: تعریف کلی Clustering خوشه‌بندی:

Clusteringیکی از روش‌های **یادگیری بدون نظارت (Unsupervised Learning)** در علم داده و یادگیری ماشین است که هدف آن **تقسیم داده‌ها به دسته ها (خوشه‌ها)** است به طوری که:

- داده‌های داخل هر خوشه **بیشتر شبیه هم باشند**
- داده‌های بین خوشه‌های مختلف **کمتر به هم شبیه باشند**

به زبان ساده‌تر:

Clustering یعنی اینکه داده‌هایی که شبیه هم هستند رو کنار هم بگذاریم، بدون اینکه از قبل بدانیم کدوم داده متعلق به کدوم دسته است.

K-Means Clustering:

K-Means یک الگوریتم **خوشه‌بندی (Clustering)** در یادگیری ماشین است که:

داده‌ها را به **kخوشه‌ی مجزا** تقسیم می‌کند، به‌طوری که **نقاط در هر خوشه به مرکز آن خوشه نزدیک‌تر** از مراکز دیگر باشند. این الگوریتم یکی از ساده ترین روش های دسته بندی محسوب می شود. یعنی نقطه ای را پیدا میکنیم که در مرکز خوشه قرار گرفته باشد و به تمام داده ها در آن خوشه نزدیکترین حالت را داشته باشد.

در کتابخانه sklearn یک تابع برای انجام clustering وجود دارد.

```
from sklearn.datasets import make_blobs
```

`make_blobs` یک تابع برای **تولید داده‌های خوشه‌ای‌شکل (blobs)** در فضای-n بعدی است.

این تابع برای ساخت **داده‌های تستی برای الگوریتم‌های Clustering** طراحی شده است.

```
X, _ = make_blobs(random_state=42)
X.shape
```

این دستور یک دیتاست مصنوعی (ساختگی) می‌سازد که به‌صورت خوشه‌ای (blobs) است. و به‌طور پیش‌فرض، ۱۰۰ نقطه داده‌ای دوبعدی ایجاد می‌کند که به ۳ خوشه تقسیم شده‌اند.

پارامتر `random_state=42` باعث می‌شود که تولید داده‌ها تصادفی ولی تکرارپذیر باشد، یعنی هر بار که کد را اجرا کنید، دقیقاً همان داده‌ها ساخته شوند.

```
X, _ = make_blobs(...)
```

تابع `make_blobs` دو خروجی دارد:

- `X`: داده‌ها (مختصات نقاط در فضای دوبعدی)
- `y`: برچسب هر نقطه (که خوشه‌اش را مشخص می‌کند)

در این خط فقط `X` را نگه می‌داریم و `y` را با `_` دور می‌اندازیم، چون به آن احتیاجی نداریم.

نکته: مقدار 42 یک عدد رایج بین برنامه نویسان است و میتوان آن را تغییر داد. این عدد برگرفته از رمان "راهنمای مسافران کهکشان به کهکشان" است که در آن 42 پاسخ نهایی به زندگی، جهان و همه چیز است. بنابراین می توان این عدد راتغییر داد اما اگر این مقدار خالی باشد None برگردانده می شود و هر بار کد اجرا شود مقدار متفاوتی بازگردانده می شود.

```
plt.scatter(X[:, 0], X[:, 1]);
```

در ادامه یک نمودار دو بعدی رسم میکنیم.

```
X[:, 0]
```

یعنی: تمام ردیف‌ها، ستون اول از آرایه‌ی `X` این می‌شود **محور افقی(X-axis)**

```
X[:, 1]
```

یعنی: تمام ردیف‌ها، ستون دوم از آرایه‌ی `X` این می‌شود **محور عمودی(Y-axis)**

به عبارتی:

```
X[:, 0]
```

فرض کن `X` یک جدول (یا آرایه) از داده‌هاست.این دستور یعنی از همه‌ی ردیف‌ها، فقط ستون اول رو بردار ستون اول معمولاً **مقدار محور x** هست.

```
X[:, 1]
```

از همه‌ی ردیف‌ها، فقط ستون دوم رو بردار. ستون دوم معمولاً **مقدار محور y** است.

```
kmeans = KMeans(n_clusters=3, random_state=42)
```

یک شی جدید از الگوریتم kmeans بساز که قرار است داده‌ها را به سه خوشه تقسیم کند و برای اینکه نتایج تصادفی قابل تکرار باشند مقدار random state را برابر با عدد چهل و دو قرار بدهیم.

عبارت n clusters برابر با سه یعنی ما می‌خواهیم سه گروه یا خوشه پیدا کنیم.

عبارت random state برابر با چهل و دو یعنی در اجرای دوباره کد نتایج یکسان بمانند چون الگوریتم kmeans به طور پیش‌فرض رفتار تصادفی دارد و با این عدد آن را کنترل می‌کنیم.

```
kmeans = KMeans(n_clusters=3, random_state=42)
```

این خط کد الگوریتم kmeans را روی داده‌های X اجرا می‌کند و به صورت هم‌زمان دو کار انجام می‌دهد. اول الگوریتم را روی داده‌ها آموزش می‌دهد یعنی خوشه‌ها را پیدا می‌کند و مراکز آن‌ها را مشخص می‌کند. دوم برای هر داده در X مشخص می‌کند که به کدام خوشه تعلق دارد و آن را برچسب‌گذاری می‌کند. خروجی این عملیات لیستی از اعداد است که نشان می‌دهد هر داده به کدام خوشه اختصاص یافته است. این لیست در متغیری به نام labels ذخیره می‌شود تا بعداً بتوانیم از آن برای تحلیل یا رسم نمودار استفاده کنیم.

```
plt.scatter(X[:, 0], X[:, 1], c=labels);
```

این خط کد یک نمودار نقطه‌ای از داده‌های X رسم می‌کند و برای هر نقطه رنگ متفاوتی بر اساس خوشه‌ای که به آن تعلق دارد در نظر می‌گیرد. محور افقی شامل مقادیر ستون اول داده‌ها و محور عمودی شامل مقادیر ستون دوم است. با استفاده از آرگومان cبرابر باlabels ، هر داده با رنگ خوشه‌ای که به آن تعلق دارد نمایش داده می‌شود. این باعث می‌شود که خوشه‌بندی الگوریتم KMeansبه صورت بصری قابل مشاهده باشد و بتوان به راحتی تشخیص داد که الگوریتم چگونه داده‌ها را گروه‌بندی کرده است.