Using Data Science on a Stroke Prediction Dataset

Author: Fauzan Amjad

About Me

My name is Fauzan Amjad and I'm a student in the Rutgers School of Arts and Sciences majoring in Computer Science while also taking pre med requirements. When others hear about my academic ambitions, it is often met with confusion as it seems like I'm combining two very distinct disciplines together. It is for that very reason why I am choosing to combine computer science with the healthcare field. The application of computer science, particularly artificial intelligence and data science, in the healthcare field has so much potential, yet many people don't see it as a lucrative or fulfilling path to pursue. These applications can come in various shapes: hospital resource allocation, clinical diagnoses, emergency services, etc. My friends and I actually won the Johnson and Johnson Black Tech hackathon where we developed a machine learning algorithm predictive analysis model to predict with 91% accuracy the leading causes of infant death based on CDC data. With these type of experiences, I aspire to be a physician scientist to further explore the power we can yield with data in the field of medicine in order to achieve a more modernized future.

Background Information

Throughout the history of our civilization, we've developed and implemented elements into our society that have either has a beneficial impact or a detrimental impact on the population. The development of vaccines to prevent harmful diseases, groundbreaking cancer research, and the commercialization of advancing computers that can process more information with more efficiency all have contributed to increasing the standard of living, providing more opportunities, and, simply put, modernizing the world we live in. Unfortunately, we've also developed innovations that may have negatively impacted the society we live in, and even with good intentions, the long-terms consequences of such distribution foster a broad spectrum of issues. Although weapons of mass destruction, imperialism, and radicalization through different mediums of what we see in our life could be explained for hours, let's mainly focus on detrimental items that affect our health. Foods marketed as low fat but contain an unhealthy amount of sugar, tobacco being constantly used especially in pop culture, electronic vape with nicotine pods being marketed to the younger generation with colorful flavors, and even unhealthy foods distributed in the public school system have contributed to obesity, cancer, stroke, and heart disease.

A stroke occurs when brain cells die due to your brain not getting enough blood (Stroke). The most recognized symptom of one that has this is that they have slurred speech coupled with paralysis of a certain body type and headaches. Heart disease can encompass a spectrum of issues that can impact the heart such as an infection on the heart, a defect in the heart structure, coronary artery disease, the irregular pattern of how a heart may beat, and a disease that impacts the muscle on the heart (Heart Disease). Heart diseases is the number one cause of death in the United States and number two cause of death globally. The reason for why I bring up these particular issues is that a lot of heart issues and strokes are caused by activities pursued by the individual or the environment the individual within. It is important to note that for some, these issues can be a result of natural causes or simply genetic factors; however, an unhealthy lifestyle can further exponentiate the problems or put the potential person in a much graver situation.

Dataset

The dataset, which was found on the Kaggle data scientist community, that will be utilized in this research is a confidential dataset that anonymizes the people utilized in the study with specific identification numbers. The rich dataset includes 5110 people with information such as their gender, their age, whether or not they have hypertension, whether or not they have ever had a heart disease, whether or not they were ever married, their type of work industry (government, self-employed, private, never), the type of location they live in (rural or urban), their average glucose level, their body mass index, their smoking status, and whether or not they have ever had a stroke.

Exploring the Dataset

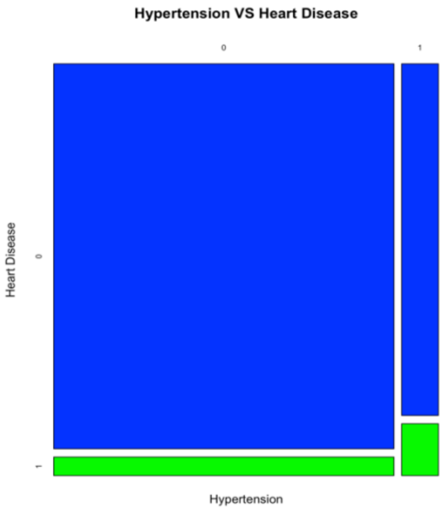


Figure 1.1

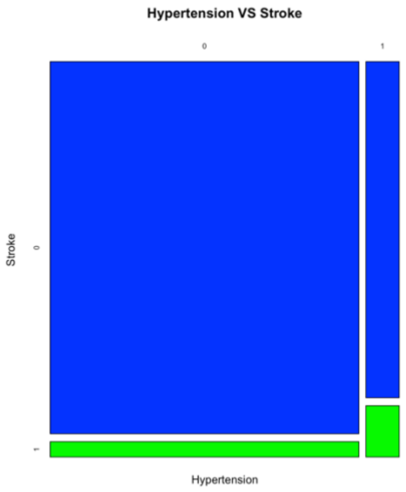


Figure 1.2

Hypertension is easily described as high blood pressure and can lead to problems like heart disease and stroke. Utilizing the dataset, a mosaic plot was developed that uses whether or not someone has hypertension and whether or not someone has had a heart disease as the categorical data. The mosaic plot is shown in Figure 1.1. Understand that for future references within this article, the number zero will represent something that is not the case while the number one will represent something that is the case. For example, if someone has had a heart disease, that'll be noted as one on the dataset. Shown in Figure 1.1, we can see a clear indication that hypertension may actually have an effect on whether or not someone has gotten heart disease. Most of the people sampled in the dataset have neither hypertension nor heart disease. But the percentage of people that have hypertension increases within people that have heart disease. Hypertension does not necessarily have to be present in order for someone to get heart disease. The same pattern applies to Figure 1.2 as well that shows a mosaic plot between hypertension and stroke. Although you do not necessarily need to have hypertension to get a stroke, the percentage of people that have hypertension increases when we look at the stroke population versus the population that doesn't have a stroke.

The dataset also provides information regarding each individual's calculated average glucose level. A higher glucose level can damage many aspects of the body and make the body more susceptible to both heart diseases and strokes. Since the average glucose level is quantitative data and whether or not someone has a heart disease is categorical, a boxplot was developed to show the distribution as well as plotting the means. As shown in Figure 3, the median average glucose levels of heart disease people was higher than the median average glucose level of people who didn't have heart disease. Interestingly, the distribution was a lot more broader for people that have had a heart disease as even people with above a 250 average glucose level were considered part of the 75th to 99th percentile. For people that have not have had a heart disease, everything after approximately 160 as an average glucose level was considered an outlier. The same pattern is exhibited when stroke versus average glucose level is plotted on a mosaic plot as shown in Figure 3. The median average glucose level of people that have had a stroke is higher than the median average glucose level of people that have not had a stroke. Furthermore, the distribution of average glucose level for people that have had a stroke is larger than the distribution for people that have not had a stroke, with quarter three of the second plot exceeding the statistical maximum (not an outlier) of the first plot shown in Figure 3.

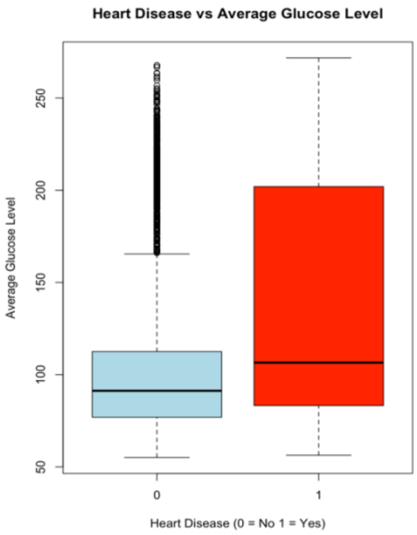


Figure 2

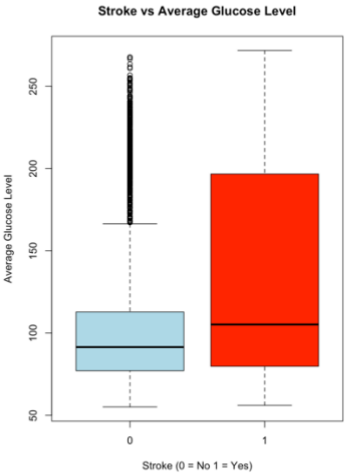


Figure 3

Once seen as a popular method relaxing, smoking tobacco has now been linked to many disease and cancers, including heart disease and strokes. The dataset includes data about each individuals smoking status. The different categorical data points that were used include never smoked, smokes, formerly smokes, and if their smoking status is unknown. As shown in Figure 4, not smoking doesn't necessarily mean that someone is going to not a heart disease. However, the percentage of heart disease people who smoke is higher than the percentage of non-heart disease people

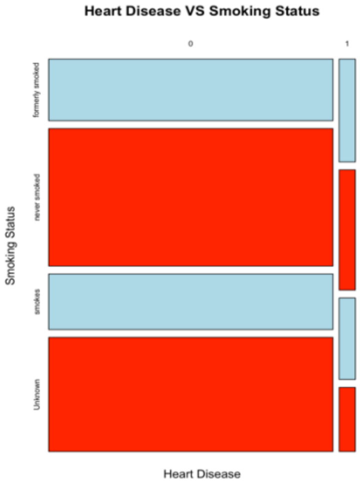


Figure 4

who smoke. The same trend is exhibited with formerly smoked, as the percentage of them is higher with people who have had heart disease versus people who have never had heart disease. Percentage of people who never smoked is roughly smaller for people who have had heart disease than it is for people that don’t have heart disease. Look to Figure 5 for the stroke versus smoking status mosaic plot. The same trend is exhibited with Figure 5 like it did in Figure 4, except the percentage of people who have never smoked remains roughly the same between people who have had a stroke and people who have never had a stroke. The percentage of formerly smoked increases with people who have had a heart disease as well as people who currently smoked. We can use Figure 4 and Figure 5 to show that smoking, even if one formerly pursued it, can have an impact on whether or not they can get a stroke or heart disease; however, one doesn’t necessarily need to have smoked to get these respective diseases, suggesting that there were other environmental factors and possibly genetics.

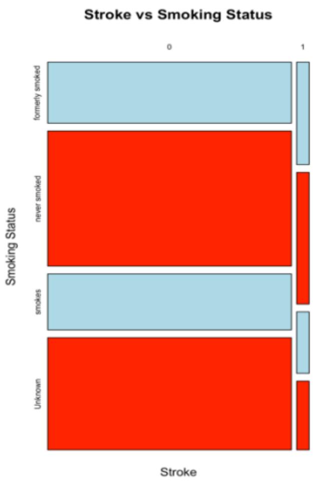


Figure 5

Permutation Tests

Through exploring the data, Figure 2 illustrated the median average glucose level was higher for people with heart disease than it was for people that have never had a heart disease. Conducting a permutation test will statistically show whether that increase in average glucose level is significant or not. The null hypothesis will be “There is no difference between the average glucose level of people with heart disease and the average glucose level of people that have never had a heart disease.” The alternate hypothesis will be “There is a difference between the average glucose level of people with heart disease and the average glucose level of people that have never had a heart disease.” Figure 6

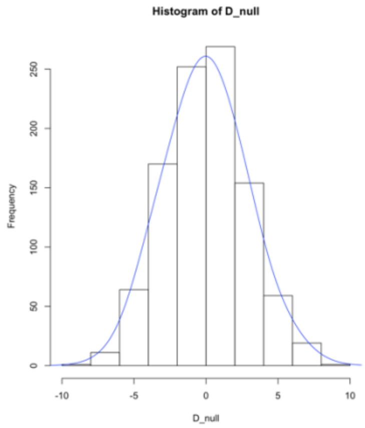
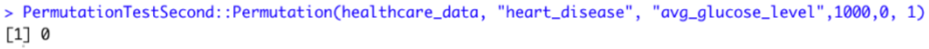


Figure 6



```
> PermutationTestSecond::Permutation(healthcare_data, "heart_disease", "avg_glucose_level",1000,0, 1)
[1] 0
```

Figure 7

illustrates the distribution of D_null and Figure 7 from RStudio, the software that ran the permutation test, shows that the permutation test yielded a p-value of 0. Since the p-value is under the 0.05 threshold, the null hypothesis is rejected and we have reasonable cause to accept the alternate hypothesis. We have significant statistical evidence that shows there is a difference between the average glucose level of people with heart disease and the average glucose level of people that have never had a heart disease.

Roughly the same permutation test can be constructed with stroke data this time. Conducting a permutation test will statistically show whether that increase in average glucose level of people that have strokes over people who have never had a stroke is significant or not. The null hypothesis will be “There is no difference between the average glucose level of people with stroke and the average glucose level of people that have never had a stroke.” The alternate hypothesis will be “There is a difference between the average glucose level of people with stroke and the average glucose level of people that have never had a stroke.” Figure 8 illustrates the distribution of D_null and Figure 9 from RStudio, the software that ran the permutation test, shows that the permutation test yielded a p-value of 0. Since the p-value is under the 0.05 threshold, the null hypothesis is rejected and we have reasonable cause to accept the alternate hypothesis. We have significant statistical evidence that shows there is a difference between the average glucose level of people with stroke and the average glucose level of people that have never had a stroke.

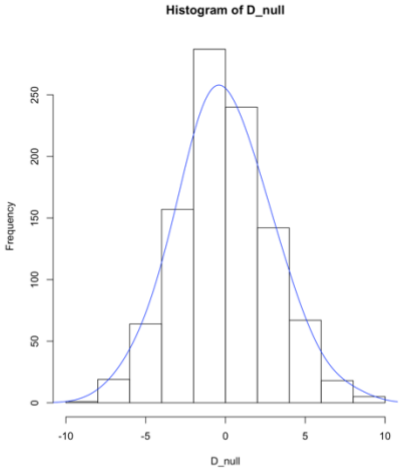
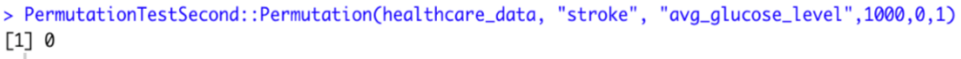


Figure 9



```
> PermutationTestSecond::Permutation(healthcare_data, "stroke", "avg_glucose_level",1000,0,1)
[1] 0
```

Figure 8

Discusson

The dataset provider notes that stroke and heart disease are the leading causes of death globally and cites statistics from the World Health Organization to prove their point. On top of that, the provider indirectly suggests that parameters such as gender, age, smoking status, body mass index, and glucose levels play a role in whether or not someone will get a heart disease and/or stroke. Although this isn't a clear indication that the data is skewed, it would be important to note that the dataset provider was collecting data with the purpose of proving that these input parameters play an impact in whether or not someone is going to get certain diseases. Furthermore, most of the dataset is filled with information of people who neither have had a heart disease issue nor stroke, meaning many conclusions or statistical plots were desiged with significantly different amounts of data for each category which is particularly evident in the mosaic plots.

The abundance of information in the rich dataset and plots constructed from the respective data itself provide insight about certain characteristics or other factors that may cause someone to get a heart disease or stroke. Certain things like having a higher average glucose level (maybe as a result of diabetes) and smoking or formerly smoking can cause people to get heart disease and stroke. As shown by the permutation tests, there is statistical evidence to believe that the average glucose levels of people with heart disease is different than the average glucose levels of people that have never had a heart disease. With the inclusion of the boxplot, the average glucose level is shown to be higher with people that have had a heart disease. The same thing is illustrates with people that have had strokes. Especially in today's world, the study of what causes these diseases are important as they are one of the leading causes of death and hardship across the globe and although genetic plays a factor, one cannot deny that the negative habits pursued in life as well as the environment one may be in play a role in whether or not someone will develop a disease/issue like heart diseases or a stroke. Other variables will definitely play a role and it is important to find those factors and see what factors impact more than others.

Data science application in finding root causes for certain diseases as well as illustrating whether or not they were influenced by environmental factors or activities pursued by the individual allow doctors and scientists in the healthcare field to effectively diagnose individuals of these respective diseases and suggest what's the best course of action to actually take in the recovery process. I'm excited to see the future of data science as well as artificial intelligence in the field of medicine.

Citations

"Heart Disease." Mayo Clinic, Mayo Foundation for Medical Education and Research, 9 Feb. 2021, www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

"Stroke." Mayo Clinic, Mayo Foundation for Medical Education and Research, 9 Feb. 2021, www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113#:~:text=A%20stroke%20occurs%20when%20the,and%20prompt%20treatment%20is%20crucial.