Predicting Stroke

Presented By: Leah Apking, Erin Clark, Nancy Gomez, & Sheila Troxel

According to the World Health Organization stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths each year.

Risk Factors Considered:

- Age
- Hypertension
- Average Glucose Level
- BMI
- Stroke
- Gender
- Maritial Status
- Employment Type
- Residence Type
- Heart Disease

Previewing the Data

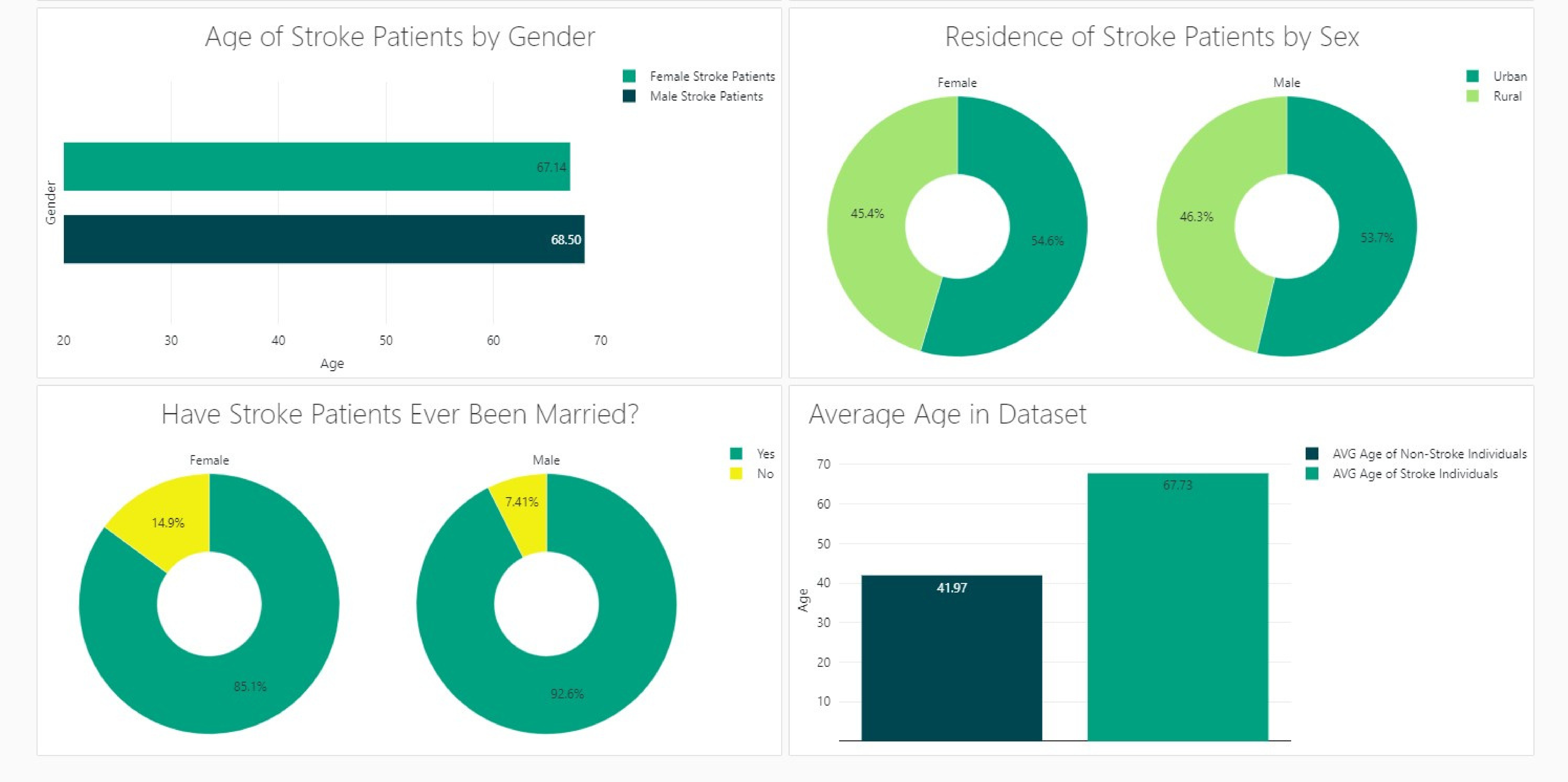
Using a DataBricks notebook, Python, and Spark SQL we were able to review and analyze the stroke prediction data to learn more about the patients included in the dataset and how the clinical features may factor into our predictions.

Spark SQL

Python

DataBricks

Visualizing Stroke & Predictive Factors



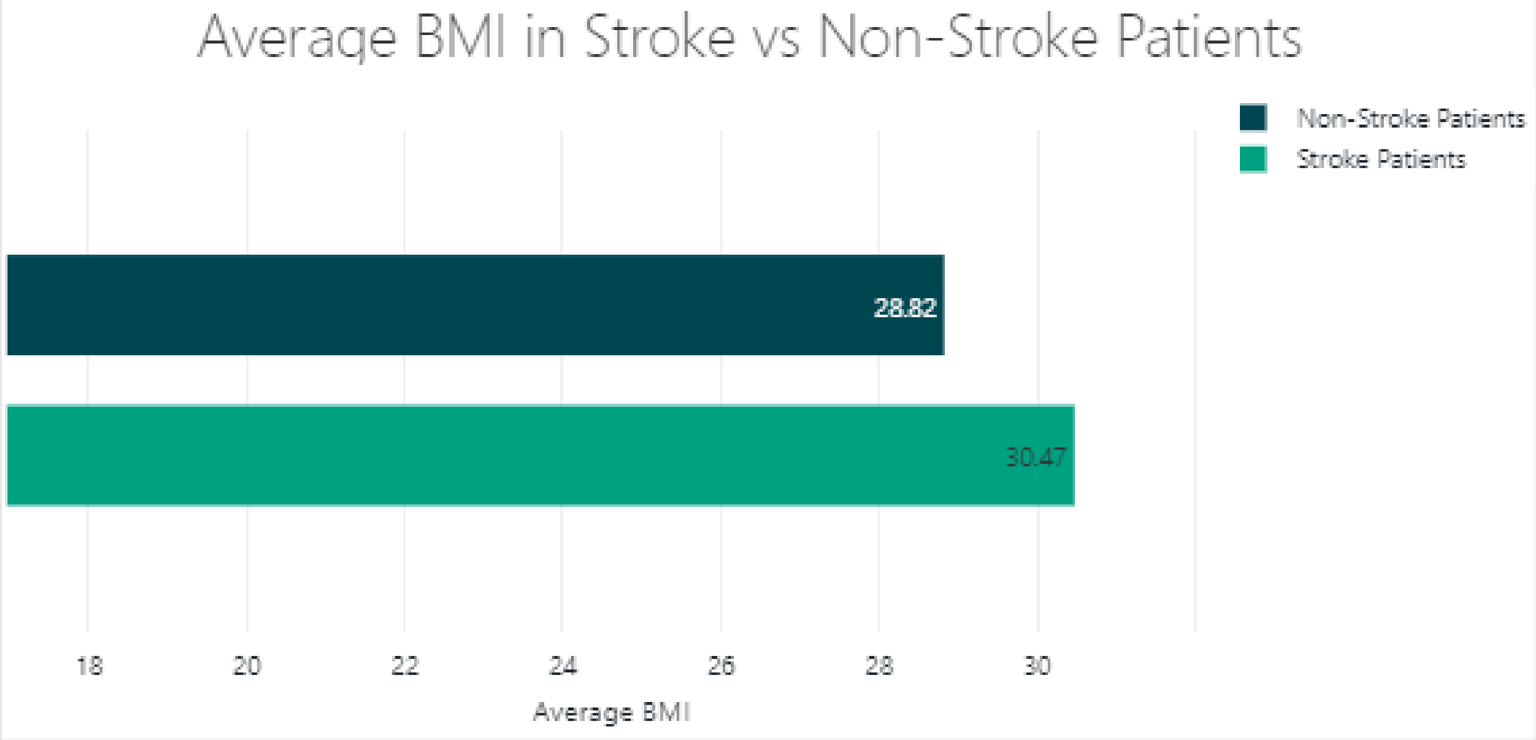
Age of Stroke Patients by Gender

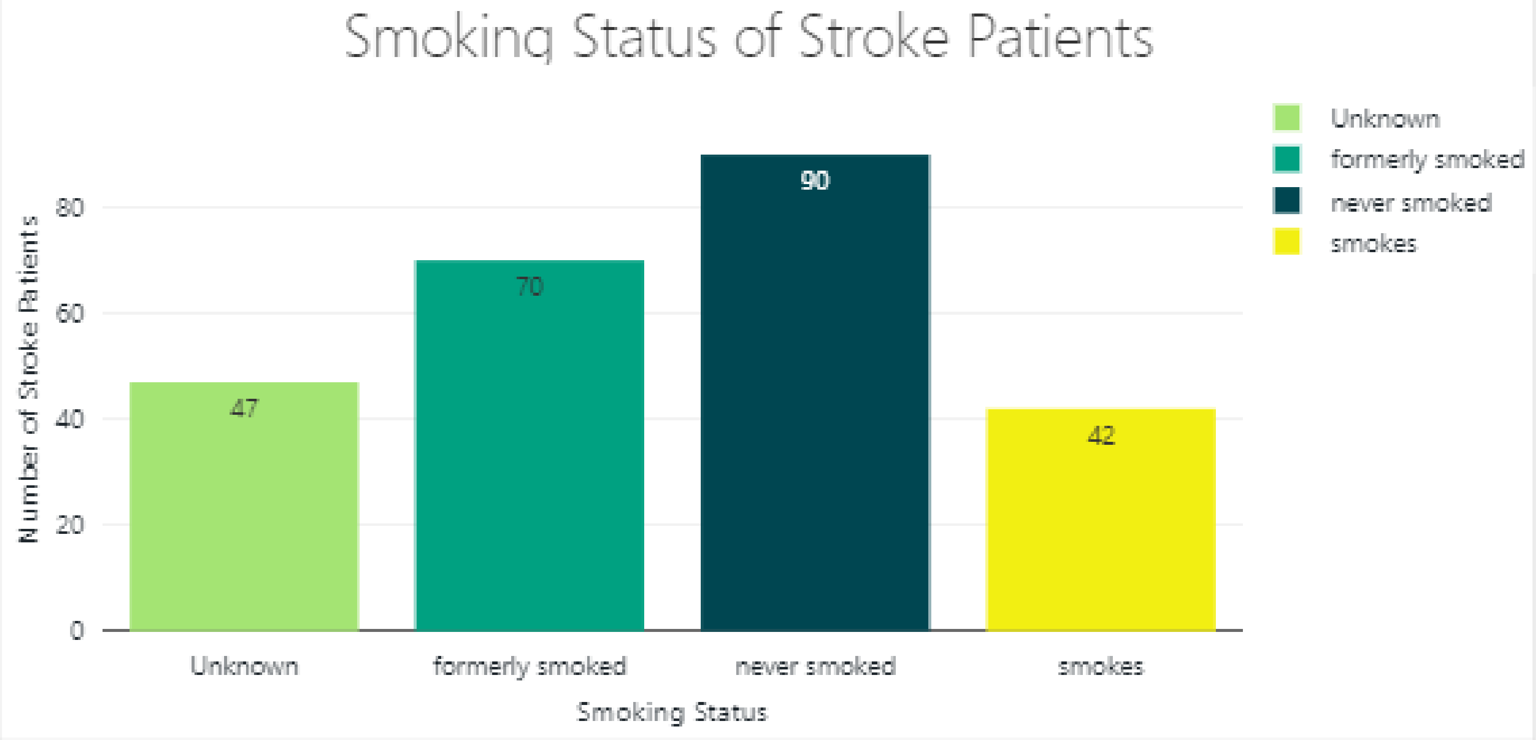
Residence of Stroke Patients by Sex

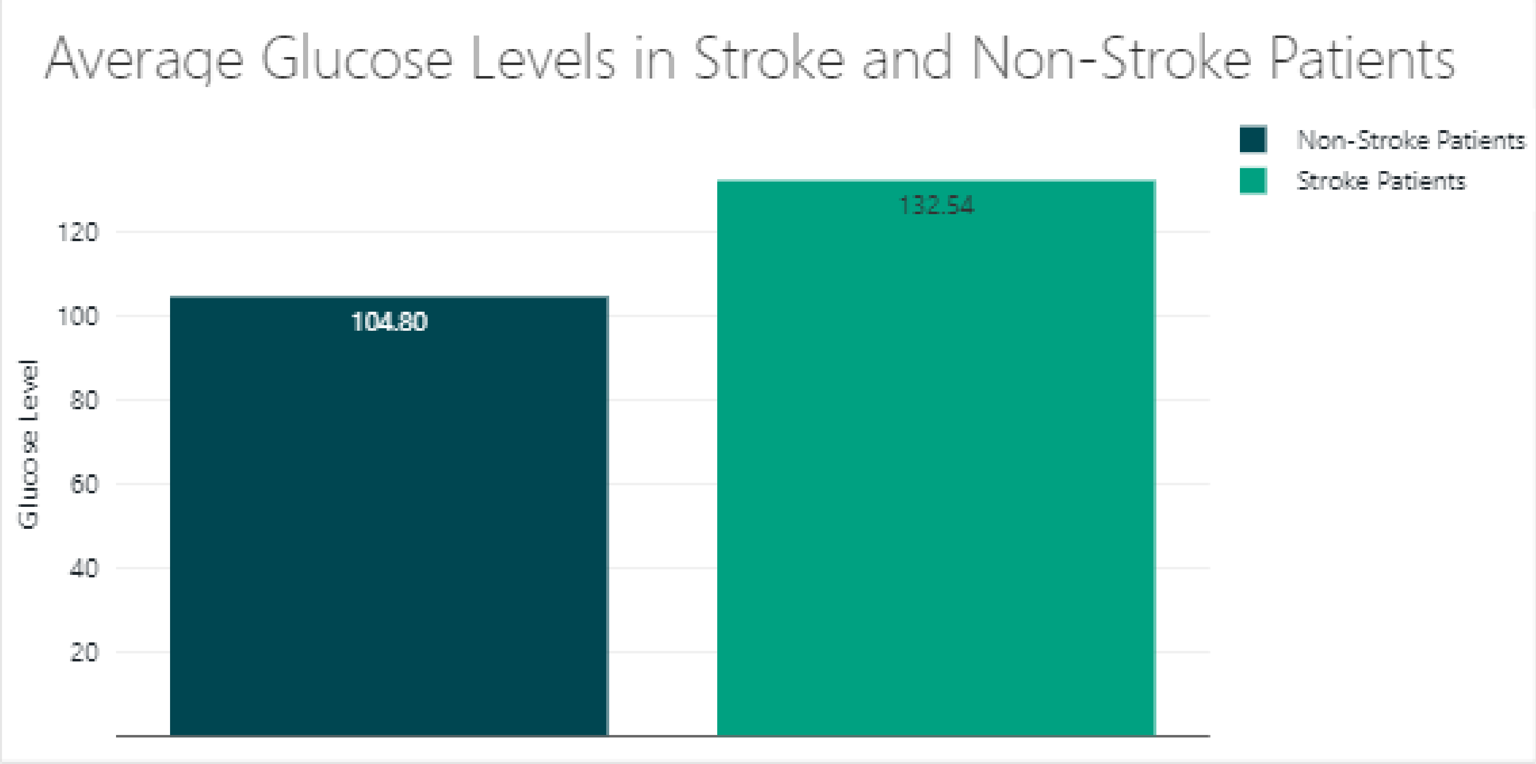
Have Stroke Patients Ever Been Married?

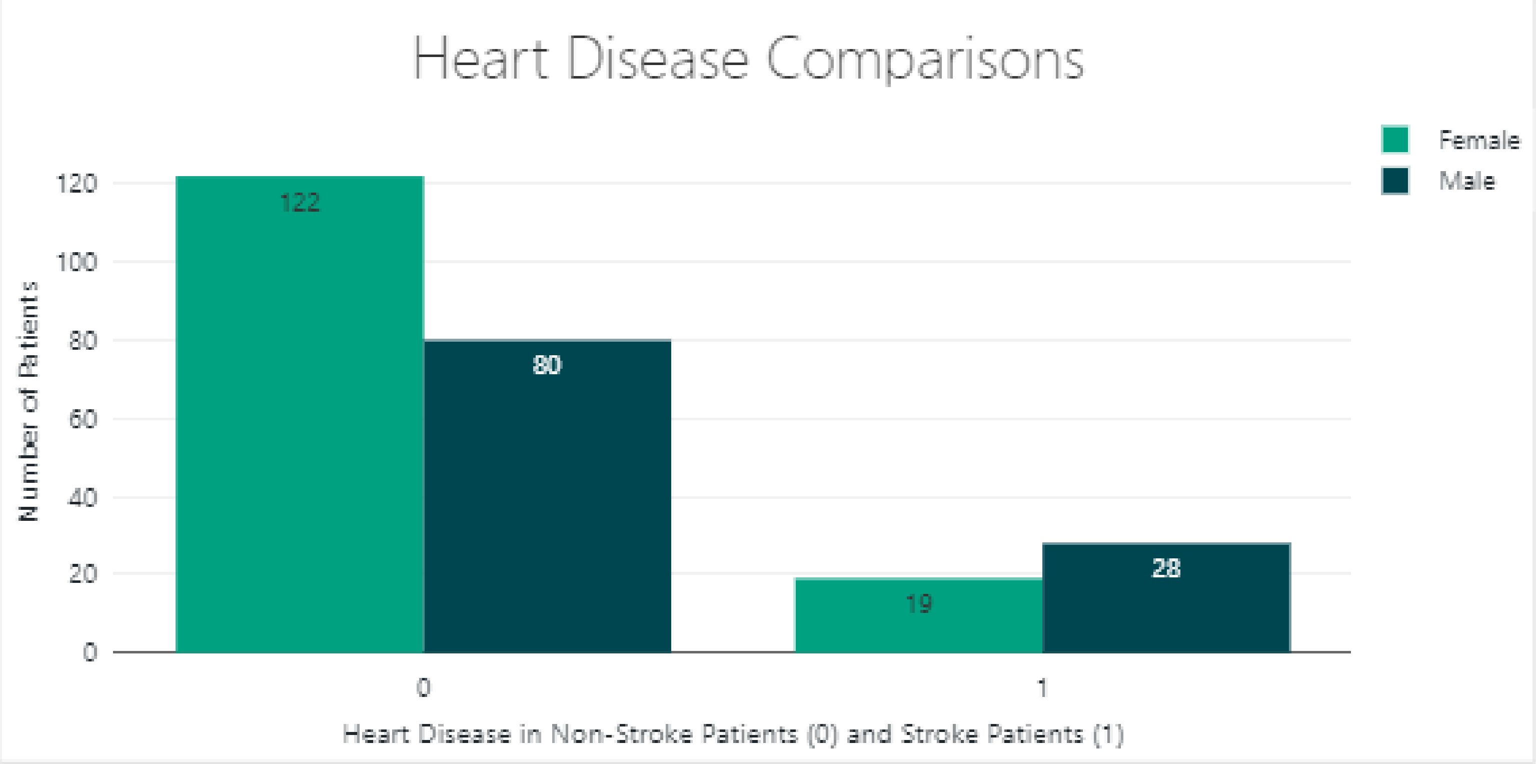
Average Age in Dataset

Visualizing Stroke & Predictive Factors

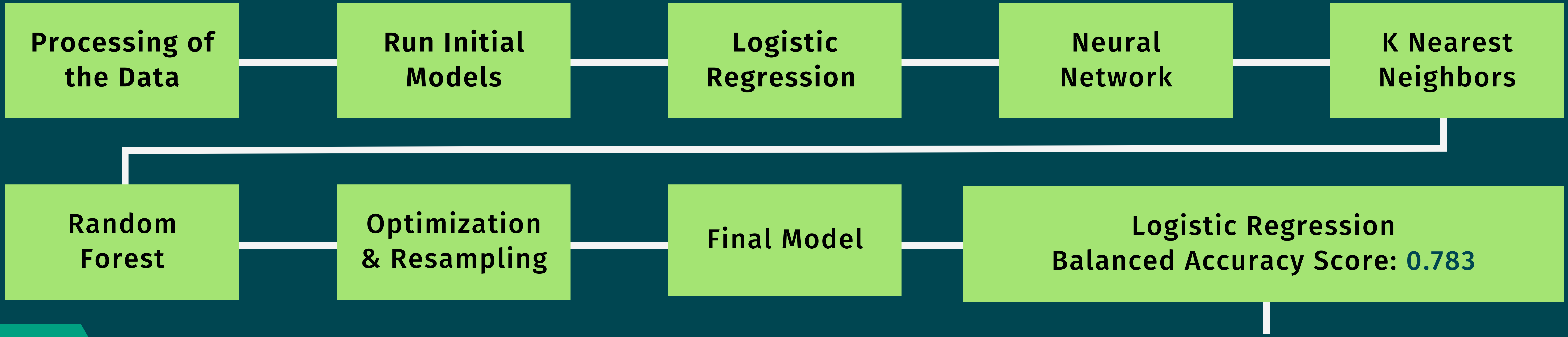

Average BMI in Stroke vs Non-Stroke Patients


Smoking Status of Stroke Patients

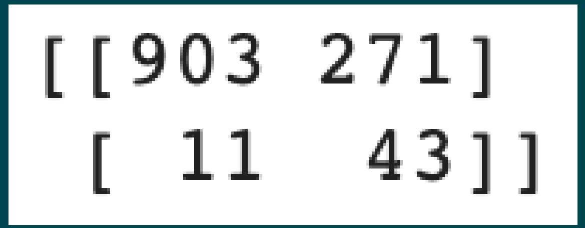

Average Glucose Levels in Stroke and Non-Stroke Patients


Heart Disease Comparisons

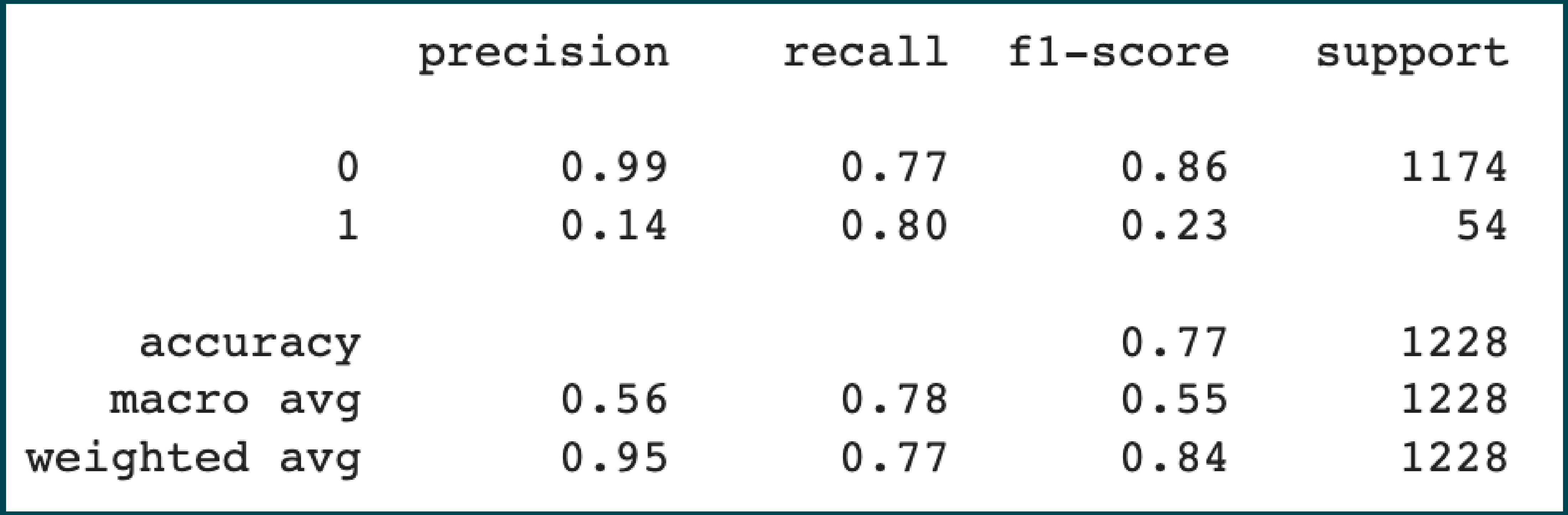
Machine Learning Model



Confusion Matrix



```
[[903 271]
 [ 11  43]]
```



```
              precision    recall  f1-score   support

           0       0.99      0.77      0.86      1174
           1       0.14      0.80      0.23        54

    accuracy                           0.77      1228
   macro avg       0.56      0.78      0.55      1228
weighted avg       0.95      0.77      0.84      1228
```

Summary of Findings

Our focus when building this model was to identify stroke patients with the hope of being able to predict which patients are at a high risk of having a stroke in the future. We did our best to accommodate the lopsided dataset, which upon further investigation was not highly representative of the demographic most likely to suffer or have suffered a stroke. Given a larger more targeted dataset, such as older adults, with additional features such as family history, LDL cholesterol levels, presence of diabetes, or race and ethnicity it is likely that further modeling can help more accurately identify patients at a greater risk for stroke.

Resource & Tools Page

Kaggle Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Google Colab

Python

Spark SQL

DataBricks

Images: Unsplash

TensorFlow

Ski-Kit Learn

Imbalanced-Learn