Start coding or generate with AI.

AIM AND GOAL:

The primary aim was to develop an effective email spam detection system using Natural Language Processing (NLP) techniques followed by machine learning model training. The main goal was to optimize the precision metric, prioritizing the accurate identification of spam emails while minimizing false positives

Importing Libraries

```
import numpy as np
import pandas as pd
import nltk
import seaborn as sns
import matplotlib.pyplot as plt
import re
```

LOADING DATASET

```
df=pd.read_csv('/content/spam.csv',encoding='ISO-8859-1')
df
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will Ã¼ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

5572 rows × 2 columns

```
df.head()
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

```
df.tail()
```

	Category	Message
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will Ã¼ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

```
df.dtypes
```

```
Category    object
Message     object
```

```
dtype: object
```

```
df.columns
```

```
Index(['Category', 'Message'], dtype='object')
```

CHECKING MISSING VALUES

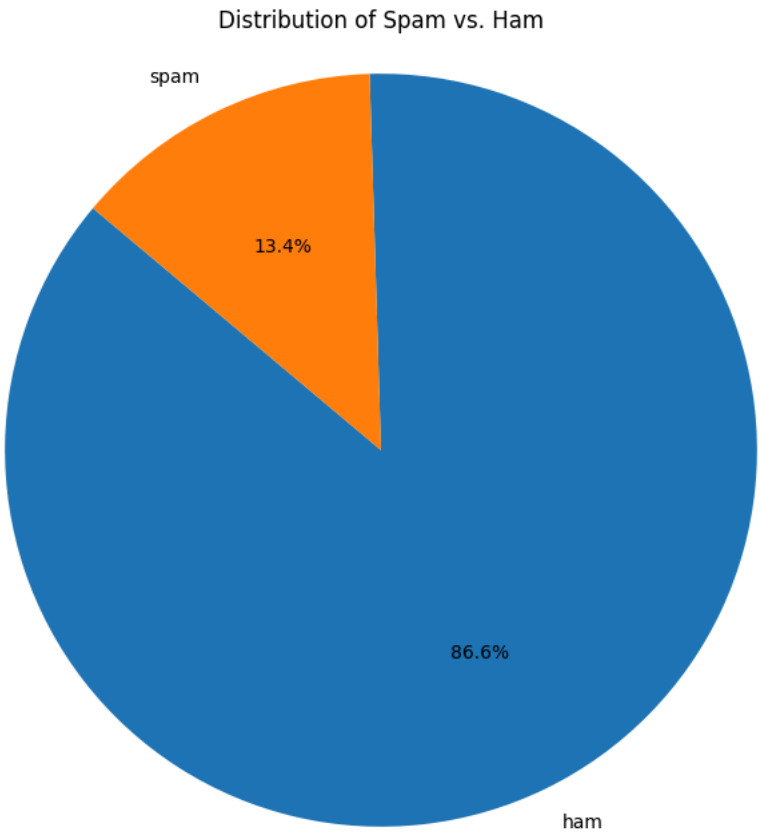
```
df.isna().sum()
```

```
Category    0
Message     0
dtype: int64
```

CALCULATE THE COUNT OF EACH LABEL

```
category_counts = df['Category'].value_counts()

# Plotting the pie chart
plt.figure(figsize=(8, 8))
plt.pie(category_counts, labels=category_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Spam vs. Ham')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```




```
df['Category']=df['Category'].map({'ham':1,'spam':0})
df
```

	Category	Message
0	1	Go until jurong point, crazy.. Available only ...
1	1	Ok lar... Joking wif u oni...
2	0	Free entry in 2 a wkly comp to win FA Cup fina...
3	1	U dun say so early hor... U c already then say...
4	1	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	0	This is the 2nd time we have tried 2 contact u...
5568	1	Will Ã¼ b going to esplanade fr home?
5569	1	Pity, * was in mood for that. So...any other s...
5570	1	The guy did some bitching but I acted like i'd...
5571	1	Rofl. Its true to its name

5572 rows × 2 columns

NLP

```
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('omw-1.4')
tweets=df.Message
tweets
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
0       Go until jurong point, crazy.. Available only ...
1                           Ok lar... Joking wif u oni...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor... U c already then say...
4       Nah I don't think he goes to usf, he lives aro...
                                  ...
5567    This is the 2nd time we have tried 2 contact u...
5568                Will Ã¼ b going to esplanade fr home?
5569    Pity, * was in mood for that. So...any other s...
5570    The guy did some bitching but I acted like i'd...
5571                            Rofl. Its true to its name
Name: Message, Length: 5572, dtype: object
```

TOKENIZATION

```
from nltk import TweetTokenizer
tk=TweetTokenizer()
tweets=tweets.apply(lambda x:tk.tokenize(x)).apply(lambda x:" ".join(x))
tweets
```

```
0       Go until jurong point , crazy .. Available onl...
1                         Ok lar ... Joking wif u oni ...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor ... U c already then sa...
4       Nah I don't think he goes to usf , he lives ar...
                                  ...
5567    This is the 2nd time we have tried 2 contact u...
5568               Will Ã ¼ b going to esplanade fr home ?
5569    Pity , * was in mood for that . So ... any oth...
5570    The guy did some bitching but I acted like i'd...
5571                          Rofl . Its true to its name
Name: Message, Length: 5572, dtype: object
```

```
tweets=tweets.str.replace('[^a-zA-Z0-9]+',' ')
tweets
```

```
<ipython-input-199-243a49c37bfd>:1: FutureWarning: The default value of regex will change from True to False in a future version.
  tweets=tweets.str.replace('[^a-zA-Z0-9]+',' ')
0       Go until jurong point crazy Available only in ...
1                             Ok lar Joking wif u oni
2       Free entry in 2 a wkly comp to win FA Cup fina...
3            U dun say so early hor U c already then say
```

```
4       Nah I don t think he goes to usf he lives arou...
                                  ...
5567    This is the 2nd time we have tried 2 contact u...
5568                   Will b going to esplanade fr home
5569    Pity was in mood for that So any other suggest...
5570    The guy did some bitching but I acted like i d...
5571                          Rofl Its true to its name
Name: Message, Length: 5572, dtype: object
```

```
from nltk.tokenize import word_tokenize
tweets=tweets.apply(lambda x:' '.join([w for w in word_tokenize(x) if len(w)>=3]))
tweets
```

```
0       until jurong point crazy Available only bugis ...
1                                     lar Joking wif oni
2       Free entry wkly comp win Cup final tkts 21st M...
3                    dun say early hor already then say
4         Nah don think goes usf lives around here though
                                  ...
5567    This the 2nd time have tried contact have won ...
5568                           Will going esplanade home
5569         Pity was mood for that any other suggestions
5570    The guy did some bitching but acted like inter...
5571                             Rofl Its true its name
Name: Message, Length: 5572, dtype: object
```

```
from nltk.stem import SnowballStemmer
stemmer=SnowballStemmer('english')
tweets=tweets.apply(lambda x:[stemmer.stem(i.lower())for i in tk.tokenize(x)]).apply(lambda x:' '.join(x))
tweets
```

```
0       until jurong point crazi avail onli bugi great...
1                                       lar joke wif oni
2       free entri wkli comp win cup final tkts 21st m...
3                    dun say earli hor alreadi then say
4           nah don think goe usf live around here though
                                  ...
5567    this the 2nd time have tri contact have won th...
5568                              will go esplanad home
5569             piti was mood for that ani other suggest
5570    the guy did some bitch but act like interest b...
5571                               rofl it true it name
Name: Message, Length: 5572, dtype: object
```

REMOVE STOPWORDS

```
from nltk.corpus import stopwords
nltk.download('stopwords')
sw=stopwords.words('english')
tweets=tweets.apply(lambda x:[i for i in tk.tokenize(x) if i not in sw]).apply(lambda x:' '.join(x))
tweets
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
0       jurong point crazi avail onli bugi great world...
1                                       lar joke wif oni
2       free entri wkli comp win cup final tkts 21st m...
3                          dun say earli hor alreadi say
4                    nah think goe usf live around though
                                  ...
5567    2nd time tri contact 750 pound prize claim eas...
5568                                    go esplanad home
5569                               piti mood ani suggest
5570    guy bitch act like interest buy someth els nex...
5571                                      rofl true name
Name: Message, Length: 5572, dtype: object
```

TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer
vec=TfidfVectorizer()
train_data=vec.fit_transform(tweets)
train_data
```

```
<5572x6885 sparse matrix of type '<class 'numpy.float64'>'
        with 44122 stored elements in Compressed Sparse Row format>
```

```
y=df['Category'].values
y
```

```
array([1, 1, 0, ..., 1, 1, 1])
```

```
x=train_data
x
```

```
<5572x6885 sparse matrix of type '<class 'numpy.float64'>'
        with 44122 stored elements in Compressed Sparse Row format>
```

```
print(train_data.shape)
print(y.shape)
```

```
(5572, 6885)
(5572,)
```

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(train_data,y,test_size=0.30,random_state=42)
```

MODEL CREATION

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix,accuracy_score
from sklearn.metrics import classification_report
k_model=KNeighborsClassifier(n_neighbors=7)
n_model=GaussianNB()
s_model=SVC()
r_model=RandomForestClassifier()
d_model=DecisionTreeClassifier(criterion='entropy')
lst_model=[k_model,n_model,r_model,s_model,d_model]
```

```
x_train = x_train.toarray()
x_test = x_test.toarray()
```

```
for i in lst_model:
  print('model is',i)
  i.fit(x_train,y_train)
  y_pred=i.predict(x_test)
  print("*"*100)
  print(confusion_matrix(y_test,y_pred))
  print("Accuracy score is",accuracy_score(y_test,y_pred))
  print(".......classification Report........")
  print(classification_report(y_test,y_pred))
```

```
model is KNeighborsClassifier(n_neighbors=7)
****************************************************************************************************
[[  43  181]
 [   0 1448]]
Accuracy score is 0.8917464114832536
.......classification Report........
              precision    recall  f1-score   support

           0       1.00      0.19      0.32       224
           1       0.89      1.00      0.94      1448

    accuracy                           0.89      1672
   macro avg       0.94      0.60      0.63      1672
weighted avg       0.90      0.89      0.86      1672

model is GaussianNB()
****************************************************************************************************
[[ 196   28]
 [ 210 1238]]
Accuracy score is 0.8576555023923444
.......classification Report........
              precision    recall  f1-score   support

           0       0.48      0.88      0.62       224
           1       0.98      0.85      0.91      1448

    accuracy                           0.86      1672
   macro avg       0.73      0.86      0.77      1672
weighted avg       0.91      0.86      0.87      1672

model is RandomForestClassifier()
****************************************************************************************************
[[ 189   35]
 [   0 1448]]
Accuracy score is 0.979066985645933
.......classification Report........
              precision    recall  f1-score   support
```

```
           0       1.00      0.84      0.92       224
           1       0.98      1.00      0.99      1448

    accuracy                           0.98      1672
   macro avg       0.99      0.92      0.95      1672
weighted avg       0.98      0.98      0.98      1672

model is SVC()
*****************************************************************************************************
[[ 181   43]
 [   1 1447]]
Accuracy score is 0.9736842105263158
.......classification Report........
              precision    recall  f1-score   support

           0       0.99      0.81      0.89       224
           1       0.97      1.00      0.99      1448

    accuracy                           0.97      1672
   macro avg       0.98      0.90      0.94      1672
```