classmate
Date
Page

Machine learning

① Model performance - Accuracy ↑↑ → High
② overfitting, underfitting -
③ Bias vs Variance -

*** imp

$$\boxed{\text{Training} \propto \frac{1}{\text{Bias}}}$$

① overfitting underfitting

Case (i)

DATASET

Books → Train → Model is trained → Accuracy ↑↑ → 95%
Exam → Test → Model is tested → Accuracy ↓↓ → 60%

Low Bias ← overfitting
High variance

Case (ii)

Train → Accuracy ↓↓ → 55%
Test → Accuracy ↓↓ → 50%

underfitting
⇓
High Bias
High Variance

Generalized Model

Train → Acc ↑↑
Test → Acc ↑↑

Low Bias
Low Variance

Missing Values

Handling Missing Dataset :

Classification → Supervised ML

⇓

Output Categorical features

⇓

2 Categorical → yes/No

1000 datapoints

900 yes   100 yes

f1	f2	O/P
-	-	No
-	-	No
-	-	Yes

⇒ 900 yes 100 No's

900:100 ⇒ 9:1 (imbalanced dataset)

⇓

500:500

600:400

700:300

900 yes   100 No's

⇓

$\boxed{\text{Dumb Model}}$

⇓

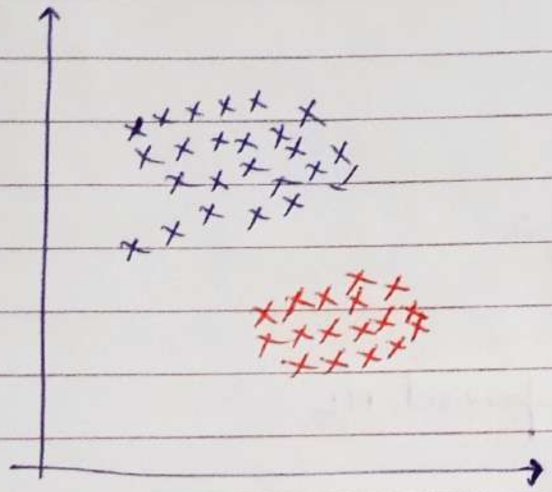
Yes

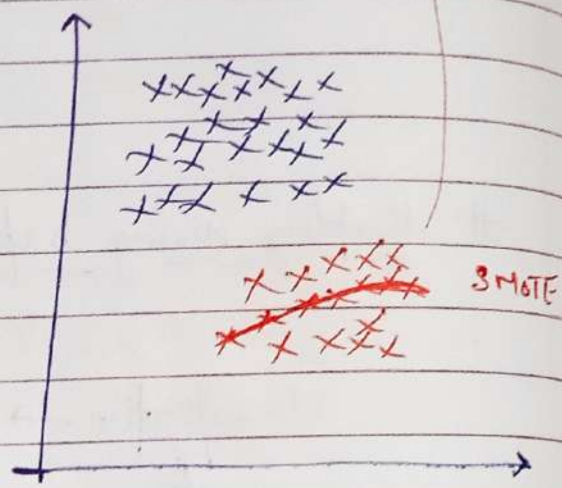
90% Accuracy

① upsampling ② Down Sampling

SMOTE

upsampling





Feature Extraction:-

Features extraction is process of selecting & extracting the most important features from raw data.

ML application ⟶ 1000 features

⇓

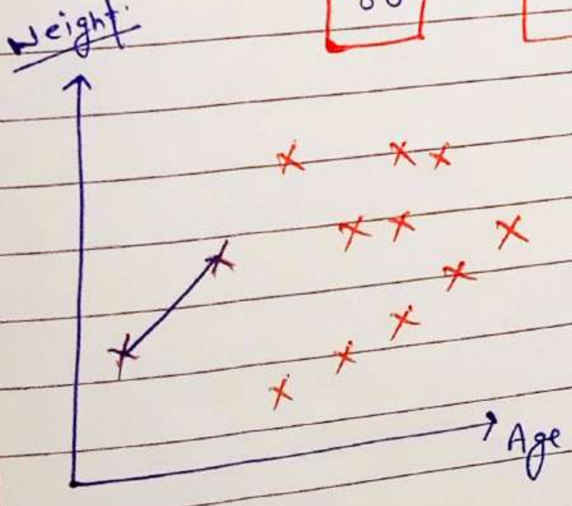
Most imp. features

⇓

M.L Algorithm.

1- Feature Scaling :-

Age (yrs)	Weight (kg)	Height (cm, inch)	BMI
32	70	140 cm	–
28	75	160 cm	–
35	80	155 cm	–



Normal or standardize

⇓

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

$$\boxed{Z\text{-score} = \left(\frac{x_i - \bar{x}}{\sigma}\right)}$$

⇓

Standardization.

Normalization: (min max Scaler)

(0,1)

unit vector.

(2) feature Selection → We just pick the most important features

500 features → Top 10 features
⇓
ML model train

(i) Filter method (ii) Embeded method

(3) PCA {Principal Component analysis}

* feature Scaling :-

(i) Standardization → Z-score
(ii) min max.

Standardization:-

Age
21
25
23
21
14
31

$$\boxed{Z\text{-}score = \frac{x_i - \bar{x}}{\sigma}}$$

$\mu = 0, \sigma = 1$

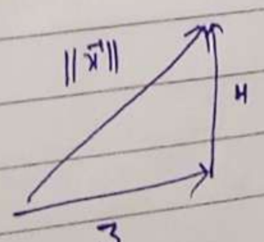
Normalization [min max scaler] - 0 to 1

Age
21
25
23
21
14
31

$$\boxed{x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}}$$

(3.) Unit vector - magnitude of 1

$\vec{x} = (3,4)$



$\|\vec{x}\| = \sqrt{(3)^2 + (4)^2} = \sqrt{25}$

5

$$\hat{u} = \left(\frac{3}{\|\vec{u}\|}, \frac{4}{\|\vec{u}\|}\right) = \left(\frac{3}{5}, \frac{4}{5}\right)$$

$$\|\hat{u}\| = \sqrt{\left(\frac{3}{5}\right)^2 + \left(\frac{4}{5}\right)^2}$$

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

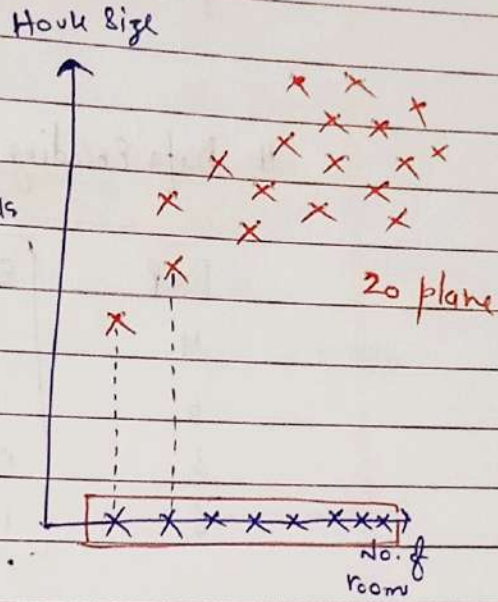
$$\sqrt{\frac{9+16}{25}}$$

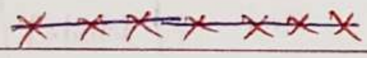
$$= \sqrt{\frac{25}{25}} = 1$$

PCA (Principal Components Analysis)

Dataset

No. of rooms	House Size	Price
2	800	40 lacs
–	–	–
–	–	–



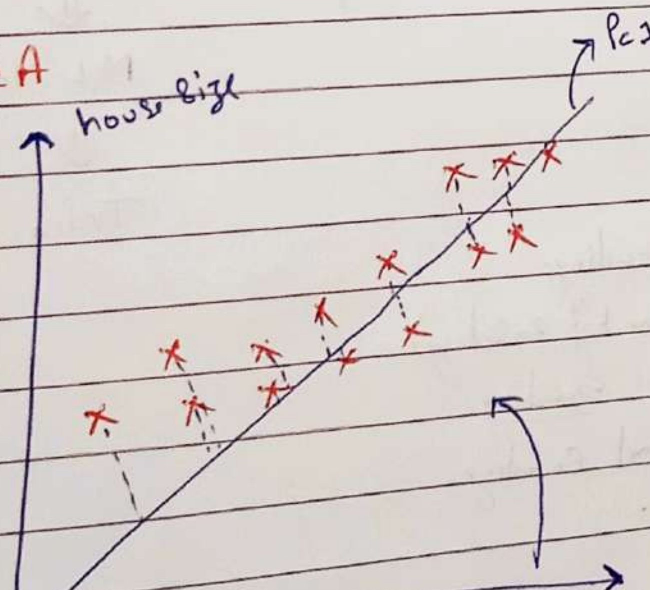


Variance ↑L   Rooms

No. of Rooms   Price   Machine Learning model

Problem :- Data Loss

PCA



2D - 1D

No. of Room House Size $\boxed{\text{Pc1} \quad \text{Price}}$ ⇒ Model train

Data Encoding :-

independent features

output features

Exp	Education	Salary
4	B.E	50K
6	PHD	100K
3	Master	80K
6	PHD	120K

Categorical features

Exp → $\boxed{\text{Model}}$ → Salary.
Education →

Data Encoding Aim → Categorical feature → Nominal
⇓
ML Model
⇓
Train.

* Types of data encoding
1. Nominal/OHE (one hot encoding)
2. ORDinal & Label Encoding.
3. Target guided ordinal Encoding.

(1) Nominal / OHE (one hot Encoding)

Nominal Encoding is a technique used to transform Categorical Variable that have no instrinsic ordering into numerical values that can be used in Machine learning models. One Common method for nominal encoding is one-hot encoding, which create a binary vector for each Category in the variable.

House Price (DATASET)

No. of Rooms	House Size	Location	Price
		Bangalore	
		Noida	
		Delhi	
		PATNA	
		Bangalore	
		Noida	

Location → OHE

Bangalore, Delhi, Noida

[1 0 0]
[0 0 1]
[0 1 0]
1 0 0]
[0 1 0]

→ Disadvantage:-

(1) sparse matrix [10 location)
↓
(overfitting)

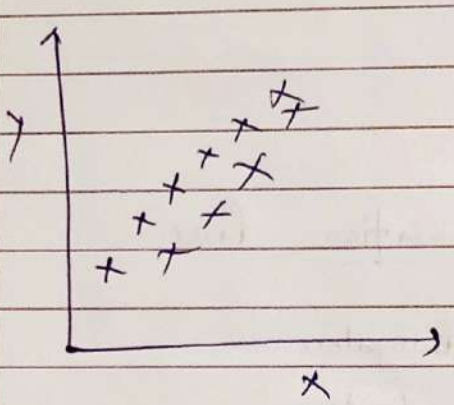
(2) 1000 features [1000 categories]
↓
Inc the no. of features

Covariance & Correlation:-

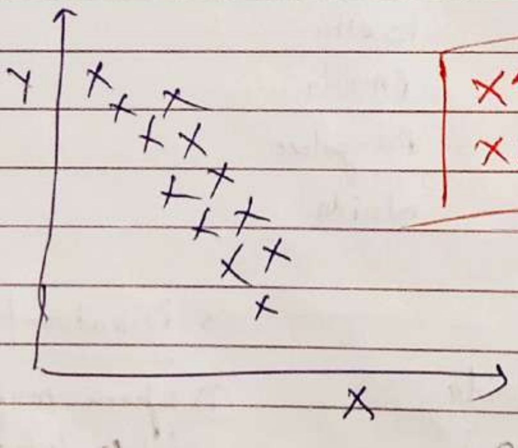
[Relationship bet^n x & y]

x	y
2	3
4	5
6	7
8	9

x↑ y↑
x↓ y↑
x↓ y↓
x↑ y↓



x↑ y↑
x↓ y↓



x↑ y↓
x↓ y↑

Ex:- ↓↑ Size of house → Price of house ↑↓ ↓ o/p

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

X	Y
2	3
4	5
6	7
8	9

$\bar{x} = 5$ $\bar{y} = 6$

$$Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$\frac{(x_i - \bar{x})(x_i - \bar{x})}{(n-1)}$$

$Cov(x,x)$ ⇒ specifying spread

14
4)24

X↑ Y↑
X↓ Y↓

+ve Covariance

X↑ Y↓
X↓ Y↑

-ve Covariance

X	Y
2	3
4	5
6	7

$\bar{x} = 4$ $\bar{y} = 5$

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{n-1}$$

= 4 +ve value

X & Y are having a +ve Covarriance

*** imp.
Advantages & Disadvantages

Advantages	Disadvantages
help us to find relationship betn X & Y +ve or -ve value	- Covarriance doesn't have specific limit value.

+150 +300

X Y Z

② Pearson Correlation Coefficient [-1 to +1)

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

① The more the value towards +1 the more +ve Correlated it is (x,y)

② The more the value towards -1 the more -ve Correlated it is (x,y)

- Spearman Rank Correlation (-1 to 1)

$$\gamma_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

X	Y	$\boxed{R(x)}$	$\boxed{R(y)}$
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

Feature Selection :-

+ve: Size of House ↑

+ve: No. of rooms ↑

+ve: Location ↑

No. of ~~people~~ staying ↑

⇓

Haunted

(-ve)

Price ↑↑

↑ -

X & Y are having a +ve Covarriance

*** imp.

Advantages & Disadvantages

Advantages	Disadvantages
help us to find relationship betn X & Y +ve or -ve Value	- Covarriance doesn't have specific limit Value.

+100 +300

X Y Z

(2) Pearson Correlation Coefficient [-1 to +1)

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

(1) The more the value towards +1 the more +ve Correlated it is (x,y)

(2) The more the value towards -1 the more -ve Correlated it is (x,y)