```
# Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score, f1_score, confusion_matrix
from sklearn.naive_bayes import MultinomialNB
from wordcloud import WordCloud


# Importing dataset
df = pd.read_csv("/content/spam.csv", encoding='ISO-8859-1')


# Display the first few rows of the dataset
df.head()
```

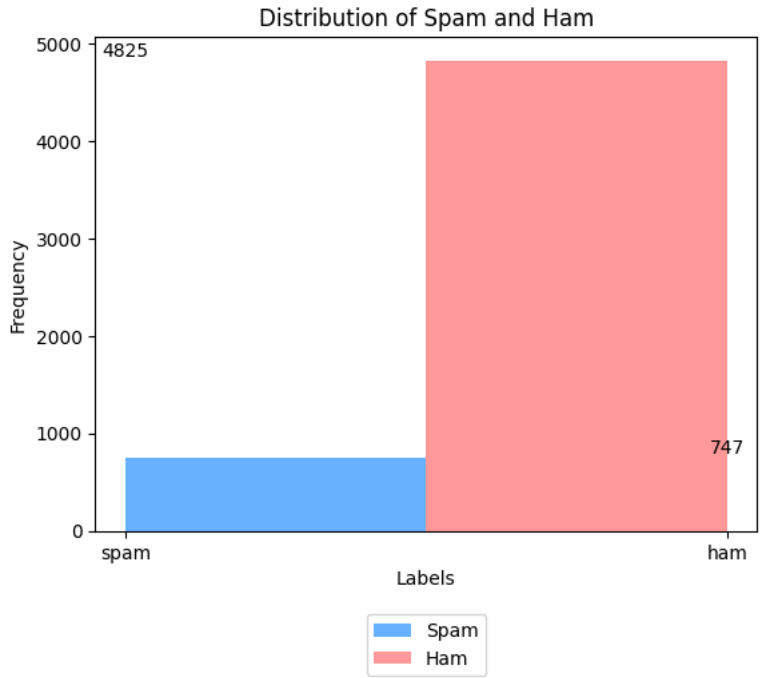
	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah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

```
# Drop the columns that do not influence the result
df = df.drop(columns=df.columns[2:5])


# Checking the result
df.head()
```

	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...

```
# Renaming the columns to make the names clearer
df.columns = ['labels', 'data']


# Visualization of spam and not spam cases

# Count the number of observations for each category
label_counts = df['labels'].value_counts()

# Define pastel colors
pastel_colors = ['#66B2FF', '#FF9999']

# Create a histogram with pastel colors and space between bars
plt.hist([df[df['labels'] == 'spam']['labels'], df[df['labels'] == 'ham']['labels']],
         bins=2, color=pastel_colors, stacked=True, label=['Spam', 'Ham'])

# Add labels and title
plt.xlabel('Labels')
plt.ylabel('Frequency')
plt.title('Distribution of Spam and Ham')

# Add text with counts on the bars
for i, count in enumerate(label_counts):
    plt.text(i, count, str(count), ha='center', va='bottom')

# Move the legend lower
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.15))
```

```
# Show the plot
plt.show()
```




```
# Binary Encoding
df['binary_labels'] = df['labels'].map({'ham': 0, 'spam': 1})
y = df['binary_labels'].to_numpy()
```

```
# Split data into training and testing sets
df_train, df_test, y_train, y_test = train_test_split(df['data'], y, test_size=0.2, random_state=42)
```

```
# Check the shapes of the data splits
df_train.shape, df_test.shape, y_train.shape, y_test.shape
```

```
((4457,), (1115,), (4457,), (1115,))
```

```
# Create a CountVectorizer for text feature extraction
featurizer = CountVectorizer(decode_error='ignore')
x_train = featurizer.fit_transform(df_train)
x_test = featurizer.transform(df_test)
```

```
# Check the result of feature extraction
x_train
```

```
<4457x7735 sparse matrix of type '<class 'numpy.int64'>'
        with 58978 stored elements in Compressed Sparse Row format>
```

```
# Create the Multinomial Naive Bayes model
model = MultinomialNB()
model.fit(x_train, y_train)
```

```
▾ MultinomialNB
MultinomialNB()
```

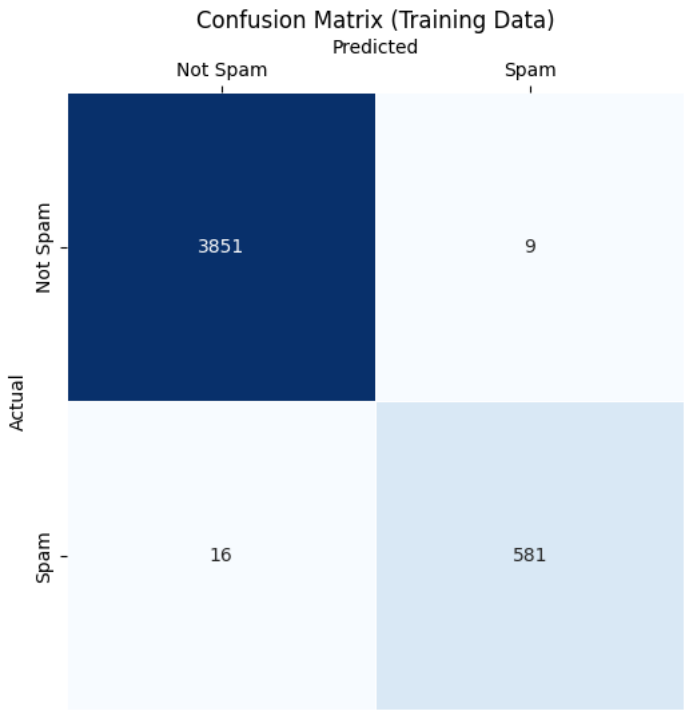
```
# Evaluate the model's performance
train_accuracy = model.score(x_train, y_train)
test_accuracy = model.score(x_test, y_test)
```

```
# Making predictions on the training and test data
Ptrain = model.predict(x_train)
Ptest = model.predict(x_test)
```

```
# Calculate F1 scores for training and test data
train_f1 = f1_score(y_train, Ptrain)
test_f1 = f1_score(y_test, Ptest)

# Calculate predicted probabilities for being in class 1 (spam) for both training and test data
Prob_train = model.predict_proba(x_train)[:, 1]
Prob_test = model.predict_proba(x_test)[:, 1]

# Calculate the AUC-ROC (Area Under the Receiver Operating Characteristic) score for both training and test data
train_auc = roc_auc_score(y_train, Prob_train)
test_auc = roc_auc_score(y_test, Prob_test)


# Create a DataFrame to display the metrics
metrics_df = pd.DataFrame({
    'Train/Test': ['Train', 'Test'],
    'Accuracy': [train_accuracy, test_accuracy],
    'F1': [train_f1, test_f1],
    'AUC': [train_auc, test_auc]
})


# Display the metrics table
print(metrics_df)
```

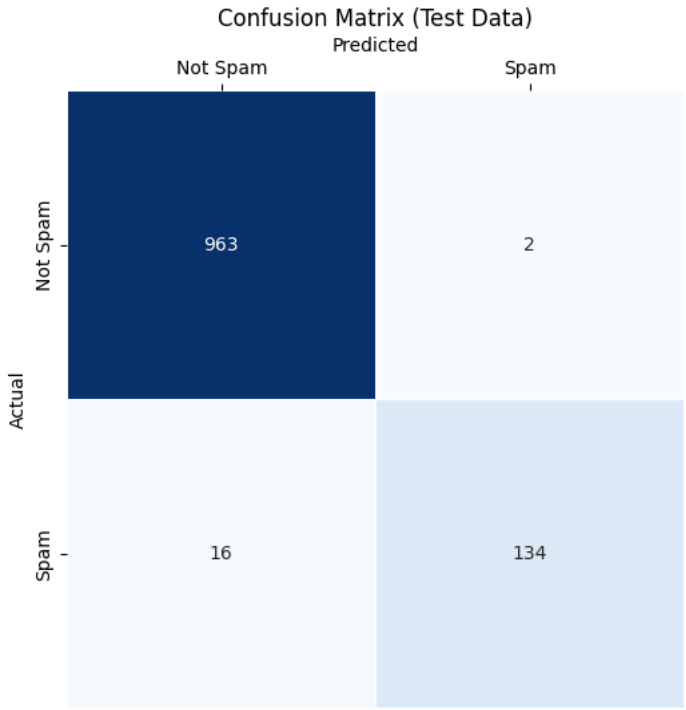
```
  Train/Test  Accuracy        F1       AUC
0      Train  0.994391  0.978939  0.994113
1       Test  0.983857  0.937063  0.975392
```

```
# Create a confusion matrix for training data
cm = confusion_matrix(y_train, Ptrain)

# Create a heatmap for the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", linewidths=.5, square=True, cbar=False,
            xticklabels=['Not Spam', 'Spam'],
            yticklabels=['Not Spam', 'Spam'])
plt.xlabel('Predicted')
plt.title('Confusion Matrix (Training Data)')
plt.gca().xaxis.tick_top()  # Put x-axis labels on top
plt.gca().xaxis.set_label_position('top')
plt.ylabel('Actual')
plt.show()
```




```
# Create a confusion matrix for test data
cm_test = confusion_matrix(y_test, Ptest)
```

```
# Create a heatmap for the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(cm_test, annot=True, fmt="d", cmap="Blues", linewidths=.5, square=True, cbar=False,
            xticklabels=['Not Spam', 'Spam'],
            yticklabels=['Not Spam', 'Spam'])
plt.xlabel('Predicted')
plt.title('Confusion Matrix (Test Data)')
plt.gca().xaxis.tick_top()  # Put x-axis labels on top
plt.gca().xaxis.set_label_position('top')
plt.ylabel('Actual')
plt.show()
```




```
# Make predictions on the entire dataset
x = featurizer.transform(df['data'])
df['predictions'] = model.predict(x)


# The messages that were spam but were treated as not spam

unident_spam = df[(df['predictions'] == 0) & (df['binary_labels'] == 1)]['data']
for msg in unident_spam:
    print(msg)
```

```
FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å
Did you hear about the new \Divorce Barbie\"? It comes with all of Ken's stuff!"
Hi I'm sue. I am 20 years old and work as a lapdancer. I love sex. Text me live - I'm i my bedroom now. text SUE to 89555. By TextOperat
Email AlertFrom: Jeri StewartSize: 2KBSubject: Low-cost prescripiton drvgsTo listen to email call 123
Do you realize that in about 40 years, we'll have thousands of old ladies running around with tattoos?
Ever thought about living a good life with a perfect partner? Just txt back NAME and AGE to join the mobile community. (100p/SMS)
Hello. We need some posh birds and chaps to user trial prods for champneys. Can i put you down? I need your address and dob asap. Ta r
Can U get 2 phone NOW? I wanna chat 2 set up meet Call me NOW on 09096102316 U can cum here 2moro Luv JANE xx Callså£1/minmoremobsEMSPOE
Hi its LUCY Hubby at meetins all day Fri & I will B alone at hotel U fancy cumin over? Pls leave msg 2day 09099726395 Lucy x Callså£1/mi
Would you like to see my XXX pics they are so hot they were nearly banned in the uk!
CALL 09090900040 & LISTEN TO EXTREME DIRTY LIVE CHAT GOING ON IN THE OFFICE RIGHT NOW TOTAL PRIVACY NO ONE KNOWS YOUR [sic] LISTENING 60
Hi ya babe x u 4goten bout me?' scammers getting smart..Though this is a regular vodafone no, if you respond you get further prem rate m
Babe: U want me dont u baby! Im nasty and have a thing 4 filthyguys. Fancy a rude time with a sexy bitch. How about we go slo n hard! Tx
Hello darling how are you today? I would love to have a chat, why dont you tell me what you look like and what you are in to sexy?
How come it takes so little time for a child who is afraid of the dark to become a teenager who wants to stay out all night?
INTERFLORA - åÒIt's not too late to order Interflora flowers for christmas call 0800 505060 to place your order before Midnight tomorrow
ROMCAPspam Everyone around should be responding well to your presence since you are so warm and outgoing. You are bringing in a real bre
Do you ever notice that when you're driving, anyone going slower than you is an idiot and everyone driving faster than you is a maniac?
LookAtMe!: Thanks for your purchase of a video clip from LookAtMe!, you've been charged 35p. Think you can do better? Why not send a vic
Sorry I missed your call let's talk when you have the time. I'm on 07090201529
LIFE has never been this much fun and great until you came in. You made it truly special for me. I won't forget you! enjoy @ one gbp/sms
Not heard from U4 a while. Call me now am here all night with just my knickers on. Make me beg for it like U did last time 01223585236 X
2/2 146tf150p
Oh my god! I've found your number again! I'm so glad, text me back xafter this msgs cst std ntwk chg å£1.50
ringtoneking 84484
In The Simpsons Movie released in July 2007 name the band that died at the start of the film? A-Green Day, B-Blue Day, C-Red Day. (Send
```

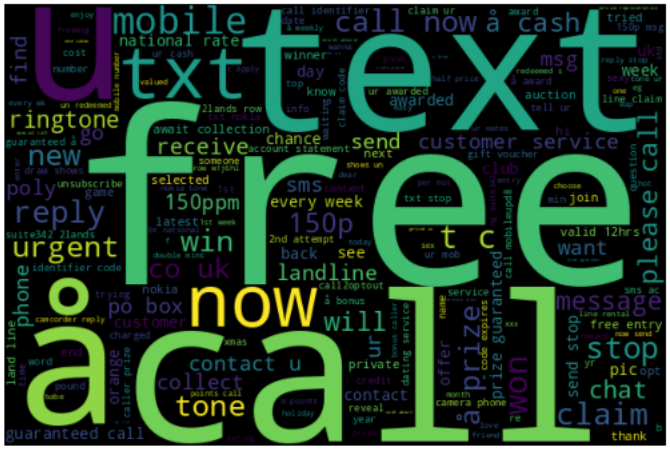
```
Missed call alert. These numbers called but left no message. 07008009200
thesmszone.com lets you send free anonymous and masked messages..im sending this message from there..do you see the potential for abuse?
Money i have won wining number 946 wot do i do next
Hi babe its Chloe, how r u? I was smashed on saturday night, it was great! How was your weekend? U been missing me? SP visionsms.com Tex
Hi this is Amy, we will be sending you a free phone number in a couple of days, which will give you an access to all the adult parties..
dating:i have had two of these. Only started after i sent a text to talk sport radio last week. Any connection do you think or coincider
```

```
# The messages that were not spam but were treated as spam

not_spam = df[(df['predictions'] == 1) & (df['binary_labels'] == 0)]['data']
for msg in not_spam:
    print(msg)
```

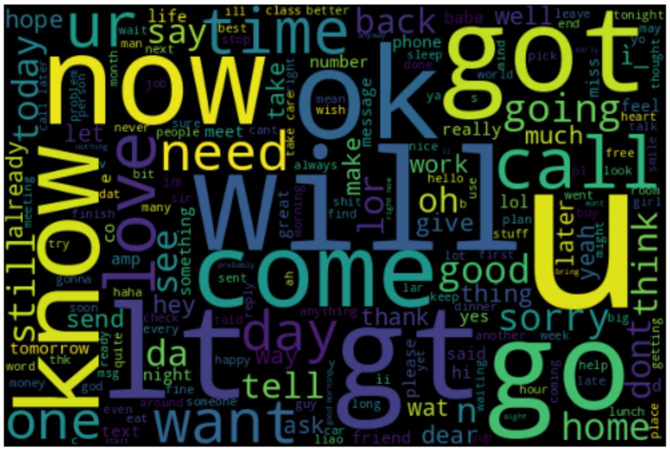
```
 Waiting for your call.
 Can u get pic msgs to your phone?
 We have sent JD for Customer Service cum Accounts Executive to ur mail id, For details contact us
 Hey...Great deal...Farm tour 9am to 5pm $95/pax, $50 deposit by 16 May
 Unlimited texts. Limited minutes.
 Mathews or tait or edwards or anderson
 Have you laid your airtel line to rest?
 I liked the new mobile
 Anytime...
 Nokia phone is lovly..
 We have sent JD for Customer Service cum Accounts Executive to ur mail id, For details contact us
```

```
# Creating a word cloud

def visualize(label):
    words = ''
    for msg in df[df['labels'] == label]['data']:
        msg = msg.lower()
        words += msg + ' '
    wordcloud = WordCloud(width=600, height=400).generate(words)
    plt.imshow(wordcloud)
    plt.axis('off')
    plt.show()
```

```
# The keywords that were more popular in spam category
visualize('spam')
```




```
# The keywords that were more popular in not spam category
visualize('ham')
```


now
ok
got
will
u
know
come
lt
gt
go
call
love
time
ur
need
going
one
want
home
day
good
sorry
think
still
already
today
back
well
say
tell
da
see
oh
work
much
take
send
hey
thank
thing
give
lor
make
later
yeah
dont
lol
hope
life
tomorrow
dear
ask
wat
friend
n