In [2]:
```
import pandas as pd
```

In [3]:
```
df=pd.read_csv("https://raw.githubusercontent.com/sunnysavita10/Naive-Bayes/main/
```

In [4]:
```
df.head()
```

Out[4]:

	label	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 [5]:
```
df['message'][0]
```

Out[5]:
```
'Go until jurong point, crazy.. Available only in bugis n great world la e buff
et... Cine there got amore wat...'
```

In [6]:
```
df['message'][10]
```

Out[6]:
```
"I'm gonna be home soon and i don't want to talk about this stuff anymore tonig
ht, k? I've cried enough today."
```

In [7]:
```
df['message'][40]
```

Out[7]:
```
'Pls go ahead with watts. I just wanted to be sure. Do have a great weekend. Ab
iola'
```

In [ ]:
```

```

In [8]:
```
import nltk
```

In [9]:
```
import re
# regular expression
```

In [10]:
```

nltk.download("stopwords")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Mukul\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[10]:
```
True
```

In [11]:
```
from nltk.corpus import stopwords
```

In [12]:
```
from nltk.stem.porter import PorterStemmer
```

In [13]:
```
ps=PorterStemmer()
```

In [14]:
```
stopwords.words('english')
```

Out[14]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
```

In [34]:
```
corpus=[]
```

In [35]:
```
# one by one step for explaning
rev=re.sub("[^a-zA-Z]",' ', df['message'][0])
rev
```

Out[35]:
```
'Go until jurong point  crazy   Available only in bugis n great world la e buff
et    Cine there got amore wat   '
```

In [36]:
```
rev.lower()
```

Out[36]:
```
'go until jurong point  crazy   available only in bugis n great world la e buff
et    cine there got amore wat   '
```

In [37]:
```
rev=rev.split()
```

In [42]:
```
rev
```

Out[42]:
```
['Go',
 'until',
 'jurong',
 'point',
 'crazy',
 'Available',
 'only',
 'in',
 'bugis',
 'n',
 'great',
 'world',
 'la',
 'e',
 'buffet',
 'Cine',
 'there',
 'got',
 'amore',
 'wat']
```

In [43]:
```
[ps.stem(word) for word in rev if not word in stopwords.words('english')]
```

Out[43]:
```
['go',
 'jurong',
 'point',
 'crazi',
 'avail',
 'bugi',
 'n',
 'great',
 'world',
 'la',
 'e',
 'buffet',
 'cine',
 'got',
 'amor',
 'wat']
```

In [48]:
```
''.join(rev)  P
```

Out[48]:
```
'Go until jurong point crazy Available only in bugis n great world la e buffet
Cine there got amore wat'
```

In [ ]:

In [49]:
```
pip install Corpus
```

```
Requirement already satisfied: Corpus in c:\users\mukul\anaconda3\lib\site-pack
ages (0.4.2)
Note: you may need to restart the kernel to use updated packages.
```

In [50]:
```
# add all the step
for i in range(0,len(df)):
    review=re.sub("[^a-zA-Z]",' ', df['message'][i])
    review=review.lower()
    review=review.split()

    review=[ps.stem(word) for word in review if not word in stopwords.words('engl
    review=' '.join(review)
    corpus.append(review)
```

In [52]:
```
corpus
```

Out[52]:
```
['go jurong point crazi avail bugi n great world la e buffet cine got amor wa
t',
 'ok lar joke wif u oni',
 'free entri wkli comp win fa cup final tkt st may text fa receiv entri quest
ion std txt rate c appli',
 'u dun say earli hor u c alreadi say',
 'nah think goe usf live around though',
 'freemsg hey darl week word back like fun still tb ok xxx std chg send rcv',
 'even brother like speak treat like aid patent',
 'per request mell mell oru minnaminungint nurungu vettam set callertun calle
r press copi friend callertun',
 'winner valu network custom select receivea prize reward claim call claim co
de kl valid hour',
 'mobil month u r entitl updat latest colour mobil camera free call mobil upd
at co free',
 'gonna home soon want talk stuff anymor tonight k cri enough today',
 'six chanc win cash pound txt csh send cost p day day tsandc appli repli hl
info',
 'urgent week free membership prize jackpot txt word claim c www dbuk net lcc
```

In [53]:
```
corpus[1]
```

Out[53]:
```
'ok lar joke wif u oni'
```

In [54]:
```
corpus[2]
```

Out[54]:
```
'free entri wkli comp win fa cup final tkt st may text fa receiv entri question
std txt rate c appli'
```

In [55]:
```
corpus[6]
```

Out[55]:
```
'even brother like speak treat like aid patent'
```

Convert the data in vector

In [57]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [58]:
```
cv=CountVectorizer()
```

In [59]:
```
X=cv.fit_transform(corpus).toarray()
```

In [60]:
```
#1 set of unique words
#2 finally it is creating a vectors
X.shape
```

Out[60]:
```
(5572, 6296)
```

In [71]:
```
X[0]
```

Out[71]:
```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [72]:
```
X[1]
```

Out[72]:
```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [61]:
```
df['label']   # target feature
```

Out[61]:
```
0        ham
1        ham
2       spam
3        ham
4        ham
        ...
5567    spam
5568     ham
5569     ham
5570     ham
5571     ham
Name: label, Length: 5572, dtype: object
```

In [62]:
```
y=pd.get_dummies(df['label'],drop_first=True)
```

In [63]:
```
X
```

Out[63]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [65]:
```
y
```

Out[65]:

	spam
0	0
1	0
2	1
3	0
4	0
...	...
5567	1
5568	0
5569	0
5570	0
5571	0

5572 rows × 1 columns

Train_test_split

In [67]:
```
from sklearn.model_selection import train_test_split
```

In [68]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25 , random_state=
```

1. Gaussian Naive Bayes

In [69]:
```
from sklearn.naive_bayes import GaussianNB
```

In [70]:
```
model=GaussianNB()
```

In [71]:
```
model.fit(X_train,y_train)
```

```
C:\Users\Mukul\anaconda3\lib\site-packages\sklearn\utils\validation.py:1111: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

Out[71]:
```
GaussianNB()
```

In [72]:
```
y_pred=model.predict(X_test)
```

In [74]:
```
y_pred
```

Out[74]:
```
array([0, 0, 0, ..., 0, 0, 1], dtype=uint8)
```

In [73]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test , y_pred)
```

Out[73]:
```
0.8729361091170137
```

2. Multinomial Navie Bayes

In [75]:
```
from sklearn.naive_bayes import MultinomialNB
```

In [77]:
```
model2=MultinomialNB()
```

In [79]:
```
model2.fit(X_train , y_train)
```

```
C:\Users\Mukul\anaconda3\lib\site-packages\sklearn\utils\validation.py:1111: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

Out[79]:
```
MultinomialNB()
```

In [80]:
```
y_predict2=model2.predict(X_test)
```

In [81]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_predict2)
```

Out[81]:
```
0.9691313711414213
```

In [ ]: