# AAI 501 - Final Project

Lauren Taylor, Kayla Wright, Ethan Schmitt

# Background

Cerebrovascular accidents (known as strokes) can be detrimental to people's lives after they occur. Strokes cause long lasting damage due to death of brain tissue and the longer treatment is waited to be given, the more damage that occurs. It is important for a patient to know that they are at risk of developing a stroke so they can adjust their lifestyle and take special precautions to avoid such a detrimental occurrence. There are many risk factors for strokes such as: chronic high blood pressure, diabetes, heart diseases, high cholesterol levels, and smoking. This is not a full list of risk factors, but a patient's awareness of risk factors can reduce the risk of strokes by employing lifestyle changes and understanding signs and symptoms of stroke for quick identification.

# Objectives

- To use exploratory data analysis to understand the dataset.
- To use logistic regression to predict people that have had strokes from this dataset.
- To use and refine a decision tree to predict people that have had strokes in the dataset.
- To have a better understanding of the types of patients at risk for developing a stroke and what the most telling risk factors are.

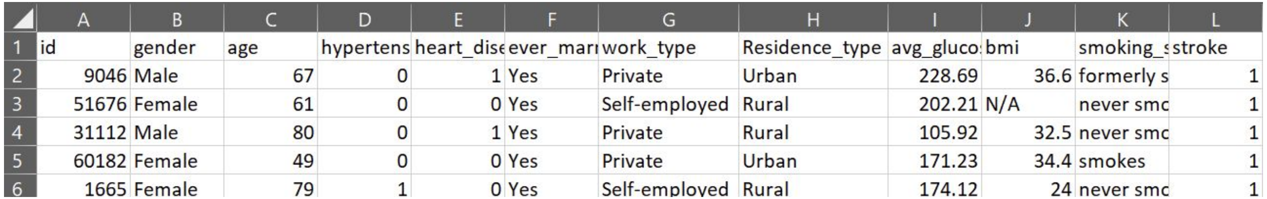
# Agenda

- Our Data Set
- Data analysis and insights
- Introduction and Research
- SMOTE Algorithm
- Our Approach - High Level Breakdown
- Logistic Regression Model
- Decision Tree Classification Model
- Results
- Conclusions and Takeaways

# Our Data Set

Link:

Stroke Prediction Dataset



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | gender | age | hypertens | heart_dise | ever_marr | work_type | Residence_type | avg_gluco | bmi | smoking_s | stroke |
| 2 | 9046 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly s | 1 |
| 3 | 51676 | Female | 61 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | N/A | never smc | 1 |
| 4 | 31112 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smc | 1 |
| 5 | 60182 | Female | 49 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| 6 | 1665 | Female | 79 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24 | never smc | 1 |

## Inputs

- ID
- Age
- Gender
- Hypertension
- Heart Disease
- Ever Married
- Work Type
- Residence Type
- Ave. Glucose Level
- BMI

## Issues

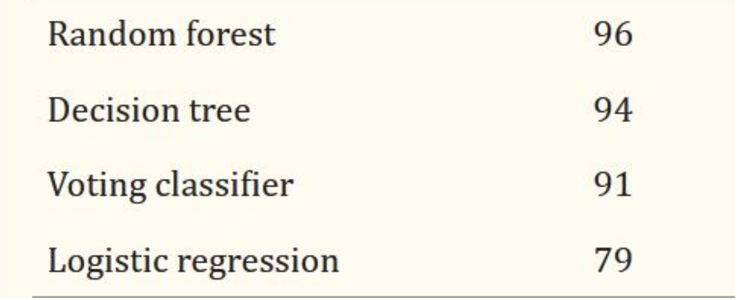
- NaNs in BMI
- Imbalanced output
- Categorical inputs

## Solutions

- Impute median
- SMOTE algorithm
- Dummy variables

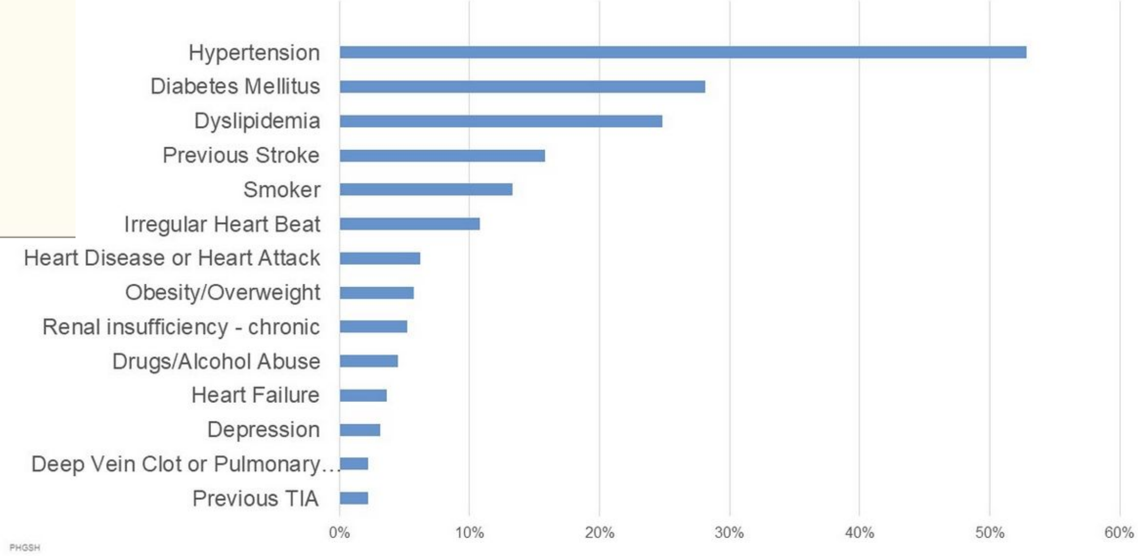
# Introduction and Research

Previous Study Link



| | |
|---|---|
| Random forest | 96 |
| Decision tree | 94 |
| Voting classifier | 91 |
| Logistic regression | 79 |

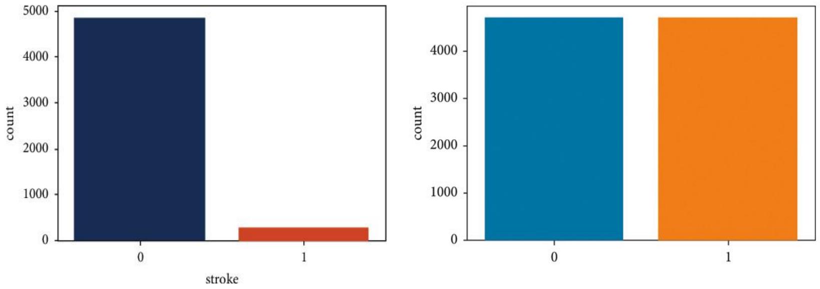
Common Stroke Risk Factors



# SMOTE Algorithm

- Used to create synthetic samples when output is imbalanced
- Can skew dataset to make false negatives/positives more/less likely
- Use imblearn Python package to implement SMOTE.

Distribution of Outputs in the Stroke Prediction Database



Graphical Representation of How SMOTE Works

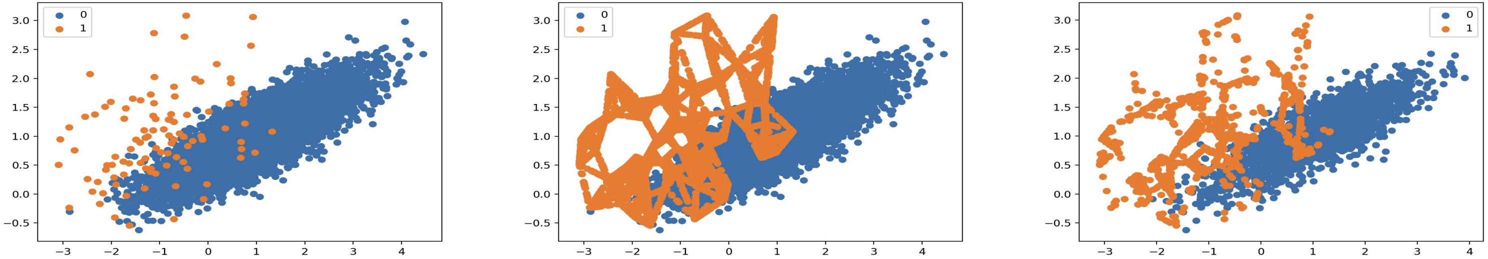
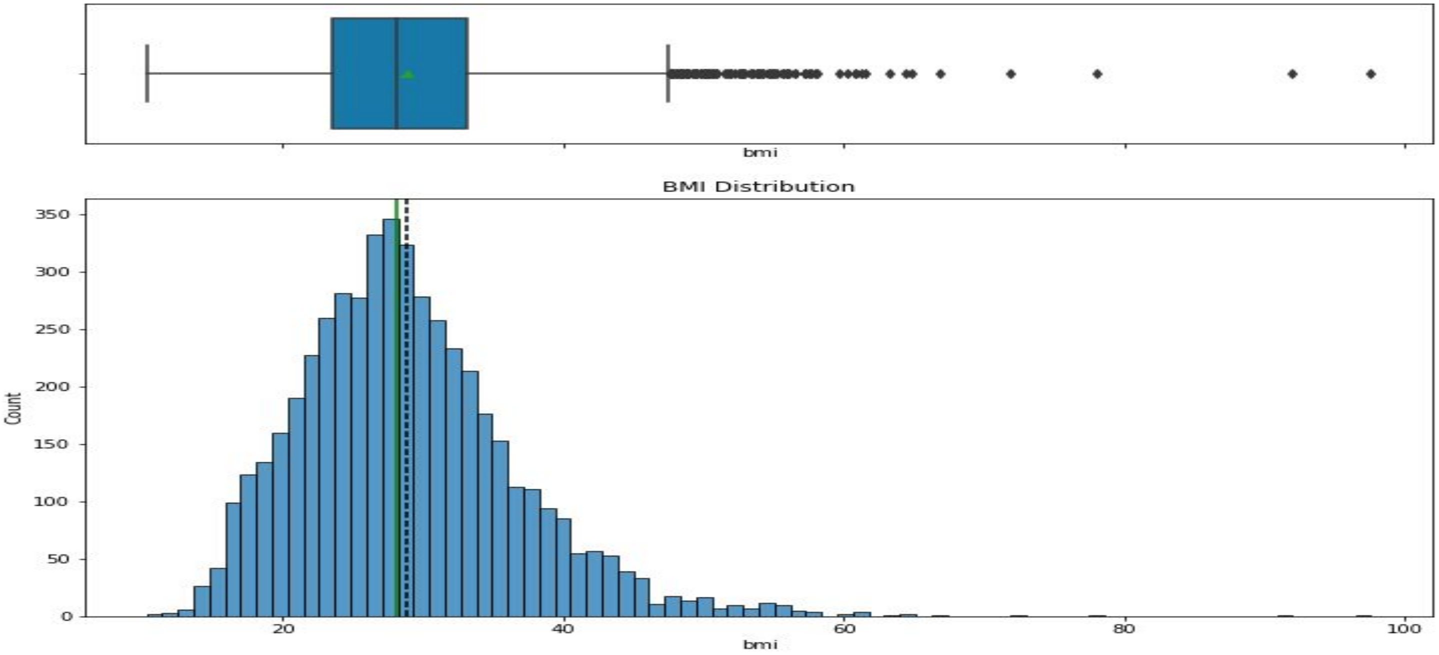


Figure 1: Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of healthcare engineering*, *2021*, 7633381. https://doi.org/10.1155/2021/7633381
Figure 2: Korstanje, J. (2021, August 30). Smote. Towards Data Science. https://towardsdatascience.com/smote-fdce2f605729

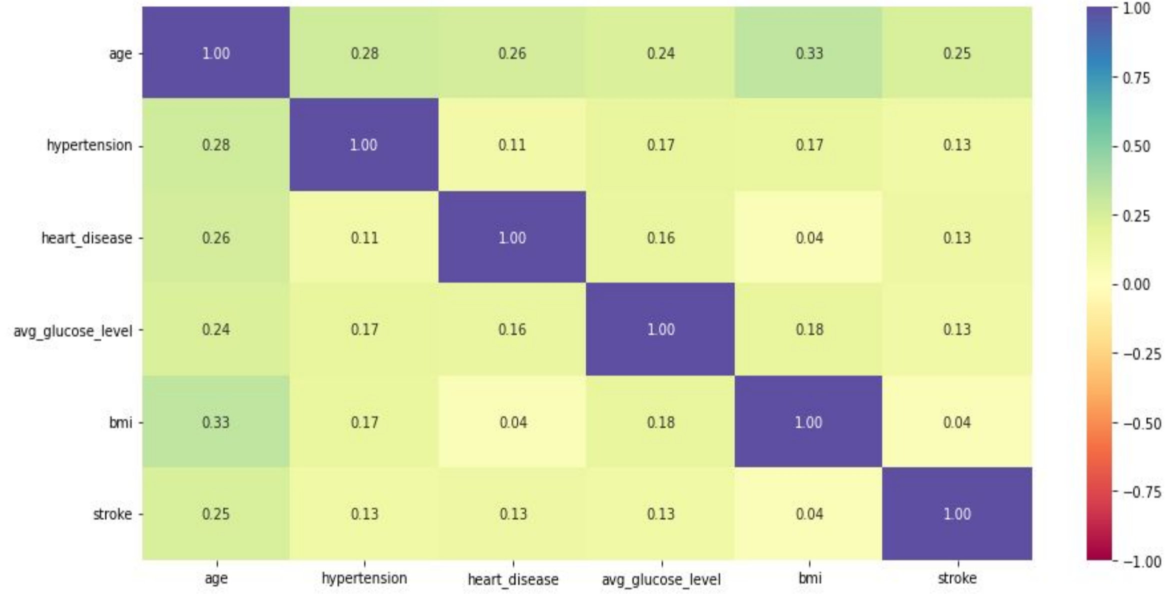
# Exploratory Data Analysis - Univariate

- Our first road bump was that BMI had missing values. In order to address this situation and impute the correct numbers, we had to take a look at the distribution.
- BMI ended up being slightly right-skewed, so we imputed the median.
- We did not want to drop the over 200 missing values so we agreed this was the best course of action.

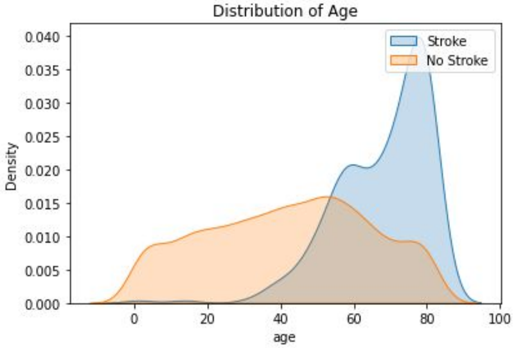


# Exploratory Data Analysis - Multivariate

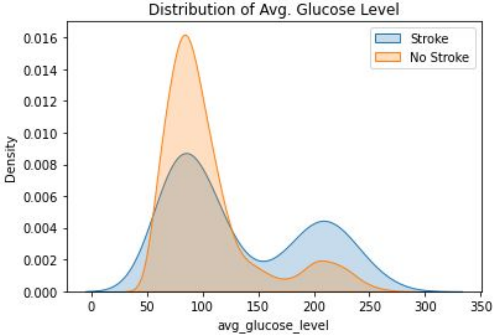
- Finding the variables that correlate with stroke was also very important to us. We started with a heatmap of all variables.
- The highest correlations with stroke were age, high blood pressure, heart disease, and average glucose level.



|  | age | hypertension | heart_disease | avg_glucose_level | bmi | stroke |
|---|---|---|---|---|---|---|
| age | 1.00 | 0.28 | 0.26 | 0.24 | 0.33 | 0.25 |
| hypertension | 0.28 | 1.00 | 0.11 | 0.17 | 0.17 | 0.13 |
| heart_disease | 0.26 | 0.11 | 1.00 | 0.16 | 0.04 | 0.13 |
| avg_glucose_level | 0.24 | 0.17 | 0.16 | 1.00 | 0.18 | 0.13 |
| bmi | 0.33 | 0.17 | 0.04 | 0.18 | 1.00 | 0.04 |
| stroke | 0.25 | 0.13 | 0.13 | 0.13 | 0.04 | 1.00 |



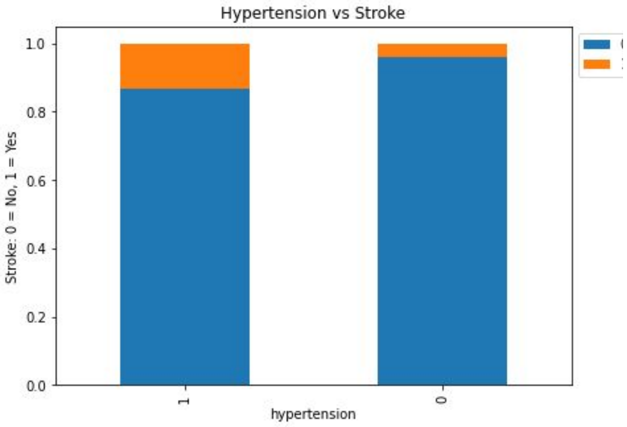
Distribution of Age

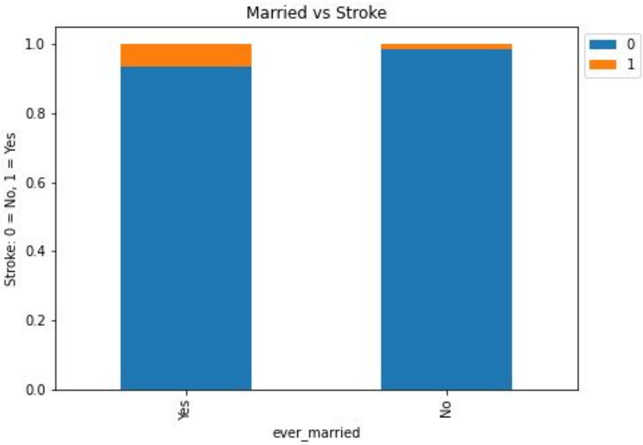


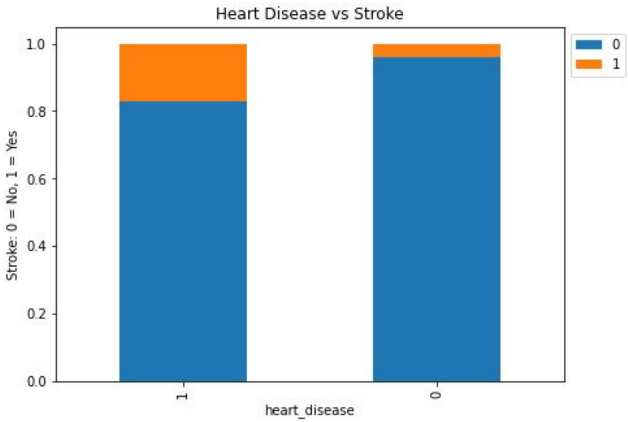
Distribution of Avg. Glucose Level

# Exploratory Data Analysis - Multivariate

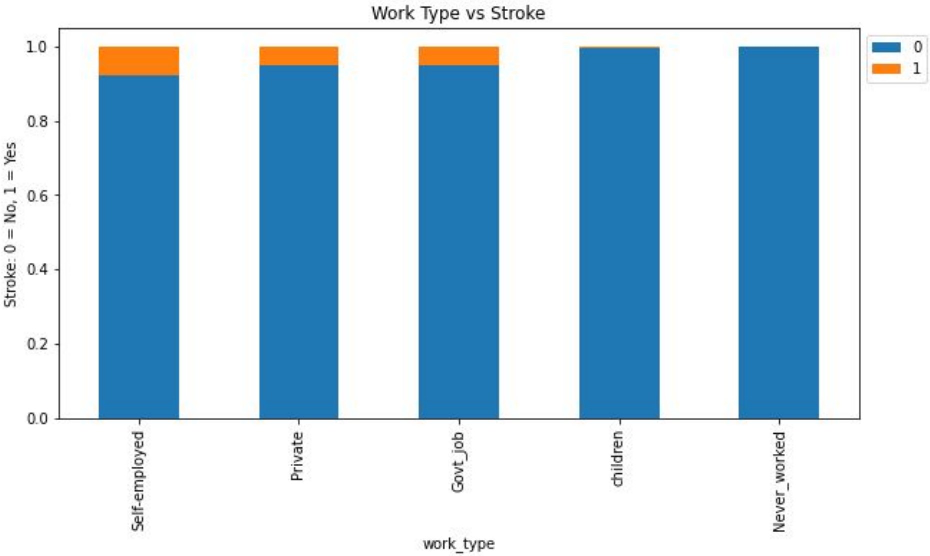
- Next we wanted to take a closer look at the categorical variables and the binary classifiers that could impact stroke.

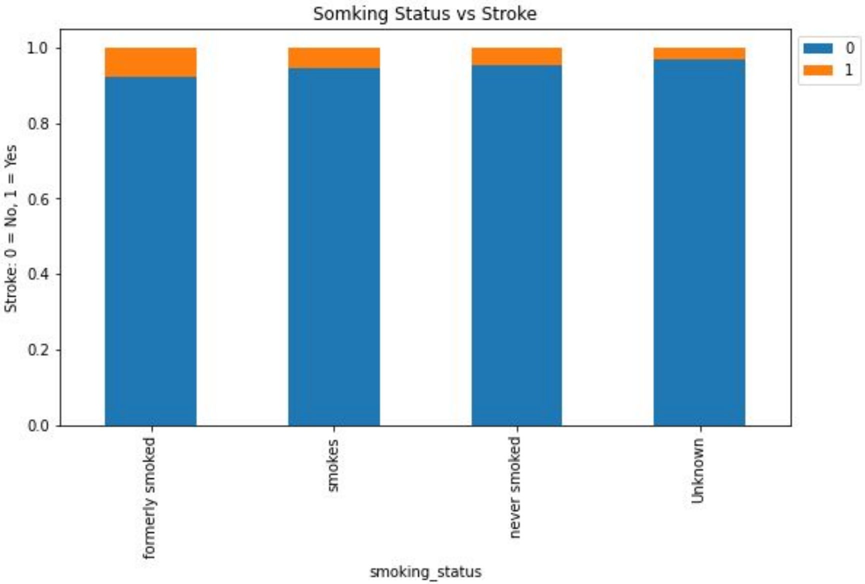






- We determined after looking at each of these types of variables, the factors that predicted stroke the most were high blood pressure, heart disease, marriage, being self employed, and being a former smoker.

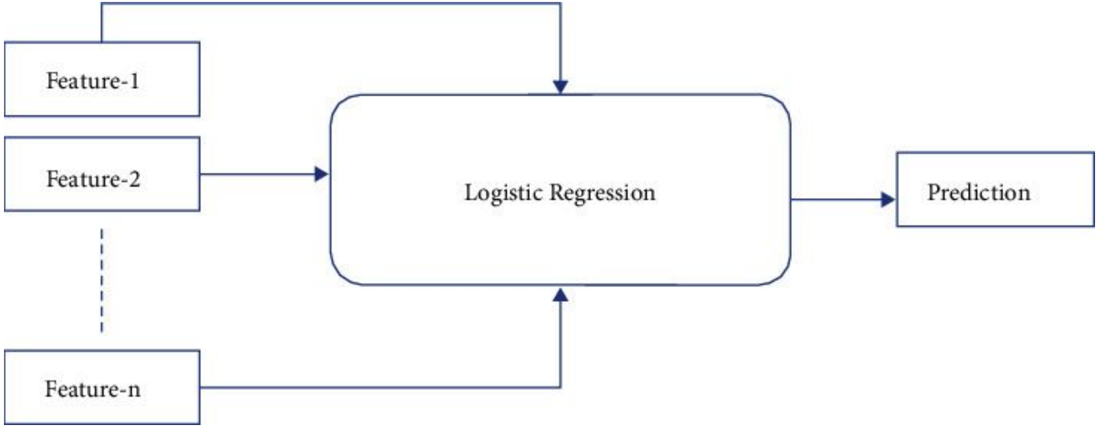




# Our Approach - High Level Breakdown

- Import dataset
- Remove ID input column
- Detect and replace BMI NaNs with median
- Address how we will deal with a 0.05 to 0.95 imbalanced dataset.
- Use SMOTE to create synthetic samples
- Create dummy variables for categoricals
- Train logistic regression model
- Remove variables with high p-values
- Train logistic regression model
- Train decision tree classification model
- Inspect results from models

# Logistic Regression



-Use statsmodels Logit function to train model.

-Use sklearn functions for preprocessing:

- Train-test split
- Standard scaler
- Confusion matrix
- Accuracy score

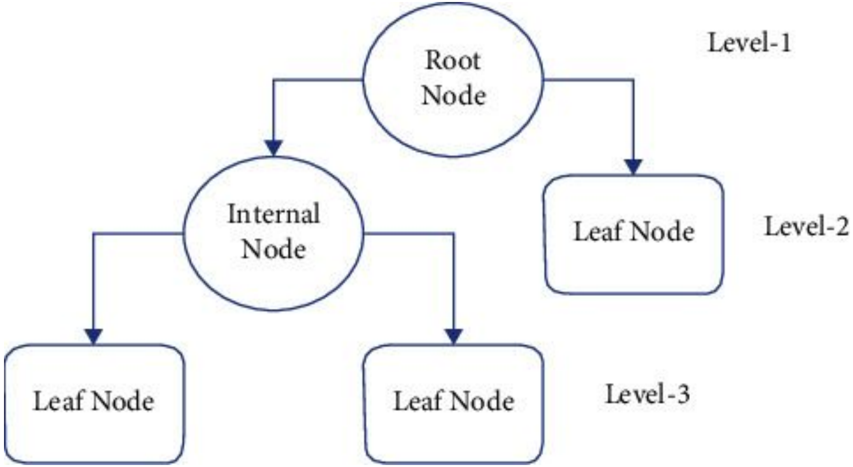
-Features:

- Gender
- Residence type
- Ever married
- Gender
- Avg glucose level
- Age
- Smoking statuses

-Threshold for classification:

- If model output is greater than or equal to .4 → classification = 1

# Decision Tree Classification



- The only column we dropped here was residence type.
- Create dummy variables for categorical values
- Split train/test by 70-30
- Create a base decision tree with gini impurity which classifies probability that a feature that will be incorrectly classified if randomly selected.
- Evaluate recall and accuracy for base tree.
- Utilize GridSearchCV which finds the best parameters for a tree for the given data.
- Understand the parameters chosen by GridSearchCV and check recall + accuracy.

# Recall, Accuracy, Precision, or F1 Score?

- Ideally, we want to reduce the number of false negatives, or the number of people who are predicted to not have a stroke and actually have one.
- False negatives are extremely costly for the healthcare industry because incorrectly classifying if a patient will not have a stroke when they do can cost the patient's trust, wellbeing, and life.

Recall Score: Measures the model's ability to predict the correct positives out of all of the positive results. This score is great for an imbalanced dataset like ours. Very good at letting us know how many false positives that are present.
Recall = TP / (FN + TP)

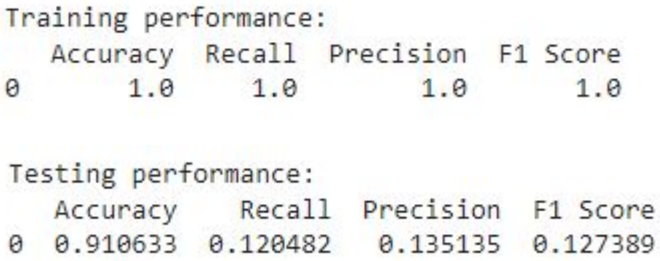
Accuracy: Measures the model's ability to predict by looking at the ratio of true positives and true negatives.
Accuracy = (TP + TN) / (TP + FN + TN + FP)

Precision: Measures the model's ability to correctly identify the proportion of correct positive classifications.
Precision = TP / (FP + TP)

F1 Score: A harmonic mean between precision and recall.
F1 Score = 2 * precision * recall / (precision + recall)

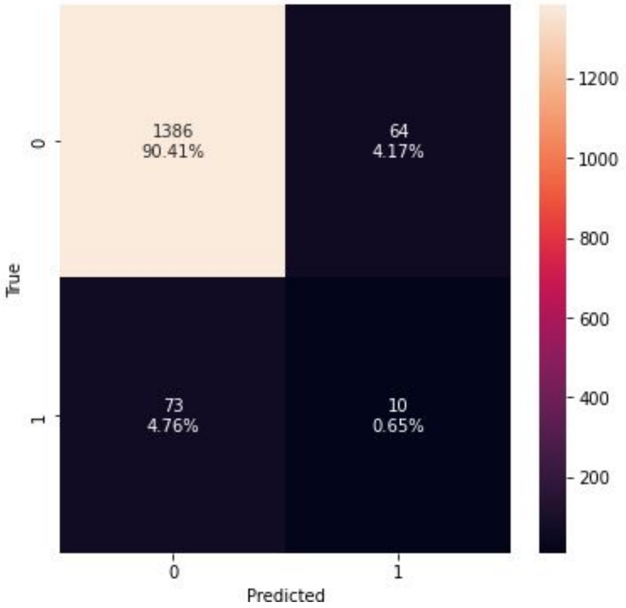
- We want to optimize our recall score without sacrificing too much.

# Decision Tree Results - Model 1



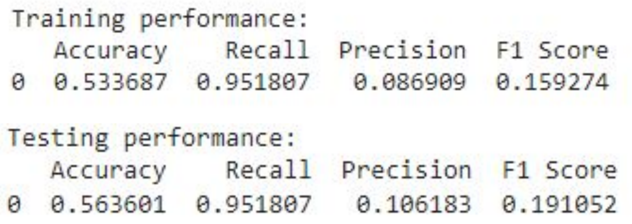
```
Training performance:
   Accuracy  Recall  Precision  F1 Score
0       1.0     1.0        1.0       1.0

Testing performance:
   Accuracy    Recall  Precision  F1 Score
0  0.910633  0.120482   0.135135  0.127389
```



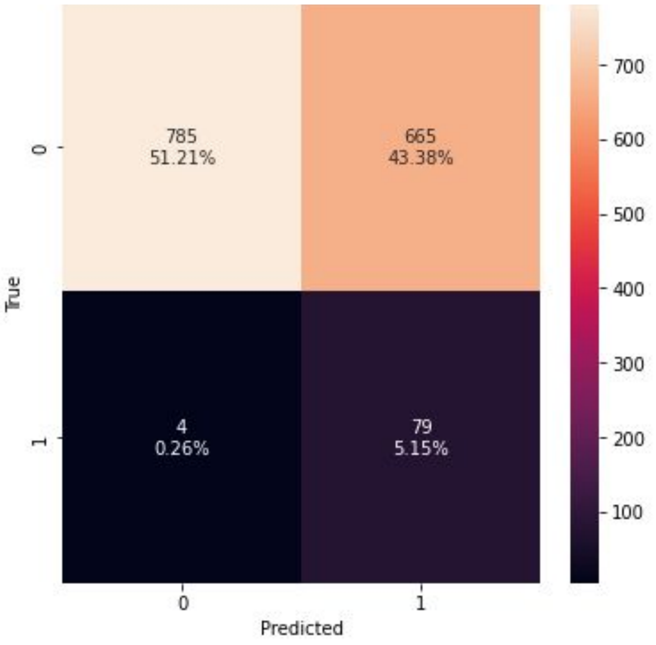
- This first model was created with the gini impurity only.
- It is heavily overfit as seen from training and testing performance.
- The accuracy is decent, but we are aiming for a smaller false negative rate. There are 73 false negatives which would not be a good model to employ in healthcare.
- Next we will try optimizing with GridSearchCV.

# Decision Tree Results - Model 2



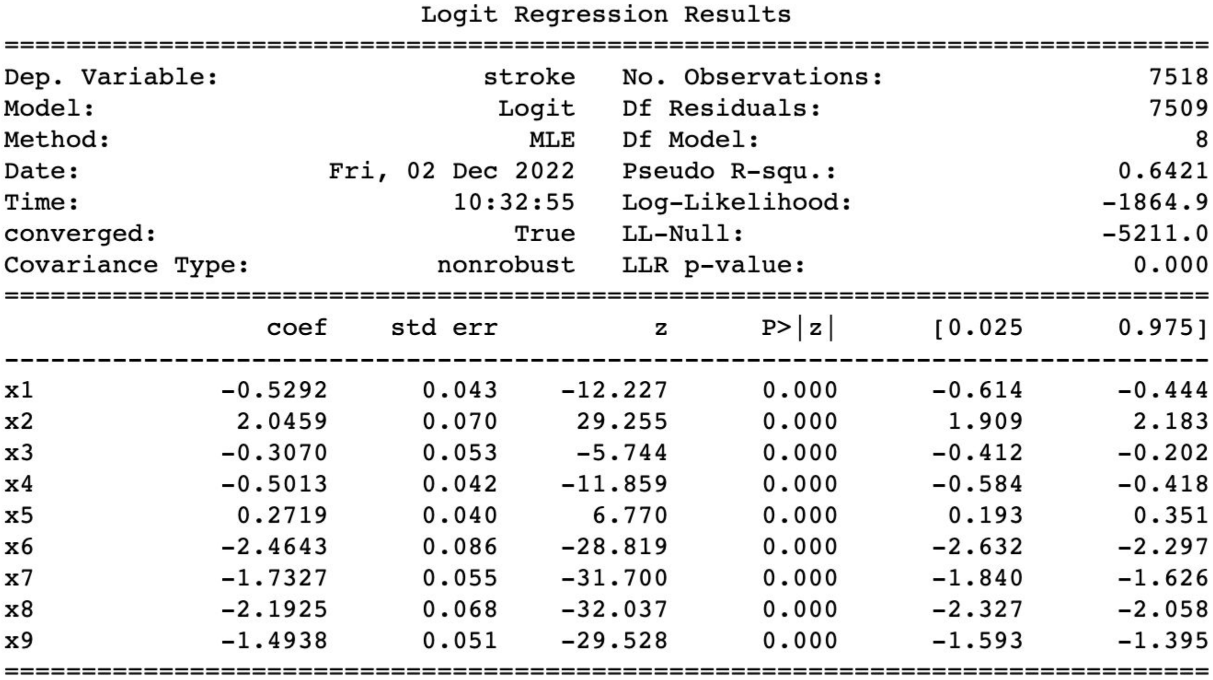
```
Training performance:
   Accuracy    Recall  Precision  F1 Score
0  0.533687  0.951807   0.086909  0.159274

Testing performance:
   Accuracy    Recall  Precision  F1 Score
0  0.563601  0.951807   0.106183  0.191052
```



- After using GridSearchCV to find optimal parameters we ended up with these metrics.
- The parameters chosen were: (class_weight={0: 0.05, 1: 0.95}, criterion='entropy', max_depth=5, min_impurity_decrease=0.01, random_state=1, splitter='random')
- The accuracy is not ideal, but the recall is great and less stroke patients are being misclassified as not having one.
- The model is not overfit anymore which is a huge concern for decision trees.

# Logistic Regression Results



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 stroke   No. Observations:                 7518
Model:                          Logit   Df Residuals:                     7509
Method:                           MLE   Df Model:                            8
Date:                Fri, 02 Dec 2022   Pseudo R-squ.:                  0.6421
Time:                        10:32:55   Log-Likelihood:                -1864.9
converged:                       True   LL-Null:                       -5211.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.5292      0.043    -12.227      0.000      -0.614      -0.444
x2             2.0459      0.070     29.255      0.000       1.909       2.183
x3            -0.3070      0.053     -5.744      0.000      -0.412      -0.202
x4            -0.5013      0.042    -11.859      0.000      -0.584      -0.418
x5             0.2719      0.040      6.770      0.000       0.193       0.351
x6            -2.4643      0.086    -28.819      0.000      -2.632      -2.297
x7            -1.7327      0.055    -31.700      0.000      -1.840      -1.626
x8            -2.1925      0.068    -32.037      0.000      -2.327      -2.058
x9            -1.4938      0.051    -29.528      0.000      -1.593      -1.395
==============================================================================
```

- Based on the p-values all the predictors are significant.

-Train accuracy = 88%

-Test accuracy = 89%

-Similarity between train and test accuracy indicate low likelihood of overfitting.

-Which predictors had the most weight?

# Logistic Regression Results Cont.

| **Confusion Matrix** | | Predicted | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | 856 | 103 |
| | Negative | 99 | 822 |

-False Negative and positive rates equal to approximately 10.7%.

-Which is worse false positive or negative?

-Think of consequences in a clinical setting.

# Conclusions and Takeaways

- Our logistic regression model outperformed the previous paper
- Main Differentiators
    - Biasing the model towards false positive
- Using new inputs and data to refine
    - Stroke History
    - MRI data
    - EEG data
    - Genetics
- New algorithms or network architectures
    - Convolutional Neural Networks
    - Deep Learning Networks / LSTM Cells
    - Transformers
- Constant Real-time Analysis for High Risk Patients

# References

About Stroke. Retrieved December 5, 2022, from https://www.cdc.gov/stroke/about.htm

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of healthcare engineering*, *2021*, 7633381. https://doi.org/10.1155/2021/7633381

Early Stroke Prediction Methods for Prevention of Strokes. (2022, April 11). Retrieved December 5, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9017592/

Stroke Prediction Dataset. (2020). Retrieved December 5, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Korstanje, J. (2021, August 30). Smote. Towards Data Science. https://towardsdatascience.com/smote-fdce2f605729