تکلیف اول درس شبکه های عصبی

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۲۶ آبان ۱۴۰۴

سوال ۲

مجموعه داده مسکن کالیفرنیا، که از سرشماری سال ۱۹۹۰ ایالات متحده استخراج شده است، یکی از مجموعه داده‌های معروف برای مسائل رگرسیون چندمتغیره است. این داده‌ها شامل ویژگی‌های اقتصادی، جمعیتی و جغرافیایی مناطق مسکونی در کالیفرنیا بوده و هدف، پیش‌بینی Median House Value است.

الف: رسم و تفسیر توزیع (هیستوگرام) هر یک از ویژگی‌ها

با استفاده از کتابخانه sklearn، داده‌ها بارگذاری شده و هیستوگرام تمام ویژگی‌ها و متغیر هدف با ۵۰ ستون (bin) رسم گردید. شکل ۱ توزیع هر یک از ویژگی‌ها را نشان می‌دهد.

تفسیر هیستوگرام‌ها

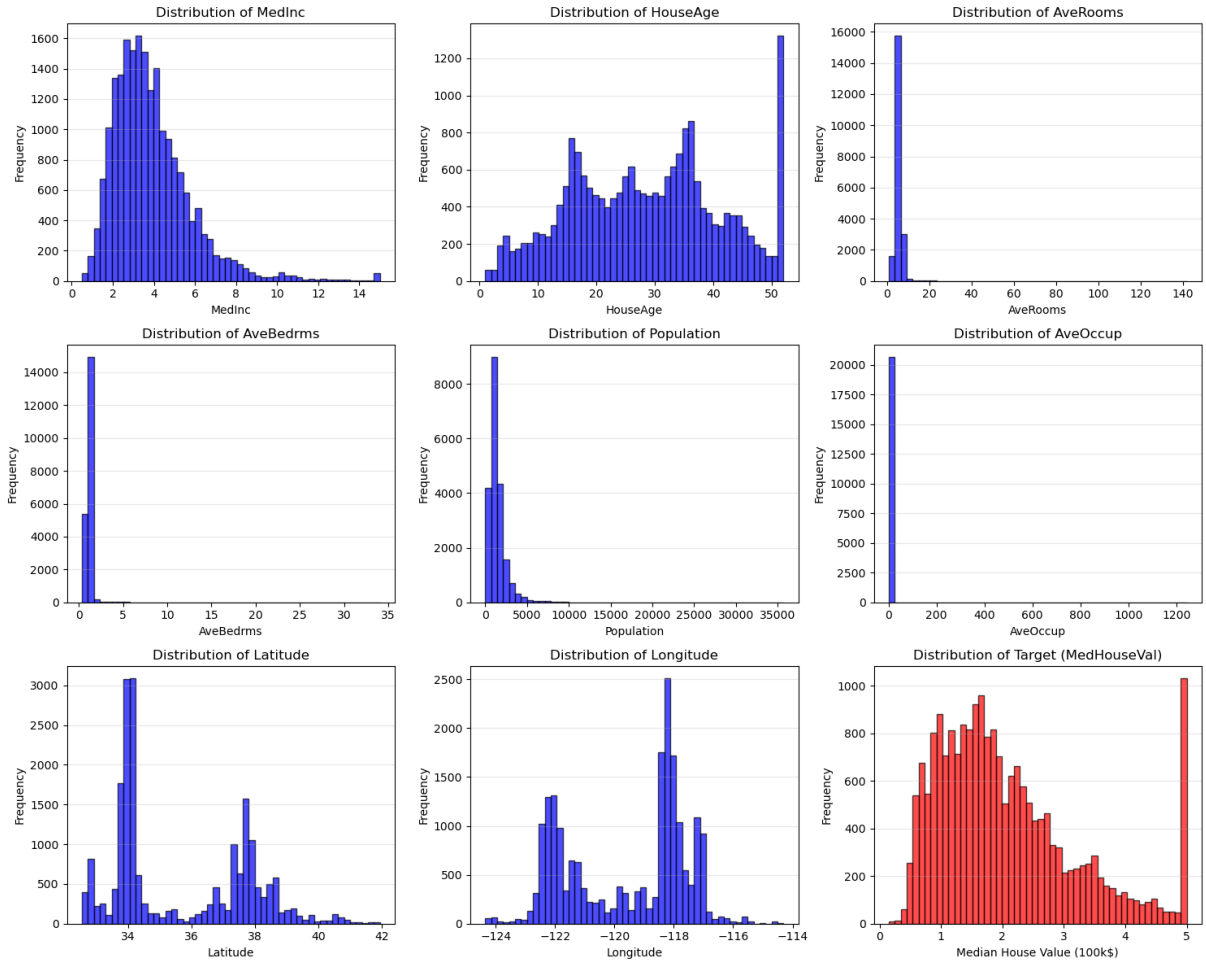
در ادامه، توزیع هر ویژگی به همراه تحلیل و اطلاعات استخراج‌شده از آن بررسی می‌شود:

۱. **MedInc (درآمد متوسط خانوار):** توزیع چپ‌کشیده (right-skewed) با پیک در حدود ۳ تا ۴ واحد (صد هزار دلار). بیشتر خانوارها درآمد متوسط پایینی دارند و تعداد کمی درآمد بسیار بالا (تا ۱۵) دارند. این نشان‌دهنده نابرابری درآمدی در مناطق است.

۲. **HouseAge (سن متوسط خانه):** توزیع نسبتاً یکنواخت با چند پیک در سنین ۱۵، ۳۰ و ۵۰ سال. این نشان می‌دهد خانه‌ها در سنین مختلف به طور نسبتاً یکسان توزیع شده‌اند، اما ساخت خانه در دهه‌های خاصی (مثلاً پس از جنگ جهانی دوم) بیشتر بوده است.

۳. **AveRooms (میانگین تعداد اتاق در هر خانه):** توزیع شدیداً چپ‌کشیده با پیک در حدود ۵ اتاق و دم بسیار طولانی تا بیش از ۱۲۰ اتاق. وجود مقادیر بسیار بالا (مثلاً خانه‌هایی با بیش از ۱۰۰ اتاق) نشان‌دهنده **داده‌های پرت (outlier)** یا احتمالاً خطا در داده‌هاست. این ویژگی نیاز به نرمال‌سازی یا حذف پرت دارد.

۴. **AveBedrms (میانگین تعداد اتاق خواب):** مشابه AveRooms، توزیع چپ‌کشیده با پیک در حدود ۱ تا ۱.۵ و دم طولانی تا بیش از ۳۰. مقادیر بسیار بالا غیرواقعی به نظر می‌رسند و احتمالاً ناشی از خطا یا مناطق خاص (مثلاً خوابگاه‌ها) هستند.



شکل ۱: توزیع (هیستوگرام) ویژگی‌های مجموعه داده مسکن کالیفرنیا و متغیر هدف

۵. **Population (جمعیت منطقه):** توزیع شدیداً چپ‌کشیده با پیک در حدود ۱۰۰۰ نفر و دم تا بیش از ۳۵۰۰۰. بیشتر مناطق جمعیت کمی دارند و تعداد کمی منطقه پرجمعیت هستند. این ویژگی نیز دارای پرت‌های قابل توجه است.

۶. **AveOccup (میانگین تعداد افراد در هر خانه):** توزیع چپ‌کشیده با پیک در حدود ۵.۲ تا ۳ و دم بسیار طولانی تا بیش از ۱۲۰۰. مقادیر بسیار بالا کاملاً غیرعادی هستند و نشان‌دهنده داده‌های پرت یا خطا در ثبت اطلاعات است.

۷. **Latitude (عرض جغرافیایی):** توزیع چندمداله با پیک‌های اصلی در حدود ۳۴، ۳۷ و ۳۹ درجه. این الگو با توزیع جغرافیایی مناطق مسکونی در کالیفرنیا (از جنوب تا شمال) همخوانی دارد.

۸. **Longitude (طول جغرافیایی):** توزیع چندمداله با پیک‌های اصلی در حدود −۱۲۲ و −۱۱۸ درجه. این نشان‌دهنده تمرکز جمعیت در نواحی ساحلی (لس آنجلس، سان فرانسیسکو) و دره مرکزی است.

۹. **Median House Value (متغیر هدف):** توزیع چپ‌کشیده با پیک در حدود ۸.۱ تا ۲ واحد (۱۸۰ تا ۲۰۰ هزار دلار) و یک **پیک غیرعادی در ۵ واحد (۵۰۰ هزار دلار).** این پیک در انتهای بالایی به دلیل **سقف گذاری (capping)** در داده‌های اصلی است — در سرشماری ۱۹۹۰، قیمت‌های بالای ۵۰۰ هزار دلار به این مقدار گرد شده‌اند. این امر مدل‌سازی را پیچیده می‌کند و ممکن است نیاز به حذف یا اصلاح این نمونه‌ها باشد.

جمع‌بندی اطلاعات استخراج‌شده از هیستوگرام‌ها

- **توزیع‌های نامتقارن و چپ‌کشیده:** اکثر ویژگی‌ها (به‌ویژه MedInc، AveRooms، AveBedrms، Population، AveOccup) توزیع چپ‌کشیده دارند و نیاز به **تبدیل لگاریتمی یا نرمال‌سازی** دارند.
- **وجود پرت‌های شدید:** ویژگی‌هایی مانند AveRooms، AveBedrms و AveOccup شامل مقادیر بسیار غیرعادی هستند که باید با روش‌های **حذف پرت** یا **جایگزینی** مدیریت شوند.
- **سقف‌گذاری در متغیر هدف:** پیک در ۵۰۰ هزار دلار نشان‌دهنده censoring است و مدل‌های رگرسیون خطی ممکن است در این ناحیه خطای سیستماتیک داشته باشند.
- **اطلاعات جغرافیایی ارزشمند:** توزیع Latitude و Longitude نشان‌دهنده الگوهای فضایی مشخص است که می‌تواند در مهندسی ویژگی (مثلاً محاسبه فاصله از ساحل) مفید باشد.

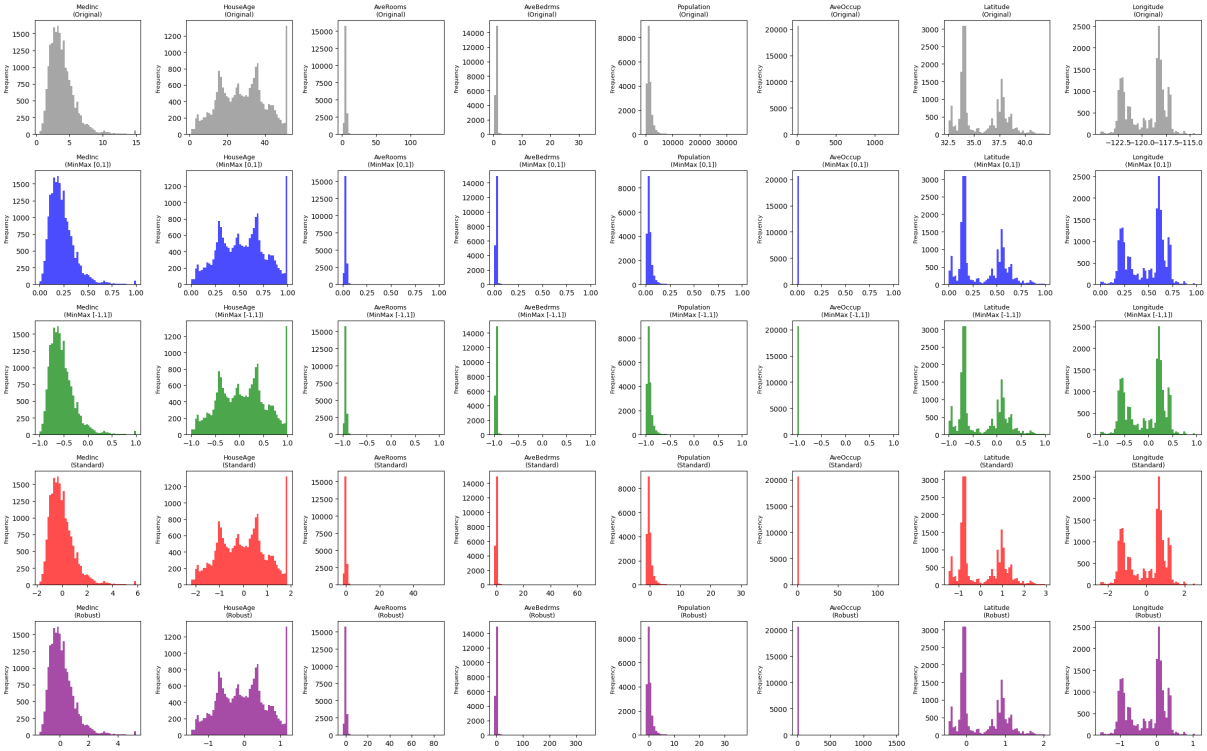
این تحلیل اولیه نشان می‌دهد که پیش‌پردازش داده (نرمال‌سازی، حذف پرت، تبدیل متغیرها) برای بهبود عملکرد مدل‌های یادگیری ماشین ضروری است.

ب: مقایسه روش‌های نرمال‌سازی داده‌ها

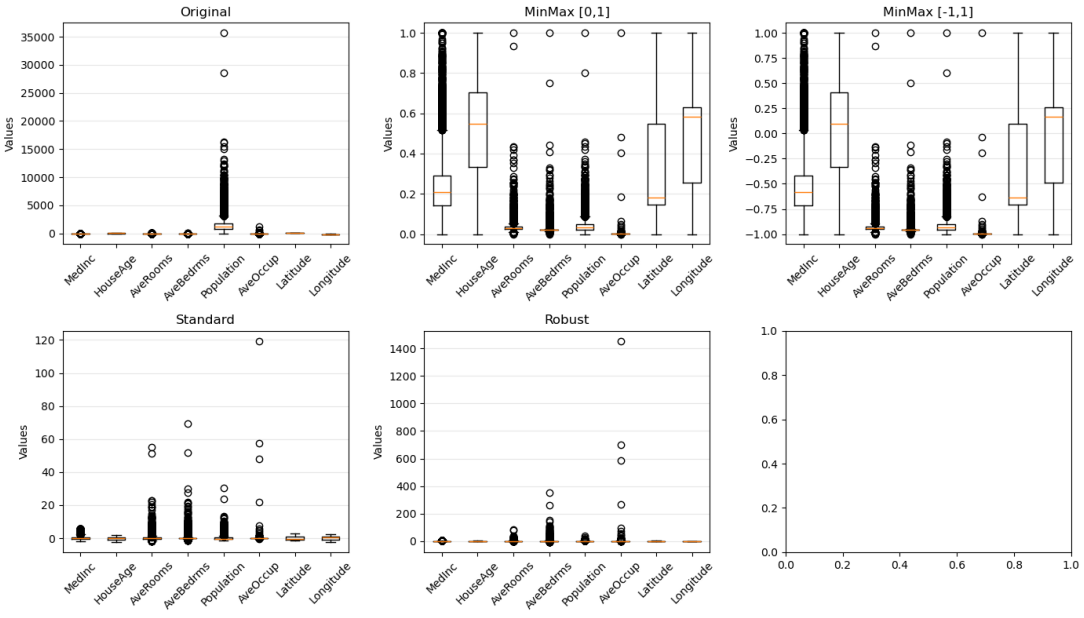
در این بخش، چهار روش نرمال‌سازی زیر بر روی ویژگی‌های مجموعه داده مسکن کالیفرنیا اعمال شده و تأثیر هر یک بر توزیع داده‌ها، مقادیر پرت و آماره‌های توصیفی بررسی شده است:

۱. **نرمال‌سازی مین‌مکس به بازه** $[0, 1]$
۲. **نرمال‌سازی مین‌مکس به بازه** $[-1, 1]$
۳. **نرمال‌سازی استاندارد (Z-score)**
۴. **نرمال‌سازی مقاوم (Robust Scaling)**

شکل ۲ توزیع هیستوگرام هر ویژگی را قبل و بعد از اعمال هر روش نشان می‌دهد. همچنین شکل ۳ با استفاده از نمودار جعبه‌ای (boxplot)، تأثیر هر روش بر مدیریت داده‌های پرت را مقایسه می‌کند.



شکل ۲: مقایسه توزیع هیستوگرام ویژگی‌ها در حالت اصلی و پس از اعمال چهار روش نرمال‌سازی



شکل ۳: نمودار جعبه‌ای (boxplot) برای مقایسه تأثیر روش‌های نرمال‌سازی بر داده‌های پرت

فرمول ریاضی و تفسیر هر روش

در ادامه، برای هر روش، **فرمول ریاضی** و **تفسیر عملی** آن (اینکه دقیقاً چه کاری انجام می‌دهد و چه تأثیری بر داده‌ها دارد) ارائه شده است:

۰.۱ **نرمال‌سازی مین‌مکس به بازه** $[0, 1]$

فرمول:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

تفسیر: این روش کوچک‌ترین مقدار ویژگی را به ۰ و بزرگ‌ترین مقدار را به ۱ نگاشت می‌کند. تمام مقادیر دیگر به صورت خطی در این بازه توزیع می‌شوند. به عبارت دیگر، **مقیاس ویژگی را به بازه استاندارد [۰،۱] تبدیل می‌کند** بدون تغییر در شکل نسبی توزیع. این روش برای الگوریتم‌هایی که به مقیاس مطلق حساس هستند (مثل شبکه‌های عصبی با تابع فعال‌سازی سیگموید) مفید است، اما **بسیار حساس به پرت** است — یک مقدار پرت می‌تواند کل داده‌ها را فشرده کند.

۲. **نرمال‌سازی مین‌مکس به بازه** $[-1, 1]$
فرمول:

$$x' = 2 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1$$

تفسیر: مشابه روش قبلی، اما بازه خروجی $[-1, 1]$ است. ابتدا داده‌ها به $[0, 1]$ نگاشت می‌شوند، سپس با ضرب در ۲ و کم کردن ۱، به بازه متقارن حول صفر تبدیل می‌شوند. این روش برای الگوریتم‌هایی که به **مقیاس متقارن حول صفر** نیاز دارند (مثل برخی روش‌های بهینه‌سازی یا شبکه‌های عصبی با تابع فعال‌سازی tanh) مناسب است. حساسیت به پرت همچنان بالاست.

۳. **نرمال‌سازی استاندارد Scaling (Standard) یا Z-score**
فرمول:

$$x' = \frac{x - \mu}{\sigma}$$

تفسیر: این روش میانگین (μ) ویژگی را به ۰ و انحراف معیار (σ) را به ۱ تبدیل می‌کند. به عبارت دیگر، **داده‌ها را به توزیع نرمال استاندارد (میانگین ۰، واریانس ۱) نزدیک می‌کند.** این روش فرض می‌کند داده‌ها تقریباً نرمال هستند و برای الگوریتم‌های مبتنی بر فاصله (مثل SVM، k-NN، رگرسیون خطی) بسیار مفید است. اما چون از میانگین و انحراف معیار استفاده می‌کند، **تأثیر پرت‌ها بسیار زیاد است** و می‌تواند توزیع را به شدت تحریف کند.

۴. **نرمال‌سازی مقاوم Scaling) (Robust**
فرمول:

$$x' = \frac{x - \text{median}}{Q_3 - Q_1} \quad (\text{IQR} = Q_3 - Q_1)$$

تفسیر: این روش از **میانه (median)** به جای میانگین و **دامنه بین‌چارکی (IQR)** به جای انحراف معیار استفاده می‌کند. میانه نماینده مرکزی مقاوم به پرت است و IQR فقط به ۵۰٪ داده‌های مرکزی وابسته است. بنابراین، **این روش در حضور پرت‌ها بسیار مقاوم است** و توزیع نسبی داده‌های اصلی را بدون تحریف حفظ می‌کند. برای داده‌های واقعی با نویز و پرت (مثل این دیتاست) بهترین انتخاب است.

تحلیل آماری کلی (از خروجی کد)

جدول ۱: مقایسه آماری روش‌های نرمال‌سازی (محدوده میانگین، انحراف معیار، حداقل و حداکثر)

روش	محدوده میانگین	محدوده انحراف معیار	حداقل	حداکثر
اصلی	$[-119.57, 1425.48]$	$[0.47, 1132.46]$	-124.35	35682.00
MinMax[۰،۱]	$[0.0019, 0.5420]$	$[0.0084, 0.2468]$	0.0000	1.0000
MinMax[-۱،۱]	$[-0.9962, 0.0839]$	$[0.0167, 0.4936]$	-1.0000	1.0000
Standard	$[\sim 10^{-15}, \sim 10^{-17}]$	$[1.0000, 1.0000]$	-2.39	119.42
Robust	$[-0.2849, 0.5125]$	$[0.5286, 12.1828]$	-7.66	1455.12

مشاهدات کلیدی از هیستوگرام‌ها و باکس‌پلات‌ها

- **MinMax [۰،۱] و [-۱،۱]:** این روش‌ها تمام مقادیر را به بازه محدود مین‌مکس می‌برند، اما **حساس به پرت** هستند. در ویژگی‌هایی مانند Population و AveRooms، پرت‌های شدید باعث فشرده شدن کل داده‌ها در نزدیکی ۰ یا −۱ می‌شوند.
- **Scaling: Standard** میانگین $0 \approx$ و انحراف معیار $1 =$ برای همه ویژگی‌ها. اما به دلیل استفاده از میانگین و انحراف معیار، **تأثیر پرت‌ها بسیار زیاد است** و توزیع پس از نرمال‌سازی همچنان چپ‌کشیده باقی می‌ماند.
- **Scaling: Robust** از میانه و دامنه بین‌چارکی (IQR) استفاده می‌کند که **مقاوم به پرت** هستند. در باکس‌پلات، مشاهده می‌شود که پرت‌ها در این روش **کمتر کشیده شده‌اند** و توزیع متعادل‌تری دارند.

—

پ: کدام روش برای داده‌های دارای پرت بهتر است؟ چرا؟

Robust Scaling بهترین روش برای داده‌های دارای پرت است.

دلیل ریاضی

- **میانه :(Median)** برخلاف میانگین، تحت تأثیر مقادیر بسیار بزرگ یا کوچک قرار نمی‌گیرد. اگر ۱۰% داده‌ها پرت باشند، میانه همچنان نماینده خوبی از مرکز داده‌هاست.
- **IQR** = $Q_3 - Q_1$: این معیار پراکندگی فقط به ۵۰% داده‌های مرکزی وابسته است و پرت‌های دور (بالای $Q_3 + 1.5 \times IQR$ یا زیر $Q_1 - 1.5 \times IQR$) در محاسبه آن دخیل نیستند.
- **نتیجه:** نرمال‌سازی با میانه و IQR، **مقیاس داده‌های اصلی را بدون تحریف توسط پرت‌ها حفظ می‌کند.**

تأیید با نمونه از داده‌ها (ویژگی MedInc)

جدول ۲: مقایسه ۱۰ نمونه اول ویژگی MedInc پس از نرمال‌سازی

نمونه	اصلی	MinMax[۰،۱]	MinMax[-۱،۱]	Standard	**Robust**
۱	۳۲۵۲.۸	۵۳۹۷.۰	۰۷۹۳.۰	۳۴۴۸.۲	**۱۹۷۶.۲**
۲	۳۰۱۴.۸	۵۳۸۰.۰	۰۷۶۱.۰	۳۳۲۲.۲	**۱۸۶۷.۲**
۳	۲۵۷۴.۷	۴۶۶۰.۰	۰۶۷۹.۰-	۷۸۲۷.۱	**۷۰۷۷.۱**
⋮	⋮	⋮	⋮	⋮	⋮
۱۰	۶۹۱۲.۳	۲۲۰۱.۰	۵۵۹۸.۰-	۰۹۴۵.۰-	**۰۷۱۷.۰**

- در MinMax، مقادیر بالا (مثل ۳۰۸) به نزدیکی ۵.۰ فشرده می‌شوند، در حالی که بیشتر داده‌ها در بازه ۰ تا ۳.۰ هستند **از دست رفتن وضوح.**
- در Standard، به دلیل وجود پرت‌های درآمد بالا، انحراف معیار بزرگ شده و مقادیر معمولی (مثل ۶.۳) به نزدیکی صفر یا منفی می‌روند **تحریف توزیع.**
- در **Robust**، مقادیر منطقی (درآمد ۳ تا ۸) به بازه منطقی $[-1, 3]$ نگاشت می‌شوند و **تفاوت نسبی بین نمونه‌ها حفظ می‌شود.**

شواهد بصری از باکس‌پلات (شکل ۳)

- در روش‌های MinMax و Standard، پرت‌ها (نقاط سیاه) بسیار دور از جعبه هستند و **توزیع نامتعادل** است.
- در Robust، جعبه‌ها متعادل‌تر و پرت‌ها **کمتر برجسته** هستند نشان‌دهنده **مدیریت بهتر نویز.**

نتیجه‌گیری نهایی

Robust Scaling به دلیل استفاده از آماره‌های مقاوم (میانه و ،(IQR بهترین عملکرد را در حضور داده‌های پرت دارد و توزیع نسبی داده‌ها را بدون تحریف حفظ می‌کند. این روش به‌ویژه برای مدل‌های حساس به مقیاس مانند شبکه‌های عصبی، SVM و k-NN توصیه می‌شود.