مصورسازی روش‌های نرمال‌سازی داده

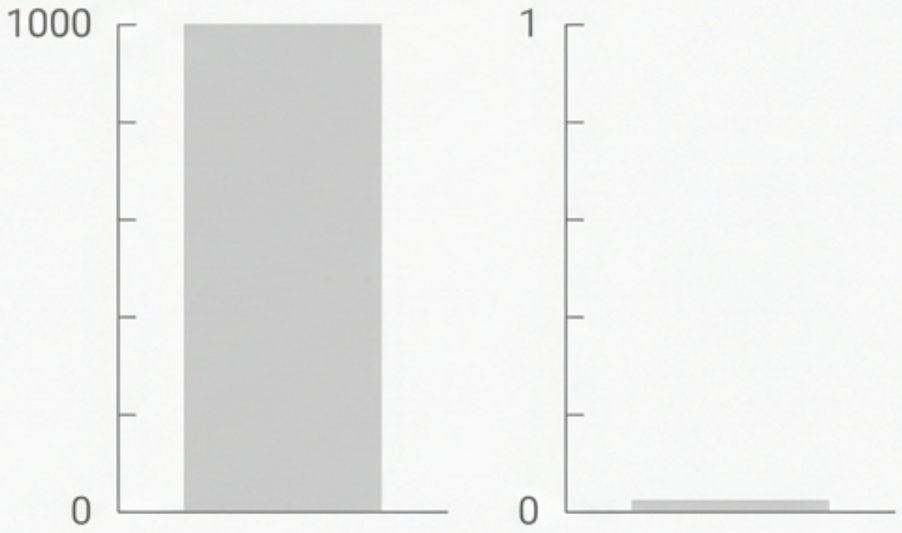
یک مقایسه بصری از تکنیک‌های رایج مقیاس‌بندی برای آماده‌سازی داده‌ها

NotebookLM

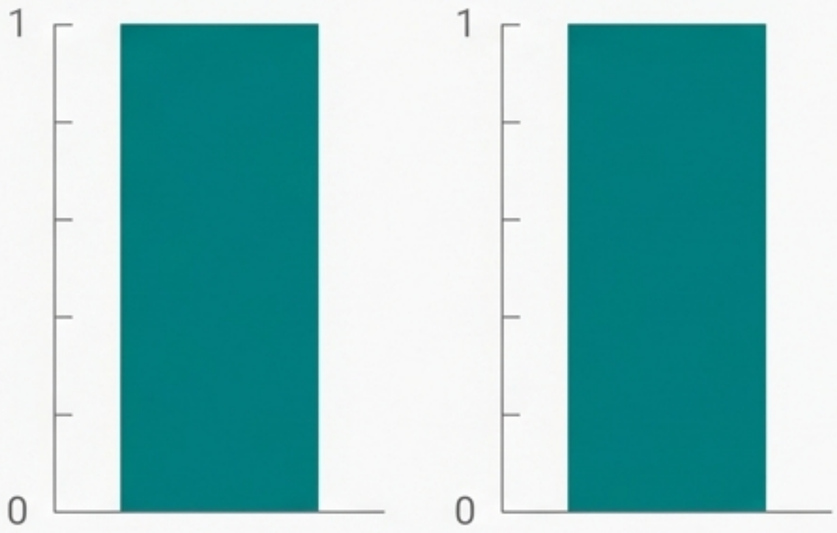
چرا مقیاس‌بندی داده‌ها اهمیت دارد؟

الگوریتم‌های یادگیری ماشین به ویژگی‌هایی که در مقیاس‌های بسیار متفاوتی قرار دارند، حساس هستند. یک ویژگی با دامنه بزرگ (مثلاً ۰ تا ۱۰۰۰) می‌تواند بر یک ویژگی با دامنه کوچک (مثلاً ۰ تا ۱) غلبه کند. مقیاس‌بندی، تمام ویژگی‌ها را در یک محدوده استاندارد قرار می‌دهد تا از این سوگیری جلوگیری شود و عملکرد مدل بهبود یابد.

مشکل: مقیاس‌های ناهمگون



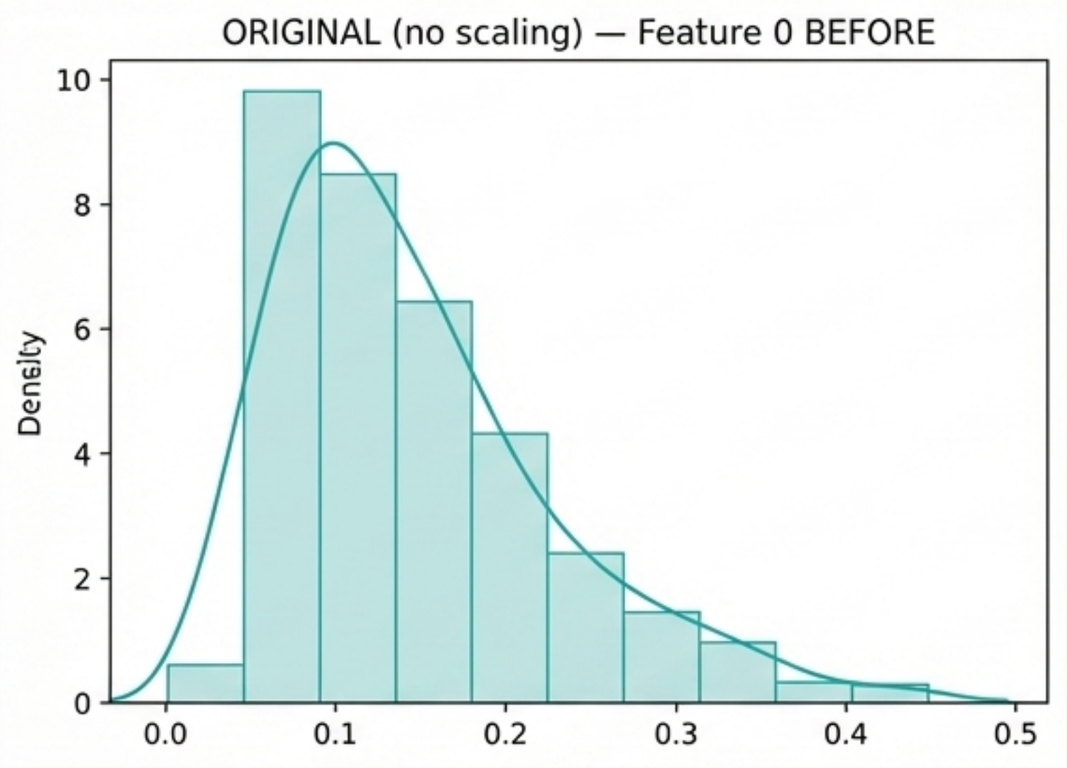
راه حل: مقیاس‌بندی



NotebookLM

نقطه شروع: نگاهی به داده‌های اولیه ما

ما از ویژگی‌های عددی استخراج‌شده از یک مجموعه داده متنی استفاده می‌کنیم. این نمودارها و آمارها، توزیع داده‌های خام را برای یکی از ویژگی‌ها (Feature 0) قبل از اعمال هرگونه تغییری نشان می‌دهند.



این نمودار، فراوانی مقادیر مختلف داده‌های خام را نشان می‌دهد.

آمار کلیدی

- میانگین (Mean): 0.061
- انحراف معیار (Std Dev): 0.044
- کمینه (Min): 0.0
- بیشینه (Max): 0.5
- چولگی (Skewness): 1.397

NotebookLM

روش اول: مقیاس‌بندی کمینه-بیشینه (Min-Max Scaling)

این روش، داده‌ها را به صورت خطی به یک محدوده ثابت، معمولاً بین ۰ و ۱، منتقل می‌کند. شکل کلی توزیع داده‌ها حفظ می‌شود، اما مقیاس محورها تغییر می‌کند.

```
X_scaled = (X - X_min) / (X_max - X_min)
```

کاربرد اصلی

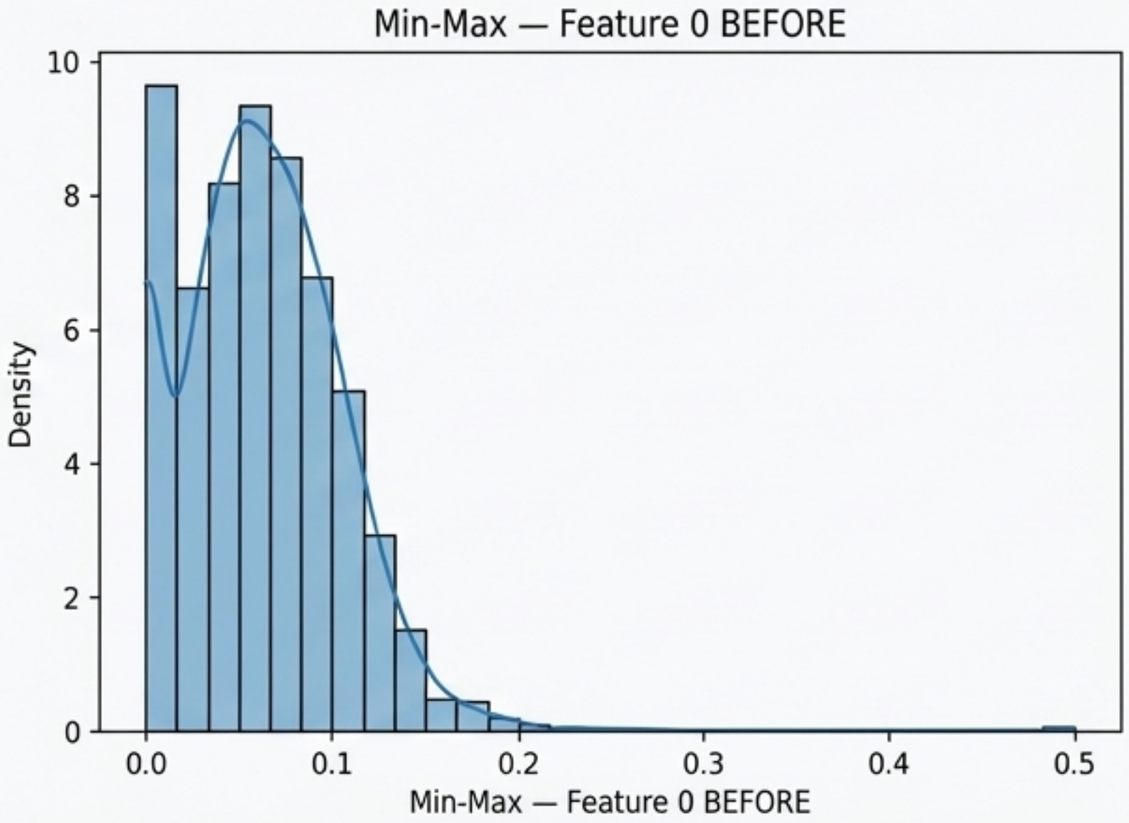
این روش برای الگوریتم‌هایی مانند شبکه‌های عصبی که به ورودی‌هایی در یک محدوده کوچک و نیاز دارند، بسیار رایج است. است.

NotebookLM

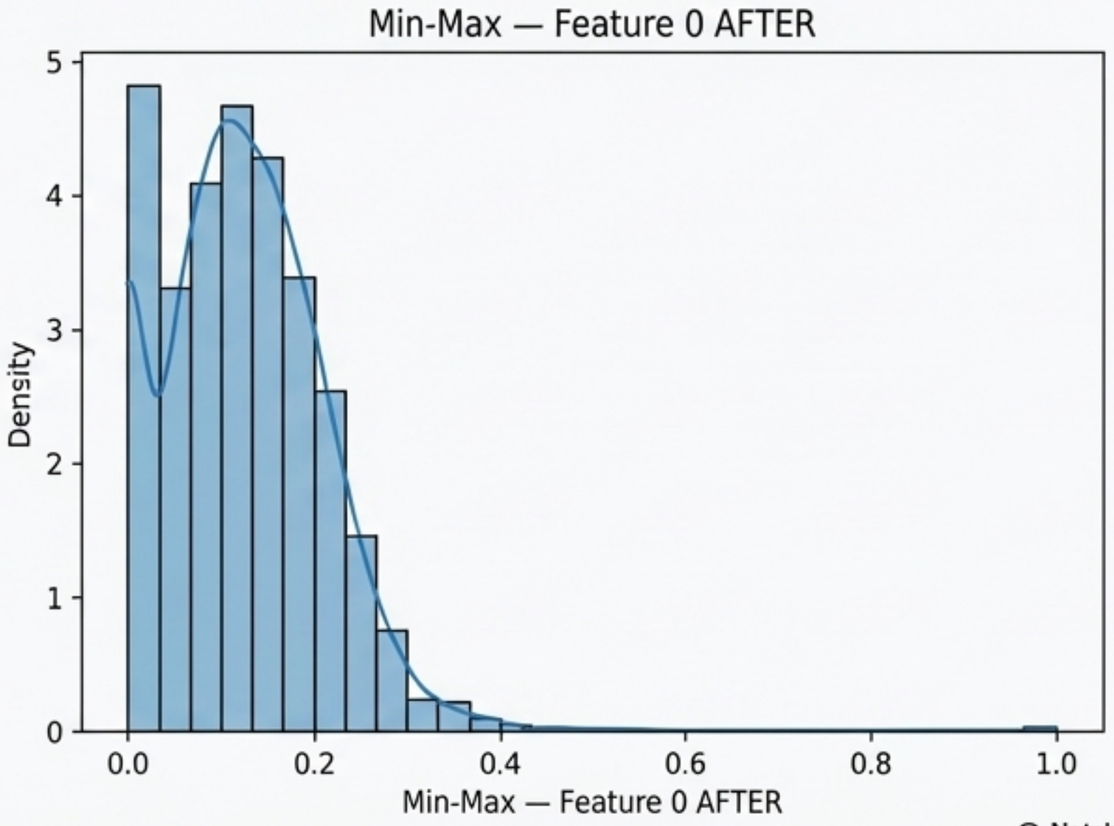
مقیاس‌بندی کمینه-بیشینه: قبل و بعد

این نمودار نشان می‌دهد که چگونه مقیاس محور افقی از [0, 0.5] به [0, 1] تغییر کرده، در حالی که شکل توزیع ثابت مانده است.

قبل از تبدیل



پس از تبدیل



NotebookLM

روش دوم: استانداردسازی Z-Score

این تکنیک، داده‌ها را طوری تبدیل می‌کند که میانگین آن‌ها برابر با ۰ و انحراف معیارشان برابر با ۱ شود. این کار به «مرکزیت‌بخشی» (centering) داده‌ها معروف است.

```
X_scaled = (X − μ) / σ
```

کاربرد اصلی

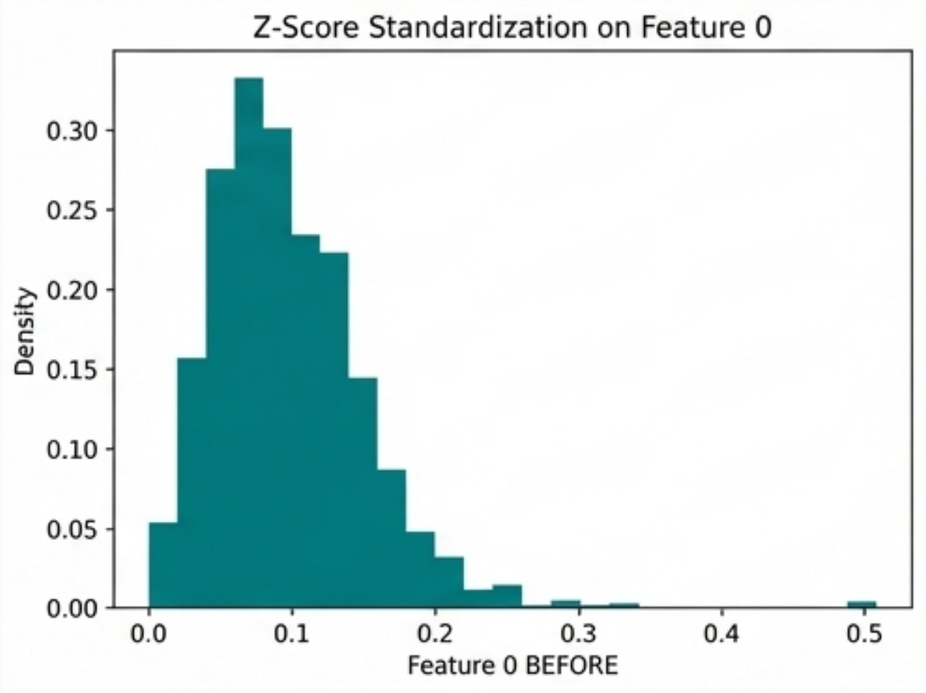
این روش الگوریتم‌هایی که فرض می‌کنند داده‌ها حول صفر توزیع شده‌اند (مانند PCA یا رگرسیون با regularizers) بسیار مفید است.

NotebookLM

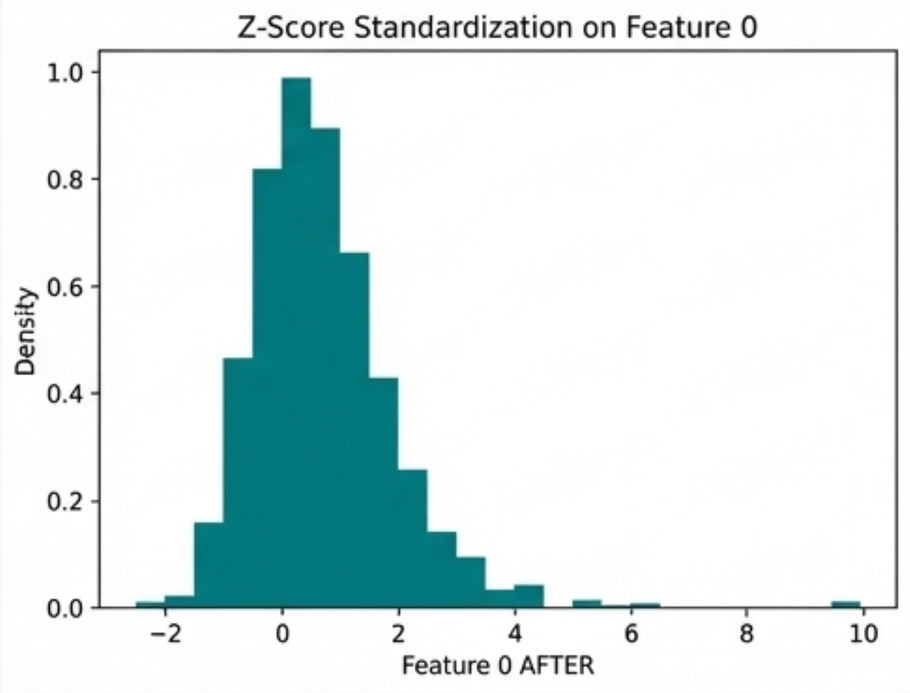
استانداردسازی Z-Score: قبل و بعد

این نمودار نشان می‌دهد که چگونه مرکز توزیع به صفر منتقل شده و مقادیر اکنون بر اساس انحراف معیار بیان می‌شوند.

قبل از تبدیل



پس از تبدیل



NotebookLM

روش سوم: مقیاس‌بندی اعشاری (Decimal Scaling)

این روش ساده، با جابجا کردن ممیز اعشاری مقادیر، آن‌ها را به محدوده [-1, 1] می‌آورد. مکان ممیز بر اساس بزرگترین قدر مطلق مقدار در داده‌ها تعیین می‌شود (در مثال ما، داده‌ها بر ۱۰ تقسیم می‌شوند).

```
X_scaled = X / 10^j
```

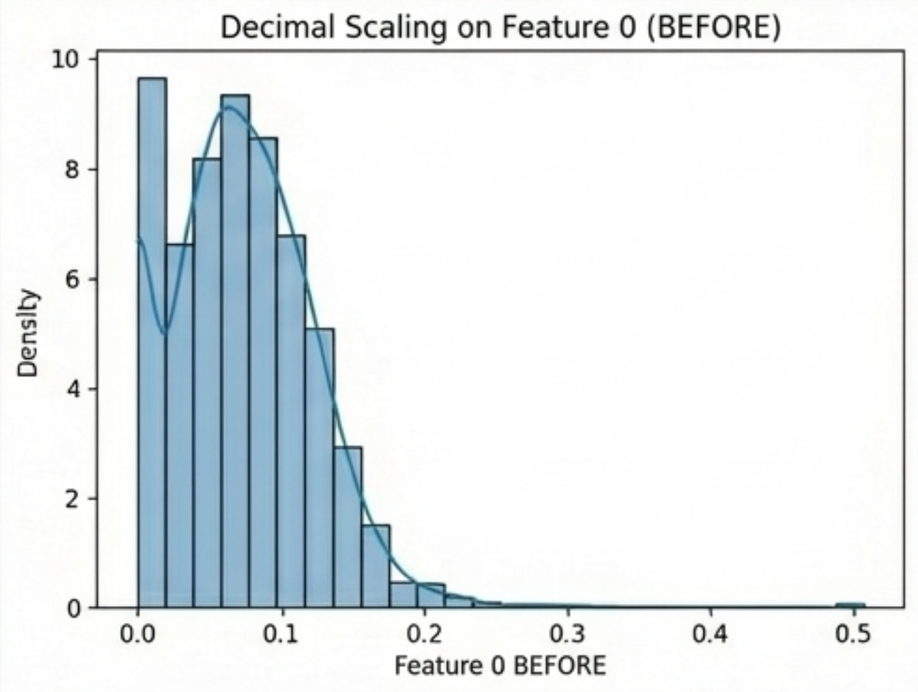
کاربرد اصلی: یک روش سریع و قابل فهم برای کاهش مقیاس داده‌ها بدون محاسبات پیچیده.

NotebookLM

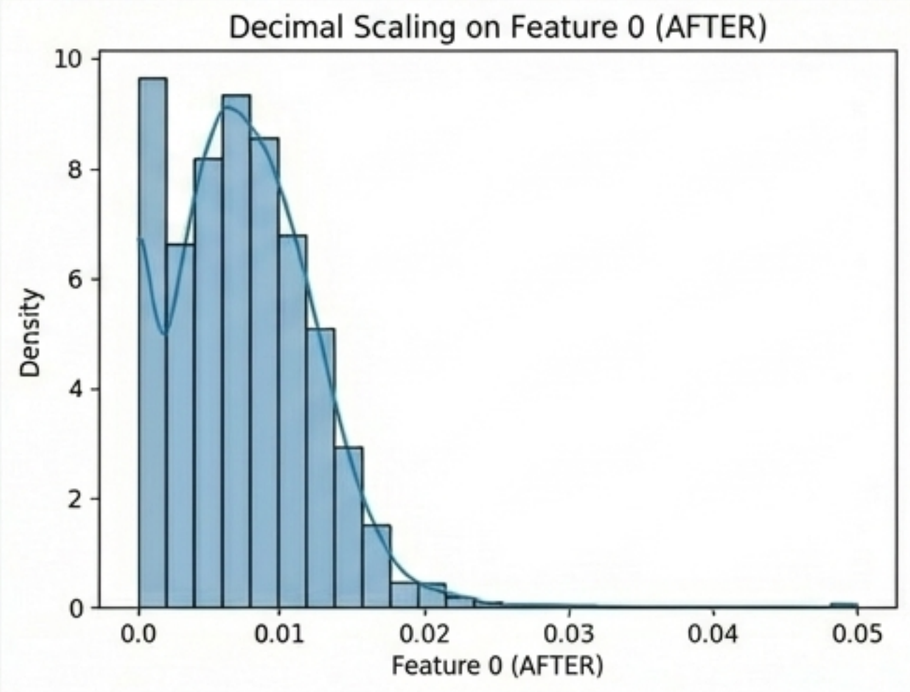
مقیاس‌بندی اعشاری: قبل و بعد

این نمودار نشان می‌دهد که مقادیر با تقسیم بر ۱۰ کوچکتر شده‌اند، اما شکل توزیع دست‌نخورده باقی مانده است.

قبل از تبدیل



پس از تبدیل



NotebookLM

روش پیشرفته: تبدیل کوانتایل (Quantile Transformation)

برخلاف روش‌های قبلی، این تکنیک **شکل توزیع داده‌ها را تغییر می‌دهد** و سعی می‌کند آن را به یک توزیع مشخص (مانند توزیع نرمال یا گوسی) نزدیک کند.

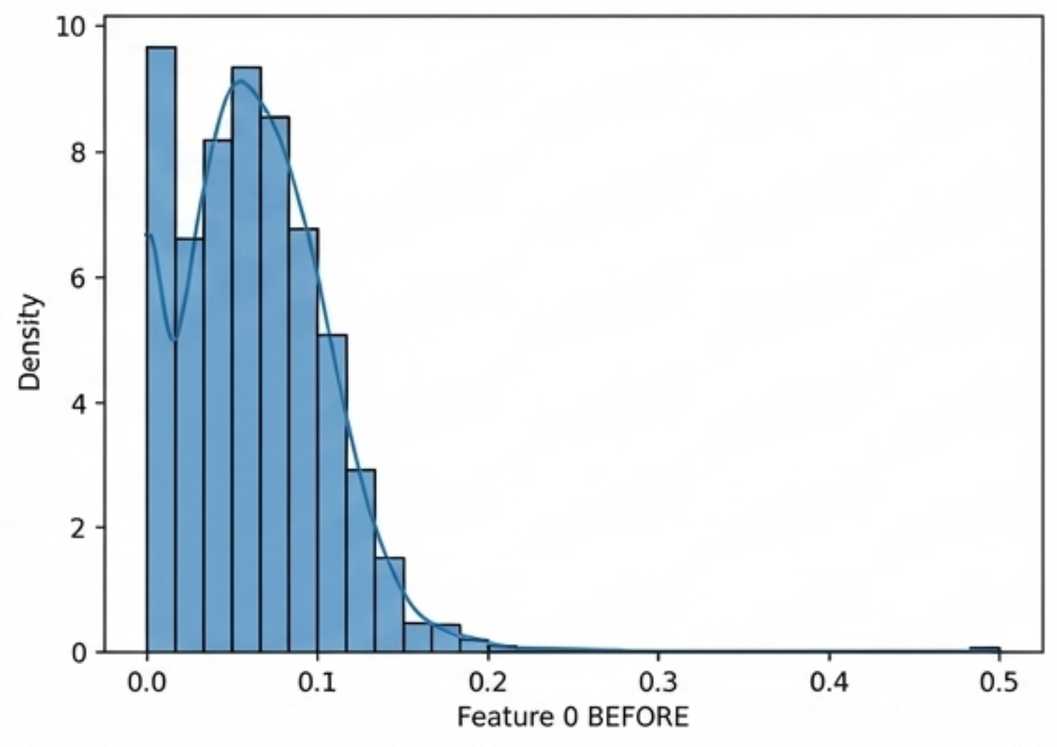
کاربرد اصلی: برای الگوریتم‌هایی که به توزیع نرمال حساس هستند یا برای مدیریت داده‌های پرت (outliers) بسیار مؤثر است.

NotebookLM

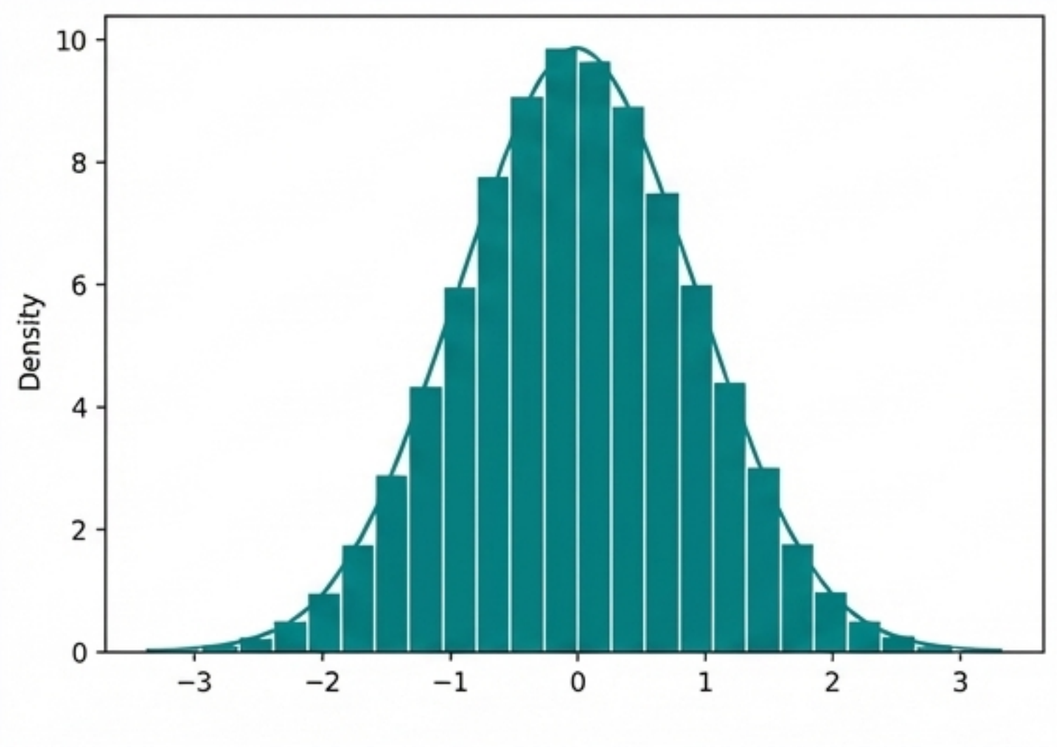
تبدیل کوانتایل: تغییر شکل توزیع

این نمودار نشان می‌دهد که چگونه توزیع نامتقارن اولیه به یک توزیع زنگوله‌ای-شکل (نرمال) تبدیل شده است.

توزیع نامتقارن اولیه



پس از تبدیل (توزیع نرمال)



NotebookLM

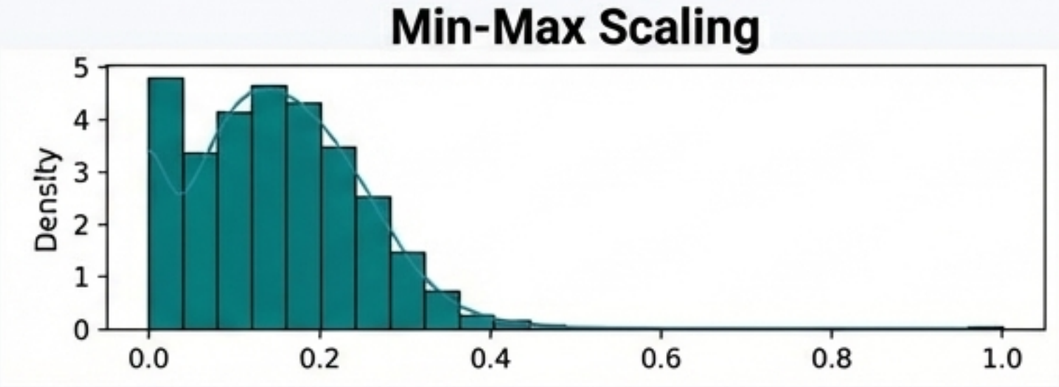
مقایسه جامع روش‌های مقیاس‌بندی

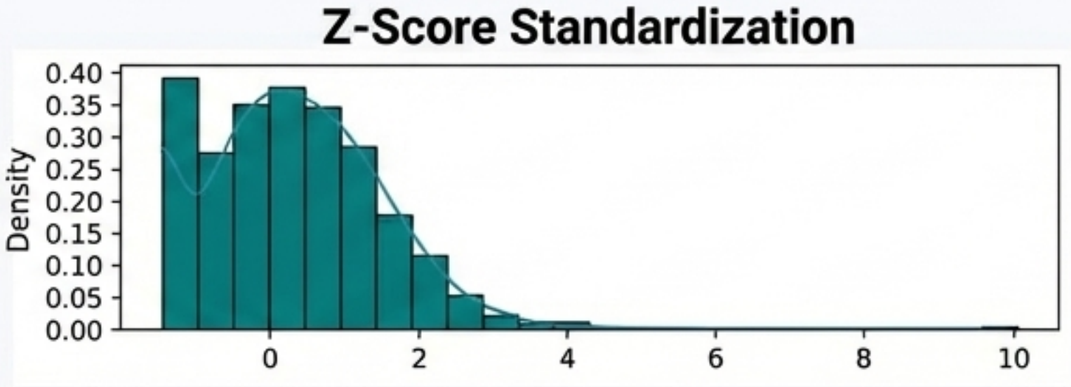
ویژگی (Feature)	مقیاس‌بندی Min-Max	استانداردسازی Z-Score	مقیاس‌بندی اعشاری	تبدیل کوانتایل
محدوده خروجی	معمولاً [0, 0]	متغیر، بدون محدوده ثابت	[-1, 1]	وابسته به توزیع هدف
اثر بر میانگین/انحراف	تغییر می‌کند	0 / 1	تغییر می‌کند	0 / 1 (برای خروجی نرمال)
تغییر می‌کند (به نرمال)	حفظ می‌شود	حفظ می‌شود	حفظ می‌شود	**اثر بر شکل توزیع**
پایین	بالا	متوسط	بالا	حساسیت به داده پرت

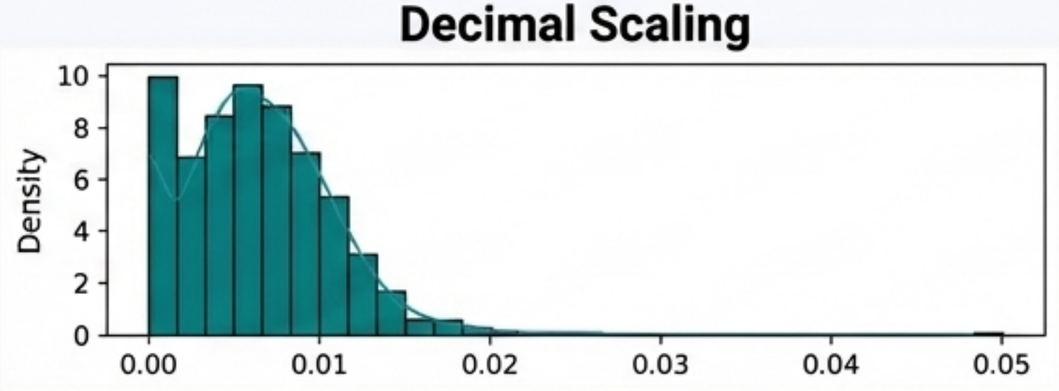
NotebookLM

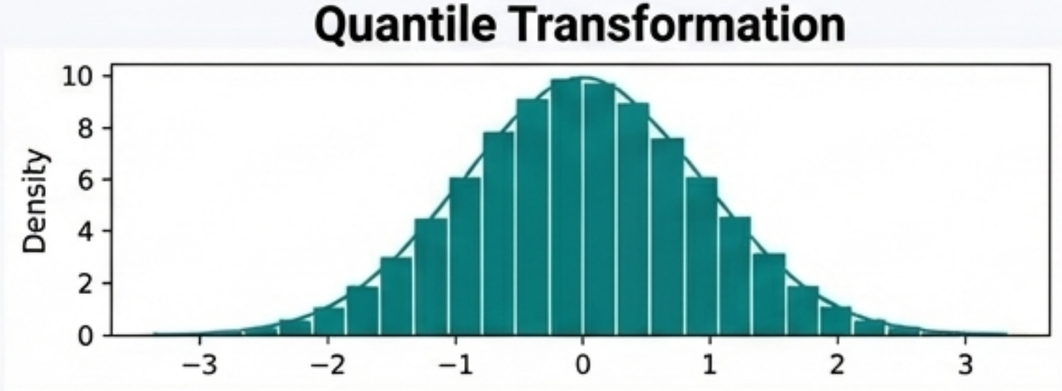
یک تصویر، هزار کلمه: تأثیر نهایی هر روش

این چهار نمودار، توزیع داده‌ها را **پس از** اعمال هر یک از روش‌های مقیاس‌بندی نشان می‌دهند. تفاوت‌ها در مقیاس و شکل به وضوح قابل مشاهده است.









NotebookLM

کدام روش را انتخاب کنیم؟

هیچ «بهترین» روشی وجود ندارد. انتخاب صحیح به دو عامل کلیدی بستگی دارد:

1. ***الگوریتم یادگیری ماشین شما*:** آیا مدل شما به داده‌های مرکزیت‌یافته (Z-Score) نیاز دارد یا به دامنه‌ای ثابت (Min-Max)؟
2. ***توزیع داده‌های شما*:** آیا داده‌های پرت زیادی دارید یا نیاز دارید توزیع داده‌ها نرمال باشد (Quantile Transform)؟

همیشه تأثیر روش انتخابی خود را بر روی داده‌ها به صورت بصری بررسی کنید تا از نتیجه کار مطمئن شوید.

NotebookLM