Import Libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt, pydotplus
%matplotlib inline

import math
```

In [2]:
```
import sklearn
from sklearn import preprocessing
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.linear_model import LogisticRegression

from pylab import rcParams
```

Load Dataset

In [3]:
```
url = "https://raw.githubusercontent.com/BigDataGal/Python-for-Data-Science/master/titanic-train.csv"
titanic = pd.read_csv(url)
```

In [4]:
```
titanic.head(10)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

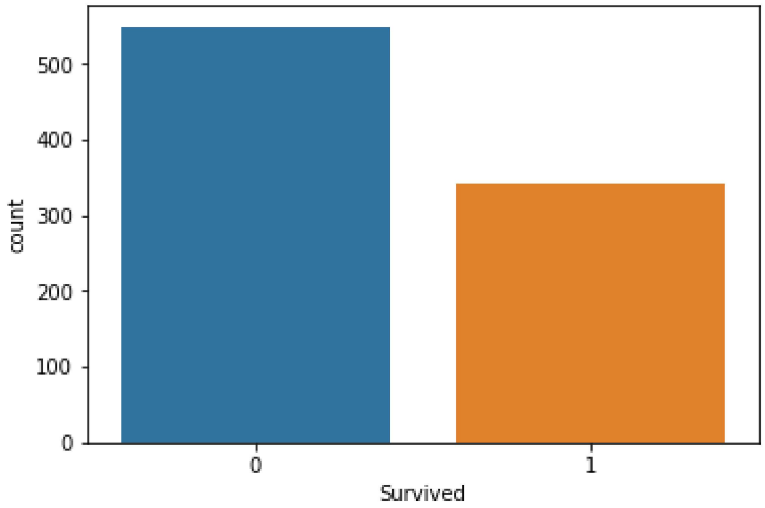
In [5]:
```
print("# of passesngers in original data: " + str(titanic.shape))
```

```
# of passesngers in original data: (891, 12)
```

Analysing Data

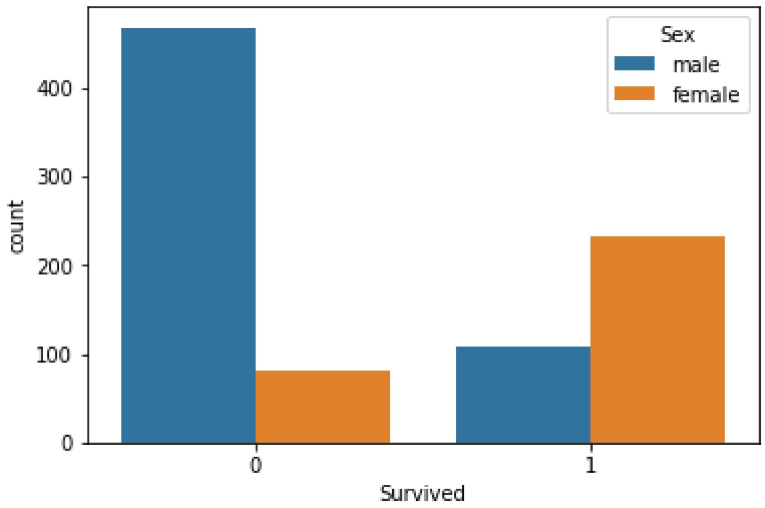
In [6]:
```
sns.countplot(x="Survived", data = titanic)
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f66574048>
```



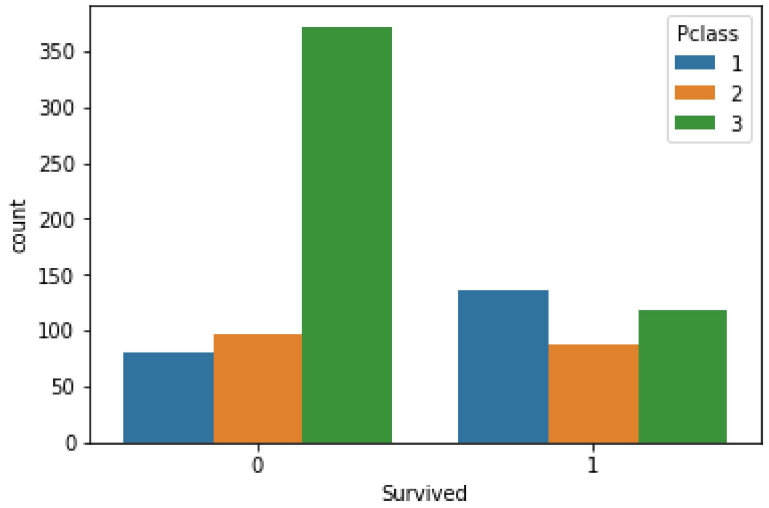
In [7]:
```
sns.countplot(x="Survived", hue="Sex" , data = titanic)
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f685f1470>
```



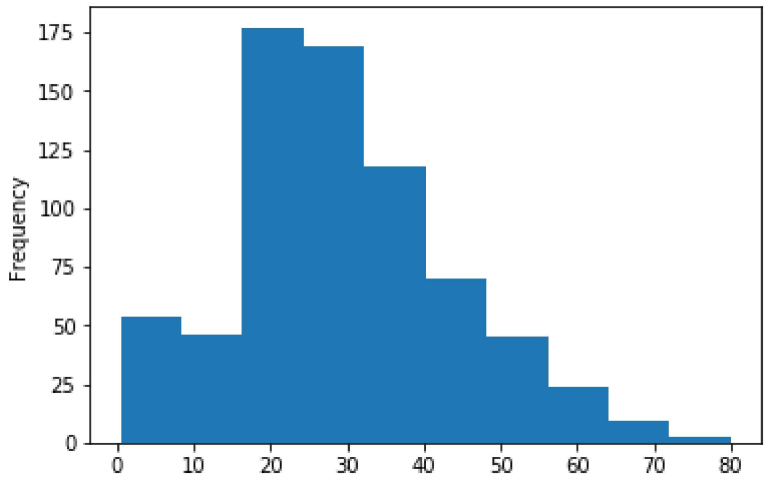
In [8]:
```
sns.countplot(x="Survived", hue="Pclass" , data = titanic)
```

Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f68628898>
```



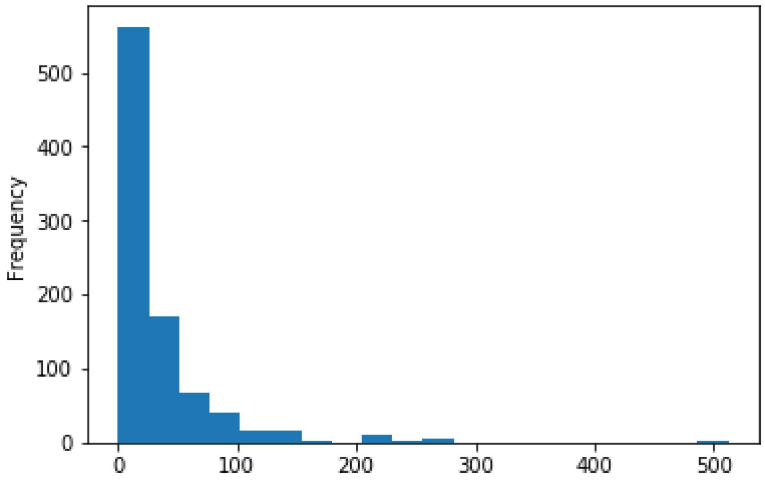
In [9]:
```
titanic["Age"].plot.hist()
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f685ddd30>
```



In [10]:
```
titanic["Fare"].plot.hist(bins=20)
```

Out[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f68669be0>
```

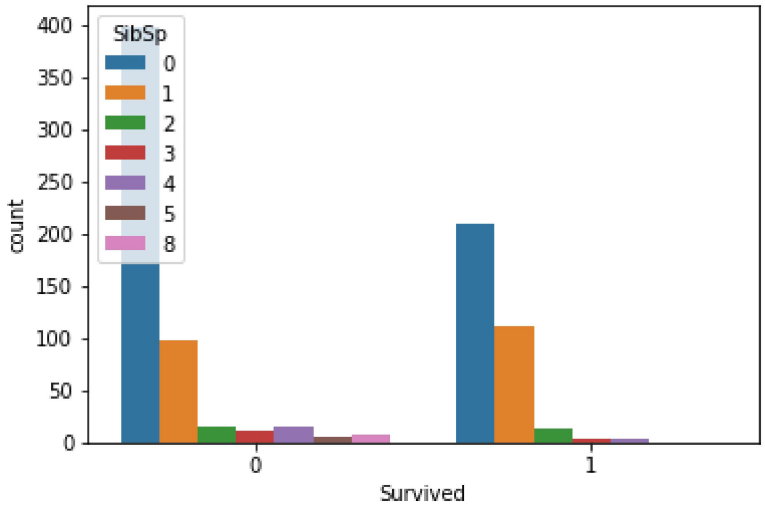


In [11]:
```
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

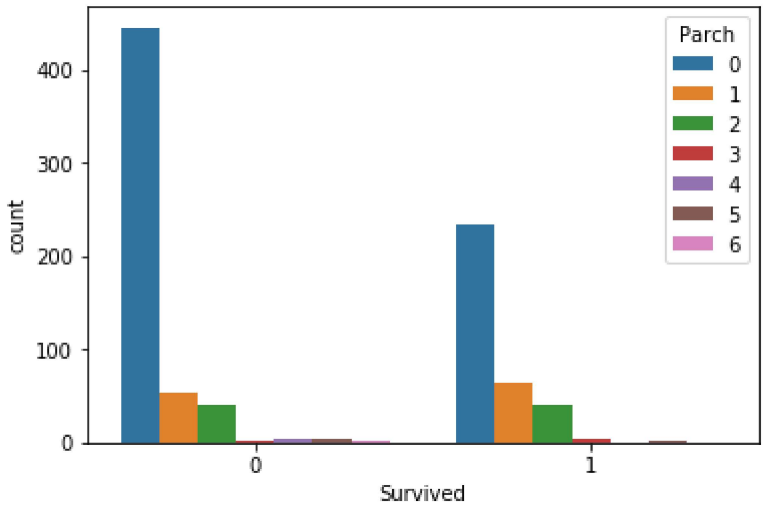
In [12]:
```
sns.countplot(x="Survived", hue="SibSp" , data = titanic)
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f687236a0>
```



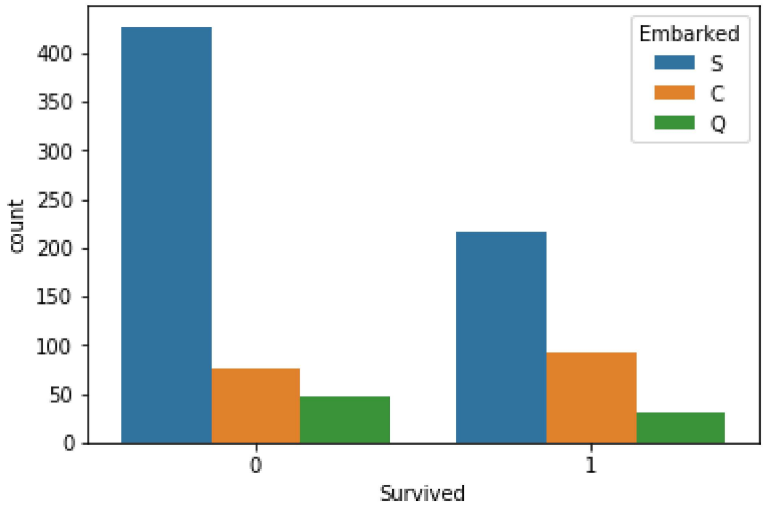
In [13]:
```
sns.countplot(x="Survived", hue="Parch" , data = titanic)
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f6866efd0>
```



In [14]:
```
sns.countplot(x="Survived", hue="Embarked" , data = titanic)
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f687cc0b8>
```



Data Wrangling.

Clean the data by removing the nan values and unnessary columns in dataset.

In [15]:
```
titanic.isnull()
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
5	False	False	False	False	False	True	False	False	False	False	True	False
6	False	False	False	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False	False	True	False
8	False	False	False	False	False	False	False	False	False	False	True	False
9	False	False	False	False	False	False	False	False	False	False	True	False
10	False	False	False	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False	False	False	False	True	False
13	False	False	False	False	False	False	False	False	False	False	True	False
14	False	False	False	False	False	False	False	False	False	False	True	False
15	False	False	False	False	False	False	False	False	False	False	True	False
16	False	False	False	False	False	False	False	False	False	False	True	False
17	False	False	False	False	False	True	False	False	False	False	True	False
18	False	False	False	False	False	False	False	False	False	False	True	False
19	False	False	False	False	False	True	False	False	False	False	True	False
20	False	False	False	False	False	False	False	False	False	False	True	False
21	False	False	False	False	False	False	False	False	False	False	False	False
22	False	False	False	False	False	False	False	False	False	False	True	False
23	False	False	False	False	False	False	False	False	False	False	False	False
24	False	False	False	False	False	False	False	False	False	False	True	False
25	False	False	False	False	False	False	False	False	False	False	True	False

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
26	False	False	False	False	False	True	False	False	False	False	True	False
27	False	False	False	False	False	False	False	False	False	False	False	False
28	False	False	False	False	False	True	False	False	False	False	True	False
29	False	False	False	False	False	True	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
861	False	False	False	False	False	False	False	False	False	False	True	False
862	False	False	False	False	False	False	False	False	False	False	False	False
863	False	False	False	False	False	True	False	False	False	False	True	False
864	False	False	False	False	False	False	False	False	False	False	True	False
865	False	False	False	False	False	False	False	False	False	False	True	False
866	False	False	False	False	False	False	False	False	False	False	True	False
867	False	False	False	False	False	False	False	False	False	False	False	False
868	False	False	False	False	False	True	False	False	False	False	True	False
869	False	False	False	False	False	False	False	False	False	False	True	False
870	False	False	False	False	False	False	False	False	False	False	True	False
871	False	False	False	False	False	False	False	False	False	False	False	False
872	False	False	False	False	False	False	False	False	False	False	False	False
873	False	False	False	False	False	False	False	False	False	False	True	False
874	False	False	False	False	False	False	False	False	False	False	True	False
875	False	False	False	False	False	False	False	False	False	False	True	False
876	False	False	False	False	False	False	False	False	False	False	True	False
877	False	False	False	False	False	False	False	False	False	False	True	False
878	False	False	False	False	False	True	False	False	False	False	True	False
879	False	False	False	False	False	False	False	False	False	False	False	False
880	False	False	False	False	False	False	False	False	False	False	True	False
881	False	False	False	False	False	False	False	False	False	False	True	False
882	False	False	False	False	False	False	False	False	False	False	True	False

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
883	False	False	False	False	False	False	False	False	False	False	True	False
884	False	False	False	False	False	False	False	False	False	False	True	False
885	False	False	False	False	False	False	False	False	False	False	True	False
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

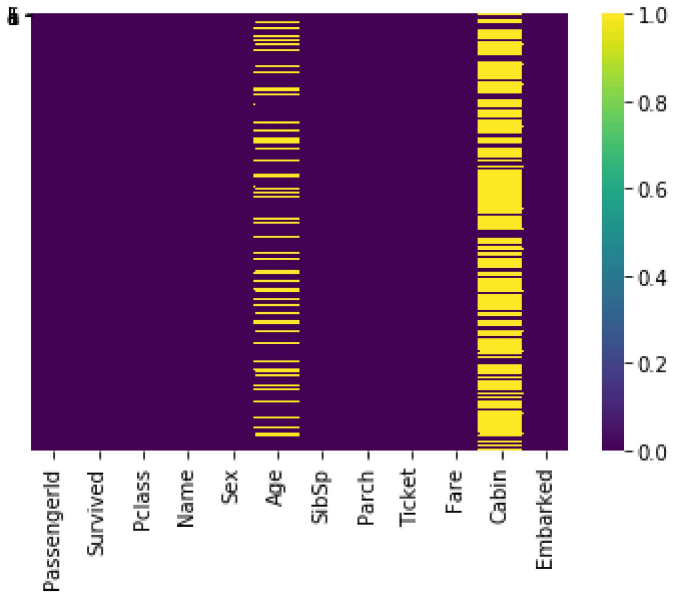
891 rows × 12 columns

In [16]:
```
titanic.isnull().sum()
```

Out[16]:
```
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
```

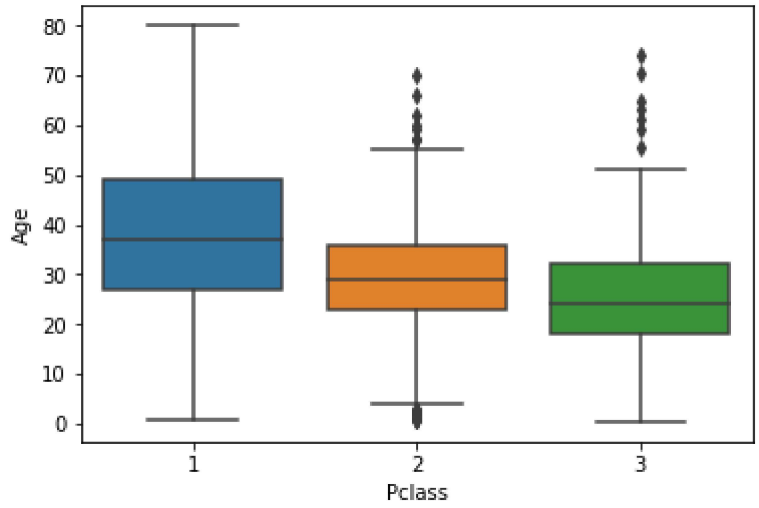
In [17]:
```
sns.heatmap(titanic.isnull(), yticklabels="False", cmap ="viridis")
```

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x16f68824780>



In [18]:
```
sns.boxplot(x="Pclass", y="Age", data=titanic)
```

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x16f68908ef0>



In [19]:
```
titanic.drop("Cabin", axis=1, inplace=True)
```

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

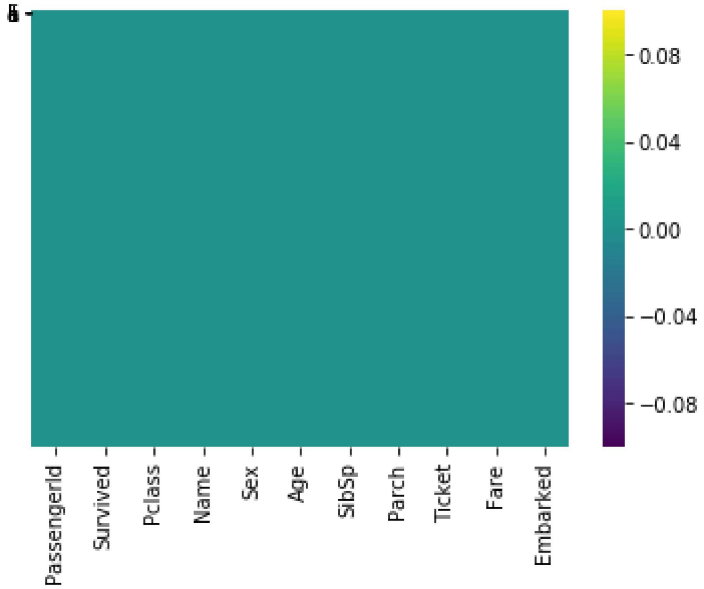
Out[20]:

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

In [21]:
```
titanic.dropna(inplace=True)
```

In [22]:
```
sns.heatmap(titanic.isnull(), yticklabels="False", cmap ="viridis")
```

Out[22]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x16f68964e80>
```



In [23]:
```
titanic.isnull().sum()
```

Out[23]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [24]:
```
sex = pd.get_dummies(titanic['Sex'], drop_first=True)
```

In [25]:
```
sex
```

Out[25]:

	male
0	1
1	0
2	0
3	0
4	1
6	1
7	1
8	0
9	0
10	0
11	0
12	1
13	1
14	0
15	0
16	1
18	0
20	1
21	1
22	0
23	1
24	0
25	0
27	1
30	1
33	1

	male
34	1
35	1
37	1
38	0
...	...
856	0
857	1
858	0
860	1
861	1
862	0
864	1
865	0
866	0
867	1
869	1
870	1
871	0
872	1
873	1
874	0
875	0
876	1
877	1
879	0
880	0
881	1

	male
882	0
883	1
884	1
885	0
886	1
887	0
889	1
890	1

712 rows × 1 columns

In [26]:
```
embark = pd.get_dummies(titanic['Embarked'], drop_first=True)
```

In [27]:
```
embark
```

Out[27]:

	Q	S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1
6	0	1
7	0	1
8	0	1
9	0	0
10	0	1
11	0	1
12	0	1
13	0	1
14	0	1
15	0	1
16	1	0
18	0	1
20	0	1
21	0	1
22	1	0
23	0	1
24	0	1
25	0	1
27	0	1
30	0	0
33	0	1

	Q	S
34	0	0
35	0	1
37	0	1
38	0	1
...	...	...
856	0	1
857	0	1
858	0	0
860	0	1
861	0	1
862	0	1
864	0	1
865	0	1
866	0	0
867	0	1
869	0	1
870	0	1
871	0	1
872	0	1
873	0	1
874	0	0
875	0	0
876	0	1
877	0	1
879	0	0
880	0	1
881	0	1

	Q	S
882	0	1
883	0	1
884	0	1
885	1	0
886	0	1
887	0	1
889	0	0
890	1	0

712 rows × 2 columns

In [28]:
```
pcl = pd.get_dummies(titanic['Pclass'], drop_first=True)
```

In [29]:
```
titanic =pd.concat([titanic, sex, embark, pcl], axis=1)
```

In [30]:
```
titanic.head()
```

Out[30]:

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

In [31]:
```
titanic.drop(['Sex','Embarked','PassengerId', 'Name', 'Pclass', 'Ticket' ], axis =1, inplace=True)
```

In [32]:
```
titanic.head()
```

Out[32]:

	Survived	Age	SibSp	Parch	Fare	male	Q	S	2	3
0	0	22.0	1	0	7.2500	1	0	1	0	1
1	1	38.0	1	0	71.2833	0	0	0	0	0
2	1	26.0	0	0	7.9250	0	0	1	0	1
3	1	35.0	1	0	53.1000	0	0	1	0	0
4	0	35.0	0	0	8.0500	1	0	1	0	1

Train & Test Data

In [33]:
```
X = titanic.drop("Survived", axis=1)
y = titanic["Survived"]
```

In [34]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

LogisticRegression

In [35]:
```
logmodel = LogisticRegression()
```

In [36]:
```
logmodel.fit(X_train, y_train)
```

Out[36]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

In [37]:
```
predictions = logmodel.predict(X_test)
```

In [38]:
```
classification_report(y_test, predictions)
```

Out[38]:
```
'             precision    recall  f1-score   support\n\n          0       0.81      0.83      0.82       126
\n          1       0.75      0.72      0.73        88\n\navg / total       0.78      0.79      0.78       21
4\n'
```

In [39]:
```
from sklearn.metrics import  confusion_matrix
```

In [40]:
```
confusion_matrix(y_test, predictions)
```

Out[40]:
```
array([[105,  21],
       [ 25,  63]], dtype=int64)
```

Accuracy

In [41]:
```
from sklearn.metrics import  accuracy_score
```

In [42]:
```
accuracy_score(y_test, predictions)*100
```

Out[42]:
```
78.50467289719626
```

Decision Tree Classification

In [43]:
```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(criterion='entropy', max_depth=3, random_state=0)
dtree.fit(X_train, y_train)
```

Out[43]:
```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=3,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

In [44]:
```
# use the model to make predictions with the test data
y_pred = dtree.predict(X_test)
```

In [45]:
```
# how did the model perform?
count_misclassified = (y_test != y_pred).sum()
print('Misclassified samples: {}'.format(count_misclassified))

accuracy = metrics.accuracy_score(y_test, y_pred)
print('Accuracy: {:.2f}'.format(accuracy))
```

```
Misclassified samples: 44
Accuracy: 0.79
```

In [46]:
```
from sklearn.cross_validation import cross_val_score

scores = cross_val_score(estimator = dtree,       # Model to test
                                 X = X,
                                 y = y,           # Target variable
                           scoring = "accuracy",  # Scoring metric
                                cv = 10)          # Cross validation folds
print("Accuracy per fold: ")
print(scores)
print("Average accuracy: ", scores.mean())
```

```
Accuracy per fold:
[0.81944444 0.80555556 0.80555556 0.86111111 0.77464789 0.78873239
 0.8028169  0.76056338 0.85714286 0.87142857]
Average accuracy:  0.8146998658618377

C:\Users\Sreekanth\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWarning: This modul
e was deprecated in version 0.18 in favor of the model_selection module into which all the refactored classes
and functions are moved. Also note that the interface of the new CV iterators are different from that of this
module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```