**Stroke Prediction Models Using Machine Learning**

Sofia Mirshed

Fullstack Academy

April 2023

Abstract

Using a dataset of patient attributes including known risk factors and demographics to determine the likelihood of stroke. Strokes often have detrimental and permanent damage as a result, a predictive model would ideally be used to take preventative measures to lower the risk of stroke. Machine learning algorithms are used to build predictive models and models are compared by reviewing the accuracies, confusion matrices, and feature importance to determine the best fitting model.

Introduction

A stroke is when blood flow to the brain is interrupted by a clot, a bleed, or in some cases both. According to the Center for Disease Control (CDC), every 40 seconds, someone in the United States has a stroke. Every 3.5 minutes, someone dies of stroke. Strokes are not only a prevalent health issue, but a deadly one as well, especially for people over 65 years old.

High [blood pressure](https://www.cdc.gov/bloodpressure/index.htm), [high cholesterol](https://www.cdc.gov/cholesterol/index.htm), smoking, obesity, and diabetes are leading [causes of stroke](https://www.cdc.gov/stroke/risk_factors.htm). One in 3 U.S. adults has at least one of these conditions or habits.

ASCVD (Atherosclerotic Cardiovascular Disease) is the currently most used risk estimator for cardiovascular disease. This estimator determines 10-year risk of cardiovascular disease and includes: age, diabetes, sex, smoker, total cholesterol, hdl cholesterol, systolic blood pressure, treatment for hypertension, and race.

There is a great need and application for accurate prediction models in healthcare.

The goal of this project was to build a predictive model, much like the ASCVD but specifically for stroke. A predictive model would be a tool used to aid in early intervention and prevention of stroke.

Dataset Description

The dataset used is from Kaggle and has a usability score of 10.0 and contains 12 attributes and 5110 observations (11 health and demographic attributes). The attributes are described as follows:

1) **id**: unique identifier  
2) **gender**: "Male", "Female" or "Other"  
3) **age**: age of the patient  
4) **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension  
5**) heart\_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease  
6) **ever\_married**: "No" or "Yes"  
7) **work\_type**: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"  
8) **Residence\_type**: "Rural" or "Urban"  
9) **avg\_glucose\_level**: average glucose level in blood  
10) **bmi**: body mass index  
11) **smoking\_status**: "formerly smoked", "never smoked", "smokes" or "Unknown"\*  
12) **stroke**: 1 if the patient had a stroke or 0 if not

\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

Most features were categorical data except for id, age, avg\_glucose\_level, and bmi which were numerical.

Method

Data Wrangling and Cleaning

In this dataset, there were 201 null values found for BMI. Because this dataset is rather small and it does not have many null values, it was decided to fill the nulls with the mean in order to preserve as much data as possible to strengthen the models. The id field was dropped because it was not necessary data for the purposes of this project.

The categorical data needed to be converted to numerical and sorted which was accomplished within Excel. The data was sorted by assigning higher number values for higher implied importance as follows:

**gender**: “male” = 0, “female” = 1

**ever\_married**: “no” = 0, “yes” = 1

**work\_type**: “Never\_worked” = 0, “children” = 1, “Govt\_job” = 2, “Self-employed” = 3, “Private” = 4

**Residence\_type**: “Rural” = 0, “Urban” = 1

**smoking\_status**: “Unknown” = 0, “never smoked” = 1, “formerly smoked” = 2, “smokes” = 3

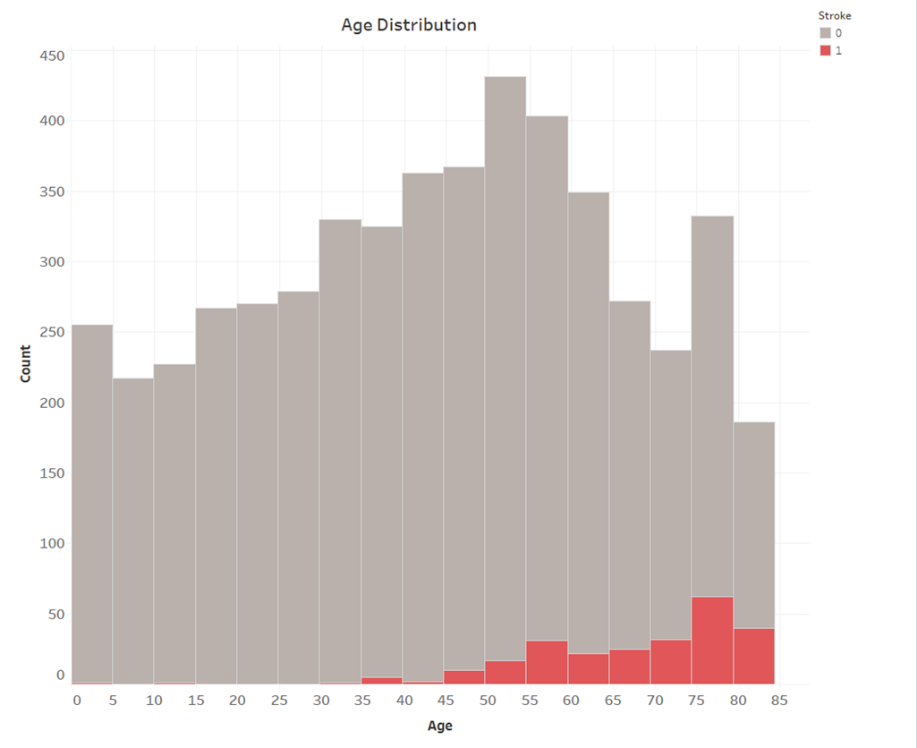
The higher implied importance was determined after an exploratory data analysis was conducted to determine higher instances of stroke by each attribute. The results of the analysis are shown in the following section.

Exploratory Data Analysis

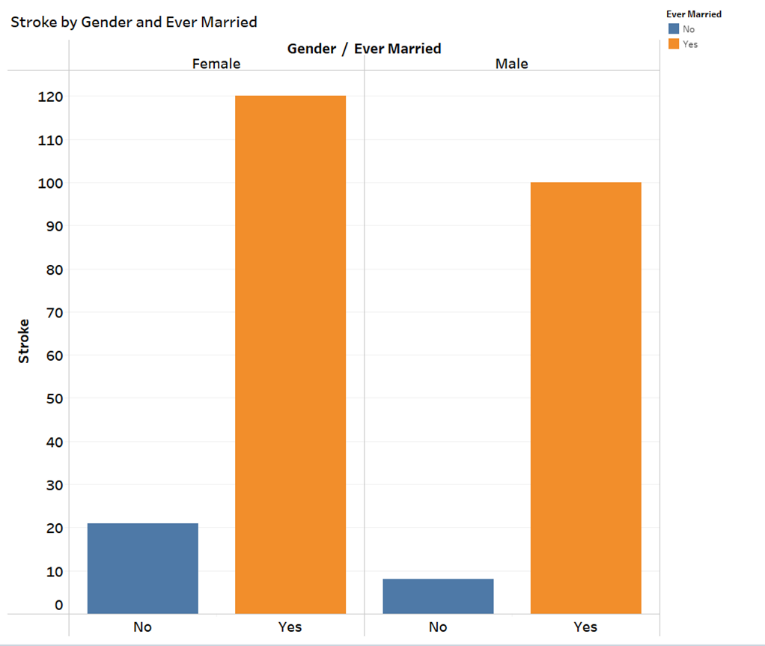
A correlation matrix was then created using Python to visualize the strength of relationships between each of the attributes. Looking at stroke specifically, it showed that age had the highest correlation with stroke in this dataset at a 0.25 correlation value. Then hypertension, heart disease and average glucose level were equally correlated with stroke with a 0.13 correlation value.



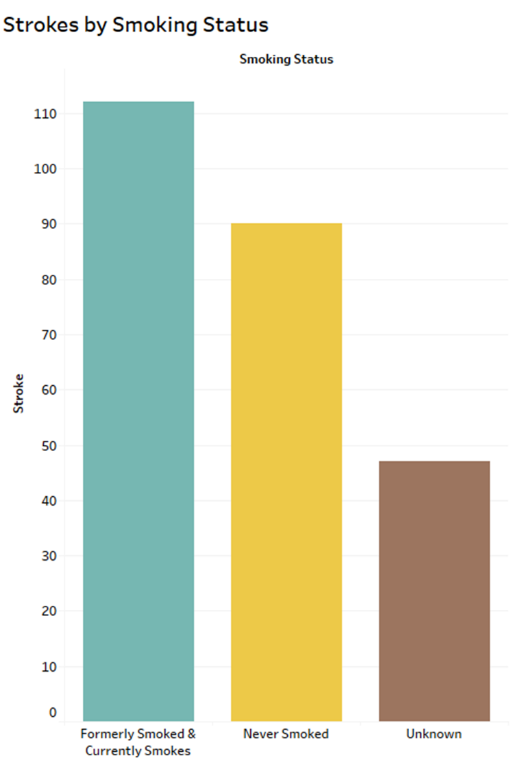
When looking at the distribution of age, it was noticed that the range was large from 0 to 82 years old. The age for all patients was normally distributed, but the distribution of age for those who experienced a stroke was skewed left as seen in red below. This confirms that older people tend to have strokes with the greatest distribution between ages 75 and 80 years old.



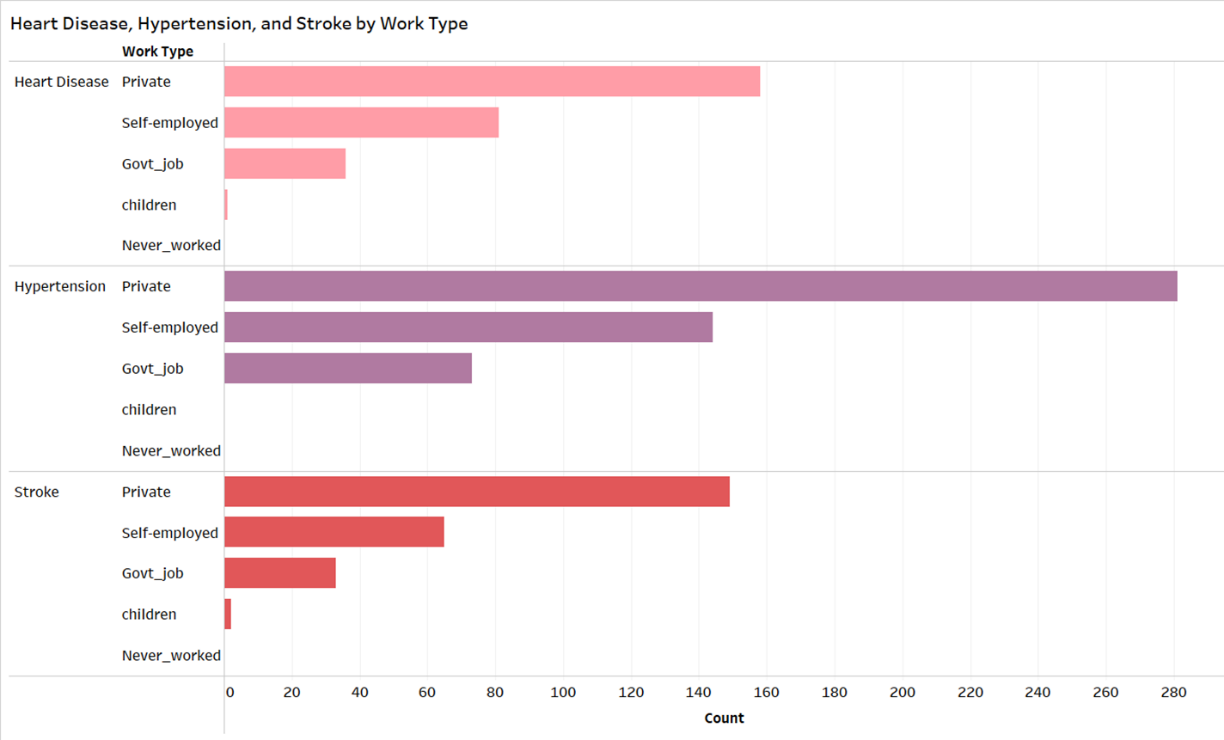
It is known that females tend to be at higher risk of stroke than males and upon an initial glance of the data, it was supported but not with much significance. There was however, a much higher instance of stroke for both males and females who have ever been married, compared to those who have not.



Smoking is a well-known health risk and is often deadly but most attributed to the risks of lung disease, especially lung cancer. Combining “formerly smoked” and “smokes” data was done to compare instances of stroke for those who have a history of smoking versus those without. This showed that any history of smoking has a higher instance of stroke than in those who have never smoked.



Looking at heart disease, hypertension, and stroke by work type gave some insight into health risks of working certain types of jobs. There was a higher prevalence of all three conditions in the private work sector which may be correlated with stress levels.



Machine Learning

Three different machine learning algorithms were used to build predictive models. These included logistic regression, random forest, and decision tree. Logistic regression had the highest accuracy at 94.12% which was too high to be correct. A regression model would not fit this data well because they work best for continuous outcomes. The outcome for this model, stroke is binary so a classifier model would work better for discrete outcomes. Random forest and decision tree are both classifiers. Random forest achieved an accuracy of 93.83% which is still very high, decision tree achieved a 90.61% accuracy.

Results

Reviewing the confusion matrices helped to determine which model worked best for this dataset and the purposes of this project. The goal was to build a model that would predict stroke, so the true positive section of the confusion matrix was what was focused on. Decision tree was the only model to have any true positives, meaning it was the only model that predicted any strokes correctly. Although the decision tree had the lowest accuracy, it is still rather high to be a good fitting model. It was the most precise in obtaining the intended result.

Chart

Description automatically generated

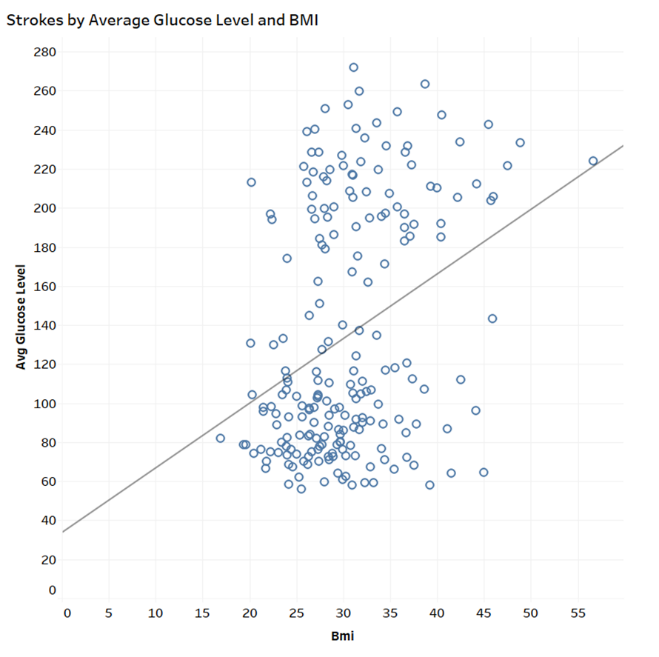
Discussion

Feature importance of the classifier models was reviewed to determine how the models were reading and using the attributes. Both models read average glucose level as having the highest importance for prediction, possibly because it is continuous data. The decision tree though, ranked age of greater importance over BMI which aligned with the earlier finding of age having a high correlation with stroke.

Chart, bar chart

Description automatically generated

Both classifier models ranked average glucose level and BMI as of high importance for prediction of stroke. There is a strong positive correlation between the two attributes which may explain why they are of similar importance in the models.



It was concluded that the decision tree was the best model for this dataset and had the strongest results. Although the accuracy was lowest, it was the only model to accurately predict any strokes. The feature importance also showed that average glucose level and age are features most important for predicting strokes within this dataset. As the CDC has said, diabetes is a leading cause of stroke. Stanford Health Care also states that most strokes occur in people 65 years old or older.

The accuracies of the models used were so high because the dataset had many limitations. The dataset was too small to run a strong predictive model. The age range was also very large and could have skewed the data. None of the attributes included in the data would account for the outliers that were children and young adults that had strokes, for example Leukemia. A decision tree would still be a good model to use for a larger dataset with more clinical attributes and more patients. It may also be strengthened if only data for older adults was used.

References

<https://www.cdc.gov/stroke/facts.htm>

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

<https://stanfordhealthcare.org/content/dam/SHC/clinics/stroke-center/docs/stroke-young-patients-qa.pdf>

<https://www.stat.cmu.edu/capstoneresearch/spring2021/315files/team16.html>