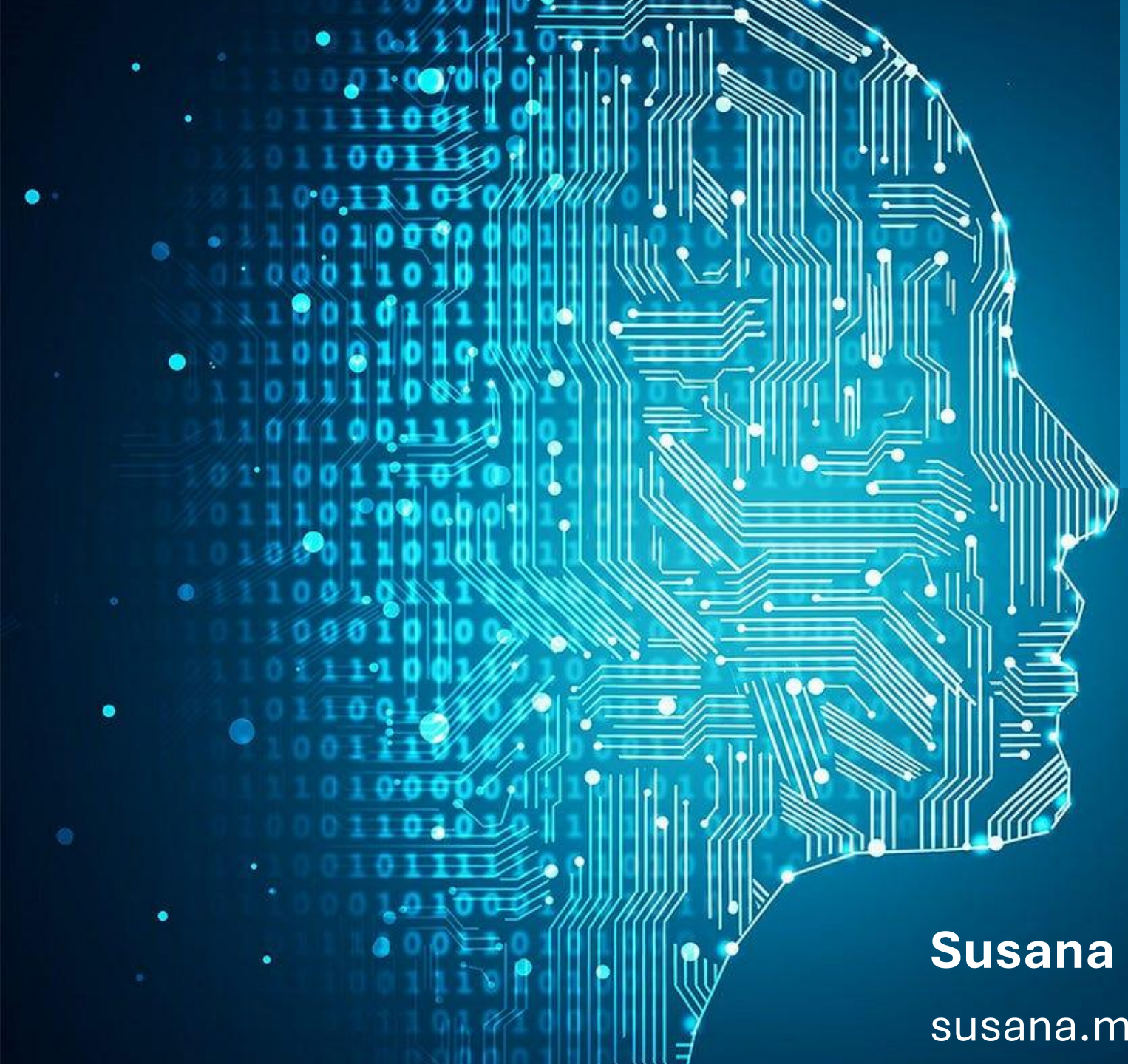
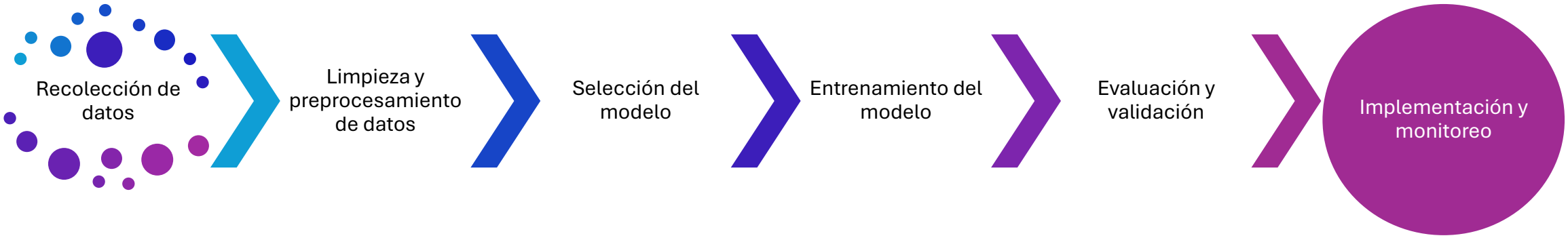


Machine Learning

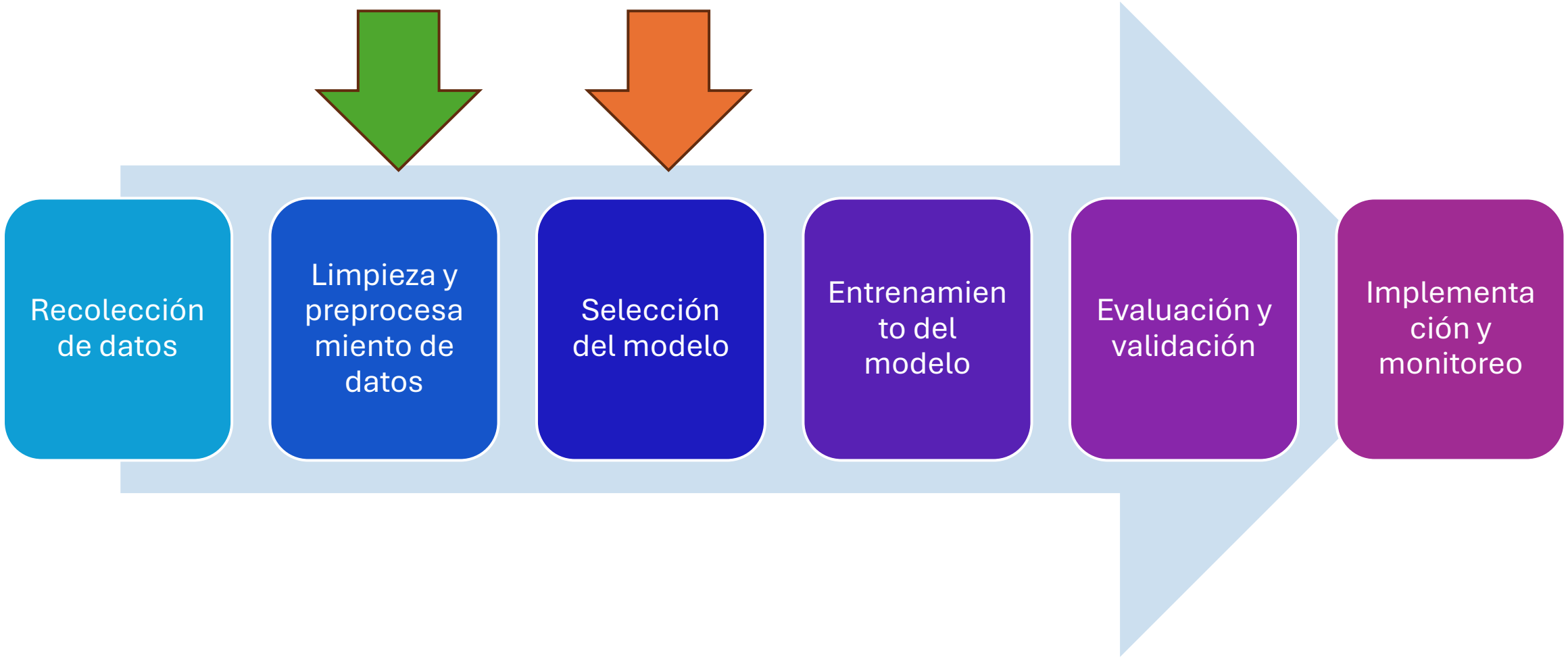
Susana Medina Gordillo

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

Métodos Avanzados de Clustering

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

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

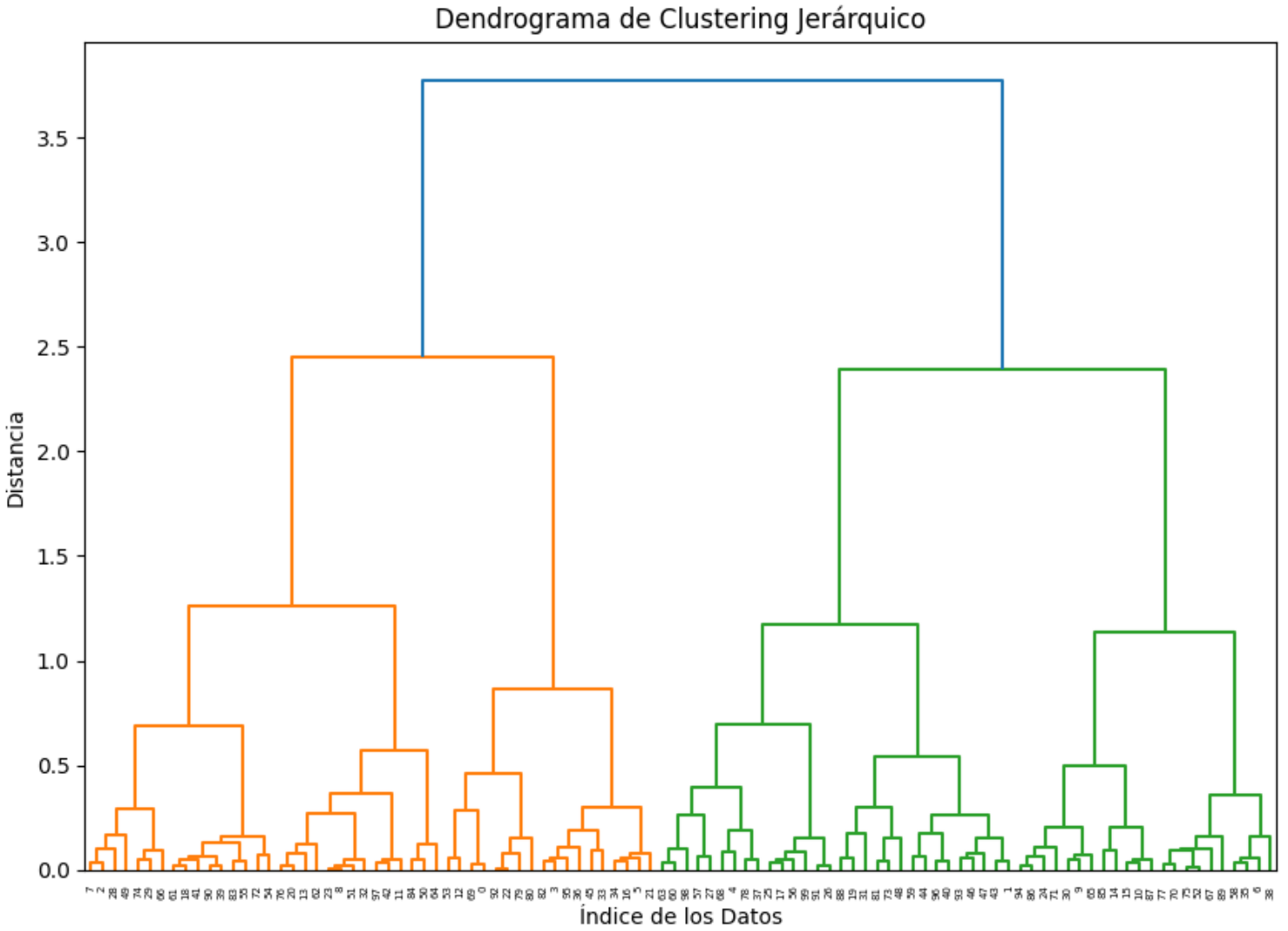
Agrupación en clústeres

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

Clustering Jerárquico

Clustering Jerárquico

Construye una jerarquía de clústeres, ya sea de abajo hacia arriba (**aglomerativo**) o de arriba hacia abajo (**divisivo**).



Clustering Jerárquico

Aglomerativo

(Bottom-up)

Cada punto comienza como su propio clúster.

En cada paso, se fusionan los dos clústeres más cercanos hasta que solo queda un clúster o se alcanza un criterio de parada.

Divisivo

(Top-down)

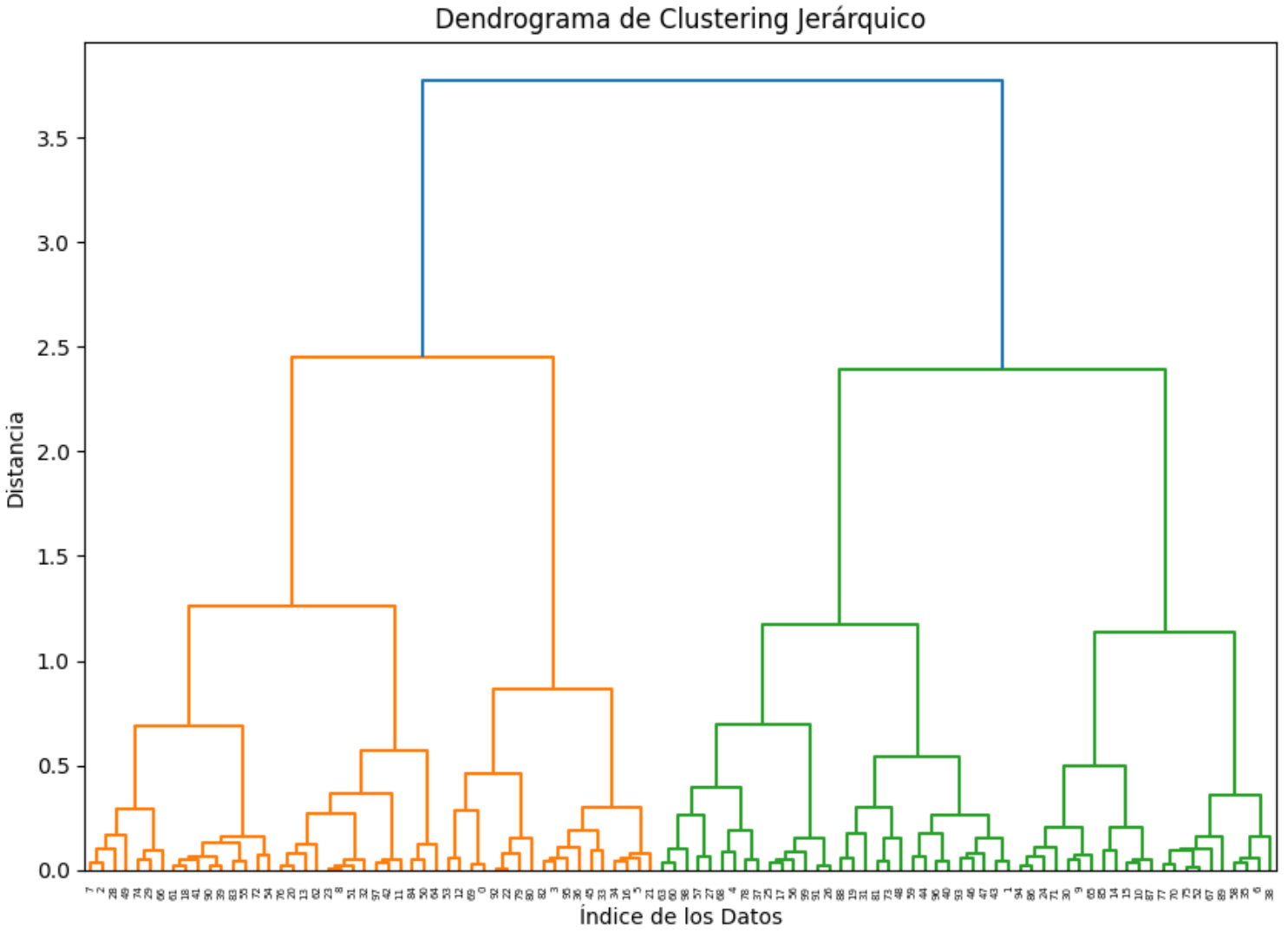
Comienza con todos los puntos en un solo clúster.

En cada paso, se divide el clúster más heterogéneo hasta que cada punto está en su propio clúster o se alcanza un criterio de parada.

Clustering Jerárquico

El **dendrograma**

es el diagrama en forma de árbol que muestra la **jerarquía** de los clústeres y las distancias a las que se fusionaron o dividieron.



Clustering Jerárquico

Ventajas

✓No requiere especificar el **número de clústeres** de antemano

✓Proporciona una estructura jerárquica **útil para la interpretación**.

Desventajas

➢Puede ser **computacionalmente costoso** para grandes conjuntos de datos.

➢La elección de la **métrica de distancia** y el método de enlace (cómo se mide la distancia entre clústeres) puede afectar significativamente los resultados.

DBSCAN : Density-Based Spatial Clustering of Applications with Noise

DBSCAN

Agrupa puntos que están **muy cerca** unos de otros, basándose en una medida de **densidad de puntos**.

Marca como ruido los puntos que se encuentran solos en regiones de baja densidad.

DBSCAN: parámetros clave

Epsilon (eps): La distancia máxima entre dos puntos para que se consideren vecinos.

MinPts: El número mínimo de puntos dentro del radio epsilon para que un punto central forme un clúster.

DBSCAN: funcionamiento

1. Para cada punto, cuenta cuántos vecinos tiene dentro de **epsilon**.
2. Si un punto tiene al menos **MinPts** vecinos, se considera un punto central.
3. Los puntos que están dentro del radio de un **punto central** se consideran **puntos alcanzables**.
4. Un clúster se forma a partir de un punto central y todos los puntos que son alcanzables por **densidad** desde él (incluyendo otros puntos centrales).
5. Los puntos que no son puntos centrales ni alcanzables se marcan como ruido (outliers).

DBSCAN

Ventajas

- ✓Puede encontrar clústeres de **formas arbitrarias**.
- ✓Es **robusto** al ruido y a los *outliers*,
- ✓NO requiere especificar el número de clústeres de antemano.

Desventajas

- ➢El **rendimiento** depende significativamente de la elección de los parámetros eps y MinPts
- ➢Puede tener dificultades con clústeres de densidades muy variables.

Dónde están los algoritmos de clustering avanzado?

scikit-learn

Machine Learning in Python

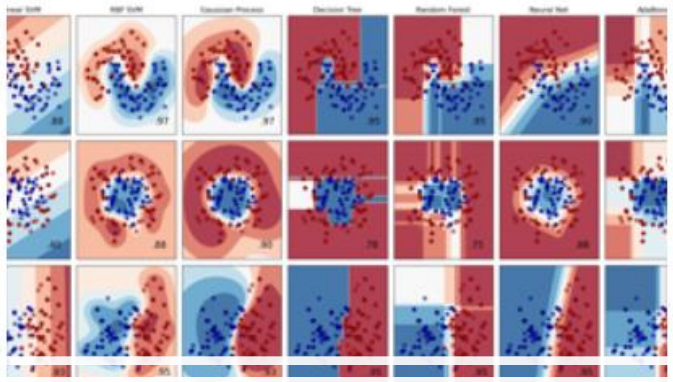
Getting Started | Release Highlights for 1.6

- 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...

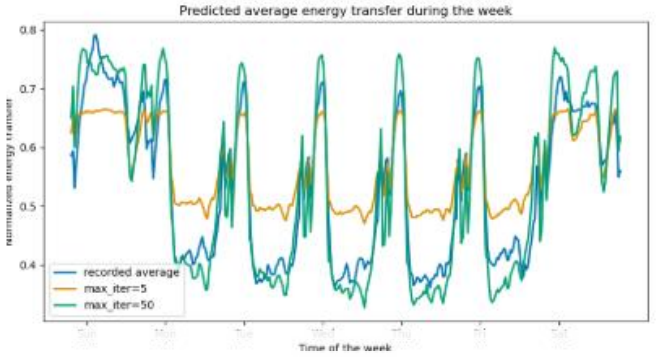


Examples

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...

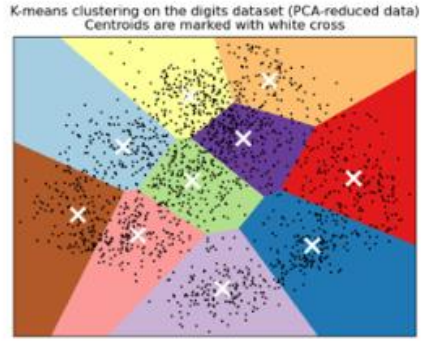


Examples

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 clustering es una técnica fundamental para agrupar datos similares.

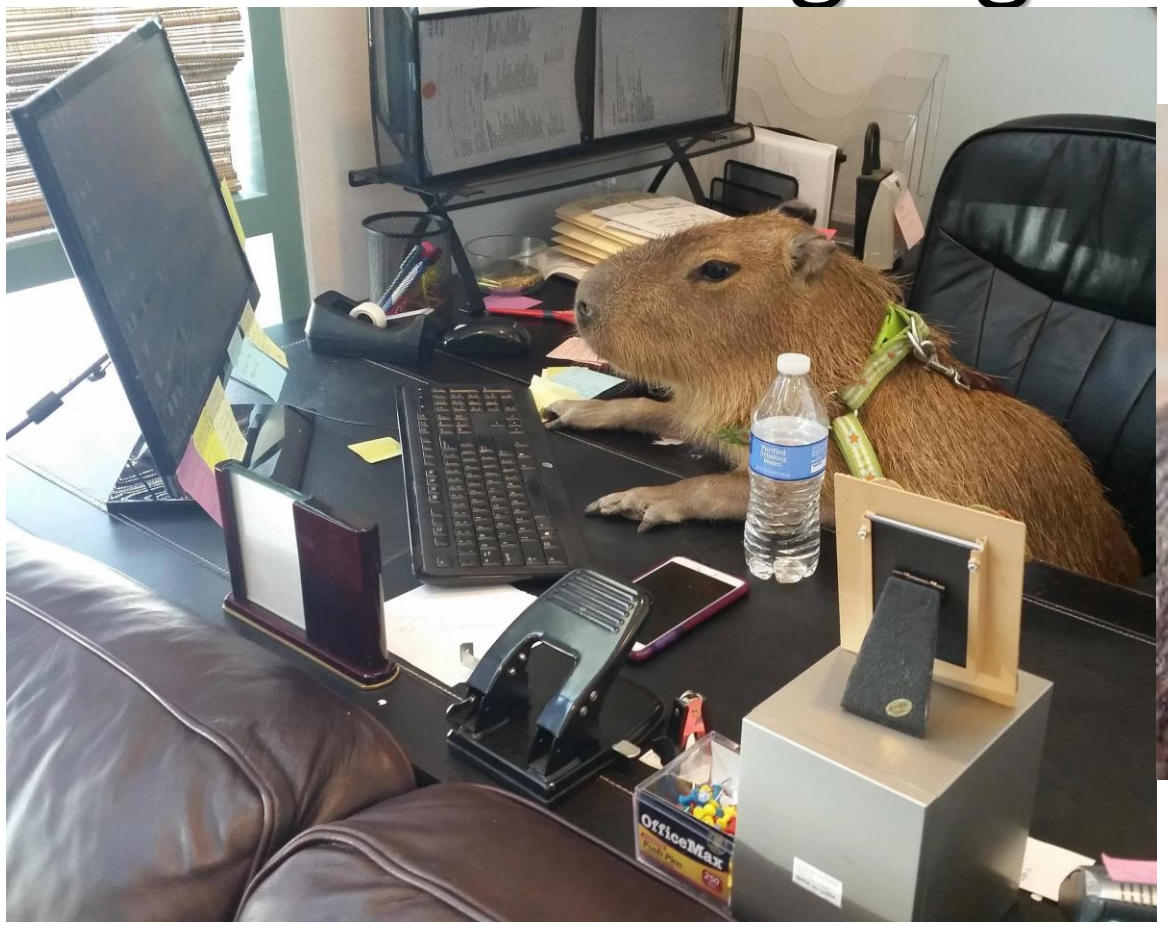
Ambos métodos ofrecen enfoques diferentes al clustering en comparación con algoritmos particionales como K-Means.

El clustering jerárquico es útil para explorar la estructura de los datos a diferentes niveles de granularidad, mientras que DBSCAN es efectivo para descubrir clústeres basados en la densidad y manejar el ruido

La elección del método depende de las características específicas de los datos y los objetivos del análisis.

Ejercicio práctico

colab.research.google.com

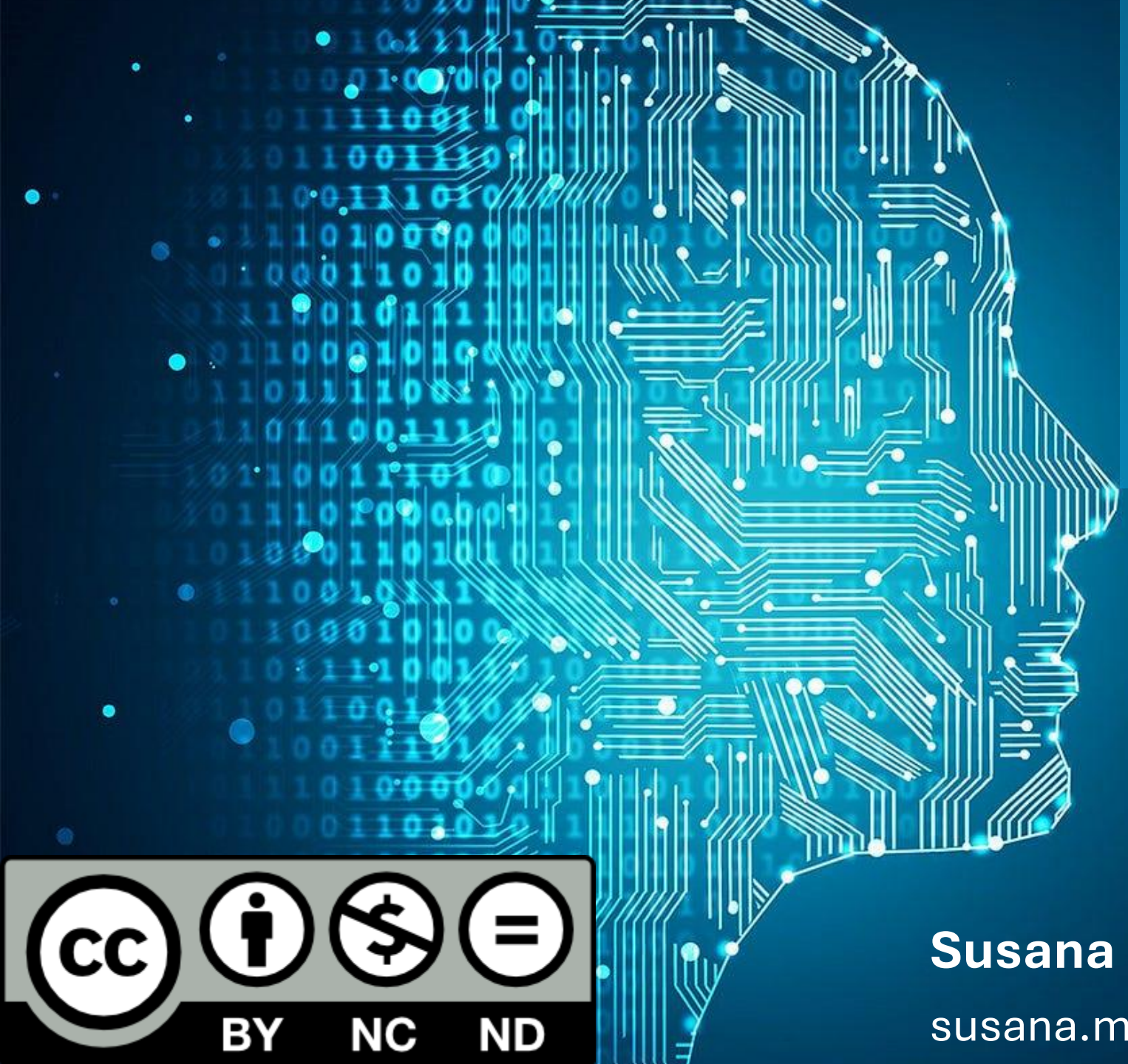


Ejercicio práctico: Google Colaboratory (*Colabs*)

- 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