Original Research Article

A machine Learning Methodology for Chronic Kidney Disease

P Manjula PriyaDarsini[1*], M S V S Bhadri Raju[2]

[1]M.Tech, Computer Science and Technology, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India

[2]Professor, Department of CSE, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India

Received: 28-11-2021 / Revised: 21-12-2021 / Accepted: 02-01-2022

Abstract

Most of the people in the world are suffering a lot with Chronic Kidney Disease (CKD) widely considered to be a global issue. People with heart disease, hypertension, diabetes or who have suffering from CKD family they are at risk.This is related to improved mortality and high risk of a few other conditions, including the raise of heart disease and human services. The adverse outcome of CKD can be prevented with proper early detection and management. Machine Learning (ML) aids to analyze tens of thousands of data points and produce outcomes, provides regular risk scores, precise resource allocation, and has many other applications healthcare. In this paper some machine learning techniques are used and compared namely Support Vector Machine (SVM), Naïve Bayes (NB), NB-TREE, and Iterative Random Forest (IRF). Among all these four algorithms IRF got the best accuracy of 99.4%. Here we are using IRF which provide the stable decision paths with high order interactions. Hence IRF could be used for any data to get good accuracy.

Keywords: CKD, ML, Iterative Random Forest

Introduction

Chronic Kidney Disease (CKD) iswidely regardedas a global issue[1]. CKD is also referred to as chronic renal failure or chronic kidney disease, which has many causes[12], and is abundantly widespread than individuals, Symptoms grow slowly and are not disease specific. Present according to medical and health statistics 10% of the population is affectedby CKD andnearly 58 million deathsoccurred in 2005around the world[14]. Where according to the World Health Organization (WHO) 35 million attributed to chronic diseases. As per the Global Burden of Disease study in 1990,CKDwas ranked 27th and it was unfortunately rose to18th in 2010[3]. More than two million people are treated with kidney transplantation or dialysis to stay alive; so far this count can only be reported to 10 per cent of peoplewho need treatment to live[15]. As indicated by the National Kidney Foundation there are twenty-six million of people adults in the US who have CKD and millions of others who are at increased risk[4].

In developed countries for self-monitor of CKD Multi care Android apps were developed based on medical recommendations [8].UCD (User-centered design) methodology identify the CKD Patients in early stage and also Very helpful t oconfirm the diagnosis of CKD[2]. Machine Learning plays very important and essential role in Health and Medical diagnosis. Several ML methods and metrics helping healthcare specialists to build alternative workforce strategies, deliver smart easy healthcare, and reduce operating and delivery costs, in addition helps to examine thousands of data points and suggest.

Related work

The cure of CKD has taken numerous forms throughout the years. So for the fast and accurate result many tools and techniques were used by researchers. Different researchers used different models, and techniques to find CKD in its early stage.

Veenita Kunwar et al[13], has used one of the Data mining (DM) techniques. Naïve Bayes and Artificial neural networks (ANN) are applied on the dataset collected from UCI which has only 220 values after cleaning.With Rapid Minor tool NB provides higher prediction accuracy than ANN.In datamining is a bit challenging because it's a manual technique and also the estimations or predictions may be incorrect.So the researchers have started to focus on ML.

Kabir Hashi et al.[5], has used two ML models DT and K-Nearest Neighbor (KNN) classification models on Pima Indians Dataset by considering WEKA tool and K value as 7 the accuracy can be improved by DT.But Even small change of data lead to change the entire optimal tree structure.

Husey in Polat et al.[10], has explained the two popular feature reduction ML techniques such as wrapper and filter. SVM improved the rate of accuracy after employing the filter subset evaluator.Fundamental fault of SVM is that it is volatile.

Iliyas Ibrahim et al.[6], worked on Deep Neural net- works (DNN) which is the part of AI to predict CKD. The selected target variable is given to DNN which yields the highly accurate results. But it's complexity of cost and time can be increase withincreasing depth. Inexplicability is the major issue for DNN.

Pasadana et al.[9], has focused on the performance of all decision treeslike consolidate tree construction J48, DStump, NBTree, RT, RF etc.along with seven metrics on UCI. RF improved the accuracy after cross validating dataset ten times. RF is computationally expensive needs more power, memory and resources.

T. Chen et al.[11], proposed a novel method for sparce data along with scalable boosting tree. Boosting is an ensemble technique of DT helps to reduce sequential model errors.which helps to save time and cost.

TheoverallaccuracyofthedevelopedmodelsshowedthattheBoosting somewhat betterthan other.

Khan Bilal et al.[7], has compared NBTree, NB, J48, SVM, MLP, and (CHIRP) are applied on processed data. The results of CHIRP and NBTree are good than other.The limitations of CHIRPisnot a parametric model so the prediction can be done early and simple.

The overall accuracy of the developed models showed that the even small change of data lead to change the entire results, so the predictions are inexact which leads to ambiguity. To avoid this an ensemble model is needed to find timely immediate and exact prediction.

*Correspondence

Dr. P Manjula PriyaDarsini

M.Tech, Computer Science and Technology, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India

E-mail: manjulapriyadarshini0607@gmail.com

Proposed Work

In this proposed paperan expert ensemblemodel has created with the combination of boosting and bagging along with three models like NB, SVM and NB Tree.

Naive Bayes (NB)

NB is a probabilistic classification model comes under supervised learning. It works on Bayestheorem.So, NB is grate and wonderful option for categorization and classification. The Bayesian model is not subject to over-fitting but simpler to use. Here set of T samples D = {D_1, D_2, D_3...D_T} it is training data and each sample D is represented as T-Dimensional vector where as{A_1, A_2,A_3…A_T} correspond to the attributes. $C_{1,}$ C_2 are the two classes ckd, not ckd. NB works based on the formula given bellow

P (C|A) = P (A|C) ·P(C) / P(A)

Where P (C) = relative frequency of Cclass and the probability of event always depends on P(C).

P(A) is constant of allclasses C_1, C_2.

A = Attribute values and C = Class

Above P(A) is constant So only the product of

P(A|C).P(C) get maximized.Then the estimate class will be calculated by Assumption of conditional independence.

E= argmax P(C)$*\prod_{k=1}^{T} P(A|C)$

Where E = Estimate of class

Generally, in NB an event occurring based on the probability of a previous event.The model involves the steps:

(a)Import CKD dataset and libraries needed.

(b)After that divided the uploaded dataset into 80% train and 20% test data.

(c) Thenconvert the loaded data to frequency table.

(d) Thenprobability for frequency table have been calculated.

(e) Apply Bayes' Theorem probability to the equation.

(f) Finally, calculate estimate Class CKD or NOT which is generated by assumption.

By applying NB we got 94.8% of accuracy.

Support Vector Machine (SVM)

SVM comes under the supervisedalgorithm which is compatible with learning counts. And it is used for classification as well as for regression. It is used to categories both linear and non- linear data. To classify the linearly separable data it use a hyper plane (i.e., a straight line). For non-linearly separable data Kernel based SVM is used to convert the low dimensional data to high dimensional data.The accuracy obtained by SVM is 80.4% along with the metrics. Steps are:

Step 1:import libraries and dataset.

Step 2:separate X and Y variable.

Step 3: Divide dataset into train and test and apply SVM

Step 4: get the results for the model.

NBTree

NBTree is the combination of Naïve Bayes and decision tree which is applicable for both classification and regression. NBTree is a hybrid model developed by combining both NB and Decision tree and is a parametric model.It converts the data into tree to solve a problem, Start or root nodes will take the data as input whereeach attribute is represented by an internal node in the tree representation, and each class label is represented by a leaf node. Scaling and normalization of data is not needed in this. Steps for NBTree are: (a) we calculate the probability for each leaf node by using naïve bayes. (b) After that categorization of each leaf node will be perform. (c) The process will be repeated until there is no leaf node left. The accuracy of NBTree is 97.6% along with metrics.

Iterative Random Forest

To increase performance, learning employs a collection of models to produce a better composite model. The main concept is to group multiple “weak learners” to come up with a “strong learner”, By combining the two models bagging and boosting a new hybrid algorithm is developed called IRF. Boosting and bagging are used for both classification and prediction.An ensemble technique which divides the data into some random samples and parallelly executes call bagging. Simply the findings of numerous independent predictors are integrated using the majorityof their priority, where as inboosting method the learners are created sequentially rather than independently. The IRF algorithm is a computationally efficient approach to search for interactions of unknown form and order in high dimensional data in three simple steps. First, the RF fitting process is adaptively regularised by iterative feature re-weighting. Second, decision rules can be extracted from a weighted RF that will be from continuous or categorical to binary features. Which helps to forms the generalization of an Intersection random Tree (RIT). It is an efficient algorithm for computations that going to look for high-order interactions in binary data.

The following are steps for IRF Algorithm:

Step 1: First, load the dataset chronic kidney disease.

Step 2: Start with the selection of random samples from a given dataset.

Step 3: Bagging and boosting will apply for every sample. Then it will get the prediction result from every sample.

Step 4: Next, the iteration method is applied on predicted tree.

Step 5: In this step, will be performed for every predicted result.

Step 6: At last, select the most affected samples predicted as the final results.

The belowd epicts flow chart shows the suggested model architecture for predicting chronic kidney disease. Which has the sub-modules like pre-processing, processed data to divide the entire data to train and test.Each module demonstrates how it contributed to the predictive model's overall accuracy.

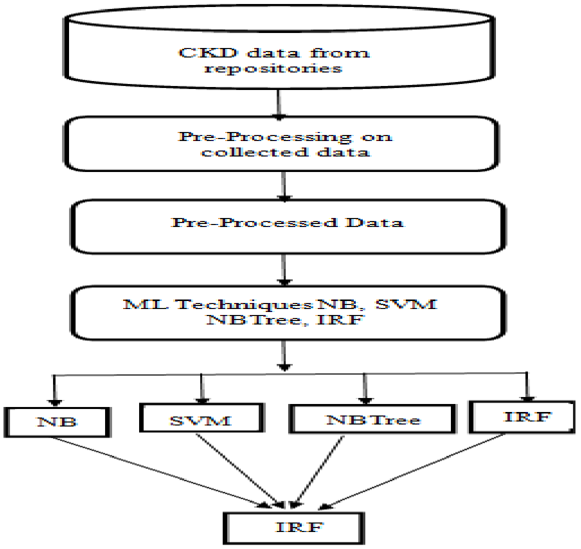
CKD data from repositories
Pre-Processing on collected data
Pre-Processed Data
ML Techniques NB, SVM NBTree, IRF
NB
SVM
NBTree
IRF
IRF

Figure 1 Work flow of proposed model

Experimental results

The proposed IRF method works on the dataset containing more than 6000 records and then analysis conducted on the dataset which has 6400 records with four different ML methods. The ages of observation

are varied from 2-90 years old. Dataset has 25 features among them 11-Numeric 13-Nominal features. we can see it contains missing values and non-numeric values and MLalgorithms will not accept such values so we need to preprocess it.

The results IRF is taken for different values are tested randomly to find how accurately the modelpredicts for each value. The proposed work can classify and predict with an accuracy of 99.2% with the metrics precision, Recall, F1-Score. The four algorithms results and their accuracies are represented in the table below:

Table4.1Performanceoffour models

Algorithm	Precision	Recall	F1-score	Accuracy in %
NB	95.7	93.02	94.8	94.8
NB-Tree	98.6	97.5	97.6	97.6
SVM	79.7	82.12	79.9	80.4
IRF	99.45	99.06	99.23	99.4

Conclusion

Existing methods for detecting CKD are limited, and most of them are based on Data Mining. ML provide remarkable results for prediction and detection of disease. The work is done on the kidney disease dataset which has 6400 records with 25 attributes.The dataset is divided as 80% for training, 20% for testing. After applying the methods, we got 94.8% accuracy for NB, for SVM 80.4%, for NBTree 97.6%, and IRF got 99.2% of accuracy.As a conclusion, IRF is the method to detect and obtained great resultswhen compared to other algorithms.

References

1. C. Webster, E. V. Nagler, ‘‘chronic kidney disease’’doi:10.1016/S0140-6736 32064-5.
2. Sobrinho, L. D. Silva,‘‘Design and evaluation of a mobile app to assist the self-monitoring of the CKD in developing countries’’jan 2019 doi:10.1186/s12911-018-0587-9.
3. A Survey on CKD Detection Using Novel MethodJayalakshmi, Lipsa Nayak,Chennai.
4. G.Murshid,T.Parvez,N.Fezal,L.``Dataminingtechniquestopredict chronic kidney disease," Apr. 2019.
5. Hashi,E.K. Zaman, Anex pert clinical decision support system to predict disease using classification techniques. (2017)doi:10.1109/ecace.2017.7912937.
6. Iliyas, IsahSaidu, “Prediction of CKD Using Deep Neural Network” Katsina State, Nigeri.
7. Khan,Naseem,(2020). An Empirical Evaluation of ML Techniques for CKDProphecy. IEEE, doi:10.1109/access.2020.2981689.
8. Levey, Andrew S; de Jong, Paul E; Coresh, “A KDIGO Controversies Conference report.Kidney International, doi:10.1038/ki.2010.483.
9. Pasadana,I.A.;Hartama,D. CKD Prediction by Using Different Decision Tree Techniques.doi:10.1088/1742- 6596/1255/1/012024 . (2019).
10. Polat, Huseyin; DanaeiMehr, Aydin Diagnosis of CKD Based on SVM by FeatureSelection Methods. doi:10.1007/s10916-017-0703 (2017).
11. T.ChenandC.Guestrin,“XGBoost:AScalableTreeBoostingSystem”22ndACM SIGKDDInt.https://arxiv.org/abs/1603.02754. (2016)
12. The global burden of kidney disease and thesustainable development goals Valerie A Luyckx,a Marcello Tonellib& John.
13. Veenita Kunwar, Khushboo “CKD analysis using data mining classification techniques” ASET,CSE Amity University Uttar Pradesh Noida,
14. V. Giannouli and N. Syrmos, ‘‘Attitudes of younger and older adults towards kidney diseases in Greece,’’ Health Psychol. Res., vol. 7, no. 2, p. 8230,2019.
15. W.G.Couser,“the contribution of CKD to the global burden of major no communicable diseases”. doi:10.1038/ki.2011.368. (2011).

Conflict of Interest: Nil Source of support: Nil