# **Project Report**

Stroke Prediction

By: Jacob Crawford, Jaimi Patel, Seif Abuhashish

**Executive Summary: (Goal, method, conclusion)**

This report presents the findings of a stroke prediction developed as part of the DSC478 final project. The dataset <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset> is utilized to foresee whether a patient is going to suffer a stroke given the information like gender, age, various diseases, and smoking status, etc. Each column in the information gives relevant data about the patient.The goal of our classification algorithms is to analyze the stroke data and predict the possibility of having a stroke or not.

Before building our models we explore and investigate the dataset for a deeper understanding. Histogram distribution with stroke incidence charts are used to compare the inter categories of each independent variable. The probability of stroke in each category of the independent variables helped us better understand the impact of certain variables on risk. Variables ‘heart\_disease’ and ‘hypertension’ stood out because they respectively increase the risk of stroke by 4 times (heart\_disease) and 3 times (hypertension).

The classification methods in the study involved Decision Tree, Logistic Regression and KNN and compared their accuracy to see which model works better with the data also helped differentiate which factors were most capable of categorizing patients having stroke or not . Initially, we describe the data to know all the properties and data types which can help to differentiate categorical and numerical variables. We replaced outlier and missing data by replacing it with mean values and visualizing the categorical and numerical variables to see their distribution.

**Dataset Overview:**

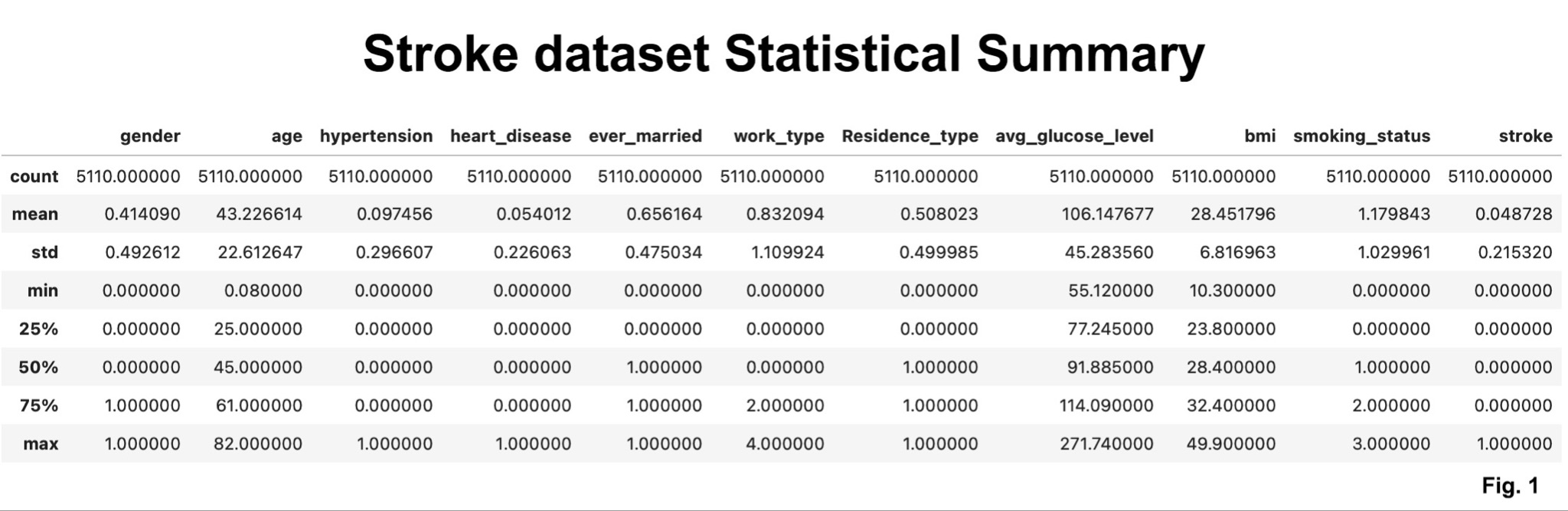
Strokes are the second leading cause of death and the third leading cause of disability in the world. According to The World Stroke Organization (WSO), about 14 million people a year suffer from strokes and about 5.5 million die. (“What We Do | World Stroke Organization”) As scary as this seems, there is hope according to the CDC (Centers for Disease Control) because a stroke can sometimes be dependent on avoidable risk factors. These risk factors can be prevented and minimized with proper health/nutritional education, healthcare monitoring, regular physician follow ups, and awareness of the risk factors. (George et al.)

This dataset is utilized to foresee whether a patient is going to suffer a stroke given information like gender, age, various diseases, and smoking status. Each column in the information gives relevant data about the patient. The idea is to analyze the stroke data and predict the possibility of having a stroke. Our data set contains 12 attributes and 5110 data points.

**Data Cleaning:**

## 

## In our dataset “healthcare-dataset-stroke-data.csv”, we had 201 missing values. We replaced the missing values with the mean of that column. We differentiate categorical and numerical variables. Also we converted categorical variables into numerical for better visualization. Along with this we found out that there is no use to column “ID”, for that reason we dropped that column. This cleaned file was used to do the data visualization and to run machine learning models. Fig. 1 below displays the datasets statistical summary.

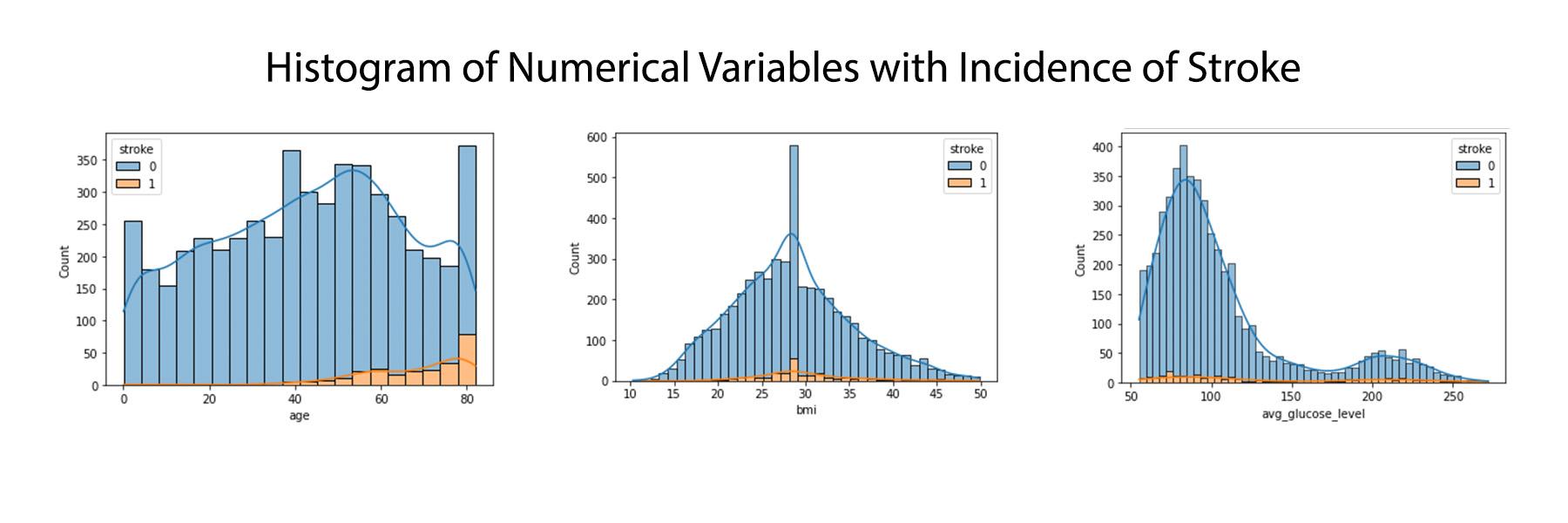


**Breakdown of all variable with categories and their stroke incidence percentage:**



**Data Visualization for Exploratory Data Analysis**

***Numeric Variables Analysis:*** *age, bmi, avg\_glucose\_level*

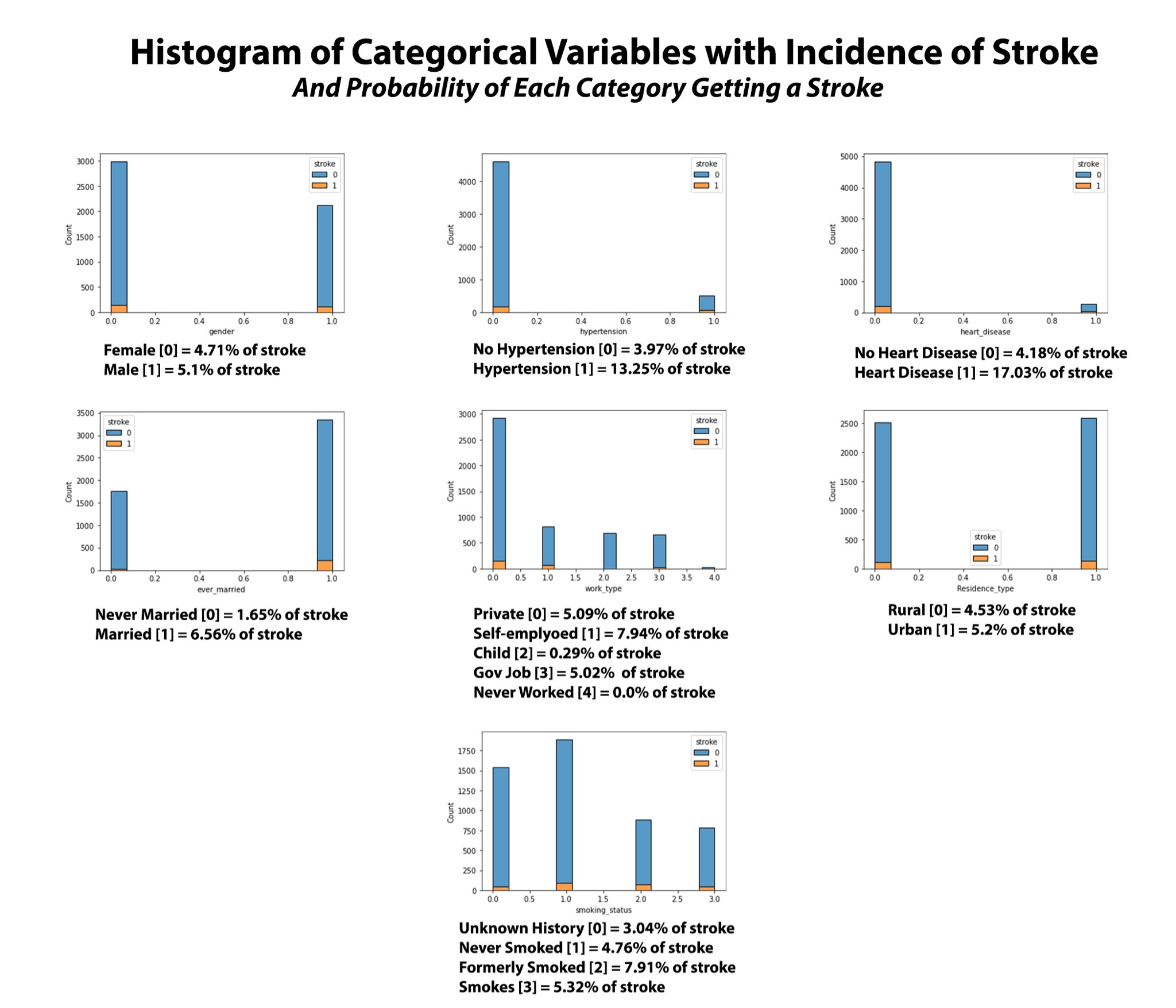


Age is normally distributed but has a much higher incidence of stroke as age increases. Ages below 40 had very few stroke incidences while ages above 60 seemed to increase in stroke. Graphical plots like histograms help us visualize this and tell us a story about our dataset that helps draw conclusions. We expect this observation because older people may suffer from a decline in health as they get older which would make them vulnerable to the risk of stroke versus a younger healthier person.

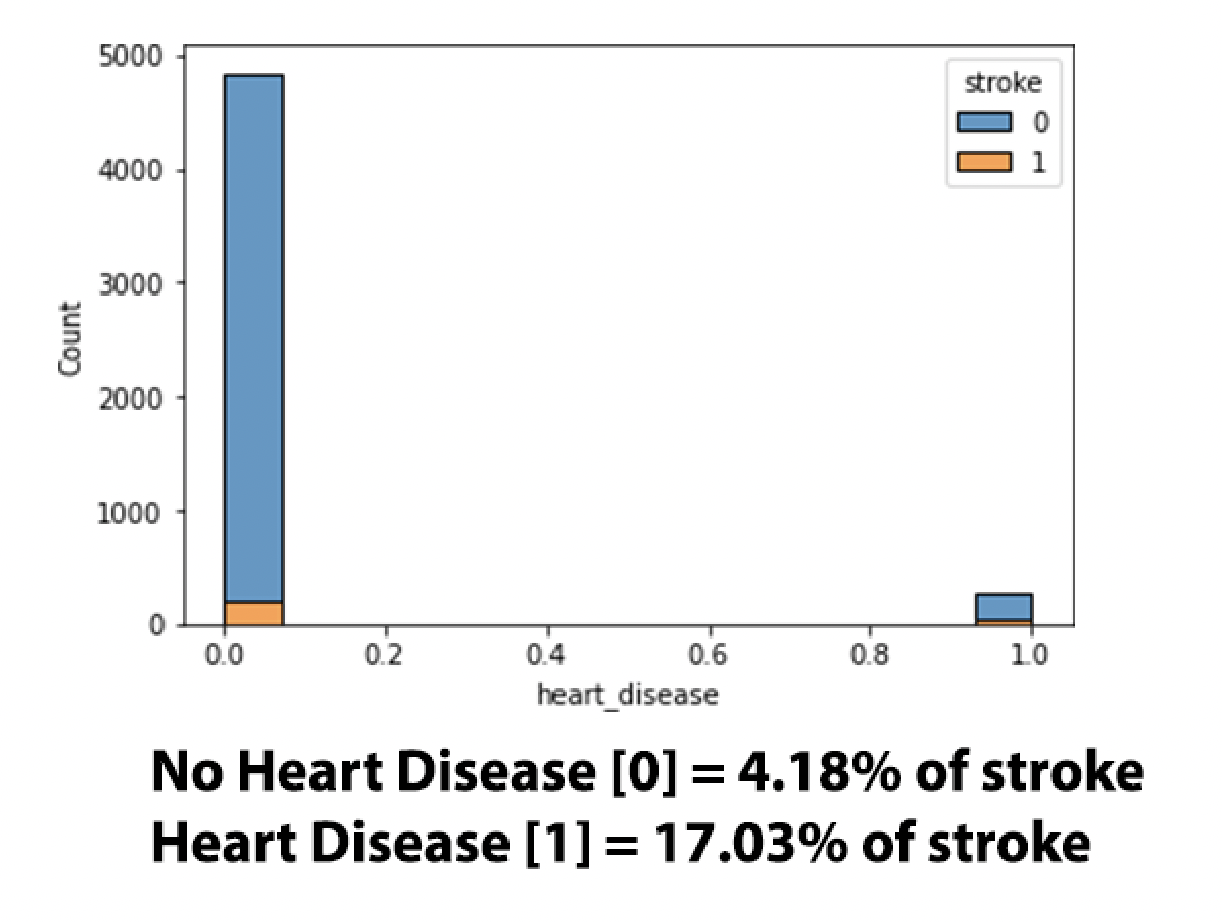
Stroke incidences increase as “bmi '' increases but eventually tapers down after bmi=40 because the likelihood of a person with a BMI over 40 is less frequent. Stroke increases with higher bmi values make logical sense because BMI is an indication of a person’s physical health, and poor health will lead to greater stroke risk.

“Avg\_glucose\_level” had two areas with increased stroke incidences. The first is within a normal glucose level (for people that suffer a stroke with no relation to glucose levels) and then interestingly when glucose levels are above 160 (which may indicate that high blood sugar increases the risk of stroke as poor glucose control leads to diabetes and a list of other health complications).

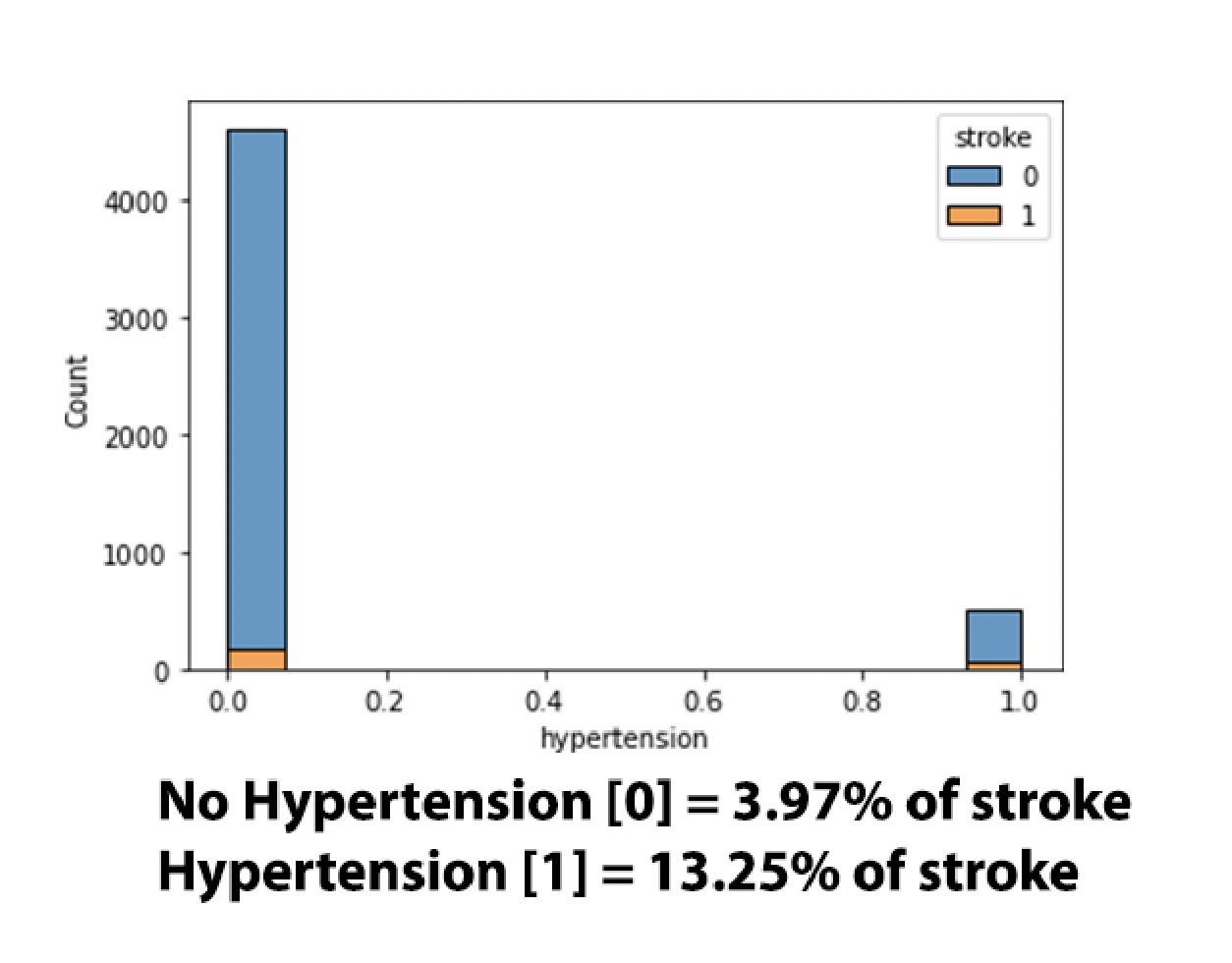
***Exploring Categorical Variables:*** *gender, ever\_married, work\_type, Residence\_type, smoking\_status, hypertension, heart\_disease*

****

As we explore the distribution and incidence of stroke in our categorical variables, we notice an interesting distribution of categories in each variable. Factors related to health such as (***heart disease and hypertension***) always increase the risk of stroke when present. For ‘heart\_disease’ we have significantly less samples of people with both stroke and heart disease which may lead you to think that hypertension is insignificant to stroke risk but with a closer look, the proportion of people with stroke and heart disease is 17.03% vs. 4.18%. The presence of ***heart disease increases stroke risk by 4 times***than with people with no heart disease.



For ‘hypertension’ there may seem to be more individuals with stroke and *no hypertension* vs. individuals with both but that is misleading because the proportion of people with hypertension and stroke is about 3 times greater than without hypertension (probability w/ hypertension 13.25% vs w/o hypertension 3.97%). This means ***hypertension increases stroke risk by 3 times***.



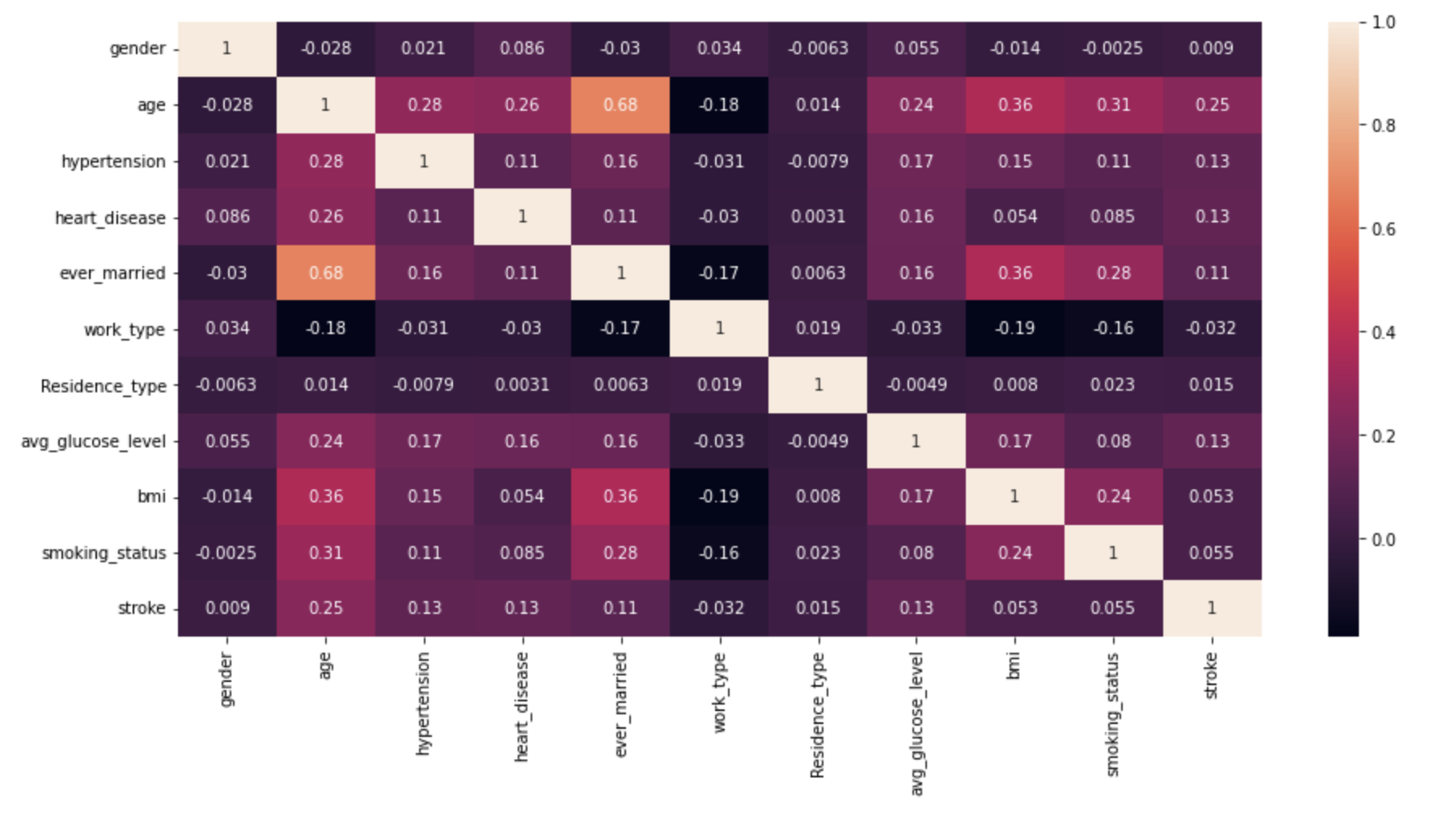
This makes logical and medical sense because all three conditions (stroke, heart\_disease, hypertension) have similar mechanisms of pathophysiology (caused by blockage in the arteries that can lead to cutting off blood supply to a major organ like the brain or heart). This information is very helpful because it can help in minimizing or preventing major health events related to stroke. Early medical intervention and medical knowledge can also help as preventive medicine.

Relationship status as a category may seem to be odd since we are exploring a health dataset but variables like this show the importance of exploring information that may or may not be useful. People who are married or were married have a stroke probability of 6.56% vs. People who have never been married have a 1.65% stroke probability. This means that people who have been in a relationship ***increase their risk of stroke by 4 times***. This makes partial sense because I can agree that relationships may be stressful at times and that pressure can potentially cause an increase in health risks but the probability of increasing risk by ***4 times*** like heart disease seems to be inaccurate. Further investigation of this variable is needed.



**Correlation Matrix:**

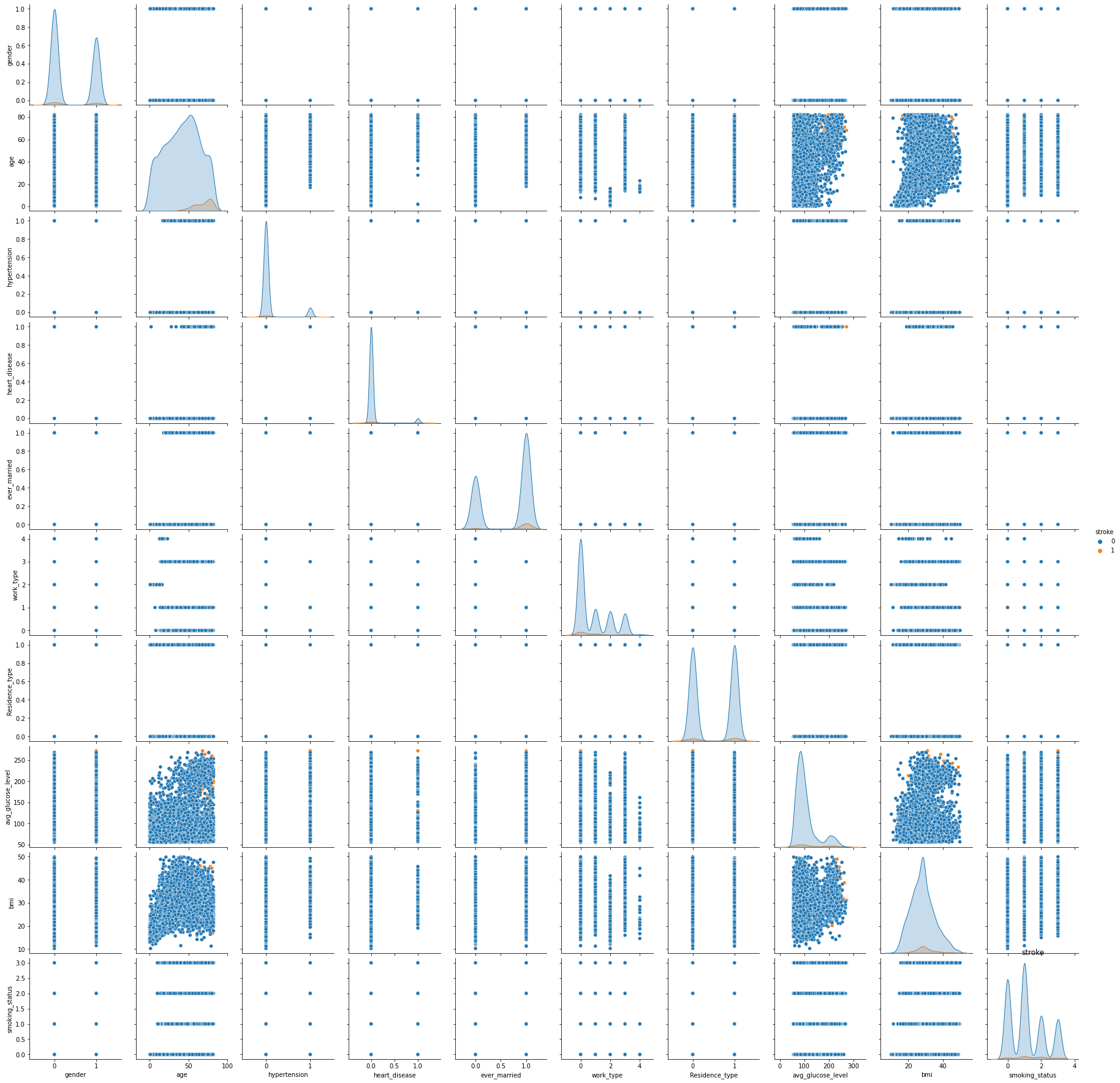
Plotting a correlation matrix of all the variables helps visualize the relationships between the independent variables. The closer the correlation coefficient of two variables to +1 the stronger the positive correlation and vice versa; the closer the coefficient is to –1 the stronger the negative correlation. Coefficient values closest to 0 have no correlation or strength.

****

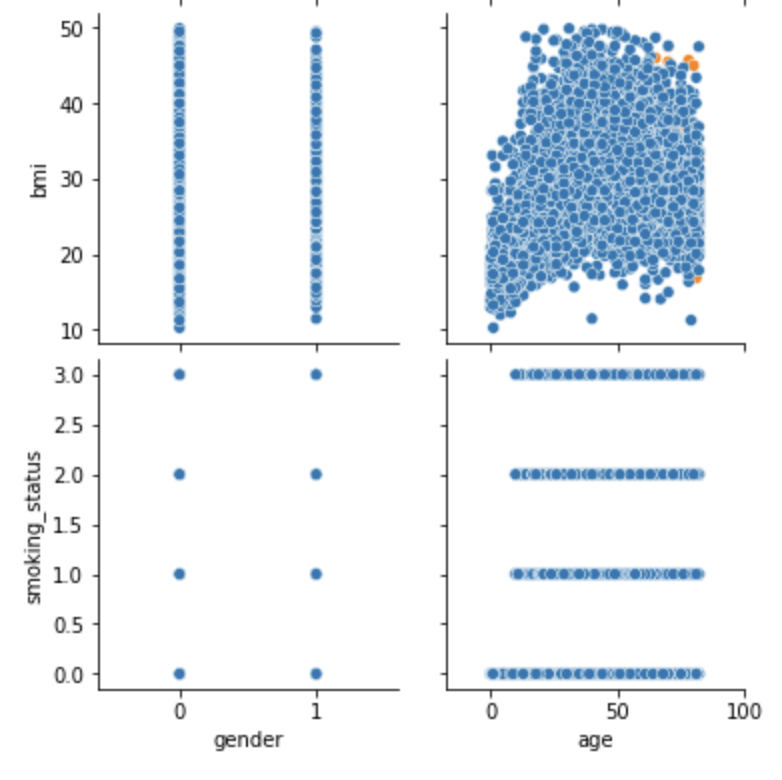
The strongest correlation was ***0.68*** which is for variables ‘age’ and ‘ever\_married’. Correlation coefficient of 0.68 would indicate a strong positive correlation, but this needs to be further investigated because it may be inaccurate due to the disproportional density of the marriage categories. The second strongest correlation was ***0.36*** which is a fairly weak positive correlation between ‘age’ and ‘bmi’.

The weak negative correlation observed is for ‘work\_type’ and ‘smoking\_status’ = ***-0.16***, ‘bmi’= ***-0.19***, ‘ever\_married’= ***-0.17***, ‘age’ = ***-0.18.*** This would mean that, as long as work type was ‘never worked’, ‘child’ or ‘government job’ the negatively correlated factors would decrease a little bit. We can see lower bmi, lower smoking status (no smoking or unknown), unmarried, and younger aged individuals.

**Scatter Plot of stroke dataset:**



The scatter plot above indicates the weak positive correlation between age and bmi as the correlation coefficient was 0.36

****

**Methods Used:**

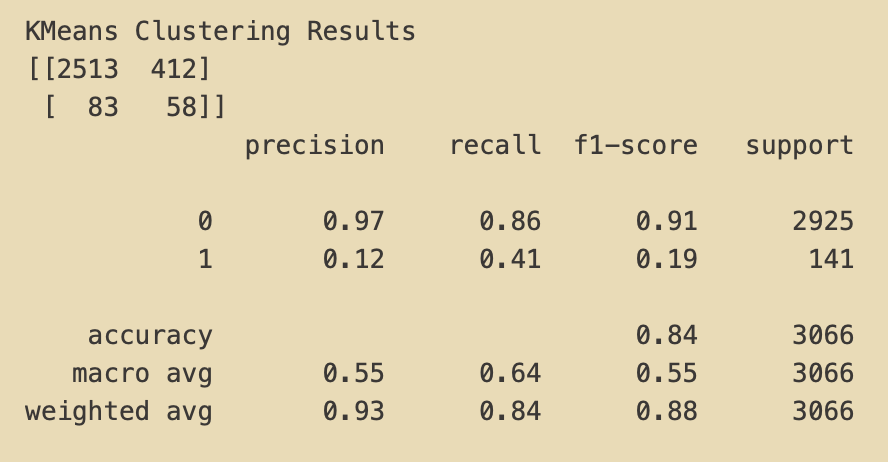
We used a variety of supervised learning algorithms and one unsupervised learning algorithm.

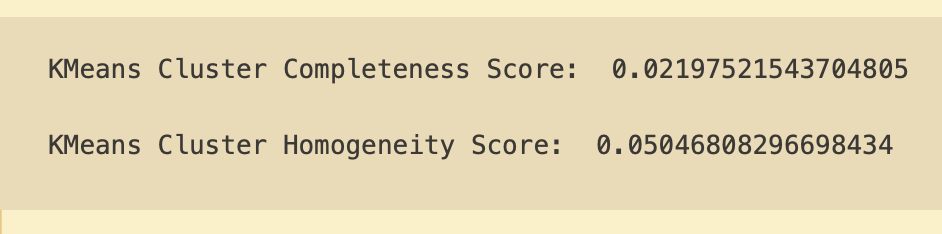
KMeans Clustering:

We wanted to try using an Unsupervised Learning Algorithm in order to explore the data and see distinctions that the algorithm found. KMeans is not a classifier.

For this part, we attempted to use KMeans Clustering to cluster the data into two groups: Stroke YES and Stroke NO. It is very important to note, we actually had the labels for this data set. However, we did NOT use them for the KMeans clustering algorithm since it is an unsupervised learning algorithm. Under normal circumstances, we use the KMeans algorithm because we do not have labels.

In this case, we will use the labels to test how well the algorithm performed, but we will not usually do this for KMeans. So, the classification report and confusion matrix at the end of this section don't truly make sense in a real-world setting. Additionally, we could not use the ROC curve to gauge the KMeans model because the sklearn accuracy plot only works for testing supervised learning algorithms that classify. Instead of the ROC, we used the homogeneity report.





The Completeness Score is when all members of a given class are assigned to the same cluster. When the score approaches 1, most of the data points that are members of a given class are elements of the same cluster. Therefore, this Completeness Score means that most of the target labels were in different clusters.

The Homogeneity Score is a measure of each cluster containing only members of a single class. The Homogeneity Score approaches 1 when all the clusters contain almost only data points that are members of a single class. Since the KMeans score is 0.05, the model’s clusters are not helpful.

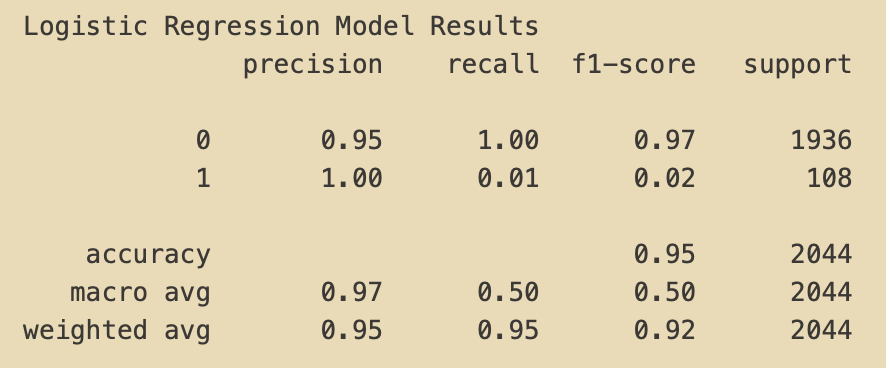
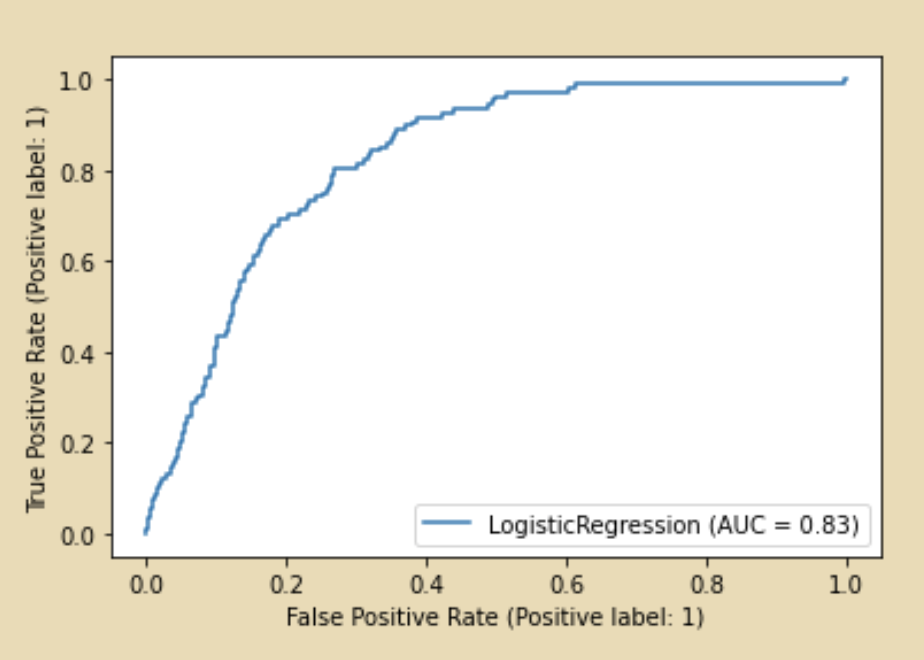
While the KMeans Cluster had a comparatively lower accuracy score (.84), it had a similar F-1 score (.91) as the classifier models.

**Logistic Regression**:

We used Logistic Regression because it is well-suited for predicting a discrete value: yes or no. Logistic Regression models the probability of some event occurring as a linear function of a set of predictor variables. In our study, we wanted to predict whether or not someone would have a stroke given their attributes.

We used the sklearn library and created a model, fit it to the training set (X\_train and y\_train), and called the prediction function. The predict function predicts using the linear model.

The logistic regression model was evaluated with the Score Method, the ROC Curve Graph, Confusion Matrix and the Classification Report. The Score Method returns the coefficient of determination of the prediction for which the model got .95. The logistic regression model got a .95 accuracy rate and an f1-score of .97. Most impressive is that the model had a score of .83 for area under the curve.

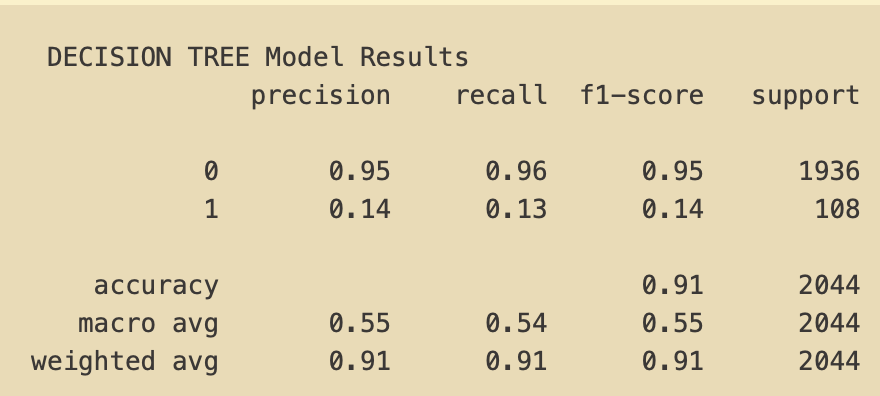


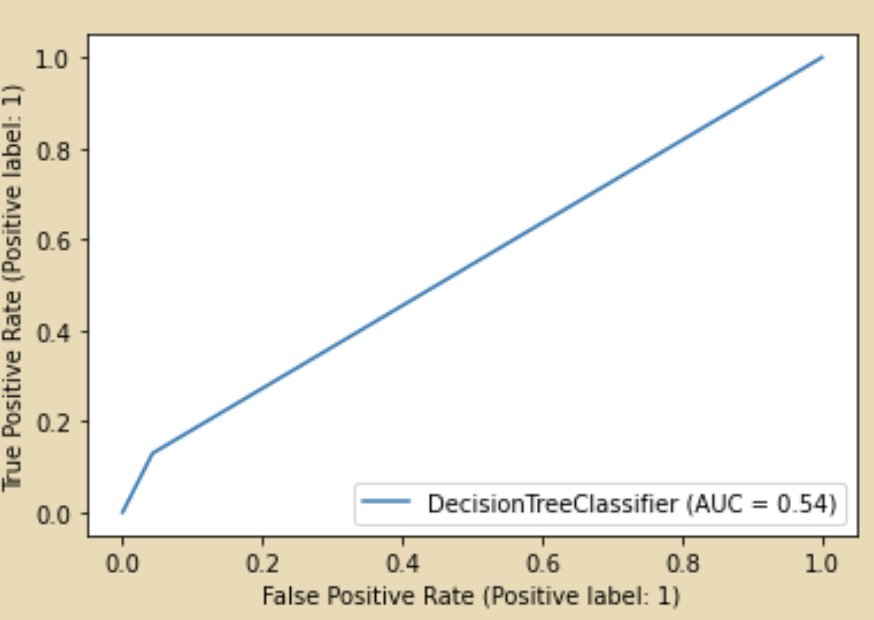
**Decision Tree**:

Decision Trees are often more accurate than linear regression when instances are not represented well by simple linear models. Decision trees are a popular supervised learning method.

We created a decision tree model and fit it to the stroke training data set. We called the predict method on the instance and assigned it to a variable.

We used the sklearn metrics and classification report library to evaluate the model. It had a lower (91%) accuracy than the logistic regression model, and it had a lower F-1 score (.95 ) than all the other models. The decision tree model had a relatively low area-under-curve score (.54) as well. A model with no skill is represented at the point (0.5, 0.5). A model with no skill at each threshold is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. The decision tree’s score of .54 is not much of an improvement from .5.



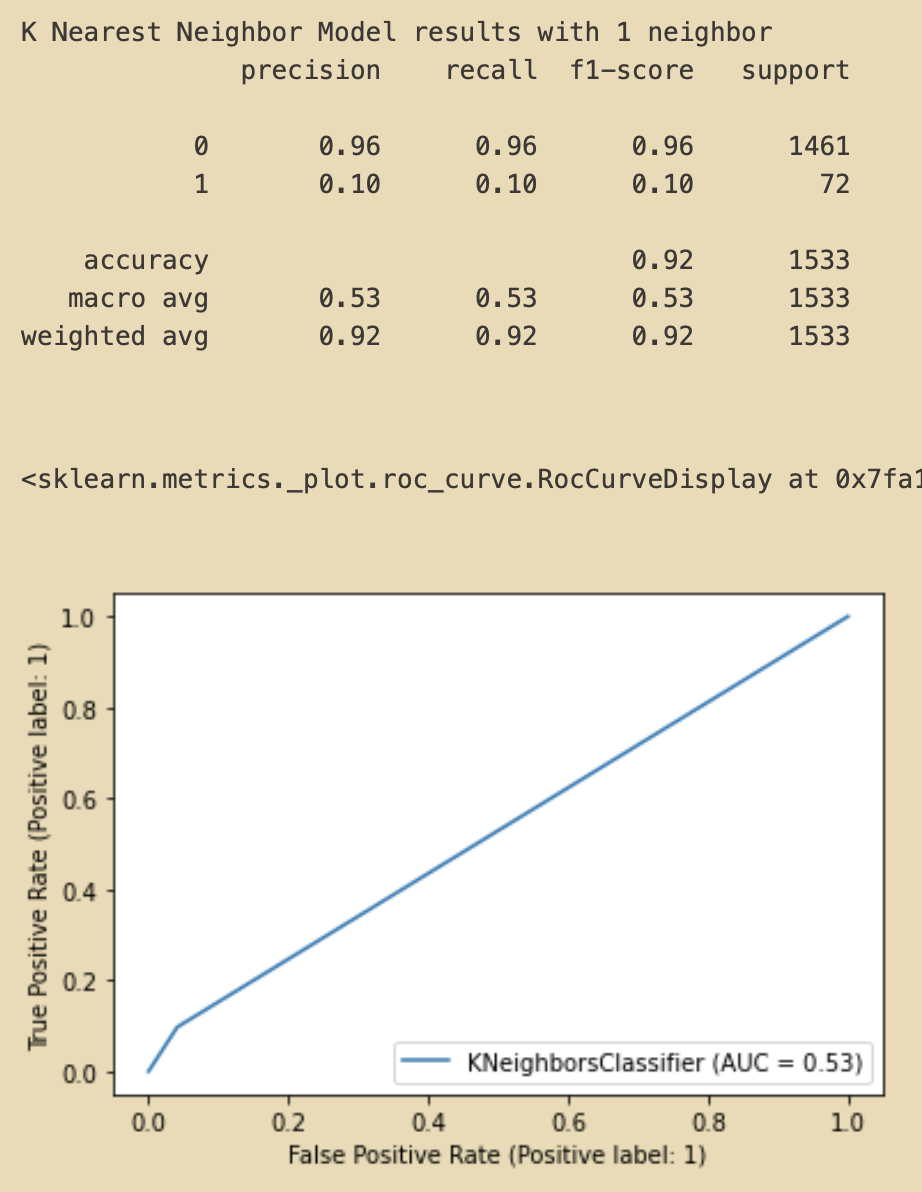


Next, we made a K Nearest Neighbor algorithm. Before we began creating the KNN model, the training set had to be scaled. We prepared the training set to be used by the KNN model by scaling all the values. We standardized the feature values because the KNN classifier predicts the class of a given test observation by identifying the observations that are nearest to it. The scale of the variables matters. Any variables that are on a large scale will have a much larger effect on the distance between the observations, and hence on the KNN classifier, than variables that are on a small scale.

We used the sklearn StandardScaler which standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. StandardScaler results in a distribution with a standard deviation equal to 1.

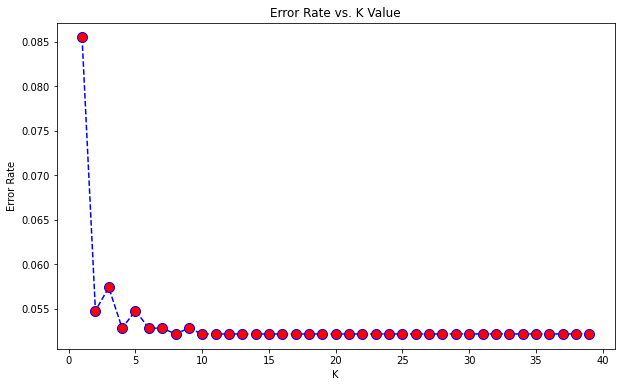
Then, like the other classifier models, wefit the model to the training set and called the prediction function.

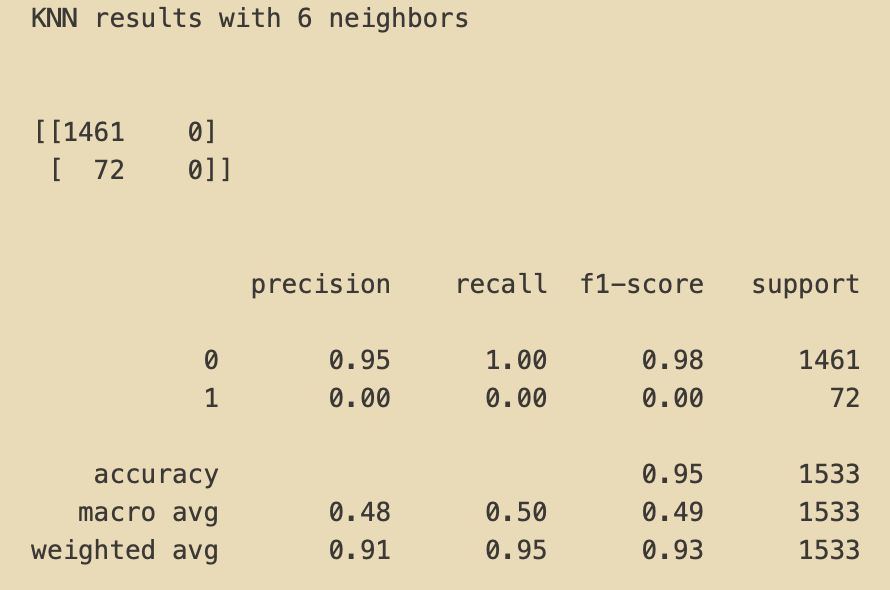
We used the classification report and the ROC curve to evaluate the model. The initial results of the KNN prediction were good, as you can see, in the figure below. Although the accuracy score was 1% higher, the area under the curve for the KNN using K=1 was not much improvement over the Decision Tree model’s AUC score.



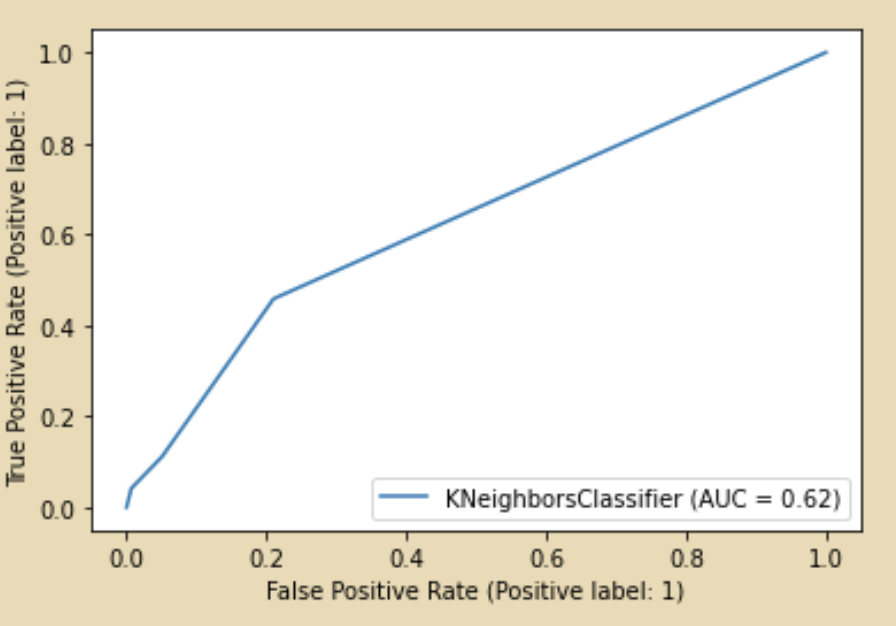
In general, different values of *k* affect the outcome of classification

If k is too small, it is sensitive to noise (noisy points). If it’s too large, it can include points from the other classes. So, it is worth experimenting with other values of k in order to find a lower error rate. In order to do this, we iterated through values of k while fitting a model with the k value, predicting and evaluating that model’s error rate. This created the figure below (Error Rate vs K Value). As you can see, the error rate goes down and does not improve past K=6.





As you can see from the classification report (above), tuning the model and changing the k value to 6 improved the accuracy 0.03 higher to 0.95. Additionally, tuning the model improved the area-under the-curve 0.09 points to 0.62 (see figure below).



**Conclusion:**

In conclusion, the best model would have a high F-1 score, ROC auc score and accuracy. If one of these is high and the others are low, something must be wrong with the model. Given that the logistic regression model had a relatively high F-1 score (0.97), a high accuracy score, (0.95), and the highest ROC auc score of 0.83, it is the most reliable model for predicting. The tuned KNN model was a close second place, but its ROC auc score was 0.19 lower than the logistic regression model’s ROC auc. Therefore we choose to use the logistic regression model to predict strokes.

**Contributions:**

Jaimi Patel:

Generated Colab file to share working code. Loaded, formatted and visualized the dataset with the correct data points and variables into our python script. Data preprocessing and replacing the found outliers and missing values in data with mean values. Separated categorical and numerical data for future analysis. Visualized the number of unique values per variable. Created dummy variables for categorical data. Created histograms for numerical variables and pie charts for categorical data to analyze distribution. Helped in preparing demo slides. Prepared Executive summary and final report.

Jacob Crawford:

Used sklearn train test split to create a training set and a test set. Explored KMeans Clustering as predictor and reported the homogeneity score. Built a Logistic Regression model to predict a binary value. Used the sklearn library to create a classification report. Built a decision tree model to predict a binary/discrete value. Prepared the training set to be used by the KNN model by scaling all the values. Built a K Nearest Neighbor (KNN) model and fit it and trained it to predict from the training set. Evaluated the KNN model and examined the accuracy with the sklearn classification report and ROC graph. Graphed the error rate vs the K (number of neighbors). Used the new K value from the graph to see if that improved the accuracy. Reviewed the ROC curves, F-1 ratings and accuracy scores to determine the best model to predict strokes.

Seif Abuhashish: Project proposal, group effort Executive Summary, data loading, cleaning, and preprocessing, Exploratory Data Analysis for numeric and categorical variables. Histogram analysis of category distribution with probability percentage of stroke. Correlation matrix and scatter plot analysis. Presentation Slides.