Day 4

What is Feature Scaling?

Feature scaling is a technique used to standardise or normalise the range of independent variables (features) in a dataset. It adjusts the values of features so they are on a comparable scale, typically within a specific range(like 0 to 1)or with a mean of 0 and standard deviation of 1.

Why is Feature Scaling Needed?

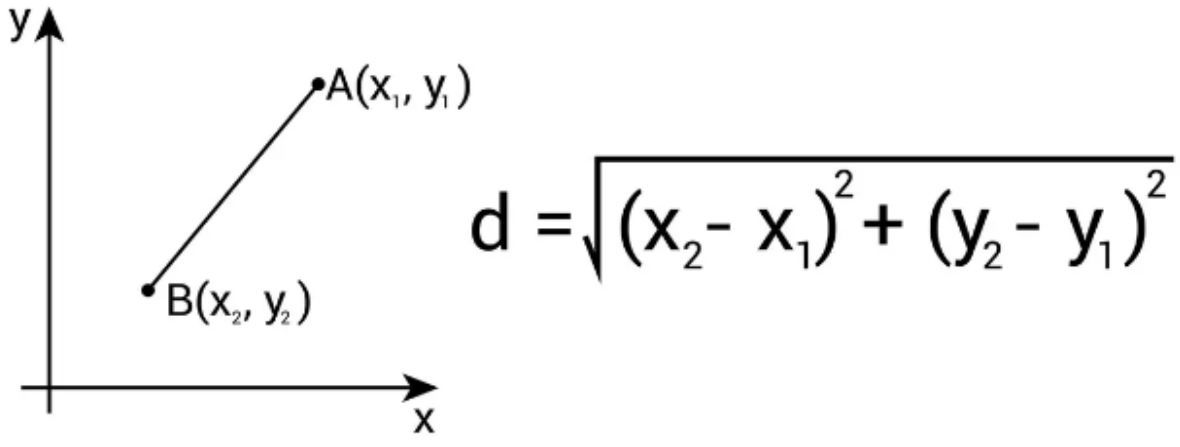
1. **Improves Model Performance**: Algorithms like gradient descent converge faster with scaled data, leading to better performance.
2. **Prevents Bias**: Algorithms that calculate distances between data points (like KNN, SVM) or that are sensitive to the magnitude of features (like linear regression) may produce biased results if features are on different scales.
3. **Enhances Interpretability**: Scaling allows features to contribute more equally to the model, improving the interpretability of results.

Simple Example:

Imagine you are predicting house prices using two features using KNN:

- **Size** (in square feet) which ranges from 500 to 2000
- **Number of bedrooms** which ranges from 1 to 5

Euclidean distance:



If A(x1,y1) >>>>> B(x2,y2)

Point A will dominant and KNN will not perform well.

Without scaling, the algorithm might give more weight to size simply because it has a larger range of values, ignoring the importance of the number of bedrooms. By scaling both features, they contribute more equally to the prediction.

Standardisation:

- **Description**: Standardisation rescales features to have a mean of 0 and a standard deviation of 1.
- **Formula**:

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

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

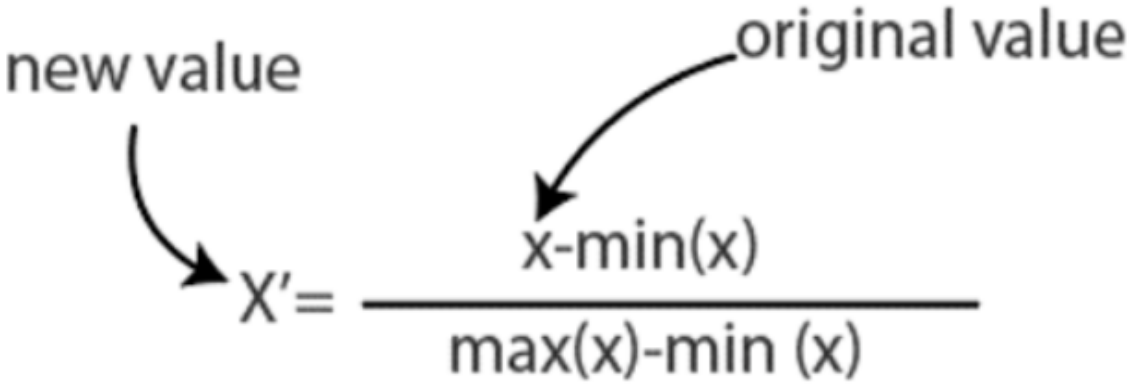
- **When to Use**: Useful when the data follows a Gaussian (normal) distribution and when algorithms assume standardised data (e.g., SVM, logistic regression).

Normalisation:

Normalisation is a technique used to adjust the values of features in a dataset to a common scale without distorting the differences in the ranges of values. It typically scales the data to a range of [0, 1] or [-1, 1].

Types of Normalisation:

- **Min-Max Normalisation**:
  - **Description**: Rescales the feature values to a specified range, usually [0, 1].
  - **Formula**:



  - **Use Case**: Useful when the distribution of data is not Gaussian and when you want to scale features to a specific range.

2. Robust Scaling is a normalisation technique that scales features using statistics that are robust to outliers. Unlike standard scaling methods that use the mean and standard deviation, robust scaling uses the **median** and **interquartile range (IQR)**, making it more resistant to the influence of outliers.

- **Formula**:

$$X_{robust} = \frac{X - X_{median}}{X_{75th} - X_{25th}}$$

$$where\ X_{75th} - X_{25th} = Interquartile\ range$$

Why Use Robust Scaling?

- **Outlier Resistance**: Since it uses the median and IQR, robust scaling is not affected by extreme values or outliers, making it suitable for datasets with outliers.
- **Stable Range**: It scales the data around the median, usually bringing the feature values to a range similar to [-1, 1] for most data points.

3. Max Abs Normalisation:

- **Description**: Scales the data by dividing by the maximum absolute value, resulting in values ranging between -1 and 1.
- **Formula**:

$$X_i' = \frac{X_i}{abs(X_{max})}$$

- **Use Case**: Useful when the data contains outliers, as it doesn't shift/center the data but only rescales.