Assignment No:-2

Page No.
Date

- Aim :- Perform data preprocessing tasks & Demonstrate following association rule mining on data sets.

- Course outcomes :-

CO2 :- Demonstrate the classification, clustering & etc. in large data sets.

CO3 :- Ability to add mining algorithm as a Component to the exiting tools.

- Software & Hardware Requirements :-

Sr_no	S/W & H/W Requriements	Specification
1)	Notpad or Text document	Notepad version VII. 2312. 18.0
2)	Weka tool	version - V3.8.6
3)	laptop	8-GB RAM, 64 bit.

- Theory :-

- Data Preprocessing in Data Mining :-

Data preprocessing is an important step in the data mining process.

- It refers to the cleaning, transforming, integrating of the data in order to make it ready for analysis.
- The goal of data preprocessing is to

improve the quality of the data and to make it more suitable for the specific data mining task.

- Some common steps in data preprocessing include :

1) Data cleaning :- This involves identifying & correcting errors or inconsistencies in the data, such as missing values, outliers & duplicates.
   - Various techniques can be used for data cleaning, such as imputation, removal & transformation.

2) Data Integration :- This involves combining data from multiple sources to create a unified dataset.
   - Data integration can be challenging as it requires handling data with different fromats structures & semantics.
   - Techniques such as record linkage & data fusion can be used for data integration.

3) Data Transformation :- This involves converting the data into a suitable format for analysis.
   - common techniques used in data transformation include normalization, standardization & discretization.

- Normalization is used to scale the data to a common range, while standization is used to transform the data to have zero mean & unit variance.
- Discretization is used to convert continuous data into discrete categories.

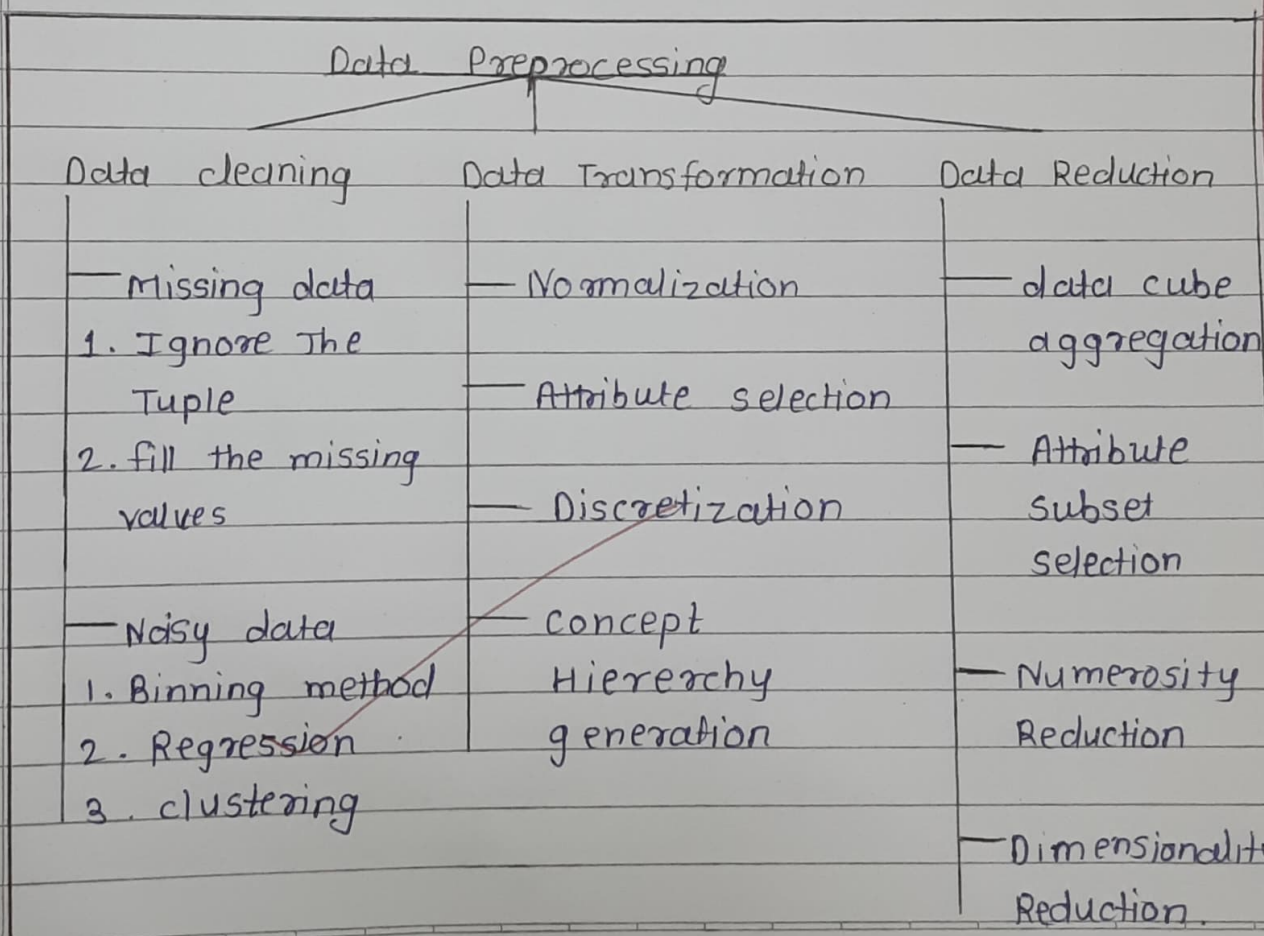
4) Data Reduction :- This involves reducing the size of the dateset while preserving the important information.
- Data reduction can be achieved through techniques such as feature selection & feature extraction.
- feature selection involes selecting a subset of relevant features from the dateset, while features from the dateset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

5) Data Discretization :- This involves dividing continuous data into discrete categories or intervals
- Discrelization is often used in data mining & machine learning algorithms that require categorical data.
- Discretization can be achieved through techniques such as equal width binning, equal frequency binning & clustering.

6) Data Normalization :- This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1.
- Normalization is often used to handle data with different units and scales.
- Common normalization techniques include min-max normalization, z-score normalization & decimal scaling.

• Preprocessing in Data Mining :-

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



Page No.
Date

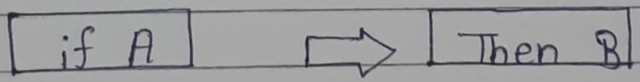
• Association Rule :-

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data items on another data item & maps accordingly so that it can be more profitable.

- It tries to find some interesting relations or associations among the variables of dataset.
- It is based on different rules to discover the interesting relations between variable in the database.
- The association rule learning is one of the very important concepts of ML, & it is employed in Market Basket analysis, web usage mining, continuous production, etc.

• How does Association Rule learning work?

Association rule learning works on the concept of if and Else statement, such as if A then B.



Here the if element is called antecedent, and then statement is called as Consequent.

- These types of relationships where we can find out some association or relation between two items is known as single cardinality.

Page No.
Date

- It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly.
- So, it measure the associations between thousands of data items, There are several metrics.
- These matrics are given below:
    1) Support
    2) Confidence
    3) lift

1) Support :-

Support is the frequency of A or how frequently an item appears in the dataset.
- It is defined as the fraction of the transaction T that contains the itemset X.
- if there are X datasets, Then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{freq}(X)}{T}$$

2) Confidence :-

Confidence indicates how often the rule has been found to be true.
- Or how often the items X & Y occur together in the dataset when the occurrence of X is already given.

- It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{confidence} = \frac{\text{freq}(X, Y)}{\text{freq}(X)}$$

3) Lift :-

It is the strength of any rule, which can be defined as below formula:

$$\text{lift} = \frac{\text{supp}(X, Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

- It is the ratio of the observed support measure & expected support if X and Y are independent of each other.
- It has three possible value.

1) If Lift = 1 :- The probability of occurrence of antecedent & consequent is independent of each other.

2) Lift > 1 :- It determines the degree to which the two itemsets are dependent to each others.

3) Lift < 1 :- It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Page No.
Date

- **FP-Growth Algorithm :-**

- An efficient and scalable method to complete set of frequent patterns.

- It allow frequent items set discovery without candidate item set generation.
  - Two step approach :
    1) Build a compact data structure called the fp-tree.
    2) Extracts frequent item set directly from the fp-Tree.

- **How to identify frequent patterns using fp-tree algorithm**

Step 1 :- Calculate Minimum support

Step 2 :- find frequency of occurrence

Step 3 :- Prioritize the items.

Step 4 :- Order the items according to priority

Step 5 :- ~~ord~~ Validation.

Page No.
Date

- Conclusion :-

Thus, I have studied about the Data preprocessing tasks.
Data preprocessing is an essential step in the data mining process & play a crucial role in ensuring that the data in a suitable format for data analysis.
And also learn about the association Rule & FP-growth algorithm.