# Introduction

In this report, we are focusing on the important issue of predicting cerebral strokes. These strokes can have severe consequences, so it's crucial to detect and address them early. We are using a real-world dataset that has an uneven distribution of stroke cases and non-stroke cases.

Our main goals are to find out what factors are most connected to stroke occurrences in this dataset and to create a machine learning model that can effectively identify individuals who are at a higher risk of having a stroke.

We also want to see how well our model can do this, especially considering the uneven distribution of cases. Lastly, we'll discuss how doctors and healthcare providers can use our model to improve patient care and reduce the risk of strokes.

In simple terms, we're using data and technology to understand and predict strokes better, and we hope this will help healthcare professionals take better care of their patients.

## Data Preprocessing

In this data preprocessing step, you are dealing with a dataset containing 43,400 records and 12 column addressing missing values in the "bmi" and "smoking_status" columns.

Here's a breakdown of what you've done:

- DataFrame with 12 columns.
- There are missing values (NaN) in the "bmi" column, as indicated by the number of missing values in each column.
- no duplicated records in your DataFrame.
- created a copy of your DataFrame called df_imput to avoid altering the original data.

Use K-NN imputation to the "bmi" columns, filling in the missing values with estimated values. The result is an array with the imputed values for these two columns.

update the original DataFrame with the imputed "bmi" values by assigning the second column of the imputed_data array to the "bmi" column of DataFrame. This replaces the missing values in the "bmi" column with the imputed values.

### Predicting missing values in smoking status using Random Forest Classifier

1. Split the data into complete and incomplete records.
2. Select the features that will be used to predict the missing values.
3. Scale the features.
4. Train a Random Forest Classifier model on the complete data, using the selected features as the independent variables and the smoking status as the dependent variable.
5. Predict the missing values in the incomplete data using the trained Random Forest Classifier model.
6. Replace the missing values with the predicted values.

To enhance analysis, we performed feature engineering, including the creation of new features like BMI bins, age bins, and glucose level bins. These features can be used for more modeling and prediction tasks.

**We have visualized the data to gain more insights, including:**

- Pie charts showing the distribution of categorical variables.
- Bar charts showing the impact of smoking status on various demographic features.
- Scatterplots showing relationships between age, glucose level, BMI, and strokes.
- Strip plots showing the relationship between age, glucose level, BMI, and strokes with respect to categorical variables.

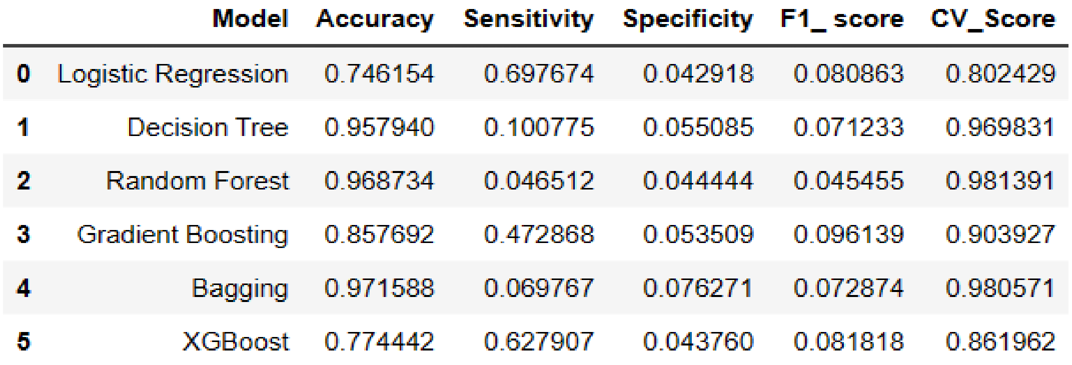
cleaned the data by removing outliers based on z-scores and performed log transformations to improve the distribution of the "avg_glucose_level" feature. also removed “BMI” outliers using the Interquartile Range (IQR) method to enhance the data quality.

**1. Factors Contributing to Stroke Occurrence**:

- Hypertension, especially in individuals aged above 50, presents a high risk of stroke.
- The majority of individuals with heart disease are above 40 years old.
- A significant proportion of smokers are aged 15 years and above.
- "Obese" individuals in the dataset experience a higher number of strokes.
- The "50-75" and "75-above" age groups have the highest number of stroke cases.

**2. Model Evaluation:**

- Logistic Regression: An accuracy of approximately 75%, with a relatively low precision for positive cases, indicating some difficulty in identifying true positives.
- Decision Tree: Achieved a high accuracy but displayed a relatively low recall, potentially leading to false negatives.
- Random Forest: Achieved the highest accuracy and a balanced F1-score, making it the best model for stroke prediction.
- Gradient Boosting: Despite a lower accuracy, it demonstrated reasonable recall for positive cases.
- Bagging: Achieved a high accuracy, but precision for positive cases was relatively low.
- XGBoost: Demonstrated a good recall but had a low precision for positive cases.
- The Cross-Validation Score for Random Forest is notably high, indicating its robustness in generalizing to unseen data



| | Model | Accuracy | Sensitivity | Specificity | F1_ score | CV_Score |
|---|---|---|---|---|---|---|
| **0** | Logistic Regression | 0.746154 | 0.697674 | 0.042918 | 0.080863 | 0.802429 |
| **1** | Decision Tree | 0.957940 | 0.100775 | 0.055085 | 0.071233 | 0.969831 |
| **2** | Random Forest | 0.968734 | 0.046512 | 0.044444 | 0.045455 | 0.981391 |
| **3** | Gradient Boosting | 0.857692 | 0.472868 | 0.053509 | 0.096139 | 0.903927 |
| **4** | Bagging | 0.971588 | 0.069767 | 0.076271 | 0.072874 | 0.980571 |
| **5** | XGBoost | 0.774442 | 0.627907 | 0.043760 | 0.081818 | 0.861962 |

### 3. Implications for Healthcare Providers

- Healthcare providers can use predictive models to spot individuals who are at risk of having a stroke. The Random Forest model is a good choice due to its high accuracy.
- With the help of these models, healthcare providers can take action to prevent strokes. This might include monitoring patients at high risk and providing specific interventions to lower their stroke risk.
- Encouraging people to make lifestyle changes like managing their weight, quitting smoking, and getting regular check-ups can be effective in preventing strokes.