An Intelligent Decision Support System for Early Detection of Chronic Kidney Disease using Machine Learning Models

Rathna R
School of Computer Science and Engineering, Vellore Institute of Technology Chennai Campus, Chennai. 600127. Tamil Nadu. India.
rathna.r@vit.ac.in

Abstract - **All Chronic Kidney Disease (CKD) has become one of the rapidly increasing and frightening non – communicable disease in the world. 10% of the population of India and more than 10% of the population of the world are affected by CKD. The population mentioned here covers people in between 20 to 80 years of age. There are even children and adolescents in quite a big number affected by CKD but not included in this percentage. So many factors are responsible for this. If this disease is detected in advance, there are many treatments available to give the good quality of life back to the affected people. The work presented here is about predicting the CKD in advance based on the clinical parameters of the patients. The Machine Learning algorithms namely Logistic Regression, Randomizable Filtered Classifier and Reduced Error Pruning (REP) Tree are implemented over a dataset with 24 parameters of 400 patients and the prediction percentages are compared. Further the results of the best classification technique is improved by applying Ensemble Bagging Algorithm. It has given a result of 99.25% accuracy in prediction. The procedure is carried out for the dataset with feature selection procedure done in initial stage.**

Index Terms – Bagging, Chronic Kidney Disease, Classification, Features, Machine Learning and Prediction.

INTRODUCTION

We all know that the human body has two kidneys. The hard truth is that we have only two. We are living in a fast, mechanical world running behind the thirst for accelerating to the next higher level so that we can have a more comfortable and more luxurious life style. But that could not be a healthy life style. Due to the changes in the way we try to live is against what we followed for decades and also due to hereditary, people are getting affected by more number of diseases nowadays.

Diabetes Mellitus is found in majority of everyone above the age of 60 and also in youngsters. Long term diabetes is one of the reason for getting CKD [1]. Similar to this, having hypertension for a long period of time results in CKD [2]. Apart from these two, there are so many other proven factors responsible for this. Like the food we eat, always living in air-conditioned rooms which prevents sweating and thereby directly giving more work for the kidneys, etc., there are many number of reasons. Because of all these comforts people are having, they have become obese and not ready to physically work. Most of the young mind set is to work hard day and night, using the brain by sitting in the same comfortable air conditioned room. That lifestyle is also a reason for the CKD. Drinking less amount of water and taking food at irregular times also has its effect. So there is a marked difference in a list of human vital parameters related to this disease in the CKD patients under observation. The data set used for this work is downloaded from the UCI Machine Learning Repository. There are data pertained to 25 parameters collected from 400 patients over a period of 2 months.

The CKD patients have either both or one kidney failure. Sometimes patients with partial kidney failure also come under this category. These patients become so thin and fragile and will be going through frequent dialysis [3]. Dialysis is the physical process of filtering the excess water and fluid waste from the blood. It is the major function of the kidneys. There are machines available which does this dialysis process for the CKD patients externally.

The number of times a patient has to undergo this dialysis depends upon the level of kidney failure. It is a time consuming and tiring process for the patients. Working people with CKD suffer the most. It is costly, painful and time consuming. But that is unavoidable for CKD patients. So detecting the CKD in an early stage would help them recover from that using proper medication or through Kidney transplantation.

Prediction system using Machine Learning algorithms are helpful in so many applications nowadays like weather forecasting [4], stock price prediction [5], traffic flow prediction [6], choosing right crop for agriculture land [7], predicting leaf diseases in plants [8], landslide like disaster detections in advance [9], etc., Prediction system for deadly diseases [10] plays a vital role in the medical field helping doctors with fast diagnosis. It helps saving many precious lives.

So in this work such a prediction system using machine learning algorithms has been proposed. Based on the performance evaluation of three algorithms namely

Randomizable Filtered Classifier, Logistic Regression and REP Tree, one of the best algorithm giving higher rate of accuracy is chosen and further ensemble bagging algorithm is applied for improving the performance.

RELATED WORKS

Machine Learning is a boon to Healthcare Industries. During the pandemic period, it was so helpful in variety of predictions in early detection of the disease using various parameters and also helped in the field of pharmaceutical industries. Similarly, prediction system for early detection of some dreadful diseases is broadly being used in many healthcare applications. Some of the works, where the ML algorithms played a major role in disease predictions are discussed in this section.

In this work [11], based on the laboratory data collected from a patient, classification algorithms namely Logistic Regression and KNN algorithms were implemented and the prediction of heart disease was made. It has been said that this worked more accurate than the Naïve Bayes and other classification algorithms. The implementation has been done on the .pynb format.

The work by Shadman Nashif et al., [12] describes about a real time cloud based heart disease prediction system. In this system, they used the Support Vector Machine (SVM) for classification and prediction of heart disease from a set of clinical data. Apart from this prediction system, they developed a real time hardware set up to monitor the elderly patients with heart problem. That was used to monitor the live vital parameters of the patient. So by comparing the real time data with prediction system, if anything goes wrong, that information will be immediately sent to the physician through phone.

In the work of Eman M Alanazi [13], many machine learning algorithms were used concentrating on 6 parameters of the collected health related data from the Government agencies related to brain stroke. They came to a conclusion that data resampling method was more fruitful than the data selection procedures over the Random Forest classification algorithm. The result percentages were discussed. They have taken data set from nearly 15000 patients, among them only 17% were actually suffering from stroke. In [14], a full statistical dataset was analysed and it has been concluded that the mental stroke occurs to patients with higher values of five important vital health parameters. Further they have used perceptron neural network for providing the highest accuracy rate in stroke prediction.

The use of Machine Learning for assisting the organ (heart, kidney, liver, etc.,) transplantation has been explained in this work [15]. It has been stated that the Artificial Neural Network and logistic regression models have proven successful in optimizing the donor - recipient match finding. In particular it was helpful for finding a matching pair of kidneys from a pool of kidney datasets. Machine Learning algorithms are useful in post-transplant care also. Normally the human body tries to repel the foreign particles if they enter. Similarly the human body of the recipient patient also will try to repel the implanted new kidney. So the ML models help in monitoring the vital parameters of the patient immediately after operation for the successful survival of the patient. Even if there is any chance of repulsion, the models will predict that in advance.

PROPOSED METHODOLOGY FOR CKD PREDICTION

The idea is to use the dataset of CKD patients and using the appropriate ML model for predicting the disease in advance. There are many types of Machine Learning algorithms available under the categories namely supervised, unsupervised, semi-supervised and reinforcement learning algorithms. In this work the predictions of Logistic Regression, Randomizable Filtered Classifier and Reduced Error Pruning (REP) Tree algorithms are studied. From the results obtained, to achieve the maximum performance, Bagging is applied over the algorithms and best optimized result has been achieved [Fig.1].

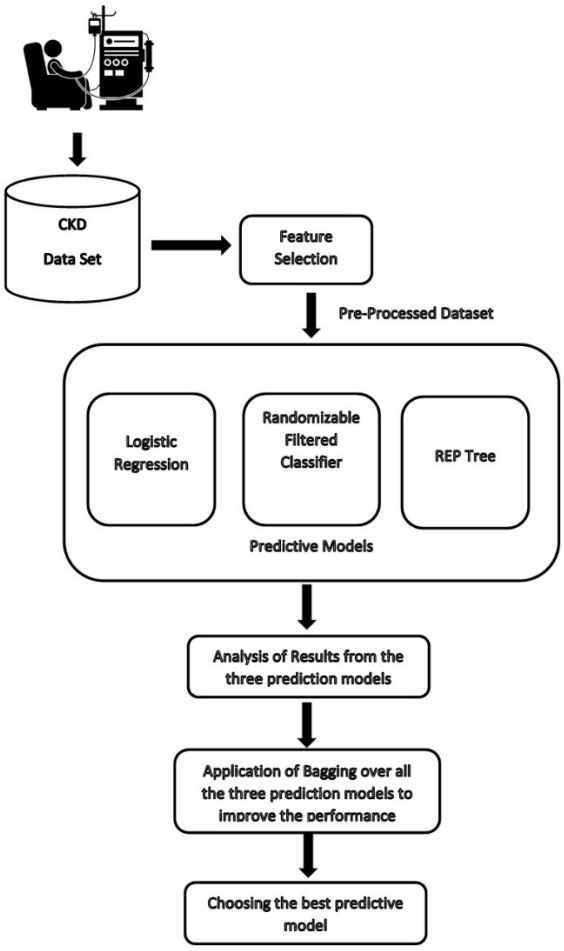

FIGURE I
DECISION SUPPORT SYSTEM FOR CKD PREDICTION

I. Data Collection and pre-processing

The first step is to decide the data set. Either the data can be collected from CKD patients through a proper communication with the hospital administration or it can be directly taken from authorized health data repository. For this work the dataset has been downloaded from the UCI (UC Irvine) repository. The dataset is about 400 CKD patients with 25 parameters. Among them 11 are numerical data and 14 are nominal data (Table I).

TABLE I
NUMERICAL AND NOMINAL CLASSIFICATION OF DATASET ON CKD PATIENTS

S.No	Parameter	Numerical/Nominal
1.	Age (age)	Numerical
2.	Blood Pressure (bp)	Numerical
3.	Specific Gravity (sp)	Nominal
4.	Albumin (al)	Nominal
5.	Sugar (su)	Nominal
6.	Red Blood Cells (sbc)	Nominal
7.	Pus Cell (pc)	Nominal
8.	Pus Cell clumps (pcc)	Nominal
9.	Bacteria (ba)	Nominal
10.	Blood Glucose Random (bgr)	Numerical
11.	Blood Urea (bu)	Numerical
12.	Serum Creatinine (sc)	Numerical
13.	Sodium (sod)	Numerical
14.	Potassium (pot)	Numerical
15.	Hemoglobin (hemo)	Numerical
16.	Packed Cell Volume (pcv)	Numerical
17.	White Blood Cell Count (wbcc)	Numerical
18.	Red Blood Cell (rbcc)	Numerical
19.	Hypertension (htn)	Nominal
20.	Diabetes Mellitus (dm)	Nominal
21.	Coronary Artery Disease (cad)	Nominal
22.	Appetite (appet)	Nominal
23.	Pedal Edema (pe)	Nominal
24.	Anaemia (ane)	Nominal
25.	Class (class)	Nominal

The original csv file had some 'Not a Number' values, some outliers and duplicate values. Those values were removed and the file was made ready for running the ML algorithms. Feature selection is used for removing the parameters which barely have any effect on the prediction of CKD. The attribute selection using classification category of supervised learning is applied. It reduced the parameters from 25 to 18 (Table II).

TABLE II
DATASET AFTER THE APPLICATION OF FEATURE SELECTION

S.No	Parameters
1.	Bp
2.	Sg
3.	Al
4.	Rbc
5.	Bgr
6.	Bu
7.	Sc
8.	Sod
9.	Pot
10.	Hemo
11.	Pcv
12.	wbcc
13.	Htn
14.	Dm
15.	appet
16.	Pe
17.	Ane
18.	class

II. Building the Predictive Models

The next step is to analyse the data under all the above refined attributes and start applying the various ML models namely Logistic Regression, Randomizable Filtered Classifier and Reduced Error Pruning (REP) Tree.

- **Logistic Regression**

The Logistic Regression is mainly used for prediction based on collected data [16]. Hence that is selected as the first model to be implemented on this CKD dataset. It is a kind of supervised learning algorithm. It does the calculation based on the success and failure rate and the result would be in the form of a probability value. So there would be only two classes- either the patient is having chronic kidney disease (ckd)or not having CKD (notckd). Hence this is used for building the first ML model.

In the Logistic Regression, a sigmoid function is used in the calculation of Odd. Odd is similar to probability. It calculates the ratio of the prediction being ckd and notckd. If the input attributes are represented as X, the class as Y, the coefficients as c, weight as wt and Z as the input to the sigmoid function, then the equation for Odd can be written as

$$\frac{P(X)}{(1-P(X)} = e^{Z} \qquad (1)$$

$$\ln\left(\frac{P(X)}{(1-P(X)}\right) = Z \qquad (2)$$

$$\ln\left(\frac{P(X)}{(1-P(X)}\right) = wt.X + c \qquad (3)$$

So the Logistic Regression equation for the taken dataset would be

$$P(X; c, wt) = \frac{1}{1+e^{-wt.X+c}}$$

Above is the equation for class 0 that is ckd, for class 1, that is nonckd, it would be $1 - P(X; c, wt)$

Implementing the Logistic Regression on the 10 fold cross validation of data gives a result of 388 (97%) instances getting correctly classified and the remaining 12 (3%) getting incorrectly classified. There are so many evaluation metrics available for analysing the built ML model. Accuracy, Precision, Recall, F1 Score, Confusion matrix are some of those evaluation metrics. Confusion matrix gives this accuracy measure in the form of a matrix of rows

representing the predicted class of the ckd samples and columns representing the actual class of the sample. Confusion Matrix gives the following data (Tables III and IV):

TABLE III
CONFUSION MATRIX STRUCTURE

True Positive	False Positive
False Negative	True Negative

TABLE IV
CONFUSION MATRIX AFTER APPLYING LOGISTIC REGRESSION

ckd	notckd
241	9
3	147

- **Randomizable Filtered Classifier**

This is also a supervised learning model coming under the 'meta' category of classification. It is a type of the filtered classifier which applies a transformation to the input data, and then implements Instance Based Learner (IBk) which is actually K-Nearest Neighbour (KNN) algorithm. It is known as instance based because it builds the prediction from the training dataset. By doing so, the accuracy would be more effective in the filtered input than in the original. In instance based learning, tuning is not required for the attributes. The process is actually tightly woven with the previous data in the form of some weight values. Here, because of this simple technique, the learning happens very slowly. By applying this model on the ckd data, an accuracy of 89.75 % (359 instances were correctly classified and 41 instances were incorrectly classified) has been obtained. The confusion matrix for this model gives the following values (Table IV):

TABLE V
CONFUSION MATRIX AFTER APPLYING RANDOMIZABLE FILTERED CLASSIFIER

Ckd	notckd
224	26
15	135

- **Reduced Error Pruning (REP) Tree**

Overfitting problem occurs in the Randomizable Filtered Classifier method. This prevents the final development of a good training model. Hence this Reduced Error Pruning Tree method is chosen next for the training. Initial step is to repeat the process of training the data and growing the tree. The next step is validation. This helps in finding which part of the tree (nodes) needs to be pruned. How the tree is behaving initially is compared with how it behaves after pruning a particular node and its branches using backfitting. Then that node becomes a normal leaf node with the label of 'ckd'. This way, the ML algorithm works. The algorithm finally gives a prediction of 97.5 % in correctly classifying the data. The confusion matrix (Table VI) of this implementation is

TABLE VI
CONFUSION MATRIX AFTER APPLYING RANDOMIZABLE FILTERED CLASSIFIER

Ckd	notckd
248	2
8	142

III. Performance Boosting

After the implementation of all the three mentioned ML algorithms, the results show that the Logistic Regression (97%) and REP Tree (97.5%) are giving an accuracy of around 97%. The Randomizable Filtered Classifier method has given a prediction accuracy of 89.75%. In an attempt to increase the performance of these algorithms, ensemble bagging is executed using the same three algorithms. It gave a productive result. Logistic Regression now gives an accuracy of 99.25%, Randomizable Filtered Classifier gives an accuracy of 92.25% and REP Tree now gives a result of 98.75%. From the results, it is obvious that Logistic Regression has proven to be the fool proof method for doing disease prediction using stored dataset. Next to that, REP Tree can also be applied as the time taken for building the REP Tree Model with bagging is only 0.22 seconds whereas the Logistic Regression Model takes 0.88 seconds with bagging for giving the prediction results.

RESULTS AND DISCUSSION

To minimise the time taken for doing the predictive analysis, feature selection is used for fine tuning the attribute numbers as the initial step. The effect of the 18 attributes which contribute the most for the classification and their percentages are given in the Fig.2. It is also showing that sugar (diabetes mellitus) contributes the most for the development of CKD. Haemoglobin (hemo), albumin (al), blood glucose random level (bgr) are always on the decreasing side for the patients. It is dangerous as it may go below normal level anytime and could be fatal. On the other side serum creatinine (sc) and potassium (pot) are on the higher side and it increases for the patients. That is also a dangerous condition indicating deterioration of health.

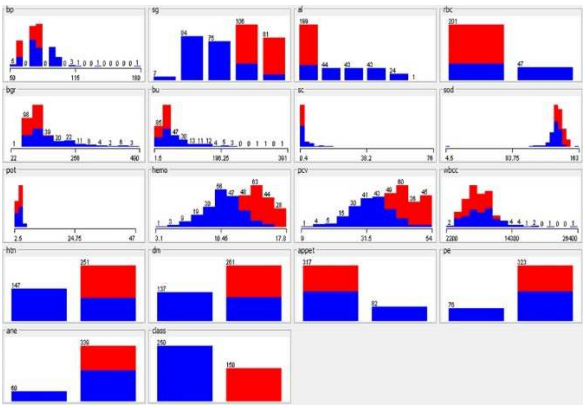

FIGURE 3
ANALYSIS OF ATTRIBUTES

After the attribute selection, the ML models were developed for the CKD dataset. Their percentage of accuracy in prediction and their result after applying bagging for improving their performances were compared and discussed in Table VII.

TABLE VII
PERFORMANCE EVALUATION

S. No	ML Model	Percentage of Accuracy	Percentage of accuracy after applying bagging
1	Logistic Regression	97	99.25
2	Randomizable Filtered Classifier	89.75	92.25
3	REP Tree	97.5	98.75

The REP Tree gave the highest accuracy in prediction while comparing with the other two algorithms which is shown in the Fig. 3. But instead of applying the bagging on the REP Tree, it has been applied on all the three models to study their performances.

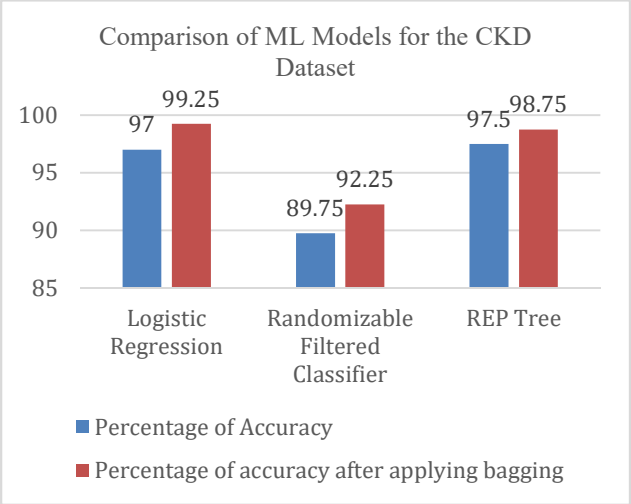

FIGURE 4
PERFORMANCE COMPARISON OF THE ML MODELS WITH AND WITHOUT BAGGING

It gave a result showing Logistic Regression Model is the best model for CKD data set. It showed a sharp increase in the accuracy after applying the bagging. So, for the prediction of chronic kidney disease based on the stored dataset, ML model built using Logistic Regression gives the best result.

CONCLUSION

The Chronic Kidney Disease is prevalently existing nowadays and due to this, the number of patients undergoing dialysis is also on the increasing pattern. Long back, it was common to elderly people who were suffering with diabetes mellitus for a long duration. But it has changed now. Regardless of age, people are getting affected by CKD. The physical pain and the financial critical situations they face is miserable. This work presented here is an attempt to identify the upcoming of CKD condition in prior and also to find out which parameters are actually worsening the CKD condition. So that, preventive steps can be taken to avoid entering the crucial stage of kidney failure. Three Machine Learning algorithms were studied in detail. Bagging is done to further improve the performance of prediction. The final results showed that the Logistic Regression gave the higher optimal prediction accuracy based on the analysis of given CKD patients' dataset. However REP Tree gives a very close percentage of accuracy with less time consumption.

ACKNOWLEDGEMENT

I acknowledge the School of Computer Science and Engineering (SCOPE) of Vellore Institute of Technology (VIT) Chennai campus for supporting this research work.

REFERENCES

[1] Thomas, Merlin C. et al., "Diabetic kidney disease", Nature Reviews Disease Primers. Vol.1, No.1, 2015. DOI: 10.1038/nrdp.2015.18.

[2] Weldegiorgis, Misghina, Woodward, Mark, "The impact of hypertension on chronic kidney disease and end-stage renal disease is greater in men than women: a systematic review and meta-analysis", BMC Nephrology. Vol.21, No.1, 2020. DOI: 10.1186/s12882-020-02151-7.

[3] Sarah Elshahat at al., "The impact of chronic kidney disease on developed countries from a health economics perspective: A systematic scoping review", Plos One journal. March 24, 2020. https://doi.org/10.1371/journal.pone.0230512.

[4] G. Hemalatha, K. Srinivasa Rao and D. Arun Kumar, "Weather Prediction using Advanced Machine Learning Techniques", *J. Phys.: Conf. Ser.* 2089 012059 **DOI** 10.1088/1742-6596/2089/1/012059

[5] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", Procedia Computer Science,Volume 167, 2020, Pages 599-606, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.326.

[6] Razali, N.A.M., Shamsaimon, N., Ishak, K.K. et al. Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning. J Big Data 8, 152 (2021). https://doi.org/10.1186/s40537-021-00542-7.

[7] Prasad, S. & Thangatamilan, M. & Aravindan, V. & Harish, A. & Janani, S. & Kausika, S.. (2021). Selection of the Best Crop for Farming Using Machine Learning. In book: Materials, Design, and Manufacturing for Sustainable Environment (pp.755-765). 10.1007/978-981-15-9809-8_55.

[8] Sunil S. Harakannanavar, Jayashri M. Rudagi, Veena I Puranikmath, Ayesha Siddiqua, R Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms". Global Transitions Proceedings, Volume 3, Issue 1, 2022, Pages 305-310, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2022.03.016.

[9] Neha Gupta, Kamlesh Kumar Rana, "Disaster Prediction And Post Disaster Management Using Machine Learning and Bluetooth". Webology (ISSN: 1735-188X). Vol. 18, No. 5, 2021.

[10] Rahatara Ferdousi, M. Anwar Hossain and Abdulmotaleb El Saddik, "Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS". IEEE Access. July, Vol.9, 2021. DOI. 10.1109/ACCESS.2021.3094063.

[11] Harshit Jindal et al., "Heart disease prediction using machine learning algorithms". 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072 DOI 10.1088/1757-899X/1022/1/012072

[12] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System". World Journal of Engineering and Technology Vol.6 No.4, November 2018. DOI: 10.4236/wjet.2018.64057.

[13] Eman M Alanazi, Aalaa Abdou, and Jake Luo, "Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models". JMIR Form Res. Vol.5(12); 2021.

[14] Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John. A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthcare Analytics, Volume 2, 2022, ISSN 2772-4425, https://doi.org/10.1016/j.health.2022.100032.

[15] Gotlieb, N., Azhie, A., Sharma, D. et al. The promise of machine learning applications in solid organ transplantation. npj Digit. Med. 5, 89 (2022). https://doi.org/10.1038/s41746-022-00637-2.

[16] Bernard X. W. Liew, Francisco M. Kovacs, David Rügamer & Ana Royuela. Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. European Spine Journal. Vol.31, No.8, March 2022. https://doi.org/10.1007/s00586-022-07188-w.

AUTHOR INFORMATION

Dr.R.Rathna, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai.