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,

# Background & 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. There are no tools publicly available which could track these indicators and create a risk profile for an individual before the individual is in grip of one 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.

# Data Set

We will use the [Heart Disease Data Set](https://archive.ics.uci.edu/ml/datasets/Heart+Disease) 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.

# 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.

# 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.

# 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.

# 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

**Project Progress Report**

Following are the updates added in project document to provide project progress report which is due on end of week 11.

# Data Preperation

The Heart Disease Data Set from the UCI Machine Learning Repository was utilized for this project. This data set consists of four distinct data sets which will be labeled in this text as the (i) Cleveland, (ii) Hungary, (iii) Switzerland and (iv) VA data sets.

The data sets share a number of common attributes, but the Cleveland set is considered to be the most complete and is typically the only one that has been used for machine learning research.

We have therefore opted to create two different datasets, on for Cleveland only and the second one a combination of all four datasets but using only the most common attributes across the four sets.

Both data sets were initially cleaned to remove all samples that contained erroneous data. The sets were then inspected for outliers based on comparison between standardized residuals and leverage, as well as Cook’s distance.

## Cleveland Data Set

For the Cleveland data set, the following attributes were used (refer to UCI reference for a detailed explanation of these attributes.):

1. #3 (age)

2. #4 (sex)

3. #9 (cp)

4. #10 (trestbps)

5. #12 (chol)

6. #16 (fbs)

7. #19 (restecg)

8. #32 (thalach)

9. #38 (exang)

10. #40 (oldpeak)

11. #41 (slope)

12. #44 (ca)

13. #51 (thal)

14. #58 (num) (the predicted attribute)

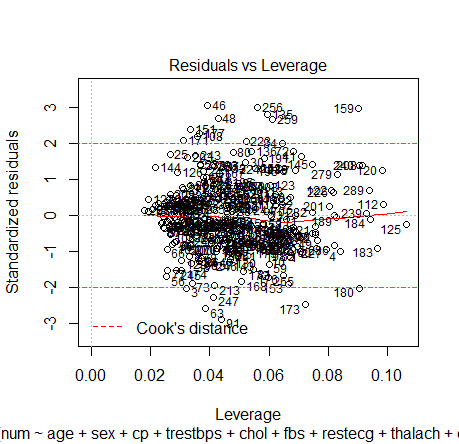
Preparation of the data consisted of two parts. Firstly, all rows containing erroneous values were removed. This reduced the data sample count from 303 to 296 which was then used to construct a linear regression model against all attributes using R.

Figure 1 shows the resulting plot with labelled data points.



Figure 1Residuals vs Leverage for complete Cleveland data set

Data point 209 was identified as a possible outlier due to higher than expected deviation from the mean and points 91 and 152 due to relatively high leverage and Cooks’ distance. Figure 2 shows the resulting plot after these three data points had been removed.

Figure 2 Residuals vs Leverage for reduced Cleveland data set

## Combined Data Set

Combining data sets proved to be somewhat of a challenge due to the high number of missing attributes in data sets (ii) – (iv) when compared with (i). Evaluating all attributes across all four data sets, the final subset decided upon included:

1. #3 (age)

2. #4 (sex)

3. #9 (cp)

7. #19 (restecg)

8. #32 (thalach)

9. #38 (exang)

14. #58 (num) (the predicted attribute)

It is interesting to note that these attributes will be the ones measured and recorded immediately on admission of a patient to the hospital as opposed to some of the other attributes that would require either fasted or non-fasted blood tests. It therefore suggests that this is a logical grouping of attributes.

In addition to consolidating the attributes across the four different data sets, we also decided to add an additional attribute, *set\_id*, that identified the dataset per sample. Since each data set represents a different geographical area, we will test to see if this has a significant outcome in predictions.

Combining these data sets resulted in a total sample count of 920. In total, 57 samples were removed that contained at least one erroneous attribute value resulting in an interim data set of 863 samples.

Similar to the Cleveland data set, we constructed a linear regression model in R to determine any obvious outliers resulting in the plot shown as Figure 3.

Figure 3 suggests that point 212 is a possible outlier in this group of samples and was removed. The result is presented in Figure 4.

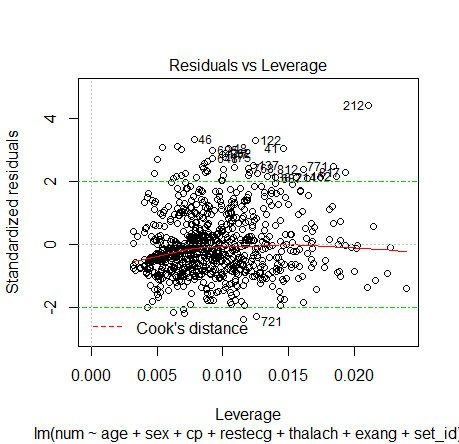


Figure 3 Residuals vs Leverage for complete Combined data set

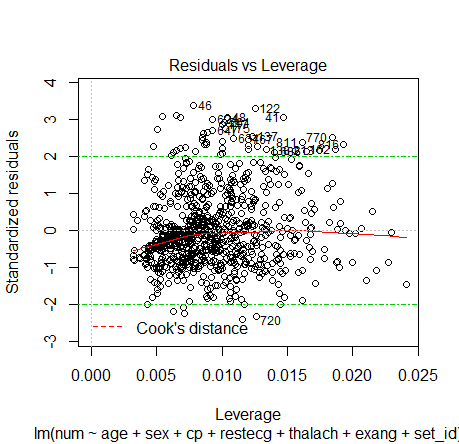


Figure 4 Residuals vs Leverage for reduced Combined data set

**Challenges**

Following were the challenges team faced in early phase of project bootstrap and team has collaborated and worked together to finalize data suitable for project

1. Finding data which can have information releated to Heart Disease and having multiple attributes for analysis. We searched on various open source platforms like Google Big Data& Kaggle to look into various categories to understand category of data which fits our criteria
2. Finding reliable data which can be trusted and backed by authority or recognized organization.
3. Research and understanding of various Machine Learning Algorithm suitable for the project. To solve this challenge we went through ML course and found that linear regression technique could be first step to get started.

**Changes**

There are no changes with respect to original project proposal.

**Prototype**

After having multiple meeting and discussion team has decided to have following project prototype in order to run analysis based on user input and display risk factor

*Step 1: Preprocess and data cleanup*

>> python preProcess.py

In this step application would cleanup data so that there is no inconsistency and no outlier and data format is adjusted and saved in separate file so that we can apply ML methodology to categorize result based on user input

*Step 2: Reading user input and providing prediction result*

Console:

>> python riskAnalysis.py input.json

Computing and matching your data …

Output of analysis:

riskResult: {

outcome: “MEDIUM”

}

The input.json file will have user input in following format

userData: {

age: 30,

sex: “M”,

cholestrol: 135.65,

sugar: 153,

ecg: 23.45

….

}

**Releated Projects**

We found following article in which author talk about disease prediction but there is no implementation and team is trying to put together pieces learned in courseowork into working application to take idea to next step.

Link:https://medium.com/@jonbaldie/predicting-heart-disease-diagnoses-with-machine-learning-2e1a8f5213f8

**Team Collaboration**

1. Working across different timezone   
    *Team has decided to have weekly meetings. Currently we have setup recurring meeting for every Saturday 10am.*

*We use Zoom as a communication tool for meetings*

1. Tasking and Resposibility

*Team has adopted Agile Methodology where we do task distribution and assign action item to each team member*

1. Project updates

*Team has integrated Github bot into Slack channel so that every push to project by any team member would notify us in Slack channel.*

**Risk & Timeline**

Currently there is no risk in project implementation   
 and team feel confident to deliver project as per   
 original timeline given in project proposal. Team has   
 finished first milestone and currently actively working  
 on Milestone on creating ML model.

*­­*