Excercise:

CSV file is available to download at
https://github.com/codebasics/py/blob/master/ML/9_decision_tree/Exercise/titanic.csv
(https://github.com/codebasics/py/blob/master/ML/9_decision_tree/Exercise/titanic.csv)

1. In this file using following columns build a model to predict if person would survive or not,

a. Pclass
b. Sex
c. Age
d. Fare

2. Define Survived as your target variable.
3. Check if there is missing data & if there is any fill it with mean.
4. Convert the text data into number by using map method.
5. Utilize Train Test and Split method.
6. Calculate the score of your model

Solution for Excercise

In [1]:
```
import pandas as pd
df = pd.read_csv('D:/Data_Science/My Github/Machine-Learning-with-Python/9. Decisi
df.head()
```

Out[1]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN

In [2]:
```
target = df['Survived']
target
```

Out[2]:
```
0      0
1      1
2      1
3      1
4      0
      ..
886    0
887    1
888    0
889    1
890    0
Name: Survived, Length: 891, dtype: int64
```

In [3]:
```
df
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N
...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	N
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	N
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C1
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	N

891 rows × 12 columns

In [4]:
```
df['Pclass'].isnull().values.any()
```

Out[4]:
```
False
```

Which means there is no missing data in 'Pclass'.

In [5]:
```
df['Sex'].isnull().values.any()
```

Out[5]:
```
False
```

In [6]:
```
df['Age'].isnull().values.any()
```

Out[6]:
```
True
```

which mean there is Nan value in 'Age' that we should handle it.

In [7]:
```
df['Fare'].isnull().values.any()
```

Out[7]:
```
False
```

In [9]:
```
#Replace mising data by mean
df['Age'] = df['Age'].fillna(df['Age'].mean())
```

In [10]:
```
df['Age']
```

Out[10]:
```
0      22.000000
1      38.000000
2      26.000000
3      35.000000
4      35.000000
         ...
886    27.000000
887    19.000000
888    29.699118
889    26.000000
890    32.000000
Name: Age, Length: 891, dtype: float64
```

In [11]:
```
# Check for missing value
df['Age'].isnull().values.any()
```

Out[11]:
```
False
```

In [12]:
```
inputs = df.drop(['PassengerId','Survived','Name','SibSp','Parch','Ticket','Cabin'
inputs
```

Out[12]:

	Pclass	Sex	Age	Fare
0	3	male	22.000000	7.2500
1	1	female	38.000000	71.2833
2	3	female	26.000000	7.9250
3	1	female	35.000000	53.1000
4	3	male	35.000000	8.0500
...	...	...	...	...
886	2	male	27.000000	13.0000
887	1	female	19.000000	30.0000
888	3	female	29.699118	23.4500
889	1	male	26.000000	30.0000
890	3	male	32.000000	7.7500

In [13]:
```
# Convert text data into numbers
inputs.Sex = inputs.Sex.map({'male':0 ,'female':1})
```

In [14]:
```
inputs.head()
```

Out[14]:

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

train and test split method

In [15]:
```
from sklearn.model_selection import train_test_split
```

In [20]:
```
X_train,X_test,y_train,y_test=train_test_split(inputs,target,test_size=0.2)
```

In [21]:
```
len(X_test)
```

Out[21]: 179

In [22]:
```
len(X_train)
```

Out[22]: 712

In [23]:
```
from sklearn import tree
model = tree.DecisionTreeClassifier()
```

In [24]:
```
model.fit(X_train,y_train)
```

Out[24]:
```
DecisionTreeClassifier()
```

In [25]:
```
model.score(X_test,y_test)
```

Out[25]:
```
0.7653631284916201
```

Is a woman in a third Pclass at the age of 28 and with the price of 71 survived?

In [26]:
```
model.predict([[3,1,28,71]])
```

Out[26]:
```
array([0], dtype=int64)
```

Date	Author
2021-09-13	**Ehsan Zia**