SDSC 2102 group project
Topic: Strokes Prediction

Group members:
Chan Yuk Yee 56230549
Choi Chun Fai 56222863
Cheung Hoi Pang 56262793
Ho Hoi Kit 56229409

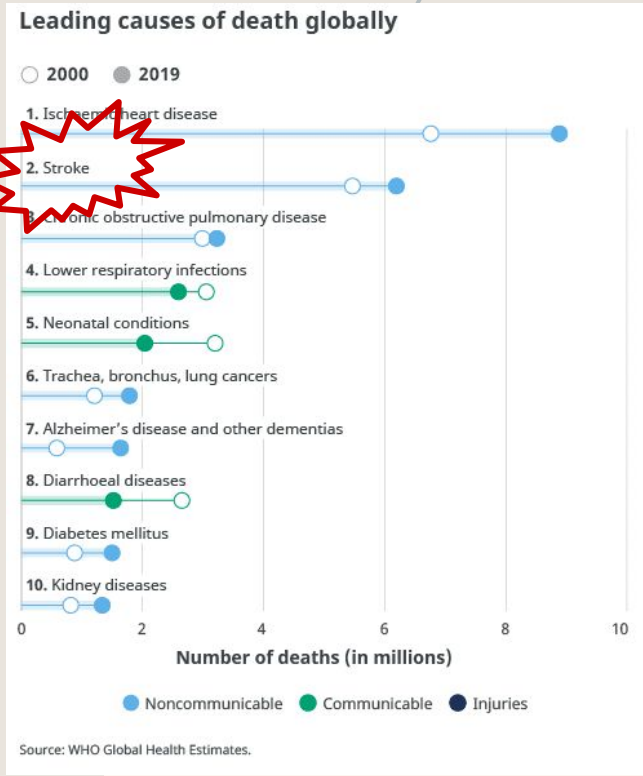
Date : 14/04/2022

1

Introduction

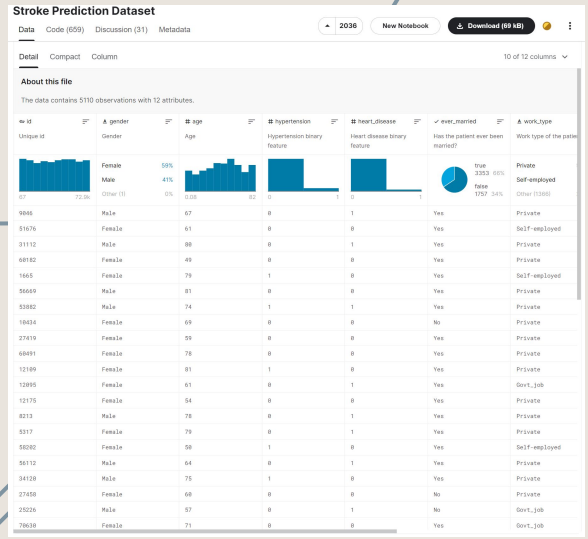
Stroke is a medical emergency !

- Stroke is the **number 2 leading cause of death** globally in 2019
- **Every 3.5 minutes** someone dies because of a stroke
- Complications includes:
  - *Difficulty talking or swallowing*
  - *memory loss or thinking difficulties*
  - *Loss of muscle movement*
  - *…..*



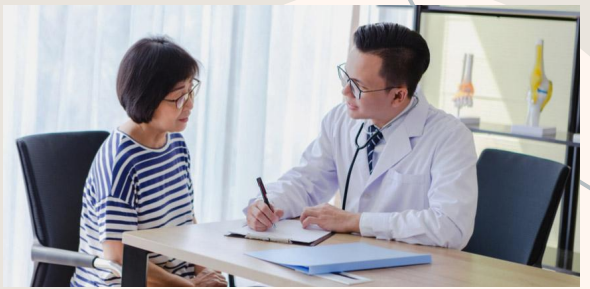
Objective

Patient's data



Analysis and Predict

Feature:
Age
Gender
Heart disease
Work type
Bmi
Average Glucose level
....



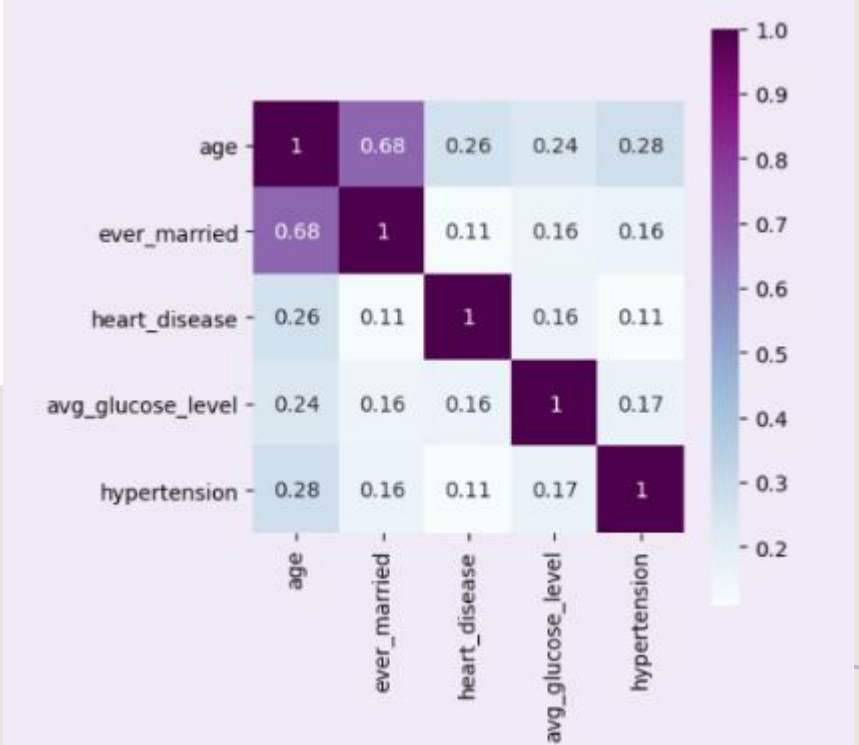
You have **stroke**
OR
You are **healthy**

2 Data Visualization

5 factor for stroke?

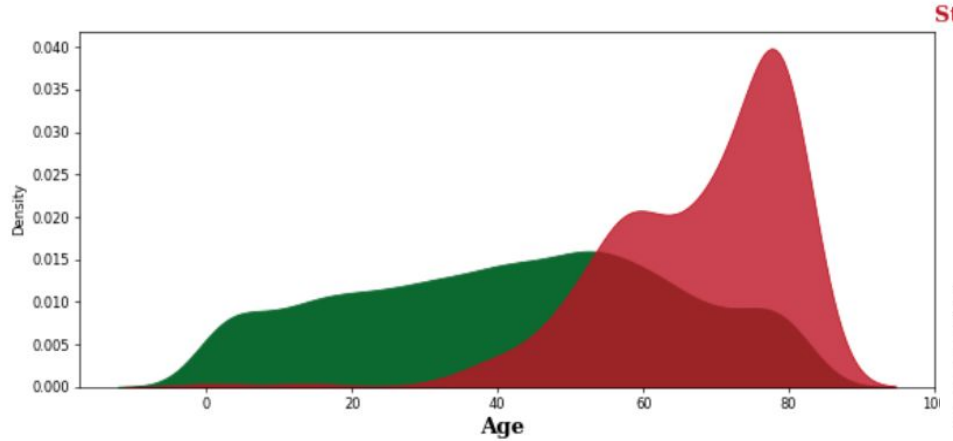
- Age
- Married or not?
- Heart disease
- Average glucose level
- hypertension

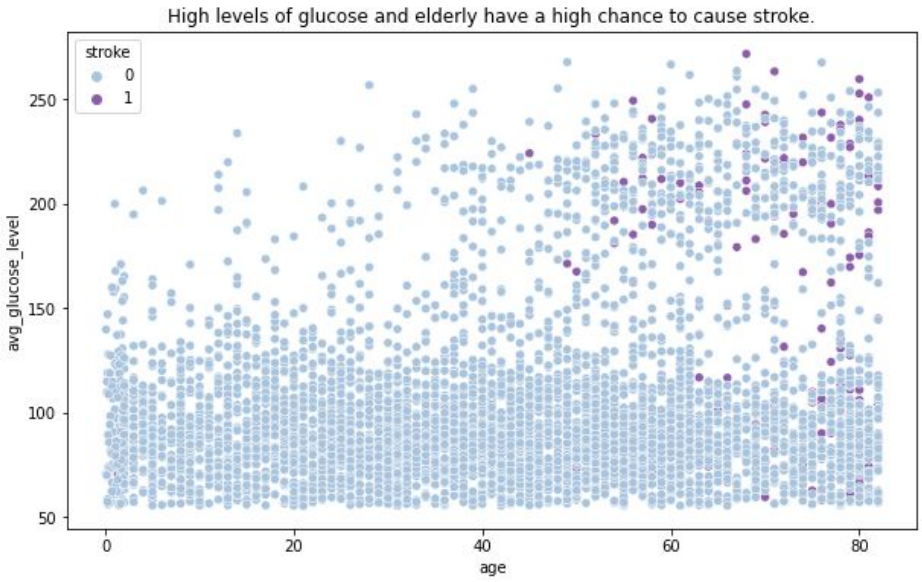
However, the state of being married increases with age. Hence, the feature of **married or not is an irrelevance** in the stroke prediction.



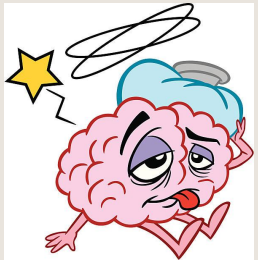
Kdeplot

The relationship between age and probability of having a stroke.



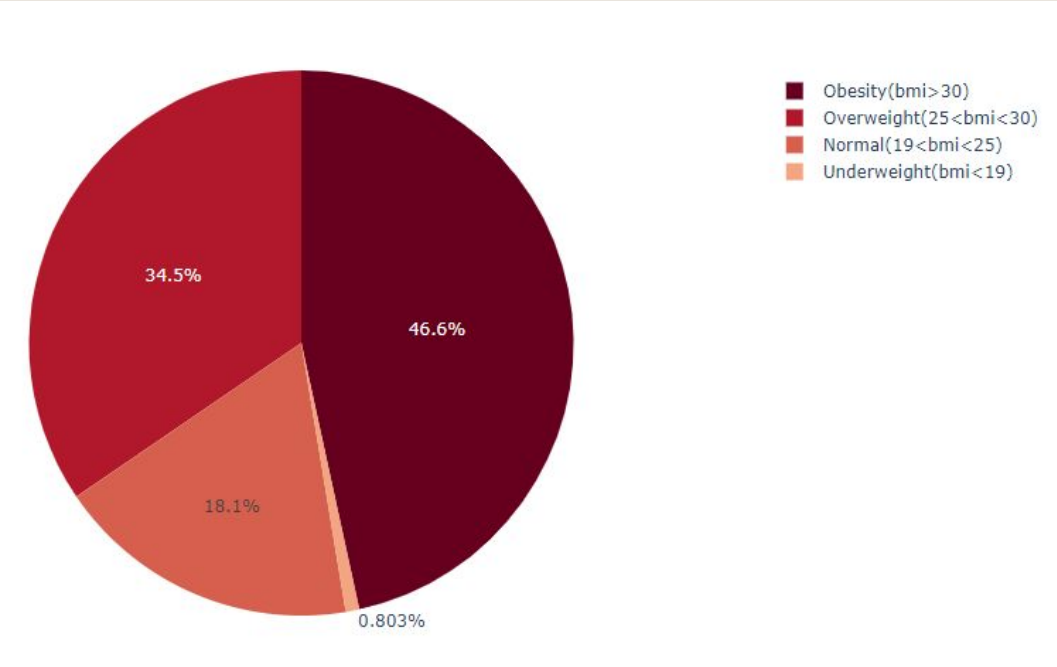


From the above graphs, we discovered those that **aged over 60** and **blood glucose levels larger than 150** are more easily to have a stroke.



With heart disease
Average glucose level
Hypertension

Issue of Obesity



For those who have a stroke, **Over 81.1% of them are fat** (overweight or obesity)

It implies **high bmi** has been linked to stroke.

3

Data Preprocessing

Handling missing data

****MICE Imputation****

Feature encoding

****LabelEncoder****

Splitting dataset

****Train-test Split****

Feature normalization

****StandardScaler****

Upsampling data

****SMOTE****

The data entered in the dataset is **incomplete**.

Machine learning models can only use numerical values. Therefore, it is necessary to **convert the categorical values** of the relevant features into numerical values.

Splitting the dataset can also be important to **detect whether the model is underfitting or overfitting**.

Feature normalization allows for **faster convergence** on machine learning.

The challenge of **imbalanced datasets** is that most machine learning techniques will ignore, and have **poor performance on the minority class**.

4
Data Modeling

Classification Model Selection and Evaluation :

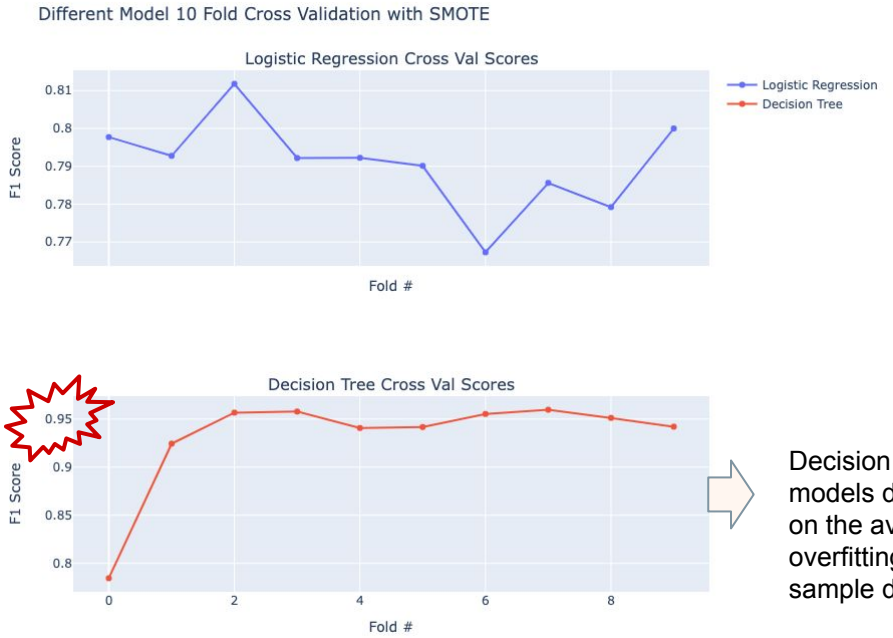
$$F\text{-}score = \frac{2Precision * Recall}{Precision + Recall}$$

Logistic Regression

```
Logistic Regreesion :
[[1070  388]
 [ 287 1171]]
Accuracy Score:  0.7685185185185185

Standard Deviation: 3.86 %

ROC AUC Score: 0.77

Precision: 0.75

Recall: 0.80

F1: 0.78
```

Decision Tree

```
Decision Tree :
[[1381   77]
 [ 363 1095]]
Accuracy Score:  0.8491083676268861

Standard Deviation: 3.86 %

ROC AUC Score: 0.85

Precision: 0.93

Recall: 0.75

F1: 0.83
```

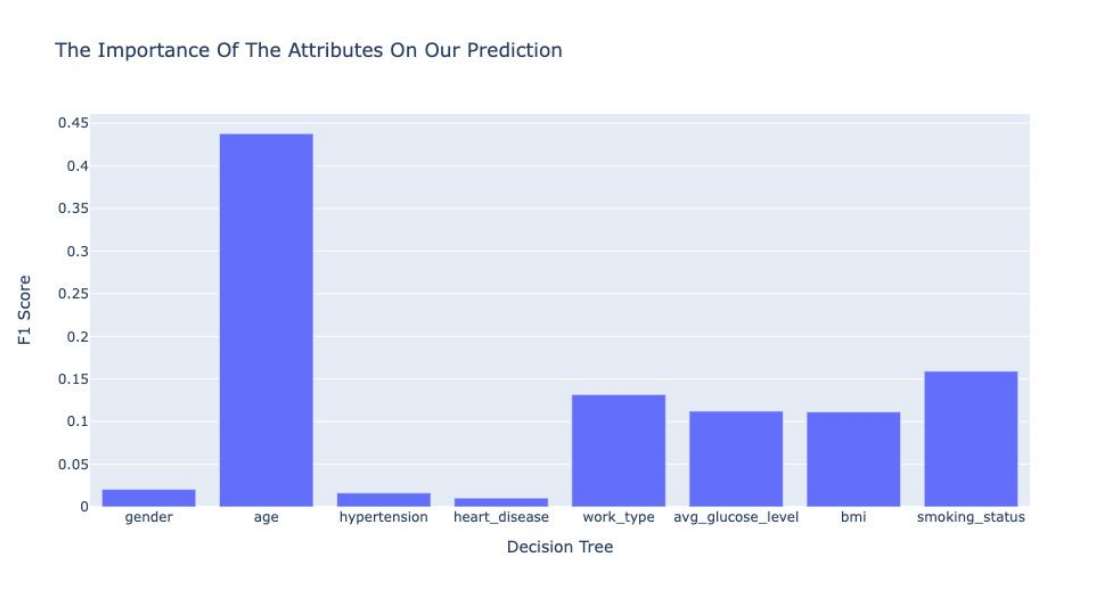


Decision tree models did the best on the average of overfitting the sample data

4

Data Modeling

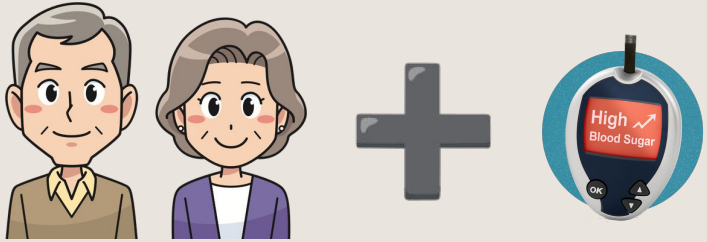
Decision Tree Importance Feature :



We saw that an individual's **age** was the most important predictors of stroke-susceptible individuals. The second important predictors is individual's **work type, average glucose level, bmi ,and smoking status.**

5

Results Interpretation, Discussion & Conclusion



➔ Higher chance to cause stroke

Over 81.1% stroke patients are **overweight**

➔ *Strongly positive correlation between **high bmi** and **stroke***

<u>Best Classification Model</u>

➢ **Decision Tree**

***#1** Important*

↘ *age*

#2 *work type, average glucose level, bmi ,and smoking status.*

6

Insights

To prevent **Blood Sugar Spikes, Overweight, Stroke**

- ★ Eat fewer carbs (Carbohydrates), esp. refined carbs (e.g. rice)
- ★ Reduce your sugar intake
- ★ Exercise more
- ★ Eat more fiber, e.g. vegetable, fruits, oat
- ★ Drink more water
- ★ Drop your cigarette, don’t drink alcohol

Thanks