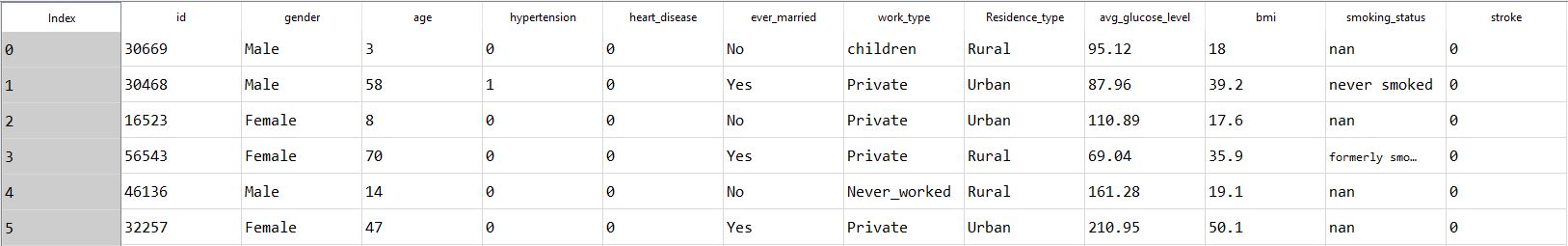
**Capstone Project: Stroke Patients Classification Modeling**

**Phase I: Data Wrangling**

Initially, train set needs to be cleaned and formatted in a right format (i.e., data types) for analysis. Dataset gets loaded and a few observations are printed to check data type on each column (i.e., feature). Currently, dataset presents a couple of problems. First, some categorical (i.e., nominal features) like “hypertension” and “heart\_disease” are encoded with 0 means “No” and 1 means “Yes” as shown in Table 1. Whereas same nominal feature ‘ever\_married’ contains value as a string “Yes or No”. Thus, it is important to format them in a consistent manner. Both hypertension and heart\_disease features’ values will be re-encoded as a string a string “Yes or No”. Second, smoking\_status is categorical but ordinal feature which means that unique values are ordered. It needs to be re-encoded as meaningful order like ‘never smoked: 0, ‘formerly smoked:1’ and ‘smokes:2’.

Third, there are missing values present in two features smoking\_status and BMI. Smoking\_status is a categorical feature and it is missing about 30.6%. Follow by BMI is a numerical feature where it is missing about 3.37% as shown in Table 2. Thus, missing values will be imputed using different basic statistic methodology. For numerical features like BMI will replaced missing values by computing a median value of BMI. Also, it could be replaced using a mean value, but median value is better as it less sensitive to provide biased results (i.e., due to presence of extreme values on that feature).

Finally, last step of checking any outliers on the data set will be performed. In training set, there are 3 independent numerical features (i.e., scale) in which are “age”, “bmi” and “avg\_glucose\_level”. In this analysis, age will be selected as a variable to detect any outliers and remove observations accordingly. This is purely based on earlier assumption that age maybe more significant feature in comparison to other two on having a stroke condition. Anyway, outliers will be ruled out by perfmoinig interquartile range (IQR) method. Where any values sit below and above the lower and upper bounds of IQR will be treated as outliers. Here are some equations related to IQR method. Where IQR = Q3 – Q1, lower bound equals Q1 – (1.5\*IQR) and upper bound equals Q3 + (1.5\*IQR).



**Table 1**. Top 5 observations of cardiac stroke training set dataframe before data cleansing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Data Type** | **Missing Value Counts** | **Missing Percent (%)** |
| smoking\_status | Object | 13292 | 30.6 |
| Bmi | Float | 1462 | 3.37 |
| Hypertension | Int | 0 | 0.0 |
| heart\_disease | Int | 0 | 0.0 |
| ever\_married | Object | 0 | 0.0 |

**Table 2**. A matrix table showing data type, missing value counts and percentage on five features

At last dataset is properly cleansed, each feature will be formatted based on their proper data type. Multiple lines of code will be executed to ensure nominal features as a category, encoded ordinal feature like “smoking\_status” as an integer and remaining numerical features as a float data type.

Now before we proceed to next phase, a cleaned dataset will be copied with two same dataframes. One will be primarily used for exploratory data analysis. While other one will be used for machine learning modeling step. For machine learning modeling, all nominal features (i.e., hypertension, heart\_disease, work\_type etc.) will be transformed into new dummy variable feature.

Dummy variable encoding is extremely crucial step for any machine learning (ML) modelling to intake as inputs. Since these features do not come in any ordered manner, ML model results can be inaccurate if these do not get encoded as a sparse matrix of features that contains (0 or 1). Also note that while performing dummy variable encoding, one of the first dummy feature needs to be omitted. This prevents having high multi-collinearity (i.e., high correlation) among dummy variables on dataset.