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

This dataset consists of various variables related to a patient’s demographic information and health data; the models and algorithms that we will go on to create will use these variables to predict the likelihood of whether or not the patient will have a stroke. 'Age' represents the numeric age of the patients, 'avg\_glucose\_level' signifies their blood sugar levels, and 'bmi' indicates their body mass index. 'Ever\_married' is a categorical variable denoting whether the patient has ever been married ('Yes' or 'No'), while 'gender' captures the patient's gender ('Male', 'Female', or 'Other'). 'Heart\_disease' and 'hypertension' are binary variables indicating whether the patient has heart disease or hypertension, respectively. 'Id' is the patient’s identification number for the experiment, and 'Residence\_type' indicates whether the patient lives in an urban or rural community. 'Smoking\_status' is the patient’s relationship with smoking - current smoker, never smoked, or used to smoke. 'Work\_type' explains their employment status, including categories like never worked, homemaker, public sector employment, private sector employment, and self-employment. Lastly, 'stroke' denotes whether a patient suffered a stroke during the sample period.

The data wrangling section of the project went smoothly overall. We did not think that ‘ever\_married’, ‘heart\_disease’, ‘hypertension’, ‘id’, ‘Residence\_type’, or ‘stroke’ needed to be cleaned. For ‘Age’, we rounded all values into integers to simplify the dataset. For ‘avg\_glucose\_level’ and ‘bmi’, we ensured that there were no missing values in the data.. For gender, we replaced the ‘Other’ values with nan. For smoking\_status, we cleaned up the labels of the unique values - “never smoked” to “Non-Smoker”, “formerly smoked” to “Previous Smoker”, and “smokes” to “Current Smoker” . Also, we initially replaced Unknown values with nan, but this resulted in a large number of rows being dropped during modeling, so we are ultimately keeping the Unknown category to preserve these rows. Imputing Unknown values with some other category was also an option that may have yielded better results. For work\_type, we cleaned up the labels of the unique values to make them easier to understand - “Private" to “Private Sector", “Govt\_job” to “Public Sector”, “children" to “Homemaker", and “never\_worked” to “Unemployed”.

Results: Two to five pages providing visualizations, statistics, a discussion of your methodology, and a presentation of your main findings.

For both the training and testing data, we created histogram plots for ‘age’, ‘avg\_glucose\_level’, and ‘bmi’. When comparing the differences between training and testing data visualizations, there are slight variations. For ‘avg\_glucose\_level’, the histograms are both heavily right-skewed. However, the most prevalent average glucose level is closer to 85 for the training data, while the testing average glucose level is closer to a value of 75. For ‘bmi’, the training data histogram is flatter, or more equalized, that the BMI values for the testing data. signifying that there is a wider spread of BMI values. For the rest of the variables, we created bar plot visualizations. For all variables, there was no significant difference aside from the overall sample size being larger in the testing dataset. The order for ‘smoking\_status’ and ‘work\_type’ is different, but that is not a significant difference.

For the models, our group started with a simple linear regression model. A linear regression model enables us to predict the likelihood of a stroke using several independent variables. Once this model is trained on existing data, then it can be used to predict new data. The linear model that we created resulted in a RMSE of 0.206 and an R-squared value of 0.0849. These values are very close to the model that was already given to us; the RMSE shows that there is a considerable deviation between predicted and actual values, and the R-squared value shows that the model does not explain a significant amount of variance in the data.

Next, we built a linear model with polynomial expansion. The difference between a regular linear model and a linear model with polynomial expansion is that a linear model simply assumes a linear relationship between the chosen variables, whereas a linear model with polynomial expansion is able to capture nonlinear relationships by inducing higher-order degrees and allows for a better prediction of more complex relationships between variables. We varied the value of the degree (1, 2, 3, 4) and examined their respective R-squared values for each degree. The highest R-squared value produced by any of our polynomial expanded models is 0.0863, which was achieved with a degree of 2. Admittedly, this R-squared value is essentially the same as the simple linear model.

We also implemented a regression tree. Each tree consists of nodes that split based on different features and values. The root node is at the top and represents the entire dataset. As the tree grows, each split creates more nodes until the terminal nodes are reached. The depth of a tree indicates how many layers of splits it contains. Greater depth means that the model can capture more complex relationships. However, greater depth can also lead to overfitting, where the model fits the training data too closely and doesn't generalize well to new, unseen data. Our best model had a depth of 2, an R-squared value of 0.0635, and a RMSE of 0.187.

Our classification tree, which is similar to the regression tree, serves a different purpose. It differs in how it does not provide information on predicting the chance of having a stroke, but instead helps determine the most influential features including age and average glucose level. Our best model had a depth of 5, and a high accuracy of 0.941.

Lastly, for our two classifier models, they only predict 0s for the stroke variable, as there is less error when every prediction is 0 than when any 1s are predicted at all. Thus, they have the same RMSE of 0.221, and they yield the same predictions that we would get from just predicting no stroke every time.