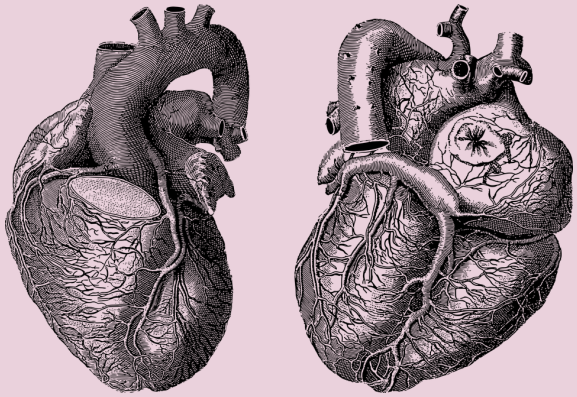


Stroke Prediction Dataset

Victoria Almazan
Coding Dojo

Project 2 - Advance Machine Learning

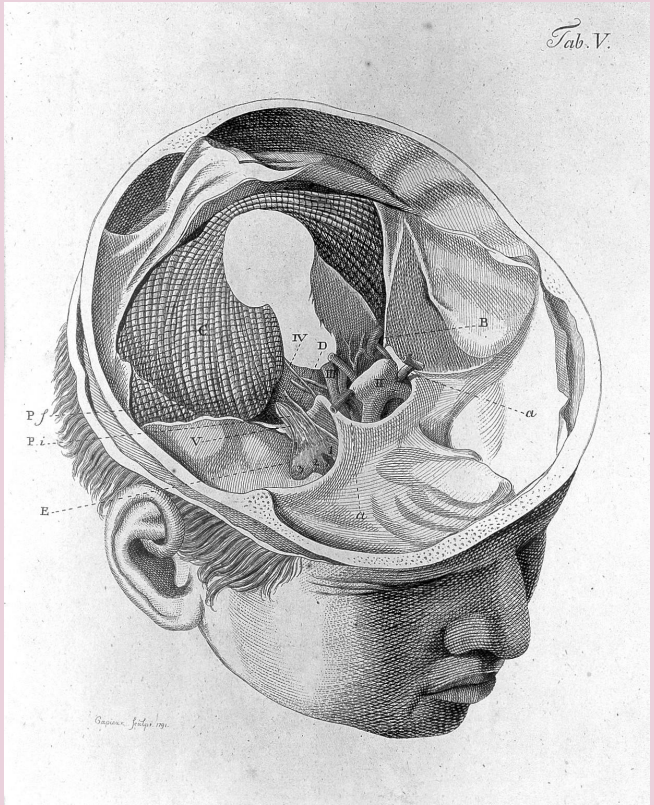
Objective:

Our Goal is to better help our partners understand what is most likely the highest leading causes that attributes to strokes. We will be diving into understanding key features that will better explain such reasons and how to proceed with moving forward with better business actions.

Data Description:

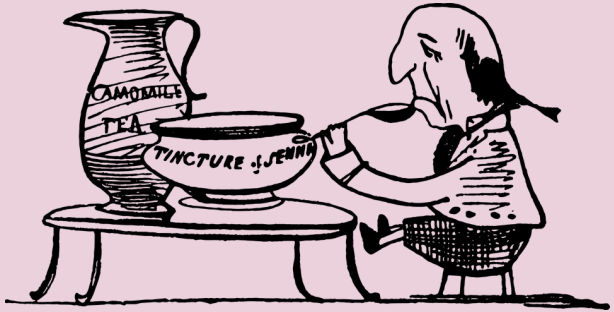
This dataset represents a study based on strokes being the leading cause of death by 11% globally per the World Health Organization.

This Data set can be found on Kaggle.



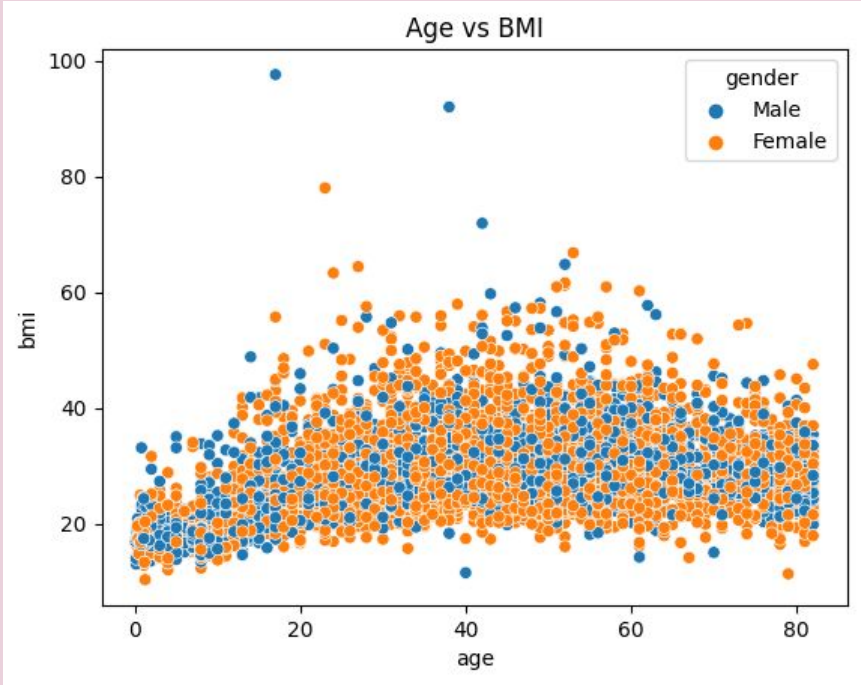
What is our target?

The target of this data is to predict whether it is likely for a patient to get a stroke based on different attributes such as age, heart disease, and smoking status(to name a few)! Additionally, outside attributes also aid in understanding our data more thoroughly, we will be applying visuals to better analyze before taking further business actions.



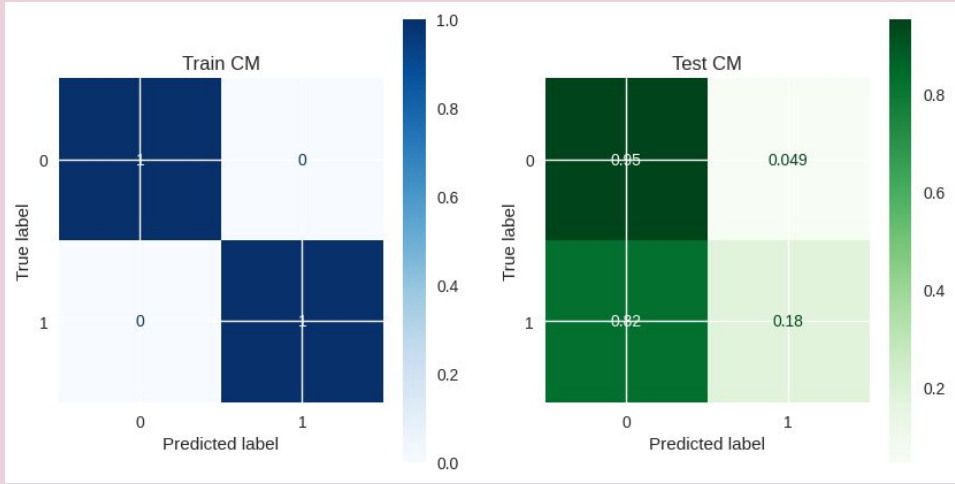
Visualizations and Understanding

- The visual on the right represents features such as 'BMI', 'Age' and 'Gender'.
- We can assume that the majority of the data has an average bmi of 20- 60. The majority seem to be more female than male. This shows us a baseline of where are our participants in the study.



Visualizations and Understanding

Our Best model was the Decision Tree Model, there is opportunity to improve our model so we will use logistic regression to hypertune our model.




```
train
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3645
           1       1.00      1.00      1.00       187

    accuracy                           1.00      3832
   macro avg       1.00      1.00      1.00      3832
weighted avg       1.00      1.00      1.00      3832

test
              precision    recall  f1-score   support

           0       0.96      0.95      0.95      1216
           1       0.16      0.18      0.17        62

    accuracy                           0.91      1278
   macro avg       0.56      0.56      0.56      1278
weighted avg       0.92      0.91      0.92      1278
```

On the next slide we will see our results using LogReg Hypertuning.

Visualizations and Understanding

[i] CLASSIFICATION REPORT FOR: Training Data

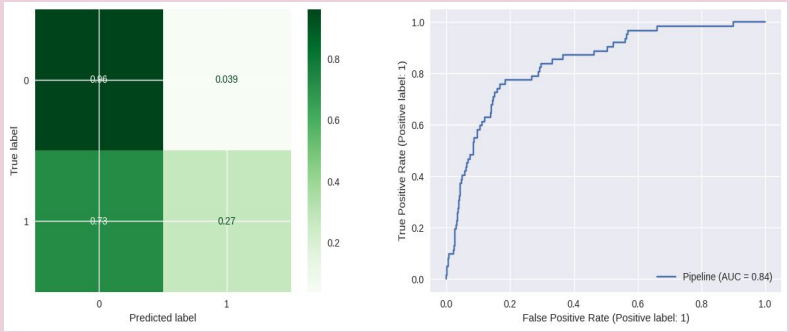
	precision	recall	f1-score	support
0	0.98	0.86	0.91	3645
1	0.17	0.58	0.27	187
accuracy			0.84	3832
macro avg	0.58	0.72	0.59	3832
weighted avg	0.94	0.84	0.88	3832

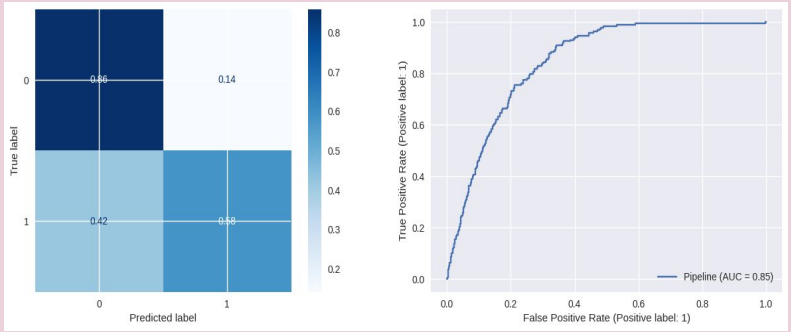
[i] CLASSIFICATION REPORT FOR: Test Data

	precision	recall	f1-score	support
0	0.96	0.96	0.96	1216
1	0.27	0.27	0.27	62
accuracy			0.93	1278
macro avg	0.61	0.62	0.62	1278
weighted avg	0.93	0.93	0.93	1278

Here we can see our LogReg Models Testing & Training Data. Our predictions may have some false negatives but with tuning we were able to lessen the margin of false negatives and positives. Then we used PCA for a final score of

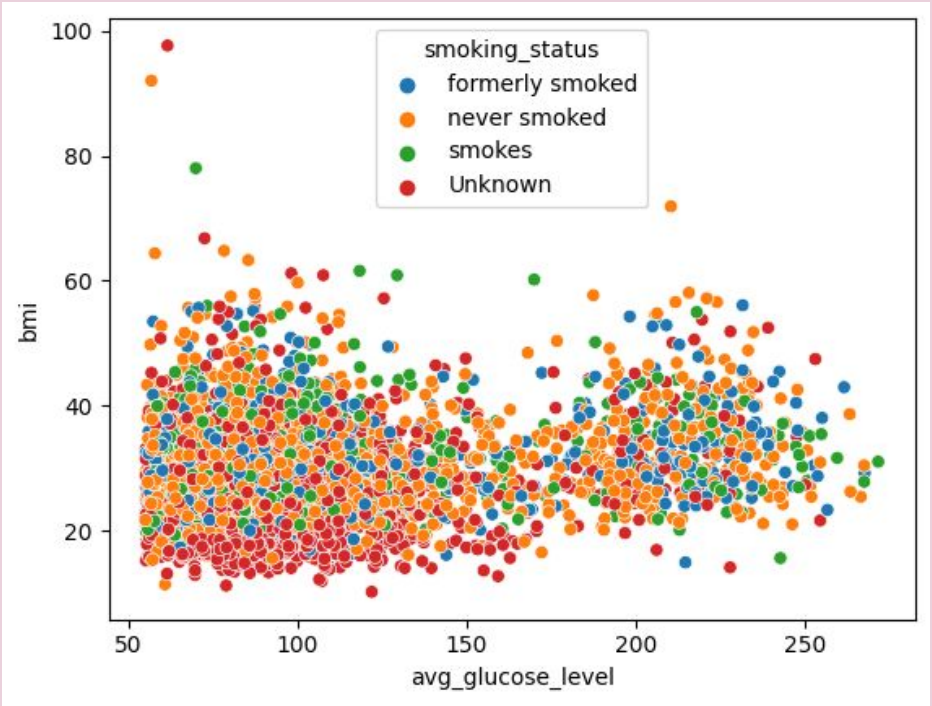
```
Testing accuracy: 0.9374021909233177
```





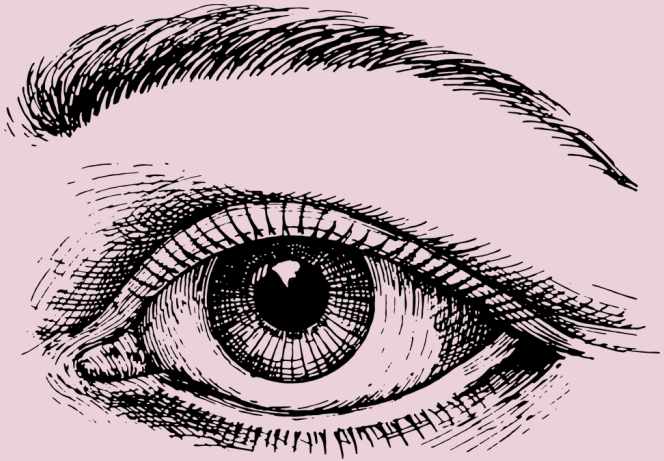
Challenges?

Based on all of our data, our modeling did have false negatives and false positives, but with modeling and hypertuning we were able to make sure that we adjusted our algorithms to be above 94% accuracy. The impacts of having false negatives and false positives means that any falsely reported information would greatly impact the overall health of patients.For example, this can impact recommendations of prescribing medicines and treatments that can interfere features such as heart disease, avg glucose level, smoking status and bmi had higher correlation. For example, as bmi increases, insulin resistance also increase which in turn increases blood glucose level to rise as we can see in the figure.



My Recommendations!

- My recommendations are to continue to collect data to have continuous understanding base line of changing data but additionally I recommend adding a column indicating if patient is on a medication plan. This would also help understand if certain medical treatments are effective in preventing or promoting strokes.
- Additionally, based on our models we can deduce that 93% of our data will predict the likelihood of patients will have a stroke based on features such as age, smoking status, bmi, heart disease and avg glucose level.



Questions?