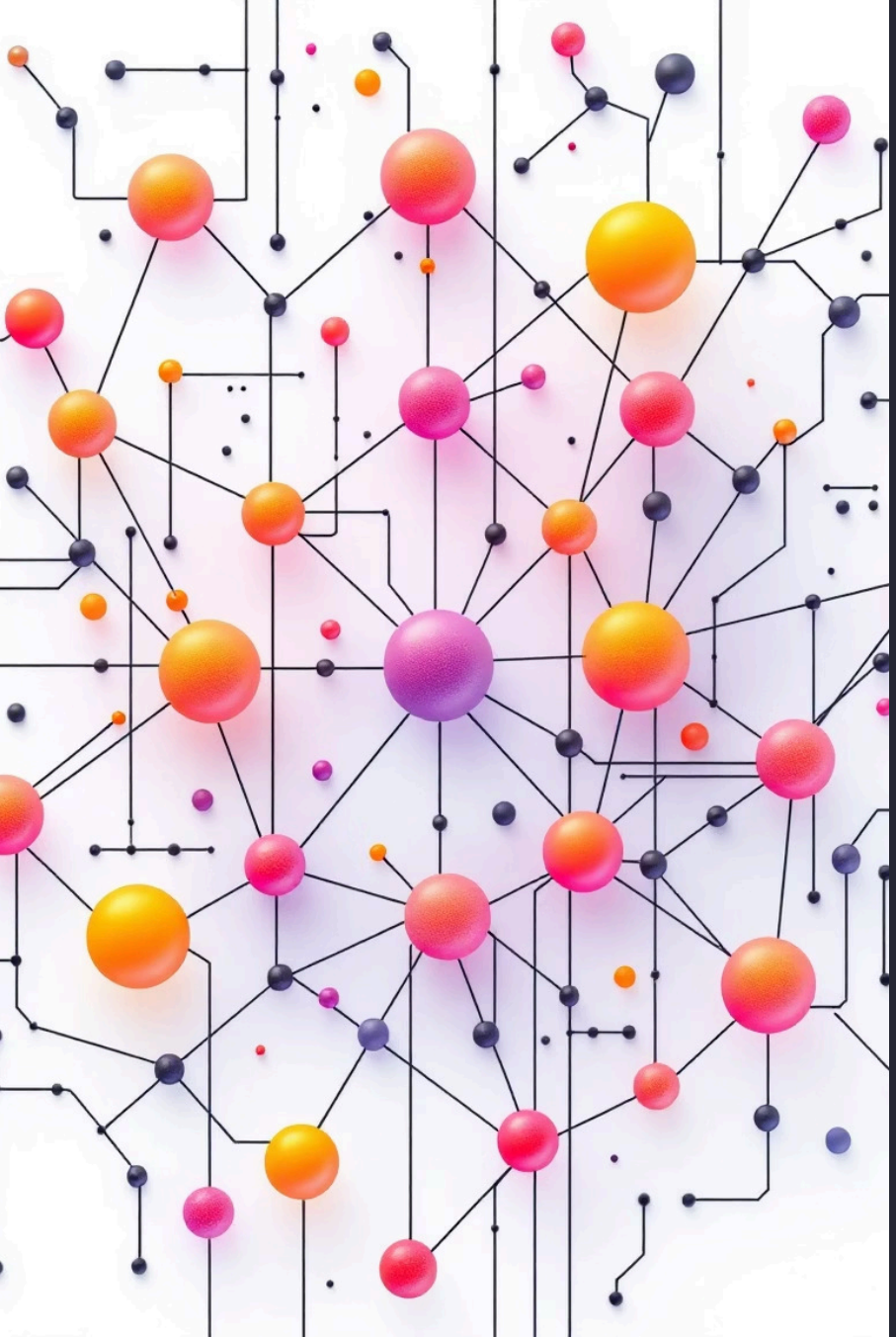


Métricas de Evaluación de Modelos

Extracción de Conocimiento en Bases de Datos

Presentan: Daron Tarín González, Ángel Ricardo Chávez Zaragoza, Mildred Villaseñor Ruiz

Docente: Ing. Luis Enrique Mascote Cano | **Grupo:** IDGS91N

Universidad Tecnológica de Chihuahua | Desarrollo y Gestión de Software

Objetivos e Introducción

Objetivo

Identificar y aplicar métricas de evaluación para modelos de agrupación y reducción de dimensionalidad mediante un caso de estudio con dataset real.

Contexto

En modelos no supervisados no existen etiquetas verdaderas. Las métricas específicas permiten medir cohesión, separación, estabilidad y fidelidad de las representaciones generadas.

Métricas de Agrupación

Índice de Silueta

Mide separación entre clústeres y cohesión interna. Valores cercanos a 1 indican excelente agrupamiento.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Calinski-Harabasz

Mide relación entre dispersión entre clústeres y dentro de clústeres. Valores altos indican mejor separación.

$$CH = \frac{\text{Entre clusters}/(k-1)}{\text{Dentro clusters}/(n-k)}$$

Davies-Bouldin

Evalúa dispersión interna vs separación entre clústeres. Valores bajos (<1) indican clústeres bien definidos.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

Métricas de Reducción de Dimensionalidad

Varianza Explicada (PCA)

Indica cuánta información del conjunto original conservan los componentes principales.

$$\text{Varianza acumulada} = \sum_{i=1}^{k} \lambda_i$$

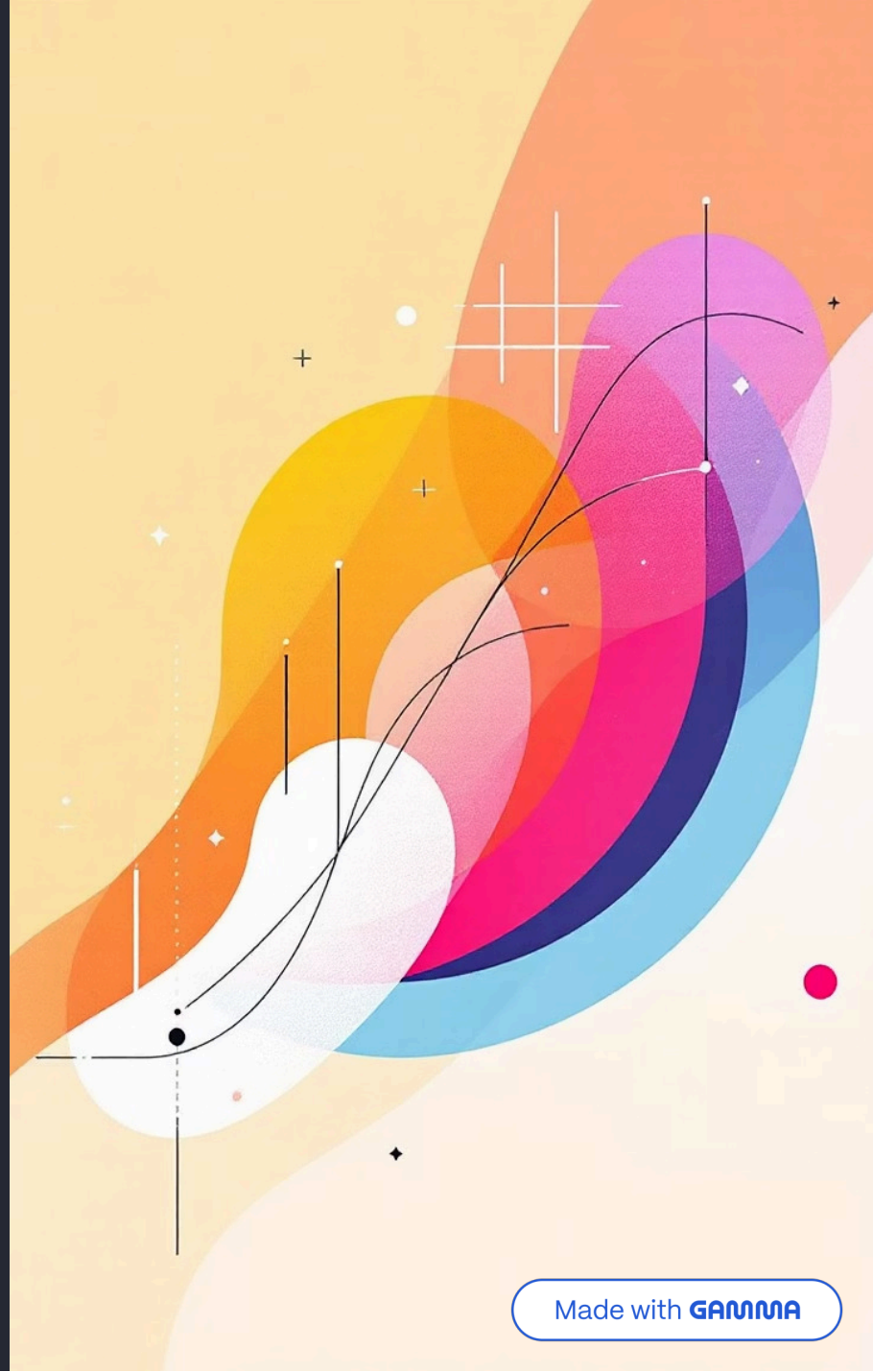
Interpretación: Alta (>80%) = buena preservación de información.

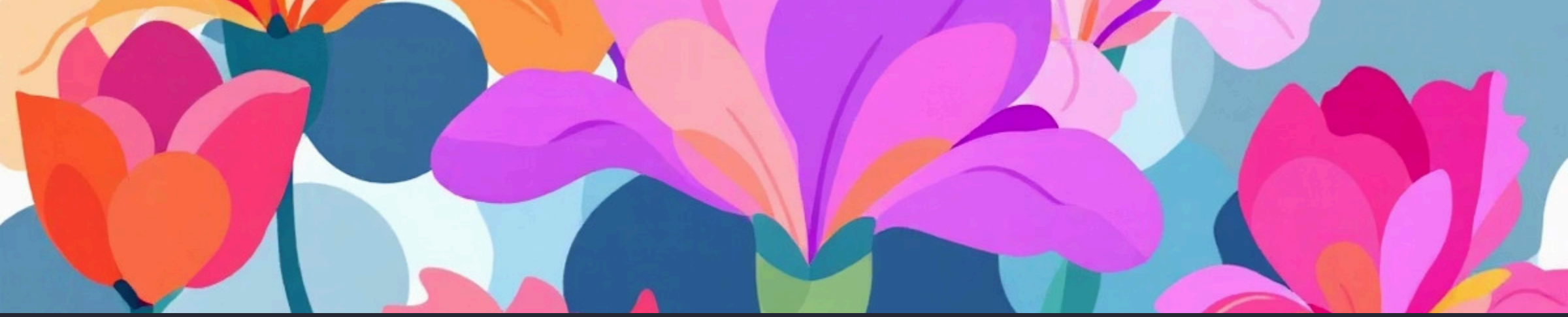
Error de Reconstrucción

Mide diferencia entre representación original y reconstruida tras reducir dimensionalidad.

$$E = |X - \hat{X}|^2$$

Interpretación: Error bajo = buena preservación de estructura.





Dataset Seleccionado: Iris

Composición

150 muestras con 4 atributos numéricos: largo y ancho del sépalo, largo y ancho del pétalo.

Parámetros K-means

k=3 porque el dataset contiene tres especies conocidas, permitiendo verificar patrones sin etiquetas.

Parámetros PCA

Dos componentes capturan ~95% de varianza, garantizando representación fiel y visualización 2D.

Resultados de Clustering

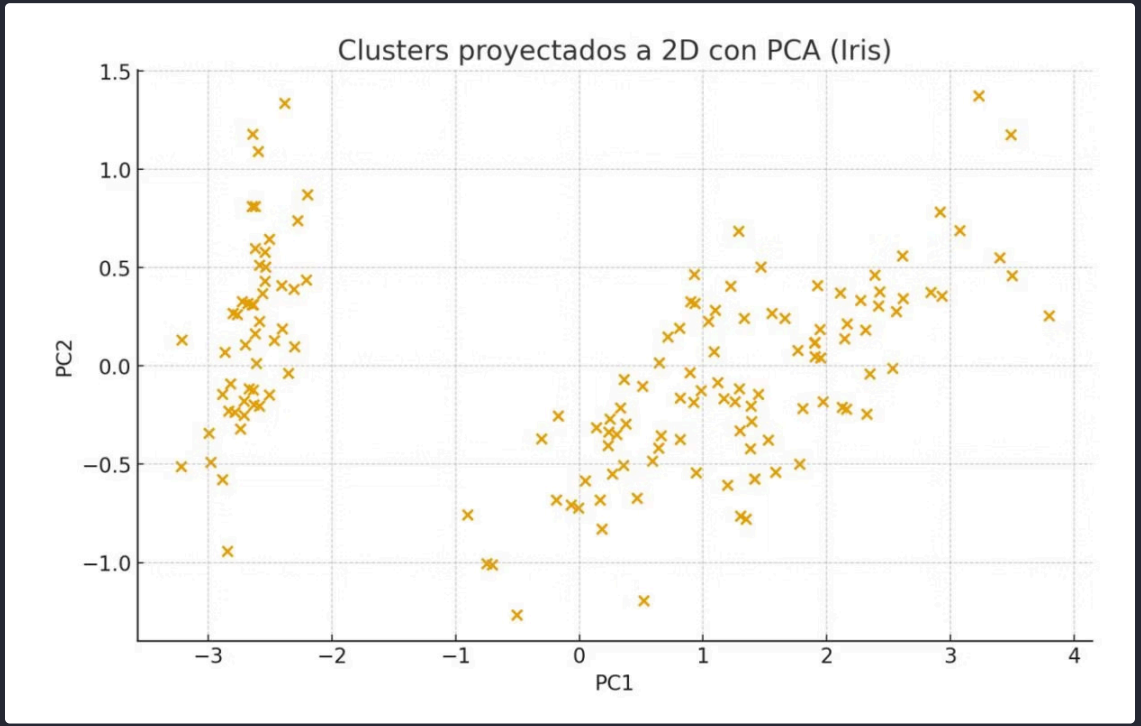


Figura 1: Proyección PCA mostrando tres agrupaciones. Setosa claramente separada, Versicolor y Virginica con superposición.

Métricas Obtenidas

Métrica	Valor
Silueta	0.54
Davies-Bouldin	0.68
Calinski-Harabasz	420.1

La silueta indica separación moderada. Davies-Bouldin <1 confirma buena definición. Calinski-Harabasz alto indica excelente agrupamiento.

Resultados de Reducción (PCA)

Varianza Explicada

Componente	Varianza
PC1	72.7%
PC2	23.0%
PC3	3.6%
PC4	0.7%
PC1+PC2	**95.7%**

Error de reconstrucción: E = 0.041

PCA conserva casi toda la información original.

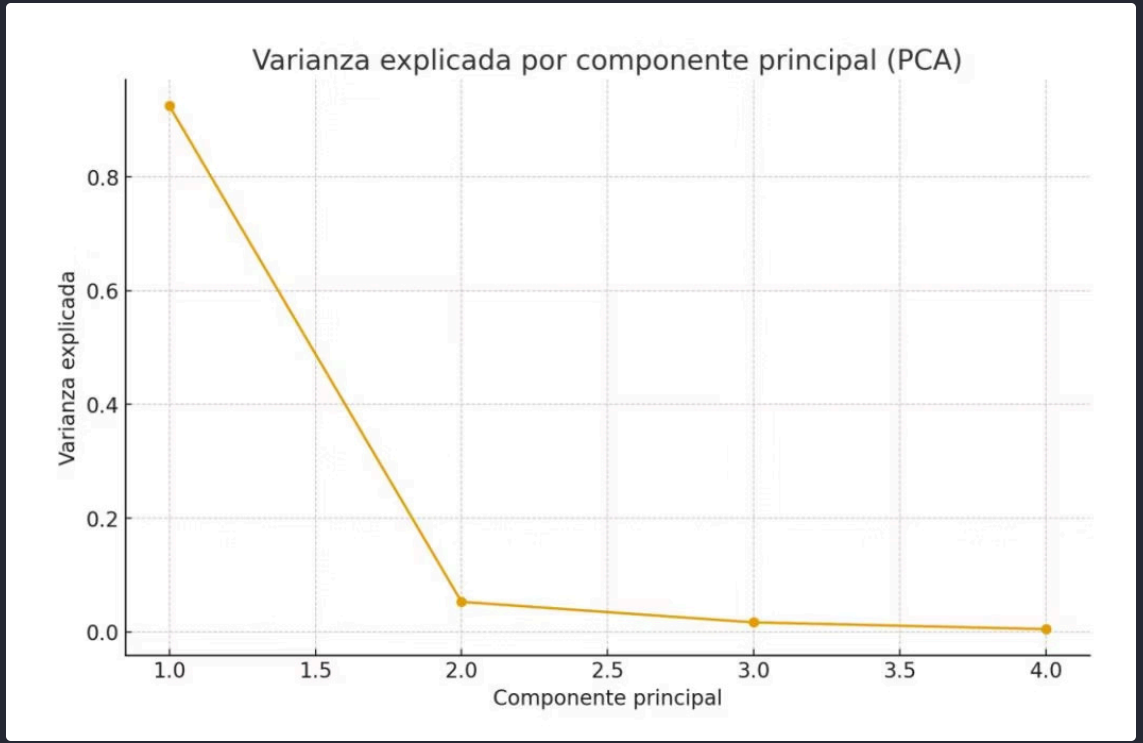


Figura 2: Los dos primeros componentes capturan más del 95% de la varianza total.

Comparativa y Análisis

Métricas de Clustering

Silueta analiza cohesión interna. Davies-Bouldin penaliza dispersión. Calinski-Harabasz mide relaciones de varianza.

Métricas de Reducción

Varianza explicada muestra que PCA captura >95% de información en dos componentes. Error de reconstrucción confirma pérdida mínima.

Validación Integral

K-means genera agrupamiento razonable. PCA permite visualización adecuada sin comprometer estructura de datos.

Conclusiones y Recomendaciones

Conclusiones

- Las métricas permiten evaluación formal y cuantitativa de modelos no supervisados
- K-means logró estructura coherente con buena cohesión y separación razonable
- PCA conservó >95% de varianza en dos componentes, manteniendo estructura esencial
- Validación mediante métricas garantiza análisis precisos y decisiones fundamentadas

Recomendaciones

- Verificar múltiples métricas antes de concluir sobre calidad del modelo
- Experimentar con diferentes valores de k o parámetros
- Utilizar PCA para visualizar y mejorar rendimiento
- Complementar métricas con interpretación visual
- Evaluar con datasets más grandes para validar estabilidad

Referencias

Scikit-learn. (2024). *Clustering Metrics*. **https://scikit-learn.org/stable/modules/clustering.html**

Fikiri.net. (2024). *Tutorial sobre múltiples métricas de evaluación de clústeres*. **https://fikiri.net/es/tutorial-sobre-multiples-metricas-de-evaluacion-de-clusteres/**

Analytics Lane. (2023). *Índice de Davies-Bouldin para K-means*. **https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/**

Analytics Lane. (2023). *Identificar el número de clústeres con Calinski-Harabasz*. https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/