SciKit Learn Preprocessing Overview

In [3]:
```
import numpy as np
```

In [4]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [11]:
```
data = np.random.randint(0,100,(10,2))
```

In [12]:
```
data
```

Out[12]:
```
array([[56, 95],
       [68, 83],
       [ 5, 62],
       [84,  1],
       [78,  4],
       [24, 56],
       [72, 42],
       [25, 32],
       [91, 20],
       [22, 48]])
```

In [13]:
```
scaler_model = MinMaxScaler()
```

In [14]:
```
scaler_model.fit(data)
```

```
C:\Users\Marcial\Anaconda3\envs\tf_1_3\lib\site-packages\sklearn\utils
\validation.py:444: DataConversionWarning: Data with input dtype int32
was converted to float64 by MinMaxScaler.
  warnings.warn(msg, DataConversionWarning)
```

Out[14]:
```
MinMaxScaler(copy=True, feature_range=(0, 1))
```

In [15]:
```
scaler_model.transform(data)
```

Out[15]:
```
array([[ 0.59302326,  1.        ],
       [ 0.73255814,  0.87234043],
       [ 0.        ,  0.64893617],
       [ 0.91860465,  0.        ],
       [ 0.84883721,  0.03191489],
       [ 0.22093023,  0.58510638],
       [ 0.77906977,  0.43617021],
       [ 0.23255814,  0.32978723],
       [ 1.        ,  0.20212766],
       [ 0.19767442,  0.5       ]])
```

In [16]:
```
# In one step
result = scaler_model.fit_transform(data)
```

```
C:\Users\Marcial\Anaconda3\envs\tf_1_3\lib\site-packages\sklearn\utils
\validation.py:444: DataConversionWarning: Data with input dtype int32
was converted to float64 by MinMaxScaler.
  warnings.warn(msg, DataConversionWarning)
```

In [18]:
```
result
```

Out[18]:
```
array([[ 0.59302326,  1.        ],
       [ 0.73255814,  0.87234043],
       [ 0.        ,  0.64893617],
       [ 0.91860465,  0.        ],
       [ 0.84883721,  0.03191489],
       [ 0.22093023,  0.58510638],
       [ 0.77906977,  0.43617021],
       [ 0.23255814,  0.32978723],
       [ 1.        ,  0.20212766],
       [ 0.19767442,  0.5       ]])
```

In [20]:
```
import pandas as pd
```

In [21]:
```
data = pd.DataFrame(data=np.random.randint(0,101,(50,4)),columns=['f1',
'f2','f3','label'])
```

In [24]:
```
data.head()
```

Out[24]:

	f1	f2	f3	label
0	79	12	96	29
1	35	75	39	84
2	5	61	62	87
3	97	85	76	69
4	67	65	30	64

In [25]:
```
x = data[['f1','f2','f3']] # Alternatively x = data.drop('label',axis=1)
y = data['label']
```

In [26]:
```
from sklearn.model_selection import train_test_split
```

In [27]:
```
X_train, X_test, y_train, y_test = train_test_split(x,y,test_size=0.3,ra
ndom_state=101)
```

In [28]:
```
X_train.shape
```

Out[28]:
```
(35, 3)
```

In [29]:
```
X_test.shape
```

Out[29]: (15, 3)

In [30]:
```
y_train.shape
```

Out[30]: (35,)

In [31]:
```
y_test.shape
```

Out[31]: (15,)

Great Job!