# Are you prone to a stroke?

by: Pat Lar

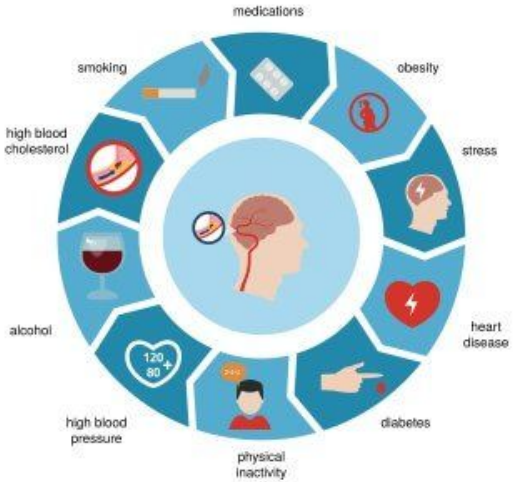
# Why There's a Need to Focus on Strokes

According to the World Health Organization (WHO) stroke is the **2nd** leading cause of death globally, responsible for approximately **11%** of total deaths.

- #1 leading cause of death in America (cdc.gov)

For stroke survivors that did not get help quickly, they may live with complications

- Memory loss, speech impairment, eating disabilities, and/or loss of normal bodily functions (source: Johns Hopkins)




YASHODA HOSPITALS

**Risk Factors For Stroke**

Source: Yashoda Hospitals

# But Prevention is Possible!

# Problem

## Can patients with an affinity to having a stroke, be identified with a high degree of accuracy and sensitivity?

As health data from smartphones, smartwatches, and even tech-friendly primary care facilities continues to accumulate, there is a real opportunity to identify those future at risk stroke victims.

- Provide patients with targeted preventative measures
- Monitor patients predicted as high risk
  - Possibly integrate this monitoring into smartwatches

# TABLE OF CONTENTS

# Data Source

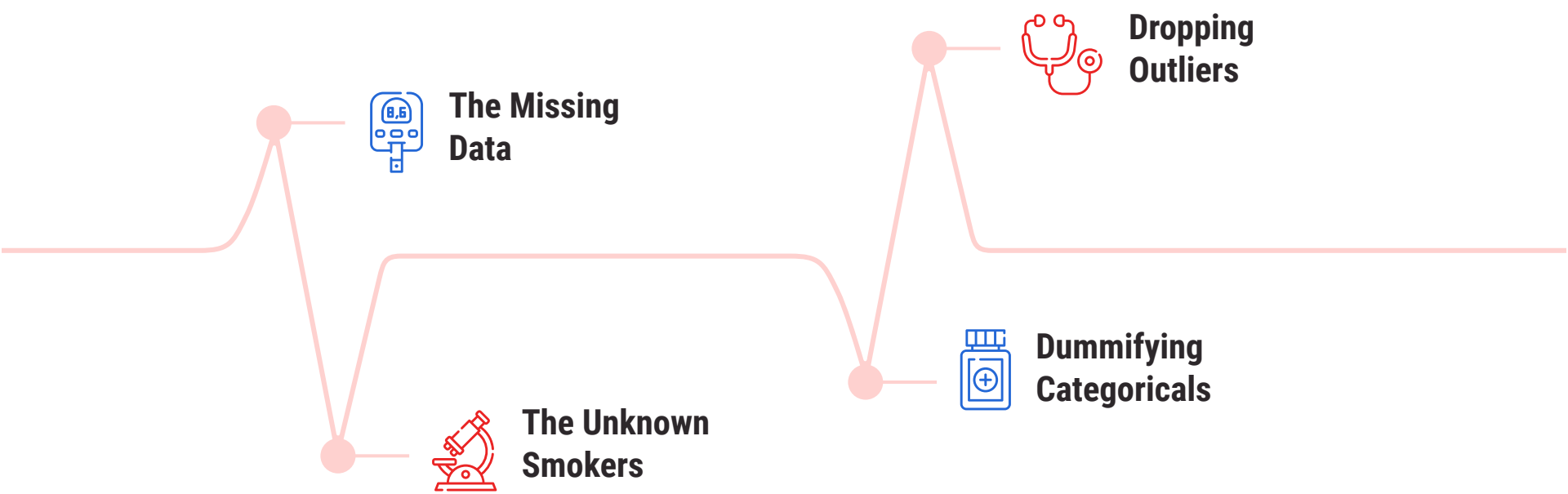
- Sourced from Kaggle competition data
  - Used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status.
- Competition stated that the sourcing on this data was confidential

**Patient Data Features**

'glucose level',
'married',
'residence',
'hypertension',
'age',
'heart disease',
'gender',
'smoker',
'stroke'
'bmi',

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

# Data Cleaning



The Missing Data

Dropping Outliers

The Unknown Smokers

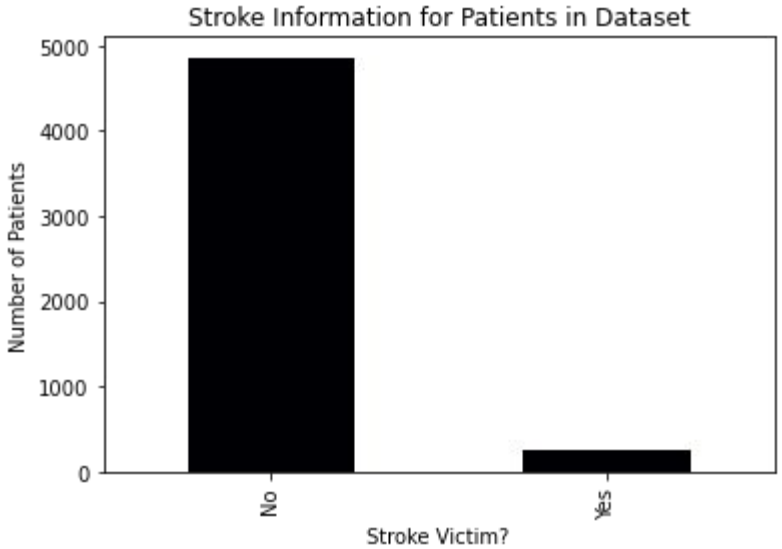
Dummifying Categoricals

# EDA- Imbalanced Classes

- Data is extremely imbalanced
  - 0.05% were stroke victims (248)

- Techniques to handle imbalances:
  - Oversampling
  - Undersampling
  - Synthetic Minority Over-sampling Technique (SMOTE)

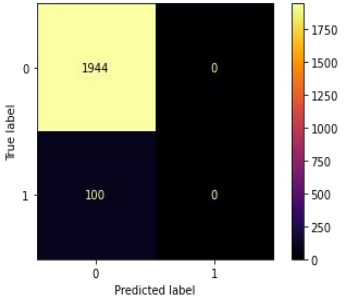
- Metrics:
  - Recall/sensitivity (reduce false negatives)
  - Accuracy



# EDA- Imbalanced Classes

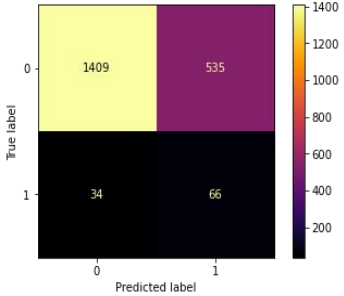
## Control

- Recall = 0%
- Accuracy = 95%



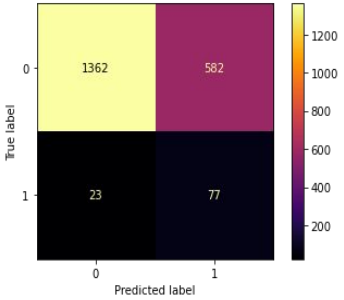

## SMOTE

- Recall = 72%
- Accuracy = 69%



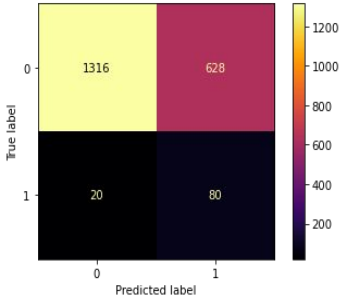

## Oversampling

- Recall = 74%
- Accuracy = 70%

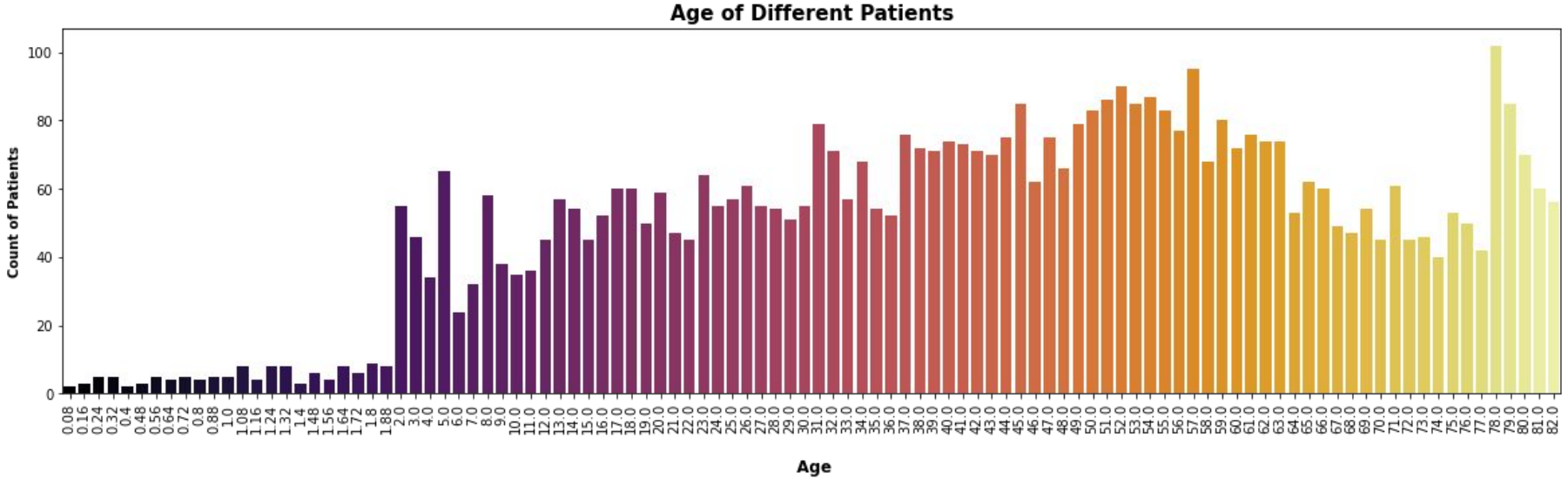



## Undersampling

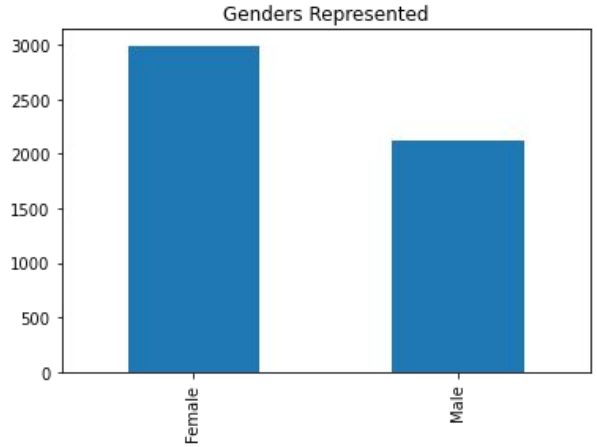
- Recall = 74%
- Accuracy = 68%

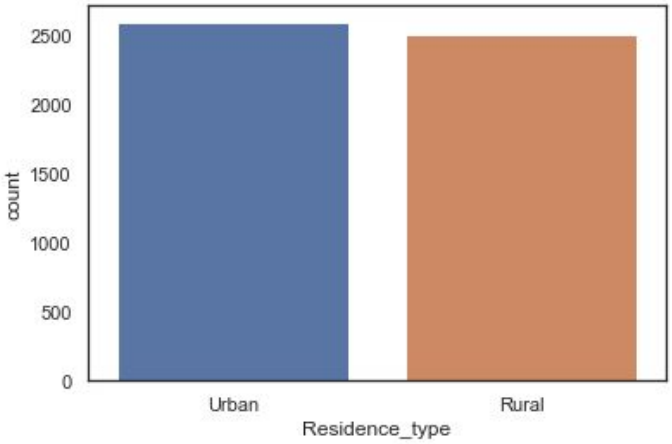


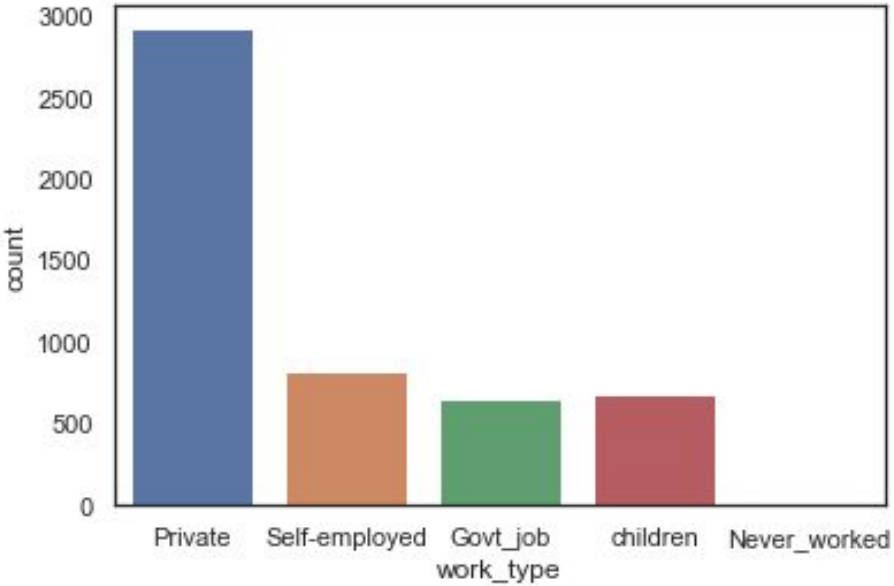
# EDA- General



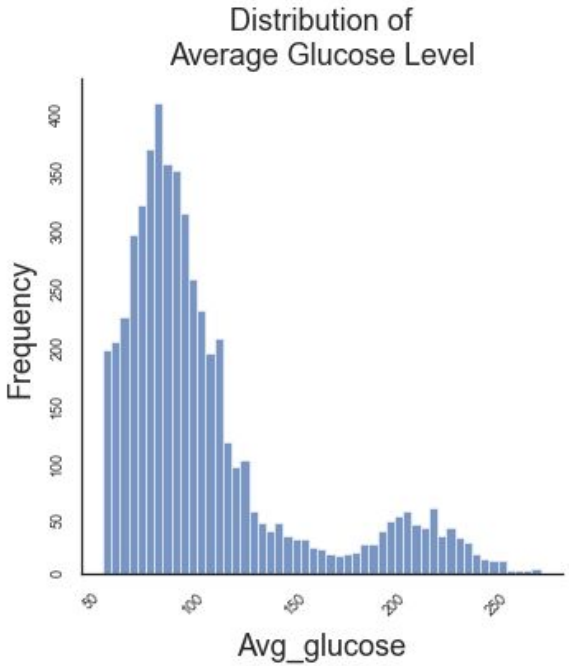
# EDA- General

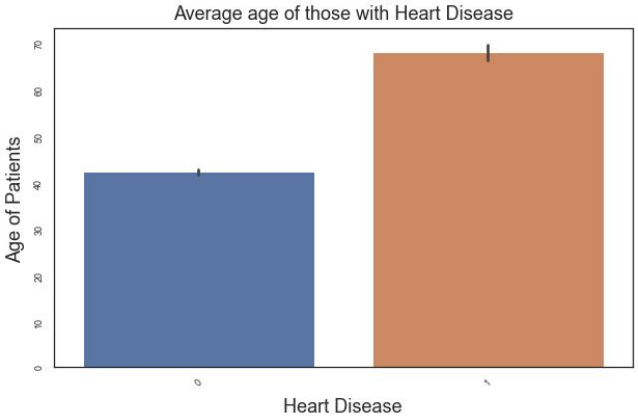

Genders Represented
3000
2500
2000
1500
1000
500
0
Female
Male


count
2500
2000
1500
1000
500
0
Urban
Rural
Residence_type

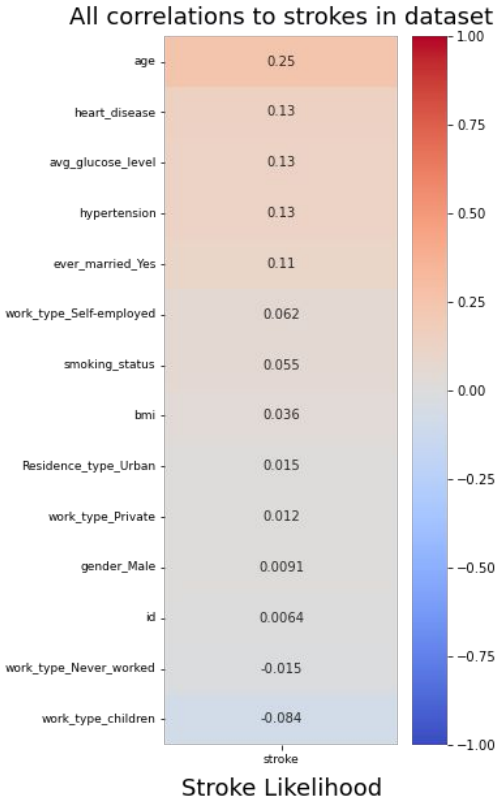

count
3000
2500
2000
1500
1000
500
0
Private
Self-employed
Govt_job
children
Never_worked
work_type

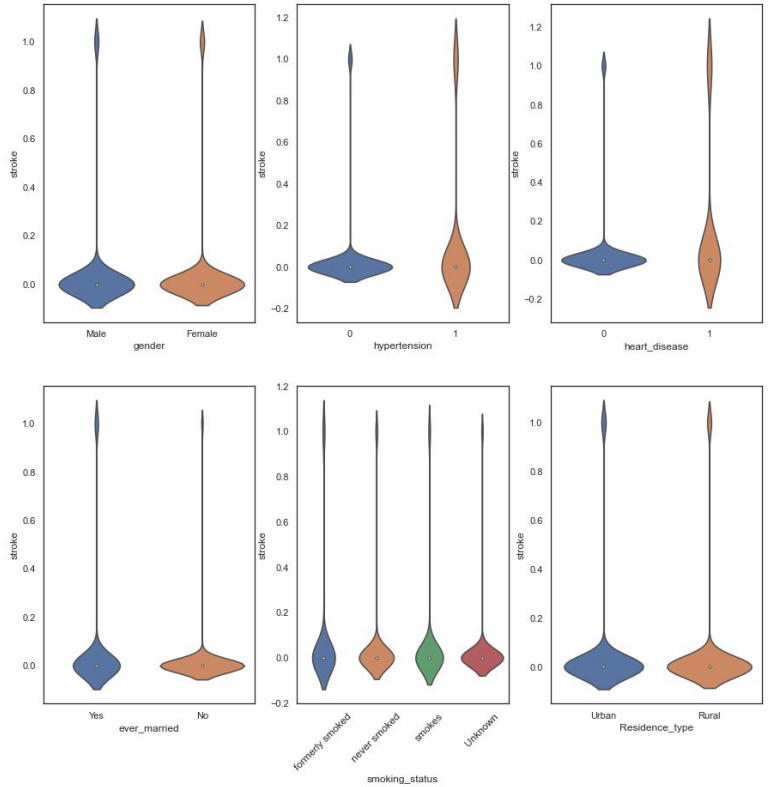
# EDA- General


Distribution of
Average Glucose Level
Frequency
0
50
100
150
200
250
300
350
400
50
100
150
200
250
Avg_glucose

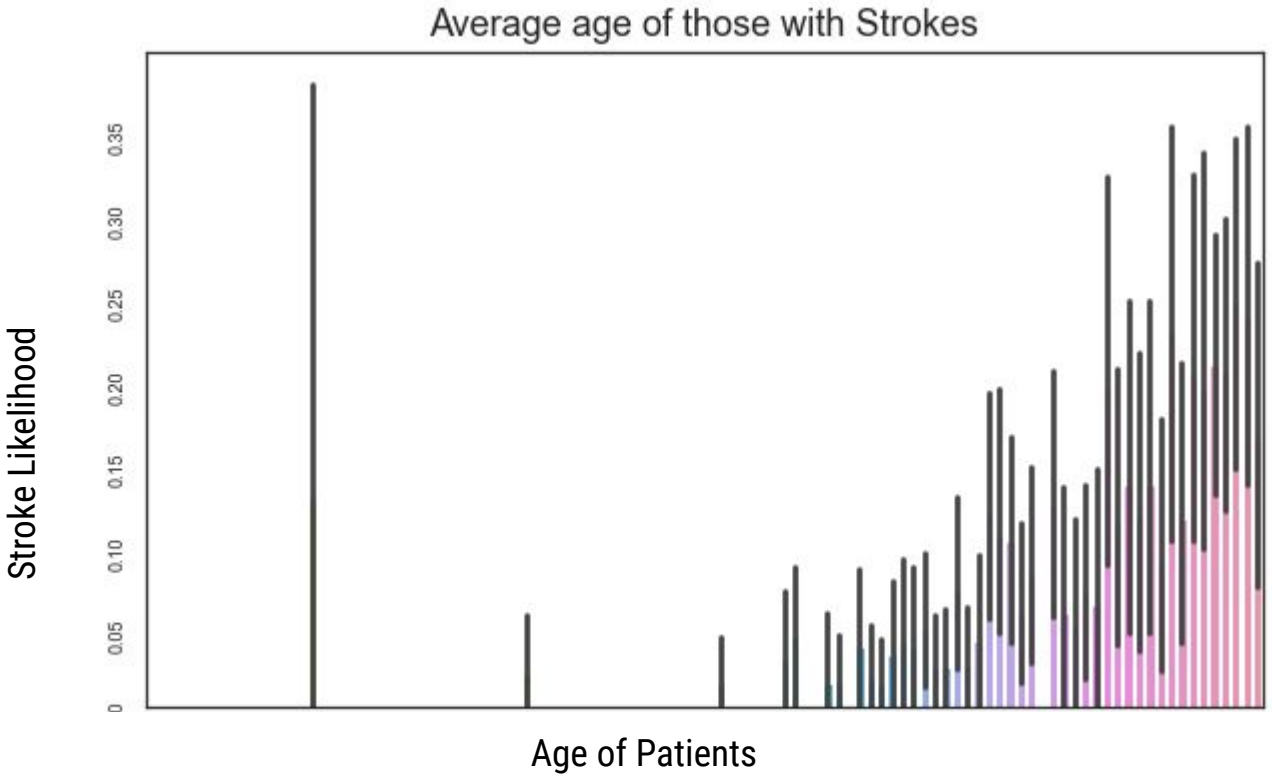

Average age of those with Heart Disease
Age of Patients
0
10
20
30
40
50
60
70
0
1
Heart Disease

# EDA- Stroke Victims



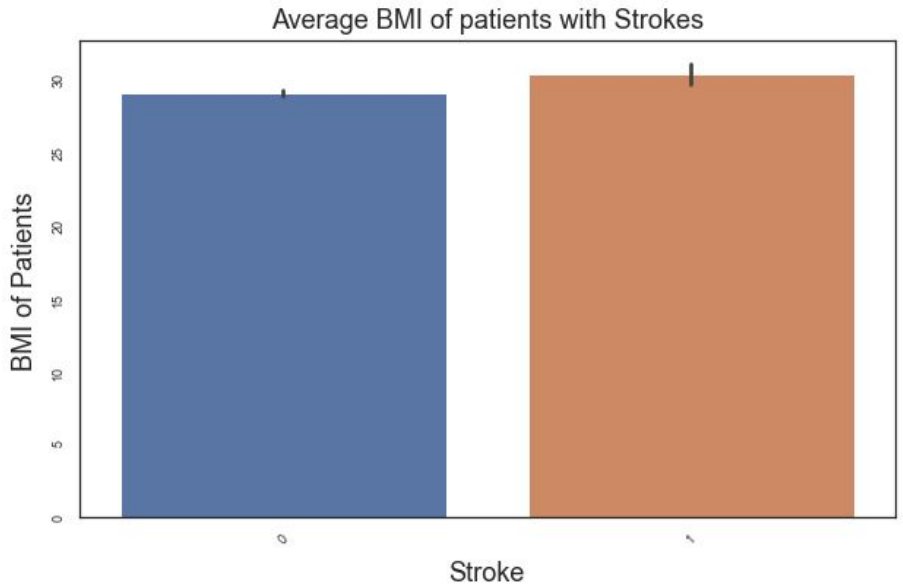



# EDA- Stroke Victims



# EDA- Stroke Victims

- BMI 18.5 to 24.9, it falls within the normal range



# Baseline Evaluation

## Baseline Accuracy

- 95% without oversampling
- 50% with oversampling

# Models

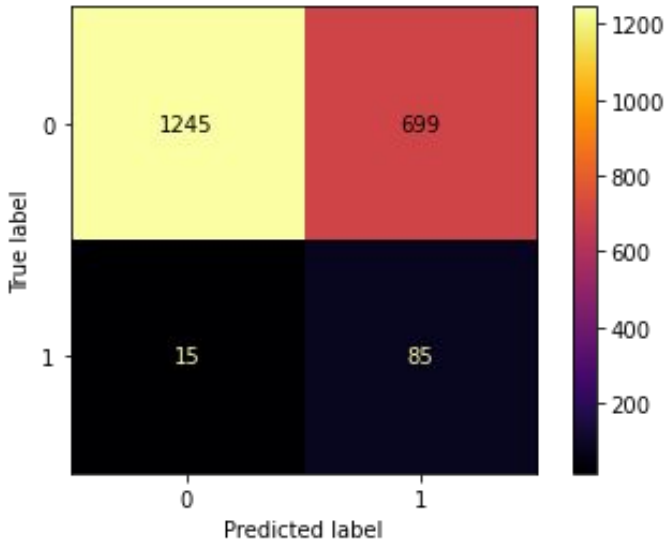
| Models Tested | Recall | Accuracy | Train & Test Scores |
|---|---|---|---|
| Logistic Regression | 78% | 73% | 0.77, 0.73 |
| Support Vector Machine | 83% | 65% | 0.78, 0.65 |
| K Nearest Neighbor | 32% | 82% | 0.94, 0.82 |
| Decision Tree | 85% | 65% | 0.80, 0.65 |
| Extremely Randomized Trees (ExtraTrees) | 84% | 63% | 0.77, 0.63 |
| Random Forest Classifier | 74% | 70% | 0.85, 0.70 |

# Decision Tree!

# Top Model- Decision Tree

- **Best Params**
  - Min Samples Split: 4
  - Min Samples Leaf: 1
  - Max depth: 3
  - Class Weight: balanced

- **Top Features**
  - Age (weight = 0.92)
  - BMI
  - Hypertension
  - (all other features ignored by model)



# Results and Conclusions

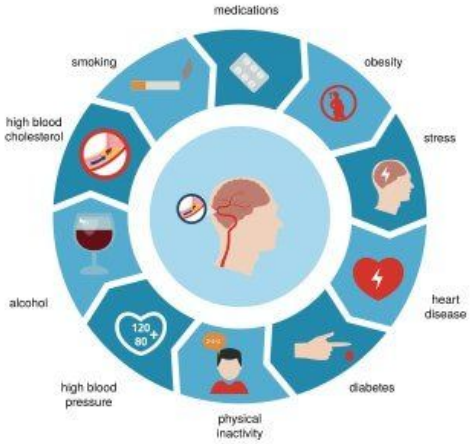
*Can patients with an affinity to having a stroke, be identified with a high degree of accuracy and sensitivity?*

- Yes!
- Models performed okay, average glucose level non-factor
- Imbalanced classes created an uphill battle for the models

# Results and Conclusions

***How can this help you?***

- Age is the strongest stroke indicator
- Focus on managing BMI and hypertension




YASHODA HOSPITALS

**Risk Factors For Stroke**

# Future Work

- Get more data!
- Create a functioning site with streamlit for patients to discover their stroke risk
- Incorporate this data with other health data to help determine stroke patient's risk