Data Reduction and Data Transformation
Value added Course
Lecture 4

Er. Brijendra Singh

Assistant Professor, Department of Bioinformatics

UIET, CSJMU University, Kanpur

Data Preprocessing

- Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Why is Data preprocessing important?

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following-

Accuracy: To check whether the data entered is correct or not.

Completeness: To check whether the data is available or not recorded.

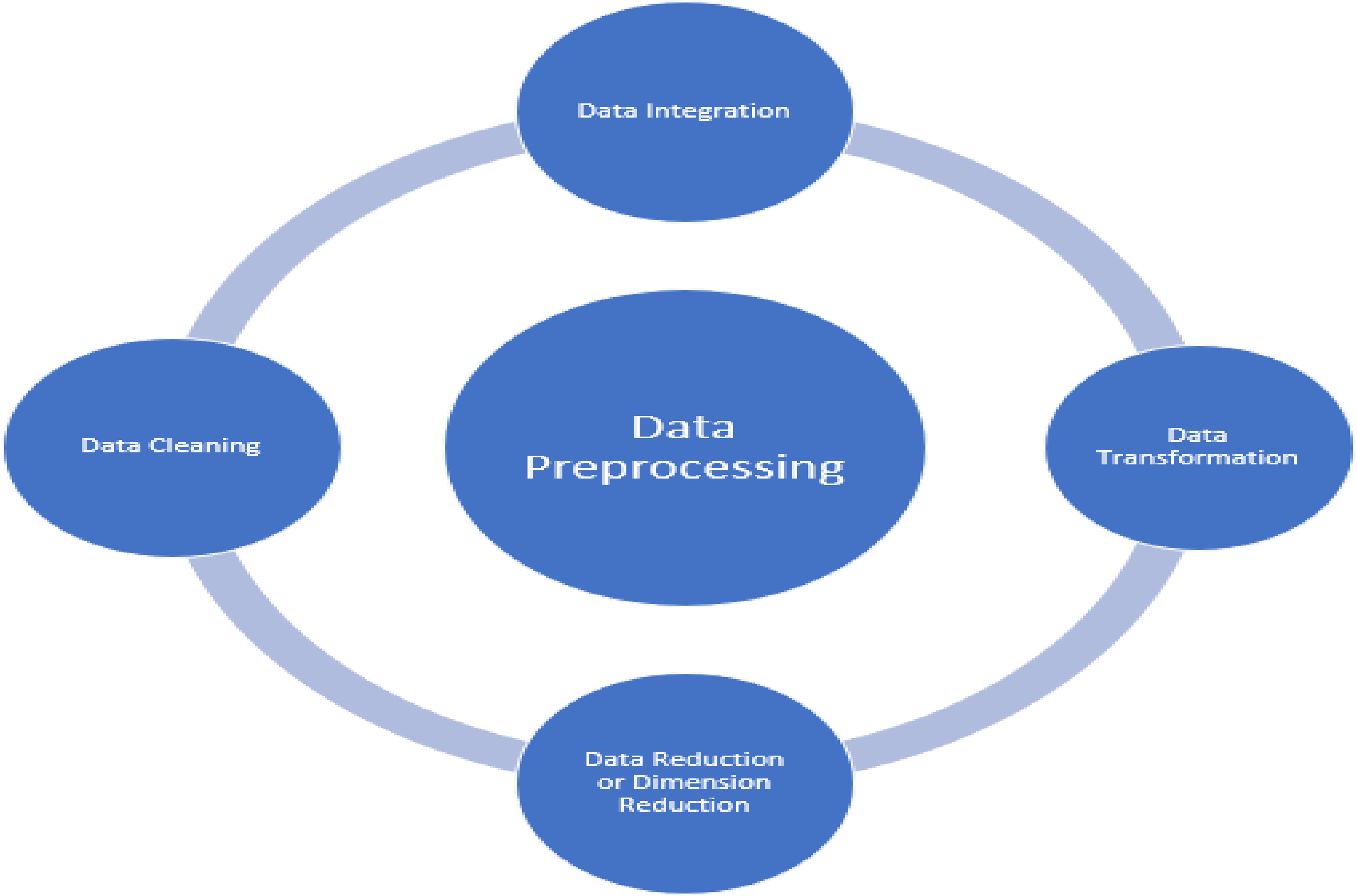
Consistency: To check whether the same data is kept in all the places that do or do not match.

Timeliness: The data should be updated correctly.

Believability: The data should be trustable.

Interpretability: The understandability of the data.

Steps involved in data pre-processing



Major Tasks in Data Preprocessing

- Data cleaning
- Data integration
- Data reduction
- Data transformation

Data cleaning:

- Data cleaning is the process to remove incorrect data, incomplete data and inaccurate data from the datasets, and it also replaces the missing values. There are some techniques in data cleaning

Handling missing values:

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually but it is not recommended when that dataset is big.
- The attribute’s mean value can be used to replace the missing value when the data is normally distributed wherein in the case of non-normal distribution median value of the attribute can be used.

Conti..

- While using regression or decision tree algorithms the missing value can be replaced by the most probable value.

Noisy Data:- Noisy generally means random error or containing unnecessary data points.

It is a random error or variance in a measured variable.

Methods to handle noisy data:-

Binning:-This method is to smooth or handle noisy data. First, the data is sorted then and then the sorted values are separated and stored in the form of bins.

- There are three methods for smoothing data in the bin. Smoothing by bin mean,

Smoothing by bin median, Smoothing by bin boundary.

- **Regression:-** This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.

Conti…

Clustering:- This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

Data integration:

- The process of combining multiple sources into a single dataset. The Data integration process is one of the main components in data management. There are some problems to be considered during data integration.

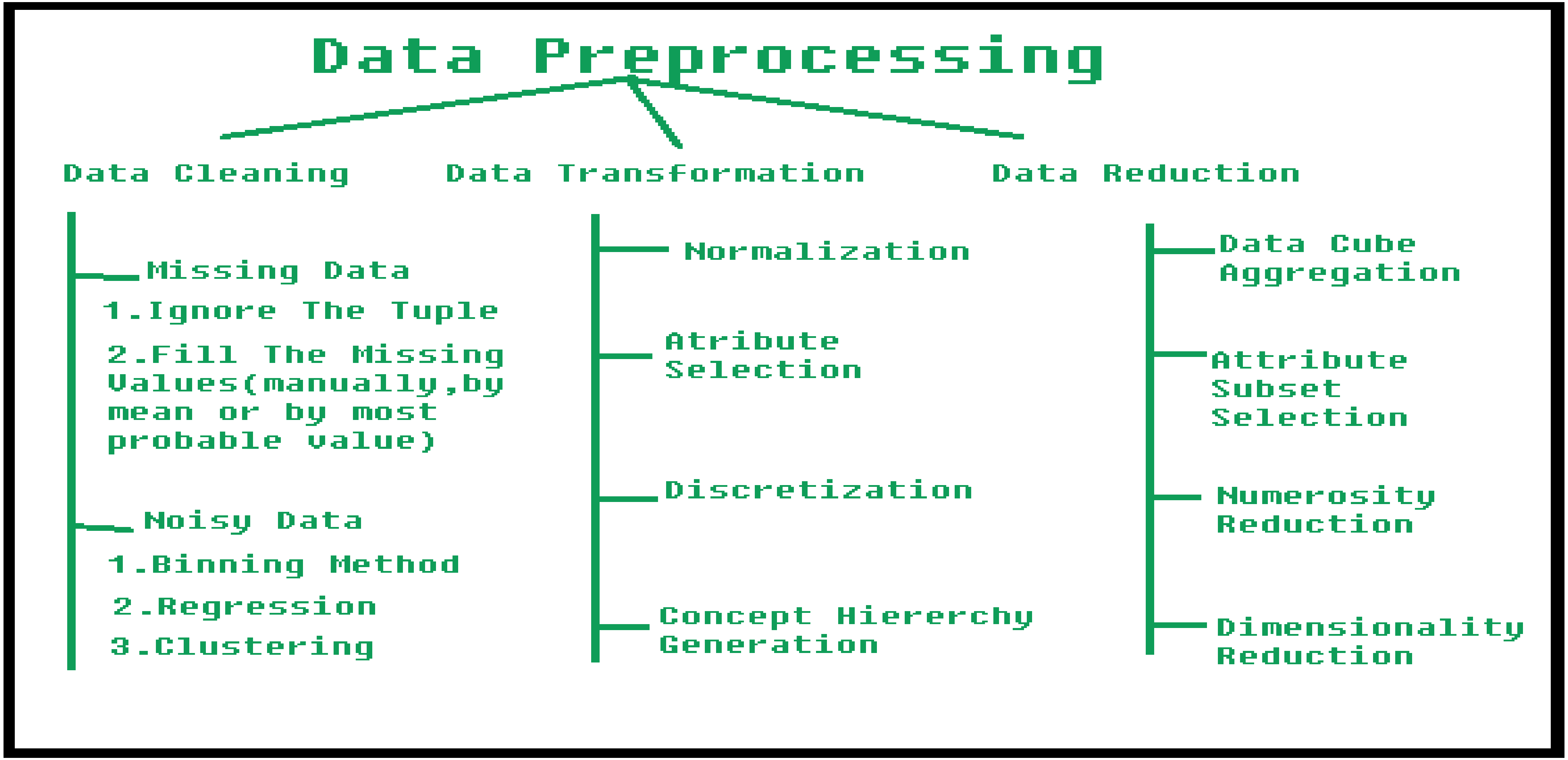
Schema integration: Integrates metadata(a set of data that describes other data) from different sources.

Entity identification problem: Identifying entities from multiple databases. For example, the system or the use should know student _id of one database and student name of another database belongs to the same entity.

- **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ. Like the attribute values from one database may differ from another database. For example, the date format may differ like “MM/DD/YYYY” or “DD/MM/YYYY”.

•

Overview of data preprocessing



Data Preprocessing

Data Cleaning

- Missing Data
  1. Ignore The Tuple
  2. Fill The Missing Values(manually,by mean or by most probable value)
- Noisy Data
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

- Normalization
- Atribute Selection
- Discretization
- Concept Hiererchy Generation

Data Reduction

- Data Cube Aggregation
- Attribute Subset Selection
- Numerosity Reduction
- Dimensionality Reduction

How to remove noisy data

- Binning methods:-

For example: consider the data- 10,2,19,18,20,18,25,28,22

i) First sort the data-2,10,18,18,19,20,22,25,28

ii) Divide the data into bins- (buckets of range of values)

iii) Three Bins

2,10,18 mean value is 10

18,19,20 mean value is 19

22 25 28 mean value is 25

(Bin Size=3)

Conti..

1. Smoothing by BIN means-

Calculate the mean value- mean value of three bins are 10, 19, 25.

Value of bin is replaced by mean value (average)

10 10 10

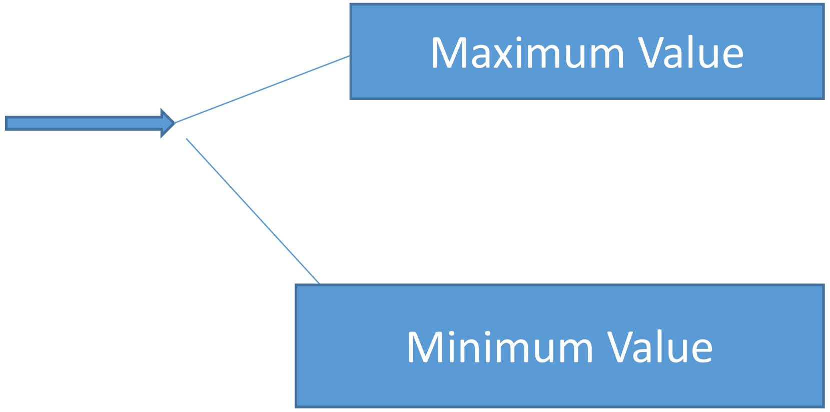
19 19 19

25 25 25

2. Bin Medians- if the value is odd (N+1/2), if the value is even (N/2) we take a value of second one.

Conti…

BIN Boundaries



Maximum Value

Minimum Value

All the values in the BINs are replaced by the closest boundary value.

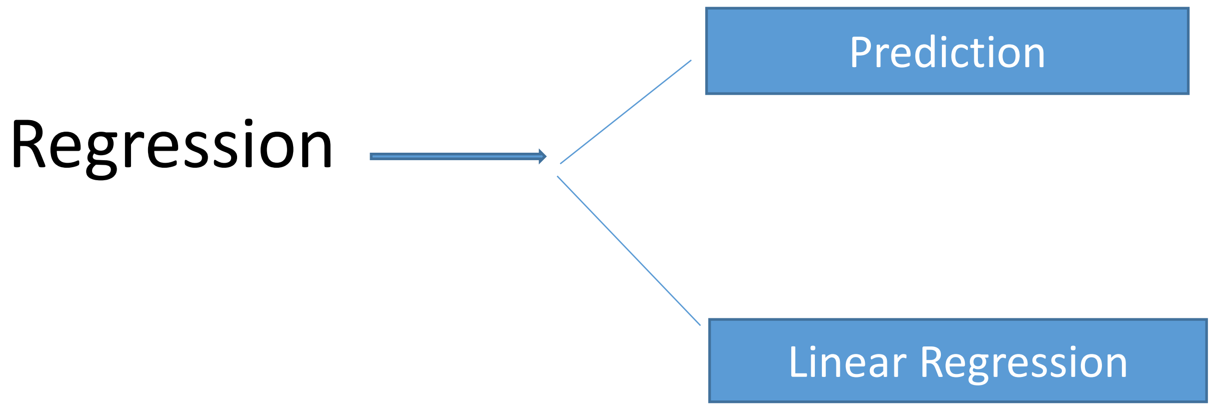
Values 2 10 18 2 2 18

18 19 25 18 18 25

22 25 28 22 22 28

Regression

- It is a data mining technique which is used to fit an equation to a data set. (data fitting function)



Regression

Prediction

Linear Regression

$$Y = b + mx$$

Y= Predicted value and mx is given value.

Clustering Method

Clustering:- Groups(clusters) are formed from the data having similar values.

Two types- unsupervised clustering – in this technique, where do not to supervise the model, instead you need to allow the model to work on its own to discover information. It mainly deals with the unlabelled data.

Supervised clustering- in this technique you train the machine using data which is well labeled.

Data transformation

- **Data transformation** is the process of changing the format, structure, or values of data.
- The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:
- **Smoothing:** It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

.

2. Aggregation

- Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.
- This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
- The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.
- For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

3. Discretization

- It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

For **example**, (1-10, 11-20) (age:- young, middle age, senior).

4. Attribute Construction:

Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

Conti..

5. **Generalization:** It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

- **6. Normalization:** Data normalization involves converting all data variable into a given range. Techniques that are used for normalization are:

Min-Max Normalization:

- This transforms the original data linearly.
- Suppose that: min_A is the minima and max_A is the maxima of an attribute, P

.

Conti…

- We Have the Formula:

$$v'= \underline{\text{v-minp}}\ /\ \ \text{maxp-minp (new\_maxp- new\_minp)+new\_mi}$$

  - Where v is the value you want to plot in the new range.
  - v’ is the new value you get after normalizing the old value.

Example: Suppose the minimum and maximum value for an attribute profit(P) are Rs. 10, 000 and Rs. 100, 000. We want to plot the profit in the range [0, 1]. Using min-max normalization the value of Rs. 20, 000 for attribute profit can be plotted to:

And hence, we get the value of v’ as 0.11

20,000-10,000/10,0000-10,000 (1-0)+0

v’= 0.11

Conti..

Z-Score Normalization: In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation

- A value, v, of attribute A is normalized to v' by computing

For **example**:
Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:

v'= v-a /Standard Deviation A

- **Example:** Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:

  85000-60000/10000 = 2.50

- And hence we get the value of v' to be 2.5

Conti..

- **Decimal Scaling:**
- It normalizes the values of an attribute by changing the position of their decimal points
- The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
- A value, v, of attribute A is normalized to v’ by computing

v’=v/10j

  - where j is the smallest integer such that Max(|v’|) < 1.

For **example**:

  - Suppose: Values of an attribute P varies from -99 to 99.
  - The maximum absolute value of P is 99.
  - For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.

.

Data Reduction

- The method of data reduction may achieve a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

Methods of data reduction:
These are explained as following below.

- **1. Data Cube Aggregation:** This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months.
- They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

2. Dimension reduction: Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.

Conti..

- **Step-wise Forward Selection –** The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.
- Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: { }

Step-1: {X1}

Step-2: {X1, X2}

Step-3: {X1, X2, X5}

Final reduced attribute set: {X1, X2, X5}

- **Step-wise Backward Selection –** This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set. Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: {X1, X2, X3, X4, X5, X6 }

Step-1: {X1, X2, X3, X4, X5}

Step-2: {X1, X2, X3, X5}

Step-3: {X1, X2, X5}

Final reduced attribute set: {X1, X2, X5}