Name: Gaurav Kumar Singh

Registration Number: 19BCE2119

Course: Information Security Analysis and Audit (CSE3501) L21+22

Digital Assignment 2

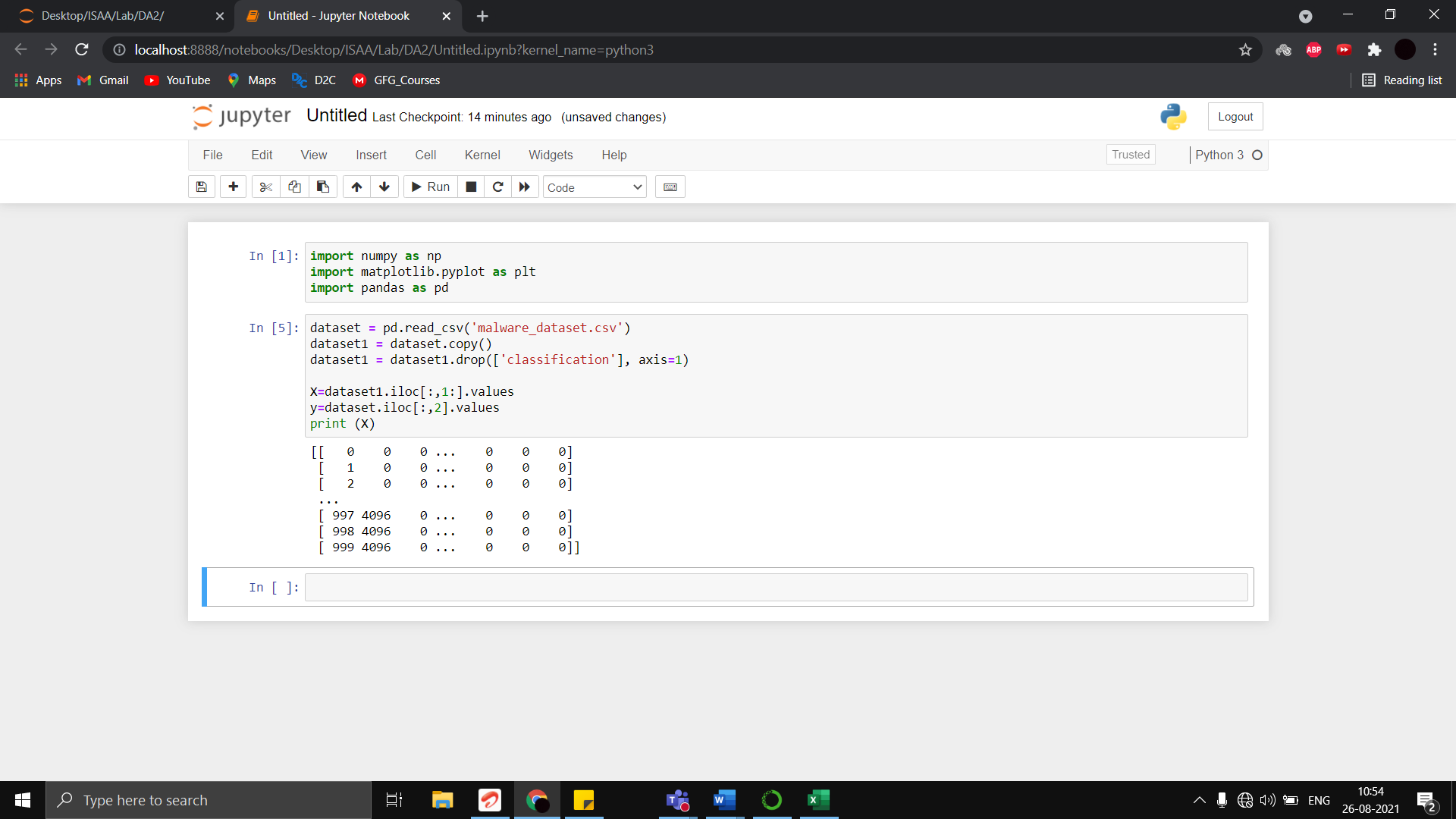
**INTRODUCTION MALWARE DATASET**

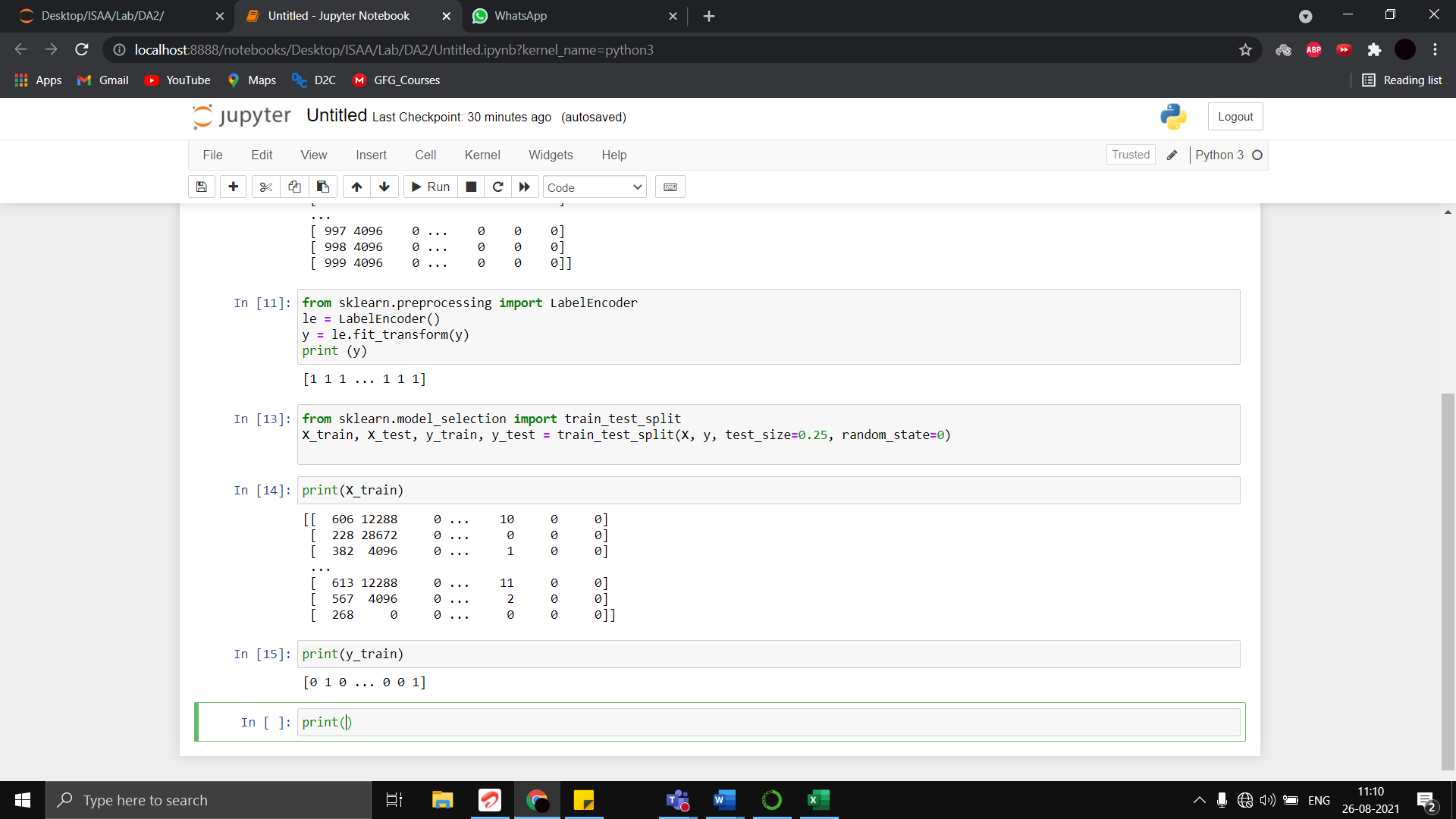
The dataset contains information about various benign and malware along with their characteristics like task priority, virtual memory statistics, etc. With the help of these statistics, we can make predictions about unknown software to predict the risk by classifying it as malware or benign. There are a lot of machine learning algorithms so we can train and test our machine learning algorithms to find out which algorithm works best for our use case.

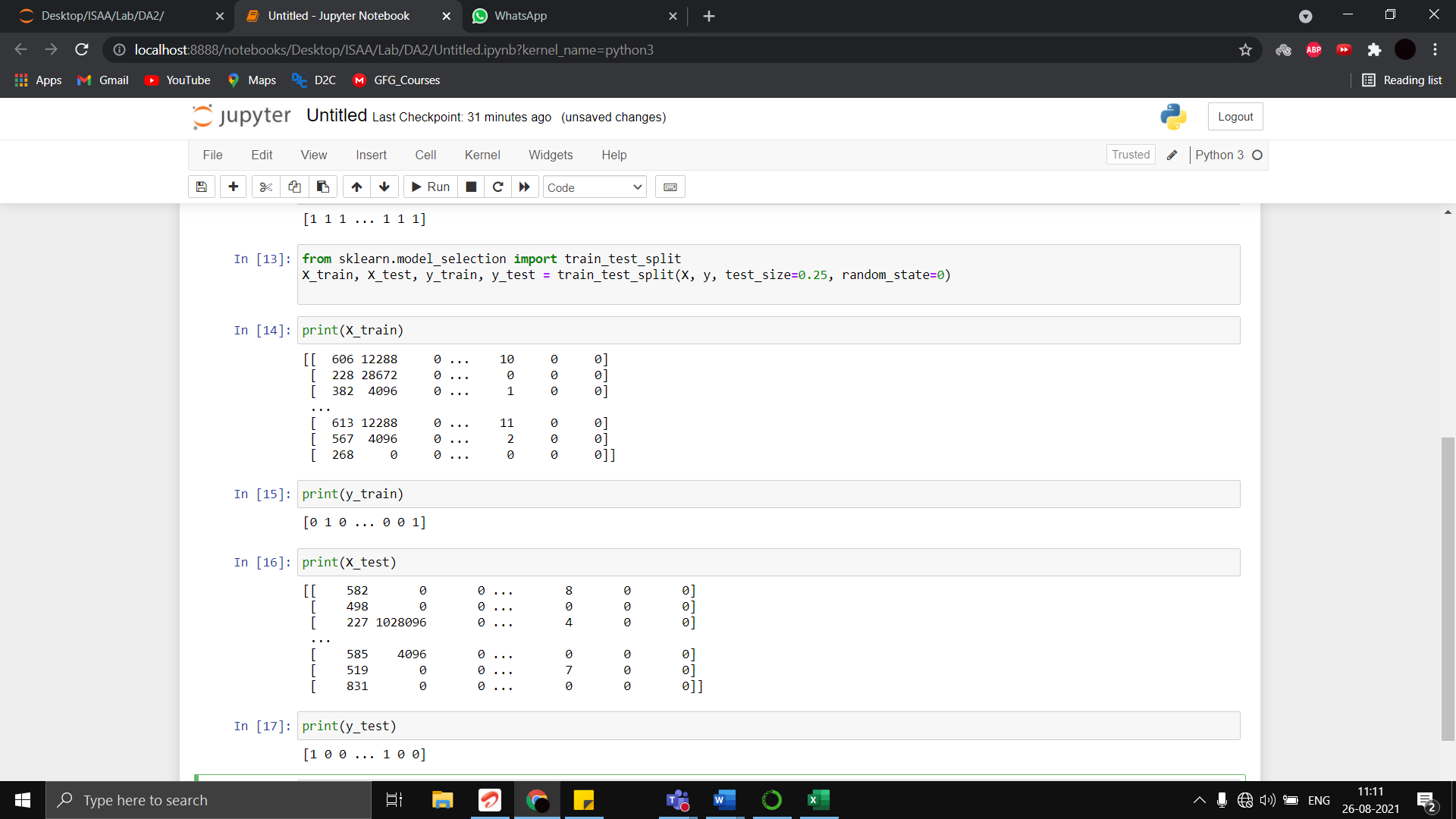
**LOGISTIC REGRESSION**

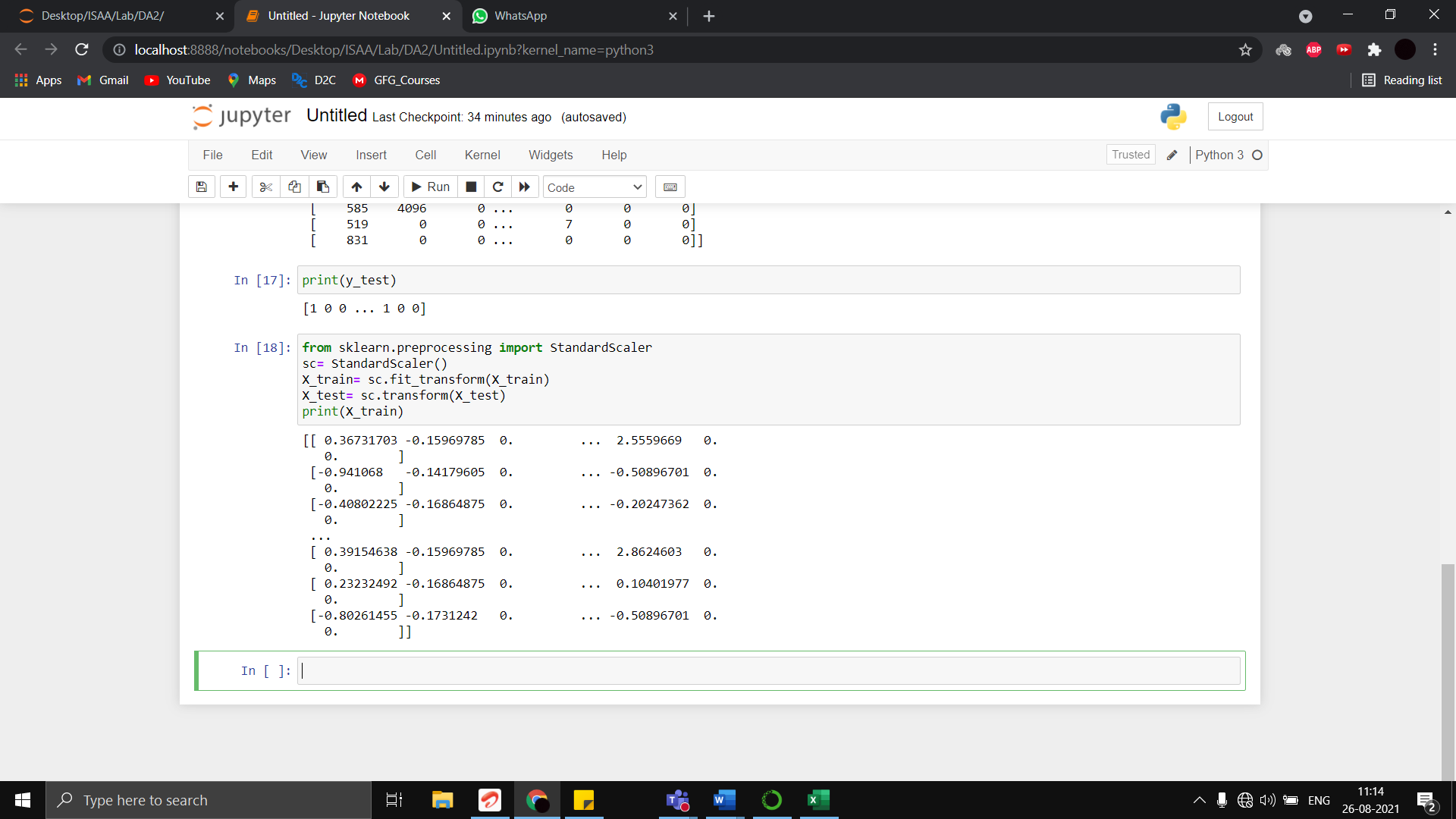
Logistic regression is a supervised machine learning algorithm that is used to classify the data. It is used to classify data in only one of the two maximum possible ways in the case of Binomial Logistic Regression which we are going to use for our test because we have to classify the data as either malware or benign only. We must choose only meaningful variables to train our model and a large sample size is required to train the algorithm. It makes use of logistic functions, also known as sigmoid functions as part of its calculations for the classification of data.

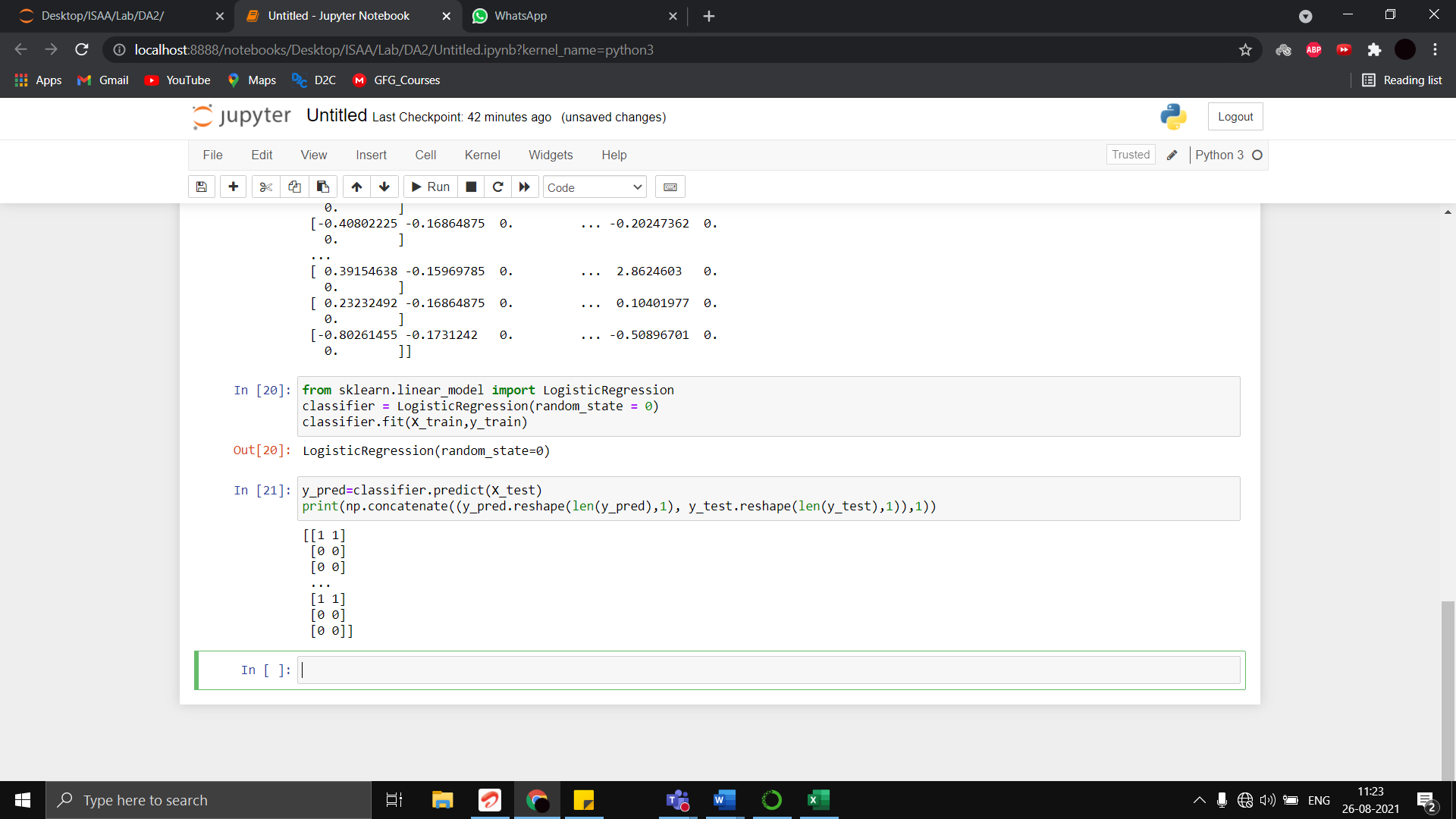
**MILESTONES WITH TIMESTAMPS**

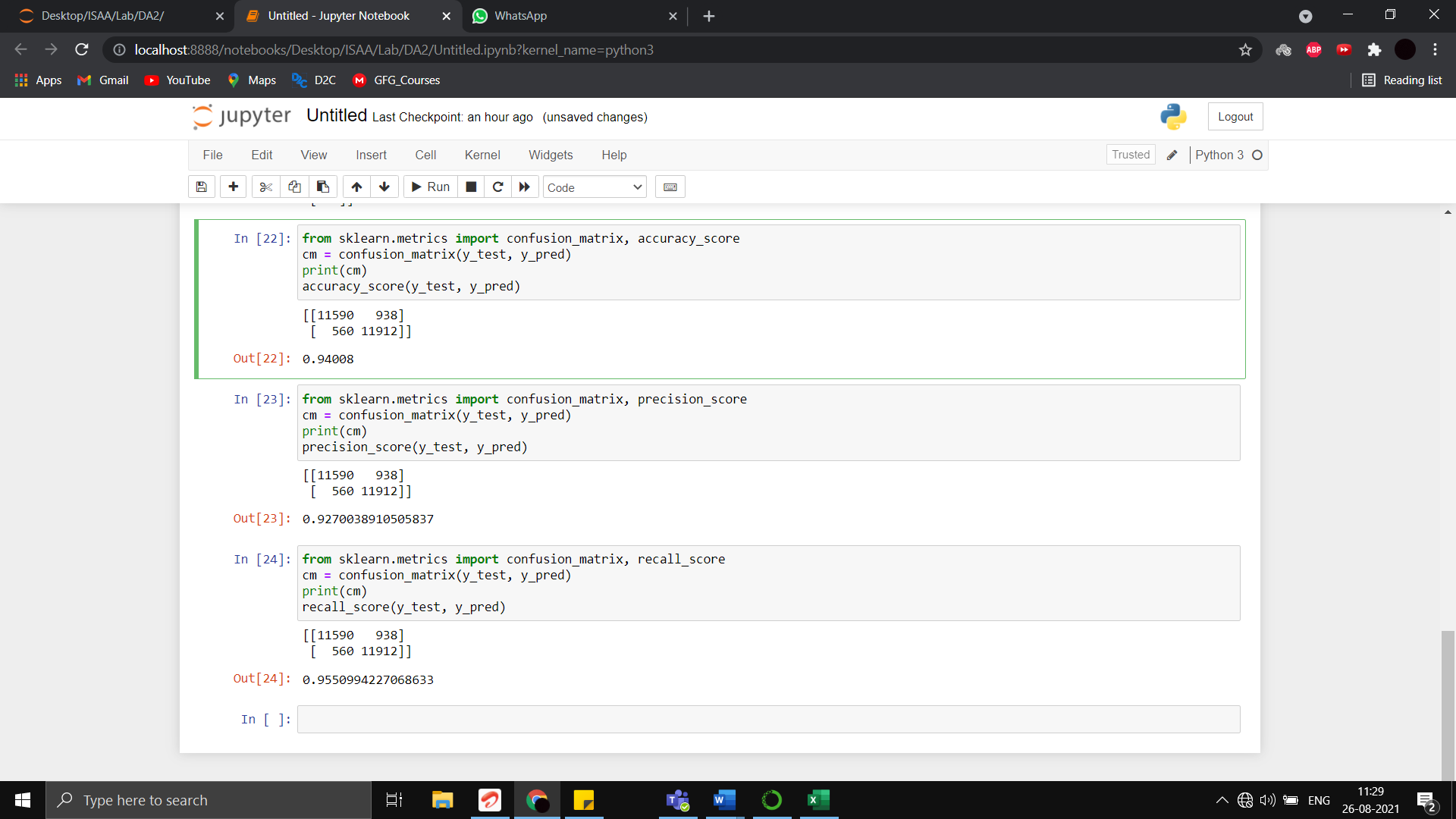








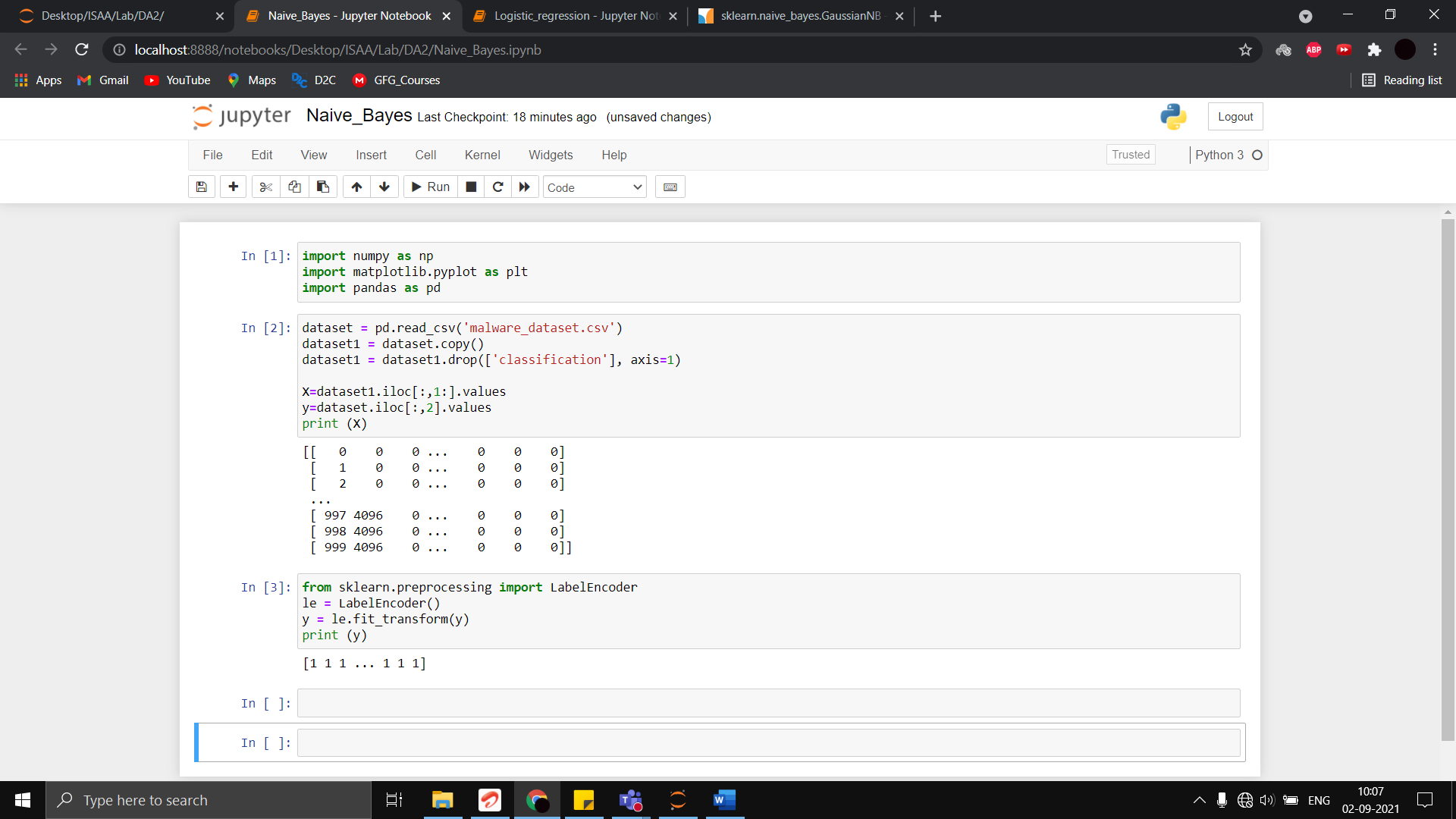


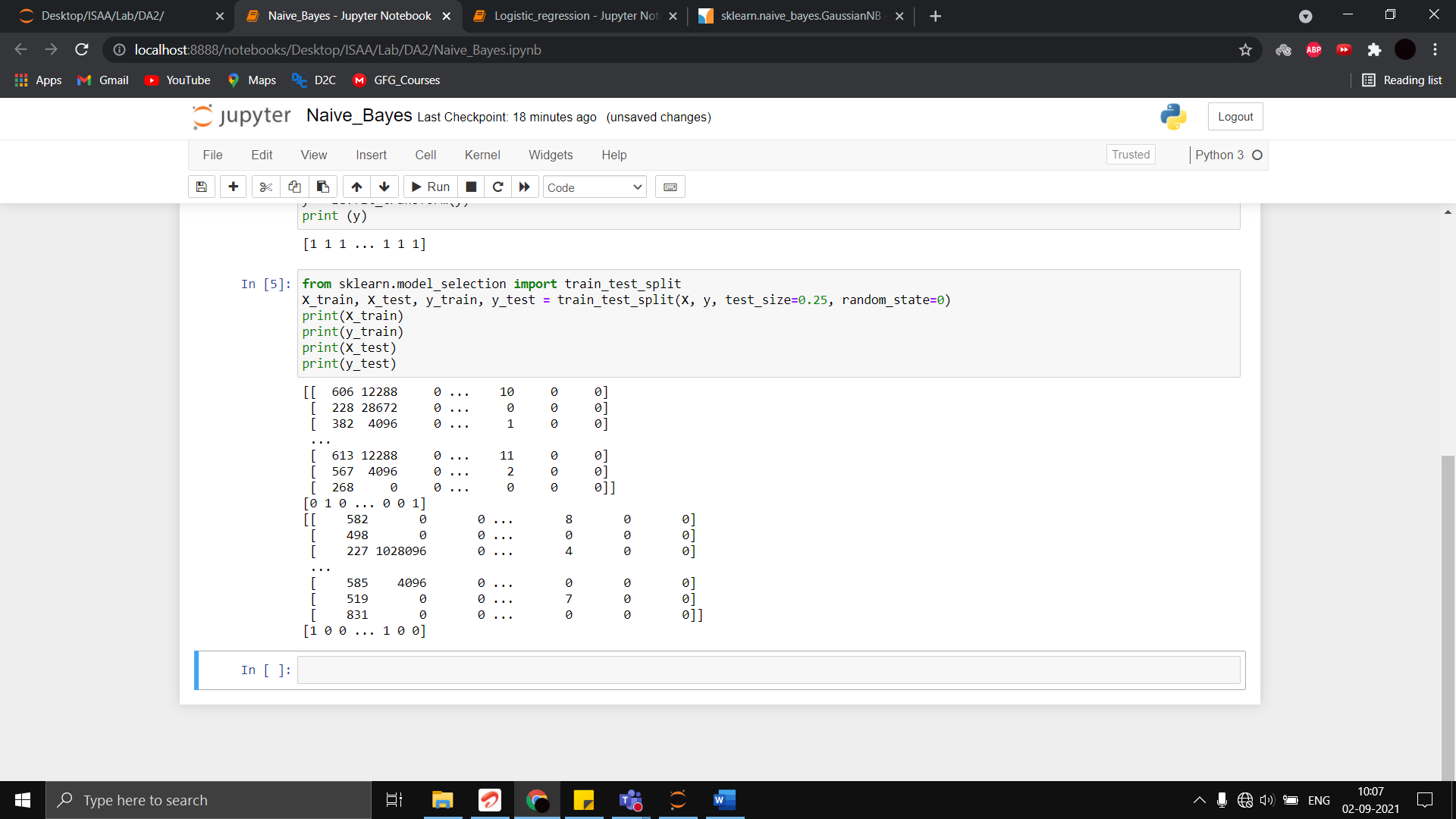


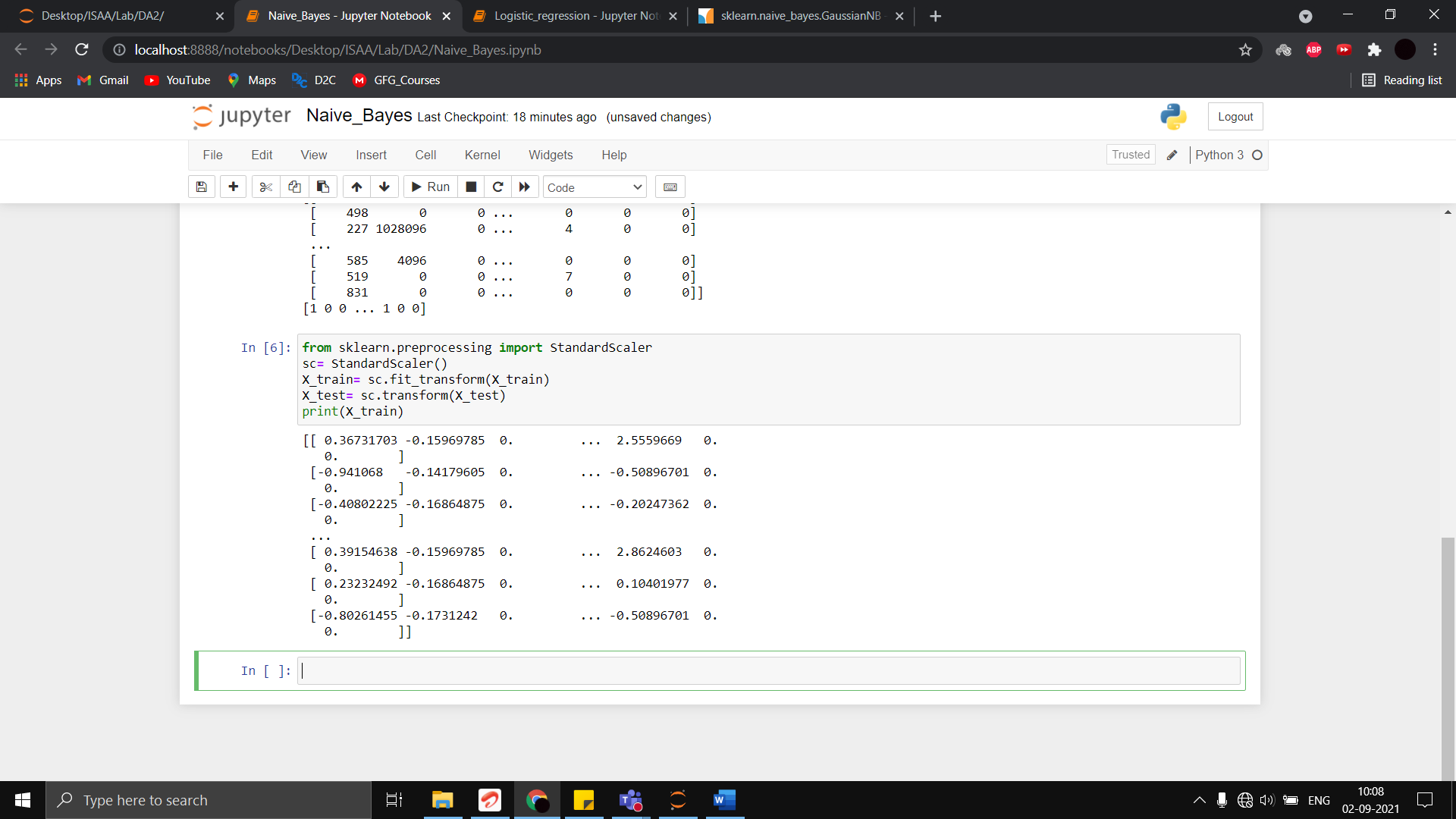
**NAÏVE BAYES**

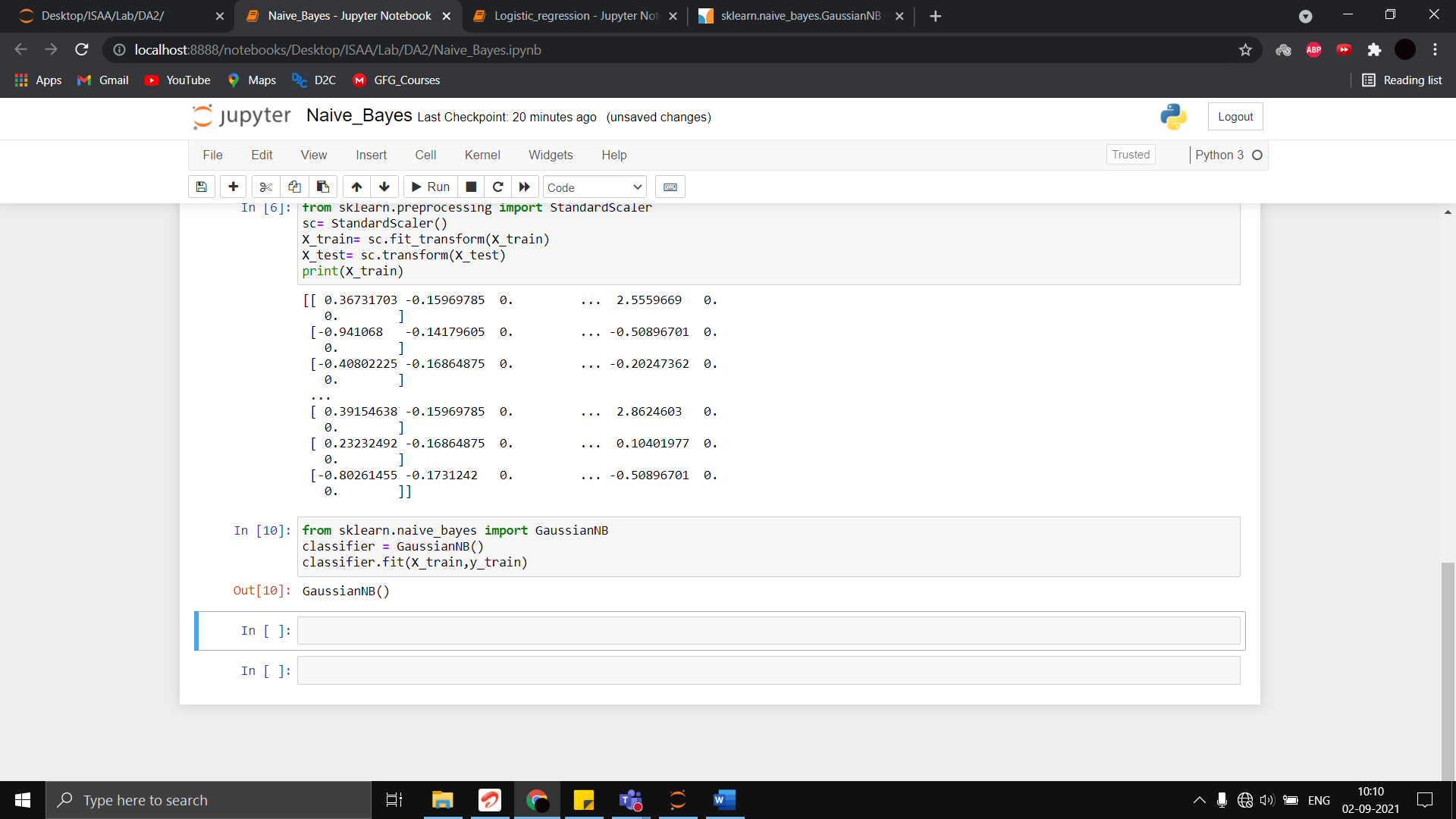
Just like Logistic regression, Naïve Bayes is also a supervised learning algorithm based on Bayes Theorem for the classification of data. It is a probabilistic classifier, implying it uses the probability of event occurrence as part of its mathematical calculations. It has many variations in itself, a few of which are Gaussian Naïve Bayes, Multinomial Naïve Bayes etc. We are going to use Gaussian Naïve Bayes for our specific use case for this activity and calculate accuracy scores, recall scores, and precision scores to compare these respective scores to that of logistic regression to find which algorithm gives us the most accurate classification for any unknown data of application which need to be classified as either malware or benign.

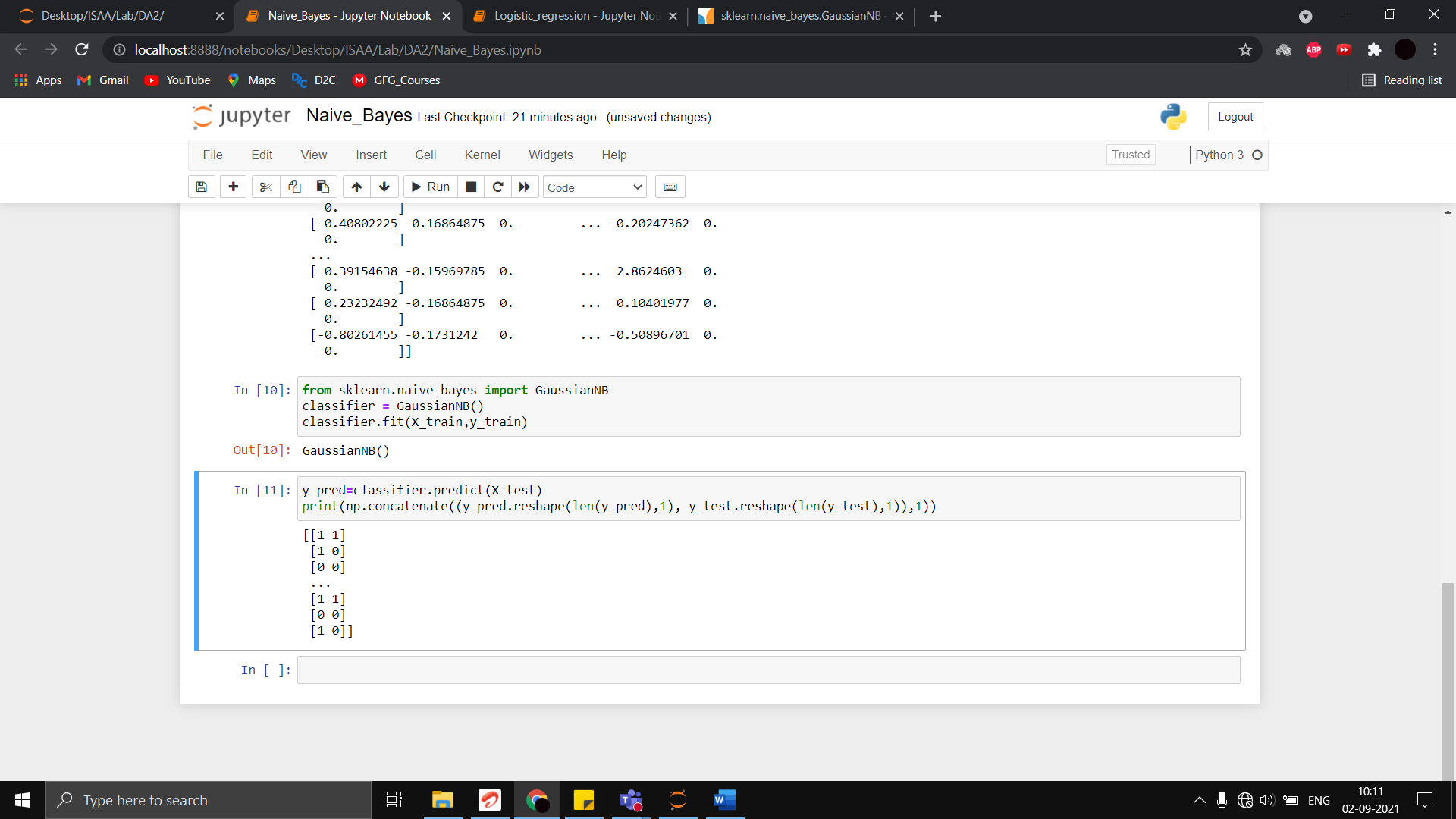
**MILESTONES WITH TIMESTAMPS**

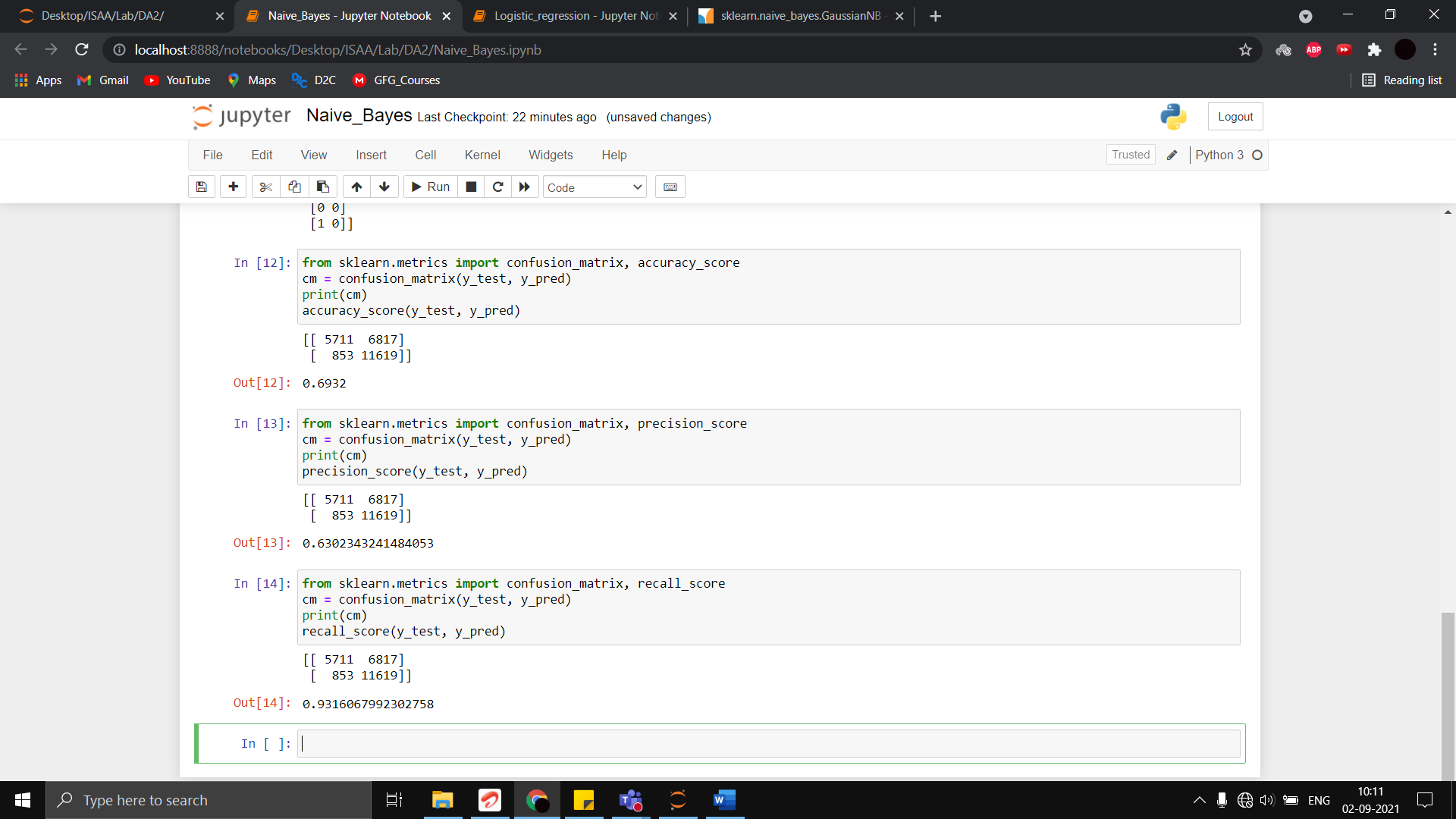












**COMPARISON**

|  |  |
| --- | --- |
| **LOGISTIC REGRESSION** | **GAUSSIAN NAÏVE BAYES** |
| Accuracy Score: 0.94008 | Accuracy Score: 0.6932 |
| Precision Score: 0.9270038910505837 | Precision Score: 0.6302343241484053 |
| Recall Score: 0.9550994227068633 | Recall Score: 0.9316067992302758 |
| Accuracy and Precision Scores for logistic regression are very high when compared to Gaussian Naïve Bayes and Recall Score is also slightly higher. | Accuracy and Precision scores for Gaussian Naïve Bayes algorithm is not very good for this case when compared to Logistic Regression. |

Logistic Regression is a better Machine Learning algorithm to train the malware dataset in our case when compared to the Gaussian Naïve Bayes algorithm because, in almost all positive performance metrics, Logistic Regression gave better scores.