PREDICTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHM

Article by :

- U.Dhulaja Bavani
- N.Jayalakshmi
- K.Moulika
- Ajitha

Prediction of Chronic Kidney Disease Using Machine Learning Algorithm

Abstract: In today's era everyone is trying to be conscious about health although due to workload and busy schedule one gives attention to the health when it shows any symptoms of some kind. But CKD is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be lead to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. By using data of CKD patients with 14 attributes and 400 record we are going to use various machine learning techniques like Decision Tree, SVM, etc. To build a model with maximum accuracy of predicting whether CKD or not and if yes then its Severity.

Keywords: CKD, Decision Tree, GFR, SVM, Machine Learning

I. INTRODUCTION

We all knows, that Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation. In simple words we can say that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, weekboans, poor nutrition health and nerve damage, Decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of your blood creatinine, age, race, gender, and other factors. The earlier disease is detected the better chance of showing or stopping its progression. Based upon GFR the renal damage severity by CKD is categorized into following five stages:

Table 1: Stages of CKD [2]

Stage	**Description**	**GFR(mL/min)**
-	At increased risk for CKD	>=90 with risk factors
1	Kidney damage with normal or increased GFR	>=90
2	Mild decrease in GFR	60-89
3	Moderate decrease in GFR	30-59
4	Severe decrease in GFR	15-29
5	Kidney Failure	<15 or dialysis

Santosh A. Shinde and Dr. P. Raja Rajeswari presented a machine learning concept map and review on applications of machine learning in healthcare domain in order to predict different disease, intellectually [5]. Machine Learning is one such tool which is widely utilized in different domains because it doesn’t require different algorithm for different dataset. In medical science CKD is one of the major challenges; because a lot of parameters and technicality is involved for accurately predicting this disease. Machine learning could be a better choice for achieving high accuracy for predicting not only CKD but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the CKD, algorithms such as Naive Bayes, Decision Tree, KNN, Neural Network, are used to predicate risk of CKD each algorithm has its specialty such as Naive Bayes used probability for predicting CKD, whereas decision tree is used to provide classified report for the CKD, whereas the Neural Network provides opportunities to minimize the error in prediction of CKD [1]. All these techniques are using old patient record for getting prediction about new patient. This prediction system for CKD helps doctors to predict heart disease in the early stage of disease resulting in saving millions of life.

II. LITERATURE SURVEY

There are many researchers who work on prediction of CKD with the help of many different classification algorithm. And those researchers get expected output of their model.

Gunarathne W.H.S.D et.al. [1] Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes.

S.Ramya and Dr.N.Radha [2] worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD according to its gravity. By analysing different algorithms like Basic Propagation Neural Network, RBF and RF. The analysis results indicates that RBF algorithm gives better results than the other classifiers and produces 85.3% accuracy.

S.Dilli Arasu and Dr. R. Thirumalaiselvi [3] has worked on missing values in a dataset of chronic Kidney Disease. Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing so they got up with unknown values. They replaced missing values with recalculated values.

Asif salekin and john stankovic [7] they use novel approach to detect CKD using machine learning algorithm. They get result on dataset which having 400 records and 25 attributes which gives result of patient having CKD or not CKD. They use k-nearest neighbours, random forest and neural network to get results. For feature reduction they use wrapper method which detect CKD with high accuracy.

Pinar Yildirim [8] searches the effect of class imbalance when we train the data by using development of neural network algorithm for making medical decision on chronic kidney disease. In this proposed work, a comparative study was performed using sampling algorithm. This study reveals that the performance of classification algorithms can be improved by using the sampling algorithms. It also reveals that the learning rate is a crucial parameter which significantly effect on multilayer perceptron.

Sahil Sharma, Vinod Sharma, and Atul Sharma [9], has assessed 12 different classification algorithm on dataset which having 400 records and 24 attributes. They had compared their calculated results with actual results for calculating the accuracy of prediction results. They used assessment metrics like accuracy, sensitivity, precision and specificity. They find that the decision tree technique gives accuracy up to 98.6%, sensitivity of 0.9720, and precision of 1 and specificity of 1.

Dataset and Attributes: In this paper CKD dataset [4] is downloaded from UCI repository. This dataset includes 400 patients' records with 25 attributes. All this 25 attributes are main attributes which are related to CKD disease. Out of 25 attributes we only use 14 attributes to build our predictive model.

III. METHODOLOGY

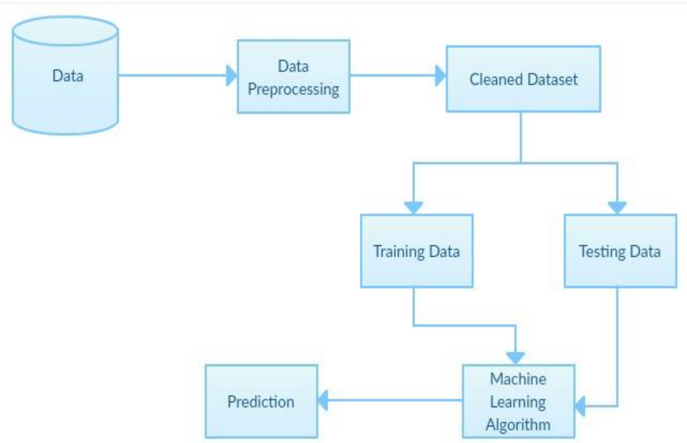

Fig 1.1 Architecture Diagram

i) Dataset - Dataset of **prediction of chronic kidney disease using machine learning algorithm** is downloaded from UCI repository. In that dataset there are 400 patient records are included. Also they include 25 attributes but we take only 14 attributes for building model. Age, Blood pressure, Albumin, Red blood cells, Pus cell, Pus cells clumps, Serum creatinine, Haemoglobin, White blood cell count, Red blood cell count, Anaemia, Classification, Appetite, Packed cell volume all this 14 attributes are used to build model.

ii) Data Prepossessing:

Data Cleaning: Gather open source raw data of CKD patients available on internet. Data obtained from internet does not contains the name of the attribute so first we assigned the names to the attribute. Missing values in the dataset like NA's or blank values are removed by using WEKA function "ReplaceMissingValues" used, which replaces NA's with the mean values of that attribute.

Data Reduction: Out of 25 attributes present in the dataset, we have selected 14 important attributes required to build predictive model. Following table shows the selected attributes:

Table 2: Attributes of Dataset [1]

Attribute	Value Used
Age	Discrete Integer Values
Blood Pressure	Discrete Integer Values
Albumin	Nominal Values
Red Blood cells	Nominal Values(Normal, Abnormal)
Pus cell	Nominal Values(Normal, Abnormal)
Pus cells clumps	Nominal Values(Present, Not-Present)
Serum creatinine	Numeric Values
Haemoglobin	Numeric Values
White blood cell count	Discrete Integer Values
Red blood cell count	Numeric Values
Anaemia	Nominal Values(Yes, No)
Classification	Nominal Values(CKD, Not CKD)
Appetite	Nominal Values(Good, Poor)
Packed cell volume	Discrete Integer Values

iii) Training and Testing Dataset: The dataset is divided into two sub datasets both containing 14 attributes.

Training data: training dataset is derived from main dataset and it contains 300 out of 400 records in main dataset of CKD.
Testing data: testing dataset is of 100 out of 400 records from main CKD dataset.

iv) Classifiers:

Decision Tree: Decision tree is a graphical representation of specific decision situation that used for predictive model, main component of decision tree involves root, nodes, and branching decision. Decision tree is used in those area of the medical science where numerous parameters involved in classification of data set. Since decision tree is most compressive approach among all machine learning algorithm. These clearly reflect important features in the data set. They can also generate the most affecting feature in the mass of population. Decision tree is based on entropy and Information gain clearly signifies the importance of dataset. Drawback of decision tree is that it suffers from two major problems overfitting and it is based on greedy method. overfitting happened due to decision tree split dataset aligned to axis it means it need a lot of nodes to split data, this problem is resolved by J48 explained in based on greedy method lead to less optimal tree, if dynamic approach is taken it lead to exponential number of tree which is not feasible[6].

Support Vector Machine: "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well .Support Vectors are simply the coordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line) [6].

Steps for Calculation of Hyperplane:
1. Set up training data
2. Set up SVM parameter
3. Train the SVM
4. Region classified by the SVM
5. Support vector

v) Prediction:

Prediction using Decision tree:

Prediction done by Decision tree model trained by training CKD dataset shows following results:

Prediction using Decision Tree

Parameter	Results
Total Number of Instances	400
Correctly Classified Instances	367
Incorrectly Classified Instances	33
Kappa	0.827
Mean absolute	0.1396
Root mean squared Error	0.2699
Relative absolute error	29.7783 %
Root relative squared error	55.7551 %
F1 Measure	0.8958990
Precision	0.8502994
Recall	0.9466666
TNR	0.9000000
TPR	0.9466666
FNR	0.0533333
FPR	0.1000000

Prediction using SVM:

Prediction done by SVM model trained by training CKD dataset shows following results:

Parameter	Results
Total Number of Instances	400
Correctly Classified Instances	387
Incorrectly Classified Instances	13
Kappa	0.9315
Mean absolute	0.0325
Root mean squared Error	0.1803
Relative absolute error	6.9308 %
Root relative squared error	37.2379 %
F1 Measure	0.9579
Precision	0.9308
Recall	0.9866
TNR	0.956
TPR	0.9866
FNR	0.0133
FPR	0.044

IV. FUTURE SCOPE

This work will be considered as basement for the healthcare system for CKD patients. Also extension to this work is that implementation of deep learning since deep learning provides high-quality performance than machine learning algorithm.

CONCLUSION

In this paper we have studied different machine learning algorithms. We have analysed 14 different attributes related to CKD patients and predicted accuracy for different machine learning algorithms like Decision tree and Support Vector Machine. From the results analysis, it is observed that the decision tree algorithms gives the accuracy of 91.75% and SVM gives accuracy of 96.75%. When considering the decision tree algorithm it builds the tree based on the entire dataset by using all the features of the dataset. The advantage of this system is that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our respected and knowledgeable guide, parents and lecturers of VPKBIET Baramati and others who provide their valuable guidance and time for guiding us.

REFERENCES

[1]. Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)”,2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[2]. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

[3]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, “Review of Chronic Kidney Disease based on Data Mining Techniques”,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505

[4]. L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,”2015. [Online].Available:http://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.

[5]. S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning : a review,” IJET, vol. 7, no. 3, pp. 1019–1023, 2018.

[6]. Himanshu Sharma,M A Rizvi,”Prediction of Heart Disease using Machine Learning Algorithms: A Survey”,International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8

[7]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[8]. Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE International Conference on Computer Software and Applications (COMPSAC), IEEE, Jul. 2017, doi: 10.1109/COMPSAC.2017.84

[9]. Sahil Sharma, Vinod Sharma, Atul Sharma, “Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis,” July18, 2016.