Diagnosing Patient's Stroke Risk Using Health Data

I. Introduction

A. Problem Statement

According to the CDC, 1 in every 6 deaths from cardiovascular disease is due to stroke -- our goal is to create a model that helps inform a patient about their personal risk for having a stroke. This model could take a generalized process and produce something much more personal for each patient.

B. Background

A common complaint about the current state of the healthcare system in the US is the lack of personalization and the overwhelming amount of generalization occurring from patient to patient. With statistics such as someone having a stroke every 40 seconds in the US, it seems like the obvious choice to make sure we are taking the time to accurately diagnose a patient's risk for stroke so that the patient can take the appropriate steps to reduce their risk.

C. Goal

This model aims to create a personalized approach to diagnosing a patient prior to the occurrence of a potentially fatal disease. The goal is to be able to gather data from a patient's chart and then have a simple model that outputs whether or not a patient is at risk for a stroke and in which percentile their risk level lies.

References:

1) *https://www.cdc.gov/stroke/facts.htm*

II. Datasets

A. The stroke prediction dataset contains patient information pulled from a kaggle dataset. The dataset aims to predict whether a patient is likely to suffer a stroke based on input parameters such as gender, age, hypertension, smoking status, and a few lifestyle type statuses. We pulled in the data from .csv format and read it in as a DataFrame.

Categorical Data	Numerical Data
• Gender • Hypertension • Heart Disease • Ever Married • Work Type • Residence Type • Smoking Status • Stroke Occured	• Age • Average Glucose Level • BMI

III. Data Cleaning and Data Wrangling

A. Data from this dataset did not require any cleaning. The data was read in as a DataFrame with 5110 rows and 12 columns. Most work was done during Exploratory Data Analysis.

IV. Exploratory Data Analysis and Initial Findings

A. Missing Values

The primary step taken during Exploratory Data Analysis was to address missing values found in the data set.

a. BMI

To deal with the 3.93% of missing values in the BMI column, we imputed the average BMI of 28.89. This imputation did not change the overall mean of the column but did bring a slight shift to the standard deviation bringing it from 7.85 to 7.70. The shift was considered negligible.

B. Distributions

The distribution for each feature was checked for both stroke positive and stroke negative patients.

a. Marital Status

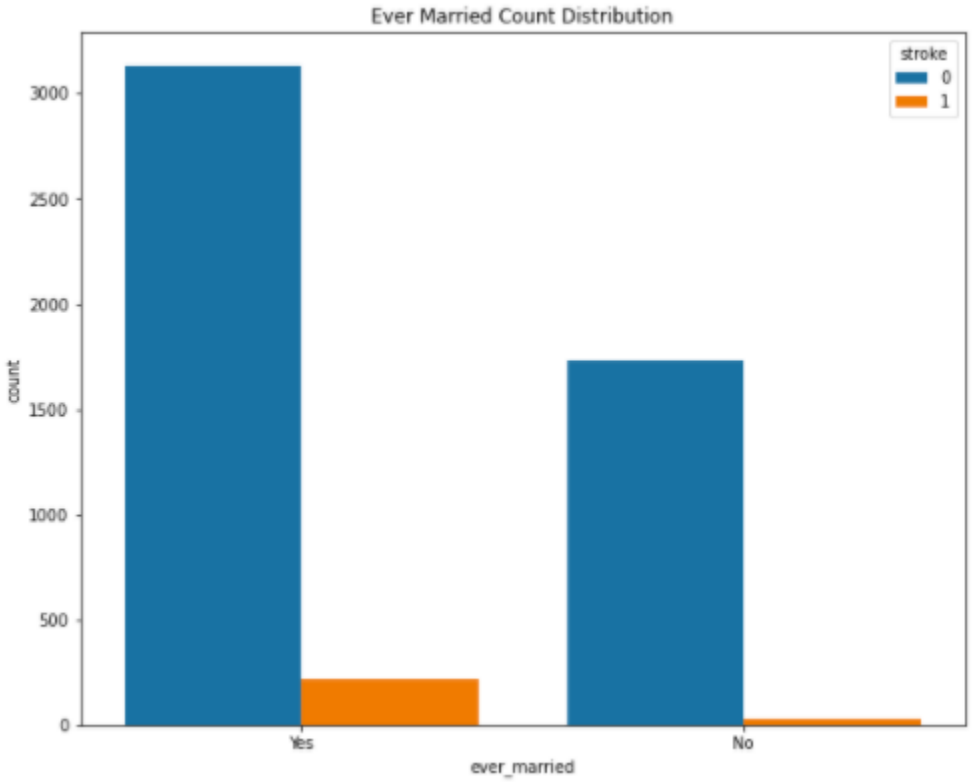


Figure 1. Marital Status count plot for both stroke positive and stroke negative patients.

The marital status distribution comparison showed that more people in the stroke positive group were married (88%); this was also true for the stroke negative patients but at a lower percentage (64.5%). Resulting in a wider difference between groups in the stroke positive group.

b. Work Type

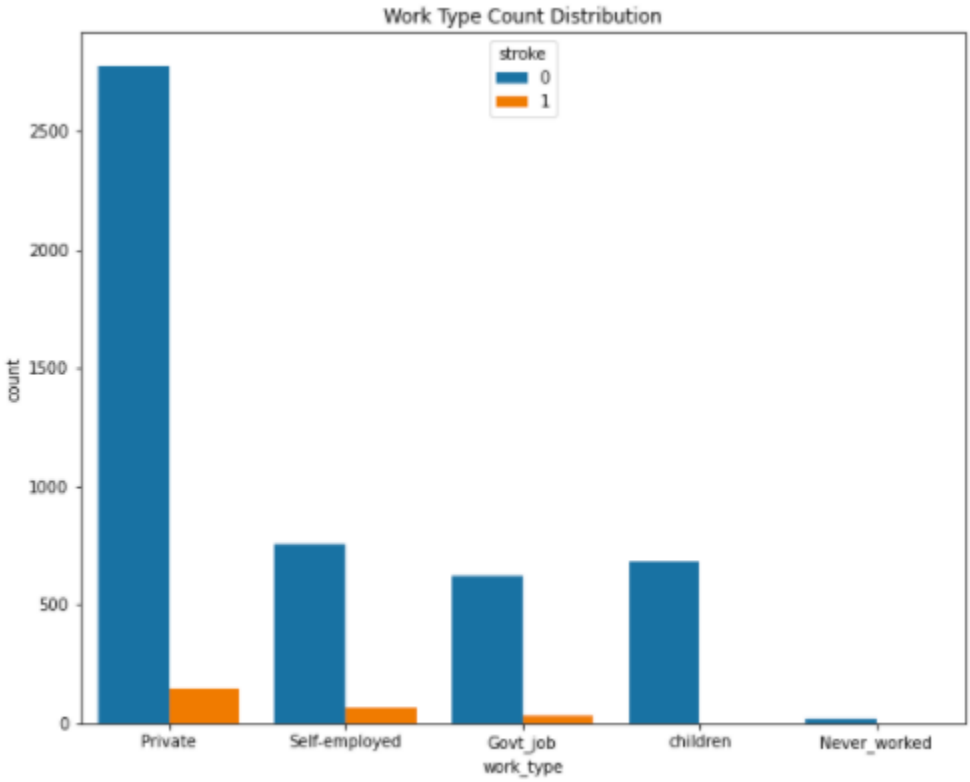


Figure 2. Work Type count plot for both stroke negative and stroke positive patients.

For both the stroke positive and stroke negative groups, patients who worked a private job were the highest population -- 59.8% for the stroke positive group and 57.1% for the stroke negative group. The private job employees were followed by those who were self-employed. This distribution was expected since most of the general population tends to be privately employed. The most significant difference between the two distributions was that none of the patients from the “Never worked” population appeared in the stroke positive group. This difference could be due to such a small sample of those who had never worked (only 0.43% of the entire dataset’s population).

c. Residence Type

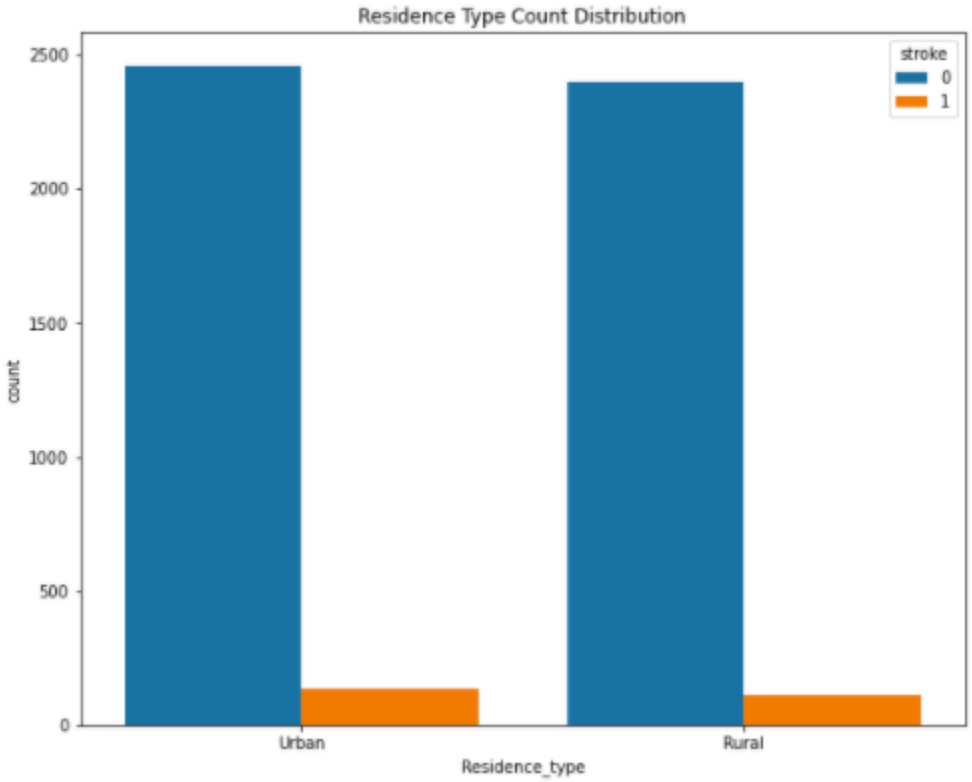


Figure 3. Residence Type countplot for both stroke negative and stroke positive patients.

For both the stroke positive and negative groups, the Urban populations hold the majority -- 54.2% in stroke positive groups and 50.6% in the stroke negative groups.

d. Smoking Status

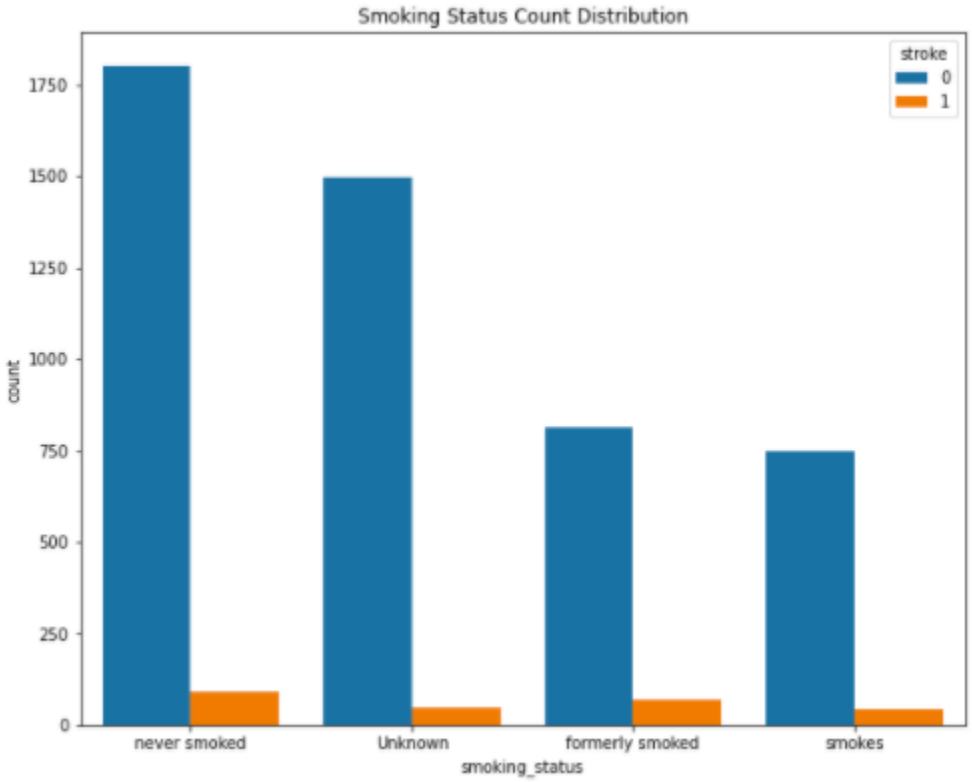


Figure 4. Smoking Status countplot for both stroke negative and stroke positive patients.

Patients that reported as “Never Smoked” held the majority in both populations -- 36.2% in the stroke positive group and 37.1% in the stroke negative group. Interestingly enough, the smallest population for the stroke positive group was the patients who reported as smokers (16.9%). This feature group did have 30.2% of its data categorized as “Unknown” which could play a part in the unexpected results found during EDA.

e. Gender

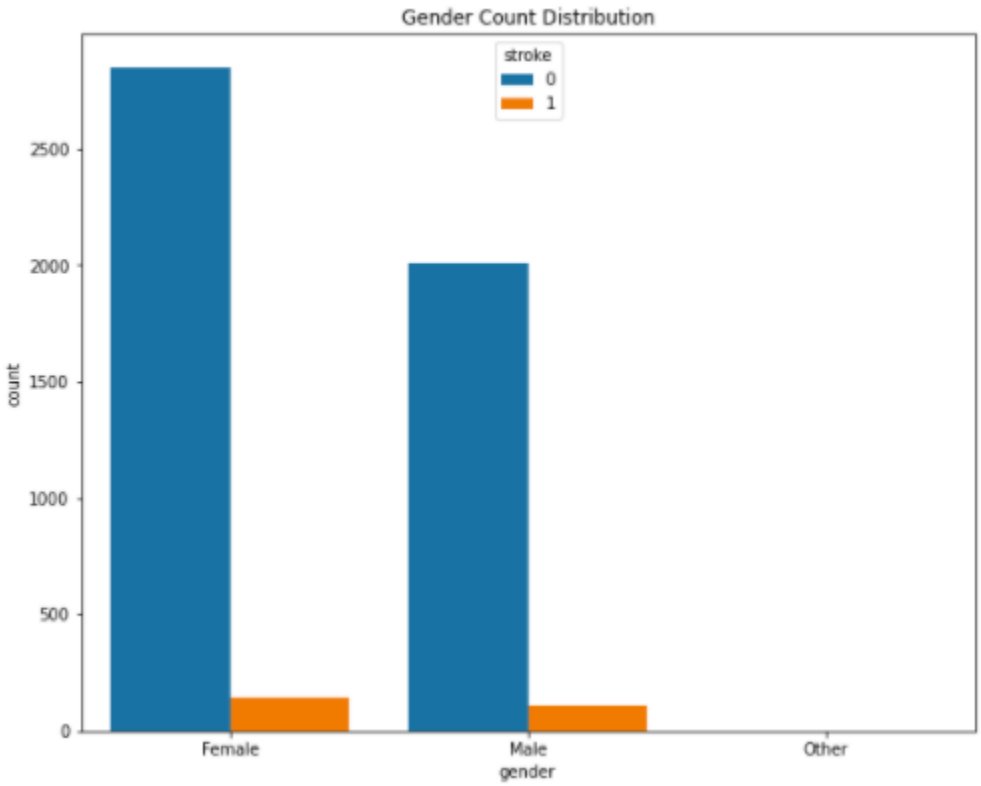


Figure 5. Gender countplot for both stroke negative and stroke positive patients.

Females were the majority in both the stroke positive and stroke negative groups -- 56.6% in the stroke positive group and 58.7% in the stroke negative group. This feature group was imbalanced with 17% more females than there were males.

f. Age

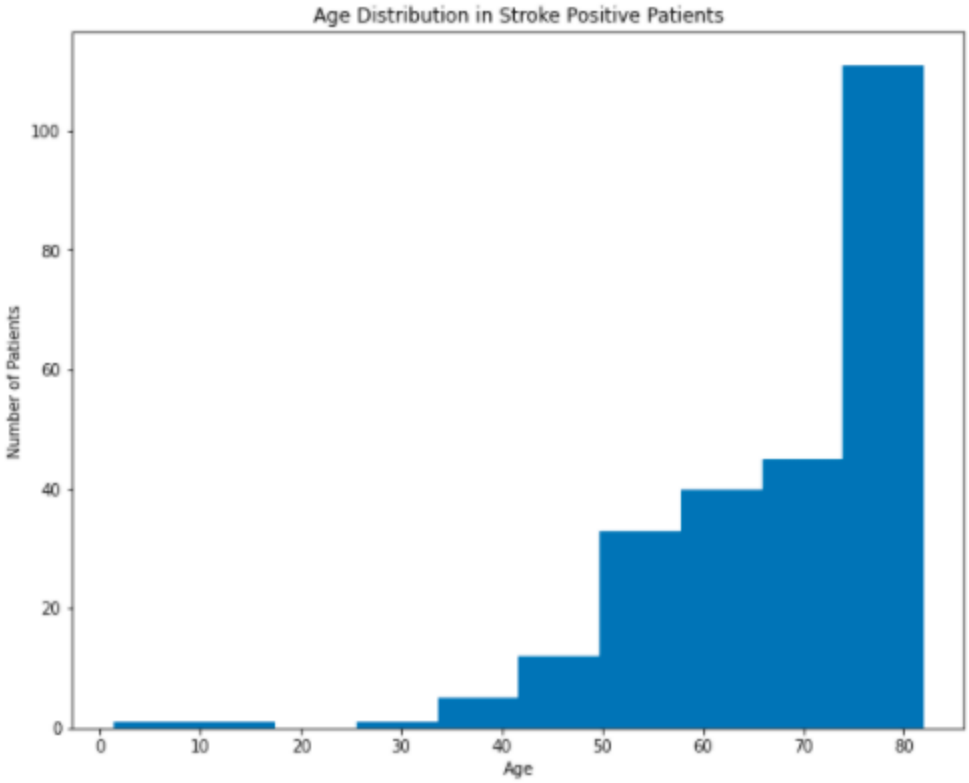


Figure 6. Age countplot for both stroke negative and stroke positive patients.

Patients who reported as stroke positive were generally found in the older groups. Distributions showed that 75% of patients in the stroke positive group were above the age of 59 with an average age of 68 years.

g. Average Glucose Levels

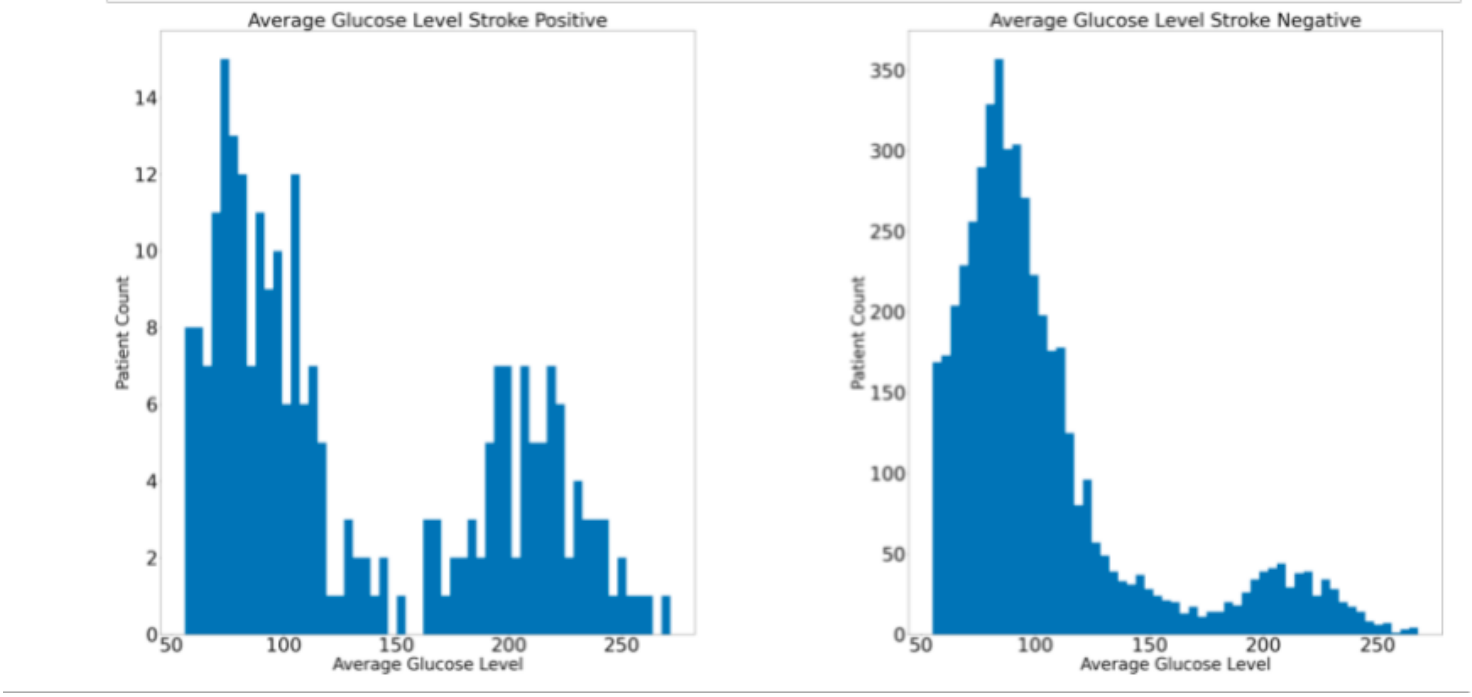


Figure 7. Average Glucose Level distribution for both stroke positive and stroke negative patients.

For our dataset as whole, the average value fell at 132.54 which is a normal glucose reading. 24.6% of patients were within the diabetic range (75th percentile = 196.71) and 10.8% of patients were within the prediabetic range (25th percentile = 79.9). When analyzing patients from the stroke positive and stroke negative groups separately, it was found that 48.1% of stroke positive patients had abnormal glucose readings.

h. BMI

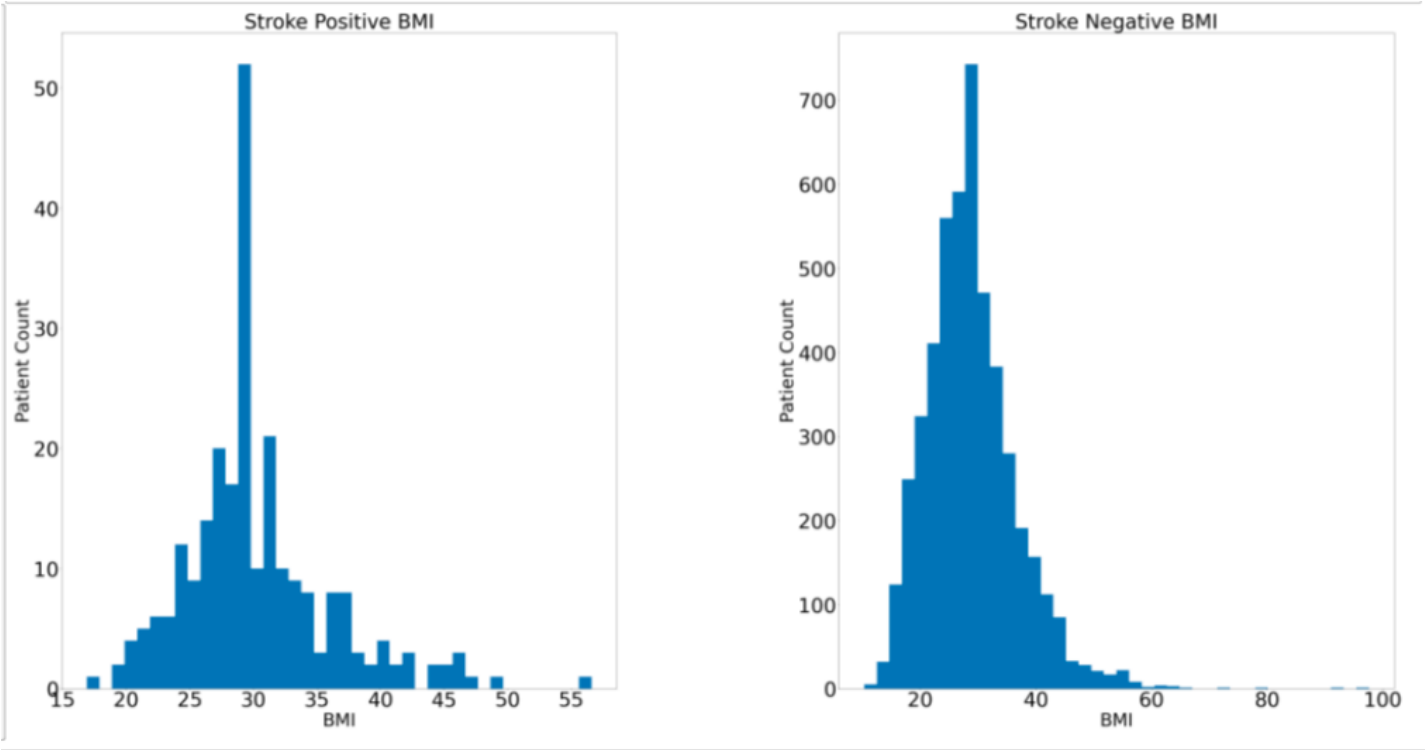


Figure 8. BMI distribution for both stroke negative and stroke positive patients

A majority of stoke positive patients were reported to be either overweight or obese (84.7%) with a larger percentage (46.2%) of this group falling within the overweight range (BMI 25.0-29.9).

i. Hypertension

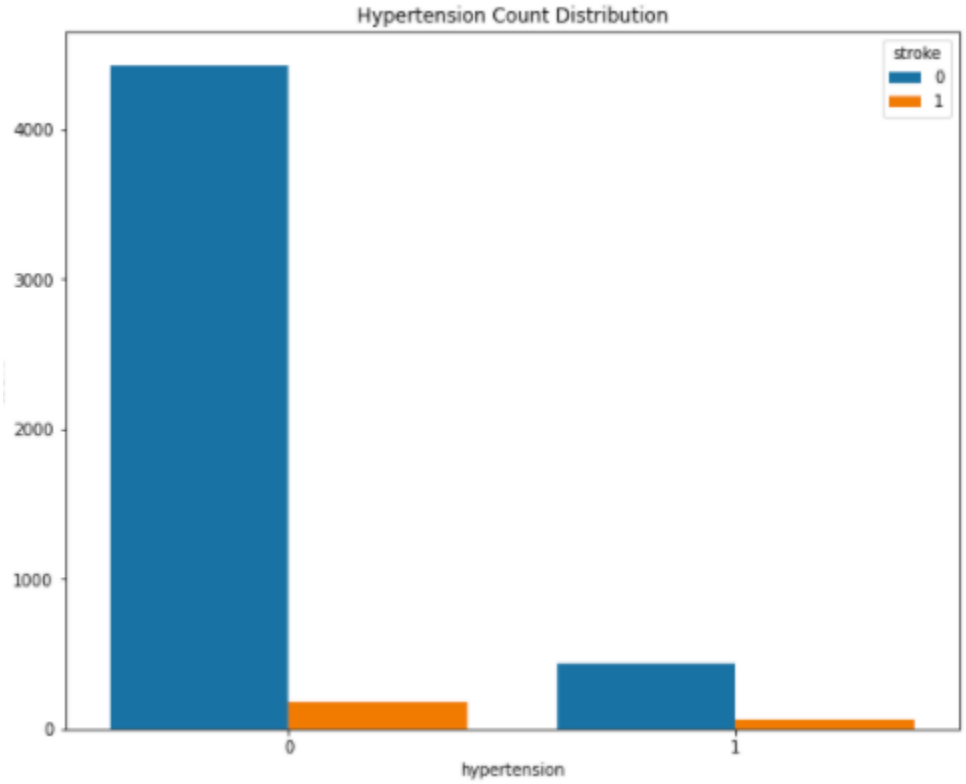


Figure 9. Hypertension countplot for both stroke negative and stroke positive patients

There was a noticeable difference between stroke positive patients with hypertension and stroke negative patients with the same diagnosis. While the majority for both groups was those who did not have hypertension, the distribution of people in the stroke positive group with hypertension showed as 26.5%, while those with hypertension in the stroke negative group only reported at 8.8%.

j. Heart Disease

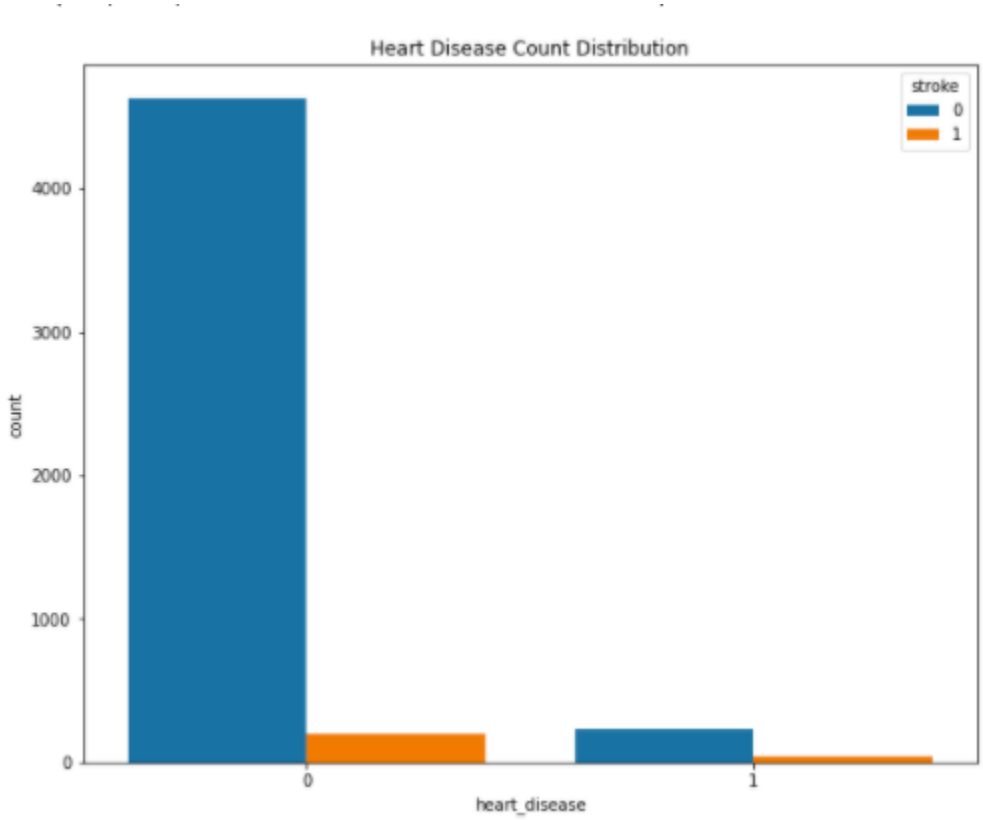


Figure 10. Heart Disease countplot for both stroke negative and stroke positive patients.

There was a noticeable difference in the heart disease distributions as well. 18.9% of patients with heart disease appeared in the stroke positive group and only 4.7% lie in the stroke negative group.

C. Statistical Significance

The features below were tested and proven to have statistical significance in the determination of a patient's stroke risk:

Statistically Significant Features
• Work Type • Smoking Status • Marital Status • Hypertension • Heart Disease • Age • BMI • Average Glucose Levels

The only two features that did not show statistical significance were Gender and Residence Type. Since there is already research showing gender is an important factor in determination of stroke risk, it was not dropped from the dataset. A model was built with the Residence Type feature included as well as a model without it included -- overall, performance was better when the Residence Type feature was kept within the data.

V. Data Pre-Processing

A. Missing Values

The two columns that did not show statistical significance, Gender and Residence Type, needed to be addressed during data pre-processing. We did not feel comfortable making the assumption that Gender played no role in the determination of stroke risk, but Residence Type did raise some eyebrows. Two dataframes were created -- one that maintained the presence of the Residence Type column and one with the Residence Type column dropped. These were both used on our model and the results were compared to determine how to move forward.

B. Binary Encoding

Indicator variables were created using a binary encoder since the data was not composed of ordered variables. Binary encoding helped to keep dimensionality down and to aid in saving space in terms of memory.

C. Train Test Split and Normalizing Numerical Features

A train and test set were created utilizing a test size of 30% for both of our dataframes. Our data was normalized following the split into training and testing sets to avoid any data leakage issues.

VI. Modeling

A. Choosing a Model

Three model's were tested to ensure we were choosing the best scoring option. Of the three models tested (Random Forest Model, Logistic Regression Model, and Naive Bayes Model), we chose to move forward with a Random Forest Model.

B. Parameter Tuning

We utilized Randomized Search to find the best parameters for using the Random Forest Classifier on our data. The best parameters are as listed:

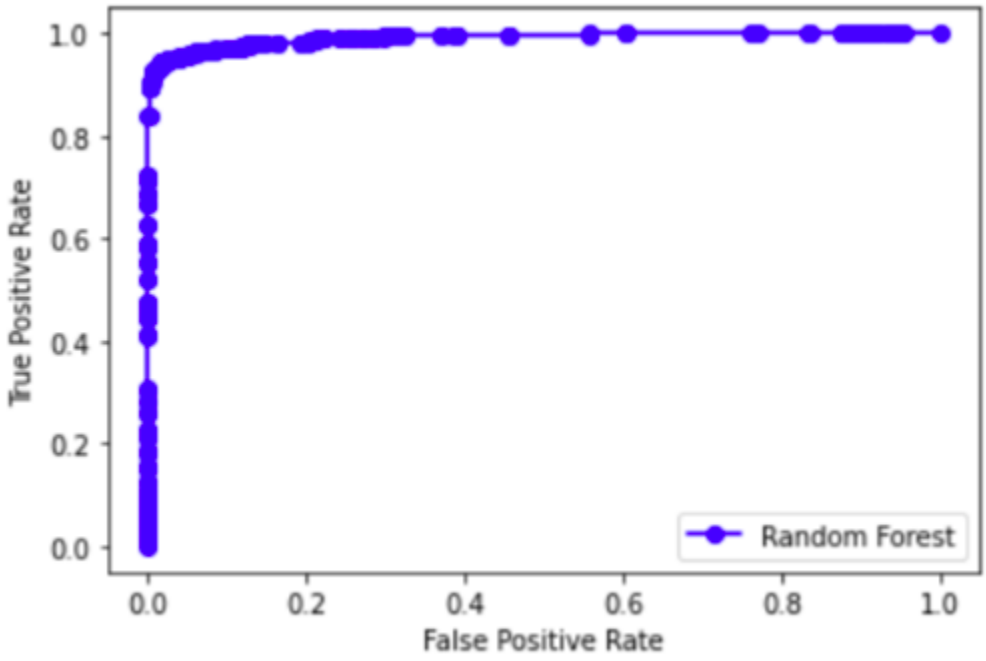
- n_estimators = 2500
- min_samples_split = 5
- min_samples_leaf = 1
- max_features = 'auto'
- max_depth = 30

C. Results

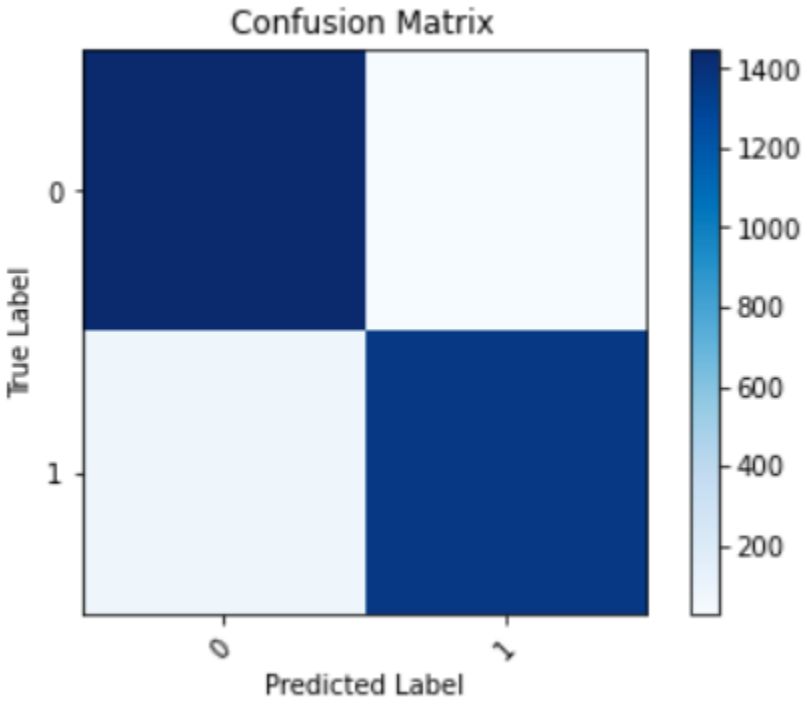
Using the tuned parameters yielded the following results:

ROC/AUC Score:

0.992



Confusion Matrix Results:
True Negative: 1449
False Positive: 26
False Negative: 80
True Positive: 1362



The above results were produced while keeping the Residence Type column within the dataframe.

VII. Future Work

Since all of this data is generally already collected from most patients that are seen by a doctor, it could become a model that is integrated in a normal chart processing system that can alert a doctor about a patient's risk level. The idea is to work on preventative care instead of risking fatality or post disease treatment.