Stroke Prediction

An Exercise in Machine Learning and Stroke Probability Predictions

Yakov Krasnikov

Charles Philpott

Troy Youngblood

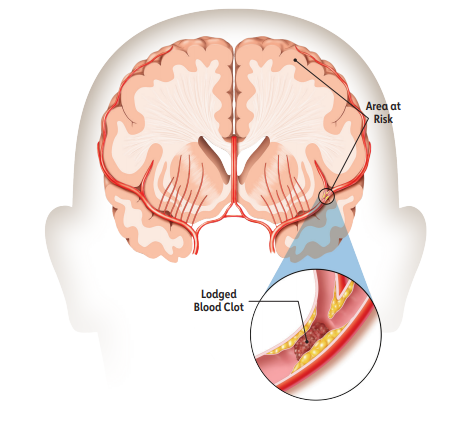
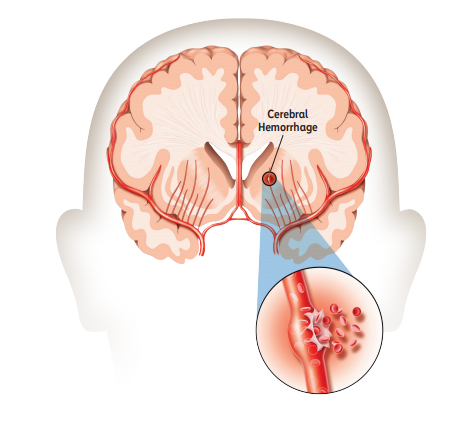
**Introduction**

The objective of this activity is to develop a preliminary screening tool which can be used to identify the likelihood of an individual having a stroke based on general contributing attributes. Data [1] from Kaggle was used as the basis for the predictive model.

Cerebrovascular accidents (strokes) in 2020 were the 5th leading causeof death [2] in the United States.

A stroke occurs when the blood supply to a region of the brain is suddenly blocked or when a rupture occurs starving the brain cells of oxygen and nutrients. Blockage obstructing the flow of blood to a region of the brain is called an ischemic stroke and accounts for 87% [3] of all strokes. The rupturing of a blood vessel is called a hemorrhagic stroke and accounts for 13% [4] of all strokes.

Ischemic Stroke Hemorrhagic Stroke

Source of Images [5]

The dataset used for the predictive model did not identify the type of stroke for each respective individual. To stay consistent with the dataset, the general word stroke will be used to describe the occurrence being predicted. A third category of stroke called a transient ischemic attack (TIA), or "mini stroke", caused by a temporary clot can also occur. The TIA has contributing factors similar to those of the ischemic and hemorrhagic stroke and is included in the general term stroke when identifying a potential outcome.

Per the American Stroke Association, 80% of strokes are preventable [6].

**Hypothesis**

By using data associated with stroke victims, a predictive model will be developed to identify the likelihood of a stroke. The hypothesis being tested is will the importance of the model features match the risk factors identified by the American Stroke Association.

Basis Risk Factors from American Stroke Association which can be controlled common to the dataset.

* High Blood Pressure
* Smoking
* Diabetes
* Obesity
* Heart Disease

Basis Risk Factors from American Stroke Association which cannot be controlled common to the dataset.

* Age
* Gender

**Data Source**

The data source used for the model prediction was from Kaggle [1].

The attributes with the dataset are:

* id: a unique identified for each set of information
* gender: “Male, “Female”, “Other”
* age: age of the patient
* hypertension: 0 assigned if hypertension not present, 1 if patient has hypertension
* heart\_disease: 0 assigned if heart disease not present, 1 if patient has 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”
* stroke: 0 if patient has not had a stroke, 1 if patient has had a stroke

**Data Review**

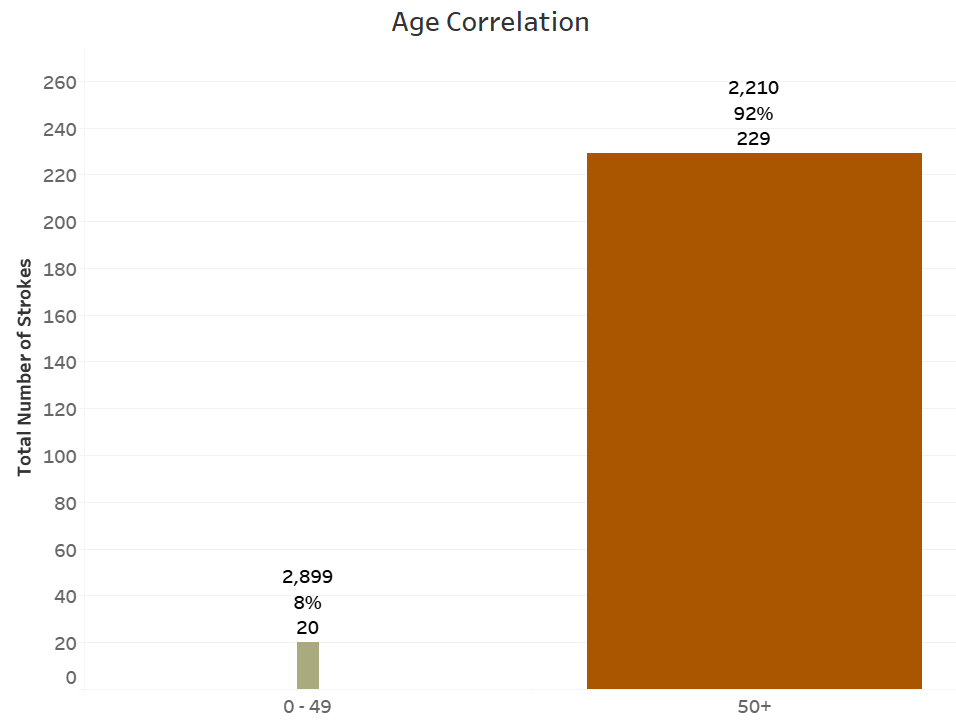
The raw dataset for machine learning consists of 5110 unique rows of information.

There were 2,994 (58.60%) “Females”, 2,115 (41.40%) “Males” and 1 “Other” in the gender attribute. The “Other” gender was dropped from the dataset for a resulting dataset of 5,109 unique rows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Review | | | | |
| Data Attribute | Female | | Male | |
| Count | Percent of gender | Count | Percent of gender |
| Had a stroke (Y) | 141 | 4.7 % | 108 | 5.1 % |
| Considered diabetic risk | 230 | 7.7 % | 204 | 9.6 % |
| Have heart disease (Y) | 113 | 3.8 % | 163 | 7.7 % |
| Have hypertension (Y) | 276 | 9.2 % | 222 | 10.5 % |
| Considered obese | 1,115 | 37.2 % | 805 | 38.1 % |
| Married (Y) | 2,001 | 66.8 % | 1,352 | 63.9 % |
| Live in Urban areas (Y) | 1,529 | 51.1 % | 1,067 | 50.4 % |
| Never smoked | 1,229 | 41.0 % | 663 | 31.3 % |
| Formerly smoked | 477 | 15.9 % | 407 | 19.2 % |
| Currently smoke | 452 | 15.1 % | 337 | 15.9 % |
| Unknown smoking status | 836 | 27.9 % | 708 | 33.5 % |
| Age: 0-19 | 480 | 16.0 % | 486 | 22.9 % |
| Age: 20-39 | 791 | 26.4 % | 412 | 19.4 % |
| Age: 40-49 | 450 | 15.0 % | 280 | 13.2 % |
| Age: 50-59 | 472 | 15.7 % | 362 | 17.1 % |
| Age: 60-69 | 352 | 11.8 % | 269 | 12.7 % |
| Age: 70-79 | 336 | 11.2 % | 233 | 11.0 % |
| Age: 80+ | 113 | 3.8 % | 73 | 3.4 % |

Trends identified in the dataset:

* 92% of strokes occur over the age of 50, The wider and darker brown bars indicate the higher normalized percentage.



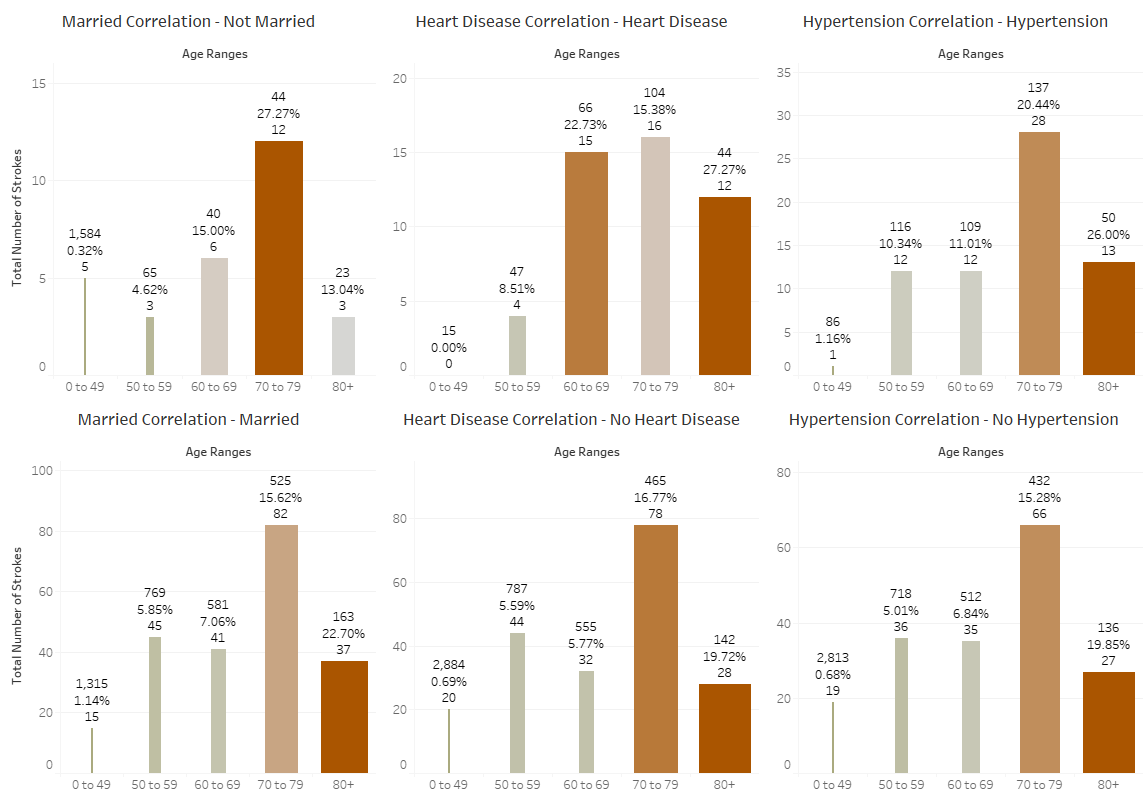
* Comorbidities Heart Disease and Hypertension, along with falling under Not Married generally have a higher percentage of strokes as age increase when the data is normalized for the specific age ranges.

The wider and darker brown bars indicate the higher normalized percentage.

The attributes are stacked for comparison purposes. The relatively equal or higher contributing factor, Not Married, Heart Disease and Hypertension are on top. The values associated with each bar in order are number of individuals in range, normalized percentage of that range and the number of individuals suffering a stroke in that range.

The lower Percentage of Not Married for Age Range 80+ could include individuals whose spouses have died and therefore where not married at the time of the stroke.

Age Range 80+ generally has the highest percentage of strokes on a normalized basis.



* Comorbidities BMI (Overweight and Obese), Glucose (Diabetic Risk and Diabetic) and Smoker Status (Formerly Smoked and Current Smoker) have the highest normalized stroke percentages. The wider and darker brown bars indicate the higher normalized percentage.

The BMI Categories are:

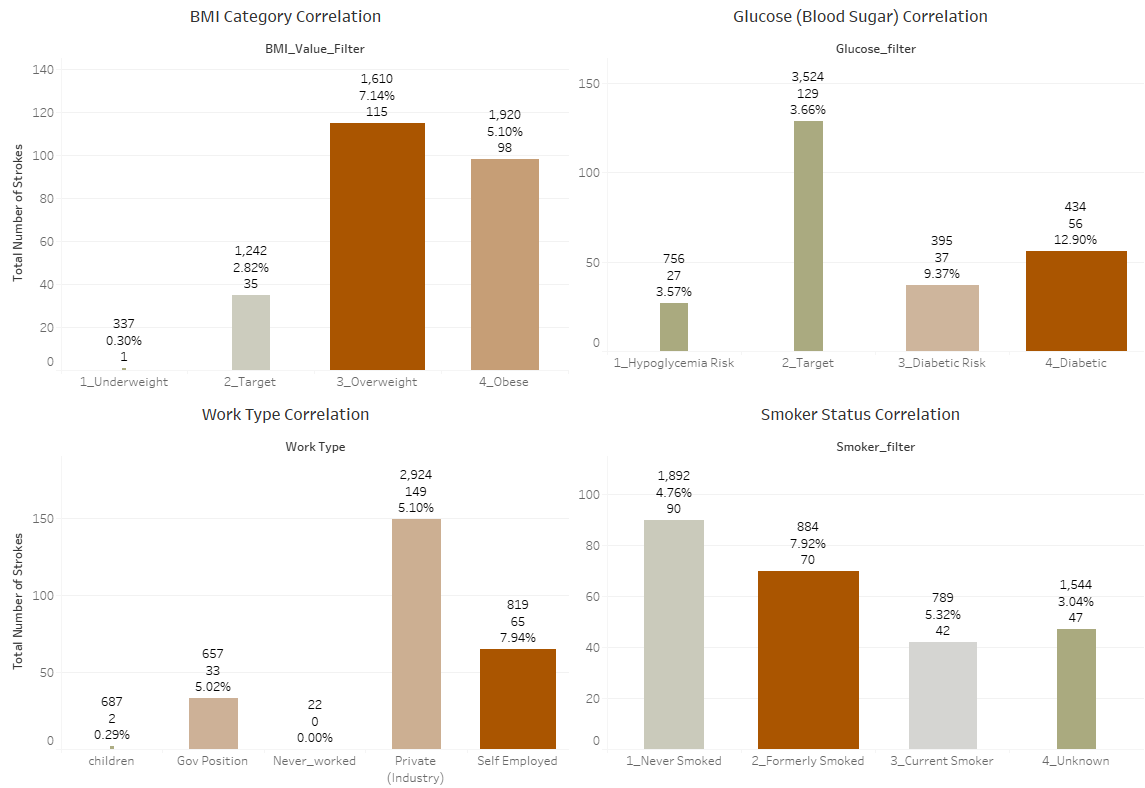
* + Underweight: BMI < 18.5
  + Target: 18.5 <= BMI < 25
  + Overweight: 25 <= BMI < 30
  + Obese: BMI > 30

The glucose data presented in the dataset is average glucose value. Depending on when a person has eaten, a glucose value will swing high and low. Therefore, to create a Glucose Category, it required a blend of Fasting, Just Eaten and Several Hours after eating ranges.

The Glucose Categories are:

* + Hypoglycemic: average glucose < = 70
  + Target: 70 < average glucose < 140
  + Diabetic Risk: 140 <= average glucose <= 200
  + Diabetic: > 200

Self Employed has the highest normalized stroke percentage for Work type.



Additional data sources [7][8] were used to supplement the stroke visualization effort. The data was used to create a map of stroke predominance (geographic location) and associated statistics.

Yakov - if desired - place holder to discuss supplemental data -

**Limitations of Dataset for Development of Model**

The dataset from Kaggle did not include ethnicity or economic status.

**Data Cleaning and Imputation**

Data cleaning was conducted in Jupyter Notebook using Python.

As previously noted, the “Other” gender category was dropped from the dataset, resulting in removing 1 row of data.

In reviewing the raw data, the bmi attribute was identified as having 201 “N/A” values. This represents 3.9% of the dataset. The mean bmi value of 28.89 was used as the replacement value for the “N/A” values.

As noted above in the representation data tables, the raw dataset has a total 1,544 “Unknown” smoking status values representing 30.4% of the dataset. A closer look at the data showed 32% of the “Unknown” values were between the ages of 0-10 and 41% was between the ages of 0-15. The Centers for Disease Control and Prevention (CDC) defines a smoker as an Adult who has previously smoked 100 cigarettes in their lifetime and who currently smokes. Research of online literature to address this issue of “Unknown” labels was conducted and it was found that “Unknown” is an accepted category. The final decision was to leave the “Unknown” smoker status values as presented in the raw dataset.

**Visualizations**

The data was represented in four different panes, Figure 1. The focal point was an area chart which presents age versus number of strokes. Most attributes are linked to this plot and are used to present characteristics of the individuals who had strokes. Three subplots are used to complement the main area plot. The first is linked to the area chart and shows the total counts per gender, number of stroke incidents and stroke percentage related to the filter settings. The second subplot is also linked to the area chart and allows the user to see the attributes of all the individuals who meet the criteria of the filter settings. The last subplot presents Age Range vs Average BMI and Number of Strokes. The line graph shows that the average BMI Category for individuals between the ages of 0 to 69 is Obese and 70+ is Overweight.

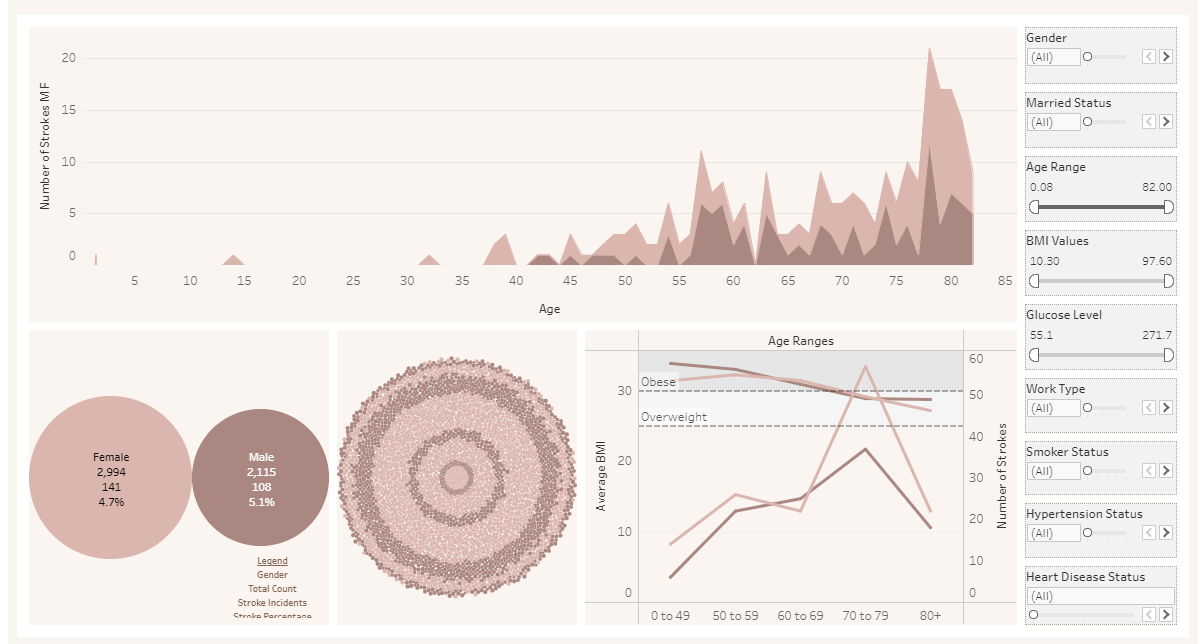


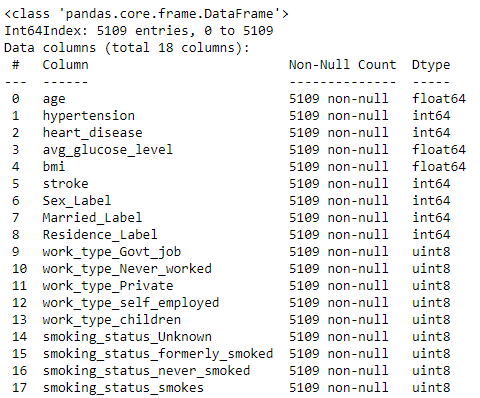
Figure 1

Yakov [7] [8]

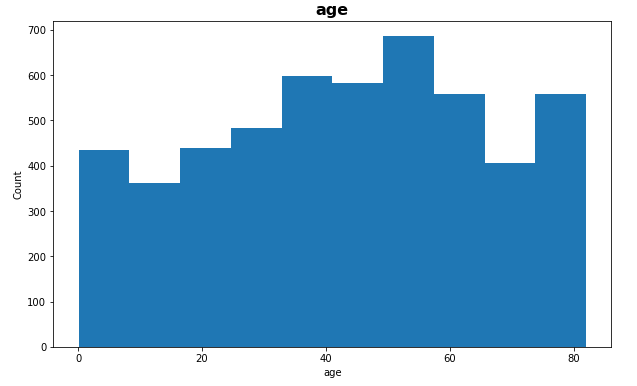
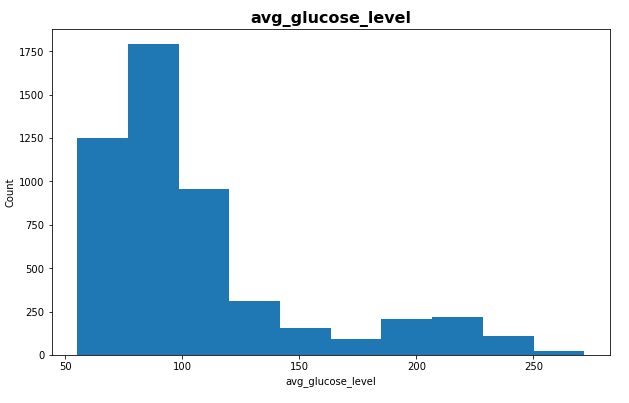
**Data Preparation for Machine Learning**

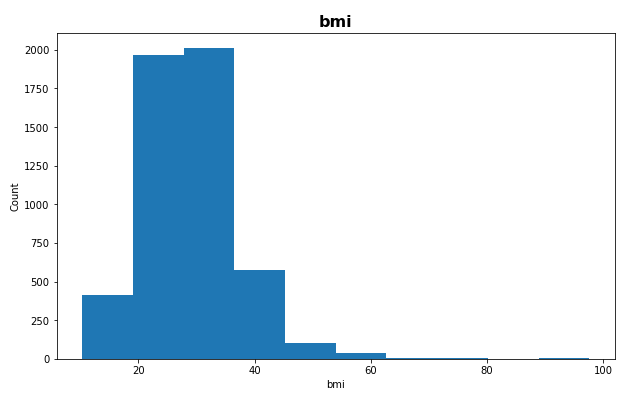
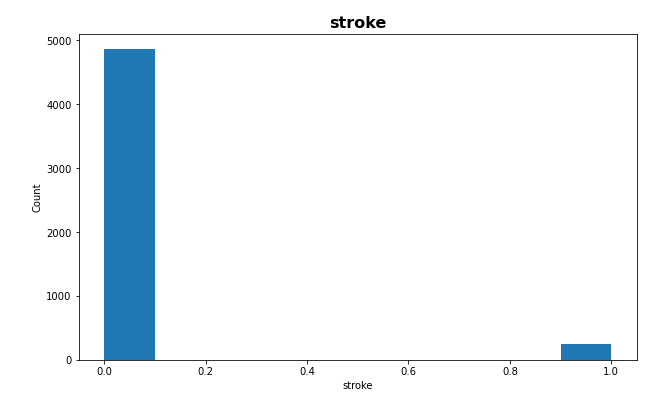
Place holder for intro discussion

One-Hot Encoding was used for categorical data work\_type, and smoking\_status to be used in the linear and tree models.

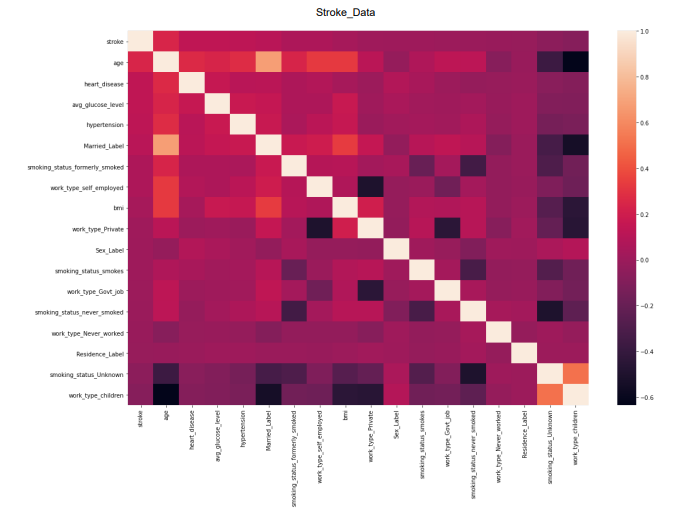


The majority of data was imbalanced, except for age and Residence\_type. The data mostly landed in the…….. Example data plots for age, bmi and average\_glucose and stroke.

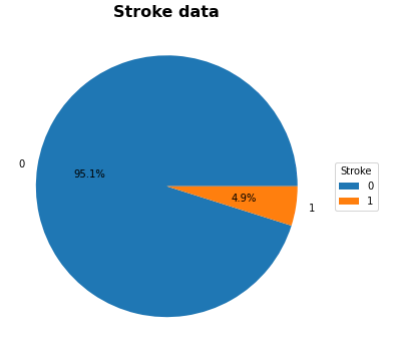
 

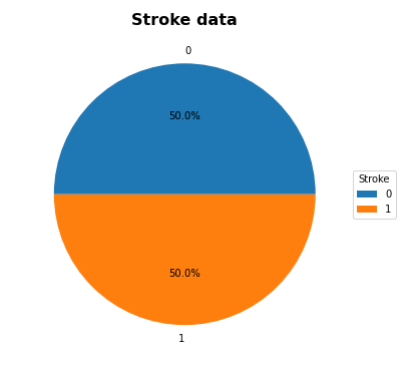
**Correlation Heat Map**



**Stroke Counts prior to synthesizing**



**Stroke data post synthesizing to manage imbalance in attribute**

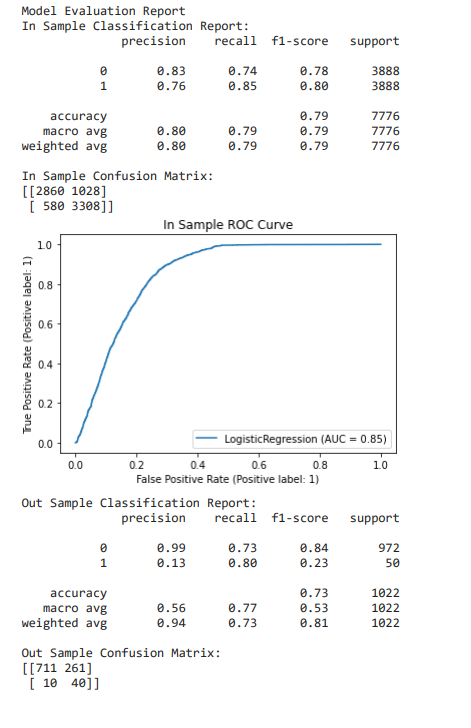


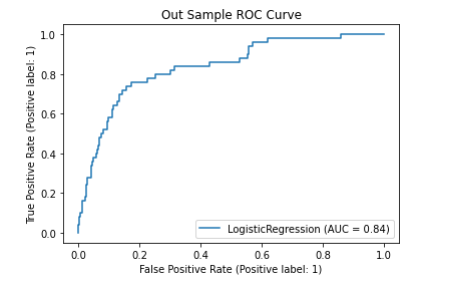
**Machine Models evaluation**

**Linear Models**

Models evaluated were Logistic Regression, KNeighbors Classifier, and SVC.

The best results of the linear models as Logistic Regression



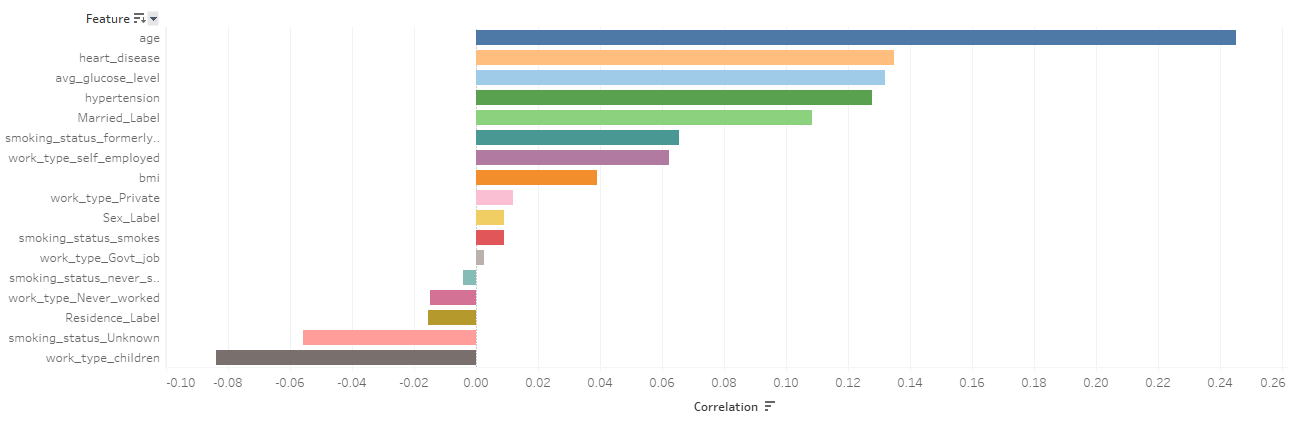


X

Y

Z

**Feature Importance**



Model x

Model y

Model z

Conclusion

Recommendations

Future Work

Appendix

References

[1] Stroke Prediction Dataset, *11 clinical features por predicting stroke events*, <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>, <https://www.kaggle.com/fedesoriano>

[2] Ahmad FB, Cisewski JA, Miniño A, Anderson RN. Provisional Mortality Data — United States, 2020. MMWR Morb Mortal Wkly Rep 2021;70:519–522. DOI: [http://dx.doi.org/10.15585/mmwr.mm7014e1external icon](http://dx.doi.org/10.15585/mmwr.mm7014e1)

[3] American Stroke Association, <https://www.stroke.org/en/about-stroke/types-of-stroke/ischemic-stroke-clots>

[4] American Stroke Association, <https://www.stroke.org/en/about-stroke/types-of-stroke/hemorrhagic-strokes-bleeds>

[5] American Stroke Association, *Explaining Stroke*, pages 1-20, <https://www.stroke.org/-/media/stroke-files/stroke-resource-center/brochures/explaining_stroke_brochure_6_25_19.pdf?la=en>

[6] American Stroke Association, <https://www.stroke.org/en/about-stroke>

[7] Stroke Mortality Data Among US Adults (35+) by State…2016, Dataset in U.S. Department of Health & Human Services, <https://data.world/us-hhs-gov/12ea7a13-b229-43b4-b19b-1459e9a64d3f>

[8] USDA Economic Research Service, U.S. Department of Agriculture, <https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>