# **Background Study**

## According to the World Health Organization, 15 million people suffer stroke worldwide each year. Of these, 5 million die and another 5 million are permanently disabled. In developed countries, the incidence of stroke is declining, largely due to efforts to lower blood pressure and reduce smoking. However, the overall rate of stroke remains high due to the aging of the population. [2]

## 2.1 Data Mining

Data mining is a process of collecting or gathering hidden data from large data sets where data set means a collection of data. Data mining process is used for finding a pattern or similarity and relationship among large data set.

Mining process can be thought as a genuine appraisement of information technologies. There are many different issues in research which can be implemented by using data mining process and techniques. [3] In health care sector, a huge amount of data is adding day by day. So, we preferred to use data mining techniques in our thesis work.

**Knowledge Discovery in Databases (KDD)**

Data mining is a part of KDD. In KDD process at first data that we have collected through database from the user interface, will be stored in data warehouse. In warehouse data will be checked or tested that whether this data is good or not for the user or the operation.

After this KDD complete the rest four steps. They are transformation, mining, interpretation or evaluation and knowledge.

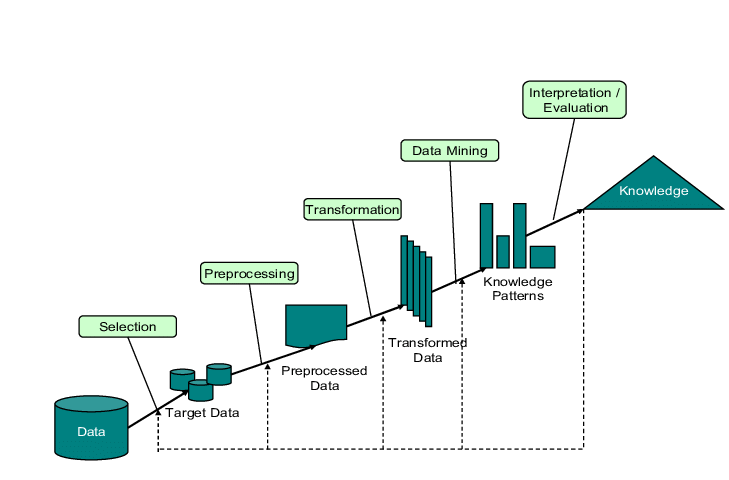


Figure 1: Steps of KDD

## 2.3.2 Naïve Bayes

Naïve Bayes is a type of classification which follows Bayes theorem and also known as probabilistic classifier method. It works with three step those are prior probability, conditional probability and posterior probability.

First step is **prior probability**. In this step the data set will convert into a frequency table. Second step is **conditional probability**. In this step, a table will generate for each train data depending on the decision class attribute. Third step is **posterior probability** and here if decision class attribute has two labels than prior will count for both class and check the value of which class is greater and finally predict the decision. There are three equations are given below which represents the classification formula of Naïve Bayes classifier. Here pos and neg represent respectively with the risk of having stroke and without the risk of having stroke.

P(pos) = Total positive risk factors.

P(neg)= Total negative risk factors.

test dataset, Ttest = a1, a2, a3………………………an.

P (Pos | Ttest) = P (a1|pos) \* P (a2|pos) \* P (a3|pos) \*………. P (an |pos) \*P (pos) (1)

P (Neg| Ttest) = P (a1|neg) \* P (a2|neg) \* P (a3|neg) \*………. P (an |neg) \*P (neg) (2)

(3)

Where n is the total number of attributes and i is an incremental number it reaches until the last number of attributes of test dataset and Rn represents the total number of risk factors.

## 2.3.3 Decision Tree

A decision tree is like a flowchart or a graph includes with a root node, leaf nodes. It can perform very well in KDD and data mining by enabling the model and knowledge extraction from the given dataset. It is able to perform with missing data or value. It can handle many types of input dataset like Numeric, Textual and Nominal.

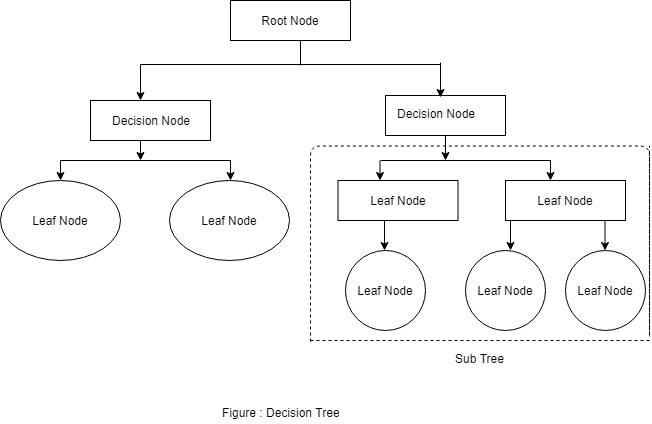


Figure 3: Decision Tree

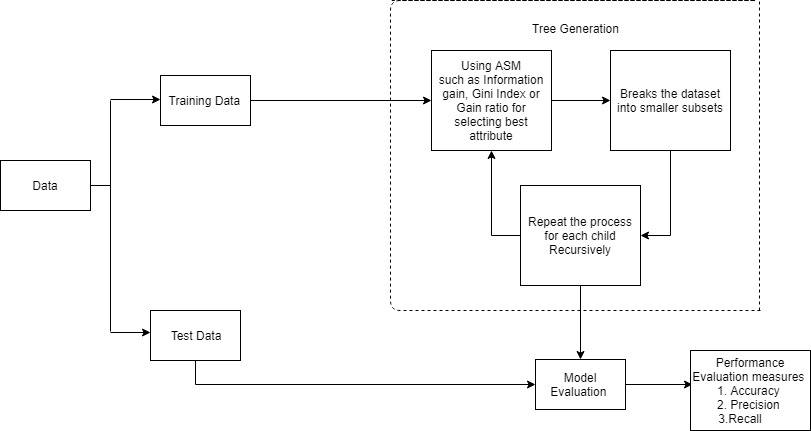
****Decision tree algorithms represents itself by following tree structure. The figure 4 shows how the decision tree algorithm works.

Figure 4: The working process of decision tree algorithm.

For selecting the best attribute, it uses ASM (Attribute Selection Measures) and for continuous-valued it uses some most popular attribute selection measures like Information gain, Gain Ration and Gini index.

**Information Gain** is one of the most popular attribute selection measure which measures the entropy also. Based on the given dataset, information gain computes the difference between the entropy before split and the average entropy after split. Information gain is used by ID3 decision tree. There are three equations from (4) to (6) are given below. Information gain follows those equation for measures.

Info(S) = (4)

Info A(S) (5)

Gain (A) = Info(S) - Info A(S) (6)

Where, Pi is the probability that an attribute tuple S belongs to class Ci.

Where, Info(S) is the required information for identifying the class label within S attribute tuple, Info A(S) is expected information, is the weight of the jth partition and i is an incremental number .Here, attribute A will be selected within the highest range of value of Gain (A) as the splitting attribute.