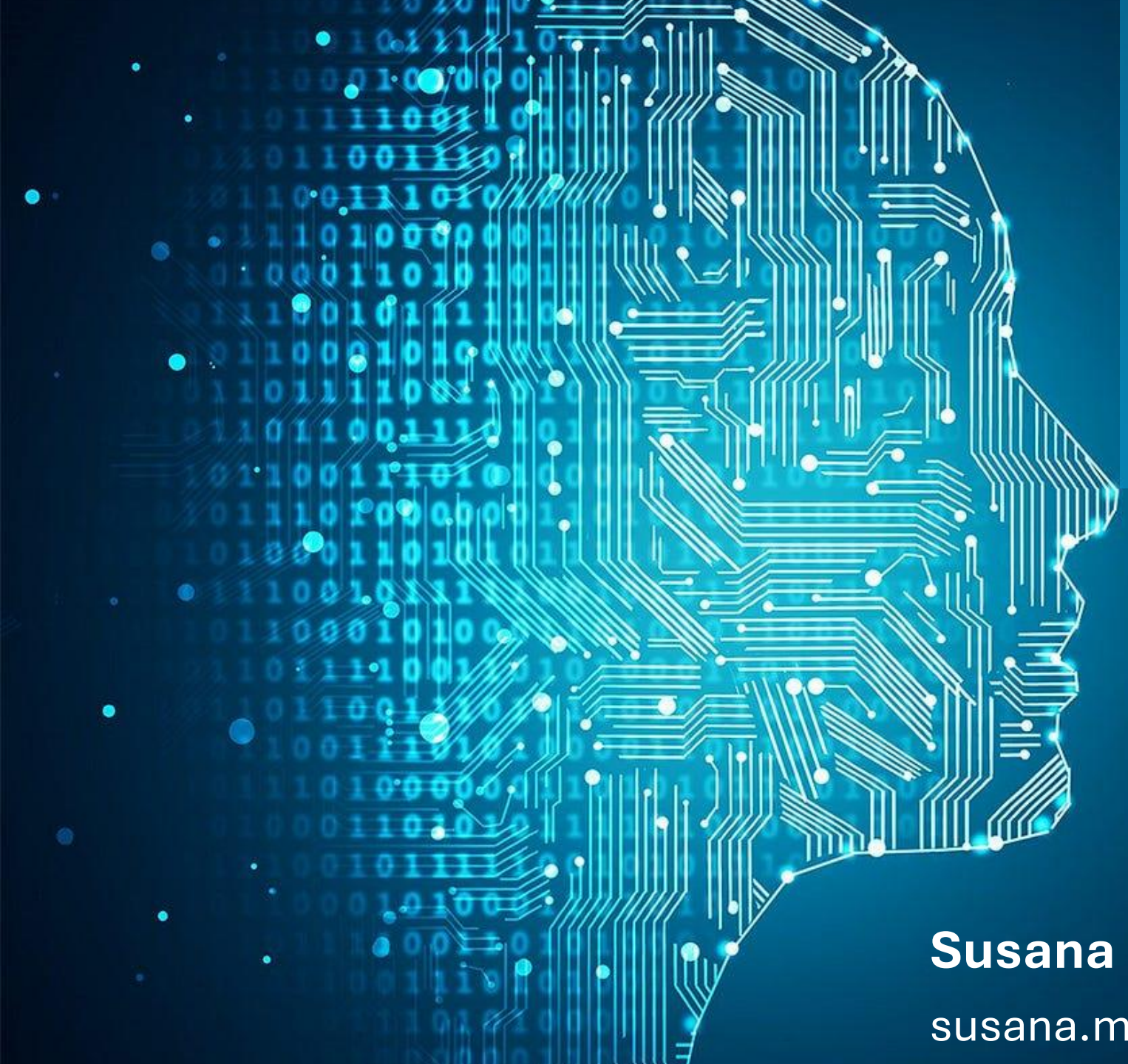
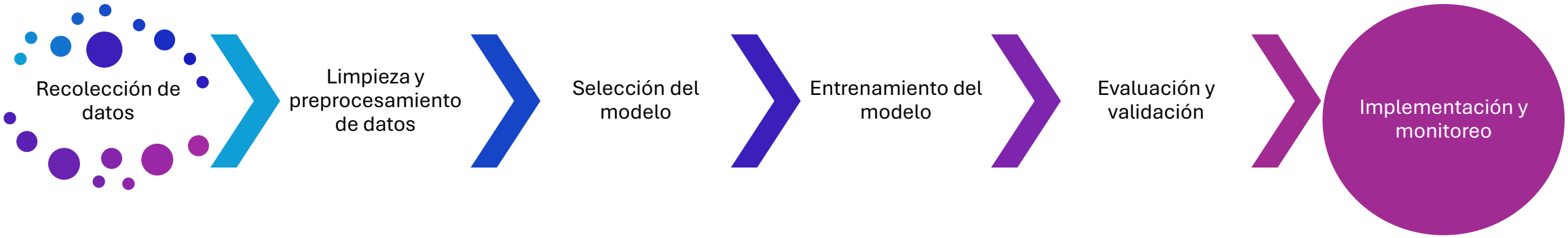


Machine Learning

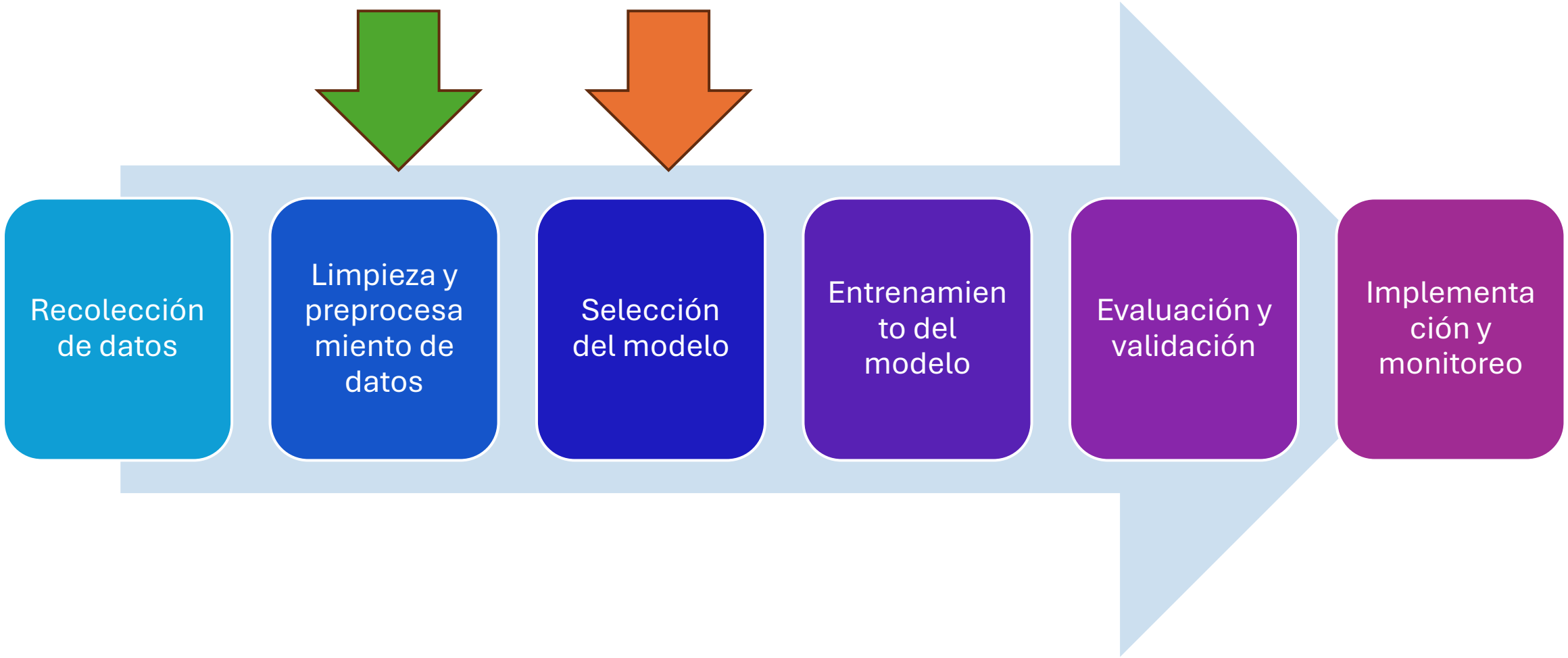
Susana Medina Gordillo

susana.medina@correounivalle.edu.co

Flujo de trabajo en Machine Learning



Flujo de trabajo en Machine Learning



Aprendizaje No Supervisado

Introducción al Aprendizaje No Supervisado: Descubriendo Patrones Ocultos

Introducción, Clustering y Métodos Avanzados

Aprendizaje No Supervisado: definición

Definición: Aprendizaje automático (ML) donde el algoritmo aprende patrones directamente de los datos sin etiquetas o sin variables objetivo predefinidas.

El objetivo es descubrir la **estructura inherente**, las **relaciones** y las **agrupaciones** (clustering) dentro de los datos.

Imagina **explorar** un nuevo conjunto de objetos sin ninguna descripción previa. Debes encontrar similitudes y agruparlos basándote en sus propias características.

¿Qué es el Aprendizaje No Supervisado?

El **Aprendizaje No Supervisado** es una rama del aprendizaje automático (Machine Learning) que se enfoca en **analizar y descubrir patrones ocultos en conjuntos de datos que no están etiquetados**.

Un algoritmo de aprendizaje no supervisado recibe datos de entrada sin ninguna guía predefinida sobre la salida correcta.

Su tarea es **encontrar estructuras inherentes, relaciones, similitudes y agrupaciones** dentro de esos datos por sí solo.

¿Qué es el Aprendizaje No Supervisado?

El **Aprendizaje No Supervisado** es una rama del aprendizaje automático (Machine Learning) que se enfoca en **analizar y descubrir patrones ocultos en conjuntos de datos que no están etiquetados**.

Un algoritmo de aprendizaje no supervisado recibe datos de entrada sin ninguna guía predefinida sobre la salida correcta.

Su tarea es **encontrar estructuras inherentes, relaciones, similitudes y agrupaciones** dentro de esos datos por sí solo.

Puntos clave sobre el Aprendizaje No Supervisado

Datos sin etiquetas

La característica distintiva es la ausencia de una variable objetivo o etiquetas preasignadas que indiquen la "respuesta correcta" para cada dato.

Descubrimiento de patrones

El objetivo principal es identificar patrones que no son evidentes a simple vista. Esto puede incluir la formación de grupos (clustering), la reducción de la dimensionalidad de los datos, o la identificación de asociaciones entre variables.

Exploratorio

A menudo se utiliza como una técnica exploratoria para obtener información sobre la estructura subyacente de los datos antes de aplicar otras técnicas de aprendizaje automático.

Flexibilidad

Los algoritmos no supervisados son más flexibles ya que no están restringidos por etiquetas predefinidas, lo que les permite descubrir patrones inesperados.

Evaluación desafiante

Evaluar el rendimiento de los algoritmos no supervisados puede ser más difícil que en el aprendizaje supervisado, ya que no hay una "verdad fundamental" directa con la cual comparar las salidas. Se utilizan métricas basadas en la estructura de los datos encontrados.

Casos de Uso del Aprendizaje No Supervisado

- Segmentación de Clientes (Marketing)
- Detección de Anomalías (Seguridad/Mantenimiento)
- Análisis de Componentes Principales (Reducción de Dimensionalidad)
- Agrupación de Documentos (Procesamiento de Lenguaje Natural)
- Sistemas de Recomendación

Casos de Uso del Aprendizaje No Supervisado: segmentación

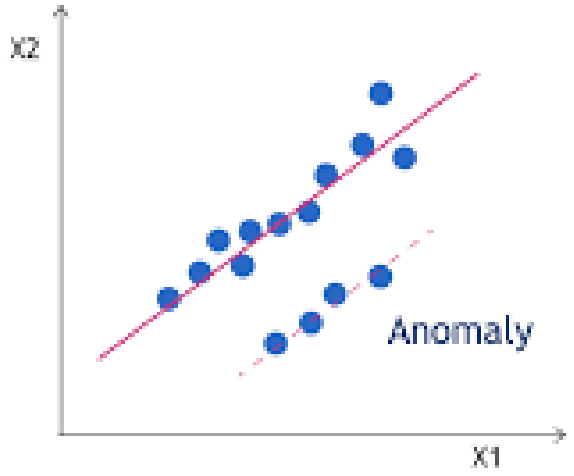
Segmentación de Clientes (Marketing)

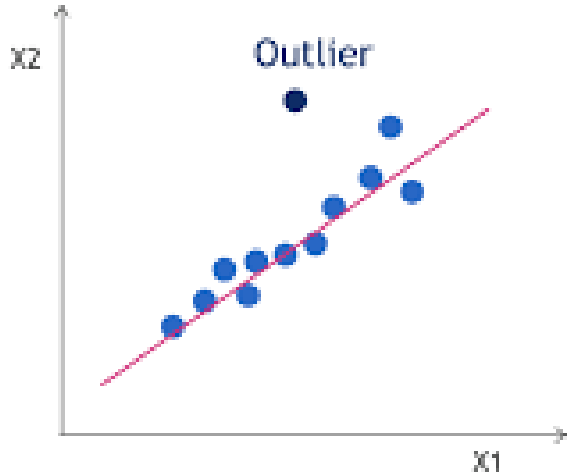
Agrupar clientes con comportamientos de compra similares para campañas personalizadas. (Icono de grupos de personas/gráficos de compra)

Casos de Uso del Aprendizaje No Supervisado: anomalías

Detección de Anomalías (Seguridad/Mantenimiento)

Identificar puntos de datos inusuales que podrían indicar fraude, errores o fallas en equipos

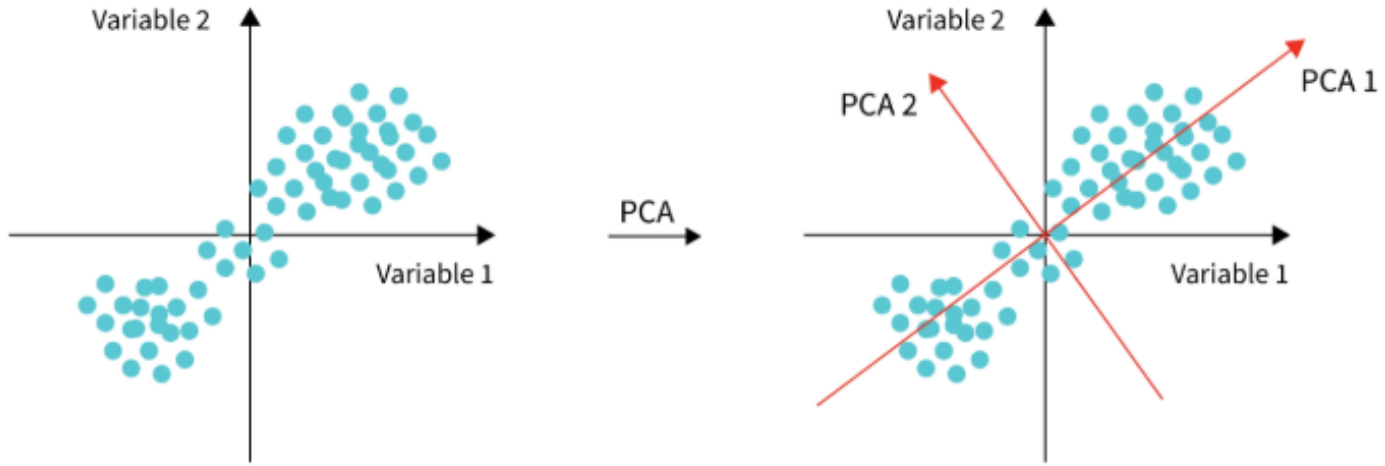




Casos de Uso del Aprendizaje No Supervisado: dimensionalidad

Análisis de Componentes Principales (Reducción de Dimensionalidad)

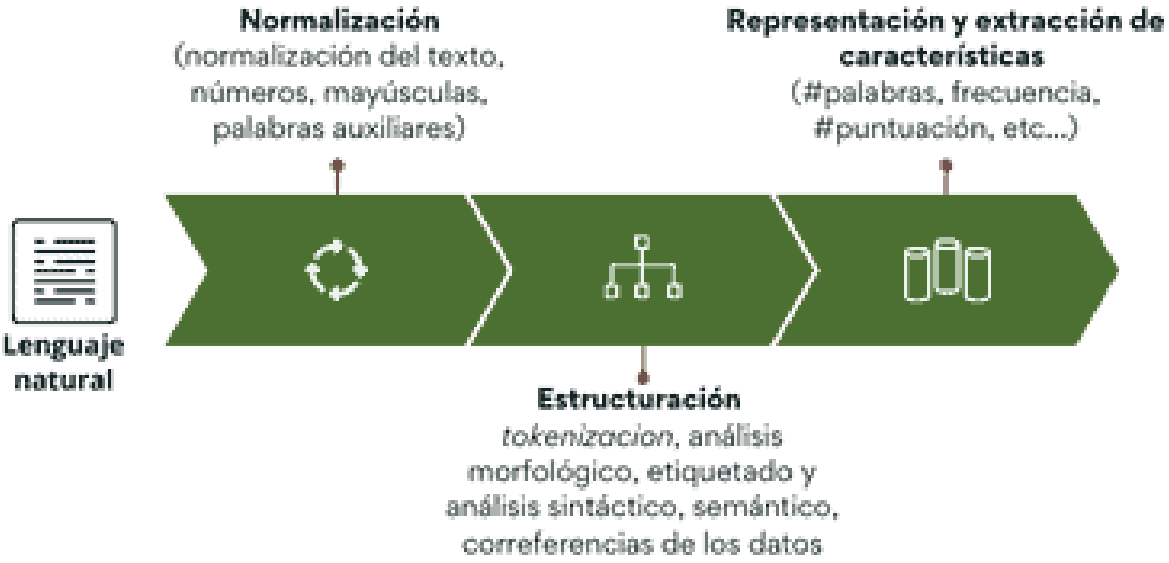
Simplificar datos complejos identificando las variables más importantes.



Casos de Uso del Aprendizaje No Supervisado: PLN

Agrupación de Documentos (Procesamiento de Lenguaje Natural)

Organizar grandes cantidades de texto en temas o categorías

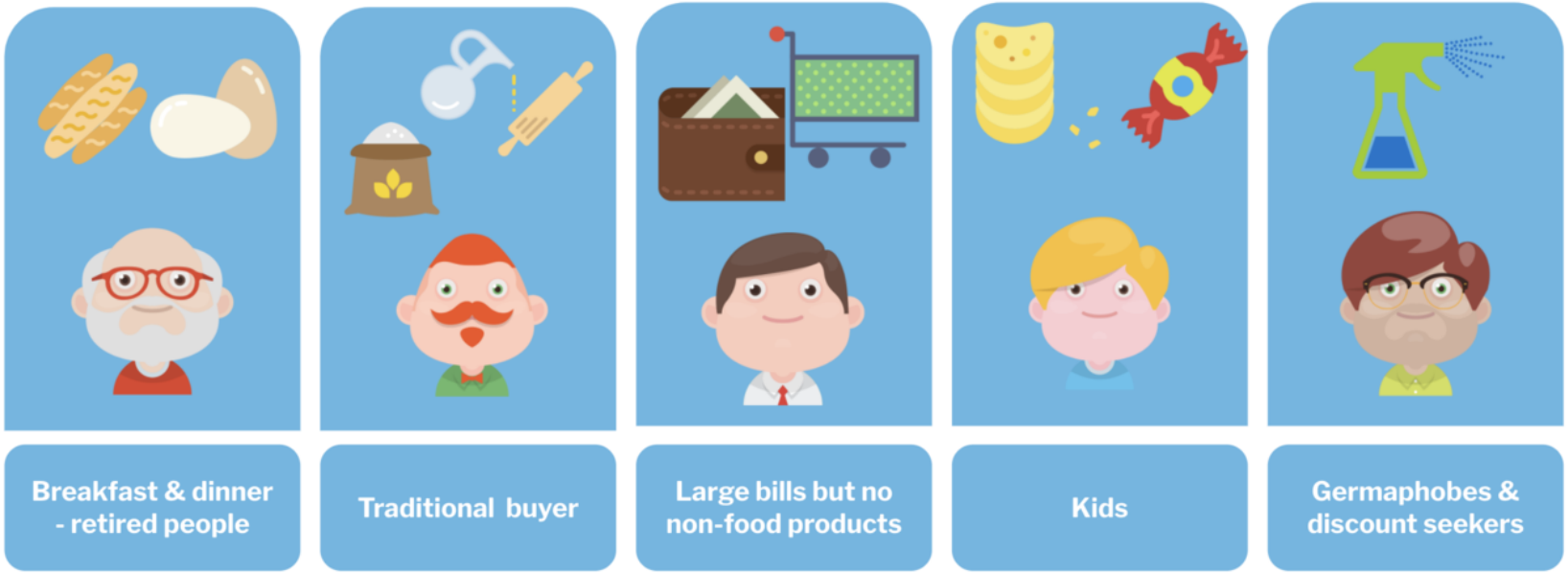


Casos de Uso del Aprendizaje No Supervisado: PLN

Casos de Uso del Aprendizaje No Supervisado: Sistemas de recomendación

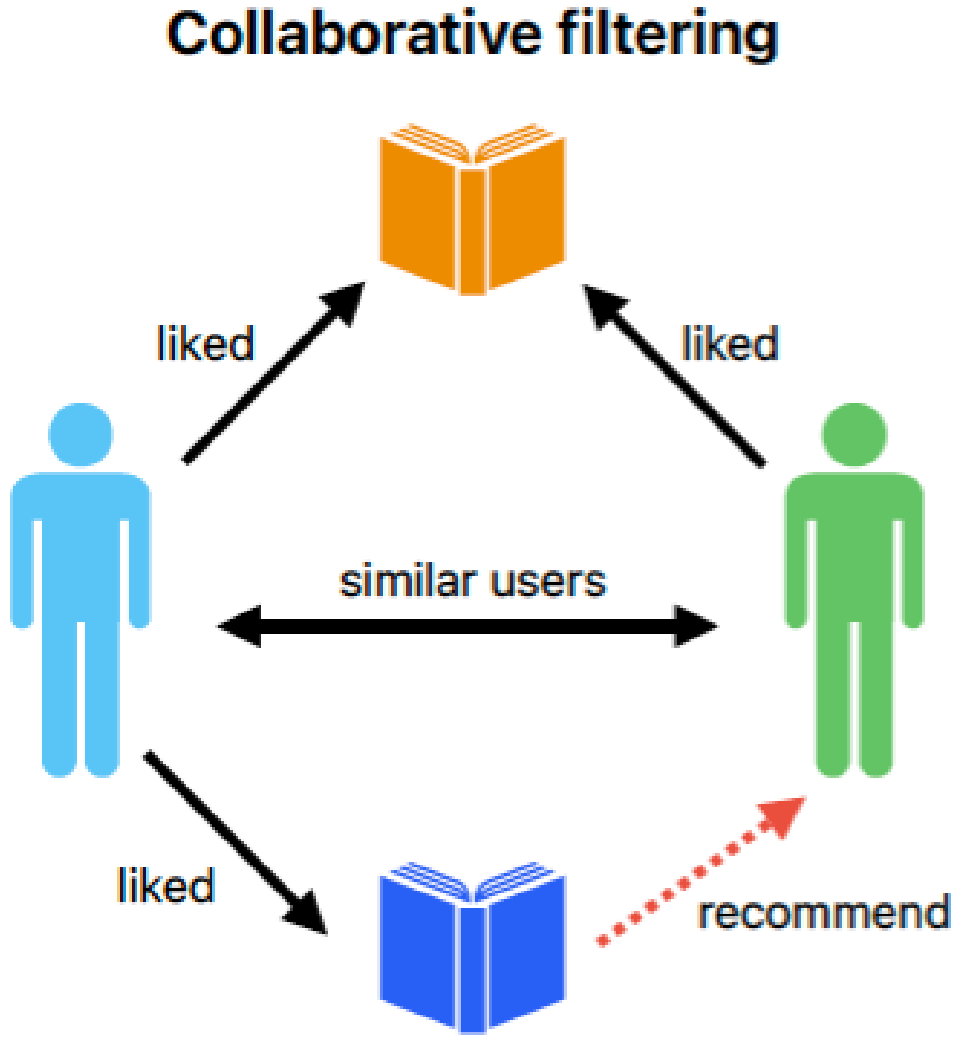
Sistemas de Recomendación

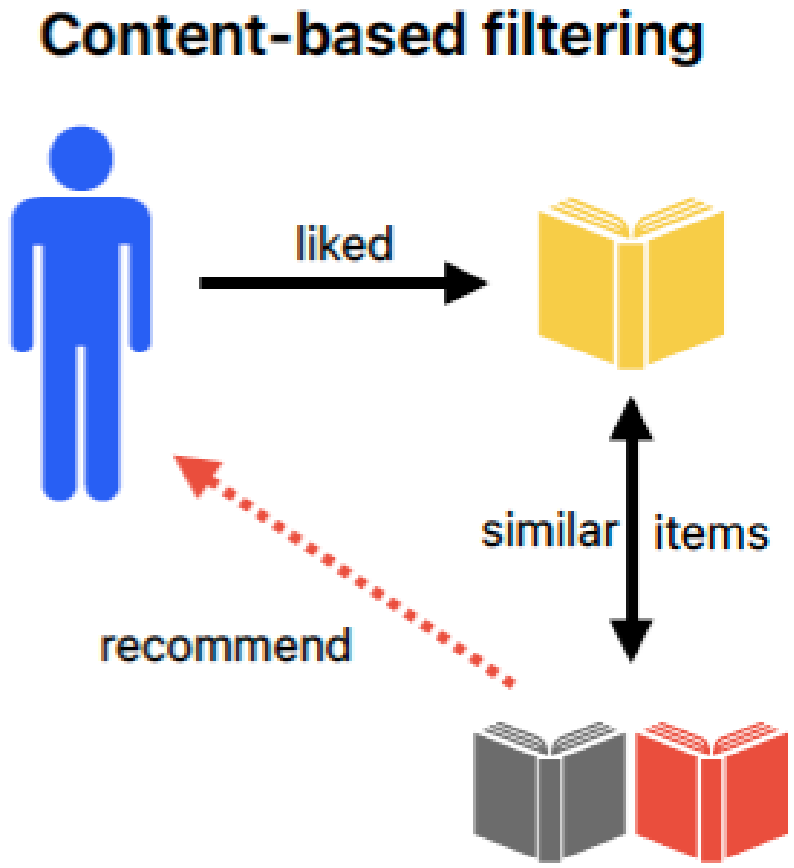
Descubrir patrones en las preferencias de los usuarios para sugerir nuevos productos o contenidos



amazon.com

Casos de Uso del Aprendizaje No Supervisado: Sistemas de recomendación





Comparación: Aprendizaje Supervisado vs. No Supervisado

Característica	Aprendizaje Supervisado	Aprendizaje No Supervisado
Datos de Entrenamiento	Datos etiquetados (entrada y salida deseada)	Datos sin etiquetas (solo entrada)
Objetivo Principal	Predecir o clasificar una salida	Descubrir estructura o patrones ocultos
Tipo de Problemas	Clasificación, Regresión	Clustering, Reducción de Dimensionalidad, Asociación
Evaluación	Métricas basadas en la verdad fundamental (precisión, error, etc.)	Métricas basadas en la estructura de los datos (inercia, silueta, etc.)
Ejemplos de Algoritmos	Regresión Lineal, Regresión Logística, Árboles de Decisión, Redes Neuronales (para clasificación/regresión)	K-Means, Clustering Jerárquico, DBSCAN, PCA, Análisis de Reglas de Asociación

Tarea #1: Preparación de datos

developers.google.com

Algoritmos de Clustering

Leer la sección y realizar el quiz al final (Verifica tu comprensión)

45 min

Agrupación en clústeres

El agrupamiento en clústeres es una estrategia clave de aprendizaje automático no supervisado para asociar elementos relacionados.

Clustering: Clustering: Agrupando Datos Similares

Clustering con K-Means: La Teoría

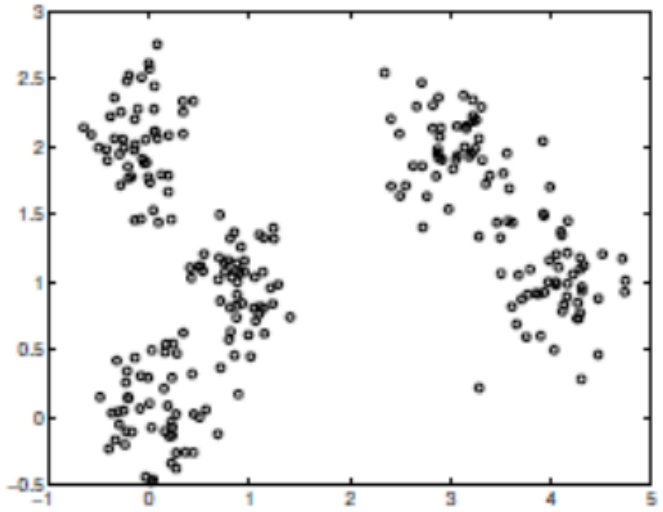
Paso 1. Algoritmo iterativo que busca particionar los datos en *k* clusters, donde *k* es un número predefinido por el usuario.

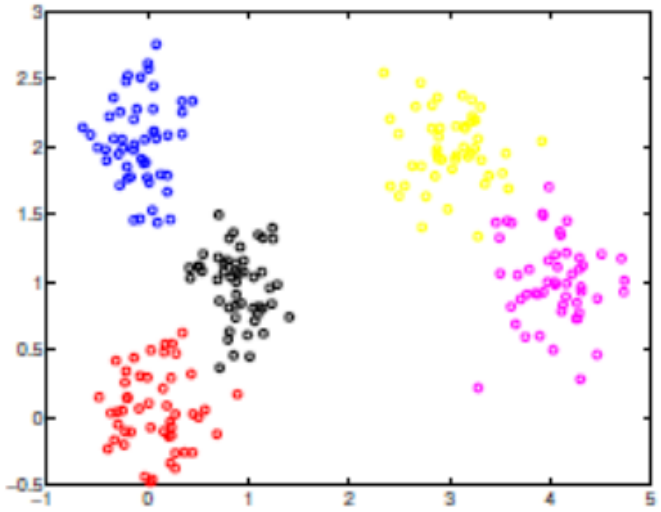
Paso 2. Pasos principales:

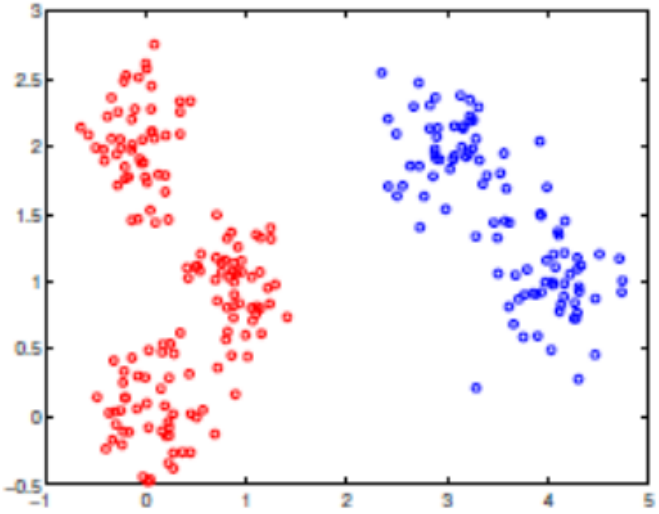
- **Inicialización**: Seleccionar aleatoriamente *k centroides iniciales*.
- **Asignación**: Asignar cada punto de datos al *cluster* cuyo centroide esté más cercano (usualmente usando distancia euclidiana).
- **Actualización**: Recalcular la posición de cada centroide como la media de todos los puntos asignados a ese *cluster*.
- **Iteración**: Repetir los pasos de asignación y actualización hasta que los **centroides converjan** (cambien muy poco) o se alcance un número **máximo de iteraciones**.

Paso 3. Objetivo: Minimizar la **inercia** (Within-Cluster Sum of Squares - **WCSS**), que es la suma de las distancias cuadradas de cada punto a su centroide asignado.

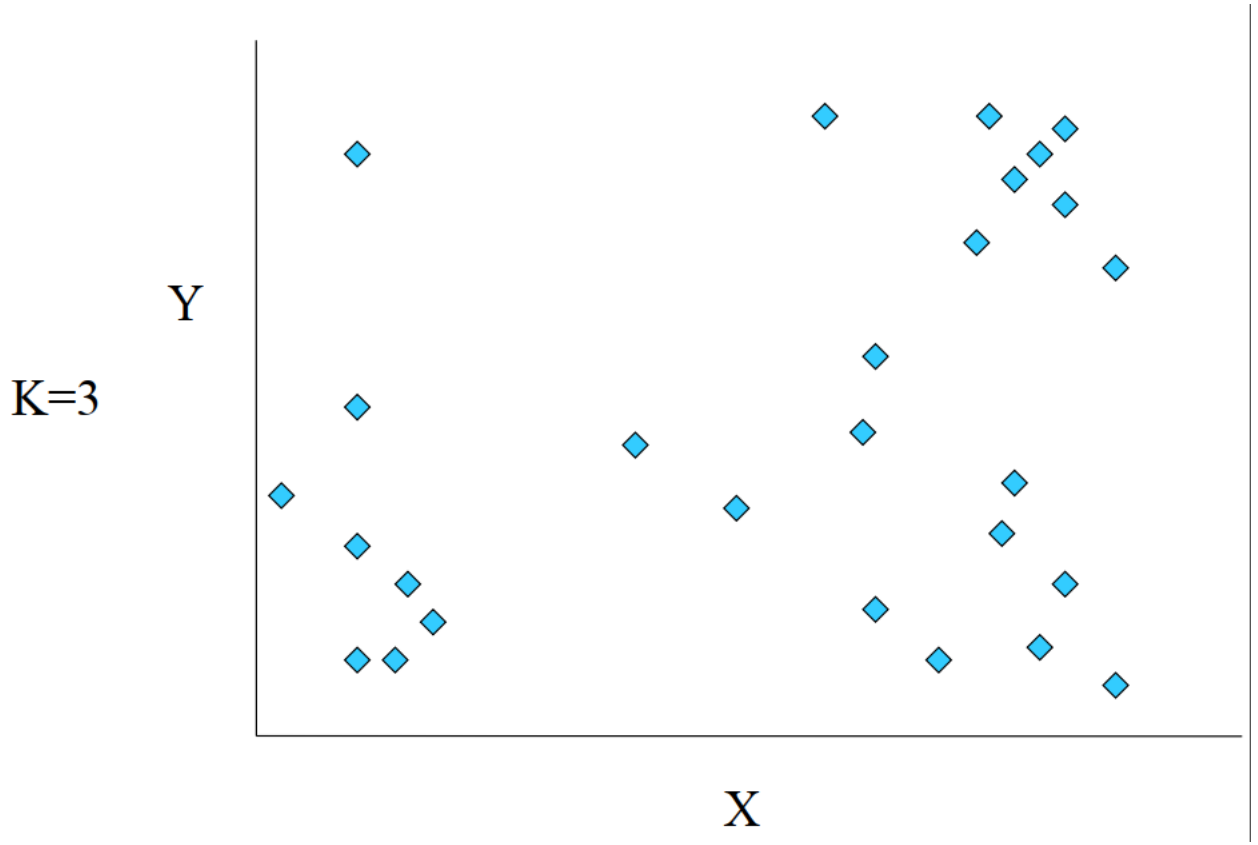
Clustering con K-Means: número k de clusters



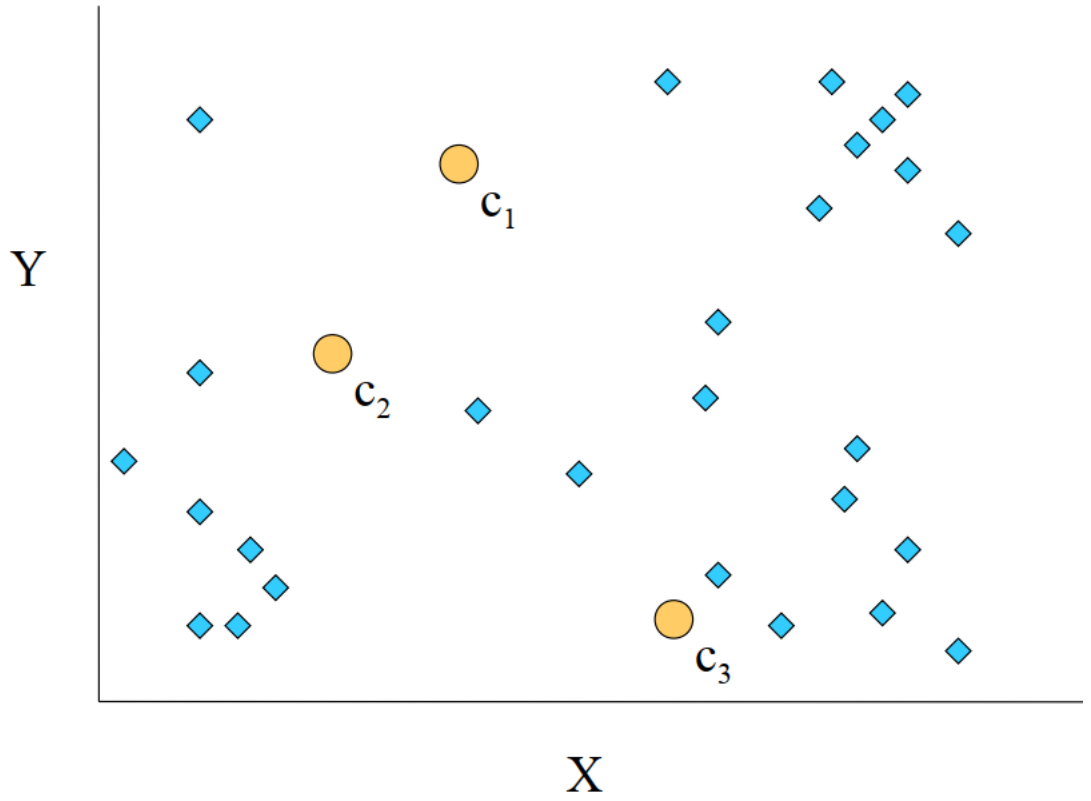




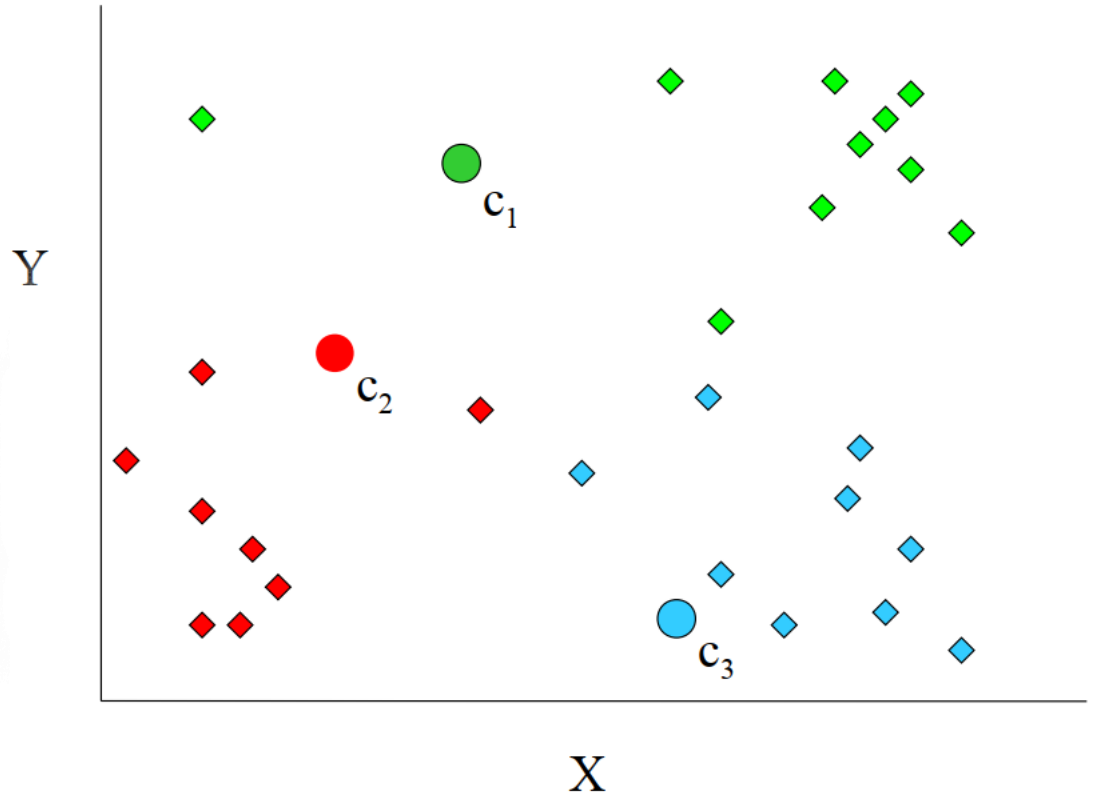
Clustering con K-Means: ejemplo



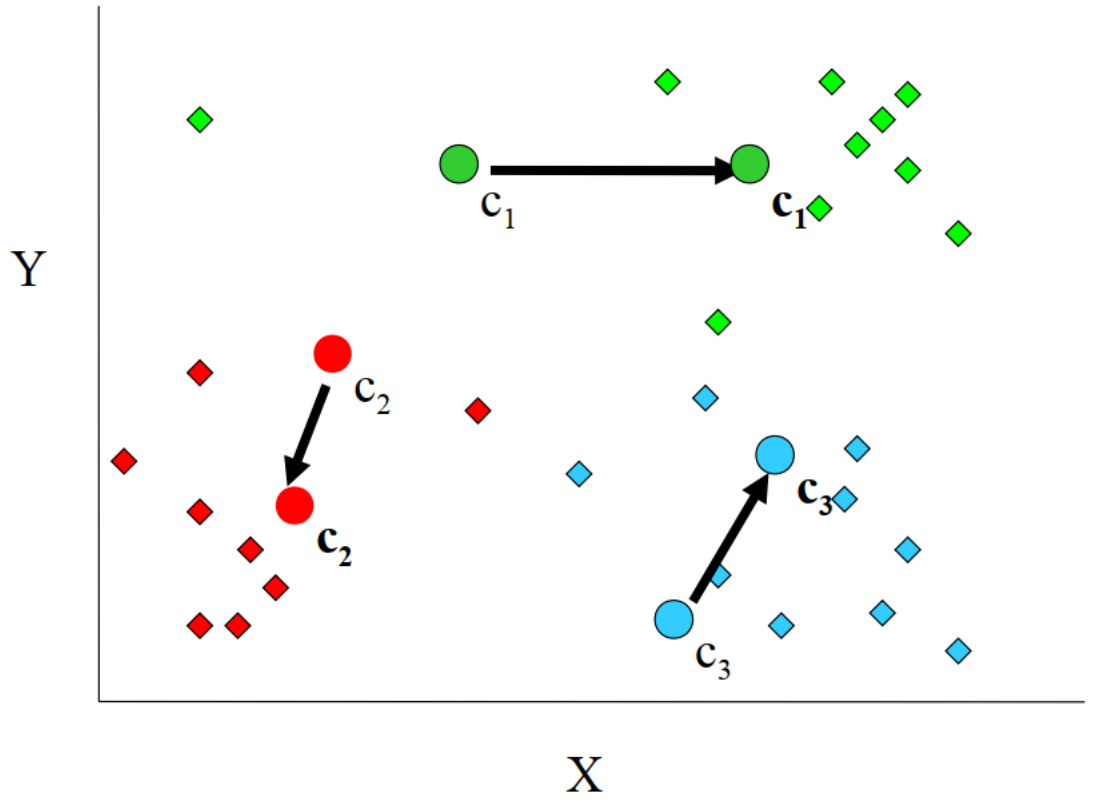
Clustering con K-Means: ejemplo



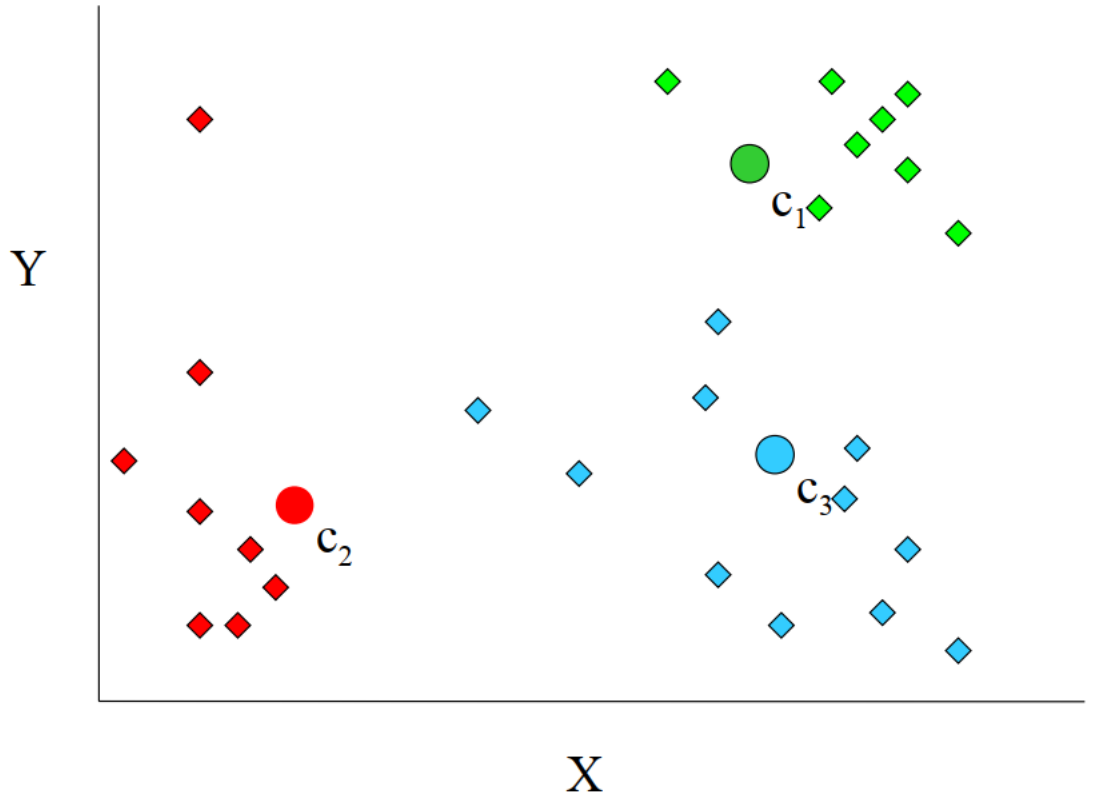
Clustering con K-Means: ejemplo



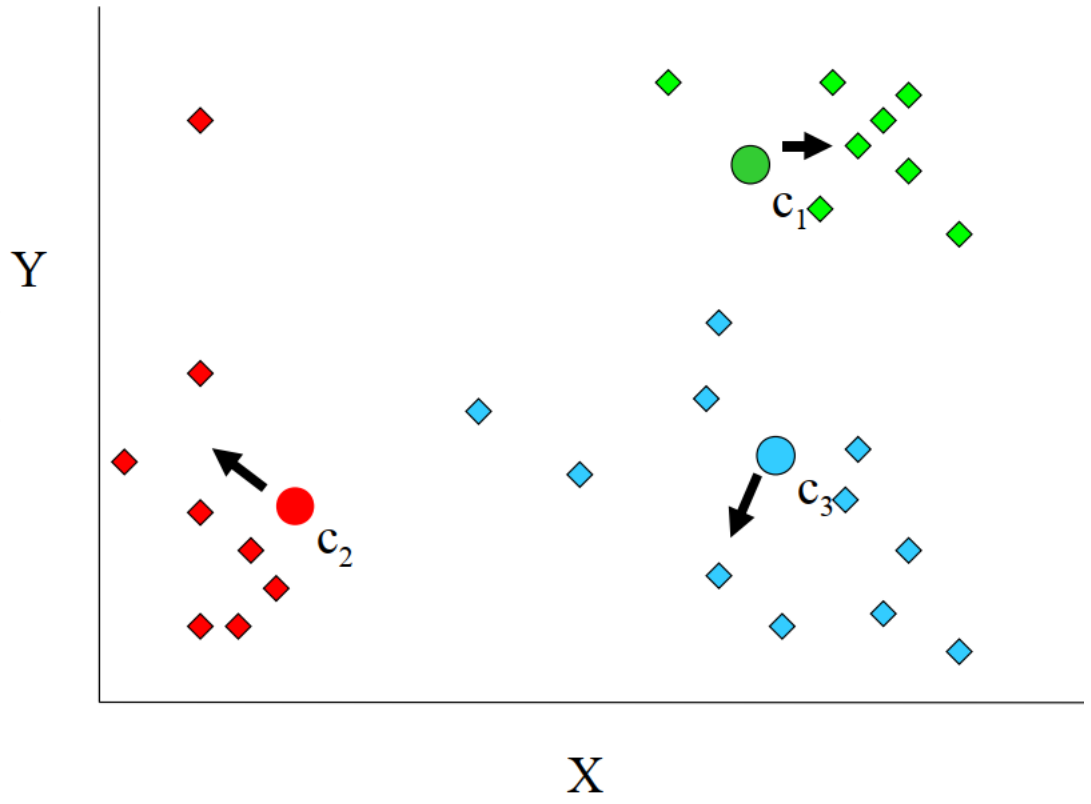
Clustering con K-Means: ejemplo



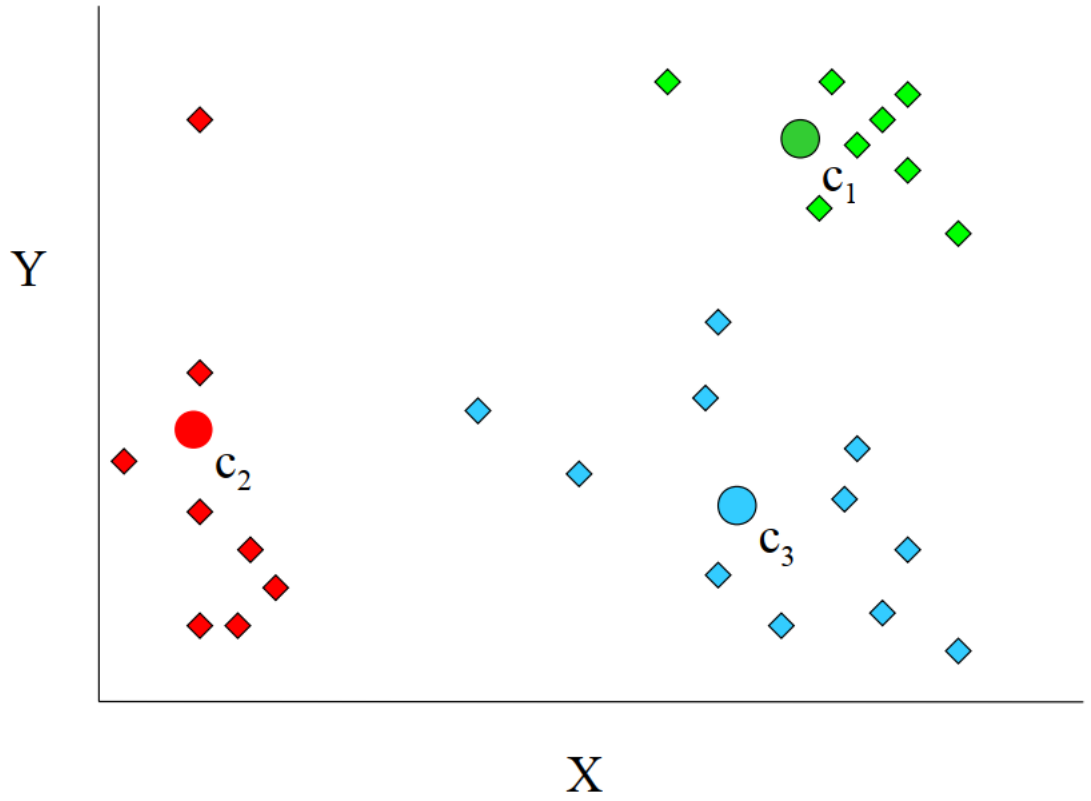
Clustering con K-Means: ejemplo



Clustering con K-Means: ejemplo



Clustering con K-Means: ejemplo



Tarea #2: K-means

developers.google.com

K-means

Leer las secciones:

- ✓ **¿Qué es el agrupamiento en clústeres k-means?**
- ✓ **Evaluación de resultados**
- ✓ **Ventajas y desventajas**

20 min

Agrupación en clústeres

El agrupamiento en clústeres es una estrategia clave de aprendizaje automático no supervisado para asociar elementos relacionados.

Cómo elegir el número k del algoritmo k-means?

Cómo elegir el número k?

Método del Codo para K-Means

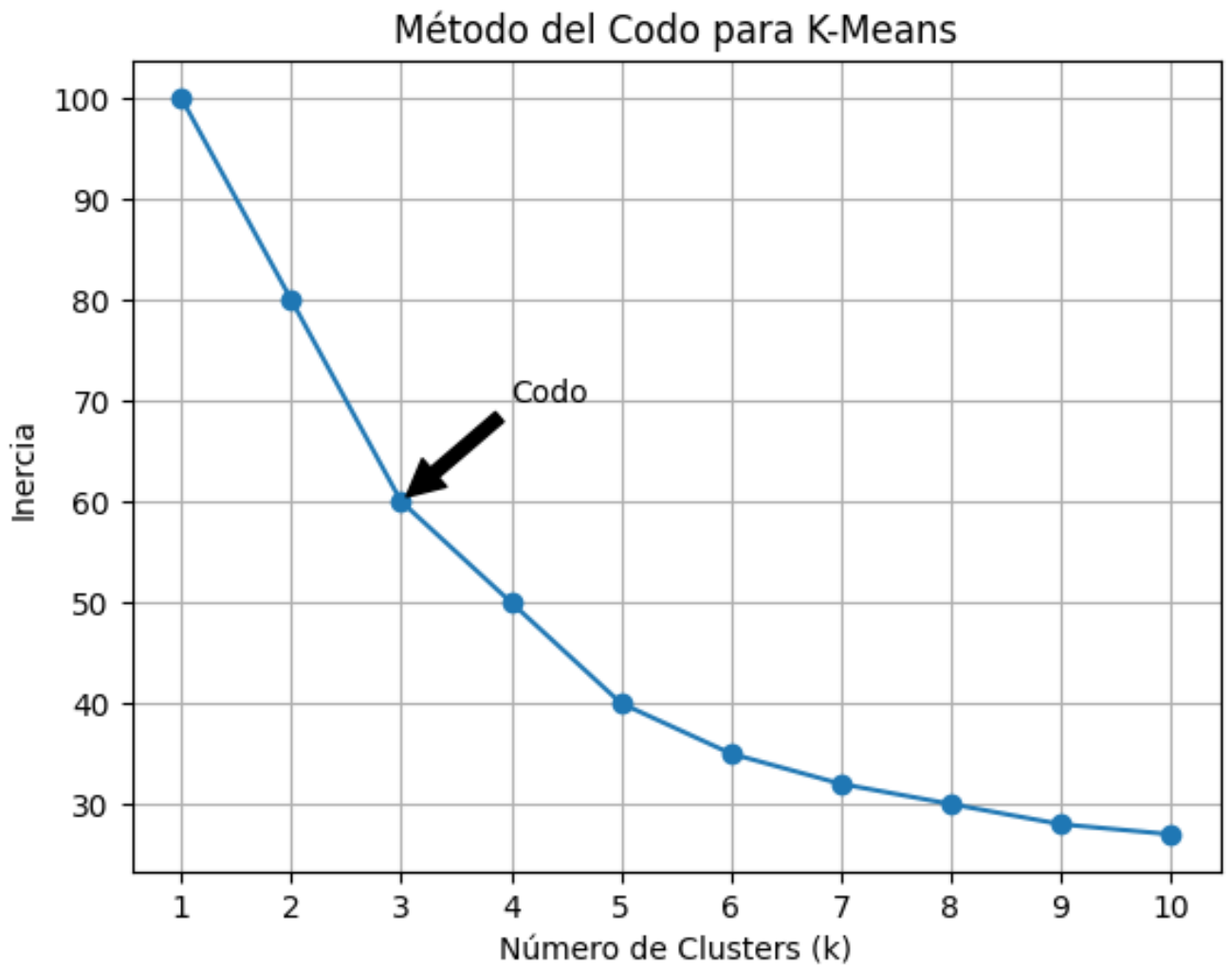
Coeficiente de Silueta (Silhouette)

Método del Codo para K-Means

El **método del codo** es una técnica visual para ayudar a determinar el **número óptimo de clusters (k)** en el algoritmo **K-Means.**

1. **Ejecutas K-Means varias veces**, probando **diferentes valores de k** (por ejemplo, desde 1 hasta un número razonable).
2. Para cada valor de k, **calcular la inercia**. La inercia es la suma de las distancias cuadradas de cada punto de datos al centroide de su cluster asignado. Una inercia menor indica que los puntos están más cerca de sus centroides y los clusters son más compactos.
3. **Graficar** la inercia en función del número de clusters (k).
4. **Buscas un "codo" en la gráfica.** Este es el punto donde la disminución de la inercia comienza a ralentizarse significativamente
5. **El valor de k en la posición del "codo"** se considera una buena estimación del número óptimo de clusters, ya que representa un equilibrio entre la compacidad de los clusters y el número de clusters.

Gráfica: Método del Codo para K-Means



Coeficiente de Silueta (Silhouette)

Silhouette se refiere a un método de interpretación y validación de la coherencia dentro del análisis de grupos. La técnica proporciona una **representación gráfica** sucinta de lo bien que se ha clasificado cada objeto.

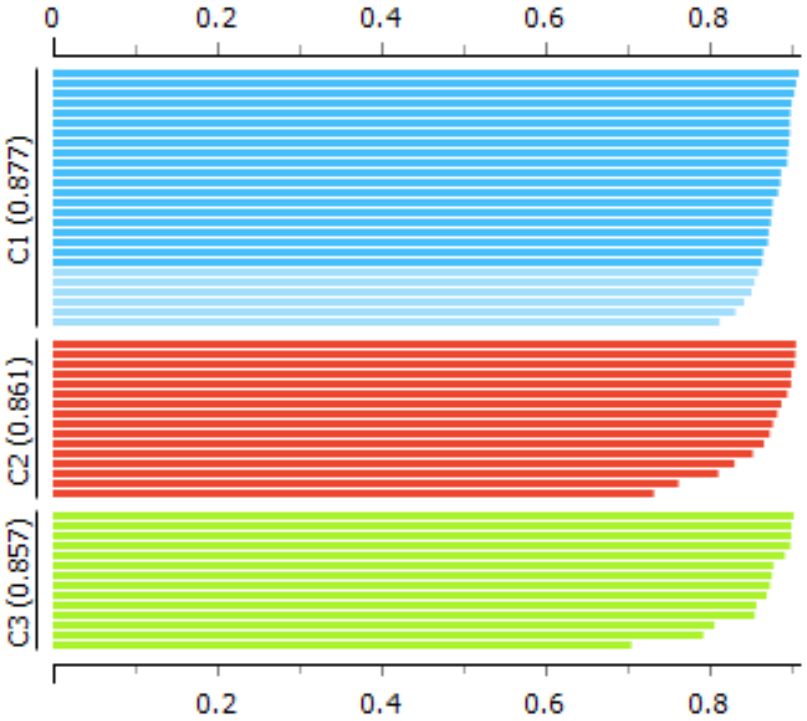
El valor de la silueta es una medida de cuán similar es un objeto a su propio cúmulo (**cohesión**) en comparación con otros cúmulos (**separación**).

La **silueta va de -1 a +1**, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con los cúmulos vecinos.

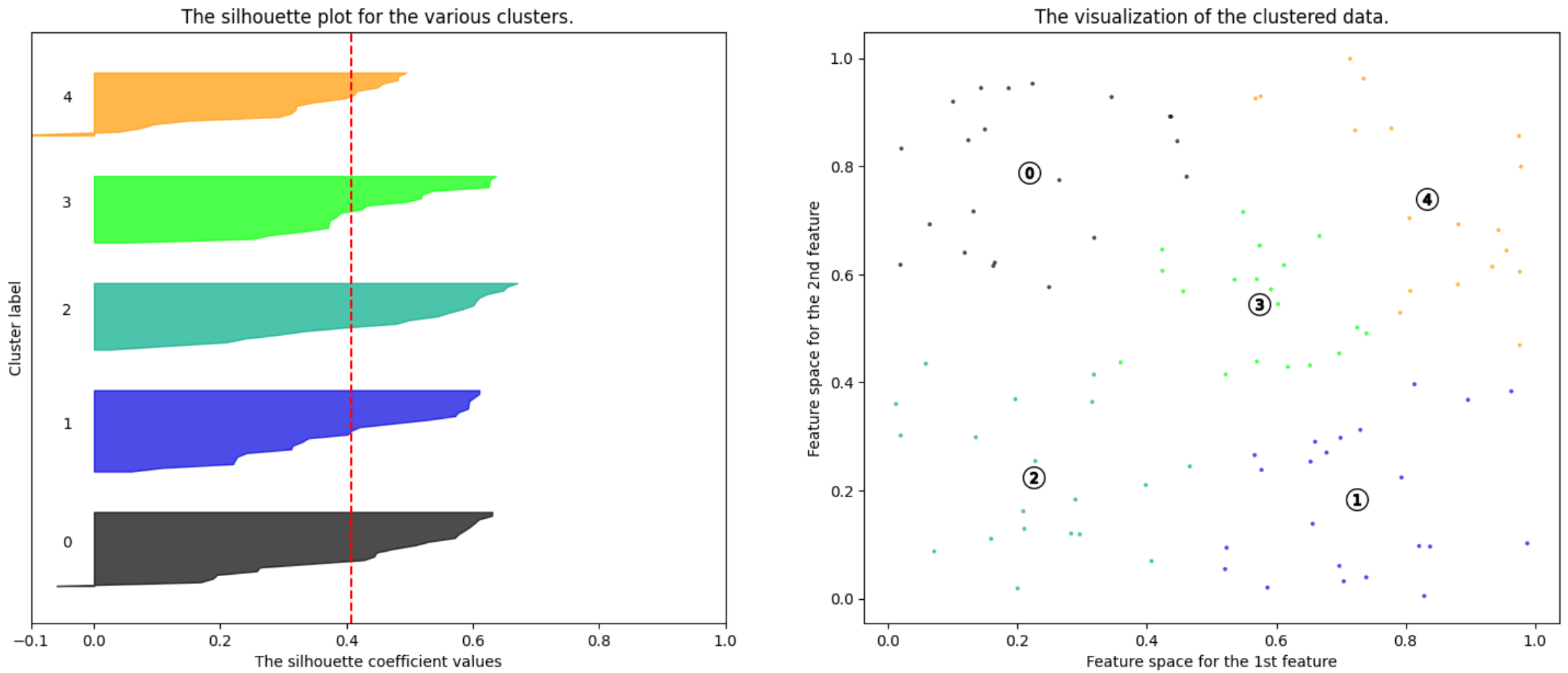
Coeficiente de Silueta (Silhouette)

Si la mayoría de los objetos tienen **un valor alto**, entonces la configuración del cúmulo es apropiada. Si muchos puntos tienen un **valor bajo** o negativo, entonces la configuración de cúmulos puede tener demasiados o muy pocos cúmulos.

La silueta puede ser calculada con cualquier métrica distancia, como la distancia euclidiana o la distancia Manhattan.



Gráfica: Silhouette



scikit-learn

Machine Learning in Python

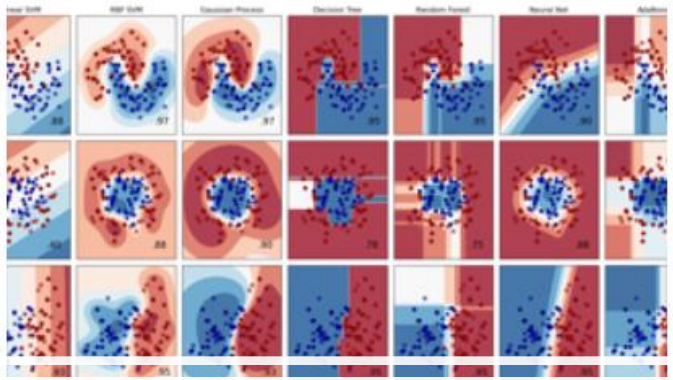
Getting Started | Release Highlights for 1.6

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

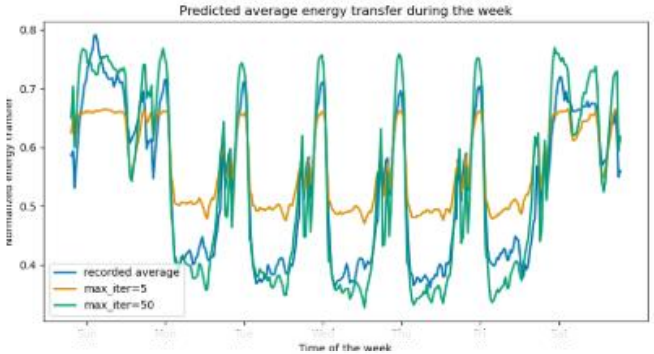


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

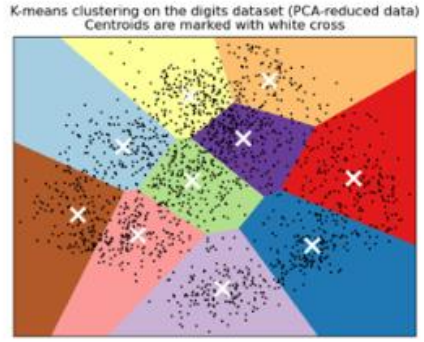


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency.

Model selection

Comparing, validating and choosing parameters and models.

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for

Conclusiones...

El aprendizaje no supervisado explora datos sin etiquetas para descubrir patrones.

El clustering es una técnica fundamental para agrupar datos similares.

K-Means es un algoritmo de clustering popular pero requiere especificar el número de clusters k.

La evaluación del clustering es más compleja que en el aprendizaje supervisado y utiliza métricas internas y externas.

Tarea #3: Resumen sobre Clustering

Clustering

Leer la sección y repasar conceptos aprendidos

5 min

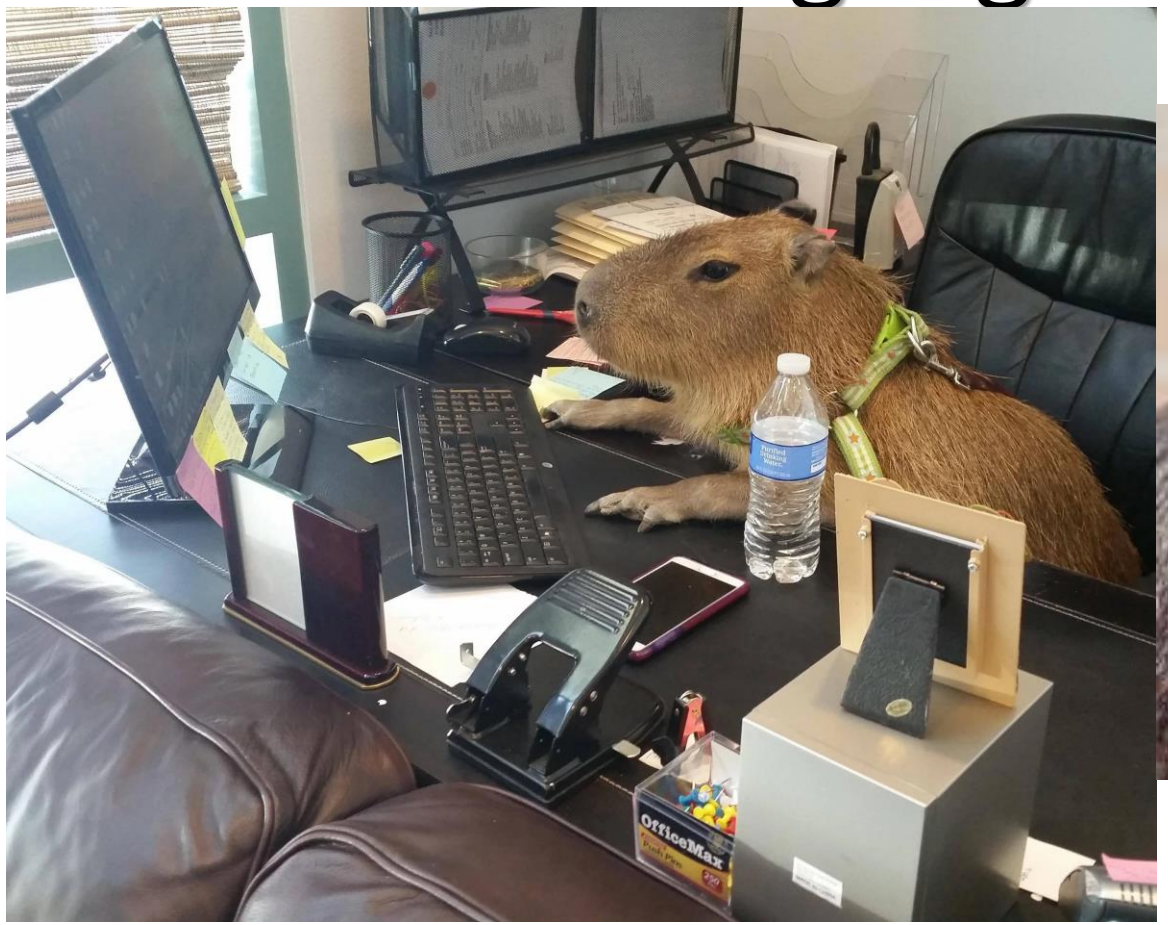
developers.google.com

Agrupación en clústeres

El agrupamiento en clústeres es una estrategia clave de aprendizaje automático no supervisado para asociar elementos relacionados.

Ejercicio práctico

colab.research.google.com

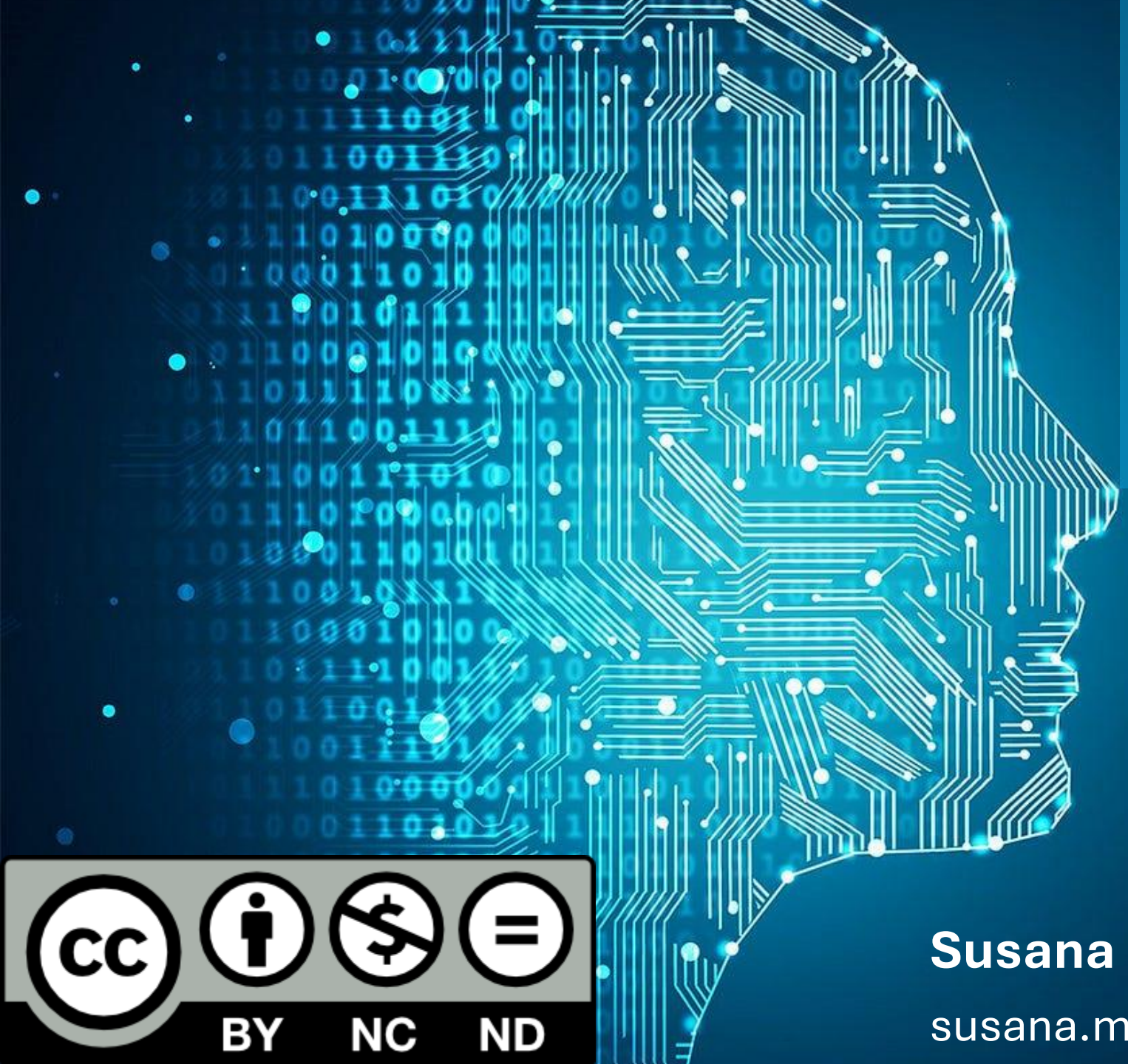


Ejercicio práctico: Google Colaboratory (*Colabs*)

- Página oficial: https://colab.google/
- Abrir Colab (incluye tutorial): https://colab.research.google.com/
- Guía para EDA: https://colab.research.google.com/github/Tanu-N-Prabhu/Python/blob/master/Exploratory_data_Analysis.ipynb
- Guía / tutorial para Selección de características con **scikit-learn**: https://www.datacamp.com/tutorial/feature-selection-python

Referencias

- “Silhouette (clustering)”, Wikipedia, la enciclopedia libre. el 10 de junio de 2024. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://es.wikipedia.org/w/index.php?title=Silhouette_(clustering)&oldid=160669974
- “Cómo crear un modelo de recomendación basado en machine learning. | Blog de Amazon Web Services (AWS)”. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://aws.amazon.com/es/blogs/aws-spanish/como-crear-un-modelo-de-recomendacion-basado-en-machine-learning/
- “Sistema de Recomendación Python y Machine Learning | Medium”. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://ivan-lee.medium.com/sistema-de-recomendacion-con-python-y-machine-learning-50858941b2bc
- “PRIVACIDAD y SISTEMAS DE RECOMENDACIÓN | LinkedIn”. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://www.linkedin.com/pulse/privacidad-y-sistemas-de-recomendaci%C3%B3n-jos%C3%A9-a-ferreira-queimada/
- “#13 El ABC del procesamiento de lenguaje natural”. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://impulsatek.com/13-el-abc-del-procesamiento-de-lenguaje-natural/
- “Tecnologías emergentes y datos abiertos: procesamiento del lenguaje natural | datos.gob.es”. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://datos.gob.es/es/documentacion/tecnologias-emergentes-y-datos-abiertos-procesamiento-del-lenguaje-natural
- S. Madala, “Principal Component Analysis (PCA)”, Scaler Topics. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://www.scaler.com/topics/nlp/what-is-pca/
- “Customer segmentation: Guide to types, tips, and strategy”, Zendesk. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://www.zendesk.com.mx/blog/customer-segmentation/
- “Customer Segmentation Analysis: Definition & Methods”, Qualtrics. Consultado: el 22 de abril de 2025. [En línea]. Disponible en: https://www.qualtrics.com/experience-management/brand/customer-segmentation/
- “scikit-learn: Machine Learning in Python”. Consultado: el 20 de febrero de 2025. [En línea]. Disponible en: https://scikit-learn.org/stable/
- Información e ideas presentadas basadas en el conocimiento general de modelos de lenguaje de IA. Gemini 2.9 Flash. Consultado: el 22 de abril de 2025. [En línea].



Machine Learning

Susana Medina Gordillo

susana.medina@correounivalle.edu.co