Group Project Description

Instructor: Cathy Poliak

Math 4322

1. Group 7: Khalyl Smith, Het Thakkar, David Oloyede, Ali Hamza Abidi Syed, Justin Wang, Mohammed Raihan Kapadia
2. Stroke Prediction Dataset:

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

   a. According to the World Health Organization (WHO) strokes are the 2nd leading cause of death in the world. We'd like to better understand what factors cause strokes to occur in so many people.
   b. 5110 observations, 12 variables
   c. Fields include:
      i. id: unique identifier
      ii. gender: "Male", "Female" or "Other"
      iii. age: age of the patient
      iv. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
      v. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
      vi. ever_married: "No" or "Yes"
      vii. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
      viii. Residence_type: "Rural" or "Urban"
      ix. avg_glucose_level: average glucose level in blood
      x. bmi: body mass index
      xi. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
         - *Note: "Unknown" in smoking_status means that the information is unavailable for this patient
      xii. stroke: 1 if the patient had a stroke or 0 if not

3. We'd like to predict if a patient is in risk of getting a stroke based on the data we produce from this dataset.
    a. Response variables will include *stroke* while *age*, *gender*, *bmi*, *hypertension*, *heart_disease*, *residence_type*, and *smoking_status* could be our predictors.
    b. We'll be using both prediction and inference to answer our question.
    c. The data question we'd like to answer is based on the data we have for each patient, what can determine if a patient had a stroke or not? Finding the answer to this can help us predict if someone is at risk of a stroke.
4. Models and methods we'll use:
    a. Since we are dealing with a qualitative response variable, we will be using logistic regression and a classification tree to answer our question.
    b. Cross-validation will be used to measure the performance of both of our models.

    Reasons: We choose logistic regression since it is easy to train, implement, interpret and efficient in use. We choose a classification tree because it doesn't require normalization or scaling of the data. We choose cross-validation as it reduces bias and gives our model the opportunity to train on multiple train-test splits.

5. Workload Distribution:

    - ➢ Het Thakkar and Syed Abidi will be focusing on the logistic regression model for our data question from 3c.
    - ➢ Khalyl Smith and David Oloyede will be implementing a classification tree on the data question from 3c.
    - ➢ Justin Wang and Mohammed Raihan Kapadia will be implementing cross-validation to measure the performance of both models.