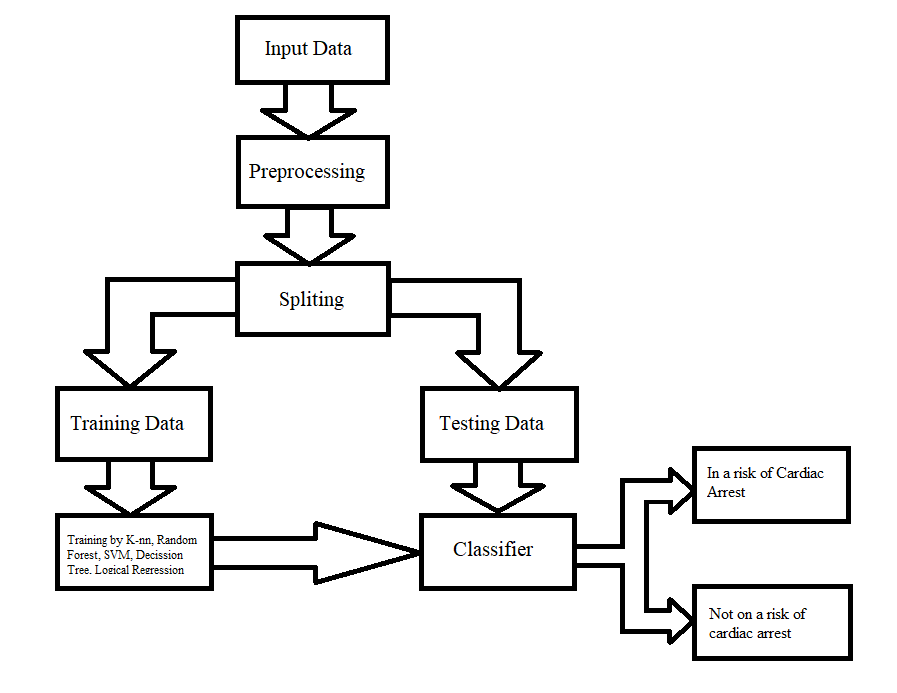
**CHAPTER 3**

**Block Diagram / Description / Flowchart / Algorithm**

**Block Diagram **

**Block diagram of ML model for Cardiac Arrest Prediction**

**Description**

**Input Data**

* Cleveland database from UCI’s repository was used for heart disease prediction system. Because Cleveland database is the most commonly used database by ML researchers. The dataset contains 303 instances and 76 attributes, but only 14 of them are referred by all published studies.
* Then dataset was split into three parts: one for training (%70), second one for testing (%15) and third one for validation (%15). There are 213 instances and 13 attributes in training data. Test data and validation data contain 45 instances and 13 attributes.
* A Cleveland dataset is a collection of information that is organized so that it can be easily accessed, managed and updated. Dataset process workloads to create and update themselves, querying the data they contain and running applications against it.
* In this project of Prediction of the Risk of Heart Attack after finding different parameters that could help analyze the person’s risk of having CVD. A huge database is required to store the information.
* Mentioned dataset is taken from different patients considering their age, gender, blood pressure, family history, cholesterol etc.

**Pre-Processing**

* Data preprocessing is an essential step use to clean the data and make it useful for any experiment associated with machine learning or data mining. In this study, multiple preprocessing steps applied on the selected dataset.
* The dataset contains 14 columns and 303 rows. We see that there are only 6 cells with null values with 4 belonging to attribute ca and 2 to thal.As the null values are very less, we can either drop them or impute them. I have imputed the mean in place of the null values however one can also delete these rows entirely.

**Splitting**

* Data is at the heart of every ML problem. Without proper data, ML models are just like bodies without soul. But in today’s world of ‘big data’ collecting data is not a major problem anymore. We are knowingly (or unknowingly) generating huge datasets every day. However, having surplus data at hand still does not solve the problem. For ML models to give reasonable results, we not only need to feed in large quantities of data but also have to ensure the quality of data.
* Data is at the heart of every ML problem. Without proper data, ML models are just like bodies without soul. But in today’s world of ‘big data’ collecting data is not a major problem anymore. We are knowingly (or unknowingly) generating huge datasets every day. However, having surplus data at hand still does not solve the problem. For ML models to give reasonable results, we not only need to feed in large quantities of data but also have to ensure the quality of data.
* The data should ideally be divided into 2 sets – namely, train and test set. The 80% data is taken for training while remaining 20% data is used for testing.

**Training Data**

* The train set would contain the data which will be fed into the model. In simple terms, our model would learn from this data. For instance, a Regression model would use the examples in this data to find gradients in order to reduce the cost function.
* Then these gradients will be used to reduce the cost and predict data effectively. The evaluation metric used is the confusion matrix. The confusion matrix displays the correctly predicted as well as incorrectly predicted values by a classifier.
* The sum of TP and TN, from the confusion matrix, is the number of correctly classified entries by the classifier. The algorithms are implemented with the default parameters only.
* In our case training of the machine is done by the help of K-nn, Decision tree, SCM, logical regression, Random forest as base classifier.

**Testing Data**

* The test set contains the data on which we test the trained and validated model. It tells us how efficient our overall model is and how likely is it going to predict something which does not make sense.
* There is a plethora of evaluation metrics (like precision, recall, accuracy, etc.) which can be used to measure the performance of our model. In simple terms it is used as to evaluate the fit machine learning model.

**Classifier**

* Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).
* A classifier utilizes some training data to understand how given input variables relate to the class. Classification belongs to the category of supervised learning where the targets also provided with the input data.
* In our case final classification is done by hyperparameter tuning of logistic regression and N-nn and prediction is done whether the subject is going to have a cardiac arrest or not (valued 0 or 1).