IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Chronic Kidney Disease Prediction

Kumar Gaurav[1*], Darshana A. Naik[2], Visesh Kumar Jaiswal[3],Manollas M[4],Ankitha V[5]

[1,2,3,4,5] Dept. of Computer Science and Engineering, Ramaiah Institue of Technology, Bangalore, India

Corresponding Author: kumargauravkumar11@gmail.com, Tel.: +91-70221-2956

DOI: https://doi.org/10.26438/ijcse/v7i4.10651069 | Available online at: www.ijcseonline.org

Accepted: 19/Apr/2019, Published: 30/Apr/2019

Abstract— Chronic kidney Disease (CKD), also known as chronic renal disease, which is continuous malfunction of kidney for months or even years. Identified based on the kidney damage or decrease in glomerular filtration rate (GFR). People with CKD are more prone to cardiovascular death than actual kidney failure. CKD is progressively predominant in patients with CVD or factors such as dyslipidemia, diabetes mellitus, hypertension and metabolic disorder. Classification models are built and are called classifiers. These classifiers will group the entered data set information to prominent classes. Chronic kidney disorder means the damage lasts and it only worsens over the period of time if not taken care of properly. This illness commonly known as kidney failure does not have any symptoms specific to the disease also sometimes the symptoms are not present and is diagnosed only by a lab test. The illness is highly diagnosed in the age of range 19-40 and higher in ages >40, here the waste starts accumulating over time as the Glomerular Filtration Rate(GFR) decreases overtime leading to increase in impurity of the blood. In this paper we are predicting the severity of kidney stage with the help of patients test report and using prediction algorithms, also we are doing a cross validation using C4.5 algorithms.

Keywords: Naïve Bayes, C4.5, Chronic Kidney Disease, Cross-Validation, Pre-processing

I. INTRODUCTION

Chronic Kidney Disease (CKD) is one of the global medical issue because of extremely high expense for the treatment and also the high death rates. World Health Organization (WHO) announced South East Asia and America witnessed highest rate of population with this illness, from a survey in 2012. Besides, the number of new patients increase yearly, while there are restrictions general medical coverage, for example, no cost or low cost remedy, absence of the vital medicinal gear and therapeutic repayment limit. As the cost for dialysis is about rupees 1,200 for every session on a normal and rupees 3,600 every week, users need to spread the cost over therapeutic repayment limit.

With respect to progression of the illness from stages 3 to 5, users should consult for suggestions to keep kidneys healthy and keep them working efficiently for a maximum amount of time from the doctor. As the measure of clients and data per client is expansive and also expanding, specialists and medicinal staffs have trouble managing the customized treatment plan. All things considered, the result is explicit to every client, and possibly regularly updatable. For the emergency clinic including military ones, this might help anticipating the need of work force and assets for dealing with future stage-5 cases. Besides, money related help can be planned and provided ahead of time for those in need. So as to accomplish this, the present study utilizes data mining methods to build a classification model that is fit and can be used for predicting kidney disease stages 3 to 5.

II. MOTIVATION

1. The training data that is present generally has missing attributes and uneven data. Hence it is required to either normalize the data in some cases or in some cases fill the missing values by taking the mean of the other values used in the data set.

2.Sometimes the data set might not contain the target class and hence it is required to find the target class based on Cross Validation.
3.Perform the detection of the level of kidney disease the patient is having based on the naive Bayes classifier after performing resampling gives more accuracy.

4. For the Medical Institutes and Government Organizations it is good have schemes generated for a specific demography of people like age or gender based on kidney chronic disease.

III. METHODOLOGY

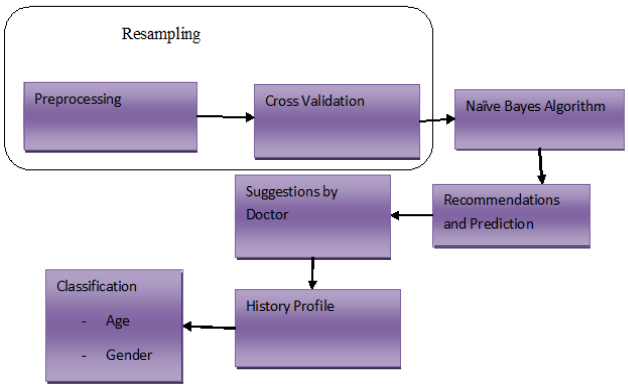

Fig. 1 Methodology

1.Resampling
Obtain the data sets from UCM Library. This library is maintained by Indian medical association. The data set is a historic data of patients who had kidney disease. The data set will have various attributes like HDL High, HDL Low, BUN High, BUN Low etc. Each row represent data for a patient with columns as attributes causing kidney disease and one attribute among those columns represent the class (level of kidney chronic disease). Among those rows few rows will have no class labels and few rows will have missing attributes and few rows will not have been normalized. Sampling process is performed in order to balance these scenarios and also cross validation is performed in order fill the class.

2. Pre-processing
If there are any missing values, then the missing values are filled based on the mean of other values. If there are any missing values and class, then data is discarded. If there are any missing class, then cross validation is performed.

3. Cross Validation
This is a process in which if the class attribute is missing in the training data then we perform computations such as information gain, gain and entropy and then form a decision tree. Once the decision tree is formed then the class is determined from the decision tree and the training data is filled.

4. Recommendations and Prediction
Naive Bayes Classifier with Resampling in order to classify the disease of the user and generate suggestions for each level of disease and appointment if the disease level is the highest.

5.Suggestions by Doctor
This module is responsible for creating the suggestions for the end user. The doctor will select each kidney stage like 3, 4 or 5 and then provide the treatment plan for each of the kidney stages.

6.History Profile
Each time whenever the user takes the test the scanning data attributes are entered and then after predicting the kidney stage then it will maintain the class. Like this when the user takes the test N number of times then a graph is generated with the number of test taken over x axis and stage of the kidney disease over the y axis. From the history profile one can come to know about the improvement in the kidney stage disease.

7.Classification by Age and Gender
During registration the user will select the gender and also provides the age. Like this there will be many users who will be registered in the system. From the set of users, the male and female users are found out and then a chart is generated based on how many males/females have kidney stage 3, 4 and 5. The age group is also divided into a set of multiple age groups and then the graphs are generated for age group 1 and level of kidney stage and age group 2 and level of kidney stage, age group 3 and level of kidney stage and finally age group4 and level of kidney stage

IV. ADVANCEMENT FROM PREVIOUS APPROACH

A. Previous Approach

In the previous approach the data sets from the historic patients are taken and then the top most attributes with highest gain are considered and then plotted as a time series. One attribute is plotted over x axis and then other attribute over y axis. If the chronic disease has to be classified into 4 different stages, then 4 random centres are picked and then clusters are created and the dataset is assigned to a specific cluster.

B. Current Approach

In the proposed approach first pre-processing is performed to clean the unwanted data, following this process the best attributes are chosen based on gain computation, classification of data set into chronic kidney disease into various stages is predicted by using Naïve Bayes Classifier. Finally, the categories are constructed based on GFR computation.

V. OVERVIEW OF IMPLEMENTATION

This overview of the project gives a concise depiction about the possibility of the project and the execution carried out. We principally portray System Architecture.

This is an essential advance in any sort of project. It is the initial step that needs to be done, it gives an earlier thought

or clear picture about if the project will succeed. Arranging a legitimate implementation and working as needs be, will be the key advance for an achievement of project.

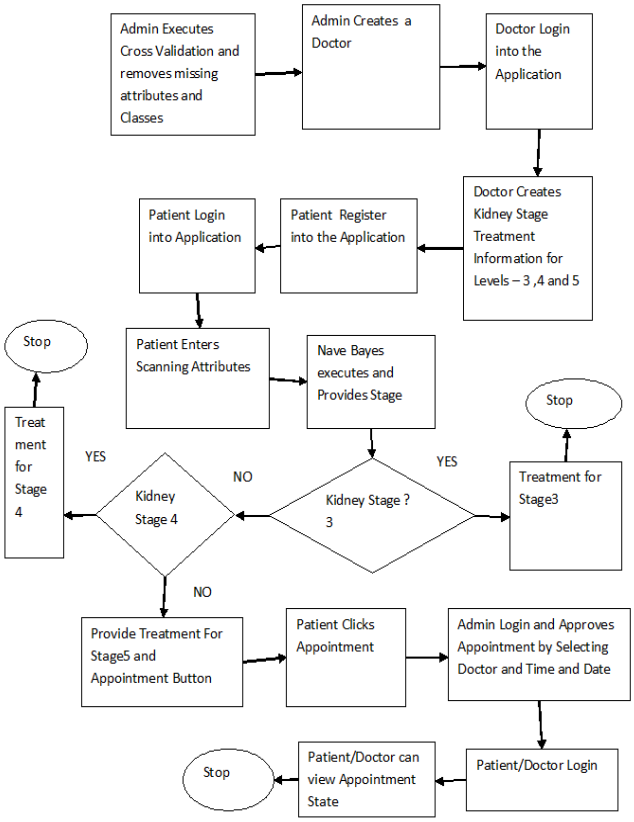

Fig. 2 Structural chart

A. Register User

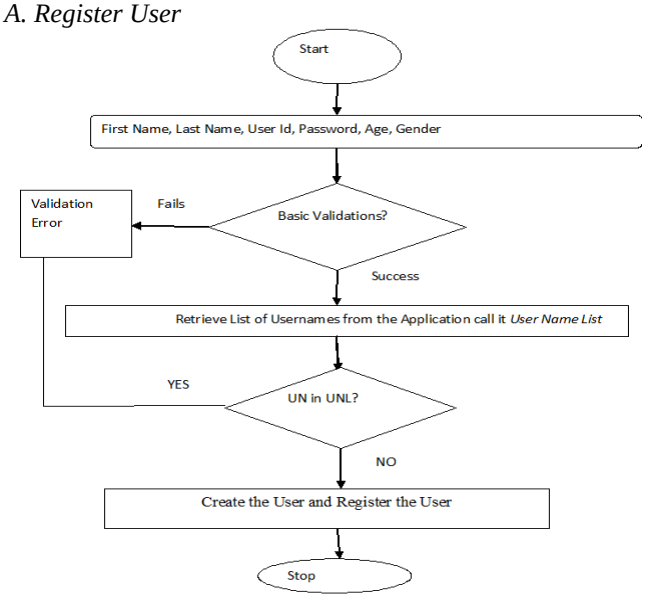

Fig: 3. Register User

B. Login

The Login is responsible for allowing the user to be able to login.

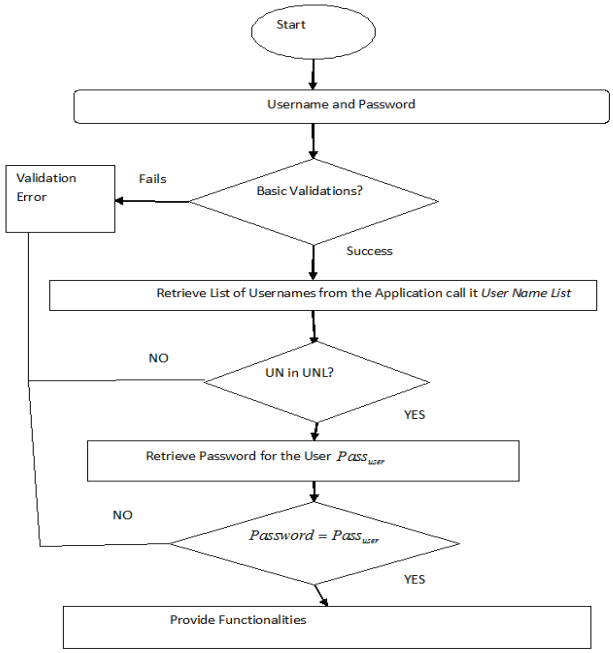

FIG: 4 LOGIN

C. Naïve Bayes Classification Algorithm

Naïve Bayes is used in constructing classifiers: it assigns class names. It is based on principle that an attribute is independent of other attribute, given the class variable. For example, a flower which is red and has thorns is considered a rose, without considering any possible correlations between the attributes.

We use Naïve Bayes Algorithm to identify to which stage (3,4,5) does the user belongs to. Naïve Bayes is best suitable as it considers that the attributes are independent of each other.

Probability computation is performed using the following equation.

$$P_{attribute} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{(\mu-T)^2}{2*\sigma^2}\right)}$$

Where,

$\sigma = s\tan dard\ \ deviation$

$\mu = mean$

$T = current\ value\ of\ \ attribute$

The total probability of No Chronic Disease is computed using:

1) Computed the given data belongs to Chronic Disease is

$$Pchronic = Pchronic + (Pchronic + PNonChronic)$$

2) Computed the given data belongs to No Chronic Disease is

$$PNonChronic = PNonChronic + (PChronic + PNonChronic)$$

If the value of Chronic Disease is higher than No Chronic Disease, then the data belongs to No Chronic Disease.

a) D. Re sampling C4.5

1. In the first phase if any row has values as zero and a specific class is present in the training data set.
2. The other rows from the training data are taken for that specific class and for that specific attribute and then the mean is computed from other attributes of the same class and then the missing values are replaced.
3. If there exist a missing class then the following steps are performed for determining the class of missing attributes.

Table 1. Training Data for Attributes

Att1	Att2	Att3	Class Label
0.5	0.4	0.6	1
0.3	0.4	0.57	1
0.5	0.7	0.78	1
0.3	0.6	0.58	1
0.6	0.8	0.2	1
0.8	0.9	0.1	1
0.6	0.7	0.5	2
0.5	0.8	0.6	2
0.8	0.7	0.7	2
0.3	0.7	0.6	2
0.5	0.6	0.7	2
0.6	0.7	0.8	2
0.6	0.8	0.7	2
0.8	0.6	0.9	2
0.6	0.9	0.8	2
0.8	0.7	0.6	2

We use Information gain to find to which class the data belongs if the class is missing in training dataset

$$Gain = Information\ Gain - Entrophy$$

Once we have gain for all the attributes, we choice the attribute having highest gain as the root and keep on repeating till the termination condition is achieved. Then based on the tree we decide our tuple belongs to which class of kidney stage.

$$Information\ Gain = -\frac{p}{P+n}\log\left(\frac{p}{p+n}\right) - \frac{n}{P+n}\log\left(\frac{n}{p+n}\right)$$

Where,

$p = count\ of\ class\ labels\ of\ 1$

$n = count\ of\ class\ labels\ of\ 0$

example, p=4 n=3

$$IG(overall) = -\frac{4}{4+3}\log(\frac{4}{4+3}) - \frac{3}{4+3}\log(\frac{3}{4+3}) = -0.2467$$

4) After that calculate entropy for each attribute using the below mentioned formula:

$$Entrophy = \sum_{i=1}^{N} stage1 + notstage1/(totalstage1 + totalotherstage)\ Information\ Gain$$

$$Gain = Information\ Gain - Entrophy$$

5) Once we have calculated entropy for all the attributes, and we have overall information gain, we calculate gain of each attributes using above mentioned formula of gain.

6) The attribute having highest gain is selected as root.

7)Keep repeating from step 3 till we meet the termination condition.

VI. EXPERIMENTED DATASET

Table 2

GENDER	AGE	FBSHIGH	FBSNORMAL	HDLHIGH	HDLNORMAL	BUNHIGH	BUNLOW	BUNNORMAL	class
1	21	0.7	0.8	0.9	0.7	0.8	0.9	0.9	3
2	23	0.7	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.7	0.8	0.9	0.7	0.8	0.9	0.9	3
2	23	0.7	0.8	0.9	0.9	0.8	0.9	0.9	3
1	23	0.9	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.8	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.8	0.8	0.9	0.8	0.8	0.9	0.9	3
1	23	0.9	0.9	0.9	0.9	0.9	0.9	0.9	3
1	23	0.8	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.9	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.6	0.8	0.9	0.7	0.8	0.9	0.9	3
1	23	0.4	0.6	0.5	0.4	0.6	0.5	0.6	4
1	23	0.4	0.5	0.5	0.4	0.6	0.5	0.6	4
1	33	0.4	0.6	0.6	0.4	0.6	0.5	0.6	4
2	23	0.4	0.5	0.5	0.4	0.6	0.5	0.6	4
2	23	0.5	0.5	0.5	0.4	0.6	0.5	0.6	4
1	23	0.4	0.6	0.6	0.4	0.6	0.5	0.6	4
1	23	0.5	0.6	0.5	0.4	0.6	0.5	0.6	4
1	43	0.5	0.6	0.5	0.4	0.6	0.5	0.6	4
2	23	0.5	0.4	0.5	0.4	0.6	0.5	0.6	4
2	23	0.5	0.4	0.5	0.4	0.6	0.5	0.6	4
1	53	0.5	0.6	0.5	0.4	0.6	0.5	0.6	4
1	43	0.2	0.4	0.3	0.4	0.4	0.4	0.3	5
1	43	0.2	0.4	0.3	0.4	0.4	0.4	0.5	5
1	33	0.2	0.4	0.3	0.4	0.3	0.4	0.5	5
2	23	0.2	0.4	0.3	0.4	0.3	0.4	0.5	5

VII. INFERENCES FROM RESULTS

1. From the various executions of Naïve Bayes and Naïve Bayes with Cross Validation, one can determine that the time taken one with cross validation is lesser.
2. The Naïve Bayes with Cross Validation is responsible for providing missing attributes using main computation and also is responsible for providing the missing class attributes.
3. The Naïve Bayes with Cross Validation Provides more accuracy as compared to only Naïve Bayes.
4. From the application one can come to know that we will able to register as patients, admin can be able to create different doctor, admin and patients.
5. The Classification of the various users are performed based on the age levels and based on the gender.

VIII. CONCLUSION

The project has 3 kinds of user the Admin, the Patient and the Doctor. The Patient is responsible for registration once the patient has registered then the user will be able to enter the scanning attributes and then get the disease level. Once the disease level is found based on the LEVEL of the disease the treatment plan is provided. If the level is highest, then an appointment button is generated. Once the Admin performs the Login the admin will be able to view the appointments and then approve an appointment. Once it is approved the user will see the notification as well as the doctor will see the client in the appointment list. The admin is be able to see the training data set and is also responsible for filling the missing values or class by executing the cross validation process. The user /client/patient can also track the history based on the number of executions the user is performing with the result of the tests taken by the patient.

IX. LIMITATION

1. The project can determine kidney stage for the 3 levels.
2. The project can handle a maximum of 5 lakh users.

X. FUTURE WORK

1. The algorithm can be extended to have more classification levels.
2. The application can be extended to tie up with government organizations to provide the classification data so that appropriate schemes can be generated for certain kinds of user.

REFERENCES

[1] Kunwar Singh Vaisla and Sithu D Sudarsan, *"Role of attributes selection in classification of Chronic Kidney Disease patients",* International Conference on Computing, Communication and Security (ICCCS), 4-5 Dec, 2015, pp 1-6.

[2] M.M. Rahman, D.N. Davis, "Addressing the class imbalance problem in medical datasets", *International Journal of Machine Learning and Computing*, vol. 3, no. 2, 2013.

[3] S. Vijayarani S. Dhayanand "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS" International Journal of Computing and Business Research (IJCBR) vol. 6 no. 2 2015.

[4] G. Kaur Er. N. Oberai "A REVIEW ARTICLE ON NAIVE BAYES CLASSIFIER WITH VARIOUS SMOOTHING TECHNIQUES" International Journal of Computer Science and Mobile Computing vol. 3 no. 10 pp. 864-868 october 2014.

[5] T. R. Patil S.S. Sherekar "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" International Journal of Computer Science and Applications vol. 6 no. 2 pp. 256-261 April 2013.

[6] "Country Statistics and Global Health Estimates. 2015" in World Health Organization (WHO)

[7] "Chronic Kidney Disease" World Kidney Day 2015 [online] Available: http://www.worldkidneyday.org/fags/chronic-kidney-disease/.

[8] "Symptoms, causes and treatment of Chronic Kidney Disease" news article 2017 [online] Available: https://www.medicalnewstoday.com/articles/172179.php

[9] Sahana B.J, "Prediction of Chronic Kidney Disease using Data Mining Classification Techniques and ANN". International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, NCETEIT - 2017 Conference Proceedings.

Authors Profile

Darshana A. Naik is working as an Assistant Professor in Computer Science Department of Ramaiah Institute of Technology. Her areas of interest include, data mining, big data networks, social network and image processing.

Ankitha V is pursing Bachelor of engineering in Computer Science and Engineering from Ramaiah Institute of Technology, Bangalore,India. Her area of interests include, data mining, web development, technical research.

Kumar Gaurav is pursing Bachelor of engineering in Computer Science and Engineering from Ramaiah Institute of Technology, Bangalore,India. His area of interests include, machine learning, artifical Intelligence, data mining, web development.

Visesh Kumar Jaiswal is pursing Bachelor of engineering in Computer Science and Engineering from Ramaiah Institute of Technology, Bangalore,India. His area of interests include, data mining,web development.

Manollas M is pursing Bachelor of engineering in Computer Science and Engineering from Ramaiah Institute of Technology, Bangalore,India. His area of interests include, machine learning, web development, data mining.

