Chronic kidney disease prediction using Machine Learning

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

E. Likhitha
likhitasrinivas444@gmail.com
NRI Institute of Technology, Vijayawada, Andhra Pradesh

J. Bhagya Lakshmi
blakshmi1108@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 stability of electrolytes, it's feasible to lose as a good process of electrolytes, it's feasible to lose as a good 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 the situation 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 predictions 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 and becomes 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 on patients health.

2. LITERATURE REVIEW

P. Sinha compared KNN classifier and support vector machine(SVM) they used matlab in which SVM performed well than 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 chronic kidney 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 chronic kidney disease include:

- Type 1 or Type 2 diabetes
- High blood pressure
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, including medical 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.

Preprocessing: Preprocessing involves two major 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 values either good or poor or we start replacing good with 0 and poor with 1 or good with 1 and poor with 0. This task is completed with the assistance of label encoder other involves replacement 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 predictive modelling 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) for the data we are giving.

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

4.2 Data set

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

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

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.
Basically random forest contains hundreds and thousands of decision trees. It trains all on a rather different set of observations, splitting nodes in each tree considering a limited number of features. The ultimate predictions of the random forest are made by averaging the predictions 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 the classification algorithms we have used like support vector classifier and random forest classifier.

### 5. PROCESSING STEPS

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

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Collect the dataset which contains the data of the patients.</td>
</tr>
<tr>
<td>2</td>
<td>Data under goes some preprocessing.</td>
</tr>
<tr>
<td>3</td>
<td>Train the model with the help of train and test data.</td>
</tr>
<tr>
<td>4</td>
<td>Take data from the user in the form of comma separated values.</td>
</tr>
<tr>
<td>5</td>
<td>Preprocess the data entered by the user.</td>
</tr>
<tr>
<td>6</td>
<td>Send data to the model.</td>
</tr>
<tr>
<td>7</td>
<td>Model will analyze that data and predict whether that person is having chronic kidney disease or not.</td>
</tr>
</tbody>
</table>

#### 6. DEMONSTRATION

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

![Fig 2: Dataset representation](image)

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 the data set collection we perform preprocessing 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_predct = model_svc.predict(x_train)
test_predct = model_svc.predict(x_test)
from sklearn.metrics import accuracy_score
print("SVC: accuracy_score(y_test, model_svc.predict(y_test))")
print("Training Accuracy: ", accuracy_score(train_predct,y_train))
print("Testing Accuracy: ", accuracy_score(test_predct,y_test))
```

SVC: 0.6566666666666667
Training Accuracy: 0.5984285714285714
Testing Accuracy: 0.6566666666666667

**RandomForestClassifier**

```
train_predct = model_rfc.predict(x_train)
test_predct = model_rfc.predict(x_test)
from sklearn.metrics import accuracy_score
print("RFC: accuracy_score(y_test, model_rfc.predict(y_test))")
print("Training Accuracy: ", accuracy_score(train_predct,y_train))
print("Testing Accuracy: ", accuracy_score(test_predct,y_test))
```

RFC: 0.9016666666666667
Training Accuracy: 1.0
Testing Accuracy: 0.9916666666666667

**Fig 3:** SVC implementation Random forest classifier and its accuracy

**Fig 4:** 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 forests assure us to provide better and accurate results.

**7. TESTING**

First step in the testing process involves getting data 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 of disease.

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 and there are some abnormalities in red blood cells and pus cells count.

Some patients are having gall stones and some are suffering with inflammation in kidney’s structure.

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

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

After preprocessing we need to see among support vector machine and random forest classifier which one has 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 classifier for our prediction, we have given our test data set to the model which have two records in it which represents two patients data.

This project features a vast scope within the future for enhancement. It is going to include 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 acquire accurate results.

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 to include 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 acquire accurate results.

9. REFERENCES