

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

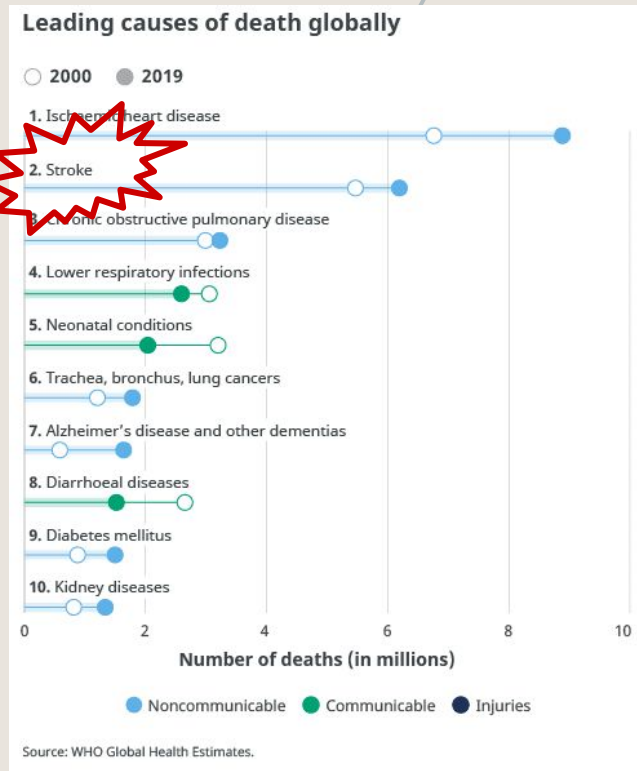
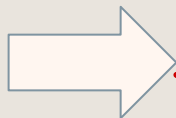


1

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



You have stroke
OR
You are healthy

Analysis and Predict

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



Stroke Prediction Dataset

Data Code (859) Discussion (31) Metadata

2038 New Notebook Download (8 kb)

Detail Compact Column 10 of 12 columns

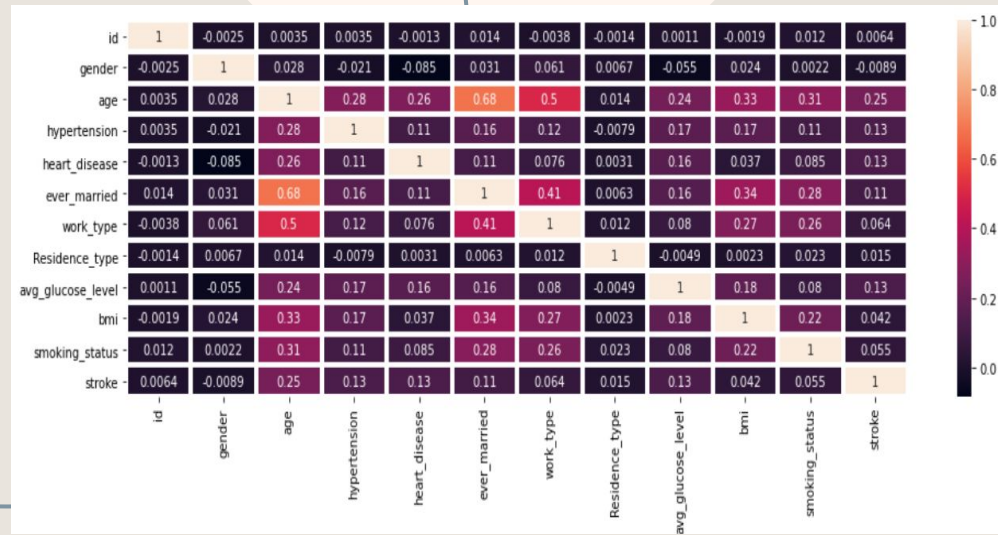
About this file

The data contains 5110 observations with 12 attributes.

id	gender	age	hypertension	heart_disease	ever_married	work_type
Unique id	Gender	Age	Hypertension binary feature	Heart disease binary feature	Has the patient ever been married?	Work type of the patient
0-12,238	Female 59% Male 41% Other (3) 0%	0-80	0-1	0-1	true 2383 60% false 1727 34%	Private Self-employed
9886	Male	67	0	1	Yes	Private
51676	Female	61	0	0	Yes	Self-employed
31112	Male	88	0	1	Yes	Private
68182	Female	49	0	0	Yes	Private
1665	Female	79	1	0	Yes	Self-employed
56669	Male	81	0	0	Yes	Private
53882	Male	74	1	1	Yes	Private
18084	Female	69	0	0	No	Private
27419	Female	59	0	0	Yes	Private
68891	Female	78	0	0	Yes	Private
12189	Female	81	1	0	Yes	Private
12895	Female	61	0	1	Yes	Govt_job
12175	Female	54	0	0	Yes	Private
8213	Male	78	0	1	Yes	Private
5317	Female	79	0	1	Yes	Private
58282	Female	58	1	0	Yes	Self-employed
56112	Male	64	0	1	Yes	Private
34128	Male	75	1	0	Yes	Private
27458	Female	68	0	0	No	Private
25026	Male	87	0	1	No	Govt_job
78638	Female	71	0	0	Yes	Govt_job

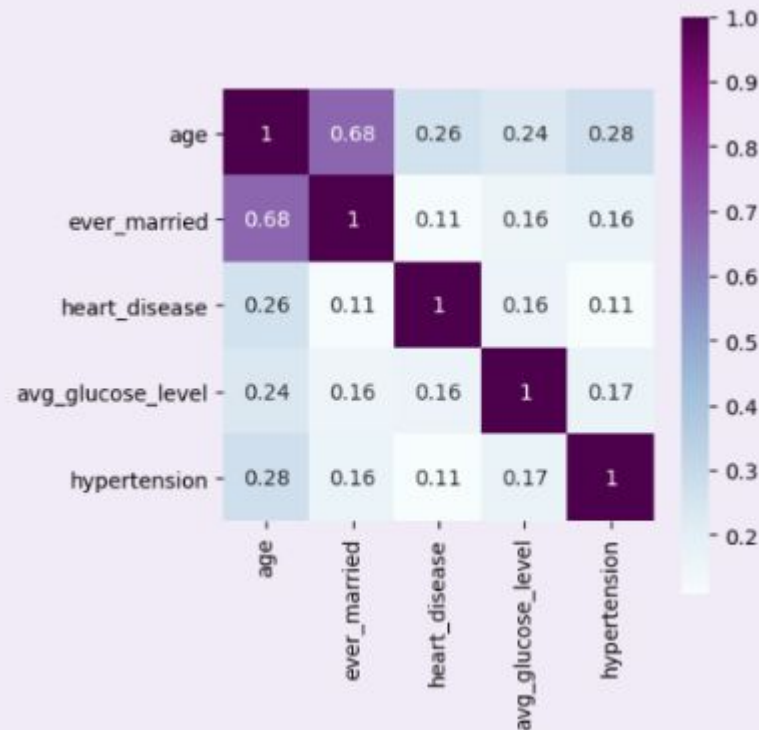
2 Data Visualization

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



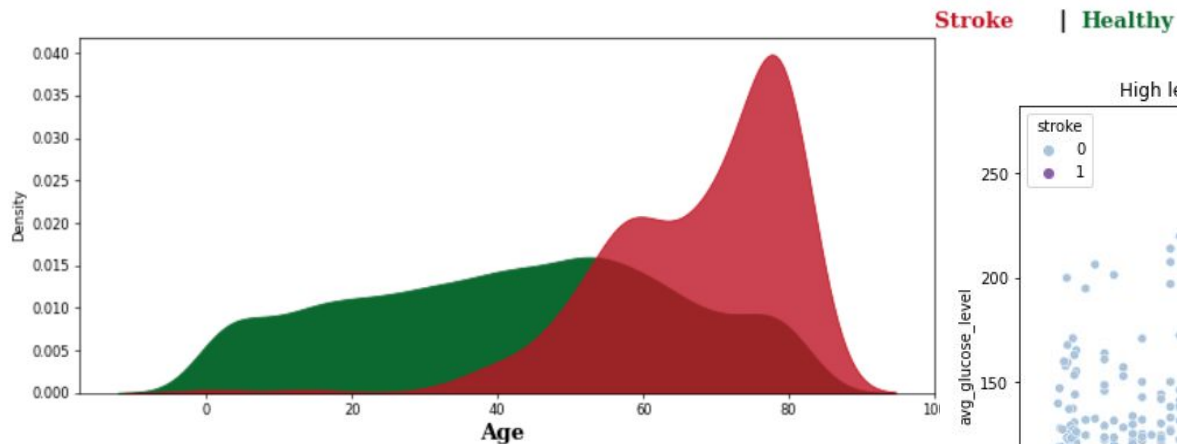
5 factor for stroke?

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

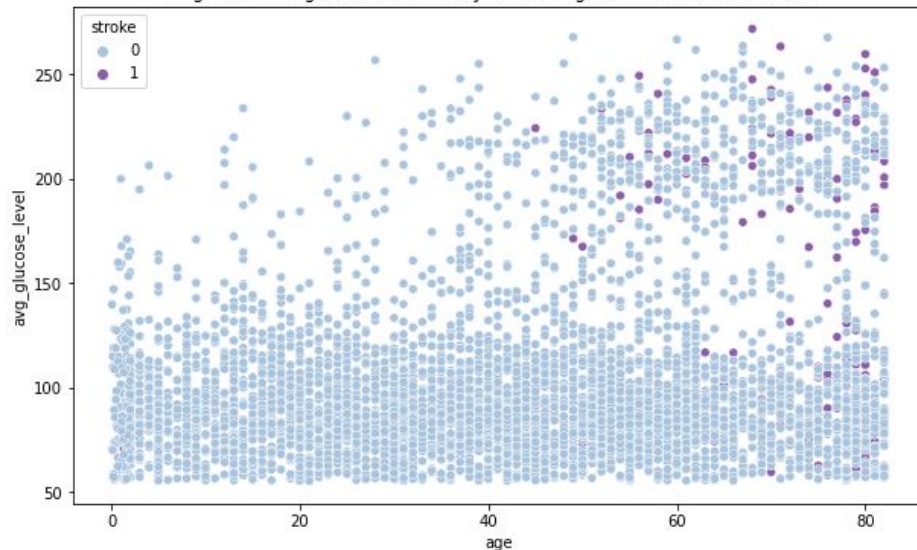


Kdeplot

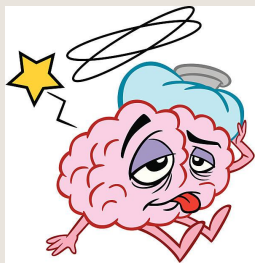
The relationship between age and probability of having a stroke.



High levels of glucose and elderly have a high chance to cause 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.

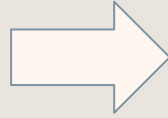




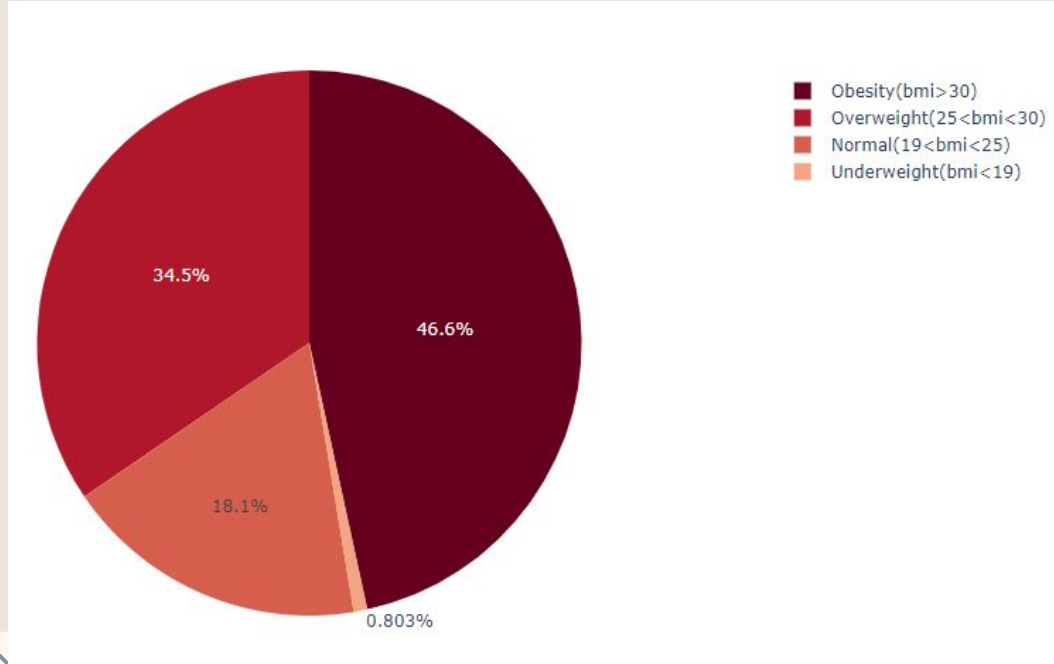
With heart disease

Average glucose level

Hypertension



Issue of Obesity



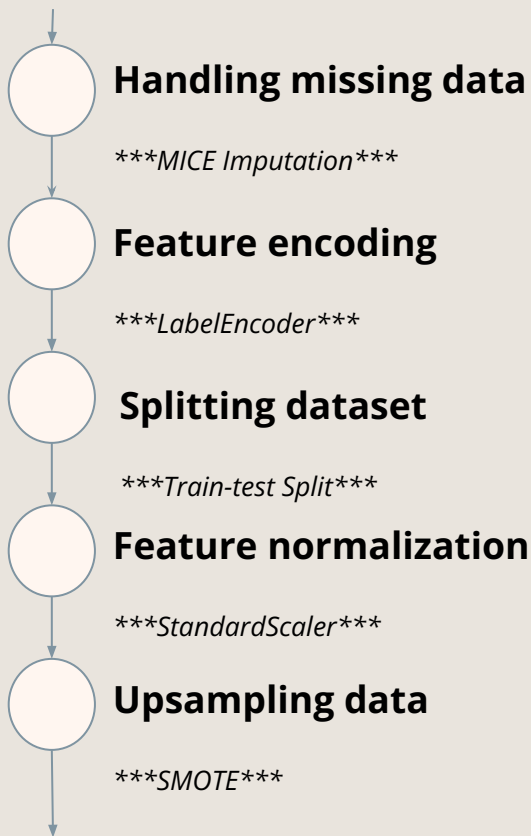
For those who have a stroke, **Over 81.1% of them are fat** (overweight or obesity)



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

3

Data Preprocessing



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 Regression :
[[1070 388]
[287 1171]]
Accuracy Score: 0.7685185185185185

Standard Deviation: 3.86 %

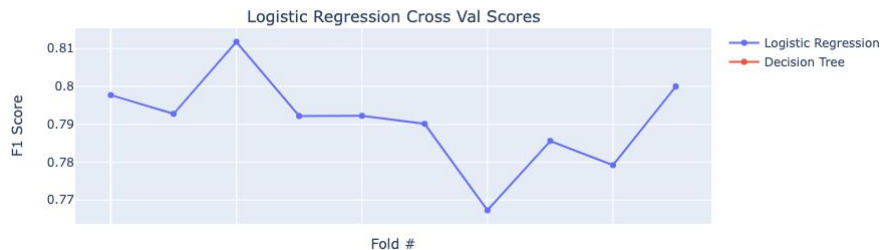
ROC AUC Score: 0.77

Precision: 0.75

Recall: 0.80

F1: 0.78

Different Model 10 Fold Cross Validation with SMOTE



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

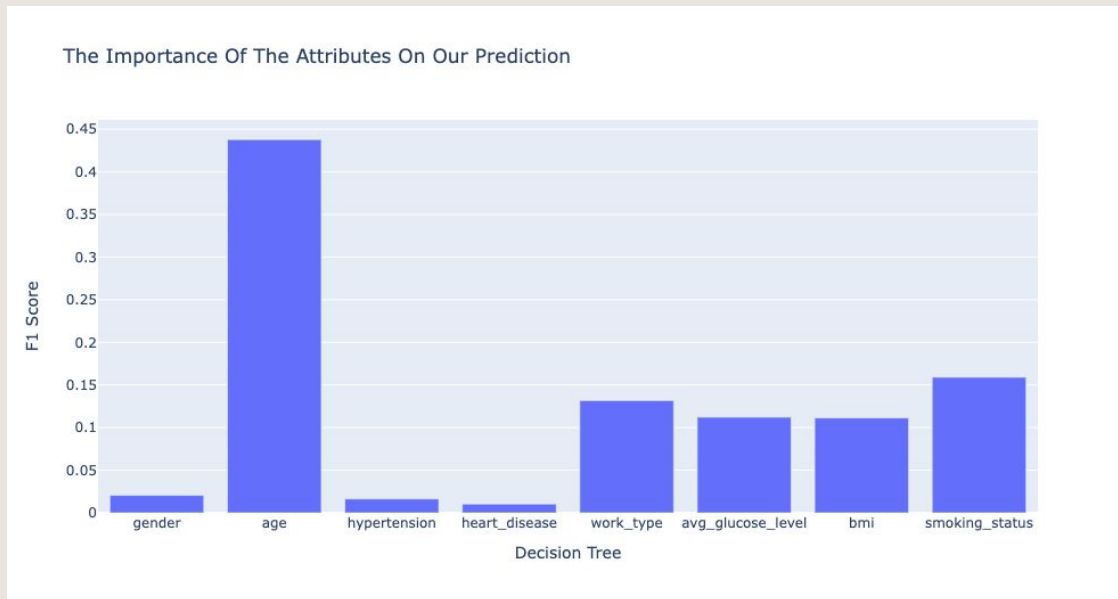


Decision tree models did the best on the average of overfitting the sample data

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



Over 81.1% stroke patients are **overweight**

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



Best Classification Model

➤ **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



The background is a solid light gray. In the top-left corner, there is a blue zigzag line. In the top-right corner, there is a thin blue arc of a circle. In the bottom-left corner, there is a small white circle and a thin blue arc of a circle. In the bottom-right corner, there is a large white circle.

Thanks