Heart Disease Prediction Using Machine Learning

Abstract: Heart diseases cause a high mortality rate in the world. Prediction and diagnosing of heart disease has become a challenging task faced by doctors and hospitals both in India and abroad. Heart Disease Prediction System is the system that helps to predict the heart disease mainly cardiovascular disease that includes Myocardial infractions. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop a software with the help machine learning algorithm which can help doctors to take decision regarding both prediction and diagnosing of heart disease. The main objective of this project is predicting the heart disease of a patient using machine learning algorithms. Comparative study of performance of machine learning algorithms is done.

Keywords-Heart Disease; Prediction System: Machine Learning; Data Mining; Classification.

Introduction:

The successful treatment of a disease is always attributed by early and accurate diagnosis. Now a days doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. We are living in an “information age” where we have large amounts of data being generated every day. This data can be used in accordance with various techniques of artificial intelligence and machine learning to effectively detect the presence of diseases in patients or the progression of certain diseases in patients. This process involves the effective recognition of information from huge amounts of data. This has been described as the process of “Knowledge Discovery from data”, which can defined as the process of converting raw data into organized form, which consists of valuable information which can be used for decision making in many applications. Data mining techniques can be used to extract this information from raw data and convert it into suitable formats to be used. This information can then be used with machine learning algorithms to make predictions and classification.

These data mining and machine learning techniques can be used in the field of medicine for diagnosis of diseases. Working on heart disease patients is an application of data mining and machine learning techniques in this field.

Fig 1: Block diagram of the prediction system

Types of Cardiovascular Diseases:

Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as a heart attack). There is another heart disease, called coronary heart disease(CHD), in which a waxy substance called plaque develops inside the coronary arteries. These are the arteries which supply oxygen-rich blood to heart muscle. When plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over many years. With the passage of time, this plaque can harden or rupture (break open). Hardened plaque eventually narrows the coronary arteries which in turn reduces the flow of oxygen-rich blood to the heart. If this plaque ruptures, a blood clot can form on its surface. A large blood clot can most of the time completely block blood flow through a coronary artery. Over time, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn’t restored quickly, the section of heart muscle begins to die. Without quick treatment, a heart attack can lead to serious health problems and even death. Heart attack is a common cause of death worldwide. Some of the common symptoms of heart attack are as follows.

Prevalence of Cardiovascular Diseases:

An estimated 17.5 million deaths occur due to cardiovascular diseases worldwide. More than 75% deaths due to cardiovascular diseases occur in the middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack . India too has a growing number of CVD patients added every year. Currently, the number of heart disease patients in India is more than 30 million. Over two lakh open heart surgeries are performed in India each year. A matter of growing concern is that the number of patients requiring coronary interventions has been rising at 20% to 30% for the past few years. The rest of the paper is organized as follows. (WRITE THE FLOW OF PAPER HERE)

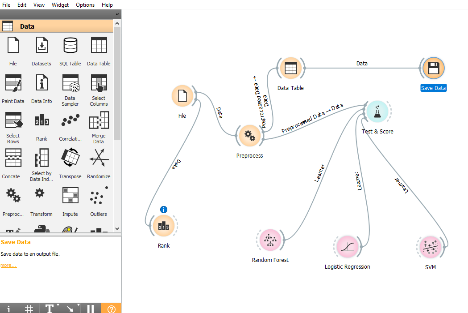
Data Mining:

Data Mining has to be done on the raw dataset to extract useful information to be used in machine learning algorithm.

Orange Data Mining Tool:

**Orange** is an [open-source](https://en.wikipedia.org/wiki/Open-source_software) [data visualization](https://en.wikipedia.org/wiki/Data_visualization), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [data mining](https://en.wikipedia.org/wiki/Data_mining) toolkit. It features a [visual programming](https://en.wikipedia.org/wiki/Visual_programming) front-end for explorative [data analysis](https://en.wikipedia.org/wiki/Data_analysis) and interactive data [visualization](https://en.wikipedia.org/wiki/Information_visualization), and can also be used as a Python library.

Orange is a component-based [visual programming](https://en.wikipedia.org/wiki/Visual_programming) software package for [data visualization](https://en.wikipedia.org/wiki/Data_visualization), [machine learning](https://en.wikipedia.org/wiki/Machine_learning), data mining, and [data analysis](https://en.wikipedia.org/wiki/Data_analysis). Orange components are called widgets and they range from simple data visualization, subset selection, and pre-processing, to empirical evaluation of learning [algorithms](https://en.wikipedia.org/wiki/Algorithms) and [predictive modelling](https://en.wikipedia.org/wiki/Predictive_modeling). Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed [widgets](https://en.wikipedia.org/wiki/Software_widget), while advanced users can use Orange as a Python library for data manipulation and widget alteration



The Dataset:

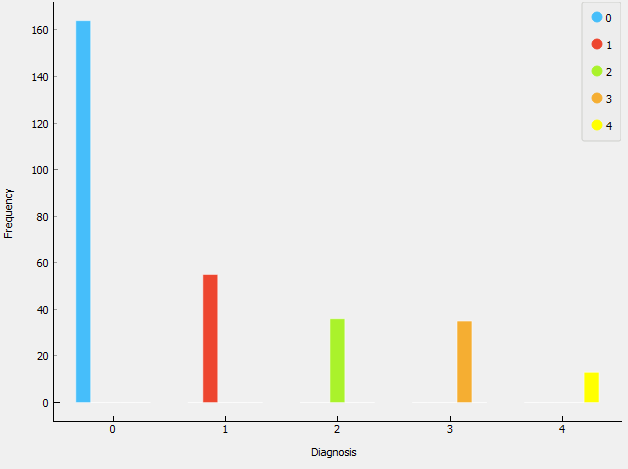
Dataset characteristics:

The dataset has been imported from the University of California, Irvine repository for machine learning. The dataset is multivariate with total 75 attributes. The attributes consist of categorical, numeric, binary and continuous attributes. Out of these 75 attributes, 14 major attributes are considered for this problem. The total number of instances are 303. There are 5 labels which are 0 for no disease and 1-4 for the progression of disease. There are multiple datasets from different sources but Cleveland dataset is used here as it has a smaller number of missing values hence is more accurate. It consists of 0.2% missing values and is used for classification problem.

Anomaly in Dataset:

During the analysis of the dataset it is found that the dataset is highly imbalanced. Apart from the label 0, the entries for the labels 1-4 are under sampled. This means that there is not enough data separately for these labels to effectively predict the presence of heart disease, which would result in low accuracy and precision. For example if a dataset consists of 100 instances out of which the tuples for label 1 are 98 and for the label 2 are only 2, we say that the data for label 2 is under sampled, so even if the classifier predicts all the inputs to be true it would be 98% accurate but actually it won’t be able to classify correctly.

To deal with this problem, synthetic minority oversampling can be done by adding the instances of the under sampled labels, or the majority class data can be under sampled by removing its entries. With the kind of data available, under sampling and over sampling will not affect the accuracy of the classification much. Another way to deal with this problem is to define two labels, i.e. for positive or negative diagnosis. This balances the dataset and increases the accuracy and precision of classification.



Conclusion:

Heart diseases when diagnosed early can be managed by various ways. By using the above approach we can predict the presence oh heart disease using the various symptoms of a patient. The two classifiers used give the most accurate predictions in this case, i.e. support vector machine and random forest classifier. Due to the unavailability of data in abundance we cannot predict the heart diseases of different kinds accurately, but the diagnosis of heat disease can be done with a fair accuracy of about 80 – 85 %. More accurate systems can be developed in future for diagnosis of diseases, when enough data is available.

SVM:

Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes.