➔ *Regular Research Paper – NS*

Diagnosis of Chronic Kidney Disease Using Various Features

Sena GORAL
Mehmet Akif Ersoy University, Turkey
scelik@mehmetakif.edu.tr

Abstract

The development of technology provides great convenience in the field of medicine, as in every field. Analyzing the patient's personal information with modern techniques allows specialists to perform faster and more effective treatment. In this study, data with 24 attributes obtained from laboratory results of 400 patients were classified and compared with KNN, decision tree classifier, random forest classifier and adaboost classifier. As a result of classification, Ada Boost Classifier showed the best performance with score value of 98.3%.

Keywords: *Chronic kidney disease, classification, classification algorithm, machine learning, artificial intellince.*

1. INTRODUCTION

Many developments in the field of medicine are realized by the application of technological developments to the field of health. Highly advanced devices are used for processes such as care, laboratory, assistive devices, diagnosis and treatment that facilitate the work of patients and healthcare professionals. Most of these devices use a common database. Thus, in addition to the personal information obtained from the patients, a lot of information about the condition of the patients is created. These data are analyzed with modern techniques and presented to the service of healthcare professionals. In this way, a faster process is created.

Chronic kidney disease is a disease that damages the kidneys that negatively affects human life. It is also a general term for disorders that affect the structure and function of the kidney[1]. Kidney failure occurs when kidney function drops below a certain point. This condition affects the whole body. Since this disease triggers many other diseases, mortality rates can also be high.

The most well-known causes of kidney disease are high blood pressure and diabetes. More than a quarter of kidney patients have high blood pressure. Diabetes is stated to be the cause of one third of all kidney patients. In most countries, diabetes has been cited as the most common cause of kidney failure. Less commonly, kidney disease may occur from other causes. Kidney disease can also be inherited. Long-term use of certain drugs can cause this disease. In such cases, doctors cannot determine the cause of the disease[2].

According to studies, approximately 10% of the world's population is affected by chronic kidney disease. Millions of people die each year because they do not have access to appropriate treatment [3]. In the vast majority of individuals in the early stages of chronic kidney disease, this disease cannot be diagnosed. Kidney disease often progresses without any symptoms. As a result, it primarily destroys most of the kidney functions. Therefore, early treatment of this disease is very important. Because with early diagnosis, appropriate treatment can be started before other symptoms of kidney damage or deterioration ocur[2].

In the field of computer science and engineering, many studies have been conducted on chronic kidney disease. In this study, it is aimed to make the correct classification of the disease by determining some important algorithms in order to diagnose chronic kidney disease early. These; KNN, decision tree classifier, random forest classifier and adaboost classifier. In the study, the correct classification performance of each algorithm was compared and the algorithm with the highest correct classification was determined.

2. RELATED WORKS

Al-Hyari et al., proposed a system for classification for diagnosing patients with kidney failure in their study. In the data set used, there are records of 15 different qualities belonging to 102 people between the ages of 11 and 81. Decision tree, naive bayes and artificial neural network algorithms are used as classifiers. From the test results, it was stated that the decision tree had a better performance than other algorithms with its classification accuracy of 92.2%[4]Gupta et al. analyzed the relationship between diseases and related diagnostic tests for 11 different chronic diseases using machine learning techniques. Using the AdaBoost technique for chronic kidney disease, the accuracy on the training dataset was calculated as 98.67% and on the test dataset as 88.66%[5].In the study conducted by Salekin and Stankovic, machine learning techniques were used in the detection of chronic kidney disease. ANN, k-Nearest Neigbours (k-NN) and C4.5 decision tree techniques were tested. In the data sets, information from 400 people, 250 of whom were chronic kidney disease and 150 of whom were healthy, was used. C4.5 technique performed better than ANN and k-NN techniques in detecting the disease[6].Ogunleye and Wang designed an automatic diagnostic system based on the XGBoost (eXtreme Gradient Boosting) technique. This system diagnosed chronic kidney with a classification accuracy of 97.58%. In the next process, the study was developed and they achieved a classification success of 100%[7].Chen et al. They made classification using KNN, SVM and Soft Independent Modeling of Class Analogy (SIMCA) algorithms. The best result, 99.7%, was obtained from KNN and SVM algorithms[8].

3. MATERIAL METHOD

There are 24 features taken from 400 people in the data set used in the study. First of all, when the data set is examined, some of the features are numerical expressions, while others are determined categorically. Table 1 shows 24 different properties with their data types.

Table 1. Features and data types found in the dataset.

	Column	Type
1	Age	float64
2	Blood Pressure	float64
3	Specific Gravity	float64
4	Albumin	float64
5	Sugar	float64
6	Red Blood Cells	object
7	Pus Cell	object
8	Pus Cell clumps	object
9	Bacteria	object
10	Blood Glucose Random	float64
11	Blood Urea	float64
12	Serum Creatinine	float64
13	Sodium	float64

14	Potassium	float64
15	Hemoglobin	float64
16	Packed Cell Volume	float64
17	White Blood Cell Count	float64
18	Red Blood Cell Count	float64
19	Hypertension	object
20	Diabetes Mellitus	object
21	Coronary Artery Disease	object
22	Appetite	object
23	Pedal Edema	object
24	Anemia	object

In the study, first of all, the data set was checked. The unique one for each feature in the dataset is listed. It was observed that some of the values taken from individuals had missing, misspelled or misspelled features. For this, it is ensured that the features in the priorities data set are categorized according to the data types. Afterwards, the status of incorrect, incomplete or misspelled expressions, which are numerical and categorical features, were determined. Operations can be made by leaving the property value as missing or incorrect, but this prevents it from being a correct application. For the missing values in the data set, first of all, all values were separated according to the data type. It is seen that there are 14 values as numeric and 10 values as object. The data set was started to be examined by changing the values whose data type was determined as wrong expression. As a result of the operations performed according to the separated groups, the missing values of the numeric values are shown as percentages in Figure 1. In Figure 2, there is a graph showing the missing values and status for categorical features.

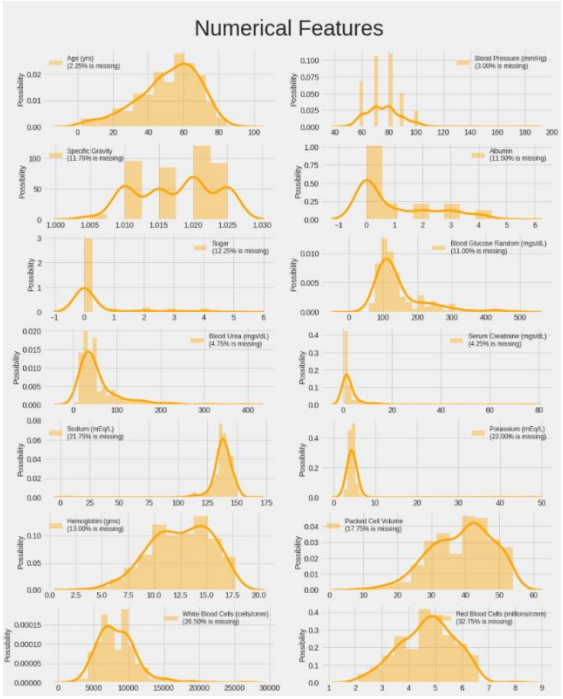

Figure 1. Numerical features status

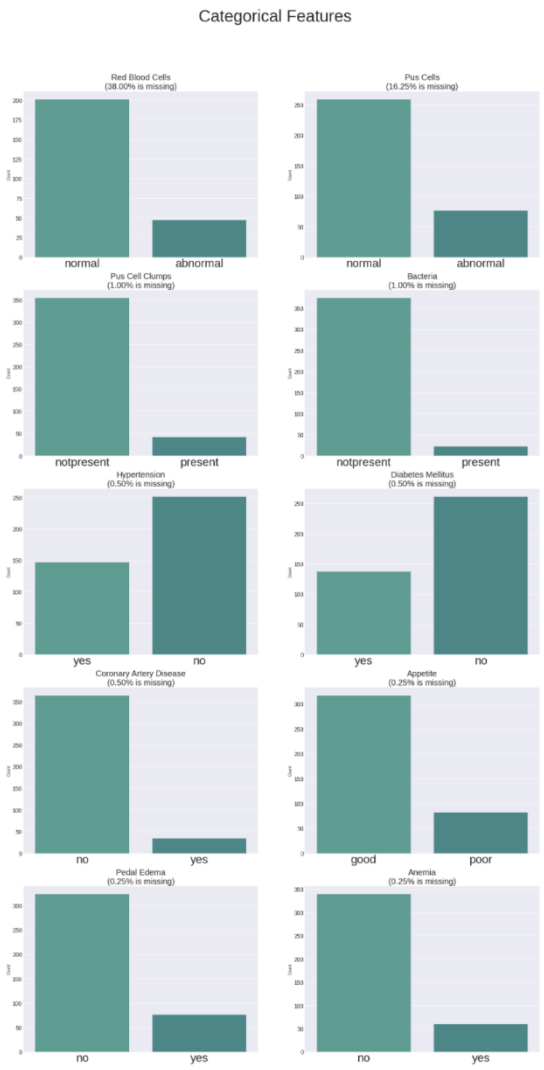

Figure 2. Categorical features status.

Random_value_imputation operations were performed for miisng values in the data set used for the system, and random values were generated in accordance with the characteristics. Comparison of all models at the end of the study is made by applying KNN, decision tree classifier, random forest classifier, island boost classifier operations on the data set prepared for applying various situations.

3.1. K-Nearest Neigbors

KNN is one of the simplest algorithms used for classification and regression. Basically, the closest points to the new point are searched. K represents the amount of nearest neighbors of the unknown point. Fitting has been done. Training Accuracy of KNN is 0.8 and Test Accuracy of KNN is 0.7166666666666667. The classification report is shown in Table 2.

Table 2. Classification Report

	Precision	Recall	F1- Score	Support
0	0.78	0.74	0.76	72
1	0.63	0.69	0.66	48
Accuracy			0.72	120
Macro avg	0.71	0.71	0.71	120
Weighted avg	0.72	0.72	0.72	120

3.2. Decision Tree Classifier

It is one of the data mining classification algorithms. It is a structure that uses large amounts of records by dividing them into very small groups of records by applying decision-making steps. Training acuracy of decision tree classifier is 1.0, test accuracy of decision tree classifier is 0.9666666666666667. The classification report is shown in Table 3.

Table 3. Classification Report

	Precision	Recall	F1- Score	Support
0	0.96	0.99	0.97	72
1	0.98	0.94	0.96	48
Accuracy			0.97	120
Macro avg	0.97	0.96	0.97	120
Weighted avg	0.97	0.97	0.97	120

Here, by increasing the data set value with GridSearchCV, the values to be formed as a result of increasing the number of combinations to be created are reviewed again. As a result of these transactions; Training Accuracy of Decision Tree Classifier is 0.9928571428571429, Test Accuracy of Decision Tree Classifier is 0.975. The classification report is shown in Table 4.

Table 4. Classification Report

	Precision	Recall	F1- Score	Support
0	0.96	1.00	0.98	72
1	1.00	0.94	0.97	48
Accuracy			0.97	120
Macro avg	0.98	0.97	0.97	120
Weighted avg	0.98	0.97	0.97	120

3.3. Random Forest Classifier

It is one of the classification methods used to increase the predictive power of the model. It is similar to the decisioc tree. But it is more suitable for preventing overfitting processes. Training Accuracy of Random Forest Classifier is 1.0, Test Accuracy of Random Forest Classifier is 0.975.

Tablo 5. Classification Report

	Precision	Recall	F1- Score	Support
0	0.96	1.00	0.98	72
1	1.00	0.94	0.97	48
Accuracy			0.97	120
Macro avg	0.98	0.97	0.97	120
Weighted avg	0.98	0.97	0.97	120

3.4. Ada Boost Classifier

Iteratively trains the machine learning model based on the correct prediction of the last training. Training Accuracy of Ada Boost Classifier is 1.0, Test Accuracy of Ada Boost Classifier is 0.975.

Table 6. Classification Report

	Precision	Recall	F1- Score	Support
0	0.96	1.00	0.98	72
1	1.00	0.94	0.97	48
Accuracy			0.97	120
Macro avg	0.98	0.97	0.97	120
Weighted avg	0.98	0.97	0.97	120

At the same time, confusion matrices were determined for all models created. Ada Boost Classifier score 0.983333, Random Forest Classifier 0.975000, Decision Tree Classifier 0.941667 and KNN 0.666667. The comparison graph of the 4 models used for the classification processes of the data set is shown in Figure 3.

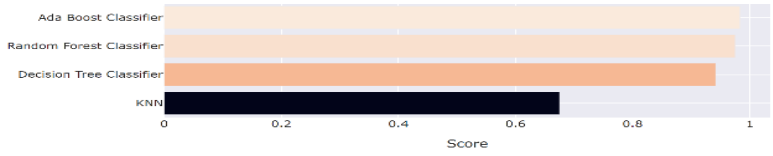

Figure 3. Model Comparison

3. CONCLUSION

Early diagnosis and treatment is very important to prevent or slow the progression of chronic kidney disease, which is one of the health problems encountered worldwide. Physicians make use of decision support systems when making decisions about disease diagnosis. In this study, KNN, decision tree, random forest and adaboost techniques were compared for chronic kidney disease. While performing the performance test, precision, recall, F1-score and support values of the techniques were used. From the results of the study, it is seen that the system is effective and can be used as an auxiliary tool by doctors.

REFERENCES

[1] Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. The Lancet, 379(9811), 165-180.

[2] World Kidney Day. (2020). Chronic Kidney Disease. Retrieved from https://www.worldkidneyday.org/facts/chronic-kidneydisease/

[3] National Kidney Foundation. (2020). Global Facts: About Kidney Disease. Retrieved from https://www.kidney.org/kidneydisease/global-facts-aboutkidney-disease#

[4] Al-Hyari, A. Y., Al-Taee, A. M., & Al-Taee, M. A. (2013, December). Clinical decision support system for diagnosis and management of chronic renal failure. In 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1-6). IEEE.

[5] Gupta, D., Khare, S., & Aggarwal, A. (2016, April). A method to predict diagnostic codes for chronic diseases using machine learning techniques. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 281-287). IEEE.

[6] Salekin, A., & Stankovic, J. (2016, October). Detection of chronic kidney disease and selecting important predictive attributes. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 262-270). IEEE.

[7] Ogunleye, A., & Wang, Q. G. (2018, June). Enhanced XGBoostbased automatic diagnosis system for chronic kidney disease. In 2018 IEEE 14th International Conference on Control and Automation (ICCA) (pp. 805-810). IEEE.

[8] Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. International urology and nephrology 2016; 48.12: 2069-2075.