

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler

import joblib
import pickle

%matplotlib inline
```

#### Loading the Dataset

```
In [2]: data=pd.read_csv("/content/chronickidneydisease.csv")
```

```
In [3]: data.head()
```

```
Out[3]:
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd

```
In [4]: data.tail()
```

```
Out[4]:
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

5 rows × 26 columns

```
In [5]: data.head(10)
```

```
Out[5]:
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
5	5	60.0	90.0	1.015	3.0	0.0	NaN	NaN	notpresent	notpresent	...	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	6	68.0	70.0	1.010	0.0	0.0	NaN	normal	notpresent	notpresent	...	36	NaN	NaN	no	no	no	good	no	no	ckd
7	7	24.0	NaN	1.015	2.0	4.0	normal	abnormal	notpresent	notpresent	...	44	6900	5	no	yes	no	good	yes	no	ckd
8	8	72.0	100.0	1.015	2.0	0.0	normal	normal	notpresent	notpresent	...	33	6500	4.0	no	no	no	good	no	no	ckd
9	9	53.0	90.0	1.020	2.0	0.0	abnormal	abnormal	present	notpresent	...	29	12100	3.7	yes	yes	no	poor	no	yes	ckd

10 rows × 26 columns

#### Drop id Column

```
In [6]: data.drop(["id"],axis=1,inplace=True)
data.columns
```

```
Out[6]: Index(['age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgn', 'bu',
'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad',
'appet', 'pe', 'ane', 'classification'],
dtype='object')
```

#### Renaming the columns

```
In [7]: data.columns=['age','blood_pressure','specific_gravity','albumin','sugar','red_blood_cells','pus_cell','pus_cell_clumps','bacteria','blood glucose ran
data.columns
```

```
Out[7]: Index(['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',
'diabetesmellitus', 'coronary_artery_disease', 'appetite',
'pedal_edema', 'anemia', 'class'],
dtype='object')
```

```
In [8]: data.info()
```

```
RangeIndex: 400 entries, 0 to 399
Data columns (total 25 columns):
#   Column                                Non-Null Count  Dtype
#   ...
#   Column                                Non-Null Count  Dtype
```

```
0 age 391 non-null float64
1 blood_pressure 388 non-null float64
2 specific_gravity 353 non-null float64
3 albumin 354 non-null float64
4 sugar 351 non-null float64
5 red_blood_cells 248 non-null object
6 pus_cell 335 non-null object
7 pus_cell_clumps 396 non-null object
8 bacteria 396 non-null object
9 blood_glucose_random 356 non-null float64
10 blood_urea 381 non-null float64
11 serum_creatinine 383 non-null float64
12 sodium 313 non-null float64
13 potassium 312 non-null float64
14 hemoglobin 348 non-null float64
15 packed_cell_volume 330 non-null object
16 white_blood_cell_count 295 non-null object
17 red_blood_cell_count 270 non-null object
18 hypertension 398 non-null object
19 diabetesmellitus 398 non-null object
20 coronary_artery_disease 398 non-null object
21 appetite 399 non-null object
22 pedal_edema 399 non-null object
23 anemia 399 non-null object
24 class 400 non-null object
dtypes: float64(11), object(14)
memory usage: 78.2+ KB
```

#### Target Column

```
In [9]: data['class'].unique()
```

```
Out[9]: array(['ckd', 'ckd\t', 'notckd'], dtype=object)
```

```
In [10]: data['class']=data['class'].replace("ckd\t","ckd")
```

```
In [11]: data['class'].unique()
```

```
Out[11]: array(['ckd', 'notckd'], dtype=object)
```

```
In [12]: catcols=set(data.dtypes[data.dtypes!='O'].index.values)
print(catcols)

{'packed_cell_volume', 'red_blood_cells', 'white_blood_cell_count', 'red_blood_cell_count', 'diabetesmellitus', 'anemia', 'bacteria', 'coronary_artery_disease', 'pedal_edema', 'appetite', 'pus_cell', 'pus_cell_clumps', 'class', 'hypertension'}
```

```
In [14]: from collections import Counter as c
```

```
In [15]: for i in catcols:
print("Columns :",i)
print(c(data[i]))
print('*'*120+'\n')
```

```
Columns : packed_cell_volume
Counter({nan: 70, '52': 21, '41': 21, '44': 19, '48': 19, '40': 16, '43': 14, '45': 13, '42': 13, '32': 12, '36': 12, '33': 12, '28': 12, '50': 12, '37': 11, '34': 11, '35': 9, '29': 9, '30': 9, '46': 9, '31': 8, '39': 7, '24': 7, '26': 6, '38': 5, '47': 4, '49': 4, '53': 4, '51': 4, '54': 4, '27': 3, '22': 3, '25': 3, '23': 2, '19': 2, '16': 1, '\t?': 1, '14': 1, '18': 1, '17': 1, '15': 1, '21': 1, '20': 1, '\t43': 1, '9': 1})
*****
```

```
Columns : red_blood_cells
Counter({'normal': 201, nan: 152, 'abnormal': 47})
*****
```

```
Columns : white_blood_cell_count
Counter({nan: 105, '9800': 11, '6700': 10, '9600': 9, '9200': 9, '7200': 9, '6900': 8, '11000': 8, '5800': 8, '7800': 7, '9100': 7, '9400': 7, '7000': 7, '4300': 6, '6300': 6, '10700': 6, '10500': 6, '7500': 5, '8300': 5, '7900': 5, '8600': 5, '5600': 5, '10200': 5, '5000': 5, '8100': 5, '9500': 5, '6000': 4, '6200': 4, '10300': 4, '7700': 4, '5500': 4, '10400': 4, '6800': 4, '6500': 4, '4700': 4, '7300': 3, '4500': 3, '8400': 3, '6400': 3, '4200': 3, '7400': 3, '8000': 3, '5400': 3, '3800': 2, '11400': 2, '5300': 2, '8500': 2, '14600': 2, '7100': 2, '13200': 2, '9000': 2, '8200': 2, '15200': 2, '12400': 2, '11800': 2, '8800': 2, '9300': 2, '6600': 2, '12100': 1, '12200': 1, '18900': 1, '21600': 1, '11300': 1, '\t6200': 1, '11800': 1, '12500': 1, '11900': 1, '12700': 1, '13600': 1, '14900': 1, '16300': 1, '\t8400': 1, '10900': 1, '2200': 1, '11200': 1, '19100': 1, '\t?': 1, '12300': 1, '16700': 1, '2600': 1, '26400': 1, '4900': 1, '12000': 1, '15700': 1, '4100': 1, '11500': 1, '10800': 1, '9900': 1, '5200': 1, '5900': 1, '9700': 1, '5100': 1})
*****
```

```
Columns : red_blood_cell_count
Counter({nan: 130, '5.2': 18, '4.5': 16, '4.9': 14, '4.7': 11, '3.9': 10, '4.8': 10, '4.6': 9, '3.4': 9, '3.7': 8, '5.0': 8, '6.1': 8, '5.5': 8, '5.9': 8, '3.8': 7, '5.4': 7, '5.8': 7, '5.3': 7, '4.3': 6, '4.2': 6, '5.6': 6, '4.4': 5, '3.2': 5, '4.1': 5, '6.2': 5, '5.1': 5, '6.4': 5, '5.7': 5, '6.5': 5, '3.6': 4, '6.0': 4, '6.3': 4, '4.0': 3, '4': 3, '3.5': 3, '3.3': 3, '5': 2, '2.6': 2, '2.8': 2, '2.5': 2, '3.1': 2, '2.1': 2, '2.9': 2, '2.7': 2, '3.0': 2, '2.3': 1, '8.0': 1, '3': 1, '2.4': 1, '\t?': 1})
*****
```

```
Columns : diabetesmellitus
Counter({'no': 258, 'yes': 134, '\tno': 3, '\tyes': 2, nan: 2, ' yes': 1})
*****
```

```
Columns : anemia
Counter({'no': 339, 'yes': 60, nan: 1})
*****
```

```
Columns : bacteria
Counter({'notpresent': 374, 'present': 22, nan: 4})
*****
```

```
Columns : coronary_artery_disease
Counter({'no': 362, 'yes': 34, '\tno': 2, nan: 2})
*****
```

```
Columns : pedal_edema
Counter({'no': 323, 'yes': 76, nan: 1})
*****
```

```
Columns : appetite
Counter({'good': 317, 'poor': 82, nan: 1})
*****
```

```
Columns : pus_cell
Counter({'normal': 259, 'abnormal': 76, nan: 65})
*****
```

### Removing the column which are not categorized

### Numerical Columns

```
for i in contcols:
    print('Continuous Columns :',i)
    print(c(data[i]))
    print('*'*120+'\n')
```

[illegible]

```

Countdown Columns : serum_creatinine
Counter({1: 40, 11: 24, 18: 23, 0.5: 23, 0.7: 22, 0.9: 22, 0.6: 18, 0.8: 17, 2.2: 10, 1.5: 9, 1.7: 9, 1.3: 8, 1.6: 8, 1.8: 7, 1.4: 7, 2.5: 7, 2.8: 7, 1.9: 6, 2.7: 5, 2.1: 5, 2.0: 5, 3.2: 5, 3.3: 5, 3.9: 4, 7.3: 4, 4.0: 3, 2.4: 3, 3.4: 3, 2.9: 3, 5.3: 3, 2.3: 3, 7.2: 2, 4.6: 2, 4.1: 2, 5.2: 2, 6.3: 2, 3.0: 2, 6.1: 2, 6.7: 2, 5.6: 2, 6.5: 2, 4.4: 2, 6.0: 2, 3.8: 1, 24.0: 1, 9.6: 1, 76.0: 1, 7.7: 1, nan: 1, 10.8: 1, 5.9: 1, 3.25: 1, nan: 1, 9.7: 1, 6.4: 1, 32.0: 1, nan: 1, nan: 1, 8.5: 1, 15.0: 1, 3.6: 1, 10.2: 1, 11.5: 1, nan: 1, 12.2: 1, 9.2: 1, 13.8: 1, 16.9: 1, 7.1: 1, 18.0: 1, 13.0: 1, 4.8: 1, 14.2: 1, 16.4: 1, nan: 1, nan: 1, 2.6: 1, 7.5: 1, 4.3: 1, 18.1: 1, 11.8: 1, 9.3: 1, 6.8: 1, 13.5: 1, nan: 1, 12.8: 1, 11.9: 1, nan: 1, nan: 1, nan: 1, 12.0: 1, nan: 1, 13.4: 1, 15.2: 1, 13.3: 1, nan: 1, nan: 1, nan: 1, nan: 1, 0.4: 1})

```

[illegible]

Continuous Columns : sodium

```

Continuous Columns : sodium
Counter({135.0: 40, 140.0: 25, 141.0: 22, 139.0: 21, 142.0: 20, 138.0: 20, 137.0: 19, 136.0: 17, 150.0: 17, 147.0: 13, 145.0: 11, 132.0: 10, 146.0: 1
0, 131.0: 9, 144.0: 9, 133.0: 8, 130.0: 7, 134.0: 6, 143.0: 4, 127.0: 3, 124.0: 3, 114.0: 2, 125.0: 2, 128.0: 2, 122.0: 2, 113.0: 2, 120.0: 2, nan: 1,
nan: 1, nan: 1, 111.0: 1, nan: 1, 104.0: 1, nan: 1, nan: 1, nan: 1, nan: 1, 4.5: 1, nan: 1, 129.0: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1,
nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, 163.0: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, n
an: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, 126.0: 1, nan: 1, nan: 1, nan: 1, nan: 1, n
an: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, na
n: 1})
=====

Continuous Columns : specific_gravity
Counter({1.02: 106, 1.01: 84, 1.025: 81, 1.015: 75, 1.005: 7, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1,
nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1,
1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1})
=====

Continuous Columns : blood_pressure
Counter({80.0: 116, 70.0: 112, 60.0: 71, 90.0: 53, 100.0: 25, 50.0: 5, 110.0: 3, nan: 1, nan: 1, 140.0: 1, 180.0: 1, nan: 1, nan: 1, nan: 1, nan: 1, n
an: 1, nan: 1, nan: 1, 120.0: 1, nan: 1, nan: 1, nan: 1})
=====

Continuous Columns : albumin
Counter({0.0: 199, 1.0: 44, 2.0: 43, 3.0: 43, 4.0: 24, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1,
nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1,
1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1, nan: 1})
=====

Continuous Columns : age
Counter({60.0: 19, 65.0: 17, 48.0: 12, 50.0: 12, 55.0: 12, 47.0: 11, 62.0: 10, 45.0: 10, 54.0: 10, 59.0: 10, 56.0: 10, 61.0: 9, 70.0: 9, 46.0: 9, 34.
0: 9, 68.0: 8, 73.0: 8, 64.0: 8, 71.0: 8, 57.0: 8, 63.0: 7, 72.0: 7, 67.0: 7, 30.0: 7, 42.0: 6, 69.0: 6, 35.0: 6, 44.0: 6, 43.0: 6, 33.0: 6, 51.0: 5,
52.0: 5, 53.0: 5, 75.0: 5, 76.0: 5, 58.0: 5, 41.0: 5, 66.0: 5, 24.0: 4, 40.0: 4, 39.0: 4, 80.0: 4, 23.0: 4, 74.0: 3, 38.0: 3, 17.0: 3, 8.0: 3, 32.0:
3, 37.0: 3, 25.0: 3, 29.0: 3, 21.0: 2, 15.0: 2, 5.0: 2, 12.0: 2, 49.0: 2, 19.0: 2, 36.0: 2, 20.0: 2, 28.0: 2, 7.0: 1, nan: 1, 82.0: 1, 11.0: 1, 26.0:
1, nan: 1, nan: 1, nan: 1, nan: 1, 81.0: 1, 14.0: 1, 27.0: 1, nan: 1, 83.0: 1, 4.0: 1, 3.0: 1, 6.0: 1, nan: 1, 90.0: 1, 78.0: 1, nan: 1, 2.0: 1, nan:
1, 22.0: 1, 79.0: 1})
=====

```

#### Removing the Columns which are not Numerical

```

In [20]: contcols.remove('specific_gravity')
         contcols.remove('albumin')
         contcols.remove('sugar')
         print(contcols)

{'blood_glucose_random', 'potassium', 'serum_creatinine', 'blood_urea', 'hemoglobin', 'sodium', 'blood_pressure', 'age'}

```

#### Adding columns which we found Continuous

```

In [21]: contcols.add('red_blood_cell_count')
         contcols.add('packed_cell_volume')
         contcols.add('white_blood_cell_count')
         print(contcols)

{'packed_cell_volume', 'blood_glucose_random', 'white_blood_cell_count', 'red_blood_cell_count', 'potassium', 'serum_creatinine', 'blood_urea', 'hemog
lobin', 'sodium', 'blood_pressure', 'age'}

```

#### Adding columns which we found Categorical

```

In [22]: catcols.add('specific_gravity')
         catcols.add('albumin')
         catcols.add('sugar')
         print(catcols)

{'red_blood_cells', 'sugar', 'diabetesmellitus', 'anemia', 'bacteria', 'coronary_artery_disease', 'pedal_edema', 'appetite', 'specific_gravity', 'pus_
cell', 'albumin', 'pus_cell_clumps', 'class', 'hypertension'}

```

#### Rectifying the Categorical Columns Classes

```

In [23]: data['coronary_artery_disease']=data.coronary_artery_disease.replace('\tno','no')
         c(data['coronary_artery_disease'])

```

#### Rectifying the Categorical Columns Classes

```

In [23]: data['coronary_artery_disease']=data.coronary_artery_disease.replace('\tno','no')
         c(data['coronary_artery_disease'])

```

```

Out[23]: Counter({ 'no': 364, 'yes': 34, nan: 2})

```

```

In [24]: data['diabetesmellitus']=data.diabetesmellitus.replace(to_replace=('\tno':'no','\tyes':'yes','yes':'yes'))
         c(data['diabetesmellitus'])

```

```

Out[24]: Counter({'yes': 136, 'no': 261, 'yes': 1, nan: 2})

```

#### Null Values

```

In [25]: data.isnull().any()

```

```

Out[25]: age                True
         blood_pressure      True
         specific_gravity    True
         albumin             True
         sugar               True
         red_blood_cells     True
         pus_cell            True
         pus_cell_clumps     True
         bacteria            True
         blood_glucose_random True
         blood_urea          True
         serum_creatinine    True
         sodium              True
         potassium           True
         hemoglobin          True
         packed_cell_volume  True

```

```

packed_cell_volume      True
white_blood_cell_count  True
red_blood_cell_count    True
hypertension            True
diabetesmellitus        True
coronary_artery_disease True
appetite                True
pedal_edema             True
anemia                  True
class                   False
dtype: bool

```

```
In [26]: data.isnull().sum()#return the count
```

```

Out[26]: age                9
blood_pressure            12
specific_gravity          47
albumin                   46
sugar                     49
red_blood_cells           152
pus_cell                  65
pus_cell_clumps           4
bacteria                  4
blood_glucose_random      44
blood_urea                 19
serum_creatinine          17
sodium                    87
potassium                 88
hemoglobin                52
packed_cell_volume        70
white_blood_cell_count    105
red_blood_cell_count      130
hypertension              2
diabetesmellitus          2
coronary_artery_disease   2
appetite                  1
pedal_edema               1

class                    False
dtype: bool

```

```
In [26]: data.isnull().sum()#return the count
```

```

Out[26]: age                9
blood_pressure            12
specific_gravity          47
albumin                   46
sugar                     49
red_blood_cells           152
pus_cell                  65
pus_cell_clumps           4
bacteria                  4
blood_glucose_random      44
blood_urea                 19
serum_creatinine          17
sodium                    87
potassium                 88
hemoglobin                52
packed_cell_volume        70
white_blood_cell_count    105
red_blood_cell_count      130
hypertension              2
diabetesmellitus          2
coronary_artery_disease   2
appetite                  1
pedal_edema               1
anemia                    1
class                     0
dtype: int64

```

```
In [27]: data.packed_cell_volume=pd.to_numeric(data.packed_cell_volume,errors='coerce')
data.white_blood_cell_count=pd.to_numeric(data.white_blood_cell_count,errors='coerce')
data.red_blood_cell_count=pd.to_numeric(data.red_blood_cell_count,errors='coerce')
```