Software Vulnerability Detection Tool Using Machine Learning Algorithms

In this project you ask to develop Software Vulnerability tool to detect SQL Injection, Cross Site Scripting (LFI & XSS same) and RFI but you not specify which algorithms or technique to use. So we have used Ensemble Machine Learning algorithm which is combination of multiple algorithms such as SVM, KNN and Naïve Bayes.

Now-a-days Machine Learning algorithms are using everywhere from Medical disease prediction to road side traffic prediction as this algorithms prediction accuracy is more than 95%.

Above success of Machine Learning algorithms are migrating us to develop vulnerability detection tool using machine learning algorithms. Machine Learning algorithms get trained on past data and then can analyse new test data to predict it class of Normal or Vulnerability type.

In propose work we are using dataset to identify 3 different classes such as ‘No Vulnerability, SQL Injection, XSS or RFI.

NO Vulnerability refers to normal SQL statement or XSS.

SQL Injection refers to abnormal commands in SQL query where attacker will intercept query and then alter or inject abnormal command

Example: select \* from users where username=’abc’ OR 1=1;

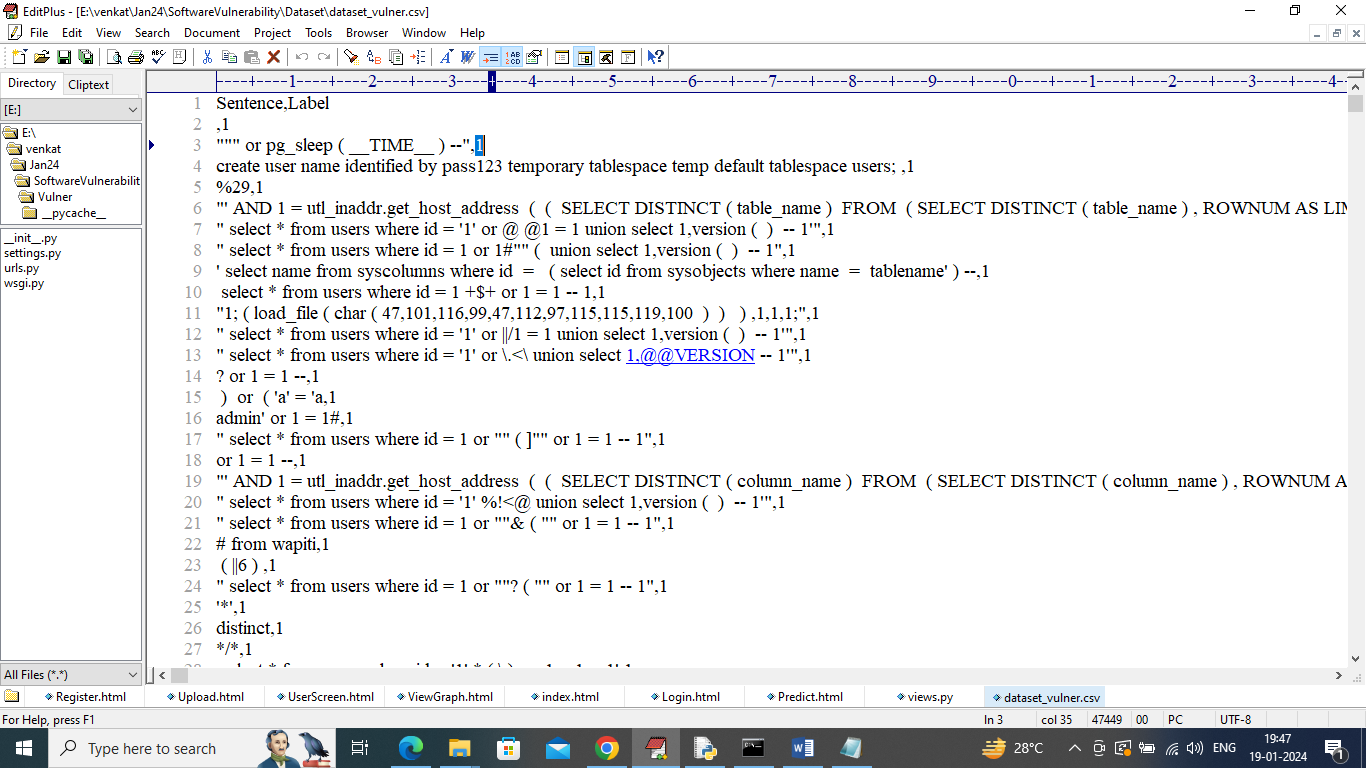
In above command ‘abc’ values will not exists in database but attacker inject ‘OR’ with 1=1 which is a true condition and attacker will get all details from database.

XSS or RFI: similarly like SQL injection attacker will update web code by altering or injecting XSS commands. XSS attacks allow attackers to inject malicious code or styles into a web page viewed by users. This can include client-side scripts.

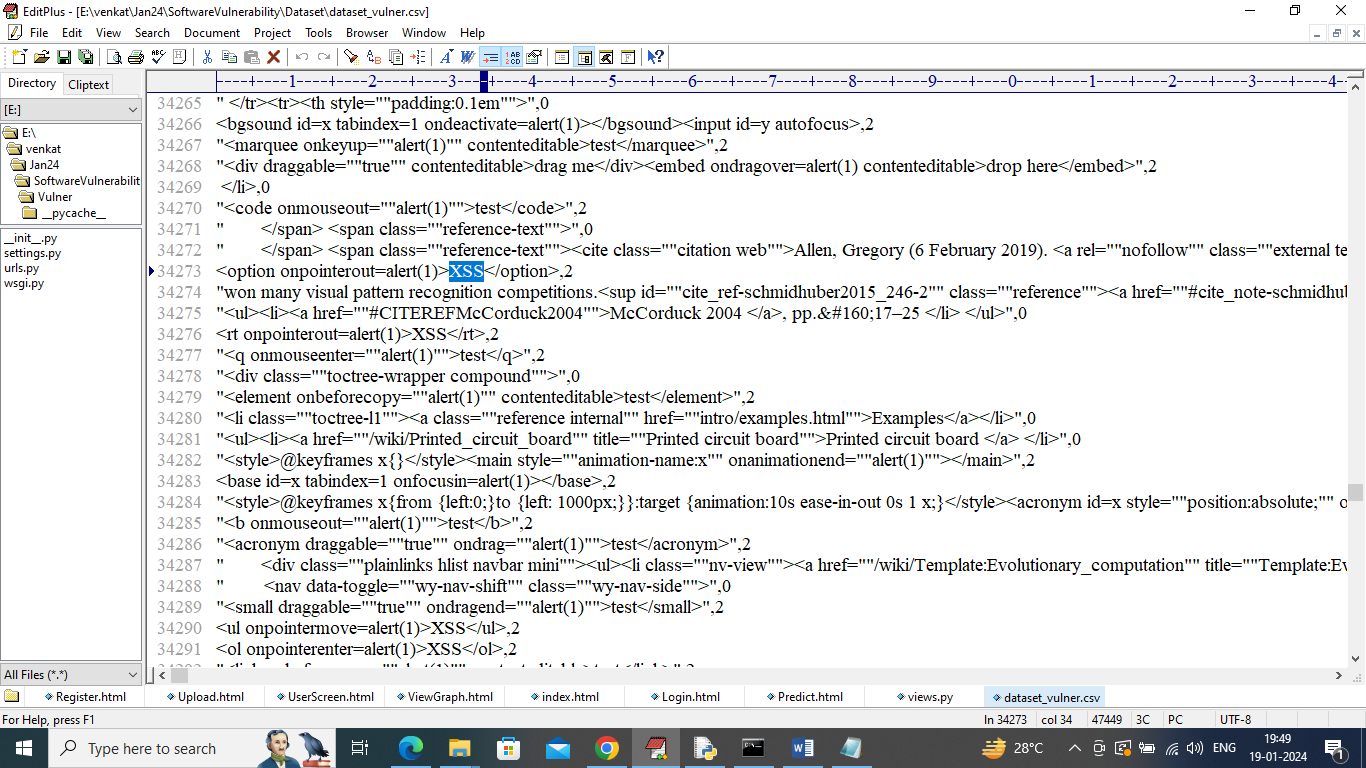
Remote File Inclusion (RFI) is a web vulnerability that allows attackers to include files from a remote location. This vulnerability is often found on websites that run PHP.

To detect all of the above Vulnerability we are using vulnerable dataset which contains all possible commands of SQL injection, XSS and RFI. Each command is associated with class label as ‘No Vulnerability, SQL injection, XSS or RFI.

In below screen showing dataset details

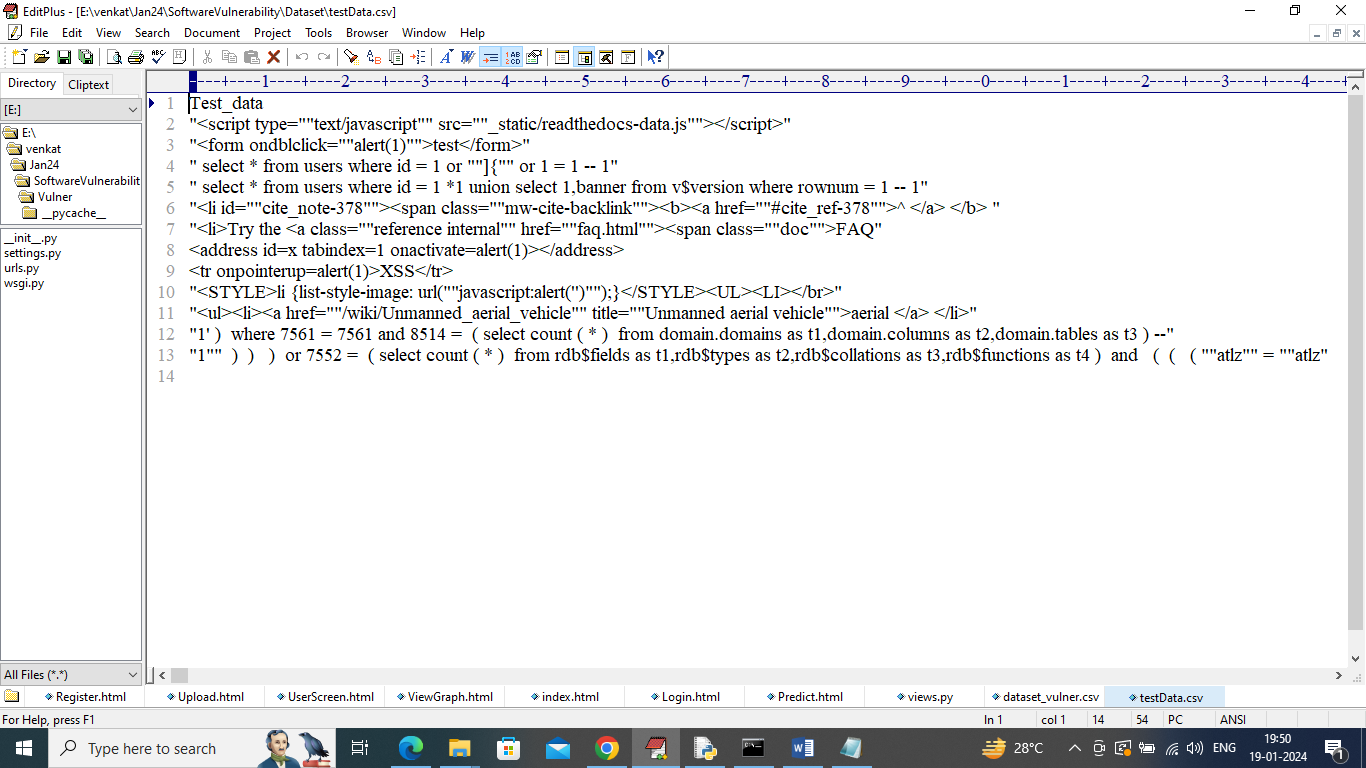


In above dataset screen showing all SQL queries with label as 1 or 0 where 1 means query contains injection and 0 means normal.



In above same dataset we have XSS and RFI coding of web pages and by using above dataset we will trained machine learning ensemble algorithms. Trained model can be applied on NEW Test Queries to detect vulnerability.

In above screen showing test queries



In above test dataset we have only queries but no classes or labels and after applying above queries on machine learning model will get predicted labels as ‘No Vulnerability or SQL or XSS’.

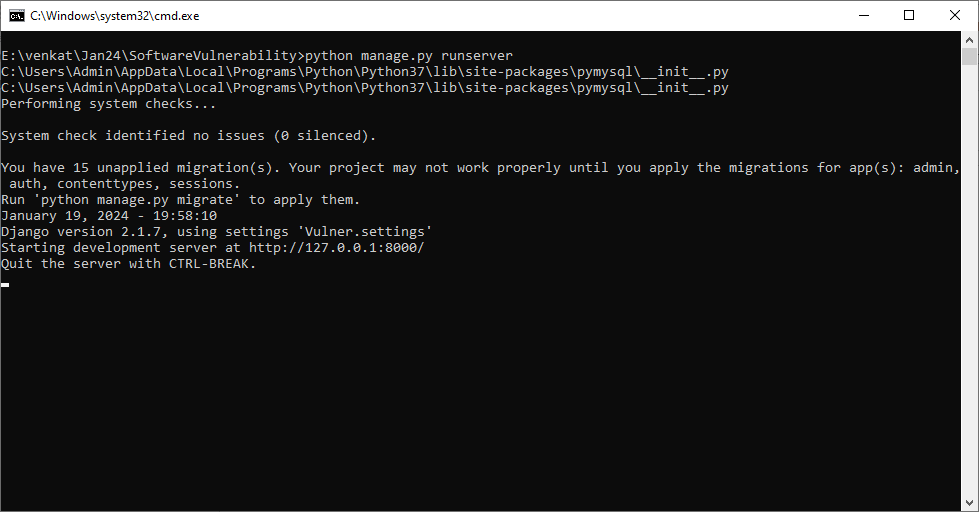
To implement this project we have designed following Modules

1. New User Register: new user can register with the application
2. User Login: after sign up user can login to application
3. Load Dataset: after login user can upload dataset to application and then extract all queries and labels from dataset and then from all queries will remove stop words like ‘and, the, or, what and many other words’. By removing stop words application will have core queries words. Dataset processing for core words will be happened using Natural Language processing toolkit
4. Run Ensemble Algorithms: processed dataset will be input to Ensemble Machine learning algorithm to train a model and this model will be applied on test data to calculate accuracy and other metrics
5. Confusion Matrix Graph: using this module we will plot confusion matrix graph of algorithm prediction capability
6. Predict Vulnerability: using this module will upload new TEST data query and then Machine learning algorithm will analyse all TEST data and predict type of vulnerability.

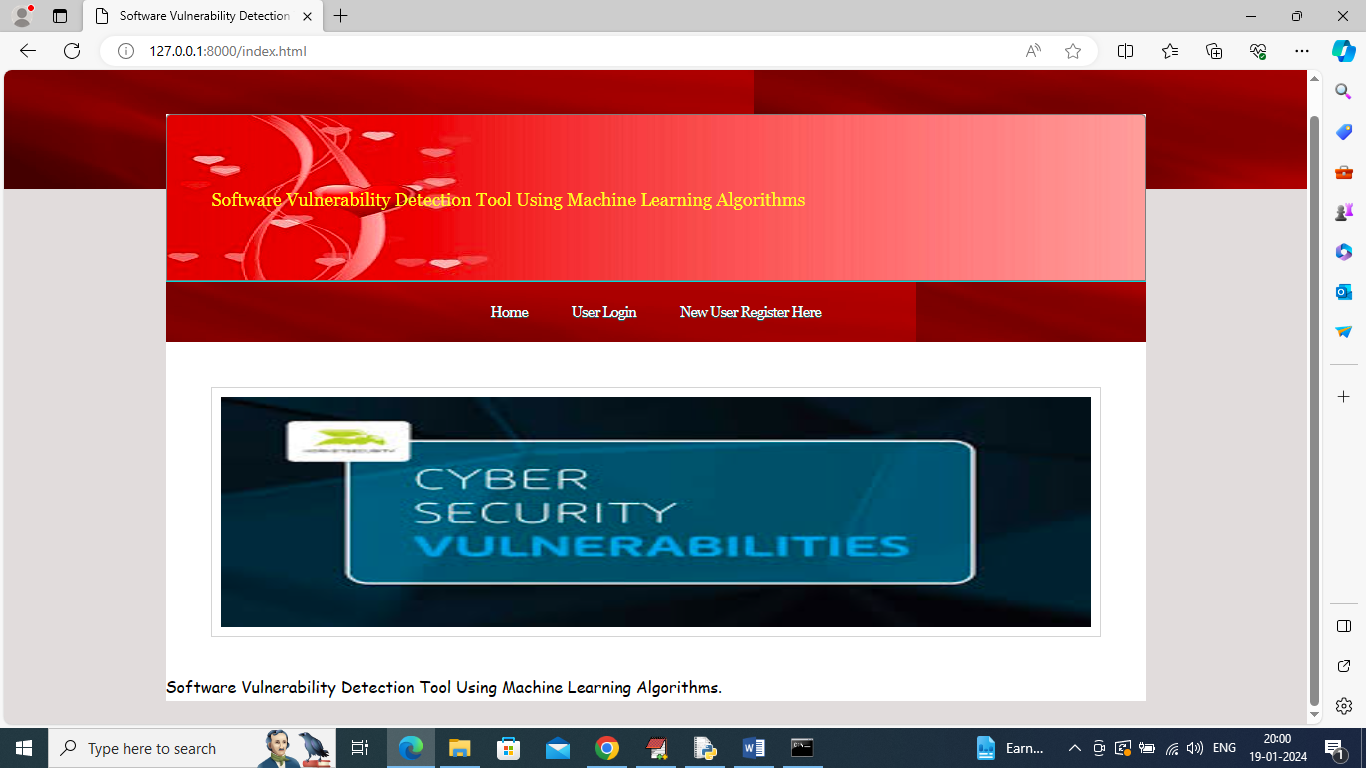
SCREEN SHOTS

To run project install python 3.7 and then install MYSQL database and then copy content from DB.txt file and paste in MYSQL to create database. Now double click on ‘installNLTK.bat’ file to download NLTK and once click then window will appear in that window click on “Download’ button to download all packages and once downloaded then window will turn to green colour and then close the window

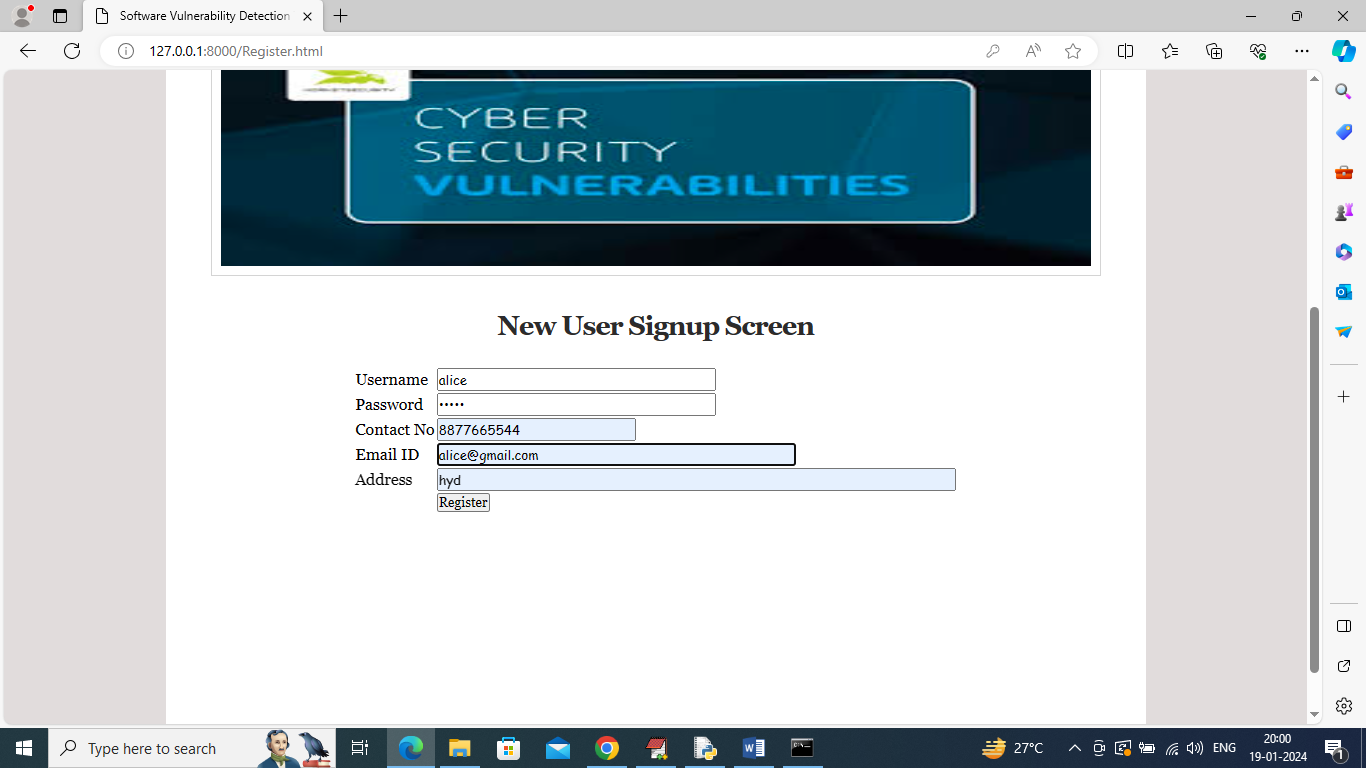
Now double click on ‘run.bat’ file to start python DJANGO web server and get below screen



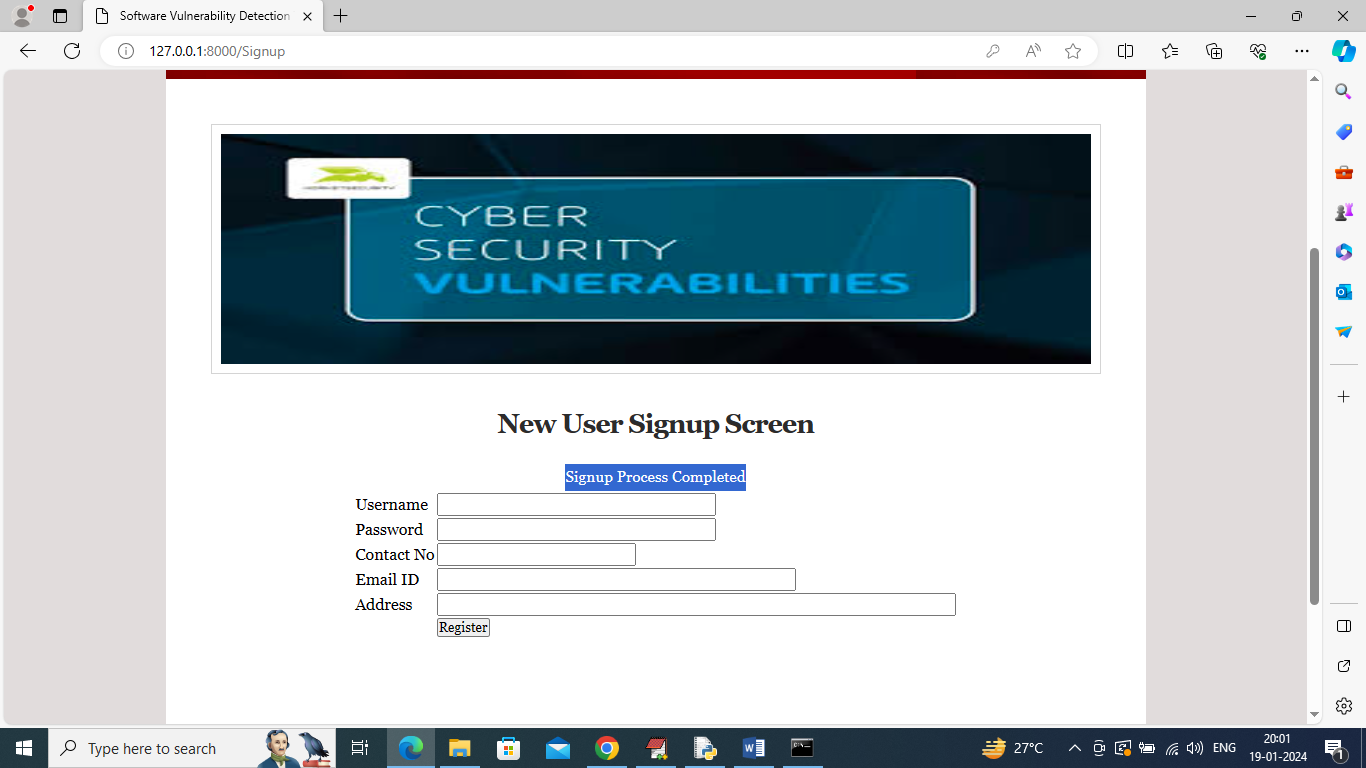
In above screen python web server started and now open browser and enter URL as <http://127.0.0.1:8000/index.html> and then press enter key to get below page



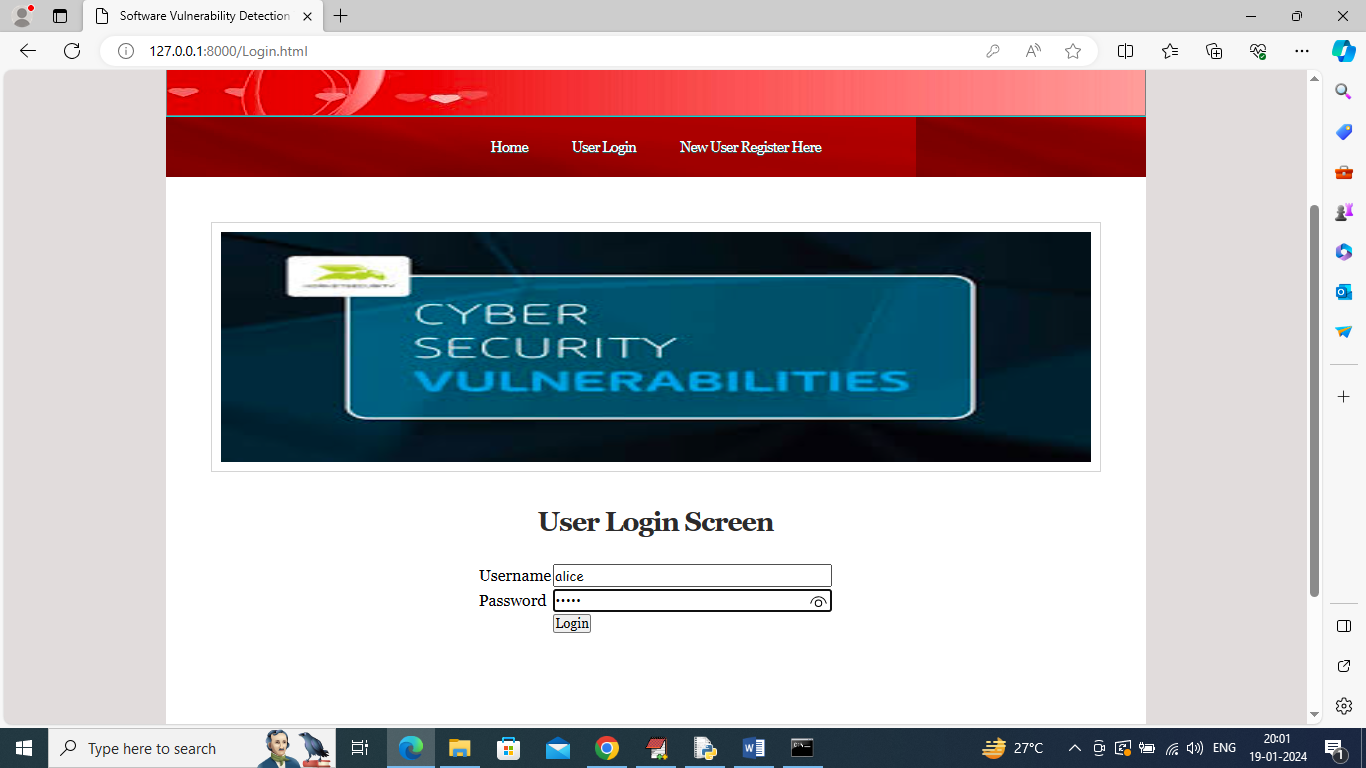
In above screen click on ‘New User Register Here’ link to get below sign up page



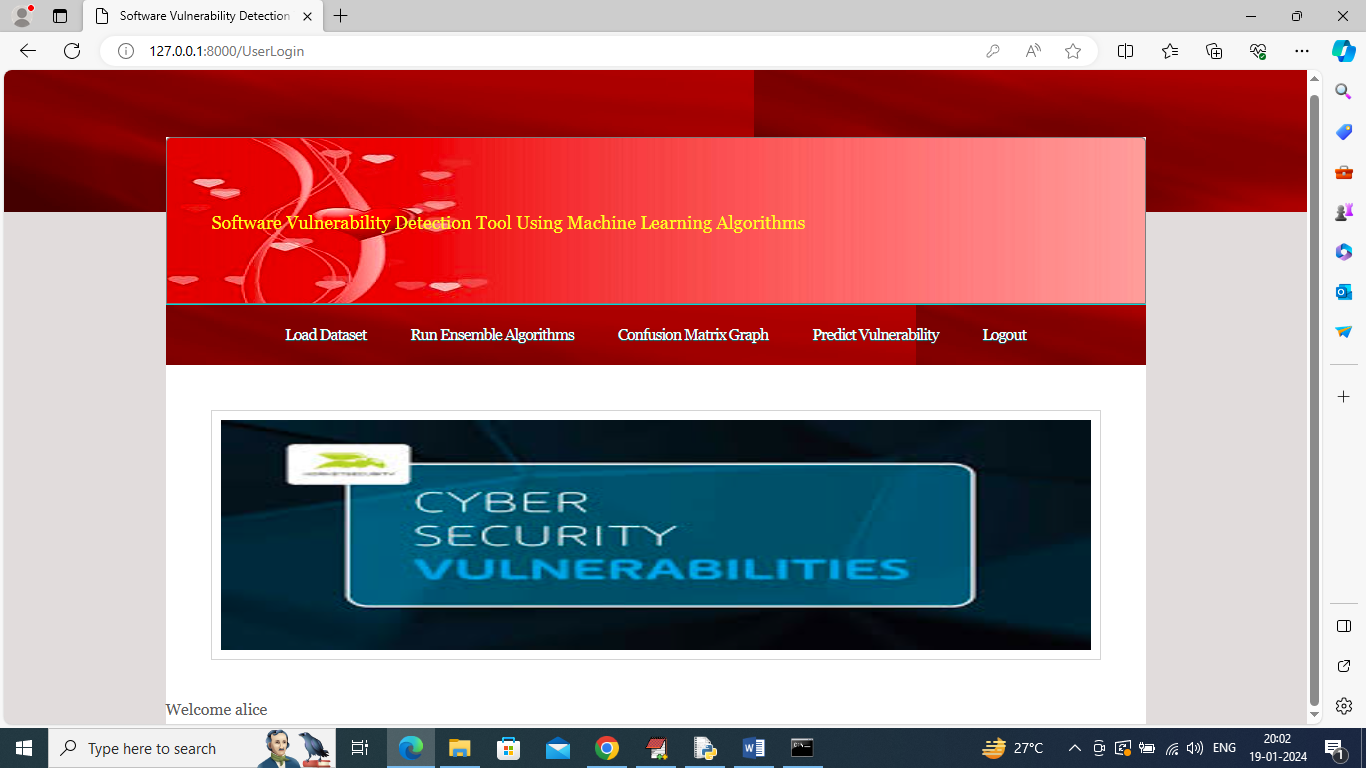
In above screen user is entering sign up details and then press button to get below page



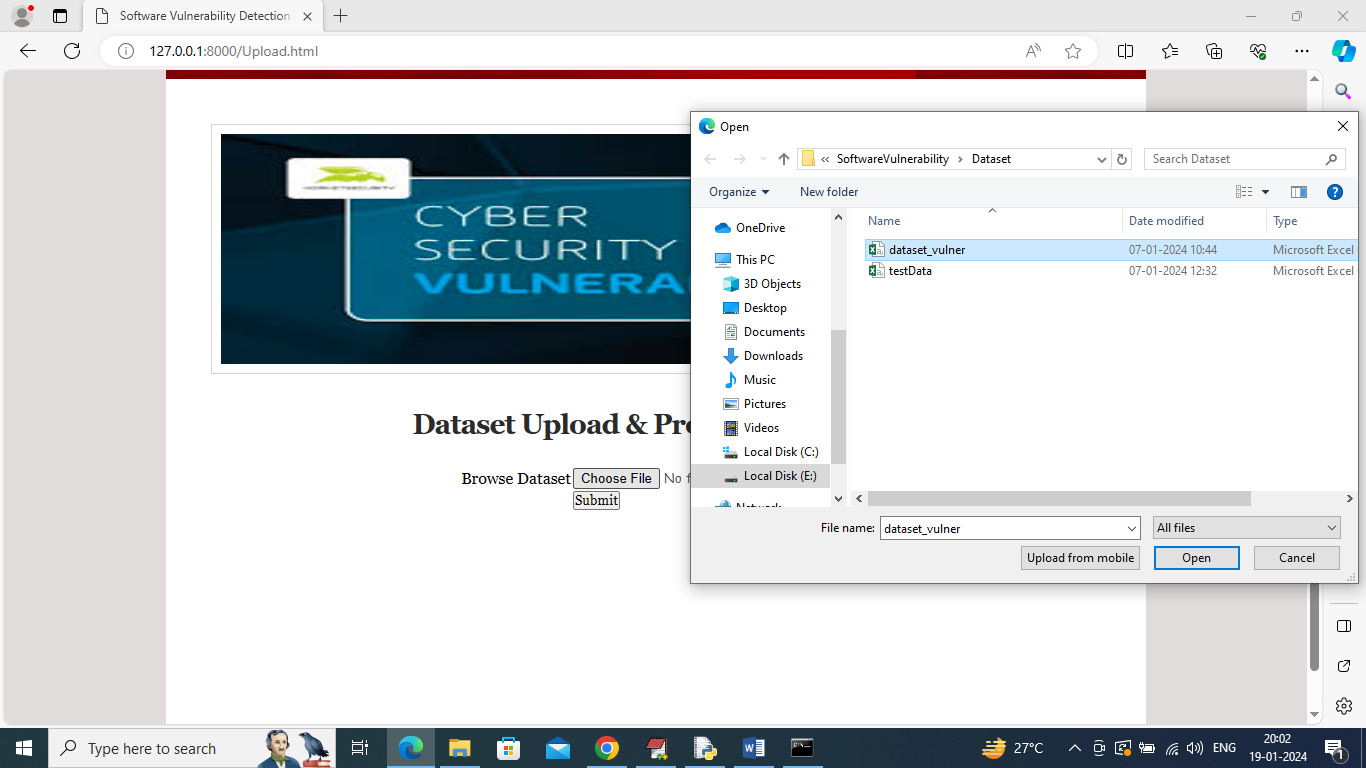
In above screen user sign up completed and now click on ‘User Login’ link to get below page



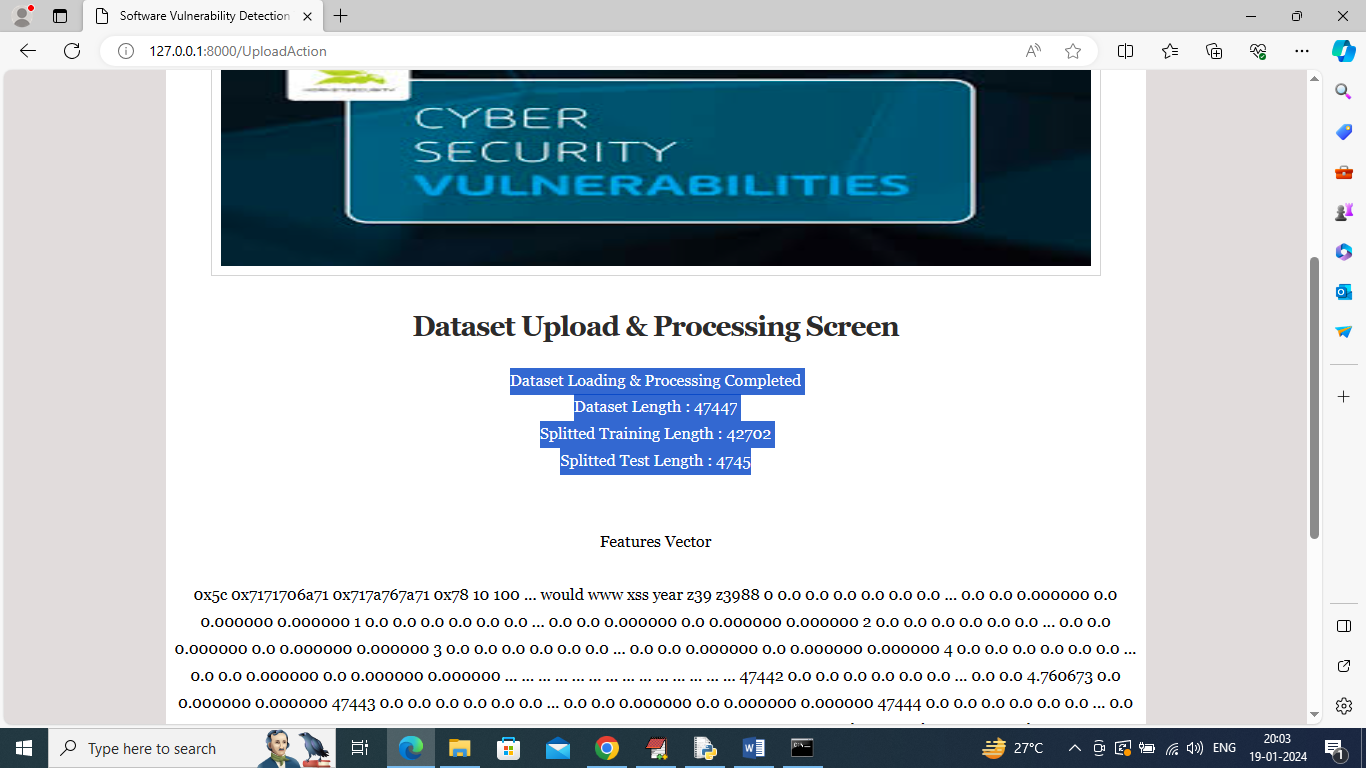
In above screen user is login and after login will get below page



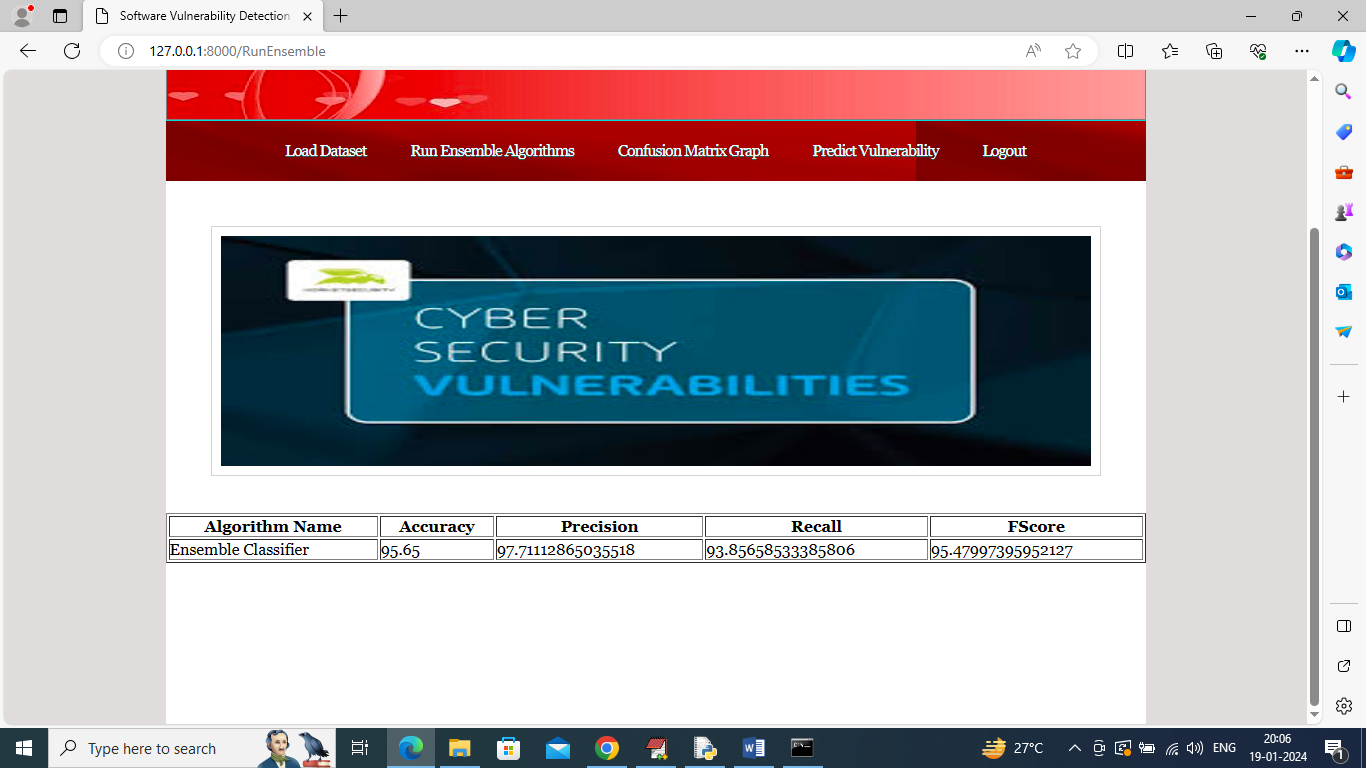
In above screen click on ‘Load Dataset’ link to get below page



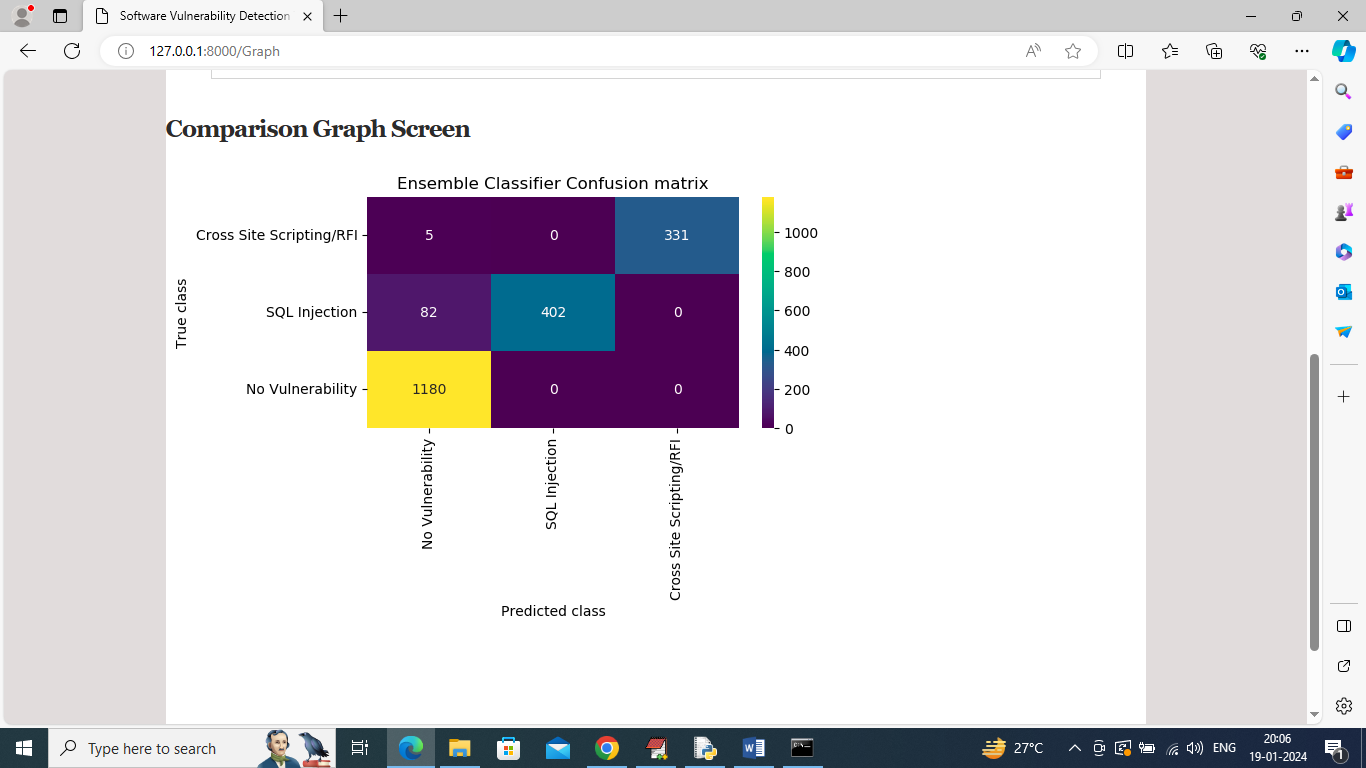
In above screen select and upload ‘dataset\_vulner.csv’ file and then click on ‘Open’ and ‘Submit’ button to load dataset and then will get below output



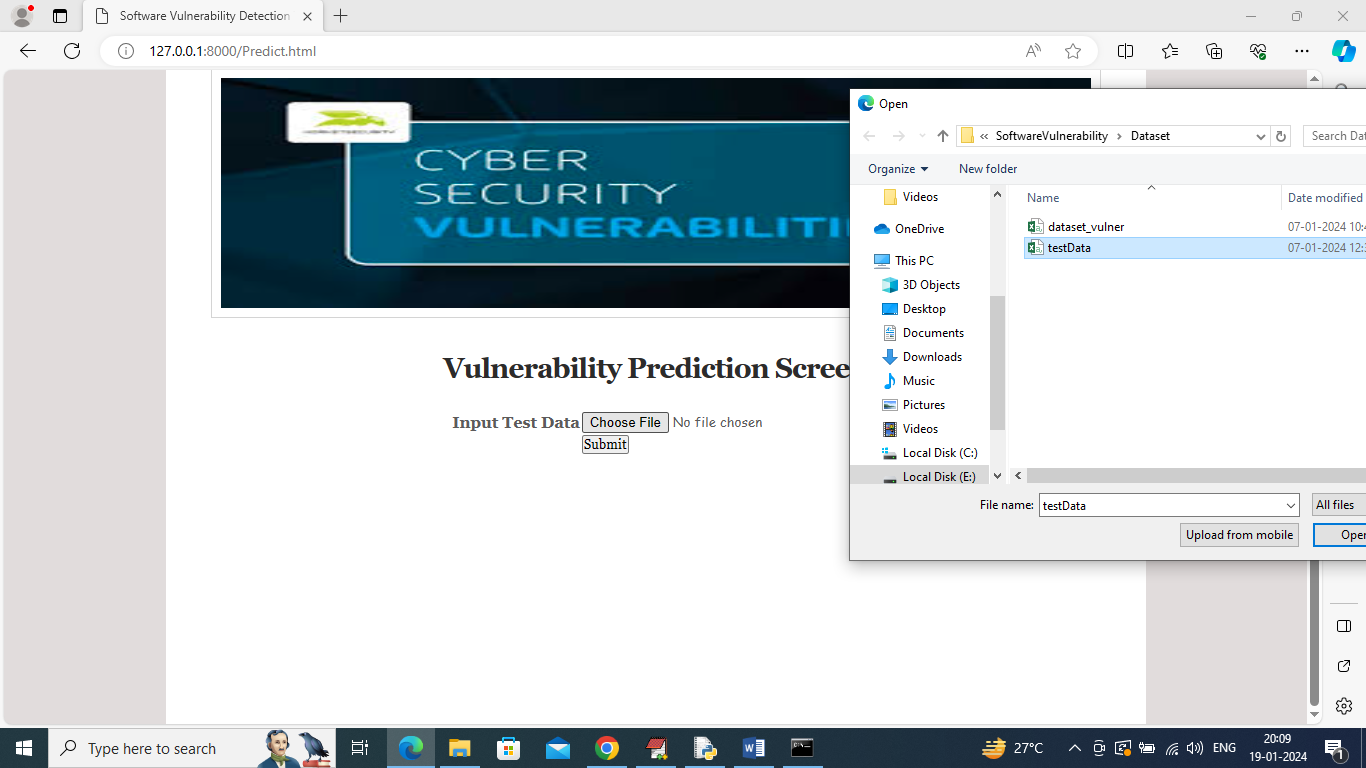
In above screen can see dataset loaded and can see total number of records available in dataset and then can see training number of records on which Machine Learning algorithm get trained and then can see number of test records on which ML will perform prediction to calculate its prediction accuracy %. Now click on ‘Run Ensemble Algorithms’ link to train ensemble algorithm and then will get below output



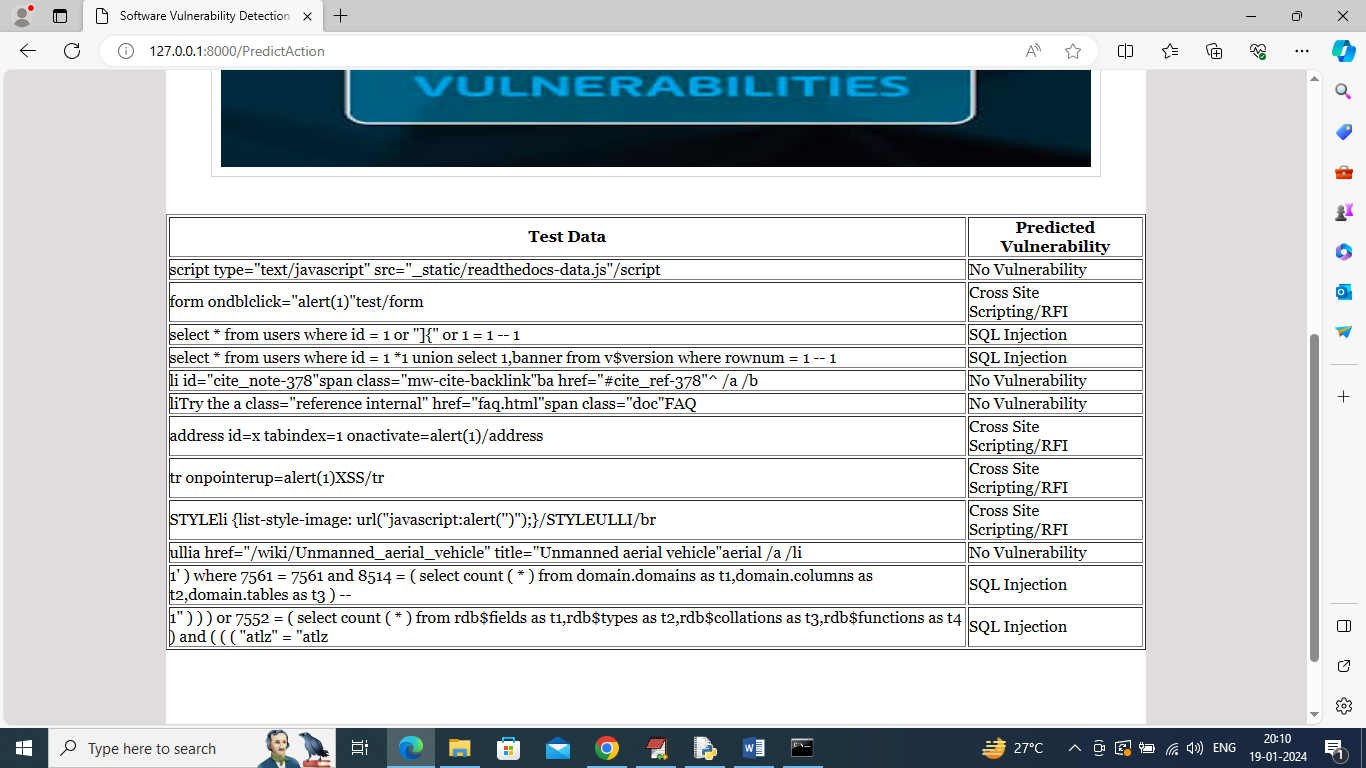
In above screen Ensemble Machine Learning algorithm training completed and can see its prediction accuracy as 95% and can see other metrics like precision, recall and FCSORE. Now click on ‘Confusion Matrix Graph’ link to view visually how many records ensemble predicted correctly and incorrectly



In above graph x-axis represents Predicted Labels and y-axis represents True Labels and then all different colour boxes in diagnol represents correct prediction count and remaining all blue boxes represents incorrect prediction count which are very few. Now click on ‘Predict Vulnerability’ link to upload test data and predict Vulnerability



In above screen selecting and uploading ‘testData.csv’ file which contains SQL, XSS and RFI coding commands and then click on ‘Submit’ button to get below output



In above table in first column can see SQL queries, XSS and RFI coding commands and in second column can see predicted vulnerability.

So by using above tool you can easily detect all vulnerability and you can add NEW test command in ‘testData.csv’ file which is available inside ‘Dataset’ folder