EMAIL SPAM DETECTION WITH MACHINE LEARNING:

In [1]:
```python
#importing basic libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer #It convert text document into numeric representation
from sklearn.linear_model import LogisticRegression   # Logistic Regression Algorithm
from sklearn.metrics import accuracy_score
```

In [2]:
```python
data_set=pd.read_csv(r"C:\Users\HP\OneDrive\Documents\oasis infobytes\spam.csv",encoding='ISO-8859-1')
```

In [3]:
```python
data_set.head()
```

Out[3]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 16	Unnamed: 17	Unnamed: 18	Unnamed: 19	Unnamed: 20	Unnamed: 21	Unnamed: 22	Unnamed: 23	Unnamed: 24	Unnamed: 25
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 26 columns

In [ ]:

In [ ]:

Data cleaning:

In [4]:
```python
dataset=data_set.drop(['Unnamed: 2','Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5',
       'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9', 'Unnamed: 10',
       'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13', 'Unnamed: 14',
       'Unnamed: 15', 'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18',
       'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22',
       'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25'],axis=1)
```

In [5]:
```python
dataset.head()
```

Out[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 [6]:
```python
dataset.tail()
```

Out[6]:

	v1	v2
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

In [7]:
```python
dataset.isnull().sum()
```

Out[7]:
```
v1    0
v2    0
dtype: int64
```

In [8]:
```python
dataset.duplicated().sum()
```

Out[8]:
```
403
```

In [9]:
```python
dataset=dataset.drop_duplicates(keep="first") #Removing the duplicate
```

In [10]:
```python
dataset.duplicated().sum()   #rechecking Duplicate values
```

Out[10]:
```
0
```

In [11]:
```python
dataset.shape
```

Out[11]:
```
(5169, 2)
```

In [12]:
```python
#rename the columns
dataset.rename(columns={'v1':'Category','v2':'Message'},inplace=True)
dataset.head()
```

Out[12]:

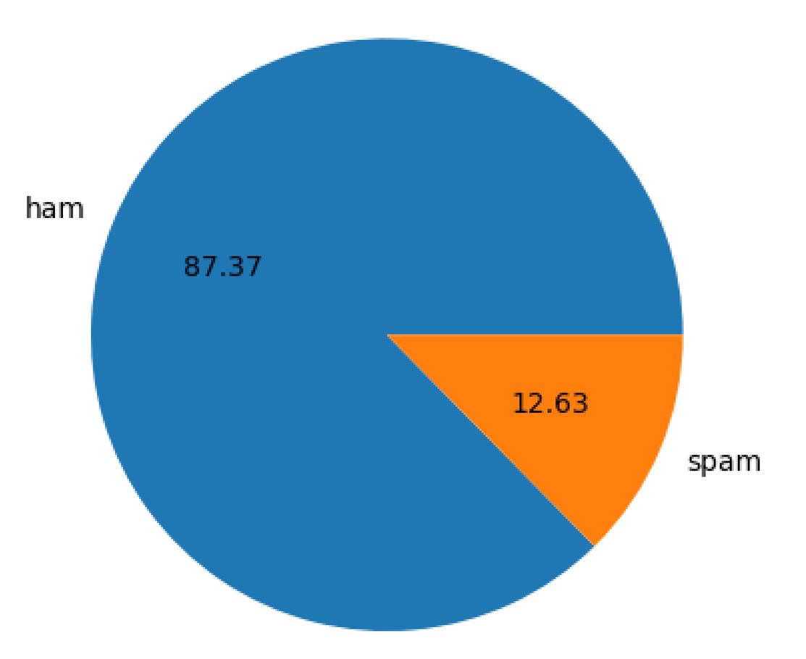
	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

In [ ]:

In [ ]:

Data Visualization

In [13]:
```python
plt.pie(dataset['Category'].value_counts(),labels=['ham','spam'],autopct="%0.2f")
plt.show()
```



In [ ]:

In [ ]:

Machine learning model

In [14]:
```python
dataset.loc[dataset['Category']=='spam','Category',]=0    # spam = 0 and Ham = 1
dataset.loc[dataset['Category']=='ham','Category',]=1
```

In [15]:
```python
X=dataset['Message']
Y=dataset['Category']
```

In [16]:
```python
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 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: Message, Length: 5169, dtype: object
```

In [17]:
```python
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: 5169, dtype: object
```

In [18]:
```python
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=3)
```

In [19]:
```python
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

```
(5169,)
(4135,)
(1034,)
```

In [20]:
```python
print(Y.shape)
print(Y_train.shape)
print(Y_test.shape)
```

```
(5169,)
(4135,)
(1034,)
```

In [21]:
```python
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)
```

In [22]:
```python
Y_train=Y_train.astype(int)
Y_test=Y_test.astype(int)
```

In [23]:
```python
Y_train
```

Out[23]:
```
4443    1
982     0
3822    1
3924    1
4927    1
       ..
806     1
990     1
1723    1
3519    1
1745    1
Name: Category, Length: 4135, dtype: int32
```

In [24]:
```python
Y_test
```

Out[24]:
```
4994    1
4292    1
4128    1
4429    1
660     1
       ..
4003    1
1107    1
5413    1
1413    0
4998    1
Name: Category, Length: 1034, dtype: int32
```

In [25]:
```python
X_train
```

Out[25]:
```
4443                        COME BACK TO TAMPA FFFFUUUUUUU
982     Congrats! 2 mobile 3G Videophones R yours. cal...
3822    Please protect yourself from e-threats. SIB ne...
3924       As if i wasn't having enough trouble sleeping.
4927    Just hopeing that wasnÕ÷t too pissed up to re...
                              ...
806      sure, but make sure he knows we ain't smokin yet
990                                      26th OF JULY
1723    Hi Jon, Pete here, Ive bin 2 Spain recently & ...
3519    No it will reach by 9 only. She telling she wi...
1745    IåÕm cool ta luv but v.tired 2 cause i have be...
Name: Message, Length: 4135, dtype: object
```

In [26]:
```python
X_test
```

Out[26]:
```
4994    Just looked it up and addie goes back Monday, ...
4292    You best watch what you say cause I get drunk ...
4128                 Me i'm not workin. Once i get job...
4429          Yar lor... How u noe? U used dat route too?
660     Under the sea, there lays a rock. In the rock,...
                              ...
4003    Well there's a pattern emerging of my friends ...
1107    From someone not to smoke when every time I've...
5413    Nite nite pocay wocay luv u more than n e thin...
1413    Dear U've been invited to XCHAT. This is our f...
4998                        Hmph. Go head, big baller.
Name: Message, Length: 1034, dtype: object
```

In [27]:
```python
print(X_test_features)
```

```
  (0, 6284)     0.43430701953285156
  (0, 4357)     0.4264504812056483
  (0, 3999)     0.4541039150126108
  (0, 3685)     0.21875536593912145
  (0, 3008)     0.3755569393427584
  (0, 796)      0.48415917776958733
  (1, 7050)     0.41978523395044104
  (1, 5656)     0.35499712111138654
  (1, 2369)     0.5617364019285492
  (1, 1608)     0.47304204171951914
  (1, 1254)     0.398046282326562
  (2, 7221)     0.7923997102028898
  (2, 3640)     0.6100022125126892
  (3, 7292)     0.4854329061592562
  (3, 6879)     0.504266116821645
  (3, 4574)     0.45819212318042857
  (3, 4009)     0.3321387620897908
  (3, 2080)     0.43567694225913534
  (4, 7218)     0.2292199492753507
  (4, 5689)     0.23694703770184997
  (4, 5540)     0.5218991333605264
  (4, 4823)     0.5112231722851113
  (4, 2508)     0.597985350067042
  (5, 3025)     0.5652583430208418
  (5, 1132)     0.8249139383264974
  :     :
  (1031, 5191)  0.3701308998039874
  (1031, 4565)  0.6439113578053595
  (1031, 4068)  0.3041309124908908
  (1031, 510)   0.42409124971092943
  (1032, 7339)  0.2632164916232764
  (1032, 7271)  0.269584829223756
  (1032, 6987)  0.25783213890193124
  (1032, 6919)  0.17642080640756436
  (1032, 6764)  0.16357649120381332
  (1032, 6298)  0.24905393871920117
  (1032, 5563)  0.2362096235154502
  (1032, 4406)  0.2773790445273517
  (1032, 3848)  0.25783213890193124
  (1032, 3538)  0.24905393871920117
  (1032, 2720)  0.23900544751939748
  (1032, 2099)  0.17318001901981067
  (1032, 1899)  0.18778753530769227
  (1032, 1660)  0.19874332825316712
  (1032, 1071)  0.21945854189397707
  (1032, 674)   0.24537378511987706
  (1032, 394)   0.24905393871920117
  (1032, 316)   0.2016158353905777
  (1032, 302)   0.1936225930341707
  (1033, 3198)  0.7337664084922214
  (1033, 1272)  0.6794018382139002
```

In [28]:
```python
model = LogisticRegression()
model.fit(X_train_features, Y_train)
```

Out[28]:
```
▾ LogisticRegression
LogisticRegression()
```

In [29]:
```python
prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train,prediction_on_training_data)
```

In [30]:
```python
print("Accuracy on training data: ",accuracy_on_training_data)
```

```
Accuracy on training data:  0.962273276904474
```

In [31]:
```python
prediction_on_test_data=model.predict(X_test_features)
accuracy_on_test_data=accuracy_score(Y_test,prediction_on_test_data)
```

In [32]:
```python
print("Accuracy on test data: ",accuracy_on_test_data)
```

```
Accuracy on test data:  0.960348162475822
```

In [ ]:

In [ ]:

Final result or prediction:

In [40]:
```python
input_mail=[" #spam 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 rate)T&C's apply 08452810075over18's"]
input_data_features = feature_extraction.transform(input_mail)
prediction=model.predict(input_data_features)
print("Our Mail is: ",prediction)
if(prediction[0]==1):
    print("Ham mail")
else:
    print("Spam mail")


# some sample messages:
    #spam 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 rate)T&C's apply 08452810075over18's
    #Even my brother is not like to speak with me. They treat me like aids patent.
    #WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
    #I HAVE A DATE ON SUNDAY WITH WILL!!
    #Nah I don't think he goes to usf, he lives around here though
    #FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
```

```
Our Mail is:  [0]
Spam mail
```

In [ ]:

In [ ]:
```python
                                      #Thanking You...
```

-Thiruvalluvan G