THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Stroke Analysis

A machine learning and visualization Capstone project.

Priya Johny

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

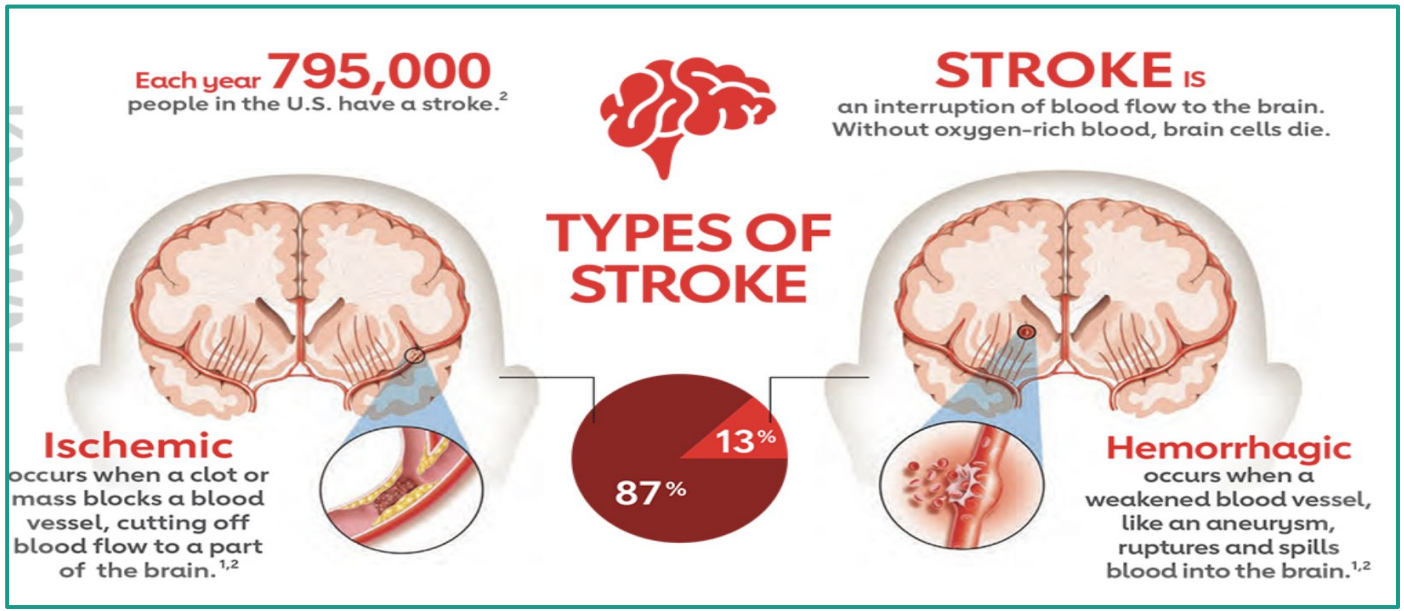
Motivation

Over 15 million suffer with stroke (the brain attack).

According to WHO:

“ Stroke is the second leading cause of death and the third leading cause of disability. One in 4 people are in danger of stroke in their lifetime. Lifestyle risk factor 4/10 for hypertension and 2/5th under age of 65 for smoking.”

Strokes Types



Credit: American Stroke Association.

Aim : To analyze and visualize on what factors of lifestyle stroke using python, ML, and Tableau.

About Dataset 1: Credit: FEDESORIANO, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

```
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             5110 non-null   object
 1   age                5110 non-null   float64
 2   hypertension       5110 non-null   int64
 3   heart_disease      5110 non-null   int64
 4   ever_married       5110 non-null   object
 5   work_type          5110 non-null   object
 6   Residence_type     5110 non-null   object
 7   avg_glucose_level  5110 non-null   float64
 8   bmi                4909 non-null   float64
 9   smoking_status     5110 non-null   object
 10  stroke             5110 non-null   int64
dtypes: float64(3), int64(3), object(5)
memory usage: 439.3+ KB
```

A binary classification problem

> We can see the min age is 0.08 indicating an infant and a max of 82 yr old adult involved in this dataset.

> We see that mean of age is over 43yrs old, mean of glucose level falls on 106.14 with a bmi of mean 29.

> Classification refers to a predictive modeling problem where a class label is predicted, here stroke.

> We find 201 null values in the 'bmi' column.

Filling them with mean values in python analysis and excluding on tableau analysis.

EDA

➢ **Categorical Features:**

01	Gender	Female > Male
02	Hypertension	No Hypertension > Hypertension
03	Heart_disease	No heart disease > heart disease
04	Ever_married	Married > Unmarried
05	Working_type	Private > Self-employed > Govt_job > children
06	Residence_type	Urban > Rural
07	Smoking_status	Never Smoked > Formerly Smoked > Smokes

➢ **Quantitative Features:**

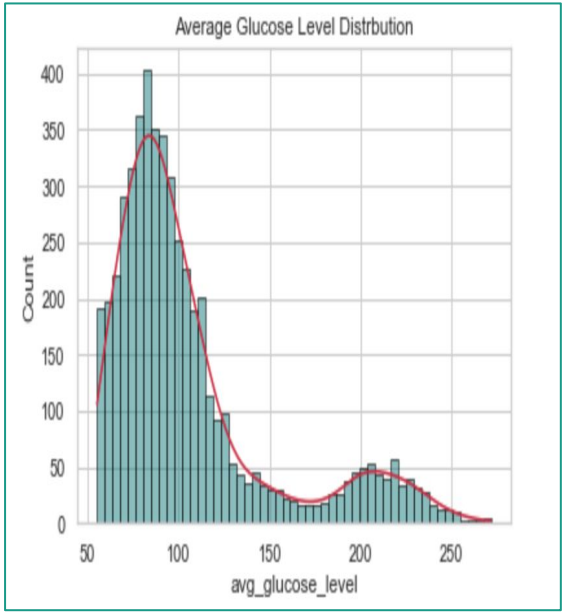
Age : The chance of having a stroke about doubles every 10 years after age 55.

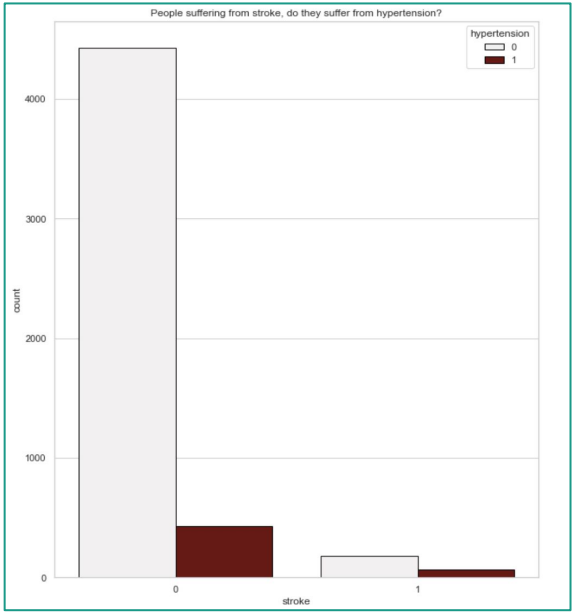
Avg_glucose_level : High blood glucose is found in stroke cases, 126+ has been observed a lot.

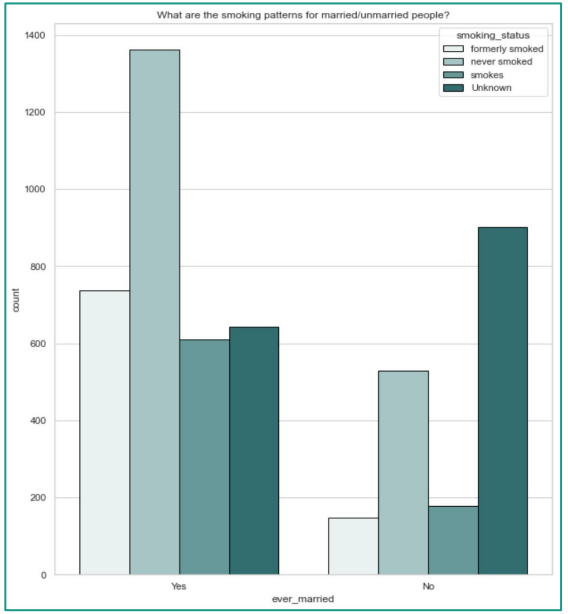
BMI : High BMI values increases the chances of Ischemic stroke.

Credit: https://lamvascular.com/healthy-lifestyle-dallas/

EDA



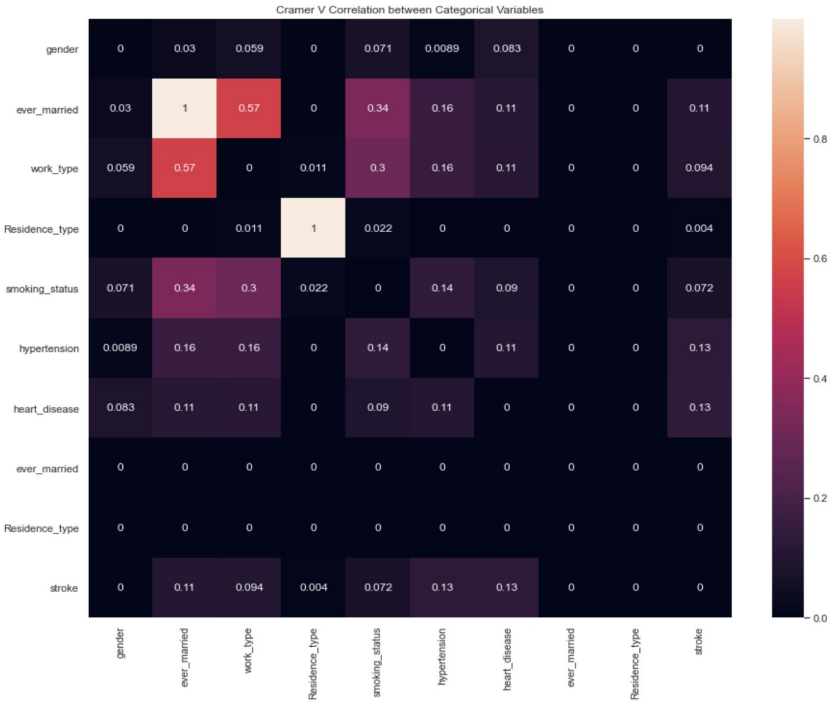




More Visualizations through tableau

- Datapoint are biased towards no stroke.
- Contradict the domain knowledge for the features : **hyper-tension(high-blood pressure)**, **heart-disease** and **smoking-status**.
- Feature engineering process, balance the dataset using SMOTE analysis and feed the balanced to some of the ML algorithm and analyze.

Correlation with Chi-Square



Feature Engineering

Smote Analysis

Oversampling : Increase the minority samples of the target variable to the majority samples

```
Before balancing the data:  Counter({0: 3887, 1: 201})
```

```
After balancing the data:  Counter({0: 3887, 1: 3887})
```

Modelling

Models used are:

1) Logistic Regression : 0.945
2) KNN: 0.940
3) Random Forest Classifier: 0.95
4) Ada Boost: 0.956
5) Decision Tree: 0.89
6) XGBOOST: 0.94

Feature Interpretation by XGBoost

Feature Interpretation:

Weight	Feature
0.1302	smoking_status_smokes
0.0981	work_type_Govt_job
0.0847	age
0.0695	work_type_children
0.0686	ever_married_Yes
0.0610	Residence_type_Urban
0.0582	smoking_status_never smoked
0.0510	work_type_Self-employed
0.0464	gender_Female
0.0405	hypertension
0.0399	smoking_status_Unknown
0.0390	gender_Male
0.0368	smoking_status_formerly smoked
0.0363	ever_married_No
0.0360	work_type_Private
0.0340	Residence_type_Rural
0.0310	heart_disease
0.0202	avg_glucose_level
0.0186	bmi
0	gender_Other

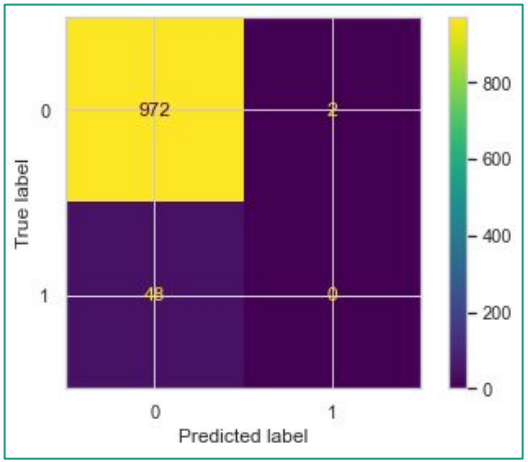
... *1 more* ...

```
F score on train set is: 0.997296949414339
F score on test set is: 0.12244897959183673

Precision on train set is: 0.998968540484786
Precision on test set is: 0.13043478260869565

Recall on train set is: 0.9956309432022616
Recall on test set is: 0.11538461538461539

Train ROC is: 0.999954028700754
Test ROC is: 0.7564036478984932
```



In AdaBoost we can see its rightly predicting 972 value as correct, which is highest by any model till now

Aim : Ischemic is considered to be the type of stroke that occurs the most. Based on previous analyzing we further use a 30-day stroke data over a period of 4 years to analyze further through tableau.

Dataset 2: https://data.chhs.ca.gov

This dataset contains risk-adjusted 30-day mortality and 30-day readmission rates, quality ratings, and number of deaths / readmissions and cases for ischemic stroke treated in California hospitals.

Outcomes Measures:

- ➔ Risk-adjustment
- ➔ Validation
- ➔ Risk factors

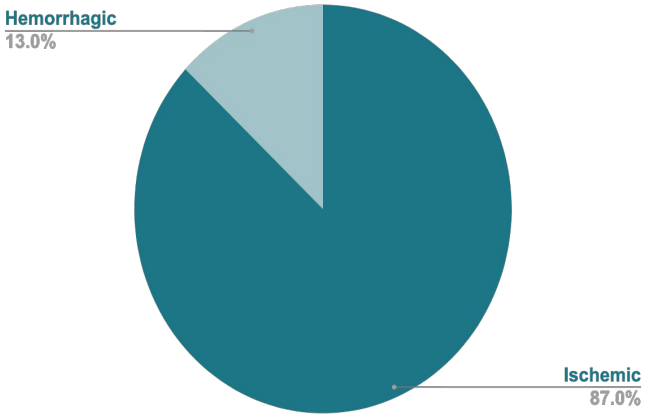


Tableau Presentation...

Thank You!