Risk Predictor - Cardiovascular Diseases

Rohit Bansal(rbansal3) Alpesh Darji (adarji2) Kathan Al Jewary (ksa2) Paul Nel (paulnel2)

***Abstract*—This electronic document is a proposal for a project to develop a risk predictor for the incidence of heart disease using on some basic health metrics based on publicly available heart disease data set.**

Keywords—machine learning, map reduce, cardiovascular,

I. Motivation

According to WHO, Cardiovascular diseases (CVD) take the lives of 17.9 million people every year, 31% of global deaths. This is the leading cause of human deaths in the world. Often the old cliché of "prevention better than cure" is very much applicable when it comes to CVD. There are multiple factors which trigger these diseases and often there are multiple indicators before the one gets into grip of these diseases.

Our idea is to indicate the risk level for a user based on the user data. This risk level will indicate the urgency of taking precautionary measures. As part of this project, we will analyze the CVD dataset to find the leading cause of these diseases using ML algorithms. Once our ML model is trained, we will predict the risk level of an individual by analyzing the individuals corresponding dataset. This tool will be available on cloud so that users can use it to find the risk level corresponding to CVD.

II. Data Set

We will use the Heart Disease Data Set available from the UCI Machine Learning Repository. This data set contains four distinct databases from Budapest, Switzerland and Cleveland respectively. A total of 75 attributes are described in these data sets including metrics such as age, gender, smoking history amongst others. We will also try considering any other publicly available datasets and include them if time permits.

III. Proposed Outcome

We propose to build a simple predictor that will provide a user their risk of heart disease based on basic health metrics such as age, gender etc. This predictor will make use of machine learning techniques to provide a user with the risk score within a defined certainty.

IV. Methodology

We will apply machine learning algorithm(s) on the publicly available dataset to train our model. This way our model will be able to establish good understanding of the underlying data. To predict any individual's risk level for a heart disease, we will gather an individual's vitals and perform a predictive analysis using the model we developed. The work is divided in four major areas research, development, validation and reporting. Since all the team members are capable of performing work in all the four specified areas, we will attempt to allocate work in such a way that everyone gets a chance to work on all four domains.

V. Timelines

Milestone 1: Due Mar 9
Data Collection and Manipulation
By this milestone we should be done with gathering and processing data for selecting and training a machine learning model. We will ensure that we are able to host the data in a cloud environment and that we are able to work with it effectively and efficiently. Concepts learned regarding cloud environment, hosted instances and virtualization, will be extremely useful in achieving this milestone.

Milestone 2: Due Mar 30
ML Model Development
By this milestone we expect that our ML algorithm will be effective and mature. The data fed to the ML model should provide enough correlation for the model to predict an outcome. We hope to refine our model during this phase and attempt to implement concepts like server hosting, big data are presented during the CCA course.

Milestone 3: Due Apr 27
Real Time Prediction Using ML Model
By this milestone we should be able to analyze test data and predict result using our ML model in a reasonable time. To achieve this, the concepts concerning parallel processing, Hadoop Map-Reduce will be utilized to the fullest. By this time we will have project almost ready. Final piece (Milestone 4) would be to add more documentation, finalize project report, setup instructions and build presentation.

Milestone 4: Due Apr 30
Project Submission along with Final Report & Presentation
Our team will work together to distribute work regarding preparation of project report and presentation. In this phase code is documented and setup instructions are clear enough for anyone to follow.

VI. Resources

We propose to develop the tools using Python and for machine learning we propose to use publicly available libraries such as sklearn or similar.

- Language: pySpark
- Cloud Technologies Used: IaaS, PaaS, Map-reduce, Hadoop, Hortonworks
- Algorithms & Concept material: Well-Known Machine Learning (ML) Concepts
- Code Management System: Github
- Libraries: Scikit-Learn, scipy, pandas, numpy