ICCS261 Term Project Report

Section 1: Topic and problem statement

What is your project title?

Heart Attack Risk Prediction

What is the major problem/research question that you seek to answer?

Given a person's lifestyle choices and current data about their health, can we reliably predict their risk of a heart attack.

What is the rationale for the problems?

According to the British Heart Foundation Global CVD Factsheet (June 2023), it's estimated that 1 in 13 people globally are living with a heart or circulatory disease. This number will continue to rise due to the current modern lifestyles and an aging and growing population. Most of the data that this Dataset contains is information which can easily be obtained from asking patients simple questions. And if these simple questions can be used to assess the risk of a heart attack for the patient, it could help in preventing/helping people who are at risk by suggesting some lifestyle changes that could be made to steer them away from it.This is why developing a machine learning model for this dataset is quite an interesting but also useful project.
Resource: https://rb.gy/dx0tzx

Section 2: Introduction and relevant literature

This is a paper on a similar topic that could be used as a reference.
https://f1000research.com/articles/11-1126
Even though it is not the same dataset, the paper still goes over the basic pipelines of what it takes to make a machine learning model, covering topics like Data Collection,EDA, and machine learning models that were used in their research.

Another paper discussing the same topic is https://www.mdpi.com/1999-4893/16/2/88

This one also uses a dataset from kaggle, which from a skim, doesn't seem to be the same dataset. However, it still contains useful steps such as outlining the method for outlier removal, Feature Selection and Reduction as well as Clustering. I'm not sure how many of these concepts will actually be used in my project. Nevertheless, it is good to have real world use cases as references.

Section 3: Describe your research design

Where is the source of data collection

https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data

What is the data about?

This dataset contains a wide range of information related to heart health and lifestyle choices. It includes specific patient details like age, gender, cholesterol levels, blood pressure, heart rate, and factors such as diabetes, family history, smoking habits, obesity, and alcohol consumption. Additionally, it covers lifestyle factors like exercise hours, dietary habits, stress levels, and sedentary hours. Medical aspects like previous heart problems, medication usage, and triglyceride levels are also taken into account. Socioeconomic factors such as income and geographic attributes like country, continent, and hemisphere are part of the dataset. With a total of 8763 patient records from all over the world, the dataset is valuable for binary classification, indicating the presence or absence of a heart attack risk. It serves as a comprehensive resource for predictive analysis and research in the field of cardiovascular health.

What was your approach to this data ?

The project will go through all the standard processes of a data science project and has been divided into these sections;

- Introduction to the Data
- Initial Cleaning
- EDA
- Data Preparation
- Model Selection
- Feature Selection
- Imbalance Removal
- Hyper-parameter Tuning

Section 4: Conclusion

Given the data that I had, which mostly consisted of superficial attributes of a person, I couldn't predict the risk of a heart attack for a given person. Maybe if the features were more scientific, like the division of Cholesterol into HDL and LDL would have helped. Something that could have improved the model more would be correctly balancing the labeled data, by not just removing a bunch of data. Furthermore, none of these methods used neural networks/deep learning models. This could be an interesting avenue to go down next time. As it stands, it is best to say that we shouldn't use this model to predict a person's risk of having a heart attack.