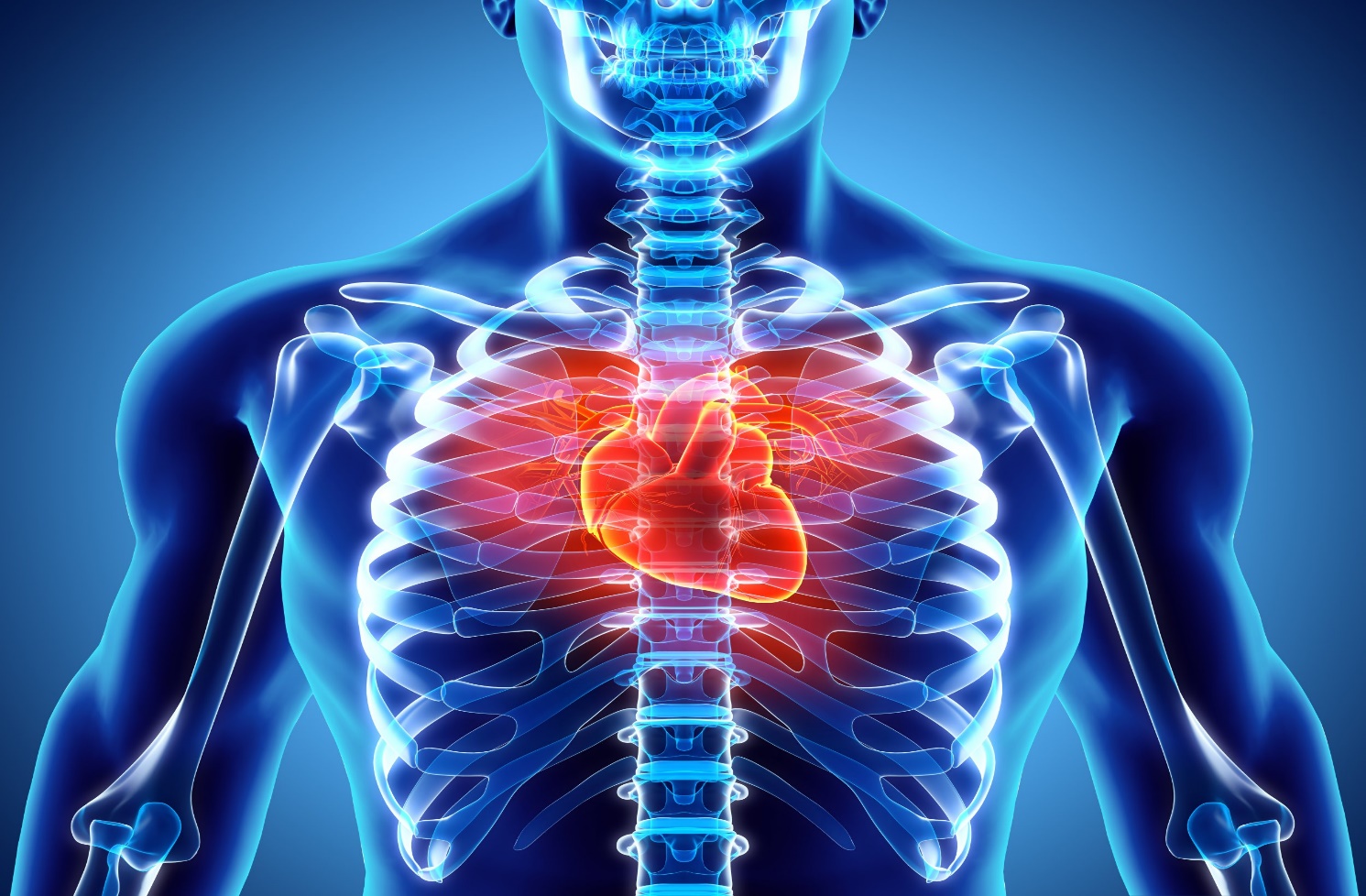
# CPTS 315

# COURSE PROJECT REPORT

# Heart Diseases Classification

# Freeman Chen



Intro:

According to WHO(world health organization), CVD can refer to a class of diseases that involves the heart or blood vessels. This disease consists of stroke, heart failure, hypertensive heart disease, rheumatic heart disease, peripheral arterial disease, and several other vascular, and cardiac problems. However, CVD which has been recognized as the leading cause of morbidity and mortality, is an important contributor to the cost of medical care [[1]](#footnote-1). In 2016, CVD was responsible for nearly one-third of all deaths across the globe, there an estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke and Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, Over the last decades, although the age-standardized mortality rates of CVD declined by 27.3%, the number of deaths increased by 42.4% from 1990 to 2015 [[2]](#footnote-2). On the other hand, CVD led to over 17 million deaths, 330 million years of life lost, and 35.6 million years lived with disability in 2017 worldwide 38% were caused by CVD[[3]](#footnote-3), “Accurate estimation of disease burden plays a key role in establishing convenient public health policies. To measure the burden of CVD, different outcome parameters can be used. There are some indices such as prevalence, incidence, mortality, and survival which can provide valuable information about the current situation and help the policy makers to organize the available resources. In general, the incidence and mortality rates of CVD vary from region to region because of several factors like lifestyle, dietary habits, appropriate health care accessibility, and so on. For example, people with a lower level of education in low-income and middle-income countries have a higher incidence of mortality from CVD”[[4]](#footnote-4), to use machine learning we can give a better aspect and evaluate the historical dataset, to find out what is the truly cause, what are those parameters are indict to, how we are going to make a change, find the difference in disease incidence and mortality rates and detect those disease as early as possible (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) so that management with counselling and medicines can begin early when the situation still under control.

With this context in mind, I set out to answer some interesting questions regrading what we were doing to detect the disease along with gaining a better understanding of how machine learning could be a key tool to help of situation we are facing. The question I stated with were simple before I did the project then I could intergrade along I do the project, some of my initial question from the project proposal were:

1. What’s the important parameters that cause CVD
2. What should I be looking at when I start with the data?
3. Which classifier will do the best job in this situation?
4. How I apply this to an application?

I set those question just by looking at the data file and did some research of how to do a data project on google, look project from GitHub, learn from the others. When I first start to do the project, I drew out my process on a white board, I decide I will first plot the graph of the dataset, using visualization to understand what’s the information that I am not catching or what’s the story behind the dataset, what’s the important information that is not provided in the context of the dataset. Second, I will apply several classify model to classify the dataset, and compare the accuracy, ROC, AUC curves, for the regression model and confusion matrix for the SVM models the selected models are:

* 1. Logistic regression model without penalty
  2. Logistic regression model with l1 penalty
  3. Logistic regression model with l2 penalty
  4. SVM-linear
  5. SVM-radial

Then I will pick the best model from the computation and make a simple API, that someone can type their health condition and machine will generate out the probability with likely hood of the heart disease predictor.

# Step 1: Visualization Dataset

For the first step, visualization dataset, I first find the shape of the data by using df.shape() function to find that the dataset has 918 observations and with 11 features and 1 binary predictor, among the features dataset, there are 6 are integer type, 1 float type and 4 object type data, the first comes up my mind is that I need to map those string data to index in order for me to go on, so I simply create a diction function for each object feature and then map to the column that contain str to index, then I use pairplot function from sns library to visualize the variables, from the graph, it could tell us what kind of the health condition relate to the heart disease the most.by looking at the graph, I found out the heart disease seems to be distinctive amongst the people with:

* more common heart diseases happen amongst elderly age however
* the people with a higher heart Rate(aka MAxHR)
* it's a higher chance for people who have high sugar compared to low
* high age with high cholesterol
* people of lower age than 45 with high hear rate have a low probability of getting a heart disease

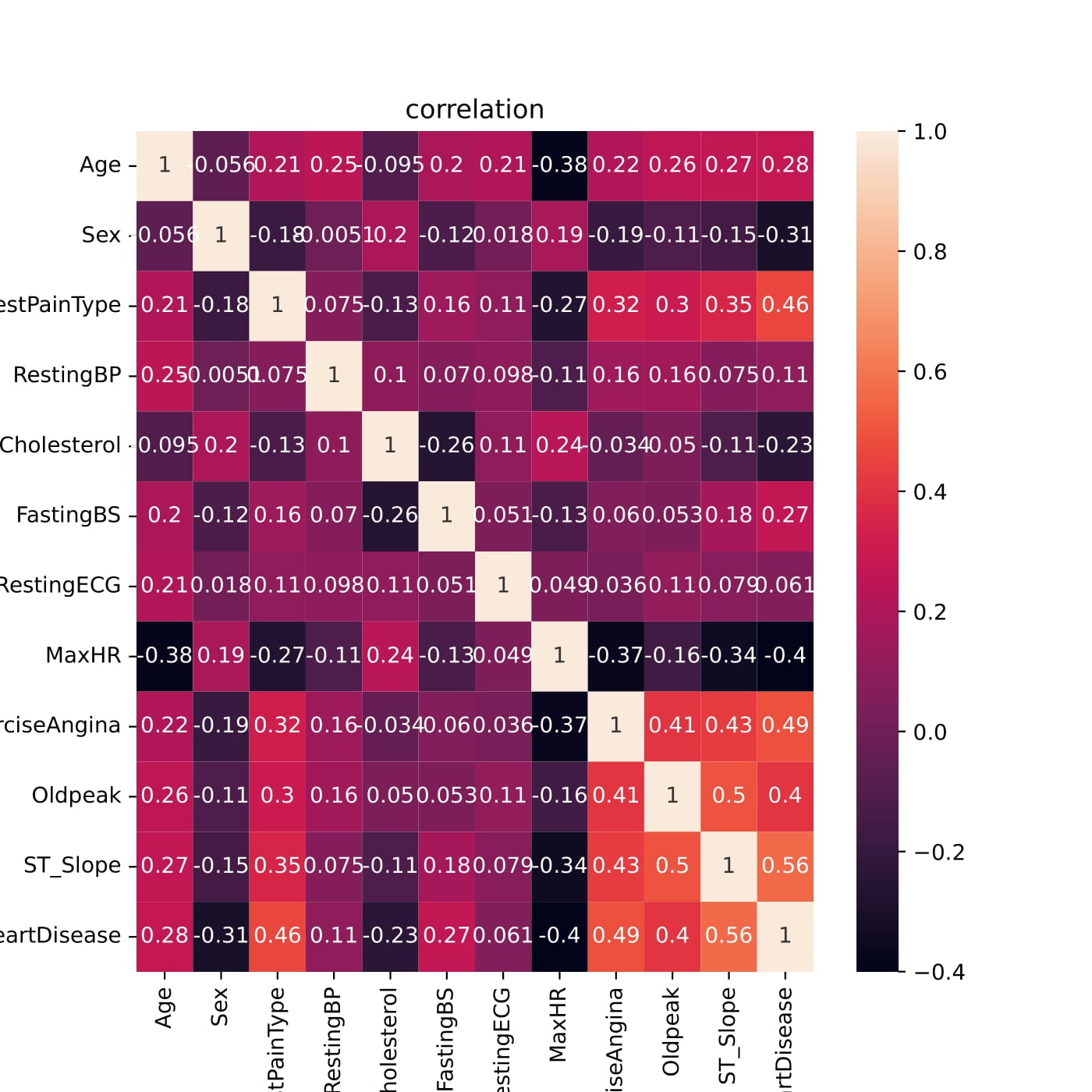
And here’s the pair graph of the plot, the plot contains a lot of information, so I have to zoom in each plot box

:

Here’s the graph: A picture containing text, furniture

Description automatically generatedforthurmore, I gmake a correlation graph across all the featrues varaibles and observed that:

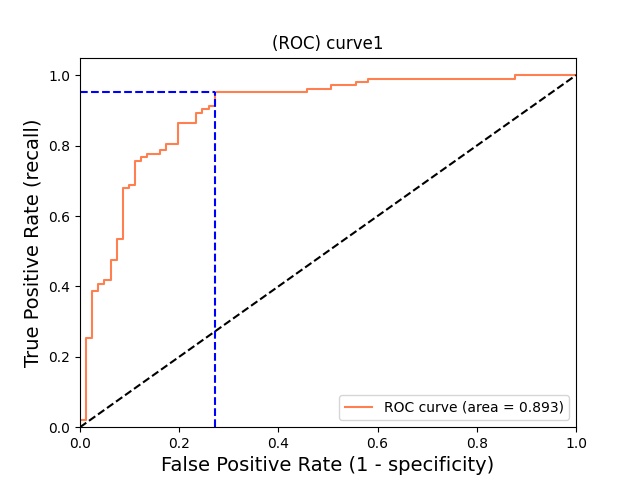
* HaxHR and Oldpeak has a highly correlated with most of features



# Step 2: training

The second step is the training process, I fist split data into training and testing with split ratio of 80:20, which training row will have 734 observations and testing set would have 184 observations, then I start to run the modes with data and print out the accuracy for each model and visualize for each model:

1. Logistic Regression without penalty: 0.832, AUC: 0.893



1. Logistic Regression with l1 penalty: 0.832, AUC: 0.895

Chart, line chart

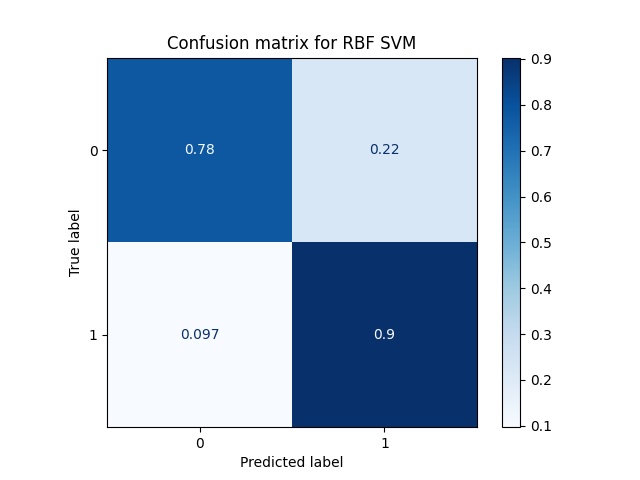
Description automatically generated

1. Logistic Regression with l2 penalty: 0.832, AUC: 0.893

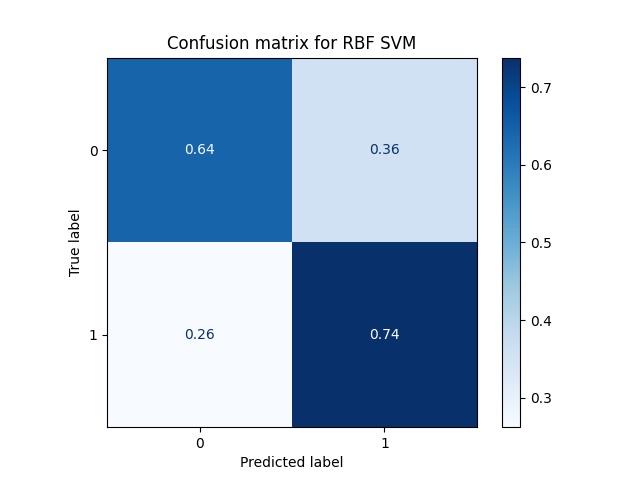
Chart, line chart

Description automatically generated

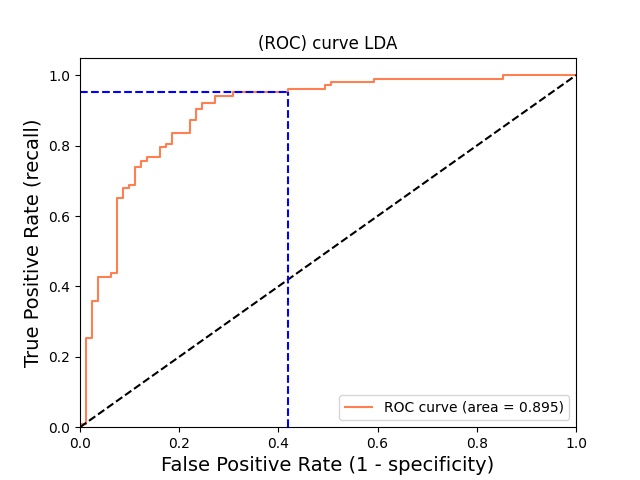
1. Linear SVM with default hyperparameters: 0.8478



1. radial svm accuracy score with default hyperparameters: 0.8478



1. LDA : 0.837, AUC: 0.895



According to each model performance, I decide use LDA model as my prediction model, because it has the highest accuracy and AUC.

# Step 3: build an application:

I saved the model as sav type file, and build a sample API that ask for people’s health record and give a prediction on the Heart condition (0: good, 1: not good) with the probability correspond to the predictor varaiable

# Conclusion:

From this project I learn a lot of the process of machine learning in gernal, unlike other computer science prject, this project I need to know not just the dataset or the process, I also need to understand how to compare the model and pick the best model that fit the data, which require a lot of the working and researching to find what is the best for this project, and there’s still a lot I need to consdier but I didn’t, such as, the balance of the dataset, normalize the dataset and lastly, using diemsion reduction method to lower the dimension of the dataset, but overall I feel I did a pretty success model and with a accury iutput,I utilized three scholarly articles to gain a better understanding of the problem .

1. Mensah GA, Roth GA, Fuster V. The global burden of cardiovascular diseases and risk factors: 2020 and beyond. J Am Coll Cardiol. 2019;74(20):2529–32. [↑](#footnote-ref-1)
2. Mensah GA, Sampson UK, Roth GA, Forouzanfar MH, Naghavi M, Murray CJ, et al. Mortality from cardiovascular diseases in sub-Saharan Africa, 1990–2013: a systematic analysis of data from the global burden of disease study 2013. Cardiovasc J Afr. 2015;26(2 Suppl 1):S6–10. [↑](#footnote-ref-2)
3. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. Lancet. 2018;392(10159):1736–88. [↑](#footnote-ref-3)
4. Rosengren A, Smyth A, Rangarajan S, Ramasundarahettige C, Bangdiwala SI, AlHabib KF, et al. Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the prospective urban rural epidemiologic (PURE) study. Lancet Glob Health. 2019;7(6):e748–e60. [↑](#footnote-ref-4)