**What question or topic will you investigate? (1-2 sentences)**

Stroke ranks as the second leading cause of death in the world with an annual mortality rate of 5.5 million. Recognizing its importance in public health studies, I would like to investigate whether a patient’s likelihood to get stroke can be predicted by his/her demographic plus clinical features.

**What data set will you use? Briefly describe it, including what the source/URL is, how many observations it has, and provide a short summary of what the columns/variables are. (2-3 sentences)**

I will use the “Stoke Prediction Dataset” downloaded from Kaggle website (URL: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>). This dataset contains 5110 observations among which 249 participants had stroke event and 4861 are healthy controls. It also provides information including gender, age, hypertension (binary variable), heart disease (binary variable), marital status (binary variable), work type, residence type (binary variable), average glucose level in blood, body mass index (BMI), and smoking status.

**List and/or describe the analyses or skills you will use to answer your question of interest. These can draw from topics we have covered or will cover in class, but they do not have to. See the grading criteria for more details. (3-5 sentences)**

1. Separate the dataset into a training and a test set (ratio 8:2) using the ***caret*** package
2. Fit logistic regression model to the training set to identify statistically significant predictors for the binary outcome variable (stroke).
3. Check for potential confounding and interaction effect and adjust model if needed.
4. Apply machine learning to predict patient’s outcome using the test set and use ***confusionMatrix*** to extract evaluation metrics: accuracy (could discuss the issue of prevalence here as healthy controls are over-represented in the original dataset), precision (specificity), and sensitivity.