Stroke Outcomes Prediction Model

## Dataset

For this project, our group is using a dataset related to patient stroke outcomes to train and compare several binary classification machine learning models to predict stroke outcomes based on input variables. We will train a variety of different model types, using several different data preprocessing techniques for each one and compare the accuracy and effectiveness of each model. The best performing ones will be selected for presentation on the project's dashboard including the ability to try out the model’s prediction with user generated input variables, as well as a presentation of some of the findings from the dataset in regards to stroke outcomes broken down by data categories.

The dataset contains a combination of medical data and non-medical lifestyle data. We intend to create a predictive model using the entire dataset, but also compare and contrast predictions when using only medical data, or only non-medical data, as well and identifying which data categories seem to be predictive that we would not have expected.

Of course, this model is not intended for any medical use, but merely high-level analysis of the limited data we have available, to find if there are any interesting correlations for stroke outcomes within the data categories of this set.

### Questions

Can this dataset be used to accurately and correctly predict whether a patient will have a stroke?

Does where you live correlate to stroke outcome?

Does where you work correlate to stroke outcome?

Does your marital status correlate to stroke outcome?

How do the non-medical factors impact our model’s accuracy, precision, and recall?

### Data Used/Exploration

Kaggle Stroke Prediction dataset: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

The original dataset has twelve columns of which 11 are features we’re using to build our model:

ID - patient ID - integer

Gender - Male/Female/Other - object

Age - Patient Age - integer

Hypertension - 1/0 - integer

Heart Disease - 1/0 - integer

Ever Married - Yes/No - object

Work type - Private/Self employed/Government Job/Never worked/children - object

Residence type - Urban/Rural - object

Average Glucose Level - integer

BMI - integer

Smoking status - Smokes/Formerly smoked/Never smoked/unknown - object

Stroke - 1/0 - integer

The BMI category has 201 null values. Rows with these null values were removed from the dataset.

The numerical categorical data was reclassified as objects.

BMI outliners (rows with BMI greater than 50 were removed).

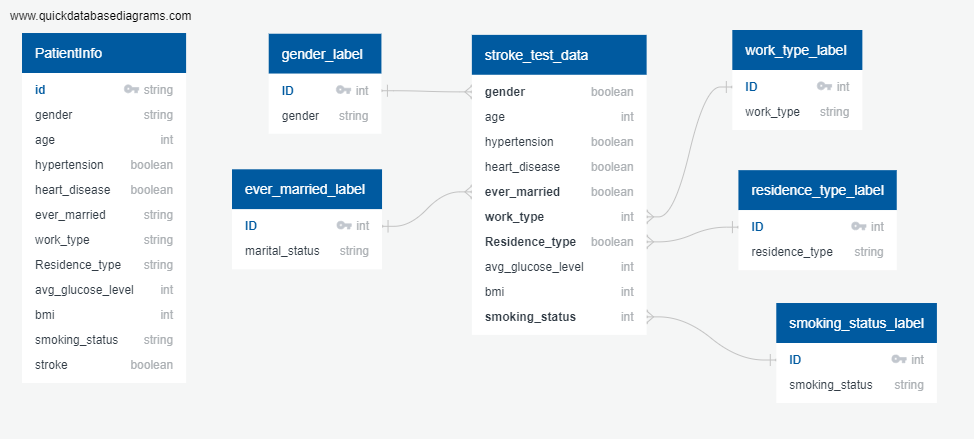
### Technologies Used

#### Data Cleaning and Analysis

Pandas will be used to clean the data, split the data (training and tsting) and Further data analysis will be completed using Python.

#### Database Storage

We’ll be using PostGres SQL database with AWS. Database ERD:

****

**Machine Learning**

SciKitLearn & Tensowflow are the ML library we'll be using to create a classifier. Our training and testing setup is over/undersampled to mitigate an umbalanced dataset.

**Dashboard**

Tableau visuals and story telling functionality will be used to provide user interactivity and identified insights.

**Stretch Goal**

Set up a webpage using Javascript forms to allow users to enter their medical, personal and lifestyle data to how likely the would be to have a stroke.