SC1015 MINI PROJECT

Stroke Prediction

Suki Ng (U2210602K) | Tay Shu Shuang (matric) | Lim Jia Jie, Isaac (U2222066A)

TABLE OF CONTENTS

01.

INTRODUCTION

Motivations & Dataset Used

OUR MOTIVATION

According to the World Health Organization (WHO) stroke is the **2nd leading cause of death globally**, responsible for approximately **11% of total deaths**. If stroke is detected or diagnosed early, the loss of death and severe damage to brain can be **prevented in 85% cases**

PROBLEM DEFINITION

We wish to find out which factors are the most important in predicting the occurrence of stroke, and how we can prevent the aggravation of such factors.

OUR DATA SET USED

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

VARIABLES PROVIDED

1) ID
2) Gender
3) Age
4) Hypertension
5) Heart Disease
6) Ever_married
7) Work_type
8) Residence_type
9) Avg_glucose_level
10) bmi
11) smoking_status
12) stroke

02.

EXPLORATORY DATA ANALYSIS

Initial data driven insights

FEATURES OF DATASET

SIZE OF DATASET:

5110

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

FEATURES OF DATASET

STATISTICAL ANALYSIS

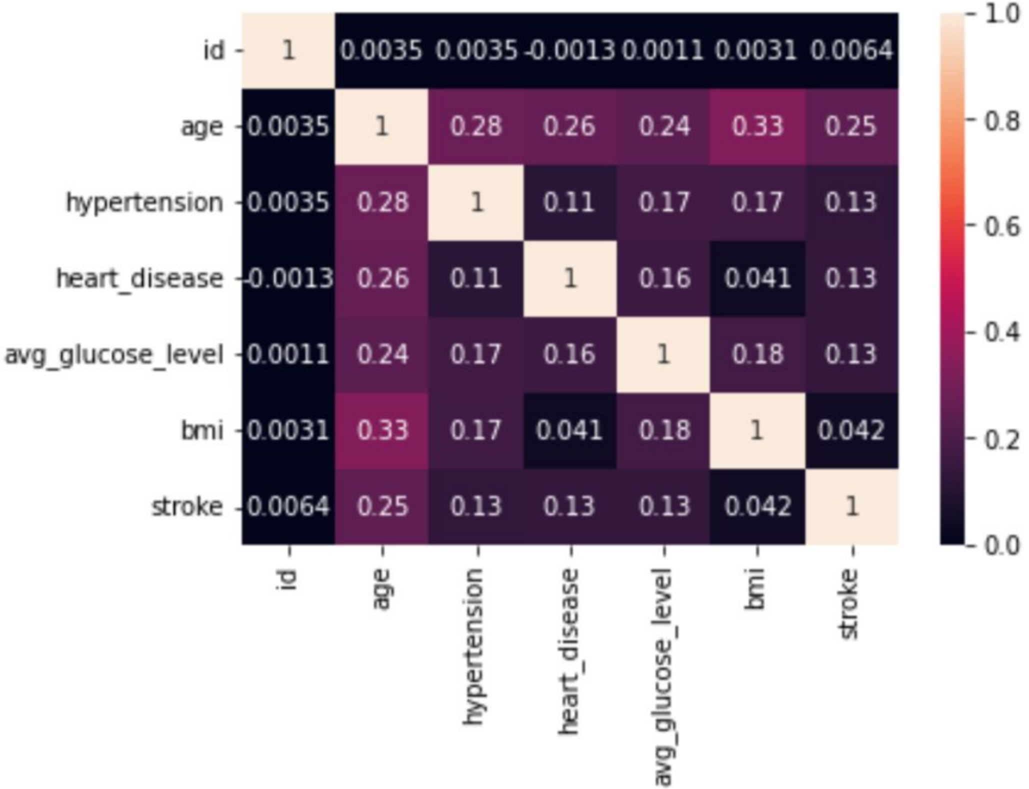
We explored the data using statistical exploration tools

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

FEATURES OF DATASET

CORRELATION MATRIX (BEFORE DATA CLEANING)



DATA CLEANING

DATA CLEANING SUGGESTIONS

```
ta cleaning improvement suggestions. Complete them before proceeding to ML modeling.
```

	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
id	5110	int64	0	0.000000	100.000000	0	possible ID column: drop
g_glucose_level	3979	float64	0	0.000000	77.866928	0	skewed: cap or drop outliers
bmi	418	float64	201	3.933464	8.180039	0	fill missing, skewed: cap or drop outliers
age	104	float64	0	0.000000	2.035225	0	
work_type	5	object	0	0.000000	0.097847	22	
smoking_status	4	object	0	0.000000	0.078278	789	
gender	3	object	0	0.000000	0.058708	1	
hypertension	2	int64	0	0.000000	0.039139	0	
heart_disease	2	int64	0	0.000000	0.039139	0	
ever_married	2	object	0	0.000000	0.039139	1757	
Residence_type	2	object	0	0.000000	0.039139	2514	
stroke	2	int64	0	0.000000	0.039139	0	

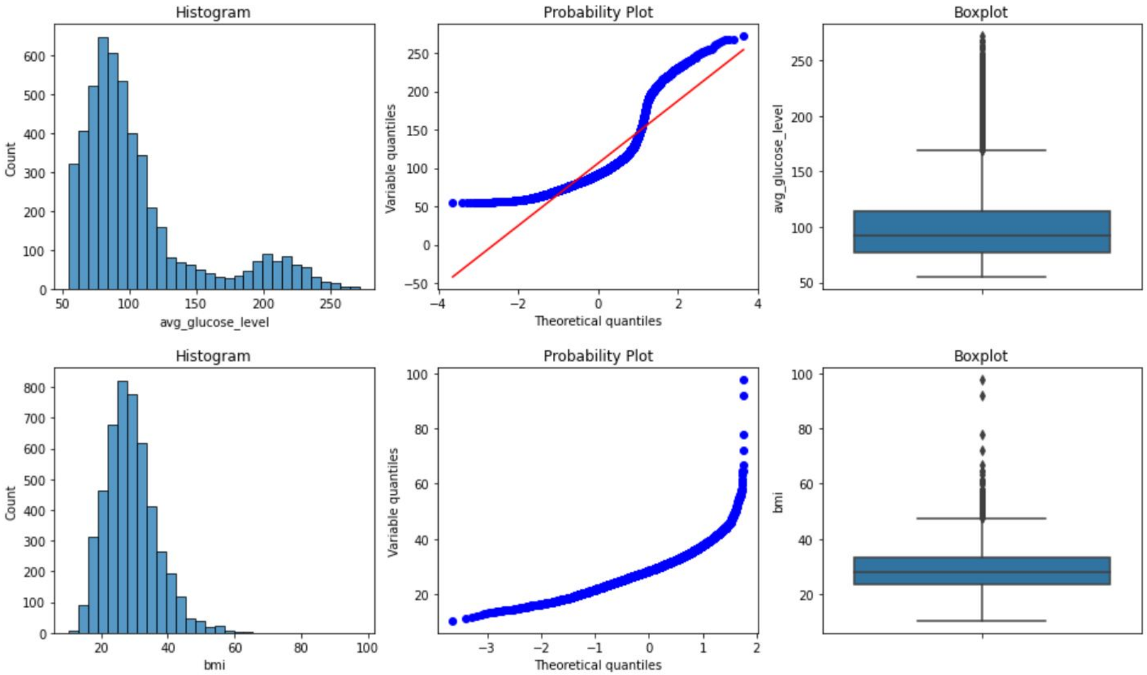
DATA CLEANING

1) DROPPED THE 'ID' COLUMNS

```
# drop ID column since it does not aid in the prediction of stroke
strokeData = strokeData.drop('id', axis = 1)
```

DATA CLEANING

2) DROP THE OUTLIERS FROM BMI AND AVG_GLUCOSE_LEVEL



DATA CLEANING

3) FOR THE BMI VARIABLE, WE REPLACED THE "NAN" AND "UNKNOWN" VALUES WITH ITS MEDIAN VALUE

```
#remove NaN in bmi and replace with median so data is still usable
bmiMedian = strokeData['bmi'].median()
strokeData['bmi'] = strokeData['bmi'].replace("unknown", "Nan")
strokeData['bmi'] = strokeData['bmi'].fillna(bmiMedian)
```

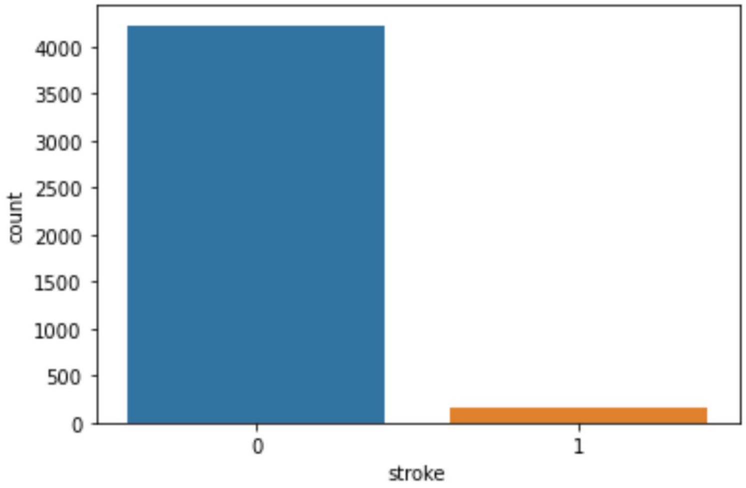
AND THEN DROPPED THE OUTLIERS

DATA CLEANING (WHAT ELSE WE TRIED)

4) SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

We have imbalanced data, with more than 4000 out of 5110 people without stroke and about 800 people with stroke.

We will be using SMOTE to oversample the data of people with stroke so that our dataset will be more balanced. This technique will be gone through more in depth later on in the presentation



AFTER DATA CLEANING

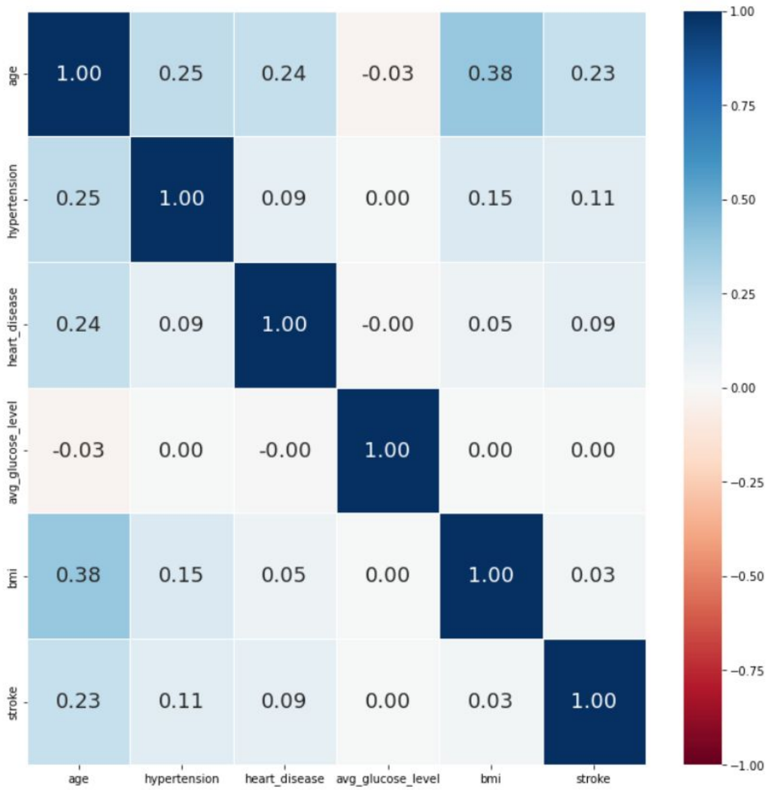
SIZE OF DATA:

BEFORE

5110

AFTER

4390



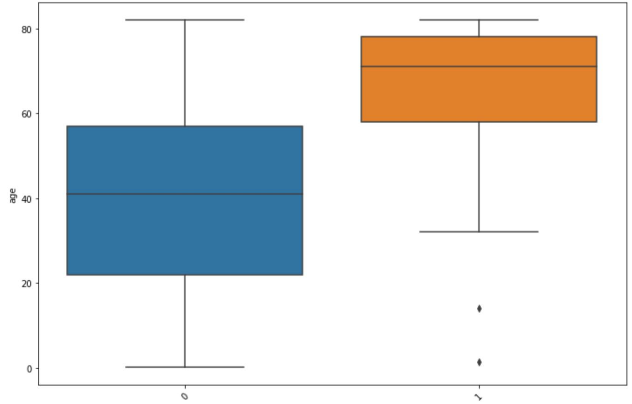
SETTING THE STAGE

HOW WE PLAN TO SET UP THE ML ANALYSIS

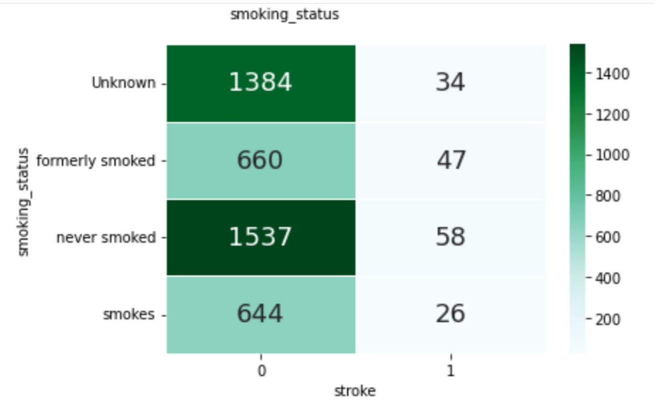
- We will start off by plotting the general distribution of variables
- We will further filter out and analyse the variables by making use of Naive Bayes and Chi Square test
- We will train and evaluate random forest classification and decision tree to predict stroke
- Analyse the coefficients and importances of various features in the different models

SETTING THE STAGE

- We will start off by plotting the general distribution of variables



Boxplot: Age against Stroke



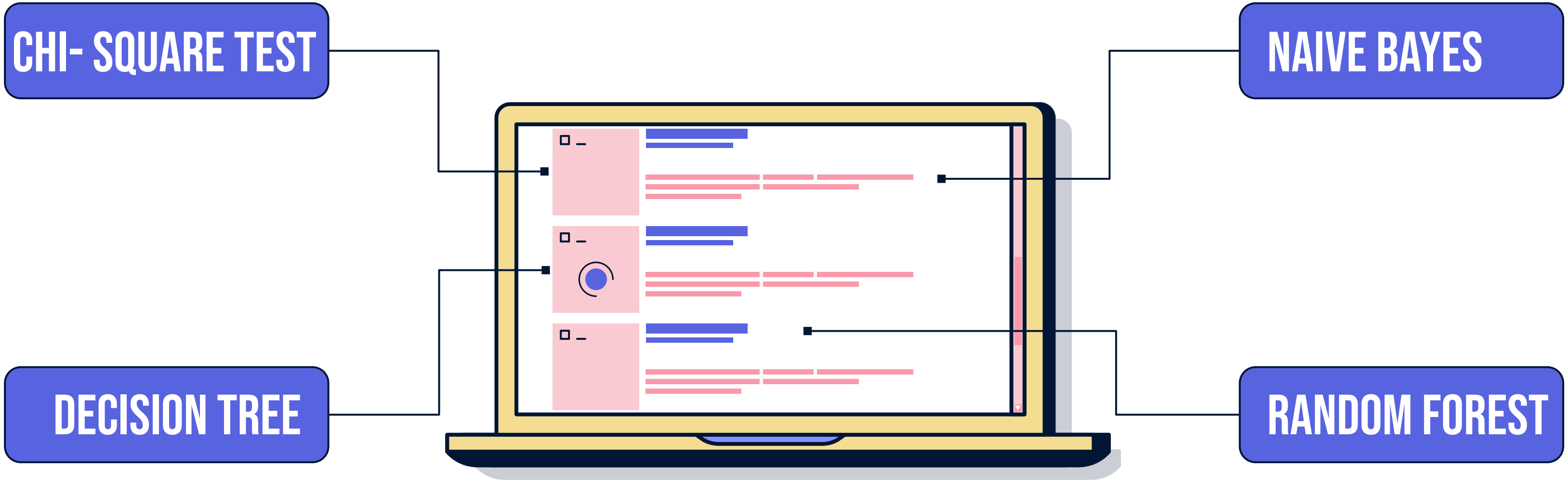
Heatmap: smoking status against stroke

03.

CORE ANALYSIS

Techniques and tools used for analysis

MACHINE LEARNING TECHNIQUES USED



CHI-SQUARE TEST

PURPOSE

tells you how likely the data you have observed occurred under the null hypothesis

P-VALUE <0.05

the data is likely to have statistical significance, and the null hypothesis is false.

P-VALUE>0.05

the data is likely to not have statistical significance, and the null hypothesis is true.

Work Type and Residence Type, BMI and Stroke, Average Glucose Level and Stroke are independent of each other

Hypertension and Heart Disease, Smoking Status and Work Type, Smoking Status and Stroke, Age and Stroke, Heart Disease and Stroke, Hypertension and Stroke are dependent of each other

NAIVE BAYES

DESCRIPTION

it is a classification technique based on Bayes' Theorem with an independence assumption among predictors, which means that this technique isolates each variable and tests it against the outcome.

Variable	Accuracy
Age	0.958997722095672
BMI	0.958997722095672
Average Glucose Level	0.958997722095672
Hypertension	0.9123006833712984
Heart Disease	0.9328018223234624

CHI-SQUARE TEST VS NAIVE BAYES

NAIVE BAYES HAVE A HIGHER ACCURACY

All 5 variables have a high accuracy of more than 90% and can help us predict stroke

OUR DATASET IS IMBALANCED

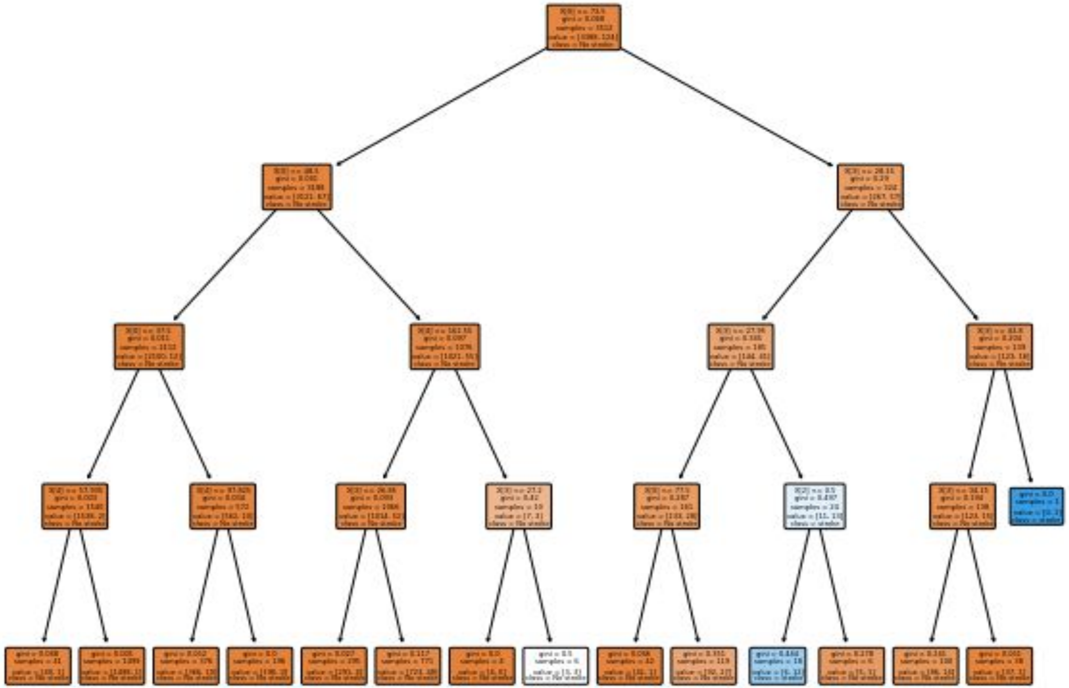
Chi-square test require a large and balanced dataset. However, our dataset is imbalanced

HENCE, WE DECIDED TO USE NAIVE BAYES TO DETERMINE THE VARIABLES TO BE USED TO PREDICT STROKE

DECISION TREE

DESCRIPTION

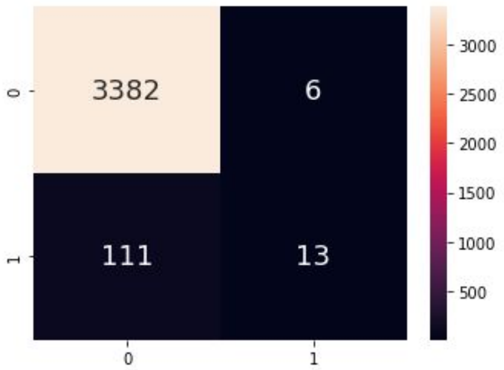
The decision tree will try to form a condition on the features to separate all the classes that are in the dataset to the fullest purity.

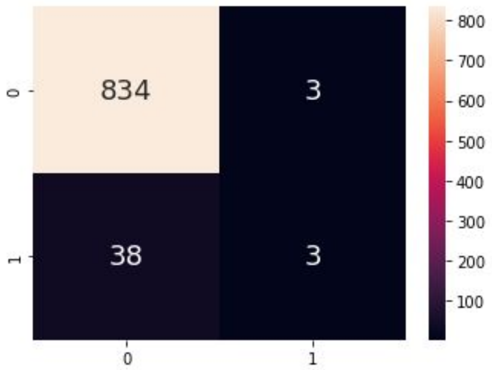


DECISION TREE CLASSIFICATION ACCURACY

TPR FOR TRAIN	TNR FOR TRAIN	TPR FOR TEST	TNR FOR TEST
0.10483870967741936	0.9982290436835891	0.07317073170731707	0.996415770609319

```
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.9533029612756264
<AxesSubplot:>
```

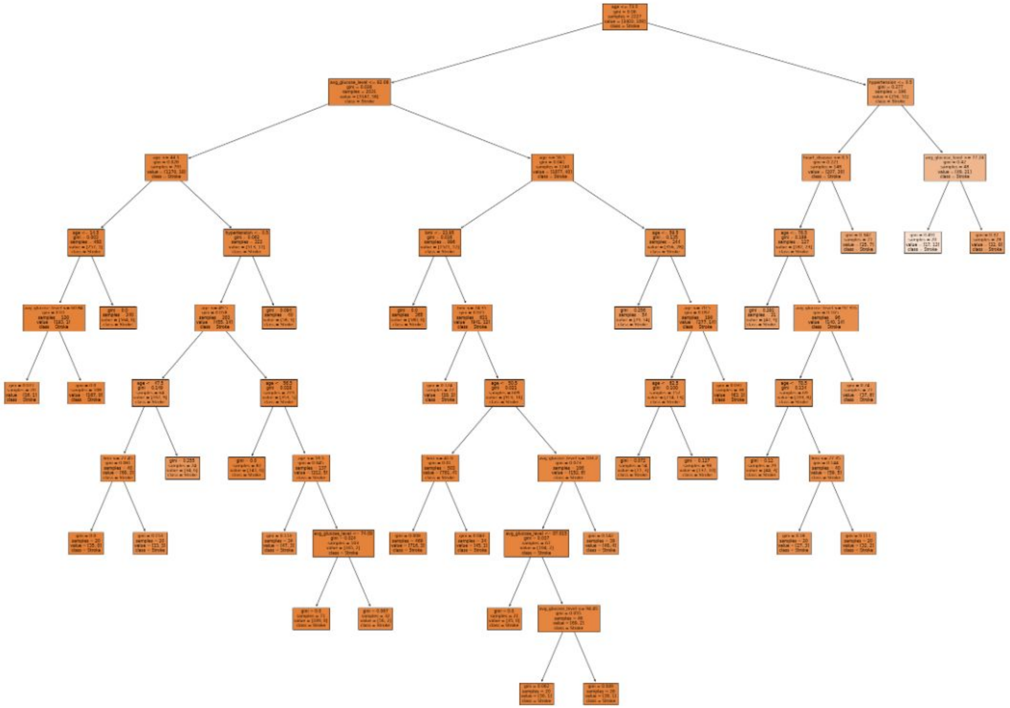




RANDOM FOREST CLASSIFICATION

DESCRIPTION

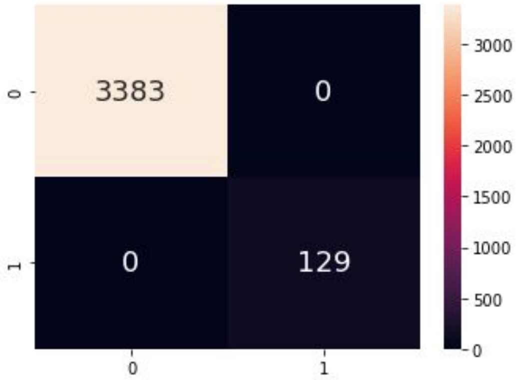
It combines the output of multiple decision trees to reach a single result.

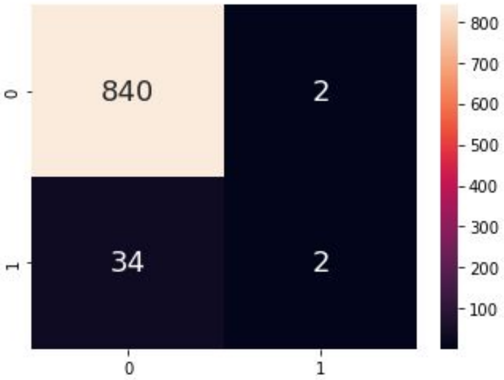


RANDOM FOREST CLASSIFICATION ACCURACY

TPR FOR TRAIN	TNR FOR TRAIN	TPR FOR TEST	TNR FOR TEST
1.0	1.0	0.05555555555555555	0.997624703087886

```
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.958997722095672
<AxesSubplot:>
```





DECISION TREE VS RANDOM FOREST REGRESSION

RANDOM FOREST HAVE A HIGHER CLASSIFICATION ACCURACY

RANDOM FOREST HAVE A HIGHER TPR AND TNR

BOTH HAVE SIMILAR TNR AND FNR

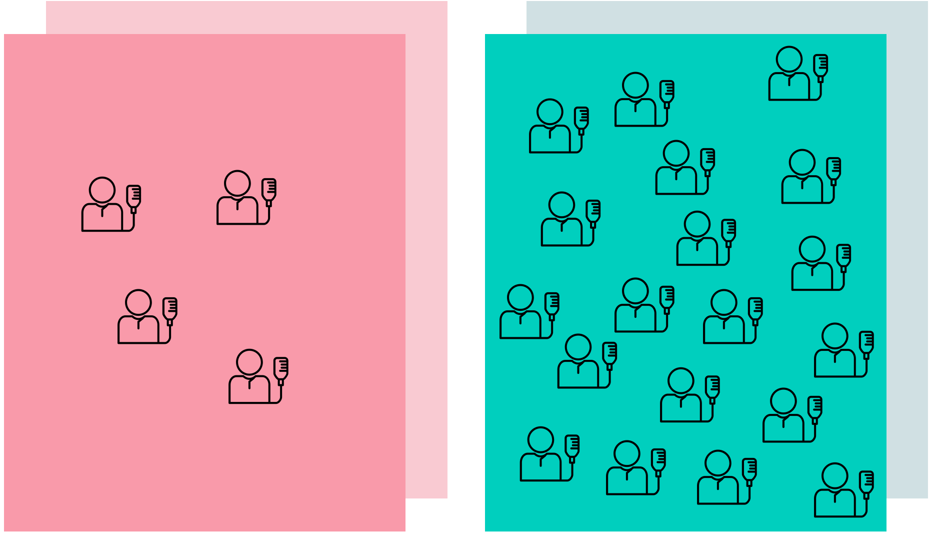
HENCE, WE DECIDED TO USE RANDOM FOREST REGRESSION TO PREDICT STROKE

04.

OUTCOME

Challenges faced and conclusion of analysis

CHALLENGES FACED



With Stroke

Without Stroke

IMBALANCED CLASSES

- More than 4000 out of 5110 patients without stroke
- Approximately 800 people with stroke
- High FNR and TNR

Data that may appear to be accurate, but is however **biased** and **useless!**

SMOTE

Synthetic Minority Oversampling Technique

DATA AUGMENTATION

A method similar to oversampling

Rather than generating identical data points, SMOTE adds small perturbations to the newly created data points

SMOTE-NC (NOMINAL & CONTINUOUS)

SMOTE-NC

NOMINAL
Creates synthetic data

CATEGORICAL
Resamples data

CHI-SQUARE & NAIVE-BAYES TEST

BASED ON THE CHI-SQUARE TEST

BMI

AVERAGE GLUCOSE LEVEL

CORROBORATED BY NAIVE-BAYES TEST

RANDOM FOREST DECISION TREE

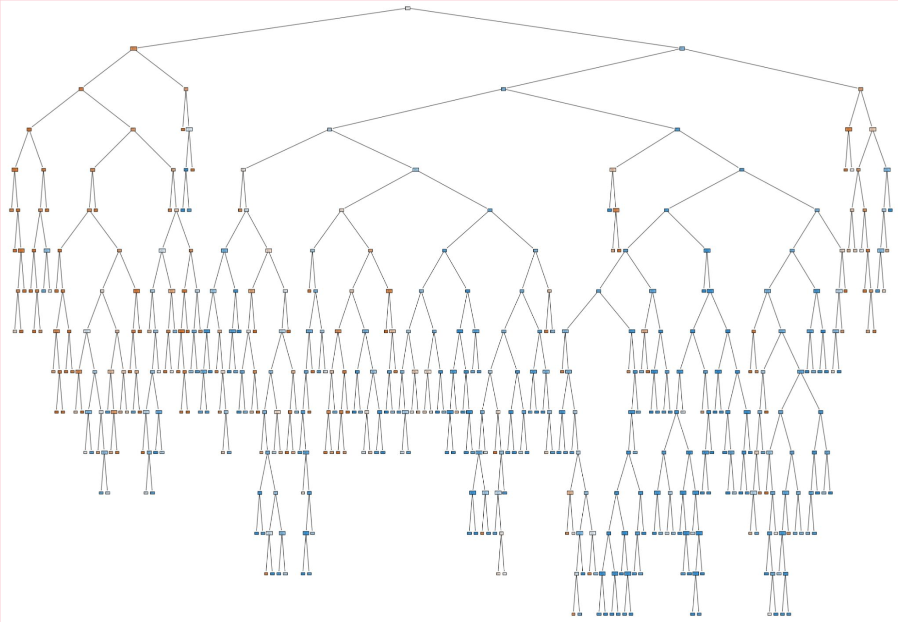
CLASSIFICATION ACCURACY

BEFORE

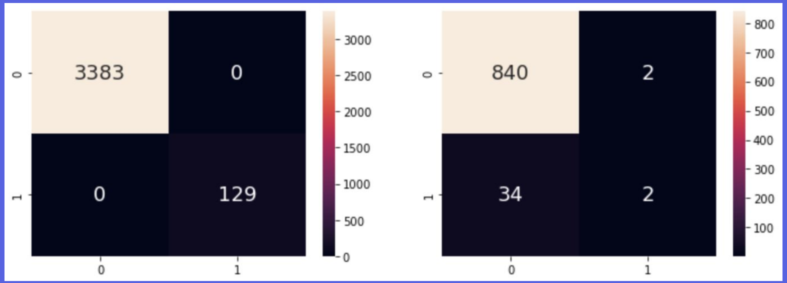
0.9632687927107062

AFTER

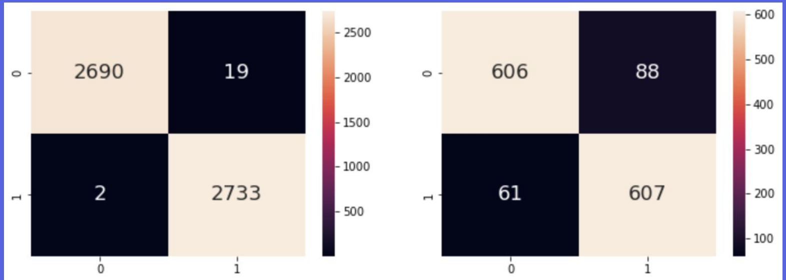
0.8906020558002937



RANDOM FOREST DECISION TREE



TRAIN BEFORE TEST



TRAIN AFTER TEST

TRAIN

TPR: 1.0
FPR: 0.0
FNR: 0.0
TNR: 1.0

TEST

TPR: 0.05555555555555555
FPR: 0.0023752969121140144
FNR: 0.9444444444444444
TNR: 0.997624703087886

TRAIN

TPR: 0.9992687385740402
FPR: 0.007013658176448745
FNR: 0.0007312614259597807
TNR: 0.9929863418235512

TEST

TPR: 0.9086826347305389
FPR: 0.12680115273775217
FNR: 0.09131736526946108
TNR: 0.8731988472622478

CONCLUSION

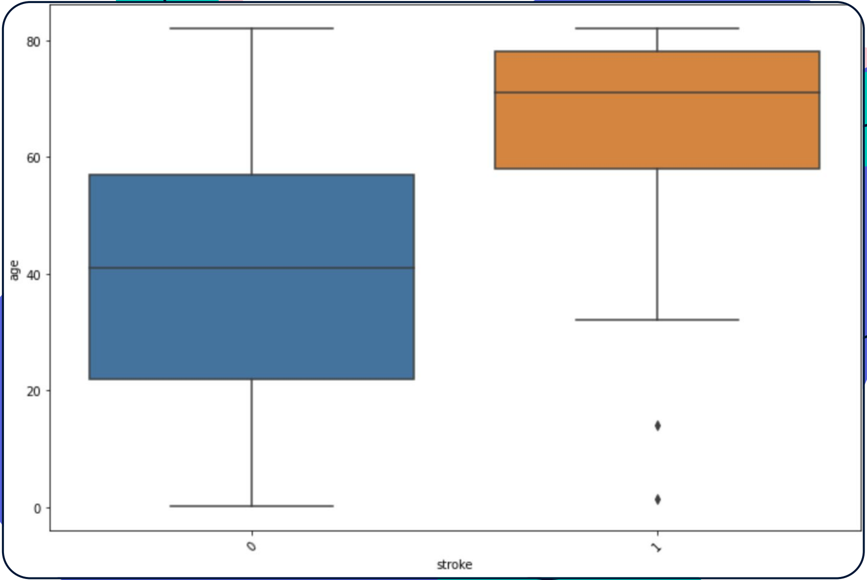
What have we learned
from our analysis?

GENERAL FINDINGS

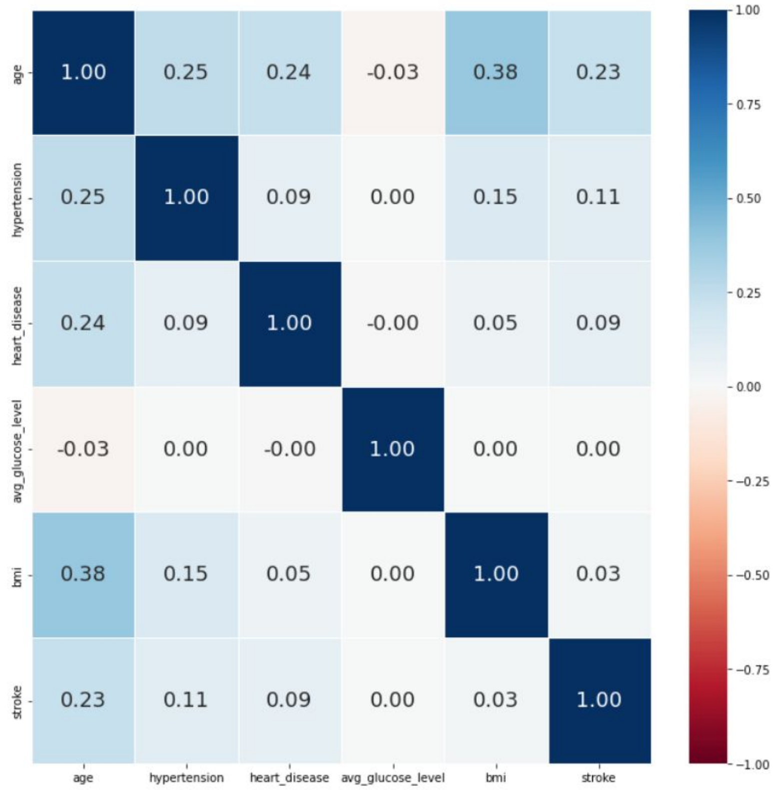
BMI IS ABOVE 30

MAJORITY OF RESPONDENTS WERE OBESE

THE DATA CAN BE OBTAINED FROM DIFFERENT GEOGRAPHICAL LOCATIONS, TO ENSURE THAT THE BMI IS MORE SPREADED OUT



CORRELATION



EXISTENCE OF HEART DISEASE

EXISTENCE OF HYPERTENSION

HIGH GLUCOSE LEVELS

AGE

Older people are more susceptible to stroke

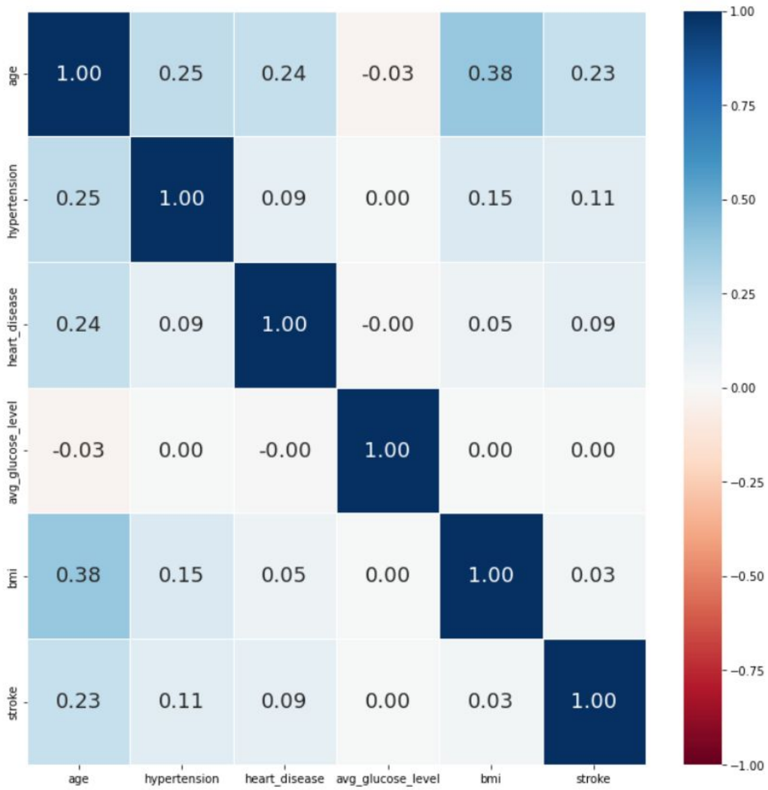
WHICH FACTORS ARE THE MOST IMPORTANT IN PREDICTING THE OCCURRENCE OF STROKE, AND HOW WE CAN PREVENT THE AGGRAVATION OF SUCH FACTORS

MOST IMPORTANT FACTOR

NO ONE CLEAR FACTOR

The correlation between stroke and the other factors is **0.38** or less

Combination of a few factors to be considered concurrently



WHICH FACTORS ARE THE MOST IMPORTANT IN PREDICTING THE OCCURRENCE OF STROKE, AND HOW WE CAN PREVENT THE AGGRAVATION OF SUCH FACTORS

HEALTH MONITORING

Encourage older patients to constantly monitor their health and visit doctors for health checkups regularly.

HEALTH MONITORING

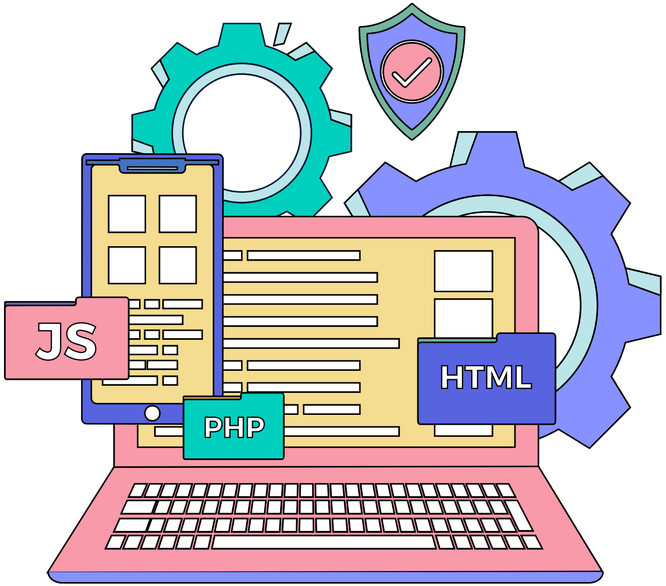
People with conditions such as hypertension, heart disease and diabetes should be encouraged to have their doctors monitor them

HEALTHY CONSUMPTION

The general public should be educated on the negative effects of excessive consumption of sugar to prevent high glucose levels.

LOSE WEIGHT

Overweight people should be encouraged to put in efforts to lose some weight in order to reduce their chances of having a stroke.



THANK YOU