

1. Reading Data

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

```
In [2]: ATT_FILE = str(input("Name of the attributes file: "))
```

Name of the attributes file: OceanProximityPreparedCleanAttributes.csv

```
In [3]: LABEL_FILE = str(input("Name of the label file (Ocean Proximity): "))  
        TRAIN_RATE = 0.8
```

Name of the label file (Ocean Proximity): OceanProximityOneHotEncodedClasses.csv

```
In [4]: attributes = pd.read_csv(ATT_FILE)  
        label = pd.read_csv(LABEL_FILE)
```

```
In [5]: attributes.head()
```

```
Out[5]:
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
0	-0.940239	0.766206	0.176471	-0.877308	-0.867474	
1	-0.569721	0.479277	-0.137255	-0.976601	-0.970205	
2	0.424303	-0.753454	-0.137255	-0.678468	-0.629112	
3	-0.043825	-0.113709	0.215686	-0.900249	-0.842644	
4	0.021912	-0.651435	-0.372549	-0.830612	-0.763811	

	population	households	median_income	median_house_value
0	-0.943833	-0.868443	-0.584047	-0.472163
1	-0.985538	-0.970728	-0.769576	-0.843295
2	-0.888954	-0.630653	-0.747907	-0.680822
3	-0.885143	-0.829962	-0.805837	-0.831749
4	-0.899437	-0.778984	-0.458035	-0.108453

```
In [6]: label.head()
```

```
Out[6]:
```

	<1H OCEAN	INLAND	NEAR BAY	NEAR OCEAN
0	0.0	0.0	0.0	1.0
1	0.0	1.0	0.0	0.0
2	1.0	0.0	0.0	0.0
3	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	1.0

```

In [7]: attributes.shape

Out[7]: (20428, 9)

In [8]: label.shape

Out[8]: (20428, 4)

In [9]: n_instances = attributes.shape[0]
        n_train = int(n_instances*TRAIN_RATE)
        n_dev = int((n_instances-n_train)/2)
        n_test = n_instances-n_train-n_dev

In [10]: n_train

Out[10]: 16342

In [11]: n_dev

Out[11]: 2043

In [12]: n_test

Out[12]: 2043

In [13]: n_train+n_dev+n_test

Out[13]: 20428

In [14]: x_train = attributes.values[:n_train]
        t_train = label.values[:n_train]

In [15]: x_train[:5]

Out[15]: array([[ -0.94023904,  0.76620616,  0.17647059, -0.8773081 , -0.86747362,
                  -0.94383251, -0.86844269, -0.58404712, -0.472163  ],
                 [-0.56972112,  0.47927736, -0.1372549 , -0.97660105, -0.97020484,
                  -0.98553771, -0.9707285 , -0.76957559, -0.84329549],
                 [ 0.42430279, -0.75345377, -0.1372549 , -0.67846788, -0.62911235,
                  -0.88895429, -0.63065285, -0.74790693, -0.68082193],
                 [-0.0438247 , -0.11370882,  0.21568627, -0.90024925, -0.84264432,
                  -0.88514252, -0.82996218, -0.80583716, -0.83174915],
                 [ 0.02191235, -0.65143464, -0.37254902, -0.83061193, -0.7638113 ,
                  -0.89943664, -0.77898372, -0.45803506, -0.10845316]])

In [16]: t_train[:5]

Out[16]: array([[0., 0., 0., 1.],
                 [0., 1., 0., 0.],
                 [1., 0., 0., 0.],
                 [0., 1., 0., 0.],
                 [0., 0., 0., 1.]])

```

```

In [17]: x_dev = attributes.values[n_train:n_train+n_dev]
         t_dev = label.values[n_train:n_train+n_dev]

In [18]: x_dev[:5]

Out[18]: array([[ -0.58565737,  0.12008502,  0.52941176, -0.95533852, -0.8603352 ,
                  -0.96754393, -0.87534945, -0.93419401, -0.5719564 ],
                 [ 0.33067729, -0.78746015, -0.45098039, -0.8768503 , -0.86530106,
                  -0.92359651, -0.86120704, -0.39439456, -0.31216366],
                 [-0.59163347,  0.06057386,  0.64705882, -0.91118572, -0.88671633,
                  -0.94500967, -0.87830949, -0.4466697 ,  0.33113266],
                 [ 0.23904382, -0.68119022,  0.60784314, -0.88793937, -0.8603352 ,
                  -0.92213907, -0.86120704, -0.46694528, -0.19010231],
                 [-0.56374502,  0.09670563,  0.37254902, -0.91266087, -0.9062694 ,
                  -0.95330587, -0.89804309, -0.45689025, -0.26391644]])

In [19]: t_dev[:5]

Out[19]: array([[0., 0., 1., 0.],
                 [1., 0., 0., 0.],
                 [0., 0., 0., 1.],
                 [1., 0., 0., 0.],
                 [0., 0., 1., 0.]])

In [20]: x_test = attributes.values[n_train+n_dev:n_instances]
         t_test = label.values[n_train+n_dev:n_instances]

In [21]: x_test[:5]

Out[21]: array([[ 0.17330677, -0.69606801, -0.33333333, -0.52042322, -0.23277467,
                  -0.80913142, -0.34155567, -0.21756941,  1.          ],
                 [ 0.2250996 , -0.73645058,  1.          , -0.86469302, -0.72315332,
                  -0.83099302, -0.74938333, -0.74750693, -0.39175096],
                 [-0.62350598,  0.09883103,  0.88235294, -0.91306781, -0.89230292,
                  -0.93637714, -0.89343858, -0.67240452, -0.13402007],
                 [ 0.22709163, -0.68331562,  0.92156863, -0.93992573, -0.92396027,
                  -0.95246504, -0.92435455, -0.61876388, -0.47381248],
                 [ 0.3247012 , -0.75770457,  0.09803922, -0.71270156, -0.66387337,
                  -0.77678747, -0.6530176 , -0.68010096, -0.39175096]])

In [22]: t_test[:5]

Out[22]: array([[1., 0., 0., 0.],
                 [0., 0., 0., 1.],
                 [0., 0., 0., 1.],
                 [1., 0., 0., 0.],
                 [1., 0., 0., 0.]])

In [23]: x_train.shape

```

```
Out[23]: (16342, 9)
```

```
In [24]: t_train.shape
```

```
Out[24]: (16342, 4)
```

```
In [25]: x_dev.shape
```

```
Out[25]: (2043, 9)
```

```
In [26]: t_dev.shape
```

```
Out[26]: (2043, 4)
```

```
In [27]: x_test.shape
```

```
Out[27]: (2043, 9)
```

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

```
Out[28]: (2043, 4)
```