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:

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. The numerous variables within this dataset reflect the numerous factors that are accounted for in stroke probability. Below is a brief summary discussing each of the variables and their significance.

Age (numeric):

- Provides the individual's age, which is significant in stroke probability in that an older person is more likely to be susceptible to stroke.

Average Glucose Level (numeric):

- Glucose level refers to blood sugar levels, which can signify other health factors or dilemmas which increase chance of stroke.

Body Mass Index (numeric):

- This variable helps signify whether one may have obesity or other underlying stroke-inducing problems.

Ever Married (dummy):

- Refers to the individual's marital status, which can relate to marital conflict increasing stress levels, thus stroke levels.

Gender (categorical):

- Refers to male, female, and other, looking at how different sexes play into stroke probability. The National Institute of Health found that men have more stroke burden, but women are associated with higher stroke occurrence in old age.

Heart Disease (dummy):

- Signals whether the individual has a present heart disease, which increases the chance of having a stroke. This is so because oxygen/blood flow is hindered if there are arteries blocked.

Hypertension (dummy):

- Does the individual have hypertension? If so, that would increase the chance of stroke due to high blood pressure causing clots, which blocks oxygen rich blood from going to the brain.

Identification (ID):

- Study identification number for each individual, which is important to have in any dataset. However, we dropped this variable in our efficient models. We dropped this variable in the second version of the KNN model, as it was deemed irrelevant.

Residence Type (dummy):

- Refers to either urban or rural living conditions for each individual. This can be applicable, and it is found that rural living conditions are associated with higher stroke occurrence.

Smoking Status (categorical):

- Asks individuals if they are a former smoker, if they never have smoked, or if they are a current smoker. Obviously, smoking leads to many health implications, thus worsening and heightening conditions of a stroke.

Work Type (categorical):

- Goes through different types of employment options, such as government, self, private sector, homeworker, and never worked. Higher stress job environments are associated with higher chances of a stroke occurring.

Stroke

- Specifies whether the individual suffered from a stroke during the sample period. It's essentially the target variable in our predictive models.

To reiterate, the above data and key variables are relevant in some way, aside from some missing data points. They all shed light on the possibility of having a stroke, in the sense that depending on the individual values of each, it might be more plausible for a stroke to occur. In regard to challenges in readings, cleaning, and preparing the data for analysis, there really weren't any. Obviously, we had to remove irrelevant features by .drop(), such as 'id', and had to convert various variables into numerics for data analysis through using dummy variables, such as 'smoking_status' and 'residence_type'. Utilizing dummy variables was a big part of our project in order to keep everything proper for analysis. Nonetheless, nothing deemed too challenging with the usage of notes and concepts learned from class. We were able to complete our original task.