Anthony Wilson

Milestone #5 – Final Paper

DSC 630 – 301

6/5/2021

Predictive Analysis - Strokes

Abstract

There is a gap in identifying who will have a stroke. Scientists have made progress in isolating certain risk factors associated with strokes like weight management, controlling blood pressure, and smoking. Scientists have also identified factors out of our control, like race and a person's family history of having a stroke. This study aims to look beyond personal and family health history, exercise, and eating healthy to determine other risk factors involved. Specifically, it will investigate those who have been married, their work type, residence type, along with previously identified risk factors like smoking, weight, and health history.

To better understand strokes, we have used the Kaggle Stroke Prediction Dataset (fedesoriano, 2021) to identify variables influencing people having strokes. There is no information on how the author collected the data. We took the dataset and split it up into two. In one dataset, we created a model to group stroke victims and non-stroke victims to identify patterns and similarities within each group. The second dataset was used to test the model's capabilities in recognizing patterns between the two groups. The results didn't show anything significant outside the Kaggle dataset.

While the model could predict with 70% accuracy on the dataset, we are not confident that we would reproduce the results outside of this sample. The model incorrectly predicted five

times more people having a stroke. More data will need to be collected to understand strokes better. This dataset is still missing information.

Introduction

Strokes are one of the leading causes of death in the United States (About Stroke, n.d.). About 795,000 people in the United States have a stroke each year (Stroke Facts, n.d.). Two different types of strokes can occur: "a blocked artery (ischemic stroke) or leaking or bursting of a blood vessel (hemorrhagic stroke)" (Stroke, n.d.). Strokes can be devastating to those affected by them and their loved ones. Understanding the risks of strokes can decrease one's chances of having a stroke. High blood pressure, obesity, smoking, and high cholesterol are common causes of having a stroke. (Stroke, n.d.) Also, a few demographic risks tied to having a stroke are people over the age of 55 are more likely to have a one, African Americans are at a higher risk, and men are more likely to have a stroke, but women are more likely to die from having a one. (Stroke, n.d.) Becoming informed is essential because there could be lifestyle changes one could make to help prevent a stroke. Beyond exercise and eating healthy, are there other influences that can cause strokes?

The general problem is that strokes endanger and affect many people in the world today. The specific issue in this project is to identify the significant influencers of a stroke.

This research project will focus on the Kaggle Stroke Prediction Dataset (fedesoriano, 2021). The dataset has eleven features listed below, along with general definitions that came from the website.

Methods

There is no information on how the author collected the data. The data owner has requested that fedesoriano, the Kaggle user, get credit. (2021) The information we have about patients comes from the Kaggle website, and the features are listed below. (fedesoriano, 2021)

Feature	Defintion
id	unique identifier
gender	Male, "Female" or "Other"
age	age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	No or "Yes"
work_type	children, "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	Rural or "Urban"
avg_glucose_level	average glucose level in blood
bmi	body mass index
smoking_status	formerly smoked, "never smoked", "smokes" or "Unknown"*
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	

Table 1 - Features

The total number of patients in this dataset is 5,110. The mean age of the patients is 43, and the minimum age is 0, with a max-age at 82. Seventy-five percent of the participants are 61 or younger. Having a majority of the participants younger than 61 is interesting because 34 percent of people hospitalized for strokes were less than 65. (Stroke Facts, 2021) The total number of patients who show as having a stroke was 249, which is 4.9% of the patients, leaving it a bit unbalanced. One hundred fifty-six of the patients who had a stroke were 65 and older; this makes up 63% of all stroke patients and 3.1% of the total sample. Patients between the ages of 50 to 59 made up 20% of all patients from the selection, and 49 had a stroke.

A total of 2,115 patients were male, and 2,994 were female, and there was one patient labeled as other. Of those who had hypertension (498), 66 of them had a stroke, and there were 183 that had a stroke but did not have hypertension. Heart disease can be scary, and having both heart disease and the risk of a stroke can be extremely scary. Two hundred seventy-six patients have heart disease, and of those, there are 47 that have had a stroke. There are 202 that have had a stroke and do not have heart disease. Three thousand three hundred fifty-three patients have been married, 220 of them have had a stroke. Of the 1,757 patients that have not been married, only 29 of them have had a stroke. There are five different categories for work types. A majority of the patients work for a private company, 2,925. Of those that work for a private company, there were 149 that had a stroke. Self-employed, government jobs and those staying home to focus on their children make up between 13-16% of the patients apiece. There are 22 patients, 0.4%, that have never worked before.

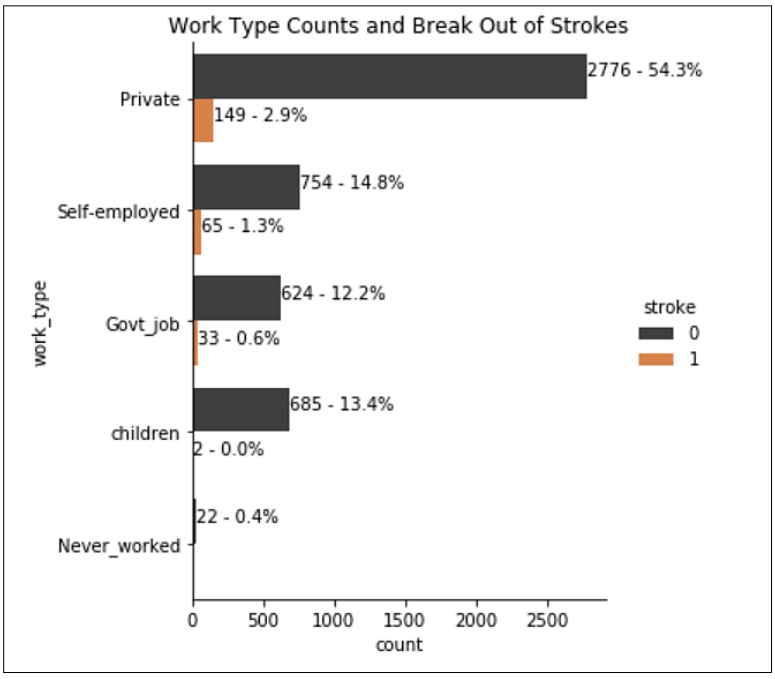

Figure 1 - Work Type Counts and Break Out of Strokes

About half of the patients live in an urban community, and the other half live in a rural community. There were no significant differences for the residence type that appeared for those who had strokes. It was close to being split down the middle for residence type. There were four different categories for smoking status see figure 2. The most significant numbers related to

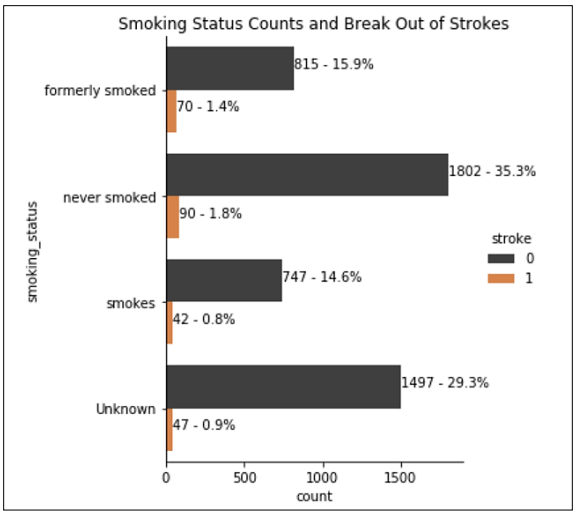

Figure 2 - Smoking Status Counts and Break Out of Strokes

smoking status were those who never smoked; they make up about 37% of the patients. There are no significant jumps for the number that smoke relative to those who have not or never have smoked.

The average body mass index (BMI) is about 28.89. The max BMI is 97.6, and 75 percent of the patients are below 33.1 BMI. In figure 3, we have the BMIs broke out into groups. Notice, several have a BMI over 30.0, and most patients who have had a stroke are in this bucket. Three thousand seven hundred seventy-six patients have a glucose level between 70-180, and 144 of them have had a stroke. There are 553 between 181-249 average glucose levels, and 73 have had a stroke. Patients who had a stroke and their glucose levels were between 181-249 account for 13.2% within the group. There is 3.8% with a glucose level between 70-180 that have had a stroke. Between the two groups, 13.2% compared to 3.8% is a significant difference.

The columns were all pretty well populated. The only one that had missing values was BMI. There was a total of 201 missing values from the BMI column. We filled the BMI columns with the average of the BMIs. In future study's we may end up considering a different value or leaving them missing. Since there was only one

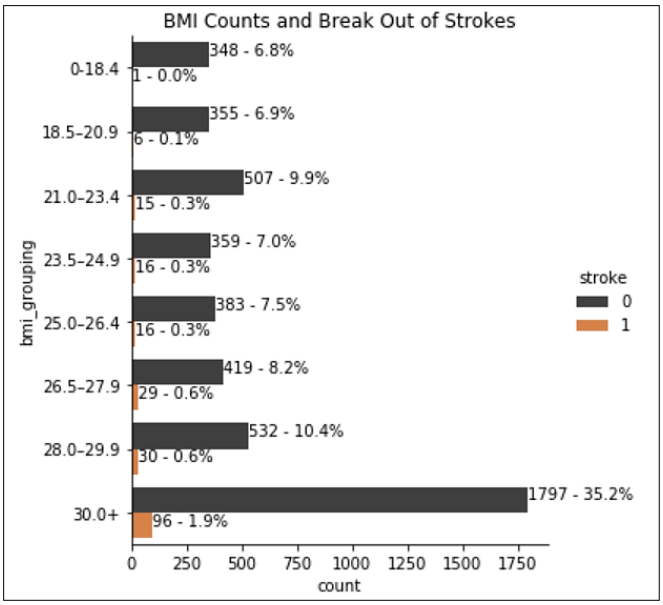

Figure 3 -BMI Counts and Break Out of Strokes

gender that was marked other, we decided to remove it. It would not have any significant effect on the model whether we kept it or left it.

Feature Engineering

We are looking to use the following model's decision trees, logistic regression, SVM, linear discriminant, naïve bayes, and random forests. We choose to take some of the features and update the columns. For example, instead of using the feature gender, we created a new feature called female. The female feature is a flag, and when marked as one, the patient is female, and if it is zero, the patient would be a male. We did the same thing with residence type by creating a feature residence_rural, where if labeled one, then the patient's residency is rural; otherwise, it is urban. We broke out the features smoking_status and work_type by their categories. For smoking, there are four categorical labels for features. The labels for smoking_status are 'unknown,' 'formerly_smoked,' 'never_smoked,' and 'smokes.' We mapped them to the following features 'smoking_status_unknown,' 'formerly_smoked,' 'never_smoked,' and 'smokes.' We did the same thing with the feature work_type, and here are the new features created 'work_govt_job,' 'work_private,' 'work_self-employed,' 'work_children,' and 'never_worked.' Doing this will make the feature selection more straightforward and also easier running the data through the models.

Feature Selection

We used the Extra-Trees Classifier and the Variance Inflation Factor (VIF) to understand better which feature to use. The Extra-Trees Classifiers will give us an idea of the most important features to use for our modeling. It uses several decision trees and sub-samples of the data while it goes through these decision trees and averages its predictions. The output gives a

value for each feature and how important they are. Figure 4 shows the feature importance using the extra-trees classifier. We can see the results of the top ten features. We have age, average glucose level, and BMI are the most important. The others features have a lower value, almost a fifth of age, glucose, and BMI.

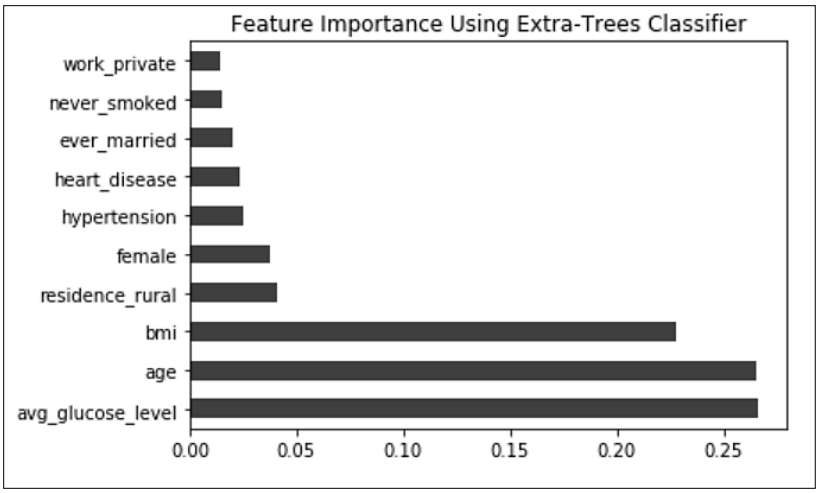

Figure 4 -Feature Importance Using Extra-Trees Classifier First Run

Let us continue to explore which features are the most important with VIF. VIF is used to identify which variables may have multicollinearity in a data set. A high VIF indicates that there might be some multicollinearity going on. Multicollinearity measures whether two independent variables correlate in regression. If this does happen, it can take away the importance of a variable in its predictive power on the dependent variable. The first time running the VIF, several features had a significantly high VIF, see Table 2. The values that were related were the work types and smoking status. To fix this, I removed the features smoking_status_unknown and never_worked. To account for the variables removed, I set the other categorical smoking and work type variables to negative ones where the removed fields were flagged. For example, if all features formerly_smoked, never_smoked, and smokes equal zero, this would

feature	VIF
age	2.868734
hypertension	1.118623
heart_disease	1.114709
avg_glucose_level	1.108213
bmi	1.300622
female	1.027906
ever_married	1.987492
work_govt_job	inf
work_private	inf
work_self-employed	inf
work_children	inf
never_worked	inf
residence_rural	1.002573
moking_status_unknown	inf
formerly_smoked	inf
never_smoked	inf
smokes	inf

Table 2 - VIF Feature Table

represent smoking_status_unknown, and we set all three variables equal to negative one. With a VIF score, we are looking for values that are less than ten. On the second run with the updated features, age and BMI both had values at twelve.

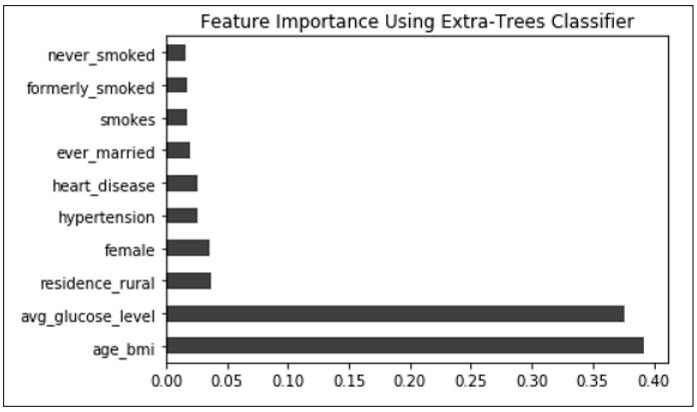

Figure 5 - Feature Importance Using Extra-Trees Classifier Second Run

To account for this high VIF, we took the patient's age minus their BMI and removed both feature's age and BMI. The results of the final VIF scores were all under six, so we felt confident in moving forward. After running the VIF calculations, we went back to the Extra-Trees Classifier to identify the top ten essential variables again and found the same information as before with age, BMI, and glucose at the top, see figure 5. Since we combined age and BMI, we added the new feature formerly_smoked. We created a correlation heatmap, see figure 6, indicating any correlation among the variables. The variables worth mentioning with a high correlation are age_bmi and ever_married, with a

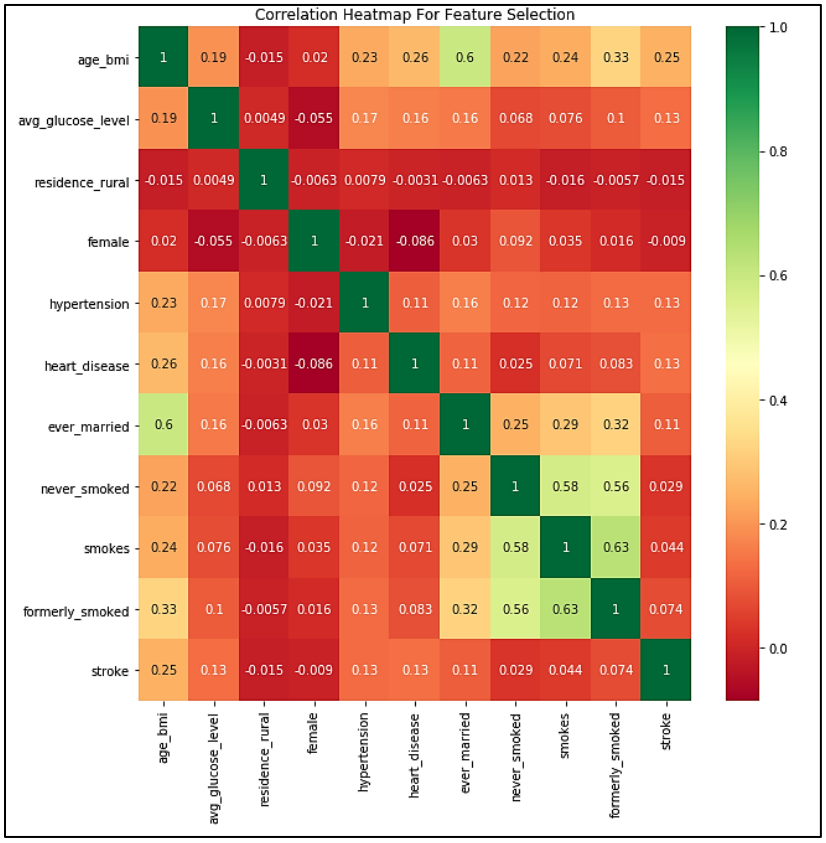

Figure 6 - Correlation Heatmap for Feature Selection

correlation of 0.6. The smoking variables compared to each other all have higher correlations as well, around 0.6. With the VIF values on our features less than six, we felt confident to move forward in our modeling.

Model

After we prepped and decided on the features, we split the data into separate data sets, test and training. Thirty percent of the data is inserted randomly into the test data set, and the rest lands in the training set. Notice in figure 7, very few patients have had a stroke. This small number causes an imbalance in the data. We used Synthetic Minority Oversampling Technique (SMOTE) to oversample the data and address the imbalance. SMOTE takes the imbalanced data, e.g., those who have had a stroke, and inserts them multiple times until the number of patients who have had a stroke equals the number of patients who have not had a stroke. If we don't address the imbalance in the data, it can cause the model to lose its accuracy in prediction.

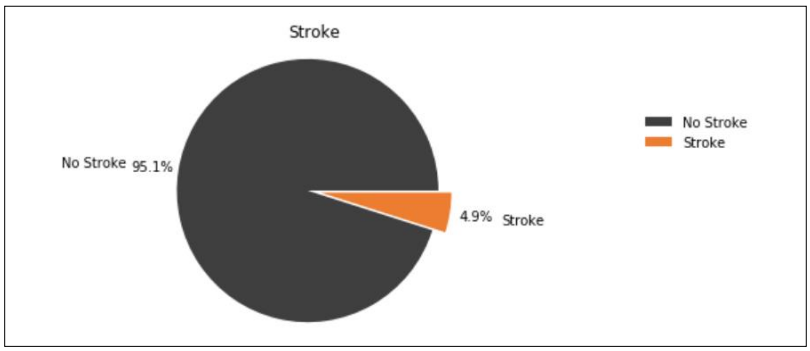

Figure 7 - Stroke Pie Chart of Patients Having Stroke

For the models, we used the package sklearn in python and the following models from the library DecisionTreeClassifier, LogisticRegression, SVM, LinearDiscriminantAnalysis, GaussianNB (Naïve-Bayes), RandomForestClassifier. As a note, we did try to tune some of the models up after running our predictive analysis. The output results were not worth noting because they did not increase in their predictive power by any means. In some instances, they decreased. We also tried to pull out all children 18 and younger to see if it differed in the results. We did this so we could look at adults. Adults 55+ are more likely to have a stroke, and children pull the means down for features like age, BMI, glucose levels, work, and smoking status.

Taking children out still left us with 82% (4,194) of the patients, but similar to the results from trying to tune the models, it did not significantly affect the results.

Results

In table 3, we have the models and their f1-scores. The three that did the best are logistic regression, SVM, and linear discriminate analysis. Below in figure 3, we have the confusion matrix for SVM. The model recall was 80%, it predicted 60 out of 75 patients having a stroke. The precision was 98.5%, which predicted 1,011 out of 1,026 patients who did not have a stroke. The total accuracy of the model was about 69.9% accurate; it predicted 1,071 out of 1,533 patients correctly from the test dataset. Other models were more accurate than the SVM. For example, the decision tree was 92% accurate, and the f1-score was 14.6%. The precision was 96%, and recall was 15%, played into the f1-score being lower for the decision tree model. Type II errors were pretty significant in all the models that we ran. The type II error in the SVM model represents the 447 patients we predicted having a stroke but did not. That is five times more patients incorrectly predicted to have a stroke. Type II errors were the most significant factor in the f1-scores being so low.

Model	f1-score
Decision_Tree	0.145695
Logistic_Regression	0.201373
SVM	0.206186
Linear Discrim	0.206593
Naive_Bayes	0.152310
Random_Forrest	0.038462

Table 3 - Model f1-scores

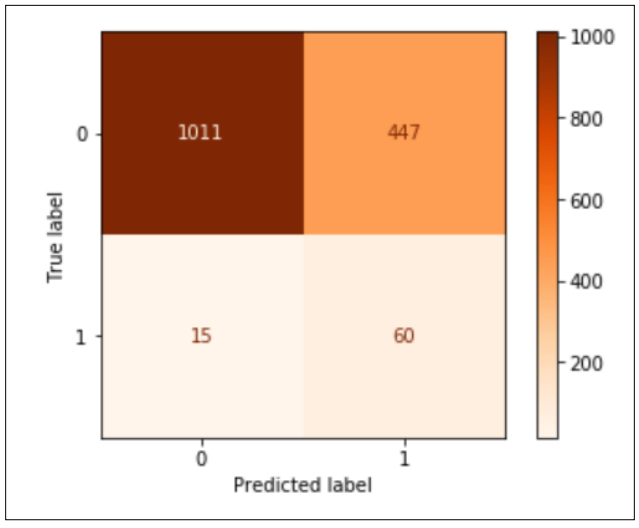

Figure 8 - SVM Confusion Matrix

Discussion/Conclusion

Our findings are interesting, with different classification models varying in output. The imbalance in the data is pretty significant.

Five percent of the patients have had a stroke, and the population is relatively young. Eighteen percent of the population is 18 or younger. With these issues, I am concerned that it will be challenging to get an accurate result. I am not confident with the prediction power these models have in identifying patients with a stroke based on the features. I want to find more datasets for strokes to understand stroke victims and potentially identify those at risk.

Acknowledgments

I would like to recognize the support of my family as I have written this paper. They have sacrificed their time to help me focus on my work. My wife Angela has taken some of the extra weight of my responsibilities to make sure I had time to work on my project. She also has proofread and helped me work through roadblocks in my paper. My professor, Dr. Brett Werner, thank you for being available to answer my questions. I appreciate the feedback on my work. It has helped me understand what I need to work on and what I have done well, giving me the knowledge and confidence I needed to complete this course. Kaustubh Singh and Michael Koffie, have been instrumental in peer reviews and ensuring that I stay on track.

References

About stroke. (n.d.). Retrieved March 28, 2021, from https://www.stroke.org/en/about-stroke

fedesoriano. (2021, January 26). *Stroke Prediction Dataset*, 1. Retrieved April 8, 2021, from Kaggle: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Stroke facts. (2021, March 17). Retrieved March 28, 2021, from https://www.cdc.gov/stroke/facts.htm#:~:text=Every%20year%2C%20more%20than%20795%2C000,are%20first%20or%20new%20strokes.&text=About%20185%2C000%20strokes%E2%80%94nearly%201,have%20had%20a%20previous%20stroke.&text=About%207%25%20of%20all%20strokes,to%20the%20brain%20is%20blocked.

Stroke. (2021, February 9). Retrieved March 28, 2021, from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113#:~:text=A%20stroke%20occurs%20when%20the,and%20prompt%20treatment%20is%20crucial.