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

Risk Factors Considered:

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

Previewing the Data

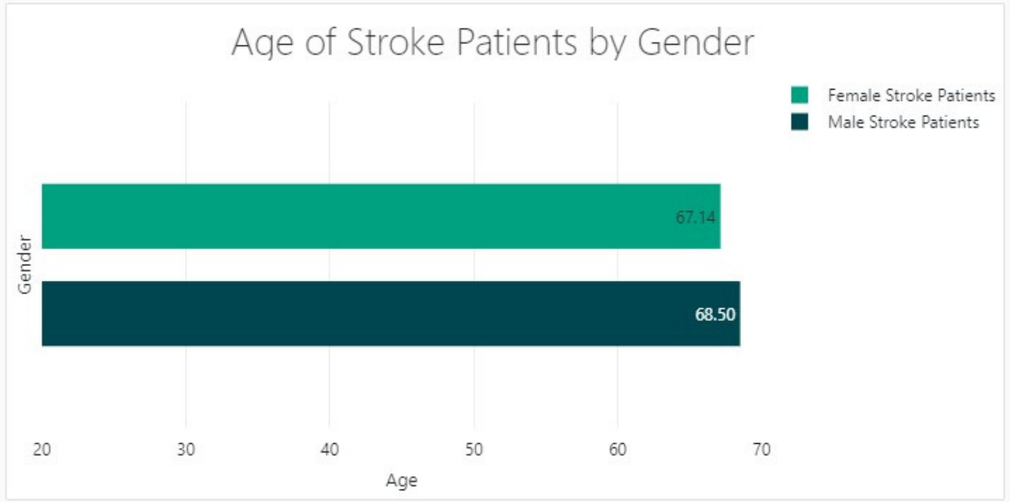
Using a DataBricks notebook, Python, and Spark SQL we were able to review and analyze the stroke prediction data to learn more about the patients included in the dataset and how the clinical features may factor into our predictions.

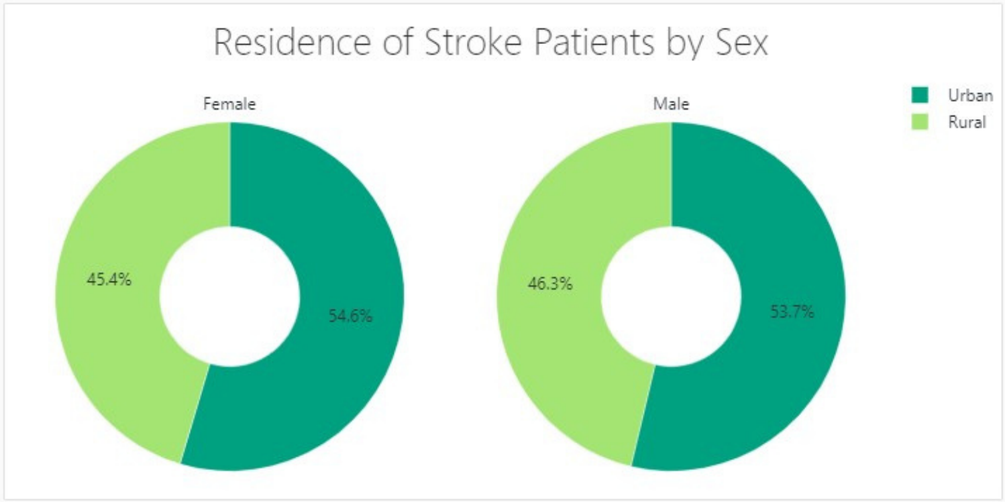
Spark SQL

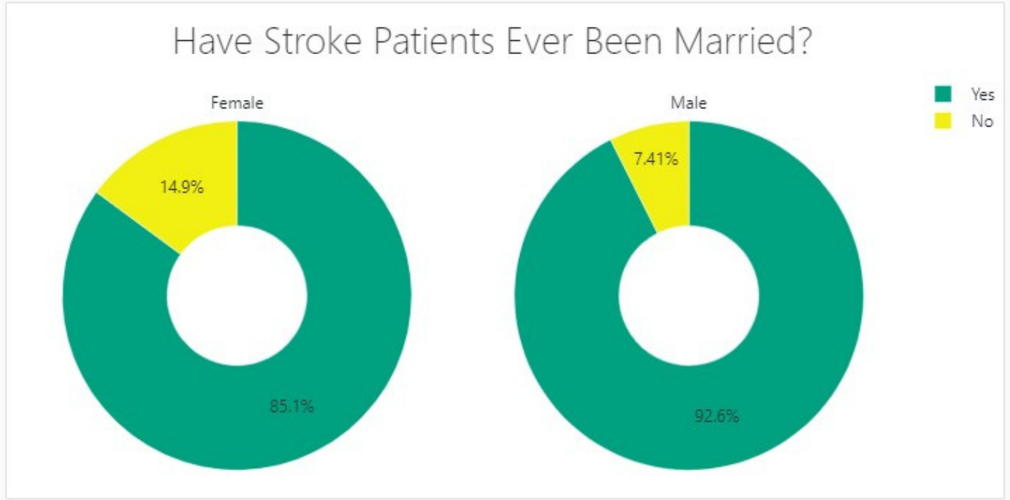
Python

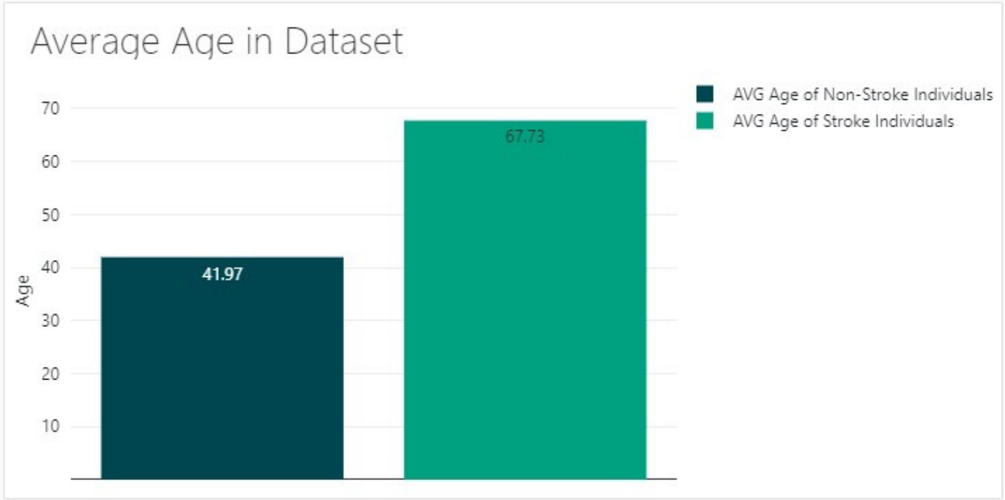
DataBricks

Visualizing Stroke & Predictive Factors

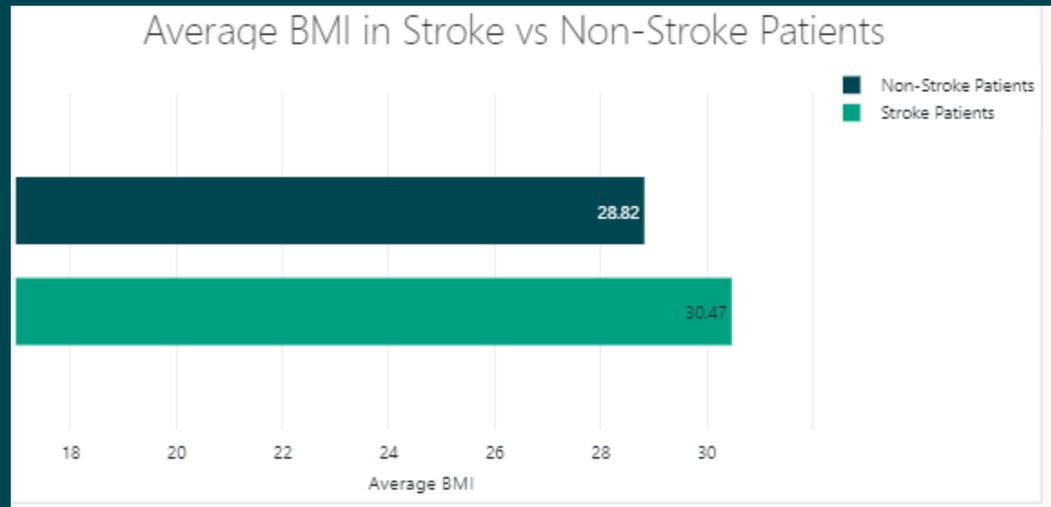


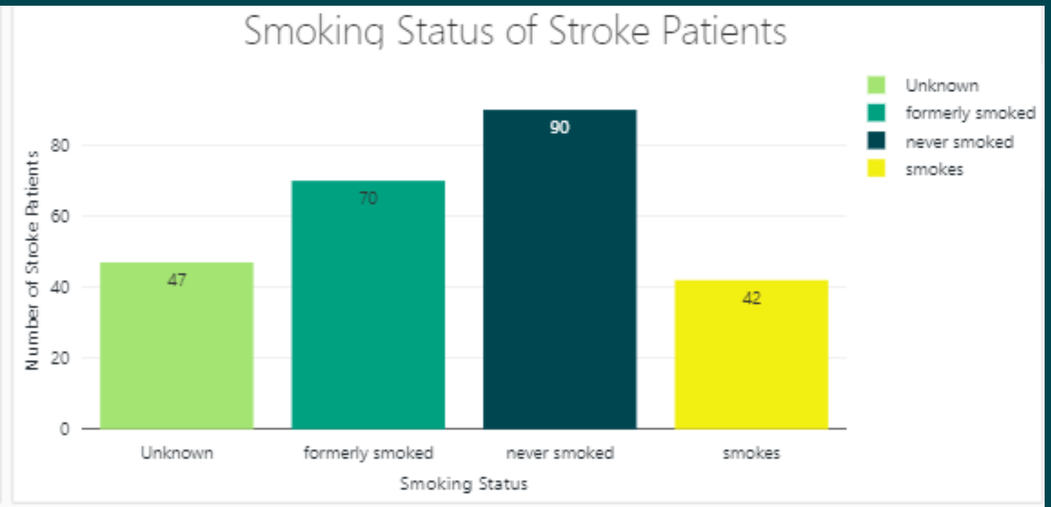


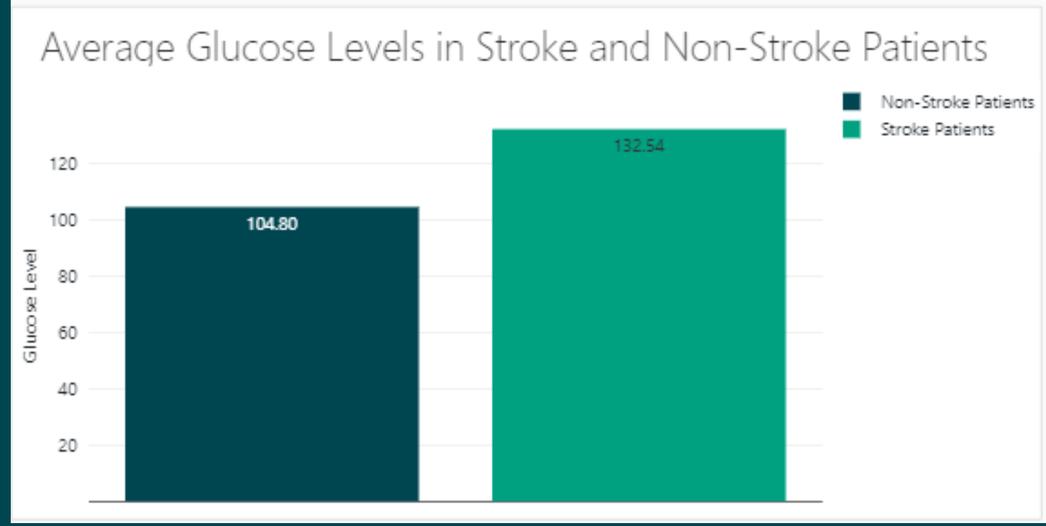


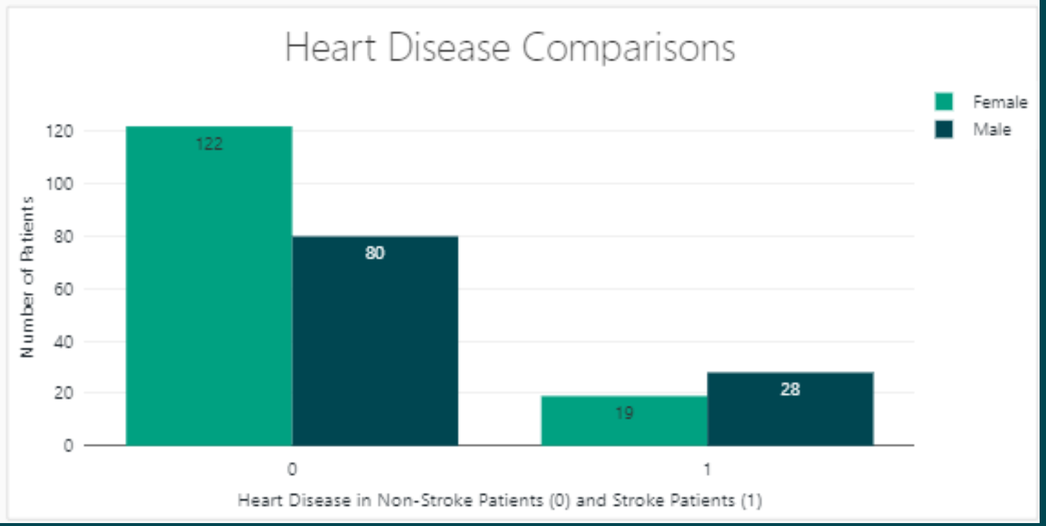


Visualizing Stroke & Predictive Factors









Machine Learning Model

Processing of the Data

Run Initial Models

Logistic Regression

Neural Network

K Nearest Neighbors

Random Forest

Optimization & Resampling

Final Model

Linear Regression Classification Report

	precision	recall	f1-score	support
0	0.99	0.73	0.84	1174
1	0.12	0.78	0.20	54
accuracy			0.73	1228
macro avg	0.55	0.75	0.52	1228
weighted avg	0.95	0.73	0.81	1228

Summary of Findings

In summarizing our findings, we found that out of the 5,110 patients, only 249 had strokes, indicating a significant imbalance of non-stroke patients. We also noticed from our visualizations that there are many different factors that can contribute to the occurrence of strokes. This makes it difficult to predict when someone might have a 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