Introduction to Data Mining (DATS 6103)

Team 2

GitHub repo: <https://github.com/QI-Xiao/T2_6103>

Data source: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Project Proposal

Title: Stroke Prediction based on person’s lifestyle

RESEARCH TOPIC:

According to the World Health Organization (WHO), stroke is the second highest cause of death around the world, leading to 11% of total deaths.

This dataset helps predict whether a patient is likely to have a stroke based on the input parameters like gender, age, various diseases, and smoking status, among others. Each row in the data provides relevant information about the patient’s lifestyle factors that may pertain to their health and the possibility of experiencing a stroke.

SMART QUESTIONS:

Overall

* What factors or variables affect or increase the likelihood of a person having a stroke?

EDA

* Is a particular gender affected more by heart disease or hypertension?
* Do marriage status and residence type contribute to the development of stroke symptoms/having a stroke?

METHODS:

The target variable in the dataset is highly unbalanced, with 5% of the participants having a stroke and the other 95% not having a stroke. To tackle this problem, we have decided to use cut-off values to decide on the favorable train-test dataset for the project.

EDA will consist of plotting graphs using Python libraries to find any relation between the possibility of a stroke with other variables.

Since the majority of the variables in the data set are categorical variables, our main focus for modeling is to use logistic regression. This also helps as the target variable is binary.

DESCRIPTION OF AVAILABLE DATA:

The source of the dataset is Kaggle Sample Dataset where it was extracted in a CSV

format. The data consists of 5110 observations and 12 variables:

ID,

Gender: Male or Female,

Age: Age of the participant,

Hypertension: Whether the participant suffers from hypertension,

Heart Disease: Whether the participant has heart disease,

Marriage Status: Is the participant married or not,

Work Type: Private, self-employed, or other,

Residence Type: Urban or Rural living,

Avg. Glucose level: glucose level of participant,

BMI: body mass index of the participant,

Smoking Status: whether the participant smokes or not,

Stroke: Did the participant suffer from a stroke or not