International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Detecting Kidney Disease using Naïve Bayes and Decision Tree in Machine Learning

Sakshi Kapoor, Rabina Verma, Surya Narayan Panda

Abstract: ***Chronic Kidney Disease (CKD) mostly influence patients suffered from difficulties due to diabetes or high blood pressure and make them unable to carry out their daily activities. In a survey , it has been revealed that one in 12 persons living in two biggest cities of India diagnosed of CKD features that put them at high risk for unfavourable outcomes. In this article, we have analyzed as well as anticipated chronic kidney disease by discovering the hidden pattern of the relationship using feature selection and Machine Learning classification approach like naive Bayes classifier and decision tree(J48). The dataset on which these approaches are applied is taken from UC Irvine repository. Based on certain feature, the approaches will predict whether a person is diagnosed with a CKD or Not CKD. While performing comparative analysis, it has been observed that J48 decision tree gives high accuracy rate in prediction. J48 classifier proves to be efficient and more effective in detecting kidney diseases.***

Keyword: ***Chronic Kidney disease (CKD), Classification Techniques- J48, machine learning, naïve Bayes.***

I. INTRODUCTION

The kidney is a pair of organs in the abdominal cavity shaped like a bean that is used to pass waste from the body in the form of urine. It helps in filtering the blood as well as sending it back to the heart. The important functions performed by kidney is to filter the waste materials from the food and maintains the fluid balance, filter minerals from blood, promote bone health, regulate blood pressure and initiating hormones which yields RBCs. Kidney disease is split into two parts, acute and chronic kidney disease [1]. Acute kidney disease causes sudden damage to the kidneys. But in few situations it is advanced to lifelong perennial kidney disease. The foremost causes includes, destruction of actual kidney tissue caused by medication, serious infection, barrier to urine leaving the kidney (due to kidney stones) which leads to perennial kidney disease. Further it can also lead to last stage, which needs dialysis or a surgical operation. The different causes of perennial kidney disease contains, blood vessels damage caused by outrageous blood pressure and diabetes, infectious attacks by different diseases on the kidney organ, the thickening of cysts, harmful effect on the kidneys because of reverse movement of urine. Figure 1 represents the difference between normal kidney and the diseased one. However, this paper attempted to focus on prophecy of perennial kidney disease which helps the nephrologists to diagnose the level of illness persisting in the patient using various machine learning classification approaches like Naïve Bayes classifier, decision tree (J48) [2].

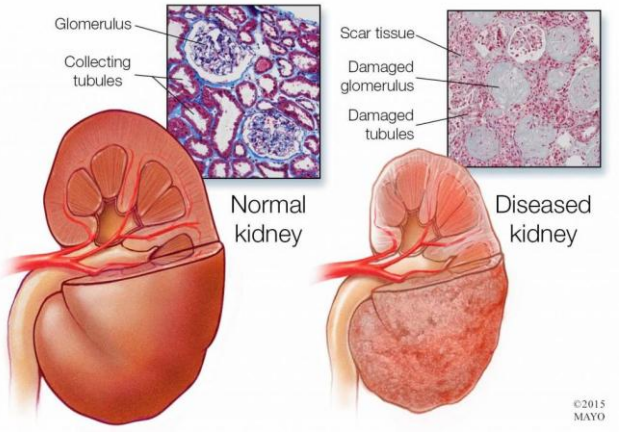

Fig.1. Structure of Kidney – Normal and diseased.

II. RELATED WORK

In literature, several researches have been carried out various techniques for the classification of kidney chronic diseases. Most of them have focused on Support Vector Machine(SVM) which is a powerful algorithm for rationalization of the model to fresh data. SVM has the ability to solve difficult, physical world problems such as categorizing the text and image, recognizing hand-writing, and analysis of biological data using computer science and mathematics approach. Another technique used by various researchers is Naïve Bayes. It converges quicker than other discriminative models like logistic regression with least requirement of training data (both continuous and discrete). It can handle binary as well as multi-class classification problems. Abeer et al. [3] have executed two approaches of information retrieval SVM and logistic regression. The observation conveys that SVM has more accurate results than LR with 93.14%. Sedighi, Z et al. [4] worked on kidney disease analysis by using two feature selection methods, which are wrapper and filter method followed by Ada Boost and Naïve Bayes classifier algorithms. K-Nearest Neighbours used the model to replace missing data. Genetic algorithm is used to select attribute used in an ensemble classification.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Sakshi Kapoor*, Department of Master of Technology in Computer Science and Engineering, Institute of Engineering and Technology Chitkara University, Punjab, India.
Rabina Verma, Department of Master of Technology in Computer Science and Engineering, Sri Sai College of Engineering and Technology (SSCET), Badhani, Pathankot, Punjab, India.
Surya Narayan Panda, Chitkara University Institute Of Engineering and Technology Chitkara University, Punjab, India

The classification result was good compared to the original classification performance accuracy. Swathi et al. [5] worked on knowledge retrieval procedures like J48, Active Directory Trees, K-Star, Radial Basis Function (RBF) for forecasting kidney disease. The experiment shows that Naïve Bayes has the greater precision. Rubini et al. [6] have surveyed a new chronic kidney disease dataset and also applied three classifiers, multilayer perceptron, radial basis function and logistics regression. It was identified that multilayer perceptron has the greater efficiency. Vijayarani et al. [7] have compared naïve bayes and Support Vector Machine classification techniques by considering six different attributes for prediction of kidney disease. The results represent that SVM is better than Naïve Bayes. Jena et al. [8] have surveyed perennial kidney disease information set with miscellaneous classification approaches like Naïve Bayes, Multilayer Perceptron, SVM, etc using weka software. There are 25 different attributes for analyzing the results of classification. Among all, Multilayer Perceptron gives higher accuracy of 99.75%. Ramya et al. [9] have applied four classification techniques like BP, Neural network, RBF and Random Forest on test data from patient medical report to predict the kidney function failure. For prediction 1000 records with 15 attributes has been used. Among all methods RBF (Radial Basis Function) give good results for estimating the disease. Manish Kumar [10], worked on various Machine Learning approaches like Naïve Bayes, Radial Basis Function, Random Forest Classifiers, Sequential Minimal Optimization, Multilayer perceptron classifier and Simple Logistic to detect chronic kidney disease. There are total 400 records for training the dataset. Among these methods Random Forest has highest accuracy. Pangong et al. [11] used as K Nearest Neighbours, Artificial Neural Network, Decision Tree and Naïve Bays to predict the intermediate pauses of kidney disease especially 3 to 5 stages. These classification models were constructed by classifying the selected or reduced set of attributes. After applying balanced classifier, attributes are reduced by the feature selection method, the performance and accuracy was secured around 85%. Padmanaban et al. [12] applied decision tree and naïve Bayes data mining approaches for prognosis of CKD for prevention of the risk factor of CKD. Chronic kidney Disease dataset was implemented on rapid miner by retrieving the dataset to set target role using 10-fold cross-validation. After classification, the performance prediction accuracy of naive Bayes was originated 86 % and the performance prediction accuracy of decision tree originates 91%. Boukenze et al. [13], developed prediction system for chronic failure disease by using different Machine Learning procedures like Support Vector Machine, K Nearest Neighbour, Multi-Layer Perceptron/Artificial Neural Network, C4.5, Bayesian network. After implementing in weka platform, compare the performance accuracy of the algorithm on the basis of different measures like execution time, sensitivity specificity and F- measure. A C4.5 decision tree is scored good accuracy i.e. 63% among other algorithms.

III. MATERIALS AND METHODS

A. Materials

The dataset is taken from the UCI machine learning repository. For eliminating the missing values, pre-processing is considered either by replacement or separation from the dataset. In order to lessen the dimensionality of the feature as well as selecting an optimal feature subset, feature selections methods are prefered .To achieve high-performance accuracy the process repeats for the same. A model is constructed using different classification algorithms like Naïve Bayes classifier, J48.The dataset contains 400 illustrations, where each illustration has 25 features: Blood Pressure, Potassium, Anemia, Albumin, Sugar, Pus Cell, Appetite, Pus Cell clumps, Bacteria, Blood Glucose Random, Hypertension, Specific Gravity, Sodium, Hemoglobin, Age, Packed Cell Volume, Red Blood Cells, White Blood Cell Count, Red Blood Cell Count, Blood Urea, Diabetes Mellitus, Serum Creatinine, Coronary Artery Disease, Pedal Edema and Class [14].The class target variable contains values “CKD” or “NOT CKD”. While “CKD” Chronic Kidney diseases specify positive test and “NOT CKD” specify negative test. At hand 250 cases in class “CKD” and 150 cases in class “NOT CKD”.

B. Methods

i. Naive Bayes classifier

Naive Bayes is a classification method based upon Bayes Theorem which computes the likelihood for every attribute. It selects the outcome with highest probability. This classifier assumes the features are independent and that the existence of a specific feature in a class is not linked to the existence of any other feature. All the properties independently make a contribution to the probability, even if the features are dependent on other features. Naive Bayes technique is mostly applicable for big datasets. It is elementary known to give exceptionally good results. Bayes theorem works on conditional probability. Conditional probability is the possibility of an occurrence to happen, given that some other event has already occurred. The equation to calculate conditional probability is given as

$$P(Hyp|Evi)=P(Evi|Hyp)*P(Hyp)/P(Evi)$$

Where, *P (Hyp)* is the possibility of hypothesis Hyp being true. *P (Evi)* is the possibility of the evidence (unrelated to the hypothesis). *P (Evi|Hyp)* is the possibility of the evidence when the hypothesis is true. *P (Hyp|Evi)* is the possibility of the hypothesis when the evidence is there [15].

ii. J48 decision tree

It is a predictive method to analyze the target value from a dataset on various given attributes. From the training data, it finds the attribute which segregate several instances. In order to achieve highest information gain, these instances are further classified. This procedure is applied over the smaller subsets in a repetitive manner until all the instances rightly placed in their class. In the given figure 1, the first level is a single header node which is a pointing node to its children. Attributes are denoted by internal nodes whereas the branches gives possible values these attributes can have.

The terminal node depicts the final value of the target variable.

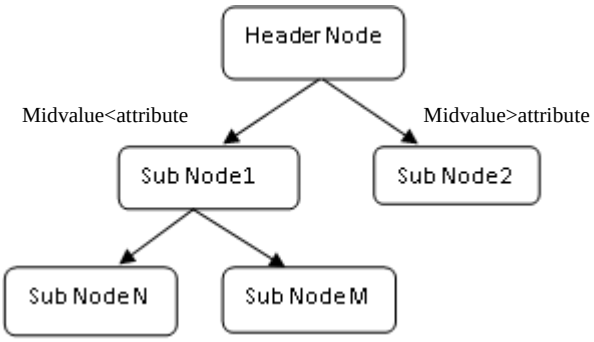

Fig. 2. Structure of the J48 decision tree [16]

IV. PROPOSED METHOD

Chronic Kidney disease dataset: It is downloaded from UCI Repository.

Classification Algorithm Applied: To predict the accuracy of CKD, two classification algorithms are applied; Naïve Bayes classifier and Decision Tree (J48) on Weka platform.

Results: The result shows how many of them are suffered from chronic Kidney disease or not. It also represents the incorrectly classified instances which are used to predict the accuracy of results in both algorithms.

Flowchart of Proposed Methodology: Figure 2 shows the flow chart of proposed work done.

- The dataset taken for prediction of CKD act as a repository which contains various instances.
- In the next step, processing on these instances is done by applying the algorithms.
- Based on certain conditions, it is predicted that particular set of instance comes under severe chronic disease or not. This prediction is further can be used by nephrologists to cure the disease [16].

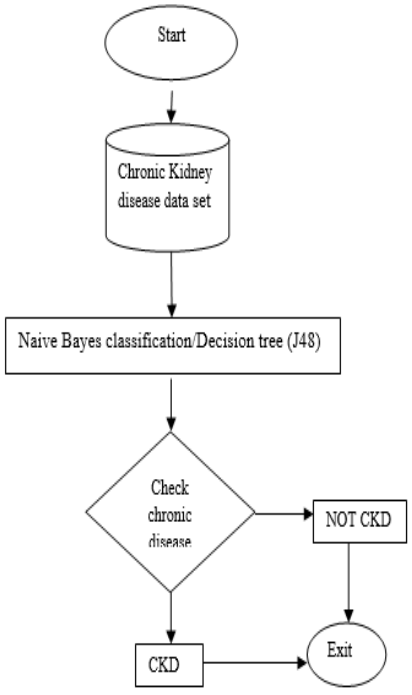

Fig. 3 Flowchart of proposed work

V. EXPERIMENTAL RESULTS

A. Result Analysis of Naive Bayes Classifier and J48 decision tree

The implementation of both classification methods is performed on WEKA platform which gives the results as shown in Table 1 and Table 2. Here, TP rate termed as True Positives that represents the instances accurately categorized in a given dataset, whereas FP rate defines False Positives i.e. the instances inaccurately arranged in a given class. Precision defines the proportion of relevant instances among the retrieved instances which is also called positive predictive value, while recall depicts the proportion of appropriate instances that have been resolved over the total number of appropriate instances which is also termed as sensitivity. Table 1 and table 2 represents the results of naive bayes and J48 decision tree respectively.

Table 1: Naïve Bayes Classifier results

True Positives Rate (%)	False Positives Rate (%)	Precision (%)	Recall (%)	Class
92	0	100	92	CKD
100	8	88.2	100	NOT CKD
96	4	94.1	96	Weighted Average

Table 2: J48 Classifier results

True Posit (%)	False Positives Rate (%)	Precision (%)	Recall (%)	Class
99.6	2	98.8	99.6	CKD
98	0.4	99.3	98	NOT CKD
98.8	1.2	99.05	98.8	Weighted Average

In naïve bayes, it has been analyzed that correctly classified instances are 380 i.e. 95% and incorrectly classified instances are 20 i.e.5%. While in J48 the correctly classified instances are 396 i.e. 99% and incorrectly classified instances are 4 i.e.1%.

B. Comparative Analysis

Precision and recall both are based on perception and measure of appropriateness. As shown in Table 3, J48 decision classifier gives better precision and recall results than naïve bayes.

Table 3: Comparative analysis

Accuracy parameters	J48 decision tree	Naïve Bayes
Precision	99.05%	94.1%
Recall	98.8%	96%

VI. CONCLUSION

The prediction of CKD can be performed using many classifiers in machine learning. Assessment results, shows that various techniques for classification have been used to recognize the analysis and prediction for chronic kidney disease. The execution accuracy of the procedures have been upgraded by using different classification algorithms like KNN, ANN, Naïve Bayes, SVM, Decision tree (J48, C4.5)and feature selection. It has been observed that J48 decision tree gives high accuracy rate in predicting kidney disease. The decision tree classifier proves to be very efficient and effective in kidney disease detection. For the future, the Implementation of prediction system can be enhanced by ensemble different classifier algorithms.

REFERENCES

1. https://www.healthline.com/human-body-maps/kidney
2. https://kidney.org.au/your-kidneys/support/kidney-disease/types
3. Abeer, Ahmad, 'Diagnosis and Classification of Chronic Renal failure Utilizing Intelligent Data Mining Classification', International Journal of Information Technology and Web Engineering, Volume 9 Issue 4, Pages 1-12, October 2014
4. Z. Sedighi, H. Ebrahimpour-Komleh, and S. J. Mousavirad, 'Feature selection effects on kidney disease analysis', International Congress on Technology, Communication and Knowledge (ICTCK), pp. 455–459, 2015.
5. Manish Kumar, 'Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm', International Journal of Compute Science and Mobile Computing, Vol 5, Issue 2, pp. 24-33, Feb-2016.
6. S. Ramya ,Dr. N. Radha, 'Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms', International Journal of Innovative Research in Computer and Communication Engineering, Vol 4, Issue 1, January 2016.
7. D. S.Vijayarani, Mr. S. Dhayanand, 'Data Mining Classification Algorithms for Kidney Disease Prediction', International journal of Cybernetics and informatics (IJCI), Vol 4, August 2015.
8. Lambodar Jeena, Narendra Ku. Kamila, 'Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease', International Journal of Engineering Research in management and Technology, Vol 4, Issue 11, Nov 2015
9. Jurlin Rubini, "Generating Comparative Analysis of Early Stage Prediction of Chronic Kidney Disease", International Journal of Modern Engineering Research, Vol 5, Issue 7, July 2015
10. P. Swathi baby, T. Panduranga Vital, 'Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms', International Journal of Engineering Research and Technology (IJERT), Vol 4, Issue 07, pp. 206-210, July -2015.
11. P. Panwong and N. Iam-On, 'Predicting transitional interval of kidney disease stages 3 to 5 using data mining method', Second Asian Conference on Defence Technology (ACDT), Vol 12, Issue 2, pp. 145–150, 2016
12. K. R. A. Padmanaban and G. Parthiban, 'Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease', Indian Journal of Science and Technology, vol. 9, pp. 29, Aug. 2016.
13. Boukenze, A. Haqiq, and H. Mousannif, 'Predicting Chronic Kidney Failure Disease Using Data Mining Techniques', in Advances in Ubiquitous Networking 2, Springer, Vol 8, pp. 701–712, Singapore, 2017.
14. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease#
15. http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning
16. http://shodhganga.inflibnet.ac.in/bitstream/10603/21206/14/14_chapter%205.

AUTHORS PROFILE

Sakshi Kapoor is currently doing her Master of Technology in Computer Science and Engineering from **Chitkara University, Punjab, India**. Her research interests are Cloud Computing, Machine learning, Scheduling, Internet of Things (IOT), CloudIot. She currently lives in Punjab, India. She has done her Bachelor of Technology from Sri Sukhmani Institute of Engineering and Technology (SSIET), Derabassi. Her Languages are English, Hindi, Punjabi. She has done her publications in many conferences. She has good knowledge of CloudSim, Matlab.

Rabina Verma currently lives in Patiala, India. She is currently working in **Chitkara University, Punjab, India.** Her Languages are English, Punjabi, Hindi. She has done her Masters of Technology in Computer Science and Engineering in the year 2014 from Sri Sai College of Engineering and Technology (SSCET), Badhani, Pathankot. Her research interests are Machine Learning, Data Mining, Cloud Computing Artificial Intelligence, Scheduling, Neural Networks, Deep Learning. She has published in various conferences and journals. She has great knowledge of Weka Tool and Matlab.

Dr. Surya Narayan Panda is a Professor in **Chitkara University, Punjab, India.** He is Director Research in Chitkara University Research and Innovation Network **(CURIN), Chitkara University**. His research areas are Internet oF Things (IOT), Network Security, Cloud Computing, CloudIoT. He currently lives in Ambala city. He has done many publications in international journals as well as conferences. He has great knowledge about latest tools in Cloud Computing and Internet of Things (IoT).