Stat 451
Group 7
Luke Fairchild, Samuel Geffers, David Kimel, Brenden Paddock, Sean Wells

Link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download

Dataset description:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for 11% of total deaths in the world.

This dataset has various input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient and we will use it to help us predict the risk of patients having a stroke.

Code:

```
import pandas as pd
url='https://github.com/sgeff/Stat451/blob/main/healthcare-dataset-stroke-data.csv?raw=true'
df = pd.read_csv(url,index_col=0)
```

Questions:

1. What model is best at predicting the risk of having a stroke?
2. What features or variables have the most weight when finding the likelihood of having a stroke?

Variables:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Methods:

- Logistic Regression (L1 + L2 regularizer methods as well)
- Decision Tree
- SVM
- KNN
- XGboost (if we learn)