

Chapter 1

INTRODUCTION

The diagnosis of the heart disease is difficult in real world and prediction of heart disease is a major challenge faced by hospitals and medical centers. Classification techniques in data mining play a significant role in prediction and data exploration. The healthcare industry collects huge amounts of healthcare data which is not feasible to handle manually. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Also, symptoms of heart disease may not be significant and thus are often neglected.

The main objective of this research project is to predict the patient getting the heart disease using machine learning algorithms. In order to handle Big Data Hadoop Map Reduce programming is used.

1.1 General Introduction

The comparative study of the machine learning algorithms is done and user can view the graphical representation of the results for easier understanding of the comparison. This application is made globally accessible by deploying it on cloud and can be accessed through any browser. It is user friendly, parallel processing is performed on the big data for faster computation. Project can be extended for predicting other diseases also.

1.2 Problem Statement

Heart Diseases remain the biggest cause of deaths for the last two decades. Recently computer technology and machine learning techniques are used to develop software to assist doctors in making appropriate decision of heart disease in an early stage. The diagnosis of heart disease depends on clinical and pathological data. Heart disease prediction system can assist medical professionals in predicting status of heart disease, based on the clinical data of patients.

Doctors may sometime fail to take an accurate decision in predicting heart disease risk level, therefore heart disease prediction systems are useful in such cases to get accurate results. There are many tools available for performing this task but all of them have some flaws. Most of the tools cannot handle big data and hence predicting heart disease would be a tedious task. In this project we are making an effort to predict the risk level of the huge datasets of patients.

1.3 Objective of the Project

The main objective of the project is to determine the heart disease risk level of a patient. The healthcare industry collects huge amounts of healthcare data which is difficult to handle with the currently existing systems. Our project is research based where the heart disease risk level of the patient is being predicted. Data mining techniques and machine learning algorithms are used for the study and cloud computing is used for deploying the same.

The tool used in development phase is Eclipse Kepler SR2 on Ubuntu operating system. The project is implemented using Java. These programs run on hadoop for parallel processing of the input data. The entire application is deployed on cloud using JELASTIC. There are many tools like WEKA for prediction but they have some limitations like, documentation for GUI is limited, scaling is a problem, can't handle Big Data , data preparation and visualization techniques offered might not be enough, it's a desktop application and not a Web application. This project is a web application since it is deployed on cloud and can be accessed through any browser. It determines the risk level of patients from Big Data and displays the graphs for easier understanding.

1.4 Current Scope

This project is useful for doctors and research centers. It is implemented using complex concepts which are gaining high scope these days. Some of these technologies include Hadoop Map Reduce programming, JFree chart API, Machine learning algorithms, Data Mining techniques and Cloud computing. Hospitals have huge datasets of patients and this project is helpful in determining the risk levels of patients.

Some of the merits of the project include:

- **Big Data:** The big data of the patient is handled using Map Reduce programming which splits the data and machine learning algorithm is applied for each line and the prediction of risk level.
- **Graphs:** In order to provide the easier understanding the results are displayed in the form of multiple graphs so that the user can get the comparative study of the algorithms.
- **Web Applications:** The tools that are available for the predictions are desktop applications our project is web applications since the project is deployed on cloud. Therefore it's easier for the user to access through all he browsers.
- **Comparative Study:** In this project we have implemented two machine learning algorithms and comparative study is done and result is shown in the form of graphs. Comparative study is done based on the accuracy of two algorithms.