CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

A ROLE OF DECISION TREE CLASSIFICATION DATA MINING TECHNIQUES TO PREDICT CHRONIC KIDNEY DISEASE

Dr. Gul Mohamed Rasitha Banu*
Assistant Professor, Department of Public Health , College of Nursing and Health Sciences, Jazan university, Jazan, Saudi Arabia

ABSTRACT

Chronic Kidney Disease is a condition in which the kidneys are damaged and cannot filter blood as well as they should. Because of this, excess fluid and waste from blood remain in the body and may cause other health problems, such as heart disease and stroke. However, it is always recommended to diagnose the disease at an earlier stage in order to prevent further harmful effects and to provide the treatment to keep the thyroid hormone at normal level. Data Mining is playing vital role in health care applications. It is used to analyze the large volumes of data. One of the important tasks in data mining is predicting disease in earlier stage, which assist physician to give better treatment to the patients. Classification is one of the most significant data mining techniques. It is supervised learning and used to classify predefined data sets. Data mining technique is mainly used in healthcare organizations for decision making, diagnosing diseases and giving better treatment to the patients. The data set used for this study on chronic kidney disease is taken from University of California Irvine (UCI) data repository. The entire research work is to be carried out with Waikato Environment in Knowledge Analysis (WEKA) open source software under Windows 7 environment. An experimental study is to be carried out using data mining techniques such as J48, Decision stump, Random Forest tree, REP tree and Random tree. As a result, the performance will be evaluated for classification techniques and their accuracy will be compared through confusion matrix. It has been concluded that the Random Forest tree gives better accuracy than other classification techniques.

Keywords: Data Mining, Chronic kidney disease, confusion matrix, Decision Tree classification

1.0 INTRODUCTION

Chronic kidney disease (CKD) is a gradually increasing global health concern. It is a condition that kidneys *slowly get damaged* and can't do important jobs like removing waste and keeping blood pressure normal. Diabetes, high blood pressure, heart disease, and a family history of kidney failure are the key risk factors for renal disease. Many CKD patients do not exhibit any symptoms until their condition reaches more advanced stages or until complications arise. If symptoms appear, they could be as follows: foamy urine, urinating (or passing gas) more or less frequently than normal, Dry or itchy skin, emesis, appetite loss, Loss of weight without attempting to reduce it. Individuals with more severe stages of CKD may also experience difficulty focusing, tingling in your feet, ankles, legs, or arms, tense or cramping muscles, breathing difficulty, throwing up, difficulty falling asleep, the breath has an ammonia scent.

Data Mining is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Classification is one of the tasks

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

in data mining. Classification is the process of assigning new objects to predefined categories or classes. In our study, the Random Forest Tree, Random Tree, J48, Decision Stump, and Random Tree algorithms are utilized to predict chronic kidney disease. For the experiment, a data set of 25 features is downloaded from the UCI repository website. [5] WEKA, open-source software, is used throughout the entire project in a Windows 7 environment. [9]. Using confusion matrix, classifiers performance is assessed.

2.0 OBJECTIVES

1. To apply various Data mining classification algorithms on dataset to predict the chronic kidney disease.
2. To assess the classifiers' performance using metrics like the confusion matrix, accuracy and error rate, and model construction time.

3.0 METHODOLOGY

3.1 Dataset Description

The chronic kidney disease dataset used in this work is collected from the website [5]. The chronic kidney disease dataset consists of 400 instances from which 250 instances belong to category chronic kidney disease and 150 instances belong to category not chronic kidney disease. In our research work we have taken 25 attributes which will be used to predict the disease. The chronic kidney disease data set is given below in Table 1.

S.No	Attribute Name	Value type
1	Age	Numeric
2	Blood Pressure	Numeric
3	Specific Gravity	Nominal
4	Albumin	Nominal
5	Sugar	Nominal
6	RBC	Nominal
7	Puscell	Nominal
8	Pus Cell clumps	Nominal
9	Bacteria	Nominal
10	Blood Glucose Random	Numeric
11	Blood urea	Numeric
12	Serum creatinine	Numeric
13	Sodium	Numeric
14	Potassium	Numeric
15	Hemoglobin	Numeric
6	Packed Cell Volume	Numeric
17	WBC	Numeric

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

18	RBC count	Numeric
19	Hypertension	Nominal
20	Diabetes	Nominal
21	Coronary Artery Disease	Nominal
22	Appetite	Nominal
23	Pedal Edema	Nominal
24	Anemia	Nominal
25	Class	Ckd, not Ckd

Table1: Chronic kidney disease dataset

3.2 Proposed System

In our work, we have used data pre-processing task to fill up the missing values and remove noisy data. We have applied J48, REP tree, Decision stump, Random tree and Random Forest tree data mining classification techniques on chronic kidney disease dataset to predict chronic kidney disease and then the classifiers performances were evaluated through confusion matrix.

4.0 RESULTS

In chronic kidney disease dataset there are 400 instances and 25 attributes . These instances are classified into 2 classes chronic kidney disease(ckd) and Not chronic kidney disease (not ckd). Out of 400 instances , 250 instances are belongs to class ckd and 150 instances are belongs to Not ckd class. The classifiers such as J48, REP tree, Decision stump, Random tree and Random Forest tree were applied on the chronic kidney disease dataset. The results of the classifiers are shown below. The following table 2 shows the confusion matrix of REP tree classifier.

The confusion Matrix of REP tree is shown below:

Target Class	ckd	Not ckd
ckd	246	4
Not ckd	9	141

Table 2: Confusion Matrix of REP tree Classifier

In REP tree classifier correctly classified instances are 387 and wrongly classified instances are 13. The following table 3 shows the confusion matrix of Random tree classifier.

The confusion Matrix of Random tree is shown below:

Target Class	ckd	Not ckd
ckd	237	13
Not ckd	5	145

Table 3: Confusion Matrix of Random tree Algorithm

In Random forest tree classifier correctly classified instances are 382 and wrongly classified instances are 18.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

The following table 4 shows the confusion matrix of J48 classifier.

Target Class	ckd	Not ckd
ckd	249	1
Not ckd	3	147

Table 4: Confusion Matrix of J48 Algorithm

In J48 tree classifier correctly classified instances are 396 and wrongly classified instances are 4.

The following table 5 shows the confusion matrix of Random Forest Tree classifier.

Target Class	ckd	Not ckd
ckd	249	0
Not ckd	1	150

Table 5: Confusion Matrix of Random Forest tree classifier

In Random Forest Tree classifier, correctly classified instances are 399 and wrongly classified instances are 1.

The following table 6 shows the confusion matrix of Decision stump classifier.

Target Class	ckd	Not ckd
ckd	224	26
Not ckd	6	144

Table 6: Confusion Matrix of Decision stump classifier

In Decision stump classifier, correctly classified instances are 368 and wrongly classified instances are 32.

The following table 7 shows the accuracy, time taken to build the model and Error Rate of REP Tree, Random Forest, Decision stump, Random Tree and J48 Algorithm.

Classifiers	Accuracy	Time Taken to build the Model	Error Rate
REP Tree	96.75%	0.02 Seconds	3.25%
Random Tree	95.5%	0 Seconds	4.5%
J48	99%	0.02 Seconds	1%
Decision Stump	92%	0 Seconds	8%
Random Forest Tree	**99.75%**	**0.11 seconds**	**0.25%**

Table 7: Accuracy, time taken to build the model and Error Rate of REP Tree, Random Forest, Decision stump, Random Tree and J48 Classifier

Table 7 shows that the accuracy of J48 (99%), accuracy of Random Forest tree (99.75%), accuracy of REP tree (96.75%), accuracy of Random Tree(95.5%), accuracy of Decision stump (92%) tree. J48 takes 0.02 seconds to build the model, Random Forest Tree takes 0.11 seconds to build the model and REP Tree takes 0.02 seconds to build the model, Random tree and Decision stump takes 0 seconds to build the model. The error rate of Random Forest tree is 0.02%, the error rate of J48 is 1%, the error rate of REP tree is 3.25%, the error rate of decision stump is 8% and the error rate of Random Tree is 4.5%. While comparing Random Forest tree with other classifier which is giving

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

highest accuracy (99.75%), less error rate (0.25%) and (0.11 seconds) taken to build the model than other classifiers.

5.0 CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is an incredibly challenging task in the field of health care. Various data mining techniques have proven to be extremely helpful in decision making. In our work, we have used data cleaning task to fill up the missing values and we have applied J48, REP tree, Decision stump, Random tree and Random Forest tree data mining classification techniques which are used to predict chronic kidney disease. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and error rate. The Random Forest tree Algorithm gives 99.75% which is providing better accuracy than other classifiers and also it gives very minimum error rate 0.25% than REP tree, J48, Decision stump and Random tree. As a future work the same technique is used to apply for other disease datasets such as heart disease, Lung cancer, cervical cancer, Anemia, liver disease, and iris and so on.

REFERENCES

1. I.A. Pasadana et al 2019 J. Phys.: Conf. Ser. 1255 012024

2. Aldhyani , Alshebami AS, Alzahrani MY. Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. *J Healthc Eng.* 2020;2020:4984967

3. Sattari M, Mohammadi M. Using Data Mining Techniques to Predict Chronic Kidney Disease: A Review Study. Int J Prev Med. 2023 Aug 28;14:110. doi: 10.4103/ijpvm.ijpvm_482_21. PMID: 37855011; PMCID: PMC10580203.

4. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inform. 2023 Jan 12;14:100189. doi: 10.1016/j.jpi.2023.100189. PMID: 36714452; PMCID: PMC9874070.

5. https://archive.ics.uci.edu/dataset/336/chronic kidney disease

6. Sahana B J, 2017, Prediction of Chronic Kidney Disease using Data Mining Classification Techniques and ANN, International Journal Of Engineering Research & Technology (IJERT) NCETEIT – 2017 (Volume 5 – Issue 20).

7. Ilyas, H., Ali, S., Ponum, M. *et al.* Chronic kidney disease diagnosis using decision tree algorithms. *BMC Nephrol* **22**, 273 (2021). https://doi.org/10.1186/s12882-021-02474-z

8. https://en.wikipedia.org/wiki/Weka_(machine_learning)

9. https://www.cs.waikato.ac.nz/ml/weka/

10. Jiawei Han, KamberMicheline, Datamining*:Concepts and Techniques*, Morgan Kaufmann Publisher, (2009).

11. V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence),* Noida, India, 2016, pp. 300-305.

12. Ana Pinto, Diana Ferreira, Cristiana Neto, António Abelha, José Machado, "Data Mining to Predict Early Stage Chronic Kidney Disease ,"Procedia Computer Science'',
Volume 177, 2020, Pages 562 - 567, ISSN 1877 - 0509.

13. Ganie SM, Dutta Pramanik PK, Mallik S, Zhao Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. PLoS One. 2023 Dec 1;18(12):e0295234.