Clasificación con **Kmedias / Aglomerativo y Moda**

Dataset: Semillas (UCI)

UNIVERSIDAD DE VALLADOLID

MINERÍA DE DATOS

4º Curso. Grado en Ingeniería Informática
5º Curso. Doble Grado Informática/Estadística (INDAT)

Departamento de Informática (ATC, CCIA y LSI)

Plataforma de trabajo – Práctica inicial

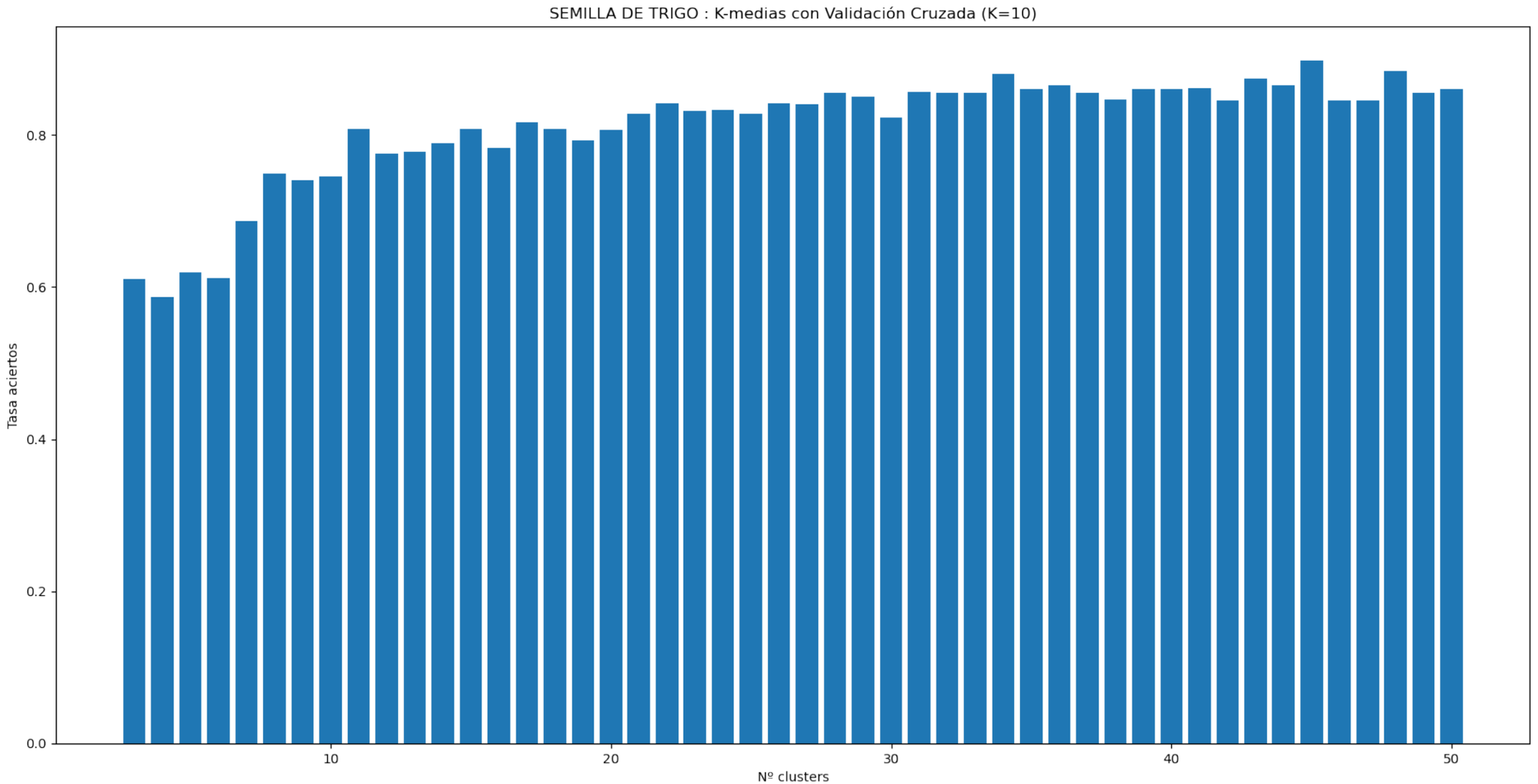
- Con el fichero de ejemplo https://archive.ics.uci.edu/ml/datasets/seeds#
  - **Estandarizar** a una distribución gausiana de media cero y desviación estándar igual a 1.
  - Aplicar clustering usando el algoritmo de las **k-medias**
  - La etiqueta de cada cluster se hará coincidir con la **moda** entre sus muestras
  - Evaluar su eficiencia usando el método de **validación cruzada** (10 carpetas) (**estratificada**)
  - Variar este K entre 3 y 50. Hallar **k óptimo**.
  - Con este k óptimo, aplicar el algoritmo **Aglomerativo** con las opciones por defecto y calcular su tasa de aciertos, que se comparará la obtenida con k-medias.
  - Obtener el dendograma.

Resultados: Kmeans y CV (K=10)

Clusters / CV	3	4	5	6	7	8	9	10	11	12	...	41	42	43	44	45	46	47	48	49	50
0	0.571429	0.761905	0.571429	0.809524	0.619048	0.666667	0.904762	0.761905	0.761905	0.761905	...	0.952381	0.809524	0.857143	0.761905	0.904762	0.952381	0.952381	0.809524	0.809524	0.904762
1	0.523810	0.809524	0.904762	0.809524	0.857143	0.809524	0.666667	0.523810	0.666667	0.904762	...	0.666667	0.952381	0.952381	0.857143	0.904762	0.809524	0.952381	0.809524	0.857143	0.904762
2	0.619048	0.619048	0.714286	0.857143	0.809524	0.714286	0.714286	0.809524	0.857143	0.761905	...	0.904762	0.857143	0.761905	0.857143	0.952381	0.809524	0.857143	0.952381	0.857143	0.952381
3	0.619048	0.714286	0.809524	0.809524	0.761905	0.857143	0.761905	0.666667	0.666667	0.761905	...	0.904762	0.904762	1.000000	0.857143	0.857143	0.809524	0.857143	0.904762	0.904762	0.809524
4	0.666667	0.619048	0.523810	0.714286	0.666667	0.714286	0.714286	0.714286	0.714286	0.809524	...	0.714286	0.904762	0.904762	1.000000	0.857143	0.904762	0.714286	0.809524	0.857143	0.809524
5	0.666667	0.809524	0.714286	0.761905	0.714286	0.619048	0.666667	0.619048	0.666667	0.761905	...	0.761905	0.809524	0.857143	0.904762	0.857143	0.857143	0.857143	0.904762	0.857143	0.904762
6	0.761905	0.761905	0.523810	0.714286	0.619048	0.571429	0.809524	0.857143	0.952381	0.857143	...	0.904762	0.666667	0.714286	0.904762	0.952381	0.761905	0.809524	0.952381	0.714286	0.904762
7	0.714286	0.761905	0.809524	0.619048	0.809524	0.571429	0.857143	0.714286	0.619048	0.666667	...	0.714286	0.904762	0.809524	0.904762	0.857143	0.904762	0.809524	0.714286	0.857143	0.809524
8	0.800000	0.750000	0.800000	0.700000	0.700000	0.800000	0.650000	0.900000	0.800000	0.850000	...	0.750000	0.850000	0.750000	0.900000	0.850000	0.900000	0.850000	0.750000	0.800000	0.900000
9	0.650000	0.600000	0.800000	0.650000	0.850000	0.700000	0.850000	0.700000	0.750000	0.750000	...	0.950000	0.900000	1.000000	0.900000	0.950000	0.850000	0.950000	0.750000	0.900000	0.800000

10 rows × 48 columns

Resultados de la Práctica Inicial




```
El k óptimo es:  40   con una tasa de aciertos de:  0.8890476190476191
```

Dendograma. Dataset "Semillas". WARD