# **Machine Learning Pipeline**

**Project Report: Stroke Prediction Using Machine Learning**

**Objective:**

To predict the likelihood of a stroke in individuals based on health-related features using multiple machine learning models, and evaluate their performance to identify the best performing model.

**Data Preprocessing**

**a. Loading the Data**

* Loaded the dataset containing heart and stroke-related attributes.

**b. Cleaning the Data**

* Handled missing values using mean imputation for numerical columns.
* Categorical variables were encoded using one-hot encoding.

**c. Feature Selection**

* Selected relevant features for model training, such as age, hypertension, heart\_disease, avg\_glucose\_level, bmi, etc.

**Class Imbalance Handling**

**a. Problem**

* Target class (stroke) was highly imbalanced (most samples were non-stroke).

**b. Solution**

* Used **SMOTE (Synthetic Minority Oversampling Technique)** to balance the dataset by generating synthetic samples for the minority class.

**Model Training**

**Trained Multiple Classification Models:**

| **Model Name** | **Used?** |
| --- | --- |
| Logistic Regression | ✅ |
| Random Forest | ✅ |
| K-Nearest Neighbors | ✅ |
| Support Vector Machine | ✅ |
| Decision Tree | ✅ |
| Gaussian Naive Bayes | ✅ |
|  |  |
|  |  |

Each model was trained on the **balanced training dataset** (X\_train\_bal, y\_train\_bal).

**Model Evaluation**

**a. Validation Approach**

* Split dataset using **train\_test\_split()** to separate training and testing data.
* Evaluation performed **only on the X\_test, y\_test** set.

**b. Evaluation Metrics Used**

For each model, the following metrics were calculated:

* Accuracy
* Precision
* Recall
* F1 Score
* ROC AUC
* Confusion Matrix
* Classification Report (including all major classification metrics)

**c. Evaluation Code**

Reusable evaluate\_model() function was created to print or return all the above metrics.

**Model Comparison**

**Metric Summary**

All model evaluation metrics were stored in a dictionary and converted to a DataFrame for easy comparison.