



STROKE PREDICTION DATASET

By Sinothile Blose

WELCOME



Sinothile Blose

Future Data
Scientist



AGENDA

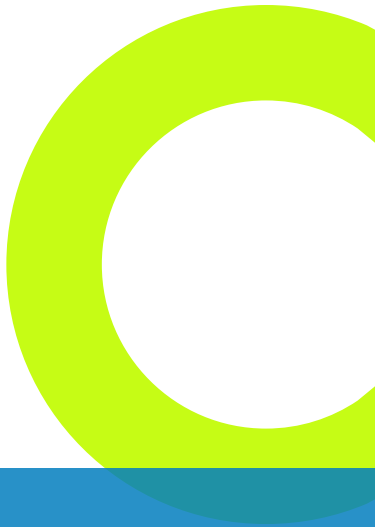
- **The stakeholder, and the problem I am solving for them**
 - **Brief introduction to data**
 - **Findings to Stake holders Via 2 Visuals**
 - **A brief description of the strengths and limitations of my model for my stakeholder.**
 - **Final recommendations based on your analysis.**
- 

Problem

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Stake holders are health organizations .

This analysis will help them so that they can control the stroke events



BUSINESS PROBLEM AND STAKEHOLDERS

The source of my data is Kaggle

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

Each row in the data provides relevant information about the patient

See the next slide for Data Dictionary

BRIEF INTRODUCTION TO DATA

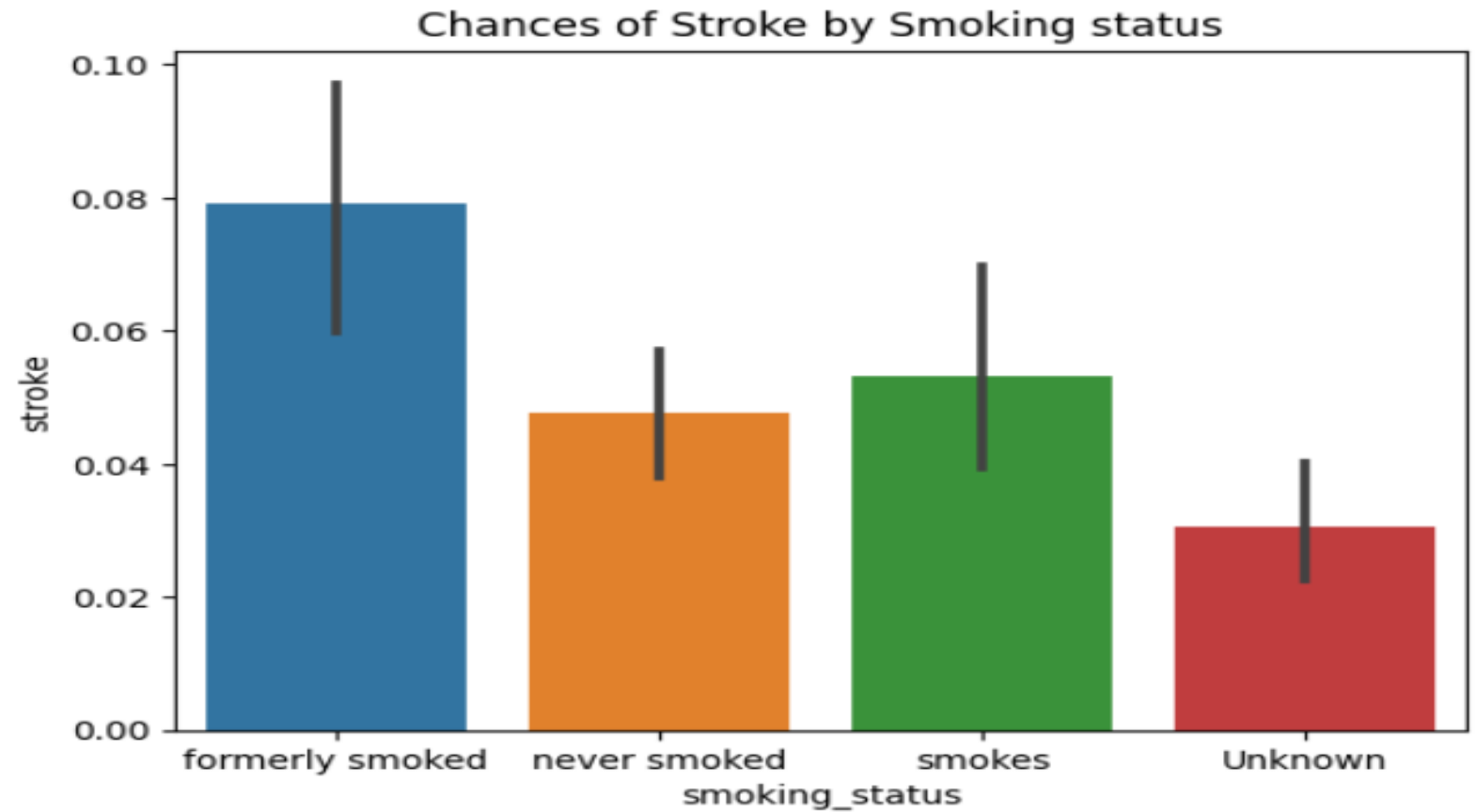
Attribute Information

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has 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

*Note: "Unknown" in means that the information is unavailable for this patient
smoking_status

BRIEF INTRODUCTION TO DATA

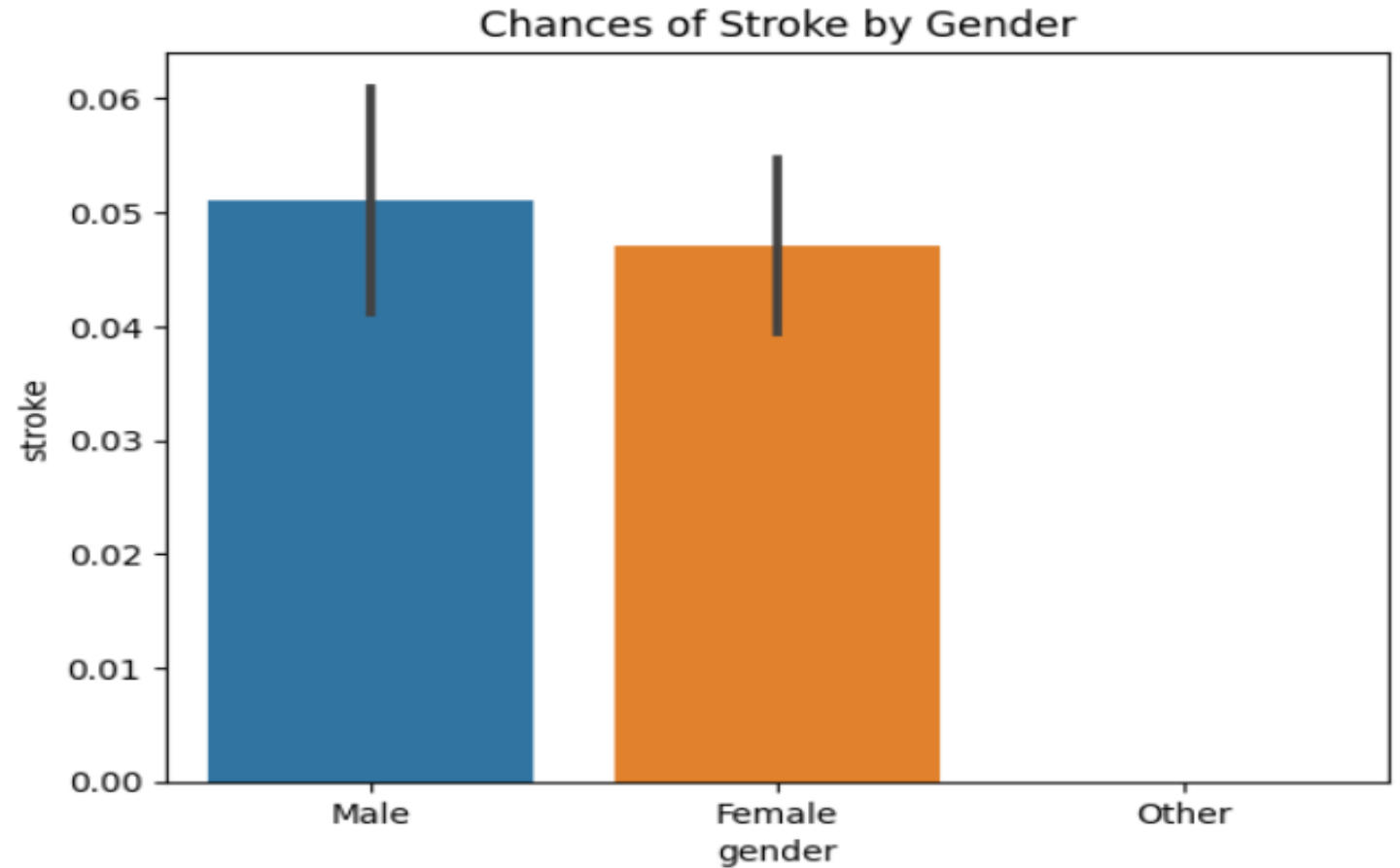
1. Smoking Status



Those that formerly smoked are at high chances of stroke

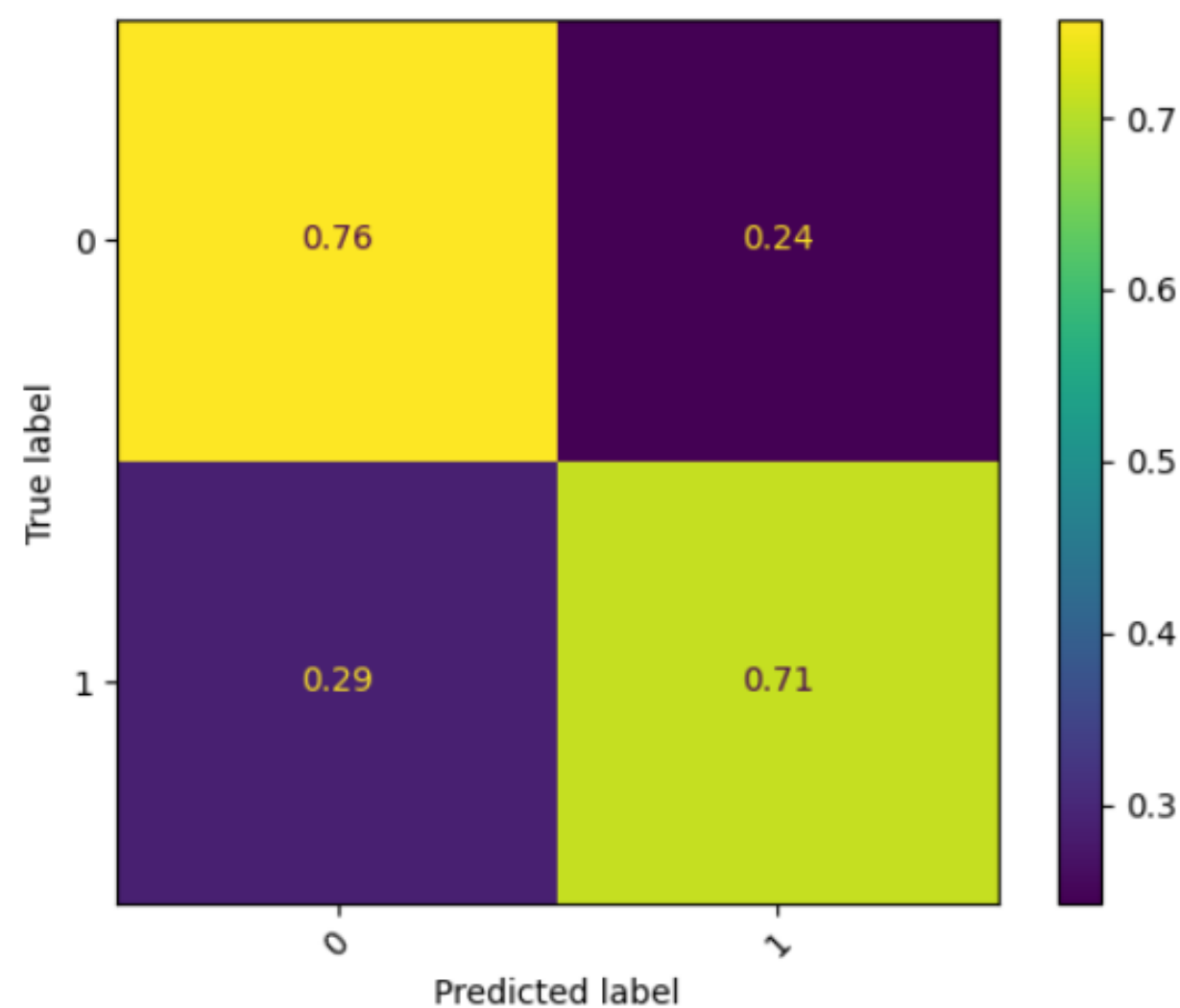
FINDINGS TO STAKE HOLDERS

2. Gender



Stroke events seems to occur more in Males

FINDINGS TO STAKE HOLDERS



```
✓ [180] print(classification_report(y_test, test_preds))
```

	precision	recall	f1-score	support
0	0.98	0.76	0.85	1198
1	0.16	0.71	0.27	80
accuracy			0.75	1278
macro avg	0.57	0.73	0.56	1278
weighted avg	0.92	0.75	0.82	1278

STRENGTHS AND LIMITATIONS

The model was able to correctly predict the 76% of patients that are not likely to have stroke.

The model was able to correctly predict the 71% of patients that are likely to have stroke.

The shortcomings was that :

The model incorrectly predicted that 29% of patients are not likely to have stroke while they are likely to have it , this is a big shortfall as these people might miss the treatment.

The model incorrectly predicted that 24% of patients are likely to have stroke while they are not it , this is also a costly error as this people might get treated while they are not at risk.

The challenge was that this data I was given was not balanced between the 2 classes of patience , I used the SMOTE to balance the data, it improved the scores, but it is still not the best.

STRENGTHS AND LIMITATIONS

```
df['stroke'].value_counts(normalize=True)
```

```
0    0.951272
```

```
1    0.048728
```

```
Name: stroke, dtype: float64
```

Data before I did SMOTE , showed 95% in the no stroke class and 4% in the stroke class

STRENGTHS AND LIMITATIONS

FINAL RECOMMENDATIONS

- The Health organizations must advise patients to never start smoking as we see from findings that the ones at the risk the most are the ones who formerly smoked and the ones that currently smoke.
- The Health organizations must also encourage males to be more careful as their at the risk the most.
- The stakeholders provide balanced data so that the models can be more accurate, as this one has some false predictions that we saw.

FINAL RECOMMENDATIONS



Q&A

Thank you!



**THANK YOU FOR YOUR
ATTENTION**