**DATA PREPROCESSING**

Data preprocessing is an important step in machine learning. The best phrase for this technique will be “Rubbish In, Rubbish Out”.

There are 7 steps for data preprocessing.

* Data Cleaning
* Data Integration
* Data Reduction
* Data Transformation
* Data Mining
* Pattern Evaluation
* Knowledge Representation

# **Data Integration:**

Collecting data or information from all sources integrated into one source. This can be done by following process,

* Data migration tool
* Data synchronization tool.
* ETL (Extract, Transform and Load) process

# **Data Cleaning:**

Data which are integrated might not be clean which may contain irrelevant data like out of range values, irrelevant data combination, missing values. This step will help the data to clear the discrepancy data.

# **Data Selection:**

This step is to will help to select the which data will be helpful for the further processing. Data selection can be done using neural network, decision trees, naive Bayes, clustering and regression.

# **Data Reduction:**

It is the process to shrink or make smaller in size with or without affecting the quality of data. The methods for the data reduction are

* Data Cube Aggregation
* Dimension Reduction (Stepwise forward selection, Stepwise backward selection)
* Data compression (Lossless Compression, Lossy Compression)
* Numerosity Compression
* Discretization (Top down, Bottom up)
* Concept hierarchy operation (Binning, Histogram analysis)

# **Data Transformation:**

Data is now ready for the transformation; we need transform the data into appropriate forms for data mining. The method used to achieve will be smoothing, aggregation and normalization.

# **Data Mining:**

We are going to apply the data mining techniques on data to discover the interesting patterns. Techniques like

* Transform task relevant data into *patterns*
* Decides purpose of model using *classification* or *characterization*

# **Pattern Evaluation:**

Process involves in to evaluate the patterns and select the which pattern will be appropriate to the user. The user will be more effective by visualizing the data.

# **Knowledge Representation:**

This is the last step for data preprocessing will be giving the data to the user by using visualization tools. Data might be in reports, tables ,etc.