School of Computer Science and Applied Mathematics
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

COMS4030A Project– Report
Investigative Analysis of Heart Disease Diagnosis

Shameel Nkosi, 1814731, Coms Hons
Molefe Molefe, 1858893, Coms Hons
Siraj Motaung, 1390537, BDA Hons

June 23, 2021

Contents

Abstract

Everything that gets damaged can reach a point of "beyond repairs," the heart is one of the things you do not wish for it to get to such a stage. Knowing beforehand that you are likely to get heart diseases in the future can help reduce the impact through taking early preventative measures. One way of finding out the likelihood of a person having a heart disease is to have a model that can accurately and intelligently tell you whether you are likely to have heart disease or not. In this report, we discuss the models we built that can aid in solving the problem of telling whether a person has or is likely to have heart diseases. We also discuss their implementations, performances, optimizations, and many more.

1 Introduction

Human beings make more sound decisions if they have experience in whatever they are trying to make decisions upon. The best way for one to gain experience is to learn. We aim to teach a computer/model to make decisions based on what it learns from the data we feed it. There are many different techniques we can use to approach the solution to our problem. Each of these techniques has its pros and cons. Below we discuss the technique we employed and reasons as to why we've chosen this particular technique.

Remember that the aim is to whether a person is most likely to have heart diseases or not. We are therefore dealing with a problem whose solution gives a "yes" or "no". In machine learning, this is popularly known as binary classification.

1.1 Machine Learning technique and datasets

1.1.1 Technique

Probability theory is a study that involves the likelihood of events occurring. Logistic regression is a machine learning technique that gives a probability of an event taking place. Probabilities allow us to make decisions while acknowledging that we could be making a bad decision though the decision taken at that time seems to be the best. We have chosen logistic regression because it appropriately spits out values of 1's for patients with heart diseases and 0's otherwise. Logistic regression is less sensitive to outliers as compared to other techniques we could have employed. It is relatively straightforward to train and is capable of giving good accuracy scores.

1.1.2 Datasets

A machine learning algorithm learns to make decisions based on the data you feed it. The more data the algorithm sees, the more accurate its decisions become. The dataset we used in our project is a combination of four datasets collected from the UCI website. The four data sets combined into one have been collected from Cleveland, Hungary, Switzerland, and the VA Long Beach. The dataset has 13 features, namely:

- Age in years
- sex/gender
- type of chest pains
- Serum cholestoral in mg/dl
- Resting blood pressure
- Maximum heart rate achieved

- Exercise induced Angina
- whether fasting blood sugar is greater than 120 mg
- Electrocardiographical Results
- ST Depression Induced by Exercise Relative to rest
- Number of Major Vessels (0-3) colored by flourosopy
- The Slope of the Peak Exercise of ST Segment
- **Target:** whether a patient has a heart disease or not.

In the subsequent chapter, we discuss data preparation, data handling, preprocessing, data standardization, and data visualizations.

2 Implementation and Results

2.1 data handling

Data handling is arguable the most important part of building a machine learning algorithm that produces the best accuracy possible. For us to learn machine learning, it had to be well presented and well documented. Similarly, for a machine to learn, data needs to be well presented and it needs to be complete. There are four important parts of data handling worth mentioning, these are:

- **Dealing with duplicate values:** *Duplicate values* may cause our model to be biased against data that doesn't have duplicate instances. To best train our model, we handle this issue by dropping all *duplicate values*, leaving only one instance of it.
- **Missing Values:** Missing values is the most common defect found in a lot of datasets. Missing values may be a result of a recording error, or maybe the respondent did not know the answer to that specific question, or any other reason not mentioned here. In our project, we deal with missing values in three different ways, namely:
  - **Drop missing values:** Delete the rows that have missing values.
  - **Replace with Mode/Mean:** In the case of replacement with mode, replace the missing values with data points that appear the most in that particular column. In the case of replacement by mean, replace the missing values with the average of all values that are not missing in that particular column. In our implementation, we only used replacement by mode.
  - **Learn missing values:** Here we resort to other machine learning algorithms to predict what the missing values would be. We make use of unsupervised learning methods to learn the missing values.

  In our project, we generated three different datasets from the above forms dealing with missing values. We then run our models on all these datasets to see which method gives the best accuracy.
- **Data standardization:** Data standardization is the rescaling of the values in the dataset so that there isn't a biased contribution in the decision-making process of our model. Take, for example, one feature that has values from 100 to 10000 and another feature that has values from 0 to 1, if the data isn't standardized, the latter feature will be insignificant in the decision-making process.
- **The bias term:** Upon dealing with duplicate values, missing values, and feature scaling, we add a column of 1's as the left-most column of our dataset. This makes it easy for us to implement the model. Adding a bias gives us the desired shape to compute the dot product between a record and the θ theta parameters.

At this point, our datasets have taken a clean and workable form. We can now implement our algorithms on these datasets.

2.2 Implementations of Logistic Regression

In this subsection, we discuss two baseline implementations upon which we make improvements. The training stage of logistic regression involves an iterative process of learning the parameters that produce the best prediction. In the baseline implementation, we train our model using stochastic gradient descent.

2.3 Analysis of baseline results

2.4 etc

3 Optimization Techniques

3.1 Employed techniques

3.2 Outcomes of optizimations

4 Conclusion