BUAN 6367 Project Report- Stroke Prediction

# Setting

The primary goal of this project is to develop a prediction model that can accurately determine whether a patient is likely to suffer a stroke based on various input parameters. Data exploration and model building will answer the following questions:

1. **How does the incidence of stroke vary based on the patient's smoking status? Are former smokers at a higher risk of stroke than current smokers or non-smokers?**
2. **Is there any difference in stroke incidence between rural and urban areas?**
3. **What is the distribution of patients with hypotension or heart disease? Is the incidence of stroke higher in patients with these conditions?**

# Data Description

**We have taken the data source from Kaggle (**[**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)**). This data source include 12 input parameters and over 5000 observations.**

|  |  |  |
| --- | --- | --- |
| **No.** | **Attribute** | **Description** |
| **1** | **ID** | **Patient unique identifier** |
| **2** | **Gender** | **“male”, “female”, or “other”** |
| **3** | **Age** | **Age of the patient** |
| **4** | **Hypotension** | **0 if patient doesn’t have hypotension, 1 if they do** |
| **5** | **Heart disease** | **0 if patient doesn’t have heart disease, 1 if they do** |
| **6** | **Marital status** | **“yes” or “no”** |
| **7** | **Work type** | **"children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"** |
| **8** | **Residence** | **“rural” or “urban”** |
| **9** | **Average 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 |

# Challenges

Major challenges faced in the project are:

Challenge 1: Gender feature with one occurrence of ‘Other’ class

One of the primary challenges is dealing with the "Other" class in the gender feature. To resolve this issue, the row containing the "Other" class is removed from the dataset, ensuring that only the male and female classes are retained.

Challenge 2: Occurrence of N/A values in BMI

Another challenge in the project is the presence of 201 N/A values in the BMI feature. These values can potentially skew the results of the analysis, leading to inaccurate predictions. To address this challenge, mean imputation is used to fill in the missing values.

Challenge 3: Imbalanced dataset

The occurrence of imbalanced data is another challenge in the project. The dependent variable stroke has a higher occurrence of 'No' than 'Yes', making it challenging to develop a model that could accurately predict the occurrence of stroke. To overcome this, we tuned hyperparameters with 5 fold cv to achieve optimal results for each model.

Challenge 4: Performance evaluation

Performance evaluation is a significant challenge in the project. The different models used, such as Logistic Regression, Decision tree, and random forest, preform differently on different evaluation metrices. For example, the Random Forest model worked better on accuracy, while Logistic Regression worked better on ROC. Since this is a classification problem, the logistic regression model is selected based on the ROC result.

# Analysis

The main goal of the project is to predict the occurrence of a stroke using machine learning models. To achieve this, dataset is and trained and evaluated on Logistic Regression, Decision Tree, and Random Forest. In addition, exploratory analysis and data cleaning is also performed to ensure the data was suitable for training the models.

Data Cleaning and Exploratory Analysis:

Before training the models, exploratory analysis and data cleaning is performed.

Data cleaning involves removing outliers, filling in missing values using techniques such as mean imputation and converting the categorical data into factors. Exploratory analysis involves visualizing the data using various plots such as bar graphs and box plots, to understand the distribution of the variables, identify outliers, and detect any missing values and finding the trend in data.

Model Training:

The first step in model training involves splitting the dataset into training and testing sets. The training set is used to train the models, while the testing set is used to evaluate their performance. The models are trained using hyperparameter tuning with 5-fold cross-validation. They affect the behavior of the model and can significantly impact its performance. This process of using hyperparameter helps to reduce the risk of overfitting and ensures that the model is not only performing well on the training set but also on the unseen data.

Model Evaluation:

After training the models, their performance is evaluated using two metrics accuracy and ROC-AUC. Accuracy measures the percentage of correct predictions made by the model. Accuracy can be misleading when the dataset is imbalanced. This is because the model can achieve high accuracy by simply predicting the majority class for all instances. ROC AUC, on the other hand, is a metric that measures the performance of a binary classifier over all possible thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different thresholds and calculates the area under the curve (AUC).

Accuracy achieved on different models :

Logistic regression: 0.9453905

Random Forest: 0.9483265

Decision Tree: 0.9465649

AUC- ROC on different models:

Logistic regression: 0.8468072

Random Forest: 0.8137218

Decision Tree: 0.7161431

# **Conclusion**

In conclusion, this project presented several challenges that requires innovative solutions to overcome. By conducting a detailed exploratory analysis, using appropriate imputation techniques, addressing the occurrence of imbalanced data, and utilizing multiple evaluation metrics, it was possible to develop a model that accurately predicted the occurrence of stroke.

* Through Exploratory Data Analysis it can be interpreted that Avg glucose level, BMI, Hypertension and heart disease are the biggest risk factors for stroke.
* From our three models i.e., Decision tree, Random Forest and Logistic regression , Random Forest gives best accuracy.
* Logistic Regression gives best ROC-AUC curve than Decision tree and Random Forest.
* Considering the significant class imbalance in the dataset, accuracy may not be an appropriate metric for evaluating the models.
* If the goal is to optimize the model's performance in identifying stroke cases, even at the cost of some false positives, ROC-AUC may be a more appropriate metric to use.