Final Project Summary

****

College of Science and Technology, Bellevue University

DSC550-T301: Data Mining (2235-1)

Sashidhar Bezawada

June 02, 2023

BRAIN STROKE PREDICTION

Final Project Summary (Milestone 1)

Introduction:

A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts (or ruptures). When that happens, part of the brain cannot get the blood (and oxygen) it needs, so it and brain cells die.

There are two types of Brain strokes:

- Ischemic Stroke – Caused by blood clots that block an artery and blood supply to brain is cut off
- Hemorrhagic Stroke – Caused by blow out of blood vessels in the brain.

Who is at risk of having a brain stroke:

1. High blood pressure over 140/90
2. High cholesterol leads to blockage in arteries that could impact the blood supply to the brain.
3. Diabetes
4. Obesity
5. Alcohol/Smoking/Tobacco
6. Sleep disorders

Apart from physical conditions, there could be non-physical factors such as the following that could be a risk factor

1. Age
2. Gender
3. Race
4. Genetics
5. History of previous strokes

Reason to choose this topic:

The main objective behind choosing this topic is irrespective of age, people these days are getting brain strokes due to our lifestyle and care towards our health. If we can predict the outcome from stroke history and various factors, we can avoid strokes ahead and lead a healthy life.

Reason for Buy-in:

We can use this product as a Self-Checker tool to predict brain strokes by the Medical organizations to predict strokes for their patients.

Dataset and Sources:

Link to Data Source: https://www.kaggle.com/code/aashidutt3/brain-stroke-prediction/data

Description: This dataset includes data related to attributes which would cause Brain Stroke in human beings. This dataset includes details like gender (Male or Female), age, hypertension (Yes or No), heart disease (Yes or No), work type (Private, Self Employed, Government or Student), average glucose level, BMI, smoking status (formerly smoked, never smoked, smokes or Unknown) and stroke (if the patient had stroke – 0 indicates no stroke and 1 indicates that patient had stroke).

In this dataset, we are going to analyze the frequency of Brain Strokes and factors which will cause Brain Strokes and the strongest predictors of it or certain attitudes towards it.

In this milestone 1, we are going to perform the following graphical analysis:

1. Comparing Brain Stroke in patients based on the gender category.
2. Comparing Brain Stroke in patients based on the hypertension.
3. Comparing Brain Stroke in patients based on the work type.
4. Comparing Brain Stroke in patients based on the smoking habits.

I have taken this analysis as a personal interest to find what factors are causing Brain Strokes as this has become one of the most common causes of death. This analysis would help the people to find out the most common causes of it and take necessary precautions to overcome it.

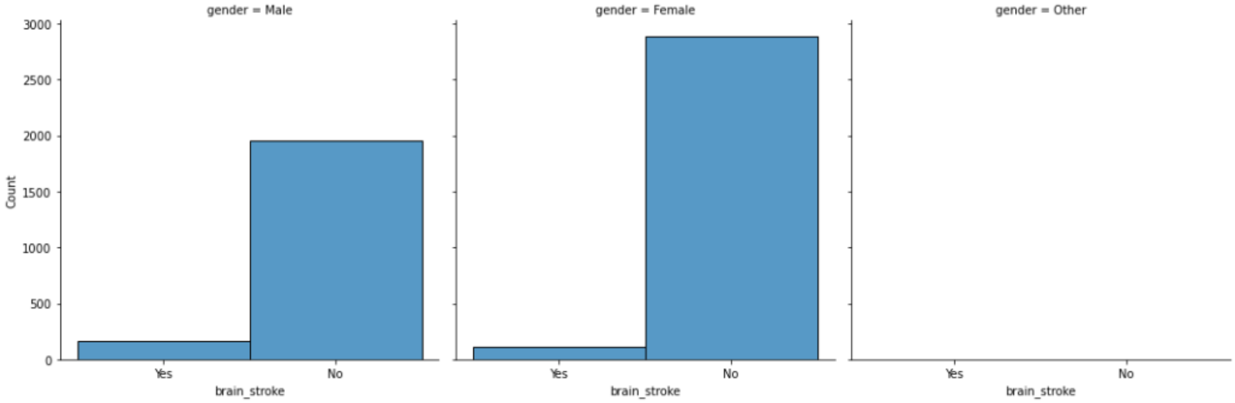
Data Transformation:

In this data Transformation phase, I have changes formatting and values for the following columns which will make job easy for the analysis:

1. work_type - Changed work type 'Children' as 'Student' as Children is not work type and Changing formatting for other work types
2. hypertension - Changed binary values to string values of Yes and No instead of 1 and 0 respectively as it is much easy to read.
3. heart_disease - Changed binary values to string values of Yes and No instead of 1 and 0 respectively as it is much easy to read.
4. stroke – Changed column name stroke to brain_stroke as this more meaningful.

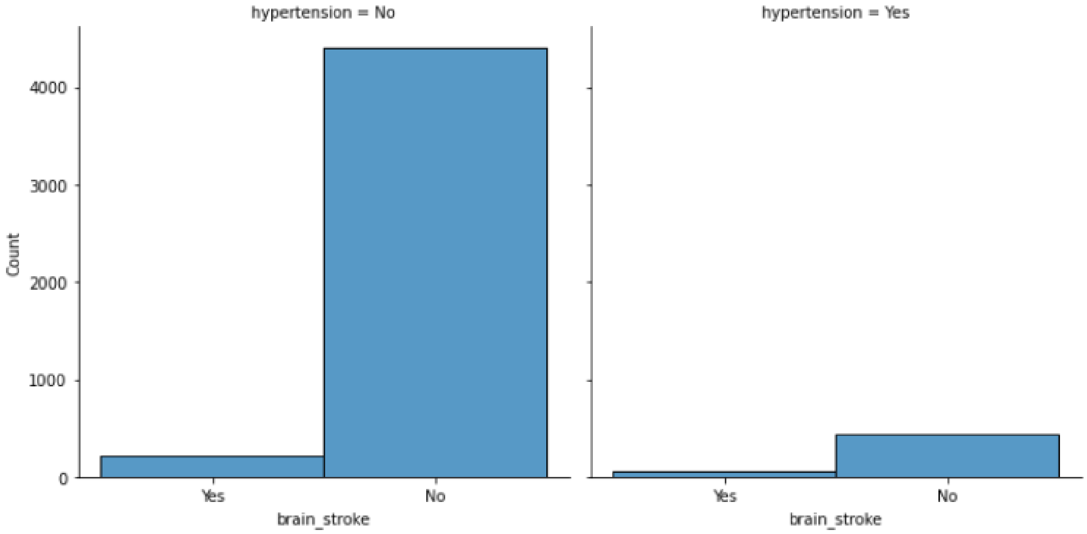
Data Visualization:

1. Comparing Brain Stroke in patients based on the gender category.



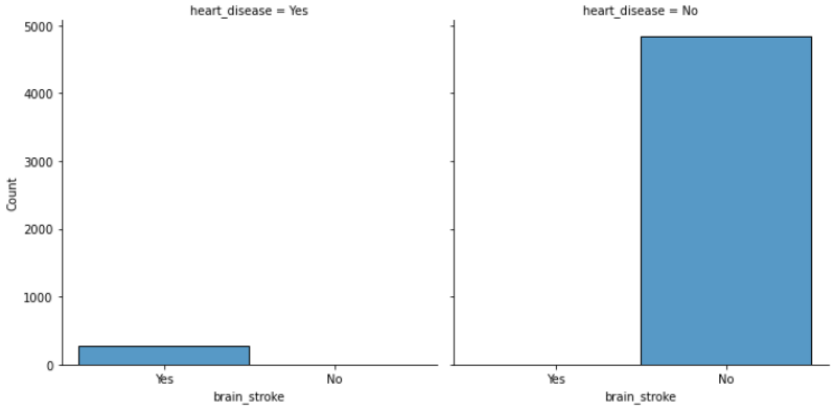
From the above graphical analysis, we can confirm that Females have more chances of suffering Brain strokes than males.

2. Comparing Brain Stroke in patients based on the hypertension.



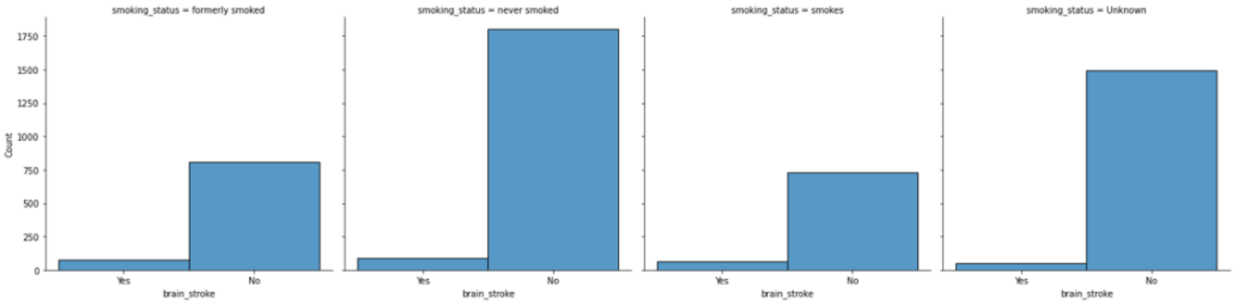
From this graphical analysis, we can confirm that with no hypertension lesser chances of Brain stroke

3. Comparing Brain Stroke in patients based on the work type.



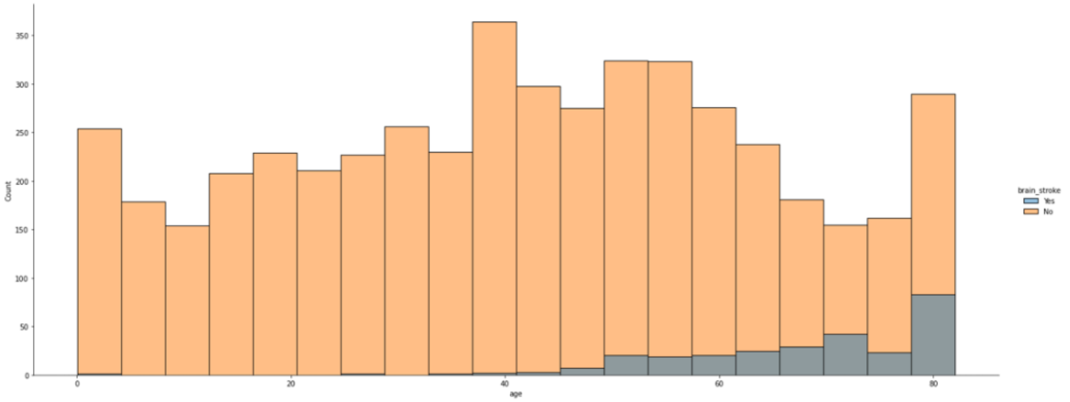
As per analysis, with heart disease history have high chances of getting Brain stroke as the graph shows heart disease is directly proportional to Brain Stroke.

4. Comparing Brain Stroke in patients based on the smoking habits.



As per graphical representation, we can confirm that non-smokers have very little chances of getting Brain stroke

5. Comparing Brain Stroke in patients based on age.



As per the graphical analysis,people over the age of 40 have high chances of getting Brain stroke

Data preparation:

As part of this project, I have worked on few data cleansing and data transformation techniques:

1. Dropping Columns/Features
2. Replacing values in a cloumn
3. Renaming column names
4. Transform features
5. Engineering new useful features
6. Replace Null values
7. String values case change

Model Building and Evaluation:

I have built KNeighborsRegression, Decision Tree Classifier and Random Forest Classifier and evaluated the model with some dummy data to it.

With classification regression and KNeighborsRegression, after evaluating the accuracy of the training model we got the output around 0.944.

With Decision Tree Classifier, after evaluating the accuracy of the training model we got the output around 0.944.

With Random Forest Classifier, after evaluating the accuracy of the training model we got the output around 0.937.

Though, there is not much difference, with 95% accuracy, classification regression and KNeighborsRegression is the best model.

Conclusion:

After implementation of above models, I can conclude that though the model building and evaluation looks good, there are some changes required. As the accuracy score for all the models is approximately 95%, I think we need to fine tune the models further to make the model more realistic than being idealistic.

These models are not ready to be deployed as they need further fine tuning. With minimal changes we can deploy to the production to be used. As the dataset is very small in size, I would recommend having a larger dataset with realistic data so that we can get more accurate predictions.

While working on this project, I have learnt building models is not easy. We need to consider many factors to make the build more accurate. Data preparation is equally important in Model building and evaluation.

There are many other models available which needs to be tried to see if any other model best suits us.

If we can maintain healthier life and good habits, we can stay away from Brain Strokes.