import numpy as np

import pandas as pd

import os

for dirname, \_, filenames **in** os.walk('/kaggle/input'):

for filename **in** filenames:

print(os.path.join(dirname, filename))

In [2]:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.utils import resample

import seaborn as sb

In [3]:

df=pd.read\_csv(r"/content/spam.csv",encoding="latin1")

df

Out[3]:

|  | 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 Ì\_ 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

In [4]:

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 [5]:

df.isnull().sum()

Out[5]:

v1 0

v2 0

Unnamed: 2 5522

Unnamed: 3 5560

Unnamed: 4 5566

dtype: int64

In [6]:

x=df.v2

x

Out[6]:

0 Go until jurong point, crazy.. Available only ...

1 Ok lar... Joking wif u oni...

2 Free entry in 2 a wkly comp to win FA Cup fina...

3 U dun say so early hor... U c already then say...

4 Nah I don't think he goes to usf, he lives aro...

...

5567 This is the 2nd time we have tried 2 contact u...

5568 Will Ì\_ b going to esplanade fr home?

5569 Pity, \* was in mood for that. So...any other s...

5570 The guy did some bitching but I acted like i'd...

5571 Rofl. Its true to its name

Name: v2, Length: 5572, dtype: object

In [7]:

y=df.v1

y

Out[7]:

0 ham

1 ham

2 spam

3 ham

4 ham

...

5567 spam

5568 ham

5569 ham

5570 ham

5571 ham

Name: v1, Length: 5572, dtype: object

In [8]:

y.replace(to\_replace='ham',value=1,inplace=True)

y.replace(to\_replace='spam',value=0,inplace=True)

y

Out[8]:

0 1

1 1

2 0

3 1

4 1

..

5567 0

5568 1

5569 1

5570 1

5571 1

Name: v1, Length: 5572, dtype: int64

In [9]:

y=df.v1

y.value\_counts()

Out[9]:

1 4825

0 747

Name: v1, dtype: int64

In [10]:

sb.histplot(y)

Out[10]:

<AxesSubplot:xlabel='v1', ylabel='Count'>

In [11]:

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,train\_size=0.9,random\_state=25)

In [12]:

gok= pd.concat([x\_train, y\_train], axis=1)

spam = gok[gok.v1==0]

ham = gok[gok.v1==1]