**Introduction**

A stroke is a life-threatening condition that occurs when the blood vessel carrying nutrients to the brain is blocked or ruptured. According to the WHO, strokes are the second leading cause of death. In addition, there is a 50% chance that the survivors of a stroke become disabled.

However, according to the CDC, 80% of strokes are preventable. The purpose of this project is to identify whether an individual is likely to get a stroke based on factors such as age, gender, pre-existing conditions, lifestyle, habits, family history, etc. Our analysis can assist the medical field in determining which factors and health conditions can more likely increase your chance of getting a stroke. This way, medical professionals can advise patients with strategies to target specific factors to decrease one’s chance of a stroke.

The project is subdivided into the following sections:

1. Obtaining and review the raw data
2. Data processing
3. Determining the correlation between variables
4. Approach
5. Exploratory data analysis
6. Supervised learning
7. Conclusion

**Data**

The data used is “Stroke Prediction Dataset” by Fedesoriano from Kaggle. The source of the data has been made confidential by the sharing party.

Dataset URL - <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv>

It contains the following columns:

1) ID: This is a unique identifier for the individual

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

3) age: Age of the individual

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

5) heart\_disease: 0 if the individual does not have any heart diseases, 1 if the individual has a heart disease

6) ever\_married: "No" or "Yes"

7) work\_type: The type of work the individual is involved in. It is "children" for kids who are not working, "Govt\_jov" for individuals in government jobs, "Never\_worked" for individuals who have never worked, "Private" for employees in the private sector, or "Self-employed".

8) Residence\_type: "Rural" or "Urban"

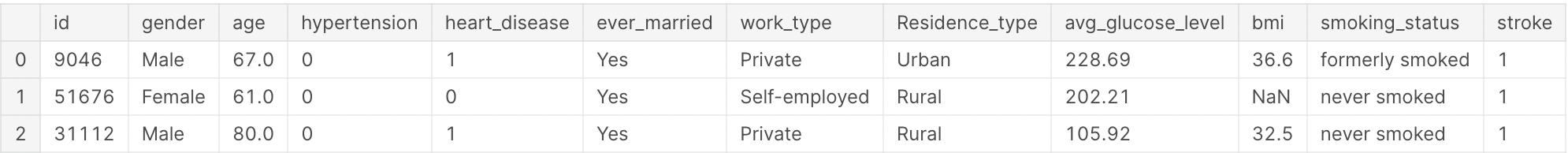
9) avg\_glucose\_level: Average glucose level in blood

10) bmi: Body mass index

11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown" (if this information for individuals is not available)

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

Figure 1: An overview of the header of our data

****

**Data Processing**

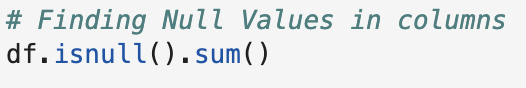
Step 1: Dropping the “Other” gender

Dropping the “Other” gender is needed because there was only one row in the dataset with a gender of “Other”.



Step 2: Finding the Null values in the BMI column

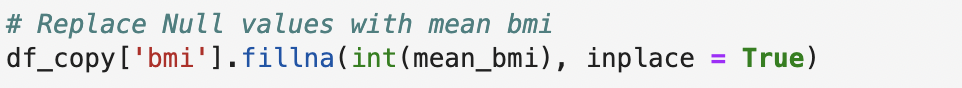
All Null values in the dataset were in the BMI column.



Step 3: Finding the average BMI in order to replace Null BMI values

Since we did not want to delete or remove any rows with a null BMI, we chose to replace the Null values with the average BMI value



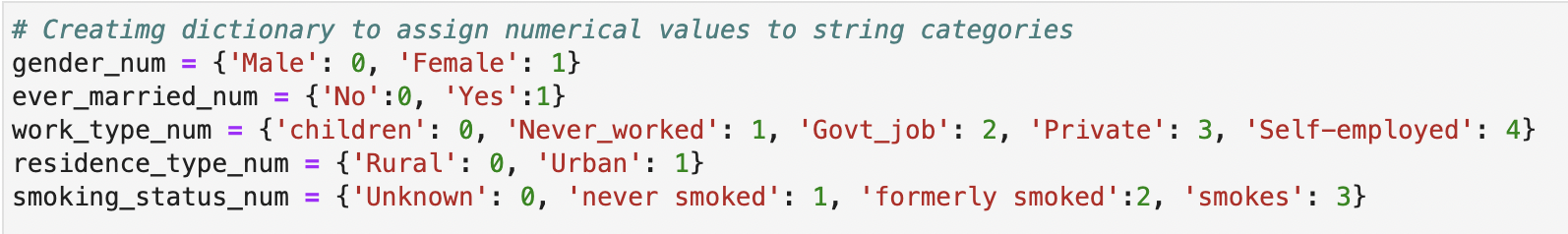




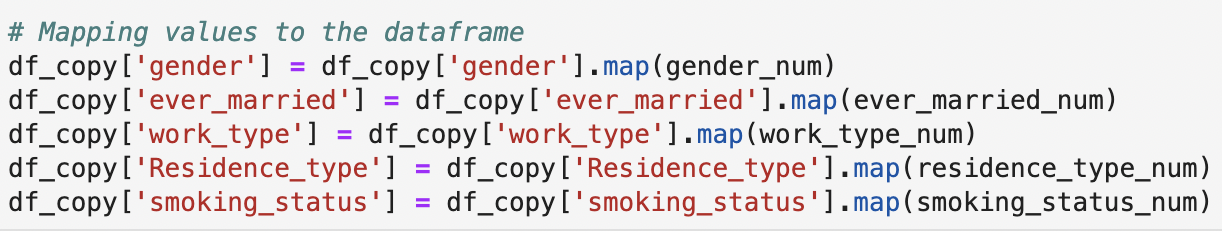
Step 4: Copy the data frame



Step 5: Assign numerical values to string categories



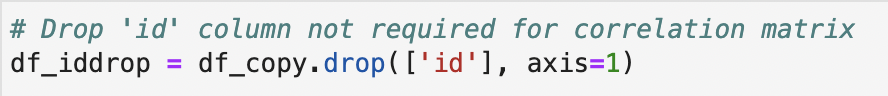
Step 6: Mapping values into the data frame



**Correlation Between Variables**

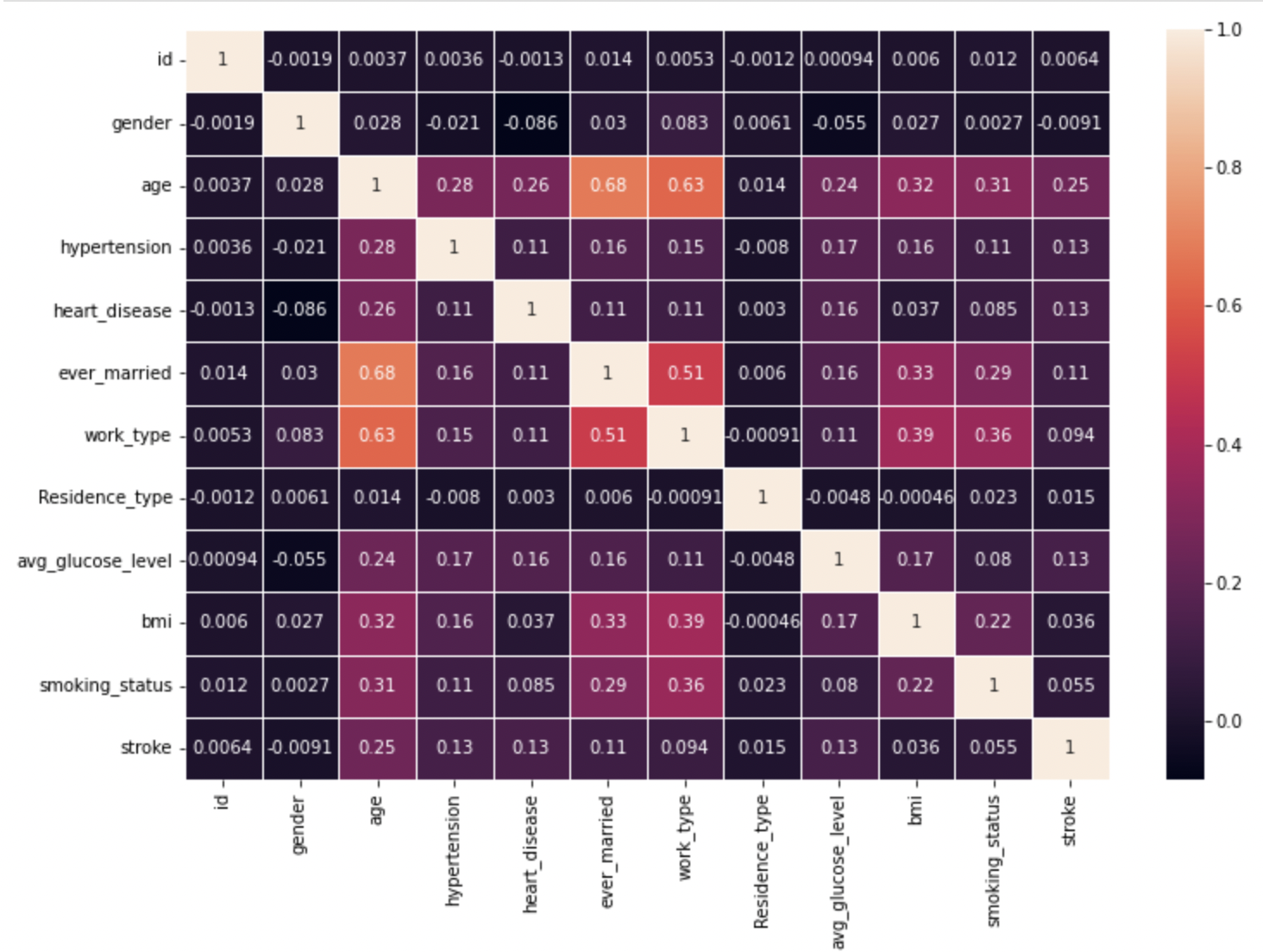
Step 1: Dropping “ID” column

As the ID column is not required for building a correlation matrix, we chose to drop the column for the correlation matrix.



Step 2: Creating the Correlation Matrix

Figure 2: Heat map correlation matrix for each variable within the dataset

****

**Approach**

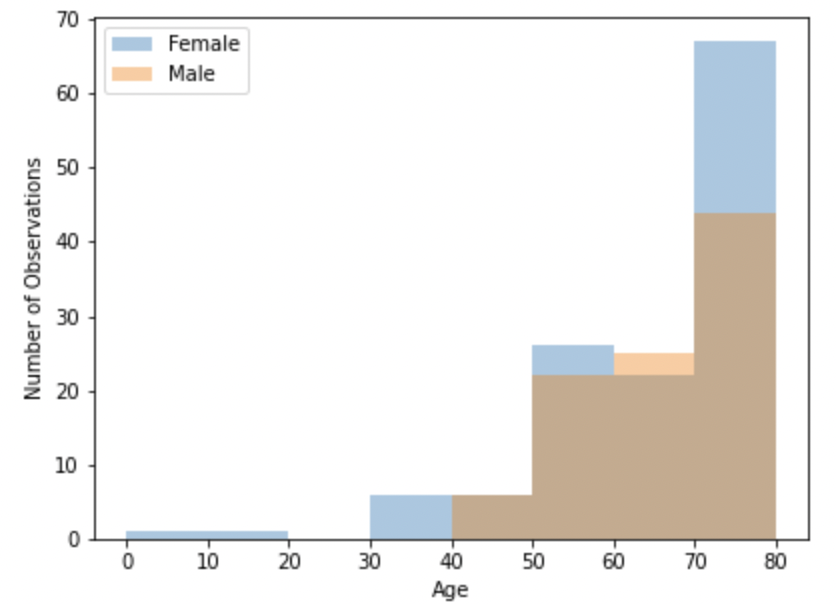
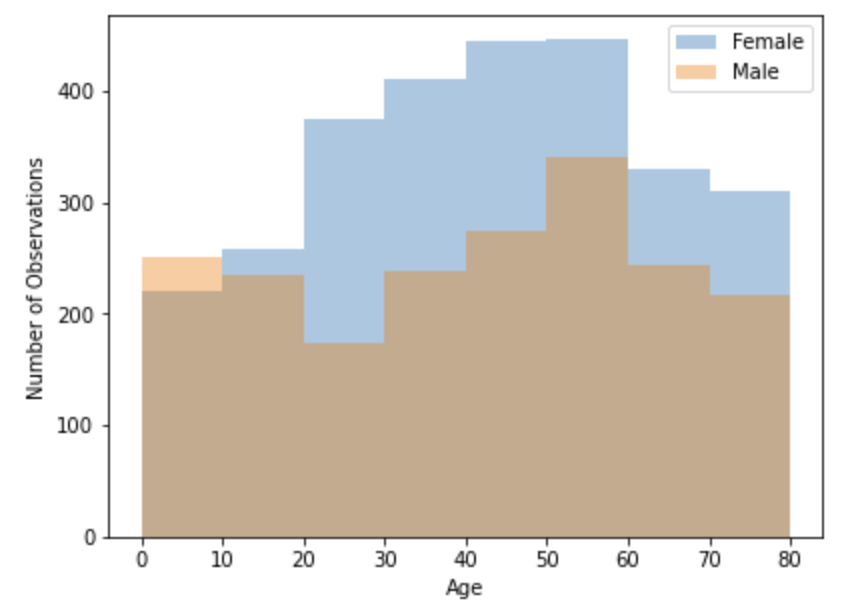
Identification of the independent variables used to predict the occurrence of stroke (dependent variable). The selection of independent variables to be used in the regression model are based on the correlation matrix. We chose to focus on the following variables:

1. Gender - Male or Female
2. Age - Age of the individual in the population
3. Heart disease - Whether or not the patient has a history of heart disease (0 or 1)
4. Work type - The type of employment held by the individual. Categories are Government job, Private job, Self Employed, Never worked, Children
5. Average glucose level - Numerical data signifying the average glucose level in an individual
6. BMI - Body Mass Index; defined as body mass divided by square of body height

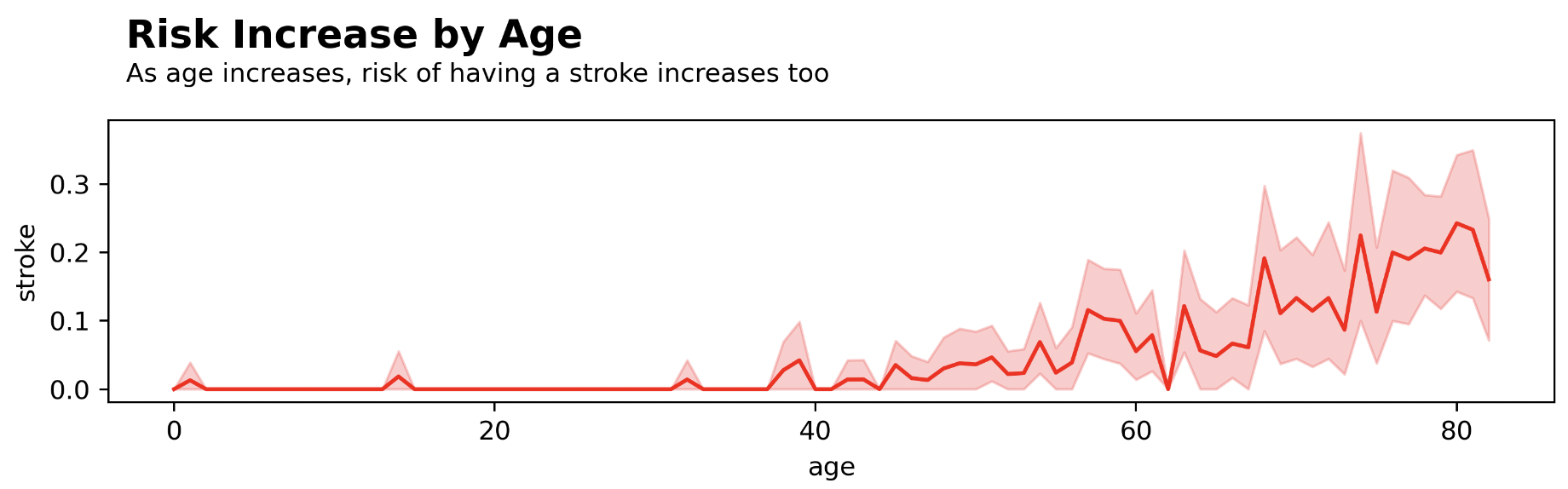
**Exploratory Data Analysis**

To identify which variables play the most part in the occurrence of a stroke, we analyzed each independent variable to see their significance. Based on the correlation matrix, we chose to focus on the following variables: age, gender, hypertension, heart disease, work type, residence type, average glucose level and BMI. The following figures shows the relationship between each variable and stroke.

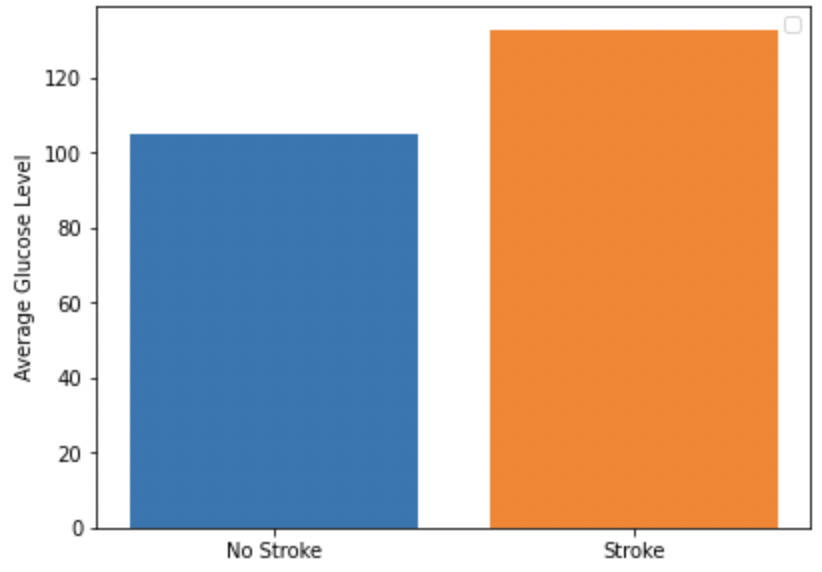
**Figure 3,4: Age vs. stroke and no-stroke**

****

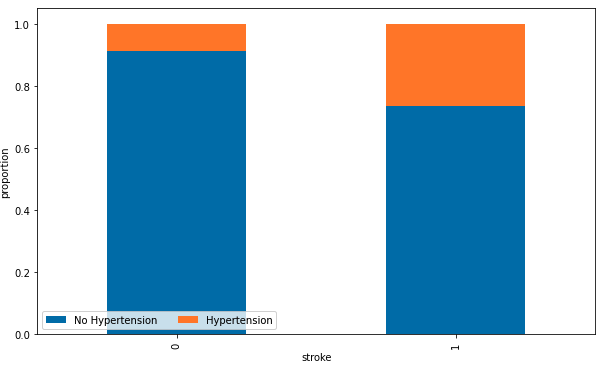
**Figure 5: Percentage of population who have had strokes vs. age**

****

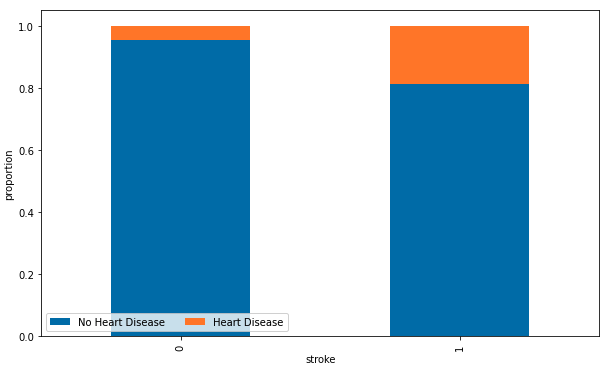
**Figure 6: Number of stroke and no-stroke patients vs. average glucose level**

****

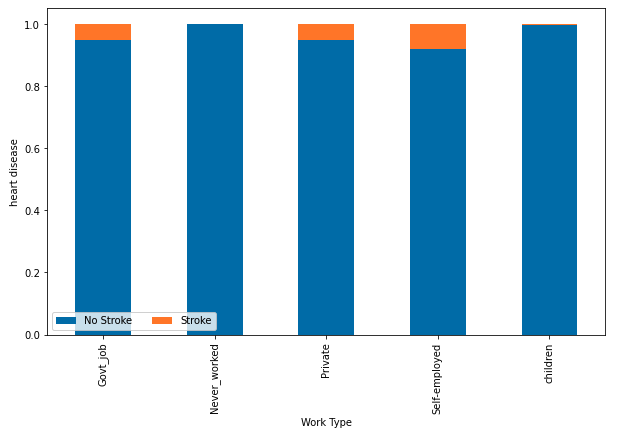
**Figure 7: Percentage of stroke and no-stroke patients vs. hypertension**

****

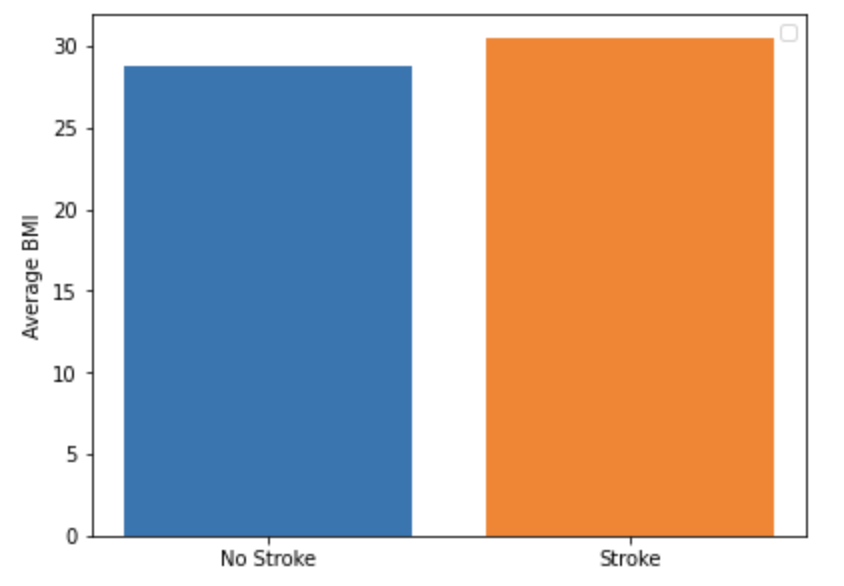
**Figure 8: Percentage of stroke patients vs. heart disease**

****

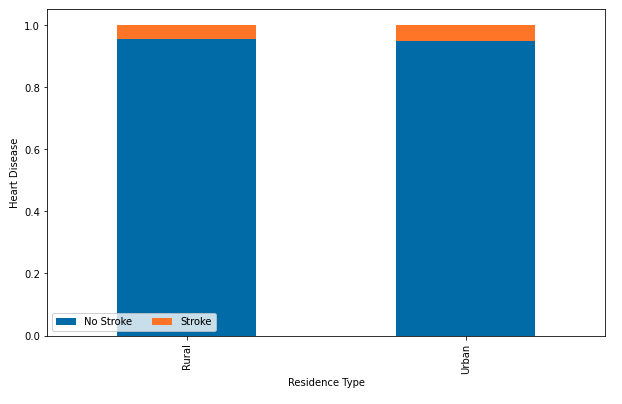
**Figure 9: Percentage of stroke and no-stroke patients vs. work type**

****

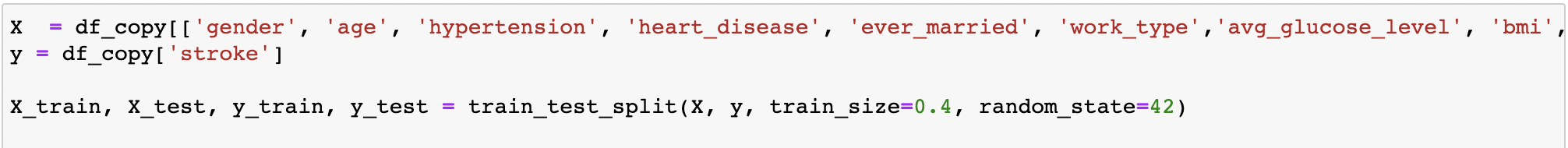
**Figure 10: Number of stroke and no-stroke patients vs. BMI**

****

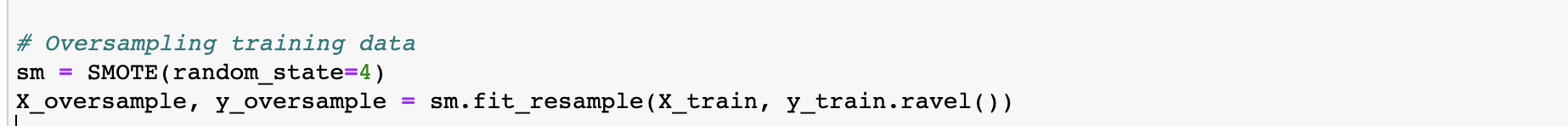
**Figure 11: Percentage of stroke and no-stroke patients vs. resident type**

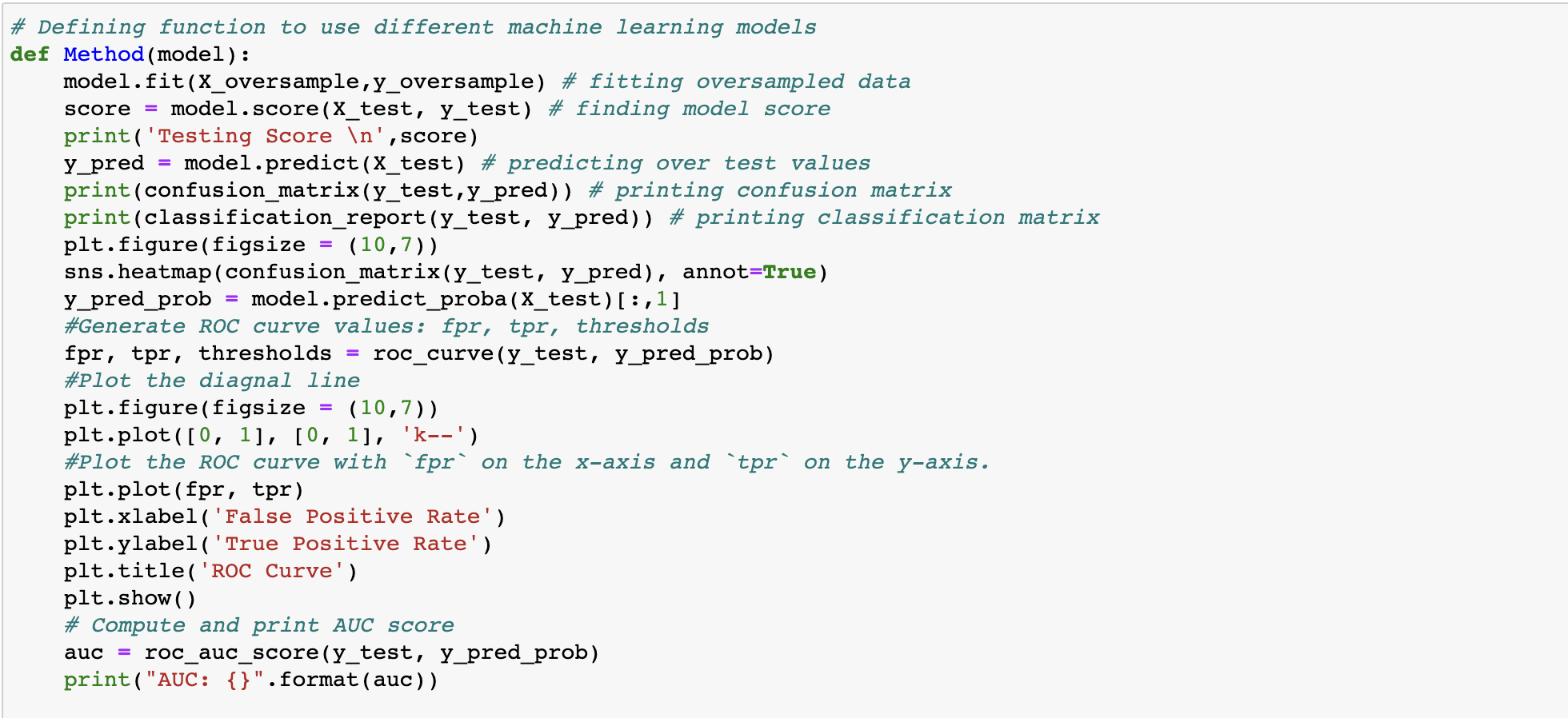
****

**Supervised Learning**

Step 1: After choosing the relevant independent variables, we separate the training and test data in a 60/40 split. 

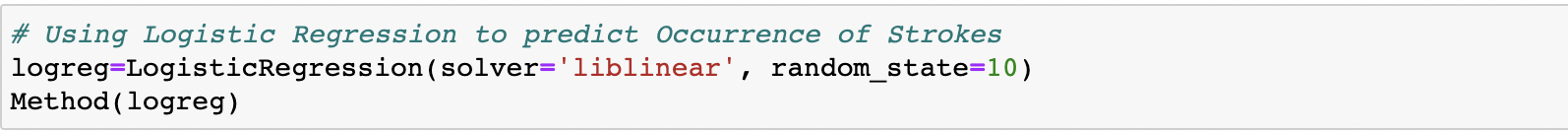
Step 2: We noticed an imbalance in our data for the number of patients who had stroke compared with the people who have not had a stroke, making the number of patients who have had a stroke the minority class. The issue with classifying with imbalanced data is that it can be ineffective in learning the decision boundary. One method to solve this problem is to oversample instances from the minority class.

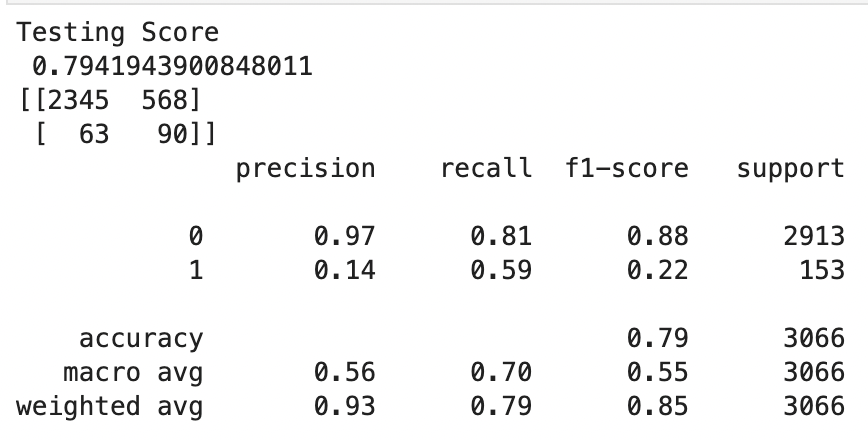
We implement Synthetic Minority Oversampling Technique (SMOTE) of oversampling for our dataset. This method synthesizes new instances from the minority class. SMOTE works by selecting a random instance from the minority class and the k of the nearest neighbors is found for that instance. Next, a randomly selected neighbor is chosen and a synthetic instance is created at some randomly selected point along the ‘line’ between these two instances.

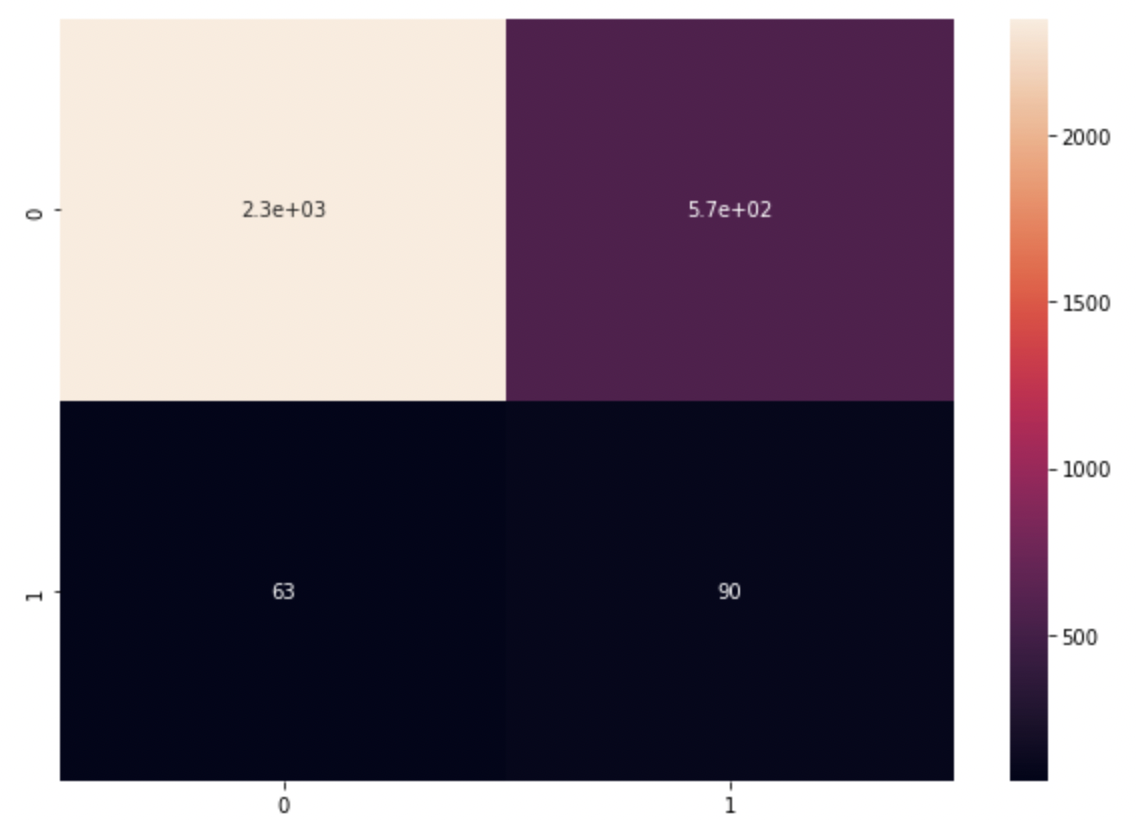
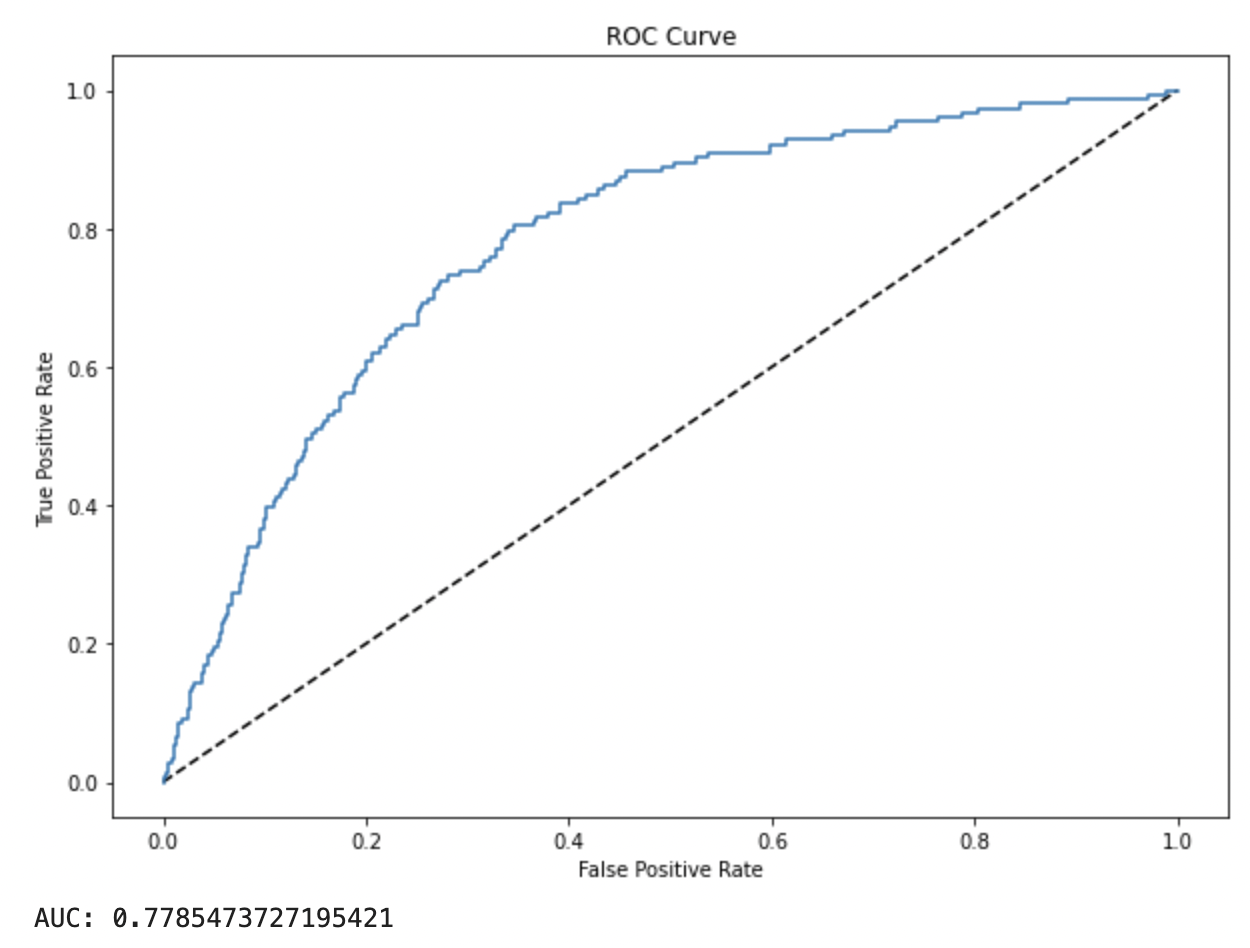
Step 3: Next, we define a function so that we can input our supervised learning models. Our function, ‘Method’, fits and predicts the data, calculates the model score, prints the confusion matrix and classification report, and generates the ROC curve for the model.****

Step 4: We use Logistic Regression to predict the occurrence of strokes.

In addition to Logistic Regression, we also used k-NN and Decision Tree but finally decided to use Logistic Regression as it had the best F1 score for stroke occurrences of 22% (F1 score for stroke occurrence using k-NN was 20% and using Decision tree was 12%)



**Figure 12: logistic regression including: testing score, ROC curve, temperature proportion**



**Conclusion**

Through our analysis of the data, we conclude that age, glucose level, hypertension and heart disease are significant factors in determining the future occurrence of a stroke.

We conclude that BMI and residence type are not significant factors in determining the future occurrence of a stroke. Work type is also not a significant factor as it is very highly correlated with age. Children are less likely to get strokes not because of their lack of occupation, but more likely because they are young. Those who work in the private sector, in government or are self employed are more likely to get strokes less because of the stress of their occupation, but more likely due to their older age.

Out of all patients who suffered from a stroke, we can predict 14% of them accurately. Accuracy can be improved if we have a larger data set. We also believe that the more data and different factors we have the more accurate our model could be.