UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 2: Agrupamiento K-medias

Integrantes:	Nicolás López
	Alberto Rodríguez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor:	Max Chacón
Ayudante:	Javier Arredondo

11 de Diciembre de 2020

Tabla de contenidos

1. Introducción

En el estudio de una población, ya sea un grupo de personas, una base de datos, entre otros, se ha necesitado con el tiempo ir agrupando los datos que se van obteniendo para encontrar características similares e ir trabajando según su clasificación. Este proceso se denomina Clustering y es fundamental para los algoritmos automatizados que se enfocan en entrenar y conocer de forma adecuada los datos con los que desarrollan sus actividades.

En el presente laboratorio se busca trabajar una base de datos ya conocida de hongos de la familia Agaricus y Lepiota. Este data set fue extraído de la guía de campo de la sociedad Audubon. Todos los atributos del data set corresponden a variables categóricas.

Primero se debe procesar este conjuntos de datos mencionado para luego aplicar el mencionado Clustering. Esto servirá para crear conjuntos de hongos con características similares o también se busca que se agrupen en cuanto a hongos venenosos y comestibles. Este último estudio se busca realizar porque se sabe que la base de datos en su principal función agrupa los datos en dos clases. Entonces, al realizar este Clustering, se debería realizar esta agrupación de forma automática.

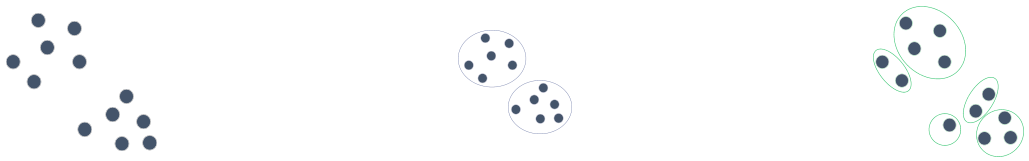
A lo largo del documento, se entrega un marco teórico para introducir en la materia al lector y luego se realiza el análisis y manejo de la base de datos de hongos. Por ultimo, se entrega una conclusión a partir de los resultados que se obtienen.

2. Marco Teórico

2.1. Clustering

El clustering son un conjunto de técnicas cuya finalidad es encontrar patrones o grupos (cluster) dentro de un conjunto de observaciones. A cada subconjunto o grupo de datos se le denomina "cluster" el cual contiene datos que son similares entre si, pero son distintos respecto a los datos de otros clusters. Se trata de un método *no supervisado*, esto quiere decir que no existe una respuesta correcta. (Amat, 2017).

Explicado de mejor manera puede ser con el siguiente en la figura 1, donde se puede agrupar en 2 grupos o 5 grupos el mismo conjunto de datos. (Martinez, 2020).



(a) Sin agrupar (b) Agrupación en 2 grupos (c) Agrupación en 5 grupos

Figura 1: Mismo conjuntos de datos con distintas agrupaciones

2.2. Algoritmo K-medias

Existen técnicas, para obtener los agrupamientos, este es un tipo de algoritmo de aprendizaje no supervisado que busca patrones en los datos sin tener una predicción específica como objetivo. En lugar de tener una salida, los datos solo tienen una entrada que serían múltiples variables que describen los datos.

Esta algoritmo necesita como dato de entrada el número de grupos en los que vamos a segmentar la población, a este valor se le denomina "K". A partir de este número "K" de clusters, el algoritmo coloca primero k puntos aleatorios (centroides). Luego asigna a

cualquiera de esos puntos todas las muestras con las distancias más pequeñas. A continuación, el punto se desplaza a la media de las muestras más cercanas. Esto generará una nueva asignación de muestras, ya que algunas muestras están ahora más cerca de otro centroide. Este proceso se repite de forma iterativa hasta que ya no hayan cambios y se logre una convergencia. Este resultado final representa el ajuste que maximiza la distancia entre los distintos grupos y minimiza la distancia intragrupo. (Lorca, 2019).

La figura 2, muestra como van cambiando las asignaciones de las observaciones a medida que se ejecuta cada paso del algoritmo (Amat, 2017).

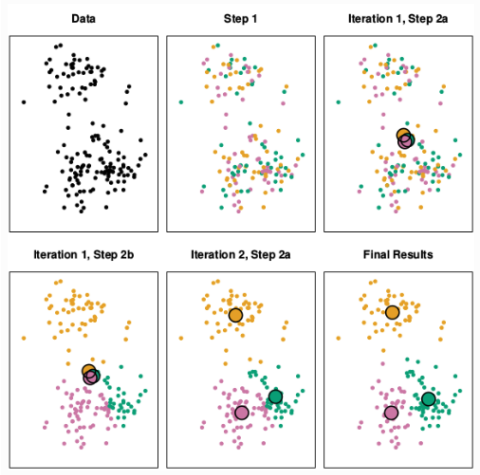


Figura 2: Función de transferencia final

2.3. Distancias utilizadas

Para realizar los agrupamientos entre los datos, se necesita calcular las distancias que tienen estos puntos a los centroides. La distancia entonces define la cuantificación de la similitud entre observaciones. Para el caso del conjunto de datos a analizar, se utiliza la distancia de Gower,ya que este set de datos solo contiene variables categóricas, debido a que se calcula como el promedio de disimilitudes parciales entre individuos. Cada disimilitud parcial (y por tanto la distancia de Gower). La ecuación 1 muestra como se calcula la distancia de gower.

$$d\,(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}(f) \tag{1}$$

3. Pre-procesamiento

3.1. Eliminación de variables

En el pre-procesamiento, se busca dejar los datos preparados para utilizarlos en el estudio, ya sea borrando variables o cambiando su tipo de dato. Para el caso de este laboratorio en particular, se busca preparar los datos para realizar un Clustering.

En materia de Clustering, se busca normalmente realizar un escalamiento de los datos. Sin embargo, la base de datos que se trabaja está conformada con variables categóricas que no se pueden normalizar, por lo que se trabajan los datos según sus valores categóricos. En un principio, los datos provenientes del dataset de hongos, vienen con nombres poco característicos ya que solo utilizan las iniciales, por esto, la primera modificación a la base de datos es otorgar nombres que identifiquen de mejor forma los atributos.

Luego de tener los atributos ordenados, se analiza que variables deben ser eliminadas de la base de datos. A simple vista, se eliminan 2 tipos de datos. El primero es la clase ("edibility"), ya que al realizar el Clustering se busca encontrar los grupos que dividan a los hongos en comestibles y venenosos, por lo que este dato afecta la agrupación de los datos. Por otro lado, se elimina el atributo "veil-type" ya que solo puede tomar un valor y no es significativo para el estudio. Una variable que se intentó inicialmente eliminar fue raíz del tallo ("stalk-root"), ya que es un atributo conflictivo, porque en varias observaciones su valor es desconocido. Esta opción se descartó porque al eliminar las filas con ese valor desconocido se pierde mucha información, además cuando se tiene variables categóricas, los valores desconocidos se ignoran no es necesario eliminarlos. (Guha, 1999).

3.2. Análisis de componentes múltiples

Este análisis tiene como objetivo resumir una gran cantidad de datos en un número reducido de dimensiones, con la menor perdida de información posible (Kassambara, 2017). Se utiliza análisis de componentes múltiple porque se tiene más de 2 atributos categóricos.

A continuación, mediante el lenguaje R se analiza que hacer con las variables restantes. Se observan distintos gráficos en dos dimensiones que muestran el porcentaje en

que se presenta esta variable.

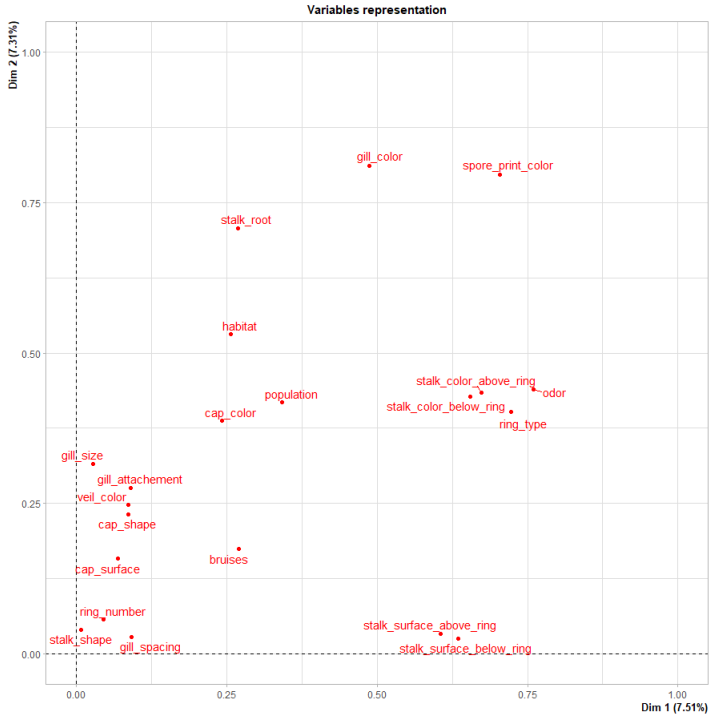


Figura 3: Análisis de componentes múltiple

El gráfico de la figura 3, se obtiene a partir de utilizar la función MCA() en R. Se puede observar como se distribuyen los componentes. Se puede notar como el color de las esporas, el olor, la raíz del tallo o el color de branquias son importantes para el estudio. Sin embargo, la mayoría de las variables se encuentran distribuidas en la parte inferior izquierda del gráfico (esto indica que su dimensión en el estudio es pequeña pero la suma de ellas forma una parte significante de la base de datos), por lo que tomar solo las variables mas importantes eliminarían una gran cantidad de datos que puede ser importante.

A partir del gráfico de la figura 4, se concluye que la dimensión más grande del estudio no llega a más de 10 % y que las 10 primeras dimensiones solo contienen el 48.1 %. Esto quiere decir que reducir la cantidad de dimensiones, la información se reduce drásticamente. Por lo que no es posible reducir las dimensiones para realizar el Clustering. Para confirmar, se realiza un último análisis de los datos para buscar un posible descarte de una variable.

En la figura 5 se obtiene un resultado similar al de la figura 3 ya que se puede identificar como 5 tipos de datos son más influyentes que los demás. También se puede ver

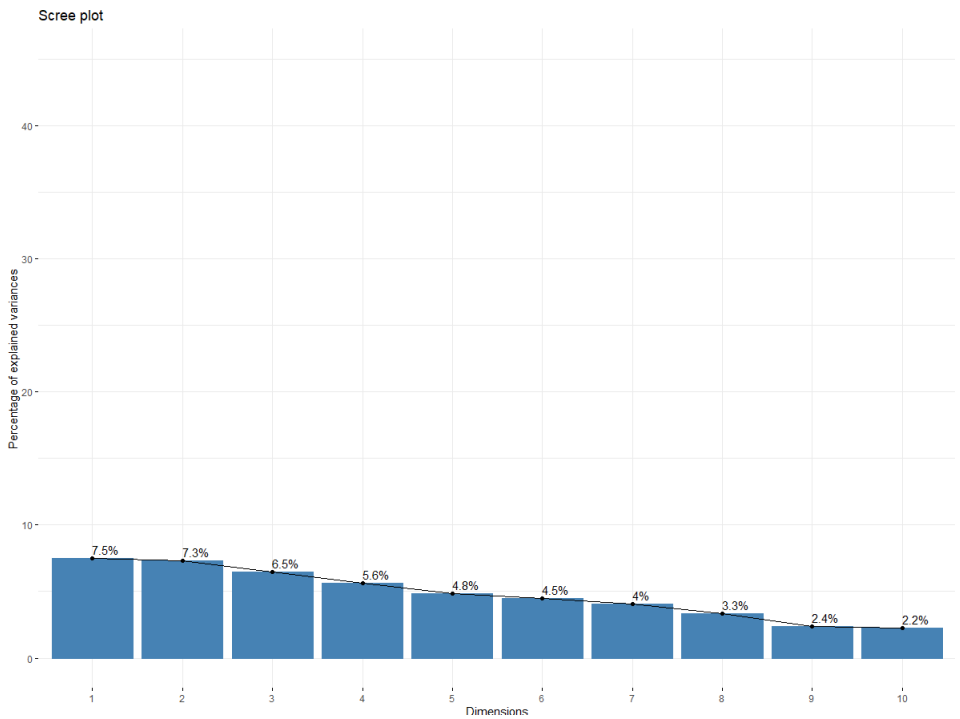


Figura 4: Dimensiones de la base de datos

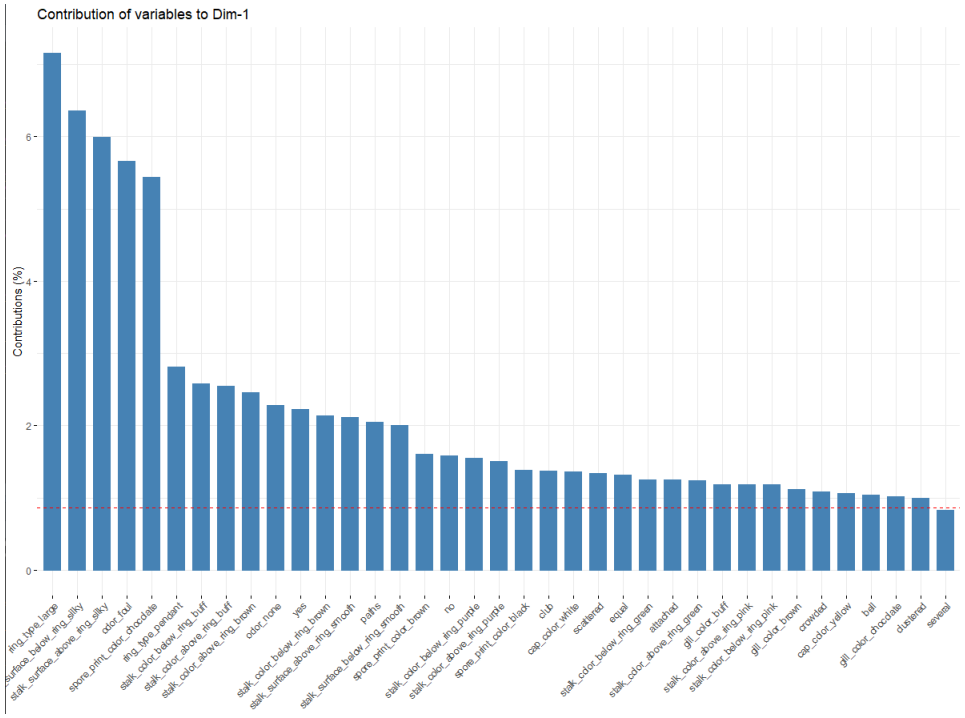


Figura 5: Contribución de las variables

como las demás variables se comportan de manera similar en cuanto a su contribución con el estudio.

Finalmente, se decide no seguir eliminado variables para el estudio, ya que dejar los atributos más contribuyentes eliminaría gran parte de la base de datos y la mayor parte de datos se comportan de manera uniforme. Por esto no se descarta otro tipo de dato para el Clustering.

4. Obtención del Clúster

4.1. Número de k grupos a formar

Para encontrar el K óptimo con el cual generar los clusters, existe distintas formas de encontrarlo. La planeada en utilizar era la métrica de el ancho promedio de las siluetas (silhouette). Sin embargo lamentablemente se intentó con varias funciones para obtener el óptimo, pero ninguna terminó de ejecutarse, estó se debió al gran tamaño de observaciones que se tienen.

Buscando otras opciones se tiene el algoritmo de ROCK (Guha, 1999), el cual calcula que al trabajar con el set de datos *mushrooms*, el valor óptimo de grupos es 21. A pesar de esto los resultados arrojaron que la mayoría de los datos solo estaban contenidos en clusters, mientras los otros tenían mucha menor cantidad de datos, incluso se dan cuenta que muchos clusters tiene información en común, es decir, que los grupos no están de todo bien separados. Por lo tanto se probará primeramente con un K=3.

4.1.1. Clustering con 3 grupos (K=3)

Si se realizan 3 clusters, al observar los valores de los atributos de cada cluster con la función de R *summary*, podemos observar que los 2 primeros clusters concentran la mayoría de la información del set de datos, siendo la primera casi el doble de la primera (observar figura 7), también de que son pocos lo atributos que tienen valores distintos entre clusters. Por ejemplo el atributo *odor* (olor) o *capShape* (superficie de tapa), pero por lo general no se ve una gran diferencia entre los atributos de cada cluster. En la figura 6 se puede ver el agrupamiento de 3 clusters, se puede apreciar como los grupos en si no están comprimidos completamente (sobretodo el cluster 1), hay partes de cada cluster que están muy dispersas, además que los cluster comparten datos en común por lo que no están muy bien separados entre ellos.

Debido a los resultados observados se intentará con un K=2.

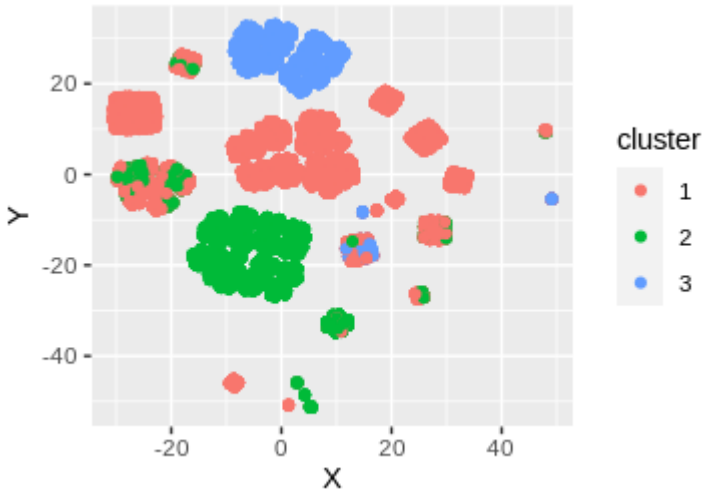


Figura 6: Agrupamiento en 3 clusters

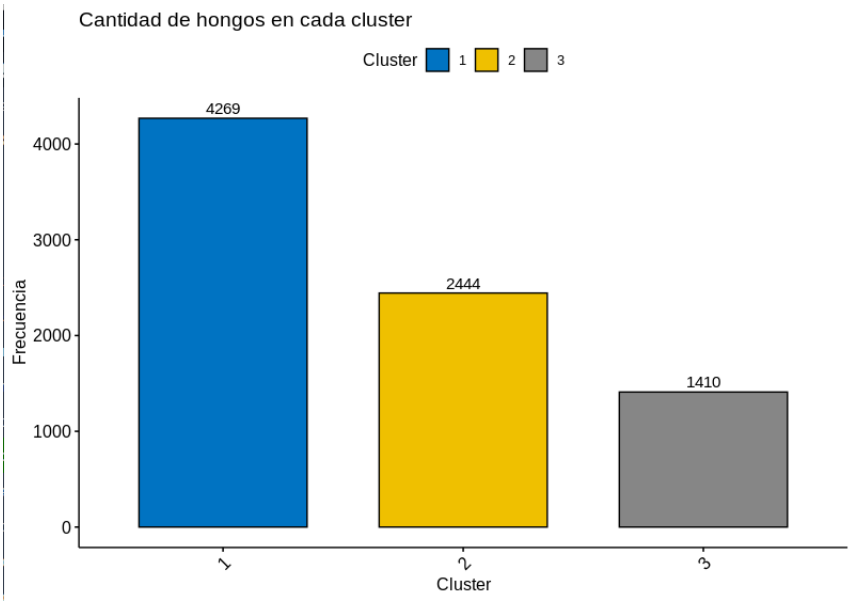


Figura 7: Cantidad de hongos en cada cluster

4.1.2. Clustering con 2 grupos (K=2)

Se realizan 2 clusters, ya que se analizó que con 3 grupos no hay una muy buena separación de los hongos. Además al tener 2 grupos podemos comparar si estos concuerdan con las clases presentes en el set de datos: comestible y venenoso.

Los valores de los atributos entre los clusters es distinto, solo existen pocos datos que son similares entre ellos. En la figura 8 se puede observar lo anterior de mejor forma.

El cluster 1 no es muy comprimido, es decir tiene atributos con valores muy distintos que si pueden estar dentro de ese grupo. En cambio el cluster 2, se puede observar que los datos estan muy comprimidos, eso si en 2 “masas” distintas, son solo muy pocos los datos que no están cerca de ninguna de estas 2.

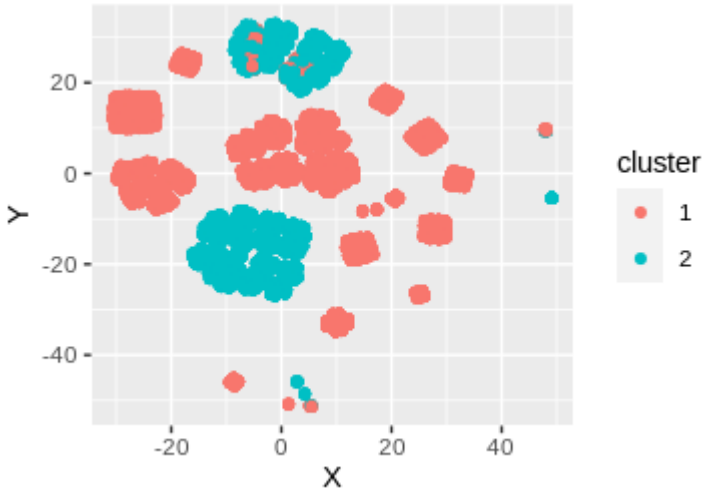


Figura 8: Agrupamiento en 2 clusters

4.2. Algoritmo k-medoids

Para obtener los resultados de la sección anterior, se usó una variación del algoritmo k-means, llamado k-medoids. La distancia de Gower encaja bien con el algoritmo de k-medoids, además este algoritmo es una alternativa más robusta y menos sensible a los valores atípicos que el algoritmo k-means. La principal diferencia entre el algoritmo de k-meoids y k-meas, es que el primero escoge como centroide un dato existente. (Filaire, 2018).

La función utilizada en R para utilizar este algoritmo, es *pam* (Maechler, 2020), perteneciente a la librería *cluster*, esta realiza el escalamiento de los datos internamente, por lo que no era necesario hacerlo. También indicar que para calcular la distancia de Gower, se utilizó la función *daisy* de la misma librería anterior. (Maechler, 2019).

5. Análisis de los resultados

Como se explicó anteriormente, cuando se utilizó un K=2, se logró mejor compresión de los cluster, es decir, los datos estaban cercas unos con los otros, pero el cluster 1 no cumplía del todo esto, ya que tenia muchos datos distribuidos en varias partes.

En la figura 9, se puede observar la cantidad de hongos que contiene cada cluster. Si comparamos esto con la clase inicial del set de datos, de saber si los hongos son comestibles o venenosos, esta distribución no se acerca. Ya que sabemos del laboratorio 1, que en el set de datos inicialmente se tiene 4208 hongos catalogados como comestibles y 3915 hongos catalogados como venenosos, esto equivale al 51,80 % y 48,2 % respectivamente. Mientras que el primer cluster contiene un 63.29 % de los datos y el segundo un 36,71 %.

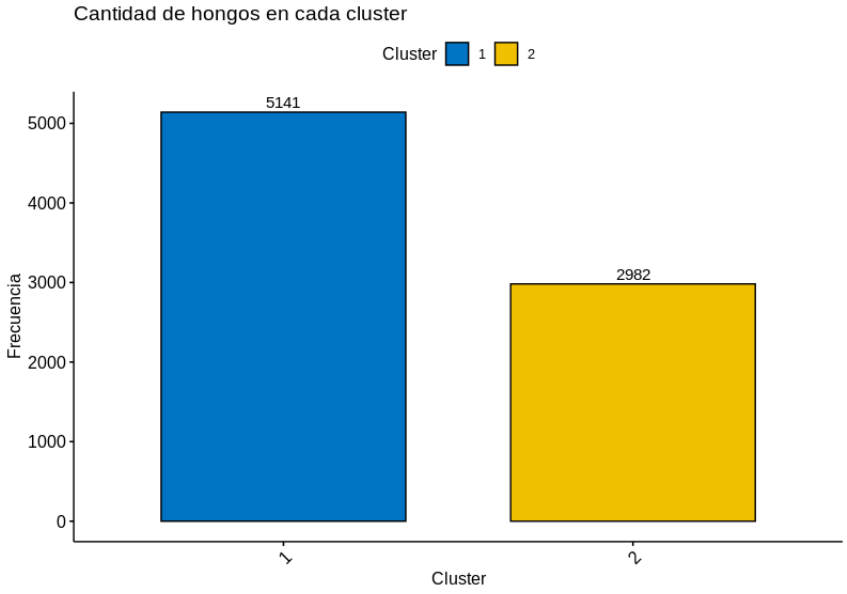


Figura 9: Cantidad de hongos en los 2 clusters

Analizando la división de los datos obtenidos, el Clustering lo que realizó fue una agrupación por similitudes en los atributos. Sin embargo, no se dividen los grupos en especies venenosas y comestibles, porque el método de Clustering no conoce los datos que son más influyentes (por ejemplo el olor) en el estudio para determinar a que pertenece un hongo. En este caso, se analiza que si se otorgara el peso de esta variable, la agrupación cambiaría y se entregaría una agrupación similar a la que se da en la experiencia del Laboratorio 1.

Si se compara con el estudio estudiado para el laboratorio 1, concuerda con lo

que se dijo, de que no hay una forma simple de clasificar los hongos (por lo menos la familia Lepiota y Agaricus). (Schlimmer, 1987). Debido a lo presentado, se considera que el clustering no es una buena herramienta para determinar el tipo de clase de los hongos.

En cuanto a K=3 se pudo notar que se crearon grupos con características similares. En este caso, no se puede decir que se clasificaron en venenosas y comestibles, ya que no se podría identificar a que pertenece cada grupo. Por lo que, realizando una clasificación notando los atributos, se podría entregar un nombre a cada grupo y decir a que pertenece cada uno.

En cuanto al procesamiento de los datos, ya que no se pudo eliminar una gran cantidad de variables o realizar un escalamiento o normalización correspondiente, se determina que la base de datos de hongos no es adecuada o debe realizarse un trabajo más profundo a los datos para que estos sean más representativos por los cluster. Esto debido a los resultados obtenidos con K=2 y K=3.

6. Conclusión

En cuanto a los objetivos, se tiene que gracias a esta experiencia se pudo enterder de mejor forma el análisis de agrupamientos visto en cátedra, ya que se lleva a la practica la aplicación de clustering sobre un conjunto de datos real. Además se utilizó otra materia vista en clases, el cual es el análisis de componentes principales, esto como opción de poder realizar un buen pre-procesamiento.

Mencionando el pre-procesamiento anteiormente, pese a que se intentaron forma para poder filtrar la información del set de datos, no fue mucho lo que se logró, solo la eliminación de 2 atributos (uno de ellos las clases), esto debido a que no existía una gran diferencia entre las dimensiones.

En base a los resultados, aunque se logró un agrupamiento en torno a 2 clusters, ya que este se analizó que era el más óptimo, aún así los resultados no fueron buenos. En un cluster se tenía el 63,29 % de las observaciones, mientras que en el otro el 36,71 % restantes. La clasificación de las clases no fueron correctas, ya que se sabía con antelación como debía ser la correcta agrupación del set de datos, con 4208 hongos comestibles y 3915 venenosos, lo cuál en porcentaje es 51.80 % y 48.2 % respectivamente. Lo que nos indica una gran diferencia entre ambas proporciones.

Aspectos positivos a destacar es que se aprendió de buena manera a realizar clustering gracias a la librería *cluster* de R. Se analizó correctamente los resultado, se internalizó de manera correcta la materia, a pesar de que el pre-procesamineto y el clustering no se pudieron realizar exitosamente con las herramientas disponibles.

Finalmente, se concluye que el agrupamiento o clustering no es una buena herramienta para este conjunto de datos, que es mejor aplicar las reglas lógicas del laboratorio anterior (Wlodzislaw D, 1997).

Bibliografía

Amat, R. (2017). *Clustering y heatmaps: aprendizaje no supervisado.*

Filaire, T. (2018). *Clustering on mixed type data.*

Guha, S. (1999). *ROCK: A Robust Clustering Algorithm for Categorical Attributes.*

Kassambara, A. (2017). *MCA - Multiple Correspondence Analysis in R: Essentials.*

Lorca, D. (2019). *K-Means Clustering: Agrupamiento con Minería de datos.*

Maechler, M. (2019). *Dissimilarity Matrix Calculation.*

Maechler, M. (2020). *Partitioning Around Medoids.*

Martinez, J. (2020). *Clustering (Agrupamiento), K-Means con ejemplos en python.*

Schlimmer, J. C. (1987). *Concept Acquisition through Representational Adjustment.* PhD thesis, University of California.

Wlodzislaw D, Adamczack R, G. K. (1997). *Extraction of crisp logical rules using constructive constrained backpropagation networks.* Proceedings of International Conference on Neural Networks (ICNN'97).