Springboard: Data Science Career Track
Second Capstone Project: Project Proposal
By Lucien Meteumba
June, 2020

PROBLEM IDENTIFICATION

The goal of this project is to diagnose if a patient in a given hospital has cardiovascular disease or not. The first thought that can come in people's mind is "Doctors with medical equipment are the only people that can diagnose such patients". But with the rise of Data Science and Machine Learning, the data scientist can use data that have been already collected to predict the state of the patient. The model to be built out of this project will help Doctors diagnose patients faster and will help the hospital save money.

CONTEXT

With the help of historical data and machine Learning we are now able to help in the diagnosis of diseases. Data science can help diagnose if a patient has a particular disease by predicting the probability that patients will suffer from certain diseases. By creating a model able to predict if a patient has cardiovascular disease or not, we will be able to save time, distance and money. Patients will be able to be diagnosed from the comfort of their home provided that they have a computer. The hospital and the patient will not have to spend a lot of money because of the medical equipment and the Doctor's time.

CRITERIA FOR SUCCESS

Our criteria for success will be to build, and evaluate the performance of, a model that will be able to predict if a patient has cardiovascular disease or not.

SCOPE OF SOLUTION SPACE

We will analyze the data taken from "CARDIOVASCULAR DISEASE dataset"[1] from Kaggle's Website.

[1] -https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

CONSTRAINTS

The main constraints in this project might be missing values and the fact that our data might not be large enough.

STAKEHOLDERS

The intended stakeholders are Health specialists and Medical Doctors.

DATA SOURCES

The dataset we are using to support this project is “CARDIOVASCULAR DISEASE dataset” from Kaggle’s website. It is an excel file that has 70,000 observations and 13 features.

PROBLEM SOLVING

After cleaning the data, we will use Machine Learning models like KNN (k-nearest neighbors), Naive Bayes, Support Vector Machines, and other algorithms to find the model that can give us the best performance. We will be using the Python programming language for this project.
The deliverables will be the code I will develop, which will be available from my GitHub repository[2]. We will also deliver a project report and a presentation slide deck.

[2] -https://github.com/KingdomDataScience/Springboard