Project Final Report

Project Title: Prediction of Getting Cardiovascular Diseases

Group Members:
Leader: Kaijing Zhang (KZ)
Member: Shiming Chen (SC)

Problem Definition

Cardiovascular Disease problems are medical problems, and computers can't solve it. This report will explore how to use computers to analyze the everyday issues that cause disease and prevent cardiovascular disease.

The purpose of this analysis is to learn more about factors that cause cardiovascular diseases and make a prediction on whether people are under the risk of having cardiovascular diseases. Early detection and management in machine learning models are 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) The following questions will be explored:

1. What is the correlation between each feature and the results of cardiovascular diseases?
2. Although computers can't solve medical problems, what model is best suited to study such medical problems and find correlations?

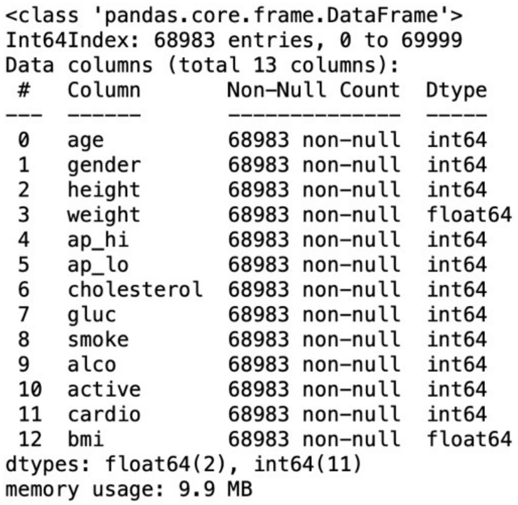
Description of Background

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. Older people, especially the age older than 50, have high prevalence, high morbidity, and high mortality. Young people also have an increasing trend of having cardiovascular diseases these days. Cardiovascular and cerebrovascular diseases are systemic vascular lesions or systemic vascular lesions in the heart and brain. The etiology mainly includes four aspects: atherosclerosis, hypertensive arteriosclerosis, arteritis, and other vascular factors; Hemodynamic factors such as hypertension; hyperlipidemia and diabetes; Leukemia, anemia, thrombocytosis, and other blood components.

Description of Dataset:

The dataset used is “Cardiovascular Disease Dataset” from Kaggle (Link). The dataset was collected at the moment of medical examination about two years ago, including 70,000 cases. This dataset was cleaned six months ago, but the old one is still chosen. It can learn more about how to solve the database while encountering incorrect data like outliers or something wrong due to collection. Kaggle.com is a hosting site for open datasets for data science. Kaggle has an exciting 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. **Description of Method Used**



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 68983 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   age          68983 non-null  int64
 1   gender       68983 non-null  int64
 2   height       68983 non-null  int64
 3   weight       68983 non-null  float64
 4   ap_hi        68983 non-null  int64
 5   ap_lo        68983 non-null  int64
 6   cholesterol  68983 non-null  int64
 7   gluc         68983 non-null  int64
 8   smoke        68983 non-null  int64
 9   alco         68983 non-null  int64
 10  active       68983 non-null  int64
 11  cardio       68983 non-null  int64
 12  bmi          68983 non-null  float64
dtypes: float64(2), int64(11)
memory usage: 9.9 MB
```

As the data source used is a .CSV file, all data will be imported to the Pandas data frame to clean the data. After checked the data and cleaned it, the project started into the part of the analysis. This process is the last step in finding and correcting identifiable errors in the data file, including checking data consistency, handling invalid and missing values, and so on. Unlike a questionnaire, data cleansing after input is done by computer rather than by hand. The process of reviewing and validating data is designed to eliminate duplication, correct errors, and provide data consistency.

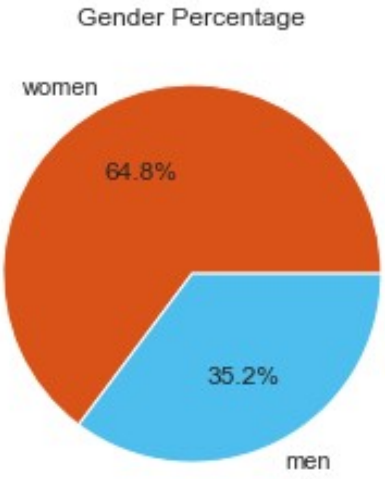
All data removed the ‘id’ variable since it is not crucial for later analysis. For the ‘age’ variable, the age data has converted the unit from days to years to get a better version for use. For blood pressure, it is abnormal if diastolic blood pressure is more significant than systolic blood pressure. Thus, drop the data which ‘ap_lo’ is more generous than ‘ap_hi’ to avoid an exceptional situation.

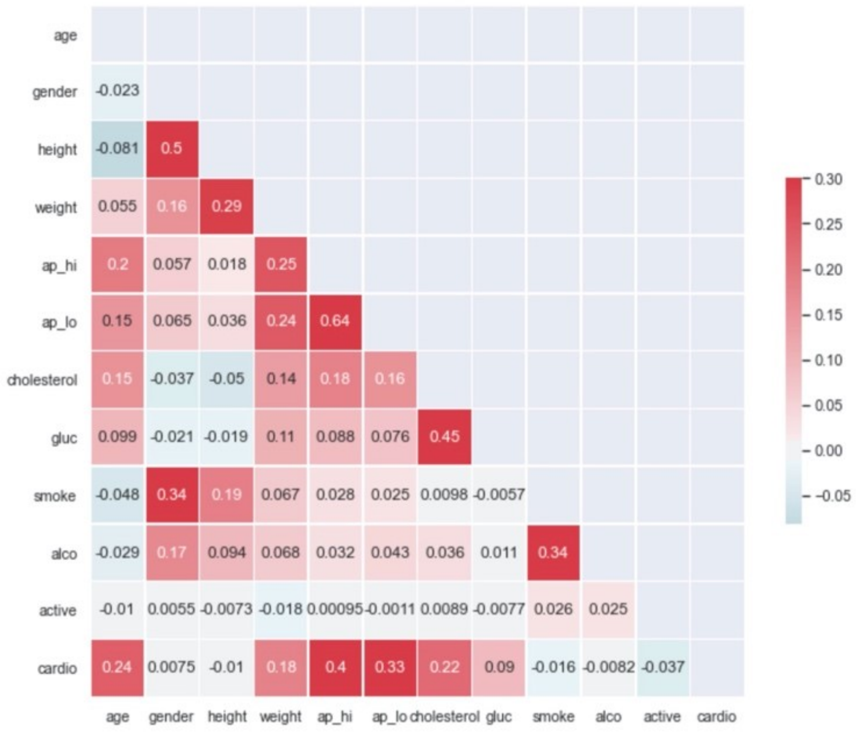
For the 70000 data, there is no null value or Nah value found. After dropping the outlier, 68983 amounts of data were kept. Due to the huge data size, dropping data will not affect the analysis result.

After data cleaning, The Program will mainly use Some methods of Pandas to analyze data. Besides, sklearn Package in Python is also a good choice to build models and analyze data. For Visualization purposes, SNS and Matplotlib will provide support.

Experiment: Analysis result

In the data set, 65 percent of the test population is female, so the report will also be explicitly conducted for women. This figure is the ratio of men to women with the disease. But that does not mean men will not be considered—women five years longer life expectancy than men. However, death from cardiovascular disease in women than men and sex on the physiological differences may play a





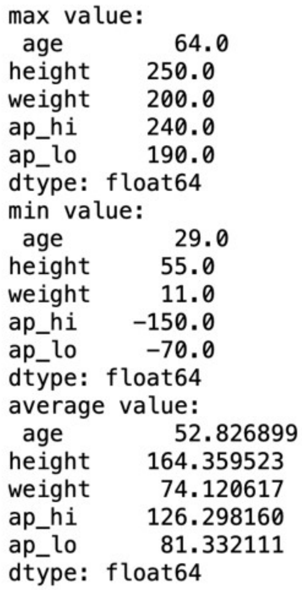
significant role in cardiovascular functions related to the physiological differences between men and women like these can lead to the onset of cardiovascular disease, susceptibility differences, the prevalence of and response to treatment. Sex differences in cardiovascular disease are associated with various factors, physiological age differences, differences in the levels of hormones (such as sex hormones), renal function difference, the difference of chromosome, mitochondrial function difference, etc. In addition, some of the physical factors may also affect the incidence of cardiovascular disease and mortality, such as living habits, consumption ideas, health level, smoking, drinking, etc.). There are also significant differences between men and women in key systems that are important in developing hypertension and cardiovascular disease, including the sympathetic nervous system, the renin-angiotensin-aldosterone system (RAAS), and the immune system. Gender-specific strategies for the treatment of hypertension and cardiovascular disease can improve outcomes in affected patients.

Making a correlation matrix by Matplotlib.pyplot to find out the correlation between features and features. Of course, the primary purpose of doing this is to explore the correlation between each segment and the Cardio target. The matrix showed that age, blood pressure (ap_hi and ap_low), and cholesterol have a higher correlation with cardiovascular diseases than other variables. The correlation values of hypertension and hypotension with cardio were 0.4 and 0.33 respectively.

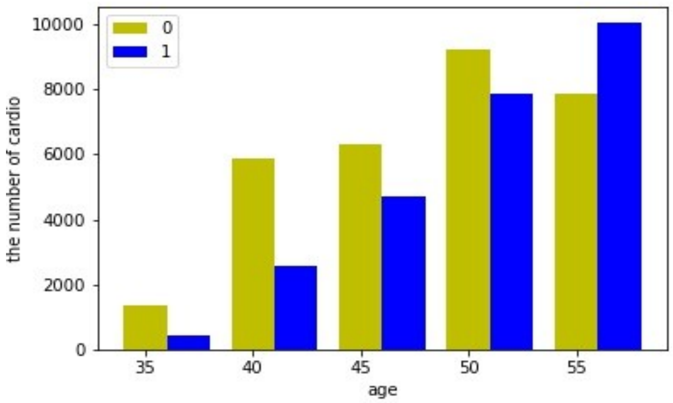
Diastolic blood pressure(ap_lo) and systolic blood pressure(ap_hi), cholesterol and glycogen, gender, and height strongly correlate. Findings of these variables may give some thoughts for analysis.

Before the correlation analysis is really started, the maximum, minimum and average values of each feature are calculated, so that the index out of range of variables is not allowed in the analysis process.

In order to find the relationship in a better way, the pictures are drawn about the feature with a higher correlation with cardio, the Bar of age. They found out that the number of people who got sick with age was much higher than the number who did not.

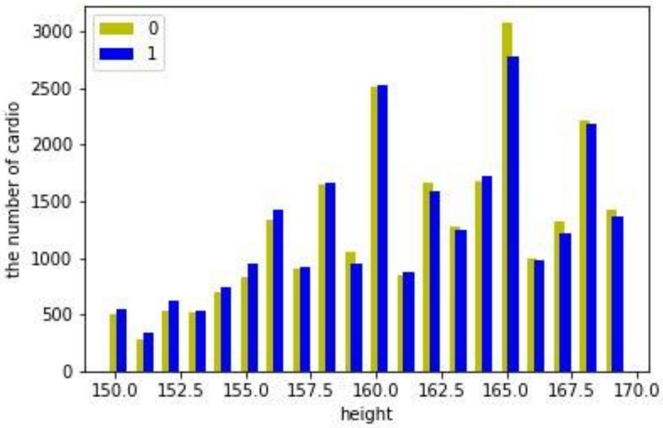


```
max value:
 age        64.0
height     250.0
weight     200.0
ap_hi      240.0
ap_lo      190.0
dtype: float64
min value:
 age        29.0
height      55.0
weight      11.0
ap_hi     -150.0
ap_lo      -70.0
dtype: float64
average value:
 age         52.826899
height     164.359523
weight      74.120617
ap_hi      126.298160
ap_lo       81.332111
dtype: float64
```

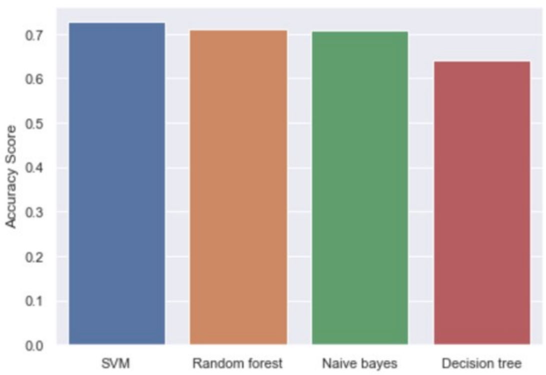


According to the correlation matrix, since height and weight are not important features for having cardio, the data are combined to Body Mass Index using the math formula and add an attribute named ‘bmi’ to react whether a person is health bass on the comprehensive consideration of height and weight.

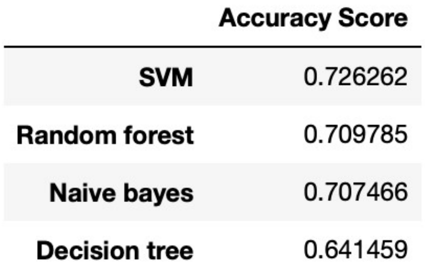
After finding the correlation matrix, the reliability of the correlation coefficients in the conclusion was verified by plotting cardio and age trends. As age increases, the proportion of people who get sick is much higher than those who don't. And height is not the more important factor affecting cardio. Both the correlation coefficient and the other graph show that, for people of different heights, the proportion of people who got sick was almost the same as the proportion who didn't.



Summary for Models



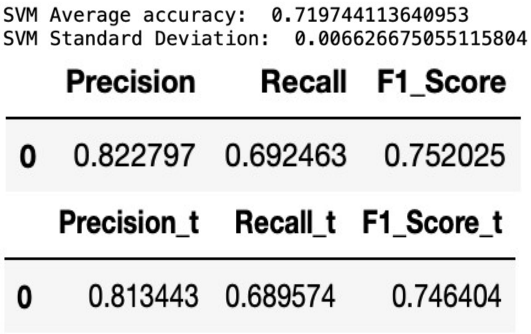
Finally, the last part of the program is a prediction model. For testing the model, the program split the data into the training set and test set with test size of 0.3. Models considered to used are Decision Tree, Random Forest, Naïve Bayes, and SVM. To choose the more precise model to use for prediction, comparing the accuracy for models is significant. By getting the accuracy score, the result found out SVM has better performance than others. Thus, the model will be used as the primary model for further evaluation. The next steps will focus on the SVM model and the decision tree model. Because SVM has a high accuracy value. The decision tree is more convenient to observe and can draw a clear conclusion.



	Accuracy Score
SVM	0.726262
Random forest	0.709785
Naive bayes	0.707466
Decision tree	0.641459

SVM model

Because of using kernel trick that maps the input data points into high-dimensional feature space, SVM has become an efficient tool to perform classification. SYM has regularization parameters so as to avoid some mistakes and overfitting. In this scenario, it's possible that some variables such as family disease background, a relative to both the occurrence of cardiovascular and other independent variables, are omitted. SVM plays a great role in decreasing the undesired impact due to these biases to the minimum. The data can get "k" pieces of outcomes from different mini-training sets from our major training set, with the help of K-Fold cross-validation. The mean of these results will be selected as the actual result. With testing the K-Fold cross-validation, the standard deviation is approximate 0.0067, which indicated our data's consistency. Using SVC from the SVM model through the test set and training set, information for accuracy, confusion matrix, precision, recall, and F1_score is gotten. For the test data set, the accuracy is about 72.63%, precision for the model is about 82.28%, recall is about 69.25%, and F1 score is about 75.2%.



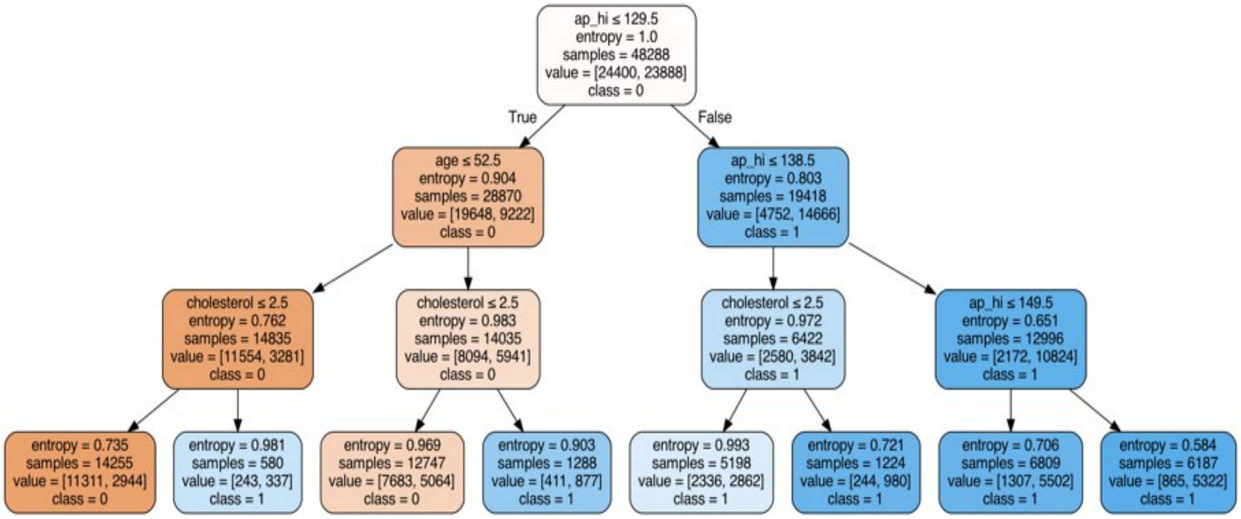
```
SVM Average accuracy:  0.719744113640953
SVM Standard Deviation:  0.006626675055115804
```

	Precision	Recall	F1_Score
0	0.822797	0.692463	0.752025

	Precision_t	Recall_t	F1_Score_t
0	0.813443	0.689574	0.746404

For the training data set, the accuracy is approximately 72.07%, precision for the model is 81.34%, recall is about 68.96%, and F1 score is about 74.64. The results were similar by comparing these data results for training and testing data sets, suggesting that the SVM model performed well in predicting Cardiovascular diseases.

Decision Tree Model



In the process of studying the model, the result shows that the decision tree has acceptable precision. After that, it was cutting off the first three lines to demonstrate. Keep using decision trees because the computational complexity is not high. The output is easy to understand, is not sensitive to the missing of intermediate values, and can deal with irrelevant characteristic data. Entropy is the uncertainty of a random variable. The higher the variable uncertainty, the higher the entropy. From the first three lines, some importance can be found that the more profound the decision tree, the lower the entropy. At the bottom, entropy is negligible. That's why to use decision trees. There are many features to train the model. Moreover, the range of these features is relatively small, and some features are even binary or ternary.

Observation and Conclusion

Base on the analysis above, higher systolic blood pressure is most significant to cause Cardiovascular diseases, and elder people has more risk to have Cardiovascular diseases. Also, higher cholesterol level has higher leading to get Cardiovascular diseases. In contrast, drinking and smoking habits which seems like bad habits for health are not much significant for having Cardiovascular diseases which are surprised. To reduce the risk of having Cardiovascular disease, individuals might control blood pressure, participate in more physical activities to get better health condition. Healthy diet is also important for reducing cholesterol level. Although drinking and smoking have little efftect on causing cardio diseases, limit the amount of drinking alcohol and smoking cigarette is effective for better physical health. Mental health is also significant, people should find happiness from their daily life. Since the data set does not consider the risk of familial inheritance for having these diseases, the prediction models may not working for individuals who has these genetic disorders.

For classification used above, both SVM and decision tree model are all getting good accuracy score which are both effective models for performing this data set. SVM get better score because it is better for classification depending on the type of the data source.

Machine learning is not an effective way to solve problems in the medical profession. But through this study, the thing learned is more about the cardiovascular disease with the help of auxiliary materials. Moreover, it is useful new knowledge to find the correlation and rule by analyzing the existing test case database. Thus, it is easy way to sort of break down the severity of cardiovascular disease into three or four levels. And each class has specific rules for the population. People can be reminded of how to prevent the disease and avoid it through these rules effectively. Then, this process can use the computer to analyze and solve.

Several of the main objectives of the project have been achieved in this project. Learn how to analyze data using Dataframe in Python. Understand how each algorithm works. Know what data is worth analyzing and useful. Find ways to explore different types of data sets.

Reference

1. “Cardiovascular Disease (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 3. “Predicting presence of Heart Diseases using Machine Learning”, https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning36f00f3edb2c