Preparatory Exam

(1) Subject: Data mining

Semester: 5th sem.

STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

I Answer any five 5×1=5

1. Define data Mining.

It is a process of discovering (or) Mining (or) Extracting the knowledge from large amount of data.

2. What is OLTP?

It is a database which is used to store day to day transaction.

3. Define Prediction.

Prediction is a set of techniques used for forecasting the future.

4. Mention any two application of Data Mining.

* Sales and Marketting
* Education
* Health care
* Banking

5. State Baysian Theorem.

Baye's Theorem is named after Thomas Bayes.

According Bay's Therorem $P(H/x) = P(x/H) \cdot P(H) / P(X)$

where, H → Hypothesis

x → data tuple.

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

6. Mention different Smoothing technique.

* Average
* Exponential function
* moving average

Binning method:- (3M) data are divided into smaller parts.

(2)

3x3=9.

II Answer any three

7. List any 3 differences b/w^n data mining and data ware house.

Data mining	Data warehouse
Data is retrived / analysed regularly	* Data is stored periodically.
Extract usefull pattern/ Knowledge from huge amount of data	* Integrate data from various source
It is focused on discovery of patterns	* It is focused on storage medium
Must take place after data warehouse	* Must take place before data mining
Used to identify the behaviour of perticular extracted information	* It is used identify the overall behaviour of the system

8. Explain Associative classification with an example

An associative classifier is a kind of supervised learning model that user association rules to assign a target value.

* Associative classifier is a promising data mining approach that integrates classification and association rule discovery to build classification model.

* An association rule has 2 parts:
  antecedent (if)
  consequent (then)

Eg: Market Basket analysis: It is the first application this rule
If a person buys bread, 70% likely of buying milk.

Here, bread → antecedent Milk → consequent

STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

(3)

9. Define Machine Learning. List its categories.

Machine learning application of artificial Intelligence that provides System ability to automatically learn and improve from experience without being explicitly programming.

Categories are : (1) Supervised Learning
(2) Unsupervised Learning
(3) Reinforcement Learning
(4) Semi Supervised Learning.

10. What is confusion matrix ? Explain.

Transformation about ~~do~~ correctly and incorrectly classified test records is tabulated in "Confusion Matrix".

Format of confusion matrix:

		Predicted class	
		Positive class 1	Negative class=0
Actual class	class=1 True	f(1,1) T.P	f(1,0) F.N
	class=0 False	f(0,1) F.P	f(0,0) TN

f(1,1) → True positive is no of records from class 1 is correctly predicated as class 1

f(1,0) → False negative is no of records from class 1 incorrectly predicated as class 0

f(0,1) → False positive is no of records from class 0 is incorrectly predicted as class 1

f(0,0) → True negative is no of records from class 0 is correctly predicted as class 0.

Inside
Silk.

(91) 376-18.

III Answer any three.

11. Explain OLAP operations.

The OLAP operations are: (1) Roll-up
(2) Drill down
(3) Slice and dice
(4) Pivot (rotate)

(1) Roll-up: It performs aggregation on a data cube in any of the following way.

* By climbing up a concept hierarchy for a dimension
* By dimensional reduction.

Location	medal
Delhi	6
New York	2
Patiala	3
USA	5

→

Location	medal
India	8
USA	7

(2) Drill down: It is reverse operation of rollup. It uses either of the way:

* By stepping down a concept hierarchy for a dimension
* By introducing a new dimension

Location	medal
India	8
USA	7

→

Location	medal
Delhi	5
Newyork	2
Patiala	3
Losangeles	5.

ROLL NO. DATE

(5) Slice and dice

The slice operation selects one particular dimension from a given cube and provides a new sub-cube.

location	medal
Delhi	5
los angles	5

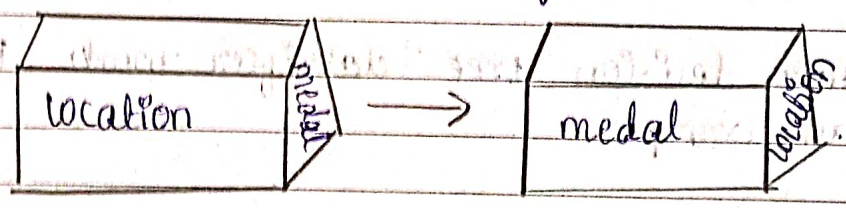
Location	Medal
New York	2
Patiala	3

Dice selects two or more dimensional from given cube and provides a new sub cube

location	medal
New York	2
Patiala	3

(6) Pivot

The pivot is also known as rotation. It rotates the data axis in view in order to provide an alternative presentation of data.



12. Explain Rule based classifier used to solve the classification problem.

Rule based classifier is a technique for classifying records using collection of IF-THEN Rules. we can express a rule in the following form.

R: IF condition THEN → conclusion

nano

Eg: R: If occupation = "Yes" and salary = "High"
Then Buys car = "Yes".

* If part of the rule is called as antecedent
* then part of the rule is called as consquent.

Characterstics:

i) Mutually exclusive rule:
It states that "every record is covered by at most one rule in the rule set".
The rules in a ruleset are mutually exclusive if no two rules in rule set are generated by the same record.

ii) Exhaustive rule:
A ruleset has exhaustive coverage if there is a rule for each combination of attribute values.
This property ensures that "every record is covered by atleast one rule in R".

13. How decision tree classifier works. Explain with an example.

A Decision Tree is a simple representation for classification. It is a supervised machine learning where the data is continuously split according to a certain parameter.

Decision Tree consist of:
* Nodes
* Edges / Branches
* Leaf nodes

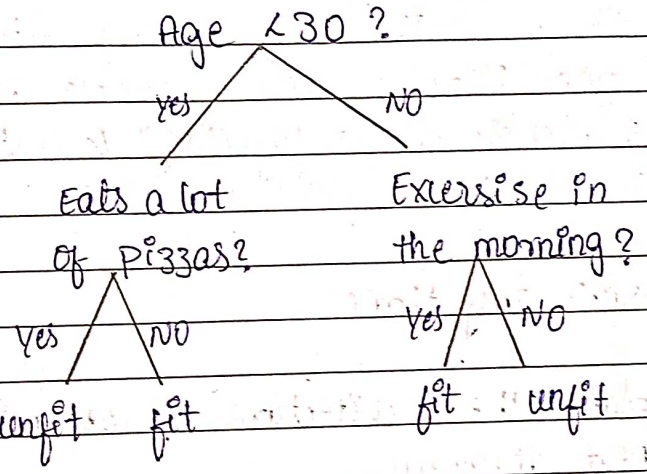
STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

* Nodes: It test the value of a certain attribute

* Edges/Branch: correspond to the outcome of a test and connect to the next node or leaf.

* leaf node: Terminal nodes that predict the outcome (it represent class label).

Eg: Is a person fit?



14. Name any 5 application area of data mining. Discuss in detail.

i) Sales/Marketting.
ii) Health care
iii) Education
iv) Banking
v) Retail industry
vi) Transportation
vii) E-Commerce
viii) Farming.

Health care: * Data mining is used for classification of diseas.
* It is also used to find the relationships bw^n disease and the effectiveness of treatements.

Education: * Data mining is used to predict student future by their learning behaviour.
* learning pattern of student can be captured.

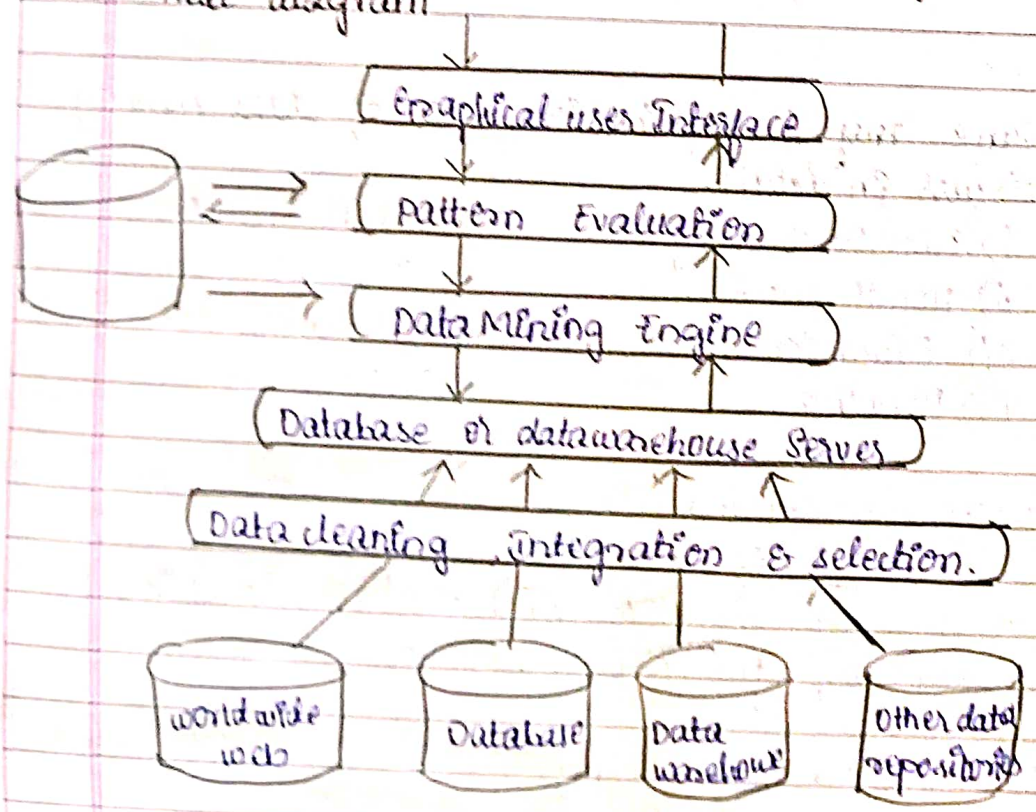
Banking: Banking and financial sector uses data mining for credit analysis, transaction analysis, cache management and to predict payments.

Transportation: Data mining helps determines the distribution schedules among warehouse and outlets and analyse loading pattern.

E-Commerce: Many E-commerce sites used data mining to offer cross-selling & upselling of the products. The shopping sites such as Amazon, Flipkart shows "people also viewed", "Frequently bought together" to customes who are interacting with the site.

IV. Answer any three 3X7=21.

15. Explain the architecture of datamining with neat diagram



STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

i) Data ware house:

Data ware house is a place which store information collected from the multiple source under unified schema.

ii) Database or Data warehouse Server:

The Server is the place that holds all the data which is ready to be processed. The fetching of data works upon the users request thus the actual datasets can be very personal

iii) Data mining Engine:

It is the core component of data mining process which consist of various modules used perform clustring, classification, prediction etc.

iv) Pattern Evaluation:

Pattern Evaluation is responsible for finding various patterns with the half of data mining Engine. Eg: pie-chart, bar-graph etc.

v) Graphical User Interface:

It provides communication b/w user and data mining system.

& it allows user to use system easily even if user don't have knowledge of system.

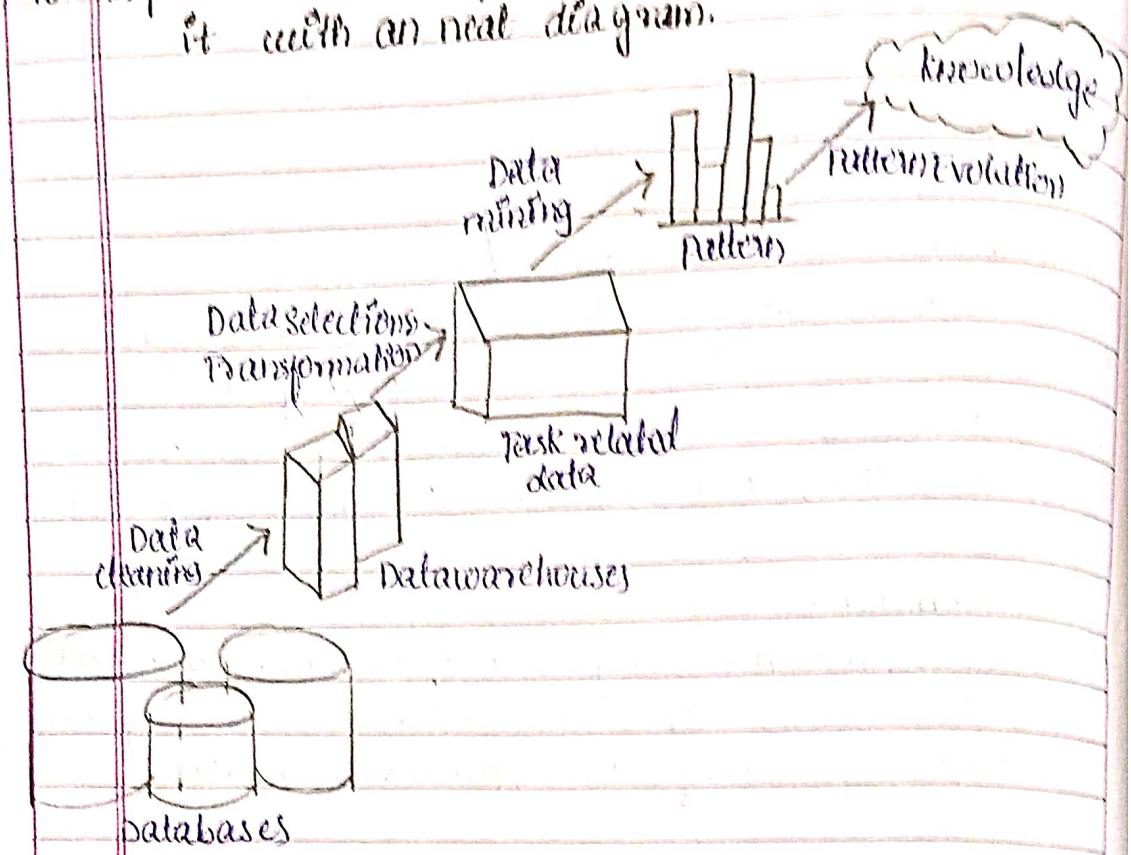
vi) Knowledge Base

It consist of Data that is very important in the process of data mining.

& It provides Input to the data mining Engine which guides data mining engine in process of pattern search.

(10)

16. Expand ~~Explain~~ KDD. Explain the steps involved in it with an neat diagram.



1) Data cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.
* noise data is random or variance error.

2) Data Integration: Data integration is defined as heterogeneous data from multiple source combined in a common source (Data warehouse)
* Data integration using data migration tools

3) Data selection: Data selection is defined as the process where data relevant to the analysis is decided and retrived from the data collection
* Data selection using Neural network

4) Data transformation: Data transformation is defined as the process of transforming data into appropriate

STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

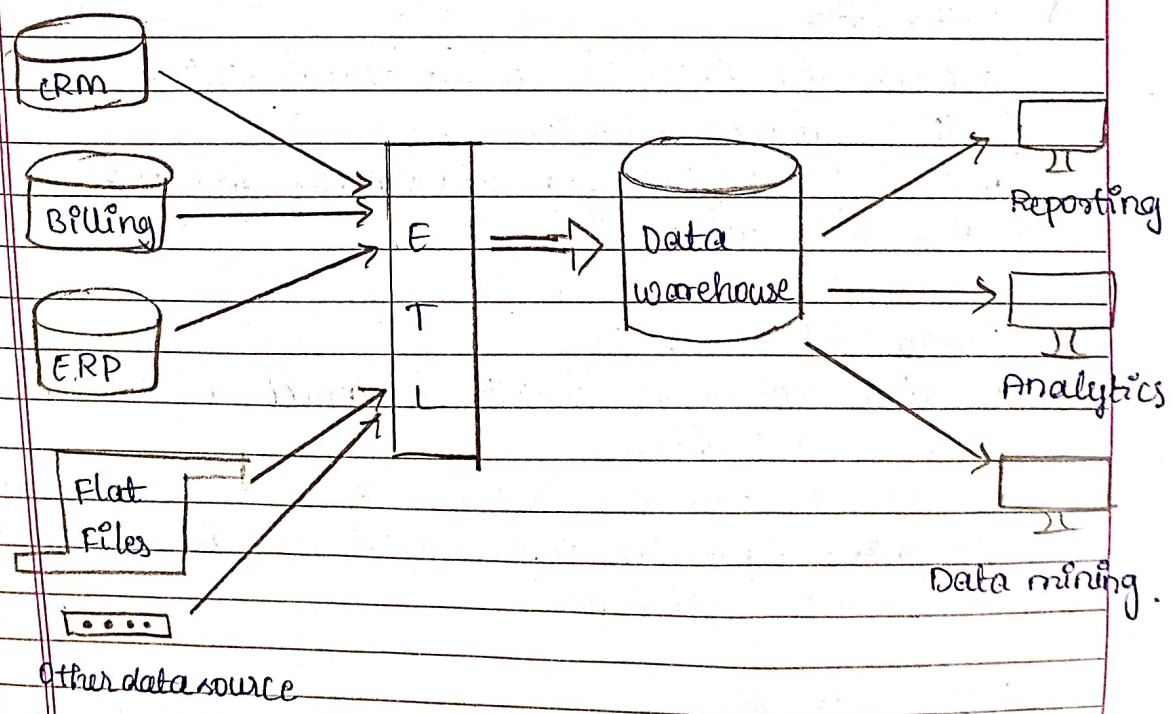
form required by mining procedure.
Data transformation has two steps:
① Data mapping ② Code generation.

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

6) Pattern Evaluation: pattern Evaluation is defined as identifying strictly increasing patterns ~~respondi~~ representing knowledge based on given measure.

7) Knowledge representation: knowledge representation is defined as technique which utilises visualization tool to present data mining result.

17. Explain Architecture of Datawarehouse with diagram.



(12)

There are mainly five component of data warehou

1) Data warehouse database:
* The central database is the foundation of the data warehousing environment.
* This database is implemented on RDBMS technology.
* This kind of implementation is constrained by the fact that traditional RDBMS systems in optimised for transactional database processing & not for data warehousing.

2) Sourcing, Acquisition, clean-up and Transformation tool
* The data sourcing transformation, and migration tools are performing all conversions.
* Summerization and all changes need to transform data unified formate in the datawarehouse.

3) Metadata
* metadata is data about data which defines the data warehouse.
* Used for building maintaining and managing the data warehouse

4) Query tool
* Query tool allows user to interact with the data warehouse system.
* One of the primary objective of data warehousing is to ~~make st~~ provide information to business to make strategic decision.

5) Data marts.
* A data mart is access layer which is used to get data out of the user. It is provided as option in large size data warehouse as it takes less time and money to build.

STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

8. For the given data set, apply Navie Baye's algorithm and predict the outcome for {yellow, Sweet, long}

Fruit	Yellow	Sweet	Long	total
Mango	350	450	0	650
Banana	400	300	350	400
Other	50	100	50	150
Total	800	850	400	1200

$$\text{Formula} = \frac{P(B/A) \cdot P(A)}{P(B)} = P(A/B).$$

i) Mango

a) yellow

A = Yellow B = Mango

$$P\left(\frac{\text{Yellow}}{\text{mango}}\right) = \frac{P(\text{mango}/\text{Yellow}) \cdot P(\text{Yellow})}{P(\text{mango})}$$

$$= \frac{\left(\frac{350}{800}\right)\left(\frac{800}{1200}\right)}{\frac{650}{1200}}$$

$$= \frac{350}{1200} \times \frac{1200}{650} = \frac{350}{650}$$

$$= 0.53$$

b) Sweet, A = Sweet B = mango

$$P\left(\frac{\text{Sweet}}{\text{mango}}\right) = \frac{P(\text{mango}/\text{sweet}) \cdot P(\text{sweet})}{P(\text{mango})}$$

$$= \frac{\left(\frac{450}{850}\right) \cdot \left(\frac{850}{1200}\right)}{\frac{650}{1200}}$$

nano

$$= \frac{450}{1200} \times \frac{1200}{650}$$

$$= \frac{450}{650} = 0.69$$

c) long

A = long B = Mango

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(B)}{P(A)}$$

$$= \frac{\left(\frac{0}{400}\right) \times \left(\frac{400}{1200}\right)}{\frac{650}{1200}}$$

$$= \frac{0}{1200} \times \frac{1200}{650} = \frac{0}{650}$$

$$= 0$$

ii) Banana

a) Yellow

A = Yellow B = Banana

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$= \frac{\left(\frac{400}{800}\right)\left(\frac{800}{1200}\right)}{\left(\frac{400}{1200}\right)}$$

$$= \frac{400}{1200} \times \frac{1200}{400}$$

$$= 1$$

STUDENT NAME		TOTAL MARKS
CLASS	SUBJECT	
ROLL NO.	DATE	

b) Sweet

A = Sweet B = Banana

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$= \frac{\left(\frac{300}{850}\right)\left(\frac{850}{1200}\right)}{\left(\frac{400}{1200}\right)}$$

$$= \frac{300}{1200} \times \frac{1200}{400}$$

$$= \frac{300}{400} = 0.75.$$

c) long A = ~~Sweet~~ long B = Banana.

$$= \frac{P\left(\frac{350}{400}\right) \cdot P\left(\frac{400}{1200}\right)}{P\left(\frac{400}{1200}\right)}$$

$$= \frac{350}{1200} \times \frac{1200}{400} = \frac{350}{400} = 0.87$$

iii) Other

a) yellow A = Yellow B = Other.

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$= \frac{\left(\frac{50}{800}\right)\left(\frac{800}{1200}\right)}{\left(\frac{150}{1200}\right)}$$

$$= \frac{50}{1200} \times \frac{1200}{150} = \frac{50}{150} = 0.33$$

nano

(16)

ii) Sweet A = Sweet B = Other

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$= \frac{\left(\frac{100}{\cancel{850}}\right) \times \left(\frac{\cancel{850}}{1200}\right)}{\left(\frac{150}{1200}\right)}$$

$$= \frac{100}{\cancel{1200}} \times \frac{\cancel{1200}}{150} = \frac{100}{150} = \underline{\underline{0.66}}$$

iii) long A = long B = Other

$$P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$= \frac{\left(\frac{50}{\cancel{400}}\right) \left(\frac{\cancel{400}}{1200}\right)}{\left(\frac{150}{1200}\right)}$$

$$= \frac{50}{\cancel{1200}} \times \frac{\cancel{1200}}{150} = \frac{50}{150} = \underline{\underline{0.33}}$$

Mango → $0.53 \times 0.69 \times 0 = 0$

Banana → $1 \times 0.75 \times 0.87 = 0.65$

Other → $0.63 \times 0.66 \times 0.33 = 0.071$