WEEK-10

Naive Bayes

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("spam.csv")
df.head()
```

```
  Class                                               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...
```

```
df.info()
```

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

```
df.describe()
```

```
       Class                     sms
count   5572                    5572
unique     2                    5169
top      ham  Sorry, I'll call later
freq    4825                      30
```

```
df.columns
```

```
Index(['Class', 'sms'], dtype='object')
```

```
df.groupby("Class").describe()
```

```
        sms                                                          
      count unique                                                top freq
Class                                                                 
ham    4825   4516                             Sorry, I'll call later   30
spam    747    653  Please call our customer service representativ...    4
```

```
df['spam']=df["Class"].apply(lambda x: 1 if x=="spam" else 0)
df.head()
```

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

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(df.sms,df.spam,test_size=30,random_state=20)
```

```
from sklearn.feature_extraction.text import CountVectorizer
v=CountVectorizer()
X_train_count=v.fit_transform(X_train.values)
X_train_count.toarray()[:3]
```

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

```
from sklearn.naive_bayes import MultinomialNB
model=MultinomialNB()
model.fit(X_train_count,Y_train)
```

```
MultinomialNB()
```

```
msgs=[
    "Hey Ram,can we get together to watch cricket match tomorrow?",
    "Upto 20% discount on parking,exclusive offer just for you,Don't miss this offer!"
]
msgs_count=v.transform(msgs)
model.predict(msgs_count)
```

```
array([0, 1], dtype=int64)
```

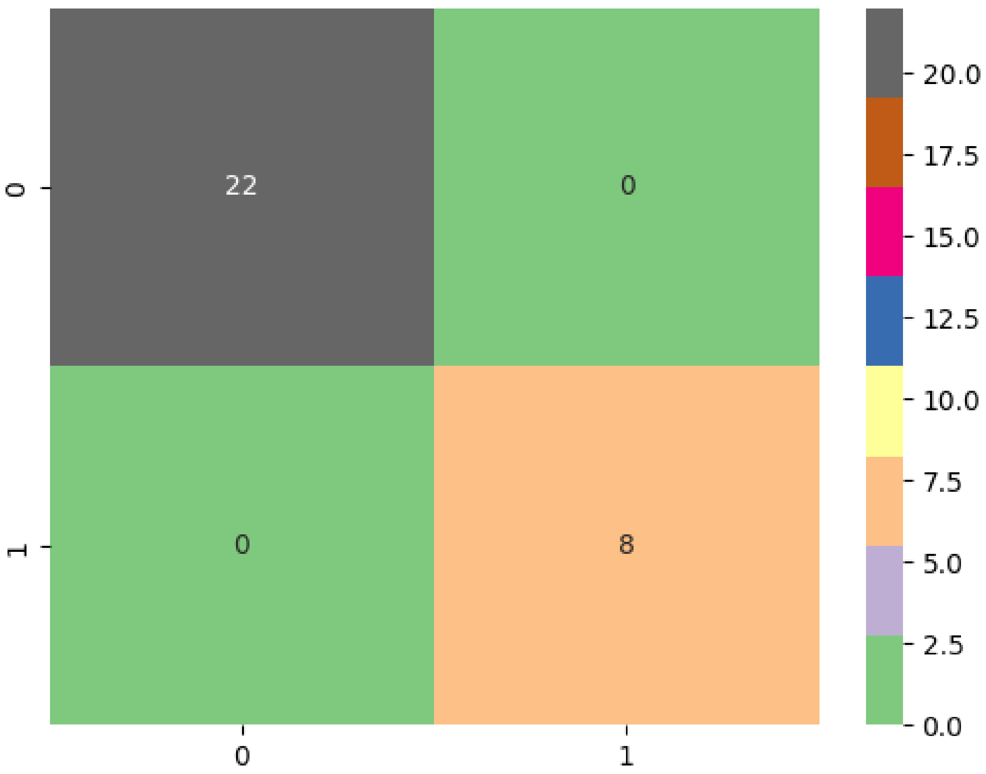
```
X_test_count=v.transform(X_test)
model.score(X_test_count,Y_test)
```

```
1.0
```

```
y_predicted=model.predict(X_test_count)
```

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(Y_test,y_predicted)
sns.heatmap(cm,annot=True,cmap="Accent")
```

```
<Axes: >
```


22
0
0
8
0
1
0
1
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0