Final Project Research Summary

Lifestyle Predictors of Stroke

Data Mining - 6103

Akshat Saini, Xiao Qi, Keelin Arseneault

2022-05-06

Table of Contents

1 Introduction

A stroke occurs when blood flow to the brain is blocked, often leading to long term disability or death. The World Health Organization states that stroke is the second highest cause of death around the world, leading to 11% of total deaths.

Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. Between 2020 and 2021, in the United States, stroke is the number two cause of death right after deaths due to the virus Covid-19. It is an important topic to study, since healthcare systems already under stress can gain insight from data and various models that show patients' likelihood of experiencing a stroke. This allows doctors to make preventative recommendations and hopefully save lives, as well as time and money for healthcare facilities.

1.1 Dataset Source

The study on lifestyle predictors of stroke was conducted using a free dataset which was sourced from Kaggle. The raw data originally contained 5,110 observations and 12 variables. The dataset was initially quite imbalanced, with very few participants that had experienced a stroke.

1.2 SMART Questions

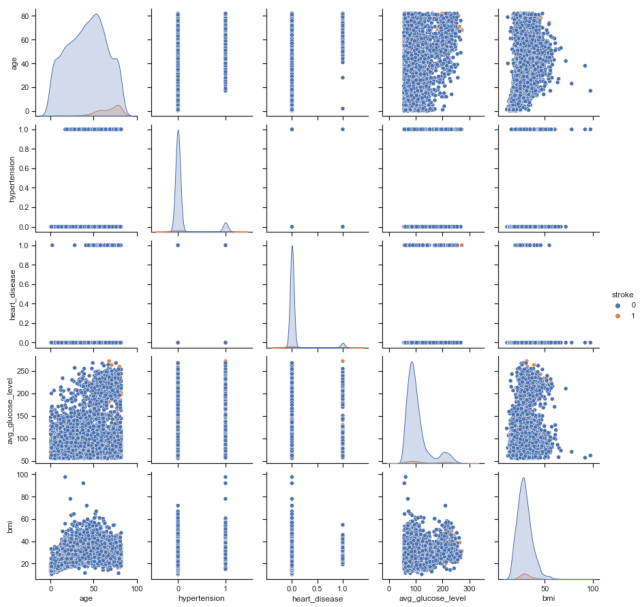
The relevancy of our topic, one of the SMART question requirements, is particularly timely during the current month of May - stroke awareness month. The main focus of

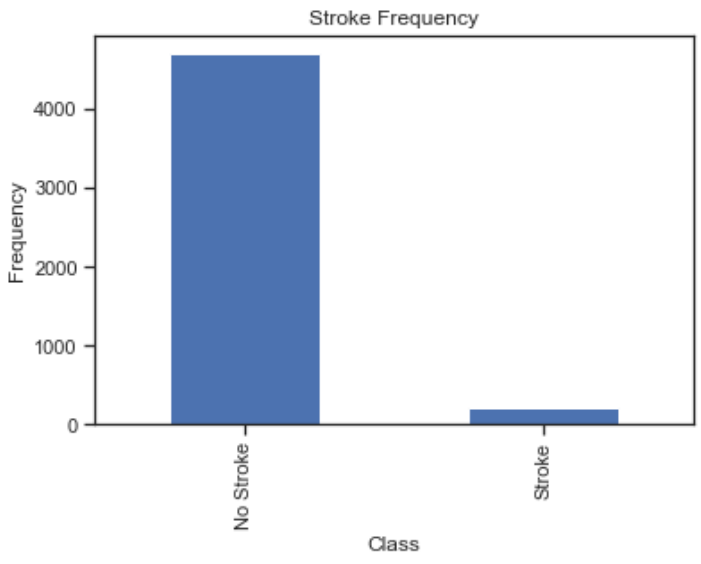
this study was to explore the below set of SMART questions, related to a mixture of both EDA and modeling processes:

- What factors or variables affect the likelihood of a person having a stroke?
- Can we predict if someone will have a stroke based on their health and lifestyle?
- What are the relationships between having a stroke and quantitative variables in the dataset (BMI, age, glucose level, etc.)?
- Is a particular gender affected more from heart disease or hypertension?
- Do marriage status and residence type contribute to having a stroke?

2 Preparation, Data Tidying, and Initial Analysis

To begin the data preparation process, we chose to drop the Patient ID variable, leaving 11 variables to explore further: *Gender, Age, Hypertension, Heart Disease, Marital Status, Work Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke*. Next, we removed rows containing NaN data, or missing values. This eliminated 201 observations from the original 5,110 rows in the dataset, leaving 4,699 patients who did not have a stroke and 209 who did have a stroke. As previously mentioned, this data is very unbalanced, even when split up by gender.





Pair plots above confirm imbalance in the dataset represented in each variable, while the histogram on the right shows this overall disproportion between the number of participants that had a stroke vs. participants that did not have one. Exploratory Data Analysis (EDA) was performed on the original unbalanced data, but we acknowledged the need to address this issue in the upcoming modeling process.

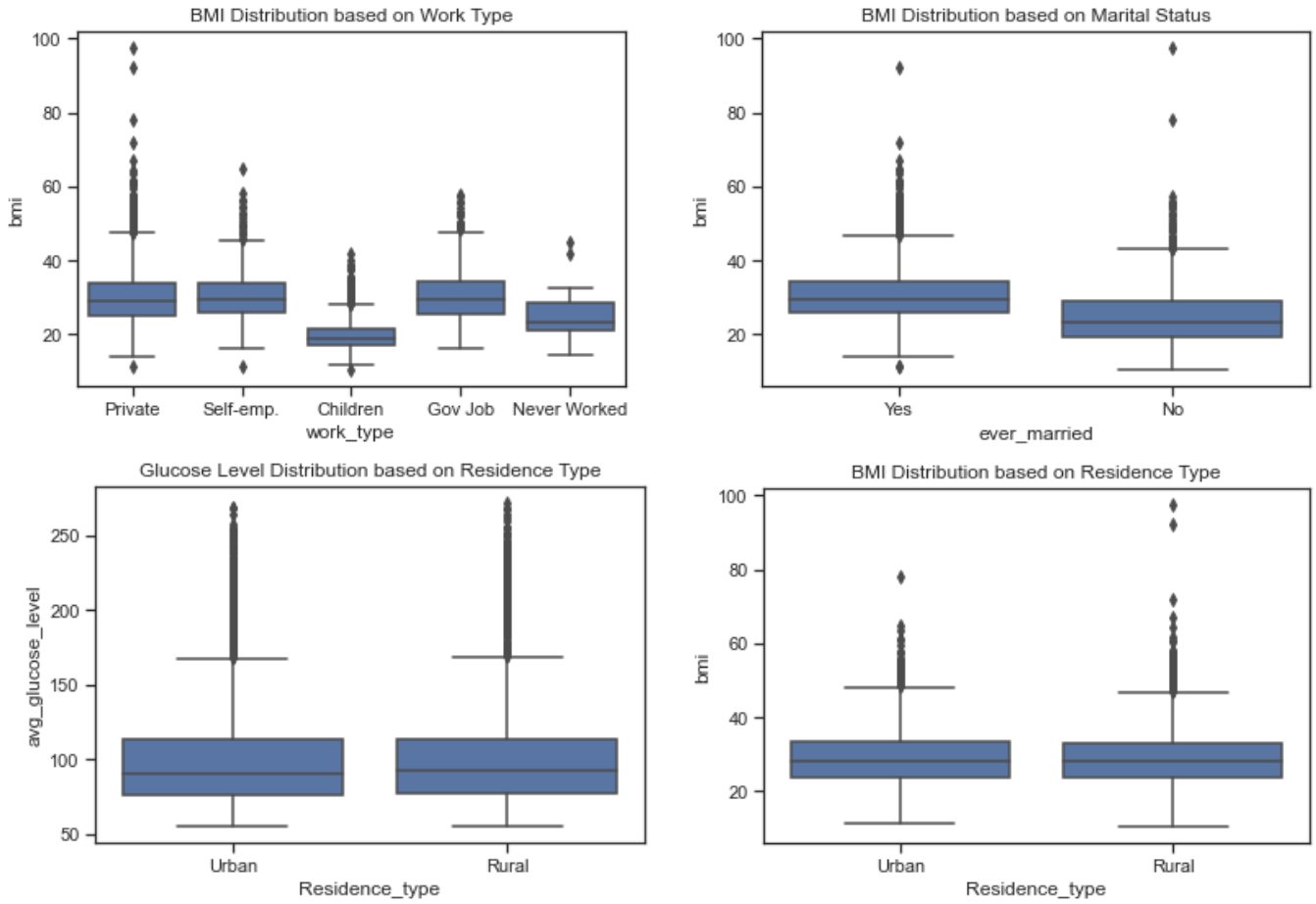
Our first EDA-related SMART question aimed to find relationships between having a stroke and quantitative variables in the dataset (BMI, age, glucose level, etc). This question was first explored with a statistical testing approach called Welch's t-test. Welch's t-test is used to test whether the mean value of a given variable is significantly different between two groups - specifically, two groups that have unequal variance and imbalanced sample sizes. This applies strongly to our dataset, since we have more non-stroke data, so we used this t-test to look for any statistically significant difference between the mean BMI, age, and glucose values of stroke vs. non stroke groups, in order to draw some initial conclusions about the relevance of these variables in whether someone has a stroke or not. For all three variables, as shown in the table below, the p-value was below 0.05, indicating there is a significant difference in the mean values of BMI, age, and glucose for stroke vs. non-stroke observations.

Welch's T-Test		
Variable	T-Statistic	P-Value
BMI	3.6374	0.0003
Age	28.2864	<0.0001
Glucose	7.0034	<0.0001

These strong differences between stroke and non-stroke groups were promising to

see for our goal of building relevant classification models that can make predictions pertaining to stroke factors.

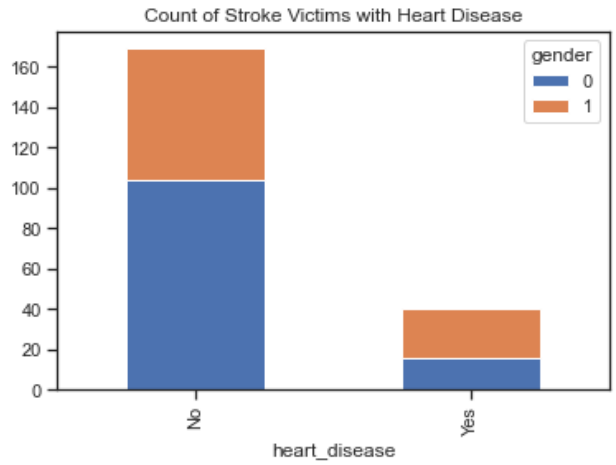
Boxplots, shown below, also revealed some interesting potential relationships between these quantitative variables that are significant to stroke status and the categorical variables in the data.



The distributions of the data show that stay-at-home parents appear to have lower BMI than patients with other types of work (perhaps from chasing the kids around?), while marriage is shown to perhaps be related to higher BMI. On the other hand, residence type did not seem to show major differences in BMI or glucose level distribution - we thought that living in an urban versus rural environment may impact nutrition access and diet choices, but there is not a strong indication of difference in residence type impacting these quantitative stroke indicators. Though no solid

conclusions can be drawn from these visualizations alone, they contributed to our initial thoughts about what the modeling process could look like and the importance of some variables versus others.

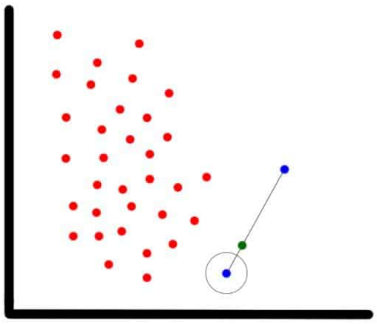
Another EDA SMART question was whether a particular gender may be affected more from heart disease or hypertension. Of the participants who had a stroke, we were wondering if there is any indication that more of those participants suffered from a comorbidity like heart disease or hypertension (high blood pressure), and if there was more representation of this occurring in one gender over the other. Interestingly there was pretty even balance between genders for no hypertension or heart disease. But for those positive for hypertension or heart disease, more women (represented with a blue bar in the visuals below) who had stroke had hypertension while more men (represented with orange) who had stroke had heart disease. This turned out to be just an additional preliminary observation that may also be of interest in our models when asking if hypertension or heart disease can contribute to predicting stroke.



Overall, EDA gave us an overall understanding of the dataset, how the imbalance of the data looks in reality, and how the health and lifestyle variables associated with stroke might also be related to each other. After balancing the dataset, we were able to proceed with modeling both unbalanced and balanced data using various methods to compare the results and see if the variables in our dataset can successfully contribute to understanding stroke factors.

3 Balancing the Dataset

The image shown to the right is usually how an imbalanced dataset appears in a Scatter Plot. As observed in the above plots and graphs, our dataset is very similar to this image and a model generated using the set would lead to a bias, or majority of the values being classified as 0 or 'no_stroke'.



To tackle the problem of an imbalanced set, we can use many techniques such as undersampling or oversampling, etc. However, general oversampling of the data set in many cases leads to creation of duplicate values which the team was trying to avoid.

We will balance the dataset using a minority oversampling technique called SMOTE: Synthetic Minority Oversampling Technique. This technique uses KNN or K-nearest neighbors algorithm to generate synthetic or artificial values of the under-represented category in the target variable. The values that are generated using SMOTE are synthetic and are close to the actual values but not duplicates.

For more information: https://github.com/scikit-learn-contrib/imbalanced-learn

4 Model Building

4.1 logistic regression with different cutoff value

We want to use the data to predict the stroke in this part. So people can take some actions like a more detailed physical examination before the conditions worsen.

We know that the data is unbalanced, so we built a logistic regression model with a different cutoff value.

4.1.1 Model Training

First, let us look at how unbalanced the dataset is. We find that 95.7% of the data are healthy instances, and only 4.3% are people with stroke.

Stroke	0	1
count	4699	209
ratio	0.957	0.043

To train and evaluate the model, we split the dataset into two parts. 80% of the dataset will be used to train the model, and the rest will be used to test the model's accuracy.
The first model is below.

Generalized Linear Model Regression Results

Dep. Variable:	stroke	No. Observations:	3926
Model:	GLM	Df Residuals:	3910
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-547.92
Date:	Tue, 26 Apr 2022	Deviance:	1095.8
Time:	14:14:22	Pearson chi2:	3.19e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.8080	0.688	-11.346	0.000	-9.157	-6.459
C(gender)[T.1]	0.0810	0.173	0.469	0.639	-0.257	0.419
C(hypertension)[T.1]	0.4558	0.198	2.298	0.022	0.067	0.844
C(heart_disease)[T.1]	0.4260	0.230	1.850	0.064	-0.025	0.877
C(ever_married)[T.1]	-0.0290	0.279	-0.104	0.917	-0.575	0.517
C(work_type)[T.1]	-0.3785	0.201	-1.881	0.060	-0.773	0.016
C(work_type)[T.2]	0.6108	1.122	0.544	0.586	-1.588	2.810
C(work_type)[T.3]	-0.4806	0.275	-1.746	0.081	-1.020	0.059
C(work_type)[T.4]	-16.3674	1.16e+04	-0.001	0.999	-2.28e+04	2.28e+04
C(Residence_type)[T.1]	0.0025	0.168	0.015	0.988	-0.327	0.332
C(smoking_status)[T.1]	-0.1543	0.256	-0.604	0.546	-0.655	0.346
C(smoking_status)[T.2]	0.0257	0.215	0.119	0.905	-0.397	0.448
C(smoking_status)[T.3]	0.4811	0.235	2.048	0.041	0.021	0.941
age	0.0716	0.007	10.233	0.000	0.058	0.085
avg_glucose_level	0.0042	0.001	2.852	0.004	0.001	0.007
bmi	0.0023	0.013	0.176	0.860	-0.024	0.028

We can find that the p-values of some variables are larger than 0.05, which means that these variables are not significant, so we just dropped these variables and built the second model.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 stroke   No. Observations:                 3926
Model:                            GLM   Df Residuals:                     3921
Model Family:                Binomial   Df Model:                            4
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -554.12
Date:                Tue, 26 Apr 2022   Deviance:                       1108.2
Time:                        14:17:10   Pearson chi2:                 3.25e+03
No. Iterations:                     8
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -7.4871      0.424    -17.656      0.000      -8.318      -6.656
C(hypertension)[T.1]             0.4671      0.196      2.386      0.017       0.083       0.851
C(heart_disease)[T.1]            0.5181      0.225      2.299      0.022       0.076       0.960
age                              0.0659      0.006     10.801      0.000       0.054       0.078
avg_glucose_level                0.0043      0.001      3.049      0.002       0.002       0.007
================================================================================================
```

The second model is reliable. And the Variance Inflation Factors (VIF) for all variables are all smaller than 10. which means there is no high correlation between variables.

variables	VIF
hypertension	1.10
heart_disease	1.08
avg_glucose_level	1.10
age	1.18
Intercept	8.80

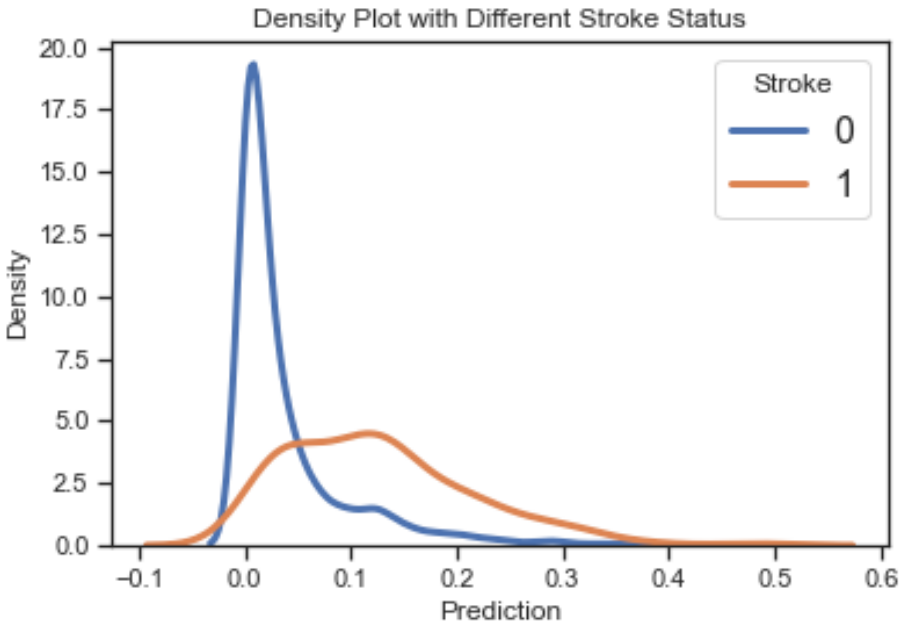
And here is the formula for the logistic regression model:

```
glm(formula='stroke ~ age + C(hypertension) + C(heart_disease) +
avg_glucose_level', data=df_unbalanced_train,
family=sm.families.Binomial())
```

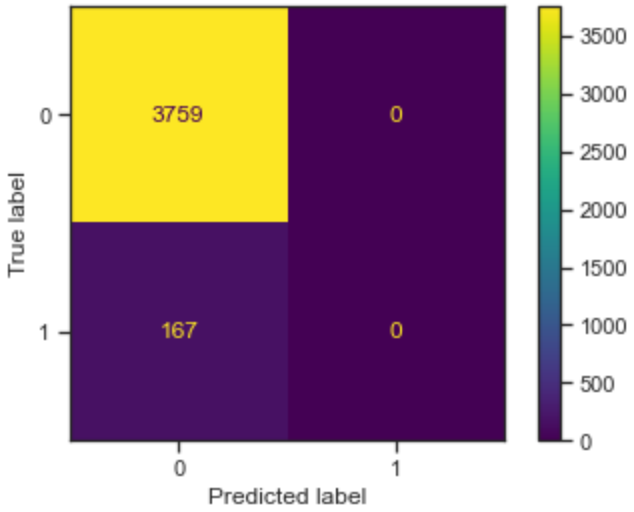
4.1.2 Predicting and Assessing the Model

Let's have a quick look at our prediction first.

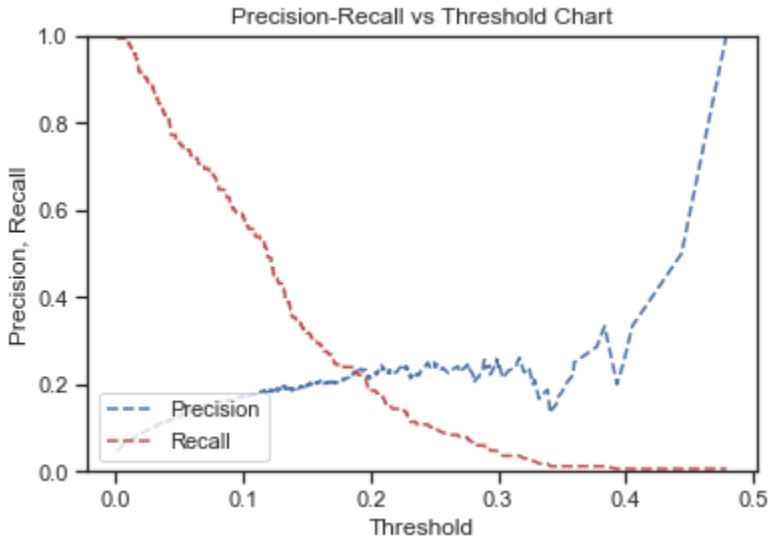
The graph here is the predicted value for healthy people and people having a stroke. The orange line is the people who have a stroke, and the blue line is the healthy people. The probabilities of people with stroke and healthy people are almost lower than 0.5. It is because most of the data are zero (healthy) instances.



Then, we can have a look using the visualization of the confusion matrix.
We can find from this picture that when the cutoff is 0.5, there is no False Positive (FP). So the precision will be 1. But the recall now is 0 because there is no True Positive (TP).
Because the number of stroke instances is low, our model will likely make a false negative mistake. Therefore, 0.5 is not a suitable cutoff value.



Then we can look at the model's precision and recall with different cutoff values. When we add the cutoff value, we increase the precision. But we decrease recall.

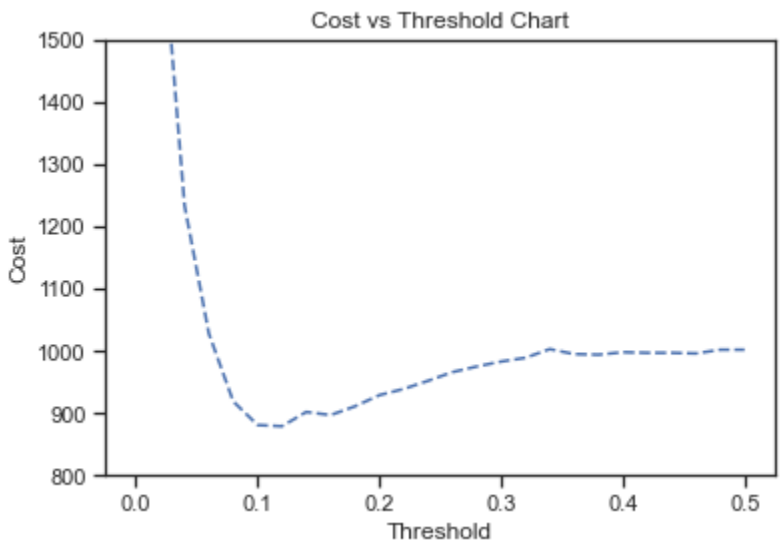


4.1.3 Choosing the Suitable Cutoff Value

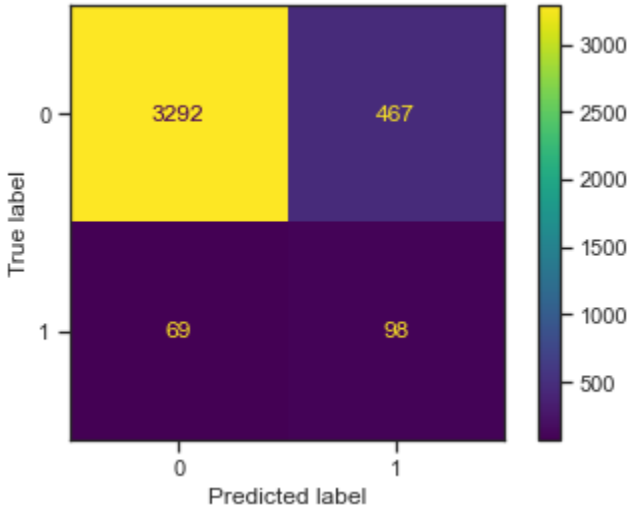
So what is a suitable cutoff value?

When we have an FN (a stroke instance but not found), the wrong prediction may injure their life. But for an FP (a healthy person but predicted stroke), they just need a more detailed physical examination. So an FN is more costly than an FP for our project. So set cost_FN = 6, cost_FP = 1. We can get a graph of the total cost in different cutoff values.

And apparently, the suitable cutoff value is 0.1.



We can re-plot the confusion matrix using the train dataset to see what happened when we switched to the new cutoff value.



Our classification model makes fewer false negative (FN) errors since its cost is six times higher than a false positive (FP).

Then we can use test data to evaluate the model. For example, here is the classification report by test dataset.

Logistic regression classification report using test dataset

	precision	recall	f1-score	support
0	0.98	0.88	0.93	940
1	0.18	0.57	0.27	42
accuracy			0.87	982
macro avg	0.58	0.73	0.60	982
weighted avg	0.94	0.87	0.90	982

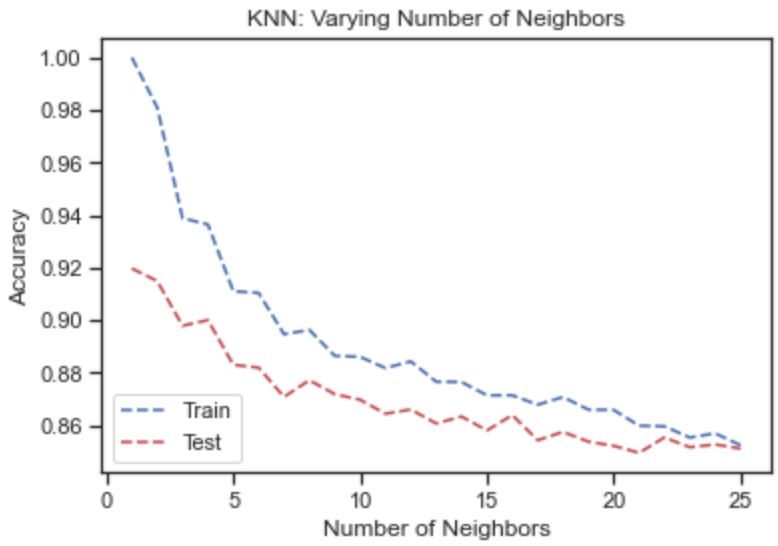
We can find from the classification report above that an unbalanced dataset doesn't have a high f-score in the model. And balancing should be the first step when handling the dataset to make the model more dependable.

4.2 KNN

Then, we built a KNN model. Unlike the previous model, this time, we used a balanced dataset by SMOTE, split the dataset into train and test, and compared the accuracy of the training dataset and test dataset to find out the best K for this model.

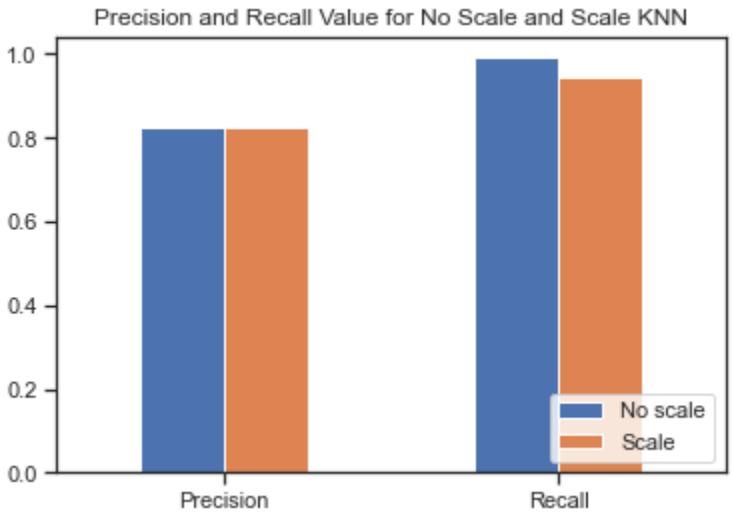
4.2.1 Choosing the suitable K

The first step using the KNN model is to choose a suitable K.
We were considering that the k should be an odd number. And a small k may cause overfitting, and a large k will cause underfitting. According to the graph about numbers of neighbors and the model's accuracy, we choose seven as the K.



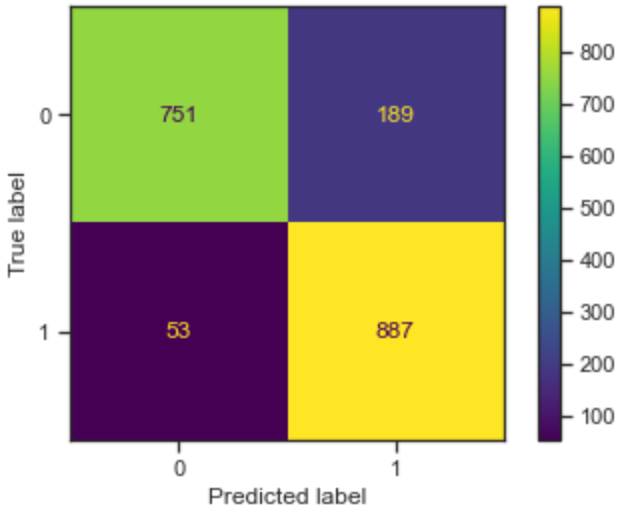
4.2.2 Scaled and unscaled KNN

Then we compared the KNN model with scaled data and without scaled data. We thought that the scaled dataset would have higher precision and recall value. But it seems that unscaled datasets perform a little better.



4.2.3 Evaluating the model

There is the confusion matrix using a scaled test dataset. Again, the precision is 0.82, and the recall value is 0.94.



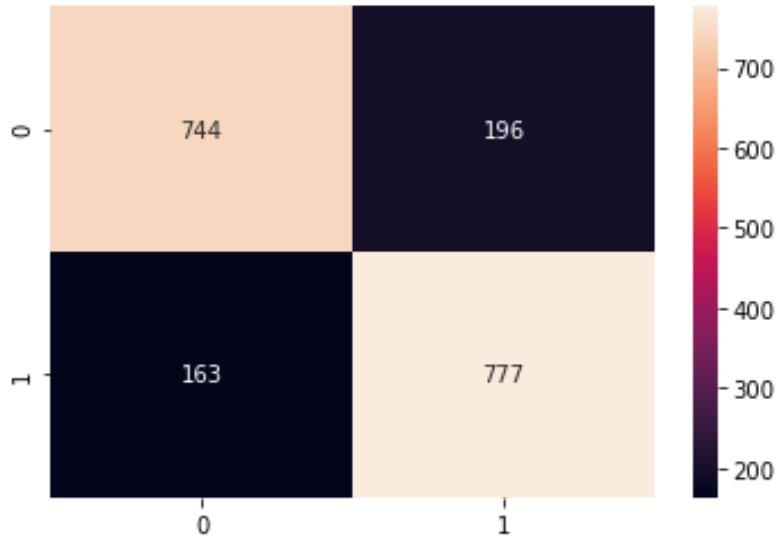
And the classification report for the dataset is as follows.

KNN classification report using test dataset

	precision	recall	f1-score	support
0	0.93	0.80	0.86	940
1	0.82	0.94	0.88	940
accuracy			0.87	1880
macro avg	0.88	0.87	0.87	1880
weighted avg	0.88	0.87	0.87	1880

4.3 Logit on Balanced Set

Here, we performed simple logistic regression on the unscaled data set balanced using the SMOTE technique. Using the Train/Test split package available from sklearn, we were able to split 80% of the data set into a training set and the rest 20% into a test set.



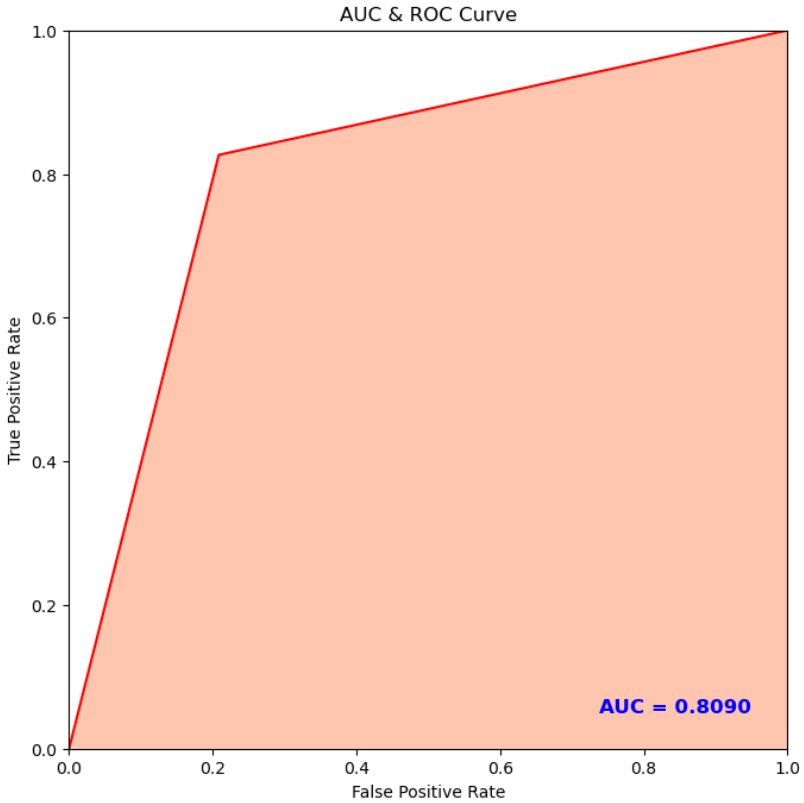
- y axis - truth values
- x axis - predicted values
- 744 participants were marked as 0 or as not having a stroke and 744 times participants were predicted as not having a stroke by the model.
- 163 participants were 1 or at risk of stroke but 163 times participants were predicted as not at risk of stroke or 0 by the model.
- 196 participants were 0 or not at risk of stroke but 196 times participants were predicted as at risk of stroke or 1 by the model.
- 777 participants were 1 or at risk of a stroke and 777 times participants were predicted as at risk of a stroke or 1 by the model.

Now, generating the classification report for the above model to get more information about the Accuracy, precision and recall:

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.83	0.80	0.82	940
1	0.81	0.84	0.82	940
ACCURACY			0.82	1880
MACRO AVG.	0.82	0.82	0.82	1880
WEIGHTED AVG.	0.82	0.82	0.82	1880

4.3.1 Generating the ROC



Area under the curve: 0.8090

A general rule to follow suggests that AUC of the ROC should be 0.80 or above for a good or reliable model. Since 0.8090 is still close to the border line, we will perform

feature selection on the data set to select the variables with the highest variance on the set and then compare the accuracy of the model again!

4.4 Feature Selection

The purpose of feature selection is the process of analyzing a subset of a dataset or identifying key elements or series of data to use in a model. Now, we want to perform feature selection on the dataset to select ideal variables that cause a variation in the target variable.

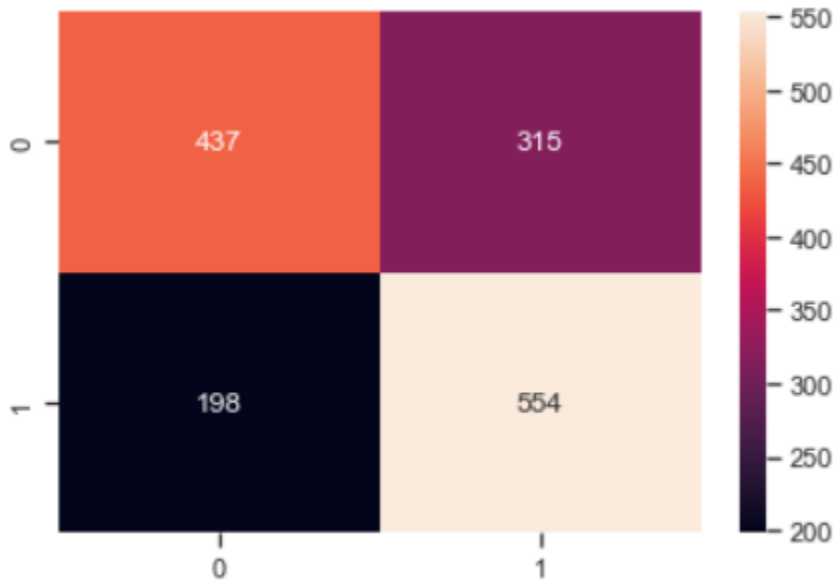
After conducting Feature Selection using the sklearn library exhaustive search model, we have surprisingly found that variables like *Residence type, work type, married status, heart disease, hypertension*, and *gender* have more variance on the target variable than other variables like BMI, avg. glucose level or even smoking.

To make sure that the above selected variables are not related to each other we will perform the test for multicollinearity by evaluating VIF values.

	variables	VIF
0	hypertension	1.104266
1	heart_disease	1.066529
2	ever_married	1.644542
3	work_type	1.293665
4	Residence_type	1.456065
5	gender	1.375509

Since all the variables had a vif value in the range of 1 to 2, the model has no problem of multicollinearity.

4.4.1 Logit modeling on selected features



- y axis - truth values
- x axis - predicted values
- 437 participants were marked as 0 or as not having a stroke and 437 times participants were predicted as not having a stroke by the model.
- 315 participants were 1 or at risk of a stroke but 315 times participants were predicted as not at risk of stroke or 0 by the model.
- 198 participants were 0 or not at risk of stroke but 198 times participants were predicted as at risk of stroke or 1 by the model.
- 554 participants were 1 or at risk of a stroke and 554 times participants were predicted as at risk of a stroke or 1 by the model.

The Feature selected Logit Model led to an AUC of 0.659. This is much below the 0.80 border. This is usual behavior when feature selection or PCA is performed on a data set. As we reduce the size of the data that is available to us, the accuracy of

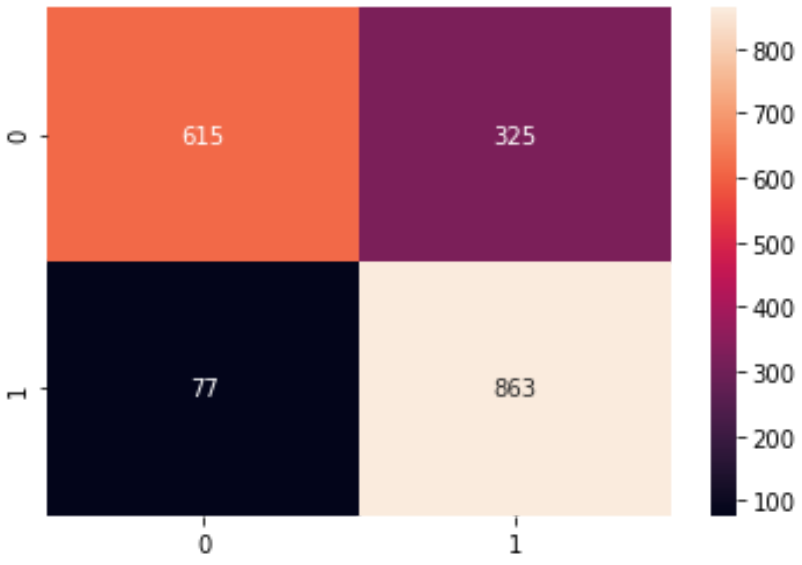
the model tends to decrease. These options tend to be favorable when the data set is extremely huge.

4.5 Decision Tree - Classification Model

Next, the team decided to build a classification model using a decision tree. As the target variable has a binary outcome, a classification tree model should be ideal in classifying whether a participant has any probability of getting a stroke or not.

This model was performed on the balanced data set. So, the model nearly had an equal number of participants that were at risk for a stroke and participants that were not at risk of a stroke.

Using the train-test split method from the sklearn library, we split 80% of the data set into a training set and 20% into a test set.



- y axis - truth values
- x axis - predicted values
- 615 participants were marked as 0 or as not having a stroke and 615 times participants were predicted as not having a stroke by the model.
- 325 participants were 1 or at risk of a stroke but 325 times participants were predicted as not at risk of stroke or 0 by the model.
- 77 participants were 0 or not at risk of stroke but 77 times participants were predicted as at risk of stroke or 1 by the model.
- 863 participants were 1 or at risk of a stroke and 863 times participants were predicted as at risk of a stroke or 1 by the model.

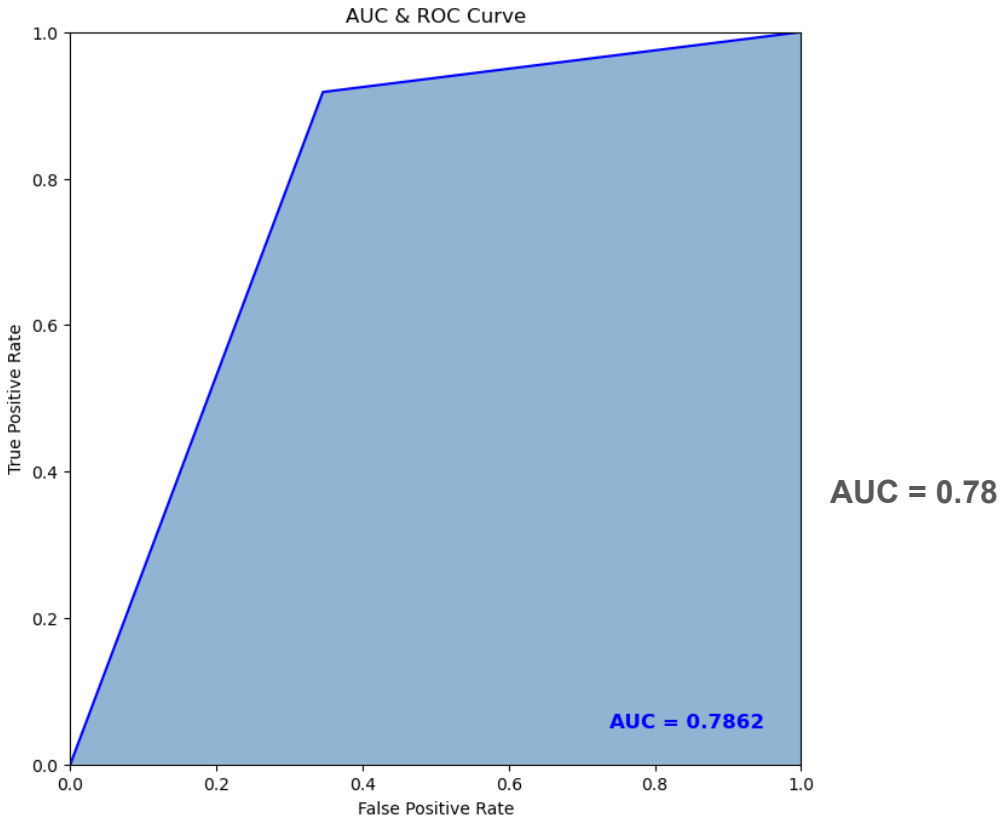
Now, generating the classification report to have a better look at the accuracy, recall and precision scores of the model.

CLASSIFICATION REPORT:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.91	0.65	0.76	940
1	0.73	0.93	0.82	940
ACCURACY			0.79	1880
MACRO AVG.	0.82	0.79	0.79	1880
WEIGHTED AVG.	0.82	0.79	0.79	1880

We can see a decent Recall value for 1 or ‘risk of stroke’ using the classification tree model.

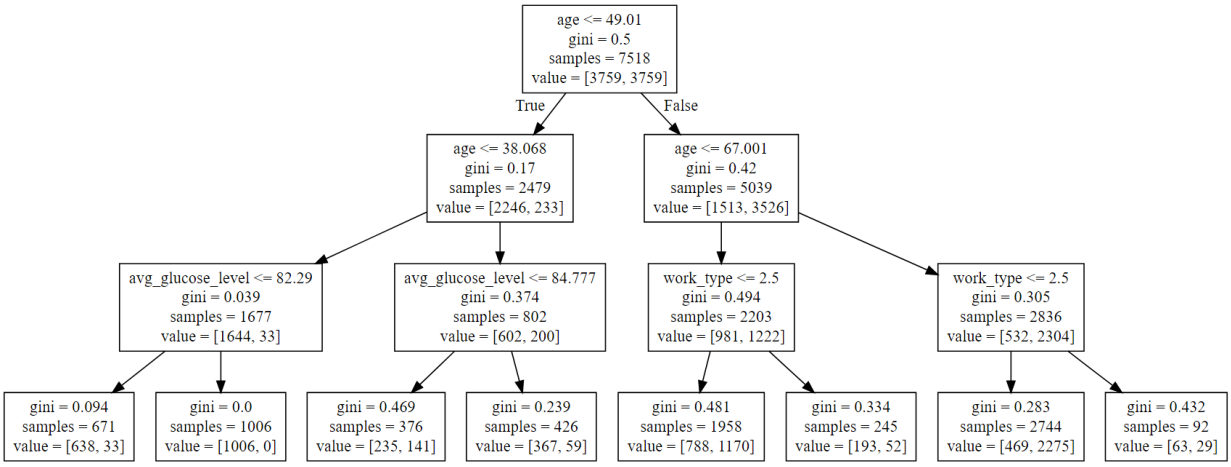
4.5.1 Generating the ROC



4.5.2 Generating the Tree Structure

The Tree structure was further generated using the available graphviz extension. This allowed us to see the division between the branches in a more clear and concrete manner.

As can be observed below, the max_depth of the tree was set to 3. Increasing the max_depth to a greater number would lead to a higher accuracy score, but that is usual behavior given the property of overfitting as every category type would essentially be its own node in the tree.



5 Model Comparison and Summary

MODEL	ACCURACY	RECALL	AUC
STAT MODEL LOGIT (Imbalanced)	0.18	0.57	0.73
KNN (Balanced, Scaled, n = 7)	0.82	0.94	0.87
KNN (Balanced, Not-Scaled, n = 7	0.83	0.99	0.89
LOGIT - 1 (Balanced)	0.81	0.83	0.81
LOGIT - 2 (Balanced + Feature Selected)	0.66	0.72	0.66
CLASSIFICATION TREE (Balanced)	0.79	0.93	0.79

In the previous part, we built six models using four different methods. And here, we listed the accuracy, recall, and AUC parameters from our models.

Without balancing the dataset, we can find that the logistic regression model performed not well compared with other models. To get a high recall value, we have to sacrifice the model's accuracy. So balancing the dataset is important when we build the model.

All the five models using a balanced dataset perform well in the test dataset. However, comparing the KNN with and without a scaled dataset, unlike the usual knowledge, the KNN model using an unscaled dataset has a better performance. This could be due to the fact that the data was already presented as values 0 or 1.

Also, the logistic regression model without feature selection has a better performance.

Among these, KNN and classification tree perform much better than the other four models and have a higher recall value for 1 or ‘at risk of stroke’, which means if we want to predict the stroke by our models, KNN and classification tree would be a better choice.

6 Conclusions of Study

- The first conclusion that came about from this particular dataset was that balancing data is important in order to build a decent model, in order to give equal priority to each class.

- We also conclude that EDA showed significant difference between stroke and non-stroke group averages of BMI, glucose levels, and other quantitative variables, but these were not considered to be as important when we used feature selection, which showed categorical variables like marital status, residence type and work type to be more significant when modeling stroke. This could be directly related to the stress levels associated with those variables.

- We also noticed there was a pretty even balance between stroke victims of both genders who had hypertension and heart disease. But for those positive for hypertension or heart disease, more women who had stroke had hypertension while more men who had stroke had heart disease. The modeling process later confirmed that *hypertension* and heart *disease* are indeed *important* stroke factors.

- Finally, after modeling with both balanced and unbalanced datasets, we can draw conclusions about the importance of balancing the data as well as the quality of our models.
    - The logistic model relying on the cut off method to address the unbalanced data did not perform as well as most of the balanced data models in general.
    - With N set to 7, the KNN model performed best with an average accuracy of 0.83 and area under the curve (AUC) value of 0.89, which exceeds the widely accepted value of 0.80 to determine a good model.

- Though we can't focus strictly on the accuracy of the models to determine their usefulness, their high accuracy and recall rate is a good indicator that stroke is related to certain factors and patients can prioritize taking care in certain areas of their life to lessen the effects of stress and other contributors to a stroke occurrence. Healthcare applications are usually of high relevance for modeling data, because even though nothing can be predicted for certain, even having a slightly improved understanding of health/lifestyle factors that relate to a potentially deadly outcome can save a lot of people's lives and hospital resources.