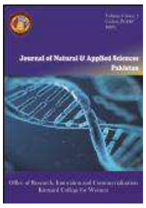
Contents lists available http://www.kinnaird.edu.pk/

Journal of Natural & Applied Sciences Pakistan

Journal homepage: http://jnasp.kinnaird.edu.pk/



A SMART METHODOLOGY FOR ANALYZING CHRONIC KIDNEY DISEASE DETECTION

Rana M. Amir Latif[1*], Muhammad Farhan[1], Farah Ijaz[1], Muhammad Umer[1], Syed Umair Aslam Shah[1]
[1]Department of Computer Science,
Comsats University Islamabad, Sahiwal Campus.

Article Info

*Corresponding Author
Tel: +92 307-6148524
Email Id:
ranaamir10611@gmail.com

Keywords
Kidney diseases; medical dataset; machine learning algorithms; chronic disease; weka.

Abstract

Chronic kidney disease is increasing day by day all over the world; this is also known as a chronic rental disease because this disease is life-threatening. To save people from this life-threatening disease, we have given suitable techniques and results in this paper for its accurate and early detection. It will not ensure 100% safety from disease but provide a suitable time to get a cure from it with its early detection. We have used 24 symptoms of chronic kidney disease in this paper which help us to accurately detect this disease with the help of two machine learning classification algorithms, i.e., C4.5 and C5.0. We conclude the results by introducing the medical datasets to all two algorithms separately with the help of the decision tree and the statistical information about the dataset, also this medical dataset is feed into the machine learning algorithms which are built-in WEKA are used for comparison and to check the accuracy of the algorithm. Algorithms used in this comparison are J48, Naïve Bayes and Logistic Model Tree (LMT) are used and them accurately to predict the chronic kidney disease is calculated. This research proved the efficiency of the C5.0 algorithm since it predicted more accuracy, short duration and less error rate as compared to the C4.5 algorithm. Also, the J48 algorithm is the best algorithm to predict chronic kidney disease. The best algorithm among different algorithms can be selected prediction model can be used for determining chronic kidney disease.

1. Introduction

At the lower back of the abdomen, kidneys are the positioned pair of organs. By using bladder through urine kidney wash blood by eliminating waste or toxin material from the body. In that circumstances when the kidney is not able to filter the waste blood through body then body fills with the waste or toxins material that causes the kidney failure, and, in the results, this failure leads to death also this kidney disease be more severe and chronic. This chronic disease harm the kidney in certain circumstances that also decrease the ability to keep us healthy. This waste toxin causes a high level of difficulties in our blood like nerve damage, poor nutritional health, weak bones, anemia, and high blood pressure, increase the risk of having heart and blood vessel disease. Also, this chronic disease causes many diseases in the human body like Bacteria and albumin in urine, anemia, lupus, coronary artery disease, hypertension, high blood pressure, and diabetes. These diseases have some complication in the medication of kidney disease when a deficiency of sodium and potassium in blood vessels and family history. Chronic kidney diseases are controlled to getting worst with early diagnose and treatment. This disease increase in the human body; ultimately it leads to kidney failure so in the treatment of this disease needs a kidney transplant or dialysis to save a life [1]. For analysis and prediction machine learning is a growing field that is concerned with several variables data and grown from computational learning theory in artificial intelligence, the study of pattern recognition having algorithms and computational methods. Machine learning plays a vital role in the medical sciences for diagnosing and prediction of many critical diseases. For prediction and diagnosing in machine learning some sets of datasets with some certain features are used for the representation of each instance in the dataset [2]. That is why professionals and experts and imperfect in finding the hidden patterns in datasets. Hence to investigate the raw data and to extract existing information from the datasets for decision makers use different computational methods.

One of ID3 extension is C4.5. This algorithm produces a decision tree by recursively splitting the data of given crop pest training. As a statistical classifier C4.5 is often referred and for classification, its generated decision trees used. By using the depth-first strategy, C4.5 decision tree grows. Resulting decision trees pruning is allowing by it. With noisy data, missing values and numeric attributes it can deal. To handle continuous attributes list is split by creating a threshold, splitting is done based on attribute value equal to or less than threshold value and value is above the threshold. Once the tree is created C4.5 goes back through it and branches that do not help, C4.5 attempts to remove them by replacing them with leaf nodes. Also, the C4.5 algorithm shows the statistical information about the dataset that will use for analysis. Correctly classified accuracy be extracted after analysis of dataset which shows correctly classified test accuracy percentage. Kappa statistics that measure inter-rater agreement for the qualitative item and in analyze algorithm classification accuracy the number of errors is shown by mean absolute error [3].

One of ID3 extension is C4.5, and C4.5 algorithm extension is C5.0. For big data set this classification algorithm is suitable. Based on efficiency, memory and speed are improved than C4.5. Crop pest training data provide maximum weight, and C5.0 model works by splitting it. The missing attribute from crop pest training data set and C5.0 easily handles the multivalued attribute. In this paper, the training pest data is used for constructing the C5.0 decision tree while for prediction testing pest data is used. By using the confusion matrix that will create after the analysis of the dataset the accuracy of this algorithm finds out. The confusion matrix is the table that is often using to describe the performance of the classification [4]. In ID3 J48 utilize an expansion. J48 includes lots of supplemental features including a derivation of regulations, steady attribute-value ranges, and conclusion bushes pruning and conclusion designs trimming. J48 be open Java source code of C4.5 algorithm at the WEKA for data mining. WEKA instrument related to shrub pruning, plus and it supplies greater alternatives to incorporate with trees. An instrument for accuracy may function as prospective overfitting pruning. To get each foliage pruning that the recursive classification is going to be achieved and the classification of this data needs to be ideal. The principle of this algorithm generates specific info. To obtain the truth and balance of endurance would be the primary purpose of all generalization of this decision shrub [5].

This Bayes classifier supposes a feature from the category does not have any connection to some different feature from this category. It might be described having a good instance that writers have an apple that it believes like an excellent fresh fruit when it is a reddish coloration and 3 inches. Even writers can view these capabilities will be contingent on the occurrence of different capabilities. However, the different capabilities will lead undependably from the chances of this fresh fruit that's the reason why writers say it is that a Naïve. Naïve Bayes version is beneficial for massive datasets. On exceptionally complex category algorithm naïve Bayes is understood as jelqing [6].

For supervised learning terminal induction and model's way is a favorite technique equally techniques utilize for numerical worth along with minimal lessons. Tress who have lining regression work in the leaves may utilize for calling numerical amounts [7]. Writers utilize logistic regression rather than linear regression with this creator might use platform shrewd fitting course of action which may select proper qualities to organize info and reveal how writers might way to generate a logistic regression version in the leaves to automatically enhance in high degrees at the shrub. Together with all the craft of finding out strategies, writers can review the operation of their algorithm using various strategies. The focus or objective of this paper is detecting the chronic kidney disease with accuracy and helping humanity to fight with this dangerous disease, but we defined the objectives of this paper in a specific manner i.e.

- To extract useful information with classification accuracy for chronic kidney disease detection.
- Classification of different machine learning algorithms
- To identify the best performing algorithm in detecting chronic kidney disease out of the three algorithms used in this paper.

We have used 24 symptoms of chronic kidney disease which are considered while detecting this disease and they are as follows shown in "Table. 1".

Table 1. The symptoms used in this paper for detecting chronic kidney disease **[8]**

Attribute Name	*Attribute Information*
Age(numerical)	age in years
Blood Pressure(numerical)	bp in mm/Hg
Specific Gravity(nominal)	sg - (1.005,1.010,1.015,1.020,1.025)
Albumin(nominal)	al - (0,1,2,3,4,5)
Sugar(nominal)	su - (0,1,2,3,4,5)
Red Blood Cells(nominal)	rbc - (normal, abnormal)
Pus Cell (nominal)	pc - (normal, abnormal)
Pus Cell clumps(nominal)	pcc - (present, not present)
Bacteria(nominal)	ba - (present, not present)
Blood Glucose Random(numerical)	bgr in mgs/dl
Blood Urea(numerical)	bu in mgs/dl
Serum Creatinine(numerical)	sc in mgs/dl
Sodium(numerical)	sod in mEq/L
Potassium(numerical)	pot in mEq/L
Hemoglobin(numerical)	hemo in gms
Packed Cell Volume(numerical)	pcv
White Blood Cell Count(numerical)	wc in cells/cumm
Red Blood Cell Count(numerical)	rc in millions/cmm
Hypertension(nominal)	htn - (yes, no)
Diabetes Mellitus(nominal)	dm - (yes, no)
Coronary Artery Disease(nominal)	cad - (yes, no)
Appetite(nominal)	appet - (good, poor)
Pedal Edema(nominal)	pe - (yes, no)
Anemia(nominal)	ane - (yes, no)
Class (nominal)	class - (ckd, not ckd)

2. Literature Review

The medical business is providing vast quantities of info that will need to become mine to detect hidden info, therefore anybody useful forecast, mining, decision-making, and investigation. To successfully manage this specific circumstance machine learning methods, and supply drugs. Moreover, persistent kidney illness forecast is just one of the full most fundamental issues in healthcare decision-making as it is but one of many main reasons for departure. Thus, an automatic instrument for the early forecast with this disorder will probably be helpful to remedy. Together with 10-fold cross validation analyzing of each classifier separately as well as in practice, the chosen characteristic can utilize. Concerning MCC, precision, and a spot under the ROC curve (AUC) with worth 1.0, 1.0 and 1.0 respectively, and the RF classifier outperforms other classifiers plus it is exhibited from the outcome [9]. Even much wisdom and practical experience are necessary for kidney disorder as it is a complicated endeavor. In illness burden in developing states, in some among the key contributors as well as in most developed states quiet killer is kidney disorder. For forecasting, the ailments by the data set chiefly data-mining are utilized from the medical market. Bipolar disorder data collection is examined with statistics mining classification methods, specifically Naive Bayes, ANN, Decision trees and so on [10].

To asses correctly classified data by applying machine learning techniques to understand kidney stone patient's statistical analysis. Due to human generations and change of climatic factors, kidney stone formation is most common. Ancient culture is losing in India, and adverse effects in humans are producing by food habits and new industrialization. By data mining and statistical analysis, a survey has been conducted on kidney stones. In the present work, authors produce a meta-analysis and systematic review. Good accuracy is predicted in present studies with C4.5, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and NNge show 100% correctly classified, with zero relative absolute error and good accuracy results. Kidney stones treatment results might be better with machine learning approaches. Due to anthropogenic climate change, for researchers in current decades, it becomes one of the challenging tasks of treatment and Diagnosis of kidney stones [11].

This paper is focusing on developing a method to enhance the ultrasound kidney images diagnosis based on association rule-mining for classification of kidney images and automated diagnosis implementing a computer-aided decision support system. Analyze the medical images, the association rule mining method is used and suggestions of diagnosis it will automatically generate. To suggest the diagnosing, it combines specialist high-level knowledge with automatically images low-level features. In data mining techniques, well-known is association rule mining because, in massive databases, it aims to find interesting patterns. Three categories of kidney namely medical renal, cortical cyst and healthy will be distinguished by our proposed method. For processing, the author takes segmented US kidney images. Then, for extraction of the feature, the process of feature extraction is applied, and only just the relevant features are selected from those extracted features. To reduce the mining complexity, the author will apply two things on extracted features one is feature selection and the second one is the discretization process. Based on the selected features, the association rules are generated. Bayesian Classifier algorithm is used in our proposed method which is based on the Bayesian theorem, and it is a new associative classifier. It suggests a diagnosis of a given image and with a high value of accuracy classifies it [12].

Often, Artificial Neural networks are used for early detection of diseases, as a powerful discriminating classifier. Over parametric classifiers, they have several advantages like as discriminate analysis. By using three different characteristics and architecture of neural network algorithms to diagnose kidney stone disease is the primary objective of this paper. Based on the training data set size, accuracy and time taken to build a model of all three neural networks, show a performance comparison is the aim of this work. For kidney stone disease diagnosis, the author will use Learning vector quantization (LVQ), two layers feedforward perceptron trained with backpropagation training algorithm and Radial basis function (RBF) networks. In this work, the author used an open source tool for simulation, the name of the tool is the Waikato Environment for Knowledge Analysis (WEKA) and I used its version 3.7.5. Real world dataset author used for diagnosis with 8 attributes and 1000 instances. The author will propose the best algorithm in the end part, for kidney stone diagnosis, on the performance comparison of different algorithms. So, this will help in reduces the diagnosis time and inpatient early identification of kidney stone will be possible [13].

In this paper pushes us to precisely identify this malady with the assistance of three arrangement calculation, i.e., J48, SMO in data mining tool WEKA and naïve Bayes. Author complete the outcomes by acquainting the therapeutic datasets to each of three processes independently with the assistance of information stream interface of WEKA data mining apparatus, the limitations which are utilized to think about the aftereffects of these three different calculations are mean kappa statistics, absolute error and the total number of instances studied either correctly or incorrectly. The interface author has utilized as a part of this paper is information stream window of WEKA data mining tool which not only gives the accurate assessments of results as well as gives a pictorial perspective of the information stream. At last, after directing all analyses, it has been inferred that J48 is the best performing calculation out of three calculations utilized as part of distinguishing chronic kidney disease accurately and at most punctual stages conceivable [14].

The potential risks to people's wellbeing from persistent diseases invariably come about and enlarging day daily on earth. Step-by-step guidelines reduce those dangers is an increasingly important dilemma in healing therapy. Together with such lines, this paper suggests a version of some chronic ailments predict and decision-making frame comprising case-based justification (CBR) and data mining

(DM). The principles processes of this frame include things like

1. Embracing data mining methods to decide on the known key fundamentals from health test statistics
2. Utilizing the split standers for its distinct persistent diseases prediction
3. Utilizing CBR to help reevaluate the continual diseases analysis and solutions greatly
4. Stretching those approaches to do the job in a frame to that capability of continual ailments maximizing, finding out, coordinating, and discussing [15].

Now, a genetic algorithm (GA) prepaid neural network (NN) established version was suggested to recognize persistent kidney disorder (CKD) that includes proven to become perhaps one among the absolute most recent threats into this generating and undeveloped states. Investigations and studies from different sections of India have suggested that CKD is changing to some remarkable concern daily. The fiscal bodyweight of this future and treatment outcomes of CKD can be extreme into most, should maybe not understood in an earlier coordinating. The more NN-GA version was suggested which defeats the matter of making use of by neighborhood search-based finding out algorithms to directly organize NNs. The enter vector of this NN is always improved using GA to organize the exact NN. The version was contrasted and undoubtedly known classifiers such as multi-layer Perception feed-forward community (MLP-FFN), as well with NN and Random Forest. The implementation of this classifiers was projected seeing exactness precision, remember, F-Measure, along with correctness. The demo is around to urge that NN-GA-based version is designed with identifying CKD additional than any other active version [16].

3. Methodology

In this paper, our focus to find that relevant features for kidney disease that will differentiate about the symptoms of the chronic kidney disease in a specific person exists or not exist. For identification of that symptoms there could be carried out the specific analysis with using different algorithms of machine learning also do some statistical Investigation for more finding more difference between this classes that it could be chronic kidney diseases (CKD) or no chronic kidney diseases (CKD).

Firstly, the author is searching for a real-time dataset on kidney disease. Finally, the author got dataset on UCI machine learning responsorial this dataset in the '.TEXT' format converts this dataset in some suitable format it could be '.CSV' or.' ARFF.' After this, the author is going to interoperate that dataset in a tool whose name is the RStudio. Download the individual packages in that tool these packages will technically support that algorithm that the author us

e in the analysis of that dataset to extract the useful information from that dataset. Algorithms author use in this paper is C4.5 and C5.0. There is some common relation in these algorithms that both will create a decision tree as in output and will show the accuracy of the algorithms. In this paper, the author is going to extract information about the existence of kidney disease in the patient or not also author will make a comparison of these two machine learning algorithms and in finally tell about will algorithm is more accurate and best based on the accuracy of that machine learning algorithm. Fig: Methodology of chronic kidney disease detection is shown in "Fig. 1".

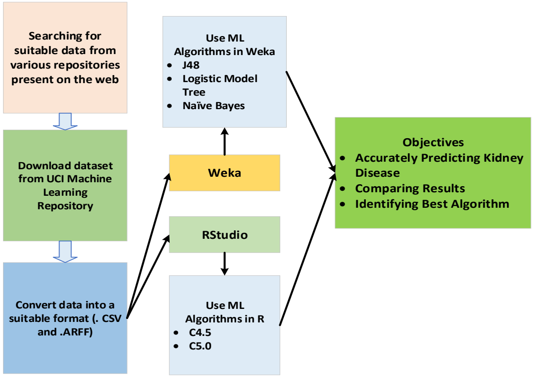
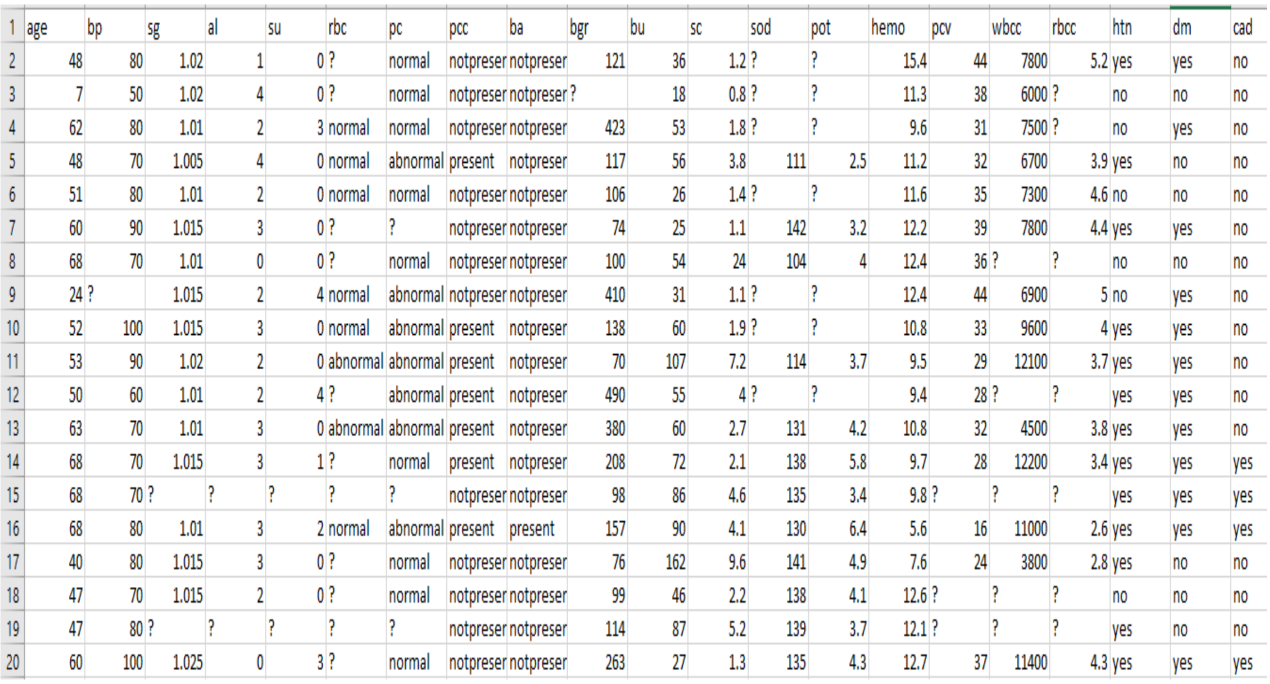

Searching for suitable data from various repositories present on the web
Download dataset from UCI Machine Learning Repository
Convert data into a suitable format (. CSV and .ARFF)
Weka
RStudio
Use ML Algorithms in Weka
• J48
• Logistic Model Tree
• Naïve Bayes
Use ML Algorithms in R
• C4.5
• C5.0
Objectives
• Accurately Predicting Kidney Disease
• Comparing Results
• Identifying Best Algorithm

Figure 1: The methodology of Chronic Kidney Disease Detection

4. Dataset Collection

In the dataset, each instance has its description, and every attribute of data represents variable; this is also called single statistical data. For classification and prediction of chronic disease, we use a dataset that uses for comparison of the accuracy by using different machine learning algorithms. The dataset used in our experiments has sown in "Fig. 2". 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 [8].



1	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad
2	48	80	1.02	1	0	?	normal	notpreser	notpreser	121	36	1.2	?	?	15.4	44	7800	5.2	yes	yes	no
3	7	50	1.02	4	0	?	normal	notpreser	notpreser	?	18	0.8	?	?	11.3	38	6000	?	no	no	no
4	62	80	1.01	2	3	normal	normal	notpreser	notpreser	423	53	1.8	?	?	9.6	31	7500	?	no	yes	no
5	48	70	1.005	4	0	normal	abnormal	present	notpreser	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no
6	51	80	1.01	2	0	normal	normal	notpreser	notpreser	106	26	1.4	?	?	11.6	35	7300	4.6	no	no	no
7	60	90	1.015	3	0	?	?	notpreser	notpreser	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no
8	68	70	1.01	0	0	?	normal	notpreser	notpreser	100	54	24	104	4	12.4	36	?	?	no	no	no
9	24	?	1.015	2	4	normal	abnormal	notpreser	notpreser	410	31	1.1	?	?	12.4	44	6900	5	no	yes	no
10	52	100	1.015	3	0	normal	abnormal	present	notpreser	138	60	1.9	?	?	10.8	33	9600	4	yes	yes	no
11	53	90	1.02	2	0	abnormal	abnormal	present	notpreser	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no
12	50	60	1.01	2	4	?	abnormal	present	notpreser	490	55	4	?	?	9.4	28	?	?	yes	yes	no
13	63	70	1.01	3	0	abnormal	abnormal	present	notpreser	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no
14	68	70	1.015	3	1	?	normal	present	notpreser	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes
15	68	70	?	?	?	?	?	notpreser	notpreser	98	86	4.6	135	3.4	9.8	?	?	?	yes	yes	yes
16	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes
17	40	80	1.015	3	0	?	normal	notpreser	notpreser	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no
18	47	70	1.015	2	0	?	normal	notpreser	notpreser	99	46	2.2	138	4.1	12.6	?	?	?	no	no	no
19	47	80	?	?	?	?	?	notpreser	notpreser	114	87	5.2	139	3.7	12.1	?	?	?	yes	no	no
20	60	100	1.025	0	3	?	normal	notpreser	notpreser	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes

Figure 2. Medical Datasets for Chronic Kidney Disease Detection

5. Results and Discussion

For that execution of this C4.5 and C5.0 algorithm, we utilize the R-Studio instrument. R language can be a statistical programming language that's an entirely free opensource package-based language. For app composing, this terminology is quite potent. Lots of roles have been already assembled inside this terminology. We will find lots of bundles utilized inside this language extend the cutting-edge search. As an alternative in line with this forecast of analyzing dataset C4.5 and a C5.0 decision tree was assembled by pest management info. Predictions are produced by people with all the WEKA data mining device for both accuracy and classification by directly employing distinct algorithmic procedures. We are assessing the consequences at both manners firstly we all discover that the optimal algorithm using the contrast of the various features like Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error and kappa statistic and so on. Second, the Truth of those calculations will examine different parameters such as TP Charge, FP Pace, Precision, Recall, F-Measure, MCC, ROC Spot, and PRC Area. That is visualized from the bar graph. The picked algorithm creates the bladder disorder discovery process automatic. Before accepting decision-related to persistent kidney disease that this prediction version utilized for specifying the kidney disorder longer exact.

5.1 C4.5

A C4.5 decision tree is using to signify the classification of pest training data that is shown in “Fig. 3”. By using the 24 attributes in the medical dataset used to create a decision tree. In this decision tree, there are three colors of scale ratio grey, white and black that start from 0 to 1. When we are level 0 on a scale ratio that will show that there is no chronic kidney disease and when this is at on ratio level 1 that shows that there is a chronic kidney disease at that scale also show with the black color in scale ratio. White color shows that there is no decision about the disease that there is exist in the patient or not.

In the C4.5 model, classification accuracy achieved shows that 98.2935% out of total 293 instances from which 288 are correctly classified, and 5 are not correctly classified, mean absolute error is 0.0331, kappa statistics is 0.9634 are outputs is shown in "Table. 2". Also, the confusion matrix that shows in "Table. 3" this matrix describes the actual accuracy of the algorithm.

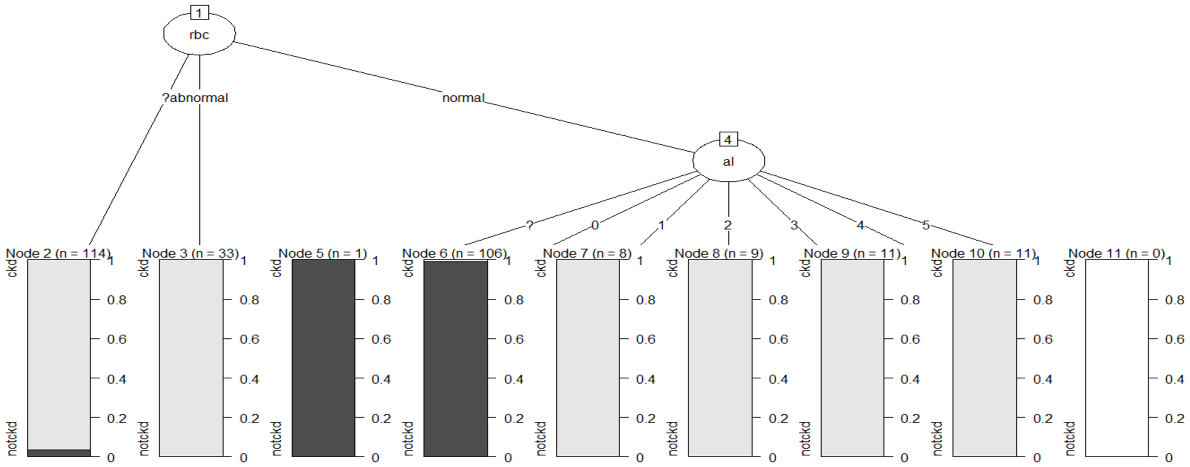


Figure 3. Decision Tree of C4.5 Algorithm

Table 2. Statistical results of C4.5 Algorithm

Correctly Classified Instances	288	98.2935%
Incorrectly Classified Instances	5	1.7065%
Kappa Statistics	0.9634	
Mean absolute error	0.0331	
Root mean squared error	0.1287	
Relative absolute error	7.0565%	
Root-relative square error	26.57%	
Total number of instances	293	

Table 3. Show the confusion matrix of C4.5 Algorithm

a	b	classified as
182	1	a = ckd
4	106	b = notckd

5.2 C5.0

A C4.5 decision tree is using to signify the classification of the pest by color that is shown in "Fig. 4". 99.49% of the data are correctly classified in the C5.0 model. The error rate in C5.0 is measured as 0.51%. By using the 24 attributes in the medical dataset used to create a decision tree. In this decision tree, there are three colors of scale ratio grey, white and black that start from 0 to 1. When we are level 0 on a scale ratio that will show that there is no chronic kidney disease and when this is at on ratio level 1 that shows that there is a chronic kidney disease at that scale show with the black color in scale ratio. White color shows that there is no decision about the disease that there is exist in the patient or not.

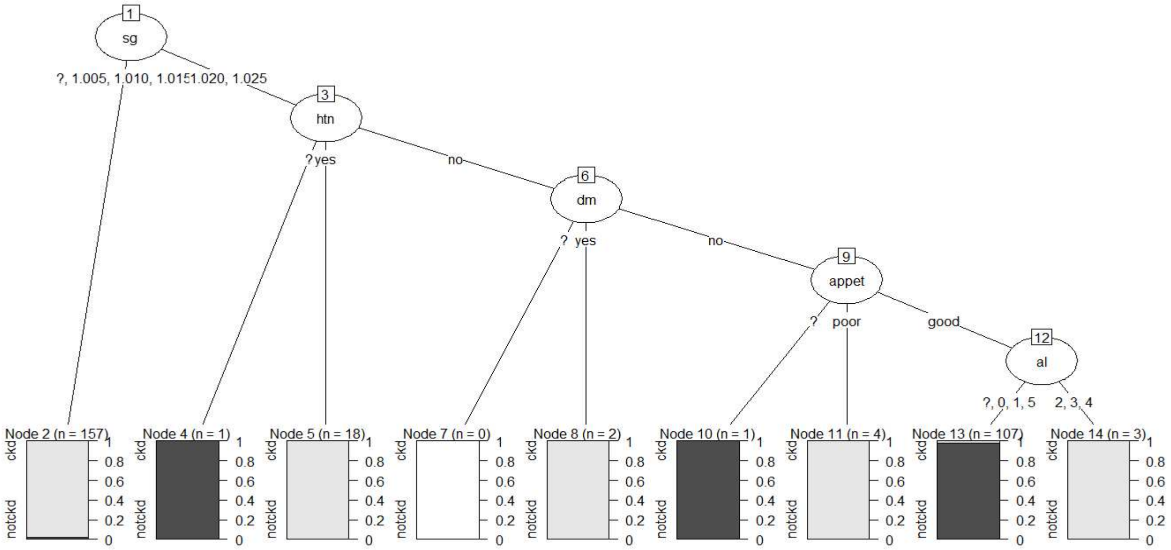


Figure 4. Decision Tree of C5.0 Algorithm

In the C5.0 model, classification accuracy achieved shows that 98.9761% out of total 293 instances from which 290 are correctly classified, and 3 are not correctly classified as shown in “Table. 4”.

Table 4. Show the confusion matrix of C5.0 Algorithm

(a)	(b)	classified as
180	3	(a): class ckd
	110	(b): class notckd

5.3 J48

We make use of the J48 algorithm in WEKA to research the degenerative kidney disease. From the outcome, we are pulling some statistical info regarding the algorithm which shows different parameters to spell out the truth of the algorithm shown at “Table. 5" classification precision realized implies that 99% from perfect 400 examples by which 396 are classified, an 1 are not properly categorized, imply total error is 0.0225, kappa stats is 0.9786 are signals. Additionally, in "Fig. 5" we have imagined different parameters at a bar chart which may demonstrate the truth of this algorithm in greater specifically. As a result, there is not a set of this bar shows in a bar chart. Series1 indicates the bar of Persistent kidney disease, series2 reveal the bar of this non-chronic kidney disease. Also, series3 indicates the weighted average of those variables defined from the bar chart.

Table 5. Statistical Information of J48 Algorithm

Correctly Classified Instances	396	99%
Incorrectly Classified Instances	4	1%
Kappa Statistics	0.9786	
Mean Absolute Error	0.0225	
Root Mean Squared Error	0.0807	
Relative Absolute Error	4.7997%	
Root Relative Squared Error	16.6603%	
Total Number of Instances	400	

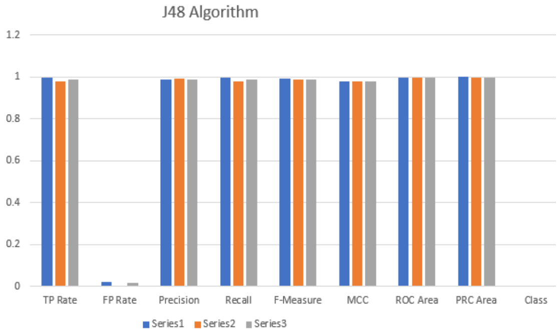


Figure 5. Barchart visualization of J48 Algorithm

5.4 Logistic Model Tree

We utilize the Logistic design Tree algorithm in WEKA to research the degenerative kidney disease. From the outcome, we're pulling some statistical info regarding the algorithm which shows diverse parameters to spell out the truth of the algorithm shown at "Table. 6" classification precision realized implies that 98% from absolute 400 examples by which 392 are classified and 8 aren't properly categorized, implies total error is 0.0222, kappa stats is 0.9577 are signals. Also, in "Fig. 6" we imagine different parameters at a bar chart which may demonstrate the truth of this algorithm in greater specifically. As a result, there is not a set of this bar shows in a bar chart. Series1 indicates the bar of Persistent kidney disease, series2 reveal the bar of this non-chronic kidney disease. Also, series3 indicates the weighted average of those variables defined from the bar chart.

Table 6. Statistical Information of Logistic Model Tree Algorithm

Correctly Classified Instances	392	98%
Incorrectly Classified Instances	8	2%
Kappa Statistics	0.9577	
Mean Absolute Error	0.0222	
Root Mean Squared Error	0.1068	
Relative Absolute Error	4.7237%	
Root Relative Squared Error	22.0573%	
Total Number of Instances	400	

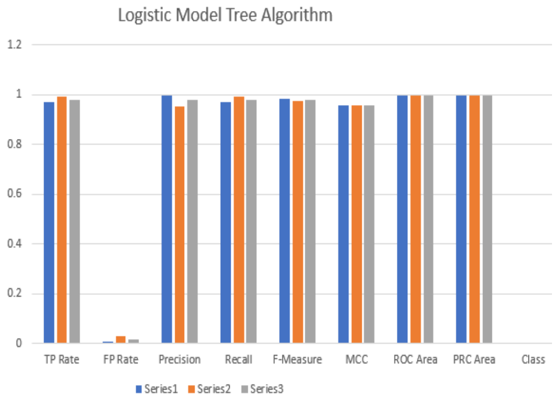


Figure 6. Barchart visualization of Logistic Modle Tree Algorithm

5.5 Naïve Bayes

We make use of the Naïve Bayes algorithm in WEKA to research the degenerative kidney disease. From the outcome, we are pulling some statistical info regarding the algorithm which shows different parameters to spell out the truth of the algorithm shown at "Table. 5" classification precision realized demonstrates 95% from perfect 400 examples by which 380 are classified, and 20 are not properly categorized, imply total error is 0.0479, kappa stats is 0.8961 are signals. Also, in "Fig. 7" we imagine different parameters at a bar chart which could demonstrate the truth of this algorithm in greater specifically. As a result, there is not a set of this bar shows in a bar chart. Series1 indicates the bar of Persistent kidney disease, series2 reveal the bar of this non-chronic kidney disease. Also, series3 indicates the weighted average of those variables defined from the bar chart.

Table 7. Statistical Information of Naïve Bayes Algorithm

Correctly Classified Instances	380	95%
Incorrectly Classified Instances	20	5%
Kappa Statistics	0.8961	
Mean Absolute Error	0.0479	
Root Mean Squared Error	0.2046	
Relative Absolute Error	10.2125%	
Root Relative Squared Error	42.2526%	
Total Number of Instances	400	

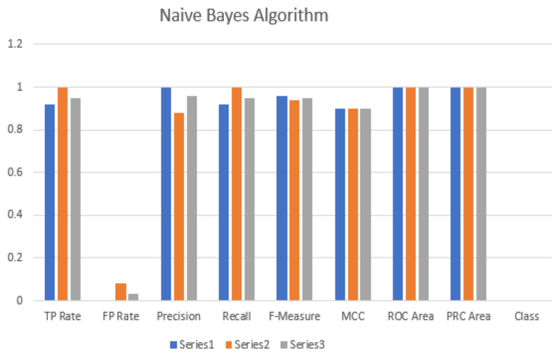


Figure 7. Barchart visualization of Naïve Bayes Algorithm

5.6 Discussion

"Table. 9" embraces the result of the decision tree, i.e., correctly classified instances, incorrectly classified instances, accuracy prediction.

Table 8. Comparison of Algorithms

Classifier	C4.5	C5.0
Correctly classified instances	288	290
Incorrectly classified instances	5	3
Accuracy Prediction	98.2935%	98.9761%

This research proved the efficiency of the C5.0 algorithm since it predicted more accuracy, short duration and less error rate as compared to the C4.5 algorithm.

Authors have used classification data mining technique in this paper using various algorithms such as Naïve Bayes, J48, and Logistic Model Tree with only one interface that is Knowledge flow. The parameters authors have used on which basis the results have been carried out are a total number of instances used either correctly classified or incorrectly classified, mean absolute error and kappa statistics. Algorithms accuracy is shown in the below Table. From the results, it is visible that j48 is the best performing algorithm with accuracy 99% also it has classified maximum number of correct instances, i.e., 396, has the least mean absolute error, i.e., 0.0225 and has maximum kappa statistics, i.e., 0.9786 as shown in "Table. 10".

Table 9. Statistical Compression Of All Algorithms

Algorithm	**Total Instances (1353)**		**Mean Absolute Error**	**Kappa Statistics**
	Correct	**Incorrect**		
J48	396	4	0.0225	0.9786
Naïve Bayes	392	8	0.0222	0.9577
Logistic Model Tree	380	20	0.0479	0.8961

6. Conclusions and Future Work

Utilizing R-Studio data mining tool forecasting persistent kidney disease is a primary goal of this paper. For the experimentation, we have utilized three calculations, i.e., C4.5 or C5.0. Afterward employed those calculations with machine learning processes from R-Studio tool for data mining. From the window conducting such calculations, precision is accessed that will be examined. Depending on the precision accomplished the presses were contrasted after conducting these calculations. These calculations assess classifier precision to each other on predicated kappa statistics, mean absolute error and correctly recorded instances and its indeed observable that C5.0 will be the greatest acting algorithm. We have utilized four calculations, i.e., Naïve Bayes, J48, and Logistic Model Tree to the experimentation. These calculations were executed employing the WEKA data mining procedure to successfully test algorithm precision that was accessed after conducting those calculations at the output. We assess the consequences at the two manners firstly we all discover the very optimal algorithm using the contrast of different features Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error and kappa statistics. Secondly, the accuracy of these algorithms will analyze with different parameters like TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area and PRC Area that's visualized from the bar chart. For identification, several different disorders like cancer therefore forth at the health care field that the software of those machine learning calculations stretched farther. Additionally, it helps people by employing different software of machine learning algorithms in resolving the medical exploration issue. For a forecast of all diseases utilizing these devices learning algorithms, yet still, another benefit is the fact that in the event of substantial data collections constituting lakhs of sufferers or a case as soon as the variety of sufferers to whom the prognosis must be achieved is how tremendous it readily introduces a disease.

Even though to analyze the overview of visualization of the result, Association Rule Mining, clustering and classification to predict disease among patient in medical field RStudio is a powerful data mining tool but to classify further different datasets such as MATLAB we can use other tools. By prediction of cancer and other diseases, in the future, we plan to extend this approach because it is used with chronic kidney disease dataset.

References

A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," The Lancet, vol. 389, no. 10075, pp. 1238-1252, 2017.

J. Coresh et al., "Prevalence of chronic kidney disease in the United States," Jama, vol. 298, no. 17, pp. 2038-2047, 2007.

S. Ruggieri, "Efficient C4. 5 [classification algorithm]," IEEE transactions on knowledge and data engineering, vol. 14, no. 2, pp. 438-444, 2002.

S.-l. Pang and J.-z. Gong, "C5. 0 classification algorithm and application on individual credit evaluation of banks," Systems Engineering-Theory & Practice, vol. 29, no. 12, pp. 94-104, 2009.

N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on a j48 algorithm for data mining," Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013.

K. Suresh and R. Dillibabu, "Designing a Machine Learning Based Software Risk Assessment Model Using Naïve Bayes Algorithm," 2018.

T. Miyato, S.-i. Maeda, S. Ishii, and M. Koyama, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," IEEE transactions on pattern analysis and machine intelligence, 2018.

L. J. R. R. Scholar). (2015-07-03). Chronic_Kidney_Disease Data Set. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#

R. Kannan and V. Vasanthi, "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease," in Soft Computing and Medical Bioinformatics: Springer, 2019, pp. 63-72.

C. L. Manske, Y. Wang, T. Rector, R. Wilson, and C. White, "Coronary revascularisation in insulin-dependent diabetic patients with chronic renal failure," The Lancet, vol. 340, no. 8826, pp. 998-1002, 1992.

P. G. Norris et al., "Immune function, mutant frequency, and cancer risk in the DNA repair defective genodermatoses xeroderma pigmentosum, Cockayne's syndrome, and trichothiodystrophy," Journal of Investigative Dermatology, vol. 94, no. 1, pp. 94-100, 1990.

J. Coresh, B. C. Astor, T. Greene, G. Eknoyan, and A. S. Levey, "Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey," American journal of kidney diseases, vol. 41, no. 1, pp. 1-12, 2003.

N. Tangri et al., "A predictive model for progression of chronic kidney disease to kidney failure," Jama, vol. 305, no. 15, pp. 1553-1559, 2011.

L. Jena and N. K. Kamila, "Distributed data mining classification algorithms for prediction of chronic-kidney-disease," International Journal of Emerging Research in Management &Technology, vol. 4, no. 11, pp. 110-118, 2015.

A. Kusiak, B. Dixon, and S. Shah, "Predicting survival time for kidney dialysis patients: a data mining approach," Computers in biology and medicine, vol. 35, no. 4, pp. 311-327, 2005.

T. B. Drüeke et al., "Normalization of hemoglobin level in patients with chronic kidney disease and anemia," New England Journal of Medicine, vol. 355, no. 20, pp. 2071-2084, 2006.

BLANK NOT TO BE PRINTED