Curse of Dimensionality

➢ What is the Curse of Dimensionality?

The **curse of dimensionality** refers to the **problems and challenges** that arise when working with **high-dimensional data** (data with many features or variables).

As dimensions increase:

- Data becomes **sparse**
- **Distance metrics become meaningless**
- Models may **overfit** or **fail to generalize**
- Computation becomes **slower and more expensive**

➢ Impact on Machine Learning:

- Harder to **visualize** and interpret data.
- **KNN**, clustering, and distance-based models become unreliable.
- Requires **more training data** to represent the space adequately.
- Increases risk of **overfitting** due to more noise and redundancy.

➢ *Advantages of High Dimensions:*

- May capture **complex patterns**.
- **More features** might mean **more information**, if features are relevant.

✗ *Disadvantages:*

- **Increased complexity** and computation.
- **Harder to detect relationships** between features.
- May include **irrelevant or redundant features**.

➢ Solution: Dimensionality Reduction

1. Feature Selection

Keep only the **most relevant features**:

- **Forward Selection**: Start with none, add one feature at a time.
- **Backward Elimination**: Start with all features, remove least important one by one.

2. Feature Extraction

Transform data into a **lower-dimensional space** while preserving key patterns:

- **PCA (Principal Component Analysis)** – Projects data along axes of maximum variance.
- **LDA (Linear Discriminant Analysis)** – Maximizes **class separability**.

- **t-SNE (t-distributed Stochastic Neighbor Embedding)** – Good for **visualizing high-dimensional data** in 2D/3D.

➢ Key Takeaway:

The curse of dimensionality can **degrade model performance** in high-dimensional spaces. Combat it using **feature selection** and **feature extraction** to **simplify your data**, improve **efficiency**, and **boost model accuracy**.
