November 12, 2022

Chronic KIdney Disease Report

UNIVERSITY OF JEDDAH

Data Mining report

PRESENTED BY

- Elaan moazen 2006947
- Raghad alQithmi 2005999
- Raneen Alshehri 2005560
- Taif Basheikh 2005890
- Ghaidaa Almaghrabi 1912386

Introduction

Chronic kidney disease (CKD) is a major public health problem with increasing challenges for early diagnosis, timely prevention and effective treatment.

The current data set for chronic kidney disease consists of 27 parameters and 390 rows. The study performs a comparative analysis of rule-based categories in order to create human-interpretable rules for a diagnosis. And use the different models to determine the best accuracy Decision tree, Naive Bayes , Neural Network, Random Forest, and K-NN.

Phase 1

Outline:

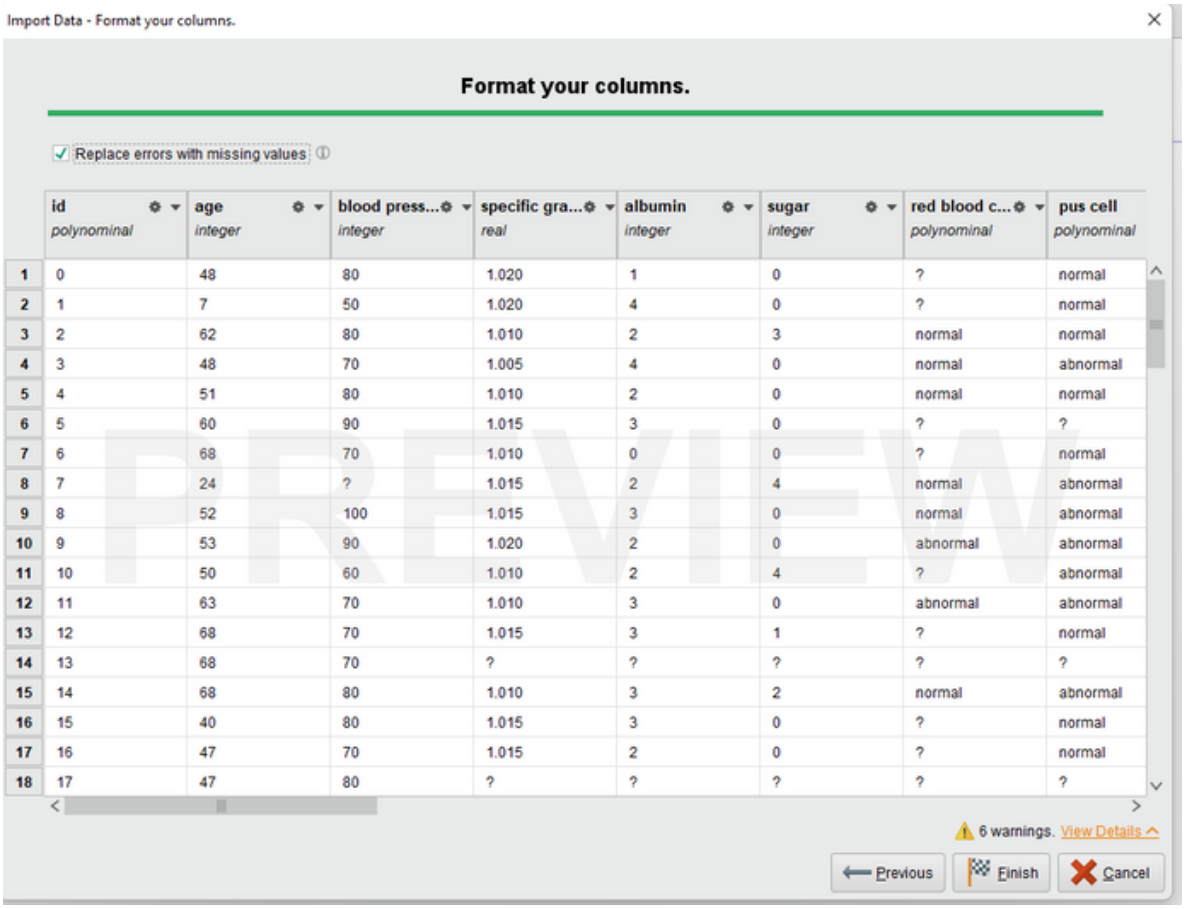
- Understand the Data
    - Types of the Attributes
    - Data Before Cleaning
    - Summarized Properties
- Visualization Techniques
    - Histogram
    - Bar Chart
    - Box plot
    - Bar Chart
    - Bell curve
- Correlated Attributes
- Data Cleaning
    - Fill the Missing Values
    - Identify outliers
    - Dealing with outliers and smooth out noisy data

Types of the Attributes

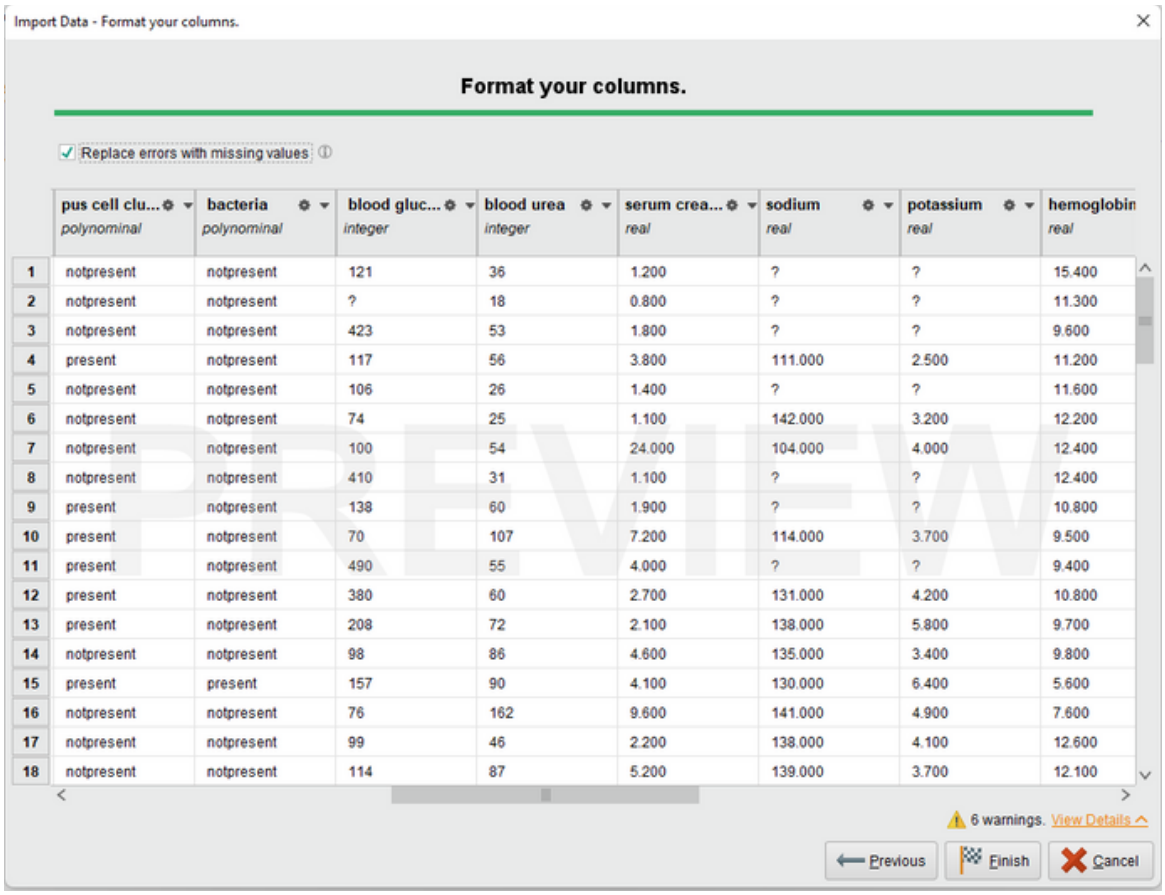
Attribute	Type	Example
Id	nominal	Id(1,2,3,4, 5,6, 7)
Age	numerical	age in years
Blood Pressure	numerical	bp in mm/Hg
Specific Gravity	numerical	sg(1.005,1.010,1.015,1.020,1.025)
Albumin	numerical	al - (0,1,2,3,4,5)
Sugar	numerical	su - (0,1,2,3,4,5)
Red Blood Cells	nominal	rbc - (normal, abnormal)
Pus Cell	nominal	pc - (normal, abnormal)
Pus Cell clumps	nominal	pcc - (present, notpresent)
Bacteria	nominal	ba - (present,notpresent)
Blood Glucose Random	numerical	bgr in mgs/dl
Blood Urea	numerical	sc in mgs/dl
Serum Creatinine	numerical	sod in mEq/L
Sodium	numerical	sod in mEq/L
Potassium	numerical	pot in mEq/L
Hemoglobin	numerical	hemo in gms
Packed Cell Volume	numerical	pcv - (44 , 38 , 31)
White Blood Cell Count	numerical	wc in cells/cumm
Red Blood Cell Count	numerical	rc in millions/cmm
Hypertension	binominal	htn - (yes, no)
Diabetes Mellitus	binominal	dm - (yes, no)
Coronary Artery Disease	binominal	cad - (yes, no)
Appetite	binominal	appet - (good, poor)
Pedal Edema	binominal	pe - (yes, no)
Anemia	binominal	ane - (yes, no)
Class	nominal	class - (ckd, notckd)

Data Before Cleaning

- **Import Data In RapidMiner :**



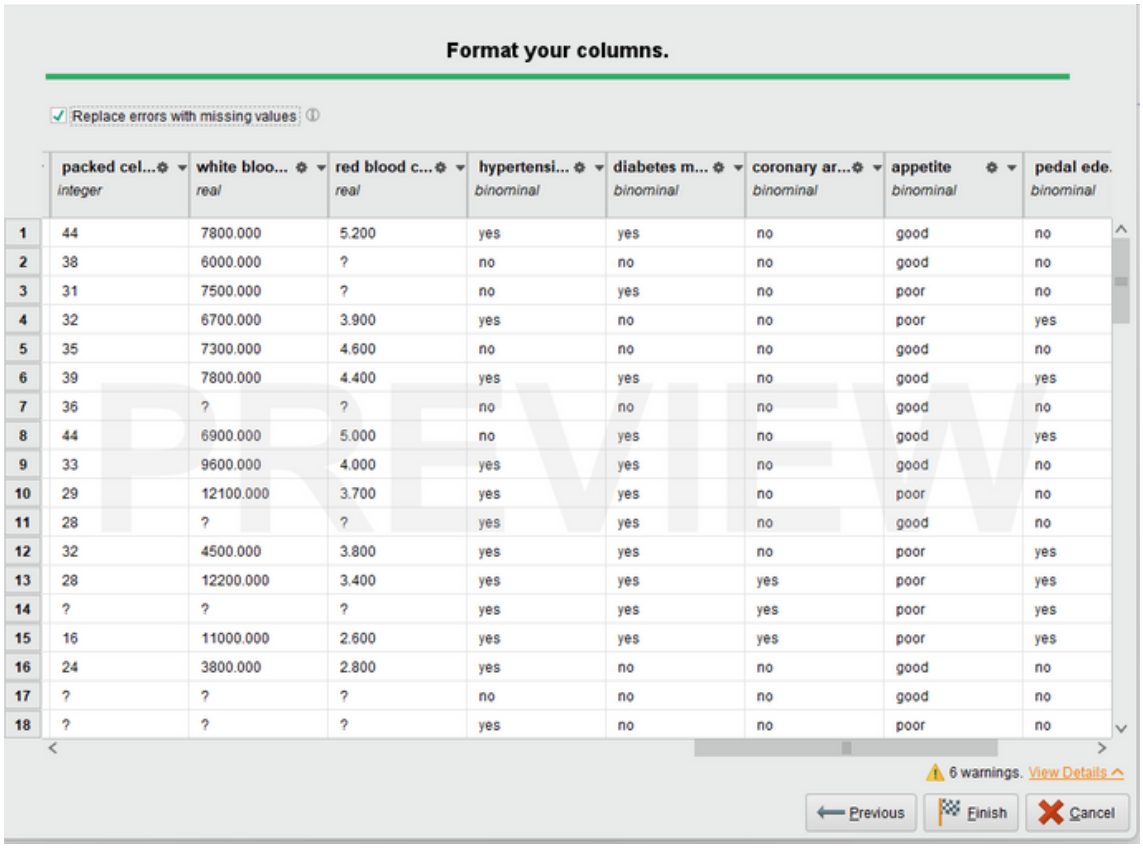
	id *polynominal*	age *integer*	blood press... *integer*	specific gra... *real*	albumin *integer*	sugar *integer*	red blood c... *polynominal*	pus cell *polynominal*
1	0	48	80	1.020	1	0	?	normal
2	1	7	50	1.020	4	0	?	normal
3	2	62	80	1.010	2	3	normal	normal
4	3	48	70	1.005	4	0	normal	abnormal
5	4	51	80	1.010	2	0	normal	normal
6	5	60	90	1.015	3	0	?	?
7	6	68	70	1.010	0	0	?	normal
8	7	24	?	1.015	2	4	normal	abnormal
9	8	52	100	1.015	3	0	normal	abnormal
10	9	53	90	1.020	2	0	abnormal	abnormal
11	10	50	60	1.010	2	4	?	abnormal
12	11	63	70	1.010	3	0	abnormal	abnormal
13	12	68	70	1.015	3	1	?	normal
14	13	68	70	?	?	?	?	?
15	14	68	80	1.010	3	2	normal	abnormal
16	15	40	80	1.015	3	0	?	normal
17	16	47	70	1.015	2	0	?	normal
18	17	47	80	?	?	?	?	?



	pus cell clu... *polynominal*	bacteria *polynominal*	blood gluc... *integer*	blood urea *integer*	serum crea... *real*	sodium *real*	potassium *real*	hemoglobin *real*
1	notpresent	notpresent	121	36	1.200	?	?	15.400
2	notpresent	notpresent	?	18	0.800	?	?	11.300
3	notpresent	notpresent	423	53	1.800	?	?	9.600
4	present	notpresent	117	56	3.800	111.000	2.500	11.200
5	notpresent	notpresent	106	26	1.400	?	?	11.600
6	notpresent	notpresent	74	25	1.100	142.000	3.200	12.200
7	notpresent	notpresent	100	54	24.000	104.000	4.000	12.400
8	notpresent	notpresent	410	31	1.100	?	?	12.400
9	present	notpresent	138	60	1.900	?	?	10.800
10	present	notpresent	70	107	7.200	114.000	3.700	9.500
11	present	notpresent	490	55	4.000	?	?	9.400
12	present	notpresent	380	60	2.700	131.000	4.200	10.800
13	present	notpresent	208	72	2.100	138.000	5.800	9.700
14	notpresent	notpresent	98	86	4.600	135.000	3.400	9.800
15	present	present	157	90	4.100	130.000	6.400	5.600
16	notpresent	notpresent	76	162	9.600	141.000	4.900	7.600
17	notpresent	notpresent	99	46	2.200	138.000	4.100	12.600
18	notpresent	notpresent	114	87	5.200	139.000	3.700	12.100

Data Before Cleaning

- **Import Data In RapidMiner :**



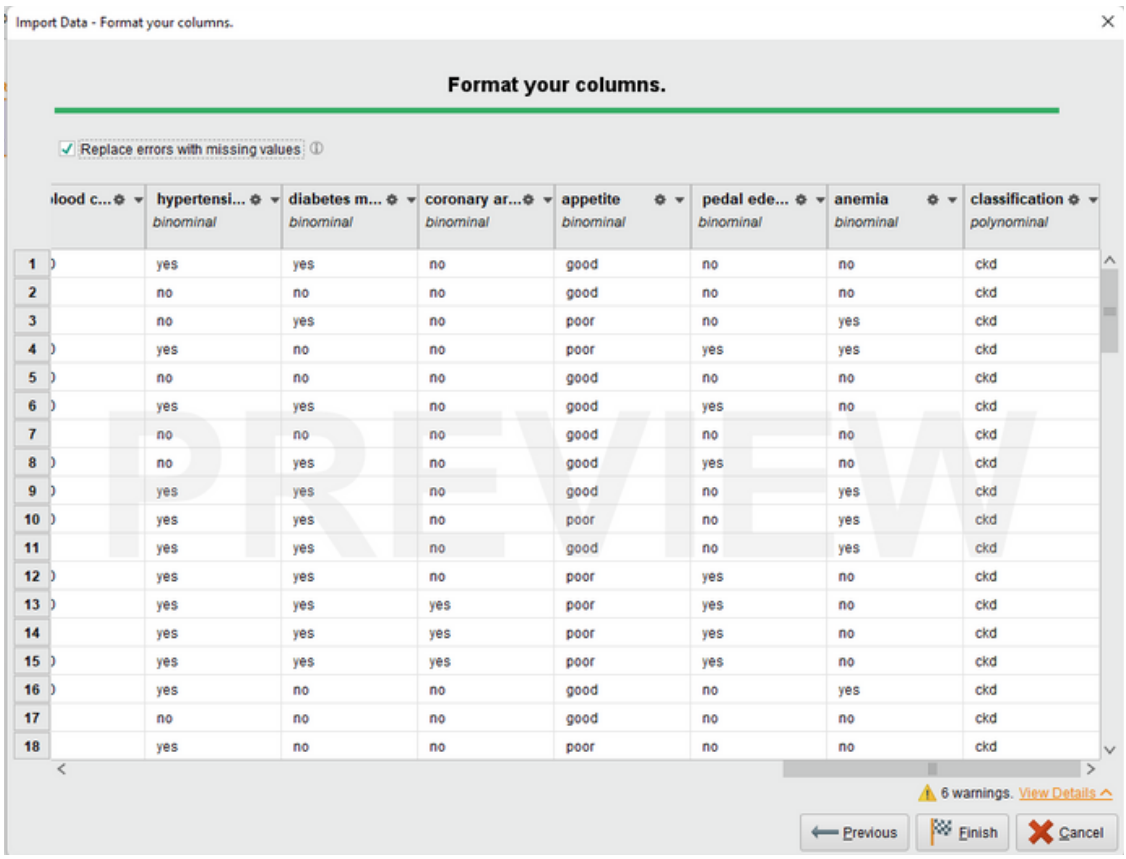
Format your columns.

Replace errors with missing values

	packed cel... integer	white bloo... real	red blood c... real	hypertensi... binominal	diabetes m... binominal	coronary ar... binominal	appetite binominal	pedal ede. binominal
1	44	7800.000	5.200	yes	yes	no	good	no
2	38	6000.000	?	no	no	no	good	no
3	31	7500.000	?	no	yes	no	poor	no
4	32	6700.000	3.900	yes	no	no	poor	yes
5	35	7300.000	4.600	no	no	no	good	no
6	39	7800.000	4.400	yes	yes	no	good	yes
7	36	?	?	no	no	no	good	no
8	44	6900.000	5.000	no	yes	no	good	yes
9	33	9600.000	4.000	yes	yes	no	good	no
10	29	12100.000	3.700	yes	yes	no	poor	no
11	28	?	?	yes	yes	no	good	no
12	32	4500.000	3.800	yes	yes	no	poor	yes
13	28	12200.000	3.400	yes	yes	yes	poor	yes
14	?	?	?	yes	yes	yes	poor	yes
15	16	11000.000	2.600	yes	yes	yes	poor	yes
16	24	3800.000	2.800	yes	no	no	good	no
17	?	?	?	no	no	no	good	no
18	?	?	?	yes	no	no	poor	no

6 warnings. View Details

Previous Finish Cancel



Import Data - Format your columns.

Format your columns.

Replace errors with missing values

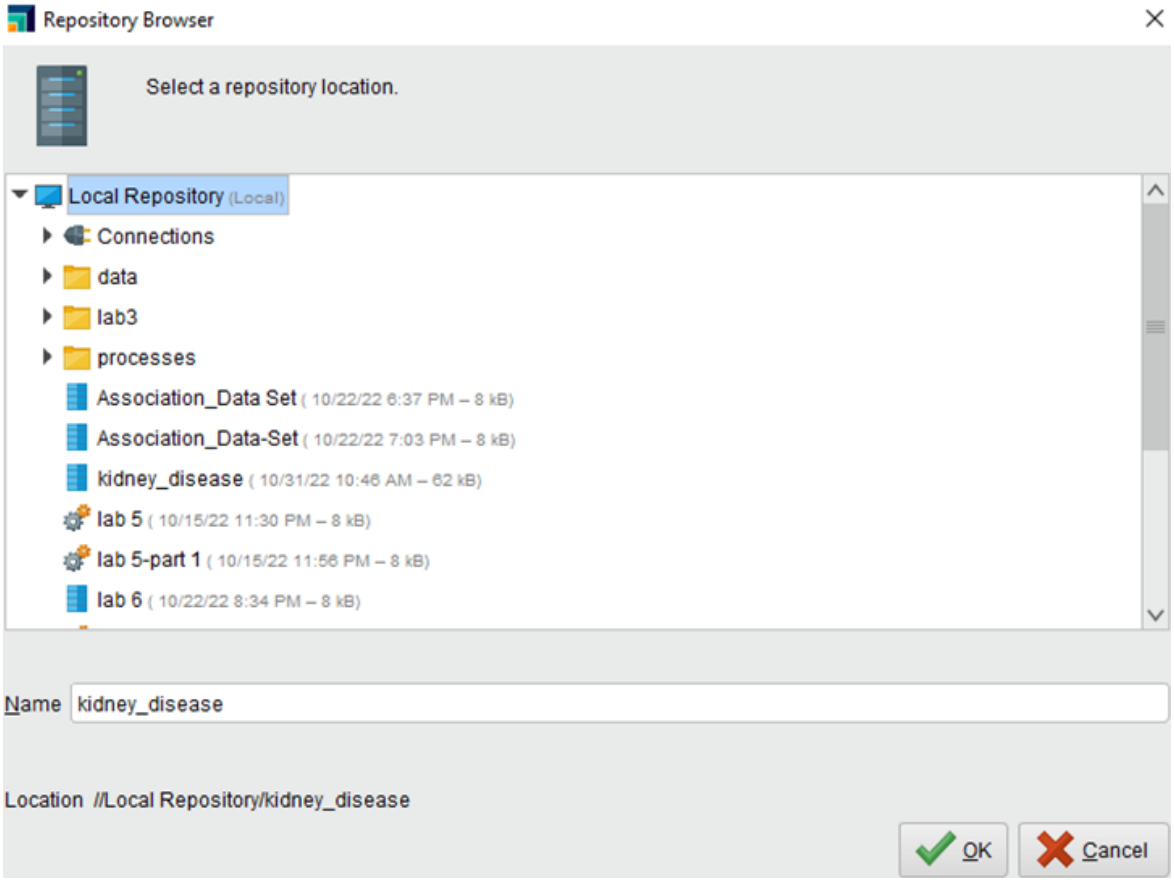
	blood c...	hypertensi... binominal	diabetes m... binominal	coronary ar... binominal	appetite binominal	pedal ede... binominal	anemia binominal	classification polynominal
1		yes	yes	no	good	no	no	ckd
2		no	no	no	good	no	no	ckd
3		no	yes	no	poor	no	yes	ckd
4		yes	no	no	poor	yes	yes	ckd
5		no	no	no	good	no	no	ckd
6		yes	yes	no	good	yes	no	ckd
7		no	no	no	good	no	no	ckd
8		no	yes	no	good	yes	no	ckd
9		yes	yes	no	good	no	yes	ckd
10		yes	yes	no	poor	no	yes	ckd
11		yes	yes	no	good	no	yes	ckd
12		yes	yes	no	poor	yes	no	ckd
13		yes	yes	yes	poor	yes	no	ckd
14		yes	yes	yes	poor	yes	no	ckd
15		yes	yes	yes	poor	yes	no	ckd
16		yes	no	no	good	no	yes	ckd
17		no	no	no	good	no	no	ckd
18		yes	no	no	poor	no	no	ckd

6 warnings. View Details

Previous Finish Cancel

Data Before Cleaning

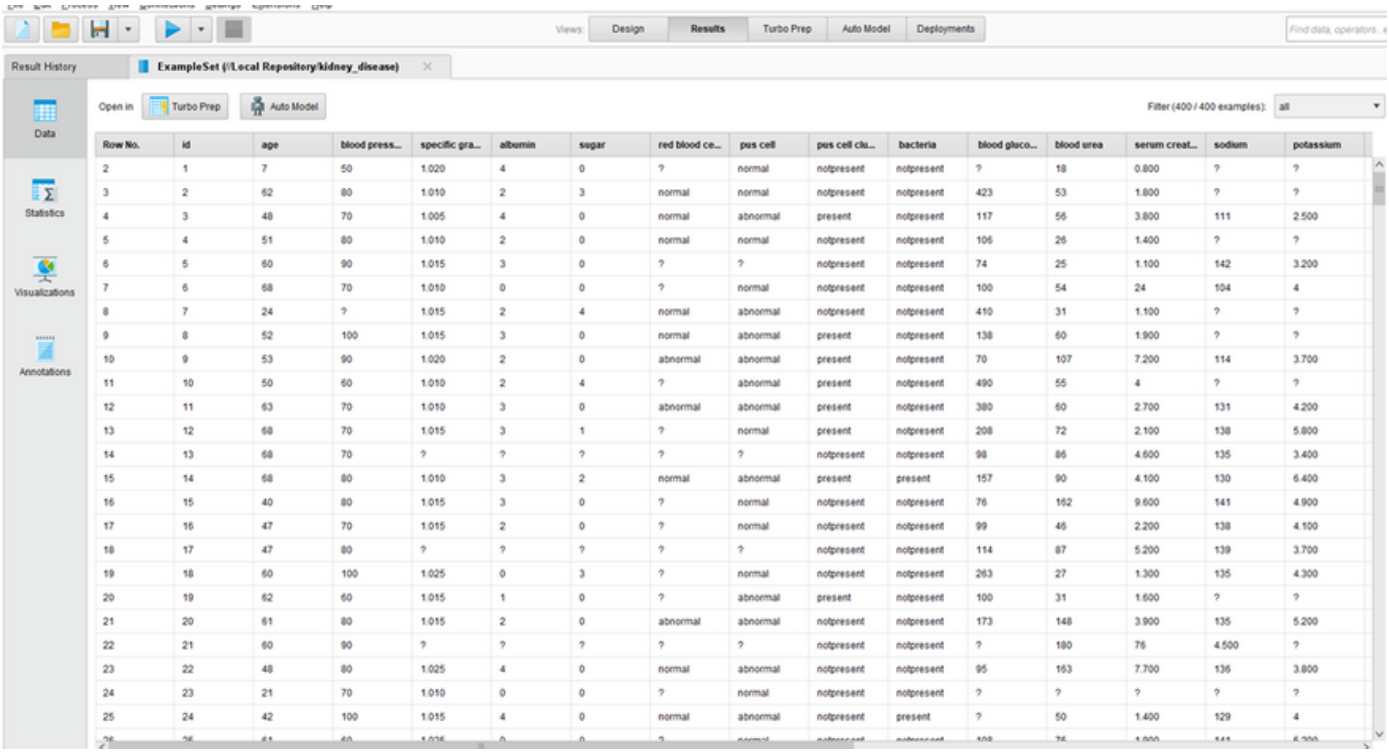
- **Select a repository location:**



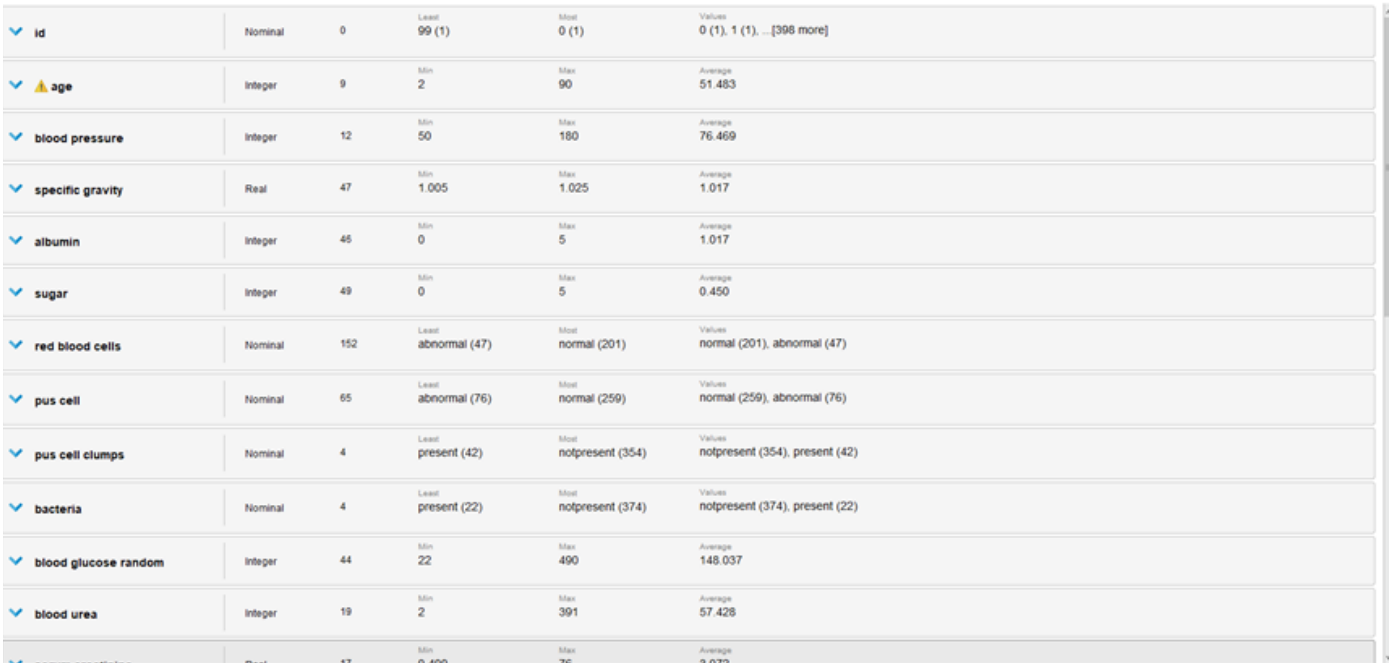
- **To display Data :**

Data Before Cleaning

- **Results:**



Row No.	id	age	blood press...	specific gra...	albumin	sugar	red blood ce...	pus cell	pus cell clu...	bacteria	blood gluco...	blood urea	serum creat...	sodium	potassium
2	1	7	50	1.020	4	0	?	normal	notpresent	notpresent	?	18	0.800	?	?
3	2	62	80	1.010	2	3	normal	normal	notpresent	notpresent	423	53	1.800	?	?
4	3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.800	111	2.500
5	4	51	80	1.010	2	0	normal	normal	notpresent	notpresent	106	26	1.400	?	?
6	5	60	90	1.015	3	0	?	?	notpresent	notpresent	74	25	1.100	142	3.200
7	6	68	70	1.010	0	0	?	normal	notpresent	notpresent	100	54	24	104	4
8	7	24	?	1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.100	?	?
9	8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.900	?	?
10	9	53	90	1.020	2	0	abnormal	abnormal	present	notpresent	70	107	7.200	114	3.700
11	10	50	60	1.010	2	4	?	abnormal	present	notpresent	490	55	4	?	?
12	11	63	70	1.010	3	0	abnormal	abnormal	present	notpresent	380	60	2.700	131	4.200
13	12	68	70	1.015	3	1	?	normal	present	notpresent	208	72	2.100	138	5.800
14	13	68	70	?	?	?	?	?	notpresent	notpresent	98	86	4.600	135	3.400
15	14	68	80	1.010	3	2	normal	abnormal	present	present	157	90	4.100	130	6.400
16	15	40	80	1.015	3	0	?	normal	notpresent	notpresent	76	162	9.600	141	4.900
17	16	47	70	1.015	2	0	?	normal	notpresent	notpresent	99	46	2.200	138	4.100
18	17	47	80	?	?	?	?	?	notpresent	notpresent	114	87	5.200	139	3.700
19	18	60	100	1.025	0	3	?	normal	notpresent	notpresent	263	27	1.300	135	4.300
20	19	62	60	1.015	1	0	?	abnormal	present	notpresent	100	31	1.600	?	?
21	20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.900	135	5.200
22	21	60	90	?	?	?	?	?	notpresent	notpresent	?	180	76	4.500	?
23	22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.700	136	3.800
24	23	21	70	1.010	0	0	?	normal	notpresent	notpresent	?	?	?	?	?
25	24	42	100	1.015	4	0	normal	abnormal	notpresent	present	?	50	1.400	129	4

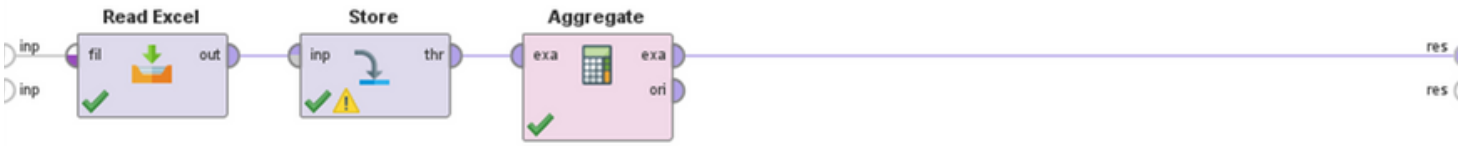


id	Nominal	0	Least: 99 (1)	Most: 0 (1)	Values: 0 (1), 1 (1), ... [398 more]
age	Integer	9	Min: 2	Max: 90	Average: 51.483
blood pressure	Integer	12	Min: 50	Max: 180	Average: 76.469
specific gravity	Real	47	Min: 1.005	Max: 1.025	Average: 1.017
albumin	Integer	46	Min: 0	Max: 5	Average: 1.017
sugar	Integer	49	Min: 0	Max: 5	Average: 0.450
red blood cells	Nominal	152	Least: abnormal (47)	Most: normal (201)	Values: normal (201), abnormal (47)
pus cell	Nominal	65	Least: abnormal (76)	Most: normal (259)	Values: normal (259), abnormal (76)
pus cell clumps	Nominal	4	Least: present (42)	Most: notpresent (354)	Values: notpresent (354), present (42)
bacteria	Nominal	4	Least: present (22)	Most: notpresent (374)	Values: notpresent (374), present (22)
blood glucose random	Integer	44	Min: 22	Max: 490	Average: 148.037
blood urea	Integer	19	Min: 2	Max: 391	Average: 57.428

As we can see in this dataset, there are many missing data, and we will address all these problems in case there are any missing values or duplicate values or outliers in the data cleaning stage.

Summarized Properties

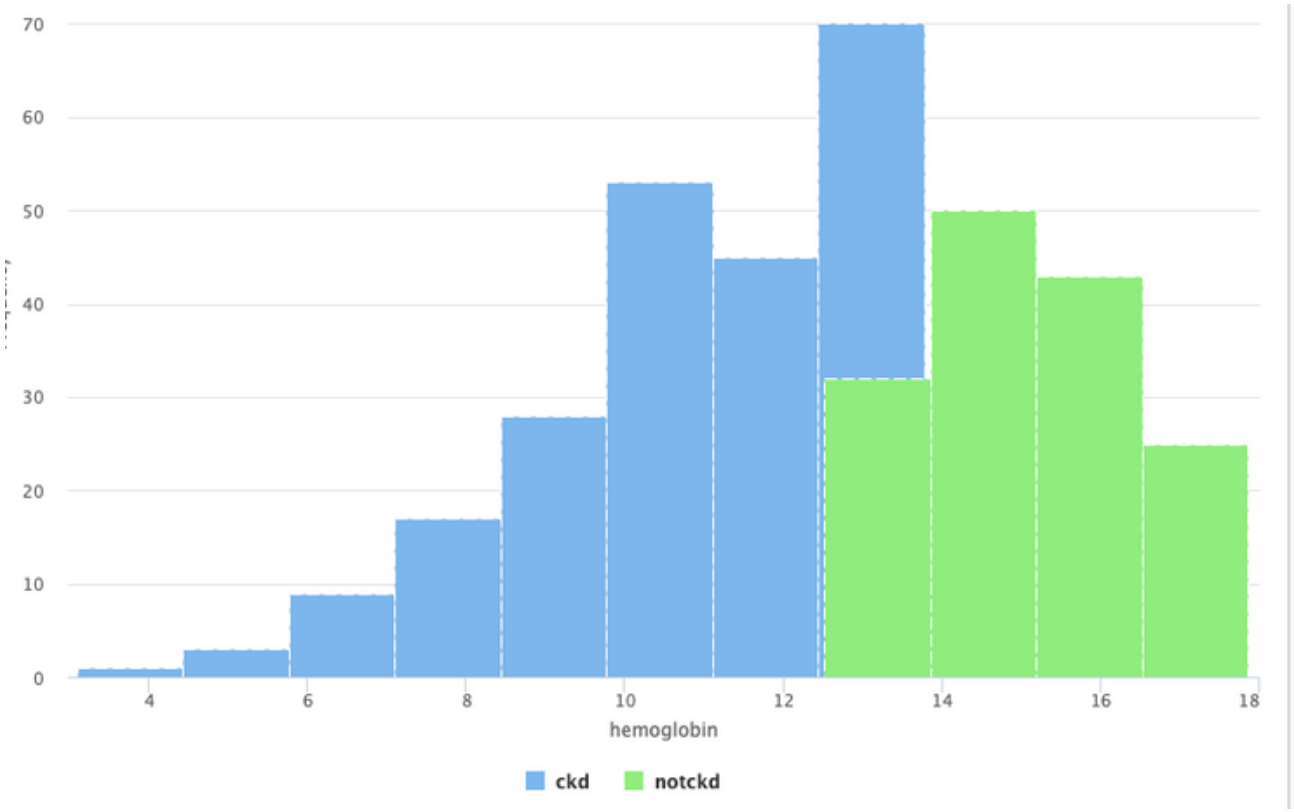
- **Make measurements for each attribute :**



Attribute	measure	result
id	Value	399
age	Min , Max ,Average ,Deviation, median , variance , mode	2 , 90 , 51.483 , 17.170 , 55 , 294.799 , 60
blood pressure	Min , Max, Average, Deviation, median , variance ,mode	50 , 180 , 76.469 , 13.684 ,80, 187.242 , 80
specific gravity	Min , Max, Average, Deviation ,median , mode	1.005 , 1.025 ,1.017 , 0.006 ,1.020 , 1.020
albumin	Min , Max, Average, Deviation, variance	0 , 5 , 1.017 , 1.353 , 1.830
sugar	Min , Max, Average, Deviation , variance	0 , 5 , 0.450 , 1.099 , 1.208
red blood cells	Mode	Normal (201) , abnormal (47)
pus cell	Mode	Normal (259) , abnormal (76)
pus cell clumps	Mode	Not present (354) , present (42)
bacteria	Mode	Not present (374) , present (22)
blood glucose random	Min , Max, Average, Deviation, median , variance ,mode	22 , 490 , 148.037 , 79.282 , 121 , 6285.590 , 99
blood urea	Min , Max, Average, Deviation, median , variance ,mode	2 , 391 , 57.428 , 50.502 , 42 , 2550.498 , 46
serum creatinine	Min , Max, Average, Deviation, median , variance , mode	0.400 , 76 , 3.072 , 5.741 ,1.300 , 32,961 ,1.200
sodium	Min , Max, Average, Deviation, median , variance ,mode	4.500 , 163 , 137.529 , 10.409 , 138, 108.342, 135
potassium	Min , Max, Average, Deviation, median , variance ,mode	2.500 , 47 , 4.627 , 3.194 , 4.400, 10.201 , 3.500
hemoglobin	Min , Max, Average, Deviation, median , variance ,mode	3.100 , 17.800, 12.526 , 2.913 , 12.650, 8.483 , 15
packed cell volume	Min , Max, Average, Deviation, median , variance ,mode	9 , 54 , 38.884 , 8.990 , 40, 80.822 , 41
white blood cell count	Min , Max, Average, Deviation, median , variance ,mode	2200 , 26400 , 8406.122 , 2944.474 , 8000, 8669928.258, 9800
red blood cell count	Min , Max, Average, Deviation, median , variance ,mode	2.100 , 8 , 4.707 , 1.025 , 4.800 , 1.051 , 5.200
hypertension	Mode	Yes (147) , No (251)
diabetes mellitus	Mode	Yes (134) , No (258)
coronary artery disease	Mode	Yes (34) , No (362)
appetite	Mode	Good (317) , poor (82)
pedal edema	Mode	Yes (76) , No (323)
anemia	Mode	Yes (60) , No (339)
class	Mode	Ckd (250) , not ckd (150)

Visualization Techniques

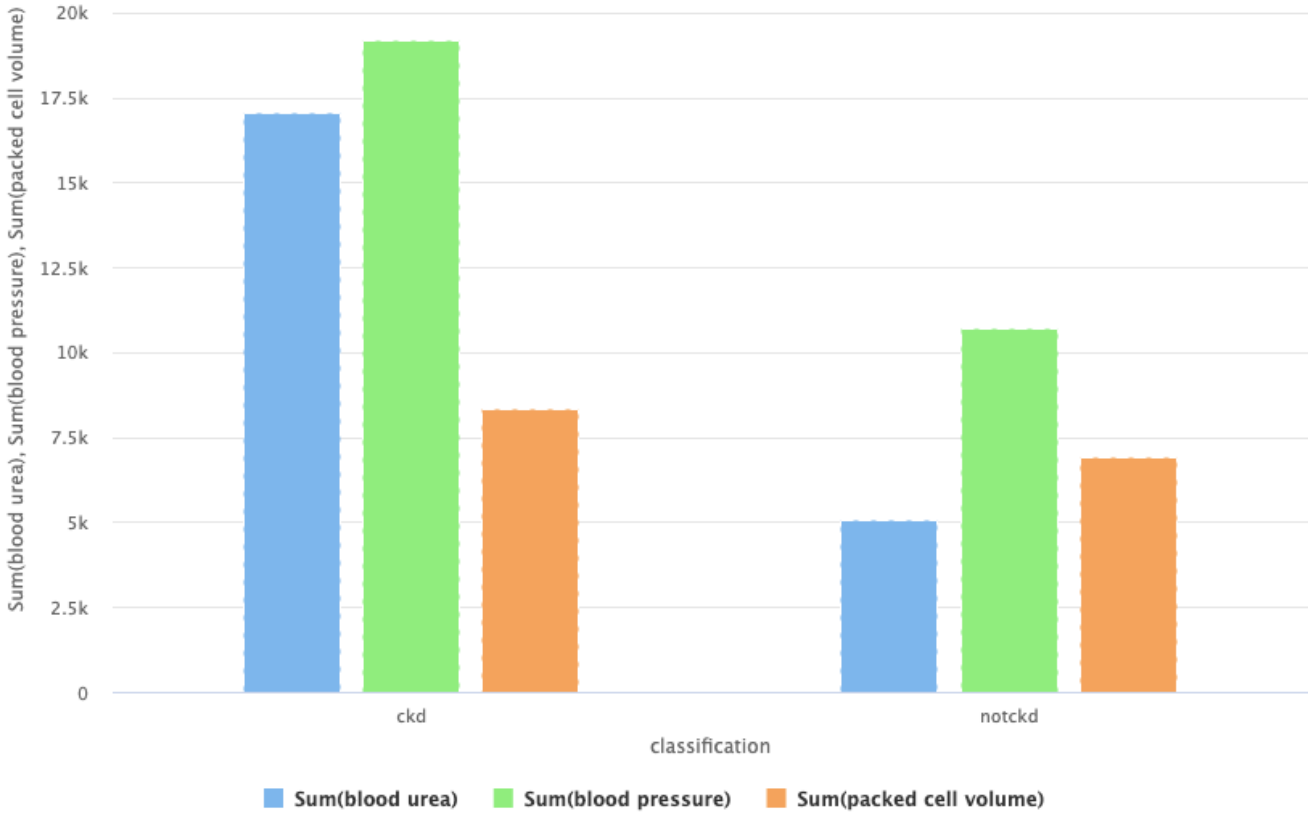
- **Histogram**



People with chronic kidney disease have more hemoglobin in their blood, but notcdk people have a moderate level of hemoglobin.

Visualization Techniques

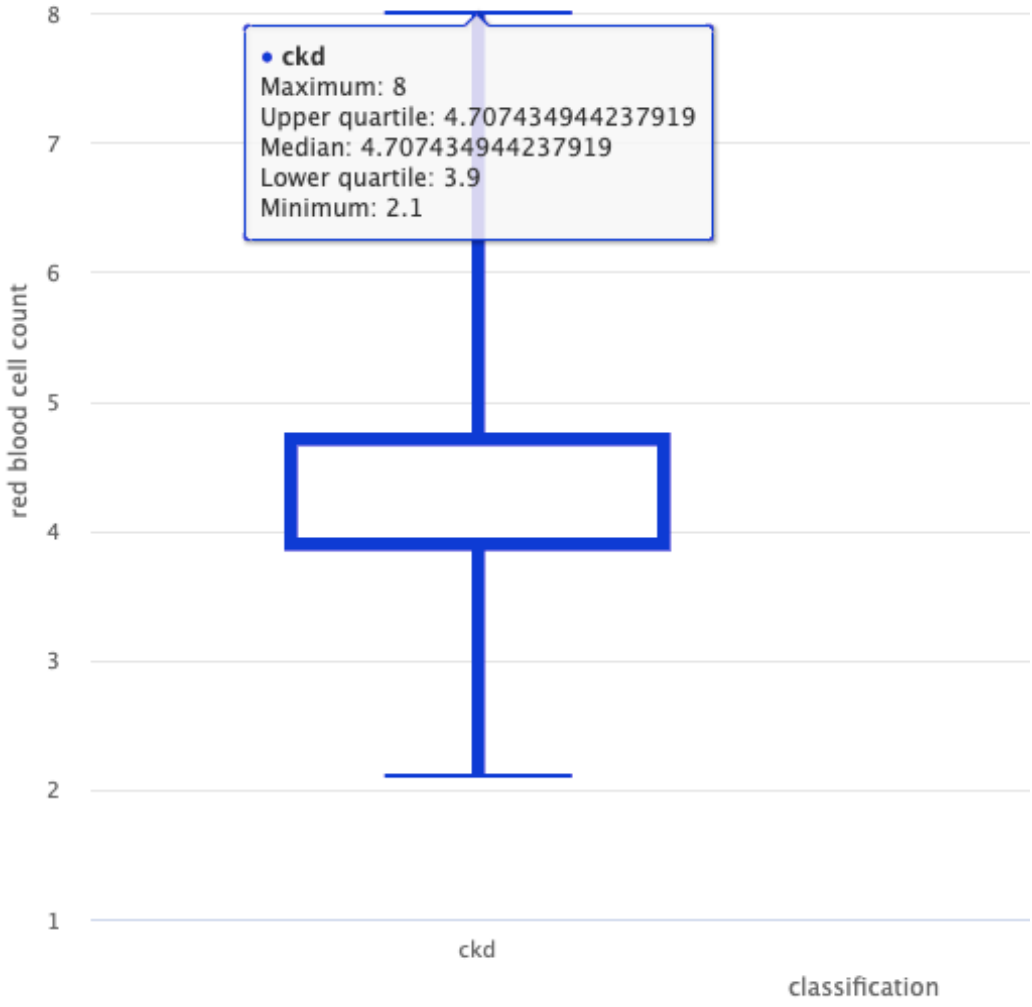
- **Bar Chart**



Blood pressure and blood urea for CKD patients is over 15k which is not the case for NotCKD people but packed cell volume has no effect for CKD patients

Visualization Techniques

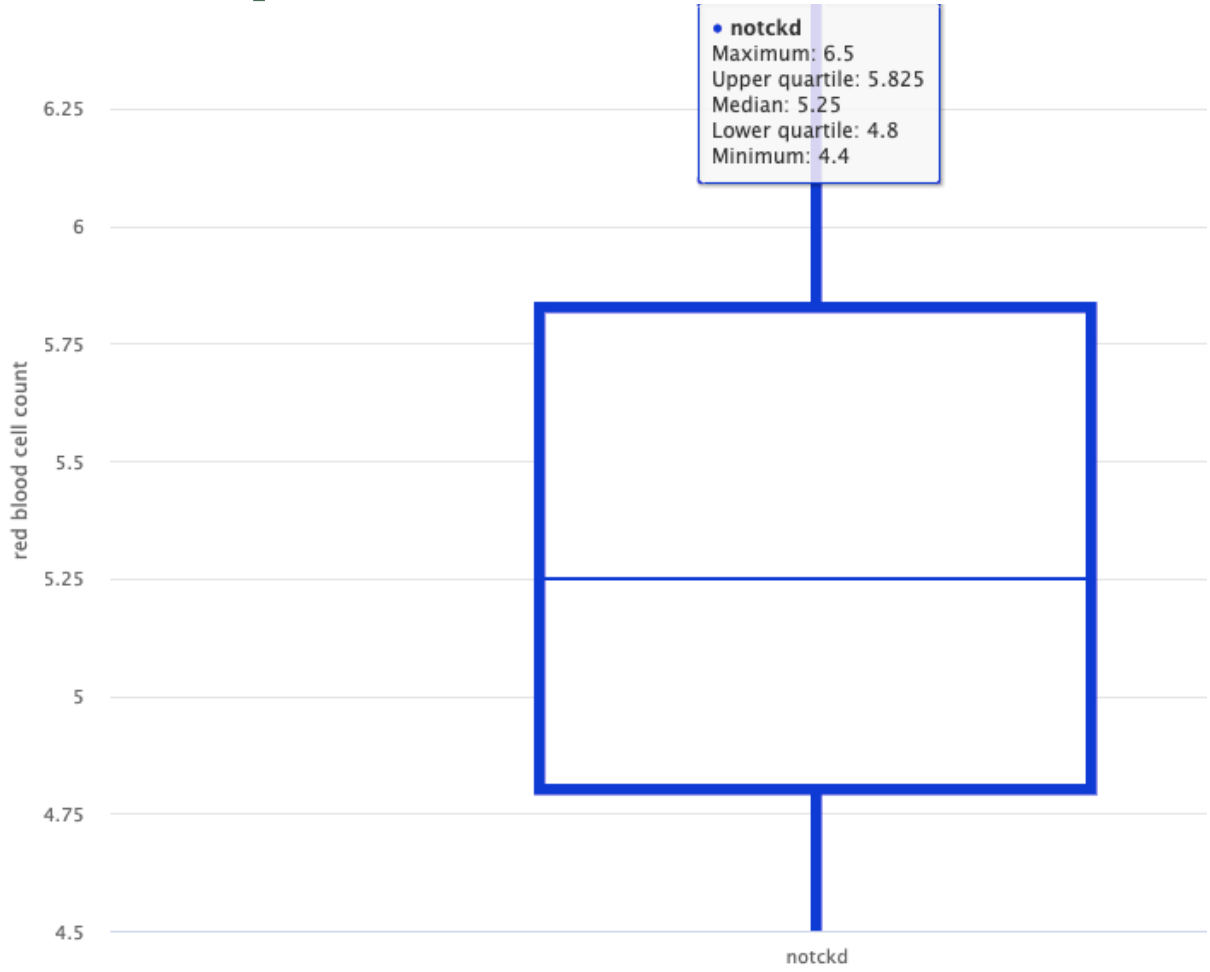
- **Box plot**



After analyzing the given data, found that patients with kidney failure had an average percentage of red blood cells 4.7,And its highest level was8 , It is supernatural, and its lowest level was. 2.1,And it's sub natural.

Visualization Techniques

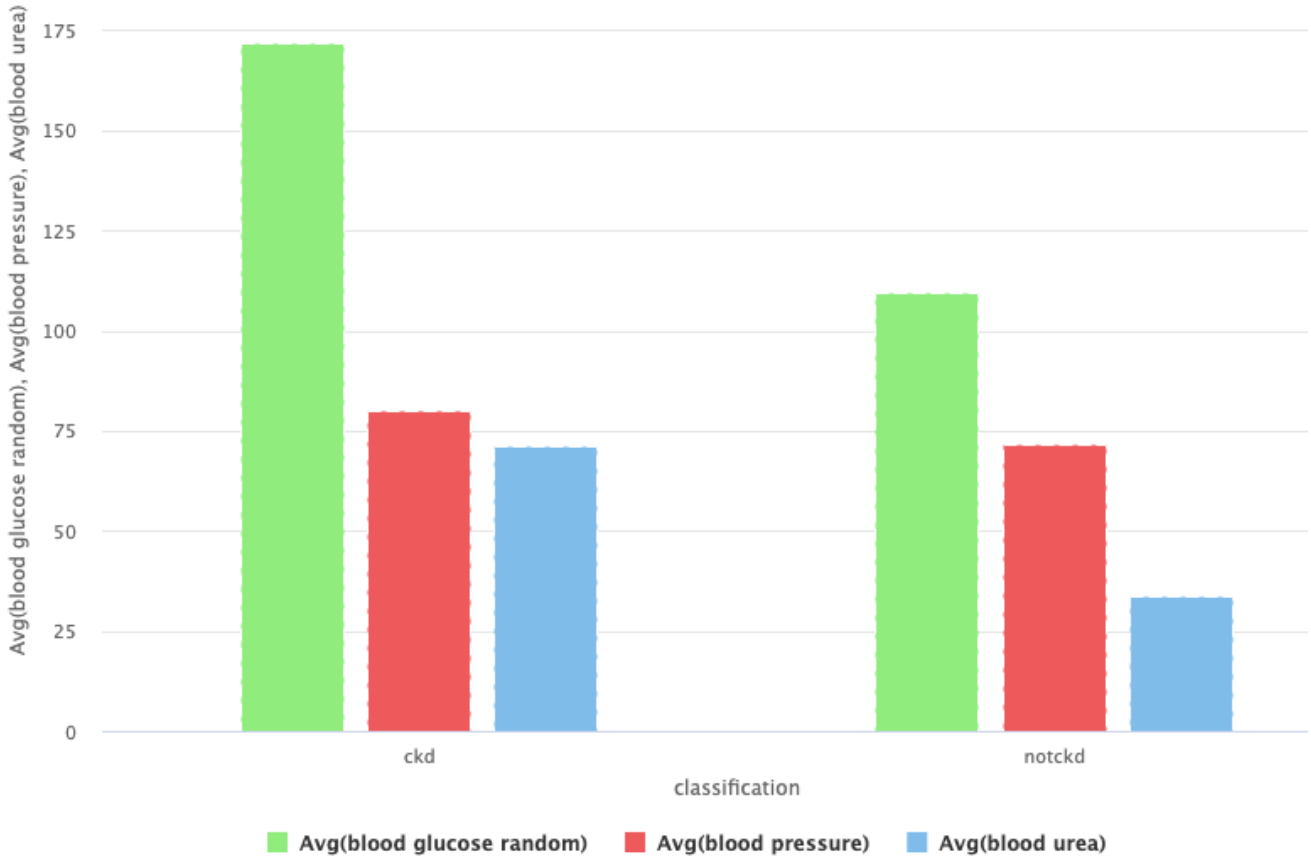
- **Box plot**



The average percentage of red blood cells in patients without kidney failure was 5.2, the greatest level was 6.5, which is the normal level, and the lowest level was 4.4, according to my analysis of the provided data.

Visualization Techniques

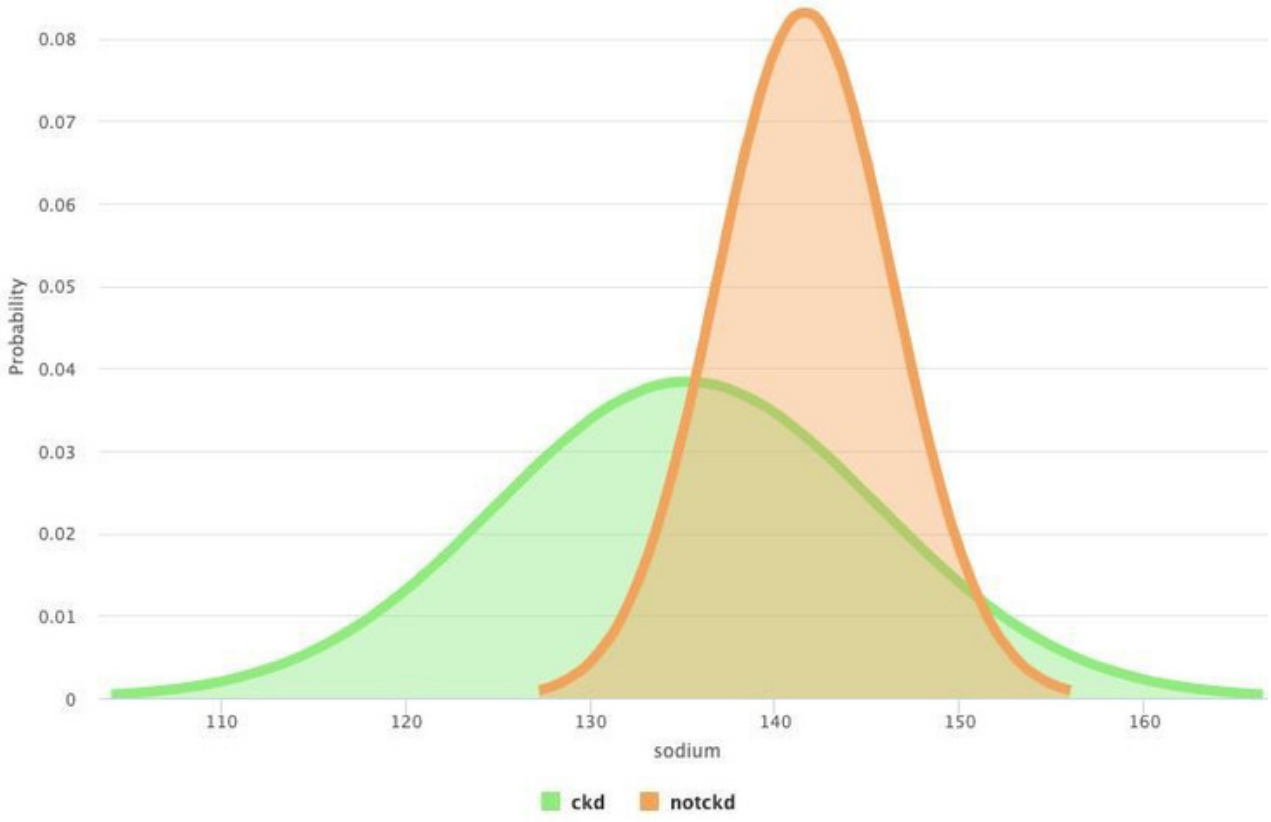
- **Bar Chart**



This chart explains how the avg of blood urea and pressure have the same average levels but for the blood glucose, the CKD patients are so high which is reasonable because having high blood sugar from diabetes can cause damage inside your kidneys.

Visualization Techniques

- **Bell curve**



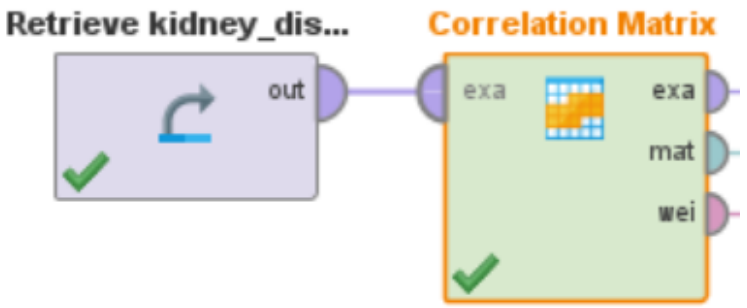
Sodium levels for CKD patients are quite low because CKD patients, elevated blood pressure is a frequent finding and is traditionally considered a direct consequence of their sodium sensitivity.

Correlation

- **Most related attributes:**

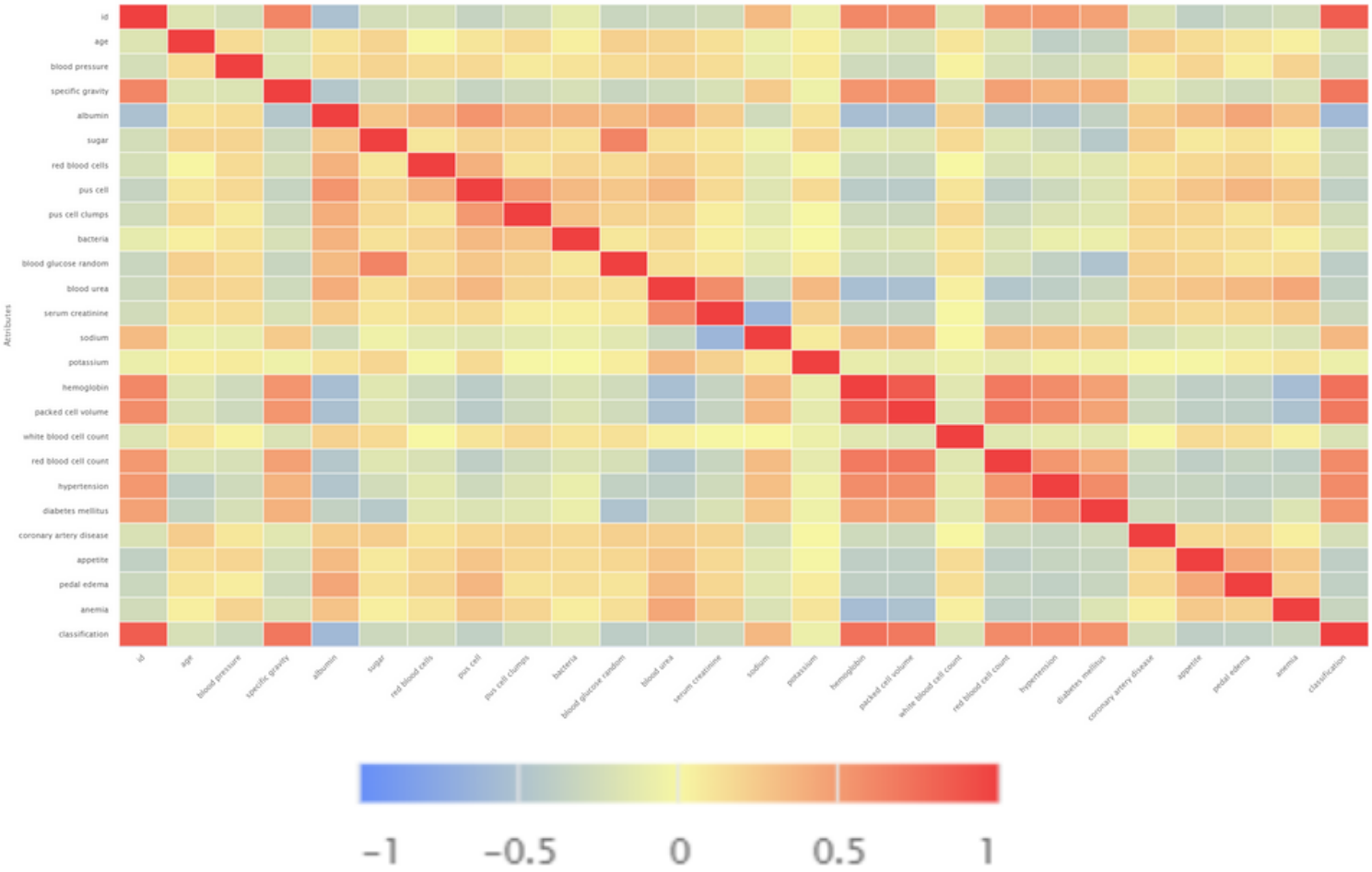
Row No.	hemoglobin	packed cell volume	red blood cell count	hypertension	diabetes mellitus	classification
245	8.600	26	2.500	yes	no	ckd
246	12.526	39	4.707	no	no	ckd
247	12.600	37	4.100	yes	yes	ckd
248	3.100	9	2.100	yes	yes	ckd
249	15	48	4.500	no	no	notckd
250	17	52	5	no	no	notckd
251	15.900	46	4.700	no	no	notckd

- **distance matrix opreation:**



Using Correlation matrix operation to display distance matrix for the related attributes.

Correlation



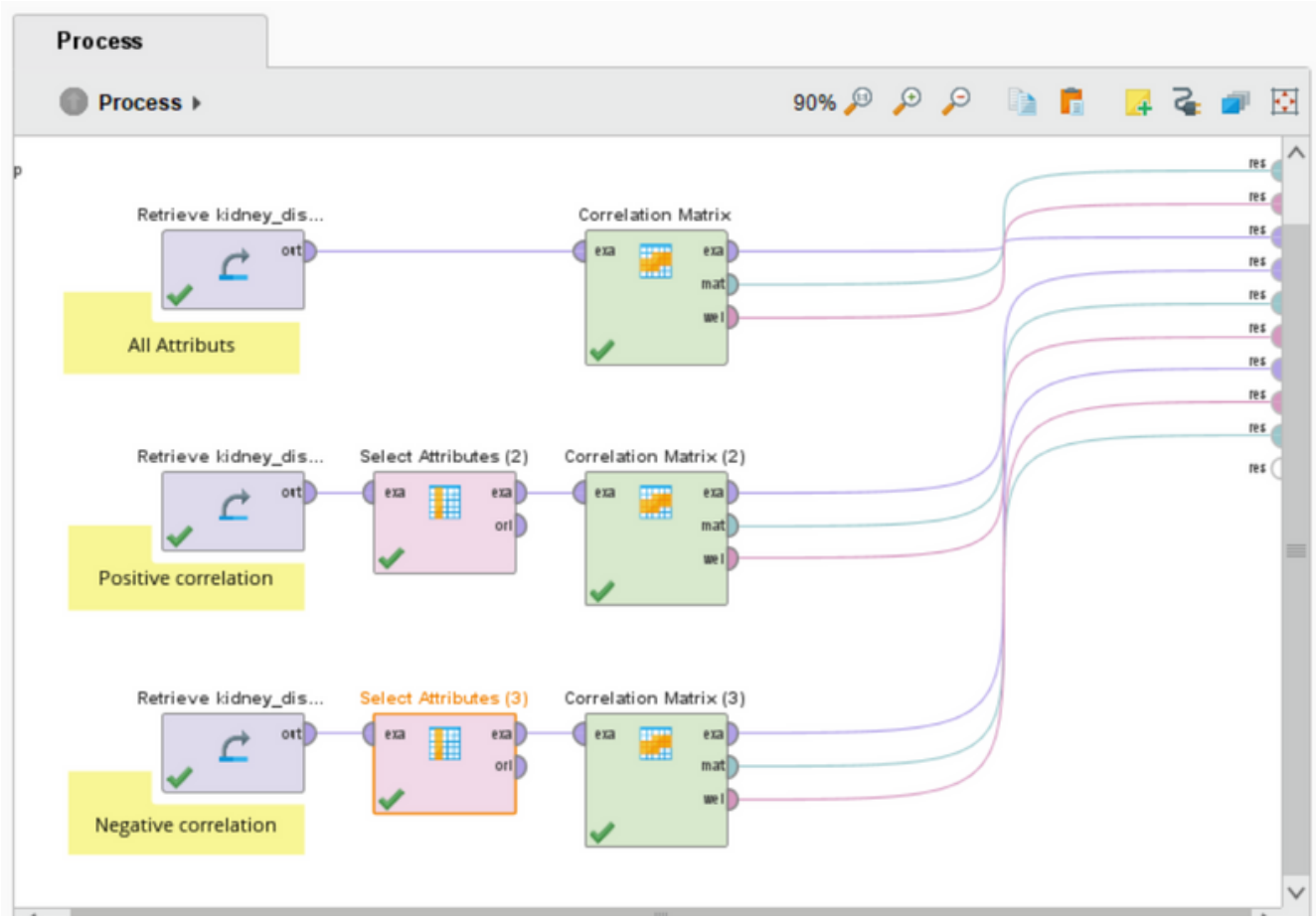
It's easy to find the related attributes from the Matrix Visualization view. each square's color from the cross point represents the degree of correlation between two attributes.

Red Points:
Red squares represent a high Positive correlation between the two attributes.

Yellow Points:
Yellow points mean there is no correlation between attribute pairs.

Blue Points:
Blue squares represent a high Negative correlation between two attributes.

Correlation

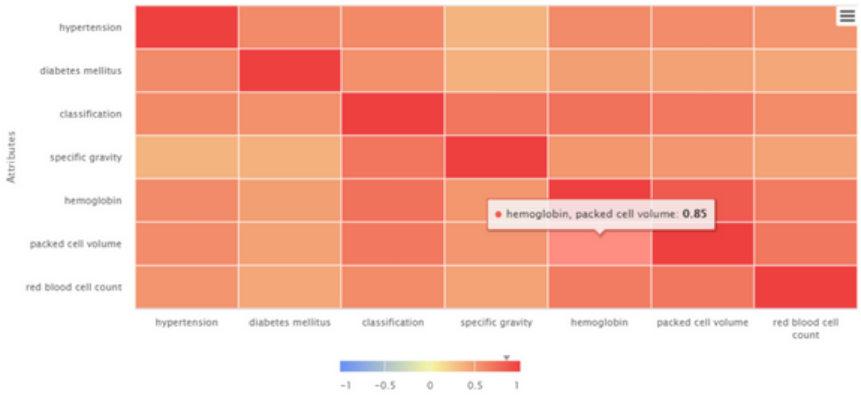


The first correlation matrix is used to discover the strong correlation between attributes pair, by showing the " Matrix Visualization ".

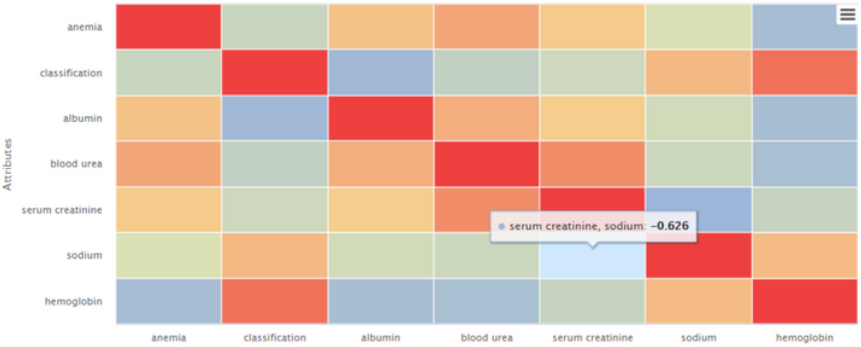
After finding the most related attributes it was selected to visualize the Positive correlation and Negative Correlation Separately in the Correlation matrix(2) and (3).

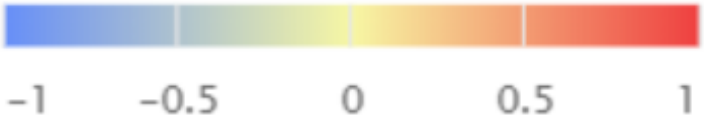
Correlation

First Attribute	Second Attribute	Correlation ↓
hemoglobin	packed cell volume	0.850
classification	hemoglobin	0.727
classification	specific gravity	0.706
packed cell volume	red blood cell count	0.693
classification	packed cell volume	0.686
hemoglobin	red blood cell count	0.674



First Attribute	Second Attribute	Correlation ↑
serum creatinine	sodium	-0.626
classification	albumin	-0.600
anemia	hemoglobin	-0.555
albumin	hemoglobin	-0.546
blood urea	hemoglobin	-0.543
classification	blood urea	-0.373





The most strongly related Attbiutes with High correlation were (hemoglobin, and packed cell volume) with a positive correlation, and (serum creatinine, and sodium) with a negative correlation.

To analyze the relationship between those attributes:

Hemoglobin and packed cell volume (**Positive correlation**): An increase in hemoglobin rate generally means there is a higher packed cell volume rate at blood production.

Serum creatinine and sodium (**Negative correlation**): if the Serum creatinine level was lower and creatinine clearance is higher in the kidneys, the sodium will be at a lower level also.

Note: All these attributes increase the severity of chronic kidney disease (CKD)

Correlation

Attributes	hemoglobin	packed cell volume	red blood cell count	hypertension	diabetes mellitus	classification
hemoglobin	1	0.854	0.682	0.581	0.470	0.730
packed cell volume	0.854	1	0.702	0.571	0.456	0.688
red blood cell count	0.682	0.702	1	0.527	0.421	0.592
hypertension	0.581	0.571	0.527	1	0.587	0.589
diabetes mellitus	0.470	0.456	0.421	0.587	1	0.548
classification	0.730	0.688	0.592	0.589	0.548	1

Attributes	anemia	classifi...	albumin	blood urea	serum creatinine	sodium	hemoglobin
anemia	1	-0.327	0.286	0.442	0.238	-0.204	-0.555
classification	-0.327	1	-0.600	-0.373	-0.295	0.337	0.727
albumin	0.286	-0.600	1	0.402	0.229	-0.267	-0.546
blood urea	0.442	-0.373	0.402	1	0.577	-0.300	-0.543
serum creatinine	0.238	-0.295	0.229	0.577	1	-0.626	-0.343
sodium	-0.204	0.337	-0.267	-0.300	-0.626	1	0.325
hemoglobin	-0.555	0.727	-0.546	-0.543	-0.343	0.325	1

Here to make the result clear by the language of the numbers you can find a distance matrix that includes the values of correlation for every two attributes.

The tables above display the correlation Matrix of strongly related attributes. hemoglobin and packed cell volume, serum creatinine, and sodium have the highest correlation at the Matrix.

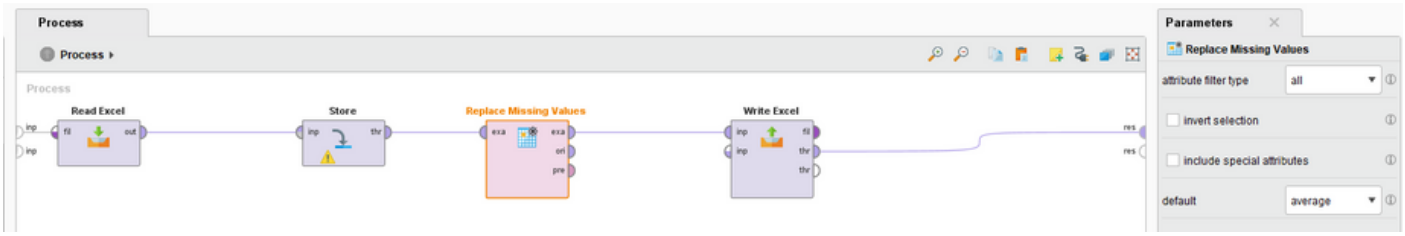
Data Cleaning

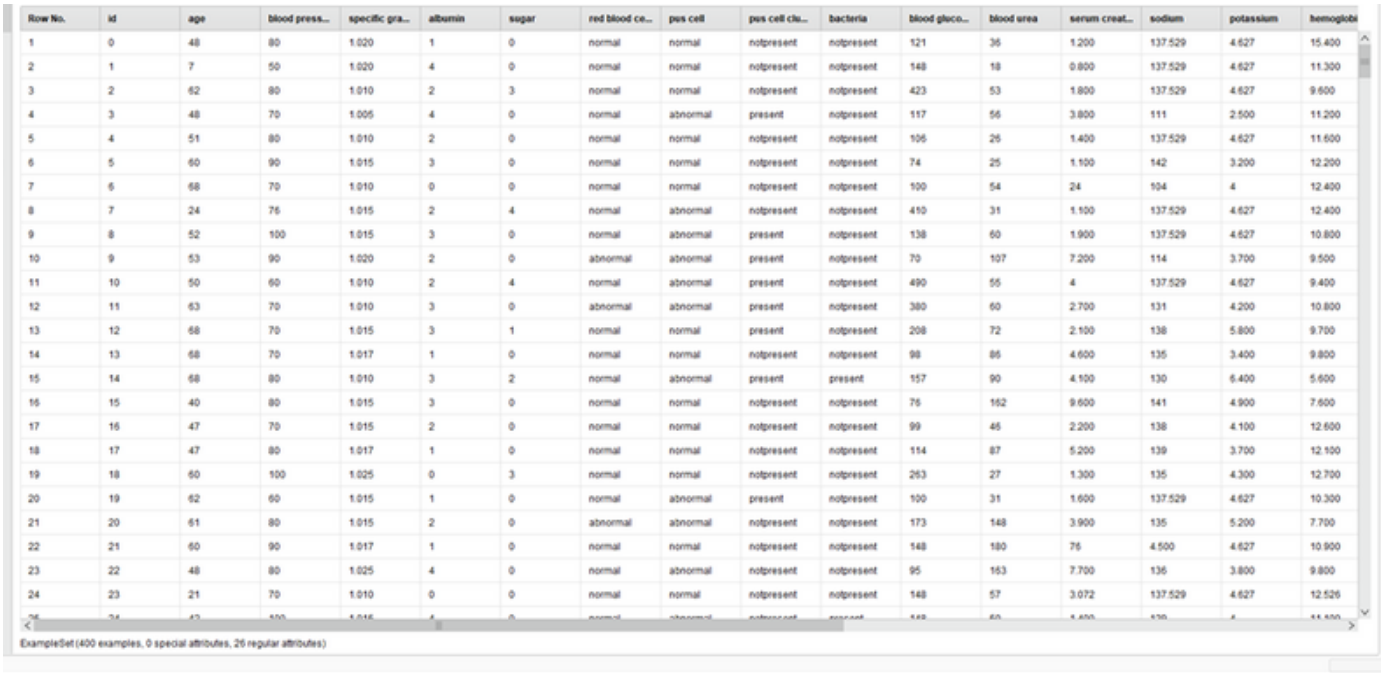
- **Fill the Missing Values:**

  As we noticed that we have a set of rows with missing data.These fields were filled in two operator , the first one by Replace Missing Values and the second Impute Missing Values . Each way gave different results .

- **Replace Missing Values:**

  This operator quickly fills in all missing values, and this missing data is replaced by one of these methods by either filling it in with the mean, maximum, minimum, or value.





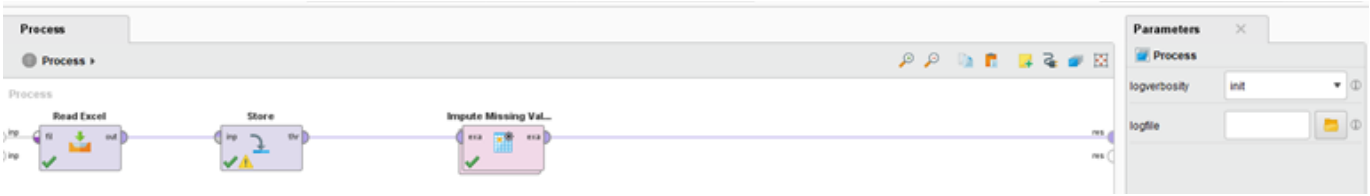
Row No.	id	age	blood press...	specific gra...	albumin	sugar	red blood ce...	pus cell	pus cell clu...	bacteria	blood gluco...	blood urea	serum creat...	sodium	potassium	hemoglobi
1	0	48	80	1.020	1	0	normal	normal	notpresent	notpresent	121	36	1.200	137.529	4.627	15.400
2	1	7	50	1.020	4	0	normal	normal	notpresent	notpresent	148	18	0.800	137.529	4.627	11.300
3	2	62	80	1.010	2	3	normal	normal	notpresent	notpresent	423	53	1.800	137.529	4.627	9.600
4	3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.800	111	2.500	11.200
5	4	51	80	1.010	2	0	normal	normal	notpresent	notpresent	106	26	1.400	137.529	4.627	11.600
6	5	60	90	1.015	3	0	normal	normal	notpresent	notpresent	74	25	1.100	142	3.200	12.200
7	6	68	70	1.010	0	0	normal	normal	notpresent	notpresent	100	54	24	104	4	12.400
8	7	24	76	1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.100	137.529	4.627	12.400
9	8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.900	137.529	4.627	10.800
10	9	53	90	1.020	2	0	abnormal	abnormal	present	notpresent	70	107	7.200	114	3.700	9.500
11	10	50	60	1.010	2	4	normal	abnormal	present	notpresent	490	55	4	137.529	4.627	9.400
12	11	63	70	1.010	3	0	abnormal	abnormal	present	notpresent	380	60	2.700	131	4.200	10.800
13	12	68	70	1.015	3	1	normal	normal	present	notpresent	208	72	2.100	138	5.800	9.700
14	13	68	70	1.017	1	0	normal	normal	notpresent	notpresent	98	86	4.600	135	3.400	9.800
15	14	68	80	1.010	3	2	normal	abnormal	present	present	157	90	4.100	130	6.400	5.600
16	15	40	80	1.015	3	0	normal	normal	notpresent	notpresent	76	162	9.600	141	4.900	7.600
17	16	47	70	1.015	2	0	normal	normal	notpresent	notpresent	99	46	2.200	138	4.100	12.600
18	17	47	80	1.017	1	0	normal	normal	notpresent	notpresent	114	87	5.200	139	3.700	12.100
19	18	60	100	1.025	0	3	normal	normal	notpresent	notpresent	263	27	1.300	135	4.300	12.700
20	19	62	60	1.015	1	0	normal	abnormal	present	notpresent	100	31	1.600	137.529	4.627	10.300
21	20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.900	135	5.200	7.700
22	21	60	90	1.017	1	0	normal	normal	notpresent	notpresent	148	180	76	4.500	4.627	10.900
23	22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.700	136	3.800	9.800
24	23	21	70	1.010	0	0	normal	normal	notpresent	notpresent	148	57	3.072	137.529	4.627	12.526

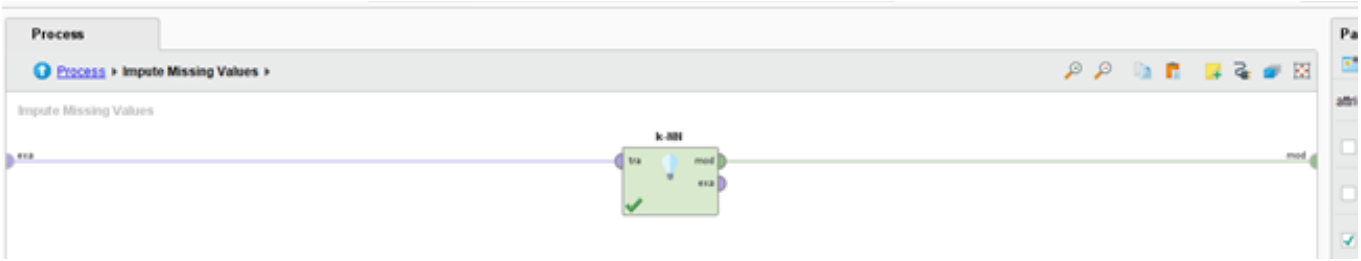
ExampleSet (400 examples, 0 special attributes, 26 regular attributes)

Data Cleaning

- **Impute Missing Value (K-NN) :**

This operator took some time and fill in all the missing values. This operator is uses the nearest neighbor algorithm, which means that it does not fill everything that is missing with one value, but with different values according to the missing value.





Row No.	id	age	blood press...	specific gra...	albumin	sugar	red blood ce...	pus cell	pus cell clu...	bacteria	blood gluco...	blood urea	serum creat...	sodium	potassium	hemoglobi
1	0	48	80	1.020	1	0	normal	normal	notpresent	notpresent	121	36	1.200	142.588	4.023	15.400
2	1	7	50	1.020	4	0	normal	normal	notpresent	notpresent	98.409	18	0.800	141.054	4.308	11.300
3	2	62	80	1.010	2	3	normal	normal	notpresent	notpresent	423	53	1.800	140.282	4.815	9.600
4	3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.800	111	2.500	11.200
5	4	51	80	1.010	2	0	normal	normal	notpresent	notpresent	106	26	1.400	145.000	4.305	11.600
6	5	60	90	1.015	3	0	normal	normal	notpresent	notpresent	74	25	1.100	142	3.200	12.200
7	6	68	70	1.010	0	0	normal	normal	notpresent	notpresent	100	54	24	104	4	12.400
8	7	24	84.370	1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.100	139.811	4.134	12.400
9	8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.900	138.555	4.033	10.800
10	9	53	90	1.020	2	0	abnormal	abnormal	present	notpresent	70	107	7.200	114	3.700	9.500
11	10	50	60	1.010	2	4	abnormal	abnormal	present	notpresent	490	55	4	133.245	4.090	9.400
12	11	63	70	1.010	3	0	abnormal	abnormal	present	notpresent	380	60	2.700	131	4.200	10.800
13	12	68	70	1.015	3	1	abnormal	normal	present	notpresent	208	72	2.100	138	5.800	9.700
14	13	68	70	1.016	1.618	0	normal	normal	notpresent	notpresent	98	86	4.600	135	3.400	9.800
15	14	68	80	1.010	3	2	normal	abnormal	present	present	157	90	4.100	130	6.400	5.600
16	15	40	80	1.015	3	0	normal	normal	notpresent	notpresent	76	162	9.600	141	4.900	7.600
17	16	47	70	1.015	2	0	normal	normal	notpresent	notpresent	99	46	2.200	138	4.100	12.600
18	17	47	80	1.017	1.818	0	normal	normal	notpresent	notpresent	114	87	5.200	139	3.700	12.100
19	18	60	100	1.025	0	3	normal	normal	notpresent	notpresent	263	27	1.300	135	4.300	12.700
20	19	62	60	1.015	1	0	normal	abnormal	present	notpresent	100	31	1.600	136.806	3.691	10.300
21	20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.900	135	5.200	7.700
22	21	60	90	1.021	0.855	0	normal	normal	notpresent	notpresent	119.495	180	76	4.500	4.614	10.900
23	22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.700	136	3.800	9.800
24	23	21	70	1.010	0	0	normal	normal	notpresent	notpresent	110.935	35.487	1.002	141.642	4.078	14.749

ExampleSet (400 examples, 0 special attributes, 26 regular attributes)

Data Cleaning

- **Results of 2 operations after Fill in missing values:**
  - **Replace Missing Values:**

ExampleSet (Replace Missing Values)

Name	Type	Missing	Statistics		
id	Polynominal	0	Least: 99 (1)	Most: 0 (1)	Values: 0 (1), 1 (1), ... [398 more]
age	Integer	0	Min: 2	Max: 90	Average: 51.472
blood pressure	Integer	0	Min: 50	Max: 180	Average: 76.455
specific gravity	Real	0	Min: 1.005	Max: 1.025	Average: 1.017
albumin	Integer	0	Min: 0	Max: 5	Average: 1.015
sugar	Integer	0	Min: 0	Max: 5	Average: 0.395
red blood cells	Polynominal	0	Least: abnormal (47)	Most: normal (353)	Values: normal (353), abnormal (47)
pus cell	Polynominal	0	Least: abnormal (76)	Most: normal (324)	Values: normal (324), abnormal (76)
pus cell clumps	Polynominal	0	Least: present (42)	Most: notpresent (358)	Values: notpresent (358), present (42)
bacteria	Polynominal	0	Least: present (22)	Most: notpresent (378)	Values: notpresent (378), present (22)
blood glucose random	Integer	0	Min: 22	Max: 490	Average: 148.032
blood urea	Integer	0	Min: 2	Max: 391	Average: 57.407
[illegible]	Real	0	Min: 0.400	Max: 76	Average: 3.072

Showing attributes 1 - 26 — Examples: 400 Special Attributes: 0 Regular Attributes: 26

  - **Impute Missing Value (K-NN) :**

ExampleSet (Impute Missing Values)

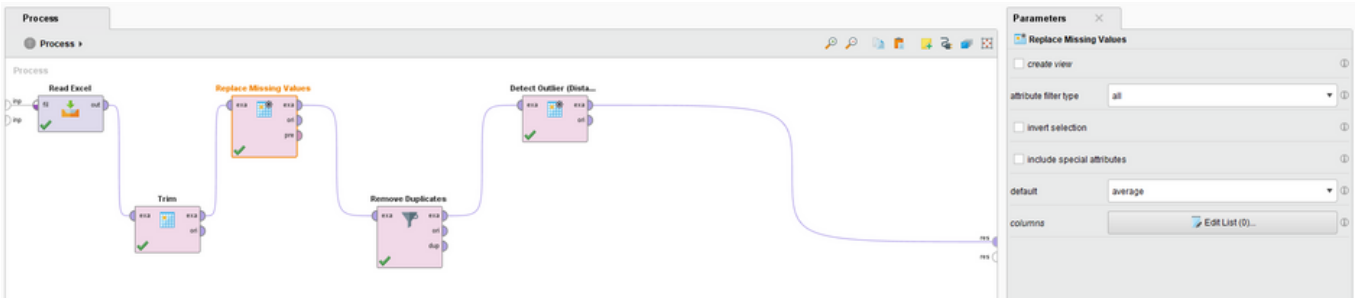
Name	Type	Missing	Statistics		
id	Polynominal	0	Least: 99 (1)	Most: 0 (1)	Values: 0 (1), 1 (1), ... [398 more]
age	Integer	0	Min: 2	Max: 90	Average: 51.527
blood pressure	Integer	0	Min: 50	Max: 180	Average: 76.460
specific gravity	Real	0	Min: 1.005	Max: 1.025	Average: 1.018
albumin	Integer	0	Min: 0	Max: 5	Average: 1.059
sugar	Integer	0	Min: 0	Max: 5	Average: 0.436
red blood cells	Polynominal	0	Least: abnormal (58)	Most: normal (342)	Values: normal (342), abnormal (58)
pus cell	Polynominal	0	Least: abnormal (82)	Most: normal (318)	Values: normal (318), abnormal (82)
pus cell clumps	Polynominal	0	Least: present (42)	Most: notpresent (358)	Values: notpresent (358), present (42)
bacteria	Polynominal	0	Least: present (22)	Most: notpresent (378)	Values: notpresent (378), present (22)
blood glucose random	Integer	0	Min: 22	Max: 490	Average: 147.850
blood urea	Integer	0	Min: 2	Max: 391	Average: 57.166
[illegible]	Real	0	Min: 0.400	Max: 76	Average: 3.042

Showing attributes 1 - 26 — Examples: 400 Special Attributes: 0 Regular Attributes: 26

Data Cleaning

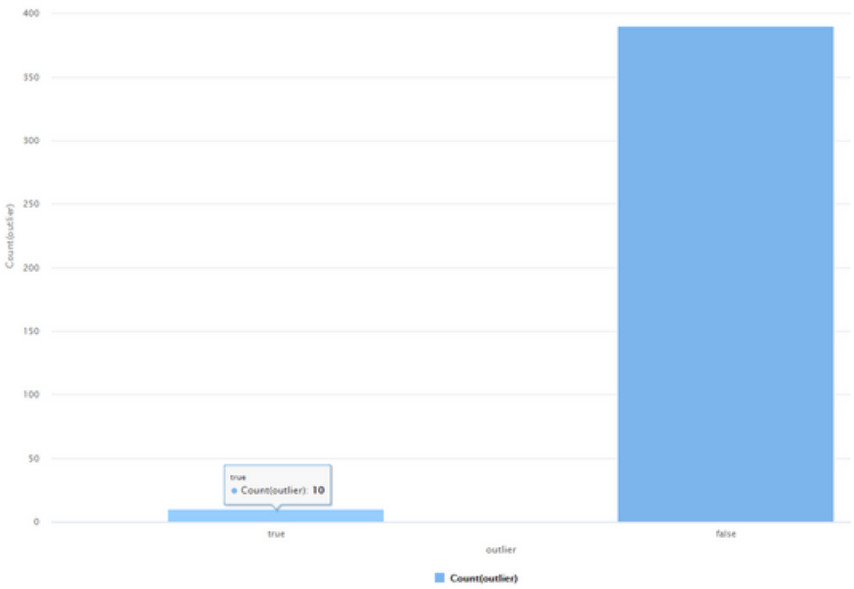
- **Replace Missing Values , Identify outliers and smooth out noisy data :**

In our project to clean the data set well, we filled in the missing values by replacing any missing value with the mean value, and also we put an operator that removes all duplicate values , also we put an operator that shows any outliers (distances) we have in the data set that was chosen for this The project.



- **Outliers results for this data set :**

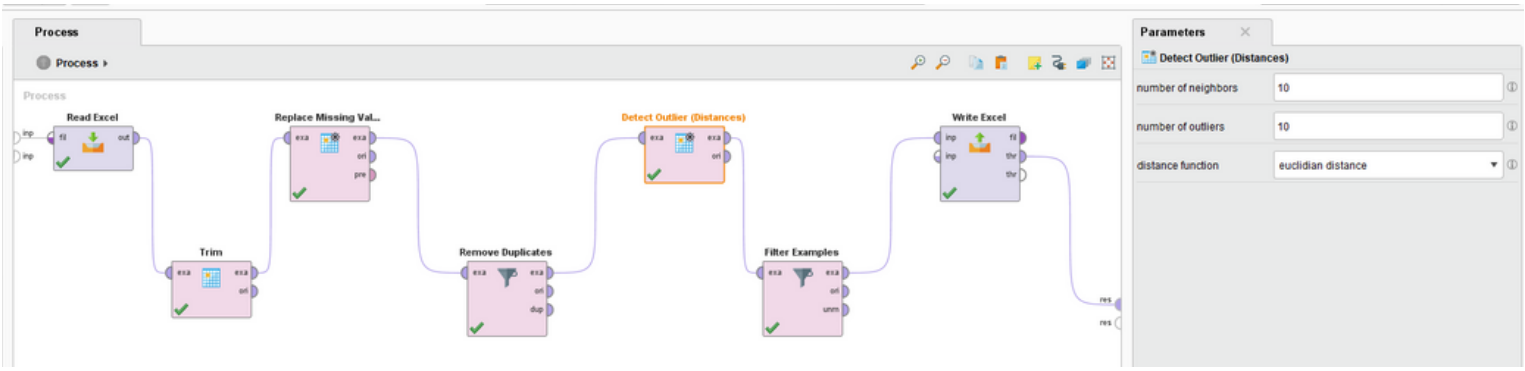
As we can see, we have 10 outliers ((TRUE)) and 390 values that are not outliers ((FALSE))

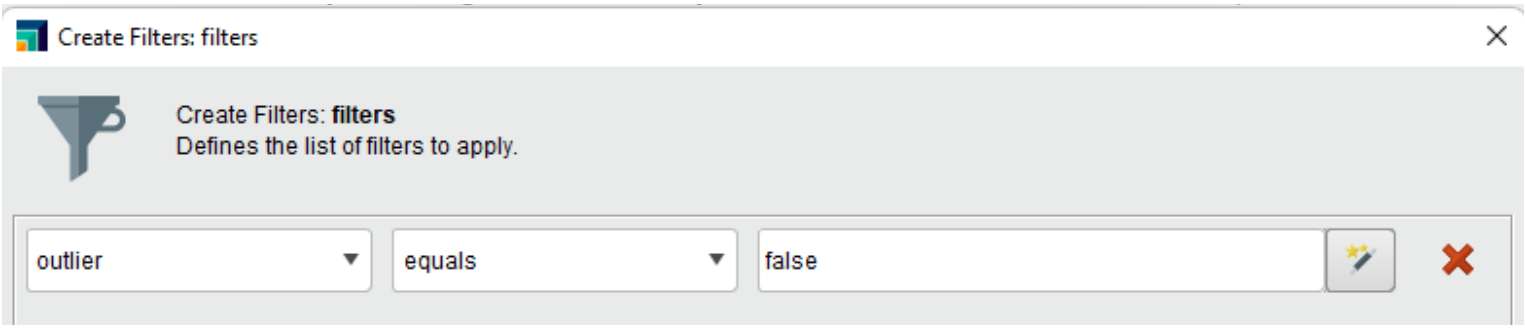


Data Cleaning

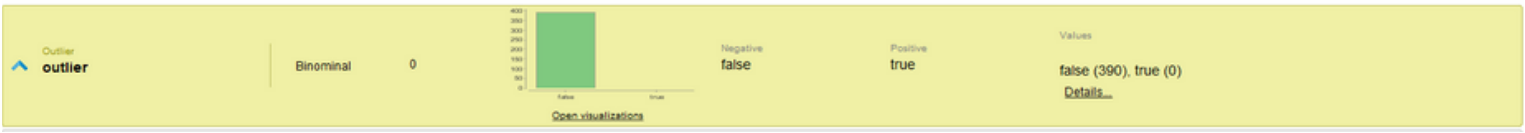
- **Dealing with outliers and smooth out noisy data :**

To deal with outliers, we added an operator that filters all the outliers .

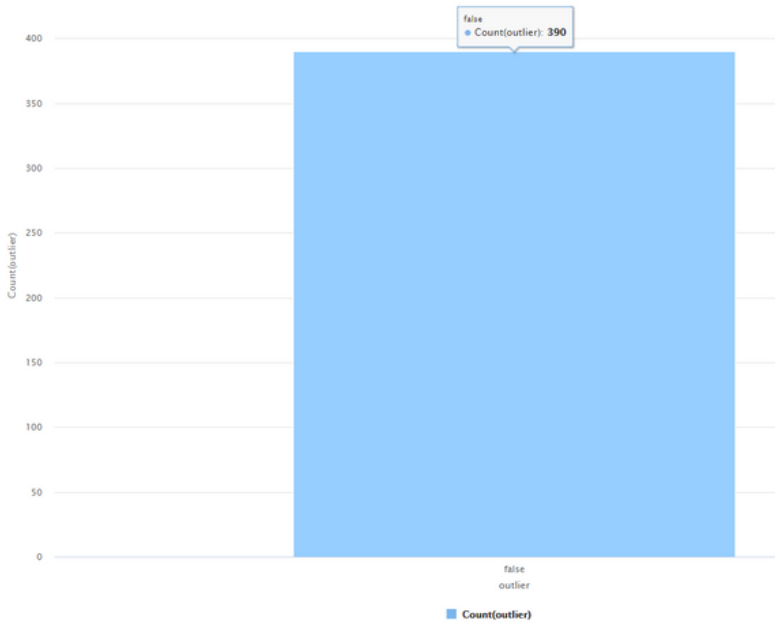




- **Outliers results for this data set :**



As we can see, we don't have any outliers ((TRUE)) and 390 values that are not outliers ((FALSE))



Phase 2

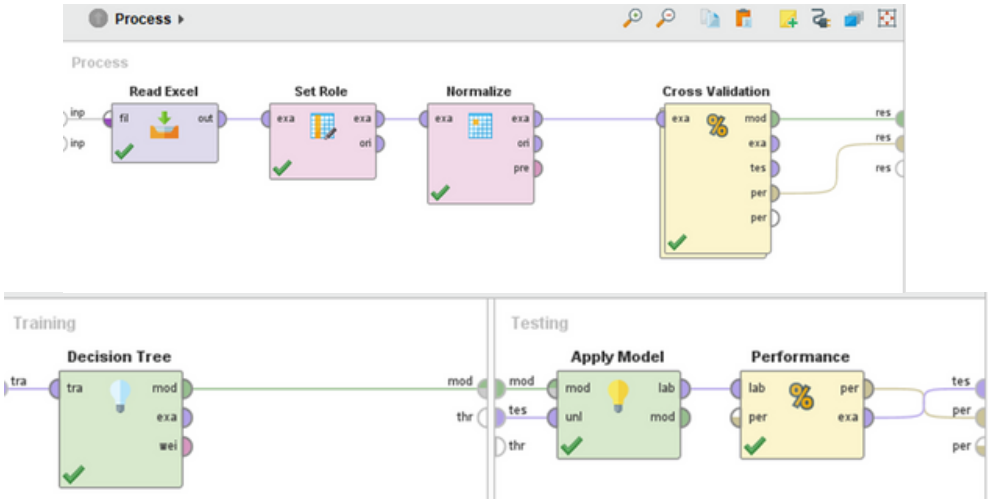
Outline:

- **Measure the accuracy**
  - Normalizing Data (Range & Z-score)
  - Imputing Missing Values
  - Discretizing the Numeric Attributes
  - Reducing dimensions
  - Pre-processing Steps :
    - Find a combination
  - Neural Network classifier
  - K-NN
  - Naive Bayes
  - Random Forest
- **Result**

Measure the accuracy

Measure accuracy for Classification and prediction with data cleaning, after the pre-pressing process we need to classify the group status to predict them using a decision tree classifier.

A.1. Normalizing the data (z-score normalizations).



accuracy: 96.67% +/- 2.43% (micro average: 96.67%)

	true ckd	true notckd	class precision
pred. ckd	237	10	95.95%
pred. notckd	3	140	97.90%
class recall	98.75%	93.33%	

We can see in the confusion matrix

Accuracy is: 96.67%
Classification_error: 3.33%
Class recall CKD: 98.75%
Class recall NotCKD: 93.33%

The accuracy here was high and excellent, the error rate was very small, and the correlation is 0.930. So, the correlation is strong

Measure the accuracy

A.2. Normalizing the data (Range normalizations).

accuracy: 96.67% +/- 2.43% (micro average: 96.67%)

	true ckd	true notckd	class precision
pred. ckd	237	10	95.95%
pred. notckd	3	140	97.90%
class recall	98.75%	93.33%	

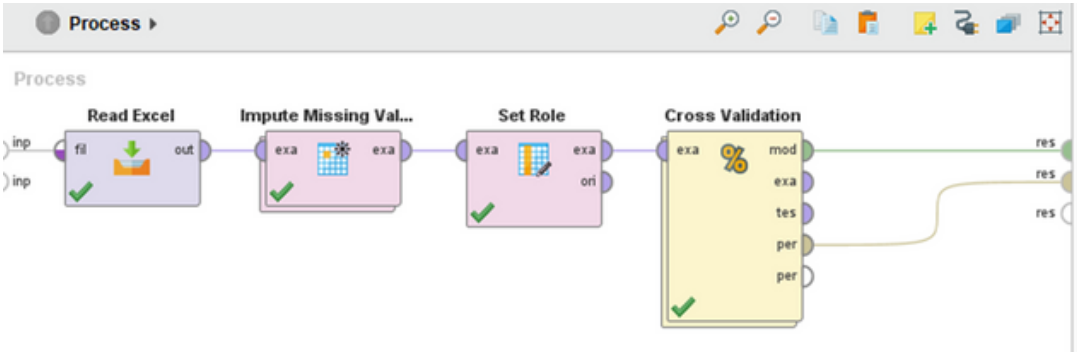
We can see in the confusion matrix

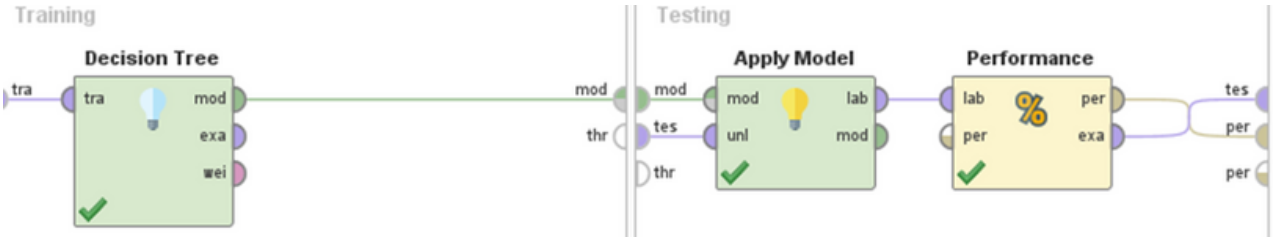
Accuracy is: 96.67%
Classification_error: 3.33%
Class recall CKD: 98.75%
Class recall NotCKD: 93.33%

The accuracy here was high and excellent, the error rate was very small, and the correlation is 0.930. So, the correlation is strong

Measure the accuracy

B.1 Accuracy after imputing missing values





accuracy: 95.90% +/- 0.73% (micro average: 95.90%)

	true ckd	true notckd	class precision
pred. ckd	231	7	97.06%
pred. notckd	9	143	94.08%
class recall	96.25%	95.33%	

We can see in the confusion matrix

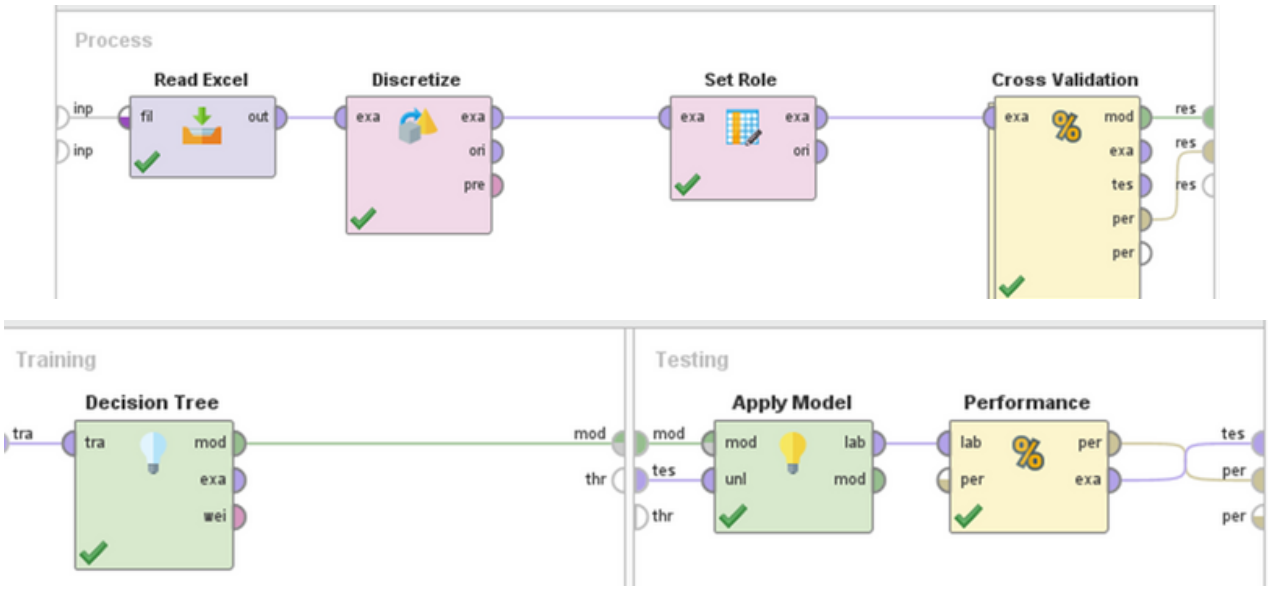
Accuracy is: 95.90%
Classification error: 4.10%
Class recall CKD: 96.25%
Class recall NotCKD : 95.33%

The accuracy here was high and excellent, and the error rate was very small ,and the correlation is 0.915, so the correlation is strong.

Measure the accuracy

C. Accuracy after discretizing the numeric attributes

The size of bins parameter is used for specifying the required size of bins. This discretization is performed by binning examples into bins containing the same, user-specified number of examples. Each bin range is named automatically



C1.Bins 2

accuracy: 91.79% +/- 11.13% (micro average: 91.79%)

	true ckd	true notckd	class precision
pred. ckd	226	18	92.62%
pred. notckd	14	132	90.41%
class recall	94.17%	88.00%	

We can see in the confusion matrix

Accuracy is: 91.79%

Classification_error: 8.21%

Class recall CKD: 96.25%

Class recall NotCKD: 95.%

Measure the accuracy

C. Accuracy after discretizing the numeric attributes

C2.Bins 3

accuracy: 92.56% +/- 5.05% (micro average: 92.56%)

	true ckd	true notckd	class precision
pred. ckd	220	9	96.07%
pred. notckd	20	141	87.58%
class recall	91.67%	94.00%	

Accuracy is: 92.56%%

Classification error: 7.44%

Class recall CKD: 91.67%

Class recall NotCKD: 94.00%

C3.Bins 4

accuracy: 91.28% +/- 10.96% (micro average: 91.28%)

	true ckd	true notckd	class precision
pred. ckd	225	19	92.21%
pred. notckd	15	131	89.73%
class recall	93.75%	87.33%	

Accuracy is: 91.28%

Classification error: 8.72%

Class recall CKD: 93.75%

Class recall NotCKD: 87.33%

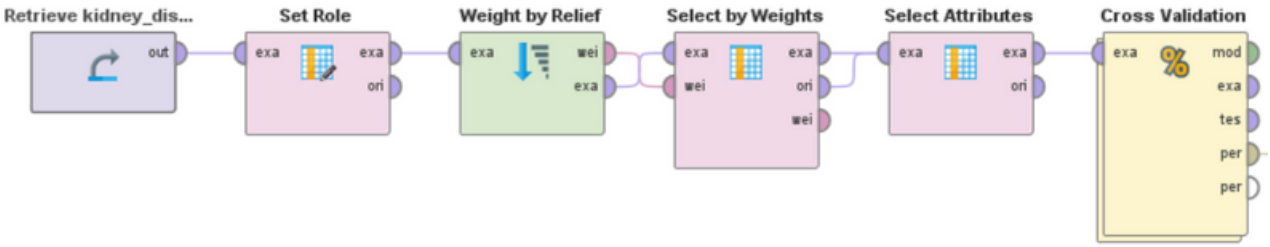
We note that the accuracy in Bins 3 is slightly higher than 2 and 4, but both have excellent accuracy and very little error.

Measure the accuracy

D1. Accuracy after reducing dimensions (Weight by Relief and Select by weights)

The Relief calculates the relevance of the attributes by Relief.

The key idea of Relief is to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes that are near each other.



accuracy: 97.18% +/- 2.82% (micro average: 97.18%)

	true ckd	true notckd	class precision
pred. ckd	231	2	99.14%
pred. notckd	9	148	94.27%
class recall	96.25%	98.67%	

Can see in the confusion matrix

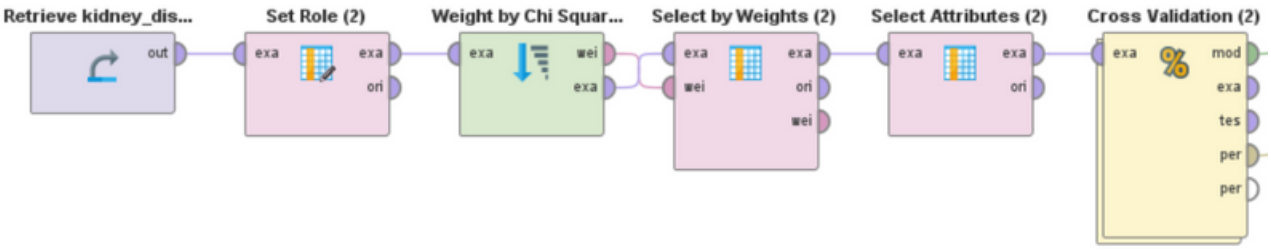
Weight is 1.0
Accuracy is: 97.18%
Classification error:2.82%
Class recall CKD: 96.25%
Class recall NotCKD: 98.67%

The accuracy here is high and excellent, the error rate is small, and the correlation is 0.942. So, the correlation is strong.

Measure the accuracy

D.2 Accuracy after reducing dimensions (Weight by Chi Squared Statistic and Select by weights)

Weight by Chi Squared Statistic calculates the relevance of the attributes selected from a dataset by computing for each attribute of the input ExampleSet the value of the chi-squared statistic with respect to the class attribute.



accuracy: 96.67% +/- 2.97% (micro average: 96.67%)

	true ckd	true notckd	class precision
pred. ckd	236	9	96.33%
pred. notckd	4	141	97.24%
class recall	98.33%	94.00%	

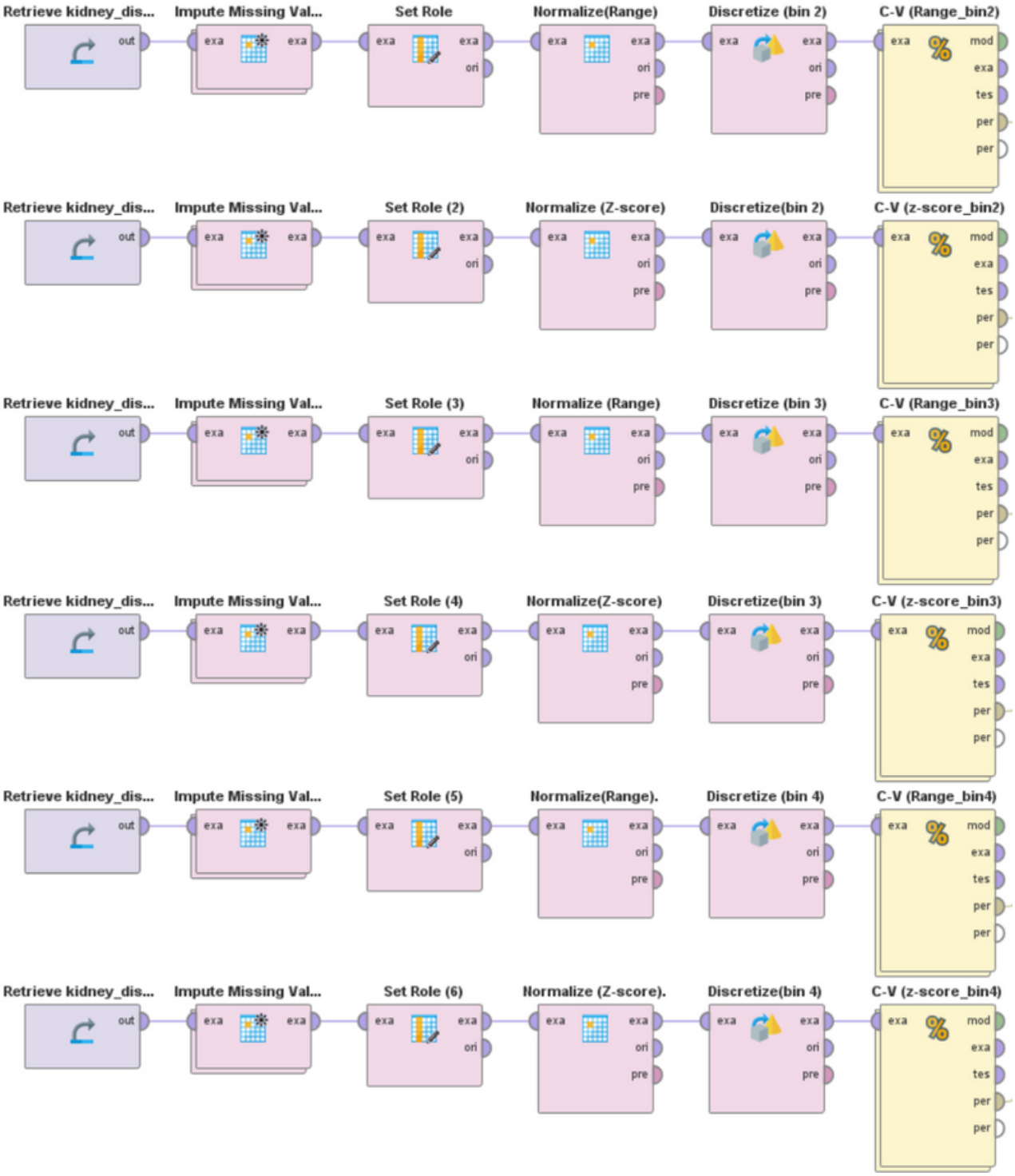
We can see in the confusion matrix

Weight is 1.0
Accuracy is: 96.67%
Classification error: 3.33%
Class recall CKD: 98.33%
Class recall NotCKD: 94.00%

The accuracy here was high and excellent, the error rate was very small, and the correlation is 0.929 . So the correlation is strong.

Measure the accuracy

E. Using all of the pre-processing steps excluding dimensionality reduction



- The information gained from all preprocessing steps **except** for dimension reduction

normalize	bin	Accuracy	Classification error	Class recall CKD	Class recall Not CKD	correlation
Range	2	**97.95%**	**2.05%**	**96.67%**	**100.00%**	**0.958**
	3	96.41%	3.59%	96.67%	96.00%	0.924
	4	96.92%	3.08%	96.25%	98.00%	0.936
Z-score	2	97.69%	2.31%	96.25%	100.00%	0.953
	3	96.41%	3.59%	94.58%	99.33%	0.927
	4	96.92%	3.08%	95.83%	98.67%	0.937

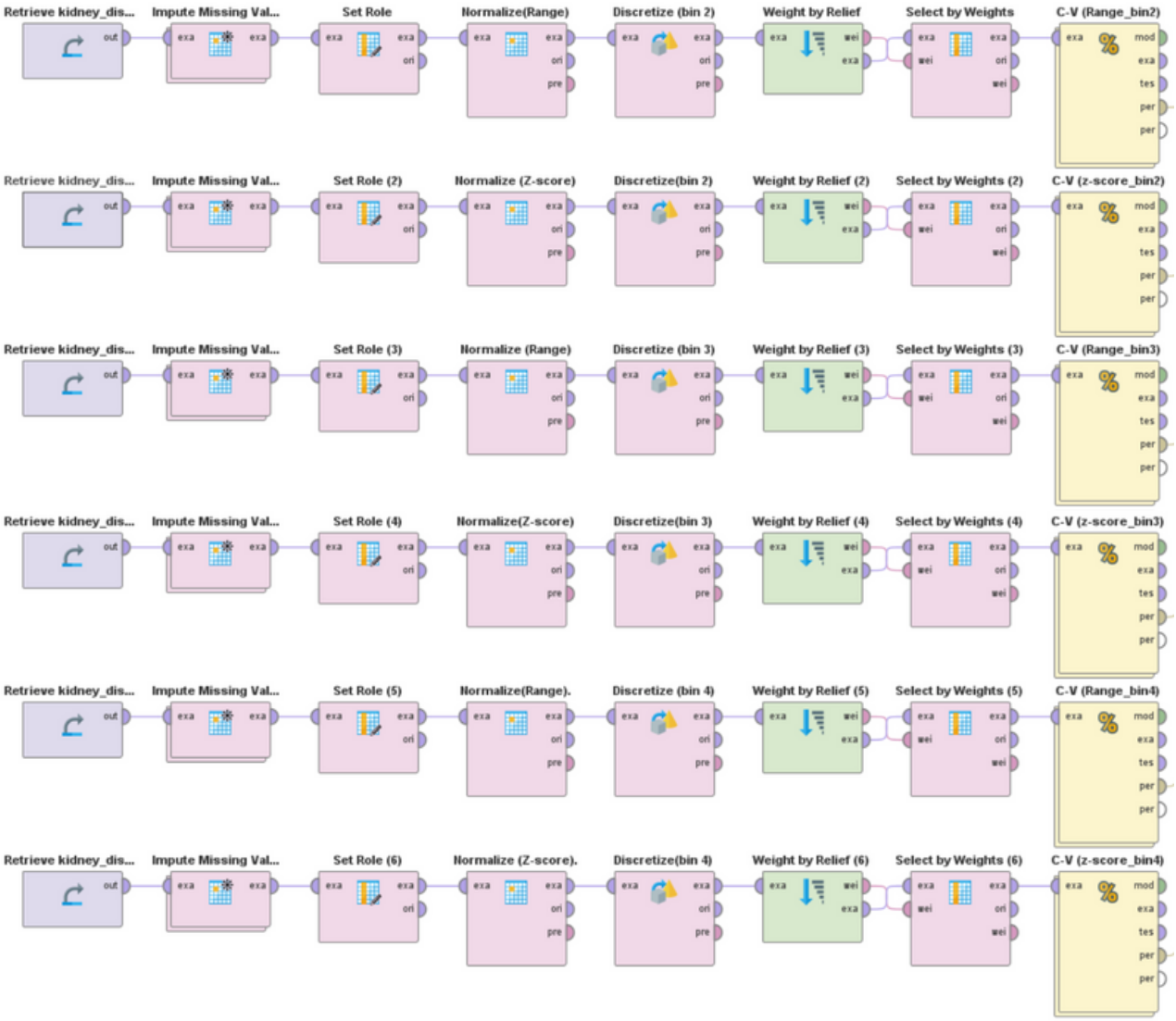
As we have seen in the result, on **Normalized Range(2Bin)** gives the best result, because it gives high accuracy.

accuracy: 97.95% +/- 2.36% (micro average: 97.95%)

	true ckd	true notckd	class precision
pred. ckd	232	0	100.00%
pred. notckd	8	150	94.94%
class recall	96.67%	100.00%	

Measure the accuracy

F. using all of the pre-processing steps including dimensionality reduction



- The information gained from all preprocessing steps **including** dimension reduction

normalize	bin	Accuracy	Classification error	Class recall CKD	Class recall Not CKD	correlation
Range	2	96.15%	3.85%	93.75%	100.00%	0.923
	3	**96.67%**	**3.33%**	**97.08%**	**96.00%**	**0.930**
	4	95.38%	4.62%	95.42%	95.33%	0.903
Z-score	2	96.15%	3.85%	93.75%	100.00%	0.923
	3	96.15%	3.85%	96.25%	96.00%	0.919
	4	96.41%	3.59%	97.50%	94.67%	0.924

As we have seen in the result, the **Normalized Range(3Bin)** gives the best result, because it gives high accuracy.

accuracy: 96.67% +/- 3.43% (micro average: 96.67%)

	true ckd	true notckd	class precision
pred. ckd	233	6	97.49%
pred. notckd	7	144	95.36%
class recall	97.08%	96.00%	

Measure the accuracy

2. Find a combination of pre-processing steps which gives the best results.

E. normalizing the data for dimension reduction (bins 2)

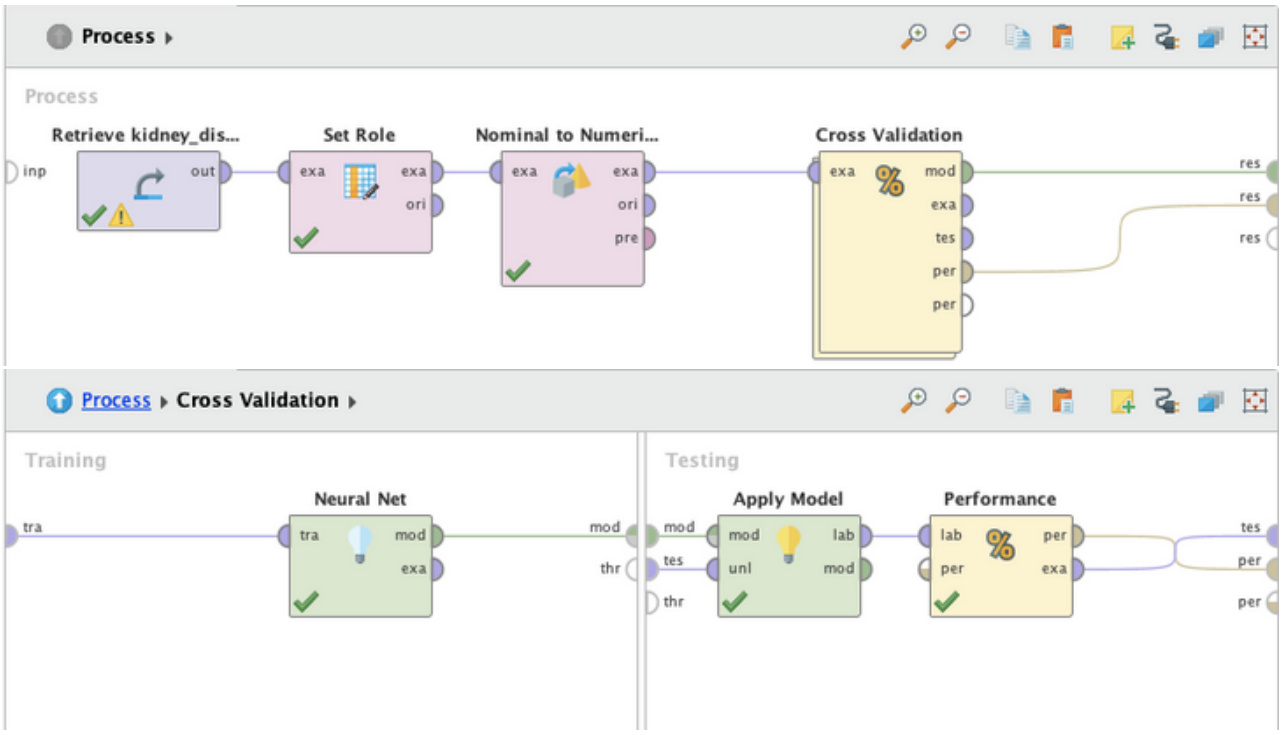
accuracy: 97.95% +/- 2.36% (micro average: 97.95%)

	true ckd	true notckd	class precision
pred. ckd	232	0	100.00%
pred. notckd	8	150	94.94%
class recall	96.67%	100.00%	

Finally, we decided to choose Normalization range (2 bins) values in pre-processing steps except for dimension reduction because it has the highest accuracy among all others to complete and the highest correlations you are prioritizing and demonstrate how they align with your own strategy and goals.

Measure the accuracy

3. Using a Neural Network classifier and suitable data pre-processing steps.



accuracy: 98.97% +/- 1.32% (micro average: 98.97%)

	true ckd	true notckd	class precision
pred. ckd	236	0	100.00%
pred. notckd	4	150	97.40%
class recall	98.33%	100.00%	

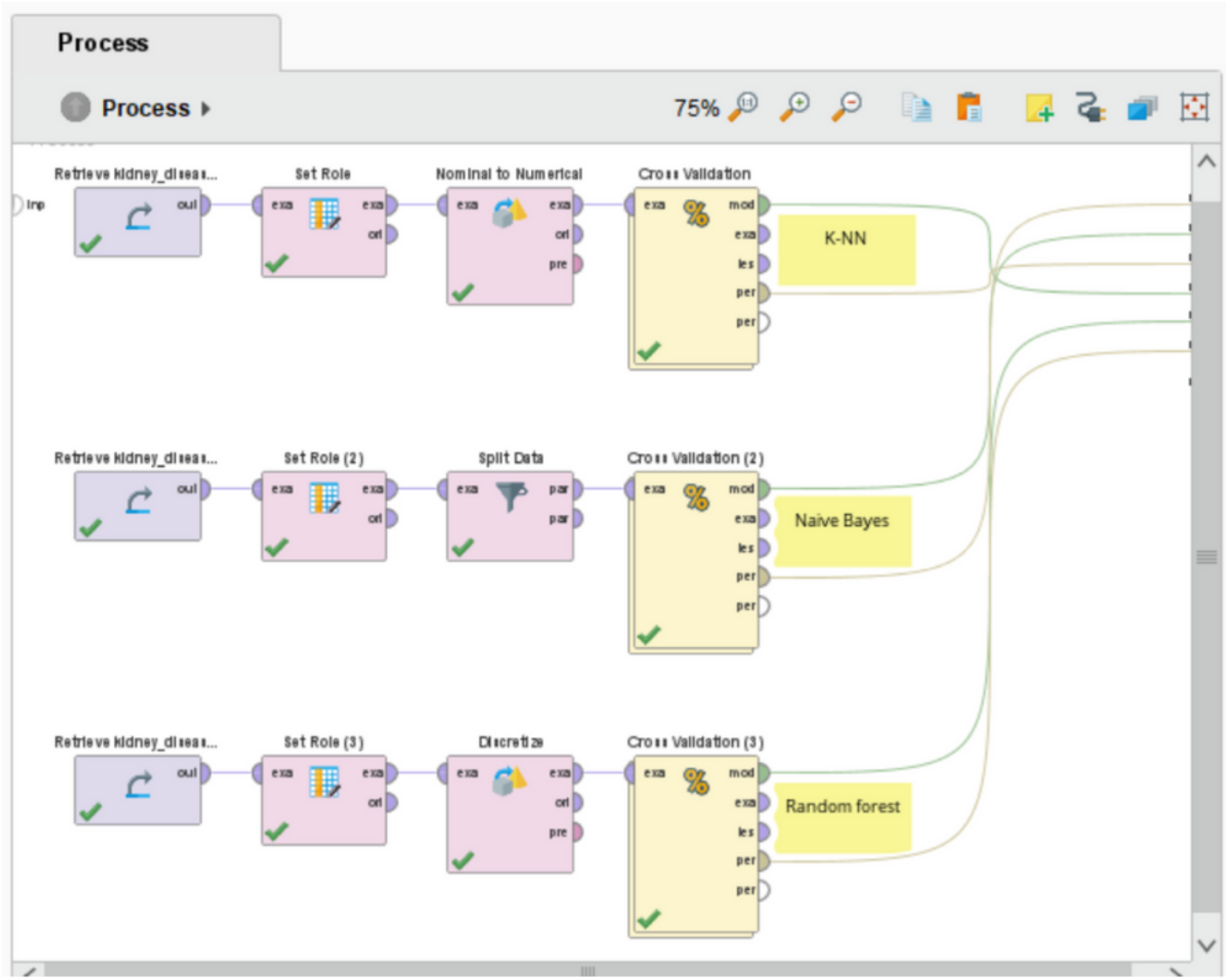
The Accuracy by using Neural Network classifier and pre-processing steps is :

Accuracy is:98.97%
Class recall CKD: 98.33%
Class recall NotCKD :100.00%

The accuracy was very high and good, the error rate was very low, and the correlation was strong at 0.976.

Measure the accuracy

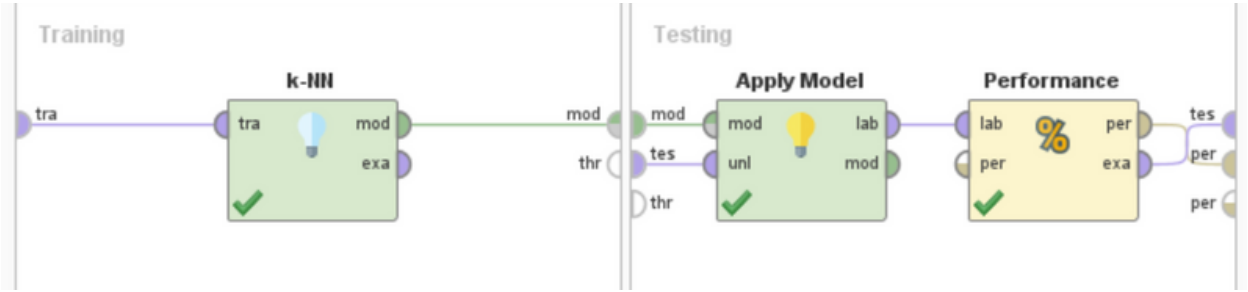
4. Using 3 classifiers not used in previous tasks using suitable data pre-processing steps.



As a different classifier from the previous methods K-NN, Naive Bayes, and Random forest are good.

Measure the accuracy

K-NN Algorithm



accuracy: 75.64% +/- 8.82% (micro average: 75.64%)

	true ckd	true notckd	class precision
pred. ckd	165	20	89.19%
pred. notckd	75	130	63.41%
class recall	68.75%	86.67%	

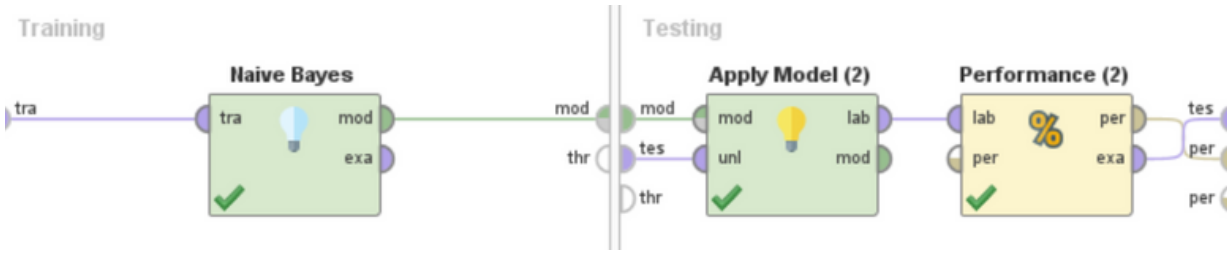
We can see in the confusion matrix

Accuracy is: 75.64%
Classification error: 24.36%
Class recall CKD: 68.75%
Class recall NotCKD: 86.67%

The accuracy here was good, the error rate was acceptable, and the correlation is 0.540. So the correlation is good.

Measure the accuracy

Native Bayes



accuracy: 95.26% +/- 3.84% (micro average: 95.24%)

	true ckd	true notckd	class precision
pred. ckd	155	0	100.00%
pred. notckd	13	105	88.98%
class recall	92.26%	100.00%	

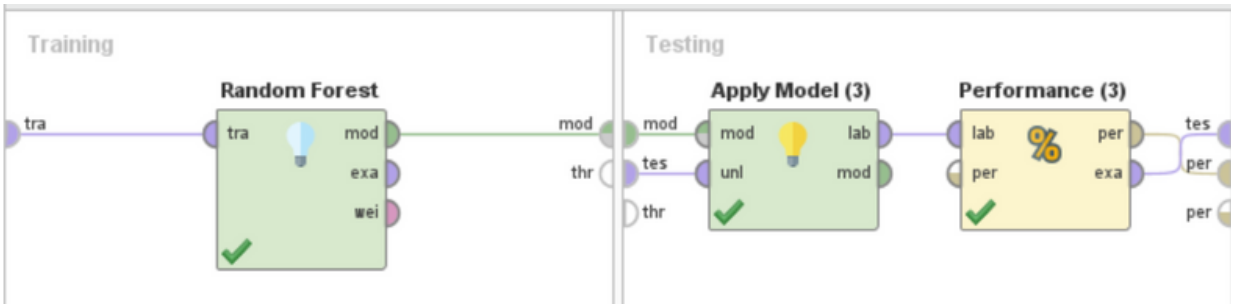
We can see in the confusion matrix

Accuracy is: 95.24%
Classification error: 4.74%
Class recall CKD: 92.26%
Class recall NotCKD: 100%

The accuracy here was High, there is no error rate in the prediction of NotCKD people. and the correlation is 0.910. So the correlation is strong.

Measure the accuracy

Random forest



accuracy: 98.72% +/- 1.35% (micro average: 98.72%)

	true ckd	true notckd	class precision
pred. ckd	238	3	98.76%
pred. notckd	2	147	98.66%
class recall	99.17%	98.00%	

We can see in the confusion matrix

Accuracy is: 98.72%
Classification error: 1.28%
Class recall CKD: 99.17%
Class recall NotCKD: 98%

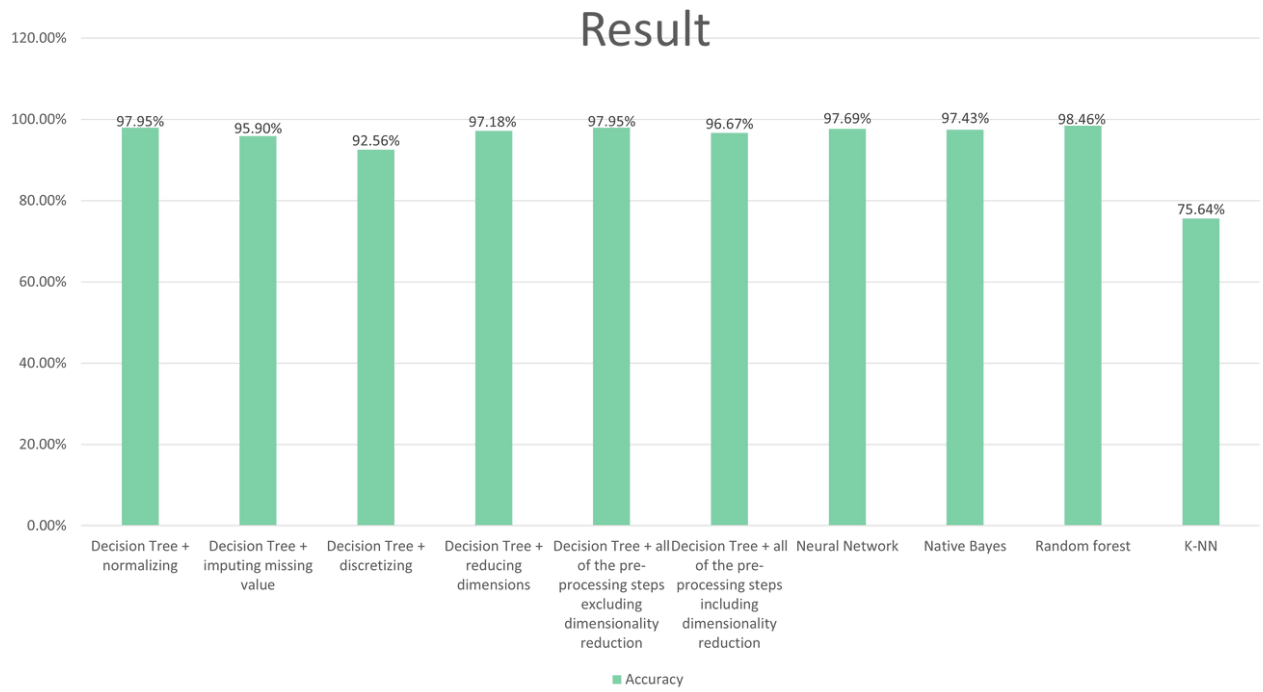
The accuracy here was High, The error rate was good enough. and the correlation is 0.974 So, the correlation is good.

Result

End your report with a review of the highlights, and a renewed commitment to continue working on making the SDGs attainable by 2030.

Experimen	Accuracy
Decision Tree + normalizing	96.67%
Decision Tree + imputing missing value	95.90%
Decision Tree + discretizing	92.56%
Decision Tree + reducing dimensions	97.18%
Decision Tree + all of the pre-processing steps excluding dimensionality reduction	97.95%
Decision Tree + all of the pre-processing steps including dimensionality reduction	96.67%
Neural Network	98.97%
Native Bayes	95.26%
Random Forest	98.72%
K-NN	75.64%

Result



We have tried so many techniques and methods to get the best accuracy such as the Decision Table-Naive Bayes, Neural Network, Random Forest, and K-NN. The best algorithm that gives us the best accuracy is **Neural Network** which gives an accuracy is **98.97%**