```
import pandas as pd
import numpy as np
p=pd.read_csv("https://github.com/YBI-Foundation/Dataset/raw/main/Spam%20Email.csv")
```

```
p.head()
```

	ID	Mail	Text	Label
0	1	ham	Subject: christmas tree farm pictures\r\n	0
1	2	ham	Subject: vastar resources , inc .\r\ngary , pr...	0
2	3	ham	Subject: calpine daily gas nomination\r\n- cal...	0
3	4	ham	Subject: re : issue\r\nfyi - see note below - ...	0
4	5	ham	Subject: meter 7268 nov allocation\r\nfyi .\r\...	0

```
p.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5171 entries, 0 to 5170
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   ID      5171 non-null   int64 
 1   Mail    5171 non-null   object
 2   Text    5171 non-null   object
 3   Label   5171 non-null   int64 
dtypes: int64(2), object(2)
memory usage: 161.7+ KB
```

```
p.columns
```

```
Index(['ID', 'Mail', 'Text', 'Label'], dtype='object')
```

```
p.shape
```

```
(5171, 4)
```

```
y=p['Label']
```

```
y.shape
```

```
(5171,)
```

```
y
```

```
0       0
1       0
2       0
3       0
4       0
       ..
5166    1
5167    1
5168    1
5169    1
5170    1
Name: Label, Length: 5171, dtype: int64
```

```
x=p['Text']
```

```
x.shape
```

```
(5171,)
```

```
x
```

```
0               Subject: christmas tree farm pictures\r\n
1       Subject: vastar resources , inc .\r\ngary , pr...
2       Subject: calpine daily gas nomination\r\n- cal...
3       Subject: re : issue\r\nfyi - see note below - ...
4       Subject: meter 7268 nov allocation\r\nfyi .\r\...
                              ...
5166    Subject: our pro - forma invoice attached\r\nd...
5167    Subject: str _ rndlen ( 2 - 4 ) } { extra _ ti...
5168    Subject: check me out !\r\n61 bb\r\nhey derm\r...
5169    Subject: hot jobs\r\nglobal marketing specialt...
5170    Subject: save up to 89 % on ink + no shipping ...
Name: Text, Length: 5171, dtype: object
```

```
from sklearn.model_selection import train_test_split
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5,stratify=y,random_st
```

```
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

```
((2585,), (2586,), (2585,), (2586,))
```

```
!pip install sklearn
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-whe
Requirement already satisfied: sklearn in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: numpy>=1.14.6 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/
```

```
Requirement already satisfied: scipy>=1.1.0 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-pac
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
t=TfidfVectorizer(min_df=1,stop_words='english',lowercase='True')
```

```
x_train_features=t.fit_transform(x_train)
x_test_features=t.transform(x_test)
```

```
x_train
```

```
3511    Subject: expense report receipts not received\...
855     Subject: wc 551 revision and notice of force m...
1417    Subject: hpl meter # 985355 brown common point...
2411    Subject: ranks communication\r\nafter getting ...
1717    Subject: cornhusker contact information - revi...
                              ...
3811    Subject: fwd : everything here . + xanax + _ v...
2728    Subject: exxon company , usa global # 96035668...
2806    Subject: fw : first delivery - rodessa operati...
2691    Subject: latest frontera doc\r\n- - - - - - - ...
331     Subject: hl & p month to date flow\r\njanet . ...
Name: Text, Length: 2585, dtype: object
```

```
x_train_features
```

```
<2585x33820 sparse matrix of type '<class 'numpy.float64'>'
        with 170047 stored elements in Compressed Sparse Row format>
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
r=RandomForestClassifier(random_state=2528)
```

```
r.fit(x_train_features,y_train)
```

```
RandomForestClassifier(random_state=2528)
```

Double-click (or enter) to edit

```
y_pred=r.predict(x_test_features)
```

```
y_pred.shape
```

```
(2586,)
```

```
y_pred
```

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

Double-click (or enter) to edit

```
r.predict_proba(x_test_features)
```

```
array([[0.15, 0.85],
       [0.96, 0.04],
       [0.95, 0.05],
       ...,
       [1.  , 0.  ],
       [0.99, 0.01],
       [1.  , 0.  ]])
```

```
from sklearn.metrics import confusion_matrix,classification_report
```

Double-click (or enter) to edit

```
print(confusion_matrix(y_test,y_pred))
```

```
[[1774   62]
 [  17  733]]
```

```
print(classification_report(y_test,y_pred))
```

```
                  precision    recall  f1-score   support

               0       0.99      0.97      0.98      1836
               1       0.92      0.98      0.95       750

        accuracy                           0.97      2586
       macro avg       0.96      0.97      0.96      2586
    weighted avg       0.97      0.97      0.97      2586
```

Double-click (or enter) to edit