International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET22924

Prediction of Chronic Kidney Disease-A Machine Learning Perspective

Kantharaju. V, R. Pavithra, Nisarga H, Karishma S
KNSIT College of Engineering, Bangalore, Karnataka, India

Article Info
Volume 9, Issue 2
Page Number : 37-43
Publication Issue :
March-April-2022
Article History
Accepted : 05 March 2022
Published: 15 March 2022

ABSTRACT

Chronic Kidney Disease is one of the most critical illnesses nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on the following factors given below: (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv)Least absolute shrinkage and selection operator regression, (v) synthetic minority over- sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority oversampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F- measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%. Along with machine learning models one deep neural network has been applied on the same dataset and it has been noted that deep neural network achieved the highest accuracy of 99.6%.

Keywords: LSVM, Machine Learning Technique, Support Vector Machine

I. INTRODUCTION

Chronic kidney Disease (CKD) means your kidneys are damaged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body[1]. This disease is chronic because of the damage gradually over a long period. It is flattering a common disease worldwide. Due to CKD may have some health troubles? There are many causes for CKD like diabetes, high blood pressure, heart disease. Along with these critical diseases, CKD also depends on age and gender. If your kidney is not working, then you may notice one or more symptoms like abdominal pain, back pain, diarrhea, fever, nose bleeds, rash, vomiting.

There are two main diseases of CKD: (i) diabetes and (ii) high blood pressure. So that controlling of these two diseases is the prevention of CKD. Usually, CKD does not give any sign till kidney is damaged badly. CKD is being increased rapidly as per the studies hospitalization cases increase 6.23 percent per year but the global mortality rate remains fixed.

There are few diagnostic tests to check the condition of CKD:

(i) Estimated glomerular filtration rate (eGFR) (ii) urine test (iii) blood pressure.

A.EGFR

eGFR value shows that how your kidney cleaning the blood. If your eGFR value is greater than90, that means the kidney is normal. If eGFR value is less than 60, that means you have CKD.

B.URINE TEST

The doctor also asks for urine test for kidney functionality because kidneys make urine.

If the urine contains blood and protein, that means your kidney is not working properly.

C.BLOOD PRESSURE

Doctor measures blood pressure as Blood pressure range shows how your heart is pumping blood[2]. If eGFR value reaches less than 15, that means the patient has end-stage kidney disease.

At this point, there are only available treatments:

(i)Dialysis and (ii) kidney transplant. If dialysis is not possible, the doctor has only one solution, i.e., kidney transplantation. However, it is extremely expensive. Therefore, it is critical noteworthiness in early recognition, monitoring and handling of the disease. It is essential to predict the striding of CKD with appropriate accuracy due to its dynamic and secretive nature in the early stages and patient abnormality. Medical treatment of CKD is prescribed by the stage. Anything other than this, it is very imperative to characterize the organization of the infection because it gives a few indications. It underpins the assurance of fundamental intercessions and medications. Medical treatment is a very significant application area of intellectual intelligent systems. Afterwards, Data mining can play a big role to find out hidden information from the huge patient medical and treatment dataset that doctors frequently obtain from patients to get pieces of knowledge about the symptomatic data and to execute precise treatment plans.

Data mining can be categorized as the method of extracting hidden information from a huge dataset. Data mining strategies are connected and utilized broadly in various contexts and areas. Using data mining methods, we may predict, classify, filter and cluster data. The objective states the algorithm processing of a training set containing a set of attributes and targets. Data mining is suitable to mining in data if the dataset is huge but we can also do it with the help of machine learning. The machine learning can also find data analysis and pattern detection. A variety of health dataset is present so machine learning algorithms are best fit to improve the accuracy of diagnosis prediction. As healthcare electronic dataset grows rapidly, machine learning algorithms are becoming more common in healthcare. This research article primarily aims to predict whether a person has Chronic Kidney Disease or not. In this perception,

seven different machine learning classifiers were applied on the dataset. All the algorithms were running with both full features and selected features. SMOTE was used for oversampling and all the results were recorded. All the machine learning model results were also compared with one deep neural network algorithm. Deep learning neural network was used with two hidden layers. IBM SPSS Modeller was applied for computational purpose. The contribution reveals the accuracy estimate of 99.6% when applying deep neural network on the dataset.

II. SYSTEM ANALYSIS

A.EXISTING SYSTEM

Three classification techniques are used: -nearest neighbor's classifier, decision tree classifier (DT), and logistic regression. Machine learning classifiers are used to forecast a data point's class, target, labels, and categories[3].

• Disadvantages:

The overall achieved accuracy is 89%.

Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector

Machine (SVM) and Radial Basis Function (RBF) algorithms

• Disadvantages:

The training time for the dataset is more so the above framework model is less efficient. Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques using Decision Stump, Hoeffding Tree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPTree, and SimpleCart.

• Disadvantages:

The above methods are concentrating only on Decision Tree Model, As Dataset size increases, the testing and validation also takes longer time to output the results.

B.PROPOSED SYSTEM

Until now, in majority of cases full features have been taken into consideration. In this research feature optimization was carried out, wherein three different feature selection algorithms were applied to find the algorithm most beneficial to extract the important feature for the prediction of Chronic Kidney Disease. As many datasets have imbalanced class, class balancing is needed for increasing the performance of classifier model. In this research SMOTE was used as a class balancer. The highest accuracy of 99.6% was achieved whereas the article provides an accuracy of 99.1% on the same dataset. According to the highest accuracy of the model was 99.7%, but they worked on risk calculation of the patient whereas the main aim of the article is to predict Chronic Kidney Disease.

Our Project primarily aims to predict whether a person has chronic kidney disease or not. In this perception, seven different machine learning classifiers were applied on the dataset. All the algorithms were running with both full features and selected features. SMOTE was used for oversampling and all the results were recorded. All the machine learning model results were also compared with one deep neural network algorithm. Deep learning neural network was used with two hidden layers. IBM SPSS Modeler was applied for computational purpose. The contribution reveals the accuracy estimate of 99.6% when applying deep neural network on the dataset.

III. SYSTEM DESIGN

A.DATASET

Chronic Kidney Disease dataset is used for this research work. Many researchers had also used this dataset. This dataset is being provided by the UC Irvine Machine Learning Repository and it is available on the UCI website. This dataset contains 400 instances and 24 attributes with 1target attribute. The target attribute has labelled in two-class to represent CKD or non-CKD. The dataset was collected from various hospitals in 2015. It contains also missing value.

B.METHODOLOGY

In this research, we have developed a model to predict CKD disease in patients. The performance of the model was tested on both all attributes and selected features. Among feature selection methods there were Wrapper, Filter and Embedded allowing selecting vital features[4]. Classifier algorithms performance was tested on the selected features. IBM SPSS tool is used for preparing the model. The machine learning classifiers such as artificial neural network (ANN), C5.0, logistic regression, linear support vector machine (LSVM), K- nearest neighbors (KNN) and random tree were used for training the model. Each classifier validation and performance matrix was computed. The procedure of this research including five stages: (i) dataset preprocessing, (ii) feature selection, (iii) classifier application, (iv) SMOTE and (v) analyzing the performance of the classifier. Along with machine learning models, a deep neural network was applied for comparing the result of machine learning models and deep neural network. Artificial Neural network classifier was used for this purpose. In this research the significance of two models were checked by statistic testing namely Mc Nemar's test.

C.PREPROCESSING OF DATA

Data preprocessing could be a strategy that is utilized to change over the raw
information into a clean dataset. It is a basic step to train every machine learning classifier algorithm. This technique concludes such actions as handle missing values, rescaling of the dataset, transform
into binary data and standardize of the dataset. When the dataset included attributes with varying scales, rescaling is used to scale the dataset. The binary transformation has been applied to convert the value into 0 and 1. All values of every attribute are considered as 1 for above the threshold and as 0 for below the threshold. Standardized method ensures that each attribute has mean 0 and standard deviation 1 [5].

D.FEATURE SELECTION

Feature selection is needed for trained each machine learning classifier because
without removing unnecessary attributes from the dataset result may be affected. The classifier algorithm with feature selection gives better performance and reduces the execution time of the model. For this process, three different feature selection methods were used in this research.

1) FILTER METHOD

The filter is one of the methods to select the appropriate feature. It selects the feature
on their integral features without integrating any learning classifier algorithm. This method gives result faster as compared to the wrapper method. The method assigns the score to every attribute based on their statistical correlator between attributes. There are many filter methods are available, but Correlation-based Feature Selection (CFS) method has been used. CFS is the algorithm to select the feature-based on the attribute ranks. It assigns the rank to attribute subset as based on the correlation heuristic evaluation function. The function works on the strategy that creates two class labels, one is correlated to class and low correlated class and selects only correlated label class attributes.

2) WRAPPER METHOD

Wrapper method selects the subset of features based on a precise machine learning algorithm. It used the greedy search method for finding a possible subset of features. The method can be implemented with using any of the following algorithms forward selection, backward elimination and recursive elimination. In the research, we used the forward feature selection method [6]. The forward feature selection iteratively selects the feature. This procedure starts with the null model and works iteratively and adds the attribute in each step.

The attribute is keeping add in the model until the attribute does not improve model performance.

3) EMBEDDED METHOD

The embedded method is decision tree algorithm for feature selection. It selects the feature in each step works recursively while the tree is growing and split the sample set into a smaller subset. The most common decision tree algorithm is: ID3, C4.5 and CART. There are other available method is creating linear models. The most common methods are LASSO [7]. with L1 penalty and Ridge with L2 penalty. In this research LASSO (least absolute shrinkage and selection operator) algorithm has been used. It performs two main tasks: regularization and feature selection. In regularization, it shrinks some feature coefficients to zero that means features are not important for the predictor model.

E.CLASSIFICATION ALGORITHMS

Classification technique is an important feature of supervised learning. Classifiers learn from the training dataset and apply on the testing dataset for finding the target attribute.

Below there are classification techniques used in research.

1) ARTIFICIAL NEURAL NETWORK

Artificial neural network is a part of artificial intelligence. It is a type of supervised machine learning. Its structure is the same as the human brain. ANN also has neurons and just like in human all neurons are interconnected to one another, ANN neurons are connected to each other in layers of the network. Neurons there are known as nodes[8].

2) C5.0

C5.0 is a type of decision tree because it creates the decision tree from the input. The tree has the number of branches. It utilizes the tree structure to model the relationship between features and potential outcomes. At each node of the tree, the attribute of the dataset is chosen. It can handle nominal and numeric features both. C5.0 is the extended version of the C4.5 classification algorithm and uses information entropy concept.

3) LOGISTIC REGRESSION

Logistic regression is also a type of supervised learning algorithm. It is a statistical model. The probability of target value is predicted from logistic regression. It is divided the target attribute into two-classes: success or not success. For success, it returns 1 whereas it returns 0 for not succeeding.

4) CHAID

Chi-square automatic interaction detection (CHAID) is a type of decision tree technique. It is used to determine the relationship between variables. Nominal, ordinal and continuous data can be used in CHAID for finding the outcome. For each categorical predictor, all possible cross- tabulation is created in the CHAID model and it process works until the best outcome is attained.

5)LINEAR SUPPORT VECTOR MACHINE (LSVM)

Linear support vector machine (LSVM) is the modern particularly fast machine learning algorithm for solving multiclass classification problem for the large dataset based on a simple iterative approach. It is created the SVM model in linear CPU time of the dataset.

LSVM can be used for the high dimensional dataset is the sparse and dense format[9].

5) K- NEAREST NEIGHBORS (KNN)

KNN is a simple type of supervised algorithm. It can be used for both classification and regression problems. However, it is largely used for classification problems. KNN does not use a particular training stage and use all the data for training so that it is a lazy learning algorithm and also it does not consider anything about the underlying data, so that is a nonparametric learning algorithm.

6) RANDOM TREE

The random tree is a type of supervised classifiers. It produces lots of distinct learners. The stochastic process is used to form the tree. It is a type of ensemble learning technique for classification. It works the same as decision tree, but a random subset of attributes uses for each split. This algorithm uses for both classification problems and regression problems.

F.VALIDATION METHOD OF CLASSIFIERS

The dataset was divided into parts: training dataset and testing dataset. IBM SPSS modeler was used for the

partition and prediction of the result. The training dataset contains 50% of the data and remaining data is considered as the testing data. The type tool of IBM SPSS was applied for changing the type of attributes [10]. The performance evaluation matrix was received for each classification algorithm.

G.PERFORMANCE EVALUATION MEASURE

Various evaluation matrices were used for checking the performance of the

classifier. For this purpose, the confusion matrix was used. It is a 2_2 matrix due to two classes in the dataset. The confusion matrix gives two types of correct prediction of the classifier and two types of incorrect prediction of the classifier.

H.CONFUSION MATRIX DESCRIPTION

TP: True Positive means output as positive such that predicted result is correctly classified.

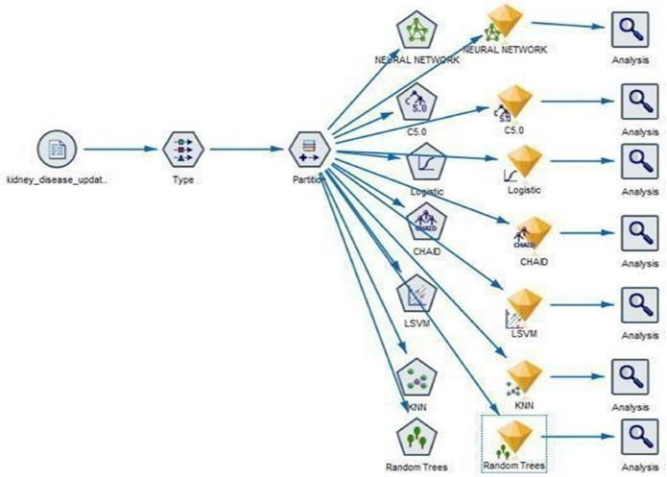
TN: True Negative means output as negative such that predicted result is correctly classified.

FP: False Positive means output as positive such that predicted result is incorrectly classified. FN: False Negative means output as negative such that predicted result is incorrectly classified [11].

- Classification Accuracy
- Classification Error
- Precision
- Recall
- F-Measure
- Roc And Auc
- Gini Coefficient
- Smote

I. STATISTICS TEST FOR MODEL COMPARISON

For the purpose of comparing two models, McNemar's test was applied on the predicted output of two models. The McNemar's test is used to determine whether there are differences On bipolar dependent variables between two related groups.



IV. CONCLUSION

This article objects to predict Chronic Kidney Disease based on full features and important features of CKD dataset. For feature selection three different techniques have been applied: correlation-based feature selection, Wrapper method and LASSO regression. In this perception, seven classifiers algorithm were applied viz. Artificial neural network, C5.0, logistic regression, CHAID, linear support vector machine(LSVM), K-Nearest neighbours and random tree. For each classier, the results were computed based on full features.

V. REFERENCES

[1]. J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, ``A machine learning methodology for diagnosing chronic kidney disease,'' IEEE Access, vol.8, pp. 20991_21002,2020.

[2]. L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino, ``Kidney failure detection using machine learning techniques,'' in Proc.8th Int. Workshop ADVANCEs ICT Infrastructures Services, 2020, pp. 1_8.

[3]. B. Deepika, ``Early prediction of chronic kidney disease by using machine learning techniques,'' Amer. J. Comput. Sci. Eng. Survey, vol. 8, no.2, p. 7, 2020.

[4]. F. Ma, T. Sun, L. Liu, and H. Jing, ``Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network, 'Future Gener. Comput. Syst., vol. 111, pp.17_26, Oct. 2020.

[5]. G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger and J. A. Recio-Garcia, ``Explainable prediction of chronicrenal disease in 31 the olombian population using neural networks and case-based reasoning,'' IEEE Access, vol. 7, pp. 152910,2019.

[6]. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Aliss, ``Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,'' Sensors, vol.20, no. 9, p. 2649, May 2020.

[7]. U. H. Amin, J. Li, Z. Ali, M. H. Memon, M. Abbas, and S. Nazir, ``Recognition of the Parkinson's disease using a hybrid feature selection approach,'' J. Intell. Fuzzy Syst., vol. 39, no. 1, pp. 1_21, Jul. 2020.

[8]. P. G. Scholar, ``chronic kidney disease prediction using machine learning,'' Int. J. Eng. Res. Technol., vol. 9, no. 7, pp. 137_140, 2020.30 .

[9]. S. Shankar, S. Verma, S. Elavarthy, T. Kiran, and P. Ghuli, ``Analysis and prediction of chronic kidney disease,'' Int. Res. J. Eng.Technol., vol. 7,no. 5, May 2020, pp. 4536_4541.

[10]. G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P.Moreno-Ger, and J. A. Recio-Garcia, ``Explainable prediction of chronicrenal disease in 31 the colombian population using neural networks and case-based reasoning,'' IEEE Access, vol. 7, pp. 152900_152910,2019.

[11]. W. D. Souza, L. C. D. Abreu, L. G. D. SilvaI, and I. M. P. Bezerra, ``Incidence of chronic kidney disease hospitalizations and mortality in Espirito Santo between 1996 to 2017,'' WisitCheungpasitporn, Univ. Mississippi Medical Center, Rochester, MN, USA, Tech.Rep., 2019, doi: 10.1371/journal.pone.0224889.

Cite this article as :

Kantharaju. V, R. Pavithra, Nisarga H, Karishma S, "Prediction of Chronic Kidney Disease-A Machine Learning Perspective", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 2, pp. 37-43, March-April 2022. Available at

doi : https://doi.org/10.32628/IJSRSET22924

Journal URL : https://ijsrset.com/IJSRSET22924