**Stroke Prediction Model**

**Introduction:**

According to the World Health Organization, strokes are one of the top 10 leading causes of death. In 2019, strokes were responsible for 11% of deaths globally. I explored various factors that possibly influence peoples' predisposition to strokes, and created a model to predict the likelihood of someone experiencing a stroke. The model can be helpful to doctors, individuals, and other healthcare professionals who want to understand who is at risk of experiencing a stroke. The data we used to build our model was called the "Stroke Prediction Dataset" found on Kaggle. I cleaned the data and created descriptive charts and graphs to visualize and understand the data. This assisted in understanding how certain factors were correlated and how specific factors may contribute to someone having a stroke.

I started by developing a benchmark model using a k-nearest neighbor model or KNN. This model gave insight into the relationships between predictor variables and strokes. Then, I built upon this model with a decision tree which yields more interpretable predictions since it has a highly interpretable output.

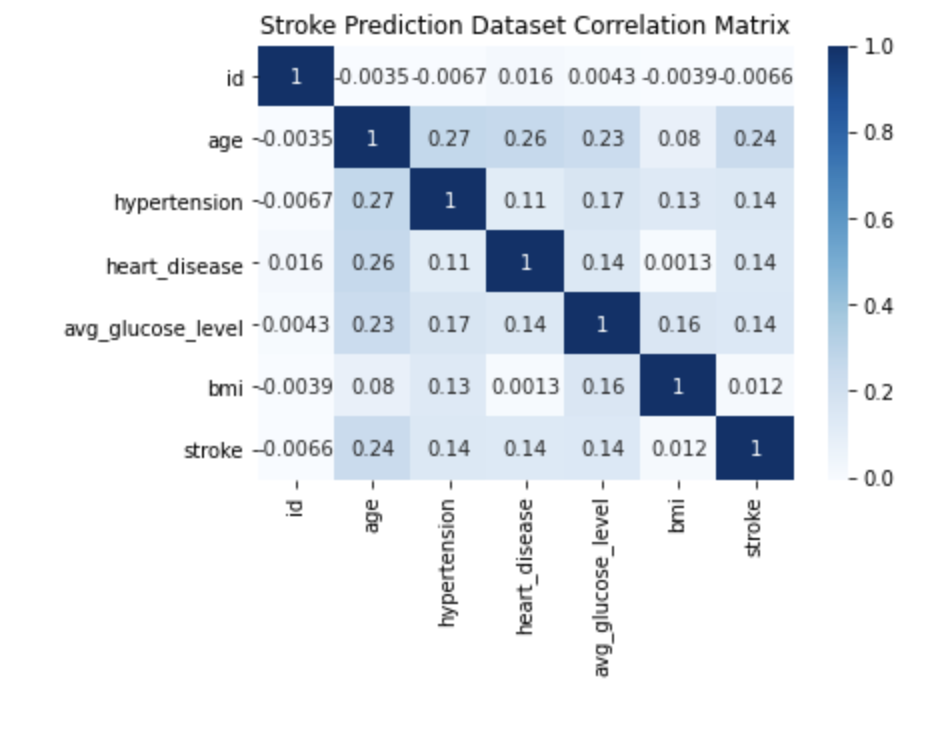
**Data Description:**

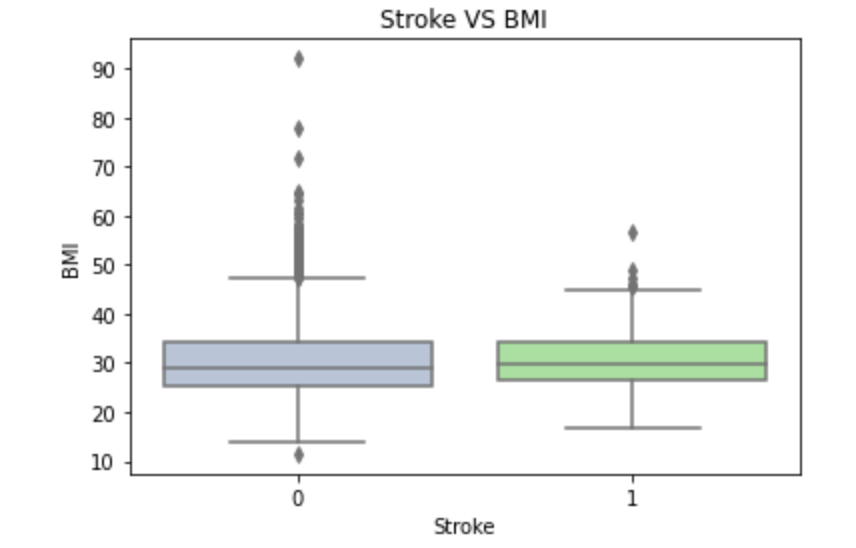
[**https://www.kaggle.com/fedesoriano/stroke-prediction-dataset**](https://www.kaggle.com/fedesoriano/stroke-prediction-dataset)

As mentioned in the introduction, the data selected was the Stroke Prediction Dataset found on Kaggle. This dataset contains 11 clinical features for predicting stroke: id, gender, age, hypertension, heart disease, marital status, work type, residence, glucose level, BMI, and smoking status. This data was collected to predict strokes and was last updated in January of 2021. The original dataset contained 5,110 observations and 11 features. During the cleaning process of the dataset, observations that had missing values for BMI and "Unknown" values for smoking status were removed. After cleaning, the dataset was reduced to 3,426 observations. The variables "Ever\_married" and "ID" were removed because the unique identifier and marital status would not be key indicators when predicting whether or not someone is predisposed to having a stroke. The predictor variables and their descriptions used in our analysis are as follows:

| **Predictor Variable** | **Description** |
| --- | --- |
| **ID** | Unique identifier |
| **Gender** | “Male”, “Female”, or “Other” |
| **Age** | Age of the patient |
| **Hypertension** | 0 if the patient does not have hypertension  1 if the patient does have hypertension |
| **Heart\_disease** | 0 if the patient does not have any heart diseases  1 if the patient has a heart disease |
| **Ever\_married** | “No” or “Yes” |
| **Work\_type** | “Children”, “Govt\_job”, “Never\_worked”, “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” |

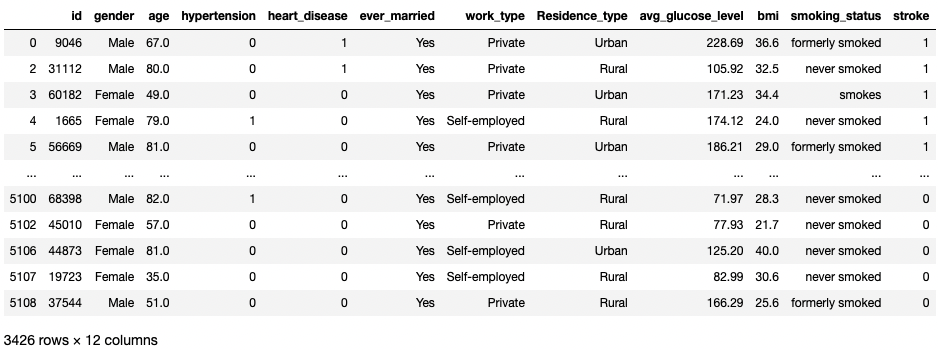
Generally, I found that the correlations between the variables were not very high. The limited number of variables in the dataset and the slightly unpredictable nature of strokes could explain this. Age seemed to have the strongest correlation (0.24) with our target variable, stroke. Hypertension and Age, Heart\_disease and Age, and Avg\_glucose\_level and Age are the predictor variable pairs with the highest correlations.

Our correlation matrix found that age had the highest correlation with our target variable, stroke. We created a boxplot to understand and visualize the variation in age between stroke and non-stroke patients. We found that the median age for stroke patients was higher than for non-stroke patients. 

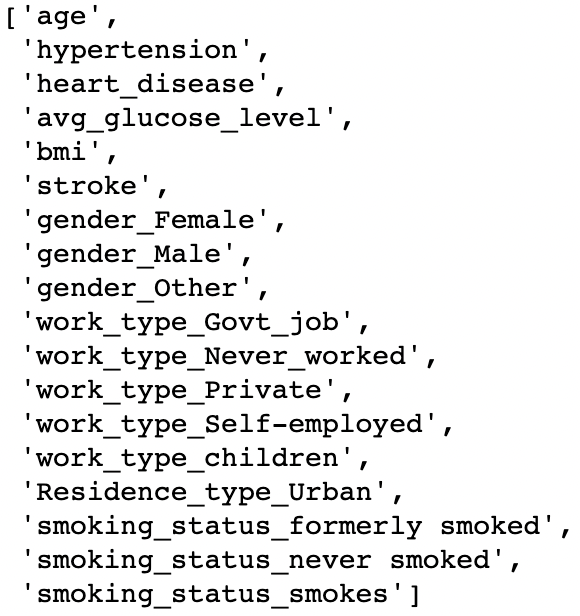
We found that the median and 3rd quartile for average glucose levels was higher among stroke patients than non-stroke patients. Although the correlation between stroke and average glucose levels was only 0.14, this graph helps visualize the difference in average glucose levels for stroke patients versus non-stroke patients. 

Furthermore, BMI has little predictive power towards predicting if someone is at risk of a stroke. We found this slightly surprising as we hypothesized that this variable would have more predictive power with strokes. In the correlation matrix, these two variables correlated 0.0012, so this output makes sense. We utilized the data visualization of our dataset by gaining a better understanding of the variables used in our dataset and their relations with other variables.

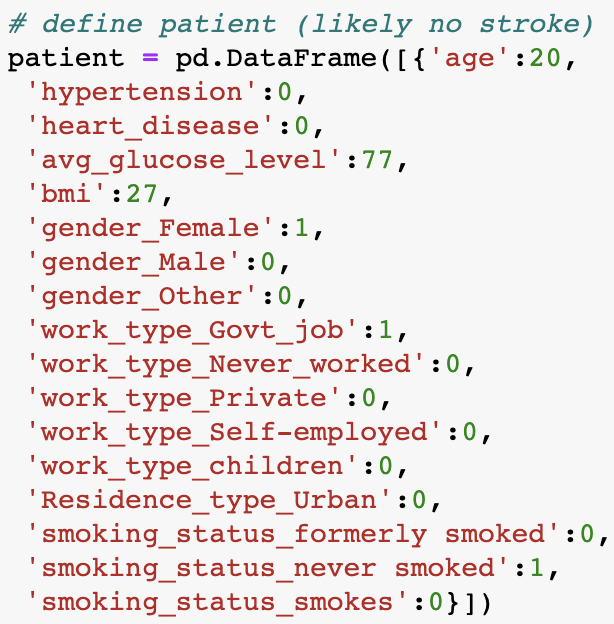
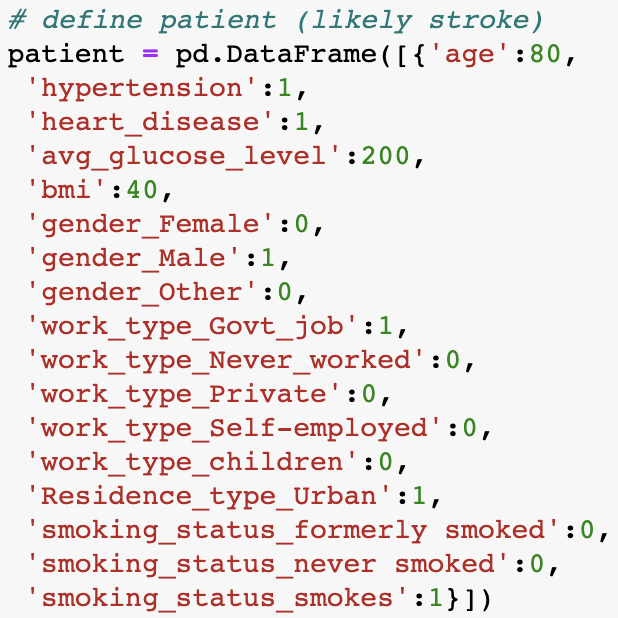
Data Cleaning & kNN Model

The first step of the data cleaning process involved creating dummy variables for the different categorical variables.

As a result, the creation of dummy created an entirely metric dataframe and deleted columns with binary counterparts. Intuitively irrelevant columns such as ID and marriage status were also deleted. The column set used for kNN analysis is shown below.

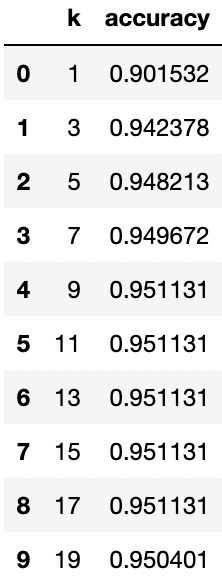


The final step of our kNN methodology was to define patients to test the model's accuracy. We created a patient that would intuitively get a stroke (left) and another that likely wouldn't (right) to see if the model would give the intuitive answer. The left dataframe output 1 (stroke) while the right output 0 (no stroke).



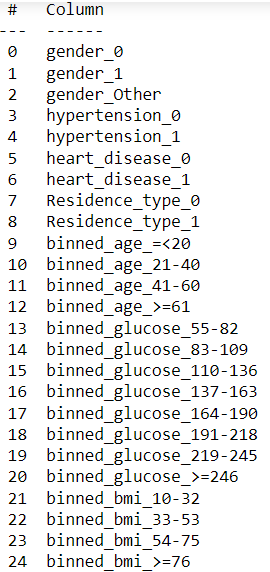
kNN Findings and Implications

We found that our kNN model was pretty accurate, with a maximum accuracy of 95.11% for k9-k17. The model also correctly predicted our two sample patients on each side of the spectrum, giving us reasonable confidence in our model. However, our kNN had its limitations which included imbalanced data (far more non-stroke than stroke patients)

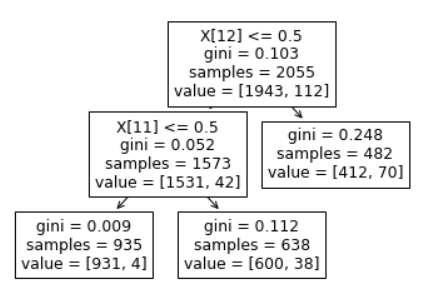


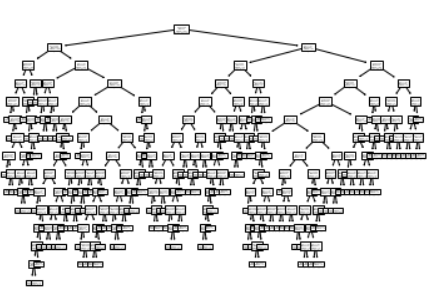
Decision Tree Methodology

Building off the understanding that kNN was able to successfully create a cluster of stroke and non-stroke patients, I now know there are key relationships within the dataset that separate stroke and non-stroke patients. Creating a decision tree will allow for more direct visibility into the factors that have the highest effect on stroke predisposition. Before using the decision tree classifier, certain variables such as gender, residence type, and smoking status had to be turned into dummy variables. In addition, the metric variables: age, BMI, and glucose level, had to be binned appropriately. And finally, the variables mentioned above, hypertension, and heart disease variables, were then changed to categorical variables. The irrelevant variables and unbinned versions of the binned variables were then removed. The list below shows the variables that are deemed as predictors of stroke.

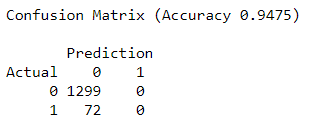


These variables (the predictors) are used to predict the stroke variable. Next, the data is partitioned into training (60%) and validation (40%) sets. Our team used the decision tree classifier to create a small and full tree utilizing the training and validation sets.

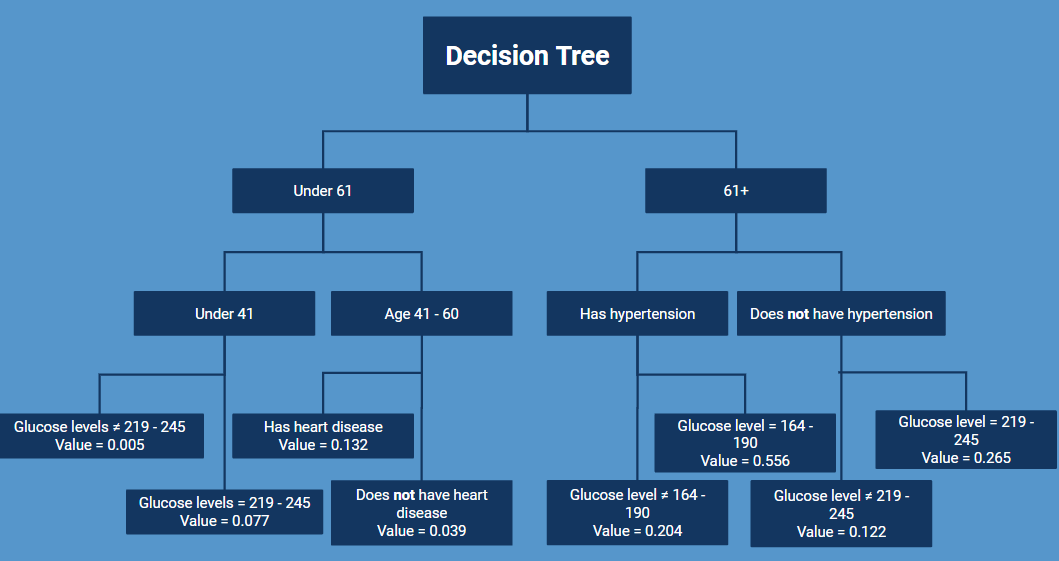




Looking at the trees and their confusion matrices, I found that the trees would initially predict no stroke for every rule. This gave the matrices high accuracy because of the dataset's low proportions of stroke victims (about 5%). However, this made the models limited and not insightful because the model does not detect strokes using the small tree. The high accuracy reflects the 95% of non-stroke patients in the dataset by predicting not having a stroke every time.

.

In addition, we discovered that the small tree was too small and was only affected by the age variable. Even after tuning the hyperparameters (within a reasonable margin), the size would not increase. Similarly, the full tree's size would not significantly decrease from changing the numbers. Removing the binned age had no real impact on the size of the tree. We believe that this was likely due to our data source's low number of stroke victims. As a result, we decided to stratify the data. Furthermore, I implemented a decision tree regressor instead of a classifier because the model only classifies groups not to have a stroke. The regressor predicts the likelihood of getting a stroke instead of predicting stroke or no stroke. Also, binned age did not become the only variable in predicting stroke. The regressor produced an appropriately-sized tree that gave more insight than the previous trees.



Findings and Implications

The decision tree provided four main takeaways. First, age, hypertension, blood glucose level, and heart disease are the most significant factors in determining risk for stroke. Second, blood glucose ranges of 164 - 190 mg/dL and 219 - 245 mg/dL indicate prediabetes and diabetes, respectively. Third, the effects of high blood glucose are amplified by hypertension. Finally, unbalanced data resulted in "no stroke" patients heavily influencing the model.

Ultimately, the tree is an easily understandable and transparent method for classification. In other words, new records can be classified or predicted. It also gives a graphical representation of the rules that follow the tree. However, single trees may be unstable and be poor at predicting. With the imbalanced data set, the magnitude of predictions may not be fully accurate, but relevant predictors are still applicable. The best way to improve the accuracy of our model would be if the dataset contained a more balanced ratio of stroke to non-stroke entries.