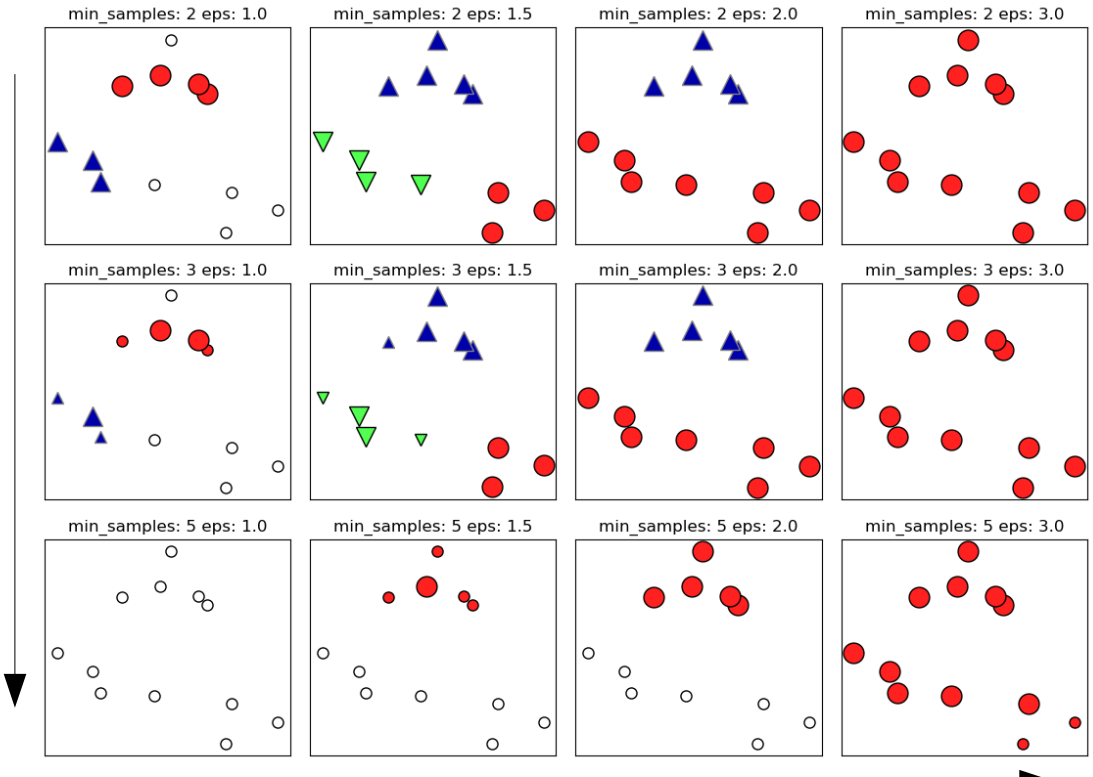
# Q5. DBSCAN

```
import numpy as np
import os

import matplotlib as mpl
import matplotlib.pyplot as plt
from matplotlib.cm import get_cmap

from sklearn.datasets import make_moons
from sklearn.cluster import DBSCAN

import mglearn

############################
### DBSCAN example display ###
############################

mglearn.plots.plot_dbscan()  # DBSCAN results according to the variations in min_samples and eps (distance threshold for cluster merging)
plt.show()

#################################################################
### DBSCAN comparison under different parameters in moon dataset ###
#################################################################

### Prepare moon dataset ##########################################################################################################################
X, y = make_moons(n_samples=1000, noise=0.05, random_state=42)  # Moon dataset with 1000 samples

### DBSCAN clustering of moon dataset using different parameters ##################################################################################

# DBSCAN 1 (eps : 0.05 / min_sample : 5 / distance metric : euclidean) ############################################################################
dbscan1 = DBSCAN(eps=0.05, min_samples=5, metric='euclidean')
db_cluster1 = dbscan1.fit_predict(X)    # Fit and cluster the dataset using DBSCAN 1

print('DBSCAN core sample num : {}'.format(len(dbscan1.core_sample_indices_)))
print('DBSCAN Labels : {}'.format(np.unique(dbscan1.labels_)))

# Plot clustering results of DBSCAN 1
plt.subplot(121)
cmap = get_cmap('Accent')   # Prepare color map / Each cluster uses an distinctive color
legend = []
for label in np.unique(dbscan1.labels_):

    # Plot only the points that correspond to certain cluster label using X[db_cluster1==label]
    # Assign the color to the points in the dataset according to their labels
    plt.scatter(X[db_cluster1==label][:, 0], X[db_cluster1==label][:, 1], c=cmap.colors[label], label='Cluster '+ str(label))
    legend.append('Cluster '+ str(label))

plt.legend(legend, loc='best')
plt.title('DBSCAN (eps : {} / min_sample : {})'.format(dbscan1.eps, dbscan1.min_samples))

# DBSCAN 2 (eps : 0.2 / min_sample : 5 / distance metric : euclidean) ############################################################################
dbscan2 = DBSCAN(eps=0.2, min_samples=5, metric='euclidean')
db_cluster2 = dbscan2.fit_predict(X)    # Fit and cluster the dataset using DBSCAN 1

print('DBSCAN core sample num : {}'.format(len(dbscan2.core_sample_indices_)))
print('DBSCAN Labels : {}'.format(np.unique(dbscan2.labels_)))

# Plot clustering results of DBSCAN 2
plt.subplot(122)
cmap = get_cmap('Accent')   # Prepare color map / Each cluster uses an distinctive color
legend = []
for label in np.unique(dbscan2.labels_):

    # Plot only the points that correspond to certain cluster label using X[db_cluster2==label]
    # Assign the color to the points in the dataset according to their labels
    plt.scatter(X[db_cluster2==label][:, 0], X[db_cluster2==label][:, 1], c=cmap.colors[label], label='Cluster '+ str(label))
    legend.append('Cluster '+ str(label))

plt.legend(legend, loc='best')
plt.title('DBSCAN (eps : {} / min_sample : {})'.format(dbscan2.eps, dbscan2.min_samples))

plt.show()
```

## DBSCAN 은 Distance Threshold (eps) 를 기준으로 국소적인 Cluster 를 만들고 주변 Cluster 가 최소 Cluster 구성조건 (min_samples) 과 Distance Threshold 를 만족할 경우 Merge 를 통해 Cluster 를 확장시켜나감

## 국소적인 Cluster 확장을 점진적으로 수행하여 전체 데이터셋을 Clustering 되게함

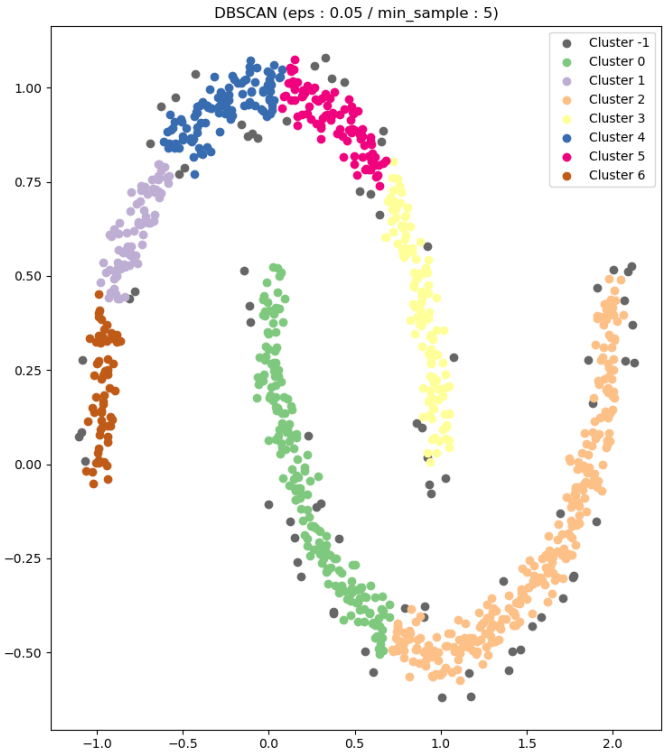


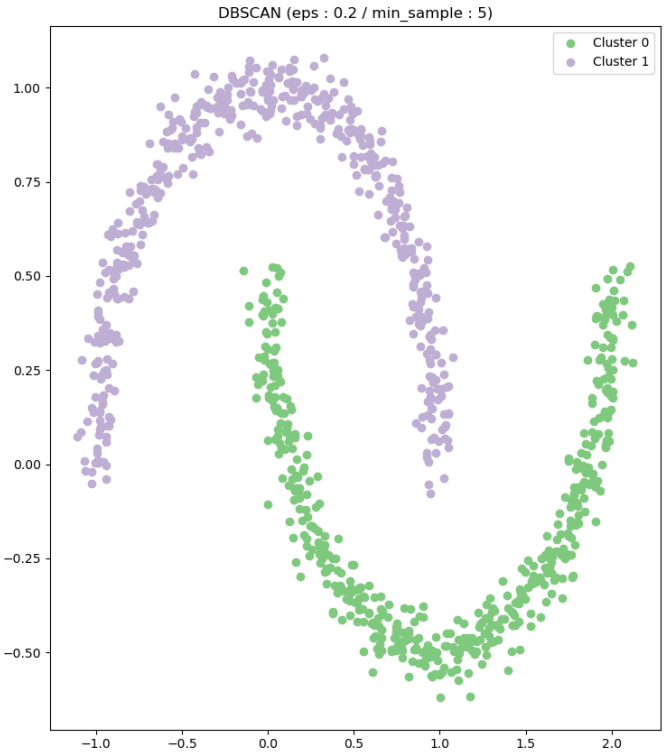

**Cluster 를 구성하는 최소 개수 기준인 min_samples 가 커질수록 Cluster 구성기준이 높아지기 때문에 Cluster 가 쉽게 생기지 않으며 Merge 를 통한 확장이 쉽게 안 됨**

**이로 인해 Cluster 가 생성이 안되어 Clustering 이 효과적이지 않게됨**

**Cluster 구성과 Merge 기준인 Distance Threshold (eps) 가 커질수록 더 쉽게 Cluster 구성과 Merge 가 발생하기 때문에 Cluster 가 쉽게 확장됨**

**이로 인해 Cluster 가 과도하게 확장되어 밀집 영역간 분리가 안될 수 있**




- **DBSCAN 에서 Distance Threshold (eps) 와 최소 Cluster 데이터 개수 (min_sample) 에 매우 민감하기 때문에 그에 따라 Clustering 결과가 매우 다름 .**
- **위 그래프에서는 min_sample 은 동일하나 eps 의 상태에 따라 생성되는 Cluster 개수 , Outlier 존재 여부 등이 결정되는 것을 볼 수 있음 .**
- **eps 가 너무 낮으면 Cluster 생성과 Merge 기준이 높아지기 때문에 여러 개의 국소적인 Cluster 가 하나의 Cluster 로 병합되지 못하는 것을 볼 수 있음 .**