Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2021;10(2):600-4

Classification of chronic kidney failure by applying different tree-based methods on a medical data set

Emek Guldogan, Zeynep Kucukacali

Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

Received 03 May 2021; Accepted 07 May 2021
Available online 11.05.2021 with doi: 10.5455/medscience.2021.05.153

Abstract

The purpose of this study is to classify chronic kidney failure (CKF) by applying different tree-based methods on the open-access CKF data set and to compare the performance of the methods used. Classification models will be created using decision trees, J48, Random Forest, and Gradient Boosted Trees from tree-based methods used in the study were applied to an open-access data set named "Chronic Kidney Disease". There are 400 patients in the data set used, 250 (62.5%) of these patients have chronic kidney failure. Different tree-based methods were implemented to classify chronic kidney failure. Among the 4 different tree-based classification models used, the model with the best classification metrics is the Random Forest model, and other models have also yielded successful results. As a result, very successful results were obtained in the study performed with the classification methods used and the chronic renal failure data set. Each model was able to classify the data with high classification performance.

Keywords: Machine learning, classification, chronic kidney failure, performance comparison

Introduction

Chronic kidney failure (CKF), which has emerged as a major public health concern around the world and in our own country, is a disorder that can develop for a variety of reasons, results in permanent kidney function loss, adversely impact people's quality of life. and necessitates lifelong treatment and follow-up [1]. The incidence of CKF is increasing rapidly nowadays, according to reports. Chronic kidney failure (CKF) is becoming a more common health condition around the world. When viewed from a prognostic standpoint, this disorder, which is very costly to treat, may have bad consequences. The development of kidney failure, acute and chronic complications due to renal dysfunction, cardiovascular mortality, and morbidity are the most serious effects [2].

Machine learning, one of the data mining techniques, is a sub-field of artificial intelligence that uses data-based learning to make predictions about new data when it is exposed to it. Machine learning systems seek to either remove the need for human intuition entirely or obtain the ability to make decisions through human-machine collaboration [3]. Classification is a supervised learning technique that classifies data according to a predetermined class label. The purpose of classification is to create a kind of model that can be applied to classify unclassified data [4]. Various methods based on statistics and machine learning have been developed for the classification process. In this study, classification models will be created using decision trees, J48, random forest, and Gradient Boosted Trees from tree-based classification methods based on machine learning principles. In classification problems, decision trees are one of the most commonly used approaches. In comparison to other approaches, decision trees are simpler to build, understand, and interpret. Another advantage of decision trees is that they generate good models in addition to these. In the decision trees model, a tree is built from the data we have, the records in the dataset are transferred to this tree, and the records are classified based on the outcome [5]. J48 is a decision tree algorithm based on the very popular C4.5 algorithm developed by J. Ross Quinlan [6]. J48 Algorithm, based on Information Gain Theory, can select relevant properties from data in an automated process. It's an iterative algorithm that divides samples based on where they obtain

***Corresponding Author:** Emek Guldogan, Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey,
E-mail: emek.guldogan@inonu.edu.tr

the information gain [7]. Breiman suggested the Random Forest (RF) approach in 2001 by introducing the Bagging method, which entails combining the decisions of several, multivariate trees. each trained with a different set of training data. rather than producing a single decision tree. Thus, the idea developed that more successful results could be obtained with many trees instead of one tree [8]. Gradient Boosting is a powerful machine learning technique. Gradient Boosting is based on boosting techniques. They are often used in conjunction with Gradient Boosting decision trees and are therefore called Gradient Boosted Trees [9, 10].

The purpose of this study is to classify CKF by applying different tree-based methods on the open-access CKF data set and to compare the performance of the methods used.

Material and Methods

Dataset

The methods used in the study were applied to an open-access data set named "Chronic Kidney Disease". The data set was obtained from https://www.kaggle.com/abhia1999/chronic-kidney-disease (11). There are 400 patients in the data set used. 250 (62.5%) of these patients have chronic kidney failure. Explanations about the variables and their properties in the data set are given in Table 1.

Table 1. Explanatıons About The Variables In The Dataset And Their Properties

Variable	Variable Description	Variable Type	Variable Role
Bp	Blood Pressure	Quantitative	Predictor
Sg	Specific Gravity	Quantitative	Predictor
Al	Albumin	Qualitative	Predictor
Su	Sugar	Qualitative	Predictor
Rbc	Red Blood Cell	Qualitative	Predictor
Bu	Blood Urea	Quantitative	Predictor
Sc	Serum Creatinine	Quantitative	Predictor
Sod	Sodium	Quantitative	Predictor
Pot	Pottasium	Quantitative	Predictor
Hemo	Hemoglobin	Quantitative	Predictor
Wbcc	White Blood Cell Co-unt	Quantitative	Predictor
Rbcc	Red Blood Cell Count	Quantitative	Predictor
Htn	Hypertension	Qualitative	Predictor
Class	Predicted Class	Qualitative	Output

Tree-Based Classification Methods

Decision Trees

One of the most common and efficient methods of knowledge discovery and data mining is deci-sion trees, which is one of the prediction methods. The rules in the data are shown in a hierar-chical and organized manner using decision trees. Decision trees are a visual modeling approach that presents the decision choices and probabilistic scenarios in a specific order by sorting and presenting the mass of knowledge about the problem faced by the decision-maker more under-standably. In this sense, decision trees can be thought of as a hierarchical model that incorporates both decisions and outcomes [12].

J48

Quinlan's J48 decision tree is a C4.5 decision tree designed for nonlinear and small data classifi-cation, J48 is a decision tree that classifies using entropy principle information. Quinlan's C4.5 algorithms is used to build a pruned C4.5 tree. To make decisions, subsets of each attribute da-taset are examined for entropy differences [13,14].

Random forest

The aim of the classifier in this algorithm, introduced by Breiman in 2001(8), is to combine the decisions of multiple trees, each trained in different training sets, rather than generating a single decision tree. While creating decision trees, when determining the attribute at each level, firstly, some calculations are made in all trees and the attribute is determined, then the attributes in other trees are combined and the most used attribute is selected. After the selected attribute is included in the tree, the same processes are repeated at other levels [15].

Gradient boosted trees (GBT)

The basic idea of the gradient boosting tree is combining a series of weak base classifiers into a strong one. It's a kind of ensemble learning that can be used to solve regression and classification problems. Leo Breiman developed the concept of gradient boosting. The approach is typically used with decision trees of a fixed size as base learners, and, in this context, is called gradient tree boosting. Gradient boosting is made up of three parts: loss function, weak learner and additive model [16].

Performance evaluation criteria

The classification matrix for the calculation of performance metrics is given in Table 2.

Table 2. Confusion matrix for calculating performance metrics

		Real		
		Positive	**Negative**	**Total**
Predicted	**Positive**	True positive (TP)	False positive (FP)	**TP+FP**
	Negative	False negative (FN)	True negative (TN)	**FN+TN**
	Total	**TP+FN**	**FP+TN**	**TP+TN+FP+FN**

Data analysis

Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. In terms of input variables, the existence of a statistically significant difference and the relationship between the categories of the output variable, "ckd" and "notckd" groups, were examined using Mann-Whitney U, Pearson Chi-square test, and Yates's correction chi-square test. $p<0.05$ values were considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

Results

Descriptive statistics related to the target variable examined are presented in Table 3 and Table 4. There is a statistically significant difference between the dependent variable classes in terms of other variables other than the “Pot” variable.

In this study, the metrics of the classification performance of the decision trees, J48, Random forest, and gradient boosted trees methods, which are among the tree-based methods used to classify the CKF dataset, are given in Table 5. below.

Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value ob-tained from the decision trees model were 96.25%, 95.33%, 96.80%, 95.14%, and 97.36% respec-tively. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value obtained from the J48 model were 97.75%, 96.00%, 98.00%, 98.08% and 97.71% respectively. Accuracy, sensitivity, specificity, positive predictive value and negative predictive value obtained from the Random forest model were 99.25%, 98.67%, 99.60%, 99.38%, and 99.26% respectively. Finally, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value obtained from the gradient boosted trees model were 98.00%, 97.33%, 98.40%, 97.46%, and 98.47% respectively.

Table 3. Descriptive statistics for Quantitative Input variables

Variables	Predicted Class		p* value
	Not-ckd	ckd	
	Median (min-max)	Median (min-max)	
Bp	70 (60-80)	80 (50-180)	<0.001*
Sg	1.02 (1.02-1.03)	1.02 (1.01-1.03)	<0.001*
Bu	33.5 (10-57)	55 (1.5-391)	<0.001*
Sc	0.9 (0.4-3.07)	2.45 (0.5-76)	<0.001*
Sod	141 (135-150)	137.53 (4.5-163)	<0.001*
Pot	4.5 (3.3-5)	4.63 (2.5-47)	0.515
Hemo	15 (12.53-17.8)	11.3 (3.1-16.1)	<0.001*
Wbcc	7750 (4300-11000)	8406 (2200-26400)	<0.001*
Rbcc	5.25 (4.4-6.5)	4.71 (2.1-8)	<0.001*

* Mann Whitney U test

Table 4. Descriptive statistics for quantitative input variables

Variables		Predicted Class		p-value
		Not ckd	**ckd**	
Al	0	145 (96.7%)	54 (21.6%)	<0.001*
	1	5 (3.3%)	85 (34.0%)	
	2	0 (0%)	43 (17.2%)	
	3	0 (0%)	43 (17.2%)	
	4	0 (0%)	24 (9.6%)	
	5	0 (0%)	1(0.4%)	
Su	0	150 (100%)	189 (75.6%)	<0.001*
	1	0 (0%)	13 (5.2%)	
	2	0 (0%)	18 (7.2%)	
	3	0 (0%)	14 (5.6%)	
	4	0 (0%)	13 (5.2%)	
	5	0 (0%)	3 (1.2%)	
Rbc	0	0 (0%)	47 (18.8%)	<0.001**
	1	150 (100%)	203 (81.2%)	
Htn	0	150 (100%)	103 (41.2%)	<0.001**
	1	0 (0%)	147 (58.8%)	

* Pearson chi-square test; ** Yates's correction chi-square test

Table 5. Classification matrices for decision trees, J48, random forest, and gradient boosted trees

Models	Metric	Value (%)
Decision trees	Accuracy	96.25
	Sensitivity	95.33
	Specificity	96.80
	Positive predictive value	95.14
	Negative predictive value	97.36
J48	Accuracy	97.75
	Sensitivity	96.00
	Specificity	98.00
	Positive predictive value	98.08
	Negative predictive value	97.71
Random Forest	Accuracy	99.25
	Sensitivity	98.67
	Specificity	99.60
	Positive predictive value	99.38
	Negative predictive value	99.26
Gradient boosted trees	Accuracy	98.00
	Sensitivity	97.33
	Specificity	98.40
	Positive predictive value	97.46
	Negative predictive value	98.47

Discussion

Chronic kidney failure (CKF) is an important public health problem with increasing frequency in the world and our country. CKF is an important health problem that is chronic and progressive impairment in the fluid-electrolyte balance, endocrine and metabolic functions of the kidney, increased mortality, and decreased quality of life. Similar findings have been found in population-based studies investigating the prevalence of CKF around the world and in our own country. Owing to its high morbidity rate and increased health costs, CKF is considered a major public health issue around the world. Therefore, it is an open area for research and new developments [17,18].

By learning the pattern in the data stack, machine learning methods perform classification and estimation. In recent years, machine learning has advanced at a breakneck rate. In recent years, machine learning approaches have been one of the tools used in disease detection and clinical decision support systems years [19].

For chronic kidney disease, a paper introduces the Density-dependent Feature Selection (DFS) with Ant Colony based Optimization (D-ACO) algorithm, which is an intelligent prediction and classification method for healthcare (CKD). When the D-ACO algorithm is compared to existing methods, the presented intelligent system outperforms them [20]. Another paper used a variety of machine learning algorithms to solve a problem in medical diagnosis for Chronic Kidney Disease and examined how effective they were at predicting the outcomes. There are 400 instances and 24 attributes in the dataset used in this analysis. The authors put 12 classification methods into the test by using data from Chronic Kidney Disease. To determine efficacy, the results of candidate methods' predictions were compared to the subject's actual medical results. The decision tree performed the highest, with an accuracy of nearly 98.6%, a sensitivity of 0.9720, a precision of 1, and a specificity of 1 [21]. A neural network-based classifier is presented in the other paper to predict whether an individual is at risk of developing chronic kidney disease (CKD). Two population groups' demographic data and medical care details are used to train the model. The model achieves 95 percent accuracy in the test data set after being trained and assessment metrics for classification algorithms are applied, making its application for disease prognosis possible. We use and verify a NN-CBR twin method to explain CKD predictions in this paper. As a result of this study, 3.494.516 people in Colombia, or 7% of the total population, were reported as being at risk of developing CKD [22]. In this study, the classification performances of tree-based methods, one of the machine learning methods, were compared. According to the findings obtained, the Random forest method gave the best classification values according to performance metrics, and other classification methods gave very high results.

Conclusion

As a result, very successful results were obtained in the study performed with the classification methods used and the chronic renal failure data set. Each model was able to classify the data with high classification performance.

Conflict of interests
There is no conflict of interest among the authors.

Financial Disclosure
All authors declare no financial support.

Ethical approval
This study does not require ethical approval and informed consent because the open-source data set is used.

References (referanslar türkçe yazılmış ingilizce olmalı)

1. Erol N. Comparison of the quality of life of patients with chronic renal failure who did not start dialysis treatment and patients who received hemodialysis treatment: Health Sciences Institute;2010.
2. Yıldırım Ü. Troponin levels in hypervolemic chronic renal failure patients and the effects of medical diuresis treatment on troponin levels: Van Yuzuncu Yil University;2020.
3. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
4. Berry MA, GS Linoff. Data mining techniques for marketing, sales and customer relationship management. 2004.
5. Silahtaroğlu G. Data mining: Concepts and algorithms: Papatya; 2013.
6. Nizam H, Akın SS. Comparison of the performance of balanced and unbalanced data sets in emotion analysis with machine learning in social media: XIX Internet Conference in Turkey; 2014.
7. Dag B, Varol A. Classification of Personality Status of Individuals According to 2D: 4D Numerical Finger Ratio: 2013.
8. Breiman L. Random forests. Machine learning. 2001;45:5-32.
9. Wang J, Li P, Ran R, et al. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Applied Sciences. 2018;8(5):689.
10. Z.S P. Evaluating XGBoost For User Classification By Using Behavioral Features Extracted From Smartphone Sensors. [Master Thesis]: KTH Royal Institute of Technology, School of Computer Science and Communication, Sweden. 2018.
11. Küçükakçalı Z, Balıkçı Çiçek İ. Performance evaluation of the ensemble learning models in the classification of chronic kidney failure. J Cognitive Systems. 2020;5:55-9.
12. Murthy SK. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery. 1998;2:345-89.
13. Salama GI, Abdelhalim M, Zeid MA-e, editors. Experimental comparison of classifiers for breast cancer diagnosis. 2012 Seventh International Conference on Computer Engineering & Systems (ICCES); 2012: IEEE.
14. Perçin İ, Yağin Fh, Arslan Ak, Çolak , editors. An Interactive Web Tool for Classification Problems Based on Machine Learning Algorithms Using Java Programming Language: Data Classification Software. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies. (ISMSIT); 2019: IEEE.
15. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports. 2020;10:1-8.
16. Biau G, Cadre B, Rouvière L. Accelerated gradient boosting. Machine Learning. 2019;108:971-92.
17. Hill NR, Fatoba ST, Oke JL, Hirst JA, O'Callaghan CA, Lasserson DS, et al. Global prevalence of chronic kidney disease–a systematic review and meta-analysis. PloS One. 2016;11:e0158765.
18. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. The lancet. 2017;389(10075):1238-52.
19. Doganer A. Prediction of renal cell carcinoma with community learning methods: İnönü University; 2020.
20. Elhoseny M, Shankar K, Uthayakumar JJSr. Intelligent diagnostic prediction and classification system for chronic kidney disease. Nature Scientific Reports. 2019;9:1-14.
21. Sharma S, Sharma V, Sharma AJapa. Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. 2016.
22. Vasquez-Morales GR, Martinez-Monterrubio SM, Moreno-Ger P, et al. Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning. IEEE Access. 2019;7:152900-10.