Requirements:

R1: Data Collection

Collect a dataset of emails, including subject lines and labels indicating whether the email is spam or ham (not spam).

	Unnamed: 0	label	text	label_num
4659	1233	ham	Subject: enron / hpl actuals for july 20 , 200...	0
679	896	ham	Subject: blue dolphin pipe line company contra...	0
1990	2249	ham	Subject: hpl meter # 981046 butane plant neche...	0
4447	4660	spam	Subject: new flat\r\ndmortgage\r\nget a zfree ...	1
1471	1281	ham	Subject: enron / hpl actuals for july 27 , 200...	0
1645	5066	spam	Subject: popularity pills , cheap ! ! ! viagra...	1
4212	911	ham	Subject: may hours survey\r\nif you guys remeb...	0
537	4173	spam	Subject: 52 - quick loan application\r\nhey\r\...	1
2152	2780	ham	Subject: base gas roll for april 01\r\ndue to ...	0
4376	5124	spam	Subject: hundreds of hours of teens porn video...	1

R2: Understand Data

Explore and understand the dataset about the subject line,labels i.e whether email is spam or ham(not spam).Check for missing values, imbalances, and patterns.

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0

R3: Preprocess and Clean Data

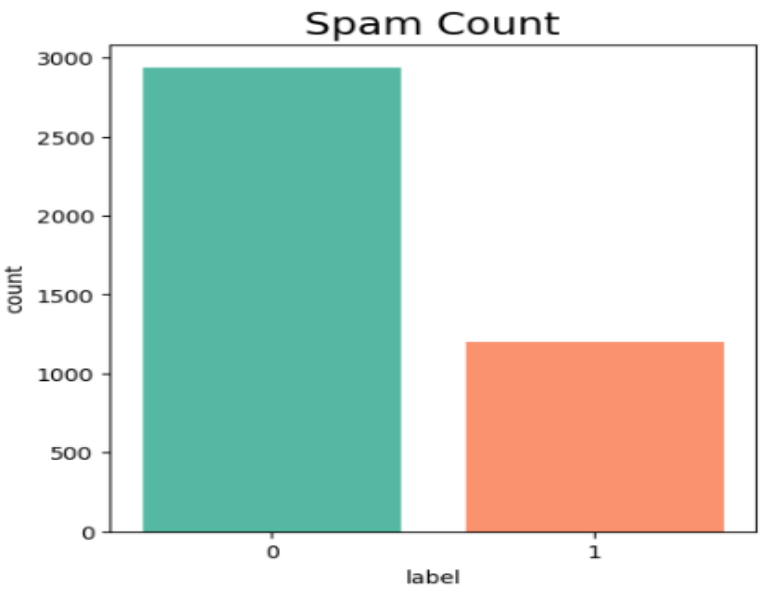
- Remove stop words, special characters, and irrelevant information.
- Convert text to lowercase and handle stemming or lemmatization.
- Optional: Remove duplicate spam emails if any.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

R4: Feature Engineering

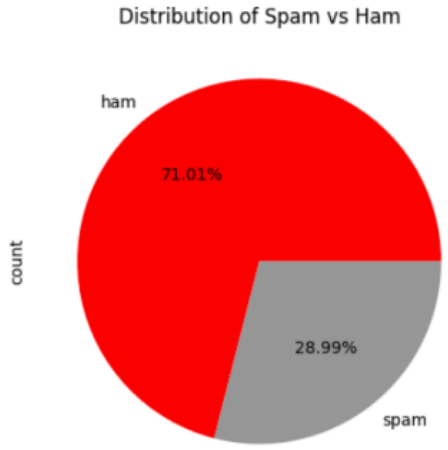
- Transform text data into numerical form using techniques like TfidVectorizer,CountVectorizer

- Considering new features Like Presence of common spam keywords (e.g., "free," "discount," "urgent").
- Data visualization to classify the ratio of spam and ham



label	count
0	3672
1	1499

dtype: int64



R5: Split Data

Divide the dataset into training and testing sets using train-test split (e.g., 80% training, 20% testing).

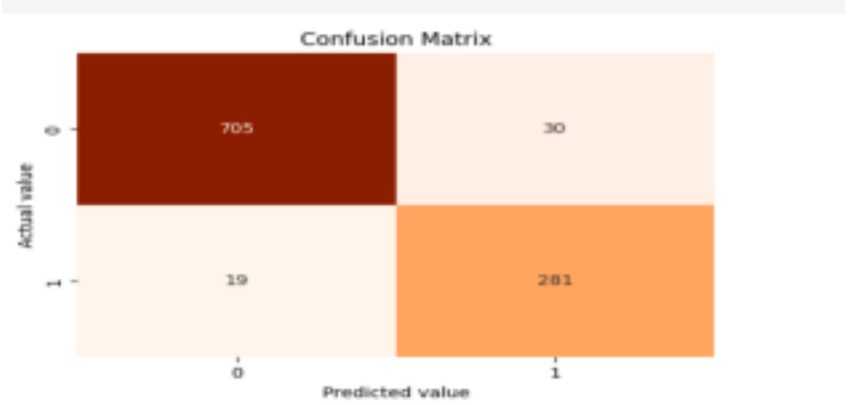
```
Number of rows in the total set: 5171
Number of rows in the training set: 3878
Number of rows in the test set: 1293
```

R6: Choose Model & Requirement and Train Model

Select machine learning models like Logistic Regression,Naive-Bayes, Support Vector Machine to Train the model on the training data.

R7: Test Model & Report Results

Evaluate the model's performance using metrics like accuracy, precision, recall, F1 score, and confusion matrix.



R8: Predict Spam Mails

Use the trained model to predict whether unseen emails are spam or ham.

```
Enter a Subject :Win a FREE iPhone now!
This email is spam.
```