11

Data Warehouse and Mining
Assignment-1

P.S.V.P.P Lavanya
18JA1A0579
III CSE-B

1 a) Explain the steps involved in the Data Mining process with the sketch of the KDD process.

Data Mining refers to extracting or "Mining" Knowledge from large amounts of data.

Steps involved in the 'Data Mining process':

→ Data cleaning: to remove noise or irrelevant data.

→ data Integration: where multiple data sources may be combined

→ Data Selection: where data relevant to the Analysis task are retrieved from the database

→ Data transformation: where data are transformed or consolidated into forms.

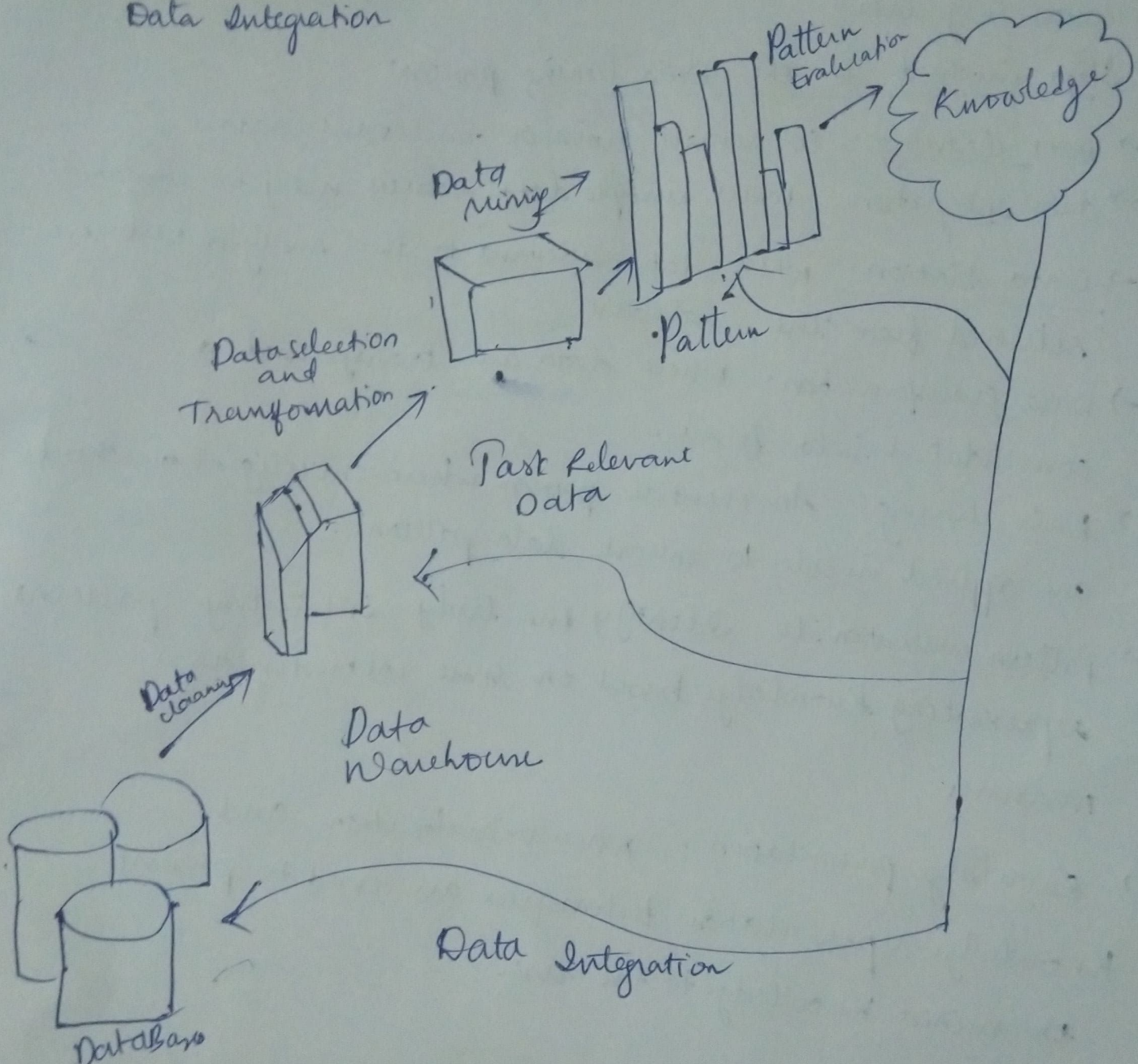
→ Data Mining: An essential process where intelligent methods are applied in order to extract data patterns.

→ Pattern evaluation: to identify the truly interesting patterns representing Knowledge based on some interestingness measures.

→ Knowledge presentation: — where visualization and Knowledge representation techniques are used to present the mined Knowledge to the user

KDP: It is an Iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed inorder to get different and more appropriate results.

→ preprocessing of Databases consists of Data cleaning and Data Integration



KDP Process

1b) A sample data containing Attributes of Mixed type.

Obj Id	Test-1 (Nominal)	Test-2 (Ordinal)	Test-3 (Numeric)
1	AA	Excellent	40
2	BB	fair	34
3	AA	excellent	45
4	CC	Good	65
5	BB	fair	36
6	DD	Bad	22

Construct dissimilarity and Similarity Matrix:

Ans. Dissimilarity Matrix of Nominal Attr:

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

where p: No. of attributes, m: no. of matches

Dissimilarity Matrix:

	1	2	3	4	5	6
1	0					
2	(2,1)	0				
3	(3,1)	(3,2)	0			
4	(4,1)	(4,2)	(4,3)	0		
5	(5,1)	(5,2)	(5,3)	(5,4)	0	
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	0

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

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

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

$d(5,1) = \frac{1-0}{1} = 1$

$d(6,1) = 1$

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

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

$d(5,2) = \frac{1-1}{1} = 0$

$d(6,2) = \frac{1-1}{1} = 1$

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

$d(5,3) = \frac{1-0}{1} = 1$

$d(6,3) = 1$

$d(5,4) = 1$

$d(6,4) = \frac{1-0}{1} = 1$

$d(6,5) = \frac{1-0}{1} = 1$

	1	2	3	4	5	6
1	0					
2	1	0				
3	0	1	0			
4	1	1	1	0		
5	1	0	1	1	0	
6	1	0	1	1	1	0

Ordinal Attributes:

To create Ordinal Matrix,

Step1: count states, There are 4 (Excellent, good, fair, bad).

Step2: Replace each Ordinal data of test 2 by Rank

fair=1 Bad=2 Good=3 Excellent=4

Step3: Normalize the Ranking $Zij = \frac{rij - 1}{mf - 1}$

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

Bad(2) = $\frac{2-1}{4-1} = \frac{1}{3} = 0.3$

Good(3) = $\frac{3-1}{4-1} = \frac{2}{3} = 0.6$

excellent(4) = $\frac{4-1}{4-1} = 1$

obj id	Test 2 (ordinal)
1	1
2	0
3	1
4	0.6
5	0
6	0.3

We make distance Matrix by using

$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$

Distance Matrix:

	1	2	3	4	5	6
1	0					
2	1	0				
3	0	1	0			
4	0.4	0.6	0.4	0		
5	1	0	1	0.6	0	
6	0.7	0.3	0.7	0.8	0.3	0

$d(2,1) = |0-1| = 1$ $d(3,2) = |1-0| = 1$

$d(3,1) = |1-1| = 0$ $d(4,3) = |0.6-1| = 0.4$

$d(5,1) = |0-1| = 1$ $d(4,2) = |0.6-0| = 0.6$

$d(5,4) = |0-0.6| = 0.6$ $d(6,4) = |0.3-0| = 0.3$

$d(5,3) = |0-1| = 1$ $d(5,1) = |0-1| = 1$

$d(5,2) = |0-0| = 0$ $d(6,3) = |0.3-1| = 0.7$

$d(5,6) = |0-0.3| = 0.3$

Now Numerical Attribute

Test Id	Test 3
1	70
2	34
3	75
4	65
5	36
6	22

formula:

$$d_{ij} = \frac{|x_i - x_j|}{max - min}$$

Here, max = 75

min = 22

Move to Normalise these rows so can be Mapped [0.0, 100]

$d(2,1) = \frac{|10-34|}{75-22}$

$= 0.64$

$d(3,1) = \frac{|70-75|}{75-22}$

$= 0.09$

$d(4,1) = \frac{|10-65|}{75-22} = 0.09$

$d(5,1) = \frac{|70-36|}{75-22}$

$= 0.641$

$d(6,1) = \frac{|70-22|}{75-22}$

$= 0.90$

$d(6,5) = \frac{|22-36|}{75-22}$

$= 0.20$

$d(3,2) = \frac{|75-34|}{75-22}$

$= 0.77$

$d(4,2) = \frac{|34-65|}{75-22}$

$= 0.58$

$d(5,2) = \frac{|34-36|}{75-22} = 0.03$

$d(6,2) = \frac{|34-22|}{75-22}$

$= 0.22$

$d(5,3) = \frac{|75-36|}{75-22}$

$= 0.73$

$d(6,4) = \frac{|22-65|}{75-22}$

$= 0.81$

$d(5,4) = \frac{|65-36|}{75-22}$

$= 0.54$

$d(6,3) = \frac{|75-22|}{75-22}$

$= 1$

	1	2	3	4	5	6
1	0					
2	0.64	0				
3	0.09	0.77	0			
4	0.09	0.58	0.18	0		
5	0.64	0.03	0.73	0.54	0	
6	0.90	0.22	1	0.81	0.26	0

Similarity Matrix:

The Matrix is formed by combining the attributes of dissimilar Matrices.

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij} d_{ij}}{\sum_{f=1}^{p} \delta_{ij}}$$

where

$\delta_{ij} = 0$ if x_{if} or x_{jf} missing

or $x_{if} = x_{jf}$ and

f is assentric

$\delta_{ij} = 1$ otherwise

$$d(2,1) = \frac{(1\times1) + (1\times1) + (1\times0.67)}{1+1+1}$$

$$= 0.89$$

$d(3,1) = 0.3$

$d(4,1) = 0.49$

$d(5,1) = 0.88$

$d(6,1) = 0.86$

$d(6,2) = 0.97$

$d(2,3) = 0.92$

$d(4,2) = 0.72$

$d(5,2) = 0.01$

$d(4,3) = 0.52$

$d(5,4) = 0.71$

$d(5,3) = 0.91$

$d(6,4) = 0.70$

$d(6,3) = 0.9$

$d(6,5) = 0.52$

Similarity Matrix:

	1	2	3	4	5	6
1	1					
2	0.87	0				
3	0.3	0.92	0			
4	0.49	0.76	0.52	0		
5	0.68	0.01	0.91	0.71	0	
6	0.82	0.17	0.9	0.70	0.52	0

Therefore for the given data with Mixed type Attributes

Dissimilarity Matrix

	1	2	3	4	5	6
1	0					
2	0.67	0				
3	0.09	0.77	0			
4	0.09	0.58	0.18	0		
5	0.64	0.03	0.73	0.54	0	
6	0.90	0.22	1	0.81	0.26	0

Similarity Matrix

	1	2	3	4	5	6
1	0					
2	0.89	0				
3	0.3	0.92	0			
4	0.49	0.76	0.52	0		
5	0.88	0.01	0.91	0.71	0	
6	0.86	0.17	0.9	0.70	0.52	0

2a) Data cleaning:

Data cleaning routines attempt to fill missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data

(i) Approaches to fill Missing values:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective.

(b) Manually filling the missing value: In general, this approach is time consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined

(c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as label like "Unknown", or $-\infty$. If missing values are replaced by, let "Unknown" then mining program things wrong. Although simple, it is not recommended.

(d) Using attribute mean for Quantitive (numeric) values or attributes mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple

(e) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction

(iii) Data Discretization:

Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals.

→ Interval labels can be used to replace actual data values.

Example:

Age: 10, 11, 13, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75

Table Before discretization:

Attribute	Age	Age	Age
	10, 11, 13, 14, 17, 19	30, 31, 32, 38,	70, 72, 73, 75
After discretization	Young	Mature	Old

Concept hierarchy Generation:

A concept hierarchy for a given numeric attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as, young, middle, or senior).

There are 5 methods for numeric concept hierarchy generation:

→ Binning
→ Histogram Analysis
→ Clustering Analysis
→ Entropy-based Discretion and
→ Data segmentation by "Natural Partitioning".

25) Interpret the below methods to Normalize the following data

data: 200, 300, 400, 600, 1000

(i) Min-max Normalization by setting min=0, max=1

min=0
max=1

V = value of Attribute

$V_1 = 200$ $V_2 = 300$ $V_3 = 400$ $V_4 = 600$ $V_5 = 1000$

new max = 1 , new min = 0

$$V' = \left\{ \frac{v - \min A}{\max A - \min A} (\text{new max}_A - \text{new min}_A) \right\} + \text{new min}_A$$

for 200: $\text{minmax} = \frac{200 - 200(1-0)}{1000 - 200} + 0 = 0.$

for 300: $\text{minmax} = \frac{300 - 200(1-0)}{1000 - 200} + 0 = \frac{100}{800} = 0.125$

for 400: $\text{minmax} = \frac{400 - 200(1-0)}{800} + 0 = 0.25$

for 600: $\text{minmax} = \frac{600 - 200(1-0)}{800} = \frac{400}{800} = 0.5$

for 1000: $\text{minmax} = \frac{1000 - 200(1-0)}{1000 - 200} + 0 = 1$

Original Data:	200	300	400	600	1000
Normalized data (min, max) (0, 1)	0	0.125	0.25	0.5	1

(iii) Z-Score Normalization:

Standard deviation = $\sqrt{\frac{\Sigma (\text{every value}) - (\text{mean of data})}{n}}$

Now Mean of Data = $\frac{200+300+400+600+1000}{5} = \frac{2500}{5} = 500$

∴ Standard deviation =

$$= \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (1000-500)^2 + (600-500)^2}{5}}$$

$$= \sqrt{\frac{10000+40000+10000+10000+25000}{5}}$$

$$= \sqrt{\frac{400000}{5}} = \sqrt{80000} = 282.8$$

Z Score of 200 = $\frac{x-M}{\sigma} = \frac{200-500}{282.8} = -1.06$

Z Score of 300 = $\frac{300-500}{282.8} = -0.7$

Z Score of 400 = $\frac{400-500}{282.8} = -0.35$

Z Score of 600 = $\frac{600-500}{282.8} = 0.35$

Z Score of 1000 = $\frac{1000-500}{282.8} = 1.78$

original data	200	300	400	600	1000
Z Score (Normalized data)	-1.06	-0.7	-0.35	0.35	1.78

(iii) Mean = $\frac{200+300+400+600+1000}{5} = 500$

Mean Absolute deviation : $\frac{1}{n}\sum_{i=1}^{n}|x_i - m(x)|$

Where

$M(X)$ = Average value of Dataset

n = no of datavalues

x_i = data values in set

$$= \frac{|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|}{5}$$

$$MAD = \frac{300+200+100+100+500}{5} = \frac{1200}{5} = 240$$

ZSCORE with MAD

for 200: $\frac{200-500}{240} = -1.25$

for 300 = $\frac{300-500}{240} = 0.83$

for 400 : $\frac{400-500}{240} = 2.45$

for 600 = $\frac{600-500}{240} = 0.41$

for 1000 = $\frac{1000-500}{240} = 2.08$

Original Data	200	300	400	600	1000
Normalized (Zscores)	-1.25	0.83	2.45	0.41	2.08

(iv) Decimal scaling: $v' = [v/10^j]$

where j = the smallest integer such that $\max(|v'|) < 1$.

$V = 200 \rightarrow [200/10^3] = 0.2$

$V = 300 \rightarrow [300/10^3] = 0.3$

$V = 400 \rightarrow [400/10^3] = 0.4$

$V = 600 \rightarrow [600/10^3] = 0.6$

$V = 1000 \rightarrow [1000/10^3] = 1$

Original data	200	300	400	600	1000
Normalized data	0.2	0.3	0.4	0.6	1

3a) List the differences between classification and prediction

3a)

Classification	Prediction
1) Classification is the process of identifying to which category, a new observation belongs to on the basis of a training data set containing observations whose category membership is known	1) Prediction is the process of identifying the missing or unavailable numerical data for a new observation
→ In classification, the accuracy depends on finding the label correctly	→ In prediction, accuracy depends on how well a given predictor can guess the value of predicted attribute for new data
→ A model or the classifier is constructed to find the categorical labels.	→ A model or a predictor will be constructed that predicts a continuous-valued function or ordered value
→ In classification, the model can be known as the classifier	→ In prediction, model can be known as the predictor
→ Classification is mostly based on our current or past assumptions	→ Prediction is based on our future assumptions

3b) Make use of the training examples shown in below table for a Binary classification problem to compute below Operations.

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a) What is the entropy of this collection of training examples with respect to the class Attribute?

$$\text{Entropy} = -\sum_{i=0}^{t-1} p(i|t) \log_2 p(i|t)$$

$$= -\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$$

$$= -\left[(-0.51997) + (-0.47111)\right]$$

$$= \underline{0.99107}$$

b) What are the Information gains of a_1 and a_2 relative to these training examples

$$\text{Entropy}(a_1) = -\sum_{i=0}^{t-1} p(i|t) \log_2 p(i|t)$$

$$= -\left[\left(\frac{3}{4}\right) * \log_2\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right) * \log_2\left(\frac{1}{4}\right)\right]$$

$$= -\left[(-0.31128) + (-0.5)\right]$$

$$= \underline{0.81128}$$

Entropy (a_2) $= -\left[\left(\frac{2}{5}\right)*\log_2\left(\frac{2}{5}\right)+\left(\frac{3}{5}\right)*\log_2\left(\frac{3}{5}\right)\right]$

$= -[(-0.52877)+(-0.44218)]$

$= 0.97095$

Entropy (a_1) $= -\left[\left(\frac{1}{5}\right)*\log_2\left(\frac{1}{5}\right)+\left(\frac{4}{5}\right)*\log_2\left(\frac{4}{5}\right)\right]$

$= -[(-0.46439)+(-0.25754)]$

$= 0.72193$

Entropy (a_3) $= -\left[\left(\frac{2}{4}\right)*\log_2\left(\frac{2}{4}\right)+\left(\frac{2}{4}\right)*\log_2\left(\frac{2}{4}\right)\right]$

$= -[(-0.5)+(-0.5)] = 1$

Information gain

a_1: $\Delta = I(\text{parent}) - \sum_{j=1}^{k}\frac{N(v_j)}{N} I(v_j) =$

$= 0.991 - \left[\left(\frac{4}{9}\right)*(0.81128)+\left(\frac{5}{9}\right)*(0.72193)\right] = 0.229$

a_2: $0.991 - \left[\left(\frac{5}{9}\right)*(0.971)+\left(\frac{4}{9}\right)*(1)\right] = 0.007$

c) For a_3, which is a continuous Attribute, Compute Information gain for every possible split

sorted values	1		3		4		5		6		7		8			
Split positions	0.5		2		3.5		4.5		5.5		6.5		7.5		8.5	
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
−	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
Gain	0		0.143		0.002488		0.0724		0.00973		0.01523		0.10211		0	

Split number1

Entropy(t) = $-\left[\left(\frac{4}{9}\right)*\log_2\left(\frac{4}{9}\right)+\left(\frac{5}{9}\right)\log_2\left(\frac{5}{9}\right)\right]$

$= -[-0.51997) + (-0.47111)]$

$= 0.99107$

Information Gain [Δ=0]

Split Number2

≤ entropy (t) = $-\left[\left(\frac{1}{1}\right)*\log_2\left(\frac{1}{1}\right)+0*\log_2(0)\right] = 0$

> entropy (t) = $-\left[\left(\frac{3}{8}\right)*\log_2\left(\frac{3}{8}\right)+\left(\frac{5}{8}\right)*\log_2\left(\frac{3}{8}\right)\right]$

$= -[(-0.53064)+(-0.42379)]$

$= 0.95443$

Weighted average: $\left[\left(\frac{1}{9}\right)*0\right]+\left[\left(\frac{8}{9}\right)*0.95443\right] = 0.84899$

Information Gain: Δ = 0.991 - (0.8439) = [0.143] = Δ

Split Number 3

≤ entropy(t) = $-\left[\left(\frac{1}{2}\right)*\log_2\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)*\log_2\left(\frac{1}{2}\right)\right]$

$= -[(-0.5)+(-0.5)] = 1$

> entropy(t) = $-\left[\left(\frac{3}{7}\right)*\log_2\left(\frac{3}{7}\right)+\left(\frac{4}{7}\right)*\log_2\left(\frac{4}{7}\right)\right]$

$= -[(-0.52388)+(-0.46135)]$

$= 0.98523$

Weighted Average = $\left[\left(\frac{2}{9}\right)*1\right]+\left[\left(\frac{7}{9}\right)*(0.98523)\right]$

$= 0.988512$

Information gain:

∴ Δ = 0.991 - (0.544407) = 0.002488

Split number 4

≤ • entropy(t) = $-\left[\left(\frac{2}{3}\right) * \log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right) * \log_2\left(\frac{1}{3}\right)\right]$

$= -[(-0.38998) + (-0.52832)]$

$= 0.9183$

> entropy(t) = $-\left[\left(\frac{2}{6}\right) * \log_2\left(\frac{2}{6}\right) + \left(\frac{4}{6}\right) * \log_2\left(\frac{4}{6}\right)\right]$

$= -[(-0.52388) + (-0.38998)]$

$= 0.9183$

Weighted Average: $\left[\left(\frac{3}{9}\right) * 0.9183\right] + \left[\left(\frac{6}{9}\right) * (0.9183)\right] = 0.9183$

Information gain:

$0.991 - (0.9183) = \Delta = 0.0727$

Split Number 5

≤ Entropy(t) = $-\left[\left(\frac{2}{5}\right) * \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) * \log_2\left(\frac{3}{5}\right)\right]$

$= -[(-0.52877) + (-0.44218)]$

$= 0.97095$

> Entropy(t) = $-\left[\left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right)\right]$

$= -[(-0.5) + (-0.5)] = 1$

Weighted Average: $\left[\left(\frac{5}{9}\right) * (0.97095)\right] + \left[\left(\frac{4}{9}\right) * 1\right] = 0.96386$

Information gain:

$\Delta = 0.991 - (0.99032) = 0.007139$

Split

Split Number 6

< Entropy (t) = $-\left[\left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right) + \left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right)\right] = 1$

> Entropy (t) = $-\left[\left(\frac{1}{3}\right) * \log_2\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right) * \log_2\left(\frac{2}{3}\right)\right]$

$= \left[(0.52388) + (-0.38998)\right]$

$= 0.9183$

Weighted Average : $\left[\left(\frac{6}{9}\right) * (1)\right] + \left[\left(\frac{3}{9}\right) * 0.9183\right] = 0.972765$

Information gain : $\Delta = 0.991 - (-0.8889) = 0.10211$ 0.018235

Split number 7

< Entropy (t) = $-\left[\left(\frac{4}{8}\right) * \log_2\left(\frac{4}{8}\right) + \left(\frac{4}{8}\right) * \log_2\left(\frac{4}{8}\right)\right]$

$= -\left[(-0.5) + (-0.5)\right] = 1$

> Entropy (t) = $-\left[\left(\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right) + \left(\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right)\right] = 0$

Weighted Average = $\left[\left(\frac{8}{9}\right) * 1\right] + \left[\left(\frac{1}{9}\right) * 0\right] = 0.8889$

Information Gain = $0.991 + 0.8889 = 0.10211$

Split Number 8

< Entropy (t) = $-\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$

$= -\left[(-0.51997) + (-0.47111)\right] = 0.99108$

> Entropy (t) = $-\left[\left(\frac{0}{0}\right) * \log_2 \cdots\right] = 0$

Information Gain $\Delta = 0$

d) What is the best split (between a_1, a_2, a_3) according to the information gain?

→ According to the information gain, the best split is a_1 due to its higher gain in comparision to a_2 and a_3.

e) What is the best split (between a_1, a_2) according to the classification error rate?

→ According to the classification error rate, the best split is a_1 due to a lower classification error in comparision to a_2.

The classification error depicts the accuracy of the sample set) the higher the classification error the more error the sample contains

a_1: classification error $(t) = 1 - \max_i [P(i|t)]$

$= 1 - \left[\frac{7}{9}, \frac{2}{9}\right] = 1 - \frac{7}{9} = \frac{2}{9} = 0.222$

a_2: classification error $(t) = 1 - \left[\frac{5}{9}, \frac{4}{9}\right]$

$= 1 - \frac{5}{9} = \frac{4}{9} = 0.444$

f) What is the best split (between a_1 and a_2) according to Gini Index?

According to Gini Index a_1 is the best split because It got lesser Gini Index value

a_1: $\text{gini}(t) = 1 - \sum_{i=0}^{i=1} [P(i|t)]^2 = 1 - \left[\left(\frac{7}{9}\right)^2 + \left(\frac{2}{9}\right)^2\right] = 0.3457$

a_2: $\text{gini}(t) = 1 - \left[\left(\frac{4}{9}\right)^2 + \left(\frac{5}{9}\right)^2\right] = 1 - \left(\frac{16}{81} + \frac{25}{81}\right) = 0.4938$