Review on Chronic Kidney Disease Dataset

RAJAT SRIVASTAVA
CSE GRADUATE VIT VELLORE
(FOR INTERNSHIP REVIEW)

About the Dataset

- **CKD (Chronic Kidney Dataset)** is a healthcare(medical) dataset of **400 patients** available under the repository of **UCI (University of California and Irvine)**
- **Link -** https://uci.edu/

- **AI in Healthcare** is growing branch with its application of treating patients with drug based monitoring of patients reaction to a specific treatment , to even operation using Robots .

- CKD has 25 features(11 numeric ,14 nominal) and 1 output feature (**total 26 features**).
- These include patient information of **physical characteristics , blood test based entity values and any other disease onset**.

- Modified version of this dataset is available at Kaggle with **14 features**.

Let’s Deep Dive into features of the dataset

age - age
bp - blood pressure
sg - specific gravity
al - albumin
su - sugar
rbc - red blood cells
pc - pus cell
pcc - pus cell clumps
ba - bacteria
bgr - blood glucose random
bu - blood urea
sc - serum creatinine
sod - sodium
pot - potassium
hemo - hemoglobin
pcv - packed cell volume
wc - white blood cell count
rc - red blood cell count
htn - hypertension
dm - diabetes mellitus
cad - coronary artery disease
appet - appetite
pe - pedal edema
ane - anemia
class - class

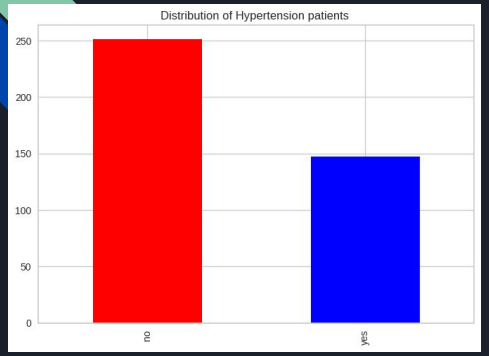
0	id	400	non-null	int64
1	age	391	non-null	float64
2	bp	388	non-null	float64
3	sg	353	non-null	float64
4	al	354	non-null	float64
5	su	351	non-null	float64
6	bc	248	non-null	object
7	pc	335	non-null	object
8	pcc	396	non-null	object
9	ba	396	non-null	object
10	bg	356	non-null	float64
11	bu	381	non-null	float64
12	sc	383	non-null	float64
13	sod	313	non-null	float64
14	pot	312	non-null	float64
15	hemo	348	non-null	float64
16	pcv	329	non-null	float64
17	wbcc	294	non-null	float64
18	bcc	269	non-null	float64
19	htn	398	non-null	object
20	dm	397	non-null	object
21	cad	398	non-null	object
22	appet	399	non-null	object
23	pe	399	non-null	object
24	ane	399	non-null	object
25	class	400	non-null	object

14 - Nominal (input is categorical value)

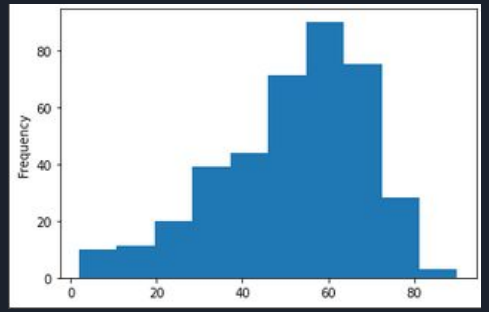
11 - Numeric (integer/float input)

1 - output class (affected or not affected by disease)

EDA of the Dataset



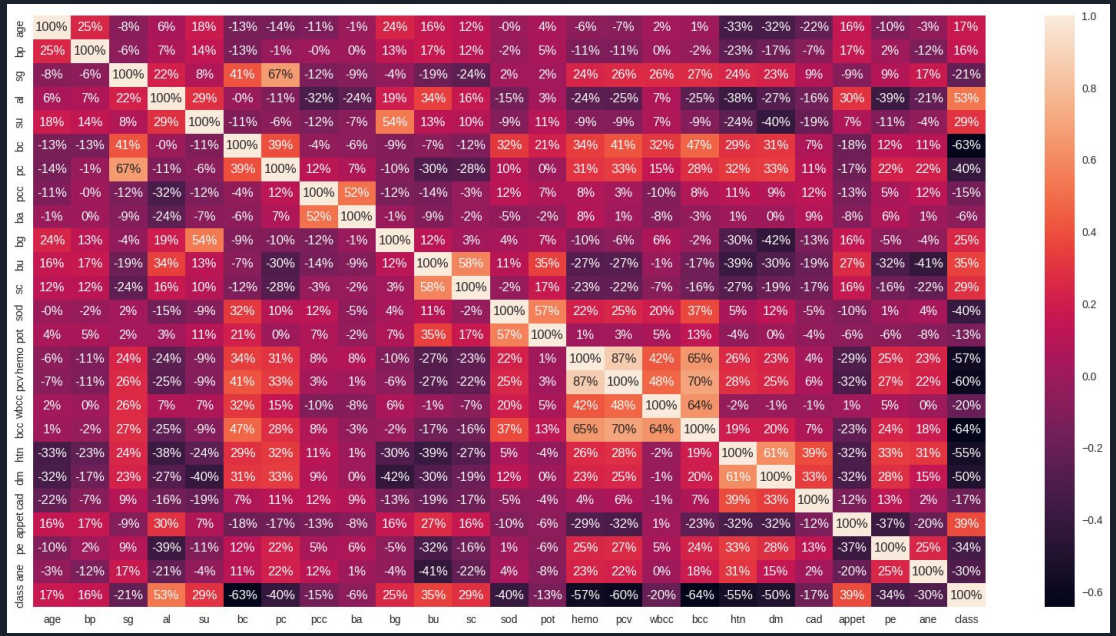
Hypertension patients



Age Distribution of patients

	id	age	bp	sg	al	su	bg	bu	sc	sod	pot	hemo	pcv	wbcc	bcc
count	400.000000	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000	329.000000	294.000000	269.000000
mean	200.500000	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437	38.884498	8406.122449	4.707435
std	115.614301	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587	8.990105	2944.474190	1.025323
min	1.000000	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000	9.000000	2200.000000	2.100000
25%	100.750000	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000	32.000000	6500.000000	3.900000
50%	200.500000	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000	40.000000	8000.000000	4.800000
75%	300.250000	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000	45.000000	9800.000000	5.400000
max	400.000000	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000	54.000000	26400.000000	8.000000

Numerical value features stats

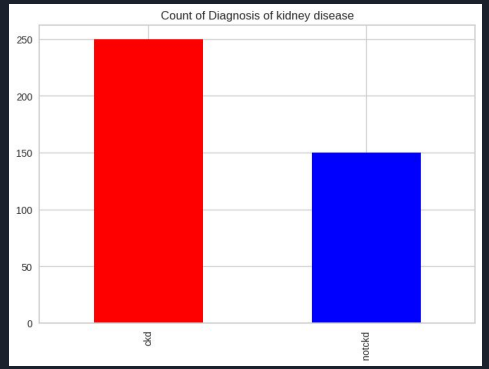


Correlation Matrix

What do we predict from the dataset ?

So the Output label -

- CKD (1) represents a person is affected by the kidney disease &
- NONCKD (0) meaning the person is healthy



Distribution of patients

Feature Engineering

- Categorical Features to Numerical (Label Encoding) - we used map
- Understand The Feature and it’s dependency on output feature as well as other features (data redundancy) - understand via pearson correlation matrix
- Remove or replace null values - replaced as new class variable in input features as (0)
- Drop un-required features

Machine Learning Algorithms to be used

- SVMs
- Decision Tree
- Bagging and Boosting
- Logistic Regression