HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that improves on DBSCAN by adding hierarchical clustering and density-based features to identify clusters of varying shapes and sizes, as well as being able to handle noise points more effectively.

Steps of HDBSCAN

Step 1: Calculate Distances Between Points
Step 2: Calculate Core Distances
Step 3: Compute Mutual Reachability Distances
Step 4: Construct the Minimum Spanning Tree (MST)
Step 5: Condense the Tree & Identify Clusters
Step 6: Extract Final Clusters

Mathematical Example:

Let's use a small toy dataset to illustrate how HDBSCAN works.

Dataset:

Example Dataset:

Consider the following 2D data points:

- $A = (1, 2)$
- $B = (2, 2)$
- $C = (2, 3)$
- $D = (7, 8)$
- $E = (8, 7)$
- $F = (25, 80)$

Let's set the parameters for HDBSCAN:

- $k = 2$ (meaning the core distance is based on the second nearest neighbor).
- We'll focus on calculating the **mutual reachability distance** and the **minimum spanning tree** to create a hierarchical structure.

Step 1: Calculate Euclidean Distances Between Points

To begin, we calculate the **Euclidean distance** between each pair of points:

1. $d(A, B) = \sqrt{(2-1)^2 + (2-2)^2} = \sqrt{1} = 1$
2. $d(A, C) = \sqrt{(2-1)^2 + (3-2)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$
3. $d(A, D) = \sqrt{(7-1)^2 + (8-2)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.49$
4. $d(A, E) = \sqrt{(8-1)^2 + (7-2)^2} = \sqrt{49+25} = \sqrt{74} \approx 8.60$
5. $d(A, F) = \sqrt{(25-1)^2 + (80-2)^2} = \sqrt{576+6241} = \sqrt{6817} \approx 82.55$
6. $d(B, C) = \sqrt{(2-2)^2 + (3-2)^2} = \sqrt{1} = 1$
7. $d(B, D) = \sqrt{(7-2)^2 + (8-2)^2} = \sqrt{25+36} = \sqrt{61} \approx 7.81$
8. $d(B, E) = \sqrt{(8-2)^2 + (7-2)^2} = \sqrt{36+25} = \sqrt{61} \approx 7.81$
9. $d(B, F) = \sqrt{(25-2)^2 + (80-2)^2} = \sqrt{529+6241} = \sqrt{6770} \approx 82.34$
10. $d(C, D) = \sqrt{(7-2)^2 + (8-3)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$
11. $d(C, E) = \sqrt{(8-2)^2 + (7-3)^2} = \sqrt{36+16} = \sqrt{52} \approx 7.21$
12. $d(C, F) = \sqrt{(25-2)^2 + (80-3)^2} = \sqrt{529+5929} = \sqrt{6458} \approx 80.33$
13. $d(D, E) = \sqrt{(8-7)^2 + (7-8)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$
14. $d(D, F) = \sqrt{(25-7)^2 + (80-8)^2} = \sqrt{324+5184} = \sqrt{5508} \approx 74.19$
15. $d(E, F) = \sqrt{(25-8)^2 + (80-7)^2} = \sqrt{2\text{[illegible]}+5329} = \sqrt{5618} \approx 74.99$

Step 2: Determine Core Distances

The **core distance** for each point is determined by the distance to its min_cluster_size-th nearest neighbor. Since we are using **min_cluster_size = 2**, for each point, **we need to find the second nearest neighbor distance.**

Core Distances:

- **A (1, 2)**: The two nearest neighbors are **B (2, 2)** with distance 1 and **C (2, 3)** with distance 1.41. The core distance is 1.41.
- **B (2, 2)**: The two nearest neighbors are **A (1, 2)** with distance 1 and **C (2, 3)** with distance 1. The core distance is 1.
- **C (2, 3)**: The two nearest neighbors are **B (2, 2)** with distance 1 and **A (1, 2)** with distance 1.41. The core distance is 1.41.
- **D (7, 8)**: The two nearest neighbors are **E (8, 7)** with distance 1.41 and **C (2, 3)** with distance 7.07. The core distance is 7.07.
- **E (8, 7)**: The two nearest neighbors are **D (7, 8)** with distance 1.41 and **C (2, 3)** with distance 7.21. The core distance is 7.21
- **F (25, 80)**: The two nearest neighbors are **E (8, 7)** with distance 74.95 and **D (7, 8)** with distance 74.22. The core distance is 74.95

Step 3: Calculate Mutual Reachability Distances

The **mutual reachability distance** between two points p and q is defined as:

$$d_{reach}(p,q)=\max(core\_distance(p),core\_distance(q),d(p,q))$$

Examples of Mutual Reachability Distances:

- Between **A (1, 2)** and **B (2, 2)**:
  - Core distances: core_distance(A)=1.41, core_distance(B)=1
  - Euclidean distance: d(A,B)=1
  - Mutual reachability distance: $d_{reach}(A,B)=\max(1.41,1,1)=$**1.41**
- Between **A (1, 2)** and **C (2, 3)**:
  - Core distances: core_distance(A)=1.41, core_distance(C)=1.41
  - Euclidean distance: d(A,C)=1.41
  - Mutual reachability distance: $d_{reach}(A,C)=\max(1.41,1.41,1.41)=$**1.41**
- Between **D (7, 8)** and **E (8, 7)**:
  - Core distances: core_distance(D)=7.07, core_distance(E)= 7.21
  - Euclidean distance: d(D,E)=1.41
  - Mutual reachability distance: $d_{reach}(D,E)=\max(7.07,1.41,7.21)=$**7.21**
- Between **D (7, 8)** and **F (25, 80)**:
- Core distances: core_distance(D)=7.07, core_distance(F)=74.95
- Euclidean distance: d(D,F)=74.22
- Mutual reachability distance: $d_{reach}(D,F)=\max(7.07,74.95,74.22)=$**74.95**

Step 4: Build Minimum Spanning Tree (MST)

We construct the **Minimum Spanning Tree (MST)** from the mutual reachability distances. In the MST, points with smaller mutual reachability distances will be connected first, forming a tree structure. Based on the distances we calculated, the MST will connect the following:

- **A (1, 2) → B (2, 2) → C (2, 3)**
- **D (7, 8) → E (8, 7)**
- **F (25, 80)** is isolated and will be a separate point, treated as noise.

Step 5: Extract Clusters

The **Condensed Tree** is then used to extract clusters. The tree shows which points are densely connected. We can cut the tree at a certain threshold to extract clusters. Since **F (25, 80)** is very far from the other points, it will likely be considered **noise**.

From the MST, we can extract the following clusters:

- **Cluster 1**: Points **A (1, 2), B (2, 2), C (2, 3)**

- **Cluster 2**: Points **D (7, 8), E (8, 7)**
- **Noise**: Point **F (25, 80)** is isolated and treated as noise.

Final Clusters:

- **Cluster 1**: {A(1,2),B(2,2),C(2,3)}
- **Cluster 2**: {D(7,8),E(8,7)}
- **Noise**: {F(25,80)}

This is how HDBSCAN would cluster the data points and identify **noise** based on density and the concept of mutual reachability distances. The result is two clusters and one outlier (noise).