$TASK-6$

- EMAIL SPAM FILTERING

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score,f1_score
from sklearn.naive_bayes import MultinomialNB
```

In [9]:
```
import pandas as pd
file_path="C:\\Users\\pedur\\OneDrive\\spam.csv"
file=open(file_path,encoding='Latin1')
spam_df=pd.read_csv(file)
spam_df['v1'].value_counts()
```

Out[9]:
```
v1
ham     4825
spam     747
Name: count, dtype: int64
```

In [10]:
```
file=open(file_path,encoding='utf-8',errors='replace')
spam_df=pd.read_csv(file)
spam_df
```

Out[10]:

	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
...	...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will □ 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

5572 rows × 5 columns

In [12]:
```
spam_df=pd.read_csv(file_path,encoding='latin1')
spam_df
```

Out[12]:

	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
...	...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will_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

5572 rows × 5 columns

In [10]:
```
spam_df.isnull().sum()
```

Out[10]:
```
v1              0
v2              0
Unnamed: 2   5522
Unnamed: 3   5560
Unnamed: 4   5566
dtype: int64
```

In [11]:
```
spam_df.dtypes
```

Out[11]:
```
v1            object
v2            object
Unnamed: 2    object
Unnamed: 3    object
Unnamed: 4    object
dtype: object
```

In [13]:
```
columns_to_drop=["Unnamed: 2","Unnamed: 3","Unnamed: 4"]
spam_df.drop(columns=columns_to_drop,inplace=True)
```

In [14]:
```
spam_df
```

Out[14]:

	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...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will □ 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 [15]:
```
spam_df.columns
```

Out[15]:
```
Index(['v1', 'v2'], dtype='object')
```

In [16]:
```
new_columns_names={'v1':"category",'v2':"Message-in-email"}
spam_df.rename(columns=new_columns_names,inplace=True)
```

In [17]:
```
spam_df
```

Out[17]:

	category	Message-in-email
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 □ 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 [18]:
```
spam_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   category         5572 non-null   object
 1   Message-in-email  5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

In [19]:
```
spam_df.head()
```

Out[19]:

	category	Message-in-email
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 [20]:
```
spam_df.tail()
```

Out[20]:

	category	Message-in-email
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will □ 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 [23]:
```
spam_df.describe()
```

Out[23]:

	category	Message-in-email
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

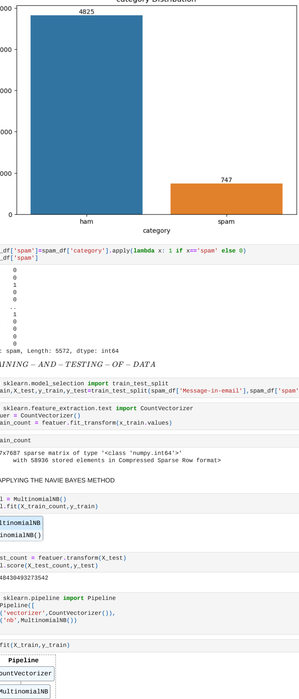
In [21]:
```
spam_df.shape
```

Out[21]:
```
(5572, 2)
```

In [22]:
```
spam_df.size
```

Out[22]:
```
11144
```

In [23]:
```
category_counts=spam_df['category'].value_counts().reset_index()
category_counts.columns=['category','count']
plt.figure(figsize=(8,6))
sns.barplot(x='category',y='count',data=category_counts)
plt.xlabel('category')
plt.ylabel('count')
plt.title('category Distribution')
for i,count in enumerate(category_counts['count']):
    plt.text(i,count,str(count),ha='center',va='bottom')
plt.show()
```



In [24]:
```
spam_df['spam']=spam_df['category'].apply(lambda x: 1 if x=='spam' else 0)
spam_df['spam']
```

Out[24]:
```
0       0
1       0
2       1
3       0
4       0
       ..
5567    1
5568    0
5569    0
5570    0
5571    0
Name: spam, Length: 5572, dtype: int64
```

$TRAINING-AND-TESTING-OF-DATA$

In [38]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(spam_df['Message-in-email'],spam_df['spam'],test_size=0.2)
```

In [39]:
```
from sklearn.feature_extraction.text import CountVectorizer
featuer = CountVectorizer()
X_train_count = featuer.fit_transform(x_train.values)
```

In [40]:
```
X_train_count
```

Out[40]:
```
<4457x7687 sparse matrix of type '<class 'numpy.int64'>'
	with 58936 stored elements in Compressed Sparse Row format>
```

- APPLYING THE NAVIE BAYES METHOD

In [41]:
```
model = MultinomialNB()
model.fit(X_train_count,y_train)
```

Out[41]:
```
▾ MultinomialNB
MultinomialNB()
```

In [42]:
```
X_test_count = featuer.transform(X_test)
model.score(X_test_count,y_test)
```

Out[42]:
```
0.8448430493273542
```

In [46]:
```
from sklearn.pipeline import Pipeline
clf=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('nb',MultinomialNB())
])
```

In [47]:
```
clf.fit(X_train,y_train)
```

Out[47]:
```
▸ Pipeline
▾ CountVectorizer
▾ MultinomialNB
```

In [48]:
```
clf.score(X_test,y_test)
```

Out[48]:
```
0.9793721973094417
```

- NOW DESIGN A PRE_BUILD MODEL TO DETECT SPAM AND NOT SPAM MESSAGE

In [50]:
```
pretrained_model=model
new_sentences=[
    "Your account have 100 debited, is waiting to be collected.Simply text the password \MIX\" to 85069 to verify. Get Usher and Britnay.FML is a sp
]
new_sentences_count=featuer.transform(new_sentences)
predictions=pretrained_model.predict(new_sentences_count)
for sentence,prediction in zip(new_sentences,predictions):
    if prediction == 1:
        print(f"'{sentence}' is a spam message.")
    else:
        print(f"'{sentence}' is not a spam message.")
```

```
'Your account have 100 debited, is waiting to be collected.Simply text the password \MIX" to 85069 to verify. Get Usher and Britnay.FML is a spam m
essage.' is not a spam message.
```