Healthcare dataset

STROKE OR NOT?

Who are our stakeholders?

- In the context of the problem presented, our stakeholders would be doctors and healthcare professionals.

What Problem are we solving?

- The model created allows the stakeholder to predict the likelihood of a stroke in patients dependent on various factors such as : Smoker status, Age, Gender, Etc.

Introduction to data

- 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

countplot for stroke VS work_type
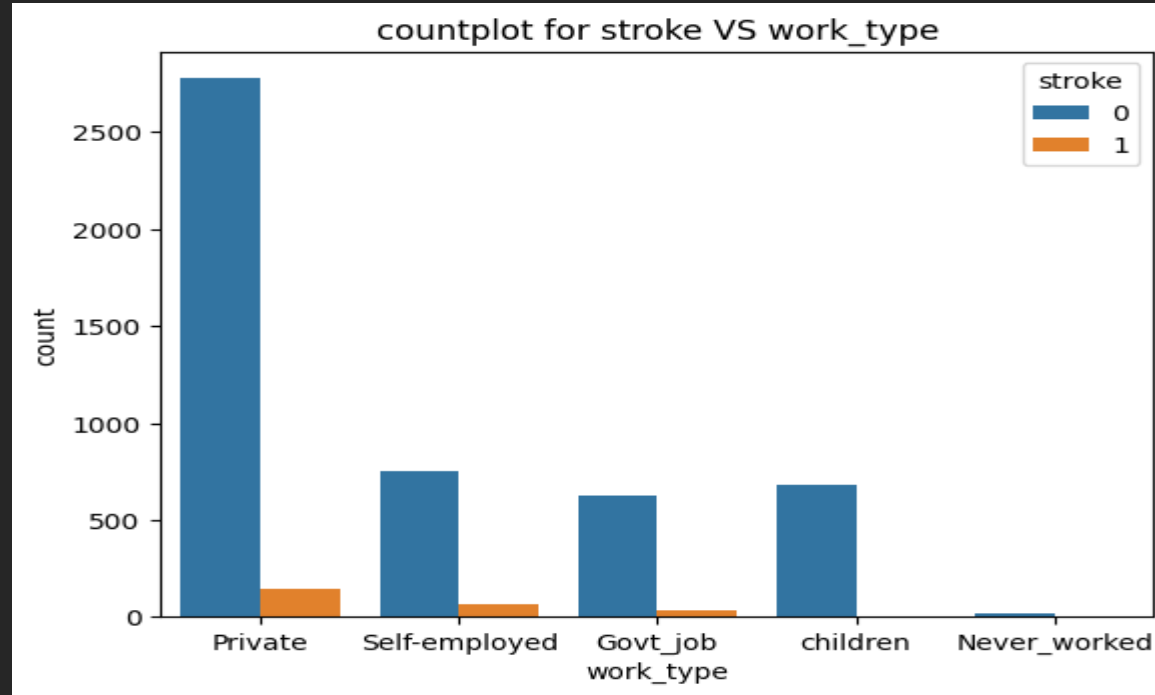
- The above depicted graph displays a count of predicted stroke VS individual work type

- As we can see the private sector has the highest as well as the lowest count for stroke and no stroke respectively.

Visual: 2
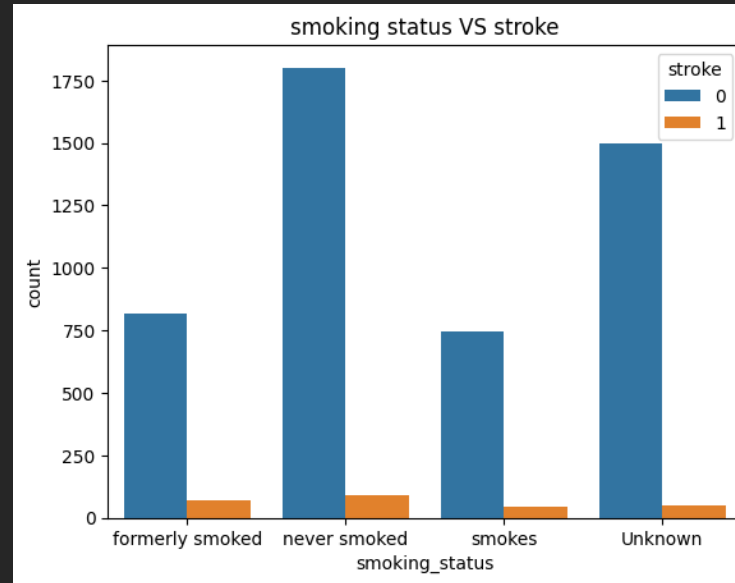
- Looking at the above visual, we can see a count plot for smoker status VS stroke(yes or no.)

- The graph shows us that individuals who have never smoked have the highest likelihood as well as the lowest for stroke and no stroke.

Strengths and limitations of my model

- The model chosen for production would be a decision tree classifier that is making use of PCA to reduce the dimensions or complexity within the data.

- 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!

- False negatives would mean that we would have patients classified as not being likely to have a stroke when they may be at risk!

- False positives would classify patients that are not likely to have a stroke as being at risk!

Final Recommendations

- My final recommendation that I feel could improve the quality of data as well as model predictions would be obtaining more data to try and balance our classes.

- Another action that can be taken may be to try and make use of a different model?

- The model had low precision and recall(correct positive/negative predictions) rates based on the technical analysis, I feel the imbalance in classes plays a big role in this. We could try and increase the weighting on our lesser class so that our model counts each prediction for that class as being greater than a count for the other class.