A Stroking Analysis of Strokes

Team 4: Andrew Kim, Benjamin Earnest, and Hunter Blum

Shiley-Marcos School of Engineering, University of San Diego

Abstract

# Predicting Stroke Risk Factors

Caused by vascular injuries in the brain, strokes are the second leading cause of death and disability in the world (Murphy and Werring, 2020, p. 1). In the United States, somebody dies of a stroke every 3.5 minutes resulting in $53 billion in costs (Centers for Disease Control and Prevention, 2022). However, prevention strategies can have a massive impact, with current prevention techniques mitigating up to 80% of strokes (Pandian et al., 2018). A major part of beginning prevention is identifying patients who are at risk of stroke. Current model accuracies for predicting strokes in patients can range from 88% to 97% (Singh and Choudhary, 2017; Emon et al., 2020). This project considers existing data based on patient risk factors, identifies the key predictor variables, and evaluates model efficacy. The goal is to identify which modeling algorithm, provides the most accurate and sensitive predictions of strokes in patients when a tested with novel data.

# Methodology

Our goal is to develop a predictive model for stroke risk, utilizing a secondary data analysis. We will use a data set consisting of 5,110 records and 12 different attributes for each record. The attributes considered for model development are gender, age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, BMI, and smoking status. The target variable in the data set is “stroke”. The data set includes continuous, categorical, and binary data. The data has been cleaned and an exploratory data analysis has been completed to support transition into the modeling phase. This has given us an initial look into which predictor variables may be best suited for use in our predictive models.

**Exploratory Data Analysis**

**Data Prepartation**

# C5.0

# CART

# Logistic Regression

# Random Forest

# Naïve Bayes

# Neural Network

# Results

# Conclusion

# References

Centers for Disease Control and Prevention. (2022, April 5). *Stroke facts*. Centers for Disease Control and Prevention. Retrieved April 13, 2022, from https://www.cdc.gov/stroke/facts.htm

Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., & Kaiser, M. S. (2020, November). Performance analysis of machine learning approaches in stroke prediction. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1464-1469). IEEE.

Murphy, S. J., & Werring, D. J. (2020). Stroke: causes and clinical features. *Medicine*, *48*(9), 561-566.

Pandian, J. D., Gall, S. L., Kate, M. P., Silva, G. S., Akinyemi, R. O., Ovbiagele, B. I., ... & Thrift, A. G. (2018). Prevention of stroke: a global perspective. *The Lancet*, *392*(10154), 1269-1278.

Singh, M. S., & Choudhary, P. (2017, August). Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 158-161). IEEE.