

Agrupamiento de imágenes

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Clustering 3

A dark blue diagonal gradient bar that starts from the bottom-left corner and extends towards the top-right corner, covering the lower half of the slide.

Métodos de validación: ¿Por qué es importante validar?

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
- Determinar el mejor número de clusters (k).
- Comparar métodos de Clustering (sobre los mismos datos).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
 - Antes de empezar a buscar es importante verificar si existe una **tendencia al clustering**.
- Determinar el mejor número de clusters (k).
- Comparar métodos de Clustering (sobre los mismos datos).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
 - Antes de empezar a buscar es importante verificar si existe una **tendencia al clustering**.
- Determinar el mejor número de clusters (k).
 - Determinar cuán buenos son los grupos (**externa/interna**).
 - Determinar cuán buena es la separación entre ellos (**externa/interna**).
 - Constatar el número de clusters con las etiquetas *a priori* (si existen) (**externa**).
- Comparar métodos de Clustering (sobre los mismos datos).
 - Determinar cuán buenos son los grupos (**externa/interna**).
 - Determinar cuán buena es la separación entre ellos (**externa/interna**).
 - Constatar la pertenencia a los grupos con las etiquetas *a priori* (si existen) (**externa**).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

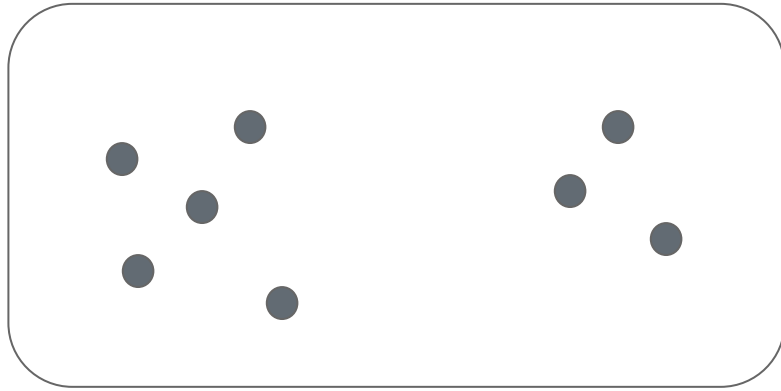
- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
 - Antes de empezar a buscar es importante verificar si existe una **tendencia al clustering**.
- Determinar el mejor número de clusters (k).
 - Determinar cuán buenos son los grupos (**externa/interna**).
 - Determinar cuán buena es la separación entre ellos (**externa/interna**).
 - Constatar el número de clusters con las etiquetas *a priori* (si existen) (**externa**).
- Comparar métodos de Clustering (sobre los mismos datos).
 - Determinar cuán buenos son los grupos (**externa/interna**).
 - Determinar cuán buena es la separación entre ellos (**externa/interna**).
 - Constatar la pertenencia a los grupos con las etiquetas *a priori* (si existen) (**externa**).

Existen diferentes medidas de validación y no existe un criterio único para determinar cuál es la mejor.

No hay una medida única que se pueda usar para todos los métodos de clustering.

Tendencia al Clustering

Estadístico (coeficiente) de Hopkins



● *elementos del dataset*

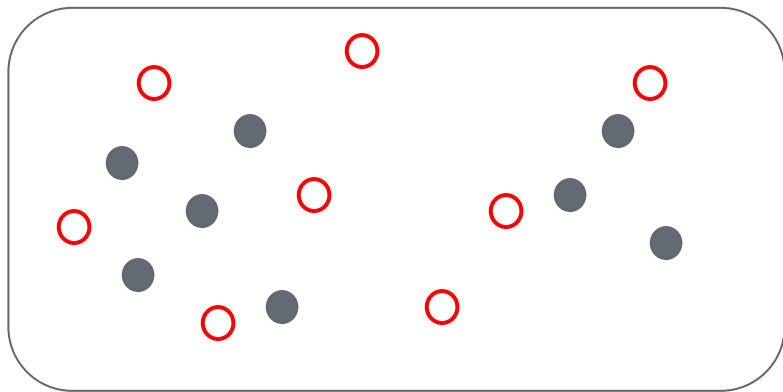
Tendencia al Clustering

Estadístico (coeficiente) de Hopkins

$$H = \frac{\sum_{i=1}^p w_i}{\sum_{i=1}^p u_i + \sum_{i=1}^p w_i}$$

w_i = distancia de un elemento i al azar al vecino más cercano

u_i = distancia de un punto i agregado al azar al vecino más cercano



● elementos del dataset

○ elementos agregados al azar

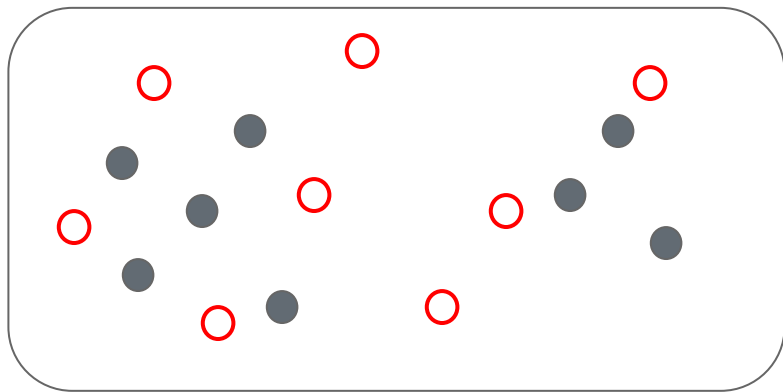
Tendencia al Clustering

Estadístico (coeficiente) de Hopkins

$$H = \frac{\sum_{i=1}^p w_i}{\sum_{i=1}^p u_i + \sum_{i=1}^p w_i}$$

w_i = distancia de un elemento i al azar al vecino más cercano

u_i = distancia de un punto i agregado al azar al vecino más cercano



● elementos del dataset

○ elementos agregados al azar

$H \sim 0$: Clusters!

$H \sim 0.5$: Distrib. homogénea



Validación externa

1. Matriz de confusión

Clusters	0	1	2	3	4
Labels					
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

Wu, J., Xiong, H., & Chen, J. (2009, June). Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 877-886).

<https://maths-people.anu.edu.au/~johnm/courses/mathdm/2009/talks/justin-paper.pdf>

Validación externa

2. Medida Normalizada de van Dongen: Medida mejorada de la pureza (que mide cuánto se aleja de tener sólo tengo un elemento por fila/columna)

Clusters	0	1	2	3	4
Labels					
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

$$VD_n = \frac{(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})}{(2n - \max_i n_{i.} - \max_j n_{.j})}$$

Van Dongen, S. (2000). Performance criteria for graph clustering and Markov cluster experiments. In *National research institute for mathematics and computer science*. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.9783&rep=rep1&type=pdf>

Validación externa

2. Medida Normalizada de van Dongen: Medida mejorada de la pureza (que mide cuánto se aleja de tener sólo tengo un elemento por fila/columna)

Clusters	0	1	2	3	4
Labels					
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

$$2 \cdot 200 - (25 + 24 + 34 + 39 + 40) - (25 + 24 + 34 + 39 + 40)$$

$$VD_n = \frac{(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})}{(2n - \max_i n_{i.} - \max_j n_{.j})}$$

$$2 \cdot 200 - 40 - 40$$

Van Dongen, S. (2000). Performance criteria for graph clustering and Markov cluster experiments. In *National research institute for mathematics and computer science*. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.9783&rep=rep1&type=pdf>

Validación externa

2. Medida Normalizada de van Dongen: Medida mejorada de la pureza (que mide cuánto se aleja de tener sólo tengo un elemento por fila/columna)

Clusters	0	1	2	3	4
Labels					
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

$$VD_n = \frac{(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})}{(2n - \max_i n_{i.} - \max_j n_{.j})}$$

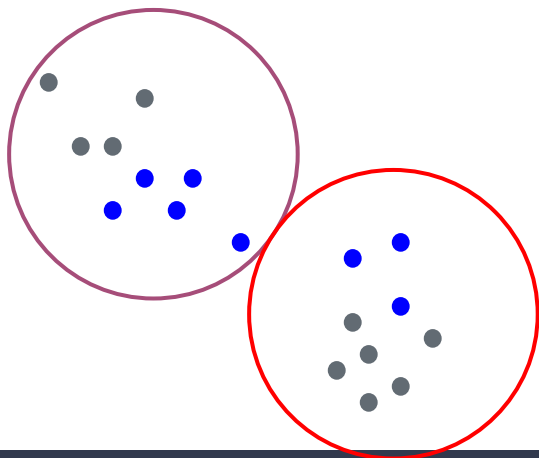
(mejor) $0 \leq VD_n \leq 1$ (peor)

Van Dongen, S. (2000). Performance criteria for graph clustering and Markov cluster experiments. In *National research institute for mathematics and computer science*. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.9783&rep=rep1&type=pdf>

Validación externa

3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI):

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$



a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.

b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.

c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.

d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.

n = número total de elementos.

Validación externa

3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI):

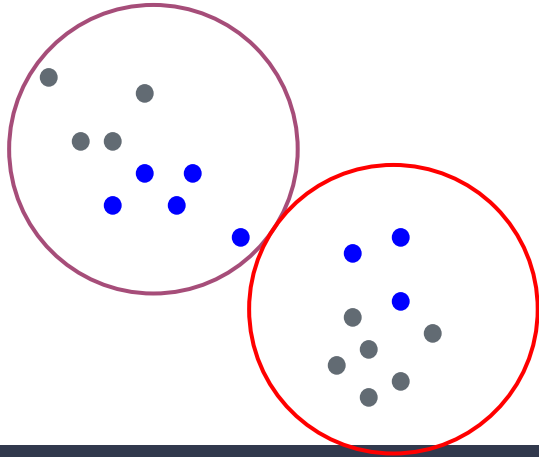
$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

$$a = 6 + 10 + 3 + 15 = 34$$

$$b = 30 + 12 = 42$$

$$n = 153$$

$$R = 76 / 153 \sim 0.5$$



Validación externa

3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI):

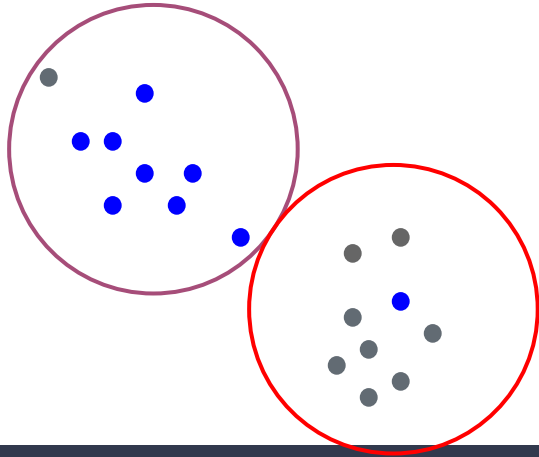
$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

$$a = 28 + 28 = 56$$

$$b = 64 + 1 = 65$$

$$n = 153$$

$$R = 111 / 153 \sim 0.73$$



Validación externa

3. Índice Rand e Índice Rand Ajustado o Normalizado (*Adjusted Rand Index, ARI*):

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.

b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.

c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.

d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.

n = número total de elementos.

Validación externa

3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI):

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

$$ARI = (R - E(R)) / (\max(R) - E(R))$$

$E(R)$ = valor esperado de R si se distribuyen al azar.

$\max(R)$ = valor máximo posible de R para los datos.



Validación Interna: Criterios general: Cohesión y Separación

- **Cohesión:** es una medida de las proximidad de los miembros de un clúster entre sí o con respecto al prototipo.
- **Separación:** es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

**Suma de los Errores al Cuadrado
(dentro de un cluster) (SSE)**

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \text{dist}(c_i, x)^2$$

**Suma de Cuadrados de Separación
(SSB)**

$$SSB = \frac{1}{2K} \sum_{i=1}^K \sum_{j=1}^K \frac{m}{K} \text{dist}(c_i, c_j)^2$$

c_i = centroide o medoide (prototipo)

Criterios general: Cohesión y Separación

- **Cohesión:** es una medida de las proximidades de los miembros de un clúster con respecto al prototipo.
- **Separación:** es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

c_i = centroide o medoide (prototipo)

**Suma Total de los Errores
al Cuadrado (TSE)**

$$TSE = SSE + SSB$$

Validación Interna

1. Coeficiente de Silhoutte:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

1. Para cada elemento i se calcula su distancia promedio a todos los otros elemento de su clúster (a_i).
2. Para el elemento i y todos los otros clústeres que no lo contienen, se calcula las distancias promedio a todos los elementos de cada clúster.

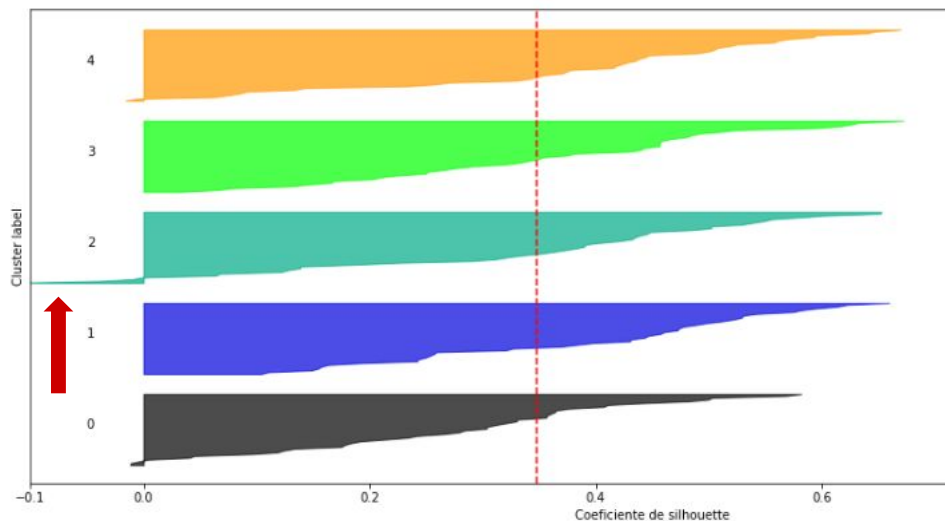
Se buscar el mínimo de esas distancias promedio a cada clúster (b_i).

3. Se calcula el coeficiente Silhouette (s_i) del elemento i .
4. Luego se puede calcular el promedio para cada cluster o el promedio global.

Validación Interna

1. Coeficiente de Silhoutte:

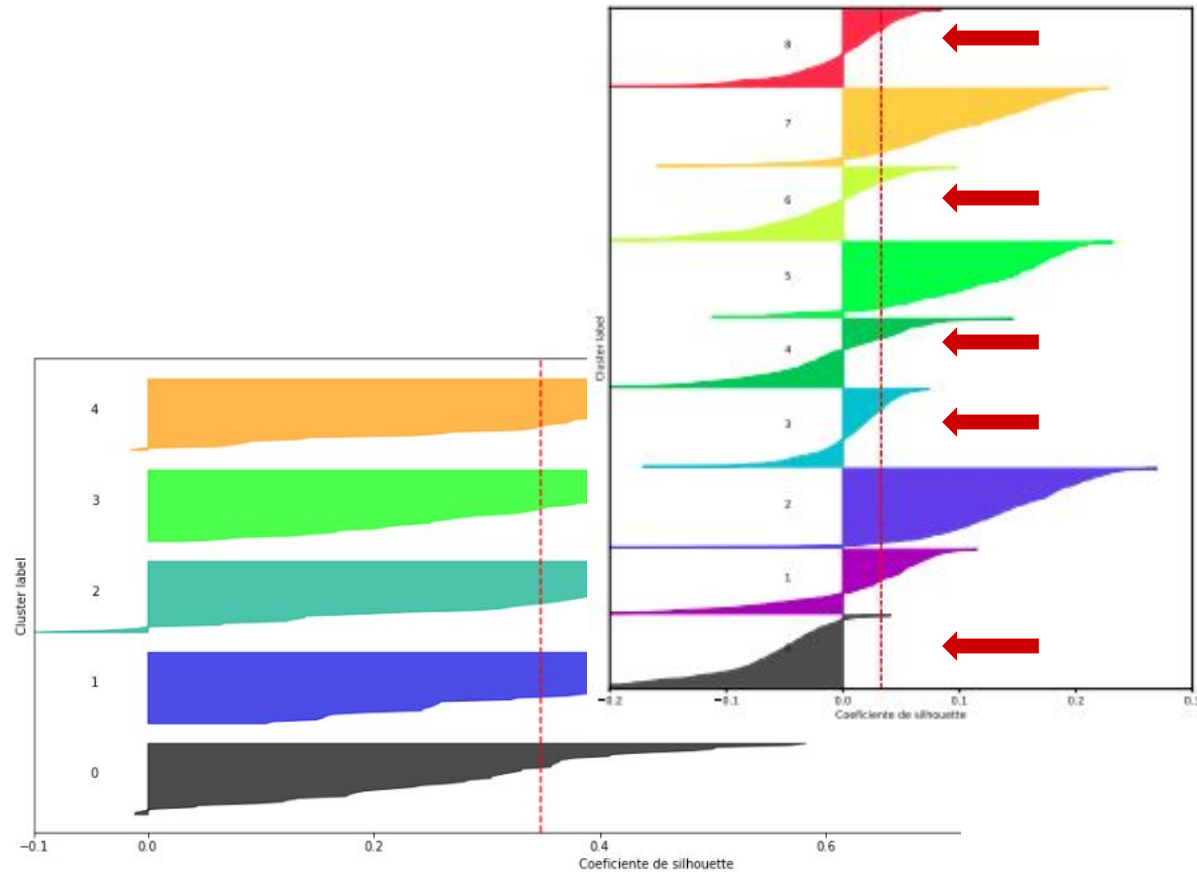
Puede ocurrir que algunos clusters tengan peor coeficiente o algunos elementos dentro del cluster. Esto puede ser indicativo de que quizás es mejor cambiar el valor de k .



Validación Interna

1. Coeficiente de Silhoutte:

Puede ocurrir que algunos clusters tengan peor coeficiente o algunos elementos dentro del cluster. Esto puede ser indicativo de que quizás es mejor cambiar el valor de k .

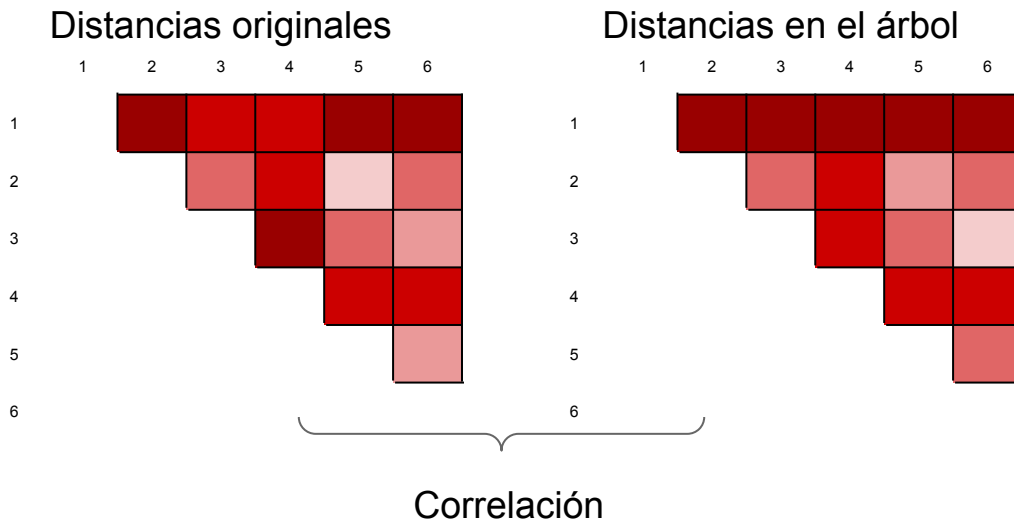
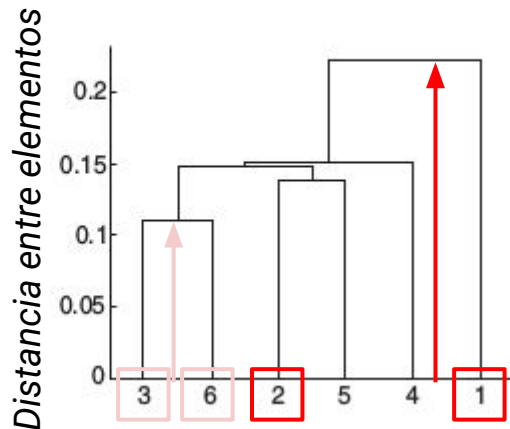


Validación Interna

2. (Jerárquico) Partición del árbol por distancia o por número de clusters: y luego se pueden aplicar las medidas de validación interna o externa igual que otros métodos.

Validación Interna

3. (*Jerárquico*) **Coeficiente de Correlación Cofenético (CoPhenetic Correlation Coefficient, CPCC)**: Mide la correlación entre la matriz de distancia que dio origen al agrupamiento y los distancias extraídas del árbol (Altura del nodo que une por primera vez dos elementos).



Validación Interna

4. *Bootstrapping*: Sirve para evaluar la estabilidad de los clusters, y así determinar cuáles son “reales” y cuáles no. Vamos a volver a estos métodos más adelante.

