ASSIGNMENT 15/ TASK 15

Shweta Jha

Registration Id-GO_STP_12574

Build a spam filter using Python and the multinomial Naive Bayes algorithm.

Check Spam or Ham? Email Classifier Using Python using MultinomialNB.

Dataset click here.

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

```
df=pd.read_csv('/content/spam.csv')
df.head()
```

	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...

```
df.tail()
```

	Category	Message
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü 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

```
df.shape
```

```
(5572, 2)
```

```
df.describe()
```

	Category	Message
count	5572	5572
unique	2	5157
top	ham	Sorry, I'll call later
freq	4825	30

```
df.sum()
```

```
Category    hamhamspamhamhamspamhamhamspamspamhamspamspamh...
Message     Go until jurong point, crazy.. Available only ...
dtype: object
```

```
df.dtypes
```

```
Category    object
Message     object
dtype: object
```

```
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   5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

```
df.size
```

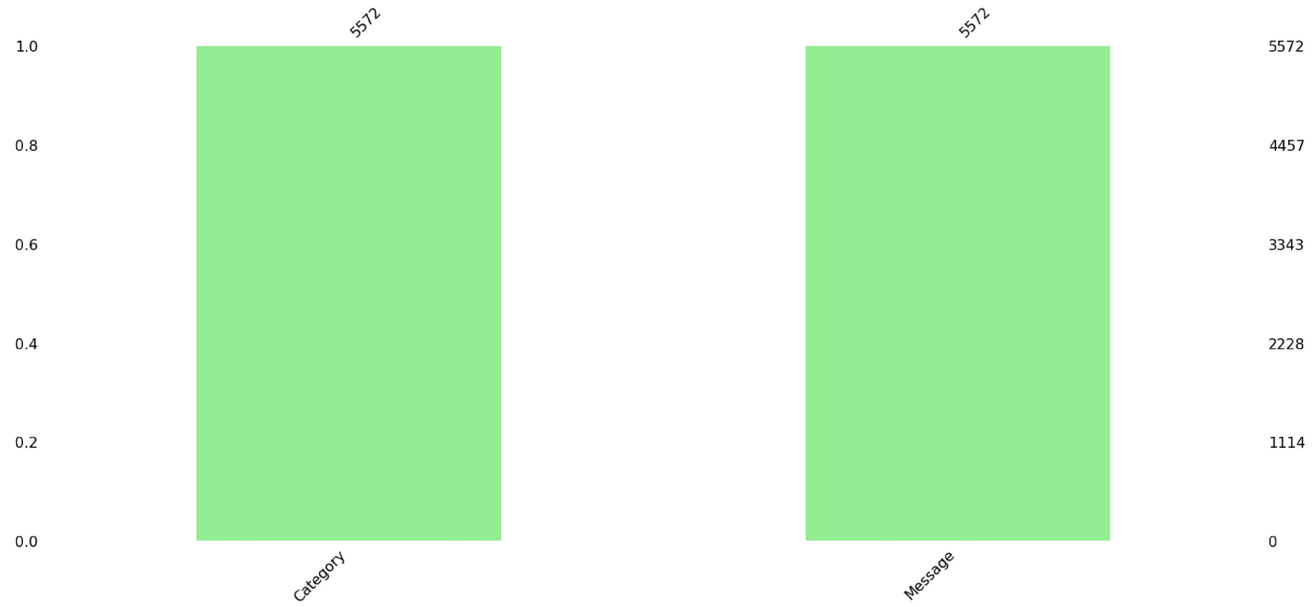
```
11144
```

```
df.columns
```

```
Index(['Category', 'Message'], dtype='object')
```

```
x=df.iloc[:,[0,1]].values
y=df.iloc[:,1].values
```

```
import missingno as msno
msno.bar(df,color="lightgreen")
plt.show()
```



```
pip install scikit-plot
```

```
Collecting scikit-plot
  Downloading https://files.pythonhosted.org/packages/7c/47/32520e259340c140a4ad27c1b
Requirement already satisfied: matplotlib>=1.4.0 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: joblib>=0.10 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: scipy>=0.9 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: scikit-learn>=0.18 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from cy
Installing collected packages: scikit-plot
Successfully installed scikit-plot-0.3.7
```

```
pip install texthero
```

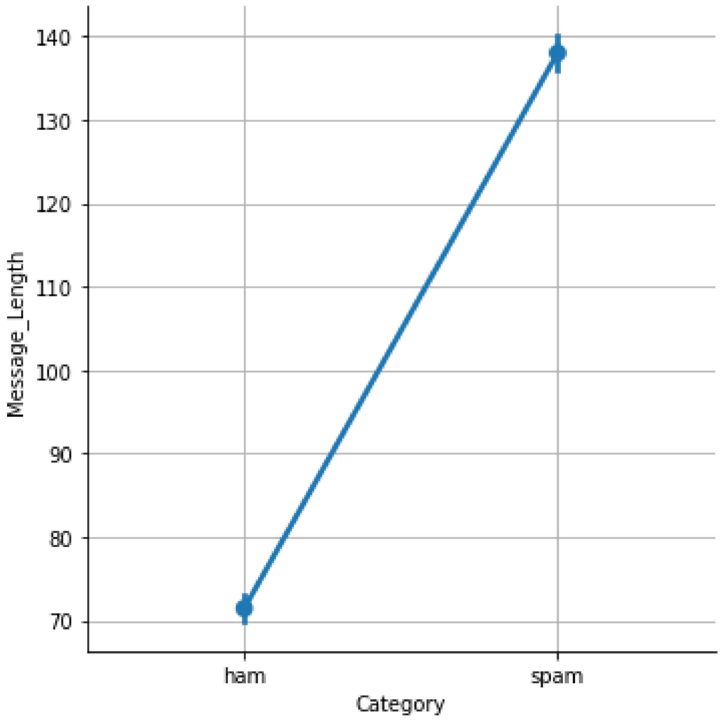
```
Collecting texthero
  Downloading https://files.pythonhosted.org/packages/1f/5a/a9d33b799fe53011de79d140a
Requirement already satisfied: tqdm>=4.3 in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: matplotlib>=3.1.0 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: spacy>=2.2.2 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: pandas>=1.0.2 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: gensim>=3.6.0 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: plotly>=4.2.0 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: wordcloud>=1.5.0 in /usr/local/lib/python3.7/dist-pack
Collecting unidecode>=1.1.1
  Downloading https://files.pythonhosted.org/packages/9e/25/723487ca2a52ebcee88a34d7c
     |████████████████████████████████| 245kB 7.7MB/s
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages
Collecting nltk>=3.3
  Downloading https://files.pythonhosted.org/packages/5e/37/9532ddd4b1bbb619333d5708a
     |████████████████████████████████| 1.5MB 37.9MB/s
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: thinc==7.4.0 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.7/
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in /usr/local/lib/python3.7/dist-p
Requirement already satisfied: plac<1.2.0,>=0.9.6 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.7/dist
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.7/dist-p
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in /usr/local/lib/python3.7/di
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.7/di
Requirement already satisfied: blis<0.5.0,>=0.4.0 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: six>=1.5.0 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: scipy>=0.18.1 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: retrying>=1.3.3 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: pillow in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: regex in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8" in /u
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: typing-extensions>=3.6.4; python_version < "3.8" in /u
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (f
Installing collected packages: unidecode, nltk, texthero
  Found existing installation: nltk 3.2.5
    Uninstalling nltk-3.2.5:
      Successfully uninstalled nltk-3.2.5
Successfully installed nltk-3.6.2 texthero-1.0.9 unidecode-1.2.0
```

```
import scikitplot as skplt
import warnings
import texthero as hero
```

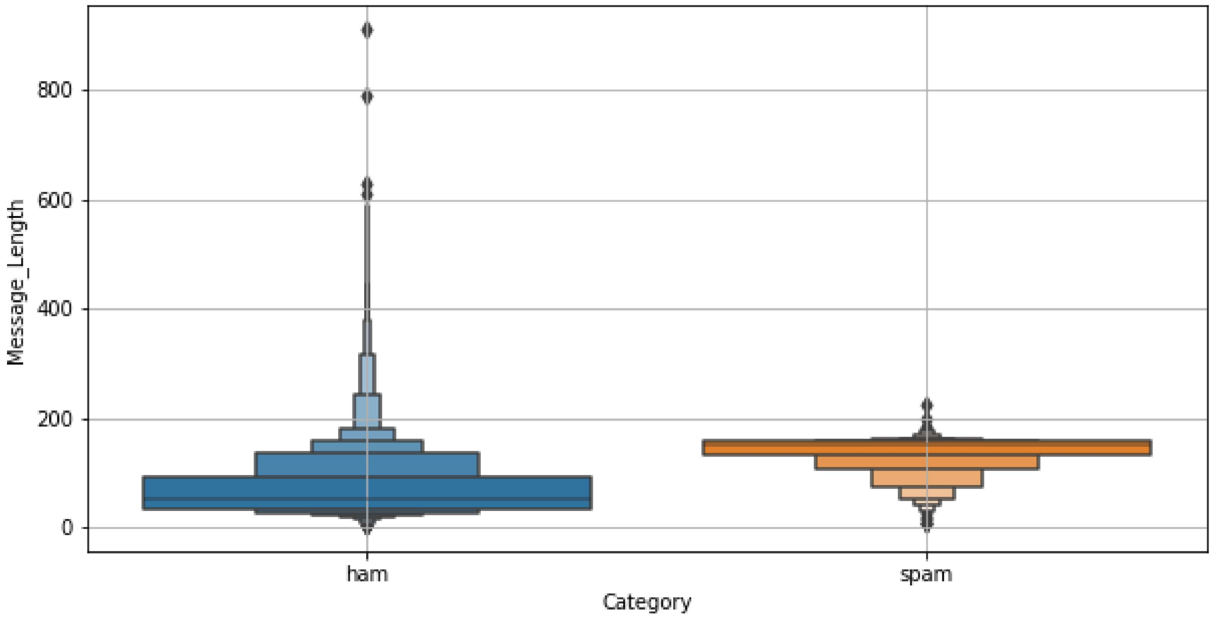
```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
df["Message_Length"]=df["Message"].apply(len)
```

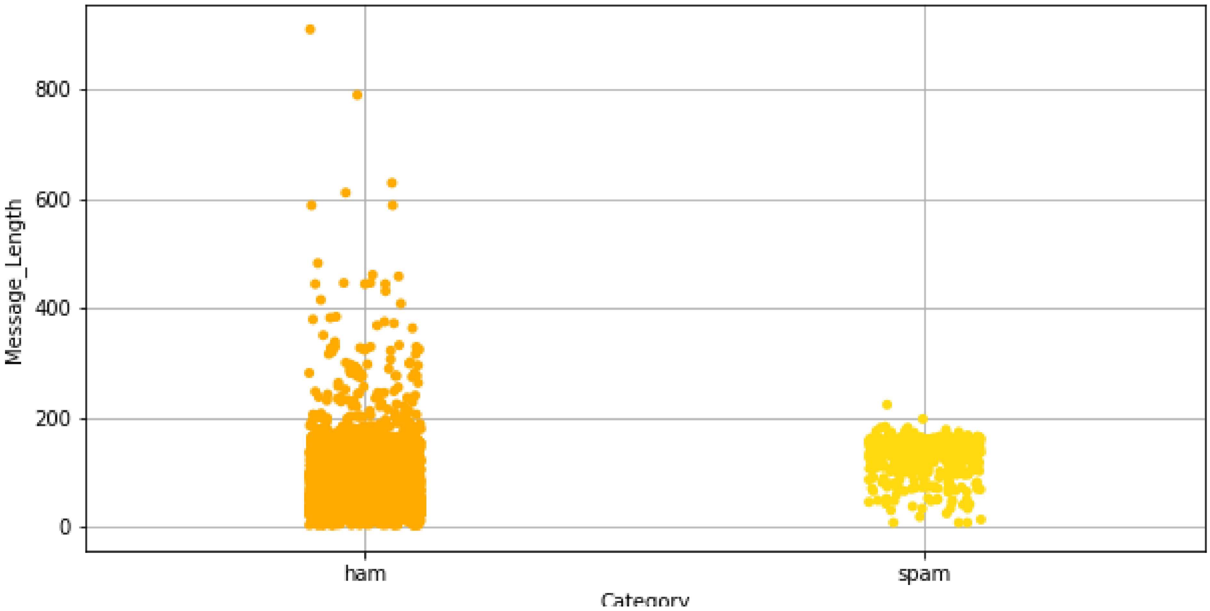
```
sns.catplot(data=df,y='Message_Length',x='Category', kind="point")
plt.grid()
plt.show()
```



```
plt.figure(figsize=(10,5))
sns.boxenplot(x='Category',y='Message_Length',data=df)
plt.grid()
plt.show()
```

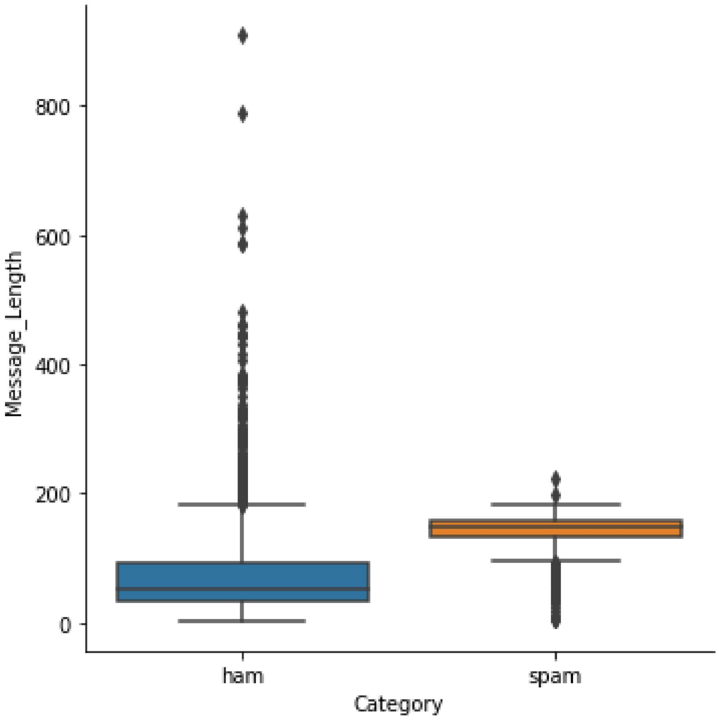


```
plt.figure(figsize=(10,5))
sns.stripplot(data=df,y='Message_Length',x='Category',palette='Wistia_r')
plt.grid()
plt.show()
```

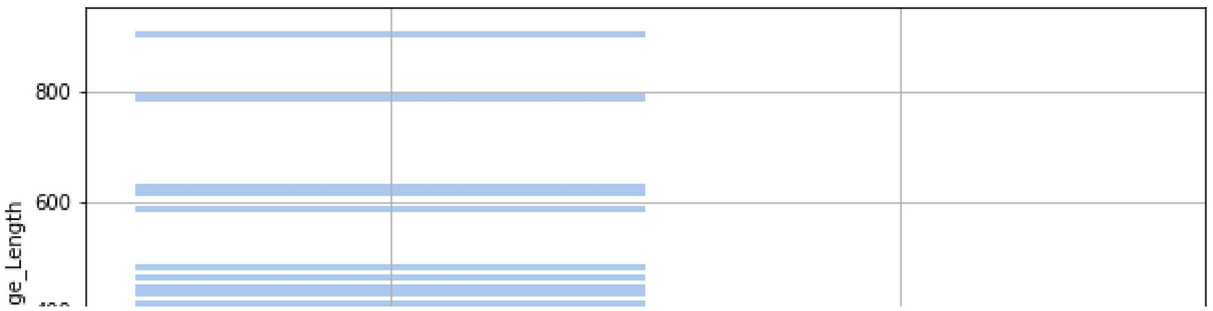


```
sns.catplot(data=df,y='Message_Length',x='Category', kind="box")
```

```
<seaborn.axisgrid.FacetGrid at 0x7fbad5861bd0>
```



```
plt.figure(figsize=(10,5))
sns.histplot(data=df,y='Message_Length',x='Category',kde=True,palette='Reds')
plt.grid()
plt.show()
```

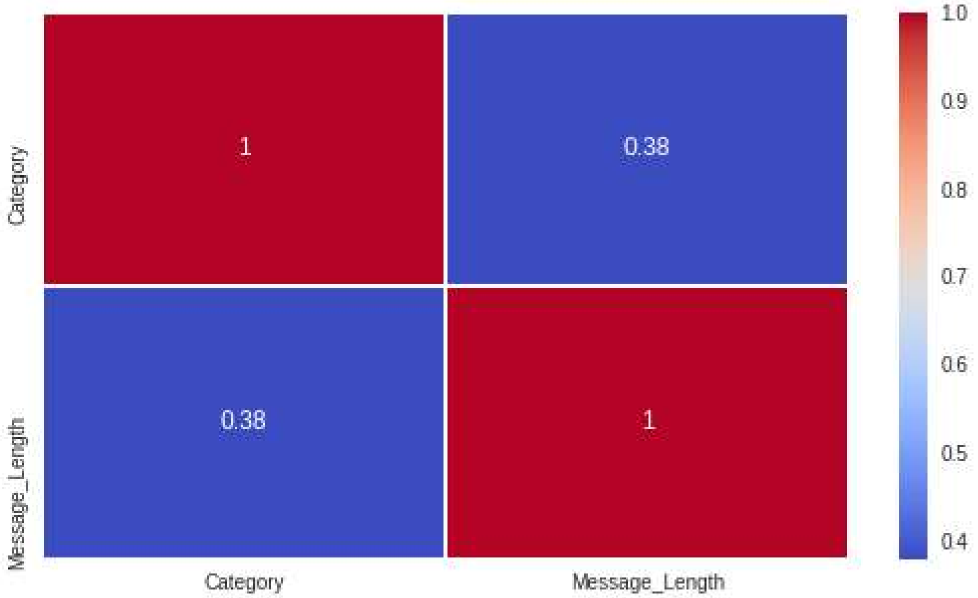


```
from yellowbrick.target import feature_correlation,BalancedBinningReference,ClassBalance
from sklearn.preprocessing import LabelEncoder
label_encoder=LabelEncoder()
```

```
df.Category=label_encoder.fit_transform(df.Category)
```

```
plt.figure(figsize=(9,5))
sns.heatmap(df.corr(), cmap='coolwarm', annot=True, linewidths=0.30)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbad579a550>
```



```
label_encoder.classes_
```

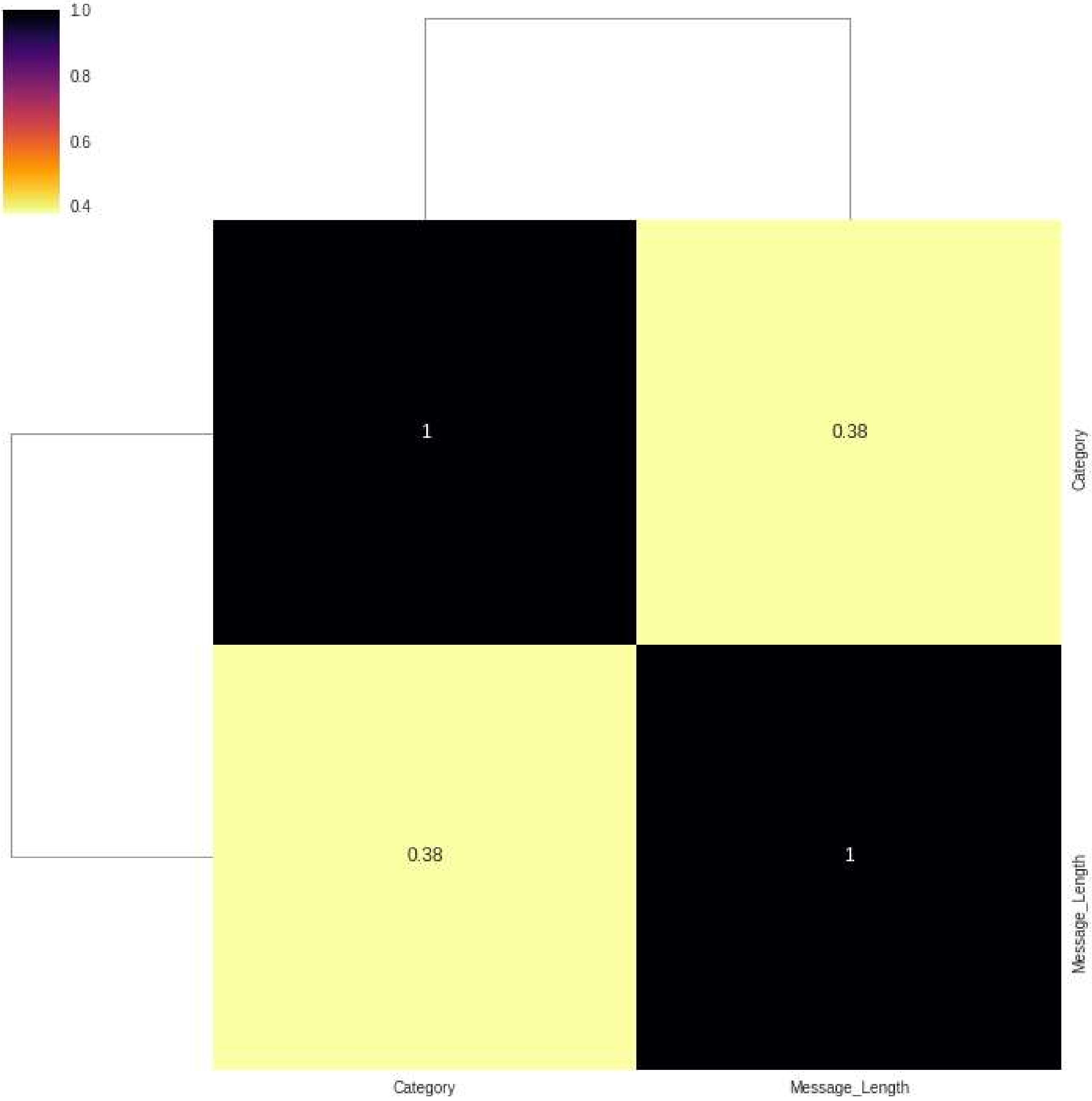
```
array(['ham', 'spam'], dtype=object)
```

```
df.head()
```

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

```
plt.figure(figsize=(3,4))
sns.clustermap(df.corr(),annot=True,cmap='inferno_r')
```

```
<seaborn.matrix.ClusterGrid at 0x7fbad55cd4d0>
<Figure size 216x288 with 0 Axes>
```



Spliting DataSet

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
textFeatures=df['Message'].copy()
vectorizer=TfidfVectorizer("english")
x=vectorizer.fit_transform(textFeatures)
y=df["Category"]
```

```
y=df['Category']
```

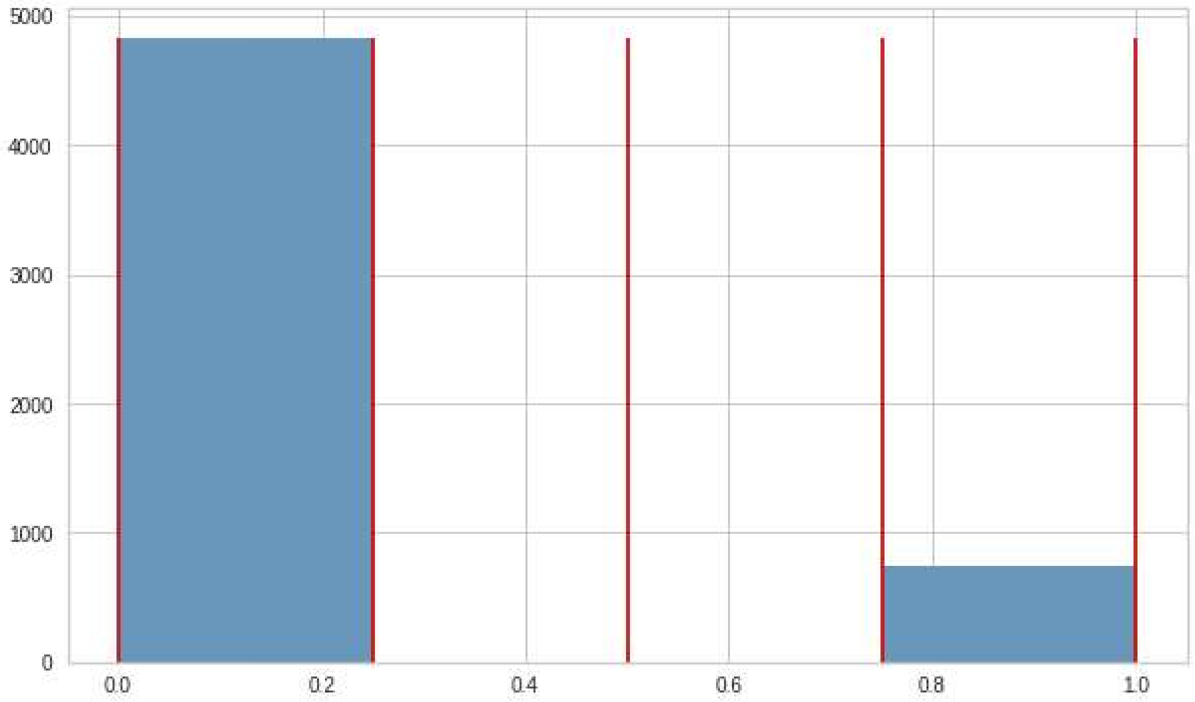
```
x[:5]
```

```
<5x8709 sparse matrix of type '<class 'numpy.float64'>'
	with 64 stored elements in Compressed Sparse Row format>
```

```
y[:5]
```

```
0    0
1    0
2    1
3    0
4    0
Name: Category, dtype: int64
```

```
plt.figure(figsize=(10,6))
visualizer=BalancedBinningReference()
visualizer.fit(y)
plt.show()
```



```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

```
X_train.shape
```

```
(4457, 8709)
```

```
y_train.shape
```

```
(4457,)
```

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix,accuracy_score,roc_auc_score
```

```
mnb_clf=MultinomialNB(alpha=0.1)
```

```
mnb_clf
```

```
    MultinomialNB(alpha=0.1, class_prior=None, fit_prior=True)
```

```
mnb_clf.fit(X_train,y_train)
```

```
    MultinomialNB(alpha=0.1, class_prior=None, fit_prior=True)
```

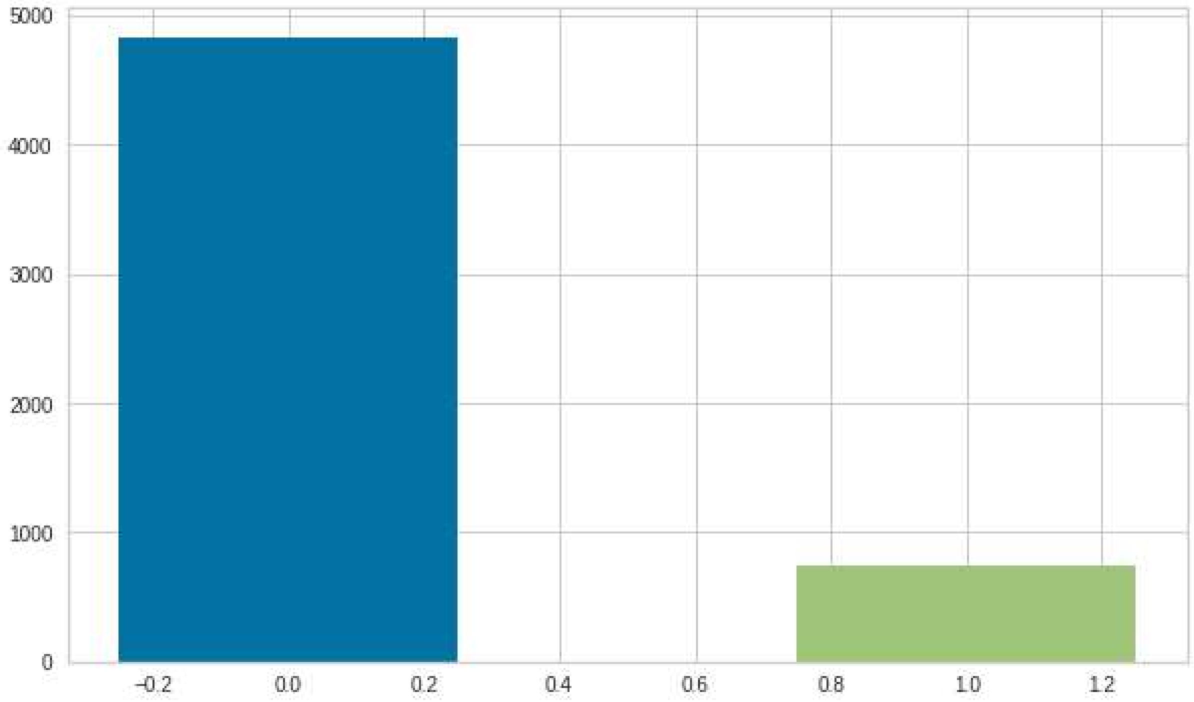
```
y_pred=mnb_clf.predict(X_test)
print(accuracy_score(y_test,y_pred))
```

```
    0.9865470852017937
```

```
classes=label_encoder.classes_
```

```
plt.figure(figsize=(10,6))
plt.figure(figsize=(10,6))
viz=ClassBalance(label=classes,colors=['lightblue','lightgreen'])
viz.fit(y)
plt.show()
```

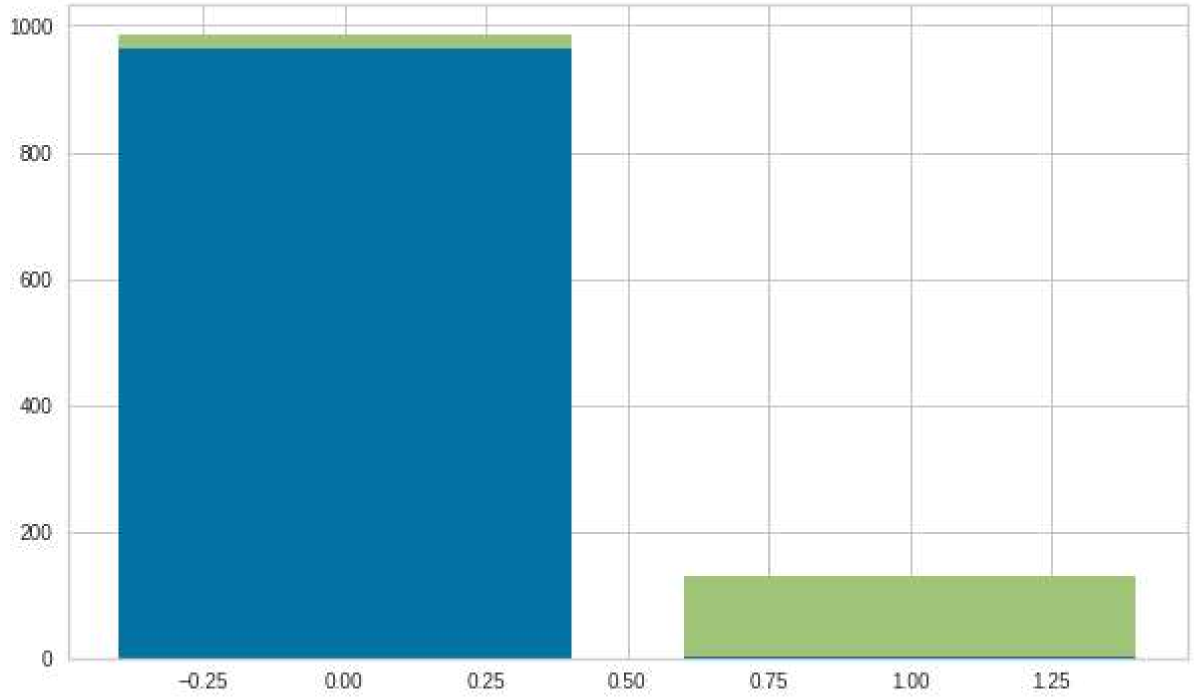
```
    <Figure size 720x432 with 0 Axes>
```



```
from sklearn.ensemble import RandomForestClassifier
from yellowbrick.classifier import ClassPredictionError
```

```
plt.figure(figsize=(10,6))
visualizer = ClassPredictionError(
    RandomForestClassifier(random_state=42, n_estimators=10), classes=classes
)
# Fit the training data to the visualizer
visualizer.fit(X_train, y_train)

# Evaluate the model on the test data
visualizer.score(X_test, y_test)
plt.show()
```



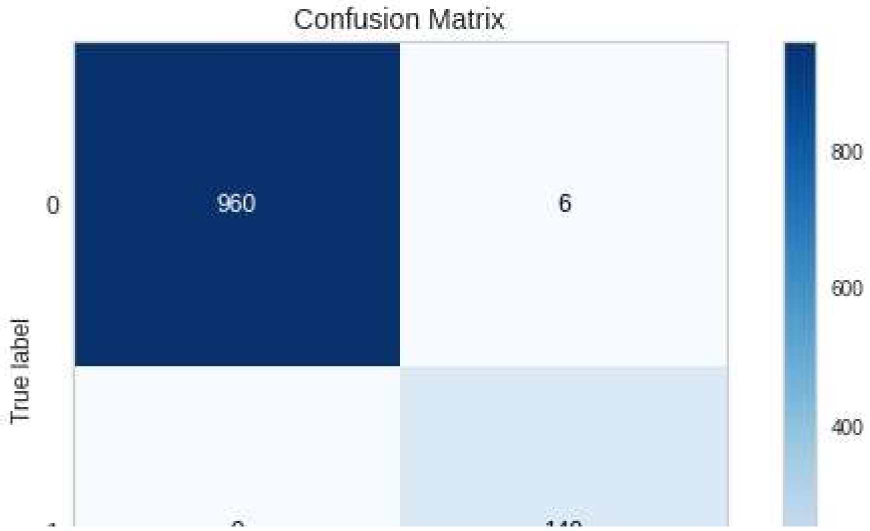
```
from sklearn.feature_extraction.text import TfidfTransformer
tf_transformer = TfidfTransformer(use_idf=False).fit(X_train_counts)
X_train_tf = tf_transformer.transform(X_train_counts)
X_train_tf.shape
```

```
     (3, 3)
```

```
confusion_matrix(y_test,y_pred)
```

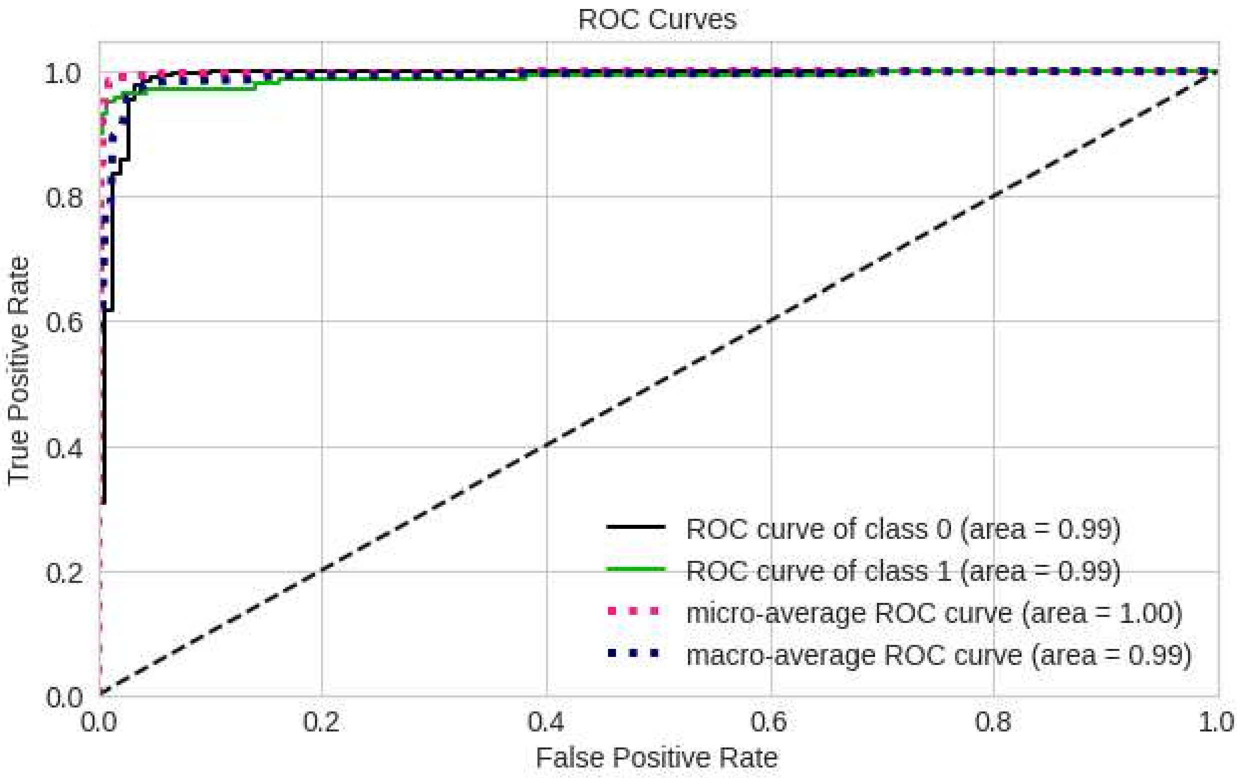
```
     array([[960,   6],
            [  9, 140]])
```

```
skplt.metrics.plot_confusion_matrix(y_test,y_pred,figsize=(10,6),title_fontsize=14)
plt.show()
```

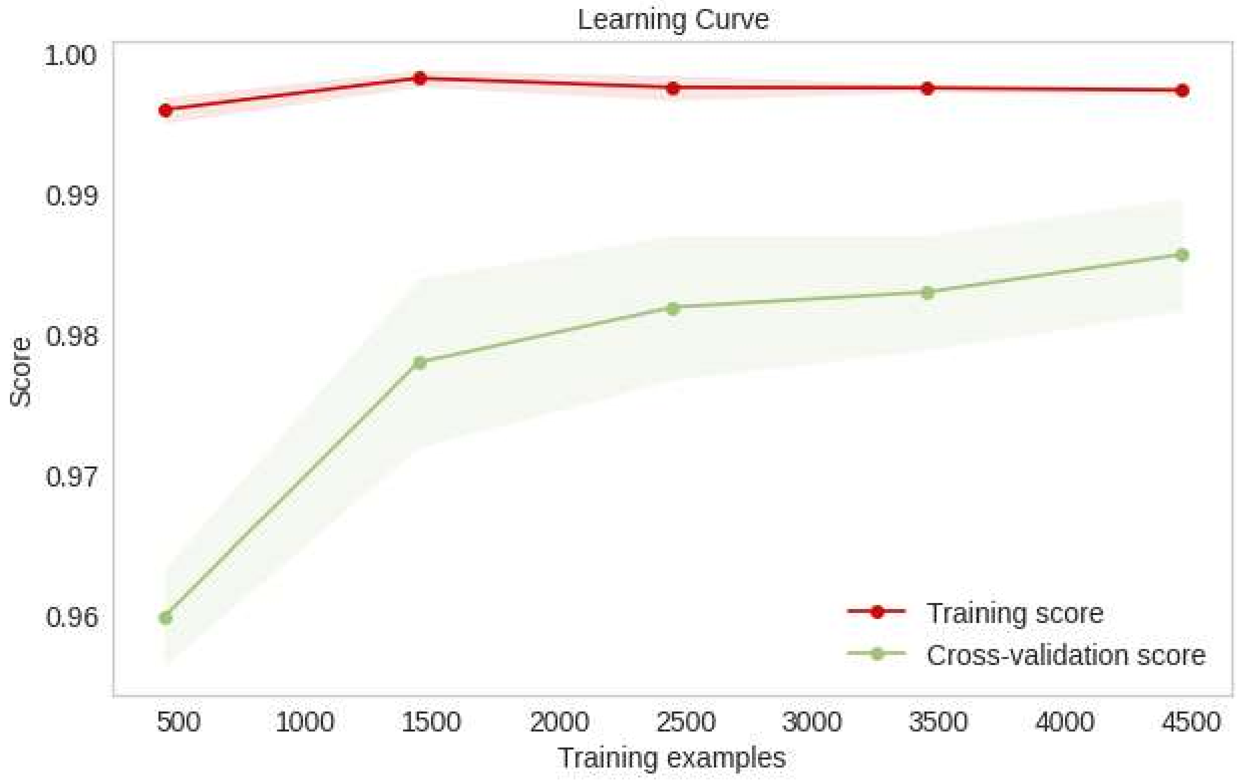


```
y_pro=mnb_clf.predict_proba(X_test)
```

```
skplt.metrics.plot_roc(y_test,y_pro,figsize=(10,6),title_fontsize=14,text_fontsize=14)
plt.show()
```



```
skplt.estimators.plot_learning_curve(mnb_clf,x,y,figsize=(10,6),title_fontsize=14,text_fon
plt.show()
```



✓ 0s completed at 11:06 AM ● ×