International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

A MACHINE LEARNING METHODOLOGY FOR DETECTING CHRONIC KIDNEY DISEASE

MURALE C[1], KAVINKUMAR P[2], MANOJ G[3], SURYA P[4]

[1]Department of Information Technology, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.
[2-4]Student, Department of Information Technology, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

ABSTRACT:

Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it induces other diseases. Since there are no obvious symptoms during the early stages of CKD, patients often fail to notice the disease. Early detection of CKD enables patients to receive timely treatment to ameliorate the progression of this disease. Machine learning models can effectively aid clinicians achieve this goal due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for diagnosing CKD. The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. KNN imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Missing values are usually seen in real-life medical situations because patients may miss some measurements for various reasons. After effectively filling out the incomplete data set, Five machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbor, naive Bayes classifier) were used to establish models. Also the stage of the disease is also predicted according to the age of the person.

Key Words: **Chronic kidney disease, KNN, Machine Learning, SVM, Random Forest**

1 INTRODUCTION

CHRONIC kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population. The percentage of prevalence of CKD in China is 10.8% , and the range of prevalence is 10%-15% in the United States. According to another study, this percentage has reached 14.7% in the Mexican adult general population. This disease is characterised by a slow deterioration in renal function, which eventually causes a complete loss of renal function. CKD does not show obvious symptoms in its early stages. Therefore, the disease may not be detected until the kidney loses about 25% of its function . In addition, CKD has high morbidity and mortality, with a global impact on the human body.Machine learning refers to a computer program, which calculates and deduces the information related to the task and obtains the characteristics of the corresponding pattern . This technology can achieve accurate and economical diagnoses of diseases; hence, it might be a promising method for diagnosing CKD.We used KNN imputation to fill in the missing values in the data set, which could be applied to the data set with the diagnostic categories are unknown.Logistic regression (LOG), RF, SVM, KNN, naive Bayes classifier (NB) were used to establish CKD diagnostic models on the complete CKD data sets.KNN imputation is used to fill in the missing values. To our knowledge, this is the first time that KNN imputation has been used for the diagnosis of CKD. In addition, building an integrated model is also a good way to improve the performance of separate individual models. The proposed methodology might effectively deal with the scene where patients are missing certain measurements before being diagnosed.

2 RELATED WORK

Chen et al. used k-nearest neighbor (KNN), support vector machine (SVM) and soft independent modelling of class analogy to diagnose CKD, KNN and SVM achieved the highest accuracy of 99.7% [1]. In addition, they used fuzzy rule-building expert system, fuzzy optimal associative memory and partial least squares discriminant analysis to diagnose CKD, and the range of accuracy in those models was 95.5%-99.6% [2]. Their studies have achieved good results in the diagnosis of CKD. In the above models, the mean imputation is used to fill in the

missing values and it depends on the diagnostic categories of the samples. As a result, their method could not be used when the diagnostic results of the samples are unknown. In reality, patients might miss some measurements for various reasons before diagnosing. In addition, for missing values in categorical variables, data obtained using mean imputation might have a large deviation from the actual values. For example, for variables with only two categories, we set the categories to 0 and 1, but the mean of the variables might be between 0 and 1. Polat et al. developed an SVM based on feature selection technology, the proposed models reduced the computational cost through feature selection, and the range of accuracy in those models was from 97.75%-98.5% [3]. J. Aljaaf et al. used novel multiple imputation to fill in the missing values, and then MLP neural network (MLP) achieved an accuracy of 98.1% [4]. Subas et al. used MLP, SVM, KNN, C4.5 decision tree and random forest (RF) to diagnose CKD, and the RF achieved an accuracy of 100% [5]. In the models established by Boukenze et al., MLP achieved the highest accuracy of 99.75% [6]. The studies of [2], [6] focus mainly on the establishment of models and achieve an ideal result. However, a complete process of filling in the missing values is not described in detail, and no feature selection technology is used to select predictors as well. Almansour et al. used SVM and neural network to diagnose CKD, and the accuracy of the models was 97.75% and 99.75%, respectively [7]. In the models established by Gunarathne et al., zero was used to fill out the missing values and decision forest achieved the best performance with the accuracy was 99.1% [8].

3 PRELIMINARIES

A. DATA DESCRIPTION AND OPERATING ENVIRONMENT

The CKD data set used in this study was obtained from the UCI machine learning repository . The data set contains 400 samples. In this CKD data set, each sample has 24 predictive variables or features (11 numerical variables and 13 categorical (nominal) variables) and a categorical response variable (class). Each class has two values, namely, ckd (sample with CKD) and notckd (sample without CKD). In the 400 samples, 250 samples belong to the category of ckd, whereas 150 samples belong to the category of notckd. It is worth mentioning that there is a large number of missing values in the data.

B. DATA PREPROCESSING

Each categorical (nominal) variable was coded to facilitate the processing in a computer. For the values of rbc and pc, normal and abnormal , pcc and ba, present and notpresent were coded as 1 and 0, respectively. For the values of htn, dm, cad, pe and ane, yes and no were coded as 1 and 0, respectively. For the value of appet, good and poor were coded as 1 and 0, respectively. Although the original data description defines three variables sg, al and su as categorical types, the values of these three variables are still numeric based, thus these variables were treated as numeric variables. After encoding the categorical variables, the missing values in the original CKD data set were processed and filled at first. KNN imputation was used in this study, and it selects the K complete samples with the shortest Euclidean distance for each sample with missing values. The values of K in this work were chosen as 3, 5, 7, 9 and 11. As a result, five complete CKD data sets were generated. In addition, we also proved the effectiveness of KNN imputation by comparing it with One other method. It is to use mean and mode of the corresponding variables to fill in missing values of continuous and categorical variables.

Variables	Explain	Class	Scale	Missing Rate
age	Age	Numerical	age in years	2.25%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020,1.025)	11.75%
al	Albumin	Nominal	(0,1,2,3,4,5)	11.5%
su	Sugar	Nominal	(0,1,2,3,4,5)	12.25%
rbc	Red Blood Cells	Nominal	(normal,abnormal)	38%
pc	Pus Cell	Nominal	(normal,abnormal)	16.25%
pcc	Pus Cell clumps	Nominal	(present,notpresent)	1%
ba	Bacteria	Nominal	(present,notpresent)	1%
bgr	Blood Glucose Random	Numerical	in mgs/dl	11%
bu	Blood Urea	Numerical	in mgs/dl	4.75%
sc	Serum Creatinine	Numerical	in mgs/dl	4.25%
sod	Sodium	Numerical	in mEq/L	21.75%
pot	Potassium	Numerical	in mEq/L	22%
hemo	Hemoglobin	Numerical	in gms	13%
pcv	Packed Cell Volume	Numerical	-	17.75%
wbcc	White Blood Cell Count	Numerical	in cells/cumm	26.5%
rbcc	Red Blood Cell Count	Numerical	in millions/cmm	32.75%
htn	Hypertension	Nominal	(yes,no)	0.5%
dm	Diabetes Mellitus	Nominal	(yes,no)	0.5%
cad	Coronary Artery Disease	Nominal	(yes,no)	0.5%
appet	appet	Nominal	(good,poor)	0.25%
pe	Pedal Edema	Nominal	(yes,no)	0.25%
ane	Anemia	Nominal	(yes,no)	0.25%
class	Class	Nominal	(ckd,notckd)	0%

FIG1.DETAILS OF THE COLUMNS IN DATASET

C. FEATURE EXTRACTION

Extracting feature vectors or predictors could remove variables that are neither useful for prediction nor related to response variables and thus prevent these unrelated variables from interfering with the model construction, which causes the models to make an accurate prediction . Herein, we used RF to extract the variables that are most meaningful to the prediction. RF detects the contribution of each variable to the reduction in the Gini index. The larger the Gini index, the higher the uncertainty in classifying the samples. The step of feature extraction was run on each complete data set. when the RF was used to extract the variables, all variables were selected expect pcc, ba , pe , ane and cad.

D. ESTABLISHING AND EVALUATING INDIVIDUAL MODELS

The following machine learning models have been obtained by using the corresponding subset of features or predictors on the complete CKD data sets for diagnosing CKD.

1) Regression-based model: LOG

2) Tree-based model: RF

3) Decision plane-based model: SVM

4) Distance-based model: KNN

5) Probability-based model: NB

1) The output of LOG was the probability that the sample belongs to notckd, and the threshold was set to 0.5.

2) RF was established using all variables. Default 500 trees is used. The RF was established and evaluated on the data sets obtained by KNN imputation.

3) The models of SVM were generated by using the RBF kernel function, where γ was set to [0.1, 0.5, 1, 2, 3, 4]. Parameter C represents the weight of misjudgment loss, and it was set to [0.5, 1, 2, 3].

4) For the NB, the value of Laplace was equal to 1.

5) For the KNN, the nearest neighbor parameter was set to [5].

CLASSIFIERS	ACCURACY
LOGISTIC REGRESSION	100%
KNN CLASSIFIER	100%
RANDOM FOREST	100%
NAÏVE BAYES	60%
SUPPORT VECTOR MACHINE	97.25%

TABLE 1.ACCURACY OF VARIOUS CLASSFIERS

E. ESTABLISHING THE INTEGRATED MODEL

LOG , SVM and RF are selected as underlying components to generate the integrated model to improve the performance of judging. The probabilities that each sample was judged as notckd in LOG and RF and SVM are used as the outputs of underlying components. The ensemble classifer is used to combine these three models.

F. PREDICTING THE STAGE OF THE DISEASE

The Stage of the disease is predicted using a dataset that contains two columns age and stage where age is used to predict the stage of the disease. The dataset is constructed using age column from which eGFR value is calculated that determines the stage of the disease.

4 RESULT AND DISCUSSION

Our results show the feasibility of the proposed methodology. By the use of KNN imputation, LOG, RF, SVM and KNN could achieve better performance than the cases when the random imputation and mean and mode imputation were used. KNN imputation could fill in the missing values in the data set for the cases wherein the diagnostic categories are unknown, which is closer to the real-life medical situation. The LOG achieved an accuracy of around 100%, KNN and RF also achieved an accuracy of 100%. Therefore, an integrated model combining LOG and RF and SVM was established to improve the performance of the component models. An integrated model is then established to improve the performance of the classifier. In this study, we used euclidean distance to evaluate the similarity between samples, and KNN could obtain a good result based on euclidean distance with the accuracy of 100%.

5 CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. The severity of the disease is found but actual severity calculation requires gender and race of the persons from which eGFR value is calculated that identifies the

stage of the disease. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the exact severity of the disease.

6 ACKNOWLEDGEMENT

First and Foremost, we are thankful to Coimbatore Institute of Technology, Department of Information Technology and Mr.C.Murale, Assistant Professor, Information Technology Coimbatore Institute of Technology. A special word of gratitude to Dr. NK.Karthikeyan , Head of Department, Information Technology Coimbatore Institute of Technology, for his continued guidance and support for our project work.

7 REFERENCES

1] Z. Chen et al., "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.

[2] A. Subasi, E. Alickovic, J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.

[3] L. Zhang et al., "Prevalence of chronic kidney disease in china: a crosssectional survey," Lancet, vol. 379, pp. 815-822, Aug. 2012.

[4] A. Singh et al., "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015.

[5] A. M. Cueto-Manzano et al., "Prevalence of chronic kidney disease in an adult population," Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.

[6] H. Polat, H.D. Mehr, A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, no. 4, Apr. 2017.

[7] C. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis," Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

[8] V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," Am. J. Med., vol. 130, no. 12, Dec. 2017.

9] Z. Chen, X. Zhang, Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," Int. Urol. Nephrol., vol. 48, no. 12, pp. 2069-2075, Dec. 2016.

[10] A. J. Aljaaf et al., "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in Proc. IEEE Congr. Evolutionary Computation, Jul. 2018.

[11] B. Boukenze, A. Haqiq and H. Mousannif, "Predicting chronic kidney failure disease using data mining techniques," in Proc. Int. Symp. Ubiquitous Networking, Nov. 2016, pp. 701-712.

[12] N. Almansour et al., "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Comput. Biol. Med., vol. 109, pp. 101-111, Jun. 2019.

[13] W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Proc. IEEE 17th Int. Conf. Bioinformatics and Bioengineering, Oct. 2017, pp. 291-296.

[14] D. Dua and C. Graff, "UCI Machine Learning Repository," Irvine, University of California, School of Information and Computer Sciences, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[15] D. Ichikawa et al., "How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach," J. Biomed. Inform., vol. 64, pp. 20-24, Dec. 2016.

[16] L. N. Sanchez-Pinto, L. R. Venable, J. Fahrenbach, M. M. Churpek, "Comparison of variable selection methods for clinical predictive modeling," Int. J. Med. Inform., vol. 116, pp. 10-17, Aug. 2018.