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

ANÁLISIS DE DATOS: LABORATORIO 2

AGRUPAMIENTO K-MEDIAS

PABLO CÁCERES LUZANTO
BENJAMÍN HERNÁNDEZ CORTÉS

Profesor:
Max Chacón
Ayudante:
Adolfo Guzmán

Santiago - Chile
19 de mayo de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

En la experiencia anterior, se estudió en profundidad la base de datos 'Zoo' creada por el investigador Richard Forsyth, pero a nivel estadístico trivial. Como es sabido, existen métodos algo más complejos para estudiar datos y extraer aún más conocimiento de ellos de lo que puede otorgar un análisis estadístico simple. Es por aquella razón, que en esta experiencia se intenta completar tres objetivos en concreto, los cuales son:

- Extraer el conocimiento de la Base de Datos 'Zoo' utilizando el algoritmo de *clustering* K-means y realizar el análisis respectivo.
- Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- Analizar por grupo e identificar aquellas características más relevantes, si un dato clasifica mejor a una clase que otra, e inferir conocimiento respecto a ello.

1.2 ORGANIZACIÓN DEL DOCUMENTO

El presente documento está estructurado básicamente en cuatro grandes capítulos. Uno de ellos corresponde al Marco Teórico, en donde se describen ciertos conceptos que tienen directa relación con el desarrollo de la experiencia como tal. Por otra parte, se tiene el capítulo denominado Pre-Procesamiento en el cual se definen ciertos criterios para eliminar registros o columnas que posean datos perdidos, inconsistentes, *outliers* o datos que simplemente no aporten información relevante. Posterior a este, se presenta el capítulo de Obtención del *Cluster*, en donde, como su nombre lo indica, se obtiene el cluster, pero de acuerdo a diversos parámetros que se van ingresando a la función escogida para obtenerlo. Luego, se realiza un exhaustivo Análisis de los Resultados obtenidos en el capítulo denominado de igual forma. Finalmente, se presentan las Conclusiones en donde se expone lo aprendido durante el desarrollo del laboratorio.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

Para el estudio y análisis de los datos se utiliza la base de datos *"Zoo Data Set"* contenida en el archivo *zoo.data* cuya información y documentación se encuentra detallada en el archivo *zoo.names*.

Por otro lado, para realizar el agrupamiento de los datos en sí, se hace uso del lenguaje de programación R junto a la IDE *RStudio* en su versión estable más reciente, e instalado en un Sistema Operativo basado en *Linux*.

CAPÍTULO 2. MARCO TEÓRICO

2.1 CLUSTERING

Clustering es un problema de aprendizaje no supervisado, que, como cualquier otro problema de este tipo, trata de encontrar una estructura en una colección de datos 'sin etiqueta'. En palabras más simples, *Clustering* corresponde al proceso de organizar objetos (datos) en grupos, cuyos miembros son similares de alguna manera. Por lo tanto, un *Cluster* es una colección o grupo de datos que son 'similares' entre ellos y 'diferentes' a los objetos que pertenecen a otros grupos o *Clusters*.

Lo anterior, se puede representar de manera gráfica con la siguiente figura.

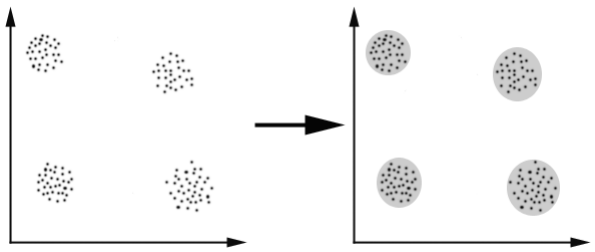

Figura 2.1: Clustering

Como se aprecia en la figura 2.1, es posible identificar cuatro grupos en los que se pueden dividir los datos. En este caso, el criterio de similitud entre ellos es la distancia, pero también pueden existir otros criterios de agrupamiento, como por ejemplo, el de tipo conceptual, en donde dos o más datos u objetos pertenecen al mismo *Cluster* si éste define un concepto común a todos los objetos. En términos simples, los objetos se agrupan de acuerdo a su ajuste a conceptos descriptivos, no de acuerdo a simples medidas de similitud. [1]

A continuación, se presenta un algoritmo de agrupamiento basado en distancias, conocido como el Algoritmo de las k-medias.

2.2 ALGORITMO K-MEDIAS

El algoritmo de las k-medias, es uno de los algoritmos de aprendizajes sin supervisión más simple que resuelven el problema de agrupamiento de datos. La idea principal es clasificar un conjunto de datos a través de un numero dado 'k' de *clusters* definido con anterioridad. Para realizar lo anterior, es necesario definir los 'k' centroides de dichos *clusters*, los cuales deben colocarse de manera astuta, debido a que los resultados que se obtengan dependerán exclusivamente de donde se ubiquen estos. De esta forma, la mejor opción, es ubicarlos lo más lejos posible el uno del otro. Luego, se toma cada punto perteneciente a un determinado conjunto de datos dado y se asocia al centroide más cercano. Cuando ya no quedan puntos pendientes, se

recalculan los 'k' nuevos centroides como baricentros [1] de los grupos resultantes del paso anterior. Después de tener los nuevos 'k' centroides, se debe realizar un nuevo enlace entre los mismos puntos y el centroide nuevo más cercano. Luego, solo resta iterar de la misma forma, hasta que no varíen los centroides.

Finalmente, cabe destacar que el algoritmo descrito, intenta minimizar una función objetivo, en este caso, una función del error cuadrático, la cual es:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{2.1}$$

En la ecuación 2.1, $||x_i^{(j)} - c_j||^2$ es una medida de la distancia elegida entre un punto de los datos $x_i^{(j)}$, y el centro del *cluster* c_j es un indicador de la distancia de los 'n' puntos de los datos respecto a sus correspondientes centros de *cluster*.

Cabe destacar, que el algoritmo de las k-medias no necesariamente encuentra la configuración óptima (que corresponde al mínimo de la función presentada en la ecuación 2.1). Además, como se menciona en la descripción del algoritmo, éste es notoriamente sensible a los centroides elegidos inicialmente para cada *cluster*. Para reducir, el efecto anterior, se puede ejecutar el algoritmo varias veces.

2.3 DISTANCIAS UTILIZADAS

2.3.1. Distancia de Manhattan

Por lo general, se utiliza cuando en la base de datos se tienen vectores de muchas dimensiones. Gráficamente, la función de distancia de Manhattan calcula la distancia que se recorrería para llegar de un punto de los datos al otro si se sigue una ruta en forma de cuadrícula, es decir, arriba, abajo, derecha o izquierda. [2] Finalmente, la fórmula que define la distancia de Manhattan entre dos vectores $\vec{x}$ e $\vec{y}$ es la siguiente:

$$||\vec{x} - \vec{y}|| = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2}$$

en donde x_i e y_i corresponden a elementos de los respectivos vectores.

2.4 MÉTODO DE LAS SILUETAS

El objetivo de este método es encontrar el número 'k' óptimo de agrupamientos y esto se hace calculando la media del coeficiente de silueta de cada objeto de la muestra. Mientras mayor es el valor de la media de un dato, mejor será la distribución de los datos en sus respectivos conglomerados una vez realizado el *cluster*.

[1] Centroide de una superficie contenida en una figura geométrica plana. Cualquier recta que pasa por él, divide a dicho segmento en dos partes de igual momento respecto a dicha recta.

CAPÍTULO 3. PRE-PROCESAMIENTO

3.1 ELIMINACIÓN DE VARIABLES

En este caso, la eliminación de variables se realiza considerando que la eliminación de ésta reducirá la dimensionalidad de la base de datos y además, considerando el aporte e impacto de la variable al momento de realizar el *cluster*.

3.1.1. Variable 'animal_name'

Esta variable es eliminada debido a que, como su nombre lo indica, solo representa el nombre de los animales presentes en la muestra, y por ende, es un dato no-numérico que no realizará ningún aporte de valor al proceso de *clustering* a realizar más adelante.

3.1.2. Variable 'type'

Por otra parte, se elimina la variable *'type'* que indica el tipo de animal, puesto que ésta de cierta forma ya agrupa los animales en grupos, lo cual a su vez podría generar un *cluster* en el que se tendrán siete grupos claramente definidos por el tipo de los animales. En consecuencia, si se incluyera esta variable, se obtendría un *cluster* altamente sesgado o incluso viciado.

3.2 ELIMINACIÓN DE INSTANCIAS

También puede darse la eliminación de instancias, bajo los mismos criterios mencionados en el inicio de la sección anterior.

3.2.1. Instancia 'girl'

En este caso, esta instancia es eliminada dado que no resulta adecuada mantenerla dentro del conjunto de datos, debido a que en sí mismo, el conjunto de datos fue destinado y confeccionado para detallar aspectos de las distintas especies animales pertenecientes a un zoológico.

3.2.2. Instancia 'frog'

En el caso de la instancia *'frog'*, se opta por eliminar una instancia de ésta, ya que se poseen dos y la única diferencia que se halla entre ambas es que una resulta ser venenosa y la otra no, atributo que no se considera importante o crucial a la hora de definir un *cluster* para dicha instancia.

CAPÍTULO 4. OBTENCIÓN DEL CLUSTER

4.1 DISTANCIA

Para calcular las distancias, entiéndase disimilaridades o proximidad, de un punto a otro del *DataSet* se utiliza la distancia de Manhattan, la cual permite realizar dicho cálculo en Bases de Datos que por lo general presentan grandes dimensionalidades, tal como ocurre en el caso actual, en el que se tienen 16 variables o dimensiones utilizadas para realizar el *cluster*. Además, no sólo se cuenta con variables del tipo binarias, sino que también se cuenta con una variable del tipo discreta (patas). Cabe destacar, que el cálculo de la Distancia de Manhattan es realizado de manera implícita en R por la función *pam()* de la librería *'Cluster'* a la cual se le pasa como parámetro en la variable *'metric'* el valor de *'manhattan'*.

4.2 NÚMERO DE GRUPOS

Como se menciona en el capítulo 2, es sumamente importante definir previo a realizar el proceso de *Clustering* de una Base de Datos, el número 'k' de grupos a utilizar para realizar dicho proceso, debido a que este número es pasado como parámetro a la función que realizará el *Clustering*. Es por esto, que en esta ocasión, se utiliza el método de las siluetas para calcular dicho número 'k'. Este método, calcula un coeficiente el cual ayuda a definir qué cantidad de grupos 'k' es la adecuada para agrupar los datos. Mientras mayor sea el coeficiente, más confiable será el numero de grupos obtenido gracias a la función. Cabe destacar, que la función a utilizar en R, para visualizar la gráfica que indica la cantidad de grupos adecuados es *'fviz_nbclust()'* y realiza el método de las siluetas descrito anteriormente de manera transparente.

Para el problema actual, tal como se aprecia en la figura 4.1, se obtienen una serie de coeficientes a través del método de las siluetas y como se menciona anteriormente, se escoge aquel numero de *clusters* cuyo coeficiente sea el más elevado, en este caso cuatro.

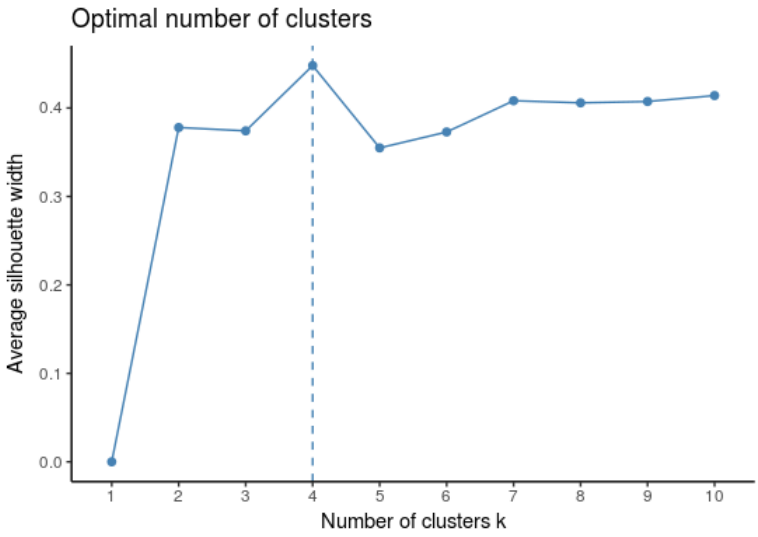

Figura 4.1: Resultado obtenido a través del método de las siluetas

4.3 CLUSTERING

Una vez se tiene el valor de 'k', sólo queda ejecutar el algoritmo de las k-medias con parámetro inicial de 'k' igual a cuatro. Al ejecutar el algoritmo a través de la función *fviz_cluster()* se obtiene lo que se muestra en la figura 4.2. Cabe destacar, que por razones de espacio y por un tema legibilidad a la hora de analizar el agrupamiento obtenido, se muestra cierto porcentaje del contenido en cada *cluster* (29,2 % para la Dimensión 1 y 20,9 % para la Dimensión 2).

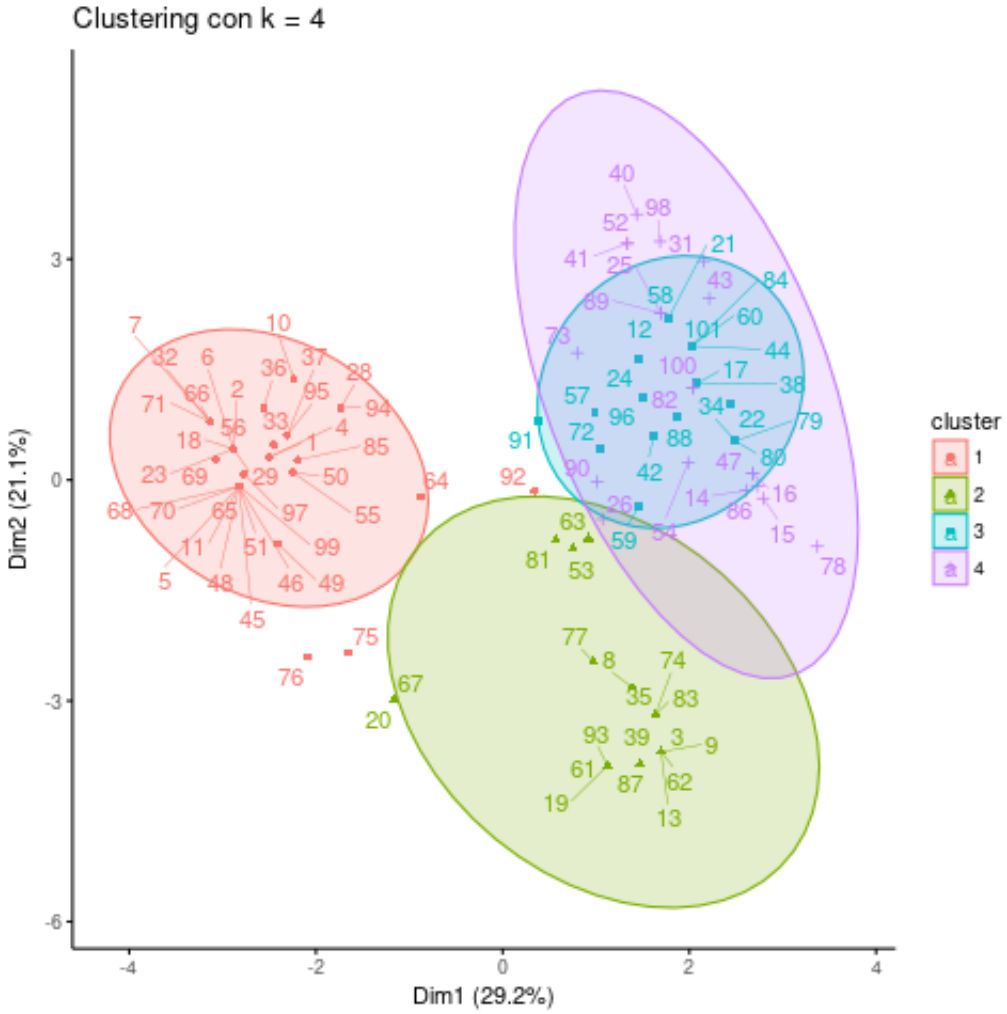

Figura 4.2: Clustering en plano bidimensional con k = 4

Por otra parte, dado que originalmente los animales del conjunto de datos están clasificados en siete distintas clases, surge la necesidad de realizar un proceso de *Clustering* para *'k'* igual a siete, resultando de esta forma lo que se muestra en la figura 4.3.

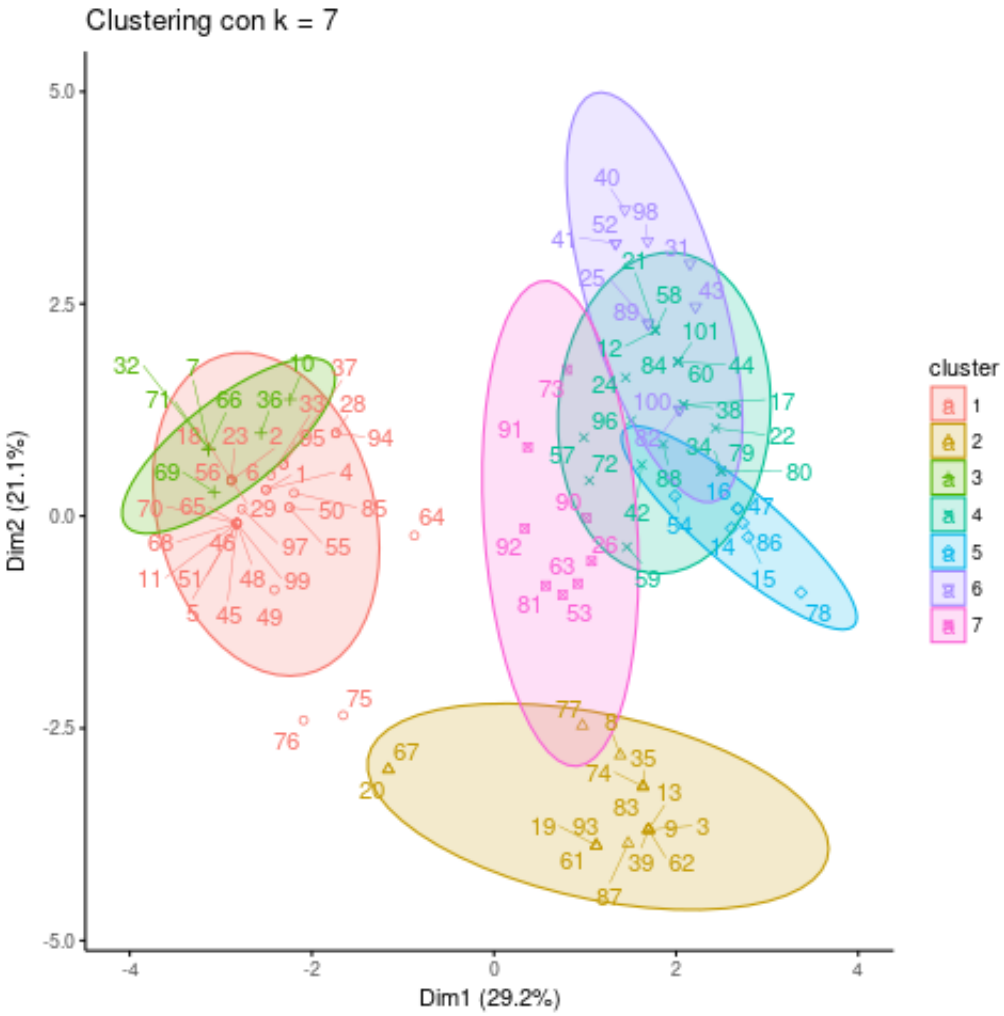

Figura 4.3: Clustering en plano bidimensional con k = 7

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS

5.1 ANÁLISIS DE AGRUPAMIENTO CON K = 4

Para realizar un correcto análisis del agrupamiento con k = 4, el cual sugiere que es óptimo para el conjunto de datos, se genera una matriz de coincidencias a modo de comparación entre el *Cluster* obtenido y el agrupamiento original del conjunto de datos. Cabe destacar, que la tabla que se presenta, no es una matriz de similaridades, sino una matriz en donde básicamente, se tienen las clasificaciones originales y se observa en que clasificación se agrupan las observaciones de dicho grupo, en el *Cluster* realizado.

$\frac{Cluster}{Original}$	type.1	type.2	type.3	type.4
type.1	38	2	0	0
type.2	0	0	20	0
type.3	1	3	1	0
type.4	0	13	0	0
type.5	0	1	0	2
type.6	0	0	0	8
type.7	0	0	0	10

Cuadro 5.1: Matriz de coincidencias entre Cluster con k = 4 y Agrupamiento Original

Como se aprecia en el Cuadro 5.1, en el caso de los mamíferos (type.1 en el *DataSet* original), estos son agrupados en dos clases en el *Cluster* generado (type.1, type.2). Lo anterior, resulta válido de cierta forma, considerando que la idea o el concepto que se tiene de un 'buen *Cluster*' para este caso es que, éste agrupe los animales en su mayoría según su clasificación original. Dicho aquello, en la tabla se pueden encontrar 'buenos agrupamientos' cuando la mayoría de datos presentes en una fila, no se encuentren dispersos por muchas columnas, porque eso sí se podría considerar sinónimo de un error en la clasificación realizada por el *Cluster*.

En general se logra apreciar que se tiene un 'buen agrupamiento' para las especies mamíferos (type.1) y aves (type.2), dado que en el *Cluster*, la mayoría de las observaciones originales fueron separadas en un único grupo, respectivamente en los grupos uno (type.1) y tres (type.3) del *Cluster*, a diferencia de las demás especies que se mezclaron entre los grupos dos (type.2) y cuatro (type.4) del correspondiente *Cluster*.

Respecto a la situación de mezcla anterior, esto es esperable considerando que el *Cluster* fue definido para que agrupara entre un total de cuatro grupos ($k = 4$), por lo que fue necesario mezclar observaciones del conjunto de datos original, aún cuando tuvieran asignado un tipo diferente antes del proceso de *Clustering*.

5.2 ANÁLISIS DE AGRUPAMIENTO CON K = 7

En este caso, también se genera una matriz de coincidencias a modo de comparación entre el *Cluster* obtenido y el agrupamiento original del conjunto de datos, con la clara diferencia de que ahora se posee siete grupos para separar los datos del conjunto original, en el *Cluster*.

$\frac{Cluster}{Original}$	**type.1**	**type.2**	**type.3**	**type.4**	**type.5**	**type.6**	**type.7**
type.1	31	2	7	0	0	0	0
type.2	0	0	0	20	0	0	0
type.3	0	1	0	0	0	0	4
type.4	0	13	0	0	0	0	0
type.5	0	0	0	0	0	0	3
type.6	0	0	0	0	0	8	0
type.7	0	0	0	0	7	2	1

Cuadro 5.2: Matriz de coincidencias entre Cluster con k = 7 y Agrupamiento Original

Como se aprecia en el Cuadro 5.2, para el caso de los mamíferos (type.1 en el *DataSet* original), estos son agrupados en tres distintas clases en el *Cluster* generado (type.1, type.2, type.3).

En general, se tiene un 'buen agrupamiento' para todas las especies, pero las especies que mejor se agruparon (o se agruparon de acuerdo a su clasificación original) en el *Cluster* fueron las aves (type.2), los peces (type.4), los anfibios (type.5) y los insectos (type.6).

Respecto a la clase type.7 (Otros tipos de animales, en el agrupamiento original), resulta lógico que esté distribuido en distintas clases por el *Cluster*, ya que en el grupo original, estos pertenecían a distintas especies no agrupadas en las anteriores y básicamente lo que hace el *Cluster* en este caso es buscar patrones de estas especies que se puedan incluir o tengan en común con las otras y los agrupa de esa forma.

CAPÍTULO 6. CONCLUSIONES

Respecto a los resultados obtenidos, se logra corroborar que si bien el método de agrupamiento de las k-medias permite obtener un *Clustering* considerablemente bueno, cuando se emplea un valor de k adecuado, al mismo tiempo se torna una dificultad analizar en base a que características realiza el agrupamiento, en especial cuando el *Cluster* se definió con un total de grupos menor (o mayor) al que se tenía presente en el conjunto de datos original, lo cual se pudo apreciar en el agrupamiento con $k = 4$, en donde se poseían algunas especies mezcladas en el *Cluster*, y que originalmente fueron clasificadas como tipos de animales distintos.

Así mismo, y considerando los resultados de los procesos de *Clustering* con valores de $k = 4$ y $k = 7$, los tipos de especies que resultaron mejor definidas en los *Clusters* fueron los mamíferos y las aves. Probablemente, esto se deba a que las características que los definían, no eran compartidas (en gran parte) por los otros tipos de especies.

En cuánto al desarrollo del laboratorio, algunos aspectos a mejorar y que quedaron pendientes son la caracterización del *Cluster* con $k = 4$, cuyo fin era, tener indicios sobre bajo que aspectos intentó agrupar las especies el algoritmo. También, otro aspecto que pudo haber mejorado es la visualización de los datos, dado que las componentes que se utilizaron para los gráficos correspondientes, solo representan alrededor del 50 % de la información presente, algo no muy representativo ni categórico para analizar. Sin embargo, dentro de los aspectos positivos, se destaca el descubrimiento de un método de búsqueda de un valor de k óptimo, en base a la medida denominada 'siluetas'.

CAPÍTULO 7. BIBLIOGRAFÍA

[1] *A Tutorial on Clustering Algorithms*, dirección: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/ (visitado 16-05-2018).

[2] *Manhattan Distance Metric*, dirección: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Manhattan_Distance_Metric.htm (visitado 17-05-2018).