**Stroke Prediction & Machine Learning**

In this project, I will perform explanatory data analysis on the Stroke Prediction Dataset and train a model to predict whether a patient is likely to experience a stroke based on several factors, such as gender, age, bmi, and more.

TABLE OF CONTENTS

1. A GRAPH TO SHOWING AGE GROUP DISTRIBUTION

2. A GRAPH TO SHOWING GENDER DISTRIBUTION

3. A GRAPH TO SHOWING HYPERTENSION DISTRIBUTION

4. A GRAPH TO SHOWING HEART DISEASE DISTRIBUTION

5. A GRAPH TO SHOWING WORK TYPE DISTRIBUTION

6. A GRAPH TO SHOWING SMOKING STATUS DISTRIBUTION

7. A MODEL FOR THIS DATA SET TO SEE IF A PATIENT HAS STROKE OR NOT, THE ACCURACY SCORE IS NOT LESS THAN 96%

1. A GRAPH TO SHOWING AGE GROUP DISTRIBUTION

Amount of Patients in Each Age Group

Ages 0-10: 507

Ages 10-20: 518

Ages 20-30: 545

Ages 30-40: 674

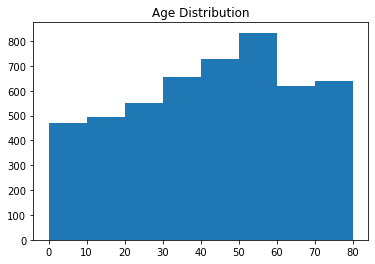
Ages 40-50: 739

Ages 50-60: 823

Ages 60-70: 594

Ages 70-80: 594

Ages 80+: 116



Observation:

The age distribution histogram shows that the age data is slightly skewed left, with a greater number of patients from older age groups. The statistics indicate that there are 507 patients ages 0-10, 518 patients ages 11-20, 545 patients ages 21-30, 674 patients ages 31-40, 739 patients ages 41-50, 823 patients ages 51-60, 594 patients ages 61-70, 594 patients ages 71-80, and 116 patients older than 80. Thus, the age group with the largest number of patients is the 50-60 group, followed behind by the 40-50 group.

2. A GRAPH TO SHOWING GENDER DISTRIBUTION



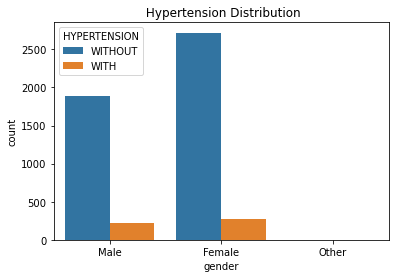
Percentage of Females: 0.5859099804305283

Percentage of Males: 0.41389432485322897

Observation:

From the histogram, we can see that there is a greater amount of female patients than male patients in the dataset. In particular, 58.60% of the dataset is made up of females, 41.40% is made up of males.

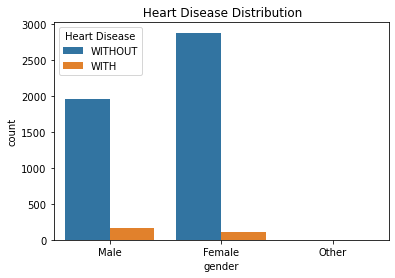
3. A GRAPH TO SHOWING HYPERTENSION DISTRIBUTION



Observation:

From the histogram, we can see that the amount of patients with hypertension is significantly less than those with hypertension. Based on the size of the bars, we can estimate that there are nine times more patients without hypertension than with hypertension.

4. A GRAPH TO SHOWING HEART DISEASE DISTRIBUTION



Observation:

Similarly to the hypertension distribution, it is clear that there are substantially more patients without heart disease than with heart disease in the dataset.

5. A GRAPH TO SHOWING WORK TYPE DISTRIBUTION

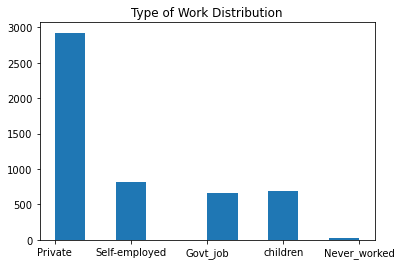
Private: 2925

Self Employed: 819

Government Job: 657

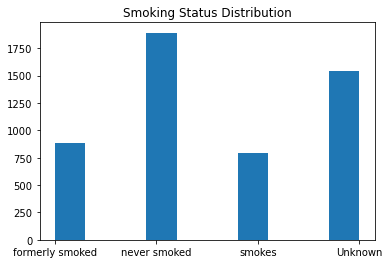
Children: 687

Never Worked: 22



Of the patients in the dataset, there is the greatest amount of patients with a private work type and the least amount of patients with no work experience. In fact, 2925 patients have a private work type, whereas 22 patients have never worked. In between the two extremes, there are 819 self-employed patients, 687 patients with children, and 657 patients with government jobs.

6. A GRAPH TO SHOWING SMOKING STATUS DISTRIBUTION



Observation:

Of the patients whose smoking status is known, the greatest amount are those who have never smoked, with lesser but similar amounts of patients that have either formerly or currently smoke. There is also a substantial amount of patients with an unknown smoking status.

## Model Creation & Selection

Given our training and testing sets, we will now try several models to see which results in the greatest accuracy. Specifically, we will try logistic regression, the RandomForest classifier, the AdaBoost classifier, the GradientBoosting classifier, and the KNeighbors classifier.

USING LOGISTIC REGRESSION.

Accuracy: 1.0

Precision recall f1-score support

0 1.00 1.00 1.00 975

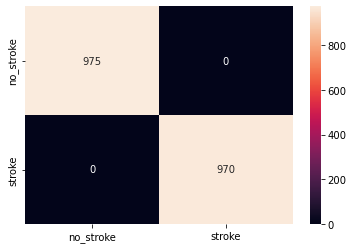
1 1.00 1.00 1.00 970

Accuracy 1.00 1945

Macro avg 1.00 1.00 1.00 1945

Weighted avg 1.00 1.00 1.00 1945

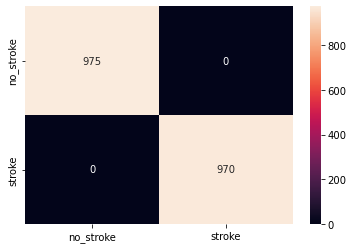
TESTING ACCURACY:



USING RANDOMFOREST CLASSIFIER.

Accuracy: 1.0

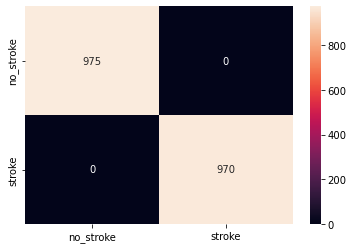
TESTING ACCURACY:



USING ADABOOST CLASSIFIER.

Accuracy: 1.0

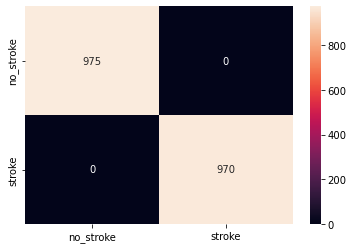
TESTING ACCURACY:



USING GRADIENTBOOSTING CLASSIFIER.

Accuracy: 1.0

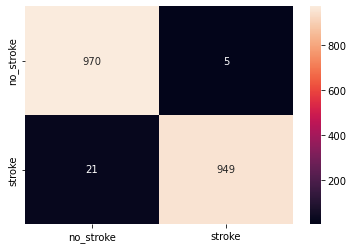
TESTING ACCURACY:



USING KNEIGHBORS CLASSIFIER.

Accuracy: 0.9866323907455012

TESTING ACCURACY:



## Results

We observe that the KNeighbors classifier results in the normal accuracy score of 98.6%, thus indicating that such a model is the most appropriate choice for determining whether a patient is likely to experience stroke.