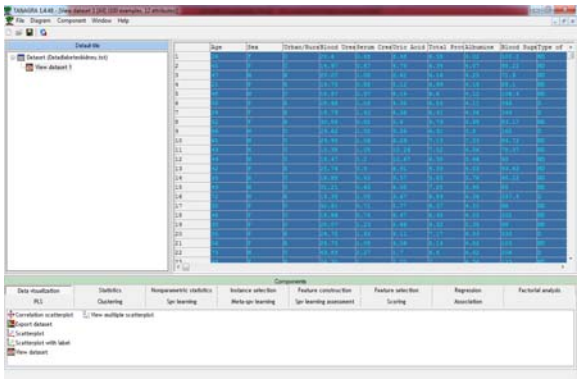

Figure 7. Viewing Dataset

- Then from the Feature Selection Tab, select "Define Status" component. Then we do the selection of parameters. We select all attributes as input except the last attribute - class. As we are interested in knowing the class (Affected or Non-affected), we set class as target.

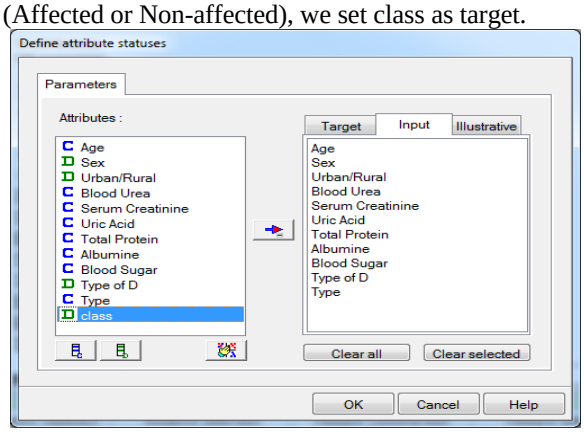

Figure 8. Selection of Input parameters

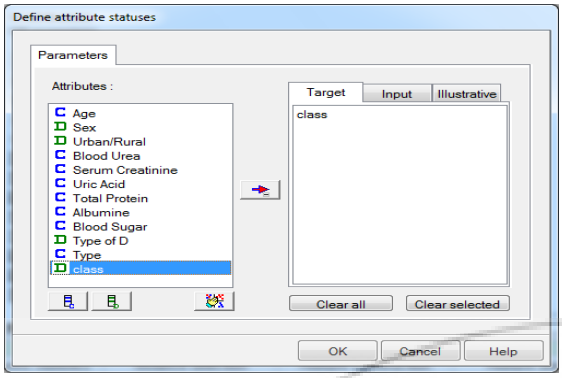

Figure 9. Selection of Output parameters

- Now we provide supervised learning using C4.5 Algorithm. For this, we add the "Supervised Learning" component present inside the Meta-Spv Learning in which we insert the "C4.5" learning algorithm (from Spv-Learning palette). On executing it, the result would be displayed. The result is shown in Fig. 10 and Fig. 11.
- Fig. 10. shows the generated decision tree in Tanagra. The root node is taken as "Total Protein" attribute and class 'A' and 'N' as the decision nodes. The tree shows that "Total Protein" is the most important attribute in the dataset that reflects greatest effect of diabetes on kidney.

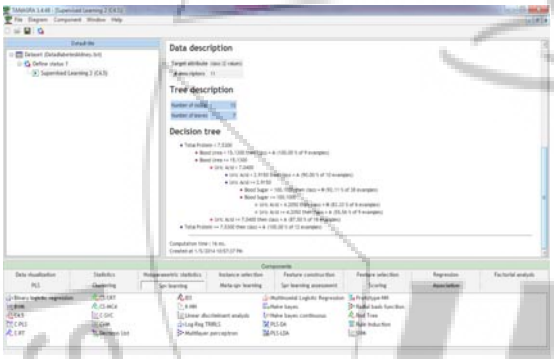

Figure 10: Generated Decision Tree

- Fig. 11. illustrates the classifier performance in Tanagra. It shows the confusion matrix and concludes that the resubstitution error rate is very less. This value is quite good for decision tree model.

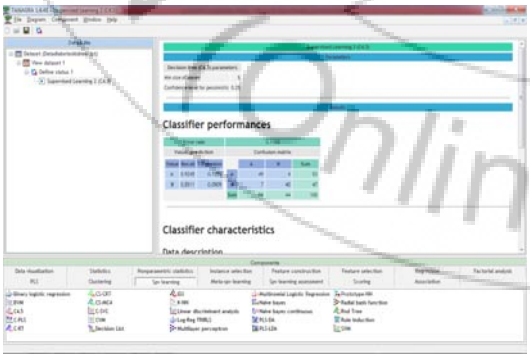

Figure 11: Interpreting Classifier Performance

- After the learning method we add a "Cross- Validation" component (from Spv Learning Assessment). We work with 10 folds and set number of repetitions to 1. We do not change the default parameters as shown in Fig. 12. The computed error rate is coming out to be 28%.

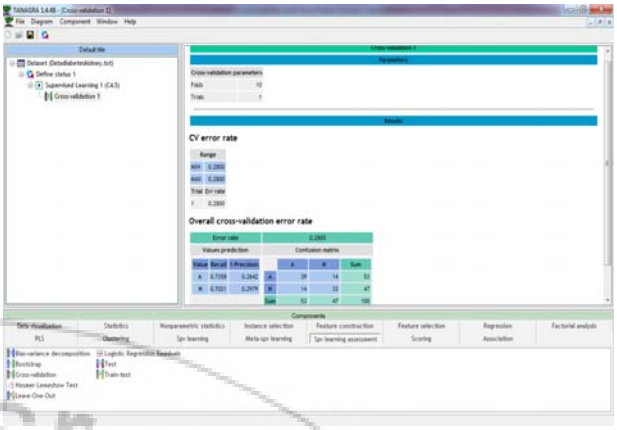

Figure 12: Supervised Learning Assessment

3.5 Comparison of classification in Tanagra & Weka

In this paper, a comparative study is made between Weka and Tanagra based on decision trees. The decision trees are generated using the application of C4.5 Algorithm that is used to generate rules signifying the effect of diabetes on kidney. The performance of classifier in both the tools is compared in terms of its accuracy, computation time and error rate.

- **Weka**

In Weka, the implementation of J48 Algorithm generates decision trees using 10-fold cross validation. Cross-validation is an efficient method for the estimation of error rate.

In Fig. 13, the decision tree has root node as "Serum Creatinine". According to the tree, Serum Creatinine determines the first decision. The numbers in parenthesis signifies the number of examples in the leaf node. The numbers after slash gives the number of misclassified examples. The decision tree includes 8 leaves and time taken to build tree model is 0.05 seconds. The error rate is 25%.

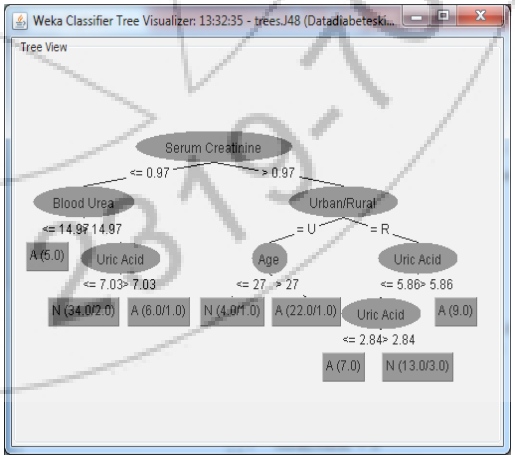

Figure 13: Decision Tree in Weka

- **Tanagra**

In Tanagra, the decision tree is generated by providing Supervised Learning using J48 Algorithm. According to the tree, "Total Protein" is taken as the root node i.e. this

attribute determines the first decision to find the diabetic effect on kidney. The tree model has 13 nodes and 7 leaves. The computation time is 0 ms. The error rate of the classifier is 11% which is lesser than Weka. So, Tanagra is more error-free than Weka.

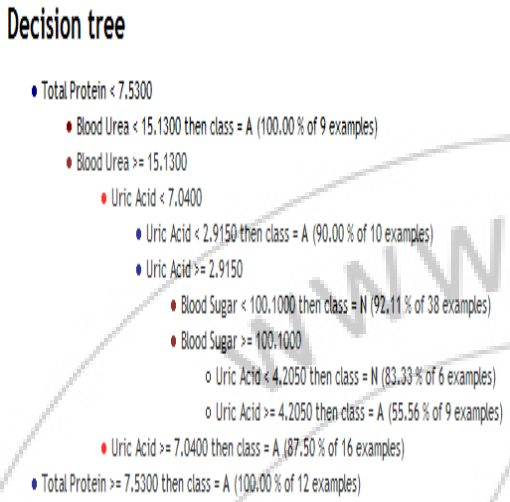

Figure 14: Decision Tree in Tanagra

4. Results and Conclusion

This research has conducted a comparative study on a dataset between two data mining toolkits (Weka and Tanagra) for classification purposes. After analyzing the results of both the tools, we found that both are able to generate tree model in very less time. Both the tools are very efficient in generating decision trees. However, in terms of classifiers` applicability, we conclude that the Weka tool is better in terms of the ability to run the classifier. However, the performance of classifier is better in Tanagra than Weka in terms of error rate. Also, Tanagra is faster than Weka in tree generation as its internal structure is organized in columns in memory. In addition, Weka tool has attained the highest performance in terms of accuracy when used with "Use Training Set" test mode than "Cross Validation" test mode followed by "Percentage Split" test mode. Through this comparative study, we conclude that Tanagra is better tool than Weka. Also, we found that c4.5 algorithm works well in decision tree induction. In future, we can implement this algorithm with more data and larger set of patient records to produce better results..

References

[1] A. Bonnaccorsi, "On the Relationship between Firm Arun K. Pujari, Data Mining Techniques
[2] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[3] http://en.wikipedia.org/wiki/Diabetes_mellitus
[4] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers,second Edition, (2006).
[5] http://www.healthline.com/health/kidney-function-tests
[6] http://www.britannica.com/EBchecked/topic/317431/kidney-function-test
[7] http://www.diabetes.ca/diabetes-and-you/living/complications/kidney/
[8] http://en.wikipedia.org/wiki/C4.5_algorithm
[9] Veronica S. Moertini,"Towards the use Of C4.5 Algorithm for Classifying Banking Dataset", INTEGRAL,Vol 8. No. 2, October 2003..
[10] Jay Gholap,"Performance Tuning of J48 Algorithm for Prediction of Soil Fertility", Innovative Journal of Medical and Health Sciences, Vol 2, No 8 (2012) .
[11] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 April 2011).
[12] http://wwww.samdrazin.com/classes/een548/project2report.pdf
[13] I.H. Witten and E. Frank, Data Mining Practical Machine Learning Tools and Techniques, Second Edition, Elsevier Inc., 2005
[14] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[15] http://en.wikipedia.org/wiki/Tanagra_(machine_learning