Summary:

Our group worked on this project to predict the likelihood of stroke occurring in individuals through utilizing different machine learning models in pertinence to specific data regarding personal, health, and demographic elements. Some of these were age, gender, marital situation, whether the individual smoked, residential status, an established heart disease, and so on. These data points, all relevant in some way, aided us in reaching our goal. Our initial method involved using a linear model with said data that was converted to numerical status, which was necessary in regard to creating such a model. However, from this initial approach, our group's findings deemed that the linear model was rather weak in terms of prediction capabilities. This was evident through the R^2 value being 0.189, and the RMSE value of 0.194. This signified that the model explained about 19% of the variability of the stroke, which leaves about 81% unexplained. The next step involved our first version of KNN, where we implemented this methodology. After doing such, we found similar results to the initial linear approximation. RMSE for this model was 0.179, while the R^2 value increased to about 0.307. The decrease in RMSE value proposes that the initial KNN approach was more effective than the linear model, though. Thus far, it seemed that our methodology was becoming more efficient. Following the first KNN model, we proceeded to perform a second one. In this version, we cleaned the data further, dropping unknown rows, and converted other categorical variables such as 'smoking_status' to numerical using dummy variables. This model also used a classification approach, rather than a regressor method utilized in the first KNN approach. In turn, we found a very high R^2 value of 0.968, and a very low RMSE value of 0.038. These results were all the more efficient than the original model provided, our initial linear model, and the first KNN model. Additionally, classification trees methodology was utilized. However, our group found that such methods weren't really useful. The tree we created only predicted two points on a graph, thus providing an RMSE of 0, and an R^2 of 1 (though there were more data points). Ultimately, the KNN model approach (specifically the second one) "wins out" over the others in predictability and accuracy of a stroke, which was our initial goal.

Data: One to two pages discussing the data and key variables, and any challenges in reading, cleaning, and preparing them for analysis.

The data sets provided numerous variables which all factored into our project and goal in predicting the likelihood of a stroke occurring in an individual. Below is a brief summary discussing each of the variables and the data overall.