**DATS 6103 – Introduction to Data Mining**

**Project Proposal**

**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**

1. What demographic and lifestyle factors (e.g., age, gender, smoking status) are most strongly associated with the occurrence of strokes?
2. How do medical conditions (e.g., hypertension, heart disease) influence the risk of a stroke?
3. Can a machine learning model be developed to predict stroke occurrence based on the dataset, and how accurate is it?

**Dataset Source**

We are using the Stroke Prediction Dataset from Kaggle, accessible at [Kaggle - Stroke Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download). 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: [GitHub - DATS6103 Project](https://github.com/AswinBalajiTR/DATS6103-Project)

**Modeling Methods**

The following modeling methods are proposed:

* **Exploratory Data Analysis (EDA**): Descriptive statistics, correlation analysis, and visualization to explore data patterns.
* **Preprocessing:** Handling missing values (e.g., bmi), encoding categorical variables, and scaling numerical data.
* **Model Building:**
  + Logistic Regression: To estimate the likelihood of a stroke.
  + Random Forest Classifier: For feature importance and predictive accuracy.
  + Support Vector Machines (SVM): To evaluate classification performance.
  + Neural Networks (optional): For advanced predictive modeling.

**Team Members**

Ashwin Balaji Thippa Ramesh, Rahul Arvind, Barath Kumar Ananthraj, Gowri Sriram