

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

FALL SEMESTER 2022-2023

## FINAL REPORT

**COURSE:** DATA MINING TECHNIQUES

**SLOT:** D1

**COURSE CODE:** SWE2009

## TITLE : MACHINE LEARNING BASED DDOS ATTACK CLASSIFIER IN SOFTWARE DEFINED NETWORK

**SUBMITTED BY** VAISHALI G-19MIS0009 VIDHYA R-19MIS0035

SNEHA K-19MIS0043

## SUBMITTED TO

FACULTY: PROF.THIPPA REDDY G

## ABSTRACT:

Intrusion Detection System (IDS) is an important tool use in cyber security to monitor and determine intrusion attacks This study aims to analyse recent researches in IDS using Machine Learning (ML) approach; with specific interest in dataset, ML algorithms and metric. Dataset selection is very important to ensure model build is suitable for IDS use. The conventional methods like firewall, which focused on the filtering of data, may not suitable to identify all types of attacks on time. For effective handling and timely identification of these types of attacks, intrusion detection systems (IDS) based on machine learning algorithms are very effective to efficiently process the large volume of data for identifying any malicious activity. Machine learning based IDS are used to analyse all network activities for any malicious behaviour. The system focus was on false negative and false positive performance metrics in order to enhance the detection rate of the intrusion detection system. The implemented experiments demonstrated that the ML model SVM classifier achieved the highest value of accuracy while the KNN classifier has achieved the lowest accuracy rate.

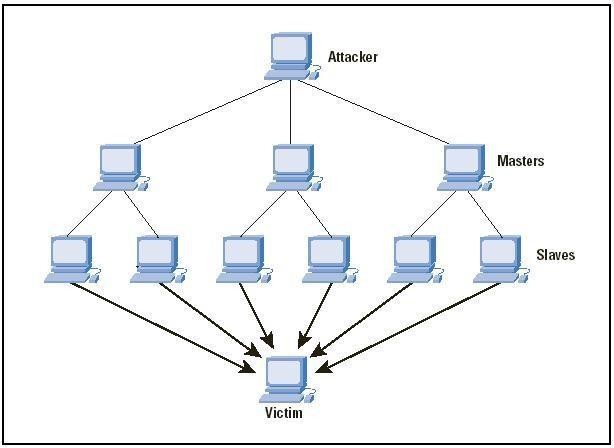
## INTRODUCTION:

Cloud computing provides its users to access the various cloud services and enables them to access the data storage and the computational resources with low data overhead. Cloud computing is rapidly growing and it has attracted the users to cloud. It offers the users to outsource the data resources from the remote locations when they need them. Cloud computing model has three effective service models and deployment models. The main characteristics provided to cloud users are: on demand service to its users, broad network access, resource pooling, rapid elasticity and measured service.

## DDoS Attacks:

DDos attacks are attempted in various forms, it can be flooding the network with a large amount of packets, , synchronization of those packets or attackers may create the zombies towards the victim machine. The most basic and effective DDos attack strategy is to flood the network with the attack packets. DDoS attacks have always been a threat to individual cloud customers as they have fewer resources to beat such attacks. But it is possible to mitigate the cloud attacks using various hazards of cloud computing.

A denial-of-Service attack (DDoS attack) makes the cloud resources unavailable or exhausts the resources for the cloud users. The common method used in such attacks is to saturate the network of the targeted machine with the fake packet requests so that legitimate traffic on the network couldn’t be responded. Such attacks usually overload the server. Distributed attacks involve the countless attackers who target a single machine.



## PROBLEM STATEMENT :

While improvement of communication and distribution of information, some disadvantage happened. Information security becomes a crucial need. Since billions of transactions occur through the Internet. Network failure in one-second impacts on millions of losses for that organization. There are lots of hacking methods are used to hack the client servers. The Intrusion detection becomes the most famous attack than the other cyberattacks nowadays. Hacking associative have lots of machines attach to their botnets. These botnets are capable to shut down any network, and it is a dangerous issue on the Internet. 3 Recognition of an Intrusion detection is not easy, but it can happen in small range of time. Attackers visible from thousands of IP addresses and Security of the Internet also increases because of there are many threats to servers and networks.

One of them is Distributed Denial of Service (DDoS) attack and it is trying to make online servers/services unavailable with massive traffic using multiple sources. Which means attackers make Servers busy or down. Attackers send malicious to personal computers and attackers remotely controlling the infected personal computers as botnets against any targets.

## SYSTEM MODULES:

* Data Loading
* Data Training
* Algorithm Applying
* Prediction

## MODULES DESCRIPTION:

* **Data Loading**

Data loading is the process of copying and loading data or data sets from a source file, folder or application to a database or similar application. It is usually implemented by copying digital data from a source and pasting or loading the data to a data storage or processing utility. Data loading is used in database-based extraction and loading techniques. Typically, such data is loaded into the destination application as a different format than the original source location. For example, when data is copied from a word processing file to a database application, the data format is changed from .doc or .txt to a .CSV or DAT format. Usually, this process is performed through or the last phase of the Extract, Transform and Load (ETL) process. The data is extracted from an external source and transformed into the destination application's supported format, where the data is further loaded.

## Data Pre-processing

Missing values were imputed to guarantee that all the algorithms would be able to handle them. Nevertheless, some algorithms could deal with missing values automatically without imputation, such as XG Boost. To restrict the comparison complexity, the missing values were imputed based on their data type. For numerical data types, the missing entries are replaced by the median value of the complete entries. For categorical data, the missing entries were replaced by the mode value of the complete entries.

## Data Cleaning

In this module the data is cleaned. After cleaning of the data, the data is grouped as per requirement. This grouping of data is known as data clustering. Then check if there is any missing value in the data set or not. It there is some missing value then change it by any default value. After that if any data need to change its format, it is done. That total process before the

30 prediction is known is data pre-processing. After that the data is used for the prediction and forecasting step.

## Data Splitting

For each experiment, we split the entire dataset into 70% training set and 30% test set. We used the training set for resampling, hyper parameter tuning, and training the model and we used test set to test the performance of the trained model. While splitting the data, we specified a random seed (any random number), which ensured the same data split every time the program executed.

## Data Training

Algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they’re given. And the better the training data is, the better the model performs. In fact, the quality and quantity of your training data has as much to do with the success of your data project as the algorithms themselves. Now, even if you’ve stored a vast amount of well-structured data, it might not be labelled in a way that actually works for training your model. For example, autonomous vehicles don’t just need pictures of the road, they need labelled images where each car, pedestrian, street sign and more are annotated; sentiment analysis projects require labels that help an algorithm understand when someone’s using slang or sarcasm; chatbots need entity extraction and careful syntactic analysis, not just raw language. In other words, the data you want to use for training usually needs to be enriched or labelled. Or you might just need to collect more of it to power your algorithms. But chances are, the data you’ve stored isn’t quite ready to be used to train your classifiers. Because if you’re trying to make a great model, you need great training data. And we know a thing or two about that. After all, we’ve labelled over 5 billion rows of data for some of the most innovative companies in the world. Whether it’s images, text, audio, or, really, any other kind of data, we can help create the training set that makes your models successful.

## Dataset

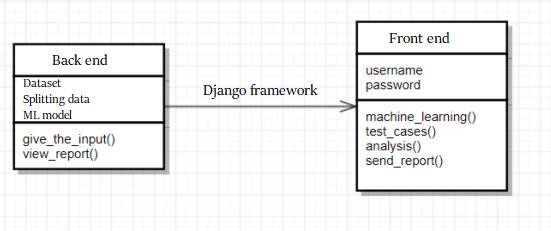
The dataset used in this project contains various field like source address, destination address, packet ID, source and destination node etc. The last column contains the type of attack for a particular row.

GIVEN INPUT: There are total five classes which are b'HTTP-FLOOD',b'Normal', b'SIDDOS', b'Smurf', b'UDP-Flood'. Our model has to predict one out of these five class.

# USE CASE DIAGRAM



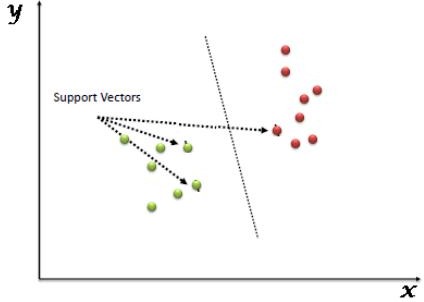
**CLASS DIAGRAM**



# TECHNIQUES:

## Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used to both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm it plots each data item as a point in n-dimensional space (where n is number of features) with the value of each feature is being in the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiate the two classes very well.



This are simply co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line)

## K-nearest neighbour (KNN)

KNN is most widely used algorithm in the field of machine learning, pattern recognition and many other areas. KNN is used for classification problems. This algorithm is also known as instance based (lazy learning) algorithm. A model or classifier is not immediately built but all training data samples are saved and waited until new observations

need to be classified. This characteristic of lazy learning algorithm makes it better than eager learning, that construct classifier before new observation needs to be classified.

That this algorithm is also more significant when dynamic data is required to be changed and updated more rapidly. KNN with different distance metrics were employed. KNN algorithm works according to the following steps using Euclidean distance formula.

It is a non-parametric classification method that calculates class memberships based on k- closest training examples. kNN is the simplest and a type of lazy learning method. Classification continues approximated locally and all computation is deferred until

classification. So, for this study, EEG data is classified with cross validation by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours

* Step I: To train the system, provide the feature space to KNN
* Step II: Measure distance using Euclidean distance formula
* Step III: Sort the values calculated using Euclidean distance using d i ≤ d i + 1, where i

= 1, 2, 3, …, k

* Step IV: Apply means or voting according to the nature of data
* Step V: Value of K (i.e., number of nearest Neighbours) depends upon the volume and nature of data provided to KNN. For large data, the value of k is kept as large, whereas for small data the value of k is also kept small.

## IMPLEMENTATION:

DATASET :

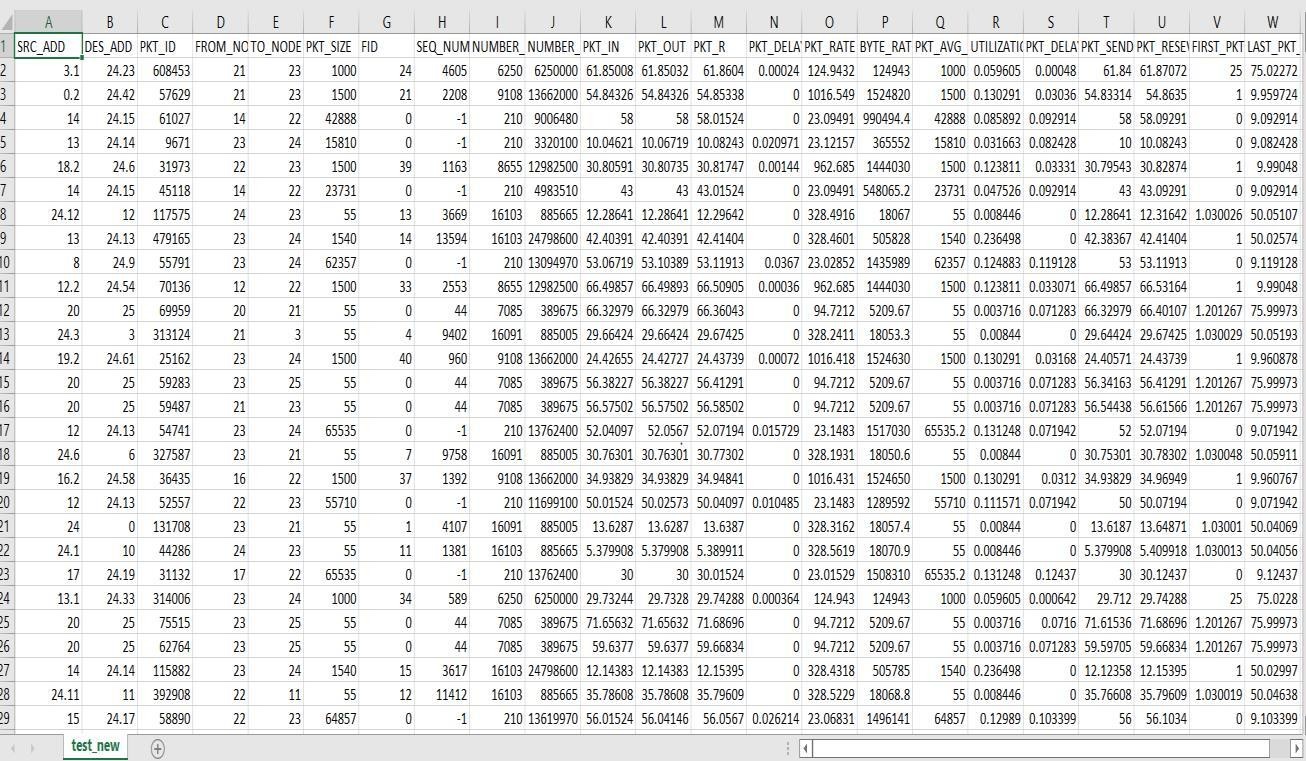
The dataset used in this project contains various field like source address, destination address, packet ID, source and destination node etc. The last column contains the type

of attack for a particular row.

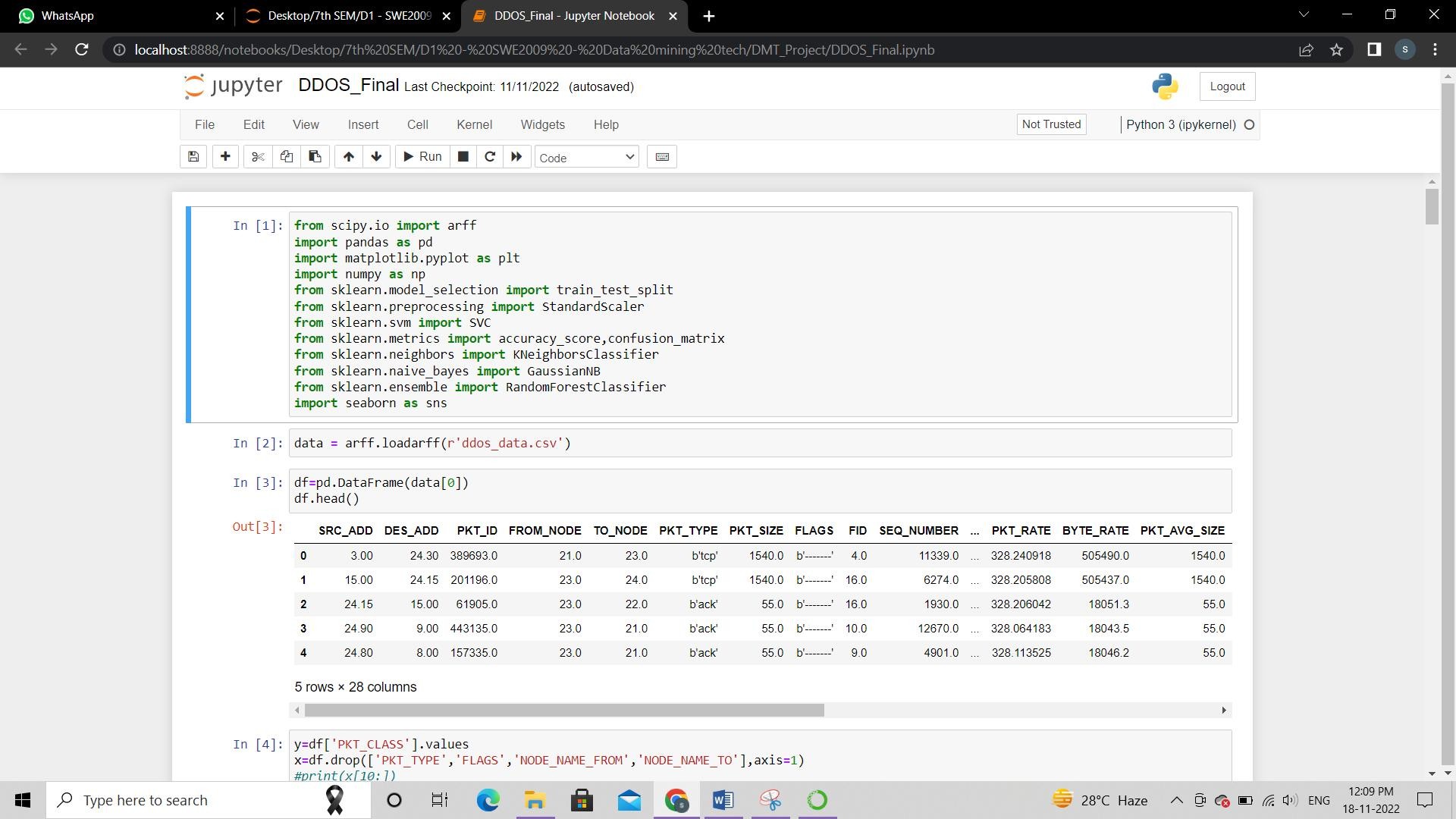
**GIVEN INPUT:** There are total five classes which are b'HTTP-FLOOD', b'Normal'

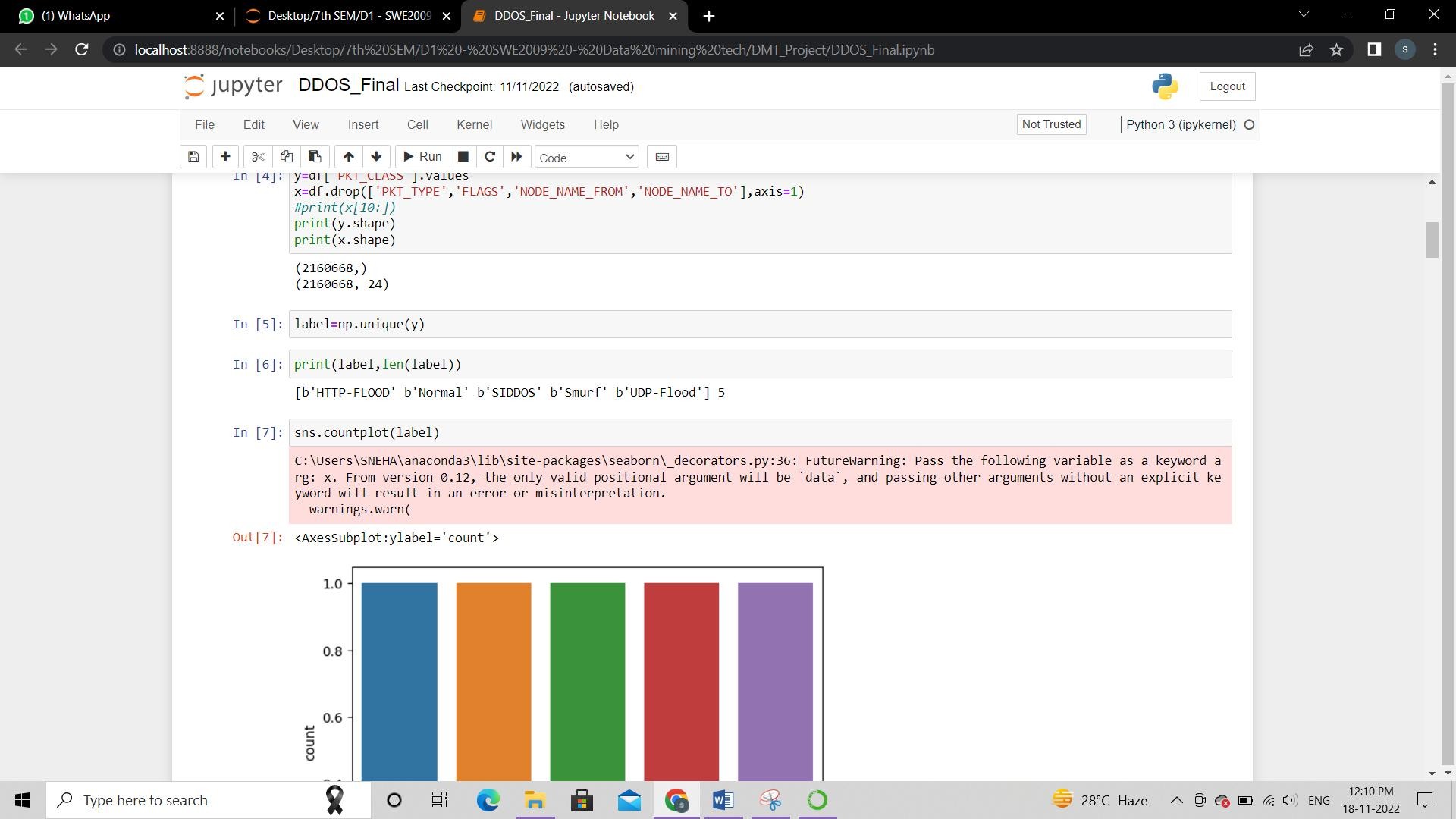
,b'SIDDOS', b'Smurf', b'UDP-Flood'.Our model has to predict one out of these five class.

Sample of our dataset:

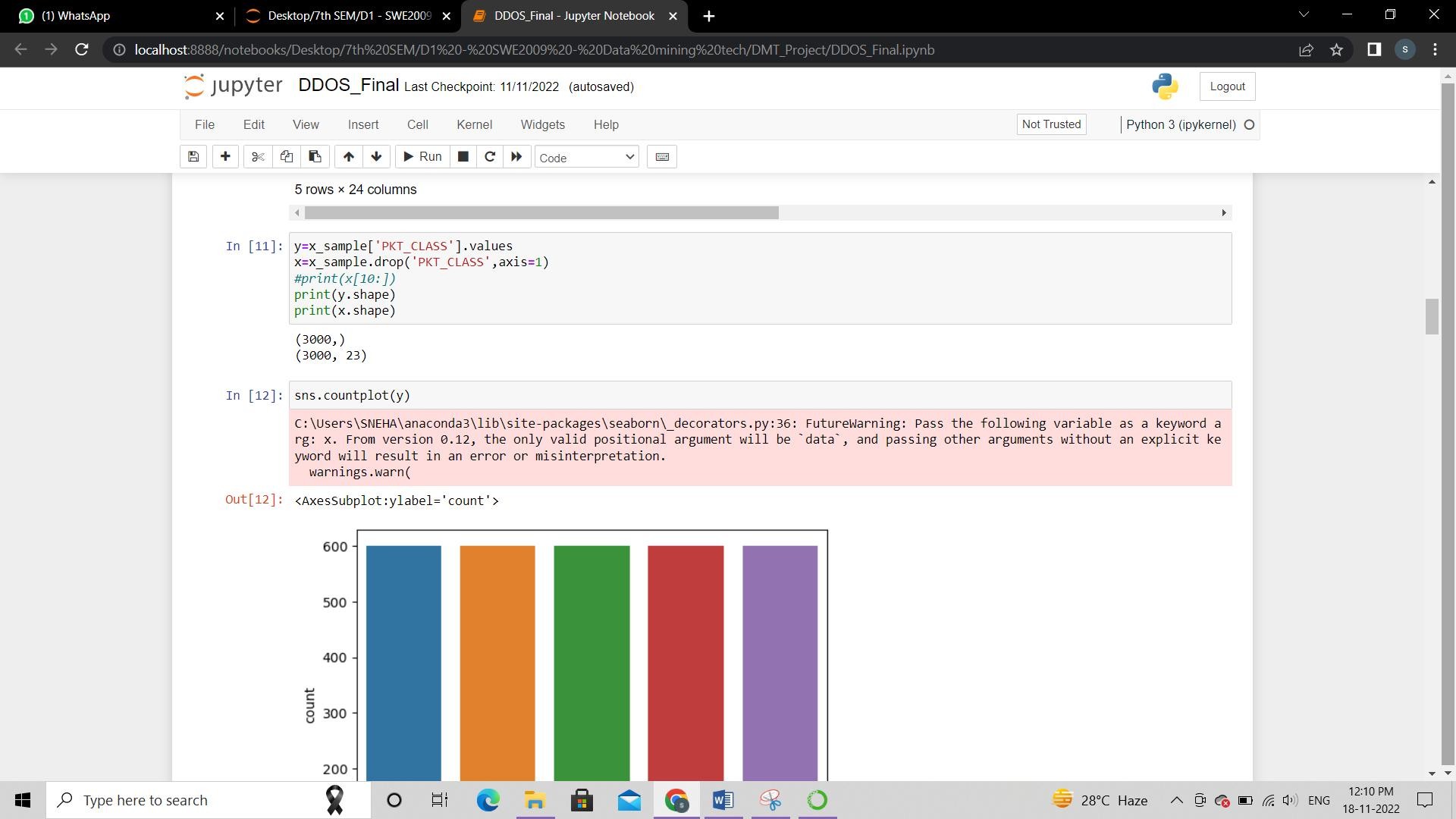


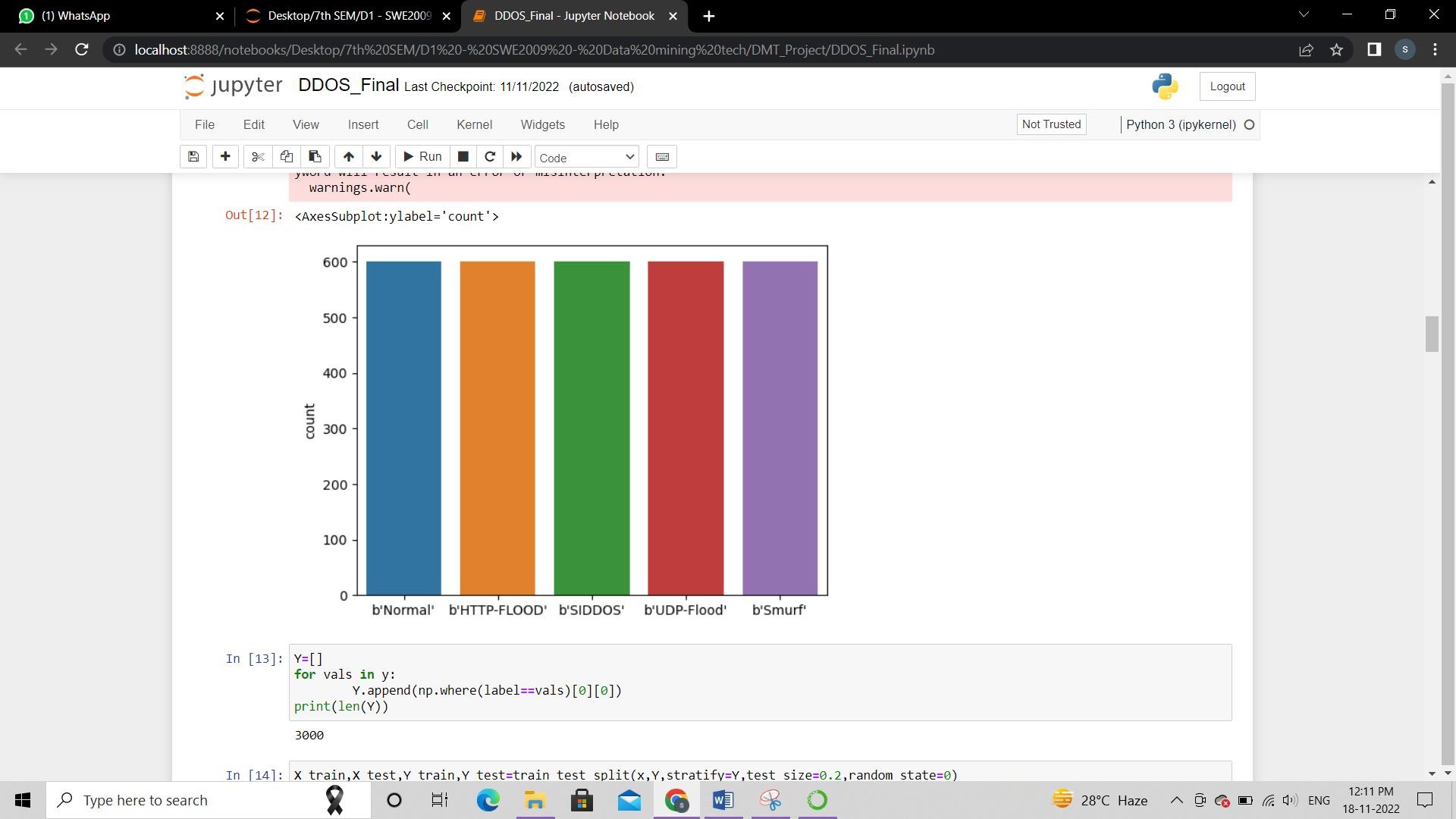
JUPYTER NOTEBOOK SOURCE CODE AND OUTPUT :

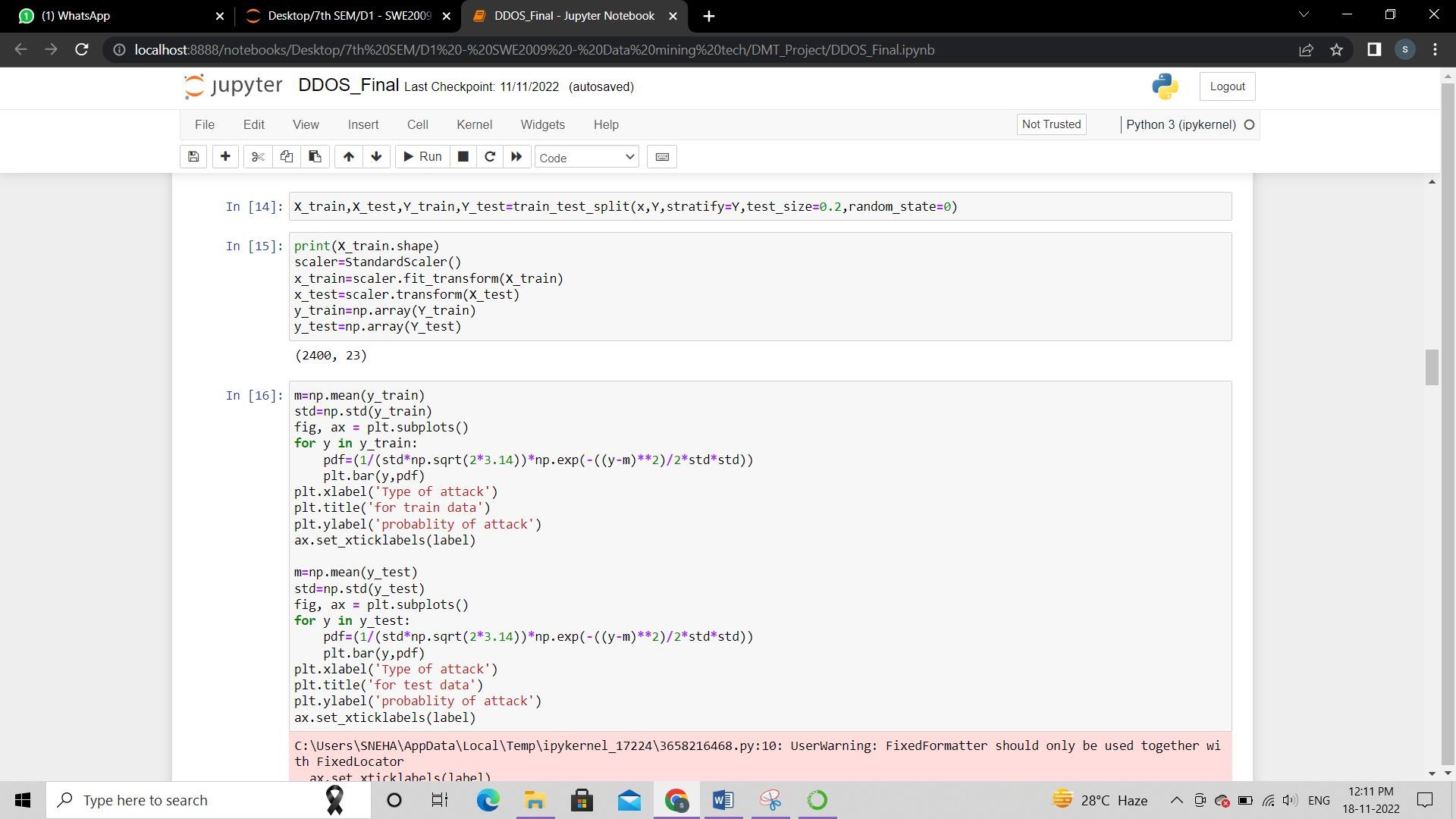


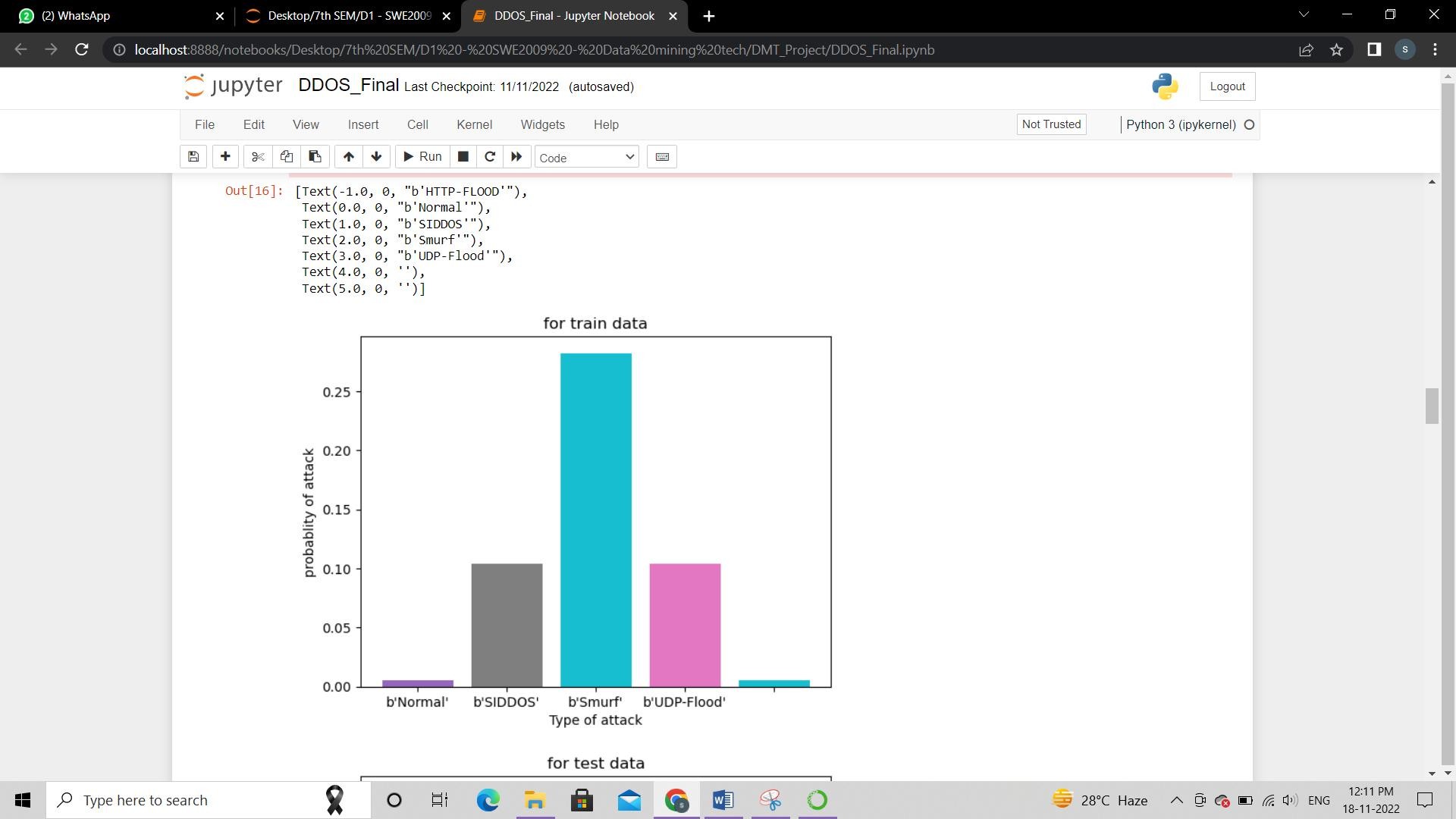


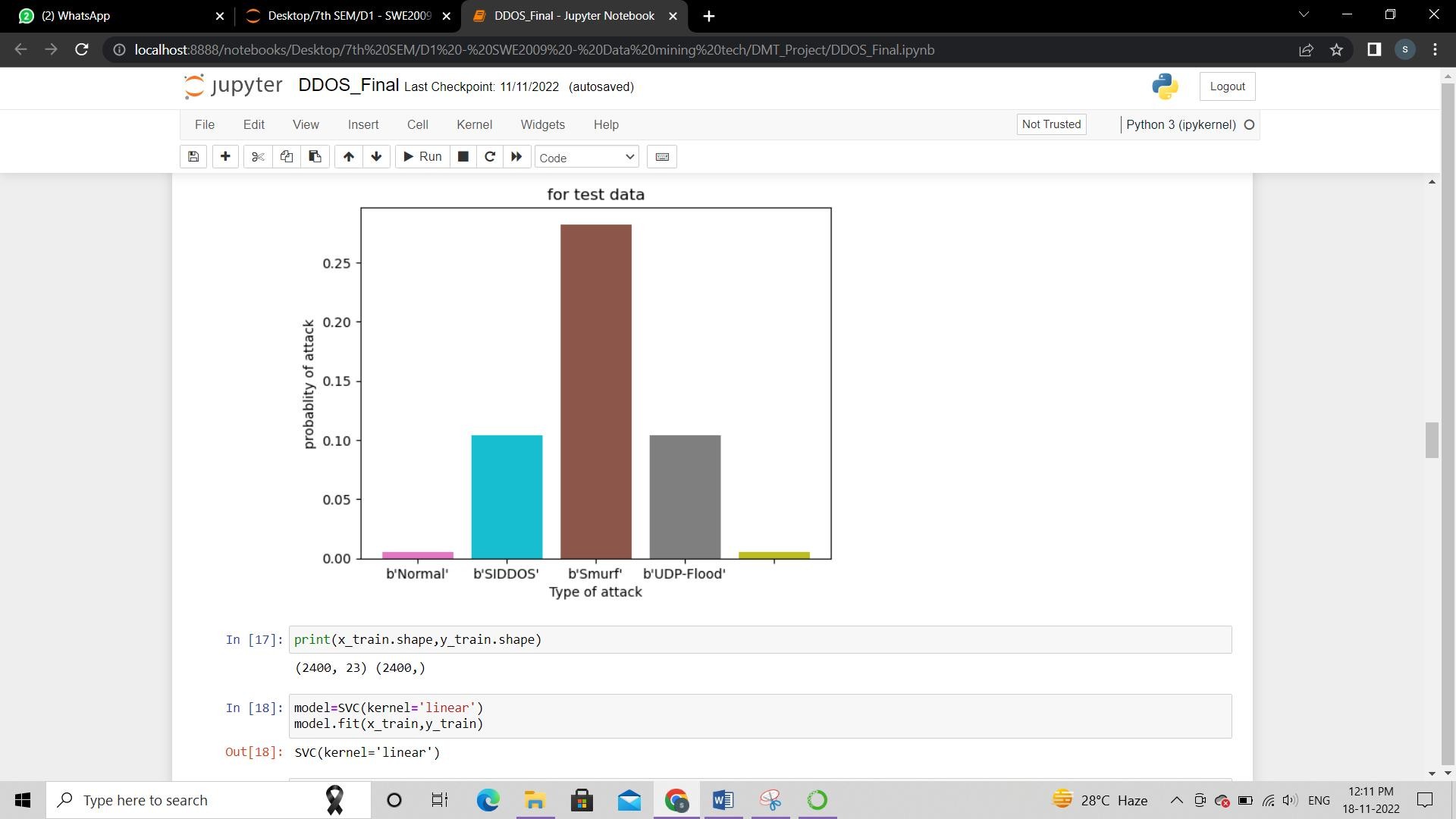


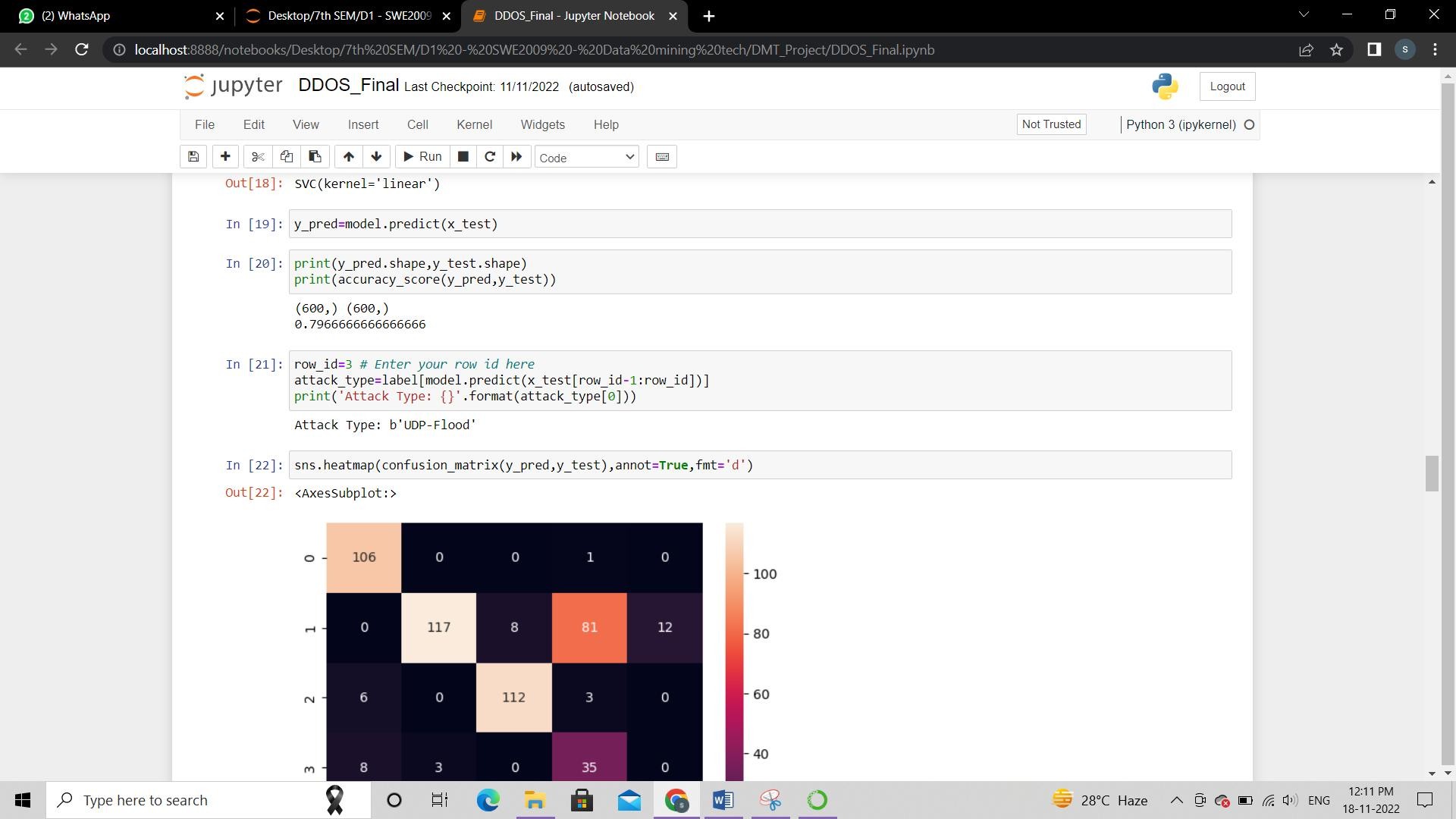


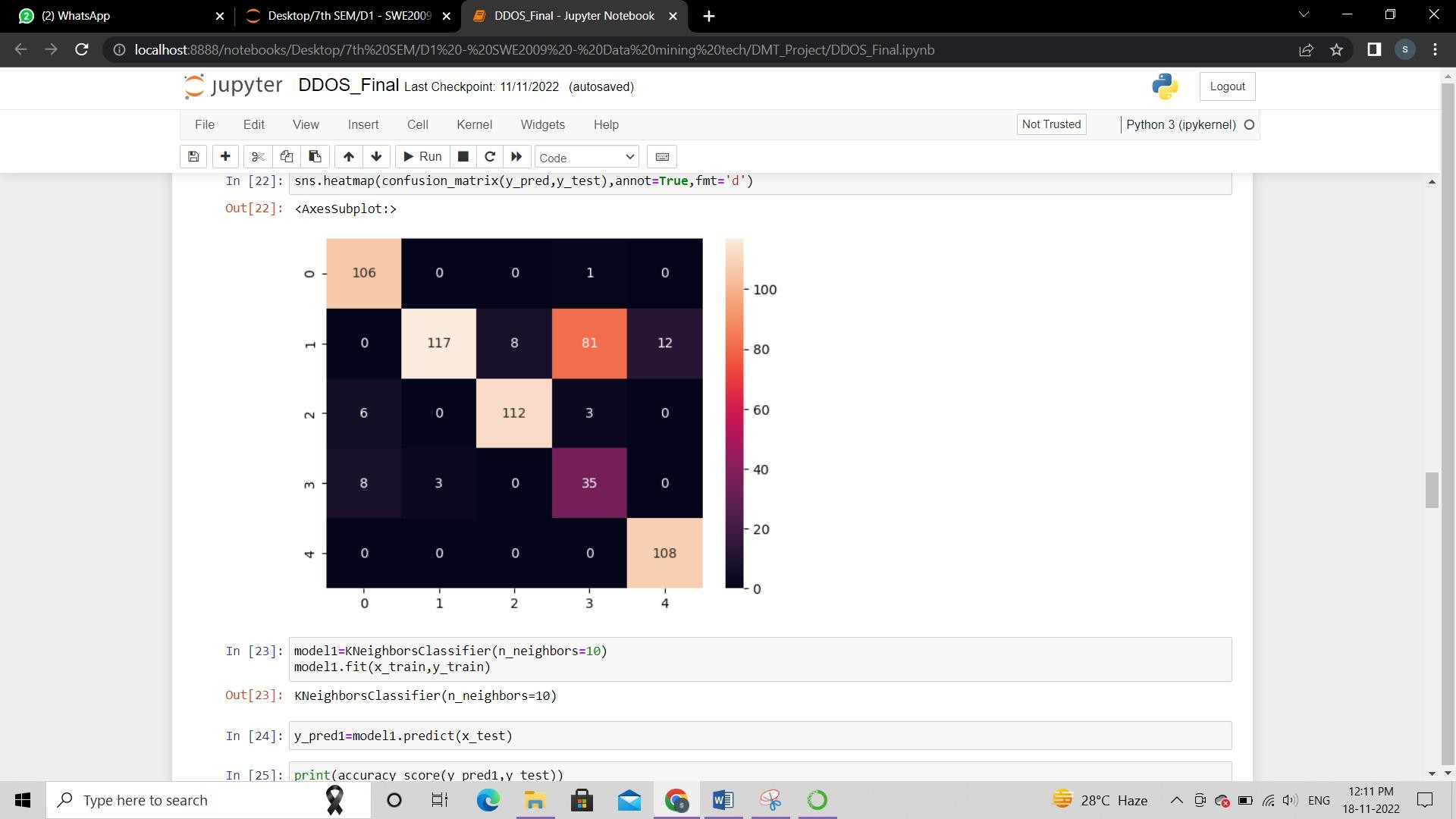


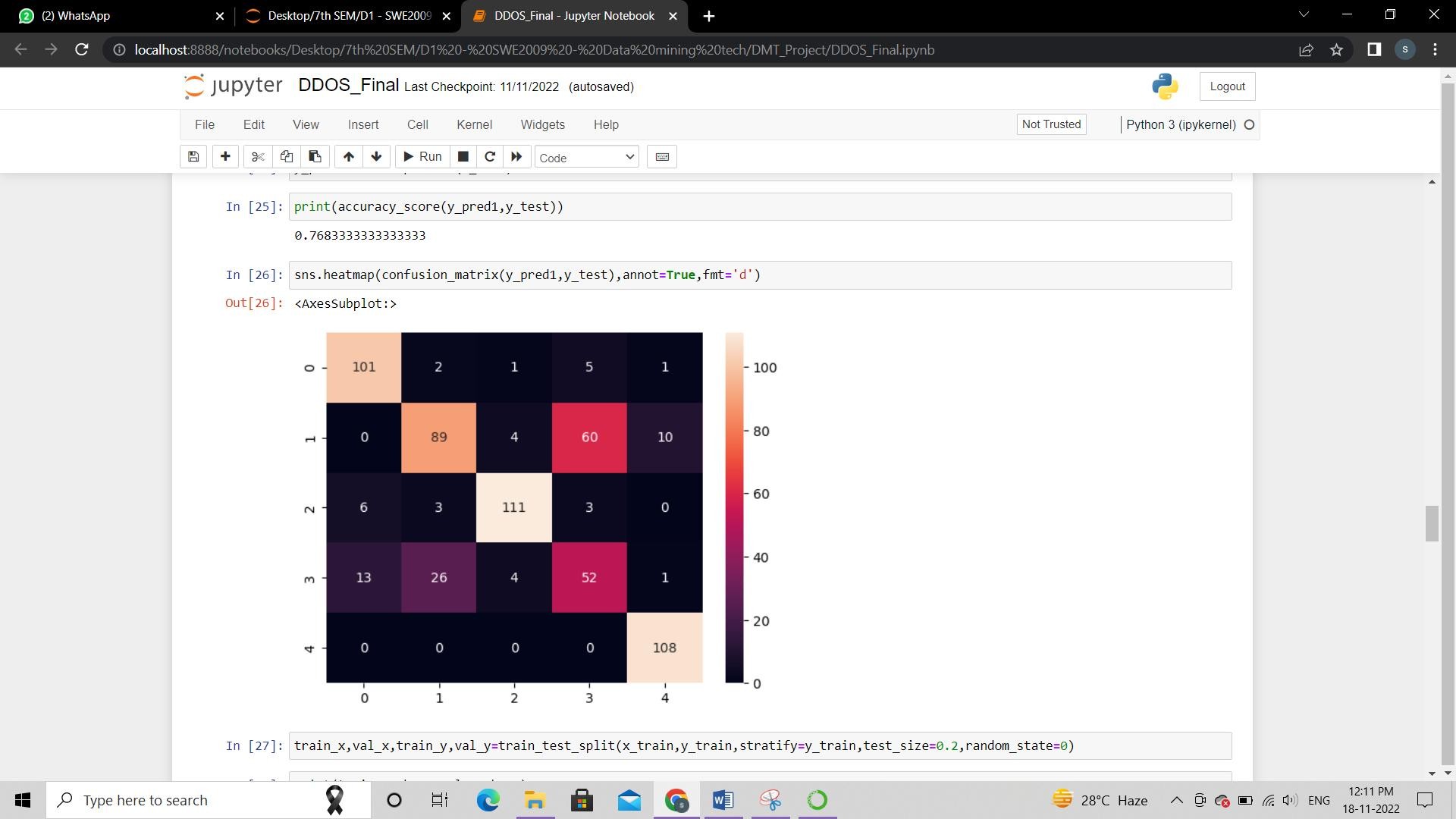


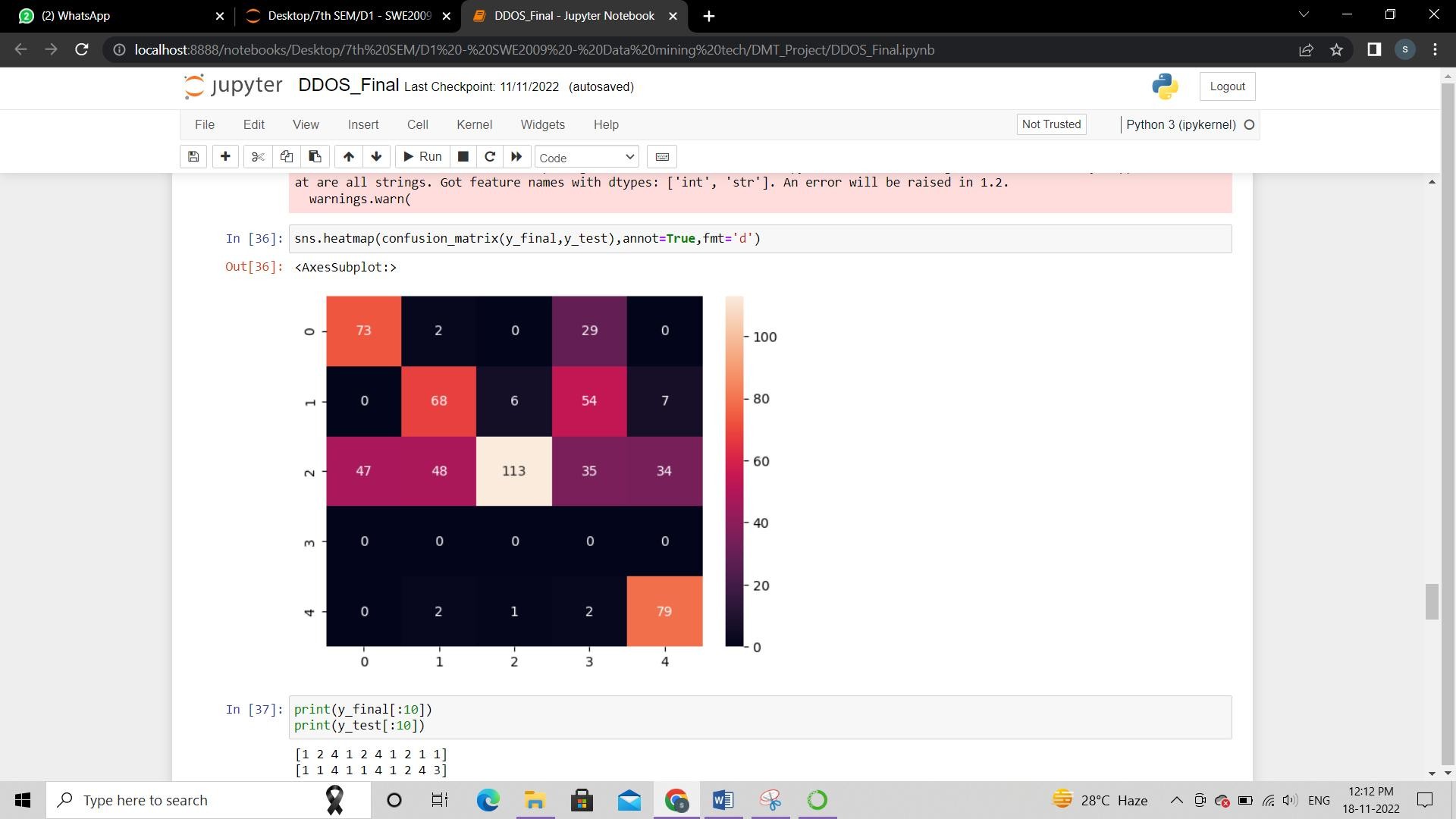


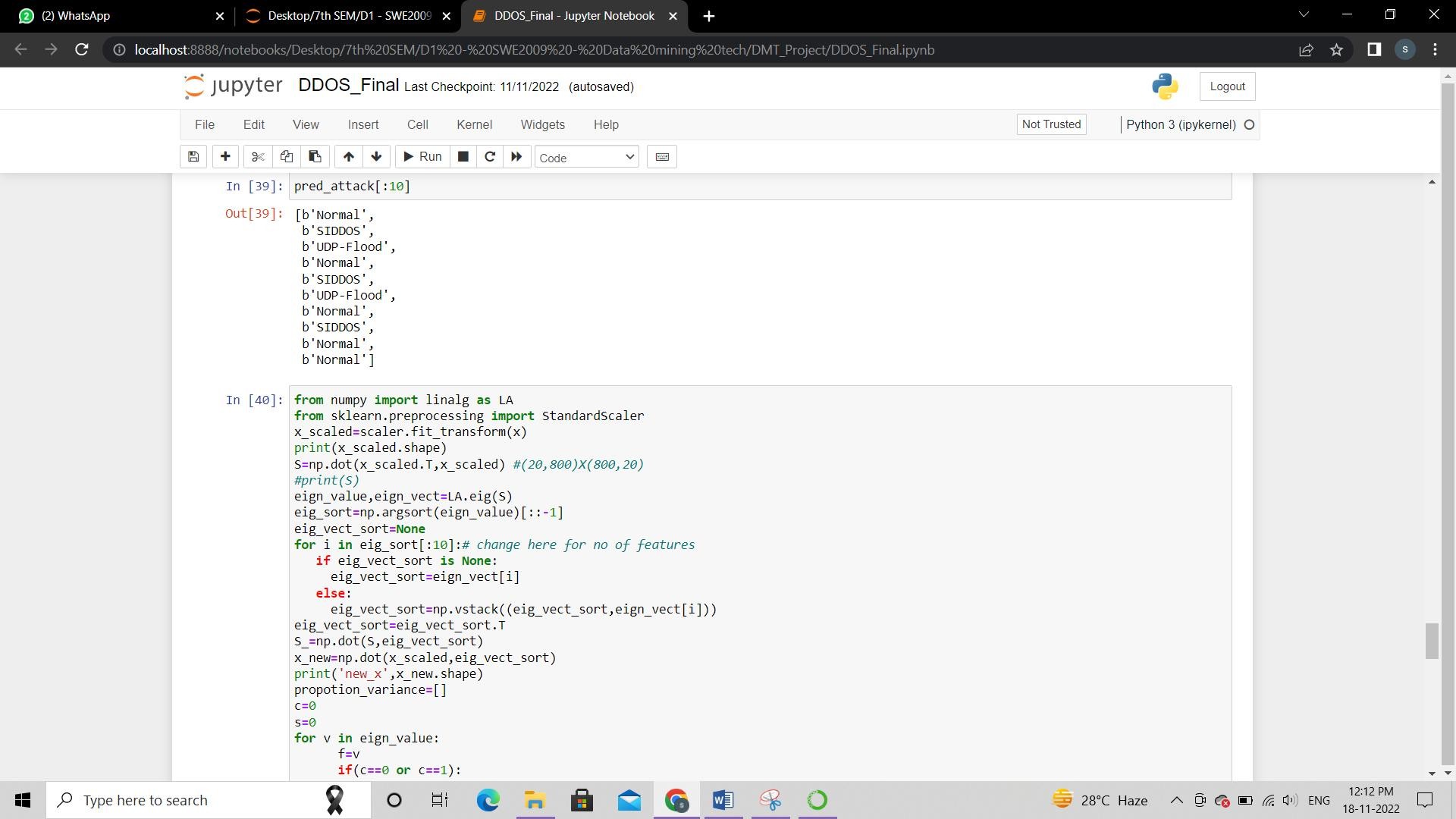


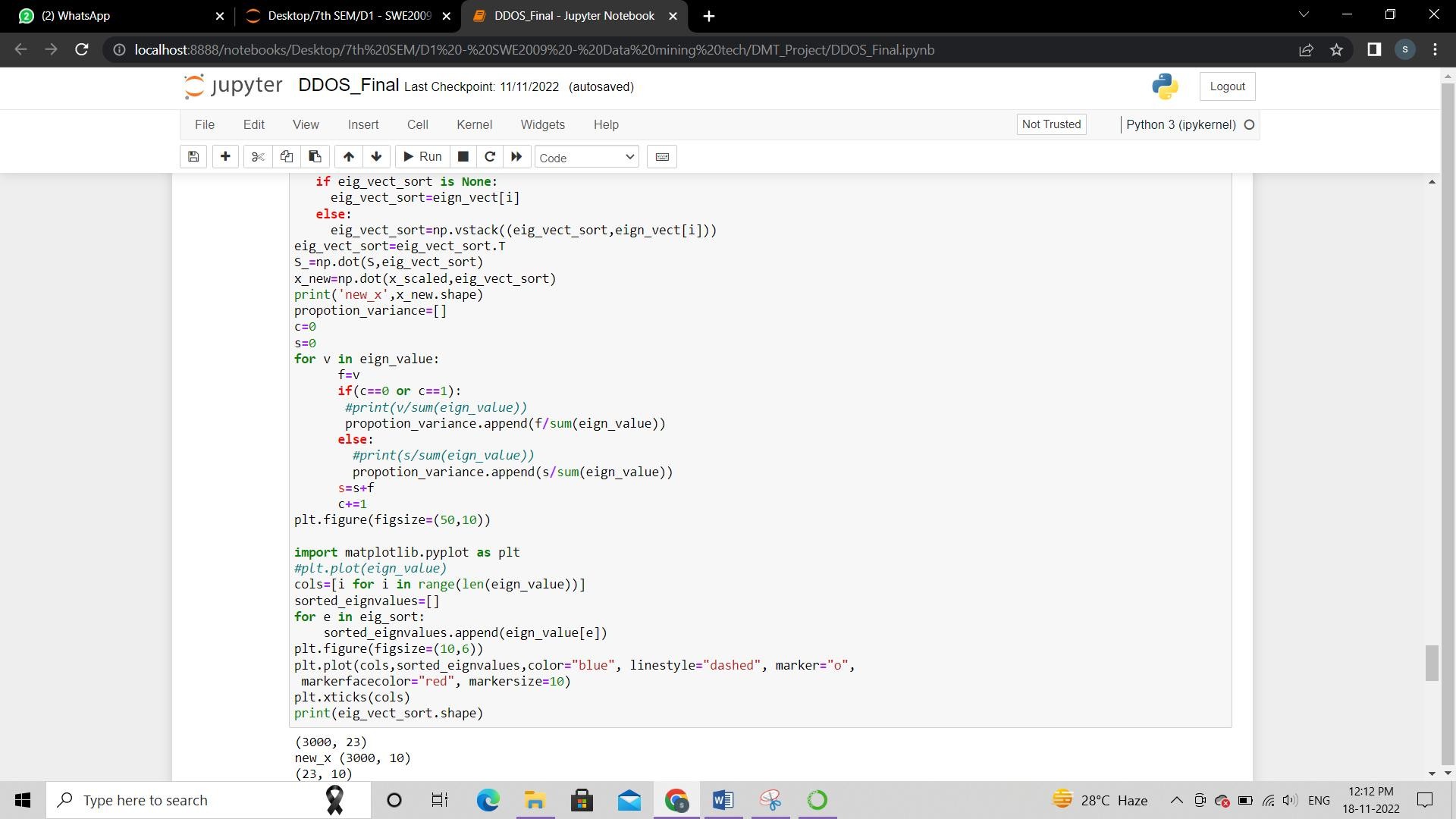


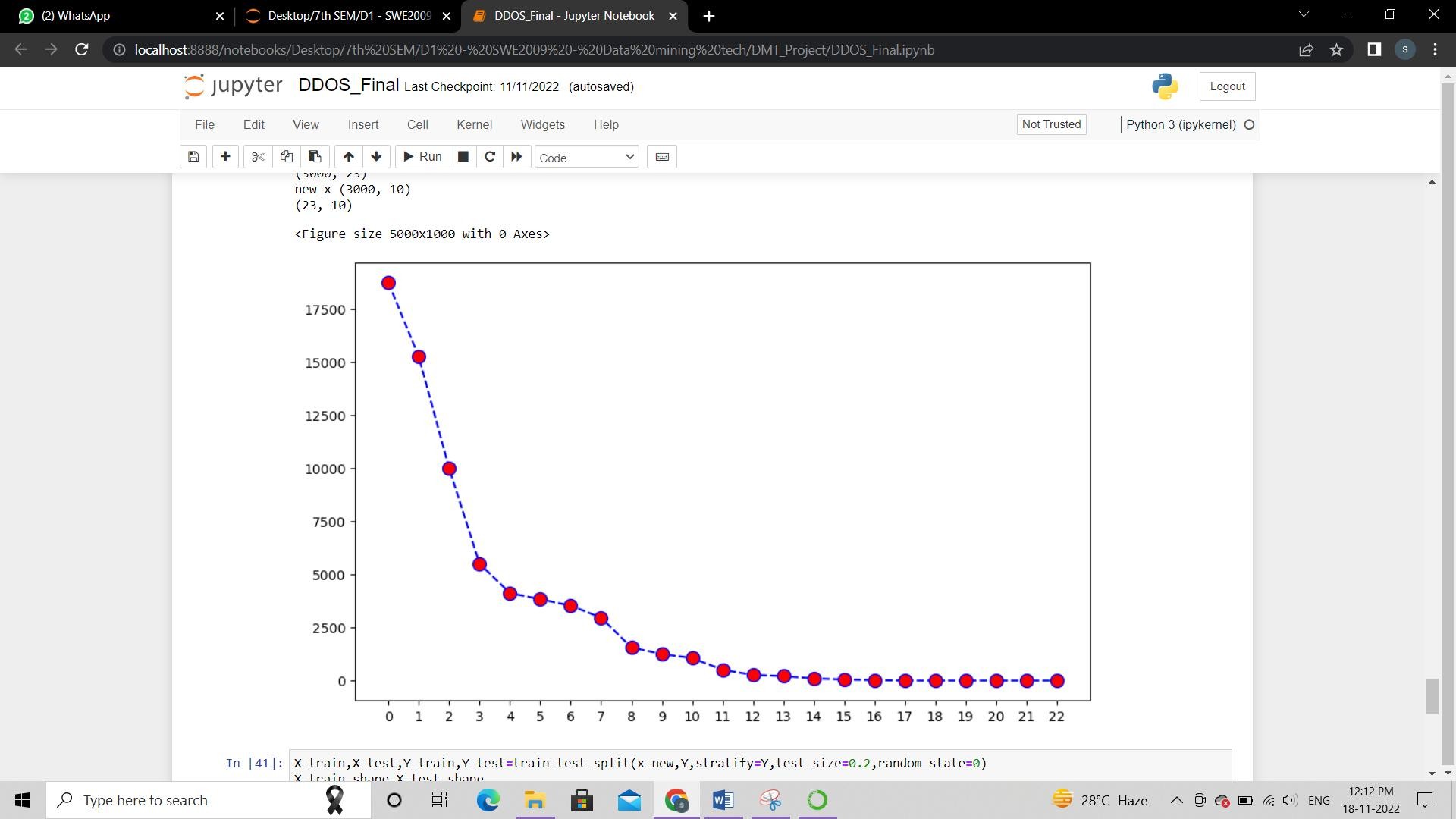


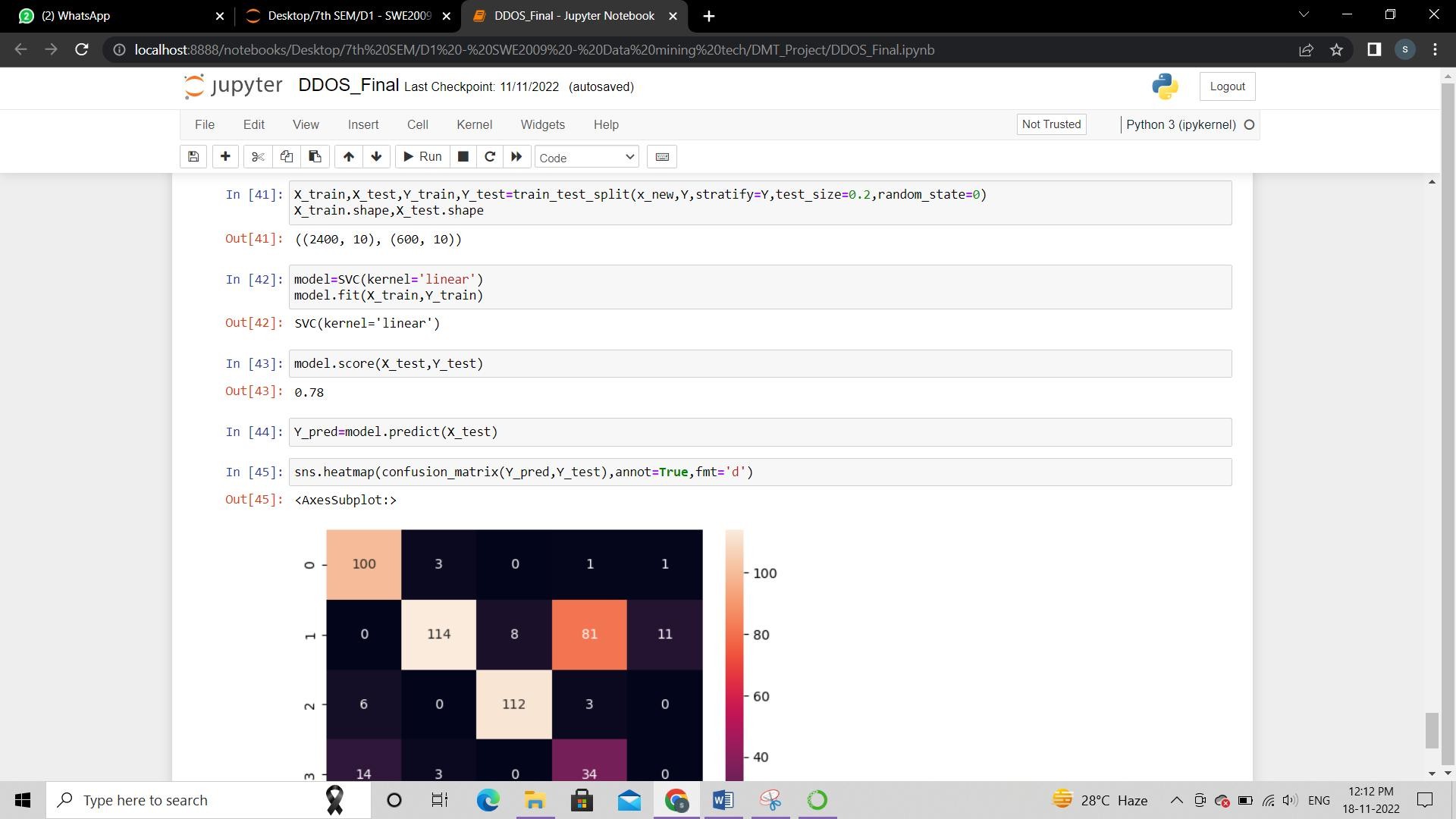




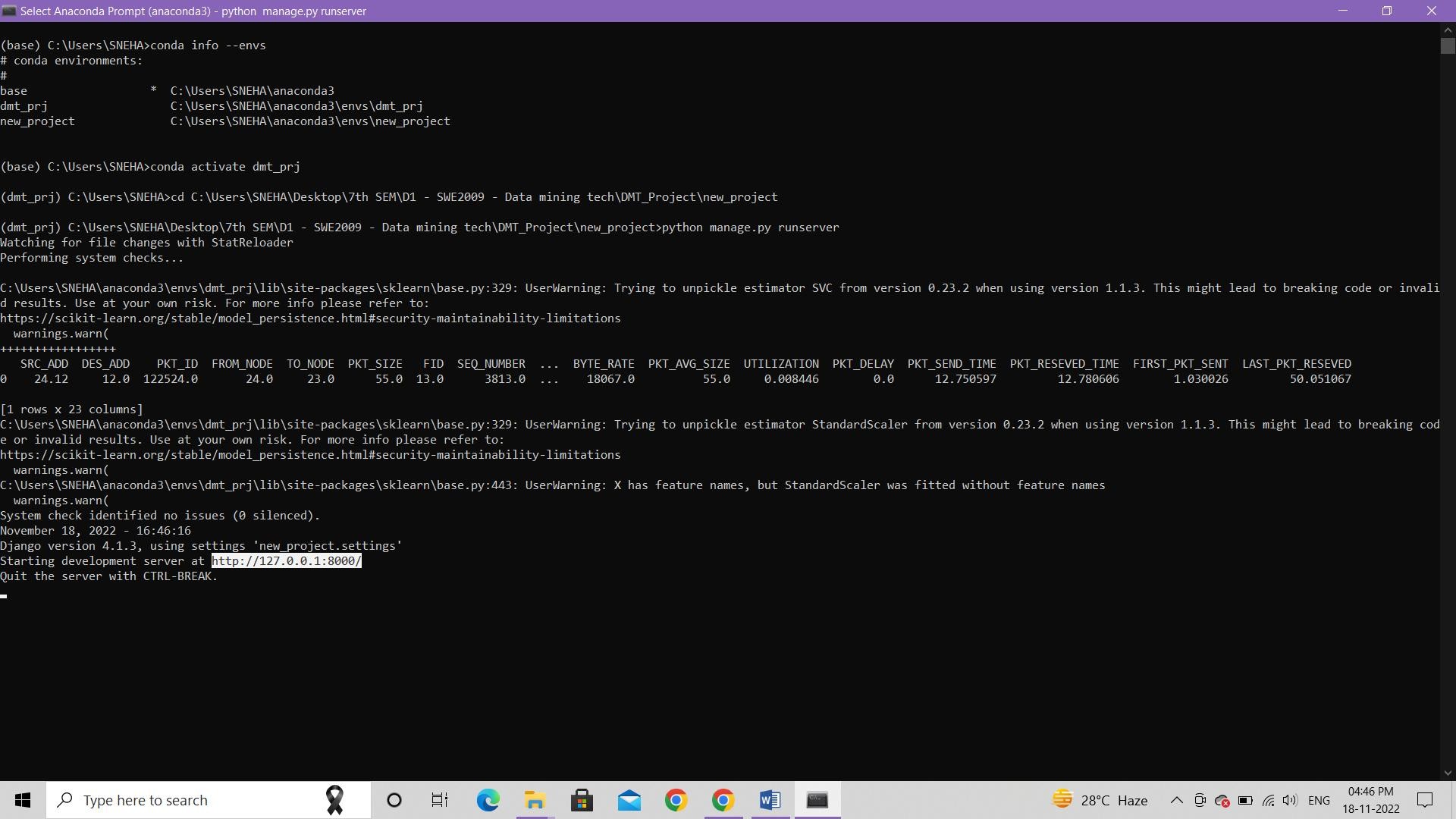


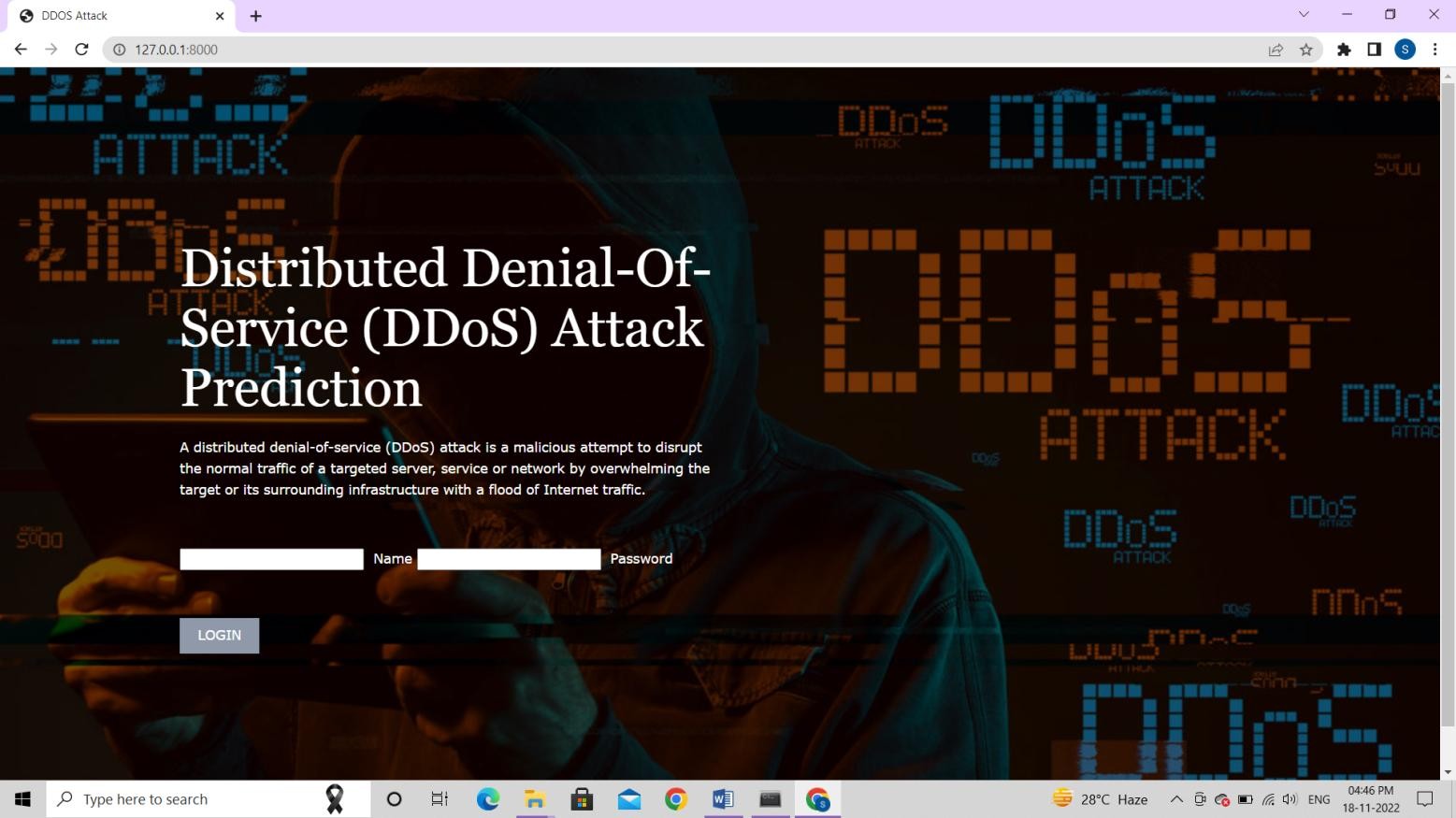


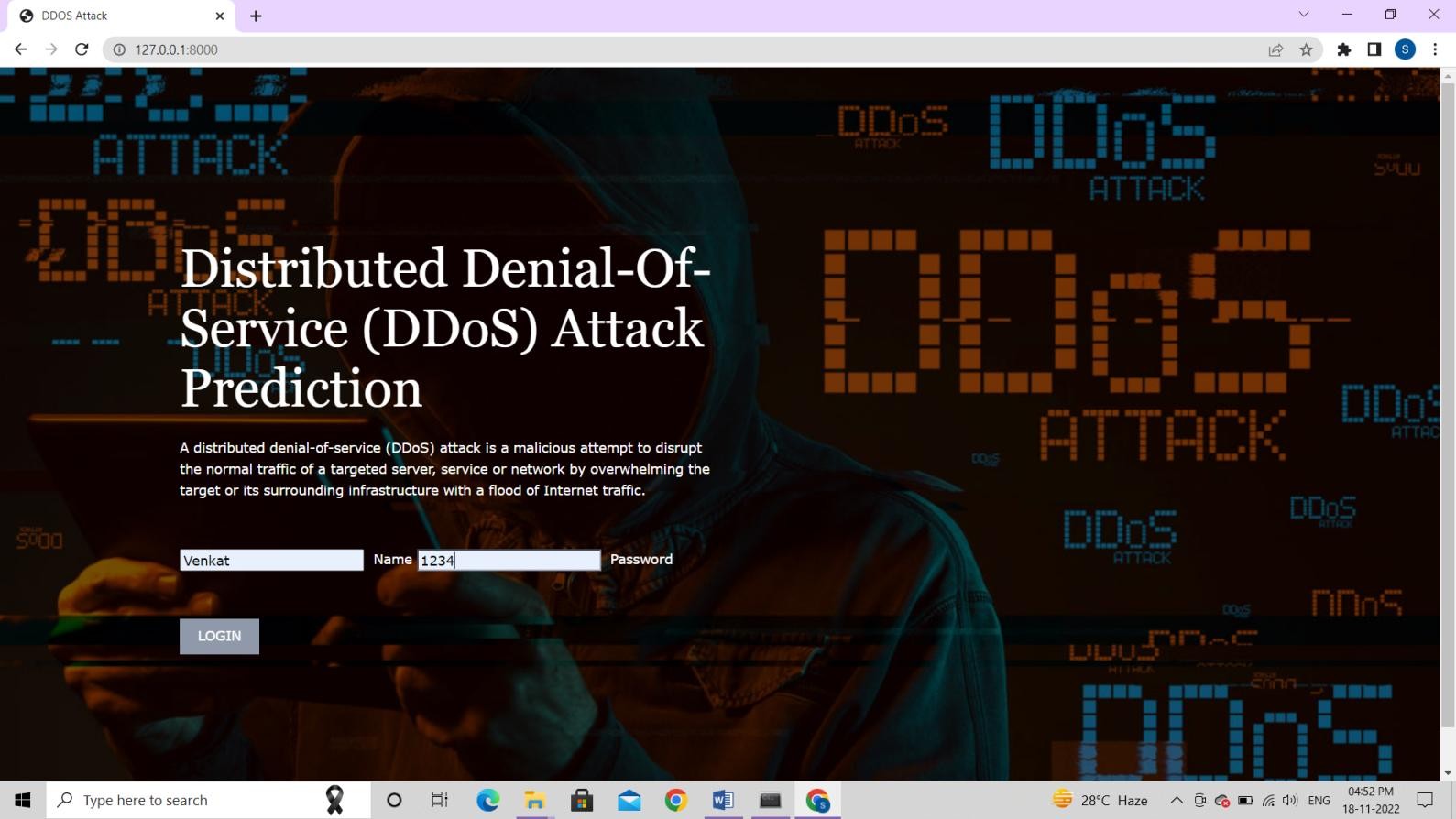


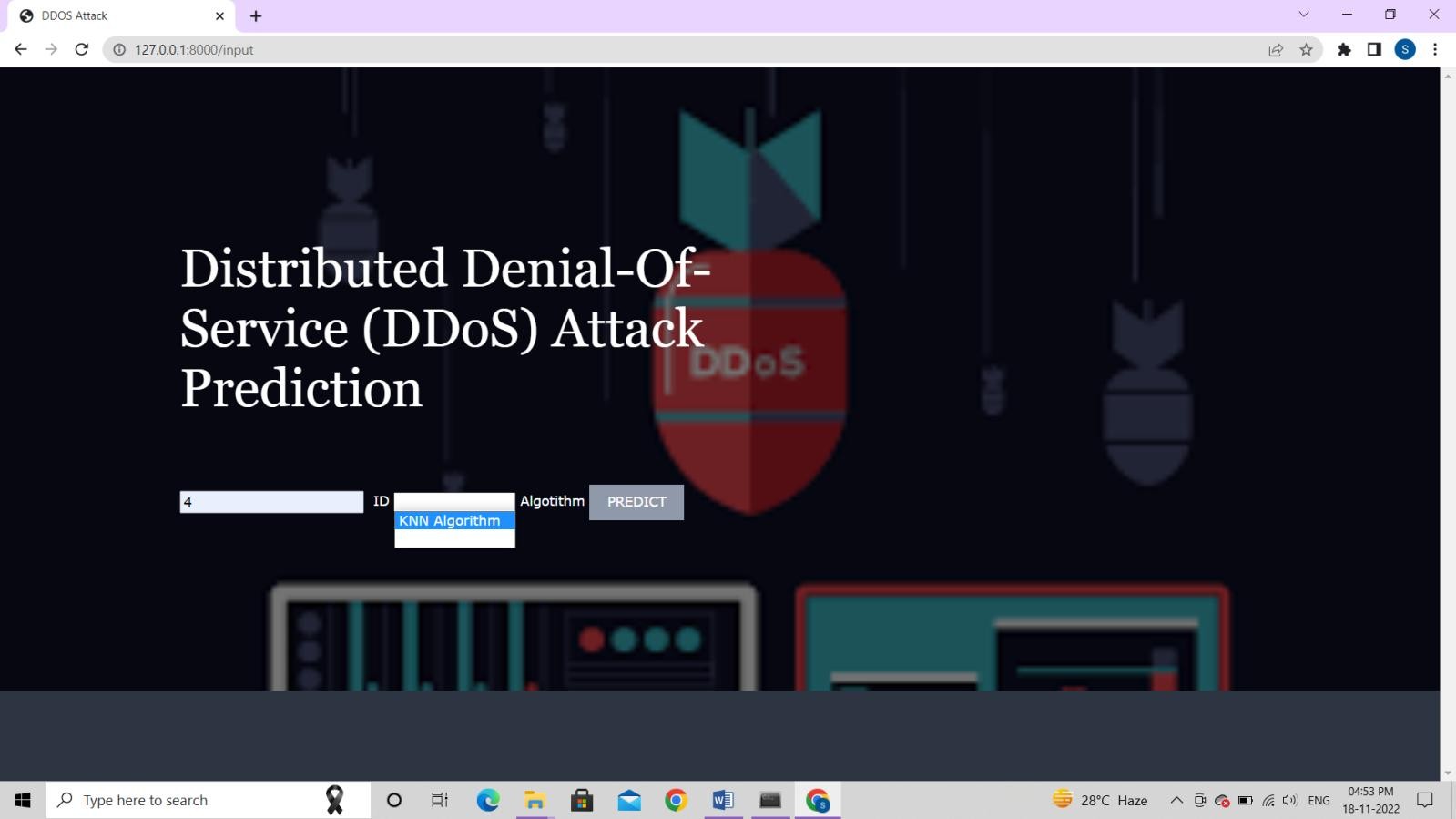


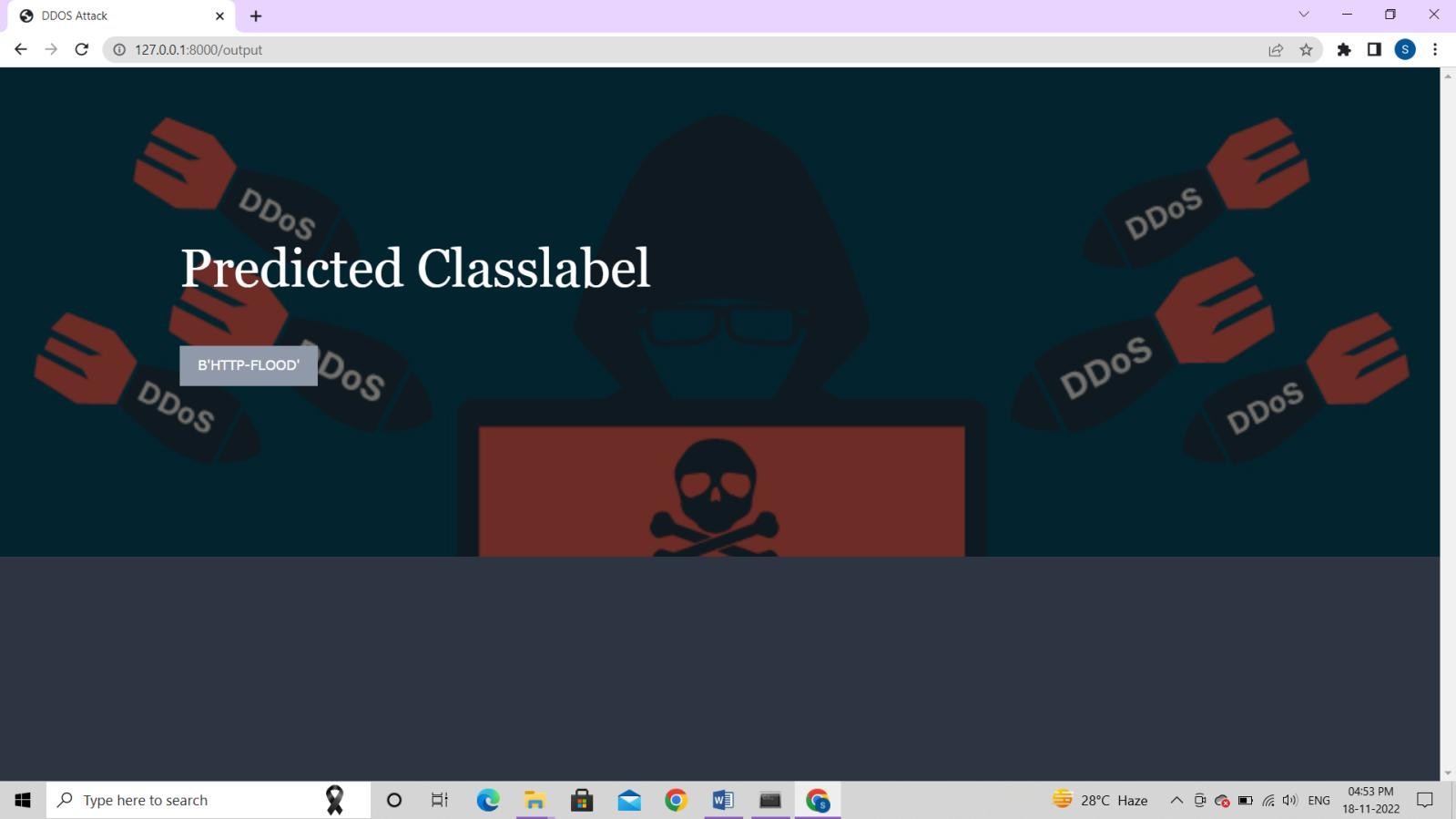
APPLICATION OUTPUT

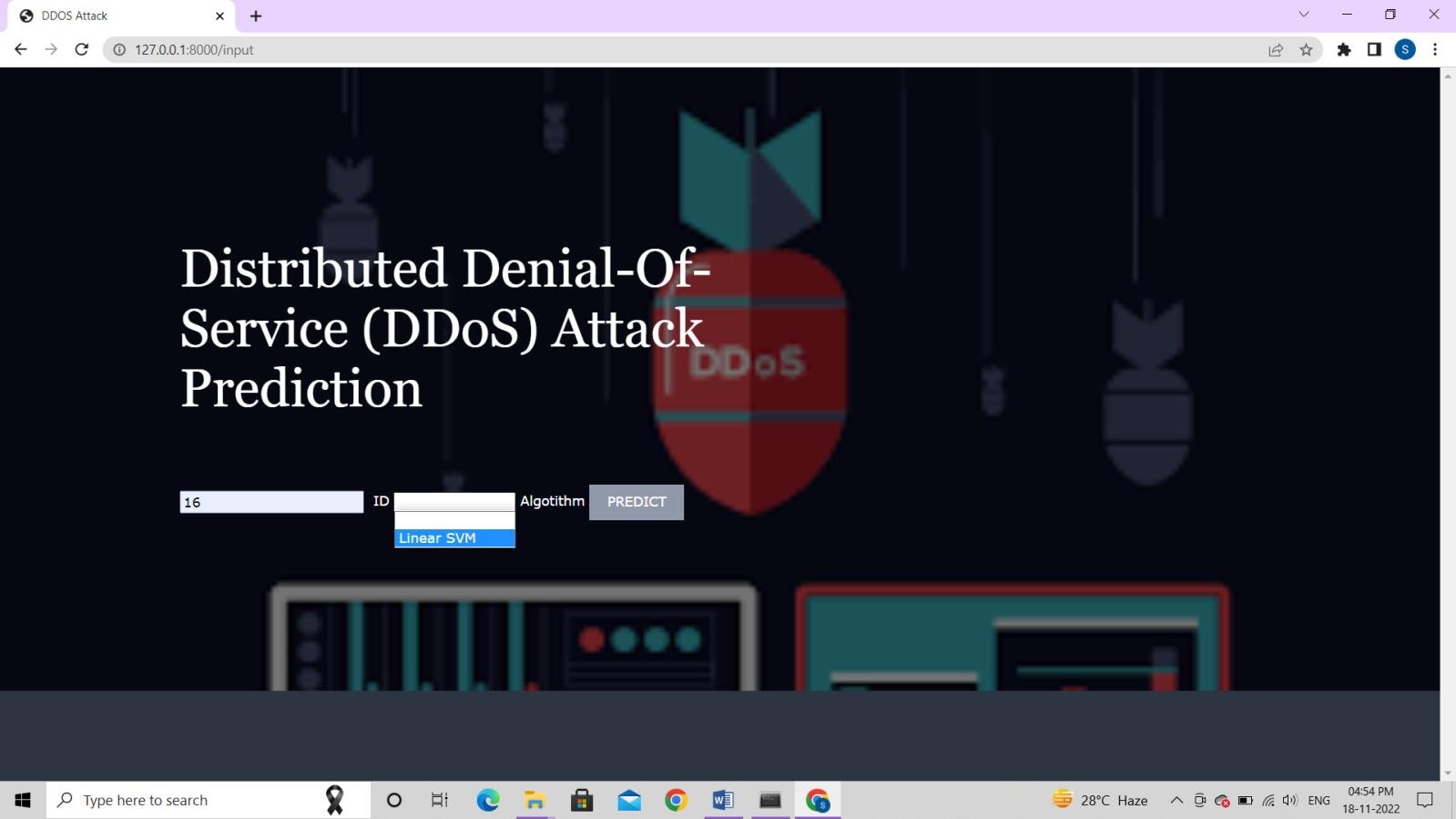


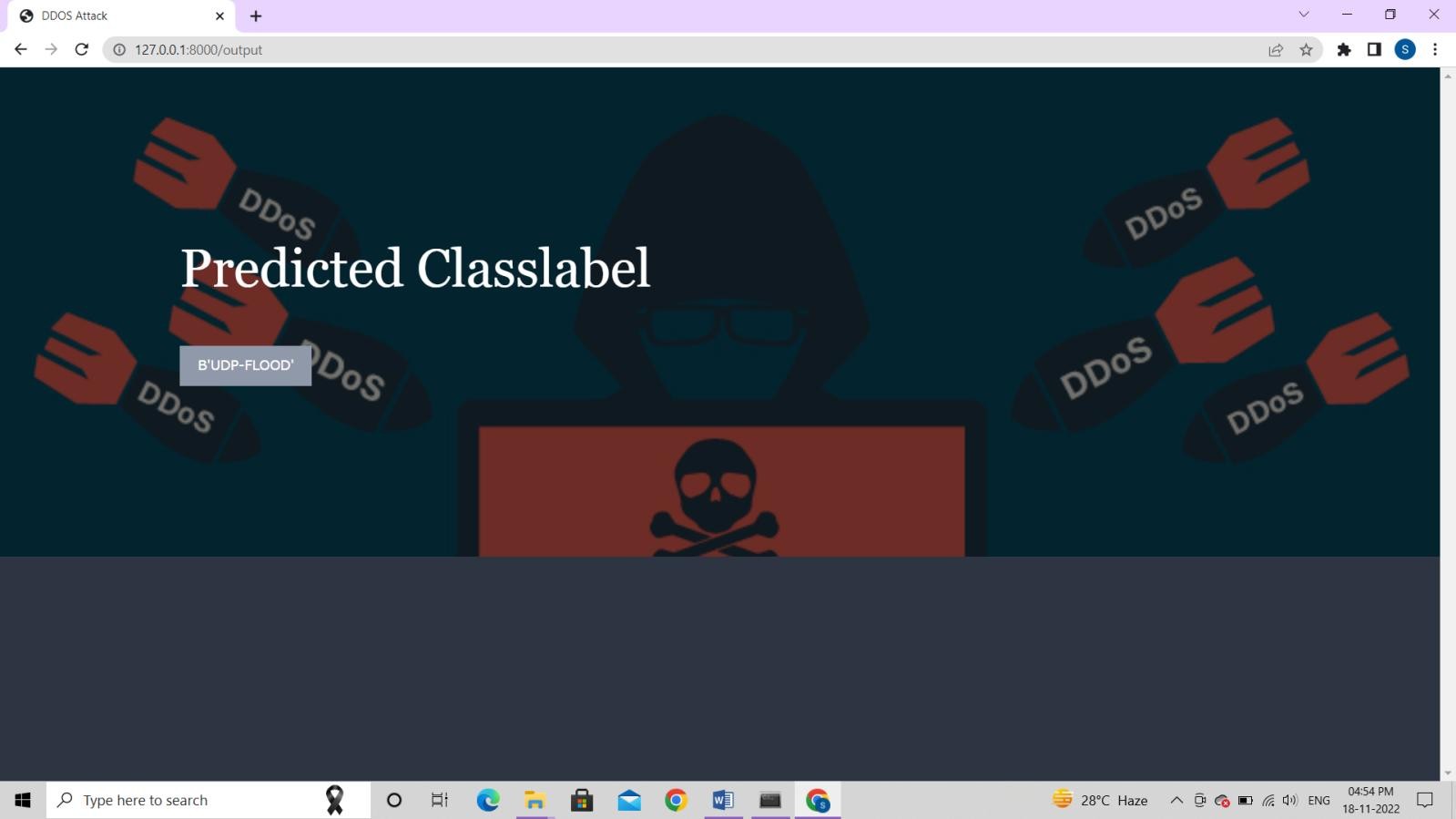


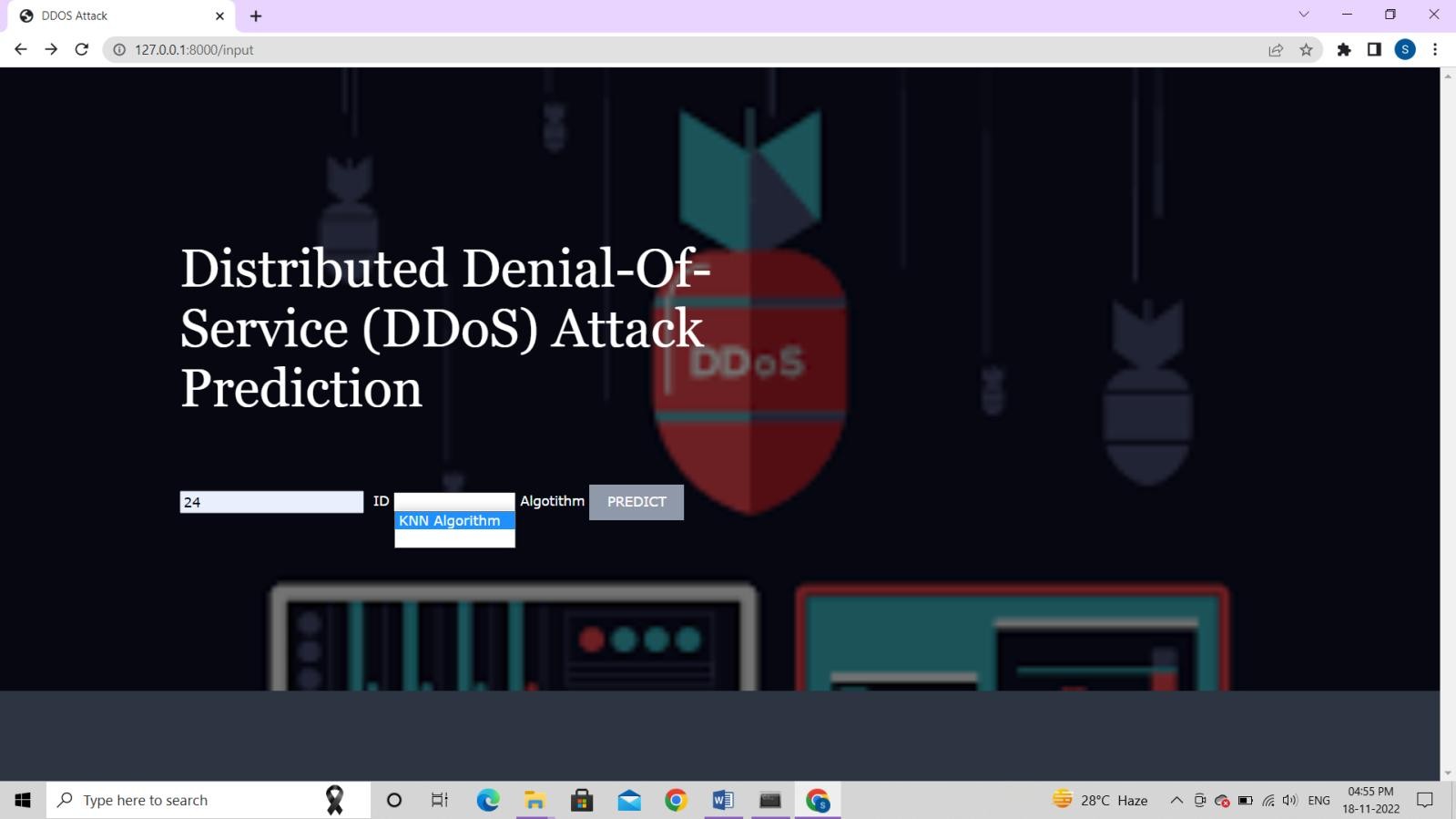


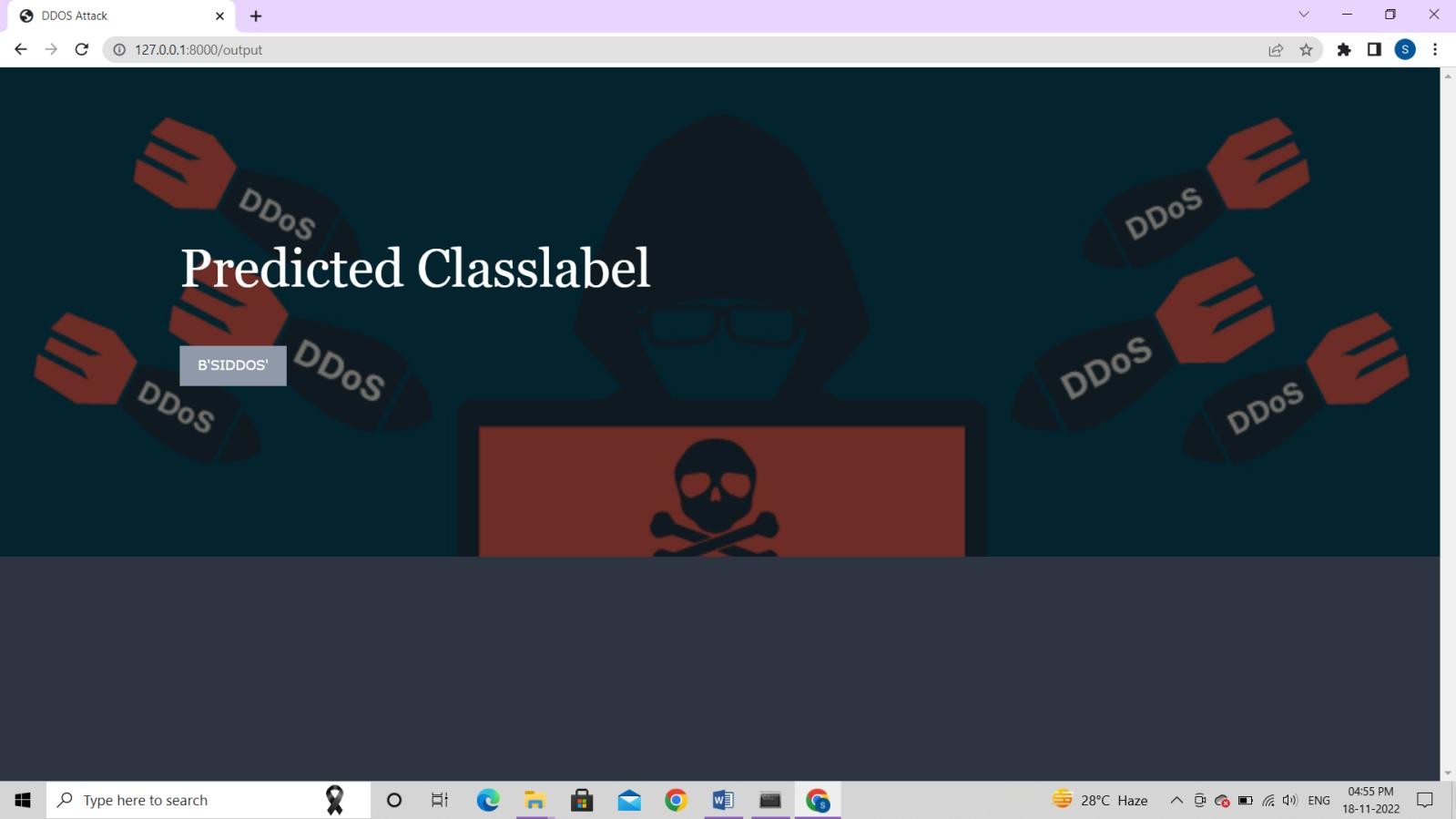


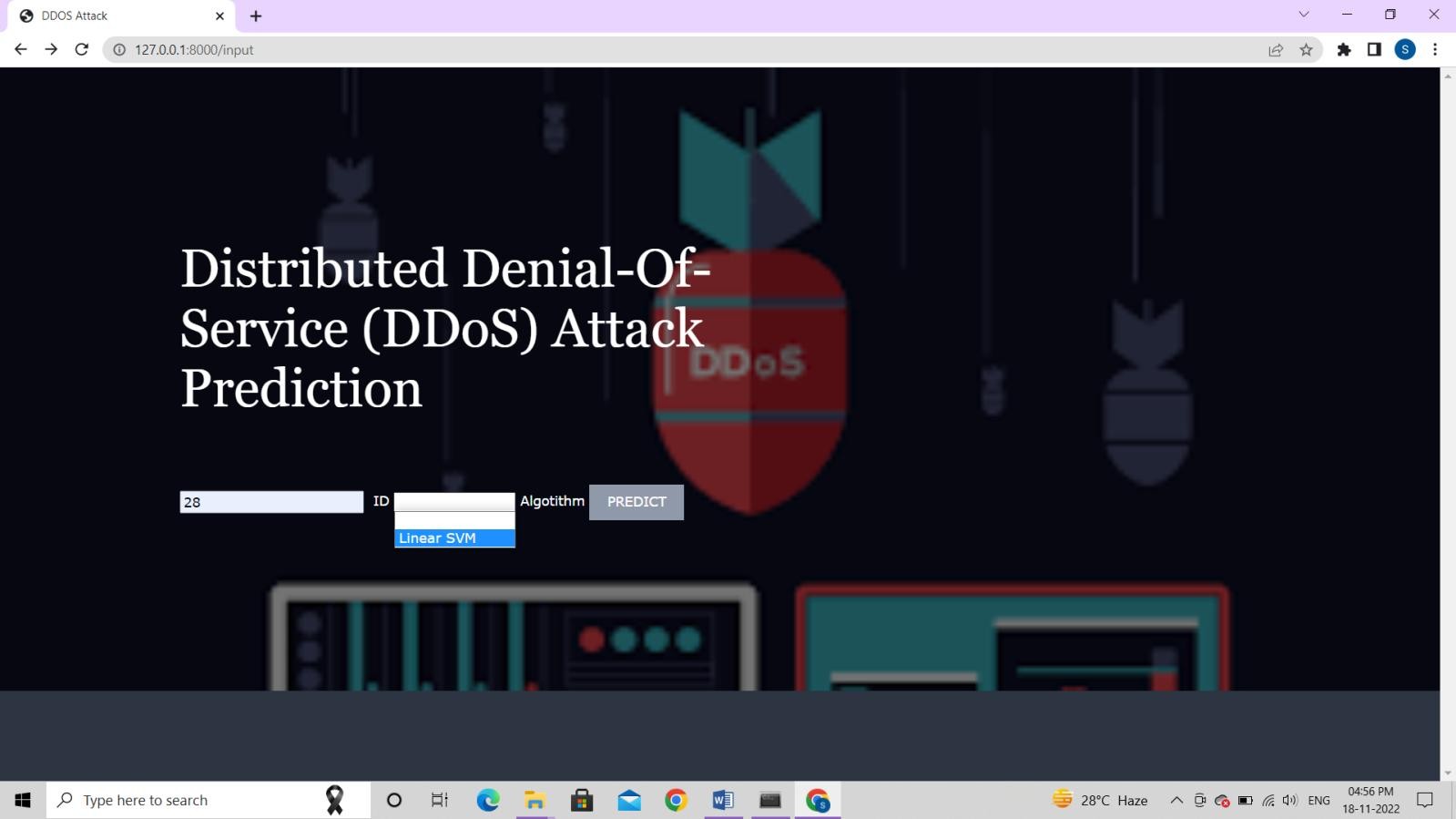


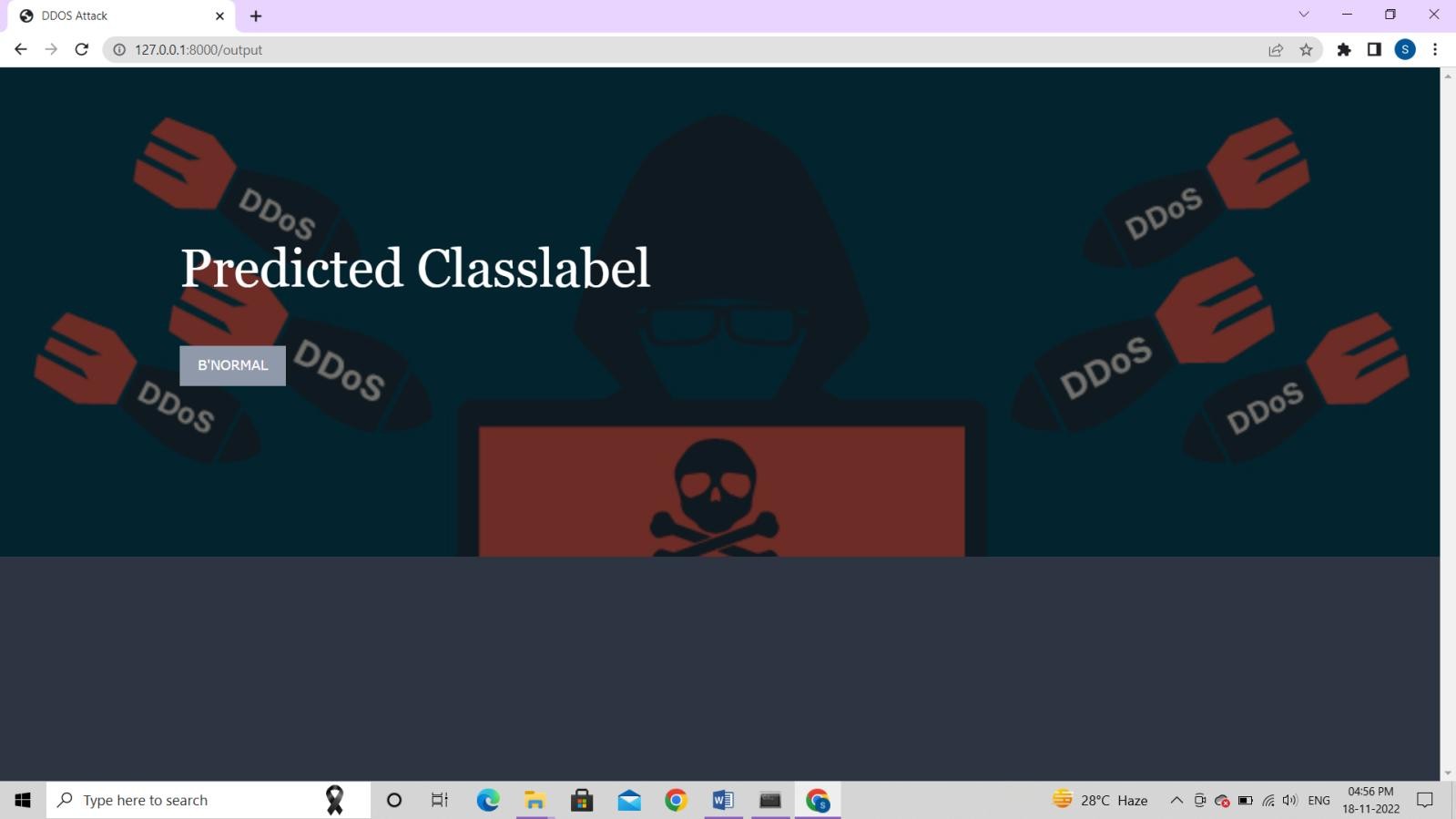


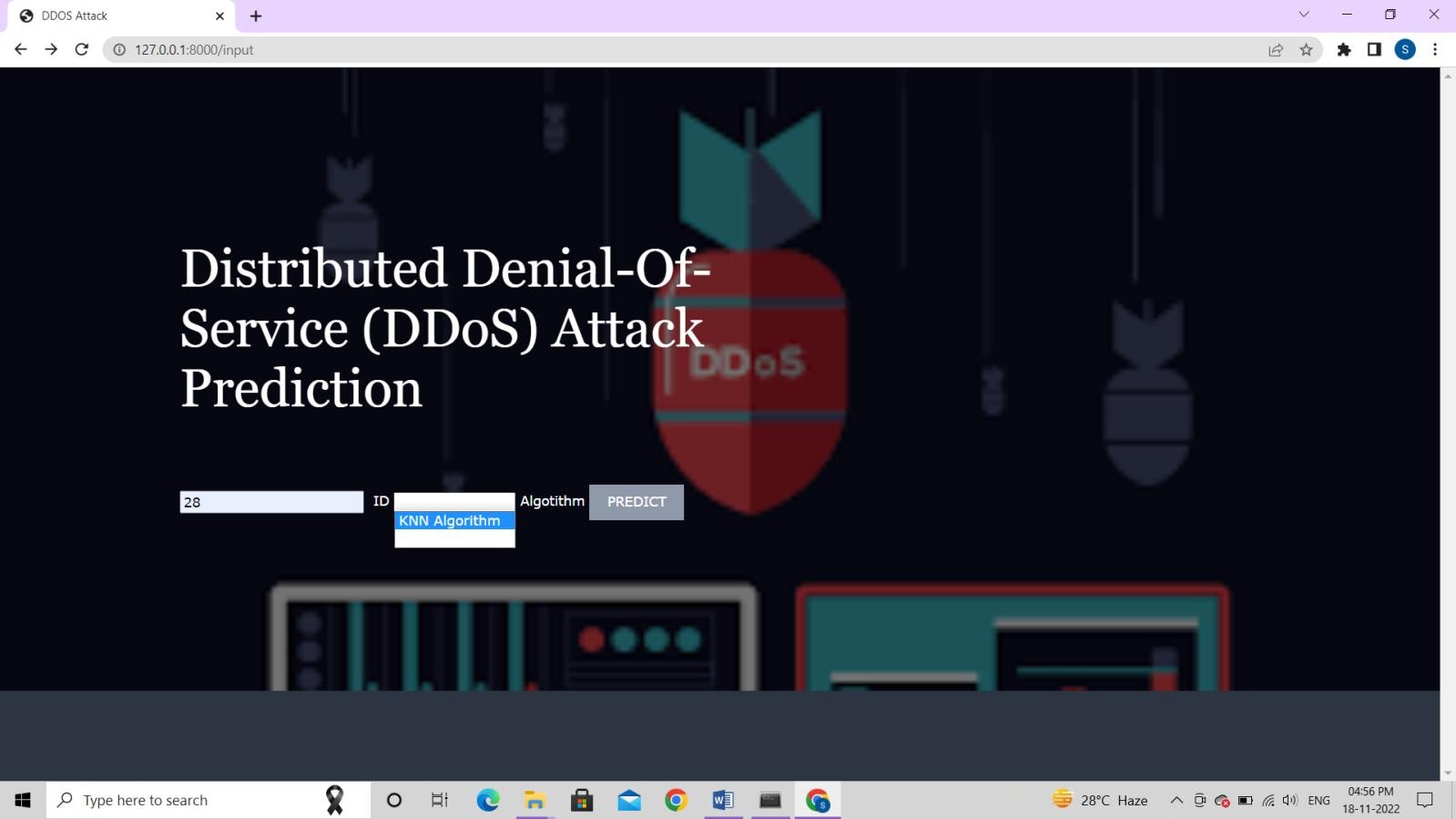


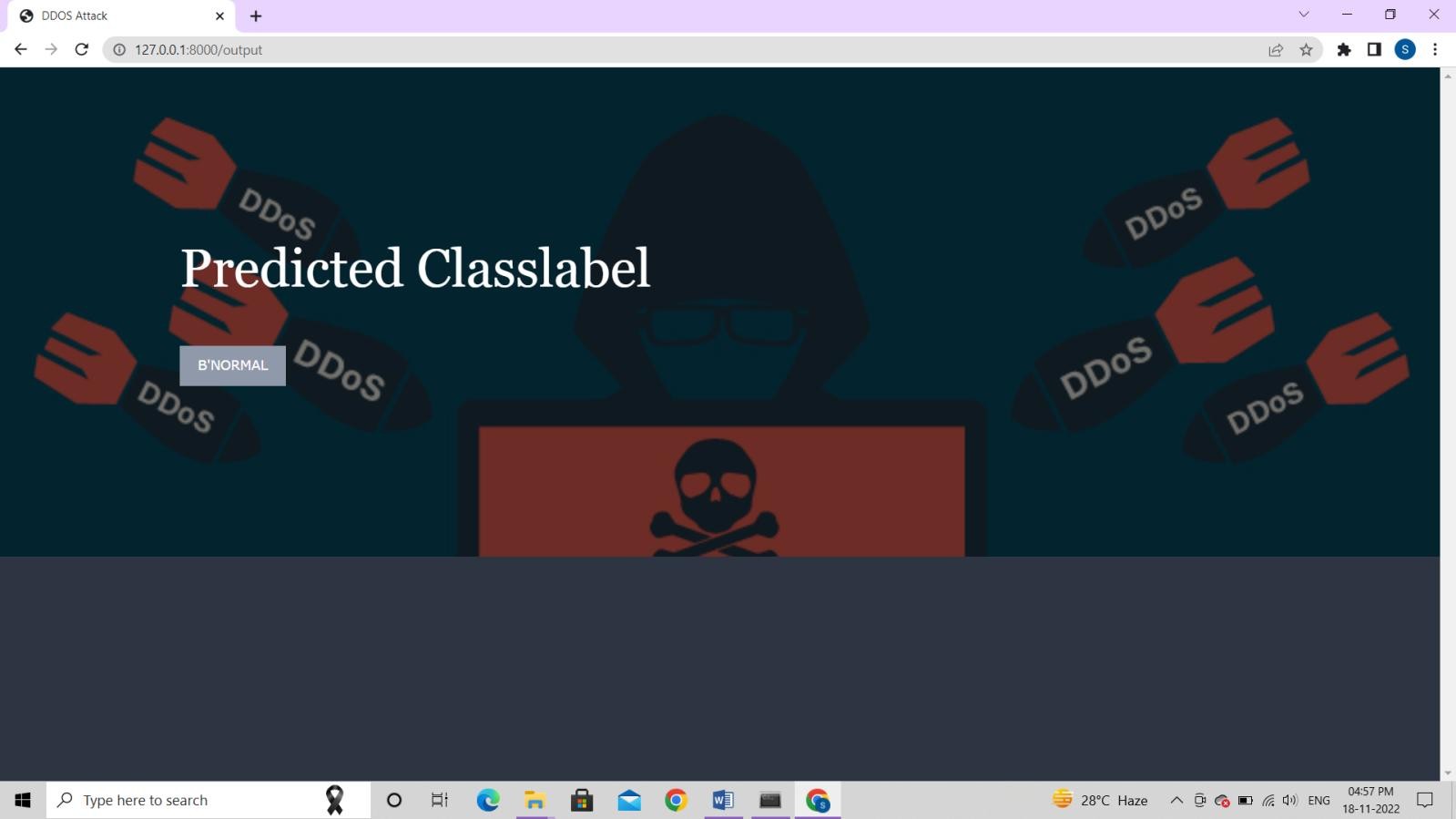


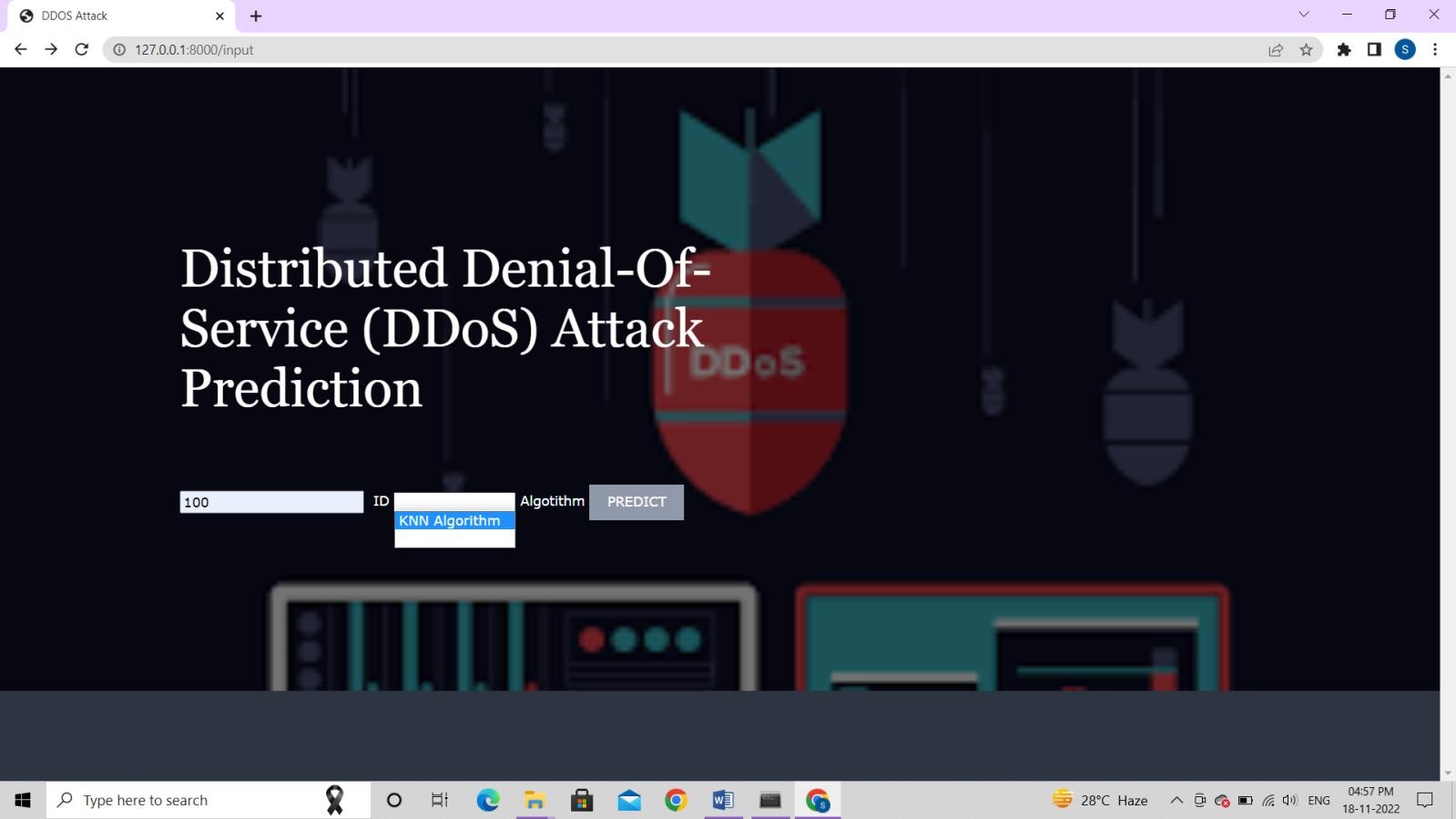














## CONCLUSION

The dataset contains 27 feature and 5 classes. Machine learning algorithm were applied to the data set namely Support Vector Machine and KNN. SVM algorithm showed greater accuracy and precision from KNN. The SVM model’s high performance and can be used in Intrusion detection Purposes. Future work will include more type of attacks and different feature selection techniques.