IST707 – Applied Machine Learning

Stroke Prediction: Identify key modifiable risk factors that cause Stroke.

Abhijith Anil Vamadev

avamadev@syr.edu

**Abstract**

This project studies the different modifiable risk factors that contribute towards a stroke. The project first explores the data, by preprocessing the variables, by conducting exploratory data analysis on all the variables to understand their significance to the dependent variable. After the key variables have been identified and explored, models are built using different Supervised classification learning machine learning models to understand the key modifiable risk factors/variables that have the most impact, in Stroke patients.

**Identify key modifiable risk factors that causes Stroke**

Stroke is a costly disease from human, family, and societal perspective. Stroke is one of the most leading causes of death and disability all around the world. Consequently, stroke is ranked as the second cause of death in the world population. Every 40 seconds, someone in the United States suffers a stroke and even if that patient can identify and able to reach medical care within the first 3 hours, they are still likely to exhibit some form of disability post stroke (Centers, 2022).

Due to the severity of strokes, patients are often not able to lead a normal life they were once able to prior to stroke. The direct costs of stroke are estimated around $35 million dollars annually as estimated in 2019 (Jessica, 2019). The stress that causes both families and individuals make this one of the deadliest diseases all around the world. There are numerous research being done to find a solutions that will help to decrease the impact of strokes and to identify the key risk factors associated with strokes (NINDS, 2022).

**Literature Review**

Stroke is defined as a brain attack, where there is a disjoint in the continuous blood flow to the brain. A stroke occurs when a blood vessel in the brain becomes blocked or narrowed, or when the blood vessel bursts and there is hemorrhaging in the brain. Stroke patients need immediate medical attention, as longer the time goes, the more likely the patient is likely to exhibit some form of disability post stroke. Common effects of stroke include problems with muscle movement (motor sensory impairment), problems with cognitive, thinking or memory, problems with understanding speech or speaking, problems with emotions and problems with pain and sensation (Risk, 2021).

Stroke risk factors are divided into two categories: Unmodifiable and Modifiable risk factors. Unmodifiable risk factors are risk factors that can’t be changed or controlled – include age, gender, race or ethnicity or family history. Modifiable risk factors are risk factors that can be changed or controlled – include High Cholesterol levels, Smoking habits, Diabetes, High blood pressure and previous history of drug abuse (Risk, 2021).

Among the variables listed above, we will primarily be focusing on modifiable risk factors. The three most common modifiable risk factors in stroke patients are High Blood Pressure, previous history of heart disease and diabetes. These risk factors are often a result of bad lifestyle and, but some factors can also be considered hereditary.

**Research Questions or Hypothesis**

This study will explore and understand what the main modifiable risk factors are for stroke patients. Are there certain number of risk factors that are more prevalent than others for a stroke patient? If so, what are the factors that contribute to Stroke?

**Method**

The data was first collected through a Kaggle source, the data has 11 variables excluding the dependent variable for this project – Stroke (Fedsoriano, 2021). Stroke as seen from the data set, is a binary variable, making this a binary classification problem. The data has a total of 5,110 rows of data. Once the data had been collected, the first primary column, ID which was used as an identifier column for the data collection was dropped as it did not add any additional value to the study.

After the ID column was dropped, all the columns were checked for null values and unique values, to identify both null values and to identify any input errors that might have been made during data entry. Overall, there was 201 null values in the BMI column, which was imputed by the mean. Most of the variables in the study were binary variables indicating if a patient had a certain type of disease or not, there were only 3 primary numerical variables: Age, Average glucose level and BMI (Body Mass Index) of patients. The variables were then separated into numerical and categorical variables (as seen in Appendix 1).

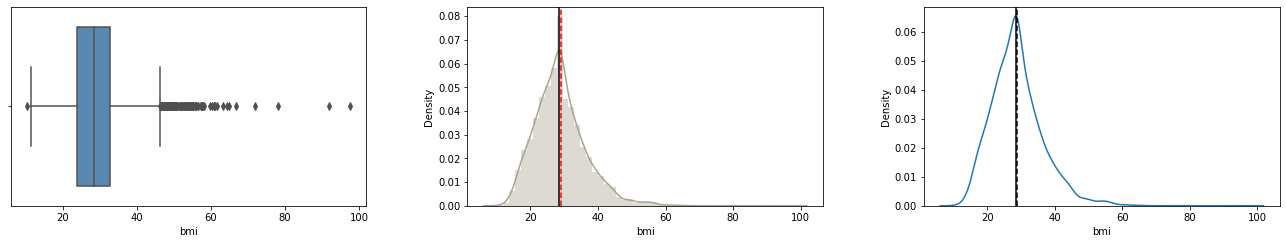
**Exploratory Data Analysis**

Exploratory data analysis (EDA) was conducted all the variables both numerical and categorical to understand their behavior better. From the EDA, the following two variables: Avg\_glucose\_level and BMI had many outliers present in them, as seen from Figure 1. To combat this, outlier detection was completed on both the variables, the interquartile range was used as a measure, so 1.5 \* Interquartile range was taken, where any number greater than this is suspected as an outlier and they were removed.

**Figure 1**

**Boxplot, histogram with density and density plot for BMI and Average glucose level**

A picture containing histogram

Description automatically generated

*Note. The first row shows the distribution of BMI, the second row shows the distribution of Average\_glucose\_level. The plot helps to understand the distribution of the variables.*

After the outlier detection was completed, the variables were again checked, to see if any further transformation needed to be applied to the variables. As seen from Figure 2, all the variables now have no outliers and look closely normally distributed.

**Figure 2**

**Boxplot of age, average\_glucose\_level and BMI after outlier detection**

Chart, box and whisker chart

Description automatically generated

*Note. The first, second and third column are the numerical variables after outlier detection as age, average\_glucose\_level and BMI respectively.*

After the numerical variables were sorted, a heat map was plotted to understand and study any existing correlation within the data. A correlation plot was made between all the numerical variables to understand the relationship between the variables. As we can see from the Figure 3, most of the variables don’t have a strong or notable relationship to one another. Only BMI and Age had a moderate relationship of 0.35.

**Figure 3**

**Correlation heat map of the numerical variables, age, average glucose level and bmi.**

Chart, treemap chart

Description automatically generated

*Note. The lighter the color, the larger the correlation magnitude. The smaller the number the lower the correlation, higher the number stronger the correlation. Correlation between two associated variables is indicated inside the shape.*

Looking at Figure 4, the data appears to be heavily imbalanced. About 95.1% of the data collected are of patients that have not had any strokes and only 4.9% of the data is about patients that have had strokes. This makes it very difficult to make a model that will predict whether a patient will have stroke or not as the data is heavily biased towards patients that has never had a stroke. Models that will be created may be biased towards predicting patients that have not had a stroke.

**Figure 4**

**Stroke – Summary of the dependent Variable.**

Chart, bar chart

Description automatically generated

*Note: The percentages included at the top indicate the number of patients for the class. 0 indicates a patient without stroke, 1 indicates a patient with stroke.*

**Model Building**

Accurate classification and reasonable intervention for high-risk population can effectively reduce the burden of stroke on families and the society. It is necessary to consider the recall of the classification model to ensure the pertinence of stroke intervention. An ROC Curve is a plot which shows the performance of a classification model at various classification threshold levels, since both classification types of strokes are equally important and since the data is heavily biased towards the majority class, ROC\_AUC\_Curve will help in assessing the model accurately than it would be accuracy. Recall and ROC\_AUC Score will be used as the primary reference when comparing models, and as a second accuracy will be used.

All the categorical variables had to be first one hot encoded and then dummy variables were created for each of the categorical variables, by dropping the first column. After creating all the dummy variables, the total number of variables for the model increased to 16. The data was then split into a 70:30 for training and testing the model.

The models that were chosen for modelling this data set, had to be based on the model’s ability to predict a reliable model, as the data was imbalanced. The following models were chosen: KNN, Decision Trees, Random Forests, Support Vector Machine (SVM) and Naïve Bayes. All models were run with a base case with the train set first, then all the models were hyper tuned with their respective parameters with Grid Search, and then tested on the train and then the test set.

**Results**

The models as specified above were run for the following metrics: Recall, ROC\_AUC\_Score and Accuracy. The models that worked best in all 3 areas, but primary on Recall and ROC\_AUC\_Score were chosen as the final model. Table 1 shows the scores for each of the hyper tuned models, in the respective areas.

**Table 1**

**Summary of scores for all the models in Accuracy, Recall and ROC\_AUC\_Score**

Application

Description automatically generated with low confidence

*Note. The numbers \* 100 is the percentage for accuracy, recall and ROC\_AUC\_score for each of the models. Boxed variable is the Naïve Bayes model, which is used as the primary model.*

Looking at the scores from Table 2, nearly all the models performed well in Test and Train accuracy, as mentioned earlier the data set is heavily biased and each of the models would be easily able to predict the majority class, and not predict the minority class. Hence, accuracy is a bad representation to primarily look for when assessing any of these models. So, Recall and ROC\_AUC score will be used as the primary measure as mentioned earlier.

**Table 2**

**Permutation Importance for variables from Naïve Bayes**

Graphical user interface, table

Description automatically generated with medium confidence  
*Note. The Importance numbers are more of an indication of how much impact each variable has, and the negative values on the Figure correspond to the variables have little to 0 impact on the dependent variable. The higher the importance value, the more impact the variable has the dependent variable.*

Nearly all the models failed to have a good recall or ROC\_AUC score but Gaussian NB (Highlighted by the purple box in Table 2) had a relatively high scores in both Recall and ROC\_AUC\_score. The Naïve Bayes model as indicated by Gaussian NB had a Test Recall score of 81.3% and Test ROC\_AUC score of 73.0% and their respective train scores were only different by 0.05% indicating a good fit from both measures and ensures that there is no overfitting with the data. Moving forward with analyzing the data Naïve Bayes will be used as the primary model, to predict the target variable Stroke.

Once the primary model was identified, Permutation Importance was used to identify the variables that had the most impact in predicting Strokes. The permutation importance function calculates the feature importance of estimators for a given dataset. The permutation importance was run 5 times and an average of the 5 scores were used to rank the variables from the most important factor to the least important factor as seen in Table 2.

As seen from Table 2 below, the 3 most important features shown by the Permutation importance of Naïve Bayes model are Heart\_disease\_1 (Patients with previous history of heart attacks) with 1.069, Hypertension\_1 (Patients with previous history of Hypertension) 0.900 and Avg\_glucose\_level (Patients average glucose level) at 0.639. These are the 3 most important variables when determining stroke according to the Naïve Bayes model.

**Discussion**

As seen previously, after running each of the models and assessing them based on their scores in Recall, ROC\_AUC score and accuracy in the respective orders, it was found that Naïve Bayes model performed the best in both Recall and ROC\_AUC score which was the primary reference for the models created. Recall scores for Naïve Bayes was 81.6% and 81.3% in the training and testing data set and ROC\_AUC score for the Naïve Bayes model was 73.5% and 73.0%, for training and testing data set. As seen from the scores, there was little to no change between testing and training data set which implies there is no overfitting.

There are several reasons why the Naïve Bayes model was able to predict with better recall and ROC\_AUC score compared to any other model as this data set suffices some of the assumptions made by Naïve Bayes. Assumption 1, In Naïve Bayes, all variables are assumed to be independent, which was fine in this case as the correlation heat map indicated that most of the variables didn’t have any strong correlation. Assumption 2, the continuous variables if any are seen as Gaussian distribution/normally distributed, which was also satisfied in this study as the variables appeared normally distributed after outlier removal. Assumption 3, Model works best when there are more categorical variables than numerical variables, as seen from the data set, there was only 3 numerical variables, 8 other variables were categorical pre-modelling. As, all these assumptions were met for Naïve bayes, and with Naïve Bayes being one of the more lenient and easier models to predict, Naïve Bayes gave the best result out of all the other models (Ghandi, 2018).

Looking at the Permutation importance of the Naïve bayes model, it was found that previous history of Heart Attack and Hypertension and High average levels of glucose were the 3 most important risk factors for determining if a patient will have stroke or not, which coincides with the findings in the literature review. Based on the literature review, previous patients with Heart Disease, Diabetes and High blood pressure were key modifiable risk factors in stroke patients. High levels of glucose or also known as Hyperglycemia is a key factor in Diabetes, and factors like Hypertension is another variant of High blood pressure. Based on the findings of this research and the literature review, the most common factors are seen as previous history of heart disease, diabetes, and high blood pressure/cholesterol, which are all seen as general indications of bad lifestyle choices (Risk Factors, 2021).

Although the study was completed, there was still some limitations. Looking back at the models, the ROC\_AUC score for Random Forests and Decision Trees were right about 50%, which indicates that there was insufficient number of features. Adding more variables to analyze during the data collection process will contribute to building a good model. Also, as suggested earlier in the exploratory data analysis section, the data was heavily imbalanced, this made it difficult to evaluate the models and build a model that would predict Stroke patients accurately, to avoid this in the future, further sampling techniques can be deployed like SMOTE, under sampling and Oversampling, that will even the difference in the classes and create better models.

**References**

1. Centers for Disease Control and Prevention. (2022, April 5). Stroke facts. Centers for Disease Control and Prevention. Retrieved May 11, 2022, from https://www.cdc.gov/stroke/facts.htm#:~:text=In%202020%2C%201%20in%206,disease%20was%20due%20to%20stroke.&text=Every%2040%20seconds%2C%20someone%20in,minutes%2C%20someone%20dies%20of%20stroke.&text=Every%20year%2C%20more%20than%20795%2C000,United%20States%20have%20a%20stroke.
2. Fedesoriano, F. E. D. E. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved May 11, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
3. Gandhi, R. (2018, May 17). Naive Bayes classifier. Medium. Retrieved May 11, 2022, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
4. Jessica. (2019, April 5). Cost analysis concludes that the direct cost of stroke, at US$35 billion annually, represents only the tip of the iceberg. NeuroNews International. Retrieved May 11, 2022, from https://neuronewsinternational.com/cost-analysis-concludes-that-the-direct-cost-of-stroke-at-us35-billion-annually-represents-only-the-tip-of-the-iceberg/
5. NINDS (Ed.). (n.d.). Stroke: Hope through research. National Institute of Neurological Disorders and Stroke. Retrieved May 11, 2022, from https://www.ninds.nih.gov/health-information/patient-caregiver-education/hope-through-research/stroke-hope-through-research
6. Risk factors for stroke. Johns Hopkins Medicine. (2021, November 15). Retrieved May 11, 2022, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke

**Appendix**

**Appendix 1**

**Table showing the variables categorized into Numerical and Categorical variables**

|  |  |
| --- | --- |
| Numerical Variables: | Categorical Variables: |
| 1. Age 2. Avg\_glucose\_level 3. BMI | 1. Gender 2. Hypertension 3. Heart Disease 4. Ever married 5. Work type 6. Residence Type 7. Smoking\_status 8. Stroke |

Note. There are 3 numerical variables and 8 categorical variables.