Special Topic → MACHINE LEARNING for NLP / Text Analytics

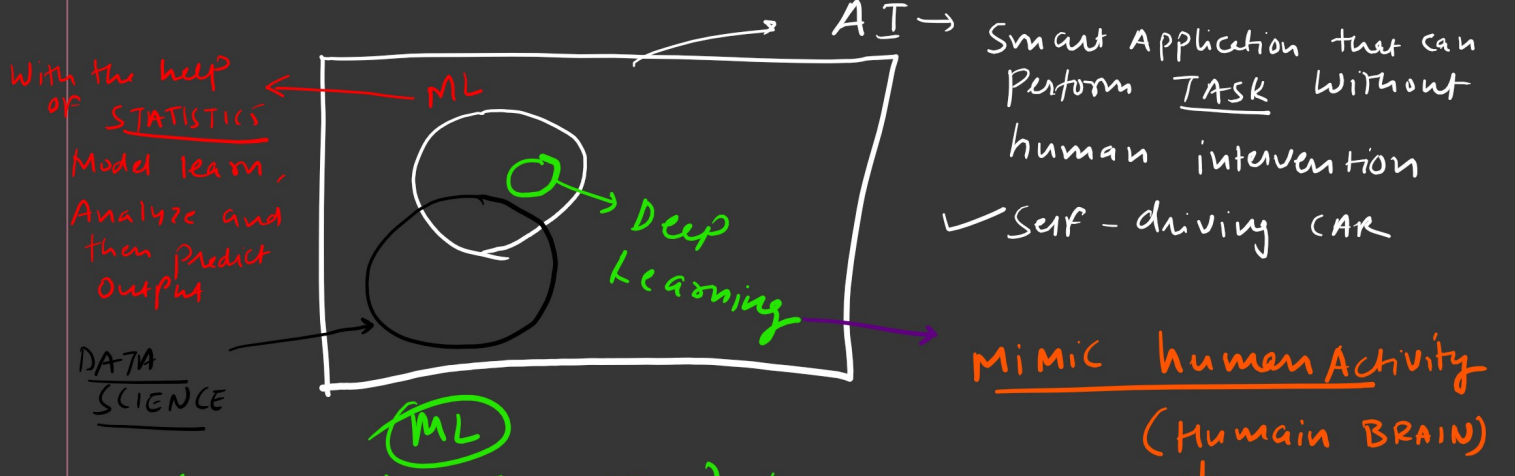
① Write
② Understand "MATHS"

AGENDA

① ABC (ML)
② Solve Algorithm (Maths and logic)

Text Analytics :

① Linear Regression
② Logistics Regression
③ Decision Tree
④ NAIVE BAYES
⑤ KNN



PRE - Requisite

1. Linear Algebra
2. Statistics
3. Probability
4. CALCULUS

STATISTICS - 101

Statistics is the science of collecting, organizing and Analyzing the DATA

"Piece of information"

STATISTICS

Descriptive

ORGANIZING + SUMMARIZING DATA

① M - Central Tendency

MEAN - MEDIAN - MODE

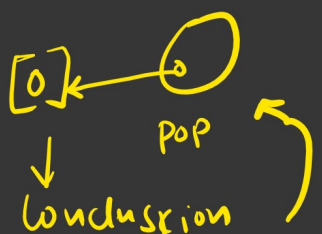
② MEASURE OF DISPERSION

VARIANCE

Std dev

Influential Stats

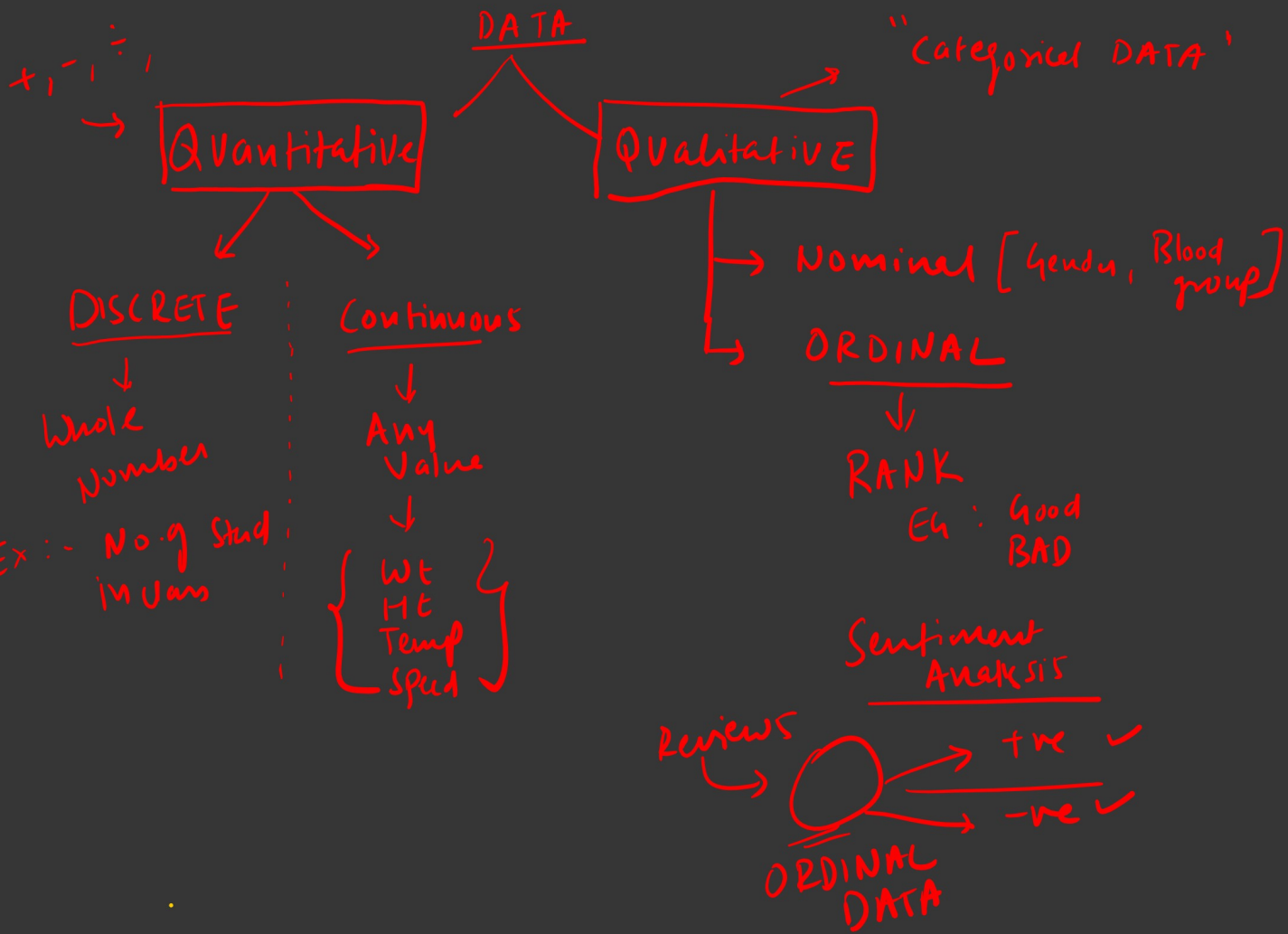
Measured DATA + Conclusion



- Z - test
- T - Test
- Hypothesis Testing

Summarizing DATA

- ✓ HISTOGRAMS
- ✓ BAR CHART
- ✓ PIE CHART



PRE Req SCALE of MEASUREMENT of DATA

1. NORMAL Scale DATA [ORDER does not Matter]
2. ORDINAL Scale DATA [RANK + Order matters]
3. Interval Scale DATA [0 → does not have zero Starting Value]
4. Ratio Scale DATA [Order/Rank matters + Ratio]

B.Tech → M.Tech → Phd → Postdoc
(4th) (3rd) (2nd) (1st)

PRE-Req from STATISTICS PERSPECTIVE

① MEAN, MEDIAN, MODE

MEAN $\mu = \sum_{i=1}^{n} \frac{x_i}{n}$

MEDIAN $X_1 = \{4, 5, 2, 3, 1\}$ ★ SORTING IS MUST

SORT = $\{1, 2, (3), 4, 5\}$ → Median Value

MODE = {Frequency Maximum}

$X = \{1, 1, 1, 2, 3, 4\}$ | Mode = 1

VARIANCE

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

x_i = Data point
μ = Population Mean
N = Size of population

STANDARD DEVIATION

$\sigma = \sqrt{\text{Variance}}$

⊘ Probability → Random Variable

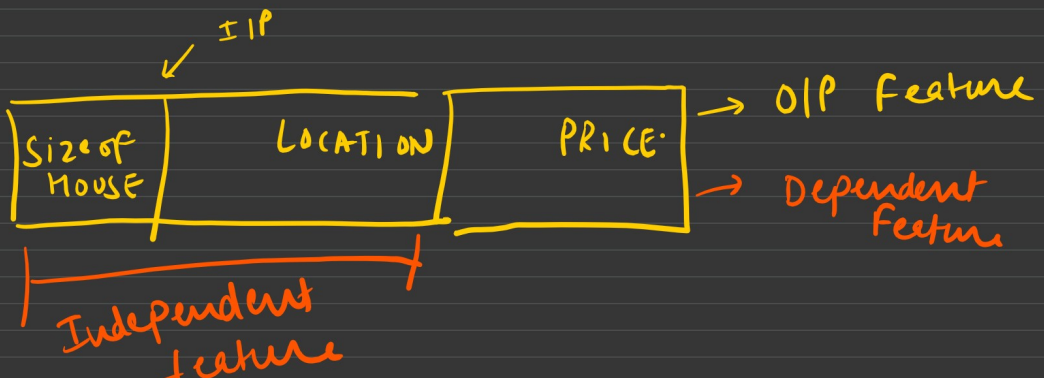
COVARIANCE - CORELATION

X	Y
2	3
4	5
6	7
8	9

Relationship btwn x and y

X ↑	Y ↓
X ↑	Y ↑
X ↓	Y ↓
X ↓	Y ↑

DATA



Independent

- Size of House ↑ → PRICE ↑
- LOCATION ↓ → PRICE ↓

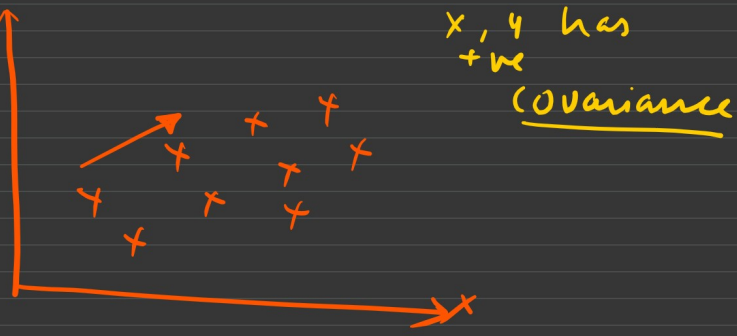
→ Size, Location

∴ Dependent feature

Scenario #1

X ↑ Y ↓

X ↓ Y ↑

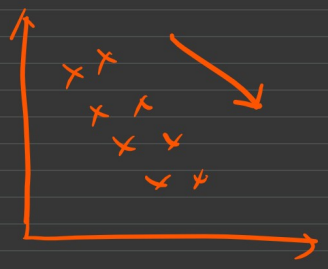


X, Y has +ve Covariance

Scenario #2

X ↓ Y ↑

X ↑ Y ↓



X, Y has Negative Covariance

COVARIANCE (X, Y)

$$\Rightarrow COV(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$ → Relationship between DATA POINTS

$$VARIANCE(X,X) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$ ↘ COV(X, Y)

$$VAR(X,X) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$ → "Spread of DATA"

X	Y
2	3
4	5
6	7

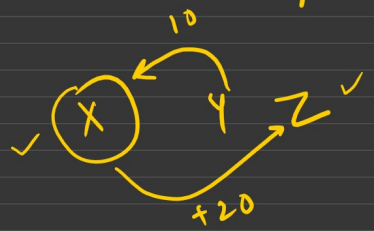
$\bar{X} = 4 \quad \bar{Y} = 5$

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

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

$$= \frac{4+0+4}{2} = 4$$

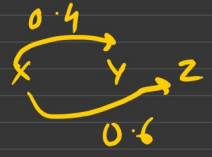
$Cov(X,Y) = 4$
$\therefore$ +ve Related



{Covariance doesnot have any specific limit}

[-1 to 1]

* PEARSON CORELATION COEFFICIENT



↓

[-1 to 1]

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = [-1 \text{ to } 1]$$

-ve correlated +ve (0-related

*

Size of House ^ Rooms ^ location X

Pearson coefficient → PRICE

Text Analytics → Pearson Correlation Coefficient

For your knowledge

(3) SPEARMAN RANK CORELATION

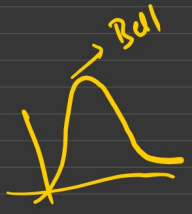
$$r_s = \frac{Cov(R(x), R(y))}{\sigma R(x) \cdot \sigma R(y)}$$

R(x) = Rank of X
R(y) = Rank of y

HISTOGRAMS

- Skewness
- Boxplot

PROBABLITY DISTRIBUTION



1. Normal / Gaussian
2. Bernoulli distribution
3. Uniform distribution
4. log Normal distribution
5. Poisson distribution
6. Power law distribution
7. Binomial distribution

PMF = Prob Mass function

PDF = Prob density function

① Percentiles and Quantiles

Percentage : 1, 2, 3, 4, 5, 6

% of No. that are odd = $\frac{3}{6} = \frac{1}{2} = 50\%$

PERCENTILE : Value below which a certain % of observation OR DATA POINT LIES

Percentile

Quartiles

$Q_1 \rightarrow$ 25 Percentile

$Q_2 \rightarrow$ 50 Percentile

$Q_3 \rightarrow$ 75 Percentile

SUMMARIZE the DATASET (Boxplot)

1. Minimum = 1
2. Q1 = 2
3. MEDIAN = 5.5
4. Q3 = 6
5. MAX = 10

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 9, 10, 29\}$

1. Outlier

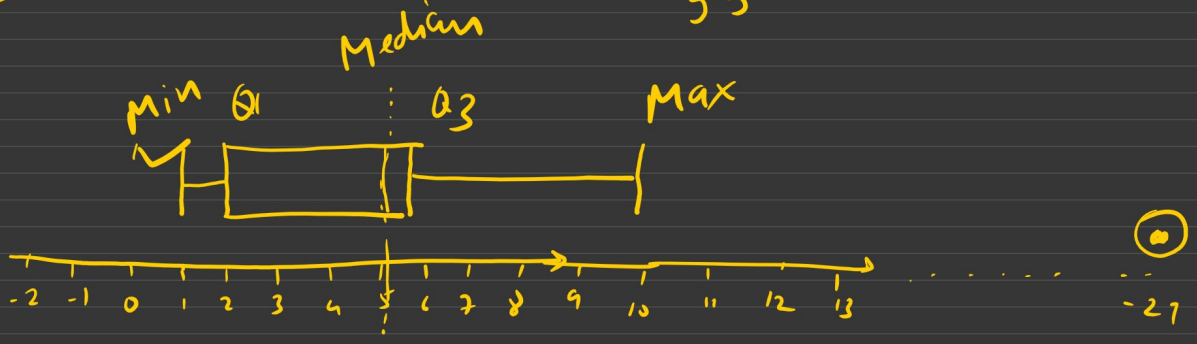
$X' = \{1, 2, 2, 2 \mid 3, 3, 4, 5 \mid 5, 6, 6, 6 \mid 7, 8, 9, 10\}$

Q1 Q2 Q3 Q4

5.5

Box plot

5



MACHINE LEARNING

Types of ML

1. Supervised ML
2. UNSUPERVISED ML
3. SEMI - SUPERVISED ML
4. Reinforcement ML

① SUPERVISED ML

- CLASSIFICATION → Categorical DATA
- REGRESSION → CONTINUOUS DATA

DATASE → Specific O/P → MODEL → O/P

★ CLASSIFICATION ML

PLAY HOURS	STUDY HOURS	PASS/FAIL
8	2	FAIL
7	3	FAIL
3	6	PASS

⇒ Categorical Feature → PASS / FAIL

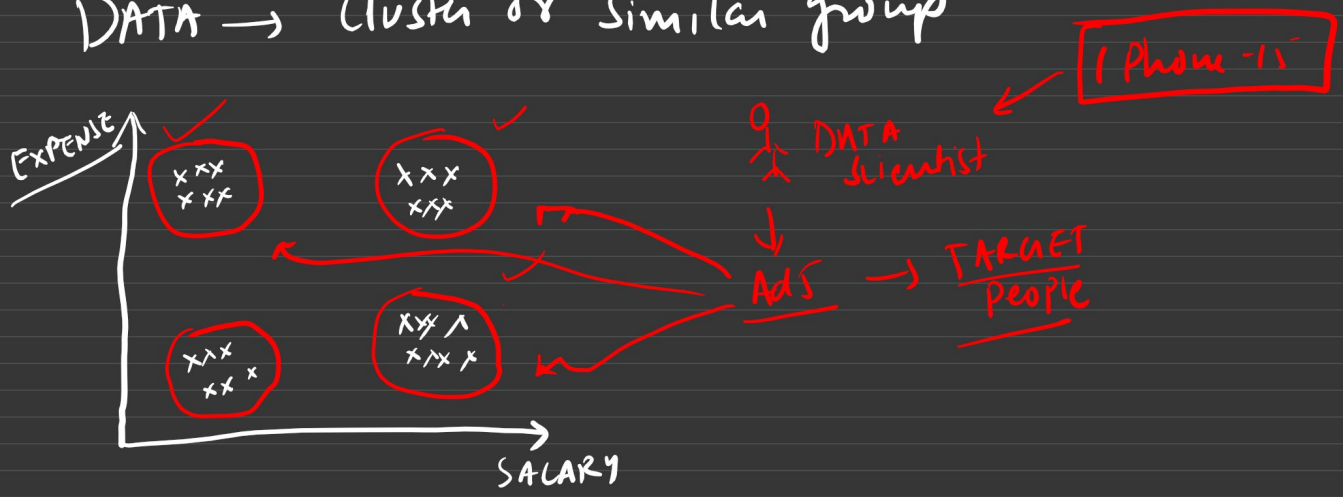
REGRESSION

Size of House	No. of Rooms	PRICE
		1M
		1.2M
		1.6M

← Continuous

② UNSUPERVISED ML

DATA → Cluster or Similar group

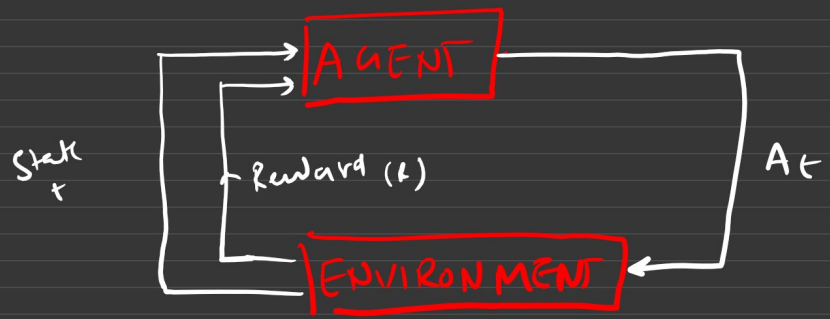


③ Semi Supervised

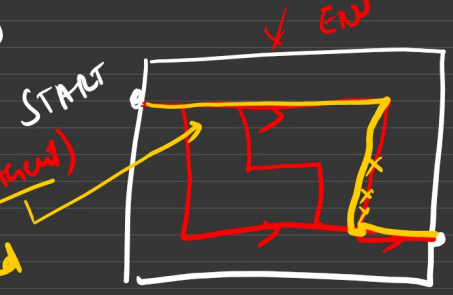
Supervised + Unsupervised

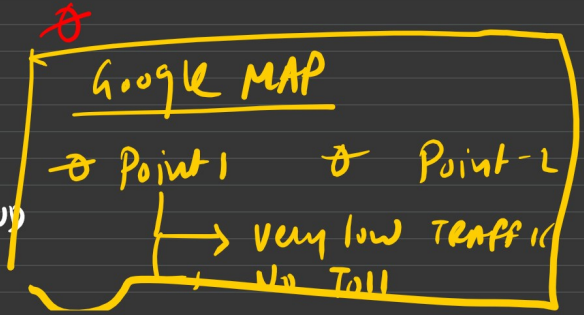
↓ Classification / Regression : / "Clustering"

④ Reinforcement learning



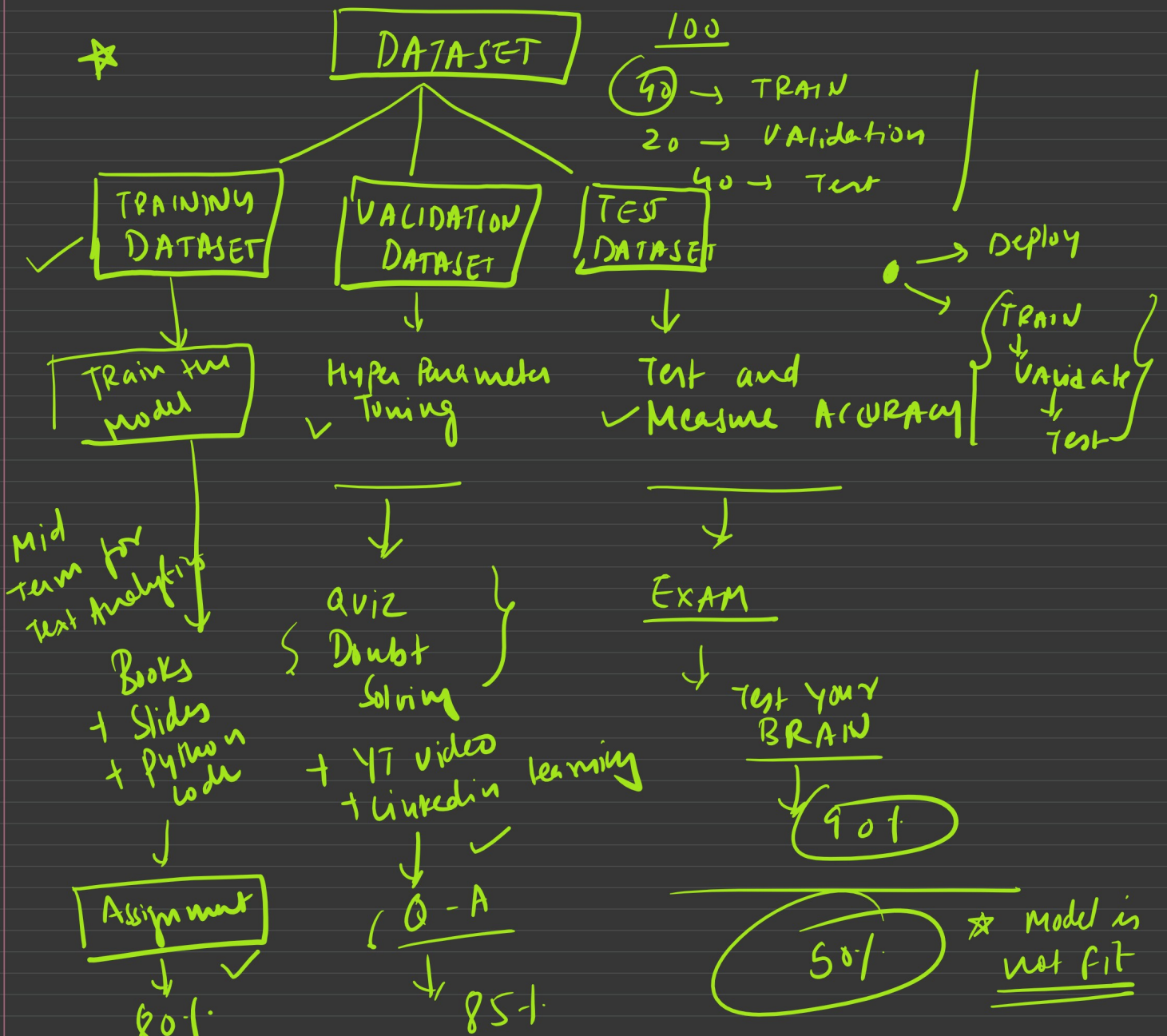
Google MAP





Reinforcement learning

Area of ML Concerned With how intelligent Agents Ought to take Actions in an Environment in Order to Maximize the notion of Cumulative Rewards.



- DATASET
- Model Performance

BOOK → TRAIN → Model is TRAINED
EXAM → TEST → Model is Tested

ACCURACY → 95%
Accuracy → 50%
} OVER FITTING

★ { low BIAS
High VARIANCE }

TRAIN → ACCURACY 55%
TEST → ACCRacy 50%
} UNDERFIT Model

{ high BIAS
high VARIANCE }

Overfit → Retrain the Model With New Algorithm/Method

Underfit → Retrain the Model With More DATA

GOAL :-

GENERALIZED Model

[TRAIN → Acc ↑↑ (85%)
TEST → Acc ↑↑ (80%)] ⟹ LOW BIAS
LOW VARIANCE

FEATURE EXTRACTION

- F.E is Process of selecting and Extracting the **Most important** Feature from Raw data

ML App → 1000 Feature
↓
All Features are Not Imp
↓
Select "Features"
↓
TRAIN the Model

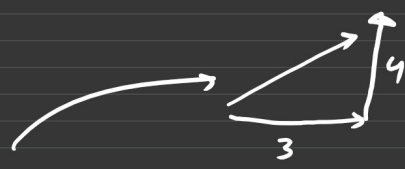
1. Feature Scaling
2. Feature Selection
3. P.C.A → PRINCIPAL Component Analysis

1. Feature Scaling
   - Standardization
     - Z-score
       - $Z\text{-score} = \frac{x_i - \bar{x}}{\sigma}$
   - Normalization
     - Min-Max Scaler
       - Transform [0,1]
   - Unit Vector ← {Text data}
     - Text → Words → Vector → ML

UNIT VECTOR

- Conversion to Unit Vector → Vector
  - ↓ Magnitude of 1
  - Magnitude of Vector
  - Vector Coordinate for Unit Vector
  - Magnitude of UV = 1

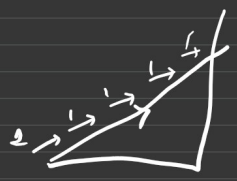
$\vec{x} = (3,4)$



Pythagoreas thm

$|\vec{x}| = \sqrt{3^2+4^2} = \sqrt{9+16} = 5$

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



Calculate Magnitude of Unit Vector = 1

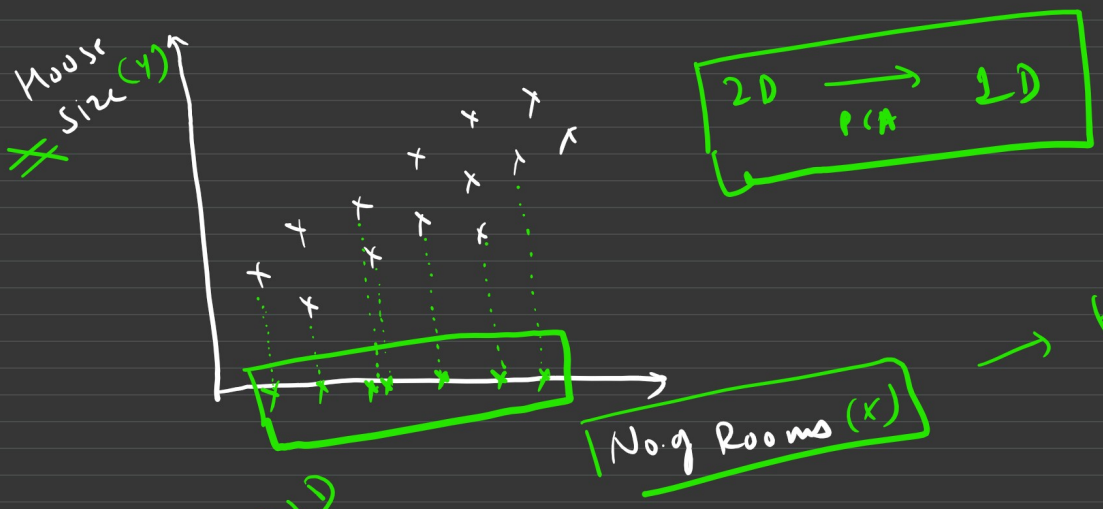
$|\hat{u}| = \sqrt{\left(\frac{3}{5}\right)^2 + \left(\frac{4}{5}\right)^2} = \sqrt{\frac{9+16}{25}} = \sqrt{\frac{25}{25}} = 1$

PCA → PRINCIPLE COMPONENT ANALYSIS

DATASET → 100 features → 10 features → ML Algorithm

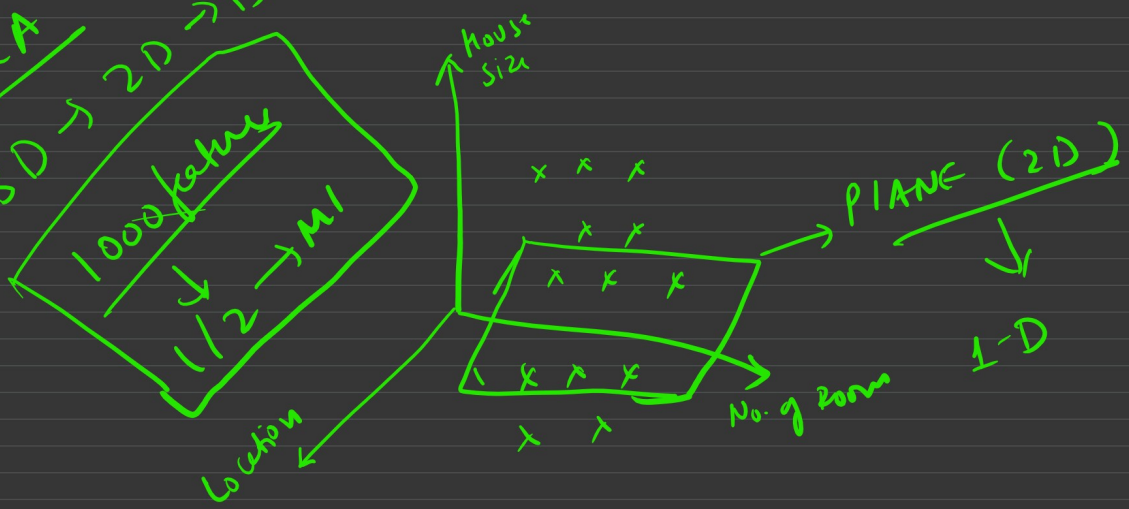
Disadvantage

- Loss of DATA



PCA

3D → 2D → 1D



⊘ loss of data

PCA → Hypothesis testing

DATA Scientist

→ COVID-19 vaccine discovery

PCA

AGE	NATIONALITY	RNA	DNA	Blood group	Wt	SURVIVAL RATE
1	2	3	4	5	6	
→	×	→	×	→	×	

⊘ Hypothesis Blood Group (A^+) → less chance of survival in covid

True / False

#1 LINEAR REGRESSION

ML Algorithms
- Supervised
  - Regression → O/P → continuous
  - Classification
    - → Multi class Classification {Blood group}
    - Binary {PASS, FAIL}
- Unsupervised

SIMPLE LINEAR REGRESSION

DATASET

Independent → I/P

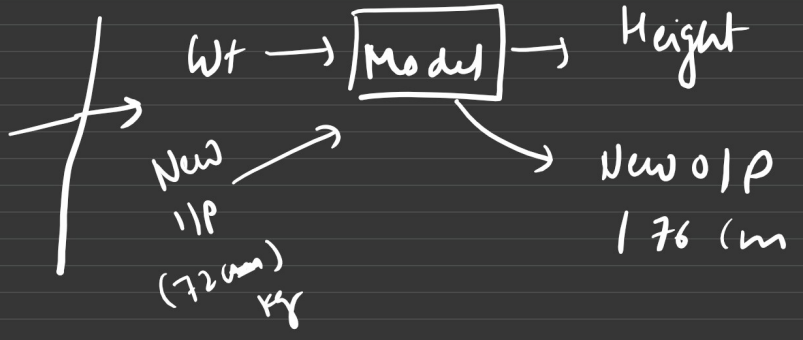
2 dependent → "O/P"

Weight	Height
74	170
80	180
75	175
-	-
-	-
-	-

72

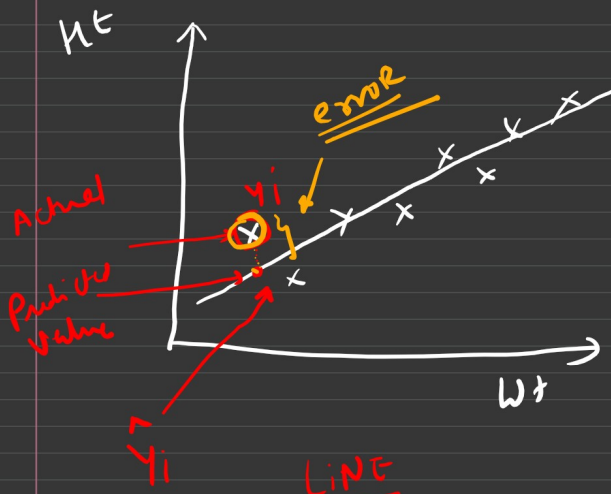
Wt (Given) → Ht (Predict)

∴ Prediction Model



Simple linear Regression

GOAL → Find **Best Fit line**, in such a way that the sum of error is ~~Zero~~ Minimum



y_i = Actual point

$\hat{y_i}$ = Predicted Point

ERROR = $y_i - \hat{y_i}$

LINE

$$\left\{ \begin{array}{l} y = mx + c \\ y = \beta_0 + \beta_1 x_1 \\ h\theta(x) = \theta_0 + \theta_1 x \end{array} \right\} \quad h\theta(x) = \theta_0 + \theta_1 x$$

θ_0 → Intercept

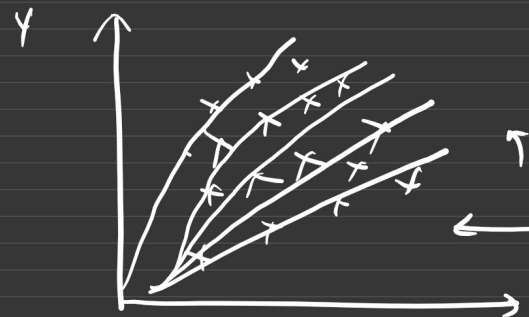
θ_1 → Slope / Coefficient

Goal → Minimize ERROR

BFL-1

ERROR = x_1

ERROR = x_2



↑ ERROR Reduced

(θ_1, θ_0)

θ_0 = Intercept

θ_1 = Slope

(θ_1, θ_0) → Min Error

* Manual (θ_0, θ_1) → Reduce ERROR

↓

Optimizer