**Coronary heart disease prediction project**

Group number : 7

Group member: Kangsheng pu

Yang Gao

Junhao Ding

Dataset link: https://github.com/junhaoding1119/ist718-finalproject-dataset

**Objective**

A large proportion of Americans suffer from chronic disease and among them, hypertension, diabetes and coronary heart disease are the most ubiquitous yet lethal diseases..National survey data reveals that by 2017 July, over 59% of American citizens are already diagnosed with at least one chronic disease, and 28% of American citizens are suffering from at least three chronic diseases.(2017, Buttorff) Among those chronic diseases, coronary heart disease can be obscure in the early stages, and patients usually will not conduct complex inspections such as Nuclear Magnetic Resonance or Invasive coronary angiography. The phenomenon eventually leads to the objective of the project, which is to diagnose coronary heart disease with fundamental examination results. The insight gained from the model will assist doctors diagnosing the disease in early phases and improve the efficiency of the examination process. The early discovery of these diseases will save the patient with more time to receive treatment, which can lead to fewer complications.

**Dataset Description:**

The data set is a record of 9999 patients’ examination results in Onondaga County, Syracuse, NY. The original data derived was provided by Stephen Wallace, the Professor in Syracuse University . The dataset is blended with synthetic data from Synthea, an open source patient data generator, to follow confidentiality agreement from the patients.

The raw dataset contains 9999 rows with 41 columns. The label variable is “Coronary\_Heart\_Disease”, which identifies whether the patient is diagnosed with CAD. Among the remaining 40 columns, 21 columns will be used as predictors and could be classified into two categories: patient’s physical records and examination results. Some examples of patient’s physical records includes the following:

Age: The age of the patients, integer.

Diabetes: Whether the patient is diagnosed with diabetes, boolean.

Hypertension: Whether the patient is diagnosed with hypertension, boolean.

The examples of examination result variables are listed below:

Glucose: The glucose level of the patient(mg/DL), float.

Heart\_rate: The heart rate of the patient(times/minute), integer

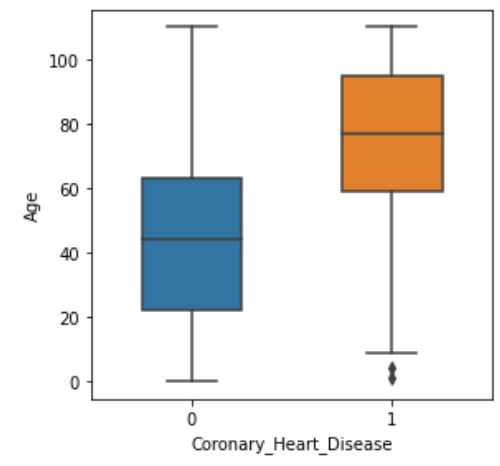
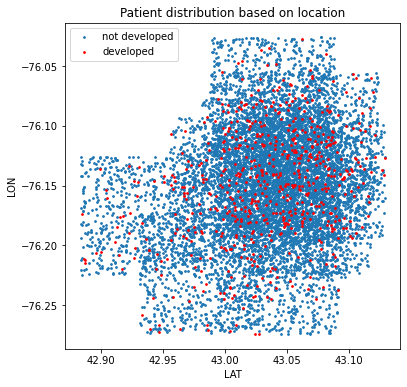
Total\_Cholesterol: The cholesterol level of the patient(mg/DL), float

Anything interesting and surprising: Before conducting any descriptive analysis, it was expected that respiratory rate and heart rate would be influenced by CAD. However, according to the classified boxplot, our group discovered that the distribution of heart rate and respiratory rate are almost identical.

**Preliminary Data Exploration**

This dataset consists of 41 variables including the label variable “cornary\_heart\_disease”, while there are also some variables that will not participate in model prediction. For example, as all units of observations are from Onondaga County, Syracuse, it might not provide any constructive information from incorporating location data into the model. After removing such data alone with functional variables such as ID and SSN, 21 variables could be potential predictors to the label variable and will participate in model prediction.

Our group conducted some descriptive analysis and attempted to discover some potential patterns. The plot below shows the distribution of patients. The graph, unfortunately, does not reveal any useful information, as both groups’ dots become denser in the middle area(LAT 43.00-43.10, LON 76.20 - 76.10).



Our group also used multiple categorized boxplots to observe whether there are any distribution differences between CAD patients and non-CAD patients. We only discovered that age has a different central tendency. The other variables have a similar central tendency and distribution for both CAD and non-CAD patients.

**Proposed Data Exploration Insights**

Our group decided to examine the central tendency and the distribution pattern based on the categorized box plots. Also, our group might conduct hypothesis tests to see whether each variable has mean differences for CAD and non-CAD patients.

**Proposed Predictions**

**T**he project aims to achieve two goal predictions: First is to predict whether the patient is diagnosed with coronary heart disease with the basic physical condition and blood examination. We plan to build models to predict whether people have coronary heart disease or not, help to assist doctors diagnosing the disease in early phases and improve the efficiency of the examination process. Next, based on the prediction and inference, we might be able to discover what factors might be the red flag of developing coronary heart disease and will be used for CAD control and prevention.

**Model Inference Insights**

Our baseline model would be a majority vote, as the data is highly unbalanced and there are only 625 units of observations that were diagnosed with CAD, which means that the baseline accuracy would be at least 93.75%. Instead of checking accuracy, assessing recall of the model would provide more valuable insights.

As the target label is a boolean class, SVM would be ideal for processing the task as it could obtain high accuracy in binary classification. Also, based on the manual diagnosing process, most doctors would check whether the patients’ certain index is within the normal range and only check on a subset of all the examination results. Decision Tree or Random Forest follow this pattern and our group would experiment on those models. After running these models, our group would check on the recall and F-1 score to determine the performance of the model, and explore which predictor values are positively correlated with CAD and which predictor values might negatively influence the possibility of developing CAD.

**Non-Spark Packages**

Our group might use the following packages for the following usages:

Collections - Counter: will be used for quick calculation for grouping data. It will be mostly used in debugging and exploratory analysis. It is also versatile in conducting descriptive statistics and assists in plot making.

Tqdm: Might be used to check the process of training models.

References

Buttorff, C. (2017, July 12). *Chronic conditions in America: Price and prevalence*. RAND Corporation. Retrieved October 20, 2022, from https://www.rand.org/blog/rand-review/2017/07/chronic-conditions-in-america-price-and-prevalence.html