Email Spam Detection

In [124...
```
#Importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score
```

In [125...
```
#importing the dataset
a=pd.read_csv("Spam.csv",encoding='ISO-8859-1')
#first five rows & columns
a.head()
```

Out[125...

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

In [126...
```
#Last 5 rows & columns
a.tail()
```

Out[126...

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will ì_ b going to esplanade fr home?	NaN	NaN	NaN
5569	ham	Pity, * was in mood for that. So...any other s...	NaN	NaN	NaN
5570	ham	The guy did some bitching but I acted like i'd...	NaN	NaN	NaN
5571	ham	Rofl. Its true to its name	NaN	NaN	NaN

In [127...
```
#Dimensions of Dataset
a.ndim
```

Out[127...
```
2
```

In [128...
```
a.size
```

Out[128...
```
27860
```

In [129...
```
a.shape
```

Out[129...
```
(5572, 5)
```

In [130...
```
a.describe()
```

Out[130...

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
count	5572	5572	50	12	6
unique	2	5169	43	10	5
top	ham	Sorry, I'll call later	bt not his girlfrnd...	GE	GNT:-)"
freq	4825	30	3	2	2

In [131...
```
a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

Data Cleaning and Pre-processing

In [132...
```
a.isnull()
```

Out[132...

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	False	False	True	True	True
1	False	False	True	True	True
2	False	False	True	True	True
3	False	False	True	True	True
4	False	False	True	True	True
...	...	...	...	...	...
5567	False	False	True	True	True
5568	False	False	True	True	True
5569	False	False	True	True	True
5570	False	False	True	True	True
5571	False	False	True	True	True

5572 rows × 5 columns

In [133...
```
a.isna().sum()
```

Out[133...
```
v1                0
v2                0
Unnamed: 2     5522
Unnamed: 3     5560
Unnamed: 4     5566
dtype: int64
```

In [134...
```
a.drop(columns=a[['Unnamed: 2','Unnamed: 3','Unnamed: 4']],axis=1,inplace=True)
print(a.head(5))
```

```
     v1                                                 v2
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...
```

In [135...
```
a.columns=['spam/ham','sms']
a
```

Out[135...

	spam/ham	sms
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 ì_ 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

In [136...
```
a.shape
```

Out[136...
```
(5572, 2)
```

In [137...
```
a.dtypes
```

Out[137...
```
spam/ham    object
sms         object
dtype: object
```

In [138...
```
a.nunique()
```

Out[138...
```
spam/ham       2
sms         5169
dtype: int64
```

In [139...
```
a.max()
```

Out[139...
```
spam/ham                                                 spam
sms         ÌÏ wait 4 me in sch i finish ard 5..
dtype: object
```

In [140...
```
a.min()
```

Out[140...
```
spam/ham                                                 ham
sms         &lt;#&gt;  in mca. But not conform.
dtype: object
```

Data Formatting

In [141...
```
a['spam/ham'].value_counts()
```

Out[141...
```
ham     4825
spam     747
Name: spam/ham, dtype: int64
```

In [142...
```
a['spam/ham']=a['spam/ham'].map({'spam': 0,'ham':1})
a
```

Out[142...

	spam/ham	sms
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 ì_ 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

In [143...
```
x=a['sms']
x
```

Out[143...
```
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 ì_ 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: sms, Length: 5572, dtype: object
```

In [144...
```
y=a['spam/ham']
y
```

Out[144...
```
0       1
1       1
2       0
3       1
4       1
       ..
5567    0
5568    1
5569    1
5570    1
5571    1
Name: spam/ham, Length: 5572, dtype: int64
```

In [145...
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=3)
```

In [146...
```
print("Training Set=(",x_train.shape,y_train.shape,")")
print("Testing Set=(",x_test.shape,y_test.shape,")")
```

```
Training Set=( (4457,) (4457,) )
Testing Set=( (1115,) (1115,) )
```

In [147...
```
feature=TfidfVectorizer(min_df=1,stop_words="english",lowercase=True)
feature
```

Out[147...
```
TfidfVectorizer(stop_words='english')
```

In [148...
```
y_train=y_train.astype("int")
y_test=y_test.astype("int")
```

In [149...
```
xtrain=feature.fit_transform(x_train)
xtest=feature.transform(x_test)
xtrain,xtest
```

Out[149...
```
(<4457x7510 sparse matrix of type '<class 'numpy.float64'>'
         with 34758 stored elements in Compressed Sparse Row format>,
 <1115x7510 sparse matrix of type '<class 'numpy.float64'>'
         with 7766 stored elements in Compressed Sparse Row format>)
```

Logistic Regression

In [150...
```
model=LogisticRegression()
model.fit(xtrain,y_train)
```

Out[150...
```
LogisticRegression()
```

In [151...
```
model.score(xtrain,y_train)
```

Out[151...
```
0.9661207889970832
```

In [152...
```
model.score(xtest,y_test)
```

Out[152...
```
0.9623318385650225
```

In [153...
```
#making predictions
y_pred=model.predict(xtest)
y_pred
```

Out[153...
```
array([1, 1, 1, ..., 1, 1, 1])
```

Predicting from random values

In [154...
```
b=["Is that seriously how you spell his name?"]
b=feature.transform(b)
predict=model.predict(b)
print(predict)
```

```
[1]
```

In [156...
```
#Evolution of model
print("Mean Squared Error : %.2f" % np.mean(y_pred-y_test)**2)
```

```
Mean Squared Error : 0.00
```

In [ ]:
```

```