# Dependencies:

* Python
* Pandas
* Numpy
* Matplotlib
* Scikit-Learn
* Imbalanced-Learn

# Preprocessing:

During data evaluation of the dataset, we observed it contained NaN, null values, categorical variables. During our data cleaning we removed the NaN and null values. Knowing some models, such as neural networking, cannot handle categorical variables in their raw form, we chose one-hot encoding for our preprocessing technique for all models used in the prediction.

One-hot encoding will identify all unique column values and will split the single categorical column into a series of columns. Each individual column will contain information about a single unique categorical value.

We then split our data into the features and target arrays. After splitting the data we use “StandardScaler” to scale or standardize the data. We then chose decision tree as our classifier before fitting or training the data.

# Model:

For the most accurate model to predict stroke we chose to use Logistic Regression