ISyE412 - Project Proposal

(a) Group members

Anthony Wu, Hema Priya Nakka, Ross Oglesby, Eira Puliyelil

(b) Dataset

Dataset: Stroke Prediction Dataset

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

(c) Dataset description

This dataset contains 5110 observations and 12 attributes. The 12 attributes which also serve as the column names include id, gender, age, hypertension, heart disease, ever married (marital status), work type, residence type, average glucose level, bmi, smoking status, and stroke. The id column contains a unique identifier. Gender is categorized as “male”, “female”, or “other”. Age lists the age of the patient. Hypertension is a 0 if the patient does not have hypertension and 1 if the patient does have hypertension. Heart disease is a 0 if the patient does not have heart disease and 1 if the patient does have heart disease. The ever-married column is categorized by either a “no” or “yes”. Work type is listed as either “children”, “Govt-job”, “Never_worked”, “Private”, or “self-employed”. Residence type is either “Rural” or “Urban”. Avg_glucose_level lists the average glucose level in the patients blood. BMI lists the patient's body mass index. Smoking_status is categorized as “formerly smoked”, “never smoked”, “smokes”, or “Unknown”. Stroke is a 0 if the patient did not have a stroke and 1 if the patient did have a stroke.

(d) Scientific research questions to address

1. Is there a stroke occurrence difference in gender?
2. Which age period has the most cases of hypertension, heart diseases or stroke?
3. Will there be a difference in stroke occurrence for people with different body mass indexes? What about glucose level and smoking status?
4. Logistic Regression: Which attributes can precisely predict one’s probability of having a stroke.