**iii 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. High blood pressure contributes to more than 12.7 million strokes worldwide. Europe averages approximately 650,000 stroke deaths each year. 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]

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. We are living in an age where data is constantly being collected. Our required and unnecessary data is being submitted too. So, we need to get help with data mining if we need to find our required data. In data mining process, Flat Files, Rational Databases, Datawarehouse, Transactional Databases, Multimedia Databases, Spatial Databases, Time Series Databases and World Wide Web (WWW) are different sources are used. Classifications, regression, clustering analysis, associations are different functionalities of data mining. Data mining techniques have great success in search engine, health care, education and business intelligence etc. It is also a very helpful techniques for researcher. So, we wanted to use it in healthcare sector to predict the risk of stroke.

Data miming techniques have a great success in research and development. Data 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, those data can be mined for different research work. So, we preferred to use data mining techniques in our thesis work.

1. **Knowledge Discovery in Databases (KDD)**

As we know in medical sectors there are many data sets are being collected and we do not need all of those data sets. So, we have to find out our necessary data sets among them. Knowledge Discovery Database can help us in that situation. Knowledge Discovery Database is a name of a process that can recognized our required data from the large data sets. It is a very helpful process for finding a pattern form large amount of data.

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. Then data will go through many different phases namely; data mining, pattern finding, related data or pattern finding, representation of data. KDD starts it’s work through data cleaning, data integration, data selection.

**Data Cleaning** is a task preformed in KDD for removing noise, handling missing data, finding error, inconsistent data etc. For example, data parsing in which data will be checked weather the data is acceptable or not.

**Data Integration** is a task used for integrating or collecting together all of those data which are coming from different sources for storing that’s data set in data warehouse.

**Data Selection -** KDD will retrieved all of the data which are related to our required data sets from the database for our analyzation.

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

**Transformation** is a step of KDD where all data which are cleaned, integrated and selected will be being consolidated or transformed from one format to another format. With the help of this step the result of data mining in KDD will be more accurate and data will be easier to understand. For transforming data into a form KDD follows some steps those are smoothing, attribute constructing, aggregation, normalization, discretization, hierarchy generation for nominal data etc. It is a very important step.

**Data Mining** is a most important step in the whole process. In there, data will be mined to find a pattern or similarity. After finding the pattern, it will be evaluated and interpreted in **Data interpretation or evaluation** steps for making the pattern more appropriate. Then KDD will show or represent the data sets to the user which is known as knowledge representation it can be as like image, graph, tree, chart, text etc. If the represented knowledge is matched with the user required data then KDD will store it as **knowledge**.

1. **Classification**

Generally, classification means classify something based on its some property or its identity. In machine learning algorithm, classification use to predict the label or the class by using a classification model. Supervised learning is the origin of the classification. Classification includes two parts. First one is learning part. A classification model is consecrated by using the training data set with the help of some classifier rules in this part. The second one is classification part in where the established model is used to predict the class (label) on the test data. Before using the constructed classification model to predict the class we need to find out the accuracy of this model. For finding accuracy at first, the mode

There are a lot of algorithms are available in classification. We can choose algorithms from there depending on the nature of our data

As classification process is used to predict the class label so the researcher used it in many times to predict the risk factors of many diseases in different research works. In machine learning, we have two types of learning methods those are supervised and unsupervised learning. Classification goes to supervised learning and clustering goes to unsupervised learning. The machine learning can give better performance with the help of both supervised and unsupervised learning. To find the patterns from the given data sets and to give the best decision by observing those data sets the process works. Now we are describing about some different algorithms used in both learning methods in prediction.

The two learning methods are-

1. Supervised learning
2. Unsupervised learning
3. Supervised learning- is a learning method where the training data set is available to built a classifier model to predict the output. It is used to predict the future output where the past data set are well defined. If anyone wants to determine the class the label from a scenario which is optimal for an unseen instance then the supervised learning methods also can do this correctly. In supervised learning

-known class available.

- the both input and output data are also known to us so we can check whether the given output is matching to our target output or not.

Supervised learning has great success in many different fields such as bioinformatics, cheminformatics, database marketing, handwriting recognition, spam detection, pattern recognition, optical character recognition etc.

In supervised learning many different algorithms are available. Those are Support Vector Machines, Linear regression, Logistic regression, Naïve Bayes, Decision trees, K-nearest neighbor, Neural Network etc.

Let us discuss about those algorithms.

**Support Vector** **Machine**

Support Vector Machine is a training algorithm works for pattern recognition problem which is a supervised algorithm also. It can work for both linear and non-linear data by finding a hyperplane for separating the data sets into two classes in the input space. Generally, SVM is a non-probabilistic binary classifier method use for finding the result in the simplest way which was based on a strong mathematical foundation. At first, it works for finding the hyperplane so that it can separate the given data set into the class labels those data belongs. For finding that hyperplane, the SVM creates some margins and checks which one is the best margin between all.

For choosing the best hyperplane, the machine checks which line is passing as far as possible from all input data. Because if the machine chose the line which one is very close to the data then the line will not be very perfect for generalizing correctly.

SVM works in two steps. First step is training and the second step is classification step. If the data is linear, in training step SVM selects the support vector for both positive and negative class and then execute the training algorithm. For constructing the hyperplane, it uses linear discriminant function, y = (w\*x) + b.

Where,

y = are labels which class the input data belongs

x = are given input data instance

w = is the weight vector

b = is the offset.

For non-linear data SVM uses kernel function and mapped the input data into high dimensional space so that the input non-linear data set becomes linear and then the SVM works. It uses some kernel function for non-linear data to transform into higher dimension space those are Polynomial kernel, Gaussian radial basis function kernel, Sigmoid kernel etc. It is very faster and powerful binary classifier method. It can generalize pattern recognize problem well even when the size of given data set is big. It has nice math property and also have the ability to handle large feature space and we can use different types of data as input like trees, strings, pixel maps etc. into SVM. Although SVM is a good supervised training approach, it can consider or give the output only for two classes. It also needs a good kernel function for non-linear data weather it cannot give the best result. Dr. S. Vijayarani and Mr. S. Dhayanand showed in their research paper which was published in 2015 that the accuracy of SVM is better than Naïve Bayes algorithms [6].

**Naïve Bayes**

Naïve Bayes is a type of classification which follows Bayes theorem and also known as probabilistic classifier method. In 18th century Thomas Bayes who was a nonconformist English clergyman worked in probability and decision theory [5]. Naïve Bayes algorithm use to predict the probability of the result or the output for an unseen or unlabeled test input. As it is a type of supervised learning method so, the training data set will available for this algorithm. It is very helpful for large data set and very easy to implement. 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. Suppose we have three attribute class and one decision class. The decision class has two labels those are positive and negative and we also have a test data set which is unseen in the train data set. By following Naïve Bayes classifier, it will work according the equation is given below [7]:

P (c|x) =

P (c|x) = P (x1|c) \* P (x2|c) \*……. P (xn|c) \*P(c)

Here,

* C = The class of our decision attribute.
* X= The test data.
* P (c|x) = Posterior probability of the class.
* P (c) = Prior Probability of class.
* P (x) = Predictor of the Prior probability.
* P (x | c) = It is known as conditional probability (likelihood).

Naïve Bayes classifier method has great success in text categorization, hand writing recognition, medical diagnosis, spam filtering, able to handle continuous data, good for both binary and multi-class classification tasks etc. Aiswarya Iyer, S. Jeyalatha et.al. made an experiment in their research paper named “Diagnosis of diabetes using classification mining techniques” and showed that the Naïve Bayes algorithm gives the better performance accuracy than J48 decision tree [8].

**Regression**

Regression is a type of supervised learning method use to predict a value. It will have training data set for implementing its algorithm to predict the output according to its test data set.

**Logistic Regression**

**Linear Regression**

**Decision Tree**

**K-nearest Neighbor**

**Neural Network**

1. Unsupervised learning- is a learning method of machine learning where the training data won’t available for the machine. The machine will not have any idea about the output. In unsupervised learning, the input data will be
2. represented as they are uncategorized, unlabeled and the machine works on the data without having any knowledge about desired the output. So, the machine or developer will try to find out the output until it cannot satisfy the machine or the developer.

When a process has more complexity, unsupervised learning can do better performance than supervised learning [4]. It is very helpful for real time prediction. K-means clustering, hierarchical clustering are very known unsupervised learning algorithms.

**K-means Clustering**

K means clustering is a very simple and the easiest clustering algorithm.

Clustering means collecting some similar objects from input space into some groups. K means clustering also collects input data into some groups. In the k means clustering, the number of groups depend on the value of k. Here k is a variable and it will be a positive integer number.

Suppose we have some data as input and now we need to separate them into some groups by using k means clustering. Then at first, we will know about the value of k.

If k= 2 then we will separate the input data into two groups depending on their similarities. If k=3 then we will make three groups and so on. We need to calculate the mean for clustering and it starts calculating from a centroid point of each cluster.

**Hierarchical Clustering**

Hierarchical clustering is a very popular example of unsupervised learning method. It is also a process of collecting all the inputs into some groups. But here it starts to collect input data into small groups and then marge all the groups together by following bottom up approach. The clustering process will not stop until all the cluster groups do not come in one big cluster. It also makes a cluster tree for each cluster or step.