```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score


# loading the data from csv file to pandas Dataframe
raw_mail_data = pd.read_csv('/content/mail_data.csv', encoding=encoding')
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<ipython-input-5-7df20e8b6112> in <cell line: 2>()
      1 # loading the data from csv file to pandas Dataframe
----> 2 raw_mail_data = pd.read_csv('/content/mail_data.csv, encoding=encoding')

                                 4 frames
/usr/local/lib/python3.10/dist-packages/pandas/io/common.py in get_handle(path_or_buf, mode, encoding, compression, memory_map,
is_text, errors, storage_options)
    857         if ioargs.encoding and "b" not in ioargs.mode:
    858             # Encoding
--> 859             handle = open(
    860                 handle,
    861                 ioargs.mode,

FileNotFoundError: [Errno 2] No such file or directory: '/content/mail_data.csv, encoding=encoding'
```

```
print(raw_mail_data)
```

```
     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 ü 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 x 2 columns]
```

```
# replace the null values with a null string
mail_data = raw_mail_data.where((pd.notnull(raw_mail_data)),'')


# printing five rows of the data frame
mail_data.head()
```

	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...

```
# checkng tmber of columns and rows in dataframe
mail_data.shape
```

```
(5572, 2)
```

Label Encoding

```
# label the spam mail as 0; ham mail as 1;


mail_data.loc[mail_data['Category'] == 'spam','Category',] = 0
mail_data.loc[mail_data['Category'] == 'ham','Category',] = 1
```

spam-0 ham-1

```
# seperate data as text and label
X = mail_data['Message']
Y = mail_data['Category']
```

```
print(X)
```

```
     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 ü 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
```

```
print(Y)
```

```
     0       1
     1       1
     2       0
     3       1
     4       1
            ..
     5567    0
     5568    1
     5569    1
     5570    1
     5571    1
     Name: Category, Length: 5572, dtype: object
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)
```

```
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

```
     (5572,)
     (4457,)
     (1115,)
```

Feature Extraction

```
#  transform the data to feature vectors that can be used us input to the logistic regression

feature_extraction = TfidfVectorizer(min_df = 1, stop_words='english', lowercase=True)

X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)

# convert Y_train and Y_test values as integers

Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

```
print(X_train)
```

```
print(X_train_features)
```

Training the Model

Logistic Regression

```
model = LogisticRegression()
```

```
# training the Logistic Regression model with the training data
model.fit(X_train_features, Y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

Evaluating the Model

```
# prediction on training data

prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)


print('Accuracy on training data : ', accuracy_on_training_data)
```

```
Accuracy on training data :  0.9670181736594121
```

```
# prediction on test data

prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)


print('Accuracy on test data : ', accuracy_on_test_data)
```

```
Accuracy on test data :  0.9659192825112107
```

Building a Predictive system

```
input_mail = ["Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rat

# convert text to feature vectors
input_data_features = feature_extraction.transform(input_mail)

# making prediction

prediction = model.predict(input_data_features)
print(prediction)

if prediction[0]==1:
  print('Ham mail')

else:
  print('spam mail')
```

```
[0]
spam mail
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```