# Stroke Prediction Project

**Eduardo Herrera, Antonio Cruz, Thanh Ha, Josh Wilks**

*May 5, 2021*

## Overview

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

## Objective

Our goal is to visualize relationships between variables to uncover unhealthy habits that may lead to strokes. In addition, we seek to build models that can accurately predict the probability a stroke will occur given a variety of health factors. We choose to run several models and hope to pick the model with the highest accuracy scores. We are using a Stroke dataset that we obtain from Kaggle.

## Attributes

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient does not have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient does not have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

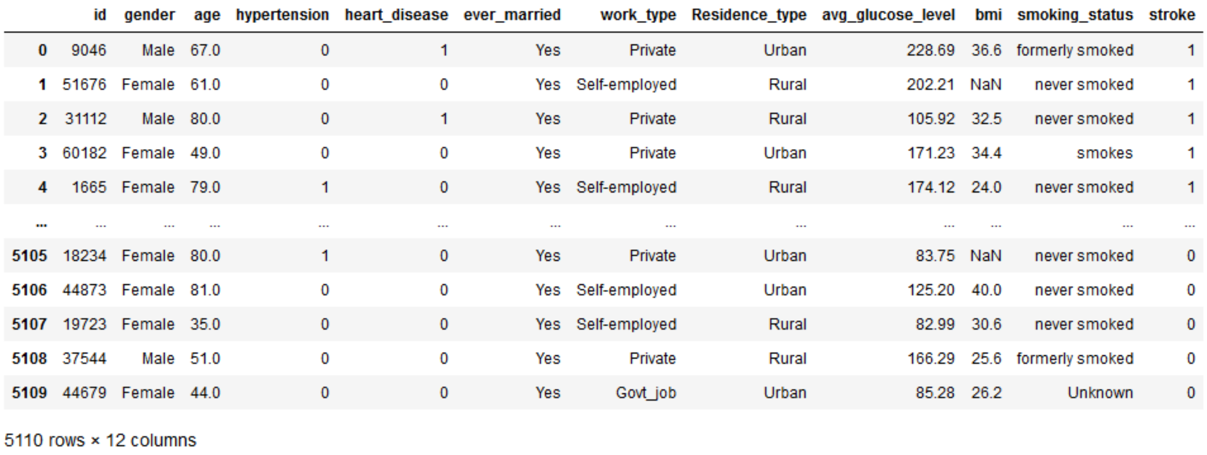
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

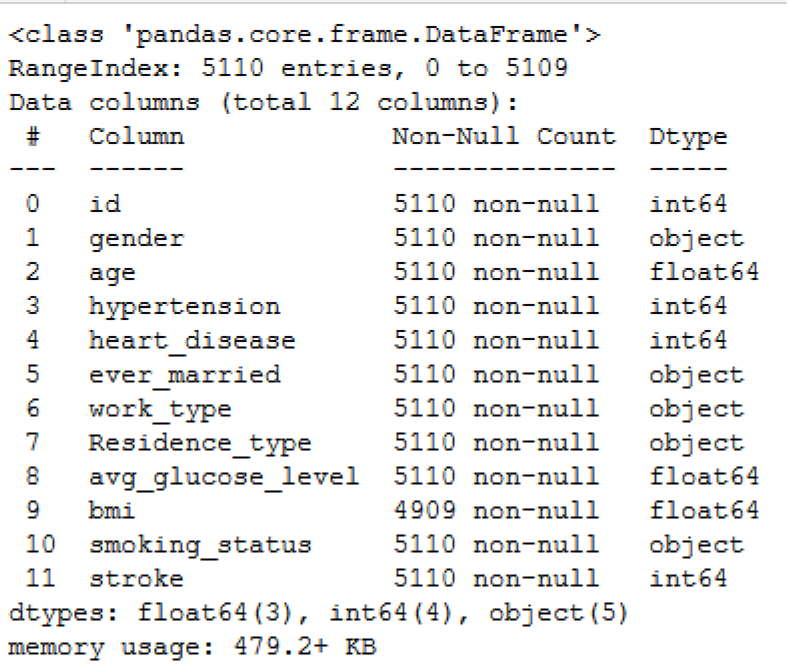
# Data Exploration

Below we take a quick look at our data set after loading into Jupyter Notebook. We see all our features as well as the total 5110 observations in our data set. We also take a look at the data types for each variable. We have 3 variables, age, bmi, and avg_gluclose_level, that are float variables. We have 3 binary variables, hypertension, heart_disease, and stroke that are int types. Stroke will be our target variable. We also have 5 categorical variables, gender, ever_married, work_type, Residence_type, and smoking_status, as object type.

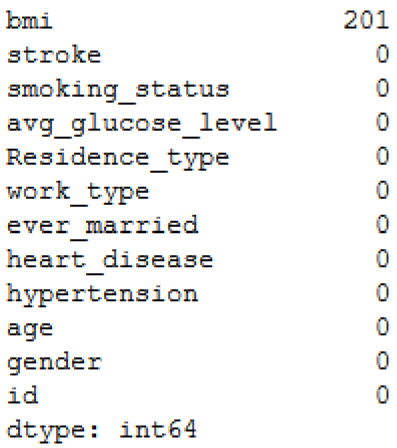


| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 9046 | Male | 67.0 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked | 1 |
| **1** | 51676 | Female | 61.0 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | NaN | never smoked | 1 |
| **2** | 31112 | Male | 80.0 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked | 1 |
| **3** | 60182 | Female | 49.0 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| **4** | 1665 | Female | 79.0 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24.0 | never smoked | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5105** | 18234 | Female | 80.0 | 1 | 0 | Yes | Private | Urban | 83.75 | NaN | never smoked | 0 |
| **5106** | 44873 | Female | 81.0 | 0 | 0 | Yes | Self-employed | Urban | 125.20 | 40.0 | never smoked | 0 |
| **5107** | 19723 | Female | 35.0 | 0 | 0 | Yes | Self-employed | Rural | 82.99 | 30.6 | never smoked | 0 |
| **5108** | 37544 | Male | 51.0 | 0 | 0 | Yes | Private | Rural | 166.29 | 25.6 | formerly smoked | 0 |
| **5109** | 44679 | Female | 44.0 | 0 | 0 | Yes | Govt_job | Urban | 85.28 | 26.2 | Unknown | 0 |

5110 rows × 12 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

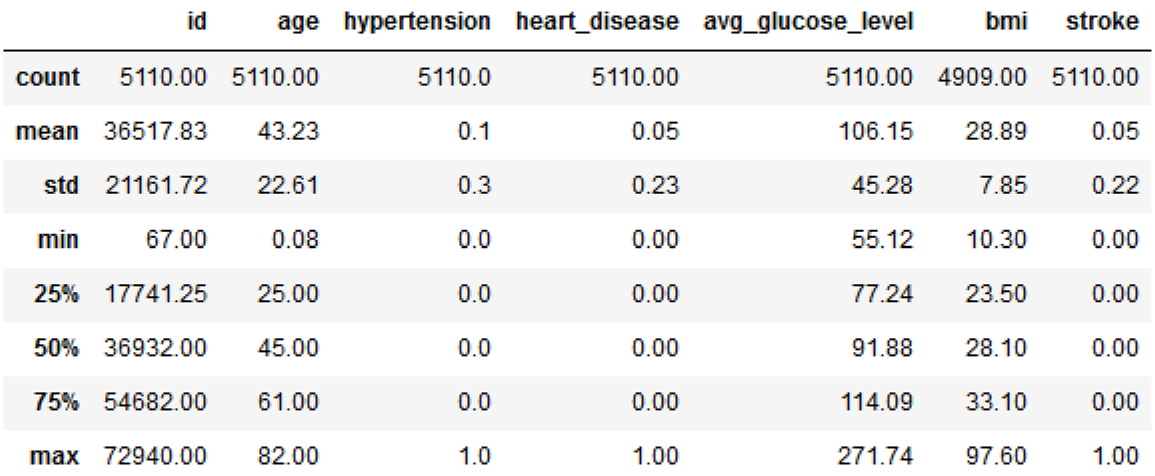


```
bmi                  201
stroke                 0
smoking_status         0
avg_glucose_level      0
Residence_type         0
work_type              0
ever_married           0
heart_disease          0
hypertension           0
age                    0
gender                 0
id                     0
dtype: int64
```

## Summary Statistics

### Numerical Variables

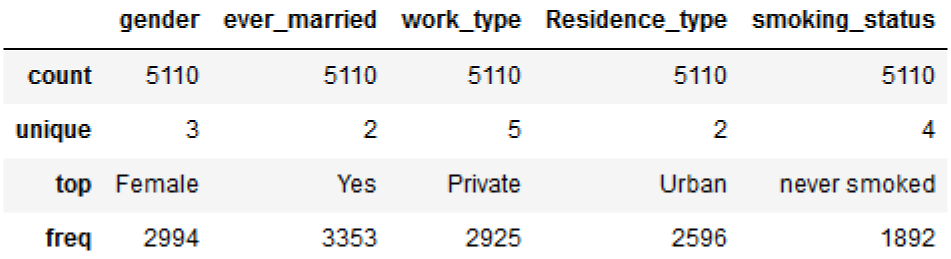
Below is the summary statistics for numerical variables. We see once again we are missing values for the bmi variable.



| | id | age | hypertension | heart_disease | avg_glucose_level | bmi | stroke |
|---|---|---|---|---|---|---|---|
| **count** | 5110.00 | 5110.00 | 5110.0 | 5110.00 | 5110.00 | 4909.00 | 5110.00 |
| **mean** | 36517.83 | 43.23 | 0.1 | 0.05 | 106.15 | 28.89 | 0.05 |
| **std** | 21161.72 | 22.61 | 0.3 | 0.23 | 45.28 | 7.85 | 0.22 |
| **min** | 67.00 | 0.08 | 0.0 | 0.00 | 55.12 | 10.30 | 0.00 |
| **25%** | 17741.25 | 25.00 | 0.0 | 0.00 | 77.24 | 23.50 | 0.00 |
| **50%** | 36932.00 | 45.00 | 0.0 | 0.00 | 91.88 | 28.10 | 0.00 |
| **75%** | 54682.00 | 61.00 | 0.0 | 0.00 | 114.09 | 33.10 | 0.00 |
| **max** | 72940.00 | 82.00 | 1.0 | 1.00 | 271.74 | 97.60 | 1.00 |

### Categorical Variables

Below is the summary statistics for categorical variables. We see the number of unique categories for each variable as well as the most frequent category and its count. We see that most people in the stroke data set are female, are married, work in the private sector, live in an urban residence, or have never smoked.

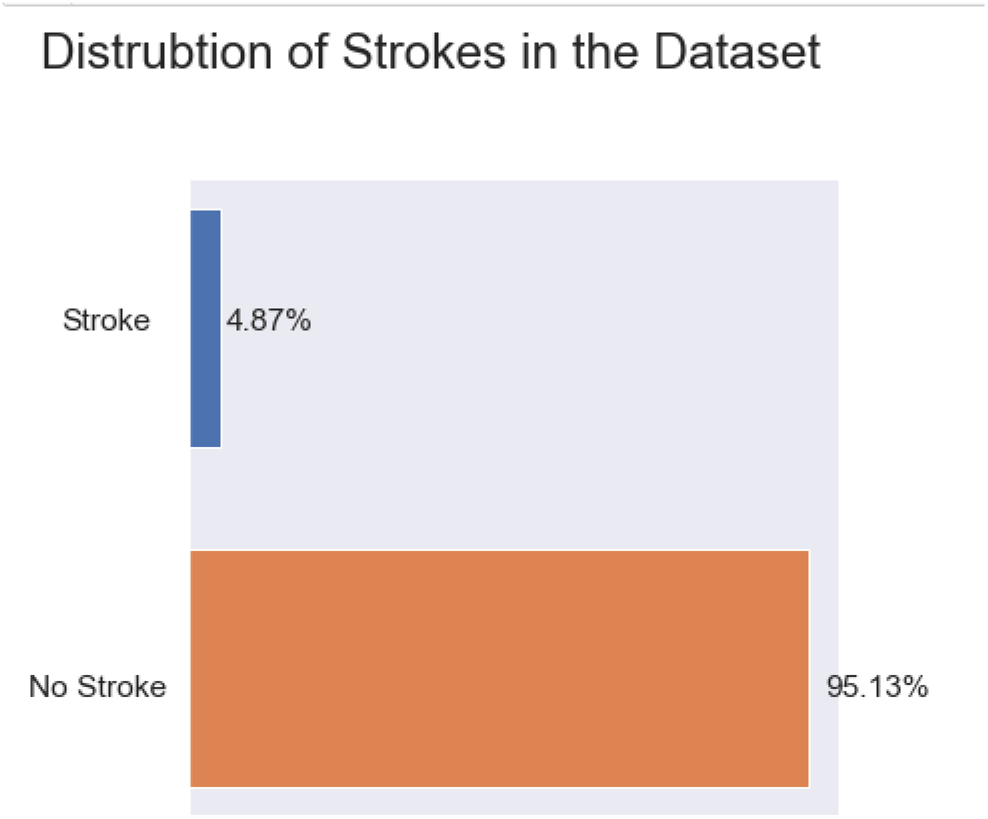


| | gender | ever_married | work_type | Residence_type | smoking_status |
|---|---|---|---|---|---|
| **count** | 5110 | 5110 | 5110 | 5110 | 5110 |
| **unique** | 3 | 2 | 5 | 2 | 4 |
| **top** | Female | Yes | Private | Urban | never smoked |
| **freq** | 2994 | 3353 | 2925 | 2596 | 1892 |

## Data Visualization

### Stroke Distribution in the data set

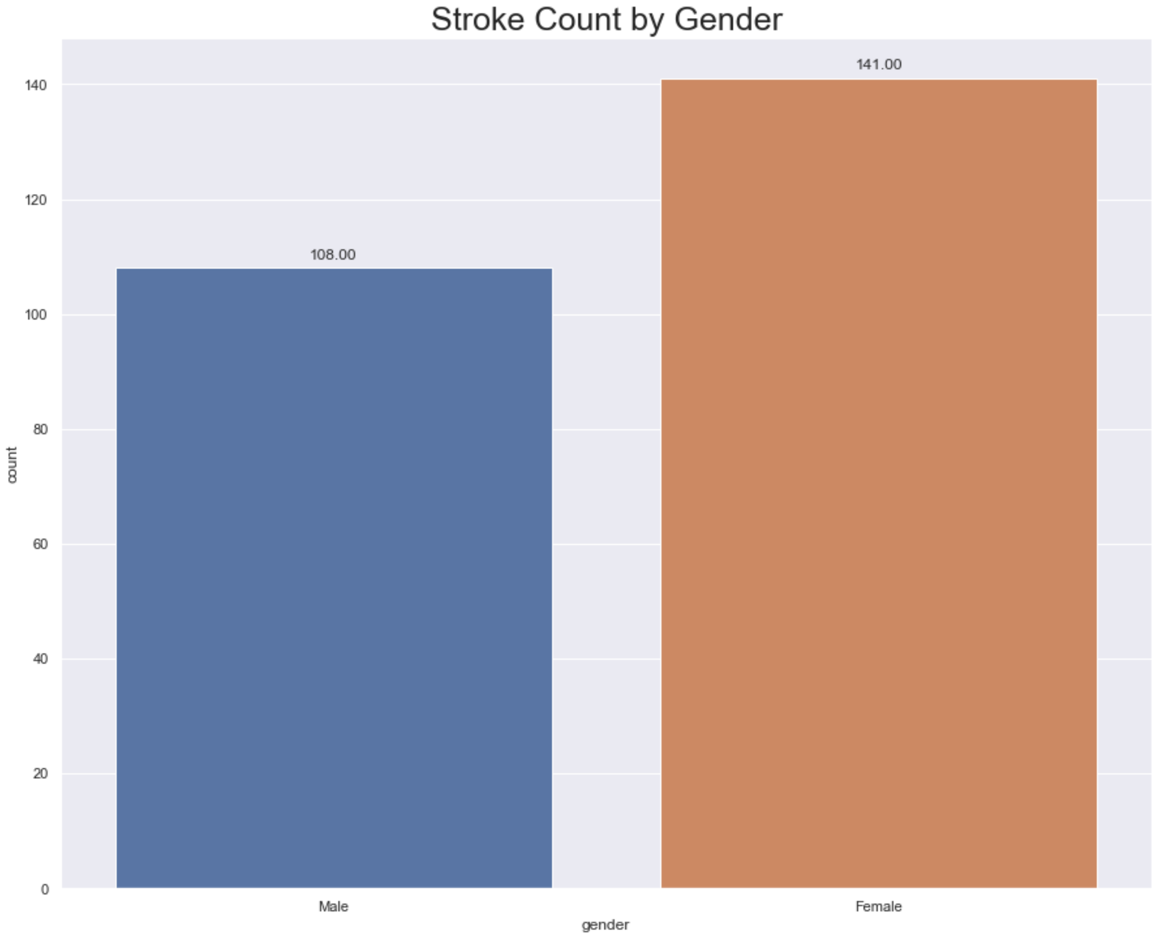
As you can see from the chart, the dataset is highly unbalanced. Most people in the dataset have not suffered a stroke. A baseline model should be able have a prediction accuracy score of 95%. A model that only predicts no strokes will have achieve this because 95% of the observations do not suffer strokes. To overcome this problem and build a proper model, we must balance our dataset, since the instance we are trying to predict occurs so rarely in our dataset. To accomplish this, we use the SMOTE package to oversample the number of strokes to achieve a balanced dataset.



Distrubtion of Strokes in the Dataset

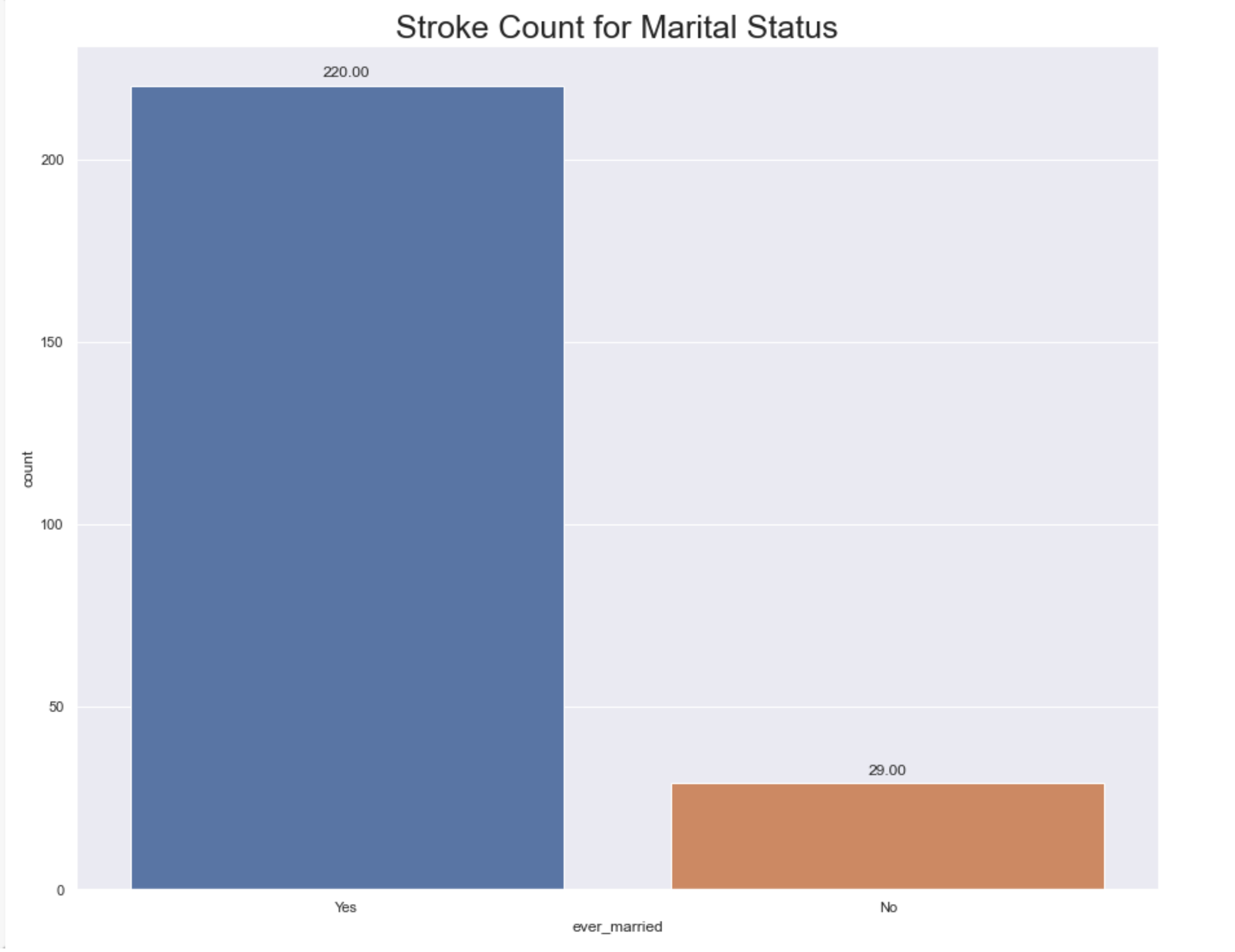
## Visualization for Gender

This graph shows the distribution of the total number of strokes across gender. So, of all the strokes that occur in our dataset the major can be attributed to women. However, to get a clearer view, we should look at the proportion of strokes within each gender. That is to say what percentage of men and women suffer strokes. We found that there are 2994 women, 2115 men, and 1 person who identified as other in the dataset. When looking at individuals who suffered strokes, there were 141 women and 108 men. We found that 4.71% of women suffered strokes while 5.11% of men suffered strokes. So, while the count plot below shows women suffer more strokes than men, we see that men suffer strokes at a higher rate compared to women.



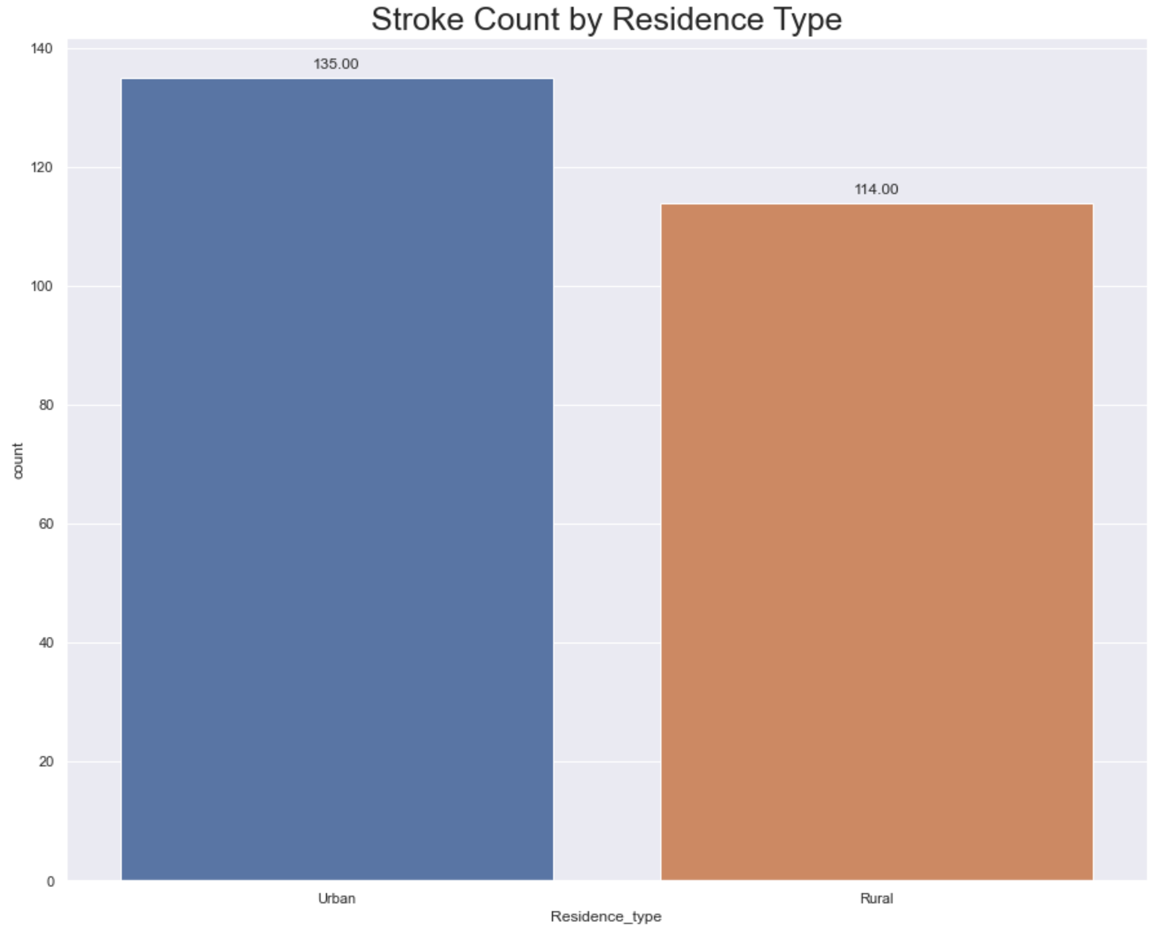
### Visualization for Marital Status

This graph below shows the total number of strokes among marriage status. So, there were 220 people who are married that suffered strokes. There are 29 people who are not married who suffered strokes. Looking at this, we are tempted to say that you are more likely to suffer a stroke if you are married. However, again we need to know the percentage of married and single people who suffer strokes instead of the total count. We found there are 3353 married people and 1757 single people in the data set. Again, 220 married suffered strokes and 29 single people strokes. So, 6.56% of married people suffered strokes and 1.65% of single people suffered strokes. In this case, married people suffer more strokes and have a higher chance to suffer a stroke compared to single people.



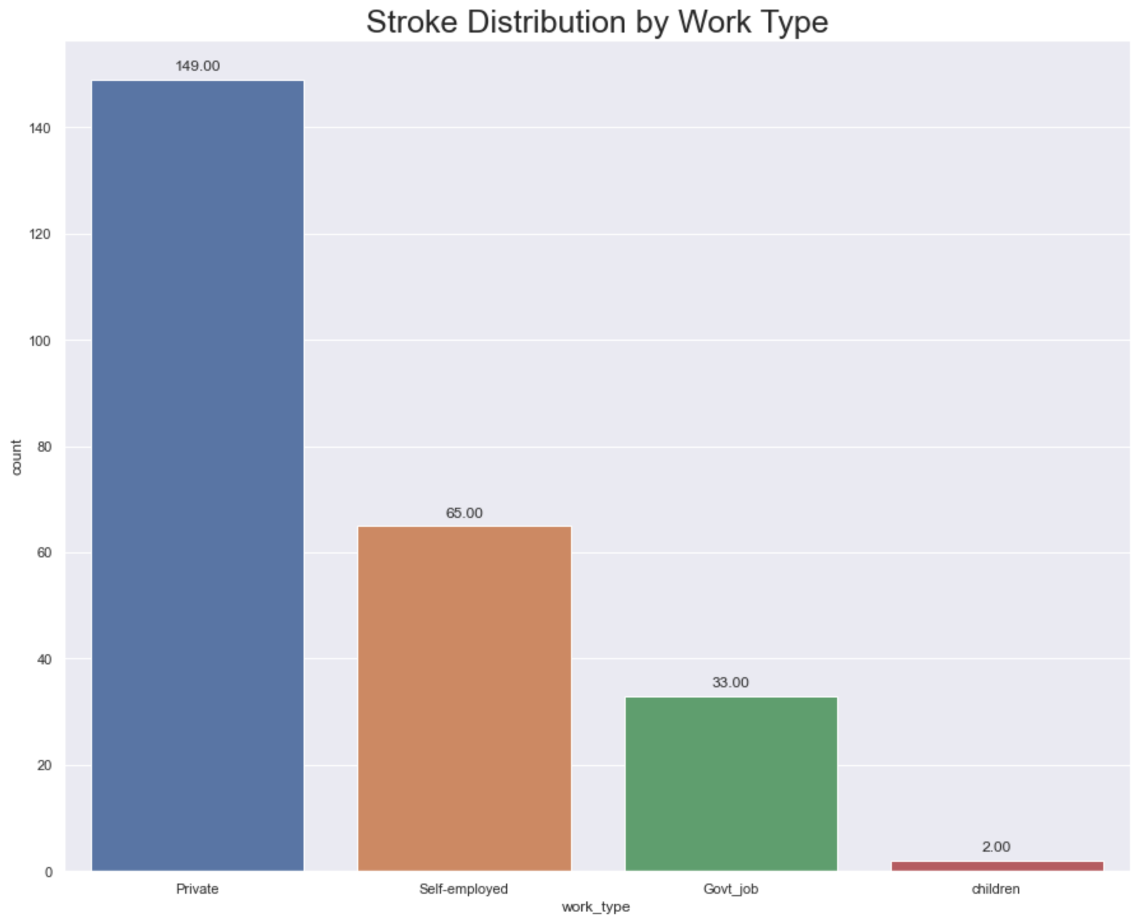
### Visualization for Urban and Rural Residences

This graph shows the overall count of strokes among residence types. There are 135 people living in an urban residence who suffered strokes. There are 114 people living in a rural residence who suffered strokes. Looking at the overall data set, there are 2596 people who live in an Urban residence and 2514 people who live in a Rural residence. So, we find that 5.2% of Urban residents suffer strokes and 4.53% of rural residents suffer strokes. Overall, we find that there are more people living in Urban residences that suffer strokes and suffer strokes at a higher rate compared to people living in Rural residences.



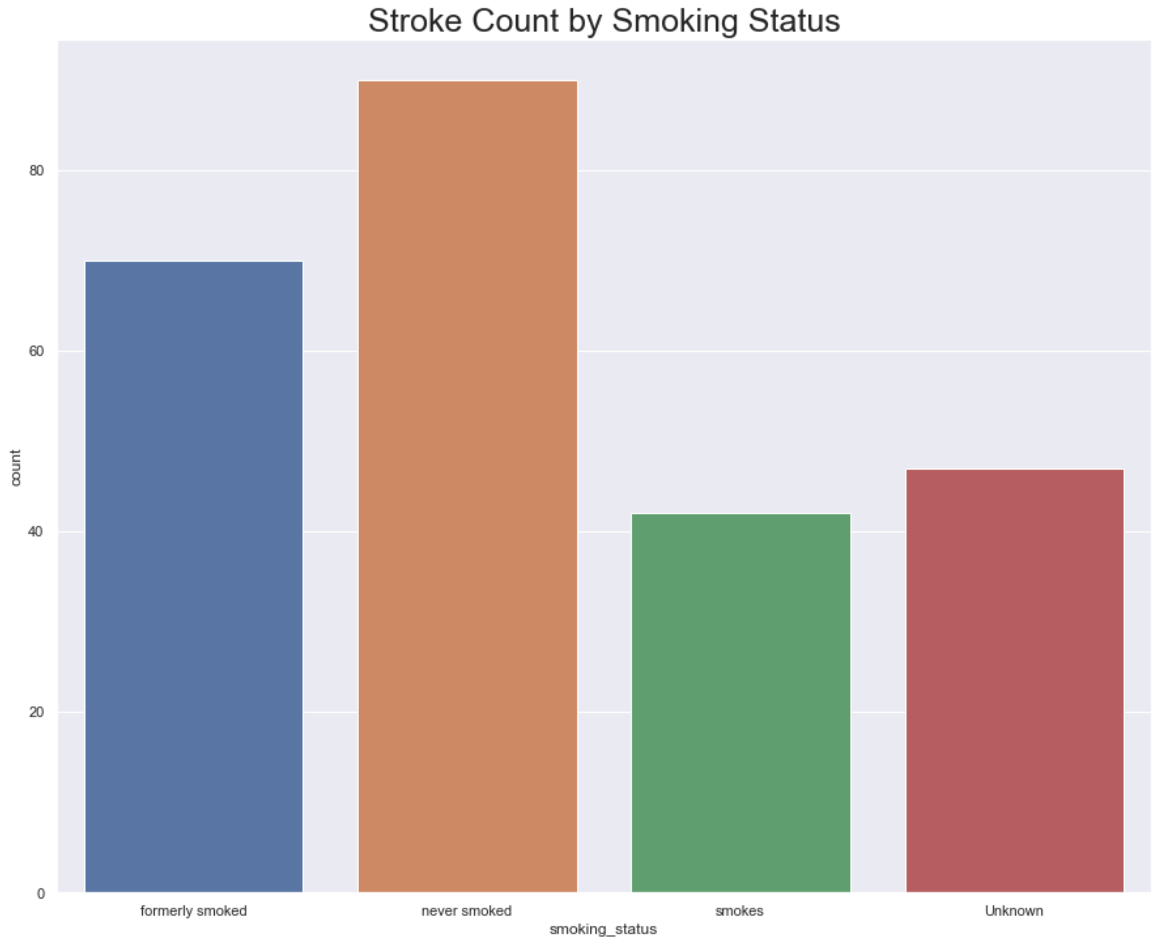
## Visualization for Work Type

The count plot below shows us the total count of strokes among work types. We see that the private sector accounts for most strokes in the dataset. However, the private sector also accounts for most of the observations in the dataset. We found that there were 2925 Private, 819 Self-employed, 657 Govt_job, 687 Children, and 22 never worked observations in the data set. So, while the graph below may lead you to believe the private sector may lead to more strokes, we find the 5.09% of private sector works suffer strokes. Meanwhile, 7.94% of self-employed people suffered strokes. We saw that 5.02% of Government works suffered strokes. Only 0.29% of children suffered strokes. We find that this graph is deceiving and while the private sector has the highest count of strokes, they do now have the highest proportion of strokes. Self-employed works suffer strokes at the highest rate.



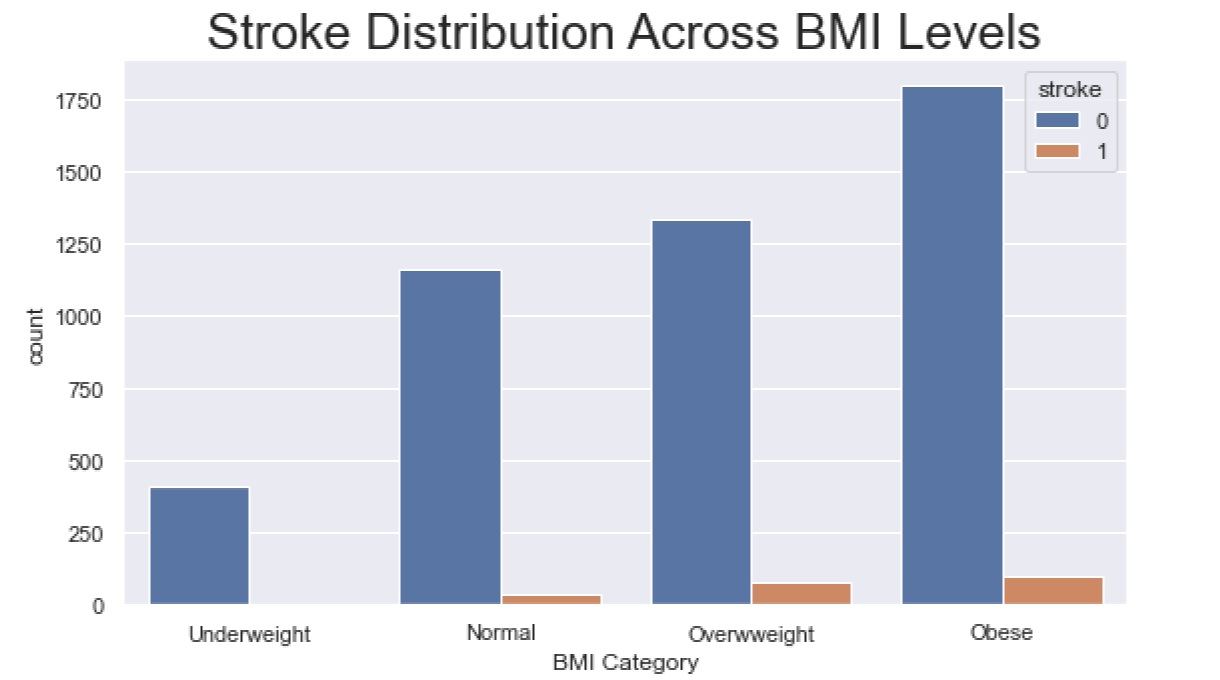
## Visualization for Smoking Status

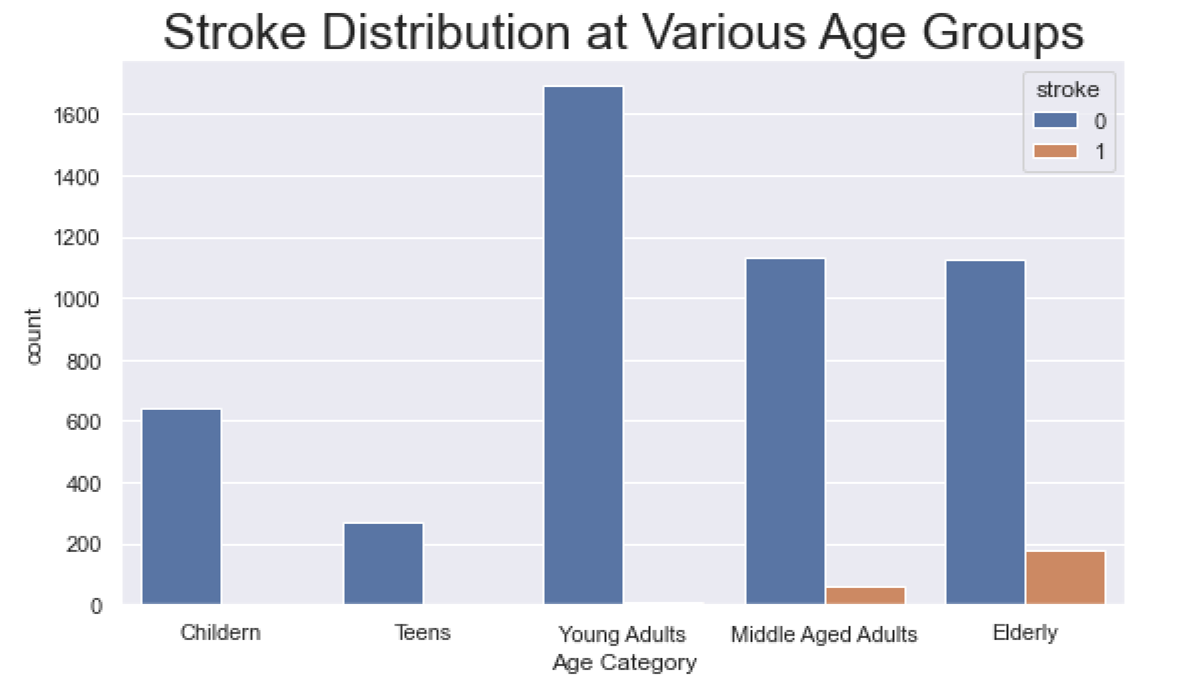
This count plot shows that people who never smoked accounts for most strokes in our data set. We need to figure out the proportion of strokes among each category to get a more accurate look at whether smoking status may have a relationship with strokes. We found that the dataset contained 1892 people who never smoked, 1544 with unknown status, 885 former smokers, and 789 smokers. We found that 90 people who never smoked, 70 former smokers, 47 people with unknown status, and 42 smokers suffered strokes. Overall, we found that 7.91% of former smokers suffered strokes, 5.32% of smokers suffered strokes, 4.76% of people who never smoked suffered strokes, and 3.04% of people with an unknown smoking status suffered strokes. So, we see that people who smoked or currently smoke do suffer strokes at a higher rate.

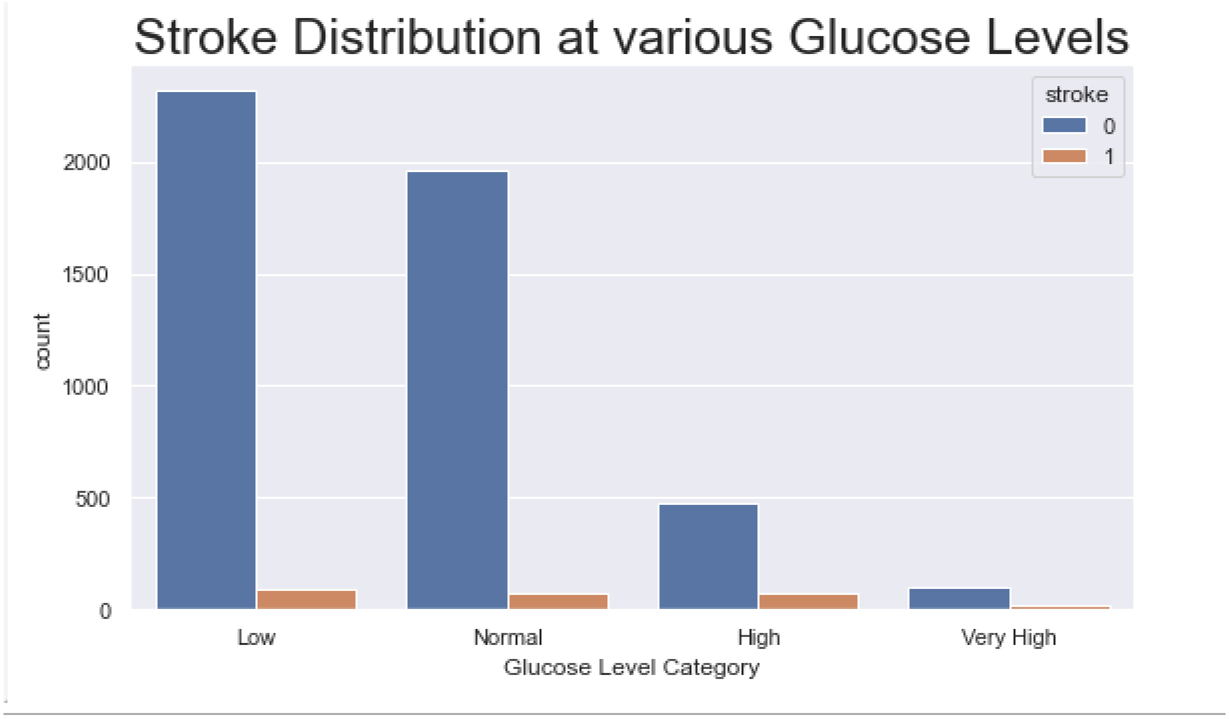


### Visualization for BMI, Age, and Average Glucose Level

We created bins to form groups for bmi, age, and avg_gluclose_level. We found that for bmi as people go up in weight into higher categories, they suffer more strokes. Similarly, as people get older, they also suffer more strokes.

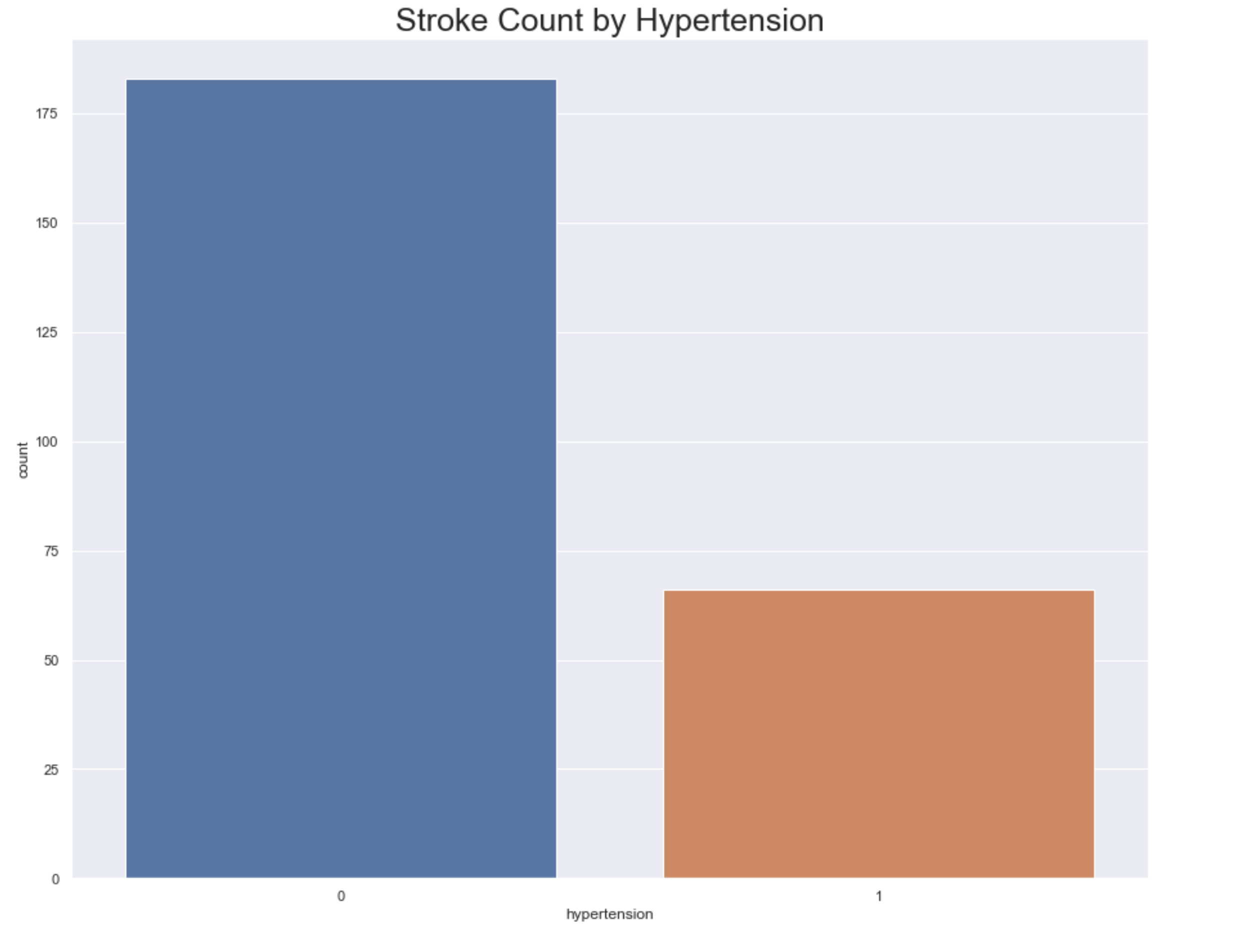





Stroke Distribution at various Glucose Levels
stroke
0
1
count
2000
1500
1000
500
0
Low
Normal
High
Very High
Glucose Level Category

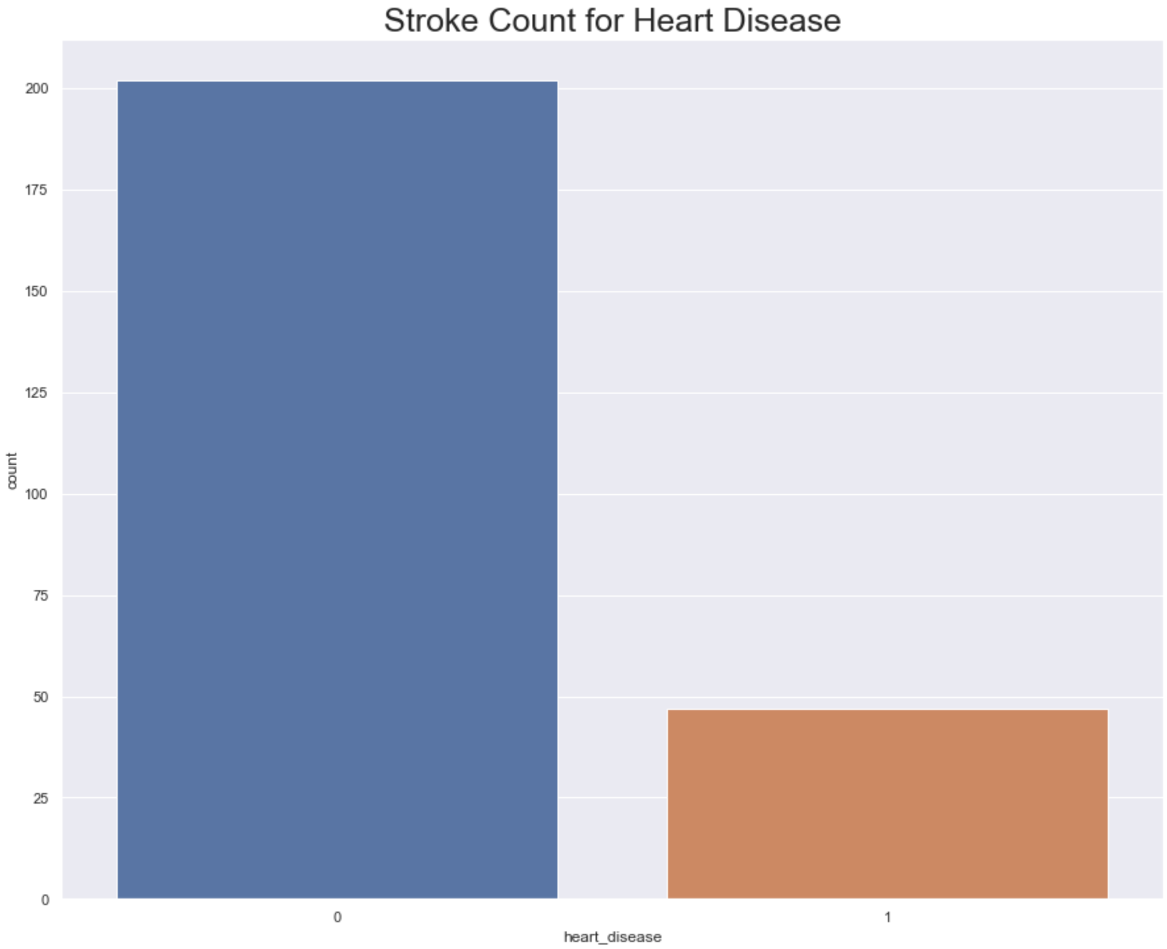
## Visualization for Hypertension

he count plot shows the count of strokes between no hypertension and hypertension. However, we found that 3.97% of people who did not suffer from hypertension suffered strokes. Meanwhile, 13.25% of people who suffered from hypertension also suffered strokes. The count plot is deceiving because people who hypertension suffer strokes at a much higher rate.



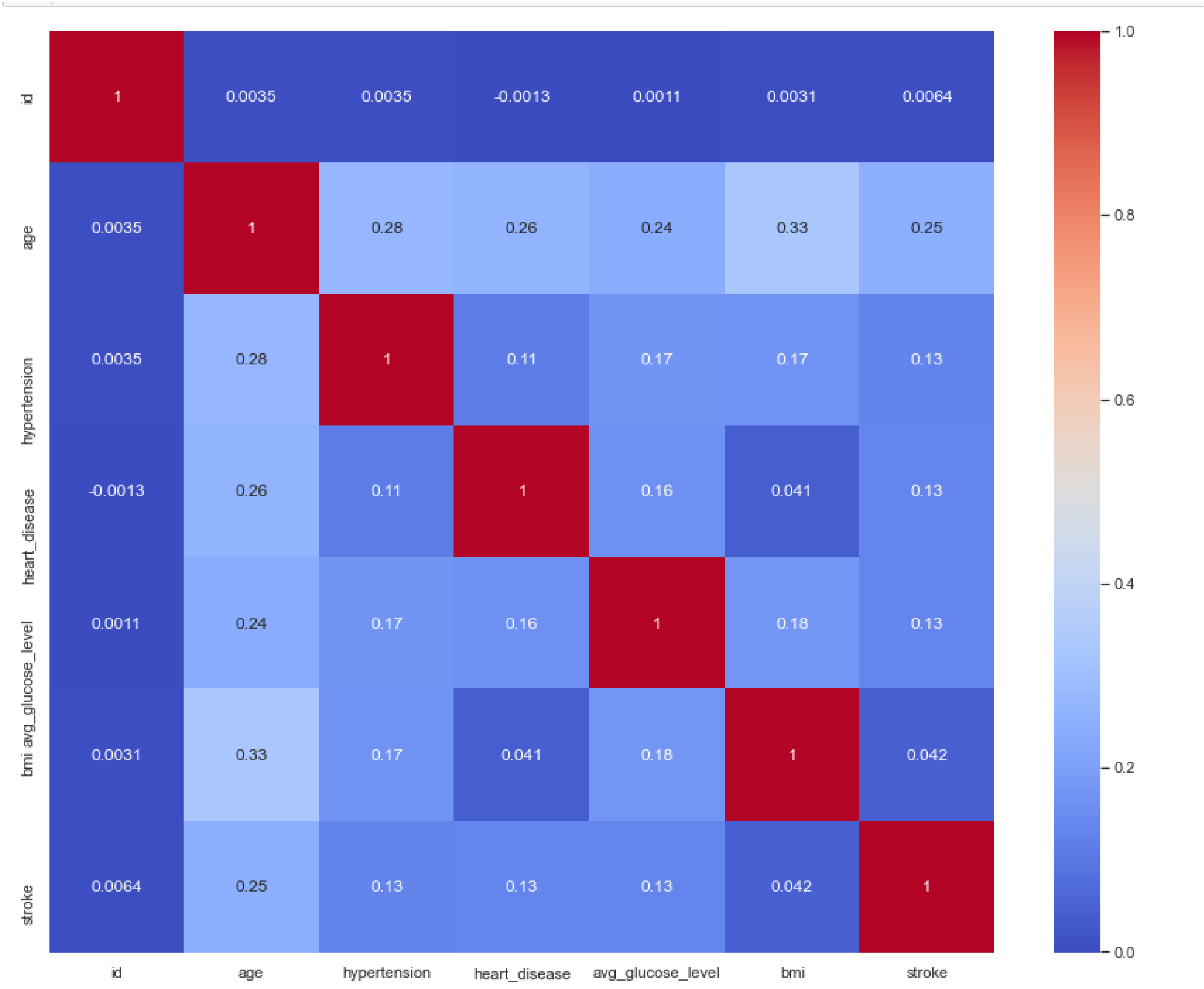
## Visualization for Heart Disease

This count plot shows that most people who suffered strokes had no history of heart disease. However, we want to find the proportion of people with heart disease or no heart disease that suffer strokes. We found that there were 4834 people who did not have a history of heart disease in the data set. While there were 276 people who did have some form of heart disease. We found that 4.18% of people with no heart disease suffered strokes. Meanwhile, 17.03% of people with a history of heart disease suffered strokes. While heart disease is a rare case in the dataset, people with heart disease suffer strokes at a much higher rate.



## Correlation Heatmap

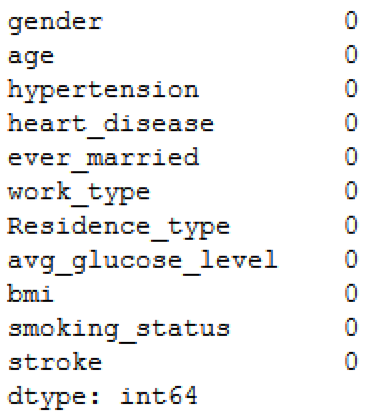
Below we see the correlation heatmap for our dataset. However, you will notice that we are missing the categorical variables. We need to process our data to get the dataset in an acceptable form and ready for modeling.



# Data Processing

## Filling in Missing Values

The first thing we did was account for the 201 missing values for the bmi variable. Looking at the summary statistics we found that mean and median for bmi were 28.89 and 28.1, respectively. We chose to use the median to fill in all the missing values in the dataset. After we checked the sum of missing values and found that there were no missing values in our dataset.



```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

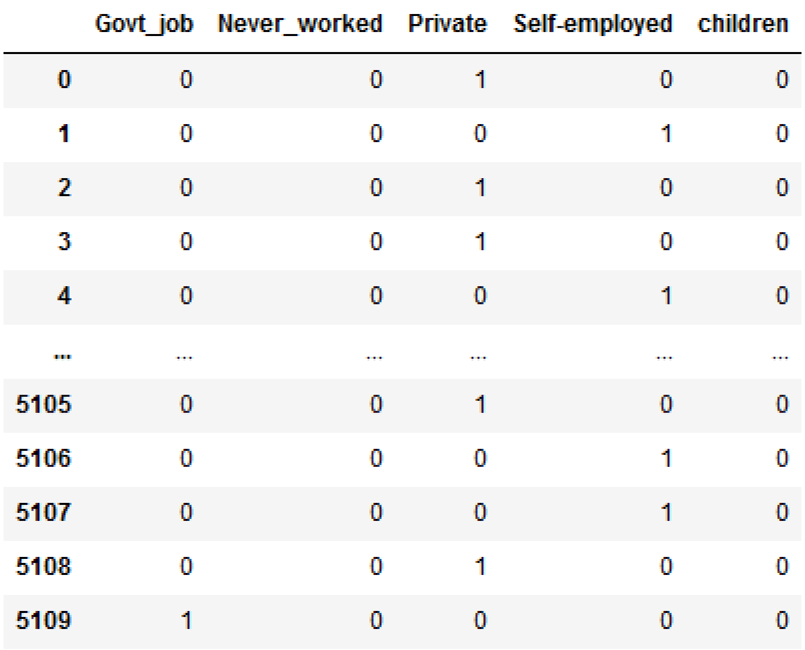
## Encoding Variables

### Label Encoding

We found that the variables with 2 unique levels and encoded them using binary. We found Residence_type and ever_married were the only two variables with 2 unique levels. For Residence_type, we encoded Urban as 0 and Rural as 1. For ever_married, we encoded No as a 0 and Yes as a 1.

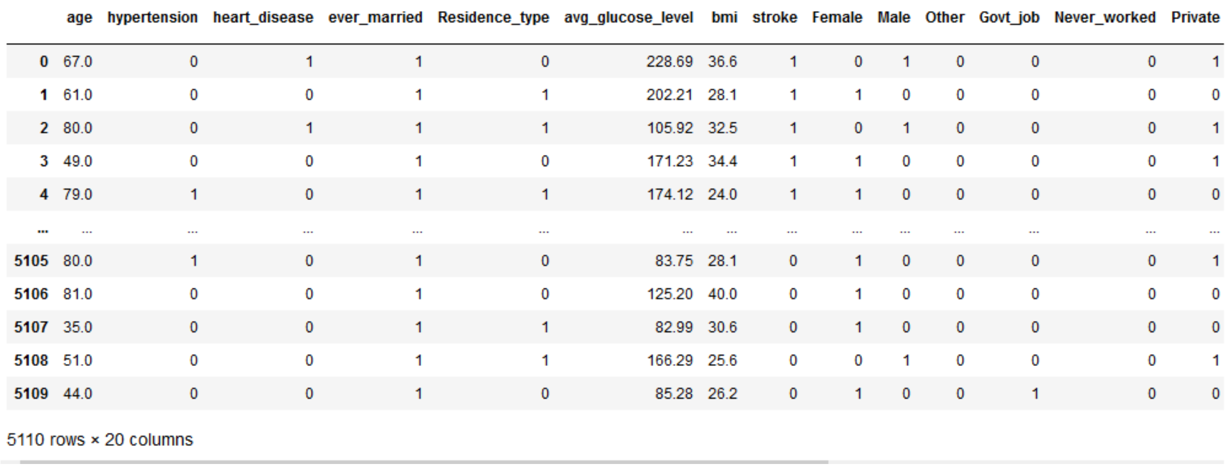
### One Hot Encoder

We used the One Hot Encoder package to encode variables with more than 2 unique levels. We chose to do this because we did not want our models to assume some form of hierarchy. We did not want models to interpret higher levels as being more important and assigning them more value because of it. Gender, smoking_status, and work_type are the variables with more than 2 unique levels. Let us look at work_type as an example. Work_type has 5 unique levels: Private, Self-employed, Gov't_job, children, and Never_worked. We create dummy variables for every level within the work_type. We add all these variables as new columns into our dataset and remove the old work_type variable from the dataset. We do the same for every variable with more than 2 unique levels.



| | Govt_job | Never_worked | Private | Self-employed | children |
|---|---|---|---|---|---|
| **0** | 0 | 0 | 1 | 0 | 0 |
| **1** | 0 | 0 | 0 | 1 | 0 |
| **2** | 0 | 0 | 1 | 0 | 0 |
| **3** | 0 | 0 | 1 | 0 | 0 |
| **4** | 0 | 0 | 0 | 1 | 0 |
| **...** | ... | ... | ... | ... | ... |
| **5105** | 0 | 0 | 1 | 0 | 0 |
| **5106** | 0 | 0 | 0 | 1 | 0 |
| **5107** | 0 | 0 | 0 | 1 | 0 |
| **5108** | 0 | 0 | 1 | 0 | 0 |
| **5109** | 1 | 0 | 0 | 0 | 0 |

Our new data set with all the new dummy variables and all the old categorical variables removed has 20 columns instead of the original 12.



| | age | hypertension | heart_disease | ever_married | Residence_type | avg_glucose_level | bmi | stroke | Female | Male | Other | Govt_job | Never_worked | Private |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 67.0 | 0 | 1 | 1 | 0 | 228.69 | 36.6 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **1** | 61.0 | 0 | 0 | 1 | 1 | 202.21 | 28.1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **2** | 80.0 | 0 | 1 | 1 | 1 | 105.92 | 32.5 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **3** | 49.0 | 0 | 0 | 1 | 0 | 171.23 | 34.4 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| **4** | 79.0 | 1 | 0 | 1 | 1 | 174.12 | 24.0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5105** | 80.0 | 1 | 0 | 1 | 0 | 83.75 | 28.1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| **5106** | 81.0 | 0 | 0 | 1 | 0 | 125.20 | 40.0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **5107** | 35.0 | 0 | 0 | 1 | 1 | 82.99 | 30.6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **5108** | 51.0 | 0 | 0 | 1 | 1 | 166.29 | 25.6 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **5109** | 44.0 | 0 | 0 | 1 | 0 | 85.28 | 26.2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |

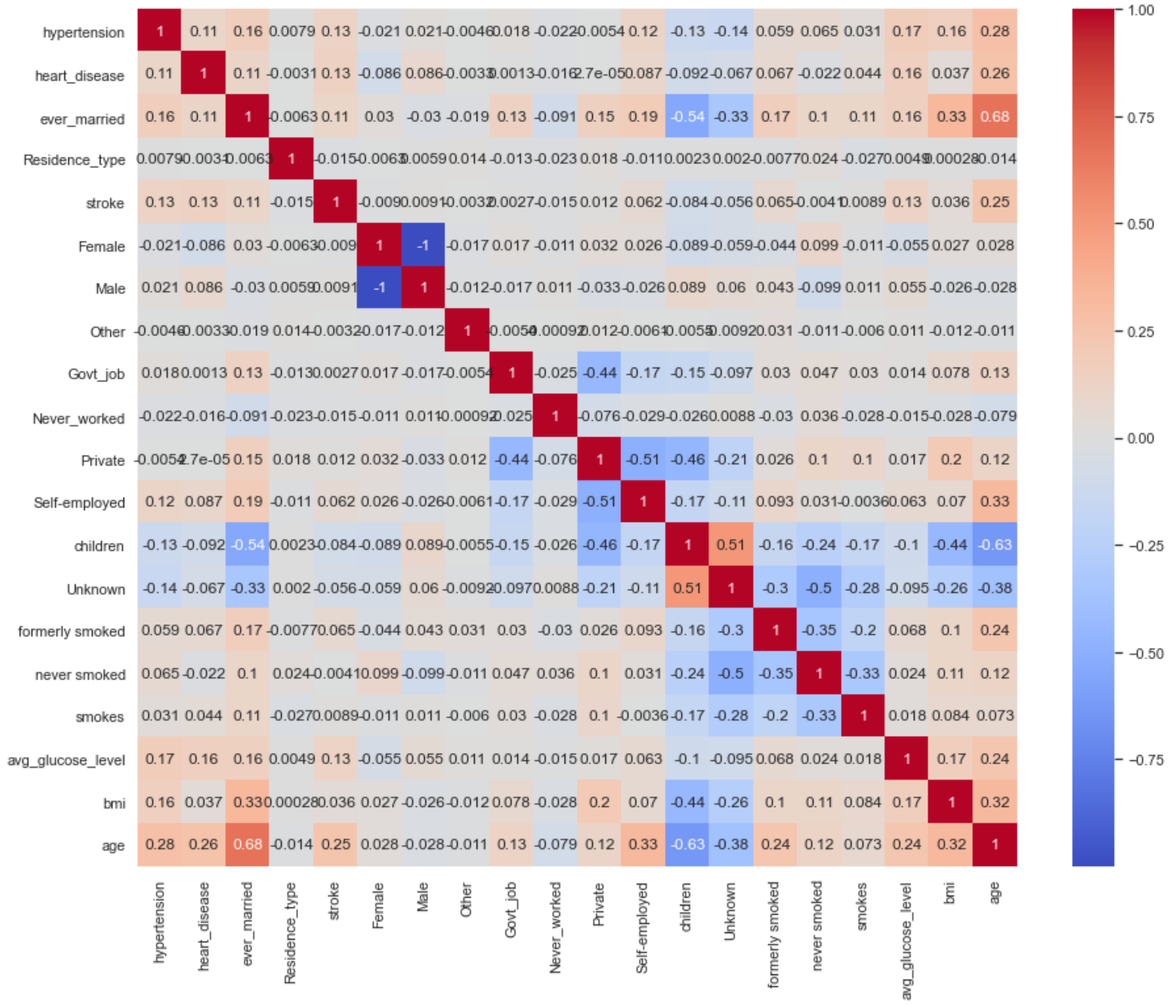
5110 rows × 20 columns

## Scaling Variables

We also used the Standard Scaler package to scale the bmi, age, and avg_gluclose_levels. We replaced the unscaled age, bmi, and avg_gluclose_level variables with their scaled counterparts.

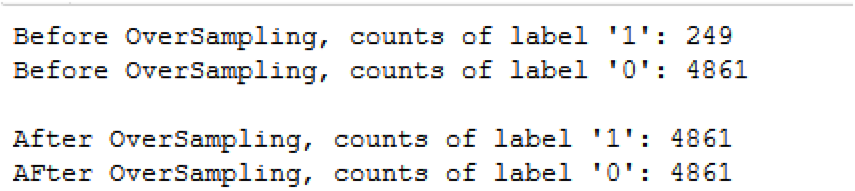
## Updated Correlation Heatmap

Below you will see the updated correlation heatmap using our final dataset. You will find it includes all our variables. We see that age is the variable with the highest correlations among all variables. We expect age to play a major factor in our model.



### Over Sampling using SMOTE

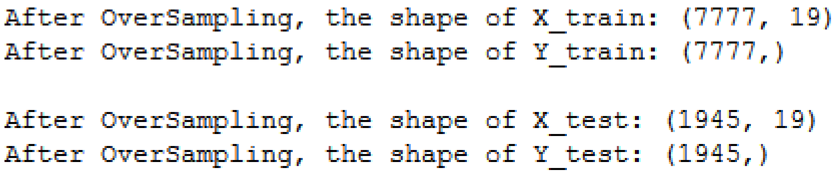
We used the SMOTE package to oversample the number of strokes in our dataset. Since, the dataset is so imbalanced and the feature of interest is so underrepresented, we chose to oversample the number of strokes to achieve a balanced dataset. Below you will see we printed the number of strokes and no strokes before and after sampling. As you see, we were successful in balancing our dataset. It is not ready to be split into test and training sets and ready for modeling.



```
Before OverSampling, counts of label '1': 249
Before OverSampling, counts of label '0': 4861

After OverSampling, counts of label '1': 4861
AFter OverSampling, counts of label '0': 4861
```

### Splitting Train and Test sets

We chose to use a 80% to 20% split for training and test sets to run our models. Below you will see the number observations for our X and Y train and test sets.

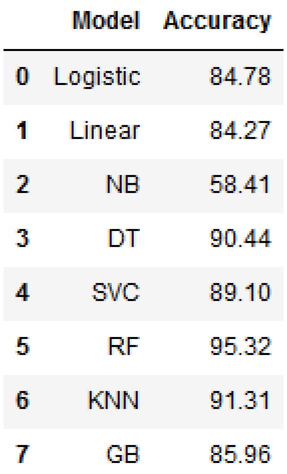


```
After OverSampling, the shape of X_train: (7777, 19)
After OverSampling, the shape of Y_train: (7777,)

After OverSampling, the shape of X_test: (1945, 19)
After OverSampling, the shape of Y_test: (1945,)
```

## Modeling

For our models we chose to use 8 different models. We chose to use Logistic Regression, Linear Regression, Naïve-Bayes, Decision Trees, Support Vector Classifier, K-Nearest Neighbors, Gradient Boosting, and Random Forest. To the right is a chart summarizing the accuracy scores of all the models we ran.

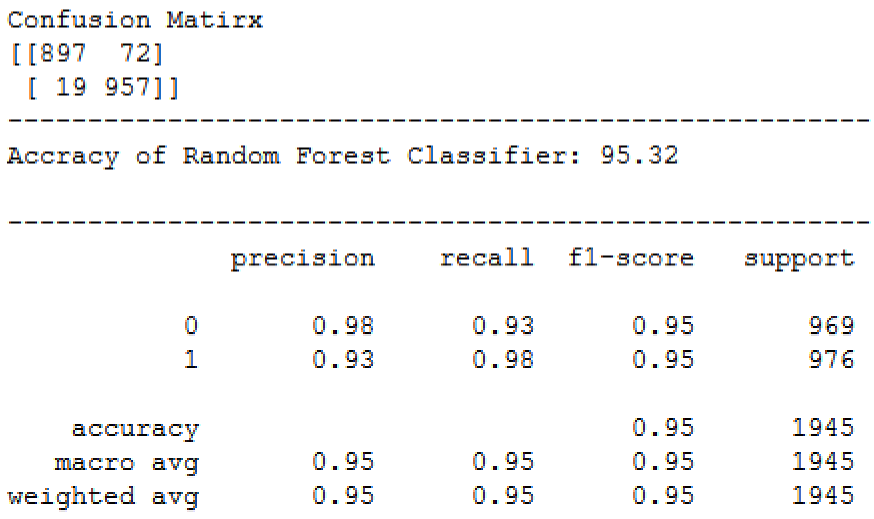
All our models delivered decent accuracy scores. We even had some models achieve higher than a 90% accuracy score. Overall, we found that the Random Forest model provided the highest accuracy score of about 95%.



| | Model | Accuracy |
|---|---|---|
| 0 | Logistic | 84.78 |
| 1 | Linear | 84.27 |
| 2 | NB | 58.41 |
| 3 | DT | 90.44 |
| 4 | SVC | 89.10 |
| 5 | RF | 95.32 |
| 6 | KNN | 91.31 |
| 7 | GB | 85.96 |

## Random Forest Model

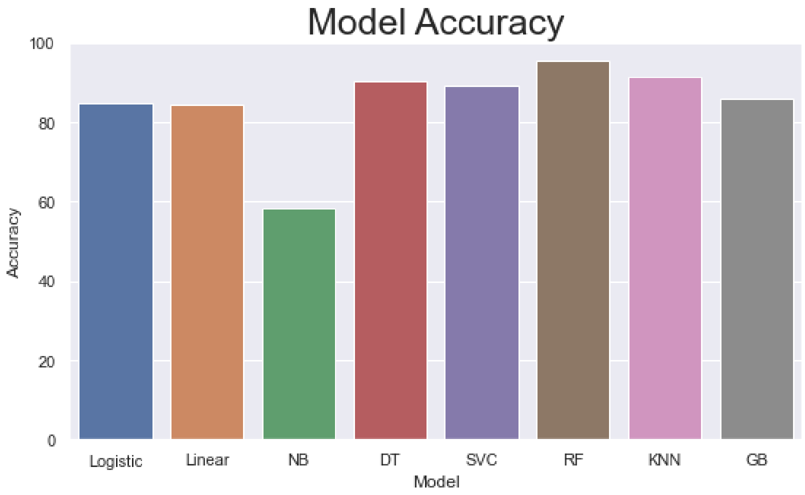
Taking a closer look at the Random Forest model, we see that the model has great precision and recall scores in addition to the accuracy score. Below is a summary of our random forest including a confusion matrix.



```
Confusion Matirx
[[897  72]
 [ 19 957]]
-----------------------------------------------------
Accracy of Random Forest Classifier: 95.32

-----------------------------------------------------
              precision    recall  f1-score   support

           0       0.98      0.93      0.95       969
           1       0.93      0.98      0.95       976

    accuracy                           0.95      1945
   macro avg       0.95      0.95      0.95      1945
weighted avg       0.95      0.95      0.95      1945
```

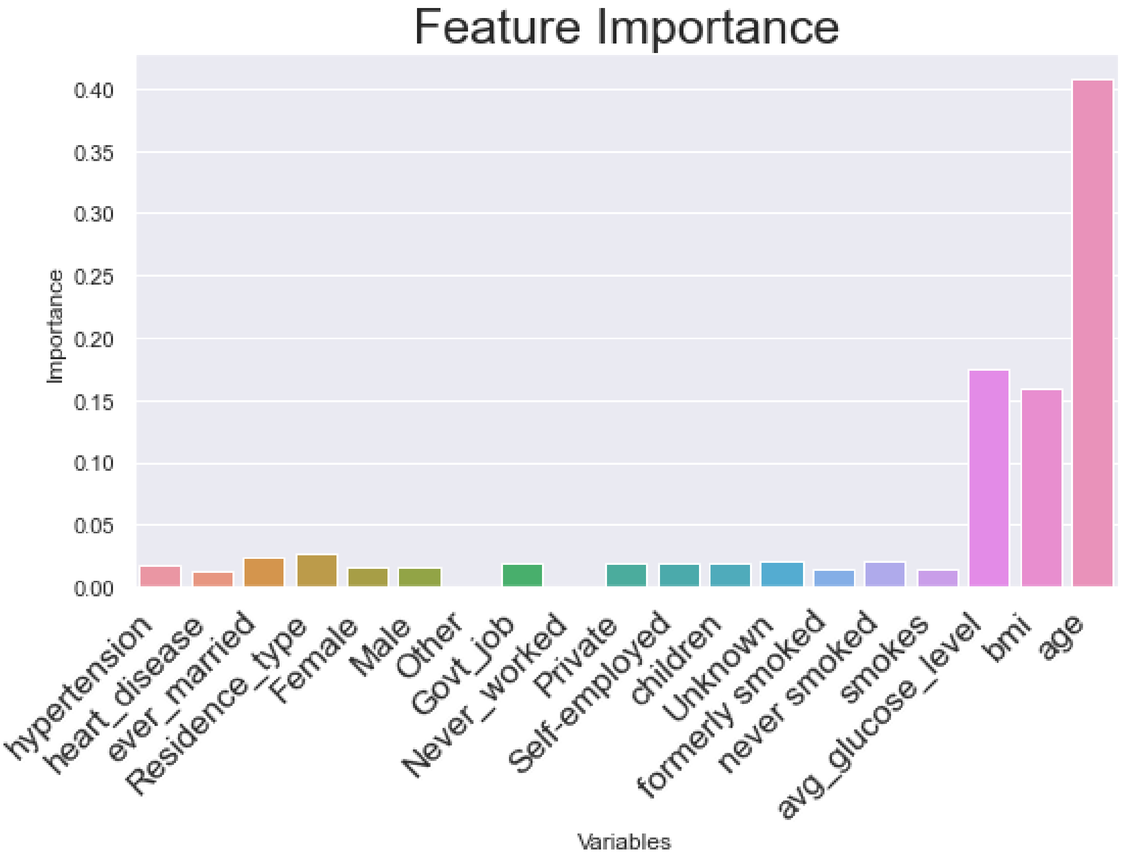
## Model Accuracy

After running various models, we found that the Random Forest models gives us the highest accuracy score among all models tested. It also had the highest F-score and Recall among all models for both instances of stroke and no stroke. Recall is the ratio of correctly predicted positive observations. In other words, of all the patients who suffered a stroke, how many did we correctly identify. Similarly, of all patients who did not suffer a stroke, how many did we label correctly. Below is a graph plotting the accuracy score of each model.



### Feature Importance

We also used the Feature Importance function to find out which of the variables in our model had the highest predictive power. The function assigns a score to each variable for its importance in prediction strokes in the data set. The higher the score the more important a variable is for prediction. Below is a graph showing each variable's feature importance. As you can see Age is by far the most important feature in predicting strokes. We also see that bmi and avg_gluclose_level are also very important in predicting strokes.



## Conclusion

In conclusion, we were able to build a random forest model with about 95% prediction accuracy. We found that age, avg_glucose_level, and bmi are the most important features for prediction. The dataset was highly imbalanced with a split of 5% for stroke and 95% for no stroke. Oversampling was used in overcoming this imbalance. The random forest model is the model with the highest prediction accuracy when compared to all the other models we tested. In addition, the Random Forest model had great Recall and Precision scores. In the end, we were able to show that strokes are more likely to occur as we age. This is something we expected, and it further goes to show the importance of living a healthy lifestyle as we get older.