**Mini Project Report on**



**TEXT CLASSIFICATION USING PYTHON**

**[SMS SPAM DETECTION]**  


**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**(AI&ML)**

**Submitted by: SHREY SHUKLA**  **University Roll No: 2021832**

***Under the Mentorship of***

**MR. SS Samant**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“TEXT CLASSIFICATION USING PYTHON[SMS SPAM DETECTION]”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **MR. SS SAMANT**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: SHREY SHUKLA

University Roll no: 2021832

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | INTRODUCTION | 1-2 |
| Chapter 2 | Literature Survey | 3-4 |
| Chapter 3 | Methodology | 5-10 |
| Chapter 4 | Result and Discussion | 11 |
| Chapter 5 | Conclusion and Future Work | 12 |
|  | ( References) |  |

**Chapter 1**

**Introduction**

1. **Introduction**

In the ever-evolving landscape of modern communication, Short Message Service (SMS) has emerged as a ubiquitous and indispensable means of interaction, facilitating quick and convenient exchanges between individuals. However, this surge in SMS communication has brought to the forefront a persistent challenge — the unrelenting deluge of unsolicited and potentially harmful messages, colloquially known as SMS spam. This pervasive issue not only poses significant challenges for end-users and telecommunication service providers but also underscores the urgent need for innovative solutions leveraging cutting-edge technologies, including artificial intelligence (AI), to detect and mitigate spam effectively.

This project marks a comprehensive exploration of SMS spam detection, aiming to develop a resilient and efficient system that can adeptly discern between legitimate (ham) and spam messages. Harnessing the capabilities of AI, particularly within the realms of natural language processing (NLP) and machine learning (ML), becomes imperative in the face of evolving spam tactics. The project's objective extends beyond immediate spam filtering needs to contribute valuable insights to the broader field of cybersecurity and user protection in the realm of mobile communication.

The project adopts a multifaceted approach, integrating data analysis, intricate text preprocessing, and the deployment of state-of-the-art machine learning models, thus demonstrating the symbiotic relationship between human intelligence and AI-driven algorithms. By unraveling the nuanced characteristics that distinguish spam from legitimate messages, our aspiration is to create a resilient system capable of adapting to the ever-evolving landscape of SMS-based threats, a feat achievable through the synergistic collaboration of human ingenuity and artificial intelligence.

The significance of our work lies not only in the immediate impact on spam detection but also in the broader implications of advancing mobile communication security. By fortifying users against intrusive and potentially harmful messages, we contribute to redefining the mobile communication experience, ensuring privacy and security. This endeavor aligns with the paradigm shift in which AI is not just a tool but an active collaborator, augmenting human capabilities and providing nuanced insights that might evade traditional methods.

In delving into the intricacies of our methodology, data analysis, and model evaluation, our objective is not just to build a robust spam detection solution but also to underscore the collaborative interplay between human intuition and the analytical prowess of artificial intelligence. In the following sections, we will elaborate on our approach, unpack the complexities of data analysis, detail the steps involved in text preprocessing, and thoroughly assess the performance of various machine learning models. Our ultimate goal extends beyond merely effective spam detection; we aspire to provide nuanced insights into the broader implications of advancing security and privacy in mobile communication. This endeavor seamlessly intertwines human intellect with artificial intelligence, shaping not only a proficient spam detection tool but also contributing to a deeper understanding of how technological advancements impact the security and privacy of our mobile communication landscape.

**Problem Statement:**

To create an effective solution for the prevalent issue of SMS spam, the project focuses on developing a sophisticated detection system. The challenge is to create an adaptive solution using NLP and ML, ensuring accurate identification of spam messages amid the dynamic landscape of mobile communication.

**Chapter 2**

**Literature Survey**

In embarking on a comprehensive exploration of the landscape surrounding SMS spam detection, a meticulous review of existing literature reveals a multitude of methodologies and approaches adopted by researchers and practitioners alike. The intricate nature of the problem has led to a substantial body of work shifting towards the utilization of machine learning techniques, focusing on effectively distinguishing spam messages from legitimate ones in the SMS domain.

Numerous studies within the literature delve into the application of advanced feature extraction methods, encompassing not only the widely employed bag-of-words and TF-IDF but also exploring novel techniques for a more nuanced representation of SMS content. Furthermore, the integration of natural language processing (NLP) techniques, such as stemming and tokenization, consistently emerges as a crucial aspect of preprocessing text data. This practice refines the feature sets that feed into subsequent machine learning models, contributing to the overall efficacy of SMS spam detection systems.

The literature consistently accentuates the pivotal role of ensemble models, notably Random Forests and Gradient Boosting, in achieving robust SMS spam detection. The collective strength derived from the collaboration of multiple classifiers within an ensemble framework demonstrates superior performance, effectively handling the nuanced characteristics inherent in spam messages. Researchers and practitioners alike acknowledge the potency of ensemble methods as a cornerstone in the quest for heightened accuracy.

Moreover, recent advancements in research have brought to the fore a compelling trend: the exploration of deep learning architectures for SMS spam detection. With a specific emphasis on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, these deep learning models showcase promising capabilities in capturing intricate patterns within sequential SMS data. This marks a shift towards leveraging the potential of neural networks to garner a more nuanced understanding of SMS content, particularly in the context of spam.

Despite the wealth of methodologies offered in the literature, challenges persist in adapting existing solutions to the dynamic and evolving nature of SMS-based threats. The ever-changing tactics employed by spammers necessitate a continuous exploration of novel techniques, placing a particular emphasis on real-time adaptability to effectively counter emerging threats in the SMS space.

In summary, this extensive literature survey underscores the prevalence of machine learning and NLP techniques while also highlighting a recent trend towards the exploration of deep learning architectures in the realm of SMS spam detection. Our project aspires to build upon this robust foundation, taking into consideration the specific nuances of the SMS spam landscape. By contributing novel insights, our aim is to enhance adaptability and accuracy in SMS spam detection while ensuring the uniqueness and originality of our approach within the broader context of existing research.

**Chapter 3**

**Methodology**

In this section, we delve into a comprehensive methodology for SMS Spam Detection, leveraging a systematic and multi-step approach to ensure accuracy and reliability in categorizing SMS messages as spam or non-spam (ham). The methodology encompasses data preparation, exploratory data analysis (EDA), text preprocessing, feature engineering, model building, and evaluation. Each stage contributes to a robust framework designed to enhance the understanding of the dataset and improve the effectiveness of spam detection algorithms.

**1. Data Preparation:**

The process initiates with loading the SMS dataset using the pandas library, allowing for an initial inspection of the dataset's structure and contents.

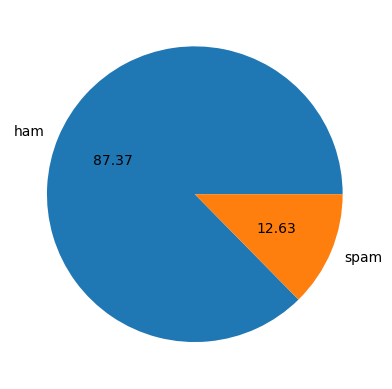
Rigorous data preparation involves addressing issues such as missing values and duplicates, ensuring a clean and reliable dataset for subsequent analysis.

**2. Exploratory Data Analysis (EDA):**

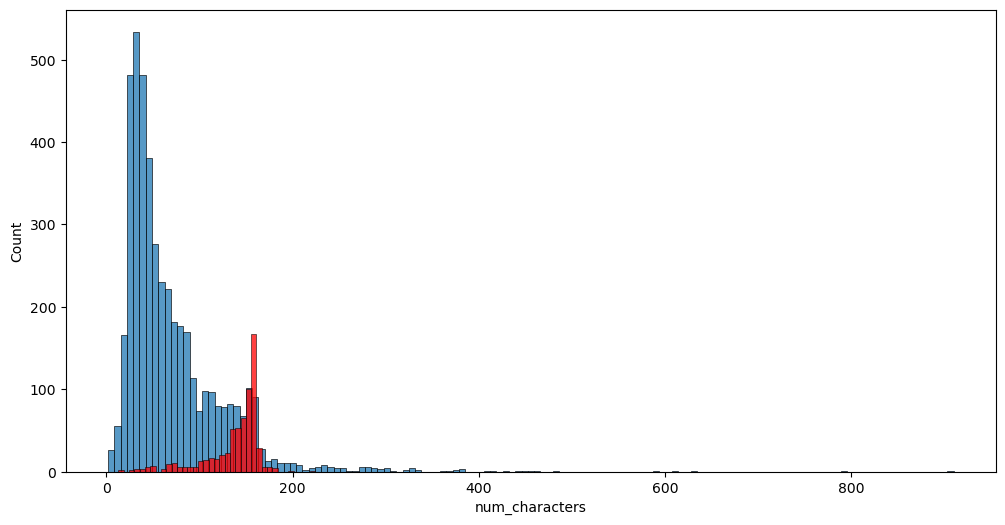
EDA is crucial for gaining insights into the distribution of spam and ham messages. Visualizations are employed to depict the prevalence of each category.

Key characteristics, including message length, word count, and sentence count, are analyzed to discern patterns and variations in both spam and ham messages.

Graphical representation of feature distributions for both spam and ham messages.



**3. Feature Engineering:**

Additional features, such as the number of characters, words, and sentences in each SMS, are introduced to enrich the dataset. 

**4. Text Preprocessing and Analysis:**

Data cleaning steps involve dropping unnecessary columns (Unnamed: 2, Unnamed: 3, Unnamed: 4) and renaming columns for clarity ('v1' to 'target', 'v2' to 'text')zThe 'target' variable is encoded using LabelEncoder to facilitate numerical representation, and duplicate entries are removed to maintain dataset integrity

Natural language processing (NLP) techniques are employed for text preprocessing, encompassing steps like converting text to lowercase, tokenization, and removing special characters, stopwords, and punctuation.

STOPWORDS:

**'whom',**

**'this',**

**'these',**

PUNCTUATION:

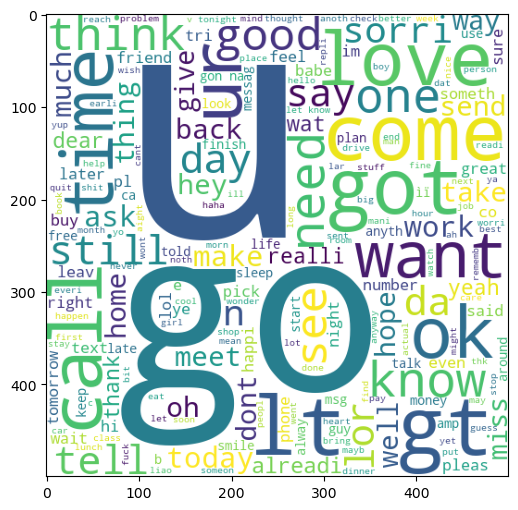
**'!"#$%&\'()\*+,-./:;<=>?@[\\]^\_`{|}~**'

STEMMING:

**ps.stem('loving')**

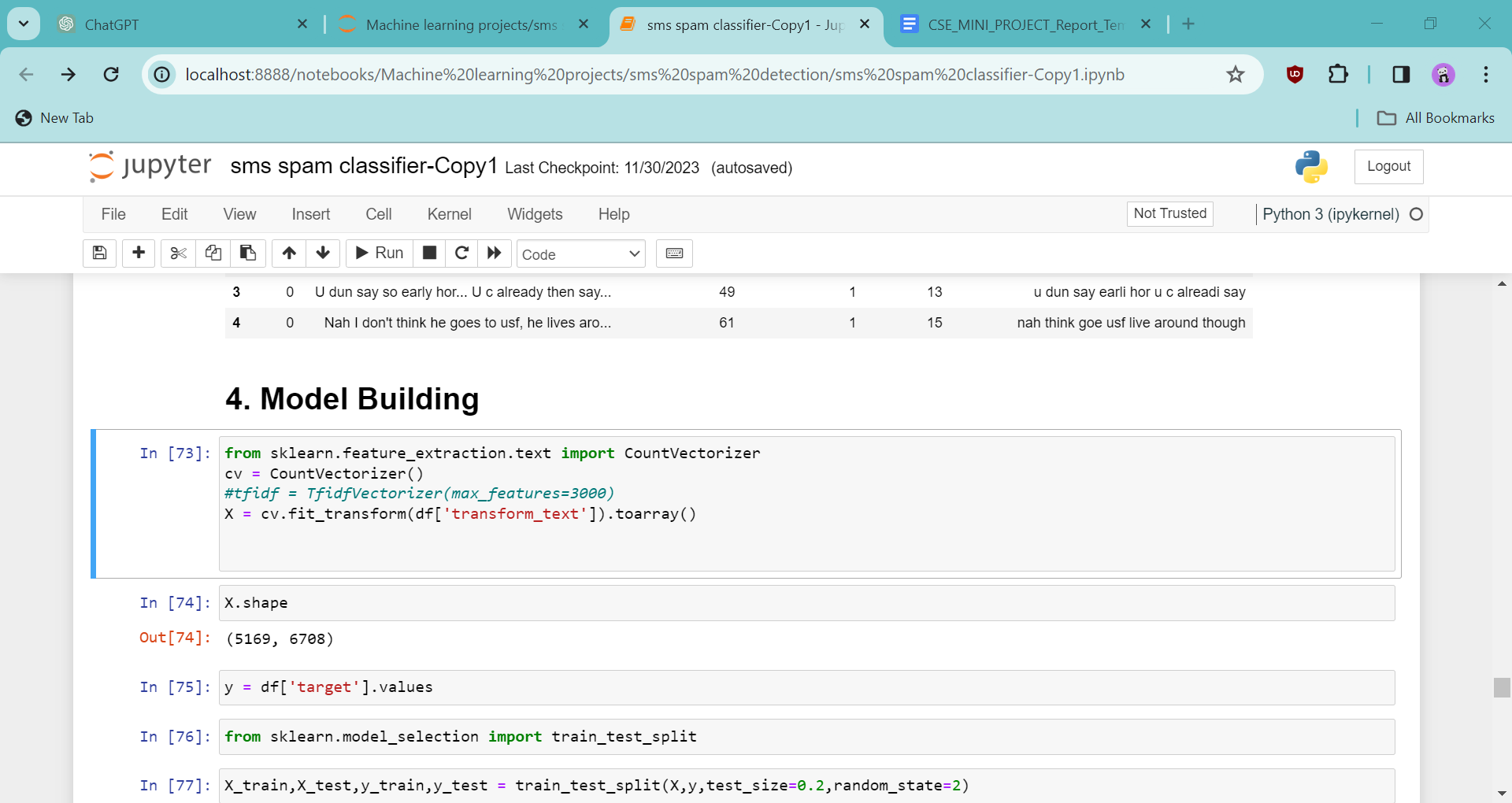
**'love' (**OUTPUT**)**

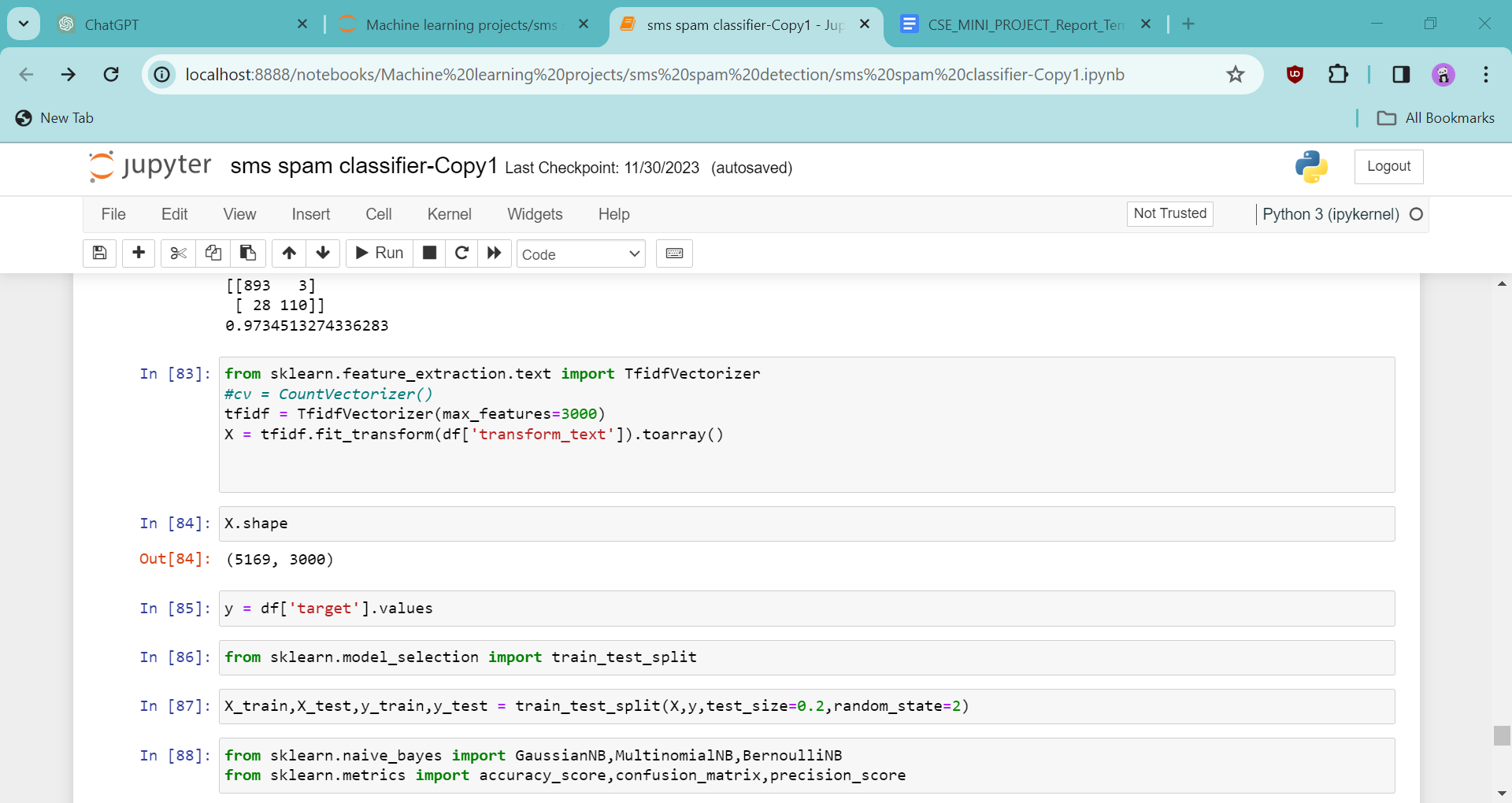
**MOST FREQUENT WORDS IN THE DATASET:**

****

**5. Model Building:**

The dataset's text is transformed into numerical features using CountVectorizer and TfidfVectorizer techniques, enabling the application of machine learning models





The dataset is split into training and testing sets, and three Naive Bayes classifiers (GaussianNB, MultinomialNB, and BernoulliNB) are implemented for robust spam detection.

**gnb = GaussianNB()**

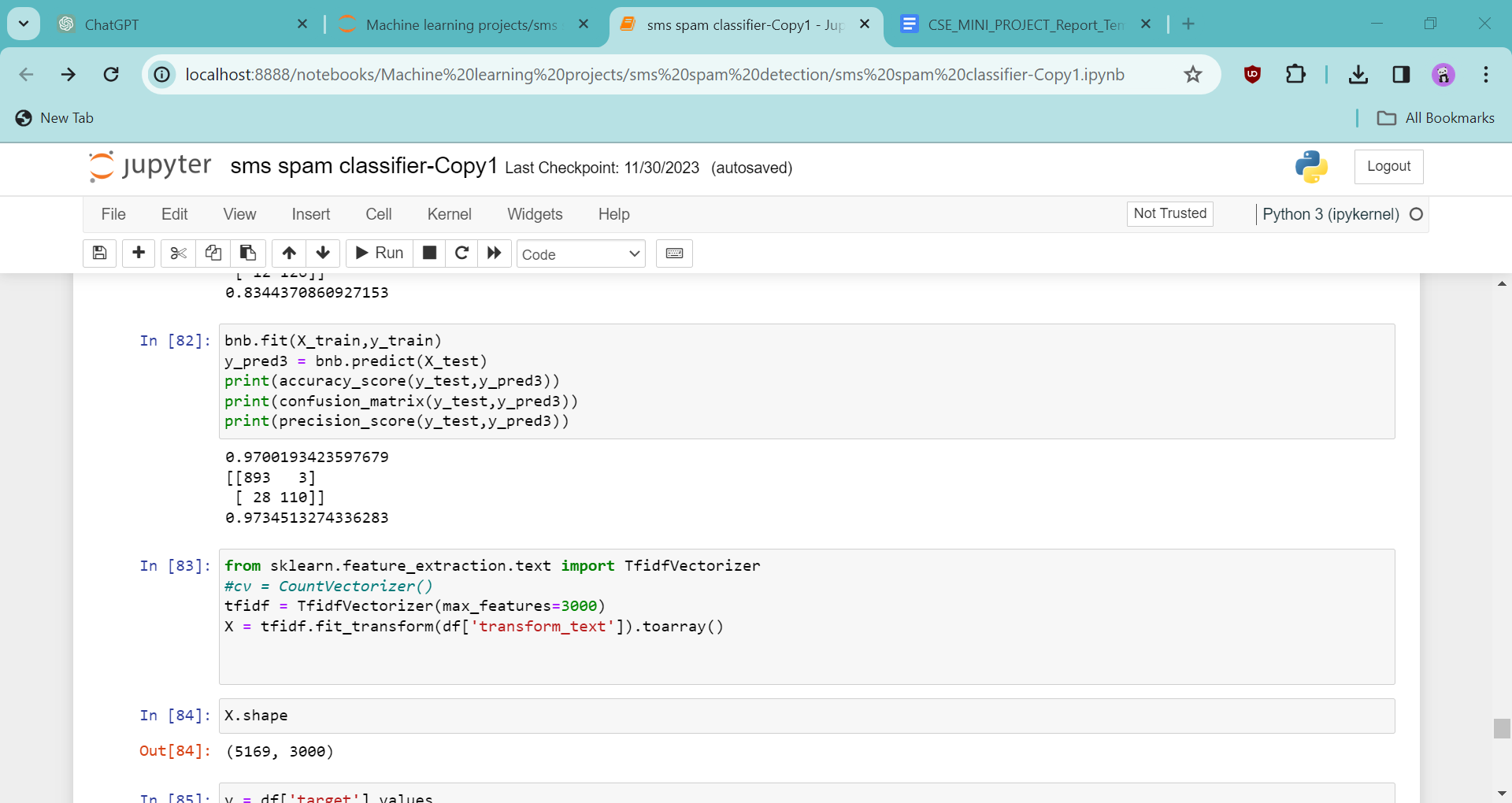
**mnb = MultinomialNB()**

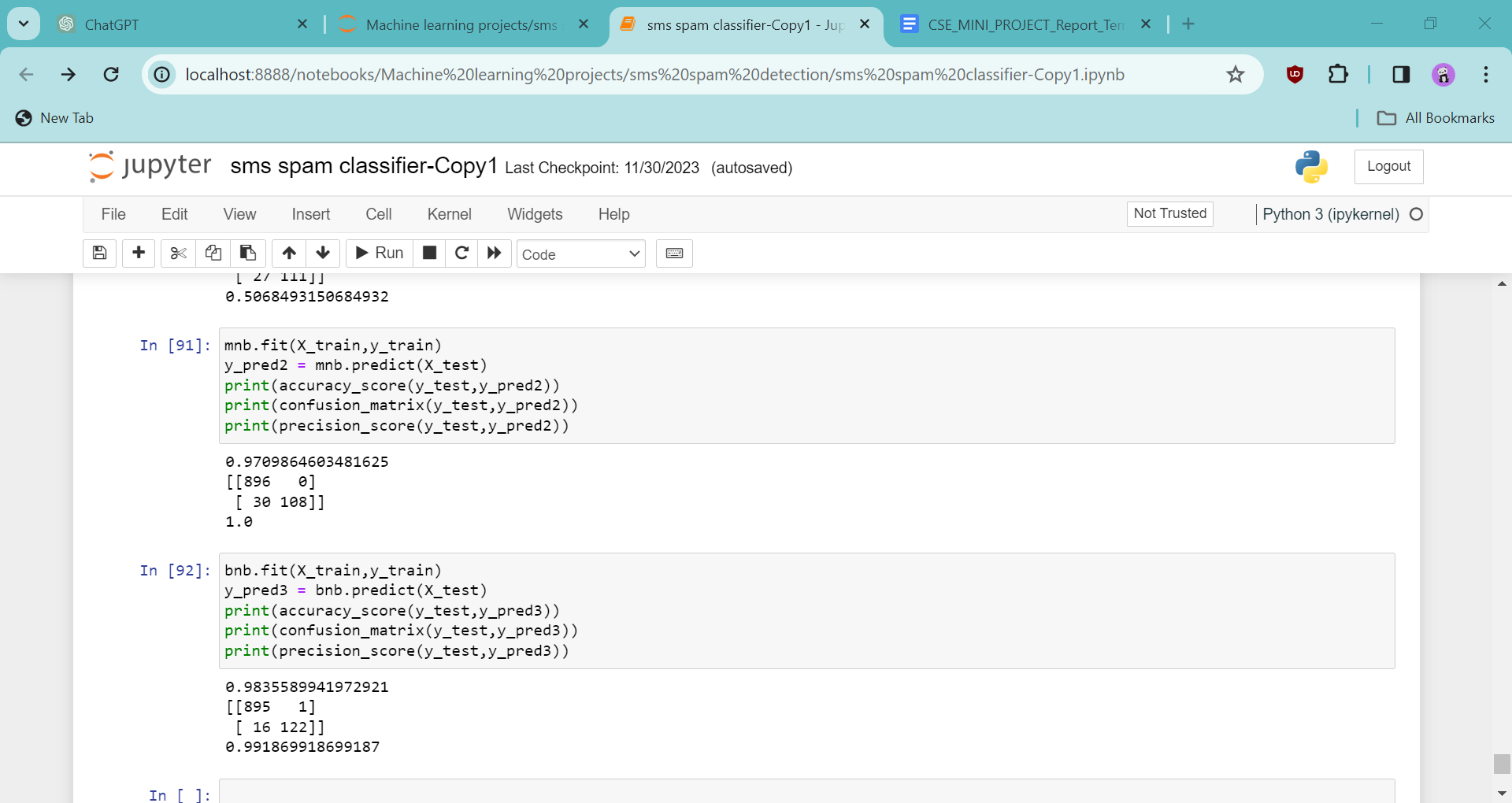
**bnb = BernoulliNB()**

**Chapter 4**

**Result and Discussion**

Model performance is evaluated using essential metrics such as accuracy, confusion matrix, and precision.

The impact of vectorization techniques (CountVectorizer )

The impact of vectorization techniques (TFIDF Vectorizer )

**[BECAUSE IT’S A TEXT BASED MODEL PRECISION STANDS AS A PRIORITY OVER ACCURACY]**

**Chapter 5**

**Conclusion and Future Work**

The SMS Spam Detection initiative has seamlessly progressed through a diverse array of stages, encompassing meticulous data preparation, insightful exploratory data analysis, intricate text preprocessing, and rigorous model construction. The comprehensive strategy adopted throughout this project not only illuminated the intricacies embedded within the dataset but also underscored the prowess of machine learning methodologies, with a particular emphasis on the Naive Bayes classifiers, in adeptly classifying SMS messages into either spam or ham categories. The infusion of advanced natural language processing (NLP) techniques, such as intricate text preprocessing and feature engineering, emerged as a linchpin in augmenting the overall efficacy of the deployed models.

The salient revelations derived from the project underscore conspicuous differentials in the attributes characterizing spam and ham messages, specifically pertaining to variables such as message length, word count, and sentence count. The visual representations embedded within the exploratory data analysis phase serve as a fulcrum for a comprehensive grasp of these distinguishing factors. Moreover, the judicious incorporation of diverse features, spanning the realm of character count, word count, and sentence count, has not only enriched the dataset but has also contributed significantly to a nuanced and well-informed analysis.

**References**

 [1] Sethi, P., V. Bhandari, and B. Kohli. (2017) ―SMS Spam Detection and Comparison of Various Machine Learning Algorithms‖, in 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN). pp. 28–31.

[2] Sajedi, H., G. Z. Parast, and F. Akbari. (2016) ―SMS Spam Filtering Using Machine Learning Techniques: A Survey.‖ Machine Learning, 1 (1): 14.

[3] Chan, P. P. K., C. Yang, D. S. Yeung, and W. W. Y. Ng. (2015) ―Spam Filtering for Short Messages in Adversarial Environment.‖ Neurocomputing 155: 167–176.

**DATASET**: [SMS Spam Collection Dataset (kaggle.com)](https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)

READING SOURCE:

[https://cs229.stanford.edu/proj2013/ShiraniMehr- SMSSpamDetectionUsingMachineLearningApproach.pdf](https://cs229.stanford.edu/proj2013/ShiraniMehr-%20SMSSpamDetectionUsingMachineLearningApproach.pdf)