International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217345

Data Mining Techniques to Predict Chronic Kidney Diseases

Saba Karim, Chaitanya Mankar
Computer Engineering, Dhole Patil College of Engineering, Pune, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 300-304

Publication Issue :
May-June-2021

Article History
Accepted :15May2021
Published :22May2021

ABSTRACT

Chronic Kidney Disease (CKD) is one of the most widespread illnesses nowadays in the world. Some statistics shows that 26 million adults in the United States have CKD and million others are at increasing risk. When condition of the kidney get worse, the wastes in the blood are formed in a high level. Data mining has been a present pattern for an accomplishing analytic outcomes. The Clinical diagnosis of CKD is based on blood and urine tests as well as removing a sample of kidney tissues for testing. By Some previous diagnosis and method of detection the kidney diseases are important to help stop the progression to kidney failure. Data mining and analytics techniques which can be used for predicting CKD by utilizing samples of patient's data and diagnosis records done previously. The aim of my project is to anticipate CKD utilizing the classification strategy Naïve Bayes. Pre-processing the data is performed to impute any missing data and identified the variables that should be considered in the prediction models. Based on the accuracy of prediction the different predictive analytics models are assessed and compared. By presenting a decision support tool which will be used to help in the diagnosis of CKD.

Keywords: *Chronic Kidney Disease, Naïve Bayes, Pre-processing*

I. INTRODUCTION

Chronic Kidney Diseases incorporate the state where the kidneys fails to function and reduces the potential to keep person suffer from the diseases healthy. When condition of the kidney get worse, the wastes in the blood are formed in a high level[2]. Data mining has been a present pattern for an accomplishing analytic outcomes. Colossal measure of un-mined data which is gathered by the human services industries so as to find concealed data for the powerful analysis and basic leadership. Data mining is the way to towards extricating concealed data from gigantic datasets. The goal of my project is to anticipate CKD utilizing the classification strategy Naive Bayes algorithm. The phases of Chronic Kidney Disease (CKD) are anticipated in the light of Glomerular Filtration Rate (GFR)[9]. This project provides a decision support tool that will help in the diagnosis of CKD.

II. METHODS AND MATERIAL

Our Aim is to predict the chronic kidney disease using the machine learning algorithm. Chronic kidney disease (CKD) means your kidneys are

damaged and can't filter blood the way they should. The disease is called "chronic" because the damage to your kidneys happens slowly over a long period of time. This damage can cause wastes to build up in your body. CKD can also cause other health problems.10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year because the doctors are unable diagnose the disease. The system is automation for predicting the CKD.

The system will be a Real-world web-based application that can be used by many hospitals. Naive Bayes is a probabilistic classifier based on Bayes theorem. It assumes variables are independent of each other. The algorithm is easy to build and works well with huge data sets. It has been used because it makes use of small training data to estimate the parameters important for classification. It performs well in multiple class prediction. When assumption of independence holds a Naive Bayes classifier perform better compare to other models like logistic regression and you need less training data. In India an automated diagnosis system would reduce the lengthy process in health care. With an improved symptom analyzing algorithm, the system can suggest diagnostic test to the users hence reducing time and cost in big hospitals.

There are several causes of kidney injury that lead to the final common pathway of ESRD, and the syndrome is characterized by anemia, hypertension, renal bone disease, nutritional impairment, , impaired quality of life and reduced life expectancy. The study and description of the epidemiology of ESRD in the United States population has been greatly enriched by the United States Renal Data System (USRDS), presence of a surveillance system that collects information about the occurrence and that outcomes of care on all incident patients receiving treatment for ESRD in the United States. Chronic kidney disease (CKD) is defined by the presence of sustained abnormalities of renal function and results from different causes of renal injury. CKD can lead to progressive loss of renal function and may terminate in ESRD after a variable period of time following the initiating injury.

This problem can be overcome by classify our dataset using different machine-learning algorithm, which includes training and testing the model. We will try to explore the correlation between the dataset attributes to find out there dependency on each other in the development of chronic kidney diseases.

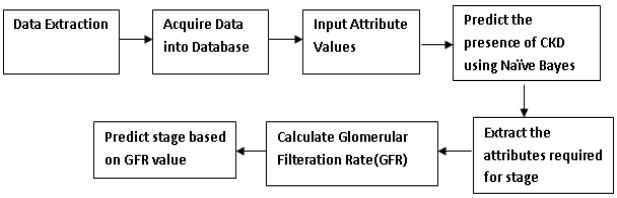

Figure1. Architectural Diagram of System

Naïve Bayes

The Naive Bayes Classifier technique based Bayesian theorem and mostly appropriate when there is high dimensionality of the inputs. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. The Naïve Bayes classifier works on a simple, but comparatively intuitive concept[7]. Also, in some cases it is also seen that Naïve Bayes outperforms many other comparatively complex algorithms. It makes use of the variables contained in the data sample, by observing them individually, independent of each other **Bayes theorem**

1. P (C|X) = P (X|C) ·P(C) / P(X).
2. P(X) is constant for all classes.
3. P(C) = relative freq of class C samples c such that p is increased=c Such that P (X|C) P(C) is increased
4. Problem: computing P (X|C) is unfeasible!

Naïve Bayes Algorithm

Proposed framework makes use of "Naïve Bayes Algorithm" This algorithm predicts whether the patient is suffering from ckd or not ckd.

Step 1: Scan the dataset (storage servers)

Step 2: Calculate the probability of each attribute value. [n, n c, m, p]

Step 3: Apply the formulae

P(attributevalue(ai) / subjectvaluevj)= (nc + mp)/(n+m)

Where*:*

- n = the number of training examples for which v = vj
- nc = number of examples for which v = vj and a = ai
- p = a priori estimate for P(aijvj)
- m = the equivalent sample size

Step 4: Multiply the probabilities by p

Step 5: Compare the values and classify the attribute values to one of the predefined set of cl.

By developing a system for the CKD prediction by classify our dataset using different machine learning algorithms which includes training and testing the model. We will try to explore the correlation between the dataset attributes to find out there dependency on each other in the development of chronic kidney disease. In India an automated diagnosis system would reduce the lengthy process in health care. With an improved symptoms analyzing algorithm, the system can suggest diagnostic test to the users hence reducing time and cost in big hospitals.

Data preprocessing

Data preprocessing prepares raw data for further processing. The traditional data reprocessing method is reacting as it starts with data that is assumed ready for analysis and there is no feedback and impart for the way of data collection. The data inconsistency between data sets is the main difficulty for the data preprocessing .

- Treating missing values
- Rule based outlier detection
- Imputation methods to treating missing value
- Attribute correction using data mining concepts
- Data integration using Knowledge repository and Jaro Winkler
- Data discretization using the Equal width methodology

  Data reduction
- Dimensionality reduction
- Numerosity reduction

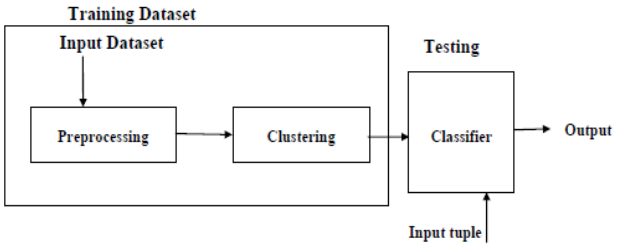

Figure 2: Disease Detection

III.RESULTS AND DISCUSSION

For study, downloaded the dataset from the UCI Machine Learning Repository named Chronic Kidney Disease uploaded in 2015. This dataset has been collected from the Apollo hospital (Tamilnadu) nearly 2 months of period and has 25 attributes, 11 numeric and 14 nominal. The attributes and its description is mentioned in Table 1. Total 400 instances of the dataset is used for the training to prediction algorithms, out of which 250 has label chronic kidney disease (CKD) and 150 has label non chronic kidney disease (NCKD) The source of the dataset we used for the proposed system has been prepared at Apollo Hospital in Tamil Nadu of India. The owner of the dataset graciously made the dataset

available in the machine learning data site Kaggle.com from which we gained access to the dataset. Followings are the information of the creative personal of this dataset.

There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.
the most basic terms, which are

True Positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True Negatives (TN): We predicted no, and they don't have the disease.

False Positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

False Negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

The formula for calculating these parameters are given below:

$$Specificity = \frac{tn}{tn + fp} * 100$$

$$MCC = \frac{(tp * tn) - (fp * fn)}{\sqrt{(tp + fn) * (tn + fp) * (tp + fp) * (tn + fn)}}$$

$$Sensitivity = \frac{tp}{tp + fn} * 100$$

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

tp is the number of true positives,
tn is the number of true negatives,
fp is the number of false positives and
fn is the number of false negatives.

Table 1: Performance measures for the test data

	True CKD	True Not CKD	
Pred CKD	44	5	49
Pred Not CKD	9	42	42
	44	4	

Table 2 : Performance measures for the test data

Method	Accuracy	Sensitivity	specificity	Precision
Naïve Bayes	94.50	100	89	89

Table 3: Performance measures for the training data

Method	Accuracy	Sensitivity	specificity	Precision
Naïve Bayes	94.49	100	89	89

IV.CONCLUSION

The chronic kidney disease can be very well predicted using many classifiers in Data Mining. One can also predict the level of chronic kidney disease using classifiers. we speculate that applyingthis methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. In the future, a large number of more complex and representative
data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.

V. ACKNOWLEDGMENT

We would like to thank the UCI machine learning repository for sharing the CKD data set.

VI. REFERENCES

[1]. Jiongming qin 1, lin chen 2, yuhua liu 1, chuanjun liu 2, changhao feng 1, and bin chen 1. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. date of current version February 4, 2020.

[2]. William M. McClellan WM. Epidemiology and risk factors for chronic kidney disease. The Medical clinics of North America. 2005;89(3):419
445.doi10.1016/j.mcna.2004.11.006

[3]. Cristóbal Romero, Data mining in course management systems : Hippisley-Cox, J., and Coupland, C., 2010, "Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the Kidney® Scores," Hippisley-Cox and Coupland BMC Family Practice, 11 -49.

[4]. Navdeep Tangri, Lesley A. Stevens, John Griffith, PhD, Hocine Tighiouart, MS Ognjenka Djurdjev, David Naimark, Adeera Levin, Andrew S. Levey, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure" JAMA the Journal of the American Medical Association · April 2011.

[5]. Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orrù, Carlo Carcassi and Giorgio La Nasa, Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome, Chapter 5, an open access article distributed under the terms of the Creative Commons Attribution License.

[6]. Ziyad, A., 2013, "Prediction of Renal End Points in Chronic Kidney Disease," Kidney International, 83(2), 189-191].

[7]. Kobayashi, T., Yoshida, T. 2014, "A Metabolomics-Based Approach for Predicting Stages of Chronic Kidney Disease," Biochemical and Biophysical Research Communications, 445, 412–416.

[8]. Kobayashi, T., Yoshida, T. 2014, "A Metabolomics-Based Approach for Predicting Stages of Chronic Kidney Disease," Biochemical and Biophysical Research Communications, 445, 412–416.

[9]. Lakshmi, K.R., Nagesh, Y., and Veera Krishna, M., 2014, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering and Technology, 7(1), 242-254.

[10]. International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307 DOI : https://doi.org/10.32628/CSEIT1952331

VII. Citation

Mr. Saba Karim

M.E Scholar, Department of Computer Engineering, DPES Dhole Patil College of Engineering, Wagholi, Pune, India, Approved by the AICTE New Delhi, & The Government of Maharashtra, Affiliated to Savitribai Phule Pune University.

Has completed Bachelor of Engineering in 2019, Computer Engineering Department, Jamia Institute of Engineering & Management Studies, Akkalkuwa, Affiliated to Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon.

Achievements: Paper Published- 04, Seminar-03

E-mail: sabakarim026@gmail.com

Prof. Chaitanya Mankar

Assistant Professor, Department of Computer Engineering, DPES Dhole Patil College of Engineering, Wagholi, Pune, India, Approved by the AICTE New Delhi, & The Government of Maharashtra, Affiliated to Savitribai Phule Pune University. Has completed Master of Engineering in Computer Networks.

Experience: 07 years of Teaching.

Area of Interest: Wireless Sensor Networks

Achievements: Paper- 06, workshop-05, Seminars-03

Email Id: Chaitanya.mankar@gmail.com

Cite this article as :

Saba Karim, Chaitanya Mankar, "Data Mining Techniques to Predict Chronic Kidney Diseases", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 300-304, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT217345

Journal URL : https://ijsrcseit.com/CSEIT217345