Linear regression

Q-1 From the following data find the best fit line using linear regression of sales on purchase.

Sale	91	97	108	121	67	124	51	73	111	57
Purchase	71	75	69	97	70	91	39	61	80	47

→ Given that

Sale	Purchase
91	71
97	75
108	69
121	97
67	70
124	91
51	39
73	61
111	80
57	47

We know that Best fit line using linear regression is:

$$y = ax + b$$

where $$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

and $$b = \frac{\Sigma y - a\Sigma x}{n}$$

n	Sales (x)	Purchase (y)	xy	x^2
1	91	71	6461	8281
2	97	75	7275	9409
3	108	69	7452	11664
4	121	97	11737	14641
5	67	70	4690	4489
6	124	91	11284	15376
7	51	39	1989	2601
8	73	61	4453	5329
9	111	80	8880	12321
10	57	47	2679	3249
	$\Sigma x = 900$	$\Sigma y = 700$	$\Sigma xy = 66900$	$\Sigma x^2 = 87360$

$n = 10$

$\Sigma x = 900$

$\Sigma y = 700$

$\Sigma xy = 66900$

$\Sigma x^2 = 87360$

For value of a

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{10 * 66900 - 900 \times 700}{10(87360) - (900)^2}$$

$$= \frac{669000 - 630000}{873600 - 810000}$$

Assignment No.:_____ Page No.:____
Subject :_____
Roll No.:_____

(3)

Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

$= \frac{39000}{63600}$

$\therefore a = 0.613 \quad ---- (i)$

For value of b

$$b = \frac{\Sigma y - a\Sigma x}{n}$$

$= (900 - 0.613(900))/10$

$= (900 - 551.07)/10$

$\therefore b = 1218.3/10 = 124.83 \ldots (ii)$

$y = ax + b \quad ---- (iii)$

Putting value of a and b in eqn (iii) from eqn (i) and (ii) we get

$y = 0.613x + 124.83$

is the best fit line using regression for given data of sale on purchase.

Q-2 Find the best line using linear regression of Mth. Math on English given the following data set also estimate the error.

English	20	60	55	45	75	35	25	90	10	50
Math	20	45	65	40	55	35	15	80	25	50

We know that

Best fit line using linear regression $= y = ax + b$

where $a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$

and $b = \frac{\Sigma y - a\Sigma x}{n}$

Formula for error estimation $\varepsilon_{rr} = \sqrt{\frac{\Sigma y - y_i}{n}}$

n	English (x_i)	Math (y)	xy	x^2
1	20	20	400	400
2	60	45	2700	3600
3	55	65	3575	3025
4	45	40	1800	2025
5	75	55	4125	5625
6	35	35	1225	1225
7	25	15	375	625
8	90	80	7200	8100
9	10	25	250	100
10	50	50	2500	2500

$n = 10$ $\quad \Sigma xy = 24150$

$\Sigma x = 465$ $\quad \Sigma x^2 = 27225$

$\Sigma y = 430$

Value of a :

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{10 * 24150 - 465 * 430}{10 * 27225 - (465)^2}$$

$$= \frac{241500 - 199950}{272250 - 216225}$$

$$= \frac{41550}{56025}$$

$$\therefore a = 0.742 \quad \text{--- (i)}$$

Value of b :

$$b = \frac{\Sigma y - a\Sigma x}{n}$$

$$= \frac{430 - 0.742 * 465}{10}$$

$$= (430 - 345.03)/10$$

$$\therefore b = 8.497 \quad \text{--- (ii)}$$

We know that $y = ax + b$, from (i) & (ii) y will be :

$$y = 0.7422x + 8.497$$

For the error estimation, we know that

$$\sigma_{est} = \sqrt{\frac{\Sigma(y-y')^2}{n}}$$

n	Math (y)	y'	y - y'	$(y-y')^2$
1	20	23.337	-3.337	11.135
2	45	53.017	-8.017	64.272
3	55	49.307	5.693	32.410
4	40	41.887	-1.887	3.650
5	55	64.147	-9.147	83.667
6	35	34.467	0.533	0.284
7	15	27.047	-12.047	145.130
8	80	75.277	4.723	22.306
9	25	15.917	9.083	82.500
10	30	45.597	-4.403	19.386

$$\Sigma (y-y')^2 = 464.75$$

$$\therefore \sigma_{est} = \sqrt{\frac{464.75}{10}}$$

$$= \sqrt{46.474}$$

$$\therefore \sigma_{est} = 6.817$$

For best line using linear regression for math on english is $y = 0.742x + 8.419$ and estimated error is 6.817.

Q-3 Find the least Square regression line, $y = ax + b$ for the following data & also estimate the value of y when $x = 10$.

x	y
0	2
1	3
2	5
3	4
4	6

Formulae :

$$y = ax + b$$

where $$a = \frac{n\Sigma xy + \Sigma x + \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

and $$b = \frac{\Sigma y - a\Sigma x}{n}$$

x	y	xy	x^2
0	2	0	0
1	3	3	1
2	5	10	4
3	4	12	9
4	6	24	16

$\Sigma x = 10$ $\quad \Sigma xy = 49$

$\Sigma y = 20$ $\quad \Sigma x^2 = 30$

(8)

Value of a

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{5(49) - 10 \times 20}{5 \times 30 - (10)^2}$$

$$= \frac{245 - 200}{150 - 100}$$

$$= \frac{45}{50}$$

$$\therefore a = 0.9$$

$$b = \frac{\Sigma y - a\Sigma x}{n} = \frac{1}{5}[20 - 0.9 \times 10] = 2.2$$

We know that

$$\therefore y = ax + b$$

$$\therefore y = 0.9x + 2.2$$

Given that $x = 10$, $y = ?$

$$y = 0.9(10) + 22$$

$$y = 11.2$$

∴ For given data value of $x = 10$ they $y = 11.2$

Assignment No. : Page No. :
Subject :
Roll No :

(9)

Bayesian Classifier

Q-1 Consider the following data of buying_computer and classify a tuple X = (age <=30 ; income=medium, student = yes, credit_rating = fair)

Age	Income	Student	Credit_rating	Buy_computer
<=30	High	no	fair	no
<= 30	High	no	excellent	no
31...40	High	no	fair	yes
>40	Medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	High	yes	fair	yes
31...40	medium	no	excellent	yes
>40	medium	no	excellent	no

P(Ei) :

P(buy_computer = "yes") = 9/14 = 0.643

P(buy_computer = "No" = 5/14 = 0.357

Formula :

Naive Bayes classifier :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Given

$x = (x \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{Fair}, \text{buy\_computer} = ?)$

Compute $P(x/Ci)$ for each class

$P(\text{age} \leq 30 \mid \text{buy\_computer} = \text{yes}) = 2/9 = 0.222$

$P(\text{age} \leq 30 \mid \text{buy\_computer} = \text{no}) = 3/5 = 0.6$

$P(\text{income} = \text{medium} \mid \text{buy\_computer} = \text{yes}) = 4/9 = 0.444$

$P(\text{income\_medium} \mid \text{buy\_computer} = \text{no}) = 2/5 = 0.4$

$P(\text{Student} = \text{yes} \mid \text{buy\_computer} = \text{yes}) = 6/9 = 0.667$

$P(\text{Student} = \text{yes} \mid \text{buy\_computer} = \text{no}) = 1/5 = 0.5$

$P(\text{credit\_rating} = \text{Fair} \mid \text{buy\_computer} = \text{yes}) = 6/9 = 0.667$

$P(\text{credit\_rating} = \text{Fair} \mid \text{buy\_computer} = \text{no}) = 2/5 = 0.4$

Now

$P(x/Ci)$:

$P(x \mid \text{buy\_computer} = \text{"yes"})$

$= 0.222 \times 0.444 \times 0.667 \times 0.667$

$= 0.044$

$P(x \mid \text{buy\_computer} = \text{No})$

$= 0.6 \times 0.4 \times 0.2 \times 0.4$

$= 0.019$

$P(x/Ci) P(Ci)$: For $P(\text{buy\_computer} = \text{yes})$

$P(x \mid \text{buy\_computer} = \text{yes}) * P(\text{buy\_computer} = \text{yes})$

$= 0.044 * 0.643 = 0.028$

For buy_computer = No

$P(x \mid \text{buy\_computer} = \text{No}) * P(\text{buy\_computer} = \text{no})$

$= 0.019 * 0.357 = 0.007$

$\therefore$ X belong to yes class to buy_computer.

Assignment No.:_____ Page No.:_____
Subject :_____
Roll No.:_____

(11)

Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

Q-2 Consider the following data of approved_loan Classand classify a tuple
x (age=young, has_job=false, own_house=false, credit_rating =good, class=?)

ID	Age	has_job	own_house	Creditrating	Class
1	Young	F	F	Fair	No
2	Young	F	F	Good	No
3	Young	T	F	Good	Yes
4	Young	T	T	Fair	Yes
5	Young	F	F	Fair	No
6	Mid	F	F	Fair	No
7	Mid	F	F	Good	No
8	Mid	T	T	Good	Yes
9	Mid	F	T	Excellent	Yes
10	Mid	F	T	Excellent	Yes
11	Old	F	T	Excellent	Yes
12	Old	F	T	Good	Yes
13	Old	T	F	Good	Yes
14	Old	T	F	Excellent	Yes
15	Old	F	F	Fair	No

P(Ci):
P(loan_approved = Yes) = 6/15 = 0.4
P(loan_approved = No) = 9/15 = 0.6

Formula:
Naive Bayes classifier:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

12

Given that

X (age = young, has_job = false, own_house = false credit_rating = good, class = ?)

Compute P(x|Ci) for each class

P(age = young | approved_loan = yes) = 2/6 = 0.33
P(age = young | approved_loan = No) = 4/9 = 0.44
P(has_job = false | approved_loan = yes) = 4/6 = 0.67
P(has_job = false | approved_loan = No) = 6/9 = 0.67
P(own_house = false | approved_loan = yes) = 3/6 = 0.5
P(own_house = false | approved_loan = No) = 6/9 = 0.67
P(credit_rating = good | approved_loan = yes) = 4/6 = 0.67
P(credit_rating = good | approved_loan = No) = 2/9 = 0.22

P(X | loan_approved = yes)
= 0.33 X 0.67 X 0.5 X 0.67
= 0.074

P(X | loan_approved = No) = 0.44 X 0.67 X 0.67 X 0.22
= 0.043

Now P(x|Ci) P(Ci)
For approved_loan = yes
P(X | approved_loan = yes) P(loan_approved = yes)
= 0.074 * 0.4 = 0.029
P(X | loan_approved = No) (loan_approved = No)
= 0.043 * 0.6
= 0.026

∴ X belongs to class = yes as P(X|Ci) of loan_approved loan is higher than loan_approved = No.

Q-3 With the help of the training sample given below using Bayes classification and can we believe that the patient having the given symptoms has Flu.

x {Chill=yes, Rainy_nose = No, Headache=Mild, Fever=Yes, Flu=?}

Chills	Rainy-nose	Headache	Fever	Flu
Yes	No	Mild	Yes	No
Yes	Yes	no	no	Yes
Yes	No	Strong	Yes	yes
No	Yes	mild	Yes	Yes
No	No	no	no	No
No	Yes	Strong	Yes	Yes
No	Yes	Strong	no	No
Yes	Yes	mild	Yes	Yes

Formula :

Naive Bayes classifier

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(Ci) :

P(Flu=Yes) = 5/8 = 0.625

P(Flu=No) = 3/8 = 0.375

Calculate P(x|ci) for each class

Given that

x {chill=yes, rainy_nose=no, headache=mild, fever=yes, Flu=?}

(14)

P(Chill=yes | flu=yes) = 3/5 = 0.6
P(chill=yes | flu=no) = 1/3 = 0.33
P(rainy.nose=yes | flu=yes) = 1/5 = 0.2
P(rainy-nose=yes | flu=yes) = 2/3 = 0.66
P(headache=mild | flu=yes) = 2/5 = 0.4
P(headache=mild | flu=no) = 1/3 = 0.33
P(fever=yes | flu=yes) = 4/5 = 0.8
P(fever=yes | flu=no) = 1/3 = 0.33

P(x | Cl):

P(x | flu=yes) = 0.6 × 0.2 × 0.4 × 0.8 = 0.0384
P(x | flu=No) = 0.33 × 0.66 × 0.33 × 0.33 = 0.0241

P(x|Cl) P(Cl)

P(x | flu=yes) P(flu=yes) = 0.038 × 0.625 = 0.0234
P(x | flu=no) P(flu=no) = 0.024 × 0.375 = 0.009

∴ With the given x symptoms patient belong to class yes of flu.

(15)

Decision Tree (ID3)

Q-1 Construct Decision Tree based on ID3 for the following training data.

Day	Outlook	Temp	Humidity	Wind	Decision (Play tennis)
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Formula :

$$\text{Gini index} = 1 - \sum (P_i)^2 \quad \ldots (1)$$

For $i = 0$ to number of classes

Step 1 : Find the Gini index for each feature.

Outlook :

	No. of Instances	Yes	No	Gini Index	Gini Index of Outlook
Sunny	5	2	3	0.48	
Overcast	4	4	0	0	0.342
Rain	5	3	2	0.48	

Gini(Outlook = Sunny) = $1-(2/5)^2-(3/5)^2 = 0.48$

Gini (Outlook = Overcast) = $1-(4/4)^2-(0/4)^2 = 0$

Gini (Outlook = Rain) = $1-(3/5)^2-(2/5)^2 = 0.48$

Gini (Outlook) = $\frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$

∴ Gini(Outlook) = 0.342

Temperature	No of Instance	Yes	No	Gini Index	Gini for Temperature
Hot	4	2	2	0.5	
Mild	6	4	2	0.445	0.440
Cool	4	3	1	0.375	

Gini (temperature = hot) = $1-(2/4)^2-(2/4)^2 = 0.5$

Gini (temperature = mild) = $1-(4/6)^2-(2/6)^2 = 0.445$

Gini (temperature = cool) = $1-(3/4)^2-(1/4)^2 = 0.375$

Gini (temperature) = $\frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.445 + \frac{4}{14} \times 0.375$

∴ Gini (Temperature) = 0.440

Humidity	No of Instances	Yes	No	Gini Index	Gini Index for humidity
High	7	4	3	0.489	0.367
Normal	7	6	1	0.245	

Gini (humidity = high) = $1-(4/7)^2-(3/7)^2 = 0.489$

Gini (humidity = normal) = $1-(6/7)^2-(1/7)^2 = 0.245$

Gini (Humidity) = $\frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.245$

∴ Gini (Humidity) = 0.367

Wind

	No. of Instance	Yes	No	Gini Index	Gini Index for wind
Strong	6	3	3	0.375	0.428
Weak	8	6	2	0.5	

$Gini(wind = strong) = 1 - (3/6)^2 - (3/6)^2 = 0.375$

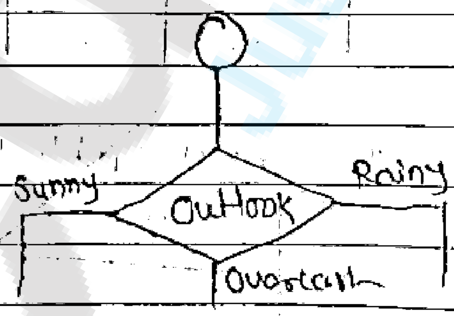
$Gini(wind = weak) = 1 - (6/8)^2 - (2/8)^2 = 0.5$

$Gini(Wind) = \frac{6}{14} \times 0.375 + \frac{8}{14} \times 0.5$

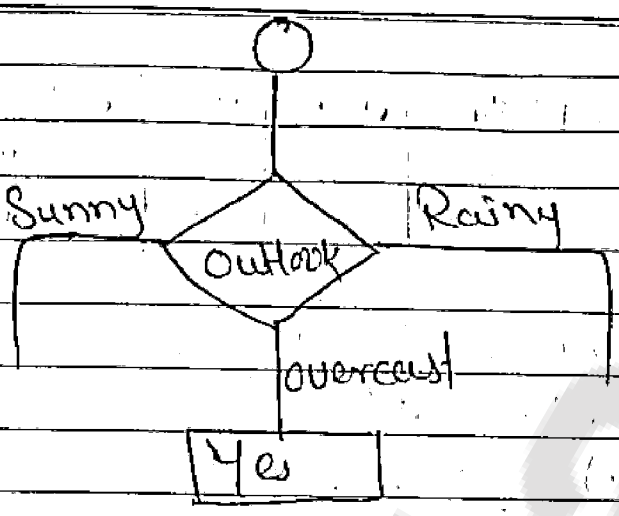
$\therefore Gini(Wind) = 0.428$

Step 2: Choose the lowest gini value feature as winner.

Feature	Gini Index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428



From the given data, overcast decision is always yes in any condition.
Hence the next feature will be overcast and decision will be "yes".



Step Now we have to find the decision for the Sunny outlook with respect to feature

Sunny Outlook

Temperature		No. of Instance	Yes	No	Gini of Instance	Gini of Sunny temperature
Gini coef	hot	2	0	2	0	
Temperature	cool	1	1	0	0	0.2
	mild	2	1	1	0.5	
Humidity	High	3	0	3	0	0
	Normal	2	2	0	0	
Wind	Weak	3	1	2	2.66	0.466
	Strong	2	1	1	0.2	

Decision for sunny outlook

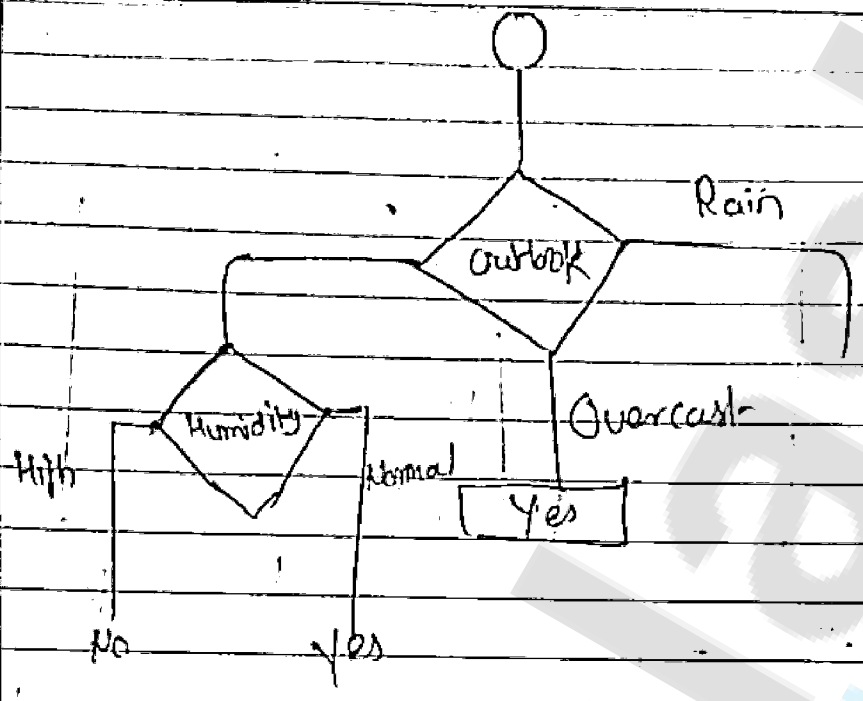
Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

Humidity win the sunny outlook

Assignment No.: _____ Page No.: ____
Subject : _____
Roll No.: _____

(14)

Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL



Rain outlook		yes	no	No. of Instance	Gini index of Instance	Gini best of feature
Temperature	hot					
	Cool	1	1	2	0.5	0.466
	mild	2	1	3	0.444	
Humidity	High	1	1	2	0.5	0.466
	Normal	2	1	3	0.444	
Wind	Weak	3	0	3	0	0
	Strong	0	2	2	0	

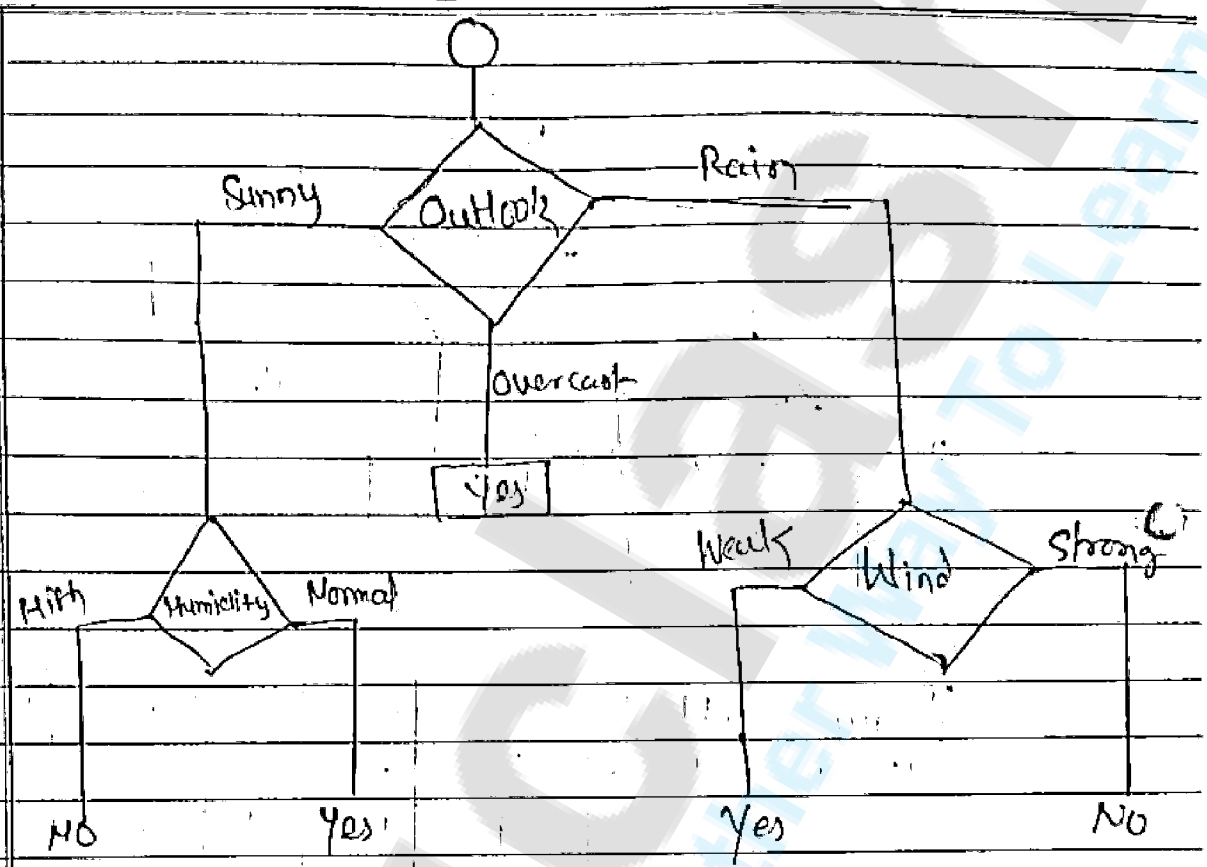
Decision For Rain outlook feature

Feature	Gini Index
Temperature	0.466
Humidity	0.466
Wind	0

For rain outlook the wind feature is important attribute.

NCRD's
Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

2.0



Q-2 Generate the decision tree for following dataset

No	Significant	Reorganised	Volume	Class
1	Yes	No	Small	Bad
2	Yes	Yes	large	Bad
3	No	Yes	medium	Bad
4	No	Yes	Medium	Good
5	Yes	No	large	Good
6	No	No	large	Good

Formula :

Gini Index $= 1 - \Sigma (Pi)^2$

For P no number of classes

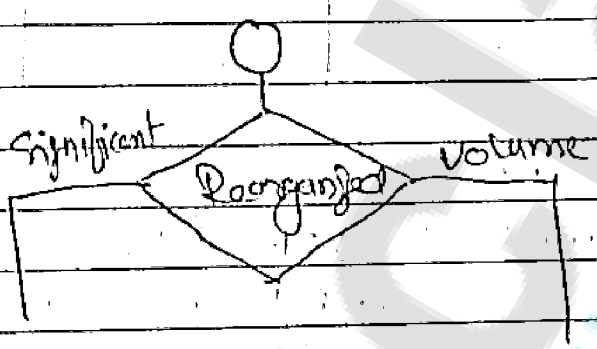
Find the gini index for each feature

Feature	No. of Instance	Good	Bad	Gini of instance	Gini of Feature
1. Significant					
Yes	3	1	2	0.444	0.444
No	3	2	1	0.444	
Reorganised					
Yes	2	0	2	0	0.25
No	4	3	1	0.375	
Volume					
Small	1	0	1	0	0.388
Mid	2	1	1	0.5	
Large	3	2	1	0.444	

9.2

Decision For Feature

Feature	Gini Index
Significant	0.444
Reorganized	0.25
Volume	0.388

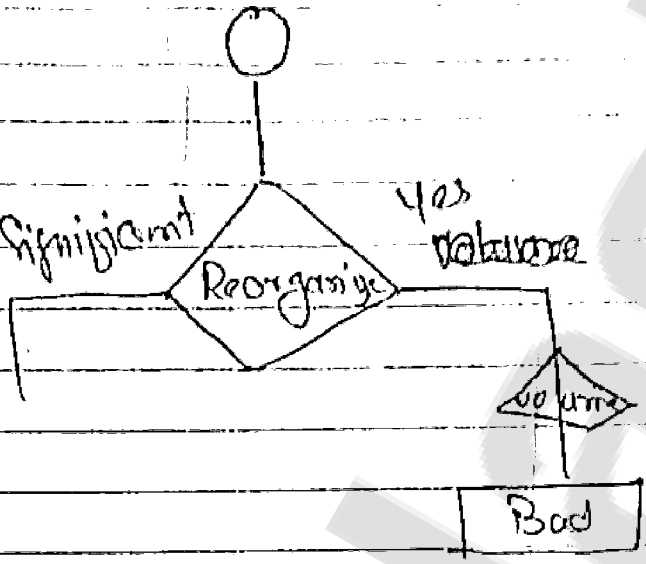


Yes reorganized

Feature	No. of	Instance	Good	Bad	Gini	Gini of
Significant						
yes	3	5	1	2		
no	3		1	2		
Volume						
small	1	5	0	1	0	0
medium	1		0	1	0	
large	1		0	1	0	

It clearly given that for yes reorganized decision is bad.

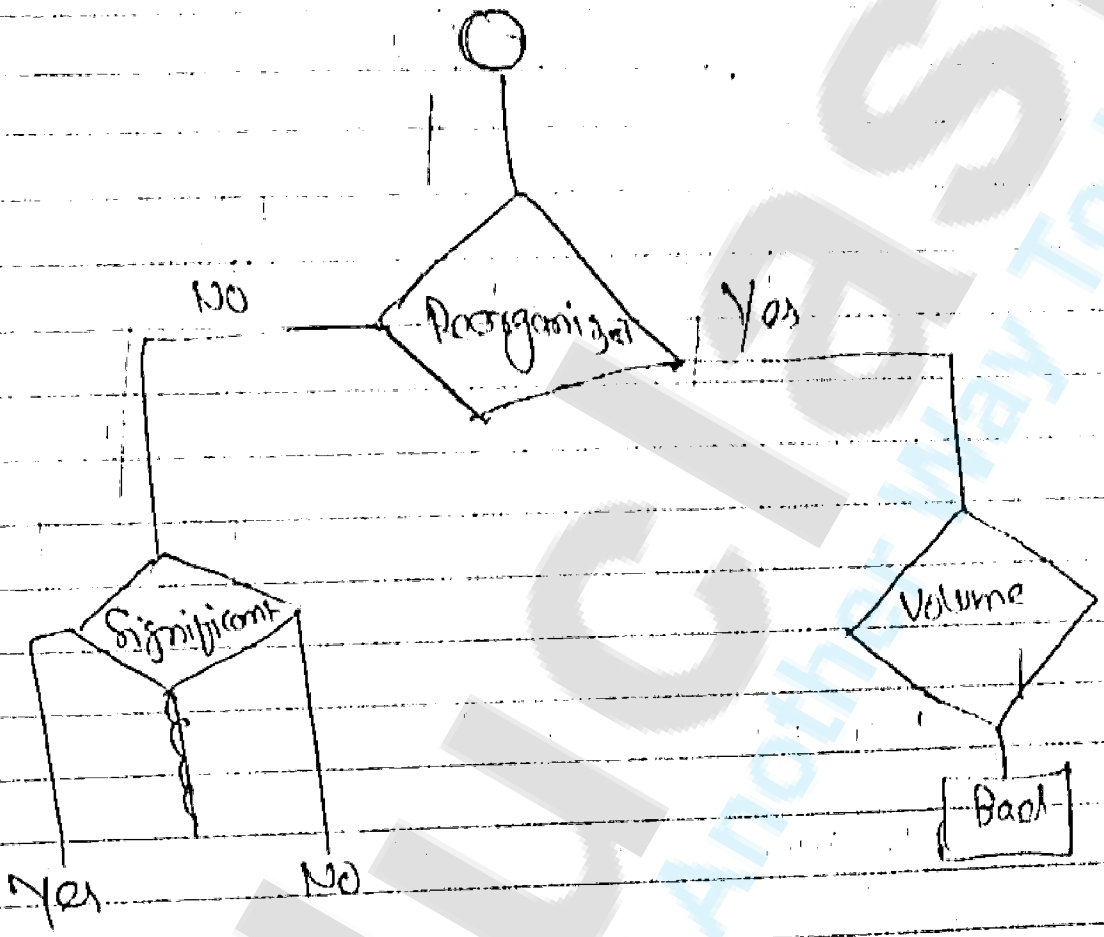
(23)



No reorganized

Feature	No. of Instance	Good	Bad	Gini of Instance	Gini of Feature
Significant					
Yes	2	1	1	0	0
No	2	2	0	0	
Volume					
Small	1	0	1	0	0
medium	1	0	1	0	
large	2	2	0	0	

Q 8)



Assignment No : Page No
Subject :
Roll No.:

05

Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

Q-3 Construct a decision tree based on ID3 for the following training data

	Chill	Rainy-nose	Headache	Power	Flu
1	Yes	No	Mild	Yes	No
2	Yes	Yes	No	No	Yes
3	Yes	No	Strong	Yes	Yes
4	No	Yes	Mild	Yes	Yes
5	No	No	No	No	No
6	No	Yes	Strong	Yes	Yes
7	No	No	Strong	No	No
8	Yes	Yes	Mild	Yes	Yes

Formula:

Gini Index = $1 - (\sum P_i)^2$

Chill

	No. of Instance	Yes	No	Gini Index	Gini Index for Chill
Yes	4	3	1	0.375	0.9375
No	4	2	2	0.5	

Rainy-nose

	No. of Instance	Yes	No	Gini Index	Gini Index for Rainy nose
No	3	2	1	0.444	
Yes	5	4	1	0.32	0.366

Headache

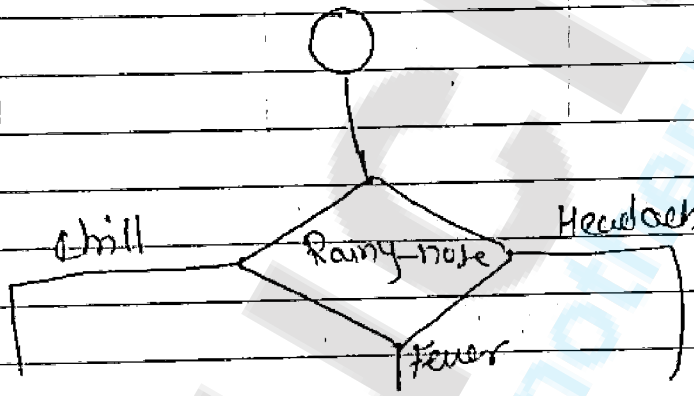
	No. of Instance	Yes	No	Gini Index	Gini for Headache
No	2	1	1	0.5	
Mild	3	2	1	0.444	0.458
Strong	3	2	1	0.444	

26

Fever

	No. of Instance	Yes	No	Gini Index	Gini Index for Fever
Yes	5	4	1	0.32	0.366
No	3		2	0.444	

Feature	Gini Index
Chill	0.4735
Rainy_nose	0.366
Headache	0.458
Fever	0.366

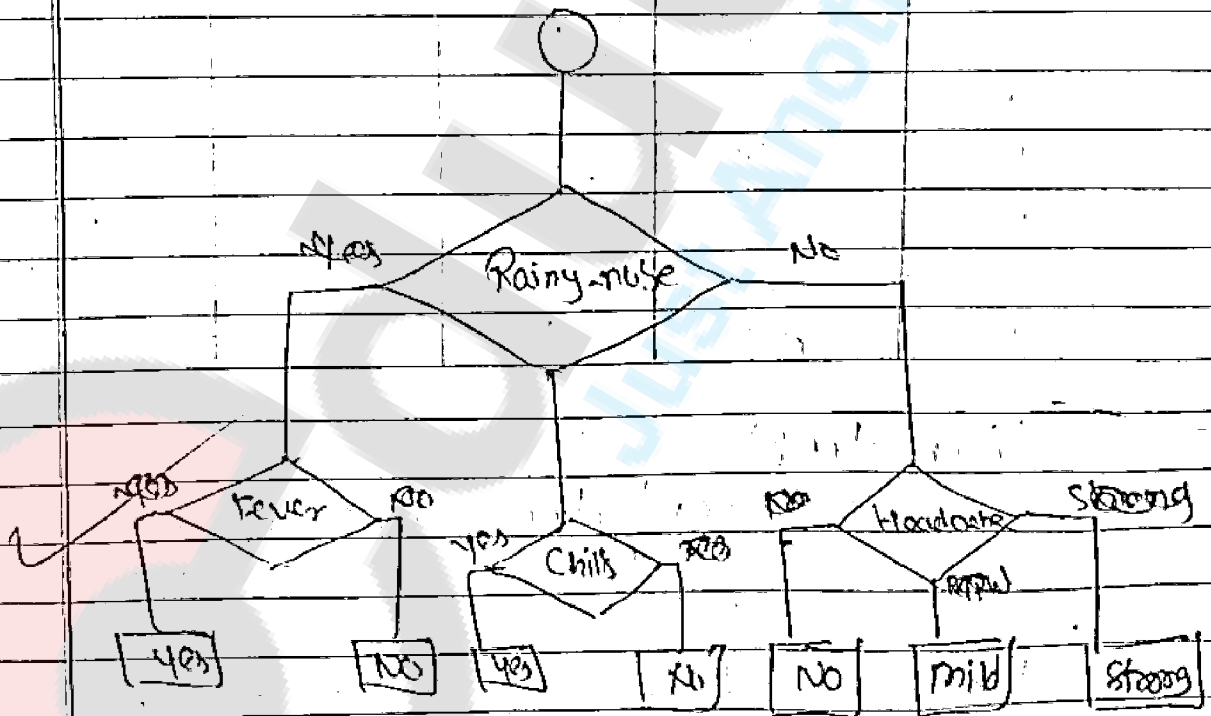


Rainy Nose = Yes

Feature	No of Instance	Yes	No	Gini of Instance	Gini of Feature
Chill					
Yes	2	2	0	0	0.266
No	3	2	1	0.444	
Headache					
No	1	1	0	0	
Mild	2	2	0	0	0.2
Strong	2	1	1	0.5	
Fever					
Yes	3	3	0	0	0.2
No	2	1	1	0.5	

Rainy rose = no

Feature	No. of Instance	Yes	No	Gini of Instance	Gini of Feature
Chill					
Yes	2	1	1	0.5	0.33
No	1	0	1	0	
Headache					
No	1	0	1	0	0
Mild	1	0	1	0	
Strong	1	1	0	0	
Fever					
Yes	2	1	1	0.5	0.33
No	1	0	1	0	



032

Q-4 Construct a decision tree based on ID3 for the following training set

ID	Age	has_job	own_house	credit_rating	Class
1	Young	F	F	Fair	No
2	Young	F	F	Good	No
3	Young	T	F	Good	Yes
4	Young	T	T	Fair	Yes
5	Young	F	F	Fair	No
6	Mid	F	F	Fair	No
7	Mid	F	F	Good	No
8	Mid	T	T	Good	Yes
9	Mid	F	T	Excellent	Yes
10	Mid	F	T	Excellent	Yes
11	old	F	T	Excellent	Yes
12	old	F	T	Good	Yes
13	old	T	F	Good	Yes
14	old	T	F	Excellent	Yes
15	old	F	F	Fair	No

Formula:

Gini Index = $1 - \sum(P_i)^2$

for i=1 to number of classes

Find Gini index for each feature.

Assignment No.:_____ Page No.:_____
Subject :_____
Roll No.:_____

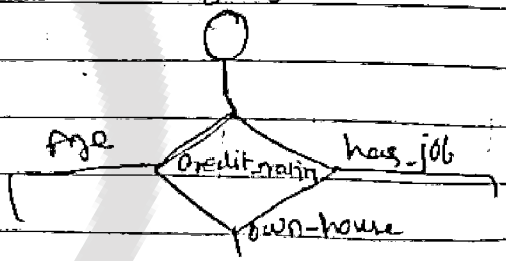
2d

NCRD's Sterling INSTITUTE OF MANAGEMENT STUDIES, NERUL

Feature	No. of Instance	Class: Yes	No	Gini Index of class	
Age					
young	5	2	3	0.48	
Mid	5	3	2	0.48	0.426
Old	5	3	1	0.32	
has_job					
F	10	4	6	0.48	0.32
T	5	5	0	0	
own_house					
F	9	3	6	0.44	0.264
T	6	6	0	0	
Credit_rating					
Fair	5	1	4	0.32	0.177
Good	6	4	2	0.44	
Excellent	4	4	0	0	

Feature	Gini Index
Age	0.426
has_job	0.32
Own_house	0.264
Credit_rating	0.177

As Credit_rating has lowest Gini Index it a root main feature in given data.



(30)

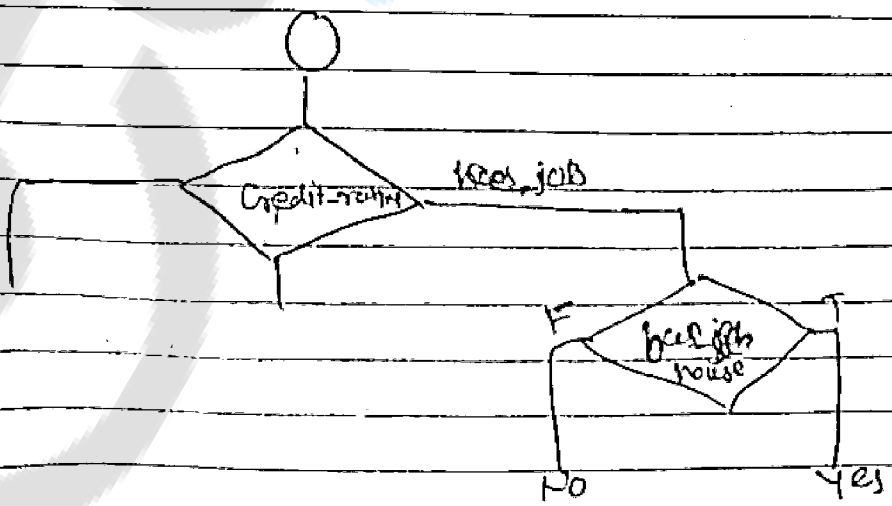
fair Credit rating

Feature	No. of instance	Yes	No	Gini index for instance	Feature Gini Index
Age					
Young	3	1	2	0.44	
mid	1	0	1	0	0.264
Old	1	0	1	0	
has_job					
False	4	0	4	0	0
True	1	1	0	0	
Own_house					
F	4	0	4	0	0
T	1	1	0	0	

→ Decision

Feature	Gini Index
Age	0.264
has_job	0
own_hous	0

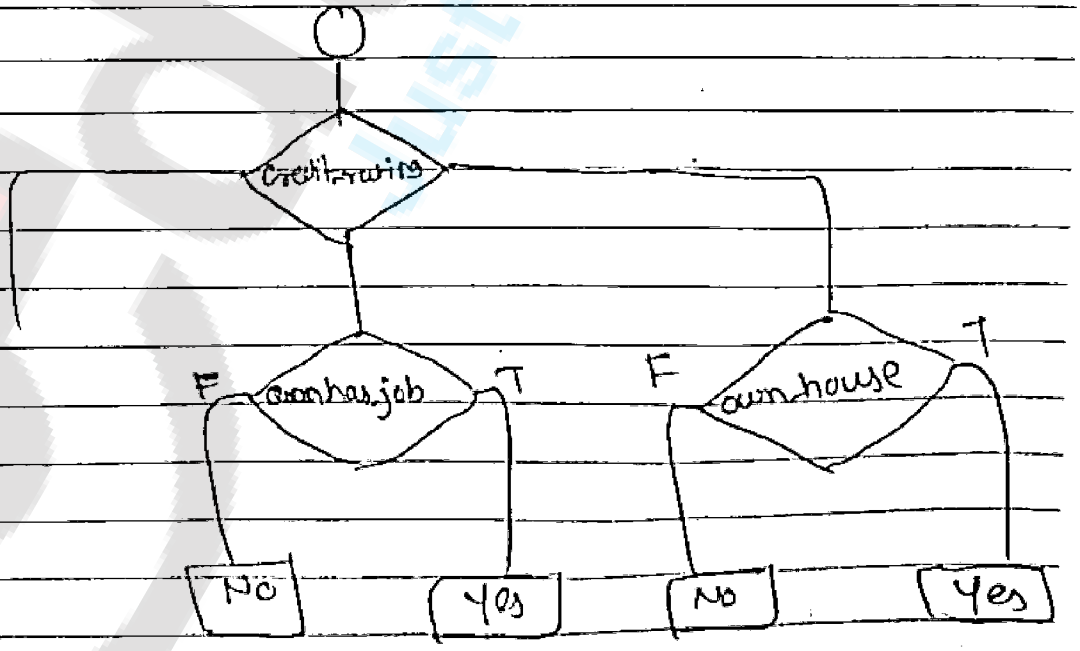
has_job and ~~own_house~~ win for Credit rating
we will chose own house here



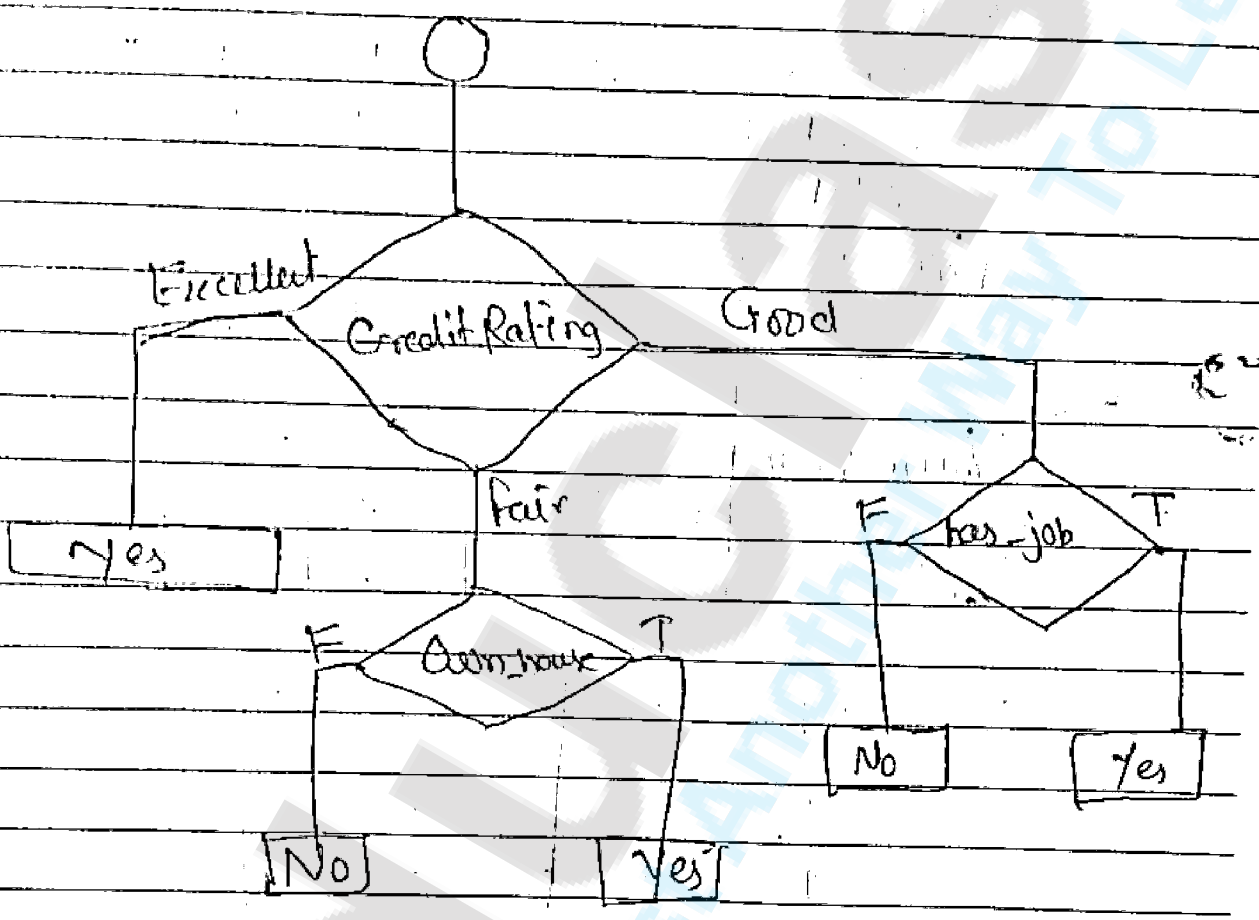
Good credit rating

Feature	No. of Instance	Yes	No	Gini index	Gini Index for Feature
Age					
young	2	1	1	0.5	
mid	2	1	1	0.5	0.33
old	2	2	0	0	
has_job					
F	3	1	2	0.44	0.22
T	3	3	0	0	
own_house					
F	4	2	2	0.5	0.33
T	2	2	0	0	

Feature	Gini Index
Age	0.33
has_job	0.22
own_house	0.33



For excellent credit rating it yes for each condition and/or feature.



Assignment No. Page No.
Subject
Roll No.

(B3)

K-mean clustering

Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

Q.1 What is clustering? Explain k-means clustering algorithm. Using kmeans clustering, cluster the following data into two cluster and show each step.

{2, 4, 10, 12, 3, 20, 30, 11, 25}

Clustering :- Clustering is a machine learning technique that involves the grouping of data point. Given a set of data points we can use a clustering algorithm to classify each data point into a specific group. In theory, data points are in the same group should have similar properties and/or features while data points in different groups should have highly dissimilar properties and/or features.

Clustering is a feature of unsupervised learning. It is a common technique for statistical data analysis used in many fields.

For Data Science, we can use clustering analysis to gain some valuable insight from our data by seeing what groups the data points fall into when we apply a clustering algorithm. There are 5 popular clustering algorithm:

1. K-mean clustering
2. Mean-shift clustering
3. Density-Based Spatial clusking of Applicwith Noise (DBSCAN)
4. Expectation-maximization (EM) cluster using Gaussian Mixture Model (GMM)
5. Agglomerative Hierarchical Clustering

(B4)

K-mean clustering algorithm:-

(1) Randomly/given select cluster centres.

(2) Calculate the differentce between each data point and cluster centre

(3) Assign data point to the cluster whose distance from the cluster centre is minimum from all the cluster centre.

(4) Recalculate new cluster center

$$V_i = \left(\frac{1}{|c_i|}\right) \sum_{j=1}^{c_i} (x_j)$$

(5) Recalculate the distance between each data point and obtain new cluster centre.

(6) If no data point was reassigned then stop other-wise repeat step 3.

K-mean clustering is pretty fast.
It has a od linear complexity O(n).

Q-8 Perform k-mean clustering for given data set
{2, 3, 4, 10, 11, 12, 20, 25, 30}

→ Let $m_1 = 4$ and $m_2 = 12$
and $K = 2$
$k_1 = \{2, 3, 4\}$
$k_2 = \{10, 11, 12, 20, 25, 30\}$

$m_1 = \frac{9}{3} = 3$ $\quad m_2 = \frac{108}{6} = 18$

$\therefore m_1 = 3$ and $m_2 = 18$

→ New cluster
$k_1 = \{2, 3, 4, 10\}$
$k_2 = \{11, 12, 20, 25, 30\}$

Mean(m_1) for $k_1 = 4.75 \cong 5$
mean(m_2) for $k_2 = 19.6 \cong 20$

→ New cluster
$k_1 = \{2, 3, 4, 10, 11, 12\}$
$k_2 = \{20, 25, 30\}$

mean (m_1) for $k_1 = 7$
mean (m_2) for $k_2 = 25$

→ New cluster
$k_1 = \{2, 3, 4, 10, 11, 12\}$
$k_2 = \{20, 25, 30\}$
mean(m_1) for $k_1 = 7$
mean(m_2) for $k_2 = 25$

Final cluster will be

Cluster (k_1)	Cluster (k_2)
2	20
3	25
4	30
10	
11	
12	

Step used in this numerical:

Step1 : Choose random cluster and initialize.
Step2 : Take a mean value.
Step3 : Find nearest number to mean and put in cluster.
Step4 : Repeat step (2) and step (3) until you get same mean value and / or you covered all the datapoint.
Step5 : Stop.

Q-21 Explain k means clustering algorithm. Perform k means clustering using Euclidean distance measure for the given data set (k = 2).

A	1.0	1.5	3.0	5.0	3.5	4.5	3.5
B	1.0	2.0	4.0	7.0	5.0	5.0	4.5

k-means clustering algorithm:

(1) Randomly select cluster centres.

(2) Calculate the distance between each data point and cluster centre.

(3) Assign data point to the cluster whose distance from the cluster centre is minimum from all the cluster centre.

(4) Recalculate the cluster centre.

$$V_i = \left(\frac{1}{U_i}\right) \sum_{j=1}^{c} (x_j)$$

(5) Recalculate the distance between each data point and obtain new cluster centre.

(6) If no data point was reassigned then stop otherwise repeat step 3.

ID	X	Y
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

C_1 = Cluster 1 = {1, 2, 3} Centroid for C_1 = [(1,1), (1.25, 1.5), (2, 1.4, 2.5)]

C_2 = Cluster 2 = {4, 5, 6, 7}

Centroid for C_2 = [(5,7), (4.25, 6), (4.375, 5.5), (3.937, 5)]

Formula:

Euclidean Distance = $\sqrt{(x_x - x_1)^2 + (y_y - y_1)^2}$

Initialize the cluster

Cluster	X	Y	Centroid value
C_1	1	1	
C_2	5	7	

→ For Row 2 using Euclidean Formula

For row we have to find Euclidean distance with respect to both cluster 1 and cluster 2

ED for C_1 = $\sqrt{(1.5-1)^2 + (2-1)^2}$

= $\sqrt{1.25}$

= 1.11

ED for C_2 = $\sqrt{(1.5-5)^2 + (2-7)^2}$

= $\sqrt{37.25}$

= 6.10

Row 2 goes in C_1.

∴ New centroid value of C_1 = $\left(\frac{1+1.5}{2}, \frac{1+2}{2}\right)$

New centroid value of C_1 = $[1.25, 1.5]$

→ For Row 3, using Euclidean formula

ED for C_1 = $\sqrt{(3-1.25)^2 + (4-1.5)^2}$

= $\sqrt{3.0625 + 6.25}$ = $\sqrt{9.3125}$

= 3.05

ED for C_2 = $\sqrt{(3-5)^2 + (4-7)^2}$

= 3.60

Row 3 goes in cluster 1 (C_1).

New centroid for C_1 = $\left(\frac{1.25+3}{2}, \frac{1.5+4}{2}\right)$

= $(4.25/2), (5.5/2)$

∴ New centroid for C_1 = $(2.12, 2.75)$

→ For Row 5 using Euclidean formula

ED for C_1 = $\sqrt{(3.5-2.12)^2 + (5-2.75)^2}$

= 2.690

ED for C_2 = $\sqrt{(3.5-4)^2 + (5-5.5)^2}$

= 2.500

Row 5 goes in cluster 2.

New centroid for $C_2 = \left(\frac{3.5+5}{2}, \frac{7+5}{2}\right)$

$\therefore$ Centroid for $C_2 = (4.25, 6)$

→ for Row 6, using Euclidean distance formula

ED for $C_1 = \sqrt{(4.5-2.14)^2 + (5-2.75)^2}$
$= 3.260$

ED for $C_2 = \sqrt{(4.5-4.25)^2 + (5-6)^2}$
$= 1.030$

Row 6 goes in cluster 2.

New centroid $= \left(\frac{4.25+4.5}{2}, \frac{6+5}{2}\right)$

New centroid for $C_2 = (4.375, 5.5)$

→ For Row 7, using euclidean formula

ED for $C_1 = \sqrt{(3.5-2.14)^2 + (4.5-2.75)^2}$
$= 2.216$

ED for $C_2 = \sqrt{(3.5-4.375)^2 + (4.5-5.5)^2}$
$= 1.328$

Row 7 goes in cluster 2.

New centroid for $C_2 = \left(\frac{4.375+3.5}{2}, \frac{5.5+4.5}{2}\right)$
$= (3.937, 5)$.

Assignment No.:_____ Page No.:_____
Subject :_____
Roll No.:_____

(41)

NCRD's Sterling
INSTITUTE OF MANAGEMENT STUDIES, NERUL

→

ID	X	Y
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

ID	cluster 1			Centroid
1	1	1	1	(1, 1)
2	1.5	2	2	(1.25, 1.5)
3	3	4	4	(2.14, 2.75)

ID	cluster 2			Centroid
4	5	7		(5, 7)
5	3.5	5		(4.25, 6)
6	4.5	5		(4.375, 5.5)
7	3.5	4.5		(3.937, 5)

Solution
V. good.

Q-3 Explain the k-means clustering and write the algorithm. Find Two clusters of the following data set.

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

K-mean clustering algorithm:

(1) Randomly select cluster centres.

(2) Calculate the difference between each data point and cluster centre.

(3) Assign data point to the cluster whose distance from the cluster centre is minimum from all the cluster centre.

(4) Recalculate the cluster centre.

$$Vi = (1/ui) \sum_{j=1} (xi)$$

(5) Recalcute the distance between each data point and obtain new cluster centre.

(6) If no data point was reassigned then stop otherwise repeat step 3.

Given that $k = 2$

Let A and C are cluster 1 and cluster 2 respectively.

Given data set

P	X	Y
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

Formula :

Euclidean Distance (ED) $= \sqrt{(x_x - x_1)^2 + (y_y - y_1)^2}$

Ans ⟹

Cluster $k_1 = \{A, B\}$

Cluster $k_2 = \{C, D, E\}$

Centroid For $k_1 = \{(1,1), (1, 0.5)\}$

Centroid For $k_2 = \{(0,2), (1,3)\}$

For Row B :

ED for cluster $k_1 = \sqrt{(1-1)^2 + (0-1)^2}$

$= 1$

ED for $k_2 = \sqrt{(1-0)^2 + (2-2)^2}$

$= 1$

Row B goes in cluster k_1.

New Centroid for $k_1 = \left(\frac{1+1}{2}, \left(\frac{1+0}{2}\right)\right)$

Centroid Row $k_1 = (1, 0.5)$

→ For Row D using Euclidean Distance Formula

ED for $K_1 = \sqrt{(2-1)^2+(4-0.5)^2}$

$= 3.640$

ED for $K_2 = \sqrt{(2-0)^2+(4-2)^2}$

$= 2.828$

Row D goes in cluster K_2.

New centroid for $K_2 = \left(\frac{0+2}{2}, \frac{2+4}{2}\right)$

∴ New centroid for $K_2 = (1, 3)$

→ For Row E using Euclidean Distance formula

ED for $K_1 = \sqrt{(3-1)^2+(5-0.5)^2}$

$= 4.924$

ED for $K_2 = \sqrt{(3-1)^2+(5-3)^2}$

$= 2.828$

Row E goes in cluster K_2.

Given Data			Cluster 1				Cluster 2			
P	x	y	P	x	y	Centroid	P	x	y	Centroid
A	1	1	A	1	1	(1,1)	C	0	2	(0,2)
B	1	0	B	1	0	(1,0.5)	D	2	4	(1,3)
C	0	2					E	3	5	
D	2	4								
E	3	5								