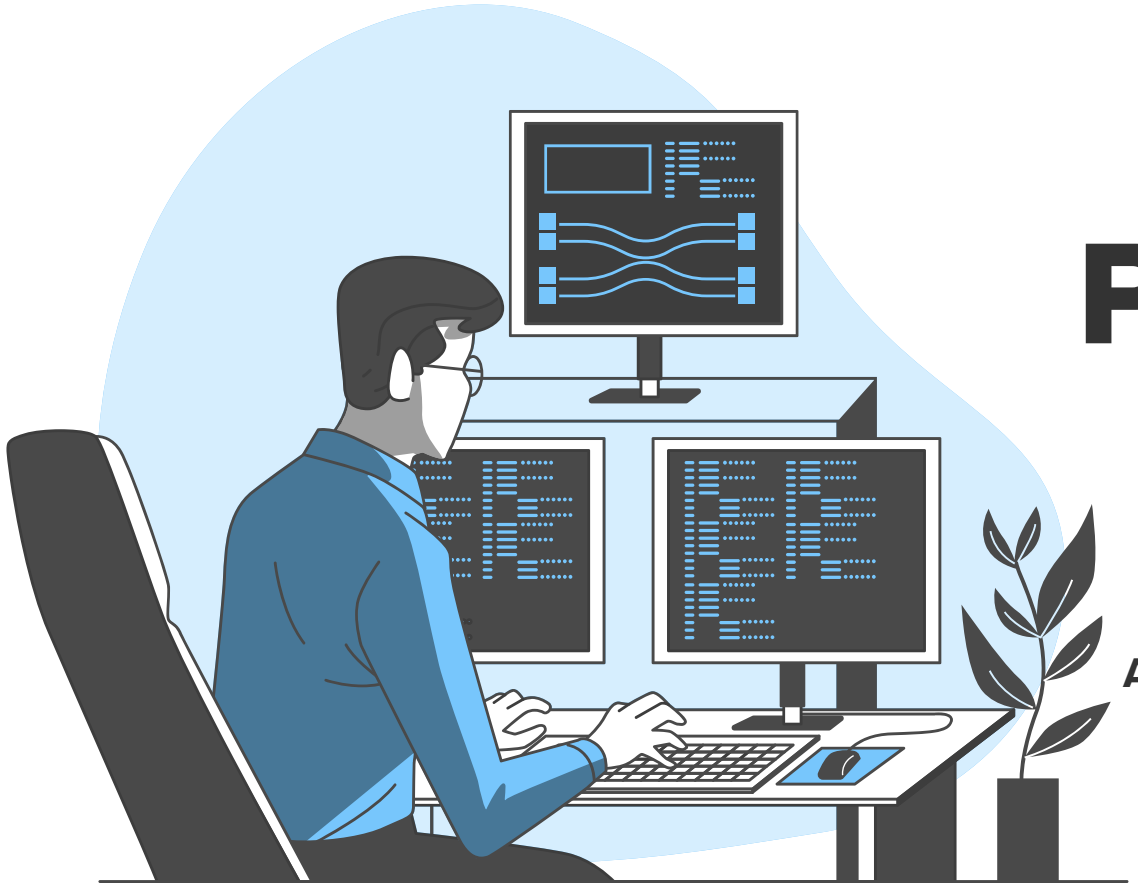


Stroke Prediction

SC1015 A133:

Teh Min Ze (U2111370H)

Teo Zhi Hao (U2222650F)

Ang Jia Wei Leon (U2222065K)

Contents

01

Practical Motivation

Practical Motivation

Stroke Prediction Dataset from Kaggle was used.

Purpose:

To accurately predict stroke based on common variables shared

Importance in today's context:

- Stroke is the 4th leading cause of death in Singapore (WSO)
- Variables are familiar and easy to manage

03

Data Preparation

Data Preparation

Remove entries specified as ‘N/A’

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4909 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 4909 non-null   int64  
 1   gender             4909 non-null   object 
 2   age                4909 non-null   float64
 3   hypertension       4909 non-null   int64  
 4   heart_disease      4909 non-null   int64  
 5   ever_married       4909 non-null   object 
 6   work_type          4909 non-null   object 
 7   Residence_type     4909 non-null   object 
 8   avg_glucose_level  4909 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     4909 non-null   object 
 11  stroke             4909 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 498.6+ KB
```

03
Exploratory Data Analysis

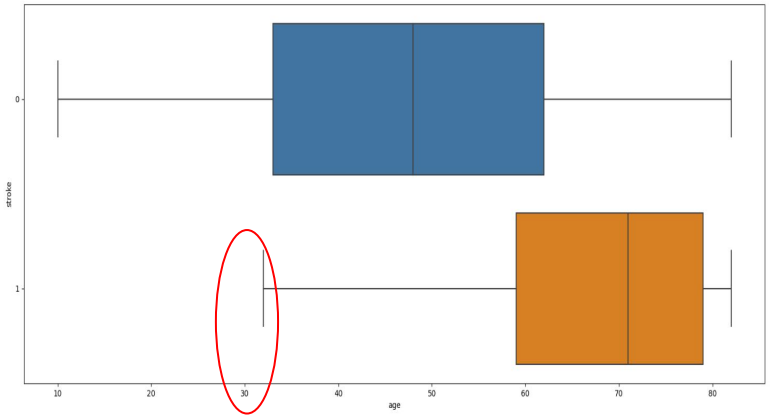
Variables in Dataset

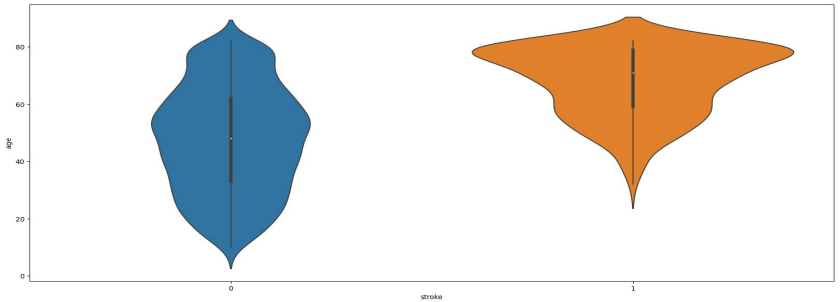
Response Variable: Stroke

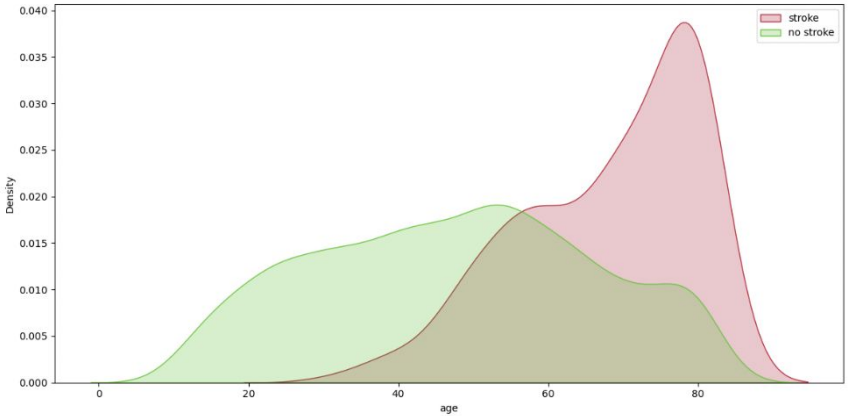
Numerical Variables
age
bmi
avg_glucose_level

Categorical Variables
gender
hypertension
heart_disease
ever_married
work_type
smoking_status
residence_type

Numerical Variables - Age



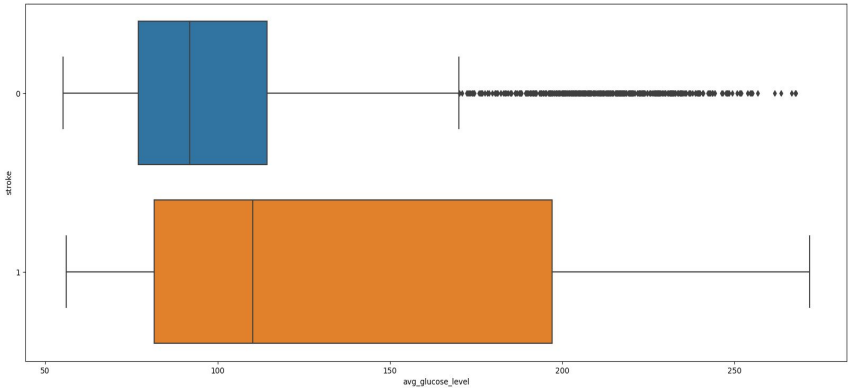


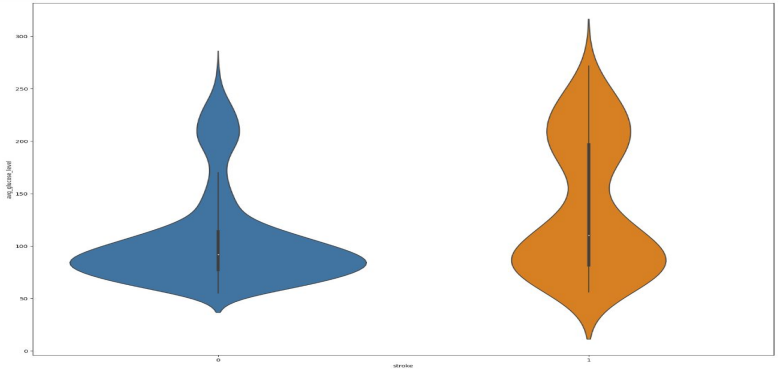


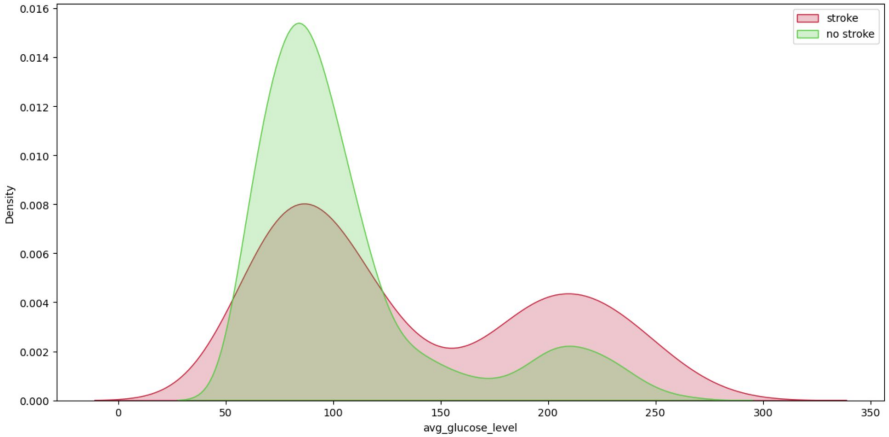
Observations:

1. **Minimum age for stroke patients > 30**
2. **Older people tend to suffer from stroke**

Numerical Variables - Average Glucose Level



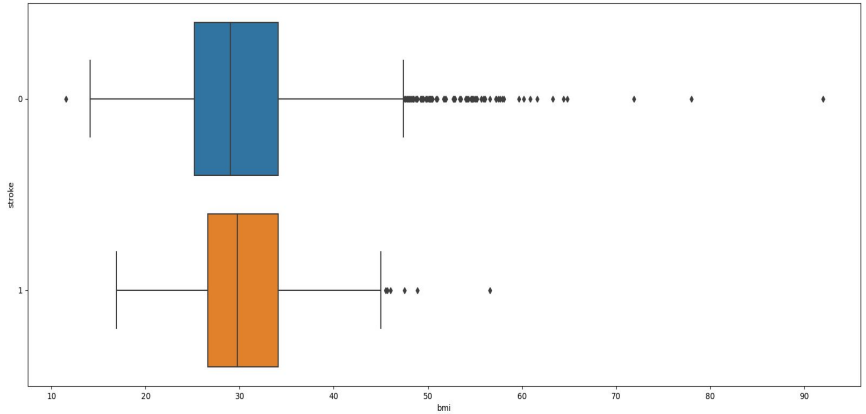


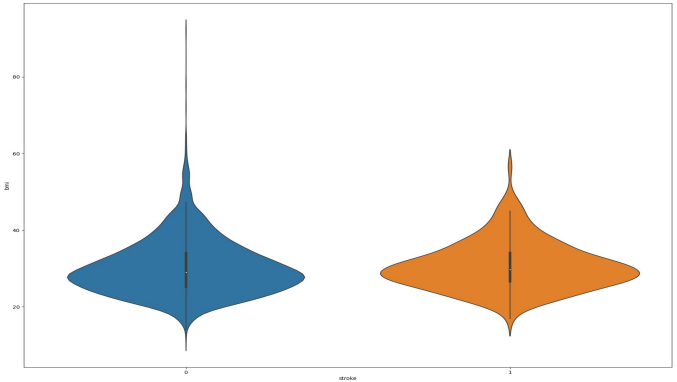


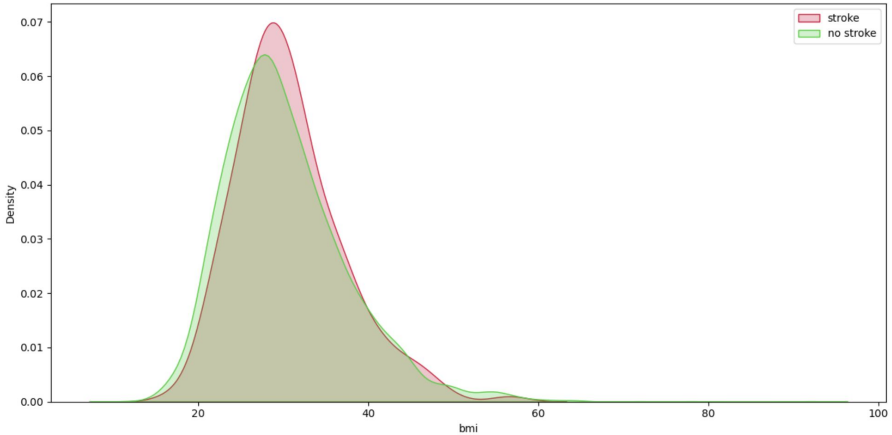
Observations:

1. **Large number of anomalies**
2. **More even distribution for patients with stroke**
3. **Patients with acceptable glucose levels are at less risk of stroke (72 - 99 mg/DL)**

Numerical Variables - Body Mass Index



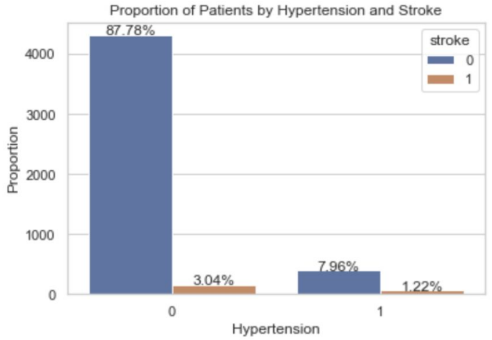


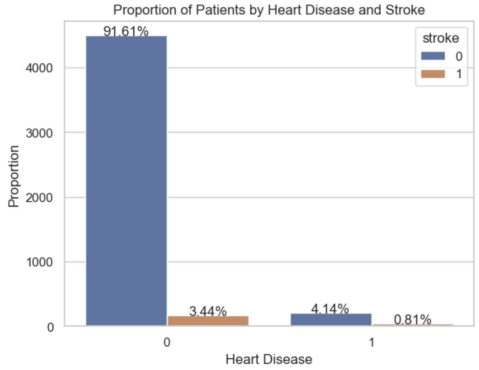


Observations:

1. **Large number of anomalies**
2. **Similar violin plot shapes → Similar distribution of data**
3. **Overlap in KDE plot**

Categorical Variables - Health

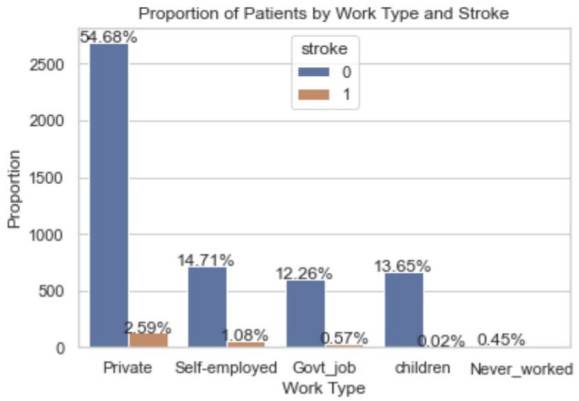


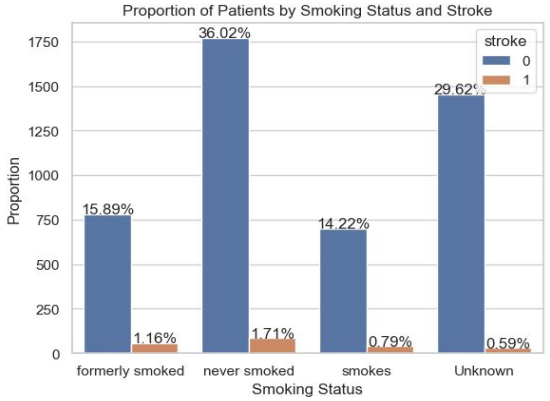


Observations:

- Higher proportions of patients with hypertension and heart disease have stroke
- Can deduce that having health related issues increase chance of stroke

Categorical Variables - Lifestyle

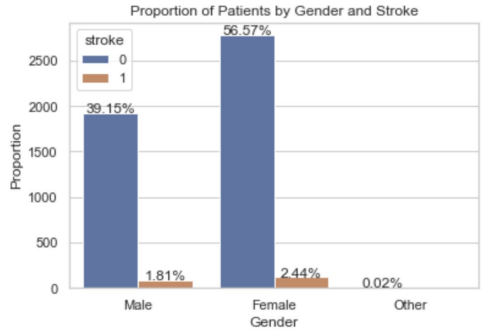


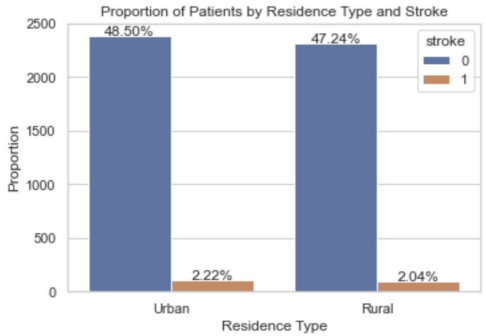


Observations:

- **Patients who are working (Private jobs, self-employed or government jobs) are more likely to be at risk of stroke compared to children and those who never worked**
- **Patients who smoke (formerly and currently) have a higher risk of suffering from a stroke**

Categorical Variables - Others



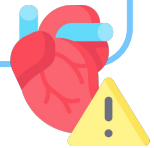


Observations:

- **Although there are more female patients than male, proportion of patients with stroke remain similar in both genders**
- **This is similar for residence type → is not an significant factor in stroke prediction**

Relationship between Variables

Positive Correlation between age and heart disease



Positive Correlation between age and hypertension

Patients who smoke tend to be from urban areas

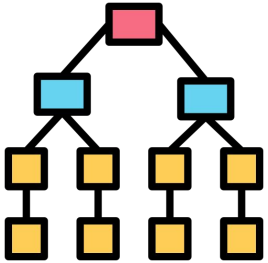
Patients who were ever married have higher BMI

Self employed individuals tend to be older than those of other categories

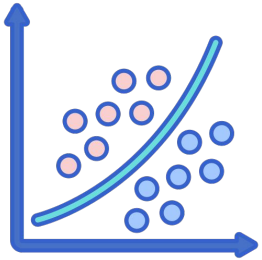
04

Machine Learning

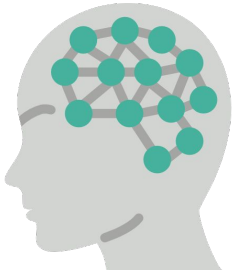
Machine Learning Models



Random Forest



Logistics Regression



Artificial Neural Network (ANN)

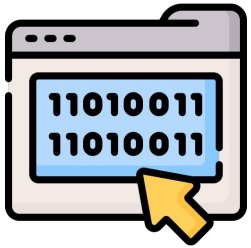
Metrics for Model Evaluation

F1 Score

Classification Accuracy

Random Forest Model

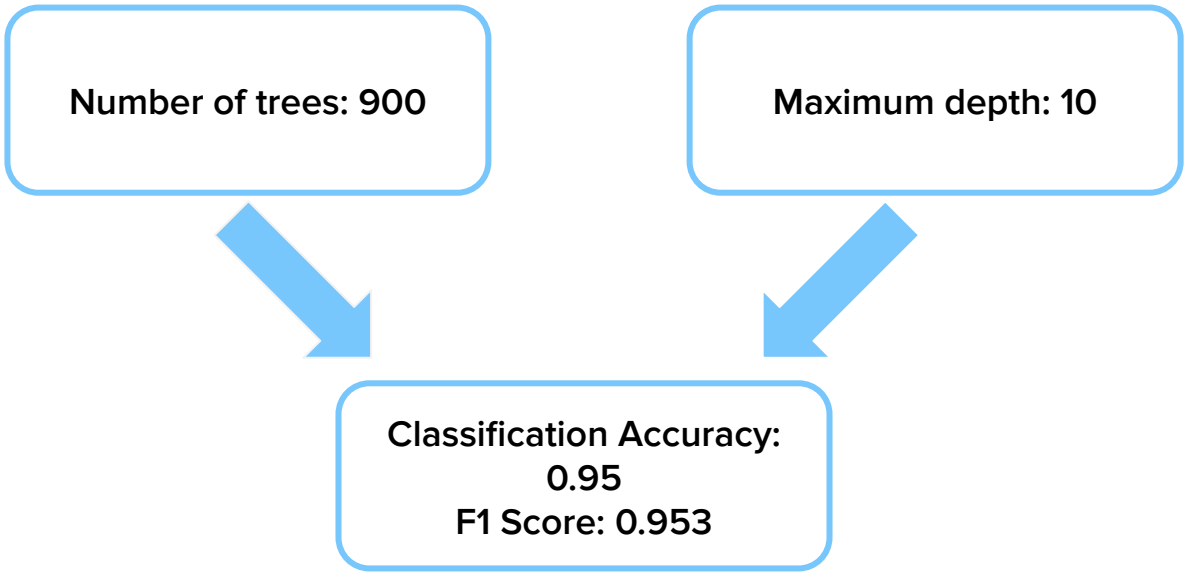
Class balancing



Encodes categorical variables via OHE (One-Hot-Encoding)

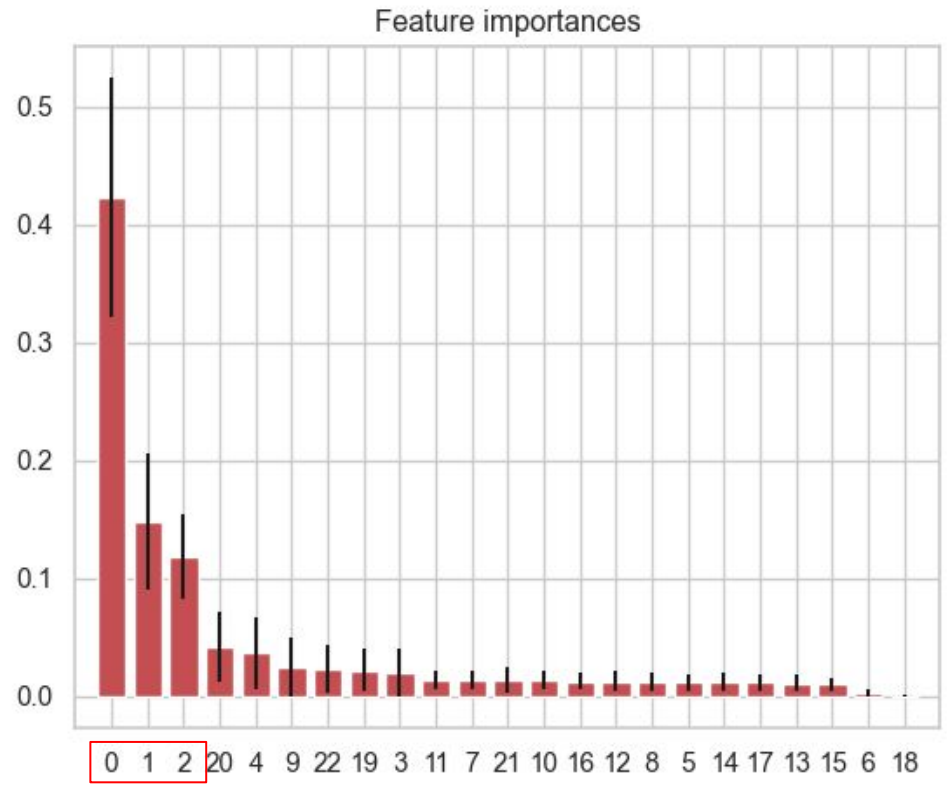
Fine tuning of number of trees and depth

Random Forest Model

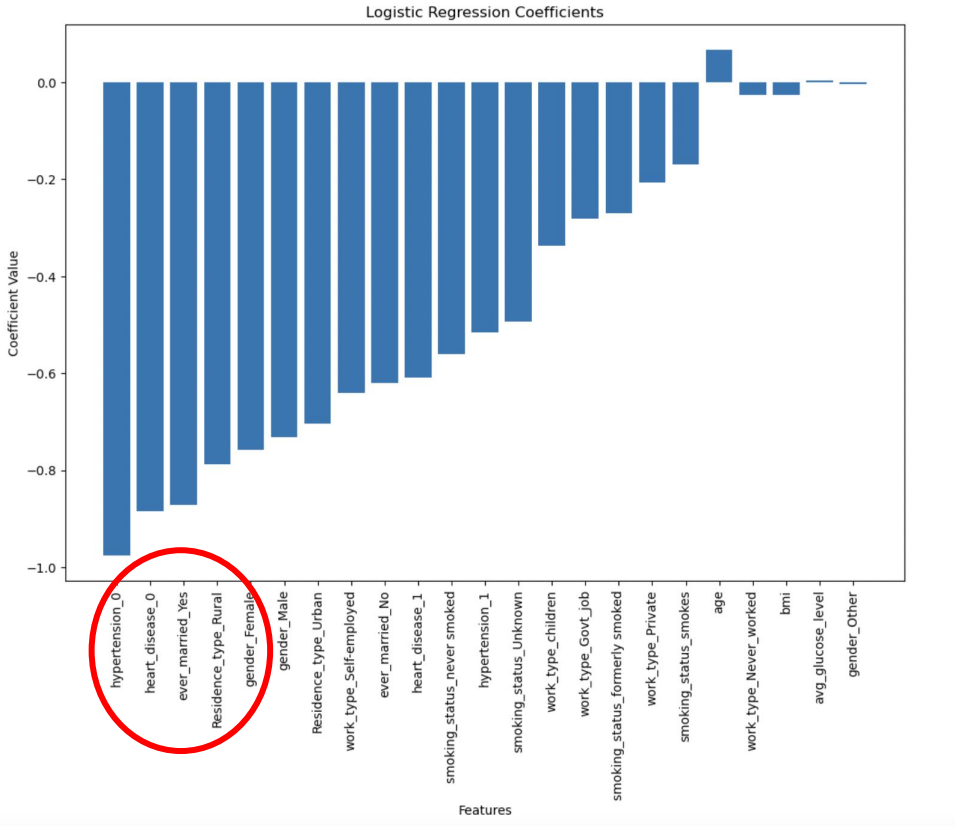


Insight - Random Forest

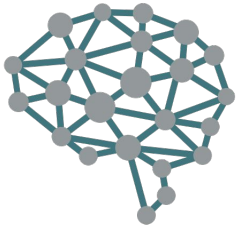
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9400 entries, 114 to 5109
Data columns (total 24 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   age                              9400 non-null   float64
 1   avg_glucose_level                9400 non-null   float64
 2   bmi                              9400 non-null   float64
 3   ever_married_No                  9400 non-null   float64
 4   ever_married_Yes                 9400 non-null   float64
 5   work_type_Govt_job               9400 non-null   float64
 6   work_type_Never_worked           9400 non-null   float64
 7   work_type_Private                9400 non-null   float64
 8   work_type_Self-employed          9400 non-null   float64
 9   work_type_children               9400 non-null   float64
 10  Residence_type_Rural             9400 non-null   float64
 11  Residence_type_Urban             9400 non-null   float64
 12  smoking_status_Unknown           9400 non-null   float64
 13  smoking_status_formerly smoked   9400 non-null   float64
 14  smoking_status_never smoked      9400 non-null   float64
 15  smoking_status_smokes            9400 non-null   float64
 16  gender_Female                    9400 non-null   float64
 17  gender_Male                      9400 non-null   float64
 18  gender_Other                     9400 non-null   float64
 19  hypertension_0                   9400 non-null   float64
 20  hypertension_1                   9400 non-null   float64
 21  heart_disease_0                  9400 non-null   float64
 22  heart_disease_1                  9400 non-null   float64
 23  stroke                           9400 non-null   float64
dtypes: float64(24)
memory usage: 1.8 MB
```



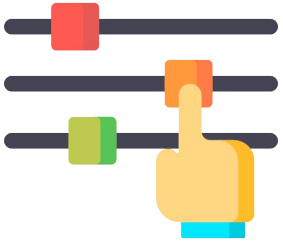
Logistic Regression



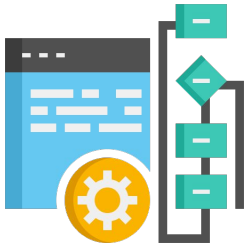
Artificial Neural Network (ANN)



Architecture



Hyperparameters

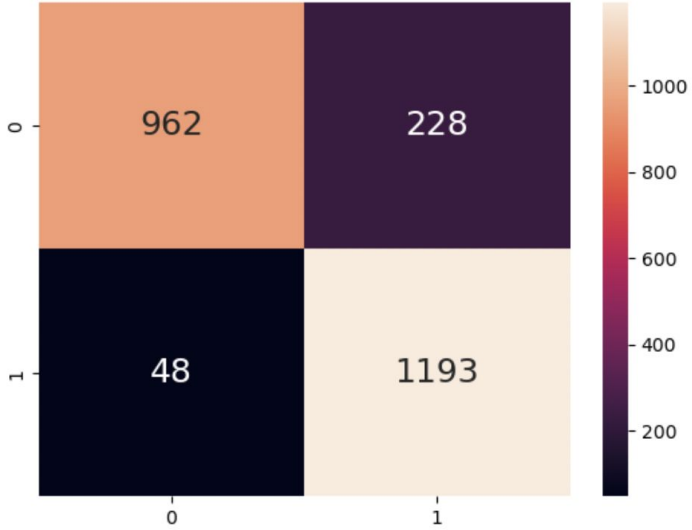


Learning Algorithms

Artificial Neural Network (ANN)

```
MLPClassifier(hidden_layer_sizes=(10, 10, 10), max_iter=1000)
```

	precision	recall	f1-score	support
0	0.95	0.81	0.87	1190
1	0.84	0.96	0.90	1241
accuracy			0.89	2431
macro avg	0.90	0.88	0.89	2431
weighted avg	0.89	0.89	0.89	2431

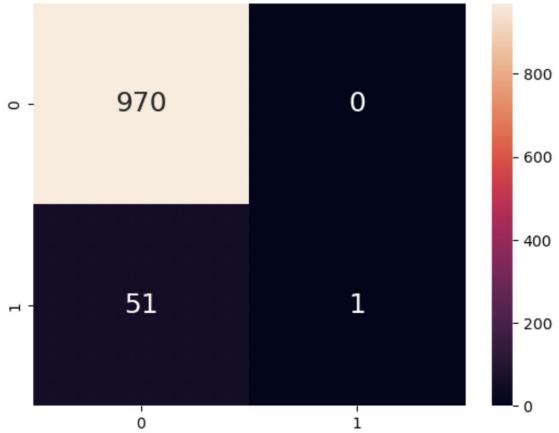


Observations:

1. Relatively high accuracy and F1 score
2. Able to be refined further by increasing layer sizes and iterations at the cost of time and computational cost
3. One drawback → Finding the right parameters has to be done iteratively

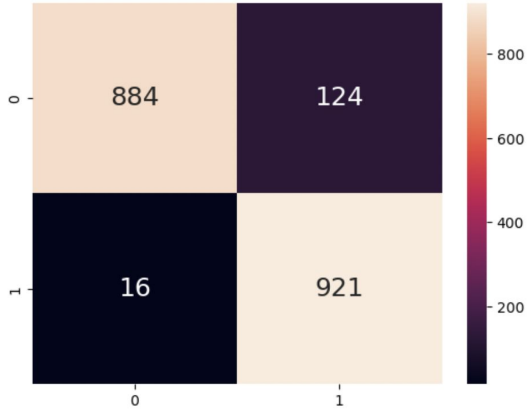
Models Evaluation

```
Test Data
Accuracy  :       0.9500978473581213
F1 score:         0.03773584905660378

TPR Test :        0.019230769230769232
TNR Test :        1.0

FPR Test :        0.0
FNR Test :        0.9807692307692307
```



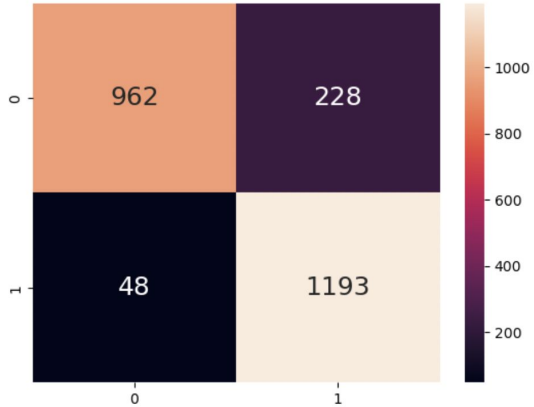
Logistic Regression

```
Test Data
Accuracy  :       0.9280205655526992

TPR Test :        0.9829242262540021
TNR Test :        0.876984126984127

FPR Test :        0.12301587301587301
FNR Test :        0.017075773745997867
F1 score :        0.929364278506559
```



Random Forest

	precision	recall	f1-score	support
0	0.95	0.81	0.87	1190
1	0.84	0.96	0.90	1241
accuracy			0.89	2431
macro avg	0.90	0.88	0.89	2431
weighted avg	0.89	0.89	0.89	2431



Artificial Neural Network

High Accuracy and F1 score, with the lowest FNR which is crucial in our problem statement.

Findings & Thoughts

With the high accuracy and f1 scores coupled with low FNR, we achieved our goal of predicting stroke

- Several unconventional data that can be used by the health industry for better prediction of stroke such as marriage status, work type & residence type.
- Our dataset does not include some variables that can be of significance when predicting stroke such as family stroke history, hours exercise per week etc.

Moving on, instead of predicting whether a person might get stroke or not, we could try to determine the probability of a person getting stroke…

Thank you!