Reconocimiento de patrones

Tarea 6

Clustering 3

Alumno:

Pérez Rodríguez Raúl Francisco

Octubre 2017

Realice la identificación de clases por k-means de los datos de los sobrevivientes del Titanic.

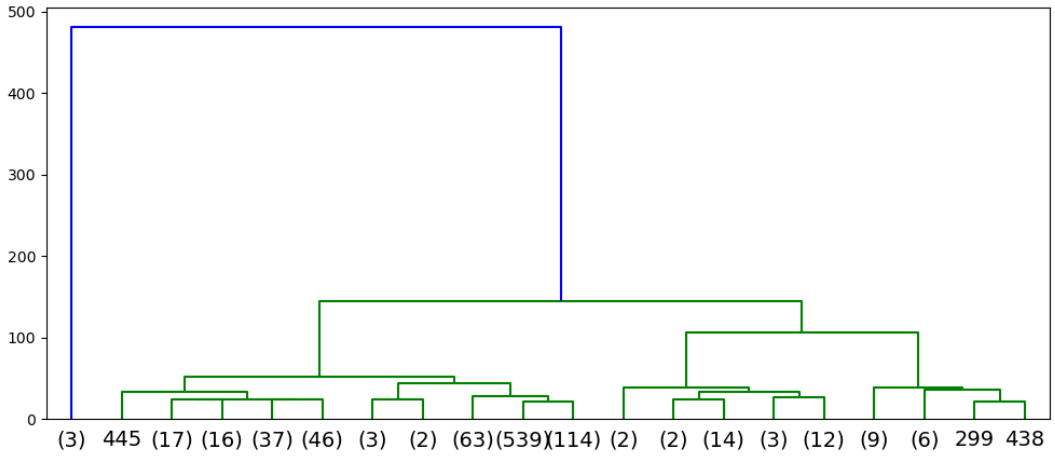
Dendrogramas

El conjunto de datos iniciales del Titanic, contienen 10 variables con 891 instancias. Eliminando las variables no numéricas, quedan las variables Survived ,Pclass, Age, SibSp, Parch, Fare.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Utilizando la técnica de dendrograma con el método de centroid y truncando para obtener los últimos 20 clusters obtenemos la siguiente gráfica.



Analizando los últimos 20 clusters del dendrograma se pueden observar que puede haber entre 3 y 5 grupos.

K-Means

Utilizando el dendrograma como una primera prueba, se observaron la formación de entre 3 y 5 grupos de datos. Con estos resultados se utilizara el k-means con k igual a 3, 4 y 5 para observar la formación de los grupos con una selección aleatoria de los clusters iniciales.

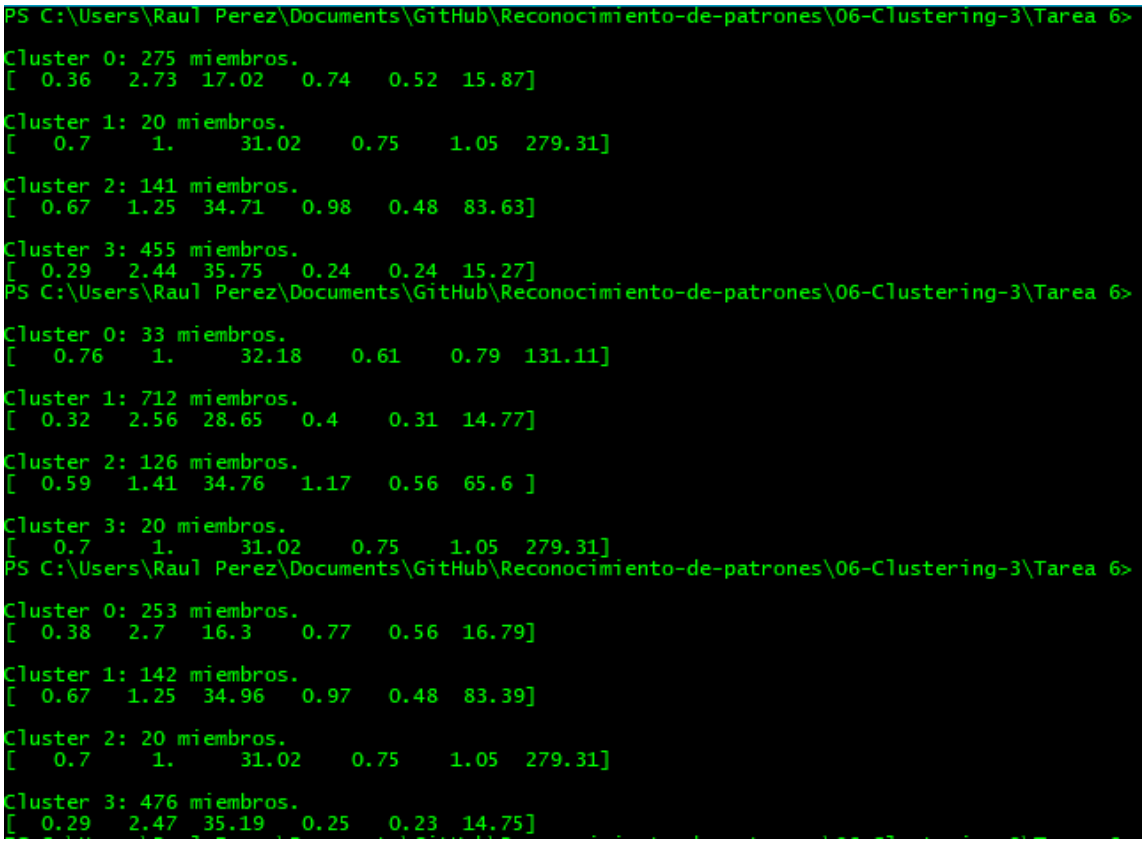
Con k igual a 3

Realizando 3 corridas con el k-means con k igual a 3, se obtuvieron los mismos clusters

```
PS C:\Users\Raul Perez\Documents\GitHub\Reconocimiento-de-patrones\06-Clustering-3\Tarea 6> python .\main.py

Cluster 0: 142 miembros.
[  0.67   1.25  34.96   0.97   0.48  83.39]

Cluster 1: 729 miembros.
[  0.32   2.55  28.64   0.43   0.34  15.45]

Cluster 2: 20 miembros.
[   0.7    1.    31.02   0.75   1.05 279.31]
PS C:\Users\Raul Perez\Documents\GitHub\Reconocimiento-de-patrones\06-Clustering-3\Tarea 6> python .\main.py

Cluster 0: 142 miembros.
[  0.67   1.25  34.96   0.97   0.48  83.39]

Cluster 1: 729 miembros.
[  0.32   2.55  28.64   0.43   0.34  15.45]

Cluster 2: 20 miembros.
[   0.7    1.    31.02   0.75   1.05 279.31]
PS C:\Users\Raul Perez\Documents\GitHub\Reconocimiento-de-patrones\06-Clustering-3\Tarea 6> python .\main.py

Cluster 0: 20 miembros.
[   0.7    1.    31.02   0.75   1.05 279.31]

Cluster 1: 142 miembros.
[  0.67   1.25  34.96   0.97   0.48  83.39]

Cluster 2: 729 miembros.
[  0.32   2.55  28.64   0.43   0.34  15.45]
```

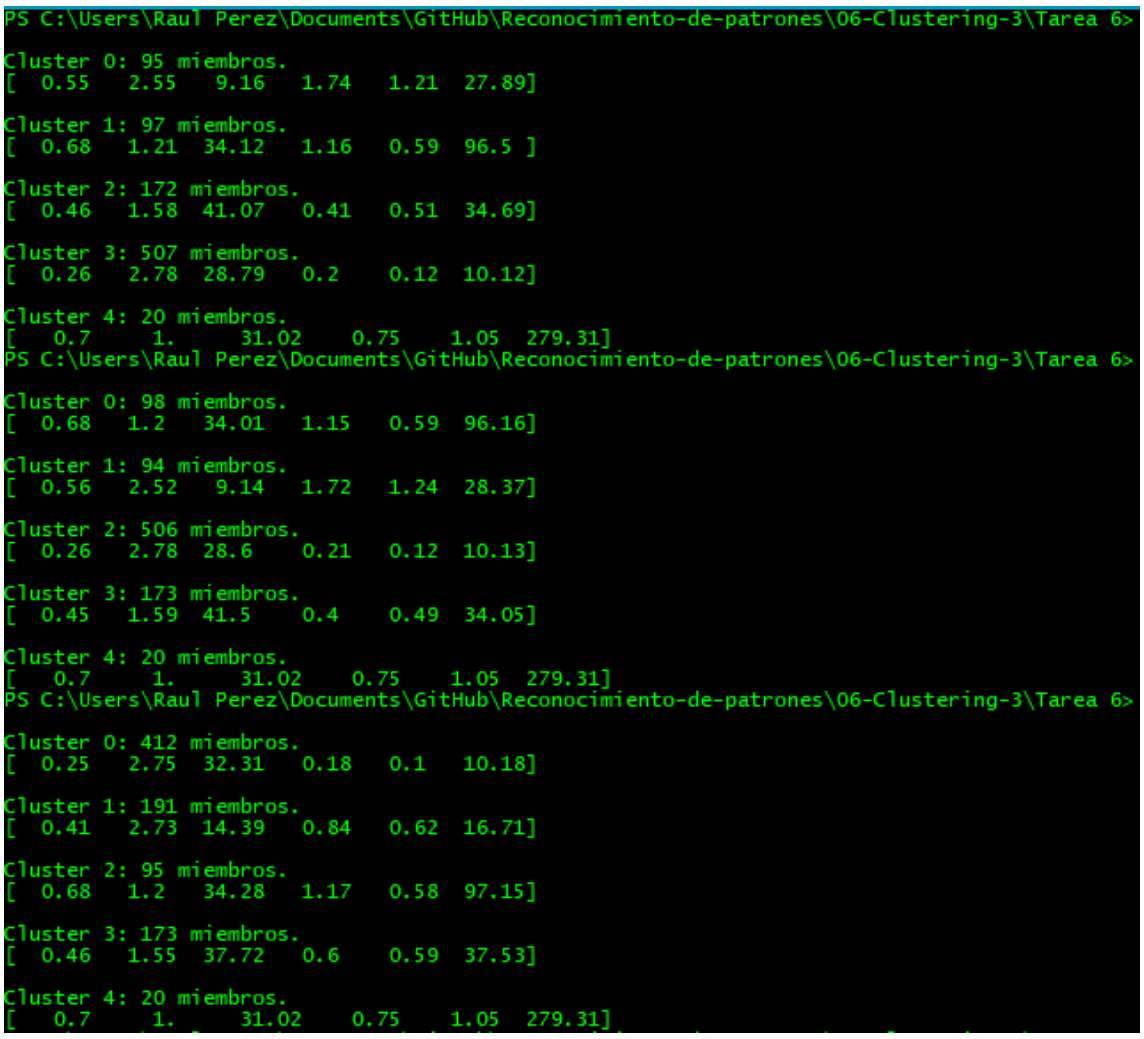
Con k igual a 4

Realizando 3 corridas con el k-means con k igual a 4, se obtuvieron clusters diferentes excepto el cluster con 20 elementos, el cual se repite con k igual a 3.



Con k igual a 5

Realizando 3 corridas con el k-means con k igual a 5, se obtuvieron clusters diferentes excepto el cluster con 20 elementos, el cual se repite con k igual a 3. Pero a comparación con k igual a 4, en este caso el número de miembros por cluster no es tan diferente.



Conclusión

Analizando los resultados al utilizar la técnica de k-means con los valores de k iguales a 3, 4 y 5. Se concluye que es más probable que haya 3 grupos de datos en el conjunto de datos.