1. Define Data and Information. Explain Kinds of Data with an example.

A) Data :-

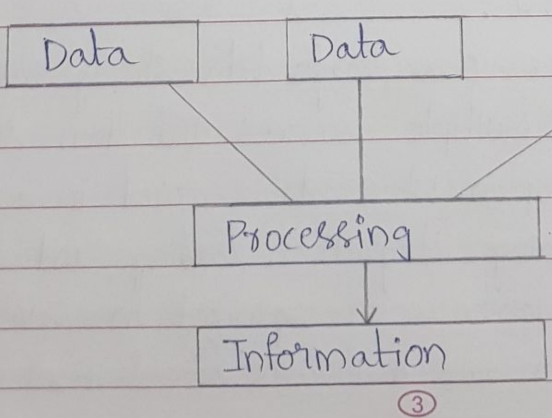
Data is a raw and unorganized fact that required to be processed to make it Meaningful. Data can be simple at the same time Unorganized unless it is organized. data comprises facts, observations, perceptions, numbers, characters, symbols, image etc..

Data is always interpreted, by a human or machine, to derive Meaning. Data contains numbers, statements and characters in a raw form.

Information :-

Information is a set of data which is processed in a Meaningful way according to the given requirement. Information is processed, structured, or presented in a given context to Make it Meaningful and useful. It improves the reliability of the data.

It helps to ensure undesirability and reduces uncertainty.



Types of data that can be Mined are:-

1. Data stored in the database:-

A database is also called a database Management System or DBMS. Every DBMS stores data that are related to each other in a way or the other. It also has a set of software programs that are used to Manage data and provide easy access to it and Managing different types of data access, such as shared, distributed and concurrent.

A relational db has tables that have different names, attributes and can store rows or records of large data sets. Every record stored in a table has a Unique Key. Entity-relationship model is created to provide a representation of a relational db that features entities and the relationships that exist between them.

Ex:-

customer (cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...)

item (item_ID, brand, category, type, price, Supplier, cost...)

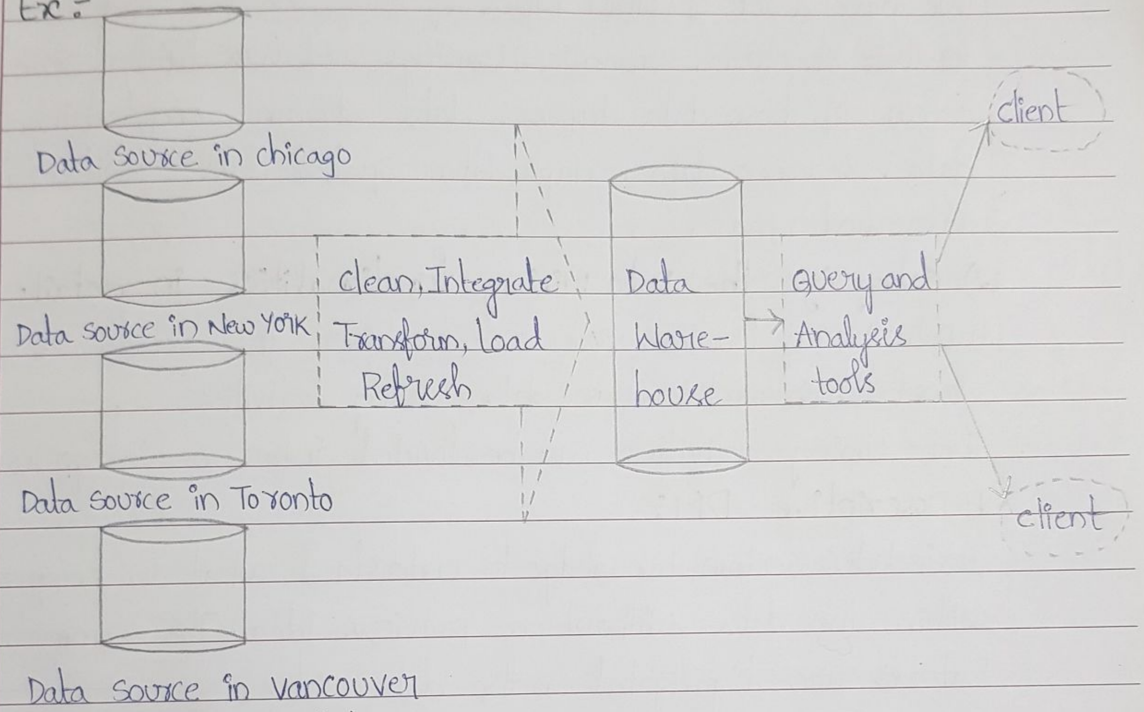
branch (branch_ID, name, address, ...)

2. Data Warehouse:-

A data Warehouse is a single data storage location that collects data from Multiple sources and then stores it in the form of a unified plan. When data is stored in a data Warehouse, it undergoes cleaning, integration, loading and refreshing. Data stored in a data Warehouse is organized in several parts. If you want information on data that was stored 6 or 12 Months back.

you will get it in the form of a summary.

Ex:-



3. Transactional data:-

Transactional database stores record that are captured as transactions. These transactions include flight booking, customer purchase, click on a website, and others. Every transaction record has a unique ID. It also lists all those items that made it a transaction.

Ex:-

trans_ID	list_of_item_IDs
T100	I1, I3, I8, I16
T200	I2, I8
...	...

4. Other types of data :
We have a lot of other types of data as well that are known for their structure, semantic meanings, and versatility. Here are a few of those data types : data streams, engineering design data, sequence data, graph data, spatial data, Multimedia data, and more.

2) Categorize the Data Mining functionalities in detail.

A) Data Mining functions are used to define the trends or correlations contained in data Mining activities.
Data Mining activities can be divided into 2 categories :

1. Descriptive DM :-
It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.
Ex:- count, average etc..

2. Predictive DM :-
It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the s/w estimates the characteristics that are absent.
Ex:- Judging from findings of a patient's Medical examinations.

DM functionality :

1. Class /concept Descriptions :-
Classes or definitions can be correlated with results. In simplified descriptive and yet accurate ways, it can be helpful to define individual groups and concepts. These class or concept definitions

are referred to as class/concept descriptions.

⇒ Data characterization:

This refers to the summary of general characteristics or features of the class that is under the study.

⇒ Data Discrimination:

It compares common features of class which is under study.

The output of this process can be represented in Many forms.

Ex: bar charts, curves, pie charts.

2. Mining Frequent Patterns, Associations and correlations:

Frequent patterns are nothing but things that are found to be most common in the data.

There are different Kinds of frequency that can be observed in the dataset.

⇒ Frequent item set:

This applies to a number of items that can be seen together regularly for ex: Milk and sugar.

⇒ Frequent Subsequence:

This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.

⇒ Frequent Substructure:

It refers to the different Kinds of data structures such as trees and graphs that May be combined with the itemset or subsequence.

Association analysis:

The process involves uncovering the relationship b/w data and deciding the rules of the association. It is a way of discovering the relationship b/w various items. Ex:- It can be used to determine the sales of items that are frequently purchased together.

Correlation Analysis :

Correlation is a Mathematical technique that can show whether and how strongly the pairs of attributes are related to each other.

Ex:- Highted people tend to have more weight.

3) Classification and prediction :-

⇒ Classification

It is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous. The derived model is established on the analysis of a set of training data.

⇒ Prediction

It defines predict some unavailable data values or pending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase/decrease trends in time-related information.

4) Clustering :

It is similar to classification but the classes are not predefined. The classes are represented by data attributes. It is unsupervised learning. The objects are clustered or grouped, depends on the principle of Maximizing the intraclass similarity and Minimizing the intraclass similarity.

5) outlier analysis :-

outliers are data elements that cannot be grouped in a given class or cluster. These are the data objects which have multiple behaviour from the general behaviour of other data objects. The analysis of this type of data can be essential to Mine the Knowledge.

6) Evolution analysis :-

It defines the trends for objects whose behaviour changes over some time.

3)(i) Elaborate the different data preprocessing techniques

Data preprocessing is a data Mining technique which is used to transform the raw data in a useful and efficient format.

Techniques involved are

1) Data cleaning :-

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of Missing data and noisy data.

a) Missing data : This situation arises when some data is Missing in the data. It can be handled in various ways.

1) Ignore the tuples :- This approach is suitable only when the dataset we have is quite large and Multiple values are missing within a tuple.

2) Fill the Missing values :- You can choose to fill the missing

Values Manually, by attribute Mean or the most probable Value.

b) Noisy data :- It is a Meaningless data that can't be interpreted by Machines. It can be generated due to faulty data collection handled by.

1) Binning Method :- It works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task.

2) Regression :- Data can be made smooth by fitting it to a regression function. Regression Used May be linear or Multiple.

3) clustering :- It groups the similar data in a cluster.

2) Data Transformation :-

It transforms data in appropriate forms suitable for Mining process.

1) Normalization :- It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2) Attribute selection :- In this, new attributes are constructed from the given set of attributes to help the Mining process.

3) Discretization :- It is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4) Concept Hierarchy Generation :- Attributes are converted from lower level to higher level in hierarchy.

3) Data reduction :- It aims to increase the storage efficiency and reduce data storage and analysis costs.

1) Data cube Aggregation :- Aggregation operation is applied to data for the construction of the data cube.

2) Attribute Subset Selection :- The highly relevant attributes

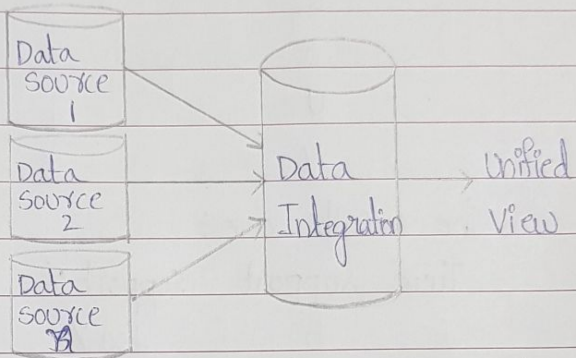
should be used, rest all can be discarded.

3) Numerosity Reduction:- This enable to store the model of data instead of whole data.

4) Dimensionality reduction:- This reduce the size of data by encoding Mechanisms.

4) Data Integration:-

It combines data from Multiple sources into a coherent data source. These sources May include Multiple databases.



ii) Apply the following rules on a database has 5 transactions. let min sup = 60% and min items bought conf = 80%.

TID	Items bought	TID	Items bought
T100	{M,O,N,K,E,Y}	T400	{M,U,C,K,Y}
T200	{D,O,N,K,E,Y}	T500	{C,O,O,K,I,E}
T300	{M,A,K,E}		

a) Find all frequent item sets using Apriori

b) list all of the strong association rules (with sup & conf)

A) 1. The database scanned once to generate the 1 itemset. The total ID is 5. Minimum support is 60%.

Item	Support	Support %
A	1	20%
C.	2	40%.

D	1	20%
E	5	100%
I	1	20%
K	4	80%
M	3	60%
N	2	40%
O	4	80%
U	1	20%
Y	3	60%

The results are :-

Item	support	Support %
E	5	100%
K	4	80%
M	3	60%
O	4	80%
Y	3	60%

⇒ 2. The database scanned once to generate the 2 Itemsets.

Item	Support	Support %
E, K	4	80%
E, M	3	60%
E, O	4	80%
E, Y	3	60%
K, M	2	40%
K, O	3	60%
K, Y	2	40%
M, O	2	40%
M, Y	2	40%
O, Y	3	60%

The result is :-

E, K	4	80%
E, M	3	60%
E, O	4	80%
E, Y	3	60%
K, O	3	60%
O, Y	3	60%

3. The database scanned once again to generate 3 itemsets. Result is:

Item	Support	Support %
E,K,O	3	60%
E,O,Y	3	60%

The itemsets (E,K,O) and (E,O,Y). So, there are 6 possible Association rules.

Association rules of {E,K,O}:

R1: E → K,O = 60% < 80%
R2: K → E,O = 75% < 80%
R3: O → E,K = 75% < 80%
R4: E,K → O = 75% < 80%
R5: K,O → E = 100% > 80%
R6: E,O → K = 75% < 80%

Association rules of {E,O,Y}:

R7: E → O,Y = 60% < 80%
R8: Y → E,O = 100% > 80%
R9: O → E,Y = 75% < 80%
R10: E,O → Y = 75% < 80%
R11: O,Y → E = 100% > 80%
R12: E,Y → O = 100% > 80%

Rules R5, R8, R11, R12 are strong association rules

∴ customers who purchase any of the 2 products from E,K,O would want to purchase the other one and all the customers who purchase 2 items from E,O,Y also want to purchase the other one.

4) What is FP-Growth tree? Explain FP-Growth tree algorithm with an Example.

A) Frequent pattern tree is a tree-like structure that is made with initial itemsets of the db. Each node of FP tree represents an item of the itemset. The rootnode represents null while the lower nodes represent the itemsets.

FP Algorithm steps

The FP growth method lets us find the Frequent pattern without candidate generation.

#1 First scan the database to find the occurences of the itemsets in db i.e support count.

#2 secondly, construct the FP tree. create root node which is null.

#3 scan the db again and examine the transactions. The itemset with the Max count is taken at top, the next set with lower count and so on. tree constructed in descending order.

#4 After ordering if any itemset of this transaction is already present in another branch then this would share a common prefix to the root. common itemsets are linked.

#5 count of IS is incremented as it occurs in transactions.

#6 Mine the created FP tree. Traverse the path which is conditional pattern

#7 construct a conditional FP tree, which is formed by a count of itemsets in the path. The itemsets Meeting threshold support are considered in the conditional FP tree.

#8 Frequent patterns are generated from this.

Ex:- Support threshold = 50%, confidence = 60%.

Table 1

Transaction	list of items
T_1	i1, i2, i3
T_2	i2, i3, i4
T_3	i4, i5
T_4	i1, i2, i4
T_5	i1, i2, i3, i5
T_6	i1, i2, i3, i4

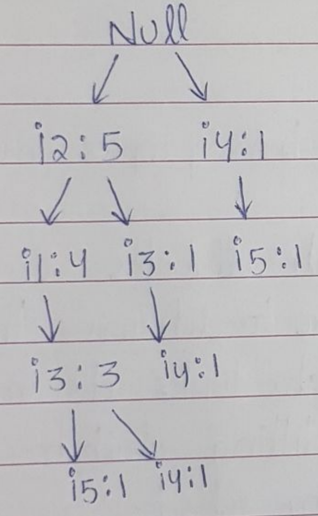
Min-supp = 3 = 0.5 * 6

1. Count of each item and sorting them in descending order

Item	count
i1	4
i2	5
i3	4
i4	4
i5	2

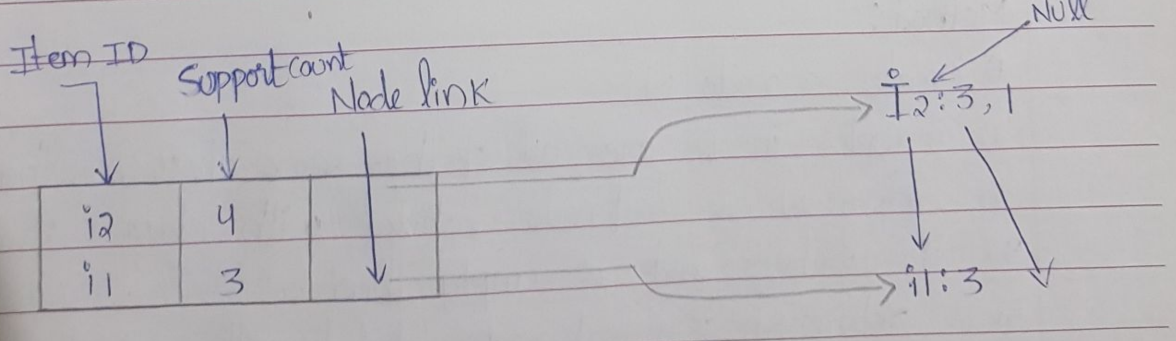
Item	count
i2	5
i1	4
i3	4
i4	4

3. Build FP tree



4. Mining of FP-tree

Item	conditional Pattern base	conditional FP-tree	Frequent Patterns generated
i4	{i2, i1, i3:1}, {i2, i3:1}	{i2:2, i3:2}	{i2, i4:2}, {i3, i4:2}, {i2, i3, i4:2}
i3	{i2, i1:3}, {i2:1}	{i2:4, i1:3}	{i2, i3:4}, {i1, i3:3}, {i2, i1, i3:3}
i1	{i2:4}	{i2:4}	{i2, i1:4}



5. Explain Decision tree induction algorithm for classification. Discuss the usage of information gain.

A machine researcher named J.Ross Quinlan in 1980 developed a decision tree Algorithm. Known as ID3 (Iterative dichotomiser). later he presented C4.5, which was the successor of ID3.

In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide and conquer manner.

Decision tree induction algorithm

Algorithm: Generate_decision_tree.

Input:

⇒ Data partition D, which is a set of training tuples and their associated class labels.

⇒ attribute_list, the set of Candidate attributes

⇒ Attribute_selection_method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting_attribute and possibly, either a split-point or splitting subset.

output: A decision tree.

Method:

1) create a node N;
2) if tuples in D are all of the same class, C, then
3) return N as leaf node labeled with class C;
4) if attribute_list is empty then
5) return N as a leaf node labeled with majority class in D;
6) apply Attribute_selection_Method (D, attribute_list) to find the

"best" splitting_criterion;
7) label node N with splitting criterions;
8) if splitting_attribute is discrete_valued and
 Multiway splits allowed then // not restricted to binary trees.
9) attribute_list ← attribute_list - splitting_attribute;
10) for each outcome j of splitting_criterion
11) let Dj be the set of data tuples in D satisfying outcome j;
12) if Dj is empty then
13) attach a leaf labeled with the majority class in D to node N;
14) else attach the node returned by Generate_decision_tree (Dj, attribute_list) to node N;
15) return N;

Usage of information gain:

Information gain helps to determine the order of attributes in the nodes of a decision tree. We can use IG to determine how good the splitting of nodes in a decision tree.

I - ASSIGNMENT

(Start Writing From Here)

1) What is Data mining? explain Data mining preprocessing Techniques.

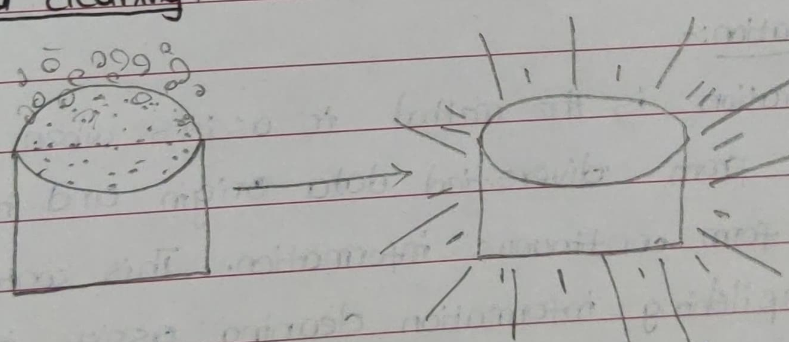
Data Mining:- is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data Analysis.

preprocessing:-

1) Incompleteness (lacking attributes)
2) Noise (Deviate from actual result)
3) Inconsistency (Discrepanies in code).

Froms of pre-processing:-

(A) Data cleaning:-



The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

• (a). Missing Data:-
This situation arises when some data is missing in the data.
It can be handled in various ways.
some of them are:

1. Ignore the tuples:-
This approach is suitable when some data is missing in the data. It can be handled in various ways. dataset we have is quite large and multiple values are missing within a tuple.

2) Fill the missing values:-
There are various ways to do this task. you can choose to fill the missing values manually, by attribute mean (or) the most probable value.

(B) Noisy Data:-
Data Integration:-
Data integration is the method to assists when the information is collected from diversified data origin and information is merging to form continuous information. This continous information after accompilshing information cleaning assists in analysis and data preparation.

Data Transformation:- Data Transformation is assisted to modify the raw information into a definite design according to the requirement of the set.

Normalization:-
In the Normalization method, numerical information is modified into the identified range.

Aggregation:-

Aggregation can be specified from the conversation itself, this process assists to merge the characteristics into one.

Generalization:-

In generalization, a lower level feature is modifying to a higher state.

Data Reduction:-

After the conversion and scaling of information duplication that is insignificancy within the information is taking away and effectively organizes the information during data preparation.

2) List out Data mining task premitives?

A) A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives.

* These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process

1. The set of task-relevant data to be mined:-
This specifies the portions of the database (or) the set of data in which the user is instrested. This includes the database

attributes (or) data warehouse dimensions of instrest.

- In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join and aggregation.

* The data collection process results in a new data relational called the "initial data relation".

2) The kind of knowledge to be mined:-

This specifies the data mining functions to be performed, such as characterization, discrimination, association (or) correlation analysis, classification, prediction, clustering, outlier analysis, (or) evolution analysis.

3- The background knowledge to be used in the discovery process.

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of of abstraction.

- Rolling up - Generlization of data:- Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.

CMR GROUP OF INSTITUTIONS
EXPLORE TO INVENT

- **Drilling Down - Specialization of data:-** concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

4) **The interestingness measures and thresholds for pattern evaluation:-**

Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. For example, interesting measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

→ The to make pattern evaluation.

- simplicity
- certainty (confidence)
- utility
- Novelty

5) **The expected representation for visualizing the discovered patterns:-**

This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, cross tabs, charts, graphs, decision trees, cubes, (or) other visual representations.

* users must be able to specify the forms of presentation to be used for displaying the discovered patterns. some representations forms may be letter suited than others for particular kinds of knowledge.

3) Illustrate with an ~~Ap~~ Apriori Algorithm for the given dataset below.

TID	list of items
001	milk, dal, sugar, bread
002	Dal, sugar, wheat, jam
003	milk, bread, curd, paneer
004	wheat, paneer, dal, sugar
005	milk, paneer, bread
006	wheat, dal, paneer, bread.

A) 1. Generating I^{st} frequent itemset pattern

Items	support count
Milk	3
dal	4
sugar	3
paneer	4
bread	4
wheat	3
Jam	1 ×
curd	1 ×

< 1

1 or < 2 will be neglected.

Items	support count.
milk	3
dal	4
sugar	3
paneer	4
bread	4
wheat	3

<1

2) Generating 2nd frequent item set pattern.

Items	Support count
$\boxed{\text{milk, dal}}$	1 ×
$\boxed{\text{milk, sugar}}$	1 ×
milk, paneer	2
milk, bread	3
$\boxed{\text{milk, wheat}}$	0 ×
Dal, Sugar	3
Dal, paneer	2
Dal, bread	2
Dal, wheat	3
$\boxed{\text{sugar, paneer}}$	1 ×
$\boxed{\text{sugar, bread}}$	1 ×
Sugar, wheat	2
paneer, bread	3
paneer, wheat	2
$\boxed{\text{bread, wheat}}$	1 ×

1 or <2 will be neglected.

Items	Support count
Milk, paneer	2
milk, bread	3
Dal, Sugar	3
Dal, paneer	2
Dal, bread	2
Dal, wheat	3
Sugar, wheat	2
paneer, bread	3
paneer, wheat	2

3) Generating 3rd frequent item set pattern

Items	Support count
Milk, paneer, bread	2
milk, paneer, dal, sugar	0
milk, panner, dal	0
milk, dal, panner, bread	0
milk, paneer, dal, wheat	0
milk, paneer, sugar, wheat	0
milk, paneer, bread	2
milk, paneer, wheat	0
Milk, bread, dal, Sugar	1
Milk, bread, dal, paneer	0
milk, bread, dal	1
milk, bread, dal, wheat	0

milk, bread, sugar, wheat	0 X
milk, bread, paneer	2
milk, bread, wheat, paneer	0 X
Dal, sugar, paneer	1 X
Dal, sugar, bread	1 X
Dal, sugar, wheat	2
Dal, Sugar, wheat	2
Dal, sugar, paneer, bread	0 X
Dal, sugar, paneer, wheat	1 X
Dal, paneer, bread	1 X
Dal, paneer, wheat	2
Dal, bread, wheat	1 X
Dal, bread, sugar, wheat	0 X
Dal, bread, panner	1 X
Dal, bread, panner, wheat	1 X
Dal, wheat, sugar	2
Dal, wheat, paneer, bread	1 X
Dal, wheat, paneer, wheat	2
Sugar, wheat, paneer, bread	0 X
sugar, wheat, paneer	1 X
paneer, bread, wheat	1 X

1 or <2 will be neglected.

Items	support count
milk, panner, bread	3
Dal, sugar, wheat	3
Dal, panna, wheat	3

Dal, paneer, wheat is the most associative frequent item set.

u) Apply the following rules on the database has five transactions let min sup = 60% and min items bought conf = 80%.

TID	Items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

(i) Apply FP-Growth Algorithm to the following transactional data to find frequent items.

(ii) List all frequent itemsets with their support count.

FP- Growth Algorithm.

1) Generating I^{st} frequent itemset pattern:-

TID	Datasets.
T100	{M, O, N, K, E, Y}
T200	{ D, O, N, K, E, Y }
T300	{ M, A, K, E }
T400	{ M, U, C, K, Y }
T500	{ C, O, O, K, I, E }

2)

Item	support count.	support %.
M	3	60%.
O	3	60%. min-
N	2	40%.
K	5	100%.
E	4	80%.
Y	3	60%.
D	1	20%.
A	1	20%.
U	1	20%.
C	2	40%.
I	1	20%.

3) The results are :-

Item	support	support %
E	5	100 %
K	4	80 %
M	3	60 %
O	4	80 %
Y	3	60 %

↑

Frequent pattern.

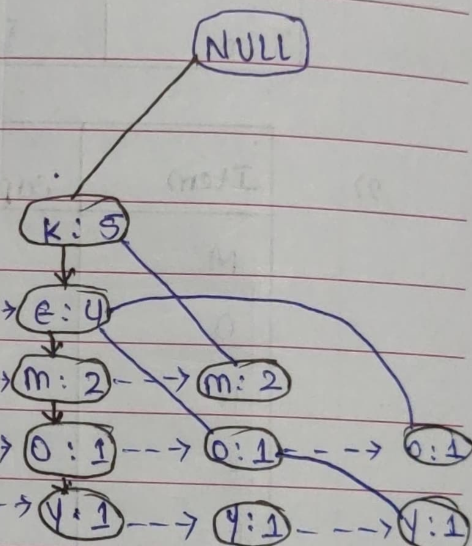
⇒ The database scanned once to generate the 1 Item sets

4) Ordered Item set :-

K, E, M, O, Y
K, E, O, Y
K, E, M
K, M, Y
K, E, O

5) F.P tree :-

Item ID	S.C	N.L
K	5	
E	4	
M	3	
O	3	
Y	3	



6) Mining of FP-tree :-

Item	conditional pattern base	conditional F.p tree.	Generated F.P.
Y	[{KEMO : 1}, {KEO : 1}, {KM : 1}]	[{K : 3}]	<K, Y : 3>
O	[{KEM : 1}, {KE : 2}]	[{KE : 3}]	<K, O : 3> <E, O : 3> <O, E, K : 3>
M	[{KE : 2}, {K : 1}]	[{K : 3}]	<M, K : 3>
E	[{K : 4}]	[{K : 4}]	<E, K : 3>
K	–	–	

5) Explain naive Bayes algorithm with an example.

A) Naive Bayes algorithm is a Supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

* The Naive Bayes algorithm is comprised of 2 words Naive and Bayes.

Baye's Theorem:-

* Baye's Theorem is a simple mathematical formula used for calculating conditional.

* Conditional probability is a measure of the probability of an event ocurring given that another event has (by assumption, presumption, assertion, or evidence) occured.

probability of B ocuring given evidence A has already ocured

probability of A occuring

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

probability of A occuring given evidence B has already occured

probability of B occuring.

Example:-

person	Height	Weight	Foot size.
male	6.00	180	12
male	5.92	190	11
male	5.58	170	12
male	5.92	165	10
Female.	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female.	5.75	150	9.

p(male) = 4/8 = 0.5

p(female) = 4/8 = 0.5

Male:-

$$\text{mean (Height)} = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$\text{Variance (Height)} = \frac{\Sigma(x_i - x)^2}{n-1} = 0.035055.$$

sex	(Height) mean	(height) variance	(weight) mean	(weight) variance	(foot size) mean	(foot size) variance
male	5.855	0.035055	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667.

p(male) = 4/8 = 0.5

p(female) = 4/8 = 0.5

New instance to be classified is:-

sex	Height	weight	foot size.
sample	6	130	8.

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2 * 0.05033}} = 1.5789.$$

$P(W|M) = 5.9881e^{-6}$	$P(H|F) = 2.2346\,e^{-1}$	$P(FS|F) = 2.8669e^{-1}$
$P(FS|M) = 1.3112e^{-3}$.	$P(W|F) = 1.6789\,e^{-2}$	

$$\text{posterior (male)} = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{\text{Evidence.}}$$

$$\Rightarrow 0.5 * 1.5789 * 5.9881\,e^{-6} * 1.3112e^{-3} = 6.198e^{-9}$$

$$\text{posterior (Female)} = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{\text{evidence.}}$$

$$\Rightarrow 0.5 * 2.2346e^{-1} * 1.6789\,e^{-2} * 2.8669e^{-1}$$

$$\Rightarrow 5.377\,e^{-4}$$

I - ASSIGNMENT

(Start Writing From Here)

1) what is KDD? Explain KDD steps with a neat diagram.

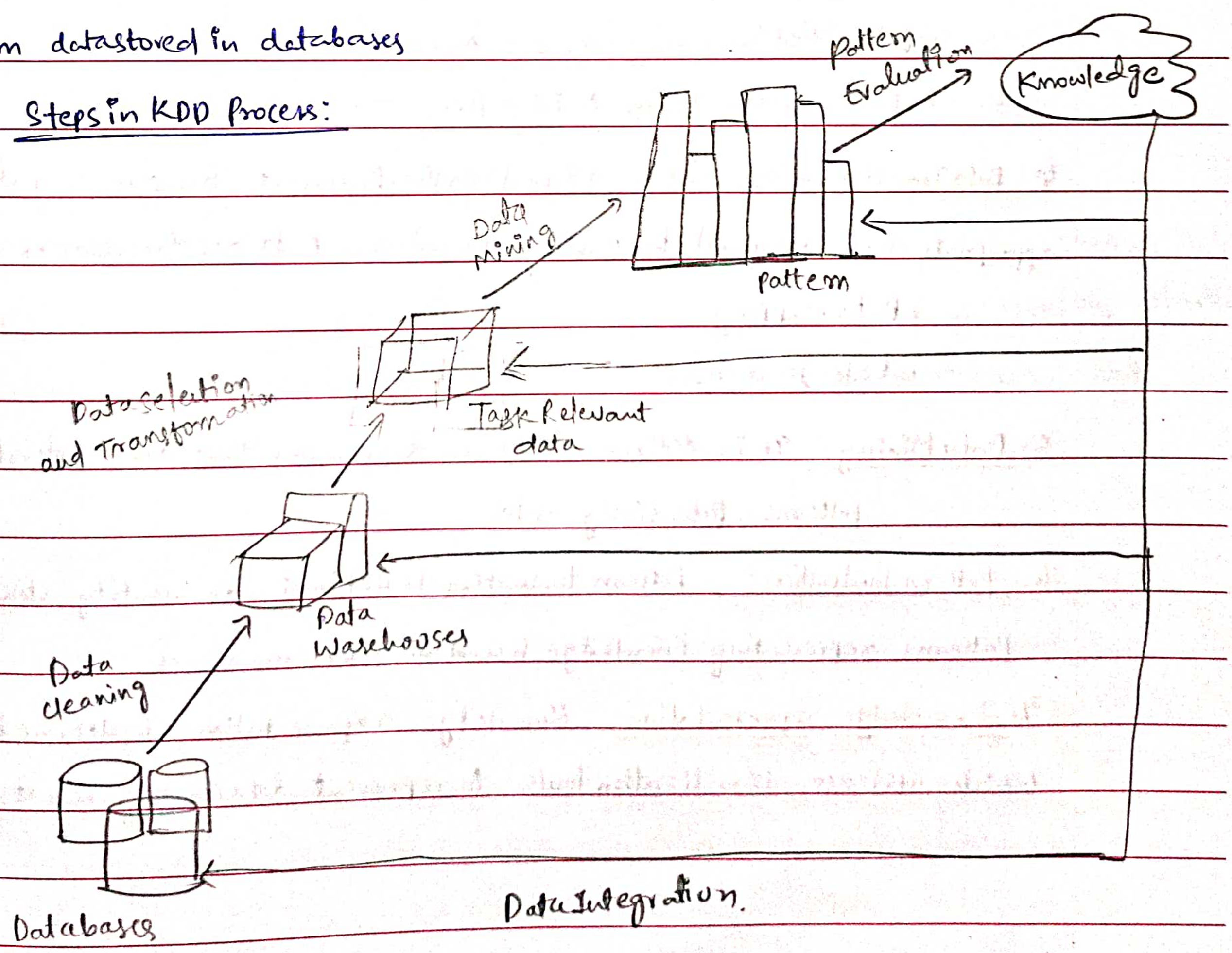
Data Mining - Knowledge Discovery in Databases (KDD).

Why we need DataMining?

volume of information is increasing everyday that we can handle from bussiness transcations, Scientific data, Sensor data, Pictures, Videos etc. So, we need a system that will be capable of extracting essence of information available and that can automatically generate report.

* Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implict, Previously Unknown and Potentially useful Info from datastored in databases

Steps in KDD Process:



1. Datacleaning: Datacleaning is defined as removal of noisy and irrevelant data from collection.

- cleaning in case of missing values
- cleaning noisy data, where noise is a random or variance error.

2. Data Integration: Data Integration is defined as heterogeneous data from multiple Sources Combined in a Common Source

- Data Integration Using Data Migration tools.
- Data Integration Using Data Synchronization tools

3. Dataselection: Dataselection is defined as the Process where data relevant to the analysis is decided and retrived from the data collection.

- Data Selection using Neural Network
- Data Selection Using Decision Trees.

4. Data Transformation: It is defined as the Process of Transforming data into appropriate form required by mining procedure. Data Transformation is a two Step Process

- Data Mapping
- Code generation

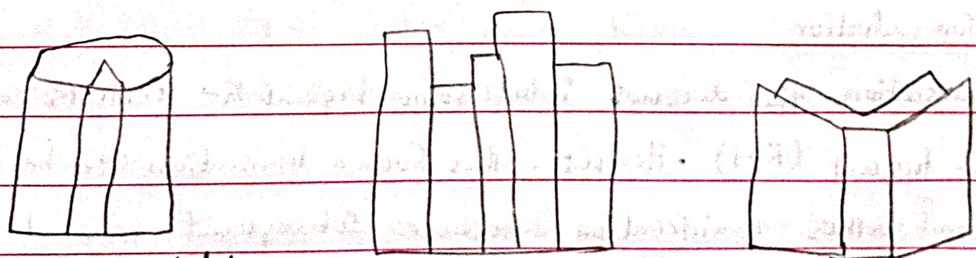
5. Data Mining: It is defined as clever techniques that are applied to extract patterns potentially useful.

6. Pattern Evaluation: Pattern Evaluation is defined as identify strictly increasing Patterns representing knowledge based on given measures.

7. Knowledge representation: Knowledge representation is defined as technique which utilizes Visualization tools to represent data Mining results.

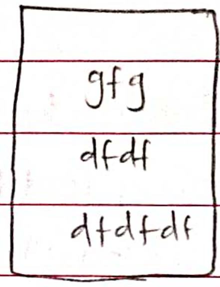
2) List out issues in Data Mining and Data Mining task Primitives.

* Data Mining Task can be Speified in the form of datamining Query which
* A data mining query is defined in terms of datamining task Premitives.
* These Primitives allow the user to interactively Communicate with the data mining system during discovery to diret the Mining Process or examine the findings from different angles or depths.
* The data Mining Primitives Specify the following
    1. Set of task-relevant data to be Mined.
    2. Kind of Knowledge to be mined.
    3. Background Knowledge to be Used in the discovery Process
    4. Interestigness measuses and thresholds for Pattern Evaluation
    5. Repersentation for Visualzing the discovered Patterns.
* A data mining Query language can be designed to incorporate these Primitives, allowing Users to interact with data Mining Systems flexibly. Having a data mining Query Language Provides a foundation on which -user friendly graphical interfaces Can be built.



Task-relevant data
Knowledge type to be mined characterization
Background Know-ledge

Pattern Interestingness measures simplicity.



Visualization of discovered Patterns, Rules, tables, reports, charts, graphs.

3) Explain improving of the effiency of Apriori Algorithm.

These are some variations of the Apriori algorithm that have been projected that target developing the efficiency of the original algorithm which are as follows

1) The hash-based technique :

A hash-based technique can be used to decrease the size of the Canditate K-itemsets, CK, for K>1. for instance, when scanning each Transcation in the database to create the frequent 1-itemsets, L1 from the Canditate 1-itemsets in C1, it can make some 2-itemsets for each transcations, hash them into several buckets of a hash table structure, and increase the equivalent bucket counts.

2) Transcation reduction :

A transcation that does not include some frequent K-itemsets cannot include some frequent (K+1)-itemsets. Thus such a transcation can be masked or deleted from further consideration because subsequent scans of the database for j-itemsets, where j>K will not need it.

3) Partioning: - A partioning techinque can be used that needed 2 database scans to mine the frequent itemsets. It includes 2 phases involving in phase 1, the algorithm subdivides the transcations of D into n non-overlapping parisitions. If the minimum support count for a parition is min_sup x the number of transcations in that parition.

4) Sampling: The fundamental idea of the sampling is to select a random samples of the given data D, and then search for frequent itemsets in S rather than D. In this method, it can trade off some degree of accuracy aganist effiency. The sample size of S is such that the search for frequent itemsets in S can be completed in main memory, and therefore only one scan of the transcations in S is needed overall.

5) Explain naive Bayes algorithm with an example.

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems

* The Naive Bayes algorithm is comprised of 2 words Naive and Baye's, which can be described as:

* **Naive:** It is called Naive because it assumes that the occurence of a certain feature is independent of the occurence of other features.

* **Bayes:** It is called Bayes because it is depends on the principle of Baye's Theorem.

Baye's Theorem:

* Baye's Theorem is also known as Baye's Rule or Baye's law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability

* The formula for Baye's theorem is given as

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

where,

* $P(A|B)$ is Posterior Probability
* $P(B/A)$ is Likelihood Probability.
* $P(A)$ is Prior Probability
* $P(B)$ is Marginal Probability.

Working of Naive Baye's classifier:

1. Convert the given dataset into frequency tables.
2. Generate likehood table by finding the Probabilites of given features.
3. Now, Use Bayes theorem to calculate the Posterior Probability.

Problem: If the weather is Sunny, then Player Should Play or not?

Solution: To Solve this, first Consider the below dataset.

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes

8	Rainy	No
9	Sonny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Frequency table for the weather conditions

Weather

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likelihood table Weather Condition:

Weather	No	Yes	
Overcast	0	5	5/14 = 0.35
Rainy	2	2	4/14 = 0.29
Sunny	2	3	5/14 = 0.35
All	4/14 = 0.29	10/14 = 0.71	

Applying Baye's Theorem:

$$P(Yes/sunny) = \frac{P(Sunny/Yes) * P(Yes)}{P(Sunny)}.$$

$P(Sunny/yes) = 3/10 = 0.3$

$P(Sunny) = 0.35$

$P(yes) = 0.7$

So, $P(Yes/Sunny) = 0.3 * 0.71 / 0.35 = 0.60$

$$P(No/Sunny) = \frac{P(Sunny/No) * P(No)}{P(Sunny)}$$

$P(Sunny/No) = 2/4 = 0.5$

$P(No) = 0.29$

$P(Sunny) = 0.35$

$$So, P(No/Sunny) = \frac{0.5 * 0.29}{0.35} = 0.41$$

So we can see from above Calulation that $P(Yes/Sunny) > P(No/Sunny)$.

Ans FP-Growth Algorithm is Frequent pattern growth algorithm to improve apriori method. FP growth algorithm represents the database in the form of a tree called as FP Tree.

FP Tree: It is a tree like data structure that is made with initial itemsets of the database.

The main purpose is to mine the most frequent pattern.

1. Count of each item

Item	Count
I_1	2
i_2	2
i_3	5
i_4	5
i_5	5
i_6	5
i_7	4
i_8	2

2. Sort the itemset in descending order

	Item	~~min~~ support count
✓	i_3	5
✓	i_4	5
✓	i_5	5
✓	i_6	5
✓	i_7	4

i_8	2
i_2	2
i_1	2

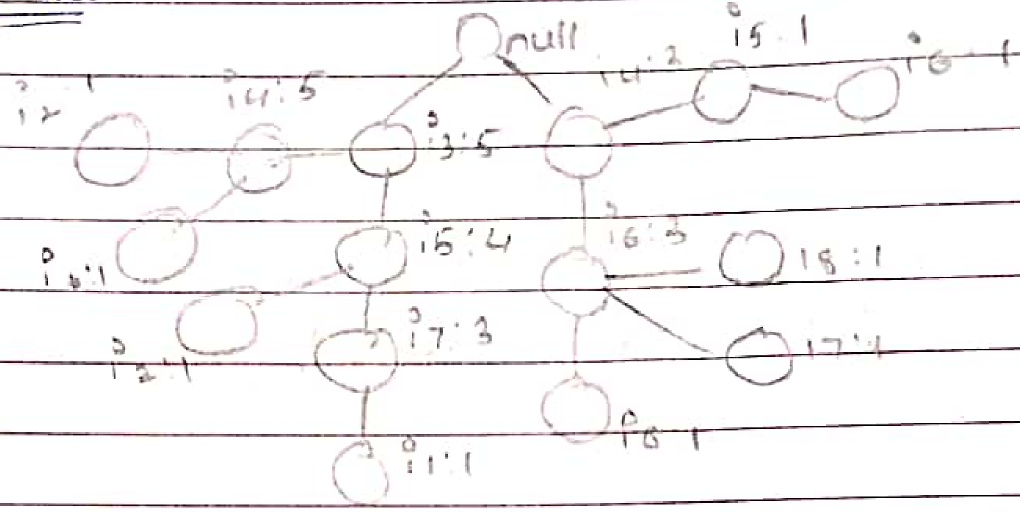
<u>FP-tree</u> A frequent pattern set (L) is built which contain all the elements whose frequency is greater than or equal to minimum support.

$L = \{ i_3 : 5,\ i_4 : 5,\ i_5 : 5,\ i_6 : 5\ ;\ i_7 : 4 \}$

Transaction ID	Items	Ordered - Item set
I_1	$\{i_3, i_5, i_7\}$	$\{i_3, i_5, i_7\}$
I_2	$\{i_4, i_6, i_8\}$	$\{i_4, i_6, \cancel{i_8}\}$
I_1	$\{i_3, i_5, i_7\}$	$\{i_3, i_5, i_7\}$
I_9	$\{i_7, i_5, i_1\}$	$\{i_5, i_7\}$
I_2	$\{i_4, i_6, i_7\}$	$\{i_4, i_6, i_7\}$
I_1	$\{i_2, i_3, i_4\}$	$\{i_3, i_4\}$
I_3	$\{i_4, i_5, i_6\}$	$\{i_4, i_5, i_6\}$
I_7	$\{i_8, i_6, i_1\}$	$\{i_6\}$
I_8	$\{i_5, i_3, i_2\}$	$\{i_3, i_5\}$
I_9	$\{i_3, i_4, i_6\}$	$\{i_3, i_4, i_6\}$

Inserting each element or item into tree

FP-tree



5Q Explain naive Bayes algorithm with an example.

Ans Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

* The Naive Bayes algorithm is comprised of 2 words Naive and Bayes.

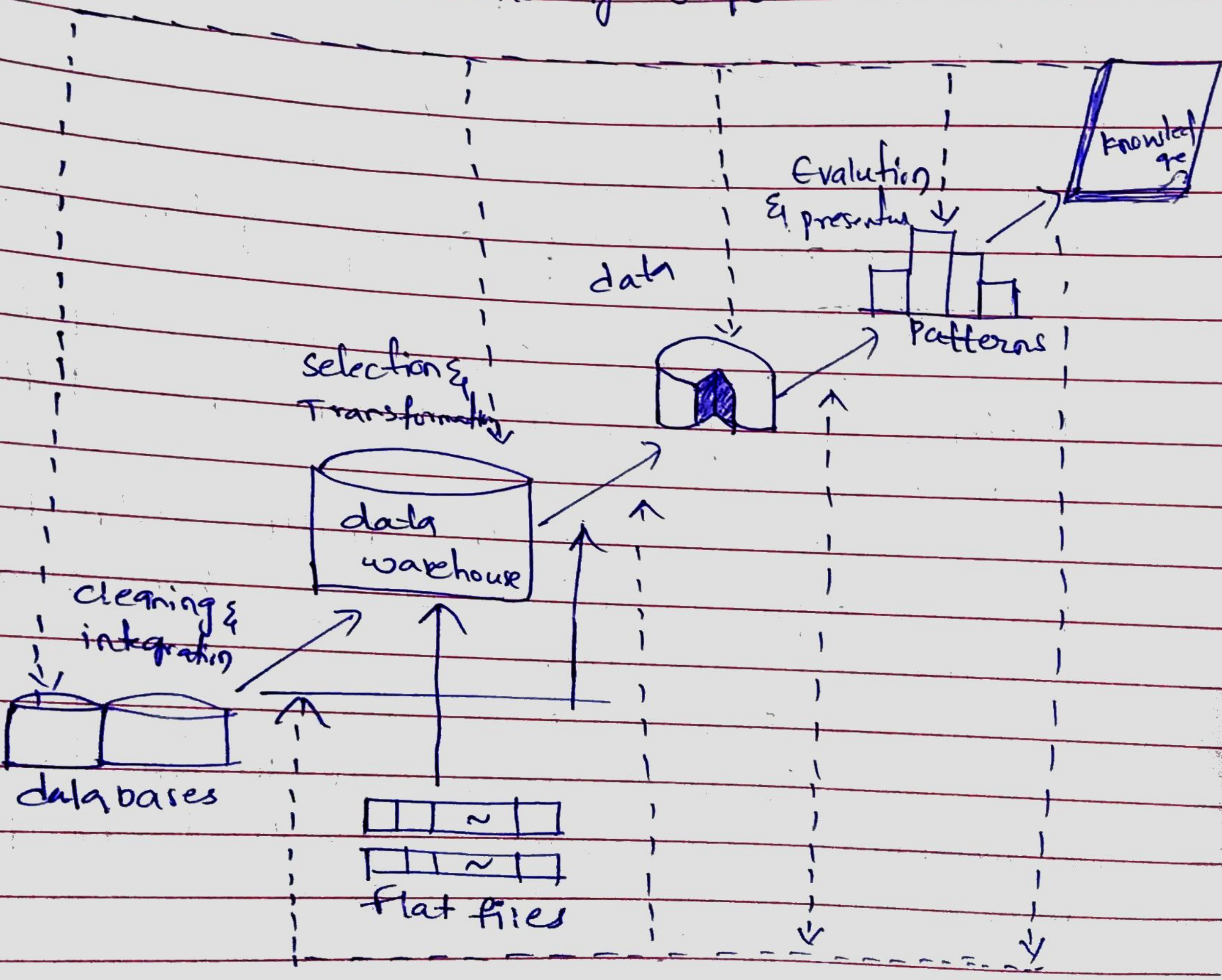
Baye's Theorem

It is used to determine the probability of a hypothesis with prior knowledge.
It depends on conditional probability

I - ASSIGNMENT
(Start Writing From Here)

① What is KDD? Explain KDD steps with a neat diagram.

Ans Data Mining is a fast growing field also known as knowledge discovery from data, or KDD or knowledge mining from data, knowledge extraction data/pattern analysis, data archaeology and data dredging
The knowledge discovery process is an iterative sequence of the following steps:



1. Data cleaning
To remove noise and inconsistent data
2. Data integration
Where Multiple data sources may be combined.
3. Data selection
Where data relevant to the analysis task are retrived from the data base
4. Data Transformation.
Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations
5. Data Mining
An essential process where intelligent Methods are applied to extract data patterns
6. Patteren evolution
To identify the truly intresting patterns representing knowledge based on intrestingness meausures.
7. Knowledge presentation
Where visualization and knowledge representation techniques are used to present mined knowledge to users.

(2) Write a short notes on
a) Data Cleaning
b) Data Transformation

An

a) Data Cleaning

it is a crucial process in data mining. it carries an impotant part in the building of a model.

Data cleaning can be regraded as the process needed, but every often neglects it.
Data Quality is the main issue in quality information Management.
Data quality problems occur anywhere in information systems.
These problems are solved by data cleaning.
Data Cleaning is fixing or remoring incorrect, corrupted, incorrectly formatted duplicate, or incomplete data with a dataset.

Steps of Data cleaning

1) Remove duplicate
2) Fix structural errors
3) filter unwanted outliers
4) Handle Missing data

5) Validate and QA

Methods of data cleaning

1) ignore the tuples
2) Fill the Missing value
3) Binning
4) Regression
5) Clustering

Characteristics of data cleaning

→ Accuracy
→ Coherence
→ Validity
→ uniformity
→ Data verification
→ Clean data back flow

b) Data Transformation

it is the process of converting data from one format to another, typically from that format of a source system into the required format of a destination system.

CMR GROUP OF INSTITUTIONS

Data Transformation is a component of most data integration and data Management tasks, such as data wrangling and data warehousing. The data are Transformed in ways That are ideal for mining the data.
The data transformation involves steps that are:

1. smoothing
2. Aggregation
3. Discretization
4. Attribute construction
5. Generalization
6. Normalization
   i) Min-Max Normalization
   ii) Z-score Normalization.
   iii) Decimal scaling.

③ Illustrate with an A-priori Algorithm for the given dataset below:

TID	list of items
001	Milk, dal, sugar, bread
002	Dal, sugar, wheat, jam
003	Milk bread curd Panner
004	wheat, Panner, dal sugar
005	Milk, Paneer, bread

006 wheat, dal, Paneer, bread.

Ans: 1. Generating Ist frequent itemset pattern

items	support count
Milk	3
dal	4
sugar	3
paneer	4
bread	4
wheat	3
Jam	1 x
curd	1 x

C1

1 or < 2 will be neglected.

items	support count
Milk	3
dal	4
sugar	3
Paneer	4
bread	4
wheat	3

L1

2) Generating 2nd frequent item set pattern.

items	support count
Milk, dal	1 X
Milk, sugar	1 X
Milk, Paneer	2
Milk, bread	3
Milk, wheat	0 X
dal, sugar	3
dal, paneer	2
dal, bread	2
dal, bwheat	3
sugar, paneer	1 X
sugar, bread	1 X
sugar, wheat	2
paneer, bread	3
Paneer, wheat	2
bread, wheat	1 X

C1

1 or < 2 will be neglected

items	support count
Milk, Paneer	2
Milk, bread	3
dal, sugar	3
dal, Panner	2
dal, bread	2
dal, wheat	3
sugar wheat	2
Panna bread	3
Panner wheat	2

L_1

3) Generating 3rd frequent item set pattern.

items	support count	
Milk, Paneer, bread	2	
Milk, Paner, dal, sugar	0	x
Milk, Panner, dal	0	x
Milk, dal, panna, bread	0	x
Milk, panna, dal, wheat	0	x
Milk, Panna, sugar, wheat	0	x
Milk, panna, bread	2	
Milk, panner wheat	0	x
Milk, bread, dal, sugar	1	x
Milk, bread, dal, panner	0	x

Milk, bread, dal	1	x
Milk, bread, dal, wheat	0	x
Milk, bread, sugar, wheat	0	x
Milk, bread, panner	2	
Milk, bread, wheat, panner	0	x
dal, sugar, paneer	1	x
dal, sugar, bread	1	x
dal, sugar, wheat	2	
dal, sugar wheat	2	
dal, sugar, panner, bread	0	x
dal sugar panner wheat	1	x
dal, panner, bread	1	x
dal, panner, wheat	2	
dal, panner, sugar wheat	1	x
dal, panner bread	1	x
dal, panner, wheat	2	
dal, bread, wheat	1	x
dal bread, sugar, wheat	0	x
dal bread panner	1	x
dal bread panner, wheat	1	x
dal wheat, sugar	2	
dal, wheat, panner bread	1	x
dal, wheat panner, wheat	2	
sugar wheat, panner, bread	0	x
sugar wheat panner	1	x

CMR

{ panner, bread, wheat } : 1) × }

C1

1 or < 2 will be neglected

items	support count
Milk, panner, bread	3
dal, sugar, wheat	3
dal, panner, wheat	3.

dal, panner, wheat is the most assosiative frequent item set

(4) What is correlation Analysis? Explain correlation Analysis from Association Analysis?

A Correlation analysis is a statistical method used to measure the strength of the linear relationship between two variables and compute their association.

correlation Analysis calculates the level of change in one variable due to the change in the other.

A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.

Types of correlation Analysis in Data mining.

1) Pearson correlation

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

2) Kendall rank correlation

$$T = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

N_c = No. of concordant

N_d = No. of dicordant

3) Spearman rank corelation.

$$P = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

P = Spearman rank correlation

di = the d/w b/w the ranks of corresponding variables

n = no. of observations

5. Explain Decision tree induction algorithm for classification. Discuss the usage of information gain.

A. Decision Tree is a supervised learning method used in data mining for classification and regression methods. it is a tree that helps us in decision-making purposes. The decision tree creates classification or regression as a tree structure. it separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed.

→ The final tree is a tree with the decision nodes and leaf nodes

→ A decision node has at least two branches

→ The leaf nodes show a classification or decision.

Key factors

Entropy:-

it refers to a common way to measure impurity. in the decision tree, it measures the randomness or impurity in data sets.

high entropy
↓
less random
dataset
↓
Less entropy

information Gain:

Information Gain refers to the decline in entropy after the dataset is split. it is also called Entropy Reduction.

Building a decision tree is all about discovering attributes that return the highest data gain

The decision tree useful for:

→ It enables us to analyze the possible consequences of a decision throughly
→ it provides us a framework to measure the values of outcomes and the probability of accomplishing them.
→ it helps us to make the best decisions based on existing data and best speculations

Advantages of using decision trees:

→ A decision tree does not need scaling of information

→ Missing values in data also do not influence the process of building a choice tree to any considerable extent.

→ A decision tree model is automatic and simple to explain to the technical team as well as stakeholders

→ Compared to other algorithms, decision trees need less extertion for data preparation during pre-processing

→ A decision tree does not require a standardization of data.