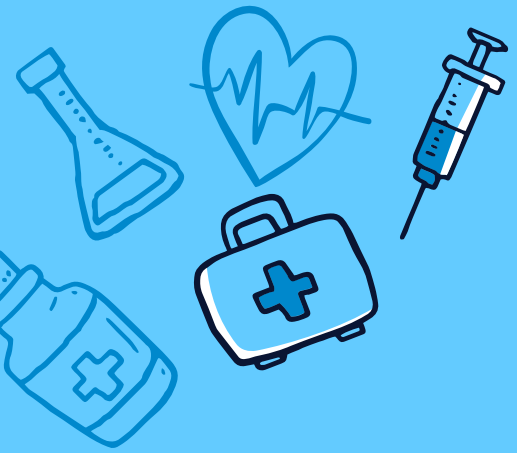
Heart Disease Prediction

Jaya, Lulu, Simon, Felix

Table of Contents



01

Introduction

690,882

People died of heart disease in 2020

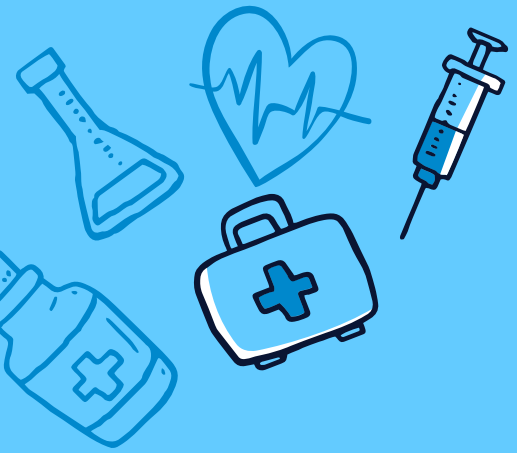

This accounts for about 25% of all deaths in the US

Data Overview

6208
4220 training observations
1808 testing observations

19
7 numerical predictors
12 categorical predictors

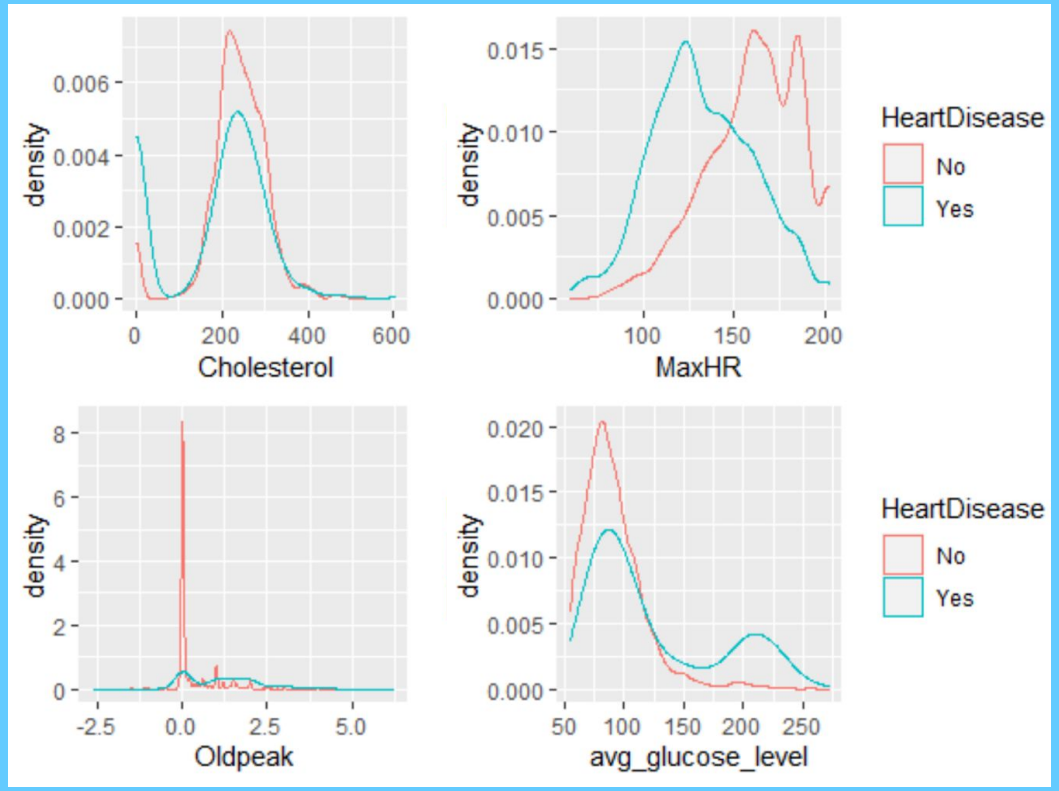
918
631 NA's in training data
287 NA's in testing data



02

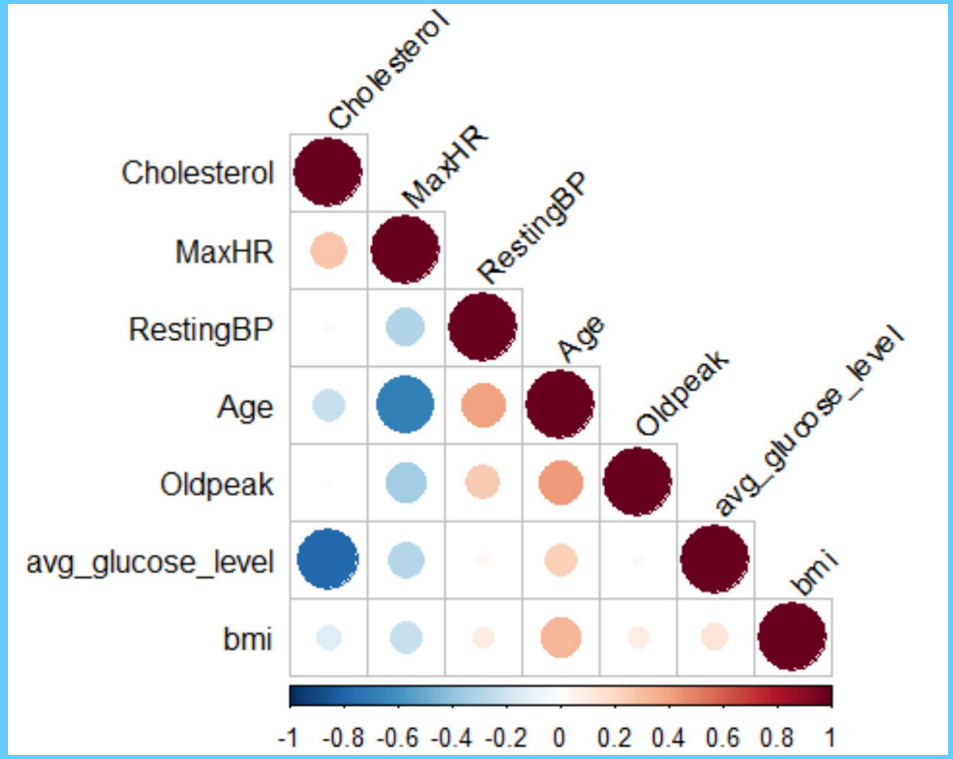
EDA and Data Cleaning

Numerical Predictors



Potential good candidates

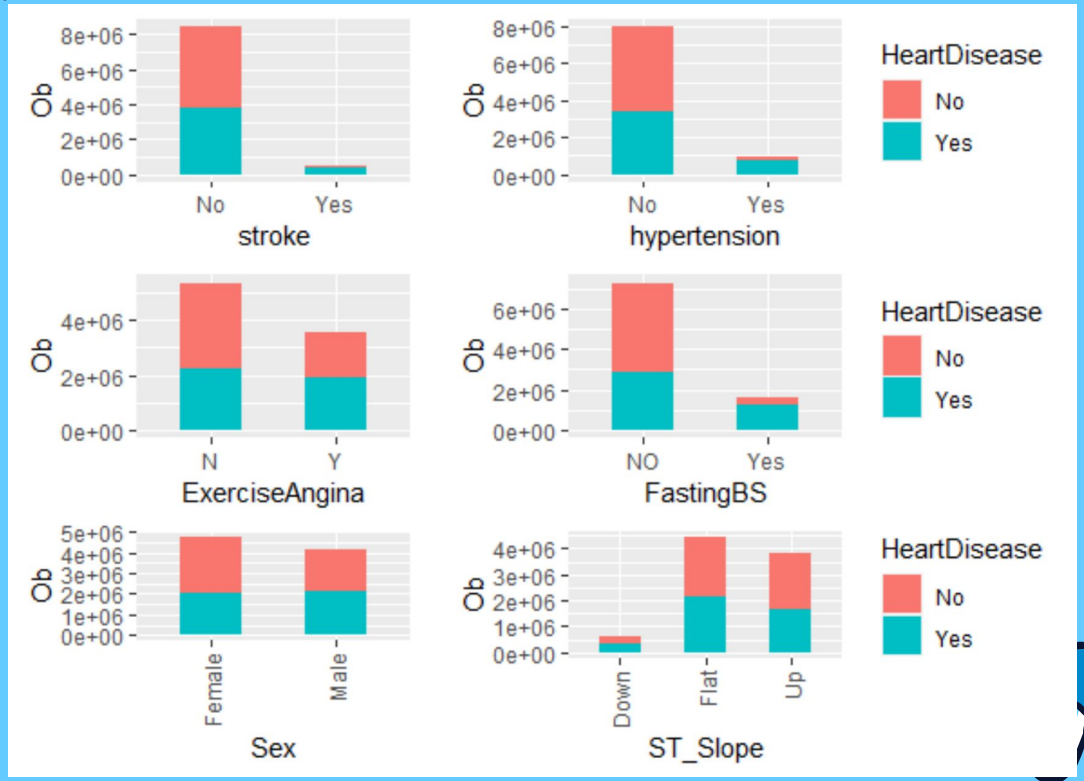
Collinearity



High correlation among predictors such as avg_glucose_level and Cholesterol

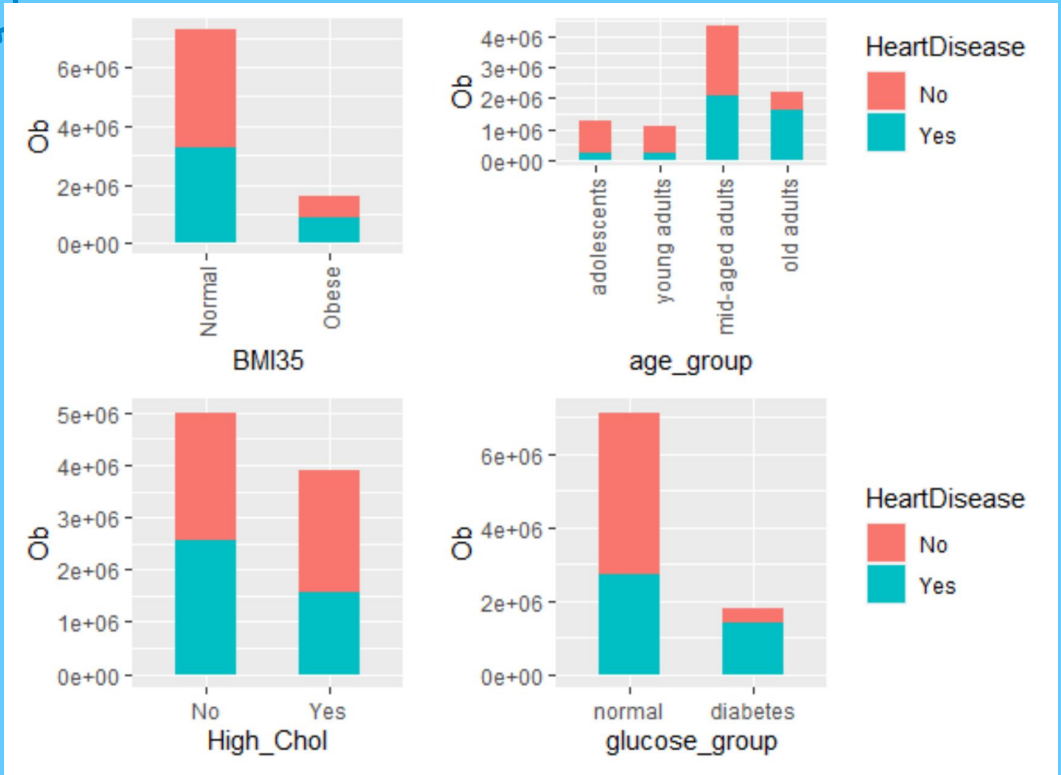
Hence, we later used PCA to create uncorrelated predictors.

Categorical Predictors



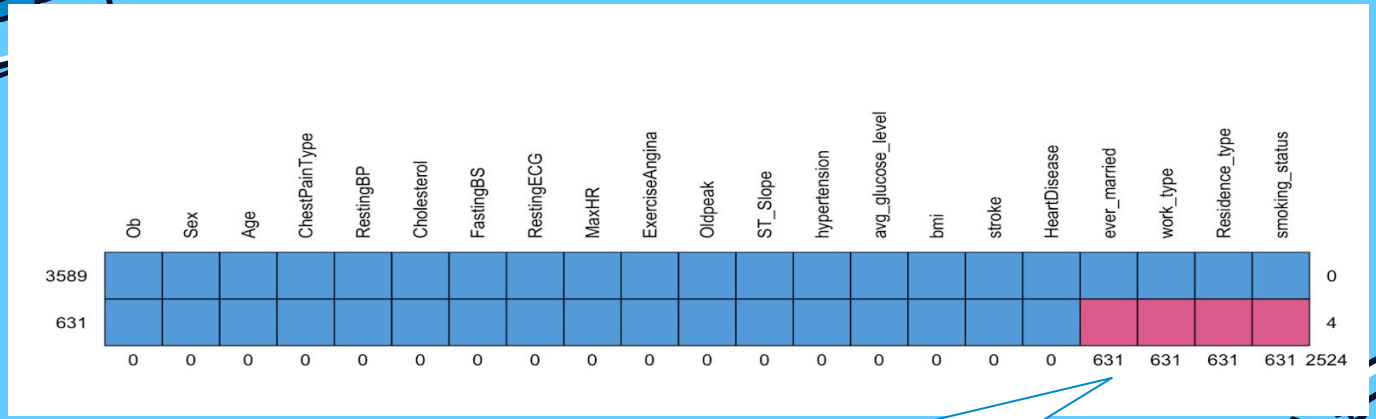
These categorical predictors showed great difference in ratios of heart disease diagnosis.

Customized Categorical Predictors



We noticed that people with higher cholesterol levels and age have a higher ratio of heart disease.

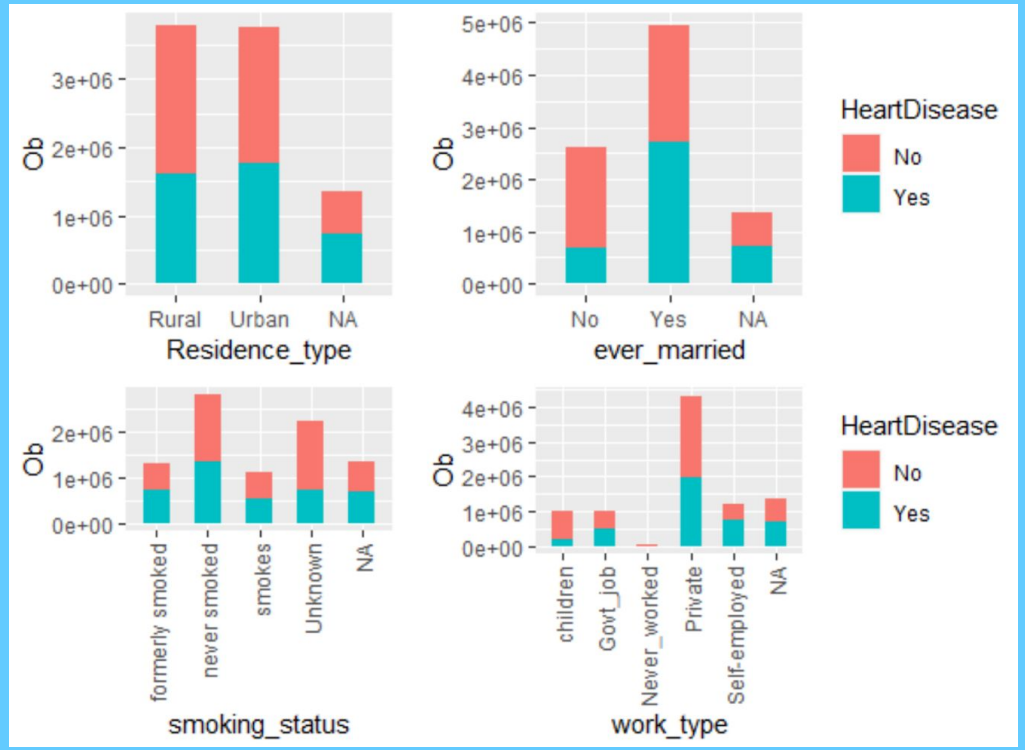
NA Imputation



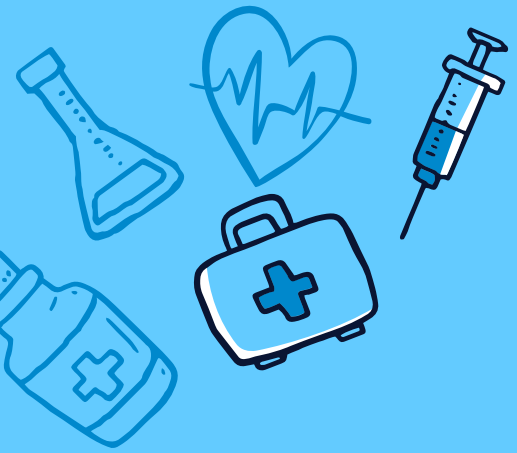
The NAs all appear in the same rows of four columns: ever_married, work_type, Residence_type, smoking_status

NA Imputation

After looking at the stack bar charts, we find the reason of unfavorable accuracy.



After trying to imputate by using Mice and Hmisc packages, the result is not ideal !



03

Feature Selection

Chi-square test

Predictors	p-val	Predictors	p-val
age_group	2.91e-139	glucose_group	3.29e-96
FastingBS	1.01e-83	hypertension	1.74e-42
stroke	2.86e-25	High_Chol	9.31e-13
ExerciseAngina	7.84e-09	Sex	4.78e-08
ST_Slope	1.13e-06	BMI35	8.65e-06
ChestPainType	4.42e-02	RestingECG	2.63e-01

By chi-square test, we found many categorical predictors and the diagnosis of heart disease are dependent, indicated by the significant p values.

t-test

Predictors	p-val
Age	0.592973
RestingBP	0.381200
Cholesterol	0.000783
MaxHR	<2e-16
Oldpeak	<2e-16
avg_glucose_level	<2e-16
bmi	0.373225

By t-test, we found that MaxHR, Oldpeak, avg_glucose_level, and cholesterol were significant predictors. The t-test was conducted using the glm function in R.

Selected predictors: Cholesterol, FastingBS, ST_Slope, stroke, ExerciseAngina, MaxHR, avg_glucose_level, Oldpeak

Backward Selection

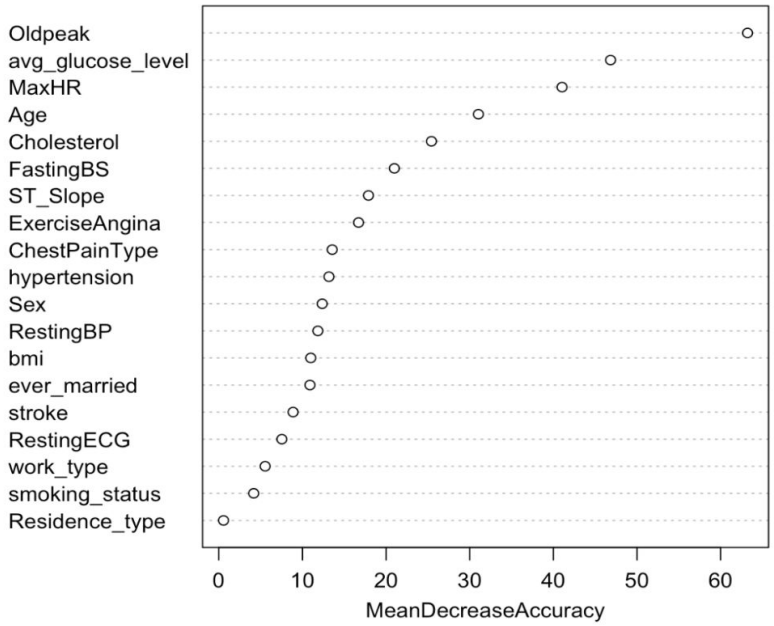
AIC

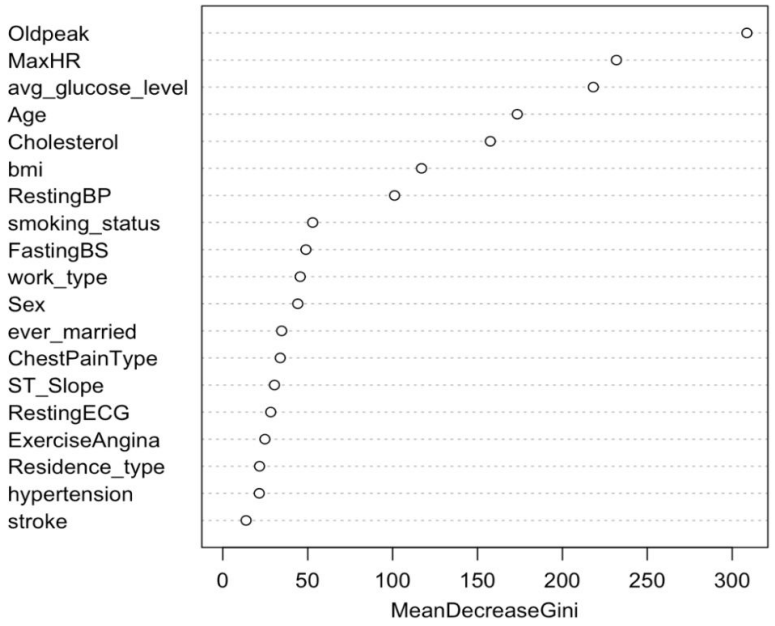
```
Step:  AIC=3618.82
as.factor(HeartDisease) ~ Sex + ChestPainType + Cholesterol +
    FastingBS + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
    avg_glucose_level + stroke

                   Df Deviance    AIC
<none>                  3584.8 3618.8
- ChestPainType     3   3592.6 3620.6
- Cholesterol       1   3596.7 3628.7
- Sex               4   3614.4 3640.4
- ST_Slope          2   3613.6 3643.6
- FastingBS         1   3624.6 3656.6
- ExerciseAngina    1   3657.3 3689.3
- stroke            1   3660.1 3692.1
- MaxHR             1   3756.8 3788.8
- avg_glucose_level 1   3823.5 3855.5
- Oldpeak           1   4238.5 4270.5
```

BIC

```
Step:  AIC=3707.46
as.factor(HeartDisease) ~ Cholesterol + FastingBS + MaxHR + ExerciseAngina +
    Oldpeak + ST_Slope + avg_glucose_level + stroke

                   Df Deviance    AIC
<none>                  3624.0 3707.5
- Cholesterol       1   3634.0 3709.1
- FastingBS         1   3667.3 3742.4
- ST_Slope          2   3690.1 3756.9
- stroke            1   3697.8 3772.9
- ExerciseAngina    1   3698.7 3773.8
- MaxHR             1   3801.9 3877.1
- avg_glucose_level 1   3862.0 3937.1
- Oldpeak           1   4271.1 4346.2
```

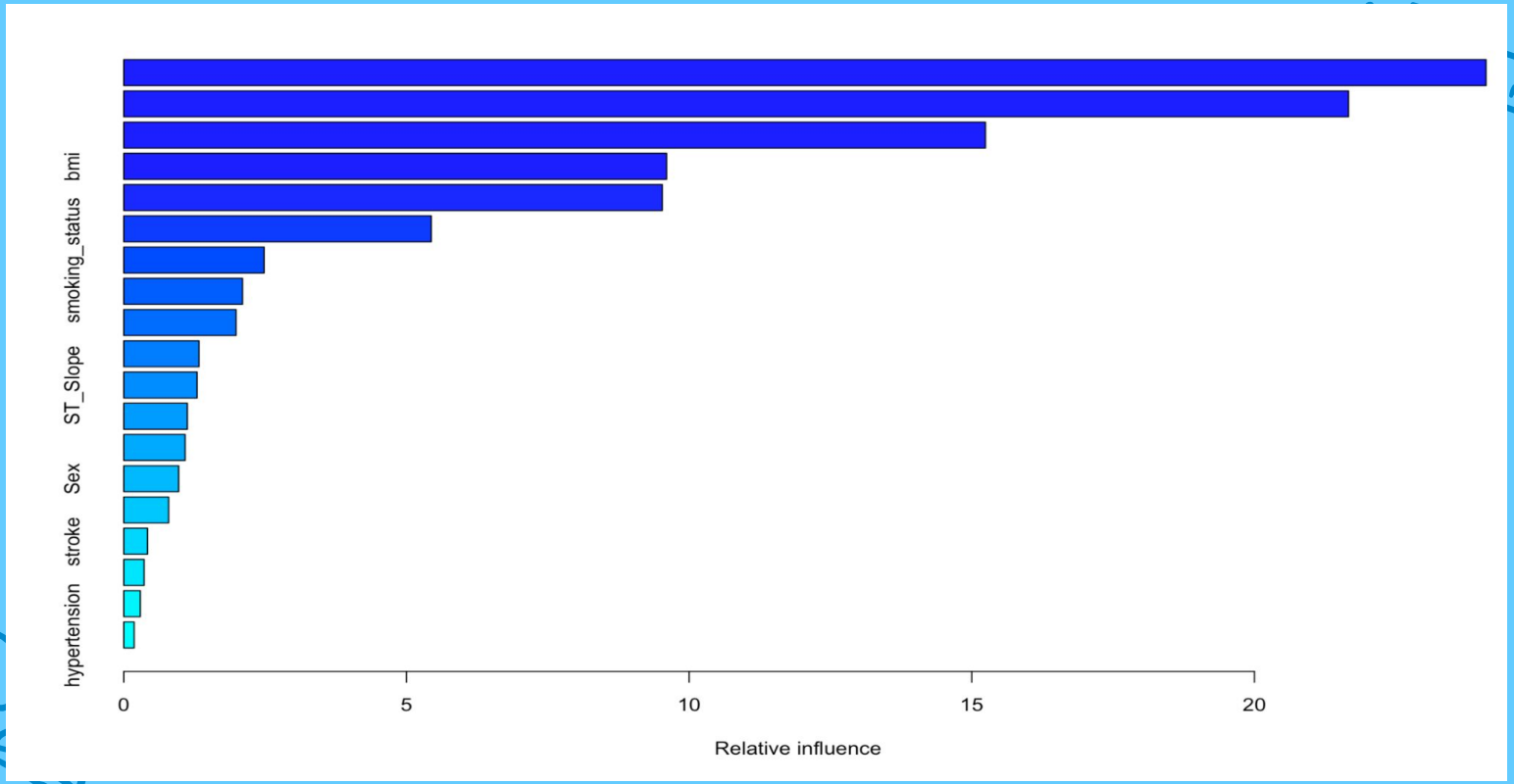
Random Forest





The top 5 predictors matches !!!

GBM


bmi
smoking_status
ST_Slope
Sex
stroke
hypertension
0
5
10
15
20
Relative influence

Important Categorical Predictors

ChestPainType
typical angina, atypical angina, non-anginal pain, asymptomatic

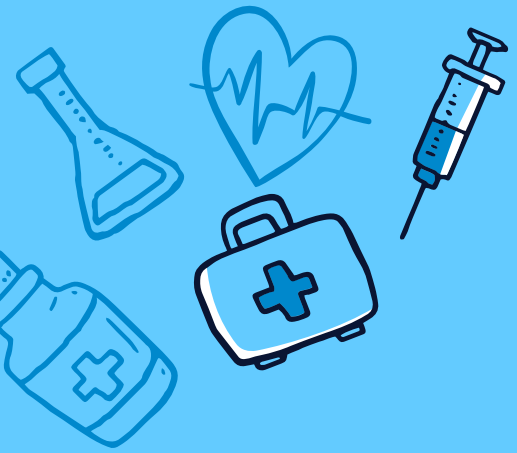
FastingBS
Fasting blood sugar more than 120 mg/dl

ST_Slope
upsloping, flat, downsloping

ExerciseAngina
exercise include angina

Stroke
Whether or not happened

Hypertension
suffering from hypertension or not



04

Modeling

Methods

PCA

Logistic Regression

LDA

Random Forest

SVM

Others

PCA

Method: We choose the four numerical predictors occurring in the backward selection, and create new independent predictors based on the linear combination of these four original predictors.
We choose to include all four principal components into our model since it will give us higher testing accuracy.

PC1+PC2+PC3 → 0.79604

PC1+PC2+PC3+PC4 → 0.81422

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.4060 1.1061 0.7636 0.46524
Proportion of Variance 0.4942 0.3059 0.1458 0.05411
Cumulative Proportion  0.4942 0.8001 0.9459 1.00000
                              PC1        PC2        PC3          PC4
Cholesterol        -0.6260296  0.2856590 -0.1765945  0.703775737
MaxHR              -0.4239441 -0.5098683  0.7483284  0.017615774
Oldpeak             0.1702114  0.7667492  0.6189535 -0.004500945
avg_glucose_level   0.6319703 -0.2655728  0.1603606  0.710189508
```

Logistic Regression

Logistic regression is very suitable for the dataset which response variables have two categories, and in our dataset, the heart disease just have two categories: Yes or No. We finally choose to use four principal components, plus some selected categorical predictors, like ChestPainType, FastingBS, ST_Slope, Excersise_Angina, and stroke.

Training accuracy = 0.8149289

Testing accuracy = 0.80632

	No	Yes
No	1912	464
Yes	317	1527

LDA

LDA is another linear classifier we applied to our data. After using PCA to resolve the collinearity among numerical predictors, its accuracy has increased greatly. Our best model includes four principal components, Sex, ChestPainType, FastingBS, ST_Slope, ExerciseAngina, stroke, hypertension and our customized categorical predictor age_group.

Training accuracy = 0.8137

Testing accuracy = 0.81422

	No	Yes
No	1948	505
Yes	281	1486

Random Forest

Random Forest is an efficient technique that can both applied to regression and classification problem. We first tune the mtry and number of trees and put all the predictors in to reduce the dimension, and finally choose four principal components, Sex, ChestPainType, FastingBS, T_Slope, ExerciseAngina, stroke, hypertension and our customized categorical predictor age_group.

Training accuracy = 0.8031

Testing accuracy = 0.7984

	No	Yes
No	1914	315
Yes	516	1475

SVM-Linear Kernel

As shown before, linear classifier generates better prediction for our data, so SVM with linear kernel is another model we've tried. After tuning parametes like gamma, cost, and degree, our best model includes four principal components, Sex, ChestPainType, FastingBS, T_Slope, ExerciseAngina, stroke, hypertension and our customized categorical predictor age_group.

Training accuracy = 0.8156

Testing accuracy = 0.81264

	No	Yes
No	1937	486
Yes	292	1505

SVM-Radial Kernel

We also tried svm model with radial kernel since the boundary may be different.
After tuning parametes like gamma(0.4) and cost(1), our best model includes four principal components, Sex, ChestPainType, FastingBS, T_Slope, ExerciseAngina, stroke, hypertension and our customized categorical predictor age_group.

Training accuracy = 0.8628

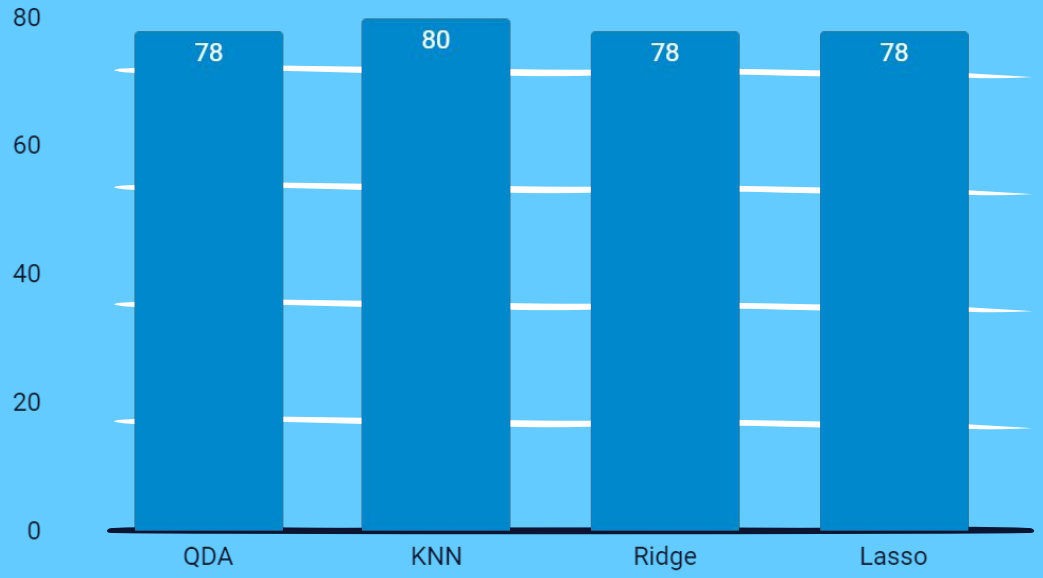
Testing accuracy = 0.805

	No	Yes
No	2070	420
Yes	159	1571

Others

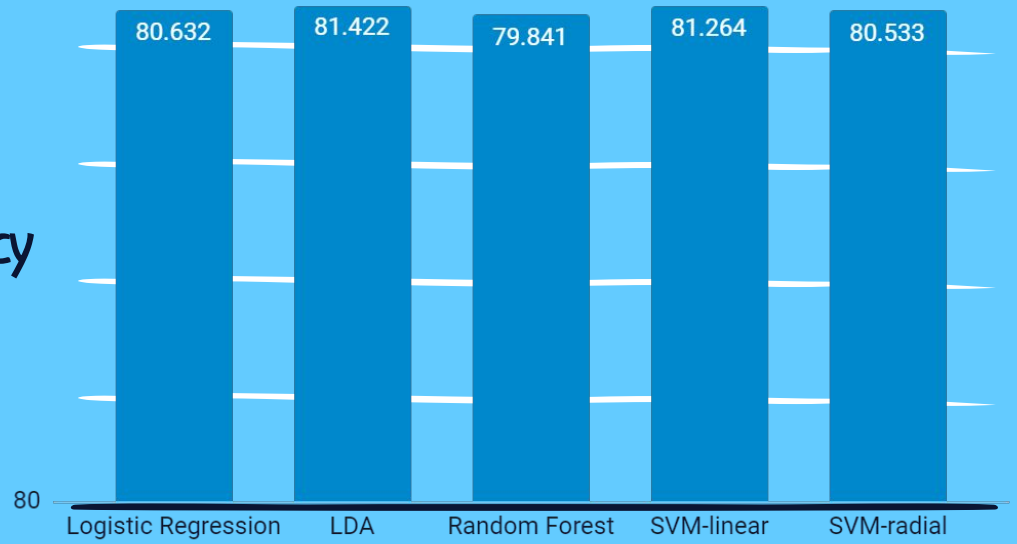
We have also attempted QDA, KNN, Lasso, Ridge, SVM-radial kernel, but they didn't generate satisfying training accuracy.

Training Accuracy



Model Evaluation

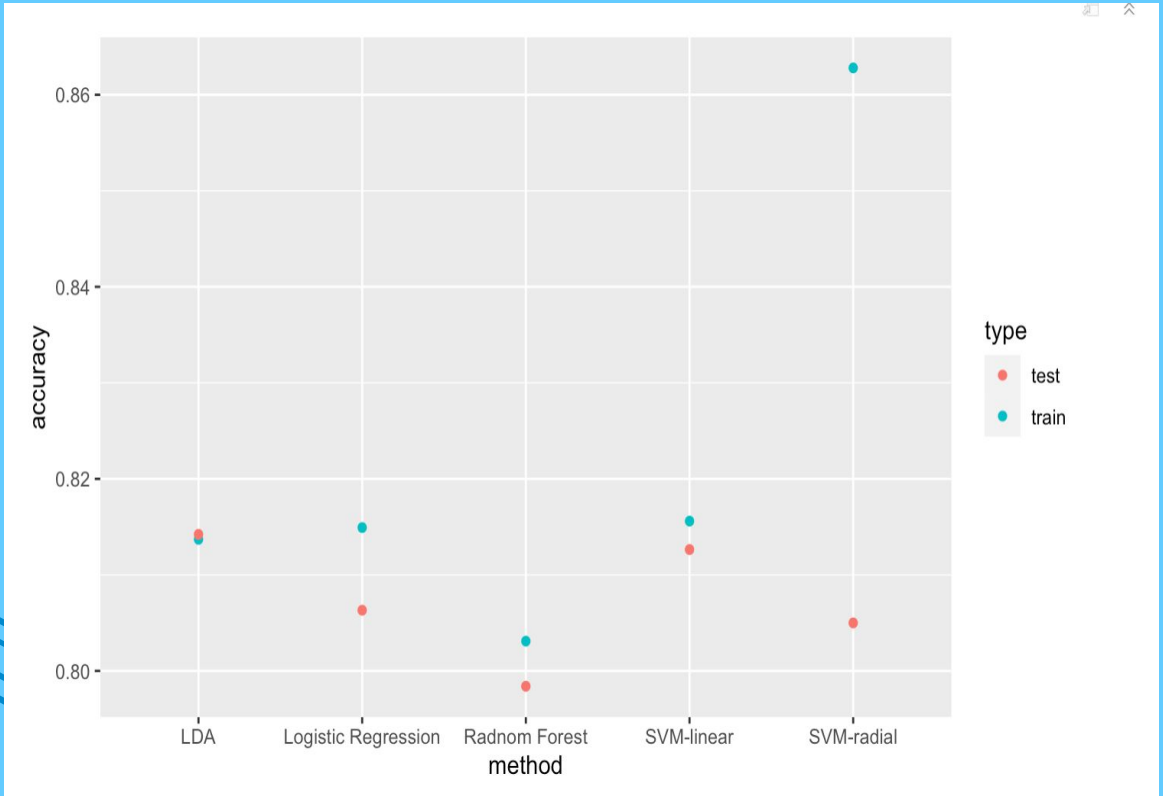
Test Accuracy



LDA

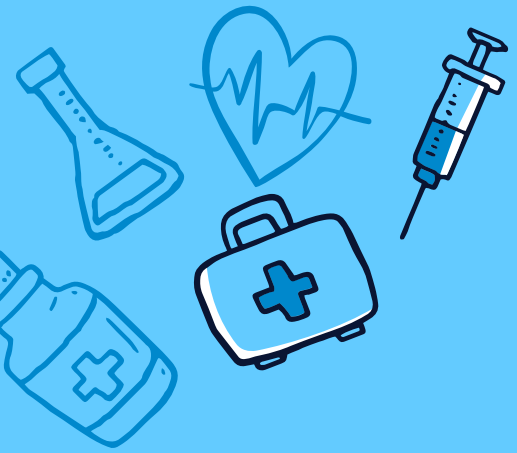
Has the best prediction accuracy

Model Evaluation



We can see that the support vector machine model with radial kernel has overfitting problems.

LDA and SVM with linear kernel performs better for our data set, which means our data set has linear decision boundary !



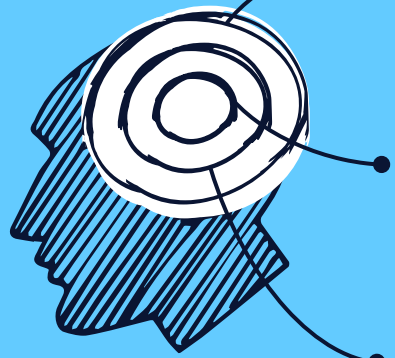
05

Discussion

Conclusion

- By developing and testing several different classification models, this study clearly established the viability of predictive diagnosis of heart disease.
- By employing several statistical techniques, we identified several significant risk factors towards the development of heart disease.

Limitations

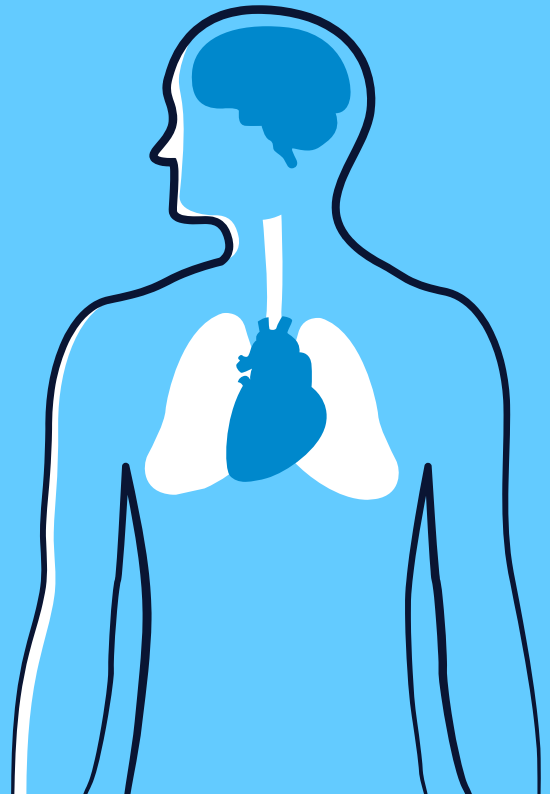


01 The size of observations in the training sample is relatively small.

02 Unable to generalize to all types of heart disease.

03 There might be more efficient ways to impute NA's.

Recommendations



A

Data Explore

More potential transformations of the variables

B

Machine Learning

Build dynamically and self-developed model

References

- Dietary Cholesterol and Cardiovascular Risk: A Science Advisory From the American Heart Association
- Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1 - BMC Public Health
- Gender differences in cardiovascular disease - ScienceDirect
- Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases
- Prediction of severity of coronary artery disease using slope of submaximal ST segment/heart rate relationship
- Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity
- Tobacco smoking and risk of 36 cardiovascular disease subtypes: fatal and non-fatal outcomes in a large prospective Australian study - BMC Medicine

Thanks