6000419005 4
Jasil Pawar
DWM

ESE - DWM - SEM V

11/02/2022

Q.2] a) Data Transformation -

The changes made in the format of the structure of data is called data transformation. These methods are used to make the data more usable for the mining process.

The various methods are:

1) Smoothing -

with the help of algorithms, we can remove the noise from the dataset & help in knowing the important features of the dataset.

2) Aggregation -

In aggregation, the data is stored and presented in the form of a summary. The dataset which is from multiple sources is integrated with the data analysis description

3) Discretization -

The continous data is split into discrete intervals. With this we reduce the size of the data. This also makes the data usable for categorical algorithms.

4) Normalization -

In normalization the data is scaled so that it can be represented within a smaller range eg: -1 to 1 or 0 to 1

Discretization :

Data discretization is the method of converting huge number of data values into smaller ones so that the execution & management of data becomes easy.

In this, we convert the attributes of continous values to a data with a finite set of intervals.

There are two types of discretization
- → Superviised
- → Unsupervised

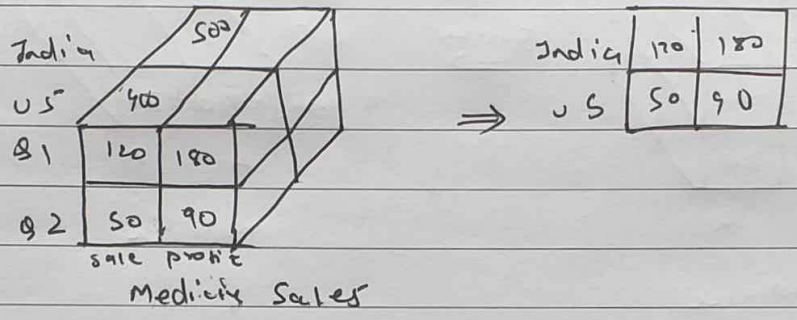
For eg: If we have age values from 0 to 80, it can be split into several ranges like 0 - 20, 21 - 40, 41 - 60, 61 - 80 //

The algorithms used to discretize the data are
- → Binning
- → Histogram Analysis
- → cluster Analysis

Q.2) b) Various OLAP operations are

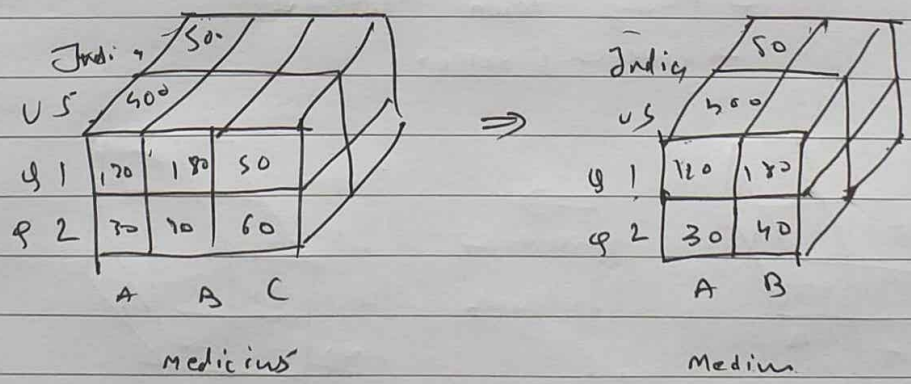
a) Slice -

In slice we select a particular dimension & provide a new cube with one dimension flat



b) Dice :-

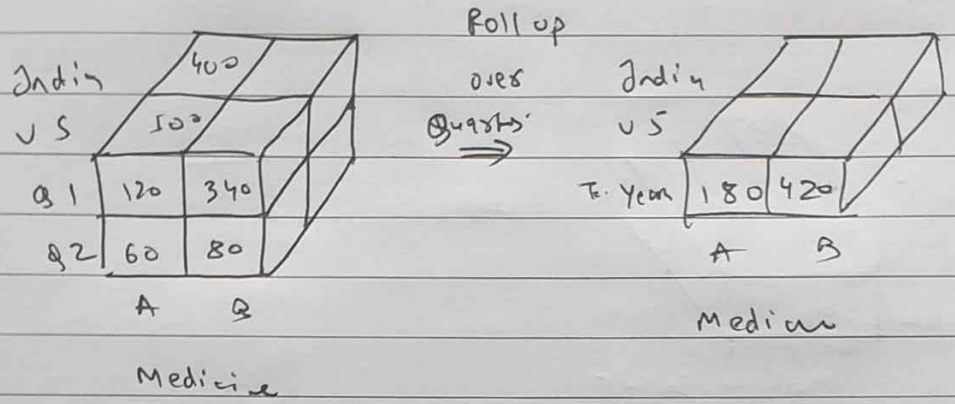
In dice, one or more dimensions are selected & a subcube is generated



iii) Roll up →

In rollup, the data cube is aggregated over a dimension

→ Moving up concept hierarchy

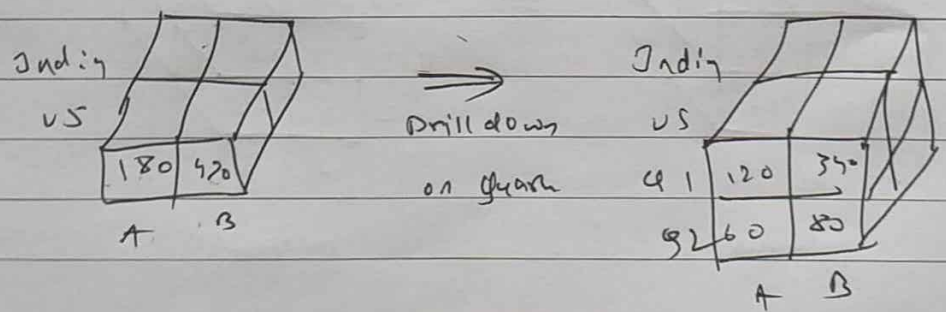


iv) Drill down -

In drill down, the data cube is further divided into detailed data

→ Moving down hierarchy

→ reverse of roll up



Q.3] a] Techniques to improve classification accuracy

1. Ensemble Methods -

An ensemble for classification is a composite model made up of a combination of classifiers. Ensembles tend to be more accurate than their component classifiers.

a) Bagging -

A bagging ensemble model trains a number of separate classifiers using bootstrap training. To classify a tuple, the bagging classifier counts the no of votes for each class by its trained models and assigns the class with the majority votes.

b) Boosting -

A boosting ensemble model assigns weights to each training tuple. A series of k classifiers are then learned. The learning process gives more weights to tuples that were incorrectly classified by previous classifiers so that the next models can pay "more attention" to those tuples. The ensemble classifier then combines the votes of its individual classifiers while predicting with the weight being a function of their accuracy.

c) Random forest -

This model generates a set of random trees with different attributes used to determine the split. During classification, the majority vote from the trees is assigned

Other methods for improving accuracy include

1. Solving Class Imbalance
   → when no of types of different classes differ in orders of magnitude
   → solved by oversampling, undersampling & hybrid sampling

2. Adding more data to reduce bias & overfitting.

3. Treating missing & outlier data

4. Feature selection from the data

6000419005 4

Q.3] b] Metadata -

Metadata can be defined as data about the data. It is a catalogue of the data & is known as the nerve center. Metadata describes all the pertinent aspects of the data warehouse fully and precisely.

Based on the functional areas in a data warehouse, metadata can be classified as:

1) Data acquisition
2) Data Storage
3) Information Delivery

1) Data Acquisition -

Metadata is recorded by processes in the data acquisition area for administering and monitoring the ongoing functions of a data warehouse. The users also use this metadata to find sources for their data elements. This metadata covers processes like data extraction, transformation, cleansing, integration and loading.

2) Data Storage -

Metada from this area is used for designing the full data refreshes and incremental data loads. The DBA uses the metadata for the processing of backup, recovery & tuning the database. The metadata is also used for purging & administration. It is also used by the users to get latest load date. It covers data loading, archiving & mangement

3) Information Delivery -

Metadata recorded in information delivery relates to predefined queries, predefined reports and input parameter dimensions for queries and reports. This metadata also includes information for OLAP. This covers report generation query processing and complex analysis.

(Q.4)

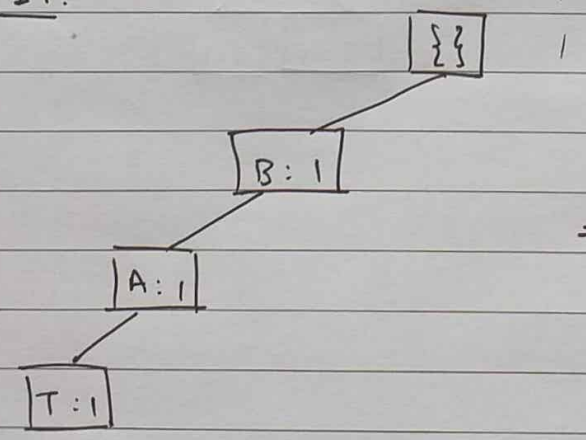
Q.4] a) FP Tree

Transaction ID	Items		TID	Items (sorted)
100	B, A, T		100	B, A, T
200	A, C		200	A, C
300	A, S		300	A, S
400	B, A, C		400	B, A, C
500	B, S	⇒	500	B, S
600	B, S		600	B, S
700	B, S, A, T		700	B, S, A, T
800	B, S		800	B, S
900	B, A, S		900	B, A, S

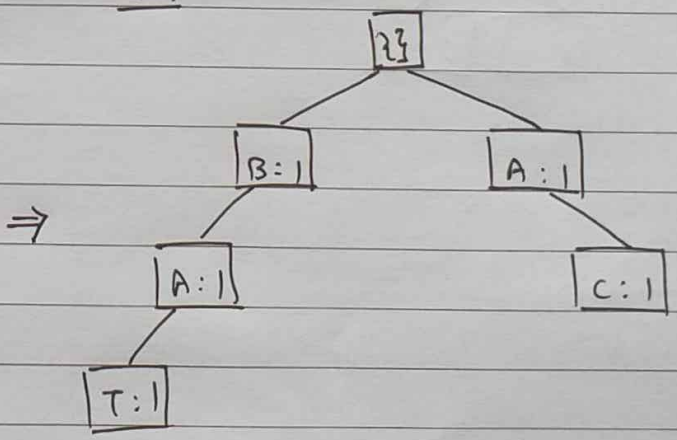
minSupport = 2 , Frequency: B = 7, A = 6, S = 6, T = 2, C = 2

⇒

I1:



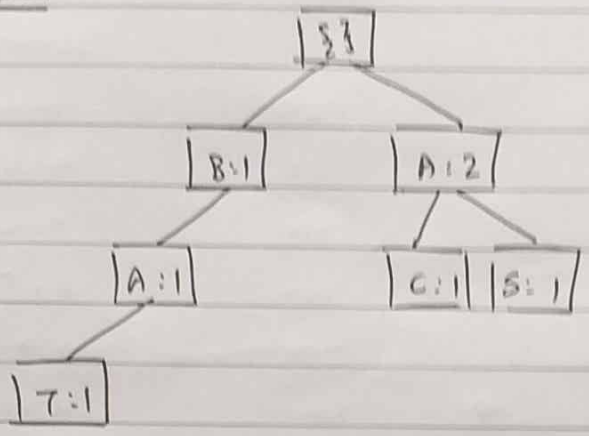
I2:



6000419005H

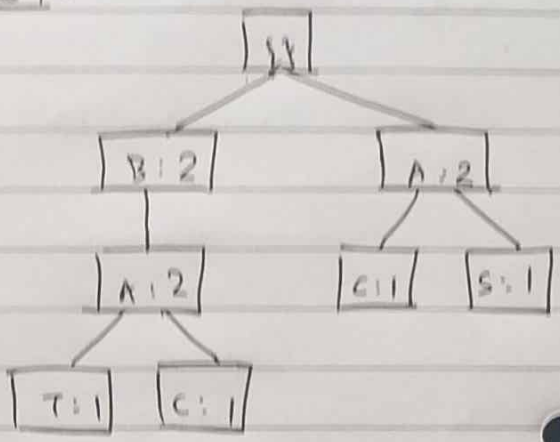
Ans

I3:

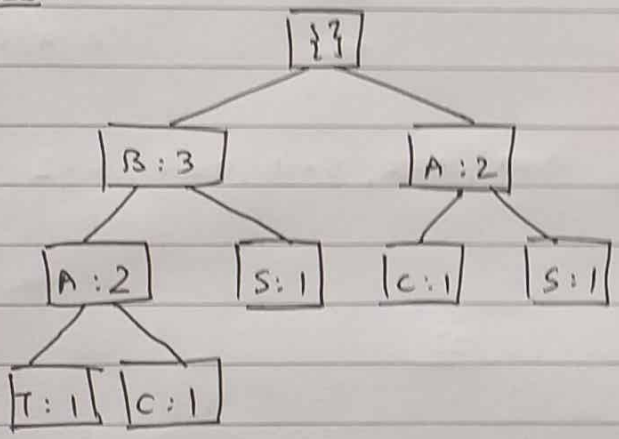


⇒

I4:

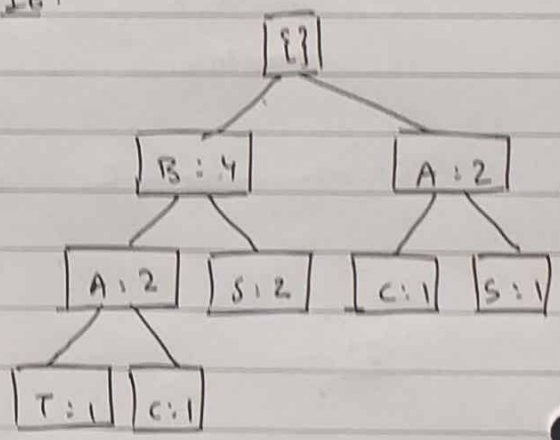


I5:



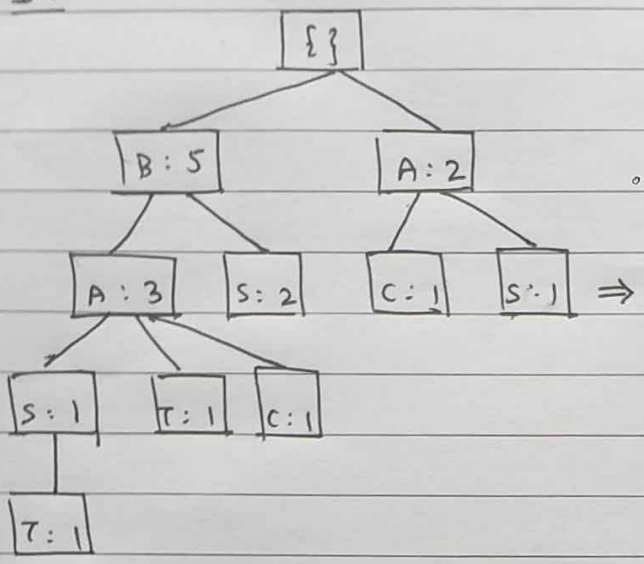
⇒

I6:



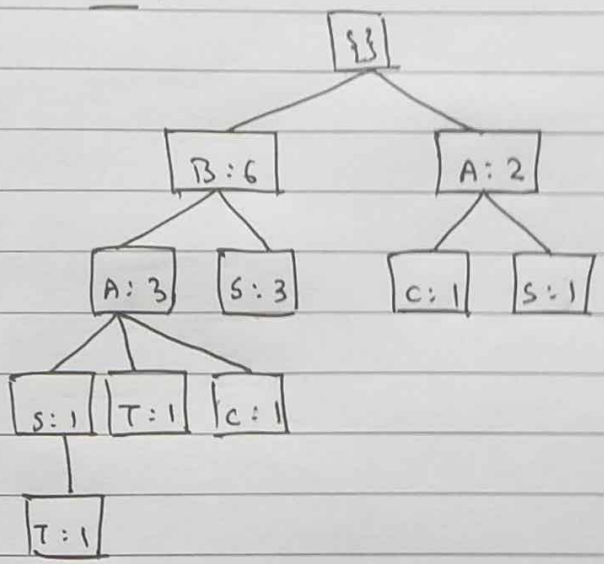
6000419005G

I7 :

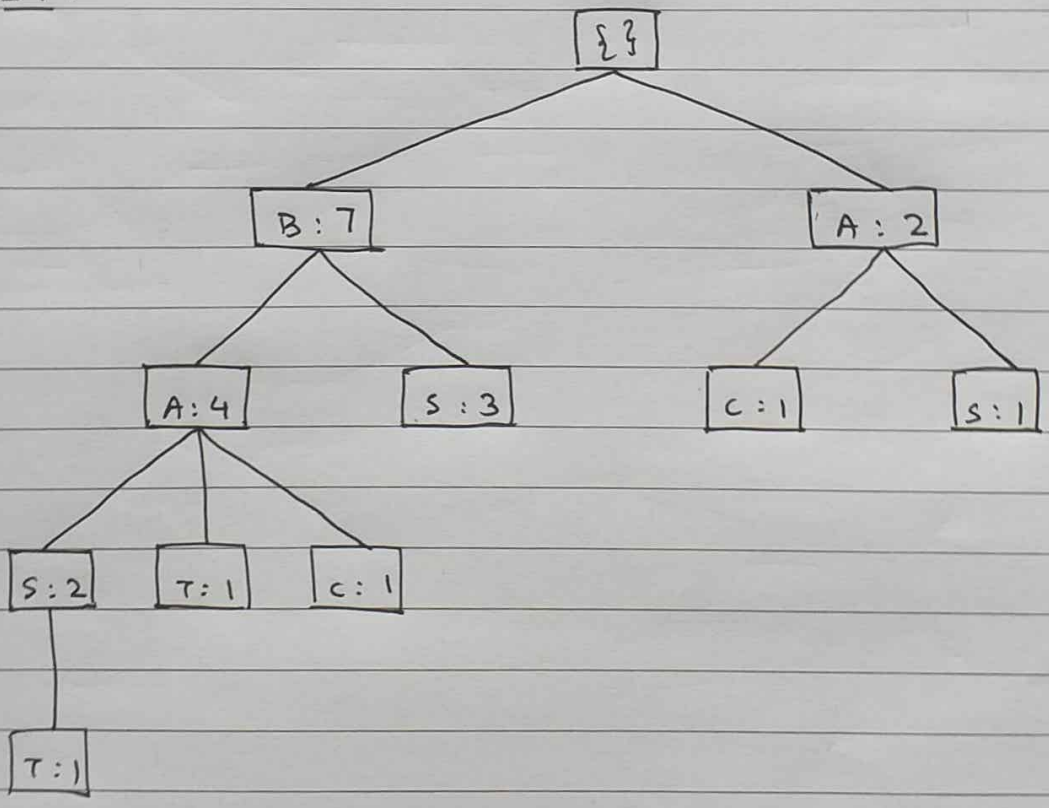


⇒

I8 :



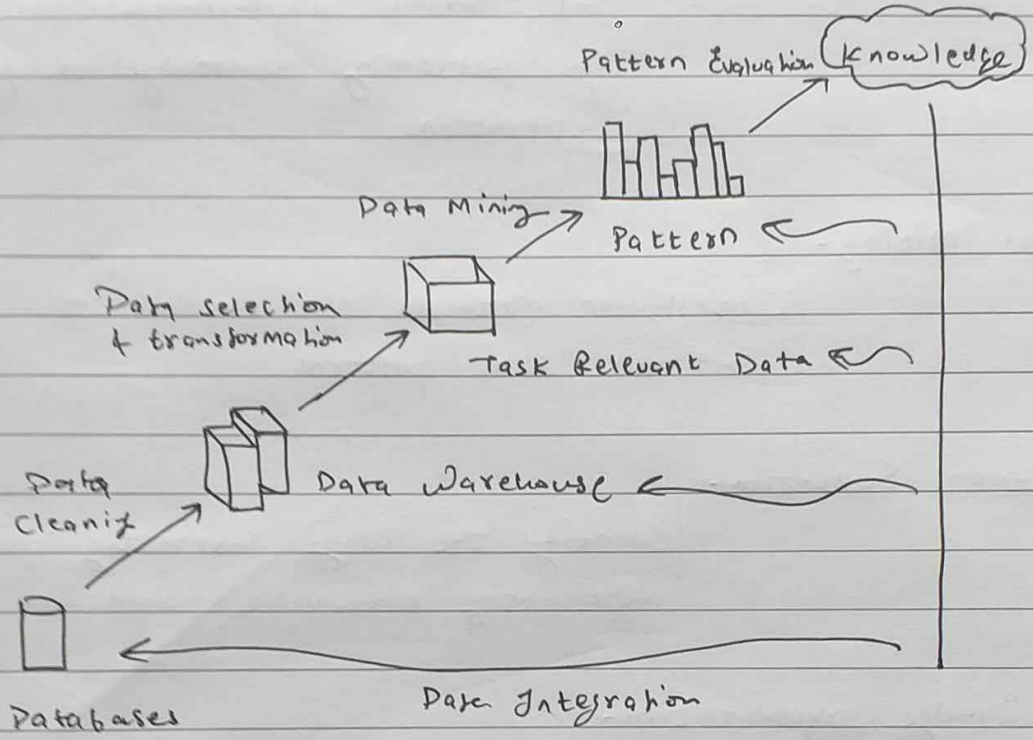
I9 :



6000419005H

Unit

Q. 4) b) Knowledge Discovery process in data mining



1. Data Cleaning –

Removing noisy & inconsistent data

2. Data Integration :-

Multiple data sources are combined

3. Data Selection –

Data relevant to the analysis task is retrieved from the database

4) Data Transformation -

Data is transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations

5) Data Mining -

Intelligent methods are applied to extract data patterns

6) Pattern Evaluation -

Identify the truly interesting patterns representing knowledge based on measures

7) Knowledge presentation -

Visualisation and knowledge representation techniques are used to present mined knowledge.

6000419005 4

Q. 5] a]

Transaction ID	Items
t1	1, 3, 4
t2	2, 3, 5
t3	1, 2, 3, 5
t4	2, 5
t5	1, 2, 3, 5

min-sup = 40% min conf = 70%.

⇒

min Sup = 40% .

$40\% = \frac{x}{5} \times 100 \quad \Rightarrow x = 2$ //

Support = 2

C1 :

Item	Count
1	3
2	4
3	4
4	1
5	4

L1 :

Item	Count
1	3
2	4
3	4
5	4

C2 :

Item	Count
{1, 2}	2
{1, 3}	3
{1, 5}	2

L2 :

Item	Count
{1, 2}	2
{1, 3}	3
{1, 5}	2

{2, 3}	3	{2, 3}	3
{2, 5}	4	{2, 5}	4
{3, 5}	3	{3, 5}	3

C3

Item	Count
{1, 2, 3}	2
{1, 2, 5}	2
{2, 3, 5}	2
{1, 3, 5}	2

L3

Item	Count
{1, 2, 3}	2
{1, 2, 5}	2
{2, 3, 5}	3
{1, 3, 5}	2

C4

Item	Count
{1, 2, 3, 5}	2

L4

Item	Count
{1, 2, 3, 5}	2

Hence, frequent itemset (largest) = {1, 2, 3, 5} with support 2

6000419005 4

Q.I] b] MOLAP vs ROLAP vs HOLAP

ROLAP	MOLAP	HOLAP
1. Relational ~~Multidimensional~~ DB is used as storage location for summary aggregation	1. Multidimensional DB is used as storage location for summary aggregation.	1. Multidimentional DB is used as storage location for summary aggregation
2. Processing time of ROLAP is very slow	2. Processing time is fast	2. Processing time is fast
3. Requires large storage	3. Requires medium-level of storage	3. Requires low storage compared to MOLAP & ROLAP
4. Relational DB is used for detailed data	4. Multidimentional DB is used for detailed data	4. Relational DB is used for detailed data
5. Low access latency	5. High access latency	5. Medium access latency
6. Slow query response	6. Fast query response	6. Medium query response

6000419005 4

Q 6] Given test record: X = (Refund = No, Married, Income = 120K)

Refund / Evade	Yes	No
Yes	0	3/7
No	1	4/7

M-Status / Evade	Yes	No
Single	2/3	4/7
Divorce	1/3	1/7
Married	0	4/7

For taxable income

if evade = No, sample mean = 110 , varn = 2975

evade = Yes, mean = 90 , VM = 25

$P(X/No) = P(\text{refund} = No/No) \times P(\text{married}/No) \times P(\text{income} = 120K/No)$

$= \frac{4}{7} \times \frac{4}{7} \times \left(\frac{1}{\sqrt{2\pi}(54.54)} \cdot e^{\left(\frac{110-120}{2(54.54)}\right)^2} \right)$

$= 0.0023$

$P(X/Yes) = P(\text{ref} = No/Ye) \times P(\text{Married}/Yes) \times P(\text{Income} = 120K/Yes)$

$= 0$

60004190054

Since,

$P(x|No) \cdot P(No) > P(x|Yes) \cdot P(Yes)$

"class = no"

Thus,

a) The given person does not evade tax

b) All tuples with status = married have outcome evade = No

∴ All married tuples don not evade tax

Q6) b) 1) Page Rank

→ Page rank is an algorithm to rank web pages
→ It is a way of measuring importance of web page
→ It count the no of quality input links to a page to determine a rough estimate of importance
→ It assumes that more important websites are likely to receive more links
→ It assigns numerical weight to each element to give importance
→ The numerical rank given is known as Page Rank.

2) HITS

→ Hyperlink induced topic search is an algorithm used in link analysis
→ It discovers ranks the webpages relevant to a search
→ HITS identifies
  → HUBS - pages that contain good information across
  → Authorities - Pages that are good for given info

→ Authority score -
  How good a page is for given query & retrieved by how many pgs

→ Hub score →
  How many good authorities it points to