spam

February 24, 2024

[1]:
```
import pandas as pd
```

[2]:
```
import numpy as np
```

[3]:
```
pip install --upgrade pandas
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(2.2.1)
Requirement already satisfied: numpy<2,>=1.22.4 in
/usr/local/lib/python3.10/dist-packages (from pandas) (1.25.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2023.4)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

[4]:
```
pd.__version__
```

[4]:
```
'2.2.1'
```

[5]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

[6]:
```
data = pd.read_csv('spam.csv', encoding='iso-8859-1')
```

[7]:
```
pip install chardet
```

```
Requirement already satisfied: chardet in /usr/local/lib/python3.10/dist-
packages (5.2.0)
```

[8]:
```
import chardet

with open('spam.csv', 'rb') as f:
    result = chardet.detect(f.read())
```

```
print(result['encoding'])
```

```
Windows-1252
```

[9]:
```
import pandas as pd

try:
    data = pd.read_csv('spam.csv', encoding='utf-8')
except UnicodeDecodeError as e:
    print(f"Error reading file: {e}")
```

```
Error reading file: 'utf-8' codec can't decode bytes in position 135-136:
invalid continuation byte
```

[10]:
```
data
```

[10]:
```
        v1                                                 v2 Unnamed: 2  \
0      ham  Go until jurong point, crazy.. Available only …         NaN
1      ham                      Ok lar… Joking wif u oni…         NaN
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…         NaN
3      ham  U dun say so early hor… U c already then say…         NaN
4      ham  Nah I don't think he goes to usf, he lives aro…         NaN
…     …                                                  …         …
5567  spam  This is the 2nd time we have tried 2 contact u…         NaN
5568   ham              Will Ì_ b going to esplanade fr home?          NaN
5569   ham  Pity, * was in mood for that. So…any other s…         NaN
5570   ham  The guy did some bitching but I acted like i'd…         NaN
5571   ham                        Rofl. Its true to its name          NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
…         …        …
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]
```

[11]:
```
data.info
```

[11]:
```
<bound method DataFrame.info of         v1
v2 Unnamed: 2  \
0      ham  Go until jurong point, crazy.. Available only …        NaN
1      ham                      Ok lar… Joking wif u oni…        NaN
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…        NaN
3      ham  U dun say so early hor… U c already then say…        NaN
4      ham  Nah I don't think he goes to usf, he lives aro…        NaN
…    …                                                  …        …
5567  spam  This is the 2nd time we have tried 2 contact u…        NaN
5568   ham              Will Ì_ b going to esplanade fr home?        NaN
5569   ham  Pity, * was in mood for that. So…any other s…        NaN
5570   ham  The guy did some bitching but I acted like i'd…        NaN
5571   ham                        Rofl. Its true to its name        NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
…          …         …
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]>
```

[12]:
```
data.head()
```

[12]:
```
     v1                                                  v2 Unnamed: 2  \
0   ham  Go until jurong point, crazy.. Available only …        NaN
1   ham                      Ok lar… Joking wif u oni…        NaN
2  spam  Free entry in 2 a wkly comp to win FA Cup fina…        NaN
3   ham  U dun say so early hor… U c already then say…        NaN
4   ham  Nah I don't think he goes to usf, he lives aro…        NaN

  Unnamed: 3 Unnamed: 4
0        NaN        NaN
1        NaN        NaN
2        NaN        NaN
3        NaN        NaN
4        NaN        NaN
```

[13]:
```
data.shape
```

[13]:
```
(5572, 5)
```

[14]:
```
data.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'],inplace=True)
```

[15]:
```
data
```

[15]:
```
        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 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 x 2 columns]
```

[16]:
```
data.rename(columns={'v1':'phone','v2':'message'},inplace=True)
data.head()
```

[16]:
```
  phone                                            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…
```

[17]:
```
data.duplicated().sum()
```

[17]:
```
403
```

[18]:
```
data1 = data.drop_duplicates()
```

[19]:
```
data1.duplicated().sum()
```

[19]:
```
0
```

[20]:
```
data1.shape
```

[20]:
```
(5169, 2)
```

[21]:
```
data['phone'].value_counts()
```

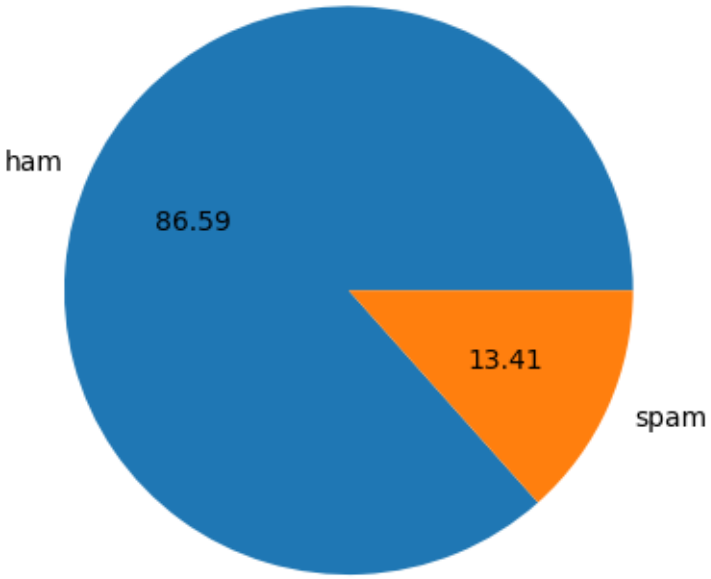
[21]:
```
phone
ham     4825
spam     747
Name: count, dtype: int64
```

[22]:
```
import matplotlib.pyplot as plt
```

[23]:
```
data['phone'].unique()
```

[23]:
```
array(['ham', 'spam'], dtype=object)
```

[24]:
```
plt.pie(data['phone'].value_counts(), labels=['ham','spam'],autopct="%0.2f")
plt.show()
```



[25]:
```
import nltk
```

[26]:
```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages
(3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages
(from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages
(from nltk) (1.3.2)
```

```
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.10/dist-packages (from nltk) (2023.12.25)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages
(from nltk) (4.66.2)
```

[27]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

[27]:
```
True
```

[28]:
```
data['num_characters'] = data['message'].apply(len)
```

[29]:
```
data
```

[29]:
```
     phone                                              message  num_characters
0      ham  Go until jurong point, crazy.. Available only …              111
1      ham                      Ok lar… Joking wif u oni…               29
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…              155
3      ham  U dun say so early hor… U c already then say…               49
4      ham  Nah I don't think he goes to usf, he lives aro…               61
…    …                                                  …              …
5567  spam  This is the 2nd time we have tried 2 contact u…              161
5568   ham              Will Ì_ b going to esplanade fr home?               37
5569   ham  Pity, * was in mood for that. So…any other s…               57
5570   ham  The guy did some bitching but I acted like i'd…              125
5571   ham                         Rofl. Its true to its name               26

[5572 rows x 3 columns]
```

[30]:
```
data['num_words'] = data['message'].apply(lambda x:len(nltk.word_tokenize(x)))
```

[31]:
```
data
```

[31]:
```
     phone                                              message  num_characters  \
0      ham  Go until jurong point, crazy.. Available only …              111
1      ham                      Ok lar… Joking wif u oni…               29
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…              155
3      ham  U dun say so early hor… U c already then say…               49
4      ham  Nah I don't think he goes to usf, he lives aro…               61
…    …                                                  …              …
5567  spam  This is the 2nd time we have tried 2 contact u…              161
5568   ham              Will Ì_ b going to esplanade fr home?               37
5569   ham  Pity, * was in mood for that. So…any other s…               57
5570   ham  The guy did some bitching but I acted like i'd…              125
5571   ham                         Rofl. Its true to its name               26
```

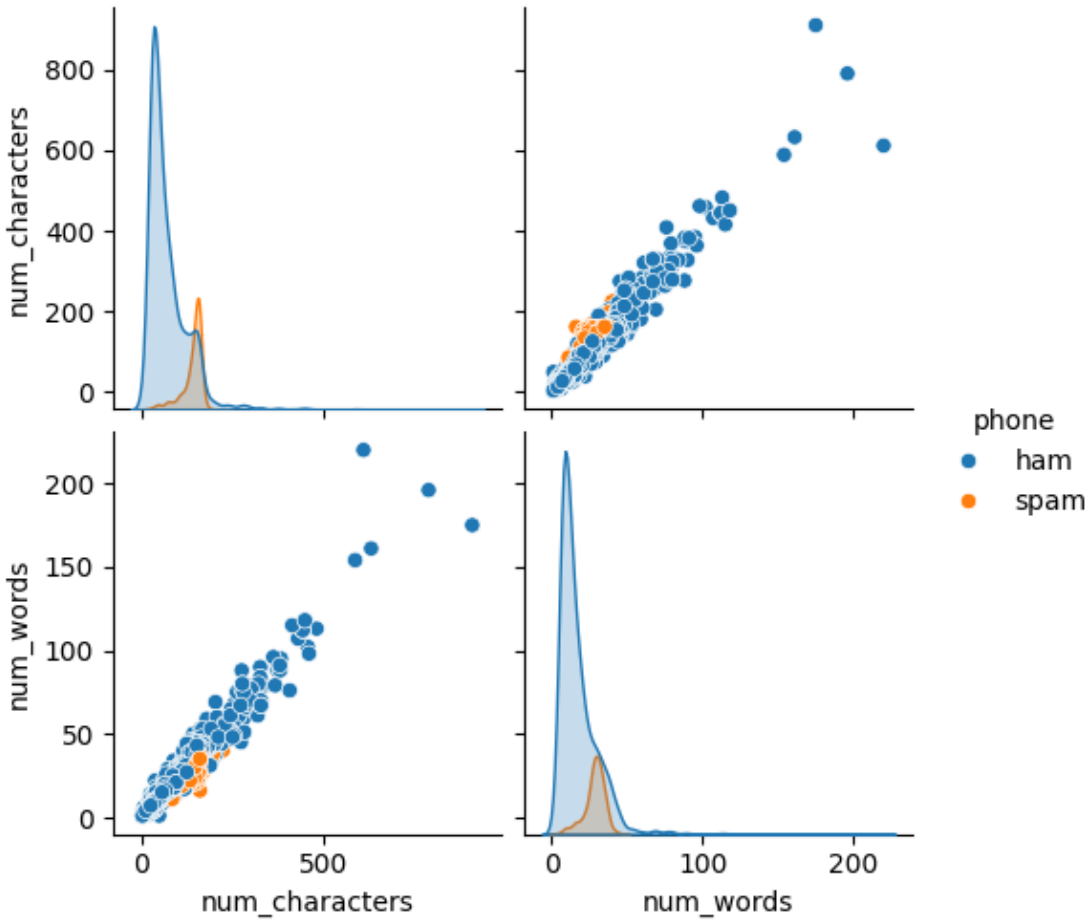
```
      num_words
0            24
1             8
2            37
3            13
4            15
…          …
5567         35
5568          9
5569         15
5570         27
5571          7

[5572 rows x 4 columns]
```

[32]:

```python
sns.pairplot(data,hue='phone')
```

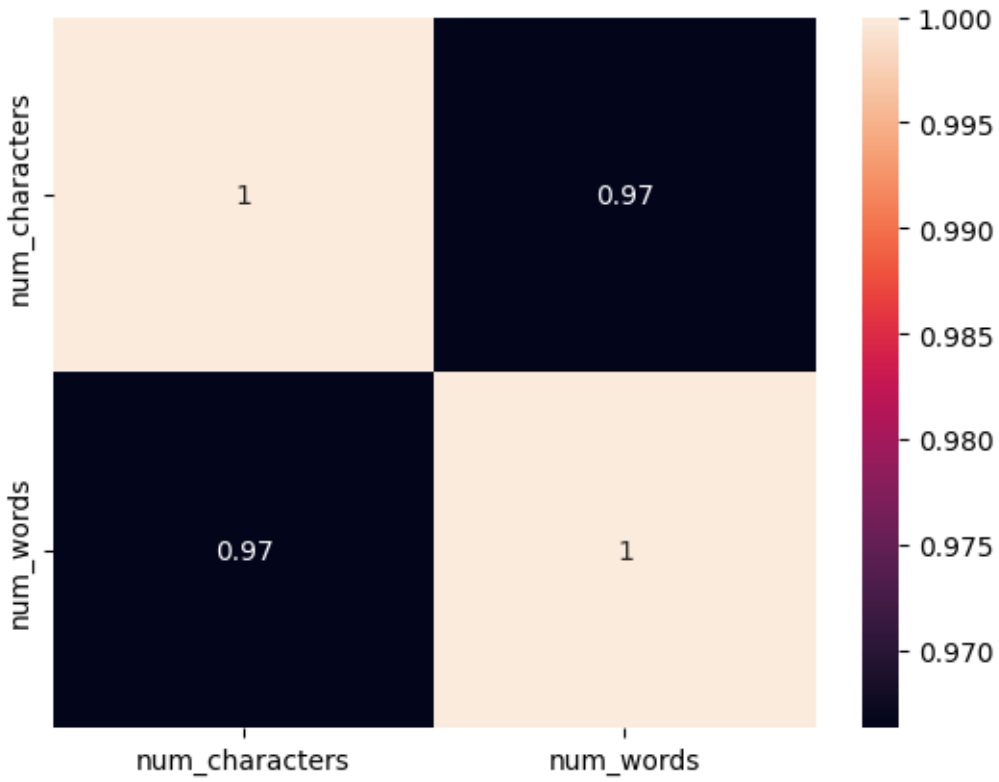
```
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

[32]: <seaborn.axisgrid.PairGrid at 0x7f7438191690>



[33]:
```
numeric_values = data.select_dtypes( include = ['number'])
sns.heatmap(numeric_values.corr(),annot=True)
```

[33]: <Axes: >



[34]:
```
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
import string
```

[35]:
```
nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Unzipping corpora/stopwords.zip.
```

[35]: True

[36]:
```
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)

    y = []
    for i in text:
        if i.isalnum():
```

```
            y.append(i)
    text = y[:]
    y.clear()

    for i in text:
        if i not in stopwords.words('english') and i not in string.punctuation:
            y.append(i)
    text = y[:]
    y.clear()
    ps = PorterStemmer()
    for i in text:
        y.append(ps.stem(i))

    return " ".join(y)
```

[37]:
```
data['message'][10]
```

[37]:
```
"I'm gonna be home soon and i don't want to talk about this stuff anymore
tonight, k? I've cried enough today."
```

[38]:
```
ps = PorterStemmer()
ps.stem('loving')
```

[38]:
```
'love'
```

[39]:
```
data['transformed_text'] = data['message'].apply(transform_text)
```

[40]:
```
data.head()
```

[40]:

	phone	message	num_characters	\
0	ham	Go until jurong point, crazy.. Available only …	111	
1	ham	Ok lar… Joking wif u oni…	29	
2	spam	Free entry in 2 a wkly comp to win FA Cup fina…	155	
3	ham	U dun say so early hor… U c already then say…	49	
4	ham	Nah I don't think he goes to usf, he lives aro…	61	

	num_words	transformed_text
0	24	go jurong point crazi avail bugi n great world…
1	8	ok lar joke wif u oni
2	37	free entri 2 wkli comp win fa cup final tkt 21…
3	13	u dun say earli hor u c alreadi say
4	15	nah think goe usf live around though

[41]:
```
from wordcloud import WordCloud
wc = WordCloud(width=500,height=500,min_font_size=10,background_color='white')
```

[42]:
```
spam_wc = wc.generate(data[data['phone'] == 1]['transformed_text'].str.
↪cat(sep=" ")) if len(data[data['phone'] == 1]['transformed_text']) > 0 else␣
↪None
```

[43]:
```
# Generate the word cloud if there is text data available
if len(data[data['message'] == 1]['transformed_text']) > 0:
    spam_wc = wc.generate(data[data['phone'] == 1]['transformed_text'].str.
↪cat(sep=" "))

    # Convert WordCloud object to array
    wordcloud_array = spam_wc.to_array()

    # Display the word cloud
    plt.figure(figsize=(15, 6))
    plt.imshow(wordcloud_array, interpolation='bilinear')
    plt.axis('off')
    plt.show()
else:
    print("No text data found for the given condition.")
```

```
No text data found for the given condition.
```

[47]:
```
spam_corpus = []
for msg in data[data['phone'] == 1]['transformed_text'].tolist():
    for word in msg.split():
        spam_corpus.append(word)
```

[48]:
```
len(spam_corpus)
```

[48]:
```
0
```

[52]:
```
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
cv = CountVectorizer()
tfidf = TfidfVectorizer(max_features=3000)
```

[53]:
```
X = tfidf.fit_transform(data['transformed_text']).toarray()
```

[54]:
```
X.shape
```

[54]:
```
(5572, 3000)
```

[56]:
```
y = data['phone'].values
```

[57]:
```
from sklearn.model_selection import train_test_split
```

[58]:
```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.
↪2,random_state=2)
```

[59]:
```
from sklearn.naive_bayes import GaussianNB,MultinomialNB,BernoulliNB
from sklearn.metrics import accuracy_score,confusion_matrix,precision_score
```

[60]:
```
gnb = GaussianNB()
mnb = MultinomialNB()
bnb = BernoulliNB()
```

[62]:
```
gnb.fit(X_train,y_train)
y_pred1 = gnb.predict(X_test)
print(accuracy_score(y_test,y_pred1))
print(confusion_matrix(y_test,y_pred1))
```

```
0.8672645739910314
[[841 116]
 [ 32 126]]
```

[63]:
```
print(precision_score(y_test, y_pred1, pos_label='spam'))
```

```
0.5206611570247934
```

[65]:
```
mnb.fit(X_train,y_train)
y_pred2 = mnb.predict(X_test)
print(accuracy_score(y_test,y_pred2))
print(confusion_matrix(y_test,y_pred2))
print(precision_score(y_test, y_pred2, pos_label='spam'))
```

```
0.9650224215246637
[[956   1]
 [ 38 120]]
0.9917355371900827
```

[67]:
```
bnb.fit(X_train,y_train)
y_pred3 = bnb.predict(X_test)
print(accuracy_score(y_test,y_pred3))
print(confusion_matrix(y_test,y_pred3))
print(precision_score(y_test, y_pred3, pos_label='spam'))
```

```
0.9748878923766816
[[955   2]
 [ 26 132]]
0.9850746268656716
```

[70]:
```
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import ExtraTreesClassifier

svc = SVC(kernel='sigmoid', gamma=1.0, probability=True)
mnb = MultinomialNB()
```

```
etc = ExtraTreesClassifier(n_estimators=50, random_state=2)
```

[71]:
```
from sklearn.ensemble import VotingClassifier
```

[72]:
```
voting = VotingClassifier(estimators=[('svm', svc), ('nb', mnb), ('et',␣
↪etc)],voting='soft')
```

[73]:
```
voting.fit(X_train,y_train)
```

[73]:
```
VotingClassifier(estimators=[('svm',
                              SVC(gamma=1.0, kernel='sigmoid',
                                  probability=True)),
                             ('nb', MultinomialNB()),
                             ('et',
                              ExtraTreesClassifier(n_estimators=50,
                                                   random_state=2))],
                 voting='soft')
```

[76]:
```
y_pred = voting.predict(X_test)
print("Accuracy",accuracy_score(y_test,y_pred))
```

```
Accuracy 0.9775784753363229
```

[ ]:
```

```