Q2]

B]

i] Gini index.

We calculate the split for all attributes.

Income :-

$$\text{Split} = \frac{5}{14}\,\text{gini(Low)} + \frac{4}{14}\,\text{gini(High)} + \frac{5}{14}\,\text{gini(medium)}$$

$$= \frac{5}{14}\left[1-\left(\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2\right)\right] + \frac{4}{14}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right] + \frac{5}{14}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]$$

$$= 0.392$$

Default finglevel :-

$$\text{Split} = \frac{4}{14}\,\text{gini(High)} + \frac{6}{14}\,\text{gini(medium)} + \frac{4}{14}\,\text{gini(low)}$$

$$= 0.438$$

Credit score:

$$\text{Split} = \frac{7}{14}\,\text{gini(High)} + \frac{7}{14}\,\text{gini(low)}$$

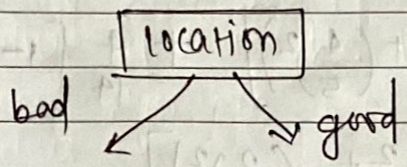
$$= 0.493$$

Location :

$$\text{Split} = \frac{8}{14}\,\text{gini(bad)} + \frac{6}{14}\,\text{gini(good)}$$

$$= \frac{5}{8}\left[1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)\right] + \frac{3}{8}\left[1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right)\right]$$

$$= 0.336$$



Location is the smallest hence it is the root node

Now we split

Income : $\text{split} = \frac{3}{8}\,\text{gini(low)} + \frac{2}{8}\,\text{gin(High)} + \frac{3}{8}\,\text{gini(medium)} = 0.295$

Defaulting : $\text{split} = \frac{3}{8}\,\text{gini[High]} + \frac{3}{8}\,\text{gini(medium} + \frac{2}{8}\,\text{gini(low)} = 0.34$

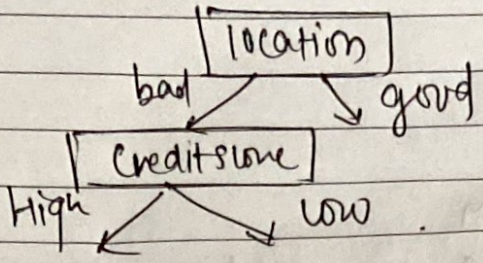
Credit score $\text{split} = \frac{4}{8}\,\text{gini(High)} + \frac{4}{8}\,\text{gini(low)} = 0.25$

G2020
SEM:06 ML

BARFI Page No. Date

Rebeca

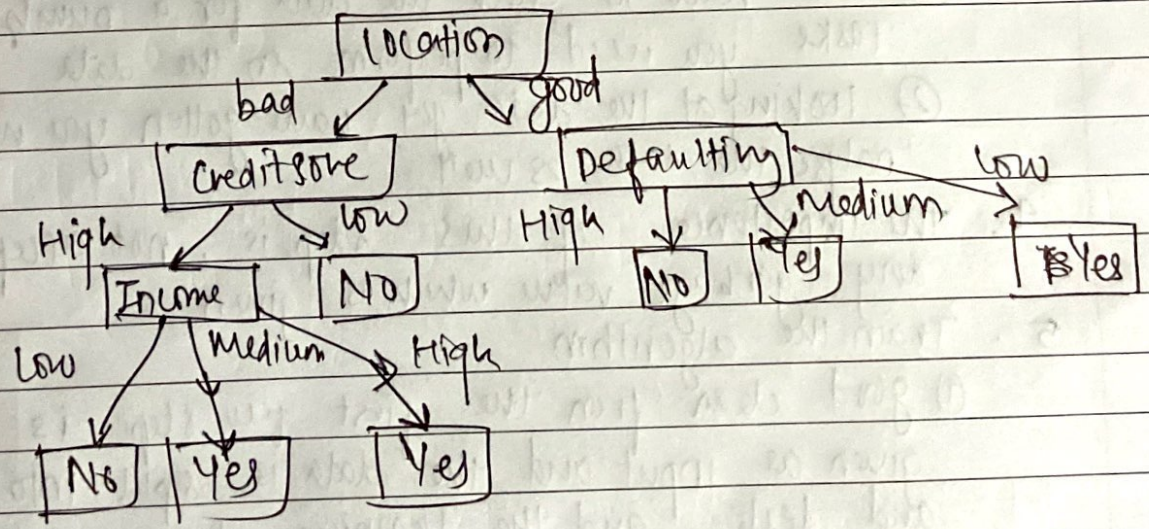
3/12



Now,

Income: split = $\frac{2}{6}$ gini(low) + $\frac{2}{6}$ gini(High) + $\frac{2}{6}$ gini(Medium) = 0.295

Defaulting: split = $\frac{4}{6}$ gini(High) + $\frac{2}{6}$ gini(medium) + $\frac{3}{6}$ gini(Low) = 0



Q2]
A]

iii) Steps in designing a machine learning model

1. Collection of data
   ① You should collect the samples from a website and extract the data
   ② From an RSS feed or an API.
2. Preparation of data
   ① Once you have the input data, you need to check whether if it useable format or not
   ② Some algorithm can accept some specific values
3. Analyse the input data
   ① You have to check the data for a number of tasks you need to perform on the data
   ② Looking at the data you have gotten you need to nalyse the data as well
4. The importance of this step is, not accepting any garbage value which is given
5. Train the algorithm
   ① good clean from the first few steps is given as input and the data is split into train and test and the training of the data is done.
6. Test the algorithm
   ① Information learned in the previous step is put to a test. In case of unsupervised learning, you have to use some other methods to do it
7. Use the gotten data and
   The program is made to do a specific task.

62020 ML
SEM:06

BARFI Page No.
Date / /

Rakesh

5 / 12

Q2]

A]

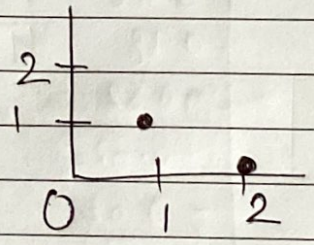
i]

① A support vector machine is a supervised learning algorithm that sorts data into 2 categories

② An SVM outputs a map of the sorted data with the margins between the two as far as possible.

③ SVM is called the maximum margin classifier because the numeric input variable (x) in your data (the columns) form an n-dimensional space

④ Consider building an SVM over a little data



⑤ The maximum margin vector will be parallel to the shortest line connecting points of the two classes. Therefore it passes through,

$$y = x_1 + 2x_2 - 5.5$$

⑥ Now, we construct a constraint over this

$a + 2a + b = -1$

$2a + 6b + b = 1$

Hence, $a = 2/5$

$b = -11/5$

62020
SEM: 06
Rebecca

ML

BARFI | Page No.
Date / /

6 / 12

(7) so the optimity plane here is given as

$\vec{w}$ (2/5, 4/5)
and $b = -11/5$

(8) The margin boundary is

$$2/|\vec{w}| = 2/\sqrt{4/25 + 16/25}$$
$$= 2/(2\sqrt{5}/5)$$
$$= \sqrt{5}$$

This answer can be confirmed geometrically looking at the graph.