Data section: Tu-Yen Dang and Victoria DaRosa

Data

The dataset used in this project examines the health of numerous individuals in relation to strokes. The given data was used to predict the likelihood an individual has had a stroke. In order to do this effectively, the data was split into two different files: a training dataset and a testing dataset. With the two different datasets, the data can be trained with the training dataset, and predictions can be made with the testing data set. There are multiple key variables within the dataset. The numeric variables include age, hypertension, heart_disease, bmi, avg_glucose_level, which analyze the medical history of the individuals. The categorical variables include work_type, Residence_type, and smoking_status, which analyze different aspects of individuals' demographics and lifestyle choices. The stroke variable, indicating whether an individual suffered a stroke during the sample period, serves as a focal point for constructing a regression model. This model involves regressing the stroke variable against all other variables, which gives a broad view of how strokes play out in this dataset.

To begin cleaning the data, there were two major things to change about the columns. Firstly, all of the column names were changed to lowercase to maintain consistency (‘Residence_type’ became ‘residence_type’). Then, the first two columns were removed because they were not useful (‘Unnamed: 0’ and ‘id’). After this, the data sets were checked for null values and outliers. It was found that only the ‘bmi’ column had null values in each data set, where the training set had 159 null values, while the testing set had 42 null values. Combined, that totaled to 201 null values out of 5,110 values, which was fairly insignificant. So, the null values were replaced with the average value of ‘bmi’ to ensure the data was not skewed.

Despite this being the only technical null values in the dataset, the ‘smoking_status’ dataset also had a lot of unusable values. This was found when looking for string outliers, where it was discovered that there were over 1200 (almost 33%) ‘unknown’ recordings in the column. This was an issue because though a value was ‘recorded’, it did not provide any information, and was the equivalent of a null value. It would be unfair to move all these values into one category since it represented such a large portion of the data. So, it was decided that nothing would be done to this. There was no explanation for why there were so many ‘Unknown’ values, however this is also one of the more delicate subjects (similar to how the ‘bmi’ may be a sensitive topic to participants), so people may not feel comfortable stating their smoking status.

Outliers were also looked for in the process of data wrangling. For continuous numeric columns (‘bmi’, ‘avg_glucose_level, age’), boxplots were depicted to see if and how many outliers were present. The ‘age’ column had no outliers, but for ‘bmi’ and ‘avg_glucose_level’, there were a significant number of outliers in the boxplot. It was decided that because of how many they were (and how they were concentrated around the same areas), they would be kept in case they provided indication of a stroke. For the discrete numeric columns (‘heart_disease’, ‘hypertension’, ‘stroke’), the ‘.value_counts()’ function was used to see if there were any outliers (values that weren’t recorded as 0 or 1). It was found that there were no outliers for the section.

Finally, string outliers were checked. After looking through each column’s unique value, it was decided that the values were not consistent with their capitalization, so all characters became lowercase. This was also when the ‘smoking_status’ issue was discovered with the unknown values. Furthermore, considering the context of the lab, it was assumed that ‘formerly smoked’ would have a similar/the same impact on the likelihood of having a stroke as the ‘smokes’ value was, so they were combined into one variable. It was also realized that the

‘ever_married’ values were binary, similar to the discrete numeric columns. So, to keep the consistent format, the column became numeric and the values were changed so that 1 would represent ‘yes’, and 0 would represent ‘no’.