Titanic project

The objective of this project is develop a predicive model that classifies passengers on the Titanic as either survivors or non-survivors based on various features.

Importing necessary libraries

In [62]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
```

Import data

In [63]:

```python
train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
```

Part 1: Data Understanding

In [64]:

```python
train.head(2)
```

Out[64]:

	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

In [65]:

```python
train.columns
```

Out[65]:

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

In [66]:

```python
train.dtypes
```

Out[66]:

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
```

```
Embarked          object
dtype: object
```

Part 2: Data Cleaning

Droping columns that clearly doesn't give any useful information

In [67]:

```
train = train.drop(columns = ["Name", "Ticket", "Cabin", "Embarked","PassengerId"])
```

Getting rid of NaN values in dataset

In [68]:

```
train.isna().sum()
```

Out[68]:

```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
dtype: int64
```

In [69]:

```
train.shape
```

Out[69]:

```
(891, 7)
```

In [70]:

```
train.dropna(inplace = True)
```

Preparation for ploting relationship graph

In [71]:

```
train.Sex = train.Sex.replace({"female" : 1 , "male" : 0})
```

In [72]:

```
train.head(5)
```

Out[72]:

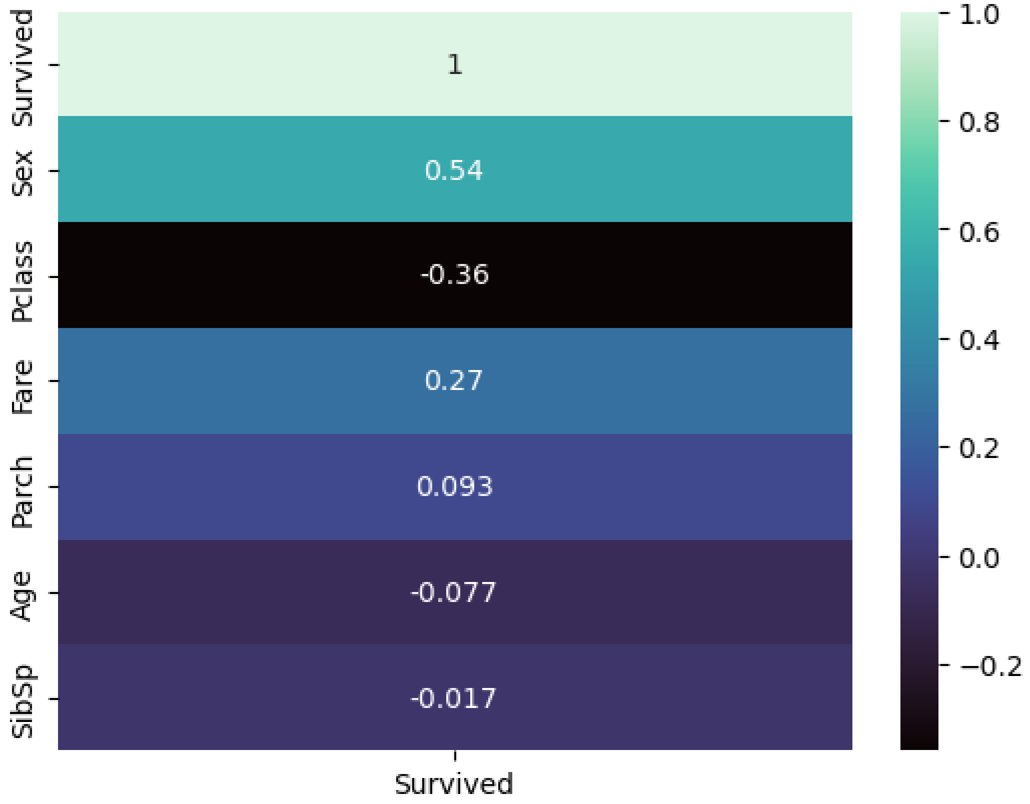
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
0	0	3	0	22.0	1	0	7.2500
1	1	1	1	38.0	1	0	71.2833
2	1	3	1	26.0	0	0	7.9250
3	1	1	1	35.0	1	0	53.1000
4	0	3	0	35.0	0	0	8.0500

Part 3: Data Visualization

Plotting relationship graph

In [73]:

```
train_corr = train.corr()
fig = sns.heatmap(train.corr()[["Survived"]].sort_values(by = ["Survived"], ascending =
plt.show()
```

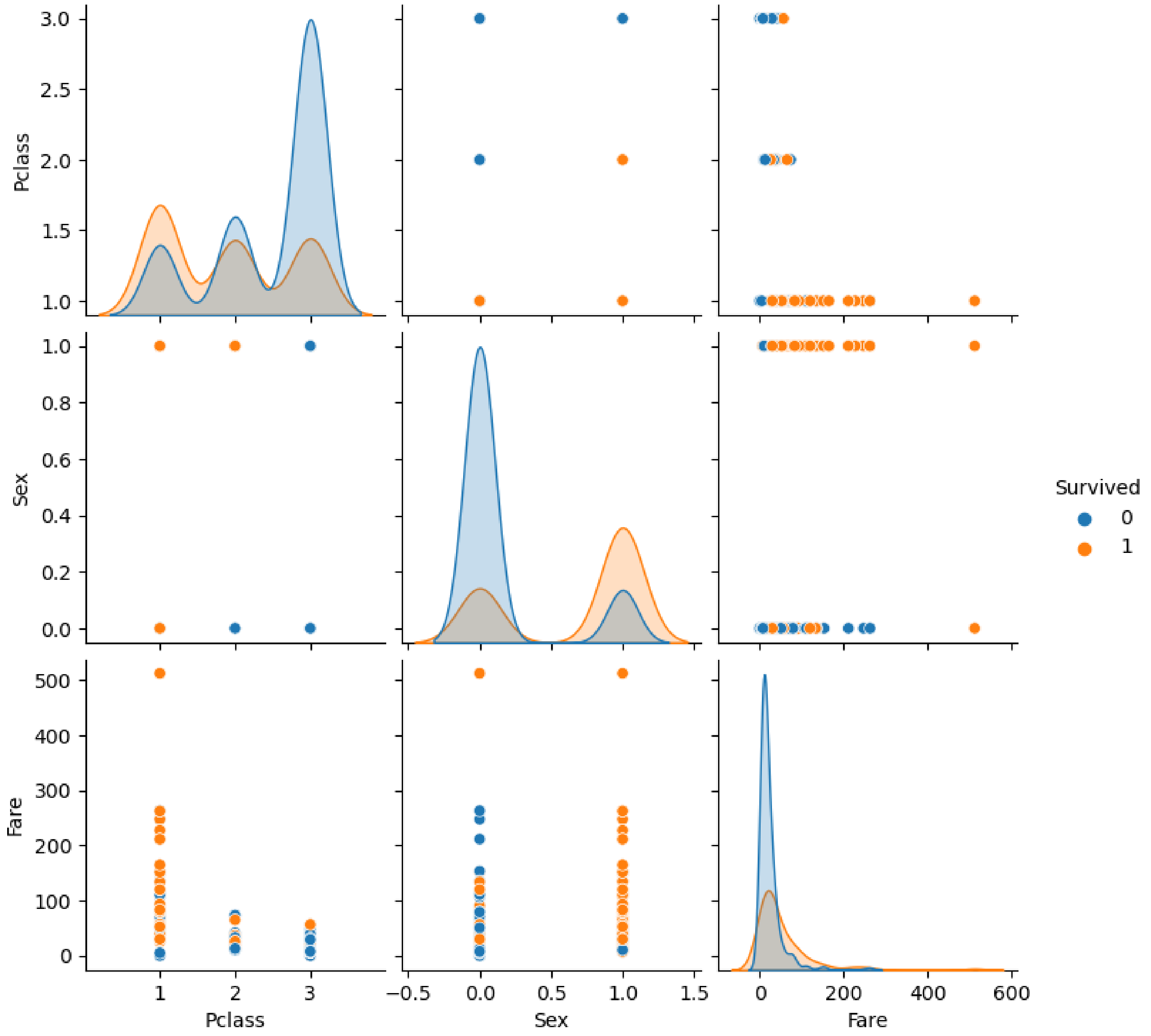


Deleting parameters which have poor correlation with survival rate

```
In [74]: train.drop(columns = ["Parch","SibSp","Age"],inplace = True)
```

Creating plots to decide which model is the best for our data

```
In [75]: warnings.filterwarnings('ignore')
         figure = sns.pairplot(train, hue = "Survived")
         plt.show()
```



Part 4 : Model Building

We will use classification model ,because data we need to predict boolean variable

From the graph is clear that Logistic Regression model is the best, because in the graphs overlapping is minimal

In [76]:
```
x_train , x_test , y_train , y_test = train_test_split(train[["Pclass","Sex","Fare"]], t
```

Logistic Regression

In [77]:
```
log_reg = LogisticRegression(random_state = 69).fit(x_train,y_train)
```

In [78]:
```
log_reg.score(x_test,y_test)
```

Out[78]:
```
0.813953488372093
```

Let check if statement above was right and Logistic Regression is the best model (We

can check it only in small datasets based on economic reasons)

Dessision tree

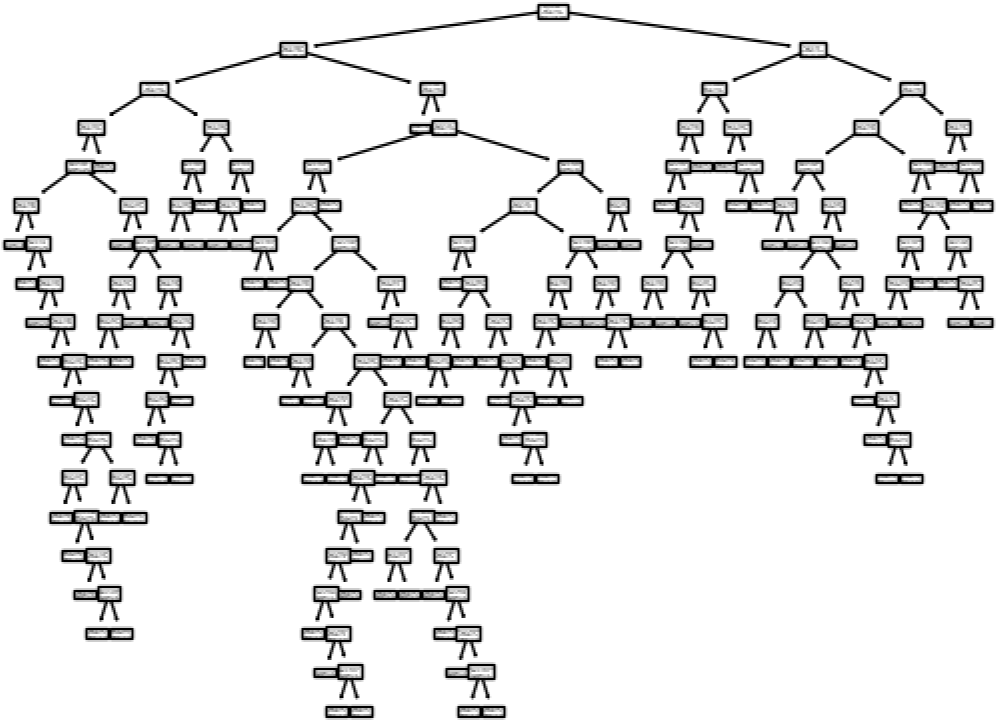
In [79]:
```
dt = tree.DecisionTreeClassifier().fit(x_train,y_train)
```

In [80]:
```
dt.score(x_test,y_test)
```

Out[80]:
```
0.786046511627907
```

In [81]:
```
tree.plot_tree(dt)
```

Out[81]:
```
[Text(0.5553101503759399, 0.9736842105263158, 'x[1] <= 0.5\ngini = 0.474\nsamples = 499
\nvalue = [306, 193]'),
 Text(0.29332706766917294, 0.9210526315789473, 'x[2] <= 15.646\ngini = 0.335\nsamples =
329\nvalue = [259, 70]'),
 Text(0.15338345864661654, 0.868421052631579, 'x[2] <= 12.5\ngini = 0.221\nsamples = 198
\nvalue = [173, 25]'),
 Text(0.09022556390977443, 0.8157894736842105, 'x[2] <= 12.413\ngini = 0.242\nsamples =
163\nvalue = [140, 23]'),
 Text(0.07819548872180451, 0.7631578947368421, 'x[2] <= 7.91\ngini = 0.235\nsamples = 16
2\nvalue = [140, 22]'),
 Text(0.02406015037593985, 0.7105263157894737, 'x[2] <= 6.862\ngini = 0.165\nsamples = 8
8\nvalue = [80, 8]'),
 Text(0.012030075187969926, 0.6578947368421053, 'gini = 0.0\nsamples = 8\nvalue = [8,
0]'),
 Text(0.03609022556390978, 0.6578947368421053, 'x[2] <= 7.01\ngini = 0.18\nsamples = 80
\nvalue = [72, 8]'),
 Text(0.02406015037593985, 0.6052631578947368, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.0481203007518797, 0.6052631578947368, 'x[2] <= 7.133\ngini = 0.162\nsamples = 79
\nvalue = [72, 7]'),
 Text(0.03609022556390978, 0.5526315789473685, 'gini = 0.0\nsamples = 11\nvalue = [11,
0]'),
 Text(0.06015037593984962, 0.5526315789473685, 'x[2] <= 7.183\ngini = 0.185\nsamples = 6
8\nvalue = [61, 7]'),
 Text(0.0481203007518797, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.07218045112781955, 0.5, 'x[2] <= 7.227\ngini = 0.163\nsamples = 67\nvalue = [61,
6]'),
 Text(0.06015037593984962, 0.4473684210526316, 'gini = 0.375\nsamples = 4\nvalue = [3,
1]'),
 Text(0.08421052631578947, 0.4473684210526316, 'x[2] <= 7.742\ngini = 0.146\nsamples = 6
3\nvalue = [58, 5]'),
 Text(0.07218045112781955, 0.39473684210526316, 'gini = 0.0\nsamples = 18\nvalue = [18,
0]'),
 Text(0.0962406015037594, 0.39473684210526316, 'x[2] <= 7.871\ngini = 0.198\nsamples = 4
5\nvalue = [40, 5]'),
 Text(0.07218045112781955, 0.34210526315789475, 'x[2] <= 7.763\ngini = 0.252\nsamples =
27\nvalue = [23, 4]'),
 Text(0.06015037593984962, 0.2894736842105263, 'gini = 0.32\nsamples = 5\nvalue = [4,
1]'),
 Text(0.08421052631578947, 0.2894736842105263, 'x[2] <= 7.785\ngini = 0.236\nsamples = 2
2\nvalue = [19, 3]'),
 Text(0.07218045112781955, 0.23684210526315788, 'gini = 0.198\nsamples = 9\nvalue = [8,
1]'),
 Text(0.0962406015037594, 0.23684210526315788, 'x[2] <= 7.798\ngini = 0.26\nsamples = 13
\nvalue = [11, 2]'),
 Text(0.08421052631578947, 0.18421052631578946, 'gini = 0.32\nsamples = 5\nvalue = [4,
1]'),
 Text(0.10827067669172932, 0.18421052631578946, 'x[2] <= 7.827\ngini = 0.219\nsamples =
8\nvalue = [7, 1]'),
 Text(0.0962406015037594, 0.13157894736842105, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.12030075187969924, 0.13157894736842105, 'gini = 0.245\nsamples = 7\nvalue = [6,
1]'),
 Text(0.12030075187969924, 0.34210526315789475, 'x[2] <= 7.892\ngini = 0.105\nsamples =
18\nvalue = [17, 1]'),
 Text(0.10827067669172932, 0.2894736842105263, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.13233082706766916, 0.2894736842105263, 'gini = 0.111\nsamples = 17\nvalue = [16,
1]'),
 Text(0.13233082706766916, 0.7105263157894737, 'x[2] <= 7.988\ngini = 0.307\nsamples = 7
4\nvalue = [60, 14]'),
 Text(0.12030075187969924, 0.6578947368421053, 'gini = 0.5\nsamples = 10\nvalue = [5,
5]'),
 Text(0.1443609022556391, 0.6578947368421053, 'x[2] <= 8.585\ngini = 0.242\nsamples = 64
\nvalue = [55, 9]'),
 Text(0.12030075187969924, 0.6052631578947368, 'x[2] <= 8.475\ngini = 0.33\nsamples = 24
\nvalue = [19, 5]'),
 Text(0.10827067669172932, 0.5526315789473685, 'x[2] <= 8.104\ngini = 0.287\nsamples = 2
3\nvalue = [19, 4]'),
 Text(0.0962406015037594, 0.5, 'gini = 0.32\nsamples = 20\nvalue = [16, 4]'),
 Text(0.12030075187969924, 0.5, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.13233082706766916, 0.5526315789473685, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.16842105263157894, 0.6052631578947368, 'x[2] <= 9.492\ngini = 0.18\nsamples = 40
\nvalue = [36, 4]'),
 Text(0.15639097744360902, 0.5526315789473685, 'gini = 0.0\nsamples = 14\nvalue = [14,
0]'),
 Text(0.18045112781954886, 0.5526315789473685, 'x[2] <= 11.317\ngini = 0.26\nsamples = 2
6\nvalue = [22, 4]'),
 Text(0.16842105263157894, 0.5, 'x[2] <= 10.817\ngini = 0.308\nsamples = 21\nvalue = [1
7, 4]'),
 Text(0.15639097744360902, 0.4473684210526316, 'x[2] <= 9.673\ngini = 0.255\nsamples = 2
0\nvalue = [17, 3]'),
 Text(0.1443609022556391, 0.39473684210526316, 'gini = 0.32\nsamples = 5\nvalue = [4,
1]'),
 Text(0.16842105263157894, 0.39473684210526316, 'x[0] <= 2.5\ngini = 0.231\nsamples = 15
\nvalue = [13, 2]'),
 Text(0.15639097744360902, 0.34210526315789475, 'gini = 0.26\nsamples = 13\nvalue = [11,
2]'),
 Text(0.18045112781954886, 0.34210526315789475, 'gini = 0.0\nsamples = 2\nvalue = [2,
0]'),
 Text(0.18045112781954886, 0.4473684210526316, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.1924812030075188, 0.5, 'gini = 0.0\nsamples = 5\nvalue = [5, 0]'),
 Text(0.10225563909774436, 0.7631578947368421, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.21654135338345865, 0.8157894736842105, 'x[2] <= 14.477\ngini = 0.108\nsamples =
35\nvalue = [33, 2]'),
 Text(0.1924812030075188, 0.7631578947368421, 'x[2] <= 13.25\ngini = 0.067\nsamples = 29
\nvalue = [28, 1]'),
 Text(0.18045112781954886, 0.7105263157894737, 'x[2] <= 12.938\ngini = 0.087\nsamples =
22\nvalue = [21, 1]'),
 Text(0.16842105263157894, 0.6578947368421053, 'gini = 0.0\nsamples = 2\nvalue = [2,
0]'),
 Text(0.1924812030075188, 0.6578947368421053, 'gini = 0.095\nsamples = 20\nvalue = [19,
1]'),
 Text(0.20451127819548873, 0.7105263157894737, 'gini = 0.0\nsamples = 7\nvalue = [7,
0]'),
 Text(0.24060150375939848, 0.7631578947368421, 'x[2] <= 14.75\ngini = 0.278\nsamples = 6
\nvalue = [5, 1]'),
 Text(0.22857142857142856, 0.7105263157894737, 'x[0] <= 2.5\ngini = 0.5\nsamples = 2\nva
lue = [1, 1]'),
 Text(0.21654135338345865, 0.6578947368421053, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.24060150375939848, 0.6578947368421053, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.25263157894736843, 0.7105263157894737, 'gini = 0.0\nsamples = 4\nvalue = [4,
0]'),
 Text(0.4332706766917293, 0.868421052631579, 'x[2] <= 16.0\ngini = 0.451\nsamples = 131
\nvalue = [86, 45]'),
 Text(0.4212406015037594, 0.8157894736842105, 'gini = 0.0\nsamples = 3\nvalue = [0,
3]'),
 Text(0.44530075187969925, 0.8157894736842105, 'x[0] <= 1.5\ngini = 0.441\nsamples = 128
\nvalue = [86, 42]'),
 Text(0.3176691729323308, 0.7631578947368421, 'x[2] <= 152.506\ngini = 0.492\nsamples =
64\nvalue = [36, 28]'),
 Text(0.3056390977443609, 0.7105263157894737, 'x[2] <= 26.419\ngini = 0.499\nsamples = 5
9\nvalue = [31, 28]'),
 Text(0.2646616541353383, 0.6578947368421053, 'x[2] <= 26.144\ngini = 0.375\nsamples = 4
\nvalue = [1, 3]'),
 Text(0.25263157894736843, 0.6052631578947368, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.27669172932330827, 0.6052631578947368, 'gini = 0.0\nsamples = 3\nvalue = [0,
3]'),
 Text(0.34661654135338343, 0.6578947368421053, 'x[2] <= 116.638\ngini = 0.496\nsamples =
55\nvalue = [30, 25]'),
 Text(0.3007518796992481, 0.6052631578947368, 'x[2] <= 29.85\ngini = 0.491\nsamples = 51
\nvalue = [29, 22]'),
 Text(0.2661654135338346, 0.5526315789473685, 'x[2] <= 27.135\ngini = 0.426\nsamples = 1
3\nvalue = [9, 4]'),
 Text(0.25413533834586466, 0.5, 'gini = 0.48\nsamples = 10\nvalue = [6, 4]'),
 Text(0.2781954887218045, 0.5, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.33533834586466166, 0.5526315789473685, 'x[2] <= 32.51\ngini = 0.499\nsamples = 3
8\nvalue = [20, 18]'),
 Text(0.3022556390977444, 0.5, 'x[2] <= 30.25\ngini = 0.278\nsamples = 6\nvalue = [1,
5]'),
 Text(0.29022556390977444, 0.4473684210526316, 'gini = 0.444\nsamples = 3\nvalue = [1,
2]'),
 Text(0.3142857142857143, 0.4473684210526316, 'gini = 0.0\nsamples = 3\nvalue = [0,
3]'),
 Text(0.3684210526315789, 0.5, 'x[2] <= 51.931\ngini = 0.482\nsamples = 32\nvalue = [19,
13]'),
 Text(0.3383458646616541, 0.4473684210526316, 'x[2] <= 37.812\ngini = 0.219\nsamples = 8
\nvalue = [7, 1]'),
 Text(0.3263157894736842, 0.39473684210526316, 'x[2] <= 34.76\ngini = 0.444\nsamples = 3
\nvalue = [2, 1]'),
 Text(0.3142857142857143, 0.34210526315789475, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.3383458646616541, 0.34210526315789475, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.35037593984962406, 0.39473684210526316, 'gini = 0.0\nsamples = 5\nvalue = [5,
0]'),
 Text(0.39849624060150374, 0.4473684210526316, 'x[2] <= 77.008\ngini = 0.5\nsamples = 24
\nvalue = [12, 12]'),
 Text(0.3744360902255639, 0.39473684210526316, 'x[2] <= 62.267\ngini = 0.444\nsamples =
12\nvalue = [4, 8]'),
 Text(0.362406015037594, 0.34210526315789475, 'x[2] <= 59.052\ngini = 0.48\nsamples = 10
\nvalue = [4, 6]'),
 Text(0.35037593984962406, 0.2894736842105263, 'x[2] <= 56.415\ngini = 0.444\nsamples =
9\nvalue = [3, 6]'),
 Text(0.3383458646616541, 0.23684210526315788, 'x[2] <= 55.671\ngini = 0.469\nsamples =
8\nvalue = [3, 5]'),
 Text(0.3263157894736842, 0.18421052631578946, 'x[2] <= 52.277\ngini = 0.408\nsamples =
7\nvalue = [2, 5]'),
 Text(0.3142857142857143, 0.13157894736842105, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.3383458646616541, 0.13157894736842105, 'x[2] <= 52.827\ngini = 0.32\nsamples = 5
\nvalue = [1, 4]'),
 Text(0.3263157894736842, 0.07894736842105263, 'gini = 0.0\nsamples = 2\nvalue = [0,
2]'),
 Text(0.35037593984962406, 0.07894736842105263, 'x[2] <= 54.271\ngini = 0.444\nsamples =
3\nvalue = [1, 2]'),
 Text(0.3383458646616541, 0.02631578947368421, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.362406015037594, 0.02631578947368421, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.35037593984962406, 0.18421052631578946, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.362406015037594, 0.23684210526315788, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.3744360902255639, 0.2894736842105263, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.38646616541353385, 0.34210526315789475, 'gini = 0.0\nsamples = 2\nvalue = [0,
2]'),
 Text(0.42255639097744363, 0.39473684210526316, 'x[2] <= 78.244\ngini = 0.444\nsamples =
12\nvalue = [8, 4]'),
 Text(0.4105263157894737, 0.34210526315789475, 'gini = 0.0\nsamples = 2\nvalue = [2,
0]'),
 Text(0.4345864661654135, 0.34210526315789475, 'x[2] <= 112.079\ngini = 0.48\nsamples =
10\nvalue = [6, 4]'),
 Text(0.42255639097744363, 0.2894736842105263, 'x[2] <= 80.754\ngini = 0.494\nsamples =
9\nvalue = [5, 4]'),
 Text(0.39849624060150374, 0.23684210526315788, 'x[2] <= 79.425\ngini = 0.444\nsamples =
3\nvalue = [2, 1]'),
 Text(0.38646616541353385, 0.18421052631578946, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.4105263157894737, 0.18421052631578946, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.44661654135338347, 0.23684210526315788, 'x[2] <= 82.667\ngini = 0.5\nsamples = 6
\nvalue = [3, 3]'),
 Text(0.4345864661654135, 0.18421052631578946, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.45864661654135336, 0.18421052631578946, 'x[2] <= 86.29\ngini = 0.48\nsamples = 5
\nvalue = [3, 2]'),
 Text(0.44661654135338347, 0.13157894736842105, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.4706766917293233, 0.13157894736842105, 'x[2] <= 89.552\ngini = 0.5\nsamples = 4
\nvalue = [2, 2]'),
 Text(0.45864661654135336, 0.07894736842105263, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.48270676691729325, 0.07894736842105263, 'x[2] <= 100.442\ngini = 0.444\nsamples
= 3\nvalue = [2, 1]'),
 Text(0.4706766917293233, 0.02631578947368421, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.49473684210526314, 0.02631578947368421, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.44661654135338347, 0.2894736842105263, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.3924812030075188, 0.6052631578947368, 'x[2] <= 134.642\ngini = 0.375\nsamples =
4\nvalue = [1, 3]'),
 Text(0.3804511278195489, 0.5526315789473685, 'gini = 0.0\nsamples = 2\nvalue = [0,
2]'),
 Text(0.4045112781954887, 0.5526315789473685, 'x[2] <= 143.592\ngini = 0.5\nsamples = 2
\nvalue = [1, 1]'),
 Text(0.3924812030075188, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
 Text(0.41654135338345866, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.32969924812030077, 0.7105263157894737, 'gini = 0.0\nsamples = 5\nvalue = [5,
0]'),
 Text(0.5729323308270676, 0.7631578947368421, 'x[2] <= 51.698\ngini = 0.342\nsamples = 6
4\nvalue = [50, 14]'),
 Text(0.524812030075188, 0.7105263157894737, 'x[0] <= 2.5\ngini = 0.311\nsamples = 57\nv
alue = [46, 11]'),
 Text(0.46466165413533833, 0.6578947368421053, 'x[2] <= 19.875\ngini = 0.444\nsamples =
24\nvalue = [16, 8]'),
 Text(0.45263157894736844, 0.6052631578947368, 'gini = 0.0\nsamples = 2\nvalue = [0,
2]'),
 Text(0.4766917293233083, 0.6052631578947368, 'x[2] <= 28.375\ngini = 0.397\nsamples = 2
2\nvalue = [16, 6]'),
 Text(0.45263157894736844, 0.5526315789473685, 'x[2] <= 26.125\ngini = 0.305\nsamples =
16\nvalue = [13, 3]'),
 Text(0.4406015037593985, 0.5, 'x[2] <= 25.0\ngini = 0.375\nsamples = 12\nvalue = [9,
3]'),
 Text(0.42857142857142855, 0.4473684210526316, 'gini = 0.0\nsamples = 3\nvalue = [3,
0]'),
 Text(0.45263157894736844, 0.4473684210526316, 'gini = 0.444\nsamples = 9\nvalue = [6,
3]'),
 Text(0.46466165413533833, 0.5, 'gini = 0.0\nsamples = 4\nvalue = [4, 0]'),
 Text(0.5007518796992482, 0.5526315789473685, 'x[2] <= 29.535\ngini = 0.5\nsamples = 6\n
value = [3, 3]'),
 Text(0.48872180451127817, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.512781954887218, 0.5, 'x[2] <= 33.41\ngini = 0.48\nsamples = 5\nvalue = [3,
2]'),
 Text(0.5007518796992482, 0.4473684210526316, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.524812030075188, 0.4473684210526316, 'x[2] <= 36.877\ngini = 0.5\nsamples = 4\nv
alue = [2, 2]'),
 Text(0.512781954887218, 0.39473684210526316, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.5368421052631579, 0.39473684210526316, 'x[2] <= 38.002\ngini = 0.444\nsamples =
3\nvalue = [2, 1]'),
 Text(0.524812030075188, 0.34210526315789475, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.5488721804511278, 0.34210526315789475, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.5849624060150376, 0.6578947368421053, 'x[2] <= 20.825\ngini = 0.165\nsamples = 3
3\nvalue = [30, 3]'),
 Text(0.5609022556390978, 0.6052631578947368, 'x[2] <= 20.55\ngini = 0.298\nsamples = 11
\nvalue = [9, 2]'),
 Text(0.5488721804511278, 0.5526315789473685, 'x[2] <= 18.394\ngini = 0.18\nsamples = 10
\nvalue = [9, 1]'),
 Text(0.5368421052631579, 0.5, 'gini = 0.0\nsamples = 6\nvalue = [6, 0]'),
 Text(0.5609022556390978, 0.5, 'x[2] <= 19.5\ngini = 0.375\nsamples = 4\nvalue = [3,
1]'),
 Text(0.5488721804511278, 0.4473684210526316, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.5729323308270676, 0.4473684210526316, 'gini = 0.0\nsamples = 2\nvalue = [2,
0]'),
 Text(0.5729323308270676, 0.5526315789473685, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.6090225563909775, 0.6052631578947368, 'x[2] <= 31.331\ngini = 0.087\nsamples = 2
2\nvalue = [21, 1]'),
 Text(0.5969924812030075, 0.5526315789473685, 'gini = 0.0\nsamples = 12\nvalue = [12,
0]'),
 Text(0.6210526315789474, 0.5526315789473685, 'x[2] <= 32.881\ngini = 0.18\nsamples = 10
\nvalue = [9, 1]'),
 Text(0.6090225563909775, 0.5, 'gini = 0.5\nsamples = 2\nvalue = [1, 1]'),
 Text(0.6330827067669172, 0.5, 'gini = 0.0\nsamples = 8\nvalue = [8, 0]'),
 Text(0.6210526315789474, 0.7105263157894737, 'x[0] <= 2.5\ngini = 0.49\nsamples = 7\nva
lue = [4, 3]'),
 Text(0.6090225563909775, 0.6578947368421053, 'gini = 0.0\nsamples = 4\nvalue = [4,
0]'),
 Text(0.6330827067669172, 0.6578947368421053, 'gini = 0.0\nsamples = 3\nvalue = [0,
3]'),
 Text(0.8172932330827067, 0.9210526315789473, 'x[0] <= 2.5\ngini = 0.4\nsamples = 170\nv
alue = [47, 123]'),
 Text(0.7172932330827068, 0.868421052631579, 'x[2] <= 29.856\ngini = 0.113\nsamples = 10
0\nvalue = [6, 94]'),
 Text(0.6932330827067669, 0.8157894736842105, 'x[2] <= 28.231\ngini = 0.165\nsamples = 4
4\nvalue = [4, 40]'),
 Text(0.681203007518797, 0.7631578947368421, 'x[2] <= 12.825\ngini = 0.13\nsamples = 43
\nvalue = [3, 40]'),
 Text(0.6691729323308271, 0.7105263157894737, 'gini = 0.0\nsamples = 9\nvalue = [0,
9]'),
 Text(0.6932330827067669, 0.7105263157894737, 'x[2] <= 26.125\ngini = 0.161\nsamples = 3
4\nvalue = [3, 31]'),
 Text(0.681203007518797, 0.6578947368421053, 'x[2] <= 20.25\ngini = 0.198\nsamples = 27
\nvalue = [3, 24]'),
 Text(0.6571428571428571, 0.6052631578947368, 'x[2] <= 13.25\ngini = 0.153\nsamples = 12
\nvalue = [1, 11]'),
 Text(0.6451127819548872, 0.5526315789473685, 'gini = 0.245\nsamples = 7\nvalue = [1,
6]'),
 Text(0.6691729323308271, 0.5526315789473685, 'gini = 0.0\nsamples = 5\nvalue = [0,
5]'),
 Text(0.7052631578947368, 0.6052631578947368, 'x[2] <= 22.0\ngini = 0.231\nsamples = 15
\nvalue = [2, 13]'),
 Text(0.6932330827067669, 0.5526315789473685, 'gini = 0.444\nsamples = 3\nvalue = [1,
2]'),
 Text(0.7172932330827068, 0.5526315789473685, 'x[2] <= 25.965\ngini = 0.153\nsamples = 1
2\nvalue = [1, 11]'),
 Text(0.7052631578947368, 0.5, 'gini = 0.0\nsamples = 4\nvalue = [0, 4]'),
 Text(0.7293233082706767, 0.5, 'gini = 0.219\nsamples = 8\nvalue = [1, 7]'),
 Text(0.7052631578947368, 0.6578947368421053, 'gini = 0.0\nsamples = 7\nvalue = [0,
7]'),
 Text(0.7052631578947368, 0.7631578947368421, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.7413533834586467, 0.8157894736842105, 'x[2] <= 149.035\ngini = 0.069\nsamples =
56\nvalue = [2, 54]'),
 Text(0.7293233082706767, 0.7631578947368421, 'gini = 0.0\nsamples = 39\nvalue = [0, 3
9]'),
 Text(0.7533834586466165, 0.7631578947368421, 'x[2] <= 152.506\ngini = 0.208\nsamples =
17\nvalue = [2, 15]'),
 Text(0.7413533834586467, 0.7105263157894737, 'gini = 0.444\nsamples = 3\nvalue = [2,
1]'),
 Text(0.7654135338345864, 0.7105263157894737, 'gini = 0.0\nsamples = 14\nvalue = [0, 1
4]'),
 Text(0.9172932330827067, 0.868421052631579, 'x[2] <= 20.8\ngini = 0.485\nsamples = 70\n
value = [41, 29]'),
 Text(0.8706766917293233, 0.8157894736842105, 'x[2] <= 10.798\ngini = 0.497\nsamples = 5
0\nvalue = [23, 27]'),
 Text(0.8135338345864662, 0.7631578947368421, 'x[2] <= 7.742\ngini = 0.483\nsamples = 27
\nvalue = [16, 11]'),
 Text(0.7894736842105263, 0.7105263157894737, 'x[2] <= 6.987\ngini = 0.32\nsamples = 5\n
value = [1, 4]'),
 Text(0.7774436090225564, 0.6578947368421053, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.8015037593984963, 0.6578947368421053, 'gini = 0.0\nsamples = 4\nvalue = [0,
4]'),
 Text(0.837593984962406, 0.7105263157894737, 'x[2] <= 9.706\ngini = 0.434\nsamples = 22
\nvalue = [15, 7]'),
 Text(0.825563909774436, 0.6578947368421053, 'x[2] <= 7.977\ngini = 0.465\nsamples = 19
\nvalue = [12, 7]'),
 Text(0.7954887218045112, 0.6052631578947368, 'x[2] <= 7.815\ngini = 0.375\nsamples = 12
\nvalue = [9, 3]'),
 Text(0.7714285714285715, 0.5526315789473685, 'x[2] <= 7.763\ngini = 0.278\nsamples = 6
\nvalue = [5, 1]'),
 Text(0.7593984962406015, 0.5, 'gini = 0.375\nsamples = 4\nvalue = [3, 1]'),
 Text(0.7834586466165413, 0.5, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(0.8195488721804511, 0.5526315789473685, 'x[2] <= 7.89\ngini = 0.444\nsamples = 6\n
value = [4, 2]'),
 Text(0.8075187969924812, 0.5, 'gini = 0.444\nsamples = 3\nvalue = [2, 1]'),
 Text(0.8315789473684211, 0.5, 'gini = 0.444\nsamples = 3\nvalue = [2, 1]'),
 Text(0.8556390977443609, 0.6052631578947368, 'x[2] <= 8.346\ngini = 0.49\nsamples = 7\n
value = [3, 4]'),
 Text(0.843609022556391, 0.5526315789473685, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.8676691729323308, 0.5526315789473685, 'x[2] <= 8.673\ngini = 0.5\nsamples = 6\nv
alue = [3, 3]'),
 Text(0.8556390977443609, 0.5, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
 Text(0.8796992481203008, 0.5, 'x[2] <= 8.767\ngini = 0.48\nsamples = 5\nvalue = [2,
3]'),
 Text(0.8676691729323308, 0.4473684210526316, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.8917293233082707, 0.4473684210526316, 'x[2] <= 9.1\ngini = 0.5\nsamples = 4\nval
ue = [2, 2]'),
 Text(0.8796992481203008, 0.39473684210526316, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.9037593984962407, 0.39473684210526316, 'x[2] <= 9.469\ngini = 0.444\nsamples = 3
\nvalue = [1, 2]'),
 Text(0.8917293233082707, 0.34210526315789475, 'gini = 0.0\nsamples = 1\nvalue = [0,
1]'),
 Text(0.9157894736842105, 0.34210526315789475, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.849624060150376, 0.6578947368421053, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.9278195488721804, 0.7631578947368421, 'x[2] <= 13.438\ngini = 0.423\nsamples = 2
3\nvalue = [7, 16]'),
 Text(0.9157894736842105, 0.7105263157894737, 'gini = 0.0\nsamples = 5\nvalue = [0,
5]'),
 Text(0.9398496240601504, 0.7105263157894737, 'x[2] <= 15.121\ngini = 0.475\nsamples = 1
8\nvalue = [7, 11]'),
 Text(0.9157894736842105, 0.6578947368421053, 'x[2] <= 14.456\ngini = 0.32\nsamples = 5
\nvalue = [4, 1]'),
 Text(0.9037593984962407, 0.6052631578947368, 'x[2] <= 14.427\ngini = 0.444\nsamples = 3
\nvalue = [2, 1]'),
 Text(0.8917293233082707, 0.5526315789473685, 'gini = 0.0\nsamples = 1\nvalue = [1,
0]'),
 Text(0.9157894736842105, 0.5526315789473685, 'gini = 0.5\nsamples = 2\nvalue = [1,
1]'),
 Text(0.9278195488721804, 0.6052631578947368, 'gini = 0.0\nsamples = 2\nvalue = [2,
0]'),
 Text(0.9639097744360903, 0.6578947368421053, 'x[2] <= 17.6\ngini = 0.355\nsamples = 13
\nvalue = [3, 10]'),
 Text(0.9518796992481203, 0.6052631578947368, 'gini = 0.0\nsamples = 6\nvalue = [0,
6]'),
 Text(0.9759398496240601, 0.6052631578947368, 'x[2] <= 18.629\ngini = 0.49\nsamples = 7
\nvalue = [3, 4]'),
 Text(0.9639097744360903, 0.5526315789473685, 'gini = 0.0\nsamples = 3\nvalue = [3,
0]'),
 Text(0.98796992481203, 0.5526315789473685, 'gini = 0.0\nsamples = 4\nvalue = [0, 4]'),
 Text(0.9639097744360903, 0.8157894736842105, 'x[2] <= 31.331\ngini = 0.18\nsamples = 20
\nvalue = [18, 2]'),
 Text(0.9518796992481203, 0.7631578947368421, 'gini = 0.0\nsamples = 12\nvalue = [12,
0]'),
 Text(0.9759398496240601, 0.7631578947368421, 'x[2] <= 32.881\ngini = 0.375\nsamples = 8
\nvalue = [6, 2]'),
 Text(0.9639097744360903, 0.7105263157894737, 'gini = 0.0\nsamples = 2\nvalue = [0,
2]'),
 Text(0.98796992481203, 0.7105263157894737, 'gini = 0.0\nsamples = 6\nvalue = [6, 0]')]
```



Random Forest Classifier

In [82]:

```
rfc = RandomForestClassifier(max_depth = 5 , random_state = 69).fit(x_train, y_train)
```

In [83]:

```
rfc.score(x_test,y_test)
```

Out[83]:

```
0.7953488372093023
```

After data visualizing and testing models we can conclude that Logistic Regression is the best model because of high accuaracy and low complexity levels, compare to another models.

In conclusion, the Titanic project successfully explored, analyzed, and modeled the dataset to predict passenger survival outcomes during the sinking of the Titanic.

In [84]:

```
test.isna().sum()
```

Out[84]:

```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [85]:

```
test.drop(columns = ["Name", "Ticket", "Cabin", "Embarked","Parch", "Age" ,"SibSp"], inp
test.Sex = test.Sex.replace({"female" : 1, "male" : 0})
```

In [86]:

```
test.Fare.fillna(value = test["Fare"].mean(), inplace = True)
test.Sex.fillna(value = test["Sex"].mean(), inplace = True)
test.Pclass.fillna(value = test["Pclass"].mean(), inplace = True)
```

In [87]:

```
test.isna().sum()
```

Out[87]:

```
PassengerId    0
Pclass         0
Sex            0
Fare           0
dtype: int64
```

In [88]:

```
test
```

Out[88]:

	PassengerId	Pclass	Sex	Fare
0	892	3	0	7.8292
1	893	3	1	7.0000
2	894	2	0	9.6875
3	895	3	0	8.6625
4	896	3	1	12.2875
...	...	...	...	...
413	1305	3	0	8.0500
414	1306	1	1	108.9000
415	1307	3	0	7.2500
416	1308	3	0	8.0500
417	1309	3	0	22.3583

418 rows × 4 columns

In [89]:

```
result = log_reg.predict(test[["Pclass","Sex","Fare"]])
```

In [90]:

```
result = pd.DataFrame(result)
```

In [91]:

```
result = pd.concat([test["PassengerId"],result],axis = 1)
```

In [92]:

```
result
```

Out[92]:

	PassengerId	0
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1
...	...	...
413	1305	0
414	1306	1
415	1307	0
416	1308	0
417	1309	0

418 rows × 2 columns

In [93]:

```
result.columns = ["PassengerId" , "Survived"]
```

In [97]:

```
result.reset_index(drop = True,inplace = True)
```