Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms

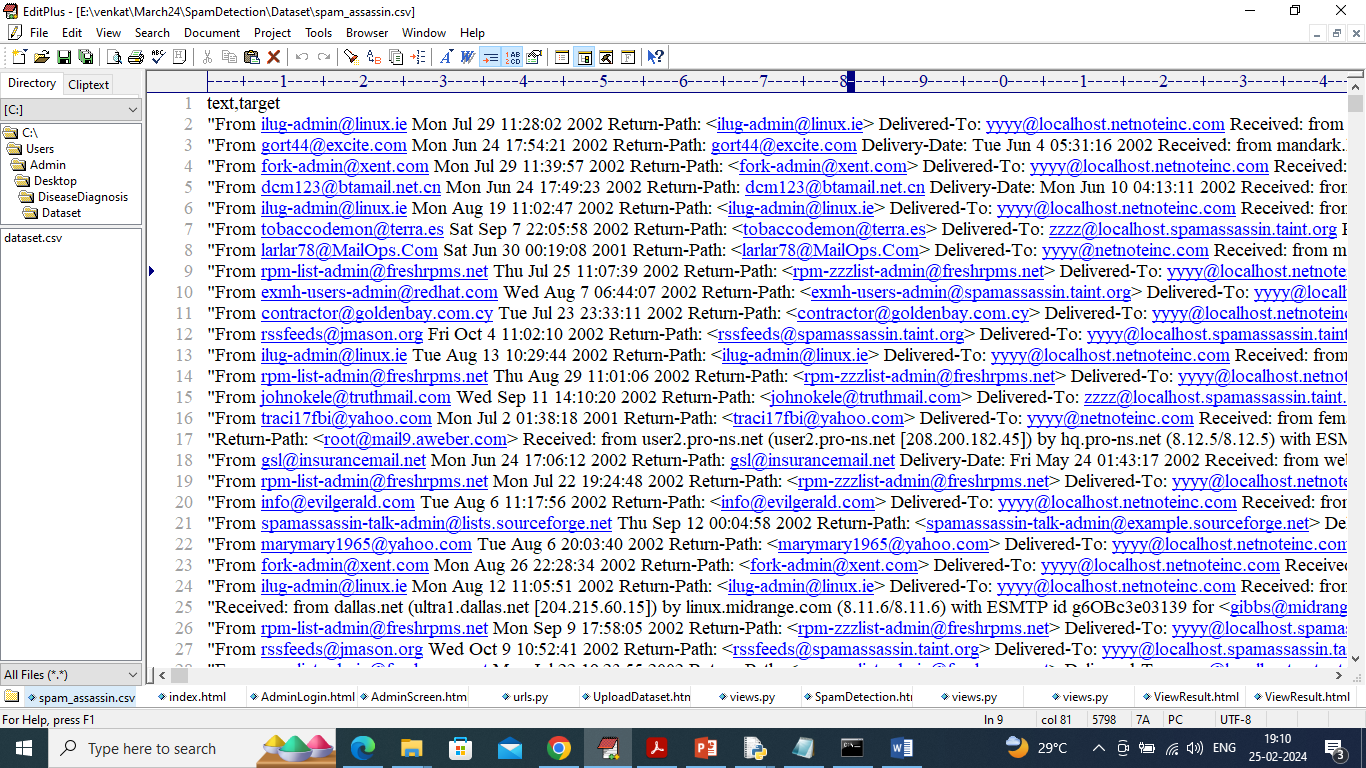
Now-a-days maximum peoples are utilizing e-mail services for business or formal communication but some malicious users will send spam e-mails with enticing links embedded in e-mails like lottery winning. Once user click on such link then malicious software embedded in link will get executed in user system and start stealing secured data and send to attacker.

To detect such attacks many signature and machine learning algorithms were developed but all those algorithms detection accuracy is not satisfactory so author of this paper employing bio-inspired Metaheuristic Algorithms like PSO and Genetic algorithm to select relevant features from training email data and then trained with various ML algorithms like Random Forest, SVM and Naïve Bayes. Each algorithm performance is evaluated using Accuracy, precision, recall and FSCORE.

Experimenting with bi-algorithms author found that some algorithms working best with GA and some working best with PSO and by using both bio-inspired technique all ML algorithms manages to obtain an accuracy between 95 to 98%.

To train above algorithms author has used 7 different datasets but training all those dataset is bit difficult so we have experimented with Spam Assassin dataset which contains email text and then label as 0 or 1 where 0 means Normal and 1 means SPAM.

In below screen showing dataset details



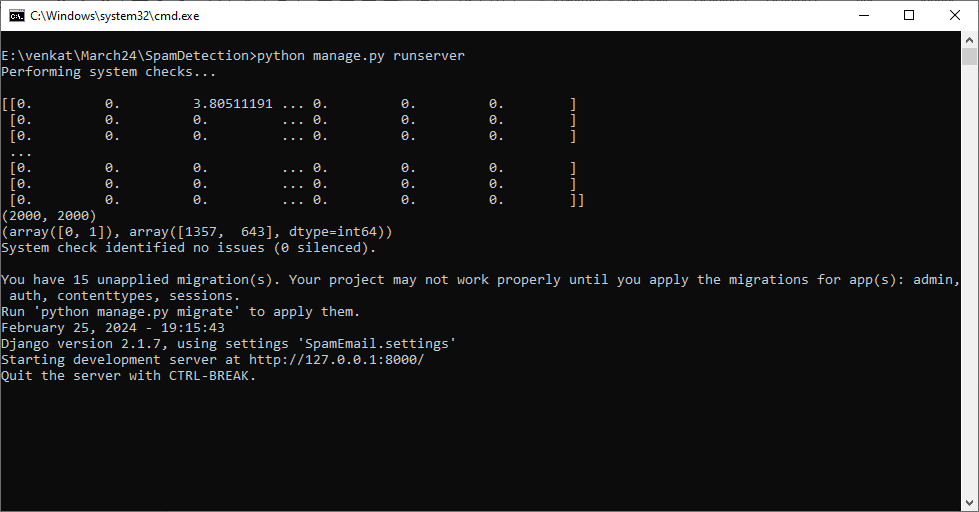
In above dataset screen first row represents dataset column names and remaining rows represents dataset values and by using above dataset will train and test all algorithms performance.

We have implemented this project we Web Application which you mention in your requirement and to implement this application we have design following modules

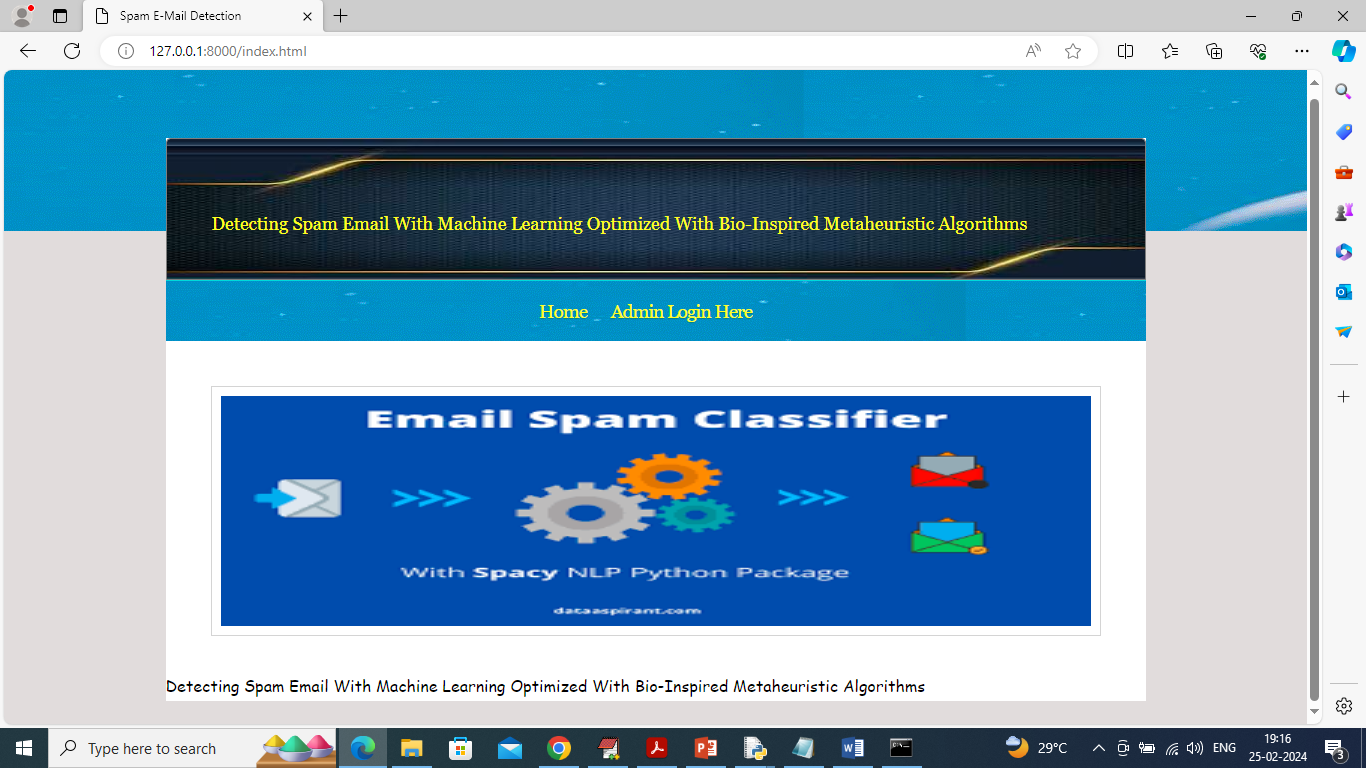
1. Admin Login: admin can login to system using username and password as admin
2. Upload Spam Assassin Dataset: after login admin will use this module to upload dataset and then convert all dataset text values into numeric TFIDF features
3. Run ML using GA: extracted features will be input to GA algorithm to select relevant features and then split them into train and test. All 3 ML algorithm will get trained on training features and then tested on test features to calculate accuracy and other metrics
4. Run ML using PSO: extracted features will be input to PSO algorithm to select relevant features and then split them into train and test. All 3 ML algorithm will get trained on training features and then tested on test features to calculate accuracy and other metrics
5. Comparison Graph: will plot comparison graph between GA and PSO
6. Spam Detection: using this module will upload test data and then ML algorithm will predict weather mail is Normal or Spam

SCREEN SHOTS

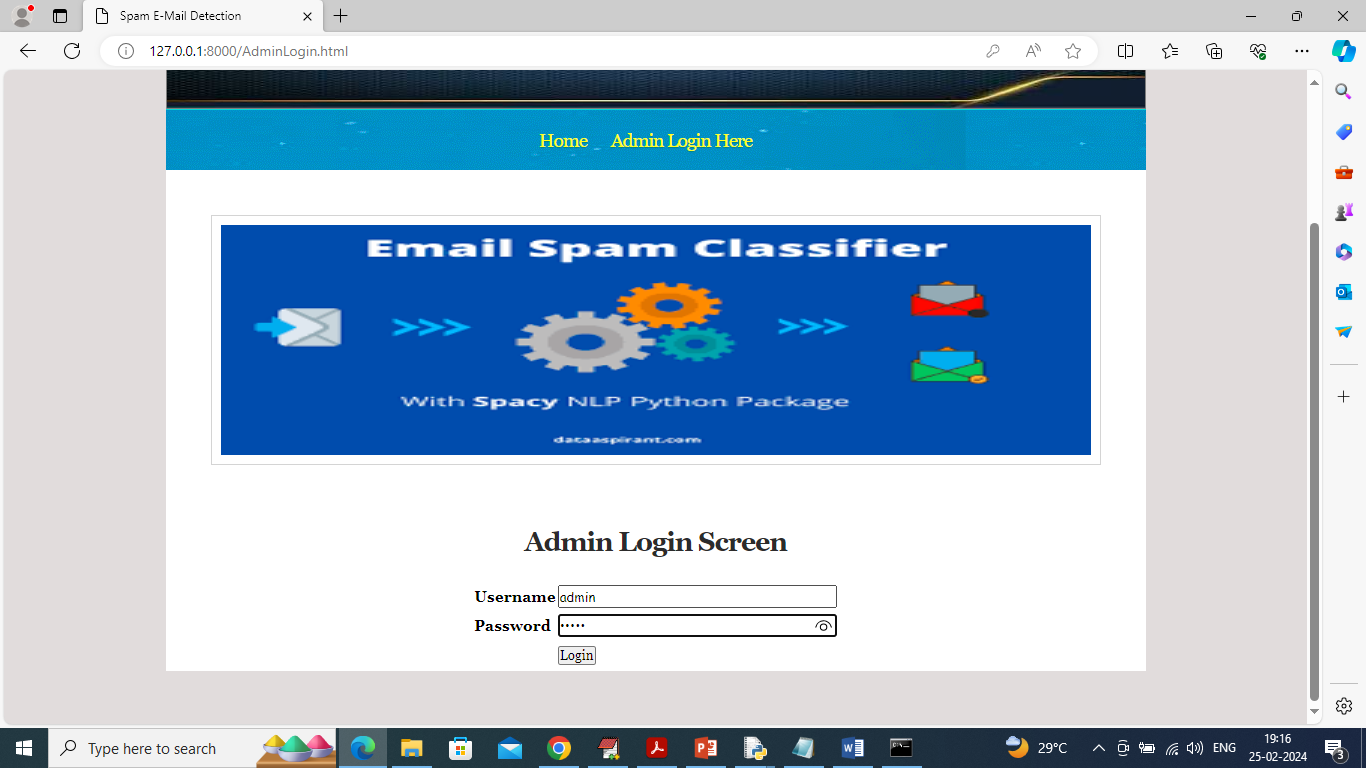
To run project double click on run.bat file to get below screen



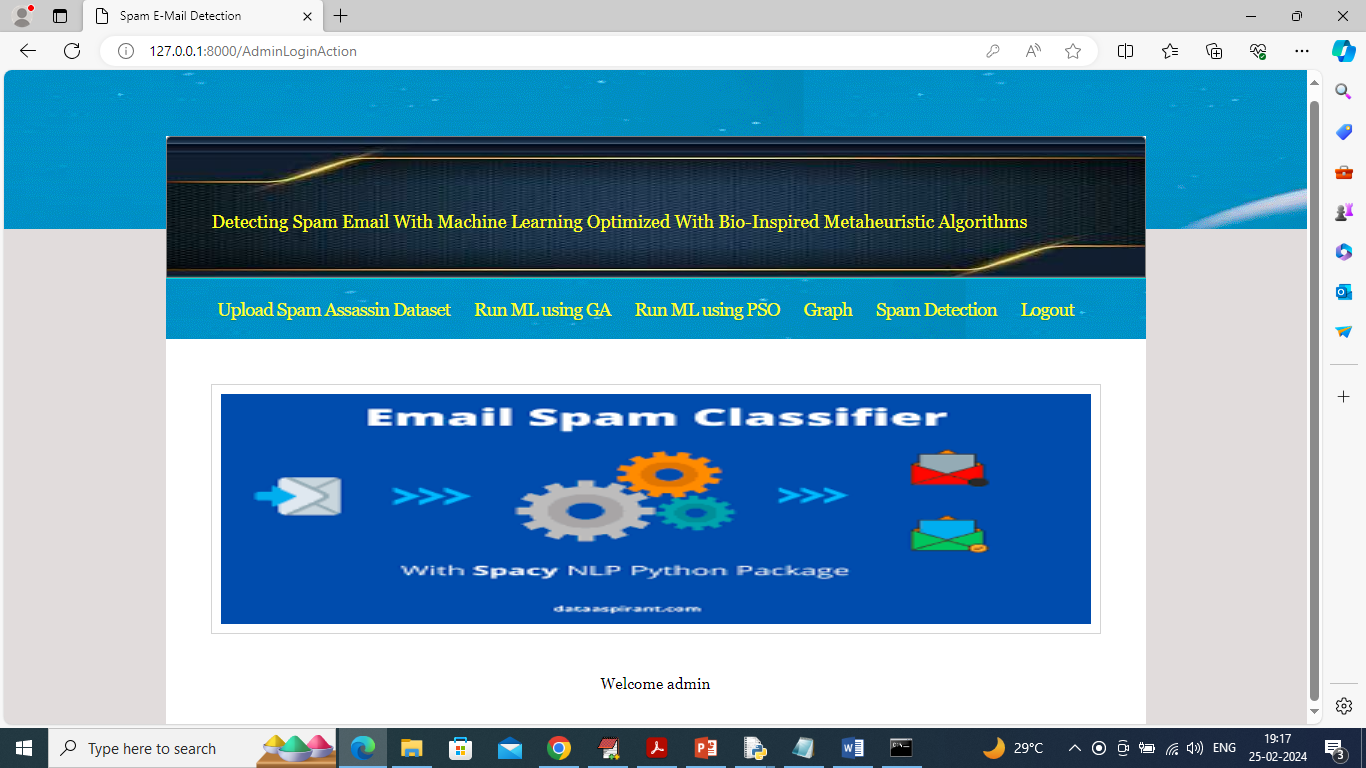
In above screen python server started and now open browser and enter URL as <http://127.0.0.1:8000/index.html> and press enter key to get below page



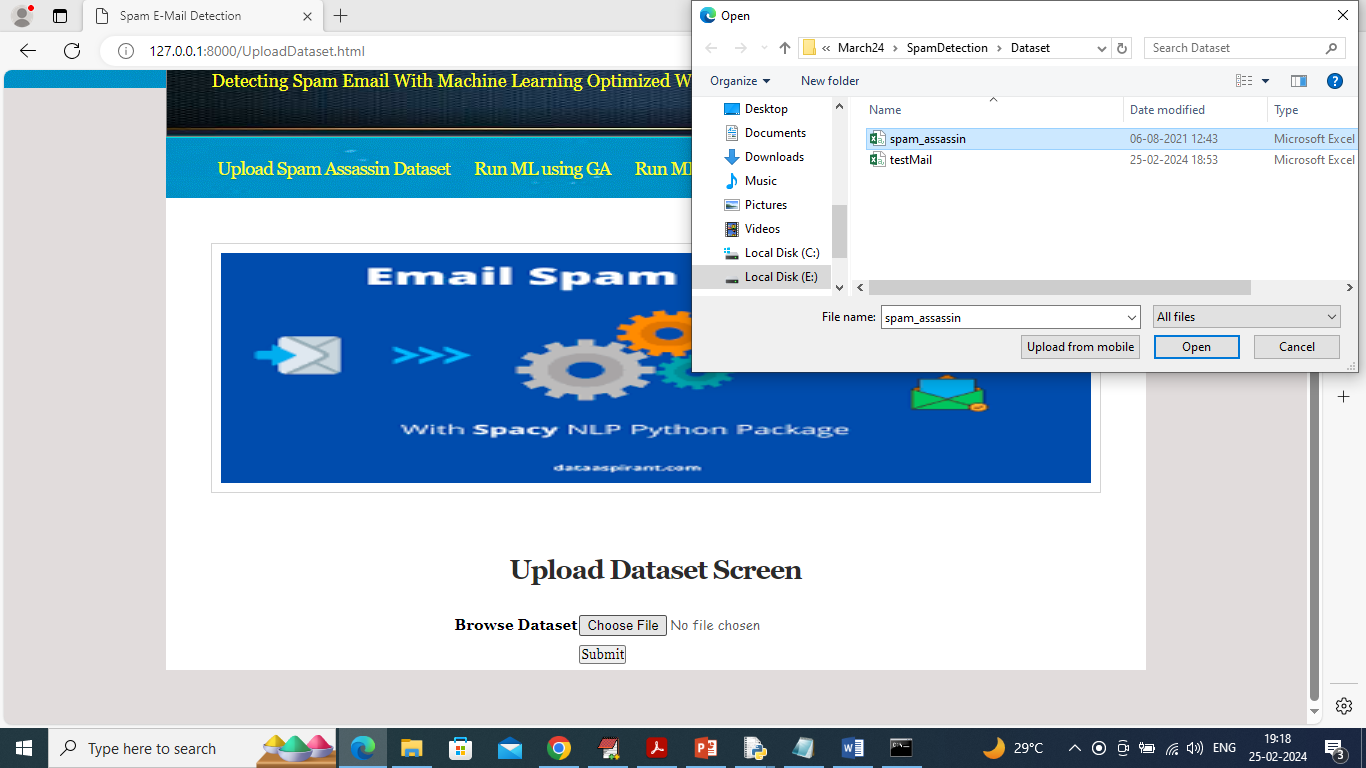
In above screen click on ‘Admin Login’ link to get below login page



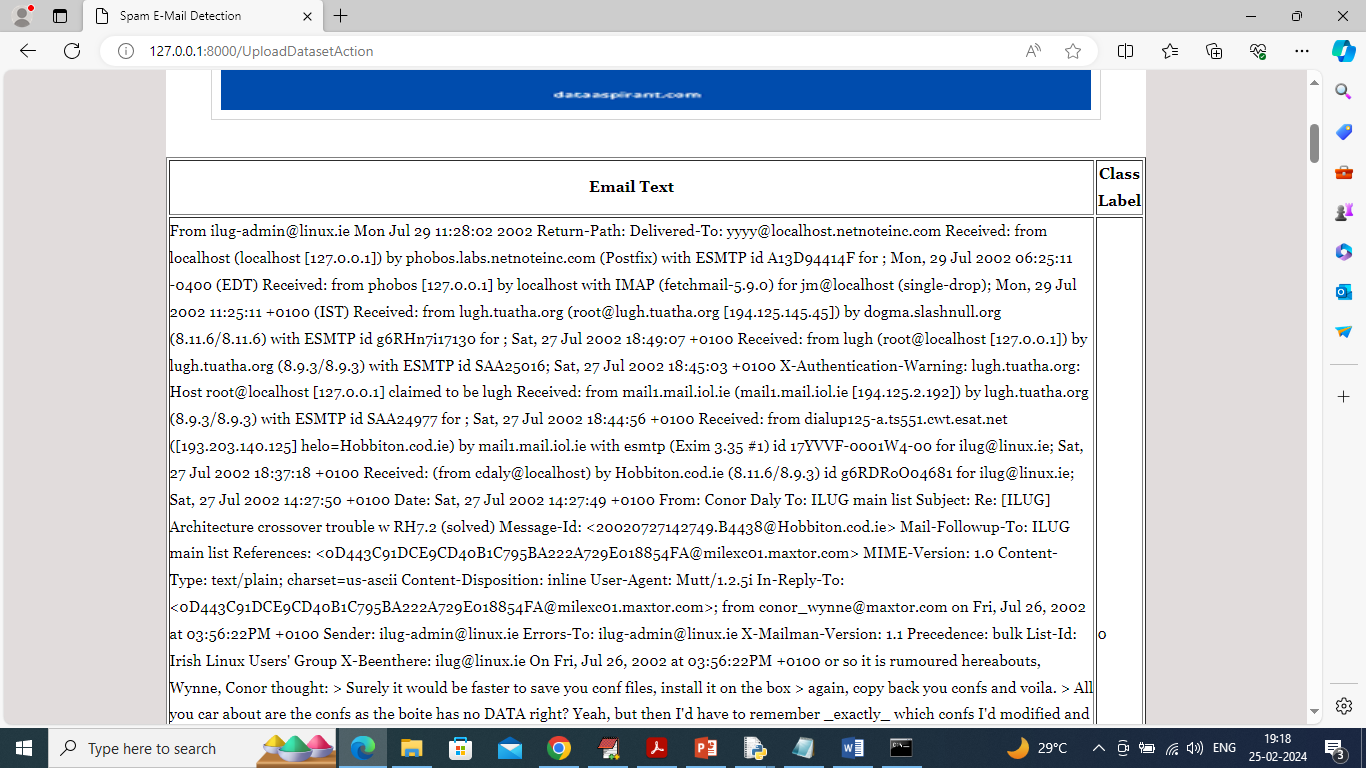
In above screen admin is login and after login will get below page



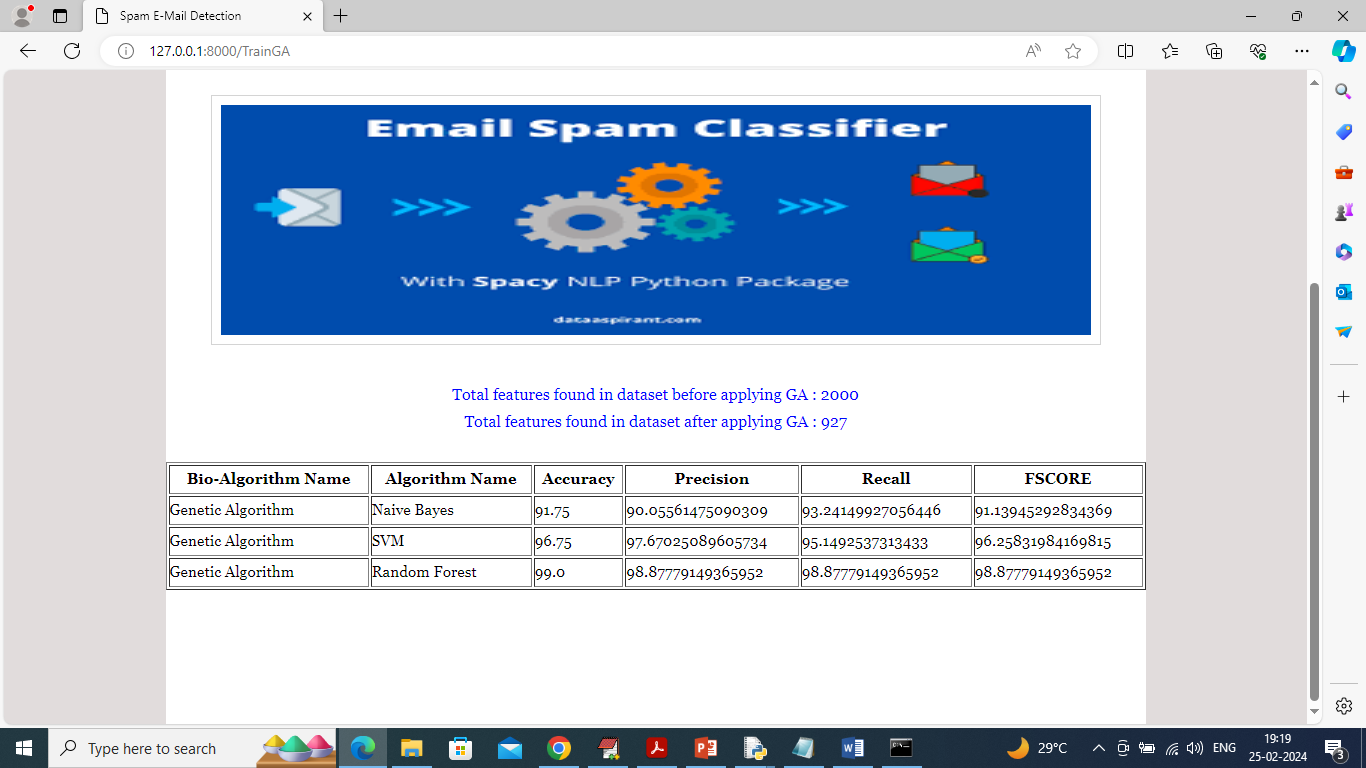
In above screen click on ‘Upload Email Assassin’ Dataset link to get below page



In above screen select and upload email assassin dataset and click on ‘Submit’ button to get below page



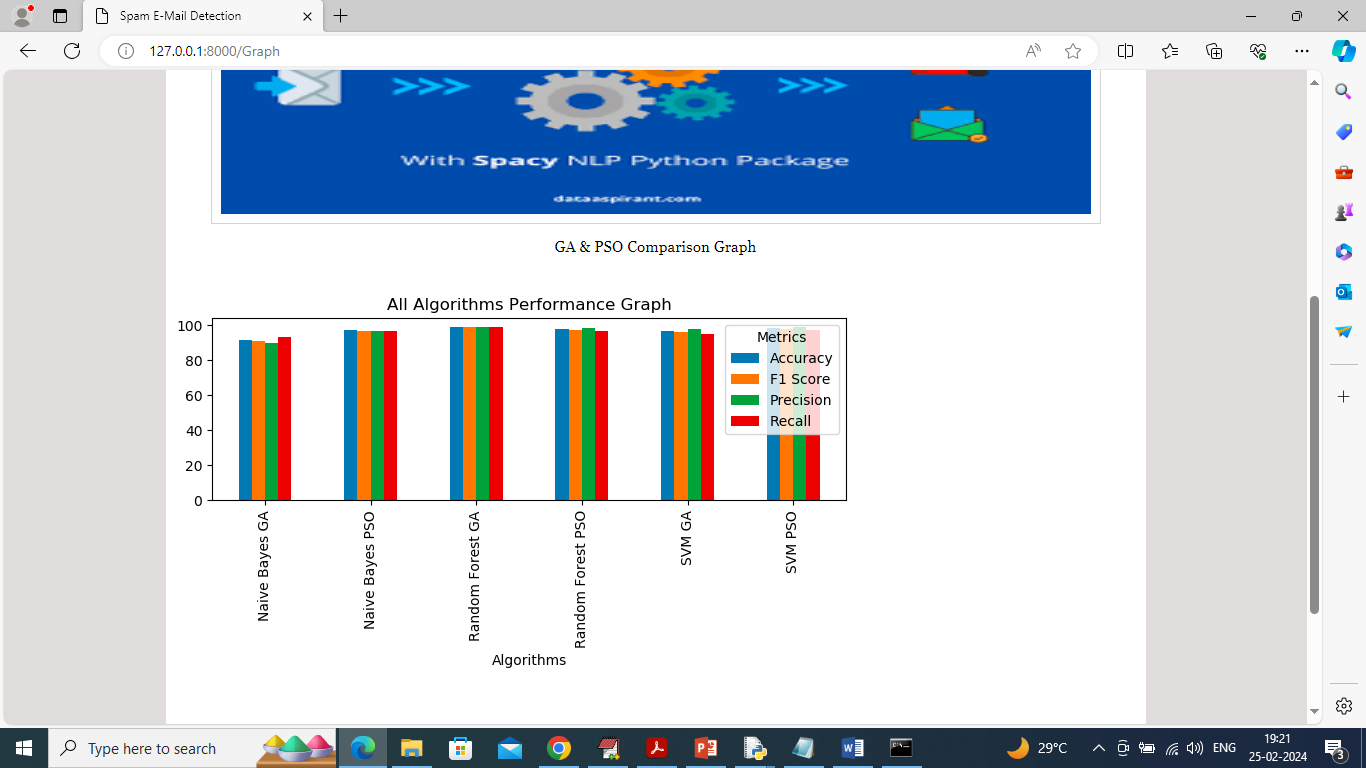
In above screen dataset loaded and now click on ‘Run ML with GA” to train ML with GA and get below page



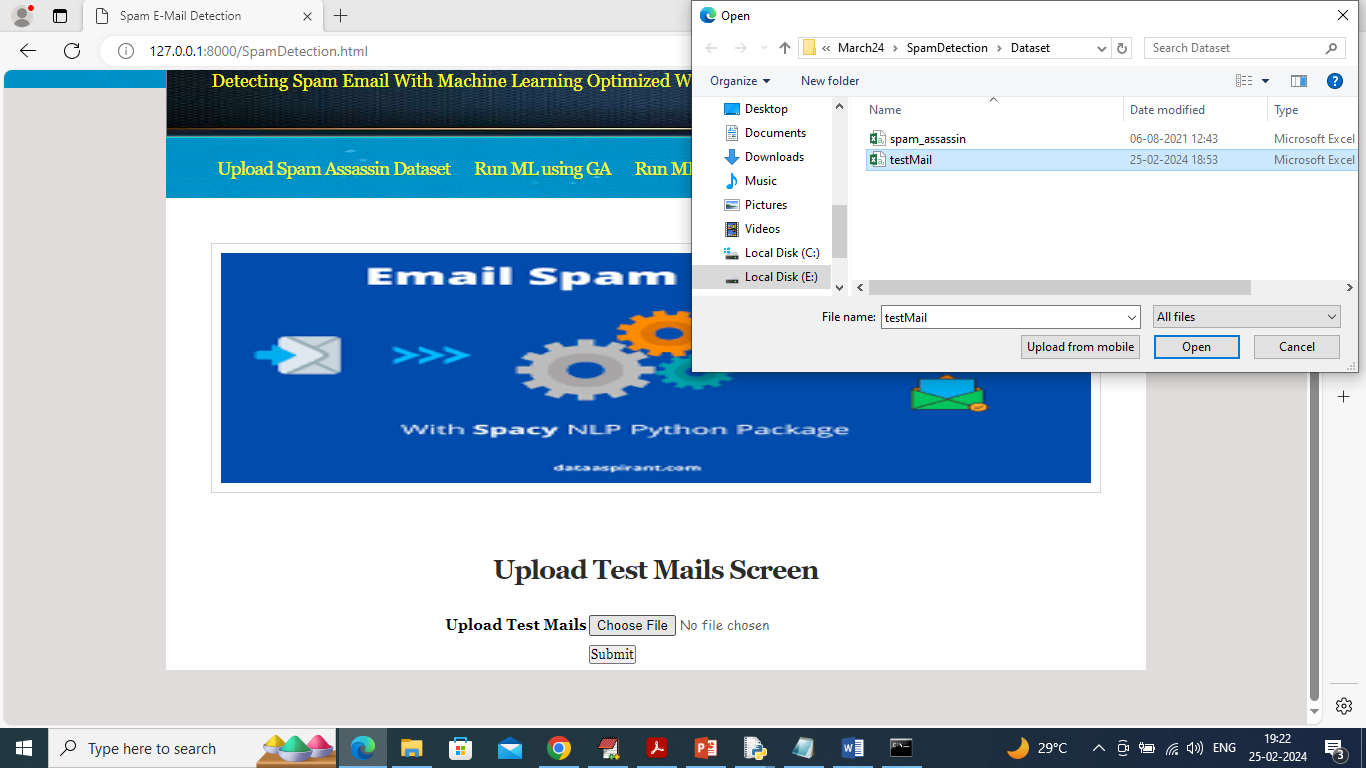
In above screen in blue colour text can see total features available in dataset and then GA selected 927 features out of 2000 and after training we can accuracy and other metrics from all 3 algorithms and now click on ‘Run ML with PSO’ link to train ML with PSO and get below page



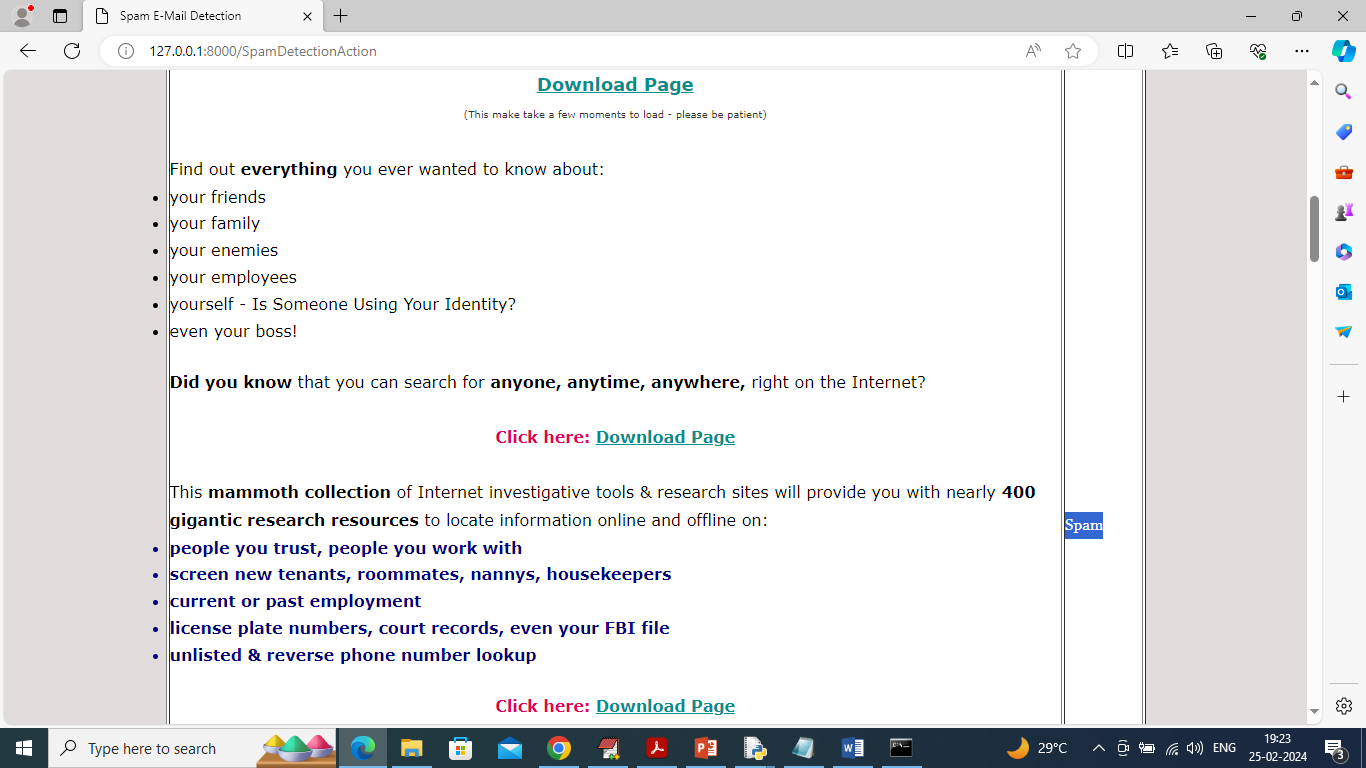
In above screen can see all ML algorithm performance on both GA and PSO and now click on ‘Graph’ link to get below page



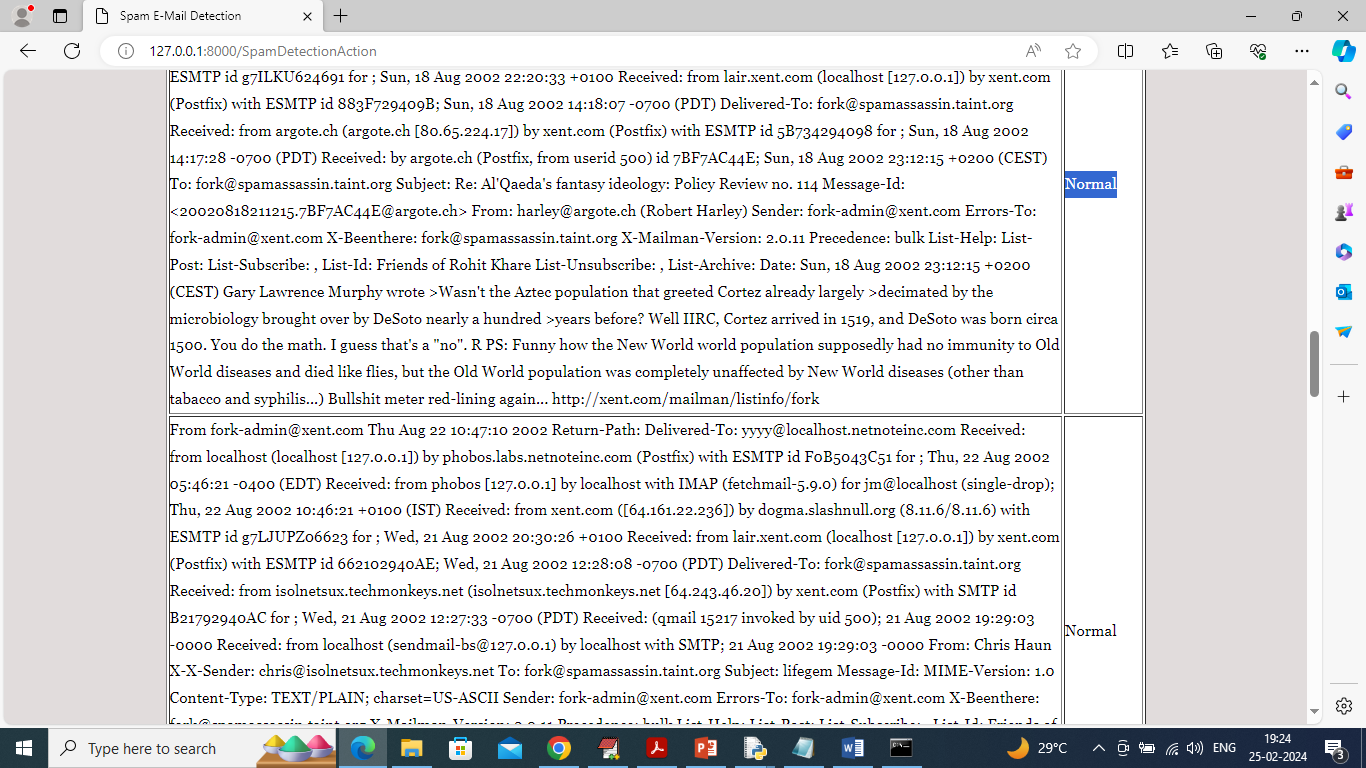
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and can see some algorithm work best with GA and some with PSO and now click on ‘Spam Detection’ link to get below page



In above screen select and upload ‘Test Mails’ file and then click on ‘Submit’ button to get below prediction



In above screen in first column can see ‘Email Text’ and in second column can see predicted label as ‘Spam’ or ‘Normal’.



In above screen can see emails without links consider as Normal.

Similarly by following above screens you can execute entire application