**Stroke Prediction Research Project**

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 predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relavant information about the patient.

### **Attribute Information**

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever\_married: "No" or "Yes"
7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
8) Residence\_type: "Rural" or "Urban"
9) avg\_glucose\_level: average glucose level in blood
10) bmi: body mass index
11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
12) stroke: 1 if the patient had a stroke or 0 if not

\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

Q1: Perform the descriptive analysis to explain and analyse all variables in the “healthcare-dataset-stroke-data.csv”:

Table1

|  |  |  |
| --- | --- | --- |
| Variable | *n* | *%* |
| gender |  |  |
|     Male |   |   |
|     Female |   |   |
| ever\_married |  |  |
|     Yes |  |  |
|     No |  |  |
| work\_type |  |  |
|     Private |  |  |
|     Self-employed |  |  |
|     Govt\_job |  |  |
|     children |  |  |
|     Never\_worked |  |  |
| Residence\_type |  |  |
|     Urban |  |  |
|     Rural |  |  |
| smoking\_status |  |  |
|     formerly smoked |  |  |
|     never smoked |  |  |
|     smokes |  |  |
|     Unknown |   |   |

Table 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | *Mean* | *SD* | *n* | Min | Max |
| age |  |  |  |  |  |
| hypertension |  |  |  |  |  |
| heart\_disease |  |  |  |  |  |
| avg\_glucose\_level |  |  |  |  |  |
| bmi |  |  |  |  |  |
| stroke |  |  |  |  |  |

Q2: Find the association between “stroke” and other Variables except the ID using CrossTab

*Note: Copy the the results table and paste it here.*

Q3: What is the conclusion from the results in the CrossTab?

Q4: How the other factors except the ID influencing and predicting the “stroke”, use the two outcome Logistic regression, what is the main finding from the Logistic regression results. (Please interpret and justify your results)

Q5: Priorities the effects from the other variables on the “stroke” (the most three risk factor using the Estimate information from the Logistic regression table)

Q6: How much the accuracy of the estimated Logistic regression model that predicting the effect from other factors on “stroke”.

Q7: Create Data Studio to present the most important results (note: paste the link)