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

Strokes are the third most common cause of death worldwide, and currently a major cause of adult neurological disability (Markus, 2008). In general, a stroke is a cerebral hemorrhage, that usually has a sudden onset and is difficult to predict. Many studies have been conducted to determine risk factors, and data has been collected to support these studies. This project considers existing data based on assume risk factors, identifies what may be the key predictor variables in that data, and then applies different modeling algorithms. The goal is to identify which modeling algorithm, assuming certain predictor variables, provides the most accurate predictions of stroke risk when a test data set is applied.

# 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.

# C5.0

# CART

# Logistic Regression

# Random Forest

# Naïve Bayes

# Neural Network

# Association Rule

# Results

# Conclusion

# References

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