

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sea
```

```
In [2]: df = pd.read_csv(r"C:\Users\user\Downloads\C2_test.gender_submission (1).csv")
df
```

Out[2]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
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 [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  -
0   PassengerId  418 non-null    int64
1   Pclass      418 non-null    int64
2   Name        418 non-null    object
3   Sex         418 non-null    object
4   Age         332 non-null    float64
5   SibSp       418 non-null    int64
6   Parch       418 non-null    int64
7   Ticket      418 non-null    object
8   Fare        417 non-null    float64
9   Cabin       91 non-null     object
10  Embarked    418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

```
In [17]: df1 = df.fillna(0)
df1
```

Out[17]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	I
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	0	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	0	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	0	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	0	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	0	
...
413	1305	3	Spector, Mr. Woolf	male	0.0	0	0	A.5. 3236	8.0500	0	
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	0	
416	1308	3	Ware, Mr. Frederick	male	0.0	0	0	359309	8.0500	0	
417	1309	3	Peter, Master. Michael J	male	0.0	1	1	2668	22.3583	0	

418 rows × 11 columns



```
In [18]: df1.columns
```

Out[18]: Index(['PassengerId', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'], dtype='object')

In [8]: df1.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  -
0   PassengerId  418 non-null    int64
1   Pclass       418 non-null    int64
2   Name         418 non-null    object
3   Sex          418 non-null    object
4   Age          418 non-null    float64
5   SibSp        418 non-null    int64
6   Parch        418 non-null    int64
7   Ticket       418 non-null    object
8   Fare         418 non-null    float64
9   Cabin        418 non-null    object
10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

In [15]: df1['Sex'].value_counts()

```
Out[15]: male      266
female    152
Name: Sex, dtype: int64
```

In [30]: df2 = df1[['PassengerId', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'Sex']]
df2

Out[30]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex
0	892	3	34.5	0	0	7.8292	male
1	893	3	47.0	1	0	7.0000	female
2	894	2	62.0	0	0	9.6875	male
3	895	3	27.0	0	0	8.6625	male
4	896	3	22.0	1	1	12.2875	female
...
413	1305	3	0.0	0	0	8.0500	male
414	1306	1	39.0	0	0	108.9000	female
415	1307	3	38.5	0	0	7.2500	male
416	1308	3	0.0	0	0	8.0500	male
417	1309	3	0.0	1	1	22.3583	male

418 rows × 7 columns

```
In [33]: x = df2[['PassengerId', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']]
y = df2['Sex']
```

```
In [34]: e1 = {"Sex":{'female':0,'male':1}}
df3 = df2.replace(e1)
df3
```

Out[34]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex
0	892	3	34.5	0	0	7.8292	1
1	893	3	47.0	1	0	7.0000	0
2	894	2	62.0	0	0	9.6875	1
3	895	3	27.0	0	0	8.6625	1
4	896	3	22.0	1	1	12.2875	0
...
413	1305	3	0.0	0	0	8.0500	1
414	1306	1	39.0	0	0	108.9000	0
415	1307	3	38.5	0	0	7.2500	1
416	1308	3	0.0	0	0	8.0500	1
417	1309	3	0.0	1	1	22.3583	1

418 rows × 7 columns

```
In [35]: from sklearn.model_selection import train_test_split
```

```
In [36]: x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.70)
```

```
In [37]: from sklearn.ensemble import RandomForestClassifier
```

```
In [38]: rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[38]: RandomForestClassifier()

```
In [42]: parameters = {
    'max_depth':[11,12,13,14,15],
    'min_samples_leaf':[15,20,25,30,35],
    'n_estimators':[10,20,30,40,50]
}
```

```
In [43]: from sklearn.model_selection import GridSearchCV
```

```
In [44]: grid_search = GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring='acc')
grid_search.fit(x_train,y_train)
```

```
Out[44]: GridSearchCV(cv=2, estimator=RandomForestClassifier(),
                      param_grid={'max_depth': [11, 12, 13, 14, 15],
                                   'min_samples_leaf': [15, 20, 25, 30, 35],
                                   'n_estimators': [10, 20, 30, 40, 50]},
                      scoring='accuracy')
```

```
In [45]: grid_search.best_score_
```

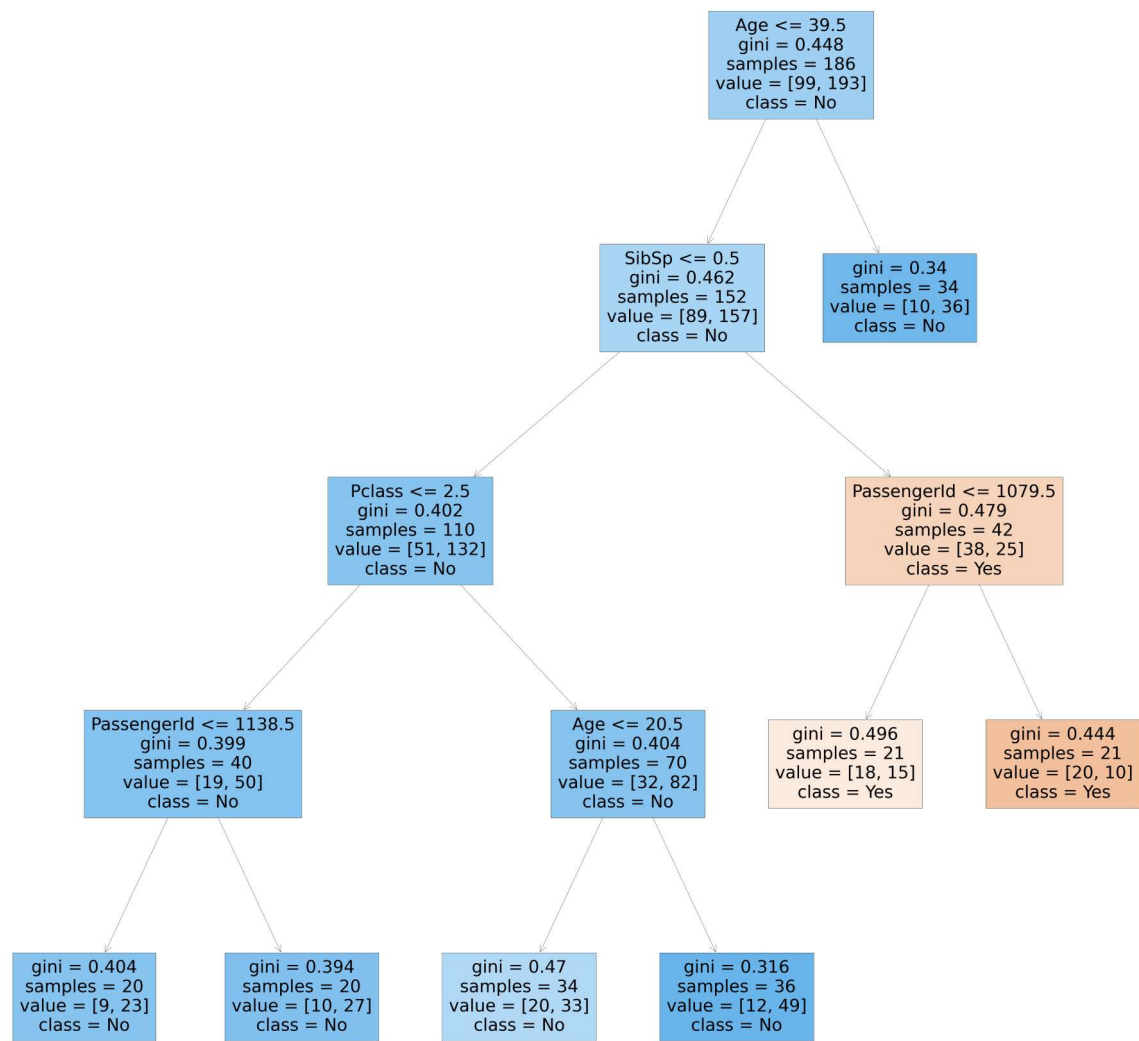
```
Out[45]: 0.6609589041095891
```

```
In [46]: from sklearn.tree import plot_tree
```

```
In [47]: rfc_best= grid_search.best_estimator_
```

```
In [50]: plt.figure(figsize=(100,100))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=["Yes","No"],
```

```
Out[50]: [Text(3804.5454545454545, 4892.4, 'Age <= 39.5\ngini = 0.448\nsamples = 186\nvalue = [99, 193]\n\nclass = No'),
Text(3297.272727272727, 3805.2, 'SibSp <= 0.5\ngini = 0.462\nsamples = 152\nvalue = [89, 157]\n\nclass = No'),
Text(2029.090909090909, 2718.0, 'Pclass <= 2.5\ngini = 0.402\nsamples = 110\n\nclass = No'),
Text(1014.5454545454545, 1630.7999999999997, 'PassengerId <= 1138.5\ngini = 0.399\nsamples = 40\n\nclass = No'),
Text(507.27272727272725, 543.5999999999995, 'gini = 0.404\nsamples = 20\n\nclass = No'),
Text(1521.8181818181818, 543.5999999999995, 'gini = 0.394\nsamples = 20\n\nclass = No'),
Text(3043.6363636363635, 1630.7999999999997, 'Age <= 20.5\ngini = 0.404\nsamples = 70\n\nclass = No'),
Text(2536.363636363636, 543.5999999999995, 'gini = 0.47\nsamples = 34\n\nclass = No'),
Text(3550.909090909091, 543.5999999999995, 'gini = 0.316\nsamples = 36\n\nclass = No'),
Text(4565.454545454545, 2718.0, 'PassengerId <= 1079.5\ngini = 0.479\nsamples = 42\n\nclass = Yes'),
Text(4058.181818181818, 1630.7999999999997, 'gini = 0.496\nsamples = 21\n\nclass = Yes'),
Text(5072.727272727272, 1630.7999999999997, 'gini = 0.444\nsamples = 21\n\nclass = Yes'),
Text(4311.818181818182, 3805.2, 'gini = 0.34\nsamples = 34\n\nclass = No')]
```



In []: