Engineering and Applied Science Letters

PSR Press

Article

Predictive analysis of chronic kidney disease based on machine learning

Huan You

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing, Jiangsu Province, China; hyoujn@163.com

Received: 21 January 2021; Accepted: 27 February 2021; Published: 11 March 2021.

Abstract: The purpose of this study is to explore the influence of factors on patients with chronic kidney disease (CKD) and to establish predictive models using machine learning methods. Data were collected from the Affiliated Hospital of Nanjing University of Chinese Medicine between January 2016 and December 2017, including 69 CKD patients and 155 healthy subjects. This study found that carotid intima-media thickness (cIMT) is the most important indicator among the top 9 important features of each model. In order to find the best model to diagnosis CKD, extreme gradient boosting (XGBoost), support vector machine (SVM) and logistic regression are established and XGBoost is the most suitable model for this study (accuracy, 0.93; specificity, 0.89; sensitivity, 0.94; F1 score, 0.91; AUC, 0.99).

Keywords: Chronic kidney disease, machine learning, prediction.

1. Introduction

The incidence of chronic kidney disease (CKD) has been increasing every year and has become a global health issue. It affects 8-16% of the world's population, especially in developing countries [1]. CKD is a kind of comprehensive kidney disease with progressive deterioration of renal function and systemic lesions. It was characterized by high prevalence, high mortality, low diagnosis and low awareness [2]. In 2002, the National Kidney Foundation (NKF) developed guidelines for CKD [3]. The criteria for this disease were as follows;

(1) The glomerular filtration rate (GFR) was less than $60mL/min$ per $1.73m^2$ for more than 3 consecutive months;
(2) Abnormal structure or function of the kidney was caused by various factors for more than 3 months;
(3) There were pathological abnormalities with or without a decrease in GFR, abnormal signs of kidney damage, or abnormalities in imaging.

Machine learning algorithm enables machines to learn from massive knowledge to the behavior rules and thinking patterns that are similar to human beings [4]. In recent years, the application frequency of machine learning algorithms in the medical field is increasing rapidly and the research depth is deepening continuously. At present, many scholars established correlation analysis and prediction models for some diseases, and the results were remarkable [5]. This study aims to compare the applicability of different algorithms to the prediction of CKD and to measure the effect of related indicators on CKD.

2. Materials and methods

2.1. Study population

This study was conducted in the Affiliated Hospital of Nanjing University of Chinese Medicine between January 2016 and December 2017. According to whether the patients had CKD, we divided them into two groups (CKD group and control group). We screened a total of 224 patients, 69 of whom were diagnosed with CKD. We selected basic clinical characteristics, laboratory findings and imaging characteristics, including a total of 23 indicators. All patients obtained written informed consent.

2.2. Statistics analysis

We used Python software (version 3.7; https://www.python.org) for statistics analysis. Continuous variables were represented by mean standard deviation, and discrete variables were represented by numbers. All features with a missing rate greater than 70% were selected to be excluded. Missing values were filled by means or modes. In statistical analysis, continuous variables and discrete variables used t test and chi-square test respectively. P value was considered statistically significant.

2.3. Model construction

Three commonly used machine learning classification algorithms were selected for modeling to predict CKD. They are extreme gradient boosting (XGBoost) [6], support vector machine (SVM) [7], and logistic regression [8]. XGBoost is a type of tree ensemble model. Its main idea is to continuously add trees and grow subtrees through feature splitting. Each time a subtree is added, a new function was learned, and the prediction results of these subtrees were used by the additive model. Adding together can continuously improve the accuracy of the model, realize the fitting of the residuals, and then accurately predict the results. Its objective function was calculated by the following formula:

$$Obj(\ominus) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where n is the number of samples, y_i is the true value of the ith sample, $\hat{y}_i$ is the predictive value of the ith sample, $l(y_i, \hat{y}_i)$ is the error function of the model and $\Omega(f_k)$ is the regularization function of the model. The SVM algorithm uses the ideas of finding the maximum interval and projecting to higher dimensions to find a hyperplane with good data classification effect to realize the classification of data. Its model was solved based on the following formula:

$$min_{w,b} max_{a \geq 0} \frac{1}{2} ||w||^2 - \sum_{i=1}^{n} \alpha_i (y_i (w^T X_i + b) + 1),$$

where w and b are the hyperplane parameter to be solved, α is langrangian, (X_i, Y_i) is the ith sample and n is the number of samples. Because SVM algorithm has a very good performance in dealing with nonlinear classification problems, so this model in medicine can be used to diagnose some diseases. Logical regression model is a very simple classification model. It introduces sigmoid function on the basis of regression model, realizes the mapping from predicted value to probability, and then can be used to deal with classification problems. The training model was calculated based on the following formula:

$$P_{y_i=1} = \frac{1}{1 + e^{-W^T + X}},$$

where ω is the model coefficient to be solved, (X_i, Y_i) is the ith sample. Although the logistic regression model has a simple structure, it is often used for its satisfactory effect on general problems and good explanation.

In the diagnosis of CKD, firstly, all 224 samples were divided into training set and test set according to the ratio of 80% and 20% respectively. The training set was used to fit and establish the above three models, while the test set was used to evaluate the effects of the established models. According to the evaluation results, the best performing model was selected among the three models as the machine learning diagnosis model for CKD.

3. Results

3.1. Clinical characteristics

According to statistical tests, 16 variables were considered statistically significant in the two groups. They were age, weight, BMI, haemoglobin, fasting blood glucose, serum creatinine, LDL-C, HDL-C, triglyceride, SBP, DBP, uric acid, leukocyte count, cIMT, PWV-ES, GFR. Summary of clinical characteristics is shown in Table 1.

Table 1. Summary of variables

Variable	CKD group	Control group	P value
Gender			
Male	37	62	
Female	33	93	0.087
Age (year)	55.768± 14.009	43.403±12.895	<0.001
Height (cm)	166.841±7.455	165.391±7.558	0.185
Weight (kg)	68.191±12.282	59.830±8.443	<0.001
BMI (kg/m^2)	24.471±4.069	21.826±2.255	<0.001
Haemoglobin, g/l	121.899±23.233	139.572±14.606	<0.001
Fasting blood glucose, mmol/L	5.768±1.343	4.876±0.692	<0.001
Serum creatinine, $\mu mol/L$	156.437±148.807	67.087±15.270	<0.001
Total cholesterol, mmol/L	5.324±2.722	4.607±0.802	0.065
LDL-C, mmol/L	2.695±0.830	2.460±0.559	0.035
HDL-C,mmol/L	1.244±0.374	1.525±0.397	<0.001
Blood pressure, mmHG			
SBP	135.217±18.189	117.628±14.399	<0.001
DBP	79.116±10.657	71.662±9.179	<0.001
Uric acid, $\mu mol/L$	396.259±146.634	278.394 ±77.579	<0.001
Leukosyte Count , $10^9/L$	6.362±2.196	5.700± 1.382	0.023
Erythrocyte Count, $10^{12}/L$	3.924±0.810	5.483±9.947	0.195
Alanine aminotransferase, μ/L	25.391±16.038	23.086±19.991	0.400
Aspartate aminotransferase, μ/L	23.493±12.830	22.314±11.286	0.491
cIMT, cm	0.062±0.030	0.045±0.002	<0.001
PWV-BS, m/s	6.401±1.465	6.105±1.140	0.140
PWV-ES, m/s	9.079±2.033	7.184±2.034	<0.001
GFR, $mL/min/m^3$	89.616±54.372	138.334±33.860	<0.001

Data were numbers or mean value standard deviation. BMI, Body mass index; LDL-C, Low-density lipoprotein-cholesterol; HDL-C, High-density lipoprotein-cholesterol; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; cIMT, Carotid intima-media thickness; PWV-BS, Pulse wave velocity-beginning of systole; PWV-ES, Pulse wave velocity-end of systole; GFR, Glomerular filtration rate.

3.2. GFR in patients

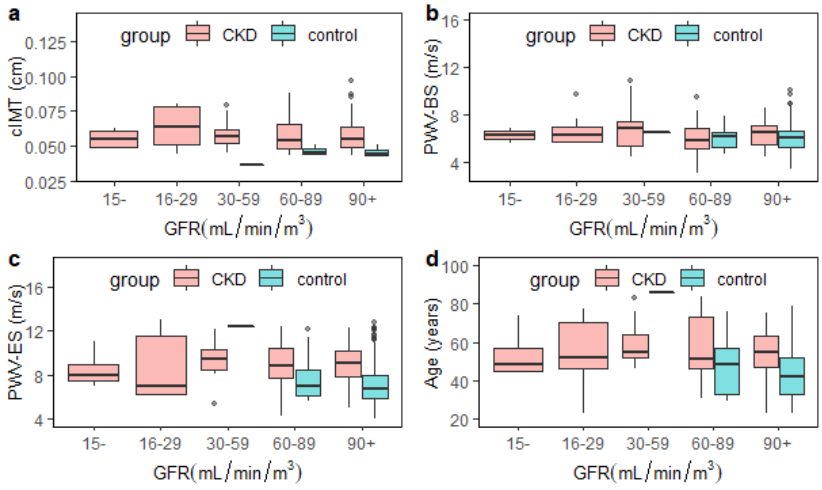

Figure 1. Box plot shows relationship of (a) carotid intima-media thickness (cIMT), (b) pulse wave velocity-beginning of systole (PWV-BS), (c) pulse wave velocity-end of systole (PWV-ES), (d) age with glomerular filtration rate (GFR).

GFR is an important indicator of CKD disease. According to the disease guidelines for CKD, patients are subdivided with GFR into five subgroups as follows' term:

1. $15mL/min/m^3$,
2. $16-29mL/min/m^3$,
3. $30-59mL/min/m^3$,
4. $60-89mL/min/m^3$,
5. $\geq 90mL/min/m^3$.

Patients with a GFR value less than 60 typically had CKD. The overall levels of cIMT, PWV-ES and age in CKD group were higher than those in control group (Figure 1). However, there was no significant difference in PWV-BS values between the two groups (Figure 1).

3.3. Important features in the model

There are many factors that affect CKD, but there is still a wide variation in the importance of features. In this study, XGBoost, SVM and logistic regression were respectively used to analyze the importance of features. The importance of 16 features in different models is displayed in Figure 2. According to the results of the three models, cIMT was the most important among features. It was considered that the common feature among the top 9 important features of the three models respectively are the features that have a relatively large influence on CKD. Through the intersection processing, factors that have a greater impact on CKD are age, GFR, haemoglobin, LDL-C, uric acid, and cIMT. Figure 3 is the analysis result of the correlation among important variables. It can be found that the correlation between GFR and uric acid is -0.557, which indicated that there might be some correlation between them.

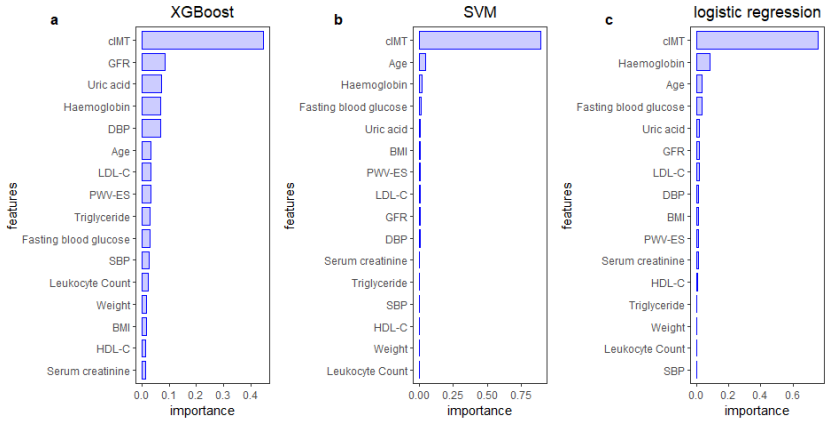

Figure 2. Feature importance analysis chart through (a) XGBooost, (b) SVM and (c) logistic regression.

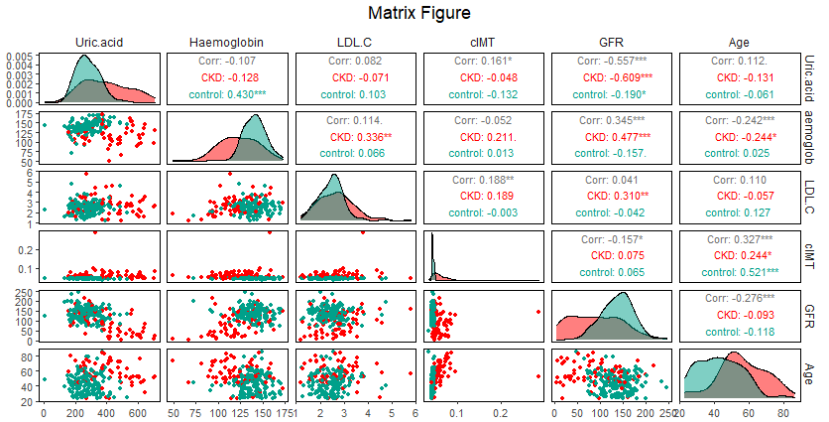

Figure 3. Matrix scatter plot among uric acid, haemoglobin, low-density lipoprotein-cholesterol (LDL-C), carotid intima-media thickness (cIMT), glomerular filtration rate (GFR) and age.

3.4. Model comparison

After modeling analysis, the performance of the three models on the test set is shown in Figure 4. In general, XGBoost is significantly better than other models (accuracy, 0.93; specificity, 0.89; sensitivity, 0.94; F1 score, 0.91; AUC, 0.99). The ROC curves of three models are shown in Figure 5. It can be seen that the effect of the logic effect is the worst, and the XGBoost effect is the best among three models.

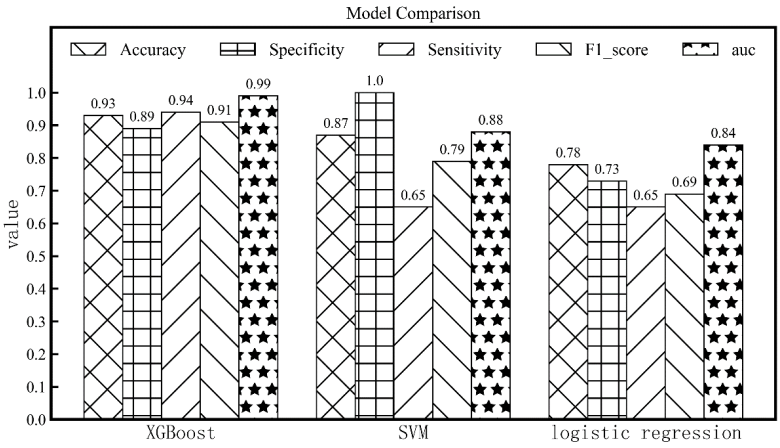

Figure 4. Model comparison among XGBoost, SVM and logistic regression according to accuracy, specificity, sensitivity, F1 score and AUC.

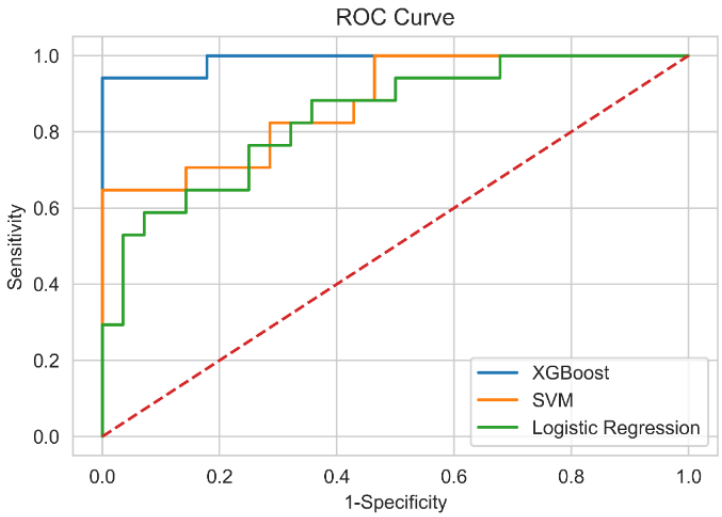

Figure 5. Roc curves among XGBoost, SVM and logistic regression.

4. Discussion

In this study, we established three CKD prediction models which are XGBoost, SVM and logistic regression. According to these three models, the ranking of feature importance under each model was analyzed. In order to measure the prediction effect of each model, we put forward four evaluation indicators (accuracy, specificity, sensitivity, F1 score and AUC) and ROC curve.

In previous studies, some scholars used artificial neural network, decision trees, and logistic regression to predict the survival of kidney dialysis [9]. A study used support vector machine and artificial neural network to construct predictive models of kidney disease. Its experimental results showed that the performance of artificial neural network is better than other algorithms, and it could obtain better accuracy and performance [10]. Baby and Vital [11] used AD trees, J48, Kstar, Naïve Bayes and Random forest in the prediction model of kidney disease and they found that the best methods were Kstar and Random forest. Although other studies

separately developed methods with strong applicability for CKD, this study comprehensively considered the effect of the model, the interpretability and the applicability of the data. Therefore, one of each type of machine learning algorithm selected in this paper is more comprehensive and the selected model could not only predict disease but also evaluate feature importance. In current study, according to the effect of three models (XGBoost, SVM and logistic regression), XGBoost had the best performance for the prediction of CKD.

The study had several advantages; Firstly, applying machine learning model to the analysis of medical problems could promote the progress of automatic disease diagnosis of CKD. Secondly, using machine learning methods to extract important features, it is possible to find indicators that has a relatively large impact on CKD from a data perspective. Thirdly, tree model, SVM model and logistic regression model are respectively used to diagnose and predict CKD. These three algorithms covered most models with relatively good explanatory ability and could comprehensively analyze the performance of machine learning algorithms in CKD diagnosis.

There were some limitations; Firstly, the amount of data used in modeling was relatively small, so the model established might be insufficient in generalization ability, and more data should be used to further verify the effectiveness of the model. Secondly, the extracted important features had not been further combined with the medical feasibility analysis, and the analyzed indicators might be uncontrollable, so the feasibility of the treatment plan needed to be further considered. In addition, since the "black box" model could not measure the importance of each feature, some "black box" models (such as artificial neural network) were not analyzed and compared. These "black box" models might had better effects, but there is no further discussion here.

5. Conclusion

Through the establishment of three predictive models, XGBoost is the most suitable for the diagnosis of CKD. In the feature importance analysis of the three models, cIMT was found to be a strong predictor of CKD. Compared with PWV-BS, PWV-ES has more important effect on CKD and stronger correlation with GFR.

Conflicts of Interest: "The author declares no conflict of interest."

References

[1] Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B. & Yang, C. W. (2013). Chronic kidney disease: global dimension and perspectives. *The Lancet, 382*(9888), 260-272.

[2] Ali, S., Dave, N., Virani, S. S., & Navaneethan, S. D. (2019). Primary and secondary prevention of cardiovascular disease in patients with chronic kidney disease. *Current Atherosclerosis Reports, 21*(9), 1-9.

[3] Levey, A. S., Coresh, J., Bolton, K., Culleton, B., Harvey, K. S., Ikizler, T. A. & Briggs, J. (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases, 39*(2 SUPPL. 1), i-ii+.

[4] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

[5] Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In 2017 *IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 291-296). IEEE.

[6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[7] Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology, 24*(12), 1565-1567.

[8] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics, 9*(4), 705-724.

[9] Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology, 7*(1), 242.

[10] Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR), 6*(2), 1-12.

[11] Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. *International Journal of Engineering Research and Technology, 4*(7), 206–210.