| STROKE PREDICTION | **Nomesh Kumar** |
| --- | --- |

# Context

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes.

A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications.

# Algorithm that mentioned in proposal but not used in project

* Naïve Bayes

I did not used naïve bayes because naïve bayes work good on data where input features are in the of categorical variable but in my data, there are certain column which has numerical data.

# Preprocessing

* Column dropped

I have dropped one column “id” from my dataset because this column is not contributing any information in predicting whether a patient will suffer from stroke or not due to the fact that “id” column has a unique value for each row.

* Filling null values

I have only one column in my dataset in which there are around 200 null values. I have filled these null values with mean of column.

* Column renaming

I have renamed one column “Residence\_type” of my dataset because all other column name were in small letters but this column start from a capital letter so I renamed it.

* One-hot-encoding

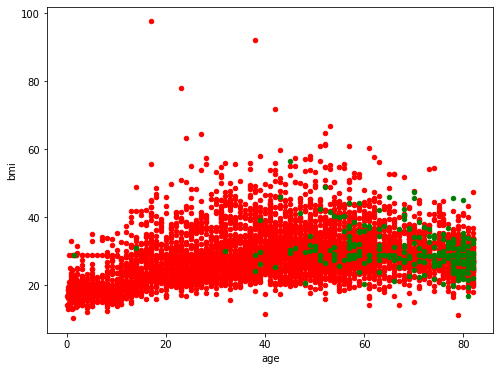
I have applied one-hot-encoding technique on categorical columns of my dataset, because to apply classification algorithm on dataset it is necessary to convert it into numerical column instead of strings values.

* Over-sampling

I had class imbalance problem in my dataset so I apply one of over-sampling technique( SMOTE ) so that my data may contains equal sample of both classes.

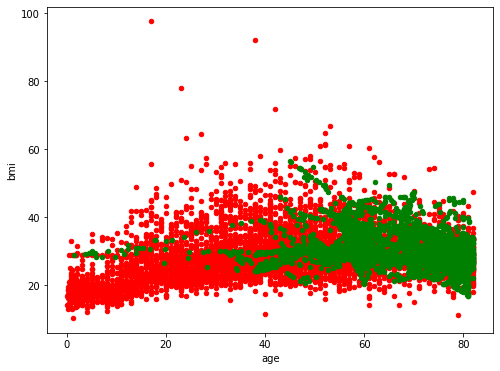
# **Before oversampling**





# **After oversampling**





# Algorithms to be used for prediction:

* LOGISTIC REGRESSION

Train accuracy= 0.9603

Test accuracy= 0.9565

* SVM(Kernel=Linear)

Train accuracy= 0.9609

Test accuracy= 0.9578

* DECISSION TREE

Train accuracy= 1.000

Test accuracy= 0.9328

* RANDOM FOREST

Train accuracy= 0.9944

Test accuracy= 0.9654

* NEURAL NETWORK

Train accuracy= 0.9724

Test accuracy= 0.9595

Technique used for selecting number of estimators for random forest:

I have used different number of estimators using for loop and then analysis train and test error using plotting and then select number of estimator for my model.

