```
# In[1]:
import pandas as pd
import numpy as np
# In[ ]:
# In[36]:
import pandas as pd
# List of possible encodings to try
encodings=['utf-8','latin1','cp1252']
file_path='spam.csv' # Change this with different encodings
# Attempt to read the CSV file with different encodings
for encoding in encodings:
    try:
        df=pd.read_csv(file_path,encoding=encoding)
        print(f"file successfully read with encoding:{encoding}")
        break
    except UnicodeDecodeError:
            print(f"failed to read with encoding:{encoding}")
            continue # Try the next coding
if 'df' in locals():
    print("CSV file has been successfully loaded.")
else:
    print("All encoding attempts failed. Unable to read the CSV file.")
# In[37]:
df.sample(5)
# In[38]:
df.shape
# In[39]:
# 1. Data cleaning
# 2. EDA
# 3. Text Processing
# 4. Model building
# 5. Evaluation
# 6. Improvement
# 7. Website
# 8. Deploy
# In[40]:
df.info()
# In[41]:
# drop last 3 cols
df.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'], inplace=True)
# In[42]:
df.sample(5)
# In[43]:
# renaming the columns
df.rename(columns={'v1':'target','v2':'text'},inplace=True)
df.sample(5)
# In[44]:
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
# In[45]:
df['target'] =encoder.fit_transform(df['target'])
# In[46]:
df.head()
# In[47]:
# missing values
df.isnull().sum()
# In[48]:
```

```
df.duplicated().sum()
# In[49]:
df=df.drop_duplicates(keep='first')
# In[50]:
df.duplicated().sum()
# In[51]:
df.shape
# In[52]:
df.head()
# In[53]:
df["target"].value_counts()
# In[54]:
import matplotlib.pyplot as plt
plt.pie(df['target'].value_counts(),labels=['ham','spam'],autopct="%0.2f"
)
plt.show()
# In[55]:
import nltk
# In[56]:
get_ipython().system('pip install nltk')
# In[57]:
nltk.download('punkt')
# In[58]:
df['num_characters']=df['text'].apply(len)
# In[59]:
df.head()
# In[60]:
df['num_words']=df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
# In[61]:
df.head()
# In[62]:
df['num_sentences']=df['text'].apply(lambda x:len(nltk.sent_tokenize(x)))
# In[63]:
df.head()
# In[64]:
df[['num_characters','num_words','num_sentences']].describe()
# In[65]:
df[df['target']==0][['num_characters','num_words','num_sentences']].descr
ibe()
# In[66]:
import seaborn as sns
# In[67]:
plt.figure(figsize=(12,6))
sns.histplot(df[df['target']==0]['num_characters'])
sns.histplot(df[df['target']==1]['num_characters'],color='red')
# In[68]:
plt.figure(figsize=(12,6))
sns.histplot(df[df['target']==0]['num_words'])
sns.histplot(df[df['target']==1]['num_words'],color='red')
# In[69]:
sns.pairplot(df,hue='target')
# In[70]:
sns.heatmap(df.corr(),annot=True)
# In[71]:
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import string
```

```
nltk.download('stopwords')
ps=PorterStemmer()
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:
        y.append(ps.stem(i))

    return " ".join(y)
transformed_text=transform_text("I'm gonna be home soon and i don't want
to talk about this stuff anymore tonight, k? I've cried enough today.")
print(transformed_text)
# In[72]:
df['text'][10]
# In[73]:
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
ps.stem('loving')
# In[74]:
df['transformed_text']=df['text'].apply(transform_text)
# In[75]:
df.head()
# In[76]:
from wordcloud import WordCloud
wc=WordCloud(width=500,height=500,min_font_size=10,background_color='whit
e')
# In[78]:
get_ipython().system('pip install wordcloud')
# In[79]:
from wordcloud import WordCloud
wc=WordCloud(width=500,height=500,min_font_size=10,background_color='whit
e')
# In[80]:
spam_wc =
wc.generate(df[df['target']==1]['transformed_text'].str.cat(sep=" "))
# In[81]:
plt.figure(figsize=(15,6))
plt.imshow(spam_wc)
# In[82]:
ham_wc =
wc.generate(df[df['target']==1]['transformed_text'].str.cat(sep=" "))
# In[83]:
plt.figure(figsize=(15,6))
plt.imshow(ham_wc)
# In[84]:
df.head()
# In[85]:
spam_corpus=[]
for msg in df[df['target']==1]['transformed_text'].tolist():
    for word in msg.split():
        spam_corpus.append(word)
# In[86]:
len(spam_corpus)
# In[87]:
ham_corpus=[]
for msg in df[df['target']==0]['transformed_text'].tolist():
    for word in msg.split():
        ham_corpus.append(word)
# In[88]:
len(ham_corpus)
# In[89]:
df.head()
# In[90]:
from sklearn.feature_extraction.text import
CountVectorizer,TfidfVectorizer
cv= CountVectorizer()
tfidf=TfidfVectorizer(max_features=3000)
# In[91]:
X= tfidf.fit_transform(df['transformed_text']).toarray()
# In[92]:
X.shape
# In[93]:
y=df['target'].values
# In[94]:
from sklearn.model_selection import train_test_split
# In[95]:
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_s
tate=2)
# In[96]:
from sklearn.naive_bayes import GaussianNB,MultinomialNB,BernoulliNB
from sklearn.metrics import
accuracy_score,confusion_matrix,precision_score
# In[97]:
gnb=GaussianNB()
mnb=MultinomialNB()
bnb=BernoulliNB()
# In[98]:
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))
print(precision_score(y_test,y_pred1))
# In[99]:
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))
# In[100]:
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))
# In[101]:
get_ipython().system('pip install xgboost')
# In[102]:
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
# In[103]:
svc=SVC(kernel='sigmoid',gamma=1.0)
knc=KNeighborsClassifier()
mnb= MultinomialNB()
dtc=DecisionTreeClassifier(max_depth=5)
lrc=LogisticRegression(solver='liblinear',penalty='l1')
rfc=RandomForestClassifier(n_estimators=50,random_state=2)
abc=AdaBoostClassifier(n_estimators=50,random_state=2)
bc= BaggingClassifier(n_estimators=50,random_state=2)
etc=ExtraTreesClassifier(n_estimators=50,random_state=2)
gbdt=GradientBoostingClassifier(n_estimators=50,random_state=2)
xgb=XGBClassifier(n_estimators=50,random_state=2)
# In[104]:
clfs={
    'SVC' : svc,
    'KN'  :knc,
    'NB'   :mnb,
    'DT' :dtc,
    'LR' :lrc,
    'RF'  :rfc,
    'AdaBoost':abc,
    'BgC':bc,
    'ETC':etc,
    'GBDT':gbdt,
    'xgb':xgb
}
# In[105]:
def train_classifier(clf,X_train,y_train,X_test,y_test):
    clf.fit(X_train,y_train)
    y_pred=clf.predict(X_test)
    accuracy=accuracy_score(y_test,y_pred)
    precision=precision_score(y_test,y_pred)
    return accuracy,precision
# In[106]:
train_classifier(svc,X_train,y_train,X_test,y_test)
# In[107]:
accuracy_scores=[]
precision_scores=[]
for name,clf in clfs.items():

current_accuracy,current_precision=train_classifier(clf,X_train,y_train,X
_test,y_test)

    print("For",name)
    print("Accuracy - ",current_accuracy)
    print("Precision- ",current_precision)
    accuracy_scores.append(current_accuracy)
    precision_scores.append(current_precision)
# In[108]:
performance_df=pd.DataFrame({'Algorithm':clfs.keys(),'Accuracy':accuracy_
scores,'Precision':precision_scores}).sort_values('precision',ascending=F
alse)
# In[110]:
performance_df
# In[111]:
performance_df1=pd.melt(performance_df,id_vars="Algorithm")
# In[112]:
performance_df1
# In[114]:
sns.catplot(x='Algorithm',y='value',
           hue='variable',data=performance_df1,kind='bar',height=5)
plt.ylim(0.5,1.0)
plt.xticks(rotation='vertical')
plt.show()
# In[117]:
temp_df=pd.DataFrame({'Algorithm':clfs.keys(),'Accuracy_max_ft_3000':accu
racy_scores,'Precision_max_ft_3000':precision_scores}).sort_values('Preci
sion_max_ft_3000',ascending=False)
# In[118]:
new_df=performance_df.merge(temp_df,on='Algorithm')
# In[119]:
new_df_scaled=new_df.merge(temp_df,on='Algorithm')
# In[120]:
temp_df=pd.DataFrame({'Algorithm':clfs.keys(),'Accuracy_num_chars':accura
cy_scores,'Precision_num_chars':precision_scores}).sort_values('Precision
_num_chars',ascending=False)
# In[121]:
new_df_scaled.merge(temp_df,on='Algorithm')
# In[122]:
svc=SVC(kernel='sigmoid',gamma=1.0,probability=True)
mnb=MultinomialNB()
etc=ExtraTreesClassifier(n_estimators=50,random_state=2)
from sklearn.ensemble import VotingClassifier
# In[123]:
voting=VotingClassifier(estimators=[('svm',svc),('nb',mnb),('et',etc)],vo
ting='soft')
# In[132]:
voting.fit(X_train,y_train)
# In[127]:
y_pred=voting.predict(X_test)
print("Accuracy",accuracy_score(y_test,y_pred))
print("Precision",precision_score(y_test,y_pred))
# In[137]:
estimators=[('svm',svc),('nb',mnb),('et',etc)]
final_estimator=RandomForestClassifier()
# In[138]:
from sklearn.ensemble import StackingClassifier
# In[139]:
clf=StackingClassifier(estimators=estimators,final_estimator=final_estima
tor)
# In[134]:
clf.fit(X_train,y_train)
y_pred=clf.predict(X_test)
print("Accuracy",accuracy_score(y_test,y_pred))
print("Precision",precision_score(y_test,y_pred))
# In[140]:
import pickle
pickle.dump(tfidf,open('vectorizer.pkl','wb'))
pickle.dump(mnb,open('model.pkl','wb'))
# In[142]:
import pickle
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
X_train=["Sample text 1","Sample text 2","Sample text 3"]
y_train=[0,1,0]
tfidf=TfidfVectorizer(lowercase=True,stop_words='english')
X_train_tfidf=tfidf.fit_transform(X_train)
mnb.fit(X_train_tfidf,y_train)
with open('vectorizer.pkl','wb')as vectorizer_file:
    pickle.dump(tfidf,vectorizer_file)
with open('model.pkl','wb')as model_file:
        pickle.dump(tfidf,model_file)

# In[ ]:
# In[ ]:
```