Popular Attribute Selection Measures are:-

1. Information Gain
2. Gini Index.
3. Gain Ratio.

Information Gain:-

$$P_i = |C_{i,D}| / |D|$$

Expected Information:-

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

Information Needed:-

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

Info. Gain :-

$$Gain(A) = Info(D) - Info_A(D).$$

Step1:- Entropy of Entire Dataset (yes → 9, No → 5)

$$S\{+9, -5\} = -9/14 \log_2 9/14 - 5/14 \log_2 5/14 = 0.94$$

Step2:- Entropy of All Attributes. (weather)

Entropy of sunny $\{+2, -3\} = -2/5 \log_2[2/5] - 3/5 \log_2[3/5]$
$= 0.97$

Entropy of ~~Rainy~~ cloudy $\{+4, -0\} = -4/4 \log_2[4/4] - 0/4 \log_2(0)$
$= 0$

Entropy of Rainy $\{+3, -2\} = -3/5 \log_2[3/5] - 2/5 \log_2[2/5]$
$= 0.97$

∴ Info Gain = Entropy(whole data) $- \frac{5}{14}$ Entro(S) $- \frac{4}{14} E(C) - \frac{5}{14} E($
$= 0.246.$

Page No.
Date

Calculate IG of Temp.

Sol:- Entropy of Entire Dataset

$S[9+, 5-] = 0.94$

Sol:- Entropy of Hot $[2+, 2-] = -2/4 \log_2 2/4 - 2/4 \log_2 2/4 = 1$

Entropy of Mild $[4+, 2-] = -4/6 \log_2 4/6 - 2/6 \log_2 2/6 = 0.91$

Entropy of Cold $[3+, 1-] = -3/4 \log_2 3/4 - 1/3 \log_2 1/3 = 0.81$

$\therefore$ Info Gain = E(whole) $- 4/14$ E(H) $- 6/14$ E(M) $- 4/14$ E(C)

$= 0.029$

Humidity

Calculate IG of ~~wind~~ :-

Sol:- E(H) $[3+, 4-] = 0.98$

E(N) $[6+, 1-] = 0.59$

IG = $0.94 - 7/14$ E(H) $- 7/14$ E(N)

$= 0.15$

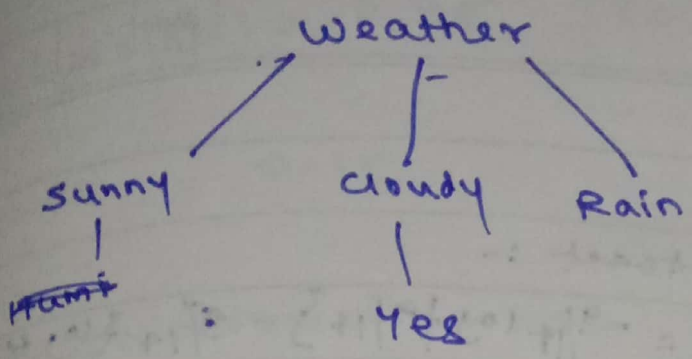
IG of Wind:-

Sol:- E(S) = 1.0

E(N) = 0.81

$\therefore$ 0.0478.

$\therefore$ Greatest Entropy is Weather



Now, Again for sunny for entire table.

Sol:- Entropy of sunny [+2, -3] = $-2/5 \log 2/5 - 3/5 \log_2 3/5$

= 0.97

IG of Temp:-

E(Hot) = 0

E(Mild) = 1.0

E(cold) = 0

∴ Info Gain = Entropy(sunny) $- \frac{2}{5}$ Ent(H) $- \frac{2}{5}$ Ent(M) $- \frac{1}{5}$ Ent(C)

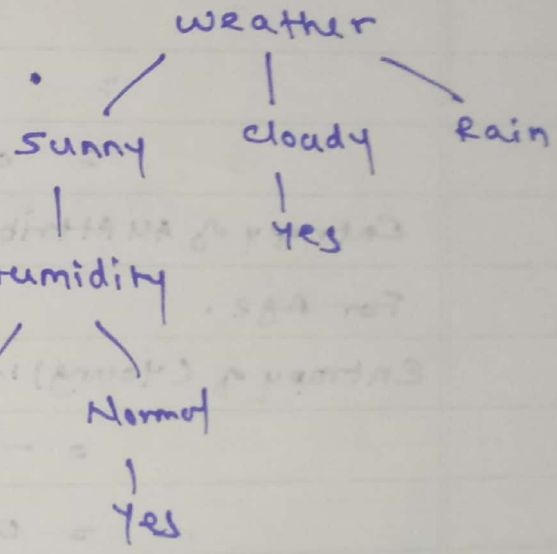
= 0.57

IG of Humidity

E(H) = 0

E(N) = 0

∴ Info Gain = 0.97



IG of wind:-

Entropy (S) = 1

E(W) = 0.918

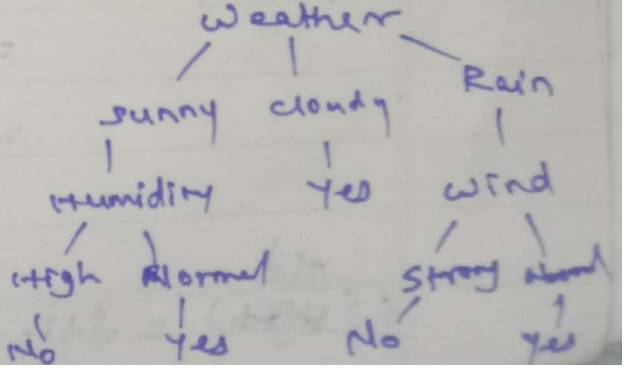
IG = 0.019

Now calculate Again for Rain:-

IG of temp = 0.019

IG of Humidity = 0.019

IG of wind = 0.97



Gain Ratio.

Gain for Entire Dataset :-

$S\{+9, -5\} = -9/14 \log\{9/14\} - 5/14 \{\log 4/14\}$

$= 0.94$

In. Gain for Income.

$0.94 - 1.557$

Class → Owns a House → Yes
→ Rented

Gain for Entire Dataset

$S\{+7, -5\} = -7/12 \log_2 7/12 - 5/12 \log_2 5/12$

$= 0.45 + 0.52$

$= 0.9798$

Entropy of All Attributes :-

For Age.

Entropy of (Young) :- $S\{+3, 0\}$

$= -3/4 \log_2[3/4] - 1/3 \log_2[1/4]$

$= 0.5$ ~~$+ 0.321$~~ $= -0.81$

Entropy of (Medium) = $S\{+3, -2\}$

$= -3/5 \log_2[3/5] - 2/5 \log_2\{2/5\}$

$= 0.970 + 0.302$

0.814

$= 1.115$

Old

$E(\text{High}) 2\{+2, +\} = -2/3 \log(1/3) - 1/3 \log_2(1/3)$

Info Gain (Age) = $0.97 - \frac{4}{12}(0.52) - \frac{5}{12}(1.115) - \frac{3}{12}(0.600)$

$= 0.15$

Info Gain for Income:-

very High:- $S(+2, 0)$.

$= -2/2 \log_2(2/2) - 0/2 \log_2(0/2)$

$= 0$

High :- $S\{4, 0\}$

$= -4/4 \log_2(4/4) - 0/4 \log_2(0/4)$

$= 0$

Medium:- $S\{+1, -2\}$

$= -1/3 \log_2(1/3) - 2/3 \log_2(2/3)$

$= 0.52 + 0.389$

$= 0.609$

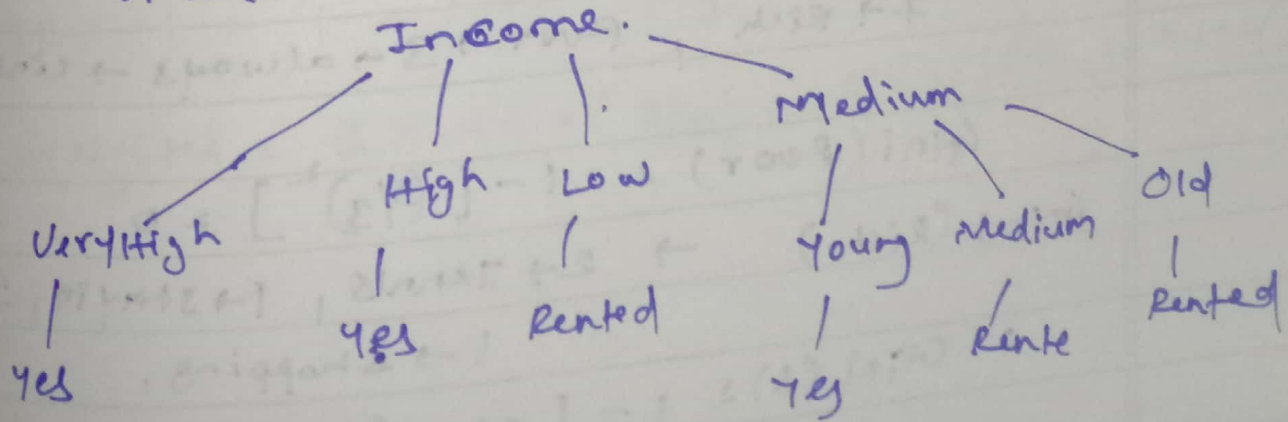
Low:- $S\{+1, -2\}$

$= 0.$

Info (Income) = $0.97 - 3/12 (0.609)$

$= 0.78.$

∴ Income.



Gini Index

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Compute Gini Indix for overall collection.

Cinema → 6, Tennis → 2, Stay in → 1, Shopping → 1

$$Gini(S) = 1 - [(6/10)^2 + (2/10)^2 + (1/10)^2 + (1/10)^2]$$

$$= 0.58.$$

Gini for Money Attribute.

7 → Rich, Poor → 3 → always → cinema

$$Gini(Poor) = 1 - [(3/3)^2] = 0$$

Gini (Rich) → 2 → Tennis, 1 → Stay in, 3 → Cinema, 1 → Shopping.

$$\therefore Gini(S) = 1 - [(2/7)^2 + (3/7)^2 + (1/7)^2 + (1/7)^2]$$

$$= 0.694$$

Weighted Avg = $0 * (3/10) + 0.694 * (7/10) = 0.486$

Gini for parents

Yes → 5, No → 5

For Yes 5 → Cinema

For No, → Tennis = 2, 1 → Staying
1 → Shopping
1 → Cinema

Gini (Parents = Yes) = $1 - [(5/5)^2]$

= 0

Gini (Parents = No) = $1 - [(1/5)^2 + (1/5)^2 + (1/5)^2 + (1/5)^2]$

= 0.72

weighted = $0 * (5/10) + 0.72 \times 5/10$

$\boxed{= 0.35}$ ← since, minimum max info.

Gini for weather :-

Sunny , Windy , Rainy.

3

2 → Cinema, 1 → Tennis

5 4

3rd Cinema 1 → Shopping

3

2 → Cinema 1 → Stay in

Gini (Sunny) = $1 - [(2/3)^2 + (1/3)^2] = 0.44$

Gini (Cinema) = $1 - [(3/4)^2 + (1/4)^2] = 0.375$

Gini (Rainy) = $1 - [(2/3)^2 + ((1/3)^2] = 0.44$.

∴ weighted = $0.44 * (3/10) + 0.375 \times (4/10) + 0.44 * (3/10)$

= 0.416

Parents

Yes → Cinema

No

W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Now for weather

Sunny → 2, Windy → 2,
Rainy → 1

Gini(Sunny) = $1 - [(1/2)^2]$

= 0

Gini (Rainy) = 0

Gini (Windy) = $1 - [(1/2)^2 + (1/2)^2]$

= 0.5

Weighted = $0 + 0 + 0.5 * 2/5$

= 0.2

For Money

Rich → 4, Poor → 1

Gini (Rich) = $1 - [(\frac{2}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{4})^2]$

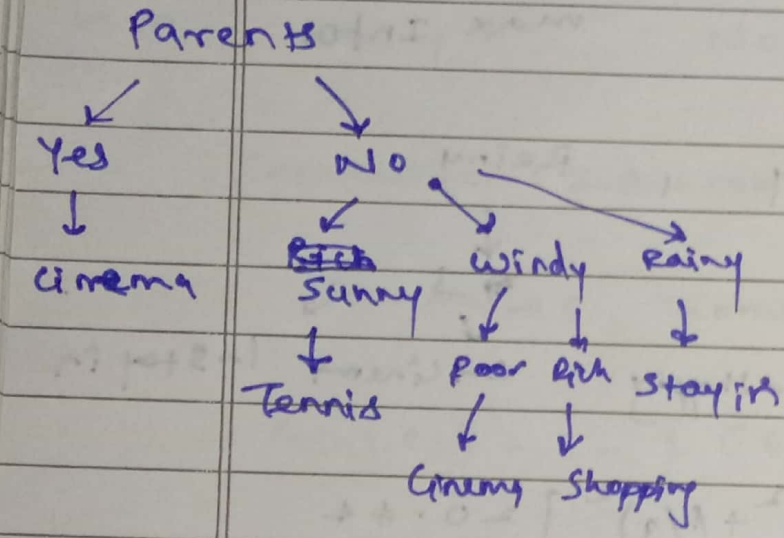
= 0.625

Gini Poor = $1 - [(1/1)^2]$

= 0

∴ Weighted = $0 * \frac{1}{1} + 0.625 (4/5)$

= 0.5



Naive-Bayes Classification :-

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

Find Y if probability of X is given.

$$P(Y|X_1, X_2, \cdots, X_n) = \frac{P(X_1|Y) * P(X_2|Y) * P(X_3|Y) \cdots P(X_n|Y) * P(Y)}{P(X_1) * P(X_2) * P(X_3) \cdots P(X_n)}$$

Person	covid (Yes/No)	flu (Yes/No)	fever (Yes/No)
1	Yes	No	Yes
2	No	Yes	Yes
3	Yes	Yes	Yes
4	No	No	No
5	Yes	No	Yes
6	No	No	Yes
7	Yes	No	Yes
8	Yes	No	No
9	No	Yes	Yes
10	No	Yes	No

Given Person (flu, covid)

$P(Yes | (flu, covid)) = P((flu|Yes) * P(covid|Yes) * P(Yes)$ → $4/7 * 3/7 * 7/10 \rightarrow 12/70 \rightarrow 0.17$

$P(No | (flu, covid)) = P(flu|No) * P(covid|No) * P(No)$ → $2/3 * 2/3 * 3/10 \rightarrow 0.13$

S1:- $P(fever = yes) = 7/10$

$P(fever = No) = 3/10$

S2:- Conditional prob.

	Yes	No
covid	4/7	2/3
flu	3/7	2/3

Play-tennis

Outlook	Temp	Humidity	windy	class
sunny	Hot	High	False	N
sunny	Hot	High	True	N
overcast	Hot	High	False	P
rain	mild	High	False	P
rain	cool	~~High~~ Normal	False	P
rain	cool	Normal	True	N
overcast	cool	Normal	True	P
sunny	Mild	~~Normal~~ High	False	N
sunny	cool	~~High~~ Normal	False	P
rain	Mild	Normal	False	P
sunny	Mild	Normal	True	P
overcast	Mild	~~Normal~~ High	True	P
overcast	Hot	~~Hi~~ Normal	False	P
rain	Mild	High	True	N

S:- P(C=P) = 9/14

P(C=N) = 5/14

	yes	No
Hot	2/9	2/5
Mild	4/9	2/5
cool	3/9	1/5

Outlook

	P	N
Sunny	2/9	3/5
Overcast	4/9	0
Rain	3/9	2/5

Humidity

	P	N
High	3/9	4/5
Normal	6/9	1/5

Windy

	P	N
False	6/9	2/5
True	3/9	3/5

$P(X|Y)$

$$\therefore \frac{2}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{2}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{3}{9} \times \frac{9}{14}$$

$$\frac{2}{9} \times \frac{4}{9} \times \frac{1}{9} \times \frac{2}{9} \times \frac{2}{9} \times \frac{4}{9} \times \frac{2}{9} \times \frac{1}{14}$$

$$\frac{2}{9} \times \frac{4}{9} \times \frac{1}{9} \times \frac{2}{9} \times \frac{2}{9} \times \frac{4}{9} \times \frac{2}{9} \times \frac{1}{14}$$

$X = \{rain, hot, high, false\}$

$$\frac{3}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{9}{14}$$

$$\frac{2}{9} \times \frac{6}{9} \times \frac{1}{14}$$

$$\frac{12}{504} \rightarrow 0.023$$

$$\frac{2}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{4}{5} \times \frac{5}{14}$$

$$\frac{320}{8750} \rightarrow 0.036$$

Hence, X is classified with class N.

Metrics for Evaluating Classifier Performance.

TP (True +ve) :- refer to tuples that were correctly labelled by classifier

TN True -ve (~~False +ve~~) :- refer to ~~False~~ -ve tuples that were correctly labelled by classifier.

FP (False +ve) :- These are -ve tuples that were incorrectly labelled

~~FN~~ FN (False -ve) :- These are +ve tuples that were incorrectly labelled.

$$\text{Accuracy} = \frac{TP + TN}{P+N}$$

$$\text{error rate} = \frac{FP + FN}{P+N}$$

$$\text{recall, TPR} = \frac{TP}{P}$$

$$\text{TNR, specificity} = \frac{TN}{N}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

predicted / actual	yes	No	Total
yes	6954	46	7000
No	412	2588	3000
Total	7366	2634	10000

Accuracy = $\frac{6954 + 2588}{10000}$ = 0.9552 %

Error = 0.0458

Sensitivity =

Modl-2

Date

1. Nominal :- categories, states or "names of things"
   → The numbers are not intended to be used quantitatively
   → Also, values do not have any specific meaningful order.

2. Binary Attributes → Two values

3. Ordinal Attributes → Values have meaningful order or a ranking among them but magnitude is not known.

4. Numeric Attributes:-
   ↳ (4.1) Interval Scaled Attributes
   (No true zero point)
   ↳ (4.2) Ratio scaled Attributes
   (Inherent zero point)

5. Discrete Attributes :-
   Has only a finite or a countably infinite set of values which may or may not be represented as integers.

6. Continous Attribute :-
   Practically, real values can only be measured and represented using finite number of digits.

3 Areas of Statistical Descriptions :-
→ Measuring central Tendencies → Mean, Median, Mode, Midrange
→ Measuring dispersion.
→ Graphic display of data.

Suppose

$x = \{2, 4, 4, 4, 5, 5, 7, 9\}$

$\therefore$ Mean = 5

$$\text{variance} = \frac{(2-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2}{8}$$

$$= \frac{32}{8} = 4$$

$\therefore$ SD $= \sqrt{\text{var}}$

$= 2$

Proximity Measure for Nominal Attribute.

Object	test-1 (Nominal)
1	code A
2	code B
3	code C
4	code A

$$d(i,j) = \frac{p - m}{p}$$

p → total attributes describing objects

m → total no. of matches

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$d(2,1) = \frac{1-0}{1} = 1$$

$$d(3,1) = \frac{1-0}{1} = 1$$

$$d(3,2) = \frac{1-0}{1} = 1$$

$$d(4,1) = \frac{2-2}{2} = 0$$

$$d(4,2) = \frac{1-0}{1} = 1$$

$$d(4,3) = \frac{1-0}{1} = 1$$

$$\begin{bmatrix} 0 & & & \\ \bullet & & & \\ & & & \\ & & & \end{bmatrix}$$

Promixity measures for binary Attributes.

A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

for symmetric binary variable $d(i,j) = \frac{r+s}{q+r+s+t}$

for assymmetric $d(i,j) = \frac{r+s}{q+r+s}$

Name	Gender	Fever	cough	Test1	Test2	Test3	Test4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

* ~~only~~ asymmetric *
* ~~Let~~ Gender is symmetric attribute
Rest all Assymmetric
* ~~Gender~~ Let Y and P be 1 and N=0
M=1, F=0

	Gender	Fever	cough	Test1	Test2	Test3	Test4
Jack	1	1	0	1	0	0	0
Mary	0	1	0	1	0	1	0
Jim	1	1	1	0	0	0	0

d(jack, mary) = q=2, r=0, s=1 = 1/3 = 0.33
d(jack, jim) = q=1, r=1, s=1 = 2/3 = 0.67
d(jim, mary) = q=1, r=1, s=2 = 3/4 = 0.75

Proximity Measures for Mixed Attributes.

No	Test I (Nominal)	Test II (Ordinal)	Test III (Numeric)
1	code A	Excellent	45
2	code B	Fair	22
3	code C	Good	64
4	Code A	Excellent	28

For Mixed Attributes calculate values for each and do the average.

For test I.

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

For test II.

3 → 1
2 → 0.33
1 → 0
3 → 1

$Z_{it} = (r_{it} - 1) / (M_{it} - 1)$

M_{it} → total no. of Attributes.

∴ $2/2 \rightarrow 1$

2 → 1

code A (1)	0			
code B (0)	1	0		
code C (0.5)	0.5	0.5	0	
code D (1)	0	1	0.5	0
	code A(1)	code B(0)	code C(0.5)	code D(1)
		0	0.5	1

For Test 3.

code A (45)	0			
code B (22)	23	0		
code C (84)	19	42	0	
code D (18)	127	6	36	0
	code A	code B	code C	code D
	45	22	84	18

Min = 6

Max = 42

Divide by 42 for Normalization.

code A	0			
code B	0.55	0		
code C	0.45	1.0	0	
code D	0.40	0.14	0.86	0

Why Data Preprocessing:-

i) incomplete
ii) inaccurate or noisy
iii) inconsistent

Well-accepted multidimensional view:-

i) Accuracy
ii) completeness
iii) consistency
iv) Timeliness
v) Believability
vi) Interpretability

ACC
TB Int

Major Tasks in Data Preprocessing:-

1) Data cleaning → fill in missing values
→ Identify outliers and smooth out noisy data
→ correct inconsistent data
→ Resolve redundancy caused by integration

2) Data Integration → Schema Integration

3) Data Transformation

4) Data Reduction

5) Data Discretization

Handle Noisy data (Binning method)
→ Smooth by bin mean
→ Smooth by bin median
→ Smooth by bin boundary

correlation analysis can be used to handle data redundancy.

Regression
• Linear regression. refers to finding the best line to fit b/w the two variables so that one can be used to predict other.
• Multiple reg:- involves more than two variable.

χ^2 Test

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- There's also something called as covariance of Numerical Data

$$Cov(A,B) = E(A \cdot B) - \bar{A} \cdot \bar{B}.$$

Explained in next page.

1) Dimensionality Red^n:-
i) Wavelet Transform
ii) Principal component
iii) Attribute selection

2) Numerosity Red^n
Data Cube Aggregation
Lossy

3) Discretization

Stock price for AllElectronics and HighTech.

Time Point	AllElectronics	HighTech.
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5
~~t6~~		

$$E(\text{AllElect}) = 6 + \frac{5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(\text{HighTech}) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80$$

$$\text{Cov}(A,B) = E(\bar{A}\cdot\bar{B}) - \bar{A}\cdot\bar{B}$$

$$= \frac{6\times 20 + 10\times 5 + 14\times 4 + 5\times 3 + 2\times 5}{5} - 4\times 10.80$$

$$= 50.2 - 43.2$$

$$= 7.$$

Since, covariance is +ve we can say that stock prices for both companies rise together.

Min-max Normalization.

$$v_i' = \frac{v_i - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A.$$

$$= \frac{200 - 200}{1000 - 200} (1 - 0) + 0$$

$$= 0.$$

For Data
200, 300, 400, 600, 1000.

0, 0.125, 0.25, 0.5, 1

$$= \frac{300 - 200}{800} (1) + 0$$

$$= \frac{100}{800} (1) + 0$$

$$= \frac{400 - 200}{800}$$

$$= \frac{200}{800} \frac{1}{4}$$

$$\frac{600 - 200}{800}$$

$$= 1.$$

0.12
8)1
-0
10
-8
20

0.2
4)1
-0
10

Naive Bayes Algorithm.

$$P(Y|X) = \frac{P(X|Y) \cdot P(X)}{P(Y)}$$

P(Default = No) =

Homeowner	Martial status	Job Experience	Defaulted
yes	Single	3.	No
No	Married	4.	No
No	Single	5.	No
yes	Married	4.	No
No	Divorced	2.	Yes
No	Married	4.	No
yes	Divorced	2.	No
No	Married	3.	Yes
No	Married	3.	No
yes	Single	2.	Yes

P(D = No) = 7/10

P(D = yes) = 3/10

	Yes	No
yes	1/3	3/7
No	2/3	4/7

	Yes	No
Married	1/3	4/7
Single	1/3	2/7
Divorced	1/3	1/7

	Yes	No
2	2/3	4/7
3	1/3	2/7
4	3/3	0/7
5	0/7	1/7

for <No, Married, 3> = 0.022

$2/3 \times 1/3 \times 1/3 \times 7/10$ (yes)

$4/7 \times 4/7 \times 2/7 \times 7/10$ (No) = 0.032

Gini Index

Gini Index for overall collection :-

$\Rightarrow 1 - [(5/14)^2 + (9/14)^2]$

$\rightarrow 0.459$

Gini Index for Age :- yes.

Youth → 5, middle aged :- 4, Senior → 5

Gini (Youth) = $1 - [(2/5)^2 + (3/5)^2]$

= 0.48

Gini (middle_aged) = 0

Gini (Senior) = $1 - [(3/5)^2 + (2/5)^2]$

= 0.48

wt. avg = $0.48 * 5/14 + 0.48 * 5/14$

= 0.342.

Gini Index Income :-

high → 4, medium → 6, low → 4

Gini (high) = $1 - [(2/4)^2 + (2/4)^2]$

= 0.5.

Gini (medium) = $1 - [(4/6)^2 + (2/6)^2]$

= 0.44

Gini (low) = $1 - [(3/4)^2 + (1/4)^2]$

= 0.375

wt avg = $0.5 \times 4/14 + 0.44 \times 6/14 + 0.375 \times 4/14 = 0.438$

student

No → 7 (Yes → 7

Gini (No) :- $1 - [(4/7)^2 + (3/7)^2]$

$= 0.489$

Gini (Yes) :- $1 - [(6/7)^2 + (1/7)^2]$

$= 0.244$

wt. avg : $0.489 \times 7/14 + 0.244 \times 7/14$

$= \frac{7}{14}[0.733]$

$= 0.3665$

credit Rating :-

fair :- 8 excellent → 6

Gini(fair) :- $1 - [(6/8)^2 + (2/8)^2]$

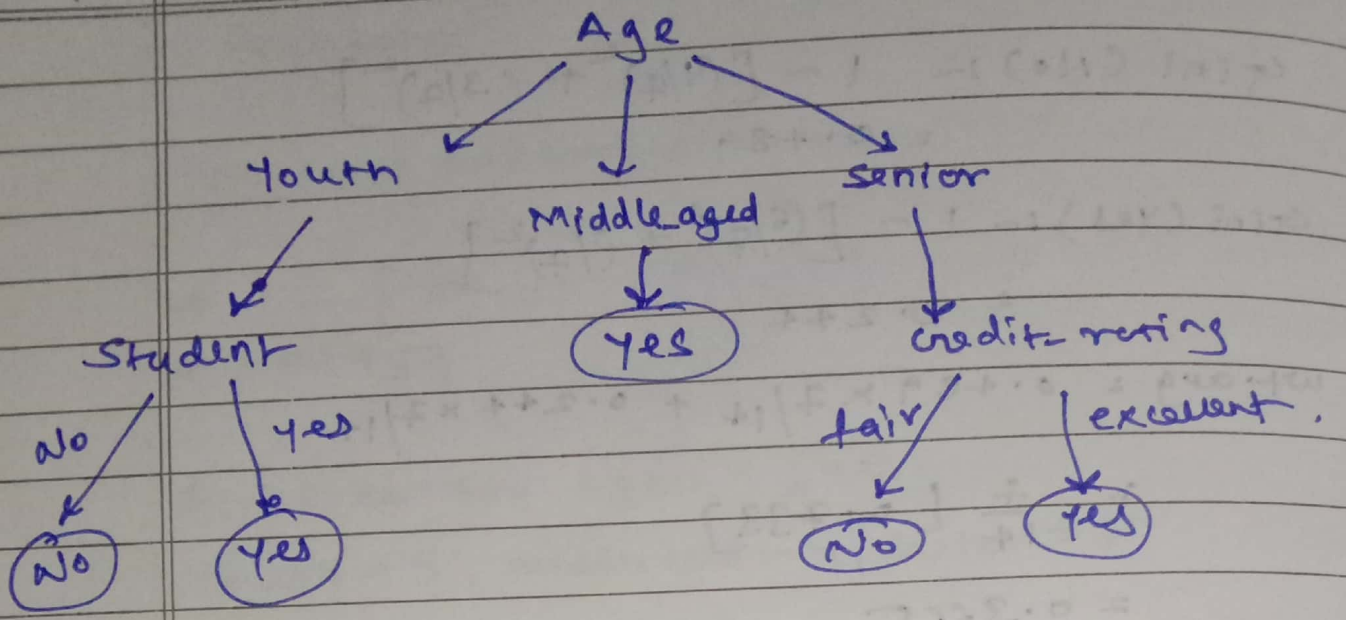
$= 0.375$

Gini excellent :- $1 - [(\frac{3}{6})^2 + (\frac{3}{6})^2]$

$= 0.5$

∴ wt. avg $= 0.375 \times \frac{8}{14} + 0.5 \times \frac{6}{14}$

$= 0.428$



	Income	Age	own_House
1.	very_High	young.	Yes
2.	High	Medium.	Yes
3.	Low	young.	Rented
4.	High	Medium.	Yes
5.	very_High	Medium.	Yes
6.	Medium	young.	Yes
7.	High	old.	Yes
8.	Medium	Medium	Rented
9.	Low	Medium	Rented
10.	Low	old	Rented
11.	High	young	Yes
12	Medium	old	Rented

Gini Index :- Yes :- 7 Rented :- 5

$1 - [(7/12)^2 + (5/12)^2] = 0.486$

For Income.

very High → 2, High :- 4, Medium → 3, low :- 3

Gini (very High) → 0 {1 - 1}

Gini (High) → 0

Gini (Medium) → $1 - [(1/3)^2 + (2/3)^2]$

$= 0.44$

Gini (Low) → 0.

wt. avg = $0.44 \times \frac{3}{12} = 0.11$

For Age.

Young → 4, Medium → 5, old → 3

Gini (Young) → $\{1 - [(3/4)^2 + (1/4)^2]$

$= 0.375$

Gini (Medium) → $1 - [(3/5)^2 + (2/5)^2]$ 0.4×5

$= 0.48$

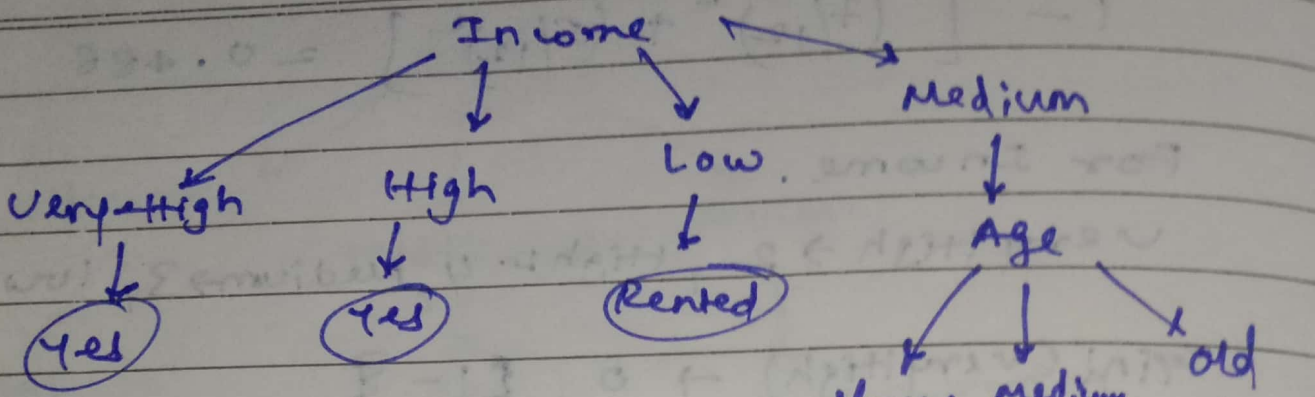
Gini (old) → $1 - [(1/3)^2 + (2/3)^2]$

→ 0.444.

$0.375 \times \frac{4}{12} + 0.48 \times \frac{5}{12} + 0.44 \times \frac{3}{12}$

$0.375 \times \frac{4}{12} + 2.0 + 0.11$

2.235.



Normalization
with z-score.