Stroke Prediction using Machine Learning

Saurabh Biswas

DSC 680 T301

Bellevue University

<https://saurabhbiswas1985.github.io/Data-Science/>

# Abstract

Stroke is the one of the leading causes of death according to the World Health Organization (WHO) and CDC data. According to WHO, almost 11% of total deaths are caused by stroke. CDC data suggests that, in United States, someone has a stroke in every 40sec, and one person dies in every 4 minutes because of stroke. From these statistics, we can understand stroke is a major health issue around the world.

Machine Learning can examine patient’s data, draw a pattern and identify major attributes that can end up causing stroke. This project will build and compare Logistic Regression, Random Forest and Artificial Neural Network in terms of f1-score, precision-recall, accuracy.

The aim of this project is to build an effective model to detect high risk patients who have more chances of having stroke. They can take proper precaution and address the concerns and avoid this dangerous health problem. I believe early detection of high-risk people will give them the opportunity to fix their lifestyle and take control of the important attributes that can end up causing stroke.

*Keywords:* FNA, Logistic Regression, Random Forest, Artificial Neural Network, Machine learning, prediction model.

Table of Contents

[Abstract 3](#_Toc56883934)

[Problem Statement 5](#_Toc56883935)

[Data 5](#_Toc56883936)

[Methodology 9](#_Toc56883937)

[Results 10](#_Toc56883938)

[Discussion 12](#_Toc56883939)

[Acknowledgments 13](#_Toc56883940)

[References 13](#_Toc56883941)

# 

# Domain Introduction

This project will use healthcare domain data. Specifically, it will use historical patients details who had stroke and some patients who didn’t have stroke.

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial.

There are many factors that increase the chance of stroke. These factors include:

* Hypertension
* Diabetes
* Obesity
* Smoking habit
* High Cholesterol
* Age
* Sex

This project will use vital statistics collected from patients to analyse and establish relationship with stroke outcome and different factors.

# Problem Statement

Stroke is the leading cause of death for Americans as per CDC data. It varies with race and ethnicity. Risk of having a stroke is twice for blacks as compared to whites.

Some stats on Stroke:

* In 2018, 1 in 6 deaths from cardiovascular disease was due to stroke.
* 2nd leading cause of death worldwide.
* In United States, there is a stroke in every 40 seconds
* Someone dies because of stroke in every 4 minutes in USA.
* It is one of the leading causes of disability.
* About 185K strokes – nearly 1 in every 4 are in people who already had a stroke.

There are some factors that can increases the chance of getting a stroke. Finding correlation among these factors and stroke manually is very difficult and requires expertise in medical field.

Machine Learning can help in this situation. Machine Learning can be used to analyse different stroke risk factor and detect a pattern between stroke and some of the risk factor. There are various classification models such as Logistic Regression, Random Forest etc. can be used to predict whether a patient will have stroke in future or not. Existing patient’s data can be used to train these models and then use the trained models to predict stroke using patient’s risk factors. It can help patients to control the factors that have greater influence in stroke. Also, high risk patient group can take proper precautions and measures so that they don’t have this life-threatening disease in future.

# Data

I’m using healthcare-dataset-stroke-data.csv from Kaggle. This dataset contains 5110 observation. It contains 11 input attributes and 1 output attribute that confirms whether the patient had stroke or not (1 – stroke, 0 – no stroke). Among these 11 input attributes, ‘id’ column identifies the patient uniquely and this doesn’t provide any value in predicting stroke. Other attributes may have some impact on stroke. These attributes are:

1. gender: "Male", "Female" or "Other"
2. age: age of the patient
3. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4. heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5. ever\_married: "No" or "Yes"
6. work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
7. Residence\_type: "Rural" or "Urban"
8. avg\_glucose\_level: average glucose level in blood
9. bmi: body mass index
10. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

Chart, bar chart

Description automatically generated

From the above output class bar chart, we can see that this dataset is highly imbalanced. Only 249 patients (4.87%) had stroke out of 5110 patients.

Snapshot of the of dataset:

Table

Description automatically generated

# Data Cleaning

I have used pandas profiling to gain basic understanding of the data. Panda’s profiling report shows that only ‘bmi’ contains 201 missing entry.

Graphical user interface

Description automatically generated

As we don’t have many observations, I decided to replace these missing bmi entries with its mean value, instead of dropping these rows completely.

Pearson’s correlation:

Chart, scatter chart

Description automatically generated

Phik’s correlation:

Chart, scatter chart

Description automatically generated

We can see that gender and residence\_type doesn’t closely correlated with outcome variable. Whereas ‘age’, ‘glucose\_level’ etc. have greater influence on stroke.

For the ease of our analysis, I have converted character string values into one-digit numbers for certain variable. E.g. ‘gender’ values have been converted into 1 (Male) and 2 (Female). Similarly ‘work-type’ values have been converted into 1 (Private), 2 (self-employed), 3(children), 4(Govt\_job) & 5(Never\_worked).

I have used boxplots to detect outliers.

Chart, box and whisker chart

Description automatically generated

This plot shows that ‘avg\_glucose\_level’ and ‘bmi’ – these two variables have outliers.

**Exploratory Data Analysis**

Performed EDA to find out impact of each risk factor on stroke.

Chart, histogram

Description automatically generated

Age v/s Stroke distribution plot shows that chances of stroke increase after 55yrs of age and it is at maximum risk at 80yrs of age.

Chart, line chart

Description automatically generated

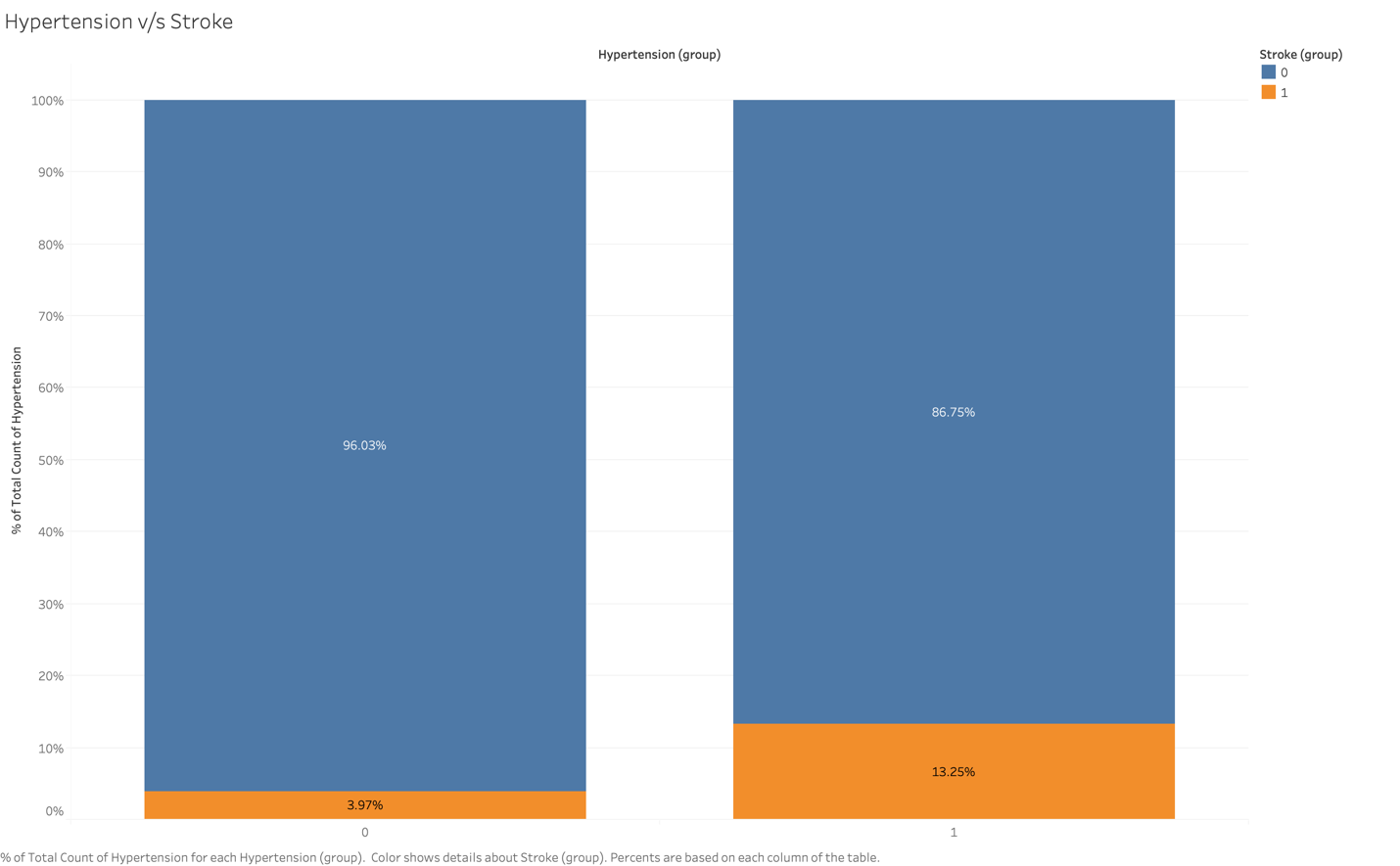
We can see if glucose level is below 100 or above 150, then chances of having stroke is more.

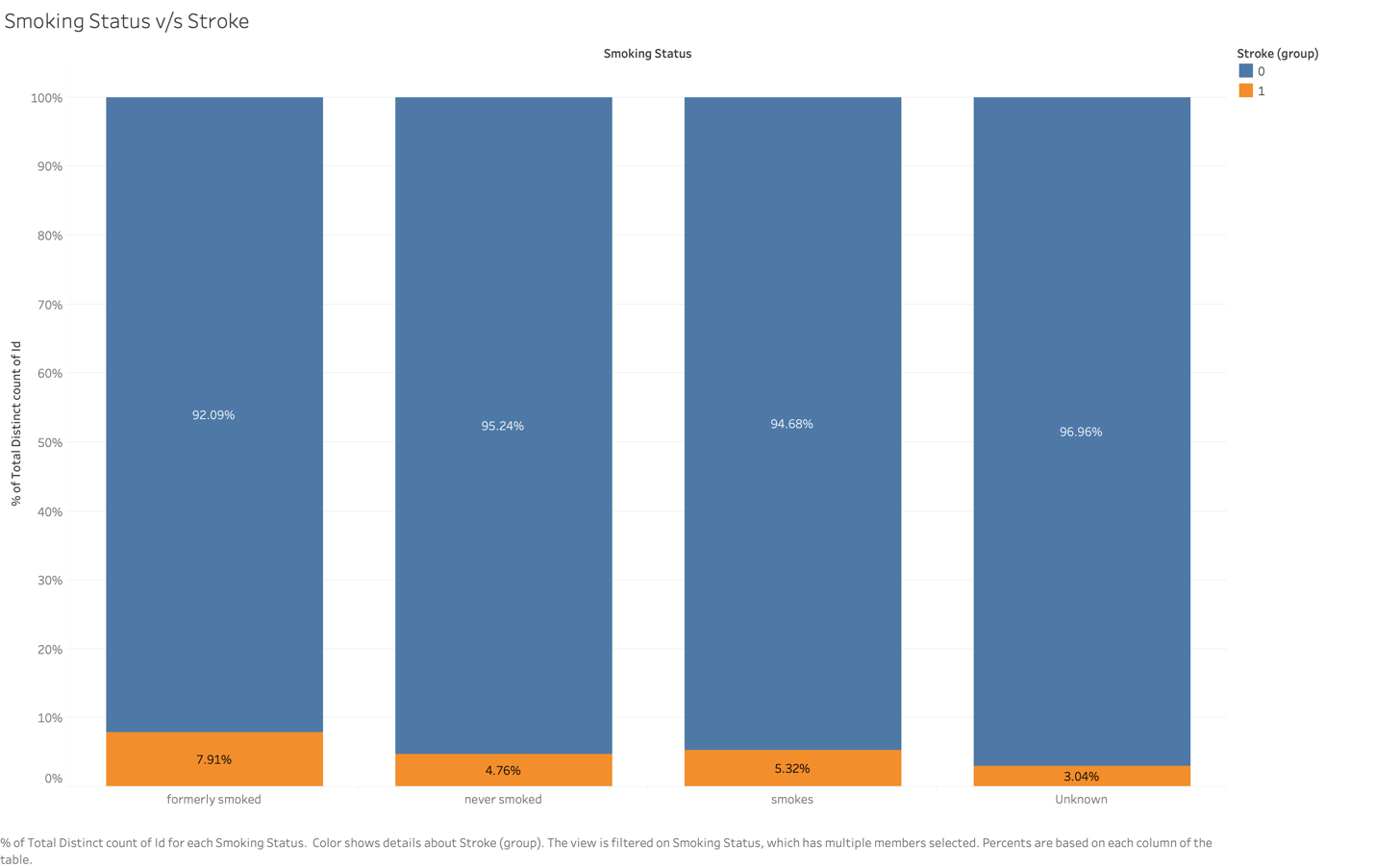
Chart, line chart

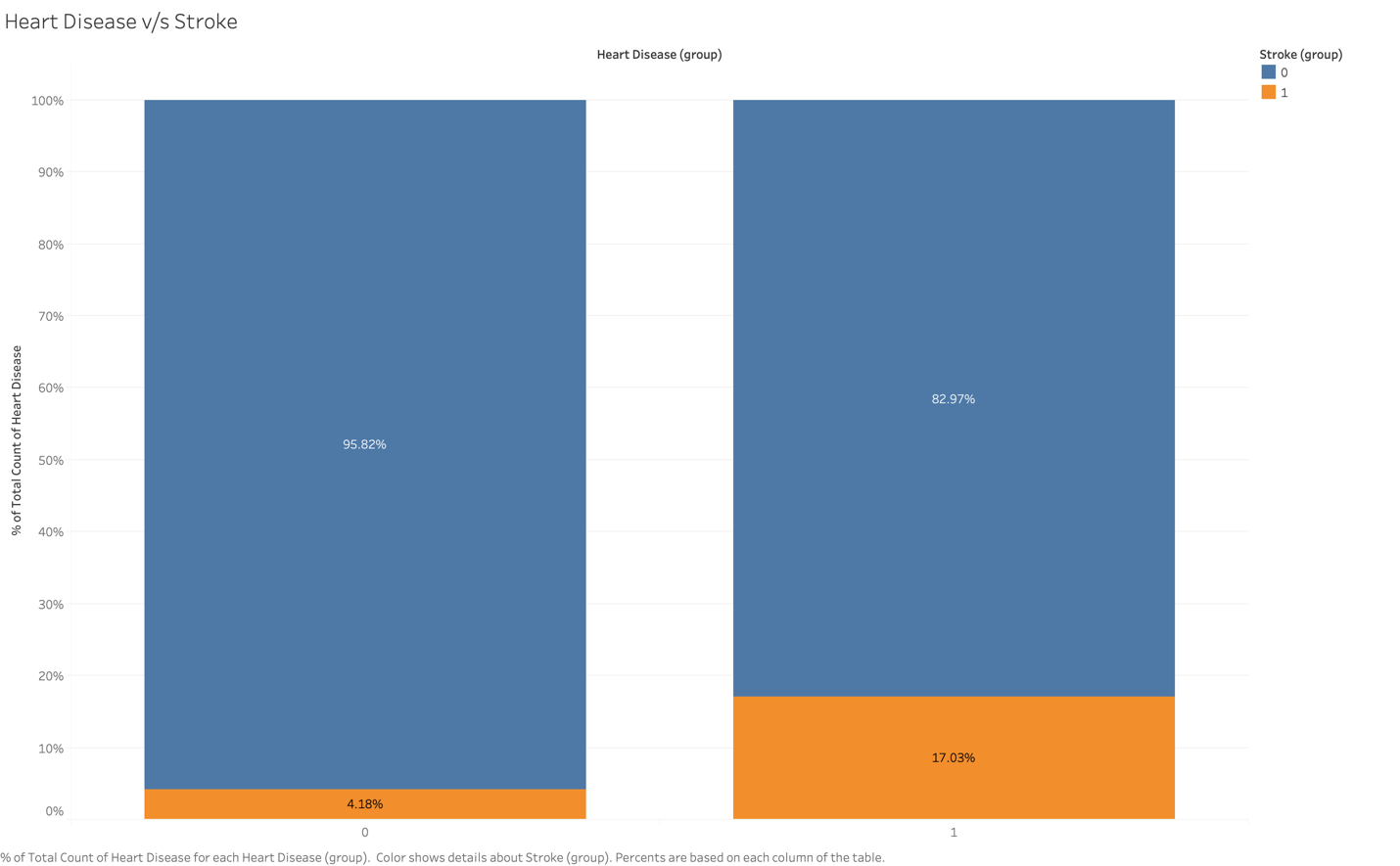
Description automatically generated

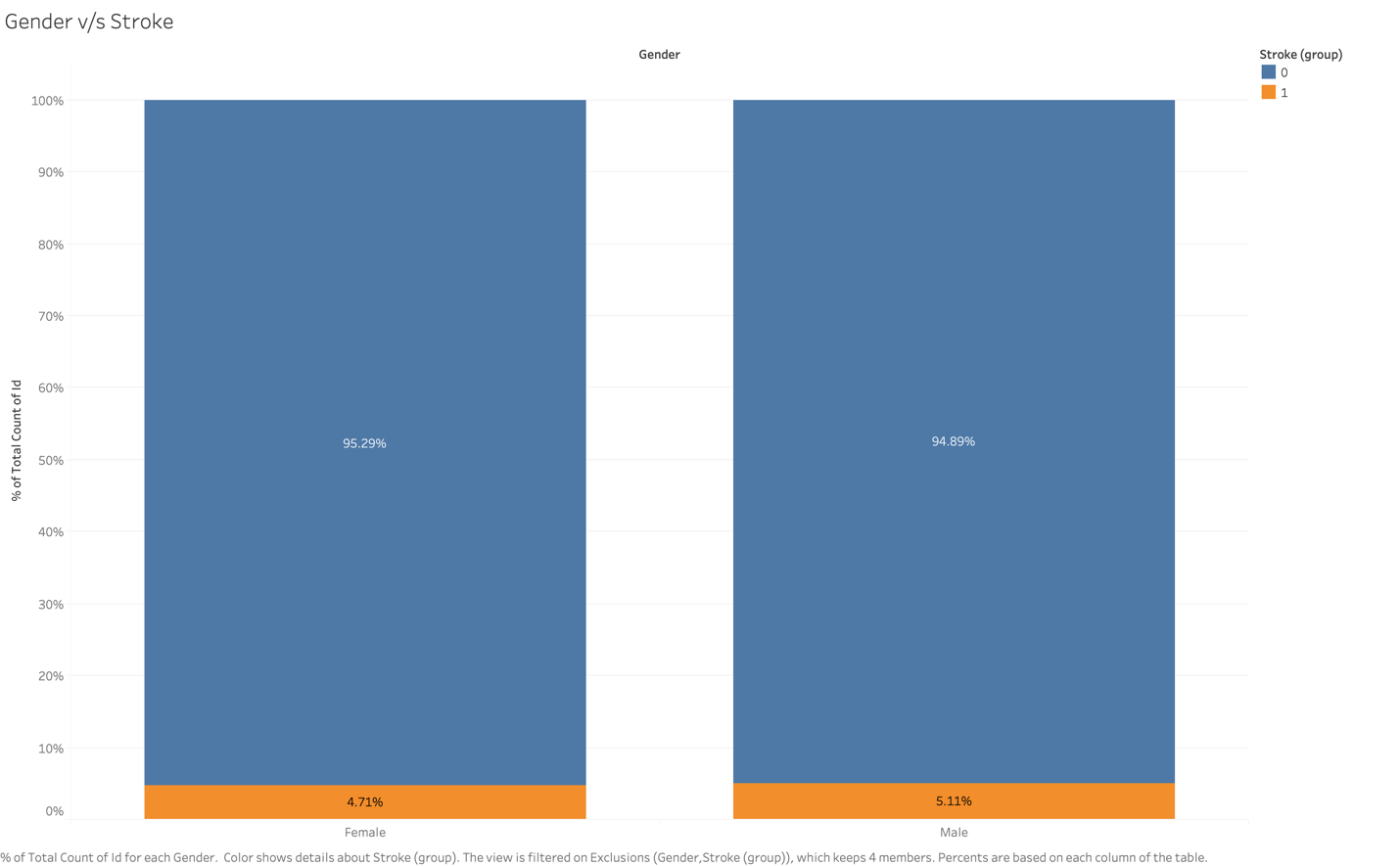
Impact of bmi on stroke isn’t very clear. We can see 28-30 bmi has increased chances of stroke.

Some other influencing factors:









**Data Preparation**

Removed outliers from glucose\_level using z\_score value. There are 4990 entries after removing outliers.

Dropped ‘id’ as this doesn’t have any impact on stroke outcome. Standardized the attributes values using RobustScaler() function.

Divided training/test dataset into 75/25 ratio after preserving the minority/majority class balance using stratify option.

Most of the models don’t do very good with imbalanced dataset. To resolve this issue, I have used SMOTE() function to oversample the minority class. After oversampling, we have 3557 observation in each class of training dataset.

# Methodology

Our aim is to build a classification model that will predict stroke using risk factors. For this purpose, we are planning to use logistic regression classifier, Random Forest classifier and Artificial Neural Network. Logistic Regression classifier will establish the baseline training results and we will compare that with results from Random Forest & Artificial Neural Network.

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least square to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for nonnormally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables, which is not always met in morphoscopic datasets. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions.

Random forest is a classifier that evolves from decision trees. It actually consists of many decision trees. To classify a new instance, each decision tree provides a classification for input data; random forest collects the classifications and chooses the most voted prediction as the result. The input of each tree is sampled data from the original dataset. In addition, a subset of features is randomly selected from the optional features to grow the tree at each node. Each tree is grown without pruning. Essentially, random forest enables a large number of weak or weakly-correlated classifiers to form a strong classifier. The Random Forests algorithm had a substantial impact on medical image computing over the last decade.

Artificial Neural Network works like human brain. It uses one input layer, one output layer and multiple hidden layers. It adjusts weights on hidden layer based on input attributes and output class variable from training datasets.

# Results

I have used sklearn’s GridSearchCV to fine tune the parameters for above model and then compared the results. We have used seeding for reproducibility.

I have plotted random forest results using CalibratedClassifierCV() function to know whether I have to perform any calibration or not. But based on the below plot, I have decided not to do any further calibration.

Chart, line chart

Description automatically generated

Our parameter tuning results are as follows:

|  |  |  |
| --- | --- | --- |
| Model | Best Parameter | Avg Precision Score |
| RF | class\_weight: none | 0.998 |
| max\_depth: 40 |
| max\_features: sqrt |
| n\_estimators: 200  bootstrap: False |
| LR | logistic\_\_C: 0.1 | 0.818 |
| select\_best\_\_k: all |

|  |  |  |
| --- | --- | --- |
| Model | Best Parameter | Avg Precision Score |
| ANN | learning\_rate\_init: 0.01 | 0.953 |
| max\_iter: 300 |
| Solver: adam |
|  |

Based on the average precision score, I have selected the best parameters for Random Forest, Artificial Neural Network and Logistic Regression classifier. Also based on the average precision score, we have selected Random Forest as our best model.

We have then tested Random Forest models using our test dataset and calculated accuracy and f1-score. f1-score is a harmonic mean of precision-recall. We have also produced confusion matrix for each test.

|  |  |
| --- | --- |
|  | **Random Forest** |
| **Accuracy** | 0.91 |
| **F1-Score** | 0.157 |
| **Avg Precision Recall Score** | 0.13 |

Random Forest Confusion Matrix:

|  |  |
| --- | --- |
| TP: 1119 | TN: 67 |
| FP: 51 | FN:11 |

**Conclusion:**

* All the Models performed Pretty Well.
* Accuracy result from Cross validation are close for all the models
* Selected Random Forest model based on the Cross-Validation accuracy score and it is easier to explain to all stakeholders.
* Results from Test Dataset show that our decision to go with Random Forest is correct.
* Trained Random Forest model on Entire Dataset using best Parameter.

# Discussion

Early detection of high risk patients group of possible stroke, can be predicted accurately by the use of machine learning techniques. This may result in the decrease of health cost and may provide ample time to patients to address their risk factors. In this project, the Random Forest has been determined to be more superior to Logistic Regression or ANN since it provided higher average precision score.

# Acknowledgments

I want to acknowledge Prof. Catie William for her guidance and valuable feedback in completing this project.

# References

CDC website: <https://www.cdc.gov/stroke/facts.htm>

Kaggle: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 183-192).

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3), 1350-1371.

Lehmann, J. F., DeLateur, B. J., Fowler Jr, R. S., Warren, C. G., Arnhold, R., Schertzer, G., ... & Chambers, K. H. (1975). Stroke rehabilitation: Outcome and prediction. Archives of Physical Medicine and Rehabilitation, 56(9), 383-389.

Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., & Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3110-3113). IEEE.

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial intelligence in medicine, 101, 101723.

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263-1265.

Lin, C. H., Hsu, K. C., Johnson, K. R., Fann, Y. C., Tsai, C. H., Sun, Y., ... & Taiwan Stroke Registry Investigators. (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Computer Methods and Programs in Biomedicine, 190, 105381.

Singh, M. S., Choudhary, P., & Thongam, K. (2019, September). A comparative analysis for various stroke prediction techniques. In International Conference on Computer Vision and Image Processing (pp. 98-106). Springer, Singapore.