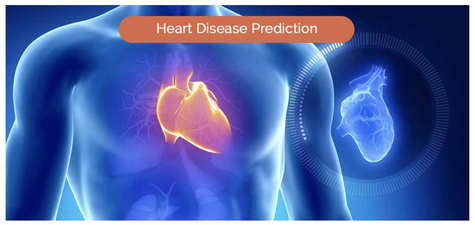
HEART DISEASE PREDICTION

Ruitong Liu, Ying Liu, Chaoying Luo, Bo Yang, Yiyang Sheng

Background

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression, decision tree and random forest.



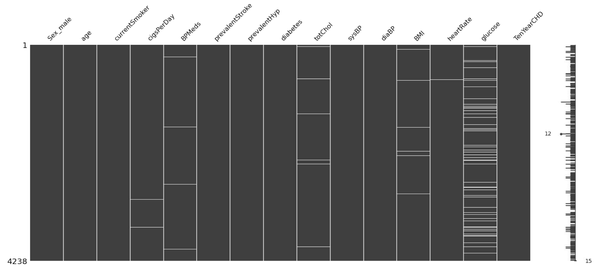
Dataset

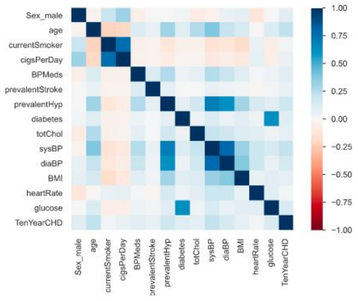
The dataset is publicly available on the Kaggle website. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

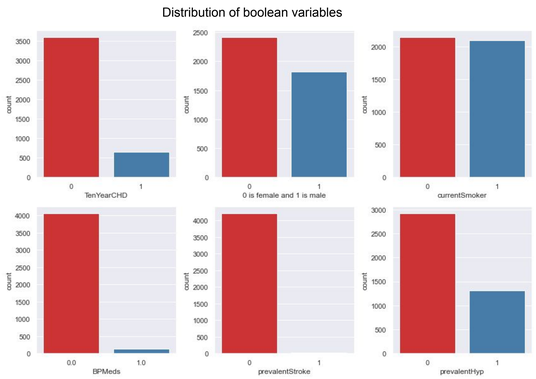
- Sex
- Age
- Current Smoker: whether or not the patient is a current smoker
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.
- BP Meds: whether or not the patient was on blood pressure medication
- Prevalent Stroke: whether or not the patient had previously had a stroke
- Prevalent Hyp: whether or not the patient was hypertensive
- Diabetes: whether or not the patient had diabetes
- Tot Chol: total cholesterol level
- Sys BP: systolic blood pressure
- Dia BP: diastolic blood pressure
- BMI: Body Mass Index
- Heart Rate: heart rate
- Glucose: glucose level
- 10 year risk of coronary heart disease CHD

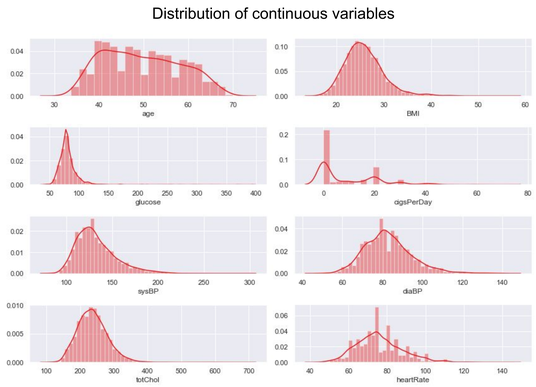
Data Cleaning

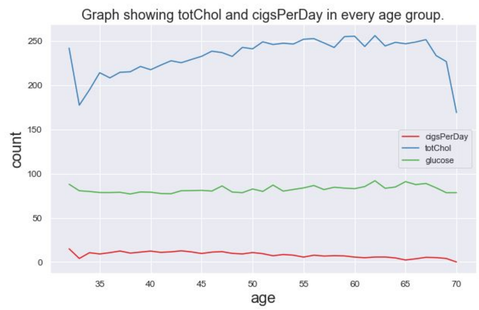




EDA







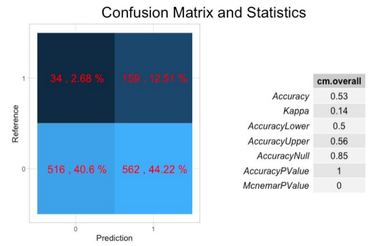
Logistic Regression

```
Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3368  -0.5844  -0.4280  -0.2956   2.8490

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.420864   0.768718 -10.954  < 2e-16 ***
Sex_male1        0.529831   0.119547   4.432 9.34e-06 ***
age              0.059878   0.007341   8.157 3.44e-16 ***
currentSmoker1  -0.105741   0.175227  -0.603 0.546209
cigsPerDay       0.022583   0.006907   3.269 0.001078 **
BPMeds1          0.388775   0.263760   1.474 0.140489
prevalentStroke1 1.004460   0.556026   1.806 0.070840 .
prevalentHyp1    0.138032   0.154691   0.892 0.372227
diabetes1       -0.165525   0.368816  -0.449 0.653575
totChol          0.001152   0.001221   0.944 0.345178
sysBP            0.016231   0.004214   3.852 0.000117 ***
diaBP           -0.000530   0.007092  -0.075 0.940425
BMI             -0.010621   0.013977  -0.760 0.447311
heartRate        0.001006   0.004615   0.218 0.827415
glucose          0.010481   0.002663   3.936 8.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2528.9  on 2966  degrees of freedom
Residual deviance: 2230.9  on 2952  degrees of freedom
AIC: 2260.9
```

```
Call:
glm(formula = TenYearCHD ~ Sex_male + age + cigsPerDay + prevalentStroke +
    sysBP + glucose, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2916  -0.5836  -0.4309  -0.2990   2.8200

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.644497   0.466318 -18.538  < 2e-16 ***
Sex_male1        0.501885   0.115848   4.332 1.48e-05 ***
age              0.061005   0.007059   8.642  < 2e-16 ***
cigsPerDay       0.019876   0.004587   4.333 1.47e-05 ***
prevalentStroke1 1.119244   0.552888   2.024   0.0429 *
sysBP            0.018655   0.002388   7.813 5.60e-15 ***
glucose          0.009474   0.001907   4.968 6.77e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2528.9  on 2966  degrees of freedom
Residual deviance: 2236.1  on 2960  degrees of freedom
AIC: 2250.1

Number of Fisher Scoring iterations: 5
```

Significant variables include Sex_male, age, cigsPerDay, prevalentStroke, sysBP, and glucose. We can run logitic regression again with these significant variables.

Since the model is predicting Heart disease too many type II errors is not advisable. A False Negative (ignoring the probability of disease when there actually is one) is more dangerous than a False Positive in this case. Hence in order to increase the sensitivity, threshold can be lowered. 0.1 threshold is adopted.



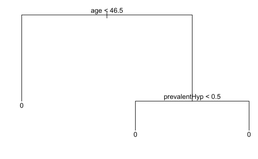
Now as we have a good accuracy for predicting people with chances of heart disease in next 10 years, we can treat them and guide them well in advance.

Decision Tree

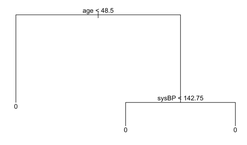
Building a decision tree with all the variables, we have an accuracy around 85.75% with a 16.14% misclassification error rate. And the variables included are age and prevalentHyp. If we use the significant variables from logistic regression, the predict accuracy is 85.18% with a 15.57% misclassification error rate. And the variables included are age and sysBP. although the first tree gave us a slightly better prediction accuracy, the second tree has a lower error rate and all the variables are significant variables. Therefore, the second tree has a better performance.

Decision Tree

```
Misclassification error rate: 0.1614 = 342 / 2119
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1817  302
         1    0    0

               Accuracy : 0.8575
                 95% CI : (0.8419, 0.8721)
    No Information Rate : 0.8575
    P-Value [Acc > NIR] : 0.5153

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16
```



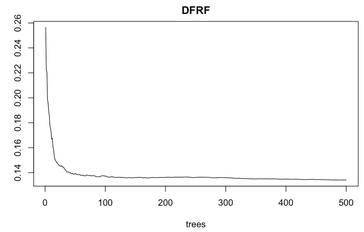

```
Misclassification error rate: 0.1557 = 330 / 2119
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1805  314
         1    0    0

               Accuracy : 0.8518
                 95% CI : (0.836, 0.8667)
    No Information Rate : 0.8518
    P-Value [Acc > NIR] : 0.5151

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16
```



Random Forest

age	11.6070253	16.7266999
currentSmoker	2.6635266	1.2461663
cigsPerDay	2.2489689	6.4644841
prevalentHyp	7.6938274	2.9433389
diabetes	2.5322369	0.4492975
totChol	1.0916327	17.0078529
BMI	1.8202488	20.5723299
heartRate	-1.0991501	13.4242186
glucose	-0.6624642	14.9095989

The importance decreases when "MeanDecreaseAccuracy" in the table aboce decreases. So set "MeanDecreaseAccuracy" in decreasing order, then it shows that "age", "prevalentHyp" and "currentSmoker" are the top 3 significant predictors.



The error is minimized at roughly 100 trees.

Conclusion

- In terms of accuracy, the best model is the decision tree. In terms of variable abundance, the best model is the logistic regression.
- The patient that was hypertensive or had diabetes shows increasing odds of having heart disease. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease.
- Men seem to be more susceptible to heart disease than women.
- Meanwhile, people under the age of 48.5 have a very low probability of suffering from heart disease.