Are we able to predict strokes based on biological and social factors?

Megan Doan and Meg Owens

Summary:

Using the provided data, we sought to create models which would accurately predict whether or not an individual would have a stroke based on multiple variables. We created models using only quantitative data as well as models which used all of the provided variables. In order to do this, we first went to work wrangling the data. We took care of missing values, ensured categorical variables were encoded as the correct data type, checked for outliers, examined quantitative data to see if transformations were necessary, and ensured there were no data inconsistencies. Following this, we built three models in an attempt to create one which would accurately predict strokes. One model was linear regression including all the provided variables, one was linear regression containing only quantitative variables, and one was a k-means clustering model using only quantitative variables. While both created linear regression models were found to be inaccurate in predicting stroke, our k-nearest neighbor model was more interesting to examine. We used this model to predict strokes based only on quantitative variables, and it showed that there was a high risk of false positives, but a low risk of false negatives. We concluded that it was a valid and accurate model, but the risk of false positives should be made clear to anyone using it. While we were not able to create a model that was entirely accurate in its predictions, we still were able to find interesting information about how one could predict, and not predict, strokes.

Data:

In order to build a model to help us predict whether or not someone will have a stroke, we looked at numerous categorical and quantitative variables. For our quantitative variables, we looked at age, BMI (body mass index), and average glucose level. The categorical variables that we chose to observe are whether or not someone has ever been married, gender, whether or not someone had heart disease, whether or not someone had hypertension, residence type, and work type. In order to prepare the original data for analysis, we split the data set into test and training set, and then we further split both datasets to separate the response variable that we are trying to predict, which is whether or not someone had a stroke (1 if the person had a stroke, 0 if the person did not have a stroke).

After splitting the data, we go into data wrangling and exploratory data analysis. We cleaned the training set first before moving on to clean the test set. The processes that we performed to clean the training set and the test set were identical as it is good practice to prepare the datasets the same way. This helps to ensure that our model built from the original test set will hopefully be more compatible when testing on the test set.

The first thing that we did in our data wrangling process was to just observe the overall shape of the predictive variables training dataset and check if there were any missing values. After we checked if there were missing values, we discovered that there were a few missing values from the BMI column. We decided to fill the missing BMI values with the average BMI of the dataset as there were not that many missing points and this is a good way to mitigate the missing values issue. We then looked at the summary statistics of the predictive variables in the training set in order to get a better grasp of the data and to see if our categorical variables were the correct data type. When we ran summary statistics and our categorical variables were outputting quantitative statistics like mean and max, we knew that the categorical variables were

not encoding properly. Therefore, we turned the hypertension and heart disease column into a factor to ensure that it was a categorical variable and our models would read it as such.

We then moved on to check for outliers by making boxplots. As depicted in Figure 1A, we saw that both the average glucose level column and the BMI column had outliers, so we proceeded to windsorize them as our method for handling our outliers. Choosing how to handle our outliers was an problem initially as we were unsure if whether or not we should remove them the rows with outliers all together, but we ultimately decided to windsorize them as there were a lot of outliers and completely removing these rows gives us less data to train our model with. We then proceeded to examine our quantitative variables, age and BMI, by making histograms in order to check if we needed to transform these variables using functions like the arcsine of a column or the logarithm of a column. The histograms of these columns appear to show the distribution of the data well and spread out as depicted in Figure 1C and Figure 1D, so we decided that we would not need to transform these columns.

After looking at the quantitative data, we decided to check the categorical variables by ensuring that there were no data inconsistencies (ie. a female is denoted only by the phrase ‘Female’ and not also by ‘F’). We found that there were no data entry inconsistencies with our categorical variables, so the last steps that we did to prepare our data was to one hot encode our categorical variables for our model building. We also removed the columns ‘Unnamed:0’ and ‘id’ as these variables are solely for enumerating our dataset and they serve no relevance as to predict whether or not someone had a stroke.

We want to reiterate that the process that was outlined above to prepare our training dataset was replicated to also prepare our test set for our model. This is to ensure consistency as

we build and test our model. Refer to Figure 2A, Figure 2B, Figure 2C, and Figure 2D to see the boxplots and histograms associated with cleaning the test set.

Results:

After data cleaning, we proceeded to build our machine learning models in order to predict whether or not someone will have a stroke. Through our different homework assignments, we experimented with adjusting and making different variations of commonly used machine learning algorithms such as decision trees, linear regression, k-nearest neighbors, etc. We want to continue our curiosity with this project in hopes of not only building a model that is the most accurate, but also testing our foundational machine learning abilities. For this reason, we decided to build two versions of the same machine learning algorithms where one model has all of the variables included, both categorical and quantitative, and the latter model has only quantitative variables. Through this project, we will attempt to build an accurate stroke prediction model while simultaneously pushing the boundaries of traditional machine learning algorithms through experimentation.

We created two models to predict stroke based on our previously specified quantitative variables- age, average glucose level, and BMI. First, we used a model based on k-nearest neighbors. When analyzing the model we found an accuracy of 0.949169110459433, a specificity of 0.0020554984583761563, a sensitivity of 0.9958890030832477, and an MCC of 0.10133669017701596. Based on these statistics, we would conclude that this is a valid and accurate model for identifying and predicting stroke patients. It is important to note that unlike k-nearest neighbors, the algorithm K-means clustering is not used to predict specific variables or groupings, but rather to understand patterns within our dataset. Because of this and the fact that

we wanted to build a predictive model, we decided to go with k-nearest neighbors over k-means clustering. It is also important to note that the very low specificity value indicates a high frequency of falsely positive results, while the MCC indicates a weak correlation between the quantitative variables and strokes. The high sensitivity value indicates there are very few falsely negative results from our model. Based on these values, we concluded that while this method is very useful in identifying stroke patients, anyone using it should be aware of the risk of false positives. The second model we created to predict based only on quantitative variables was a linear regression model. Our analysis showed the model to have an r-squared value of 0.08882323527317315 and an RMSE value of 0.22107884414269094. When attempting to analyze the accuracy, specificity, sensitivity, and MCC of the model as we did for the previous one, we were unable to find accurate results due to the differing dimensions of created confidence tables. In summary, using the found r-squared and RMSE values, we are able to conclude that this is not a very accurate model, and we would not recommend its use for predicting strokes. For our last model, we again used linear regression but this time we included both quantitative and qualitative variables. We found the model to have an r-squared value of 0.08092637612157427 and an RMSE value of 0.2067002204971974. We were very surprised when comparing this model to our previous linear regression that was composed of only quantitative variables. We believed that the addition of the categorical variables would produce a model with higher accuracy and better R-Squared, but we found that the regression model with both quantitative and qualitative variables actually resulted in a lower R-Squared and RSME value. Based on these values, we can conclude that this is not a very accurate model and that it should not be used to predict stroke. In conclusion, we found that our basic linear regression algorithms do not accurately predict whether or not someone will have a stroke.

Conclusion:

From our three models, we found two of them to be inaccurate and not of use in predicting strokes. The third one was a bit harder to have a firm conclusion on. Our k- nearest neighbor model was found to be accurate in identifying stroke patients, but with a weak correlation and a high chance of false positives. However, there was a low risk of false negatives. Essentially, this model could identify and indicate stroke and non-stroke patients, but caution should be exercised when drawing conclusions about patients identified as having a stroke.
In regards to including only quantitative variables versus both quantitative and categorical, we found little difference between the two linear regression models predicting based on both groups respectively, both concluded to be inaccurate and not of use. For those who would say that our model should not be used due to the risk of false positives, we would concede that our models are not ideal for stroke prediction, and should someone be looking for a model to accurately predict strokes, they should not use ours.

While we enjoyed the process of building our stroke prediction models, we also believe that the work we have done may be built upon and continued in the future. One of the things that we would definitely like to consider and gather data for is medical history. There are multiple genetic diseases that can make someone more likely to have a stroke in their life, so we think that added columns that indicate whether or not someone has a specific type of disease could be very valuable. Another aspect that we would be interested in looking into is alcohol usage. While we have a variable that alludes to whether or not someone is a smoker, having a column that is also dedicated to alcohol usage could be very helpful as drinking heavily can cause blood pressure to increase thus leading to a stroke.

Appendix:

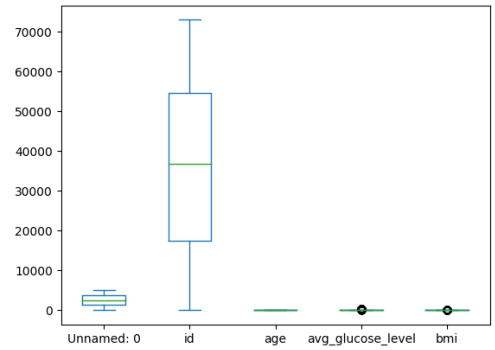

Figure 1A: Outliers within the training set before windsorizing

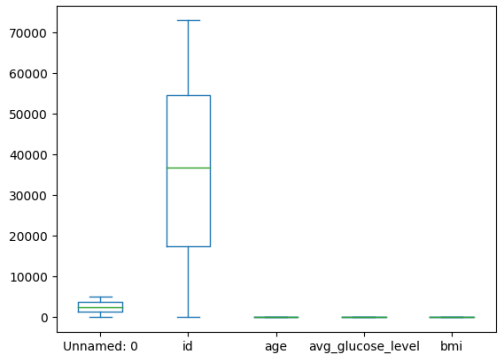

Figure 1B: Windsorizing outliers within the training set

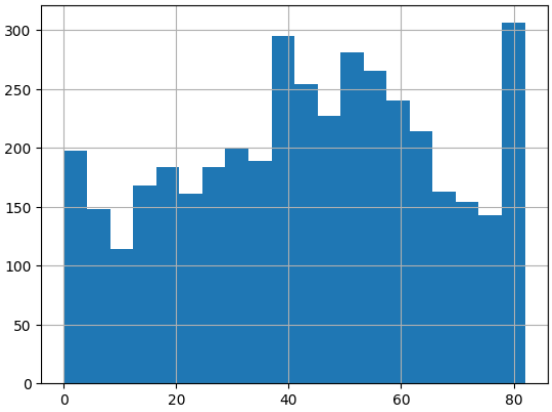

Figure 1C: Histogram of age variable within the training set

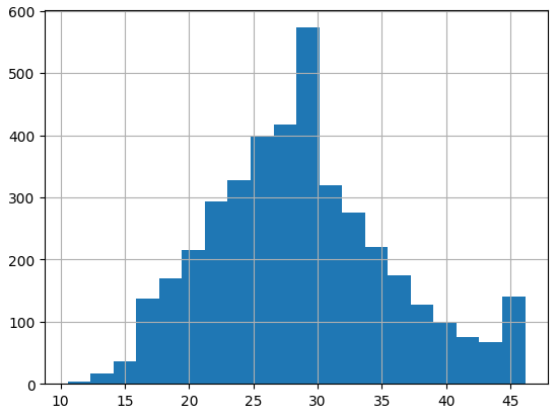

Figure 1D: Histogram of BMI variable within the training set

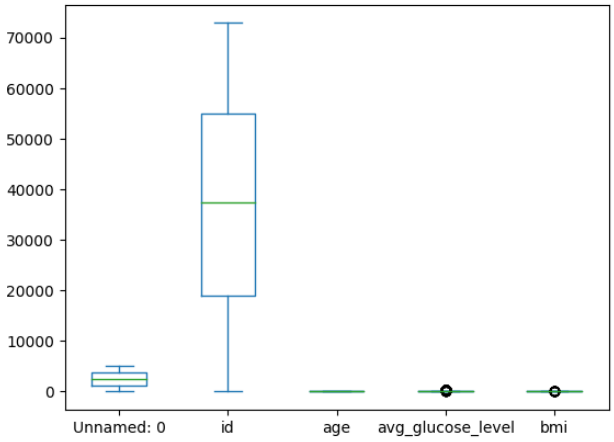

Figure 2A: Outliers within the test set before windsorizing

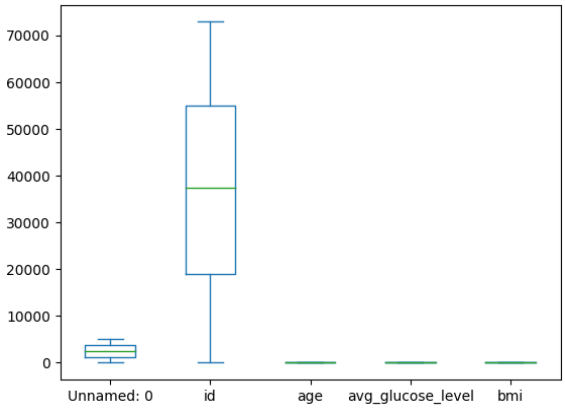

Figure 2B: Windsorizing outliers within the test set

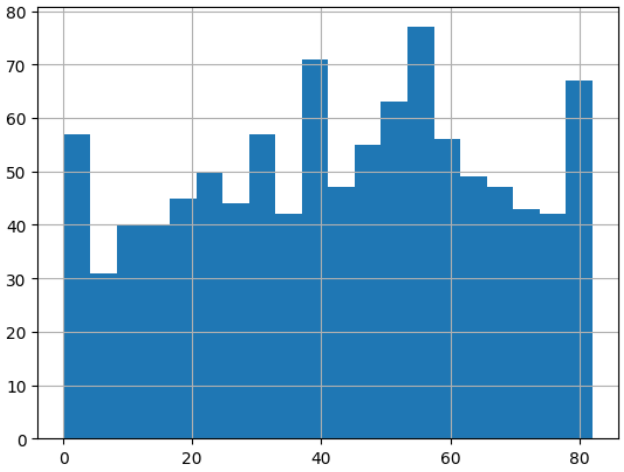

Figure 2C: Histogram of age variable within the test set

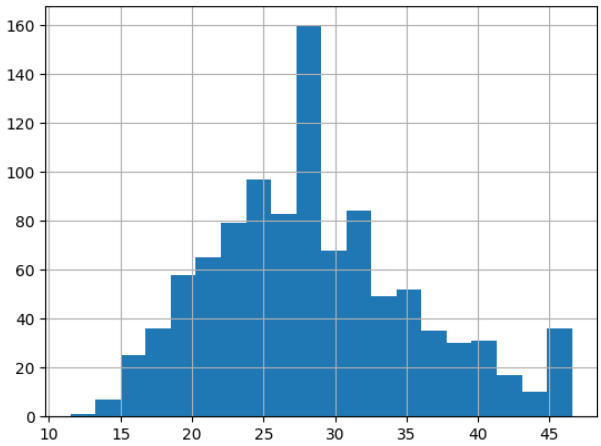

Figure 2D: Histogram of BMI variable within the test set