Cardiovascular-disease Study

TEAM NAME: Group for Cardiovascular disease data analysis

TEAM MEMBERS:
Grace Cheong A16176085
Syed Zain Ali Baquar A12732391
Jiahe Feng A15507377
Shengjie Mao A15531892
John Ge A15541533

TEAM MEMBER GITHUB IDS: gracecsyy, zainbaq , jeffrey7377, ShengjieMao, yuu6883

DATA SCIENCE QUESTION(S) & HYPOTHESIS:

Our aim is to further understand the causes of cardiovascular disease as well as to develop an algorithm that is able to non-invasively predict and diagnose such conditions. Our study will be centered around these two questions: What are the principal indicators of cardiovascular disease? To what accuracy are we able to diagnose these diseases?

We postulate that the dataset we have acquired is sufficient to make such predictions, and that there exist at least three principal indicators of heart disease.

BACKGROUND:

Cardiovascular diseases (CVDs) have emerged as the leading cause of death. According to the World Health Organization(WHO), 17.9 million people die from CVDs every year, which is approximately 31% of deaths worldwide. Through research, it has shown that the occurrence of CVDs is not random; they are often correlated with various factors such as smoking, family history, diet and access to health care services. By studying past record, it will aid us in the understanding of factors correlated with CVDs.

Symptoms of CVDs include chest discomfort and shortness of breath which are often overlooked by many patients themselves. Many patients only find out about their condition after experiencing near-fatal symptoms. With proper technological data science tools, medical professionals are able to turn large datasets into crucial information that aid the process of making informed decisions and prediction. Early diagnosis during regular check-ups is integral in saving many lives through prevention and treatments. While researching for our topic, we came across a similar project “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms” https://doi.org/10.1155/2018/3860146. Their project tested different approaches and algorithms with the goal of evaluating their performance. Obesity has been known to cause diabetes, high blood pressure, high cholesterol, all of which are factors leading to CVDs. Hence, We decided to focus our project on the correlation of obesity and CVDs.

ETHICAL CONSIDERATIONS:

In order to practice data science without doing anything ethically questionable, we carefully considered the legality, privacy concern, the bias involved in the process of data using. We examined the slides about ethical consideration from COGS 9 and COGS 108 lectures, and decided to use Deon’s data science ethics checklist (http://deon.drivendata.org/) and the project guideline to help us address all the issues with ethical considerations:

Deon’s ethics checklist is composed of five parts: data collection, data storage, analysis, modeling, and deployment:

A. Data Collection:

We did not participate in the collection of data, therefore we have no interaction with the people whose data were collected, and we do not know if there was bias in the process of data collection. We just simply obtaining data from the Internet. The data set we use is public on Kaggle and said it was "collected at the moment of medical examination." According to the term of use on Kaggle, "The materials displayed or performed or available on or through the Services, including text, graphics, data… are protected by copyright and other intellectual property laws. The Services may allow you to copy or download certain Content; please remember that just because this functionality exists, doesn't mean that all the restrictions above don't apply — they do!" We will download the dataset and perform data analysis using our software. However, we will not distort the data nor use the data for illegal purposes. The purpose of this project is just to explore the data set and answer simple questions.

Personal health is a very private subject. We also want to minimize the exposure of personally identifiable information in the dataset. The columns of the data set are age, height, weight, gender, and etc. All this information has corresponded with the patient's ID. No other personal details, including the name, or other identifiable information was given in the dataset. The dataset is totally anonymous.

B. Data Storage

Since we just explore the public data set from Kaggle. We don't necessarily have a plan to secure the dataset. However, we can assume that the data set is safe because we have downloaded the data set to local computers and if we want to share our foundings with other members in the group, we utilize the Team Google Drive associated with UCSD emails. The data set may be removed from Kaggle for various reasons or if people involved in the data collection requested it. We will delete the data set from our local computer after we finished this project.

C. Analysis

Because we are looking at the patients, not the whole population, we must have a clear picture of what hypothesis can we make, and what conclusion can we draw. For example, we will not conclude that smoking cause the disease because of general knowledge without exploration, but we will rather separate people into groups by the values of the column and find the indicator.

The visualization will correctly represent the data. We will not remove any rows and columns to maintain maximum neutrality. Our process and codes that lead to the visualization and results will be along with the conclusion and will be detailed.

D. Modeling

We will not discriminate. If we have to group people by gender for this data set, it will only be for biological exploration purposes. The groups we will divide will only reflect that they all have some common factors. Since we are doing data on biological statistics, no conclusion that discriminates people socially, racially, or politically will be made.

E. Deployment

We think that our data visualization and exploration can only lead to the benefit of the readers who are trying to find an association between different factors and cardiovascular diseases. People might find the conclusion to be helpful for giving beneficial health advice. The data and the exploration will not be used for any unethical purpose.

DATA:

Dataset Name: Cardiovascular Disease dataset

Link: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Number of Observations: 70000

Cardiovascular Disease dataset includes possible factors that may cause the disease. There are 13 columns to this dataset. They are ID, age (in days), height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol (degree compared with normal), GLUC level, smoking (whether patients smoke or not), alcohol intake (in binary feature), physical activity (in binary feature), and presence or absence of cardiovascular disease. All of these values were collected at the moment of the medical examination.

Since we are going to find out principal indicator(s) of cardiovascular disease, we will use all the data included in the dataset and will try to find out which variable(s) is/are the most direct or important indicator(s).

Team Expectations Agreement

Read over the COGS108 Team Policies individually. Then, include your group's expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member's name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team's expectations below, and have every intention to fulfill them.

These expectations are for your team's use and benefit—they won't be graded for their details. Goals should be realistic: "No group member will never miss a meeting and everyone will always show up early" is probably unrealistic, but "Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting" and "When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting" are realistic expectations. Expectations for deadlines, how you'll work together, meeting attendance and participation, and project completion should all be considered.

INCLUDE YOUR TEAM'S EXPECTATIONS HERE

1. We will all be respectful towards the opinions of others
2. We will all take ownership of being part of a team and contribute eagerly
3. We will all try our best to make it to meetings and give notice upfront if they are unable to attend
4. If there are any disagreements, we will talk it out and find a solution to it
5. After completing any individual work, we will take the effort to write an accompanying description so that other group members will be able to follow

Once completed, save this document as a PDF and submit on Gradescope (Code: 9JN2Z4). Be sure to add each team member's name to the Gradescope submission.