IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906238|

Prediction of Chronic Kidney Disease Using Data Science

S.Jayanth Reddy, S.Mohammad Ismail, R.Sai Ram Bhaskar Chowdary, P.Haradeep Kumar,

Dr.MD.Shakeel Ahamad

Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Namburu, Guntur, Andhra Pradesh, India

ABSTRACT: Chronic Kidney Disease (CKD) is a perennial condition where the kidneys deteriorate and stop functioning gradually. This disease has become one of the major public health concerns worldwide. It is insidious, often recognizable only by laboratory abnormalities until its latest stages. The main motive of this work is to ascertain the existence of chronic kidney disease by imposing various classification algorithms on the patient medical record. This research work is primarily concentrated on finding the best suitable classification algorithm which can be used for the diagnosis of CKD based on the classification report and performance factors. Empirical work is performed on different algorithms like J48 DecisionTree, Random Forest, Simple Logistic classifier, Naive Bayes Classifier. The experimental results show that Random Forest gave better results when compared to other classification algorithms and generates 98.75% accuracy.

KEYWORDS : chronic kidney disease; healthcare; data science; classification accuracy.

Project Goal - The main goal of this research work is to predict Chronic Kidney Disease in more accurate and faster way with reduced attributes by imposing best ***classification algorithm*** on the patient medical record.

I. INTRODUCTION

Data science is concerned with procedures and systems, that are used to extract knowledge or insights from large amounts of structured or unstructured data. Today, data science has far-reaching inferences in many fields, both academic and applied research domains like image recognition, machine translation, speech recognition, digital economy on one hand and fields like healthcare, social science, medical informatics etc.

Classification is a known method of data mining in healthcare domain. It is used to predict the target class for each data point. The classification methods include Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes etc.

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.CKD generally develops slowly with few symptoms, and many people cannot discern that they have it until the disease is usually in its last stages. 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. This disease kills a greater number of people each year compared to people with breast or prostate cancer. The prediction of the existence of this disease plays an important role to take necessary preventive measures. A comparative analysis is done on different algorithms like support Random Forest

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906238|

(RF), Naive Bayes (NB) classifier,J48 Decision Tree, Simple logistic classifier to check the classification accuracy for diagnosing CKD.

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.

Boukenze, B. et al.[1] In order to predict Chronic Kidney disease, they interpreted discrimination between ANNs, SVM, KNN, NB algorithms. The comparison has been done based on the accuracy of prediction using WEKA tool. In their analysis, the best algorithm was ANN and SVM with 62.5% accuracy.

Shaikhina, T. et al. [2] analyzed the DT and RF ML algorithms. These ML algorithms are used for prediction of ckd. For the classification process, they used a data mining tool named MATLAB. From the experimental result, they absorbed that both Decision tree and Random Forest algorithms performed well with the same accuracy.

Polat, H. et al. [3] developed a SVM mechanism to predict chronic kidney disease in the human body. They have made use of the WEKA tool and obtained an accuracy of 98.5% through SVM.

Panwong P. et al.[4] performed comparative analysis on KNN, ANN, RF, J48, NB algorithms using WEKA tool. They have obtained the highest accuracy with Random Forest Classifier with an accuracy of 86.6%.

Padmanaban, K. A. et al. [5] have implemented two classifiers: DT and NB using Rapid minor data mining tool. Finally, they discovered that the Decision Tree has the highest accuracy when compared to Naive Bayes classifier.

Dulhare, U. et al. [6] analyzed the prediction of CKD using NB classifier using WEKA tool and obtained an accuracy of 97.5%.
Ramya, S& Radha, N[7] They have applied classification techniques on test data using R Tool. They compared the four techniques: RF, Backpropagation, radial basis function (RBF) and ANN. Their results determine RBF has higher accuracy for prediction of the CKD.

Abeer et. al. [8] made use of Classifiers like SVM and LR are compared based on their performance. They observed that the performance of the SVM has shown better outcome when compared to the other algorithm with an accuracy of 93.1%.

Chetty, N. et al. [9] have implemented three classifiers NB, KNN and SVM. They have used WEKA tool and found that SVM has the highest accuracy among the three.

Dhruvi, R. et al. [10] they interpreted the prediction of CKD using J48 and Naive Bayes algorithms using Java. Their analysis described that the best algorithm was J48 with 62.5% accuracy.

III. METHODOLOGY

The following are the steps followed for the prediction of CKD.

Dataset and Attributes:
The database for prediction of CKD is obtained from patient medical reports which are obtained from different laboratories and is made available in kaggle website. There are 400 records with 25 various attributes related to kidney disease like age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, diabetes mellitus, anemia. Out of these 25 attributes we only use 7 attributes to build our predictive model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906238|

Data preprocessing:

i) Data Cleaning: The CKD dataset class consists of 'ckd' and 'nockd' as target labels. In order to replace strings with numbers, we assigned numbers like '1' and '0' respectively so that the algorithms can work on the data. Similarly, this process is done for the remaining attributes possessing strings. In the CKD dataset, there are many unknown values discovered. Initially, the strings are replaced with integers. Then the missing values in all the columns are filled with medians of that column respectively.

ii) Data Reduction and Feature selection: Out of 25 attributes present in the dataset, we have selected 7 important attributes required to build predictive model. The selected attributes are Blood Pressure, Specific Gravity, Albumin, Blood Sugar Level, Red Blood Cells Count, Pus Cell Count,Pus Cell Clumps.

Training and Testing Dataset:

The dataset is divided into two sub datasets both containing 7 attributes:

i) Training data: training dataset is derived from main dataset and it contains 320 out of 400 records in main dataset of CKD.

ii) Testing data: testing dataset is of 80 out of 400 records from main CKD dataset.

Classifiers and Model selection:

One among these classification models i.e J48 Decision tree, Random Forest , Simple Logistic Regression , Naive Bayes classifier is selected for final class label prediction with high accuracy.

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.

Random Forest: This is a supervised classification algorithm. It operates by the construction of many decision trees at training time and determining the class of the individual trees. They are an amalgamation of tree predictors so that all trees depend on the values of a random vector sampled independently and with a similar distribution for all trees in the forest. The higher the number of trees in the forest gives high accuracy results.

Naive Bayes Classifier: It is a model which is based on Bayes theorem. NB classifier forebodes membership chances, like the possibility that a given tuple belongs to a class. The primary purpose of using NB is that it only needs a few numbers of training data to estimate the parameters required for classification.

Simple Logistic Classifier: It is a technique in the field of statistics used in machine learning. This technique uses an equation as the representation like linear regression. It is used to delineate data and the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables.

IV.RESULTS

Comparison of Classifiers:

A *comparative analysis* is done on different algorithms like Random Forest , J48 decision tree, Logistic Regression and Naive Bayes classifier to check the classification accuracy for diagnosing CKD using WEKA tool. The results are shown below:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906238|

Classification Algorithm	Average Accuracy(%)
Random Forest	98.75%
Simple Logistic Classifier	98.25%
Naïve Bayes Classifier	97.5%
J48 Decision Tree	96.5%

Table.1 Comparison of Classifiers based on accuracy.

Model Selection:

The experimental results shown that *Random Forest* give better results when compared to other classification algorithms and generates 98.75% accuracy. Hence Random Forest Classification algorithm can be effectively used for predicting the class label ckd with high accuracy and reduced attributes.

V. 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 and data mining algorithms.

VI.CONCLUSION

The objective of this work is to observe classification algorithms to analyses and predict CKD.A ***comparative analysis*** is done on different algorithms like Random Forest , J48 decision tree, Logistic Regression , Naive Bayes classifier to check the classification accuracy for diagnosing CKD. The experimental results shown that ***Random Forest*** give better results when compared to other classification algorithms and generates 98.75% accuracy and hence Random Forest is used effectively for predicting the existence of Chronic kidney disease.

REFERENCES

[1] Boukenze, Basma, Abdelkrim Haqiq, and Hajar Mousannif. "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques." Advances in Ubiquitous Networking 2. Springer, Singapore, 2017. 701-712.
[2] T Shaikhina, Torgyn, et al. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." Biomedical Signal Processing and Control (2017).
[3] Polat, Huseyin, Homay Danaei Mehr, and Aydin Cetin. "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods." Journal of medical systems 41.4 (2017): 55.
[4] Panwong, Patcharaporn, and Natthakan Iam-On. "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method." 2016 second Asian conference on defence technology (ACDT). IEEE, 2016.
[5] Padmanaban, KR Anantha, and G. Parthiban. "Applying machine learning techniques for predicting the risk of chronic kidney disease." Indian Journal of Science and Technology 9.29 (2016): 1-6.
[6] Dulhare, Uma N., and Mohammad Ayesha. "Extraction of action rules for chronic kidney disease using Naïve bayes classifier." 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2016.
[7] Ramya, Sm, and N. Radha. "Diagnosis of chronic kidney disease using machine learning algorithms." International Journal of Innovative Research in Computer and Communication Engineering 4.1 (2016): 812-820.
[8] Abeer, Ahmad, "Diagnosis and Classification of Chronic Renal failure Utilizing Intelligent Data Mining Classification".
[9] Chetty, Naganna, Kunwar Singh Vaisla, and Sithu D. Sudarsan. "Role of attributes selection in classification of Chronic Kidney Disease patients." 2015 International Conference on Computing, Communication and Security (ICCCS). IEEE, 2015.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906238|

[10] Dhruvi. R, Yavnika P, Nutan R, " Prediction of Probability of Chronic Diseases and Providing Relative Real-Time Statistical Report using data mining and machine learning techniques". International Journal of Science, Engineering and Technology Research (IJSETR) vol. 5, no. 4. 2016.
[11] Sirage Zeynu & Shruthi Patil, "Survey on Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques and Feature Selection", International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 149-156.
[12] "Prediction Model and Risk Stratification Tool for Survival in Patients With CKD", Alexander S. Goldfarb-Rumyantzev, Shiva Gautam, Ning Dong and Robert S. Brown.
[13] "Prediction of Chronic Kidney Disease Stage 3 by CKD273, a Urinary Proteomic Biomarker", Claudia Pontillo, Zhen-Yu Zhang .
[14] Divya Jain & Vijendra Singh, "Feature selection and classification systems for chronic disease prediction: A review"
[15] S. Vijayarani, S. Dhayanand, and M. Phil. "Kidney disease prediction using SVM and ANN algorithms." International Journal of Computing and Business Research (IJCBR) vol. 6, no. 2, 2015.
[16] Ruey Key, "Constructing Models for Chronic Kidney Disease Detection and Risk Estimation", IEEE International Symposium on Intelligent C16ontrol.
[17] Dataset: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
[18] Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "Authorship Attribution of Telugu Texts based on Syntactic Features and Machine Learning Tchniques." Journal of Theoretical & Applied Information Technology 85.1 (2016).

.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details