- Summary: A one paragraph description of the question, methods, and results (about 350 words).
- Data: One to two pages discussing the data and key variables, and any challenges in reading, cleaning, and preparing them for analysis.
- Results: Two to five pages providing visualizations, statistics, a discussion of your methodology, and a presentation of your main findings.
- Conclusion: One to two pages summarizing the project, defending it from criticism, and suggesting additional work that was outside the scope of the project.
- Appendix: If you have a significant number of additional plots or table that you feel are essential to the project, you can put any amount of extra content at the end and reference it from the body of the paper.

MaryGrace Gozzi
Victoria DaRosa
Tu-Yen Dang
Charlie Perez

Summary:

The goal driving this project was to create a model that predicted the likelihood that different patients were going to suffer a stroke. A stroke, a serious medical emergency resulting from a blocked or ruptured artery, can be fatal if not treated immediately, and often causes lasting health problems. The severity of this condition emphasizes the importance of predicting stroke likelihood, because preventative measures and increased awareness have the potential to save lives. To build our model, we took both demographic information as well as pre-existing health conditions into consideration. Some of the most notable factors included 'heart-disease' and 'hypertension', both dummy variables representing whether or not a patient has heart disease or high blood pressure. Additional variables included 'avg_glucose_level' and 'bmi', both numeric variables representing their respective measures of health. The 'stoke' variable, the predictive goal of the model, was also a binary dummy variable. A value of 1 indicated that the patient had suffered a stroke during the data sample period, and value of 0 indicated that the patient had not suffered a stroke during the same period. Based on the provided variables and nature of the project's guiding question, we chose to create a classification tree to determine the likelihood of patient strokes. The most accurate classification trees had lower depth values. The trees with depth values 1 through 4 had the highest R squared value and the lowest root mean square error (RMSE) values, equalling 0.951 and 0.221 respectively. When the depth value increased beyond 4, the R squared value decreased and the RMSE value increased, indicating that the model had been overfit for the question. Working through each variable and determining the impacts of each patient's identity and health, the trees split patients into two categories, which are effectively 'stroke likely and 'stroke not likely'. Cases counted in the first column of each tree node matrix are considered 'stoke not likely', and cases counted in the second column of each tree node matrix are considered 'stoke likely'.