Práctica 2: Tipificación de Usuarios

Social Web Behaviour

Alumnos: *Sergi Albiach Caro, Daniel Constantín Birdici & Stéphane Díaz-Alejo León*

ÍNDICE

1.- INTRODUCCIÓN

En esta práctica se nos requería descargar información de los usuarios relacionados con un *hashtag*, palabra clave o seguidores de una cuenta para después realizar un proceso de análisis de la propia información obtenida. Para ello, nuestro equipo decidió utilizar el primer método, con lo cual, descargamos la información mediante Twitter Archiver utilizando la siguiente regla:

"#PlayStation OR #PS4 OR #PS5 OR #PSVR OR #PS4share OR #PS5share"

Esta regla simplemente busca aquellos tweets en los que aparezca uno de los *hashtags* mencionados. Elegimos la misma debido a que el tema de nuestro proyecto serán los usuarios y/o fans de PlayStation, la serie de videoconsolas (PlayStation 4 y PlayStation 5 en particular) que creó Sony.

2.- ANÁLISIS DE DATOS

En este apartado decidimos agrupar la información recogida utilizando una técnica no-supervisada llamada *K-means clustering*. Se trata de un algoritmo que, dado un número de grupos o clases, intenta minimizar la suma de cuadrados (la diferencia entre un punto y el promedio de la clase al cuadrado) dentro del mismo grupo. El vector de características utilizado en este algoritmo consta de cinco campos:

- *Retweets*: las veces que los tuits de este usuario se han vuelto a publicar por otros usuarios.
- *Favorites*: las veces que los tuits de este usuario han sido añadidos a favoritos.
- *Followers*: los seguidores que tiene la cuenta.
- *Follows*: el número de cuentas seguidas por el usuario.
- *Listed*: el número de listas personalizadas a las que la cuenta de este usuario fue añadido.

En el informe proporcionado por XLSTAT, podemos observar la siguiente información sobre los datos proporcionados:

Variable	Observaciones	Obs. con datos perdidos	Obs. sin datos perdidos	Mínimo	Máximo	Media	Desv. típica
Retweets	55665	0	55665	0.000	579.000	0.130	3.480
Favorites	55665	0	55665	0.000	991.000	0.811	7.167
Followers	55665	0	55665	0.000	2754074.000	1246.328	20302.640
Follows	55665	0	55665	0.000	180379.000	453.575	1702.642
Listed	55665	0	55665	0.000	26332.000	12.947	186.337

Esta información nos puede dar una idea de la distribución de población de cada característica. Por ejemplo, la población de la variable followers a pesar de tener una cantidad máxima de 2.754.074 debe de tener una enorme cantidad de muestras con valores muy bajos, debido a que la media es tan solo de 1.246,328, tres órdenes de magnitud inferior, y unas pocas muestras con valores muy altos, cómo nos muestra la desviación típica. Por norma general, se puede observar que las poblaciones presentan unas pocas muestras con valores muy altos y el resto bajos, ya que las medias tienden a ser bajas y la desviación típica un orden de magnitud o varios por encima.

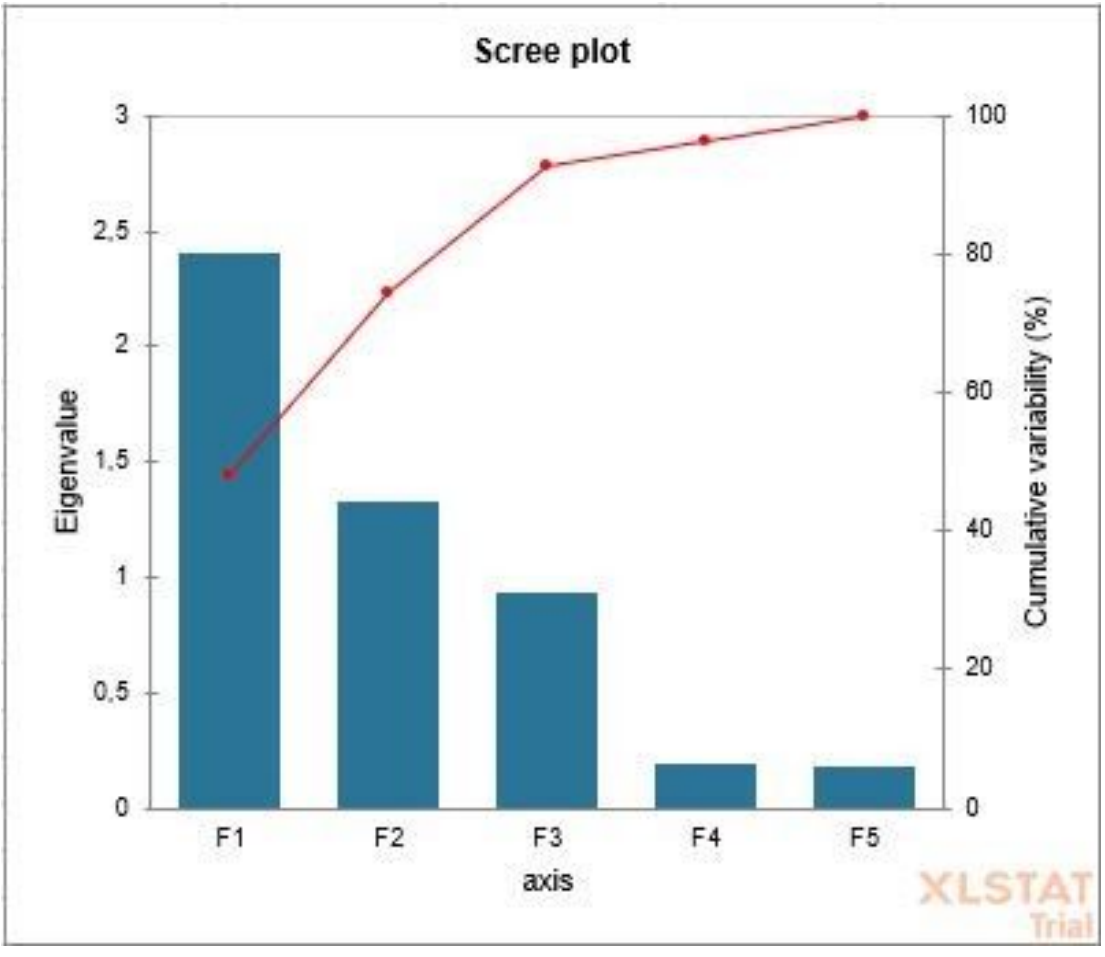
Para analizar la naturaleza de los datos hemos realizado un test de correlación entre las variables para ver si existen dependencias entre ellas.

Correlation matrix (Pearson):

Variables	Retweets	Favorites	Followers	Follows	Listed
Retweets	**1**	**0,822**	**0,309**	**0,050**	**0,237**
Favorites	**0,822**	**1**	**0,307**	**0,054**	**0,226**
Followers	**0,309**	**0,307**	**1**	**0,154**	**0,819**
Follows	**0,050**	**0,054**	**0,154**	**1**	**0,179**
Listed	**0,237**	**0,226**	**0,819**	**0,179**	**1**

Values in bold are different from 0 with a significance level alpha=0,05

Los valores obtenidos siguen lo que la lógica nos puede decir en un primer momento, existe una fuerte relación entre *retweets* y *favorites*, ya que a más *retweets*, el mensaje llega más lejos y es más probable que la gente le de favorito. Por otra parte existe una fuerte relación entre los seguidores que tienes y la pertenencia a listas, lo cual es lógico ya que los perfiles con más seguidores suelen tratarse de medios de comunicación, divulgación o perfiles verificados de la propia PlayStation, por lo que tienden a estar en listas. Existe también una mínima relación entre los *retweets* y *favorites* y los *followers*, también una correlación lógica y esperable, ya que a más followers más probabilidad de que den retuit y favorito.



Contribution of the variables (%):					
	F1	F2	F3	F4	F5
Retweets	23,849	25,256	0,597	30,266	20,032
Favorites	23,598	25,691	0,745	24,914	25,053
Followers	26,666	16,691	5,406	23,292	27,945
Follows	2,416	9,051	88,484	0,012	0,038
Listed	23,471	23,312	4,768	21,516	26,933

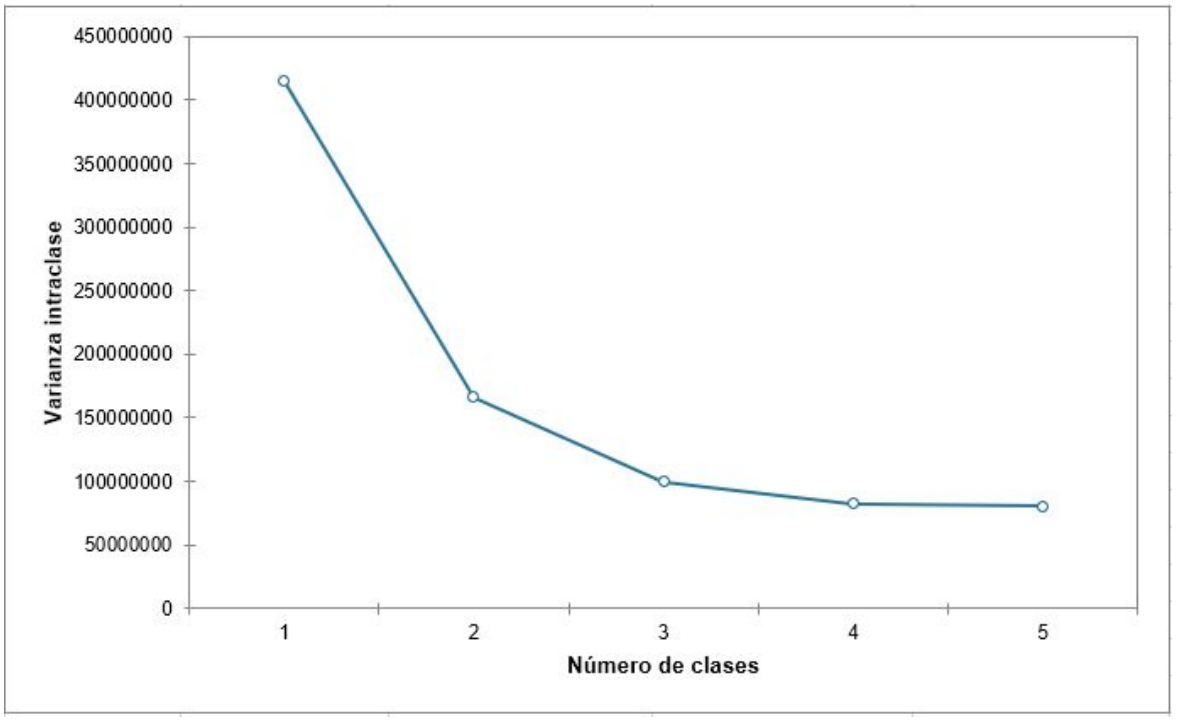
Con el objetivo de ver qué variables son más significativas, hemos realizado un análisis PCA, en el que intentamos eliminar variables del conjunto original de modo que obtengamos un conjunto más reducido pero manteniendo la mayor parte de la información del grupo de variables original.

Como podemos ver en la gráfica resultante de este análisis, con el primer factor podemos explicar hasta el 50% de la variabilidad de los datos. Este factor está compuesto mayoritariamente por las variables *retweets*, *favorites*, *followers* y *listed*, como nos indica la tabla.

Añadiendo el segundo factor, compuesto por las variables *retweets*, *favorites* y *listed*, podemos llegar hasta un 75% de variabilidad más o menos. De este modo, podemos comprobar como la variable *follows*, que representa el número de cuentas seguidas por un usuario concreto no es muy representativa. Siguiendo una lógica intuitiva, esto tiene sentido ya que el número de cuentas que un usuario sigue no impacta la popularidad de su cuenta; es decir, no tiene gran impacto en el número de *retweets*, seguidores o favoritos que pueda obtener.

3.- CLUSTERING

En los resultados proporcionados por *K-means clustering* se nos presenta la siguiente gráfica con la varianza intraclase en base al número de clusters:

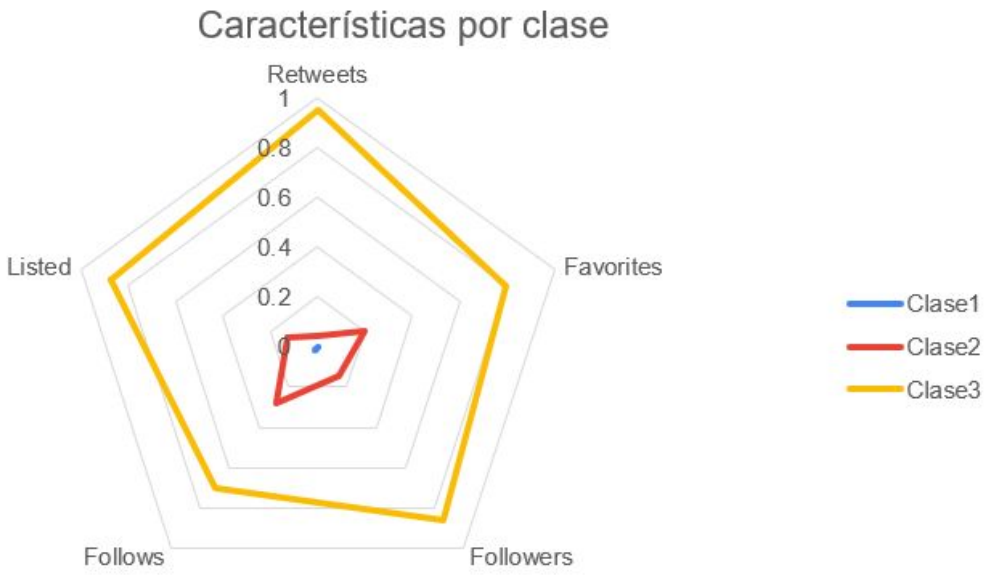


Siguiendo la “regla del codo”, podemos concluir que con estos datos la mejor agrupación serían tres clusters, ya que es el punto en el que la disminución de la varianza ya no es significativa; es decir, no vamos a conseguir una clasificación significativamente mejor utilizando más grupos.

En la siguiente tabla podemos ver los resultados de los tres clusters o clases:

Clase	1	2	3
Objetos	55564	92	9
Suma de los pesos	55564	92	9
Varianza intraclase	18140491.237	11716557031.673	434484813253.139
Distancia mínima al centroide	10.164	5482.366	164646.114
Distancia media al centroide	1242.312	82778.183	487842.228
Distancia máxima al centroide	119417.653	446397.196	1516077.368

Se puede apreciar una reducción significativa de muestras conforme cambiamos de clase, siendo la primera clase la que más muestras posee y la tercera la que menos. Para poder razonar a qué se debe esto hemos realizado la siguiente gráfica en el que se puede observar los valores normalizados del centroide de cada clase:



El grupo 1 es el usuario mayoritario, cuentas con pocos *retweets*, *favorites*, *followers*, *follows* y *listed*, en comparación con cuentas más grandes. Los usuarios de la clase 2 son usuarios medianos con una contribución moderada sobre el total, tratándose principalmente de cuentas de empresas que se dedican al mundo de los videojuegos, como *Best Buy Canada*. Por último, la clase 3 comprende las cuentas con un mayor número de seguidores, seguidos, *retweets* y favoritos, es decir, sobre todo, las cuentas oficiales de *PlayStation*, y algunas revistas como *CNET* o *famitsu*.

La siguiente tabla muestra los centroides de cada clase normalizados por característica:

Clase	Retweets	Favorites	Followers	Follows	Listed
1	0.00102653	0.00411926	0.00049287	0.01955169	0.0007631
2	0.04449321	0.19956296	0.14116623	0.28081515	0.12373949
3	0.95448026	0.79631778	0.8583409	0.69963316	0.87549741

La característica principal de la clase 1 son los follows. Los usuarios de este grupo siguen a muchas cuentas en comparación con sus seguidores. Puesto que no son cuentas populares, tampoco tienen demasiados *retweets*, favoritos ni aparecen en listas personalizadas.

La característica representativa del grupo 2 son los follows también, seguido de los favoritos y followers. Estas cuentas suelen ser de tiendas de videojuegos, revistas y otros tipos de empresas que tienen más presencia en la red social, por lo que tiene sentido que tengan más seguidores y más favoritos.

La característica principal del grupo 3 son los *retweets* seguido de los followers. Un dato que está en concordancia con lo dicho anteriormente, ya que se trata de cuentas de grandes empresas con una gran cantidad de seguidores que proporcionan muchos *retweets*.

Finalmente, un dato insólito es que el grupo 2 y 3, a pesar de estar compuestos por cuentas grandes, también cuentan con la mayor cantidad de *follows* o usuarios seguidos. Si analizamos las cuentas pertenecientes a estos grupos, podremos observar que algunas de ellas disponen de bots que siguen automáticamente a una gran cantidad de cuentas relacionadas con el sector de los videojuegos. Como ejemplo, solo la revista japonesa *famitsu* sigue a alrededor de 45 mil otras cuentas.