**Project stage I report**

**Dataset Selection & Project Setup**

Group # 10

Hung Nguyen

Thanh Van Thai

Yajaira Alonso-Camarillo

Terrance Lee

Project title/topic: Stroke Prediction

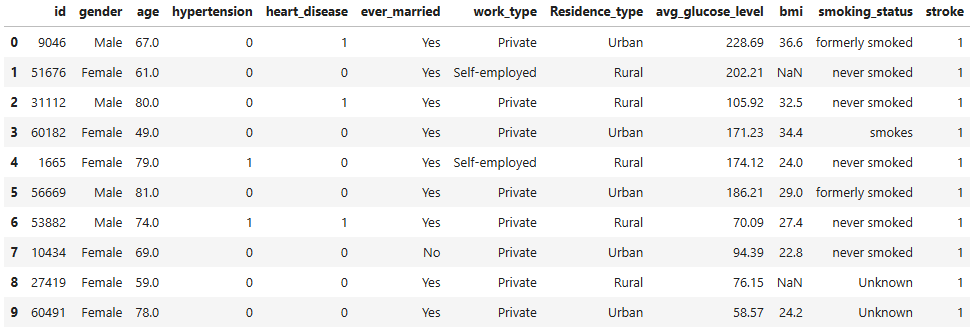
Dataset source (url or resource): [Stroke Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

**Task 1 Problem framing (10 points)**

According to the World Health Organization (WHO), stroke is the second leading cause of death globally, accounting for approximately 11% of total deaths. This project aims to build a classification model that predicts the likelihood of stroke based on patient health data. The primary goal is to identify individuals at elevated risk by using features such as age, hypertension, heart disease, BMI, and glucose levels.

Healthcare providers and public health researchers can benefit from this study by gaining a tool that supports early intervention and helps prioritize resources effectively. The dataset is well-suited for this task, as it contains more than 5,000 entries and includes a mix of numerical and categorical features that reflect real-world risk factors.

**Task 2 Dataset exploration (20 points)**

Show the first 10 rows of your dataset; create a dataset dictionary:  
  
  
  
  
  
**First 10 rows of dataset:**  
  
  
**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Type: Int / float /… | examples | Missing % |
| Gender | object | “Male”, “Female”, or “Other” | 0% |
| Age | float64 | 67.0, 61.0, 80.0 | 0% |
| Hypertension | int64 | 0 , 1 | 0% |
| Heart disease | int64 | 0 , 1 | 0% |
| Ever married | object | Yes, No | 0% |
| Work type | object | Private, Self-employed | 0% |
| Residence type | object | Urban, Rural | 0% |
| Average glucose level | float64 | 228.69, 202.21, 105.92 | 0% |
| BMI | float64 | 36.6, 32.5, 34.4 | 3.93% |
| Smoking status | object | Formerly smoked, never smoked, smokes | 0% |
| Target variables: Stroke | int64 | 0, 1 | 0% |

**Task 3 Feature exploration (40 points)**

Deal with the missing values/outliers; and then create a clean dataset for the next stage analysis.

**Summary:** The dataset consists of 11 distinct features including 8 categorical features: +gender, hypertension, heart\_disease, ever\_married, work\_type, residence\_type, smoking status, stroke. And 3 numerical features:

+age, avg\_glucose\_level, bmi. The following members will be responsible for the features mentioned above

**+Thanh Van Thai: age**

**+Hung Nguyen: bmi**

**+Yajaira Alonso-Camarillo: smoking\_status**

**+Terrance Lee: avg\_glucose\_level**

**Task 4 Save cleaned data (10 points)**

After processing, the final data frame is saved to newdata.csv in Data folder.