Healthcare dataset

STROKE OR NOT?

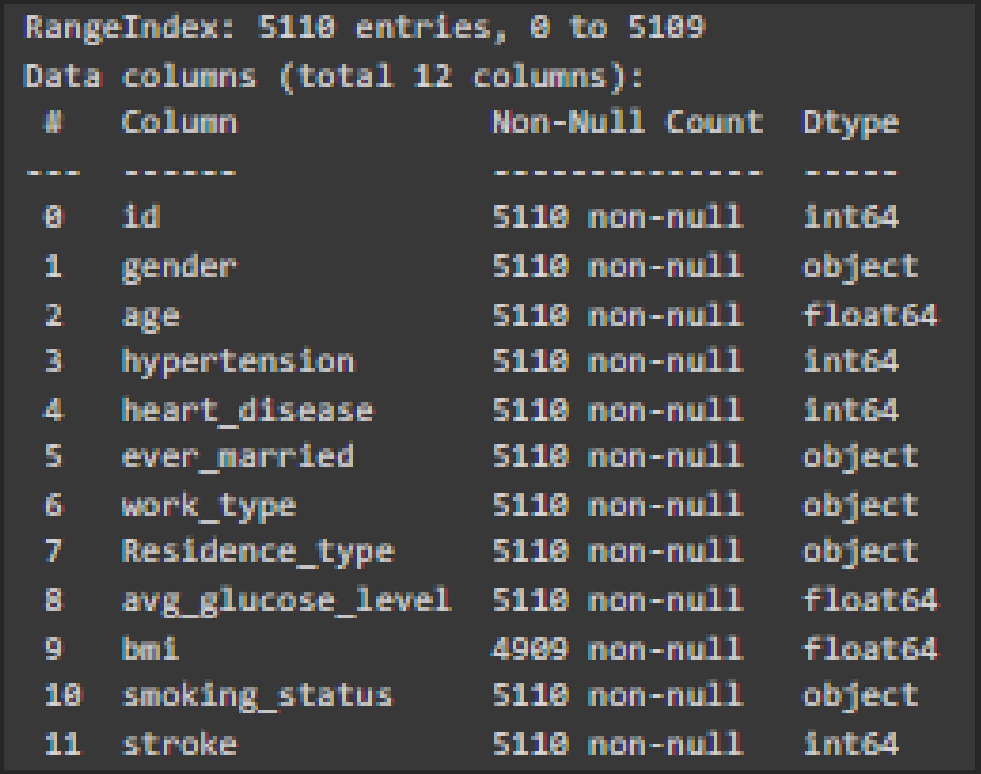
Who are our stakeholders?

- In the context of the problem presented, our stakeholders would be doctors healthcare professionals and medical insurance companies. The analysis can be used by the stakeholders to provide patients with adequate preventative care and allow insurance companies to provide clients with cover at a lower risk factor.

What Problem are we solving?

- The purpose of this project is to predict the likelihood of a stroke in patients dependent on various factors such as : Smoker status, Age, Gender, Etc.

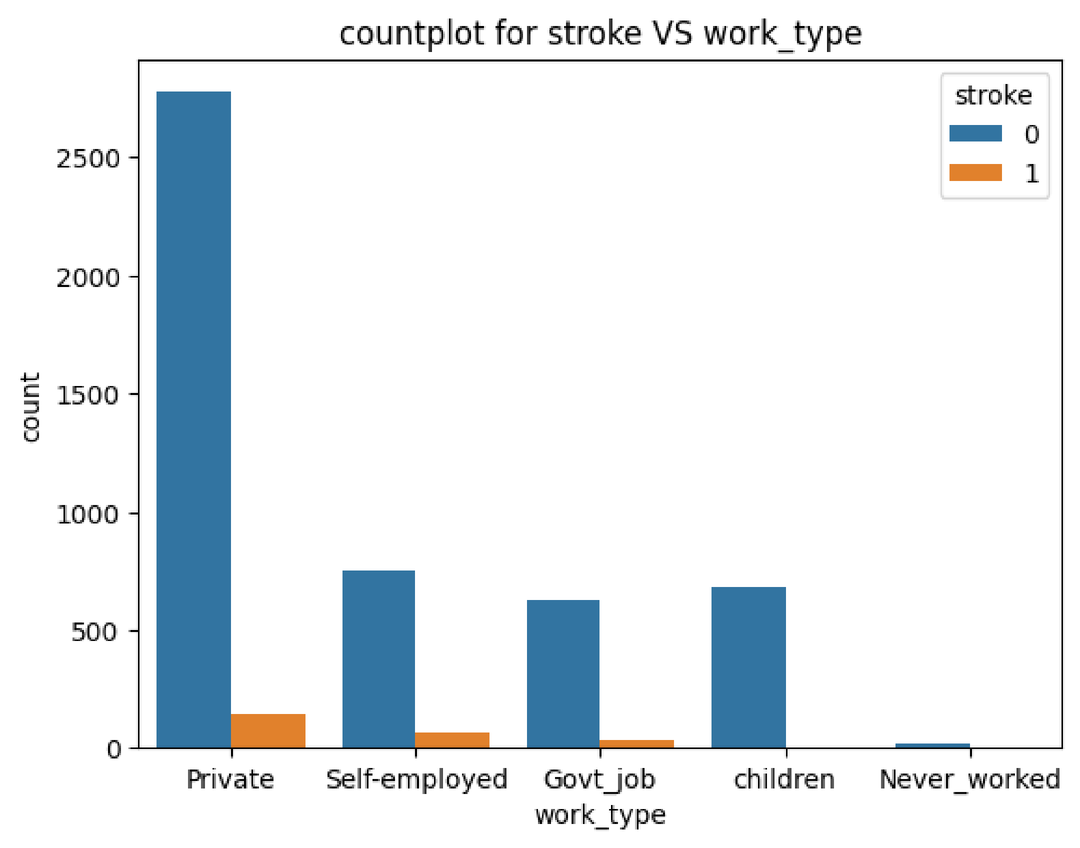
Features in data



```
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
```

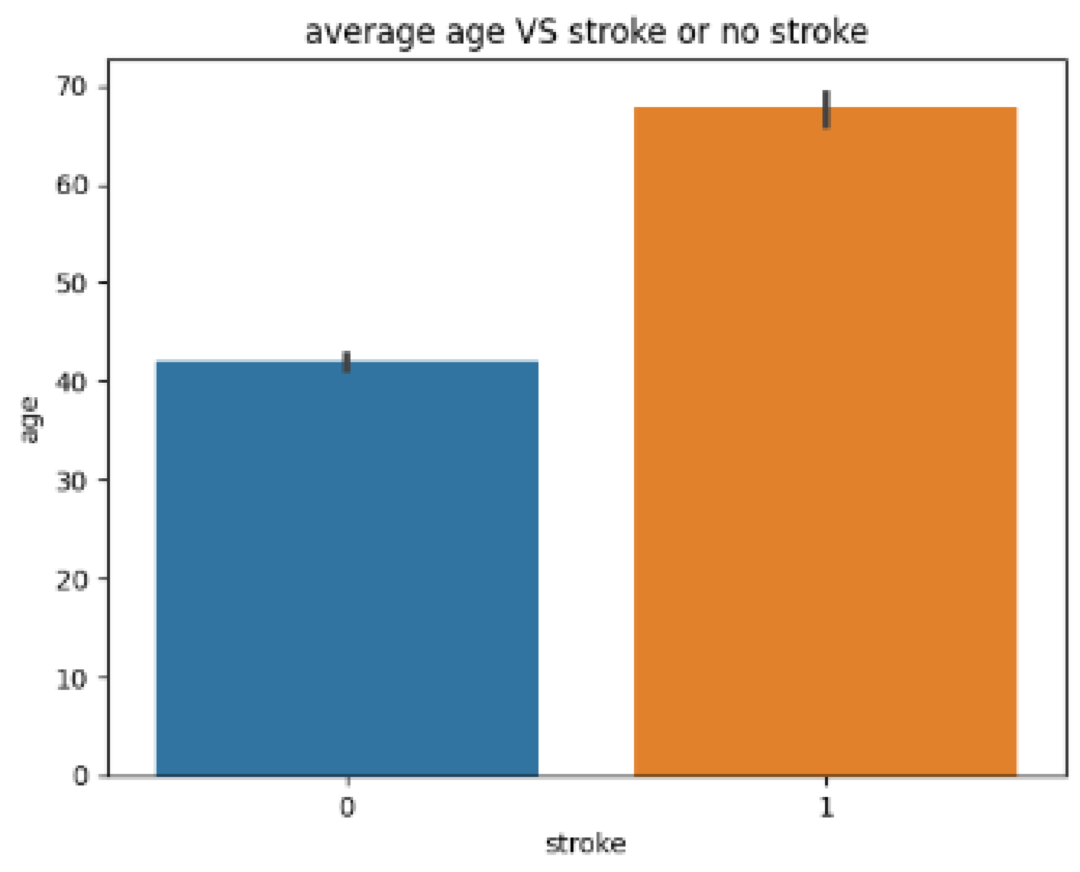
Introduction to data

- A stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications.
- 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.



Visual: 1

- The above depicted graph displays a count of predicted stroke VS individual work type
- As we can see the private sector has the highest counts for stroke and no stroke.



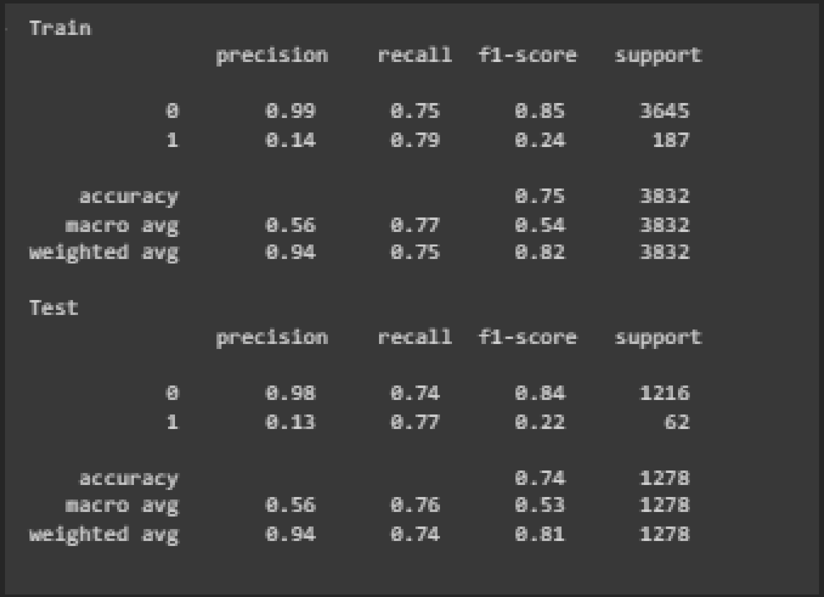
Visual: 2

- Looking at visual 2, we can see a bar plot for age vs stroke(yes or no)
- The graph shows us that individuals who are older are at greater risk of experiencing a stroke.

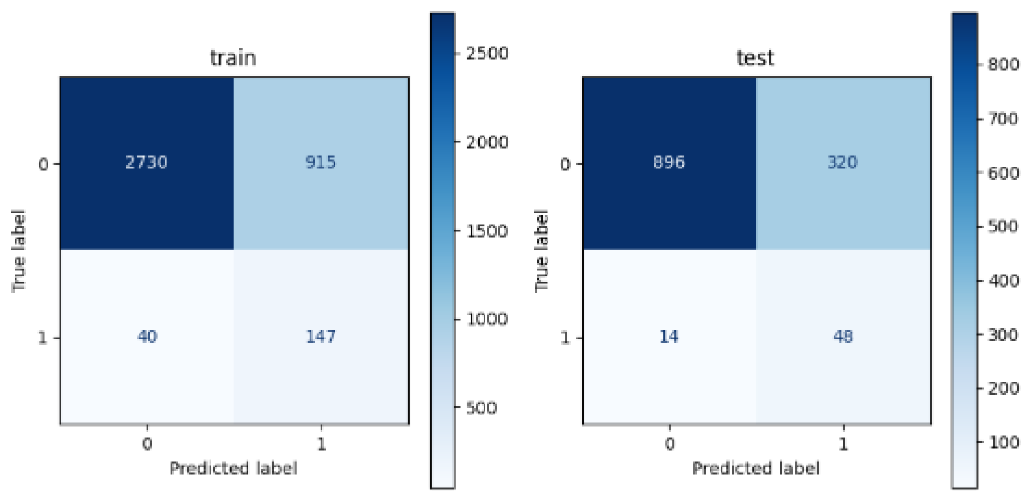
Strengths and limitations of my model

- The model chosen for production would be a logistic regression model.
- The limitations within the data set would be mainly the class imbalance, seeing as this would be a binary classification problem (0 or 1! / Yes or NO) we had many no stroke predictions and few stroke predictions.
- This could cause Bias towards a no stroke prediction within our model!
- In seeing the class imbalance we've also made use of SMOTE to balance our classes.
- False negatives would mean that we would have patients classified as not being likely to have a stroke when they may be at risk! Thus, I've chosen to reduce the number of false negatives in the model
- False positives would classify patients that are not likely to have a stroke as being at risk!

Model metrics and evaluation



```
Train
              precision    recall  f1-score   support

           0       0.99      0.75      0.85      3645
           1       0.14      0.79      0.24       187

    accuracy                           0.75      3832
   macro avg       0.56      0.77      0.54      3832
weighted avg       0.94      0.75      0.82      3832

Test
              precision    recall  f1-score   support

           0       0.98      0.74      0.84      1216
           1       0.13      0.77      0.22        62

    accuracy                           0.74      1278
   macro avg       0.56      0.76      0.53      1278
weighted avg       0.94      0.74      0.81      1278
```



Final Recommendations

- The logistic regression model was able to predict the probability of stroke with an accuracy of 74%
- The false negative rate for this model is still present and may require patients to be examined manually to pick up any indicators that may have been overlooked by the model.
- Patients that are older should take extra as they may be at greater risk for having a stroke. This may be amplified in patients that suffer from heart disease, high blood pressure, or high blood glucose levels.
- A patients work type may also play a role in indicating whether they are at risk of a stroke or not.