International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING IN THE R-ENVIRONMENT

Nikhita Borde[1], Prof. Supriya Shanbhag[2]

[1]Student, Dept. of Electronics and Communication, KLS Gogte Institute of Technology, Karnataka, India
[2]Professor, Dept. of Electronics and Communication, KLS Gogte Institute of Technology, Karnataka, India

---

Abstract - *Chronic kidney disease refers to a condition where in the kidneys cannot perform its regular function of filtering blood. Based on how severe the Chronic Disease is, one has to opt for regular dialysis or at the end one should follow the procedure for a kidney transplant. Patients that are detected with CKD are the ones most likely suffering with high blood pressure or diabetes. Because of the change in the chores and lifestyle of people, there is a hike in the number of patients suffering with CKD. The risk of kidney transplant can be bypassed if the CKD can be detected at an earlier stage. By using the classification model built by study of machine learning and R-programming we can detect if an individual has CKD based on various parameters, thereby reaching our goal. And as a conclusion, we can predict the presence of CKD in a patient. And also take care of all the preventive measures.*

Key Words*: *Chronic kidney disease, regular dialysis, kidney transplant, machine learning, R-programming

1. INTRODUCTION

Chronic kidney disease also known as chronic renal disease is a very hazardous and life ominous disease which is becoming very common now a days. In order to deal with it, either it has to be detected at earliest stage possible or has to be given suitable treatment if detected [2]. There are around 1 million cases witnessed of Chronic Kidney Disease (CKD) by the patients per year in India [4]. Usually patients suffering with CKD have common symptoms like exhaustion, poor metabolism, shooting pain, inflamed feet as well as ankles. But we cannot detect the disease effectively based on only these symptoms [2]. But these symptoms are not useful to predict whether a patient is suffering from CKD as they are categorized to be common symptoms. The parameters such as age, blood pressure, diabetes, rbc count, wbc count, albumin etc. help us in determining the presence of CKD in any patient [1]. Machine learning is considered as a domain concerned with the study of several variable data and grown from the study of pattern recognition. It also has mathematical methods, algorithms and techniques for analysis and prediction [1]. In total there are 24 symptoms/parameters depending on which we can detect CKD in any patient with better accuracy [8].

1.1 The 24 Parameters used for Prediction

The below diagram depicts the symptoms in tabular format as follows:

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

Table -1: The 24 parameters

The last row (25th row) signifies the class, meaning whether the patient is suffering with CKD or not. It is not a symptom/parameter. Hence, the Dataset used comprises of 24 independent variables and 1 dependent variable. These 24 independent parameters are passed as input to the classifier model from machine learning library. The output of the classifier model is the class which is a dependent variable [2].

2. METHODOLOGY

The code for the Detection of CKD based on the 24 parameters was written in R-language. It is a free software environment. It is used for mathematical calculations and graphics.

2.1 Algorithm to predict presence of CKD:

The block diagram given below shows the flowchart of the Project.

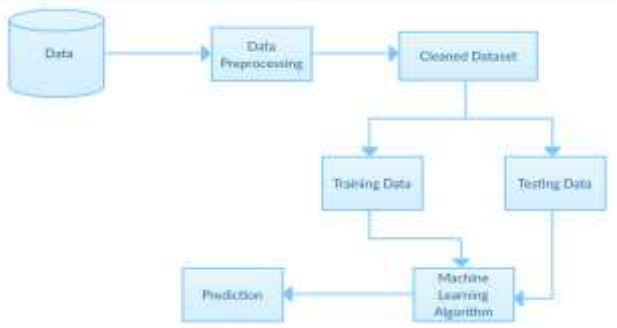

Fig -1: Block diagram depicting the steps involved in the prediction of CKD

2.1.1: Getting the Data:

The raw dataset was downloaded from the UCI repository. It consists of 400 observations. Out of which, 250 patients suffer with CKD and 150 do not [2]. The dataset comprises of 400 rows and 26 columns. The 1st column is the Id of the patient, the columns 2 to 25 determine the various parameters on which the prediction of CKD depends, and the 26th column depicts the class or the output.

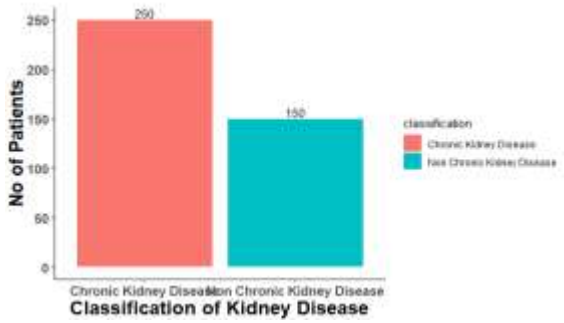

Chart -1: Graphical representation of the dataset

The parameters hold values of two different types:
1. nominal (yes/no) or (present/not-present)
2. numeric values (1,2,3..)

The data is said to be raw because it has many missing values and needs to be cleaned, manipulated and pre-processed for the next following steps. The above graph was obtained by making use of various visual packages present in R-studio.

2.1.2: Clean, Prepare and Manipulate data:

As the data is raw it needs to be cleaned, manipulated and pre-processed. This involves:
1. Finding all the null values
2. Finding all the missing values
3. Finding all the wrong values
4. Imputing the correct values in the respective places

To help the cleaning process of the null values and missing values, much graphical visualization are considered. An example is given below:

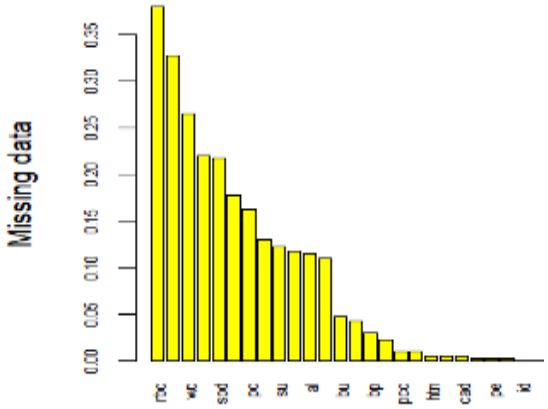

Chart -2: Graphical representation of the missing values in the columns

The above graph shows us the missing values in the respective columns, which can be imputed using packages present in R-studio.

2.1.3: Training the Model:

The dataset has been divided into training set and testing set. The split ratio is such that, 75% of the dataset is used as training set and the rest 25% is used as testing set. The split ratio is explicitly mentioned. The shuffling is done randomly by making use of suitable libraries and packages present in R. The training set is used to train the classifier model to understand the pattern of the data and the corresponding output (Supervised learning).
Here, the classifier model used to predict the presence of CKD is the Naïve Bayes algorithm.

Naïve Bayes Algorithm in R:

As we know, Bayes theorem is based on conditional probability and uses the formula:

P(B | A) = P(B) * P(A | B) / P(A)

A is the first event (person suffering with CKD)

P(A) is the probability of occurrence of event A

B is the second event (person not suffering with CKD)

P(B) is the probability of occurrence of event B

P(B/A) is the occurrence of event B given that event A has already occurred (also called as the conditional probability.

2.1.4: Testing the Model:

After the training of the model is complete, the model has learnt almost all patterns in the given training dataset. Now the model is provided with the remaining 25% of the dataset that was considered as test set. This step works as a validation process for the classifier model that was used to predict the CKD.

The output of this step gives us the Confusion matrix (for ease in mapping of the correctly predicted value and wrongly predicted value), the Accuracy, the Sensitivity and the Specificity of the prediction made by the classifier model.

The formulas used are:

2.1.4.1. Accuracy: $\frac{TP+TN}{TP+FP+FN+TN}$

TP = True positive (The output of the classifier model that stands true for the prediction of positive value)

TN = True Negative (The output of the classifier model that stands true for the prediction of negative value)

FP = False Positive (The output of the classifier model that stands false for the prediction of the positive value)

FN = False Negative (The output of the classifier model that stands false for the prediction of the negative value)

Accuracy takes values from 0 to 1. 0 being the least accurate prediction and 1 being the most accurate prediction.

2.1.4.2. Sensitivity: $\frac{TP}{TP+FN} * 100$

It is known as the rate of occurrences of True-Positives or Rate of correct Detection of positive values. Sensitivity measures the proportion of actual positives that are correctly identified.

Sensitivity takes values from 0 to 1. Sensitivity 0 means that the rate of wrong prediction of positive value is high. And Sensitivity 1 means that the rate of correct prediction of positive value is high.

2.1.4.3. Specificity: $\frac{TN}{TN+FP} * 100$

It is known as the rate of occurrences of False-Negatives or Rate of correct Detection of negative values. Specificity measures the proportion of actual negatives that are correctly identified.

Specificity takes values from 0 to 1. Specificity 0 means that the rate of wrong prediction of negative value is high. And Sensitivity 1 means that the rate of correct prediction of negative value is high.

2.1.5: Improving the performance of the model:

After the testing stage we obtain certain results as the Accuracy, Sensitivity and Specificity. To improve the results or to improve the functioning of the model for better results, certain procedure was followed.

Out of the 24 parameters that were used for prediction purpose, the model selects only the best suited parameters whose contribution in detecting the CKD is maximum. Hence, as a result of which we are left with only the most important parameters on which the prediction of CKD can be done with higher levels of Accuracy, Sensitivity and Specificity.

3. RESULTS

3.1. The Obtained **Confusion Matrix** is:

X	0	1
0	60	0
1	2	38

From the confusion matrix we understand that out of the 100 observations that were taken as the test set, the algorithm has obtained the following results:

- In the place [0,0] of the Confusion matrix we see the number 60, this means that the model has correctly predicted the True-Positives.
- In the place of [0,1] of the Confusion matrix we see the number 0, this means that the model did not predict any False-Positives.
- In the place of [1,0] of the Confusion matrix we see the number 2, this means that the model has predicted 2 False-Negatives.
- And, in the place of [1,1] of the Confusion matrix we see the number 38, this means that the model has correctly predicted 38 True-Negatives.

3.2. The **Accuracy** of the CKD prediction: 0.98 (**98%)**
The accuracy of the model is 98%, the remaining 2% was lost because the 2 wrong predictions (False-Negatives).

3.3. The **Sensitivity** of the CKD prediction: 0.9677 (**96.77%**)
Sensitivity of the model refers to the accuracy in predicting the True-Negatives. Because we see 2 False-Negatives, we obtain the sensitivity of the model as 96.77%.

3.4. The **Specificity** of the CKD prediction: 1 (**100%**)
The Specificity refers to the accuracy in predicting the True-Positives. Because we see that all of the 60 values are True-Positive, the Specificity of the model is 100%.

3.5. The results in tabular format

PARAMETERS	RESULTS IN %
1. Accuracy	98%
2. Sensitivity	96.77%
3. Specificity	100%

Table -2: Obtained results

4. CONCLUSIONS

The dataset was split into training set and testing set in the ratio 75:25. The training set was used to train the classifier model. The testing set was used to validate the classifier model. And the above results were obtained.

Another procedure of validation was performed where in any 1 random row of the test set was chosen manually and given as input to the classifier model. The obtained result stood true for both True-Positive as well as True-Negative. Hence this model works efficiently provided we have values for all parameters listed above using R-programming language libraries and packages. And the above stated results are observed.

REFERENCES

[1] Manish Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm "International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, February- 2016, pg. 24-33

[2] "A Review Paper on Chronic Kidney Disease Detection" Milandeep Arora , Er. Ajay Sharma, International Journal of Engineering Trends and Applications (I JETA) – Volume 3 Issue 4, Jul-Aug 2016

[3] " Detecting Chronic Kidney Disease Using Machine Learning", Manoj Reddy, John Cho, ICT Pillar, http://dx.doi.org/10.5339/qfarc.2016.ICTSP1534

[4] "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm", Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar, International Journal of Advanced Research in Computer and Communication Engineering Vol.7, Issue 10

[5] " Review on Data Mining Techniques for Prediction of Chronic Kidney Disease" Pallavi Sharma, Gurmanik Kaur, International Journal of Engineering Trends and Technology (IJETT) – Volume 63 Number 1 – September 2018

[6] " Data Mining Classification Algorithms For Kidney Disease Prediction "Dr. S. Vijayarani, Mr.S.Dhayanand, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015

[7] "Diagnosis of Chronic Kidney Disease by Using Random Forest Abdulhamit Subas, Emina Alickovic, and Jasmin Kevric College of Engineering, Effat University, Jeddah, 21478, Saudi Arabia"

[8] "Kidney Disease Prediction Using SVM and ANN Algorithms", Dr. S. Vijayarani, Mr.S.Dhayanand, International Journal of Computing and Business Research (IJCBR)
ISSN (Online): 2229-6166, Volume 6 Issue 2 March 2015

[9]https://en.wikipedia.org/wiki/Chronic_kidney_disease