

```
import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import pandas as pd

import numpy as np

import nltk

from nltk.corpus import stopwords

import matplotlib.pyplot as plt

import warnings

warnings.filterwarnings('ignore')

df=pd.read_csv('/content/email.csv')

df.tail()

df = df.iloc[:-1, :]

df.tail()

df.isnull().sum()

df.duplicated().sum()

df.drop_duplicates(inplace=True)

df.duplicated().sum()

df.dropna(inplace=True)

df.shape

X=df['Message']

X.head()

y=df['Category']

y.head()

df['Category'].replace({'spam': 1, 'ham': 0}, inplace=True)

y.head()

from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer

import seaborn as sns
```

```
%matplotlib inline

cv=CountVectorizer()

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=34)

X_train.shape

X_train.head()

X_train_t=cv.fit_transform(X_train)

X_test_t=cv.transform(X_test)

print(X_train_t)

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

from sklearn.metrics import confusion_matrix

model=LogisticRegression(penalty='l1', solver='liblinear')

unique_values = y_train.value_counts()

print(unique_values)

model.fit(X_train_t,y_train)

y_pred=model.predict(X_test_t)

acc=accuracy_score(y_test,y_pred)

print(confusion_matrix(y_test,y_pred))

print("Accuracy of the model is:",acc)

print("precision_score is:",precision_score(y_test,y_pred))

print("recall_score is:",recall_score(y_test,y_pred))

print("f1_score is:",f1_score(y_test,y_pred))

print()
```