Stroke Prediction by Machine Learning

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

# Which Domain?

I’m planning to use data from healthcare domain. I will gather details about patient’s history and whether he/she had stroke or not.

Stroke is the one of the leading causes of death according to the World Health Organization (WHO) and CDC data. According to WHO, almost 11% of total deaths are caused by stroke. CDC data suggests that, in United States, someone has a stroke in every 40sec, and one person dies in every 4 minutes because of stroke. From these statistics, we can understand stroke is a major health issue around the world.

Machine Learning can examine patient’s data, draw a pattern and identify major attributes that can end up causing stroke.

Below are some statistical sources of stroke related medical data and some research work done in this area to predict stroke based on various attributes using machine learning models.

CDC website: <https://www.cdc.gov/stroke/facts.htm>

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

Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 183-192).

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3), 1350-1371.

Lehmann, J. F., DeLateur, B. J., Fowler Jr, R. S., Warren, C. G., Arnhold, R., Schertzer, G., ... & Chambers, K. H. (1975). Stroke rehabilitation: Outcome and prediction. Archives of Physical Medicine and Rehabilitation, 56(9), 383-389.

Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., & Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3110-3113). IEEE.

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial intelligence in medicine, 101, 101723.

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263-1265.

Lin, C. H., Hsu, K. C., Johnson, K. R., Fann, Y. C., Tsai, C. H., Sun, Y., ... & Taiwan Stroke Registry Investigators. (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Computer Methods and Programs in Biomedicine, 190, 105381.

Singh, M. S., Choudhary, P., & Thongam, K. (2019, September). A comparative analysis for various stroke prediction techniques. In International Conference on Computer Vision and Image Processing (pp. 98-106). Springer, Singapore.

# Which Data?

I will be using stroke related dataset healthcare-dataset-stroke-data.csv from Kaggle.

Link to the dataset: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

It contains 11 input attributes and 1 output attribute that confirms whether the patient had stroke or not (1 – stroke, 0 – no stroke). Among these 11 input attributes, ‘id’ column identifies the patient uniquely and this doesn’t provide any value in predicting stroke.

Other attributes may have some impact on stroke. These attributes are:

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"\*

# Research Questions? Benefits? Why analyze these data?

Analyzing attributes from patient’s data can help us isolate high risk group of people who are more prone to stroke. With the help of machine learning we can establish a pattern between various input attributes such as age, bmi, smoking status etc. and outcome variable stroke.

This dataset contains mixed data – patients with stroke and patients without stroke. So, it provides us a good mixture. Also, it contains most of the attributes those are responsible for stroke. My aim is to find one/two key attribute from these and using that predict stroke in future.

I will start with cleaning the dataset. Then I will find correlation between input attributes and output variable. Once I have the key attributes, I will build a model and train using these key inputs and output data.

# What Method?

I will use correlation heatmaps to find correlated inputs and keep only one correlated field. Also, using this plot, I will take the most important attributes those have higher correlation values.

Once, I have this cleaned and EDA transformed dataset, I will start building models. As this is a classification problem, I’m planning to use Logistic Regression, Random Forest and Artificial Neural Network.

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.

ANN acts like a human brain. ANN classifier can analyze the attributes to find a pattern after passing the inputs through input/hidden/output layers and adjusting the weights. I like to compare the performance of ANN with RF and LR.

# Potential Issues?

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.

# Concluding Remarks

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. I believe early detection of high risk people will give them the opportunity to fix their lifestyle and take control of the important attributes that can end up causing stroke.