**Stroke Prediction**

Brooke Burgess

1. Introduction

Strokes are a leading cause of death for Americans with over 795,00 annual occurrences. This is a disease that occurs when a blood vessel within the brain is blacked by a clot of ruptures causing brain cells to die due to a lack of oxygen. Strokes carry a high risk of death and is a leading cause of serious disability, making this disease a huge burden on families and communities. Risk of stroke varies based on many factors. High blood pressure and tobacco use have shown to be the most significant indicators. Risk of stroke also increases with increased age, Black ethnicity, and presence of comorbidities, especially heart diseases.

Identifying one’s risk of stroke can be useful to implement preventative measures, ultimately saving lives of many and decreasing the financial burden of stroke-related costs. Utilizing a dataset that contains stroke risk factors, we aim to predict whether a patient is likely to experience a stroke through machine learning.

1. Description of the Dataset

We will be using the Stroke Prediction Dataset sourced from Kaggle for this study. This dataset contains 5,110 observations with 12 attributes. Attributes that contained categorical data was converted to numerical for analysis purposes. The complete index and description of the observations within this dataset have been detailed below.

Graphical user interface

Description automatically generated

* Mean Age: is approximately 43 years old and Minimum: 8 month and Maximum: 82 years old.
* Males: 41.7%, Females: 58.6%, Other: 0.0196%
* Hypertension: 90%, No Hypertension 10%
* Mean Resting Blood Pressure: is approximately 106.1 and Minimum: 55.1 and Maximum: 82.
* Mean Resting Blood Pressure: is approximately 106.1 and Minimum: 55.1 and Maximum: 82.
* No heart disease: 94%, heart disease: 6%
* Mean Resting Blood Pressure: is approximately 106.1 and Minimum: 55.1 and Maximum: 82.
* 13.4% of patients are children, 12.9% of patients are Govt\_jov, 0.4% of patients are Never\_worked, 57.2% of patients are Private, 16% of patients are Self-employed
* Ever married: 65%, Not ever married: 35%
* Urban: 51%, Rural: 49%
* 17.3% of patients are formerly smoked, 37% of patients are never smoked, 15.4% of patients are smokes, 30.2% of patients are Unknown

In the following figure we are able to visualize the distribution of the data within each variable.

Chart, box and whisker chart

Description automatically generated

To analyze the correlation of data between non-target features, we reviewed correlation analysis plots between our numerical variables. Through this analysis, we observed that with an increased BMI, there was an increase in average glucose level. We also observed that BMI increased with age.

Diagram

Description automatically generated

To observe potential feature correlations with the target (stroke), we produced an interfeature correlation plot. We see that Ever Married and Age is correlated, which is not obscure. We also ran a pairwise correlation analysis which indicated age was the most correlated to stroke and work type was negatively correlated to stroke.

Chart, treemap chart

Description automatically generatedText

Description automatically generated

Within our dataset, there is a total of 4,861 patients who did not experience a stroke compared to 249 patients who did suffer from a stroke. As our target feature that we would like to predict is very imbalanced, we applied a Synthetic Minority Oversampling Technique (SMOTE) which aims to balance class distribution by randomly increasing the minority class (stroke) examples through replication. We confirmed the technique worked by checking the distribution of the two target groups pre and post oversampling. Pre oversampling distribution of stroke to not stroke was 249:4861. Post oversampling distribution of stroke to not stroke was 4861:4861.

Prior to implementing any of the machine learning methods, the dataset was processed in the following matter. The ID variable was removed as it does not supply necessary information for the analysis of this dataset. The dataset was checked for missing values and we located 201 missing variables within the BMI category. These missing variables were replaced with the average BMI of the entire dataset. Categorical data was converted to numerical data using LabelEncoder(). Features were standardized to the same scale for improved analysis using StandardScaler().

1. Description of the Machine Learning Network

We will use various machine learning models to try and predict whether a patient has a stroke or not based on the risk factors outlined in the dataset including Logistic Regression, KNN, SMV, Naïve Bayes, Random Forest, and Decision Tree.

We preformed a logistic regression analysis to estimate the probability of an event (stroke) occurring given our dataset of variables. The outcome is calculated on the probability of success divided by the probability of failure.

Diagram

Description automatically generated with medium confidence

The use of K-Nearest Neighbors algorithm tries to classify a datapoint based on its proximity to others. In order to determine distance, often the distance metric, Euclidean distance, is used to measure between the query point and other points.

A picture containing graphical user interface

Description automatically generated

Support Vector Machine is another method that can be used to classify a datapoint. This method works to find a hyperplane that has the maximum margin between the datapoints of both classes (i.e. Stroke and Not Stroke).

Naïve bayes uses conditional probability to predict whether a stroke occurred given the other conditions within the dataset did/did not occur.

Text

Description automatically generated with medium confidence

Another way to make classifications is with a Decision Tree which uses data features to split the dataset. This method utilizes a measure of homogeneity, aka entropy, and a measure of information gain to split the dataset. For our model, we utilized the Gini Index which evaluates how well a decision tree was split by measuring the frequency of mislabeled elements within the dataset.

Text, letter

Description automatically generated

Finally, we performed a Random Forest analysis which is a collection of decision trees that utilizes bootstrapping (random sampling with replacement) or bagging (bootstrap aggregation). Ultimately, this method makes a classification based of the average output of the decision trees.

1. Experimental Set-Up

Performance of the machine learning methods will be measured by their accuracy score, confusion matrix output, and classification report.

The accuracy score tells us the percentage of accurate predictions and is calculated by dividing eh number of correct predictions by the total number of predictions. It is very important when dealing with an imbalanced dataset to analyze other performance metrics as accuracy score does not predict the minority group (not stroke). This can lead to very high accuracy scores for the stroke prediction for most models if not sampled correctly.

The confusion matrix summarizes the performance of our models by outlining the predicted and actual values. This information lets us visualize if the predictions were right or wrong. True positives represent the cases that were stroke and predicted stroke. True negative represents the cases that were not stroke and predicted not stroke. False Positive represents the cases that were not stroke but predicted stroke. False negative represents the cases that were stroke but predicted not stroke.

Table

Description automatically generated

We ran a classification report on each of our models which reports precision, recall and F1-score. Precision is this percentage of correct predictions calculated by the number of true positives divided by the sum of true positives and false positives. Recall is the proportion of correct predictions of a class calculated by the number of true positives divided by the sum of true positives and false negatives. F1-score is the percentage of correct positive predictions calculated as the mean of precision and recall.

1. Results
   1. Logistic Regression

The ability for logistic regression to accurately classify if a patient would have a stroke was not extremely effective with accuracy, precision, recall and F1 scores at about 80%.

Calendar

Description automatically generatedChart, bar chart, treemap chart

Description automatically generated

* 1. K Nearest Neighbor

The KNN method preformed better than logistic regression in classifying if a patient would have a stroke with 5% higher accuracy and improved scores on the classification report and confusion matrix.

Calendar

Description automatically generatedChart, treemap chart

Description automatically generated

* 1. Support Vector Machine

SVC method was about equally as accurate as the KNN method for all performance measures .

Calendar

Description automatically generatedChart, treemap chart

Description automatically generated

* 1. Naïve Bayes

Naïve Bayes was the worst preforming method used to predict stroke.

Calendar

Description automatically generatedChart, treemap chart

Description automatically generated

* 1. Decision Tree

The decision tree method proved to be a very effective tool to predict stroke and had high performance scores across the board.

Calendar

Description automatically generated

* 1. Random Forest

The most effective machine learning tool to predict stroke was random forest with the highest performance scores compared to all other methods.

Calendar

Description automatically generatedChart

Description automatically generated

* 1. Voting Classifier

A soft voting classifier was used to train an an ensemble of the two best preforming machine learning models, the decision tree and random forest to to predict stroke based on their highest probability of a chosen class which resulted with high preformance scores.

Calendar

Description automatically generatedA picture containing graphical user interface

Description automatically generated

1. Conclusion

My experience with this data was challenging, first to convert the variables into a uniform and numerical form, second to standardize the variables and third and most challenging to find the correct way to sample the data. I ran and re-ran the machine learning methods with various sampling and splitting scenarios to find the best outcomes. . It was very important to oversample this data prior to applying the machine learning methods so they could run the dataset in a balanced format, and I was ultimately happy with the output and results of the machine learning performance results.

It was unsurprising that different methods yielded different results, but I was surprised to see them all preform above a 75% accuracy with decent classification report scores. I am unsure if the voting classifier was most effective to preform on the decision tree and random forest models due to their similarities. Ultimately, the most effective machine learning model that could accurately predict whether a patient would experience a stroke was the random forest.

Further work could be preformed on this dataset to train a machine learning algorithm to accurately predict whether a patient would experience a stroke. Namely the use of hyperparameters to tune the models.

1. References
2. Facts and statistics about stroke in the United States. Centers of Disease Control. <https://www.cdc.gov/stroke/facts.htm>
3. Stroke Prediction Dataset. Kaggle. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Other Kaggle Examples:

1. <https://www.kaggle.com/code/ahedjneed/stroke-eda-prediction-acc-97/notebook>
2. <https://www.kaggle.com/code/ahedjneed/stroke-eda-prediction-acc-97>
3. <https://www.kaggle.com/code/moduscelarent/stroke-prediction-imbalanced-data>
4. <https://github.com/Devashree21/stroke-prediction-machine-learning/blob/main/Stroke_Prediction_DSM_V1.ipynb>