Project 2: Report/Milestone 3

Jahedur Rahman

Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catie Williams

October 23, 2022

Brain Stroke Prediction

In this project, I predicted the possibility of a brain stroke occurring in a person based on the conditions they have.

Background/History

A brain stroke is a medical condition where there is a lack of blood flow to the brain which can cause cell death. There are two main types of stroke. One is called an ischemic stroke. This is caused when there is a lack of blood flow. Another stroke is called a hemorrhagic stroke. This is caused when there is bleeding. Both types of stroke cause parts of the brain to stop functioning properly. A stroke may exhibit some signs and symptoms which include the inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side.

Business Problem

There are many risk factors that can cause a stroke. Some of these are high blood pressure, high blood cholesterol, tobacco smoking, obesity, diabetes mellitus, a previous TIA, end-stage kidney disease, and atrial fibrillation. This model will use some of these risk factors, and other factors, to predict if a stroke will happen to a person or not.

Data Explanation

Dataset

The Brain stroke prediction dataset will be the primary dataset used for this project. This dataset has all the needed columns listed below.

- gender: "Male", "Female"
- 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", "Govtjov", "Neverworked", "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

Data Preparation

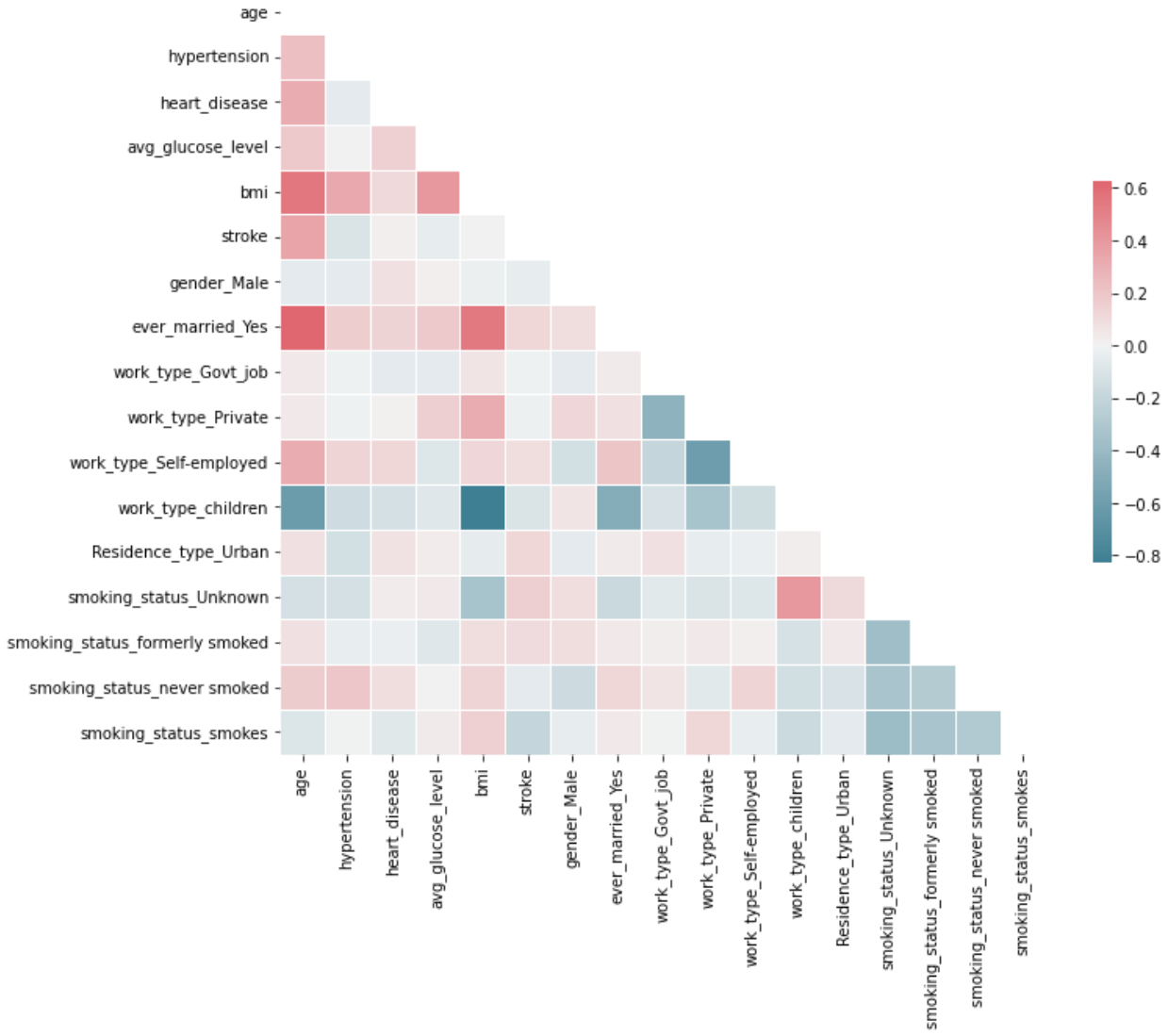
There was not a lot of preparation needed for this dataset. Since all the columns were required, no columns were removed. Also, I found no null values so I didn't need to remove any records. I had to create dummy variables for certain columns. After creating the dummy variables I removed columns that seemed repetitive to other columns. Finally, I split the dataset into training and testing datasets to fit the models.

Methods

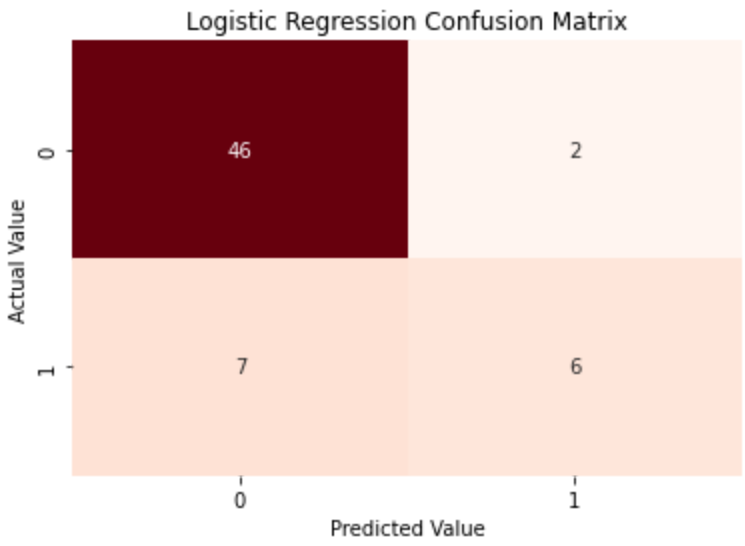
The target variable in the dataset is stroke. The model should be able to predict if a brain stroke will occur using the person's factors. There are two values for the stroke column. The value of 1 if the patient had a stroke or 0 if not. Since this is a classification scenario, a logistic regression model was used. In addition, other models such as a decision tree classifier and a random forest classifier were used to compare the results of the logistic regression.

Analysis

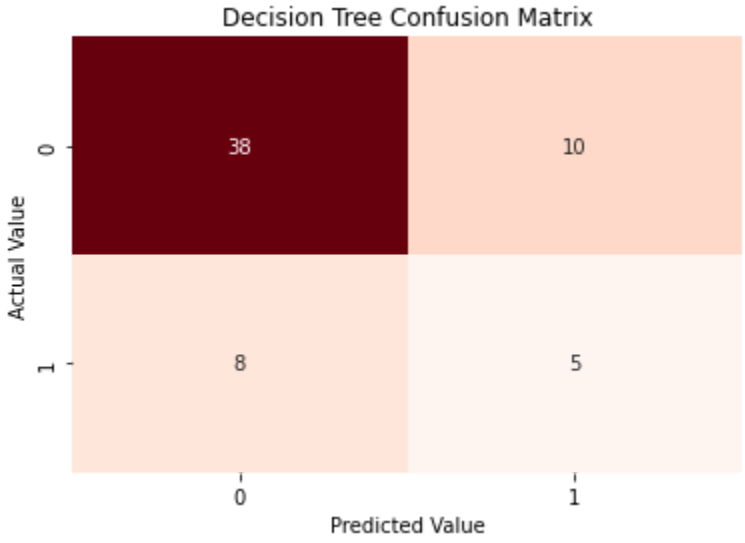
To start off with my analysis, I started with a heatmap. I wanted to see if there was any correlation between the variables.



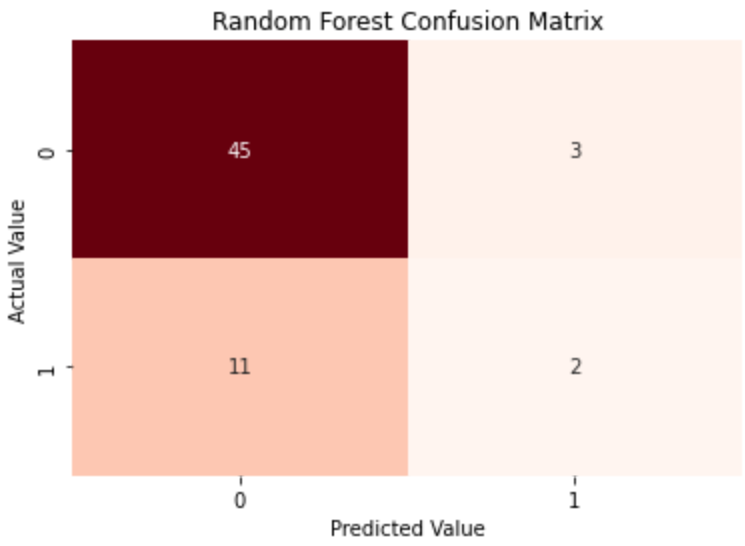
In this heatmap, we can see there is a very high correlation between ever_married_Yes and age. Unfortunately, this doesn't help us with figuring out if this affects a stroke or not. Additionally, stroke has a correlation with age. However, since age is a continuous variable and stroke is a categorical one, it is hard to tell if this correlation has any meaning. Next, I built a logistic regression model and computed the accuracy score, and created a confusion matrix.



The accuracy score for the logistic regression model was about 85%. The confusion matrix shows us that the model correctly predicted 52 out of 61 values and incorrectly predicted 9 out of 61 values. These results are really good. To compare these results I built a decision tree model.



The accuracy score for the decision tree model was about 70%. The confusion matrix shows us that the model correctly predicted 43 out of 61 values and incorrectly predicted 18 out of 61 values. These results are worse compared to the logistic regression model. Next, I built a random forest model and compared the results.



The accuracy score for the random forest model was about 77%. The confusion matrix shows us that the model correctly predicted 47 out of 61 values and incorrectly predicted 14 out of 61 values. These results are a little better than the decision tree model results. However, the results from the logistic regression model were still better.

Conclusion

Based on the model evaluation results, the logistic regression model will be used to predict a brain stroke. It has the highest accuracy and a higher number of correctly predicted values.

Assumptions and Limitations

There is a chance that the dataset has some incorrect information. Since for many of the columns there are values like 0 and 1, wrong values could be accidentally inputted. In addition, upon further investigation of the records in the dataset, I observed some age values were not whole numbers. Additionally, there are many other variables that I wish were included such as obesity, and diabetes.

Challenges

After analyzing the dataset with the heatmap, I find it weird that variables such as hypertension and heart disease had no correlation with stroke. These two variables should have a bigger effect on having a stroke or not.

Future Uses/Additional Applications and Recommendations

The model can be used to predict a brain stroke before it occurs in a person. This could be helpful to people that are already experiencing some symptoms.

Implementation Plan

The implementation plan is to use this model as an initial decision for patients experiencing some of the symptoms. This initial decision can help determine what steps could be taken to reduce the likelihood of a brain stroke occurring.

Ethical Assessment

Since the dataset uses patient data, it is important to be confidential. Looking at the dataset there are no specific identifiers to find the identity of the patient.

References

Akbasli, I. T. (2022, July 16). *Brain stroke prediction dataset*. Kaggle. Retrieved September 30, 2022, from https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset?select=full_filled_stroke_data+%281%29.csv

Appendix A

Under the Assumptions and Limitations section on page 6, it is written, “In addition, upon further investigation of the records in the dataset, I observed some age values were not whole numbers.” Removing the partial age values will most likely not change the results by much since there are very few of these values.