**Ensemble Machine Learning Analysis on Stroke Data**

**Abstract**

***Background***: According to the WHO’s data from 2019, stroke is the 2nd leading cause of death worldwide, accounting for approximately 11% of total deaths (*The top 10 causes of death,* 2020). While trends have been decreasing over the past half century, research has shown many of these deaths could have been prevented. The purpose of this analysis is to evaluate how well machine learning algorithms classify stroke patients, and assess which algorithm performs the best.

***Methods***: A set of models predicting stroke was trained and tested using a Kaggle dataset including stroke, along with health and demographic variables. A set of K*-*Nearest Neighbors (KNN), logistic regression, Support Vector Machines (SVM) and Random Forest (RF) algorithms were evaluated according to the accuracy to predict stroke.

***Results***: Among the four models, the logistic regression model demonstrated the best performance in terms of both overall accuracy (accuracy=0.755, SE=0.04) and the accuracy for predicting stroke (accuracy=0.750, SE=0.07). The second best performance was from the SVM model (overall: accuracy=0.750, SE=0.03; stroke: accuracy=0.739, SE=0.10). Although the random forest algorithm performed the best in terms of overall accuracy (accuracy=0.769, SE=0.05), it had a low accuracy for predicting stroke (accuracy=0.639, SE=0.11). Lastly, the KNN model had the worst performance of the four models with an overall accuracy of 0.754 (SE=0.03) and accuracy of 0.60 for predicting stroke (SE=0.07).

***Conclusions***: While the logistic regression model had the best performance, it still did not reach the 0.8 threshold. This suggests further research is needed to train and test a model, however one limitation of this analysis was the large amount of missing data.

**Introduction**

As the 2nd leading cause of death worldwide, and the 5th leading cause in the United States, the CDC reports nearly 800,000 strokes are reported each year in the U.S. (*The top 10 causes of death,* 2020*;* George et al, 2017). Although the stroke mortality rates have been drastically decreasing over the past 50 years, the decline is largely attributed to reductions in smoking and hypertension, and approximately 80% of strokes are preventable (George et al, 2017). Additionally, the CDC reports strokes put a significant financial burden on the U.S. healthcare system; as one of the main causes of long-term disability, stroke accounts for $34 billion in costs each year (Shugalo, 2019). Moving forward, public health experts can continue addressing these preventable deaths by screening for and addressing risk factors such as hypertension, smoking, diabetes, high cholesterol, and excessive alcohol consumption (George et al, 2017). For example, the CDC found 75% of people who had a stroke also have hypertension. As an important issue in public health, numerous researchers have used machine learning techniques to try to predict strokes. For instance, Sirsat, M. S., Fermé, E., & Câmara, J. (2020) evaluated 39 studies from ScienceDirect and found support vector machine and random forest algorithms most efficiently predicted strokes. To further explore these relationships and expand upon the work of prior stroke research, this article evaluates the effectiveness of four machine learning algorithms to predict a stroke, using various demographic and health measures.

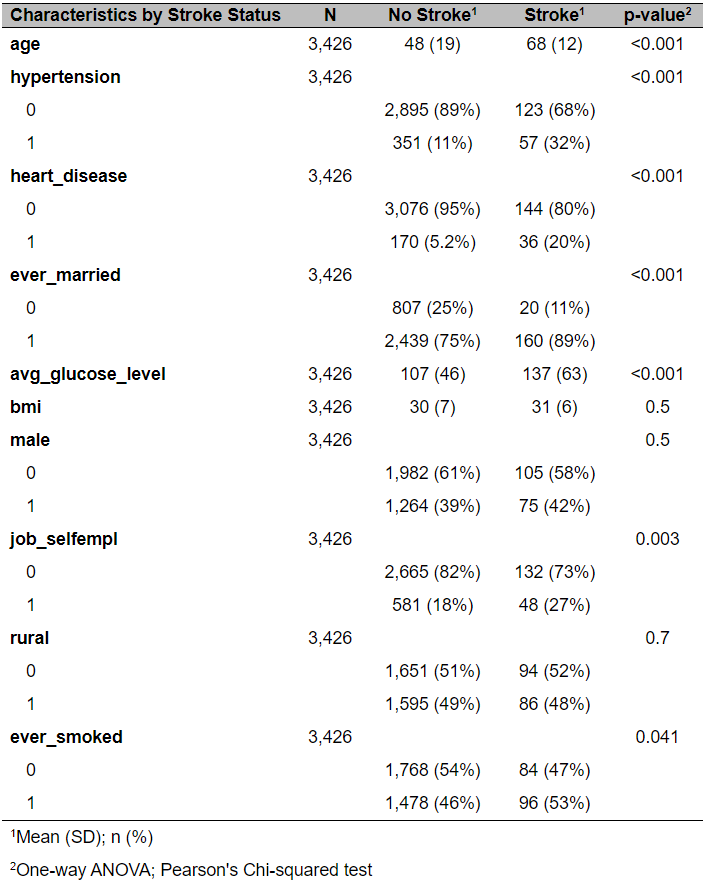
**Methods**

We used R.4.0.3 to complete all analyses; the specific packages and versions are presented in the citations. This dataset is available on Kaggle.com and can be used to predict whether an individual will have a stroke based on demographic and health-related variables. The other variables, used as predictors in our models, included age, history of hypertension, history of heart disease, whether they are married or not, whether they are self-employed, residence type (urban or rural), average glucose level, BMI and smoking status. Work type was originally a categorical variable, however we decided to create a new variable that considered whether an individual was self-employed or not (and dropped all other categories). We also combined smoking status into a binary variable to indicator whether the individual smokes or formerly smoked. These variables were selected based on availability and knowledge of risk factors. The entire dataset includes 5110 individuals, however only 3426 observations were used for our analyses due to missing data (primarily for BMI and smoking status). There were 1684 individuals with missing data for at least one of the variables. Ultimately, of the 3426 observations we considered, 180 individuals had a stroke (249 in total from the original dataset). Since only ~ 5% of the observations had a stroke, we applied SMOTE in order to adjust for the imbalanced data.

Each of the four algorithms, KNN, logistic regression, SVM (linear kernel) and RF, were trained and tested using 10-fold cross validation, using the SMOTE adjusted data. For each algorithm, we calculated overall accuracy and class specific accuracies (those the algorithms predicted a stroke and those without). For the logistic regression model, a 0.5 threshold was used for classification. The random forest and KNN models used 5-fold cross-validation to choose the best tuning parameter to train the model, and the SVM used a 10-fold cross-validation, as is the default. The random forest model used the best tuning parameters selected from a grid search including: 50, 250, 500 and 750 trees, along with 3.16, 5, 6.32, and 10 predictors (sqrt(p), p/2, 2\*sqrt(p), and p). The linear SVM model used the best tuning parameters selected from a grid search including values of 0.001, 0.05 and 0.1 for gamma, values of 1-5 for cost, and values of 0.1, 0.2, 0.3, 0.4, and 0.5 for epsilon.

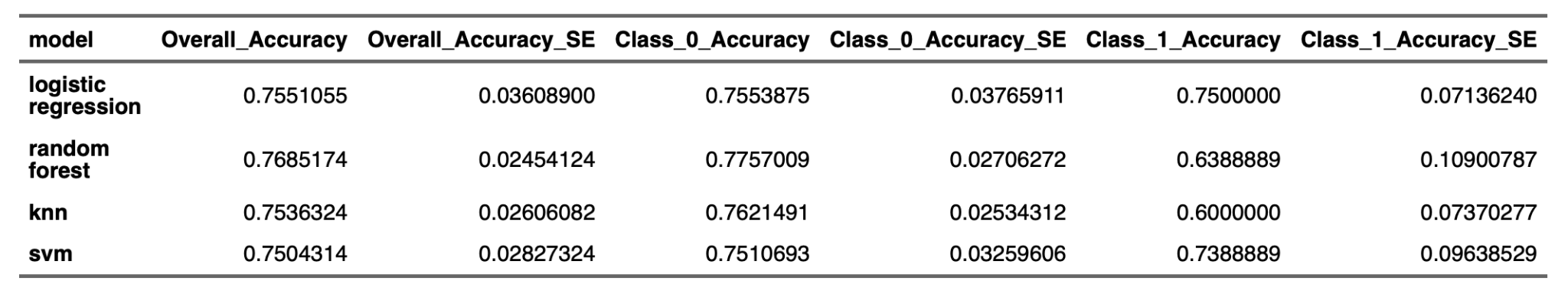
**Results**

After excluding missing data, the dataset included 3426 observations with ages ranging from 10 to 82 years old. Participants who had a stroke were generally older, had hypertension and heart disease, had a higher average glucose level and smoked. Participants characteristics for stroke were shown below (in Table 1).

*Table 1: Characteristics and Independent Variables Used in This Study*

In Table 2, we see the results for the best models obtained in our study during the training and predicting process for each algorithm. Among all the trained models, the logistic regression model presented the best performance on the test set according to the mean accuracy of stroke classification, with a value of 0.750 and standard error of 0.071. Next, the mean stroke classification accuracy for SVM is 0.739 with a standard error of 0.096. Lastly, the accuracies of stroke classifier are 0.639 (SE: 0.109) and 0.600 (SE: 0.073) for random forest model and KNN model, respectively. The mean overall accuracies for these three models are basically the same, all of them are close to 0.75 with an approximate standard error of 0.03, however the random forest model performs slightly better than the rest with an overall accuracy of 0.769 (SE: 0.04). From table 2 we can see that all of these four models performed equally well in terms of mean overall accuracy and mean accuracy for no-stroke classification, but their strengths decreased significantly when it comes to the stroke classifier accuracy which is close to 0.6 for the KNN and random forest models. Since the primary outcome for our study is to predict the stroke classifier label, in general, we conclude that the logistic model provided the best prediction accuracy for stroke classification in our study.

*Table 2: Comparison of the Classification Results for the Best Models Obtained During This Study*



**Discussion**

The main contribution of this study is to apply various machine learning models to predict the stroke classification and determine which model demonstrated the best performance. In terms of predicting stroke, the logistic regression model had the highest accuracy. Compared with other complicated machine learning methods, the logistic model is the most straightforward and easiest to utilize. The logistic regression model may have also performed better due to the large number of observations we had to exclude due to missing values. In addition, the independent variables involved in our study are easily accessed so that we can apply our logistic model to a general population with such information. However, aside from an author, Kaggle did not report specific citation information or further information about the dataset source. As a result, the specific target population is unclear.

Other studies (Siddhesh, S., 2021; Kurth, T., 2002; Park, J., 2008) indicate that BMI is associated with stroke, while ours did not. This might be because for those who have missing BMI values we simply dropped them, which leads to a loss of the number of stroke cases. And dropping observations with missing BMI potentially bias the prediction results. Future studies may consider imputation methods to impute those missing values to achieve a better prediction performance. Additionally, Siddhesh, S. (2021) also states that accuracy is not a really useful metric in the context of strongly imbalanced data. Future studies could consider additional metrics, such as AUC, sensitivity and specificity, rather than just accuracy alone.

**Citations**

1. *The top 10 causes of death*. (2020, December 9). World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>
2. George, M., Fischer, L., Koroshetz, W., Bushnell, C., Frankel, M., Foltz, J., & Thorpe, P. (2017, May 12). *CDC Grand Rounds: Public Health Strategies to Prevent and Treat Strokes*. Centers for Disease Control and Prevention.
3. Fedesoriano (2020). Stroke Prediction Dataset. *Kaggle*.Retrieved from: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>
4. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/).
5. Shugalo, I. (2019, May 17). *How Artificial Intelligence Can Predict and Detect Stroke*. Diagnostic and Interventional Cardiology. <https://www.dicardiology.com/article/how-artificial-intelligence-can-predict-and-detect-stroke>
6. Siddhesh Sawan (2021). Stroke: EDA + SMOTE + 9 Models (90% Accuracy). *Kaggle.* Retrieved from: <https://www.kaggle.com/docxian/stroke-prediction>,
7. Kurth, T., Gaziano, J. M., Berger, K., Kase, C. S., Rexrode, K. M., Cook, N. R., ... & Manson, J. E. (2002). Body mass index and the risk of stroke in men. *Archives of internal medicine*, *162*(22), 2557-2562.
8. Park, J. W., Lee, S. Y., Kim, S. Y., Choe, H., & Jee, S. H. (2008). BMI and stroke risk in Korean women. *Obesity*, *16*(2), 396-401
9. Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: A review. *Journal of Stroke and Cerebrovascular Diseases*, *29*(10), 105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162

**Package Citations**

1. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
2. Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.0. <https://CRAN.R-project.org/package=GGally>
3. Daniel D. Sjoberg, Michael Curry, Margie Hannum, Karissa Whiting and Emily C. Zabor (2021). gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.3.6. <https://CRAN.R-project.org/package=gtsummary>
4. David Gohel (2021). flextable: Functions for Tabular Reporting. R package version 0.6.2. [https://CRAN.R-project.org/package=flextable](https://cran.r-project.org/package=flextable)
5. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2020). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4. https://CRAN.R-project.org/package=e1071
6. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. [https://CRAN.R-project.org/package=caret](https://cran.r-project.org/package=caret)
7. Torgo, L. (2010). Data Mining with R, learning with case studies Chapman and Hall/CRC. <URL:http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR>
8. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>