Data Poison Detection Schemes for Distributed Machine Learning

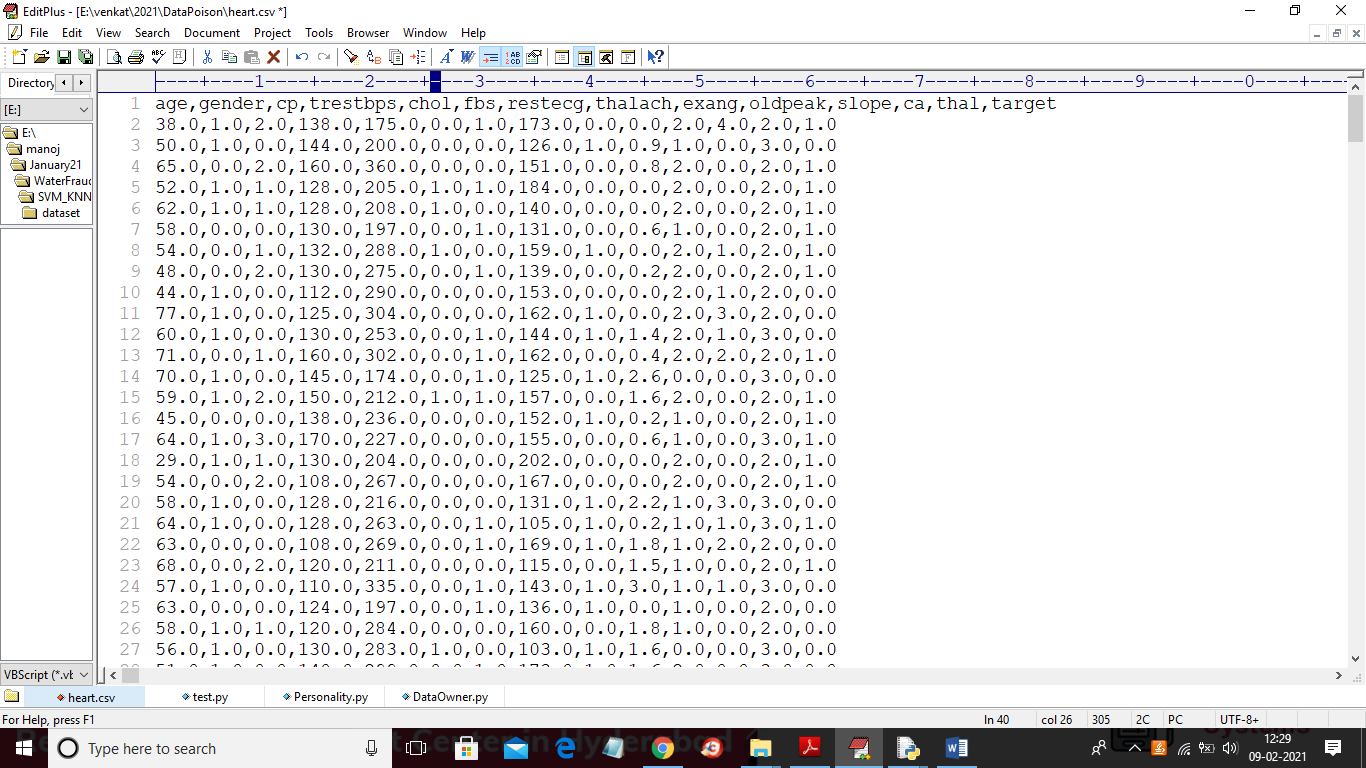
In machine learning applications we will train model by using dataset and then this model can be used to predict classes from new data. Machine learning model can be trained in two ways such as in local machine or distributed machine and in distributed machines dataset will be divided and share between multiple nodes and all nodes will process the dataset and then build ML model and send result back to centralized server.

In distributed environment sometime attackers may modify training data and then make ML to predict wrong result and to detect and remove such modified data author is using Data Poison Detection technique. This technique will inspect training dataset to identify odd values and then remove it. By applying Data Poison technique we can improve accuracy of ML algorithms.

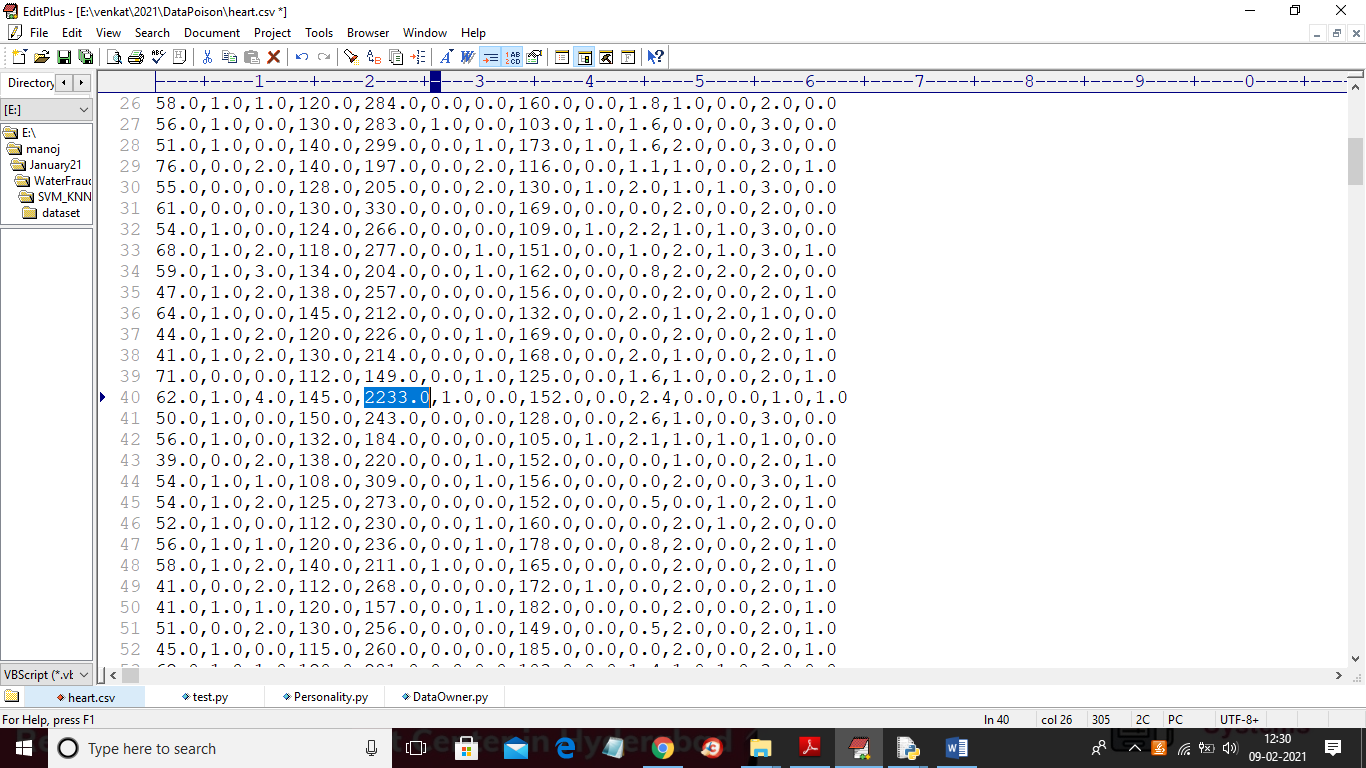
In propose paper author using two distributed techniques called Basic DML and Semi DML where Basic DML will divide dataset into multiple parts and send to worker nodes and worker nodes build ML model and send result back distributed center server. In Semi DML center server itself will devote resource to ML model to train dataset.

To implement this project we have compared performance of existing SVM and DML where existing SVM will not apply Data Poison Detection technique and DML technique will apply Data Poison technique.

To implement this project we have used heart disease dataset and in below dataset screen we can see dataset contains invalid data which called as Data Poision.



In above screen heart dataset first row contains column names and remaining rows are the column values and in below dataset screen we can see odd or invalid value



In above screen in selected blue value we can see recorded blood pressure value as 2233 which is wrong value and if ML train on such data then it may predict wrong result and it will reduce prediction accuracy and to avoid such problem we can apply Data Poison Detection technique. In python we can ‘IsolationForest’ class to detect and remove such poison data.

To implement this project we have designed following modules.

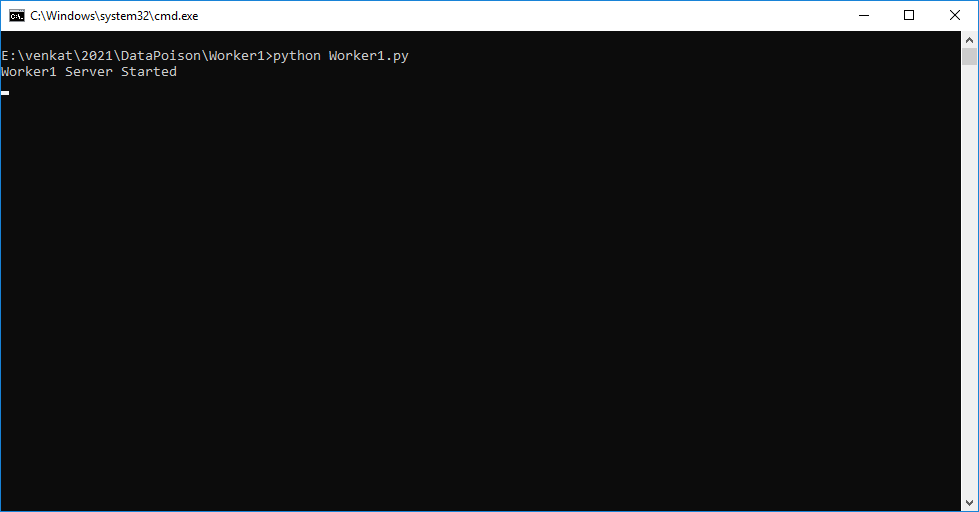
Worker1: This is a worker node which accept divided dataset from center server and then build existing SVM model and Basic DML model and then calculate accuracy of both algorithms and send result back to center server

Worker2: This is another worker node which accept other half of dataset and then run existing SVM and Basic DML and send accuracy back to center server.

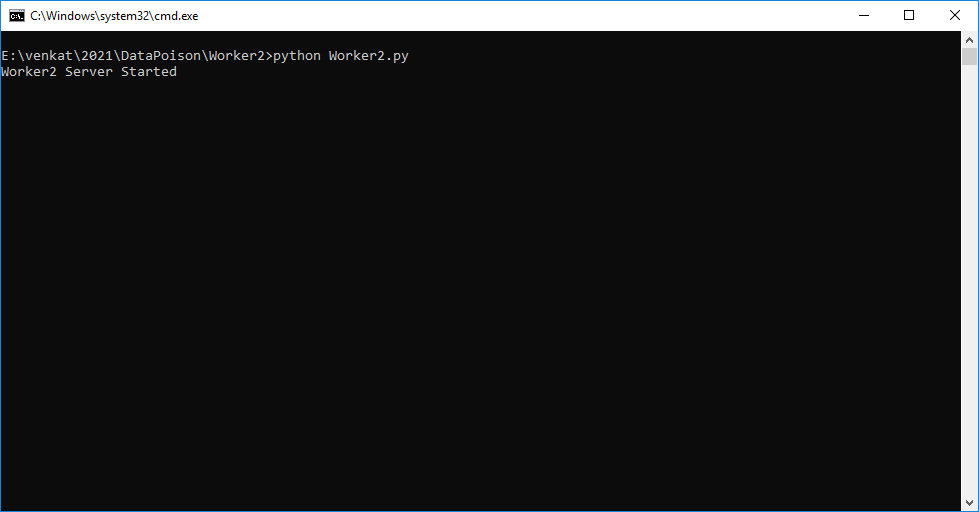
CenterServer: This is a center server which upload dataset to application and then divide dataset into two equal parts and then distribute each part to worker 1 and 2 and then collect result. This server will run semi DML and calculate its accuracy also.

SCREEN SHOTS

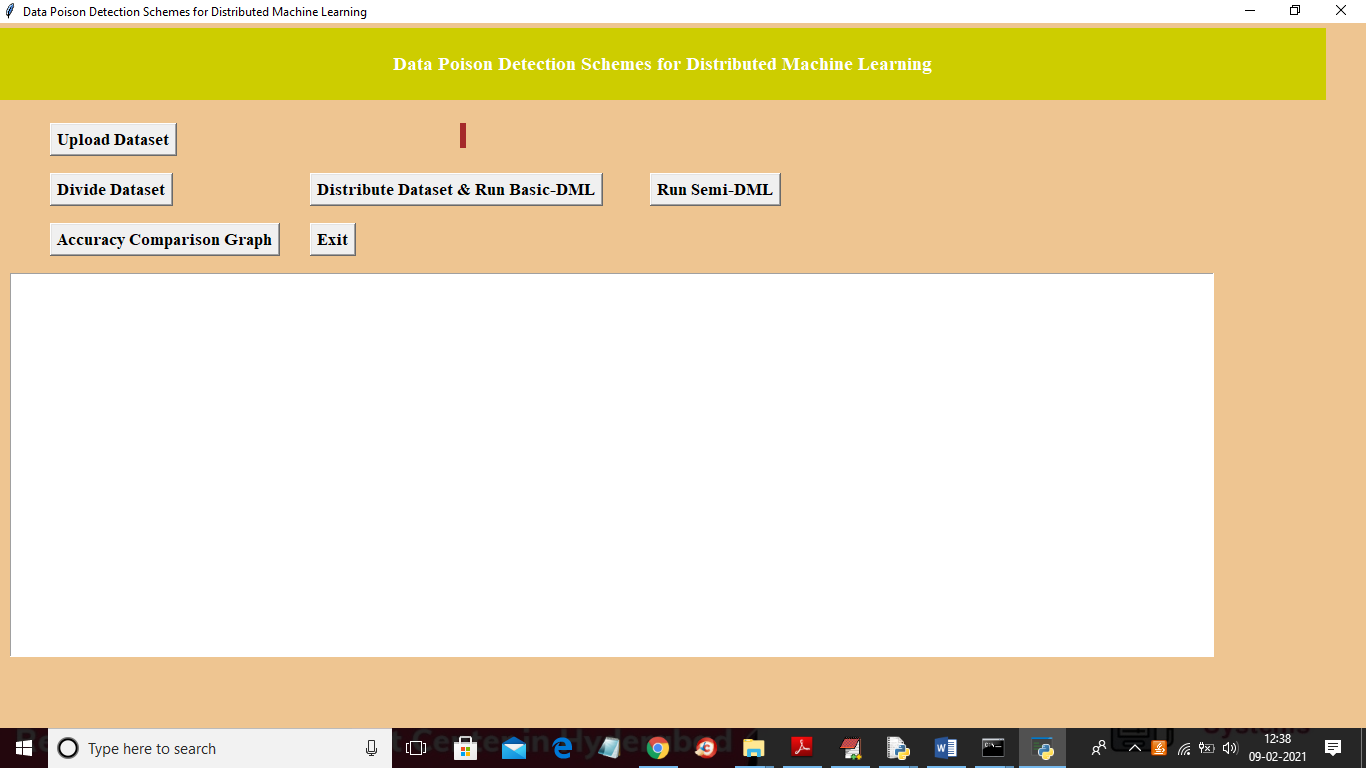
To run project first double click on ‘run.bat’ file from Worker1 folder to start worker 1 node and to get below screen



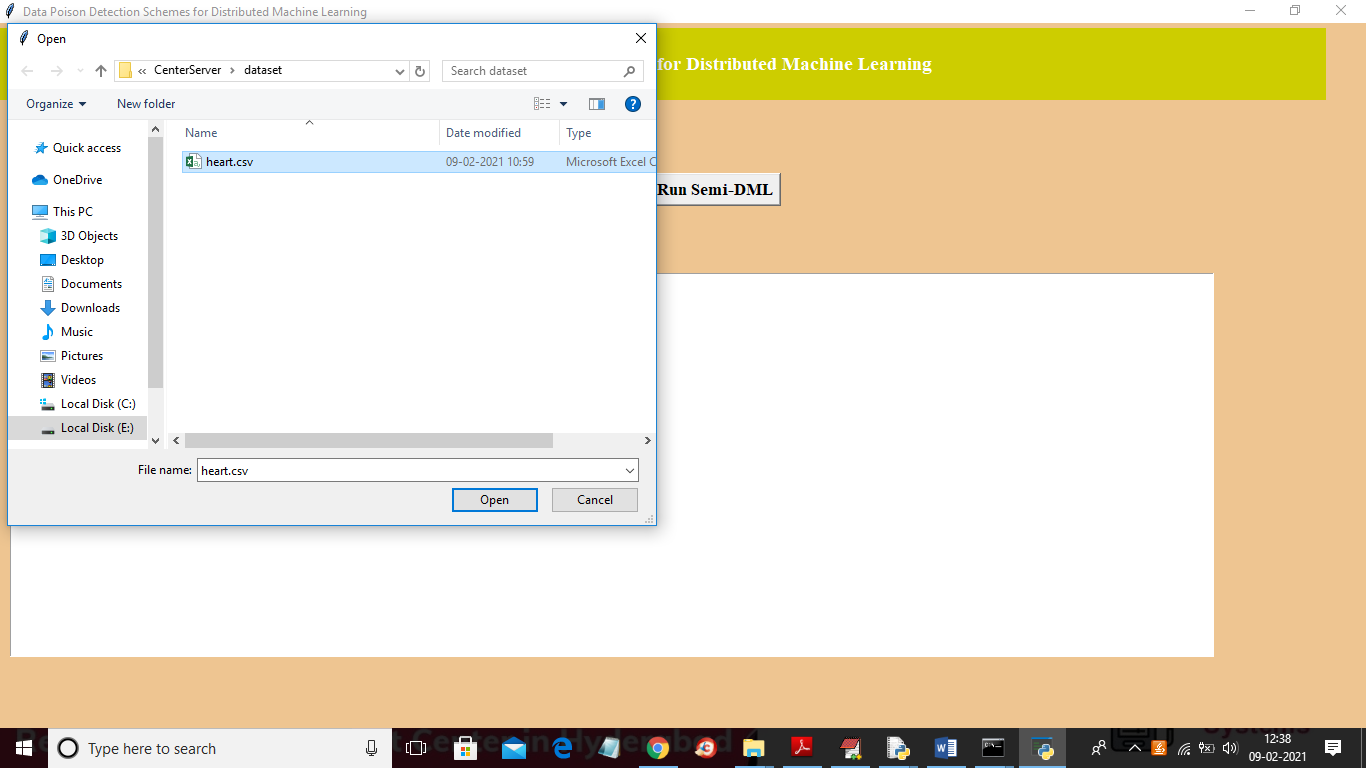
In above screen worker 1 server started and now double click on ‘run.bat’ file from worker2 folder to start worker 2



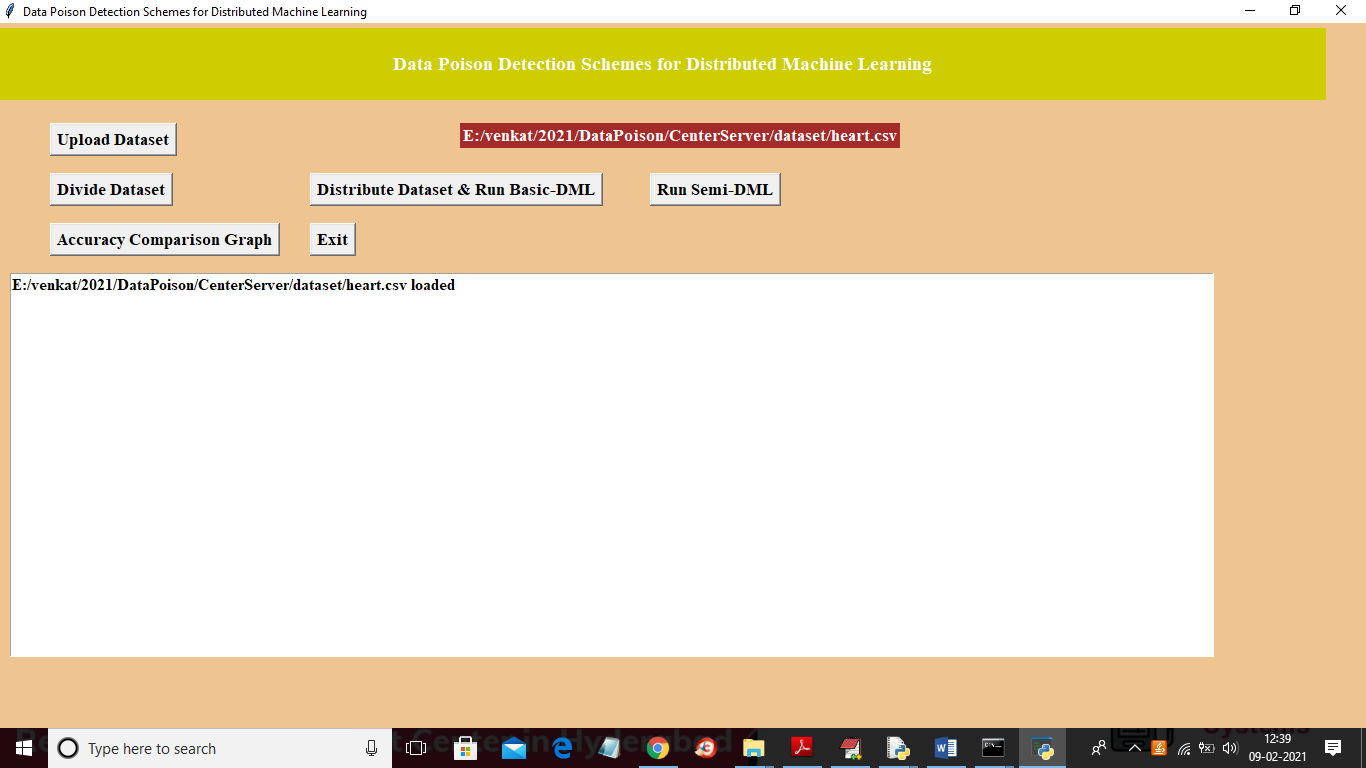
In above screen worker2 server started and now double click on ‘run.bat’ file from ‘CenterServer’ folder to start distributed server and to get below screen



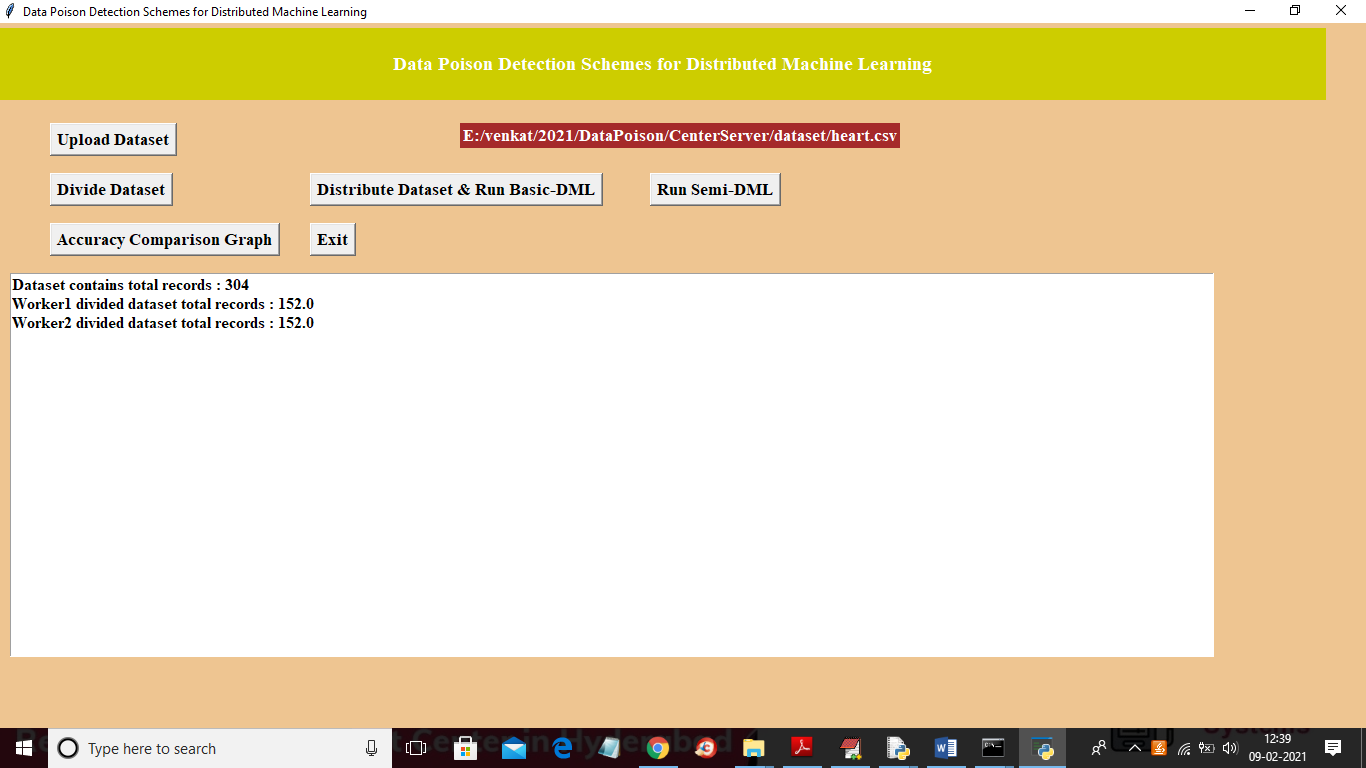
In above screen click on ‘Upload Dataset’ button to upload dataset and to get below screen



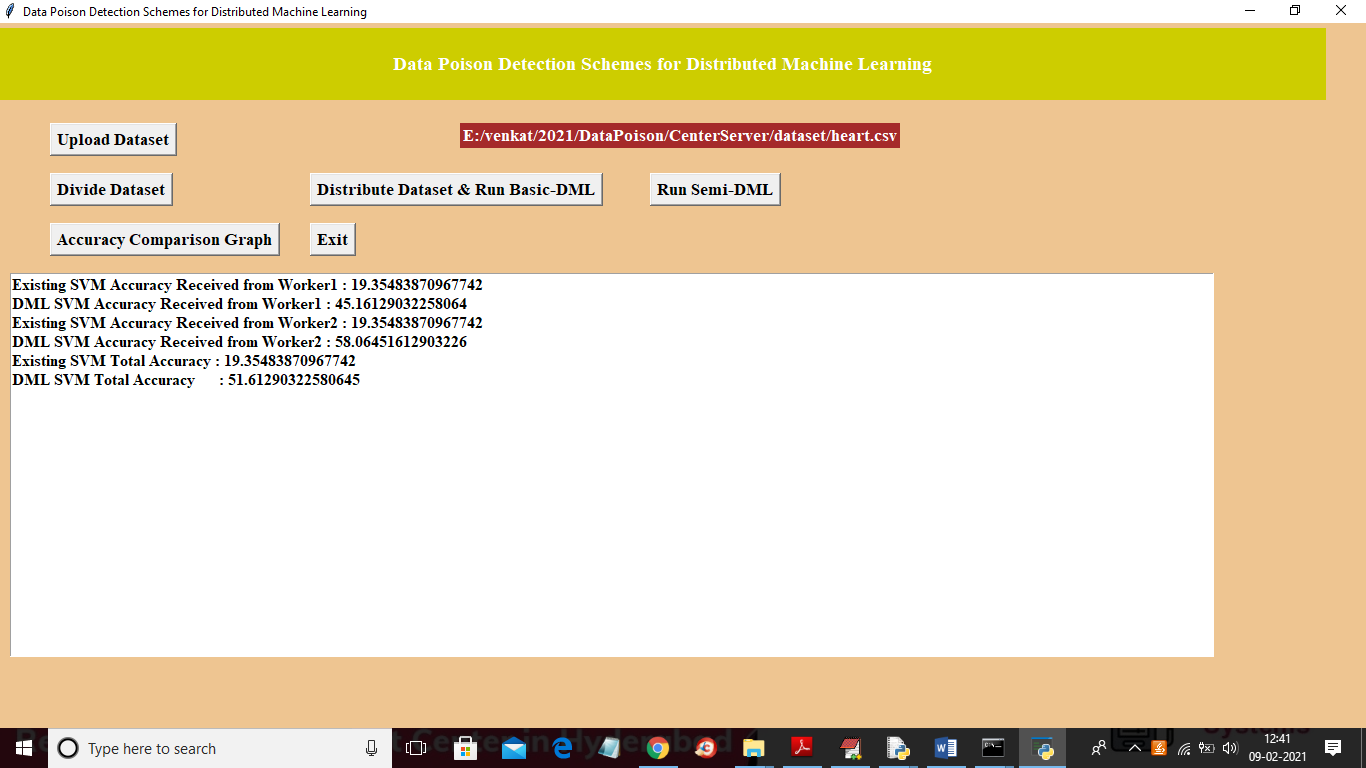
In above screen selecting and uploading ‘heart.csv’ file and then click on ‘Open’ button to load dataset and to get below screen



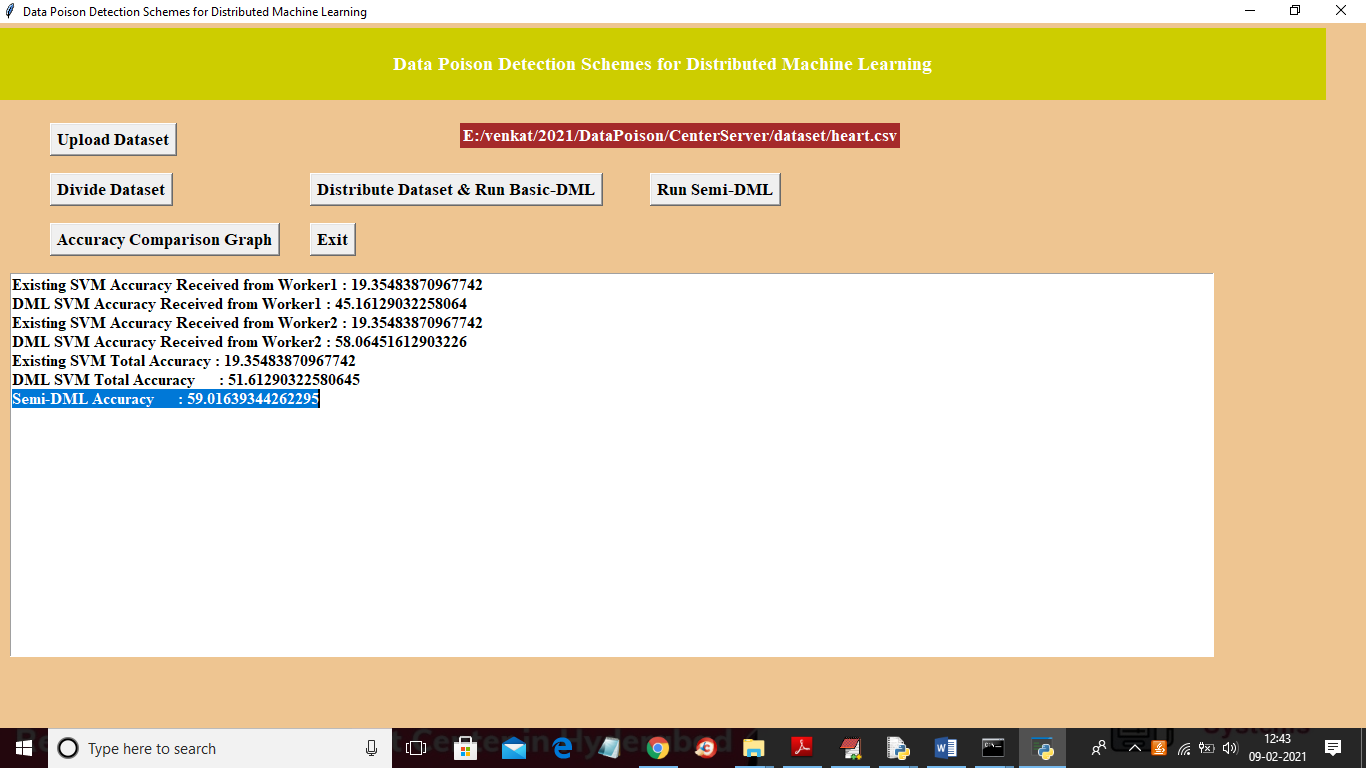
In above screen dataset loaded and now click on ‘Divide Dataset’ button to divide dataset into 2 equal parts



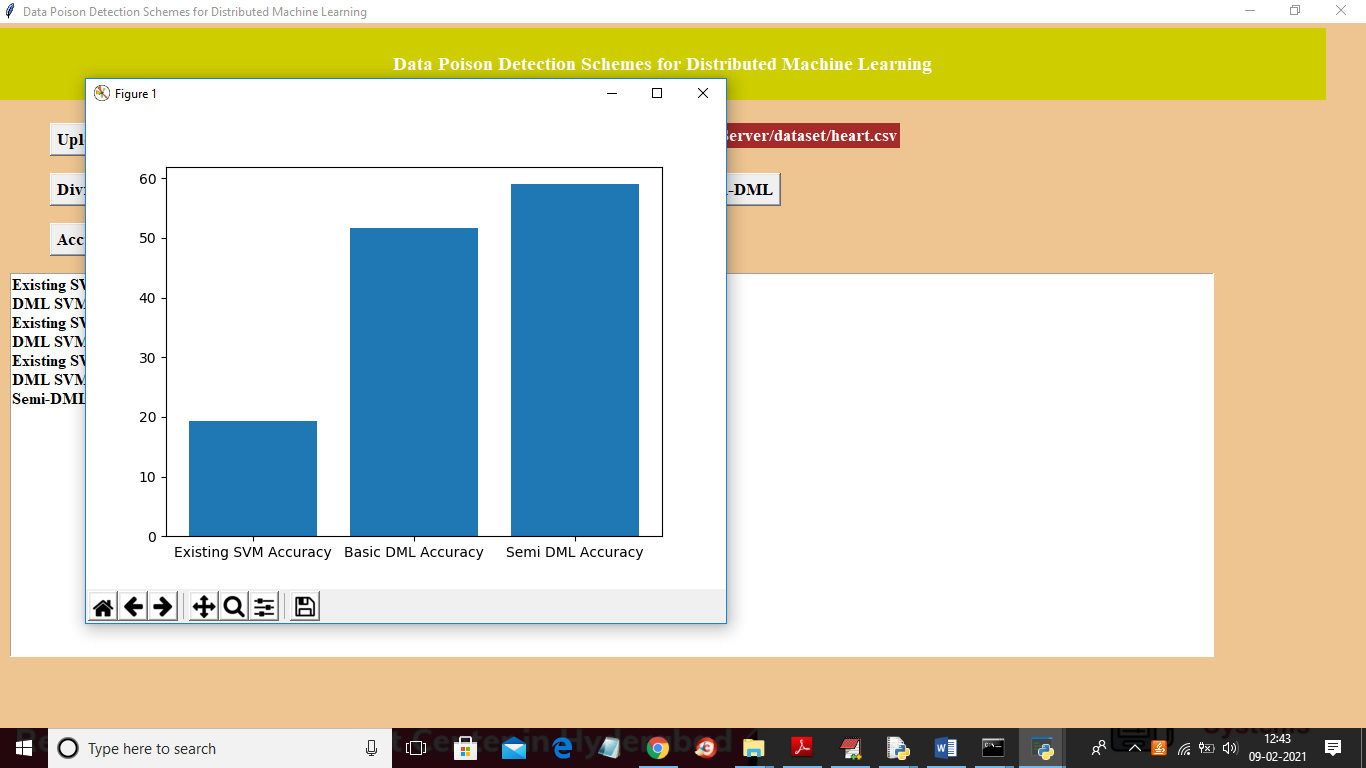
In above screen dataset contains 304 records and equally distributed to 2 parts and now click on ‘Distribute Dataset & Run Basic-DML’ button to distribute dataset to 2 workers and then get accuracy result



In above screen we got result from 2 worker nodes for existing SVM accuracy and propose DML accuracy and in above screen we can see existing SVM accuracy is 19% when data poison exists in dataset and after removing data poison using DML technique we got 51% accuracy and now click on ‘Run Semi-DML’ button to allow center server to devote resources to DML and then remove poison from dataset and then calculate accuracy



In above screen Semi-DML accuracy is 59% and now click on ‘Accuracy Comparison Graph’ button to get below graph



In above screen x-axis contains algorithm name and y-axis represents accuracy and from above graph we can conclude that Basic-DML and Semi-DML accuracy is better than existing SVM accuracy. In below worker screens also we can see accuracy values

