**Abstract**

Over the last decade heart disease is the main reason for death in the world. Almost one person dies of heart disease about every minute in India alone. In order to lower the number of deaths from heart diseases, there has to be a fast and efficient detection technique. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart-related diseases. Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical centre to predict various diseases. A huge amount of patient-related data is maintained on monthly basis. The stored data can be useful for the source of predicting the occurrence of future diseases. Some of the data mining and machine learning techniques are used to predict heart diseases, Decision Tree is one of the effective data mining methods till this date. The algorithms used in this project is namely are Decision Tree, Support vector machine (SVM), Logistic regression, Random Forests. Heart disease defines several healthcare conditions that are vast in nature which is related to the heart and has many basic causes that affect the entire body. The Dataset is collected form Kaggle Repository which contains 303 Instances with 14 features. This study's goal is to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence. The researchers accelerating their research works to develop software with the help of machine learning algorithms which can help doctors to decide both prediction and diagnosing of heart disease. The main objective of this research project is to predict the heart disease of a patient using machine learning algorithms.

**Existing System**

In this system, the input details are obtained from the patient. Then from the user inputs, using ML techniques heart disease is analysed. Now, the obtained results are compared with the results of existing models within the same domain and found to be improved. The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 87% for F-measure, competing with the other existing methods.

**Problem Identification**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data. There are 3 areas that challenge me, ethics, data and adoption. I think these are issues that the industry must tackle, we are going to be on the leading edge and have to help sort this out. Ethics is around the ownership and use of the data, we are at best the custodians of the data which belongs to the patients themselves, how do we safeguard it, how do we manage consent, and how do we ensure that we are doing things for good? Most data sets are not large, in Big Data terms in the EMR and the quality of the data is not always good so we have to spend a lot of time cleaning the data and working on techniques to use the data efficiently. Then there are a variety of tools and we should choose the right one for the right job, the old saying that “if all you have is a hammer then everything looks like a nail”, is so true in this field. This sometimes makes the results of studies hard to replicate, so we need to do “good science” here and make sure we are not finding patterns that don’t exist.

**Problem Objective**

The Main Objective of Developing this project is:

Predict whether a patient should be diagnosed with Heart Disease. This is a binary outcome.

* Positive (+) = 1, patient diagnosed with Heart Disease.
* Negative (-) = 0, patient not diagnosed with Heart Disease
* Experiment with various Classification Models & see which yields greatest accuracy.
* Examine trends & correlations within our data
* Determine which features are most important to Positive/Negative Heart Disease diagnosis

1. To Implement machine learning model to predict future possibility of heart disease by various machine learning algorithms.
2. To determine significant risk factors based on dataset which may lead to heart diseases.
3. To analyse feature selection methods and understand their working principle.

**Project Outcomes**

The main purpose of designing this system is to predict the risk of future heart disease. we have used Various Machine Learning algorithms to evaluate and train the model, such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine (SVM). These algorithms are discussed below in detail.

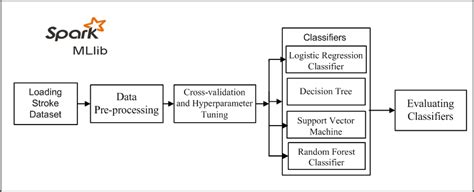
1. **Logistic Regression**: Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward elimination and recursive elimination techniques. This type of statistical analysis (also known as logit model) is often used for predictive analytics and modeling, and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C or D (multinomial regression). It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made. For example, you may want to know the likelihood of a visitor choosing an offer made on your website — or not (dependent variable). Your analysis can look at known characteristics of visitors, such as sites they came from, repeat visits to your site, behavior on your site (independent variables). Logistic regression models help you determine a probability of what type of visitors are likely to accept the offer — or not. As a result, you can make better decisions about promoting your offer or make decisions about the offer itself.
2. **Random Forest Classifier**: Random Forest is a supervised machine learning algorithm. This Technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on the belief that a greater number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality. The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “[the random subspace method](https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf)”(link resides outside IBM) (PDF, 121 KB), generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features. If we go back to the “should I surf?” example, the questions that I may ask to determine the prediction may not be as comprehensive as someone else’s set of questions. By accounting for all the potential variability in the data, we can reduce the risk of overfitting, bias, and overall variance, resulting in more precise predictions.
3. **Support Vector Machine**: Support vector machines exist in different forms; linear and non-linear A support vector is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line", A SVM can make some errors to avoid over-fitting, it tries to minimize the number of errors that will be made Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the Other learning algorithms used and the SVM classifier, the SVM has been worked very well with our dataset and produced good results. Support vector machine is able to generalize the characteristics that differentiate the training data that is provided to the algorithm. This is achieved by checking for a boundary that differentiates the two classes by the maximum margin. The boundary that separates the 2 classes is known as a hyperplane. Even if the name has a plane, if there are 2 features this hyperplane can be a line, in a 3-D it will be a plane, and so on. In the sample space, there is a possibility of zillions of hyperplanes but the objective in a support vector machine is to find that hyperplane through an iterative process that determines the hyperplane that has the maximum margin from all the corresponding training points and this will facilitate any new point coming in to fall into the accurate class on the basis of the features.
4. **Decision Tree:** A decision tree is a [decision support](https://en.wikipedia.org/wiki/Decision_support_system) tool that uses a [tree-like](https://en.wikipedia.org/wiki/Tree_(graph_theory)) [model](https://en.wikipedia.org/wiki/Causal_model) of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). It is one way to display an [algorithm](https://en.wikipedia.org/wiki/Algorithm) that only contains conditional control statements. Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](https://en.wikipedia.org/wiki/Goal), but are also a popular tool in [machine learning](https://en.wikipedia.org/wiki/Decision_tree_learning). A decision tree is a [flowchart](https://en.wikipedia.org/wiki/Flowchart)-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), a decision tree and the closely related [influence diagram](https://en.wikipedia.org/wiki/Influence_diagram) are used as a visual and analytical decision support tool, where the [expected values](https://en.wikipedia.org/wiki/Expected_value) (or [expected utility](https://en.wikipedia.org/wiki/Expected_utility)) of competing alternatives are calculated.

A decision tree consists of three types of nodes

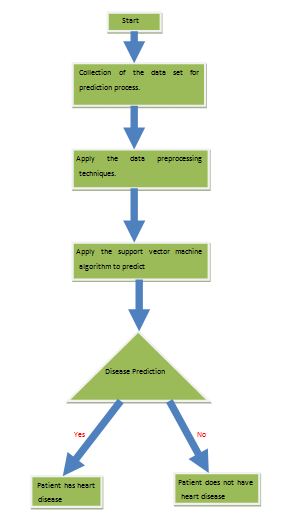
1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research) and [operations management](https://en.wikipedia.org/wiki/Operations_management). If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a [probability](https://en.wikipedia.org/wiki/Probability) model as a best choice model or online selection model [algorithm](https://en.wikipedia.org/wiki/Algorithm). Another use of decision trees is as a descriptive means for calculating [conditional probabilities](https://en.wikipedia.org/wiki/Conditional_probability). Decision trees, [influence diagrams](https://en.wikipedia.org/wiki/Influence_diagrams), [utility functions](https://en.wikipedia.org/wiki/Utility_function), and other [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis) tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or [management science](https://en.wikipedia.org/wiki/Management_science) methods.

**Architectural Design**



**Flow Chart**



**Advantages**

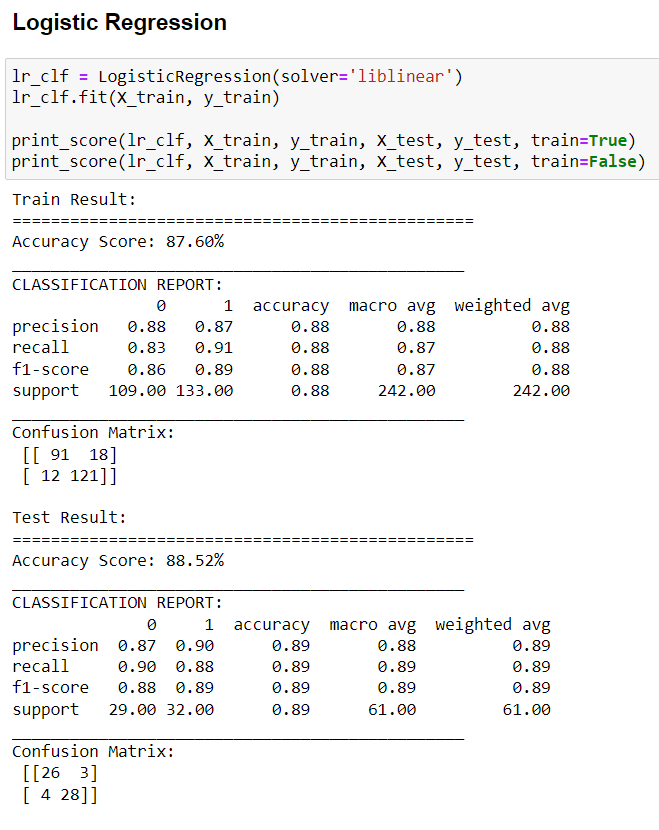
1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

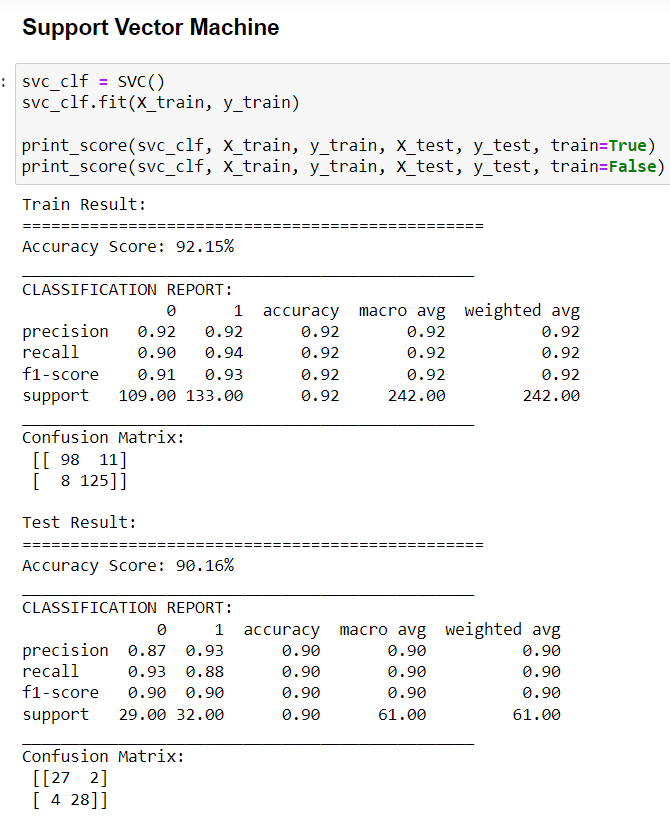
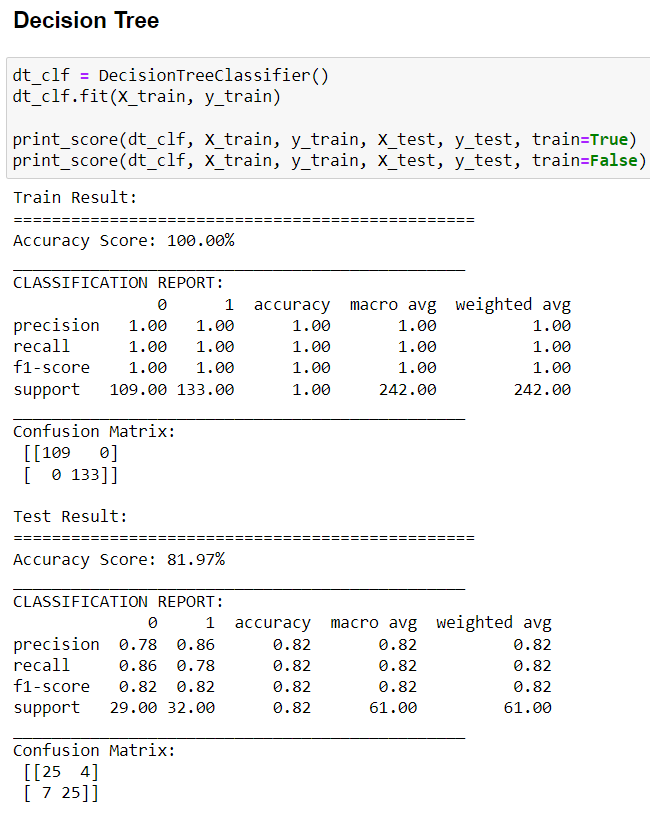
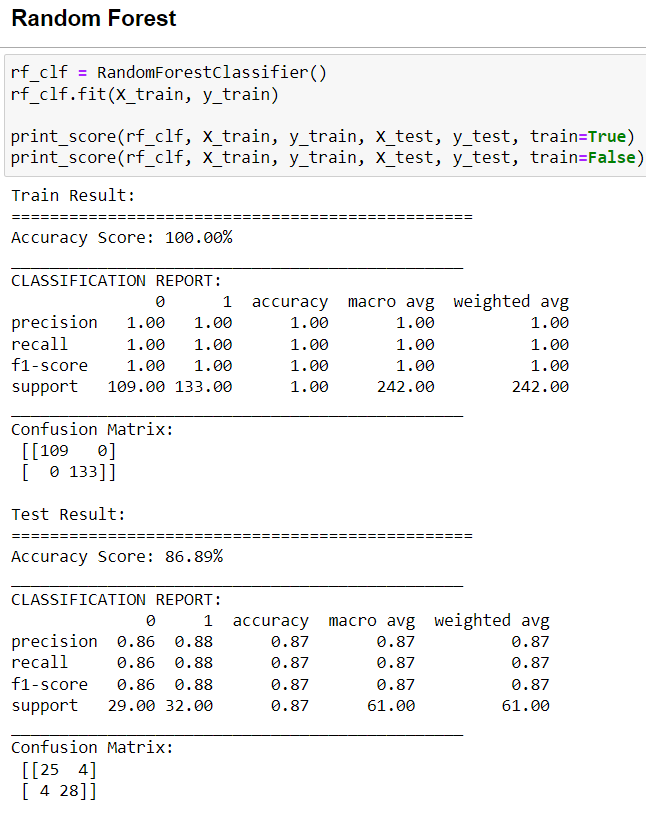
**Disadvantages**

1. Prediction of cardiovascular disease results is not accurate.
2. Data mining techniques does not help to provide effective decision making.
3. Cannot handle enormous datasets for patient records.

**Experimental Results of Algorithms**

1. **Logistic Regression**

****

1. **Support Vector Machine**
2. **Decision Tree**
3. **Random Forest Classifier**

**References**

1. <https://www.irjet.net/archives/V7/i4/IRJET-V7I4993.pdf>
2. <https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease>
3. <https://en.wikipedia.org/wiki/Cardiovascular_disease>
4. <https://www.kaggle.com/rishidamarla/heart-disease-prediction>