

Radom Forest

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
from sklearn.linear_model import LogisticRegression
```

In [3]:

```
df=pd.read_csv(r"C:\Users\user\Downloads\C2_test.csv")[0:100]
df
```

Out[3]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
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	
...	
95	987	3	Tenglin, Mr. Gunnar Isidor	male	25.0	0	0	350033	7.7958	NaN	
96	988	1	Cavendish, Mrs. Tyrell William (Julia Florence...)	female	76.0	1	0	19877	78.8500	C46	
97	989	3	Makinen, Mr. Kalle Edvard	male	29.0	0	0	STON/O 2. 3101268	7.9250	NaN	
98	990	3	Braf, Miss. Elin Ester Maria	female	20.0	0	0	347471	7.8542	NaN	
99	991	3	Nancarrow, Mr. William Henry	male	33.0	0	0	A./5. 3338	8.0500	NaN	

100 rows × 11 columns

In [4]:

```
df.columns
```

Out[4]:

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

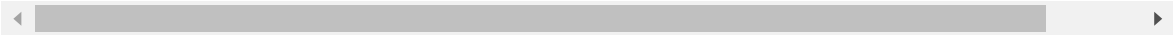
In [5]:

```
df.fillna(value=1)
```

Out[5]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	1	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	1	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	1	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	1	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	1	
...
95	987	3	Tenglin, Mr. Gunnar Isidor	male	25.0	0	0	350033	7.7958	1	
96	988	1	Cavendish, Mrs. Tyrell William (Julia Florence...)	female	76.0	1	0	19877	78.8500	C46	
97	989	3	Makinen, Mr. Kalle Edvard	male	29.0	0	0	STON/O 2. 3101268	7.9250	1	
98	990	3	Braf, Miss. Elin Ester Maria	female	20.0	0	0	347471	7.8542	1	
99	991	3	Nancarrow, Mr. William Henry	male	33.0	0	0	A./5. 3338	8.0500	1	

100 rows × 11 columns



In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  -
 0   PassengerId     100 non-null    int64
 1   Pclass          100 non-null    int64
 2   Name            100 non-null    object
 3   Sex             100 non-null    object
 4   Age             82 non-null     float64
 5   SibSp           100 non-null    int64
 6   Parch           100 non-null    int64
 7   Ticket          100 non-null    object
 8   Fare            100 non-null    float64
 9   Cabin           24 non-null     object
10   Embarked        100 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 8.7+ KB
```

In [7]:

```
d=df[['PassengerId', 'Pclass', 'SibSp', 'Parch','Sex']][0:50]
```

In [8]:

```
df['Sex'].value_counts()
```

Out[8]:

```
male      61
female    39
Name: Sex, dtype: int64
```

In [9]:

```
x=d.drop('Sex',axis=1)
y=d['Sex']
TenYearCHD1={"Sex":{"male":0,'female':1}}
d=d.replace('Sex')
print(d)
```

	PassengerId	Pclass	SibSp	Parch	Sex
0	892	3	0	0	male
1	893	3	1	0	female
2	894	2	0	0	male
3	895	3	0	0	male
4	896	3	1	1	female
5	897	3	0	0	male
6	898	3	0	0	female
7	899	2	1	1	male
8	900	3	0	0	female
9	901	3	2	0	male
10	902	3	0	0	male
11	903	1	0	0	male
12	904	1	1	0	female
13	905	2	1	0	male
14	906	1	1	0	female
15	907	2	1	0	female
16	908	2	0	0	male
17	909	3	0	0	male
18	910	3	1	0	female
19	911	3	0	0	female
20	912	1	1	0	male
21	913	3	0	1	male
22	914	1	0	0	female
23	915	1	0	1	male
24	916	1	1	3	female
25	917	3	1	0	male
26	918	1	0	1	female
27	919	3	0	0	male
28	920	1	0	0	male
29	921	3	2	0	male
30	922	2	1	0	male
31	923	2	2	0	male
32	924	3	1	2	female
33	925	3	1	2	female
34	926	1	1	0	male
35	927	3	0	0	male
36	928	3	0	0	female
37	929	3	0	0	female
38	930	3	0	0	male
39	931	3	0	0	male
40	932	3	0	1	male
41	933	1	0	0	male
42	934	3	0	0	male
43	935	2	0	0	female
44	936	1	1	0	female
45	937	3	0	0	male
46	938	1	0	0	male
47	939	3	0	0	male
48	940	1	0	0	female
49	941	3	0	2	female

In [10]:

```
from sklearn.model_selection import train_test_split
```

In [11]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

In [12]:

```
from sklearn.ensemble import RandomForestClassifier  
rfc=RandomForestClassifier()  
rfc.fit(x_train,y_train)
```

Out[12]:

```
RandomForestClassifier()
```

In [13]:

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

In [14]:

```
from sklearn.model_selection import GridSearchCV
```

In [15]:

```
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accuracy")
```

In [16]:

```
grid_search.fit(x_train,y_train)
```

Out[16]:

```
GridSearchCV(cv=2, estimator=RandomForestClassifier(),  
             param_grid={'max_depth': [1, 2, 3, 4, 5],  
                         'min_samples_leaf': [5, 10, 15, 20, 25],  
                         'n_estimators': [10, 20, 30, 40, 50]},  
             scoring='accuracy')
```

In [17]:

```
grid_search.best_score_
```

Out[17]:

```
0.6290849673202614
```

In [18]:

```
rfc_best=grid_search.best_estimator_
```

In [19]:

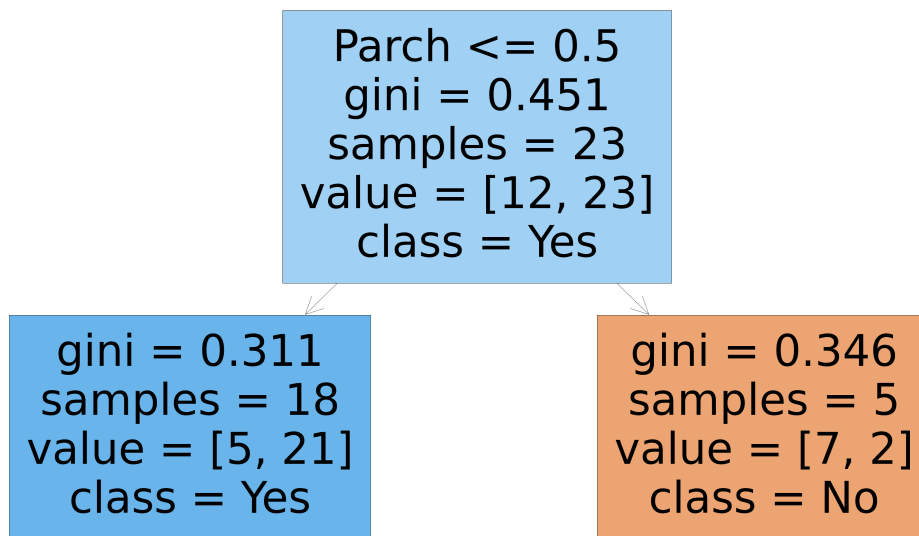
```
from sklearn.tree import plot_tree
```

In [20]:

```
plt.figure(figsize=(80,40))  
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['No','Yes'],filled
```

Out[20]:

```
[Text(2232.0, 1630.8000000000002, 'Parch <= 0.5\n gini = 0.451\n samples = 23\n value = [12, 23]\n class = Yes'),  
Text(1116.0, 543.5999999999999, 'gini = 0.311\n samples = 18\n value = [5, 21]\n class = Yes'),  
Text(3348.0, 543.5999999999999, 'gini = 0.346\n samples = 5\n value = [7, 2]\n class = No')]
```



In []: