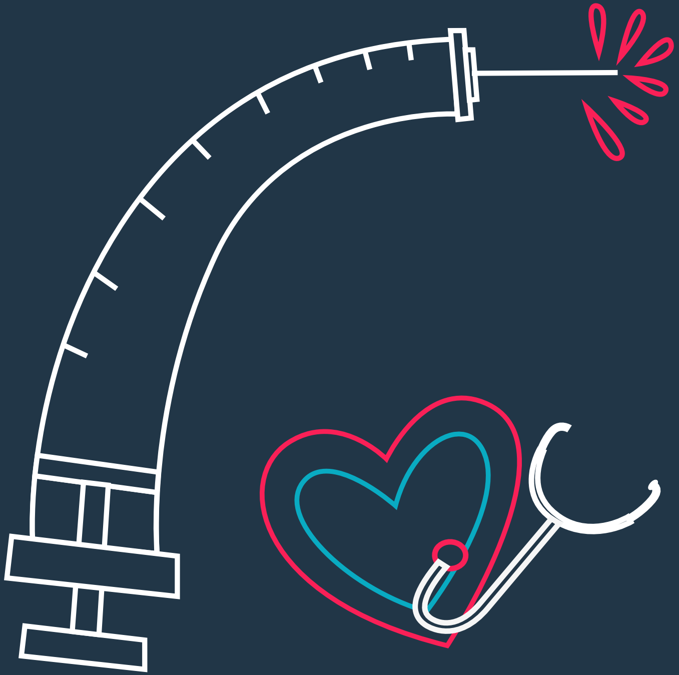


Stroke Prediction

Team2 SC1015 Project

Douglas Toh,
Ivan Ong,
Yeoh Yu Shyan

TABLE OF CONTENTS

01 Ideation and problem statement

02 Data preparation

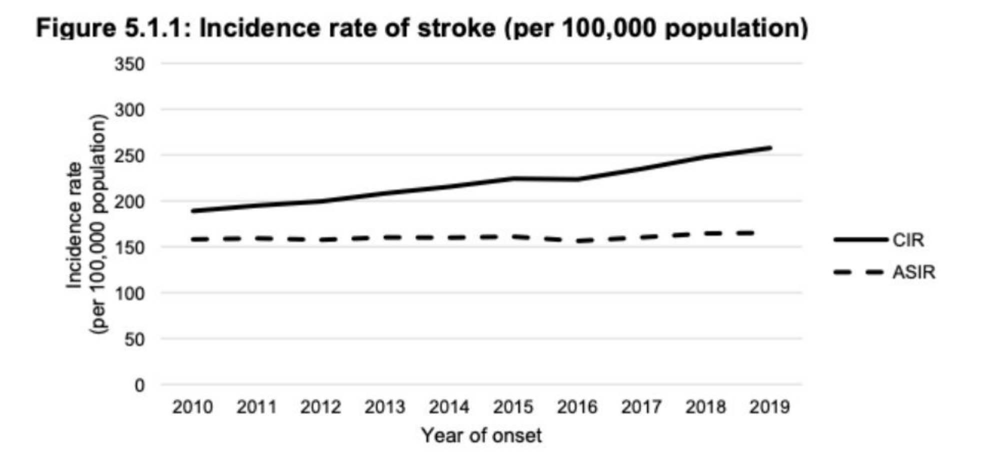
- Data Cleaning

03 Exploration Data Analysis

- Logistic regression & Random forest classifier
- Smote
- Ada Boost

04 Outcome and conclusion

Ideation/Problem statement



- It has been noted that the growth rate of stroke has been increasing at an alarming rate in Singapore
- What are the highest factors that contributes to stroke.

Data set

Factors Affecting Stroke

- Gender
- Age
- Hypertension
- Heart-Disease
- Marriage status
- Work type
- Residence type
- Smoking status
- Average Glucose level
- BMI

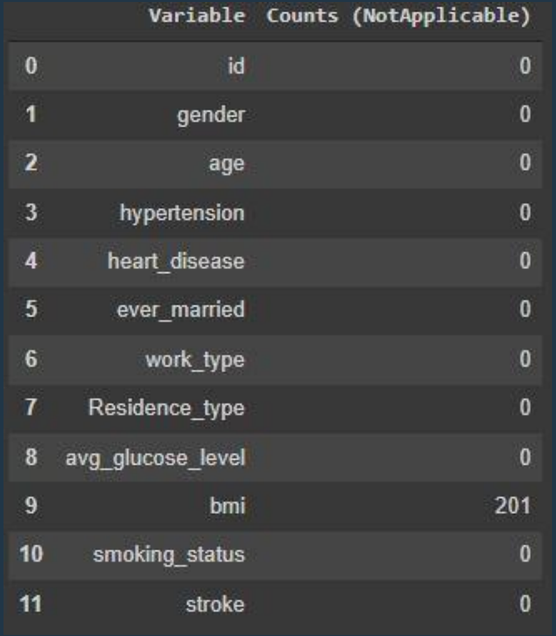
Total of 11 Columns

Categorical: 8 Variables

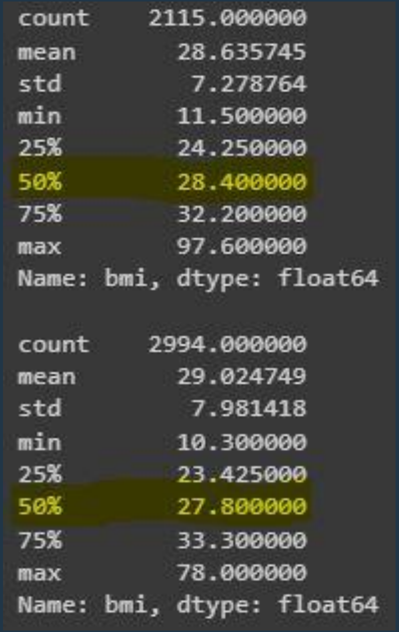
Numerical: 3 Variables

5110 Rows of Data

Data Cleaning & Preparation



	Variable	Counts (NotApplicable)
0	id	0
1	gender	0
2	age	0
3	hypertension	0
4	heart_disease	0
5	ever_married	0
6	work_type	0
7	Residence_type	0
8	avg_glucose_level	0
9	bmi	201
10	smoking_status	0
11	stroke	0



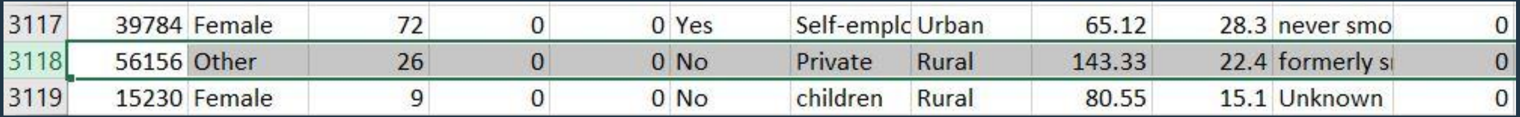
```
count    2115.000000
mean       28.635745
std         7.278764
min        11.500000
25%        24.250000
50%        28.400000
75%        32.200000
max        97.600000
Name: bmi, dtype: float64

count    2994.000000
mean       29.024749
std         7.981418
min        10.300000
25%        23.425000
50%        27.800000
75%        33.300000
max        78.000000
Name: bmi, dtype: float64
```

201 missing BMIs values

- Use medium value of male and female to replace the missing BMI values

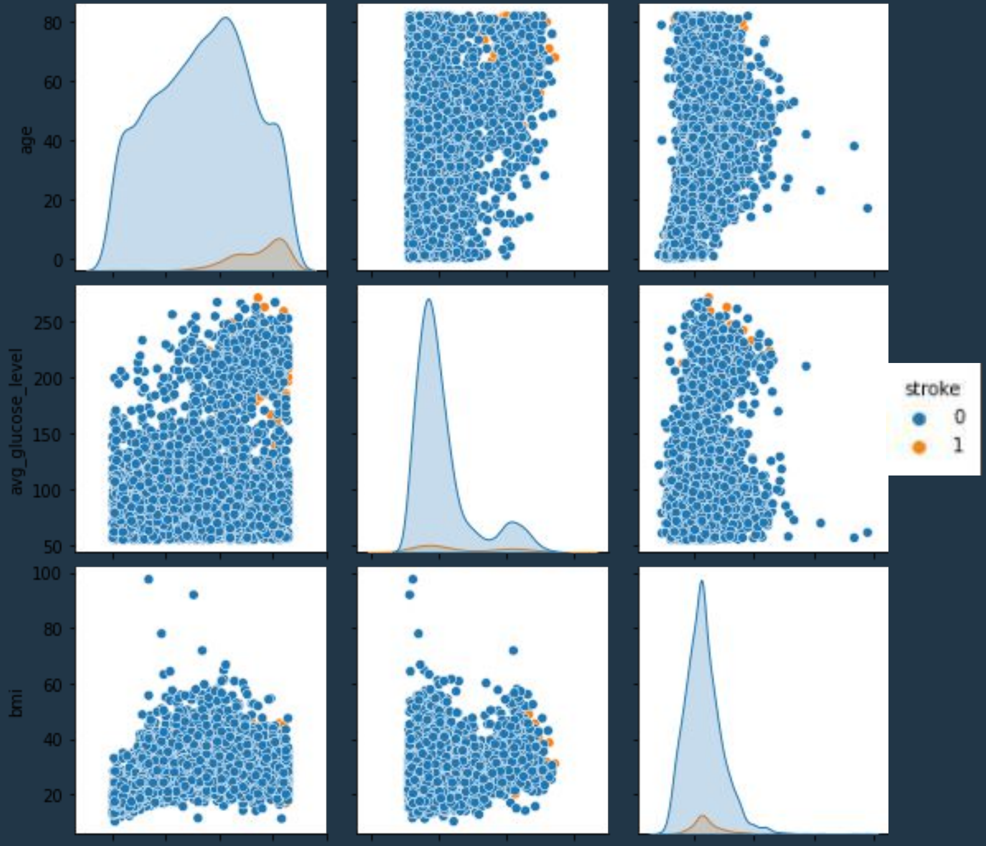
1 instance of gender labeled ‘others’

- Removed



3117	39784	Female	72	0	0	Yes	Self-emplc	Urban	65.12	28.3	never smo	0
3118	56156	Other	26	0	0	No	Private	Rural	143.33	22.4	formerly sı	0
3119	15230	Female	9	0	0	No	children	Rural	80.55	15.1	Unknown	0

Exploratory Data Analysis



Conduct a high-level Descriptive Statistical analysis on the dataset.

Patients with stroke are overlapping with patients without stroke which indicates that they are not clearly separated.

Exploratory Data Analysis

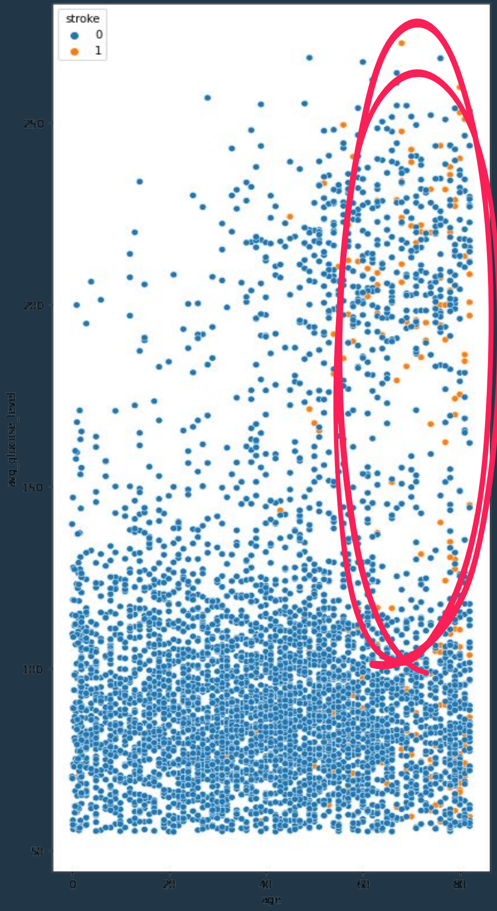
Scatterplot

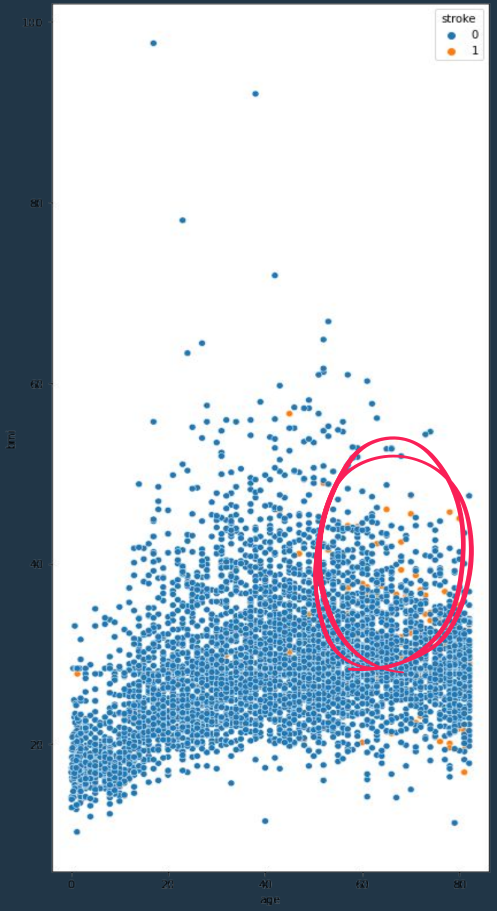
Age with Avg_glucose_level

- Older age and higher avg_glucose level has higher stroke count

Age with BMI

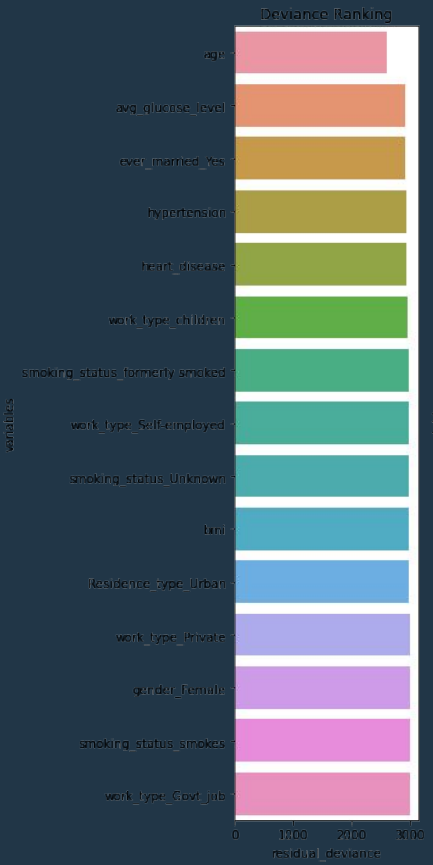
- Older age and higher BMI has higher stroke count



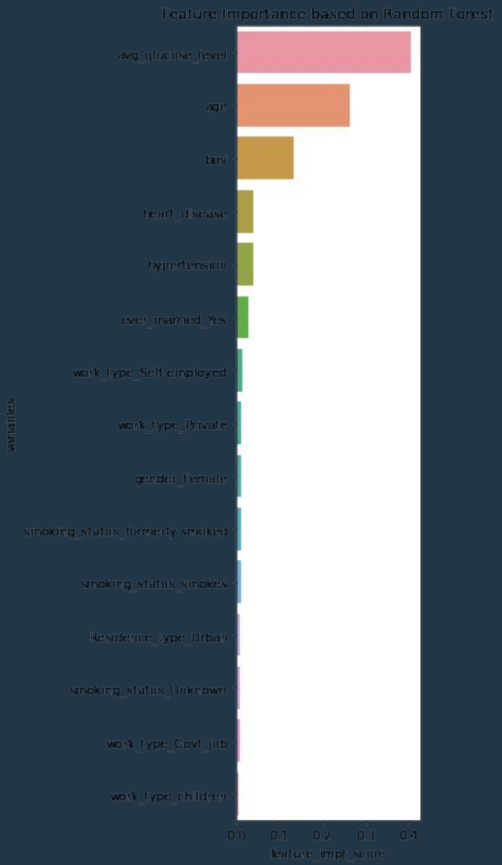


Logistic Regression & Random Forest Classifier

- Feature Importance Ranking using Deviance(LR)
  1. Deviance Ranking
  2. Random Forest Classifier
- Note “age”, “avg_glucose_level”, “ever_married”, “hypertension” & “heart disease”
- “bmi” is a stronger factor than “ever_married” for our data study.



Deviance Ranking



Feature Importance based on Random Forest

Applying SMOTE

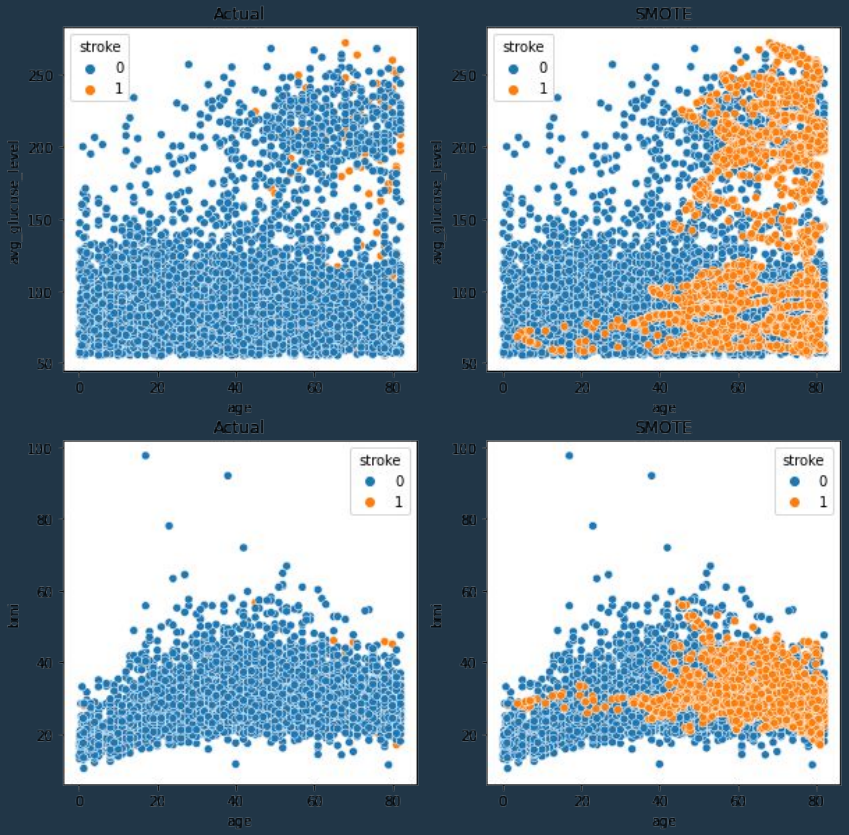
SMOTE

Synthetic Minority Oversampling Technique

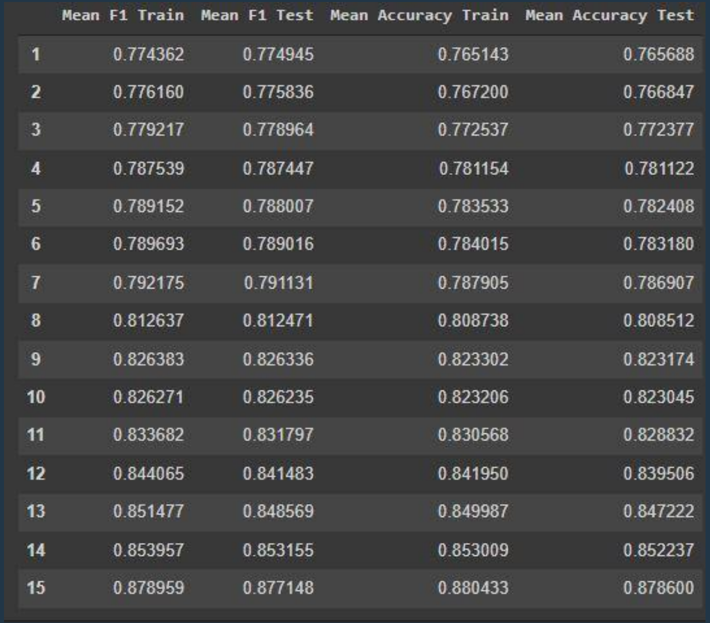
- To address imbalance of data
- K-nearest neighbour
- 4611 instances with stroke is generated

Synthetic Data was generated to the distribution of the original samples.

- Age >40 is seen to have more stroke cases.
- Small number of instance age 40<

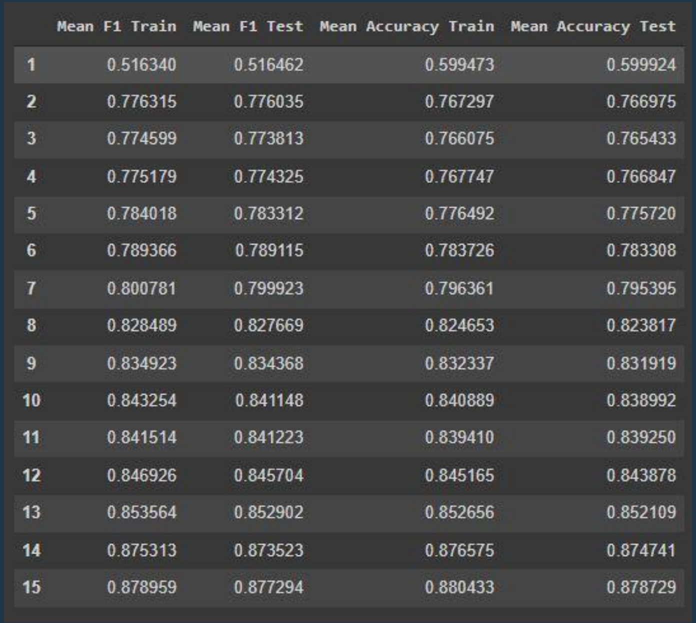


(Deviance) Logistic Regression



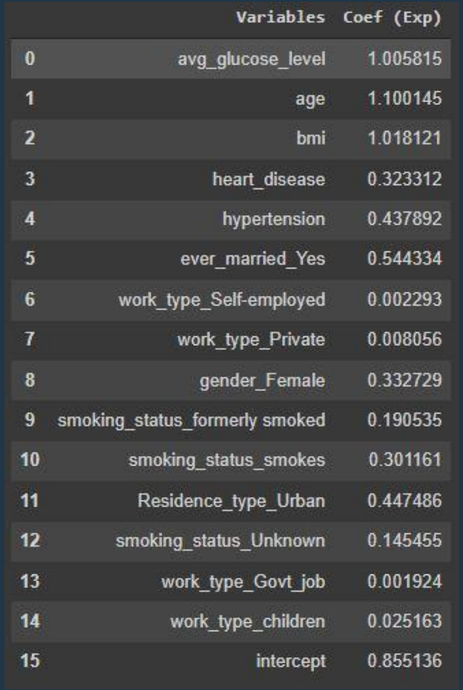
	Mean F1 Train	Mean F1 Test	Mean Accuracy Train	Mean Accuracy Test
1	0.774362	0.774945	0.765143	0.765688
2	0.776160	0.775836	0.767200	0.766847
3	0.779217	0.778964	0.772537	0.772377
4	0.787539	0.787447	0.781154	0.781122
5	0.789152	0.788007	0.783533	0.782408
6	0.789693	0.789016	0.784015	0.783180
7	0.792175	0.791131	0.787905	0.786907
8	0.812637	0.812471	0.808738	0.808512
9	0.826383	0.826336	0.823302	0.823174
10	0.826271	0.826235	0.823206	0.823045
11	0.833682	0.831797	0.830568	0.828832
12	0.844065	0.841483	0.841950	0.839506
13	0.851477	0.848569	0.849987	0.847222
14	0.853957	0.853155	0.853009	0.852237
15	0.878959	0.877148	0.880433	0.878600

Deviance Ranking Table



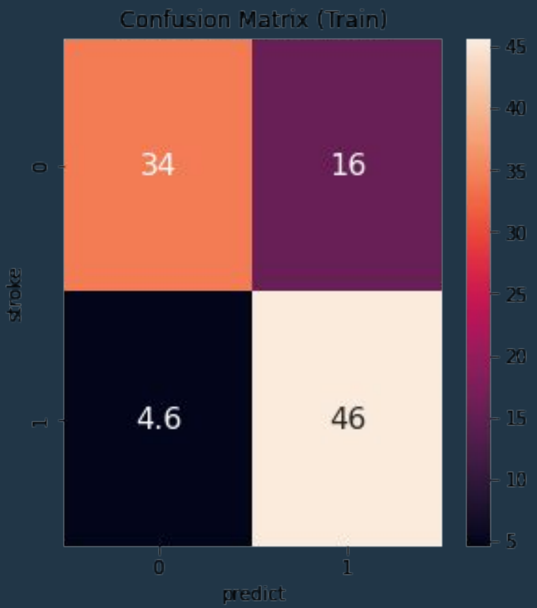
	Mean F1 Train	Mean F1 Test	Mean Accuracy Train	Mean Accuracy Test
1	0.516340	0.516462	0.599473	0.599924
2	0.776315	0.776035	0.767297	0.766975
3	0.774599	0.773813	0.766075	0.765433
4	0.775179	0.774325	0.767747	0.766847
5	0.784018	0.783312	0.776492	0.775720
6	0.789366	0.789115	0.783726	0.783308
7	0.800781	0.799923	0.796361	0.795395
8	0.828489	0.827669	0.824653	0.823817
9	0.834923	0.834368	0.832337	0.831919
10	0.843254	0.841148	0.840889	0.838992
11	0.841514	0.841223	0.839410	0.839250
12	0.846926	0.845704	0.845165	0.843878
13	0.853564	0.852902	0.852656	0.852109
14	0.875313	0.873523	0.876575	0.874741
15	0.878959	0.877294	0.880433	0.878729

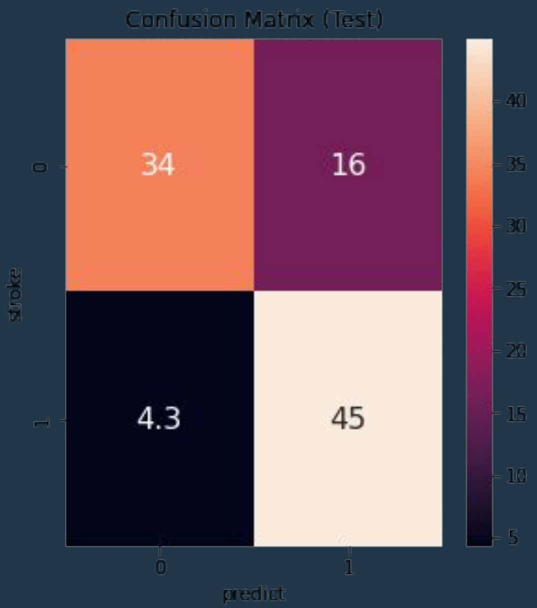
Feature Ranking Table



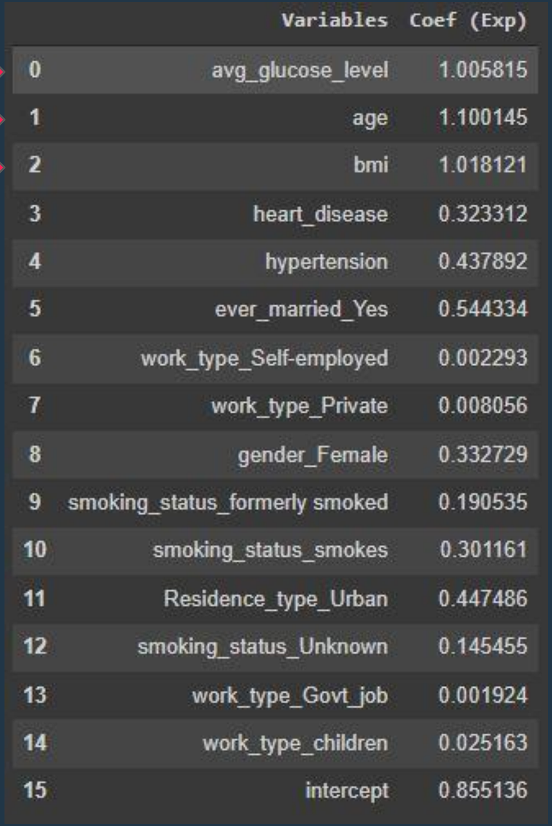
	Variables	Coef (Exp)
0	avg_glucose_level	1.005815
1	age	1.100145
2	bmi	1.018121
3	heart_disease	0.323312
4	hypertension	0.437892
5	ever_married_Yes	0.544334
6	work_type_Self-employed	0.002293
7	work_type_Private	0.008056
8	gender_Female	0.332729
9	smoking_status_formerly smoked	0.190535
10	smoking_status_smokes	0.301161
11	Residence_type_Urban	0.447486
12	smoking_status_Unknown	0.145455
13	work_type_Govt_job	0.001924
14	work_type_children	0.025163
15	intercept	0.855136

Confusion Matrix & Coefficient Table





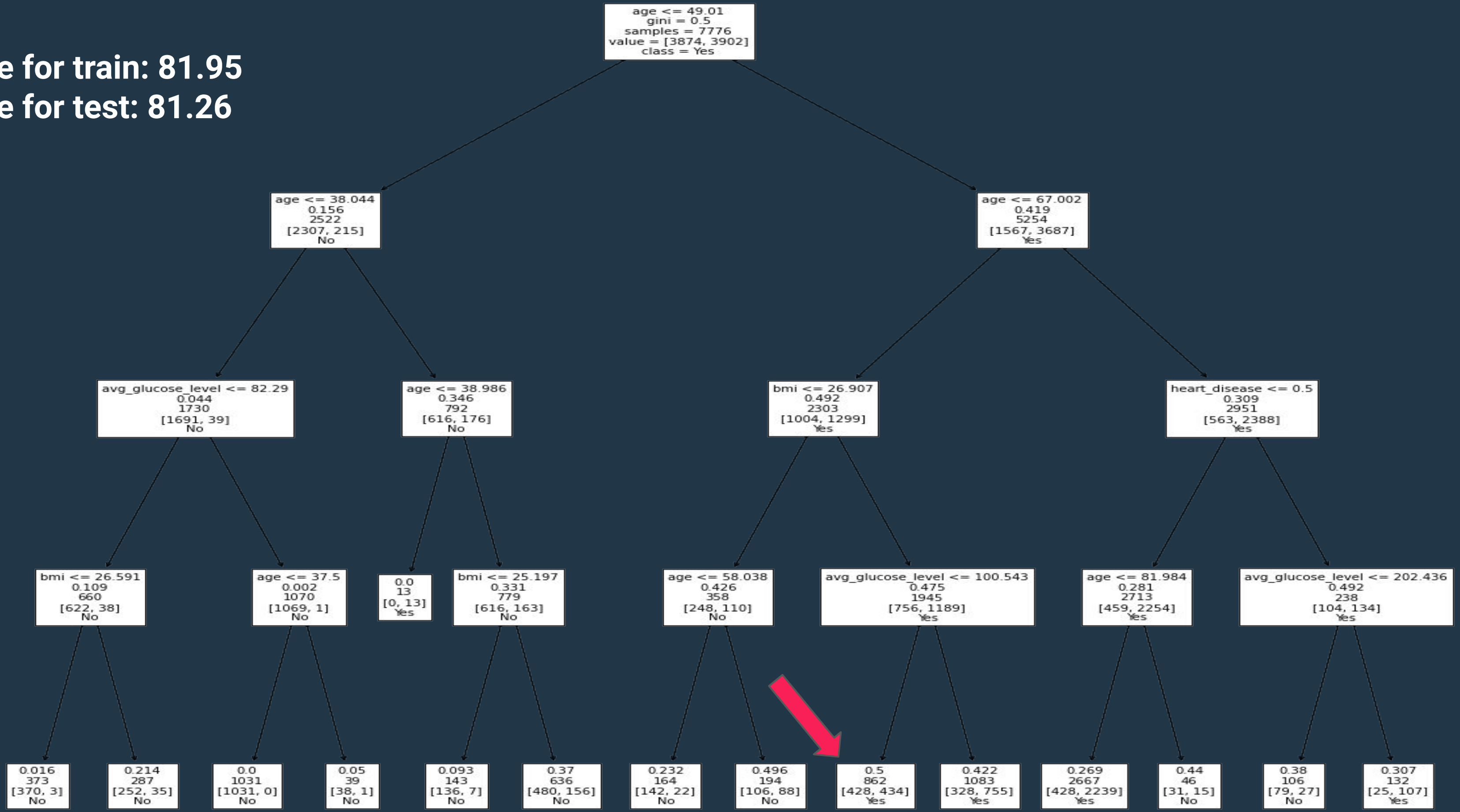
F1 Score for train: 87.84
F1 Score for test: 87.52



	Variables	Coef (Exp)
0	avg_glucose_level	1.005815
1	age	1.100145
2	bmi	1.018121
3	heart_disease	0.323312
4	hypertension	0.437892
5	ever_married_Yes	0.544334
6	work_type_Self-employed	0.002293
7	work_type_Private	0.008056
8	gender_Female	0.332729
9	smoking_status_formerly smoked	0.190535
10	smoking_status_smokes	0.301161
11	Residence_type_Urban	0.447486
12	smoking_status_Unknown	0.145455
13	work_type_Govt_job	0.001924
14	work_type_children	0.025163
15	intercept	0.855136

Decision Tree

F1 Score for train: 81.95
F1 Score for test: 81.26



AdaBoost

age <= 49.01
gini = 0.5
samples = 7776
value = [3874, 3902]
class = Yes
age <= 49.01
gini = 0.5
samples = 7776
value = [3874, 3902]
class = Yes
age <= 49.01
gini = 0.5
samples = 7776
value = [3874, 3902]
class = Yes
age <= 38.044
0.156
2522
[2307, 215]
No
age <= 38.044
0.156
age <= 38.044
0.156
age <= 67.002
0.419
age <= 67.002
0.419
age <= 67.002
0.419
5254
[1567, 3687]
Yes
avg_glucose_level <= 82.29
0.044
1730
[1691, 39]
No
age <= 38.986
0.346
792
[616, 176]
No
heart_disease <= 0.5
0.309
2951
[563, 2388]
Yes
age <= 37.5
0.002
1070
[1069, 1]
No
bmi <= 25.197
0.331
779
[616, 163]
No
0.109
660
[622, 38]
No
0.426
358
[248, 110]
No
0.002
1070
[1069, 1]
No
0.331
779
[616, 163]
No
0.281
2713
Yes
avg_glucose_level <= 100.543
0.475
1945
[756, 1189]
Yes
age <= 81.984
0.281
2713
[459, 2254]
Yes
0.0
13
[0, 13]
Yes
0.214
287
[252, 35]
No
0.0
1031
[1031, 0]
No
0.016
373
[370, 3]
No
0.05
39
[38, 1]
No
0.093
143
[136, 7]
No
0.214
287
[252, 35]
No
0.0
1031
[1031, 0]
No
0.05
39
[38, 1]
No
0.093
143
[136, 7]
No
0.37
636
[480, 156]
No
0.232
164
[142, 22]
No
0.496
194
[106, 88]
No
0.5
862
[428, 434]
Yes
0.422
1083
[328, 755]
Yes
0.269
2667
[428, 2239]
Yes
0.44
46
[31, 15]
No

MULTIPLE DECISION TREES

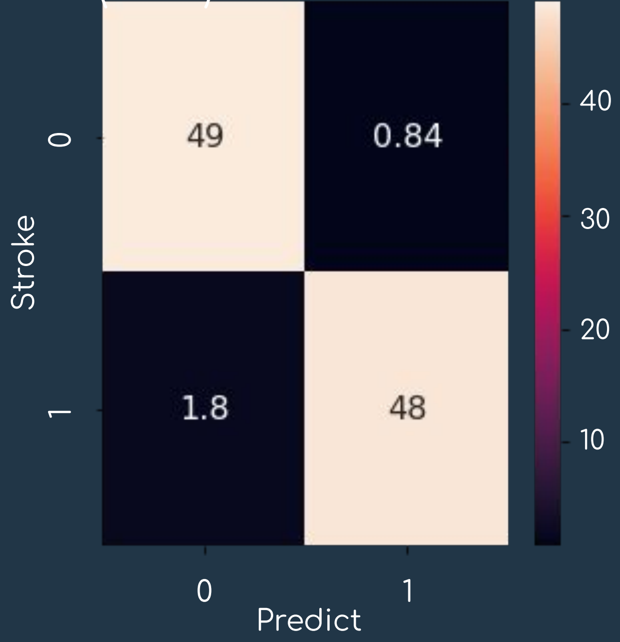
AdaBoost

Learning Rate → [0.2,0.4,0.8,1.0]

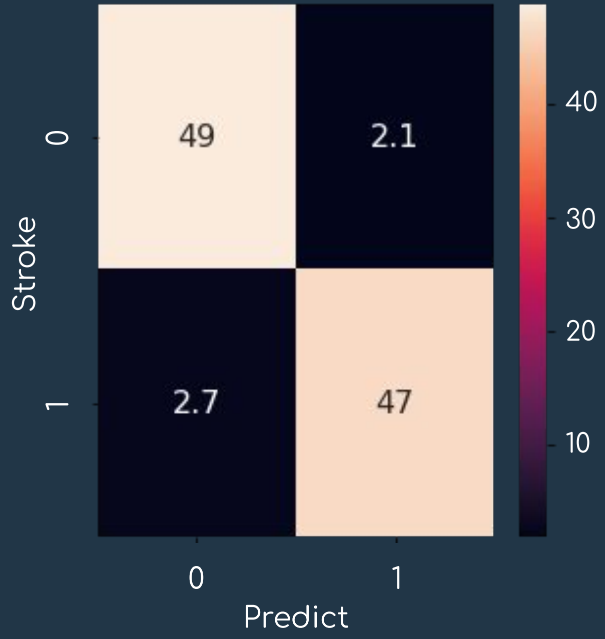
Grid Search

Best Parameter = 0.2

Confusion Matrix (Train)



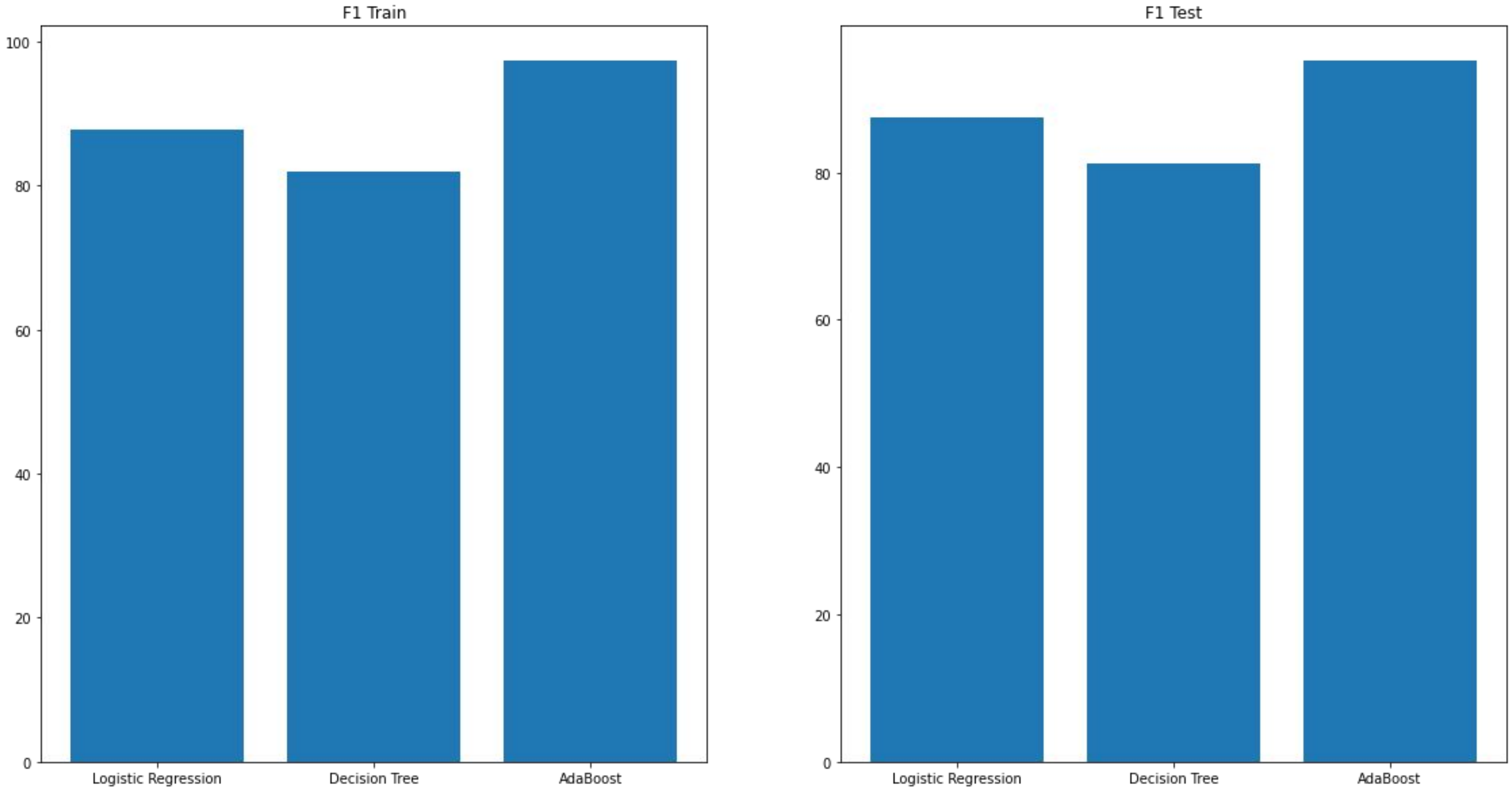
Confusion Matrix (Test)



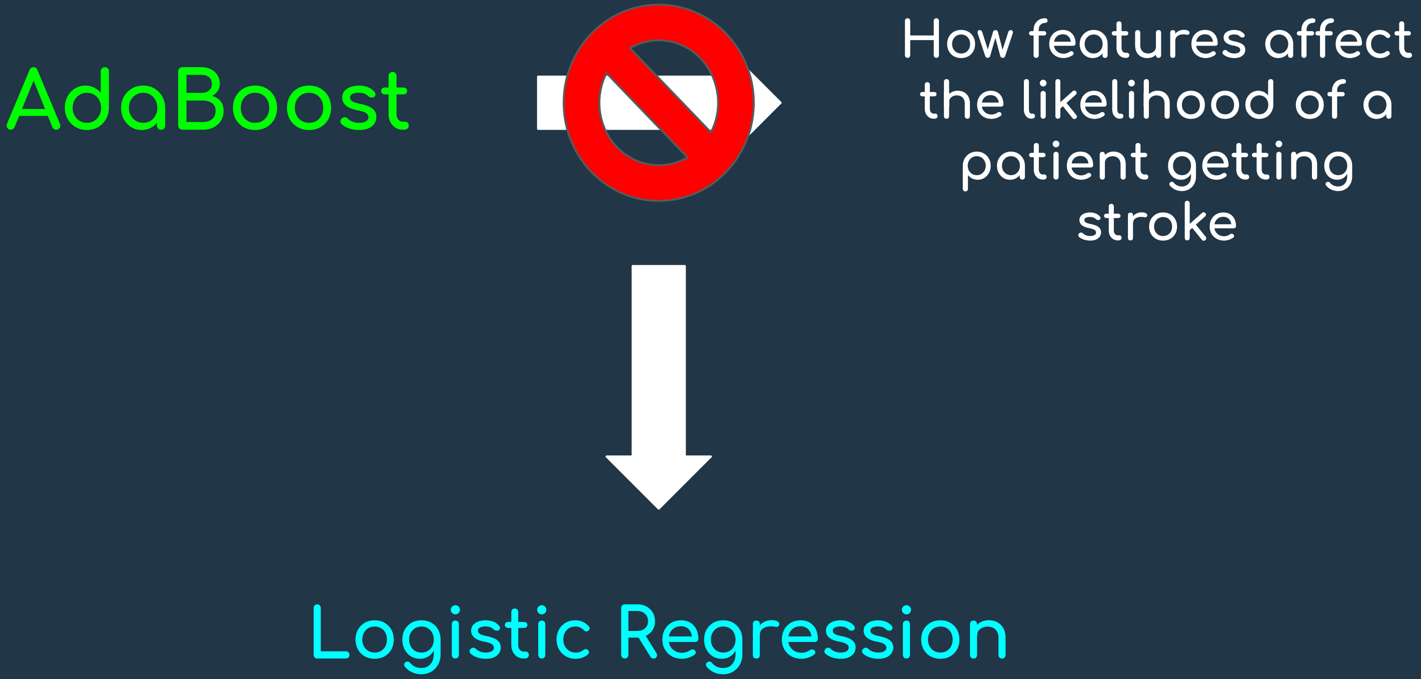
F1 Score for train: 97.32

F1 Score for test: 95.17

AdaBoost


F1 Train
100
80
60
40
20
0
Logistic Regression
Decision Tree
AdaBoost
F1 Test
80
60
40
20
0
Logistic Regression
Decision Tree
AdaBoost

AdaBoost



AdaBoost

How features affect the likelihood of a patient getting stroke

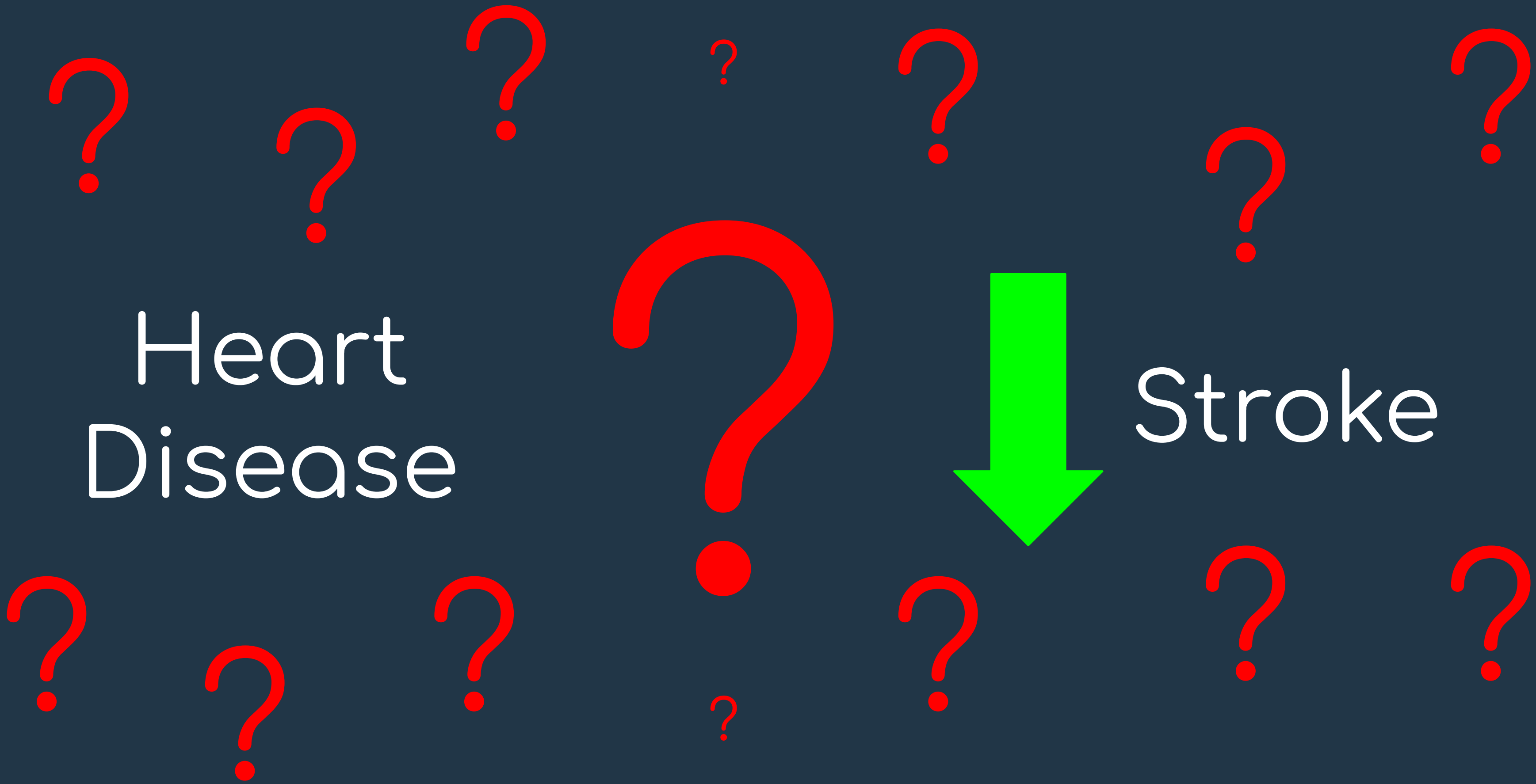
Logistic Regression

Conclusion

Logistic Regression

1. Age
2. BMI
3. Average Glucose Level
4. Heart Disease

Conclusion



Heart Disease

Stroke

Conclusion

THANK YOU!