

```
In [2]: import pandas as pd
import numpy as np
df=pd.read_csv(r"C:\Users\evang\Downloads\DATASETS\archive.zip")
df
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns



```
In [3]: df.describe()
```

```
Out[3]:
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
In [4]: df.isnull().sum()
```

```
Out[4]: 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 [5]: for i in df.columns:
        print(i)
        print(df[i].nunique())
```

```
PassengerId
891
Survived
2
Pclass
3
Name
891
Sex
2
Age
88
SibSp
7
Parch
7
Ticket
681
Fare
248
Cabin
147
Embarked
3
```

```
In [6]: df1=df.copy()
```

```
In [7]: df1['Age']=df1['Age'].interpolate()
df1.isnull().sum()
```

```
Out[7]: PassengerId      0
Survived      0
Pclass      0
Name      0
Sex      0
Age      0
SibSp      0
Parch      0
Ticket      0
Fare      0
Cabin      687
Embarked      2
dtype: int64
```

```
In [8]: df.corr()
```

```
Out[8]:
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

```
In [9]: df.columns
```

```
Out[9]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',  
              'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],  
              dtype='object')
```

SUMMARIZED ACCORDING TO ECONOMIC STATUS (CLASS) , SEX , AGE

```
In [10]: from sklearn.preprocessing import LabelEncoder  
le_Sex=LabelEncoder()  
df1['Sex']=le_Sex.fit_transform(df1['Sex'])
```

```
In [11]: x=df1.drop(['PassengerId', 'Survived', 'Name', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'], axis=1)  
x.isnull().sum()
```

```
Out[11]: Pclass      0  
Sex            0  
Age            0  
dtype: int64
```

```
In [12]: y=df1['Survived']  
from sklearn.model_selection import train_test_split  
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)  
X_train.isnull().sum()
```

```
Out[12]: Pclass      0  
Sex            0  
Age            0  
dtype: int64
```

```
In [13]: from sklearn.tree import DecisionTreeClassifier
```

create a decision tree classifier object

```
In [14]: clf=DecisionTreeClassifier()
```

fit the classifier on the training data

```
In [15]: clf.fit(X_train,y_train)
```

```
Out[15]: > DecisionTreeClassifier
```

make predictions on the test data

```
In [16]: y_predict=clf.predict(X_test)
```

```
In [17]: y_predict
```

```
Out[17]: array([0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1,
                0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1,
                0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
                1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0,
                0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0,
                0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0,
                0, 1, 1], dtype=int64)
```

```
In [19]: clf.score(X_train,y_train)
         clf.score(X_test,y_test)
```

```
Out[19]: 0.7653631284916201
```

```
In [ ]:
```