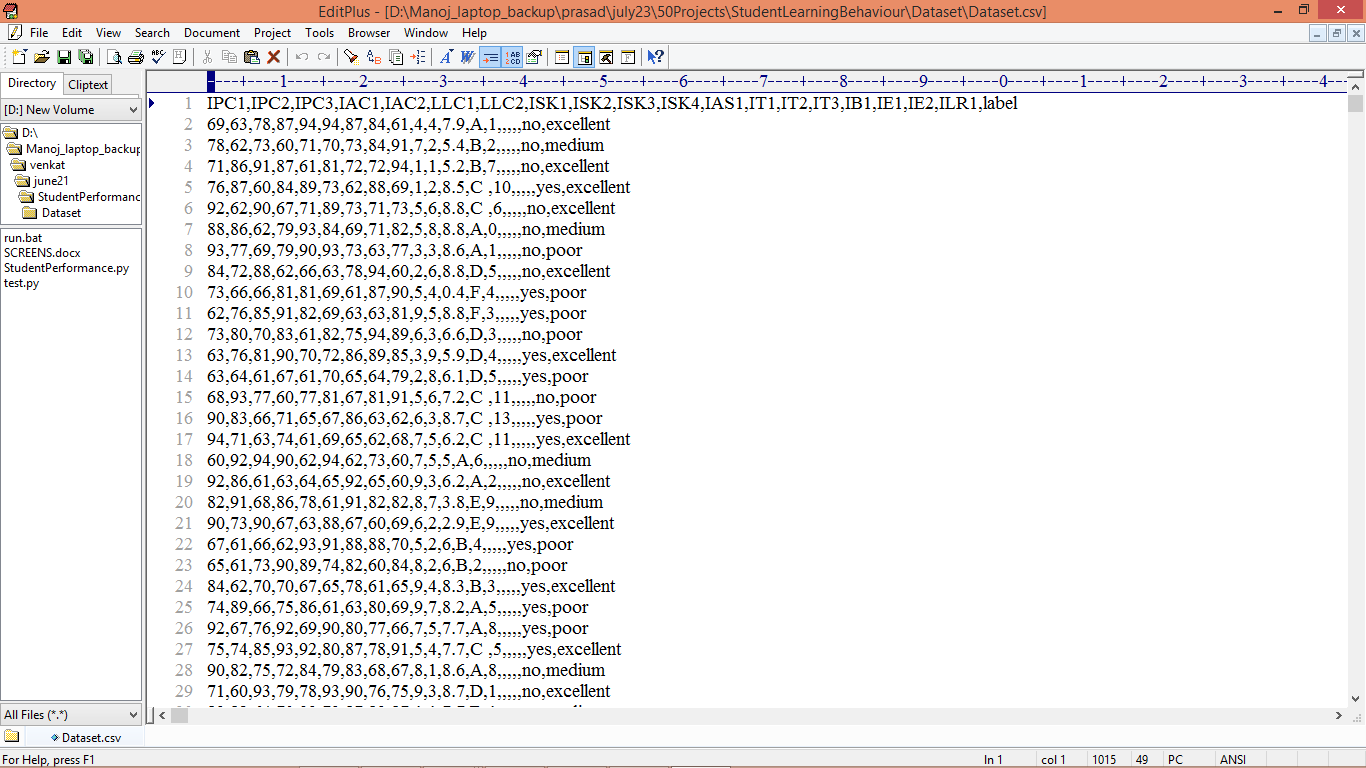
Analysis of Learning Behaviour Characteristics and Prediction of Learning Effect for Improving College Students’ Information Literacy Based on Machine Learning

Information literacy refers to student abilities on problem solving, understanding, critical thinking and many more brain based activities. In propose paper author saying 21st century student must have information literacy knowledge in solving problems, so academic tutors must have knowledge of those Students who are best or worst in information literacy, so tutors can take necessary decision to work on underperforming students to enhance their problem solving literacy.

Propose paper analyzes the characteristics of college students’ learning behaviours and explores the predictive learning effect by constructing a predictive model of learning effect based on information literacy learning behaviour characteristics. In propose paper to predict college student behaviour author employing Pearson Correlation algorithm to select relevant features from the dataset and then training this dataset with multiple machine learning algorithms such as Decision Tree, Naive Bayes, KNN, Random Forest and Neural network and then evaluate each algorithms performance in terms of accuracy, precision, recall and FSCORE. In all algorithms Random Forest is giving best accuracy.

To train all algorithms author is using student behaviour dataset and then dividing that dataset into 3 different INDICATORS such as ‘Knowledge & Skills, Application & Innovation and Morality Responsibility’ and all this indicators are further divided into second level indicators such as ‘IPC1,IPC2,IPC3,IAC1,IAC2,LLC1,LLC2,ISK1,ISK2,ISK3,ISK4,IAS1,IT1,IT2,IT3,IB1,IE1,IE2,ILR1’. All this indicators refers to student ability of problem thinking, solving, understanding and many more. All this indicator details you can read from base paper. Each record in dataset having class label as Excellent, Medium and Poor.

Below screen showing dataset details used to implement this project

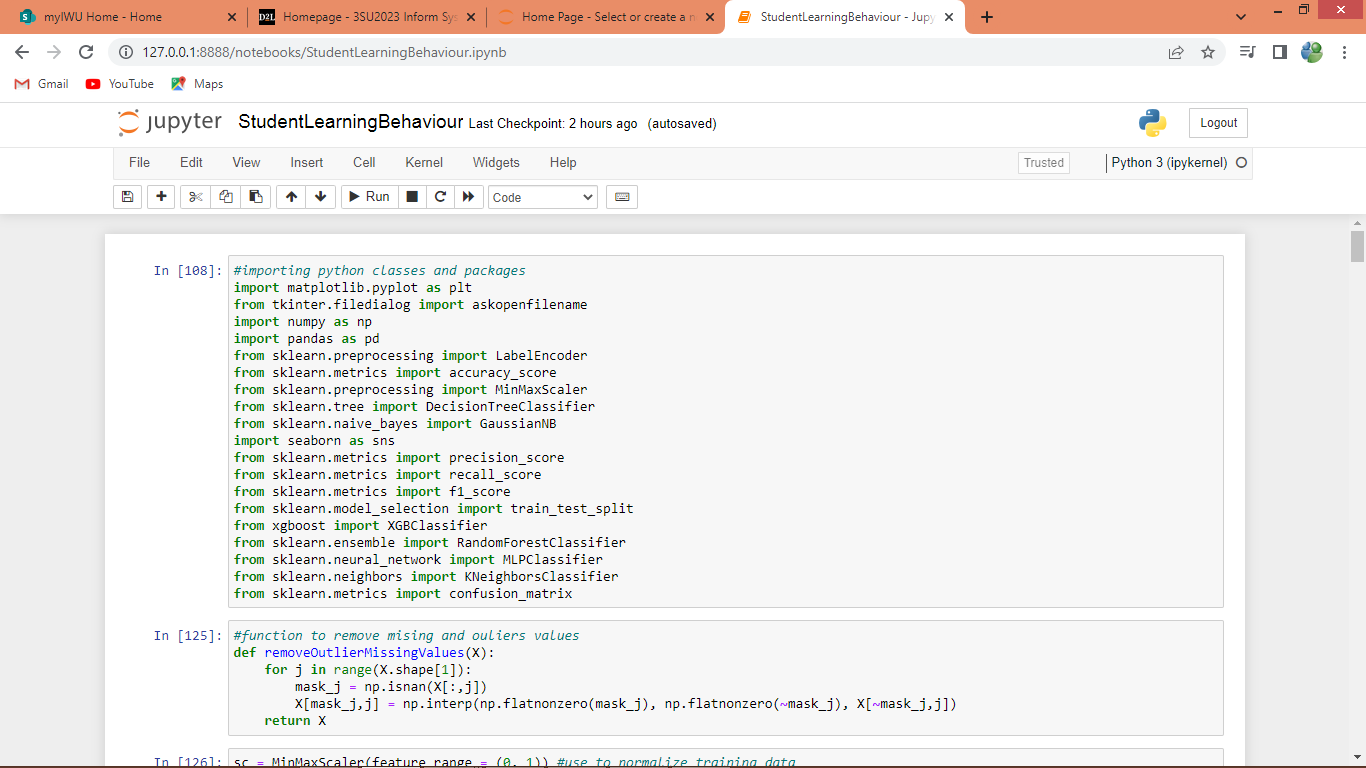


In above dataset screen first row represents dataset column names and remaining rows represents dataset values and in last column we have class label of the performing student. So by using above dataset we will train and test each algorithm performance.

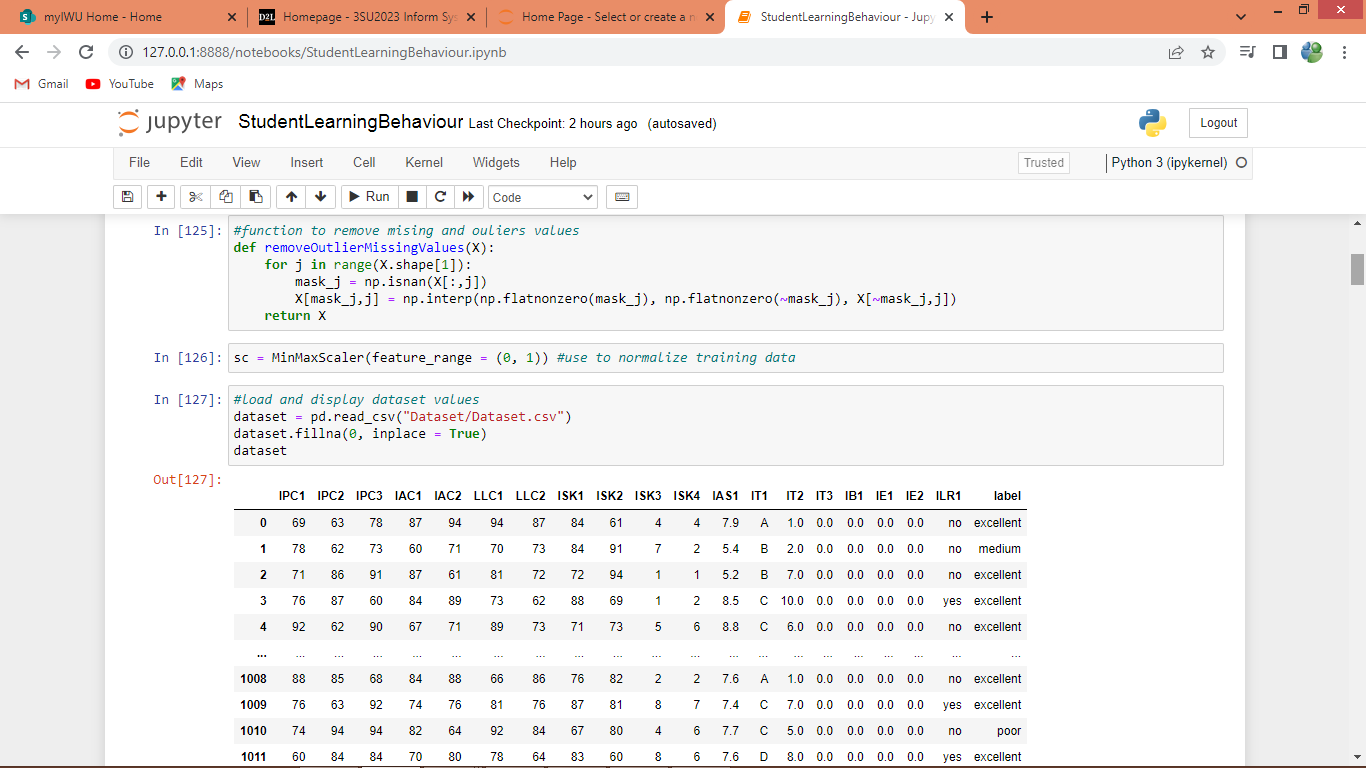
Extension Concept

In propose paper author has used all traditional machine learning algorithm but not used any advanced algorithms like CNN or LSTM or XGBOOST. So we have experimented with all algorithms but XGBOOST able to achieve high accuracy so as extension we have applied XGBOOST algorithm.

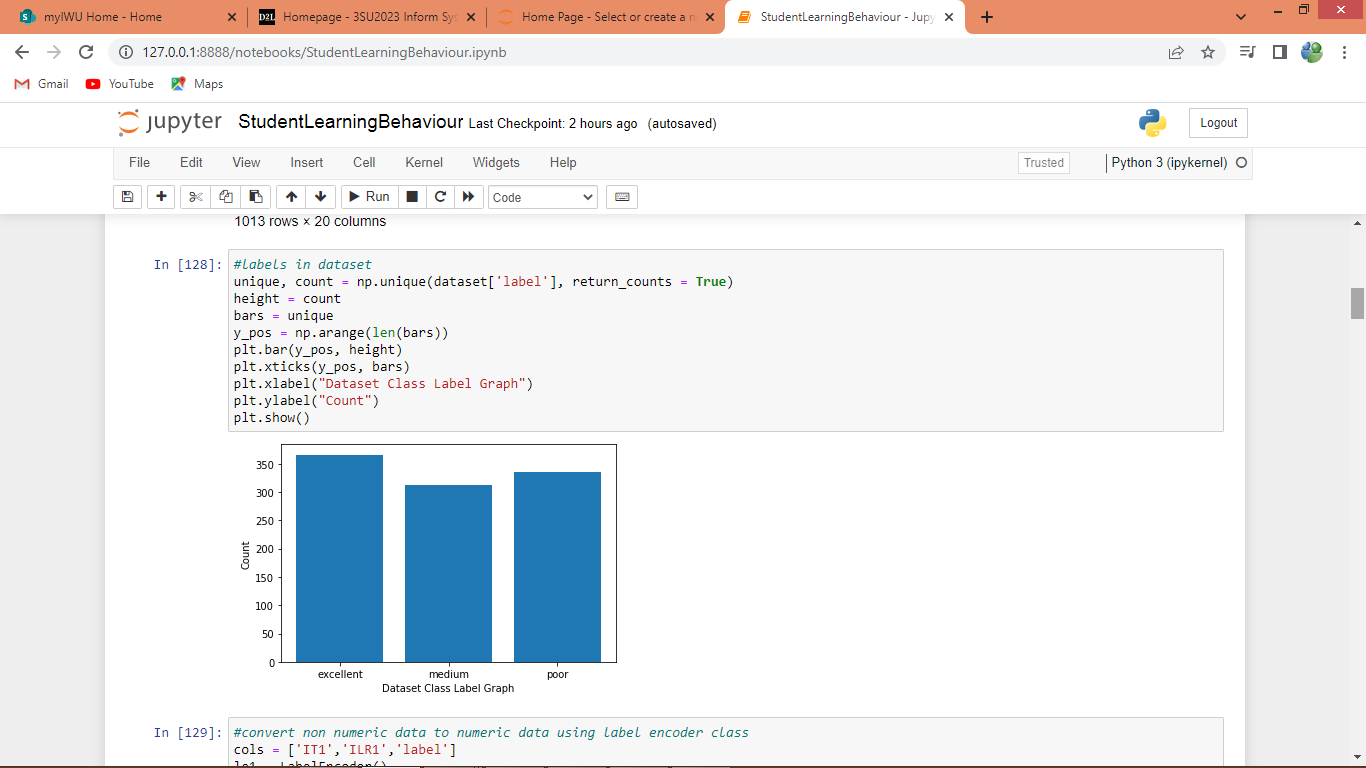
We have implemented this project using JUPYTER notebook and below are the code and output screen with blue colour comments



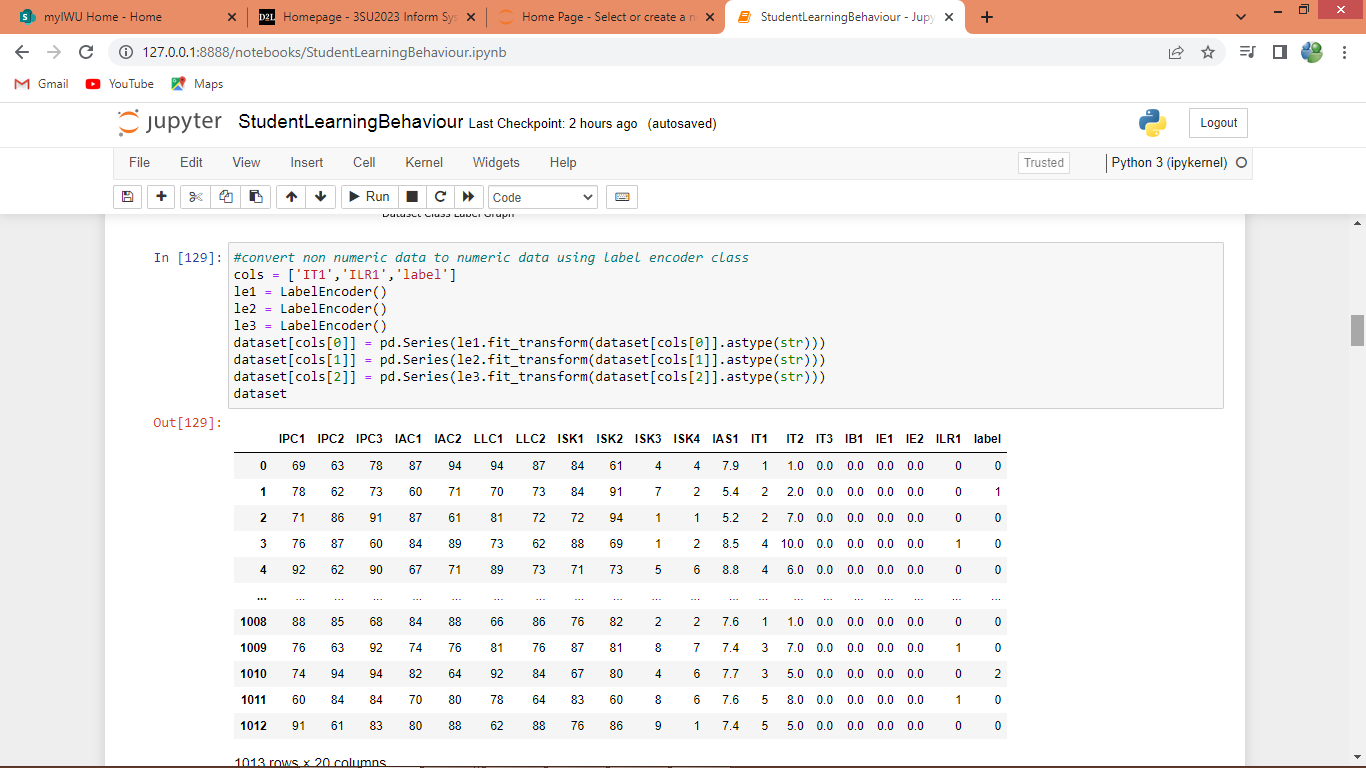
In above screen importing require classes and packages



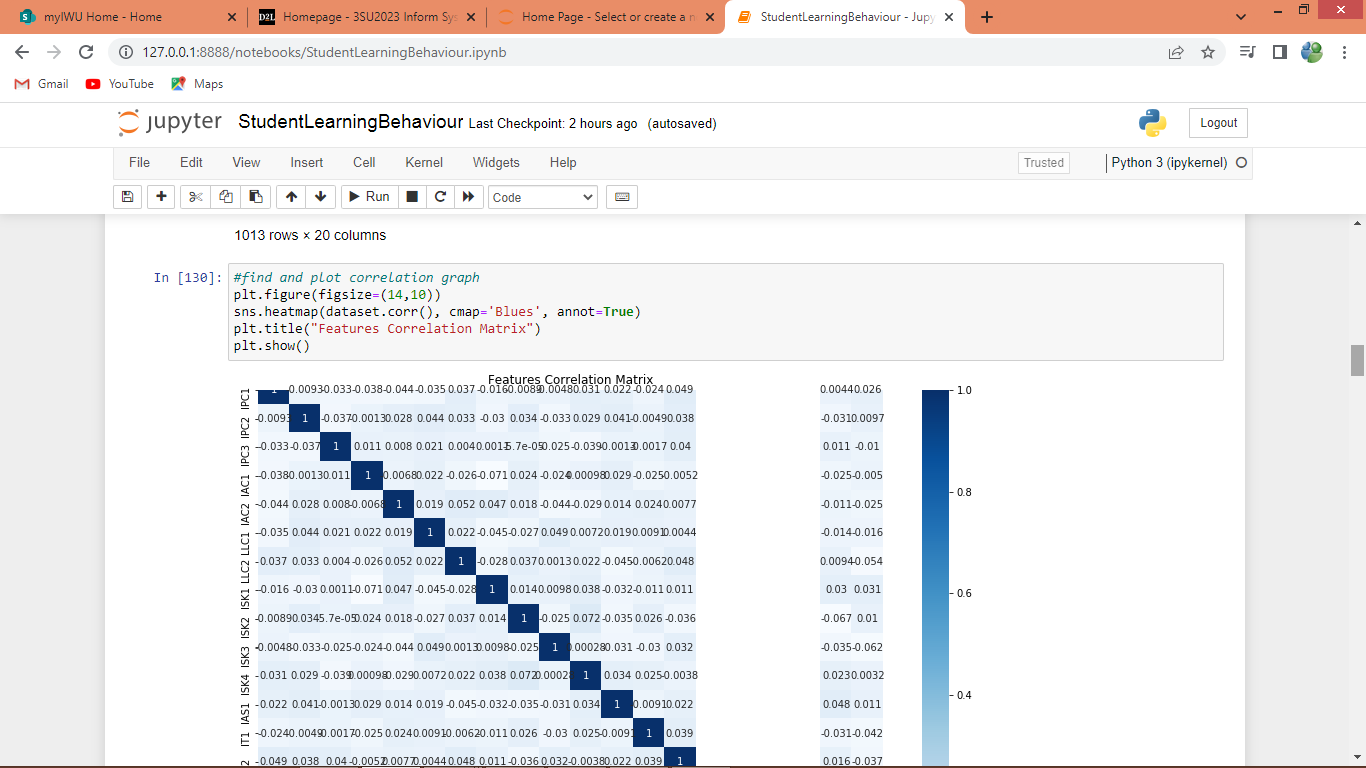
In above screen defining function to remove missing, outlier values and then defining function to normalize features and then reading and displaying dataset values



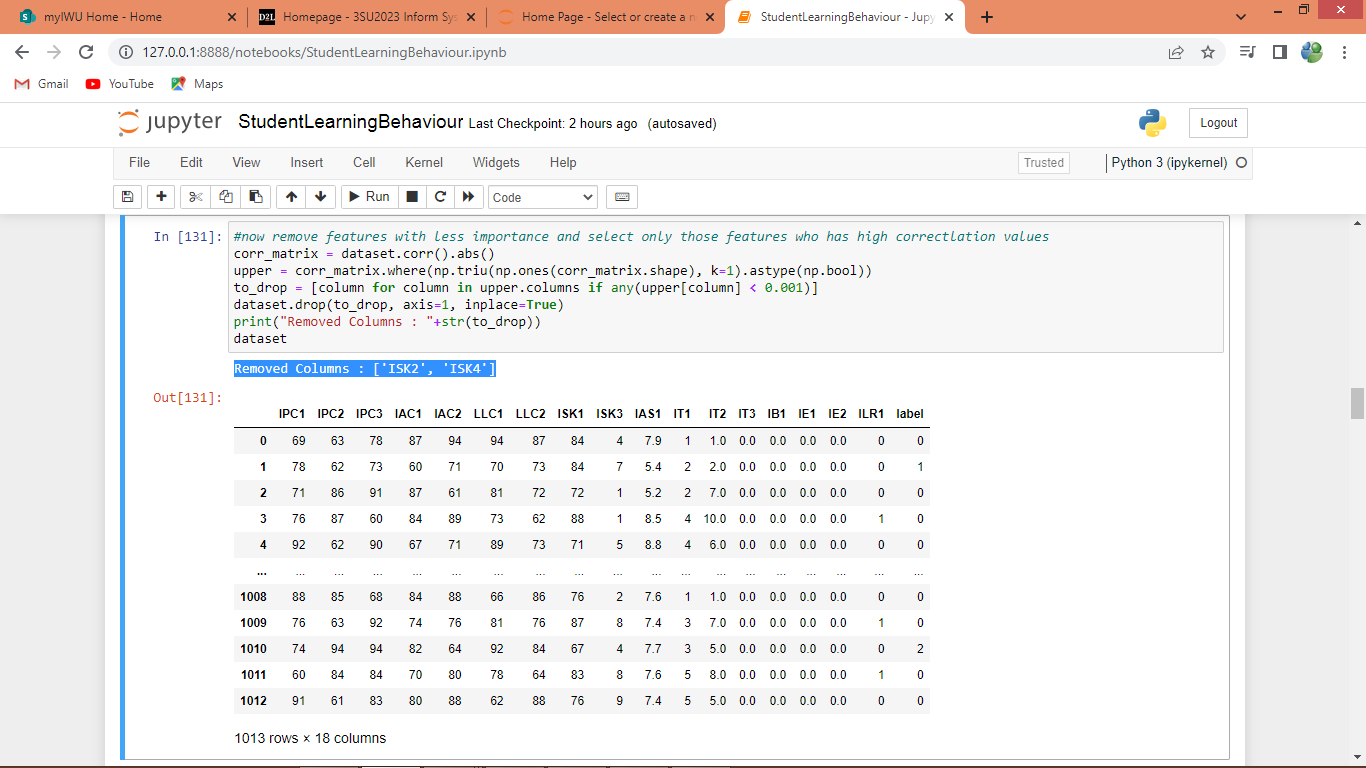
In above screen finding and plotting graph of different labels found in dataset where x-label refer to class name and y-axis represents count of class label



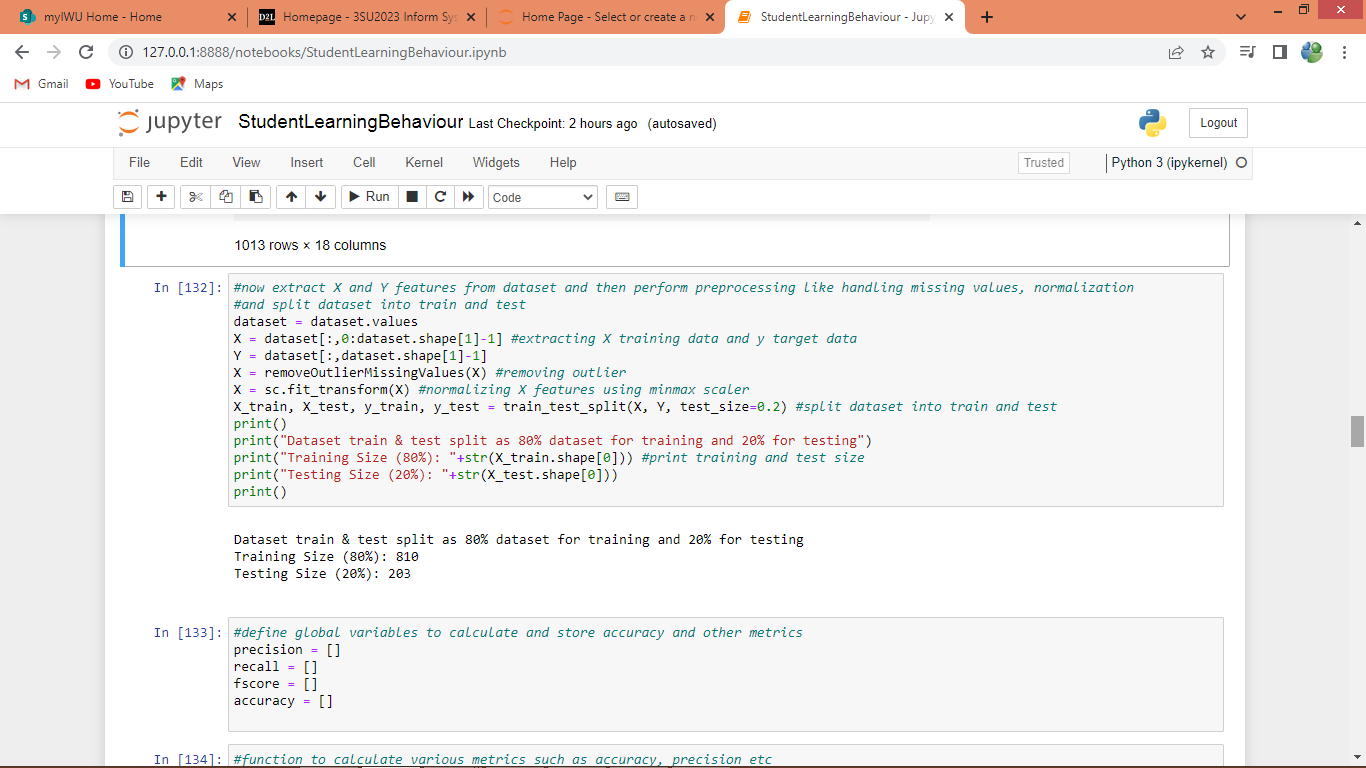
In above screen we are applying label encoder class to convert all non-numeric data into numeric values as machine learning accept only numeric features and in above screen we can see all values converted to numeric format



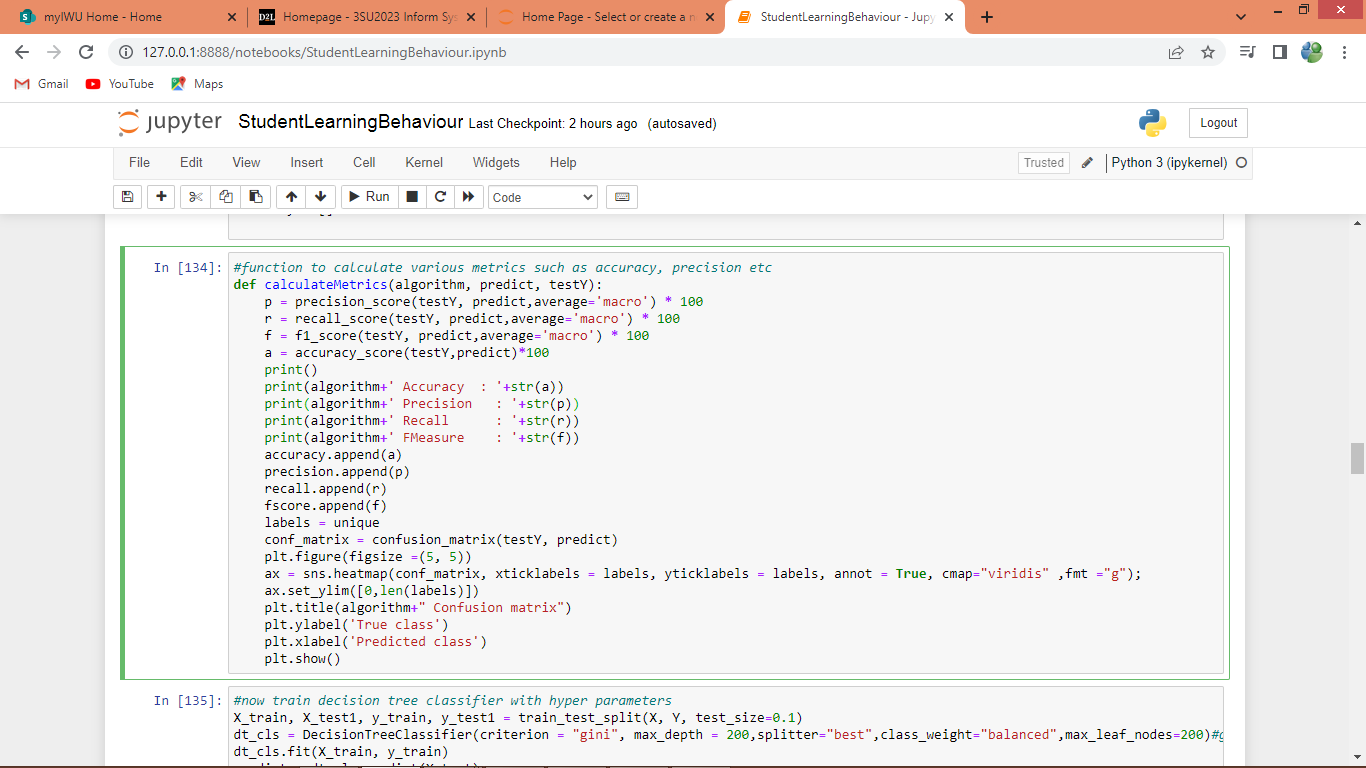
In above screen finding and plotting correlation graph where each box represents correlation value of the columns, so if features are important then its correlation value will be high.



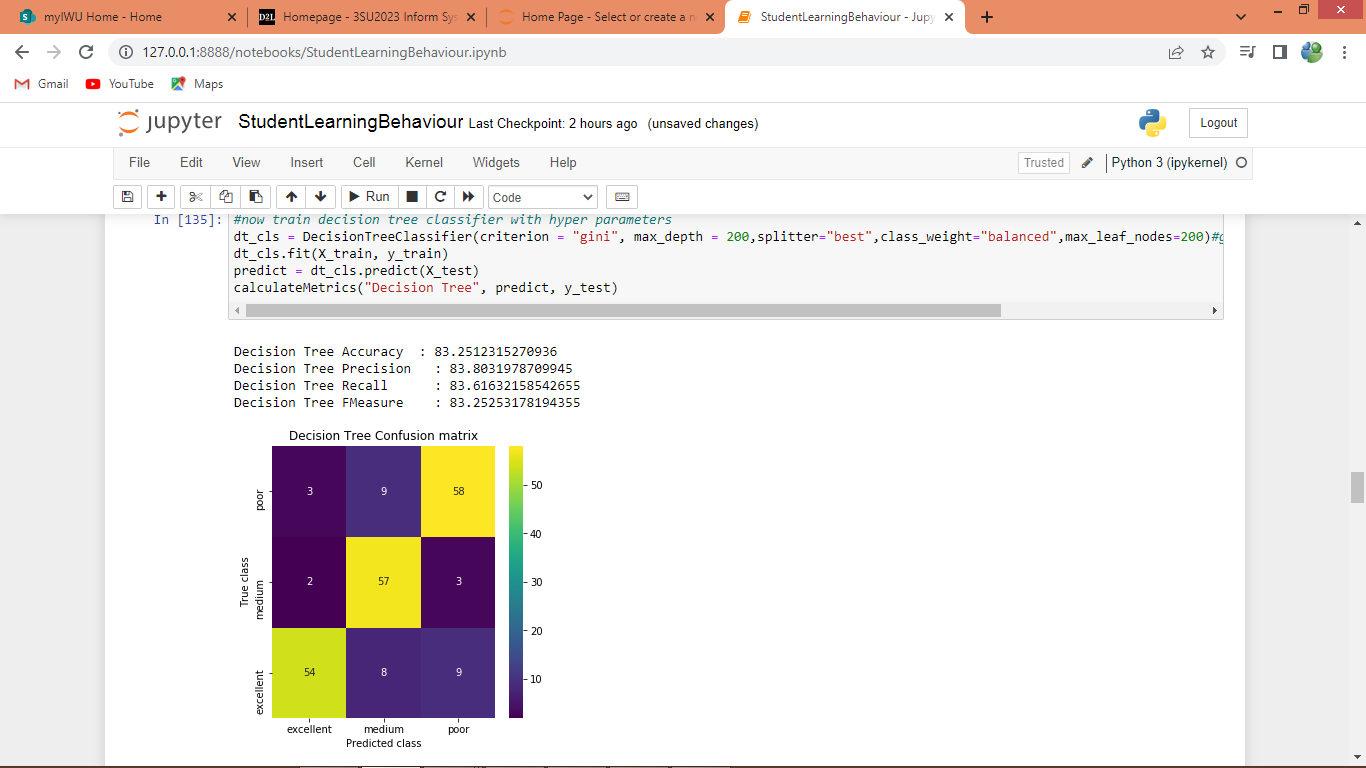
In above screen by using correlation algorithm we are removing least correlated values and the selecting only high correlated features

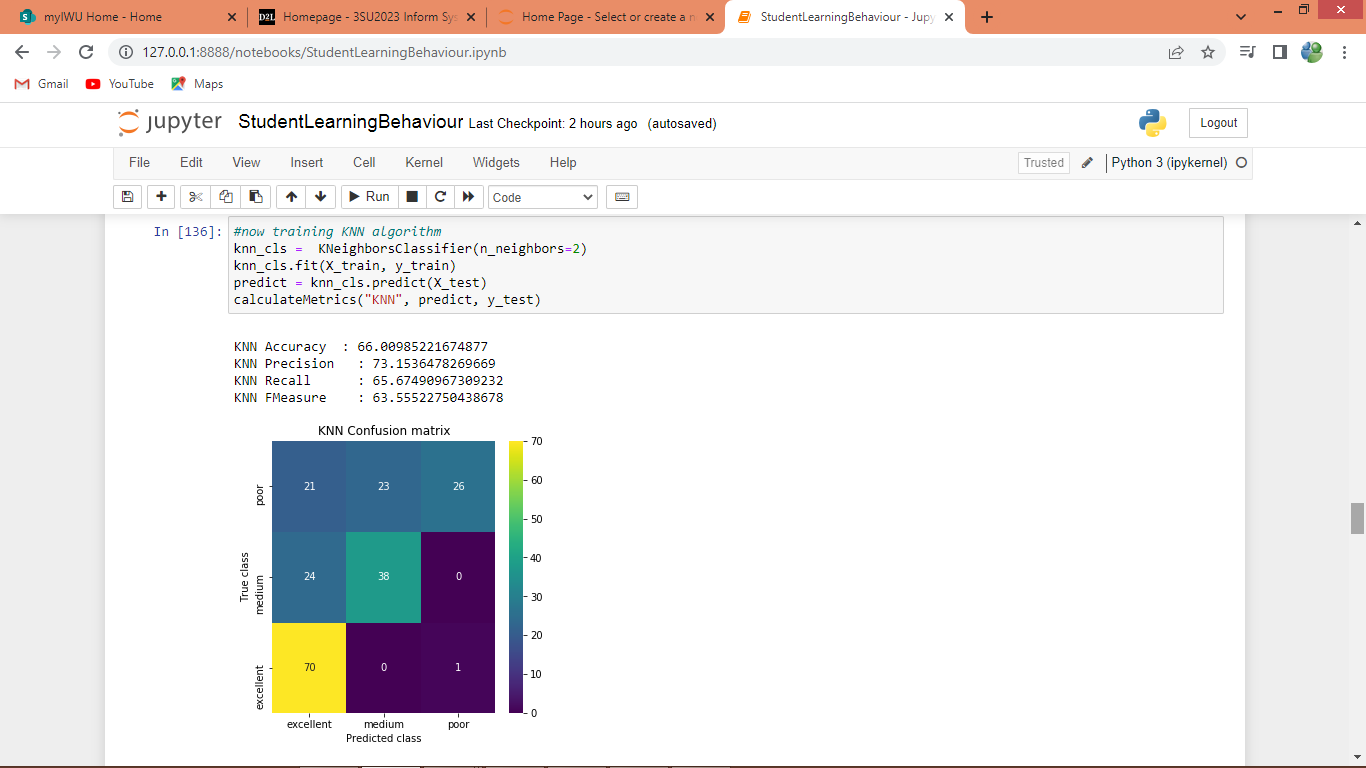


In above screen defining function to split dataset into train and test where application using 80% dataset for training and 20% for testing

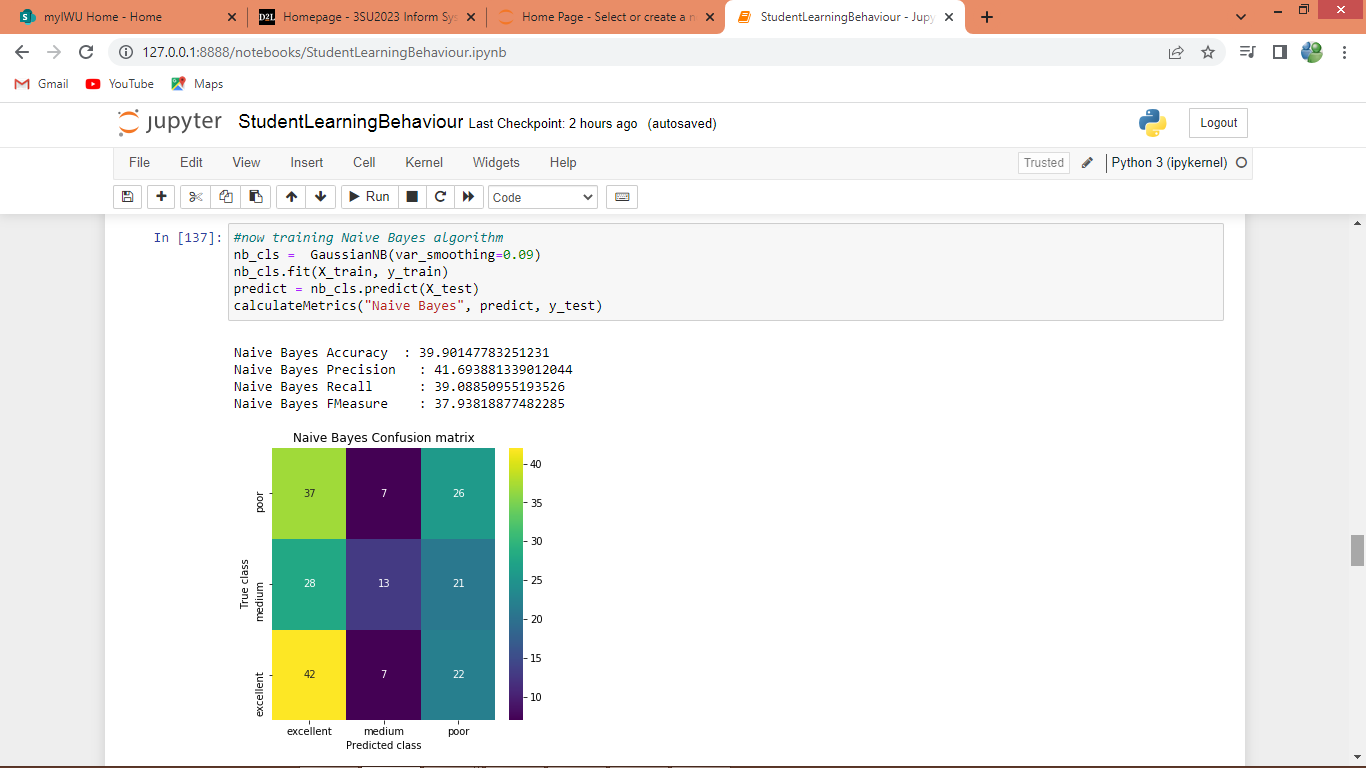


In above screen defining function to calculate accuracy and other metrics

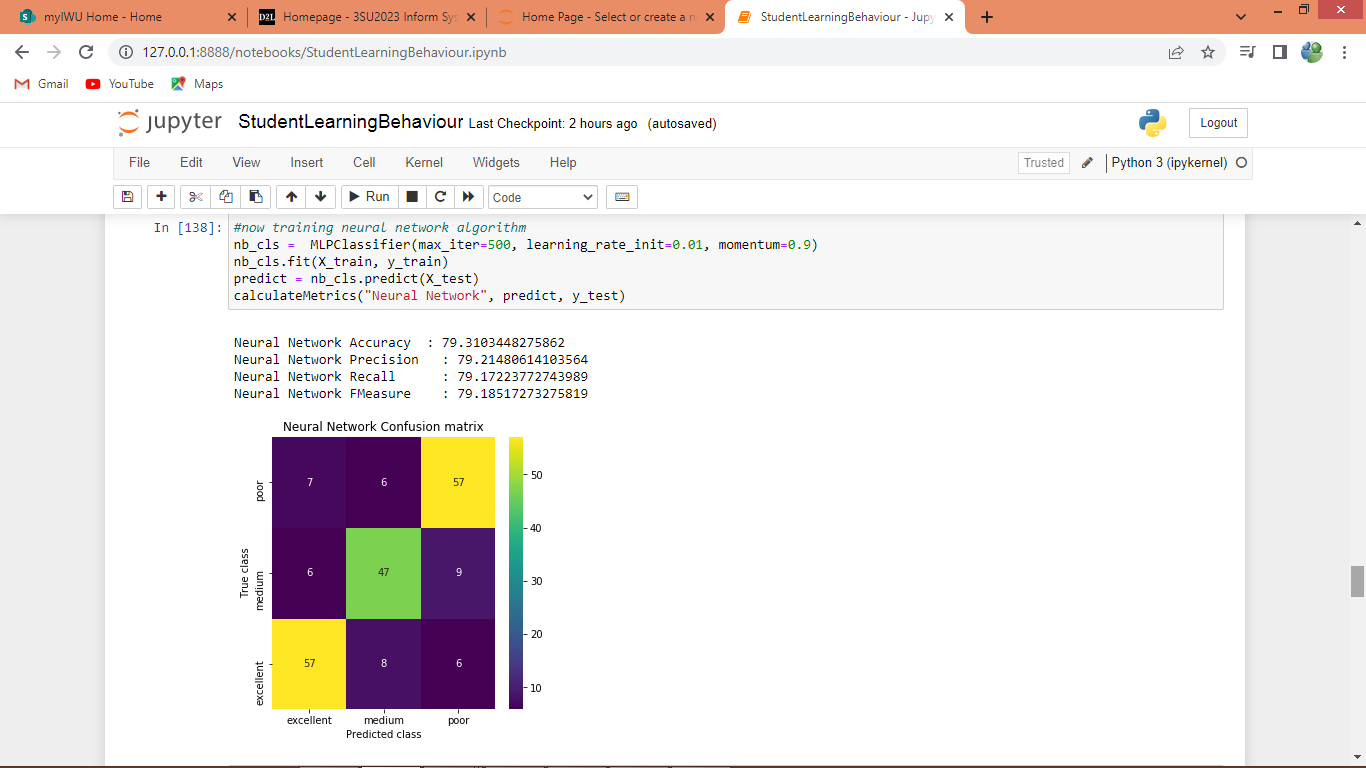
in above screen training decision tree and decision tree got 83% accuracy and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where all yellow boxes represents correct prediction count and remaining colour boxes represents incorrect prediction count which are very few.



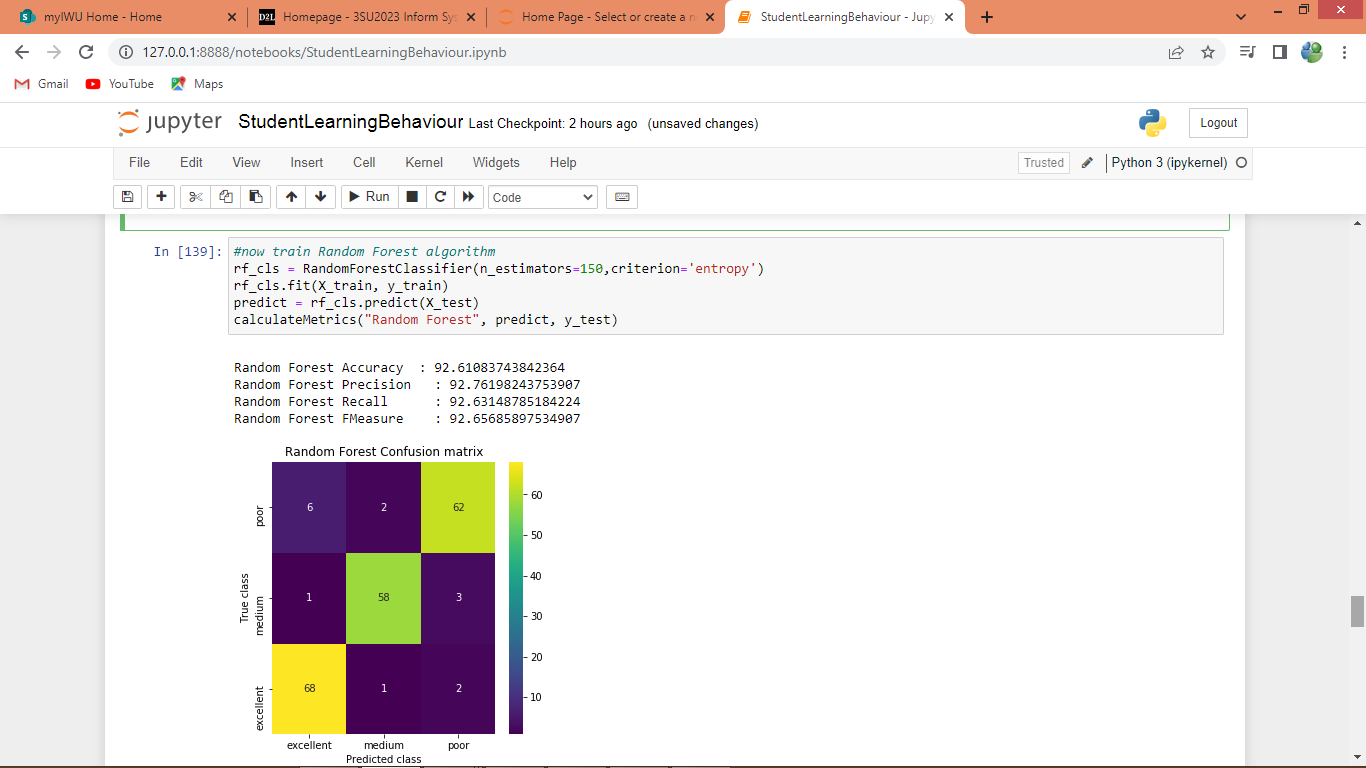
In above screen training with KNN and got 66% accuracy



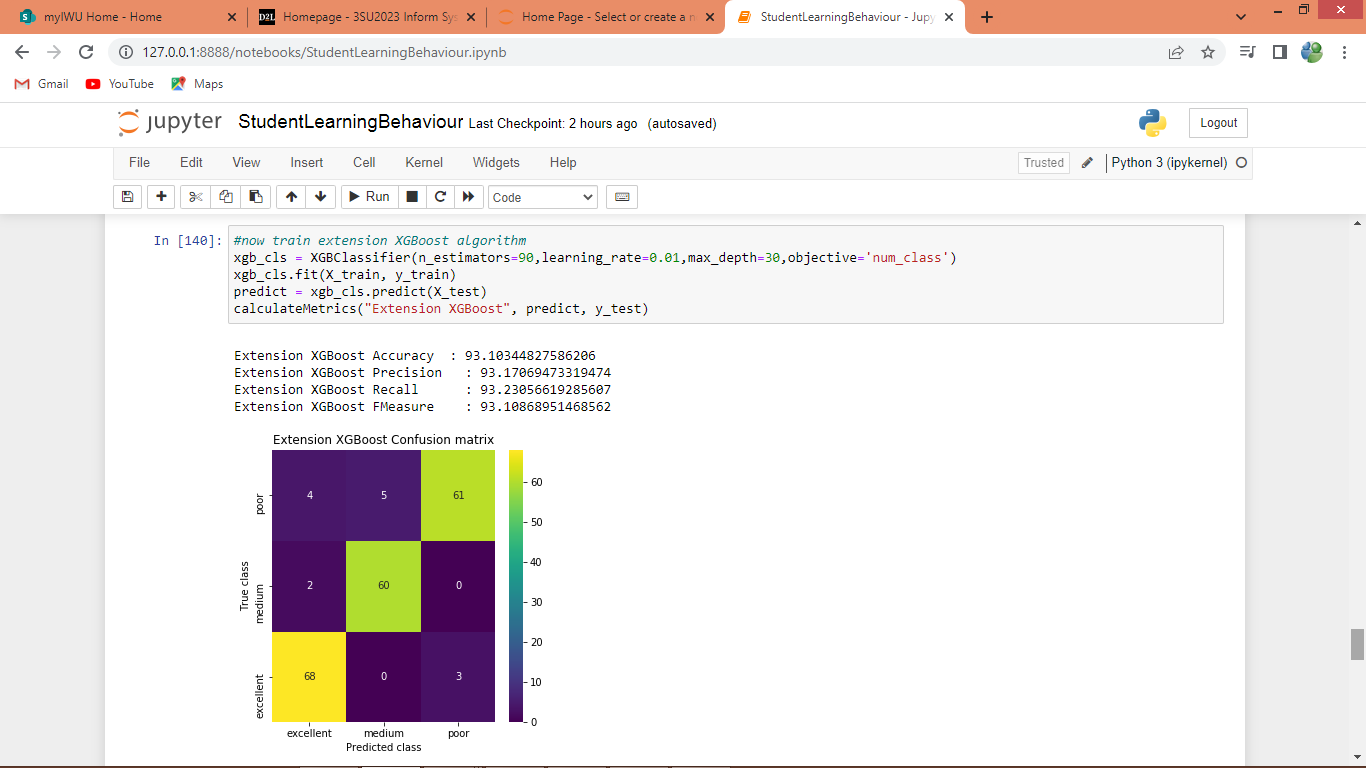
In above screen training with Naive Bayes and got 39% accuracy



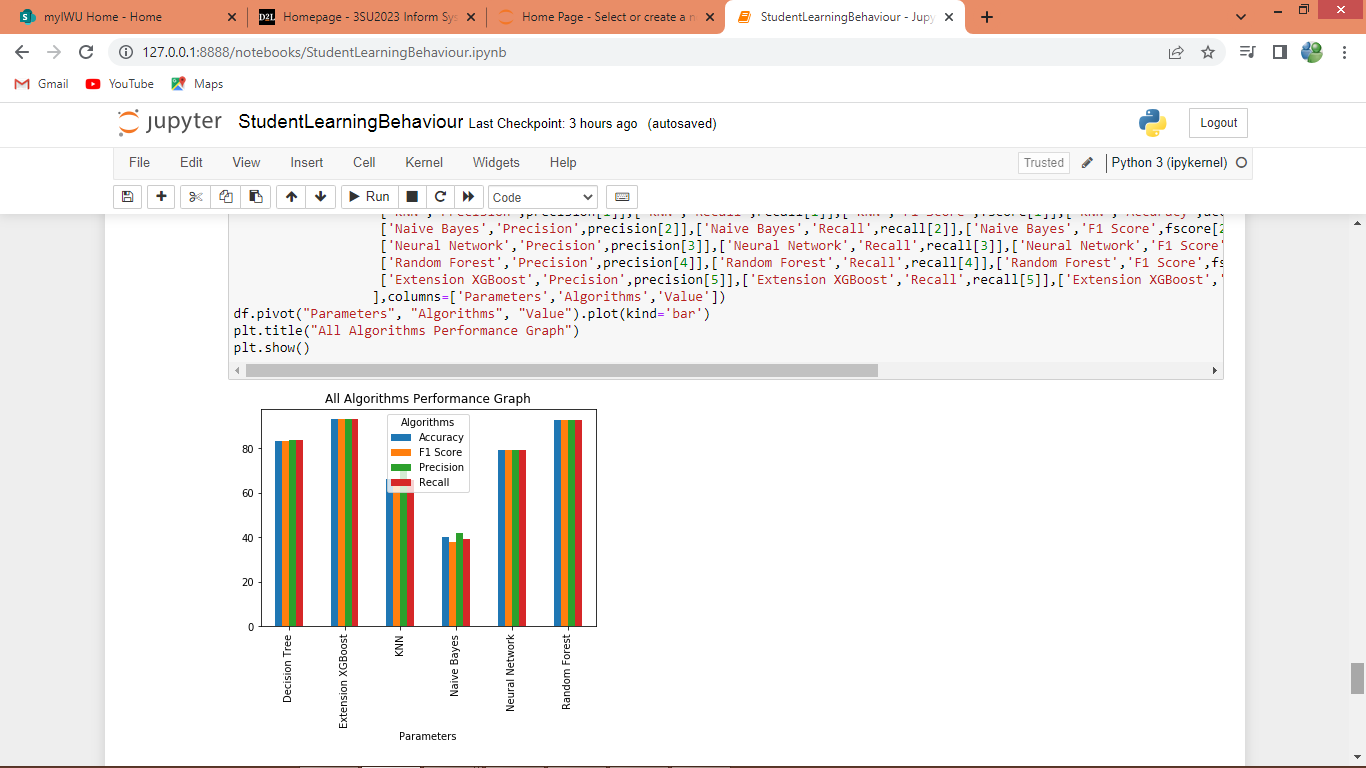
In above screen training with Neural Network and got 79% accuracy



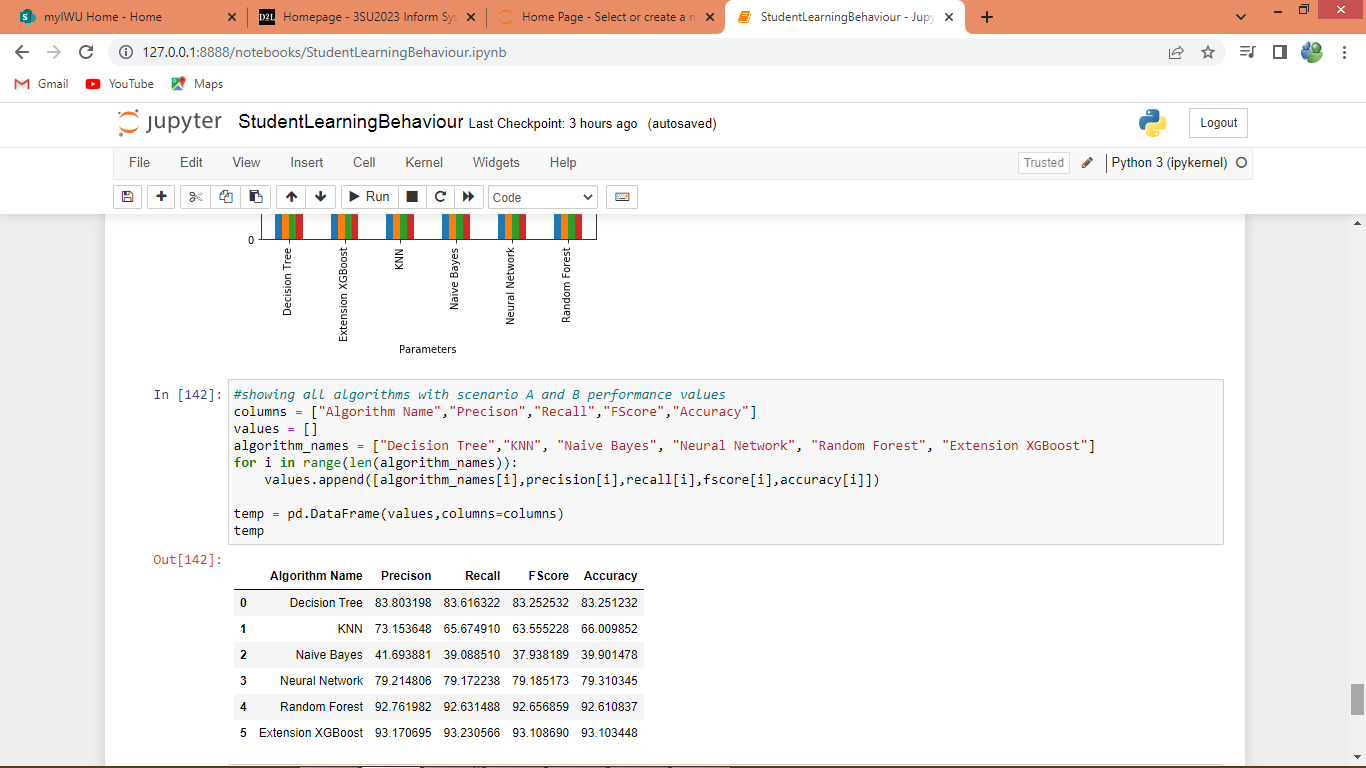
In above screen training with Random Forest and got 92% accuracy



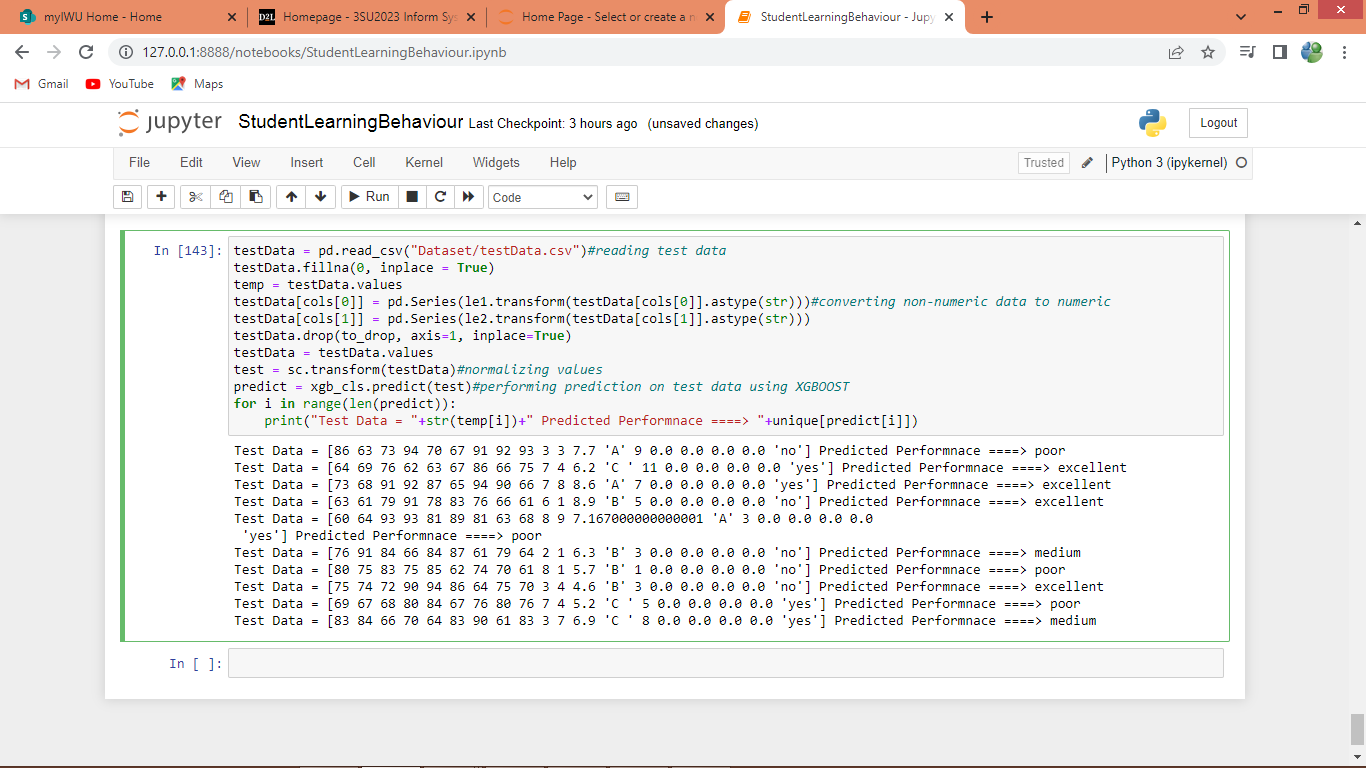
In above screen training with Extension XGBOOST algorithm and got 93% accuracy, so in all algorithms XGBOOST got high accuracy



In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics and in all algorithms XGBOOST got high performance



In above screen showing all algorithm performance in tabular format



In above screen uploading test data and then performing prediction on test data using extension XGBOOST algorithm and for each test data after arrow symbol =🡺 we can see predicted output.

Note: in propose paper author has used some dataset but not publish on internet so whatever dataset available on internet we have used that dataset. For all algorithms accuracy may vary for each run as train and test data splitted randomly