INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1781)
Available online at: https://www.ijariit.com

Chronic kidney disease prediction using Machine Learning

Dr. K. V. Sambasiva Rao
kvsambasivarao65@gmail.com
NRI Institute of Technology, Vijayawada, Andhra Pradesh

J. Bhagya Lakshmi
blakshmi1108@gmail.com
NRI Institute of Technology, Vijayawada, Andhra Pradesh

E. Likhitha
likhithasrinivas444@gmail.com
NRI Institute of Technology, Vijayawada, Andhra Pradesh

Ch. Lakshmi Kalyani
kalyani.chilakapati99@gmail.com
NRI Institute of Technology, Vijayawada, Andhra Pradesh

ABSTRACT

The feature of Kidneys is to clearout the blood and to do away with wastes from the body. All of the blood in our body passes through the kidneys several instances a day.The kidneys get rid of wastes, manage the body's fluid stability, and modify the stabilityof electrolytes, it's feasible to lose as a good deal as 90% of kidney function without experiencing any signs and symptoms or problem. Kidney ailment is a silent killer detection and prevention need to be achieved before the scenario receives even worse and thesituation is known as continual kidney sickness also known as chronic kidney disease (CKD) or chronic renal disorder. It is essential to have powerful methods for early prediction of CKD, machine learning strategies are powerful in CKD prediction , In this paper statistics is acquired from the patients and then predictionis done whether that individual is having any continual kidney disorder or not.

Keywords: *Machine Learning, SVM, Random Forest, Chronic Kidney Disease, Disease Prediction*

1. INTRODUCTION

The kidneys are one among the foremost important organs of human body; they filter waste and excess fluids from the blood. If kidneys fail to figure , waste build up within the body. Symptoms of renal failure aren't very specific to the disease. Some symptoms may include body pains, back pains, anemia, and weak bones. Some people haven't any symptoms in the least and are diagnosed by a lab test. Medication helps manage symptoms. But what if an individual doesn't have any quite symptoms, this might cause individual fatal. So detection and prevention of the disease is necessary.

Generally the disease **are often** identified when it crosses the initial stages andbecomes severe, severity of disease may results in death, therefore the patient should get to understand about the disease in earlier stages, it's going to not be possible through the optical observation of symptoms by the doctors. Now the work is to form this process as easy as possible to assist doctors to in identifying the disease in order that it helps to scale back the damage caused by the disease.
This paper mainly specialize in disease prediction because the disease prediction should be done as quickly as possible otherwise it shows negative results onpatients health

2. LITERATURE REVIEW

P. Sinha compared KNN classifier and support vector machine(SVM) they used matlab in which SVM performed well that KNN [1]

K.R. Lakshmi studied the capability of kidney dialysis using Artificial Neural Network (ANN), Decision Tree, and logistic regression [2]

D. Sunil used data mining technologies such as Naïve Bayes and Artificial Neural Networks for the chronic kidney disease prediction [3]

The analysis system proposed by S. Dayanand to predict internal organ diseases with the help of support vector machine and Artificial neural network in 2015 [4]
H. Zhang investigated the performance of Artificial Neural Network (ANN)models while applying to the prediction on chronic kidney disease patients [5]

The existing system uses some traditional data mining techniques for the prediction which gave the scope for building a model that uses latest machine learning like SVM and Random forest algorithm for prediction of chronickidney disease

3. FACTORS CAUSING KIDNEY DISEASE

Kidney disease occurs when a disease or condition impairs kidney function, causing kidney damage. Diseases and conditions that cause chronickidney disease include:

- Type 1 or Type 2 diabetes
- High blood pressure

- Inflammation on the kidney's filtering units.
- Inflammation of the kidney's tubules and surrounding structure.
- Reflux is a condition where it causesurine to back up into the kidneys.
- Kidney stones.
- Some type of cancers.
- Recurrent kidney infection

4. MODEL DEVELOPMENT

In this paper, it is proposed that, the machine learning models are used to detect the Chronic Kidney Diseases. Machine learning algorithms are best used for the early diagnosis of CKD. The utilization of machine learning models to detect Chronic Kidney Diseases is proposed during this study. Machine learning techniques are better at detecting CKD early on. SVM and Random Forest are two of the foremost used algorithms. These approaches are extremely useful during a sort of disciplines, includingmedical diagnostics. It is proposed to train the Machine Learning Algorithm with the pre- existing data sets to acknowledge the CKDs.

4.1 Architecture

Here the model is built by taking data from the patient and applying machine learning algorithms like SVM, and Random forest for prediction. The architecture of the system is as shown in the figure 1.

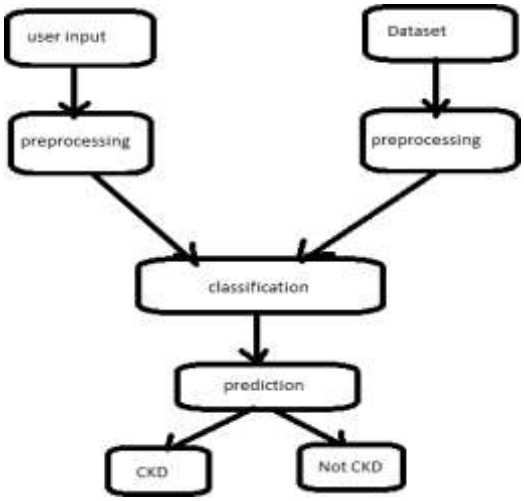

Fig 1: Architecture

4.2 Data set

The dataset is taken from UCI Machine learning repository. The data contains400 records and 25 attributes those 400 record contain 400 patients data

Attribute name	Attribute value	Attribute code
Age	Years	age
Blood Pressure	Mm/hg	bp
Specific gravity	1.005, 1.010, 1.015, 1.020, 1.025	sg
Albumin	0, 1, 2, 3, 4, 5	Al
Sugar	0, 1, 2, 3, 4, 5	Su
Red blood cells	normal, abnormal	rbc
Puss cell	normal, abnormal	Pc
Puss clell clumps	present, notpresent	pcc
Bacteria	present, notpresent	Ba
Bloodglucoserandom	mgs/dl	bgr
Blood urea	mgs/dl	Bu
Serum creatine	mgs/dl	Sc
Sodium	mEq/l	sod
Potassium	mEq/l	pot
Hemoglobin	gms	hemo
Packed cell volume	-	pcv
Whiteblood cellcount	cells/cum m	wbcc
Red blood cellcount	millions/cumm	rbcc
Hypertension	yes, no	htn
Diabetes Mellitus	yes, no	dm
Coronary Arterydisease	yes, no	cad
Appetite	good, poor	appet
Pedal edema	yes, no	pe
Anemia	yes, no	ane
Class	ckd, notckd	class

Preprocessing: Preprocessing involves twomajor tasks First one is to suit transform the information means transforming the nominal values into numeric data, for instance take a column Appetite this column contains valueseither good or poor we start replacing goodwith 0 and poor with 1 or good with 1 and poor with 0. This task is completed with theassistance of label encoder other involvesreplacement of null values this two concepts they are:

Replacing null numeric values: We replace null numeric values of a particular column with the mean of that whole column.

Replacing null nominal values: We replace all the nominal values of a particular column with the majorly repeated value of that particular column.

Classification: Classification is a predictivemodelling problem in which we predict the class label for the given data. Here we use classification technique because we need to find the class label (whether ckd or not ckd) forthe data we are giving.

Here we use two classification algorithms they are support vector machine (SVM) and Random forest algorithm

Support vector machine (SVM): Support vector machine is a supervised machine learning algorithm which is capableof performing both classification and regressionbut here we would like to perform onlyclassification so that's why we elect support vector classifier for the analysis and therefore the prediction purpose.

The linear support vector classifier works by drawing a line between two classes (here between ckd class and not ckd class). All the data points that lie on one side of the line are going be labelled as one class (ckd) and the points that lie on side are going to be labelled as second class (not ckd).

Algorithm will select a line that not only separates the two classes but stays as far away from the closest samples possible

Random forest algorithm: Random forest algorithm is also capable of performing classification as well as regression but here we need to perform only classification that's why we use random forest classifier algorithm.

Basically random forest contains hundreds and thousands of decision trees. It trains all on a rather different set of observations, splittingnodes in each tree considering a limited number of features. The ultimate predictions of the random forest are made by averaging thepredictions of every individual tree.

Prediction: The prediction is done on the information given by the user which can be done through preprocessing and applying the model that was built supported theclassification algorithms we have used like support vector classifier and random forest classifier.

5. PROCESSING STEPS

This project is developed by followingsequence of steps which are mentioned below:

- Step 1: collect the dataset whichcontains the data of the patients.
- Step 2: Dataunder goes somepreprocessing.
- Step 3: Train the model with thehelp of train and test data.
- Step 4: Take data from the user in the form of comma separated values.
- Step 5: Preprocess the data enteredby the user.
- Step 6: send data to the model.
- Step 7: Model will analyze that data and predict whether that person is having chronic kidney disease or not.

6. DEMONSTRATION

The dataset collected will contain 400 records and 25 attributes as mentioned before the dataset representation is like

Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba	Bgr	Bu	Sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
48	80	1.02	1	0		normal	notpreser	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
7	50	1.02	4	0		normal	notpreser	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
62	80	1.01	2	3	normal	normal	notpreser	notpresent	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
51	80	1.01	2	0	normal	normal	notpreser	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
60	90	1.015	3	0			notpreser	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
68	70	1.01	0	0		normal	notpreser	notpresent	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
24		1.015	2	4	normal	abnormal	notpreser	notpresent	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
50	60	1.01	2	4		abnormal	present	notpresent	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
68	70	1.015	3	1		normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd

Fig 2: Dataset representation

This data set contains all the patient's data who has chronic renal disorder and who is not having chronic renal disorder. These patients data is given as input for splitting into training and testing data which are helpful in training the model for prediction.

After data set collection we performpreprocessing for which we need to read the dataset as a csv file with .csv extension, and then we perform all the three major tasks in preprocessing.

Fit transforming, Replacing null numeric values with mean of that column, replacing null nominal values with majorly repeated values of that column

SVC classifier and its accuracy

```
train_predict = model_svc.predict(x_train)
test_predict = model_svc.predict(x_test)
from sklearn.metrics import accuracy_score
print('SVC',accuracy_score(y_test, model_svc.predict(x_test)))
print('Training Accuracy :', accuracy_score(train_predict,y_train))
print('Testing Accuracy: ',accuracy_score(test_predict,y_test))
```

```
SVC 0.6916666666666667
Training Accuracy : 0.5964285714285714
Testing Accuracy:  0.6916666666666667
```

Fig 3: SVC implementation Random forestclassifier and its accuracy

```
RandomForestClassifier()
```

```
train_predict = model_rfc.predict(x_train)
test_predict = model_rfc.predict(x_test)
from sklearn.metrics import accuracy_score
print('RFC',accuracy_score(y_test, model_rfc.predict(x_test)))
print('Training Accuracy :', accuracy_score(train_predict,y_train))
print('Testing Accuracy: ',accuracy_score(test_predict,y_test))
```

```
RFC 0.9916666666666667
Training Accuracy : 1.0
Testing Accuracy:  0.9916666666666667
```

Fig4: Random forest classifier implementation

Here SVC classifier got the training accuracy of 59% and testing accuracy of 69%, and Random forest classifier got the training accuracy of 100% and testing accuracy of 99%,among these two random forest assures us to provide better and accurate results.

7. TESTING

First step in the testing process involves gettingdata from the user in the form of csv file.

Profile of Patient:

Here we have taken 15 patients data from a hospital in Vijayawada for the prediction ofdisease**.**

Among them there are patients with type-1 diabetes and type-2 diabetes who are of age 54 and 67.

A patient of age 48 is suffering with anemia, and some with poor appetite.

A patient of age 34 is suffering with cancer andthere are some abnormalities in red blood cells and puss cells count.

Some patients are having gall stones andsome are suffering with inflammation in kidney's structure.

There are some patients with the age of 48 and 54 who are suffering with coronary arterydisease.

After collecting data some preprocessing techniques are applied on the test data file so that the test data file is compatible with thetraining which is used while training the model.

After preprocessing we need to see among support vector machine and random forest classifier which one have highest accuracy.

Here in our project we got 69 percent accuracy in support vector classifier and 99 percent accuracy in random forest classifier.

So we have used random forest classier for our prediction, we have given our test data set to the model which have two records in it which represents two patients data .

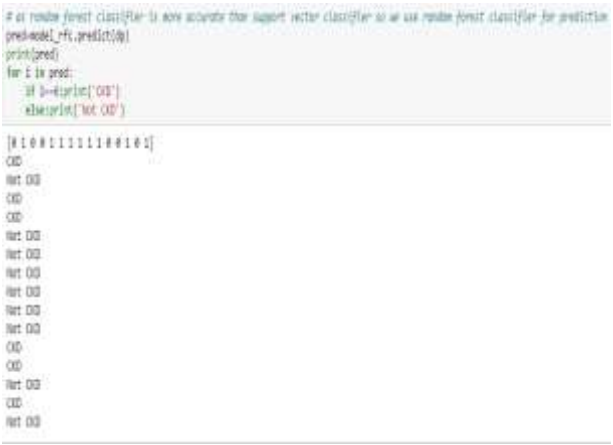

Fig 5: Prediction of CKD

We got the results as CKD for the patient having chronic kidney disease and Not CKDfor the patient who is not having chronic kidney disease.

8. CONCLUSION AND FUTURE SCOPE

It is observed from the results that the project is functioning well so as to realize specified goals and provides appropriate results. If the patient have the disease then it will say he/she has the disease, if the person don't have the disease then it will say he/she don't have the disease. system is employed to decrease the cost and offer better results.

This project features a vast scope within the future for enhancement. It is going toinclude more number of patient's data in order that the project can give more and more accurate results. We will use latest machine learning algorithms to aquire accurate results.

9. REFERENCES

[1] P. Sinha 2015. Comaparative Study of Chronic renal Disease Prediction by Using KNN and SVM. International Journal of Engineering Research and Technology (IJERT), Volume 4, Issue No. 12.

[2] K.R. Lakshmi, Y. Nagesh, and M. VeeraKhrisna. 2014. Performance Comparison of Three Data MiningTechniques for Predicting Kidney DialysisSurvivability. International Journal of Advances in Engineering and Technology, Volumen 7, Issue No.1, pp. 242-254.

[3] D. Sunil 2017. Chronic Kidney Disease Analysis Using Data Mining.International Journal of Scientific Research in Computer Science, Engineering, and Information Technology (IJSRCSEIT), Volume 2, IsssueNo. 4.

[4] H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang "chronic kidney disease survival prediction with artificial neuralnetworks" 2018 IEEE International

[5] Conference in Bioinformatics and Biomedicine (ICBB), Madrid, Spain, 2018.

[6] S. Dayanand 2015, Kidney diseaseprediction using SVM and ANN algorithms. International Journal of Computing BusinessResearch (IJCBR), Volume 6, Issue No.2.

