COLLEGE OF ENGINEERING AND MANAGEMENT, KOLAGHAT

NAME: - PUSPITA PANJA

COLLEGE ROLL NUMBER: - CSE/22/065

SUBJECT: - MACHINE LEARNING

SUBJECT CODE: - PEC-CS701E

UNIVERSITY RAGISTRAION NO.: - 221070110076

UNIVERSITY ROLL NO.: - 10700122071

DEPARTMENT: - COMPUTER SCIENCE AND ENGINEERING

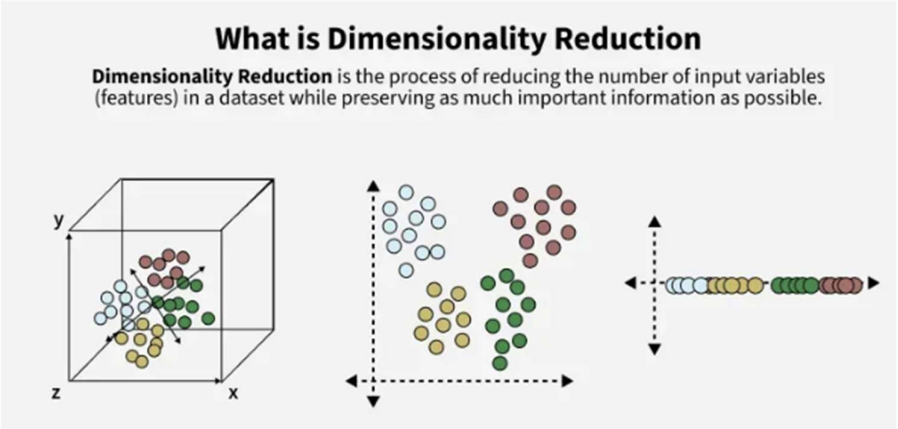
SECTION: - 'C'

ACADEMIC SESSION: - 2022-2026

Table of Contents:

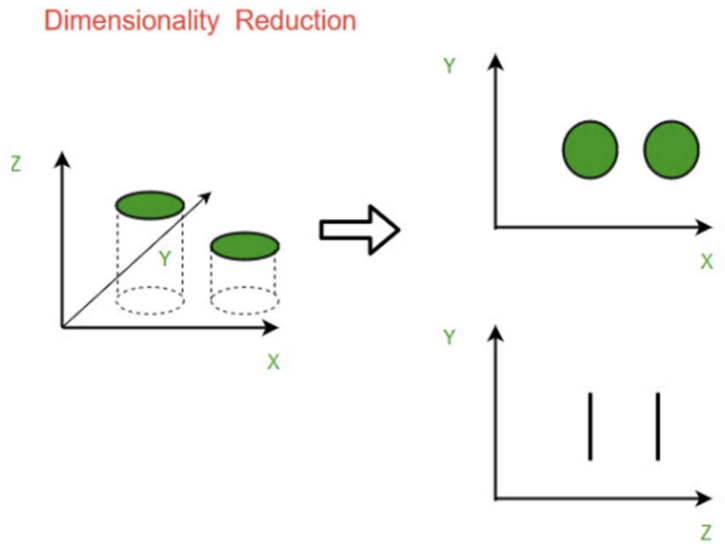
Introduction to Dimensionality Reduction:

When working with machine learning models, datasets with too many features can cause issues like slow computation and overfitting. Dimensionality reduction helps to reduce the number of features while retaining key information. Techniques like principal component analysis (PCA), singular value decomposition (SVD) and linear discriminant analysis (LDA) convert data into a lower-dimensional space while preserving important details.



Working Principle of Dimensionality Reduction:

Let's understand how dimensionality Reduction is used with the help of example. Imagine a dataset where each data point exists in a 3D space defined by axes X, Y and Z. If most of the data variance occurs along X and Y then the Z-dimension may contribute very little to understanding the structure of the data.



- Before Reduction we can see that Data exist in 3D (X, Y, Z). It has high redundancy and Z contributes little meaningful information
- On the right after reducing the dimensionality the data is represented in **lower-dimensional spaces**. The top plot (X-Y) maintains the meaningful structure while the bottom plot (Z-Y) shows that the Z-dimension contributed little useful information.

This process makes data analysis more efficient, improving computation speed and visualization while minimizing redundancy.

Dimensionality Reduction Techniques:

Dimensionality reduction techniques can be broadly divided into two categories:

1. Feature Selection

Feature selection chooses the most relevant features from the dataset without altering them. It helps remove redundant or irrelevant features, improving model efficiency. Some common methods are:

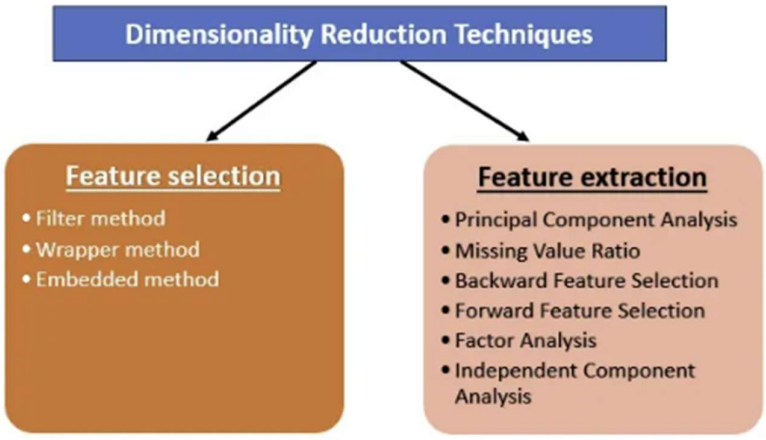
- Filter methods rank the features based on their relevance to the target variable.
- Wrapper methods use the model performance as the criteria for selecting features.

- Embedded methods combine feature selection with the model training process.

2. Feature Extraction

Feature extraction involves creating new features by combining or transforming the original features. These new features retain most of the dataset's important information in fewer dimensions. Common feature extraction methods are:

1. **Principal Component Analysis (PCA):** Converts correlated variables into uncorrelated 'principal components, reducing dimensionality while maintaining as much variance as possible enabling more efficient analysis.
2. **Missing Value Ratio:** Variables with missing data beyond a set threshold are removed, improving dataset reliability.
3. **Backward Feature Elimination**: Starts with all features and removes the least significant ones in each iteration. The process continues until only the most impactful features remain, optimizing model performance.
4. **Forward Feature Selection:** Forward Feature Selection Begins with one feature, adds others incrementally and keeps those improving model performance.
5. **Random Forest**: Random forest Uses decision trees to evaluate feature importance, automatically selecting the most relevant features without the need for manual coding, enhancing model accuracy.
6. **Factor Analysis**: Groups variables by correlation and keeps the most relevant ones for further analysis.
7. **Independent Component Analysis (ICA)**: Identifies statistically independent components, ideal for applications like 'blind source separation' where traditional correlation-based methods fall short.



Dimensionality Reduction Real World Examples:

Dimensionality reduction plays an important role in many real-world applications such as text categorization, image retrieval, gene expression analysis and more. Here are a few examples:

1. **Text Categorization:** With vast amounts of online data dimensionality reduction helps classify text documents into predefined categories by reducing the feature space like word or phrase features while maintaining accuracy.
2. **Image Retrieval**: As image data grows indexing based on visual content like colour, texture, shape rather than just text descriptions has become essential. This allows for better retrieval of images from large databases.
3. **Gene Expression Analysis**: Dimensionality reduction accelerates gene expression analysis help to classify samples like leukaemia by identifying key features, improve both speed and accuracy.
4. **Intrusion Detection**: In cybersecurity dimensionality reduction helps analyse user activity patterns to detect suspicious behaviours and intrusions by identifying optimal features for network monitoring.

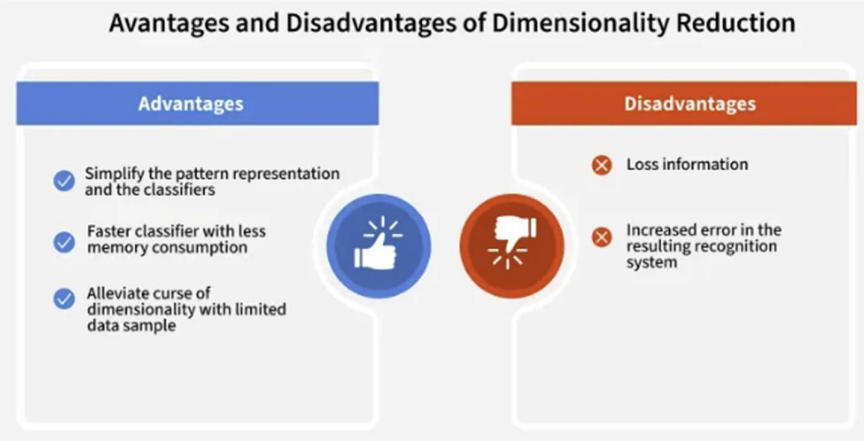
Advantages of Dimensionality Reduction:

As seen earlier high dimensionality makes models inefficient. Let's now summarize the key advantages of reducing dimensionality.

- **Faster Computation**: With fewer features machine learning algorithms can process data more quickly. This results in faster model training and testing which is particularly useful when working with large datasets.
- **Better Visualization**: As we saw in the earlier figure reducing dimensions makes it easier to visualize data and reveal hidden patterns.
- **Prevent Overfitting**: With few features models are less likely to memorize the training data and overfit. This helps the model generalize better to new, unseen data improve its ability to make accurate predictions.

Disadvantages of Dimensionality Reduction:

- **Data Loss & Reduced Accuracy:** Some important information may be lost during dimensionality reduction and affect model performance.
- **Choosing the Right Components:** Deciding how many dimensions to keep is difficult as keeping too few may lose valuable information while keeping too many can led to overfitting.



Conclusion:

Dimensionality reduction plays a crucial role in real-world applications by improving the efficiency, accuracy, and interpretability of machine learning models, as well as enabling better visualization and analysis of complex datasets.

References:

- https://www.geeksforgeeks.org/machine-learning/dimensionality-reduction/
- https://www.datacamp.com/tutorial/understanding-dimensionality-reduction