Feature Scaling in ML

- Feature Scaling का मतलब होता है **डेटा के अलग-अलग features को एक ही scale पर लाना**, ताकि कोई feature ज़्यादा बड़ा या छोटा होने की वजह से model को bias न करे।
- जैसे: Age → 0 से 100 and Salary → 10,000 से 10,00,000
- Salary का effect ज़्यादा हो जाएगा, इसलिए scaling ज़रूरी है।
- Jaise school mein sabhi bachoo ko ek dress mein padhne jana hota hai. Taki PM/CM/DR./Poor kisi ka beta ho sabko ek saman ek dress mein sako ahmiyat mitle nhi to Bade log chote ho daba denge dominate kr denge chote logo ko ahmiyaat ko khatam kr denge.

Types of Feature Scaling

- MinMax Scaling(Normaliztion) → Set the feature range between [0, 1]
- Standardization(z-score Scaling) → Mean =0, Standard deviation =1
- Robust (scaling) → Handle the outlier's data

MinMax Scaling(Normaliztion):

- **MinMax Scaling(Normaliztion):** Min-Max Scaling एक technique है जिसमें हम data को **fixed range** में बदल देते हैं, आमतौर पर **0 से 1** के बीच।
- इसमें सबसे छोटा value **0** बन जाता है
- सबसे बड़ा value **1** बन जाता है
- बाकी values इनके बीच आ जाती हैं
- Jaise school mein sabhi bachoo ko ek dress mein padhne jana hota hai. Taki PM/CM/DR./Poor kisi ka beta ho sabko ek saman ek dress mein sako ahmiyat mitle nhi to Bade log chote ho daba denge dominate kr denge chote logo ko ahmiyaat ko khatam kr denge.

Formula of Min-Max Scaling

- Min-Max Scaling का Formula

Min-Max Scaling का Formula

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- जहाँ: X = original value
- Xmin = feature का minimum value
- Xmax = feature का maximum value
- Xscaled = scaled value

Example of Min-Max Scaling

- **Machine Learning** = Data + Learning + Prediction
- मान लो हमारे पास **Age** का data है:
- Age
  10
  15
  20
  25

Min = 10
Max = 25
अब अगर Age = 15 को scale करना है:

$$\frac{15-10}{25-10} = \frac{5}{15} = 0.33$$

तो scaled value ≈ **0.33**

Where we use Min-Max Scaling

- KNN (K-Nearest-Neighbors)
- SVM(Support Vector Machines)
- K-means Clustering
- PCA(Principle component analysis)
- Gradient Descent Based Algorithm
- GMM(Gaussion-Mixture Model)

Standardization(Z-Score Scaling)

- Standardization (Z-Score Scaling) एक feature scaling technique है जिसमें data का mean 0 और standard deviation 1 कर दिया जाता है, ताकि सभी features समान scale पर आ जाएँ और ML models बेहतर perform करें।
- Algorithms हर feature को बराबर importance देते हैं।
- Data का **mean (औसत) = 0**
- Data का **standard deviation = 1**
- यानी data को एक common scale पर ले आते हैं।

Formula

- Z-Score Scaling का Formula (Math Formula)
- 📌 **Formula:**

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

- जहाँ:
- **X** = original value
- **μ (mu)** = mean (औसत)
- **σ (sigma)** = standard deviation

Example on Formula

- मान लो हमारे पास ये data है (Feature: Marks)
- 50, 60, 70, 80, 90
- Step 1: Mean (औसत) निकालो

$$\text{Mean} = \frac{50 + 60 + 70 + 80 + 90}{5} = \frac{350}{5} = 70$$

Example of SD

- Feature data(Marks) →50, 60, 70, 80, 90
- **Step 2: Standard Deviation निकालो**

📌 Formula:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

अब हर symbol का मतलब:

- X → data की value
- μ (mu) → mean (औसत)
- n → total values
- Σ → सबको जोड़ना
- √ → square root

Example of SD

- 50 60 70 80 90

Step 1: Mean (औसत) निकालो

$$\text{Mean} = \frac{50 + 60 + 70 + 80 + 90}{5}$$

$$= \frac{350}{5} = 70$$

👉 **Mean = 70**

Example of SD

- Step 2: Mean से दूरी (X – Mean)

X	X – 70
50	-20
60	-10
70	0
80	10
90	20

Example of SD

- Step 3: Square करो

X – 70	$(X - 70)^2$
-20	400
-10	100
0	0
10	100
20	400

Example of SD

- **Step 4: Average लो (Variance)**

$$\text{Variance} = \frac{400 + 100 + 0 + 100 + 400}{5}$$

$$= \frac{1000}{5} = 200$$

- **Step 5: Square Root लो (Standard Deviation)**

$$\text{Standard Deviation} = \sqrt{200}$$

$$\approx 14.14$$

✅ **Final Answer:**

📌 Standard Deviation ≈ 14.14

Now Finally Z-Score

Data:

```
50    60    70    80    90
```

हम पहले ही निकाल चुके हैं:

- **Mean (μ) = 70**
- **Standard Deviation (σ) = 14.14**

Z-Score का Formula

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

Now Finally Z-Score

- **Step-by-Step Z-Score Calculation**
- 1 X = 50

$$Z = \frac{50 - 70}{14.14} = \frac{-20}{14.14} \approx -1.41$$

- 2 X = 60

$$Z = \frac{60 - 70}{14.14} = \frac{-10}{14.14} \approx -0.71$$

- 3 X = 70

$$Z = \frac{70 - 70}{14.14} = 0$$

- 4 X = 80

$$Z = \frac{80 - 70}{14.14} = \frac{10}{14.14} \approx 0.71$$

- 5 X = 90

$$Z = \frac{90 - 70}{14.14} = \frac{20}{14.14} \approx 1.41$$

✅ Final Z-Score Table

Original Value	Z-Score
50	-1.41
60	-0.71
70	0
80	0.71
90	1.41

Now Mean of Z-Score column

- **Machine Learning** = Data + Learning + Prediction
- Mean of Z-score column = -1.40-0.71+0+0.71+1.40/5 → 0 → Mean = 0
- Mean = 0
- Now Z-Score ka Standard Deviation : $\sigma = \sqrt{\frac{\sum(X-\mu)^2}{n}}$

✅ **1-Step Calculation**

इन values का **Mean (μ) = 0**
(क्योंकि values symmetric हैं: − और + बराबर)

अब सीधे formula में डालते हैं:

$$\sigma = \sqrt{\frac{(-1.40)^2 + (-0.71)^2 + 0^2 + (0.71)^2 + (1.40)^2}{5}}$$

$$= \sqrt{\frac{1.96 + 0.5041 + 0 + 0.5041 + 1.96}{5}}$$

$$= \sqrt{\frac{4.9282}{5}}$$

$$= \sqrt{0.98564} \approx 0.99\ (\approx 1)$$

When we use Standardization

- जब data **normal distribution** के आस-पास हो
- जब outliers हों
- जब ML algorithm distance पर depend करता हो
- **Distance-based Algorithms के लिए ज़रूरी**
- **Example:** Linear Regression, Logistic Regression, KNN, SVM, Neural Networks ये **distance** पर काम करते हैं।

Where No need Standardization

- कब Standardization नहीं करना चाहिए?
- ❌ Decision Tree
- ❌ Random Forest
- क्योंकि ये **splitting** पर काम करते हैं, distance पर नहीं।

Feature Scaling → Robust Scaling

- **Robust Scaling** एक **feature scaling technique** है
- जो **outliers (बहुत ज़्यादा या बहुत कम values)** से ज़्यादा प्रभावित **नहीं होती**।
- जब data में **outliers हों**, तब RobustScaler सबसे अच्छा रहता है।
- Outlier क्या होता है? मान लीजिए आपके पास Salary Data hai: 150, 152, 155, 158, 160, 162, 3000
- यहाँ **3000** एक **outlier** है, क्योंकि यह बाकी मानों से बहुत अलग है।
- Robust Scaling Median (मध्य मान), **IQR (Interquartile Range = Q3 – Q1)** पर आधारित होता है
- इसलिए अगर डेटा में outliers हों, तो **Robust Scaler सबसे अच्छा विकल्प** होता है।

Formula of Robust Scaler

- **Machine Learning** = Data + Learning + Prediction
- **Robust Scaling का फ़ॉर्मूला**

$$X_{scaled} = \frac{X - Median}{IQR}$$

- जहाँ: **Median** = बीच का मान
- **IQR** = Q3 – Q1 (डेटा का बीच का फैलाव)

Formula of Median

- **Step 1:** सभी संख्याओं को **ascending order** (छोटे से बड़े) में लिखो,
- **Step 2:** कुल संख्याएँ = n
- **(a) जब nOdd (विषम) हो**

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \ term$$

- **Example:**
  Data: 3, 5, 7, 9, 11, $n = 5$

$$\text{Median} = \frac{5+1}{2} = 3^{rd} \ term = 7$$

Formula of Median

- **(b) जब nEven (सम) हो**

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \ term + \left(\frac{n}{2}+1\right)^{th} \ term}{2}$$

- Data: 2, 4, 6, 8 and $n = 4$

$$\text{Median} = \frac{4+6}{2} = 5$$

Quartiles

- Quartiles एक **data set** को चार बराबर हिस्सों (quarters) में बाँटते हैं। यानी डेटा के 4 हिस्सों में से हर हिस्सा कुल डेटा का 25% होता है।
- **Minimum --- Q1 --- Q2 --- Q3 --- Maximum**
- **Q1 (First Quartile)** → Data का 25% हिस्सा Q1 से नीचे होता है
- **Q2 (Second Quartile / Median)** → Data का 50% हिस्सा Q2 से नीचे होता है
- **Q3 (Third Quartile)** → Data का 75% हिस्सा Q3 से नीचे होता है
- Q1 = 25%, Q2 = 50%, Q3 = 75%

Quartiles

- Quartiles निकालने का तरीका (Un-grouped Data)
- **Step 1:** Data को **ascending order** में लगाओ
- **Step 2:** Total number of observations = n
- **Step 3:** Formula use करो

$$Q_k = \frac{k(n+1)}{4} \text{ th term जहाँ } k = 1,2,3$$

- $Q_1 = \frac{1(n+1)}{4}$ th term
- $Q_2 = \frac{2(n+1)}{4} = \text{Median}$
- $Q_3 = \frac{3(n+1)}{4}$ th term

Example:

Data: 2, 4, 6, 8, 10, 12, 14, 16

- $n = 8$
- $Q_1 = \frac{1(8+1)}{4} = \frac{9}{4} = 2.25$th term → 2nd term + 0.25*(3rd - 2nd) = 4 + 0.25*(6-4) = 4.5
- $Q_2 = Median = \frac{9}{2}$th term = 4.5th term → 8 + 0.5*(10-8) = 9
- $Q_3 = \frac{3*9}{4} = 6.75$th term → 6th + 0.75*(7th-6th) = 12 + 0.75*(14-12) = 13.5

What

Age	Qualification	Salary
2	0	20
4	1	25
6	0	30
8	1	35
10	0	40
12	1	45
14	0	200
16	1	220

👉 मैंने **Salary** में बड़े values (outliers) जानबूझकर रखे हैं, क्योंकि **Robust Scaling outliers के लिए best होता है**।

What

Robust Scaling Formula

$$X_{scaled} = \frac{X - \text{Median}}{IQR}$$

जहाँ

$$IQR = Q_3 - Q_1$$

---

3 Step-1: Median निकालो

- **Age**

Data: 2,4,6,8,10,12,14,16

Median (Q2) =

$$\frac{8 + 10}{2} = 9$$

What

4 Step-2: Q1 और Q3 निकालो

- **Age**

Lower half: 2,4,6,8

$$Q_1 = \frac{4+6}{2} = 5$$

Upper half: 10,12,14,16

$$Q_3 = \frac{12+14}{2} = 13$$

$$IQR_{Age} = 13 - 5 = 8$$

What

- Age Scaling Formula

$$Age_{scaled} = \frac{Age - 9}{8}$$

Age	Calculation	Scaled
2	(2−9)/8	−0.875
4	(4−9)/8	−0.625
6	(6−9)/8	−0.375
8	(8−9)/8	−0.125
10	(10−9)/8	0.125
12	(12−9)/8	0.375
14	(14−9)/8	0.625
16	(16−9)/8	0.875

What

- **Machine Learning** = Data + Learning + Prediction