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 model to predict strokes in patients, utilizing a secondary data analysis. We utilized R to read in a data set from Kaggle.com, consisting of 5,110 records and 12 different attributes (R Core Team, 2021; Palacios, 2021). The arbitrary “id” attribute was removed from the data and attributes were converted to the correct type. 201 observations were missing BMI values and were subsequently removed from the data.

To explore the data, we created a regular and normalized bar plot of each categorical attribute overlaid with the target stroke attribute. When normalized the presence of heart disease or hypertension exhibited the largest relationship with having a stroke (Figure if we have room). For numeric attributes a regular and normalized histogram was coded. After normalizing, age had a clear, direct relationship with a stroke occurring in patients (Figure if we have room).

The data was split into a training and testing dataset, making up 80% and 20% of the data respectively. The training data had a large imbalance for the target stroke attribute, with over 95% of observations not having a stroke. The stroke class was balanced in the training data using random duplicate oversampling, making each class of stroke equal. We selected important explanatory attributes by using the Boruta algorithm on the balanced data. Each variable had a much higher importance than the shadow data, so every attribute was considered in modelling (Figure 2.X).

Chart, waterfall chart

Description automatically generated

Figure 2.X. Output of Boruta on balanced data. Explanatory attributes are in green and the shadow data are in blue

The attributes considered for model development were gender, age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, BMI, and smoking status. A variety of models will be evaluated on a many metrics, particularly accuracy and sensitivity.

# C5.0

# CART

# Logistic Regression

# Random Forest

Two random forest models were created utilizing the “rf” method in the caret package (Kuhn, 2021). The first model utilized the unbalanced data, while taking into account the class probabilities of the stroke. The final model had an accuracy of 95.8% on the test data. However, the unbalanced model had a sensitivity of zero. In hopes of improving the sensitivity, another random forest model was produced using the balanced data set. However, this model resulted in the same output, resulting in an accuracy of 95.8% and a sensitivity of zero.

# 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.

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

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

Palacios, F. S. (2021, January 26). *Stroke prediction dataset*. Kaggle. Retrieved April 14, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

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.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

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.