Feature Selection And PCA In Data Mining

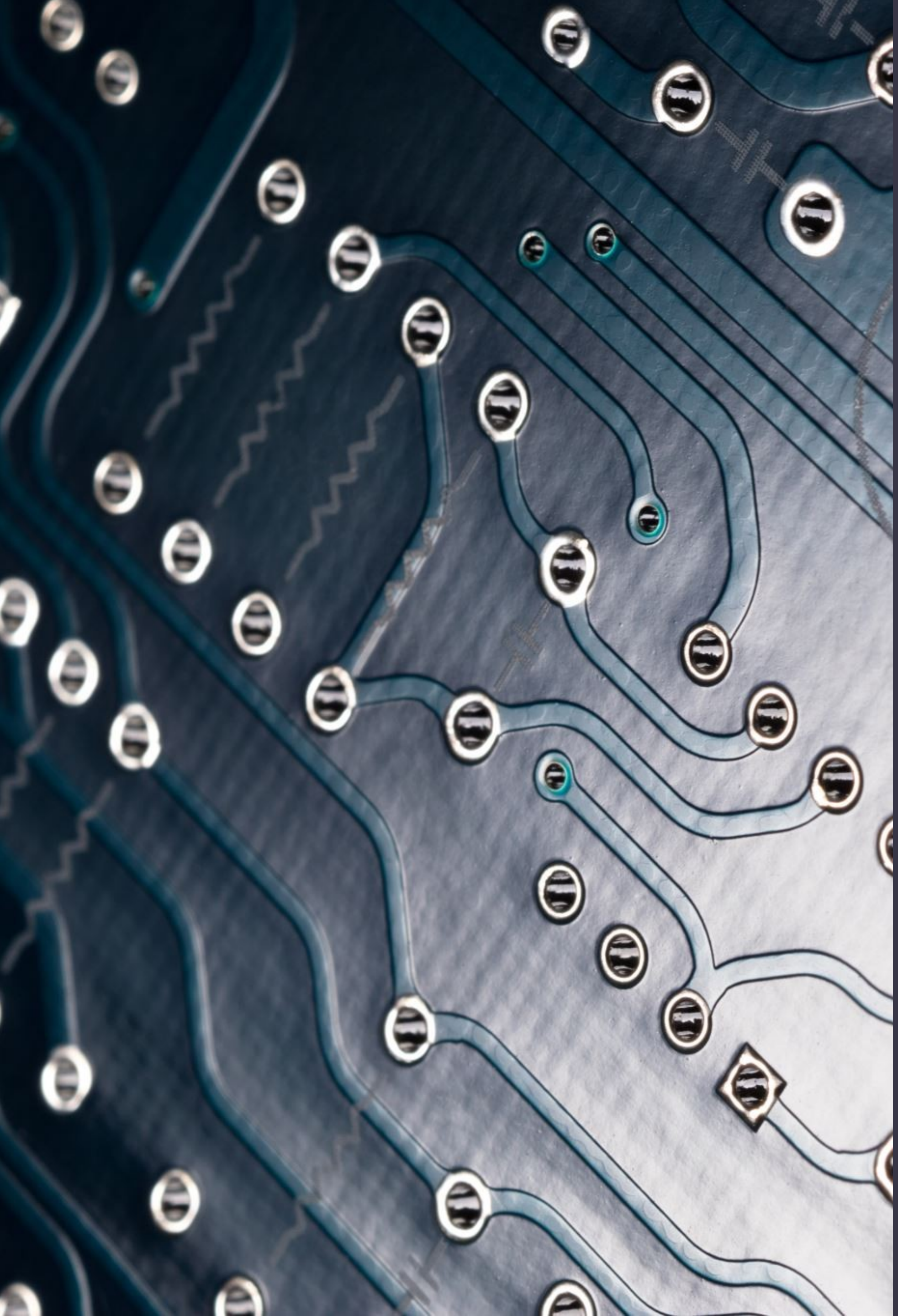
UTKU ÖZYIĞIT

İSHAK AYDEMIR

Unveiling The Power Of Dimensionality Reduction In Data Mining

Introduction- The Challenge of High-Dimensional Data The Data Deluge

- Modern datasets often contain a vast number of features (dimensions).
- **Curse of Dimensionality:**
- Increased computational cost and time.
- Difficulty in visualization.
- Degraded model performance (overfitting, sparsity).
- More data required to achieve statistical significance.



What is Dimensionality Reduction?

The Core Idea

- Process of reducing the number of random variables under consideration.
- Transforms data from a high-dimensional space to a low-dimensional space.
- Aims to preserve meaningful properties of the original data as much as possible.

Two Main Categories

Two Main Categories

1. **Feature Selection:** Selecting a subset of the original features.
2. **Feature Extraction:** Transforming features into a new, smaller set of features.

Feature Selection- Definition & Goal

Feature Selection

Definition: The process of choosing a subset of relevant features for use in model construction.

Goal:

- Improve model accuracy.
- Reduce training time.
- Enhance model interpretability.
- Mitigate the curse of dimensionality.
- Reduce overfitting.

Types of Feature Selection Methods

1. Filter Methods

Concept: Select features based on their intrinsic properties (e.g., correlation, statistical scores) without involving any learning algorithm.

Pros: Computationally efficient, independent of the learning algorithm.

Cons: May select redundant features, ignores interaction with the model.

Examples:

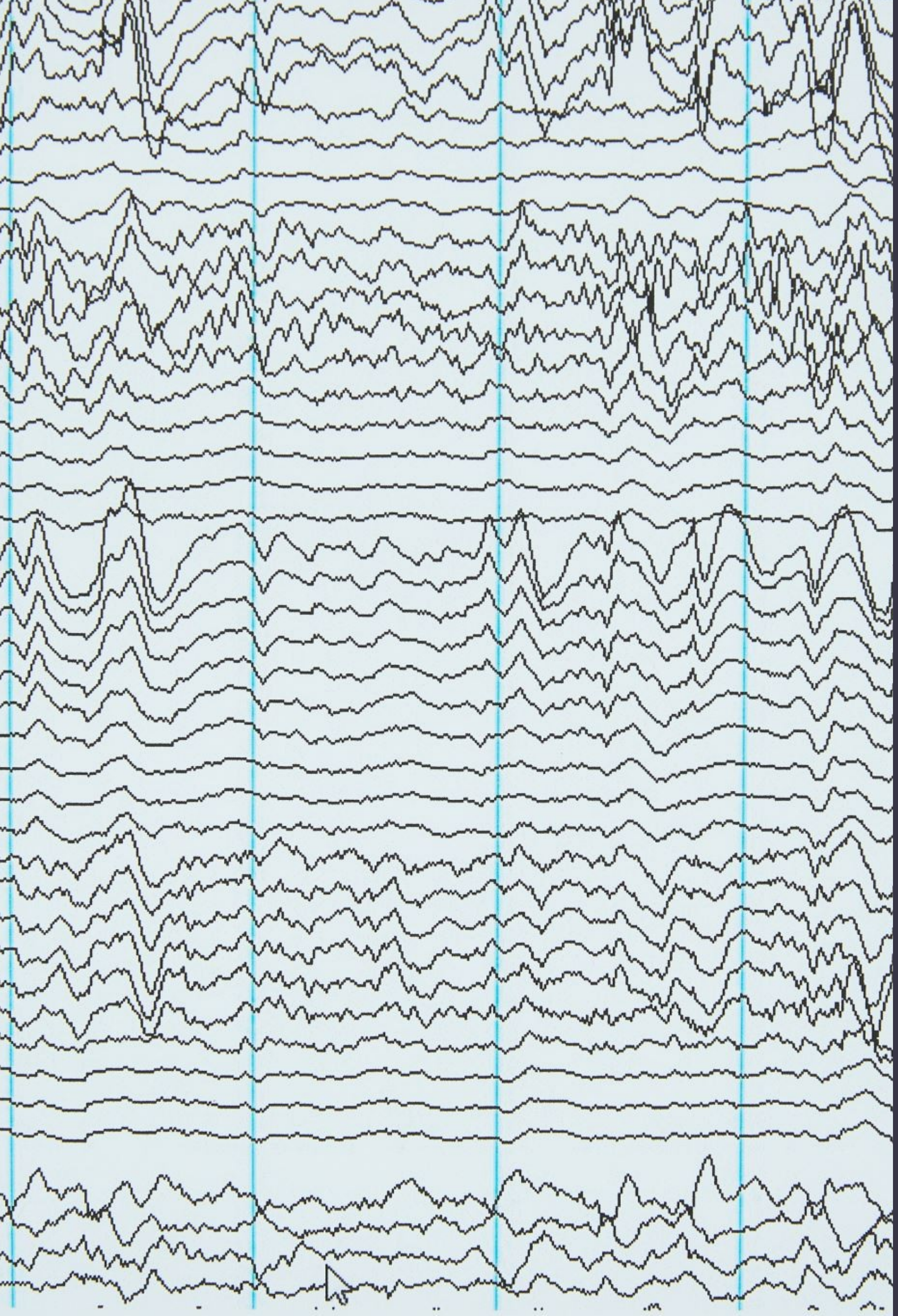
- Variance Threshold
- Chi-squared test (χ2)
- Information Gain
- Correlation Coefficient

Wrapper Methods

- **Concept:** Use a specific machine learning algorithm to evaluate the performance of different subsets of features.
- **Pros:** Account for model bias, often yield better predictive accuracy.
- **Cons:** Computationally intensive, prone to overfitting to the specific model.
- **Examples:**
- Forward Selection
- Backward Elimination
- Recursive Feature Elimination (RFE)

Embedded Methods

- **Concept:** Feature selection is integrated into the model training process itself.
- **Pros:** Less computationally expensive than wrappers, considers feature interactions.
- **Cons:** Method-dependent.
- **Examples:**
  - Lasso Regression (L1 regularization)
  - Ridge Regression (L2 regularization)
  - Decision Trees (feature importance)



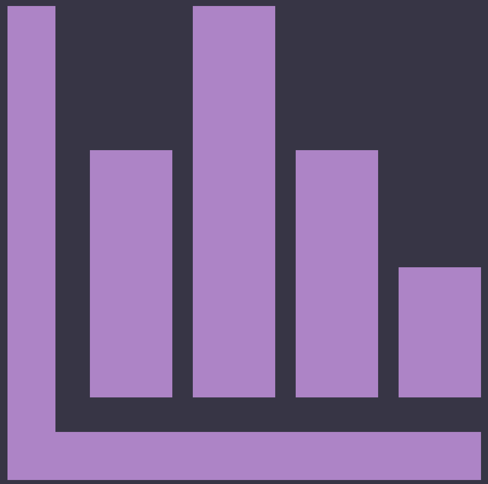
Principal Component Analysis (PCA)

- **Definition:** A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- **Goal:**
- Reduce dimensionality while retaining as much variance as possible.
- Visualize high-dimensional data.
- Remove noise from data.
- Pre-processing for other machine learning algorithms.

How PCA Works (Simplified)

The Core Steps

1. **Standardize the Data:** Ensure all features contribute equally by scaling them.
2. **Calculate the Covariance Matrix:** Understand how features vary with respect to each other.
3. **Compute Eigenvectors and Eigenvalues:**
    1. **Eigenvectors:** Represent the directions (principal components) of maximum variance.
    2. **Eigenvalues:** Indicate the magnitude of variance along each eigenvector.
4. **Sort Eigenvalues:** Order principal components by their explained variance (from highest to lowest).
5. **Select Principal Components:** Choose the top 'k' eigenvectors that capture a significant amount of variance.
6. **Transform the Data:** Project the original data onto the new subspace defined by the selected principal components.



PCA- Key Concepts

Principal Components

- New variables that are linear combinations of the original variables.
- Orthogonal (uncorrelated) to each other.
- The first principal component accounts for the largest possible variance, and each succeeding component accounts for the highest remaining variance.

Explained Variance Ratio

- Indicates the proportion of variance in the original data that is captured by each principal component.
- Helps in determining the optimal number of components to retain.