AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR 80th anniversary
Touching lives through innovation

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

# Standardizing Explanation

DIRISA Datathon | Instructor: Kgaugelo Mmakola

# Standardizing = Standard Scaling

- Standardizing means:
  - **Making the mean of each feature 0**
  - **Making the standard deviation of each feature 1**

# Why are we standardizing?

- We **standardize** data to make **machine learning models** work better, faster, and more fairly.

**Easy analogy:**

Think of it like racing three different vehicles:

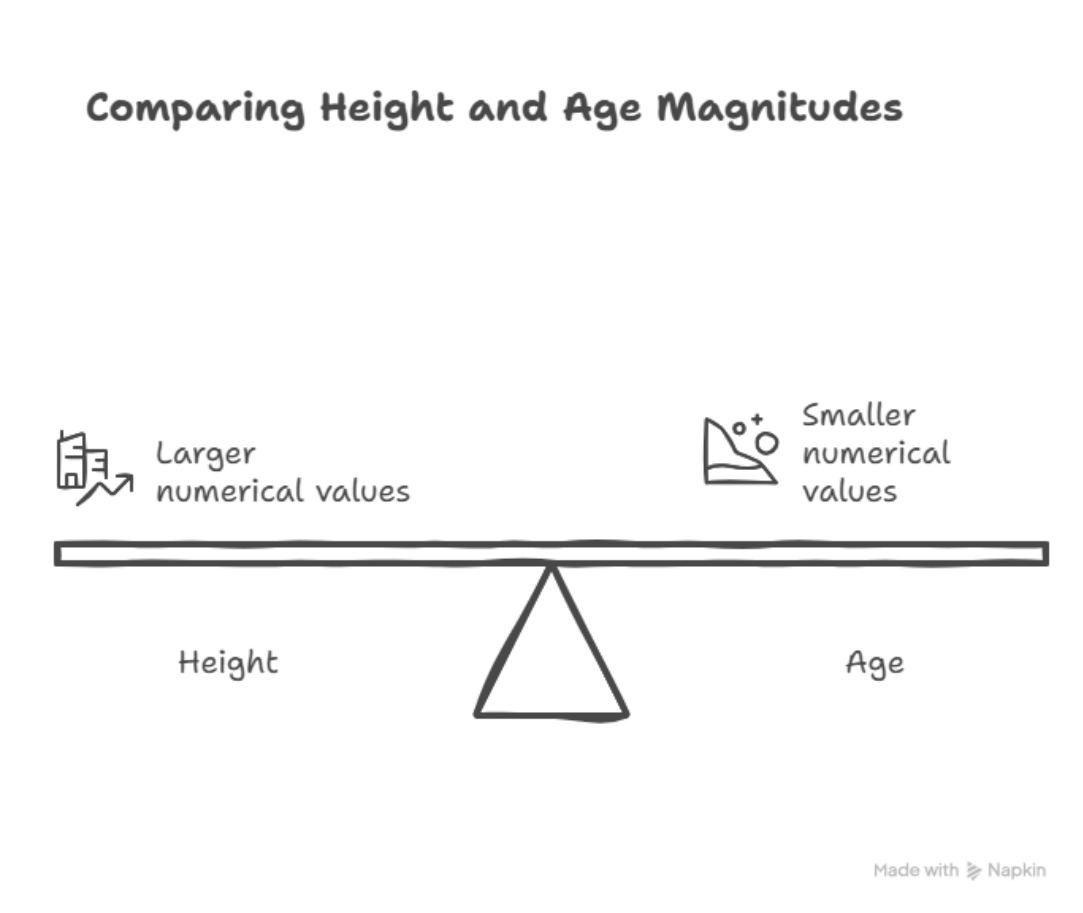
- A car (speed in km/h)
- A bicycle (speed in m/s)
- A plane (speed in Mach)
- If you compare their speeds **directly**, the plane wins every time, but that's not fair!

- You must convert everything to the **same unit** first.

**Standardizing = Putting everything on the same "unit" or "level"** so you can compare them fairly.

# Standard Scaler

- You have a list of students, and for each student, you have two things:
- Their **height** (in centimeters) and their **age** (in years).
- Height might be around 150 to 190 cm.
- Age might be around 13 to 18 years.
- These numbers differ significantly in size; height numbers are larger, and age numbers are smaller.



# Why is that a problem?

- If a computer is trying to learn something from this data (like guessing if a student will play basketball well), it might get confused because height numbers are much bigger than age numbers. The computer could think height is *more important* just because the numbers are bigger.

# What does fit_transform do?

- It **changes all the numbers** so they are on the same scale, like making height and age both be measured in the same “units” so the computer treats them fairly.

# How?

It looks at all the heights and finds the *average height* and how much the heights spread out (called ***standard deviation***).

- It does the same for ages.
- Then it changes each height and age by subtracting the average and dividing by how much they spread out.
- After this, the average for both height and age becomes zero, and the numbers are all about the same size.

# Why do we do this?

- To help the computer learn better and faster.

- To make sure no one feature (like height or age) dominates just because of bigger numbers.

# QUICK DEMO

| Student | Height (cm) | Age (years) |
|---|---|---|
| A | 170 | 15 |
| B | 180 | 17 |
| C | 160 | 14 |

# Step 1: Calculate the average (mean) for each feature

- Average Height = (170 + 180 + 160) ÷ 3 = 170 cm
- Average Age = (15 + 17 + 14) ÷ 3 ≈ 15.33 years

# Step 2: Calculate how much the numbers spread out (standard deviation)

- **For Height:**

Differences from **mean:** (170-170)=0, (180-170)=10, (160-170)=-10
Square differences: $0^2=0$, $10^2=100$, $(-10)^2=100$
Average square difference = (0+100+100)/3 = 66.67
**Standard deviation** = $\sqrt{66.67} \approx 8.16$ cm

- **For Age:**

Differences from mean: (15-15.33)=-0.33, (17-15.33)=1.67, (14-15.33)=-1.33
Square differences: $0.33^2=0.11$, $1.67^2=2.79$, $1.33^2=1.77$
Average square difference = (0.11 + 2.79 + 1.77)/3 = 1.56
**Standard deviation** = $\sqrt{1.56} \approx 1.25$ years

# Step 3: Scale each number

- For each value:

**Scaled value = (original value - mean) ÷ standard deviation**

| Student | Height scaled | Age scaled |
|---|---|---|
| A | (170 - 170) ÷ 8.16 = 0 | (15 - 15.33) ÷ 1.25 ≈ -0.26 |
| B | (180 - 170) ÷ 8.16 ≈ 1.22 | (17 - 15.33) ÷ 1.25 ≈ 1.34 |
| C | (160 - 170) ÷ 8.16 ≈ -1.22 | (14 - 15.33) ÷ 1.25 ≈ -1.07 |

# Why is it important?

- **Fair comparison** – Features with bigger numbers don't dominate.
- **Faster training** – Models learn quicker on standardized data.
- **Better accuracy** – Especially for models like:
  - Linear regression
  - Logistic regression
  - K-Nearest Neighbors (KNN)
  - Support Vector Machines (SVM)

# Recap

- We standardize data so that all features are treated **equally and fairly** when training a machine learning model. It helps the model **learn better and make smarter decisions.**


Thank you!!
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA