DOI 10.2478/ama-2022-0022

DATA MINING APPROACH IN DIAGNOSIS AND TREATMENT OF CHRONIC KIDNEY DISEASE

Andreea S. TURIAC*, Małgorzata ZDRODOWSKA**

*Faculty of Medical Engineering, University Politehnica of Bucharest, 1-7, Gh Polizu, 011061, Bucharest, Romania
**Faculty of Mechanical Engineering, Institute of Biomedical Engineering,
Bialystok Technical University, ul. Wiejska 45C, 15-351 Bialystok, Poland

andreea.s.turiac@gmail.com, m.zdrodowska@pb.edu.pl

received 2 November 2021, revised 14 March 2022, accepted 15 March 2022

Abstract: Chronic kidney disease is a general definition of kidney dysfunction that lasts more than 3 months. When chronic kidney disease is advanced, the kidneys are no longer able to cleanse the blood of toxins and harmful waste products and can no longer support the proper function of other organs. The disease can begin suddenly or develop latently over a long period of time without the presence of characteristic symptoms. The most common causes are other chronic diseases – diabetes and hypertension. Therefore, it is very important to diagnose the disease in early stages and opt for a suitable treatment - medication, diet and exercises to reduce its side effects. The purpose of this paper is to analyse and select those patient characteristics that may influence the prevalence of chronic kidney disease, as well as to extract classification rules and action rules that can be useful to medical professionals to efficiently and accurately diagnose patients with kidney chronic disease. The first step of the study was feature selection and evaluation of its effect on classification results. The study was repeated for four models – containing all available patient data, containing features identified by doctors as major factors in chronic kidney disease, and models containing features selected using Correlation Based Feature Selection and Chi-Square Test. Sequential Minimal Optimization and Multilayer Perceptron had the best performance for all four cases, with an average accuracy of 98.31% for SMO and 98.06% for Multilayer Perceptron, results that were confirmed by taking into consideration the F1-Score, for both algorithms was above 0.98. For all these models the classification rules are extracted. The final step was action rule extraction. The paper shows that appropriate data analysis allows for building models that can support doctors in diagnosing a disease and support their decisions on treatment. Action rules can be important guidelines for the doctors. They can reassure the doctor in his diagnosis or indicate new, previously unseen ways to cure the patient.

Key words: feature selection, classification, classification rules, action rules, data mining, chronic kidney disease

1. INTRODUCTION

Chronic kidney disease (CKD) is a common disease that affects between 8% and 16% of the population worldwide [1]. It is often misdiagnosed or underdiagnosed in the earlier stages because there are no particular evident symptoms in these stages of development, but can it can be detected through laboratory testing. Due to the lower rate of proper identification in the incipient phases of the disease, Kidney Disease Outcomes Initiative of the National Kidney Foundation has recently proposed guidelines to describe CKD. In these guidelines it is stated that CKD is characterised by structural or functional abnormalities that last and/or progress for more than 3 months, with or without decreased glomerular filtration rate (GFR), manifest by either pathological abnormalities or markers of kidney damage, including abnormalities in the specific blood or urine tests or in medical imaging tests or by a GFR less than 60 mL/min/1.73 m2 for more than 3 months, with or without kidney damage [2,3].

Early detection is extremely important to minimise the chances of progression to kidney failure. There are initiating factors that can contribute to increasing the risk of developing CKD, and some of them are related to ethnicity and family health record, whether or not end-stage kidney disease is present along with high-risk factors. Other aspects that can be taken into consideration are age, stating that the number of nephrons that loss function is increased with ageing, gender, some studies implying that the progression of CKD is more rapid at men [4]. Diabetes and hypertension are the two main causes of CKD which are responsable for up to two-third of the cases: diabetes, due to the presence of too much glucose in the blood that damage the filtering function of the kidneys and high blood pressure that can affect the blood vessels that irrigate the kidneys. Apart from them, glomerulonephritis and unknown causes are more common in countries of Asia and sub-Saharan Africa [5].

With the possibility of early detection, it will be a transition from a life-threatening condition that requires lifelong care and the imminent occuring of dialysis to a more common condition that focuses on prevention and slowing the loss of kidneys functionality [6]. Data mining is an effective instrument to extract useful hidden information from voluminous datasets. Health industry provides a large amound of complex data about patients and diseases that requires preparation, processing, modelling and evaluation for knowledge extraction that can be used by the healthcare professionals when making diagnosis decisions and treatment plans [7]. Medical data is loaded with structured and unstructured information and it is characterised by an inconvenient aspect: high dimensionality. Feature selection is a common processing procedure for dimensionality reduction, so the algo-

rithm is modelled for better understanding of the underlying trends within the dataset [8]. The analysis of medical data allows to reduce its dimensionality and also to extract certain rules that may be relevant to the diagnosis and treatment processes. It also gives the possibility to modify these rules by replacing some flexible attributes. It allows to reclassify the patient from one group to another. These rules can help doctors in their work by giving some guidance, for example on treatment options [9].

Various automatic diagnosis methods have been proposed and tested to detect the early stage of CKD. Avci et al. provides a performance comparison using different classifiers: Naïve Bayes, Support Vector Machine, K-Star and the famous J48 [10]. A. Rady et al. analyse the alternative of using artificial neural network algorithms and support vector machine to determine which algorithm display the best classification results [11]. Attribute sellection and clustering methods were used by S.B. Akben to create subsets of the dataset to be further evaluated with K-Nearest Neighbour. The attributes were divided into three main categories, those related to blood tests, urine tests and other parameters, and different combinations of subsets were tested [12]. By considering that some elements of information are hidden from clinical data, these techniques can facilitate, as well as lower the cost of, a less invasive approach of diagnosis.

The aim of this article is to analyse and select features and to investigate the impact of feature selection on the selected classifiers accuracy. The paper will also show the extracted classification and action rules.

2. MATERIALS AND METHODS

"Chronic Kidney Disease" is the dataset used in this paper, extracted from UCI Machine Learning Repository. It is a collection of 400 instances with 24 attributes plus the class attribute, registered during a period of aproximatively 2 months at the Apollo Hospitals. The characteristics of all attributes are shown in Tab. 1. The original dataset contains missing values which can lead to inaccurate results and reduce the model accuracy. Instead of eliminating the instance from the dataset, we opt for replacing the missing values using statistical methods. Supervised attribute filter ReplaceMissingValues [40] from Weka software, was used to fill the unknown values by calculating the mean of all values for a specific attributes – the mean of the column. The attributes are numeric and nominal and they indicate the results of a range of blood and urine tests and the presence or absence of common diseases that increase the risk of developing CKD. There are two classes: 250 instances distributed for ckd which means a high probability of have chronic kidney disease in early stages and 150 instances for notckd, the patients that are generally not prone to chronic kidney disease.

Tab. 1. Attribute information

No.	Attribute name	Description	Average value
1	age	age of the patient (num) in years	51
2	bp	blood pressure (num) in mm/hg	76
3	sg	specific gravity (nom)	-
4	al	albumin (nom)	-
5	su	sugar (nom)	-
6	rbc	red blood cells (nom)	normal/abnormal
7	pc	pus cell (nom)	normal/abnormal
8	pcc	pus cell clumps (nom)	present/not present
9	ba	bacteria (nom)	present/not present
10	bgr	blood glucose random (num) in mgs/dl	148
11	bu	blood urea (num) in mgs/dl	57
12	sc	serum creatinine (num) in mgs/dl	3
13	sod	sodium (num) in meq/l	138
14	pot	potassium (num) in meq/l	4.62
15	hemo	hemoglobin (num) in gms	12.5
16	pcv	packed cell volume	39
17	wbcc	white blood cell count (num) in cells/cum	8406
18	rbcc	red blood cell count(num) in millions/cmm	4.7
19	htn	hypertension (nom)	yes/no -values
20	dm	diabetes mellitus (nom)	yes/no -values
21	cad	coronary artery disease (nom)	yes/no -values
22	appet	appetite (nom)	good/poor
23	pe	pedal edema (nom)	yes/no -values
24	ane	anemia (nom)	yes/no -values
25	class	class (nom)	ckd/notckd

In order to ease the learning procedure, the raw values were transformed into descriptive data which can better express the medical information. All the numerical data were saved as nominal data and then, the individual test values were divided into specific ranges accordingly with the ones reported in literature:

- blood presure (bp) [13]:
  - less that 60 mm/Hg – low (0);
  - 60-80 mm/Hg – normal (1);
  - 80-90 mm/Hg – prehypertension (2);
  - higher than 90 – hypertension (3);
- blood glucose random (bgr) [14]:
  - less than 70 mgs/dl – hypoglicemia (0);
  - 70-125 mgs/dl – normal (1);
  - 125-200 mgs/dl – high (2);
  - 200-350 mgs/dl – extremely high (3);
  - higher than 380 – metabolic consequences (4);
- blood urea (bu) [15]:
  - 8-21 mgs/dl – normal (1);
  - higher than 21 mgs/dl – high (2);
- serum creatinine (sc) [16]:
  - less than 1.2mg/dl – normal (1);
  - 1.2–2mg/dl – mild renal (2);
  - 2–3mg/dl – moderate renal (3);
  - higher than 3 mg/dl – severe renal (4);
- sodium (sod) [17]:
  - lower than 135 mEq/L – hyponatremia (0);
  - 135-145 mEq/L – normal (1);
  - higher that 145 mEq/L – high (2);
- potasium (pot) [18]:
  - lower than 3.5 mEq/L – close to hypokalemia (0);
  - 3.5-5 mEq/L – normal (1);
  - higher than 5 mEq/L – high (2);
- hemoglobin (hemo) [19]:

DOI 10.2478/ama-2022-0022

- lower than 12.5 gms – low (0);
- 12.5-17.5 gms – normal (1);
- higher that 17.5 gms – high (2);

– packed cell volumes (pcv) [20]:
- lower that 36% – low (0);
- 36-53% – normal (1);
- higher than 53% – high (2);

– white blood cell count (wbcc) [21]:
- lower than 4000 cells/cum – low (0);
- 4000-11000 cells/cum – normal (1);
- higher than 11000 cells/cum – high (2);

– red blood cell count (rbcc) [22]:
- lower than 3.92 millions/cmm – low (0);
- 3.92-5.65 millions/cmm – normal (1);
- higher than 5.65 millions/cmm – high (2);

The first step of the research was feature selection and study their effect on the accuracy of classifying patients into healthy and chronic kidney disease groups. In this case, some of the attributes need to be removed due to their little relevance. Following methods were used for attribute reduction:

– Correlation Based Feature Selection is a fully automatic algorithm used to determine a good feature subset that contains the attributes highly predictive of the class correlated and, simultaneously, uncorrelated with each other. All the features and the class are treated in a uniform manner and the merit of each attribute is calculated using Ranker Search Method. Irrelevant features should be neglected because they have low correlation to the class and redundant features should be removed as they are highly correlated with at least one of the remaining features [23].
– Chi-Square Test for Feature Selection was used to test the relationship between the features. It starts from the assumption that two characteristics are independent of each other, and then evaluate whether this hypothesis is correct by calculating the statistics, the magnitude of the deviation between the actual and theoretical values [24].

The second part of this study was building models including all attributes and attributes extracted by feature selection. To carry out the second part of this study, models were built using either all the attributes or attributes extracted by features selection methods. Then we applied the classification to check if the selection of features gave the expected results. For classification, we used the following algorithms:

– AdaBoostM1 algorithm generates a strong classifier using a linear combination of member classifiers and selects a member classifier to minimize the error and to maximize the diversity among the member of the classifiers in each cycle [25]. It is particularly common to use decision stumps, small decision trees with two leaves, to build more complex base learners that provides good classifiers when boosted [26].
– Sequential Minimal Optimization (SMO) split the quadric programming problems into a series of smallest possible QP subproblems which are solved analytically. The main advantages of this problems are related to small memory space because it solves the problem without any extra matrix storage, and it requires less computing time due to a non-iterative routine for each small problem [27].
– Multilayer Perceptron is an artificial feed-forward neural network based on a three layers architecture: the input layer, the hidden layer with a non-linear activation function and an output layer where the classification task is performed. Each layer has a various number of neurons that are trained using back propagation learning algorithm [28]. Each neuron has a mathematical function that gain the input from a previous layer and produce the output for the following layer [29].
– Naïve Bayes Classification is based on the Bayes' Theorem and provides a way of combining prior probability and conditional probabilities into a single formula and then choose the classification with the highest values. The premise of this algorithm is that all the attributes contribute equally and independently to the model. In practice, this assumption is not correct but Naïve Bayes algorithm has become an important probabilistic model with remarkable success in practice [30].
– J48 decision tree in which every detail of the information is split into minor subsets by choosing an attribute. The principle of J48 decision tree is to split every detail present in the information into minor subsets by choosing an attribute of reference. At each node, the algorithm chooses the highest worthy information-gain attribute to split the data. This process is stopped when a subset has a place with a similar class in all the instances [31].
– JRip main premise is to produce error reduction at each incremental pruning and it consists on two phases: the grow phase when it continues to add terms to the rule until it is accurate and the incrementally pruning phase of each rule [32].
– CART is a tree-building technique structured as a binary recursive portioning as each node from the decision tree can be split in only two groups. It is a practical algorithm used in clinical setting because it creates uncomplicated rules that have a common point with the perspective of the clinicians [33].
– The idea behind the PART algorithm is to build a partial tree with a separate-and-conquer strategy: when it creates a rule, the instances covered in it are removed and the process continues with the remaining instances until there are none left [34].
– Random Trees uses a collection of tree classifiers and produces a random set of data to build a decision tree. The input data is classified at each tree and the overall decision is made by so called "votes". At a node, a random subset of training data is analysed and the best split is made for that particular subset [35].

We used a 10-fold cross-validation for testing, training and validation. The basic principle is to divide the data: a high percent of the data is used to build the model and then use the left-out samples to be predicted as unseen data [36]. In a 10-fold cross-validation the dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The model is trained and tested each time, it is trained on the entire dataset leaving out the specific fold and then it is tested on the leave-out subset. The accuracy estimated is the overall number of correct classifications divided by the number of the instances in the dataset [37].

To evaluate the above classifiers, we used Total Accuracy (ACC) and F1-Score.

ACC is the total efficiency of the classifier, which determines the probability of correct classification, i.e. the ratio of correct classification s to all classifications. It is expressed by the equation [38]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

A considerable disadvantage about the ACC as it does not take into consideration the differences between the types of error,

it does not punish the fact that the model classifies i.e a patient as a false negative, meaning that he is diagnosed as not having the disease when he actually has the disease (false negative values). Another aspect is related to the case when there is an unbalanced dataset, as ACC does not provide a realistic measurement – it is more effective to consult the confusion matrix (Tab. 2) [38].

Recall is the ability of a model to find all the relevant cases within the dataset, being defined as the number of true positives divided by the sum of true positives and false negatives. Precision is the ability of a classification to identify the positive features and it is defined as the number of true positives divided by the number of all instances that were classified as positives. F1-Score is an optimal blend, the harmonic mean of Recall and Precision [10]:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (2)$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (3)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4)$$

Tab. 2. Confusion Matrix

Hypothesized class	Actual class	
	Positives	Negatives
Yes	True Positives	False Pozitives
No	False Negatives	True Negatives

The final steps of this study were construction of classification rules and extraction the action rules from classification rules.

Action rules are constructed from classification rules which suggest an alternative to reclassify the instances. These rules indicate the changes in an attribute that need to be made to integrate an instance, in this case, a patient into a different category, all accordingly to the information from the clinicians. It is crucial to find useful rules from analysing the data and identify the relevant patterns that best describes the instances [9].

An action rule can be presented in the following form [41,42]:

$$[(\omega) \wedge (\alpha \rightarrow \beta)] \Longrightarrow (\Psi \rightarrow \Omega) \quad (5)$$

where ω indicates a fixed condition features conjunction, that is a part of both groups, (α → β) is recommended changes in flexible features value and (Ψ → Ω) means an effect of the action, which the user wants to achieve.

3. RESULTS

To enhance the performance of the model, first, the dimension of the data set needs to be reduced, the irrelevant features or features that are little correlated with the class label should be neglected. [39] As mentioned earlier, two feature selection methods were used:

- Correlation Based Feature Selection with Ranker Search Method, taking into consideration the first 6 attributes with a merit of approximate 0.5 or higher
- Chi-Square Test choosing the first 8 most independent features, listed in the table bellow:

In addition, we also took into one of the models only those attributes indicated by doctors as most important in chronic kidney disease. These features include blood pressure, specific gravity, diabetes mellitus, hypertension, albumin, blood urea, serum creatinine, sodium, potassium, haemoglobin, red blood cell count and packed cell volume [1,2,4]. The results of attribute selection are shown in Tab. 3. It is important to note that 6 of all features were able to be extracted for each of the methods mentioned above. These features are highlighted with the same color in Tab. 2.

The results of classification for each of analyzed models are shown in Figs. 1-3.

Tab. 3. Feature selection results

Attributes indicated by doctors	**Correlation Based Feature Selection**	**Chi-Square Test**
blood pressure	hemoglobin	serum creatinine
specific gravity	hypertension	specific gravity
diabetes mellitus	diabetes mellitus	hemoglobin
hypertension	serum creatinine	albumin
albumin	albumin	hypertension
blood urea	packed cell volume	diabetes mellitus
serum creatinine		red blood cell count
sodium		packed cell volume
potassium		
hemoglobin		
red blood cell count		
packed cell volume		

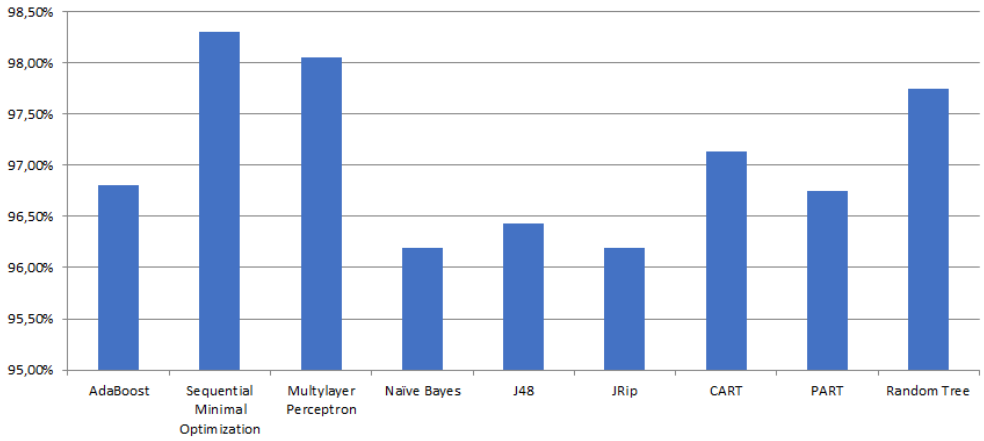
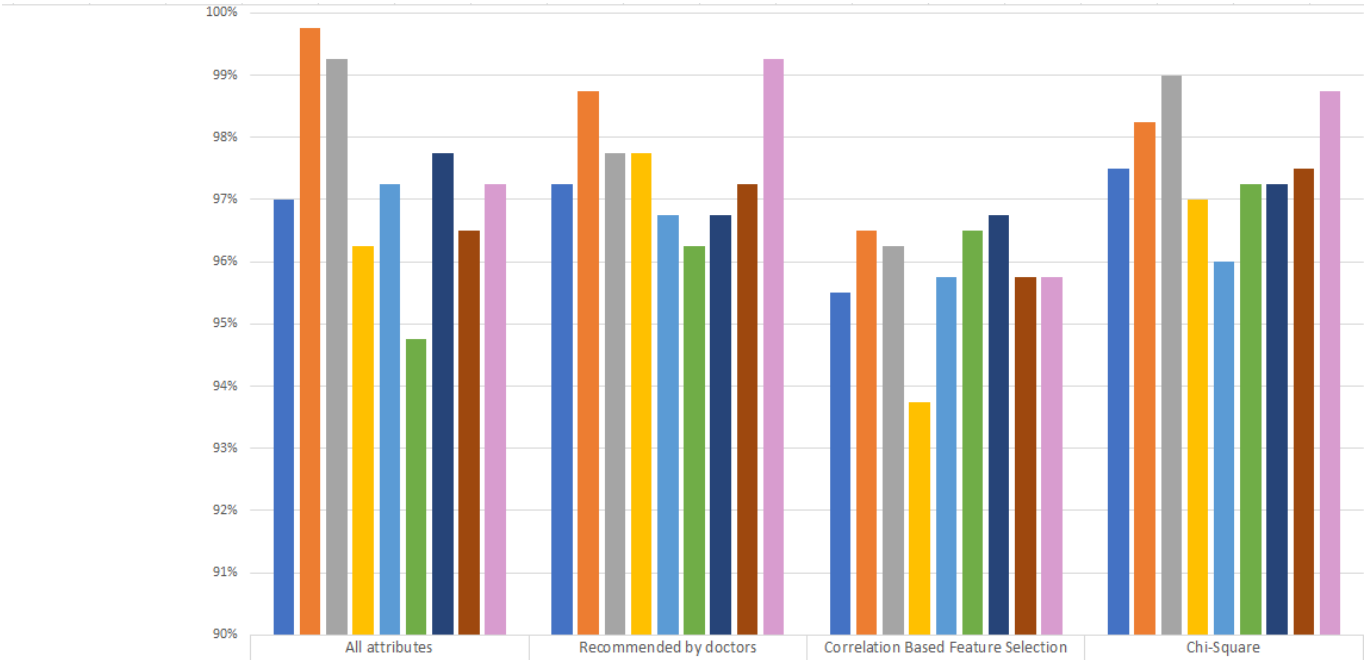


Fig. 1. Average accuracy

DOI 10.2478/ama-2022-0022



	All attributes (24 attributes)	Recommended by doctors (12 attributes)	Correlation Based Feature Selection (6 attributes)	Chi-Square (8 attributes)
AdaBoost	97%	97,25%	95,50%	97,50%
Sequential Minimal Optimization	99,75%	98,75%	96,50%	98,25%
Multylayer Perceptron	99,25%	97,75%	96,25%	99%
Naïve Bayes	96,25%	97,75%	93,75%	97%
J48	97,25%	96,75%	95,75%	96%
JRip	94,75%	96,25%	96,50%	97,25%
CART	97,75%	96,75%	96,75%	97,25%
PART	96,50%	97,25%	95,75%	97,50%
Random Tree	97,25%	99,25%	95,75%	98,75%

Fig. 2. Total Accuracy (ACC) Results

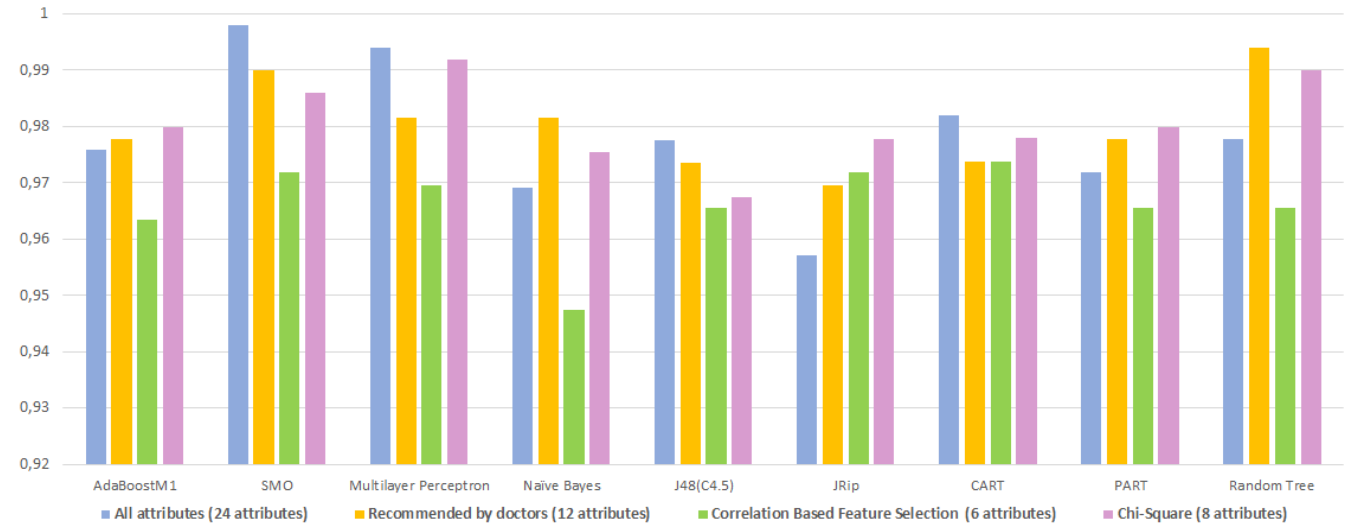


Fig. 3. F1-Score Results

The next step of the research was to identify the classification rules. These rules were obtained using the classifier algorithms mentioned earlier. A few dozen rules were obtained for each model. We compared the rules extracted for the model containing all features and for the models after feature reduction. In the following, we present dozens of rules that classify patients into a group at high risk for chronic kidney disease (underlined are those rules that were obtained in both models: the model containing all features and the models after feature selection):

- IF albumin = 0 AND hemoglobin = 1 AND serum creatinine = 1 AND hypertension = yes THEN ckd
- IF albumin = 0 AND hemoglobin = 1 AND serum creatinine = 1 AND hypertension = no AND packed cell volume = 0 THEN ckd
- IF albumin = 0 AND hemoglobin = 1 AND serum creatinine = 1 AND hypertension = no AND packed cell volume = 1 AND diabetes mellitus = yes THEN ckd

- IF albumin = 0 AND hemoglobin = 1 AND serum creatinine = 2 THEN ckd
- IF albumin = 0 AND hemoglobin = 1 AND serum creatinine = 3 THEN ckd
- IF albumin = 0 AND hemoglobin = 2 THEN ckd
- IF hemoglobin = 0 THEN ckd
- IF hemoglobin = 1 AND diabetes mellitus = no AND albumin = 0 AND serum creatinine = 2 or 3 or 4 THEN ckd
- IF hemoglobin = 1 AND diabetes mellitus = no AND albumin = 2 or 3 or 4 THEN ckd
- IF hemoglobin = 1 AND diabetes mellitus = no AND serum creatinine = 1 AND specific gravity = 2 or 3 THEN ckd
- IF hemoglobin = 1 AND diabetes mellitus = no AND serum creatinine = 2 or 3 THEN ckd
- IF hemoglobin = 1 AND diabetes mellitus = yes THEN ckd
- IF hemoglobin = 1 AND specific gravity = 1 or 2 or 3 THEN ckd
- IF hemoglobin = 1 AND specific gravity = 4 AND serum creatinine = 1 AND albumin = 0 AND packed cell volume = 0 THEN ckd
- IF hemoglobin = 1 AND specific gravity = 5 AND serum creatinine = 2 or 3 THEN ckd
- IF hemoglobin = 1 AND specific gravity = 5 AND serum creatinine = 1 AND albumin = 1 or 2 or 3 or 4 or 5 THEN ckd
- IF hemoglobin = 2 THEN ckd
- IF hypertension = no AND albumin = 0 AND diabetes mellitus = no AND specific gravity = 2 or 3 THEN ckd
- IF hypertension = no AND albumin = 0 AND diabetes mellitus = yes THEN ckd
- IF hypertension = no AND albumin = 1 or 2 or 3 or 4 THEN ckd
- IF hypertension = yes THEN ckd
- IF serum creatinine = 1 AND albumin = 0 or 5 AND hemoglobin = 0 THEN ckd
- IF serum creatinine = 1 AND albumin = 0 or 5 AND hemoglobin = 1 or 2 AND diabetes mellitus = yes THEN ckd
- IF serum creatinine = 1 AND albumin = 1 or 2 or 3 or 4 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 1 or 2 or 3 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 AND hypertension = yes THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 AND hypertension = no AND albumin = 0 AND potassium (pot) = 1 AND blood urea (bu) = 1 AND sodium = 0 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 AND hypertension = no AND albumin = 0 AND potassium (pot) = 1 AND blood urea (bu) = 2 AND sodium = 0 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 AND hypertension = no AND albumin = 0 AND potassium (pot) = 2THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 AND hypertension = no AND albumin = 1 or 2 or 3 or 4 or 5THEN ckd
- IF serum creatinine = 1 AND specific gravity = 4 or 5 AND albumin = 1 or 2 or 4 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 5 AND albumin = 1 or 2 or 3 or 4 or 5 THEN ckd
- IF serum creatinine = 1 AND specific gravity = 5 or 4 AND albumin = 1 or 3 or 4 THEN ckd
- IF serum creatinine = 2 or 3 or 4 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 1 or 2 or 3 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 4 AND hypertension = yes THEN ckd
- IF serum creatinine = 4 AND specific gravity = 4 AND hypertension = no AND sodium = 0 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 4 AND hypertension = no AND sodium = 1 AND albumin = 0 AND diabetes mellitus = yes or no THEN ckd
- IF serum creatinine = 4 AND specific gravity = 4 AND hypertension = no AND sodium = 1 AND albumin = 1 or 2 or 3 or 4 or 5 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 4 AND hypertension = no AND sodium = 2 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 5 AND blood pressure = 0 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 5 AND blood pressure = 1 AND albumin = 1 or 2 or 3 or 4 or 5 THEN ckd
- IF serum creatinine = 4 AND specific gravity = 5 AND blood pressure = 2 or 3 THEN ckd

The classification rules read as follows: e.g. the last rule means that if serum creatinine is at level 4, specific gravity is at level 5 and blood pressure is at level 2 or 3, the patient is classified in the high probability of having chronic kidney disease in early stages group.

The final step of the study was to extract action rules that would allow the reclassification of patients from the group at high risk for chronic kidney disease to the group that are generally not prone to chronic kidney disease. Here, we also received over a hundred rules. Below there are some selected action rules extracted from the classification rules (again, underlined are those rules that were obtained in both models: the model containing all features and the models after feature selection):

- [albumin=0] ∧ [hemoglobin=1] ∧ [serum creatinine=1] ∧ [hypertension=no] ∧ [packed cell volume=1] ∧ [diabetes mellitus, yes→no] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [diabetes mellitus=no] ∧ [albumin, 4→5] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [diabetes mellitus=no] ∧ [albumin=0] ∧ [serum creatinine, 2→1] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [diabetes mellitus=no] ∧ [serum creatinine=1] ∧ [specific gravity, 2→1] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [diabetes mellitus=no] ∧ [serum creatinine=1] ∧ [specific gravity, 3→4] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [diabetes mellitus=no] ∧ [serum creatinine=4] ∧ [rbcc, 0→2] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [specific gravity=4] ∧ [serum creatinine=1] ∧ [albumin=0] ∧ [packed cell volume, 0→2] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [specific gravity=4] ∧ [serum creatinine=1] ∧ [albumin=0] ∧ [packed cell volume=1] ∧ [rbcc, 0→1] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [specific gravity=5] ∧ [albumin, 2→4] ⟹ [class, ckd → notckd]
- [hemoglobin=1] ∧ [specific gravity=5] ∧ [serum creatinine=1] ∧ [albumin, 1→0] ⟹ [class, ckd → notckd]
- [hypertension=no] ∧ [albumin, 3→5] ⟹ [class, ckd → notckd]

DOI 10.2478/ama-2022-0022

- [hypertension=no] ∧ [albumin, 4→5] ⟹ [class, ckd → notckd]
- [hypertension=no] ∧ [albumin=0] ∧ [diabetes mellitus=no] ∧ [specific gravity, 2→1] ⟹ [class, ckd → notckd]
- [hypertension=no] ∧ [albumin=0] ∧ [diabetes mellitus=no] ∧ [specific gravity, 3→4] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [albumin=0] ∧ [hemoglobin=1] ∧ [diabetes mellitus, yes→no] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [albumin=0] ∧ [hemoglobin=2] ∧ [diabetes mellitus, yes→no] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=4] ∧ [albumin, 1→0] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=4] ∧ [albumin,3→2] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=4] ∧ [albumin,4→5] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=4] ∧ [hypertension=no] ∧ [albumin=0] ∧ [pot=1]∧[bu =1]∧[sod, 0→1] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 1→0] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 1→2] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 1→5] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 3→0] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 3→2] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 3→5] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 4→0] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 4→2] ⟹ [class, ckd → notckd]
- [serum creatinine=1] ∧ [specific gravity=5] ∧ [albumin, 4→5] ⟹ [class, ckd → notckd]

The action rules read as follows: e.g. the last action rule means that if serum creatinine is at level 1, specific gravity is at level 5, and albumin level is changed from 4 to 5, then we can reclassify the patient from the high probability of having chronic kidney disease in early stages group to the patients that are generally not prone to chronic kidney disease.

4. DISCUSSION

Chronic Kidney Disease means a chronic disease associated with kidney failure. Currently, kidney function is traditionally assessed by blood and urine tests. However, it is important to develop a CKD system to recognize the early stages of CKD and its symptoms. In this way, preventive measures can be taken to manage the disease at an early stage and avoid its complications.

Classification, one of the methods of data mining that involves finding a way to map data into a set of predefined classes, can be helpful here. Based on test results, we can assign a given patient to the appropriate disease class. In our work, we made a classification this for four models: containing all available patient data, containing features identified by doctors as major factors in chronic kidney disease, and models containing features selected using Correlation Based Feature Selection and Chi-Square Test.

Here we see that for each model, the highest accuracy was obtained for the Sequential Minimal Optimization and Multilayer Percepton algorithms. These results are also confirmed by the F1-Score.

For the model with all the attributes Sequential Minimal Optimization and Multilayer Perceptron performs particularly well, with an accuracy of 99.75%, respectively 99.25% and an F1-Score of 0.998, respectively 0.994.

For the model of attributes with references in the literature, those recommended by clinicians Random Tree has the highest accuracy of 99.25% and F1-Score of 0.994, followed by Sequential Minimal Optimization – 98.75% and F1-Score of 0.989 and Multilayer Perceptron – 97.75% and F1-Score of 0.981.

For the model with attributes reduced with Correlation Based Feature Selection CART and JRIP have the highest accuracy with a 96.75% and 0.973 F1-Score, respectively 96.5% and 0.971 F1-Score, closely followed by Sequential Minimal Optimization – 96.5% and Multilayer Perceptron 96.25%.

For the model with attributes reduced with Chi-Square Test Multilayer Perceptron has the best results with an accuracy of 99% and a F1-Score of 0.991, followed by Random Tree 98.75% accuracy and 0.989 F1-Score and Sequential Minimal Optimization with 98.25% accuracy and 0.979 F1-Score.

It is worth mentioning that other authors [43, 44] also worked on the same database. They often used other feature selection methods and other classification methods and also achieved high accuracy rates.

Based on ours models, we also extracted dozens of decision rules and then action rules that that would allow the reclassification of patients from the group at high risk for chronic kidney disease to the group that are generally not prone to chronic kidney disease.

The overall objective of this analysis is to find methods to correctly predict the presents of chronic kidney disease in early stages and to find optimal guidance and particular treatment for each patient based on the results. Our study was conducted using the results of a survey of 400 individuals. This may be an insufficient research sample. In order to use intelligent algorithms for optimal diagnosis, larger verified datasets are required. The feature selection made with Chi-Square Test performed very well with Multilayer Perceptron, Random Tree and SMO algorithms and this hybrid method can be improved by creating larger datasets with the attributes identified with feature selection.

5. CONCLUSION

Chronic kidney disease is a lifestyle disease that affects more and more people. This disease is special because it can be a consequence or complication of all other diseases of civilization, from obesity, diabetes, hypertension and cardiovascular diseases. Too rarely diagnosed, it occurs much more frequently than previously thought. Its course is very often hidden, therefore this chronic kidney disease is a real challenge for the XXI century medicine.

Therefore, a very important aspect is the proper and early diagnosis and the support of doctors in the process of diagnosis and treatment. Data mining, which is increasingly used in medicine and its related fields, can be helpful here. Data mining allows for a different way of looking at the disease and the factors causing it. It is based on the research of doctors, but also looks for

completely different correlations. It can connect the features that are not obvious. Appropriate data analysis allows for building models that can support doctors in diagnosing a disease, support their decisions on treatment or rehabilitation of a patient. An important aspect of data mining are classification rules and action rules. Especially the latter can be an important guideline for doctors. They can reassure the doctor in his diagnosis or indicate new, previously unseen ways to cure the patient.

REFERENCES

1. Chen TK, Knicely DH, Grams ME. Chronic Kidney Disease Diagnosis and Management: A Review. JAMA - Journal of the American Medical Association. 2019;322(13):1294–1304. doi: 10.1001/jama.2019.14745
2. Coresh J. Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey. American Journal of Kidney Diseases. 2003;41(1):1–12. doi: 10.1053/ajkd.2003.50007.
3. Tuominen TK, Jämsä T, Oksanen J, Tuukkanen J, Gao TJ, Lindholm TS, Jalovaara PK. Composite implant composed of hydroxyapatite and bone morphogenetic protein in the healing of a canine ulnar defect. Annales Chirurgiae et Gynaecologiae. 2001;90(1):32-36 .
4. Evans PD, Taal MW. Epidemiology and causes of chronic kidney disease. Chronic Renal Failure. 2011;39(7):402–406.
5. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, et al. Chronic kidney disease: Global dimension and perspectives. The Lancet: series Global Kindey Disease. 2013;382(9888):260–272.
6. Levey AS, Astor BC, Stevens LA, Coresh J. Chronic kidney disease, diabetes, and hypertension: What's in a name. Kidney International. 2010;78(1):19–22. doi: 10.1038/ki.2010.115.
7. Kunwar V, Chandel K, Sabitha AS, Bansal A. Chronic Kidney Disease analysis using data mining classification techniques. 6th International Conference - CloudSystem and Big Data Engineering (Confluence). 2016;300–305. doi: 10.1109/CONFLUENCE.2016.7508132.
8. Manonmani M, Balakrishnan S. Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset. Procedia Computer Science. 2020;171(2019):1660–1669. doi: 10.1016/j.procs.2020.04.178
9. Dardzińska A. Action Rules Mining. Springer-Verlag, Berlin. 2013.
10. Avci E, Karakus S, Ozmen O, Avci D. Performance comparison of some classifiers on Chronic Kidney Disease data. 6th International Symposium on Digital Forensic and Security (ISDFS). 2018;1-4. doi: 10.1109/ISDFS.2018.8355392.
11. Rady EHA, Anwar AS. Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked. 2019;15:100178. doi: 10.1016/j.imu.2019.100178.
12. Akben SB. Early Stage Chronic Kidney Disease Diagnosis by Applying Data Mining Methods to Urinalysis, Blood Analysis and Disease History. IRBM. 2018;39(5):353–358. doi: 10.1016/j.irbm.2018.09.004.
13. Simunovic VL. Basic & General Clinical Skills. CreateSpace Independent Publishing Platform. 2013.
14. Freeth A. Diabetes Causes, Myths, Treatment, and Home Care. eMediHealth. 2019.
15. Jujo K, Minami Y, Haruki S, Matsue Y, Shimazaki K, Kadowaki H, Ishida I, Kambayashi K, Arashi H, Sekiguchi H, Hagiwara N. Persistent high blood urea nitrogen level is associated with increased risk of cardiovaserum creatinineular events in patients with acute heart failure. ESC Heart Failure, 2017;4(4):545–553.
16. Piñol-Ripoll G, De La Puerta I, Purroy F. Serum creatinine is an inadequate screening test for renal failure in ischemic stroke patients. Acta Neurologica Scandinavica. 2009;120(1):47–52. doi: 10.1111/j.1600-0404.2008.01120.x.
17. Strazzullo P, Leclercq C. Nutriente information: Sodium. Advances in Nutrition. 2014;5(2):188–90 doi: 10.3945/an.113.005215.
18. Kardalas E, Paschou SA, Anagnostis P, Muscogiuri G, Siasos G, Vryonidou A. Hypokalemia: A clinical update. Endocrine Connections. 2018;7(4):135–146. doi: 10.1530/EC-18-0109.
19. Walker HK, Hall WD HJ. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. 1990.
20. Fairbanks VF, Tefferi A. Normal ranges for packed cell volume and hemoglobin concentration in adults: Relevance to "apparent polycythemia." European Journal of Haematology. 2000;65(5): 285–296. doi: 10.1034/j.1600-0609.2000.065005285.x.
21. White Blood Cell Count. Nursing Critical Care. 2019;14:1-40. doi: 10.1097/01.CCN.0000549633.67301.6d
22. Red Blood Cell Count. Nursing Critical Care. 2020;15(1):1-38. doi: 10.1097/01.CCN.0000612852.86589.d2
23. Hall MA. Correlation-based Feature Selection for Machine Learning. Doctoral thesis. University of Waikato. 1999.
24. Sun J, Zhang X, Liao D, Chang V. Efficient method for feature selection in text classification. 2017 International Conference on Engineering and Technology (ICET). 2017;1–6. doi: 10.1109/ICEngTechnol.2017.8308201.
25. An TK, Kim MH. A new Diverse AdaBoost classifier. Artificial Intelligence and Computational Intelligence. 2010;1:359–363. doi: 10.1109/AICI.2010.82.
26. Kegl B, Introduction to AdaBoost. 2014. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.8866&rep=rep1&type=pdf, 10 October 2021.
27. Zeng ZQ, Yu H Bin, Xu HR, Xie YQ, Gao J. Fast training Support Vector Machines using parallel Sequential Minimal Optimization. rd International Conference on Intelligent System and Knowledge Engineering. 2008;997–1001. doi: 10.1109/ISKE.2008.4731075.
28. Abirami S, Chitra P. Energy-efficient edge based real-time healthcare support system. Advances in Computers. 2020;117(1):339–368. doi: 10.1016/bs.adcom.2019.09.007
29. Kumar Y, Sahoo G. Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. International Journal of Information Technology and Computer Science 2012; 4(7):43–9. doi: 10.5815/ijitcs.2012.07.06.
30. Humphris CW. Computer Science Principles V10. CreateSpace Independent Publishing Platform. 2013.
31. Saravana N, Gayathri V. Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48). International Journal of Computer Trends and Technology. 2018;59(2):73–80. doi: 10.14445/22312803/ijctt-v59p112.
32. Waseem S, Salman A, Muhammad AK. Feature subset selection using association rule mining and JRip classifier. International Journal of Physical Sciences. 2013;8(18):885–96. doi: 10.5897/ijps2013.3842.
33. Lewis RJ, Ph D, Street WC. An Introduction to Classification and Regression Tree (CART) Analysis. 2000. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf, 10 October 2021.
34. Frank E, Witten IH. Generating accurate rule sets without global optimization. Hamilton, New Zealand: University of Waikato, Department of Computer Science. 1998.
35. Kalmegh S. Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News. International Journal of Innovative Science, Engineering & Technology. 2015;2(2):438–446.
36. Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: A critical look at current methods. Analytical and Bioanalytical Chemistry. 2008;390(5):1241–1251. doi: 10.1007/s00216-007-1790-1.
37. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Morgan Kaufmann. 1995.
38. Novakovic J, Veljovi A, Iiic S, Papic Z, Tomovic M. Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science. 2017;7(1):39–46.

DOI 10.2478/ama-2022-0022

39. Aggarwal CC. [ed.] Data Classification - Algorithms and Applications, Chapman and Hall/CRC. 2014.
40. Maimon O, Rokach L. [ed.] Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Berlin, Springer. 2005.
41. Ras ZW, Dardzinska A. Action Rules Discovery Based on Tree Classifiers and Meta-actions. Lecture Notes in Artificial Intelligence. 2009;5722;66–75.
42. Ras ZW, Dardzinska A. Action Rules Discovery without Pre-existing Classification Rules. Lecture Notes in Computer Science, 2008; 5306:181-190
43. Jongbo OA. Adetunmb AO, Ogunrinde RB, Badeji-Ajisafe B. Development of an ensemble approach to chronic kidney disease diagnosis. Scientific African, 2020;8:e00456. doi: 10.1016/j.sciaf.2020.e00456
44. Senan EM, Al-Adhaileh MH, Alsaade FW, et al. Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. Journal of Healthcare Engineering. 2021;2021:1004767. doi: 10.1155/2021/1004767.

Acknowledgements: This work is supported by the Ministry of Science and Higher Education of Poland under research project No. WZ/WM-IIB/3/2021.

Andreea S. Turiac: https://orcid.org/0000-0002-1548-5640
Małgorzata Zdrodowska: https://orcid.org/0000-0003-4383-5713