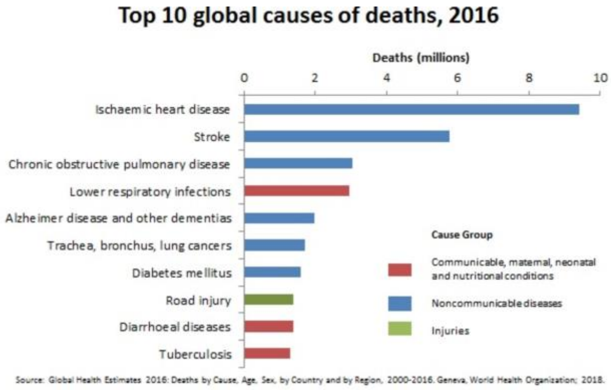
## Framing the problem:

Heart disease is one of the deadliest disease worldwide, taking away the lives of many yearly. According to World Health Organisation (WHO), Heart disease ranks first in being the main cause of death during the year 2016. This can be seen in the graph below that was taken from the WHO website.



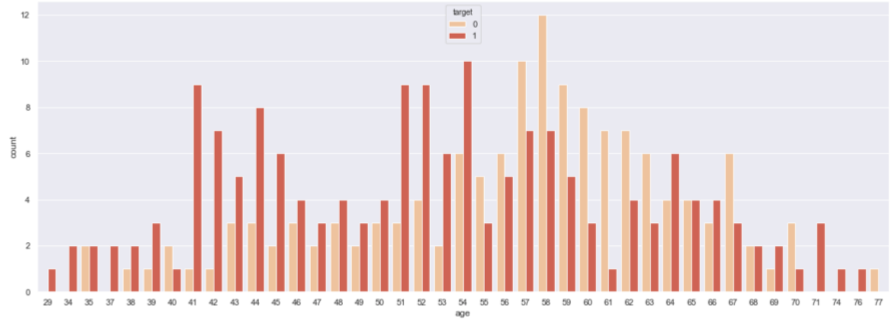
Top 10 global causes of deaths, 2016

Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

As seen from the graph, the death rate due to heart disease is quite high, hitting a toll of approximately 9.6 million in 2016. (Organisation, 2018)

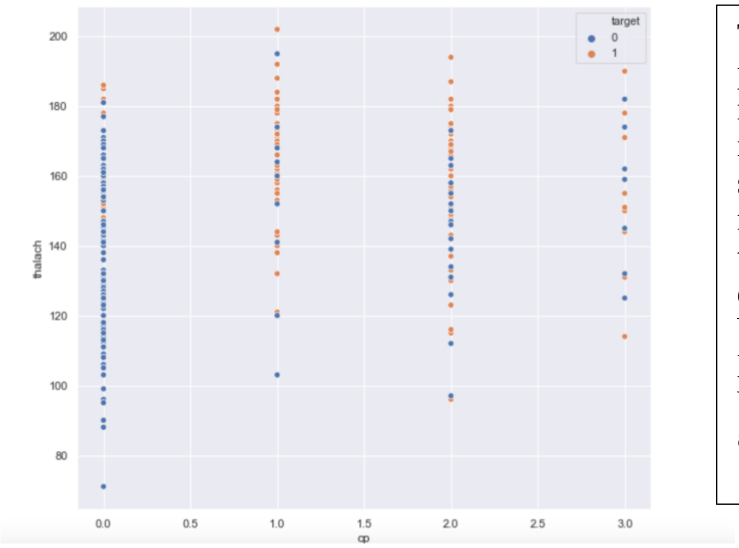
There have been many cases of sudden cardiac arrest whereby a normal, healthy person could just collapse and die from heart disease. However, cardiac arrest is not a form of heart attack, but it can happen during a heart attack. Hence, it is crucial to identify the possible factors that could lead to a high risk of heart disease. Once we have that, we could predict the likelihood of a person to have a heart attack and prevent it from occurring as early as possible. (WebMD, n.d.)

## Exploration of Data

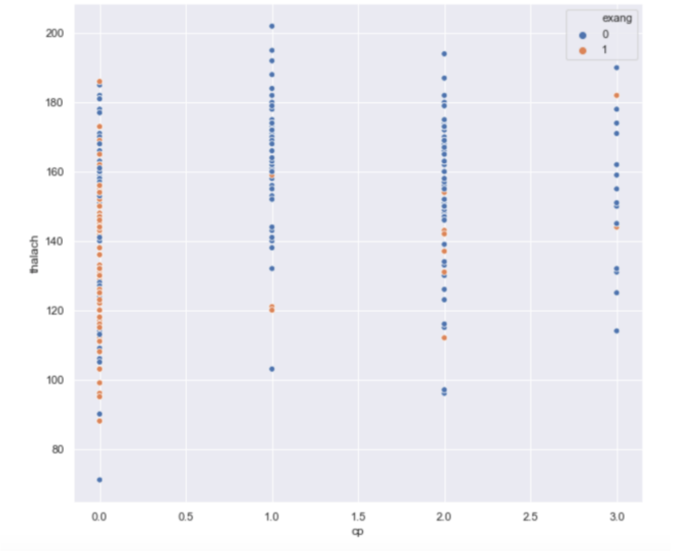


From the above graph, it shows that age group is not a very good indicator of the occurrence of heart disease, since people that have heart disease are not centered around a particular age group. Age groups of 41, 42, 44, 51, 52, 54 have particularly high counts of people that have heart disease as compared to the other age groups.

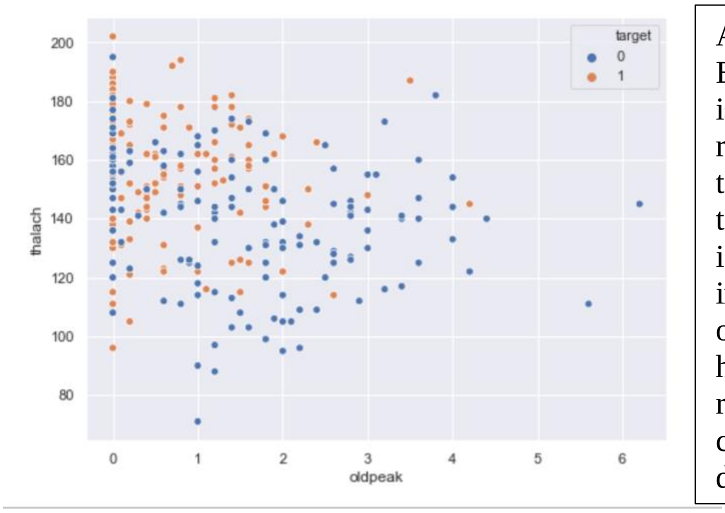
Let’s investigate the relation between chest pain (cp) and maximum heart rate achieved (thalach).



The graph on the left shows that chest pain has a dependent relationship with max heart rate. The way to calculate maximum heart rate would be to subtract your age from 220. Hence, max heart rate alone would not be able to indicate if someone has heart disease, as seen for cp=0.0. It needs to be coupled with other features. Hence, let take a look at the relationship between exang (exercise induced angina), cp and max heart rate.

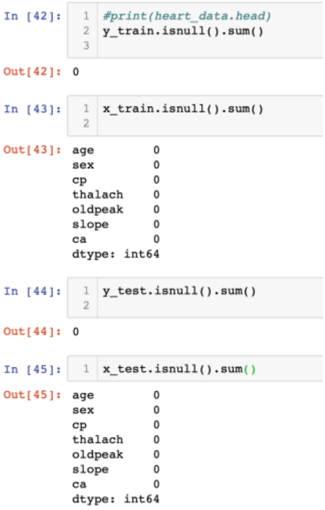


As seen from the graph on the left, if you were to compare it to the one plotted against target, having exercise induced angina does not indicate if a person has heart disease, hence it could be a good option to drop the exercise induced angina.



A ST depression test is done during an ECG to indicate the presence of ischaemia and determine if it is reversible. During the ST depression test that is induced by exercise relative to rest, a value of >2 is needed to indicate a high chance of having irreversible ischaemia. Hence a value of < 2 would indicate the presence of heart disease/ischaemia. With reference to the graph on the left, it is clear that old peak is critical in determining heart disease.

Based on the co-relation values, I will exclude exang and use cp, thalach, old peak. I will also include slope, since slope shows the slope of the peak exercise in ST segment. Ca (which is the number of major vesselscolored by flourosopy) will also be included as it shows which arteries are blocked and it's also part of the ECG test. I will include sex and age inside too, just to have a more comprehensive analysis even though it's co-relation with having heart disease is not high. Thus, I will only use cp, thalach, old peak, slope, ca, age, sex. (wikipedia-Electrocardiography, n.d.)



```
In [42]:  #print(heart_data.head)
          y_train.isnull().sum()

Out[42]: 0

In [43]:  x_train.isnull().sum()

Out[43]: age         0
         sex         0
         cp          0
         thalach     0
         oldpeak     0
         slope       0
         ca          0
         dtype: int64

In [44]:  y_test.isnull().sum()

Out[44]: 0

In [45]:  x_test.isnull().sum()

Out[45]: age         0
         sex         0
         cp          0
         thalach     0
         oldpeak     0
         slope       0
         ca          0
         dtype: int64
```

By using the .isnull().sum() function, I have also checked that the dataset for training and testing do not have null values, hence it is quite clean. This is a binary classification problem since I am only trying to predict between 2 classes. Hence, I would first use SVC to do the predictions. After that, I would use the random forest classifier and a lastly DNN.

## Discussion on ML Algorithm:

I have decided to use the Support Vector Classifier (SVC), Random Forest Classifier (RFC) and Deep Neural Networks (DNN) to analyze the heart disease dataset.

The comparative table below shows the comparisons between

| | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|
| SVC | 0.5333 | 1 | 0.6956 | 0.51 |
| RFC | 0.75 | 0.75 | 0.75 | 0.75 |
| DNN | 0.52 | 1 | 0.69 | |

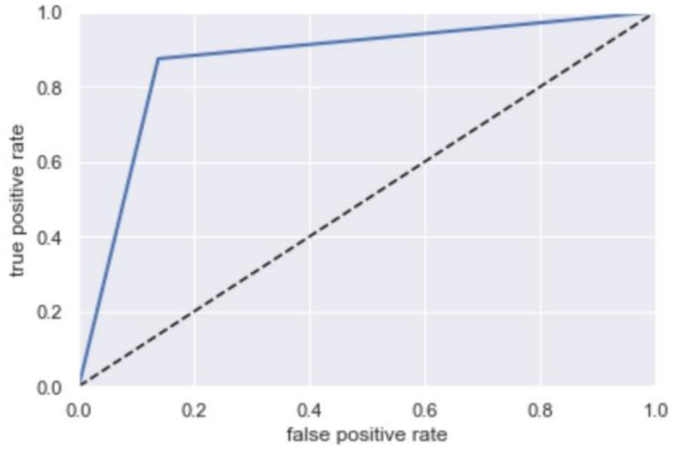
I have chosen to start with the SVC as it is one of the simplest linear models for solving both regression and classification. SVC works on the theory that a line will be created to separate the data into classes.

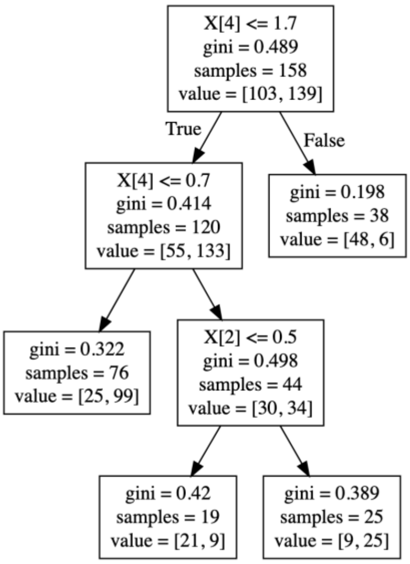
Based on the above comparative table, SVC has a precision of 0.533, recall of 1, F1 Score of 0.69 and AUC of 0.51. This indicates that the SVC is not performing very well in predicting this data set. Based on the results, it indicates that the SVC does poorly in correctly classifying those that has heart disease and those that does not. However, since its recall is 1, this indicates that SVC is able to correctly identify patients that actually have heart disease. Also, since its AUC is only 0.5, this indicates that only half of the area is being covered under the ROC curve.

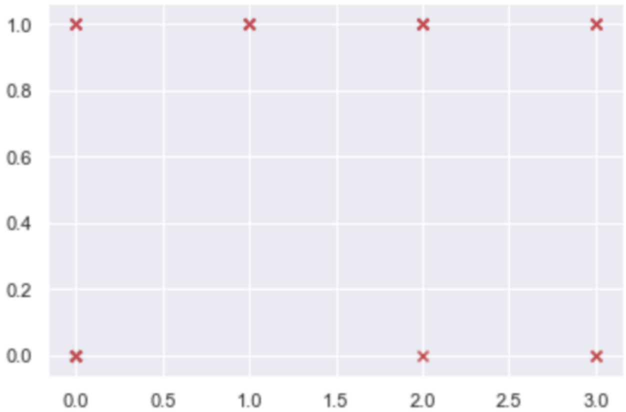
The reason that SVC did poorly with the dataset could be because there is a little randomness. Hence, I decided to use the RFC, which will add more randomness when it grows the trees and nodes. The most important features will be at the highest node/top of the tree. For the RFC, the number of estimators and maximum leaf nodes have been set at 100 and 4 respectively. The number of estimators indicates the number of trees that will be used and the higher the number of estimators, the better the classification results will be. As expected, RFC is able to correctly classify 75% of the dataset according to whether the patient has heart disease. The recall is also relatively good, giving a value of 0.75. This means the RFC is able to correctly detect that a patient has heart disease. As such, its F1 score and AUC score are much higher than SVC. Lets take a look at the DNN performance.

DNN works by having multiple hidden layers that connects between the neurons from its neighboring layers. There is a weight associated to each connection where the most important feature will have the largest weight. Each neuron uses the ReLU activation function to standardize the output. Similar to the SVC, the precision, recall of DNN is quite bad. As such, the best performing algorithm based on the precision and recall, would be RFC.

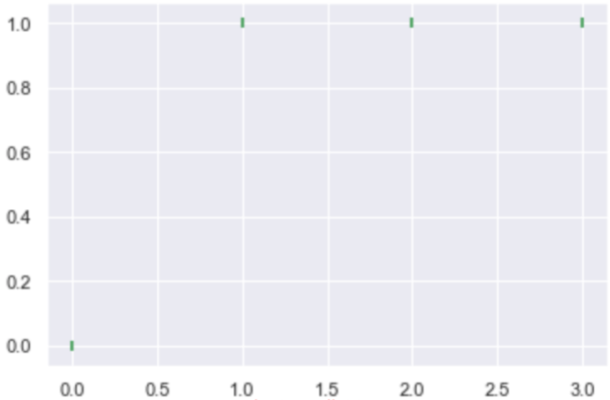
**<u>Graphs on Best performing algorithm:</u>**







cp against actual target



cp against predict target