**Final Project**

**Introduction**

A stroke is a form of brain damage that occurs when there is a disruption in the blood flow to the brain. This disruption ultimately results in the death of brain cells. The World Health Organization identifies stroke as the second leading cause of death worldwide, with the condition being responsible for approximately 11% of all deaths (WHO). Stroke is the most significant cause of long-term adult impairment and the fifth highest cause of death in the United States, with over 795,000 stroke occurrences occurring each year, with roughly 75% of them being first-time strokes.

It is both damaging and pervasive, which means that it has the potential to have a significant influence on the quality of life of a person. Potential risk factors include advancing age, hypertension, elevated cholesterol levels, smoking history, diabetes, and a general high body mass index.

One study suggests that having appropriate knowledge of the warning signs of a stroke might reduce the likelihood of having one by as much as 80 percent.

* As a direct result, hospitals must participate actively in stroke prediction to reduce the time between when patients get treatment and when they have improved outcomes.

**Project Objectives**

We want to get a model that healthcare givers can use to make early predictions about the patients they are handling. In addition, we want to obtain the correlation coefficients to determine the most significant variables. We will also output a graphical representation of the relationships between different variables. For instance, we would want to see how many females are married and have had a smoking history. These graphical charts are essential in understanding data flow at a glance.

**Dataset**

The dataset was obtained via participation in a Kaggle competition centered on predicting strokes *(https://www.kaggle.com/fedesoriano/stroke-prediction-dataset).* The dataset contains information on various topics, including the patient's age, body mass index, glucose level, gender, kind of job, and occupation, among other things. The dataset contains 5110 observations, including 249 people who suffered from strokes. Consequently, the data could be more balanced, as stroke episodes only account for 4.87 percent of all statements.

**Materials and methods used**

In this project, we used and imported the following libraries;

* dplyr
* tidyverse
* ggplot2
* gridExtra
* gplots
* dummies
* gains
* caret
* rpart
* random-forest

The following snapshot shows how to have imported the libraries on R studio.



**Exploration of Data Analysis**

**Importing Data**

Let’s now see how the data was imported on the snapshot below;



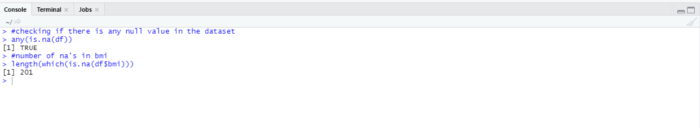
**Look for any numbers that aren't there (missing values)**

The absence of a value may present a significant challenge to modeling, and the result may not be as precise and accurate as we had hoped it would be. This is something that we had hoped would not be the case. Therefore, the first thing that must be done is to address the missing value present in the dataset. Within the data we access, approximately 201 deals for the characteristic BMI cannot be accounted for.

Before going on to modeling, the first thing that must be done is to deal with the missing value. There are three ways to accomplish this goal:

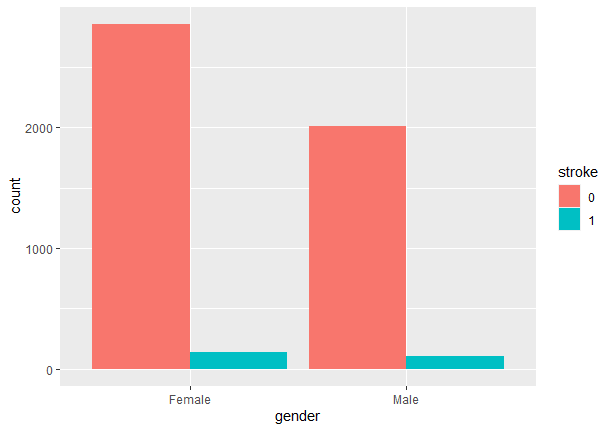
* Take out the column that contains nothing but blanks.
* calculating an estimate of the value's median
* It is estimated that the median value is

I used all three methods to meet this research's requirements. To begin, I eliminated a column from the dataset to construct a random forest model, which I referred to as model 1. After that, I completed a random forest model (model 2) by substituting mean values for absent values, and so on.

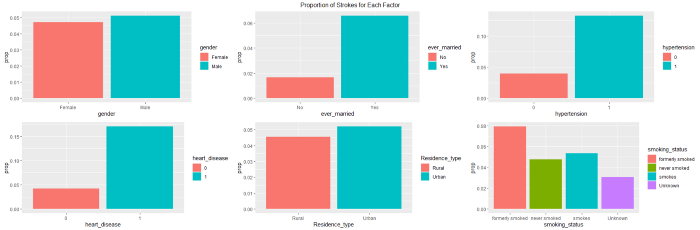


**Examine the Data to see if it is unbalanced**

The term "unbalanced data" refers to a situation in which the frequency of several categories does not correspond proportionally with one another. The following bar graph will allow us to conclude that the number of patients suffering from strokes is quite low (relatively few). The data are distributed differently.



**Investigate the possibility of a connection between the patients who had strokes**



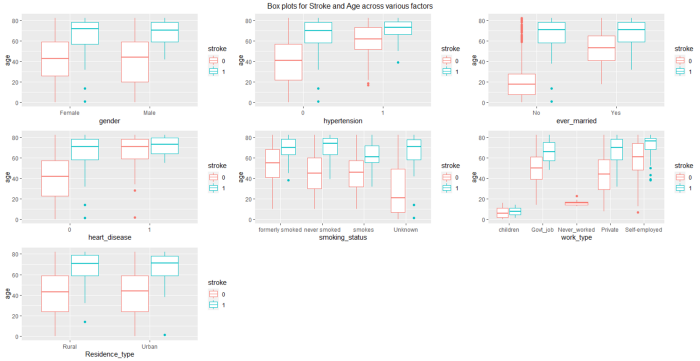
**The plots that have been shown allow for the following deductions to be made:**

There is not much of a correlation between the gender of the patient and the kind of housing in which they live. Those with a history of marriage, high blood pressure, or heart disease, and those with a greater risk of stroke.

Stroke is uncommon among children and persons who have never held a job, including those who have never worked. People who are their bosses have a higher risk of stroke than those employed by others.

Smokers have a greater chance of suffering a stroke than the overall population. This is because smoking dramatically raises the likelihood of having a variety of cardiovascular diseases.

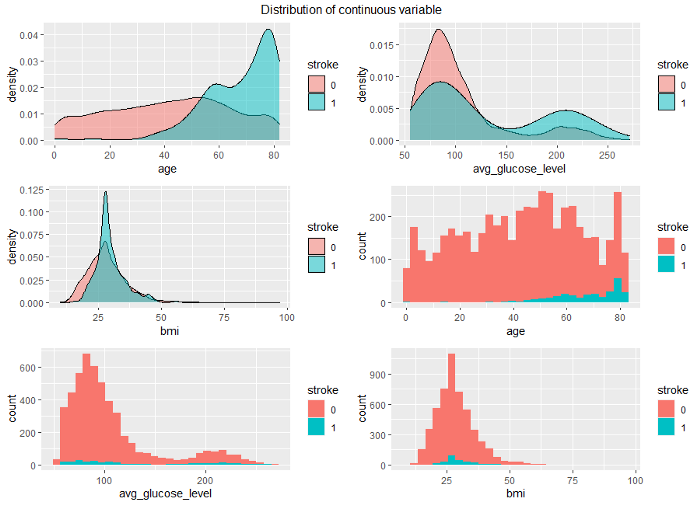
**Investigate the possibility of a correlation between age and the other factors**



**Inferences that may be drawn from the above plot are as follows:**

* Most of those who had a stroke were elderly folks in their later years.
* People who work for themselves tend to be older than the average population.
* Those who have smoked throughout their lives are at a lower risk of having a stroke than those who have never smoked.

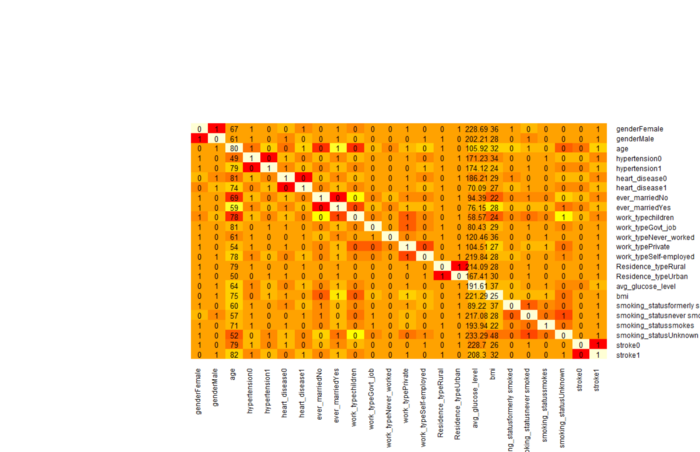
**Conduct a thorough analysis of the distribution of all continuous variables, including (BMI, average, and Glucose level)**



**From the distribution over continuous variables, one may draw the following inferences:**

* As people become older, they have a higher incidence of strokes.
* There is a bimodal distribution for the average number of strokes that occurs for persons who had strokes and those who did not get strokes.

**Examine the matrix that shows the correlation between the components**



• Age and the risk of having a stroke are strongly related to one another in a good way.

• There is a marginally significant positive connection between hypertension and stroke.

• There is a marginally significant positive association between cardiovascular illnesses and strokes.

• There is a significant inverse relationship between the occurrence of strokes and the age of children.

**Exploration of Data Analysis Conclusion**

• Based on the preliminary research conducted by EDA, we discovered that the data are very unbalanced, and the developed model can only accurately forecast those individuals who would not have a stroke—otherwise known as a guess.

• Either we should oversample the members of the minority class, or we should under-sample the members of the majority class to correct this imbalance in the data.

**Preparation of Data**

Because we know that the data are not distributed evenly, the first thing that needs to be done to be ready is to choose an acceptable sampling strategy. The size of the initial dataset is going to alter as a result of the sample, which will result in the original imbalanced dataset being balanced. The main problem arising from an unbalanced dataset is that the machine learning model cannot discriminate the minority class, which in this example is stroke; as a result, performance and accuracy are poor. This is the primary issue that develops due to an unbalanced dataset.

We have used the ROSE program to perform an oversampling procedure on the training set to rectify the unequal distribution of the data. More specifically, for the sake of this illustration, we have raised the total number of training set observations by 3000.

**The Modelling**

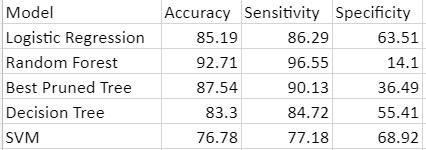
Because we are asking yes or no questions regarding the possibility of a stroke happening, we shall refer to stroke prediction as a classification issue. When it comes to determining which model performs better than others when it comes to predicting strokes in patients, the following algorithms are utilized to make that determination:

* Logistic Regression
* SVM, or Support Vector Machine.
* A Tree of Decisions
* Random Forest
* The Most Perfectly Shaped Tree

The selection of the model with the best performance is contingent on several different aspects, including the following:

* Accuracy is considered to be the single most significant criterion to consider in almost all of the models. Because the hospital wants the model to be able to forecast which patients will experience a stroke and which patients will not, accuracy is also vital in our situation. The hospital wants the model to be able to determine which patients will suffer a stroke.
* When the cost of a false positive is enormous, the measure that needs to be considered is one's ability to be precise. In our scenario, the patient who wouldn't have a stroke is forecasted as patients who would have a stroke. This would result in a significant amount of additional expenditures on the part of the patients for the treatment of stroke, which is a false alarm in this scenario.
* When a high cost is associated with false negatives, it is vital to consider metrics such as sensitivity, miss rate, and recall. In this instance, an unfair negative means that the patient at risk of a stroke was incorrectly identified as not at risk. This may result in the patient's death or cause the patient to be permanently disabled. Either outcome is possible. As a result, this is the aspect of the situation that has to be emphasized the most.

**The following table is presented an overview of the performance metrics for each algorithm:**



**The interpretation of the results from the above table:**

The preceding table allows us to conclude that all models perform admirably, with three out of the five models achieving an accuracy of more than 85%.

To choose the most significant statistic, we need to consider the ramifications that will occur in the actual world. In the circumstances presented before us, a false negative result will result in the patients' deaths. In light of this, we need to choose the model with the most fantastic accuracy while also considering the test's specificity and sensitivity.

The SVM model has the most excellent specificity, measured at 68.92%; sensitivity, measured at 77.18%; and accuracy, measured at 76.78%. Even though the precision of this model although it has a higher learning curve than the other model, we choose this one since identifying patients who have had a stroke is our first goal.

**Conclusion**

A stroke is a kind of brain damage that occurs when the normal flow of blood to the brain is interrupted. This may happen for several reasons. Strokes account for more than 795,000 cases of death each year in the United States, placing them in the sixth spot on the list of main causes of mortality. A history of smoking, diabetes, high blood pressure, elevated cholesterol levels, and diabetes in general, as well as a high public body mass index, are all risk factors. These conditions are related with aging.

There are around 201 values that the BMI characteristic cannot explain. Due to the absence of a matter, the modeling process may prove to be much more challenging, and the result may need to be more precise and accurate. It is possible to accomplish this objective in one of three ways: Delete the column that does not include any data at all. Choose whether you want to remove a cue from the dataset or try your hand at estimating the value that falls in the middle. Strokes occur far less often in younger people and those who have never been employed.

People who smoke cigarettes have a significantly increased chance of acquiring various cardiovascular diseases. People who work for themselves have a higher risk of suffering a stroke than those who another company hires. Surprisingly few people in the population will suffer from a stroke at some point in their lives (that is, relatively few). Most of those who passed away from strokes were senior citizens within a few years of their death. People who have smoked throughout their lives have a lower risk of experiencing a stroke compared to those who have never engaged in the habit. Due to the inherent flaws in the data, the constructed model could only accurately predict those individuals who would not have a stroke. This is because the model was built to account for the significant imbalance in the data. We oversampled the training set by utilizing the ROSE tool so that we could address the problem of the uneven distribution of the data.

The effect of the sample on the size of the first dataset will result in the first dataset, which was previously unbalanced, becoming balanced. When determining which of the five models offers the highest level of performance, one of the considerations that will go into the decision is that three of the models achieved an accuracy of more than 85%.

There is almost complete consensus among the models that determining the level of accuracy required is the most critical decision to make. A patient at risk of having a stroke was incorrectly diagnosed as not being at risk, and as a result, the patient was given the incorrect impression that they were not at risk of a stroke. As a result, the patient could pass away or become permanently disabled due to the false negative. This demonstrates that the patient was incorrectly classified as not being at risk of having a stroke in the first place.

**References**

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9(3), 1350-1371.

Moulton, E., Magno, S., Valabregue, R., Amor-Sahli, M., Pires, C., Lehéricy, S., ... & Rosso, C. (2019). Acute diffusivity biomarkers for motor and language outcome prediction in mild-to-severe stroke patients. Stroke, 50(8), 2050-2056.