ICCS261 Proposal report term project

Section 1: Topic and problem statement

1.1 What is your project title? (You can change later)

Heart Attack Risk Prediction

1.2 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.

1.3 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 ageing 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: Global Heart & Circulatory Diseases Factsheet.

Section 2: Introduction and relevant literature

Briefly describe the introduction and related literature reviews related to your topic

This is a paper on a similar topic that could be used as a reference.

Machine learning-based heart attack prediction: A... | F1000Research

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

In this section, you need to describe a plan for how the research activities will be carried out, for example,

- Where is the source of data collection

Kaggle Dataset

- 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.

- Define variables/features in the data you are going to use to perform the analysis/build the model

Patient ID - Unique identifier for each patient

Age - Age of the patient

Sex - Gender of the patient (Male/Female)

Cholesterol - Cholesterol levels of the patient

Blood Pressure - Blood pressure of the patient (systolic/diastolic)

Heart Rate - Heart rate of the patient

Diabetes - Whether the patient has diabetes (Yes/No)

Family History - Family history of heart-related problems (1: Yes, 0: No)

Smoking - Smoking status of the patient (1: Smoker, 0: Non-smoker)

Obesity - Obesity status of the patient (1: Obese, 0: Not obese)

Alcohol Consumption - Level of alcohol consumption by the patient (None/Light/Moderate/Heavy)

Exercise Hours Per Week - Number of exercise hours per week

Diet - Dietary habits of the patient (Healthy/Average/Unhealthy)

Previous Heart Problems - Previous heart problems of the patient (1: Yes, 0: No)

Medication Use - Medication usage by the patient (1: Yes, 0: No)

Stress Level - Stress level reported by the patient (1-10)

Sedentary Hours Per Day - Hours of sedentary activity per day

Income - Income level of the patient

BMI - Body Mass Index (BMI) of the patient

Triglycerides - Triglyceride levels of the patient

Physical Activity Days Per Week - Days of physical activity per week

Sleep Hours Per Day - Hours of sleep per day

Country - Country of the patient

Continent - Continent where the patient resides

Hemisphere - Hemisphere where the patient resides

Heart Attack Risk - Presence of heart attack risk (1: Yes, 0: No)

Some of these variables that are highly correlated/unique columns will be dropped. In addition, there will be new features constructed that could provide a more accurate representation of the data, in order to give the model the most amount of data to work with as possible.

Section 4: Exploratory data analysis

- Perform EDA with your data and report the summary statistics/interesting pattern with appropriated graphical representation and adequate description.

Proposal_EDA.ipynb