**HEART DISEASE PREDICTION SYSTEM**

**Reg No : 2019506036 Date : 11-10-2021**

**Name : P Indu Team No : 05**

**Module: Data Cleaning, Preprocessing and Exploratory Data Analysis**

**Data Cleaning:**

Data cleansing/Data cleaning is a process of detecting and correcting the corrupt or the inaccurate records from the record set, table, or database and it refers to identifying the incomplete, incorrect, inaccurate or irrelevant parts of the data and then by replacing, modifying, or deleting the dirty or coarse data.

Data cleaning is a process of ensuring (checking) that your data is to be correct, consistent and usable.

You can clean data by identifying the errors or corruptions, correcting or deleting them, or manually processing the data as needed to prevent the same errors from the occurring.

## Data cleaning in six steps

### Monitor errors, Standardize your process, Validate data accuracy, Scrub for duplicate data, Analyze your data and Communicate with your team.

**Benefits of Data Cleansing**

* Improved the decision making. Quality data deteriorates at an alarming rate.
* Boost the results and revenue.
* Save the money and reduce waste.
* Save the time and increase productivity.
* It protects the reputation.
* It minimise the compliance risks.

**Data preprocessing**:

Data preprocessing is the data mining technique which is used to transform the raw data in a useful and to be in efficient format.

**Steps Involved in Data Preprocessing:**

1. Data Cleaning:   
The data can have many irrelevant and the missing parts. To handle with this part, data cleaning is to be done. It involves the handling of missing data, noisy data etc.

2. Data Transformation:   
This step has to be taken in order to transform the data in appropriate forms which is suitable for the mining process. It involves the Normalization, Attribute Selection, Discretization and Concept Hierarchy Generation.

3. Data Reduction:   
Since Data Mining is the technique which is used to handle a huge amount of data. While working with this huge volume of data, analysis became the harder in such cases. In order to get rid of this, we use the data reduction technique. It aims to increase the storage efficiency and to reduce the data storage and to analysis it’s costs.

The various steps to data reduction are: Data Cube Aggregation, Attribute Subset Selection, Numerosity Reduction, and Dimensionality Reduction.

Data preprocessing is an important task. It is nothing but a **data mining technique which transforms the raw data into a more understandable, useful and efficient format**. Data has a better idea. So, this idea will be clearer and understandable after performing the data preprocessing.

**Exploratory Data Analysis:**

Exploratory Data Analysis is an approach of analyzing the data sets to summarize their main characteristics, often using the statistical graphics and other data visualization methods. The primary goal of EDA is **to maximize the analyst's insight into a data set and into the underlying structure of a data set**, while providing all of the specific items that an analyst would want to extract from a data set, such as: a good-fitting, parsimonious model. It can also help to determine if the statistical techniques you are considering for the data analysis are appropriate. EDA techniques is widely used method in the data discovery process today.

**Steps in Exploratory Data Analysis**

1. Importing Data from Module.
2. Analyzing the Individual Feature Patterns using Visualization.
3. Descriptive Statistical Analysis.
4. Basics of Grouping.
5. Correlation and Causation.
6. ANOVA.

The purpose of exploratory data analysis is to check for the missing data and other mistakes. Missing values need to be handled with carefully because they will reduce the quality of any of our performance metrics which can also lead to wrong prediction or classification and can also cause a high bias for any given model which is being used.

Some advantages of Exploratory Data Analysis are improved the understanding of variables by extracting the averages, mean, minimum, and maximum values,etc.

**Implementation of the Module:**

The implementation of the module is given below:

<file:///C:/Users/HP/Downloads/Data%20Cleaning%20%20(1).html>