An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

DBSCAN

fig ref: pierian data

DBSCAN

Density-based spatial clustering of applications with noise is a powerful technique which can be used for clustering and outlier detection.

- Intuition of DBSCAN
- DBSCAN vs. K-Means Clustering
- DBSCAN Hyperparameters Theory
- DBSCAN Hyperparameters Coding

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Theory and Intuition

Presented by:
Shreyas Shukla

fig ref: pierian data

DBSCAN stands for **D**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise.

fig ref: pierian data

Some Questions:

- How does DBSCAN work?
- Advantages and disadvantages of DBSCAN?
- How does it deal with outliers and noise?

fig ref: pierian data

Key Ideas

- DBSCAN focuses on using **density** of points as its main factor for assigning cluster labels.

- This creates the ability to find cluster segmentations that other algorithms have difficulty with.

fig ref: pierian data

Consider the following data set:

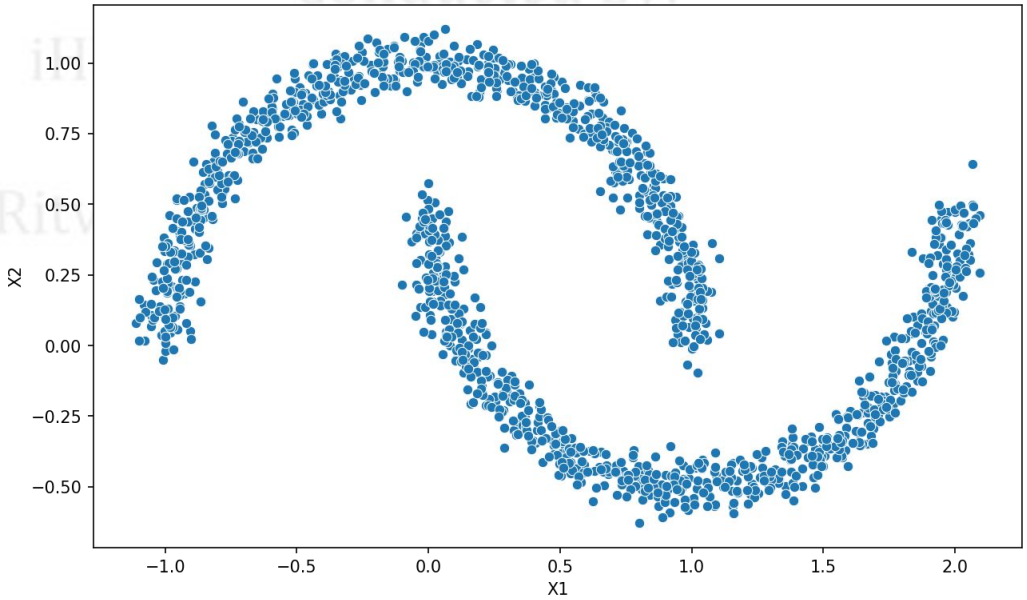


fig ref: pierian data

Cleary two “moon” shaped clusters:

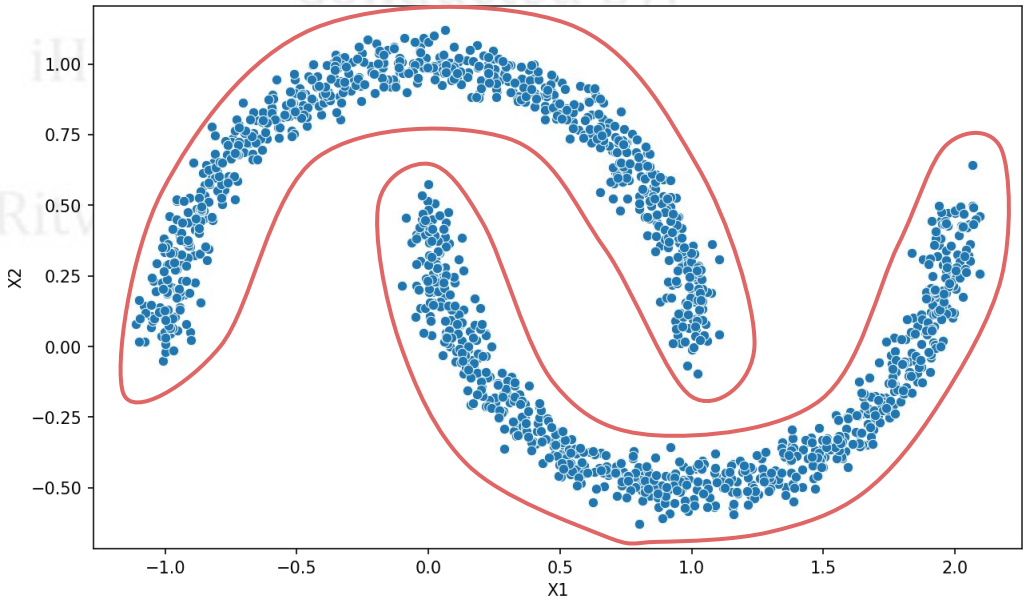


fig ref: pierian data

But distance based clustering has issues:

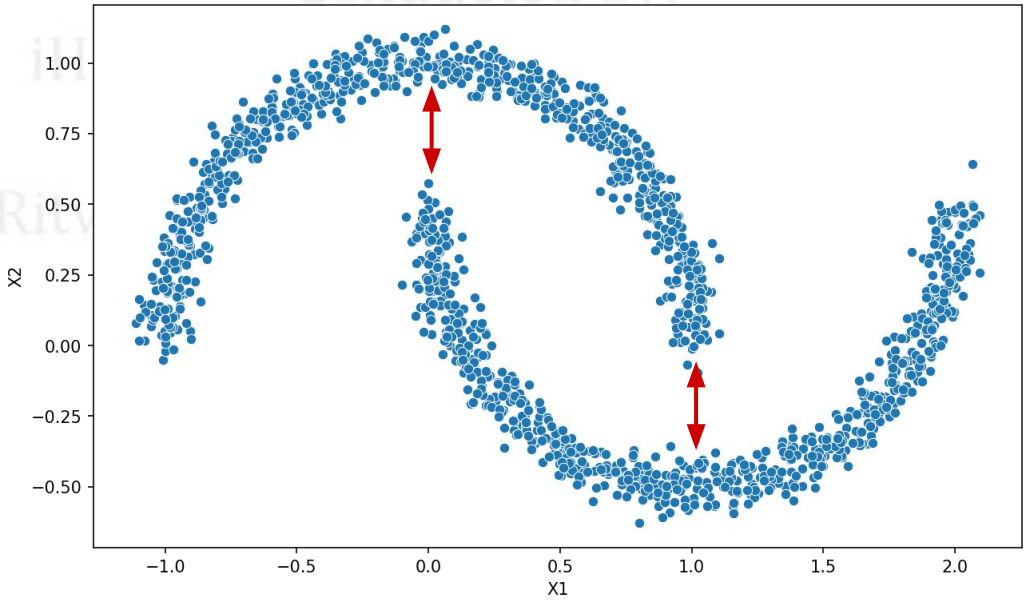


fig ref: pierian data

But distance based clustering has issues:

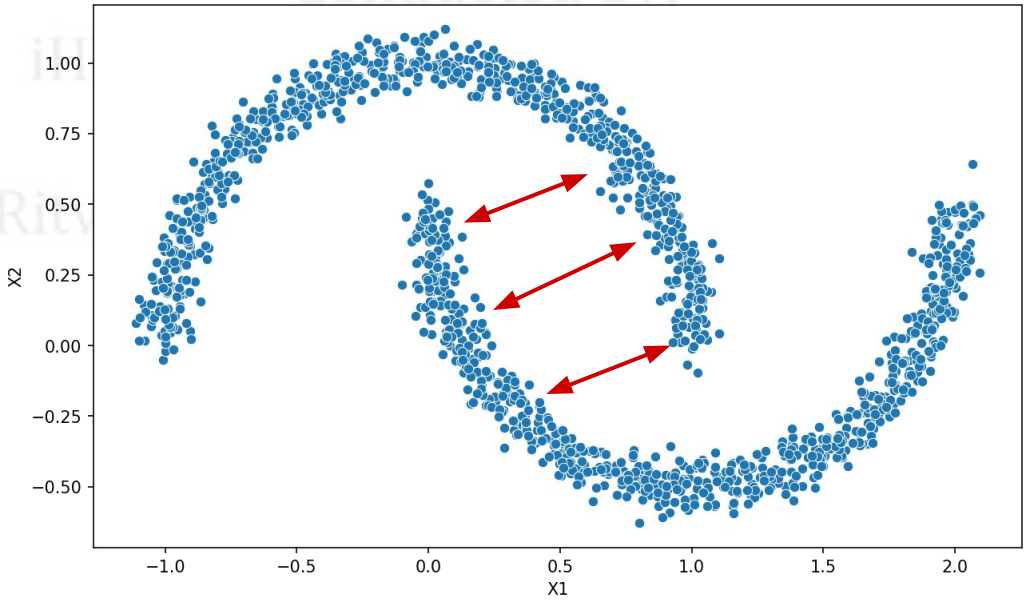


fig ref: pierian data

Results of K-Means:

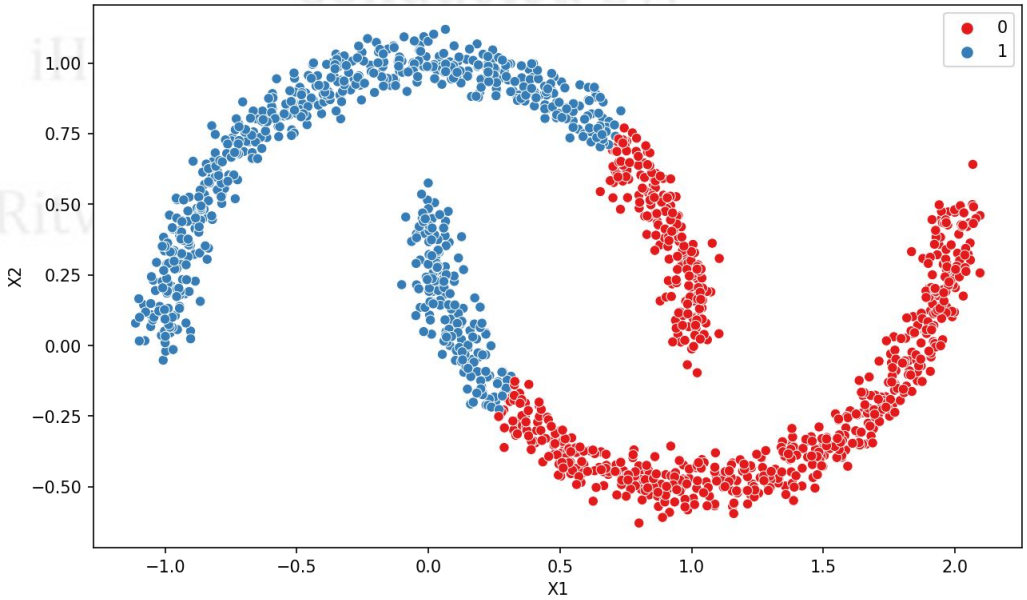


fig ref: pierian data

Results of DBSCAN:

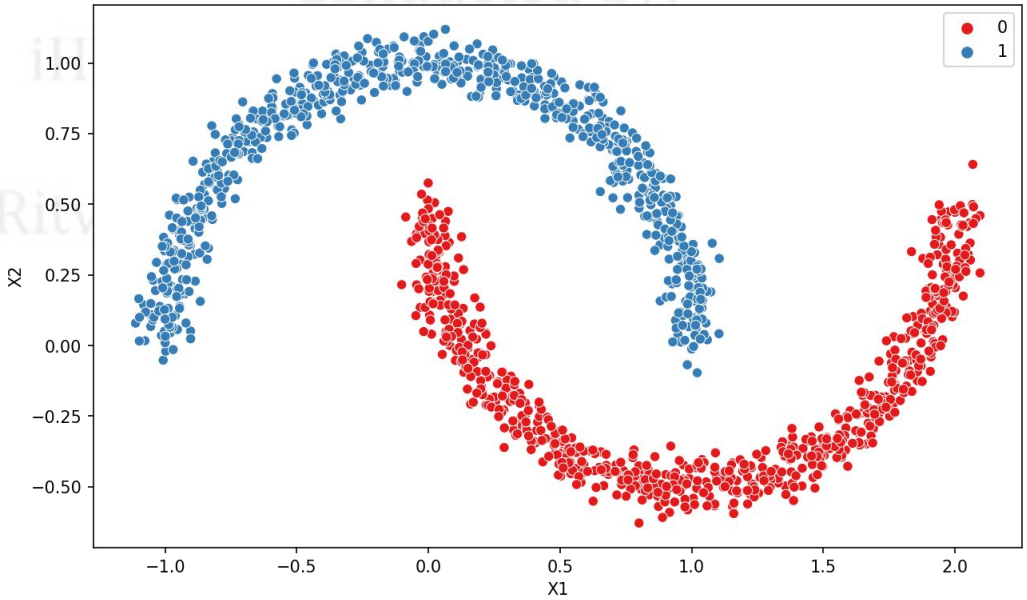


fig ref: pierian data

DBSCAN iterates through points and uses two key hyperparameters (epsilon and minimum number of points) to assign cluster labels.

Unlike K-Means, it focuses on density as the main factor for cluster assignment of points.

fig ref: pierian data

Key Hyperparameters:

- Epsilon:
  - Distance extended from a point.
- Minimum Number of Points:
  - Minimum number of points in an epsilon distance.

fig ref: pierian data

DBSCAN Point Types:

- Core
- Border
- Outlier

fig ref: pierian data

DBSCAN Point Types:

- Core
- Border
- Outlier

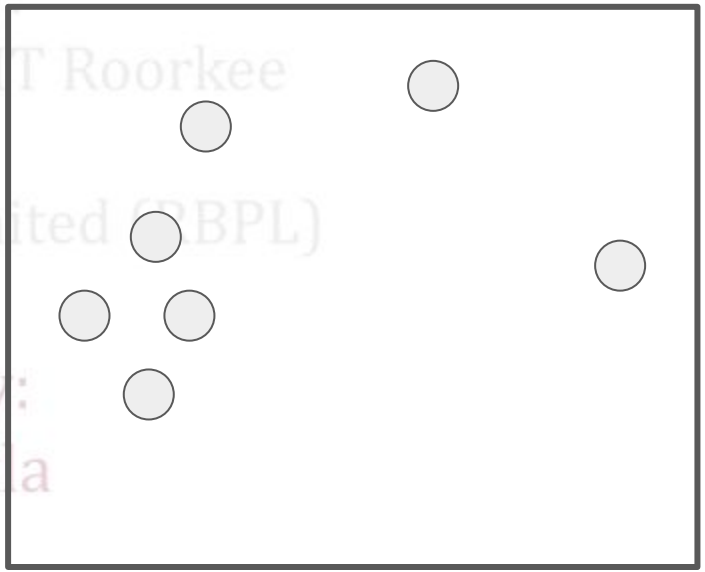


fig ref: pierian data

DBSCAN Point Types:

- Core

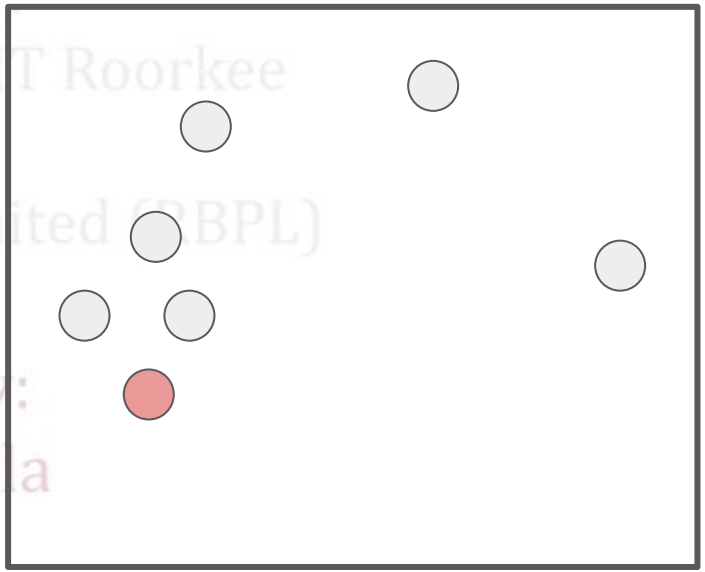


fig ref: pierian data

DBSCAN Point Types:

- Core

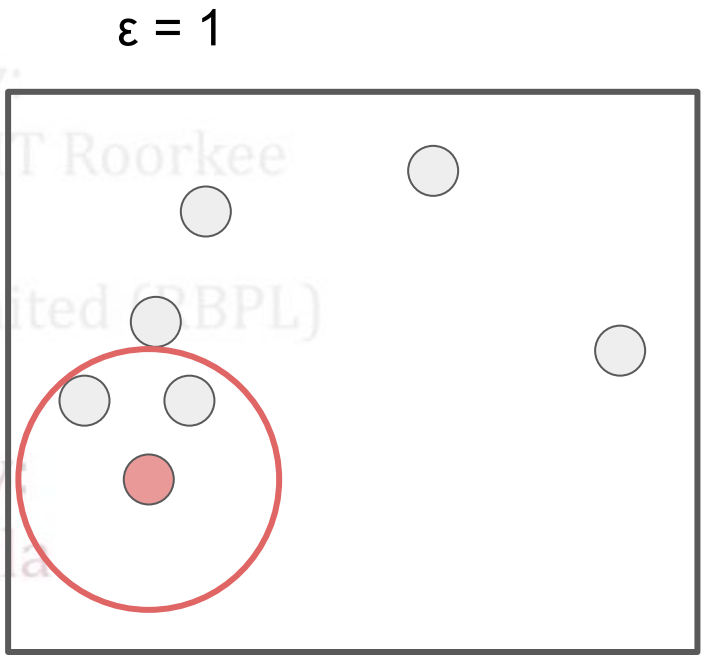


fig ref: pierian data

DBSCAN Point Types:

- Core

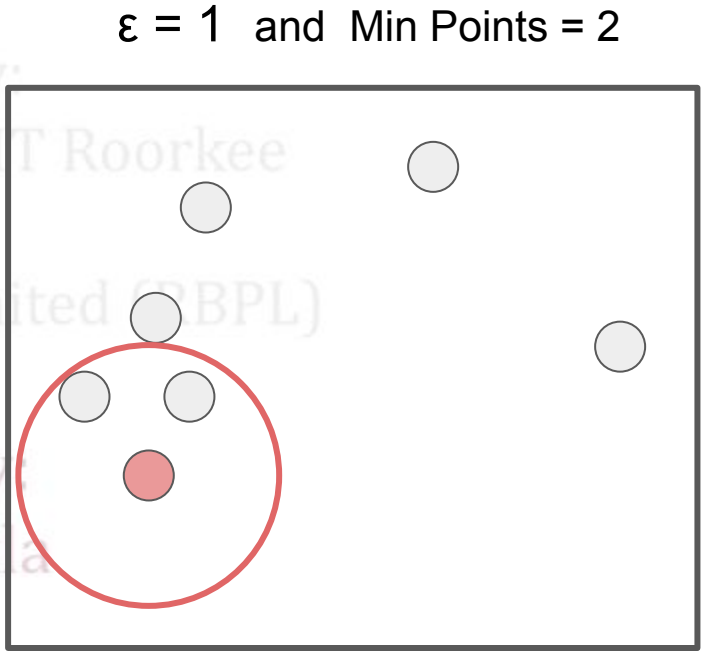


fig ref: pierian data

DBSCAN Point Types:

- Core

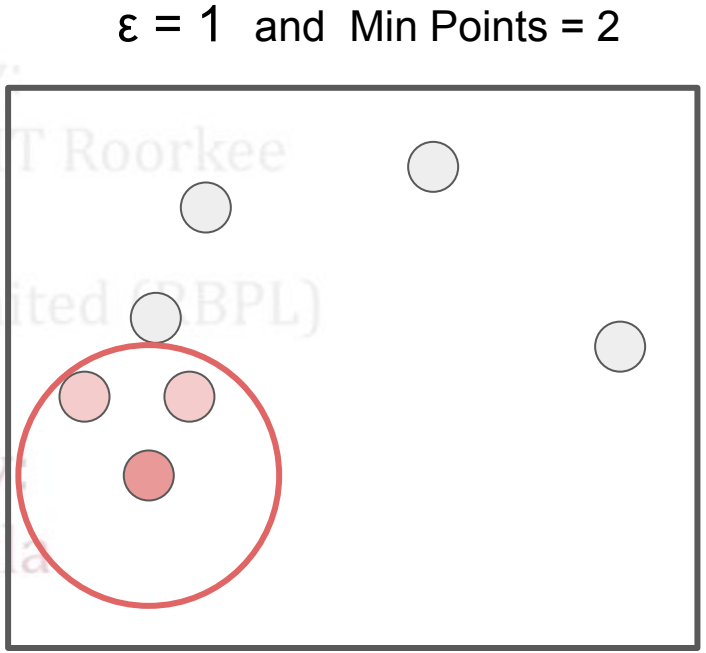


fig ref: pierian data

DBSCAN Point Types:

- Core:
  - Point with min. points in epsilon range.

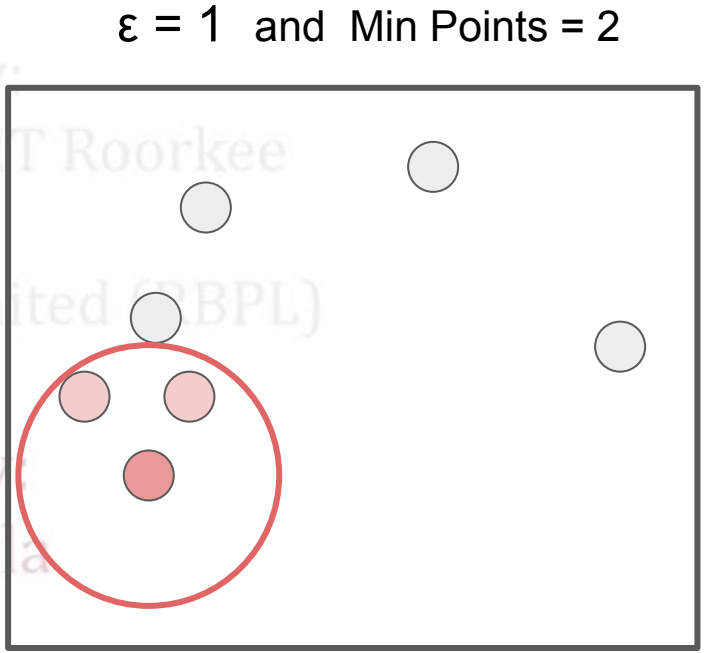


fig ref: pierian data

DBSCAN Point Types:

- Core:
  - Point with min. points in epsilon range.

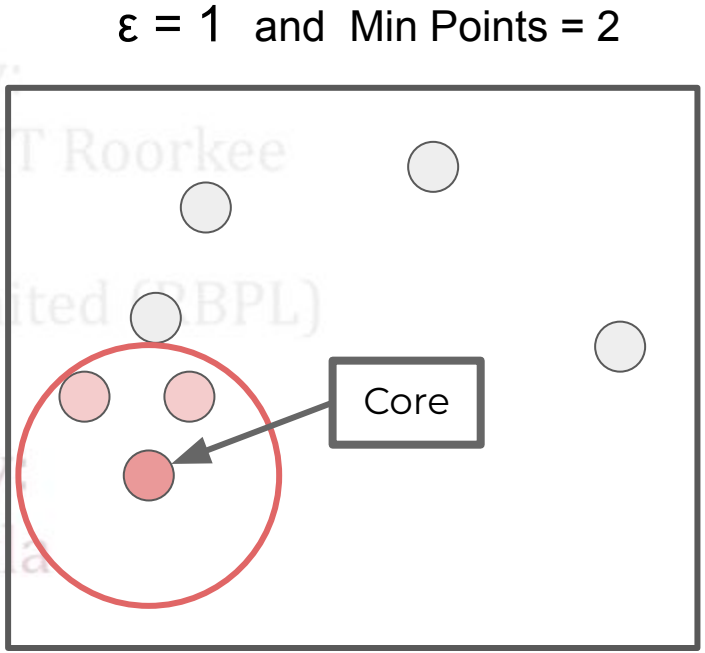


fig ref: pierian data

DBSCAN Point Types:

- Core:
  - Point with min. points in epsilon range (including itself).

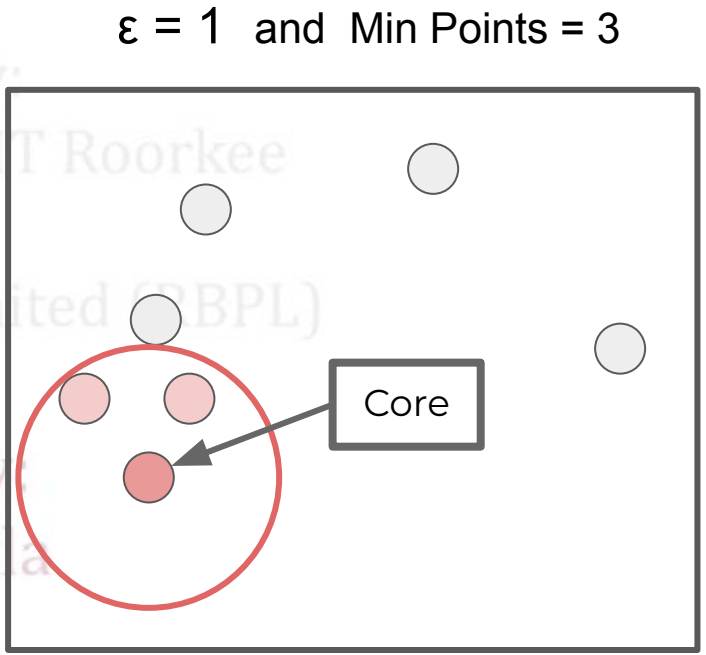


fig ref: pierian data

DBSCAN Point Types:

- Border:
  - In epsilon range of core point, but does not contain min. number of points.

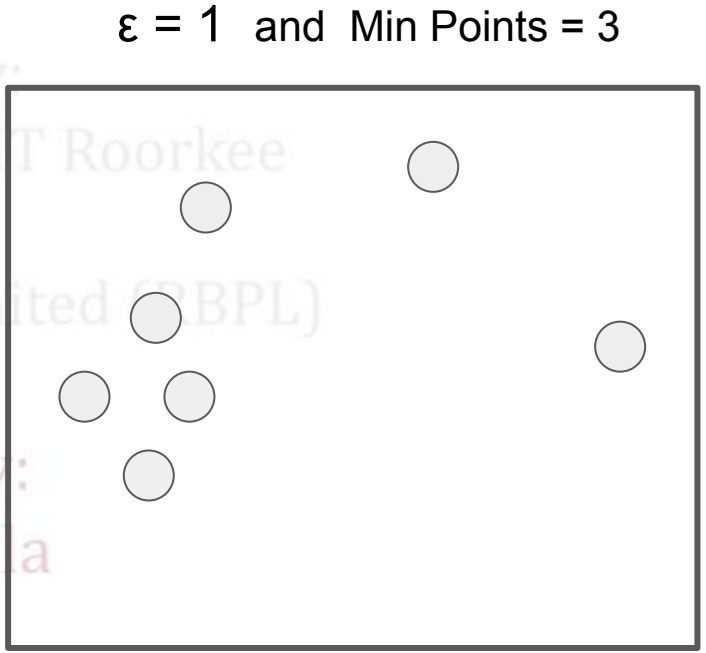


fig ref: pierian data

DBSCAN Point Types:

- Border:
  - In epsilon range of core point, but does not contain min. number of points.

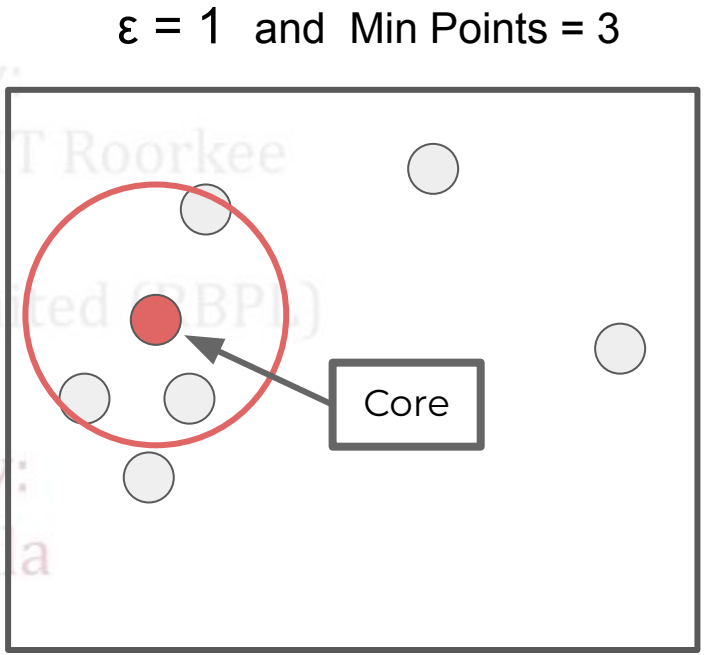


fig ref: pierian data

DBSCAN Point Types:

- Border:
  - In epsilon range of core point, but does not contain min. number of points.

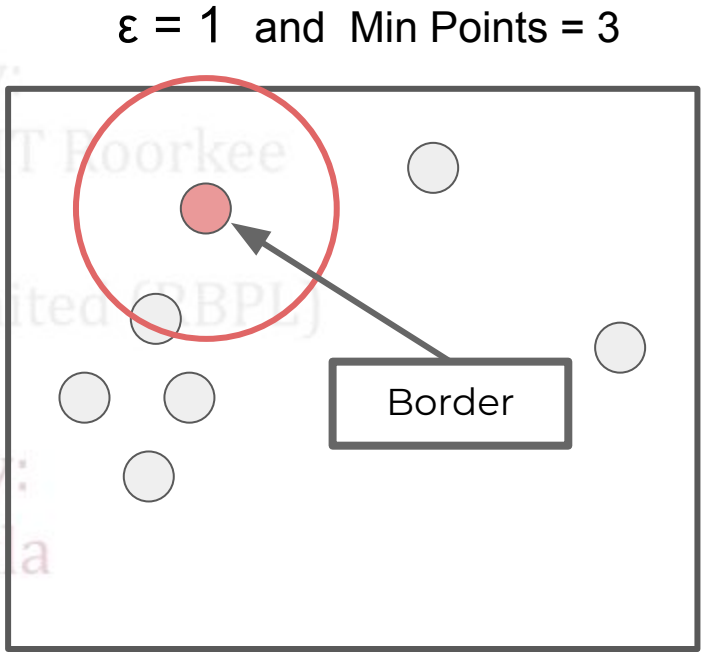


fig ref: pierian data

DBSCAN Point Types:

- Border:
  - In epsilon range of core point, but does not contain min. number of points.

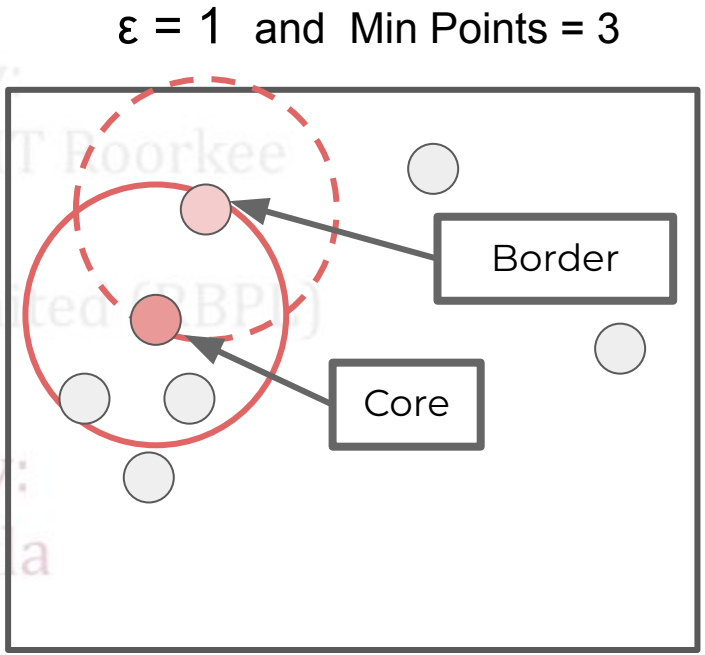


fig ref: pierian data

DBSCAN Point Types:

- Outlier:
  - Can not be "reached" by points in a cluster assignment.

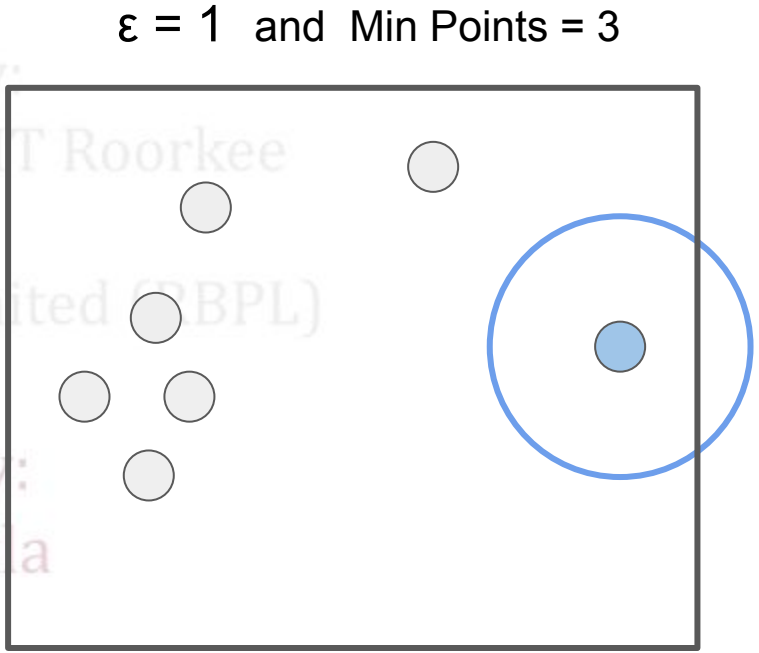


fig ref: pierian data

Let’s review the actual process of DBSCAN for assigning clusters.

fig ref: pierian data

DBSCAN Procedure:

- Pick a random point not yet assigned.
- Determine the point type.
- Once a **core** point has been found, add all directly reachable points to the same cluster as core.
- Repeat until all points have been assigned to a cluster or as an outlier.

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Rityu Bharat Private Limited (RBPL)

Coding Example on Data Sets

Presented by:
Shreyas Shukla

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Key Hyperparameters

Presented by:
Shreyas Shukla

fig ref: pierian data

Two key hyperparameters for DBSCAN:

- Epsilon:
  - Distance extended from a point to search for Min. Number of Points.
- Min. Number of Points:
  - Min. Number of Points within Epsilon distance to be a core point.

fig ref: pierian data

Adjusting these hyperparameters have two main outcomes:

- Changing number of clusters.
- Changing what is an outlier point.

fig ref: pierian data

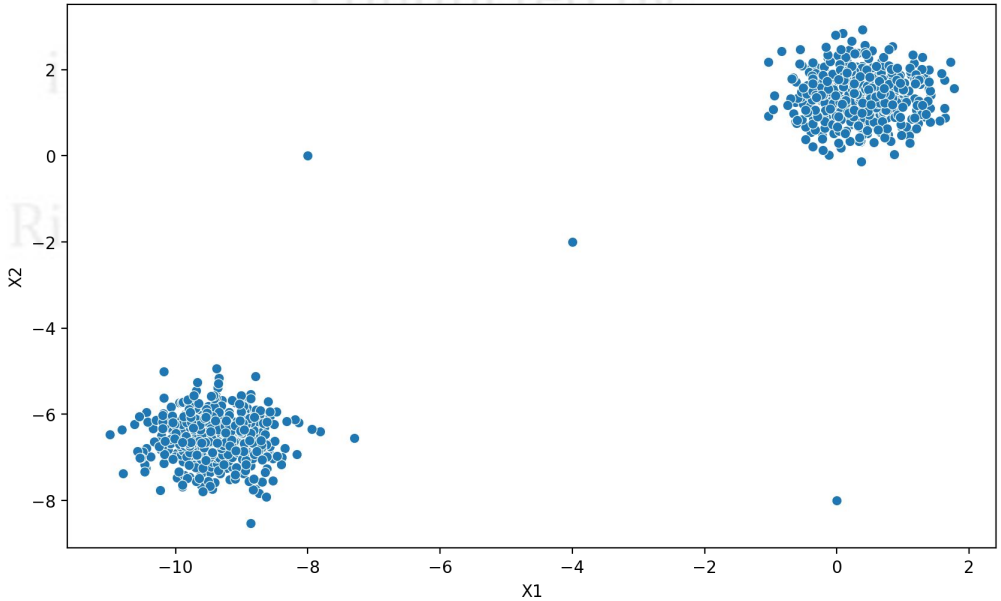


fig ref: pierian data

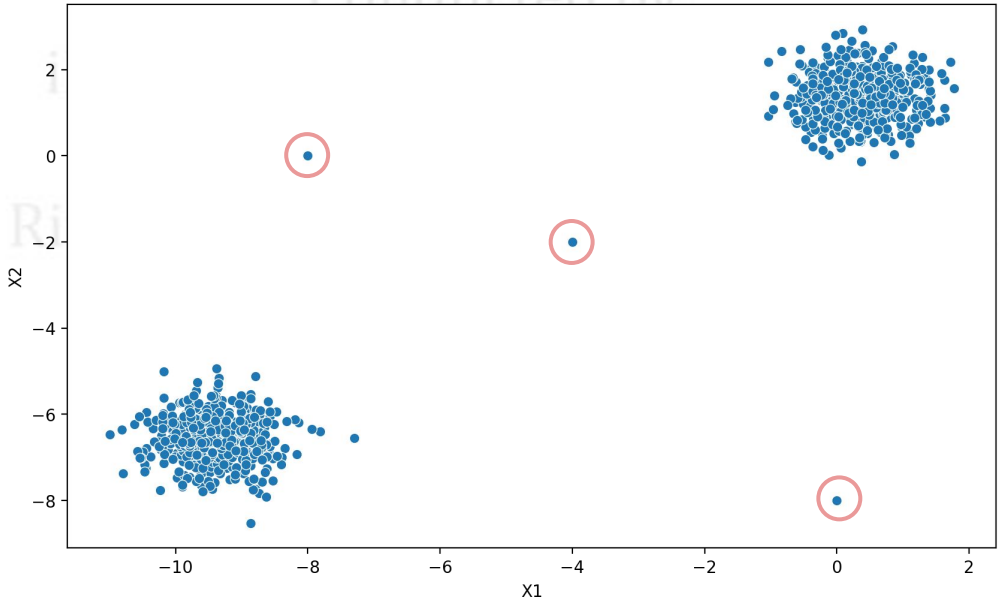


fig ref: pierian data

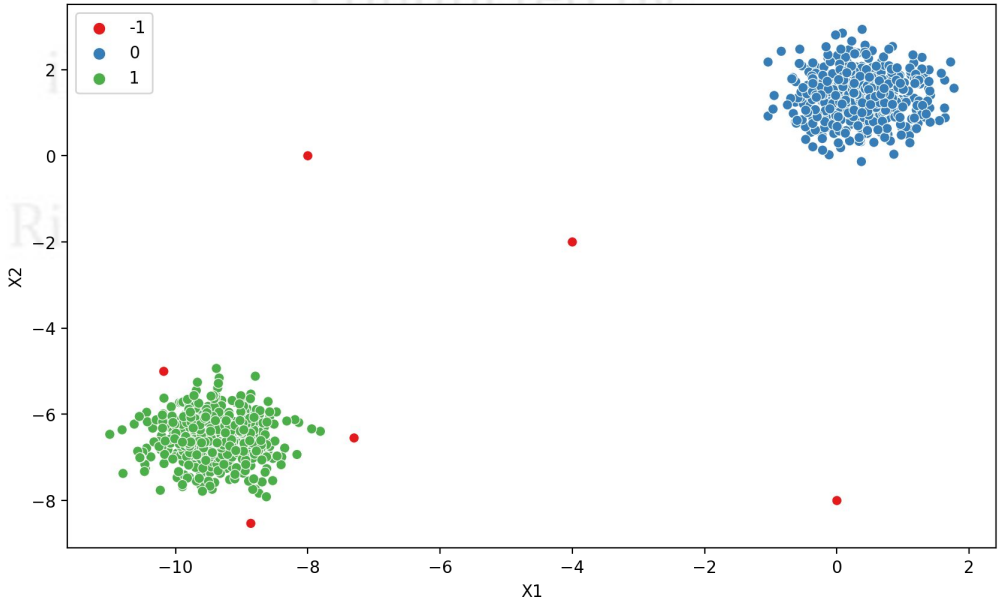


fig ref: pierian data

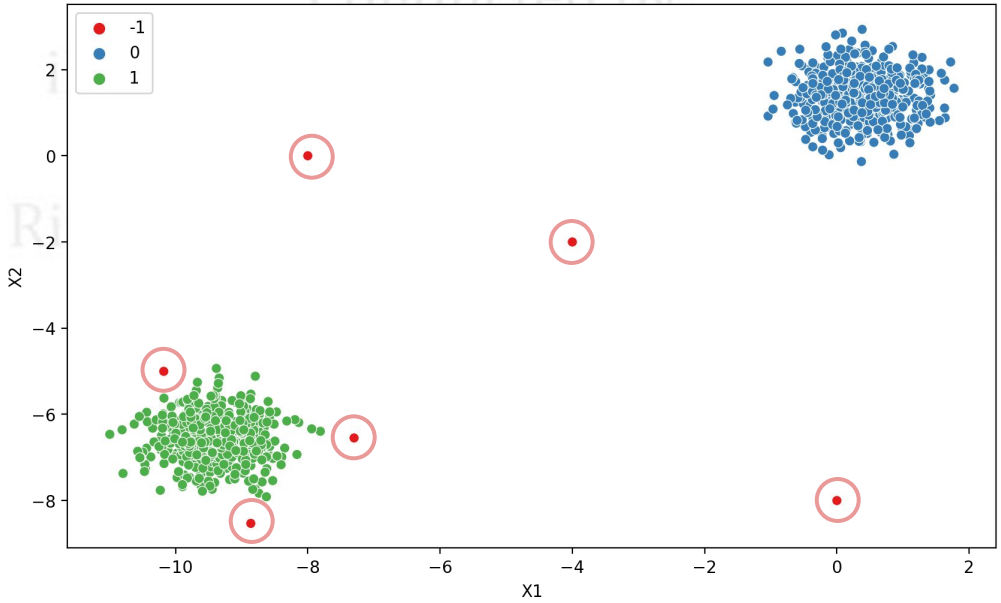


fig ref: pierian data

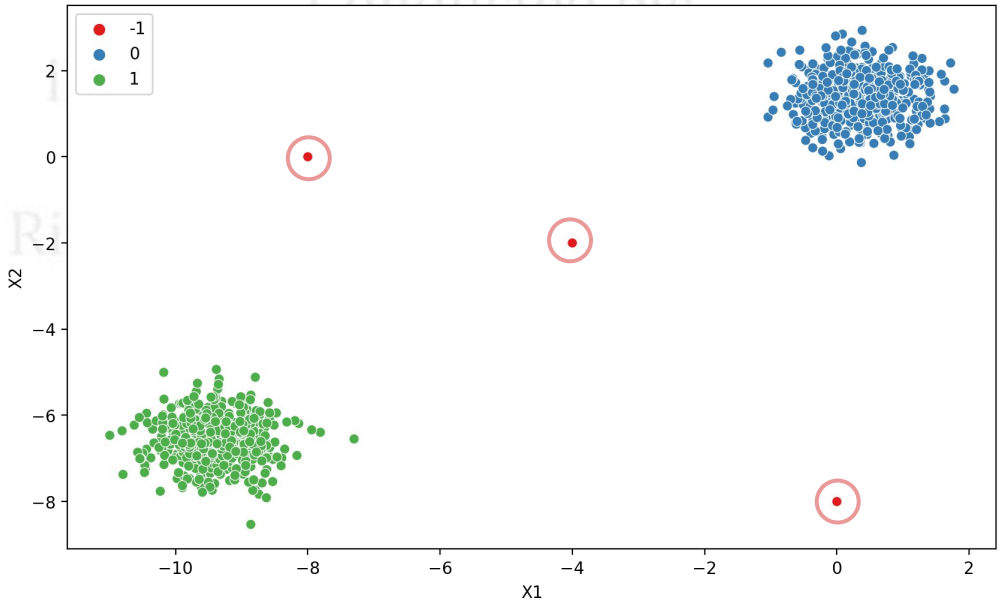


fig ref: pierian data

Epsilon Intuition:

- Increasing epsilon allows more points to be **core** points which also results in more **border** points and less outlier points.
- Imagine a huge epsilon, all points would be within the neighborhood and classified as the same cluster!
- Decreasing epsilon causes more points not to be in range of each other, creating more unique clusters.
- Imagine a tiny epsilon, the range would not extend far out enough to come into contact with any other points!

fig ref: pierian data

Methods for finding an epsilon value:

Run multiple DBSCAN models varying epsilon and measure:

- Number of Clusters
- Number of Outliers
- Percentage of Outliers
-

Extremely dependent on the particular data set and domain space.

Requires user to have some expectation or intuition about number of clusters and relative percentage of outliers.

fig ref: pierian data

Plot “elbow/knee” diagram comparing epsilon values:

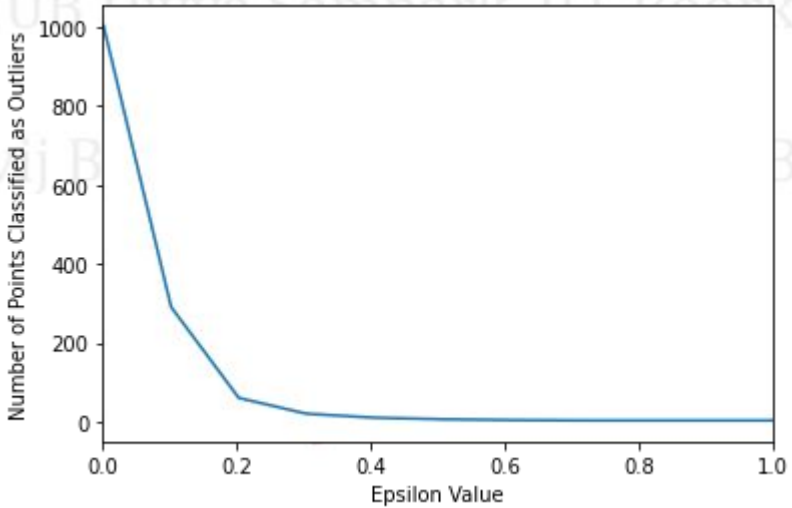


fig ref: pierian data

Minimum Number of Samples/Points:

- Number of samples in a neighborhood for a point to be considered as a **core** point (including the point itself).

fig ref: pierian data

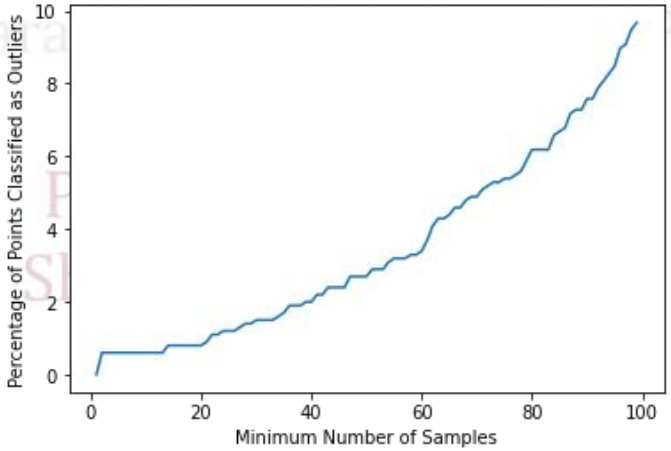
Min. Number of Samples Intuition:

- Increasing to a larger number of samples needed to be considered a core point, causes more points to be considered unique outliers.
- Imagine if min. number of samples was close to total number of points available, then very likely all points would become outliers.

fig ref: pierian data

Choosing Min. Number of Samples:

- Test multiple potential values and chart against number of outliers labeled.



Min. Number of Samples Note:

- Useful to increase to create potential new small clusters, instead of complete outliers.

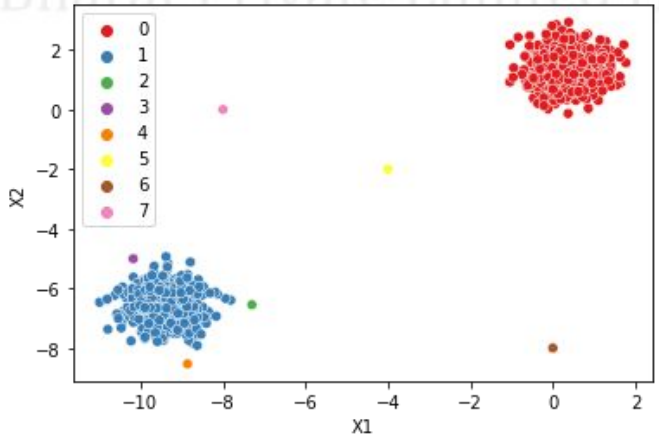


fig ref: pierian data

Let’s continue by exploring hyperparameters with code and data examples!

fig ref: pierian data