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).

**Figure 2.X.**

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

Chart, waterfall chart

Description automatically generated

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

# The C5.0 model utilized all 11 variables, which generated multiple decision nodes. Figure TBD below features the C5.0 decision tree output from the Stroke data set. Based on the model, the “age” variable serve as the root node, which splits and determines if a patient is above or not above the age of a certain value. In this case, the first age split occurs at 44. The decision tree indicates that if a patient is less than or equal to the age of 44, it proceeds to the work type of the. If a patient is greater than the age of 44, the decision tree proceeds to another age value of 66, which determines the patients’ BMI and average glucose level. shows that the variable “age” serves as the root variable, which is split and determines the types of condition or employment people of certain age ranges have and whether it leads to having or not having a stroke. 66 nodes in total. Since the variables are packed tightly, a data frame was created that included the predictor variables of the records to classify. The C5.0 model was shown to generate a 78% accuracy with a sensitivity rate of approximately 54% and a specificity rate of approximately 80%.

**Figure TBD**

*C5.0 Decision Tree Predicting Stroke Causes.*

A picture containing chart

Description automatically generated

# CART

# Logistic Regression

# Random Forest

# Naïve Bayes

# Several Naïve Bayes models were created and evaluated utilizing different predictor variables. When assessing the Naïve Bayes model, the first model that evaluated “stroke” in association with “gender” and “hypertension” generated the highest accuracy rating of 88% and the highest specificity rating of 91%. Based on this, this indicates that evaluating the gender type and hypertension level of each patient are leading indicators of a patient to have a stroke. The second model that evaluated “stroke” in association with “heart disease” and “ever married” generated the highest sensitivity rating of approximately 88%.

# Neural Network

# Association Rule

# Results

# A model evaluation was performed on all the models (regular or balanced) composed throughout the stroke evaluation to determine which was the strongest. The evaluation measured accuracy, sensitivity, specificity, precision, and F1 score of the models. The baseline models that included Artificial Neural Network (ANN) Regular, Artificial Neural Network Z, Baseline, and Random Forest Regular models produced the highest accuracy rate of approximately 96% and a 100% specificity rating, but no precision, sensitivity or F1 score ratings. The Random Forest Balanced model generated the second-highest accuracy rate of approximately 95%. Together, all models generated the highest proportion of predicted positives to actual positives. One of the strongest models through this study was the CART Cost Balance, which generated the highest sensitivity rating at approximately 98%, making it a strong model if false negatives are more crucial to avoid than false positives. The highest precision rating was found through Logistic Regression at approximately 75%, which indicated the highest proportion of predicted positives to actual positives. The highest F1 score rating was found through Logistic Regression at 74%.

**\*Note: All were made through the final model evaluations and when evaluating the comparison table.**

**Table 1**

*Model Evaluation Table for Stroke Prediction*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **F1** |
| **Baseline** | 0.9581633 | 0.0000000 | 1.0000000 | 0.0000000 | 0.0000000 |
| **Log Reg.** | 0.7600000 | 0.7900000 | 0.7400000 | 0.7500000 | 0.7400000 |
| **ANN Reg.** | 0.9581633 | 0.0000000 | 1.0000000 | 0.0000000 | 0.0000000 |
| **ANN Bal.** | 0.7612245 | 0.8292683 | 0.7582535 | 0.1302682 | 0.2251656 |
| **ANN Z** | 0.9581633 | 0.0000000 | 1.0000000 | 0.0000000 | 0.0000000 |
| **ANN Z Bal.** | 0.7479592 | 0.8780488 | 0.7422790 | 0.1294964 | 0.2257053 |
| **RF Reg.** | 0.9581633 | 0.0000000 | 1.0000000 | 0.0000000 | 0.0000000 |
| **RF Bal.** | 0.9479592 | 0.0000000 | 0.9893504 | 0.0000000 | 0.2010870 |
| **CART Bal.** | 0.7000000 | 0.9024390 | 0.6911608 | 0.1131498 | 0.2010870 |
| **CART Cost** | 0.8459184 | 0.5121951 | 0.8604899 | 0.1381579 | 0.2176166 |
| **CART Cost Bal.** | 0.5163265 | 0.9756098 | 0.4962726 | 0.0779727 | 0.1444043 |
| **C5.0 Bal.** | 0.7846939 | 0.5365854 | 0.7955272 | 0.1028037 | 0.1725490 |
| **NB Gender** | 0.8846939 | 0.2682927 | 0.9116081 | 0.1170213 | 0.1629630 |
| **NB Heart + Marry** | 0.3755102 | 0.8780488 | 0.3535676 | 0.0559876 | 0.1052632 |
| **NB Resident** | 0.5051020 | 0.5121951 | 0.5047923 | 0.0432099 | 0.0796964 |
| **NB Smoke + Work** | 0.3295918 | 0.8536585 | 0.3067093 | 0.0510204 | 0.0962861 |

# Conclusion

The predictive model is one recommended for further use in evaluating and determining the leading causes of strokes. This model has been proved to be useful in analyzing factors that can lead to certain groups of people most likely to have or not have a stroke.

# 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.

Markus, H. (2008). Stroke: causes and clinical features. Medicine. 36(11). Elsevier Ltd.

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.