SC1015 Mini Project

Identifying Hypertension

Lab Group B133, Team 5
Yong Shao En Ernest (U2221153B), Wu Rixin (U2221172G) & Li Liyi (U2220985F)

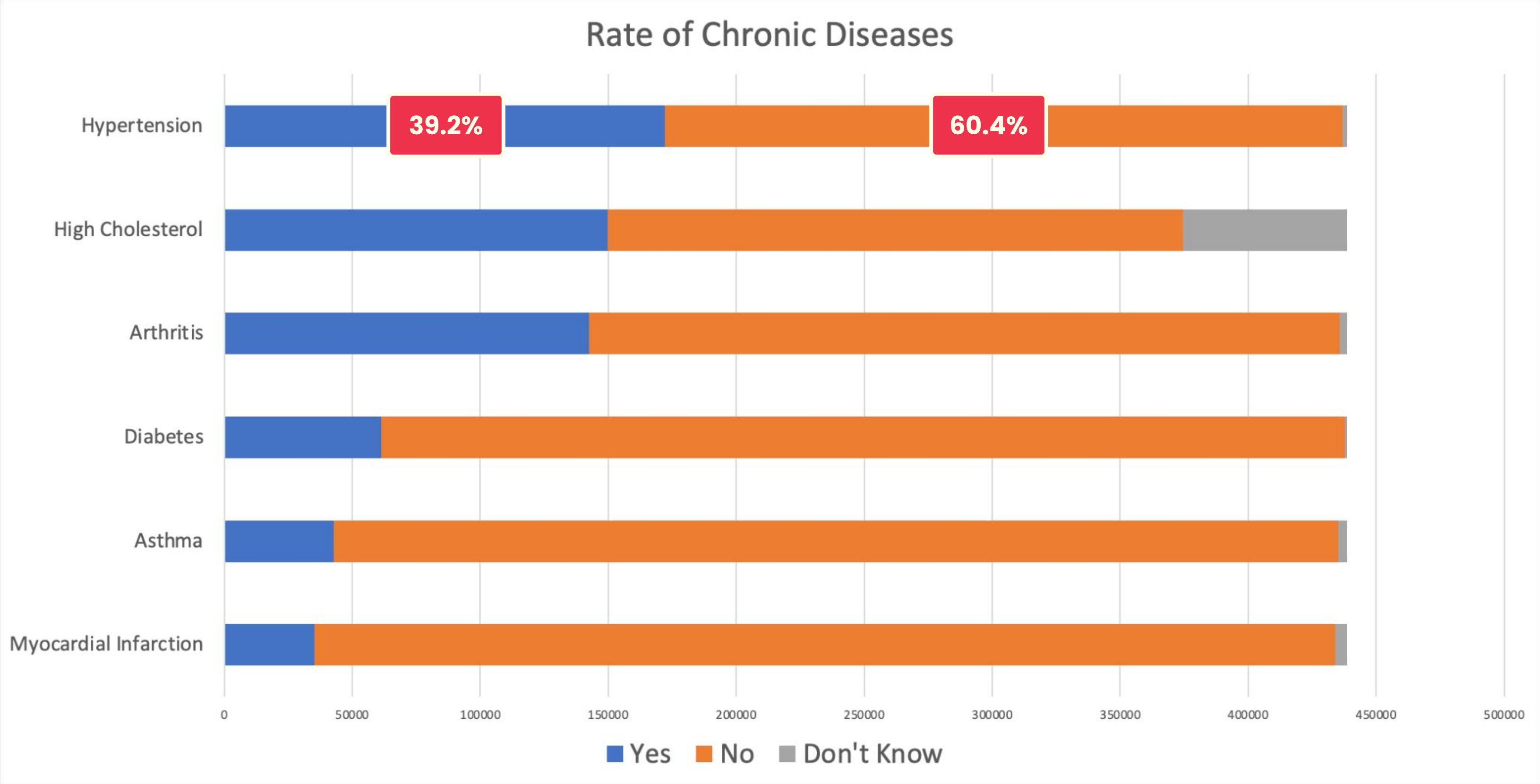
Dataset Used

We selected data from the 2021 Behavioural Risk Factor Surveillance System Survey Data and Documentation conducted by US Centers for Disease Control and Prevention(CDC).

CDC Centers for Disease Control and Prevention

CDC 24/7: Saving Lives, Protecting People™

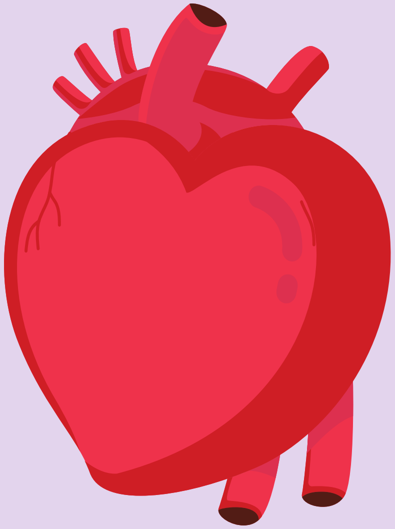
Dataset Used



Our Motivation

35.5% of Singaporeans have hypertension in 2020

No. 1 risk factor of death globally



Problem Definition

What are the variables correlated with hypertension, and how can we identify undiagnosed individuals suffering from hypertension?

Data Extraction

Factors	
BMI	Alcohol
Physical Exercise	Smoker
Diabetes	Mental Health
High Cholesterol	Physical Health
Junk Food Intake	Race
Fruit Intake	Vegetables Intake
Education Level	

Relevant Variables	
_TOTINDA	FTJUDA2_
_BMI5	VEGEDA2_
DROCDY3_	MENTHLTH
AVEDRNK3	PHYSHLTH
_RFBING5	_RFHYPE6
CHOLMED3	_EDUCAG
FRENCHF1	DIABETE4
FRUTDA2_	_RFCHOL3

Data Cleaning

01 Tackle missing and irrelevant values

	_TOTINDA	_BMI5	DROCDY3_	AVEDRNK3
0	2.0	1454.0	0.0	NaN
1	1.0	NaN	0.0	NaN
2	2.0	2829.0	0.0	NaN
3	1.0	3347.0	14.0	3.0
4	1.0	2873.0	0.0	NaN

Data Cleaning

02 Create new variables by combining existing ones

DROCDY3_
Drink occasions per day

AVEDRNK3
Number of drinks consumed

AlchoIntake
Weekly alcohol consumption

Data Cleaning

03 Standardise units of measurement & adjust the decimal places for numeric variables

Value	Value Label
101 - 199	Days
201 - 299	Weeks
300	Less than once a month
301 - 399	Month / Year
555	Never
777	Don’t know/Not sure
999	Refused
BLANK	Not asked or Missing

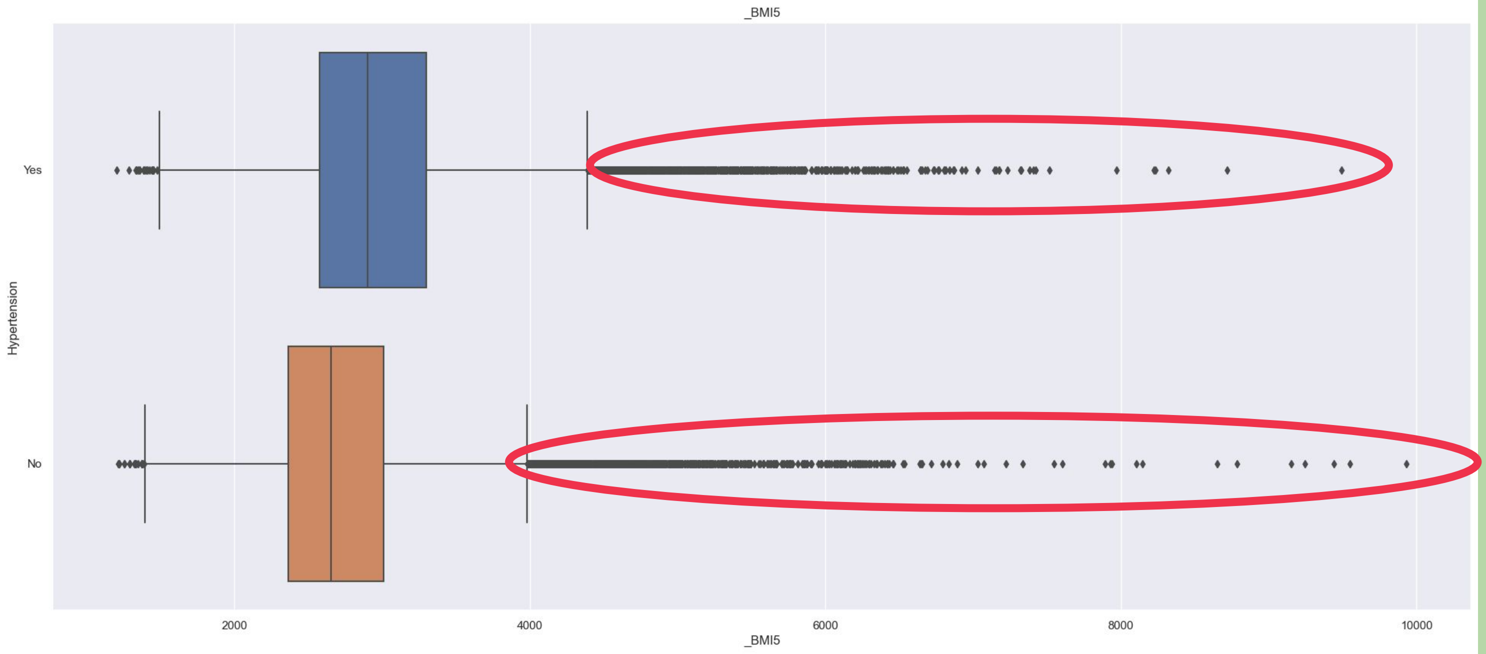
Data Cleaning

04 Decode categorical variables based on data description

Question: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	330,738	75.39	75.96
2	No	107,027	24.40	23.87
9	Don't know/Refused/Missing Notes: EXERANY2 = 7 or 9 or Missing	928	0.21	0.17

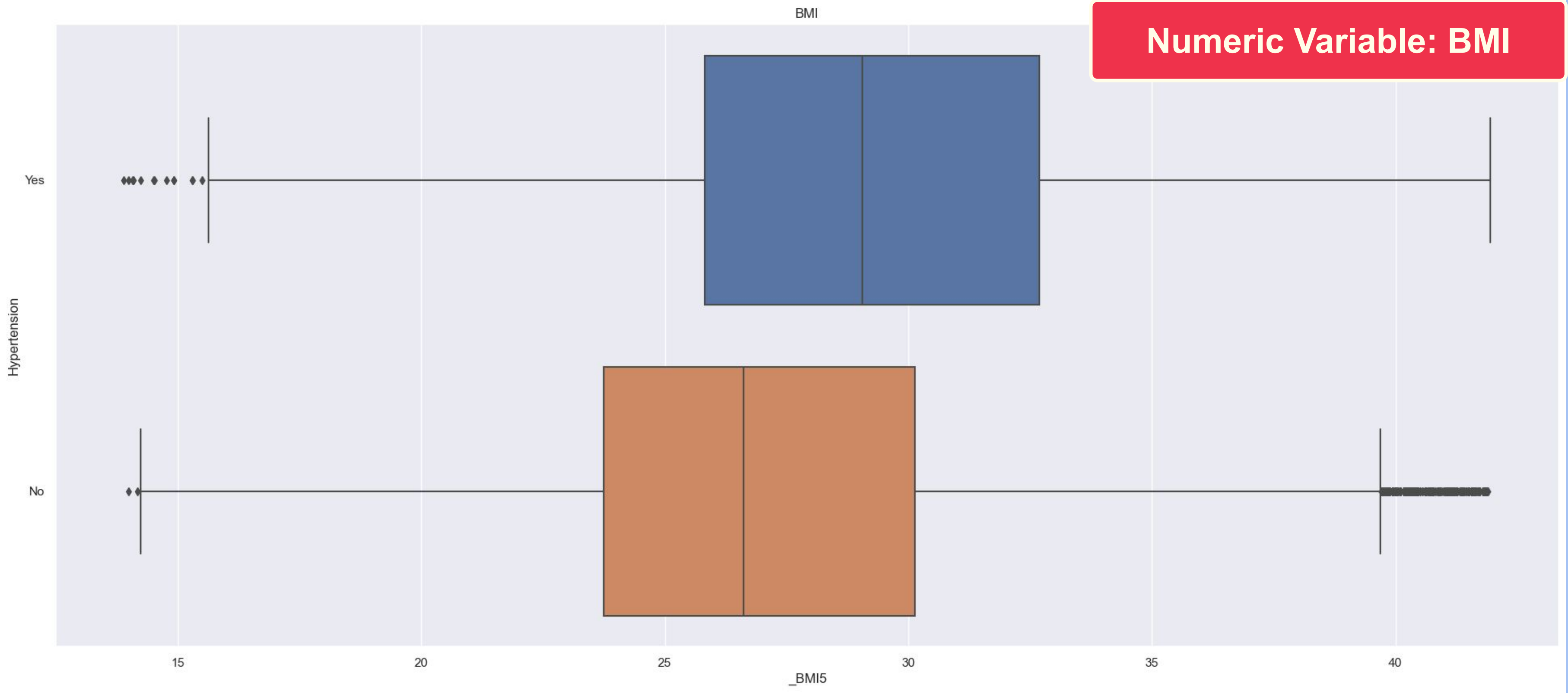
Data Cleaning

05 Identify and remove outliers for numeric variables

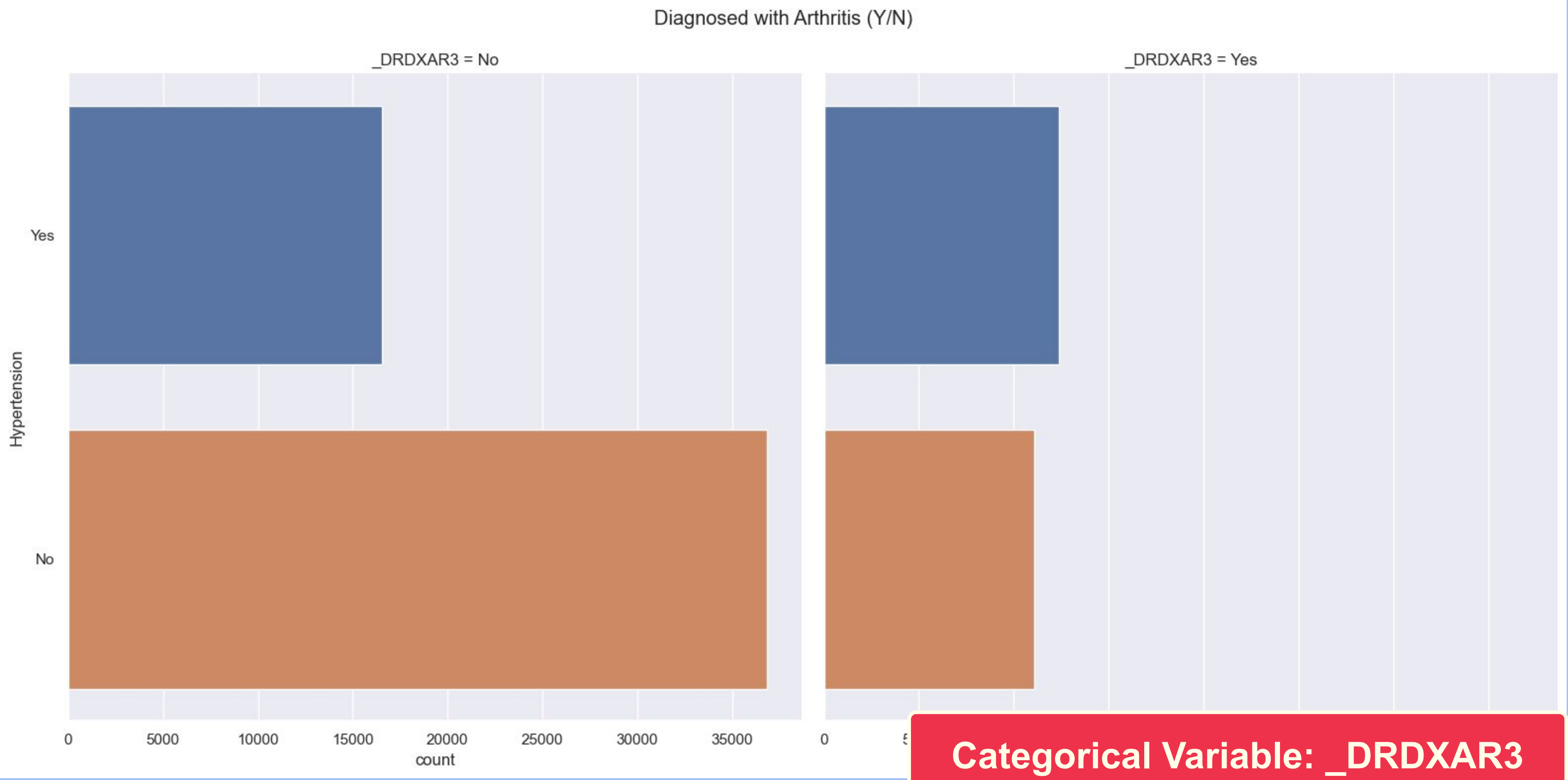


Exploratory Data Analysis

Numeric Variable: BMI



Exploratory Data Analysis



Categorical Variable: _DRDXAR3

Exploratory Data Analysis

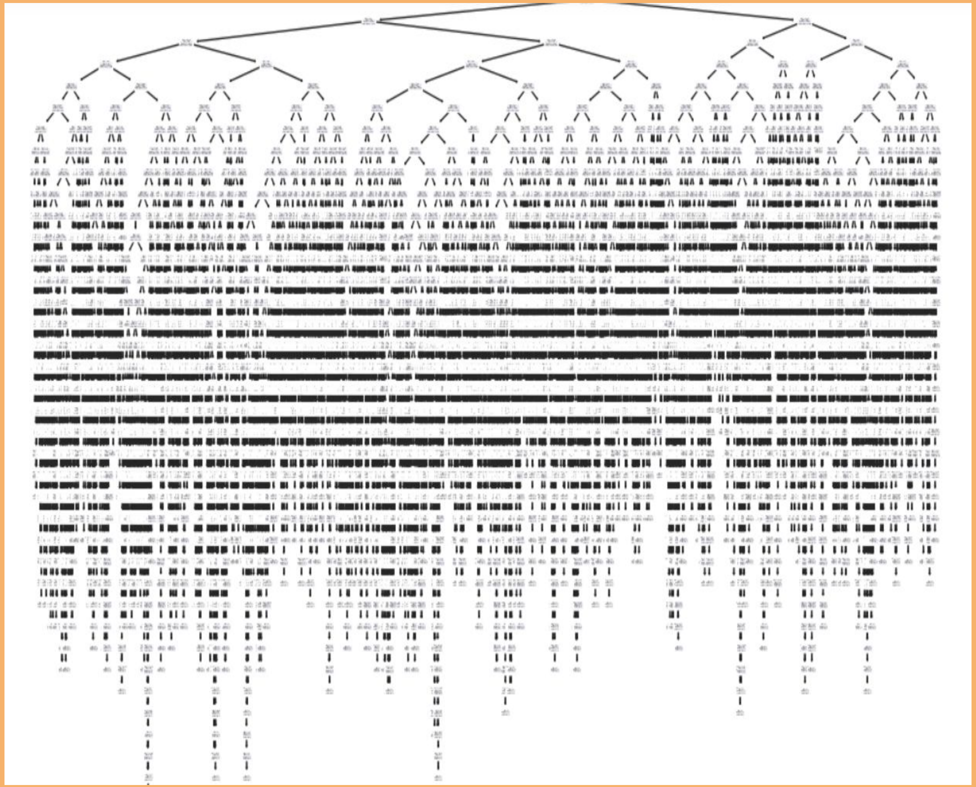
Response Variables
Hypertension

Numeric Variables
_BMI5
AlchoIntake
PHYSHLTH
_AGE80

Categorical Variables
_TOTINDA
CHOLMED3
DIABETE4
_RFCHOL3
_MICHD
_EDUCAG
_DRDXAR3

Model 1: Decision Tree

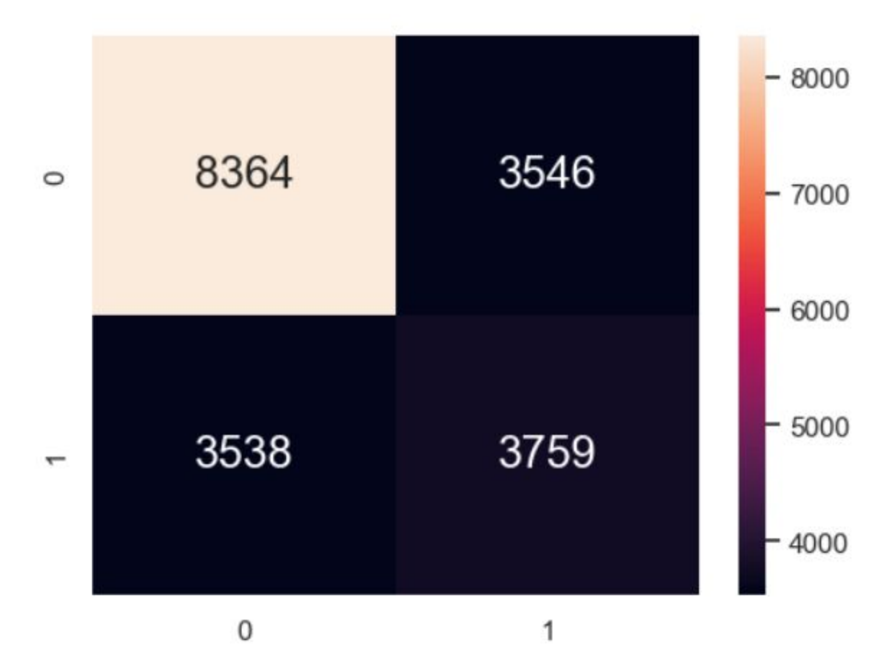
Initial Decision Tree



- Decision tree constructs a **model of decisions** and their **possible consequences**
- The tree can accurately **predict** the class or value of new, unseen instances
- It can handle **both numerical and categorical** variables
- Decision trees may suffer from overfitting when the depth level or the stopping criterion is not well-defined

Model 1: Decision Tree

Confusion Matrix of
Initial Model

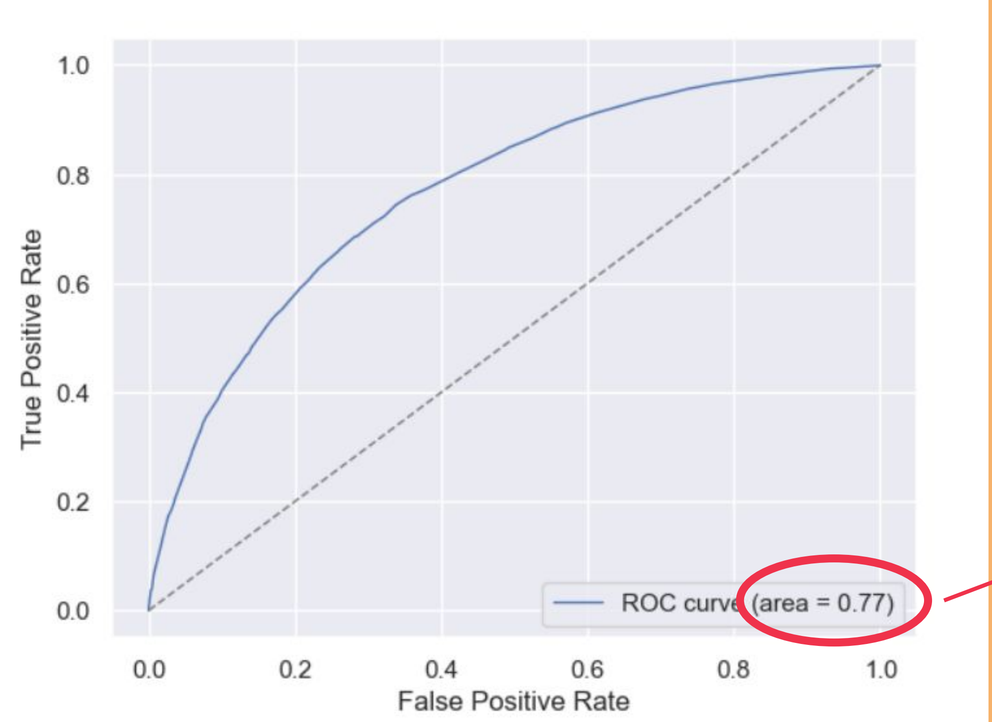


TPR = 0.51514
FPR = 0.29773

Prediction
Accuracy
= 0.60871

Model 1: Decision Tree

ROC Curve

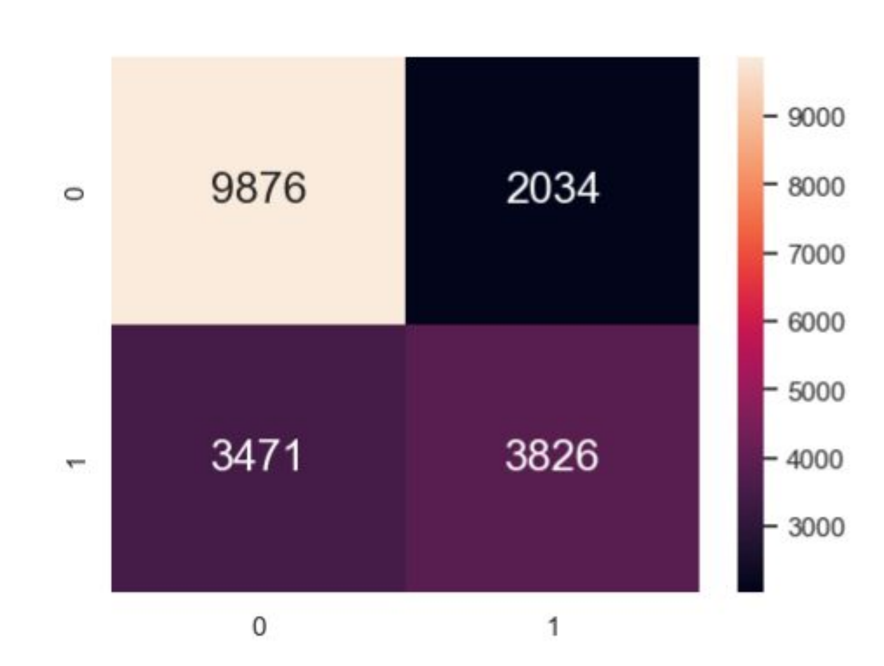


Max_depth = 5

Largest Area

Model 1: Decision Tree

Confusion Matrix after Optimisation

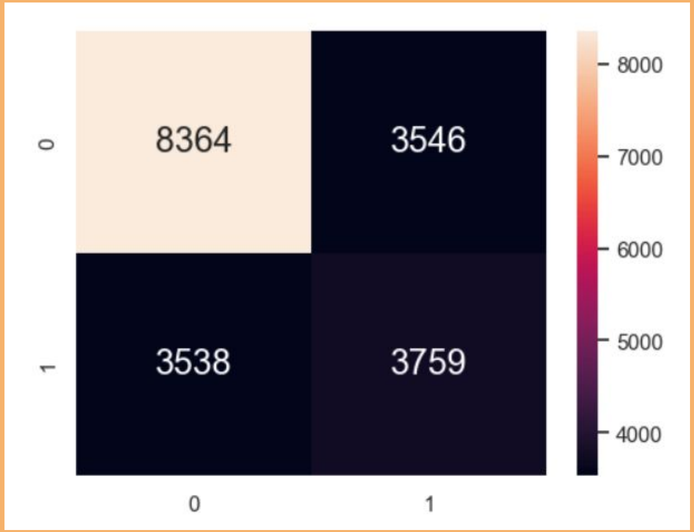


TPR = 0.52433
FPR = 0.17078

Prediction Accuracy = 0.67677

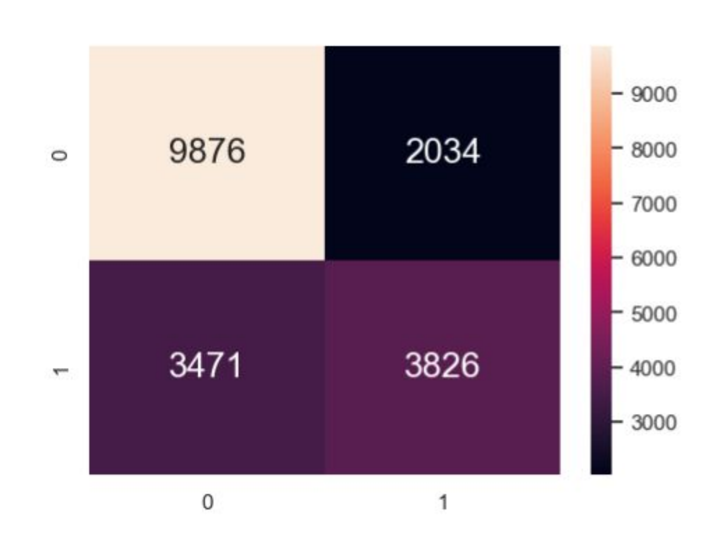
Model 1: Decision Tree

Confusion Matrix of Initial Model



Prediction Accuracy = 0.60871

Confusion Matrix after Optimisation

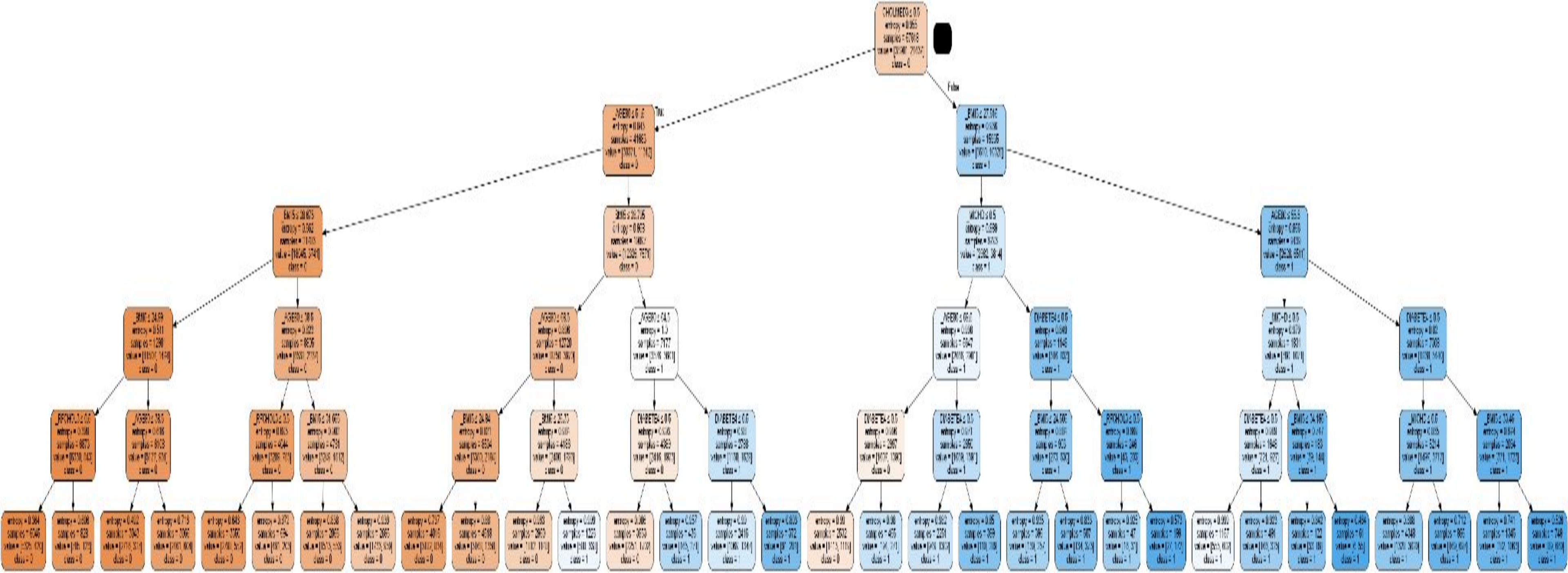


Prediction Accuracy = 0.67677

Improvement: 6.88%

Model 1: Decision Tree

Optimised Decision Tree

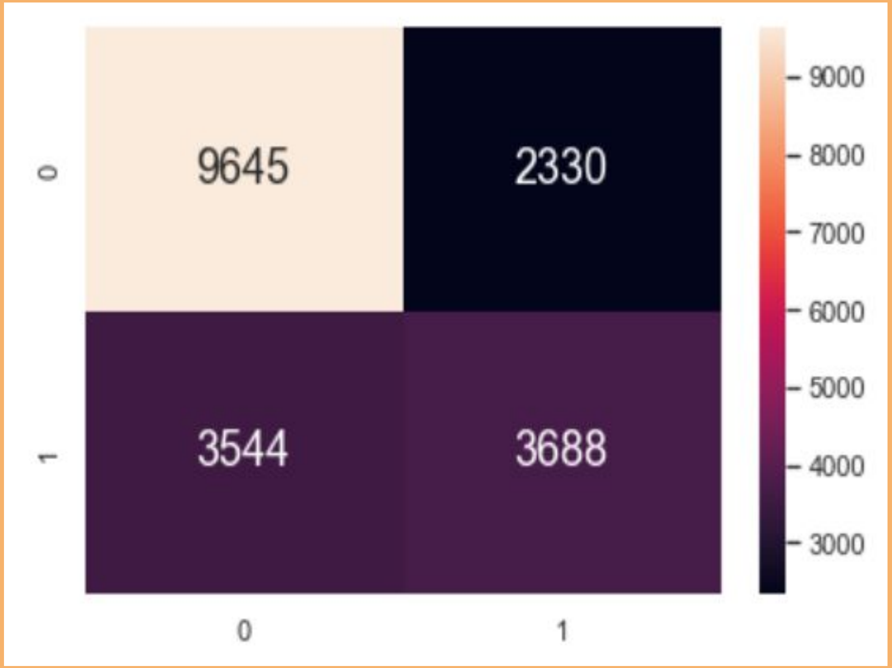


Model 2: Random Forest

- Random Forest takes random data points from random variables to come up with **multiple decision trees**.
- Multiple decision trees allows the **strengths and weaknesses** of each tree to be balanced out by the other trees.
- The output of each tree is then combined to make a final prediction with a **greater accuracy** than a single decision tree
- Suitable for **large** datasets with a **mix** of categorical and numerical data

Model 2: Random Forest

Confusion Matrix of
Initial Model

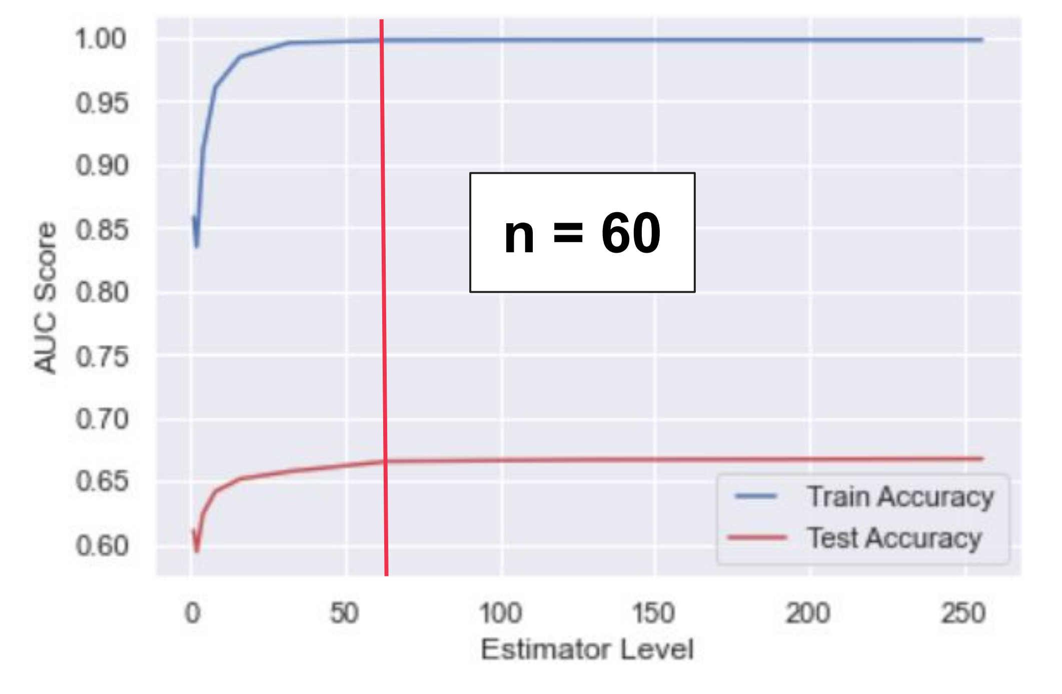


TPR = 0.50996
FPR = 0.19457

Prediction
Accuracy
= 0.69417

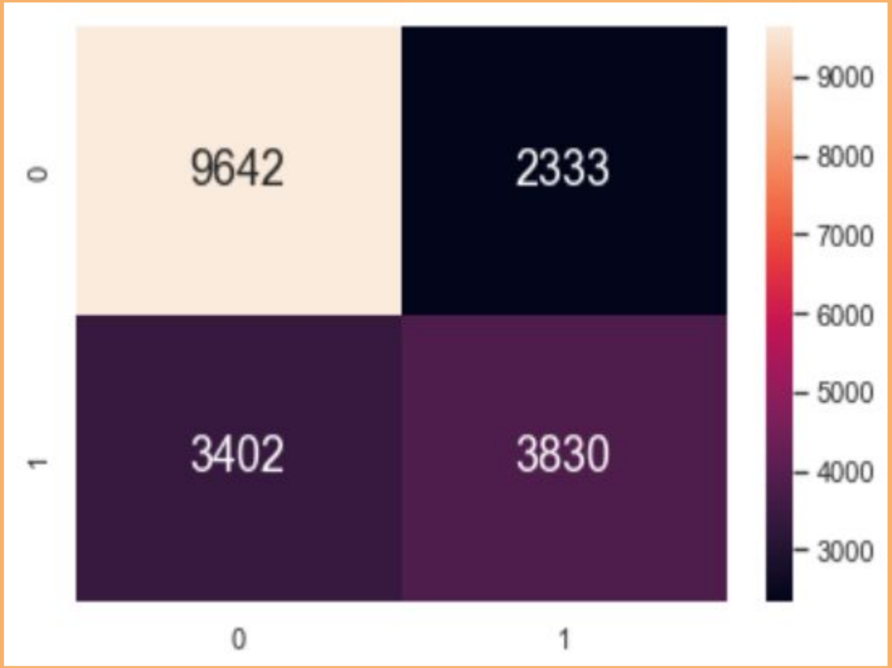
Model 2: Random Forest

ROC Curve & AUC Score



Model 2: Random Forest

Confusion Matrix after Optimisation

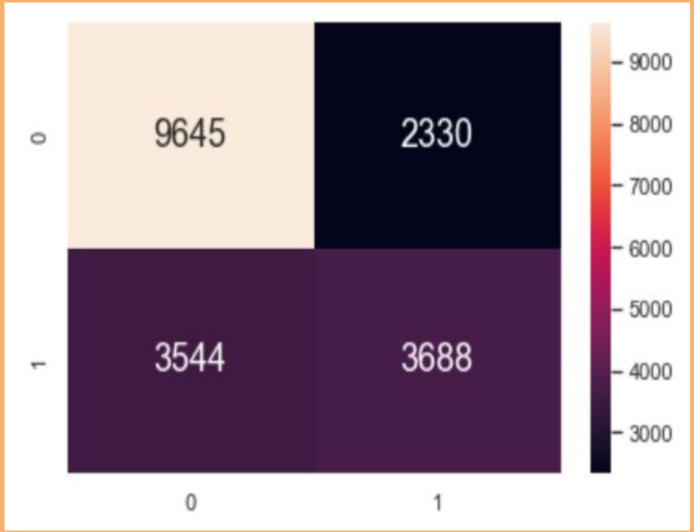


TPR = 0.52959
FPR = 0.19482

Prediction Accuracy = 0.70141

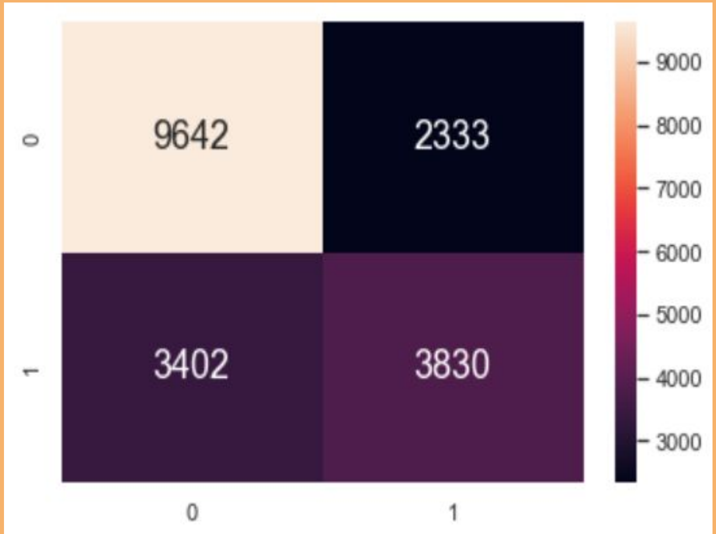
Model 2: Random Forest

Confusion Matrix of Initial Model



Prediction Accuracy = 0.69417

Confusion Matrix after Optimisation



Prediction Accuracy = 0.70141

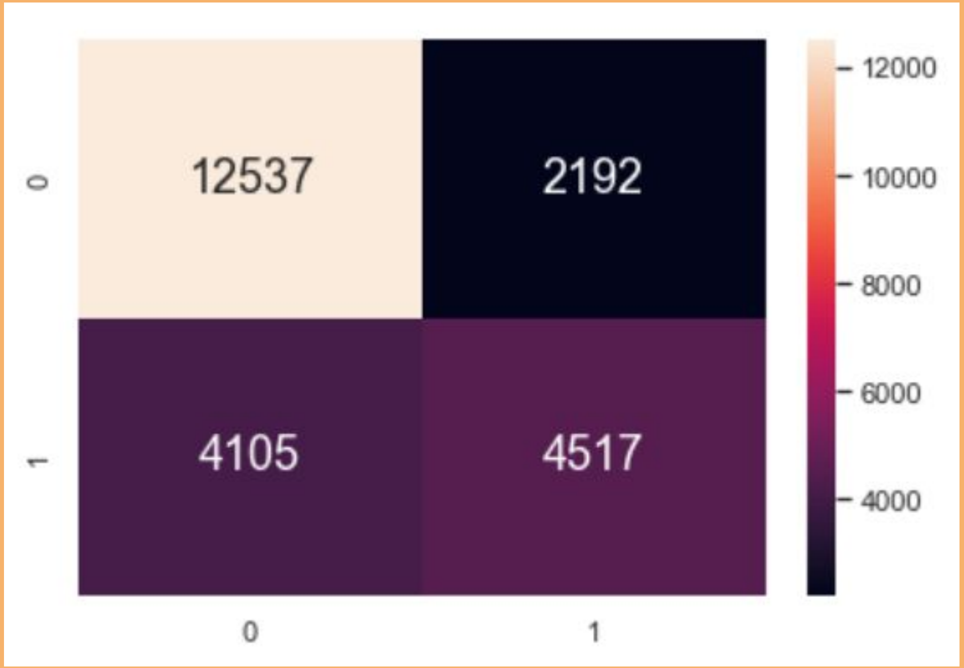
Improvement: 0.72%

Model 3: Logistic Regression

- Logistic Regression **predicts** the output of a categorical variable based on one or more independent variables.
- It **reveals** the interrelationships between different variables and their impact on outcomes
- This helps us make **accurate predictions**

Model 3: Logistic Regression

Confusion Matrix of
Initial Model



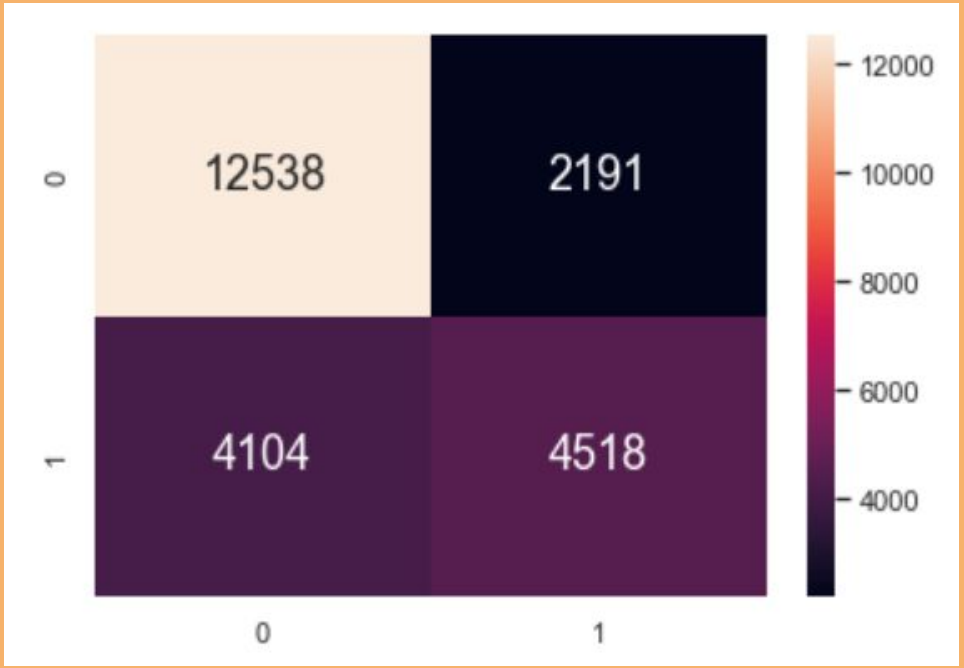
TPR = 0.52389
FPR = 0.14882

Prediction
Accuracy
= 0.73033

Model 3: Logistic Regression

Confusion Matrix after
Hyperparameter Optimisation

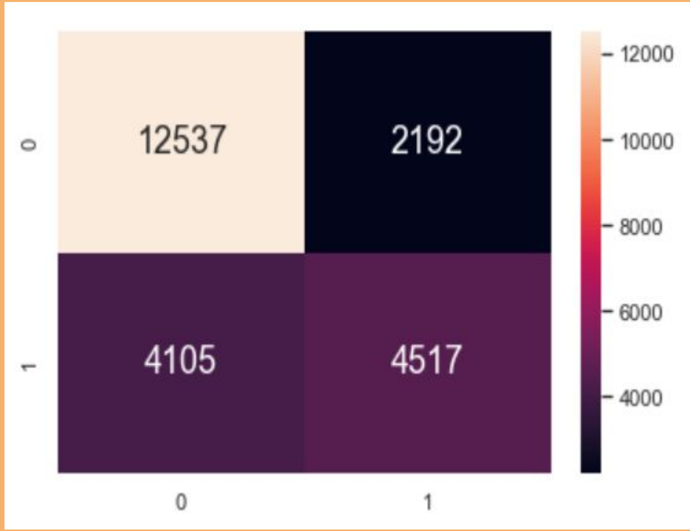
TPR = 0.52401
FPR = 0.14875



Prediction
Accuracy
= 0.73042

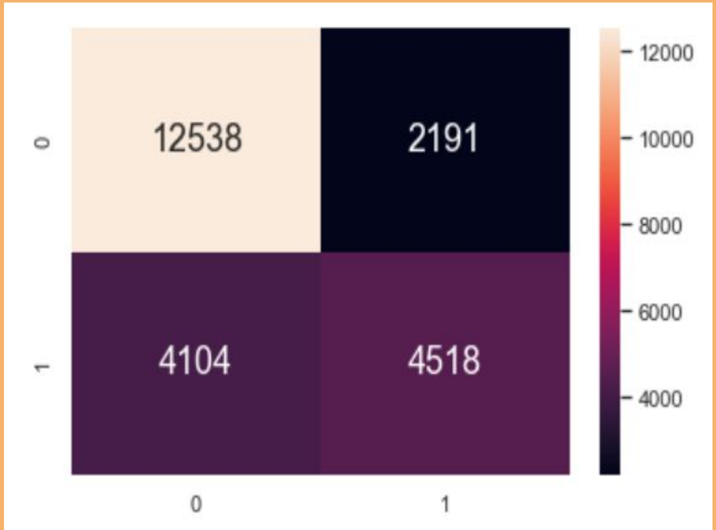
Model 3: Logistic Regression

Confusion Matrix of
Initial Model



Prediction Accuracy = 0.73033

Confusion Matrix after
Optimisation



Prediction Accuracy = 0.73042

Best Model: Logistic Regression

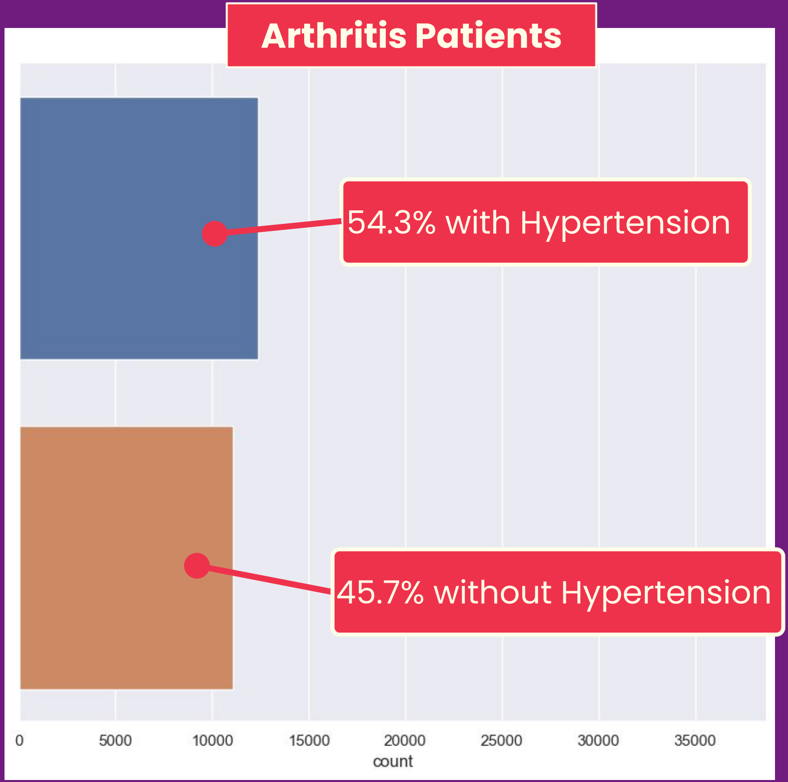
	Decision Tree	Random Forest	Logistic Regression
Accuracy	0.67677	0.70141	0.73042
TPR	0.52433	0.52959	0.52401
FPR	0.17078	0.19482	0.14875

Findings & Data-Driven Insights

Insight 1

Affirm

Arthritis is highly correlated with hypertension, but no explanation can be found as of now

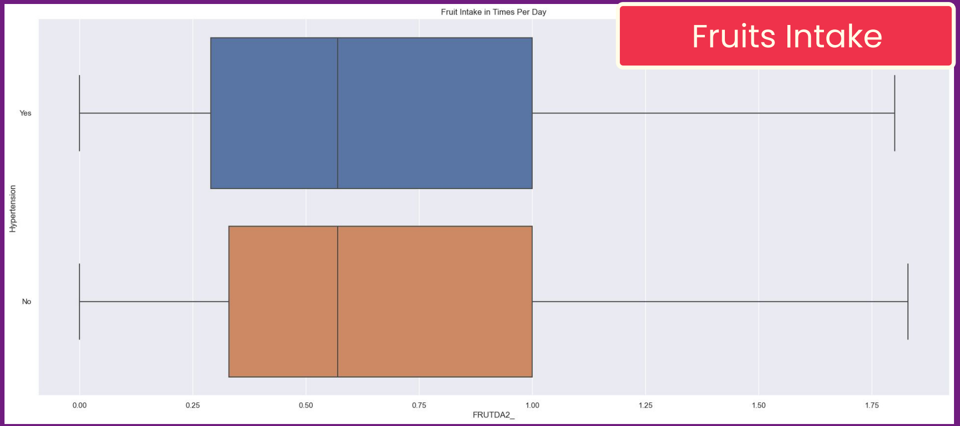


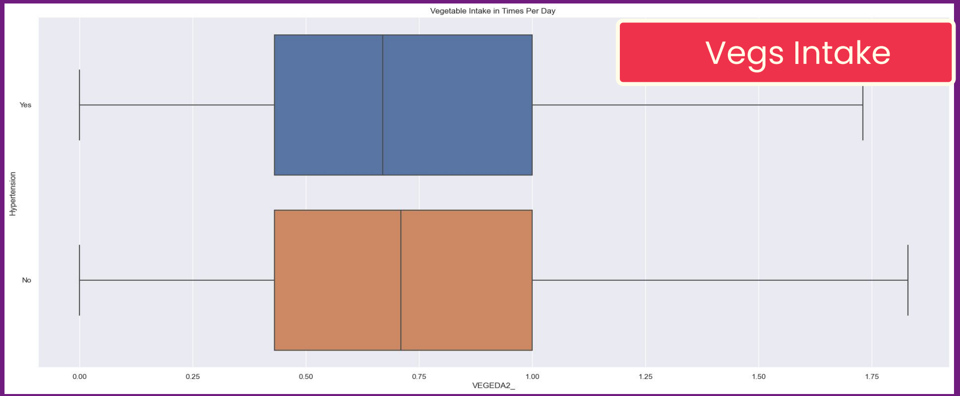
Findings & Data-Driven Insights

Insight 2

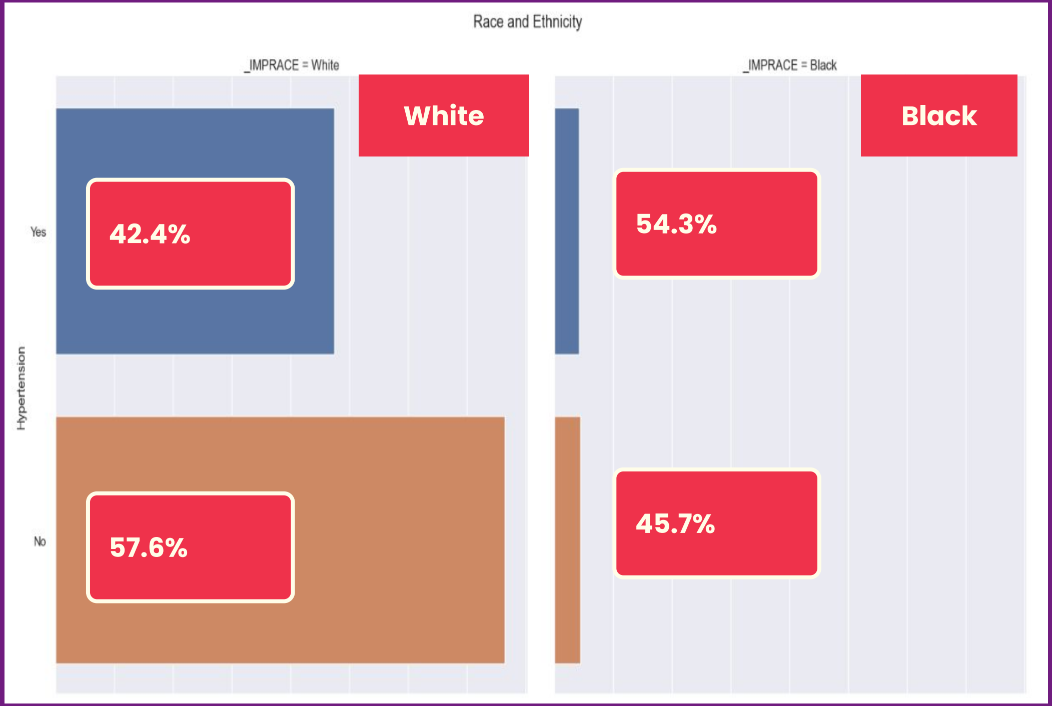
Debunk

People with and without hypertension have the same average intake of fruits and vegetables





Findings & Data-Driven Insights

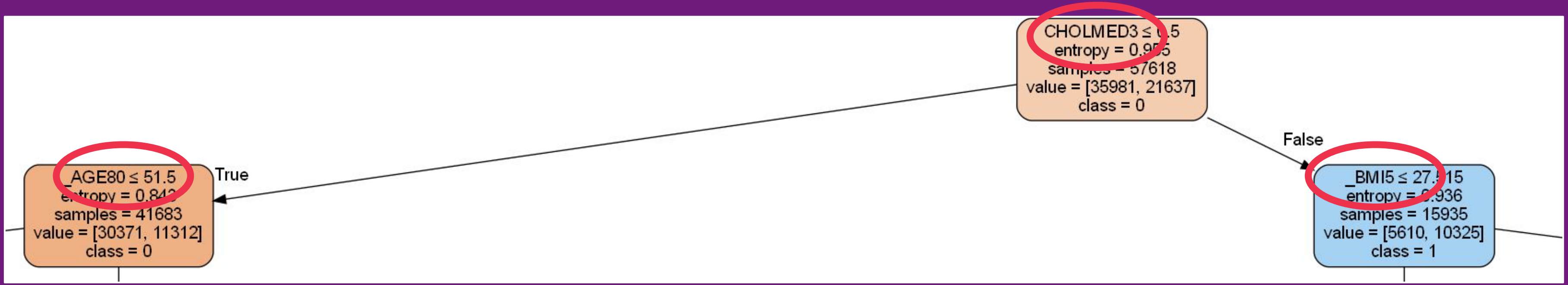


Insight 3

Discover

For all the races included in the survey except Black, the proportion of people with hypertension is smaller.

Findings & Data-Driven Insights



Insight 4

Observe

Most frequently used factors for classification models:

1. Cholesterol
2. BMI
3. Age

Future Recommendations

1. **Cholesterol Levels**

   Focus on reducing consumption of alcohol and food high in saturated fats, thus reducing cholesterol levels

2. **Body-Mass-Index**

   Encourage healthy eating habits and physical exercise

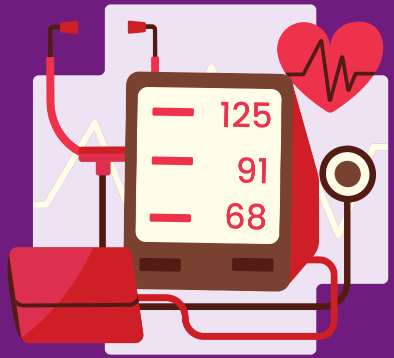
3. **Age**

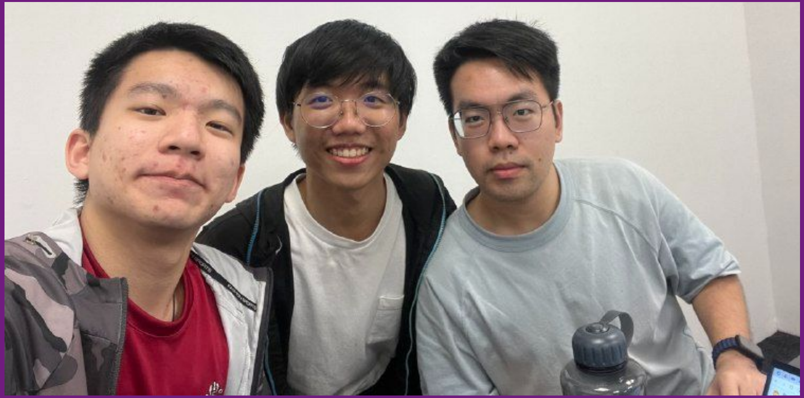
   Health campaigns can be targeted more towards the elderly

Limitations & Recommendations

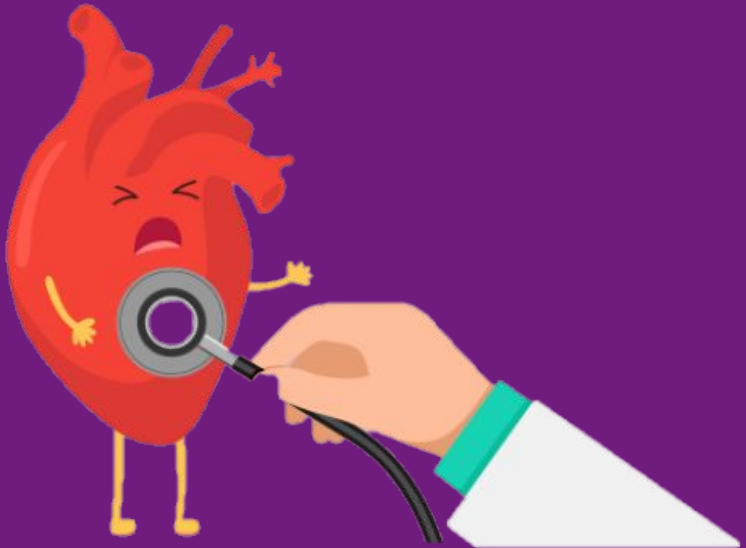
- **Data is collected from the US:**
  Demographic and lifestyle factors may differ from Singapore

- **Genetic factors:**
  A more in-depth survey can be conducted to identify factors contributing to hypertension in Singapore





Thank You!



Done by:

Yong Shao En Ernest (U2221153B)

Wu Rixin (U2221172G)

Li Liyi (U2220985F)