In [82]:

```python
# data Manipulation
import pandas as pd

# Mathematical operation
import numpy as np

# data Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# machine learning algorithm
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import string
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score
```

## Load and Explore the Dataset

In [83]:

```python
df = pd.read_csv(r"D:\cipherbyte internship\Spam Email Detection - spam
```

In [84]:

```python
df
```

Out[84]:

| | 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 |
| **...** | ... | ... | ... | ... | ... |
| **5567** | spam | This is the 2nd time we have tried 2 contact u... | NaN | NaN | NaN |
| **5568** | ham | Will �_ b going to esplanade fr home? | NaN | NaN | NaN |
| **5569** | ham | Pity, * was in mood for that. So...any other s... | NaN | NaN | NaN |
| **5570** | ham | The guy did some bitching but I acted like i'd... | NaN | NaN | NaN |
| **5571** | ham | Rofl. Its true to its name | NaN | NaN | NaN |

5572 rows × 5 columns

In [85]:

```
df.head()
```

Out[85]:

| | 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 |

In [86]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

In [87]:

```
df.isnull().sum()
```

Out[87]:

```
v1               0
v2               0
Unnamed: 2    5522
Unnamed: 3    5560
Unnamed: 4    5566
dtype: int64
```

In [88]:

```
# Drop unnecessary columns
df = df[['v1', 'v2']]  # Keep only the relevant columns
```

In [89]:
```
df
```

Out[89]:

| | 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 �_ 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 × 2 columns

In [90]:
```
# Rename columns for better clarity
df.columns = ['label', 'message']
```

In [91]:
```
df
```

Out[91]:

| | label | 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... |
| **...** | ... | ... |
| **5567** | spam | This is the 2nd time we have tried 2 contact u... |
| **5568** | ham | Will �_ 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 × 2 columns

In [92]:
```
# Check for missing values
df.isnull().sum()

```

Out[92]:
```
label      0
message    0
dtype: int64
```

In [93]:

```
# Check the distribution of labels
df['label'].value_counts()
```

Out[93]:

```
label
ham     4825
spam     747
Name: count, dtype: int64
```

## Preprocess the Data

**Convert the labels into binary format (spam → 1, ham → 0).**

**Tokenize and clean the text (remove punctuation, lowercase, stopwords, etc.).**

**Split the data into training and test sets.**

In [94]:

```
# Transform 'spam' to 1 and 'ham' to 0
df['label'] = encoder.fit_transform(df['label'])

# Verify the encoding
df['label'].value_counts()
```

```
C:\Users\Dell\AppData\Local\Temp\ipykernel_2308\978868311.py:2: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df['label'] = encoder.fit_transform(df['label'])
```

Out[94]:

```
label
0    4825
1     747
Name: count, dtype: int64
```

In [95]:

```
# Function to clean text
def clean_text(text):
    text = text.lower()
    text = ''.join([char for char in text if char not in string.punctua
    return text

df['message'] = df['message'].apply(clean_text)
```

```
C:\Users\Dell\AppData\Local\Temp\ipykernel_2308\4138295138.py:7: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df['message'] = df['message'].apply(clean_text)
```

In [96]:

```python
df
```

Out[96]:

| | label | message |
|---|---|---|
| **0** | 0 | go until jurong point crazy available only in ... |
| **1** | 0 | ok lar joking wif u oni |
| **2** | 1 | free entry in 2 a wkly comp to win fa cup fina... |
| **3** | 0 | u dun say so early hor u c already then say |
| **4** | 0 | nah i dont think he goes to usf he lives aroun... |
| **...** | ... | ... |
| **5567** | 1 | this is the 2nd time we have tried 2 contact u... |
| **5568** | 0 | will � b going to esplanade fr home |
| **5569** | 0 | pity was in mood for that soany other suggest... |
| **5570** | 0 | the guy did some bitching but i acted like id ... |
| **5571** | 0 | rofl its true to its name |

5572 rows × 2 columns

In [97]:

```python
# Split data into training and test sets (80/20 split)
X_train, X_test, y_train, y_test = train_test_split(df['message'], df[
```

## Convert Text to Numerical Features

In [98]:

```python
# Initialize TF-IDF Vectorizer
vectorizer = TfidfVectorizer(max_features=3000)

# Fit and transform the text data
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

## Train a Machine Learning Model

In [99]:

```python
# Initialize and train the model
model = LogisticRegression()
model.fit(X_train_tfidf, y_train)
```

Out[99]:

```
LogisticRegression()
```

**In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.**
**On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.**

In [101]:

```python
# Make predictions
y_pred = model.predict(X_test_tfidf)
y_pred
```

Out[101]:

```
array([0, 0, 0, ..., 0, 0, 0])
```

## Evaluate the Model

In [102]:

```
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pre
```

```
Accuracy: 0.967713004484305

Classification Report:
               precision    recall  f1-score   support

           0       0.96      1.00      0.98       965
           1       1.00      0.76      0.86       150

    accuracy                           0.97      1115
   macro avg       0.98      0.88      0.92      1115
weighted avg       0.97      0.97      0.97      1115
```