Stroke Prediction by Machine Learning

# **Saurabh Biswas DSC680 T301** <https://github.com/saurabhbiswas1985/DSC680>

# Any surprises from your domain from these data?

I’m using data from healthcare domain. My dataset contains details of patients vital stats like age, gender, bmi, hypertension etc. and stroke event. The aim of this project is to build an effective model to detect high risk patients who have more chances of having stroke. They can take proper precaution and address the concerns and avoid this dangerous health problem.

This dataset contains details about 5110 patients out of which 249 patients had stroke which is about 4.87%. I can see this dataset highly imbalanced. I was expecting a more balanced dataset.

Working with imbalanced dataset has its won challenge. It requires different metrics to compare the efficiency between different models. Building a model that always predicts ‘no stroke’ would achieve 95.13% accuracy. But that won’t solve the business need.

# The dataset is what you thought it was?

I wanted to gather the information regarding patient’s vital stats like age, hypertension, heart disease, bmi, smoking status etc. My aim is to find the impact of these attributes on stroke. My collected dataset contains 5110 patient’s information.

One attribute ‘id’ is irrelevant for my research as it is only used to uniquely identify patients. Other than that, the following attributes has some sort of impact on stroke:

1. gender: "Male", "Female" or "Other"
2. age: age of the patient
3. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4. heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5. ever\_married: "No" or "Yes"
6. work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
7. Residence\_type: "Rural" or "Urban"
8. avg\_glucose\_level: average glucose level in blood
9. bmi: body mass index
10. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

I don’t see any missing data in any of these attributes except for bmi, which has 201 missing entries. So overall, this dataset is good to use for carrying out subsequent analysis.

# Have you had to adjust your approach or research questions?

This dataset is highly imbalanced in nature. The minority class is harder to predict because there are few examples of this class, by definition. This means it is more challenging for a model to learn the characteristics of examples from this class, and to differentiate examples from this class from the majority class.

I’m planning to oversample the minority class to get a balance, before I train any model.

My plan was to use Artificial Neural Network. But this dataset is not very large. Generally, ANN training requires large datasets. So, in this case, ANN may not provide competitive performance. I will stick to logistic regression and random forest.

After plotting the heatmap, I observed that bmi and gender has very little correlation with Stroke outcome. I was expecting ‘bmi’ will have greater influence on stroke prediction. But based on my EDA result, I will ignore these two attributes.

# Is your method working?

I used pandas profiling to get an overview of the data. I saw ‘bmi’ attribute has 201 missing values. This dataset contains small number of records and dropping missing bmi data records will reduce the record count further. So, I replaced them with mean bmi instead of dropping these records.

I have changed values of few attributes such as gender, ever\_married, work\_type etc from string to number, but kept the datatype as categorical. This will help me in further model training.

From correlation heatmap, I have identified key attributes that have higher impact on stroke outcome such as age, hypertension, heart\_disease, avg\_glucose\_level. I will pay more attention to these fields for further analysis.

Boxplot revealed outliers in avg\_glucose\_level and bmi. I have removed outliers from avg\_glucose\_level using z\_score.

I will oversample minority class i.e., patients with stroke to geta more balanced dataset. Then I will split the dataset into test and training dataset and use the training dataset to train my models.

Logistic Regression is ideal for classification problems like true/false, stroke/ no-stroke, yes/no. It will provide me good baseline.

Random Forest classifier is also very good a classification prediction problem and free from any overfitting problem. As outcome of RF ensemble model is a probability score, I have a plan to calibrate the results to fit true yes/no classification problem.

I would like to compare the performance of RF with LR.

# What challenges are you having?

The research dataset contains only 5110 patient’s data. Training an ANN model requires large datasets, otherwise the performance of this model won’t be good. As I have a small training dataset, I could impact ANN model performance. So, I won’t be using ANN model. I need to find a relevant classification model beside LR and RF and train that to compare the performance.

I have found that this dataset is highly imbalanced. Based on my research, imbalanced classification problem has challenges in training the model successfully. My solution is to use over-under sampling. But I haven’t used this technique before. So, it will be a new learning for me.