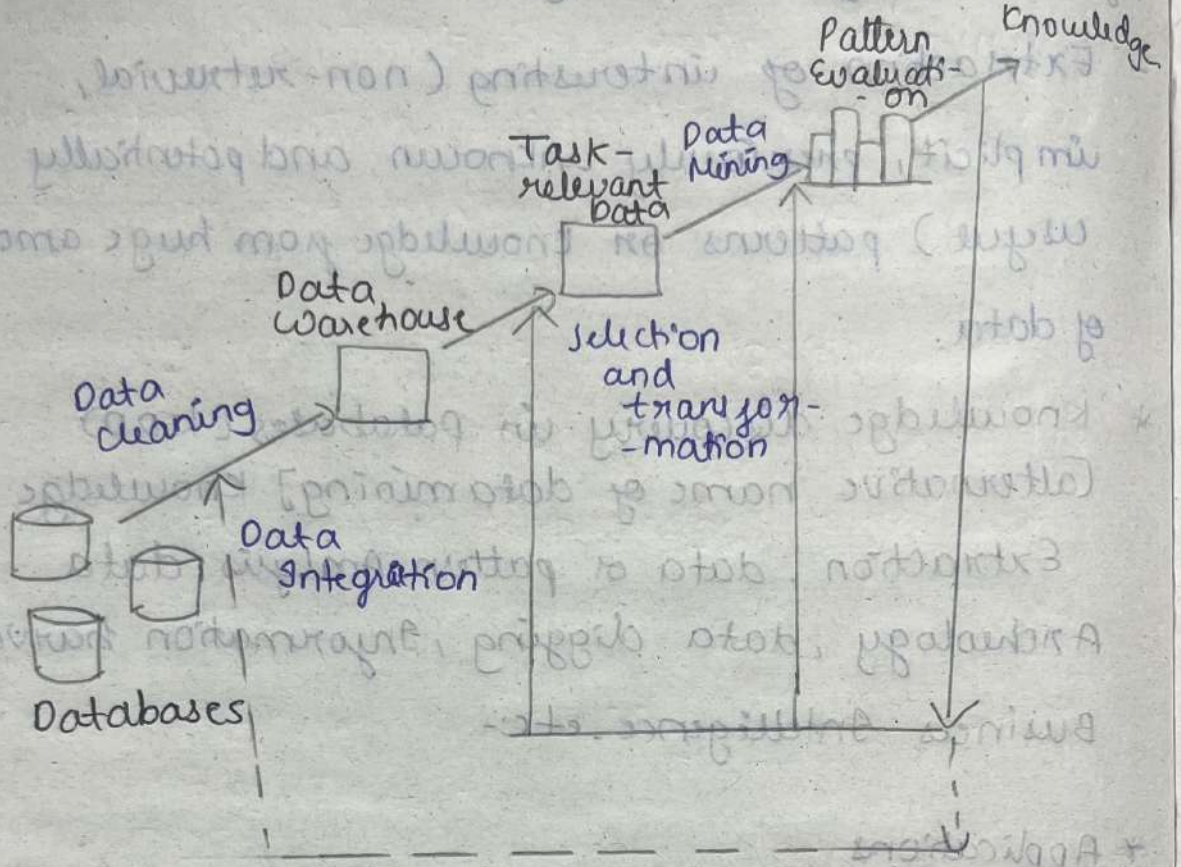
5/8/24

UNIT-I

* What is Data Mining :- Extracting information from huge amount of data (or)
- Extraction of interesting (non-retrivial, implicit, previously unknown and potentially usyul) patterns or knowledge from huge amount of data.
* Knowledge discovery in Databases (KDD) [alternative name of data mining] Knowledge Extraction, data or pattern Analysis, data Archeology, data digging, Information harvesting Business Intelligence. etc:-

* Applications
- Database Technologies
- pattern recognization
- Data visualization
- statistics
- Algorithms
- etc:-

* Data Mining - core of knowledge discovery process



* Data Mining on what kinds of data.

- Data warehouses
- Relational Databases
- Traditional Databases
- Advanced Database
- Information repositories
- Object Oriented and object Relational databases
- Spatial Database
- Time series data
- Temporal data
- Text database
- Multimedia database.

- heterogeneous and legacy databases
- World wide web, etc:-
- Data Mining Applications :-

*Customer segmentation :- Busines use data mining techniques to understand customers.

*Market basket Analysis :- In Retail out of 10 customers, if 8 are purchasing two items or both items then both will be kept near.

*Risk management :- In LIC's, Banks Loans.

* Fraud detection :- In Banks to take loan.

* Demand Prediction.

6/8/24

- KDD steps (Process)

- Data cleaning :- To remove noise and inconsistent data (may take 60% of effort).
- Data Integration :- when 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 appropriate for mining by performing summary or aggregation.

- Data Mining :- search for patterns of interest As essential process where intelligent methods are applied in order to extract data patterns.
- pattern evaluation :- To identify the truly Interesting pattern representing knowledge based on some interestingness measures.
- Knowledge presentation :- visualization and knowledge representation techniques are used to present the mined knowledge to the user.

7/8/24

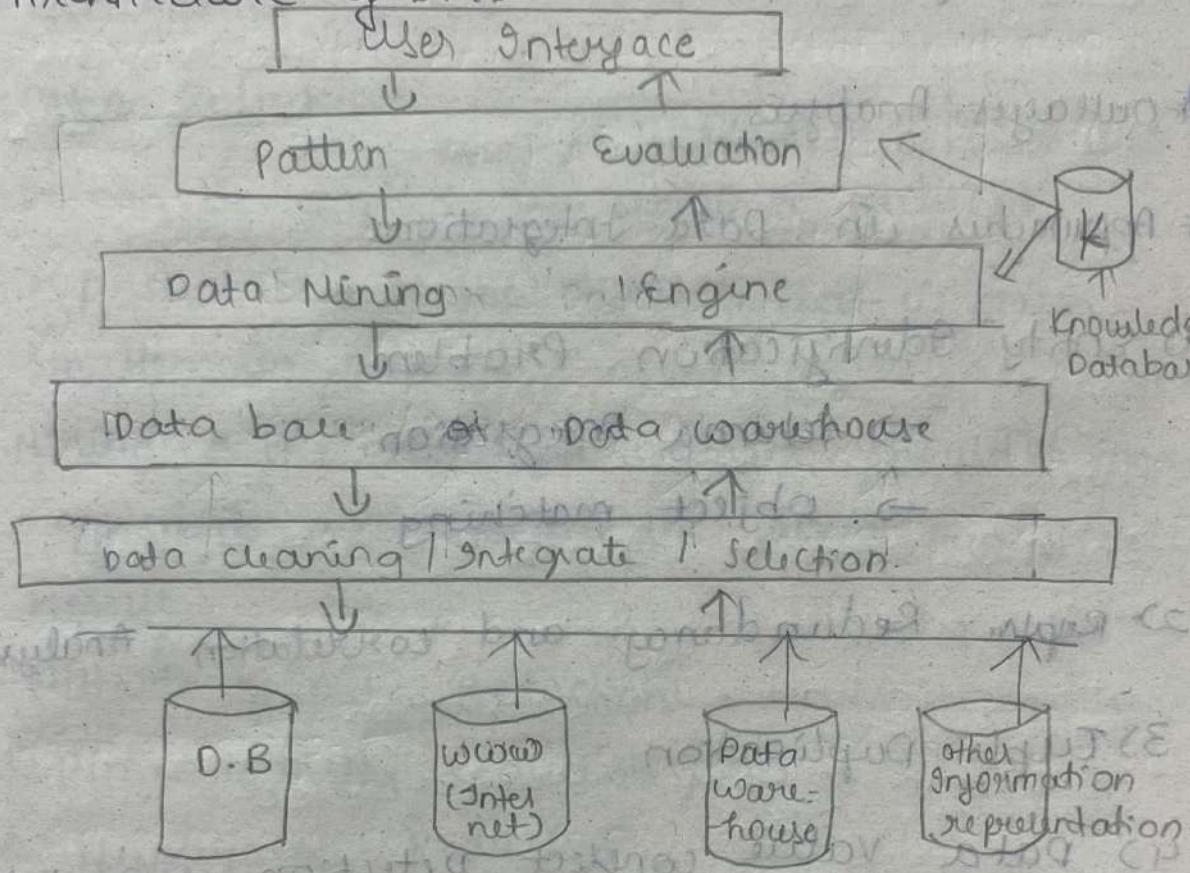
- Architecture of Data Mining
- Data Mining Tasks :-
- Predictive :-
- It is a supervised learning
- Predictive values of data by making use of known results from a different set of sample data.

Predictive
- Classification
- Regression
- Prediction
- Time series

- Descriptive :-
- Enables you to determine patterns and relationship in a sample data

Descriptive
- Association rules
- Clustering
- Sequence discovery
- Summarization.

- Architecture of DM:-

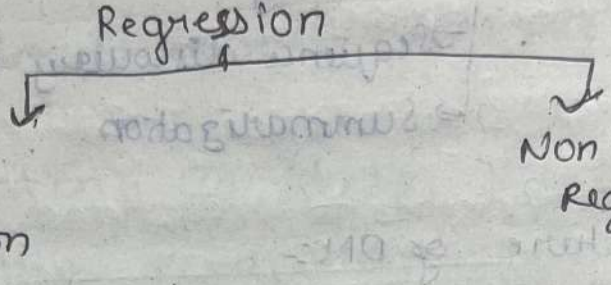


* Data cleaning :- Whatever the noisy data that cleans by using filling the missing values smoothening noisy data, resolving the inconsistency and removing the out layers.

* Binning :- Partitioning into equal sets, smoothening the bins by replacing by mean or average of null values. & mothing by bin boundaries

12/8/24

2) * Regression :-



3) * Outlayes Analysis

* Approaches in Data Integration.

1) Entity Identification Problem
   → Schema Integration
   → object matching

2) Replace Redundancy and correlation Analysis

3) Tuple Duplication

4) Data value conflict Detection of Resolution.

* Data Integration :-
- Merging of data collected from multiple sources, careful integration can help reduce redundancies, and in-consistencies in the resulting dataset.

- Tuple duplication.
- The use of denormalized table (often done to improve peryormance by avoiding joins)

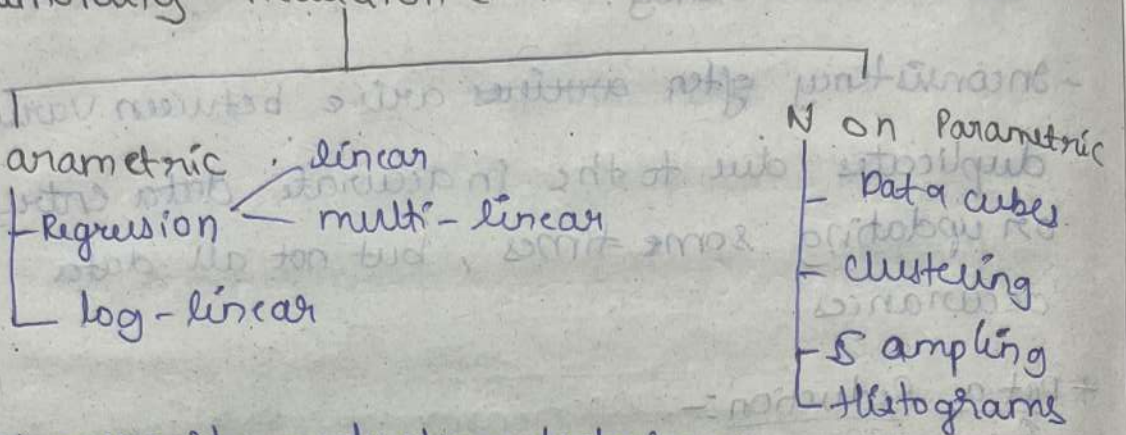
- Denormalized table is another source of data redundancy.
- Inconsistency often arise between various duplicates due to the inaccurate data entry or updating some times, but not all data occurances

* Data Reduction :- 1) Dimensionality 2) Numerosity
- It can be applied to obtain a reduced representation of the dataset, that is much smaller in volume.
- Mining on the reduced dataset should be more efficient and produce the same Analytical result.
- Methods of Data Reduction :-
1) Dimensionality Reduction (DR)
2) Numerosity Reduction
3) Data compression.

* DR :- It eliminates the redundant attributes which are weakly important across the data.
1) Step wise ~~pivot~~ forward selection
2) Step wise backward elimination
3) Decision tree induction.

13/8/24

* Numerosity Reduction (NR)



* Numerosity reduction techniques
  1) Parametric 2) Non Parametric.

* Parametric methods:- In this data is represented using some model.
- In this, model is used to estimate the data so that only parameters of data are required to be stored instead of actual data.
- The two different methods are:-
  1) Regression
  2) log-linear method.
- Used for creating such models
- Regression:- Regression can be a simple linear regression or multi-linear regression, when there is only single independent attribute. such model is called simple linear regression.
- In this the data are modelled to fit a straight line.

* Multi- linear Regression :- their are multiple Independent attributes used such regression models are called Multi-linear Regression.

- log- linear model :- In this model used to estimate the probability of each data point in a multi-dimensional space, for a set of discrete attributes based on a smaller set of dimensional combinations

* Non Parametric methods :- This methods are used for storing reduced representations of data include Histograms, clustering, sampling and data cube aggregation.

- Histogram :- This is the data representation in terms of frequency.
- clustering :- This technique partitions the whole data into different groups or clusters.
- Sampling :- This can be used for data reduction because it allows a larger dataset to be reduced? represented by a much smaller random data sample. (Subset of original data).
- Data Cube aggregation :- It involves moving the data from detail level to a fewer number of dimensions. The resultant data set is smaller in volume without loss of info, necessary further analysis tasks.

* Data compression:- In this modification, encoding or converting, the data structure of data in a way that consumes less space.
- we can divide it into a two types based on their compression techniques 1) loss less compression 2) lossy compression.

1) loss less compression:- Encoding techniques allow a simple and minimal data size reduction.
- It uses algorithms to restore the precise original data from the compressed data.

2) lossy compression:- the compressed data may be differ to the original data, but are useful enough to retrieve info from them.
- They are 1) Discrete wavelet transforms.
2) Principle component analysis.

14/8/24

- Discrete wavelet transformation:- This is a signal processing technique that transforms linear signals.
- the wavelet transforms the data can be concated and this is helpful in data reduction.

- If we store a small fraction of a strongest wavelet coefficient then the compressed approximation of the original data can be ~~uptime~~ obtain.
- Example :- An Image of size 100 MB compressed to 100 KB. Here picture quality decreases.
- Principle component analysis (PCA) :- In this analysis is a
- This method extracting the important variables from a large number of variables available in a data set, It extracts a set of low dimensional features from a high dimensional dataset with a goal of capturing as much information as possible in data.

* Steps involved in

1) Standardize the dataset
2) compute the covariance matrix for features in data sets.
3) Compute the Eigen values and Eigen vectors for the covariance matrix.
4) Sort the Eigen values and their corresponding Eigen vectors. choose K Eigen values to form an Eigen vector matrix.
5) Transform the original matrix