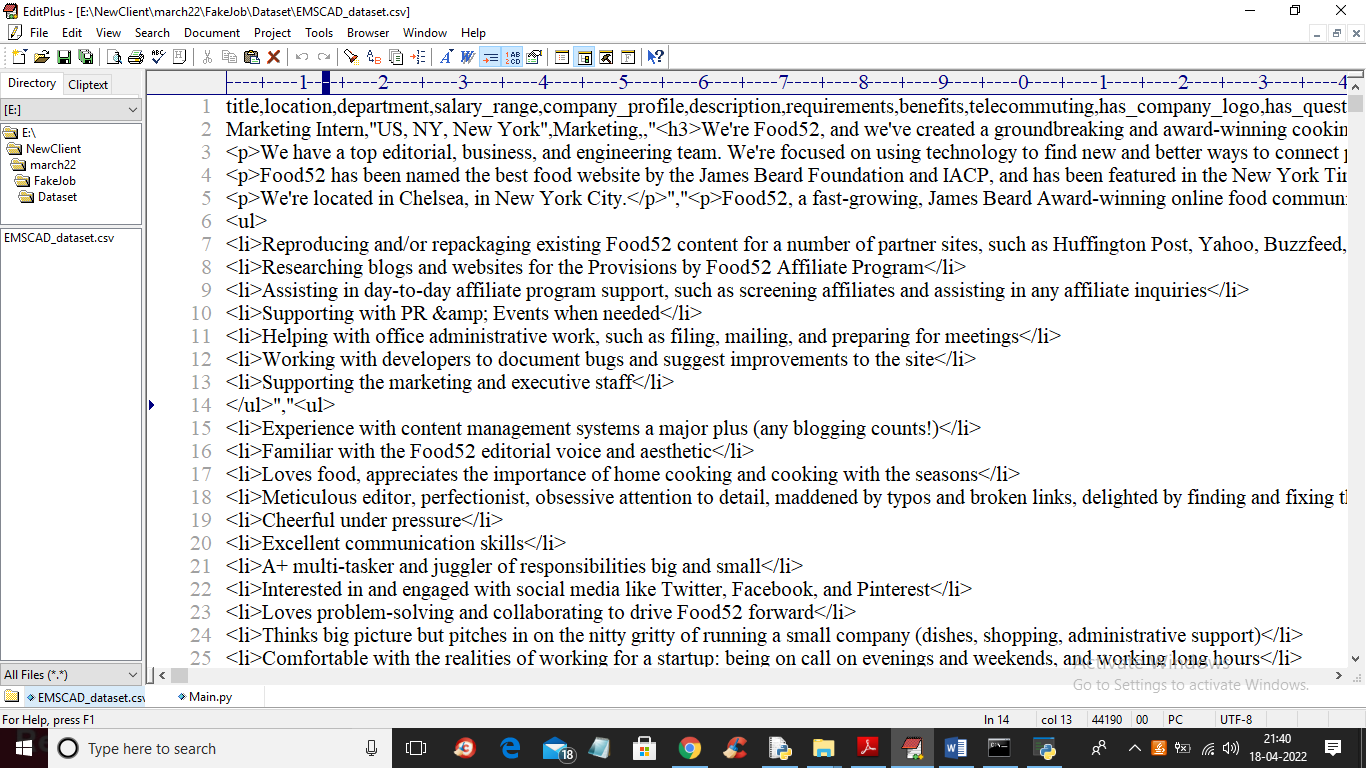
A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Earlier for jobs we need to wait for newspaper advertisement but now-a-days advancement in technologies has reduced that wait in seconds as companies can publish news jobs on internet and will be available on Google search but this advancement has led some malicious users to publish fake jobs and such fake jobs can increase frustration level in job hunters.

To overcome from this problem author of this paper are using many machine learning and deep learning algorithms which get trained on TRUE and FAKE jobs and if any fake job publish then this machine learning algorithm will analyse that job and predict it as fake. In machine learning algorithms author has used SVM, KNN, Naïve Bayes, Decision Tree, Random Forest and Multilayer Perceptron. In deep learning author has used Deep Neural Network algorithm called ANN and in both algorithms ANN is giving better accuracy.

To trained above algorithms author has used ‘Employment Scam Aegean Dataset (EMSCAD)’ dataset and this dataset can be downloaded from below link

Below screen showing dataset details



In above dataset screen first row represents dataset column names and remaining rows contains dataset values such as Company profile, job description, salary and etc. In dataset last column contains ‘fraudulent’ values as ‘f’ for Fake and ‘t’ for “True” jobs.

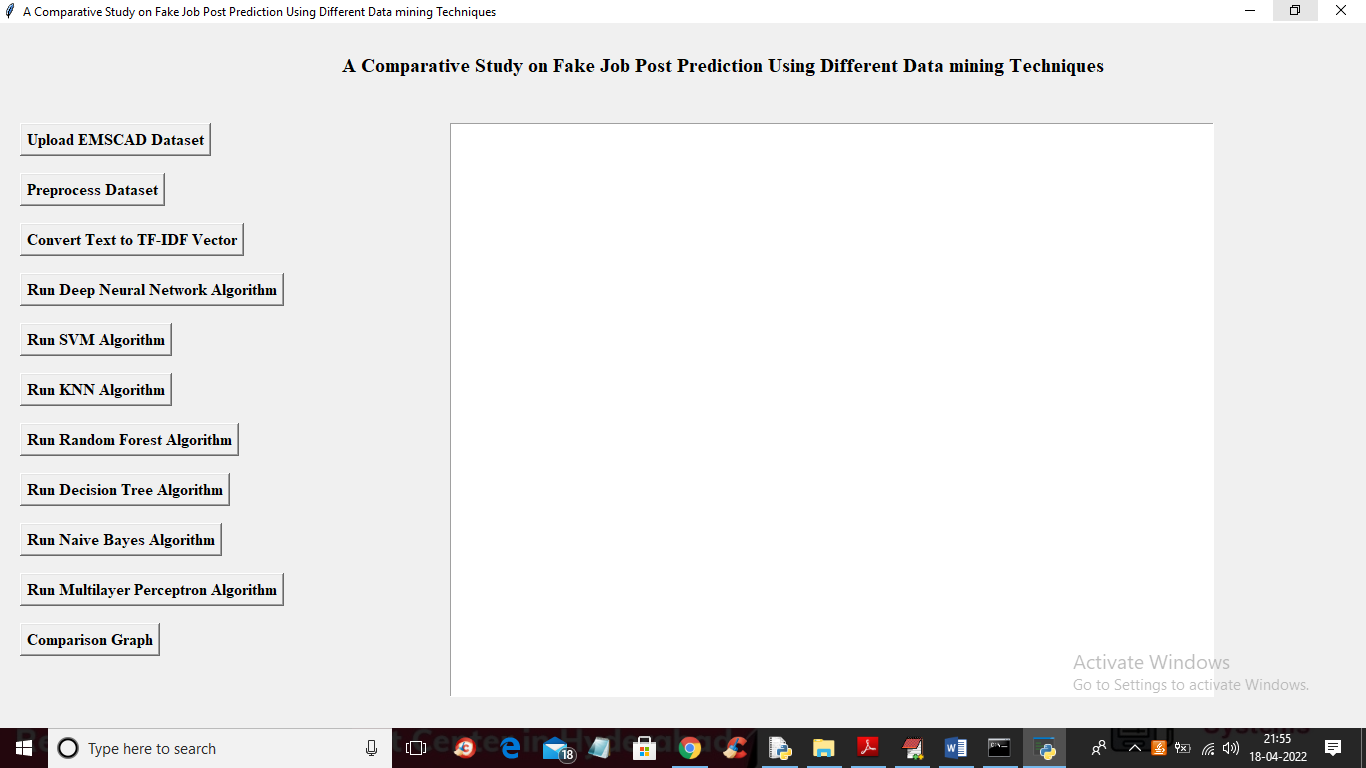
We will used above dataset to trained all algorithms and evaluate their performance in terms of accuracy, precision, recall, FSCORE and confusion matrix graph

To implement this project we have designed following modules

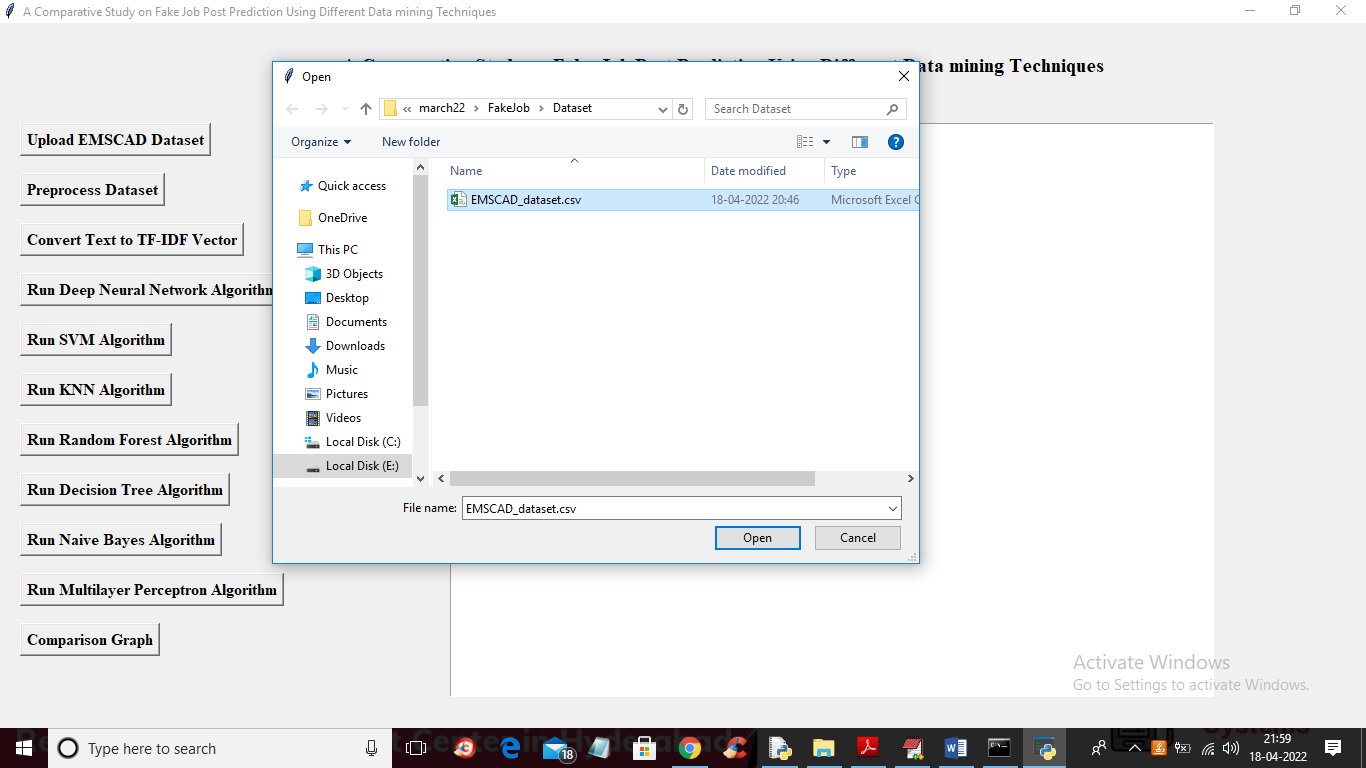
1. Upload EMSCAD Dataset: using this module we will upload ‘EMSCAD’ fake jobs Dataset
2. Preprocess Dataset: dataset often contains missing values and STOP WORDS so by using processing technique we will remove all missing values and stop words and make dataset values clean.
3. Convert Text to TF-IDF Vector: this dataset contains all non-numeric data and ML algorithms will not take such data so by applying TF-IDF (term frequency-Inverse document frequency) algorithm we will convert all words into numeric values by taking its average frequency of appearance in dataset
4. Run Deep Neural Network Algorithm: using this module we will input above vector to Neural network to trained a model and this model will be applied on test data to calculate prediction accuracy.
5. Run SVM Algorithm: using this module we will trained SVM algorithm
6. Run KNN Algorithm: using this module we will trained KNN algorithm
7. Run Random Forest Algorithm: using this module we will trained Random Forest algorithm
8. Run Decision Tree Algorithm: using this module we will trained Decision Tree algorithm
9. Run Naïve Bayes Algorithm: using this module we will trained Naïve bayes algorithm
10. Run Multilayer Perceptron Algorithm: using this module we will trained MLP algorithm
11. Comparison Graph: using this module we will plot accuracy comparison between all algorithms

SCREEN SHOTS

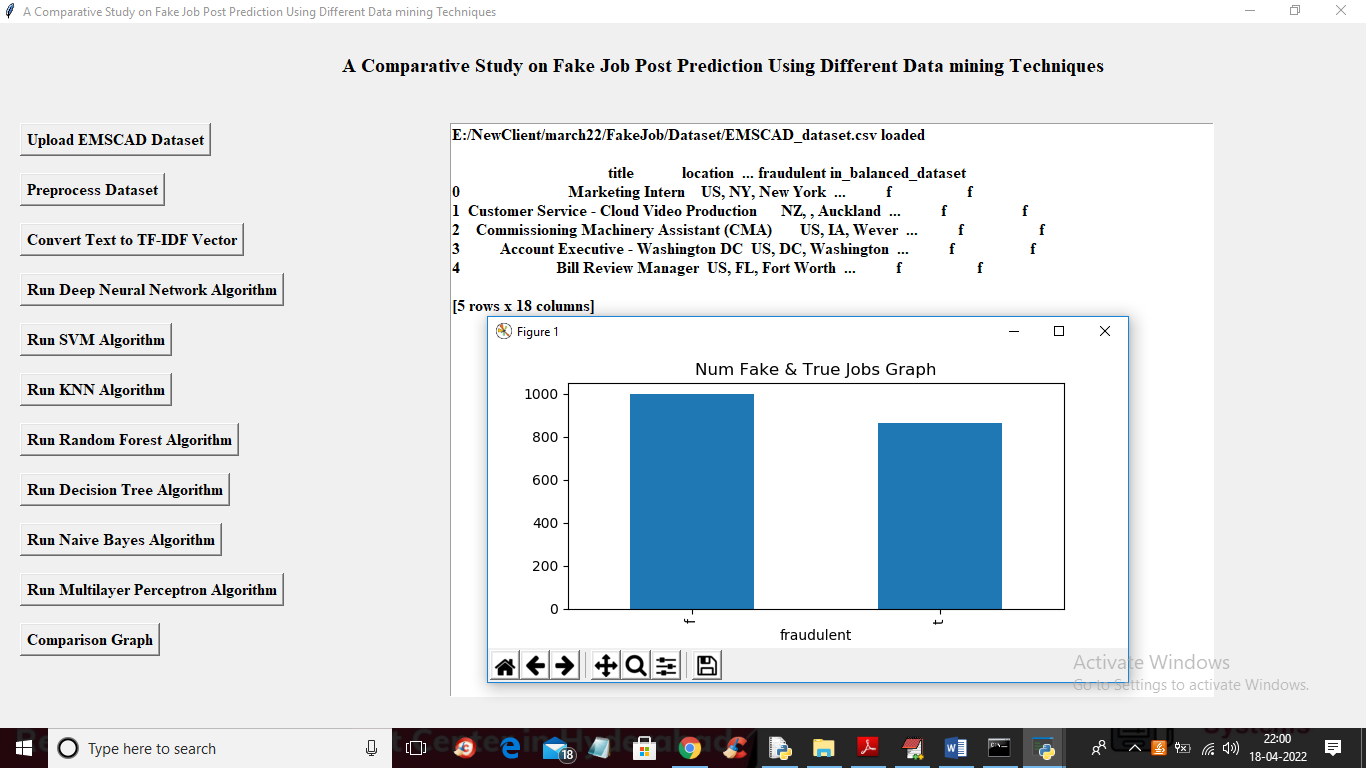
To run project double click on ‘run.bat’ file to get below screen



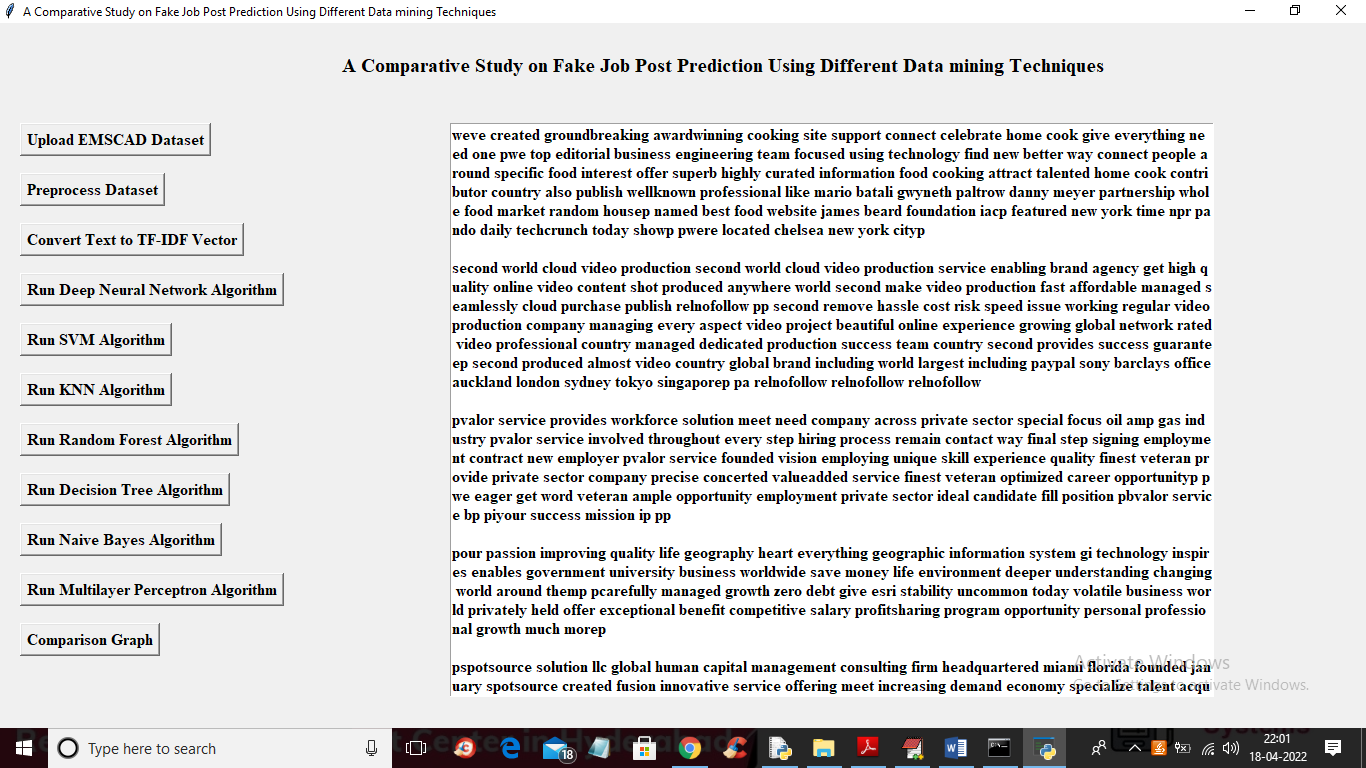
In above screen click on ‘Upload EMSCAD Dataset’ button to upload dataset to application



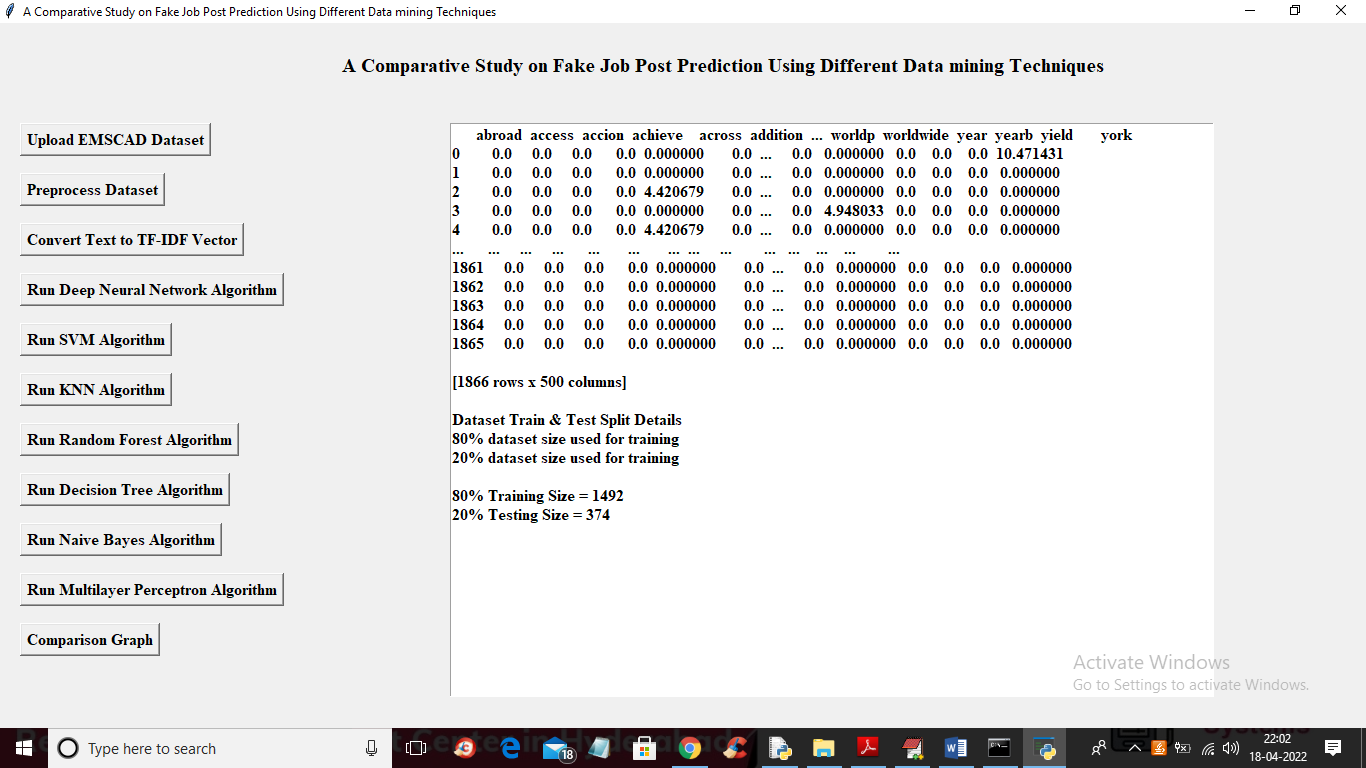
In above screen selecting and uploading dataset file and then click on ‘Open’ button to load dataset and to get below screen



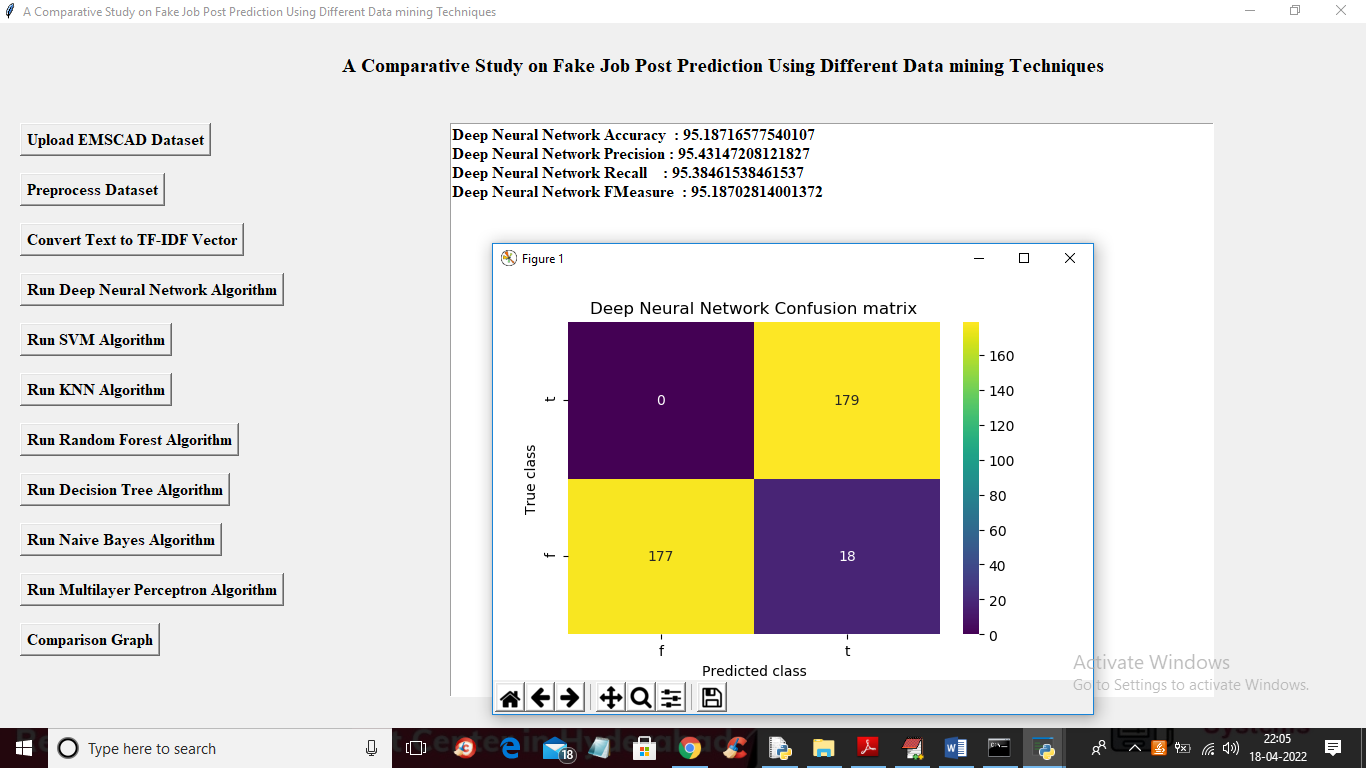
In above screen we can see dataset loaded and in graph x-axis represents ‘f’ and ‘t’ and y-axis represent counts of fake and true jobs and now close above graph and then click on ‘Preprocess Dataset’ button to remove missing and stop words from dataset



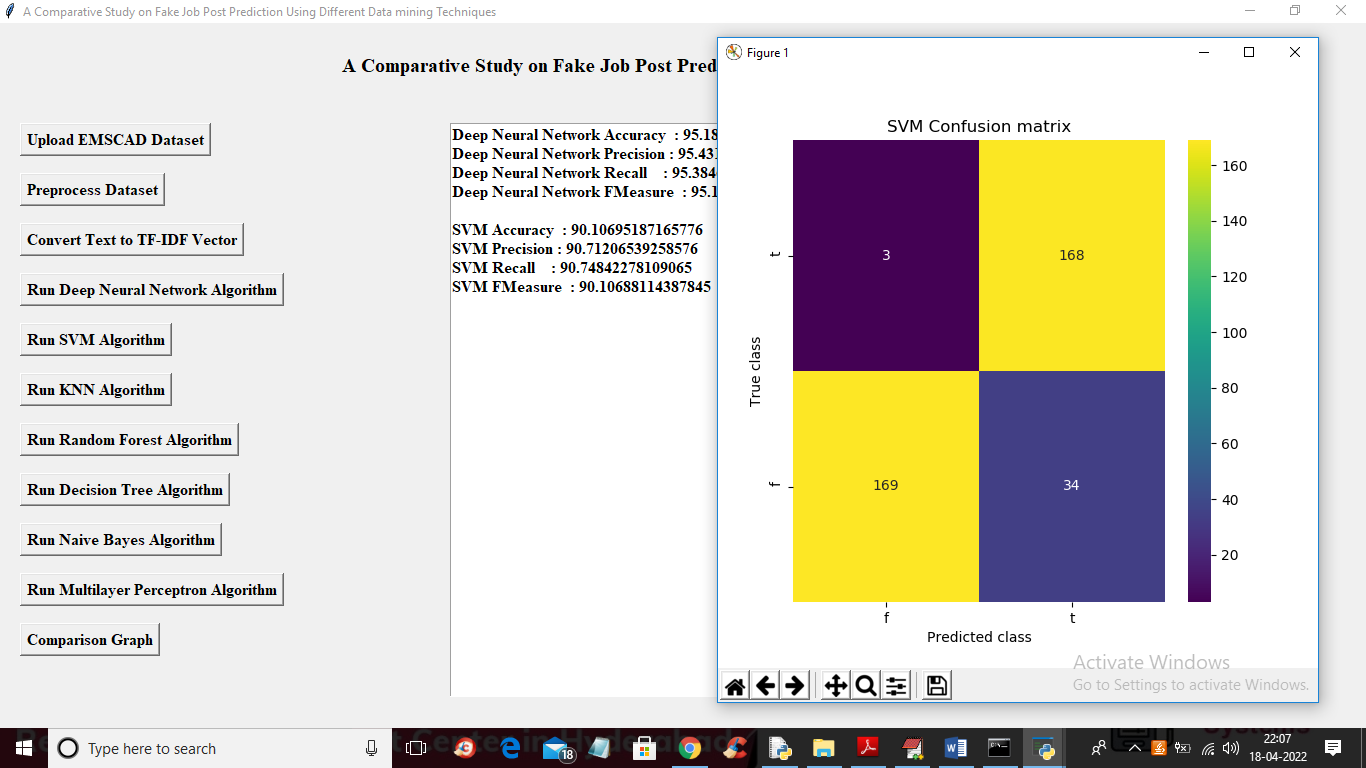
In above screen from dataset TEXT all special symbols and missing values and stop words like ‘and the this etc.’ are removed out and now click on ‘Convert Text to TF-IDF Vector’ button to convert text to numeric vector like below screen



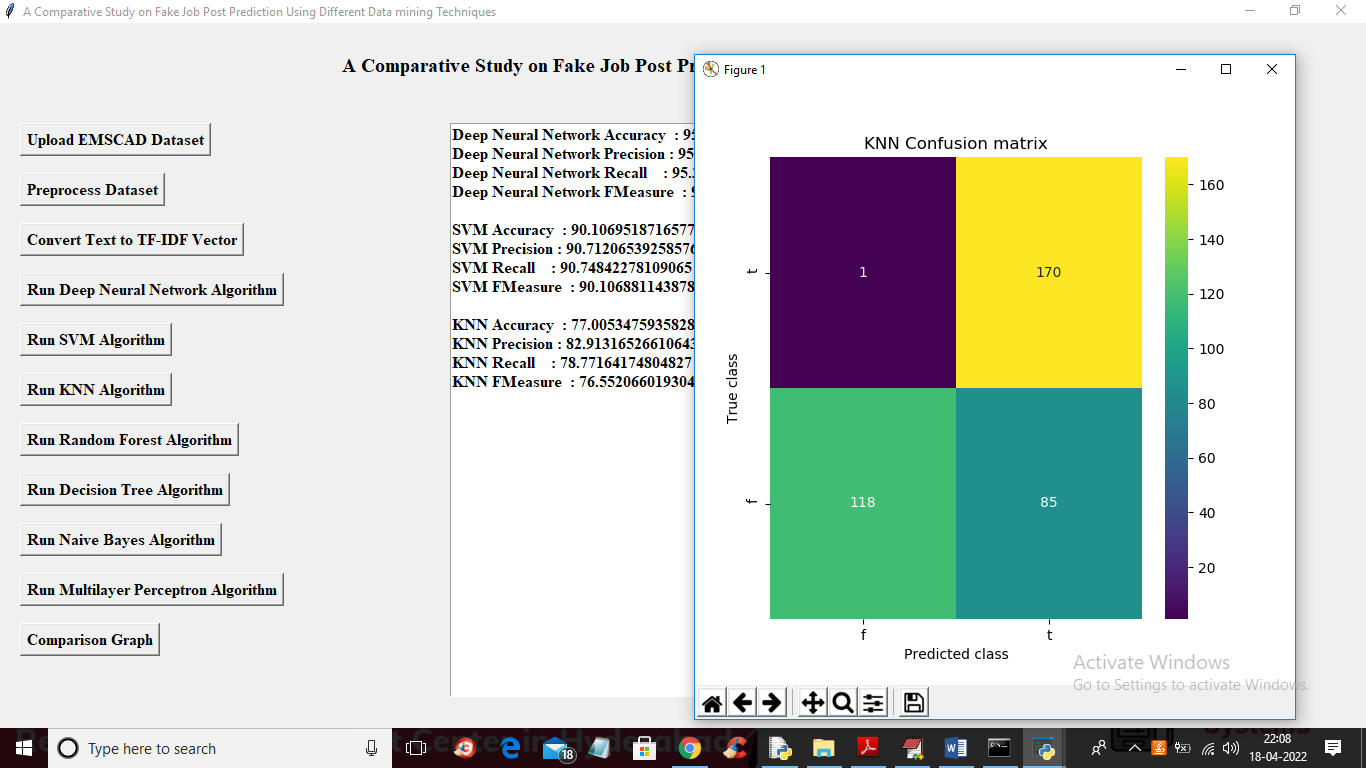
In above screen we can see dataset values are converted to vector where first row contains WORDS and remaining rows contains average frequency of those words. In above screen in last lines we can see dataset train and test split details and now click on ‘Run Deep Neural Network Algorithm’ button to train neural network and get below output



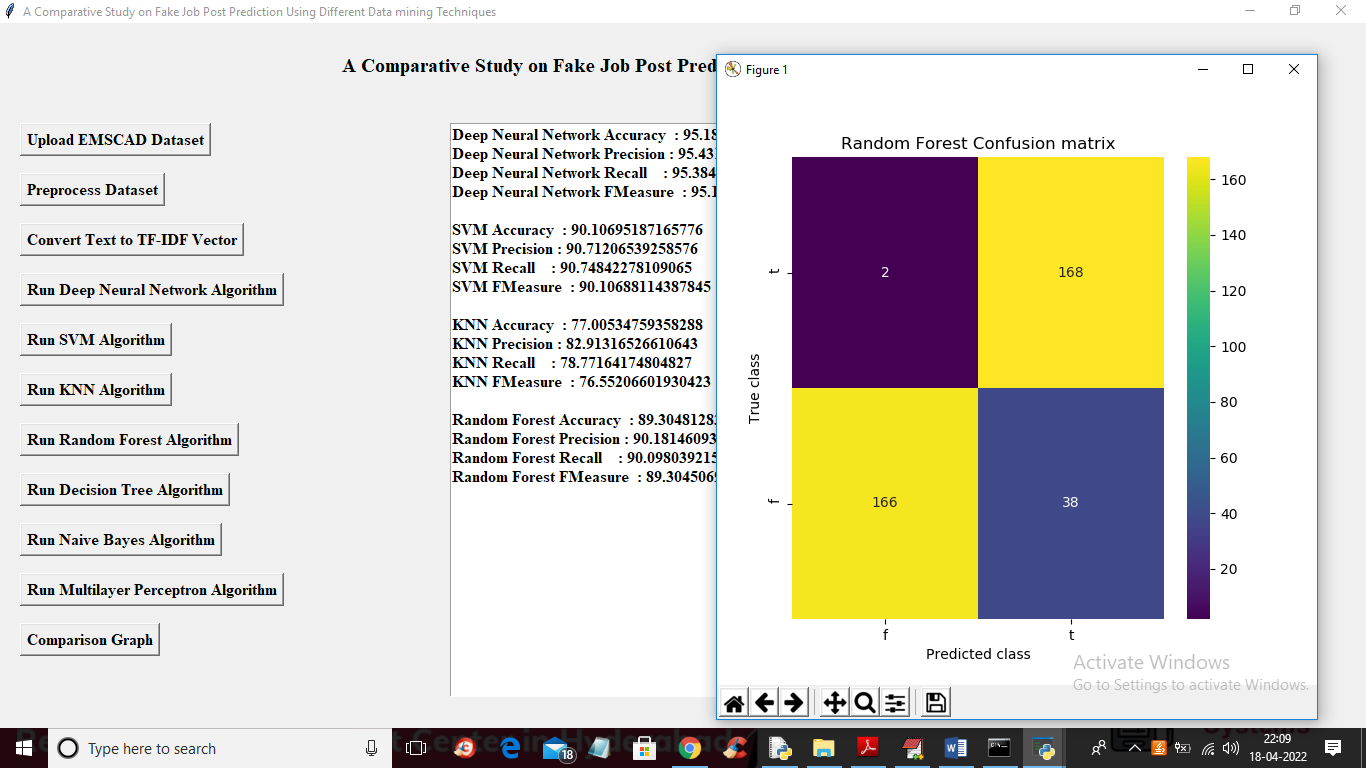
In above screen with Neural Network we got 95% accuracy and in confusion matrix graph x-axis represents PREDICTED CLASSES and y-axis represents TRUE TEST CLASSES and we can neural network predict 177 instances as FAKE correctly and 0 incorrectly and same we can see for true jobs. Now close above graph and then click on ‘Run SVM Algorithm’ button to get below output



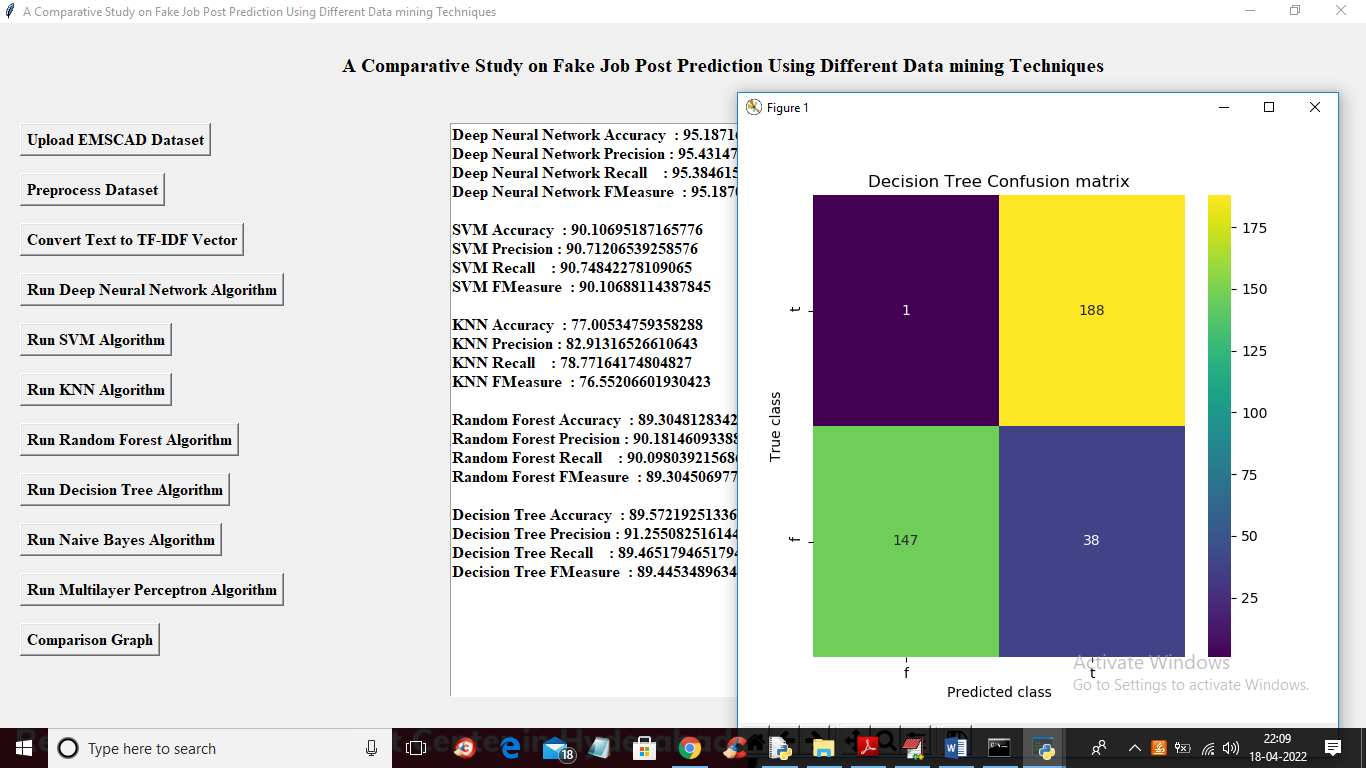
In above screen with SVM we got 90% accuracy and we can see its confusion matrix graph and now close above graph and then click on ‘Run KNN Algorithm’ button to get below output



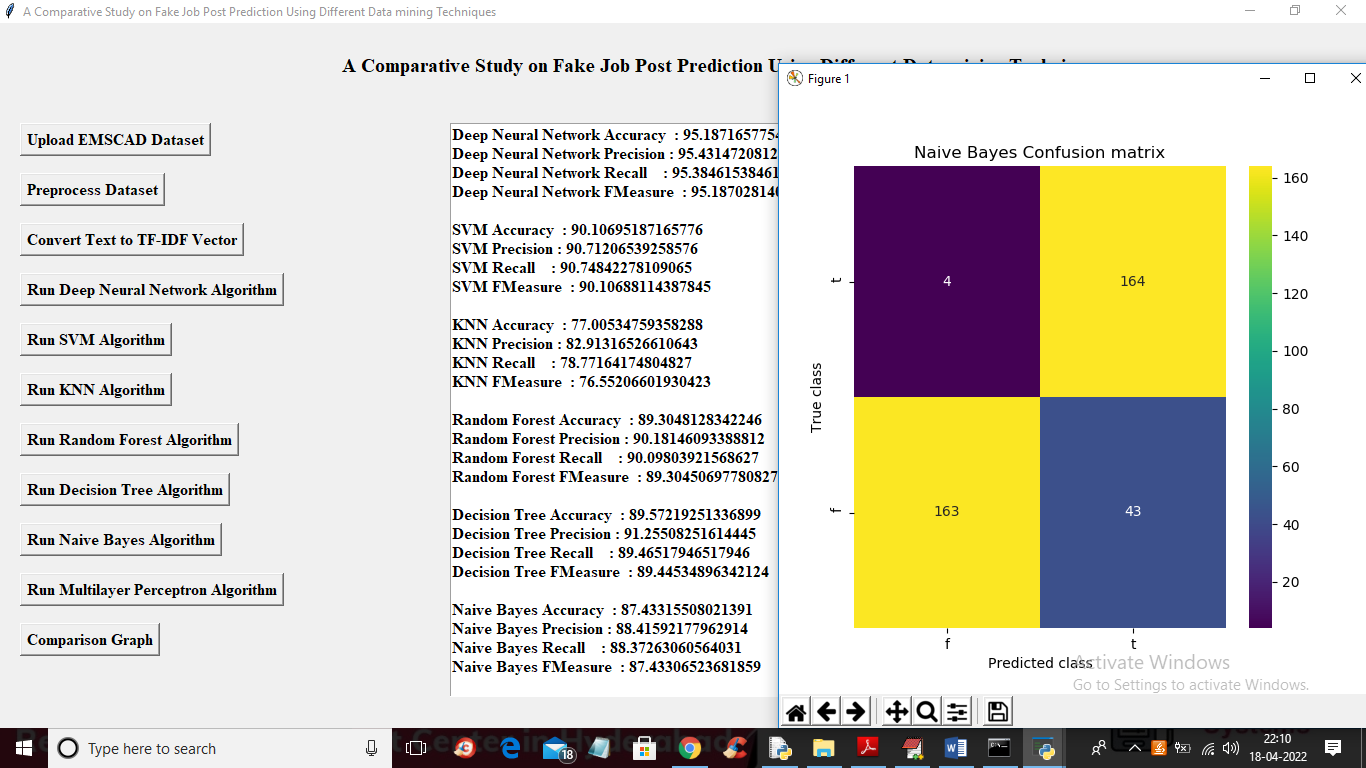
In above screen with KNN we got 77% accuracy and now close above graph and then click on ‘Run Random Forest’ button to get below output



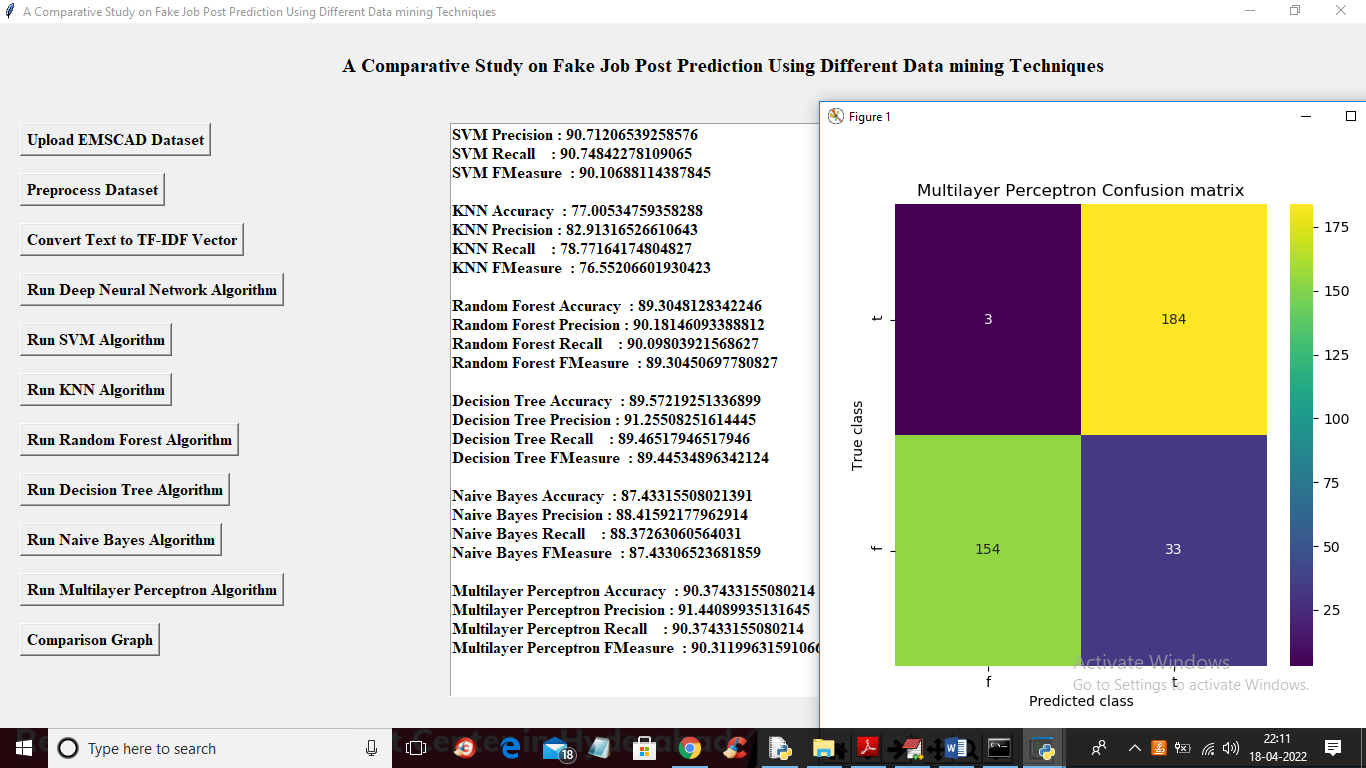
In above screen with Random Forest we got 89% accuracy and now run decision tree button



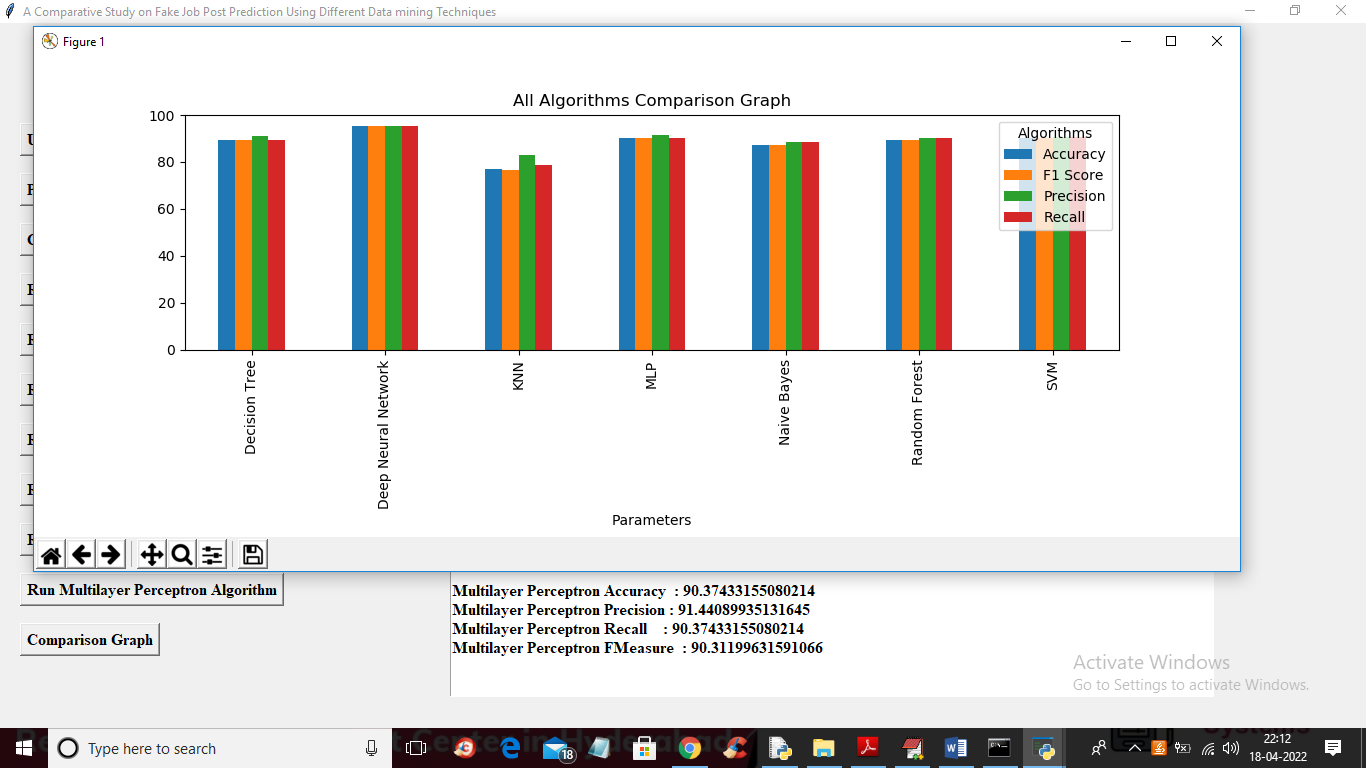
In above screen with decision tree we got 89.57% accuracy and now click on ‘Naïve Bayes’ button to get below output



In above screen with Naïve Bayes we got 87% accuracy and now click on ‘Run Multilayer Perceptron Algorithm’ button to get below output



In above screen with MLP we got 90% accuracy and now click on ‘Comparison Graph’ button to get below graph



In above graph x-axis represents algorithms names and y-axis represents accuracy, precision, recall and FSCORE in different colour bars and in above graph we can see in all algorithms Deep Neural Network got high accuracy and other values.