Final Project

1. Title: Stroke prediction data

2. Project title: Predict the ratio to become stroke patient in male depend some factors like age, heart disease, etc

3. Purpose and Outcome:

3.1. Purpose: To predict whether a male patient is likely to get a stroke based on the 5 important input parameters like age, avg\_glucose, bmi, heart\_disease, hypertension.

3.2. Outcome: A model that can predict stroke prediction with high accuracy, along with actionable insights to reduce Stroke rates. So that can forecast patent can be stroke or not.

4. Dataset:

4.1. Source: Public dataset from Kaggle (Stroke prediction data:https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data)

5. Description: According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to Each row in the data provides relevant information about the patient.

6. Structure:

* Id: unique identifier
* Gender: "Male", "Female" or "Other"
* Age: age of the patient
* Hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
* Heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
* Ever\_married: "No" or "Yes"
* Work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
* Residence\_type: "Rural" or "Urban"
* Avg\_glucose\_level: average glucose level in blood
* Bmi: body mass index
* Smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
* Stroke: 1 if the patient had a stroke or 0 if not

7. Initial Analysis Plan:

* Import necessary libraries, import data
* Data Cleaning: Handle missing values, correct data types, and remove duplicates. Moreover, execute the IQR to eliminate the outliers data.
* EDA: Finding the relationship between 5 factors that interact with each other and affect the stroke by Power Bi.
* Analysis: Perform correlation analysis, build and evaluate a logistic regression model, and identify key factors contributing to predict stroke by seaborn, numpy, matplotlib.
* Data Storytelling: Present the findings with a narrative explaining the key which affect stroke disease ratio, and give the conclusion what we need to reduce stroke disease.