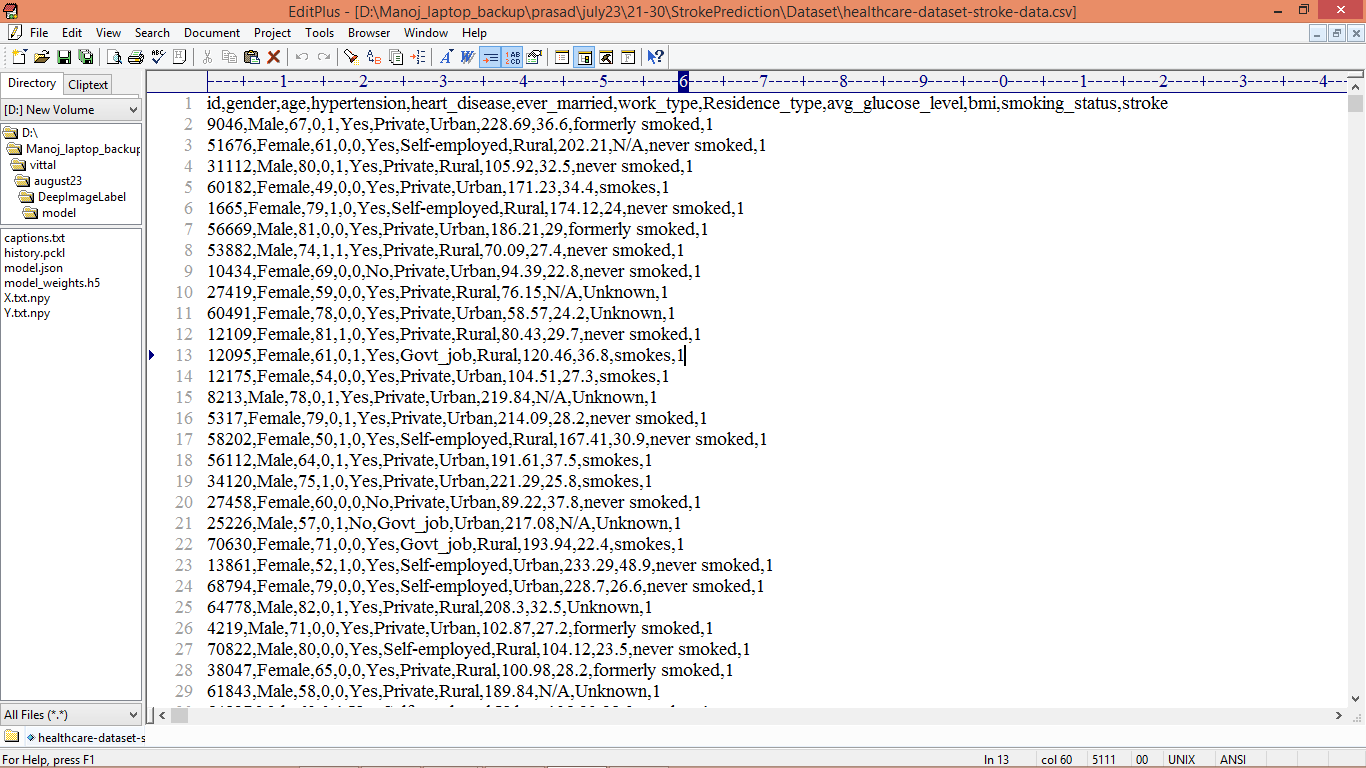
Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention

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



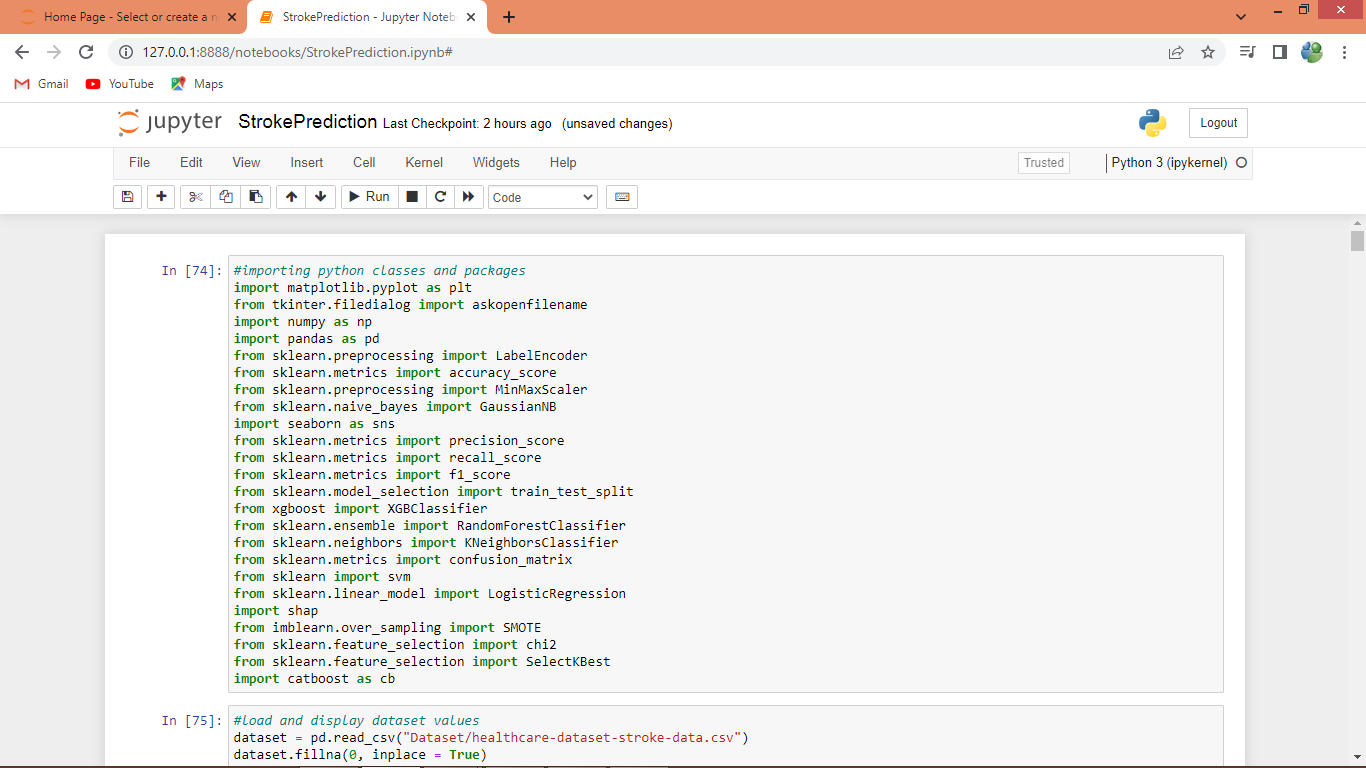
In above screen first row represents dataset column names and remaining rows represents dataset values and by using above dataset we will test all algorithm performance.

Extension Concept

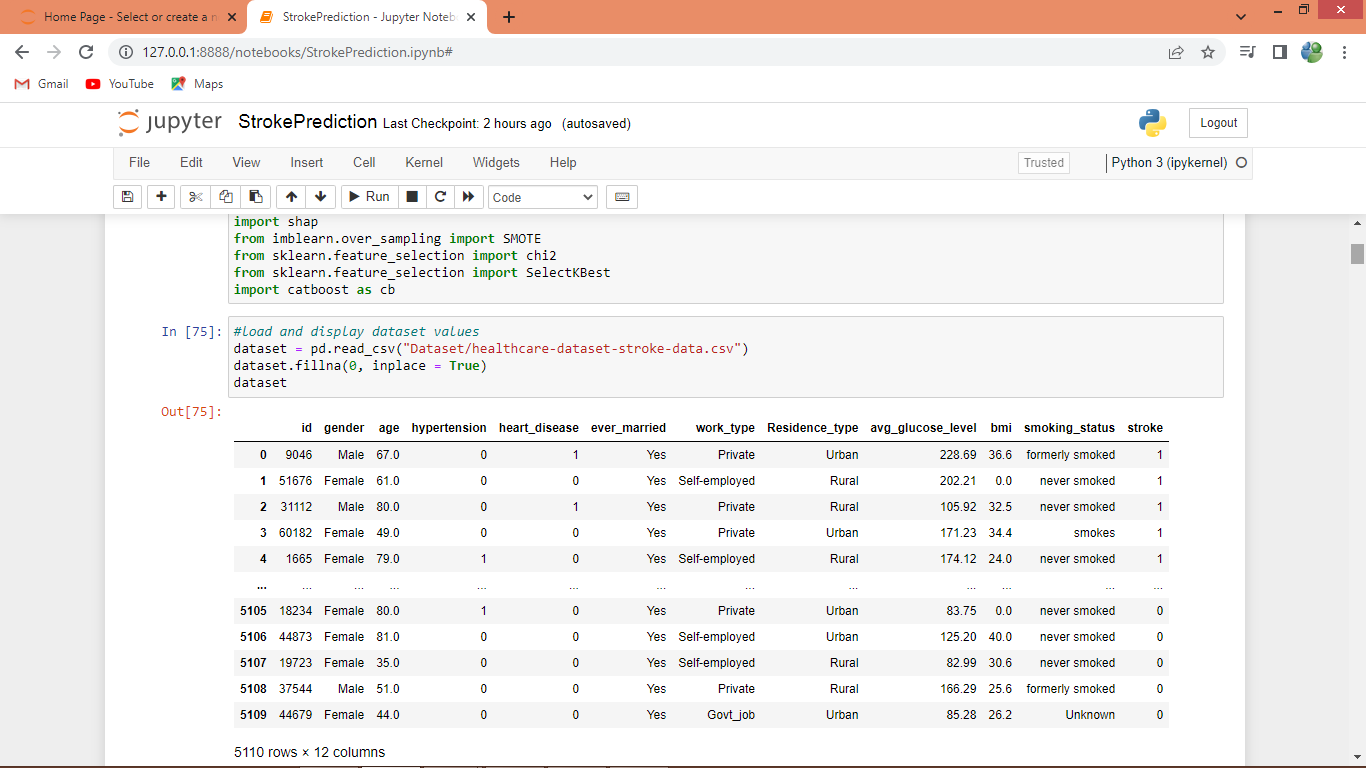
As extension we are employing CATBOOST classifier which will use forest of weak classifiers or group of multiple classifiers and then train each classifier and vote out best classifier for final prediction and using multiple classifier will help CATBOOST in enhancing prediction accuracy.

SCREEN SHOTS

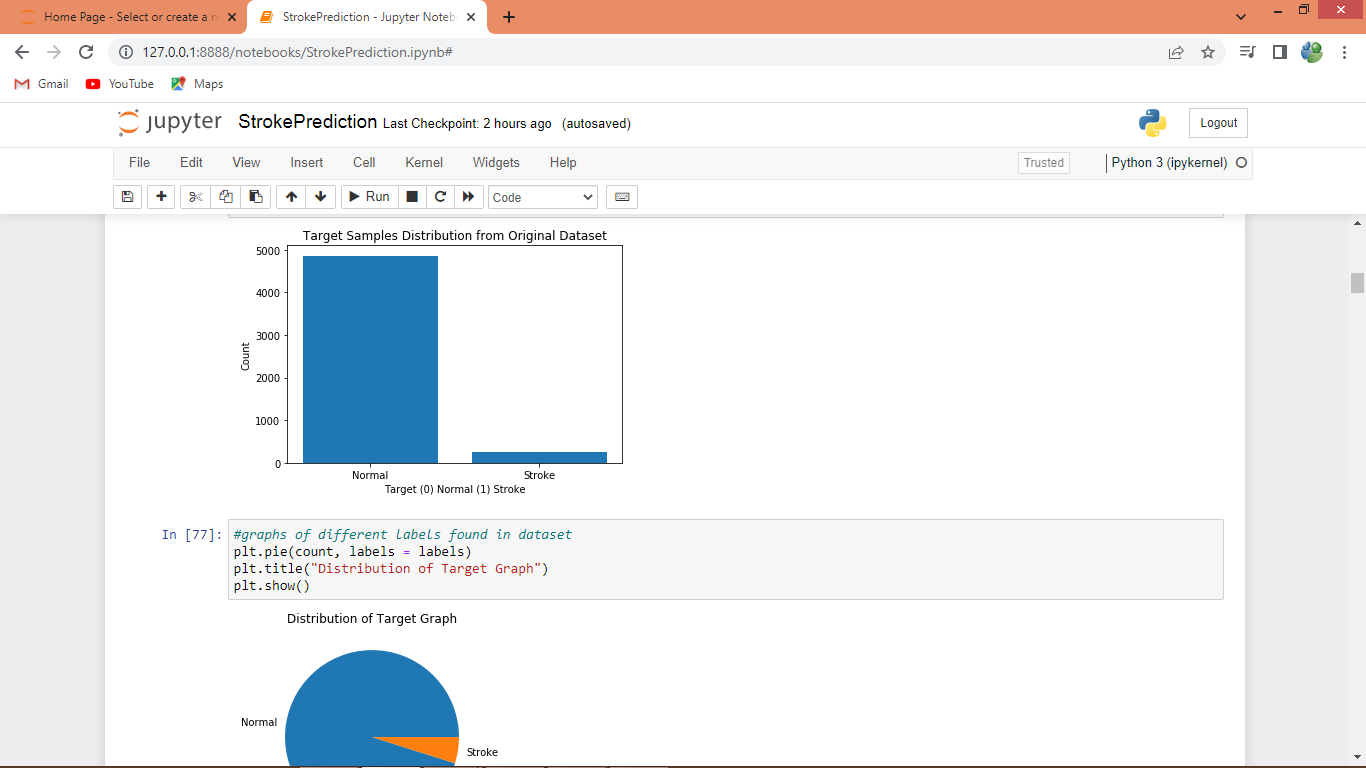
We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments



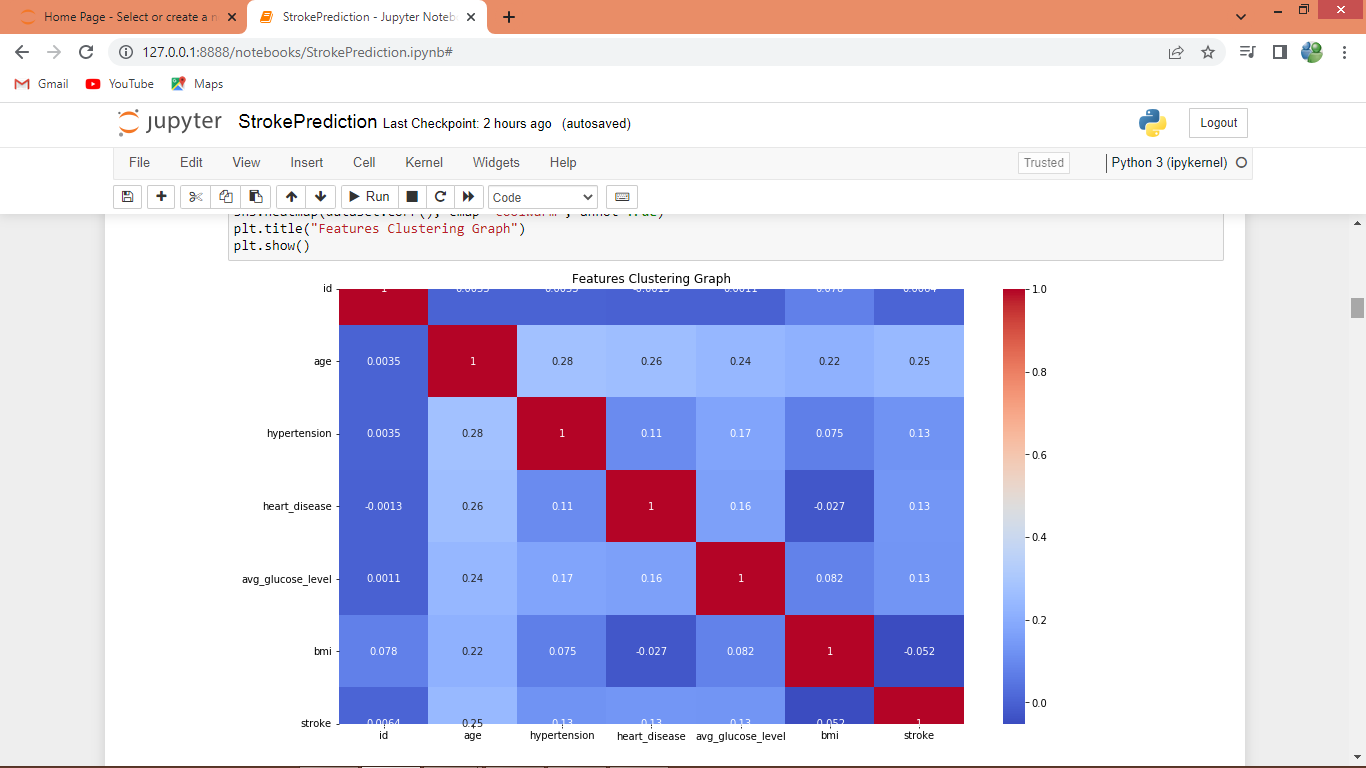
In above screen importing require python classes and packages



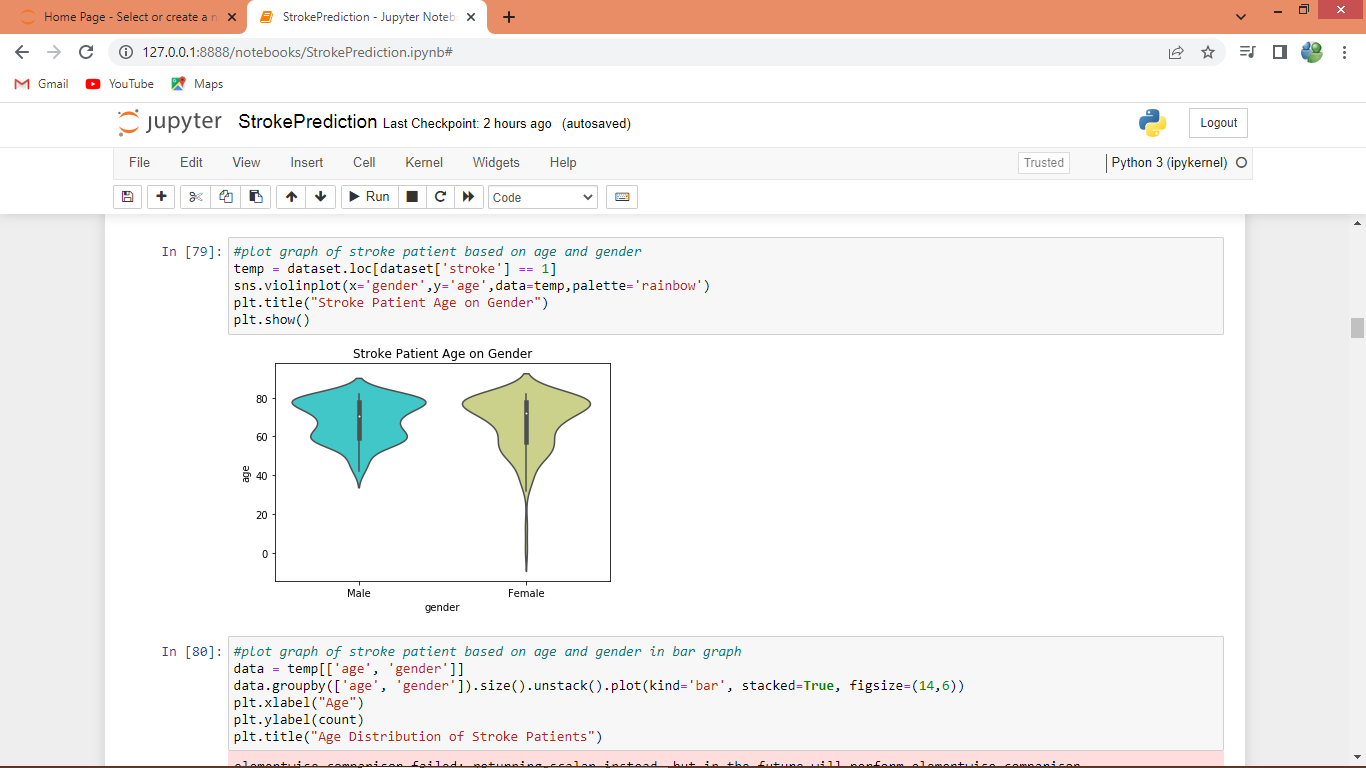
In above screen reading and displaying dataset and then removing missing values and above dataset contains both numeric and non-numeric data so by using label encoder class will convert non-numeric data to numeric data



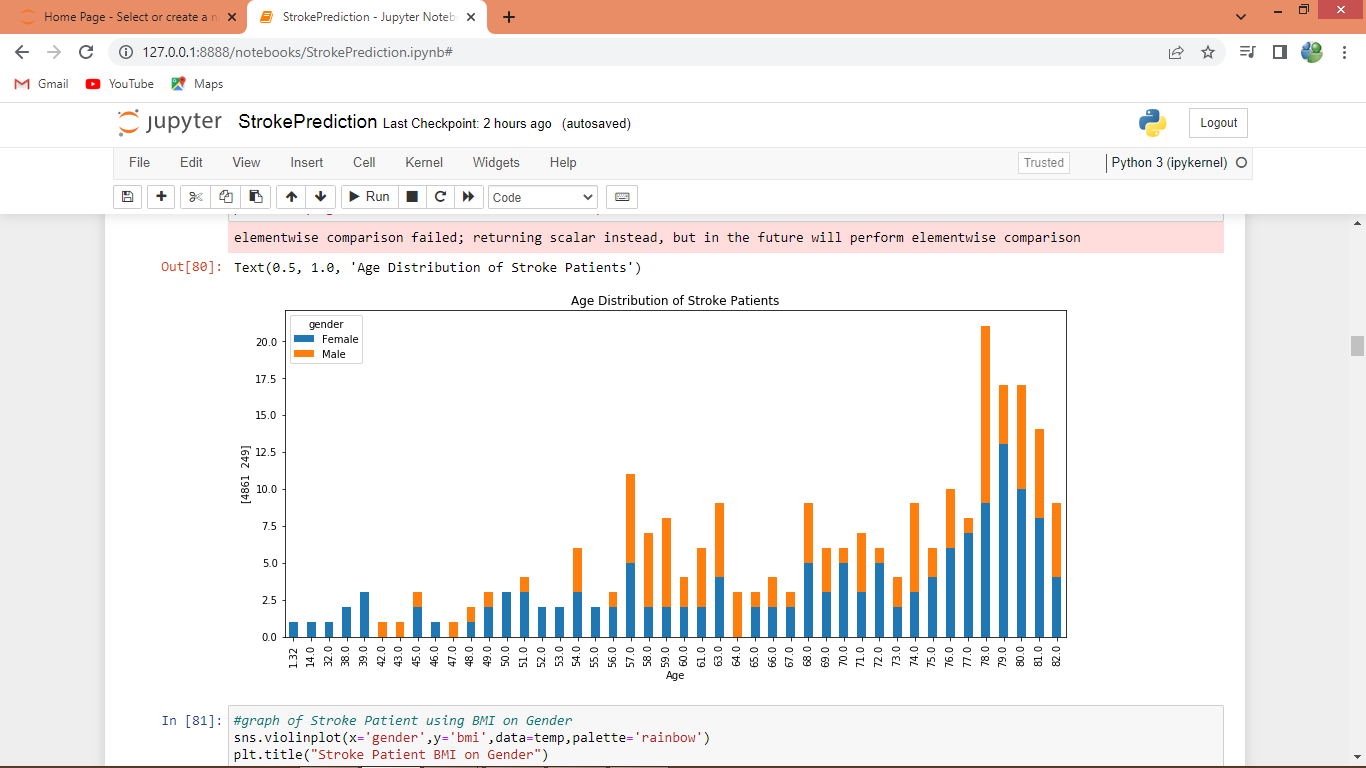
In above screen finding and plotting graph of Normal and Stroke label where x-axis represents Normal and Stroke and y-axis represents count and we can see one class contains so many records and other class contains few records only so data is highly imbalance which we can handle using SMOTE and same can see PIE graph



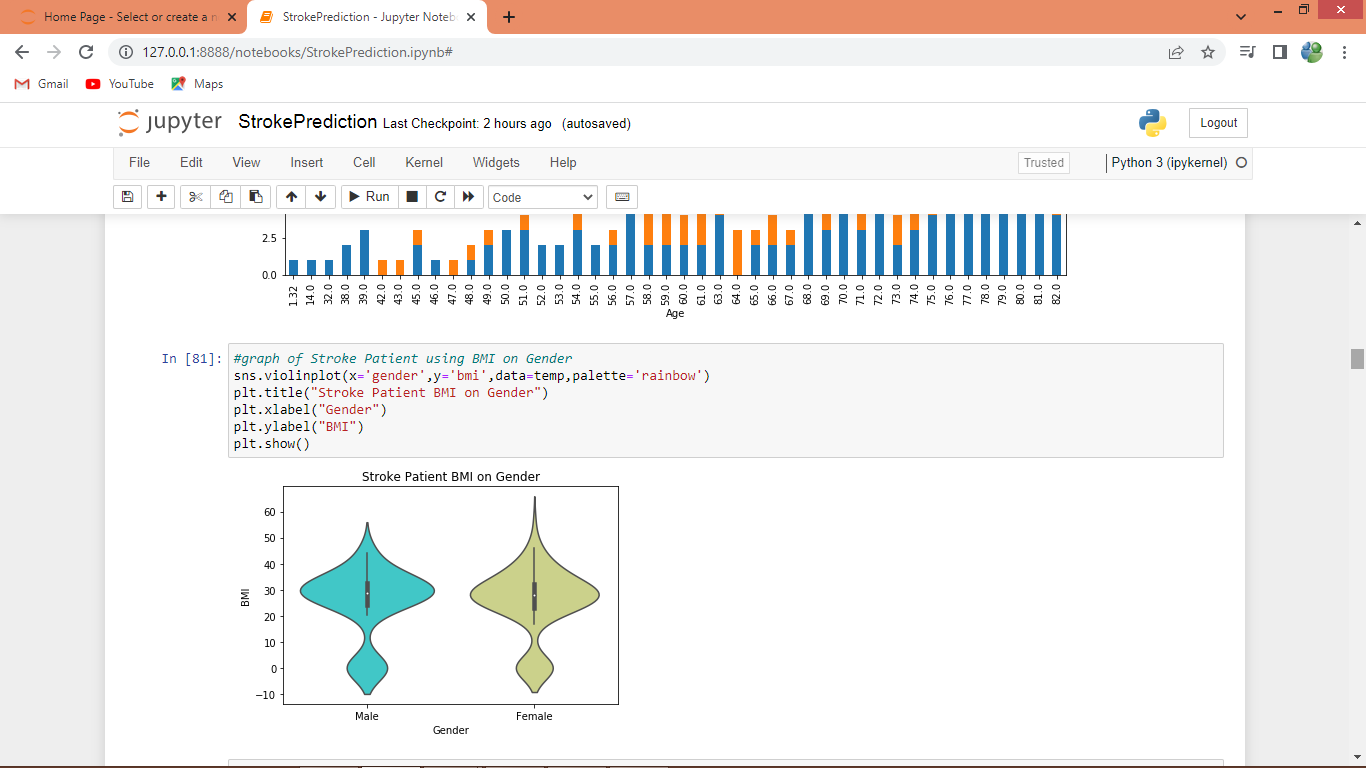
In above graph finding and displaying cluster features correlation graph and all values are not highly correlates. High correlated means features will have score more than 90%



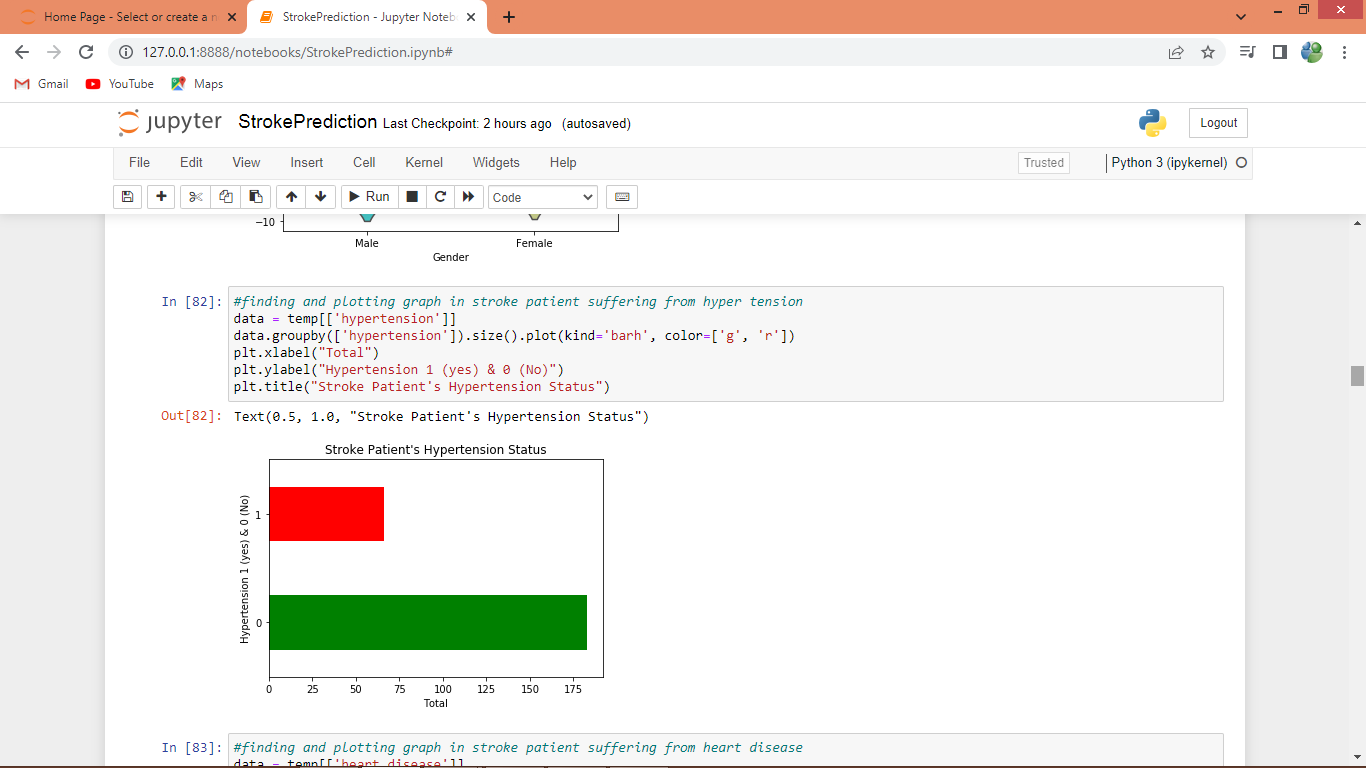
In above graph we displaying gender with strokes on different age where x-axis represents Gender and y-axis represent age



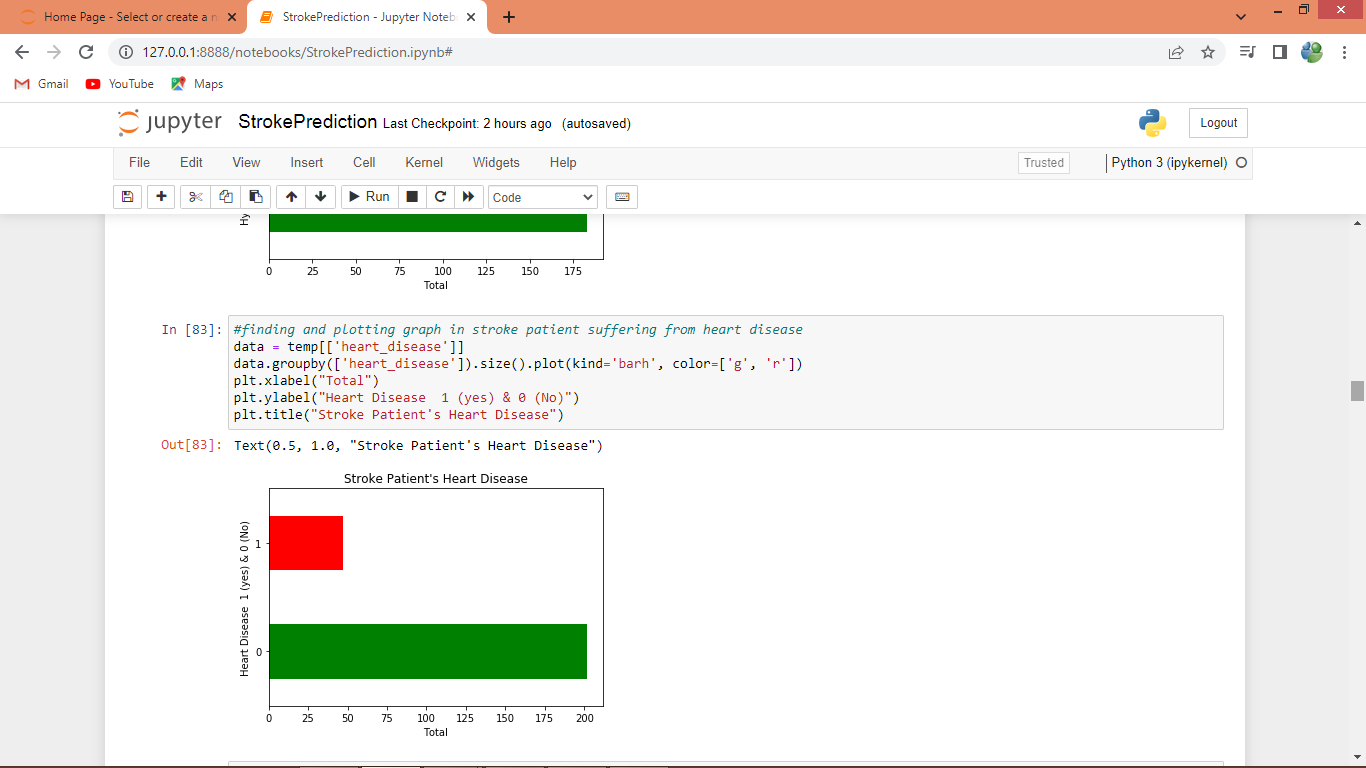
In above bar graph x-axis represents Age and y-axis represents stroke count where blue stack part is for Female and orange for Male



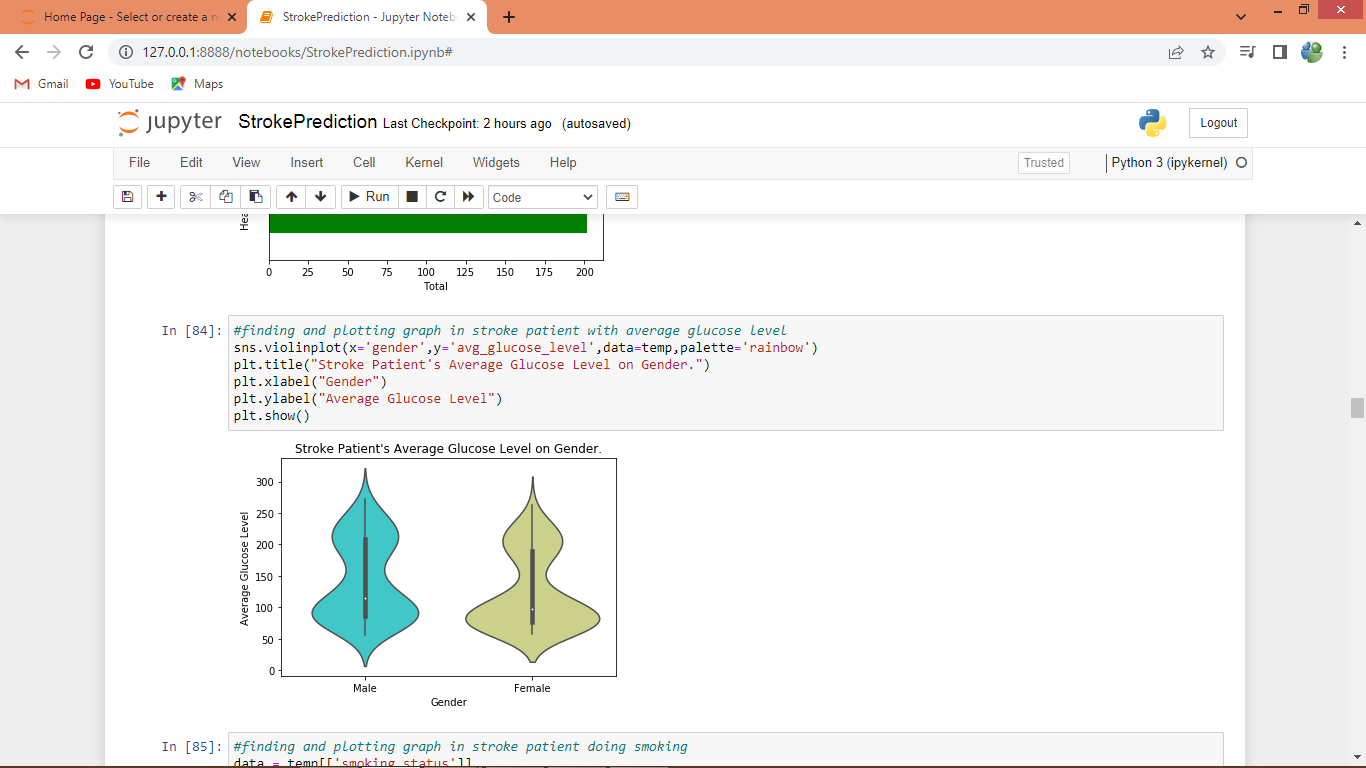
Above graph displaying gender and BMI on stroke patients



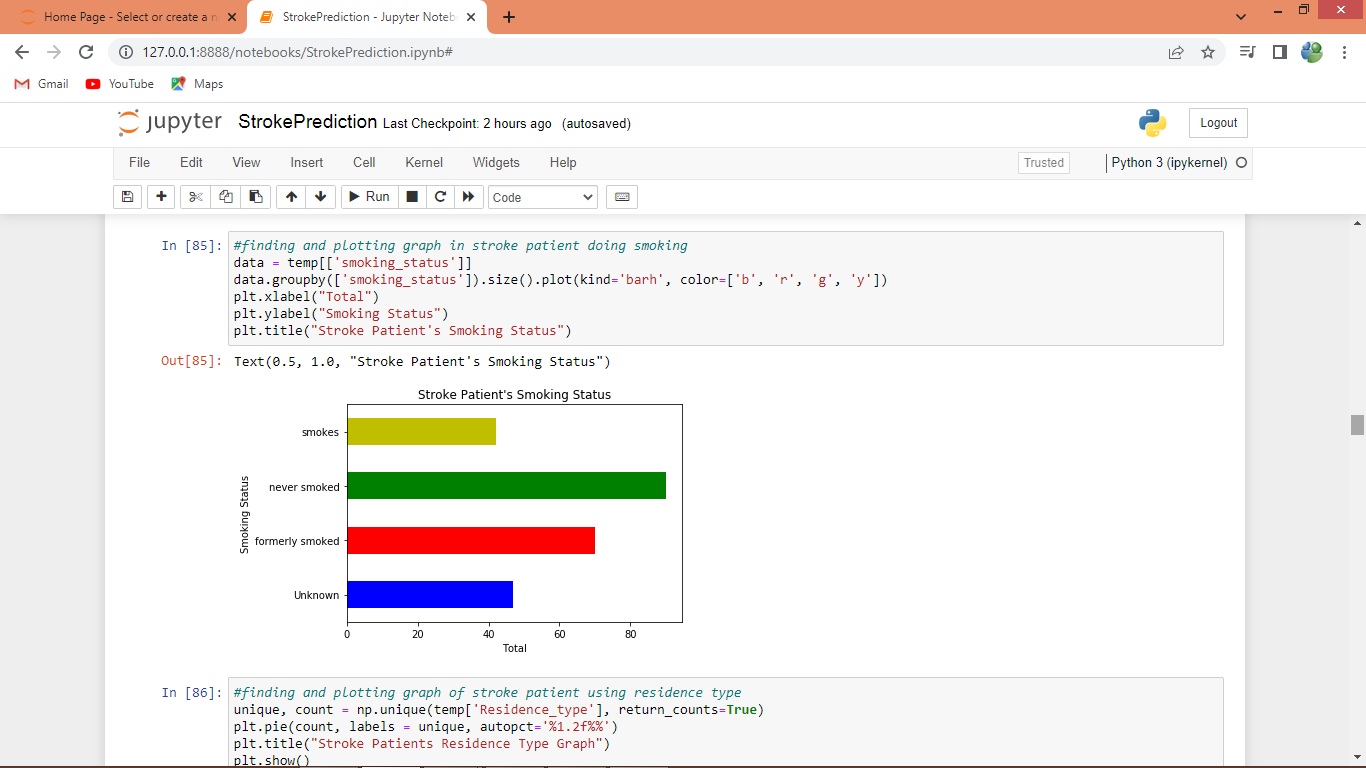
In above graph displaying number of stroke patients suffering from hypertension where in Y-axis 0 means No hyper tension and 1 means hyper tension and x-axis represents count



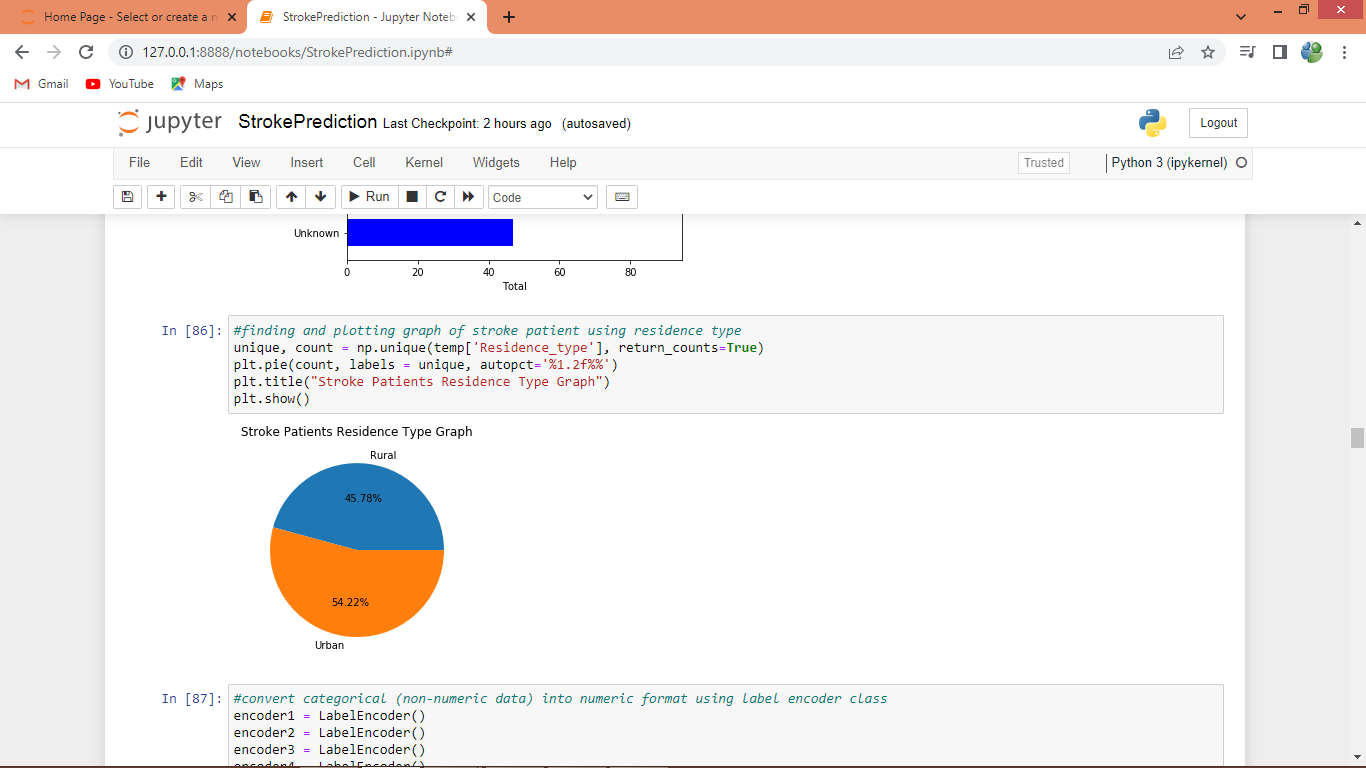
In above graph displaying number of stroke patients suffering from heart disease where in Y-axis 0 means No Heart Disease and 1 means Heart Disease and x-axis represents count



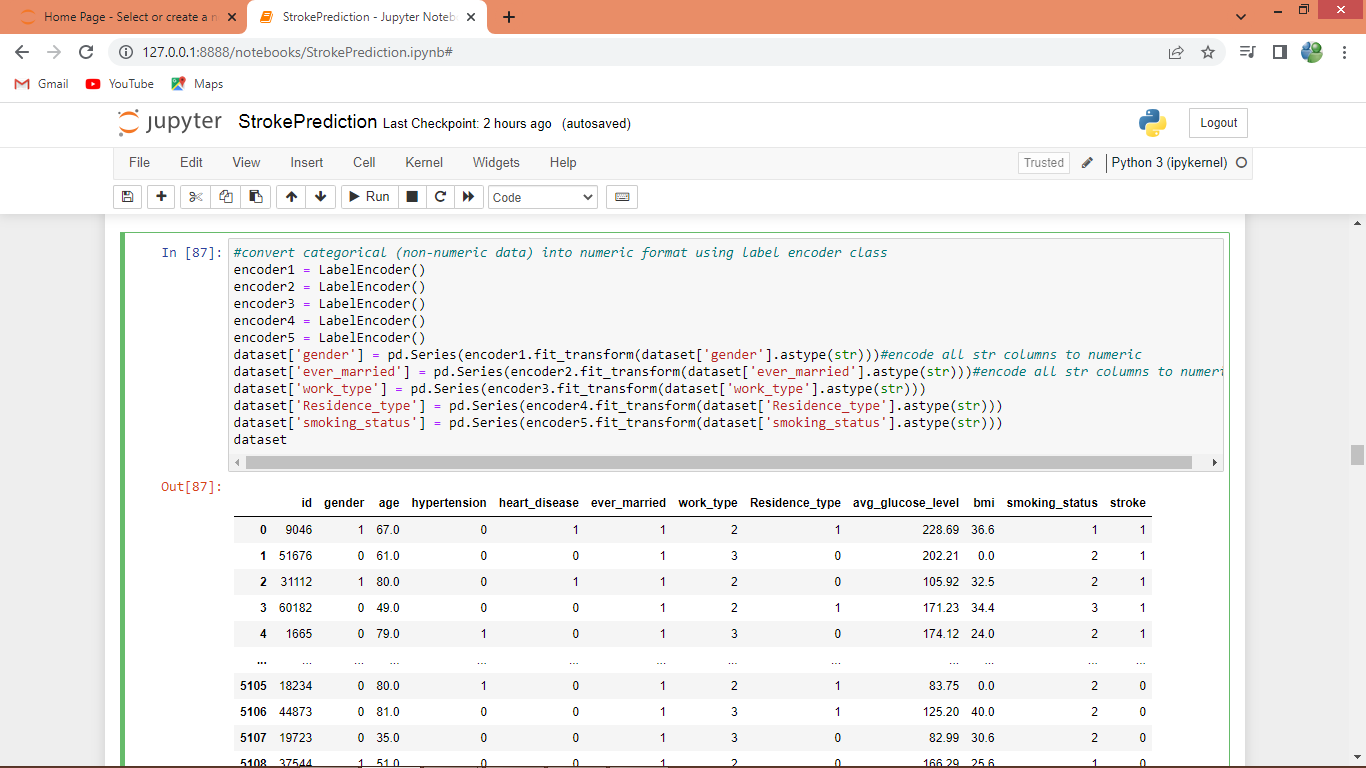
In above graph showing average glucose level on gender in stroke patients



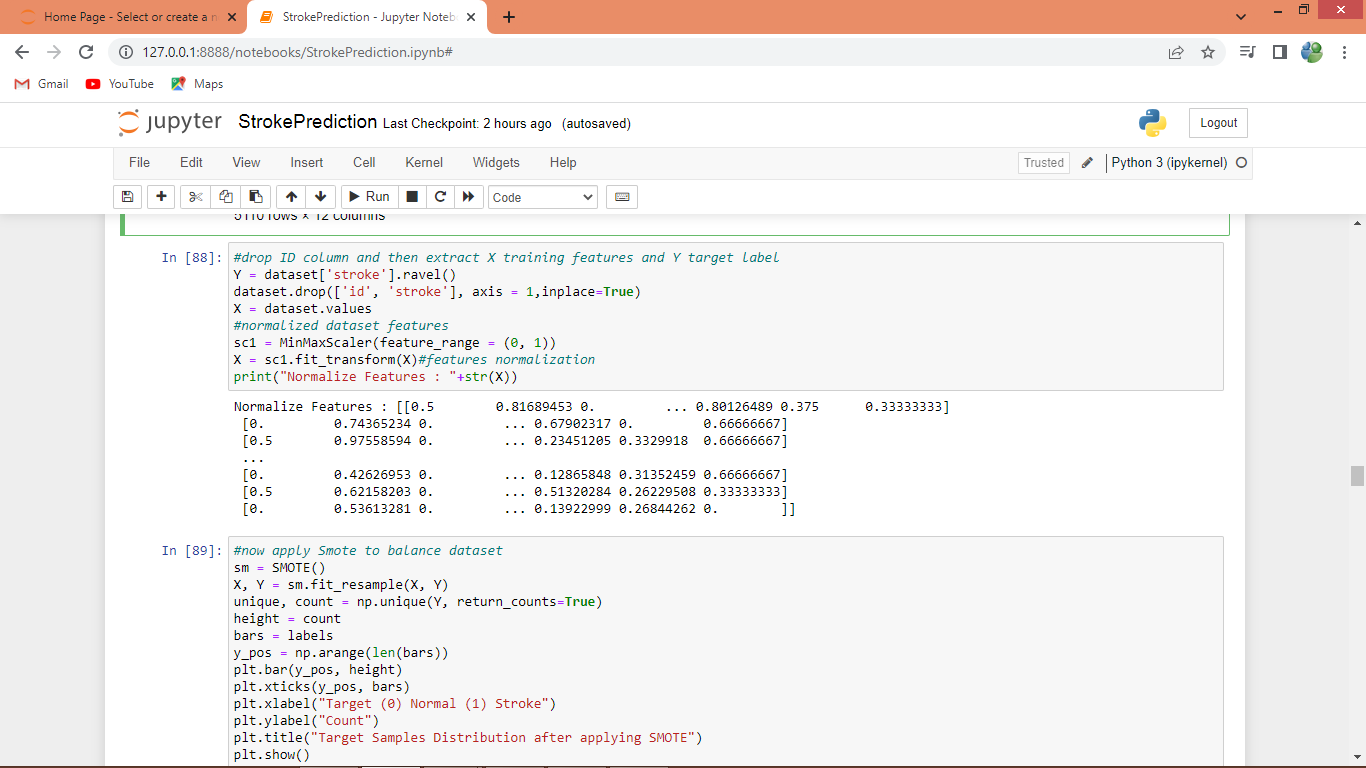
In above graph displaying number of stroke patients with smoke status



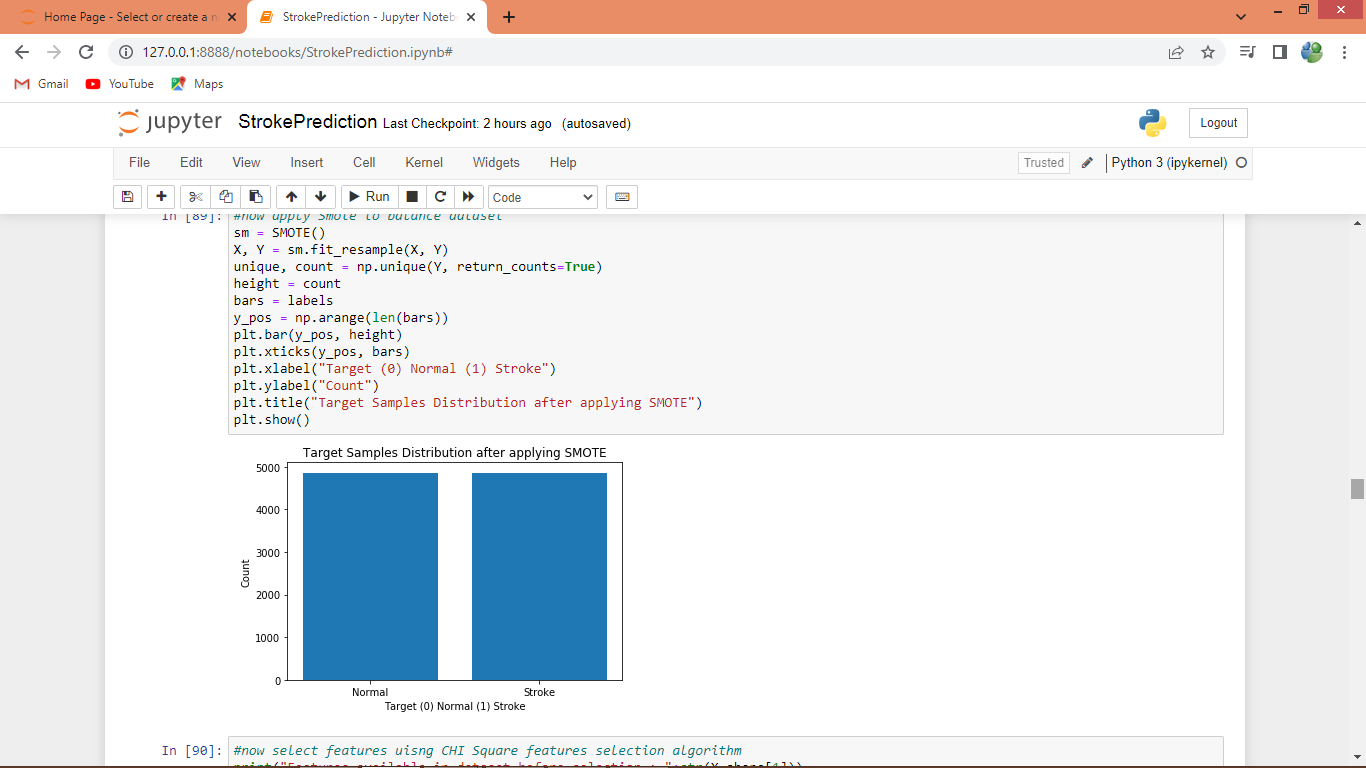
In above graph displaying residence type of stroke patients



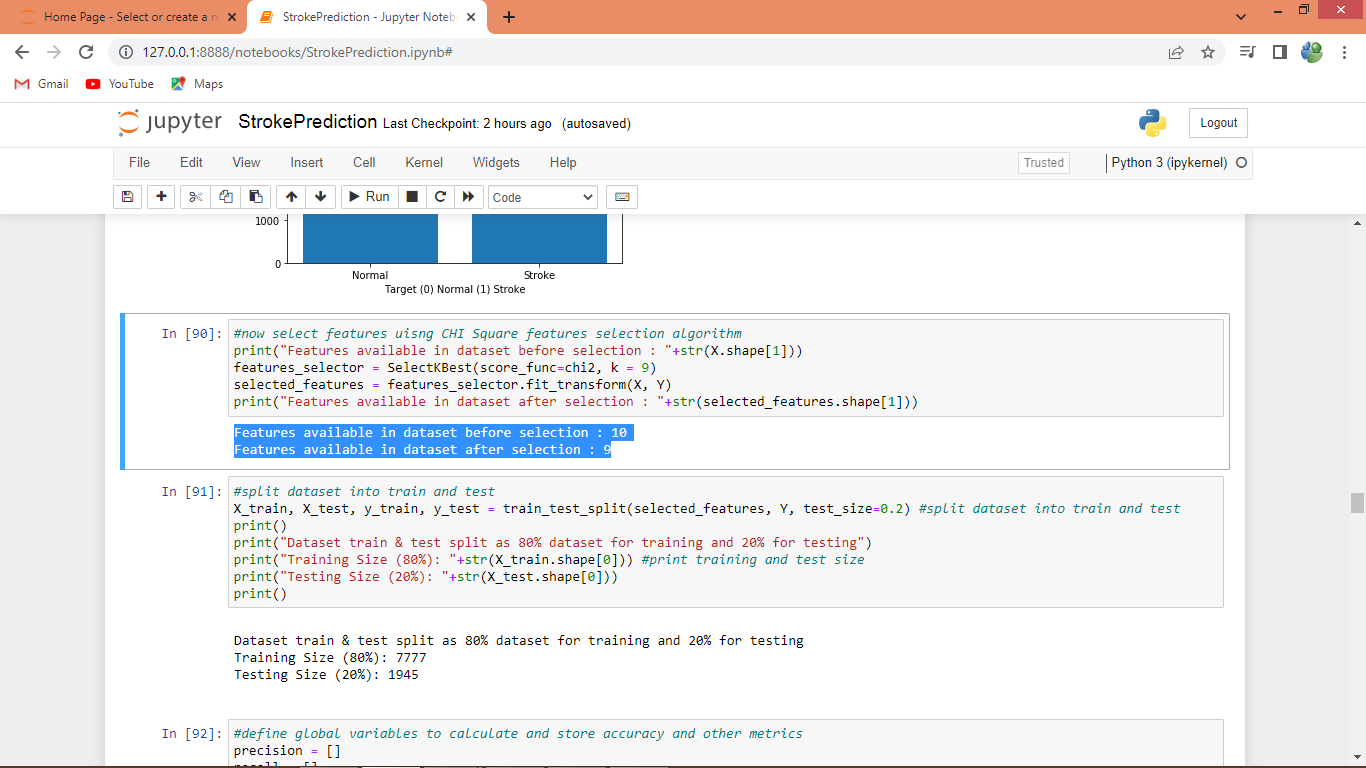
In above screen converting all categorical data to numeric format to avoid data leakage



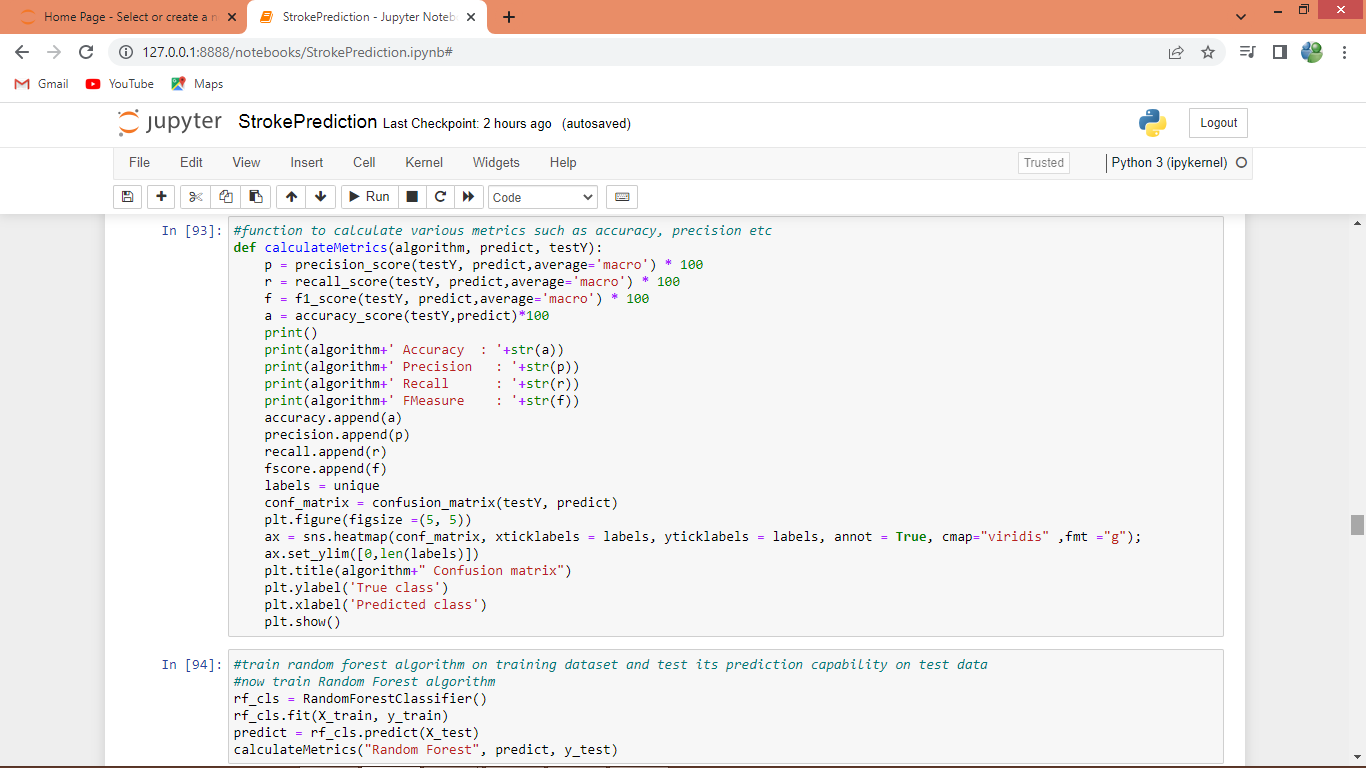
In above screen normalizing the dataset features



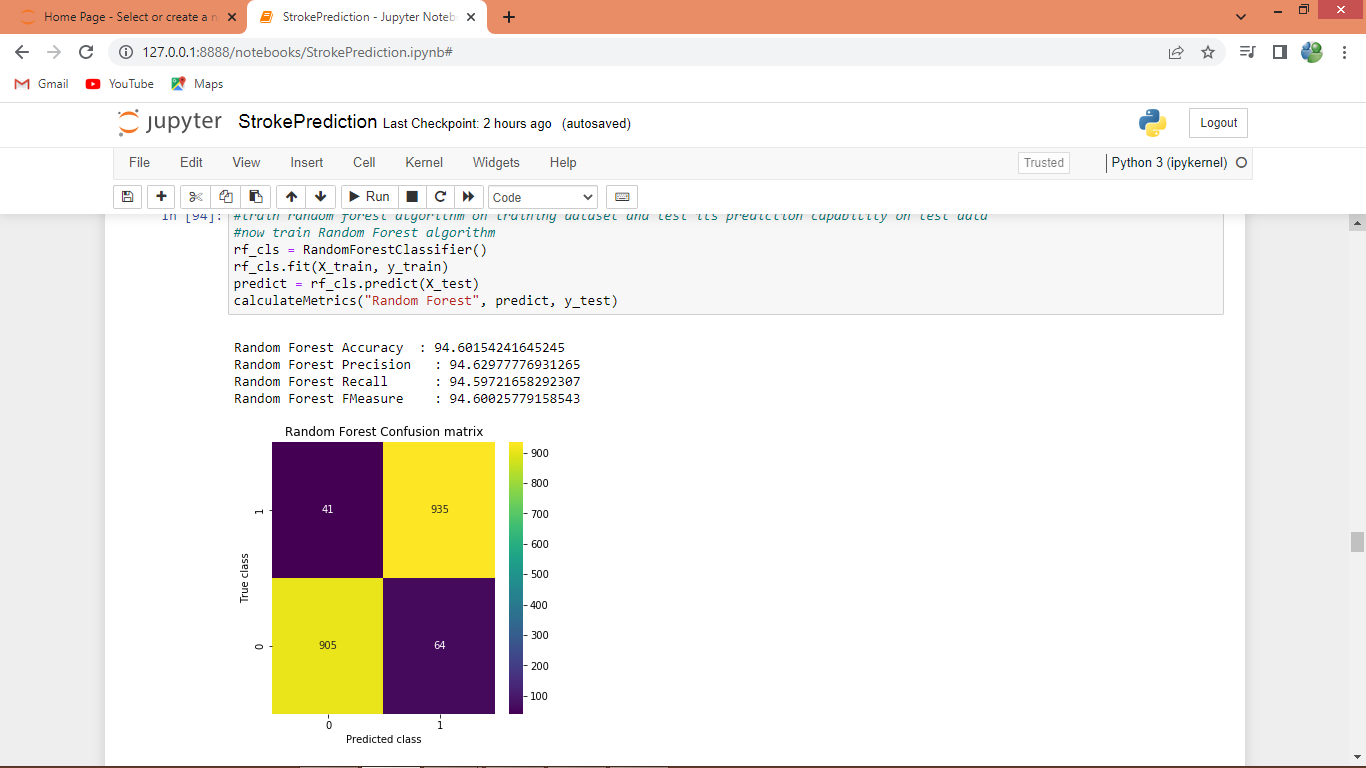
In above screen after applying SMOTE we can see both classes has equal number of records



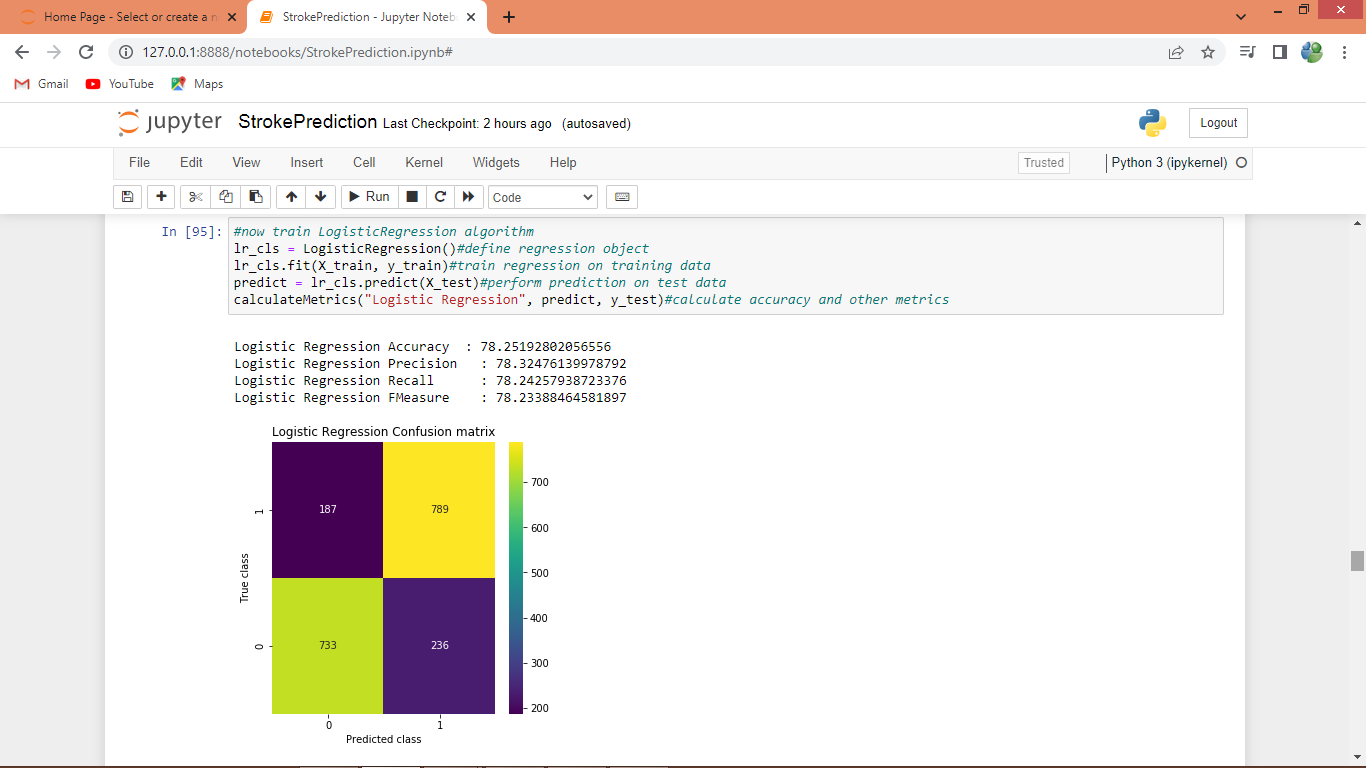
In above screen in blue colour text displaying number of features before and after CHI2 execution and then splitting dataset into train and test



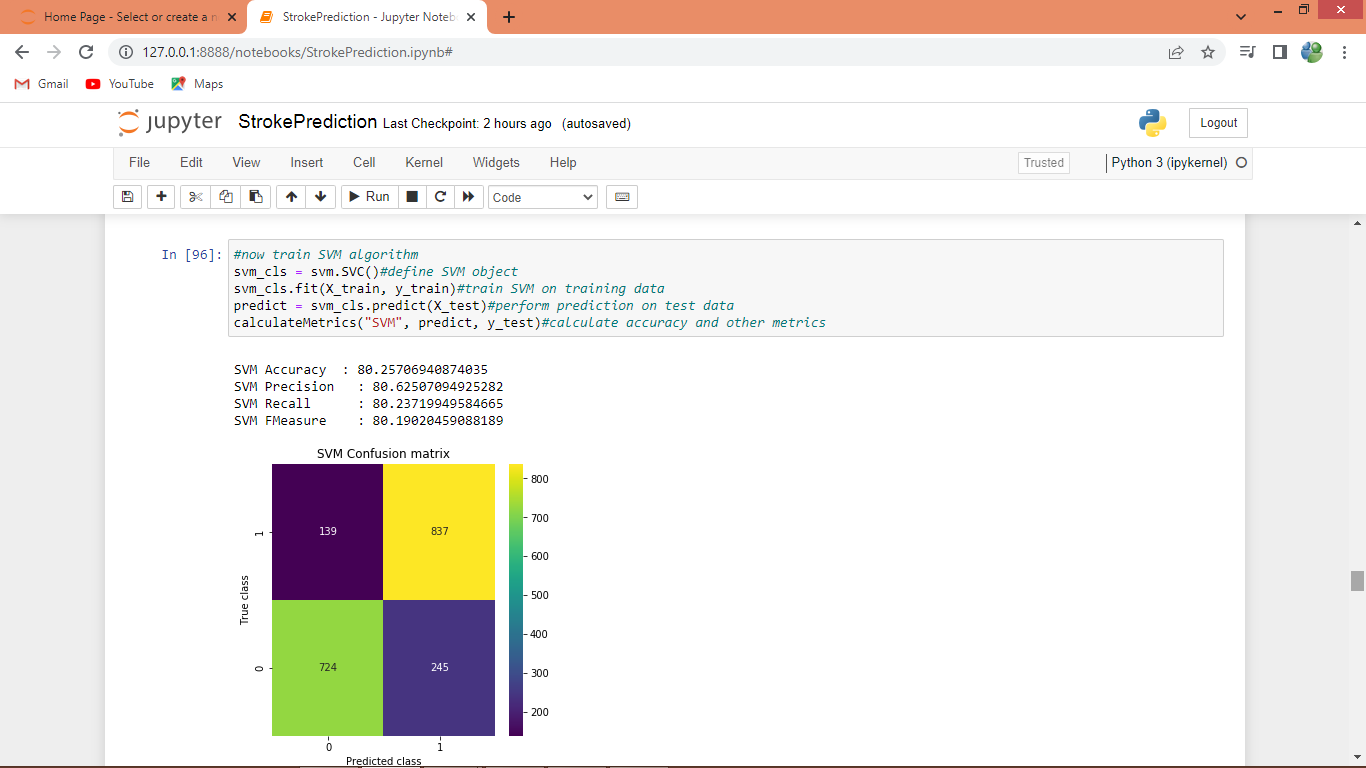
In above screen defining function to calculate accuracy and other metrics



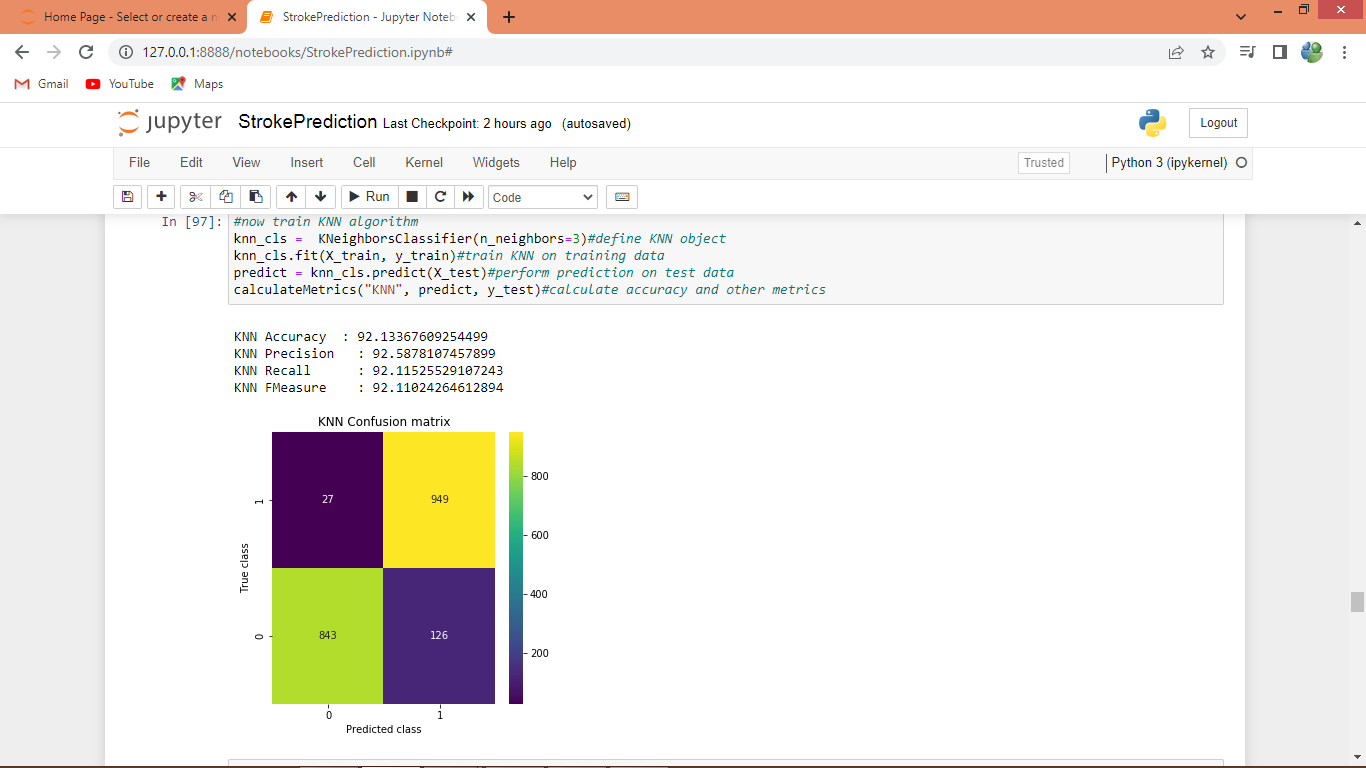
In above screen training Random Forest algorithm on training data and then performing prediction on test and then random forest got 94% accuracy and can see other metrics like precision, recall etc. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count which are very few

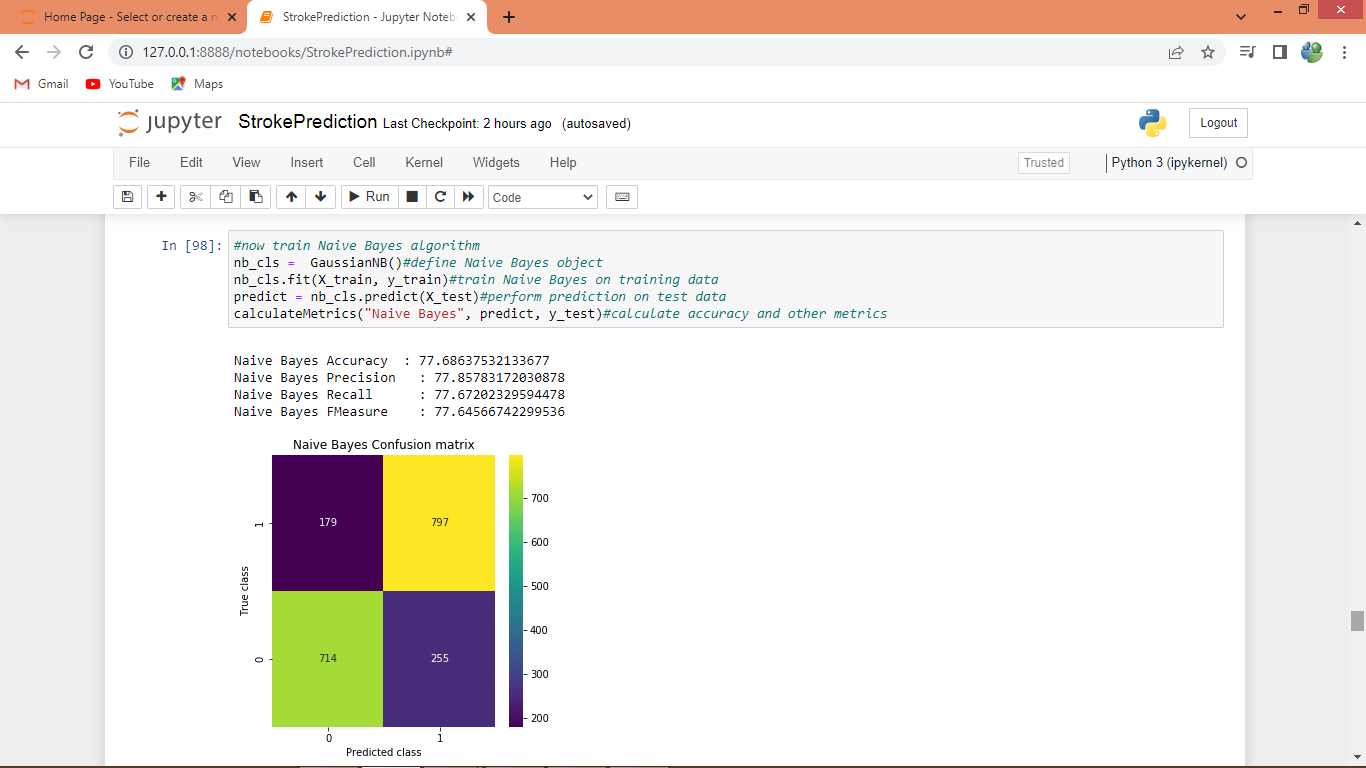


In above screen Logistic Regression got 78% accuracy

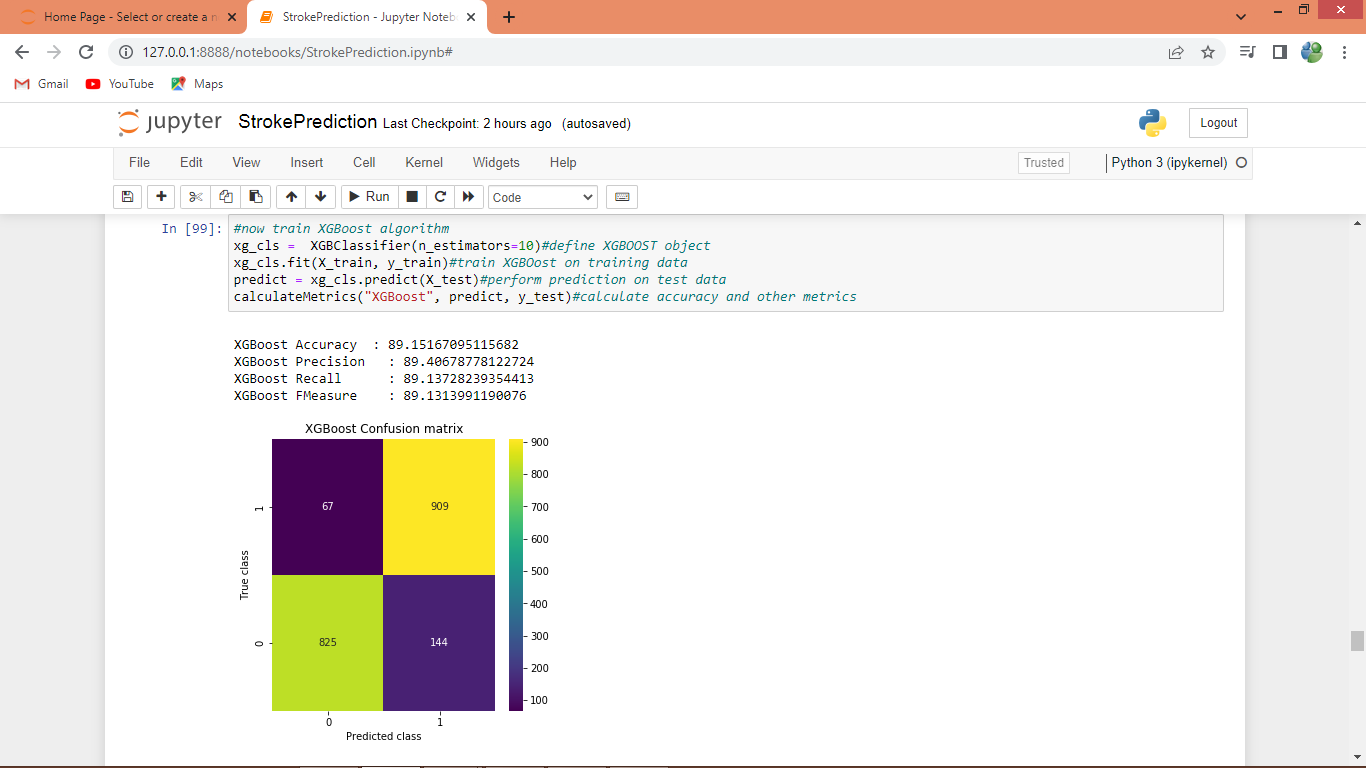


In above screen SVM got 80% accuracy

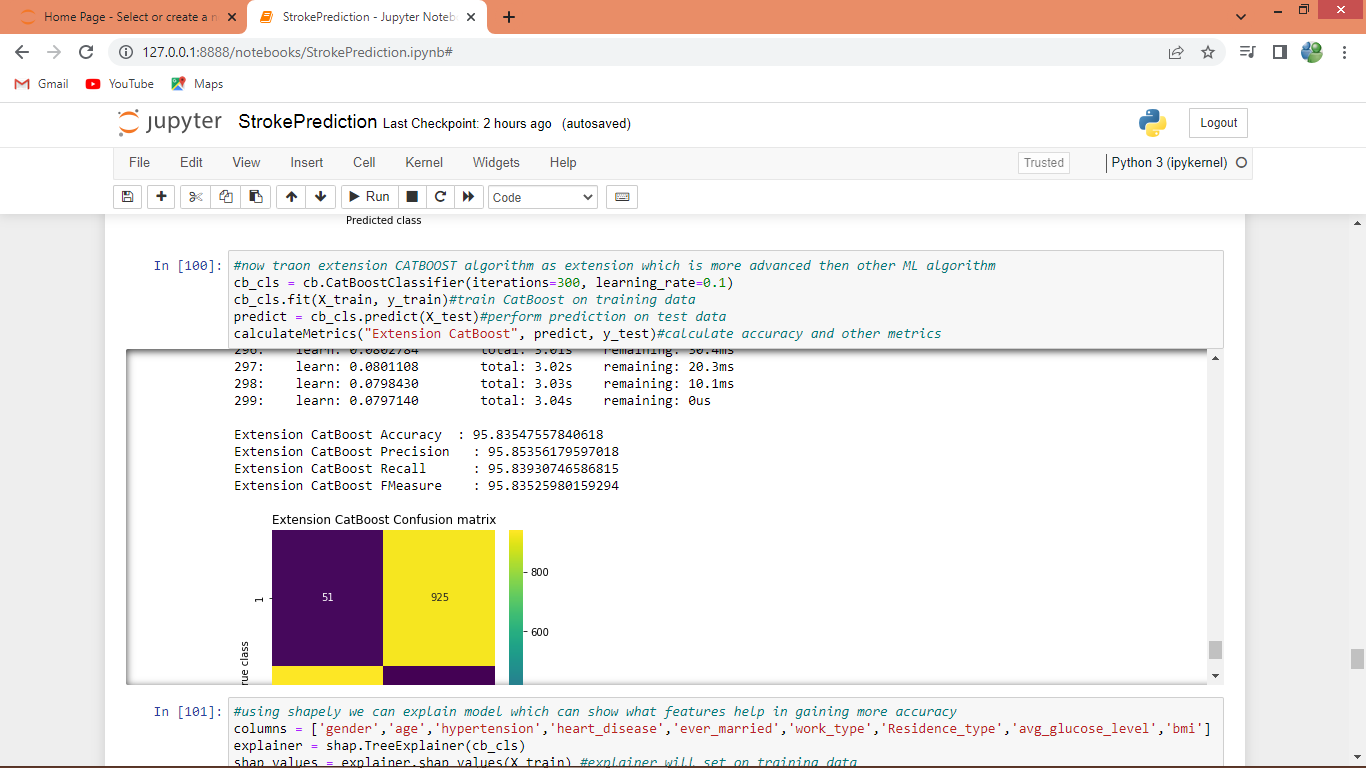
In above screen KNN got 92% accuracy



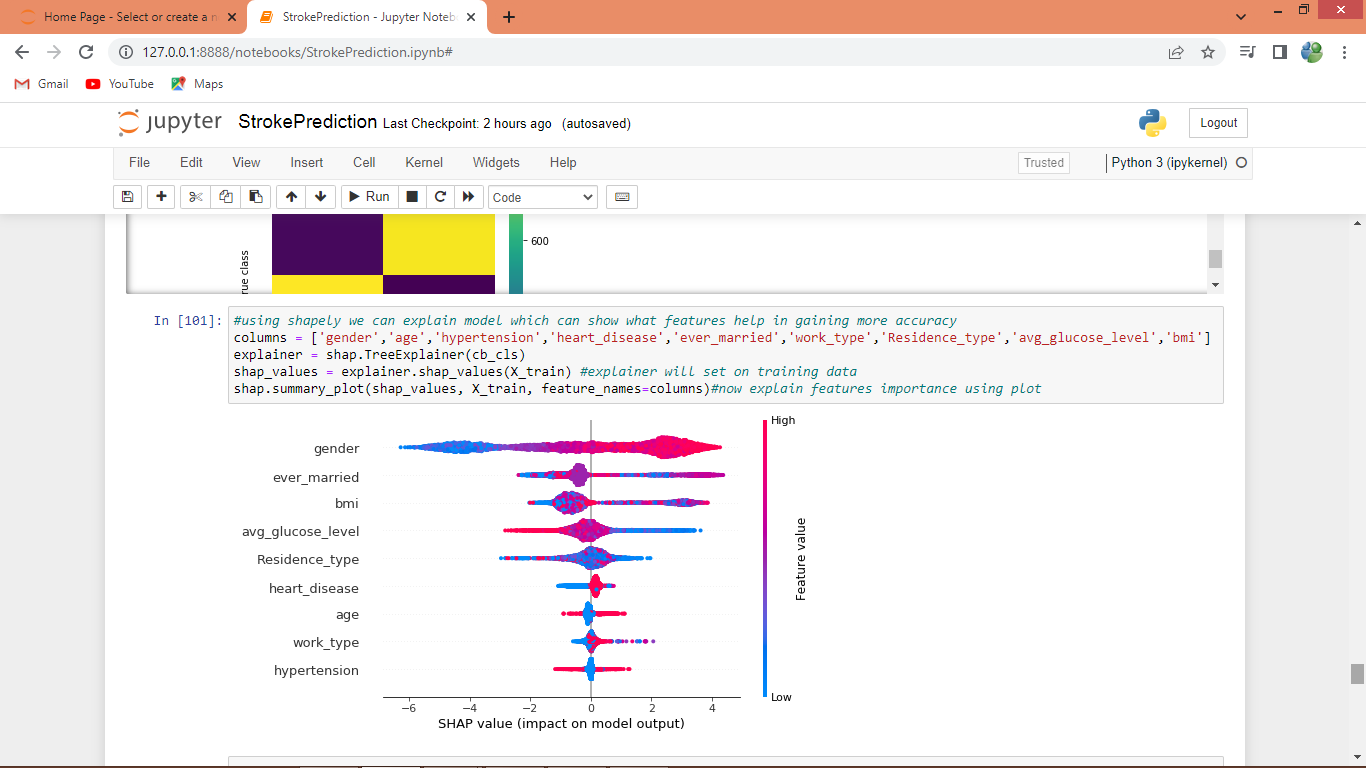
In above screen Naive Bayes got 77% accuracy



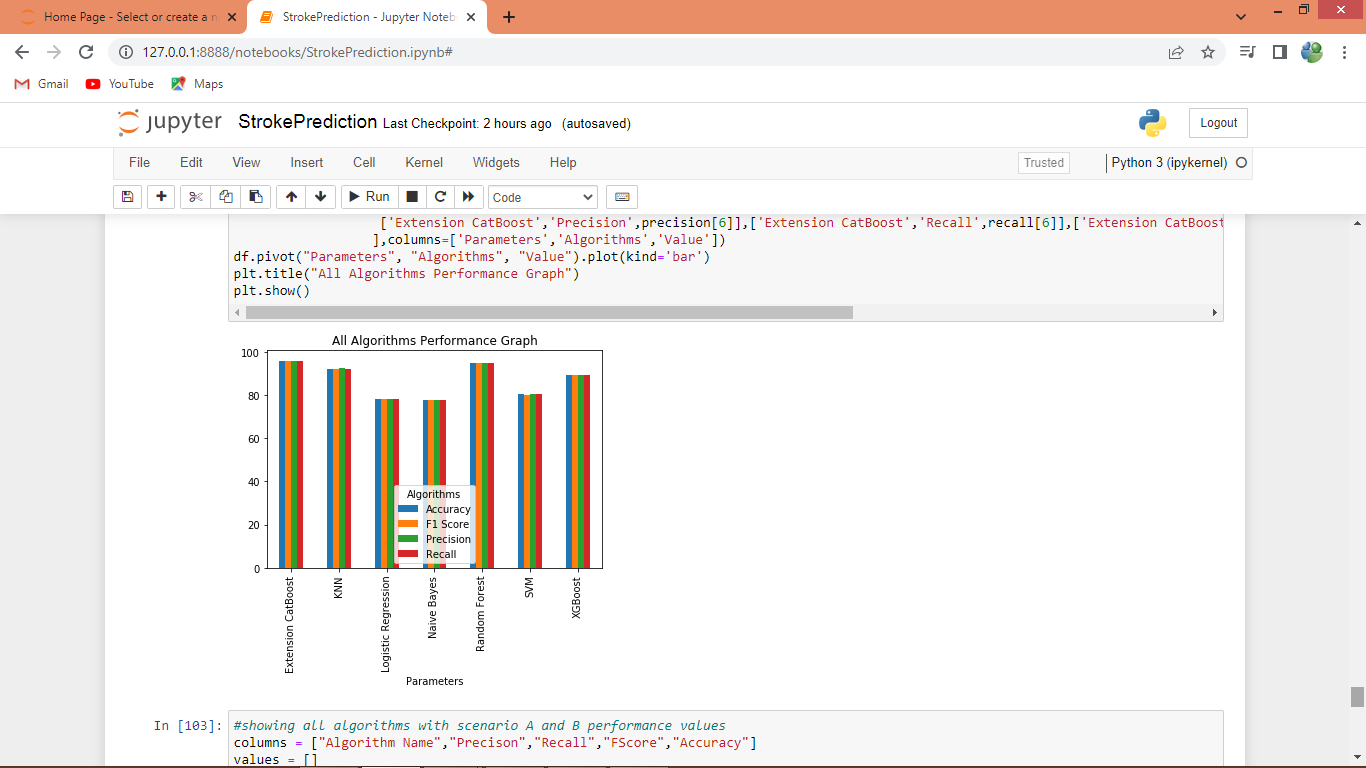
In above screen XGBOOST got 89% accuracy



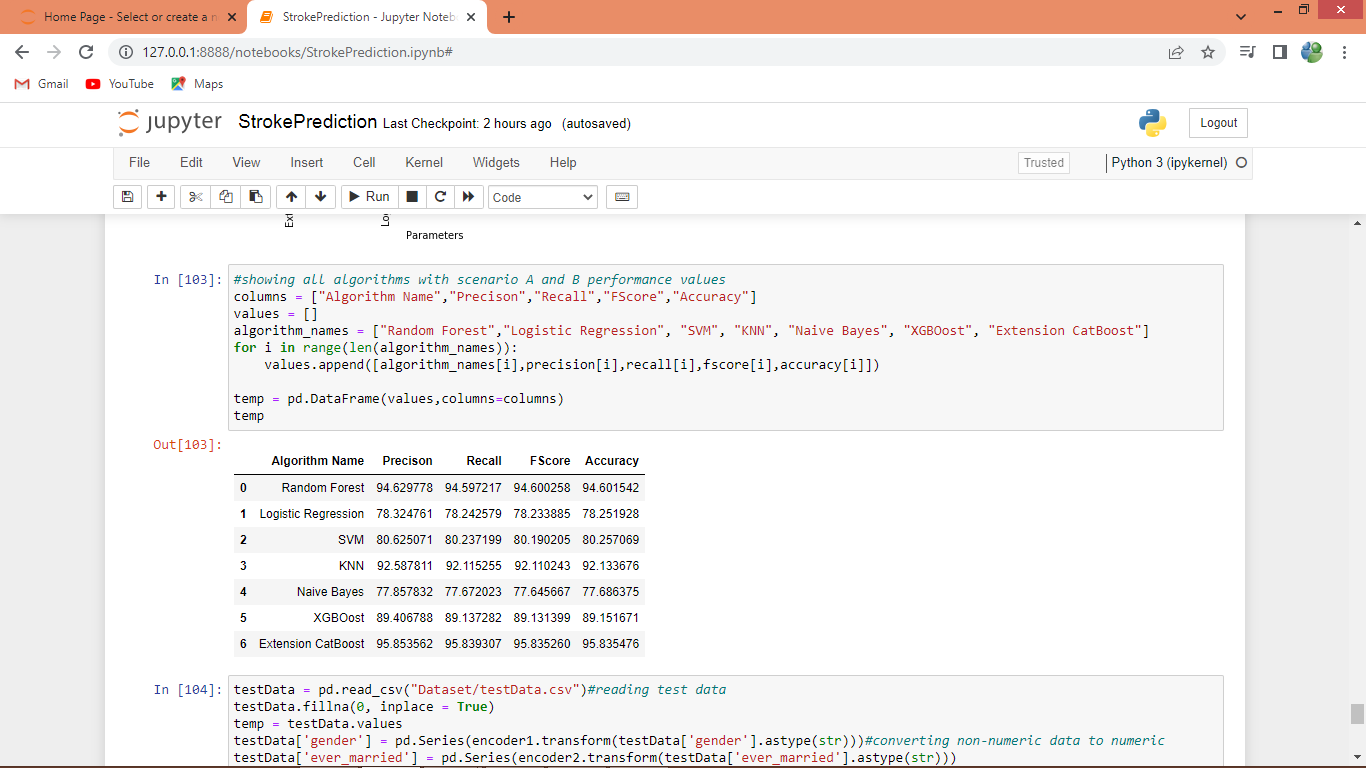
In above screen extension CATBOOST got 95% accuracy which is higher than other algorithms



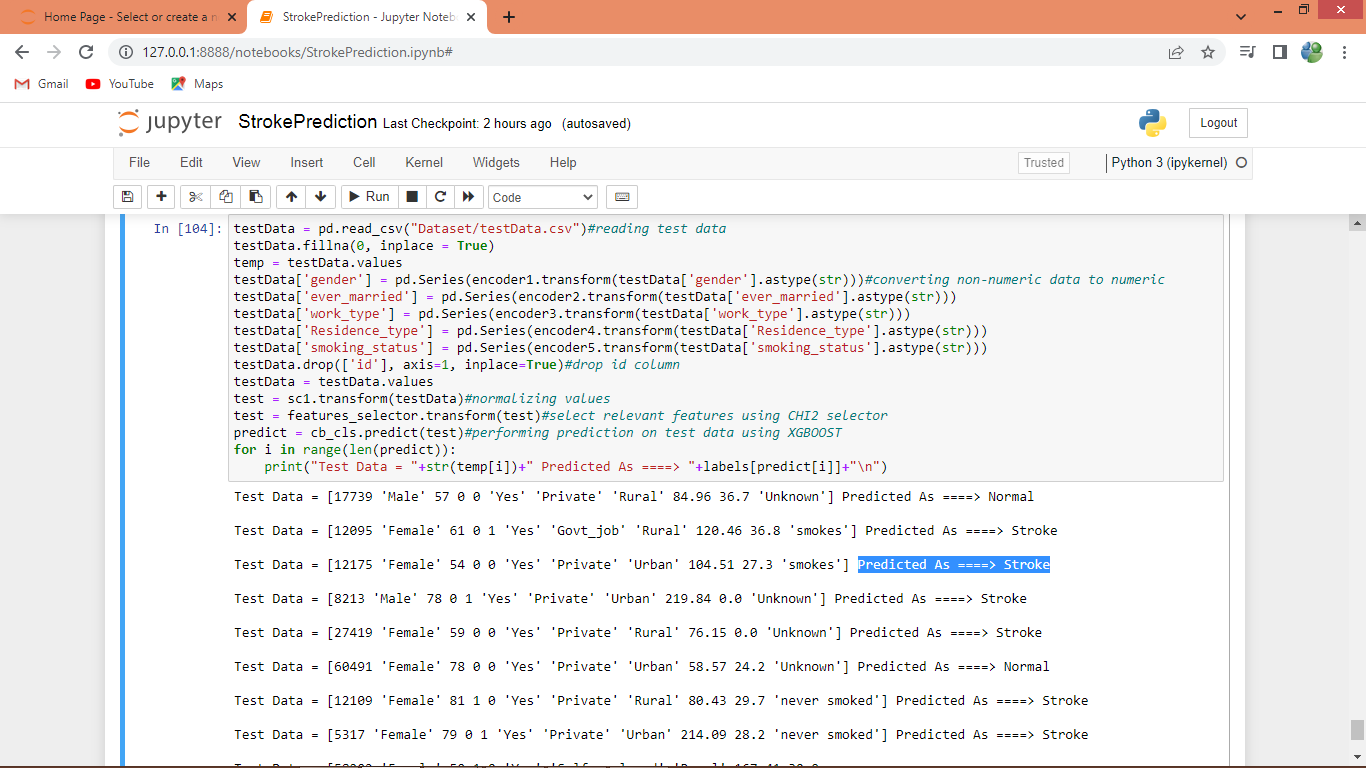
In above screen shapely explaining about features which are contributing most in correct prediction and then features whose graph reaching to high are the most relevant features used for prediction



In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms extension CATBOOST got high accuracy



In above screen displaying all algorithms performance in tabular format



In above screen reading test data, normalizing, features encoding from categorical to numeric format, removing missing values, features selection and normalization and then processed features are predicting with extension CATBOOST algorithm and in output before =🡺 arrow symbol we can see TEST data and after arrow symbol we can see predicted data as ‘Normal or Stroke’