Project Statement: Using the provided dataset, you are asked to train a model that predicts whether a patient has a stroke or not. The project can be submitted as a Jupyter Notebook and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation.

Approach: As recommended in the Project Description, Project Report consists of following sections

1) Data Analysis consisting of Brief Description of data
2) Feature Selection
3) Model Training
4) Model Evaluation

Data Analysis:

Stroke data consists of 5110 rows, each row describing 1 patient and each patient has values corresponding to 12 columns. Out of these 12 columns, id is the identifier of the patient. Hence, we can say 10 as independent variables and 1 as dependent variable. On looking at details, following observations were made

a. 5 columns are categorical (gender, ever_married, work_type, Residence_type, smoking_status)
b. 3 columns are continuous i.e., float (age, avg_glucose_level, bmi)
c. 3 columns are integer (hypertension, heart disease and stroke)

On going deeper into categorical values

- Gender – has 3 unique values (Male : 2994, Female : 2115 and Other : 1). Looks like other is a data entry other and hence is removed from the data in the later part of the analysis
- Ever_Married – has 2 unique values (Yes : 3353 and No : 1757)
- Residence_Type – has 2 unique values (Urban: 2596 and Rural : 2514)
- Work_Type – has 5 unique values (Private: 2925, Self-employed : 819, Children : 687, Govt_Job : 657 and Never_Worked : 22)
- Smoking_Status – has 4 unique values (never smoked: 1892, formerly smoked : 885, smokes : 789 and Unknown : 1544). Unknown is a significant chunk and hence is considered as a legitimate value

On going deeper into binary values, only 2 values are observed (0 and 1) for hypertension, heart disease and stroke

Univariate analysis

- Incidence rate of stroke is 249/ 5110 = 4.87%
- Some variability in incidence rate is observed over Gender (Male : 5.1% and Female : 4.7%)
- Some variability in incidence rate is observed over Residence Type (Urban : 5.2% and Rural: 4.5%)
- Some variability in incidence rate is observed over Work Type (Self employed : 7.9%, Govt_Job and Private : 5.0% and 5.1% respective, Children: 0.3% and Never worked : 0%)
- Large variability is observed over hypertension (Yes : 13.3% and No : 4.0%)
- Large variability is observed over heart disease (Yes : 17% and No : 4.2%)
- Large variability is observed over ever married (Yes: 6.6% and No: 1.7%)

- People with higher age tend to have higher chances of stroke: Concluded on basis of both scatter chart and box plot
- BMI doesn't show so much variability in box plot. However, with scatter plot till bmi value of 35 stroke is independent of bmi but post 35 stroke increases with bmi

Obviously, age can influence a lot of variables such as marital status, employment, its correlation is seen with all the variables. Before building correlation plot following transformations were made

- All categorical variables were encoded using dummy values, and
- Null values of BMI were imputed with mean values

Following variables (ever_married, work_type_children, smoking_status_unknown) were found to be correlated with age

Feature Selection

Considering incidence rate is 4.9% our data is an unbalanced data set. Hence, as first step we need to balance the data. For this we are leveraging SMOTE from library imblearn. Post balancing the data, we have used 80% data as train and 20% data as test. Feature Selection was done using SelectKBest through ANOVA. Following features were selected (Age, avg_glucose_level, gender, smoking_status). New dataframe is created considering only above-mentioned features.

Model Training and Evaluation

Following models are tried

a) RandomForestClassifier
b) ExtraTreesClassifier
c) DecisionTreeClassifier
d) Support Vector Classification
e) XG Boost
f) Logistics Regression – On basis of model suggestion, preprocessing using StandardScaler was used for Logistics regression

For each of the above model, following metrics are evaluated

a) Cross Validation Score using RepeatedStratifiedKFold : In this data is splitted into multiple folds and in every run 1 fold is used as test dataset and remaining folds are used as train dataset. This process is repeated multiple times and average of scoring metric is used for model efficiency.
b) ROC AUC Curve : This is a measure of classification that how well model is able to classify 1 against 0. In ROC AUC Curve higher the value, better is the model able to classify. X-axis is False Positive rate and Y-axis is true positive rate
    a. True Positive (TP) = 1 (actual value) is being classified as 1 (predicted value)
    b. False Positive (FP) = 0 is being classified as 1
    c. True Negative (TN) = 0 is being classified as 0
    d. False Negative (FN) = 1 (actual value) is being classified as 0 (predicted value)
    e. False Positive Rate = FP / (FP + TN) i.e., Out of total actual 0 values, how many are we predicting as 0

f. True Positive Rate = TP / (TP + FN) i.e., Out of total actual 1 values, how many are we predicting as 1

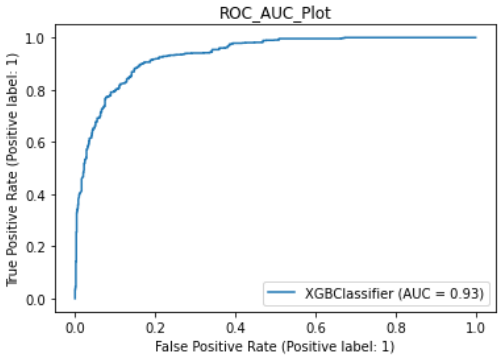
c) Precision = TP / (TP + FP) i.e., Out of total predicted values how many are we correctly predicting
d) Recall = TP / (TP + FN) i.e., is same as True Positive Rate and is also called Sensitivity
e) F1 = 2*Precision*Recall / (Precision + Recall)

On Comparing various models, we realized that XG Boost is the best model (on basis of F1 Score)

Models	Cross Validation	ROC AUC Curve	True Positive	False Negative	False Positive	True Negative	Precision	Recall	F1
RandomForestClassifier	98%	74%	29%	21%	4%	46%	87%	58%	69%
ExtraTreesClassifier	98%	72%	27%	23%	4%	46%	86%	53%	66%
DecisionTreeClassifier	91%	74%	29%	21%	5%	45%	86%	58%	69%
Support Vector Classification	85%	79%	43%	7%	14%	36%	75%	86%	80%
XG Boost	95%	87%	44%	6%	8%	42%	85%	88%	87%
Logistics Regression	92%	85%	43%	7%	9%	7%	83%	87%	85%

Sharing Results of XG Boost Model

RUC Curve



Confusion Matrix (X-axis is actual and Y-axis is predicted)