```
array([[0.80377277, 0.55160877, 0.22064351, 0.0315205 ],
       [0.82813287, 0.50702013, 0.23660939, 0.03380134],
       [0.80533308, 0.54831188, 0.2227517 , 0.03426949],
       [0.80003025, 0.53915082, 0.26087943, 0.03478392],
       [0.790965  , 0.5694948 , 0.2214702 , 0.0316386 ]])
```

Binarization

Binarization is a transformation technique for converting a dataset into binary values by setting a cutoff or threshold. All values above the threshold are set to 1, while those below are set to 0. This technique is useful for converting a dataset of probabilities into integer values or in transforming a feature to reflect some categorization. Scikit-learn implements binarization with the **Binarizer** module.

```
# import packages
from sklearn import datasets
from sklearn.preprocessing import Binarizer

# load dataset
data = datasets.load_iris()
# separate features and target
X = data.data
y = data.target

# print first 5 rows of X before binarization
X[0:5,:]
'Output':
array([[5.1, 3.5, 1.4, 0.2],
       [4.9, 3. , 1.4, 0.2],
       [4.7, 3.2, 1.3, 0.2],
       [4.6, 3.1, 1.5, 0.2],
       [5. , 3.6, 1.4, 0.2]])

# binarize X
scaler = Binarizer(threshold = 1.5).fit(X)
binarize_X = scaler.transform(X)
```

```
# print first 5 rows of X after binarization
binarize_X[0:5,:]
'Output':
array([[1., 1., 0., 0.],
       [1., 1., 0., 0.],
       [1., 1., 0., 0.],
       [1., 1., 0., 0.],
       [1., 1., 0., 0.]])
```

Encoding Categorical Variables

Most machine learning algorithms do not compute with non-numerical or categorical variables. Hence, encoding categorical variables is the technique for converting non-numerical features with labels into a numerical representation for use in machine learning modeling. Scikit-learn provides modules for encoding categorical variables including the **LabelEncoder** for encoding labels as integers, **OneHotEncoder** for converting categorical features into a matrix of integers, and **LabelBinarizer** for creating a one-hot encoding of target labels.

LabelEncoder is typically used on the target variable to transform a vector of hashable categories (or labels) into an integer representation by encoding label with values between 0 and the number of categories minus 1. This is further illustrated in Figure 18-1.

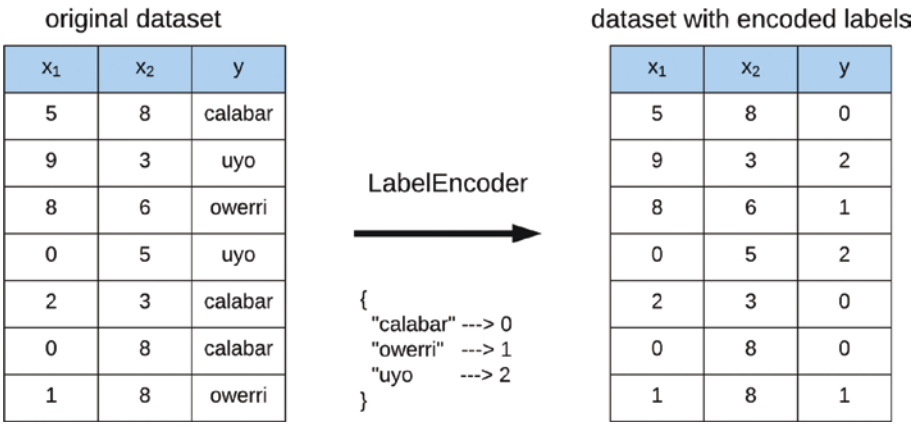

Figure 18-1. *LabelEncoder*