Data Mining Approaches for Early Prediction and Management of Chronic Kidney Disease

Beatriz Amorim[1†], Carolina Santos[1†], Catarina Nunes[1†]

[1]Universidade do Minho, Mestrado em Engenharia Biomédica - Informática Médica.

Contributing authors: pg56112@alunos.uminho.pt; pg56116@alunos.uminho.pt; pg56117@alunos.uminho.pt;
[†]These authors contributed equally to this work.

Abstract

Chronic kidney disease (CKD) poses a significant global health challenge, requiring accurate and early detection. In this study, we investigate the relationship between clinical features such as hemoglobin levels, blood urea levels and packed cell volume with CKD, applying advanced machine learning techniques to analyze the data. Our analysis highlights the effectiveness of the Random Forest model, which achieved an impressive accuracy of 98.5%. This high performance underscores the importance of pre-processing and feature selection, which played a crucial role in enhancing the model's predictive capabilities. These findings suggest that machine learning approaches, particularly Random Forest, have great potential for advancing CKD prediction, enabling earlier intervention and more personalized treatment plans. The results also emphasize the need for further validation of these techniques in diverse populations to ensure their generalization and clinical applicability.

Keywords: Chronic Kidney Disease, Data Mining, Feature Selection, Data Analysis

1 Introduction - Business Understanding

Chronic Kidney Disease (CKD) is a progressive condition affecting millions of people worldwide and it is most commonly attributed to diabetes and hypertension [1]. This disease is characterized by the gradual and irreversible loss of kidney function. Early diagnosis and effective management of CKD are critical for improving patient quality

of life and reducing associated complications. However, early detection of CKD remains challenging due to its often asymptomatic nature in the initial stages [2].

CKD is closely linked to cardiovascular diseases (CVD), including coronary artery disease and heart failure. The prevalence of cardiovascular events is significantly higher in patients in early CKD stages compared to the general population, and this risk escalates markedly in advanced CKD stages [3].

Given this context, this project follows the CRISP-DM methodology to address the problem of CKD detection and progression prediction. The general business objective is to enhance understanding of CKD progression and improve decision-making in clinical settings. The specific data mining objectives include preprocessing the dataset, applying appropriate models for pattern recognition, and evaluating the performance of these models to ensure reliability, in order to correctly predict the occurrence and development of CKD.

2 State of the art

Numerous studies highlight the importance of early CKD detection. As mentioned, early-stage CKD is often asymptomatic, making timely diagnosis difficult. Traditional diagnostic approaches rely on clinical markers such as glomerular filtration rate (GFR) and albumin [1]. However, these markers alone may not adequately capture the complexity of CKD progression [4]. Over the past decades, extensive research has focused on identifying new biomarkers in blood and urine to improve diagnostic and prognostic precision [5].

This has led to increased interest in using data-driven approaches to identify patterns in patient data that could indicate early disease.

A critical factor influencing the success of machine learning models is the preprocessing of data. Effective handling of missing values and class imbalance is essential for robust model performance. In this context, research leveraging KNN imputation and SMOTE preprocessing has demonstrated the significance of addressing these challenges. For instance, in study [6], researchers combined SMOTE preprocessing with a TrioNet model — an ensemble approach combining Extreme Gradient Boosting, Random Forest, and Extra Trees classifiers — achieving outstanding results, including an accuracy of 94.88% and an F1-score of 97.96% . These metrics surpassed those of other models such as Random Forest (92.85%) and Decision Tree (87.28%), underscoring the importance of advanced preprocessing techniques and ensemble learning for CKD prediction.

Another innovative approach, study [7], utilized a hybrid model combining Gaussian Naïve Bayes, Gradient Boosting, Decision Trees, and a Random Forest meta-classifier. This methodology achieved a remarkable 100% accuracy in CKD prediction, outperforming individual classifiers such as Gaussian Naïve Bayes (93%), Gradient Boosting (99%), Decision Tree (96%) and Random Forest (98%) . The hybrid model's ability to mitigate overfitting using k-fold cross-validation, while delivering remarkable performance highlights its potential as a reliable and robust diagnostic tool.

In the study [8], a variety of ML algorithms, including Gradient Boosting, Decision Tree, K-nearest Neighbor, Random Forest, Histogram Boost, and XGBoost, were employed to predict kidney illness. The study highlighted key clinical features such as serum creatinine level, blood pressure, and age, which played an important role in enhancing prediction accuracy. The models demonstrated strong performance, with KNN and RF achieving high accuracy (98.75%) and KNN achieving 97,50%. The research emphasized the importance of these features in improving early detection, thereby reducing healthcare burdens and enhancing patient outcomes.

Recent advancements also include the adoption of explainable AI (XAI) models to enhance transparency in CKD diagnosis. The study [9] applied XAI techniques using Random Forest, Extra Trees, LightGBM, Decision Tree, Logistic Regression and ANN. It achieved high scores across all evaluation metrics, specifically, Random Forest with 99.07% accuracy, Decision Tree with 97.22%, and Logistic Regression with 99.07%. The study also utilized hyperparameter tuning (GridSearchCV) to adjuste key parameters, optimizing model performance.

3 Materials - Data Understanding and Data Preparation

3.1 Dataset Description

The dataset used in this study is the Chronic Kidney Disease dataset (CKD) [10], obtained from UCI Machine Learning Repository. It contains records of patients with relevant clinical and laboratory information for the detection and monitoring of Chronic Kidney Disease.

- Number of instances: 400
- Number of attributes: 26
- Attribute types: integer, real and nominal attributes

Table 1 provides information about each attribute, including their data types, the number of missing values, description as well as the minimum, maximum, and average values.

Table 1 Summary of Attributes in the CKD Dataset

Attribute	Type	Missing	Description	Min	Max	Average/Mode	Deviation
id	Integer	0	Identification number	0	399	199.500	115.614
age	Integer	9	Patient's age	2	90	51.483	17.170
bp	Integer	12	Blood pressure	50	180	76.469	13.684
sg	Real	47	Specific gravity	1.005	1.025	1.017	0.006
al	Integer	46	Albumin levels in urine	0	5	1.017	1.353
su	Integer	49	Sugar levels	0	5	0.450	0.450
rbc	Nominal	152	Red blood cells	-	-	Normal	-
pc	Nominal	65	Pus cell count	-	-	Normal	-
pcc	Nominal	4	Pus cell clumps in blood	-	-	Notpresent	-
ba	Nominal	4	Bacterial presence	-	-	Notpresent	-
bgr	Integer	44	Blood glucose random count	22	490	148.037	79.282
bu	Integer	19	Blood urea level	2	391	57.428	50.502
sc	Real	17	Serum creatinine level	0.4	76	3.072	5.741
sod	Real	87	Sodium level in blood	4.5	163	137.529	10.409
pot	Real	88	Potassium level in blood	2.5	47	4.627	3.194
hemo	Real	52	Hemoglobin level in blood	3.1	17.8	12.526	2.913
pcv	Integer	71	Packed cell volume in blood	9	54	38.884	8.990
wc	Integer	106	White blood cell count	2200	26400	8406.122	2944.474
rc	Real	131	Red blood cell count	2.1	8	4.707	1.025
htn	Nominal	2	Hypertension (Yes/No)	-	-	No	-
dm	Nominal	2	Diabetes mellitus (Yes/No)	-	-	No	-
cad	Nominal	2	Coronary artery disease (Yes/No)	-	-	No	-
appet	Nominal	1	Appetite	-	-	Good	-
pe	Nominal	1	Peda edema (Yes/No)	-	-	No	-
ane	Nominal	1	Anemia (Yes/No)	-	-	No	-
classification	Nominal	0	CKD diagnosis (Yes/No)	-	-	CKD	-

3.2 Data Understanding

Chronic Kidney Disease (CKD) is influenced by various health conditions and biological markers. Analyzing these factors helps to understand their relationship with CKD and its progression. This section explores key indicators and their potential impact on CKD, as well as the ways CKD may influence them.

Hypertension is a prevalent comorbidity among CKD patients. Approximately 147 individuals with hypertension, indicated by the attribute "htn", also have CKD, suggesting a strong link between elevated blood pressure and kidney damage. On the other hand, hypertension can also be an effect of CKD, as the kidneys play a crucial role in regulating blood pressure. Impaired kidney function can lead to fluid and sodium retention, contributing to elevated blood pressure, thus creating a vicious cycle that worsens both conditions. [11]

Diabetes is another key factor in the development of CKD, as individuals with diabetes are at an increased risk of renal complications, including diabetic nephropathy, which can progress to end-stage renal disease. This explains why a notable proportion of CKD patients (137 individuals) also have diabetes (attribute "dm"). Furthermore, CKD can disrupt glucose regulation, potentially leading to insulin resistance and the onset of diabetes. However, CKD can also occur in individuals without diabetes, indicating that other factors contribute to the progression of kidney disease. [12]

Blood urea levels, denoted by the attribute "bu", are commonly elevated in individuals with CKD, particularly as the disease progresses. Urea, a waste product of protein metabolism, is normally filtered and excreted by healthy kidneys. However, in CKD, as kidney function declines, the ability to eliminate urea diminishes, leading to its accumulation in the bloodstream. Urea is considered both a direct and indirect

uraemic toxin, meaning it contributes to the toxic effects seen in CKD [13]. When examining the dataset, a similar pattern of elevated blood urea levels was observed. Among 132 patients without CKD, blood urea levels ranged from 10 to 48.7 mg/dL, which falls within the 'normal' range. However, in individuals diagnosed with CKD, urea levels increased with the severity of the disease. For instance, 112 CKD patients had urea levels ranging from 40.9 to 118.7 mg/dL. A group of 87 CKD patients had normal urea levels, between 2 and 40.9 mg/dL. Furthermore, some patients had urea levels exceeding 118.7 mg/dL, which suggests more advanced kidney dysfunction.

Looking at the "hemo" attribute, it's possible to observe the hemoglobin levels in patients and analyze how they vary based on the presence of CKD. Typically, hemoglobin levels below 14 g/dL in men and 12 g/dL in women are considered low and indicative of anemia. In CKD patients, anemia is a common complication, particularly in advanced stages of the disease. It primarily occurs due to reduced erythropoietin (EPO) production, a hormone produced by the kidneys that stimulates red blood cell production in the bone marrow. When kidney function declines, EPO production decreases, leading to insufficient red blood cells, which results in anemia [14]. The dataset clearly shows a distinction in hemoglobin levels between patients with and without CKD. Non-CKD patients tend to have hemoglobin levels greater than 13 g/dL, which is within the normal values. In contrast, 147 CKD patients exhibit a wider range of hemoglobin levels, from as low as 3 g/dL to 11.92 g/dL, with only 57 CKD patients having hemoglobin levels exceeding that value 11.92 g/dL. This suggests that anemia is a prevalent issue among CKD patients. Additionally, the "hemo" attribute is closely related to the "pcv" (packed cell volume) attribute, which reflects the proportion of red blood cells in the blood. Like hemoglobin, the "pcv" values also exhibit a similar trend: lower values are seen in CKD patients, while normal or higher values are found in non-CKD individuals. This pattern further supports the association between anemia and kidney disease, as reduced red blood cell production in CKD contributes to both low hemoglobin and low packed cell volume.

Although there are additional attributes that contribute to the understanding of CKD, these were the key factors we found most relevant to discuss regarding the disease and its progression.

3.2.1 Correlation

To analyze the relationships between numerical attributes, the correlation matrix for these attributes was computed. In order to achieve this, nominal attributes were filtered out, as correlation for such attributes is not well-defined and typically results in missing values. The resulting correlation matrix is presented in Figure 1.

Attribut...	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc
age	1	0.164	-0.181	0.115	0.229	0.242	0.203	0.137	-0.097	0.059	-0.197	-0.252	0.125	-0.287
bp	0.164	1	-0.209	0.141	0.235	0.161	0.184	0.147	-0.123	0.074	-0.311	-0.330	0.018	-0.292
sg	-0.181	-0.209	1	-0.465	-0.316	-0.384	-0.316	-0.367	0.412	-0.074	0.607	0.614	-0.241	0.604
al	0.115	0.141	-0.465	1	0.286	0.389	0.449	0.398	-0.457	0.129	-0.642	-0.618	0.230	-0.622
su	0.229	0.235	-0.316	0.286	1	0.724	0.172	0.225	-0.137	0.220	-0.229	-0.242	0.187	-0.240
bgr	0.242	0.161	-0.384	0.389	0.724	1	0.152	0.118	-0.262	0.066	-0.311	-0.308	0.147	-0.291
bu	0.203	0.184	-0.316	0.449	0.172	0.152	1	0.585	-0.328	0.358	-0.613	-0.612	0.054	-0.613
sc	0.137	0.147	-0.367	0.398	0.225	0.118	0.585	1	-0.696	0.326	-0.402	-0.406	-0.004	-0.416
sod	-0.097	-0.123	0.412	-0.457	-0.137	-0.262	-0.328	-0.696	1	0.100	0.364	0.379	0.003	0.359
pot	0.059	0.074	-0.074	0.129	0.220	0.066	0.358	0.326	0.100	1	-0.134	-0.163	-0.105	-0.164
hemo	-0.197	-0.311	0.607	-0.642	-0.229	-0.311	-0.613	-0.402	0.364	-0.134	1	0.896	-0.177	0.819
pcv	-0.252	-0.330	0.614	-0.618	-0.242	-0.308	-0.612	-0.406	0.379	-0.163	0.896	1	-0.201	0.811
wc	0.125	0.018	-0.241	0.230	0.187	0.147	0.054	-0.004	0.003	-0.105	-0.177	-0.201	1	-0.168
rc	-0.287	-0.292	0.604	-0.622	-0.240	-0.291	-0.613	-0.416	0.359	-0.164	0.819	0.811	-0.168	1

Fig. 1 Correlation Matrix with colors, created using AI Studio.

Several pairs of attributes exhibited strong to very strong correlations, which are worth noting for their potential significance in both clinical assessment and predictive modeling.

The attributes with the highest correlation coefficient (module) are hemo, pcv, and rc, with correlation coefficients of 0.896, 0.819, and 0.811 between them. This strong correlation is expected, as all three are critical indicators of blood health and oxygen transport capacity. They are closely linked physiologically, with changes in one often reflecting corresponding changes in the others.

Additionally, the negative correlation between serum creatinine and sodium is particularly interesting. Serum creatinine is a key marker of kidney function decline, while sodium levels are indicative of electrolyte balance, which often becomes disrupted in CKD. Also, the inverse relationship between albumin and hemoglobin is worth highlighting. Elevated urinary albumin is a well-known indicator of CKD progression, whereas hemoglobin levels tend to decrease due to anemia, a common complication associated with kidney dysfunction.

On the other hand, certain attributes such as potassium (pot), white blood cell count (wc), blood pressure (bp), and age exhibited weak or near-zero correlations with most other variables in the dataset.

This observation may suggest that these attributes are less directly involved in the relationships with kidney function as compared to other variables. However, it is important to note that correlation does not imply causality and that correlations can exist without carrying any qualitative meaning. Therefore additional analysis will be conducted to better understand the true nature of these relationships.

Moreover, the correlation between the attributes and the label was calculated, revealing that ”hemo” exhibited the highest value, with a coefficient of 0.730. This was followed by ”sg” at 0.699, pcv at 0.690, and ”rc” at 0,591.

3.2.2 Association Rules

A search for association rules was also carried out, aiming to find relationships between other clinical conditions, such as diabetes mellitus, coronary artery disease, hypertension or anemia, that may exist on their own, but that can also present as symptoms of CKD or increase its risk.

Using AI Studio and the FP-Growth operator, this analysis was conducted, discovering very interesting associations between hypertension and diabetes mellitus, as well between these two comorbidities and CKD. These associations were shown to have a minimum confidence percentage of 87%, being supported by at least 30% of the observations and, in the case of the association between hypertension and history of diabetes, this was supported by more than 50% of the records in the dataset. It is important to highlight that these high levels of support may be the result of missing values treatment by replacing them with the mode of the attribute in question. However, as can be seen in Table 1, the attributes mentioned here have very few missing values and, even when these records were removed, the association between these conditions remained. Additionally, the Lift value for each of these associations was greater than 1, demonstrating a positive influence between these factors.

These results corroborate the potential relationships between these attributes explored in previous sections, demonstrating that patients suffering from hypertension often also exhibit diabetes, and that both of these medical conditions are quite common in cases of CKD, either as comorbidities or symptoms. Finally, it is important to note that this does not mean that all patients suffering from CKD have or will develop these conditions, but rather that these two conditions are warning signs or possible risk factors for developing CKD.

3.3 Data Preparation

To prepare the data for use in subsequent models, a comprehensive preprocessing pipeline was implemented. This process resulted in the creation of four distinct datasets, each prepared using different strategies. These datasets will be used to evaluate which preparation approach yields the best performance for the models.

First, the data was loaded from a CSV file. An initial inspection identified that the "pcv" attribute was incorrectly classified as categorical. This was corrected by converting it to a numeric type. Additionally, the "cad" and "dm" attributes contained extra tab characters, which were removed to ensure consistency.

3.3.1 Label Class Distribution

The class distribution of the target variable, "classification," was analyzed to identify any potential imbalance. The analysis revealed that the dataset consisted of 62.5% cases labeled as "ckd" and 37.5% as "notckd". This represents a notable class imbalance, which can introduce bias into the machine learning models. Such imbalances may result in models that are more likely to predict the majority class, thereby compromising the detection accuracy for the minority class. To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied later in the pipeline to generate synthetic samples for the minority class, ensuring a more balanced distribution and improving the model's ability to generalize.

3.3.2 Handling Missing Values

Subsequently, the missing values in the dataset were examined, and their proportion per attribute was calculated. Two distinct approaches were applied to handle missing data, resulting in the first two datasets:

1. Substitution with mean/mode: Missing values in numerical attributes were replaced with the mean, while categorical attributes were filled with the mode. This method ensured no data loss while maintaining statistical consistency.
2. Attribute Removal and Substitution: Attributes with more than 30% missing data were removed, and missing values in the remaining attributes were handled using the same mean/mode substitution strategy. This approach aimed to minimize the impact of attributes with excessive missing data. The attributes removed due to excessive missing data were "rbc" and "rc".

3.3.3 Outliers Treatment

Afterwards, outliers in numerical attributes were addressed using the interquartile range (IQR) method. Any values outside the defined bounds were capped at the lower or upper limits. This step was applied to both datasets generated in the previous stage, producing two additional variants.

3.3.4 Normalization

Finally, all datasets were scaled or normalized to improve compatibility with machine learning algorithms. Two different transformations were applied:

- Standardization (Z-score): Attributes were scaled to have a mean of 0 and a standard deviation of 1.
- Min-Max Scaling: Attributes were normalized to a range of [0, 1].

Additionally, categorical attributes were encoded using label encoding to convert them into numeric format.

3.3.5 Feature Selection

Certain attributes, including "ba", "cad", "pot", "age", "pcc" and "wc", were excluded from the dataset due to their low attribute weights, as determined using the "Weight by Information Gain" operator in Altair AI Studio. Each of these attributes had a weight below 0.079, indicating an insignificant contribution to the predictive model's performance.

Other attributes, such as "pcv" and "sg", were also removed, despite presenting a high correlation with the target variable, due to the fact that these attributes exhibited high correlations with each other and the "hemo" attribute. When multiple attributes convey essentially the same information, it can cause redundancy within the model, making it more difficult for the algorithm to distinguish their individual contributions to the prediction. This redundancy can lead to unstable coefficients, overfitting, and reduced model interpretability.

On the other hand, the attribute "hemo" was maintained due to its higher correlation with the target variable, "classification", and its clinical significance. As "hemo" encapsulates much of the information that would be redundant in "pcv" and "sg", it allows the model to focus on the most important variables, ensuring better performance while avoiding overfitting.

The process resulted in four prepared datasets, each employing a unique combination of handling missing data techniques, outlier treatment, scaling, and feature selection. These datasets will be evaluated to determine the most effective preparation method for the models.

4 Methods - Modeling

4.1 Classification

Classification is a supervised learning technique aimed at predicting the class label of an instance based on its features. In this study, the target variable, "classification," was initially categorical, with instances labeled as either "ckd" or "notckd". To facilitate the modeling process, the label was transformed into binary numbers, as mentioned, with "ckd" encoded as 0 and "notckd" encoded as 1.

Various machine learning models were evaluated to determine the most effective approach for classifying the target variable "classification", including:

- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- Logistic Regression
- K-Nearest Neighbors (KNN)

The effectiveness of these models will be evaluated on the datasets prepared earlier, considering metrics such as accuracy, precision, recall, and F1-score. This evaluation will provide insight into how the various data preparation strategies, such as handling missing values, balancing class distribution, and feature selection, affect model performance.

In the following sections, we will discuss the results of applying these classification models, comparing their performance and determining the best approach for predicting chronic kidney disease.

4.2 Clustering

In order to evaluate the performance and behavior of this dataset with an unsupervised learning method, the K-Means Clustering algorithm was applied. Before subjecting the various datasets to this process, an extra preparation step was necessary to remove the target variable.

5 Results

5.1 Classification

As previously mentioned, four datasets resulting from different pre-processing methods are used for the tests:

- Dataset 1 (D1): employs substitution with scaling;
- Dataset 2 (D2): uses substitution with normalization;
- Dataset 3 (D3): applies the removal of values greater than 30% with scaling;
- Dataset 4 (D4): uses removal of values greater than 30% with normalization.

The results here presented for each model were obtained by calculating the previously mentioned statistic metrics based on the vales given by the confusion matrix. This matrix provides the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), which allow the calculation of metrics such as accuracy, precision, specificity, recall and F1-score.

Tables 2, 3, 4, 5 and 6 present the models that, for each DM technique, achieved the best accuracy, precision, recall, specificity and F1-score results, respectively.

Table 2 Best accuracy results for each DM technique

DM Technique	Dataset	Sampling Method	SMOTE use	Accuracy (%)
Random Forest	D1	Cross Validation	with SMOTE	98.50
SVM	D2	Holdout Sampling	with/without SMOTE	98.33
Logistic Regression	D1	Cross Validation	without SMOTE	97.75
Decision Tree	D3, D4	Cross Validation	with SMOTE	97.00
KNN	D2, D4	Holdout Sampling	with/without SMOTE	96.67

Table 3 Best precision results for each DM technique

DM Technique	Dataset	Sampling Method	SMOTE use	Precision (%)
Random Forest	D1	Cross Validation	with SMOTE	98.44
SVM	D1	Cross Validation	without SMOTE	98.05
Logistic Regression	D1	Cross Validation	without SMOTE	97.75
Decision Tree	D3, D4	Cross Validation	with SMOTE	96.62
KNN	D2	Holdout Sampling	with/without SMOTE	96.12

Table 4 Best recall (sensitivity) results for each DM technique

DM Technique	Dataset	Sampling Method	SMOTE use	Recall (%)
Random Forest	D1	Cross Validation	with SMOTE	98.40
SVM	D2	Holdout Sampling	with/without SMOTE	98.67
Logistic Regression	D2, D3	Holdout Sampling	with/without SMOTE	98.00
Decision Tree	D3, D4	Cross Validation	with SMOTE	97.20
KNN	D4	Holdout Sampling	with/without SMOTE	97.33

Table 5 Best specificity results for each DM technique

DM Technique	Dataset	Sampling Method	SMOTE use	Specificity (%)
Random Forest	D1, D2	Cross Validation	without SMOTE	99.20
SVM	D1	Cross Validation	without SMOTE	98.40
Logistic Regression	D3	Cross Validation	without SMOTE	98.00
Decision Tree	D3, D4	Cross Validation	with SMOTE	96.40
KNN	D2	Holdout Sampling	with/without SMOTE	96.00

Table 6 Best F1-Score results for each DM technique

DM Technique	Dataset	Sampling Method	SMOTE use	F1-Score (%)
Random Forest	D1	Cross Validation	with SMOTE	98.40
SVM	D2	Holdout Sampling	with/without SMOTE	98.24
Logistic Regression	D1	Cross Validation	without SMOTE	97.64
Decision Tree	D3, D4	Cross Validation	with SMOTE	96.83
KNN	D4	Holdout Sampling	with/without SMOTE	96.50

5.2 Clustering

In order to evaluate the success of the k-means algorithm in dividing the data into clusters, the two obtained clusters were compared with the target variable of the initial dataset, also producing a confusion matrix. The evaluation was performed on the four different datasets resulting from the preprocessing techniques. The performance results from these evaluations are presented in Table 7.

Table 7 Comparison of K-Means Clustering Performance across Datasets

Dataset	Precision	Recall	F1-Score	Accuracy
D1	0.90	0.86	0.86	0.86
D2	0.93	0.91	0.91	0.91
D3	0.89	0.84	0.84	0.84
D4	0.93	0.92	0.92	0.92

6 Discussion – Evaluation

6.1 Classification

According to the values presented in Tables 2, 3, 4, 5, and 6, we can conclude that different machine learning models exhibit varying performance depending on the preprocessing technique, the use of SMOTE, and the sampling method applied. The following sections discuss the results achieved by each model in detail.

6.1.1 Random Forest

Random Forest consistently delivers top-tier results across several metrics. Using Cross Validation and SMOTE with Dataset 1 (D1), it achieved the highest accuracy (98.50%), precision (98.44%), recall (98.40%), and F1-score (98.40%). Additionally, without SMOTE, it attained the best specificity (99.20%) in both D1 and D2. These results demonstrate that Random Forest is particularly effective with balanced datasets, especially when combined with substitution and scaling (D1).

6.1.2 Support Vector Machine (SVM)

SVM also exhibits strong performance, excelling in recall (98.67%) and F1-score (98.24%) in Dataset 2 (D2) using Holdout Sampling, regardless of SMOTE usage. Notably, it achieved high precision (98.05%) and specificity (98.40%) with Cross Validation and no SMOTE on D1. These results suggest that SVM is an excellent choice for detecting true positives, particularly in normalized datasets (D2).

6.1.3 Logistic Regression

Logistic Regression shows balanced performance across all metrics. It achieved accuracy (97.75%), precision (97.75%), and F1-score (97.64%) on Dataset 1 (D1) with Cross Validation without SMOTE. For recall (98.00%), it excelled in D2 and D3 using Holdout Sampling, with or without SMOTE. Furthermore, it performed well in specificity (98.00%) on D3 with Cross Validation and no SMOTE. These findings indicate that Logistic Regression is still efficient even when applied to slightly imbalanced datasets, considering that the best results were achieved when SMOTE was not used.

6.1.4 Decision Tree

The Decision Tree model performs best on datasets involving value removal techniques, such as D3 and D4. Using Cross Validation and SMOTE, it achieved strong results in accuracy (97.00%), precision (96.62%), recall (97.20%), and F1-score (96.83%). Its specificity (96.40%) was also competitive in these datasets. This demonstrates that Decision Tree benefits significantly from balanced datasets and data cleaning strategies, allowing the algorithm to select the most informative features at each split without any potential bias introduced by missing value treatment techniques.

6.1.5 K-Nearest Neighbors (KNN)

KNN performs competitively in normalized datasets, such as D2 and D4, using Holdout Sampling with or without SMOTE. It achieved respectable accuracy (96.67%), precision (96.12%), recall (97.33%), and F1-score (96.50%). However, its specificity (96.00%) was slightly lower, although still acceptable, especially considering that in these medical cases it is often preferable to deal with false positives than false negatives.

Overall, Random Forest emerges as the most consistent model, followed closely by SVM. Other models provide effective solutions tailored to specific datasets and pre-processing techniques. This analysis underscores the importance of matching model selection and data pre-processing methods to the specific characteristics of the data.

When comparing the results obtained in this study with those reported in the literature, we observe some notable differences in the classification performance.

Qin et al. [15] reported an accuracy of 99.75% using Random Forest, which is higher than the 98.50% achieved in this study with SMOTE applied to the Random Forest classifier. Furthermore, the Support Vector Machine (SVM) in our study achieved slightly lower accuracy compared to the 99.25% reported by *Qin et al.* Similarly, the K-Nearest Neighbors (KNN) classifier also showed a lower accuracy than the 99.25% achieved in their study.

Almasoud and Ward [16] also reported the performance of several machine learning classifiers, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). In their study, Random Forest achieved an accuracy of 98.75%, which is very close to the 98.50% achieved in this study, showing comparable performance between the two. However, the SVM in their study reported an accuracy of 97.5%, which is lower than the 98.33% achieved in our study, indicating a slightly better performance in our case. On the other hand, the KNN classifier in their study performed better, with an accuracy of 99.25%, while our study achieved 96.67%.

These differences in performance may be due to variations in the pre-processing techniques, or classifier implementation details, such as hyperparameter tuning and feature selection. However, the results in this study remain competitive, with accuracy levels close to those of state-of-the-art methods.

6.2 Clustering

The results suggest that the method used for handling missing values has a minimal impact on clustering accuracy, as well as on other performance metrics. For instance, D1 (accuracy 0.86) and D3 (accuracy 0.84), which use scaling but differ in their missing value treatment, exhibit only slight differences. Similarly, D2 (accuracy 0.91) and D4 (accuracy 0.92), both normalized but with distinct approaches to missing data, demonstrate similar performance.

However, there was a noticeable trend where datasets using normalization (D2 and D4) performed better than those using scaling (D1 and D3). This suggests that the K-Means algorithm, being sensitive to the scale of the data, may benefit more from normalization techniques, such as *MinMaxScaler*, that ensure all features contribute equally to the distance calculation. In contrast, scaling methods like *StandardScaler*, which center data around a mean of 0 with unit variance, may not provide the same level of uniformity for clustering tasks.

7 Conclusion

This study aimed to develop a predictive model to identify the risk of developing chronic kidney disease (CKD) through multiple symptoms, using various Data Mining

techniques, based on the CRISP-DM methodology. Ideally, the obtained results can be used to contribute to early diagnosis and clinical decision-making.

The best-performing model, based on the Random Forest algorithm, achieved an accuracy of 98.50%, sensitivity of 98.40%, and precision of 98.44%. These results indicate that the model appears to be capable of identifying patients at risk of developing CKD, minimizing false negatives, and thereby improving the potential for early interventions.

Comparing these results with the existing literature around the subject, the results achieved and described here are similar to the state-of-the-art ones. Although some metrics present slightly lower values in some models, the difference is very small for most cases.

Furthermore, exploratory data analysis revealed that variables such as hemoglobin and creatinine levels are among the most influential factors in predicting CKD risk. Additionally, strong associations between high blood pressure, history of diabetes and CKD were also found. These findings align with previous studies in literature, reinforcing the validity of the obtained results.

However, some limitations were also identified. The quality of the data used presented constraints, such as large amounts of missing values in certain attributes. In addition to that, it would have been interesting to have more data regarding demographic attributes, such as gender or race/ethnicity, since multiple medical conditions, including CKD, can present differently among this groups. Therefore, the generalizability of the results to different populations should be evaluated in future studies, with the inclusion of new data and variables.

In conclusion, this work highlights the potential of Data Mining as a powerful tool in fighting the progression of CKD as well as making a timely diagnosis, contributing to public health and advancing the integration of technology and medicine.

References

[1] Chen, T.K., Knicely, D.H., Grams, M.: Chronic kidney disease diagnosis and management: A review. JAMA **322**, 1294–1304 (2019) https://doi.org/10.1001/jama.2019.14745

[2] National Health Service: Chronic Kidney Disease (CKD). Accessed: 2025-01-09 (n.d.). https://www.nhs.uk/conditions/kidney-disease/

[3] Vallianou, N.G., Mitesh, S., Gkogkou, A., Geladari, E.: Chronic kidney disease and cardiovascular disease: Is there any relationship? Current Cardiology Reviews **15**(1), 55–63 (2019) https://doi.org/10.2174/1573403X14666180711124825

[4] Turin, T.C., Tonelli, M., Manns, B., Ravani, P., Ahmed, S.B., Hemmelgarn, B.R.: Proteinuria and rate of change in kidney function in a community-based population. Journal of the American Society of Nephrology (JASN) **24**(10), 1661–1667 (2013) https://doi.org/10.1681/ASN.2012111118

[5] Frederik, P., Peter, R.: Diagnosis of diabetic kidney disease: state of the art and future perspective. Kidney International Supplements **8**(1), 2–7 (2018) https://doi.org/10.1016/j.kisu.2017.10.003

[6] Alturki, N., Altamimi, A., Umer, M., Saidani, O., Alshardan, A., Alsubai, S., Omar, M., Ashraf, I.: Improving prediction of chronic kidney disease using knn imputed smote features and trionet model. Computer Modeling in Engineering & Sciences **139**(3), 3513–3534 (2024) https://doi.org/10.32604/cmes.2023.045868

[7] Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G., Shuaib Qureshi, M.: Machine learning hybrid model for the prediction of chronic kidney disease. Computational Intelligence and Neuroscience **2023**(1), 9266889 (2023) https://doi.org/10.1155/2023/9266889

[8] Dubey, Y., Mange, P., Barapatre, Y., Sable, B., Palsodkar, P., Umate, R.: Unlocking precision medicine for prognosis of chronic kidney disease using machine learning. Diagnostics **13**(19) (2023) https://doi.org/10.3390/diagnostics13193151

[9] Moreno-Sánchez, P.A.: Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable ai model. IEEE Access **11**, 38359–38369 (2023) https://doi.org/10.1109/ACCESS.2023.3264270

[10] Rubini, L., Soundarapandian, P., Eswaran, P.: Chronic Kidney Disease. UCI Machine Learning Repository (2015). https://doi.org/10.24432/C5G020

[11] Hamrahian, S.M., Falkner, B.: In: Islam, M.S. (ed.) Hypertension in Chronic Kidney Disease, pp. 307–325. Springer, Cham (2017). https://doi.org/10.1007/5584_2016_84

[12] Kumar, M., Dev, S., Khalid, M.U., Siddenthi, S.M., Noman, M., John, C., Akubuiro, C., Haider, A., Rani, R., Kashif, M., Varrassi, G., Khatri, M., Kumar, S., Mohamad, T.: The bidirectional link between diabetes and kidney disease: Mechanisms and management. Cureus **15**(9), 45615 (2023) https://doi.org/10.7759/cureus.45615

[13] Laville, S.M., Couturier, A., Lambert, O., Metzger, M., Mansencal, N., Jacquelinet, C., Laville, M., Frimat, L., Fouque, D., Combe, C., Robinson, B.M., Stengel, B., Liabeuf, S., Massy, Z.A., collaborators, C.-R.: Urea levels and cardiovascular disease in patients with chronic kidney disease. Nephrol Dial Transplant **38**(1), 184–192 (2022) https://doi.org/10.1093/ndt/gfac045

[14] National Institute of Diabetes and Digestive and Kidney Diseases: Anemia in chronic kidney disease. Accessed: 2025-01-15 (Last Reviewed September 2020). https://www.niddk.nih.gov/health-information/kidney-disease/anemia

[15] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., Chen, B.: A machine learning methodology for diagnosing chronic kidney disease. IEEE Access **8**, 20991–21002 (2020) https://doi.org/10.1109/ACCESS.2019.2963053

[16] Almasoud, M., Ward, T.E.: Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Advanced Computer Science and Applications **10**(8) (2019) https://doi.org/10.14569/IJACSA.2019.0100813