

---

# Project Machine Learning : Diagnosis of Chronic Kidney Disease

---

**Prepared by :**

Zeineb Mbarki  
Omayma Djebali  
Maryem Elkemel  
Ibtissem Ben Dhiab  
Iheb Jeridi  
Iheb Akrimi

**Class : 4DS6**

**School Year : 2022/2023**

---

## table of contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Business Understanding</b>	<b>4</b>
<b>3</b>	<b>Data Understanding</b>	<b>5</b>
3.1	Data visualisation . . . . .	5
3.1.1	The correlations between the target and other features . . . . .	6
3.1.2	Visualisation of missing values . . . . .	8
<b>4</b>	<b>Data Preparation</b>	<b>8</b>
4.1	Data Cleaning . . . . .	8
4.1.1	Missing values . . . . .	8
4.2	Data Transformation . . . . .	9
4.2.1	Encoding . . . . .	9
4.2.2	Feature selection . . . . .	9
<b>5</b>	<b>Modeling and Evaluation</b>	<b>10</b>
5.1	Scaling . . . . .	10
5.2	Splitting . . . . .	10
5.3	Applying The Models On The Base . . . . .	10
5.3.1	k-Nearest Neighbor (KNN) . . . . .	11
5.3.2	Support Vector Machine ( SVM ) . . . . .	12
5.3.3	Naive Bayes : . . . . .	12
5.3.4	Decision Tree : . . . . .	13
5.3.5	Random Forest : . . . . .	14
5.3.6	AdaBoost : . . . . .	14
5.3.7	Adaboost with Decision Tree : . . . . .	14
5.4	Evaluation . . . . .	15
5.5	Modeling with feature selection RFE . . . . .	16
5.5.1	Evaluation with feature selection RFE . . . . .	16
5.6	Modeling with feature selection CFS . . . . .	17
5.6.1	Evaluation with feature selection CFS : . . . . .	18

---

6	Deployment	18
7	Conclusion	18

---

## Table des figures

1	The original dataset . . . . .	5
2	The features . . . . .	5
3	Percentage of ckd and notckd people . . . . .	6
4	correlation between the classes and hemoglobin . . . . .	6
5	correlation between the classes and serum creatinine . . . . .	7
6	correlation between the classes and serum blood glucose . . . . .	7
7	Age Histogram . . . . .	7
8	detection of missing values . . . . .	8
9	hemo feature before and after treating missing values . . . . .	8
10	rbc feature before and after treating missing values . . . . .	9
11	appet feature before and after encoding . . . . .	9
12	Output of RFE algorithm . . . . .	10
13	Confusion matrix . . . . .	11
14	Confusion matrix . . . . .	12
15	Confusion matrix . . . . .	13
16	Confusion matrix . . . . .	13
17	Confusion matrix . . . . .	14
18	Confusion matrix . . . . .	15
19	summary table of basic metrics . . . . .	15
20	summary table of basic metrics . . . . .	16
21	Summary table of the metrics . . . . .	17
22	Summary table of the scores . . . . .	17
23	Summary table od the metrics . . . . .	18
24	Summary table of the scores . . . . .	18

# 1 Introduction

Today thanks to computer and technological developments, we are able to use machine learning algorithms in various fields. Indeed machine learning is the study of using algorithms and data that allow computers to perform tasks without instructions or input from human users.

Learning is useful when there is a lack of human expertise, a large amount of data to process or even when humans can't explain their expertise. So we use machine learning to develop systems that can automatically discover new knowledge from large databases, with an ability to mimic humans and replace certain monotonous tasks.

For this project, we are going to deal with the Diagnosis of Chronic Kidney Disease dataset, on which we are going to achieve the reproducibility of the scientific experiments covered in the two given articles.

For that we will adopt the methodology CRISP-DM short for Cross Industry Standard Process for Data Mining.

First of all we will begin with the Business Understanding; the project objectives and requirements understanding. Then the data understanding which is the initial data collection and familiarization. Followed by the data preparation for the feature selection, data transformation and cleaning. The modeling, where we apply modeling techniques selection and application. The evaluation of the model and finally the deployment .

## 2 Business Understanding

The first stage of the CRISP-DM process is to understand what you want to accomplish from a business perspective. In this case, this study aims to enhance the quality of chronic kidney disease classification with feature selection method and ensemble learning.

Chronic kidney disease (CKD) is among the top 20 causes of death worldwide and affects approximately 10 % of the world adult population. CKD is a condition in which the kidneys are damaged and cannot filter blood as well as they should. Because of this, excess fluid and waste from blood remain in the body and may cause other health problems, such as heart disease and stroke.

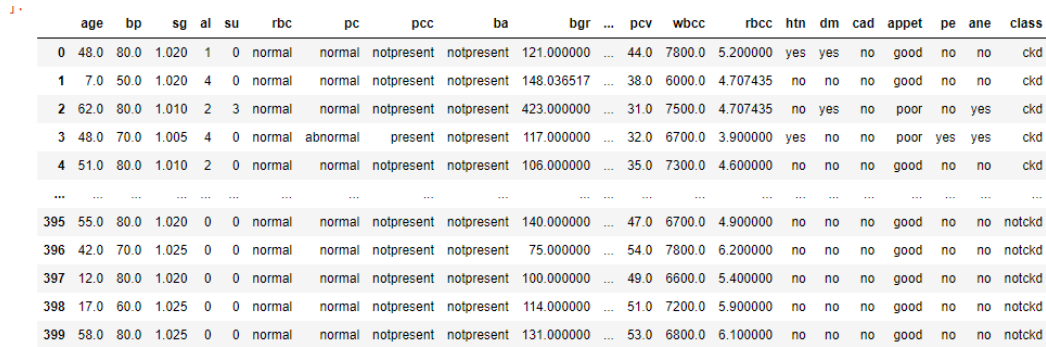
Some health consequences of CKD include :

- Anemia or low number of red blood cells
  - Anemia or low number of red blood cells
  - Increased occurrence of infections
  - Low calcium levels, high potassium levels, and high phosphorus levels in the blood of appetite or eating less or lower quality of life
  - high and low blood pressure
  - diabetes , nerve damage and bone problems
-

## 3 Data Understanding

### 3.1 Data visualisation

The CKD dataset was collected from 400 patients from the University of California, Irvine Machine Learning Repository.



	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	1.020	1	0	normal	normal	notpresent	notpresent	121.000000	...	44.0	7800.0	5.200000	yes	yes	no	good	no	no	ckd
1	7.0	50.0	1.020	4	0	normal	normal	notpresent	notpresent	148.036517	...	38.0	6000.0	4.707435	no	no	no	good	no	no	ckd
2	62.0	80.0	1.010	2	3	normal	normal	notpresent	notpresent	423.000000	...	31.0	7500.0	4.707435	no	yes	no	poor	no	yes	ckd
3	48.0	70.0	1.005	4	0	normal	abnormal	present	notpresent	117.000000	...	32.0	6700.0	3.900000	yes	no	no	poor	yes	yes	ckd
4	51.0	80.0	1.010	2	0	normal	normal	notpresent	notpresent	106.000000	...	35.0	7300.0	4.600000	no	no	no	good	no	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	55.0	80.0	1.020	0	0	normal	normal	notpresent	notpresent	140.000000	...	47.0	6700.0	4.900000	no	no	no	good	no	no	notckd
396	42.0	70.0	1.025	0	0	normal	normal	notpresent	notpresent	75.000000	...	54.0	7800.0	6.200000	no	no	no	good	no	no	notckd
397	12.0	80.0	1.020	0	0	normal	normal	notpresent	notpresent	100.000000	...	49.0	6600.0	5.400000	no	no	no	good	no	no	notckd
398	17.0	60.0	1.025	0	0	normal	normal	notpresent	notpresent	114.000000	...	51.0	7200.0	5.900000	no	no	no	good	no	no	notckd
399	58.0	80.0	1.025	0	0	normal	normal	notpresent	notpresent	131.000000	...	53.0	6800.0	6.100000	no	no	no	good	no	no	notckd

400 rows x 25 columns

FIGURE 1 – The original dataset

The dataset comprises 24 features divided into 11 numeric features and 13 categorical features, in addition to the class features, such as “ckd” and “notckd” for classification. Features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia.

#	Column	Non-Null Count	Dtype
0	age	391 non-null	float64
1	bp	388 non-null	float64
2	sg	353 non-null	object
3	al	354 non-null	object
4	su	351 non-null	object
5	rbc	248 non-null	object
6	pc	335 non-null	object
7	pcc	396 non-null	object
8	ba	396 non-null	object
9	bgr	356 non-null	float64
10	bu	381 non-null	float64
11	sc	383 non-null	float64
12	sod	313 non-null	float64
13	pot	312 non-null	float64
14	hemo	348 non-null	float64
15	pcv	329 non-null	float64
16	wbcc	294 non-null	float64
17	rbcc	269 non-null	float64
18	htn	398 non-null	object
19	dm	398 non-null	object
20	cad	398 non-null	object
21	appet	399 non-null	object
22	pe	399 non-null	object
23	ane	399 non-null	object
24	class	400 non-null	object

FIGURE 2 – The features

The diagnostic class contains two values : ckd and notckd .It contains 250 cases of “ckd” class by 62.5% and 150 cases of “notckd” by 37.5%.

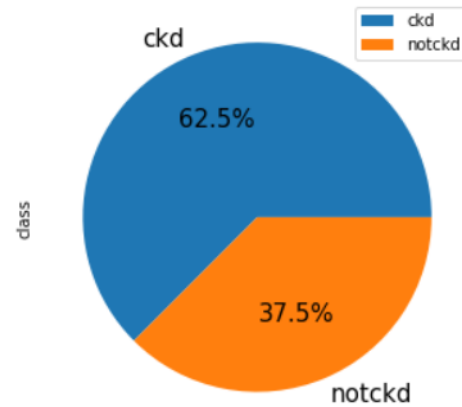


FIGURE 3 – Percentage of ckd and notckd people

To more understand the dataset, we did representative figures of the different features in function of the diagnostic class.

### 3.1.1 The correlations between the target and other features

In this figure below , we note that the red curve which represents the ckd patients , shows that Low hemoglobin can be a marker for the severity of kidney disease

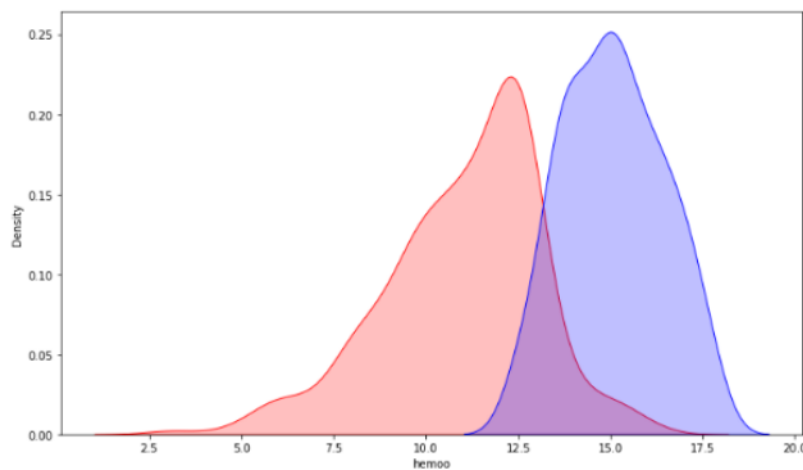


FIGURE 4 – correlation between the classes and hemoglobin

Creatinine is a waste product found in the blood during muscle activities. The kidney is involved in removing this waste material out from the body and when the kidney function is compromised, the amount of creatinine remaining in the blood will be high.

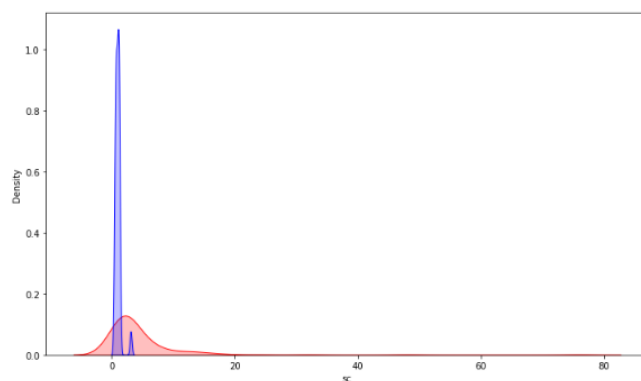


FIGURE 5 – correlation between the classes and serum creatinine

There is a risk of low blood glucose in patients with chronic kidney disease as kidney function declines insulin and if the patient suffers from diabetes, the diabetes medications will remain in the system longer because of decreased kidney clearance

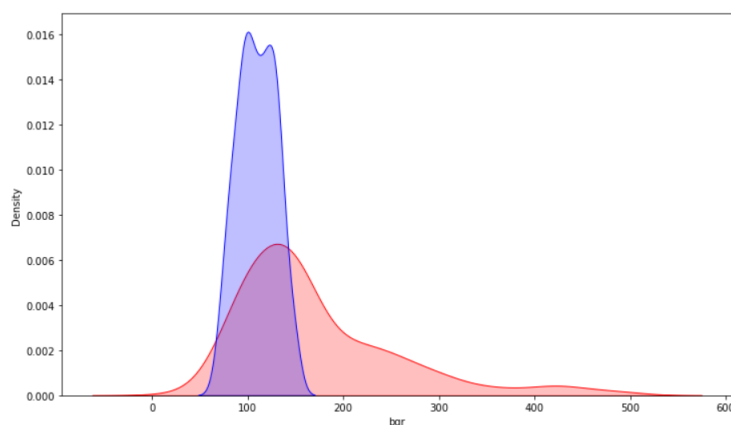


FIGURE 6 – correlation between the classes and serum blood glucose

According to current figure , CKD is more common in people aged between 60 and 75 years old than in people aged 45-60 years or 18-20 years old

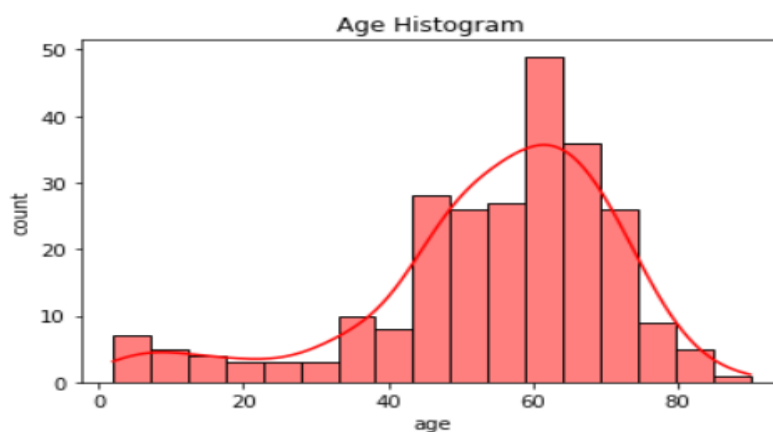


FIGURE 7 – Age Histogram



### 3.1.2 Visualisation of missing values

This figure below show that we have missing values ,The Red blood cells has the highest percentage of missing values with 38% And the lowest percentage are the appetite and anemia with 0.25%

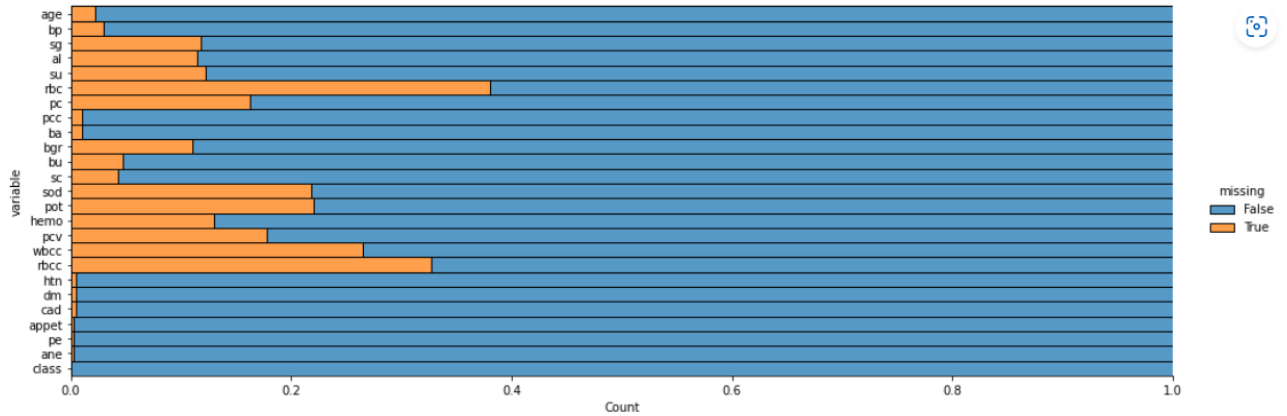


FIGURE 8 – detection of missing values

## 4 Data Preparation

### 4.1 Data Cleaning

Data cleaning is the process of correcting or removing corrupt, incorrect, or unnecessary data from the dataframe .

#### 4.1.1 Missing values

To deal with the numerical missing values we choose to apply the statistical method mean.This two figures below shows an example of a numerical feature hemoglobin before and after the transformation

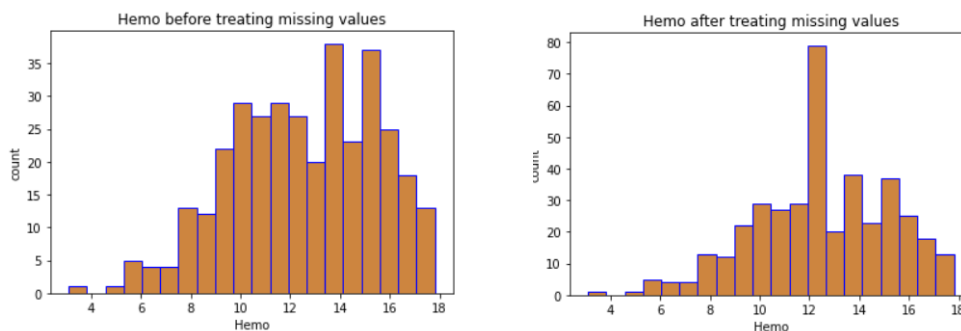


FIGURE 9 – hemo feature before and after treating missing values

To deal with the nominal missing values we choose to apply the most frequent values. This two figures below shows an example of a numerical feature rbc before and after the transformation

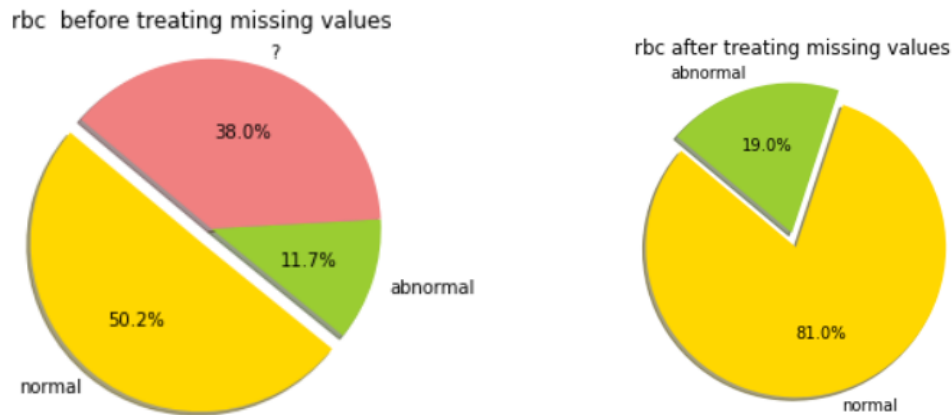


FIGURE 10 – rbc feature before and after treating missing values

## 4.2 Data Transformation

### 4.2.1 Encoding

For Encoding , we used to replace nominal values like ('normal', 'yes', 'present', 'good') with 1 and all negative data ('abnormal', 'non', 'nopresent', 'poor') with 0 .

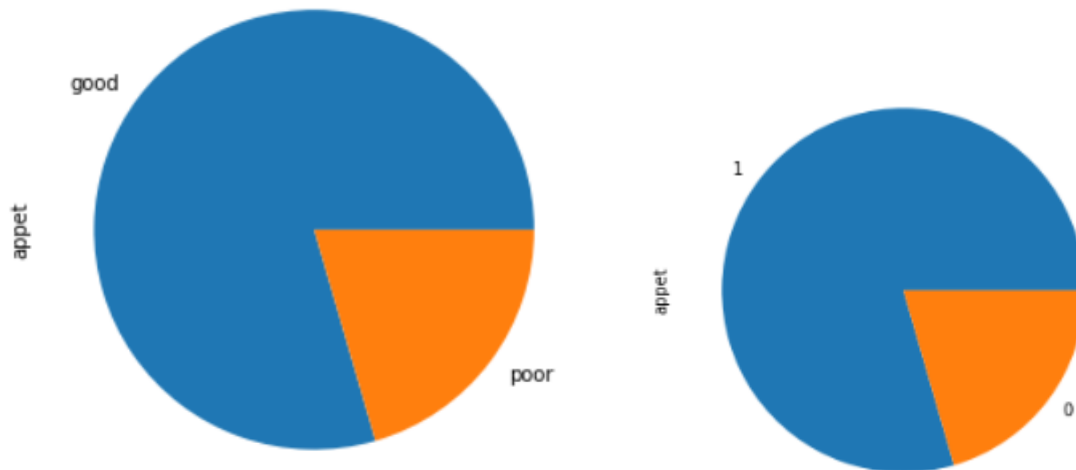


FIGURE 11 – appet feature before and after encoding

### 4.2.2 Feature selection

Feature selection is used to reduce the dimensions of the dataset and to select the most relevant features associated with CKD . We have used RFE for feature selection for

ensemble learning to improve CKD diagnosis. As mentioned in these two articles , We used integrated model to select the most significant representative features by using the Recursive Feature Elimination (RFE) algorithm

Selected feature	
0	al
1	su
2	rbc
3	pc
4	sc
5	pot
6	hemo
7	rbcc
8	htn
9	dm
10	appet
11	pe

FIGURE 12 – Output of RFE algorithm

## 5 Modeling and Evaluation

Modeling includes selecting, configuring and testing various algorithms, as well as deciding on their sequence, which provides the best results. The process is initially a descriptive one that generates knowledge and explains why things happened. It then becomes predictive and explains what will happen, and later prescriptive as it helps optimise future situations.

### 5.1 Scaling

Feature scaling is the process of normalising the range of features in the dataset, in order for machine learning models to interpret these features on the same scale.

### 5.2 Splitting

Whenever we use a machine learning model, we can't train that model on a single dataset. If we train it on a single dataset then we will not be able to assess the performance of our model. For that reason, we split our source data into training and testing datasets.

### 5.3 Applying The Models On The Base

First of all we are going to use the dataset without any features selection.

### 5.3.1 k-Nearest Neighbor (KNN)

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

- Grid Search

To properly use the KNN algorithm, we applied the Grid Search method to define the best parameters of the model. The Grid Search tries out different values and then pick the value that gives the best score :

Number of neighbors 3 and the metric is euclidean

- Then we applied the algorithm using those parameters :

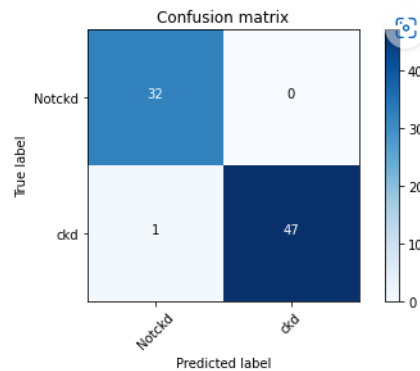


FIGURE 13 – Confusion matrix

The result of the algorithm is based on the output of the confusion Matrix as mentioned below :

- True Positive : We predicted 47 positive ( Ckd ) and it's true
- True Negative : We predicted 32 negative ( Not Ckd) and it's true
- False Positive : we predicted one false positive
- False Negative : We didn't predict any false negative.

### 5.3.2 Support Vector Machine ( SVM )

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of the SVM algorithm is to find a hyper-plane in an N-dimensional space that distinctly classifies the data points.

The result of the algorithm is based on the output of the confusion Matrix as mentioned below :

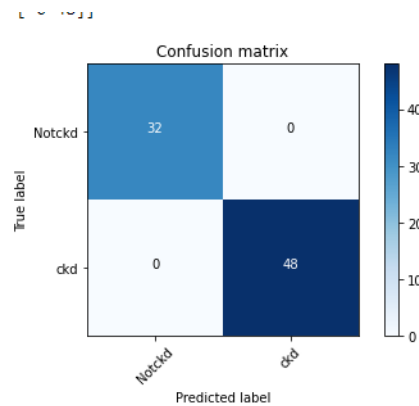


FIGURE 14 – Confusion matrix

- True Positive : We predicted 48 positive ( Ckd ) and it's true
- True Negative : We predicted 32 negative ( Not Ckd) and it's true
- False Positive : We didn't predict any false negative.
- False Negative : We didn't predict any false negative.

### 5.3.3 Naive Bayes :

The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. The result of the algorithm is based on the output of the confusion Matrix as mentioned below :

- True Positive : We predicted 46 positive ( Ckd ) and it's true
- True Negative : We predicted 32 negative ( Not Ckd) and it's true
- False Positive : we predicted two false positive
- False Negative : We predicted one predict any false negative.

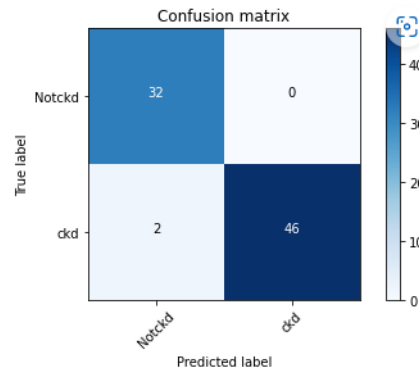


FIGURE 15 – Confusion matrix

### 5.3.4 Decision Tree :

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute

We defined a parameter and then applied the Grid Search to find the best values of this parameter is entropy

The confusion matrix :

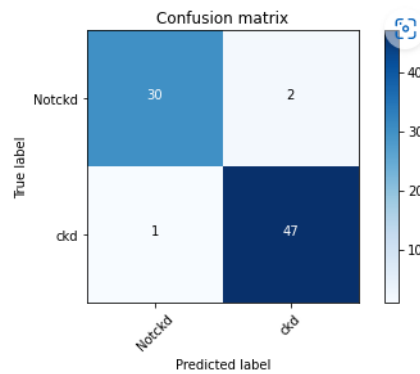


FIGURE 16 – Confusion matrix

- True Positive : We predicted 47 positive ( Ckd ) and it's true
- True Negative : We predicted 30 negative ( Not Ckd) and it's true
- False Positive : we predicted two false positive
- False Negative : We predicted one false negative.

### 5.3.5 Random Forest :

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble consists of a data sample drawn from a training set with replacement, called the bootstrap sample.

Before applying the Random Forest algorithm, we used the Grid Search method to define the best parameters :

The confusion matrix :

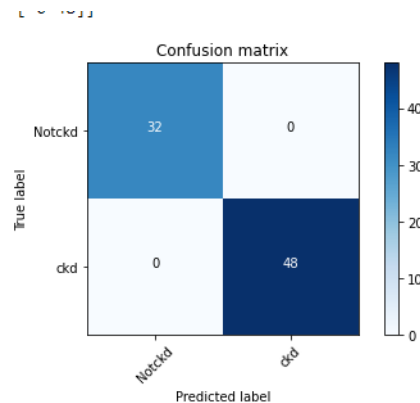


FIGURE 17 – Confusion matrix

- True Positive : We predicted 48 true positive.
- True Negative : We predicted 32 true negative.
- False Positive : We didn't predict any false positive.
- False Negative : we didn 't predict any false negative.

### 5.3.6 AdaBoost :

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted.

### 5.3.7 Adaboost with Decision Tree :

A decision tree is boosted using the AdaBoost. As we can notice from the matrix above, the use of the this model gives the same result as the SVM.

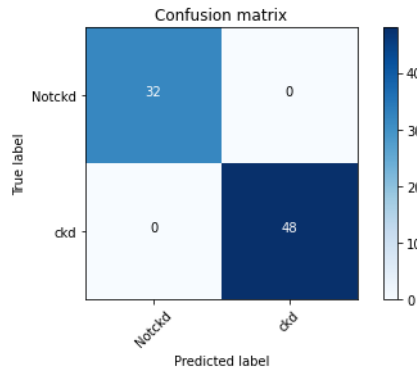


FIGURE 18 – Confusion matrix

## 5.4 Evaluation

After applying all the models treated above, this is a summary table which brings together all the results.

Based on the table, it can be noted that Random Forest and SVM have the highest accuracy (100%), highest recall (100%), highest precision (100%) and highest F1-score (100%) of all the other algorithms. The accuracy on 1st method using base classifier are quite high, the lowest accuracy is 0.96 accuracy rate with decision tree , both Naive bayes and adaboost with Naive bayes accuracy rate is 0.958.

	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
<b>KNN</b>	0.9875	1.000000	0.979167	0.989474
<b>SVM</b>	1.0000	1.000000	1.000000	1.000000
<b>NB</b>	0.9750	1.000000	0.958333	0.978723
<b>RF</b>	1.0000	1.000000	1.000000	1.000000
<b>DT</b>	0.9625	0.959184	0.979167	0.969072
<b>AB+NB</b>	0.9750	1.000000	0.958333	0.978723
<b>ADB+DT</b>	1.0000	1.000000	1.000000	1.000000

FIGURE 19 – summary table of basic metrics

From the metrics recorded which is unrealistically good, there are some improvements that can be done to improve the fitting performance like feature selection so we are going to apply RFE and CFS.

the figure below shows the score of each classifier , as we can notice Knn , Adaboost , Random forest and SVM have the highest score .



	name	score
0	Nearest_Neighbors_basique	1.000000
1	Linear_SVM	1.000000
2	Decision_Tree	0.962500
3	Random_Forest	1.000000
4	AdaBoost	1.000000
5	Naive_Bayes	0.975000

FIGURE 20 – summary table of basic metrics

## 5.5 Modeling with feature selection RFE

Now, we are going to use RFE feature selection to increase the performance of base classifier.

Recursive Feature Elimination (RFE) reduces the model complexity by removing features one by one until the optimal number of features is left. It is one of the most popular feature selection algorithms due to its flexibility and ease of use.

The features identified by the RFE are :  
albumin,sugar,red blood cells,pus cell,serum creatinine potassium,hemoglobin,red blood cell count ,hypertension,diabetes mellitus,appetite and edema.

**k-Nearest Neighbor (KNN) :** Like we already did when applying the KNN model before, we used the Grid Search to identify the best parameters to use : Number of neighbors are 2 and the metric is euclidean.

**Support vector machine (SVM) :** The best parameters for the model after the Grid Search is : linear kernel

**Decision tree (DT) :** We defined a parameter and then applied the Grid Search to find the best values of this parameter which is entropy.

**Random forest :** We defined a parameter and then applied the Grid Search to find the best values of this parameter : 30 estimators .

### 5.5.1 Evaluation with feature selection RFE

After applying all the models treated above, this is a summary table which brings together all the results.

Based on the table, it can be noted that Random Forest has the highest accuracy (100%), highest recall (100%), highest precision (100%) and highest F1-score (100%) of all the other algorithms.

The accuracy on 1st method using RFE base classifier are quite high, the lowest accuracy is 0.93 accuracy rate is SVM and NB , both adaboost with Random forest ans decision tre have accuracy rate is 0.97.

	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
<b>KNN</b>	0.9625	1.0	0.941176	0.969697
<b>SVM</b>	0.9375	1.0	0.901961	0.948454
<b>NB</b>	0.9375	1.0	0.901961	0.948454
<b>RF</b>	1.0000	1.0	1.000000	1.000000
<b>DT</b>	0.9750	1.0	0.960784	0.980000
<b>AB+NB</b>	0.9750	1.0	0.960784	0.980000
<b>AD+DT</b>	0.9750	1.0	0.960784	0.980000

FIGURE 21 – Summary table of the metrics

the figure below shows the score of each classifier , as we can notice SVM and Adaboost have the highest score .

	name	score
0	Nearest_Neighbors_basique	0.987500
1	Linear_SVM	1.000000
2	Decision_Tree	0.975000
3	Random_Forest	0.987500
4	AdaBoost	1.000000
5	Naive_Bayes	0.987500

FIGURE 22 – Summary table of the scores

## 5.6 Modeling with feature selection CFS

Correlation-based Feature Selection (CFS) is suitable to be applied to multivariate data. CFS works by calculating the interaction between features. CFS evaluates a subset of features taking into account predictive capabilities of each level of redundancy among features and those features.

The features identified by the CFS are : hypertension ,Red Blood Cell Count ,packed cell volume ,Pus Cell clumps ,serum creatinine ,anemia ,diabetes mellitus

**k-Nearest Neighbor (KNN)** Like we already did when applying the KNN model before, we used the Grid Search to identify the best parameters to use : Number of neighbors are 1 and the metric is euclidean

**Support vector machine (SVM)** The best parameters for the model after the Grid Search are : linear kernel

### 5.6.1 Evaluation with feature selection CFS :

After applying all the models treated above, this is a summary table which brings together all the results. Based on the table, it can be noted that Naive bayes and SVM has the highest accuracy (95%). The accuracy on second method using RFE base classifier are quite high, the lowest accuracy rate is 0.84 accuracy with KNN .

```

>]:

```

	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
<b>CFS+NB</b>	0.958333	1.000000	0.930556	0.964029
<b>CFS+SVM</b>	0.950000	0.971429	0.944444	0.957746
<b>CFS+knn</b>	0.841667	0.895522	0.833333	0.863309

FIGURE 23 – Summary table od the metrics

the figure below shows the score of each classifier , as we can notice SVM and Naive bayes have the highest score .

	name	score
0	Nearest_Neighbors_basique	0.841667
1	Linear_SVM	0.950000
2	Naive_Bayes	0.958333

FIGURE 24 – Summary table of the scores

## 6 Deployment

The final step in the CRISP DM methodology is the deployment in which the data mining pays off. In this final phase, it doesn't matter how brilliant your discoveries may be, or how perfectly your models fit the data, if you don't actually use those things to improve the way that you do business. In our case for this project, we are not going to elaborate this phase.

## 7 Conclusion

This report represents our work in exploring and training the dataset of the chronic kidney disease and trying different models.