**Data Description**

In this document, we will describe the results of the analysis for the Stroke Prediction dataset from the website https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. Kaggle is an online platform and community that focuses on data science, machine learning, and artificial intelligence by providing access to data, tools, and a platform for collaboration. The Stroke Prediction Dataset is used for research and analysis to understand the factors contributing to strokes and to build predictive models for identifying individuals at risk of experiencing a stroke. It includes qualitative variables like a person’s marital status, and quantitative variables like body mass index (BMI), for each individual. The data consists of 5,110 observations with 12 quantitative and qualitative variables.

The 12 variables are the following:

ID: An anonymous identifier for each individual.

Gender: The gender of the individual (e.g., Male, Female, Other).

Age: The age of the individual.

Hypertension: Indicates whether the individual has hypertension.

Heart Disease: Indicates whether the individual has heart disease.

Marital Status: The marital status of the individual (e.g., Married or Single).

Work Type: The type of work the individual is engaged in.

Residence Type: Whether the individual lives in an urban or rural area.

Average Glucose Level: The average glucose level of the individual.

BMI (Body Mass Index): The Body Mass Index of the individual.

Smoking Status: The smoking status of the individual (e.g., Smoker, Non-smoker, etc.).

Stroke: The binary response variable, indicating whether the individual had a stroke (0 = No, 1 = Yes).

ID has been removed for modeling purposes.

**Data Cleaning**

Most of the qualitative variables were re-coded to be binary. Given the large sample size of the data, missing values were removed.

-maybe we have to run some outlier tests to see if there are any outliers to remove

Analyze importance of each variable

Test for multicollinearity

**Exploratory data analysis through numeric summary and/or graphs**

Create graphs of each explanatory variable to those that have strokes and those that didn't.

Graph possibilities:

1. Stroke = 0, histogram of bmi (or other quantitative variable) overlap with histogram of bmi for stroke = 1. Or do stroke = 0 and 1 separately and group by gender or any other binary variable
2. Graph frequencies of the binary variables for stroke = 0 and 1

**Predictive modeling**

Logistic regression

Random Forest: Might have to compare model with missing values vs. without, if we dont, i think we have to address why we didnt. I dont think its necessary bc the dataset is so large and dont think its accurate in the dataset to replace the missing values by the median/mean since the observations are different individuals.