Train Test Split

Data(100%) ==> Independent_Data(x)(90%) + Dependent_data/Target_data(y)(10%)

Independent_Data(x) = Training_data(x_train), Testing_data(x_test)

Dependent_Data(Y) = Training_data(y_train), Testing_data(y_test)

Data = x,y ==> (x_train, x_test), (y_train, y_test)

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
df = pd.read_csv("C:\\Users\\91636\\OneDrive\\Desktop\\Regex ML\\Data\\covid_toy.csv")
```

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

Out[3]:

	age	gender	fever	cough	city	has_covid
0	60	Male	103.0	Mild	Kolkata	No
1	27	Male	100.0	Mild	Delhi	Yes
2	42	Male	101.0	Mild	Delhi	No
3	31	Female	98.0	Mild	Kolkata	No
4	65	Female	101.0	Mild	Mumbai	No

In [4]:
```
df.shape # Total data 100 rows and 6 columns
```

Out[4]:
```
(100, 6)
```

In [5]:
```
# Step-1   Divide data into Independent and Dependent data
x = df.drop(columns = ['has_covid'], axis=1) # Independent data
y = df['has_covid'] # Target Data
```

In [6]:
```
print("Independent Data Shape = ",x.shape)
print("Dependent Data Shape = ",y.shape)
```

```
Independent Data Shape =  (100, 5)
Dependent Data Shape =  (100,)
```

In [7]:
```
from sklearn.model_selection import train_test_split
```

In [8]:
```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.2, random_state = 42)

# random_state work as a seed(), It will fix random numbers.
# test_size = 0.2 = 20% data test, 80% data training

# train_test_split ==> models(training_data on trained), test_data check performance
# class 100 students ==> 80 students(training_data), 80 students ==> 20 students(test_data) ==> Performance accurate
# 20 students ==> Performance poor
```

In [9]:
```
print("X_train data shape = ",x_train.shape)
print("X_test data shape = ",x_test.shape)
print("Y_train data shape = ",y_train.shape)
print("Y_test data shape = ",y_test.shape)
```

```
X_train data shape =  (80, 5)
X_test data shape =  (20, 5)
Y_train data shape =  (80,)
Y_test data shape =  (20,)
```