Predicting Chronic Kidney Disease By Using Classification Algorithms In WEKA

Muhammad Sohail, Hafiz Muneeb Ahmed, Mehwish Shabbir, Kainat Noor.

Abstract— Data mining basically used to extract the meaning of the data. In data mining different techniques are used and in this paper the classification is used to determine the predicting performance of patients either has chronic kidney disease or not. Different classification algorithms are applied in WEKA. The dataset CKD used in this research and it was taken from the UCI repository. The analysis was done on the base of accuracy. The final result of this research was that BayesNet and REP Tree algorithm showed the best accuracy among all the algorithms used in this paper.

Index Terms— Classification algorithms, Weka, CKD.

1. Introduction

Chronic kidney disease is a very dangerous disease and it disturbs the normal working of the kidney. It destroys all the normal function of the kidney. There are many symptoms of the CKD including blood pressure, diabetes. The idea of the research was to predict the chronic kidney disease in the patients by using classification algorithms. There are many classification algorithms used in this research to identify the CKD. To determine the accuracy WEKA tool was used in this research. Fourteen classifications algorithms used in this research including BayesNet, NaiveBayes, IBK, DecesionTable, JRip, ZeroR, ADTree, BFTree, DecisionStump, NBTree, FT, RandomTree, REPTree and RandomForest. Classification is a data mining method and it is mainly used to for predictive analysis on large as and small datasets[1]. It is also used to develop the models on the base of given dataset. The best tool used for classification is Weka and it provides the best running time and accuracy.

2. Literature Review

In this paper there were different algorithms were used and a analyzed the CKD dataset, the main focus of this paper was on the decision tree algorithms the algorithms used in this research paper were J48, LMT, RandomForest, NBTree, CTC J48graft, REPTree, SimpleCart, DecesionStump and HoeffdingTree and the results was measured using statistical methods and run time [2]

In this paper the predictive analysis was done and the result of this paper that LMT and RandomForest were showed the best predictive analysis, and the result was generated on the base of test mode, the classifiers showed the good predictive analysis [1].

- *Muhammad Sohail is currently pursuing MSCS degree program in Riphah international University, Pakistan, PH- +92-345-0692672, E-mail: maliksohail284@yahoo.com*
- *Mehwish Shabbir is currently pursuing MSCS degree program in Riphah international University, Pakistan, PH- +92-316-3951435, E-mail: Mehwish.shabbir22@gmail.com*

The methodology in this paper was applied to find using six different methodologies such as F-Measure, Precision, Accuracy and classifiers algorithms applied on it such as KNN, ANN, SVM, Naïve Bayes, Fuzzy and Decision Tree and fuzzy tree showed the best results among all the other classifiers[2].

In this paper predictive analysis was done by comparing different technique such as SVM, KNN and the experimental results of the SVM were the best among all.The tool used for this research was MATLAB[3].

Support Vector Machine and ANN were used on CKD dataset to classify kidney disease like Acute Nephritic Syndrome, Chronic Glomerulonephritis and Actual Renal Failure. The final result of this research was that SVM showed the best accuracy among all the other classifiers and this was done by comparing their accuracy[4].

Subhas used the classifiers ANN, KNN, SVM, C.45, Decision Tree and RandomForest, the results showed that RandomForest shows the best results among all the classifiers [5].

3. Methodology

This aim of this research paper is to predict the best classification algorithms on the base of accuracy among all the other others algorithms which were used in this research. The fourteen algorithms are used and all of those algorithms belongs to the classifications. Weka tool is used in this research to predict the analysis that patients has CKD or not.

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [6]. The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.10-fold validation is applied on dataset. Dataset used in this research was chronic Kidney disease (CKD) and it was taken

from the UCI machine learning repository. The dataset contained 15 attributes and it contained 400 instances.Following were the attributes in the dataset age, blood pressure, specific grativity, albumin, sugar, red blood cells, plus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, appetite, pedal edema, and anemia.

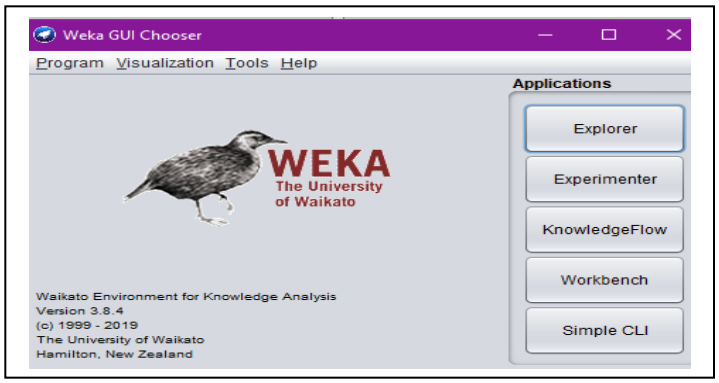

Fig 1. Weka GUI Interface

In order to start the working in Weka the explorer opened and then dataset was selected having arff file extension. After that in classification menu there are different classifications algorithms and all the algorithms applied on the dataset which were described in this research paper. The reading of all the algorithms noted manually to compare their performance easily. The accuracy, correctly clustered, incorrectly clustered and the running time of all the algorithms noted fairly so that the final results will be accurate. It was also analyzed using graphical format in excel on the base of accuracy and the running time. The detailed accuracies of all the algorithms is given below in the tabular format. In this table the percentage is also provided of the correctly and incorrectly classified clusters. All the accuracies were noted form the Weka graphical interface.

Correctly classified instance	398	99.5%
Incorrectly classified instance	2	0.5%

Fig 2. BayesNet Classified Instances

Correctly classified instance	382	95.5%
Incorrectly classified instance	18	4.5%

Fig 3. Naïve Bayes Classified Instances

Correctly classified instance	390	99.5%
Incorrectly classified instance	10	0.5%

Fig 4. NBT Tree Classified Instances

Correctly classified instance	396	99%
Incorrectly classified instance	4	1%

Fig 5. Random Forest Classified Instances

Correctly classified instance	250	62.5%
Incorrectly classified instance	150	37.5%

Fig 6. ZeroR Classified Instances

Correctly classified instance	390	97.5%
Incorrectly classified instance	10	2.5%

Fig 7. BFTree Classified Instances

Correctly classified instance	386	96.5%
Incorrectly classified instance	14	3.5%

Fig 8. Random Tree Classified Instances

Correctly classified instance	395	98.75%
Incorrectly classified instance	5	1.25%

Fig 9. FT Tree Classified Instances

Correctly classified instance	396	99%
Incorrectly classified instance	1	1%

Fig 10. IBK Classified Instances

Correctly classified instance	390	97.5%
Incorrectly classified instance	10	2.5%

Fig 11. Decision Table Classified Instances

Correctly classified instance	390	97.5%
Incorrectly classified instance	10	2.5%

Fig 12. JRip Classified Instances

Correctly classified instance	394	98.5%
Incorrectly classified instance	6	1.5%

Fig 13. ADTree Classified Instances

Correctly classified instance	368	92%
Incorrectly classified instance	2	8%

Fig 14. DecesionStump Classified Instances

Correctly classified instance	384	96%
Incorrectly classified instance	16	0.5%

Fig 15. REP Tree Classified instances

The information about practi of some algorithms are following.

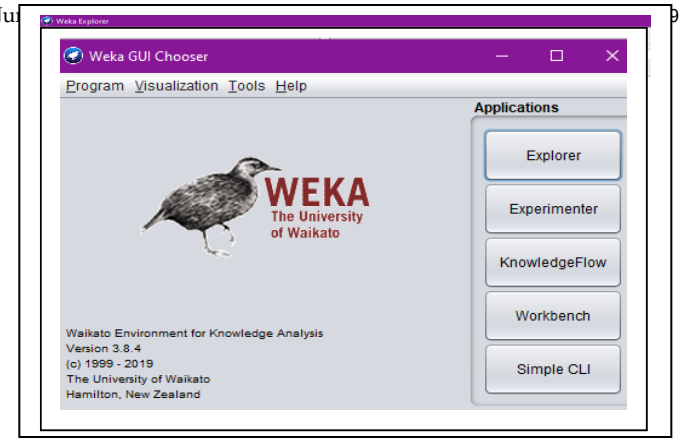

Fig 16. Random Forest

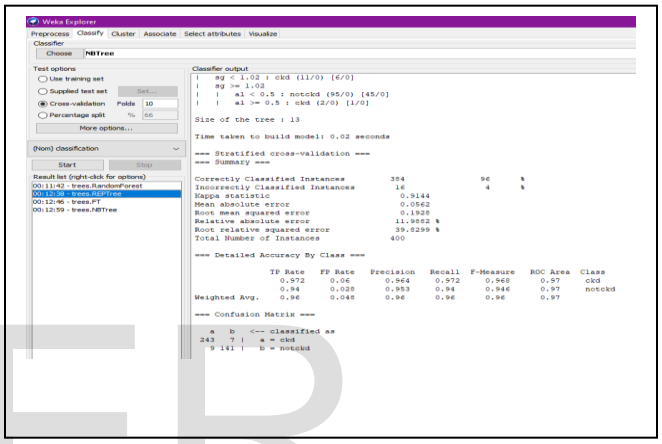

Fig 17. REPTree

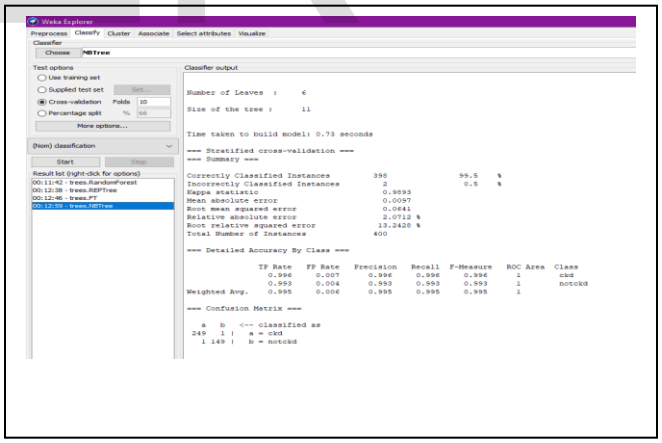

Fig 18. NBT Tree

4. RESULTS

The above all the tables clearly demonstrate the results of all the algorithms with respect to their detailed accuracy. It is cleared that the accuracy of the BayesNet and the NBT Tree are equal which shows equal correctly and incorrectly instances. Both of these algorithms show the 99.5 % accuracy. From all the accuracy it is also observed that ZeroR has the worst accuracy which is 62.5 %. The accuracy of NaiveBayes 95.5%,

IBK has 99%, DecesionTable, BFTree and JRip has 97.5%, AD-Tree has 98.5%, DecesionStump has 92%, FT has 98.75%, RandomTree has 96.5%, REPTree has 96% and RandomForest has 99%.

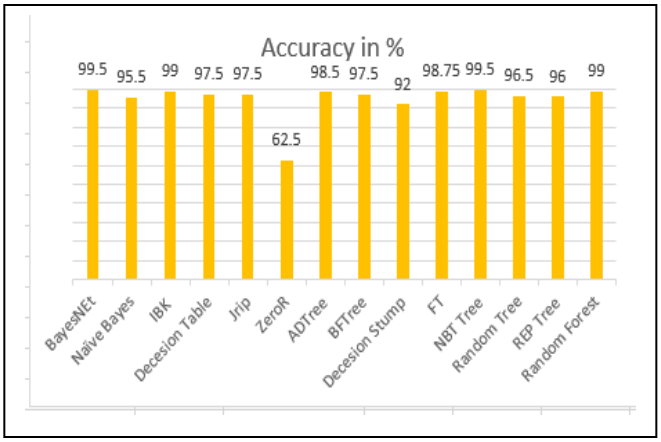

Fig 5: Accuracy in Percentage

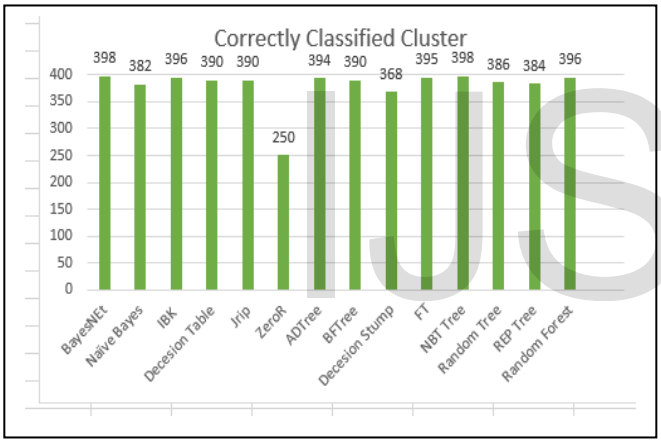

Fig 5: Correctlty Classified Instances

5. Conclusion

This document showed hat the BayesNet and NBT Tree provide the best accuracy among all the other algorithms. This research will be helpful for health department. There are many other algorithms in weka which will be helpful in predicting the chronic kidney disease in future. It will be helpful for other researchers and they can use use different tools and algorithms to check their accuracy in future. The doctors can easily diagnose the symptoms and delivers the best treatment easily.

References

[1] F. M. Ali, E.-B. E. Fgee, and Z. S. Zubi, "Predicting performance of classification algorithms," Int J Comput Eng Technol (IJCET), vol. 6, no. 2, pp. 19-28, 2015.

[2] I. Pasadana et al., "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques," in Journal of Physics: Conference Series, 2019, vol. 1255, no. 1: IOP Publishing, p. 012024.

[3] P. Sinha and P. Sinha, "Comparative study of chronic kidney disease prediction using KNN and SVM," International Journal of Engineering Research and Technology, vol. 4, no. 12, pp. 608-12, 2015.

[4] S. Vijayarani, S. Dhayanand, and M. Phil, "Kidney disease prediction using SVM and ANN algorithms," International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.

[5] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in CMBEBIH 2017: Springer, 2017, pp. 589-594.

[6] R. Arora, "Comparative analysis of classification algorithms on different datasets using WEKA," International Journal of Computer Applications, vol. 54, no. 13, 2012.