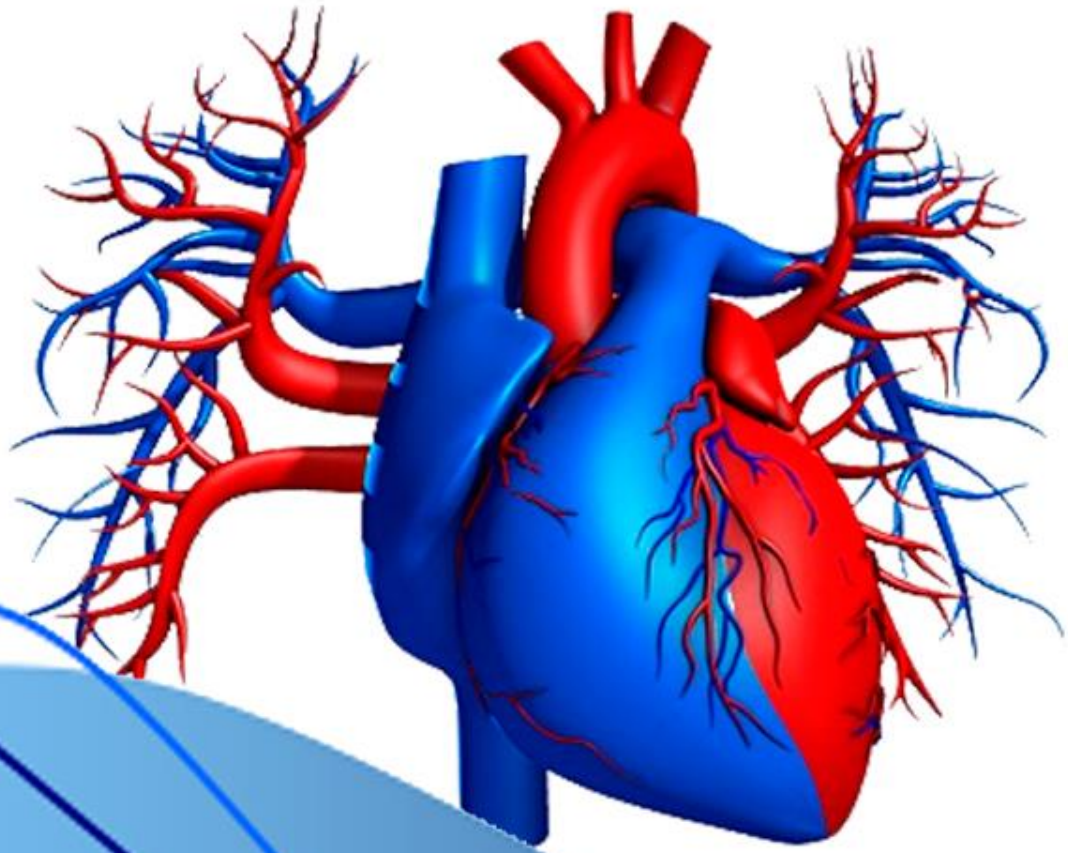
Prediction Of Patient Risk Of Heart Disease within 10 years



Presented By

Alan Yeh

Problem Statement

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in other developed countries are due to cardio vascular diseases.

As a Cardiologist, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications.

Base on the most relevant/risk factors of heart disease to make prediction on the overall risk of Heart disease within 10 years using Machine Learning Algorithm.

Data Preparation

Data Source : Kaggle Heart Disease datasets

Data At A Glance

Gender	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4240 entries, 0 to 4239
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Gender           4240 non-null   int64
 1   age              4240 non-null   int64
 2   education        4135 non-null   float64
 3   currentSmoker    4240 non-null   int64
 4   cigsPerDay       4211 non-null   float64
 5   BPMeds           4187 non-null   float64
 6   prevalentStroke  4240 non-null   int64
 7   prevalentHyp     4240 non-null   int64
 8   diabetes         4240 non-null   int64
 9   totChol          4190 non-null   float64
 10  sysBP            4240 non-null   float64
 11  diaBP            4240 non-null   float64
 12  BMI              4221 non-null   float64
 13  heartRate        4239 non-null   float64
 14  glucose          3852 non-null   float64
 15  TenYearCHD       4240 non-null   int64
dtypes: float64(9), int64(7)
```

EDA & Data Processing

Missing Data : 489 rows, constitute about 12% of entire dataset as shown below. Will drop the missing data

Drop Column : ‘education’ column is drop as not a important variable

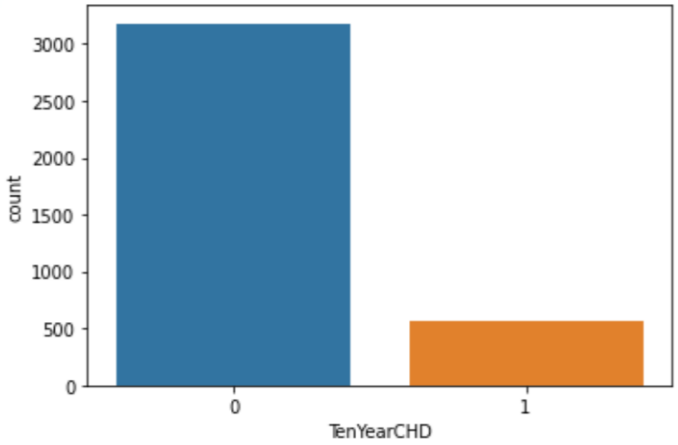
```
Gender               0
age                  0
education          105
currentSmoker        0
cigsPerDay          29
BPMeds              53
prevalentStroke      0
prevalentHyp         0
diabetes             0
totChol             50
sysBP                0
diaBP                0
BMI                 19
heartRate            1
glucose            388
TenYearCHD           0
dtype: int64
```

EDA & Data Processing

Imbalance Class Target Problem :

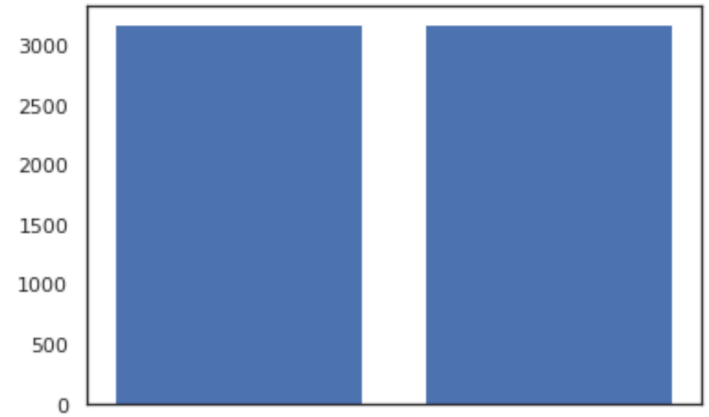
Class variable is skewed towards the ‘0’ class.

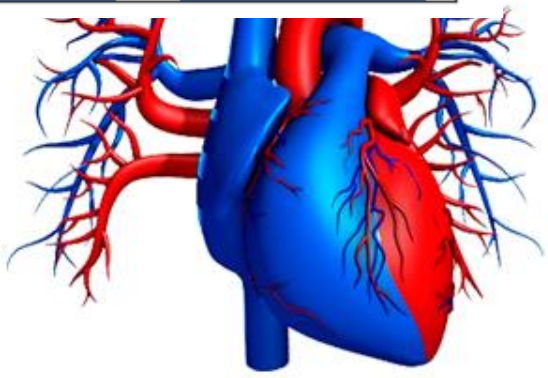
Imbalance Class Problem



SMOTE Oversampling

```
Counter({0: 3179, 1: 3179})
<BarContainer object of 2 artists>
```





Data Preparation, Training and Testing

Divide Independent and dependent variable into separate variable, set y as ‘TenYearCHD’ and rest as x.

Hyperparameter fine tuning GridSearchCV technique for 4 ML models, Logistic Regression, KNN, DecisionTree and Random Forest before training the models.

Use Best estimator from GridSearchCV

To train the models

Test the model by making prediction

on test data

Model Evaluation

Model scoring metrics base on Classification report and Confusion matrix.

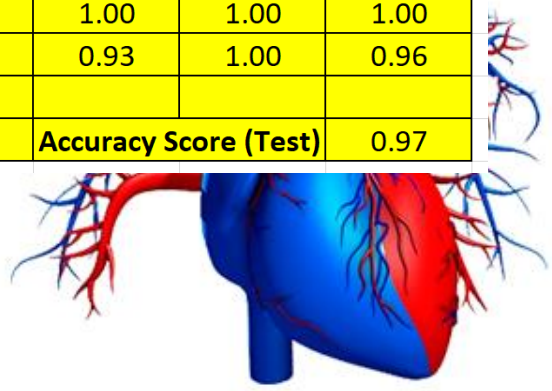
Summary of model scoring as shown.

	Logistic Regression		
Target	precision	recall	f1 score
0	1.00	1.00	1.00
1	0.93	0.92	0.96
	Accuracy Score (Test)		0.97

	K Nearest Neighbors		
Target	precision	recall	f1 score
0	1.00	1.00	1.00
1	0.92	0.92	0.92
	Accuracy Score (Test)		0.97

	Decision Tree		
Target	precision	recall	f1 score
0	1.00	1.00	1.00
1	0.93	1.00	0.96
	Accuracy Score (Test)		0.97

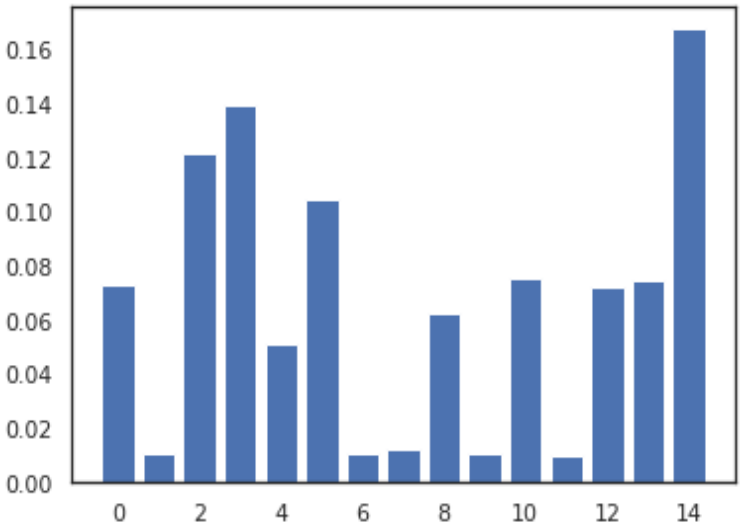
	Random Forest		
Target	precision	recall	f1 score
0	1.00	1.00	1.00
1	0.93	1.00	0.96
	Accuracy Score (Test)		0.97



Conclusion

Base on scoring metrics, Random Forest Model serve the best ML model to make prediction on the patient's risk of Heart Disease within 10 years

Features that are of great importance in Random Forest classification model is glucose (feature 14) follow by current smoker (feature 3)



Future Opportunities

If I have more time…….

I would probably choose stocks that are of great potential for growth and predict the price growth