**CSE523 Machine Learning**

**Prof. Mehul Raval**

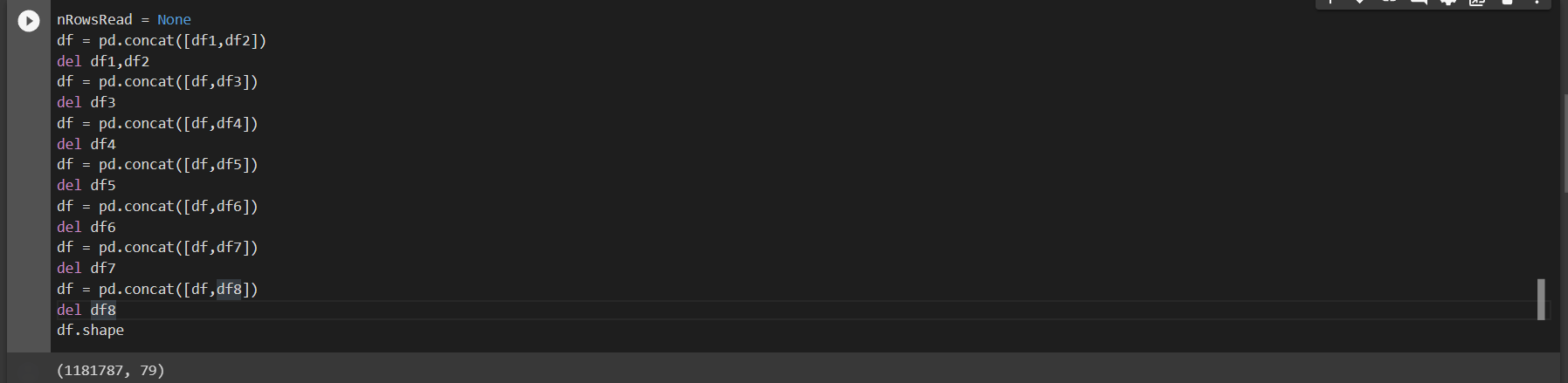
**Anomaly detection in computer networks to identify unusual activity or potential security threats**

**Week 3 Report**

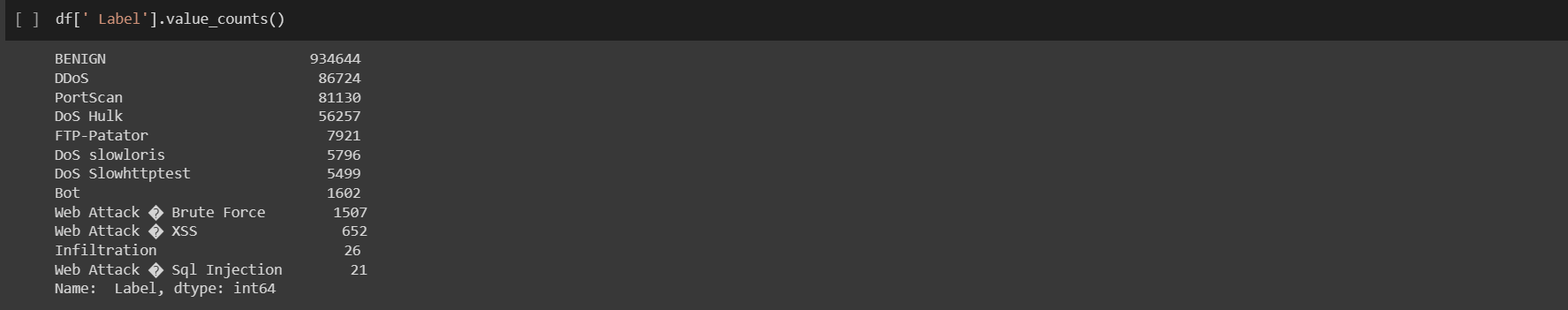
| **Name** | **Enrolment Number** |
| --- | --- |
| Vatsal Patel | AU2040043 |
| Shreya Karia | AU2040076 |
| Digant Patel | AU2040086 |
| Samarth Chauhan | AU2040097 |

Last week we implemented KNN to classify the attacks, and it was taking a lot of computation and was not giving any results. This week we worked on it further.

Our dataset consists of a total of 8 dataframes. We have concated them together. This results in a total of **1181787** records.



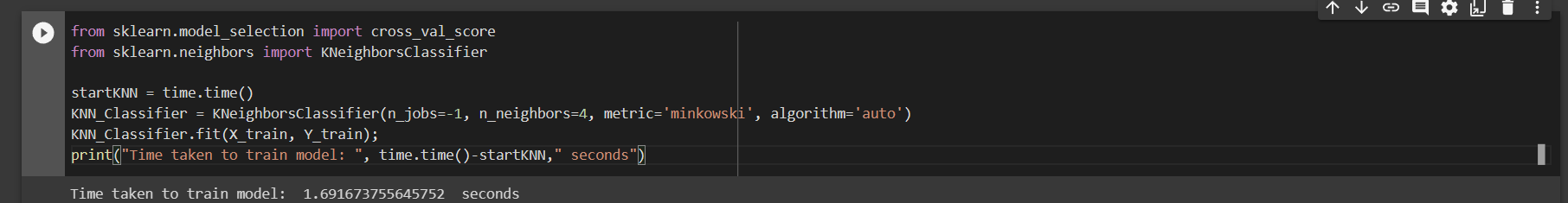
One of the 79 columns of the dataset is the “Label” column. This is used to classify the type of attack. We use this to train our model. As we can see below the dataset has a lotal of 12 types of attacks.



Post this, we have cleaned the data by dropping duplicates, dropping NaN, too high or too low values and mono-variables. We also megred similar classes like Web Attack Brute Force, Web Attack XSS and Web Attack SQL Injection to increase the class size.

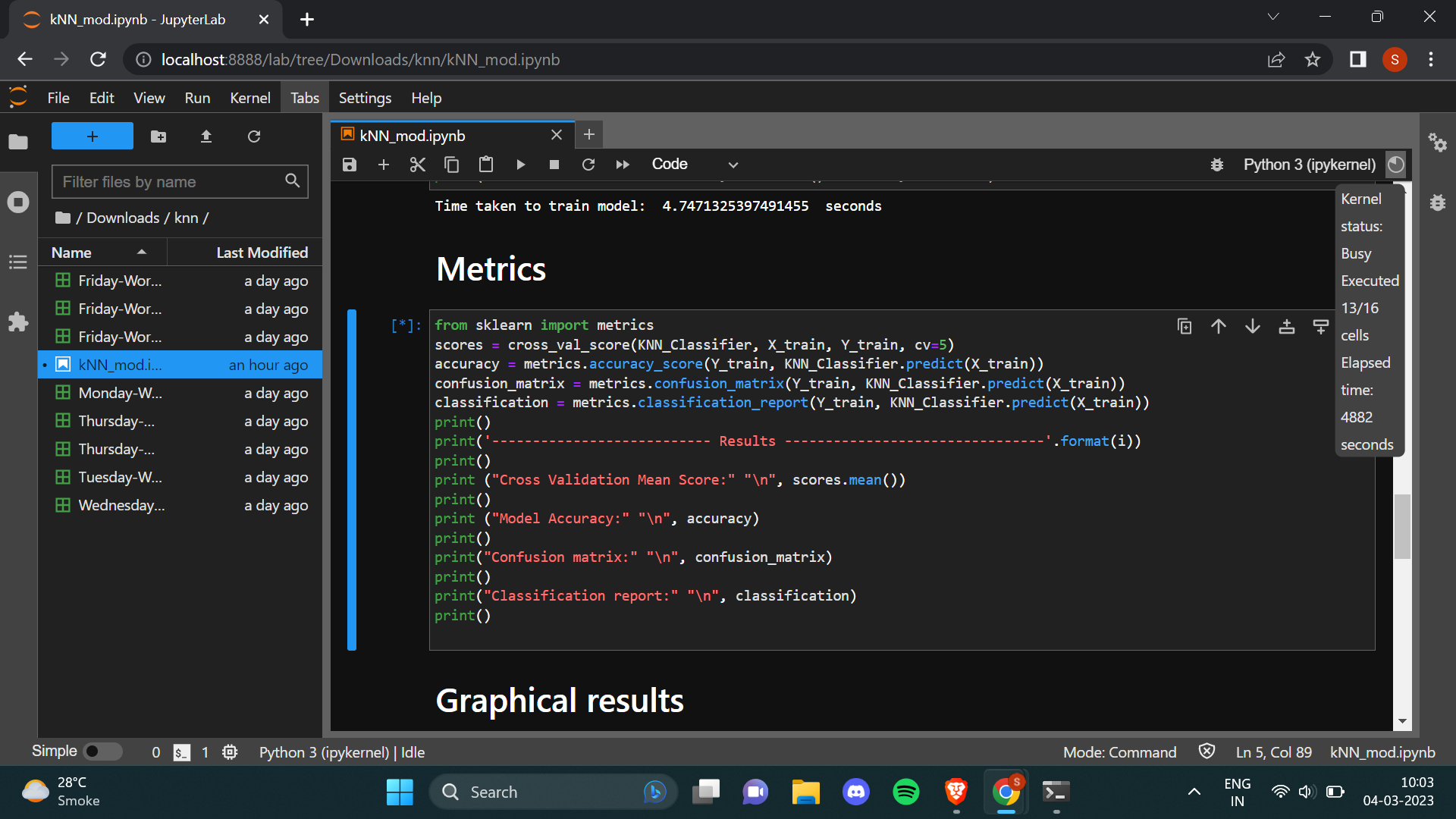
Following this, we divided our data into x and y, where x contains all the inputs and y consists of the “Label” column. This is followed by normalization of the data by applying Z-score normalization to each column in the dataframe, this prepares the data for further analysis or modeling, where having comparable features can improve the accuracy and reliability of the results.

Following this we split the data into training and test set. And use KNN model to build the classification model.



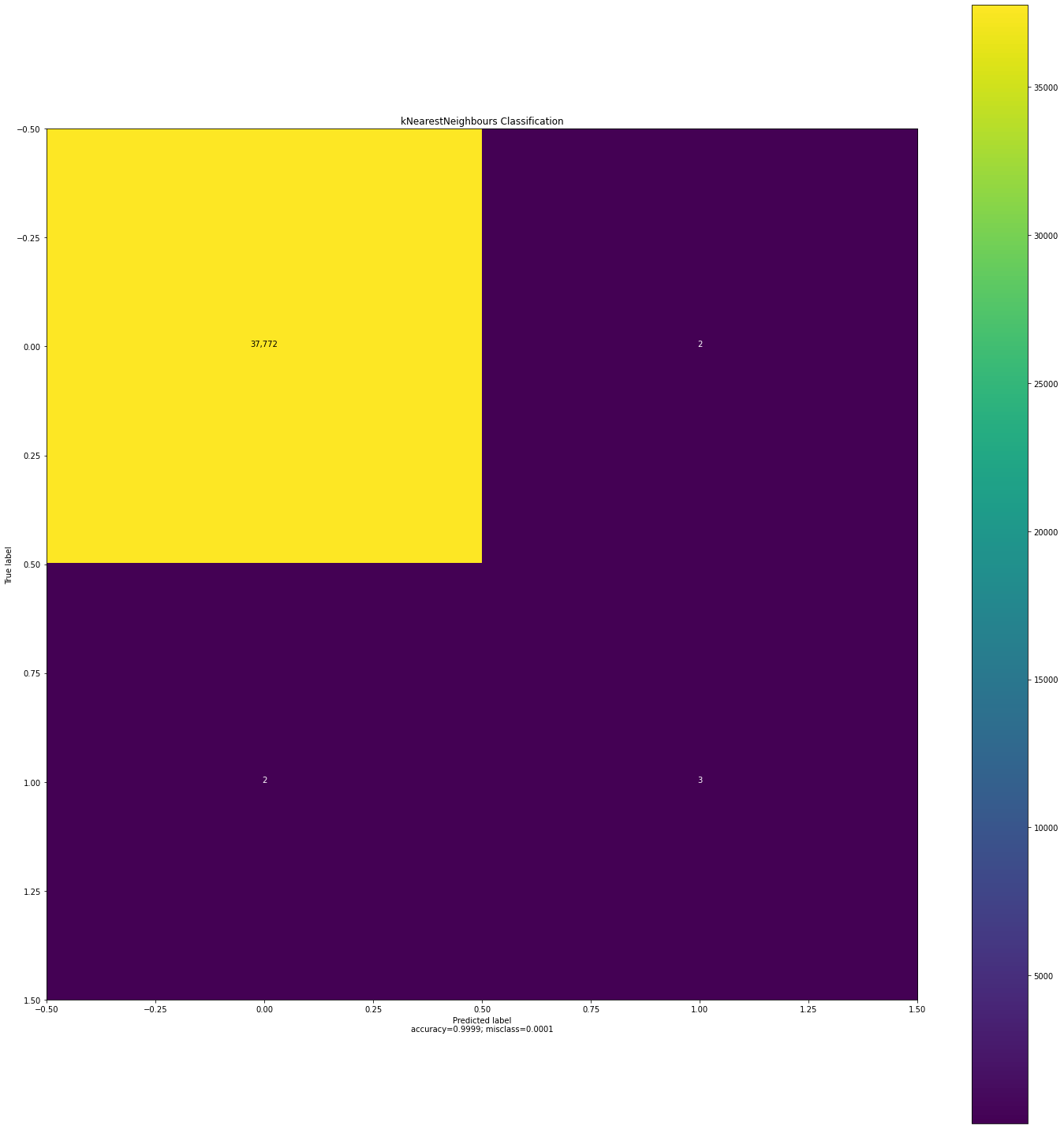
After training the model, we measure the performance of the trained KNN classifier on the training data using cross-validation and multiple metrics. It helps to assess the quality of the trained model and identify potential issues such as overfitting or underfitting. But here we face issue as mentioned below.

*Plotted confusion matrix*



*Execution time : more than 4880 seconds. Google colab disconnects after some time on its own.*

*Jupyter-lab takes too long to execute.*



*For reduced data size (almost 38,000 rows) we get this confusion matrix as a plot.*

*Overall kNN has a better prediction rate and accuracy than most of the other classification algorithms but the execution time is too much.*