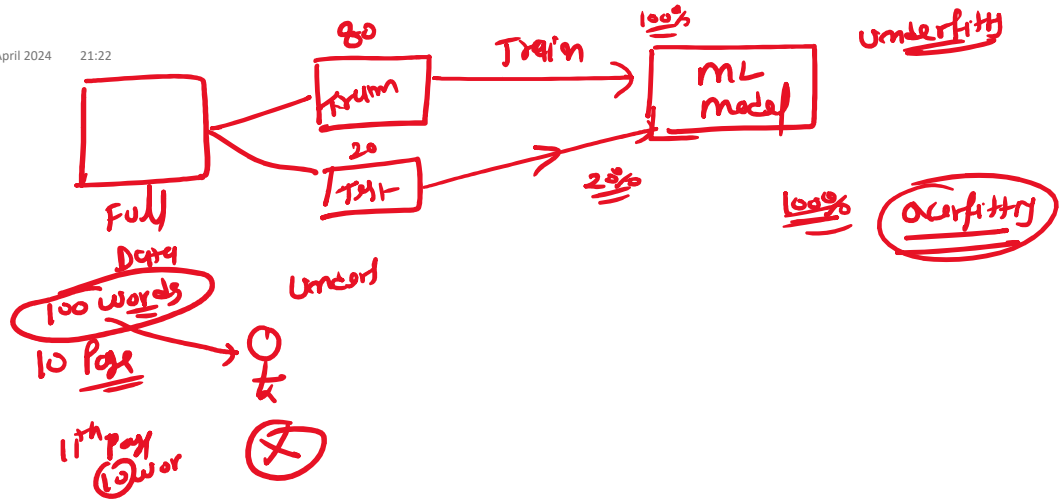
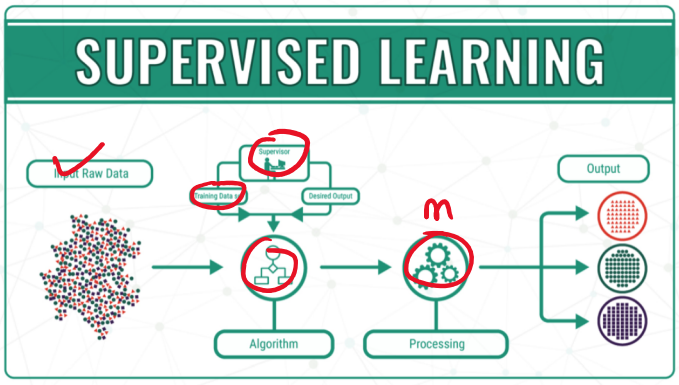
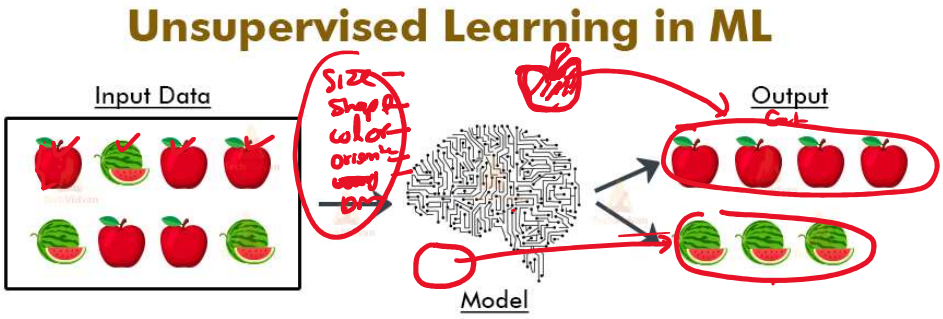
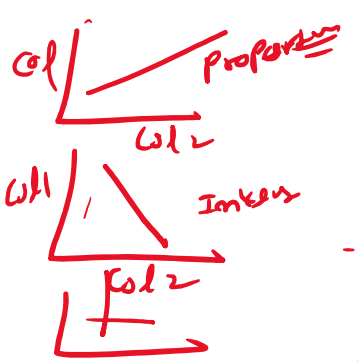
03 April 2024 21:22

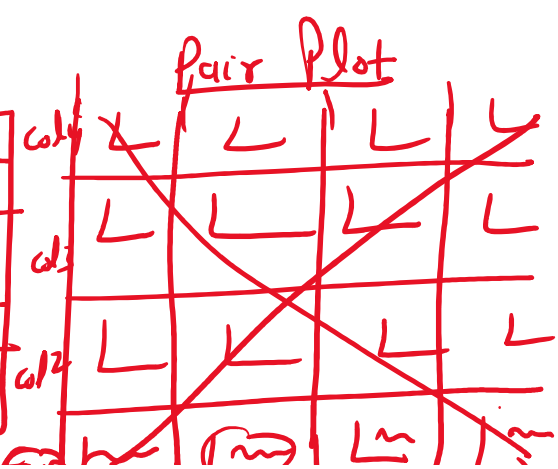


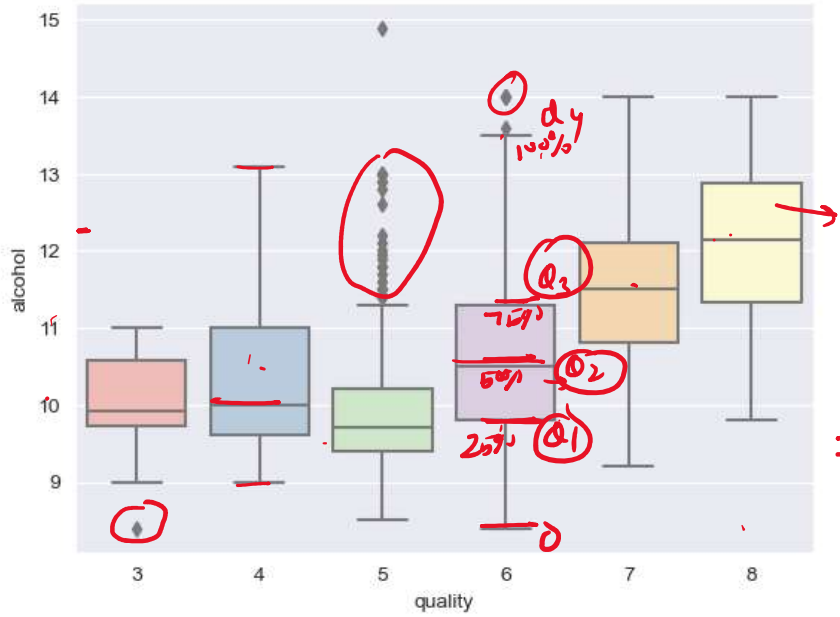




| Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|
| - | - | - | - |
| - | - | - | - |
| - | - | - | - |
| - | - | - | - |
| - | - | - | - |



max − min




$Q_4 - Q_0$

~~Variable~~ Variance

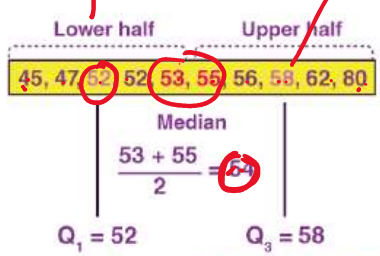
IQR

IQR
Inter Quartile Range

$Q_3 - Q_1$ = Median
$Q_2 =$

$Q_1 = 52$ $54 = Q_2$ $Q_3 = 58$

Asc
Desc




$Q_1$ 52
$Q_2$ 54
$Q_3$ 58 Median
$Q_4$ 80

$IQR = Q_3 - Q_1 = 58 - 52$
$= 6$

$IQR = 6$

12 −

$Q_3 + 1.5\ IQR$

$15 + 17.5$
32.5

−20.5 2, 9, 3, 4, 11, 17, 18, 13, 10, 6, 8, 150

−14

−20.5 2, 3, 4, 6, 8, 9, 10, 11, 13, 17, 18, 150

32.5 Out

Outlier

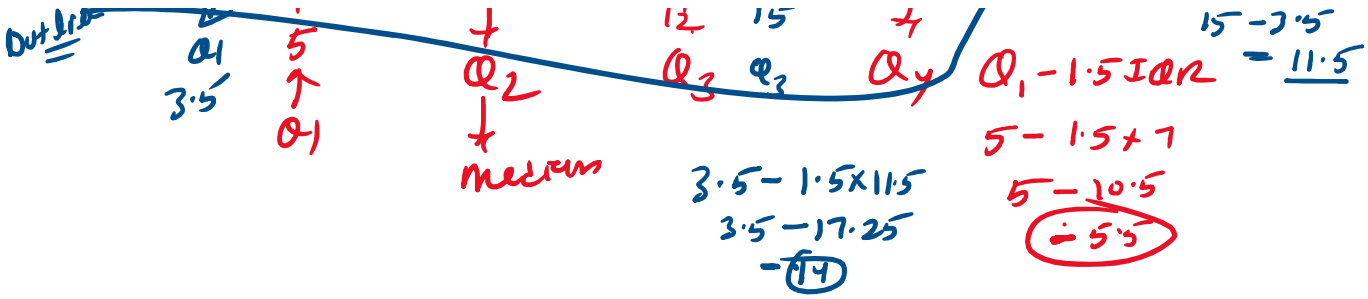
$Q_1$ 3.5 5 $Q_2$ 12 $Q_3$ 15 $Q_3$ $Q_4$

$Q_1 = 3.5$
$Q_2 = 9$ 9
$Q_3 = 12$ 15
$Q_4 = 18$ 150

$IQR = Q_3 - Q_1 = 7$

$15 - 3.5$
$= 11.5$

$Q_1 - 1.5\ IQR$

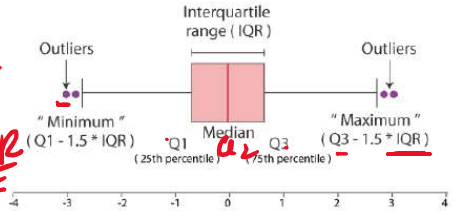



LB $Q_1 - 1.5\,IQR$

UB $Q_3 + 1.5\,IQR$

**Parts of Box Plots**

Check the image below which shows the minimum, maximum, first quartile, third quartile, median, and outliers:




Different parts of boxplot

$Q_1 - 1.5\,IQR$

Range 11 - 8 = 3 D1

8, 10, 8.5, 9, 9.5, 11, 10.5, 9.25, 8.75, 10.25, 10.75

Range 700 - 202 = 698 D2

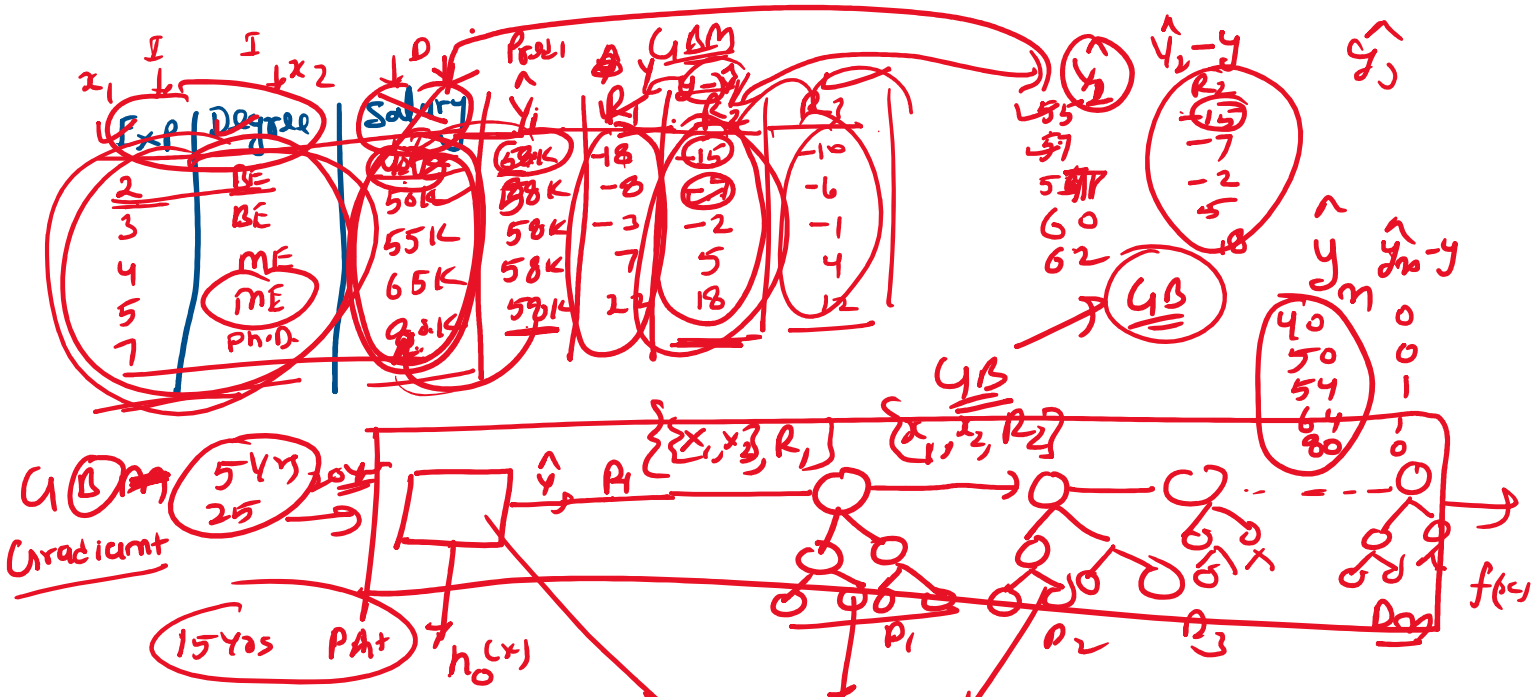
10, 3, 200, 15, 19, 22, 700, 350, 320

SL

GBM

Gradient Boosting Machine Learning Algorithm

R.F




$$f(x) = h_0(x) + l_1 h_1(x) + l_2 h_2(x) + l_3 h_3(x) \cdots l_m h_m(x)$$

$l_1$ = learning Rate

$0 - 1 = 0.05$

$$f(x) = h_0(x) + \sum_{i=1}^{n} l_i h_i(x)$$

$\hat{y}_n - y$

$\hat{y}_n - y \approx 0.5$

$\rightarrow F(x) = 58 + 0.1 \times (-18) + \cdots$

$= 58 - 1.8 - 3.0 - 2.5 - 1.8 - 0.9 - 0.6$

$= 56.2$

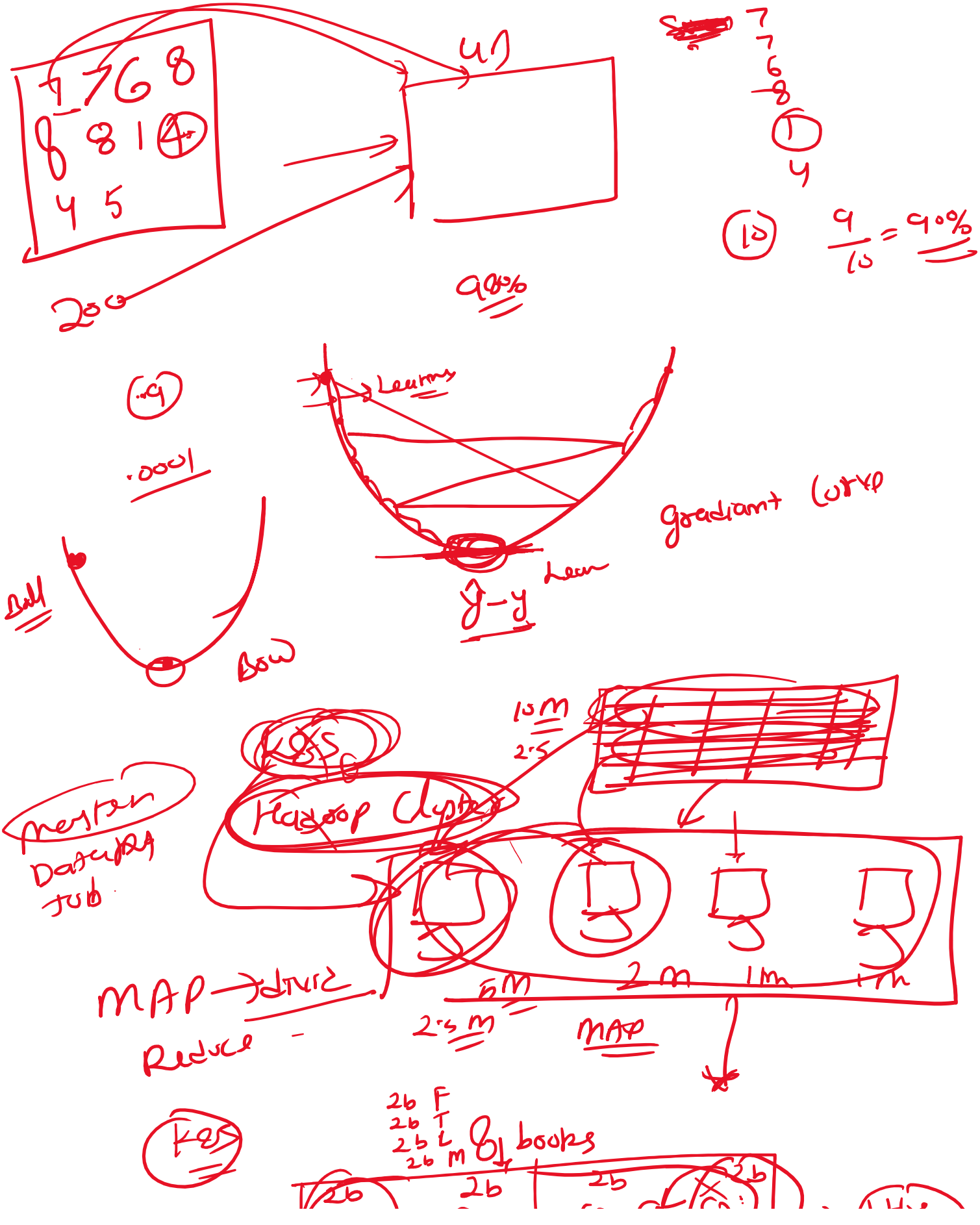
$40 = f(x) = 40$

$58 \rightarrow 40$

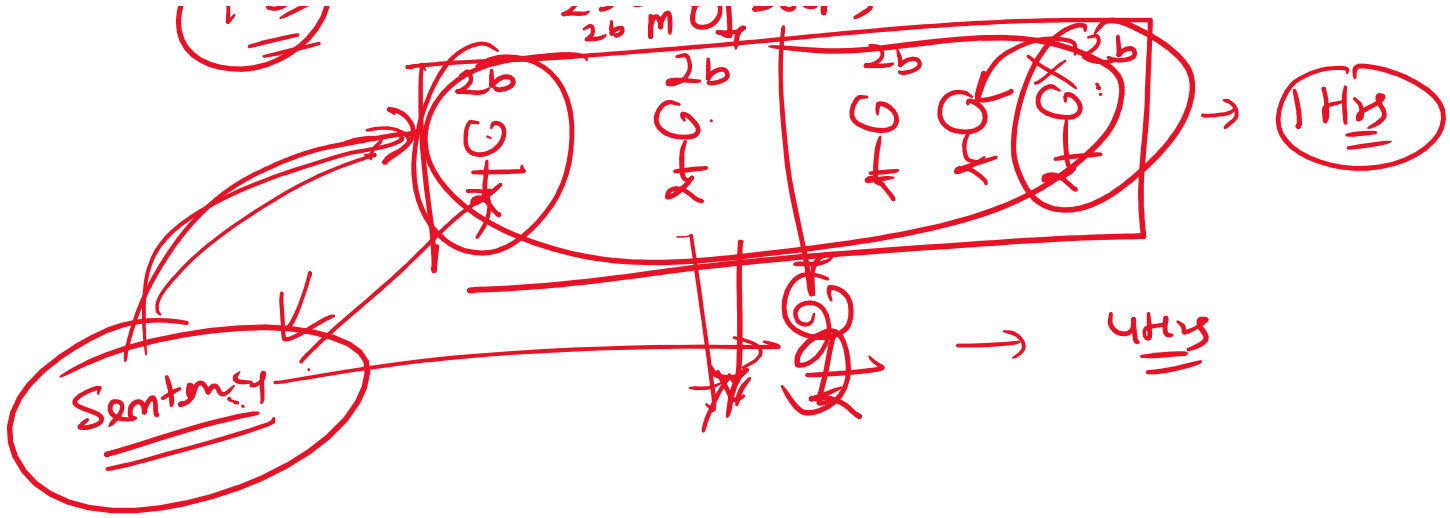
100000 DT

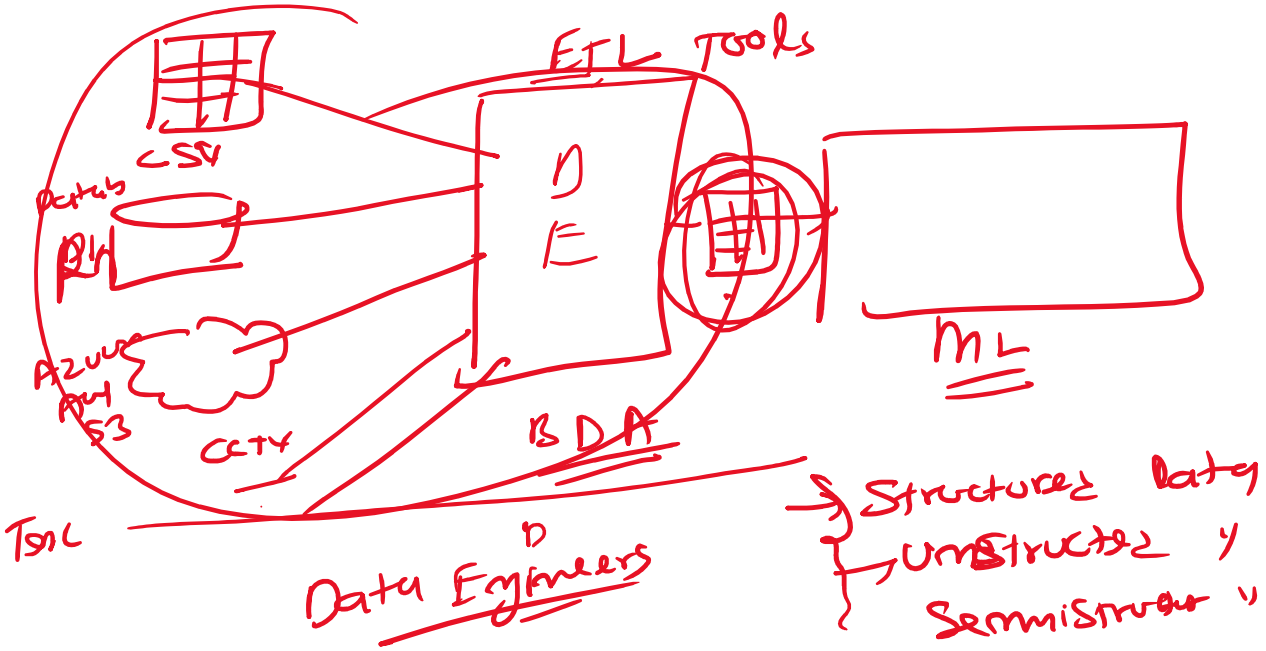
$= 62 + \cdots 0 = 1$

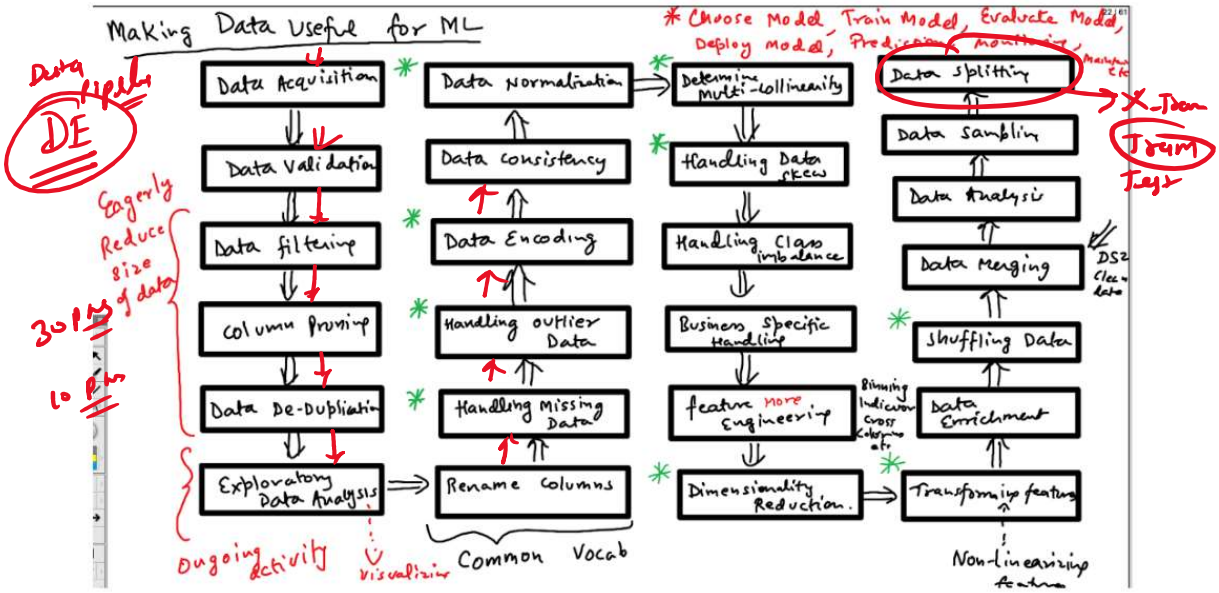
$= 82$

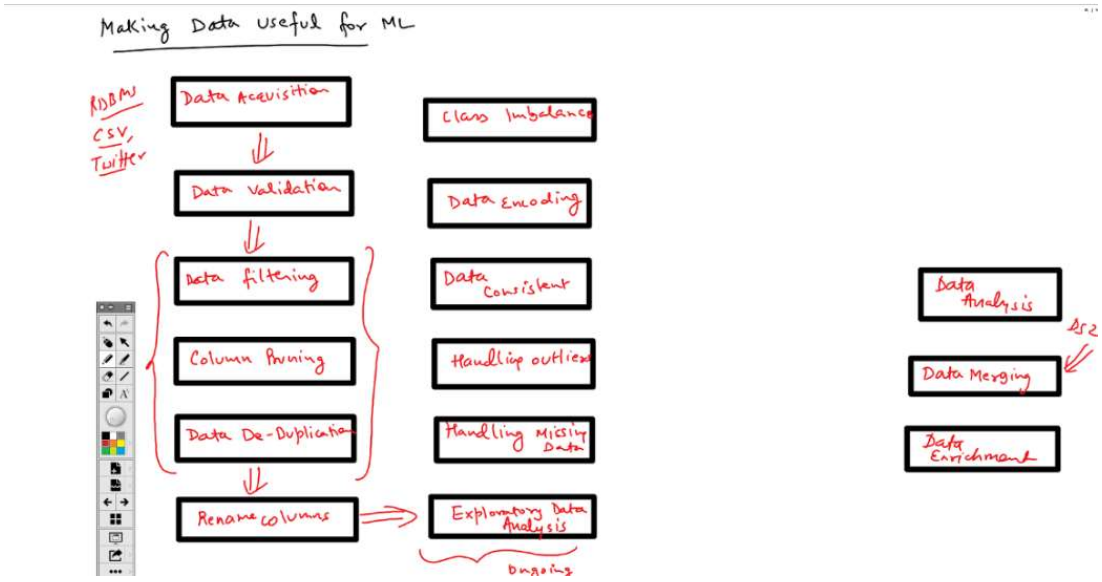


7 7 6 8
8 8 1 4
4 5
200
40
7
7
6
8
1
4
10
9/10 = 90%
90%
.9
.0001
Learning
Learn
ŷ - y
gradiant curve
Ball
Bow
10 m
2.5
Master
Data
Job
Hadoop Cluster
5 m
2 m
1 m
1 m
2.5 m
MAP
MAP → divide
Reduce
2b F
2b T
2b L
2b M
8 books
2b
2b
2b
2b


2b
2b
2b
2b
1 Hrs
4 Hrs
Sentency


ETL Tools
CSV
Patab
RH
Azure
Aux
S3
CCTV
D
E
BDA
ML
Tool
Data Engineers
Structured Data
Unstructed "
Semistructured "


Making Data Useful for ML
* Choose Model, Train Model, Evaluate Model, Deploy Model, Prediction, Monitoring, Maintenance etc
Data Pipeline
DE
Data Acquisition
Data Validation
Data filtering
Column Pruning
Data De-Duplication
Exploratory Data Analysis
Eagerly Reduce size of data
30 PMs
10 PMs
Ongoing activity
Visualizing
Data Normalization
Data consistency
Data Encoding
Handling outlier Data
Handling Missing Data
Rename Columns
Common Vocab
Determine Multi-collinearity
Handling Data skew
Handling Class imbalance
Business Specific Handling
feature more Engineering
Binning Indicator cross Colorings etc
Dimensionality Reduction.
Transforming features
Non-linearizing features
Data Splitting
X_Train
Train
Test
Data Sampling
Data Analysis
Data Merging
DS2 Clean data
Shuffling Data
Data Enrichment


Making Data Useful for ML
RDBMS
CSV,
Twitter
Data Acquisition
Data Validation
Data filtering
Column Pruning
Data De-Duplication
Rename columns
Class Imbalance
Data Encoding
Data Consistent
Handling outliers
Handling Missing Data
Exploratory Data Analysis
Ongoing
Data Analysis
DS2
Data Merging
Data Enrichment