**DATS6103 - INTRODUCTION TO DATA MINING  
PROJECT PROPOSAL**

**Title:** STROKE PREDICTION

**Team Members:** Aswin Balaji Thippa Ramesh, Rahul Arvind, Gowri Sriram Lakshmanan, Barathkumar Anantharaj

**Research Topic:**

This project aims to analyze the relationship between demographic, lifestyle, and medical factors and the likelihood of stroke occurrence. By utilizing the "Stroke Prediction Dataset," this study seeks to identify key predictors of stroke and assess their significance, enabling insights that could aid in early intervention and health risk management.

**Research Questions:**

The following questions are the main subjects which this project focuses on,

1. What are the key factors most strongly associated with the occurrence of strokes?
2. How does each predictor's distribution differ between different stroke cases ?
3. Can we determine statistically significant correlations or patterns in stroke occurrence based on categorical variables?
4. Can we identify trends or variations in stroke risk based on time-related factors like age?
5. How accurate is it on building a machine learning model predict stroke occurrence based on the dataset?

**Source of Data Set(s):** We are using the Stroke Prediction Dataset from Kaggle, accessible at [Kaggle - Stroke Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset).

This dataset contains 5110 observations, with each

entry detailing demographic, medical, and lifestyle factors such as age, gender, hypertension, heart disease, smoking status, and BMI, alongside the `stroke` variable, which indicates whether the individual has had a stroke.

**Github Repository**: Our team’s GitHub repository will serve as the central platform for collaboration and code sharing throughout the project: [Stroke Prediction](https://github.com/AswinBalajiTR/DATS6103-Project)

**Modeling Methods:**

The following modeling methods are proposed:

* **Logistic Regression:** Insights via coefficients.
* **Ensemble Methods:** Interpretability via Gini Index/Entropy, accuracy via ensemble learning, and top predictors via feature importance.
* Compare models using Accuracy, Precision, and Recall.