What is a Stroke? Disease Prevention and Prediction

**Yi Ren   
Masters of Science, Business Analytics  
California State University, East BayHayward, California**[**yren12@horizon.csueastbay.edu**](mailto:yren12@horizon.csueastbay.edu)

**Beatrice Nicole Tanlapco   
Masters of Science, Business Analytics  
California State University, East BayUnion City, California**[**btanlapco@horizon.csueastbay.edu**](mailto:btanlapco@horizon.csueastbay.edu)

**Victory Ehizonomen   
Masters of Science, Business Analytics  
California State University, East BayHayward, California**[**oehizonomen@horizon.csueastbay.edu**](mailto:oehizonomen@horizon.csueastbay.edu)

***Abstract*—This paper discusses factors and potential prevention attributes of stroke prediction using data provided by Kaggle gathered from medical institutions that has been made available for educational purposes. The goal of this project is to identify reasonable stroke factors using data analysis through visual data exploration and modeling with machine learning. We hope to elevate the active utilization of data analytics in healthcare to prioritize stroke awareness amongst providers, patients, and research institution owners.**

# Introduction

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Stroke is the leading cause of long-term disability. Because of its high morbidity, mortality, disability, recurrence, and complications, stroke is ranked by the medical community as one of the top three diseases that threaten human health, along with coronary heart disease and cancer.

Although stroke is irreversible, early recognition of the warning signs at different stages of stroke can minimize the probability of stroke. Identifying and analyzing the causes that would contribute to the onset, as well as the timely control and condition is the main reason why we would like to develop a stroke prediction model. The goal of this research is to utilize different patient attributes such as age, hypertension status, heart disease status, BMI values and smoking status and apply different machine learning methods to predict whether a patient is likely to have a stroke or not. We obtained the Healthcare Stroke Dataset provided from Kaggle.com to predict whether a patient is likely to get a stroke based on the input parameters from the dataset like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patients.

Our effort is to push for a priority on recognizing early-stroke signs to providers, patients, and further research in institutions by using analytical methods and observations to narrow down on the likely causes of these life-threatening damages.

Stroke prediction analysis is oriented towards three main groups:

1. Those who have had the disease before, and high-risk groups who have not yet developed the disease. Data visualization provides an edge to helping patients understand their current situation and alarm for necessary actions to be taken. For example, when considering factors of reducing obvious risk, if a patient has a history of heart disease, one should not drink alcohol and must control weight and average glucose levels. These reactive measures taken from visual education will help reduce the overall chance of having a stroke.

2. Pharmaceutical and medical research institutions. Solving, predicting, and understanding the methods taken to identify the risk probability of potential patients who may develop the disease in the future period allows institutions to create a foundation of additional measures for physicians to change the medical and preventive care of their patients.

3. Commitment to artificial intelligence. Artificial intelligence and doctors are working increasingly close with each other. AI systems have demonstrated a high potential for diagnosing diseases, analyzing medical images, and predicting health outcomes.

According to a survey conducted in the United Kingdom, the new system developed by researchers at the University of Nottingham can scan a patient's daily medical data and predict that people are likely to suffer a heart attack or stroke within 10 years. Compared to standard measurement and prediction methods, the AI system correctly predicted 355 more patients than the standard method. About half of these heart disease and stroke patients had not previously been flagged as "at risk". With the benefit of stroke prediction, physicians can therefore take preventive measures, such as lowering cholesterol levels through prescription drugs. Successful predictive models will hopefully move from laboratory studies to clinical applications. In that case, busy primary physicians can use artificial intelligence tools with pattern recognition capabilities. The tools could be used to flag patients and remind other doctors to focus on them.

# Data Exploration

The Healthcare Stroke Dataset is an educational resource available on Kaggle. This dataset we have selected has 5110 observations, which is not large, but is fine for building a stroke prediction model. The 12 variables selected by the creator all have some correlation with the items to be predicted, but it is still not certain that an absolutely accurate and reliable accuracy can be derived. Preparing multiple algorithmic classifiers for training data is the way to arrive at reliable accuracy. If the accuracy of multiple classifiers can reach about 80%~90%, then it proves that the model has more accurate prediction results. Of course, if we encounter an unbalanced data set, accuracy cannot be the only criterion for detection, and the priority is to integrate the data.

The attribution list is as follows:

| **Attribute** | **Attribute Description** |
| --- | --- |
| id | Unique patient identifier |
| gender | “Male”, “Female”, or “Other” |
| age | Age of patient |
| hypertension | 0 if patient doesn’t have hypertension, 1 if patient does |
| heart\_disease | 0 if patient doesn’t have any heart disease, 1 if patient has a heart disease |
| ever\_married | “No” or “Yes” |
| work\_type | “children”, “Govt\_job”, “Never\_worked”, “Private”, or “Self-employed” |
| residence\_type | “Rural” or “Urban” |
| avg\_glucose\_level | Average glucose level in blood |
| bmi | Body mass index |
| smoking\_status | "formerly smoked", "never smoked", "smokes" or "Unknown"\* *\* = Status Unavailable* |
| stroke | 0 if patient never had a stroke, 1 if patient had a stroke |

Table 1 : Data attributes

# Goals of analysis and benchmark

We will first perform our analysis by using graphical representation. For example, an analysis by gender will depict which has a higher probability of having a stroke - men or women. The analysis by age can find out which age group is at high risk of getting the stroke, too. From hypertension, smoking status, to heart disease, we can also observe if other existing conditions are common within stroke victims. Visualizing our dataset can serve as an exploration and give us a common ground on any misconceptions we may have before diving deep into our analysis.

However, a single visualization practice between our dataset’s variables is not enough. There cannot be only one reason a patient will have a stroke. For example, if we take two patients with the same BMI - one patient may have had a stroke and one may have not, causing an overlap. Does this mean the hypothesis of the high BMI factor is incorrect? Not really, because BMI is not a clear indication of stroke in this example. This justifies our reason for further analysis by overlapping other conditions.

The next step is to look for other interesting patterns in the data. The different classifiers will be used to make the prediction of strokes in patients. By comparing our results between different ML methods, we can pick out which will achieve the highest accuracy, lowest false positive rate, and false negative rate.

The next step is to look for other interesting patterns in the data. The different classifiers will be used to make the prediction of strokes in patients. By comparing our results between different ML methods, we can pick out which will achieve the highest accuracy, lowest false positive rate, and false negative rate.

# Data Analysis Process and Tools

There will be four stages to this project. This project will include the main steps of data pre-processing and cleaning, exploratory data analysis, feature engineering and modeling using machine learning, and followed by the conclusion of our analysis in the next section of this report. The main tool we use is Python, while Microsoft Power BI will be used for our graphical analysis.

*A. Data Pre-processing and Cleaning*

The dataset is composed of numeric and categorical variables. It has a total of 12 columns and 5111 rows, containing the medical records of 5110 stroke patients. According to the test, this dataset has 201 missing values in the BMI column, and this column also has several abnormally high outliers. In Python we imported and installed the required packages to begin working on the dataset such as Numpy and Pandas.

*a) Changing data types*

The original attribute types of our dataset also contained a mix of int, object, and float attribute types (Fig. 1). We’ve reconciled our data set types and changed objects to categories, and float categories like age and stroke indication to integers (Fig. 2). We also seemed to face an imbalanced dataset where stroke patients make up for only 4.87% of our dataset. In this preprocessing step, we just focused on handling missing data.

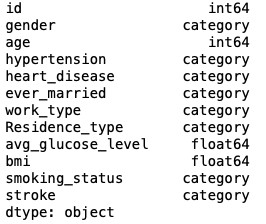
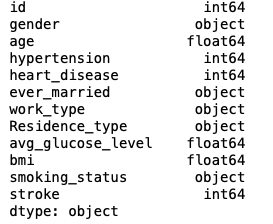


Fig. 1. Before changing data types Fig. 2. After changing data types

*b) Handling the missing values*

For patient values that are duplicate, they will be dropped. This means that if there is a patient that appears twice, such that the Patient ID is repeated, the row will be dropped. However, our dataset did not have that issue. For the BMI column with 201 missing values (Fig. 3), we imputed the median for NaN. BMI was the only attribute that contained missing values (Fig. 4).

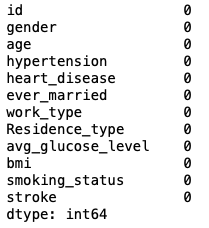
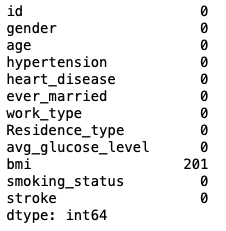


Fig. 3. Before Imputing Fig. 4. After Imputing

# *Exploratory Data Analysis*

A stroke is a life-threatening medical condition that needs urgent treatment to avoid further complications. Using the publicly accessible stroke prediction dataset to analyze and create metrics using different Power BI visuals to aid early detection and factors that will or could contribute to the illness. It was segmented according to a specific parameter by grouping the age; Marital status to stroke, avg glucose level, BMI by age group, stroke by age group, stroke by smoking status concerns, analyzed using visualizations like the pie chart, donut chart, clustered column chart, line chart.

*a) Correlation Matrix*

Before we explored the data further, we used the correlation matrix to roughly analyze the correlation between features (Fig. 5). The graph shows that age, job type, marital status, heart disease, hypertension, and smoking status have relatively strong correlations.

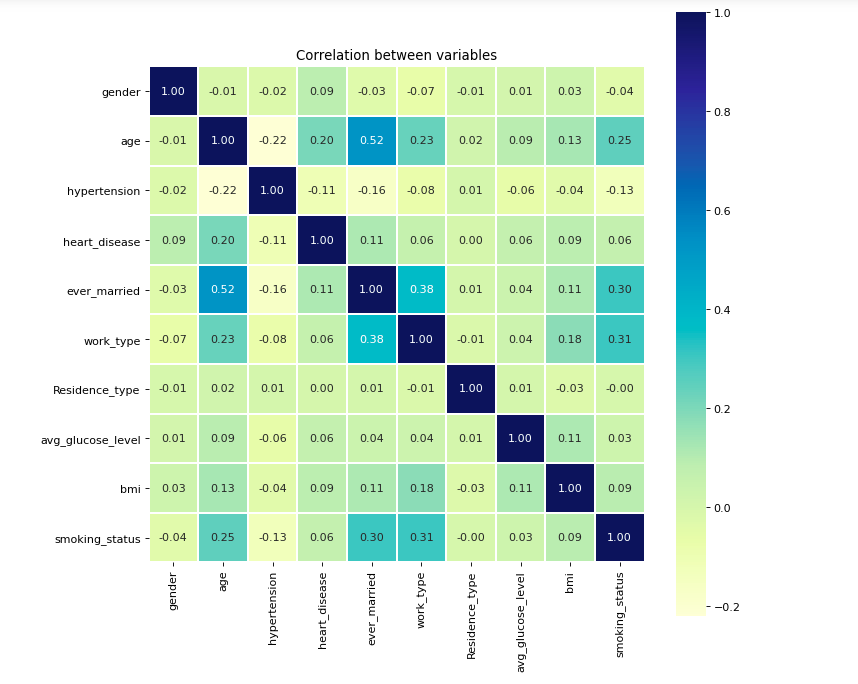


Fig. 5. Correlation Matrix *(Correlation between variables)*

* + - 1. *One-hot Relationship*

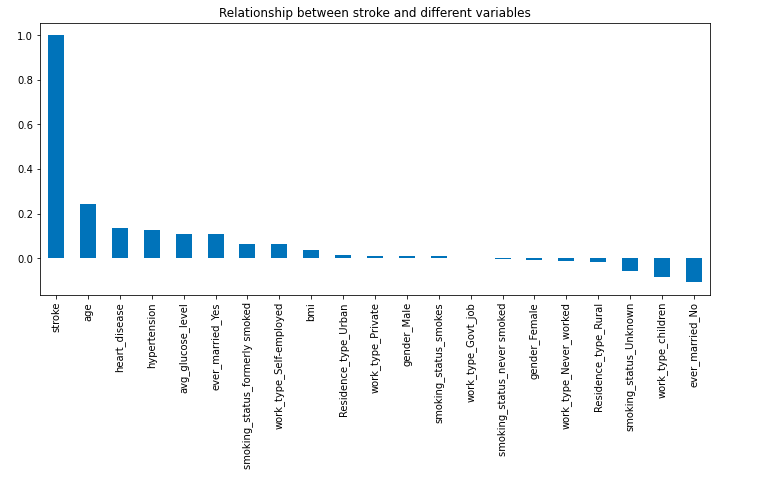
 Use one-hot encoding to view the relationship between each feature and the predicted target (Fig. 6). Stroke is less related to gender and type of residence site and can be directly discarded, while work status is retained due to its high correlation coefficient with freelancing.

Fig. 6. One-hot Relationship *(Relationship between stroke and different variables)*

* + - 1. *Stroke by Gender*

Using Microsoft Power BI, we have dove into further understanding of our own dataset and have compared the different attributes against if the patient had a stroke or not. As a simple first example in “Fig. 7”, we see that the dataset shows that females have a slightly higher representation in profile compared to their male counterparts. Females accounted for 58.55% of the profile.

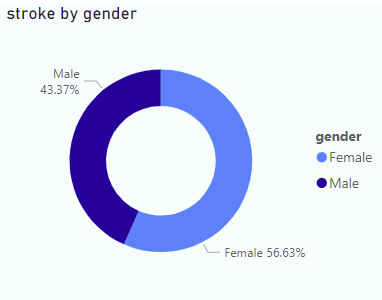


Fig. 7. Stroke by Gender

* + - 1. Stroke by Age Group

Per “Fig. 8”, we noticed that stroke risk can also vary by age. Although we are aware that stroke risk increases with age, our visualization allows us to see that strokes can - and do - occur at any age. According to the CDC, 38% of people hospitalized for stroke were less than 65 years old in 2014. At 105, the 71-80 Age Group, compared to the 18-30 Age Group, compared to the 18-30 Age Group which had the lowest stroke at 2. This age group accounted for 42.17% of stroke patients in the dataset.

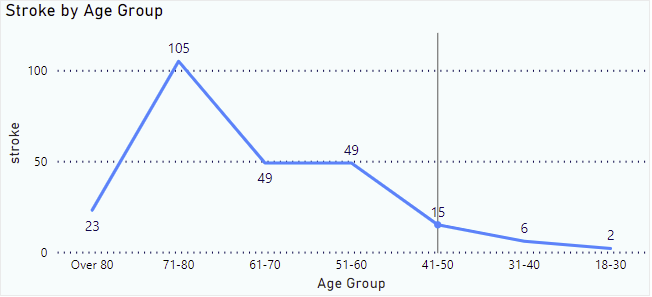


Fig. 8. Stroke by Age Group

* + - 1. *Stroke by Smoking Status*

In “Fig. 9”, we observe the stroke victim Smoking Status. According to the data available, people who never smoked had the highest Stroke Attack at 90, followed by Formerly Stroke, Unknown, and Smokes. Never Smoked accounted for 36.14% of Stroke Attack. This observation suggests that people who never smoked can also be a victim of a stroke

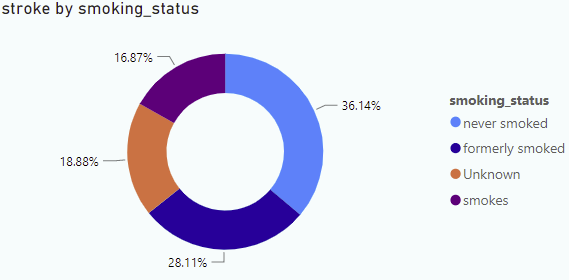


Fig. 9. Stroke by Smoking Status

* + - 1. *Correlation between Stroke and other leading cause of death*

“Fig. 10” depicts the three attributes Heart Disease, Hypertension, and Stroke in respect to Age Group. At 105, the 71-80 Age Group had the highest stroke count and was 11.4% higher than the 31-40 Age Group. The 71-80 Age Group accounted for 42.2% of stroke. Across all 7 Age Groups, strokes ranged from 2 to 105. heart disease ranged from 1 to 115, Hypertension ranged from 11 to 144, which shows that stroke and heart disease are positively correlated.

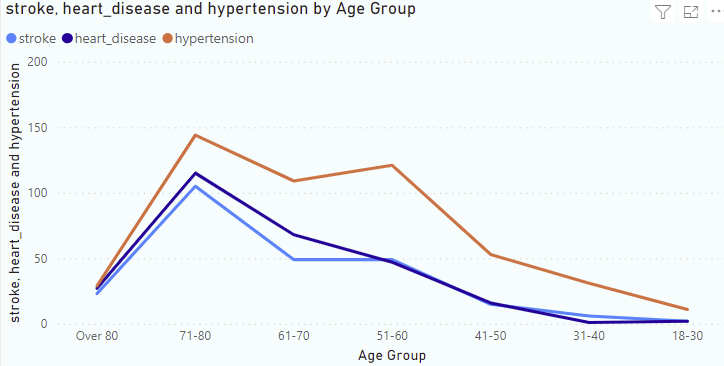


Figure 10: Stroke by Age Group

* + - 1. *Avg Glucose by BMI*

From “Figure 11”, At 31.75, the 51-60 Age Group had the highest Average BMI and was 29.67% higher than 18-30, with the lowest Average BMI at 24.48. The average Glucose Level and Average BMI diverged the most when the Age Group was Over 80 and the Average Glucose Level was 88.98, higher than the average BMI.  Across all 7 Age groups, the average BMI ranged from 24.48 to 31.75, and the average glucose level ranged from 94.04 to 116.98.

***Chart, line chart

Description automatically generated***

Figure 11: Avg of bmi and Avg Glucose Level by Age Group

* + - 1. *Heart Disease, Hypertension and Stroke by Gender*

“Fig. 12” reflects that heart disease and hypertension are negatively correlated with each other - Males had 163 counts of heart disease, 222 counts of Hypertension and 108 counts of Stroke. While females had 113 counts of heart disease, 276 counts of hypertension, and 141 counts of stroke.

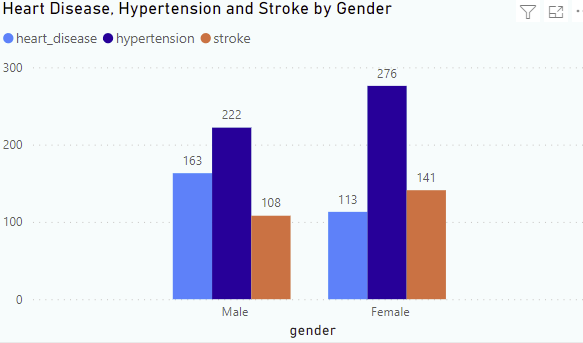


Figure 12: Stroke by Age Group

* + - 1. *Stroke by Ever Married*

“Fig. 13”: Stroke by Ever Married. Stroke risk also varies by Marital status; marriage is one of the most critical decisions in an adult person's life, as it interlaces with almost every aspect of life. It can be a source of happiness and security and also induce stress, anxiety, and heartache. There have been contrasting arguments on whether married people may be less likely to develop a stroke than individuals who aren't. But according to observations from the available data, stroke risk was higher for married or somewhat married, either still married, divorced, or separated. As the data suggest, 88.4% of the people account for stroke victims, and 11.6% for the never-married population.

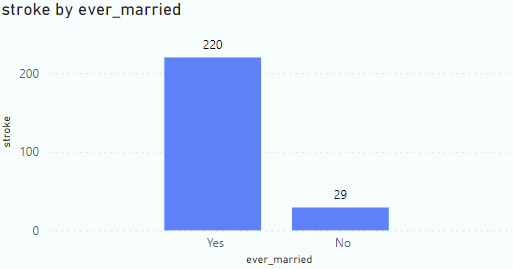


Figure 13: Stroke by Ever Married

# *Handling Unbalanced Data (SMOTE)*

When cleaning the dataset, we found that the dataset was imbalanced. If there is a serious imbalance in the data, the predictions are often also biased, i.e., the classification results will be biased towards the more observed classes. So, we chose to use the Synthetic Minority Oversampling Technique method to deal with the unbalanced data.

SMOTE is a modified scheme based on the random oversampling algorithm. Since random oversampling adopts the strategy of simply copying samples to add a few classes of samples, it is prone to the problem of model overfitting, even if the information learned by the model is too specific and not generalized enough. The basic idea of the SMOTE algorithm is to analyze the minority samples and add new samples to the dataset artificially based on the minority samples, so that the categories in the original data are not seriously imbalanced.

“Fig. 14” and “Fig. 15” show a clear comparison, where the data were smoothly transformed into balanced data by using the SMOTE method.

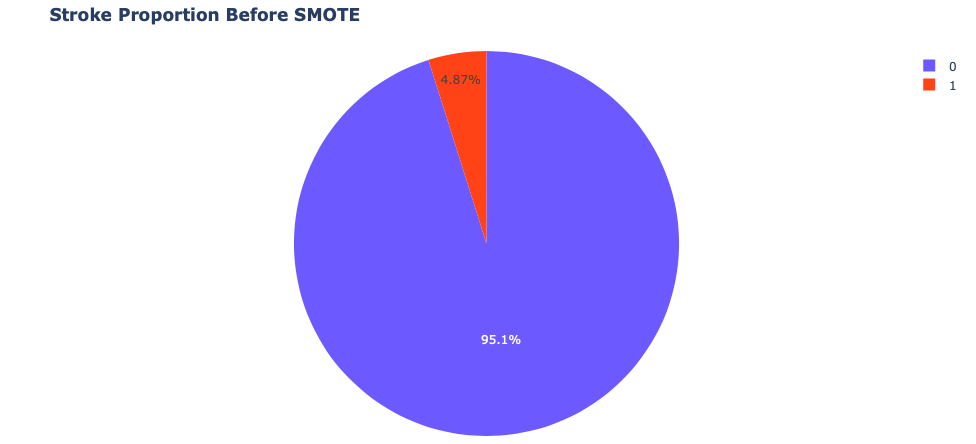
****

Fig. 14. Stroke Proportion Before SMOTE

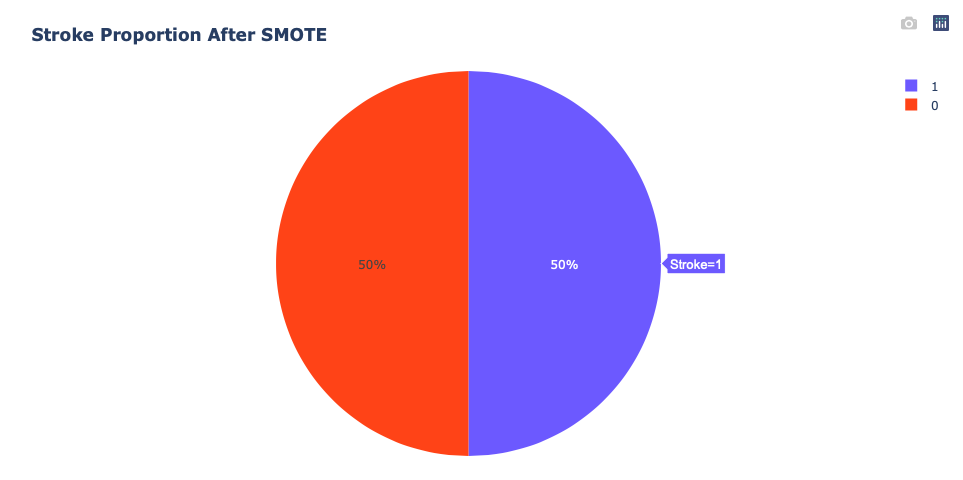
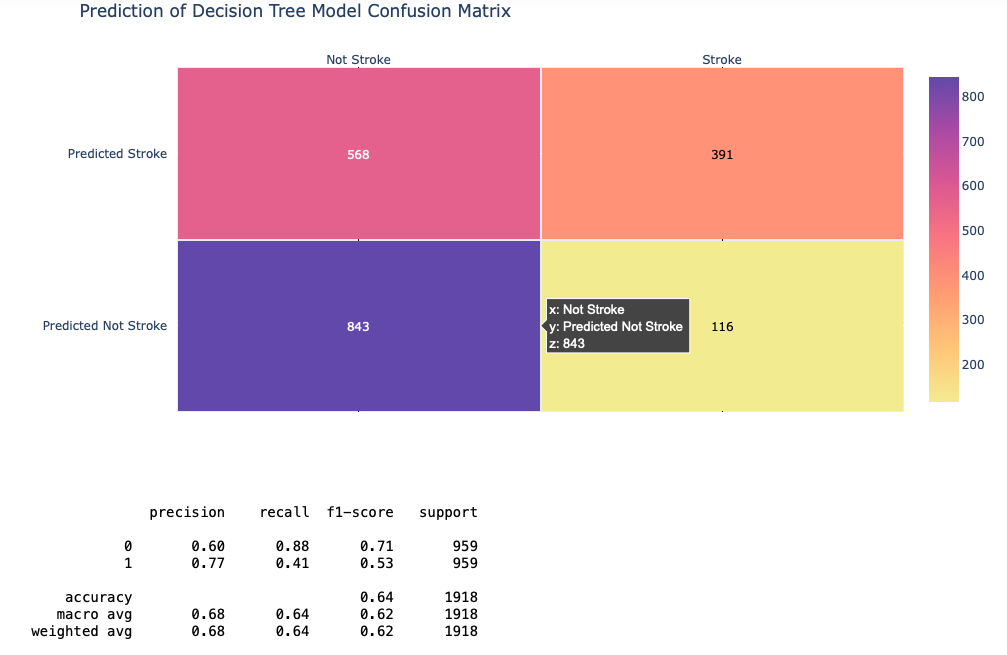
******

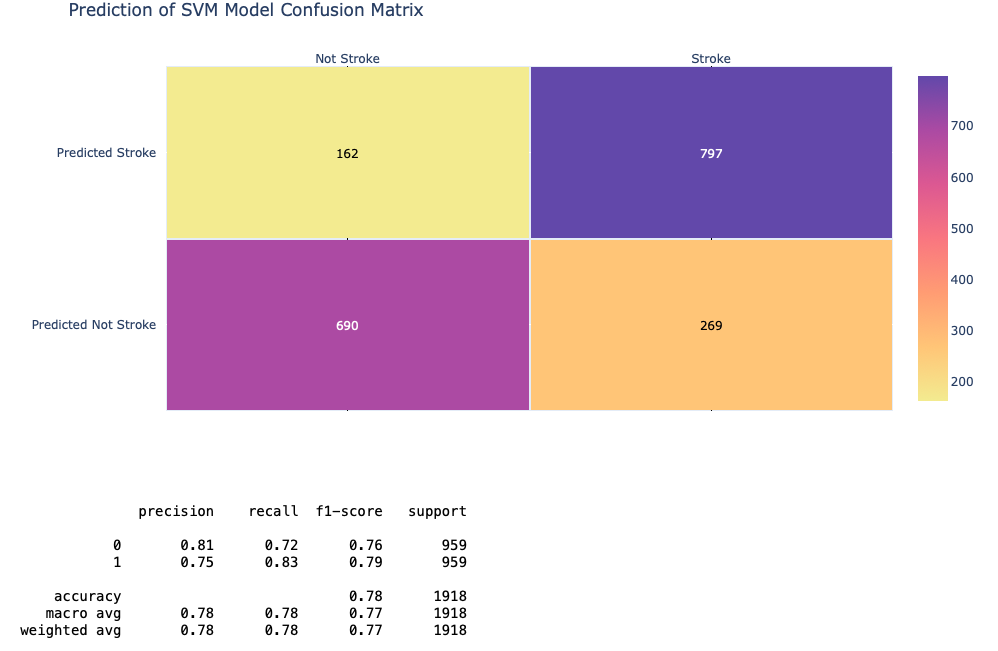
Figure 15: Stroke Proportion After SMOTE

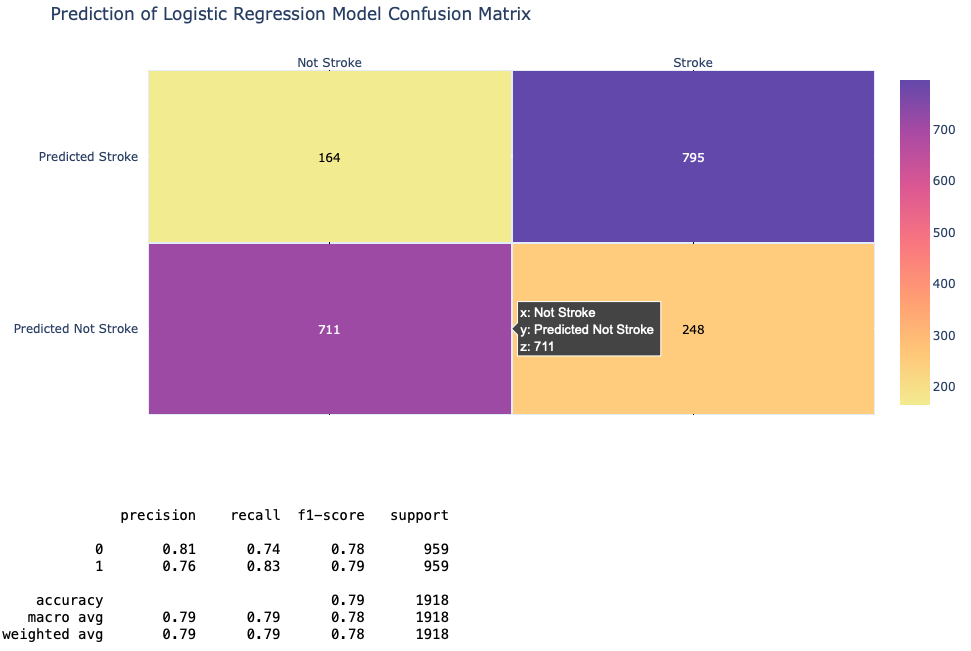
# *Detecting the Best Model with Confusion Matrix*

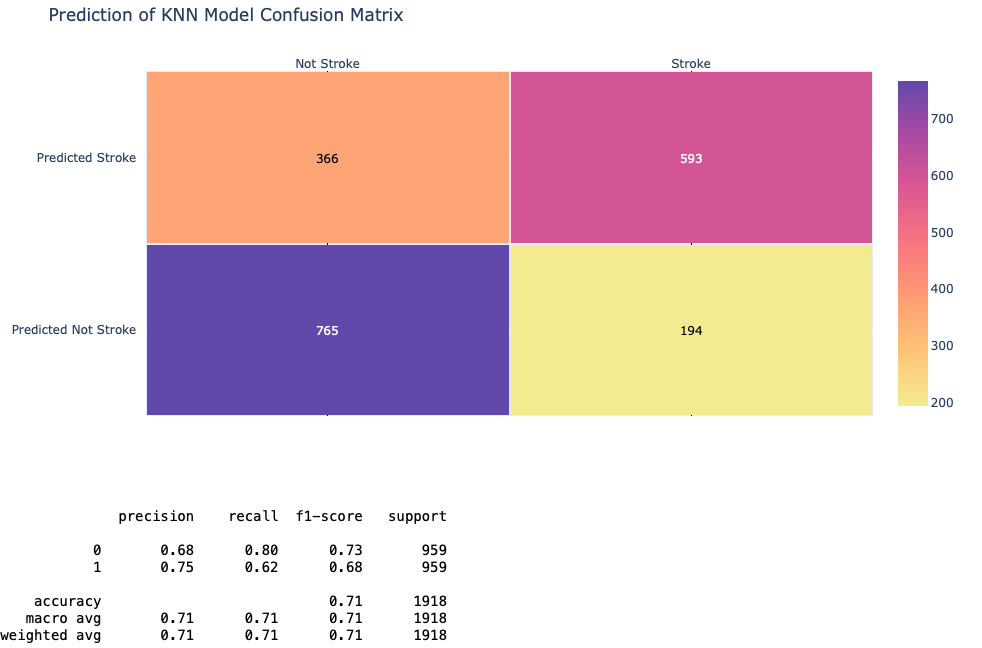
We first chose six different models to be used for prediction, they are ﻿ Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, KNN and Naïve Bayes. The best model in comparison is selected by using a confusion matrix approach. “Fig. 16” displays the combination of confusion matrices which includes the six different models. This approach allows us to further investigate which models are most suitable for prediction.











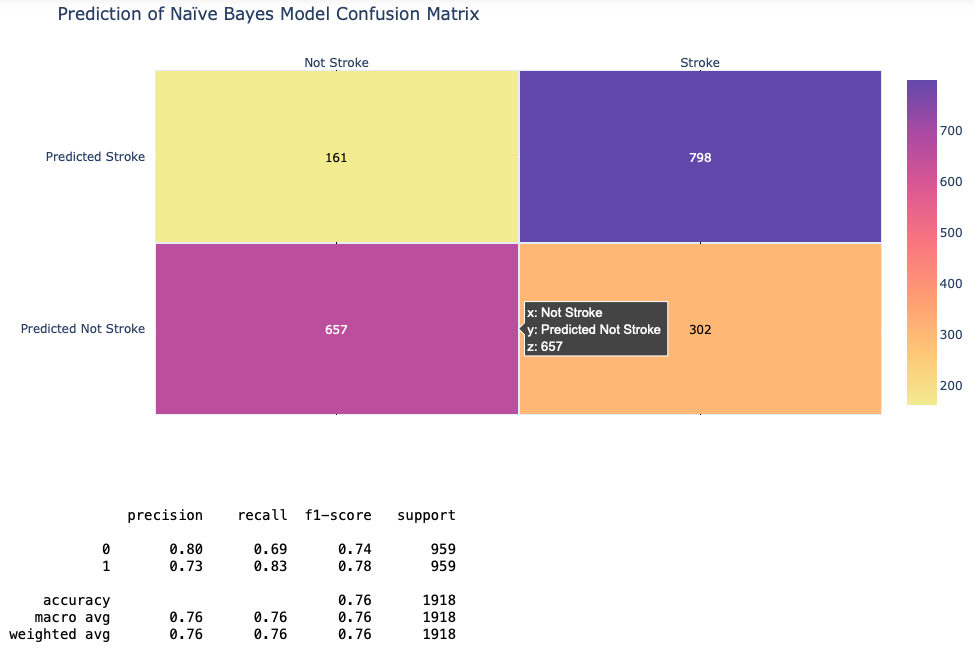


Fig. 16. Confusion Matrix of All Different Models

We further analyzed these models with the help of recall, precision, F1-score, accuracy score. Recall is the percentage of predicted pairs among the original pairs which is the correct positive predictions relative to total actual positives (the larger the value, the better, 1 is the ideal). The Precision score is the proportion of the predicted pairs that were originally correct (the larger the value, the better, 1 is ideal). Accuracy is the percentage of correct predictions. The F1-Score is an indicator that combines the output results of Precision and recall, and its value ranges from 0 to 1. 1 represents the best model output result, and 0 represents the worst model output result. To be able to analyze more finely, we chose F1 score and Accuracy as our main reference quantities, and we observed by the following “Fig. 17”, “Fig. 18” and “Fig. 19” that Support Vector Machine, Logistic Regression and Naïve Bayes methods are more appropriate for us.

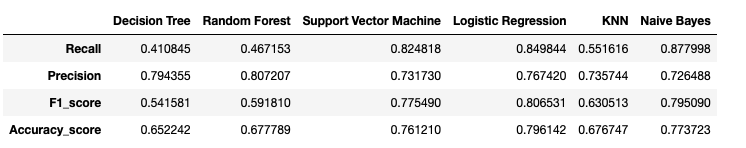


Fig. 17: Comparison of Different Methods

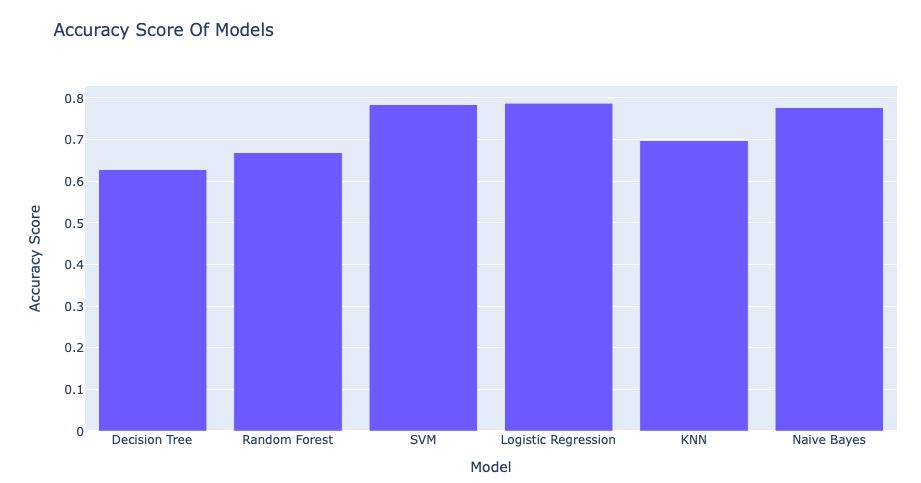
******

Fig. 18: Accuracy of Different Models

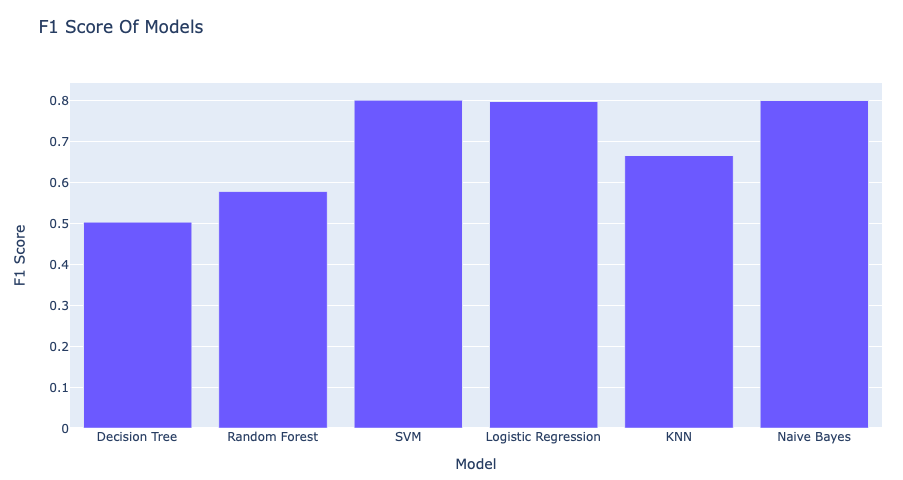


Fig. 19. F1 Score of Different Models

1. *Logistic Regression as the final predicting model*

After selecting the model, we used the best-performing Logistic Regression model to make predictions. Since we did not have any other dataset that we could use, we decided to take the last 100 rows of this dataset for prediction. As can be seen from “Fig. 20”, after cleaning the data and dealing with the unbalanced data set, the results we were able to predict became more readable and the accuracy rate improved significantly.

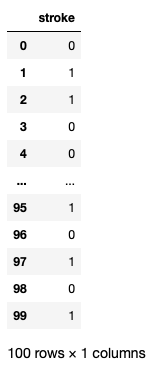


Figure 20: Predicted Result for Last 100 Rows

# v. Conclusion

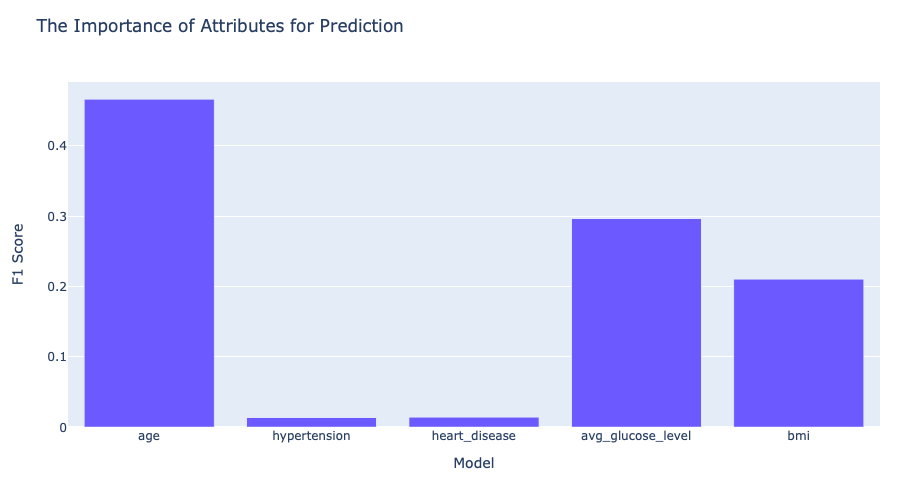
 Above’s analysis confirms that age has the greatest influence on whether one will suffer from stroke, followed by the risk of having pre-existing diseases and conditions. The accuracy of the stroke prediction models based on age and other diseases are high, and they have the heaviest weight when predicting stroke. It is demonstrated that, excluding the effect of age, people can reduce the probability of stroke by first preventing other prevalent conditions. We observed that age, BMI, and average glucose levels were the most important characteristics when predicting an individual’s prone to stroke, based on the current models (Fig. 21).

Fig. 21: The Importance of Attributes for Prediction

In reference to “Fig. 21”, we noted that the mean age is higher in women compared to men. There was a significant correlation between stroke and age. With an observed cut-off point of 40 years old, we notice that stroke risks increase as age increases. It is recommended that individuals, especially after the age of 40, should be mindful of stroke risk avoidance in two dimensions: by paying attention to diet, and to increase the frequency of regular medical checkups for proactive awareness. Individuals who already have heart disease, hypertension, or are chronic smokers, will have a higher chance of developing the stroke than the general population. Also, although the average glucose level and BMI of stroke patients are relatively concentrated in a certain range, we’ve recommended that those individuals in this range who did not have stroke should be more cautious from the fact that although their chances for stroke are much lower, they are increasingly more vulnerable to developing other conditions and diseases, which then leads to an increase the risk of stroke.

During our process of analysis, we’ve noted that there exist some bias issues with the dataset. First, the amount of data is not enough, we only have about 5000 data. Secondly, the original data contained serious imbalance, and it was necessary for us to balance the data to make better predictions. However, balancing it will cause distortion of the data. Lastly, we’ve noticed that there are not enough predictors, and it is in question the number of additional diseases that will cause stroke. Our team believes that if we obtained had a time series dataset, an additional variety of predictors, and a longer follow-up, we would build a better prediction model.