PAACDA: Comprehensive Data Corruption Detection Algorithm

Now-a-days Data is not just values as this can be used to identify hidden information from many virtual areas such as detecting attacks from online application, customer purchase behaviour in online application, data sharing or friend’s community in graph based social network and many more areas. Sometime because of system error or malicious attackers this data will be corrupted and while analysis data engineers may get wrong result.

Data corruption or outliers refers to abnormal values which may fall between normal data.

To detect corrupted or outliers many machine learning based algorithms are introduced such Local Outlier Factor (LOF), One class SVM, KMEANS or DBSCAN clustering, Isolation Forest and many more. All this algorithms detection accuracy is satisfactory but not up to the mark.

To overcome from above issued author of this paper introducing novel algorithm called PAACDA (Proximity based Adamic Adar Corruption Detection Algorithm) which is based on ADAMIC ADAR index. Adamic is graph based algorithm which evaluate local similarity measurement that predicts links in a social network. If similarity is too high then index will be consider as corrupted otherwise normal. By using Adamic ADAR PACCDA will calculate accuracy of each value by using below functions

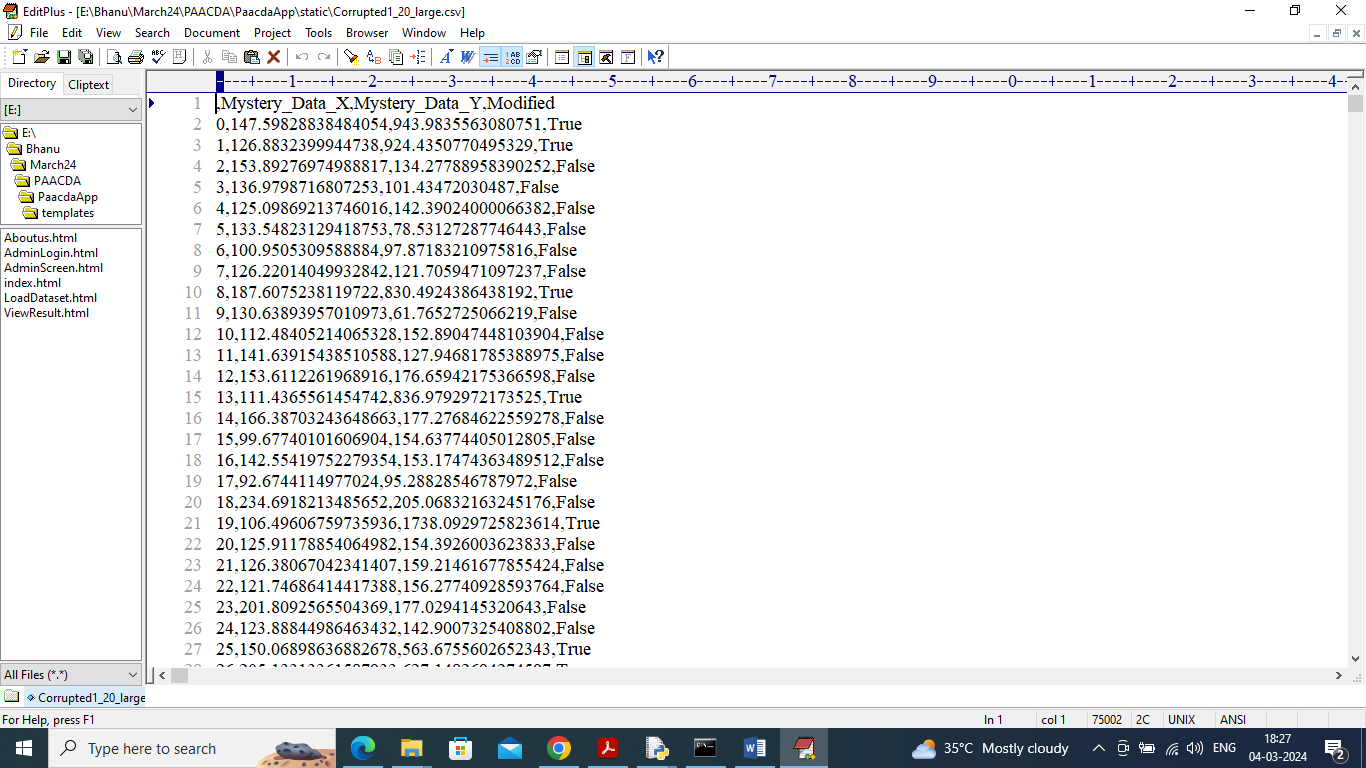
1. The mean is calculated for the column being analysed.
2. The range is set as mean/4.
3. Each data instance is iterated where x is each data instance.
4. If the instance is missing then the PAACDA Index value is set to infinity.
5. The PAACDA Index is compared amongst each other and the set of corrupted values is determined.
6. The accuracy metric is obtained by comparing predicted corrupted value with dataset true values.

In propose paper author has used many algorithms but implementing all this algorithms is quite difficult so we have implemented LOF, Isolation Forest, One Class SVM and Propose PAACDA algorithm.

To train and test each algorithm author has publish ‘Corrupted1\_20\_large.csv’ file which contains some normal and corrupted values and will show this whiles as output during execution.

Each algorithm performance is evaluated using accuracy, precision, recall and FSCORE.

In below screen showing dataset details



In above dataset screen first row contains dataset column names and remaining rows contains values. In last column we have valued called ‘True’ which mean row has corrupted values and False means ‘Normal’ values. This true and false will compare with predicted values to calculate accuracy and other metrics.

Extension Concept

In propose paper author has trained algorithm to detect corrupted values but not given any trained model to predict corrupted data from future values so we designed Hybrid PAACDA algorithm as extension which will trained Random Forest by using features from PAACDA algorithm and this Random Forest can predict corrupted data from any future dataset. Random Forest after training with PAACDA features giving 99 to 100% accuracy.

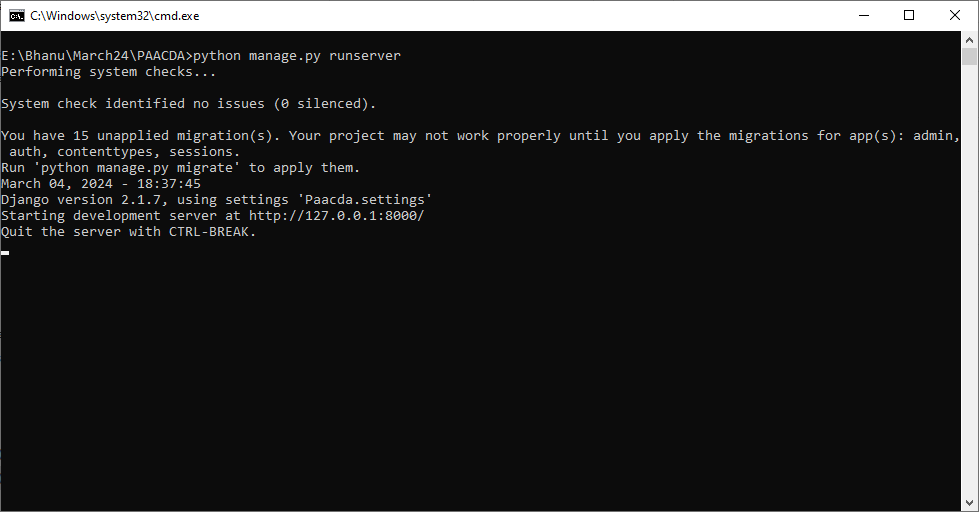
SCREEN SHOTS

We have implemented this project using DJANGO WEB framework which will contains following modules

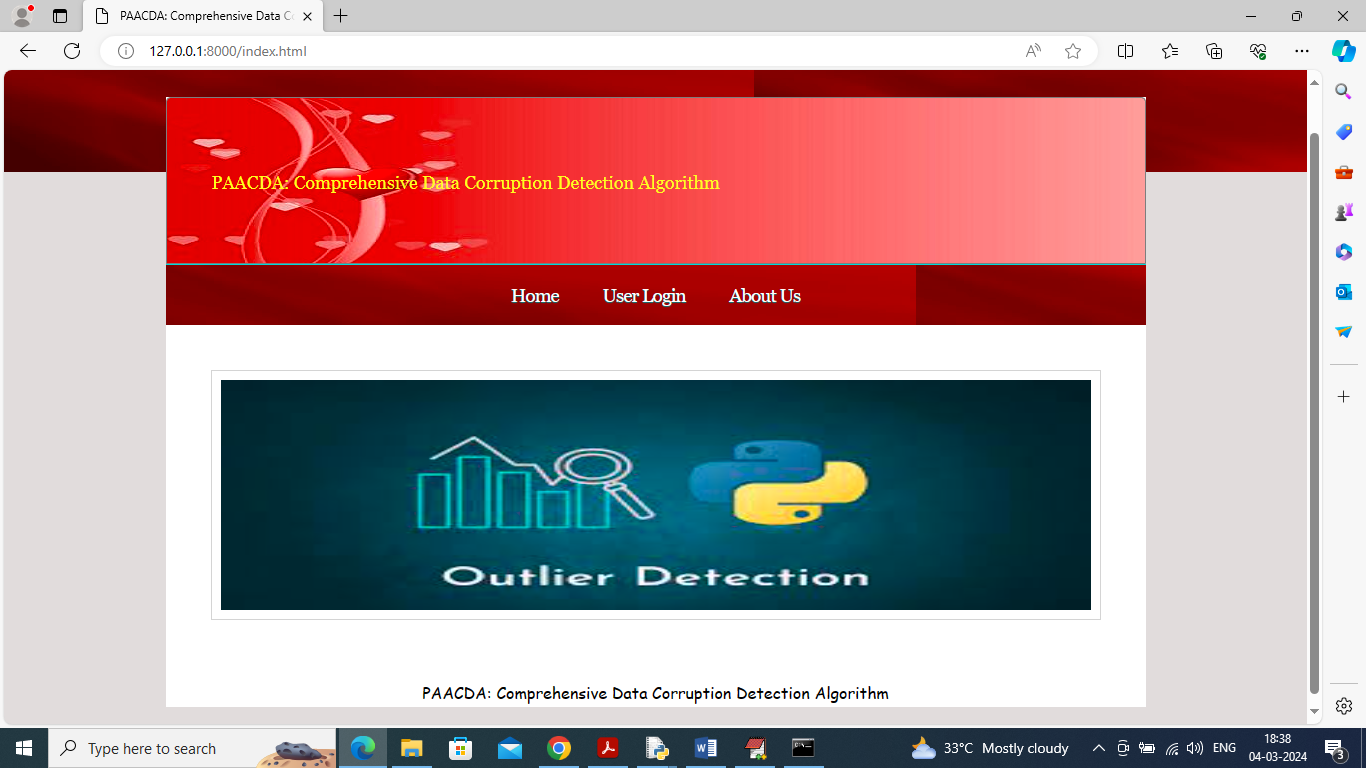
1. User Login: user can login to system using username and password as ‘admin and admin’
2. Load dataset: after login user will upload and process dataset
3. Run LOF: loaded dataset get train with ‘Local Outlier Factor’ algorithm to detect corrupted values and then detected corrupted data will be compare with true values to calculate accuracy
4. Run Isolation Forest: loaded dataset get train with ‘Isolation Forest’ algorithm to detect corrupted values and then detected corrupted data will be compare with true values to calculate accuracy
5. Run One Class SVM: loaded dataset get train with ‘OCS’ algorithm to detect corrupted values and then detected corrupted data will be compare with true values to calculate accuracy
6. Run PAACDA: loaded dataset get train with ‘PACCDA’ algorithm to detect corrupted values and then detected corrupted data will be compare with true values to calculate accuracy
7. Run Extension Hybrid PAACDA: loaded dataset get train with ‘PACCDA and Random Forest’ algorithm to detect corrupted values and then detected corrupted data will be compare with true values to calculate accuracy

SCREEN SHOTS

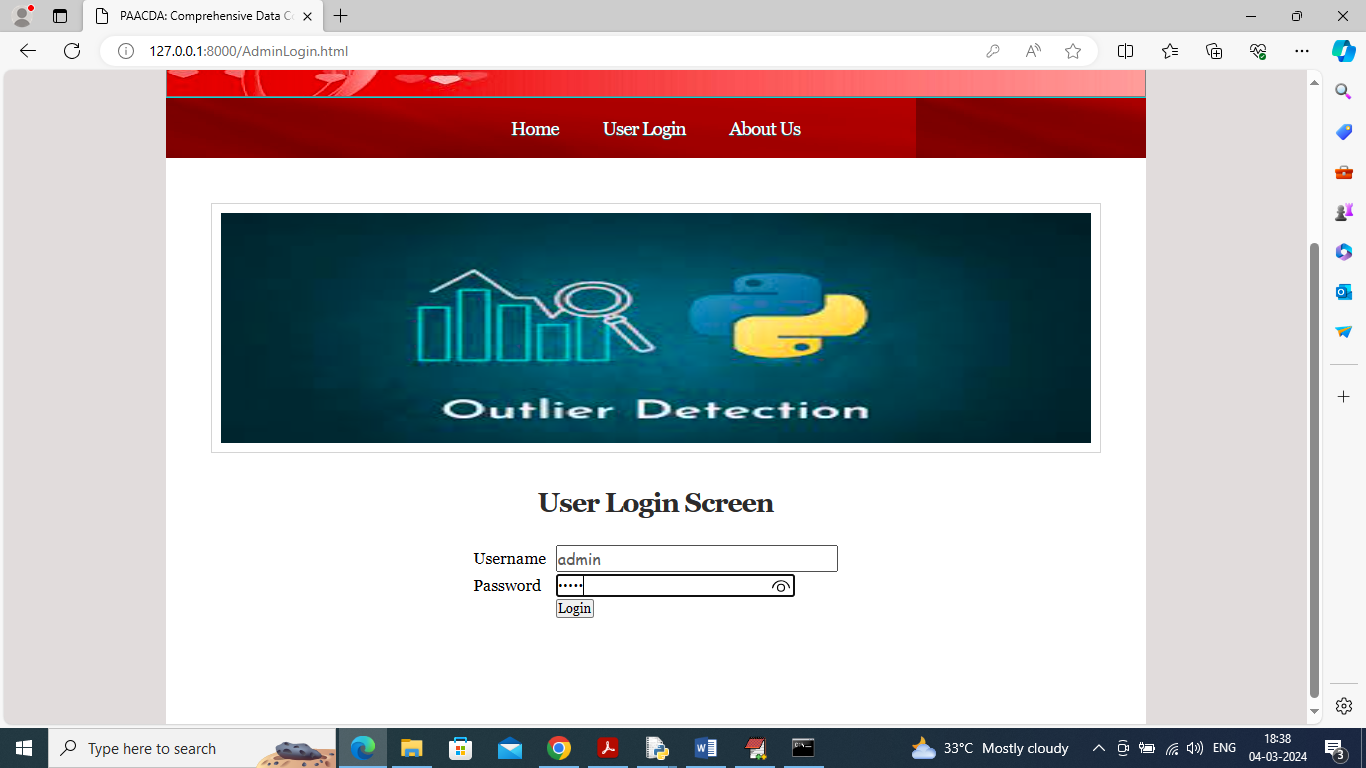
To run project double click on run.bat file to start python server and get below page



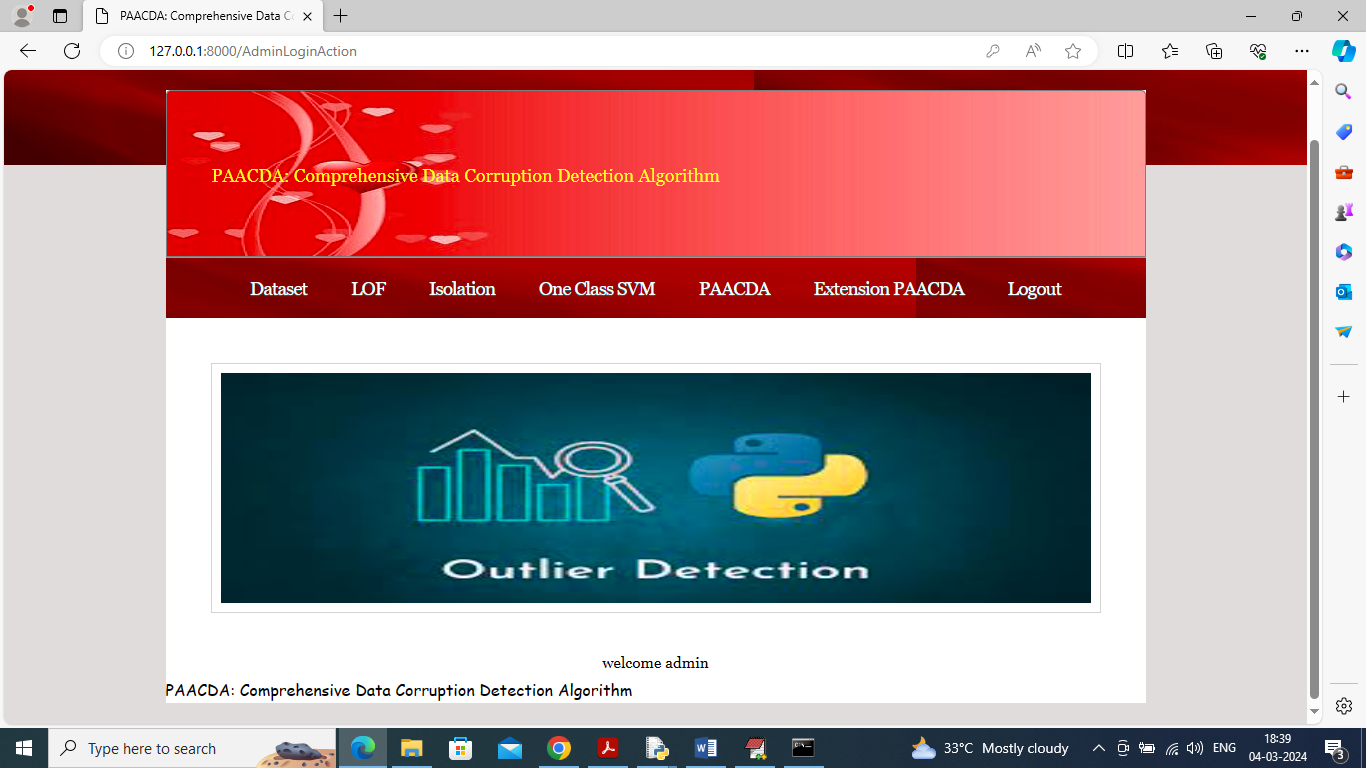
In above screen web server started and now open browser and enter URL as <http://127.0.0.1:8000/index.html> and then press enter key to get below page



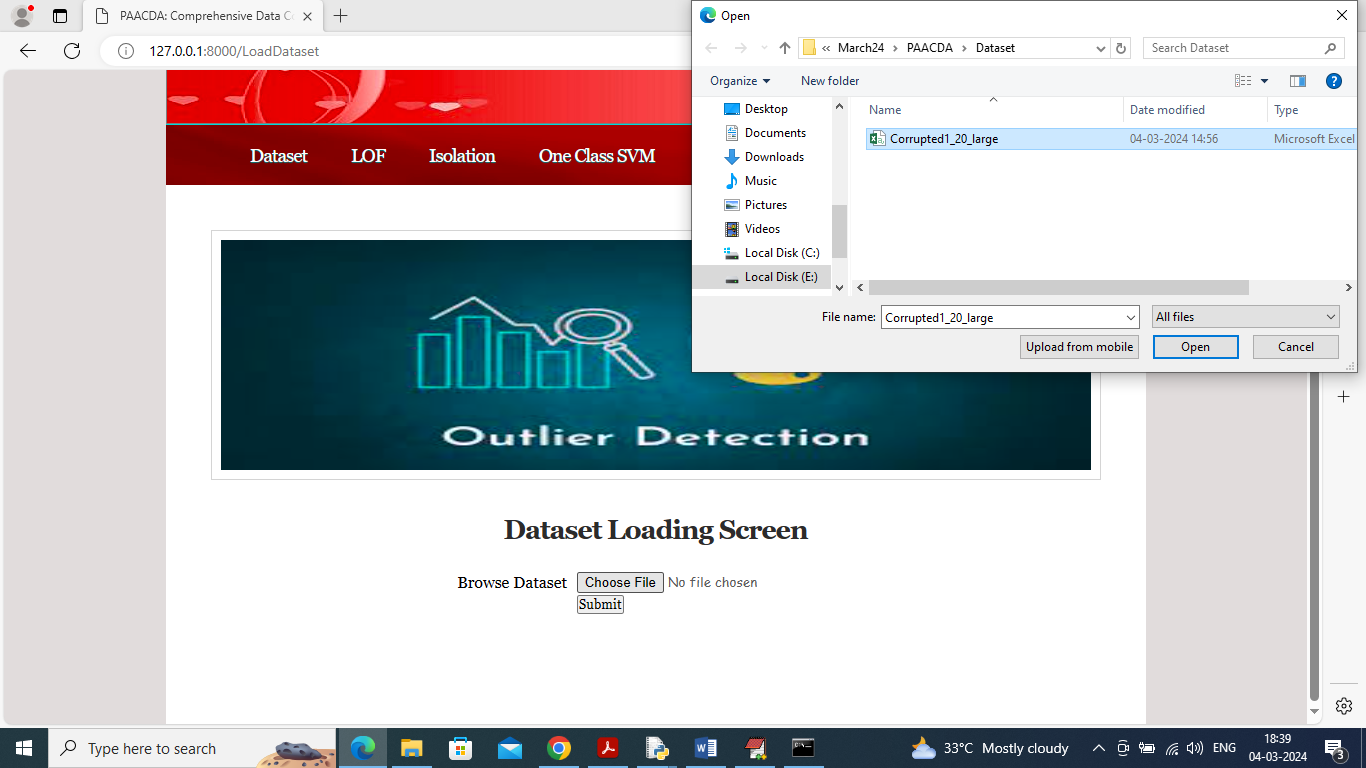
In above screen click on ‘User Login’ link to get below page



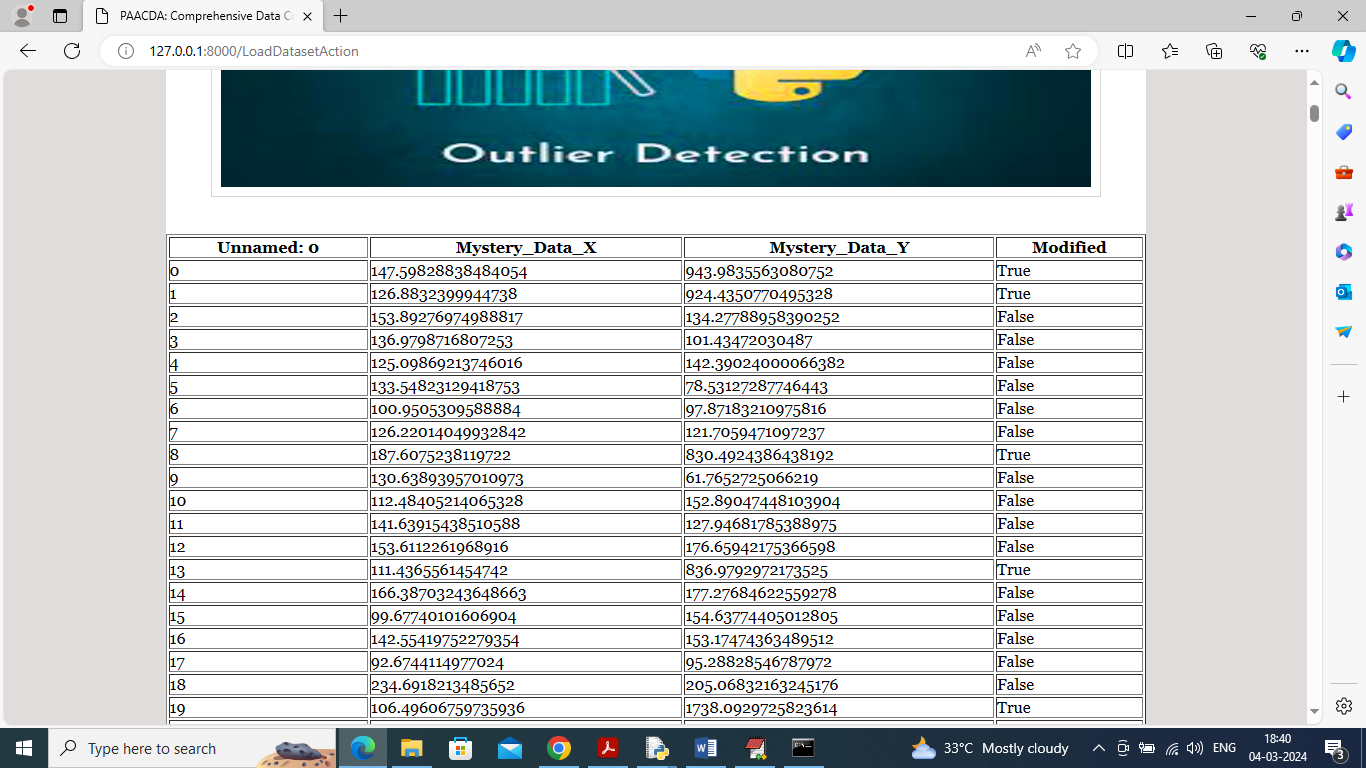
In above screen admin is login and after login will get below page



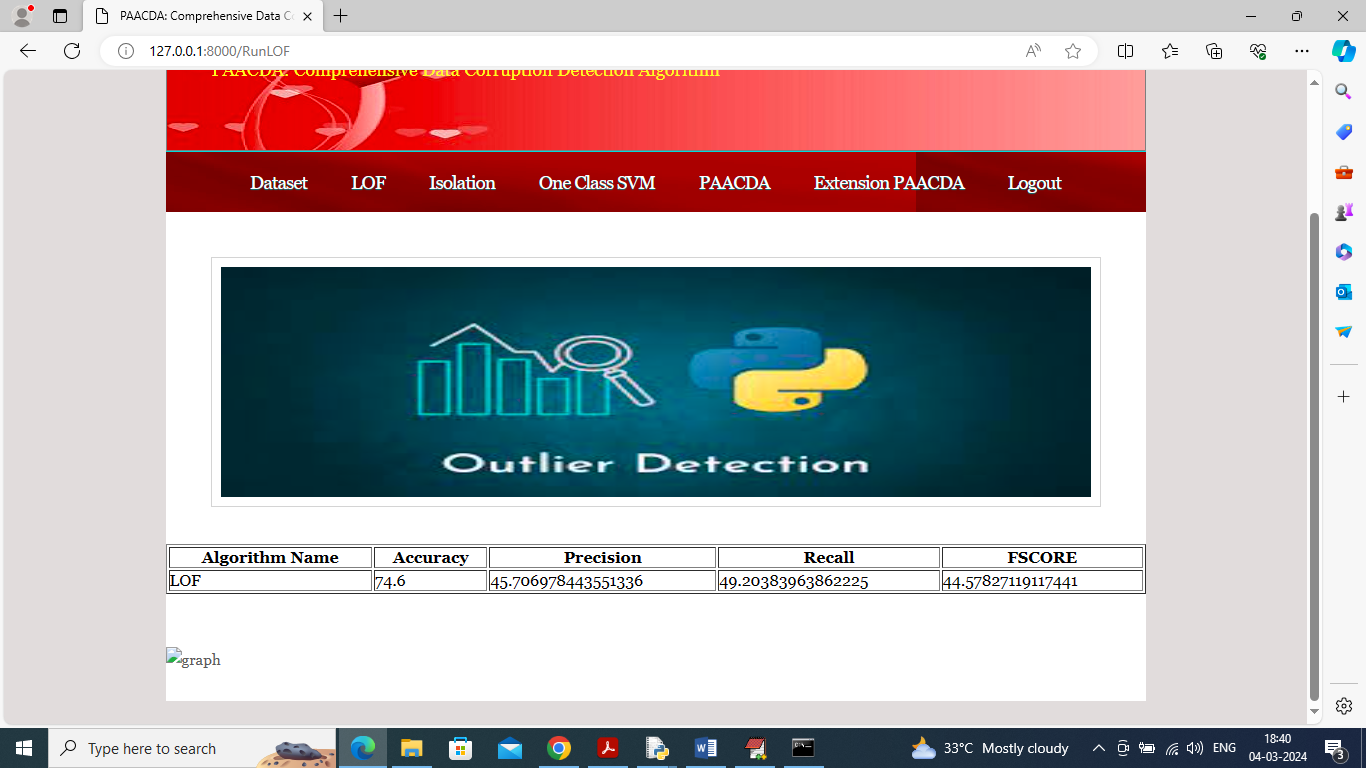
In above screen click on ‘Dataset’ link to get below page



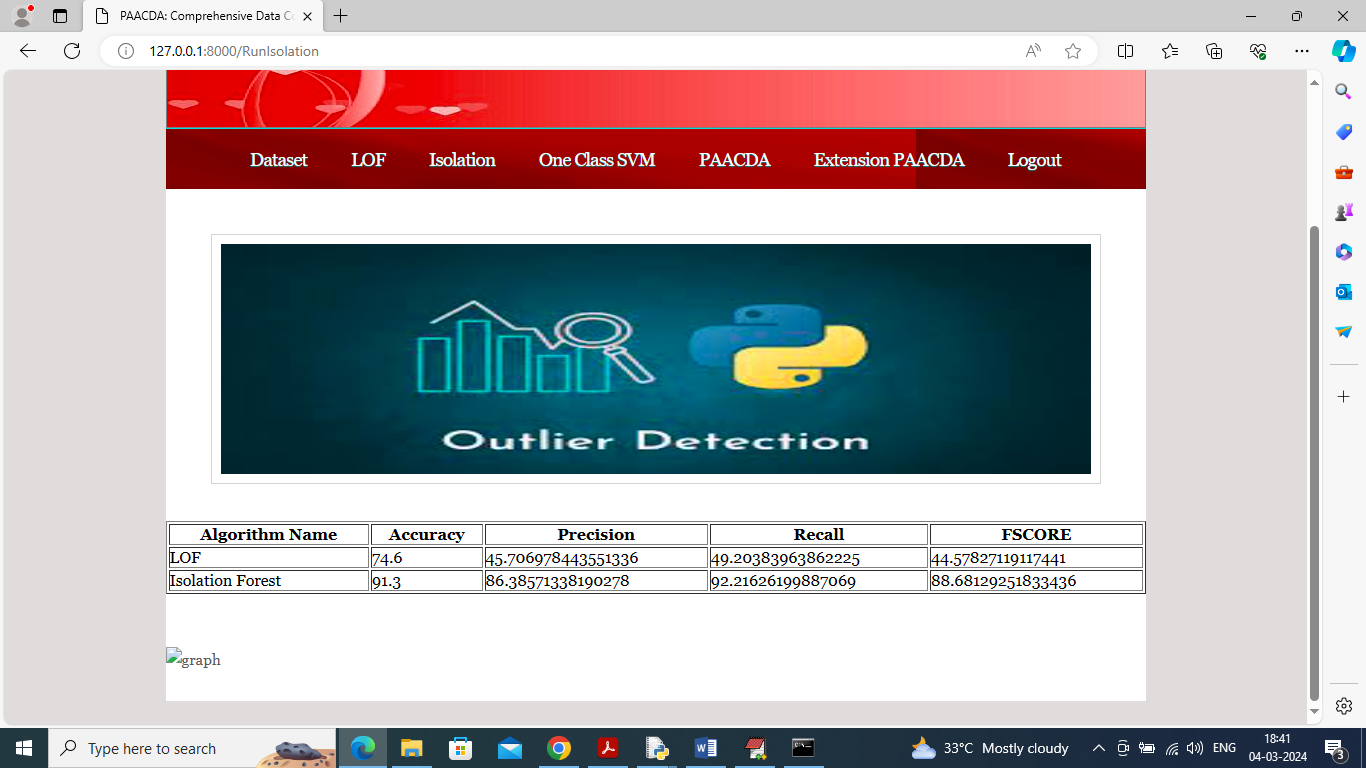
In above screen selecting and uploading ‘Dataset’ file and then click on ‘Open’ button to get below page



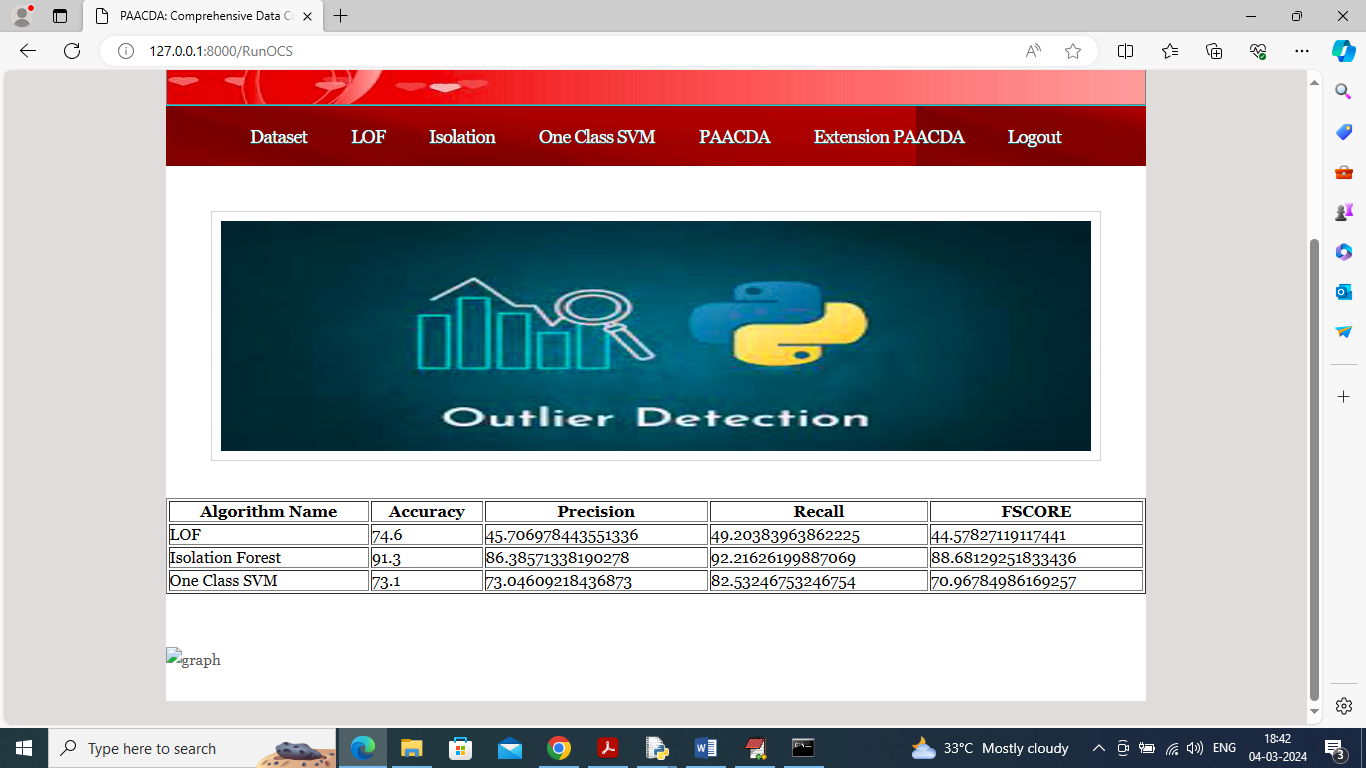
In above screen dataset loaded and now click on ‘Run LOF’ link to train LOF algorithm and get below page



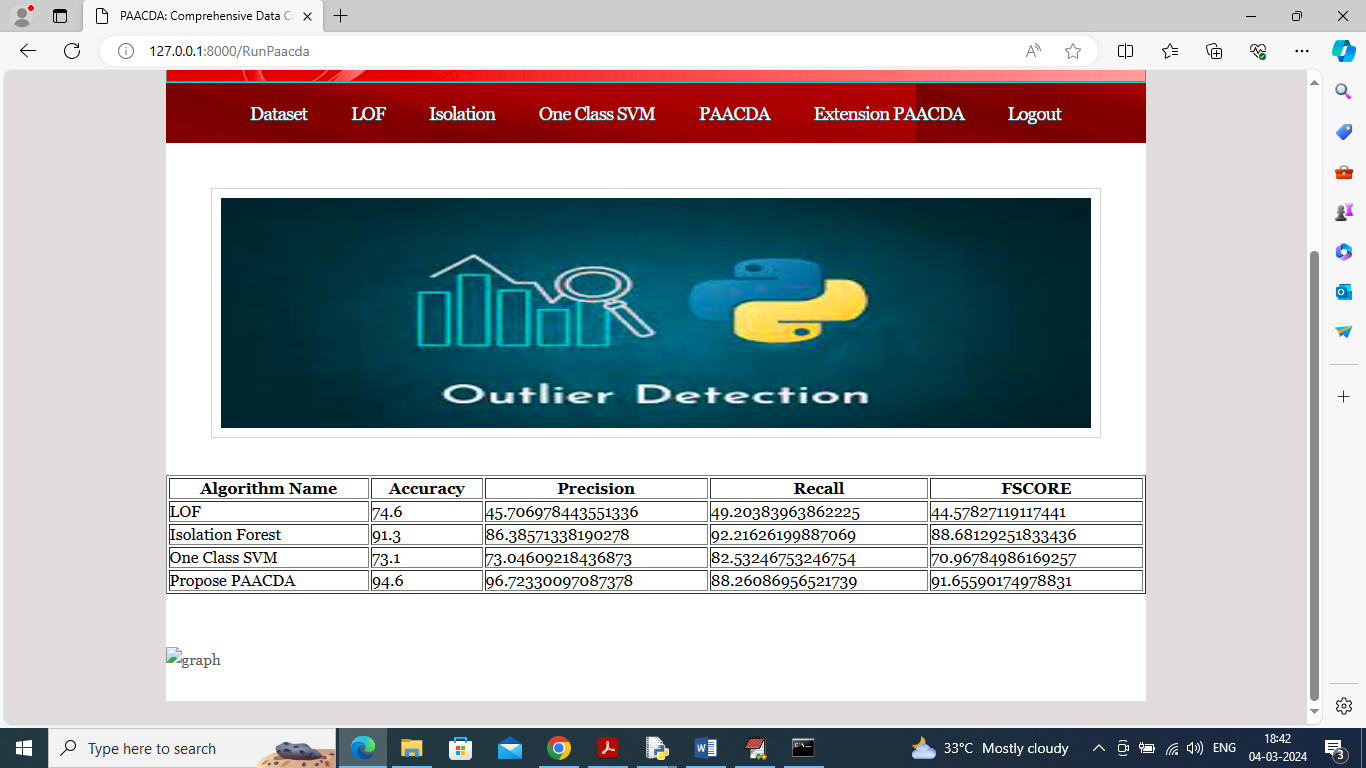
In above screen LOF training completed and got 74% accuracy for detected corrupted data and now click on ‘Run Isolation Forest’ link to train Isolation and get below page



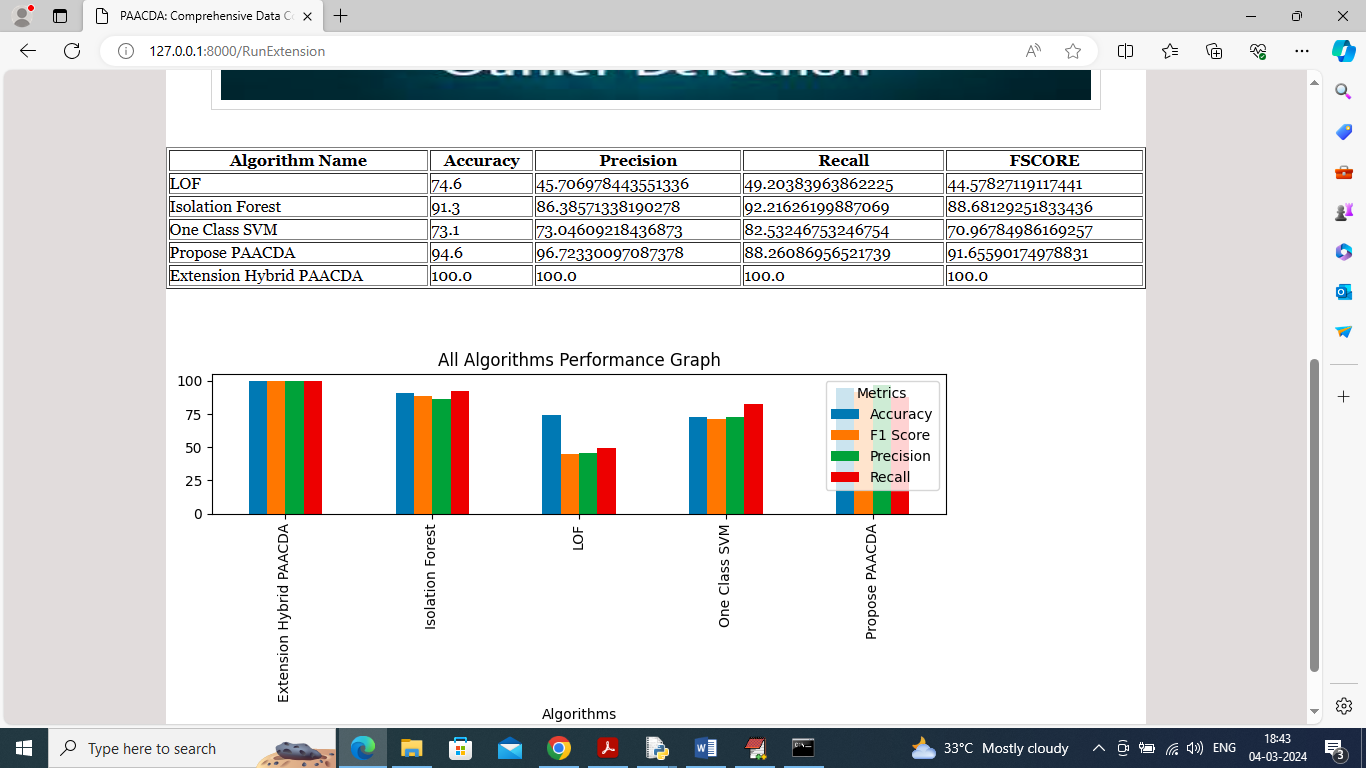
In above screen Isolation can detect corrupted values with 91% accuracy and now click on ‘One Class SVM’ to get below page



In above screen One Class SVM got 73% accuracy and now click on ‘Run PAACDA’ link to run propose algorithm and get below page

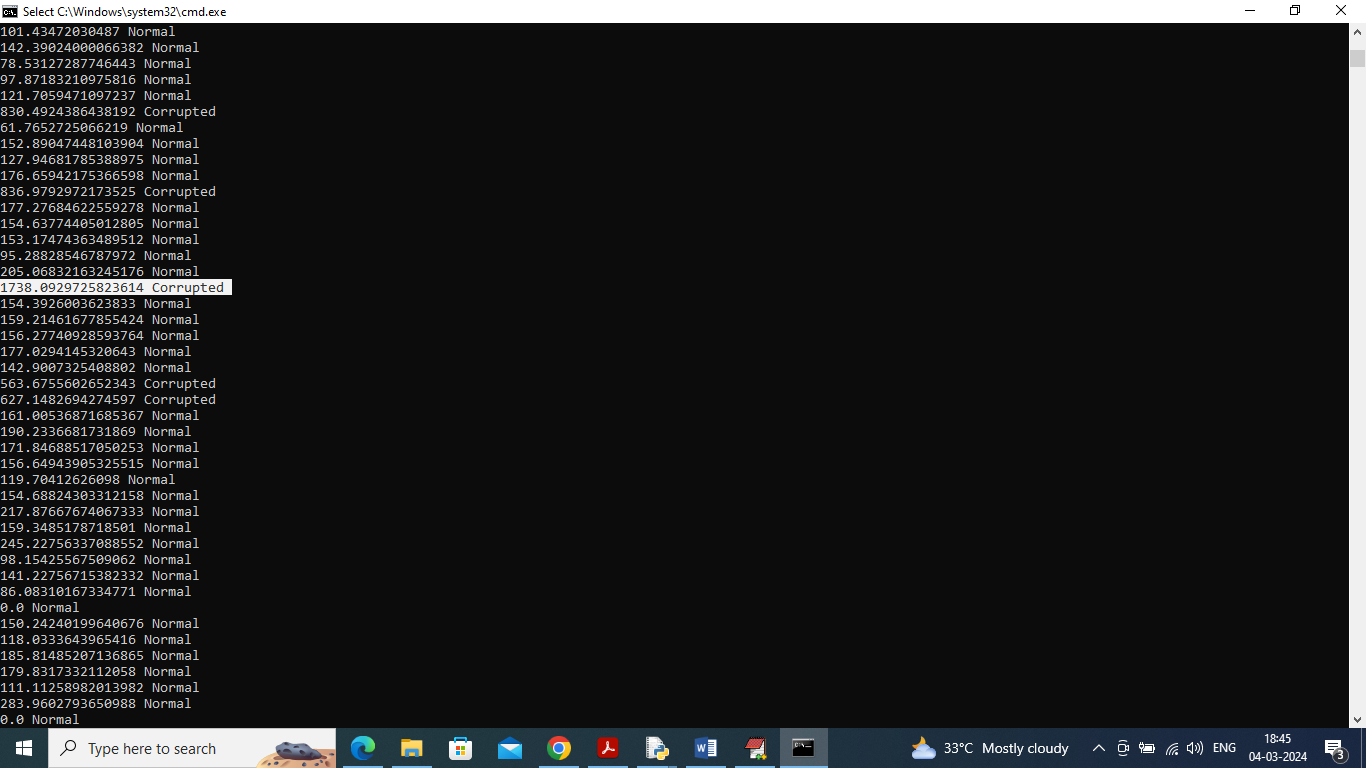


In above screen propose PAACDA got highest accuracy as 94% and now click on ‘Extension Hybrid PAACDA’ link to get below page



In above screen can see accuracy and other metrics from all algorithm and among all algorithms extension got 100% accuracy and in graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars.

In below screen showing normal and corrupted detected values



In above screen displaying values from dataset whose normal range is in between 0 to 300 but some values are 850 or more than 1300 and such values algorithm detecting as corrupted.