Online Fraud Payment Detection using Balanced ML Algorithms

In propose work we are utilizing ML algorithm to detect fraud payments from online transactions. To train ML algorithms we have used same fraud dataset which is given by you and this dataset contains only 1% records as fraud and remaining records are non-fraud which is making dataset highly imbalanced.

In propose work we have used Naïve Bayes and Random Forest algorithms for fraud detection on original and Smote based oversampling data. Smote algorithm will generate synthetic records for under-sampling classes to make dataset balanced.

After balancing dataset both algorithms got little improvement in accuracy. Each algorithm performance is evaluated in terms of accuracy, precision, recall, FSCORE and confusion matrix classification graph. Among all algorithms Random Forest with SMOTE giving high accuracy.

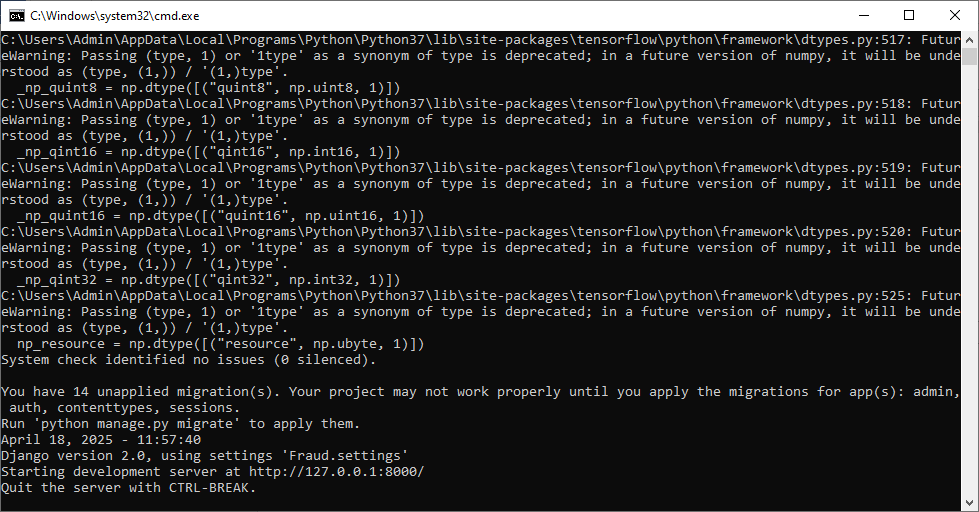
To implement this project we have designed following modules

1. New User Sign up: using this module user can sign up with the application
2. User Login: user can login to system
3. Load & Processed Dataset: after login user can run this module to load and normalized dataset using Standard Scaling and then convert all non-numeric data to numeric values using Label Encoder class. This module will split dataset into train and test where application using 80% data for training and 20% for testing and then will plot fraud and non-fraud transaction to show data imbalance
4. Balanced Data using ML: all processed training data will be input to SMOTE algorithm balanced dataset and then plot graph of balanced dataset class labels
5. Train ML Models: 80% training data will be input to both ML models with and without smote training data and then apply trained model on 20% test data to calculate prediction accuracy
6. Detect Fraud: using this module user can upload test data and then best ML model will be applied to predict weather test data contains normal or fraud payments

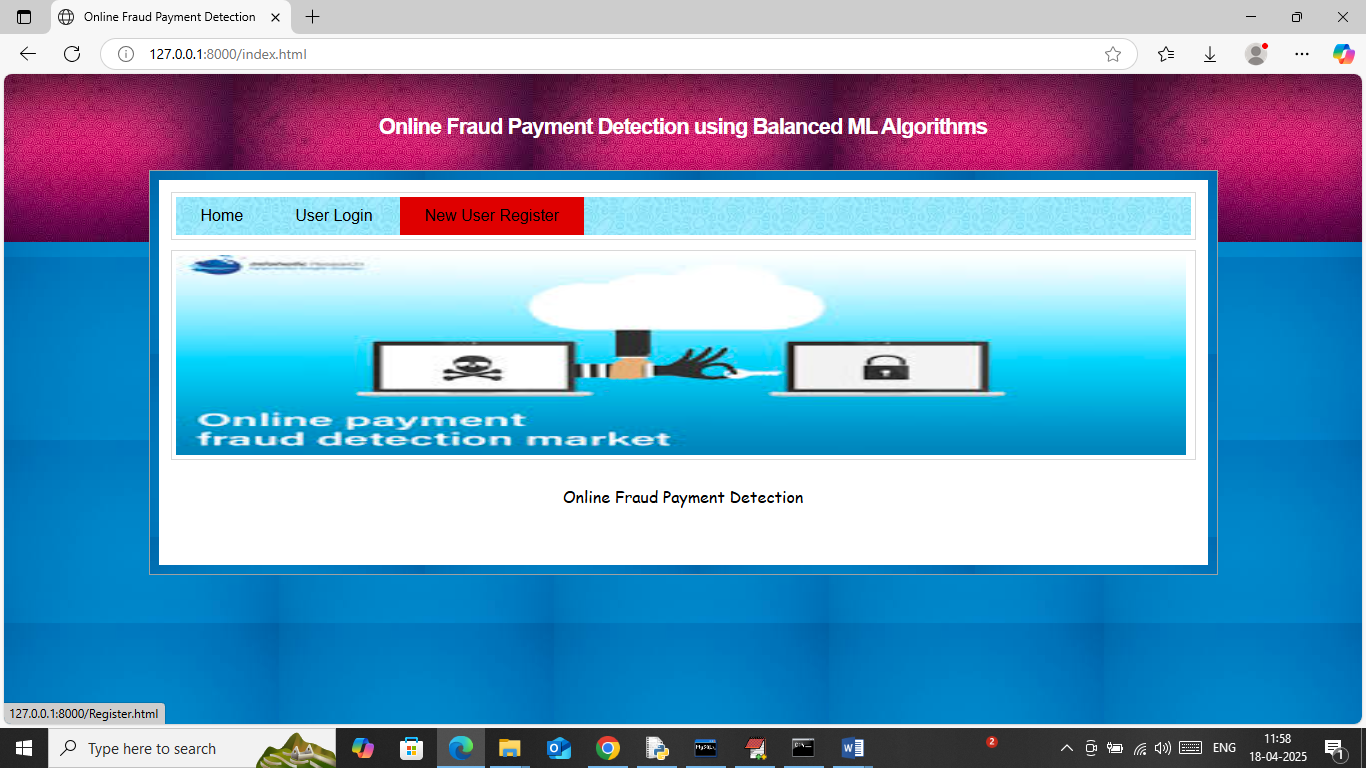
SCREEN SHOTS

Install python 3.7.2 and then install all packages given in requirements.txt file and then install MYSQL database and then copy content from ‘database.txt’ file and paste in MYSQL console to create database

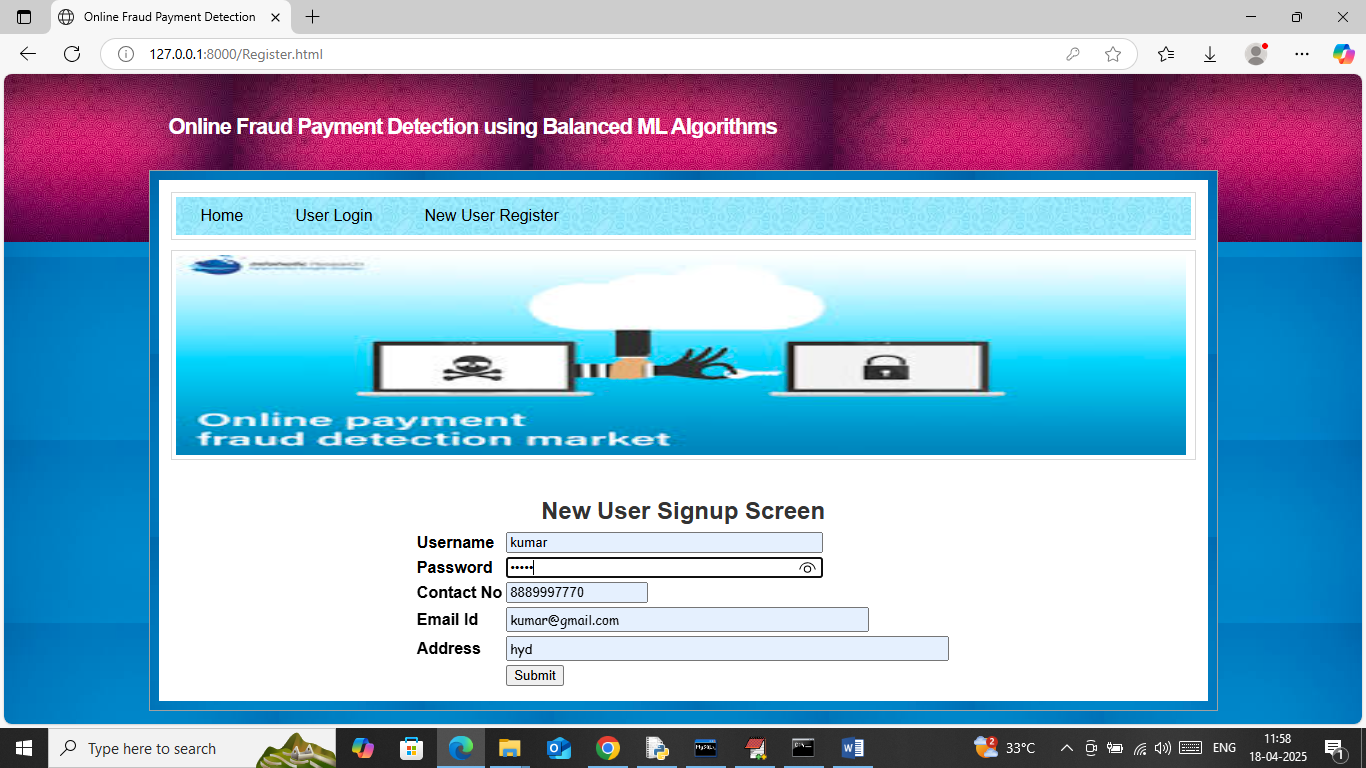
To run project double click on ‘runWebServer.bat’ file to start python server and then will get below page



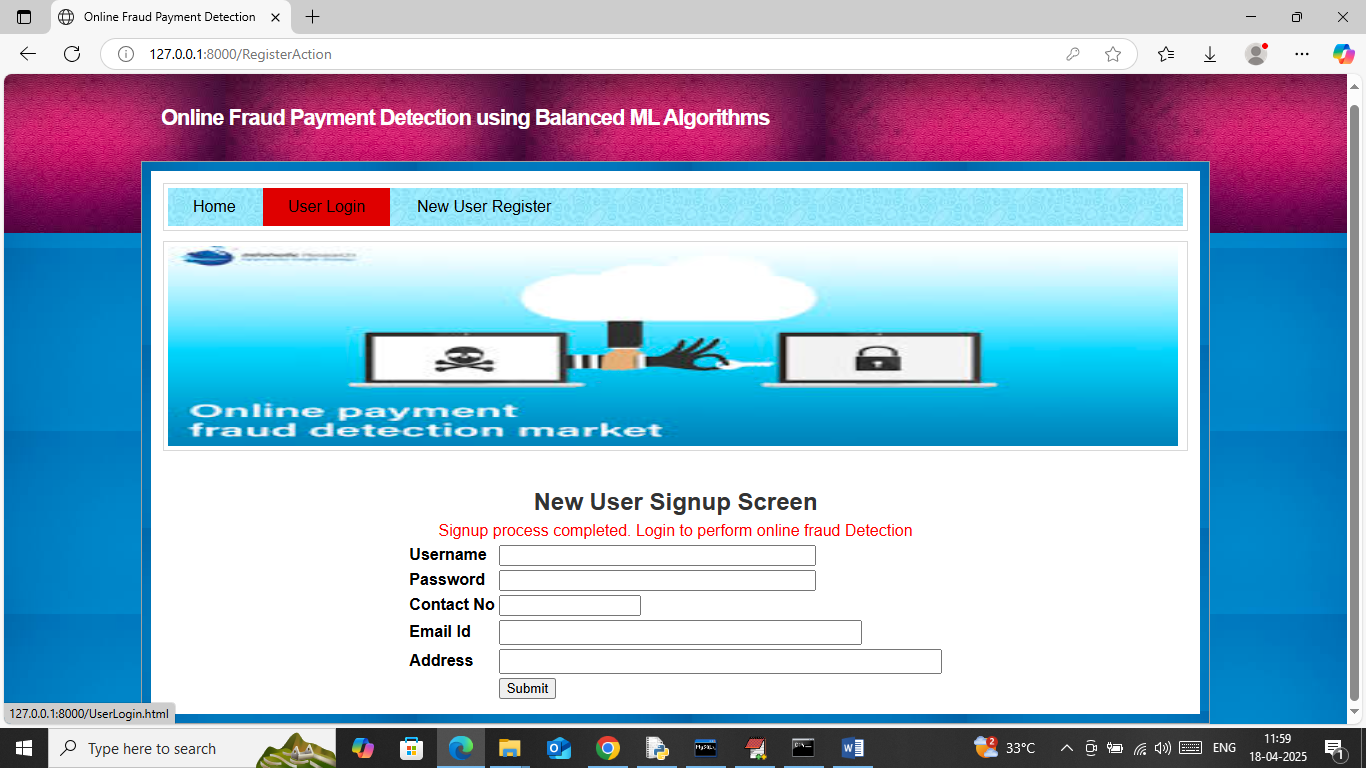
In above screen python server started and now open browser and enter URL as <http://127.0.0.1:8000/index.html> and then press enter key to get below page



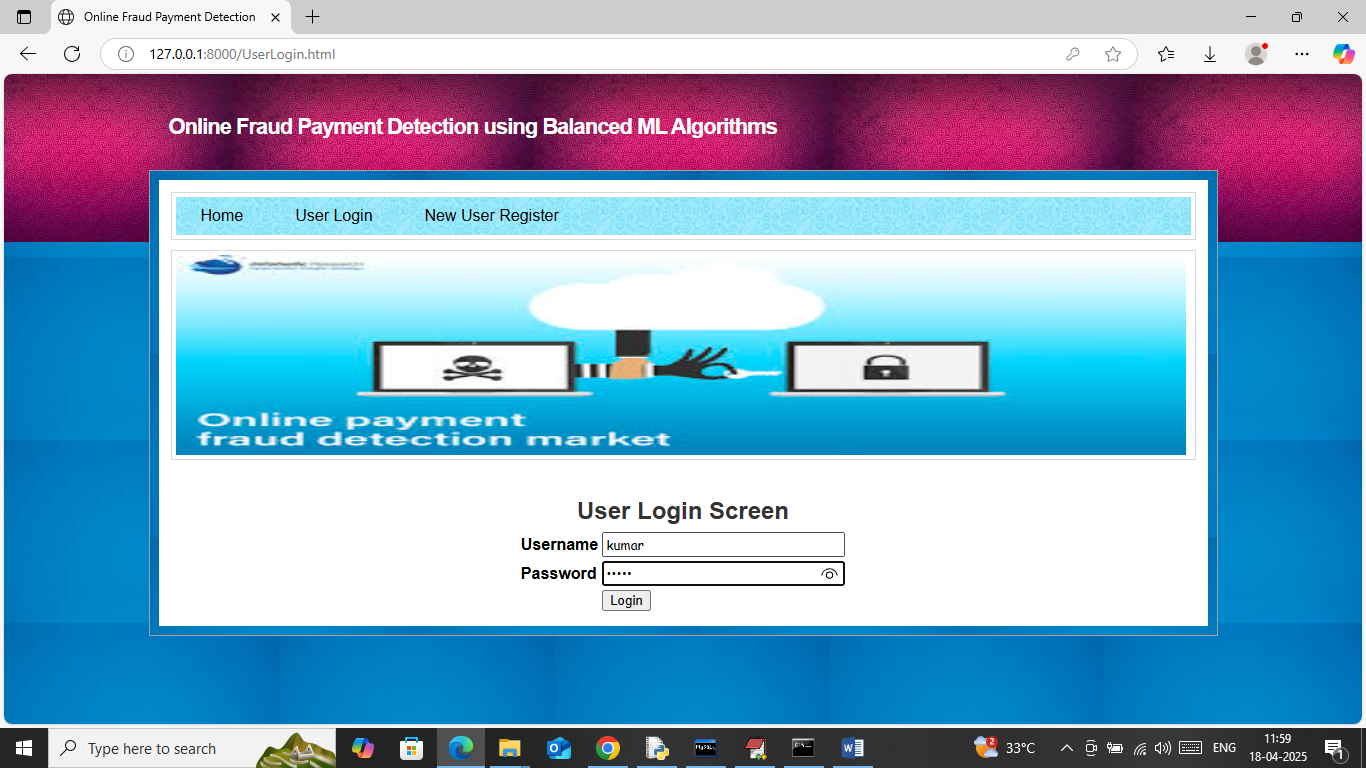
In above screen click on ‘New User Register’ link to get below page



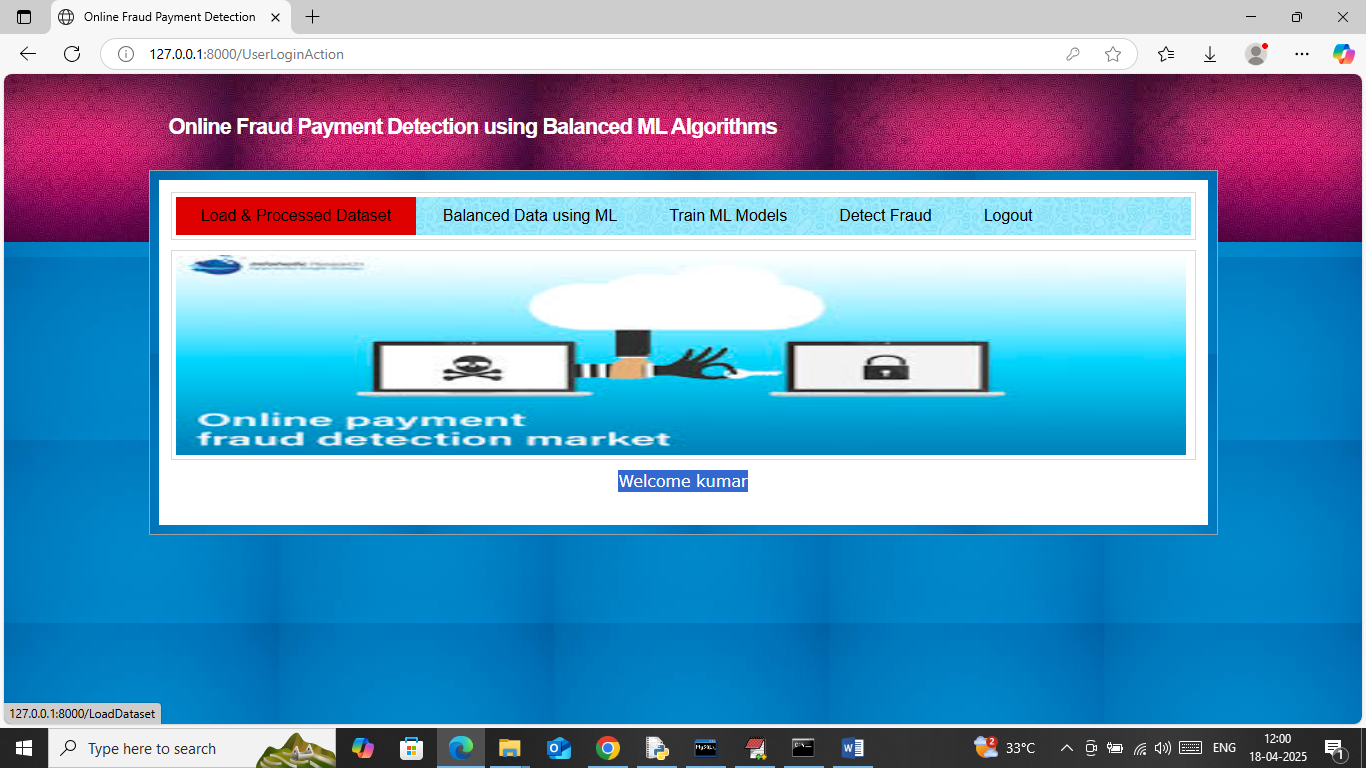
In above screen user is entering sign up details and then press button to get below page



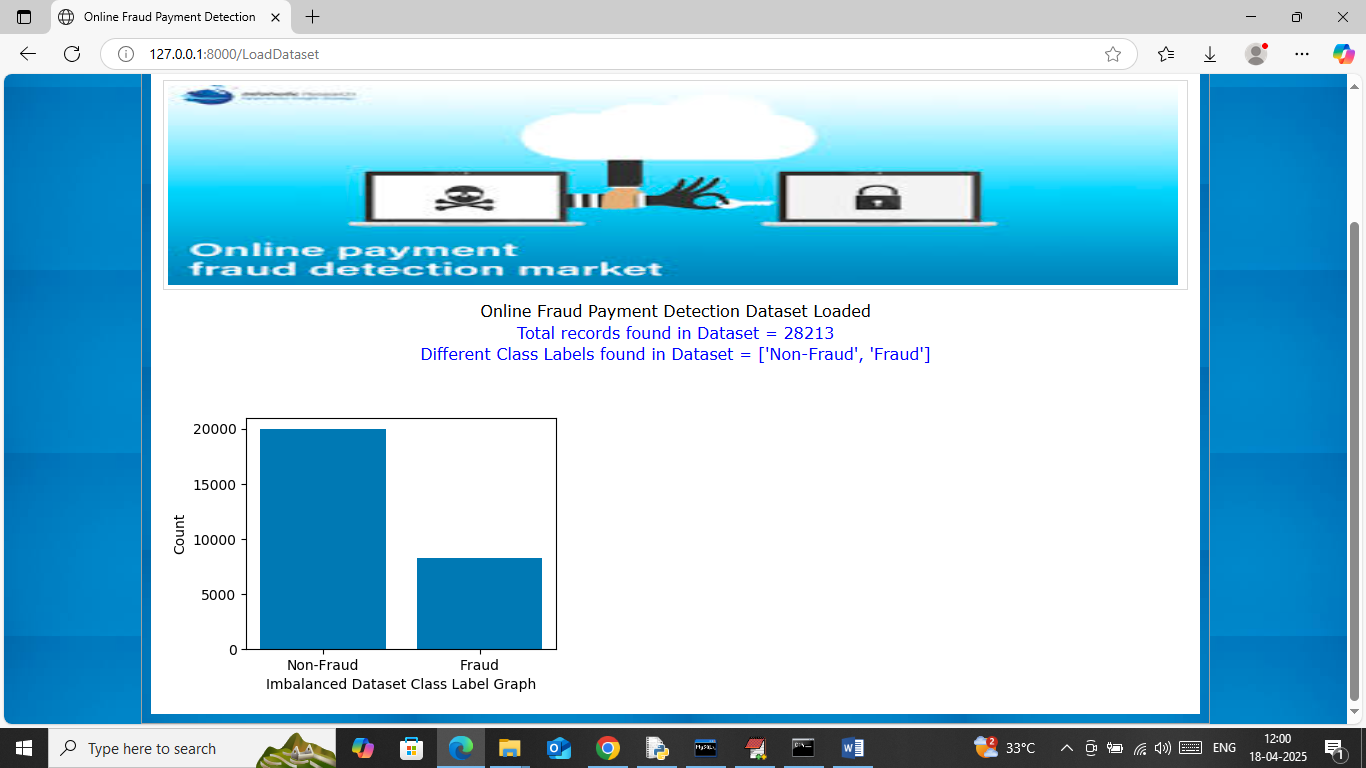
In above screen user sign up completed and now click on ‘User Login’ link to get below page



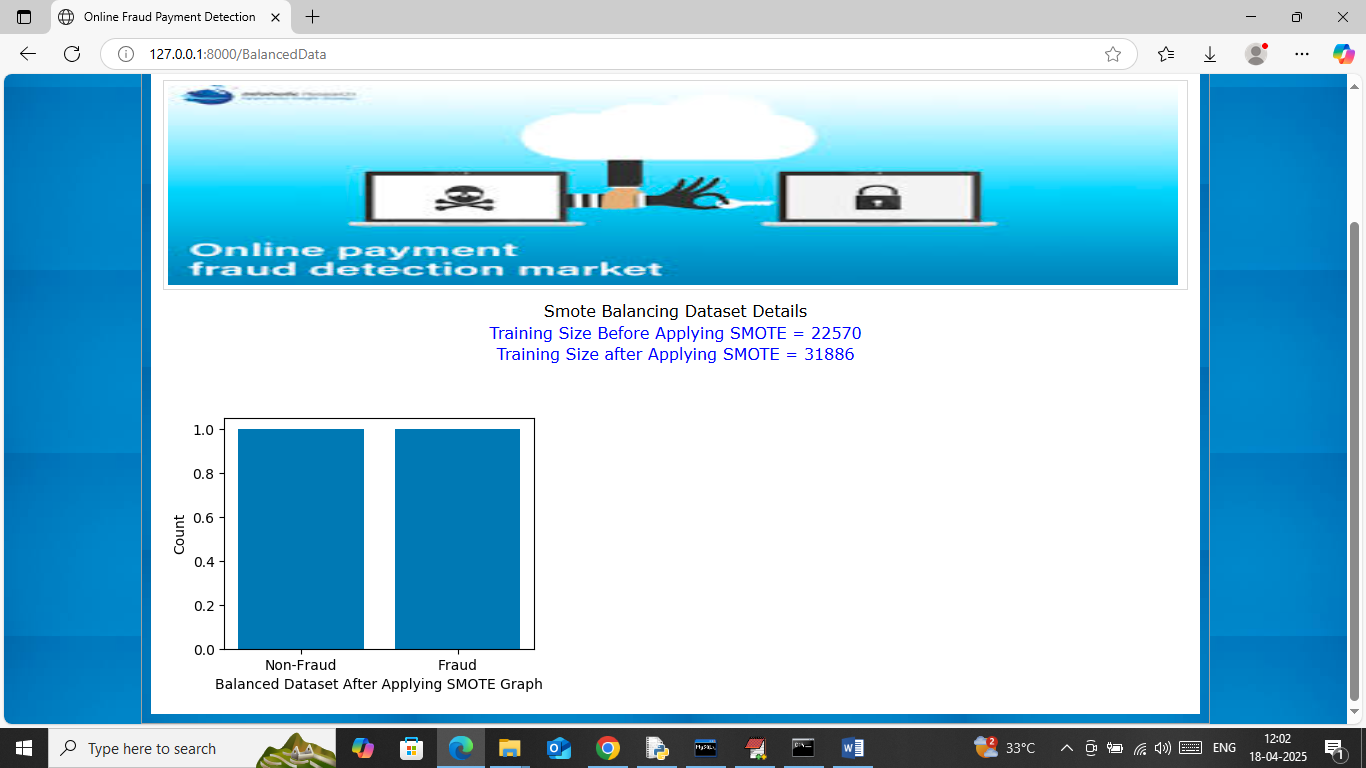
In above screen user is login and after login will get below page



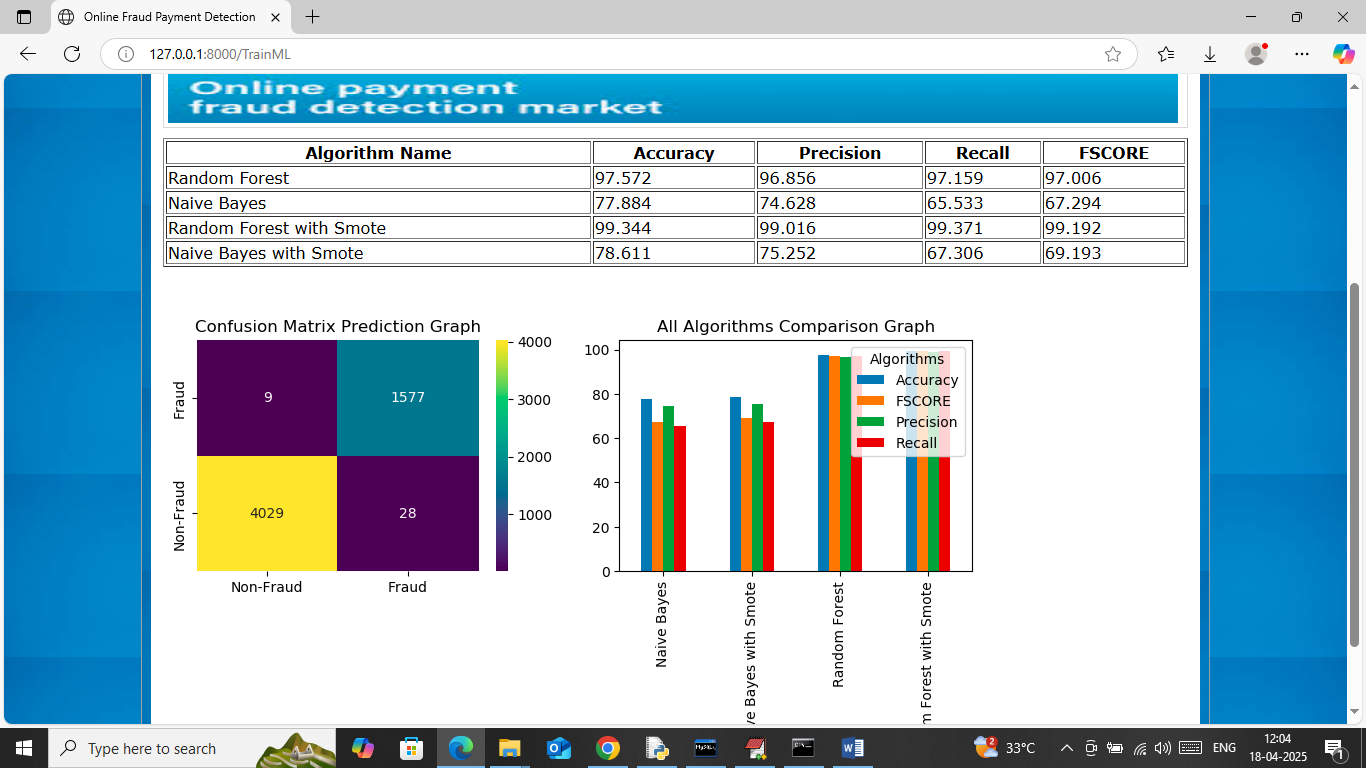
In above screen click on ‘Load & Processed Dataset’ link to load and process data and then will get below page



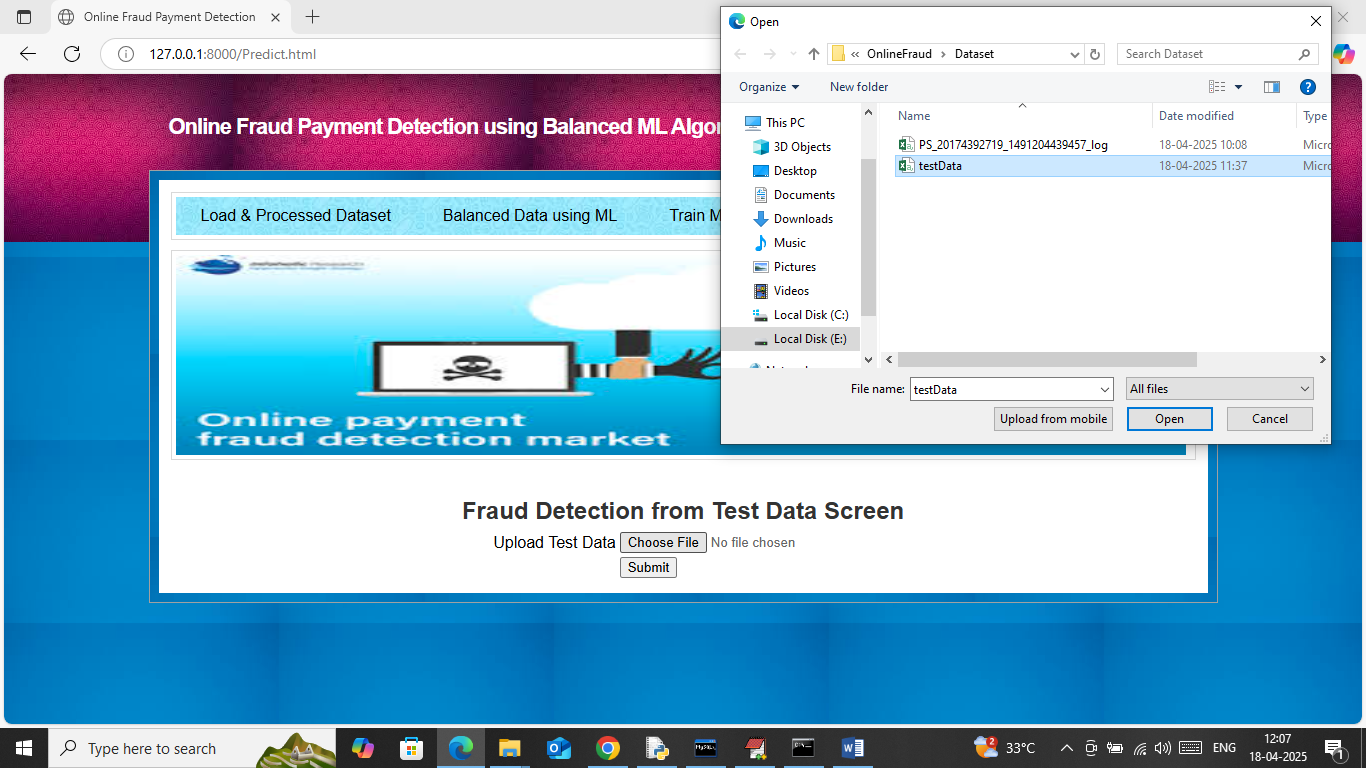
In above screen dataset loaded and can see dataset contains nearly 28000 records and can see available class labels as ‘Fraud and Non-Fraud’. In graph can see dataset contains more number of ‘Non-Fraud’ and less number of fraud payments which make dataset highly imbalance. Now click on ‘Balanced Data using ML’ link to balanced dataset and then will get below page



In above screen can see training records before smote is 22570 and after applying smote data size increased to 31886 and in graph can see both class labels have equal number of records. Now click on ‘Train ML Models’ link to train ML algorithms with and without smote and then calculate prediction accuracy



In above screen in table format can see ML algorithm accuracy with and without smote and in above screen can see Random Forest with smote got more than 99% accuracy. In confusion matrix graph x-axis represents ‘Predicted Labels’ and y-axis represents True Labels and then yellow and light green boxes in diagonal represents correct prediction count. Both blue boxes got incorrect prediction count which are very few. In bar graph showing comparison between all algorithms where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In all algorithms Random Forest with smote got high accuracy. Now click on ‘Detect Fraud’ link to get below page



In above screen selecting and uploading ‘testData.csv’ file and then click on ‘Open and submit’ button to load test data and then will get below prediction output



In above screen in first column can see ‘Test Data Values’ and then in second column can see predicted payments as ‘Fraud or Normal’.