FUNDACIÓN YPF

Módulo VI - Aprendizaje NO supervisado.

Clase 22: Clustering - Métricas

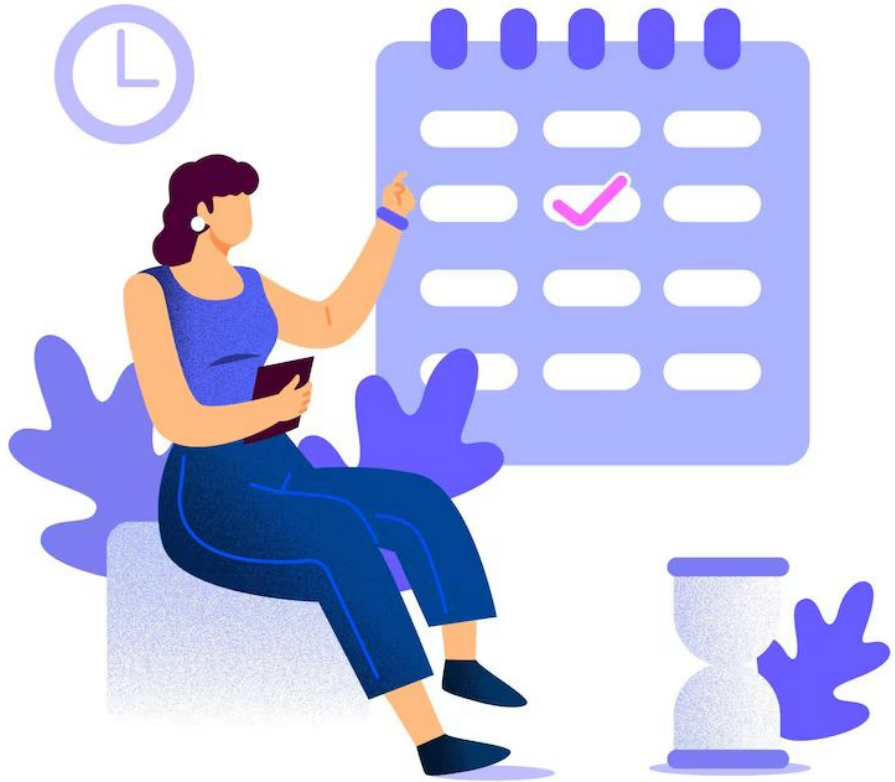
FUNDACIÓN YPF

¿Ponemos a grabar el taller?

FUNDACIÓN YPF

¿QUÉ VAMOS A VER HOY?

- Métricas para Clustering
  - Método Elbow
  - Método Silhouette

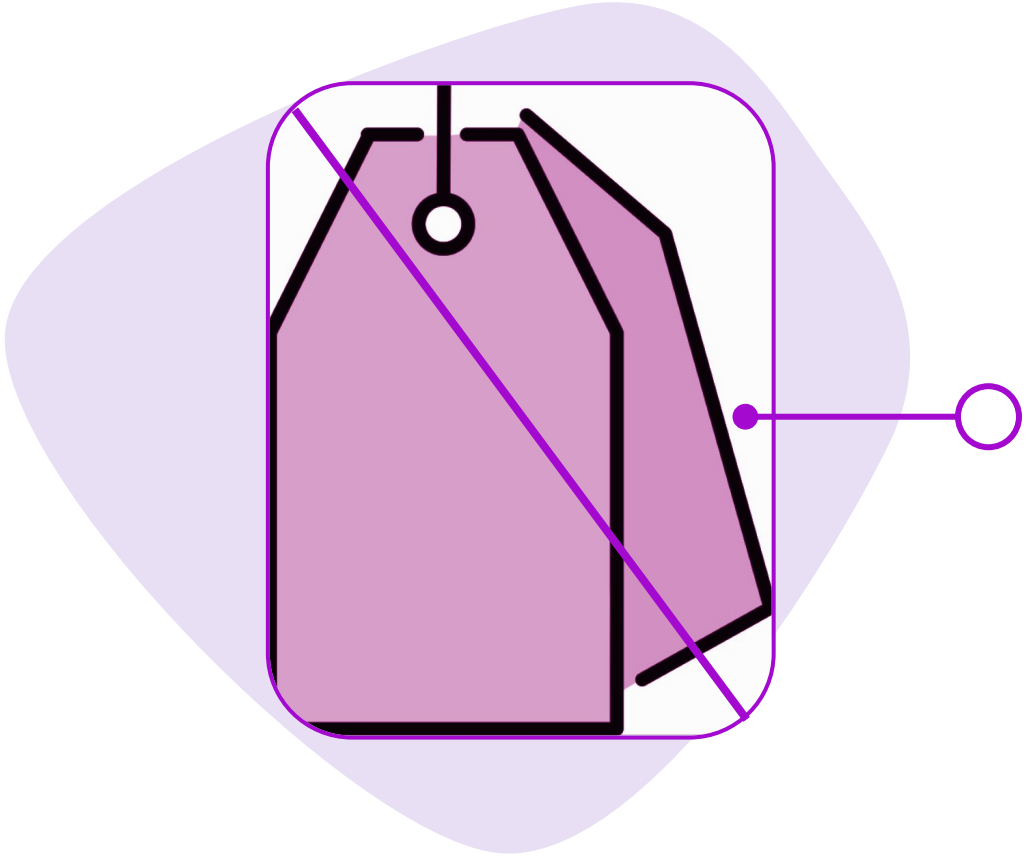


FUNDACIÓN YPF

Métricas

FUNDACIÓN YPF

Clustering



Los datos **NO** están **etiquetados**.

No hay un valor **esperado**.

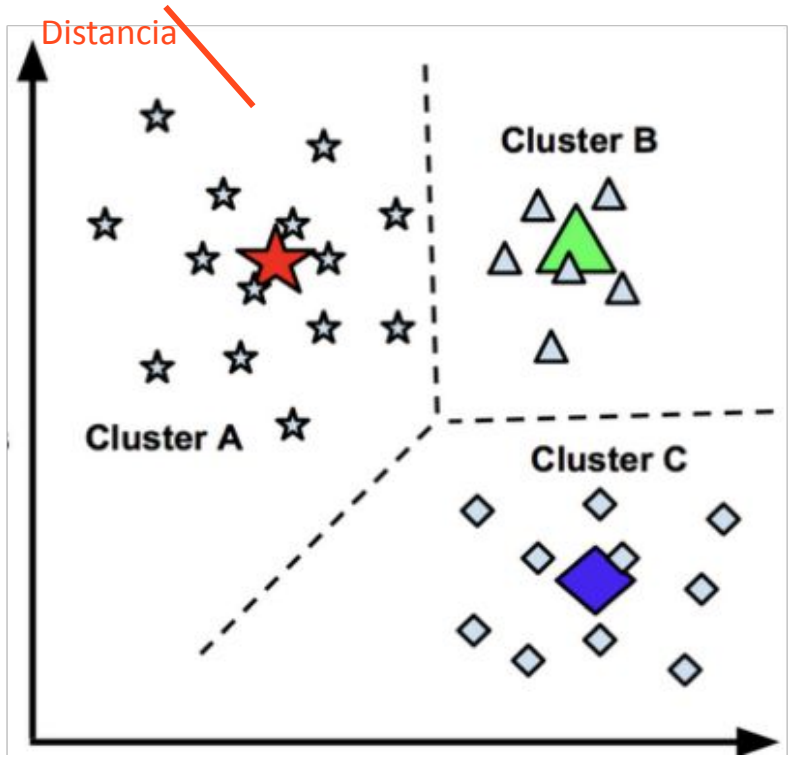
FUNDACIÓN YPF

¿CÓMO SABEMOS QUE NUESTRO MODELO FUNCIONA CORRECTAMENTE?

FUNDACIÓN YPF

Clustering

Debemos encontrar una medida para la **validación e interpretación de clusters en un dataset.**

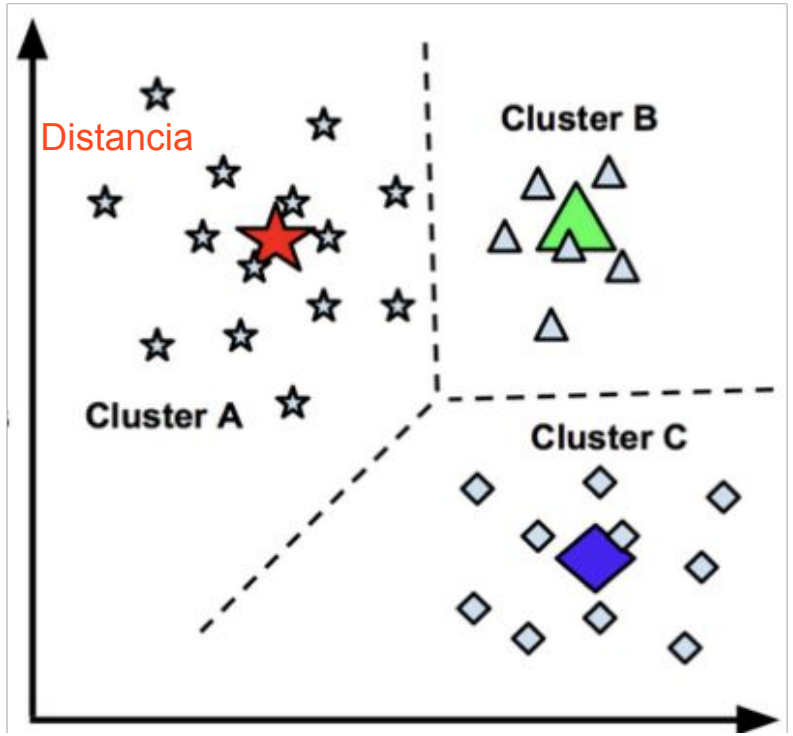


Podemos medir cuál es la distancia media de cada dato al centroide más cercano.

FUNDACIÓN YPF

Clustering

Debemos encontrar una medida para la **validación e interpretación de clusters en un dataset.**



$$d(i) = \|\mathbf{X_i} - \mathbf{Cj}\|^2$$

Distancia para cada punto

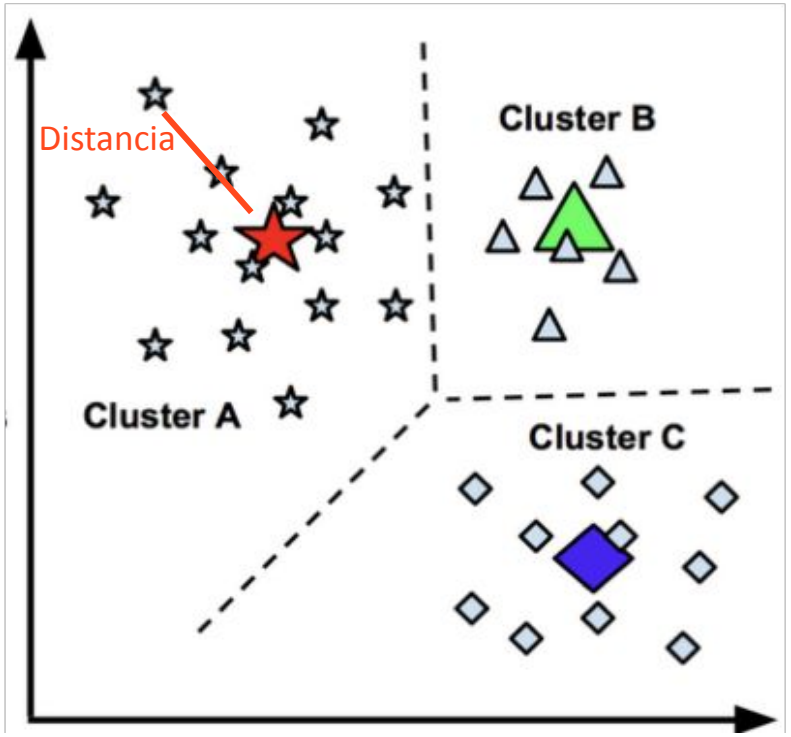
Posición del punto i

Posición del centroide más cercano

FUNDACIÓN YPF

Clustering (K-Means)

Buscamos una medida para evaluar qué tan buena resulta la **designación de clusters** en **K-Means**.



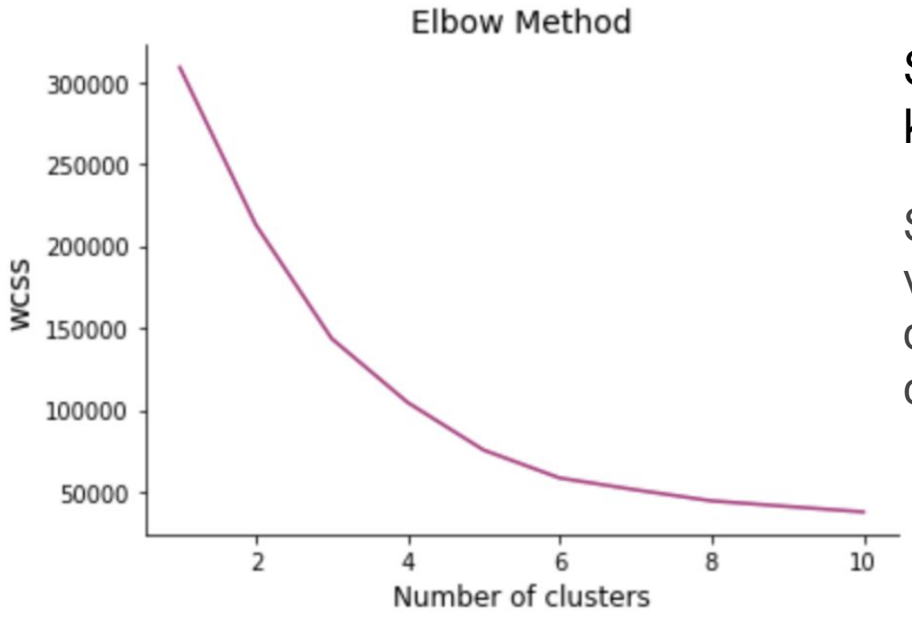
$$D = \frac{1}{N} \sum_{i=1}^{N} d(i)$$

Distancia media total

FUNDACIÓN YPF

Método Elbow

Buscamos una medida para evaluar qué tan buena resulta la **designación de clusters** en **K-Means**.



Se gráfica distancia (o inertia en sklearn) vs. k y se encuentra donde esta el codo

Se busca dónde está el **codo** de la curva. El valor de la distancia o inercia siempre desciende cuando aumenta el número de clusters.

FUNDACIÓN YPF

Método Silhouette

Medida para la validación e interpretación de clusters en un dataset (para cualquier método de clustering).

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Se mide qué tan parecidos son los datos con respecto a otros de su propio cluster (cohesión) en comparación con qué tan parecidos son a los datos en otros clusters (separación)

s(i): Valor de silhouette para el dato i.
a(i): Distancia media del dato i con el resto de su cluster
b(i): DIstancia media del dato i con el cluster más cercano

FUNDACIÓN YPF

Método Silhouette

Medida para la validación e interpretación de clusters en un dataset (para cualquier método de clustering).

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

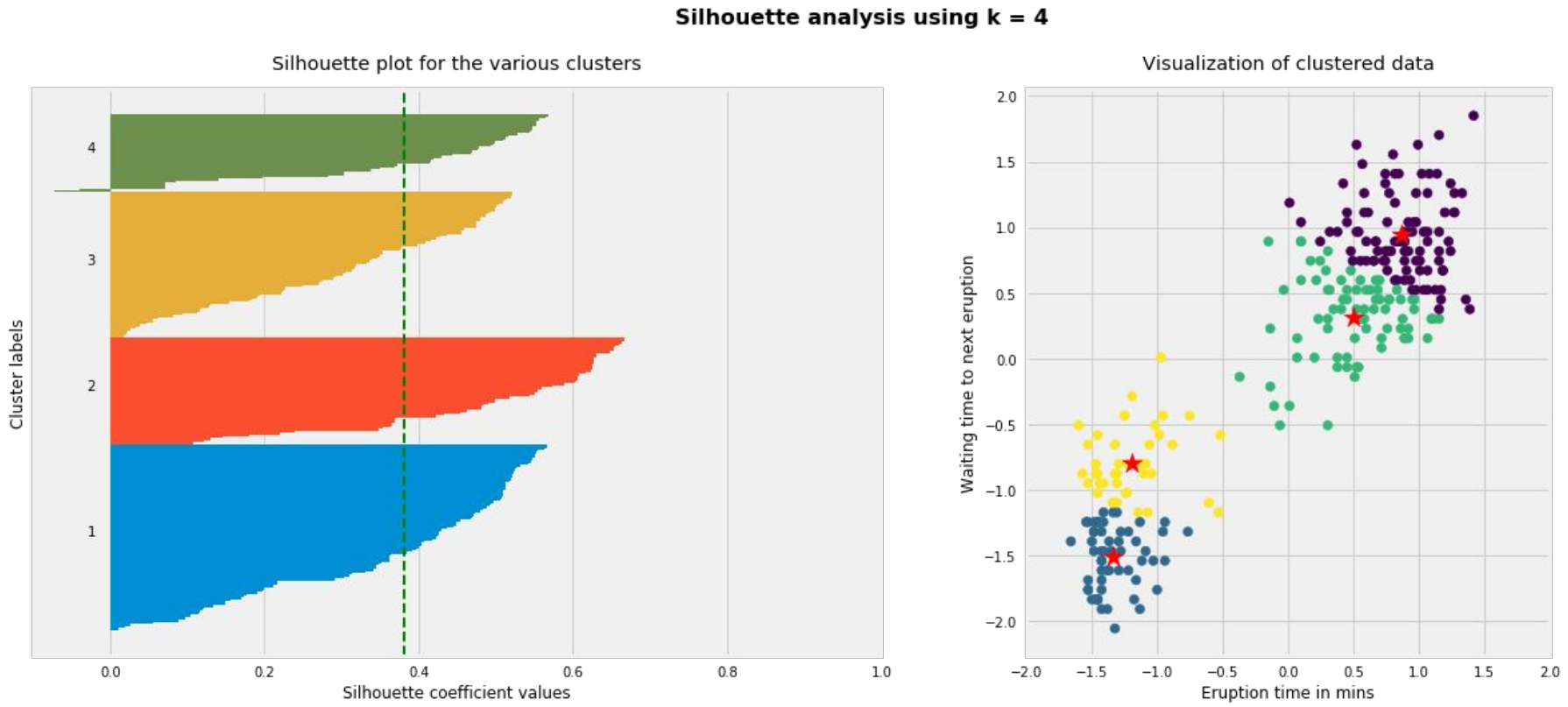
Nos da una medida para cada punto de qué tan bien están ubicados en sus clusters. Para una medida de **todo el conjunto**, se toma la **media** de todos los valores.

s(i): Valor de silhouette para el dato i.
a(i): Distancia media del dato i con el resto de su cluster
b(i): DIstancia media del dato i con el cluster más cercano

FUNDACIÓN YPF

Método Silhouette

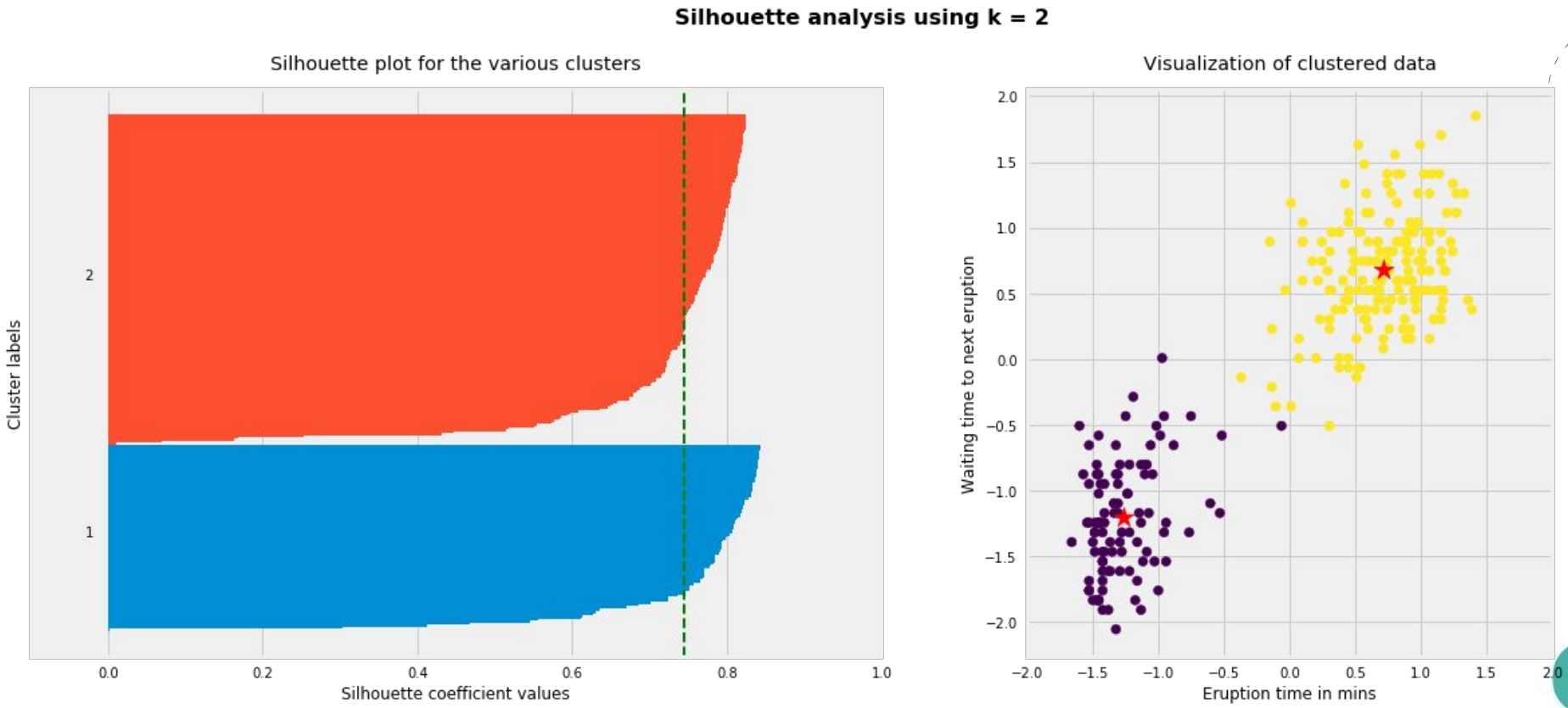
Se espera que si están bien asignados y armados los clusters, el perfil de todos los datos sea parejo.



FUNDACIÓN YPF

Método Silhouette

Se espera que si están bien asignados y armados los clusters, el perfil de todos los datos sea parejo.



FUNDACIÓN YPF

Sección Práctica

Trabajamos con métricas en la Notebook 21

TRABAJAMOS EN SALAS

FUNDACIÓN YPF

Sección Práctica

Trabajamos con métricas en la Notebook 21

TRABAJAMOS EN SALAS

FUNDACIÓN YPF

Trabajamos en salas de zoom

Métricas

Trabajaremos con la Notebook 21

En los grupos establecidos, ejercitamos como se evalúan los modelos no supervisados

50 minutos de actividad

FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Repasamos dudas

Trabajamos con la Notebook 21

Revisamos los conceptos y el código trabajados en la notebook 21

FUNDACIÓN YPF

Desafío 15 (continuación)

Para la siguiente repasar la notebook 21

FUNDACIÓN YPF

¿Alguna consulta?

FUNDACIÓN YPF

Repasemos

FUNDACIÓN YPF

Eligiendo algoritmo

Tarea

Definir de forma clara el objetivo

1

Información

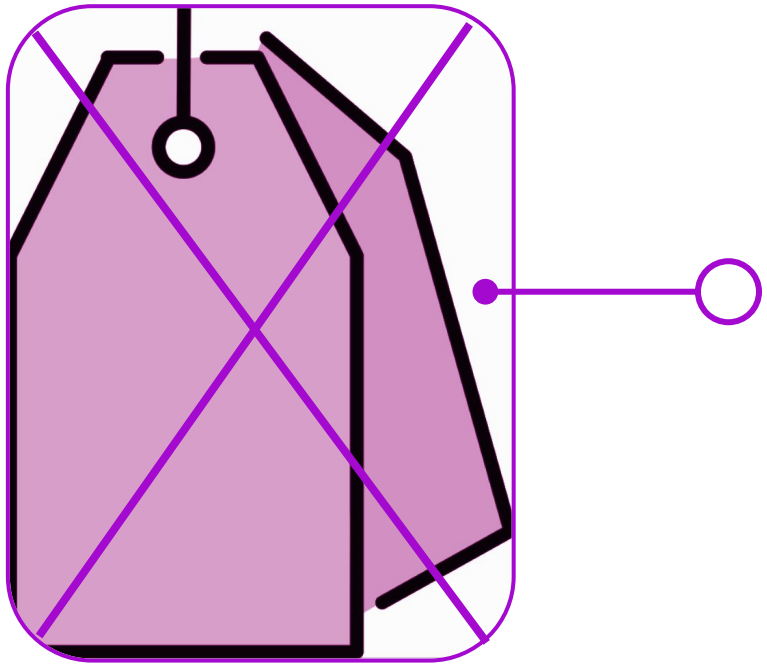
Con qué datos se cuenta para lograr el objetivo

2

- APRENDIZAJE **SUPERVISADO**
- APRENDIZAJE **NO SUPERVISADO**

FUNDACIÓN YPF

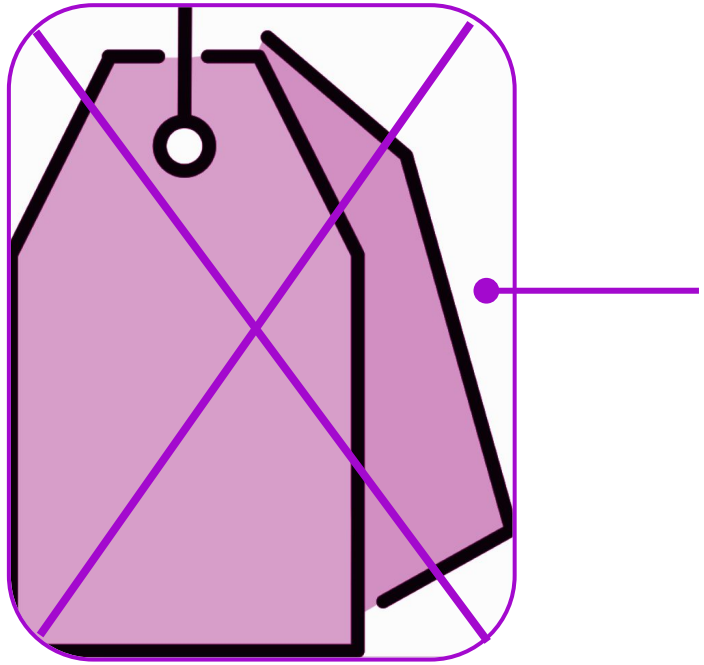
Aprendizaje No Supervisado



El algoritmo infiere patrones de un conjunto de datos que, a diferencia del aprendizaje supervisado, **no** están **etiquetados**. Puede utilizarse para descubrir la estructura subyacente de los datos

FUNDACIÓN YPF

Aprendizaje No Supervisado



- Clustering
- Reducción de la Dimensionalidad

FUNDACIÓN YPF

Reducción de la dimensionalidad

FUNDACIÓN YPF

Reducción de la dimensionalidad

El objetivo consta de **reducir la cantidad de features** de un dataset, pero **reteniendo** la **mayor** cantidad de **información** posible.

FUNDACIÓN YPF

Reducción de la dimensionalidad

Consiste en **reducir** la cantidad de **features** de un dataset, pero **reteniendo** la mayor cantidad de **información** posible.

- **Aplicaciones**
  - Reducir la complejidad del input en un modelo de regresión o clasificación
  - Visualización
  - Detectar features relevantes en datasets
- **Algoritmos**
  - **Principal Component Analysis**
  - Multidimensional Scaling
  - t-SNE: t-distributed Stochastic Neighbor Embedding
  - LDA: Linear Discriminant Analysis

FUNDACIÓN YPF

Principal Component Analysis

Técnica utilizada para describir un conjunto de datos en términos de nuevas variables o componentes no correlacionados.

- El algoritmo encuentra nuevos componentes que describen los datos
- Los componentes se ordenan por la cantidad de varianza original que describen, reduciendo la dimensionalidad del conjunto de datos.

- Utiliza **Single Vector Decomposition (SVD)**

FUNDACIÓN YPF

¡Muchas gracias!