☆ Feature Engineering (Handling missing value)

→ missing value occurs in dataset when some of the informations is not stored for a variables there are 3 mechnanism

① missing completyat Random (MCAR)

→ missing completly at Random is a type of missing data mechnaism in which the probability of value being missing is unrelated to both the observed data and the missing data. in other words, if the data is MCAR the missing values are randomly distrubuted throughout the dataset, and there is no systmatic reason for why they are missing.

for Example:- in a survey about the prevalance of certain diseases the missing data might be MCAR if they Survery participate with missing value for certain questions were selected randomly and their missing responses are not related to thir disease status or any other variables measured in the survey.

② missing At Random (MAR).

→ If one data is empty thn another data is also emply
→ age of women X → Salary of men —> X

③ missing data not a random (MNAR)

→ Because of one field It docs not find another data.
→ Job Statisfication —> income.
depended.

① MCAR —> no specific reason of missing value
→ SURVEy form

② MAR —> data is missing due to one data is depend on another data
→ Systmic Relationship → income of men → age of women

③ MdnaR —> missing is not Random and is dependent on unobserved or unmeasured factors that are associate with the missing value

Import Seaborn as sns

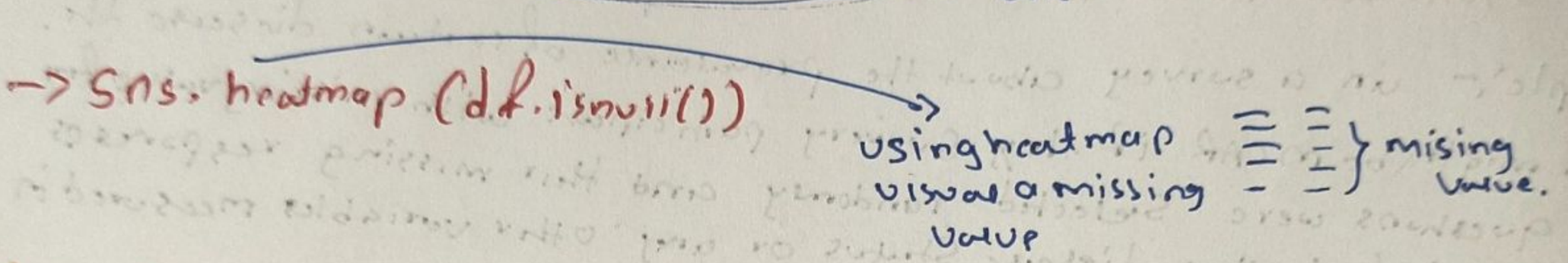
★ checking missing value in titanic dataset

-> df = sns.load_dataset ('titanic') → dataset

-> df.isnull().sum() → this fun is for find null value → Show All in One

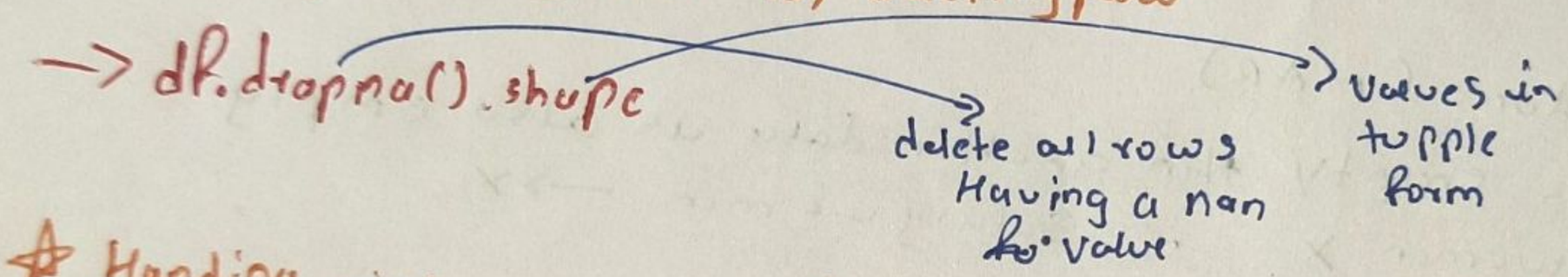
True -> missing value
False -> no missing value

-> sns.heatmap (df.isnull())



★ Handing missing value By deleting Row

-> df.dropna().shape



★ Handing missing value By deleting columns

-> df.dropna(axis=1)

-> sns.distplot(df['age']) → new column for mean.

-> df['Age_mean'] = df['age'].fillna(df['age'].mean())

-> df['Age_median'] = df['age'].fillna(df['age'].median())

-> mode = df[df['age'].notna()]['embarked'].mode()[0]

-> df['Embarked_mode'] = df['embarked'].fillna(mode)

Imbalanced dataset Handling →

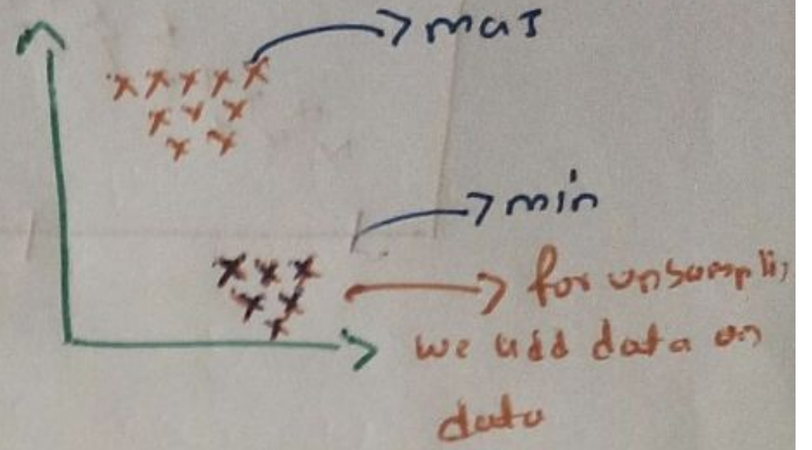
100, 900 → minority, → majority

Increasing → upsampling

decrease → downsampling

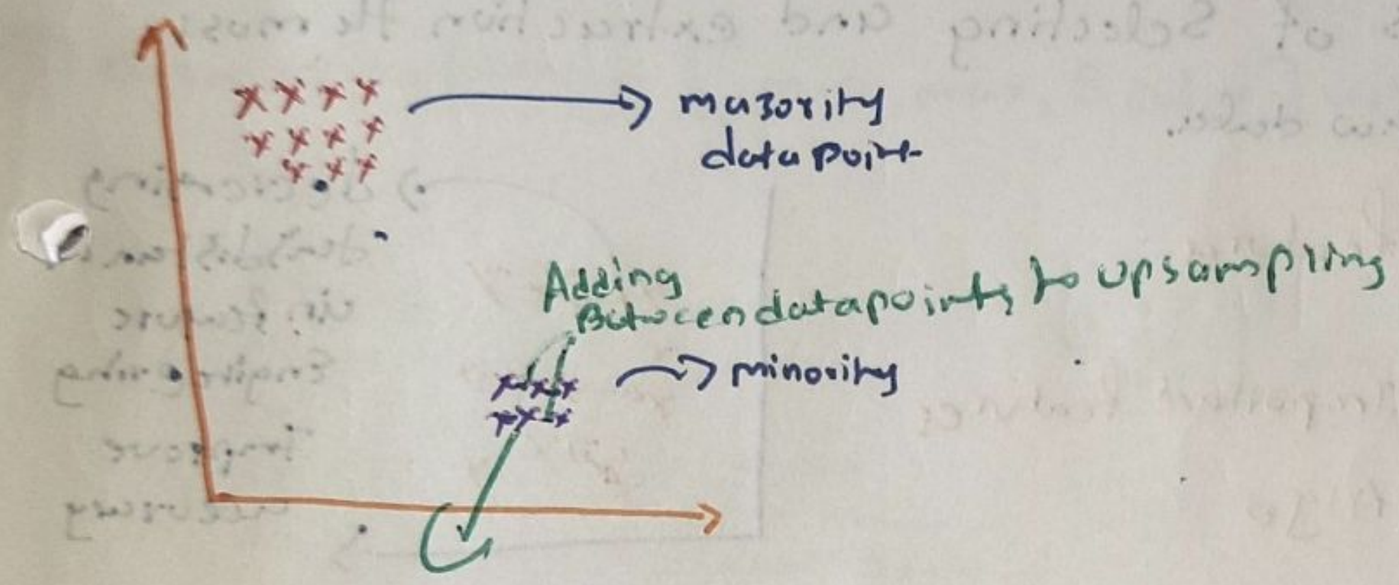
Balanced dataset.

(100, 100) (900, 900)



★ Smote (synthetic minority oversampling technique) → used in upsampling in minor

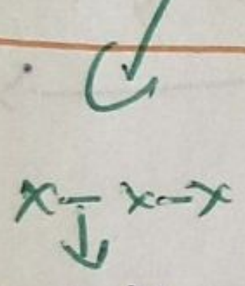
Adding datapoint in Between datapoints in minority

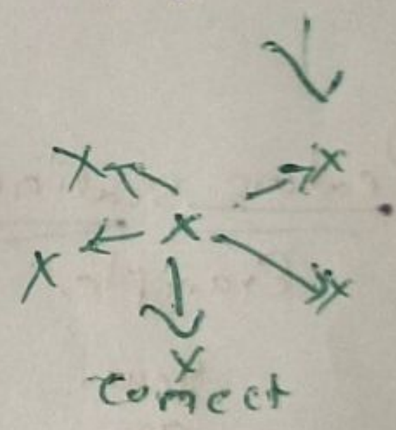


for applying Smote technique

→ pip install imblearn

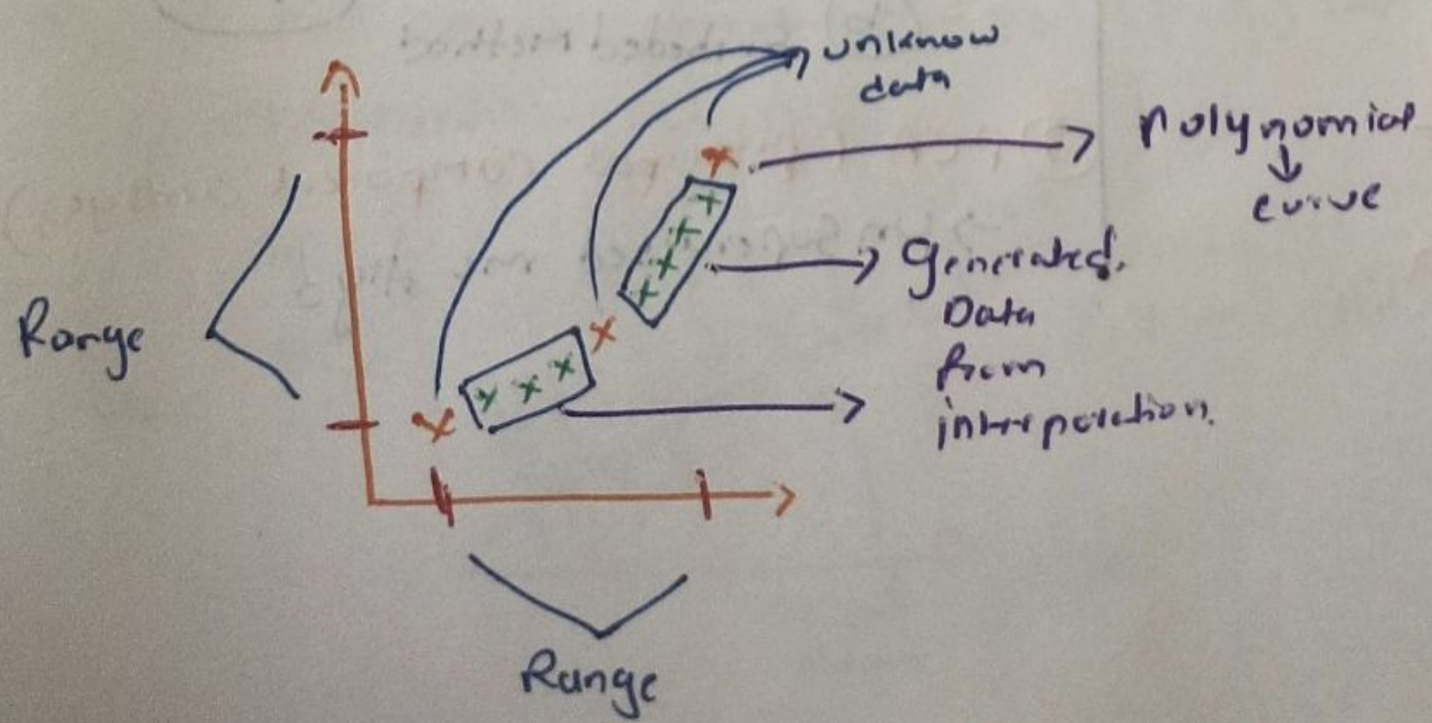
from imblearn.over_sampling import SMOTE





★ Data interpolation.

→ Data interpolation is the process of Estimating unknown values within a dataset Based on the known values. In Python there are various libraries available that can be used for data interporation, such as numpy scipy, pandas, Here is an Example of How to perform data interpolation using numpy library

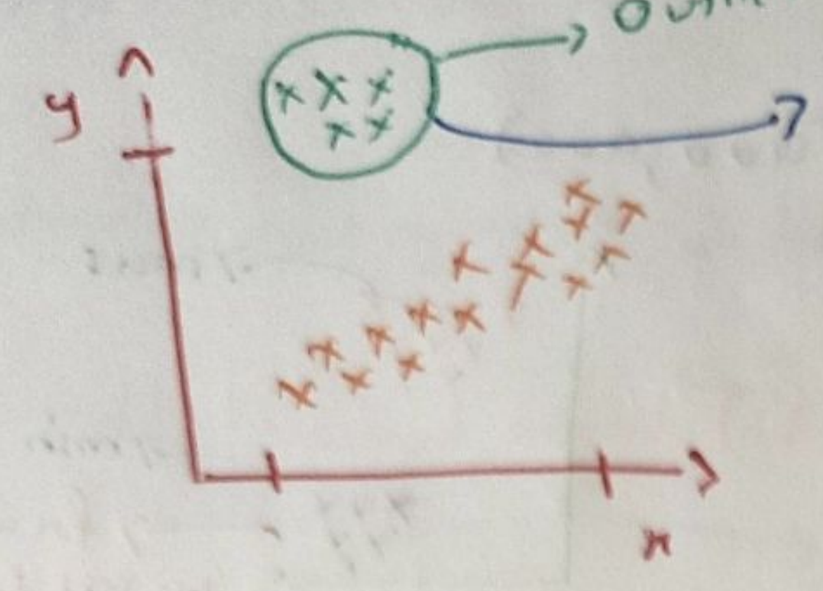


★ linear interpolation

★ Cubic interpolation

★ polynomial.

☆ Handling outlier



completely different from distribution

☆ 5-number summary

Calculating outlier

1. minimum value
2. Q1 - 25 percentile
3. median
4. Q3 - 75 percentile
5. maximum

☆ feature Extraction ⟶ data Scale down.

→ feature Extraction is process of selecting and extraction the most important features from raw data.

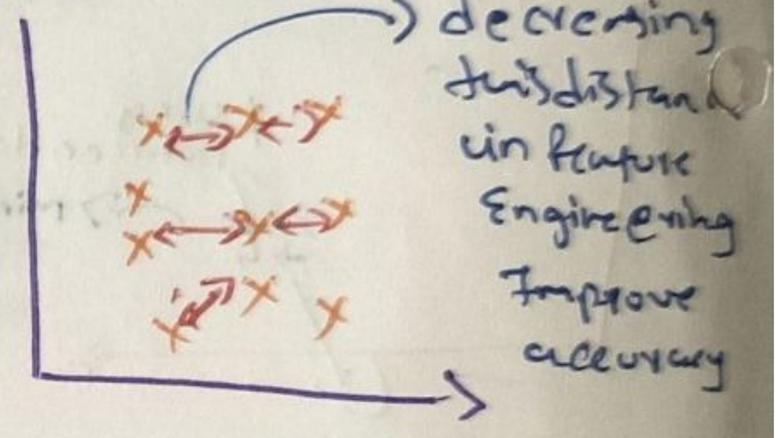
ML Applicatio ⟶ 1000 features

↓

most Important features

↓

ML Algo



① Feature scaling → [Scale] → [1 to 0]

Example:

Year age	weight	Height	BMI
32	70	140	—
28	75	160	—
35	80	155	—

normalize and standardize.

↓

Z-Score = $\mu=0$ $\sigma=1$

$$Z\text{-Score} = \frac{x_i - \mu}{u}$$

↓

Standardization

☆ normalization

[0, 1]

[min, max, Scaler]

unit vector

② Feature Selection:-

→ we just pick the most important features

(a) filter method

(b) Embeded method

top 10

③ PCA (principal component analysis)

→ Un Supervised ml Algo

Feature Scaling

① Standardization → $[\overset{\mu}{0}, \overset{\sigma}{1}]$

→ Z-Score

Feature (Age)

24
25
26
30
35
36

→ Z-Score $= \frac{x_i - \bar{x}}{\sigma}$ → $\mu = 0$, $\sigma = 1$

↓

$\mu = 0$
$\sigma = 1$

→ Age' $\mu = 0$ $\sigma = 1$

② normalization → [min, max, Scaler] → 0 to 1

Age

24
25
26
28
29
30

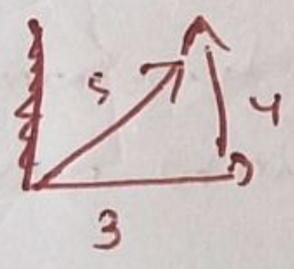
transform →

Age' [0-1]

$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$

③ Unit vectors → magnitude → 1

$\hat{a} = \left(\frac{3}{\|\vec{a}\|}, \frac{4}{\|\vec{a}\|}\right)$



$\|\hat{a}\| = \sqrt{(3/5)^2 + (4/5)^2}$

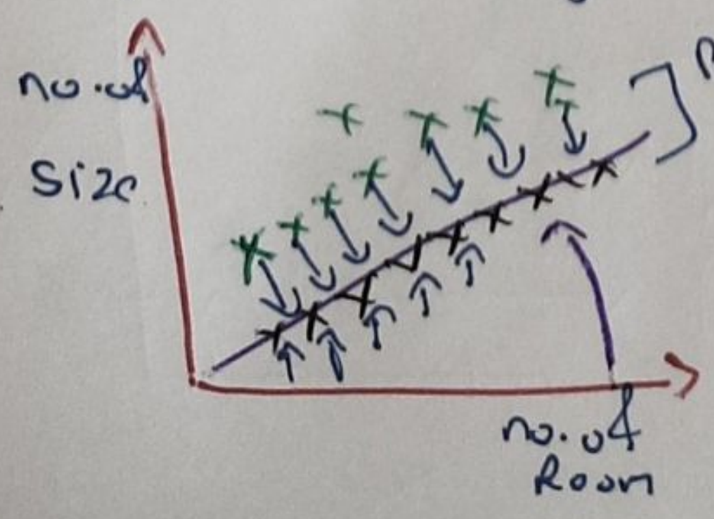
$= \sqrt{\frac{9+16}{25}}$

$= \sqrt{\frac{25}{25}} = 1$

★ Principal component analysis (PCA) → 2D → 1D also, 3D → 1D

100D → 1D

→ Unsupervised ML Algo



2D → 1D
X, Y → Mid

→ PC1 → Price.

Train mode.

★ Data encoding ⇒ categorical data ⟶ numerical data

Experience (IP)	education (IP)	Salary (OP)
4	B.E	50k
6	MASTER	60k
3	P.HD	100k
6	P.HD	120k

numerical data. | Categorical data

Data Encoding

Aim → categorical feature →

numerical

↓

ML model

↓

Train

★ type of Data encoding

(a) nominal (one Hot Encoding)
(b) ORDINAL and LABEL ENCODING
(c) Target Guided Ordinal Encoding

} Categorical → numerical.

(a) - Nominal encoding is a technique used to transform Categorical variables that have no intrinsic ordering into numerical value that can be used in ML models

f_1	f_2	f_3
	A	
	B	
	C	

Data Set → apply one Hot Encoding →

A	B	C
1	0	0
0	1	0
0	0	1

} added another three feature

Categorical.

Disadvantage.

① Spurce matrix. - [overfitting]

② 1000 feature. - 1000 matrix

Note :- don ot when there are many feature.

⑤ Target Guided ordinal encoding

-> Targeted variable in a mean replaced and make c. new datat for categoricated data