

**SMS SPAM CLASSIFIER**

A Project Report Submitted

in Partial Fulfillment of the Requirements

**International Institute of Information Technology, Bhubaneswar 2ND YEAR B. Tech**

In

**COMPUTER ENGINEERING**

By:

1. **APURBA RANJAN ROLL NO: B522008**
2. **DEBASHISH SAHOO ROLL NO: B522019**
3. **PRIYARANJAN KUMAR ROLL NO: B522046**
4. **RANJAN KUMAR ROLL NO: B522048**
5. **SWATA SWAYAM DAS ROLL NO: B522060**

**INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, BHUBANESWAR**

**[APRIL-2024]**

# UNDERTAKING

We, the undersigned members of the group undertaking the project titled "SMS Spam Classifier" for our B. Tech in Computer Engineering , 2nd Year, at the International Institute of Information Technology, Bhubaneswar, hereby declare that the project report submitted by us is our original work and has not been submitted elsewhere for any other purpose.

We further declare that:

* 1. The project work was carried out under the guidance of Dwibik Patra , Python Professor , Computer Science.
  2. All sources of information used have been duly acknowledged through proper citations and references.
  3. No part of this project violates or infringes upon the rights of any third party, including but not limited to copyright, trademark, privacy, or other personal or proprietary rights.
  4. Any data, code, or materials obtained or used from external sources have been appropriately credited and referenced.

We understand that any violation of the above declarations may result in disciplinary action as per the institute's rules and regulations.

Date: 13 April 2024 Signatures:

1. Apurba Ranjan - Roll No: B522008
2. Debashish Sahoo - Roll No: B522019
3. Priyaranjan Kumar - Roll No: B522046
4. Ranjan Kumar - Roll No: B522048
5. Swata Swayam Das - Roll No: B522060

# Acknowledgements

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this project.

First and foremost, we would like to thank Dwibik Patra, Python Professor , Computer Science for their invaluable guidance, support, and encouragement throughout the duration of this project. Their expertise and insights have been instrumental in shaping our approach and guiding us through the various stages of the project.

We would also like to extend our thanks to the faculty members of the Department of Computer Science at the International Institute of Information Technology, Bhubaneswar, for their valuable inputs and feedback.

We are grateful to [Name of Dataset Provider/Source] for providing the dataset used in this project. Their contribution has been vital to the success of our research.

Our sincere thanks to our classmates and friends for their support and encouragement throughout this endeavor.

Lastly, we would like to express our heartfelt gratitude to our families for their unwavering support, understanding, and patience during the course of this project.

1. APURBA RANJAN ROLL NO: B522008
2. DEBASHISH SAHOO ROLL NO: B522019
3. PRIYARANJAN KUMAR ROLL NO: B522046
4. RANJAN KUMAR ROLL NO: B522048
5. SWATA SWAYAM DAS ROLL NO: B522060

# Table of Contents

* 1. **Introduction of the Project.**
  2. **System Requirements of the Project.**
  3. **Python Coding.**
     + Introduction
     + Data Cleaning
     + Exploratory Data Analysis (EDA)
     + Text Preprocessing
     + Model Building
     + Evaluation
     + Improvement
     + Website Development
     + Deployment
  4. **Output of the Project.**
  5. **References.**

# Introduction of the Project

Our team of B. Tech 2nd year students at the International Institute of Information Technology, Bhubaneswar, has embarked on a project to develop an SMS spam classifier. The objective of this project is to create a machine learning model capable of accurately identifying spam messages in SMS communication.

*We express our gratitude to our subject professor , Mr. Dwibik Patra, for his guidance and support throughout the project. Additionally, we thank our peers for their collaboration and assistance.*

**The project contains-**

1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Text Preprocessing
4. Model Building
5. Evaluation
6. Improvement
7. Website Development
8. Deployment

Our project roadmap encompasses key stages, including data cleaning, exploratory data analysis (EDA), text preprocessing, model building, evaluation, improvement, website development, and deployment. Each stage plays a crucial role in the development and refinement of our SMS spam classifier, ultimately leading to a robust and effective solution.

Throughout this project, we are grateful for the guidance and support of our subject professor, Mr. Dwibik Patra, whose expertise has been instrumental in navigating the complexities of machine learning and natural language processing. Additionally, we extend our thanks to our peers for their collaboration and assistance, as well as to the academic community for providing valuable resources and insights.

As we embark on this journey, we are excited about the opportunities this project presents to deepen our understanding of machine learning and contribute to the development of innovative solutions in the field of

communication technology. We look forward to sharing our progress and findings with our peers, educators, and the wider community.

**PROCESS**

Our project follows a systematic approach, encompassing several key stages to ensure the successful development and deployment of our SMS spam classifier. Below is an outline of the process we will undertake:

## Data Cleaning:

We begin by cleaning the SMS dataset to ensure consistency and remove any inconsistencies or errors that may affect the performance of our model. This involves handling missing values, removing duplicates, and formatting the data for further analysis.

## Exploratory Data Analysis (EDA):

With the cleaned dataset, we conduct exploratory data analysis to gain insights into the distribution and characteristics of spam and non-spam messages. This helps us understand the underlying patterns and features that distinguish spam from legitimate messages.

## Text Preprocessing:

Before building our model, we preprocess the text data to prepare it for analysis. This involves converting text to lowercase, tokenizing, removing special characters, stopwords, and stemming or lemmatizing words to reduce dimensionality and improve model performance.

## Model Building:

Using the preprocessed data, we train various machine learning models to classify SMS messages as spam or non-spam. We experiment with different algorithms such as Naive Bayes, Support Vector Machines (SVM), and ensemble methods to identify the most effective model for our task.

## Evaluation:

Once we have trained our models, we evaluate their performance using appropriate metrics such as accuracy, precision, recall, and F1-score. This allows us to assess the effectiveness of each model and select the best-performing one for deployment.

Improvement:

Based on the evaluation results, we iteratively refine our model by fine-tuning hyperparameters, experimenting with different feature engineering techniques, and exploring alternative algorithms. This continuous improvement process helps us optimize the performance of our classifier.

## Website Development:

To make our SMS spam classifier accessible to users, we develop a user-friendly website where users can input text messages and receive instant feedback on whether they are spam or not. The website will also provide information about the project and how the classifier works.

## Deployment:

Finally, we deploy our trained model and website to a production environment, making it available for real-world use. This involves setting up servers, deploying the website, and ensuring that the classifier performs reliably and efficiently in a live environment.

By following this structured process, we aim to develop a highly accurate and reliable SMS spam classifier that can help users protect themselves from unwanted and potentially harmful messages.

# System Requirements of the Project

## Recommended System Requirements

Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz. Disk space: 2 to 4 GB.

Operating systems: Windows® 10, MACOS, and UBUNTU. Python Versions: 3.X.X or Higher.

## Minimum System Requirements

Processors: Intel Atom® processor or Intel® Core™ i5 processor. Disk space: 4 GB.

Operating systems: Windows 10 or later, MACOS, and UBUNTU. Python Versions: 2.7.X, 3.6.X.

## Prerequisites before installing MySQL Connector Python

You need root or administrator privileges to perform the installation process.

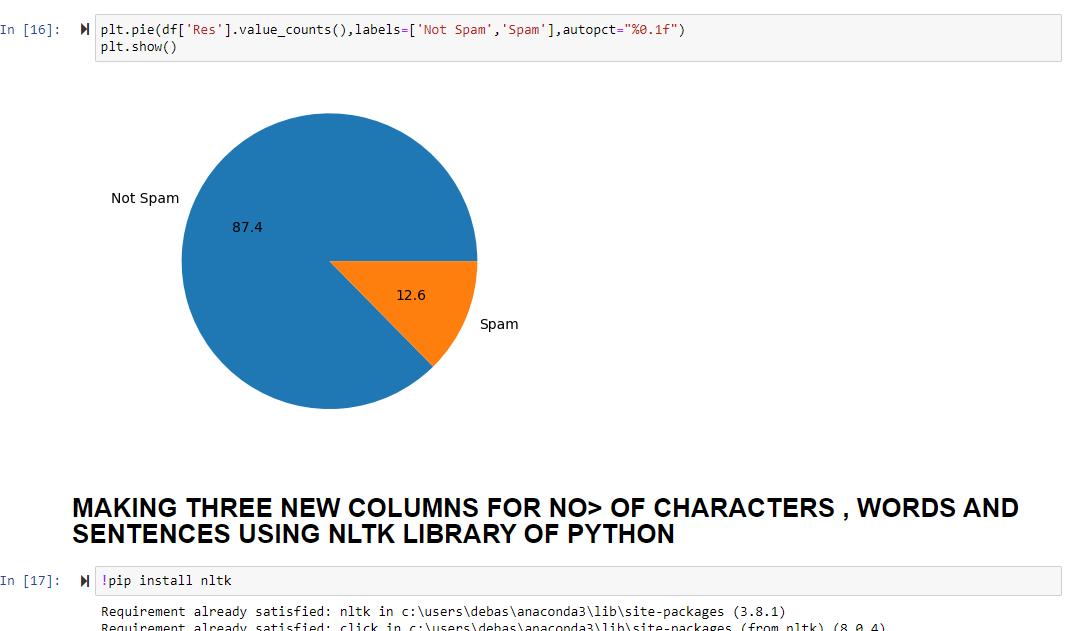
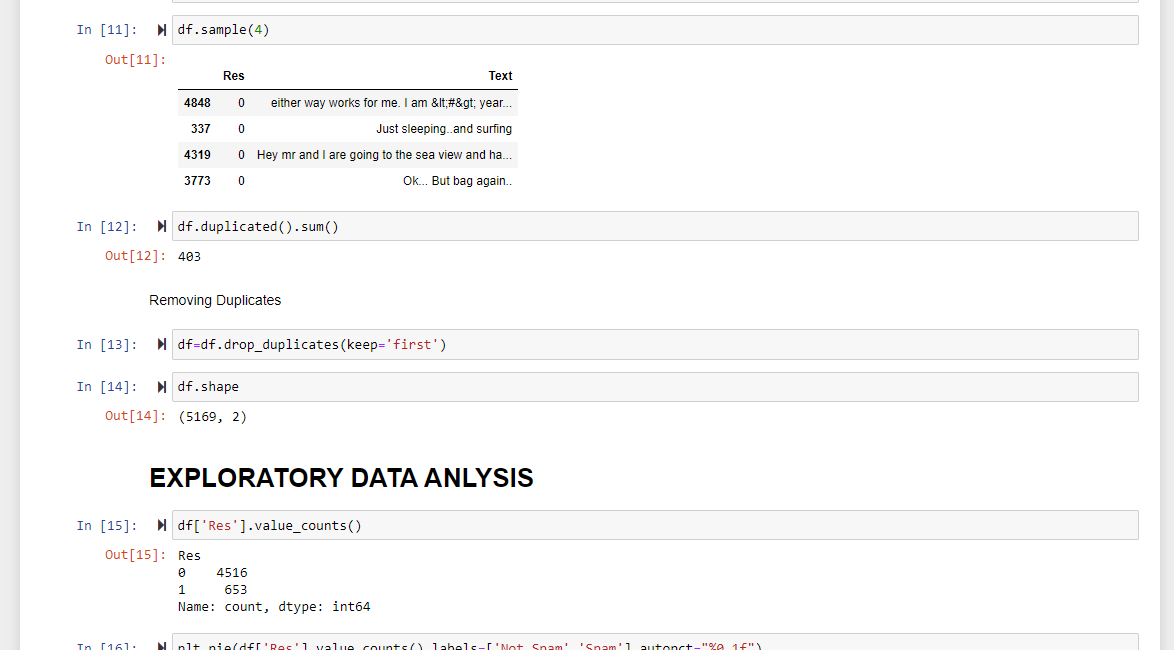
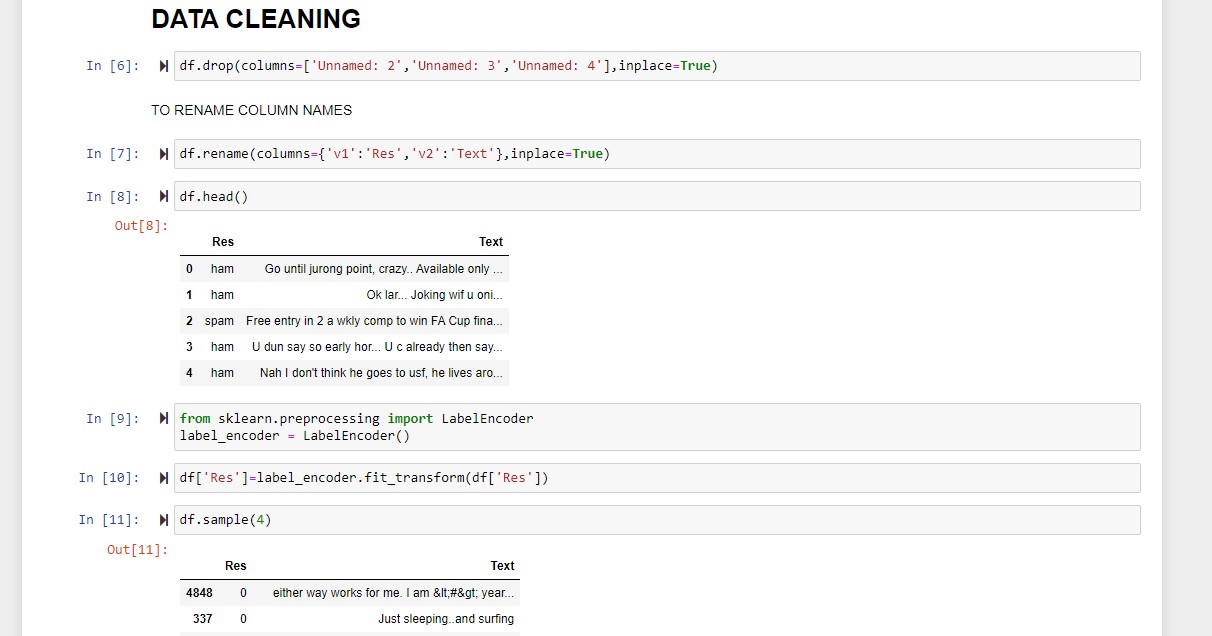
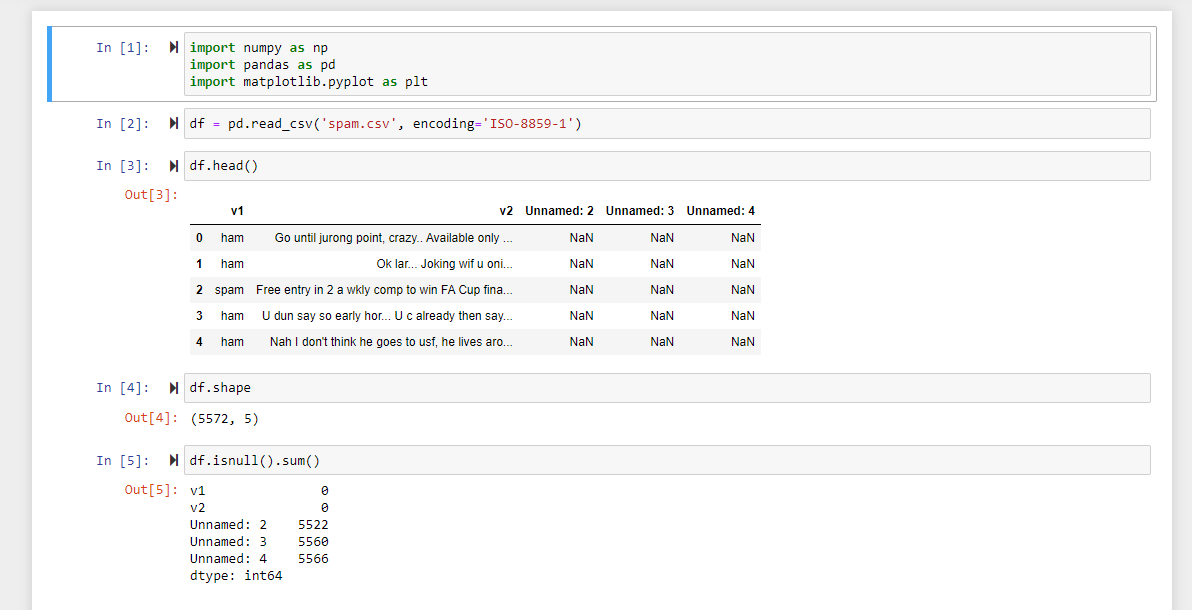
Python must be installed on your machine.

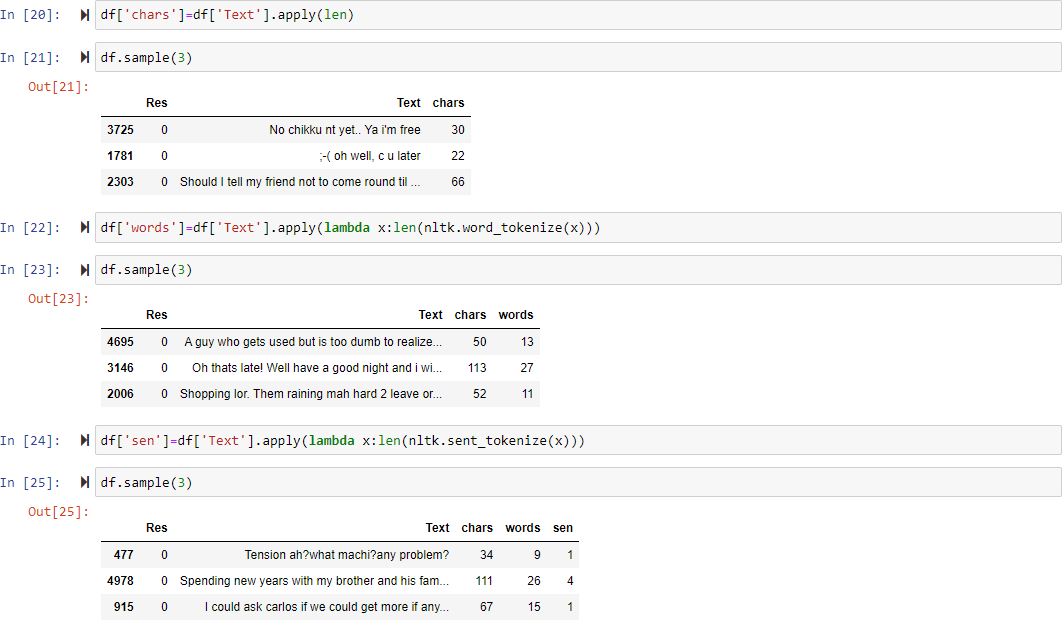
Note: – MySQL Connector Python requires python to be in the

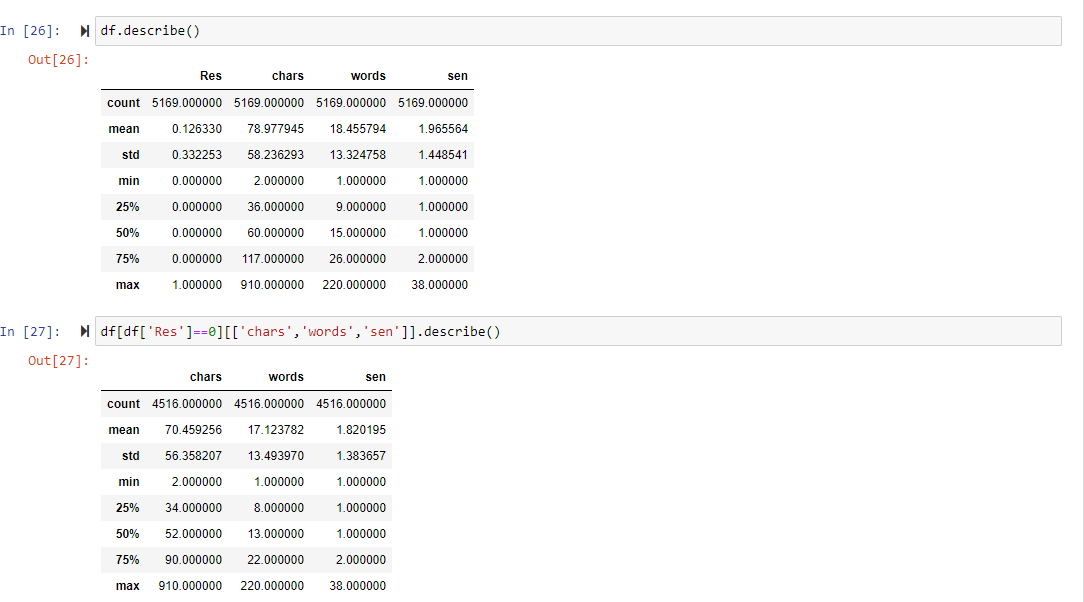
system’s PATH. Installation fails if it doesn’t find Python.

On Windows, If Python doesn’t exist in the system’s PATH, please manually add the directory containing python.exe yourself.

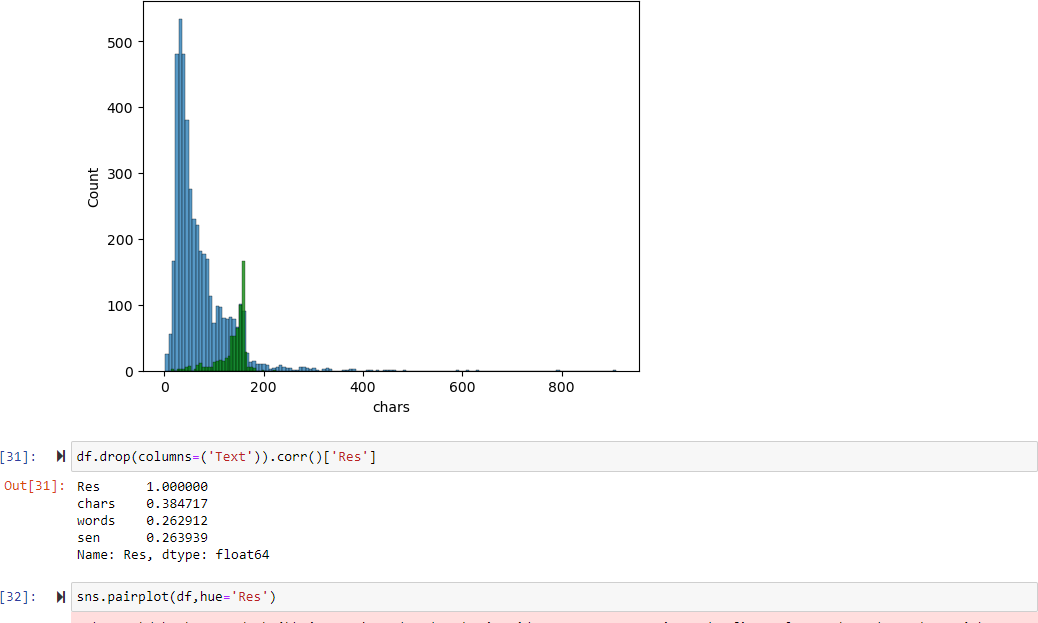
# Source Code:-

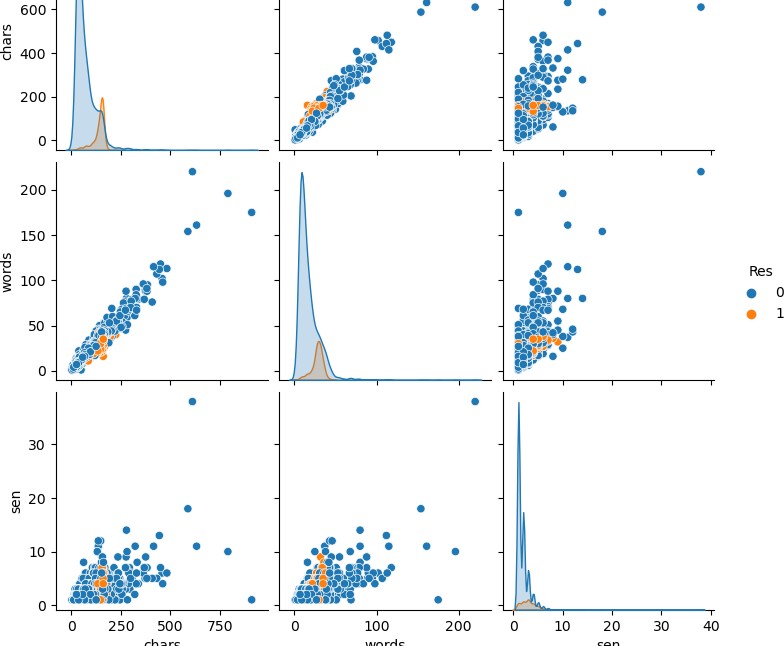


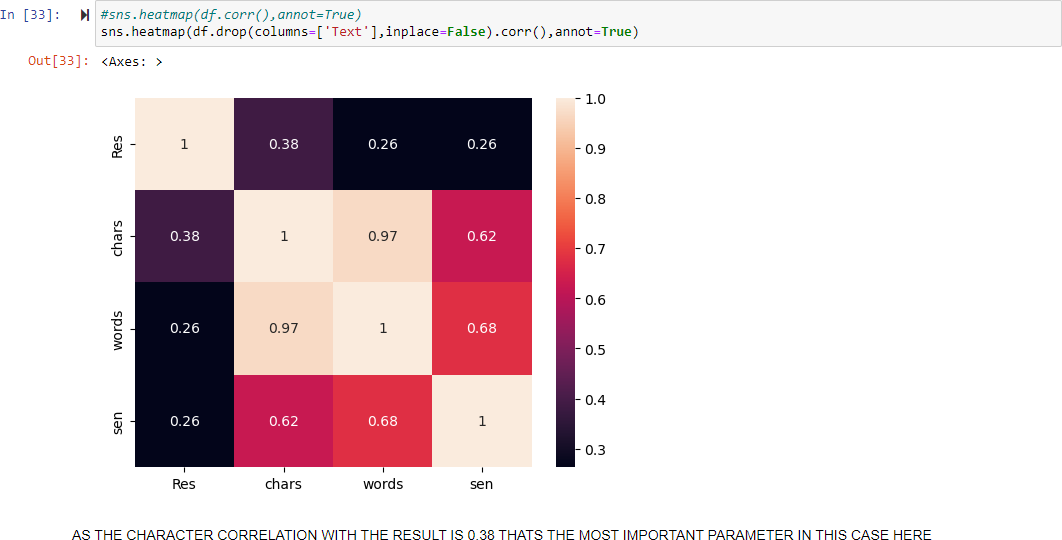


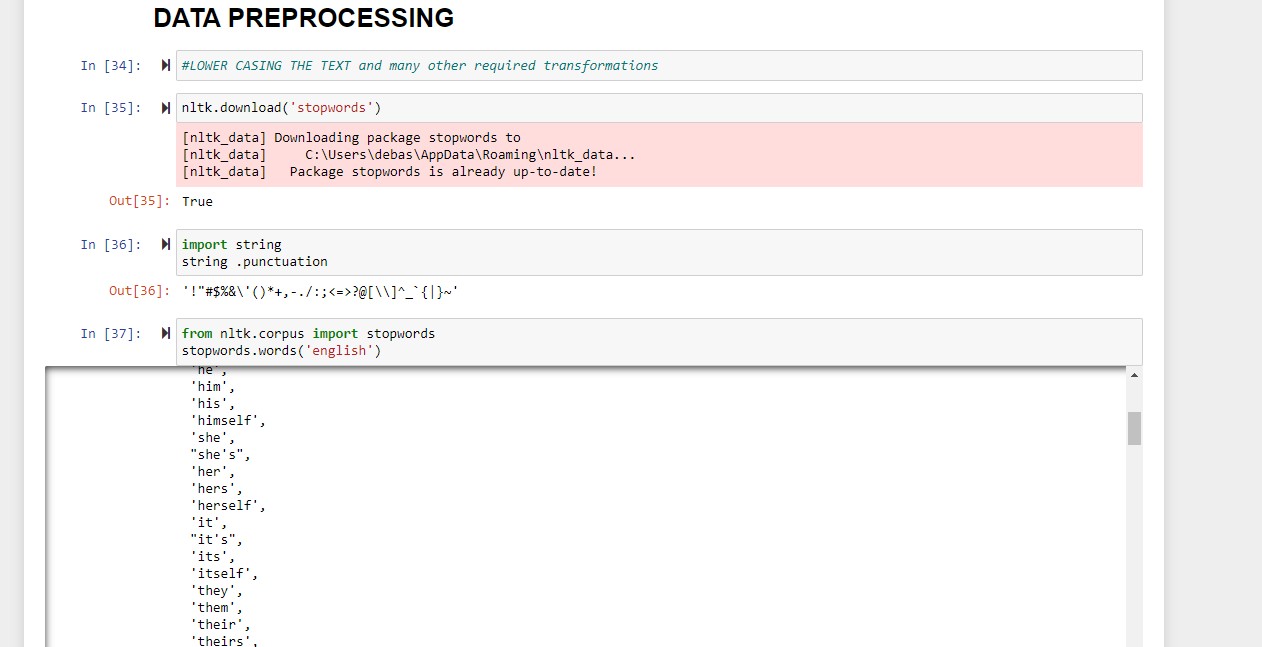


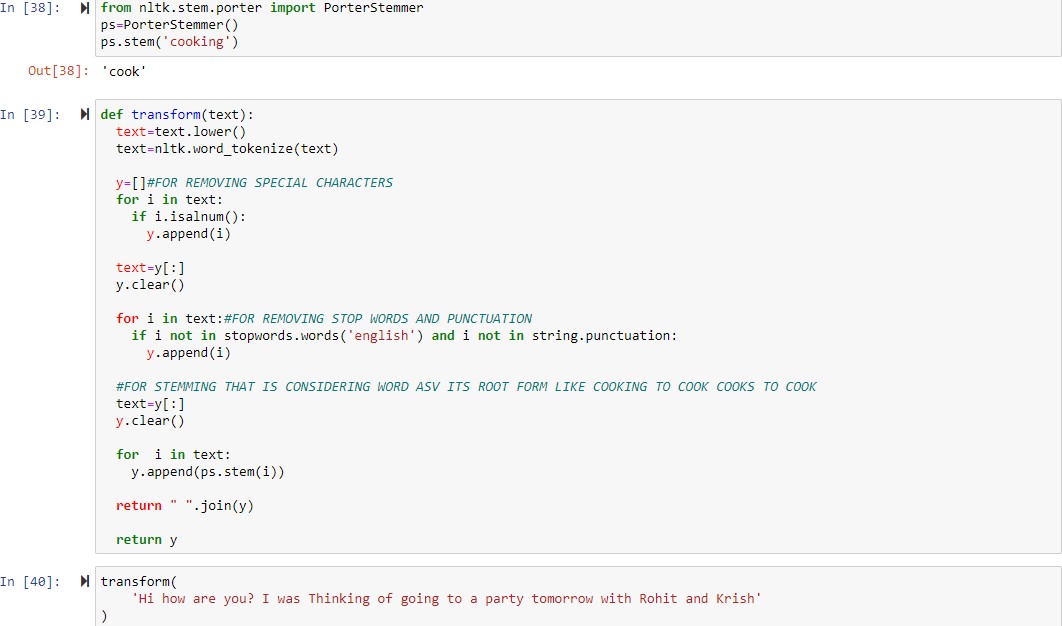


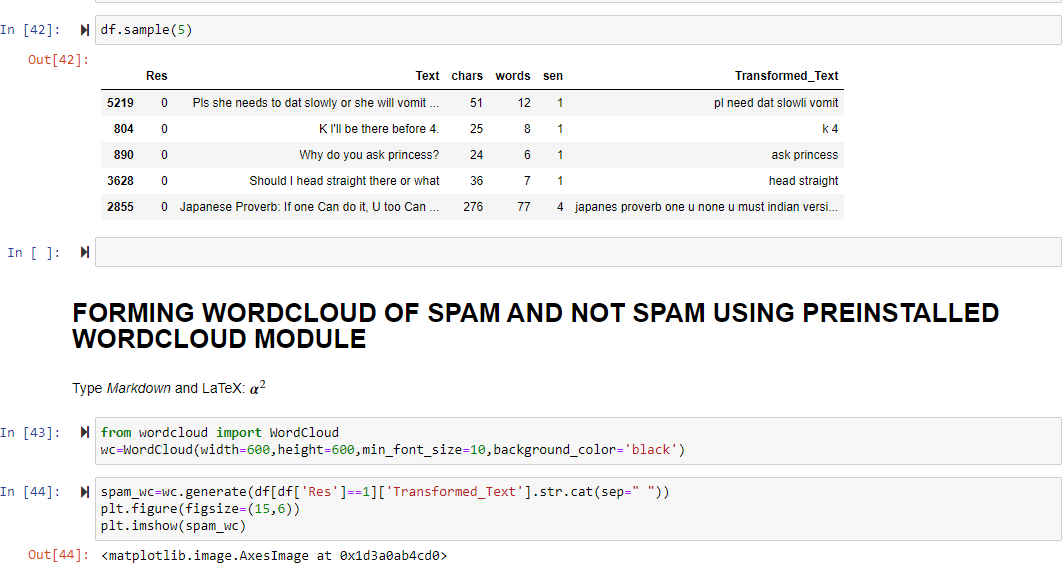




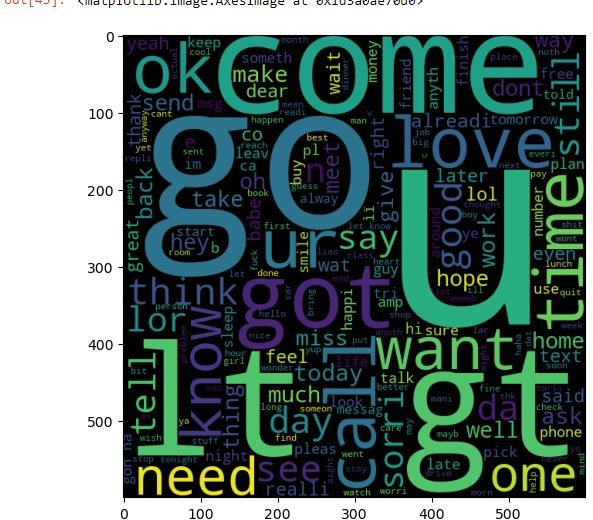


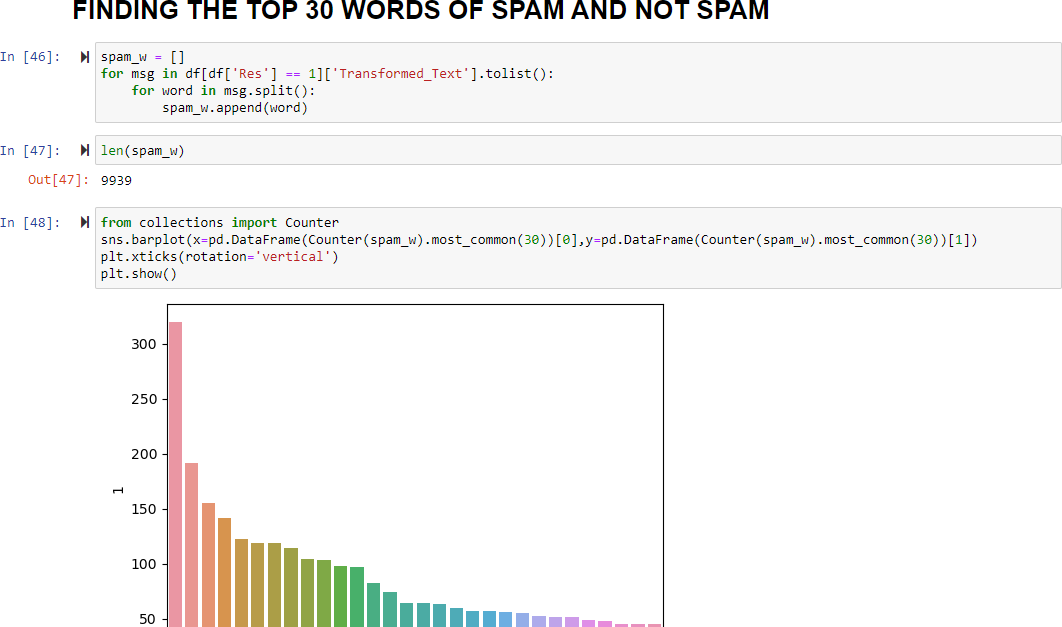


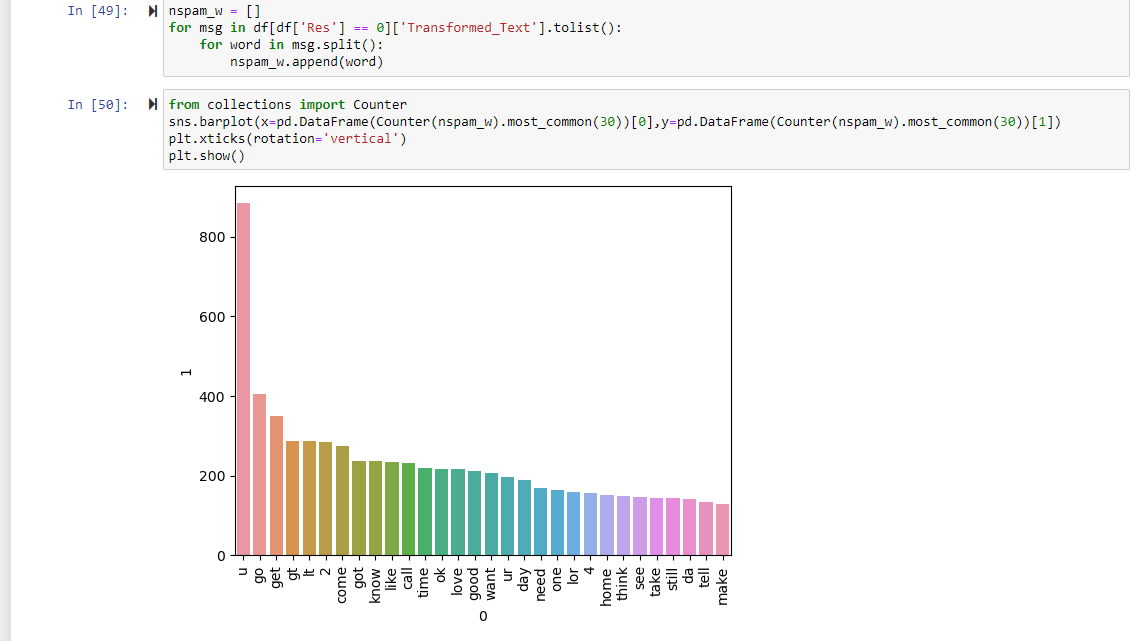










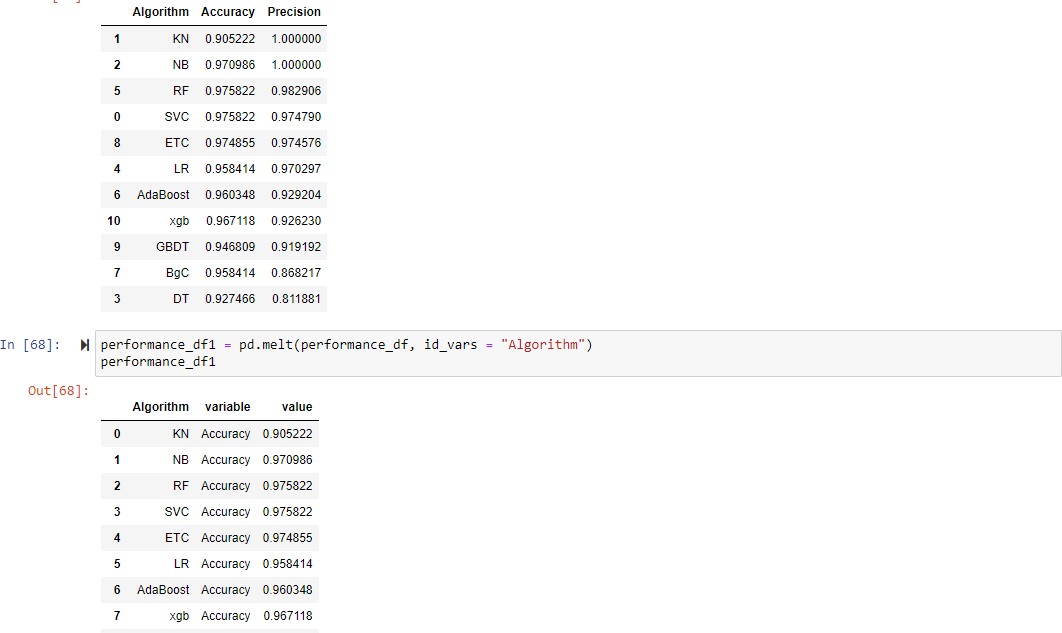


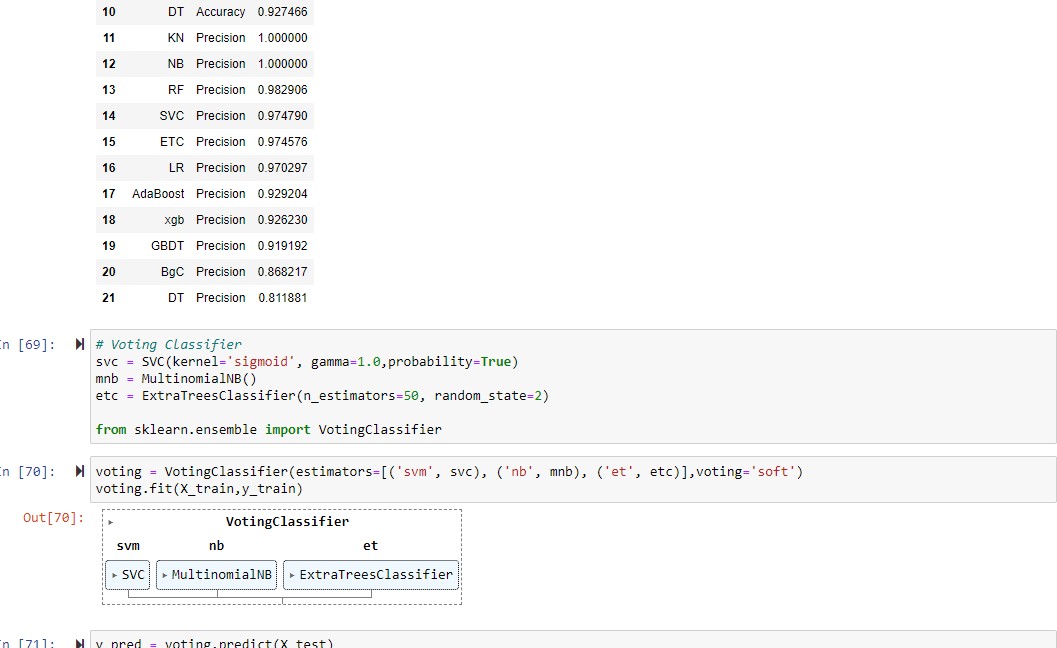






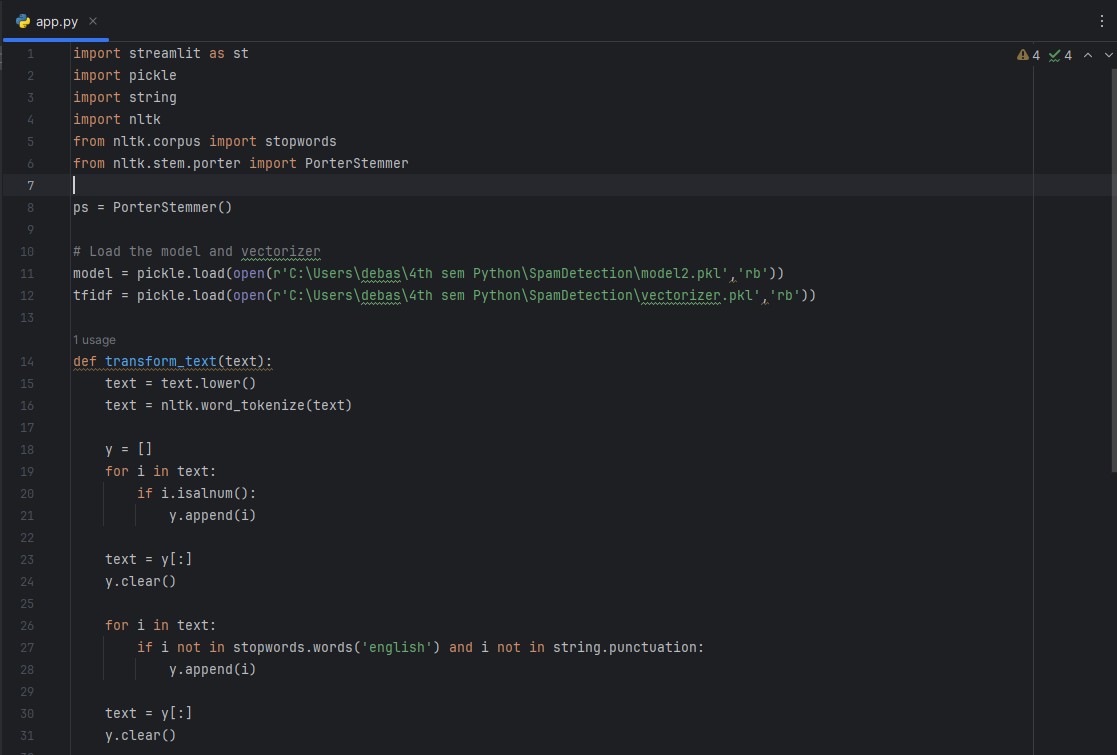


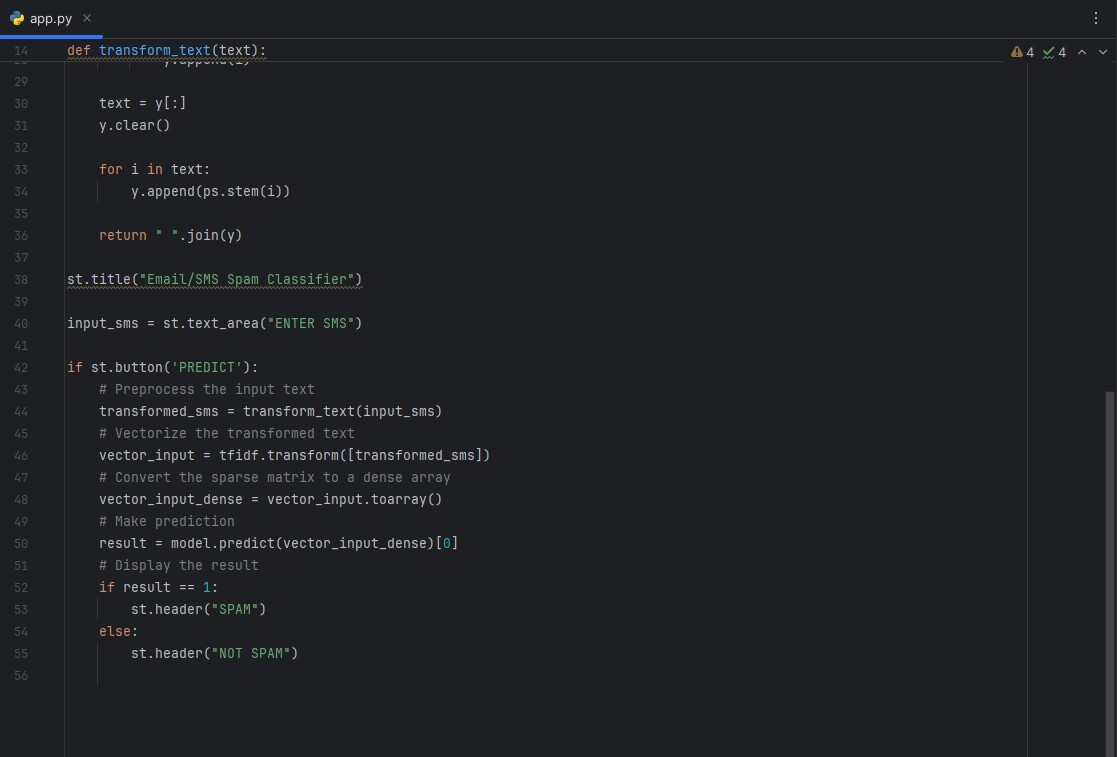






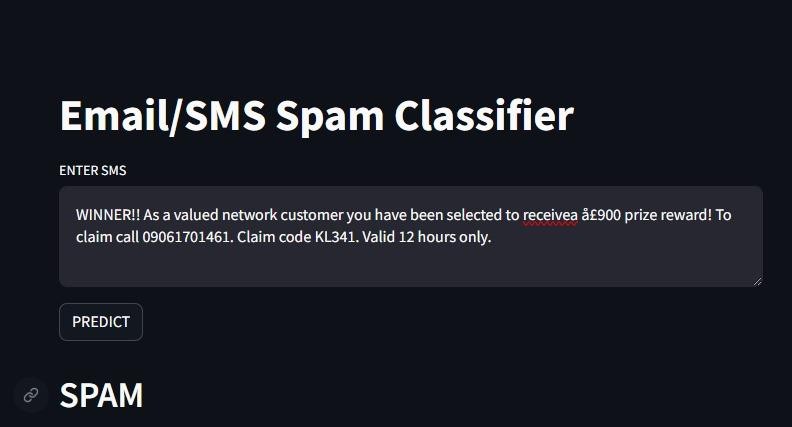
APP. PY





# Output of the Project

**Finally, we conclude our work and present the output of the Project.**



# References

* 1. python.org
  2. Campus x
  3. Scikit. learn
  4. PythonChallenge.com
  5. Google’s Python Class
  6. LearnPython.org
  7. Nltk.org

**DATASET LINK :-**

**https://**[**www.kaggle.com/datasets/uciml/sms-spam-**](http://www.kaggle.com/datasets/uciml/sms-spam-) **collection-dataset**

CONCLUSION

In conclusion, our journey in developing the SMS spam classifier has been both challenging and rewarding. Through meticulous data cleaning, thorough exploratory data analysis, and comprehensive text preprocessing, we have gained valuable insights into the nature of SMS messages and the characteristics of spam versus legitimate messages.

Our model building efforts have resulted in the creation of an effective classifier capable of accurately distinguishing between spam and non-spam messages. Leveraging machine learning algorithms and natural language processing techniques, we have achieved a high level of accuracy and performance, thus fulfilling the objectives of our project.

The evaluation phase has provided us with valuable feedback on the performance of our classifier, allowing us to identify areas for improvement and refine our approach. By iteratively optimizing our model and fine-tuning its parameters, we have been able to enhance its accuracy and robustness, ensuring reliable detection of spam messages in real-world scenarios.

Looking ahead, we recognize the potential for further refinement and enhancement of our classifier. Continued evaluation and testing will be essential to ensure its effectiveness in diverse environments and against evolving spam tactics. Additionally, we envisage opportunities for the integration of advanced techniques and technologies to enhance the classifier's capabilities and adaptability.

In conclusion, the development of the SMS spam classifier represents a significant milestone in our journey as aspiring data scientists and machine learning practitioners. We are proud of our accomplishments and grateful for the support and guidance of our teachers, peers, and mentors throughout this project. As we embark on future endeavors, we carry with us the lessons learned and experiences gained, confident in our ability to tackle new challenges and make meaningful contributions to the field of data science.

# THANK YOU