Predicting Personality from Twitter

In this paper author is asking to predict human personality by taking 5 features such as Openness, Agreeable, Neuroticism, Extroversion, and Conscientious and this features can be identified by social media dataset called ‘Twitter Profile’.

Openness means intelligent peoples who express their view in bold or open manner. This user expression can be identified by analysing his twitter profile and twitter messages, if person is intelligent then he will use open (open words also called as swear words) or bold words in his tweets. By looking for such words we can identify this person as Openness personality. LIWC dictionary contains all open or swear words by applying this dictionary on tweets messages we can predict Openness personality score. If predicted score > 0.1 then this person will put under this category.

Agreeable means peoples who use words such as ‘am, will have and this words also refers as ARTICLES or AUXILIARY VERBS’ etc will come in this category. MRC dictionary contains all words of this categories and by applying this dictionary on user’s tweets we can predict person category as agreeable.

Neuroticism means peoples in this category is consider as sentiment or emotion, peoples who use words such as ‘ugly, nasty, sad’ etc will come under this category. By looking for such words in tweets we can predict score of this category.

Extroversion means peoples of this category are friendly and person who has many number of friends or followers or following in twitter profile will comes under this category.

Conscientious means peoples who express hard working ideas in their post will come under this category.

So by analysing above 5 features O (openness), C (Conscientious), E (Extroversion), A (Agreeable), N (Neuroticism) from twitter profile and post we can predict personality of a person.

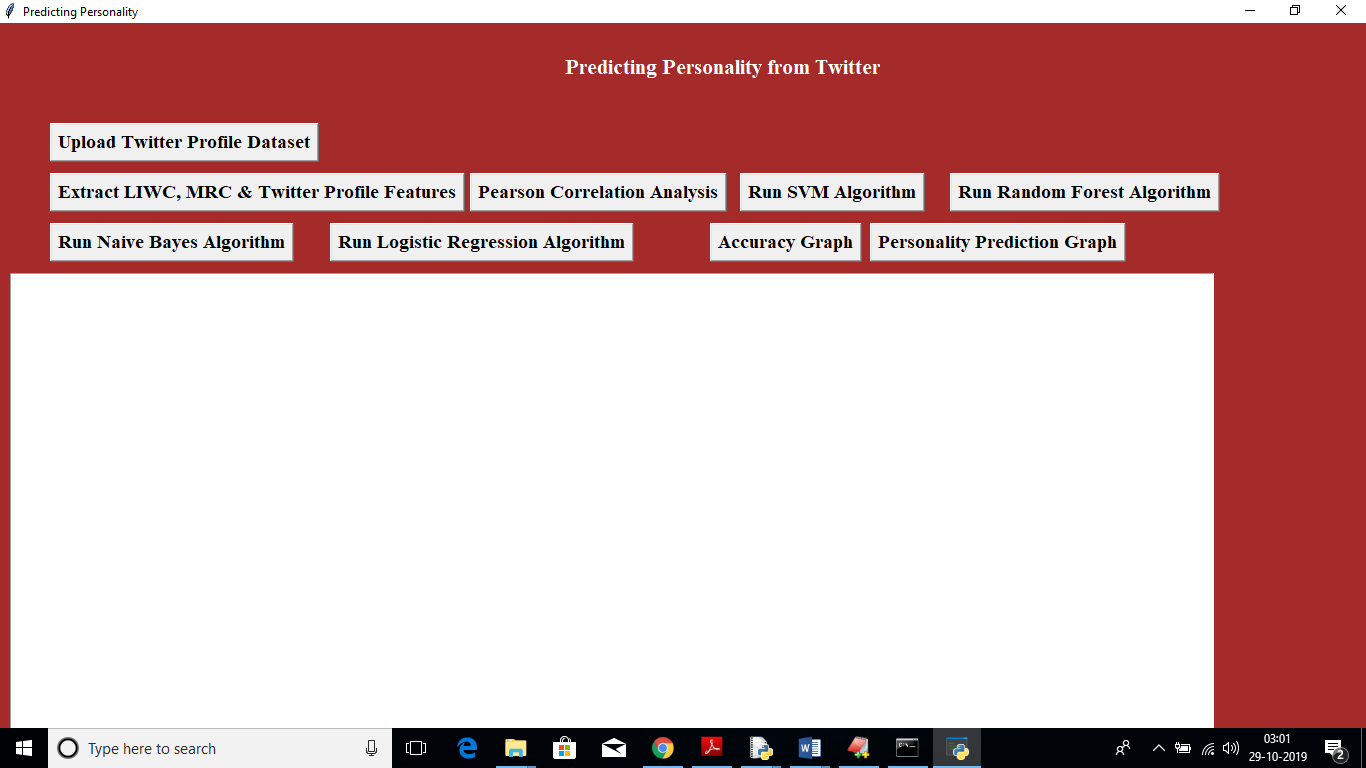
We will find average of each features from tweets and then apply Pearson Correlation formula to get score for all five features. If score > 0.1 for any feature then person belongs to that category. If person has 0.1 value for more than 1 features then that person personality belongs to that many categories. For example same person can be predicted as openness and conscientious etc.

All features values we will apply using SVM, Random Forest, Naïve Bayes & Logistic Regression algorithms to calculate accuracy of dataset and algorithms

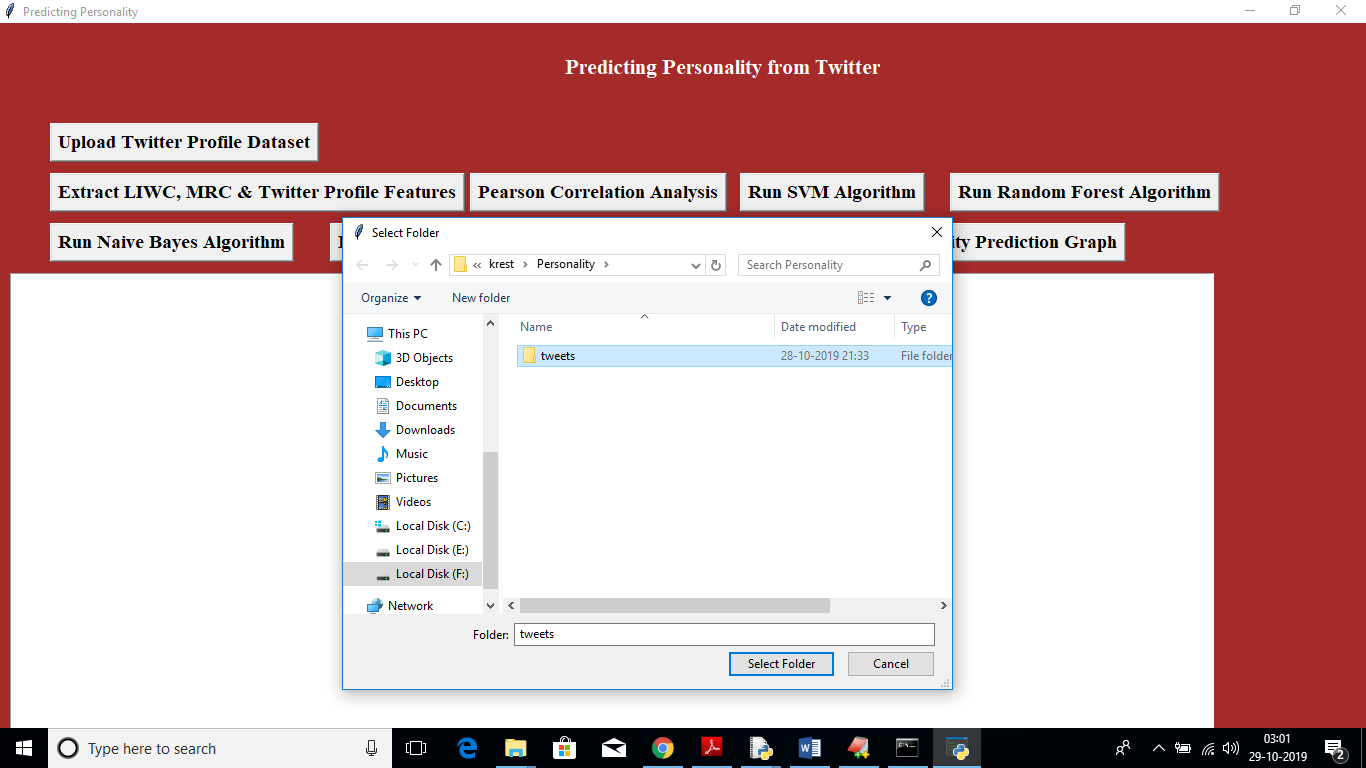
Screen shots

To run this project we used twitter dataset which contains tweets and user details in JSON format. This dataset available inside ‘tweets’ folder and each file contains tweets from 1 user

Double click on ‘run.bat’ file to get below screen



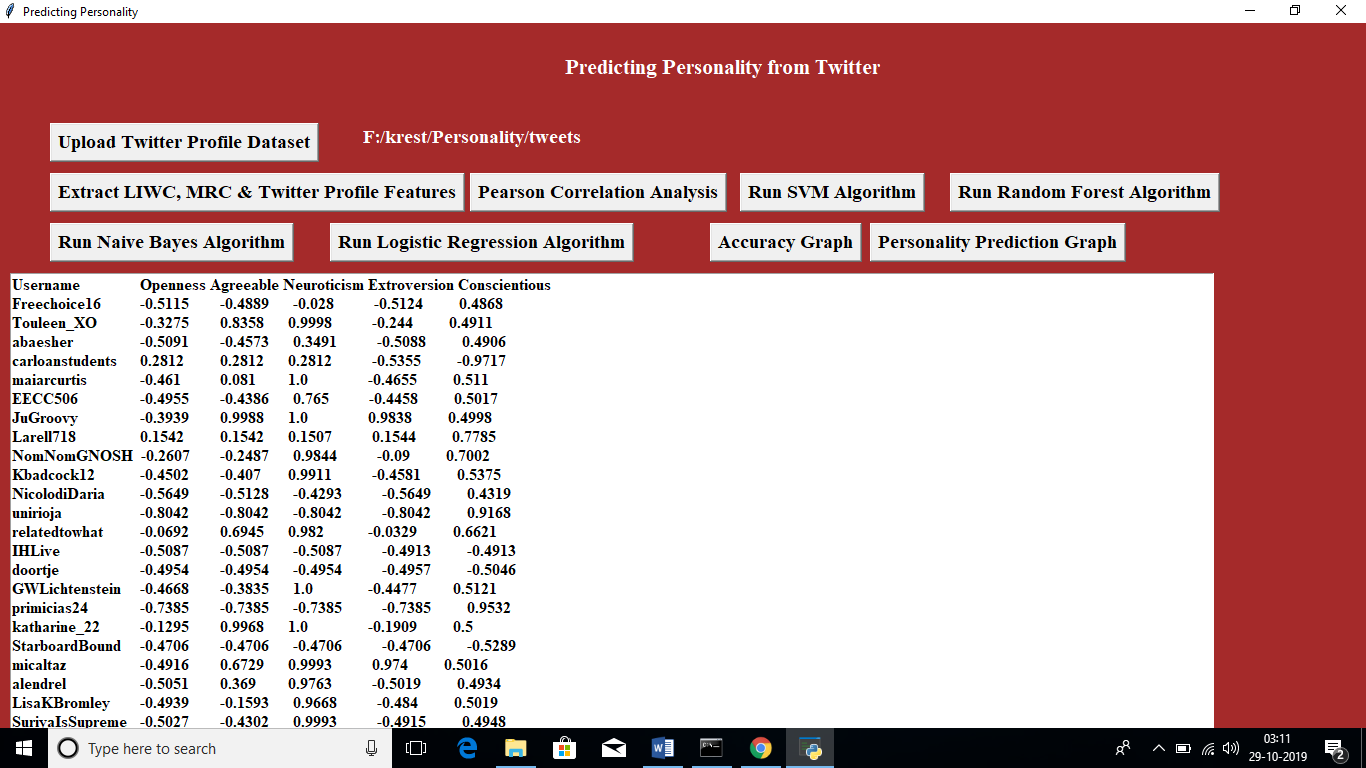
In above screen click on ‘Upload Twitter Profile Dataset’ button to upload dataset



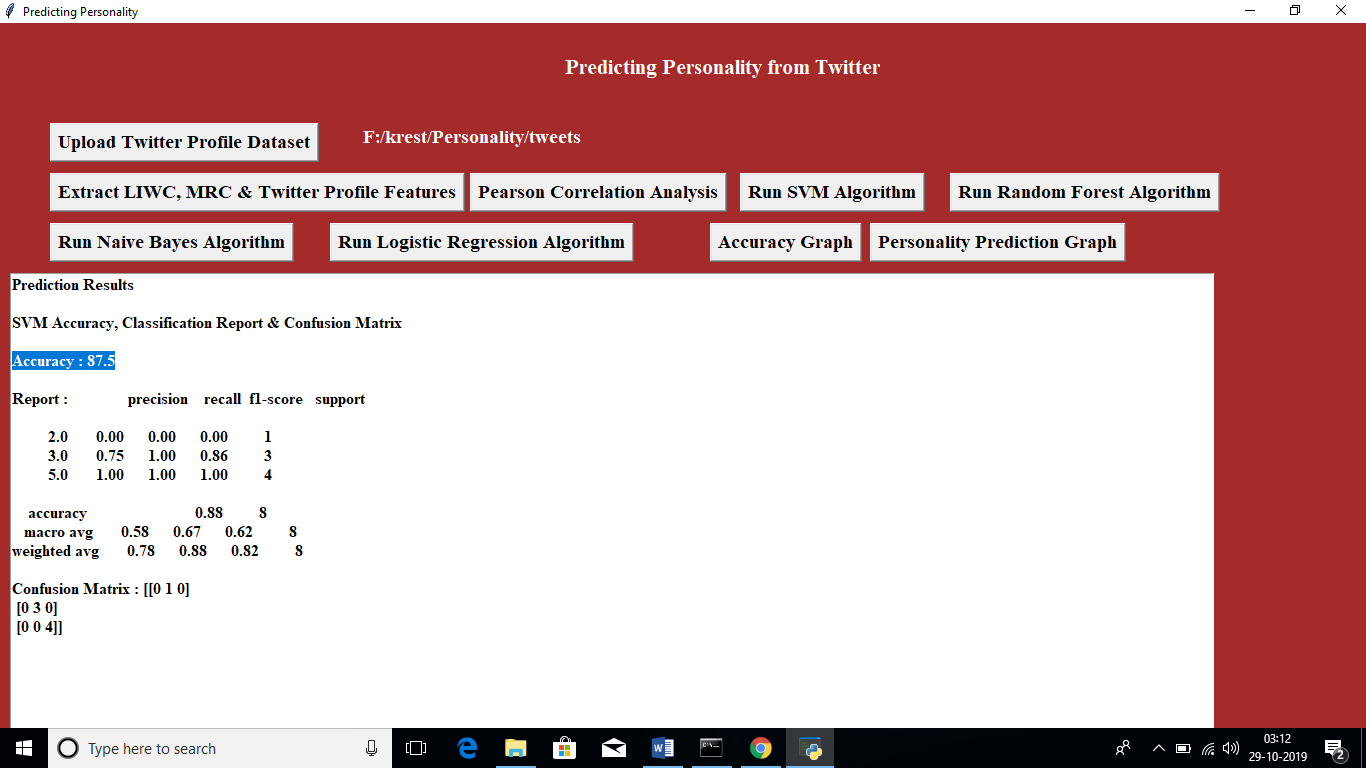
In above screen uploading tweets folder now click on ‘Extract LIWC Features’ button to extract features from tweets and profile



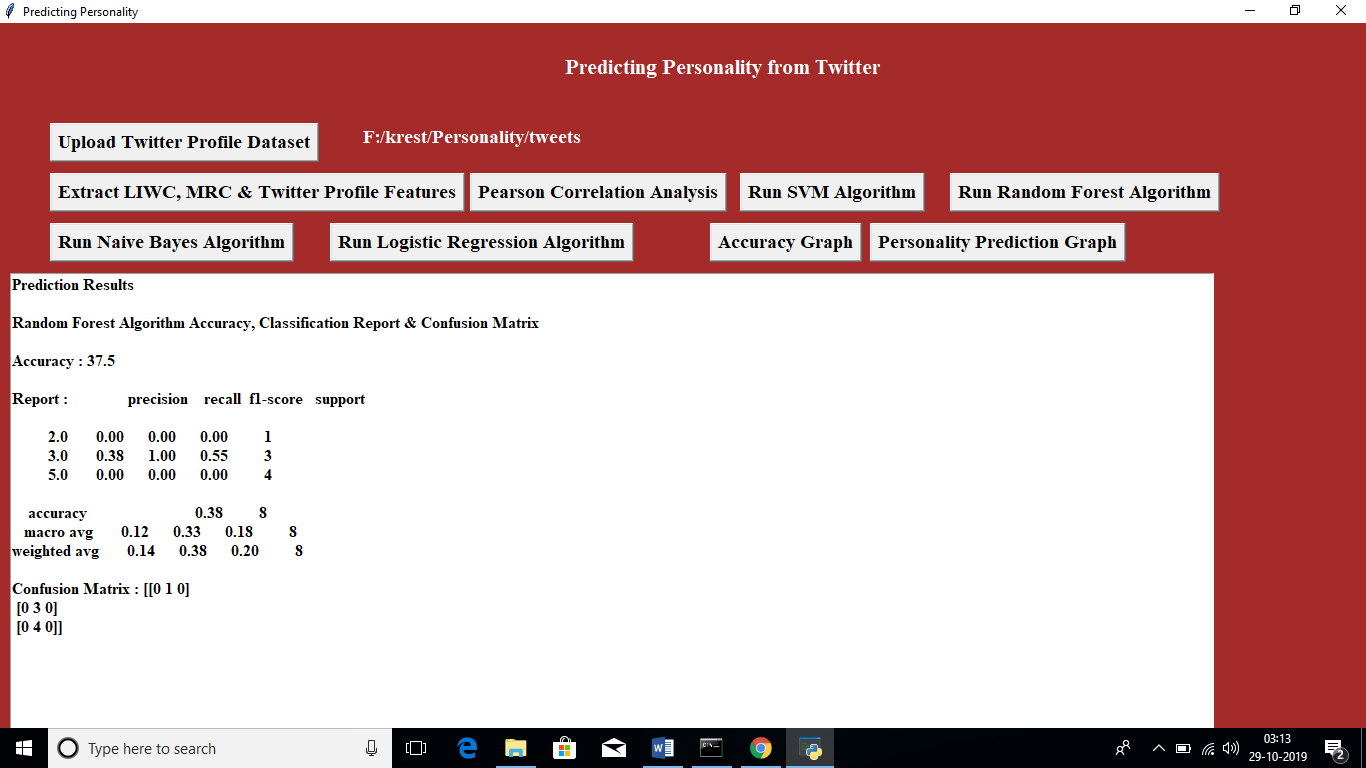
In above screen we can see we extract features from tweets and profile such as tweet text, following, followers etc. scroll down to see all profiles data. Now click on “Pearson Correlation Analysis’ to calculate scores for all five features



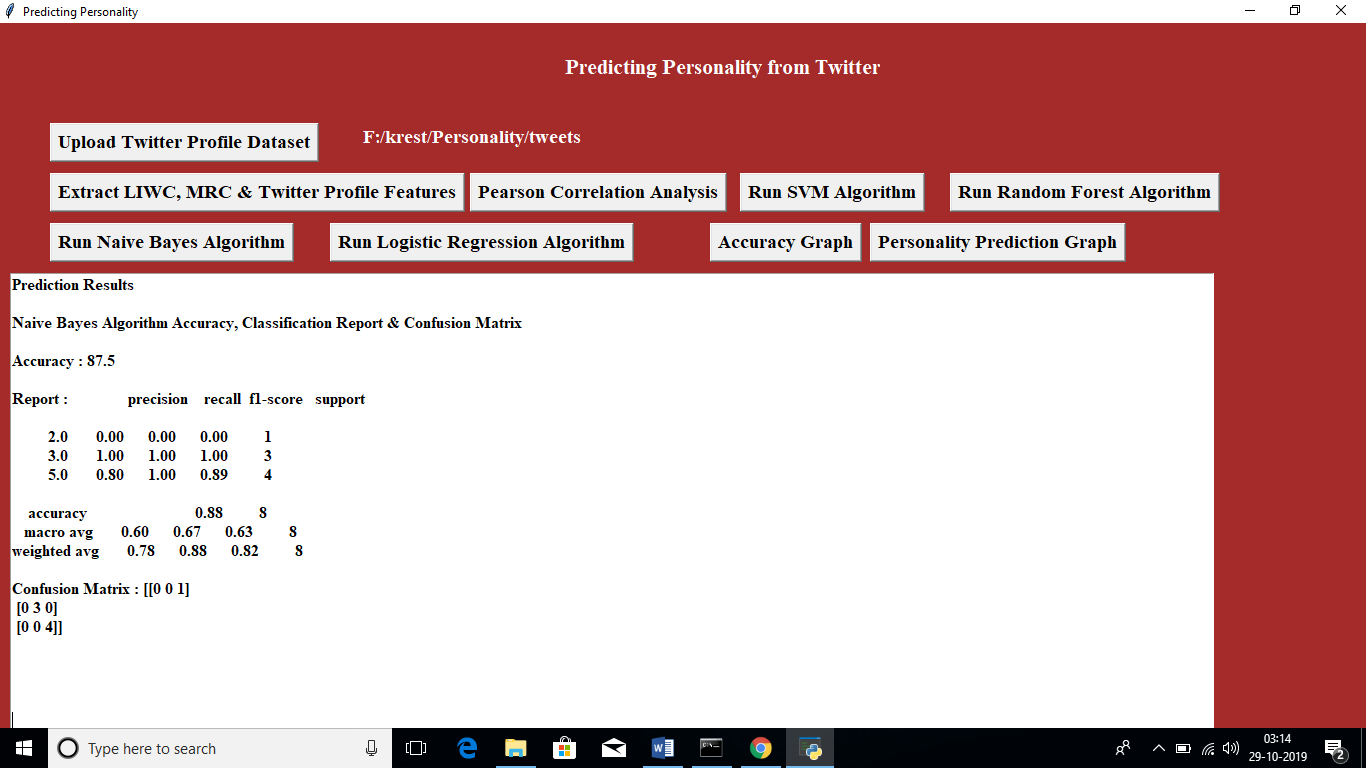
in above screen first columns contains username and rest of the columns are for features score. Now click on ‘Run SVM Algorithm’ button to get SVM Accuracy



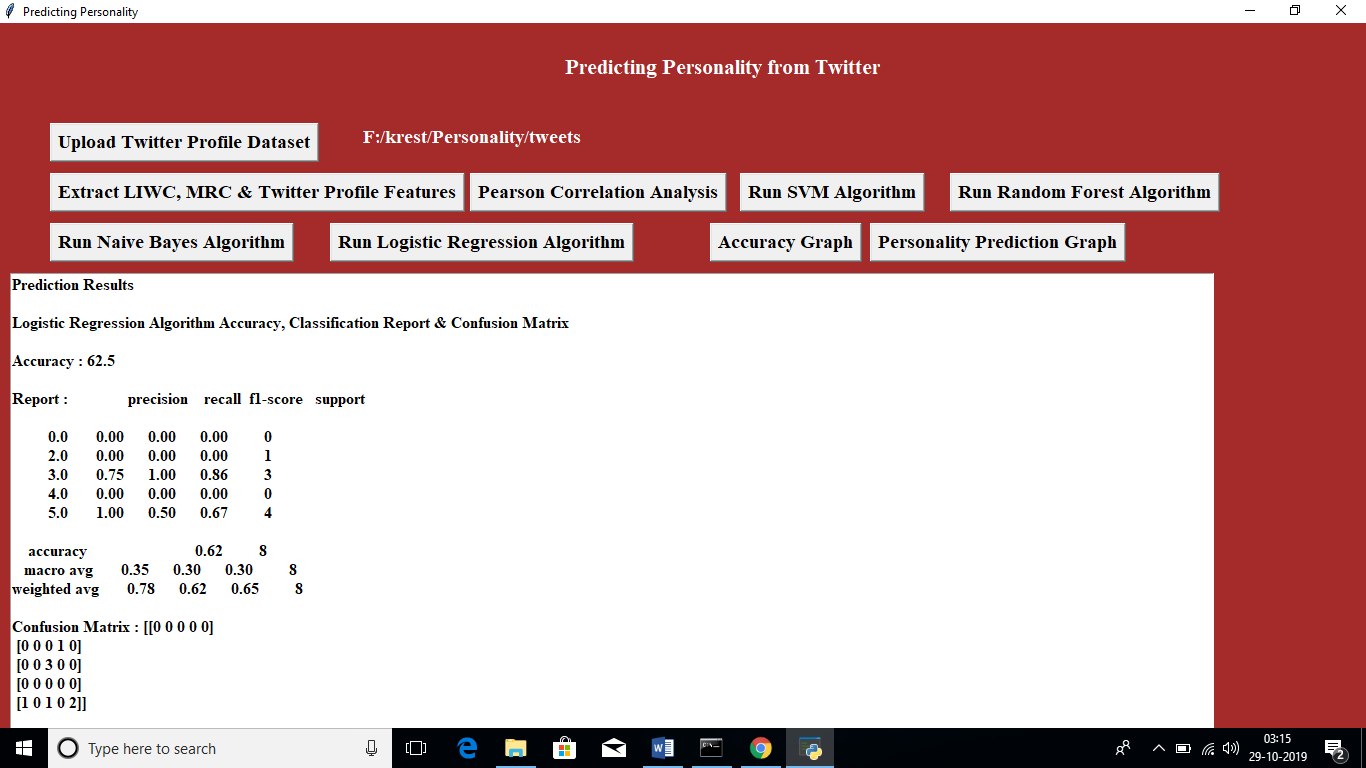
In above screen SVM accuracy is 87.5%, similarly click on ‘Run Random Forest Algorithm’ to get its accuracy



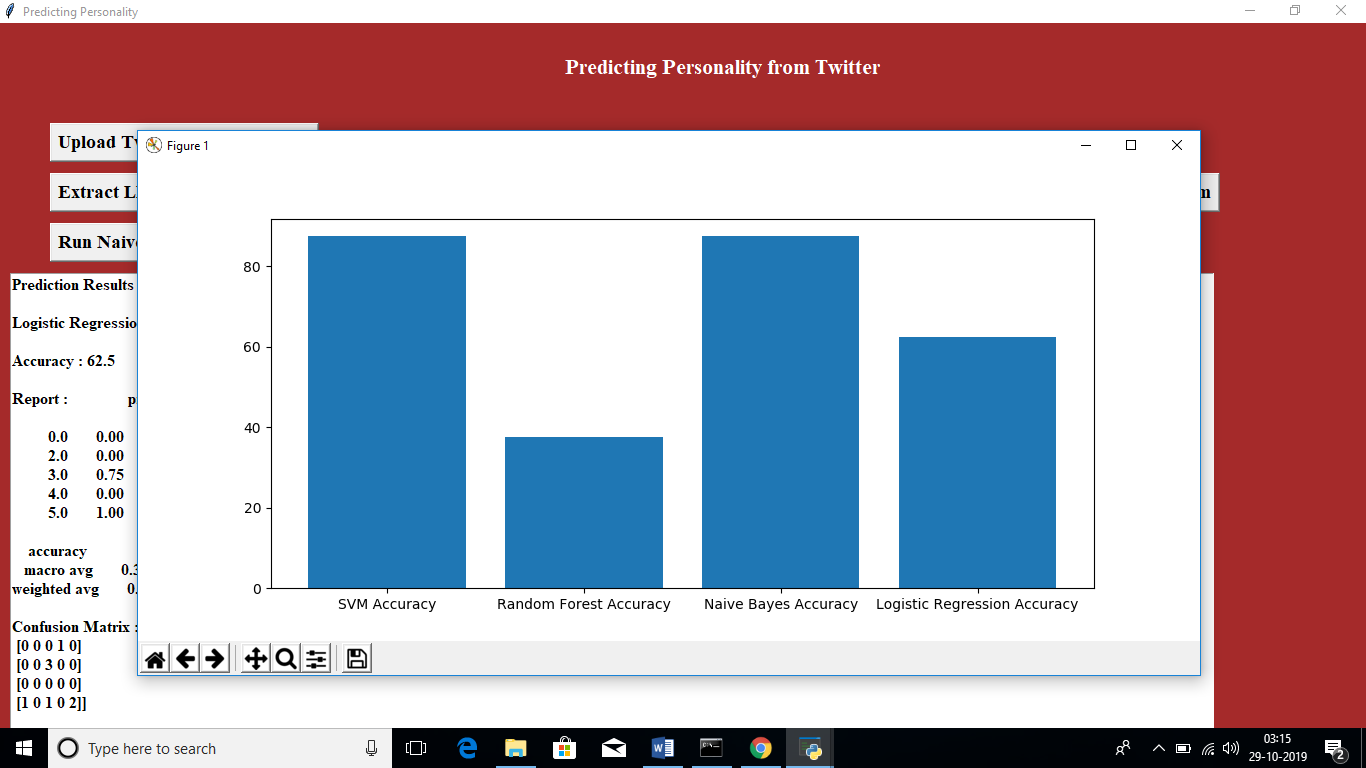
Now run ‘ Naïve Bayes’ to get its accuracy



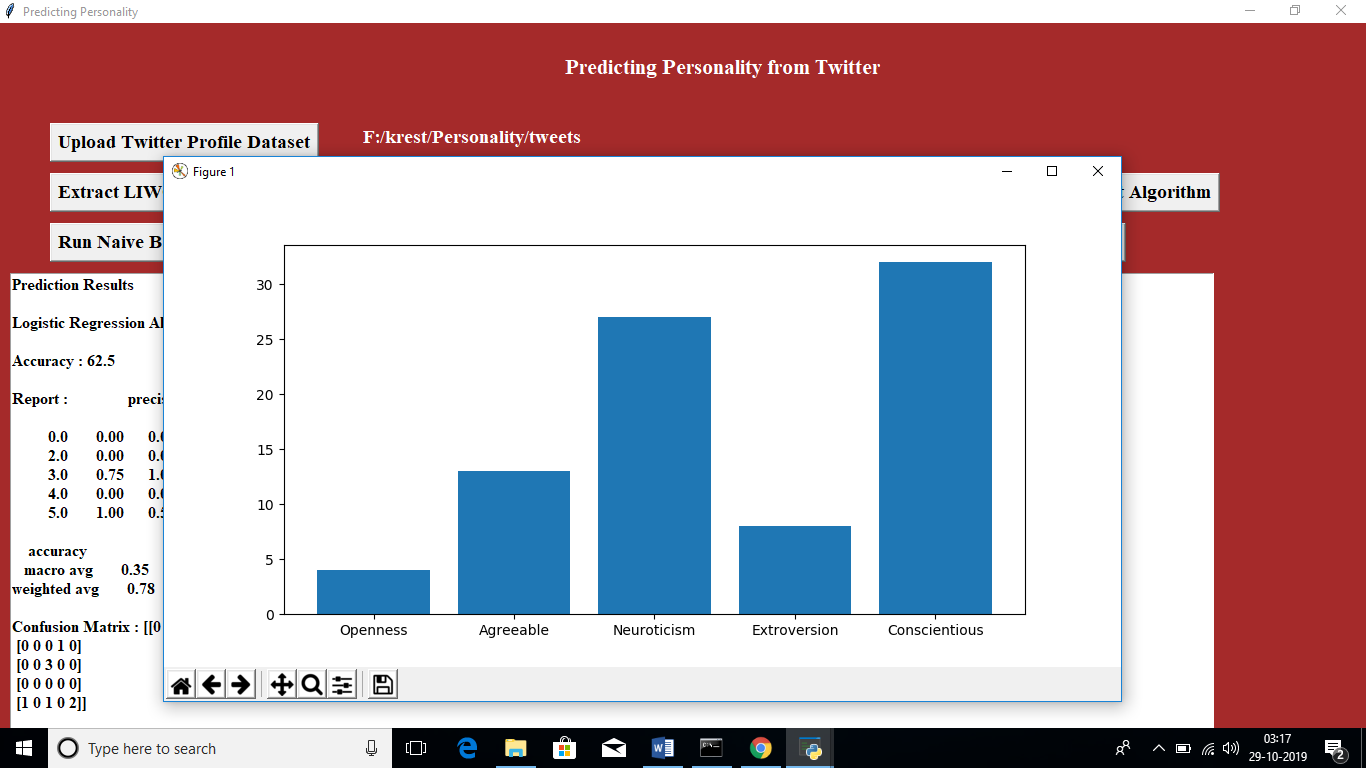
Now run ‘Logistic Regression’ Algorithm to get its accuracy



Now click on “Accuracy Graph’ button to get below accuracy graph



In above graph x-axis represents algorithm name and y-axis represents accuracy of that algorithm. Now click on ‘Personality Prediction Graph’ to get number of peoples in each category graph



In above graph x-axis represents feature category name and y-axis represents number of peoples in that category