Milestone 4: White Paper

Stroke Predictions Using Machine Learning

Monica Santana

Bellevue University

DSC 680: Applied Data Science

Professor Amirfarrokh Iranitalab

September 22, 2024

Business Problem

CDC states that “every 40 seconds, someone in the United States has a stroke.” This is why this problem is very important to help solve and possibly reduce one’s chance of getting a stroke before it’s too late. “80% of strokes are preventable”, according to the Stroke Awareness Foundation. Therefore, this project will go over data that can help answer the next few questions of analysis that may predict a patient having a stroke. This dataset will be using demographics and medical records of patients to make predictions whether certain factors may have caused them to have had a stroke or not. Research questions I will explore include what factors from the dataset are most significant in a patient having a stroke. Are there significant correlations between certain factors and non-stroke patients? Is smoking a large factor in a patient getting a stroke? Is heart disease a large factor in a patient getting a stroke? What about diabetes/high glucose or BMI levels? What age or gender is more likely to get a stroke?

Background/History

Strokes are becoming more prominent in people of all ages each year. There has been an increase in strokes in younger people, so younger people are no longer safe from getting one. We are all prone to getting one no matter the age. There are many risk factors for all ages including high blood pressure, high cholesterol, diabetes, smoking, and obesity (Cleveland Clinic, 2019). Taking care of one’s health early on will prevent bad habits in the future that may cause long term health problems such as strokes. There are two different types of strokes, hemorrhagic or ischemic. If caused by a clot obstructing the flow of blood to the brain, it’s an ischemic stroke but if it is caused by a blood vessel rupturing and preventing blood flow to the brain, it is called a hemorrhagic stroke, stated by the American Stroke Association. Both types have the same risk factors and can happen at any age. The dataset I will be using does not specify which form of stroke the patient got, but for this analysis it does not matter since the risk factors are the same. My analysis goes over what types of factors are seen more in stroke or non-stroke diagnosis.

Data Explanation

The dataset I will be using is from Kaggle, and it has 5110 rows and 12 columns. I plan on only keeping some of the columns that I find necessary for my analysis. All the columns/factors are:

1. ID: number of row/patients
2. Gender: “male”, “female”, or “other”
3. Age: age of the patient
4. Hypertension: 0 for no hypertension, or 1 for hypertension in the patient
5. Heart disease: 0 for no heart disease, or 1 for heart disease in the patient
6. Ever married: “yes” or “no”
7. Work type: "children", "govt_job", "never_worked", "private" or "self-employed"
8. Residence type: “rural” or “urban”
9. Avg glucose level: average glucose level in the patient’s blood
10. BMI: body mass index
11. Smoking status: "formerly smoked", "never smoked", "smokes" or "unknown"
12. Stroke: 0 for non-stroke, 1 for stroke in the patient

The columns I plan on keeping are gender, age, hypertension, heart disease, average glucose level, BMI, smoking status, and stroke. I don’t plan on keeping ID, ever married, work type, or residence type as I want to keep the analysis health related only. I chose this specific dataset as it has enough medical information, and a good amount of patients to compare too. I like how it includes non-stroke diagnosis as well to see how those patients differ from the factors in the stroke patients for comparison.

To prepare the dataset for the modeling, I will check for any missing data, which was 201 NaN's in the BMI column. The rest of the columns were clean of any missing values. I removed them by dropping those rows from the dataset. I want this analysis to only include personal and medical demographics of the patients, so I removed columns "id", "ever_married", "work_type", and "residence_type". Lastly, I will be checking the data types for the remaining columns I will be using. Most of the columns have the correct data type, but there is 1 column that I will be converting from a float to an integer. It is the age column, since the dataset shows ages as "64.0", I want to have it say only "64". The age column can be left as a float to account for months and half of an age but for the purpose of this project, I don't really need to look at that so I will convert to whole integers for easier use.

Methods/Analysis/Conclusion

The methods I will be using include 2 types of models. Since I am working with a binary categorical variable as my target, the best suited models will be logistic regression and random forest classifiers. I will compare the metrics from both and see which one performs better of the two. I split the data into test and training sets to fit both models by dropping the outcome variable of stroke. The metrics I will be using to evaluate the models are f1-score, precision and recall due to their nature of performing accuracy for classification models. Between the two models, the random forest classifier model performed better.

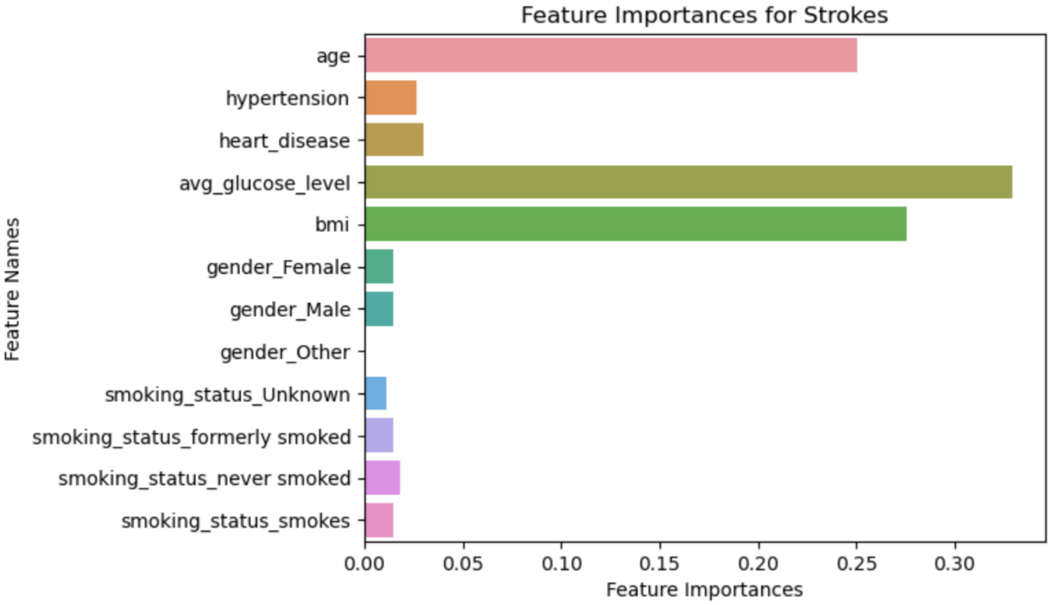
For the logistic regression model, the results showed the accuracy score is 94.81% with the accuracy score function. Overall, my model is close to average at predicting whether patients will not get a stroke with a f1-score of 97%. Since the random forest classifier model performed better, I will go into further detail there with the other metrics.

Random forest model results showed the accuracy score was 95.01% with the accuracy score function. Random forest model shows the precision for the non-stroke patient is 0.95, recall is 1.0, and the f1-score is 0.97. For the stroke outcome, precision score is 0.62, recall is 0.09, and f1-score is 0.16. This indicates that out of all the stroke patients, the model predicted, 62% had one. Out of all the patients that did get a stroke, the model predicted this outcome correctly for 9% of those patients. This model is also better at predicting non-stroke patients like the logistic regression model with precision of 95%. Out of all the non-stroke patients, the model predicted their outcome correctly very accurately. The f1-score, mean of precision and recall, is also better for non-stroke patients, 97%, compared to the stroke patients of 16%.

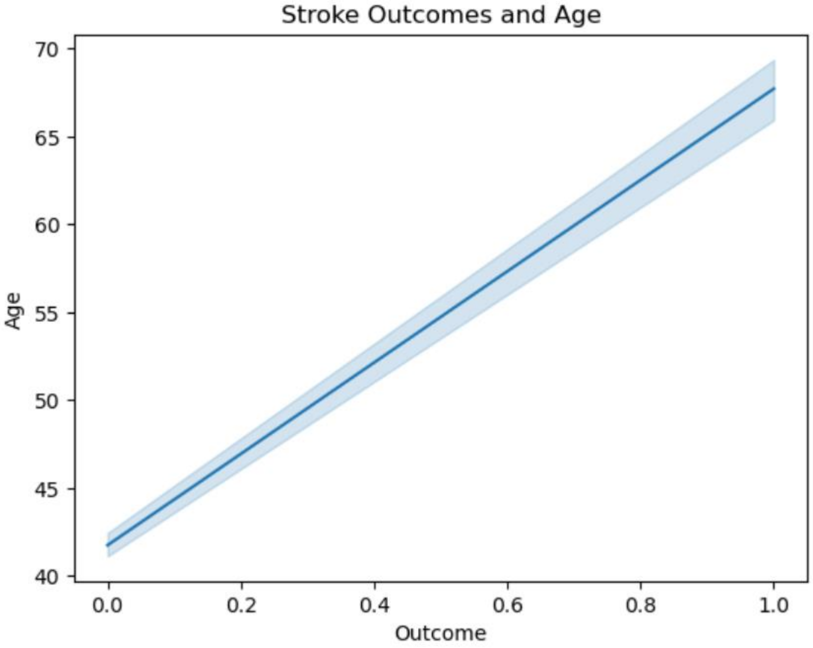
After looking at a correlation matrix and creating a bar graph with the top features that affect a stroke or non-stroke outcome, there were 3 that stood out substantially. The top 3 features were average glucose levels, BMI, and age. I also created a few visualizations to compare the glucose levels with a stroke outcome vs a non-stroke outcome. There will be another visualization to see the correlation between age and strokes, and BMI levels since they affected the stroke outcome the most.

Assumptions

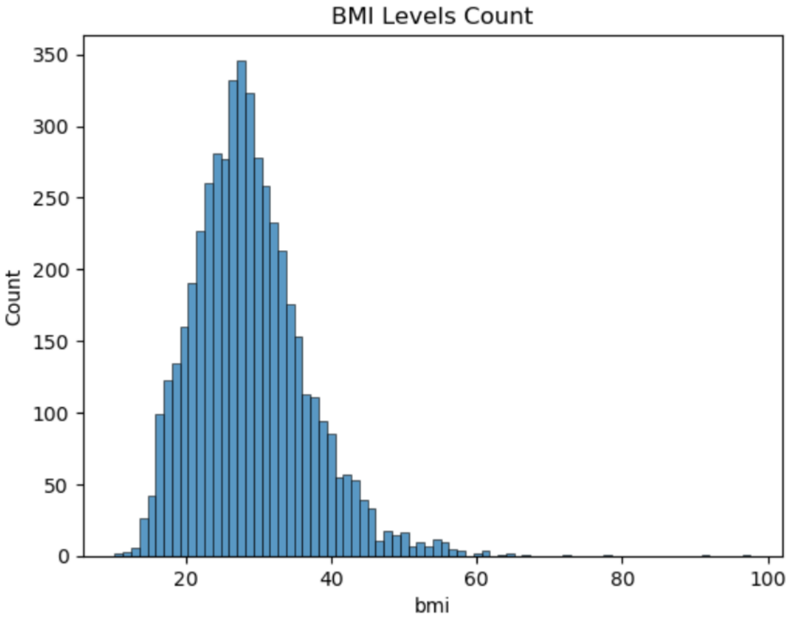
Assumptions I had made were the stroke outcome predictions accuracy and precision to be a bit higher. I did not expect it to be so low, but it makes sense since there are less patients in the dataset that have a stroke outcome. I also assumed that smoking was going to play a large role for the top features in stroke, but it was very low. Formerly smoking and smoking had the same level of importance to strokes which was still close to 0. I also assumed that heart disease would be high up in the level of importance, but it was not compared to the rest of the top 3 features shown in the bar chart below as well as the rest of the analysis with visualizations.



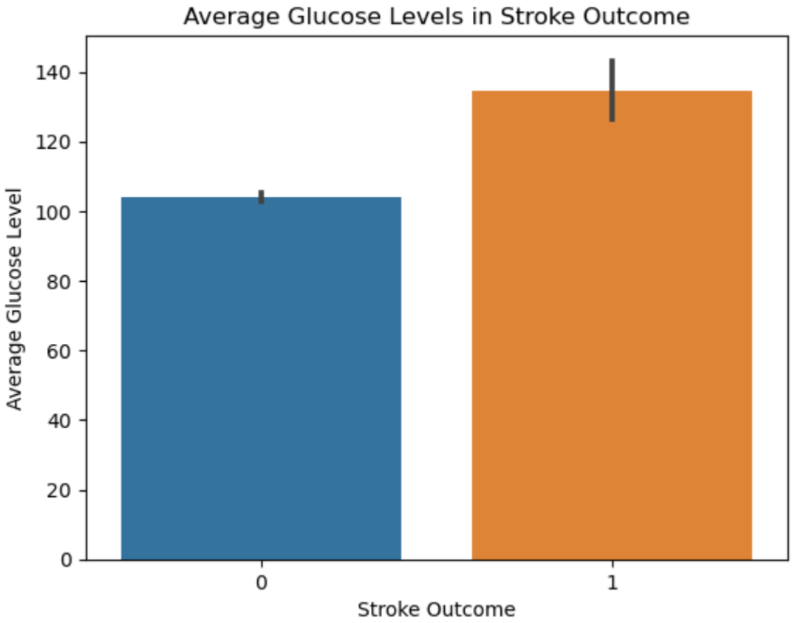
The line graph below shows the correlation between age and a stroke outcome. The 40-year-old patients were at the non-stroke outcome vs the older a patient got, the more likely they got a stroke. This provides insight on strokes not only happening to the elderly over 65-years-old.



The next visualization goes over the BMI levels and counts. Most of the patients had BMI levels between 26 and 28. According to WHO, World Health Organization, BMI greater than or equal to 25 is considered overweight. Which means most patients in this dataset are overweight.



This visualization shows that the average glucose levels for a non-stroke patient was a little over 100, while the stroke patients had closer to 135 glucose levels. A drastic difference between both.



Limitations / Challenges / Future Uses

Limitations to the analysis was not being able to conduct survival modeling. The dataset has no time variables or data for the duration value to predict the survival of said stroke outcome. Challenges faced were the dataset having more non-stroke outcomes vs stroke which I was not anticipating but it still provided valuable insights. The model can have future use for non-stroke outcome predictions since it was better for that and can be of use in the medical/healthcare field. More future uses include educational purposes and showing students what puts a stroke patient at risk, so they can start thinking about their own health and for their loved ones.

Recommendations / Implementation Plan / Ethical Assessment

Recommendations are to use this model to spread awareness for the risks of stroke and how one can help prevent one from early on. The model would be implemented in a healthcare setting after extensive review from professionals. Possible ethical concerns may include the privacy of the patient's information since their medical and demographic information is included in the dataset/models. For the presentation, the strokes in the U.S. statistics slide from outside sources must be correct to not mislead and spread inaccurate information. The references I have included here and in the presentation are trustworthy sites so that should not be an issue. My model's predictions are not 100% accurate without extensive review from professionals therefore this project cannot raise ethical concerns from medical professionals who may question the legitimacy of the project's analysis on stroke predictions. The top 3 risk factors in this analysis have only strengthened what have been risk factors for stroke for years, and nothing is controversial or questionable. Even the low risk factors cannot be proven without more review so that cannot raise any concerns. If the project was live in production, the models can raise ethical implications with the patient's information, but their names are not included for this reason.

References

(2019, September 13). *Why are strokes on the rise in younger people?* Cleveland Clinic. Why Are Strokes on the Rise in Younger People? (clevelandclinic.org)

(2020). *Stroke prediction dataset*. Kaggle. Stroke Prediction Dataset (kaggle.com)

(2024, May 15). *Stroke facts.* CDC. Stroke Facts | Stroke | CDC

About stroke. American Stroke Association. About Stroke | American Stroke Association

Stroke facts & statistics. Stroke Awareness Foundation. Stroke Facts & Statistics (strokeinfo.org)

Body mass index (BMI). World Health Organization (WHO). Body mass index (BMI) (who.int)