RANDOM FOREST- 2

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
df=pd.read_csv(r"C:\Users\BHOOMISH\Downloads\C2_train.gender_submission.csv")
df
```

Out[3]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [5]:
```
df=df.drop('Cabin',axis=1)
```

In [6]:
```
df=df.dropna()
```

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
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 [8]:
```
df.describe()
```

Out[8]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	712.000000	712.000000	712.000000	712.000000	712.000000	712.000000	712.000000
mean	448.589888	0.404494	2.240169	29.642093	0.514045	0.432584	34.567251
std	258.683191	0.491139	0.836854	14.492933	0.930692	0.854181	52.938648
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	222.750000	0.000000	1.000000	20.000000	0.000000	0.000000	8.050000
50%	445.000000	0.000000	2.000000	28.000000	0.000000	0.000000	15.645850
75%	677.250000	1.000000	3.000000	38.000000	1.000000	1.000000	33.000000
max	891.000000	1.000000	3.000000	80.000000	5.000000	6.000000	512.329200

In [9]:
```
df["Survived"].value_counts()
```

Out[9]:
```
0    424
1    288
Name: Survived, dtype: int64
```

In [10]:
```
df1=df[['PassengerId','Survived','Pclass','Age','SibSp','Parch','Fare']]
```

In [11]:
```
x=df1.drop("Survived",axis=1)
y=df1["Survived"]
```

In [12]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

In [13]:
```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[13]:
```
RandomForestClassifier
RandomForestClassifier()
```

In [14]:
```
parameters={'max_depth':[1,2,3,4,5],
            'min_samples_leaf':[5,10,15,20,25],
            'n_estimators':[10,20,30,40,50]}
```

In [15]:
```
from sklearn.model_selection import GridSearchCV
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accuracy")
grid_search.fit(x_train,y_train)
```

Out[15]:
```
GridSearchCV
estimator: RandomForestClassifier
RandomForestClassifier
```

In [16]:
```
grid_search.best_score_
```

Out[16]:
```
0.7329317269076305
```

In [17]:
```
parameters=dff
```

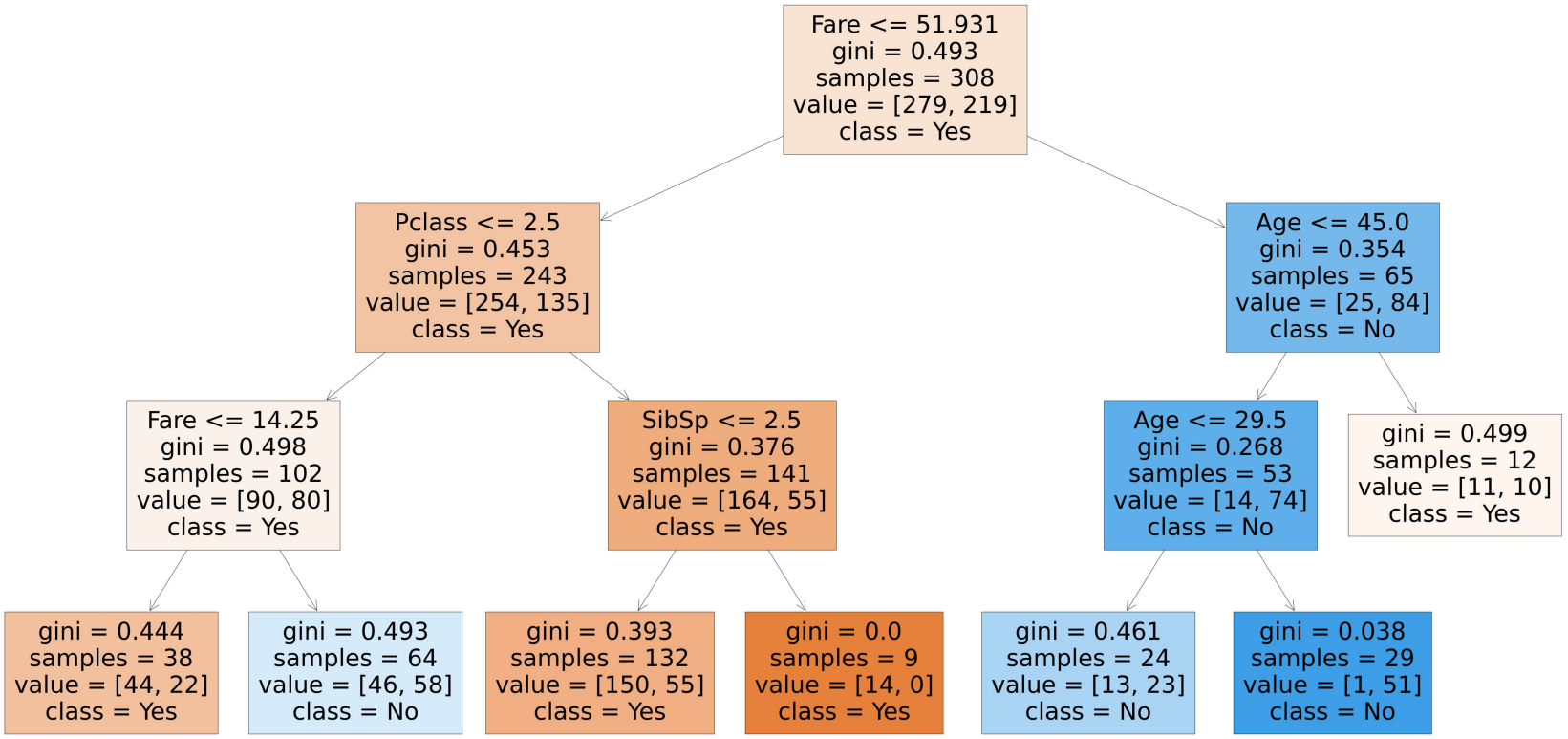
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[17], line 1
----> 1 parameters=dff

NameError: name 'dff' is not defined
```

In [18]:
```
rfc_best=grid_search.best_estimator_
```

In [19]:
```
from sklearn.tree import plot_tree
plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Yes','No'],filled=True)
```

Out[19]:
```
[Text(0.5769230769230769, 0.875, 'Fare <= 51.931\ngini = 0.493\nsamples = 308\nvalue = [279, 219]\nclass = Y
es'),
 Text(0.3076923076923077, 0.625, 'Pclass <= 2.5\ngini = 0.453\nsamples = 243\nvalue = [254, 135]\nclass = Ye
s'),
 Text(0.15384615384615385, 0.375, 'Fare <= 14.25\ngini = 0.498\nsamples = 102\nvalue = [90, 80]\nclass = Ye
s'),
 Text(0.07692307692307693, 0.125, 'gini = 0.444\nsamples = 38\nvalue = [44, 22]\nclass = Yes'),
 Text(0.23076923076923078, 0.125, 'gini = 0.493\nsamples = 64\nvalue = [46, 58]\nclass = No'),
 Text(0.46153846153846156, 0.375, 'SibSp <= 2.5\ngini = 0.376\nsamples = 141\nvalue = [164, 55]\nclass = Ye
s'),
 Text(0.38461538461538464, 0.125, 'gini = 0.393\nsamples = 132\nvalue = [150, 55]\nclass = Yes'),
 Text(0.5384615384615384, 0.125, 'gini = 0.0\nsamples = 9\nvalue = [14, 0]\nclass = Yes'),
 Text(0.8461538461538461, 0.625, 'Age <= 45.0\ngini = 0.354\nsamples = 65\nvalue = [25, 84]\nclass = No'),
 Text(0.7692307692307693, 0.375, 'Age <= 29.5\ngini = 0.268\nsamples = 53\nvalue = [14, 74]\nclass = No'),
 Text(0.6923076923076923, 0.125, 'gini = 0.461\nsamples = 24\nvalue = [13, 23]\nclass = No'),
 Text(0.8461538461538461, 0.125, 'gini = 0.038\nsamples = 29\nvalue = [1, 51]\nclass = No'),
 Text(0.9230769230769231, 0.375, 'gini = 0.499\nsamples = 12\nvalue = [11, 10]\nclass = Yes')]
```


Fare <= 51.931
gini = 0.493
samples = 308
value = [279, 219]
class = Yes
Pclass <= 2.5
gini = 0.453
samples = 243
value = [254, 135]
class = Yes
Age <= 45.0
gini = 0.354
samples = 65
value = [25, 84]
class = No
Fare <= 14.25
gini = 0.498
samples = 102
value = [90, 80]
class = Yes
SibSp <= 2.5
gini = 0.376
samples = 141
value = [164, 55]
class = Yes
Age <= 29.5
gini = 0.268
samples = 53
value = [14, 74]
class = No
gini = 0.499
samples = 12
value = [11, 10]
class = Yes
gini = 0.444
samples = 38
value = [44, 22]
class = Yes
gini = 0.493
samples = 64
value = [46, 58]
class = No
gini = 0.393
samples = 132
value = [150, 55]
class = Yes
gini = 0.0
samples = 9
value = [14, 0]
class = Yes
gini = 0.461
samples = 24
value = [13, 23]
class = No
gini = 0.038
samples = 29
value = [1, 51]
class = No

In [ ]: