[7 pages]

OM RAWAL
1811037

12/3/21 ISE DMBI

Q1 MCQ

1) A attribute subset selection

2) C 0.5.

3) C Classification

4) A Ordinal

5) B Entropy of a node decreases as we go down a tree

6) B By slope.

7) A Overfitting

8) B Mode.

9) B Sampling.

10) B 6. $\because \frac{10+2}{2} = \frac{12}{2} = 6.$

OM RAWAL
1811037

Q2

a)

Bin NO.	Values	Smooth by means
1	13, 15, 16	14.67, 14.67, 14.67
2	16, 19, 20.	18.33, 18.33, 18.33
3	20, 21, 22	21, 21, 21
4	22, 25, 25.	24, 24, 24
5	25, 25, 25	25, 25, 25
6	30, 33, 33	32, 32, 32.
7	35, 35, 35	35, 35, 35
8	35, 36, 40, 45	40.33, 40.33, 40.33
9.	46, 52, 70.	56, 56, 56.

Z score normalization = $\frac{V - \bar{A}}{\sigma_A}$

$\sigma_A = 12.94$

$V = 35$

Mean = 29.592

Z score of 35 = $\frac{35 - 29.592}{12.94}$

$= 0.4179$

OM RAWAL
1811037

(Q2b)

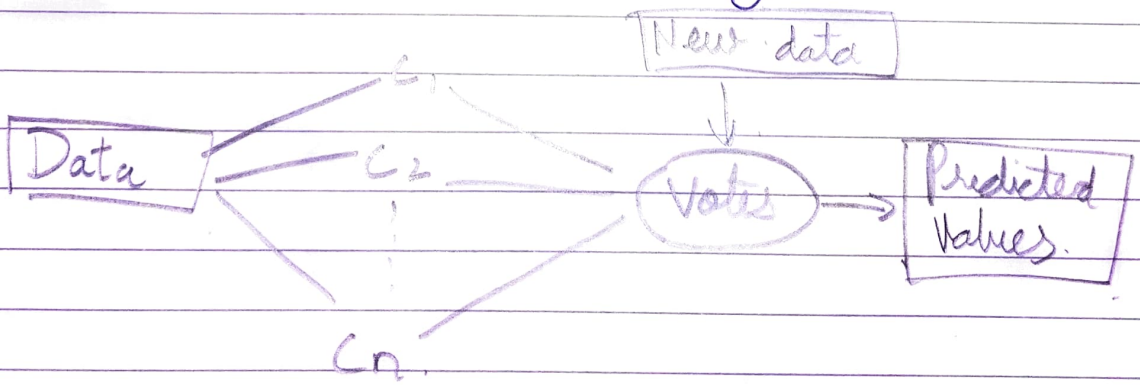
Bagging is also known as Bootstrap Aggregation.

It averages the prediction from the collection of various classifiers used.

A dataset D of n tuple for each iteration n tuples are sampled with replacement from D.
A classifier model M from each iteration is learned for each training set.
The bagged classifier M* uses voting method. for continuous values we take average.

Accuracy:

Uses voting so better than single classifier alone.
For noisy data it is robust
So overall accuracy is better.



OM RAWAL
1811037

Q2c

Count	X	Y	(A) $(x_i - \bar{x})$	(B) $(y_i - \bar{y})$	(A)×(B)	$(x_i - \bar{x})^2$
1	0	1	-1.375	-1.375	1.891	1.891
2	1	1.9	-0.375	-0.475	0.178	0.141
3	2	2.2	0.625	0.825	0.516	0.391
4	2.5	3.4	1.125	1.025	1.153	1.266
Sum	5.5	9.5			3.738	3.689
Mean	1.375	2.375				

Formula:-

$$y = a + bx$$

$$b = \frac{\Sigma [(x_i - \bar{x})(y_i - \bar{y})]}{\Sigma (x_i - \bar{x})^2}$$

$\bar{x}$ = mean of x.
$\bar{y}$ = mean of y

$$a = \bar{y} - b\bar{x}$$

Solution:

$$b = \frac{3.738}{3.689} = 1.0134$$

$$a = \bar{y} - b\bar{x} = 2.375 - 1.0134 \times 1.375$$
$$= 0.982$$

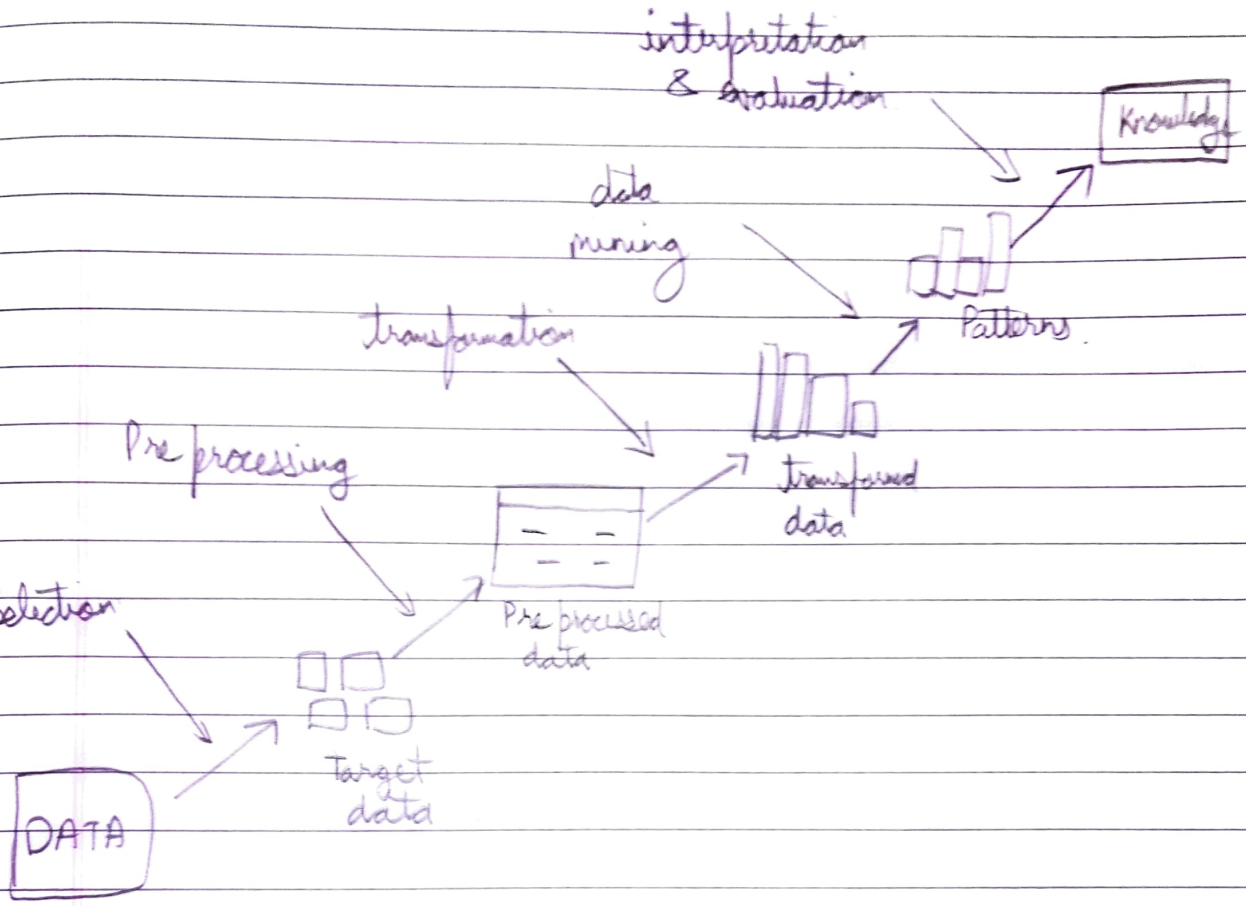
∴ Linear regression line equation is =.

$$y = 0.982 + 1.0134x$$

OM RAWAL
1811037

Q3.

KDD Process



- Steps are . assuming the COVID prediction (growth of cases)

1) Developing an understanding of application domain, prior knowledge & goals of end user.

Here application domain & goal is predicting the growth trend of cases & prior knowledge is past data from the start of pandemic.

OM RAWAL
181103J

Q3

2) Creating a target dataset.
the dataset will be patient info who contract the virus state & country level.

3) Data cleaning & preprocessing.

Noise & outliers like negative tested patients are removed.
None, Contact details are unnecessary
Missing data fields / test results are to be handled.

4) Data reduction & projection.
Removing unwanted info like gender.
dimensionality reduced data as the number of tests might be in lakhs.

5) Choosing data mining task.
Moreover we might use linear regression to predict the cases over the next couple of months.

6) Choosing data mining algorithm.
we might consider the parameters foreign travel, contact with infected, health worker, etc.

7) Data mining.
We find the growth by proper regression method.

8) Interpreting the patterns: Might increase or decrease depending on the parameters selected.
active cases

OM RAWAL
18110 37

Q3

9) The discovered knowledge of growth trends are represented using various data visualization tools.

Here in case of COVID cases we might have line graph, barchart on daily cases, logarithmic chart of cumulative value & also a pie chart of active, recovered & deaths.

—X—