In [1]:
```
import numpy as np
import pandas as pd
import nltk
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import RandomizedSearchCV
from sklearn import metrics
from nltk import word_tokenize, FreqDist
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import re
from nltk.stem.porter import PorterStemmer
from sklearn import svm
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score
```

In [2]:
```
data = pd.read_csv(r"C:\Users\adity\Downloads\titanic\train.csv")
data.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Er
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	

In [3]:
```
data.columns
```

Out[3]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [4]:
```
data.isnull().sum()
```

Out[4]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [5]:
```
data.drop(['PassengerId','Name','Ticket'],axis=1,inplace=True)
```

In [6]:
```
data
```

Out[6]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	male	22.0	1	0	7.2500	NaN	S
1	1	1	female	38.0	1	0	71.2833	C85	C
2	1	3	female	26.0	0	0	7.9250	NaN	S
3	1	1	female	35.0	1	0	53.1000	C123	S
4	0	3	male	35.0	0	0	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	NaN	S
887	1	1	female	19.0	0	0	30.0000	B42	S
888	0	3	female	NaN	1	2	23.4500	NaN	S
889	1	1	male	26.0	0	0	30.0000	C148	C
890	0	3	male	32.0	0	0	7.7500	NaN	Q

891 rows × 9 columns

In [7]:
```
data['Cabin'].ffill(inplace=True)
data['Cabin'].replace(np.nan,"Z10",inplace=True)
data.isnull().sum()
```

Out[7]:
```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Cabin         0
Embarked      2
dtype: int64
```

In [8]:
```
data['Age'] = data["Age"].fillna(data.Age.mean())
data["Embarked"] = data["Embarked"].fillna("C")
```

In [9]:
```
data.head()
```

Out[9]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	male	22.0	1	0	7.2500	Z10	S
1	1	1	female	38.0	1	0	71.2833	C85	C
2	1	3	female	26.0	0	0	7.9250	C85	S
3	1	1	female	35.0	1	0	53.1000	C123	S
4	0	3	male	35.0	0	0	8.0500	C123	S

In [10]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Survived  891 non-null    int64  
 1   Pclass    891 non-null    int64  
 2   Sex       891 non-null    object 
 3   Age       891 non-null    float64
 4   SibSp     891 non-null    int64  
 5   Parch     891 non-null    int64  
 6   Fare      891 non-null    float64
 7   Cabin     891 non-null    object 
 8   Embarked  891 non-null    object 
dtypes: float64(2), int64(4), object(3)
memory usage: 62.8+ KB
```

In [11]:
```
data.isnull().sum()
```

Out[11]:
```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Cabin       0
Embarked    0
dtype: int64
```

In [12]:
```
data.Sex.nunique()
```

Out[12]:
```
2
```

In [13]:
```
data.Cabin.nunique()
```

Out[13]:
```
148
```

In [14]:
```
data.Embarked.nunique()
```

Out[14]:
```
3
```

In [15]:
```
data.Sex = data.Sex.replace(data.Sex.unique(),[0,1])
```

In [16]:
```
data.Embarked = data.Embarked.replace(data.Embarked.unique(),[0,1,2])
```

In [17]:
```
data.head
```

Out[17]:
```
<bound method NDFrame.head of      Survived  Pclass  Sex        Age  SibSp  Parch
Fare Cabin  Embarked
0           0       3    0  22.000000      1      0   7.2500   Z10         0
1           1       1    1  38.000000      1      0  71.2833   C85         1
2           1       3    1  26.000000      0      0   7.9250   C85         0
3           1       1    1  35.000000      1      0  53.1000  C123         0
4           0       3    0  35.000000      0      0   8.0500  C123         0
..        ...     ...  ...        ...    ...    ...      ...   ...       ...
886         0       2    0  27.000000      0      0  13.0000   C50         0
887         1       1    1  19.000000      0      0  30.0000   B42         0
888         0       3    1  29.699118      1      2  23.4500   B42         0
889         1       1    0  26.000000      0      0  30.0000  C148         1
890         0       3    0  32.000000      0      0   7.7500  C148         2

[891 rows x 9 columns]>
```

In [18]:
```
data
```

Out[18]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	0	22.000000	1	0	7.2500	Z10	0
1	1	1	1	38.000000	1	0	71.2833	C85	1
2	1	3	1	26.000000	0	0	7.9250	C85	0
3	1	1	1	35.000000	1	0	53.1000	C123	0
4	0	3	0	35.000000	0	0	8.0500	C123	0
...	...	...	...	...	...	...	...	...	...
886	0	2	0	27.000000	0	0	13.0000	C50	0
887	1	1	1	19.000000	0	0	30.0000	B42	0
888	0	3	1	29.699118	1	2	23.4500	B42	0
889	1	1	0	26.000000	0	0	30.0000	C148	1
890	0	3	0	32.000000	0	0	7.7500	C148	2

891 rows × 9 columns

In [19]:
```
data.Cabin,_ = pd.factorize(data.Cabin)
```

In [20]:
```
data.head()
```

Out[20]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	0	22.0	1	0	7.2500	0	0
1	1	1	1	38.0	1	0	71.2833	1	1
2	1	3	1	26.0	0	0	7.9250	1	0
3	1	1	1	35.0	1	0	53.1000	2	0
4	0	3	0	35.0	0	0	8.0500	2	0

In [21]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Survived  891 non-null    int64  
 1   Pclass    891 non-null    int64  
 2   Sex       891 non-null    int64  
 3   Age       891 non-null    float64
 4   SibSp     891 non-null    int64  
 5   Parch     891 non-null    int64  
 6   Fare      891 non-null    float64
 7   Cabin     891 non-null    int64  
 8   Embarked  891 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 62.8 KB
```

In [22]:
```
train_y = data['Survived']
data.drop(['Survived'],axis=1,inplace=True)
```

In [23]:
```
data.head()
```

Out[23]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	3	0	22.0	1	0	7.2500	0	0
1	1	1	38.0	1	0	71.2833	1	1
2	3	1	26.0	0	0	7.9250	1	0
3	1	1	35.0	1	0	53.1000	2	0
4	3	0	35.0	0	0	8.0500	2	0

In [24]:
```
train_x = data
```

In [25]:
```
data_test = pd.read_csv(r"C:\Users\adity\Downloads\titanic\test.csv")
data_test
```

Out[25]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	
...	...	...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	

418 rows × 11 columns

In [26]:
```
passenger = data_test['PassengerId']
data_test.drop(['PassengerId'],axis=1,inplace=True)
data_test.drop(['Name','Ticket'],axis=1,inplace=True)
passenger
```

Out[26]:
```
0       892
1       893
2       894
3       895
4       896
       ...
413    1305
414    1306
415    1307
416    1308
417    1309
Name: PassengerId, Length: 418, dtype: int64
```

In [27]:
```
data_test['Cabin'].ffill(inplace=True)
data_test['Cabin'].replace(np.nan,"Z10",inplace=True)
data_test.isnull().sum()
```

Out[27]:
```
Pclass       0
Sex          0
Age         86
SibSp        0
Parch        0
Fare         1
Cabin        0
Embarked     0
dtype: int64
```

In [28]:
```
data_test['Fare'].replace(np.nan,data_test['Fare'].mean(),inplace=True)
data_test['Age'].replace(np.nan,data_test['Age'].mean(),inplace=True)
```

In [29]:
```
data_test.isnull().sum()
```

Out[29]:
```
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Cabin       0
Embarked    0
dtype: int64
```

In [30]:
```
data_test.Cabin,_ = pd.factorize(data_test.Cabin)
```

In [31]:
```
data_test
```

Out[31]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	3	male	34.50000	0	0	7.8292	0	Q
1	3	female	47.00000	1	0	7.0000	0	S
2	2	male	62.00000	0	0	9.6875	0	Q
3	3	male	27.00000	0	0	8.6625	0	S
4	3	female	22.00000	1	1	12.2875	0	S
...	...	...	...	...	...	...	...	...
413	3	male	30.27259	0	0	8.0500	6	S
414	1	female	39.00000	0	0	108.9000	76	C
415	3	male	38.50000	0	0	7.2500	76	S
416	3	male	30.27259	0	0	8.0500	76	S
417	3	male	30.27259	1	1	22.3583	76	C

418 rows × 8 columns

In [32]:
```
x1_test = data_test
```

In [33]:
```
data_test.Sex = data_test.Sex.replace(data_test.Sex.unique(),[0,1])
data_test.Embarked = data_test.Embarked.replace(data_test.Embarked.unique(),[0,1,2])
```

In [34]:
```
data_test
```

Out[34]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	3	0	34.50000	0	0	7.8292	0	0
1	3	1	47.00000	1	0	7.0000	0	1
2	2	0	62.00000	0	0	9.6875	0	0
3	3	0	27.00000	0	0	8.6625	0	1
4	3	1	22.00000	1	1	12.2875	0	1
...	...	...	...	...	...	...	...	...
413	3	0	30.27259	0	0	8.0500	6	1
414	1	1	39.00000	0	0	108.9000	76	2
415	3	0	38.50000	0	0	7.2500	76	1
416	3	0	30.27259	0	0	8.0500	76	1
417	3	0	30.27259	1	1	22.3583	76	2

418 rows × 8 columns

In [35]:
```
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score
```

In [36]:
```
model = XGBClassifier(learning_rate = 0.05, gamma = 0.05, n_estimators = 120,random_st
model.fit(train_x,train_y)
```

Out[36]:
```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=N
one,
              enable_categorical=False, eval_metric=None, feature_types=N
one,
              gamma=0.05, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.05, max_bin=N
one,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
```

In [37]:
```
x1_test=data_test
y1_test=model.predict(x1_test)
final = pd.DataFrame(data=(passenger),columns=['PassengerId'])
final.loc[:,"Survived"]=y1_test
final.head(20)
```

Out[37]:

	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	1
5	897	0
6	898	0
7	899	0
8	900	1
9	901	0
10	902	0
11	903	0
12	904	1
13	905	0
14	906	1
15	907	1
16	908	0
17	909	0
18	910	1
19	911	1

In [38]:
```
final.to_csv('gender_submission',index=False)
sub=pd.read_csv('./gender_submission')
sub.head(25)
```

Out[38]:

	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	1
5	897	0
6	898	0
7	899	0
8	900	1
9	901	0
10	902	0
11	903	0
12	904	1
13	905	0
14	906	1
15	907	1
16	908	0
17	909	0
18	910	1
19	911	1
20	912	0
21	913	0
22	914	1
23	915	0
24	916	1

In [ ]: