# Stroke Prediction In the Hospital Setting

Is it Possible?

# Introduction to the Data

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

# Introduction to the Data

❖    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 each patient.

# Introduction to the Data

❖ The American Heart Association estimates the mean lifetime cost of a stroke per person to be $228,030 for SAH, $123,565 for ICH, and $90,981 for ISC.  The average across all individuals and types of stroke, the lifetime cost per person is $130,576.
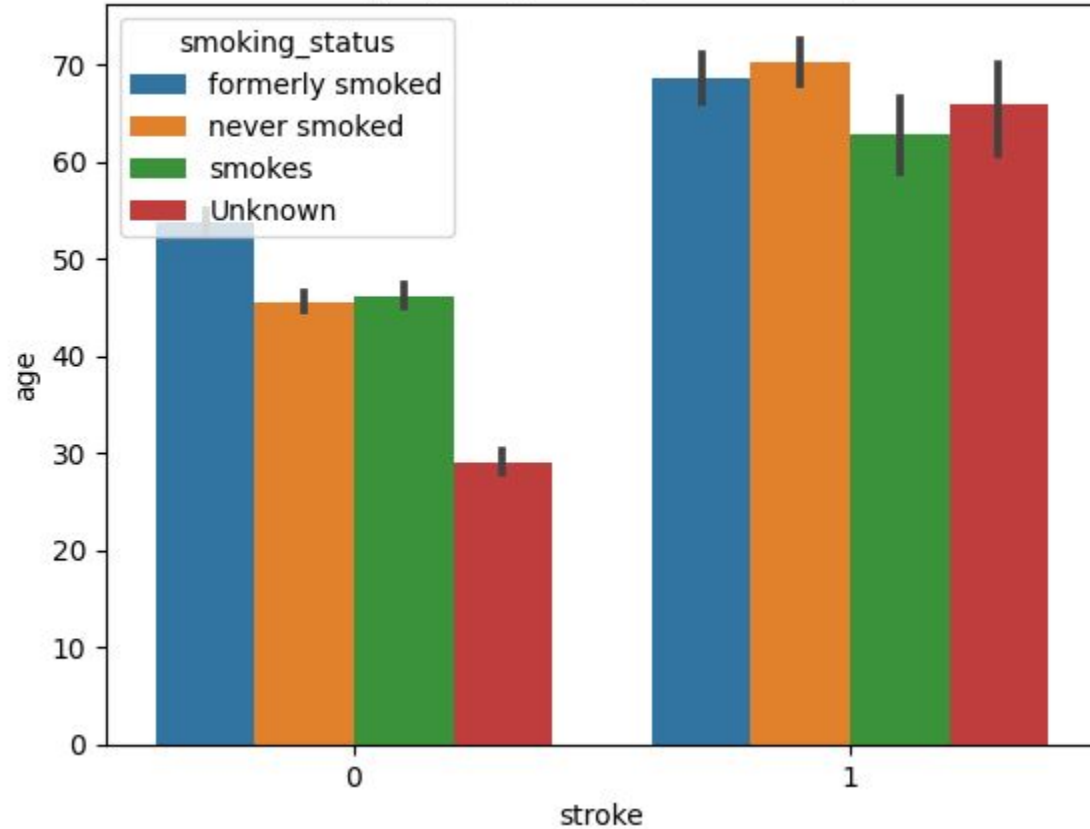
# Stroke by Age
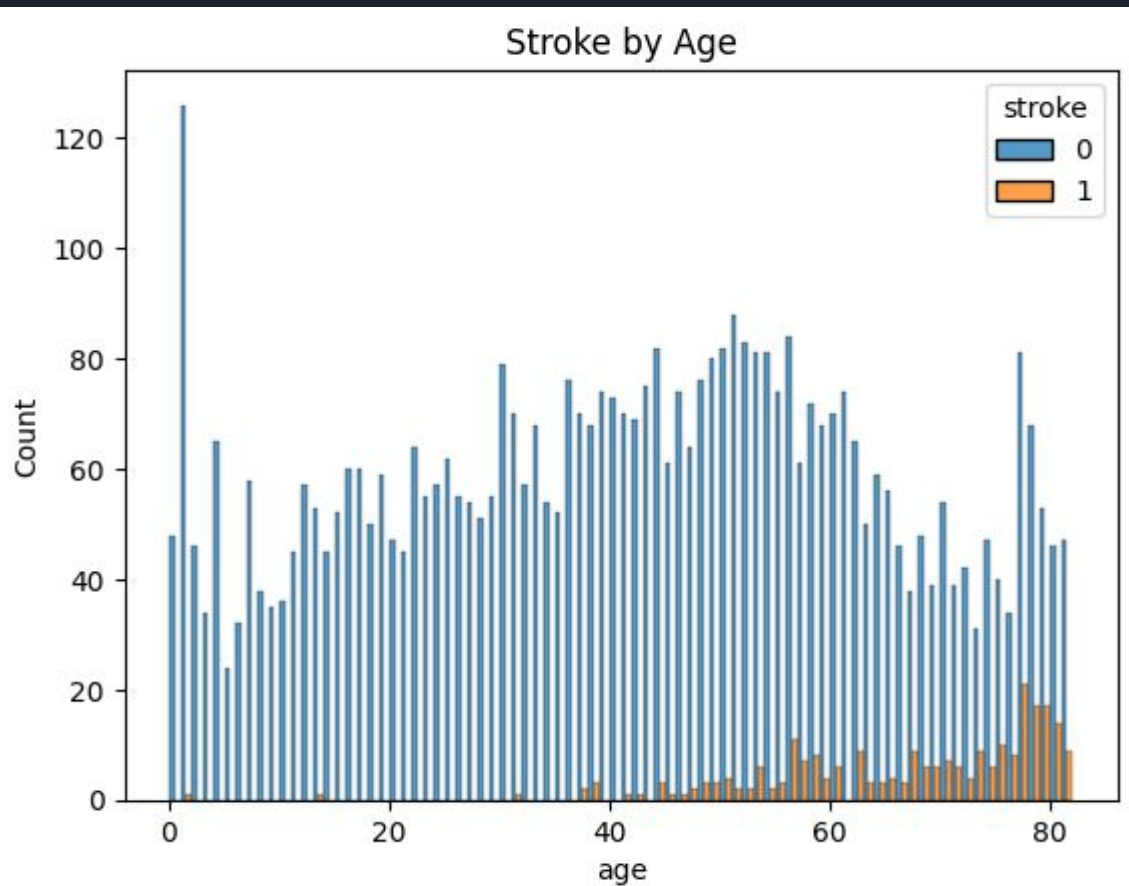
# Previous Graph

❖ The previous graph shows the correlation of strokes with age. Most of the strokes take place after the age of 40. There are some rare strokes under the age of 40. The majority of the strokes happened around the age of 80.

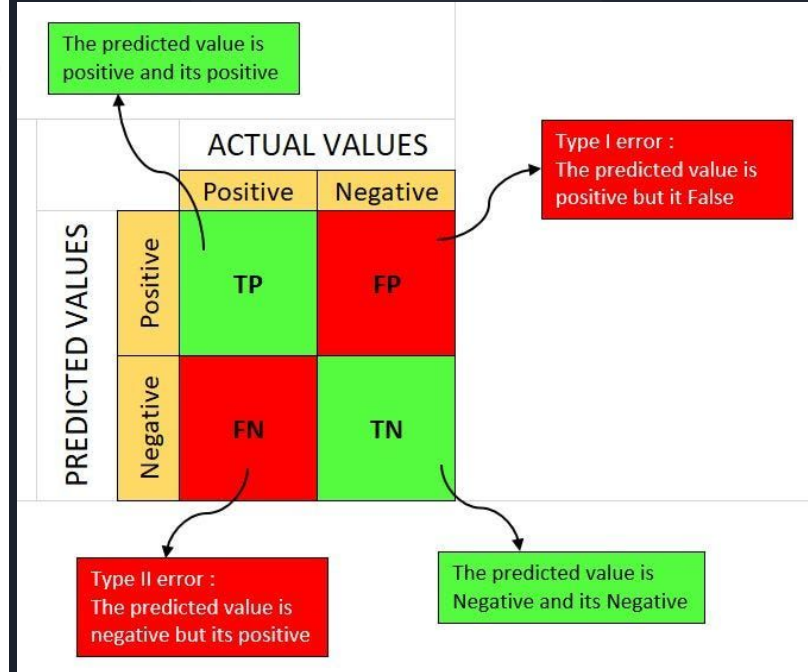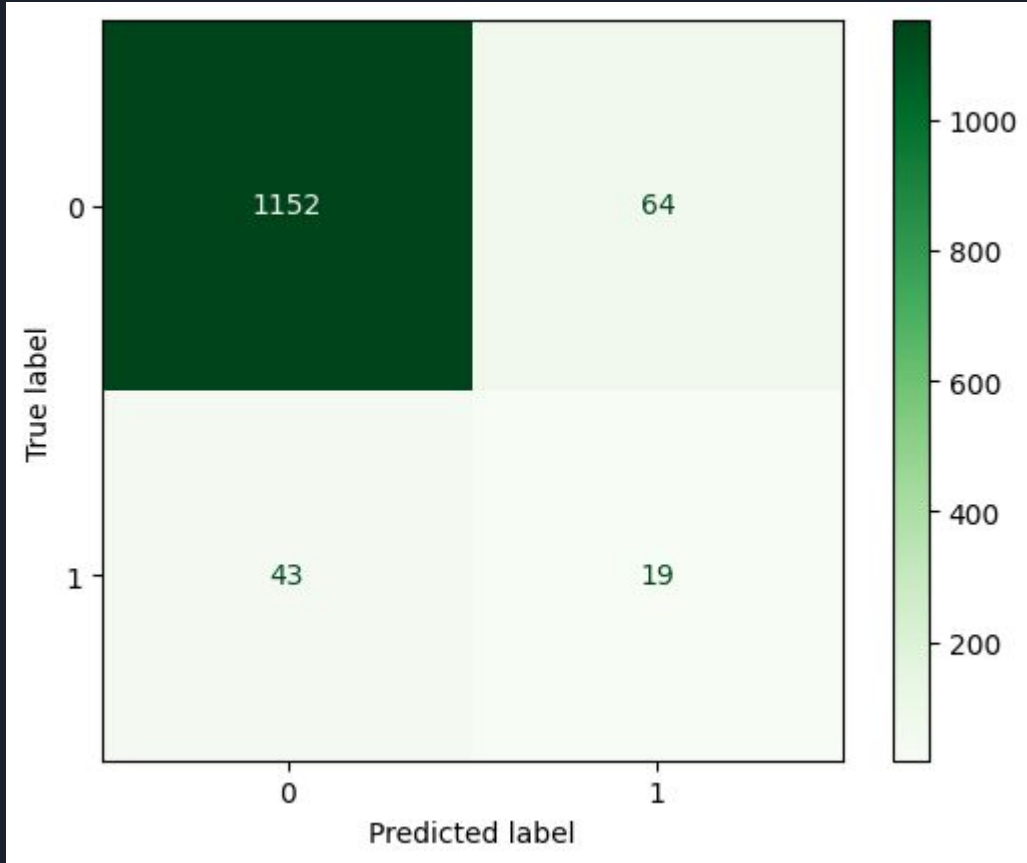Stroke By Age depending on Smoking Status

# Previous Graph

❖ The previous graph shows the correlation of strokes by age and smoking history.  Most of the strokes take place between the age of 60 and 70.  The best thing this graph shows is how not smoking puts off the onset of stroke by almost ten years.

# Best Model

# False Negatives

❖ **False Negatives**: When it comes to our model False Negatives are the worst outcome.  This is a prediction of a patient not having a stroke but they do.  This would affect the stakeholders negatively because of the cost of taking care of this stroke patient.

# False Positives

❖ **False Positives**: When it comes to this model False Positives are second worst outcome. This a prediction of a patient having a stroke but they do not. This would affect the stakeholders negatively just on reputation of the model alone. Just some happy patient's after getting the worst news possible.

# Final Recommendation

This data was very unbalanced. I had to use boosting techniques to increase the dataset for better modeling. Over sampling along with SMOTE and Logistic Regression is the best model that I found. This model ended up with decent accuracy of about 92%.