Inicialización de k-means - Clustering

Inteligencia Computacional

Cipolatti Edgardo, Rosales Mario, Santellán Franco

Director: Leandro Di Persia

Tabla de contenidos

Introducción

Introducción

- La idea es inicializar el k-Means con otros métodos más eficientes que elegir semillas al azar.

- Ejecutar los métodos propuestos en diferentes bases de datos y evaluar su desempeño según diferentes medidas e índices.

Métodos de inicialización de k-means

Métodos de inicialización de k-means

Los métodos de inicialización utilizados son:

- BallHall
- Etiquetado
- Forgy
- k-means++
- McQueen
- McRae

Métodos de inicialización de k-means

- **BallHall:** la primer semilla es el centro de masa de todo el dataset; posteriormente se seleccionan las restantes semillas dependiendo de una distancia d.

- **Etiquetado:** se seleccionan patrones según la etiqueta $\left[\frac{\alpha m}{k}\right]$.

- **Forgy:** forma clusters con patrones al azar y asigna sus medias como semillas.

Métodos de inicialización de k-means

- **K-means++:** se eligen como semillas patrones en relación a una probabilidad.
- **McQueen:** toma los primeros k patrones como semillas.
- **McRae:** se seleccionan k semillas al azar sin repetición.

Medidas e Índices de evaluación de Clusters

Medidas e Índices de evaluación de Clusters

Las medidas e índices utilizados son:

- Tiempo
- Iteraciones
- Inter-Cluster
- Intra-Cluster
- Intra/Inter
- Indice Davies-Boulding
- Indice Dunn

Medidas e Índices de evaluación de Clusters

- **Tiempo:** Suma de tiempo requerido en crear las semillas y ejecutar k-means.
- **Cantidad de Iteraciones:** Iteraciones realizadas por k-means.
- **Inter-cluster:** Indica qué tan dispersos están los clusters.
- **Intra-cluster:** Indica cuán compacto es un cluster.

Medidas e Índices de evaluación de Clusters

- **Intra/Inter:** Cociente entre los índices Intra e Inter Cluster.

- **Indice Davies–Bouldin:** Relación entre Intra e Inter-cluster.

$$\frac{1}{K}\sum_{i,j=1}^{K} max\left\{\frac{Intra(C_i) - Intra(C_j)}{Inter(z_i, z_j)}\right\}$$

- **Indice Dunn:** Relación entre diámetro (Δ) y distancias (δ).

$$min_{1\leq i,j\leq K; j\neq i}\left\{\frac{\delta(C_i, C_j)}{max_{1\leq k\leq K}\triangle(C_k)}\right\}$$

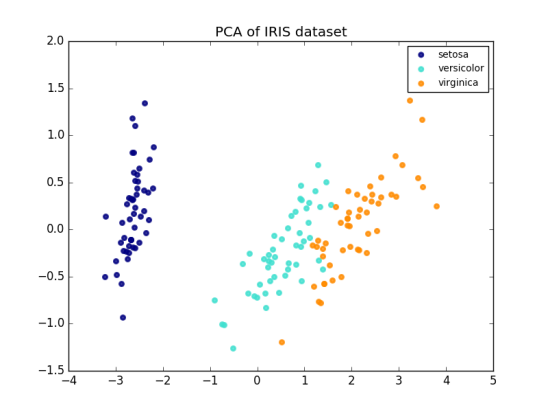
Bases de Datos

Bases de Datos

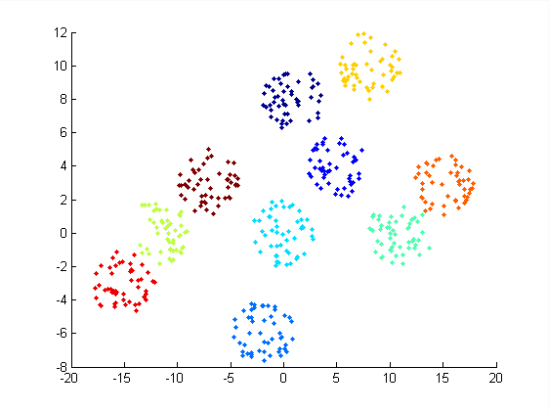
Características de las bases de datos utilizadas

Bases de datos	Dimensiones	Clases	Cantidad de Datos
Iris	4	3	150
Nubes	2	10	500
Glass Identification	10	6	214
Ionosphere	34	2	351
Doughnut	12	2	500
White Wine	11	7	500 - 4897

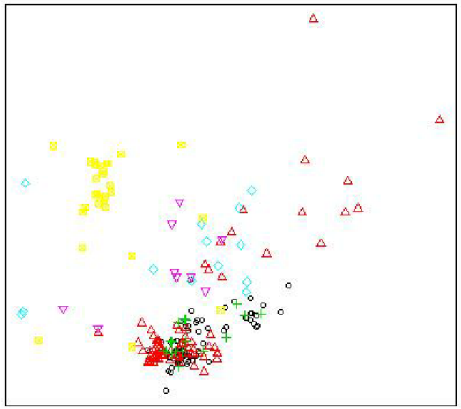
Bases de datos - Representación 2D



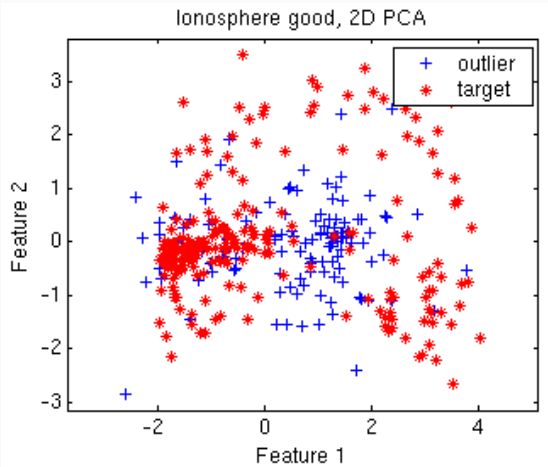
(a) Iris



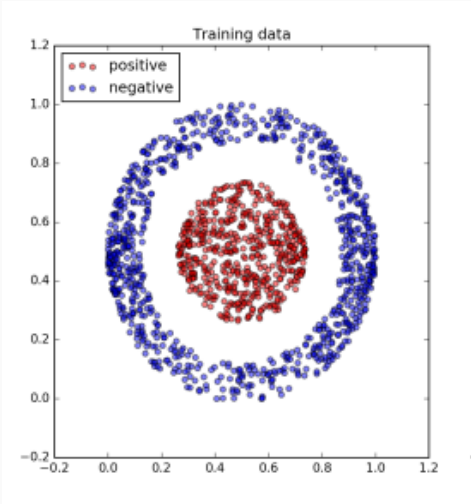
(b) Nubes



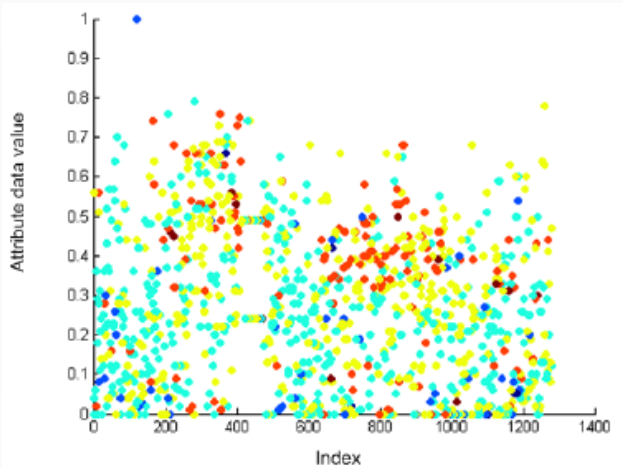
(c) Glass



(d) Ionosphere



(e) Doughnut

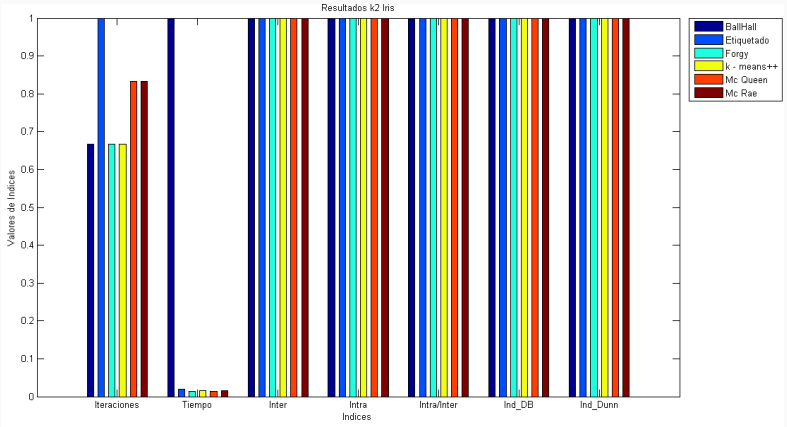


(f) White Wine

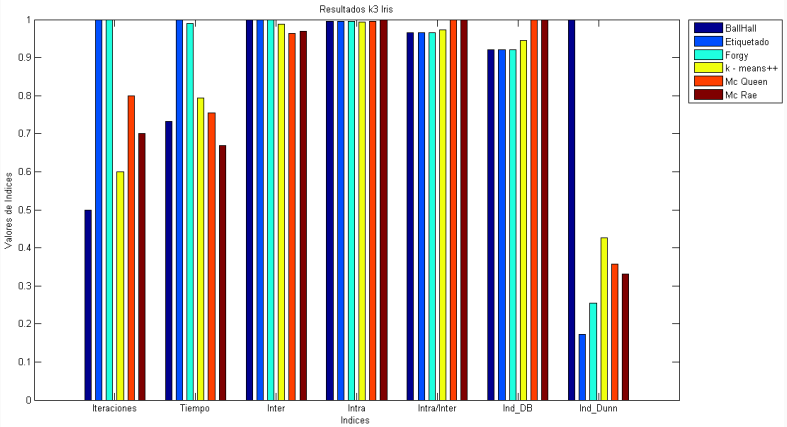
Resultados

Resultados - Iris

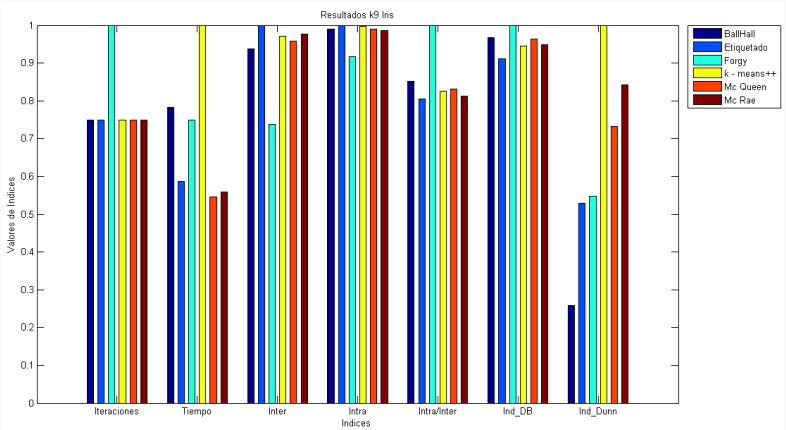
- **Iris**: 4 atributos, 3 clases y 150 elementos.



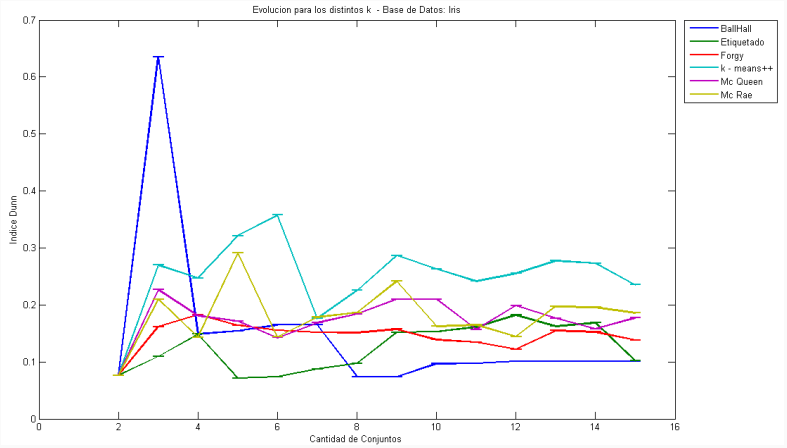
(g) Iris K=2



(h) Iris K=3



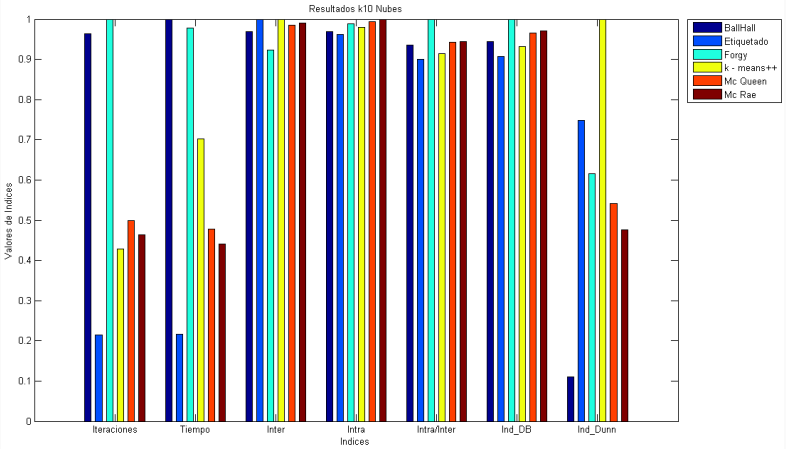
(i) Iris K=9



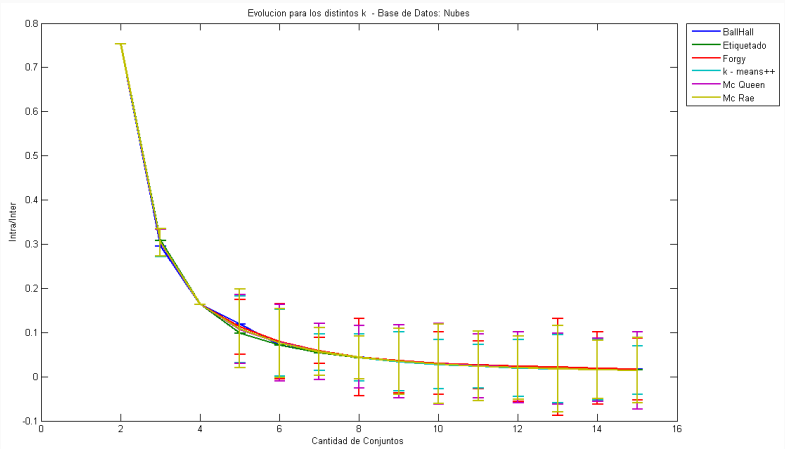
(j) Índice Dunn

Resultados - Nubes

- **Nubes-10**: 2 atributos, 10 clases y 500 elementos.



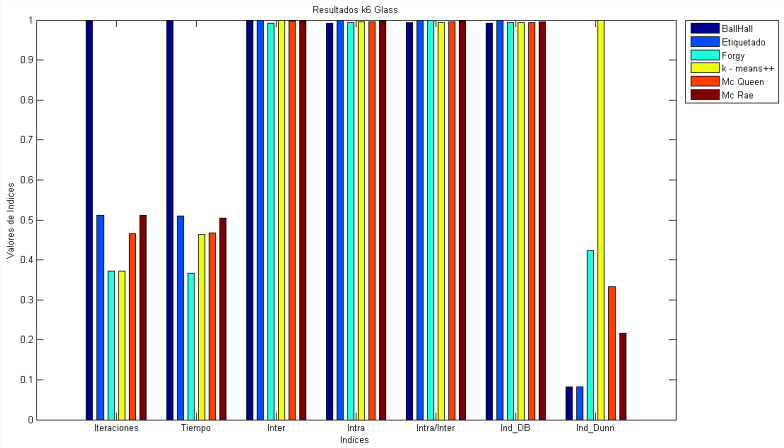
(k) Nubes K=10



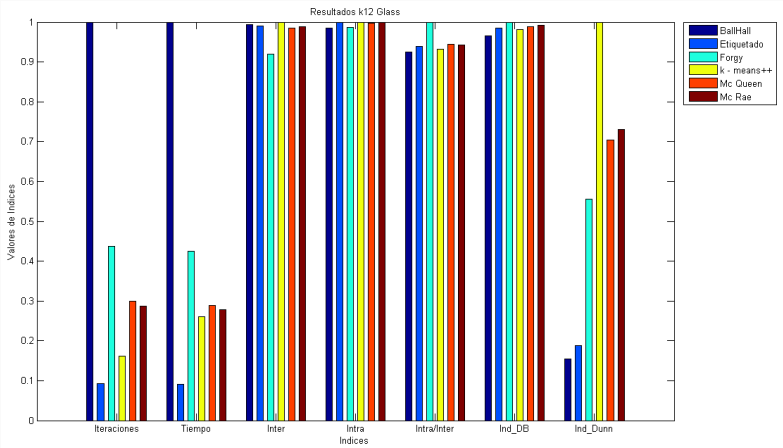
(l) Índice Intra/Inter

Resultados - Glass

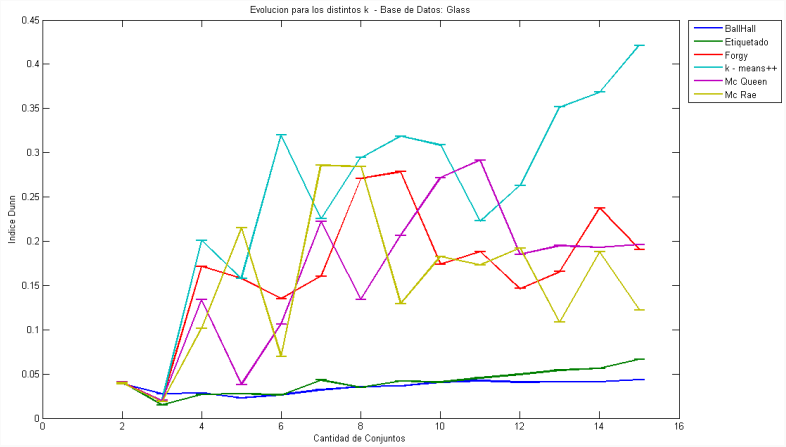
- **Glass**: 10 atributos, 6 clases y 214 elementos.



(m) Glass K=6



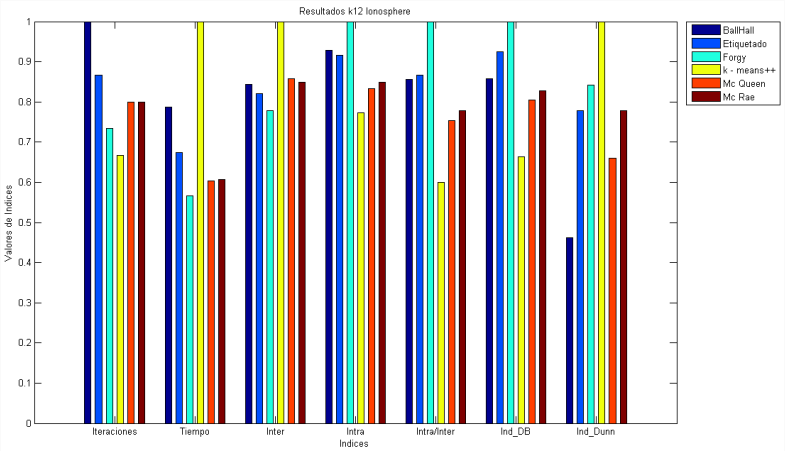
(n) Glass K=12



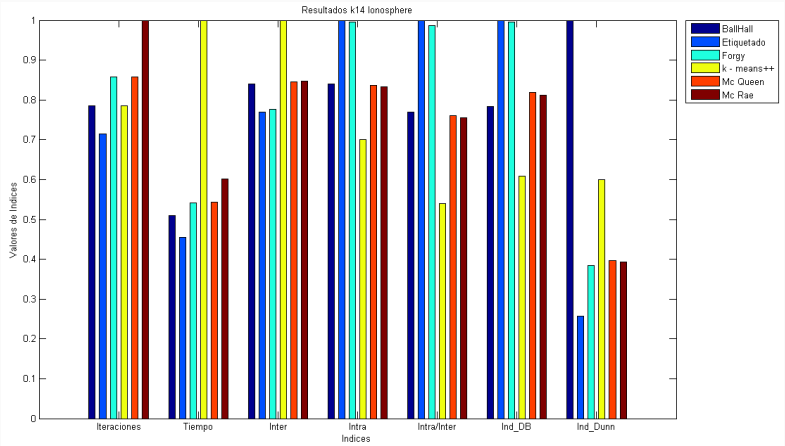
(ñ) Índice Dunn

Resultados - Ionosphere

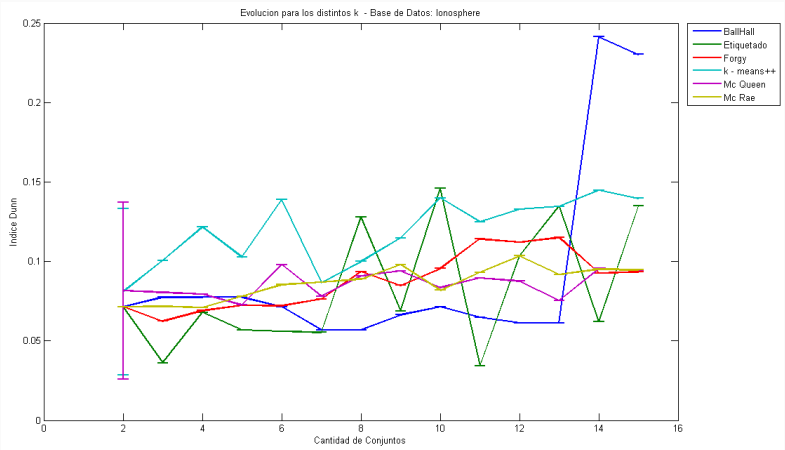
- **Ionosphere**: 34 atributos, 2 clases y 351 elementos.



(o) Ionosphere K=12



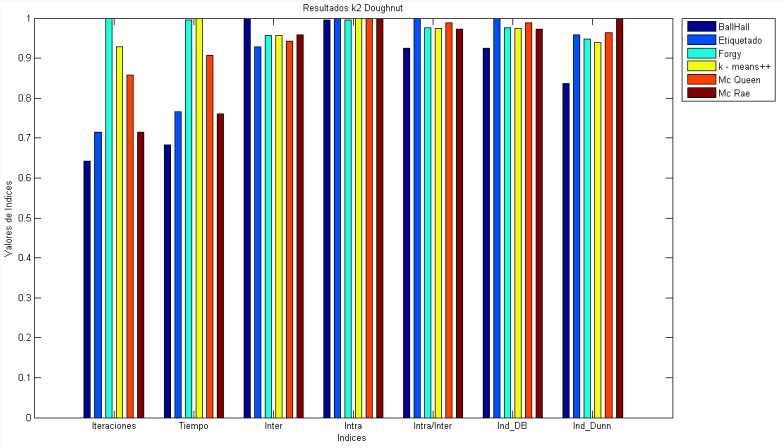
(p) Ionosphere K=14



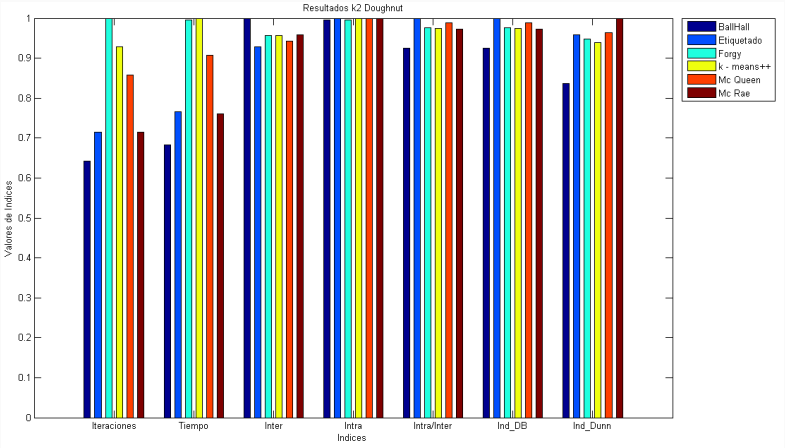
(q) Índice Dunn

Resultados - Doughnut

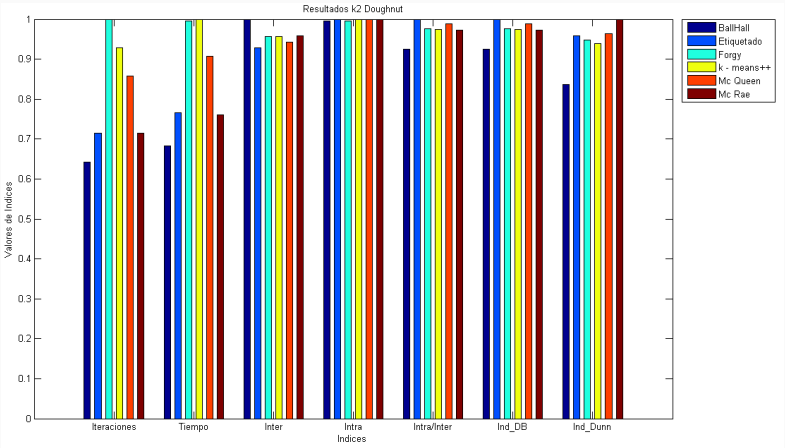
- **Doughnut**: 12 atributos, 2 clases y 500 elementos.



(r) Doughnut K=2



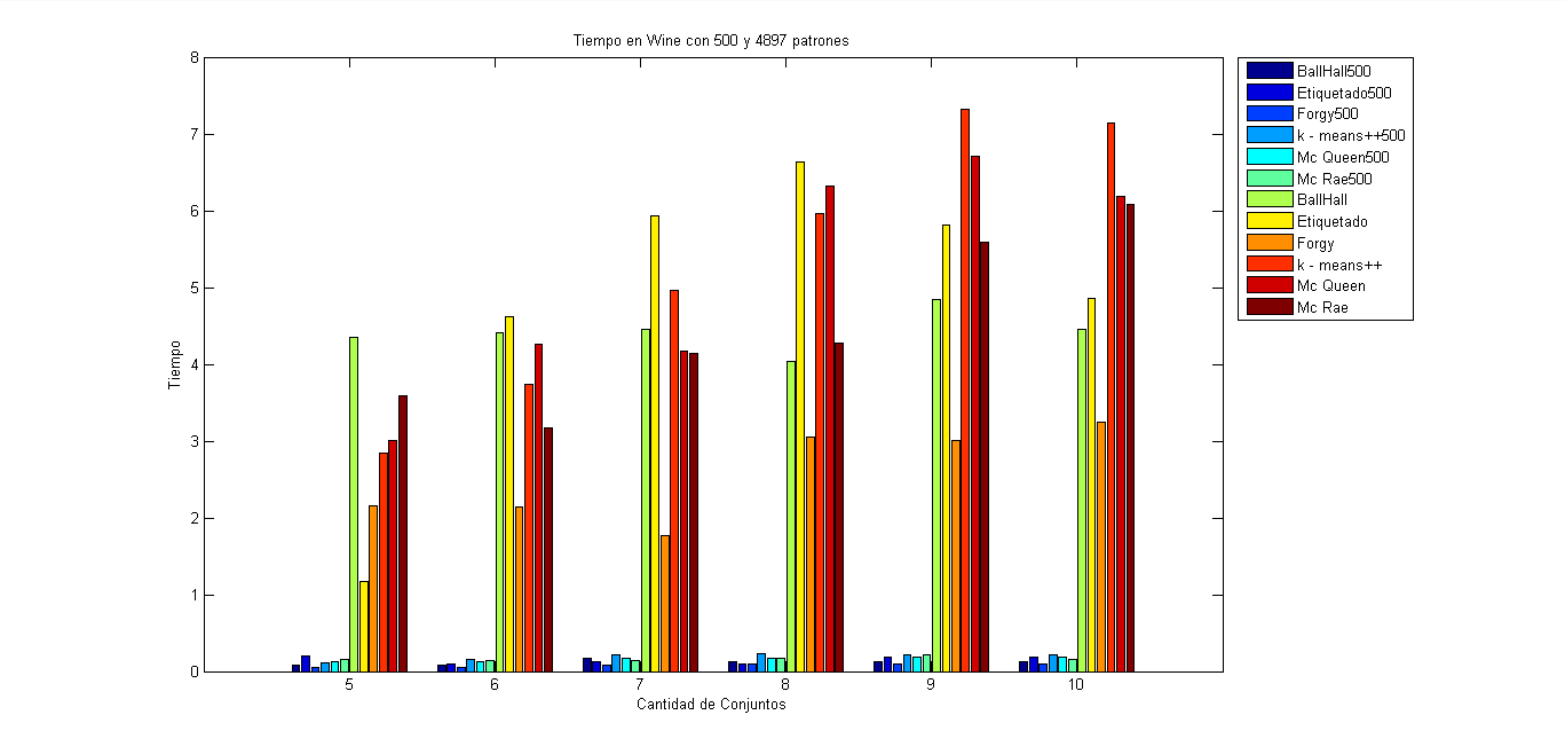
(s) Doughnut K=12



(t) Doughnut K=15

Resultados - White Wine

- **White Wine**: 11 atributos, 7 clases y 500 - 4897 elementos.



(u) Tiempo White Wine

¿Preguntas?