This in an electronic reprint of the original article (final draft).

Chronic Kidney Disease Early Diagnosis Enhancing by using Data Mining Classification and Features Selection

Pedro A. Moreno-Sanchez [1][0000-0003-3987-1789]

[1] School of Health Care and Social Work, Seinäjoki University of Applied Sciences,
60100,
Seinäjoki, Finland
pedro.morenosanchez@seamk.fi

Abstract. Chronic Kidney Disease (CKD) is currently a worldwide chronic disease with an increasing incidence, prevalence and high cost to health systems. A delayed recognition and prevention often lead to a premature mortality due to progressive and incurable loss of kidney function. Data mining classifiers employment to discover patterns in CKD indicators would contribute to an early diagnosis that allow patients to prevent such kidney severe damage. Adopting the cross Industry Standard Process of Data Mining (CRISP-DM) methodology, this work develops a classifier model that would support healthcare professionals in early diagnosis of CKD patients. By building a data pipeline that manages the different phases of CRISP-DM, an automated data transformation, modelling and evaluation is applied to the CKD dataset extracted from the UCI ML repository. Moreover, the pipeline along with the Scikit-learn package's GridSearchCV is used to carry out an exhaustive search of the best data mining classifier and the different parameters of the data preparation's substages like data missing and feature selection. Thus, AdaBoost is selected as the best classifier and it outperforms with a 100% in terms of accuracy, precision, sensivity, specificity, f1-score and roc auc, the classification results obtained by the related works reviewed. Moreover, the application of feature selection reduces up to 12 out of 24 features which are employed in the classifier model developed.

Keywords: Chronic Kidney Disease, Early Diagnosis, Data Mining, Classification, Feature Selection.

1   Introduction

Chronic kidney disease (CKD) is a worldwide public health problem with an increasing incidence, prevalence, and high cost to health systems. Globally, in 2017, 1.2 million people died from CKD, increasing the all-age mortality rate up to 41.5% since 1990. The same year, a number of 697.5 million cases of all-stage CKD were recorded that implies a global prevalence of 9,1% [1]. CKD is the most common type of kidney diseases that lead a vast majority of CKD patients to suffer premature mortality due to cardiovascular disease and the progressive loss of kidney function; as well as other types of kidney-injured syndromes with significant negative effects on their quality of life and survival rate [2].

Typically, CKD presents no symptoms in an early stage, but later, symptoms may appear like leg swelling, extreme fatigue and generalized weakness, shortness of breath, loss of appetite, or confusion. Slowing the progression of the kidney damage, usually by controlling the underlying causes, is the main focus of CKD treatment. A delayed recognition and prevention often lead to further kidney injury and health problems where hemodialysis or even kidney transplantation are the only way to keep the patient alive [3], [4].

However, the diagnosis of CKD is a process of 3 months where the level Glomerular Filtration Rate (GFR) is assessed, although is not practical for daily clinical use due to complexity of the measure procedure [5], [6]. Therefore, other estimation approaches of GFR, like Cockcroft-Gault equation or Modification of Diet in Renal Disease equation [7], are widely accepted by using filtration markers or risk factors which are easily collected like hypertension, obesity, heart disease, age, diabetes, drug abuse, family history of kidney disease, race/ethnicity [8]. By having the disease diagnosed at the beginning phase the corresponding treatments can be initiated and the patient can live longer even with these insufficient kidney functions.

However, an opaque relationship between CKD and various symptoms exists, thus, data mining is appropriate to discover the latent correlation between them contributing significantly to assess individuals with potential CKD risk. Data mining provides useful tools for multivariate data analysis, namely classification and regression, allowing predictions based on the established models and hence offering a suitable advantage for risk assessment of many diseases including CKD [9]. Therefore, as early detection and proper treatments are the cornerstone to prevent CKD, automated and accurate diagnosis methods of CKD based on data mining are necessary to assist medical personnel to early discover patients at risk and so increase their quality of life expectation.

Large amounts of complex data are being generated by healthcare stakeholders about patients, diseases, hospitals, medical equipment, claims, treatment cost, etc. that requires processing and analysis for knowledge extraction [10]. Machine learning and data mining had been successfully applied, over the past few decades, to build computer-aided diagnosis (CAD) systems for diagnosing complex health issues with good accuracy and efficiency by recognizing potentially useful, original, and comprehensible patterns in health data [11], [12]. Data Mining is particularly useful in medicine when no availability of evidence favoring a certain treatment option is found. Classification is a data mining technique, which belongs to supervised learning methods, with the primary objective of forecasting target classes precisely and accurately for a given case.

This paper aims at enhancing the quality of CKD early diagnosis by developing an automated and accurate classifier model of CKD patients based on data preprocessing and feature selection techniques, as well as an exhaustive search of the best data mining classifier. The main contributions achieved are: a data management pipeline that provides an automated control of classification task and its previous data preparation; a classifier model that outperforms the related works reviewed not only in the training but also in testing phase with new unseen data; and a reduced group of features from the original dataset which are employed by the model to obtain high accurate results in classifying CKD patients.

The next sections of the paper are organized as follows: Section II shows related works in the CKD diagnosis field, Section III discusses the methodology employed to build the classifier model, Section IV and V shows and discusses the results obtained respectively, and Section VI points the conclusions drawn in this research.

2   Related Works

Several data mining approaches have been considered for the detection of CKDs in the literature dealing with either medical images or clinical indicators. In these works, different classifiers have been mainly used such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees (DT), Näive Bayes (NB), Random Forest (RF), Ensemble Learning (Adaboost, Bagging, etc.), and Artificial Neural Networks (ANN).

Despite good accurate results achieved in detecting CKD through data mining classifiers by Chiu et al. (94,75% accuracy) [13], Baby et al. (100% accuracy) [14], or Lakshim et al. (93.85% accuracy) [15]; a comparison cannot be carried out due to different datasets employed in the classification task. However, other different studies that employed, like this research, the CKD dataset from UCI repository [16] are described as following with the aim at comparing them to our results.

Different classifiers as Radial Basis Function (RBF), LR and Multilayer perceptron (MLP) were assessed by Rubini et al. [17] being the MLP the best one with results as: 99.75% accuracy, 99.66% F1-score, 99.33% recall and 100% specificity. Ani et al. [18] built a clinical decision support system for CKD risk prediction comparing several clas-sifier and ranking its accuracy: BackProp neural network (81.5%), NB (78%), LDA (76%), KNN(90%), DT (93%), and Random subspace classification algorithms (94%). Other classifiers (KNN, SVM, LR, DT) were explored by Charleonnan et al. [19], being SVM the most accurate (98,3%) with a sensivity of 99% and specificity of 98%. Chen et al. [2] also demonstrated SVM had better accuracy (99.7%) over other methods as KNN or soft independent modeling of class analogy (SIMCA) in classifying CKD. In their research, Kunwar et al. [10] showed that NB outperforms ANN in accuracy (100% over 72.73%). Jeewantha et al. applied the percentage split method on the dataset, demonstrating most classifiers have better accuracy when percentage of training data is higher, with the MLP as the most accurate model (98.66%). The only study identified where cross-validation technique were not applied was performed by Imran et al. [11] obtaining a 99% of F1-score, precision, recall and area under the curve ROC (Roc Auc) with a model based on Feedforward neural networks over unseen samples of the test set. In addition, Van Eyck et al. [20] achieved in 2016 the best results so far with a 100% in terms of accuracy, precision, sensivity and specificity by using RF.

On the other hand, other studies explored the influence of feature selection in the classifiers result. Thus, Chetty et al. [21] applied different classifiers along with wrapper feature selection methods demonstrating that the classifiers tested performed better on reduced dataset than the original with an accuracy of 100% by using best first search strategy in wrapper feature selection and KNN classifier. Salekin et al. [7] found RF had better accuracy (99%) than KNN and ANN when wrapper feature selection or Lasso with 12 or 10 features respectively was applied. The combination of RF with feature selection as the most accurate (99.75%) was confirmed by Siyad et al.[22] among other as NB (97.5%), LR (98%) or DT(98%) . Feature selection was also tested by Basar et al. [23] on ensemble classifiers like AdaBoost, Bagging or Random Sub-spaces, being the latter the best one with 100% of accuracy by considering only 10 features of the original UCI dataset. Wibawa et al. [24] added another research to works on testing ensemble classifiers with feature selection, having an accuracy of 98,1% and 98% as F-score, prediction and recall in a resultant dataset of 17 features with Ada- Boost-KNN classifiers. In the same line, Zubair et al. [25] obtained an accuracy of 99% by using AdaBoost classifier plus ExtraTree to select the 13 most important features.

Table 1. Classification results (expressed in %) of related works. *: *cross-validation technique not applied*

Article	Accu-racy	F1-Score	Precision	Specificity	Recall	Roc Auc
Rubini [17]	100	100	-	100	99	-
Basar [23]	100	-	-	-	-	-
Van Eyck [20]	100	-	100	100	100	100
Ani [18]	94	95	97	-	93	-
Chen [2]	100	-	-	-	100	-
Chetty [21]	100	-	-	-	-	-
Kunwar [10]	100	-	-	-	-	-
Jeewantha [8]	99	-	-	-	-	100

Salekin [7]	-	99	-	-	-	-
Wibawa [24]	98	98	98	-	98	-
Zubair [25]	99	99	-	-	-	-
Charleonnan [19]	98	-	-	98	99	-
Siyad [22]	100	-	-	-	-	-
Imran [11]*(*)	99	97	-	-	99	99

As Table 1 shows, the results obtained by the different studies are almost perfect in terms of accuracy (values close to 100%). However, it must be noted that all papers reviewed, except one (Imran et al.[11]), performed the cross-validation technique to obtain their results. This technique allows using every sample of the dataset to train the model. Only Imran's model was performed over unseen data samples. Therefore, the rest of models' performance would be unknown in a deployment phase with data that have not been used for training.

3 Material and methods

For this study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) has been adopted [26]. CRISP-DM gives a methodological way to manage data mining development. As shown in Fig. 1, CRISP-DM establishes a continuous loop composed of 6 steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.

Fig. 1. CRISP-DM methodology [26].

With the aim at improving the automation and efficiency in building the classifier model as well as deploying it in a real-world scenario, developers usually combine the phases of data preparation, modelling and evaluation into a data management pipeline that controls the data flow through all algorithms applied.

3.1 Business Understanding

This stage is the most important because the intention of the project is outlined here. The main objective of this research work is to achieve a data mining model that guarantees a highly accurate and efficient classification of CKD patients.

3.2 Data Understanding

This step begins with an underlying data gathering and continues with actions to facilitate the understanding of what the project wants and needs in terms of data.

As mentioned before, the CKD dataset used in this research was extracted from the UCI Machine Learning Repository [16]. The data set, collected from the Apollo Hospitals, Karaikudi, India during a nearly 2-month period in 2015, includes a total of 400 samples depicted by 11 numeric, 13 nominal attributes and a class attribute (ckd/notckd). Out of 400 samples, 250 samples belonged to the CKD group (62.5%), and the other 150 samples to the non-CKD group implying an imbalanced dataset. Table 2 lists the attributes from the original data set. It must be noted every attribute contained missing values except the class attribute, due to possibly to the fault of the receiver input, sensor error or reluctance on data resource. The indicators considered in this dataset are feasible to collect [7] in clinical routine favoring an early diagnosis of CKD.

Table 2. Attributes description of CKD dataset

	Attributes [*Acronym*]	Indication	Average / nominal values
1	Age (year) [*age*]	Numerical	51.5 (avg)
2	Blood pressure (mm/Hg) [*bp*]	Numerical	76.5 (avg)
3	Specific gravity [*sg*]	Nominal (1.005,1.010, 1.015, 1.020, 1.025)	7, 84, 75, 106, 81
4	Albumin [*al*]	Nominal (0,1,2,3,4,5)	199,44,43,43,24,1
5	Sugar [*su*]	Nominal (0,1,2,3,4,5)	290,13,18,14,13,3
6	Red blood cells[*rbc*]	Normal or abnormal	47 abnormal
7	Pus cell [*pc*]	Normal or abnormal	76 abnormal
8	Pus cell clumps [*pcc*]	Present or not present	42 present
9	Bacteria [*ba*]	Present or not present	22 present
10	Blood glucose random (mgs/dl) [*bgr*]	Numerical	148.04 (avg)
11	Blood urea (mgs/dl) [*bu*]	Numerical	57.43 (avg)
12	Serum creatinine (mgs/dl) [*sc*]	Numerical	3.07 (avg)
13	Sodium (mEq/l) [*sod*]	Numerical	137.53 (avg)
14	Potassium (mEq/l) [*pot*]	Numerical	4.63 (avg)
15	Hemoglobin (gms) [*hemo*]	Numerical	12.53 (avg)
16	Packed cell volume [*pcv*]	Numerical	38.88 (avg)
17	White blood cell count (cells/cumm) [*wc*]	Numerical	8406.12 (avg)
18	Red blood cell count (cells/ cumm) [*rc*]	Numerical	4.71 (avg)
19	Hypertension [*htn*]	Yes or no	147 yes
20	Diabetes mellitus [*dm*]	Yes or no	137 yes
21	Coronary artery disease [cad]	Yes or no	34 yes
22	Appetite [*appet*]	Good or poor	82 poor
23	Pedal edema [*pe*]	Yes or no	76 yes
24	Anemia [*ane*]	Yes or no	60 yes
25	Target class	ckd or notckd	250 ckd

3.3 Data preparation

Once the data has been collected, it must be transformed or preprocessed into a usable subset by checking for questionable, missing or ambiguous cases.

Missing value imputation is one of the important tasks in data mining especially in the cases where the data is small and there is a need of using all available data, as occurs with CKD dataset [9]. For handling data missing values several approaches can be followed depending on the type of attribute or feature. Regarding numerical features, replacement can be done by Bayesian imputation with median or mean of the rest of feature's values; or applying multivariate imputation through techniques as KNN or iterative correlation among all features. In case of nominal features, the most common approach is to substitute missing value for the most common value of the feature.

Relatively many features can overload the classifier contributing negatively towards the calculation of the classification as well as increasing the computational time. Feature subset selection aims to reduce computing time and improve the results of prediction by removing the features/attributes in a dataset that are considered unimportant or unable to contribute to accuracy of the classification [17]. Features selection method also depends on the input feature category and the target class' category, although there are methods that can be used for both like mutual information or recursive feature elimination (RFE). Apart from the two latter, this study will use ANOVA and Chi-Squared(Chi^2) for numerical and nominal categories respectively.

Another technique used in data preparation is feature scaling to allow the model to process the samples of numerical features with a normalized range of values by applying for instance minmax scaling (used here) or standard scaling. On the other hand, nominal features are usually encoded into numbers to allow the model to perform correctly.

3.4   Modeling

Once data is prepared for being processed, several data mining classifiers can be applied in order to discover underlying patterns and so to gain meaningful insights. This is the purpose of data mining: to create knowledge information that has meaning and utility. Depending on the data mining tasks, models used can be classifiers or regressors. As the goal of this research was to enhance early diagnosis of CKD patients through classification, the following classifiers employed were: Logistic Regression, Support Vec tor Machine, Decision Trees, Random Forest, Multilayer Perceptron, Naïve Bayes, K-Nearest Neighbors, and AdaBoost with Decision Tree as base classifier. These classifiers have been employed in related works described previously, and their usage will allow to compare the performance of the model developed in this research.

3.5   Evaluation

Classifier model selection must be done by dealing with a portion of the data and adjustments are made if necessary. Therefore, splitting the training set is recommended in this phase to divide in into a training subset, to decide which model performs better, and a validation subset, to tweak the hyperparameters of the selected model refining the classification accuracy. The k-fold cross-validation technique allows using each sample of the dataset for training k-1 times and testing 1 time [27]. Therefore, the variance of classification result can be minimized. However, the dataset should have been previously split to save a test set with the aim to run the model on unseen data, thus, ensuring new samples will be classified as expected in the next deployment phase. In this research, a test set is firstly generated, and the cross-validation technique is used on training set to select the model and the parameters of the data preparation phase

To estimate classification performance, several metrics are used in this research, namely: accuracy, precision, recall/sensitivity, specificity, f1-score and roc-auc. Accuracy describes the rate of true predictions and it is suitable for balanced data among classes. However, because the data on CKD dataset is not balanced, the other metrics will be used to assess the model classifier. As following, the formulas of the measures used are shown considering the acronyms depicted in the confusion matrix.

Table 3. Confusion matrix layout

		Predicted class	
		0	1
Actual class	0	TN (True Negative)	FP (False Positive)
	1	FN (False Negative)	TP (True Positive)

Table 4. Classification metrics formula

Classification Metrics	Formula
Accuracy: the overall success rate of true prediction	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$ (1)
Sensitivity/Recall: fraction of positive instances predicted correctly	$\frac{TP}{(TP+FN)}$ (2)
Specificity: fraction of negative instances predicted correctly	$\frac{TN}{(FP+TN)}$ (3)
Precision: fraction of true positive data given all true predicted data.	$\frac{TP}{(TP+FP)}$ (4)
F1-Score: harmonic mean from precision and recall	$2 * (\frac{Precision*Recall}{Precision+Recall})$ (5)
Roc-Auc: Area under curve ROC (Receiver Operating Characteristic). Values between 0 and 1 and higher values imply better classification performance.	

3.6 Deployment

This stage is envisioned to put the selected model to perform on new data in a production environment in line with the project's objectives. Concerning this research, in this phase the model selected would be performed in clinical routine. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

3.7 Data mining software

In this research, Python [28] has been used as language programming along with scikit-learn package [29] that allows to develop every stage of the CRISP-DM methodology. In particular, by using the scikit-learn's module GridSearchCV, multiple combinations of classifiers, data missing imputation, scaling and feature selection techniques have been tested to find the best model to classify CKD.

4 Results

4.1 CKD Classifier experimental setup

As mentioned before, developers are encouraged to build pipelines that manage the data operations tackled in the data preparation, modelling and evaluation phases of CRISP-DM methodology.
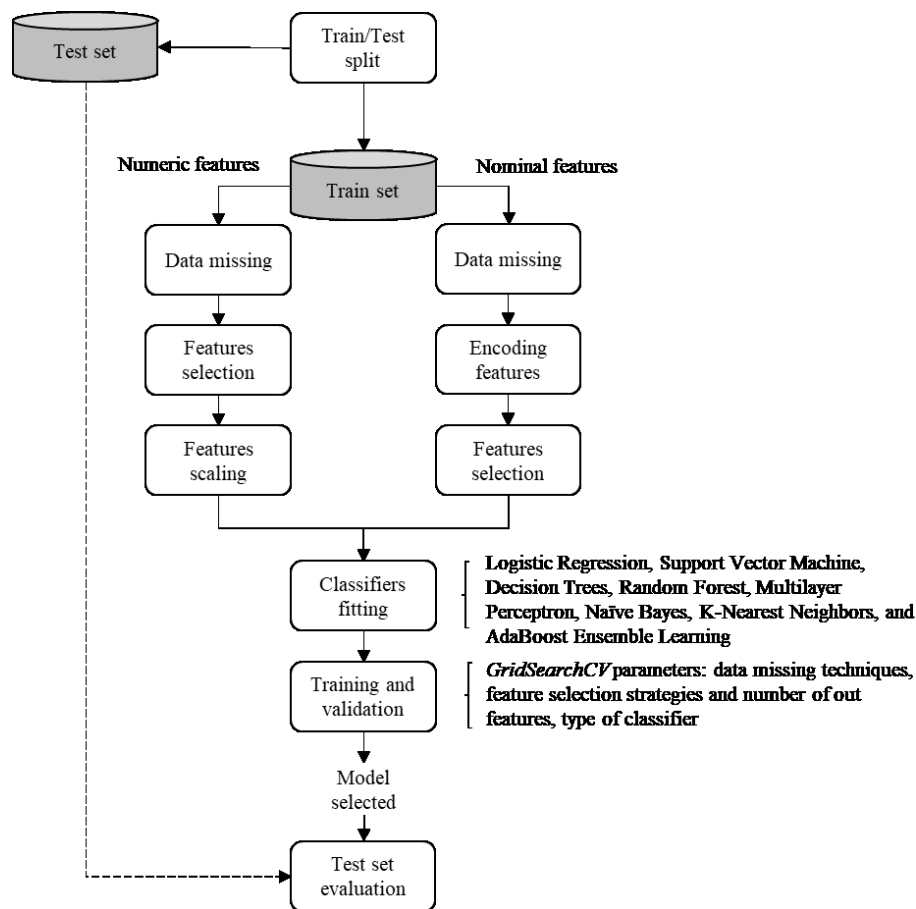
Fig. 2. CKD classifier model's pipeline

Figure 2 shows the different steps of the data management pipeline developed in this research that allows to find the best model to classify CKD patients. As a previous step, the original dataset was split for training (70% i.e. 280 samples) and testing (30% i.e. 120 samples), maintaining the same proportion of ckd/non-ckd in both sets. Next, the first step of this two-branch pipeline entails the separation of numerical and nominal features. Regarding numeric features, data missing techniques are applied first and then continuing with selection of those relevant features and a further scaling applying mixmax normalization. Data missing is also tackled first in nominal features and before applying feature selection, these features are encoded into numbers to ensure a correct performance in further steps. In order to select the model which performs the best classification of CKD patients, the classifier is trained and then validated by using 5-fold cross-validation. For that purpose, the scikit-learn's module GridSearchCV was employed since it allows to find the best classifier applying cross-validation as well as trying a grid of parameters for every stage of the pipeline. Finally, samples from the test set were used with the best model found to evaluate its classification performance with unseen data. This last evaluation gives a real notion about the selected model's performance with new data (i.e. data not used for training).

GridSearchCV allows developers to find the best combination of a model's parameters by applying an exhaustive search with multiple candidates generated from a predefined grid of parameters. Therefore, in this research the parameters needed to be optimized for the best resultant model corresponded to data missing techniques, feature selection strategy and its number of output features, as well as the type of data mining classifier employed. Table 5 shows the different values considered for these parameters.

Table 5. GridSearch CV parameters employed

GridSearchCV parameters	Values
Data missing strategy for numeric features	Mean, median, KNN, iterative.
Feature selection strategy	ANOVA (only numeric features), Chi^2 (only nominal features), mutual information, Recursive Feature Elimination (RFE).
Number of output features	1 to 11 for numeric; 1 to 13 for nominal.
Classifiers	Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, Multilayer Perceptron, Naïve Bayes, K-Nearest Neighbors, and AdaBoost with decision tree as base classifyier.

4.2 CKD Classification and feature selection results

According to GridSearchCV results, the best model's parameters found were: median for data missing strategy; RFE and 4 output features for numeric features; RFE and 5 output features for nominal features; and AdaBoost classifier. The first 3 best combination found by GridSearchCV results are shown in Table 6.

Table 6. Best 3 models found by GridSearchCV (all cells expressed in %)

Best Model *(classifier, data missing, numeric feature selection, nominal feature selection)*	Accu-racy	F1-Score	Pre-cision	Specifi-city	Recall	Roc Auc
AdaBoost, median, RFE#4, RFE#5	100	100	100	100	100	100
AdaBoost, median, ANOVA#7; Chi^2#7	99.64	99.53	99.09	99.42	100	99.71
AdaBoost, median, ANOVA#7, RFE#5	99.64	99.51	100	100	99.04	99.52

Moreover, Table 7 shows a comparison of the cross-validation results using the different classifiers but maintaining the best model's parameters related to data missing and feature selection.

Table 7. Comparison of best model parameters with all classifier considered (all cells expressed in %)

Classifier	Accu-racy.	F1-Score	Pre-cision	Specifi-city	Recall	Roc Auc
AdaBoost	100	100	100	100	100	100
Random Forest	99.29	99.00	100	100	98.10	99.05
Multilayer Perceptron	98.57	98.16	96.44	97.71	100	98.86
Logistic Regression	98.21	97.70	95.53	97.14	100	98.57
Decision Trees	98.21	97.67	97.50	98.29	98.10	98.19
Support Vector Machine	97.86	97.29	94.85	96.57	100	98.29
K-Nearest Neighbors	97.86	97.29	94.85	96.57	100	98.29
Naïve Bayes	95.71	94.61	89.78	93.14	100	96.57

The final step of the pipeline proposed in this research entailed the evaluation of the best model achieved on the test set's samples to see its performance with new unseen data. The 3 best models extracted in GridSearchCV results were evaluated in these conditions and classification results are shown in Table 8:

Table 8. 3 Best models found by GridSearchCV evaluated on test set (all cells expressed in %)

Best Model	Accu-racy	F1-Score	Pre-cision	Specifi-city	Recall	Roc Auc
AdaBoost, median, RFE#4, RFE#5	98.33	98.34	98.40	98.33	98.83	98.67
AdaBoost, median, ANOVA#7; Chi^2#7	99.17	99.17	99.18	99.17	99.17	99.33
AdaBoost, median, ANOVA#7, RFE#5	100	100	100	100	100	100

The confusion matrix of these models by using the samples of test set (120 samples) are depicted in Table 9.

Table 9. Confusion matrix of best selected models with samples of test set.

			Predicted class	
			ckd	*notckd*
AdaBoost, median, RFE#4, RFE#5	**Actual class**	*ckd*	73	2
		notckd	0	45
AdaBoost, median, ANOVA#7; Chi^2#7	**Actual class**	*ckd*	74	1
		notckd	0	45
AdaBoost, median, ANOVA#7, RFE#5	**Actual class**	*ckd*	75	0
		notckd	0	45

In addition, the best models indicated that only 9,12 and 14 out of 24 features were considered as relevant to achieve such results. The features selected for the 3 best models are shown in Table 10.

Table 10. Features selected in the 3 best models found by GridSearchCV

Best Models	Numeric features	Nominal features
AdaBoost, median, RFE#4, RFE#5	Serum creatinine, Potassium, Hemoglobin, Red blood cell count.	Specific gravity, Albumin, Hypertension, Diabetes mellitus, Pedal edema.
AdaBoost, median, ANOVA#7; Chi^2#7	Blood glucose random, Blood urea, Serum creatinine, Sodium, Hemoglobin, Packed cell volume, Red blood cell count.	Specific gravity, Albumin, Sugar, Hypertension, Diabetes mellitus, Appetite, Pedal edema.
AdaBoost, median, ANOVA#7, RFE#5	Blood glucose random, Blood urea, Serum creatinine, Sodium, Hemoglobin, Packed cell volume, Red blood cell count.	Specific gravity, Albumin, Hypertension, Diabetes mellitus, Pedal edema.

5   Discussion

The pipeline developed in this research offers the possibility to automate not only the training and testing of the model but also the searching of best parameters involved in the data preparation phase as well as the data

mining classifier employed. Furthermore, this pipeline would manage classification of new samples in case it appeared, aswell as the consequent model retraining and adaptation to new incoming data.

The classification results achieved by this research after applying cross-validation technique through GridSearchCV manifested that the classifier AdaBoost performed a better classification task compared to other classifiers considered. Moreover, such classifier along with the parameters selected by GridSearchCV (median, RFE#4, RFE#5) obtained results of 100% in terms of accuracy, precision, sensivity, specificity, f1-score and Roc Auc. Compared to the results from other related works, this research reached the most accurate figures so far like research developed by Van Eyck et al [20].

However, the best model selected with cross-validation did not perform as the best with the new data belonging to the test set. Here, a clear example of overfitting existedsince the best trained model was not the most accurate in classifying new unseen data. Consequently, other model that involved a bigger number of selected features (ANOVA#7, RFE#5) was evaluated and it classified better since the new information of features added allowed to achieve results of 100% in every classification metric considered. For the best of our knowledge, this research outperforms the rest of studies developed in CKD patients classification so far, because not only equalizing the best model obtained by using cross validation, but also achieving a perfect classification with unseen data which has not been found in any related work. The split of the datasetinto a training/validation subset, on one side, and test subset on the other, with a ratio of 70/30 could negatively affect the model performance since a small group of samples were dedicated for training. However, the classification results demonstrated the decision made about developing the pipeline and using the cross-validation strategy developed was correct.

Moreover this study contributes to the state-of-art by proposing a reduced group among the entire dataset's features with several implications: higher feasibility in classifying CKD patients since the number of features to be collected are lower; and a decreasing cost to healthcare systems as extracting less clinical indicators proposed bysuch features selected.

Due to the fact that the classifier selected is AdaBoost with Decision Tree as base classifier, an exploration of features importance in the classification task could be carried out with the aim at giving healthcare professionals an easier understanding and interpretability of the outcomes generated by the model. By doing so, not only would clinicians achieve an early diagnosis with a reduced group of indicators but also, they could focus on treatment for those important features to the risk of suffering CKD or even revert the disease progress returning to an earlier CKD stage.

6   Conclusion

This article shows a development of a classifier model aimed at early diagnosis of Chronic Kidney Disease (CKD) patients. CKD is a worldwide chronic disease with an increasing incidence that leads patients to a premature mortality if it is detected in later stages. A review of the related works has been carried out by depicting the classification results achieved by different authors. The CRISP-DM methodology has been adopted in the classifier model development to ensure data is properly processed. Moreover, a data management pipeline has been developed for automating all stages of data preparation, modeling and evaluation. After applying cross-validation technique through scikit-learn package's GridSearchCV, the best model comprises AdaBoost, as best classifier; and median, RFE#4, RFE#5 as best data preparation's parameters. Next, this best model is also tested with new unseen data by using the test set that has been previouslysplit from the original dataset before using the pipeline developed. Moreover, an exploration of the features selected during the data preparation phase are carried out to depict those dataset's attributes that contribute to the model performance. A case of overfitting is identified since the best trained model

performs worse than the other model with more features selected when dealing with unseen data in the testing phase. To the best of our knowledge, the classification results obtained either in cross-validation or in testing phase outperforms the existing results of the related works reviewed.

References

1. B. Bikbov *et al.*, "Global, regional, and national burden of chronic kidney disease, 1990– 2017: a systematic analysis for the Global Burden of Disease Study 2017," *The Lancet*, vol. 395, no. 10225, pp. 709–733, Feb. 2020, doi: 10.1016/S0140-6736(20)30045-3.
2. Z. Chen, X. Zhang, and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," *Int Urol Nephrol*, vol. 48, no. 12, pp. 2069–2075, Dec. 2016, doi: 10.1007/s11255-016-1346-4.
3. D. S. Keith, G. A. Nichols, C. M. Gullion, J. B. Brown, and D. H. Smith, "Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization," *Arch. Intern. Med.*, vol. 164, no. 6, pp. 659–663, Mar. 2004, doi: 10.1001/archinte.164.6.659.
4. A. Levin *et al.*, "Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in patients with chronic kidney disease: results of the study to evaluate early kidney disease," *Kidney Int.*, vol. 71, no. 1, pp. 31–38, Jan. 2007, doi: 10.1038/sj.ki.5002009.
5. M.-T. Liao, C.-C. Sung, K.-C. Hung, C.-C. Wu, L. Lo, and K.-C. Lu, "Insulin Resistance in Patients with Chronic Kidney Disease," *Journal of Biomedicine and Biotechnology*, Aug. 07, 2012. https://www.hindawi.com/journals/bmri/2012/691369/ (accessed Aug. 05, 2020).
6. M. A. Perazella and R. F. Reilly, "Chronic kidney disease: a new classification and staging system," *Hospital Physician*, vol. 39, no. 3, pp. 18–22, 2003.
7. A. Salekin and J. Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, Oct. 2016, pp. 262–270, doi: 10.1109/ICHI.2016.36.
8. R. A. Jeewantha, M. N. Halgamuge, A. Mohammad, and G. Ekici, "Classification Performance Analysis in Medical Science: Using Kidney Disease Data," in *Proceedings of the 2017 International Conference on Big Data Research*, Osaka, Japan, Oct. 2017, pp. 1–6, doi: 10.1145/3152723.3152724.
9. K. Kumar and B. Abhishek, *Artificial neural networks for diagnosis of kidney stones disease*. GRIN Verlag, 2012.
10. V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 300–305, doi: 10.1109/CONFLUENCE.2016.7508132.
11. A. A. Imran, M. N. Amin, and F. T. Johora, "Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide Deep Learning," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, Dec. 2018, pp. 1–6, doi: 10.1109/CIET.2018.8660844.
12. S. Dhamodharan, "Liver Disease Prediction Using Bayesian Classification," p. 3, 2014.
13. R. K. Chiu, R. Y. Chen, Shin-An Wang, and Sheng-Jen Jian, "Intelligent systems on the cloud for the early detection of chronic kidney disease," in *2012 International Conference on Machine Learning and Cybernetics*, Jul. 2012, vol. 5, pp. 1737–1742, doi: 10.1109/ICMLC.2012.6359637.
14. P. S. Baby and T. P. Vital, "Statistical analysis and predicting kidney diseases using machine learning algorithms," *International Journal of Engineering Research and Technology*, vol. 4, no. 7, 2015.
15. K. Lakshmi, Y. Nagesh, and M. V. Krishna, "Performance comparison of three data mining techniques for predicting kidney dialysis survivability," *International Journal of Advances in Engineering & Technology*, vol. 7, no. 1, p. 242, 2014.
16. D. Dua and C. Graff, *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017.

17. L. J. Rubini and P. Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease," *International Journal of Modern Engineering Research (IJMER)*, vol. 5, no. 7, pp. 49–55, 2015.

18. R. Ani, G. Sasi, U. R. Sankar, and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 1287–1292, doi: 10.1109/ICACCI.2016.7732224.

19. A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, Oct. 2016, p. MIT-80-MIT-83, doi: 10.1109/MITICON.2016.8025242.

20. J. V. Eyck *et al.*, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," 2016. /paper/Prediction-of-Chronic-Kidney-Disease-Using-Random-Eyck-Zadeh/c8f5ed96b924f00c729a1a3ff79ead91a8418dc7 (accessed Jul. 30, 2020).

21. N. Chetty, K. S. Vaisla, and S. D. Sudarsan, "Role of attributes selection in classification of Chronic Kidney Disease patients," in *2015 International Conference on Computing, Communication and Security (ICCCS)*, Dec. 2015, pp. 1–6, doi: 10.1109/CCCS.2015.7374193.

22. B. MohammedSiyad and Manoj, "Fused Features Classification for the Effective Prediction of Chronic Kidney Disease," 2016.

23. M. D. Basar and A. Akan, "Detection of chronic kidney disease by using ensemble classifiers," in *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2017, pp. 544–547.

24. M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2017, pp. 1–6, doi: 10.1109/CITSM.2017.8089245.

25. K. M. Zubair Hasan and Md. Zahid Hasan, "Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease," in *Emerging Re-search in Computing, Information, Communication and Applications*, Singapore, 2019, pp. 415–426, doi: 10.1007/978-981-13-5953-8_34.

26. R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," p. 11.

27. T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Stat Comput*, vol. 21, no. 2, pp. 137–146, Apr. 2011, doi: 10.1007/s11222-009-9153-8.

28. T. E. Oliphant, "Python for Scientific Computing," *Computing in Science Engineering*, vol. 9, no. 3, pp. 10–20, May 2007, doi: 10.1109/MCSE.2007.58.

29. F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*, p. 6.