2)

نرمال‌سازی Min-Max

- داده‌ها را به یک محدوده ثابت مقیاس‌بندی می‌کند، معمولاً [0, 1] یا [0, 10].

- فرمول:

$$X_{\text{scaled}} = a + \left(\frac{(X - X_{\min}) \times (b - a)}{X_{\max} - X_{\min}} \right)$$

- تعیین محدوده: زمانی که نیاز دارید داده‌ها در محدوده خاصی باشند، مانند 0 تا 1، که برای الگوریتم‌هایی که نیاز به ورودی نرمال شده دارند، مانند شبکه‌های عصبی، بسیار مفید است.

- تفسیر شهودی: داده‌های مقیاس‌بندی شده هنگامی که همه مقادیر در یک محدوده شناخته شده باشند، آسان‌تر تفسیر می‌شوند.

- حساس به نقاط outlier: وجود نقاط دورافتاده می‌تواند به طور قابل توجهی بر مقادیر کمینه و بیشینه تأثیر بگذارد و منجر به نرمال‌سازی ضعیف داده‌های باقی‌مانده شود.

نرمال‌سازی Z-Score

تعریف:

- داده‌ها را بر اساس میانگین (μ) و انحراف معیار (σ) مجموعه داده‌ها مقیاس‌بندی می‌کند.

- فرمول:

$$Z = \frac{(X - \mu)}{\sigma}$$

- بهتر بودن در برابر outlier: نسبت به نرمال‌سازی min-max کمتر حساس به نقاط دورافتاده است.

- مقیاس استاندارد: برای الگوریتم‌هایی که فرض می‌کنند داده‌ها به صورت نرمال توزیع شده‌اند (میانگین = 0، انحراف معیار = 1)، مانند تحلیل مؤلفه‌های اصلی (PCA) و برخی الگوریتم‌های خوشه‌بندی، مفید است.

- محدوده ثابت نیست: مقادیر حاصل در یک محدوده خاص محصور نیستند و می‌توانند بر اساس توزیع اصلی داده‌ها از [-1, 1] فراتر بروند.

انتخاب روش مناسب

1. ماهیت داده‌ها: اگر داده‌های شما شامل نقاط دورافتاده است نرمال‌سازی Z-score معمولاً ترجیح داده می‌شود زیرا کمتر به نقاط دورافتاده حساس است. اگر داده‌های شما شامل نقاط دورافتاده قابل توجهی نیست نرمال‌سازی min-max می‌تواند بسیار موثر باشد، به ویژه هنگامی که نیاز دارید داده‌ها در یک محدوده خاص قرار گیرند.

2. نیازهای الگوریتم: برای الگوریتم‌هایی که نیاز دارند داده‌ها در یک محدوده خاص باشند (مانند شبکه‌های عصبی) نرمال‌سازی min-max معمولاً مناسب‌تر است. برای الگوریتم‌هایی که فرض می‌کنند داده‌ها به صورت نرمال توزیع شده‌اند (مانند PCA، خوشه‌بندی) نرمال‌سازی Z-score معمولاً مناسب‌تر است.

3. قابلیت تفسیر: اگر نیاز دارید نتایج به راحتی در یک محدوده مشخص تفسیر شوند نرمال‌سازی min-max مقادیر را در محدوده‌ای که شما مشخص کرده‌اید (مثلاً [0, 1]) فراهم می‌کند.

نتیجه‌گیری:

برای حل این سوال نرمال‌سازی min-max مناسب‌تر به نظر می‌رسد مگر اینکه داده‌ها شامل نقاط دورافتاده قابل توجهی باشد.

(3

ضریب همبستگی پیرسون برابر با 0.19092012847711884 به این معناست که بین دو ستون `num_Grade_Values_Normalized` و `Days_for_Final_Test` همبستگی ضعیف و مثبتی وجود دارد. ضریب همبستگی پیرسون یک مقدار بین -1 و 1 است که نشان می‌دهد دو متغیر چگونه به هم مرتبط هستند:

- 1: همبستگی مثبت کامل. یعنی هر دو متغیر با هم افزایش یا کاهش می‌یابند.

- 0: عدم وجود همبستگی. یعنی تغییرات یک متغیر هیچ تاثیری بر متغیر دیگر ندارد.

- -1: همبستگی منفی کامل. یعنی وقتی یک متغیر افزایش می‌یابد، متغیر دیگر کاهش می‌یابد و بالعکس.

ضریب همبستگی پیرسون نزدیک به 0 (مثل 0.19) نشان می‌دهد که رابطه بین دو متغیر بسیار ضعیف است. در این مورد، اگرچه همبستگی مثبت است (یعنی با افزایش یکی، دیگری نیز کمی افزایش می‌یابد)، اما این همبستگی بسیار ضعیف است و به معنای ارتباط قوی بین این دو متغیر نیست.

4)

0: عدم وجود همبستگی.

1: همبستگی کامل.

مقدار 0.612 نشان می‌دهد که همبستگی بین این دو ستون نسبتاً قوی است، اما همچنان به 1 نزدیک نیست، بنابراین این همبستگی بسیار قوی نیست ولی بیشتر از یک همبستگی ضعیف است.

5)

همبستگی متوسط