Team Sage

# Topic Proposal

The Brain Stroke data set is from Kaggle. It provides information on how various factors affect a person causing a stroke. Presently there are a total of 4,981 entries in the dataset with 10 features which includes: gender, age, hypertension, heart disease, marital status, type of work, residence type, average glucose level, body mass index, smoking status, and stroke. Among these, 5 features are categorical, and 6 are numeric values. The research topic for this project is: What are the most accurate predictors of the likelihood of having a brain stroke?

# SMART Questions

1. What are the main factors that need to be included to predict a likelihood of a stroke?
2. How to address the data imbalance issue in the dataset?
3. What is the best machine learning model that can predict the likelihood of a stroke?
4. What evaluation metrics can be used to find the best model and why?

# **MODELING METHODS**

1. Logistic Regression
2. Decision Tree classifier
3. Random  Forest Classifier

# Source

**The source of our data set:** Brain Stroke Prediction on Kaggle: (<https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset?datasetId=2343381&sortBy=voteCount>)

**Number of Observations:** 4,981.

# GitHub

The link to our GitHub repository: <https://github.com/sowmyamaddali/DATS6101_Project_Team_Sage>