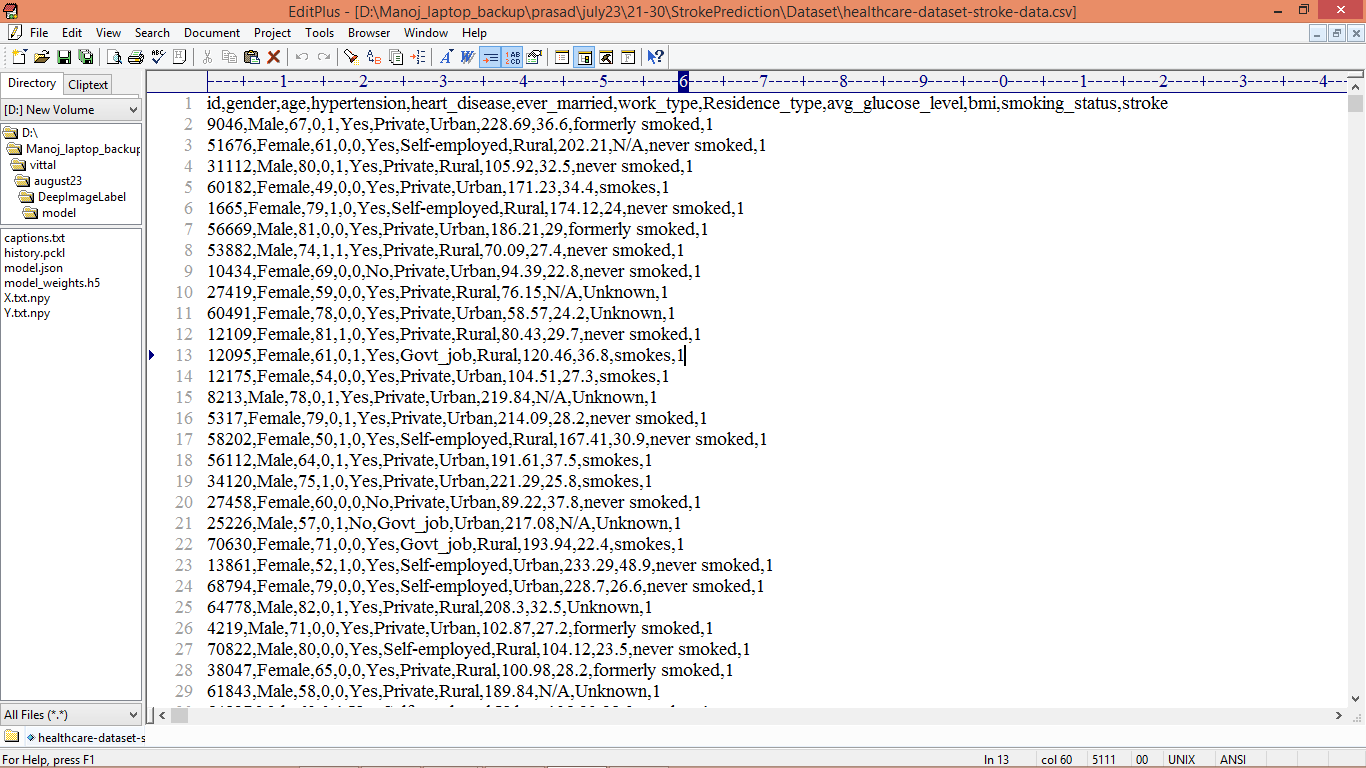
Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention

Stroke often causes due to blood flow stop to brain and this is one of the deadly diseases. Patient life can be saved and stroke can be avoided by timely and accurate detection. Existing detection technique requires heavy resources and they make time for prediction. To overcome from this problem many machine learning algorithms were introduced as they are very accurate in medical diseases prediction but existing techniques were suffering from data leakage such as improper handling or missing values, improper categorical data calculation etc. No existing techniques were employing any Explainable model (XAI) which can show which features are helping most in detecting stroke so doctor can give priority on such features for faster recovery. These explainable features can be Smoking, Age, BMI and may be other features.

So author of this paper employing different processing techniques such as Removing missing values, Imbalance data handling using SMOTE and relevant features selection using CHI2 algorithm. All this processed features will get trained on 6 different algorithms such as Random Forest, KNN, SVM, Logistic Regression, XGBOOST and Naive Bayes. In all algorithm Random Forest is giving high accuracy and each algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE.

For easy understanding of features author employing various graph on Strokes patient data. Best algorithm will be input to SHAPELY Explainable (XAI) algorithm to explain about features which are contributing most in predicting correct label.

To train all algorithms we are using STROKE dataset from KAGGLE and below screen showing all dataset details



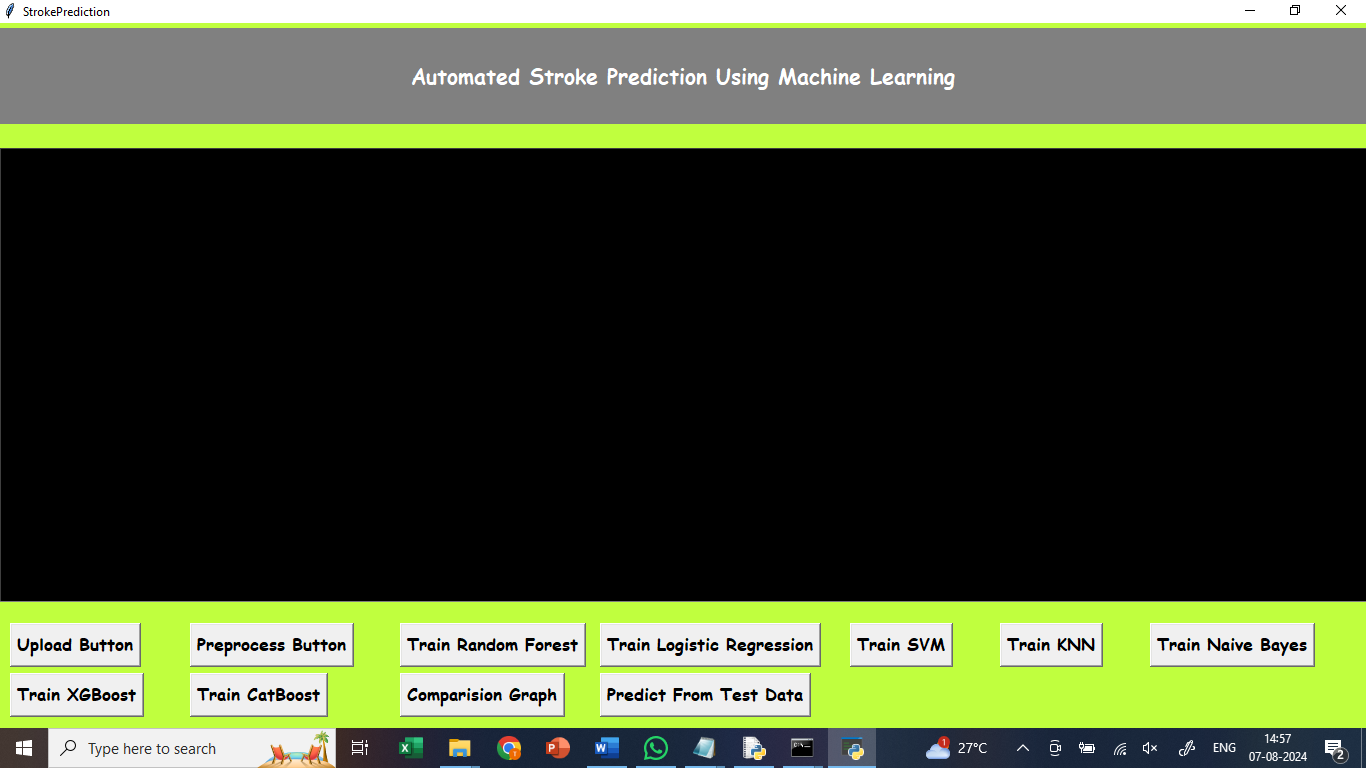
In above screen first row represents dataset column names and remaining rows represents dataset values and by using above dataset we will test all algorithm performance.

Extension Concept

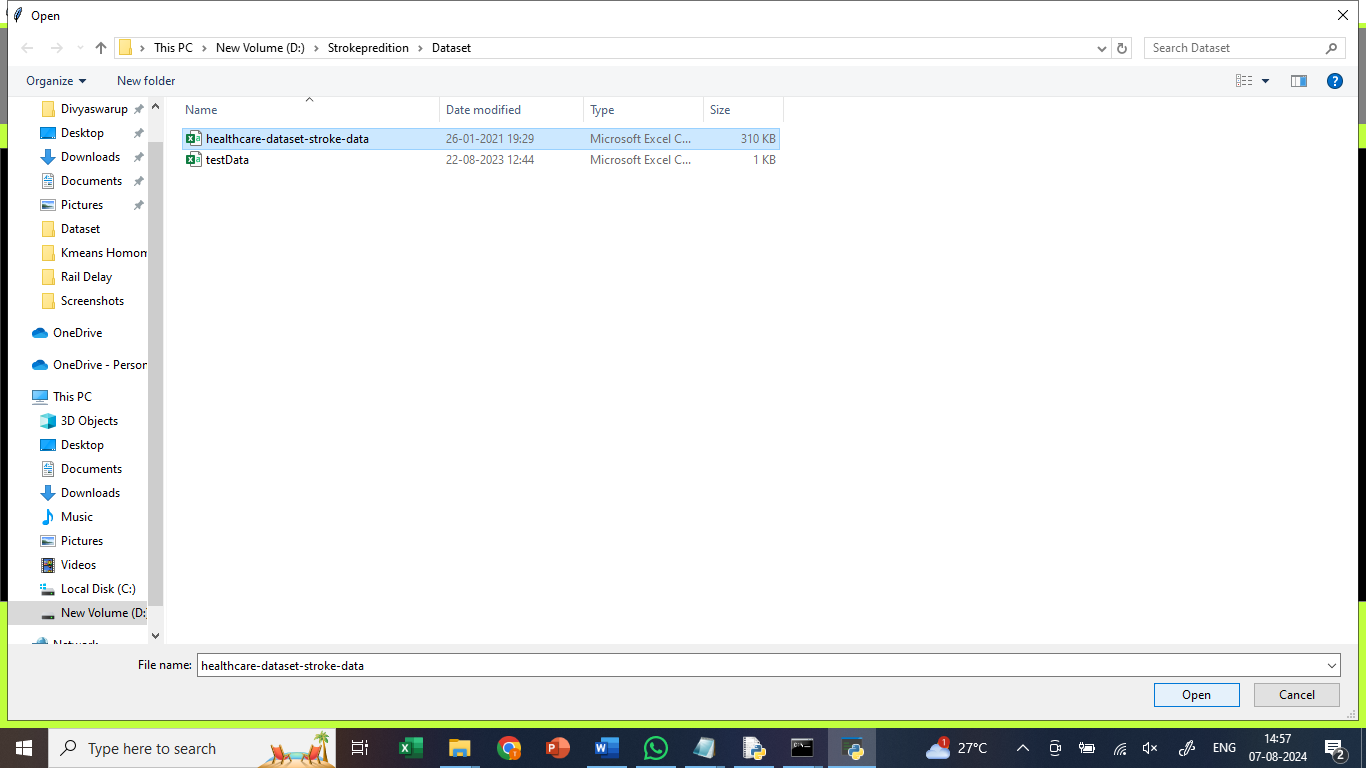
As extension we are employing CATBOOST classifier which will use forest of weak classifiers or group of multiple classifiers and then train each classifier and vote out best classifier for final prediction and using multiple classifier will help CATBOOST in enhancing prediction accuracy.

SCREEN SHOTS

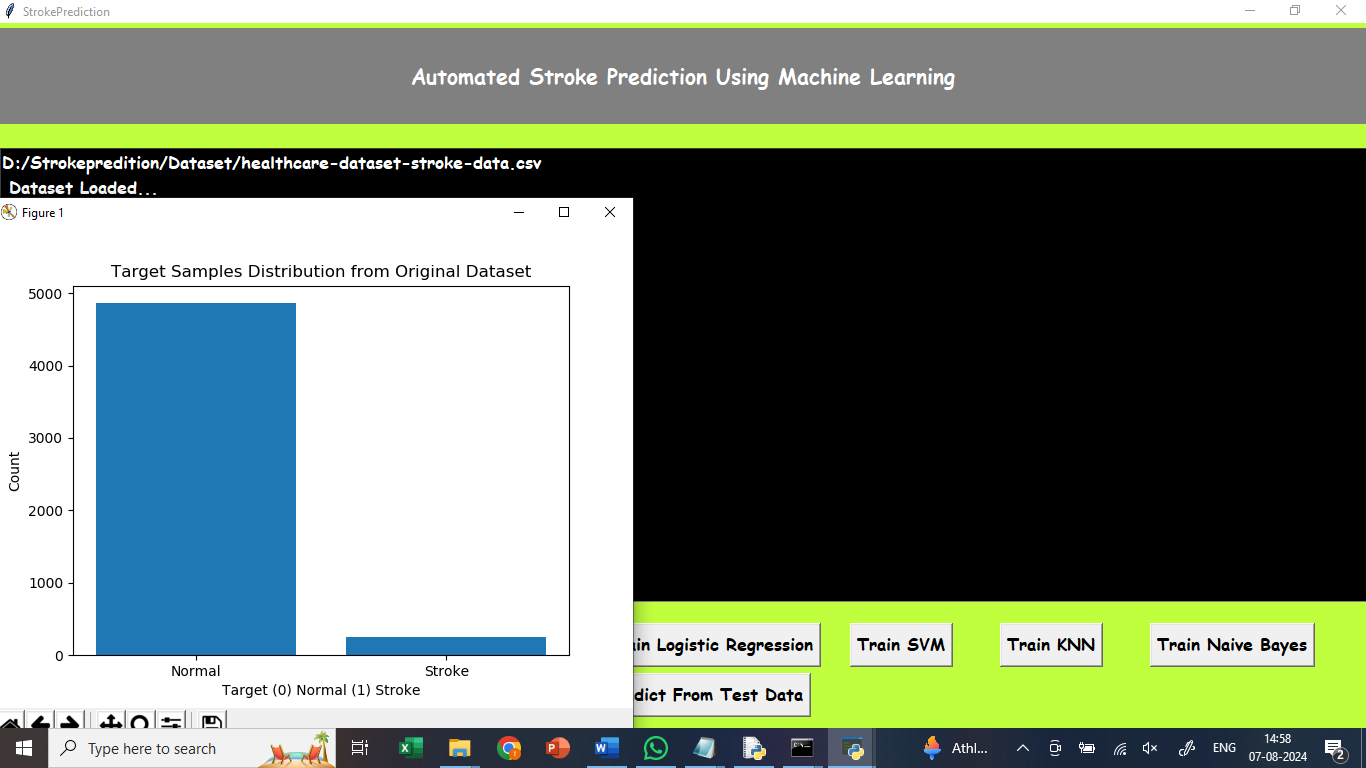
To run project double click on ‘run.bat’ file to get below screen



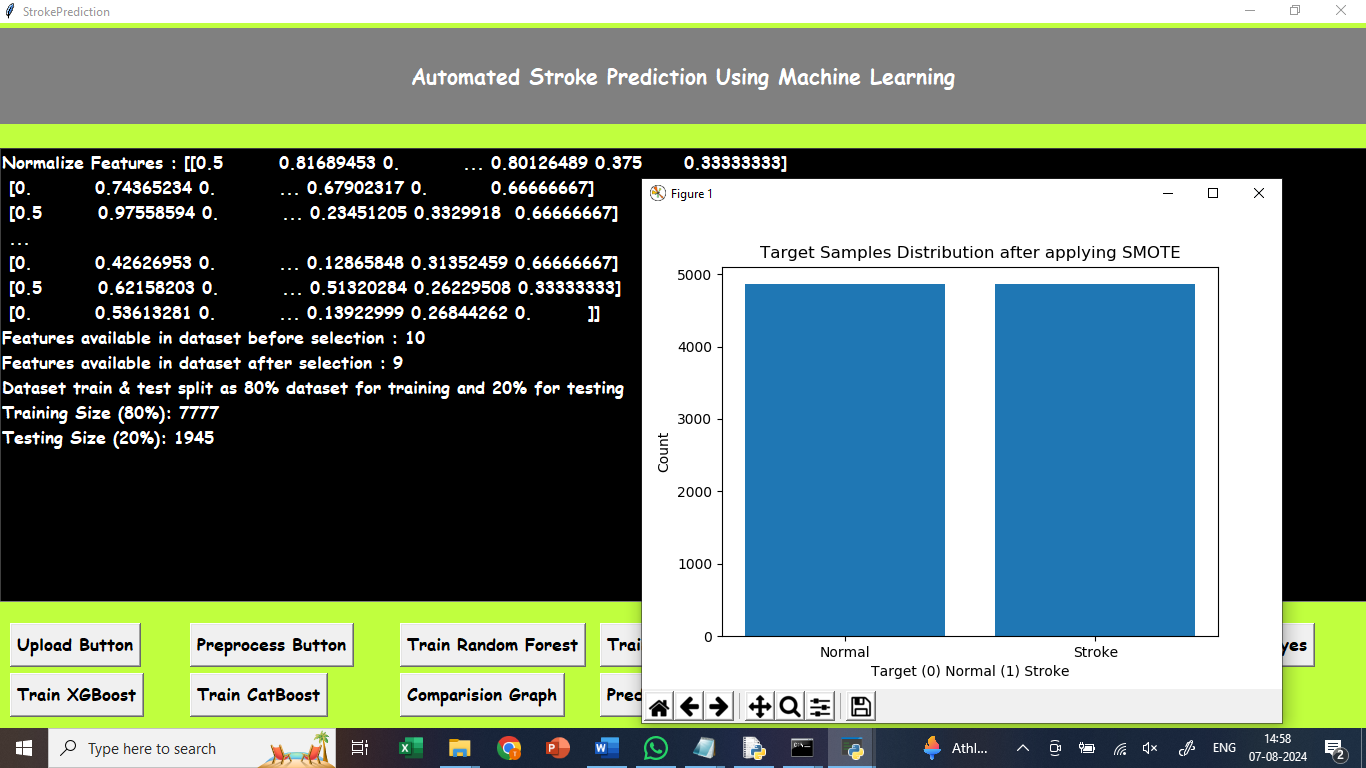
In above screen, click on upload button



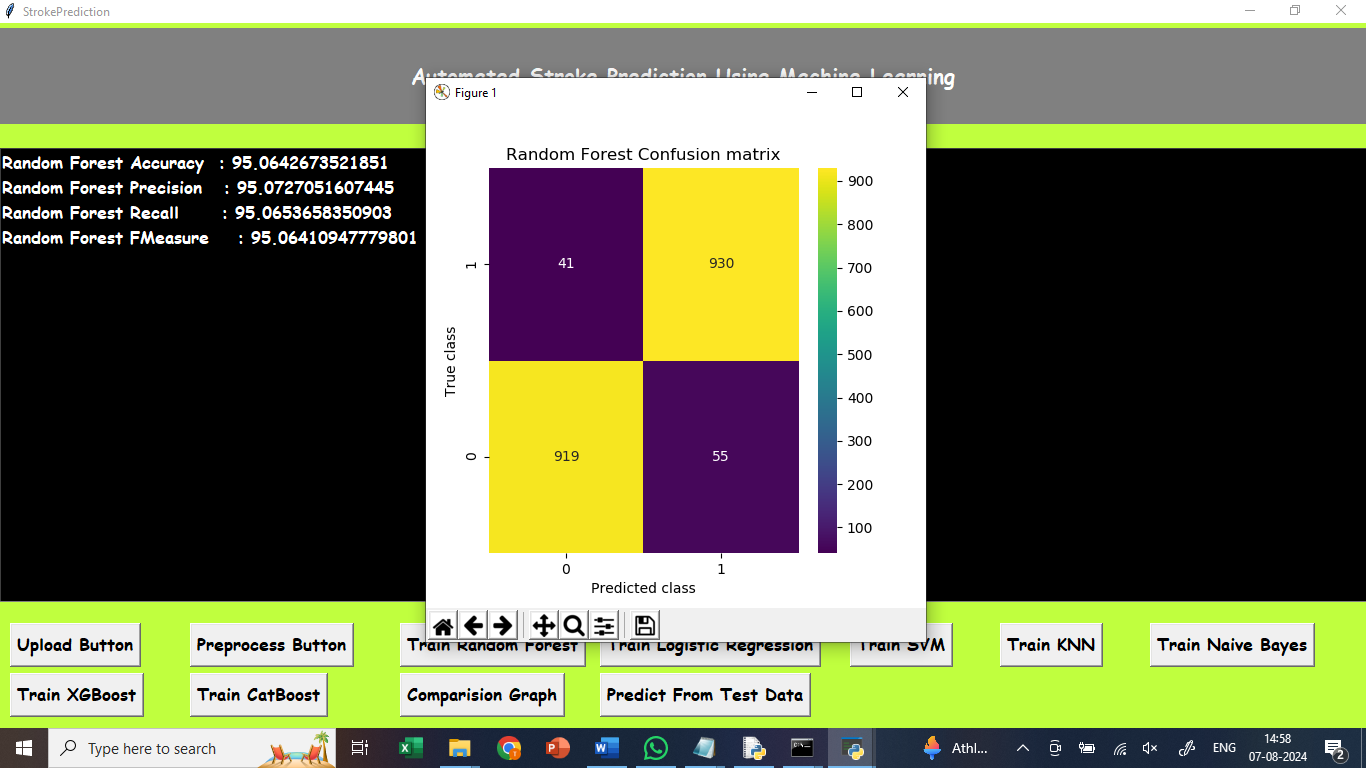
In above screen, upload healthcare dataset



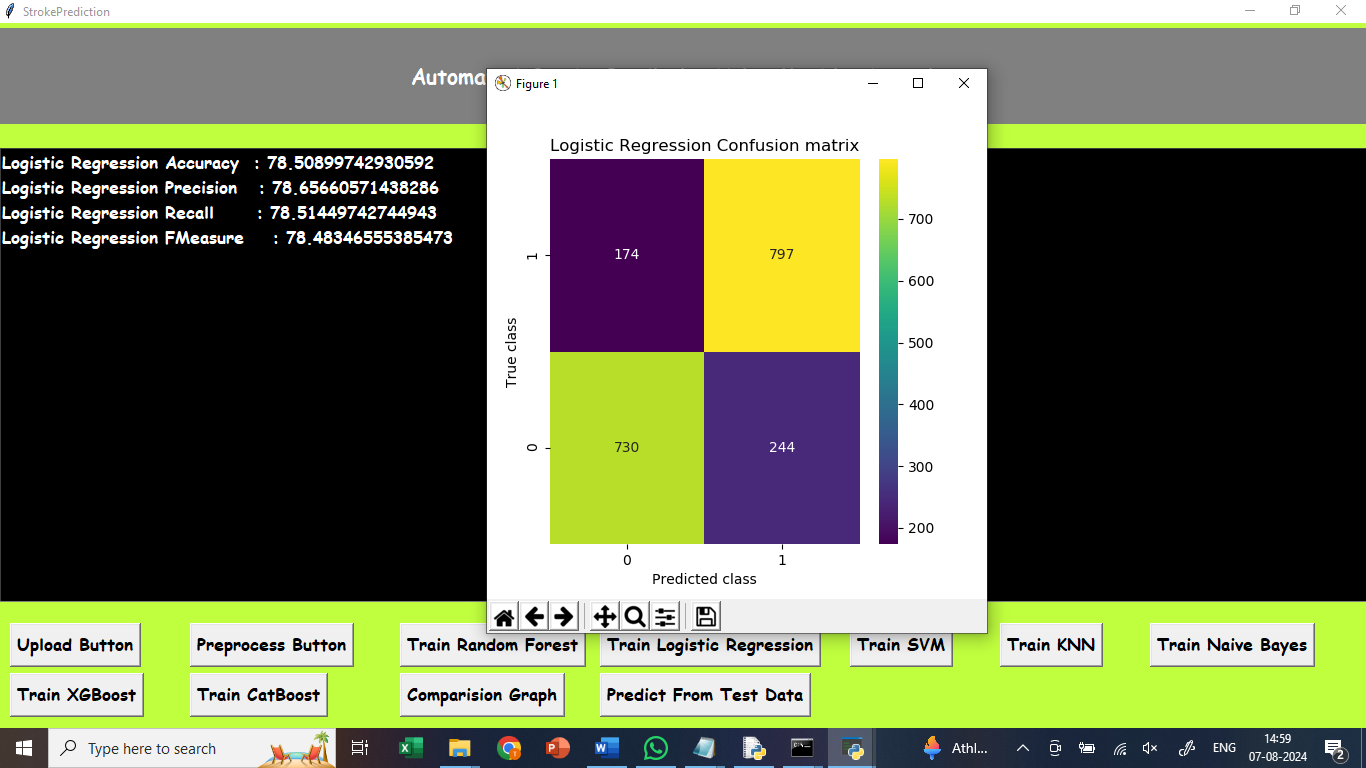
In above screen, we are getting the graph which displays the data related to normal and stroke.



In above screen, click on preprocess button, for splitting the dataset, 80% for training and 20% for testing.



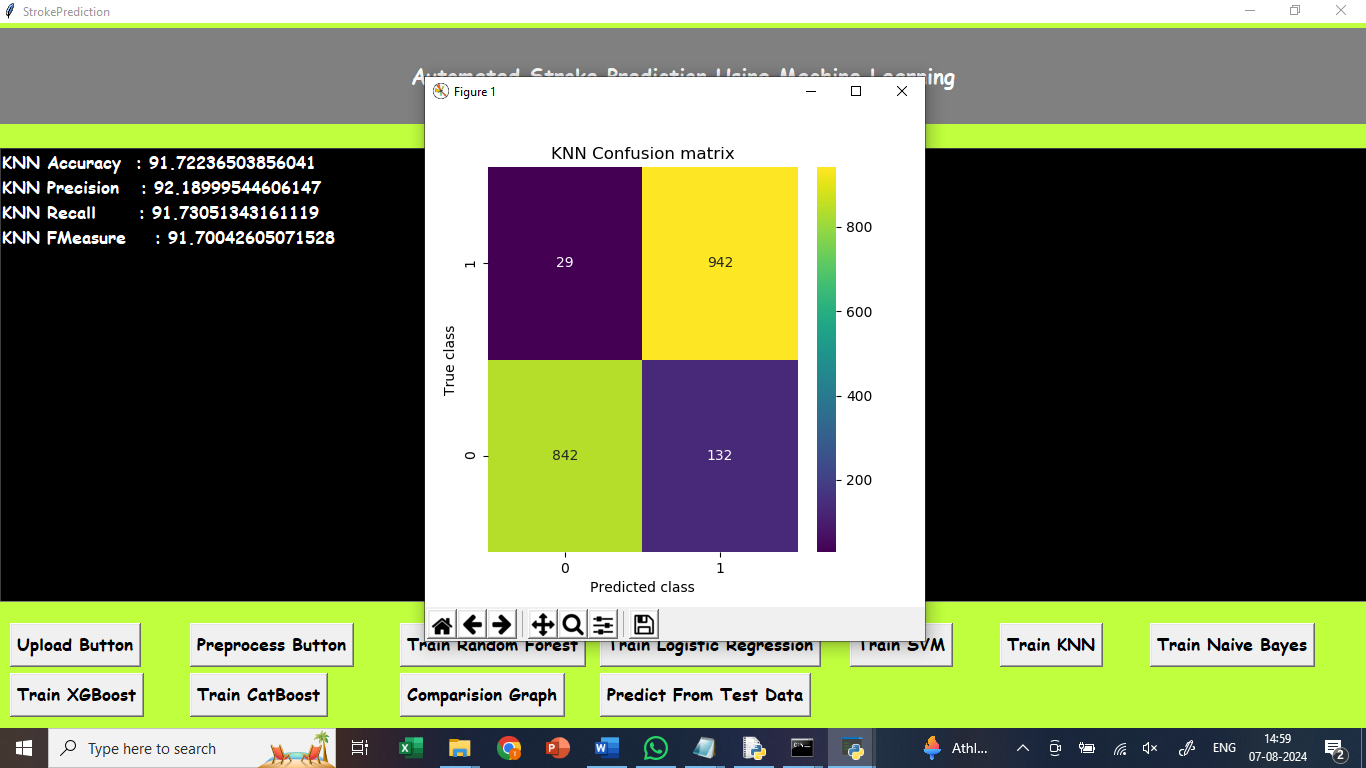
In above screen, click on Train Random Forest button, to train Random Forest and got 95% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



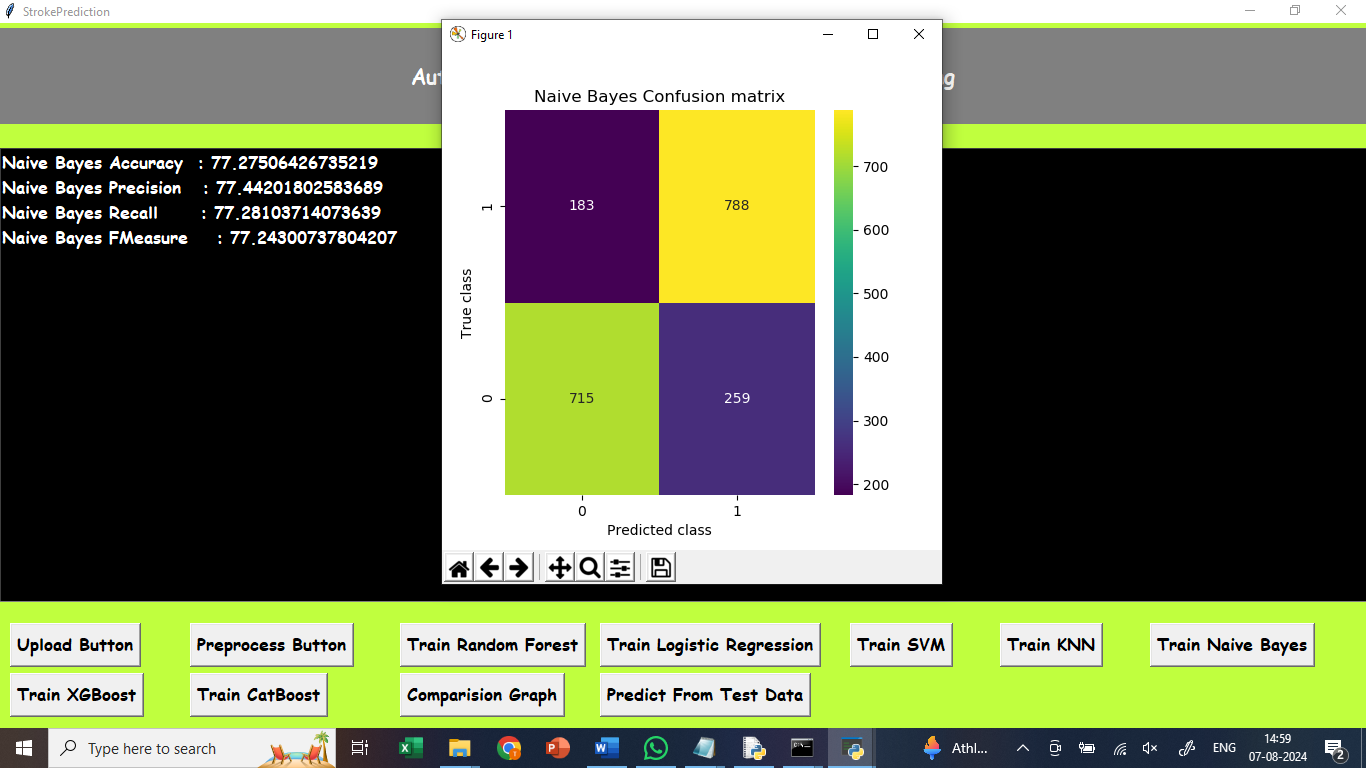
In above screen, click on Train Logistic Regression button, to train Random Forest and got 78% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



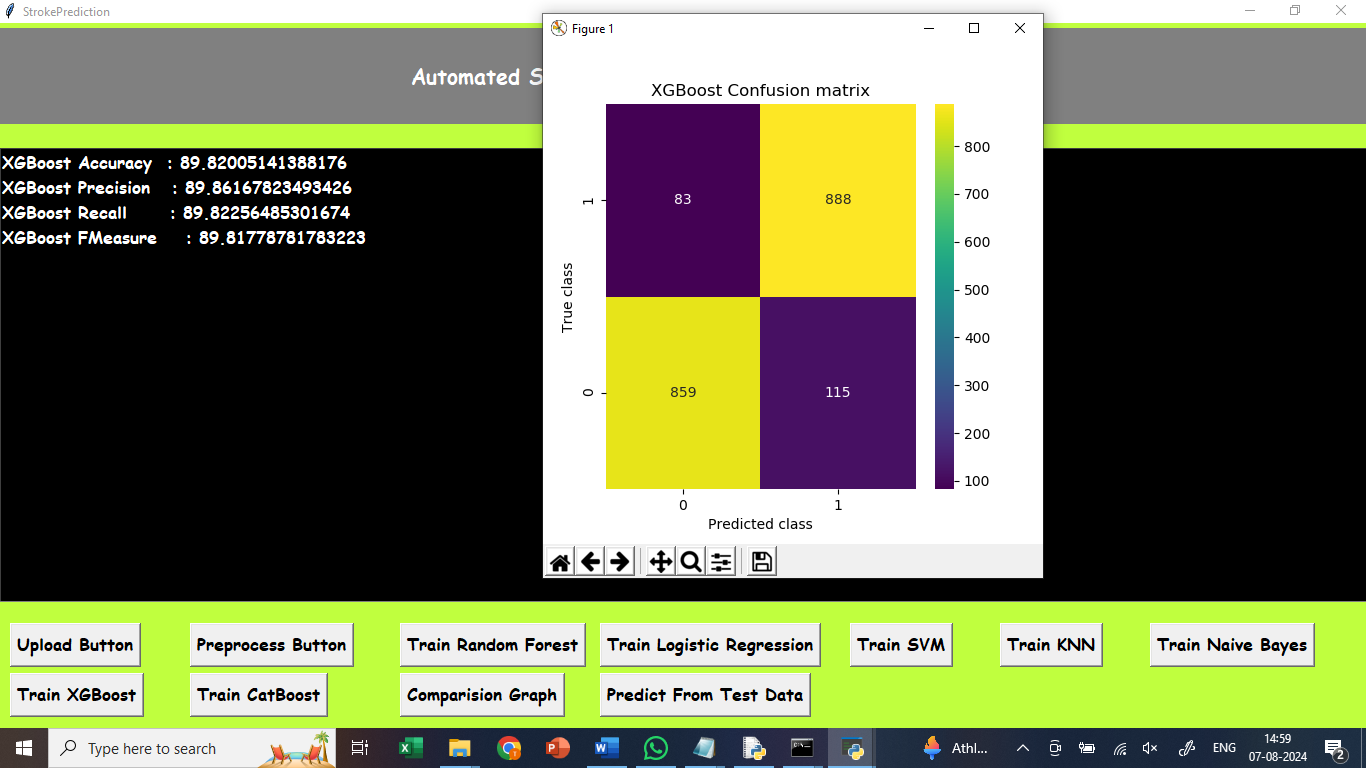
In above screen, click on Train SVM button, to train SVM and got 81% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



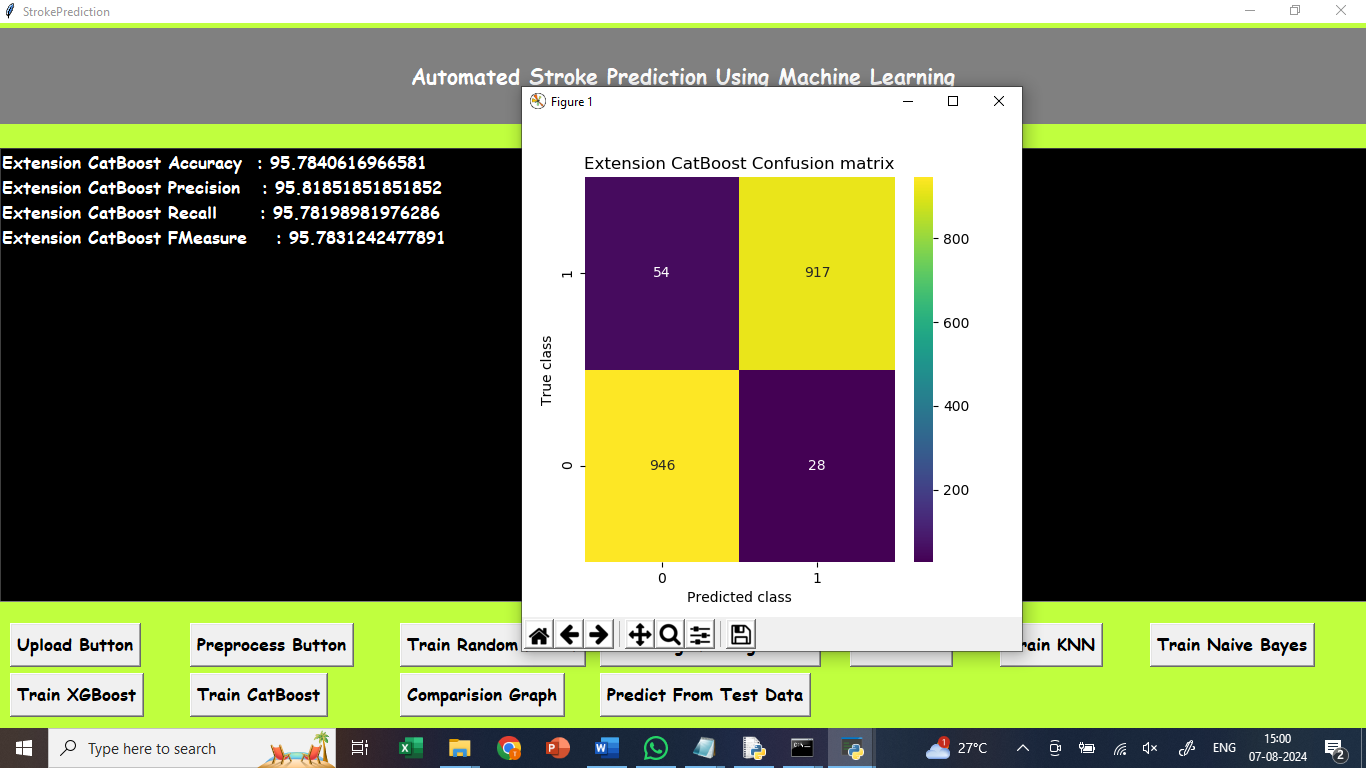
In above screen, click on Train KNN button, to train KNN and got 91% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



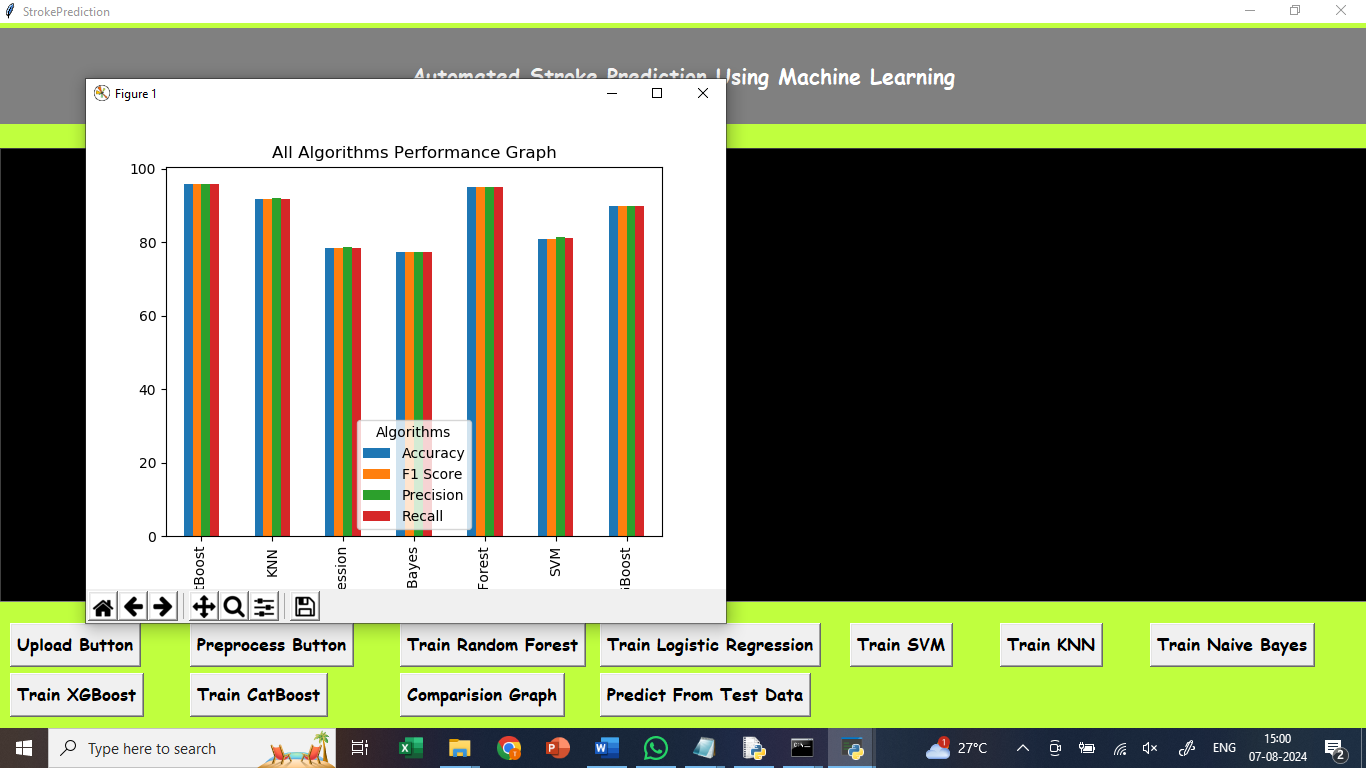
In above screen, click on Train Naïve Bayes button, to train Naïve Bayes and got 77% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



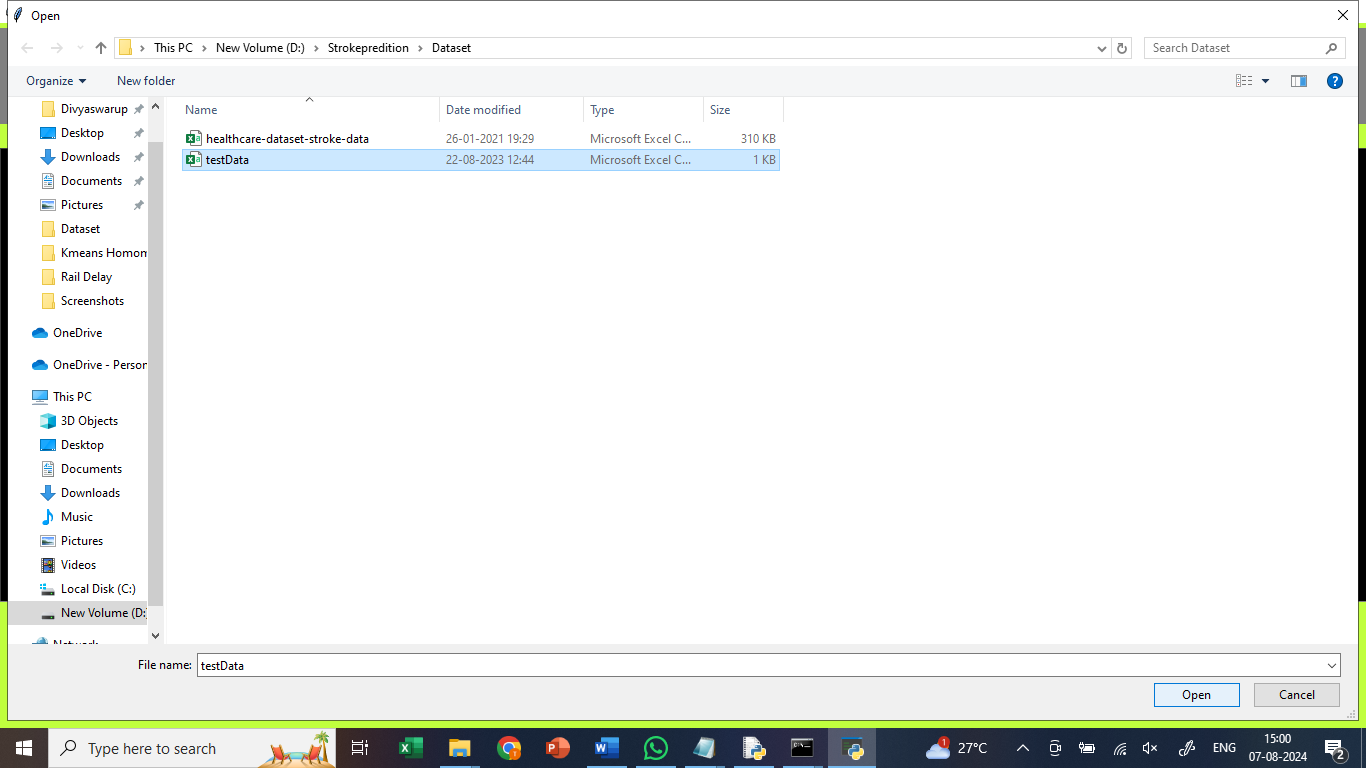
In above screen, click on XGBoost button, to train XGBoost and got 89% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



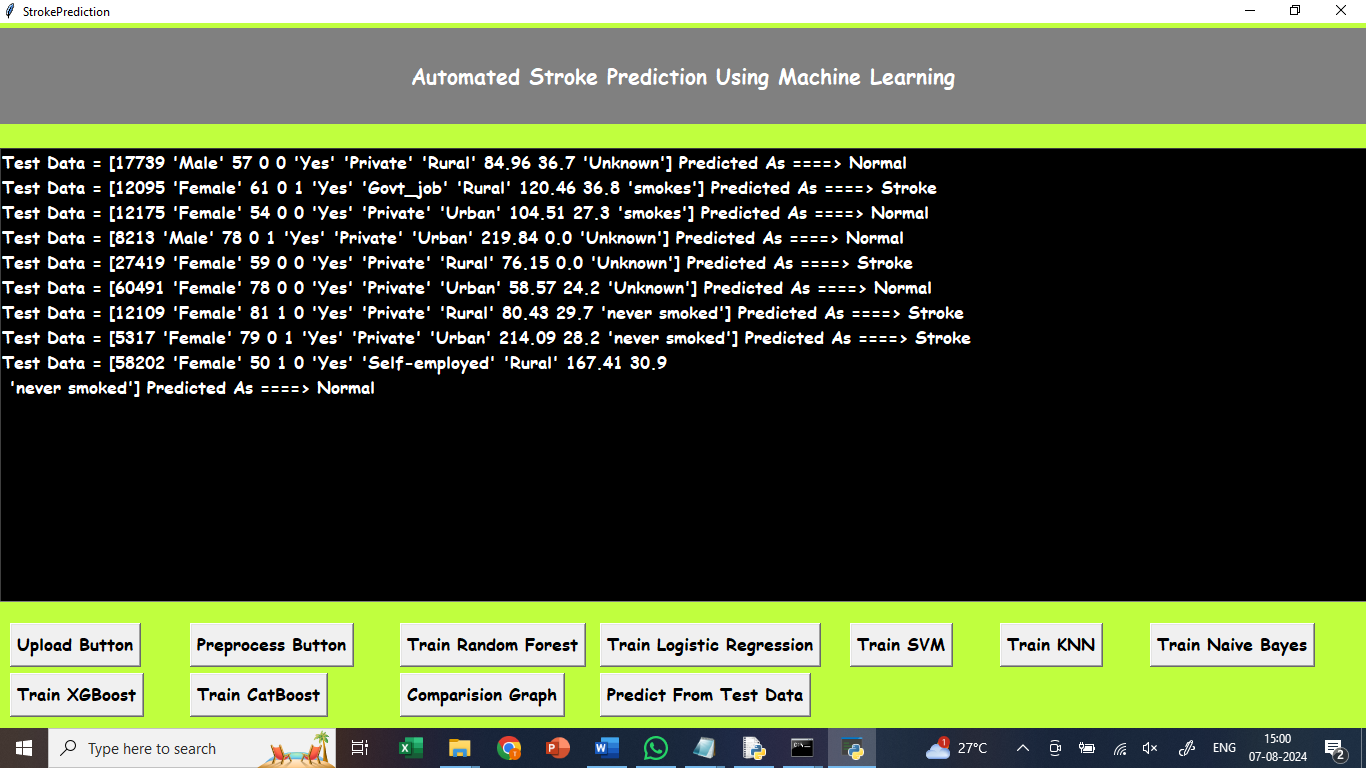
In above screen, click on Train CatBoost button, to train CatBoost and got 95% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where blue boxes represents incorrect prediction count which are very few and yellow and light green boxes represents correct prediction count.



In above screen, click on comparision graph button, to compare all algorithms performance with four different parameters.



In above screen, click on predict from test data button, upload test data to predict whether data relates to stroke or normal



In above screen, after uploading test data, it classified data as normal or stroke.