**Data science summary, findings, and insights’**

According to the result generated from our analysis, the distribution plot from the EDA (exploratory data analysis) reveals that after standardizing the variables that are integer in other to have a uniformly distributed data, age, resting blood pressure, serum cholesterol, and maximum heart rate achieved of patients has a bell shape distribution plot which means they are normally distributed. But ST depression induced by exercise related rest is skewed to the left.

The EDA result further reveals that age, resting blood pressure, serum cholesterol, maximum heart rate achieved, and ST depression induced by exercise related rest of patients has 0, 4, 3,0,4 outlier respectively as shown by the boxplot. This means that none of the values is out of proportions for the patients; all figures are close and no extremely large figure that outshines the rest for each health factors.

EDA result using correlation to generate heat map plot in bid to check multicolinearity (the test checking if there is relations between the predicting features reveals that there is correlation between slope of peak exercise (-.39) and maximum heart rate achieved, ST depression induced by exercise related rest and maximum heart rate achieved (-.34), exercise induced angina and maximum heart rate achieved (-.38), and chest pain type and maximum heart rate achieved (-.32). The figures above show that all exercise related factors collinear with maximum heart rate achieved. We could remove some of the features and keep just one to avoid multicolinearity.

Model processing

The first part of the modeling is feature engineering using RFE, here are 5 most important features selected by RFE ('sex', 'cp', 'exang', 'slope', 'ca').

Using random forest method ('thalach', 'cp', 'thal', 'ca', 'oldpeak', 'age', 'chol', 'trestbps', 'exang', 'slope') were the 10 best feature for predicting patient heart problem.

Features like age, **chol** (serum cholesterol in mg/dl , cp (chest pain type, categorized from 0 to 3), trestbps (resting blood pressure in mm Hg), thalach (maximum heart rate achieved), exang (exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), ca (the number of major vessels, and thal (thalassemia, categorized as 1 for normal, 2 for fixed defect, and 3 for reversible defect) were best in predicting heart problems. Although, this conclusion was made using randomforest.

I would say the data is a well fit data for this kind of analysis because it encapsulate most important figure needed for predicting heart disease like chest pain, cholesterol, maximum heart rate among others. But it was surprising to see that fasting blood sugar could not make list of important feature generated by RFE and randomforest.

In conclusion, the feature with the most predictive of heart disease is maximum heart rate achieved as seen in the feature importance plot from the feature engineering. Taking a peep at how the models performed, we would see that random forest performed the best with F1=.88, accuracy=.89 which is the highest among all models. Reflecting on the importance of the various health metrics in predicting heart disease, maximum heart rate seems the most crucial. Final, the limitation of this analysis is that the data used for the analysis has good factors but the sample size is small and has really messed values.