Кластеризация — это набор методов без учителя для группировки данных по определённым критериям в так называемые кластеры, что позволяет выявлять сходства и различия между объектами, а также упрощать их анализ и визуализацию. Из-за частичного сходства в постановке задач с классификацией кластеризацию ещё называют unsupervised classification.

**Области применения кластеризации и её разнови**дности

Кластеризация широко применяется в машинном обучении для решения различного спектра задач:

* классификация (определение к какому классу относится каждый объект или же выделение новых классов, которые не были известны заранее);
* сегментация рынка (разделение потенциальных клиентов на группы по их характеристикам для разработки более эффективных стратегий в маркетинге и продажах);
* сегментация изображений (разделение изображения на сегменты или группы пикселей);
* кластеризация геоданных (группировка данных по их географическому расположению, например, разделение районов на безопасные и опасные, богатые и бедные, и так далее);
* понижение размерности (уменьшение количества признаков путем объединения схожих в один кластер).

Существует множество различных типов кластеризации, которые можно разделить по следующим критериям:

### Области применения кластеризации и её разновидности

Кластеризация широко применяется в машинном обучении для решения различного спектра задач:

* классификация (определение к какому классу относится каждый объект или же выделение новых классов, которые не были известны заранее);
* сегментация рынка (разделение потенциальных клиентов на группы по их характеристикам для разработки более эффективных стратегий в маркетинге и продажах);
* сегментация изображений (разделение изображения на сегменты или группы пикселей);
* кластеризация геоданных (группировка данных по их географическому расположению, например, разделение районов на безопасные и опасные, богатые и бедные, и так далее);
* понижение размерности (уменьшение количества признаков путем объединения схожих в один кластер).

Существует множество различных типов кластеризации, которые можно разделить по следующим критериям:

* **По способу формирования кластеров:**
	+ Разделительные (partitioning) — разбивают данные на заданное число кластеров, минимизируя расстояние внутри кластера и максимизируя расстояние между кластерами (например, K-means).
	+ Основанные на плотности (density-based) — группируют точки, которые находятся в областях с высокой плотностью и отделяют их от областей с низкой плотностью (например, DBSCAN).
	+ Основанные на сетке (grid-based) — разбивают пространство на ячейки сетки и анализируют плотность данных в каждой ячейке (например, STING).
	+ Основанные на модели (model-based) — предполагают, что данные порождены некоторой статистической моделью и пытаются подобрать параметры этой модели (например, смеси Гауссианов).
	+ Основанные на графах (graph-based) — используют графовое представление данных и разбивают его на подграфы, соответствующие кластерам (например, спектральная кластеризация).
	+ Основанные на подпространствах (subspace-based) — ищут кластеры в подпространствах признаков, а не во всём пространстве (например, CLIQUE).
	+ Основанные на ансамбле (ensemble-based) — комбинируют результаты различных алгоритмов кластеризации, чтобы получить более стабильное и надёжное разбиение (например, CSPA).
* **По степени вложенности кластеров:**
	+ Плоские (flat) — разбивают данные на один уровень кластеров, не учитывая их иерархию (например, K-means).
	+ Иерархические (hierarchical) — разбивают данные на несколько уровней кластеров, учитывая их иерархию. Существуют два основных подхода к иерархической кластеризации: агломеративный (начинается с того, что каждый объект является отдельным кластером, а затем постепенно наиболее близкие кластеры объединяются в более крупные) и дивизивный (начинается с того, что все объекты составляют один кластер, а затем постепенно разделяются на более мелкие кластеры).
* **По степени пересечения кластеров:**
	+ Исключающие (exclusive) — каждый объект принадлежит только одному кластеру (например, K-means).
	+ Перекрывающие (overlapping) — каждый объект может принадлежать нескольким кластерам (например, MCOKE).
	+ Нечёткие (fuzzy) — каждый объект принадлежит каждому кластеру с некоторой степенью принадлежности (например, fuzzy K-means).