**Topic** - Stroke Prediction using Machine Learning

**Introduction**

Stroke is a severe cerebrovascular disease caused by an interruption of blood flow from and to the brain. As a direct consequence of this interruption, the brain is not able to receive oxygen and nutrients for its correct functioning. The other way around, the brain is not able to drain and expel through blood vessels all of its waste, like dead cells. In a question of minutes, the brain is in a critical condition as brain cells will imminently begin to die.

The American Stroke Association indicates that stroke is the fifth cause of death and disability in the United States. For this reason, stroke is considered a severe disease and has been the subject of extensive research, not only in the medical field but also in data science and machine learning studies.

**Dataset Description** -

The dataset used in this article contains 5110 records of patients. Each patient has 12 columns each referring to a concrete attribute. Most of these attributes correspond to medical records or the results of clinical trials. Some of the key attributes are hypertension, heart diseases, average glucose levels in the blood, and body mass index (BMI). As we can observe from these first attributes, the dataset provides relevant data regarding the likelihood of patients suffering from stroke disease. It is easy to understand that a patient with high glucose levels and BMI, who has suffered from heart diseases and/or hypertension, is more likely to suffer from stroke. In fact, stroke is also an attribute in the dataset and indicates in each medical record if the patient suffered from a stroke disease or not.

**Data Dictionary** -

a)Id - Unique Id
b)gender
c) age
d) hypertension - Binary Feature
e)Heart_disease - Binary Feature
f)Ever_married - Has the patient ever been married
g)Work_type - Work type of a patient
h)Residency_type - Residency type of patient

i)avg_glucose_level - Average Glucose level in the blood
j)bmi - Body mass Index
k)smoking_status - Smoking status of the patient
l)stroke - stroke event

**Questions** -

1. Read the dataset and view the first 10 rows of it.
2. Check the shape/dimension of the dataset
3. Check for the missing values.Display number of missing values per column.
4. Investigate and predict the missing BMI Value.
5. Check the datatype, number of non null values and name of each variable in the dataset.
6. Check the descriptive statistics of the dataset.
7. Visualize the proportion of Stroke samples in the dataset.
8. Visualize the Distribution of Male and Female Ages. Write the Observation.
10. Visualize the stroke sample based on
   a) BMI and Glucose Level
   b) BMI and Age

11. Using the pie chart visualizes the proportion of different smoking categories among the stroke population.
12. Perform hypothesis testing to find the significant variables.
13. Drop the unnecessary columns.
14. Write the code to replace following categories columns in integer format as follow –
   a) work_type('Private':0,'Selfemployed':1,'Govt_job':2,'children':1,'Never_worked':-2)
   b) ever_married('Yes': 1, 'No': 0)
   c) smoking_status('never smoked':0,'Unknown':1,'formerly smoked':2,'smokes':-1)
15. Check the distribution of 'bmi' and 'stroke' columns in the dataset.
16. List down columns that are highly skewed.
17. List down the columns that are highly kurtosis.
18. Find the distribution of all variables with respect to the outcome 'stroke' variable.
19. Plot the heatmap for correlation matrix for the given dataset. Write the observation. Especially note down columns that are highly correlated (Positive or negative correlation, consider 0.7 to 1 as high)
20. Split the dataset randomly into train and test dataset. Use a train ratio of 70:30 ratio.
21. Check the dataset is balanced or imbalanced. If it is highly investigated, a different approach to balanced the dataset by using the correct technique.
22. Model Selection/hyperparameter tuning
- Try different models and fine tune their performance until you get the desired level of performance on the given dataset.

Model Evaluation
- Evaluate the models using appropriate evaluation metrics.