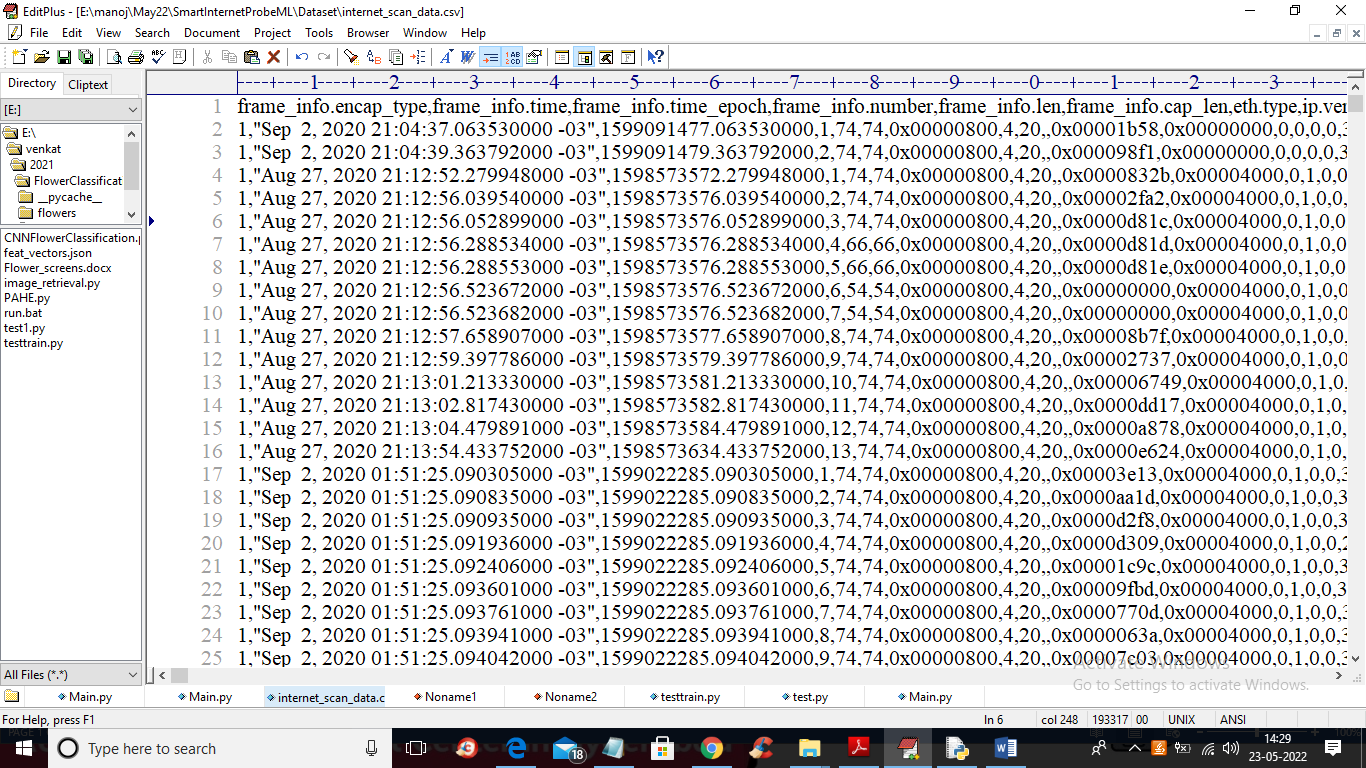
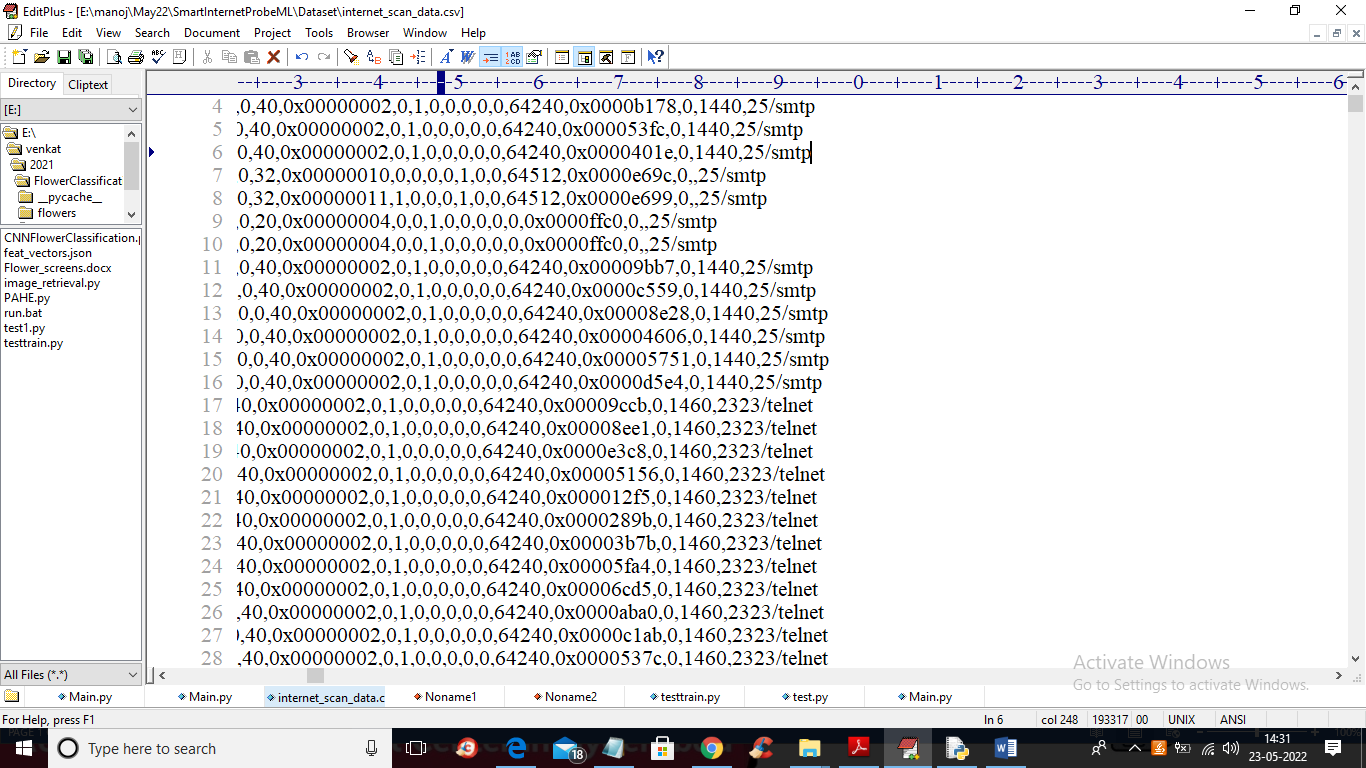
Smart Internet Probing: Scanning Using Adaptive Machine Learning

On internet all services are running on servers and this servers runs on IP address and PORT numbers and to access this services client will probe or scan PORT numbers to check servers are running or not. Sometime some servers will not run and probing such servers will increase network bandwidth and consume resources and to avoid such wastage of bandwidth and resources author of this paper employing machine learning algorithms to predict servers are running or not. Author saying PORTS and other details of all servers will share similar features and if one server is not running then machine learning can predict servers with similar features may also not run. Prediction of working port prior to scan/probe may save resources and bandwidth.

In propose paper author is using CENSYS dataset which scan 23 different PORTS and we are using same dataset. Below screen showing dataset details



In above dataset screen first row contains dataset column names and remaining rows contains dataset values and in last columns we can see labels as PORT and in below screen you can see those ports



In above screen in last column we can see smtp and telnet with port numbers are the labels used to predict weather port with similar features are working or not.

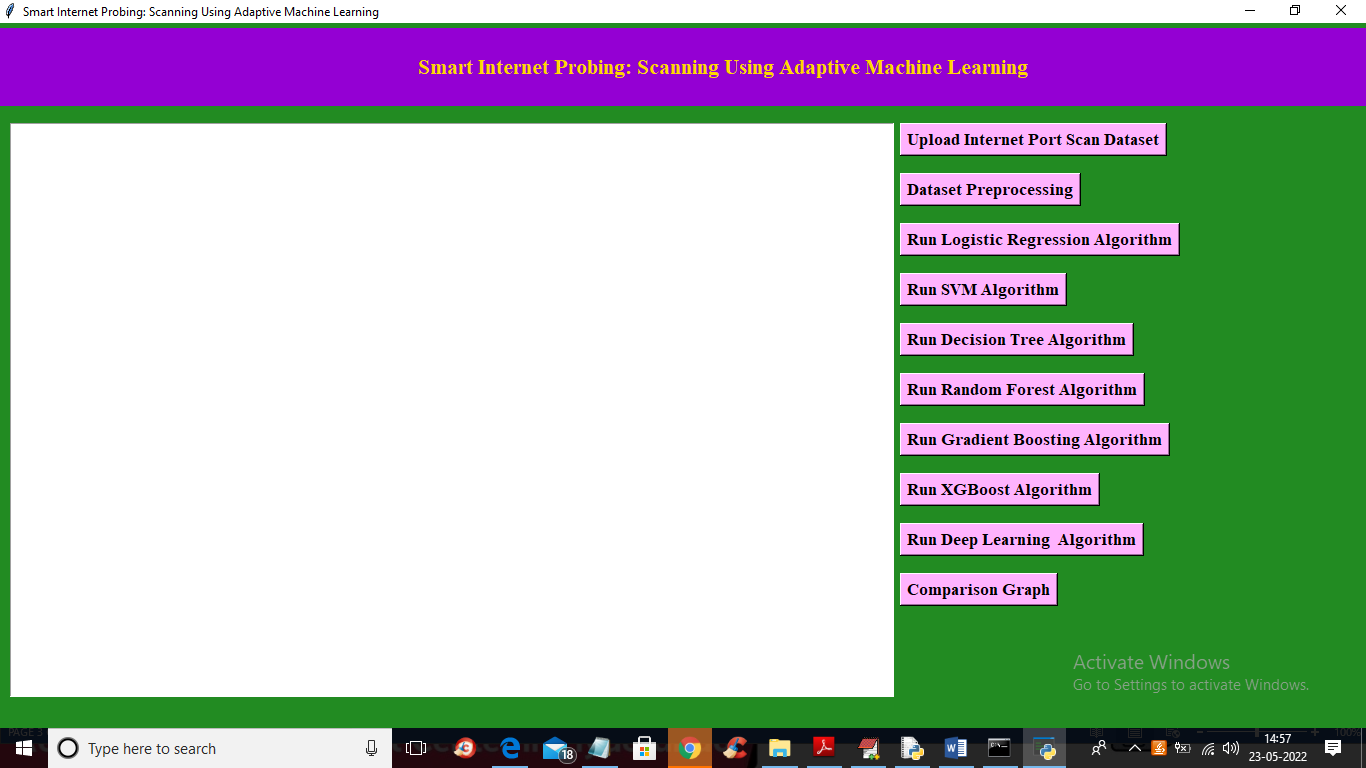
In propose paper for port classification author is using various machine learning algorithms called SVM, Logistic Regression, Decision Tree, Random Forest, Deep Learning Neural Network, Gradient Boosting and XGBOOST. Author evaluating each algorithm performance in terms of confusion matric and TPR (true positive rate). Algorithm with higher TPR will be consider as best. This algorithms can scan port PARALLELY and SEQUENTIALLY.

To implement this project we have designed following modules

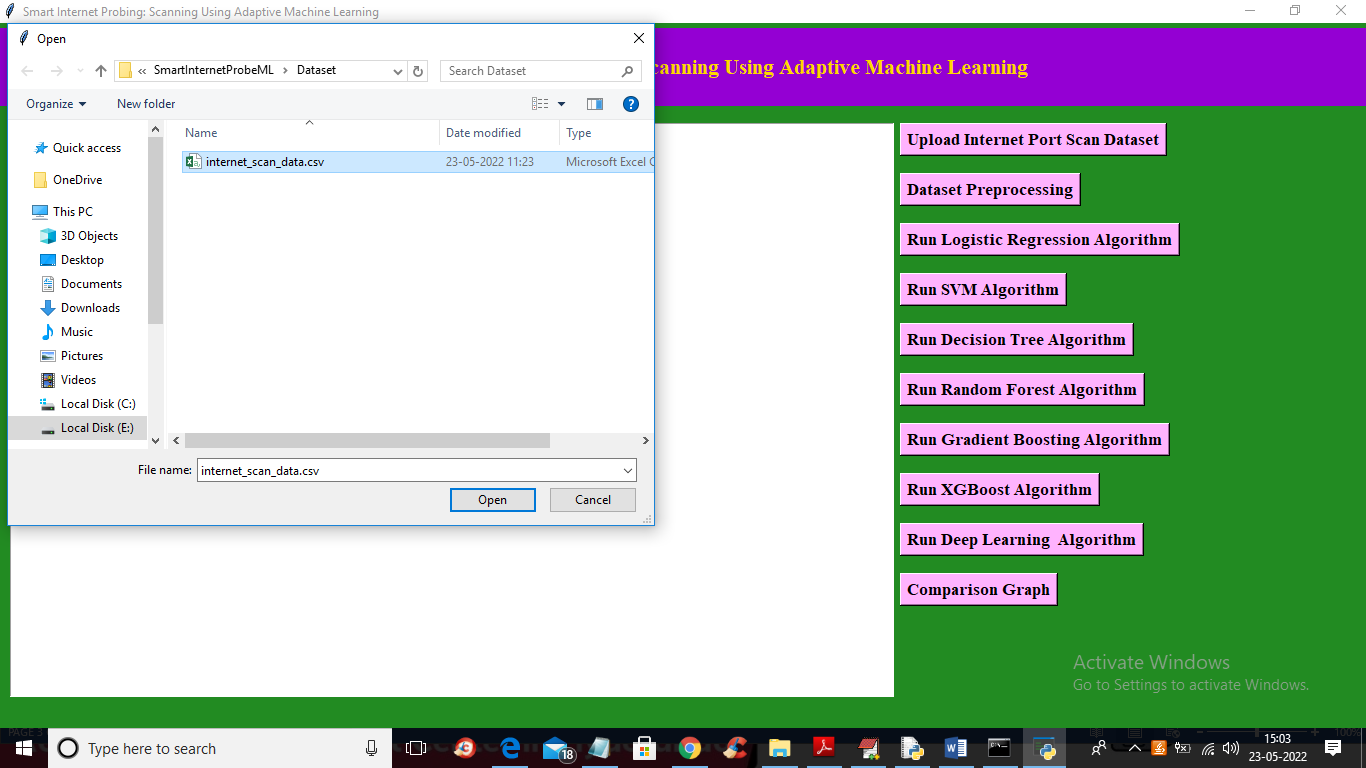
1. Upload Internet Port Scan Dataset: using this module we will upload dataset to application and then display types of PORTS found in dataset
2. Dataset Preprocessing: dataset contains non-numeric and numeric data and machine learning algorithms only accept numeric data so by using module we will encode non-numeric data into numeric and then split dataset into train and test where application used 80% dataset for training and 20% for testing
3. Run Logistic Regression Algorithm: using this module we will train logistic regression by using 80% dataset and then apply regression model on 20% dataset to calculate TPR value. The higher the TPR the better is the algorithm
4. Similarly we will train decision tree, random forest, gradient boosting, XGBOOST, deep learning neural network and SVM and then calculate its TPR.
5. Comparison Graph: using this module we will plot TPR graph of all algorithms

SCREEN SHOTS

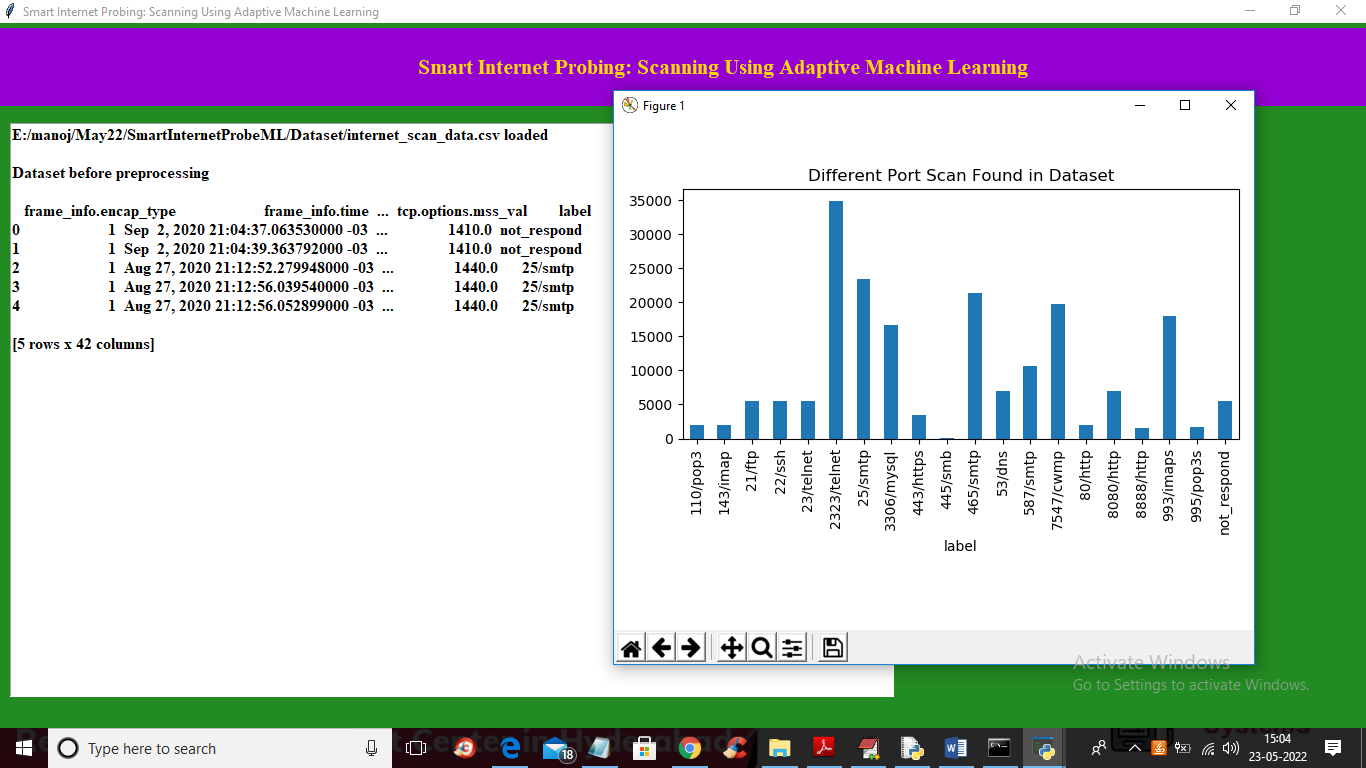
To run project double click on ‘run.bat’ file to get below output



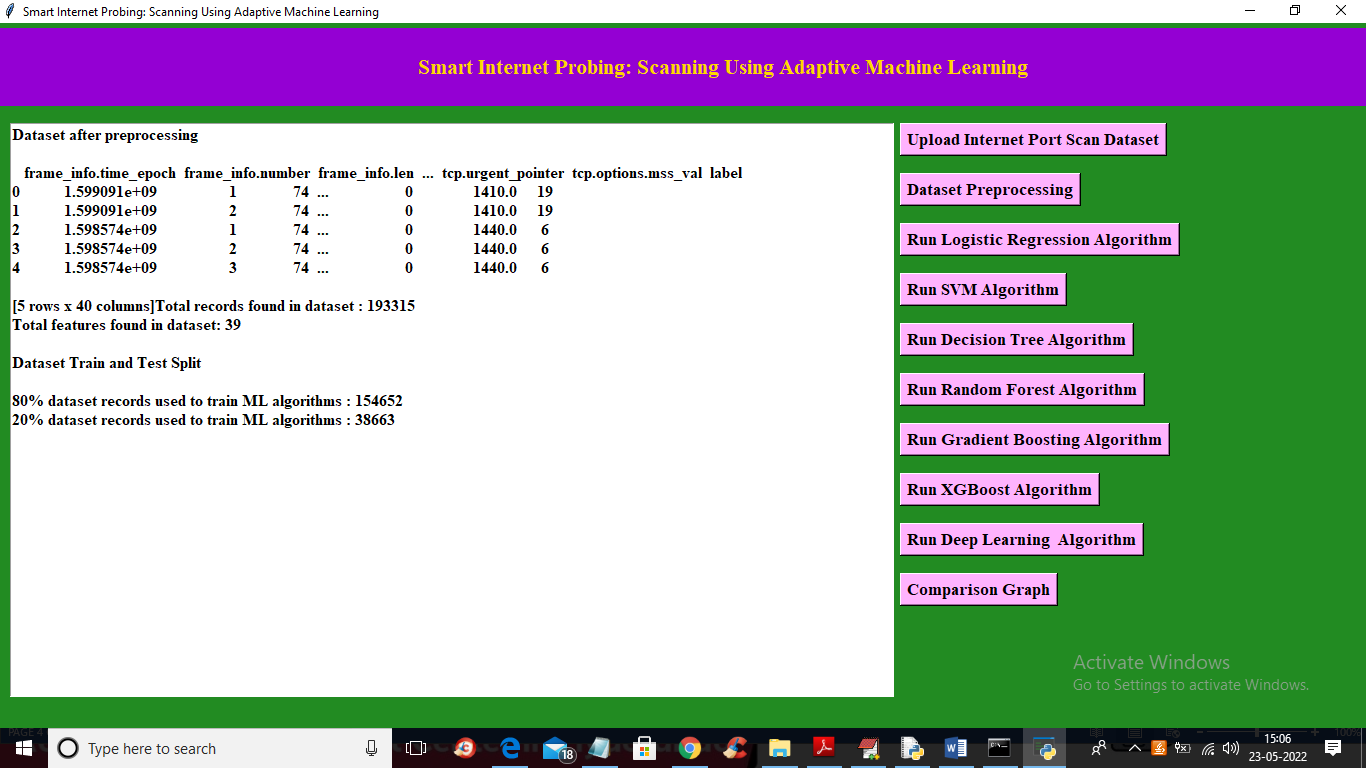
In above screen click on ‘Upload Internet Port Scan Dataset’ button to upload dataset and get below screen



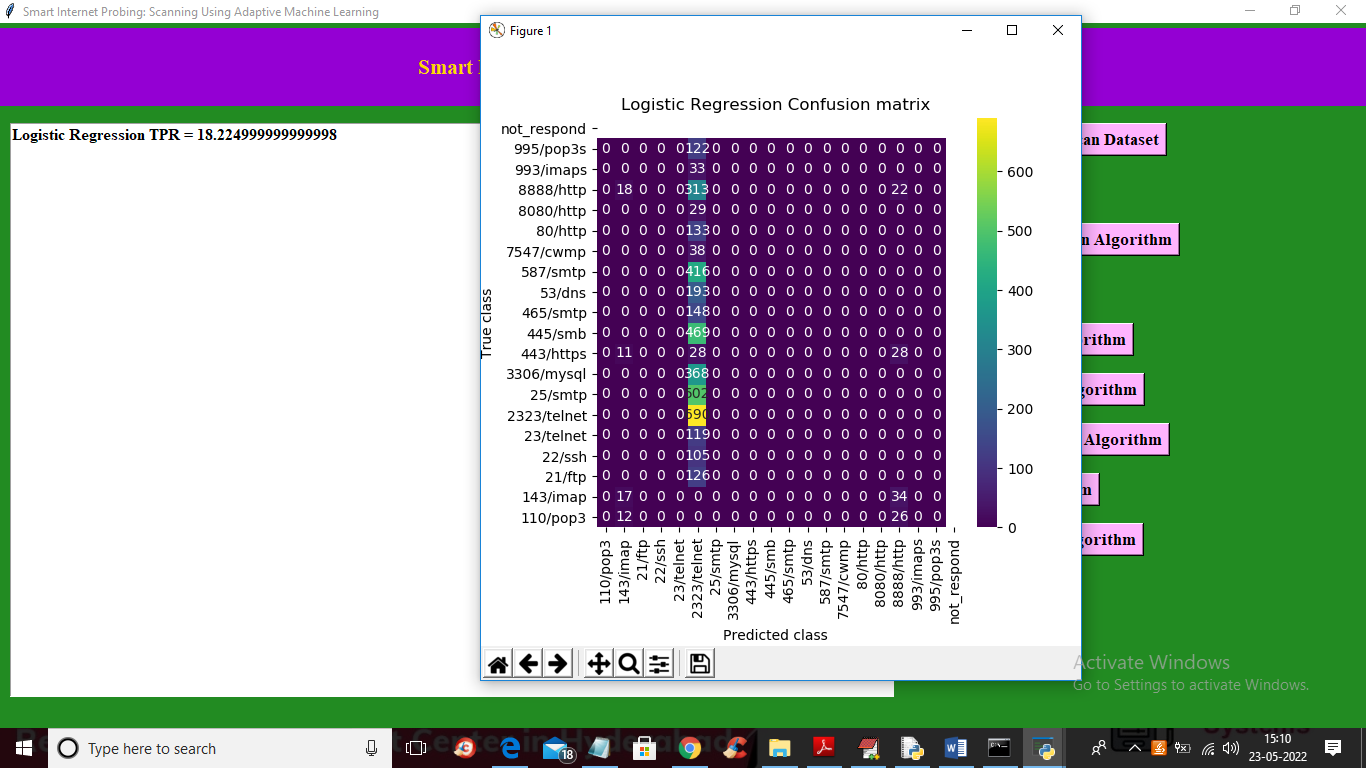
In above screen selecting and uploading dataset and then click on ‘Open’ button to load dataset and get below output



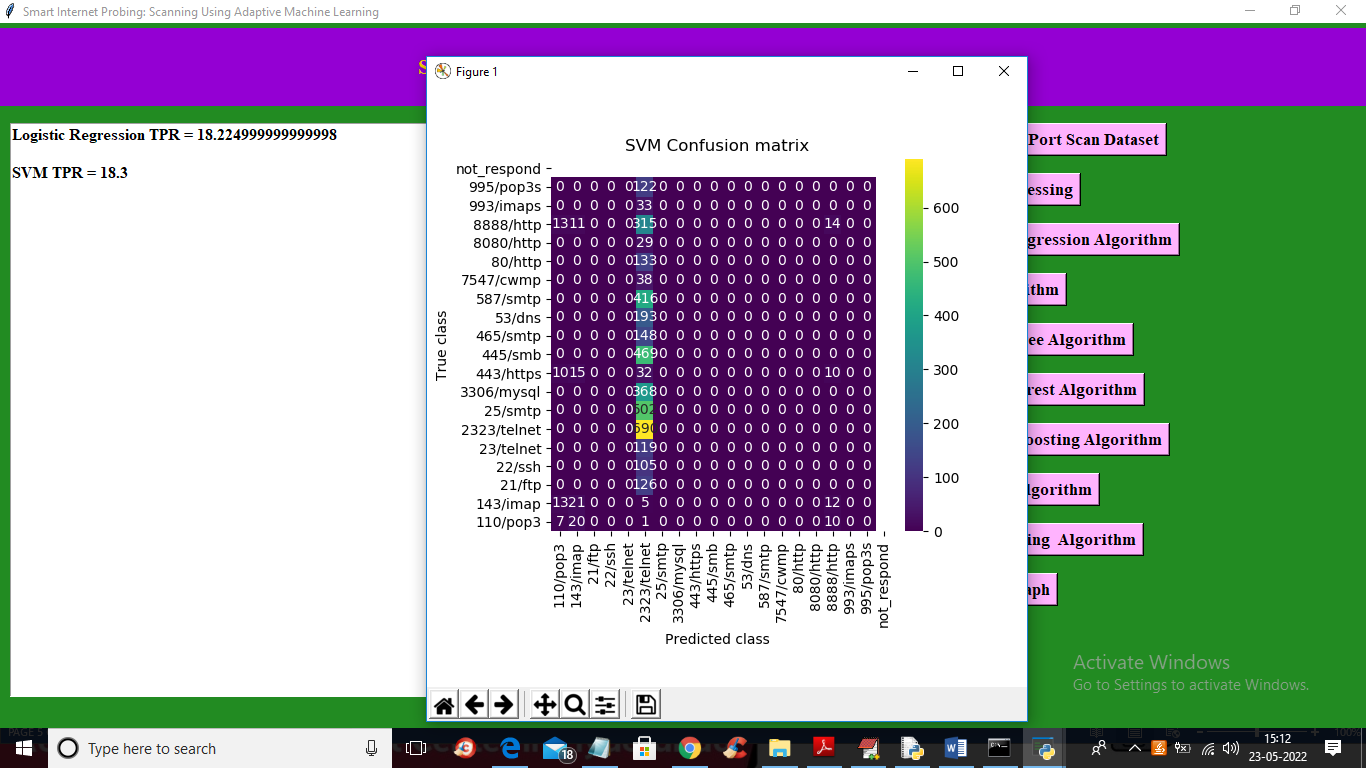
In above screen we can see dataset loaded and in dataset we can see it contains numeric and non-numeric data so we need to encode it to numeric by applying data Preprocessing. In above graph X-AXIS contains different ports and y-axis contains count of each port in dataset and now close above graph and then click on ‘Dataset Preprocessing’ button to process dataset and get below output



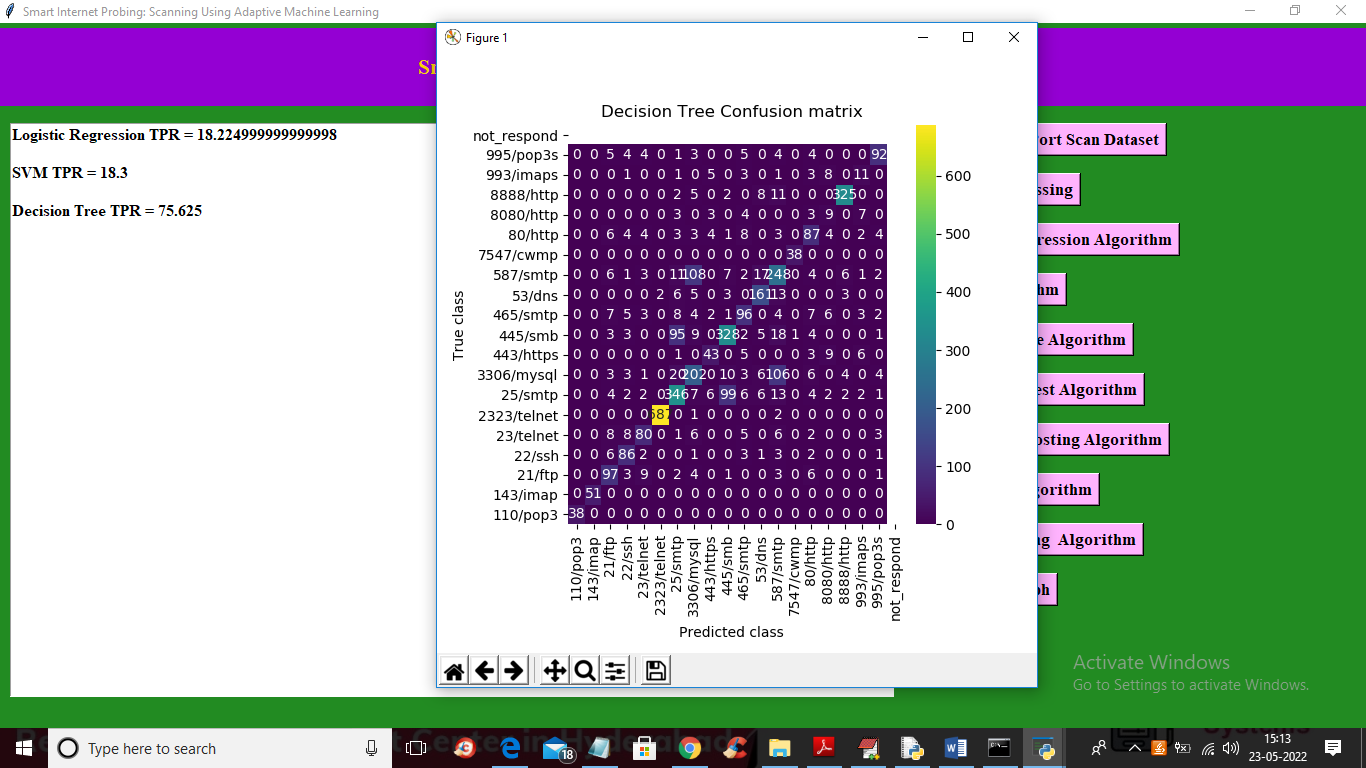
In above screen we can see all dataset values converted to numeric and we can see total records and column found in dataset and then we can see 80% training records size from dataset and 20% for testing. Now train and test data is ready and now click on ‘Run Logistic Regression Algorithm’ button to train algorithm and get below output



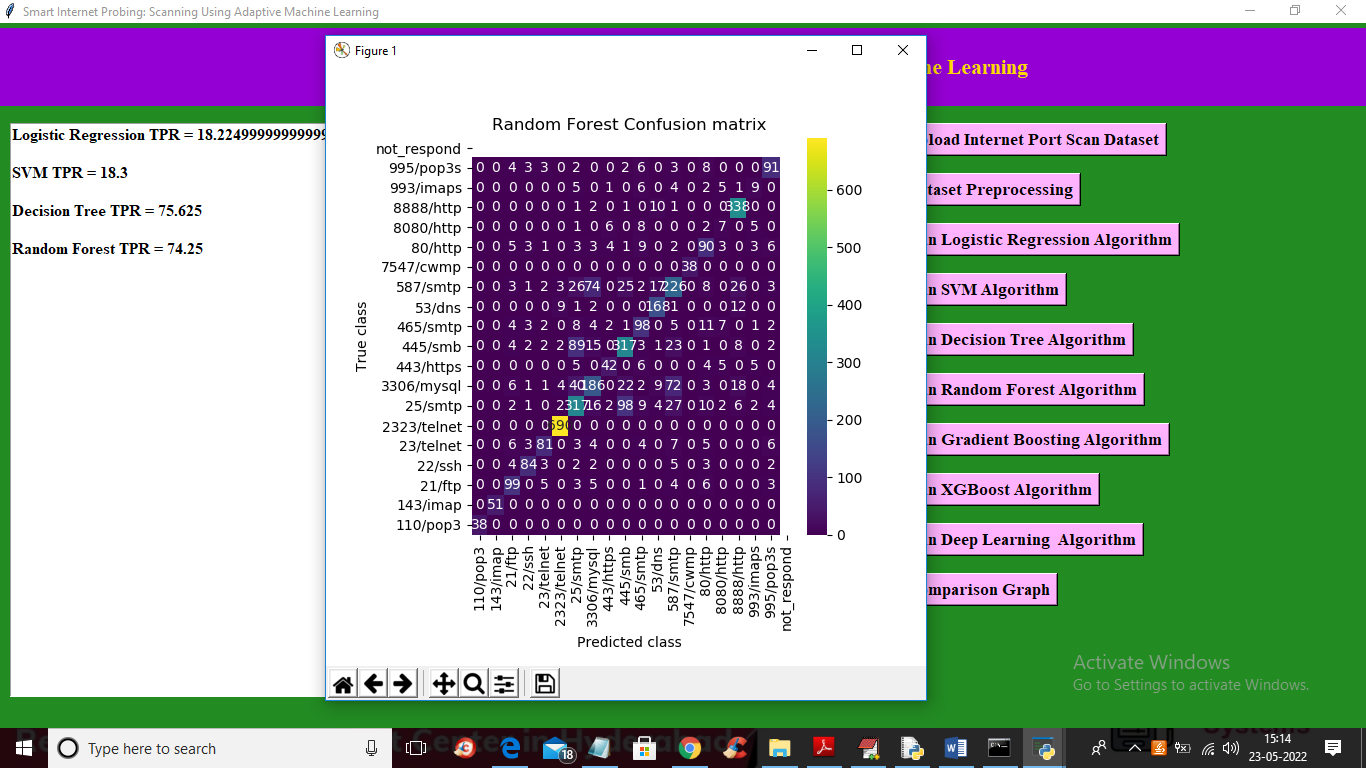
In above screen with logistic regression we got TPR as 18% which is not good and in confusion matrix x-axis represents predicted labels and y-axis represents TRUE labels and we can see all labels are predicted only in one class and now close above graph and then click on ‘Run SVM Algorithm’ to train SVM



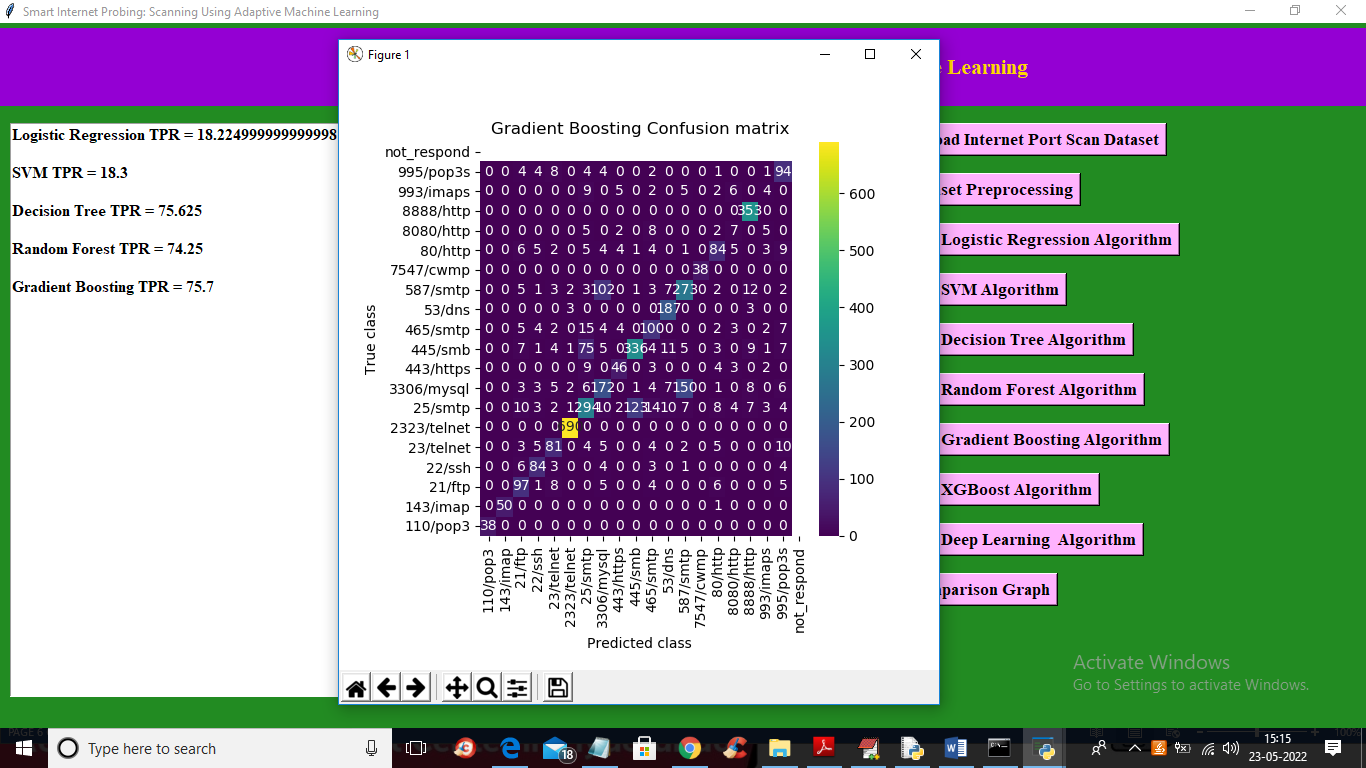
In above screen with SVM also we got only 18% TPR and now run decision tree algorithm and get below output



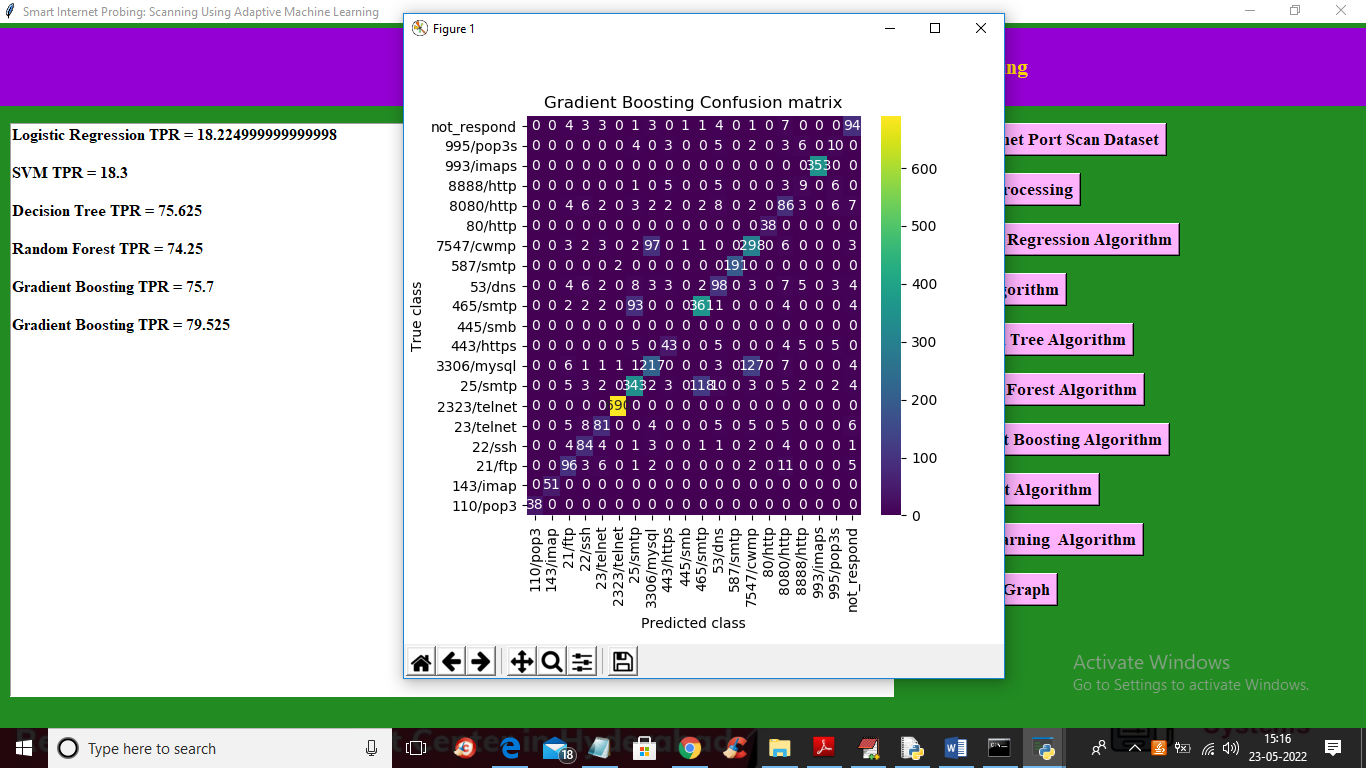
In above screen with decision tree we got 75% TPR and in confusion matrix in diagnol you can see more number of predicted and true classes are correct so its TPR is 75%. Now run Random Forest algorithm



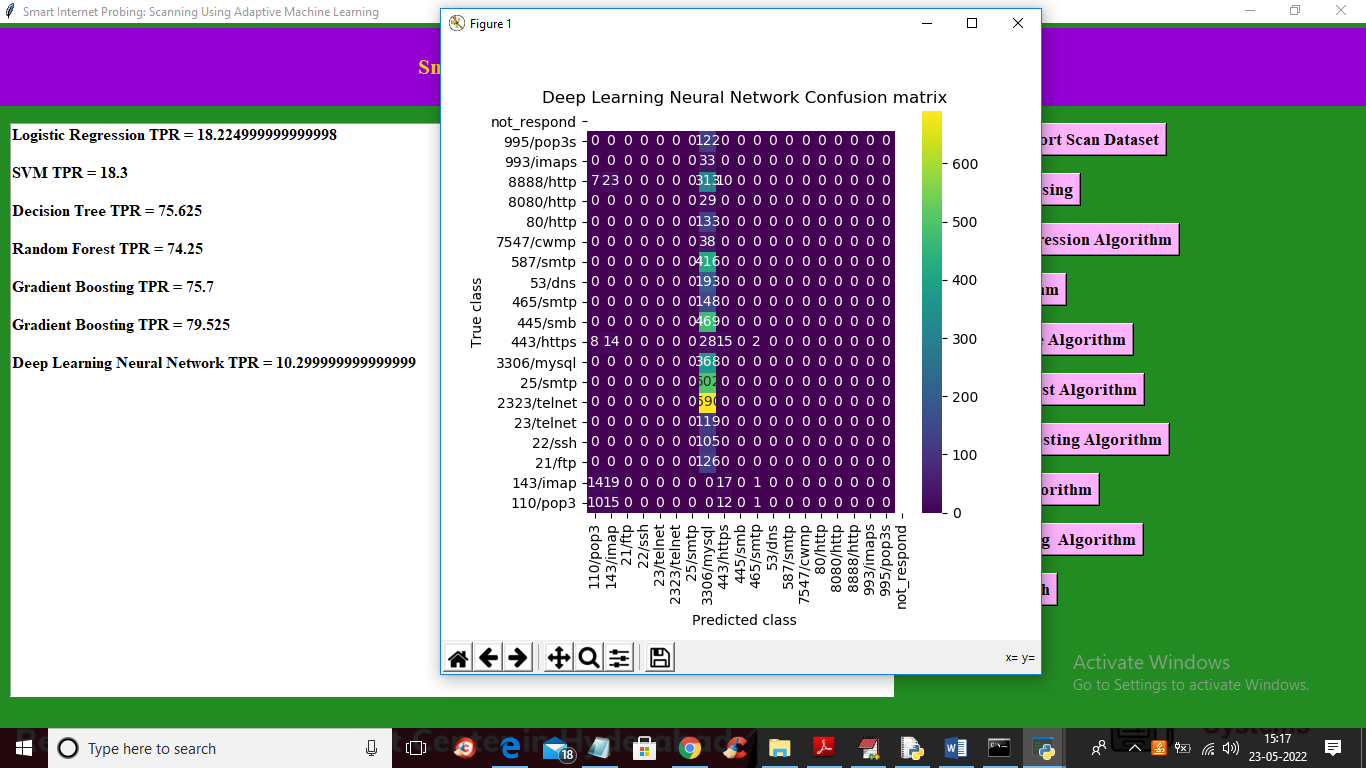
In above screen with random forest we got 74% TPR and now run gradient boosting algorithm



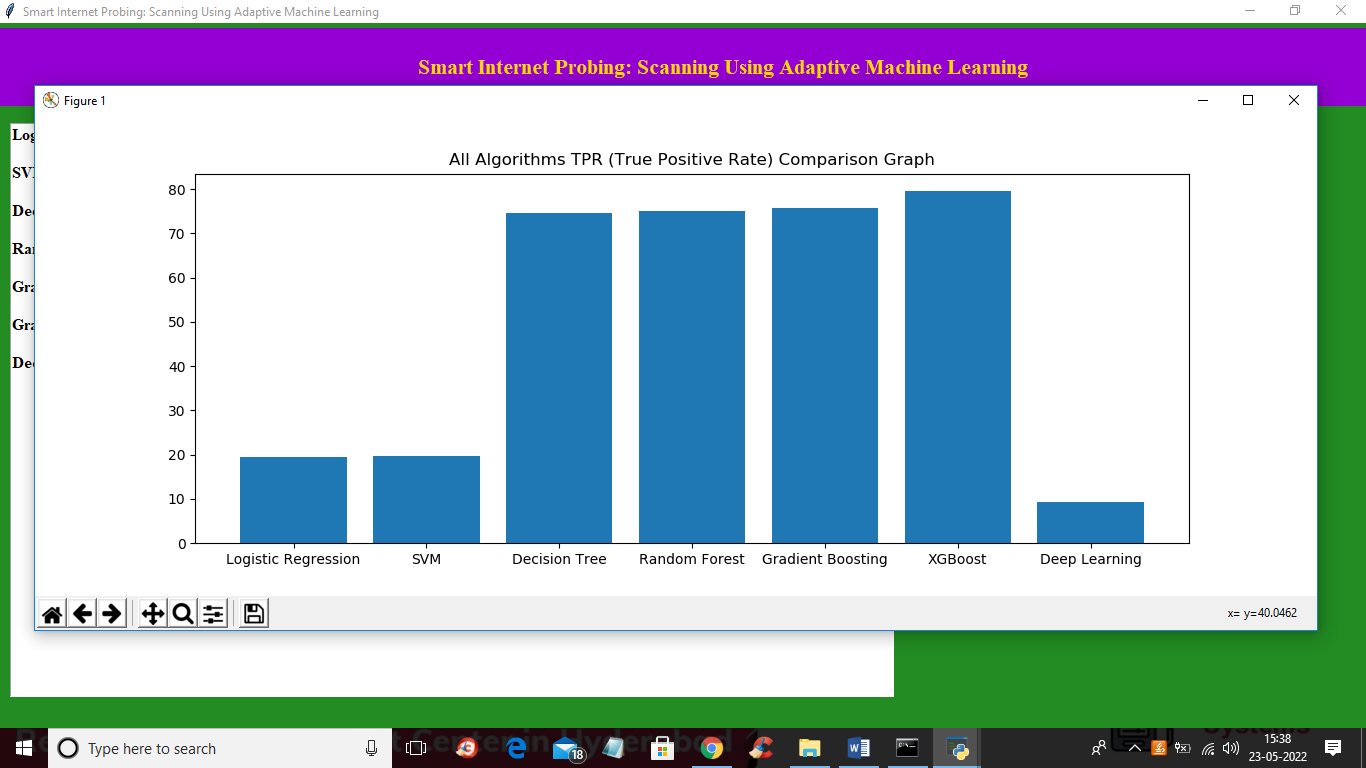
In above screen with gradient boosting we got 75.7% TPR and now run ‘XGBoost algorithm’



In above screen with XGBOOST we got 79% TPR and run Deep learning MLP algorithm



In above screen with deep learning algorithm we got 10% TPR and now click on ‘Comparison Graph’ button to get below graph



In above graph x-axis contains algorithm names and y-axis contains TPR and in all algorithms XGBOOST has got highest TPR. So by employing ML algorithms we can predict working port without probing/scanning