Scaling

Sunday, April 2, 2023 10:36 PM

Data Preprocessing

: Make changes before giving to the Model

KM	M
5	5000
6	6000
8	8000

age	Salary
59	45000
21	100000

Range → Min, Max

20 – 60 (Min, Max) 2K – 2CR (Min, Max)

Scaling : Will Bring values in same Range

- Image
- PCA
- Normal Distribution
- Standarization (Mean-0, Std-1)

Min Max Nor

$$\frac{x - a_{min}}{a_{max} - a_{min}}$$

0 to 1

$$Z = \frac{x - \mu}{\sigma}$$ (Mean, Std dev)

KM	M
5	5000
6	6000
7	7000

$$\frac{7000 - 5000}{7000 - 5000} = \frac{2000}{2000} = 1$$

KM	M
0	0
0.5	0.5
1	1

D = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

0	0
0.5	0.5
1	1

7, 8, 9, 10

Mean → 5.5

$\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n}$

$= \frac{(1-5.5)^2 + (2-5.5)}{10}$

$\sigma^2 = 9.6$

Std dev $= \sqrt{9.6}$

Mean = 0

Std dev = 1

1	-2
2	-1
3	0
4	
5	
6	
7	
8	
9	-3.7

1 → $\frac{1-5.5}{\sqrt{9.6}}$

2 → $\frac{2-5.5}{\sqrt{9.6}}$

3

4

5

$\frac{10-5.5}{\sqrt{9.6}}$

what is difference in regression and logistics
Balanced or Imbalanced, train n test
In sigmoid curve

ary for linear regression.
>>> model = sm.OLS(y, x.assign(const=1)) #statsmodels
>>> results = model.fit()
>>> print(results.summary())
Can you please explain all the summarised parameters and their uses. Also, is there any same summary function in sklearn module??

2) Is Train and testing same as Cross Validation?

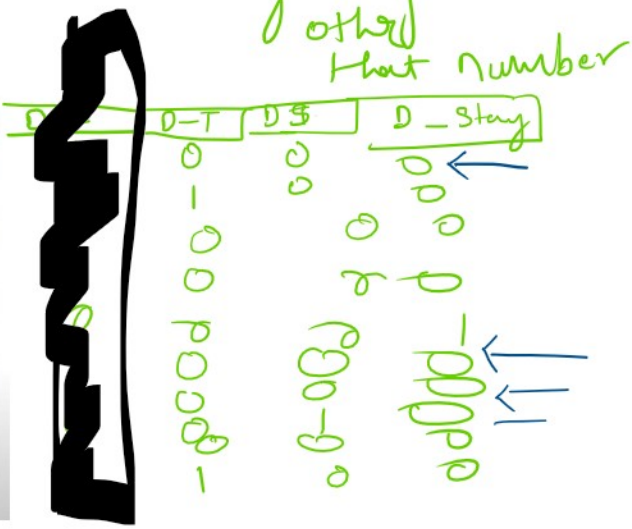
Encoding } Done

Encoding } Done

Categorical { String }

Computer cannot understand anything other that number

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

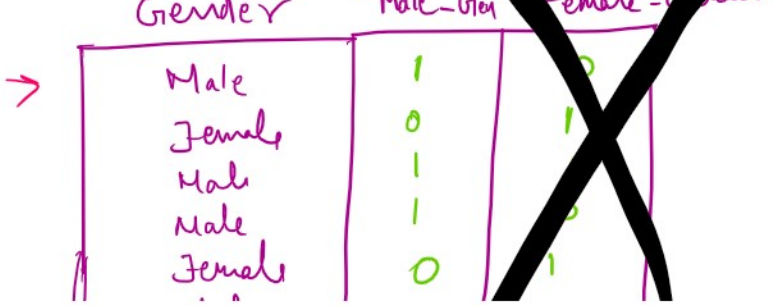


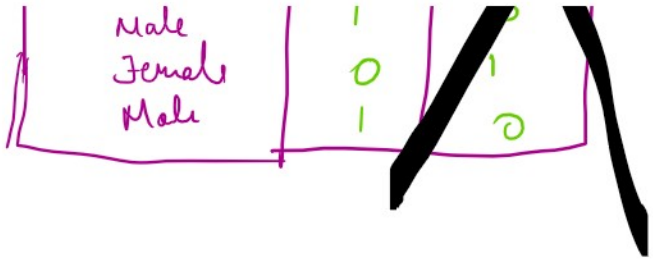
Categorical

Dummy encoding

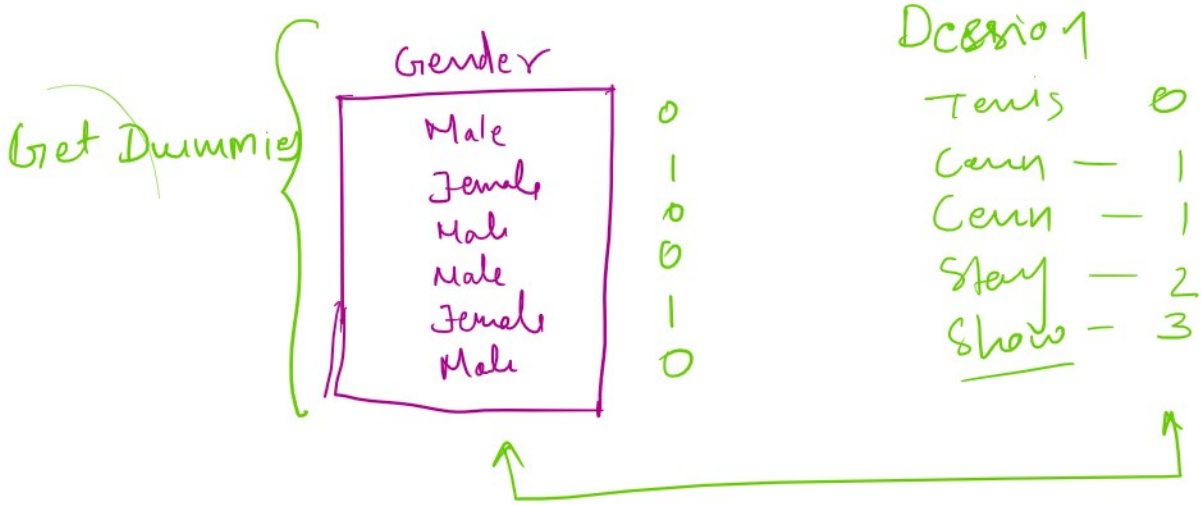
One hot encoding

Label encoding





Labelencoder



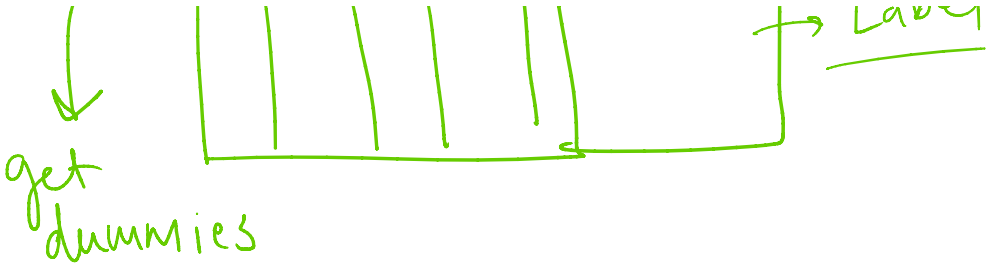
Male 0
Female 1

Female > Male
Tennis < Shopping

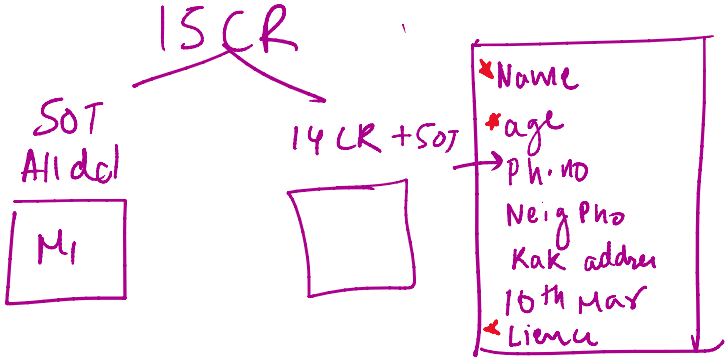
→ order

PHD > MTECH > BTECH

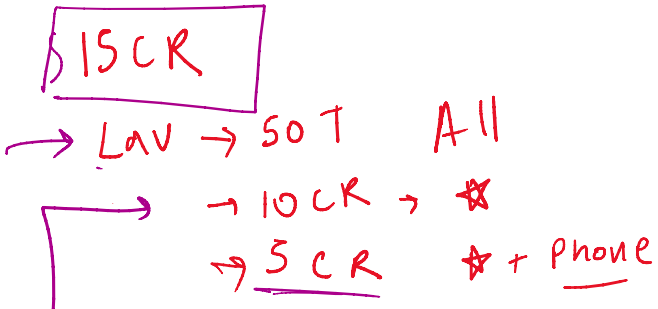
Target → Label

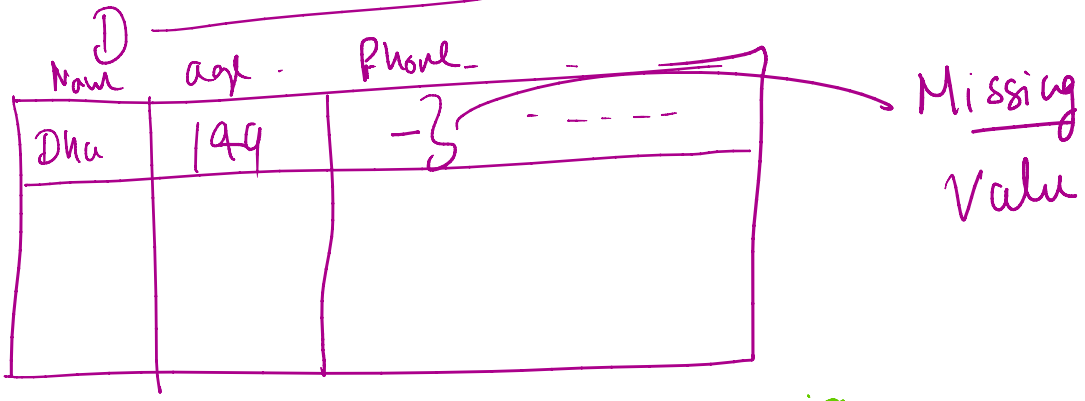


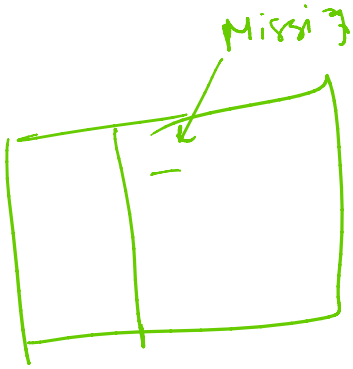
Missing Value



important values have to fill complsory and rest depends upon us if we want to fill



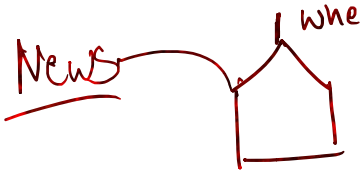




① Human Mistake

Rubbish } wrong
→ Noise

② Instrument error → Stopped



Day
week
} No data will collect

Models which cannot work with Missing Value

✡ Treatment

Row

S.No	Name	Occ	Age	Gender
1	Gaju	DS	27	M
2	Raju	SE	—	M
3	Maju	—	23	F
4	Kohar	—	—	—
5	ohar	Java	17	F
6	Raju	GS	21	M

① (Row 2)

Delete the Missing Value

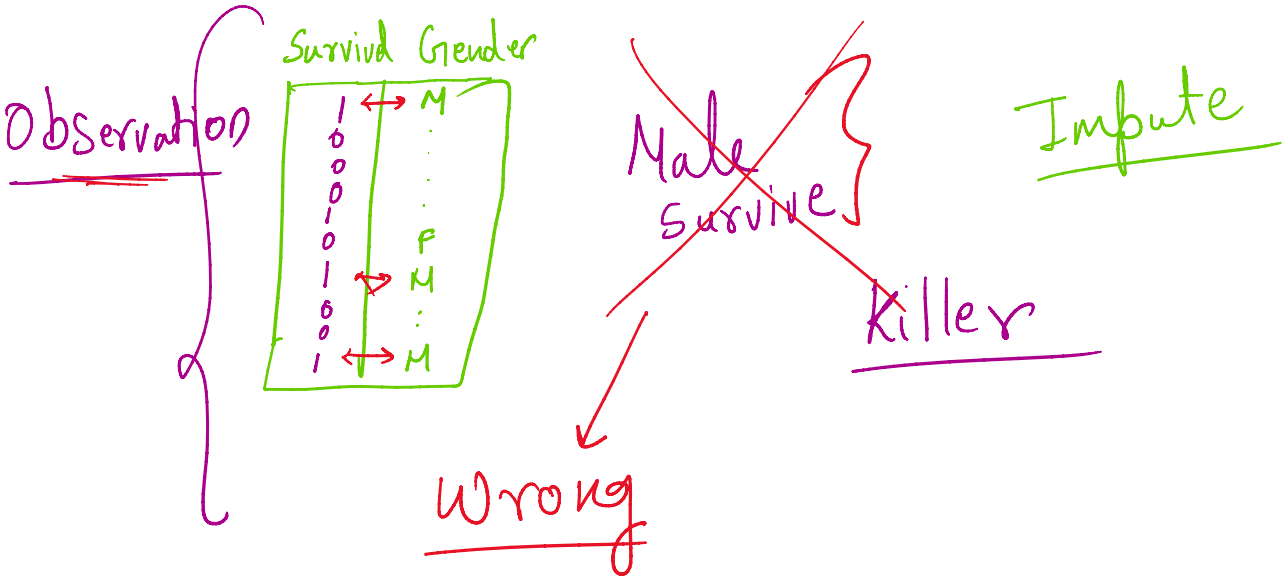
- Row
- Column

Disadvantage

- Data loss
- To be very very carefull before

1055 ✓ carefully before droping the row

: Row if has many values missing then drop the row



Drop Column 40%>

10 6%.

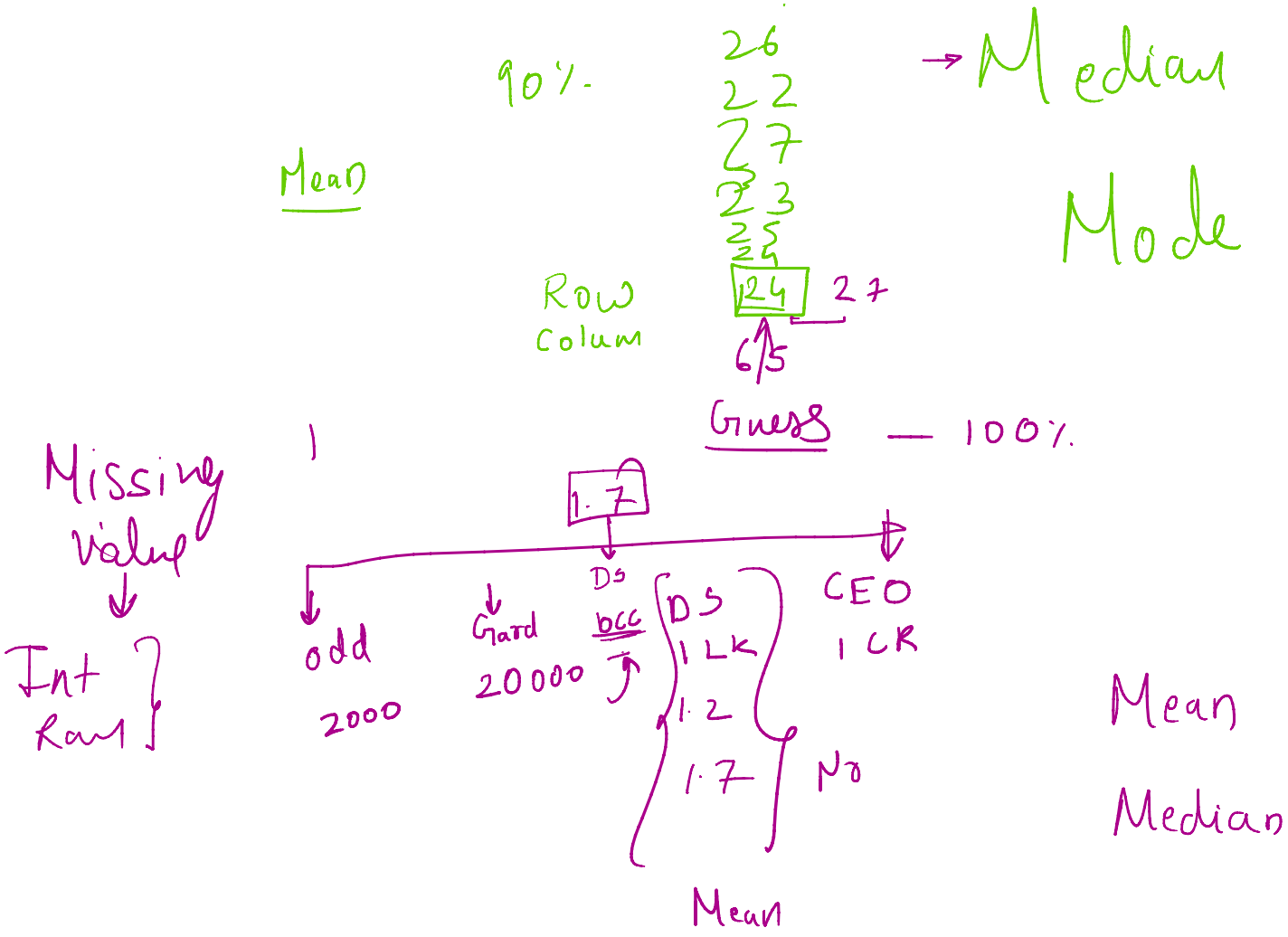
→ Impute

→

100 26
21
27
26
130 23
27
13 24
26
90%.

Mean

→ Median


90%.
26
22
27
23
25
24
27
Row
Colum
6/5
→ Median
Mode
Mean
Guess — 100%.
Missing Value
Int
Ram
1
1.7
DS
odd
2000
Gard
20000
bcc
DS
1 LK
1.2
1.7
CEO
1 CR
No
Mean
Mean
Median
Mean
Mean —
Median — outlier
GIGO
Mode

Category

Blu
white
green
org
Blac
Blan
White
White

Mode

White

9:45