Stroke Prediction Based on Lifestyle

Presented by:
Soh Han Yu Brian (U2223002H)
Ng Yee Shem (U2222651H)
Seah Angelo Michael (U2220896J)

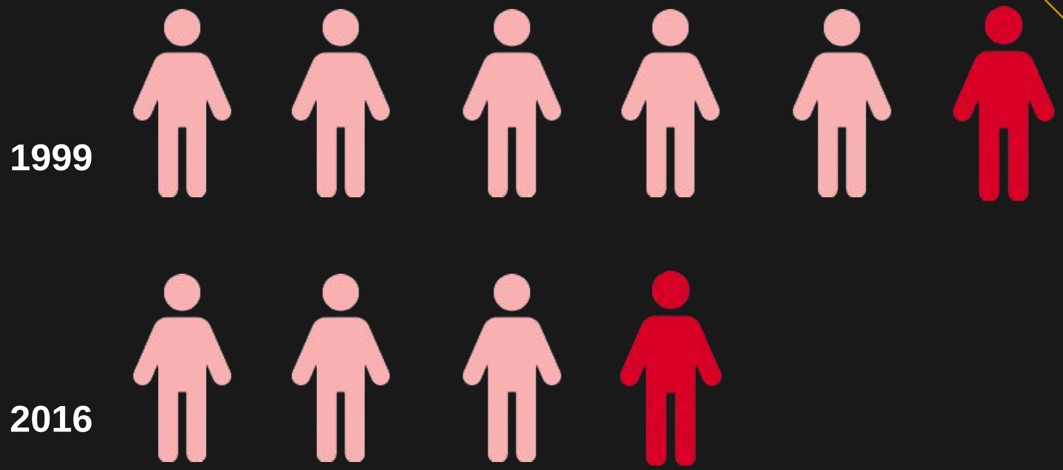
Stroke Prevalence

2nd

leading cause of death

3rd

leading cause of disability



Sample
COLLECTION

Practical
MOTIVATION

Stroke Prevalence

70%

Stroke incidence

43%

Casualty

102%

Stroke prevalence

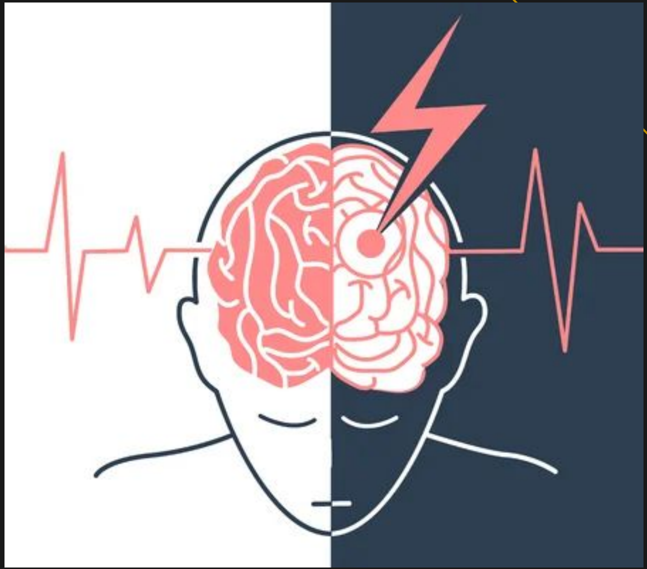
What is Stroke?

Stroke

When blood supply to part of the brain is blocked or when blood vessels in the brain burst.

Effects

- Memory loss
- Emotional problem
- Paralysis



Problem definition

CAN WE USE ONE'S LIFESTYLE TO PREDICT IF THEY WILL HAVE A STROKE?

Sample
COLLECTION

Practical
MOTIVATION

Brief Overview

Size: 5110
Variables: 11

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

Attribute information

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if doesn't have hypertension, 1 if have hypertension
5. heart_disease: 0 if doesn't have any heart diseases, 1 if have a heart disease
6. ever_married: "No" or "Yes"
7. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. Residence_type: "Rural" or "Urban"
9. avg_glucose_level: average glucose level in blood
10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12. stroke: 1 if the patient had a stroke or 0 if not

Data Preparation

Gender:

- Male: 2115
- Female: 2991
- Other: 1

BMI:

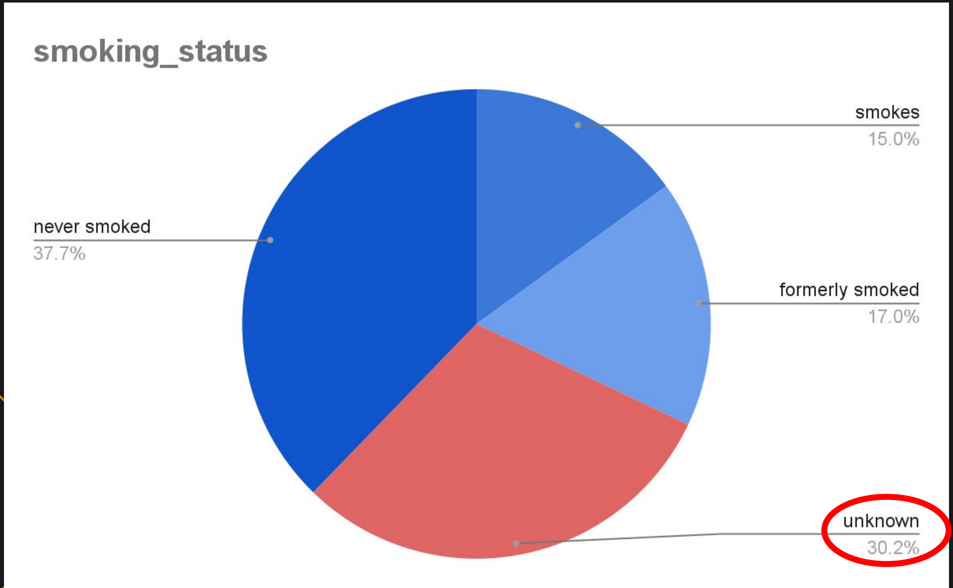
- NULL: 201

Rows of data dropped: **202**

```
df.isnull().sum()
```

```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

Data Preparation



- Statistically significant
- Categorical
- Can be treated as a unique value

Data Preparation

Value count:

DIABETES

>126 mg/dl

914

PREDIABETES

100-125 mg/dl

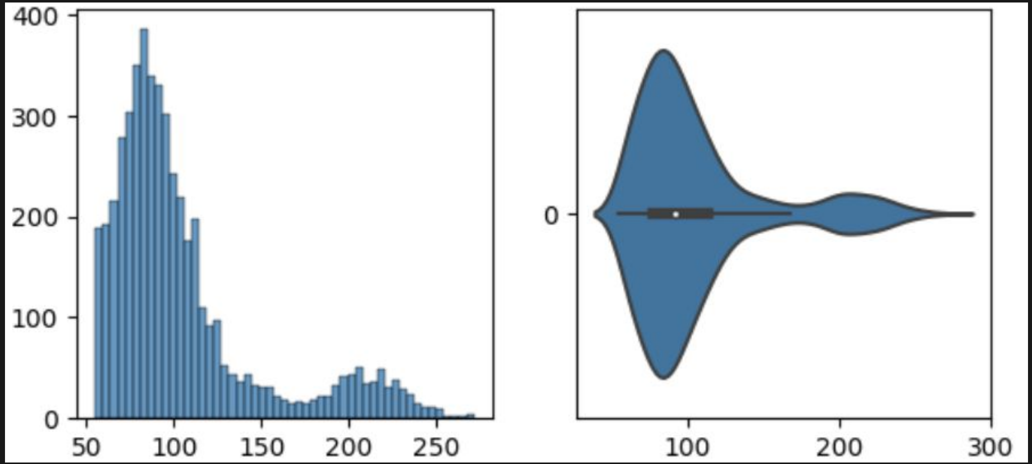
1022

NORMAL

<99 mg/dl

2972

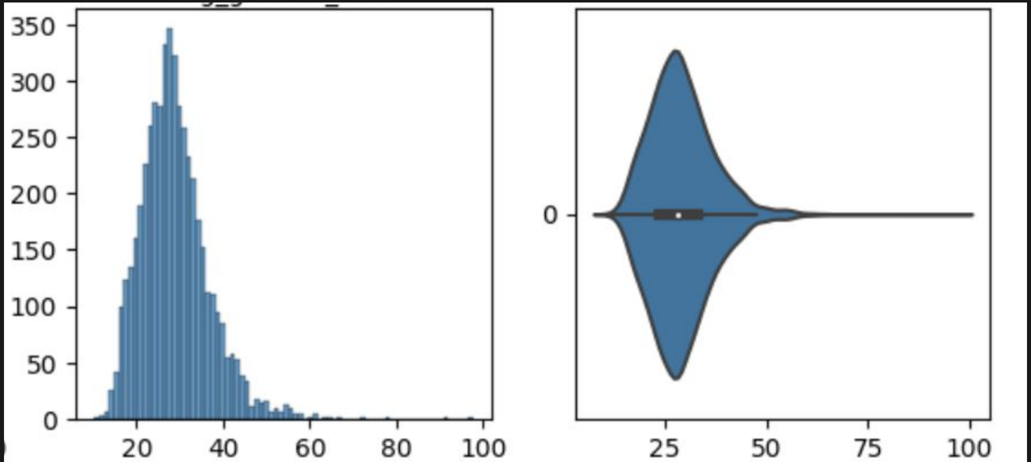
Univariate Visualisation *(continuous data)*



avg_glucose_level

Analytic VISUALIZATION Pattern RECOGNITION

Univariate Visualisation *(continuous data)*

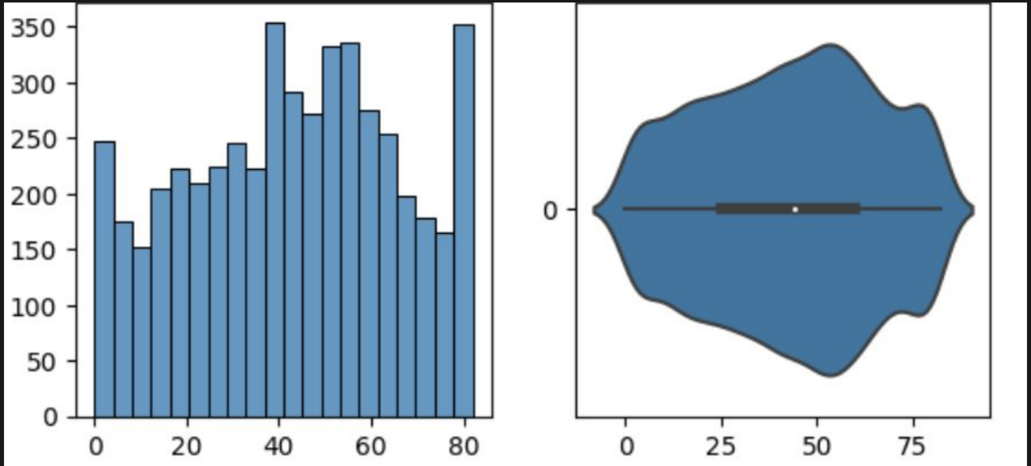


BMI

Analytic
VISUALIZATION

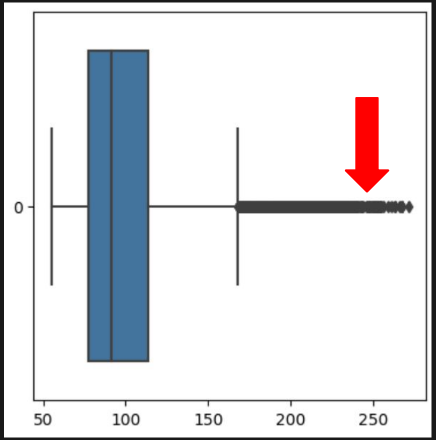
Pattern
RECOGNITION

Univariate Visualisation *(continuous data)*

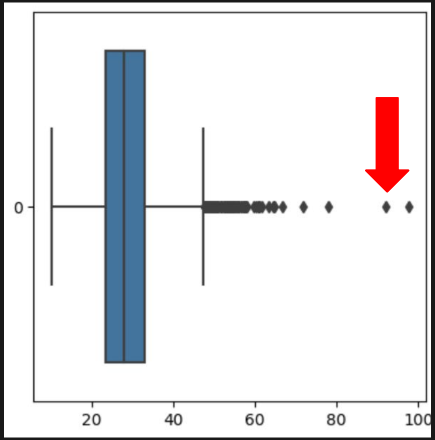


Age

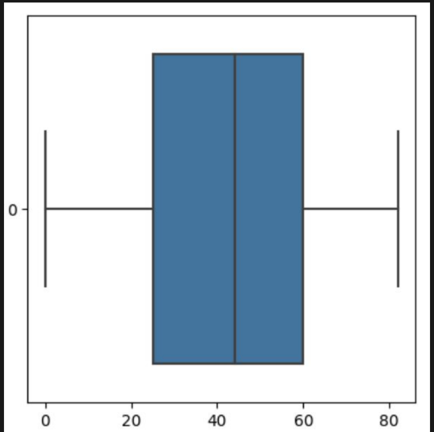
Univariate Visualisation *(continuous data)*



avg_glucose_level



BMI



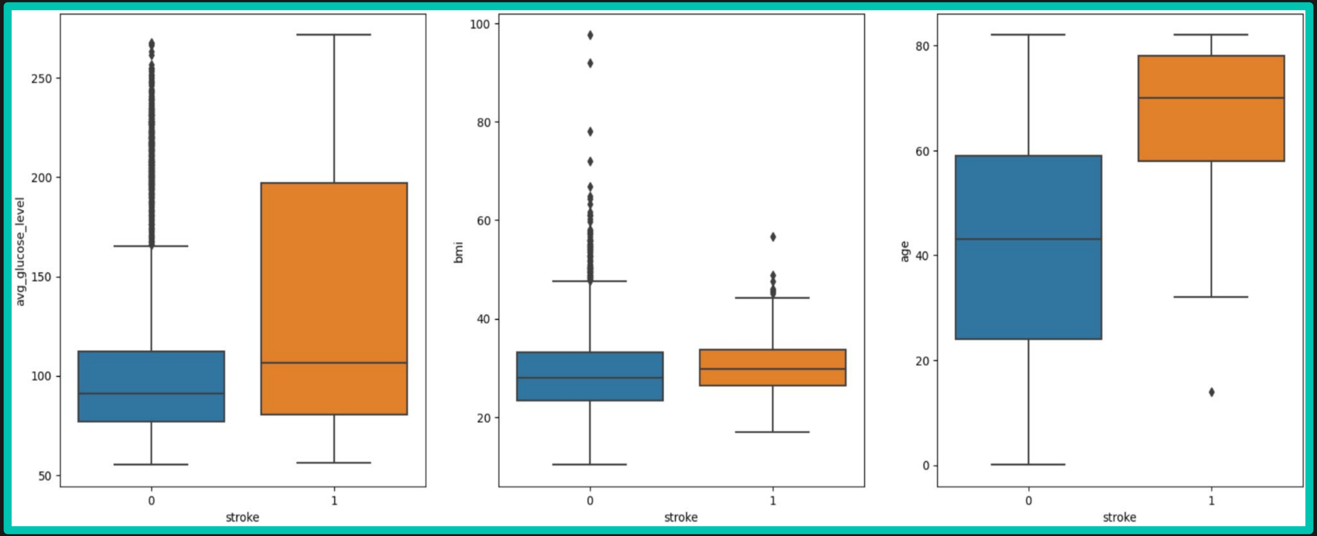
age

Boxplot helps visualise distribution and outliers

Analytic VISUALIZATION

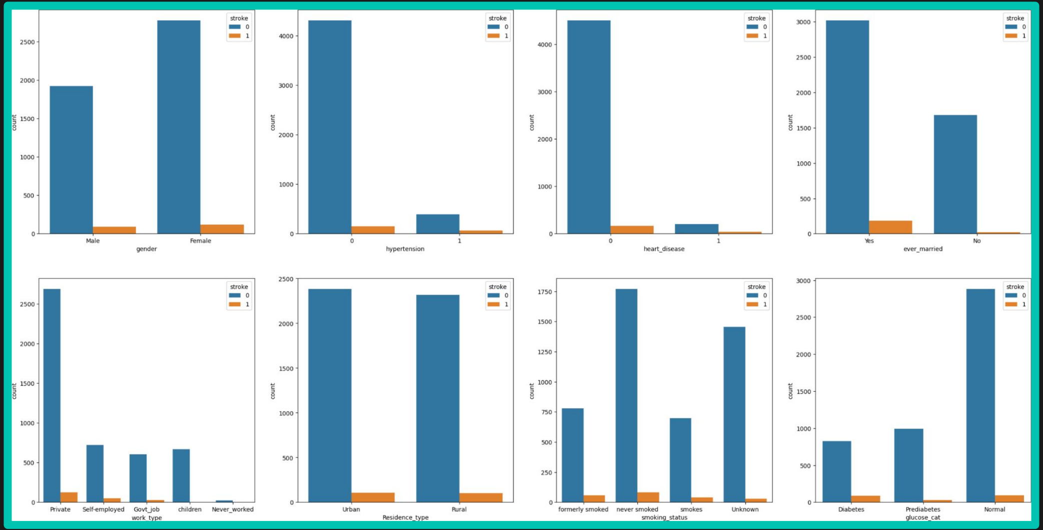
Pattern RECOGNITION

Bivariate Visualisation *(continuous data)*



Continuous data vs Stroke (BoxPlot)

Bivariate Visualisation *(categorical data)*



Categorical data vs Stroke (Countplot)

Analytic VISUALIZATION Pattern RECOGNITION

Updated attribute information

1. gender: "Male" or "Female"
2. age: age of the patient
3. hypertension: 0 if doesn't have hypertension, 1 if have hypertension
4. heart_disease: 0 if doesn't have any heart diseases, 1 if have a heart disease
5. ever_married: "No" or "Yes"
6. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
7. Residence_type: "Rural" or "Urban"
8. avg_glucose_level: average glucose level in blood
9. glucose_cat: “Normal”,”Prediabetes” or “Diabetes” (newly added)
10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12. stroke: 1 if the patient had a stroke or 0 if not

Data cleaning and EDA summary

- Our data is very imbalanced
- Non-stroke data far outweighs stroke data
- Unable to employ feature selection as we cant tell which features are helpful

Analytic VISUALIZATION Pattern RECOGNITION

Model Preparation

Represent categorical string data in integer form

Removed Residential_type as it doesn't affect data

```
new_df=pd.DataFrame.copy(df)
new_df.gender[new_df.gender == 'Male'] = 1
new_df.gender[new_df.gender == 'Female'] = 0
new_df.ever_married[new_df.ever_married == 'Yes'] = 1
new_df.ever_married[new_df.ever_married == 'No'] = 0
new_df.work_type[new_df.work_type == 'children'] = 0
new_df.work_type[new_df.work_type == 'Private'] = 1
new_df.work_type[new_df.work_type == 'Self-employed'] = 2
new_df.work_type[new_df.work_type == 'Govt_job'] = 3
new_df.work_type[new_df.work_type == 'Never_worked'] = 4
new_df.Residence_type[new_df.Residence_type == 'Urban'] = 1
new_df.Residence_type[new_df.Residence_type == 'Rural'] = 0
new_df.smoking_status[new_df.smoking_status == 'never smoked'] = 0
new_df.smoking_status[new_df.smoking_status == 'formerly smoked'] = 1
new_df.smoking_status[new_df.smoking_status == 'smokes'] = 2
new_df.smoking_status[new_df.smoking_status == 'Unknown'] = 3
new_df.glucose_cat[new_df.glucose_cat == 'Normal'] = 0
new_df.glucose_cat[new_df.glucose_cat == 'Prediabetes'] = 1
new_df.glucose_cat[new_df.glucose_cat == 'Diabetes'] = 2
```

Model Preparation

Random_state required for our data set

```
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=42)
```

X is independent variable and Y is dependent variable

Analytic VISUALIZATION

Pattern RECOGNITION

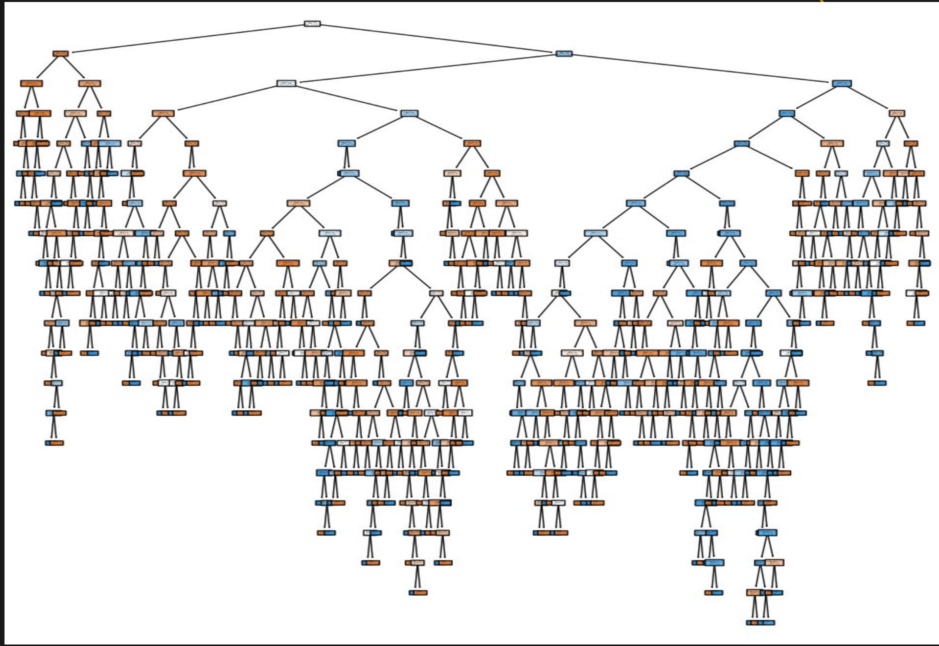
Model Preparation (SMOTE)

209:4699 (total) $\xrightarrow{\text{Train-test split}}$ 167:3759 (train set) $\xrightarrow{\text{SMOTE}}$ 3770:3770 (train set with SMOTE)

SMOTE helps to make data set 50:50
by using the minority class
(non-stroke data)

Binary Classification Tree

Binary Classification with depth of 20 (addresses problem of overfitting)

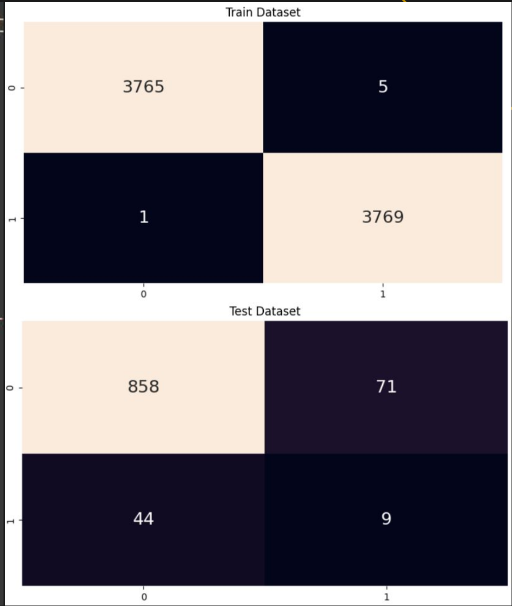


Algorithmic OPTIMIZATION Machine LEARNING

Analysis for Binary Classification Tree

Used the binary classification tree to predict stroke for the train and test dataset

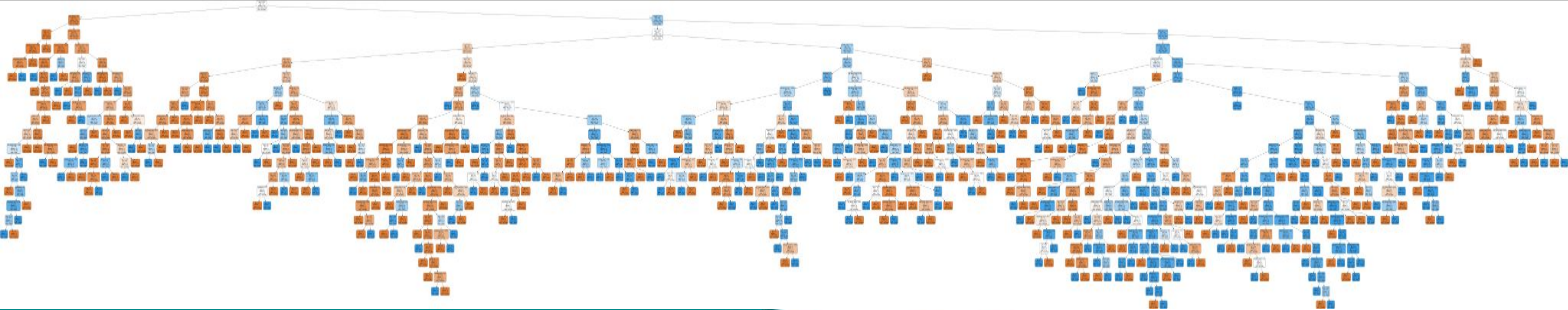
```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.999

True Positive Rate              : 1.000
False Negative Rate             : 0.000

True Negative Rate              : 0.999
False Positive Rate             : 0.001

---------------------------------------------
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.883

True Positive Rate              : 0.170
False Negative Rate             : 0.830

True Negative Rate              : 0.924
False Positive Rate             : 0.076
```



Information PRESENTATION Statistical INFERENCE

Random Forest Classification

- Multiple decision trees
- Does not overfit
- Keep classification accuracy high



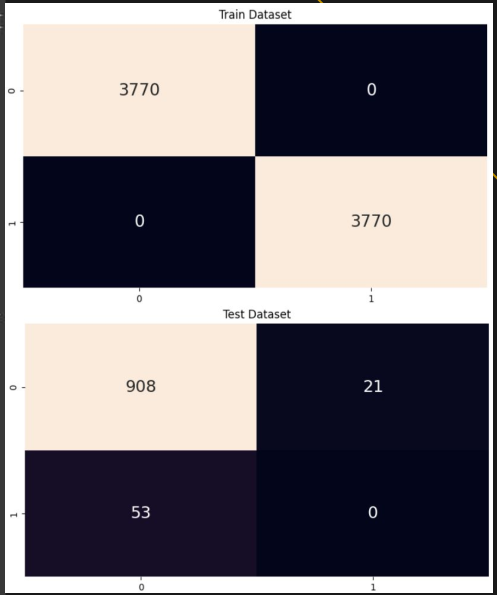
Algorithmic OPTIMIZATION

Machine LEARNING

Analysis for Random Forest Classification

Used the Random Forest Classification to predict stroke for the train and test dataset

```
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 1.000

True Positive Rate                : 1.000
False Negative Rate               : 0.000

True Negative Rate                : 1.000
False Positive Rate               : 0.000

---------------------------------------------
Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.925

True Positive Rate                : 0.000
False Negative Rate               : 1.000

True Negative Rate                : 0.977
False Positive Rate               : 0.023
```



Information PRESENTATION

Statistical INFERENCE

Hyper-parameter Tuning

Randomized Grid Search is a form of hyper-parameter optimization where hyper-parameters are randomly selected so that the search for the best hyper-parameters is less time consuming

Hyperparameters found to be the best:

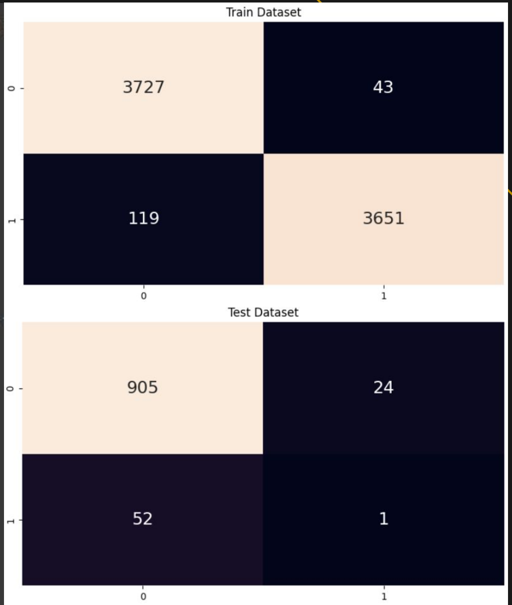
```
RandomForestClassifier(criterion='entropy', max_depth=83, max_features=None,
                       max_leaf_nodes=86, n_estimators=150, random_state=0)
```

Analysis for Grid Search

No real improvement on overall model accuracy.

However, True Positive Rate slightly improved compared to previous Random Forest Model

```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.979

True Positive Rate              : 0.968
False Negative Rate             : 0.032

True Negative Rate              : 0.989
False Positive Rate             : 0.011

---------------------------------------------
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.923

True Positive Rate              : 0.019
False Negative Rate             : 0.981

True Negative Rate              : 0.974
False Positive Rate             : 0.026
```



Information
PRESENTATION
Statistical
INFERENCE

Conclusion

From Random Forest Classification

Goodness of Fit of Model	Train Dataset
Classification Accuracy	: 1.000
True Positive Rate	: 1.000
False Negative Rate	: 0.000
True Negative Rate	: 1.000
False Positive Rate	: 0.000

Goodness of Fit of Model	Test Dataset
Classification Accuracy	: 0.925
True Positive Rate	: 0.000
False Negative Rate	: 1.000
True Negative Rate	: 0.977
False Positive Rate	: 0.023

Unsuccessful!

Classification accuracy high but true positive rate in test set were low

Dataset had no clear indications on which features could affect stroke

From Grid Search

Goodness of Fit of Model	Train Dataset
Classification Accuracy	: 0.979
True Positive Rate	: 0.968
False Negative Rate	: 0.032
True Negative Rate	: 0.989
False Positive Rate	: 0.011

Goodness of Fit of Model	Test Dataset
Classification Accuracy	: 0.923
True Positive Rate	: 0.019
False Negative Rate	: 0.981
True Negative Rate	: 0.974
False Positive Rate	: 0.026

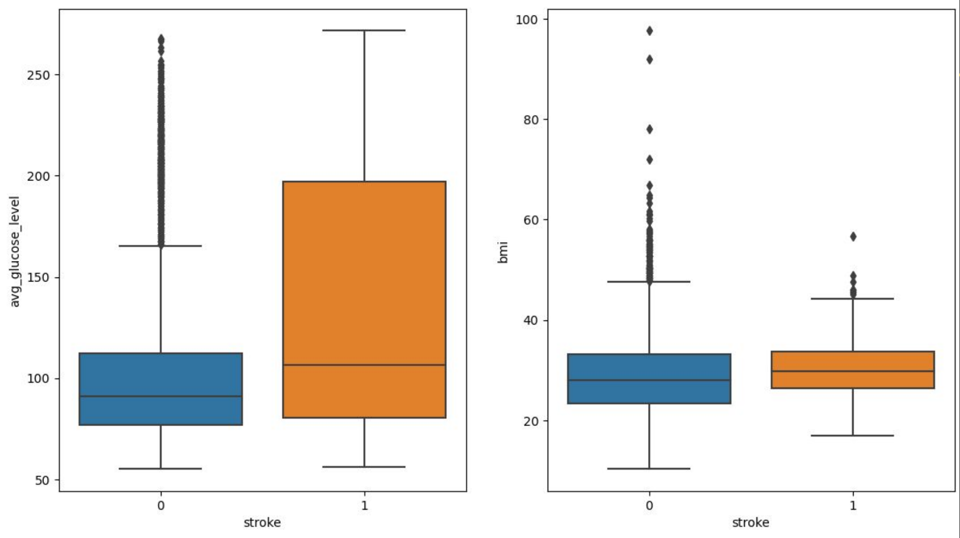
Ethical CONSIDERATION Intelligent DECISION

Conclusion

Extreme features still raise the chance that one may get a stroke

However, still difficult to predict accurately

Dataset did not contain preexisting conditions

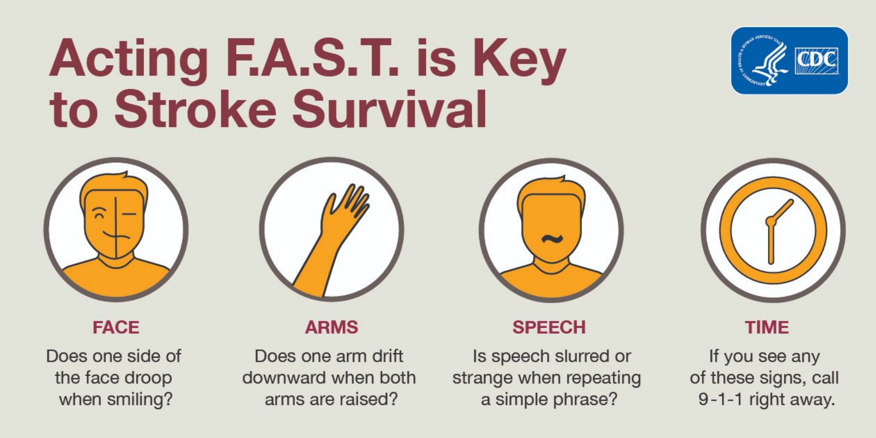


Ethical CONSIDERATION

Intelligent DECISION

Conclusion

Even we were unable to predict strokes based on one's lifestyle, it should still serve as a reminder of what we should look out for if we want to prevent stroke happening to ourselves.



Thank You

Image References:

https://www.nia.nih.gov/health/stroke

https://www.istockphoto.com/vector/human-brain-stroke-illustration-gm1296351746-389799325

https://www.123rf.com/photo_88598133_concept-of-the-disease-is-a-stroke-life-after-before-and-after-a-strokein-the-form-of-a-silhouette.html?vti=mcg1hmgxnxyofeaqjk-2-14

Additional References:

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

https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022#:~:text=Stroke%20is%20the%20leading%20cause,a%20stroke%20in%20their%20lifetime.

https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=A%20fasting%20blood%20sugar%20level,higher%20indicates%20you%20have%20diabetes

https://medium.com/analytics-vidhya/ways-to-handle-categorical-column-missing-data-its-implementations-15dc4a56893

https://www.mygreatlearning.com/blog/cross-validation/

https://www.pluralsight.com/guides/cleaning-up-data-from-outliers

https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

https://scikit-learn.org/stable/modules/tree.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/