challenge
pondering
victorious
thrilling

Saathi

Breezed
Suppose

Page No. :
Date :

Unit → 1

⇒ The two
Main task → The two "high-level" primary goal of data mining is Prediction and description.

☆ Data Mining :-

- Data Mining is mining knowledge from data.
- It is define as a process used to extract usable data from a larger set of any raw data.
- It is implies analysis data pattern in large batches of data using one or more s/w. data mining is also known as knowledge discovery in data (KDD).

☆ Advantage and Disadvantage of Data Mining :-

* Advantage :-

1. Data Mining technology help company to get knowledge based information.
2. Data Mining helps with the decision making process.
3. Data Mining helps organisations to make the profitable adjustments in operations and production.
4. It can be implement in new system as well as existing plateform.
5. It is a speed process which make to easy for the users to analysis huge amount of data in less time.
6. The data mining is a cost effective and efficient Solution compared to the other data applications.

Page No. :
Date :

* Disadvantages :—

1. There are chances of company many sales usefull information of their customer to their companies for Memory.
2. Many data mining analysis slw is difficult to operate and require advantage training to work on.
3. The data mining technology are not accurate. So it can cause serious Company Seq. in certain conditions.
4. Different data mining tools work in different manner due to different algorithum are designed.

★ Data Mining applications :—

Industry	Application
finance, insurance, Tele-communication, transport, consumer good, scientific research, utilities.	credit card analysis, claims fraud analysis, call record analysis, logistics mgmt, promotion analysis, Image, video, speech, power usage analysis.

1. Data Mining app. in sales and marketing.
2. Data Mining app. in Banking and finance.
3. Data Mining app. in Health care and insurance.
4. D.M.A in Transportations.
5. D.M.A in Medicines.
6. D.M.A in Educations.
7. D.M.A in Manufacturing engineering, research analysis.

challenge
pondering
victorious
thrilling

Saathi

Squeezed
Supper

8. D·M·A in fraud detection.
9. D·M·A in credit card spending by customer group can be identified by using Data Mining.
10. D·M·A in from historical market data.
11. D·M·A enables to identifies stock trading rules.
12. D·M is used to identifies customer loyalty by analysis the data of customer purchasing activity.

(a) Data Mining app. in Business app:-
database marketing is one of the most successful and popular business app. of d.M.

(b) Retail database contain customer shopping transaction.

(c) Using D·M technology investor can build models which can be used to predict the performance of stock.

(d) app. for credit or loan are decide base on the app. information.

→ Science app. → D·M technology have been used to medic biology many more.

→ other app. → Data Mining technology has also used man other areas, such as healthcare, sports.

☆ Data objects and attribute types:-

1. Data sets are made up of data objects.
2. A data object represent an entity in a sales databa
3. The object may be customer store items and sales

a medical database.
4. The object may be paitents.
5. In a university database the object may be students professors.
6. Data objects are typical describe by the attribute.
7. Data object are store in database.
8. They are data tuples row of a database corroshondence to the data objects and column corroshondence to the attribute. <ex→ array, pointers, records, files.

★ Attribute :-

- Attribute is a property of objects. the attribute represent the different features of the object.

Ex→

Roll NO	Name	Result
1	Muskan	Pass
2	Kusum	fail

★ Types of Attribute :-

1. Binary
2. Nominal
3. Numeric
* Internal - scaled
* Ratio - scaled

2. Nominal data :-

- It is in the alphabetical form and not in integer.

Ex→ Attribute → value
categorical → lectuar, Assistant, proffesor
data → proffessor
states → New, pending, working
colors → Black, Brown, Red

1. Binary Data :-

How only two values states :-

Ex→ Attribute → value
HIV detected → yes, no
Result → Pass, fail

☆ Binary attributes is of two types :-

1. Symmetric binary
2. Asymmetric binary

1. Symmetric binary :-

· Both values are equally important.

Ex→ Attribute → value
Gender → Male, female

2. Asymmetric binary :-

· Both values are not equally imp.

Ex→ Attribute → value

HIV detected → yes, no
Result → Pass, fail

☆ ordinal data:-

- All value has a meaningful order.
- That is call to the ordinal data.

Ex→ Attribute → value
Grade → A, B, C
BPS: Basic Pay Scale → 16, 17, 18

☆ Discreate Data:-

- This data have finite value.
- It can be in numerical form and can also be in categorical form.

Ex→ Attribute → value
Profession → Teacher, Doctor etc
Postal code → 42200, 42300 etc.

☆ Continous data:-

- This is technically have an infinite at steps.
- Continous data in float type there can be many number in b/w 1 and 2 consine.

Ex→ Attribute → value
Height → 5.4, - - - 6.5 etc
weight → 50-3, etc.

Page No. :
Date :

★ Recalling Mean :-

- Mean is the average of the number.

Ex→ 3, 5, 6, 9, 8
Mean = all value / Total no. of value
Mean = $\frac{3+5+6+9+8}{5} = \frac{31}{5} = 6.2$

Q. How to calculate the mean for data with frequencies :-

Ans.

Age	Frequency	Age * freq
22	5	110
33	2	66
44	6	264
66	4	+ 264
	17	704
Total		

Mean = $\frac{704}{17} = 41$

* Median :-

It is the middle value among all the values.

Q. How to calculate medium for an odd no. of values

Ex→ 9, 8, 5, 6, 7
Arrange value in order
~~3, 5, 6, 8, 9~~ 5, 6, 7, 8, 9
~~Median = 6~~ Median = 7

Page No.:
Date:

Q. How to calculate median for an even number of values?

Ex→ 9, 8, 5, 6, 3, 4
Arrange → 3, 4, 5, 6, 8, 9
Add two middle values and calculate their mean
median = $\frac{5+6}{2} = \frac{11}{2} = 5.5$.

* Mode :-
The mode is the most occuring value calculate the mode.

Ex→ 3, 6, 6, 8, 9
Mode = 6 because 6 is occuring two times and all other values occuring one time.

* The Weighted Arithmatic Mean :-

- It is the similar to an ordinary arithmatic mean (the most common type of average).
- Except the instead of each the data points contributing equally to the final average.
- Since, data point contribute more then others.

* The Arithmatic Mean :-

- when you find a mean for a set of numbers, all the numbers carry an equally weight.
for ex → If you want to find the arithmatic mean of 1, 3, 5, 7 and 10.

1. Add up your data points = 1+3+5+7+10 = 26
2. Divide by the number of items in the set $\frac{26}{5} = 5.2$

Page No.:
Date:

☆ Data Mining as a step of knowledge discovery process:-

- Data Mining as a steps of knowledge discovery process in database is the process of discovery a useful knowledge from a collections of data.
- This process is widely used data mining technology is a process that includes cleaning and the knowledge on the data set and accurate solution on the observed result.
- Major KDP application areas includes marketing, fraud detection, Telecommunication and manufacturing.

☆ The steps are involve in the entire KDP process are:-

1. Identify the goal of the KDD process from the customers perspective.
2. understand Application domain involved and the knowledge that's required.
3. select a target data set or subset of data samples on which discovery is be performed.
4. cleans and prepares data by deciding strategies to handle missing fields and alter the data as per the requirements.
5. Simplify the data sets by removing unwanted variables. Then, analyze useful features that can b used to represent the data depending on the goal o task.
6. Match KDD goals with data mining method to Sugg hidden pattern.

7. Interpret essential knowledge from the mined patterns.
8. use the knowledge and incorporate it into another System for future action.

steps → 1. Data cleaning → In this step, the noise and inconsistent data is removed.

2. Data integration → In this step, multiple data sources are combined.

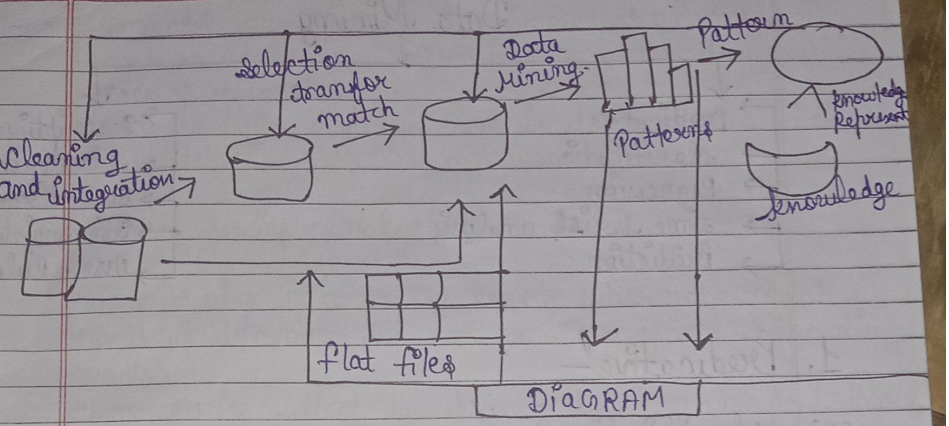
3. Data selection → Data selection is defined as the process where data relevant to the analysis is decided and retrieved from data collection.
- Data selection using neural network, Decision Tree, Naive bayes, clustering regression etc.

4. Data transformation → In this step data is transformed into appropriate forms for mining by performing summary operation.

5. Data Mining → Intelligent methods are applied in order to extract data patterns.

6. Data Evaluation → Data patterns are evaluate.

7. Knowledge Represent → Knowledge is represented.



DIAGRAM

* Data Mining tasks :-

The data mining task can be classify generally two categories :-

1. Descriptive task
2. Predicative task

1. Descriptive task :-

- The descriptive data mining task characterise the generally prospectives of data.

2. Predicative task :-

- Predictive data mining task perform inference on the available data set to predict how a new data set will be behave on work.

Page No. :
Date :

Data Mining

- Predicative
  - classification
  - Regression
  - time series analysis
  - Prediction
- Descriptive
  - clustering
  - Seq. discovery
  - Summarization
  - Asso. Rules

1. Predicative :-

(a) Classification :-

- classification derived a model to decide the class of an object base on its attribute.
- A collections of reports will be available and each record with a set of attributes.
- one of these attributes is a class.
- classification can be used direct marketing that is reduce.
- Marketing cast by target of set of customer.

Ex→ Pattern, recognization.

(b) Prediction :-

- Predicative task predict the possible value of missing or future data. prediction involves develop.
- This model is used in predicting future value of a new data set of itemset.

- A model can predict the income of an employ based on education, expression and other demographics place of stay etc.
- Also prediction analysis used in different are include fraud detection, Medical diganses.

(c) Time series Analysis :-

- Time series is a seq. of events where the next event is determine by one or more of the preceding events.
- Time series reflect the process being major and there are certain components that affect the behaviour of a process.
- Time series analysis include method to analysis time series data in order to extract useful patterns, trends and rules.
- Stock market prediction is an imp. app. of time series analysis.

(d) Regression :-

- In any regression tasks of supervised learning, the model learns to predict numeric scores.
  for ex → when an individual tries to predict the price of the stock in the coming days, gives the past history of the company and the market, it can be treated as a regression task.
- RMSE and quantiles error are the major evaluating metric for regression Quantile plots are used for univariate data distribution.

* RMSE → RMSE is defined as the square root of the average squared distance Gw.
- The actual score and the predicted score RMSE = $\sqrt{\Sigma i (y_i - \hat{y}_i)^2}$.
- where y_i is the true score for the ith data point, $\hat{y}_i$ is the predicted value, and n is the no. of data points.
- RMSE measures the standared deviation of the predictions from the ground truth.
- The RMSE is one way to measure the performance of a classifier.
- Error rate (or no. of mis. classification is another one.)

2. Descriptive task :-

1. clustering :-

- clustering is used to identify the data objects that are similar to one another the similarity can be decide based on no. of factors like such as behaviour geographical location response to contain action.

Ex→ An insurance company can cluster its custome based on age, residence, income, this group inf. will be helpful to understand the customer better and provide better customerized service.

→ clustering means grouping the object based on the inf. found in the data describe the object and their relationship.

→ The Goal is an object in a group will be similar and

one the other different object in other groups.
→ It works on the bottom up approach.

2. Summarization :-

- Summarization is the generations of data.
- A set of related data summarization which result in a small set that give aggregated inf. of the data.

Ex→ The shopping done by a customer can be summarized into total products, total spending offered used etc.
→ Such a high level summarized inf. sales can be use-full for sales or customer relationship team for detailed customer and the purchase behaviour analysis.
→ Data can be summarized in different abstraction levels and from different angles.

3. Association Rules (AR) :-

- AR are in the DM if then statements that help to show the probability of relationship blw data items with in the large data set in various types of database.
- Association rules mining has a no. of app. and it is widely used to help the discover sales co-relations in transaction data AR mining at a basic level.
- It involve the use of machine learning model to analysis the data for patterns.
- It identify frequent if then association which are called association rules.

* AR has two parts :-
1. Antecedent (if)

2. consequent (then)

1. A (if) → is an item found with in the data.
2. C (then) → is an item found combination with the antecedent.

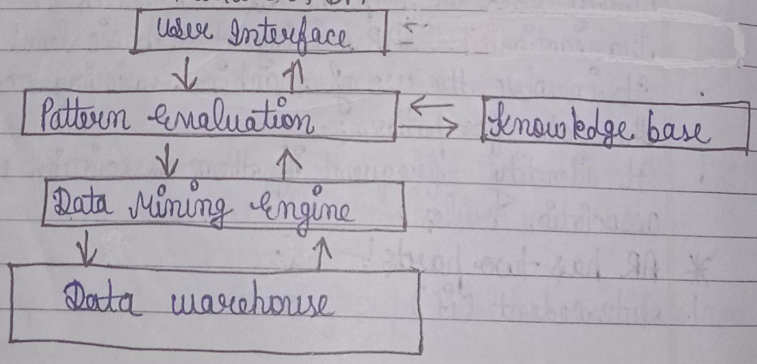
4. Seq. discovery :-

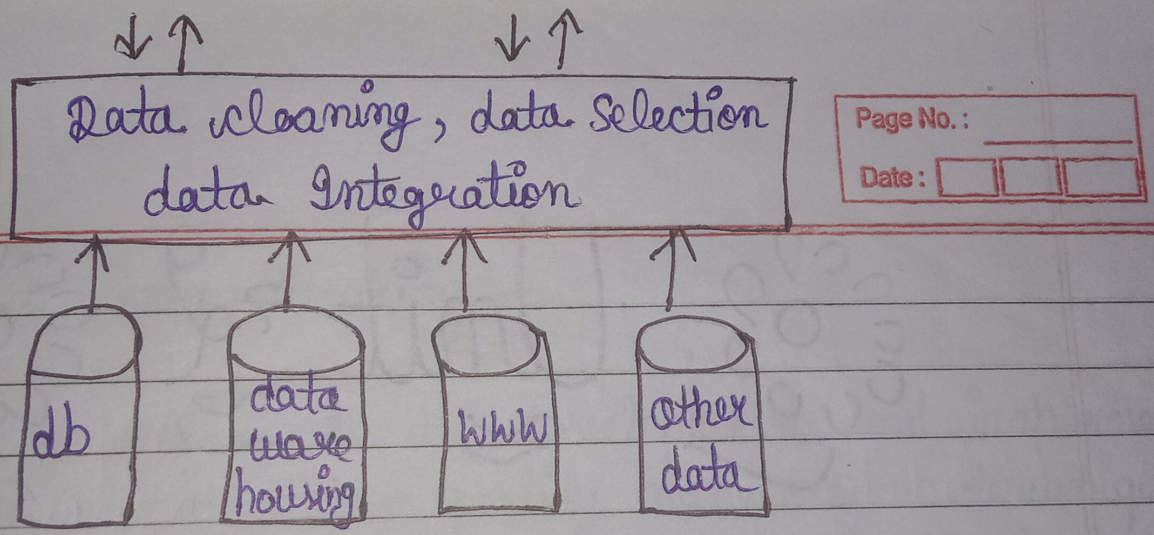
- Seq. discovery or Sequential pattern mining, is a data mining tech. that discovery statistically relevant patterns in sequential data.
- This mining program evaluates certain criteria, such as occurence frequency, duration or values in a set of sequences to find interesting hidden patterns.

☆ Use of Data Discovery :-

- The data miner simply provides a Seq. database and chooses a search parameter called the minimum support.
- This specifies the minimum no. of subsequences in which a pattern must appear to be considered relevant.
- The result are presented in a column format, with the no. of subsequent instances on the right.

Structure of DM




Data cleaning, data Selection
data Integration
db
data wave housing
www
other data