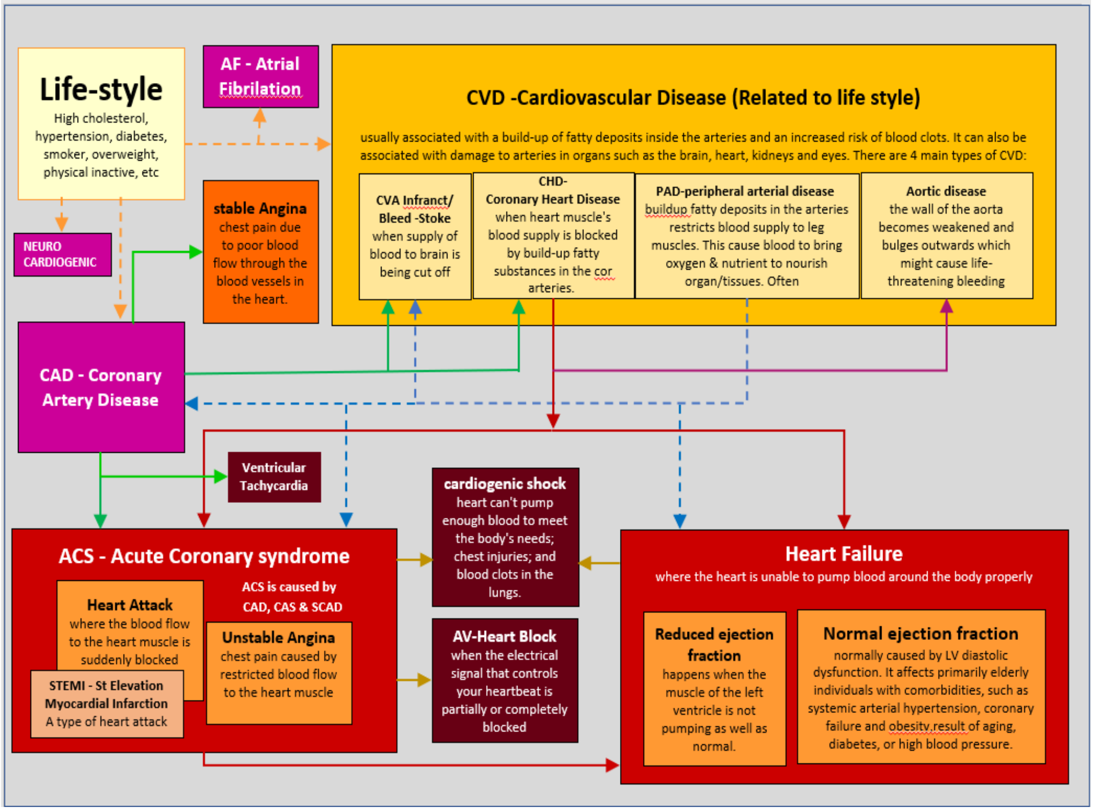
Project 1, Topic: Coronary Artery Disease Analysis & Prediction with ANN

Kimia Gholami, 903984681

Abstract:

My research focuses on biomedical and using machine learning in biomedical applications. Currently, I am working on understanding the effects of lifestyle on coronary artery disease (CAD). According to the Singapore Heart Foundation, cardiovascular disease accounted for 31.7% of all deaths in 2020 in Singapore. CAD occurs when the arteries that supply blood to the heart muscle (the coronary arteries) become hardened and narrowed due to a build-up of fatty, calcification deposits and other substances.

To understand the diseases in CAD, some conditions are mapped out in Figure 1. It looks like some diseases such as heart failure, heart attack, CHD, Stroke, cardiogenic shock, and AV heart block are caused by CAD directly/indirectly. It is said that CAD is due to lifestyle, which, if detected early, might allow monitoring and change in lifestyle to reduce the risk of cardiovascular disease. Therefore, I want to expand my knowledge in this area by focusing on predicting CAD risk.



Life-style: High cholesterol, hypertension, diabetes, smoker, overweight, physical inactive, etc

stable Angina: chest pain due to poor blood flow through the blood vessels in the heart.

CVD -Cardiovascular Disease (Related to life style)

usually associated with a build-up of fatty deposits inside the arteries and an increased risk of blood clots. It can also be associated with damage to arteries in organs such as the brain, heart, kidneys and eyes. There are 4 main types of CVD:

CVA Infranct/ Bleed -Stoke: when supply of blood to brain is being cut off

CHD- Coronary Heart Disease: when heart muscle's blood supply is blocked by build-up fatty substances in the cor arteries.

PAD-peripheral arterial disease: buildup fatty deposits in the arteries restricts blood supply to leg muscles. This cause blood to bring oxygen & nutrient to nourish organ/tissues. Often

Aortic disease: the wall of the aorta becomes weakened and bulges outwards which might cause life-threatening bleeding

ACS - Acute Coronary syndrome

ACS is caused by CAD, CAS & SCAD

Heart Attack: where the blood flow to the heart muscle is suddenly blocked

STEMI - St Elevation Myocardial Infarction: A type of heart attack

Unstable Angina: chest pain caused by restricted blood flow to the heart muscle

cardiogenic shock: heart can't pump enough blood to meet the body's needs; chest injuries; and blood clots in the lungs.

AV-Heart Block: when the electrical signal that controls your heartbeat is partially or completely blocked

Heart Failure: where the heart is unable to pump blood around the body properly

Reduced ejection fraction: happens when the muscle of the left ventricle is not pumping as well as normal.

Normal ejection fraction: normally caused by LV diastolic dysfunction. It affects primarily elderly individuals with comorbidities, such as systemic arterial hypertension, coronary failure and obesity,result of aging, diabetes, or high blood pressure.

Figure 1. CAD – Cardiovascular Disease (Related to lifestyle)

Data:

The data set is located at: https://www.kaggle.com/code/homelysmile/coronary-artery-disease-analysis-prediction/notebook.

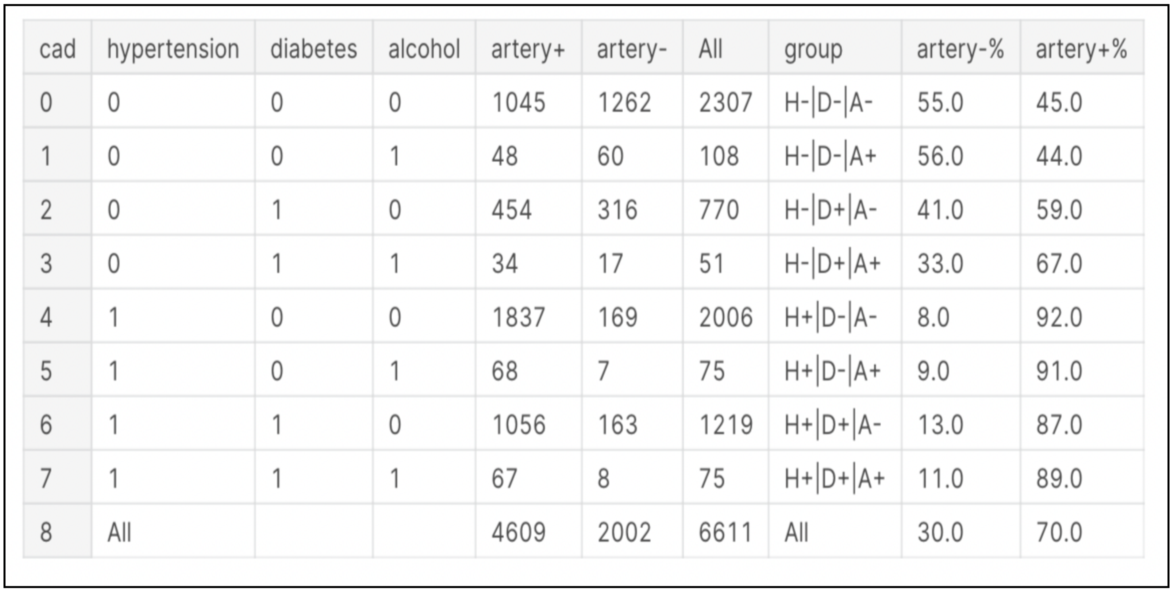
This dataset contains the CSV file of 6,612 patients with 58 columns showing different parameters such as age, gender, alcohol, smoking, etc. the data is already labeled as any sign of CAD (Labeled by 1) and no CAD (marked by 0).
20 features out of the 57 features are selected to screen their correlation with CAD patients.
The features I have used to train are: 'sno', 'type', 'day_icu', 'outcome', 'acs', 'stemi', 'heart_failure', 'hfref', 'hfnef', 'chb', 'group_age', 'group_leuk', 'group_plate', 'group_ejectf', 'cva_infract', 'cva_bleed', 'age', 'smoking', 'alcohol' and 'diabetes'.

CAD Correlation with Hypertension, Diabetes & Alcohol:

Figure 2 shows that when there is no hypertension, diabetes, or alcohol, there is a 45% risk of having CAD. With alcohol alone, the risk of CAD drops slightly to 44%. Diabetes alone increases the risk of CAD to 59% while hypertension alone increases the risk of CAD to more than 92%.

When a patient has a combination of diabetes and alcohol, the risk of CAD increases from 45% to 67% and further increases to 89% when the patient has hypertension as well.

Why does alcohol reduce the risk of CAD and why hypertension alone has a higher risk of CAD compared to a patient who also has diabetes and consumes alcohol? The risk dropped from 92% to 87%-91%.



cad	hypertension	diabetes	alcohol	artery+	artery-	All	group	artery-%	artery+%
0	0	0	0	1045	1262	2307	H-|D-|A-	55.0	45.0
1	0	0	1	48	60	108	H-|D-|A+	56.0	44.0
2	0	1	0	454	316	770	H-|D+|A-	41.0	59.0
3	0	1	1	34	17	51	H-|D+|A+	33.0	67.0
4	1	0	0	1837	169	2006	H+|D-|A-	8.0	92.0
5	1	0	1	68	7	75	H+|D-|A+	9.0	91.0
6	1	1	0	1056	163	1219	H+|D+|A-	13.0	87.0
7	1	1	1	67	8	75	H+|D+|A+	11.0	89.0
8	All			4609	2002	6611	All	30.0	70.0

Figure 2. CAD - CAD correlation between Hypertension, Diabetes and Alcohol table

Based on Figure 5, alcohol did look like it had some influence on diabetes, hypertension and CAD till age 60. When patients are in the age group 61-75, the population who consume alcohol dropped substantially. This is shown by the constant level of alcohol but the population is much higher. There were much more patients with CAD which might indicate that age factor override alcohol reduction effect. This needs further study to confirm.

If I were to look at the level of alcohol, diabetes, and hypertension when a patient didn't have CAD, it shows the same trend as those with CAD. Figure 5 might confirm that age does have a higher weight in certain age groups compared to alcohol.

Figure 3 and Figure 4 shows that the count of alcohol patients is very low. Diabetes & hypertension counts are also much lower compared to CAD. This shows that alcohol, diabetes & hypertension do have some influence on CAD. Same at the age group above 60, lower alcohol didn't reduce the count of diabetes or hypertension and this is probably when the age factor comes in.

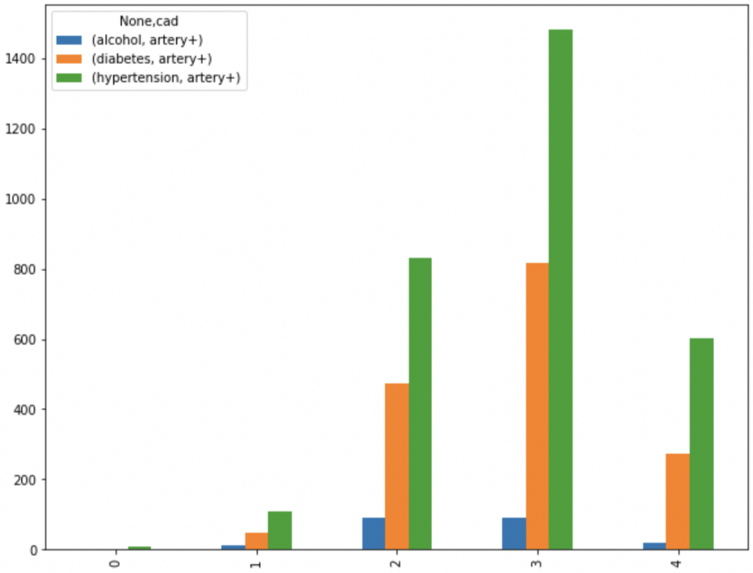


Figure 3. CAD - Alcohol, Diabetes and hypertension with group ages increasing CAD chart

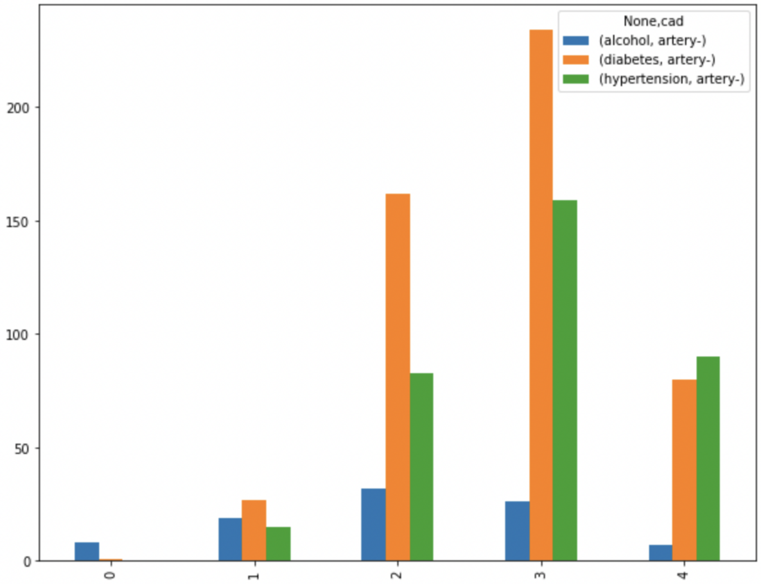


Figure 4. CAD - Alcohol, Diabetes and Hypertension with group ages decreasing CAD chart

Another possibility is, there might be some discrepancy in the definition of alcohol. In reality a small glass of wine might be good for health. Heavy drinking on the other hand will damage the heart. If there is no consistency in recording as to what is the definition of alcohol, the data might provide distorted information. This is a reminder to healthcare personnel to be standardized and have a clear definition before capturing data.

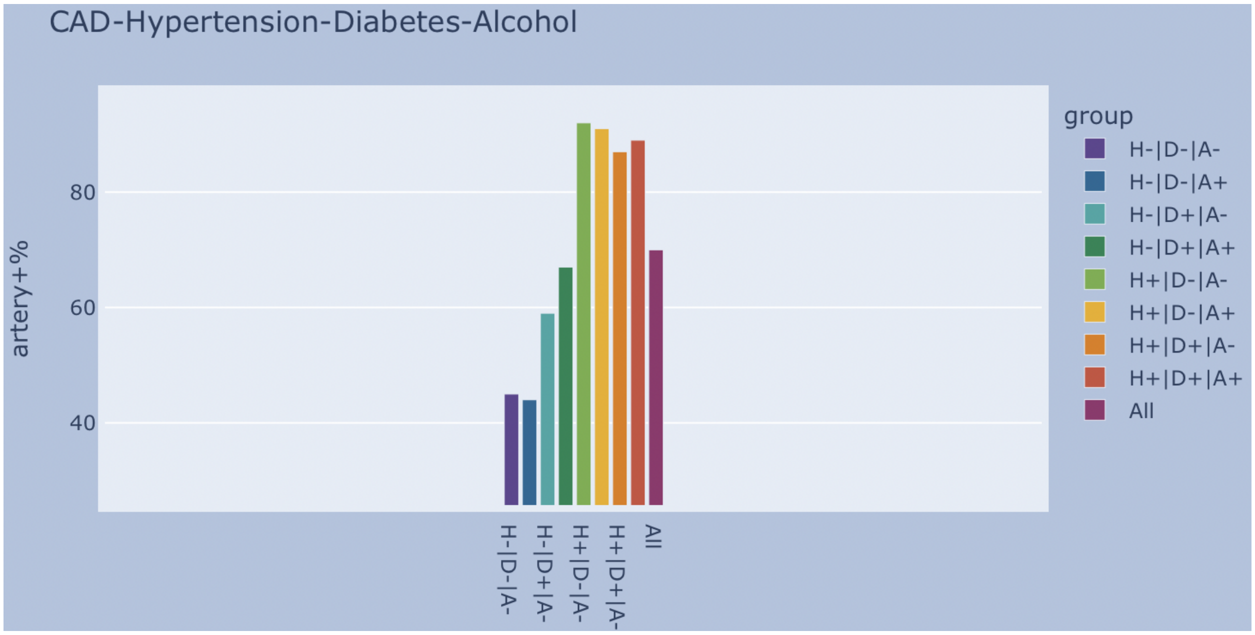


Figure 5. CAD - chart for CAD correlation with Alcohol, Diabetes and Hypertension

Glucose level for CAD patient:
A blood sugar of less than 140 mg/dL is considered a normal level. A reading between 140 mg/dl and 199 mg/dL indicates prediabetes. From the box plot below, it is obvious that glucose level is more of an indication of diabetes rather than CAD. However, the median of glucose does look higher for CAD patients (128mg/dl) compared to those without CAD (118 mg/dl) when there is no diabetes.

Hemoglobin level for CAD patients:
Hemoglobin is a protein in red blood cells that carries oxygen in the body and gives blood its red color. Below box plot shows that there are not many differences in hemoglobin levels between CAD patients and non-CAD patients.

Raised Cardiac Enzymes:
Cardiac biomarkers in the blood are a sign of heart damage, stress, or inflammation. The box plot shows that patients without CAD have no biomarkers while CAD has biomarkers.

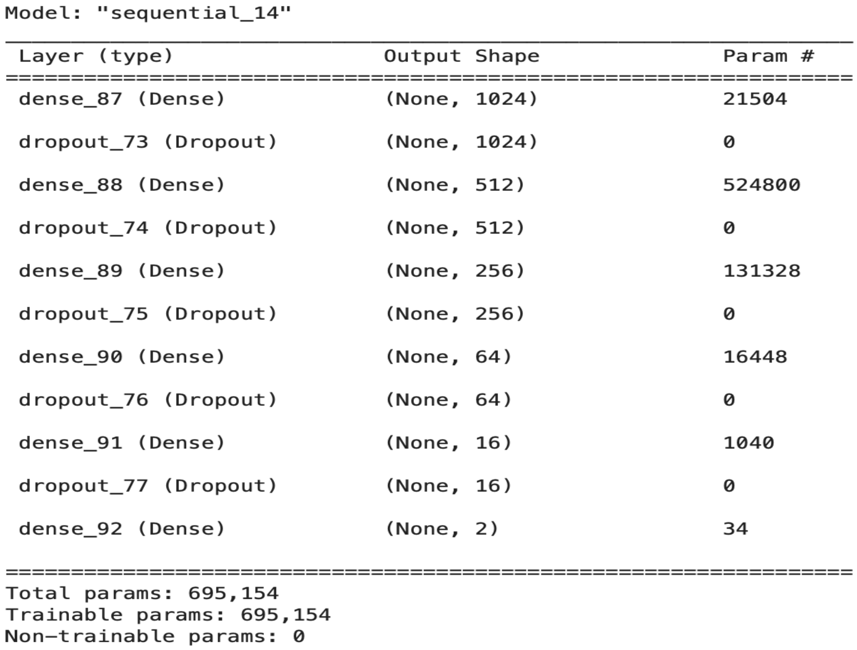
Modeling:

The network is trained by the data of 20 features in 6,612 patients with 100 epochs for each model architecture. The output is 2 classes of 0 and 1 (0 means the patient does not have CAD, and 1 means the patient has CAD). The dataset split into 60% for training and 40% for testing. Figure 6 illustrates some models and hyperparameters used to tune the network. At first, a small network was used, and I tried to find the optimization that worked better with my network. Adam's optimizer was good, but SGD had even better results, with a learning rate of 0.01. I tried different loss functions and different activation functions. I realized that by increasing the hidden layers in the network, the results might get better at some point. As I added more hidden layers, I realized that the accuracy of training data increased after some point, but the accuracy of test data or validation accuracy did not change. A 25% dropout has been used to ensure the network does not memorize the results. I also realized that regularized L2 helps the model to have better accuracy.

test	hidden layer	activation	optimization	loss	other hyperparameter	accuracy	validation accuracy
1	64,16	sigmoid	SGD(0.001)	mean square		55	56
2	64,32,16	sigmoid, softmax(last)	SGD(0.01)	mean square		56	56
3	128,64,16	relu,softmax(last)	SGD(0.01)	mean square	dropout(0.2)	60	60
4	512,256,64,16	relu,softmax(last)	adam	cross entropy	kernel_regularizer(L1=0.0,L2=0.1)	68	67
5	1024,512,256,64,16	relu,softmax(last)	SGD(0.5)	mean square	bias_regularizer(L1=0.0,L2=0.5)	70	68
6	1024,512,256,64,16	elu,softmax(last)	SGD(0.01)	mean square	bias_regularizer(L1=0.0,L2=0.5)	73	70
7	1024,512,256,128,64,32,16	relu,softmax(last)	SGD(0.01)	mean square	bias_regularizer(L1=0.0,L2=0.5)	74	68
8	1024,512,256,64,16	elu,softmax(last)	SGD(0.01)	mean square	kernel_initializer(min=-0.1, max=0.1), bias_regularizer(L1=0.0,L2=1.0), dropout(0.25)	76	72

Figure 6. CAD - different network models and hyperparameter

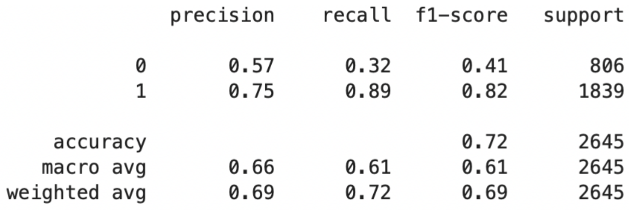
Figure 7 is the sequential model I used to train the data.



```
Model: "sequential_14"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_87 (Dense)            (None, 1024)              21504

 dropout_73 (Dropout)        (None, 1024)              0

 dense_88 (Dense)            (None, 512)               524800

 dropout_74 (Dropout)        (None, 512)               0

 dense_89 (Dense)            (None, 256)               131328

 dropout_75 (Dropout)        (None, 256)               0

 dense_90 (Dense)            (None, 64)                16448

 dropout_76 (Dropout)        (None, 64)                0

 dense_91 (Dense)            (None, 16)                1040

 dropout_77 (Dropout)        (None, 16)                0

 dense_92 (Dense)            (None, 2)                 34

=================================================================
Total params: 695,154
Trainable params: 695,154
Non-trainable params: 0
```

Figure 7. CAD - final model to train the data

Figure 8 is the best accuracy result I got on validation data.



```
              precision    recall  f1-score   support

           0       0.57      0.32      0.41       806
           1       0.75      0.89      0.82      1839

    accuracy                           0.72      2645
   macro avg       0.66      0.61      0.61      2645
weighted avg       0.69      0.72      0.69      2645
```

```
: <matplotlib.legend.Legend at 0x17c1e30d0>
```

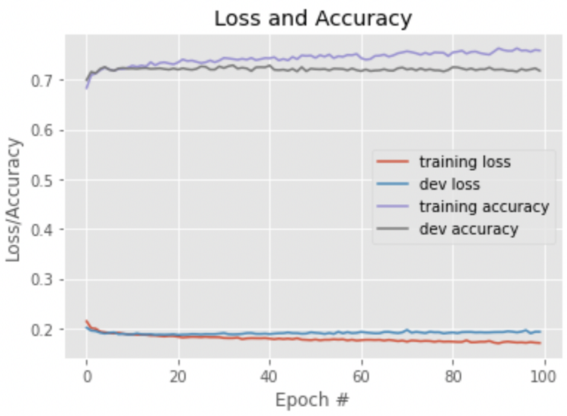


Figure 8. CAD - final result

I also included a heatmap for the binary classification in Figure 9.

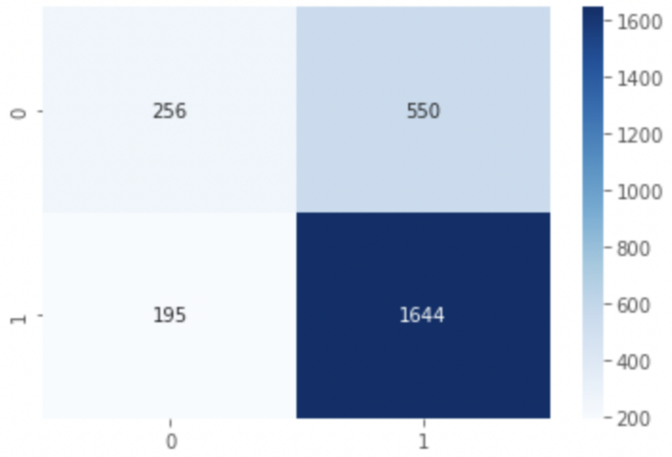


Figure 9. CAD - Heatmap visualization of the classification

Conclusion:

This project is related to the prediction of CAD by understanding the correlation with potential features. The specification and sensitivity rate is above 70% but I feel that they can be further improved if there are better features available such as cholesterol level, BMI info, more information on the type of pain or discomfort, alcohol level, exercise level, etc. This project serves as a testing ground as to whether machine learning without AI can be a reasonable potential prognosis tool to reduce machine learning time, reduce reliance on complicated/expensive/sensitive information, and reduce the costs of infrastructure for training the machine & implementation costs.