WITS
UNIVERSITY

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

2.2 Implementations of Logistic Regression

2.3 Analysis of baseline results

2.4 etc

3 Optimization Techniques

3.1 Employed techniques

3.2 Outcomes of optizimations

4 Conclusion