VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANASANGAMA, BELAGAVI – 590018



An Internship Report

on

SPAM MAIL PREDICTION USING MACHINE LEARNING

**Submitted in partial fulfillment for the award of degree of**

**Bachelor of Engineering**

**in**

**Computer Science and Engineering**

***Submitted by***

**Rhea Benedicta D’souza**

**4SO18CS097**

***Internship Carried Out***

**at**

**Cognitive Solution**

**5th Floor, M.K. Shalimar Complex**

**Kankanady, Mangalore - 575002**

 ****

**Internal Guide External Guide**

**Ms Supriya Salian Mrs. Sibby Susan**

**Assistant Professor Technical Trainer**

**St. Joseph Engineering College Cognitive Solution**

**Department of Computer Science and Engineering**

**St Joseph Engineering College**

**Mangaluru - 575028**

**2021-22**

VISVESVARAYA TECHNOLOGICAL UNIVERSITY JNANASANGAMA, BELAGAVI – 590018



An Internship Report

on

SPAM MAIL PREDICTION USING MACHINE LEARNING

**Submitted in partial fulfillment of the requirements for the degree**

**Bachelor of Engineering**

**in**

**Computer Science and Engineering**

***Submitted by***

**Rhea Benedicta D’souza 4SO18CS097**



**Department of Computer Science and Engineering**

**St Joseph Engineering College**

**Mangaluru - 575028**

**2021-22**

**St Joseph Engineering College**

**Mangaluru – 575 028**

**Department of Computer Science and Engineering**



#### CERTIFICATE

Certified that the Internship Work titled **“SPAM MAIL PREDICTION USING MACHINE LEARNING”** was carried out by **Ms. Rhea Benedicta D’souza**, bearing USN **4SO18CS097**, a bonafide student of final year B.E. in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi, during the year 2021-22. Further, it is certified that all corrections/suggestions indicated during Internal Evaluation have been incorporated in this report.

**------------------------ -------------------------------- -------------------------------**

**Ms. Supriya Salian Dr Sridevi Saralaya Dr. Rio D’Souza**

**Internal Guide Head of the Department Principal**

**External Viva Voce Examination**

**Name of the Examiners Signature with Date**

1. -------------------------------------- ---------------------------------

2. -------------------------------------- .--------------------------------

### NAME OF THE INDUSTRY / ORGANIZATION Address with pin code

### (may be given on company letterhead)

#### CERTIFICATE

Certified that the Internship Work titled………………………………………….. ………………………………………………………………………….....………………..was carried out by Mr/Ms ………………………………………….., bearing USN…………………..………, a student of St Joseph Engineering College, Mangaluru, during the year 2021-22. Further, it is stated that he/she has satisfactorily completed the work on our company premises, between ............. <start month> and ............ <end month> 2021/(2022).

**Name & Signature of the External Guide Name & Signature of the CEO/Division Head**

**Date: Date:**

**DECLARATION**

I, **Rhea Benedicta D’souza,** bearing USN **4SO18CS097**, student of final year B.E. in Computer Science and Engineering, St Joseph Engineering College, Mangaluru, hereby declare that the Internship Work titled **“SPAM MAIL PREDICTION USING MACHINE LEARNING”** has been duly executed by me from 8th September – 8th October 2021, at Cognitive Solution, Kankanady, Mangalore. Further, the “Tasks Performed” section of this report represents the work done solely by me and does not contain any statements falsely claiming work done by others, as my own.

**Date: 30/4/2022**

**Place: Mangalore** **Rhea Benedicta Dsouza**

**ACKNOWLEDGMENT**

I dedicate this page to acknowledge and thank those responsible for shaping this project. Without

their guidance and help, this experience would not have been so smooth and efficient.

I would like to extend my sincere gratitude to Mrs. Sibby Susan, Technical Trainer at Cognitive Solution for giving me the opportunity to complete my internship, his guidance and encouragement helped me throughout the internship.

I sincerely thank Ms Supriya Salian, Assistant Professor, Department of Computer Science and Engineering for her guidance which helped us fulfill the requirements prescribed by the university and her valuable suggestions which brought this internship to fruition.

I am indebted to Dr Sridevi Saralaya, Head of the Department of Computer Science and Engineering, whose kind consent and guidance helped us complete this internship successfully.

I am grateful to our Director, Rev. Fr Wilfred Prakash D’Souza, our Assistant Director, Rev. Fr Alwyn Richard D’Souza and our Principal, Dr Rio D’Souza for their support and encouragement.

I am grateful to the staff of Computer Science and Engineering Department for their encouragement and support.

I would also like to thank my friends for their valuable suggestions and family members for their continued support.

**Executive Summary**

I carried out my internship in Machine Learning using Python at Cognitive Solution, Kankanady, Mangalore from 8th September 2021 to 8th October 2021.

Cognitive Solutions a service provider of Web‐based Development & Web based Software Development Solutions, Mobile Application Development, Graphic Design and Windows Applications.Cognitive Solutions is headquartered in Mangalore, with the Business development in UAE, Saudi Arabia and Qatar. In a short span of 8+ years, our products as well as services & solutions have been widely accepted by the global market. Today, Cognitive Solution has the experience to undertake any IT development or deployment works on a single point responsibility basis. Their Products and Services are user friendly with easy controls and are of superior specifications. They are always proactive to fulfill client’s needs and requirements to the best possible extent of their satisfaction. They manage interactive sessions with clients throughout the project development.

The Objective of the internship was to gain knowledge in machine learning and build a project for better understanding of concepts and to enhance my skill. The languages and tools used were python,

This internship gave me an opportunity to enhance my professional skills and gain industry

experience. It also encouraged me to learn new technologies and tools like Machine Learning. Further, it helped to develop my presentation skill and time management. I was able to complete the task in a short duration of time. This internship also provided me an opportunity to apply acquired knowledge to real work experiences. Overall, the internship helped me gain valuable work experience and enhanced my skills which is very essential to my career.

RHEA BENEDICTA D’SOUZA

**CONTENTS**

**SL No. Title Page No.**

1 About the company 1

* 1. Brief history 1
  2. Service offered by the company 1
  3. Contact details 1

2 About the department 2

2.1 Introduction 2

2.2 Roles and responsibilities 3

3 Task Performed 4

3.1 Daily work 4

3.2 Project Implementation 7

3.3 Snapshots 8

4 Reflection notes 12

4.1 Technical outcomes 12

4.2 Non-technical outcomes 13

References 14

**Figure Index**

**SL No. Title Page No.**

**CHAPTER 1**

**ABOUT THE COMPANY**

**CHAPTER 2**

**ABOUT THE DEPARTMENT**

**CHAPTER 3**

**TASKS PERFORMED**

**3.1 Daily Work Schedule**

**Week 1 (08/09/2021 - 11/09/2021) :**

In week 1, in order to move ahead to machine learning with AI, basic python concepts were brushed up, like syntax, data types operators etc and a few exercises were given so that I was well versed with the basic concepts.

**Week 2 (13/09/2021 - 16/09/2021):**

In week 2, concepts such as loops, object oriented programming, JSON, and Regular Expressions. I was given exercises on these concepts. The facilitator then introduced me to the basics of NumPy, Pandas and Matplotlib.

**Week 3 (20/09/2021 - 23/09/2021):**

In week 3, I was introduced to machine learning basics. Difference between supervised and unsupervised learning techniques were explained. Linear regression was taught by giving a demo exercise. I was then made to implement it in my system and analyse it.

**Week 4 (27/09/2021 - 30/09/2021):**

In week 4, classification algorithms were introduced. Other algorithms introduced were KNN (K Nearest Neighbor), SVM (Support Vector Machine) and Naive bayes. Exercises on the same were taught and implemented.

**Week 5 (01/10/2021 - 04/10/2021):**

In week 5, Logistic regression, decision tree classification and Kmeans clustering was taught. Problems on decision tree classification were given . Kmeans algorithm was taught and implemented. I was tasked with completing a project by selecting a topic of my choice on any of the algorithms I was aware of.

**Week 6 (05/10/2021 - 08/10/2021):**

In week 6, I I went through some of the topics taught and chose my topic of interest and started working on spam mail detection. This project used a logistic regression model to train the mail dataset. I created a kaggle account and chose one of the dataset which contained a good number of spam and ham mails. I altered the dataset by taking some mails from my gmail account and adding them onto the dataset. The next two days I spent on my project, by performing the following steps: importing dependencies, data collection and pre processing, label encoding, splitting data into training and testing, feature extraction, training the model, evaluating the trained model and building a predictive system. I successfully completed my project on time by the 7th of October 2022.

**3.2 Project Implementation**

PROBLEM STATEMENT: In the new era of technical advancement, e-mails have gathered significant users for professional, commercial, and personal communications. Because of the high demand and huge user base, there is an upsurge in unwanted emails, also known as spam emails. Even in the current date, people lose a lot of money to frauds every day.

OBJECTIVE: The objective of identification of spam e-mails are:

1. To give knowledge to the user about the fake e-mails and relevant e-mails
2. To identify if the mail is spam or not.

WORKFLOW :

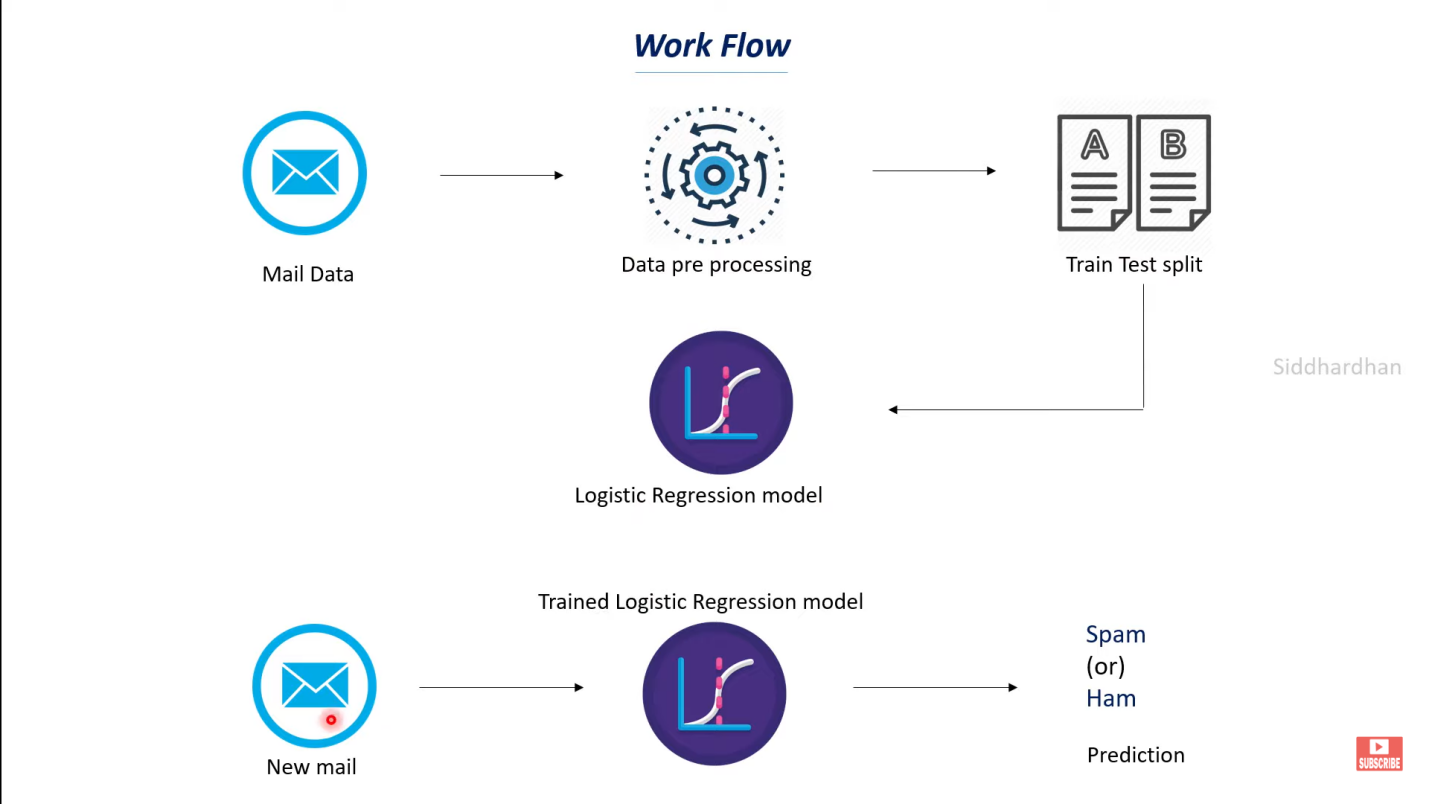


Fig 3.2a : Training Testing phase

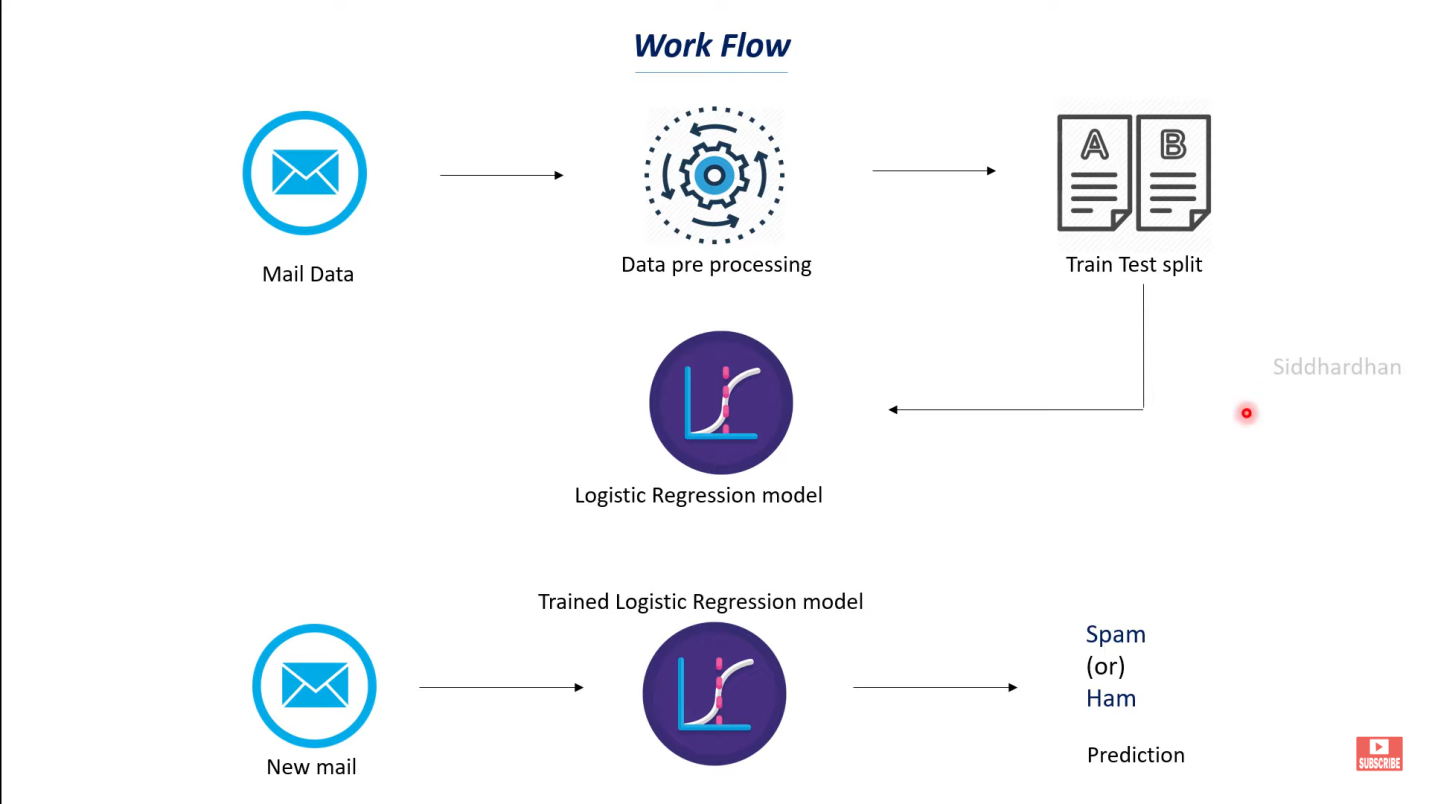


Fig 3.2b : New email classification

**3.3 SNAPSHOTS:**

**3.3.1 Importing the dependencies**

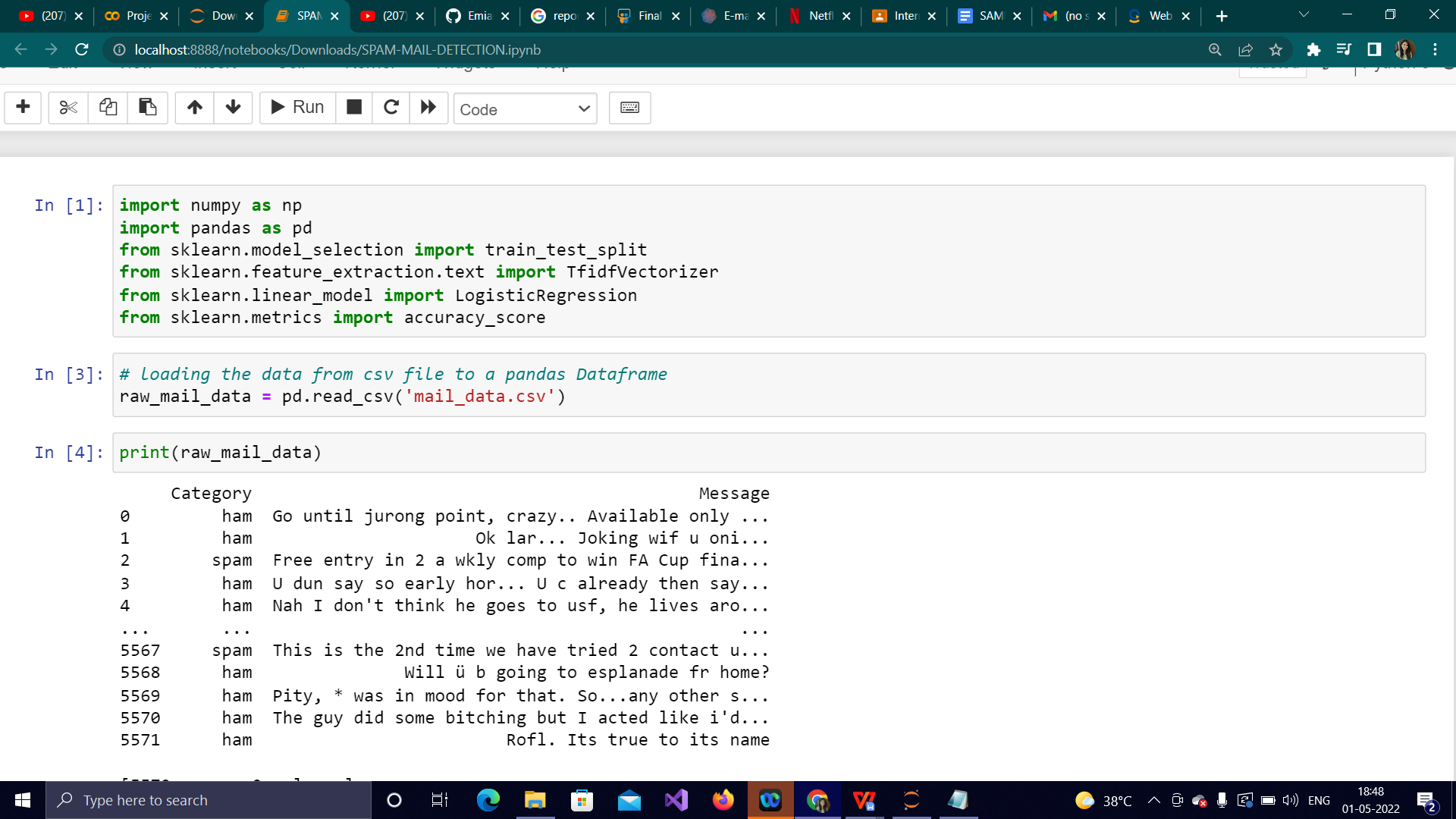


Fig 3.3.1 : Importing the dependencies

Here, numpy is used to create an array,pandas is used to create data frame, train\_test\_split is used to split our data into training and testing, TfidfVectorizer is used to convert text data(mail data) into numerical value so our machine learning model can understand it, LogisticRegression is used to classify mails into spam mail or ham mail, and accuracy score is used to know how well our model is predicting.

**3.3.2 Data Collection and Pre-Processing**

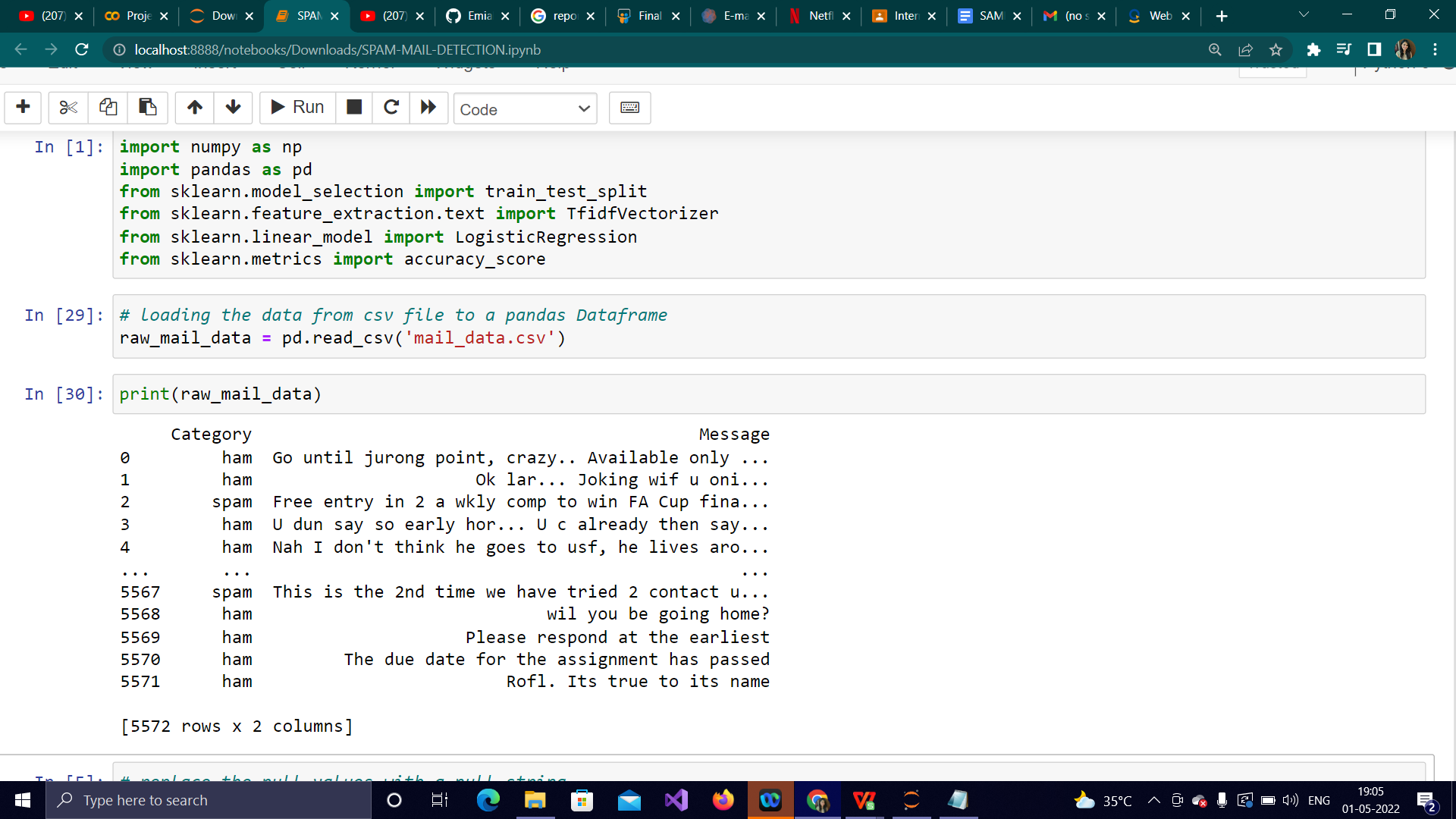


Fig 3.3.2a: Display of raw mail data

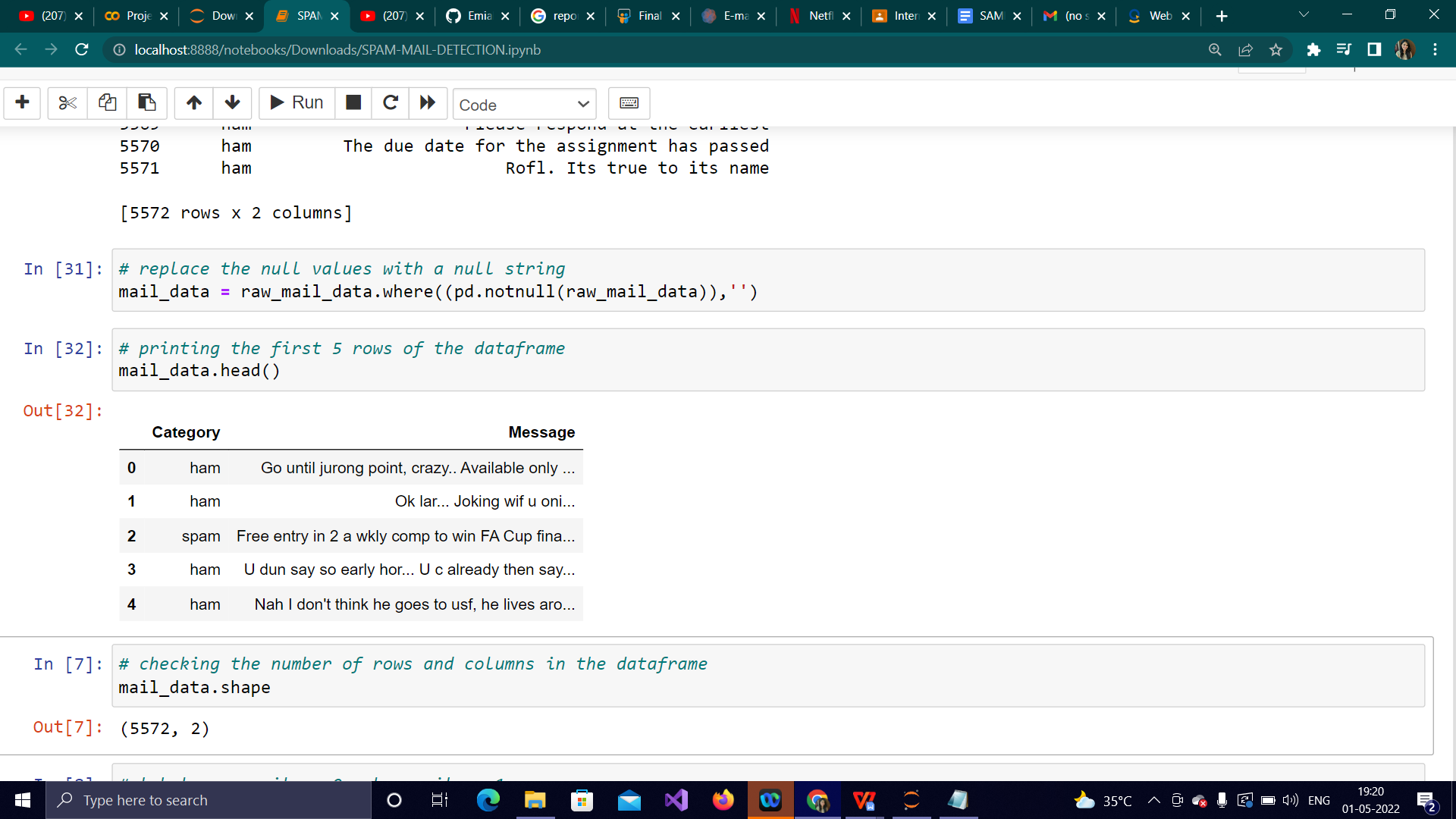


Fig 3.3.2b: Display first 5 rows

**3.3.3 Label Encoding:**

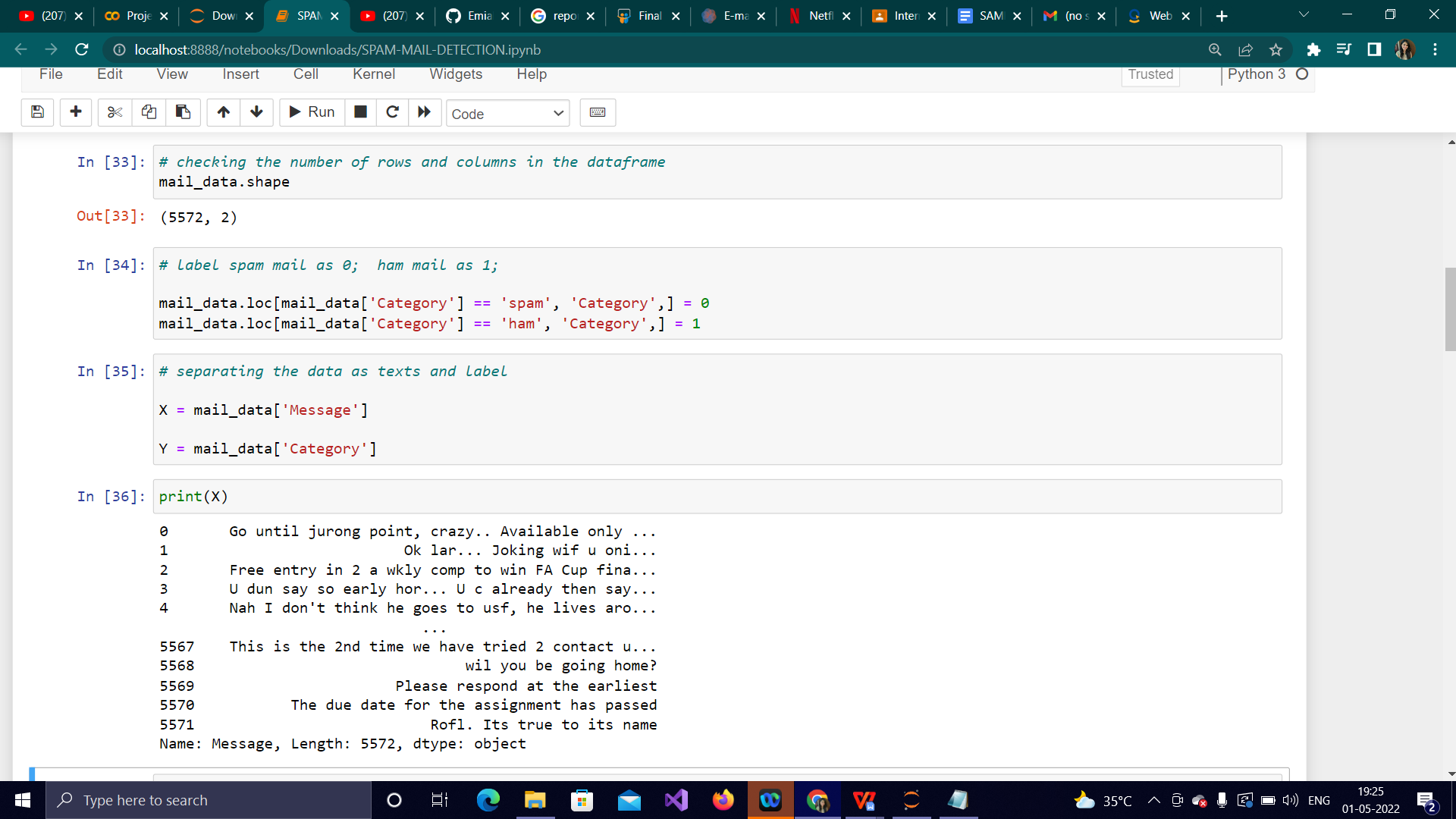


Fig 3.3.3a: Display of messages after separation

Here, the number of rows and columns consisting in the data frame are displayed. This is followed by label encoding which means labeling the spam mails as ‘0’ and ham mails (non-spam mails ) as ‘1’. The data is then separated as X and Y, where X contains the mails and Y contains the category of the mail which is nothing but 1 (ham) and 0 (spam).

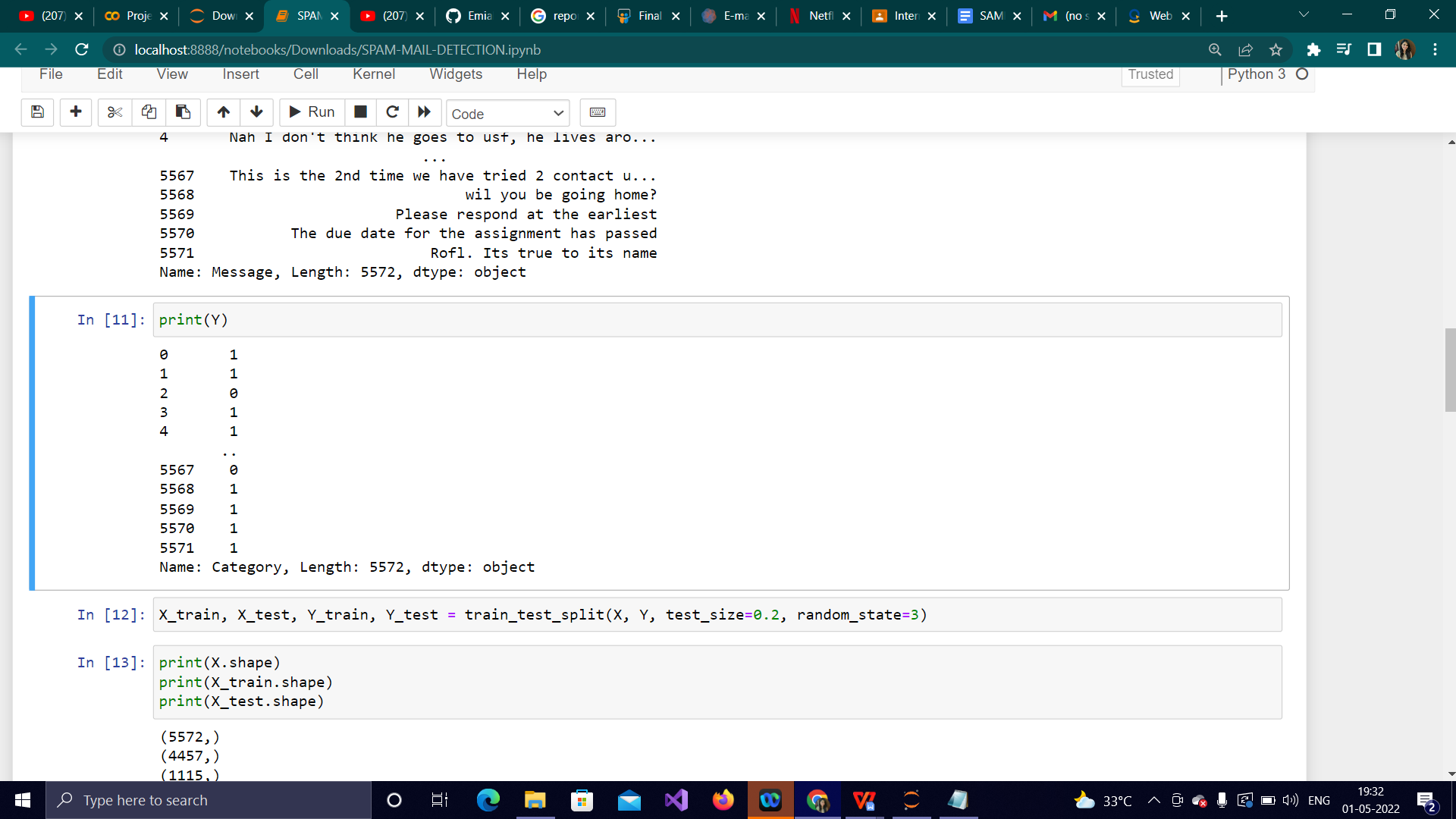


Fig 3.3.3b: Displaying the category after separation

**3.3.4 Splitting the data into training data and test data:**

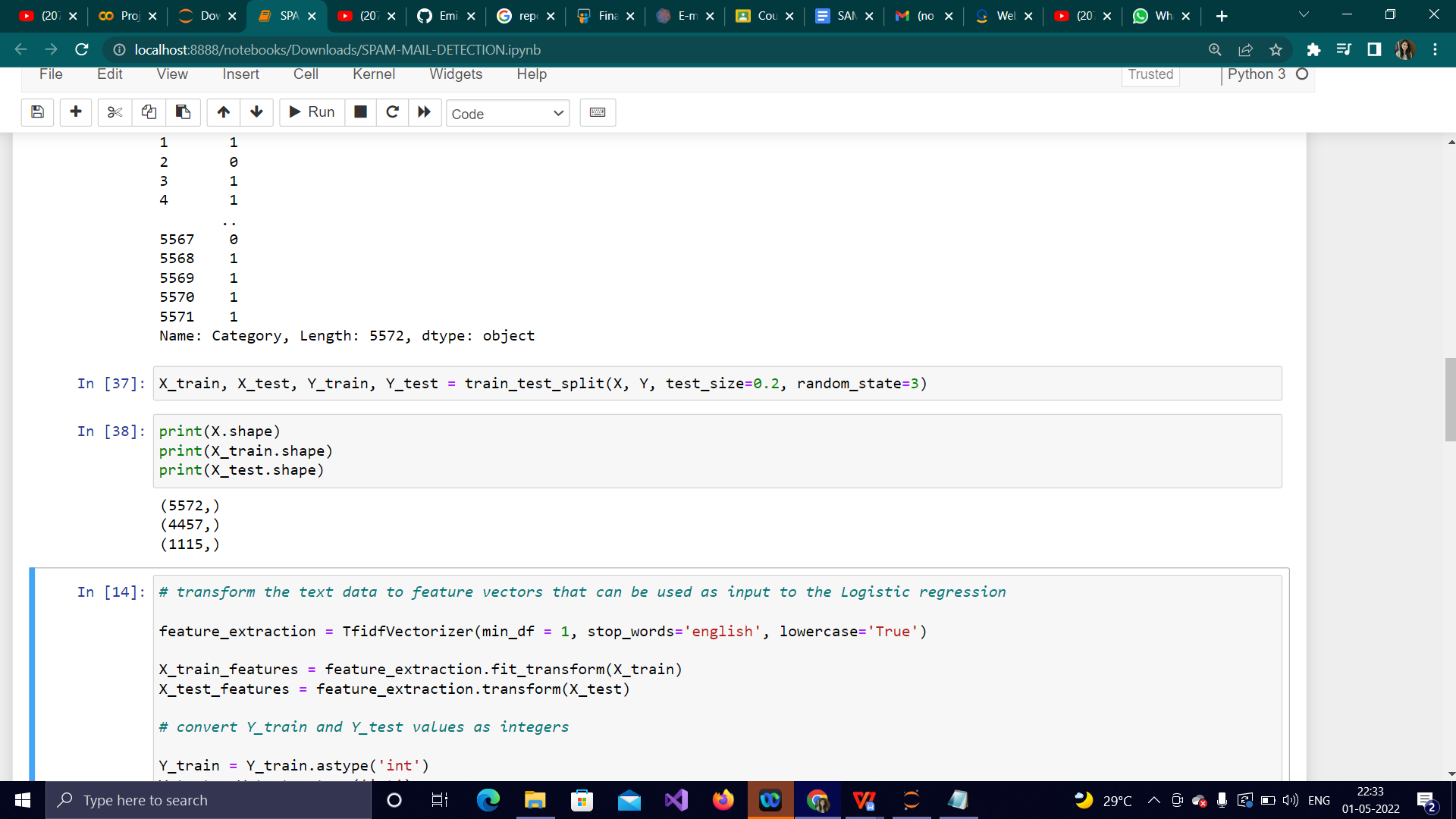


Fig 3.3.3: Displaying number of data split into training and test

Here, 80% of the data is going into training and 20% of the data is going into testing.

**3.3.5 Feature Extraction:**

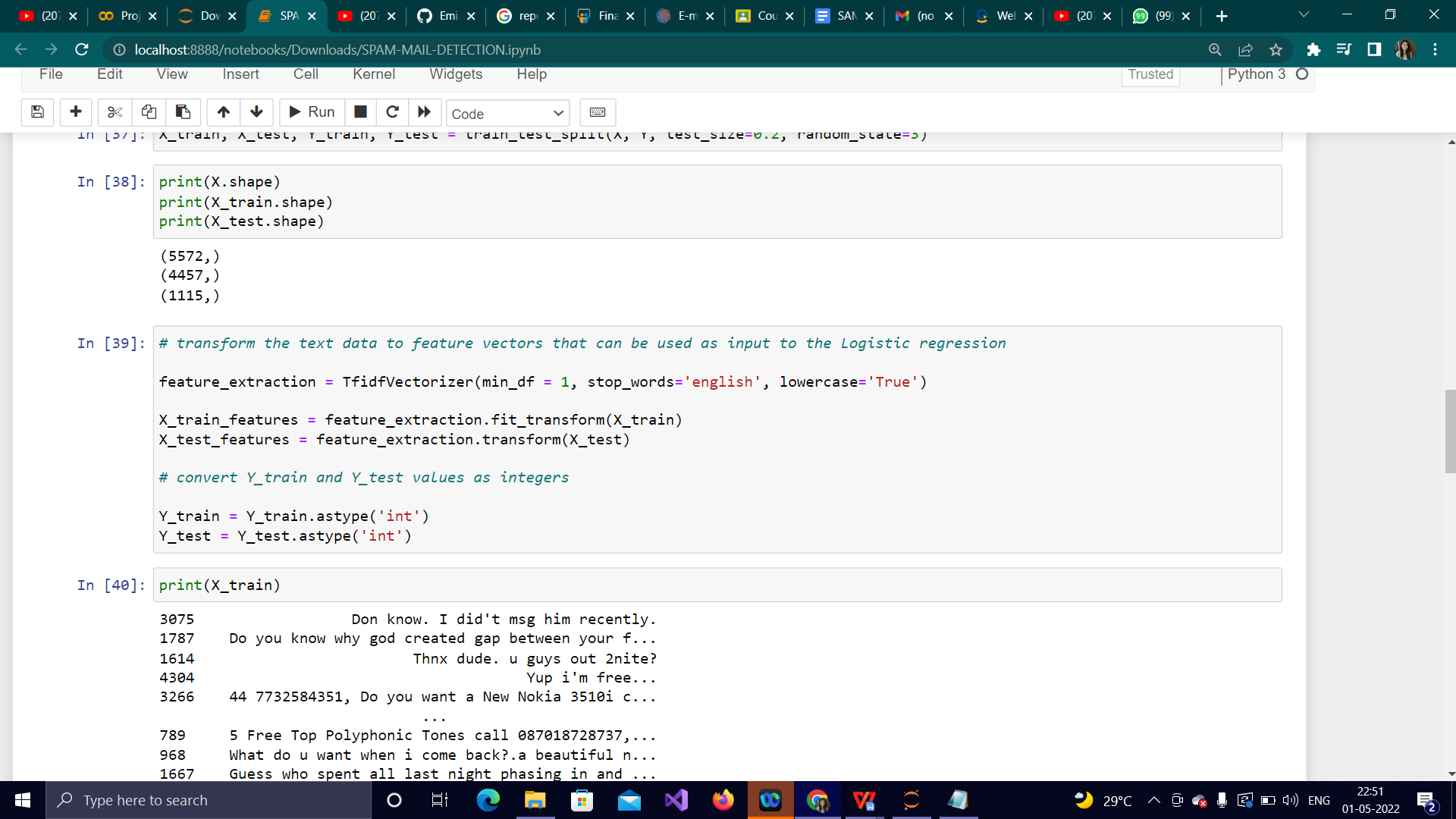


Fig 3.3.5a: Transforming text data into feature vectors

Here, we load TfidVectorizer into a variable called feature extraction. It is used to transform

Text data into feature vectors. The two important steps taking place here are: fitting all our

training data into vectorizer and following this it will transform all the x\_train data into feature vectors which are nothing but numerical values.

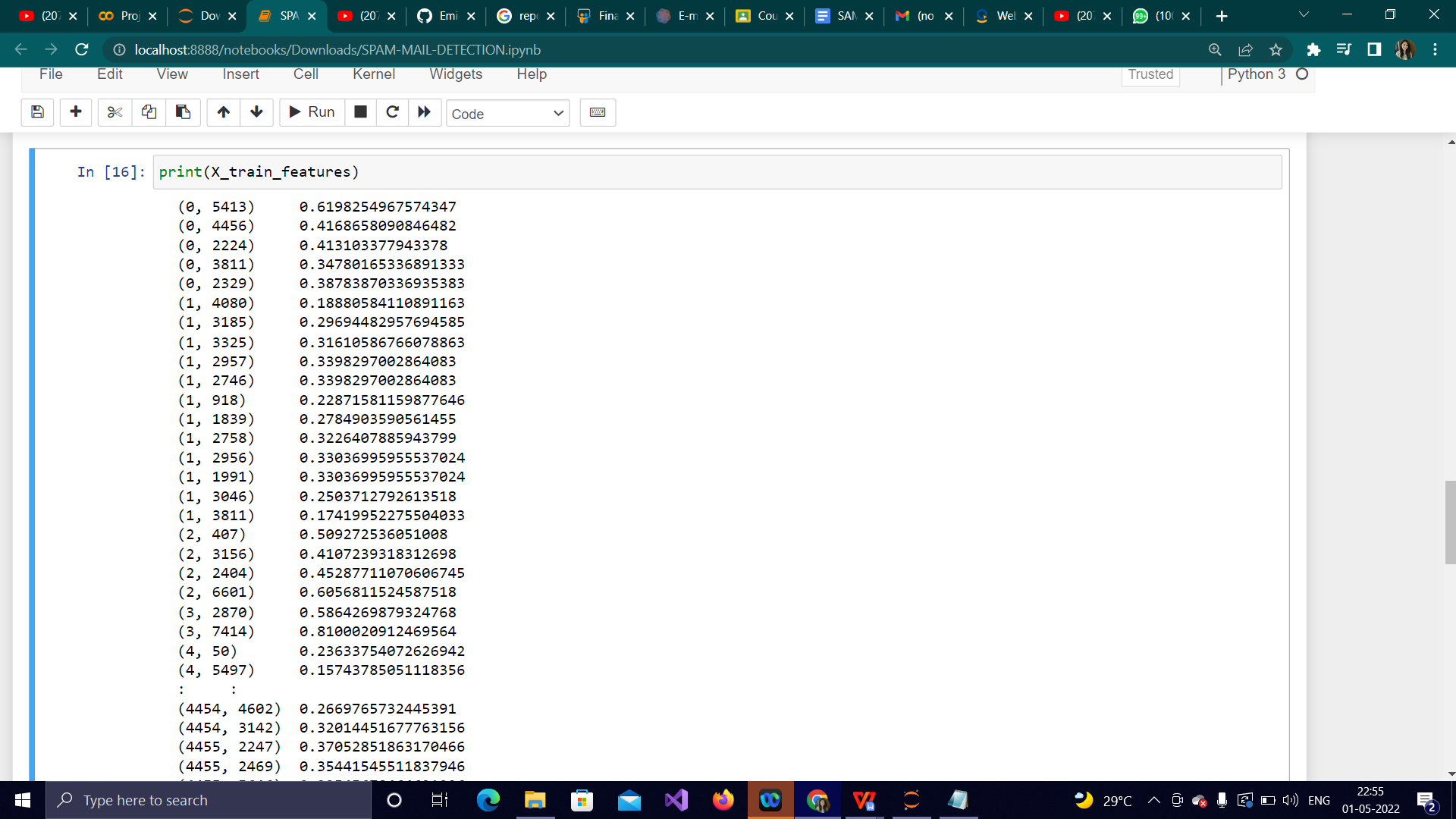


Fig 3.3.5b: Displaying x\_train\_features

**3.3.6 Training the model**

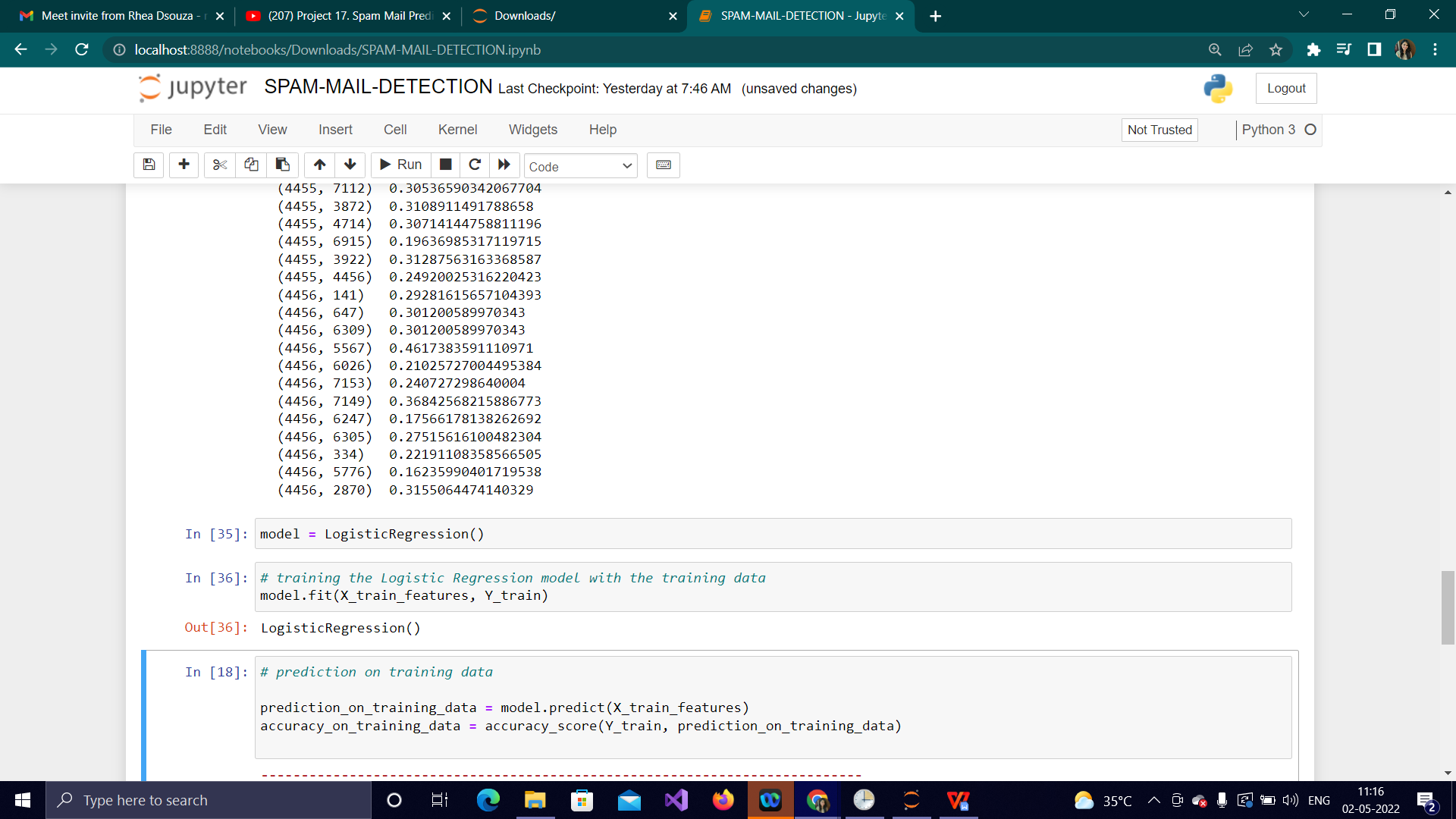


Fig 3.3.6: Training the model with the training data

**3.3.7 Evaluating the trained model**

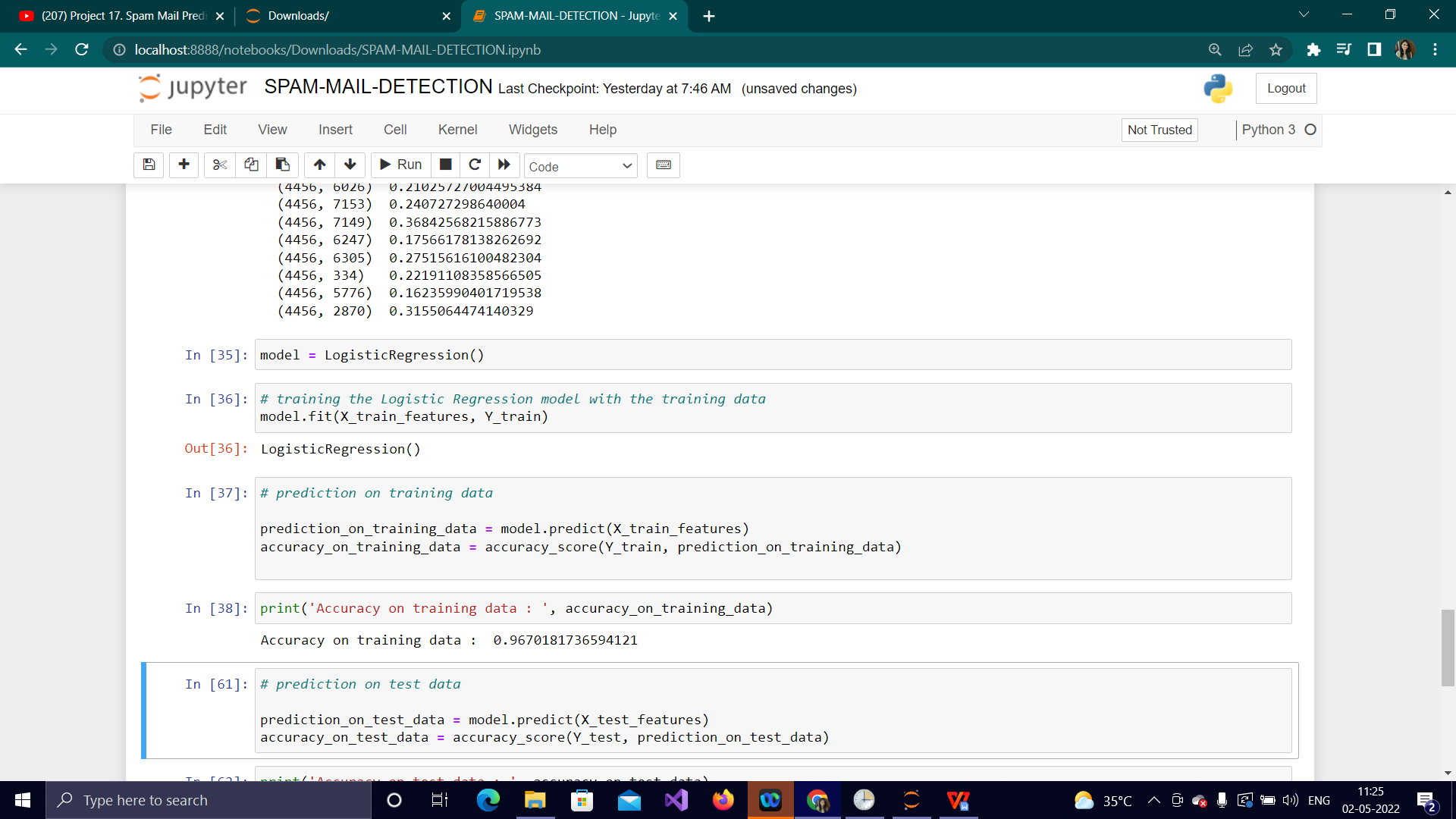


Fig 3.3.7a: Accuracy on training data

Here, we can observe that if we use this model to predict 100 different mails, it will predict them with 96% accuracy.

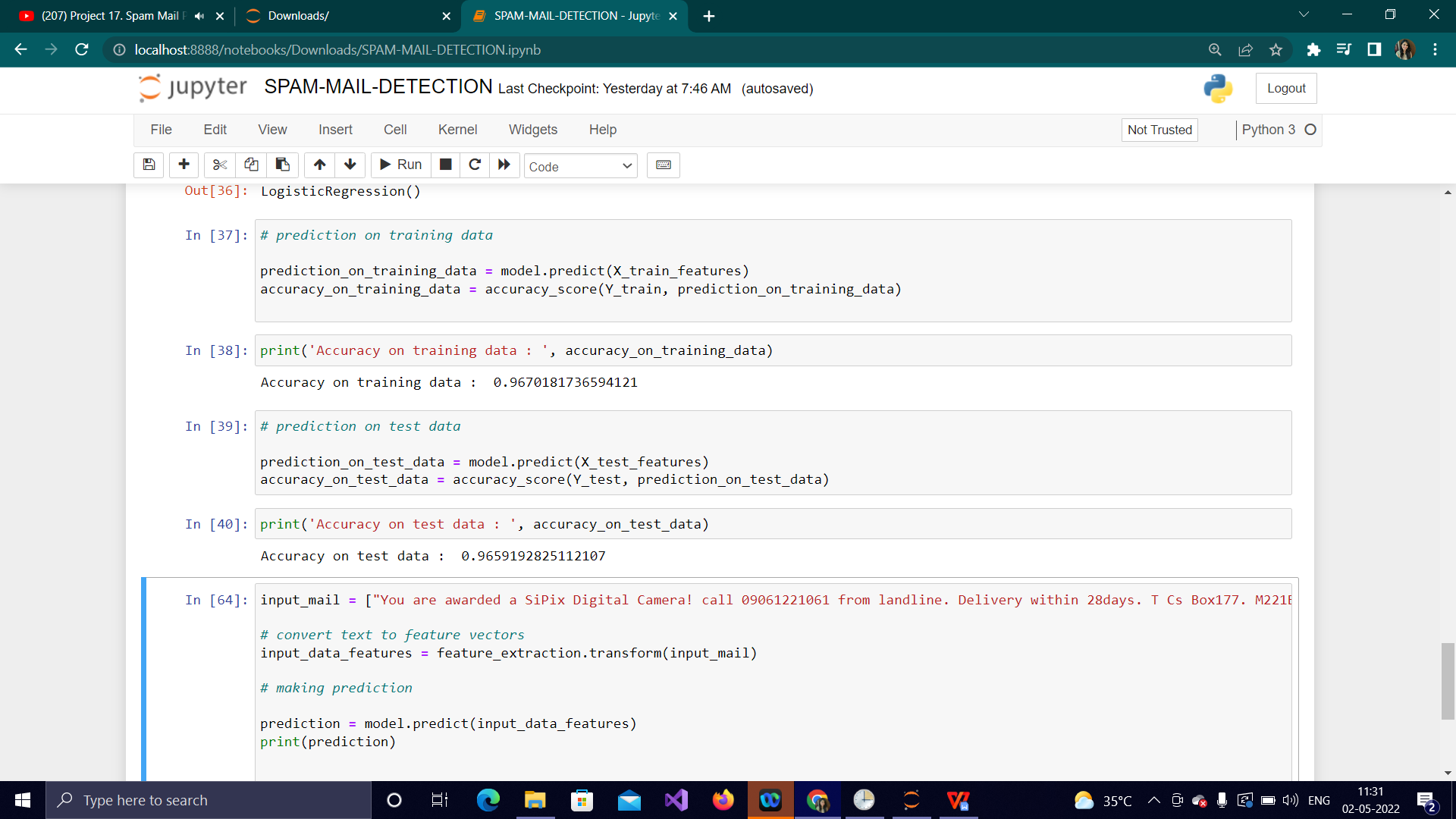


Fig 3.3.7b: Accuracy on test data

**3.3.8 Building a predicive system**

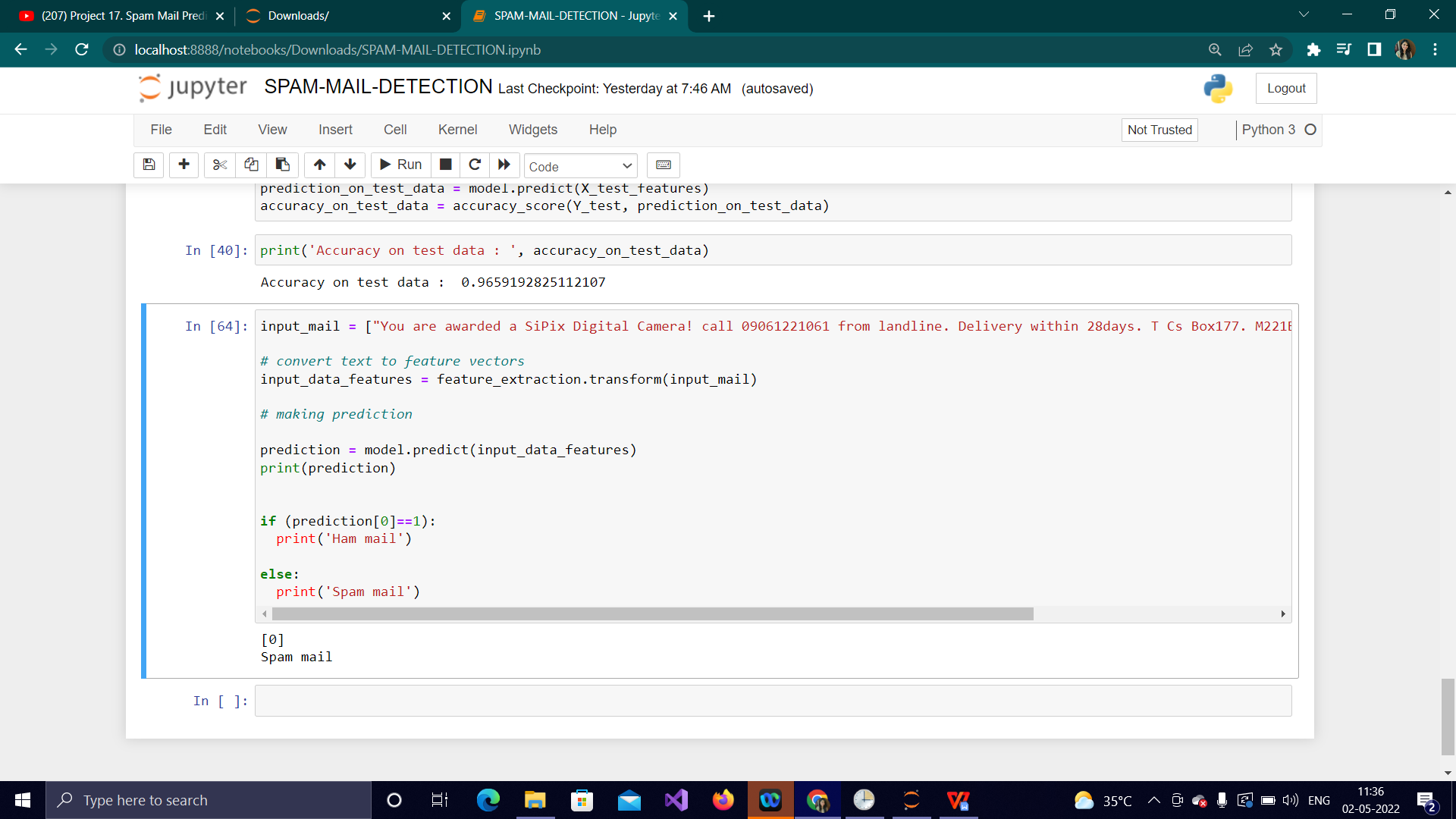


Fig 3.3.8a: Spam mail prediction

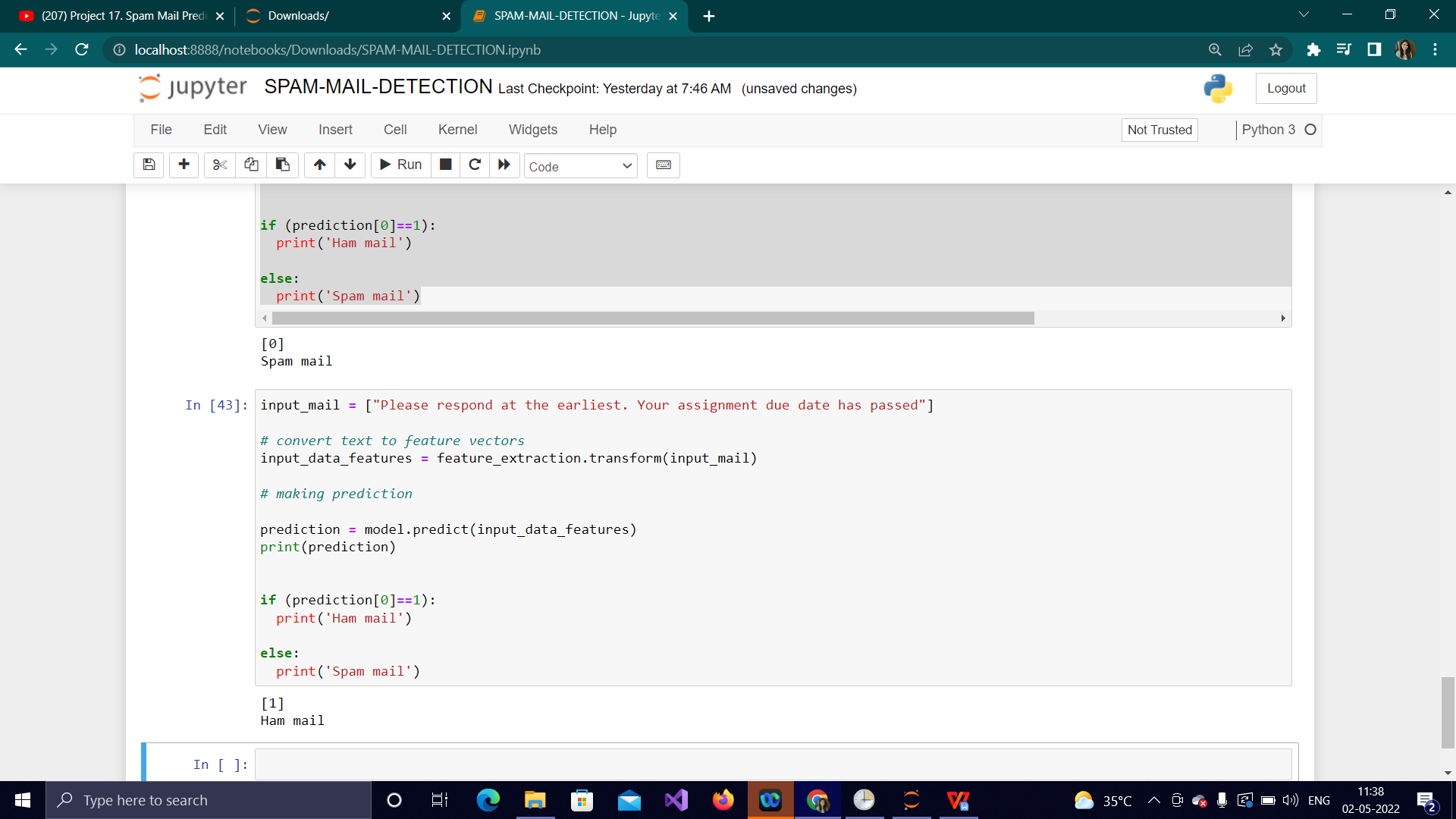


Fig 3.3.8b: Ham mail prediction

**CHAPTER 4**

**REFLECTION NOTES**