Daffodil International University

Assignment on

Why it is important to learn data pre-processing?

&

What are the things to be considered for pre-processing of data?

Course Code: CSE-450

Course Title: Data Mining

Submitted To

Asst. Prof. Subhenur Latif

Daffodil International University

Submitted By

Md. Mubassir Hasan

141-15-3354

Sec: C

Data Pre-Processing:

In Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. For this type of data, we need a data mining technique that involves transforming raw data into an understandable format. In one word we can say Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

Important to learn data pre-processing:

We know in our Real world data are generally

- ❖ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- ❖ Noisy: containing errors or outliers
- ❖ Inconsistent: containing discrepancies in codes or names

To get fresh clean required information from this huge incomplete noisy data, we need to learn Data pre-processing.

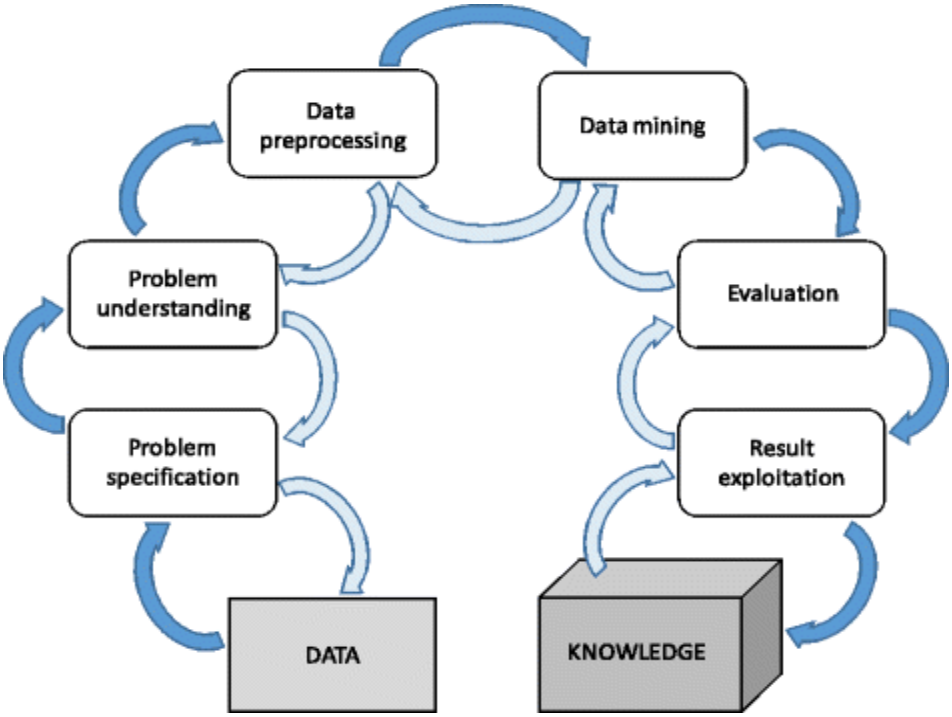
Things to be considered for pre-processing of data:

- ❖ Data Cleaning: Data cleaning is a procedure to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliners, and resolving inconsistencies

- ❖ Data Integration: Integrating multiple databases, data cubes or files, this is called data integration.

❖ Data Transformation: Data transformation operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the mining process.

❖ Data Reduction: Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results.



❖ Data Summarization: It is the process of representing the collected data in an accurate and compact way without losing any information, it also getting an information from collected data. Ex: Display the data as a graph and get the mean, median, mode etc.

Reference’s:

- Technopedia
- Cs