17/20

Project Proposal Report

DSCI550

Project Title: Analysis of Cardiovascular Disease Examination ? not

Group Members:

Leader: Kaijing Zhang (KZ)

Member: Shiming Chen (SC)

Background:

According to the World Health Organization, cardiovascular diseases (CVDs) are the most cause of death that take an estimated 17.9 million lives annually, which accounts for 31% of all deaths globally (Ref 1). Cardiovascular diseases generally refer to the ischemic or hemorrhagic disorders caused by hyperlipidemia, blood viscosity, atherosclerosis, hypertension and other heart, brain and systemic issues. These diseases seriously threaten people's lives. Not only older people, especially the age older than 50, have the characteristics of high prevalence, high morbidity and high mortality, young people also have an increasing trend of having cardiovascular diseases these days.

Need medical science?

The purpose of the analysis is to learn more about factors that cause cardiovascular diseases and determine if people are under the risk of having cardiovascular diseases.

Questions to explore in this analysis are:

1. What factors may cause cardiovascular diseases? Which factor is most related to cardiovascular diseases ?
2. Are other factors related to that factor?

Early detection and management in machine learning models is really helpful for people who have had cardiovascular disease or may have a high risk of getting it, caused by the presence of disease such as hypertension, diabetes, hyperlipidemia or other established disease, to protect themselves by reducing risks that cause diseases.(Ref 1)

Need a clear problem definition

Description of datasets:

The dataset we used is "Cardiovascular Disease Dataset" from Kaggle (Link). The dataset values were collected at the moment of medical examination about two years ago, which includes 70,000 cases. This dataset was cleaned 6 month ago, but we still choose the old one, because we

can learn more about how to solve the database while we encounter some wrong datas like outliers or something wrong due to collection. Kaggle.com is a hosting site for open datasets for data science. Kaggle has an interesting exploration facility for the dataset.

For each case, it provides 12 features: Age (age in the dataset) is objective feature and type in data set is int (days); Height (height) is objective feature and type in data set is int (cm); Weight (weight) is objective feature and type in data set is float (kg); Gender (gender) is objective feature and type in data set is categorical code; Systolic blood pressure (ap_hi) is examination feature and type in data set is int; Diastolic blood pressure (ap_lo) is examination feature and type in data set is int; Cholesterol (cholesterol) is examination feature and its meaning in dataset is 1: normal, 2: above normal, 3: well above normal; Glucose (gluc) is examination feature and its meaning in dataset is 1: normal, 2: above normal, 3: well above normal; Smoking (smoke) is subjective feature type in data set is binary; Alcohol intake (alco) is subjective feature type in data set is binary; Physical activity (active) is subjective feature type in data set is binary; Presence or absence of cardiovascular disease (cardio) is target variable type in data set is binary.

Project Plan:

Our project design drew on other data from Kaggle, which have some charts. First of all, we want to count the number of each case type for each feature and draw a chart for the appropriate data. For example, age, the number of people at each age with cardiovascular disease, and this can be represented in a pie chart as the percentage of cardiovascular disease. So you can look at the data more intuitively and come to a conclusion. Secondly, we may compare the relationship between multiple features to find the connection,trying to find out how different features relate to each other, to help people better understand cardiovascular disease. Therefore, we might use Pandas Dataframe in Python to help us process the data. We will import the CSV file into Pandas Dataframe. Data cleansing and data analysis are then performed. In addition, R language is also a good choice for data statistical analysis.

For the work division, Kaijing will focus more on the programming part and Shiming will focus more on reporting the data. My team members and I are better at different fields, so we want to give full play to our strengths to make this project as perfect as possible.

Reference

1. “Cardiovascular diseases(CVDS)”, World Health Organization, 17 May. 2017, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2 . “Cardiovascular disease”,Wikipedia. Https://en.wikipedia.org/wiki/Cardiovascular_disease

– 1 page limit.