

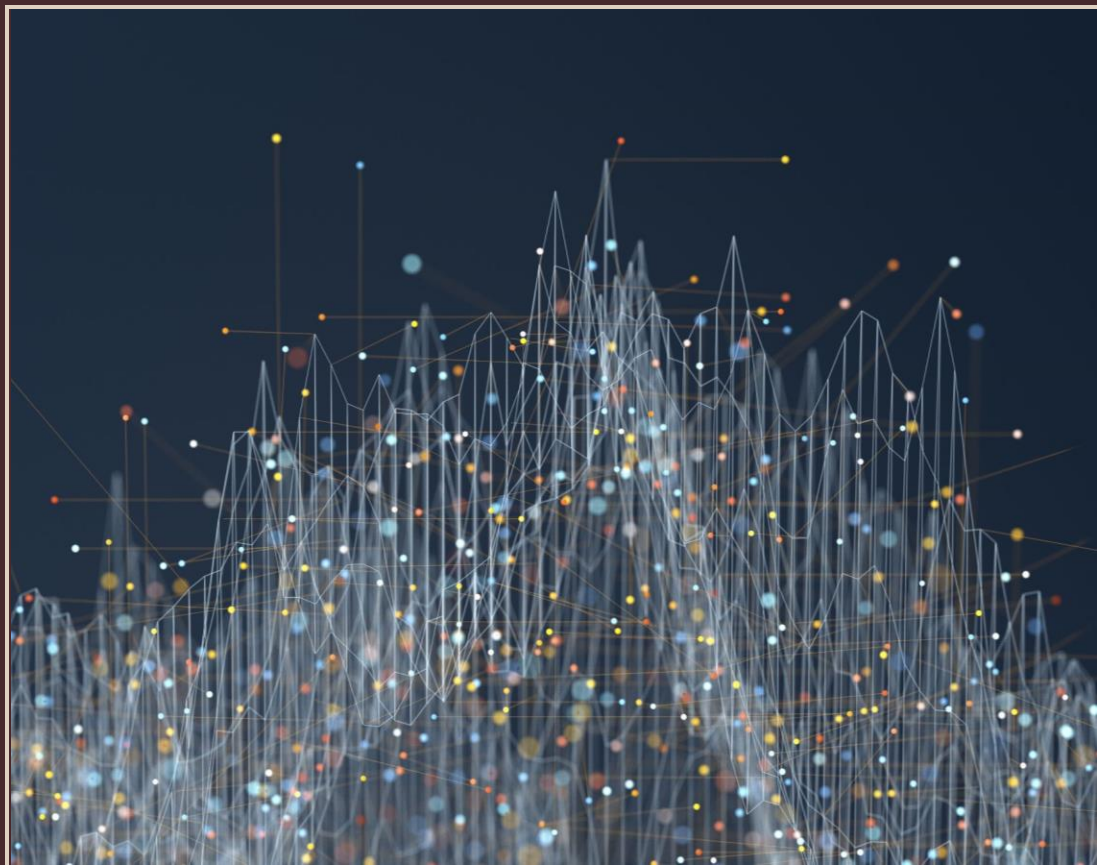


IV.2. Métricas de evaluación de modelos

Indicadores clave para medir la
precisión del modelo

Introducción

Objetivo y relevancia



Métricas para Clustering

Las métricas de evaluación permiten validar la calidad y separación coherente de los grupos formados en modelos de clustering.

Métricas para Reducción de Dimensionalidad

Estas métricas verifican que la transformación conserve la estructura y la información relevante del conjunto de datos original.

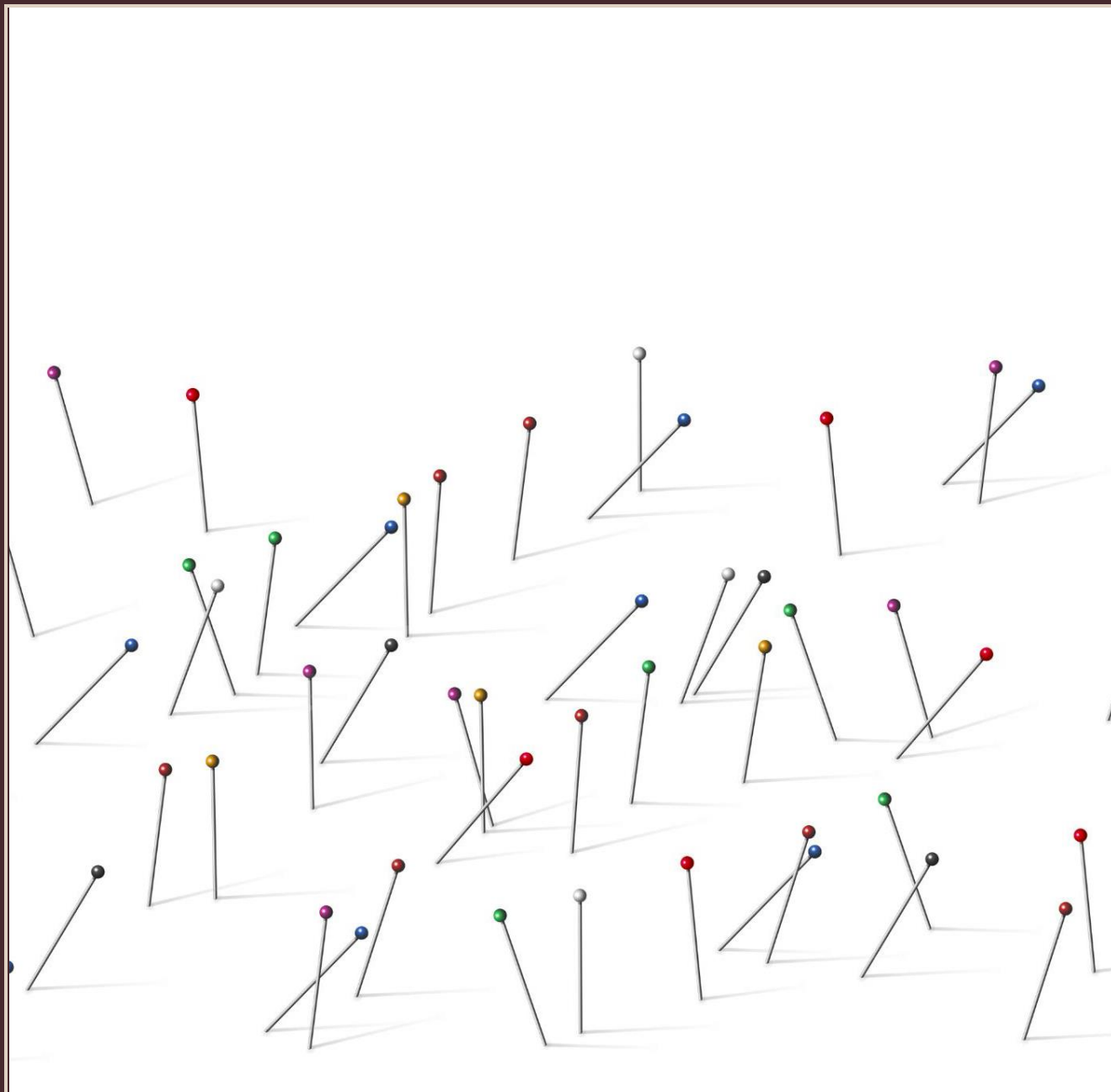
Importancia en Ciencia de Datos

Evaluar modelos es clave para garantizar interpretabilidad, optimización y toma de decisiones fundamentadas en análisis no supervisados.

Aplicación Práctica

Se aplica la evaluación usando el dataset Wine con algoritmos K-means para clustering y PCA para reducción de dimensionalidad.

Métricas de agrupación



Índice de Silueta

Definición y Fórmula

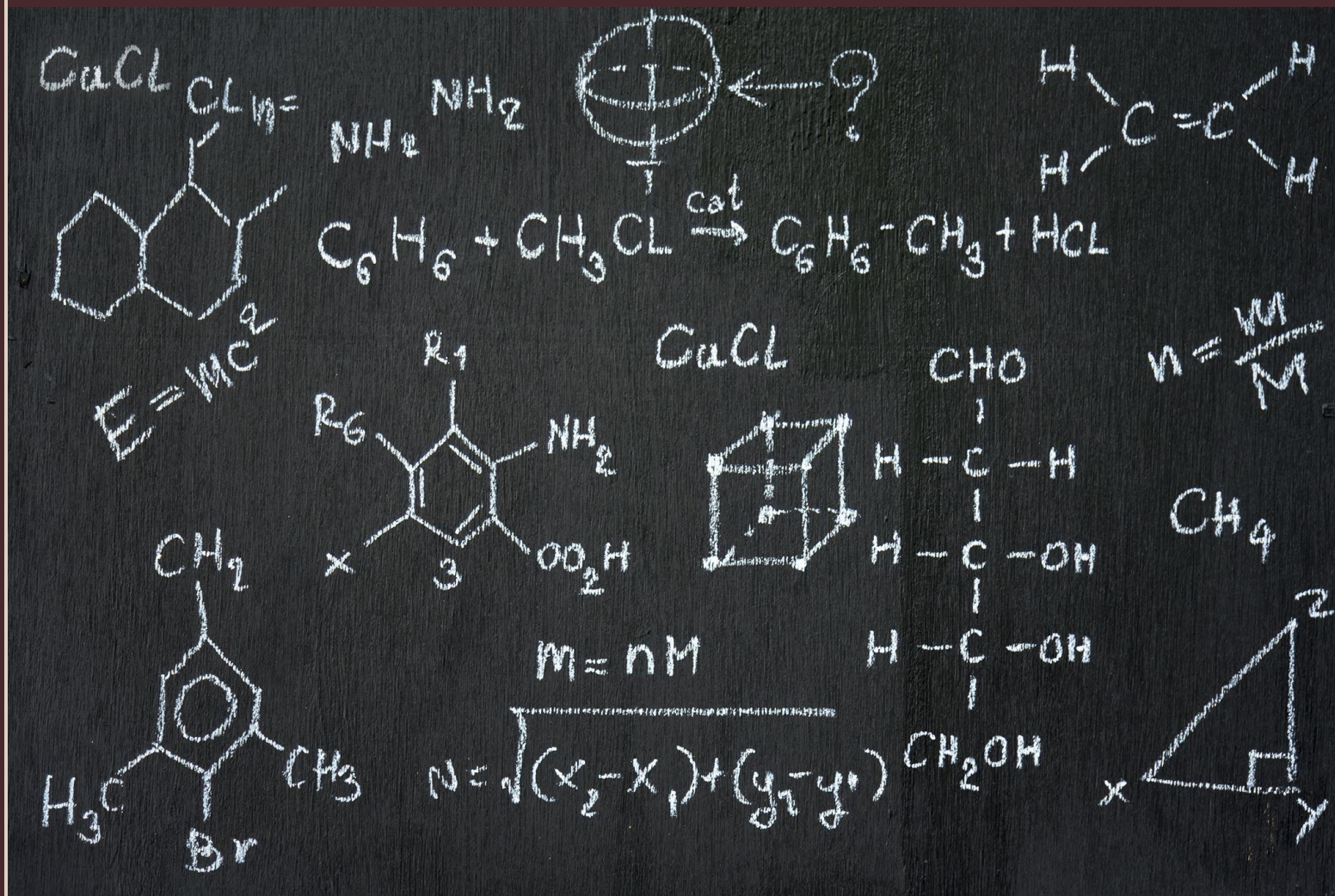
El índice de silueta evalúa la calidad de agrupamiento considerando distancias internas y externas entre puntos.

Interpretación de Valores

Valores cercanos a 1 indican buen agrupamiento y valores negativos señalan asignaciones incorrectas.

Ventajas y Limitaciones

Es fácil de interpretar, pero sensible a formas y densidades y puede ser costoso en datasets grandes.



Índice Davies–Bouldin

Evaluación de Clústeres

El índice mide la compacidad y separación de los clústeres para evaluar su calidad.

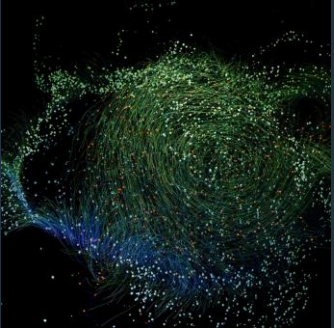
Fórmula del Índice

El índice se calcula promediando la máxima razón entre dispersión intra-clúster y distancia entre centroides.

Ventajas y Limitaciones

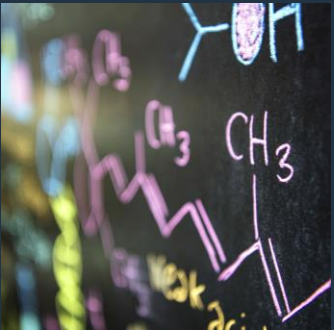
No requiere etiquetas y permite comparar configuraciones, pero es sensible a escala y outliers.

Índice Calinski–Harabasz



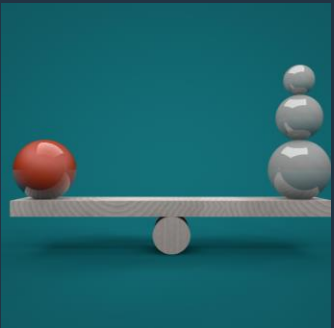
Definición del índice

Mide la relación entre dispersión entre clústeres y dispersión dentro de ellos para evaluar agrupamientos.



Fórmula matemática

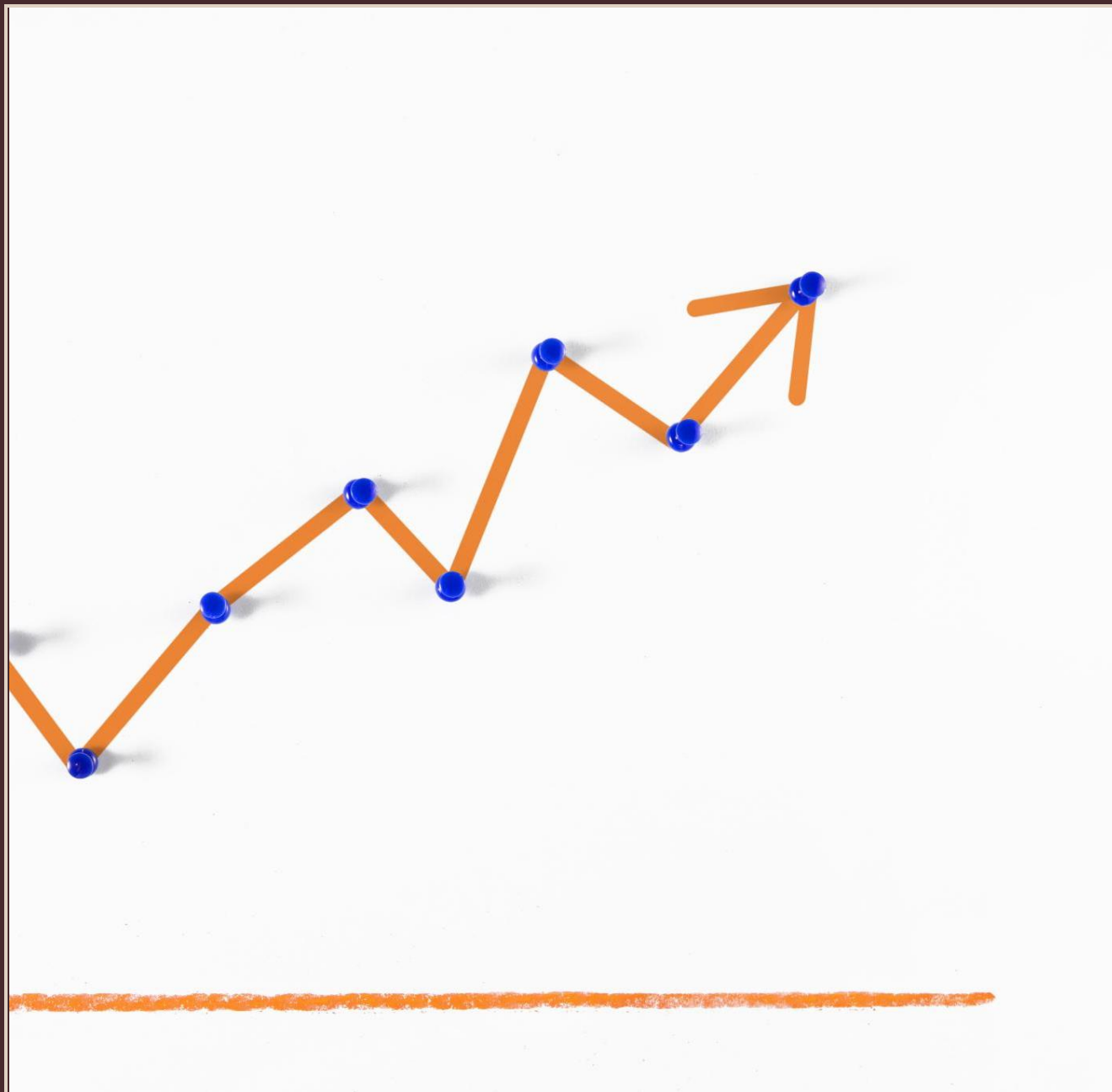
$CH = (\text{tr}(B_k)/(k-1)) / (\text{tr}(W_k)/(n-k))$, donde B_k y W_k representan matrices de dispersión.



Ventajas y limitaciones

Es eficiente y popular, pero puede favorecer más clústeres si no se controla correctamente.

Métricas de reducción de dimensionalidad



Varianza explicada acumulada

Concepto de Varianza Explicada

Mide la proporción de información retenida por componentes principales en PCA, mostrando la calidad de la reducción.

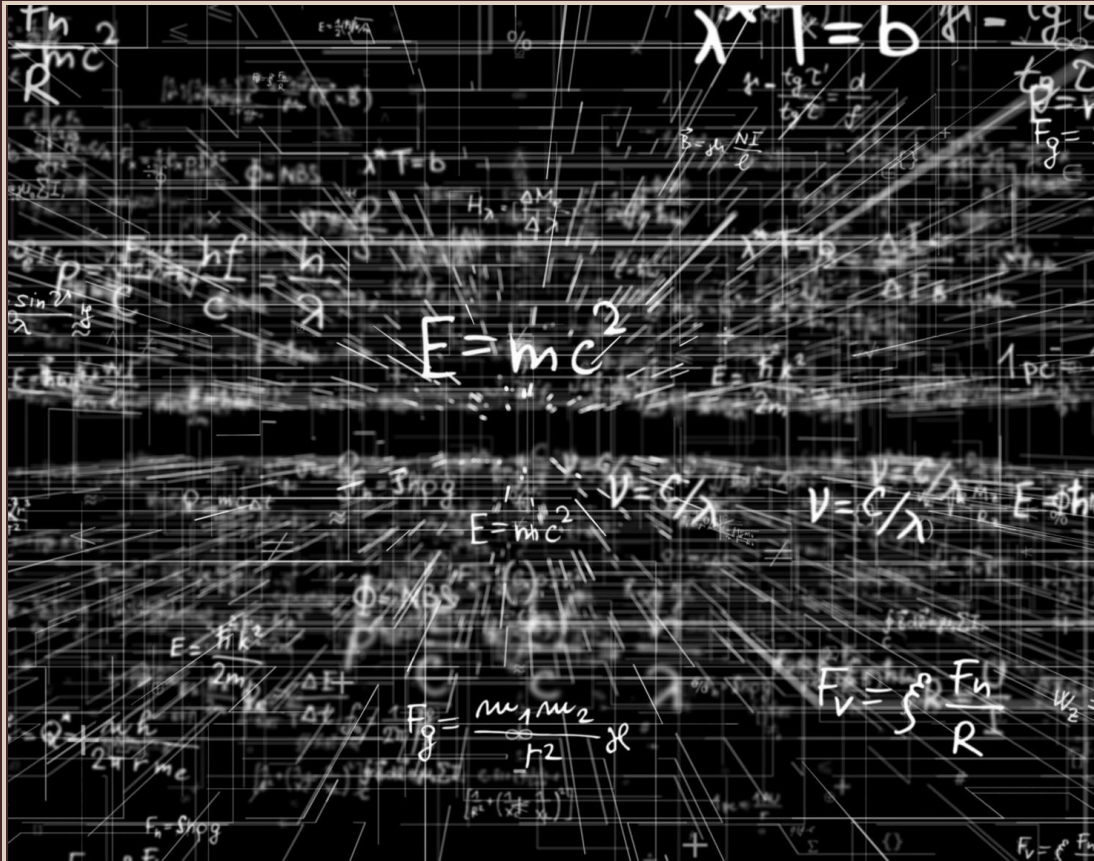
Fórmula de Varianza Acumulada

La fórmula es la suma de valores propios de componentes dividida por la suma total de valores propios.

Ventajas y Limitaciones

Fácil de interpretar y calcular, pero solo válida para métodos lineales y no siempre preserva relaciones locales.

Error de reconstrucción



Definición del error

El error de reconstrucción mide la diferencia entre datos originales y reconstruidos tras reducción y proyección inversa.

Interpretación de valores bajos

Un error bajo indica que la transformación conserva bien la estructura original de los datos.

Ventajas y aplicaciones

El error es útil en métodos lineales y no lineales, como autoencoders, para evaluar calidad de reconstrucción.

Limitaciones del método

Requiere reconstrucción explícita y puede ser costoso computacionalmente en grandes conjuntos de datos.

Caso práctico: Dataset y resultados



Descripción del dataset

Dataset Wine

El dataset contiene 178 muestras y 13 atributos numéricos sobre características químicas del vino.

Análisis multivariado

Ideal para clustering y PCA por su naturaleza multivariada y sin etiquetas para análisis no supervisado.

Aplicación de K-means y PCA

Se aplicó K-means con $k=3$ y PCA para reducción a dos dimensiones y evaluación de métricas.

Resultados de clustering

MÉTRICA	VALOR
Silueta	0.285
Davies–Bouldin	1.480
Calinski–Harabasz	561.63

Resultados de reducción

COMPONENTE	VARIANZA EXPLICADA
PC1	36.98%
PC2	19.20%

Comparativa y
conclusiones

Análisis comparativo



Métricas de Clustering

Silüeta y Calinski–Harabasz miden la compacidad y separación de clústeres para evaluar calidad en agrupación.

Métrica Davies–Bouldin

Davies–Bouldin penaliza la falta de separación entre grupos, complementando la evaluación de clústeres.

Varianza Explicada en PCA

Varianza explicada indica cuánta información conserva PCA tras reducción, clave para visualización y análisis.

Roles de las Métricas

Métricas validan segmentación o fidelidad de representación reducida según su función específica.

Conclusiones y recomendaciones



Importancia de métricas combinadas

Combinar métricas de agrupación y reducción garantiza resultados más confiables en análisis no supervisados.

Métricas recomendadas para clustering

Silueta y Calinski–Harabasz son preferidas por su facilidad de interpretación en el análisis de clústeres.

Métricas para reducción de dimensionalidad

Varianza explicada es útil para PCA, mientras que error de reconstrucción es mejor para métodos no lineales.

Ventajas prácticas de reducción previa

Reducir dimensionalidad antes del clustering mejora la calidad de los clústeres y reduce el costo computacional.