# Predicting Strokes from Medical and Lifestyle data

Lukas Kania

DATA 1030 Fall 2021: Andras Zsom

Brown University - October 13, 2021

https://github.com/lkania1/data1030project

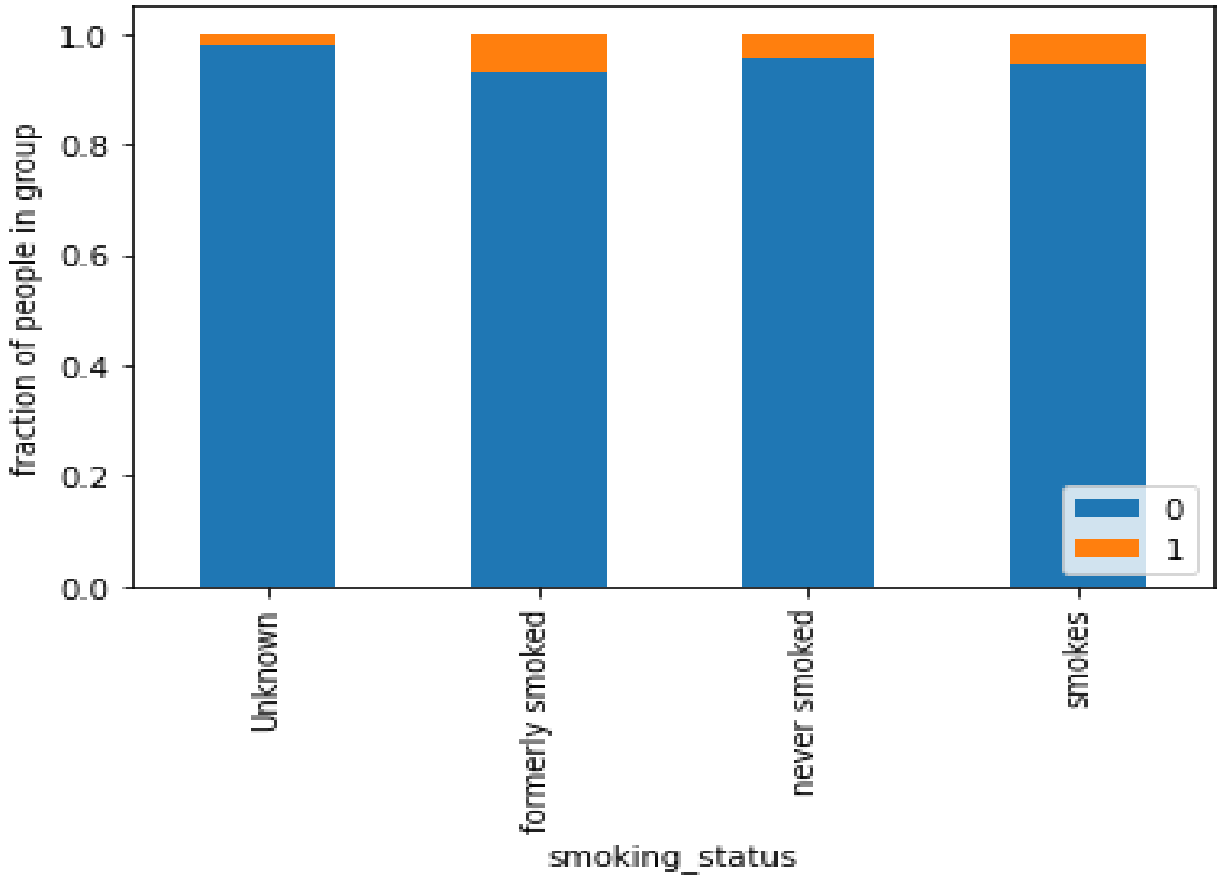
# Project Motivation

- World Health Organization states that Stroke is the second highest cause of death in the world (11% of deaths).
- Goal is to predict stroke outcomes of patients
  - Classification (1 = had a stroke, 0 = did not have a stroke)
- Accurate model could help provide insights to future patient's health outcome.

# The Dataset:

- Kaggle dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- 5,110 patients, with 10 different feature columns
- **Target Variable:** Stroke outcome
  - 1 = Had a Stroke
  - 0 = Did not have a Stroke
- **Problems:** Missing values for feature columns (bmi)

# Stroke Outcome compared with Smoking Status

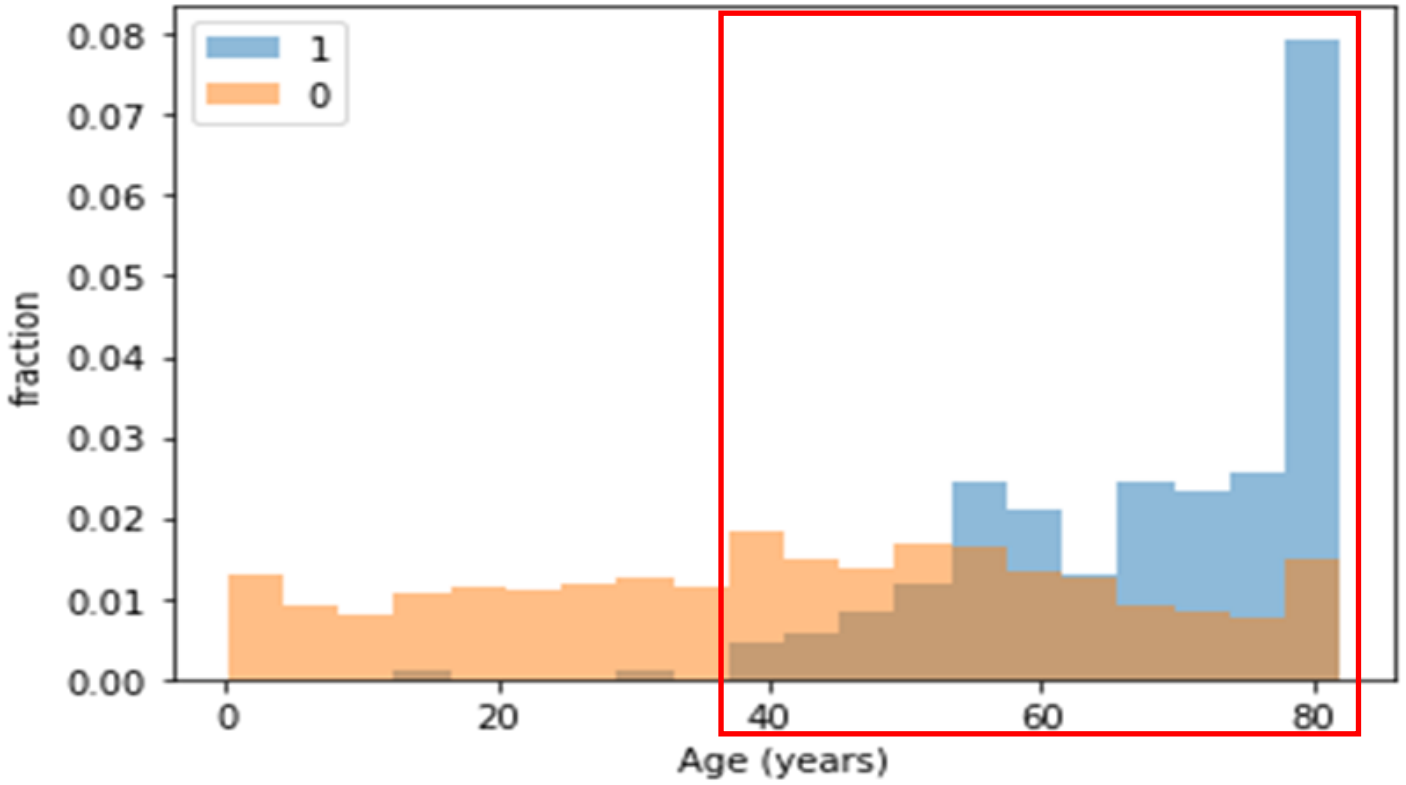



**Is smoking a strong indicator of positive stroke outcome?**

Expected to have stronger fractions for stroke positive patients.

“Unknown” status indicates information was not provided for given patient

# Patient compared to their risk of having a stroke

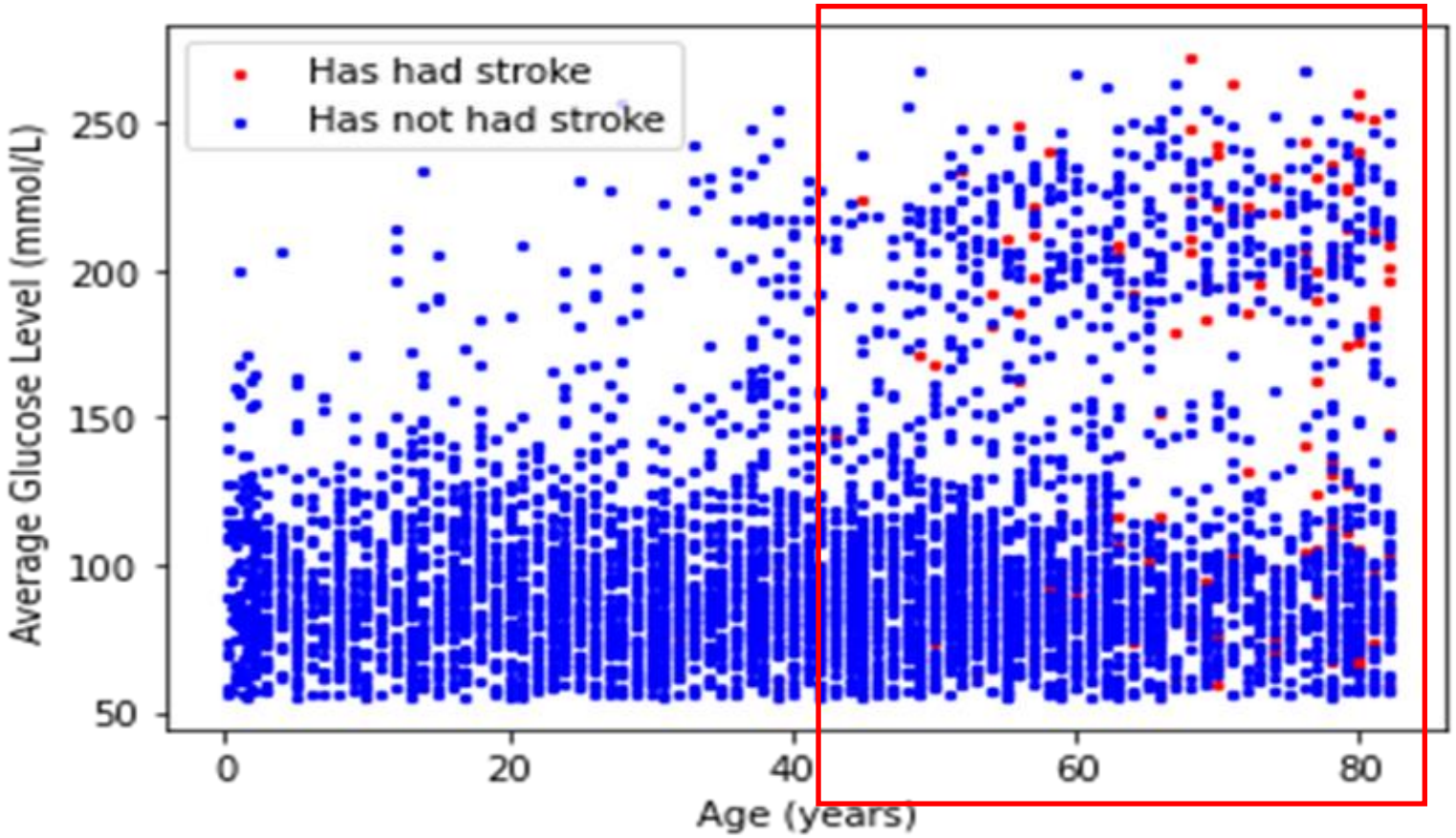



**Does age play a role in the likelihood of having a stroke?**

Age does play a role in having a stroke

Patients that did have a stroke are primarily above 40 years old.

# Age compared with Average Glucose Level with Stroke outcome distinguished



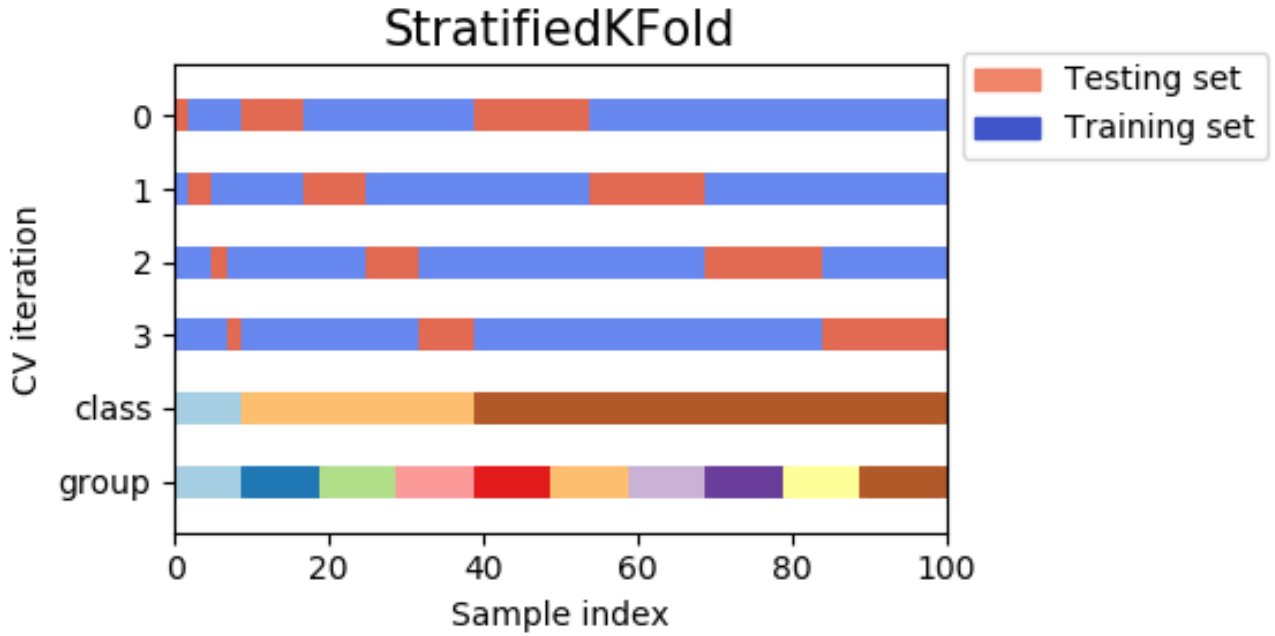

**What does multiple features say about stroke outcome?**

As discussed previously, age does impact stroke outcome.

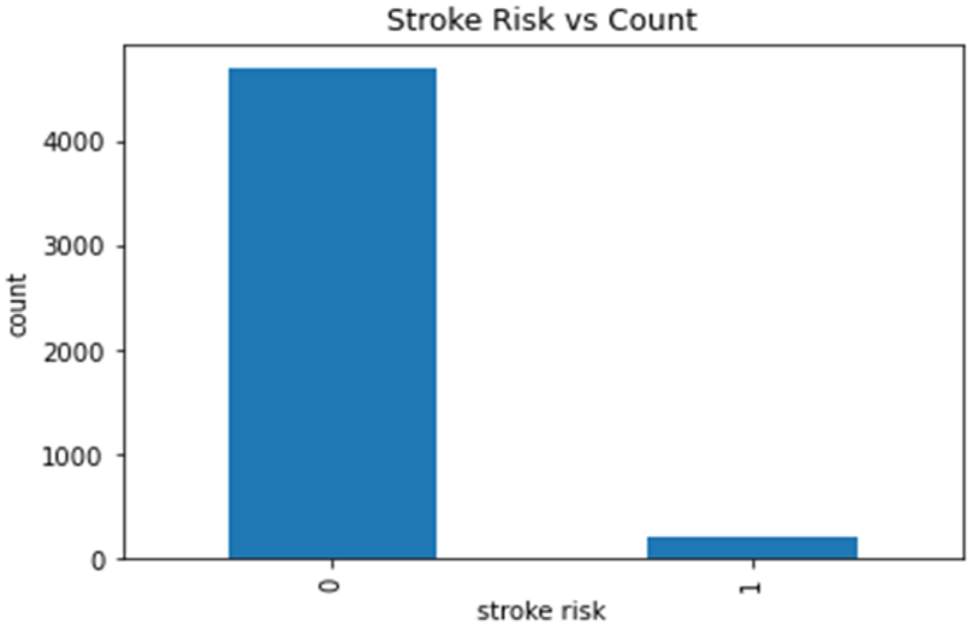
Average Glucose level does not provide as strong of insights.

# Splitting the Data:

- Unbalanced data (209 Patients had a stroke. 4901 patients did not.)
- Dataset is IID, therefore Stratified Kfold Split was utilized.
- 80/10/10 train/test/val split



# Data Preprocessing:




**StandardScaler:** Age

Continuous, roughly normally distributed

**OneHotEncoder:** work_type, smoking_status, gender, ever_married, and Residence_type

Categorical, unrankable values

**Already Processed:** Hypertension, Heart_Disease, Avg_glucose_level, Bmi

Boolean (1 or 0) values or Float values

**Missing Data:** Body Mass Index contained 201 missing values which were eliminated.

4.26% of dataset

# Questions?

Thank you!