UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERIA

DEPARTAMENTO DE INGENIERIA EN INFORMATICA

Análisis de Clustering K-means para la Base de Datos ZOO utilizando la herramienta de software "R"

AUTOR:

LUIS ORELLANA ALTAMIRANO

UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERIA

DEPARTAMENTO DE INGENIERIA EN INFORMATICA

TABLA DE CONTENIDO

1 Resumen

Este trabajo de investigación consiste en estudiar e interpretar mediante el algoritmo de K-means los agrupamientos y relaciones con las clasificaciones originales de los datos correspondientes a la base de datos ZOO. Inicialmente, se describen los objetivos generales y específicos de este trabajo de investigación. Luego, se describe la motivación que inspira este trabajo, el dominio del problema y la descripción del problema a resolver. A continuación, se presenta una descripción de la solución propuesta indicando las características, propósitos, alcances y limitaciones de la solución a implementar. Luego, en la metodología, herramientas y experimentación se describe el marco teórico, la base de datos ZOO, su normalización y la presentación de las técnicas aplicadas al análisis de K-means, tales como: las medidas de similaridad de Manhattan , Euclidiana, Minkowski y distancia de Hamming, además de las evaluación de los cluster mediante los coeficientes entregados por la gráfica de siluetas y la utilización de técnicas de visualización tales como la tabla de características en base a centroides, gráfico de siluetas, diagrama de Venn y dendrograma, entre otras, haciendo uso del software "R" V3.3. A continuación, se muestran tablas y gráficos resultantes y su respectivo análisis detallado. Y finalmente, se entregan las conclusiones respecto al problema presentado.

Palabras clave: estadística descriptiva, Clustering, algoritmo K-means, base de datos Zoo, software "R", similitud por distancia

2 Objetivos

2.1 Objetivos Generales

El objetivo general de este trabajo de investigación es realizar una búsqueda, clasificación y análisis de clustering con el algoritmo K-means sobre la Base de Datos ZOO utilizando la herramienta de software "R" en base a su paquete "cluster" y función PAN, y así poder visualizar los cluster y relacionarlos con las características y clasificaciones de las diferentes especies de animales.

2.1 Objetivos Específicos

Los objetivos específicos planteados para este trabajo de investigación son:

- Revisar y normalizar la base de datos ZOO.

- Investigar y estudiar el algoritmo de K-medias, los diferentes métodos de medición de distancias y funciones de evaluación de métodos de agrupamiento y su implementación con el software "R".

- Realizar análisis deductivo a partir de los cluster encontrados y las hipótesis y/o problema planteado a solucionar en el desarrollo del estudio, además de la comparación en relación a los resultados del análisis de reglas de asociación del laboratorio 2.

3 Descripción del Problema

3.1 Motivación

La principal motivación de este trabajo de investigación es encontrar agrupamientos inherentes en la base de datos ZOO y que tienen relación con sus características y su clasificación. Además, de la búsqueda de conocimiento que permita inferir en conceptos, clasificaciones y relaciones existentes entre las diferentes características de especies de animales.

3.2 Literatura Relevante

Dentro de la literatura relacionada con la implementación y análisis de clustering para la base de datos ZOO, destaca lo escrito por Tao Li (2010), quien presenta primero un modelo de agrupación binario de datos genérico en base a sus relaciones de asociaciones simétricas, y en segundo lugar establecer las conexiones entre el modelo propuesto con otros métodos de agrupación existentes. También, está lo escrito por Guntur (2003) quien utiliza el algoritmo de K-Medias Modificado para agrupación de datos y simultaneidad de patrones, esto es, un nuevo método que agrupa simultáneamente patrones descubiertos y sus datos asociados.

3.3 Definición del Problema

El problema principal planteado radica en encontrar agrupaciones dentro de la base de datos ZOO, esto es para poder determinar patrones entre los cluster que permitan las agrupaciones de especies según sus características morfológicas, y así afirmar y abalar las hipótesis demostradas de clasificaciones de animales en los anteriores laboratorios.

También, se hace necesario aclarar algunas asociaciones implícitas en el laboratorio 1 que no eran completamente claras al momento de utilizar técnicas de estadística descriptiva tales

como moda, media, medidas de dispersión, etc. y el problema de clasificar especies de forma más precisa por medio de analizar reglas de asociación entre características morfológicas vistas en el laboratorio 2. Así pues, se hace necesario, mediante las agrupaciones, obtener conocimiento por medio de clustering y sus características relevantes entre variables que están presentes en el dataset ZOO.

Así pues, sería de mucho interés responder algunas preguntas como:

- ¿ Los cluster resultantes tienen características asociada a un tipo de clase?
- ¿ Es posible distinguir la(s) variable(s) que distinguen a cada cluster?

4 Descripción de la Solución Propuesta

La solución para el problema planteado será la investigación e implementación de la búsqueda y análisis de clusters para la base de datos ZOO utilizando el algoritmo K-means, y que permita aplicar métodos y/o técnicas de clustering.

Para tal escenario, se ha planteado hacer uso del software "R" y su librería "cluster" y su función PAN, la cual permite visualización, obtención y por consiguiente el análisis de clusters encontrados, además del uso de la funciones para evaluación del método de agrupamiento (gráficos y coeficientes de rendimiento).

4.1 Características de la Solución.

La solución contempla el uso de K-means y la experimentación de las técnicas de medida de similitud por distancia para la base de datos ZOO mediante el software "R", que no es posible inferir o extraer de forma natural. Para ello, se utilizarán los métodos contenidos y/o incorporados dentro del paquete de software llamado "cluster" y la función PAN que es utilizado por el software "R" para entrega y visualización de resultados en reportes, gráficos y tablas.

4.2 Propósitos de la Solución.

Al termino del trabajo de investigación y una vez analizados los resultados obtenidos de cada experimento se espera encontrar clusters dentro de la base de datos ZOO y así determinar patrones entre las especies de animales y sus agrupaciones según sus características morfológicas, y también poder realizar un comparativo con las conclusiones dadas en el laboratorio 1 referentes a los análisis de estadísticas descriptivas y laboratorio 2 referente a las reglas de asociación.

4.3 Alcances y Limitaciones de la Solución.

Este laboratorio contempla sólo el análisis y extracción de cluster con K-medias para la base de datos ZOO utilizando la herramienta de software “R” y su librería “cluster” con sus métodos integrados. El laboratorio no contempla realizar análisis a la base de dato ZOO con otros métodos de análisis de datos y/o minería de datos.

5 Metodología, Herramientas y Experimentación

5.1 Marco Teórico

K-medias (K-means)

K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de N observaciones en K grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Así pues, el criterio es asignar las N observaciones a los K clusters de modo que dentro de cada cluster el promedio de las diferencias de cada observación a la media del cluster, definido por los puntos del cluster, sea mínima.

Así se tiene que dado un conjunto de observaciones (x1, x2, …, xn), en donde cada observación es una vector real d-dimensional, así el clustering k-means ayuda a particionar las n observaciones en k (≤ n) conjuntos S = {S1, S2, …, Sk} a fin de minimizar la suma de cuadrados dentro del cluster (es decir, varianza). Formalmente, el objetivo es encontrar:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

En donde μi es la media de puntos en Si. Esto es equivalente a minimizar las desviaciones cuadradas pares de puntos en el mismo cluster.

$$\sum_{\text{Cluster } C_i} \sum_{\text{Dimension } d} \sum_{x,y \in C_i} (x_d - y_d)^2$$

Debido a que la varianza total es constante, esto también equivale a maximizar las desviaciones al cuadrado entre puntos en diferentes cluster (suma de cuadrados entre clústers)

Medidas de Similaridad por Distancias

Las medidas de similaridad más conocidas son las de distancia. Para dos vectores $\bar{x}$ e $\bar{y} \in \Re^n$, se tiene:

$$\|\bar{x} - \bar{y}\| = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

Dependiendo del valor de p se generan los siguientes casos particulares que serán utilizados en este trabajo de investigación:

- Para p=1 Distancia de Manhattan (block):

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

- Para p=2 Distancia Euclediana:

$$\|\bar{x} - \bar{y}\| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

- p->infinito Distancia de Schebyshev:

$$\|\bar{x} - \bar{y}\| = \max_{i=1,2\ldots n} |x_i - y_i|$$

- Distancia de Hamming:

  Si los atributos son categóricos, entonces la distancia Hamming propone una distancia entre dos cadenas: la distancia es 1 por cada elemento diferente y 0 por cada elemento idéntico en la cadena, ejemplo:

  La distancia de “toned” a “roses” es 3.

  La distancia de 1011101 a 1001001 es 2.

Método Silhouette (silueta)

Silueta se refiere a un método de interpretación y validación de la coherencia dentro de los clusters. La técnica proporciona una representación gráfica que cuanto están afianzado

cada objeto a su cluster. Fue descrita por primera vez por Rousseeuw (1986). Así pues, el valor de la silueta es una medida de la similitud de un objeto con su propio cluster (cohesión) en comparación con otros clusters (separación). La silueta varía de -1 a 1, donde un valor alto indica que el objeto está bien adaptado a su propio cluster y mal adaptado a los clústeres vecinos.

5.2 Descripción de la Base de Datos ZOO (Dataset ZOO)

El Dataset ZOO cuenta con 101 animales extraídos desde una colección Zoológica. Hay 23 variables con una serie de rasgos que describen a los animales. Se definen 7 Tipos de Clases: Mamíferos, Aves, Reptiles, Peces, Anfibios, Insectos (voladores) e Invertebrados.

Originalmente, las características disponibles para esta dataset era un conjunto de 16 morfologías, las cuales son dicotómicas, excepto la cantidad de patas las que pueden ir desde 0 hasta 8 patas. Sin embargo, con el propósito de poder disponer de datos normalizados, se divide esta variable en 6 distintas variables, las que al igual que las demás variables de este dataset, son binarias.

Con la finalidad de disponer del contenido del Dataset ZOO, y hacer más fácil su análisis en interpretación de estos, se dispone del siguiente archivo, el cual se detalla a continuación:

zoo-cluster.csv: Corresponde a una muestra de 101 animales. Por otro lado, el primer registro que se encuentra en este archivo es el encabezado, el cual corresponde a las características más relevantes de cada especie, pues cobra sentido al momento de agrupar y clasificar cada animal según sus especies por medio de estas características morfológicas.

Las características morfológicas disponibles para cada animal presente en el dataset son:

- hair $\in \{0, 1\}$: Posee pelaje (si, no)
- feathers $\in \{0, 1\}$: Posee plumas (si, no).
- eggs $\in \{0, 1\}$: Nace por medio de huevos (si, no)
- milk $\in \{0, 1\}$: Capacidad de amamantar (si, no)
- airborne $\in \{0, 1\}$: Capacidad de volar (si, no).
- aquatic $\in \{0, 1\}$: Vive en medio acuático (si, no)

- predator ∈ {0, 1}: Es depredador (si, no)
- toothed ∈ {0, 1}: Posee dentadura (si, no)
- backbone ∈ {0, 1}: Posee columna vertebral (si, no)
- breathes ∈ {0, 1}: Es pulmonado (si, no)
- venomous ∈ {0, 1}: Es venenoso (si, no)
- fins ∈ {0, 1}: Posee aletas natatorias (si, no)
- legs_0 ∈ {0, 1}: No posee patas (si, no)
- legs_2 ∈ {0, 1}: Posee dos patas (si, no)
- legs_4 ∈ {0, 1}: Posee cuatro patas (si, no)
- legs_5 ∈ {0, 1}: Posee cinco patas (si, no)
- legs_6 ∈ {0, 1}: Posee seis patas (si, no)
- legs_8 ∈ {0, 1}: Posee ocho patas (si, no)
- tail ∈ {0, 1}: Posee cola (si, no)
- domestic ∈ {0, 1}: Puede ser domesticado (si, no)
- catsize ∈ {0, 1}: Posee el tamaño de un gato doméstico (si, no)

Se han utilizado estas dos alternativas de clasificación de animal. Una de ellas es prescindir del siguiente campo, el cual a diferencia de los campos anteriores no es dicotómico.

- class_type ∈ {1, 2, 3, 4, 5, 6, 7}: Clasificación del animal, donde:
  - 1 = Mamífero
  - 2 = Ave
  - 3 = Reptil
  - 4 = Pez
  - 5 = Anfibio

- 6 = Insecto (volador)
- 7 = Invertebrado

La otra alternativa que será considerada al momento de crear los clusters, es transformar el campo de tipo de animales a variables binarias con la idea de contrastar los resultados de clustering y las verdaderas clasificaciones de cada animal:

- class_mammal $\in \{0, 1\}$: Es mamífero (si, no)
- class_bird $\in \{0, 1\}$: Es ave (si, no)
- class_reptile $\in \{0, 1\}$: Es reptil (si, no)
- class_fish $\in \{0, 1\}$: Es pez (si, no)
- class_amphibian $\in \{0, 1\}$: Es anfibio (si, no)
- class_bug $\in \{0, 1\}$: Es insecto (volador) (si, no)
- class_invertebrate $\in \{0, 1\}$: Es invertebrado (si, no)

5.2.1 Herramientas de Software

La herramienta de software utilizada para realizar los experimentos presentados en este documento es "R" versión (3.3.3). Esta herramienta está bajo licencia GNU y cuenta con su propio lenguaje de programación con un enfoque al análisis estadístico.

"R" entre sus características principales permite: modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, generar gráficos, etc.

Como paquetes de software dentro de "R", se utilizó "Clúster", "FactoExtra" y "FactoMineR" y "Bootstrap", los cuales permiten crear clusters, agrupaciones y evaluar los métodos de cluster aplicados con la finalidad de realizar inferencias de características comunes a cada agrupación, además, estos paquetes permiten la visualización de resultados en reportes, gráficos y tablas.

5.3 Experimentación.

Los siguientes comandos en "R" fueron utilizados para realizar el análisis de los clusters generados por medio del algoritmo K-Medias:

- `library('')`

Permite Agregar nuevas librerías, las cuales incluyen funciones no disponibles de forma nativa al repositorio local del Software "R".

- `Nombre_Variable = read.csv("")`

Lee en memoria volátil un documento con extensión "csv", y lo almacena en la variable "Nombre_Variable".

```
summary(data_kmeans)
```

Entrega detalles globales estadísticos de los datos del clusters (vector del clúster, desviación estándar entre miembros y centroides, medidas por clúster, etc).

```
names(data_kmeans)
```

Campos disponibles en matriz generada por medio de algoritmo K-Means.

```
table(data_kmeans$clustering)
```

Número de miembros por clúster.

```
data_norm["nombre_campo"] <- NULL
```

Es eliminado el campo indicado.

```
data_cluster <- dist(data_norm, method = "binary", diag = FALSE,
upper = FALSE, p = 2)
```

Se crean vectores de distancia para los datos disponibles en dataset, por medio de métrica de distancia “Hamming” (distancia para datos dicotómicos).

```
data_matrix <- as.matrix(data_cluster)
```

Conversión de vectores de distancia “Hamming” en matriz, con la finalidad de aplicar sobre esos datos, el algoritmo de aglomeración K-Means.

```
data_kmeans <- pam(data_matrix, 7)
```

Utilizando la matriz de distancia “Hamming”, se aplica el algoritmo de aglomeración K-Means.

```
data$class_type[data_kmeans$id.med]
```

Se visualiza los centroides de cada uno de los clúster.

```
for (i in 1:ncol(data_norm))
data_norm[,i]=as.factor(data_norm[,i])
```

Permite discretizar cada una de las columnas disponibles en el dataset.

```
mca1 = MCA(data_norm, graph = FALSE)
```

Son transformados los datos dicotómicos a variables cuantitativas por medio de "Análisis de correspondencia múltiple", con la finalidad de aplicar sobre los datos, métricas de distancia como “Euclidean”, “Manhattan”, “Minkowski”, etc.

```
mca1 = MCA(data_norm, graph = FALSE)
```

Los datos dicotómicos disponibles en el dataset, son transformados a datos decimales.

```
data_kmeans <- pam(mca1$ind$cos2, 7, metric = "manhattan")
```

Se crea una cantidad determinada de clusters por medio de algoritmo K-Medias.

```
fviz_silhouette(silhouette(data_kmeans))
```

Gráfico de barras de siluetas (métrica de pertenencia de miembros a sus correspondientes clusters).

```
sil <- silhouette(data_kmeans)
neg_sil_index <- which(sil[, 'sil_width'] < 0.2)
sil[neg_sil_index, , drop = FALSE]
```

Visualización de miembros con alta probabilidad de encontrarse en el clúster equivocado.

```
fviz_cluster(data_kmeans, stand = FALSE, geom = "point",
frame.type = "norm")
```

Diagrama de Venn para los clusters creados por medio de algoritmo K-Means.

Extracción de conocimiento del DataSet ZOO

Normalización de los datos

Tal como se ha mencionado anteriormente, el dataset ZOO consta de un total de 101 animales con 16 características morfológicas. Todas esas características son descriptivas, es decir todos los valores disponibles son binarios, excepto la variable “patas” con valores desde 0 hasta 8.

Con la finalidad de poder normalizar todos los datos de la base de datos a un tipo en particular (binarias para este dataset), se separó la variable “patas” en 6 distintas variables, las cuales son transformadas a descriptivas, de esta forma de tener una base de datos con solo valores descriptivos.

Además de lo anteriormente mencionado, se presentan dos alternativas en la experimentación con los datos al momento de normalizarlos y luego comparar cada agrupación utilizando el coeficiente de medición entregado por la componente silhouette (silueta).

En la primera alternativa de normalización de los datos se eliminó las variables "animal_name" y "class_type", las cuales cumplen la función de individualizar a cada uno de los miembros (animales) del dataset:

En la segunda alternativa de normalización de datos se eliminó solamente la variable “animal_name”, pero por otro lado, se dividió la variable “class_type” en 7 nuevas variables con valores binarios. Tal como se había mencionado anteriormente, el objetivo de utilizar esta alternativa es entregar al algoritmo K-Medias más recursos discriminadores al momento de agrupar los animales en sus respectivos clusters.

De esta forma, solo se dispondrá de una matriz apropiada para aplicar distintas métricas de distancia según sea el caso, las cuales serán analizadas más a fondo en la siguiente sección.

Selección de métrica de distancia

Como un primer acercamiento a la métrica de distancia a utilizar, se analizarán clusters producidos por medio de un conjunto de métricas distintas. Tales métricas empleadas para el análisis son las siguientes:

- Manhattan
- Euclidea
- Minkowski
- Hamming

Para las primeras tres métricas mencionadas, fue necesario la conversión del dataset con valores cualitativos a una matriz de valores cuantitativos. Para realizar la conversión de datos cualitativos a cuantitativos se utilizó la técnica estadística llamada "análisis de correspondencia múltiple". Esta técnica es mencionada por Chavent M. (1998) con la finalidad de realizar una transformación de los datos ya existentes, en unos que permitan una mayor variedad de pruebas de métricas de distancia.

El Análisis de correspondencias es una técnica de reducción de dimensiones, una técnica para visualizar una nube de puntos multidimensional en dos dimensiones. Consiste, como las demás técnicas de reducción de dimensiones, en un procedimiento de traslado de una nube de puntos definida en un espacio de muchas dimensiones a un espacio de dos dimensiones donde poder visualizar la posición relativa de unos puntos. Este traslado se hará respetando al máximo las posiciones relativas de los puntos en la nube de puntos original. Esta técnica puede ser aplicada a datos cualitativos con la finalidad de obtener una matriz que describa características cualitativas por medio de variables cuantitativas.

Una vez obtenida la matriz con descriptores morfológicos cuantitativos, se aplica la función “pam”, que permite realizar clustering con un número predeterminado de agrupaciones.

Como primera instancia de acercamiento al problema de obtención de conocimiento por medio de la agrupación K-Medias, se emplearan las siguientes propiedades:

- Métrica de distancia: Manhattan
- Numero de clusters: 7 grupos

Como es posible observar, se ha decidido realizar siete clusters, los cuales se espera que cada una de estas agrupaciones, contenga miembros de cada una de las siete especies.

Es necesario mencionar que tanto para la matriz de datos normalizada separando la variable "class_type" en 7 nuevos campos, y la exclusión de esta variable, se obtuvo el mismo resultado de la misma especie de animal del centroide, el cual se muestra a continuación:

Tabla 1. Característica de los centroides de cada clúster

Clúster	Cantidad Miembros	Tipo de clase	Nombre de Clase	Nombre Centroide
[1]	13	1	Mamífero	Reno
[2]	16	1	Mamífero	Jabalí
[3]	12	4	Pez	Piraña
[4]	16	5	Anfibio	Rana
[5]	18	7	Invertebrado	Almeja
[6]	16	2	Ave	Flamenco
[7]	10	6	Insecto	Mosca

Como es posible observar en la tabla 1, efectivamente se crearon siete distintos clusters, sin embargo, por medio del análisis de los centroides o dicho de otra manera, los representantes de cada uno de los grupos, es posible inferir que los integrantes de cada uno de estos grupos serán del mismo tipo (nombre de clase) que su representante. Esto se infiere por variables con gran preponderancia encontradas en laboratorios anteriores tales como:

- Amamantar

- Nace por medio de huevos
- Son dentados, etc.

Por lo tanto, para los clúster 1 y 2, sus centroides son mamíferos (reno y jabalí), lo cual indica que se ha dividido al grupo de mamíferos en dos, de esta forma dejando sin agrupación visible a los animales del tipo de los reptiles.

Por otro lado, se ha efectuado otro análisis de los clusters generados con los siguientes parámetros:

- Métrica de distancia: Euclidiana
- Numero de clusters: 7 grupos

Por medio de la métrica de distancia euclidiana y la misma cantidad de clusters que el análisis anterior, se obtuvo los mismos datos sin ninguna variación con respecto a la métrica de distancia “Manhattan” (ver tabla 1).

Debido a lo anteriormente, se efectuó el análisis de las agrupaciones obtenidas por medio de los siguientes parámetros y de esta forma observar variaciones en los centroides de los grupos:

- Métrica de distancia: Hamming
- Numero de clusters: 7 grupos

La métrica “Hamming” está especialmente formulada con la finalidad de realizar mediciones de diferencias entre vectores dicotómicos, es decir variables cualitativas. Por tal motivo, para el caso de esta métrica no es necesario recurrir a la técnica de "análisis de correspondencia múltiple", ya que por medio de la normalización del dataset “Zoo”, solo se cuenta con variables dicotómicas y es posible utilizar “Hamming” de forma transparente y directa para los datos disponibles.

Debido a que el método “pam” solo puede aplicar métricas de distancia “Euclidiana” y “Manhattan”, no es posible entregar a este método la matriz de dataset Zoo. Para ello, es necesario realizar la normalización de tales datos. Para solucionar el problema anterior, se creó una matriz de distancia, la cual puede utilizar la métrica de distancia “Hamming”. De esta forma, el método “pam” realizará la agrupación predeterminada utilizando la matriz de distancia generada.

Como ya se ha mencionado, para este caso se ha decidido obtener siete clusters con métrica “Hamming”, sin embargo, al igual que con las distancias “Euclidiana” y “Manhattan”, los centroides no han cambiado significativamente, siendo de las mismas clases anteriormente mencionados (ver tabla 1), cambiando solamente el animal que caracteriza a los grupos.

Ahora bien, con la finalidad de realizar la agrupación de animales por clusters de manera más transparente, sin la necesidad de realizar numerosas normalizaciones de los datos presentes en el dataset, se ha escogido la métrica “Hamming”.

Por tal motivo, se ha tenido en consideración algunas observaciones, las cuales se mencionan en la siguiente sección.

Justificación de cantidad de clusters

Como se mencionó anteriormente, como primer acercamiento al problema de agrupación por medio del algoritmo K-Medias, la métrica de análisis de distancia escogida es “Hamming”, con un total de 7 agrupaciones correspondiente a cada uno de los tipos de animales y por otro lado, se ha excluido la variable “class_type” con la idea de ver cuán acertada es la predicción del agrupamiento de cada animal según su respectiva especie. De esta misma manera, los siguientes análisis de esta sección también incluyen las 7 subdivisiones de la variable “class_type”, encontrando de esta forma que con 7 clusters, no hay diferencia en la agrupación de especies, por lo que los siguientes análisis incluyen tanto a la exclusión de la variable mencionada, como también la división de esta en 7 nuevas variables.

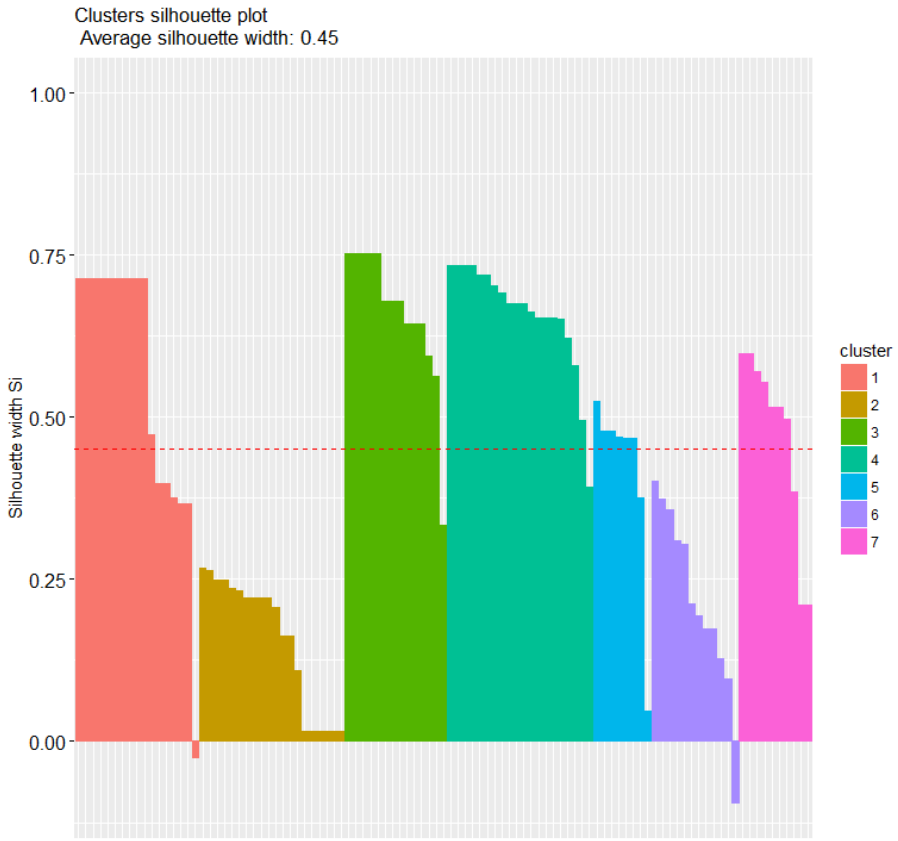

Ilustración 1. Gráfico de silueta de miembros de cada clúster excluyendo e incluyendo "class_type".

El objetivo general perseguido por las técnicas de clustering consiste en identificar grupos o clusters compactos, pero también se hace necesario la evaluación para medir la eficiencia de las técnicas de clustering aplicado. Para el caso de evaluación de clusters en nuestro trabajo utilizaremos el concepto de silueta de un clúster y la entrega de sus coeficientes de medición. Así pues, se tiene que mientras más cercana este la silueta del individuo a 1 (eje Y), quiere decir que el miembro del grupo tiene mayor afinidad al clúster asignado. Por el contrario, si la silueta del miembro del clúster está cercana a 0 (eje Y), quiere decir que tiene afinidad morfológica con otro grupo. Por otra parte, si la silueta del individuo es menor a 0 (eje Y), la probabilidad de que el miembro del clúster pertenezca a otro grupo es muy alta.

Además de lo mencionado anteriormente, esta técnica de medición de la eficiencia de clustering es apropiada para todas las métricas anteriormente mencionadas, lo que se ve respaldado por el trabajo efectuado por Yoon S. (2007), el cual trata acerca de la validación de clusters referentes a la agrupación de moléculas asociadas a funciones biológicas, donde los datos tratados fueron nominales y además la medición de calidad de las agrupaciones se efectuaron en base a la silueta promedio de la formación de clusters.

Teniendo esto presente, es posible inferir la calidad de los clusters mediante el análisis de medición de su silueta como es mostrado en la tabla 2.

Tabla 2. Miembros con silueta menor a 0.2 excluyendo e incluyendo "class_type".

ID Animal	Clúster	Clúster Vecino	Silueta
76	**[1]**	**[6]**	-0.02567631
37	**[2]**	**[1]**	0.16340873
95	**[2]**	**[1]**	0.16340873
30	**[2]**	**[1]**	0.11020649
2	**[2]**	**[1]**	0.01489273
6	**[2]**	**[1]**	0.01489273
18	**[2]**	**[1]**	0.01489273
23	**[2]**	**[1]**	0.01489273
29	**[2]**	**[1]**	0.01489273
56	**[2]**	**[1]**	0.01489273
73	**[5]**	**[7]**	0.04713528
75	**[6]**	**[1]**	0.19392192
20	**[6]**	**[3]**	0.17333612

Como es posible observar por medio de la tabla 2, todos los miembros del clúster 2 tienen una baja silueta, lo cual indica que las probabilidades son altas de que pertenezcan a otro clúster. El clúster al cual tienen mayor afinidad es al 1.

Los primeros dos clusters se componen de “mamíferos” (ver tabla 1). Por otra parte, todos los miembros del segundo clúster tienen una gran afinidad al centroide del primer clúster. Con esto es posible determinar la cantidad de clúster a trabajar para el análisis de tales agrupaciones.

Es posible inferir que con la finalidad de cumplir con la cantidad de clusters predispuestos para el método “pam”, los cuales son siete. Así pues, el tipo de animales “mamíferos” fueron subdivididos en dos grupos debido a que representan la gran mayoría de animales encontrados en el dataset “Zoo” (41 mamíferos).

Tal como fue posible determinar en el laboratorio anterior, las reglas más predominantes y significativas dentro del dataset están relacionadas con los “mamíferos”. Ante esto, el tipo animal más predominante fue subdividido en dos, tomando como centroide a dos mamíferos donde la gran diferencia radica en la presencia de pelaje entre ellos y si son depredadores.

Por otra parte, un grupo minoritario el cual es “reptil” fue absorbido por otros clusters. Detalles sobre esto se podrán ver en las siguientes secciones, sin embargo por el momento es posible inferir tal acción efectuada por el algoritmo K-Medias debido a que en el primer y segundo laboratorio se determinó que no fueron encontradas variables concluyentes para determinar de forma certera, animales pertenecientes a la especie “reptil”.

También, se pudo inferir que todos los miembros del segundo clúster en realidad tienen una baja silueta y que son muy próximos al primer clúster. En vista de lo anterior, se decidió experimentar para 6 agrupaciones, considerando los siguiente parámetros a utilizar sobre el método “pam” con el algoritmo K-Medias:

- Métrica de distancia: “Hamming”
- Numero de clusters: 6 grupos

Con el número reducido de clusters, se tienen los siguientes centroides para cada uno de los grupos excluyendo la variable “class_type”:

Tabla 3. Característica de los centroides de cada clúster excluyendo "class_type".

Clúster	Cantidad Miembros	Tipo de clase	Nombre de Clase	Nombre Centroide
[1]	36	1	Mamífero	Antílope
[2]	14	4	Pez	Piraña
[3]	20	2	Ave	Halcón
[4]	8	7	Invertebrado	Estrella de Mar
[5]	13	5	Anfibio	Tritón
[6]	10	6	Insecto	Mosca

Al disminuir la cantidad de clusters, es posible observar que los datos expuestos en la tabla 3 en comparación a los resultados obtenidos anteriormente, se visualiza que los miembros del clúster 2 habían sido absorbidos por el primer clúster, el cual también correspondía a "mamíferos". Por otra parte, es posible entender la poca capacidad de agrupación de "reptiles" ya que variables tales como nacen de huevos=1 y dan de amamantar=0, son características compartidas también con animales pertenecientes a las "aves", o también la variable "depredador = 1" es transversal para muchas especies, por lo que se concluye que con la finalidad de poder agrupar de forma más acertada la especie "reptil", es necesario más variables tales como "sangre caliente", "lengua bífida", etc.

Por otra parte, al analizar el grafico de silueta para 6 clusters se tiene:

Tabla 4. Siluetas menores a 0.2 para métrica "Hamming" y 6 clusters excluyendo "class_type".

ID Animal	Clúster	Clúster Vecino	Silueta
73	[4]	[6]	-0.02567631
20	[5]	[2]	0.16340873
67	[5]	[2]	0.16340873
81	[5]	[2]	0.11020649
63	[5]	[2]	0.01489273

76	**[5]**	**[1]**	0.01489273
64	**[5]**	**[1]**	0.01489273

Con los nuevos parámetros, se observa que algunos de los animales categorizados como “anfibios” tienen probabilidades de pertenecer a otras clasificaciones tales como “aves” y “mamíferos”. Por otra parte, al analizar la gráfica de silueta se tiene que:

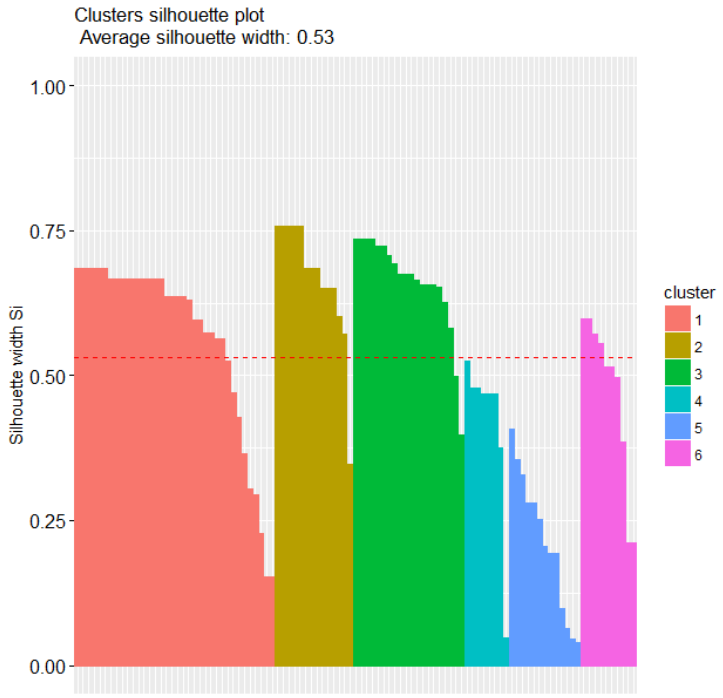

Ilustración 2. Gráfico de silueta de miembros de cada clúster para “Hamming” y 6 clusters.

Al analizar la ilustración 2, es posible observar que el clúster más numeroso, es el primero, el cual tiene como centroide a un “mamífero”. También se tiene que la silueta promedio del primer clúster subió a un 0.53, lo cual refleja que el modelo para reflejar a los “mamíferos” ha mejorado en gran manera.

Sin embargo, al observar al clúster 5 el cual tiene por centroide a un “anfibio”, es posible inferir que la silueta de este grupo es baja debido a que comparten grandes rasgos con animales de la especie de “reptiles”, los cuales parte de ellos fueron asimilados por este clúster y el de las “aves” (esta hipótesis será analizada en secciones posteriores).

Por otra parte, se ha realizado el mismo análisis constatando la silueta de los clusters generados, pero esta vez incluyendo las 7 subdivisiones de la variable “class_type”, de esta forma aportando más variables discriminatorias, con la intención de obtener una mayor afinidad de los miembros de cada grupo con sus respectivos clusters.

Por otra parte, es necesario comprender que se incluye el análisis del clustering añadiendo las variables referentes a “class_type” con la idea de contrastar tales resultados con los obtenidos por el clustering no incluyendo tales variables. Lo que se espera lograr es que el análisis de la clasificación de animales según características morfológicas (lo cual es el objetivo de este laboratorio) sea lo más próximo a la clasificación de animales dada en el dataset por medio de la variable “class_type”.

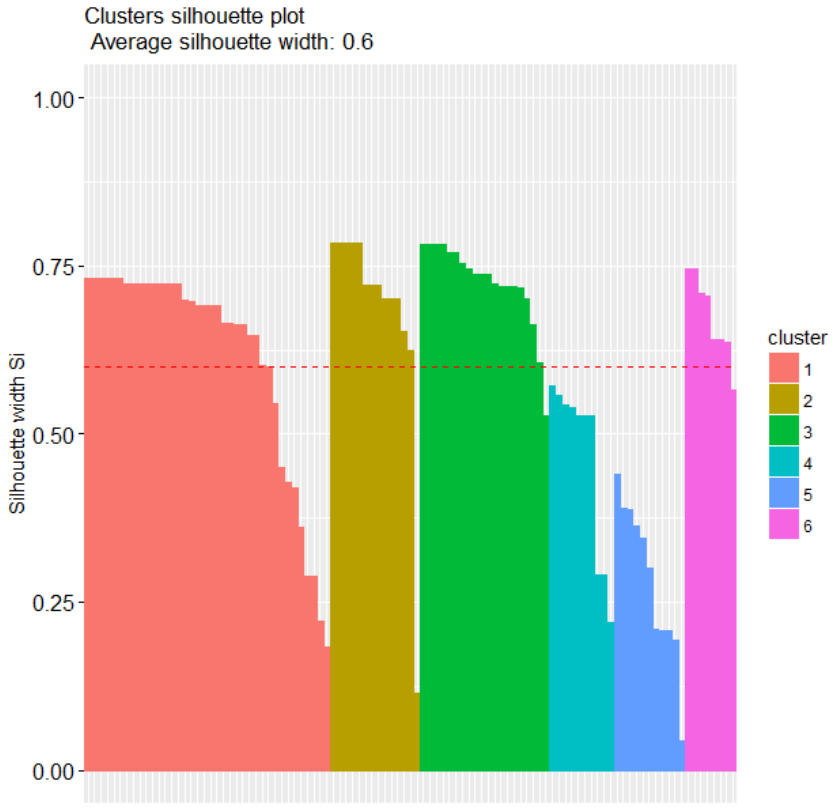

Ilustración 3. Gráfico de silueta de miembros de cada clúster para “Hamming” y 6 clusters incluyendo subdivisiones de “class_type”.

Tabla 5. Característica de los centroides de cada clúster incluyendo subdivisiones de "class_type".

Clúster	Cantidad Miembros	Nombre de Clase	Nombre Centroide
[1]	38	Mamífero	Lobo
[2]	14	Pez	Piraña
[3]	20	Ave	Halcón
[4]	10	Invertebrado	Estrella de Mar
[5]	11	Anfibio	Tritón
[6]	8	Insecto	Mosca

Al observar la ilustración 3, es posible observar que al incluir dentro de los datos las subdivisiones de la variable "class_type", la afinidad de los integrantes de cada grupo con sus respectivos centroides (medida de silueta) ha aumentado un 0,07 en comparación con la exclusión de la variable "class_type" en el análisis, lo cual indica que al incluir nuevos discriminantes de los datos para el proceso realizado por K-medias, se obtienen clusters más acertados.

Ante los análisis efectuados hasta ahora, es posible determinar que la utilización de la métrica "Hamming", 6 clusters para el algoritmo K-Medias es la manera más efectiva de agrupación para las alternativas evaluadas en este documento. Por otra parte, como se mencionó anteriormente, el análisis se efectuara sobre los parámetros mencionados, pero no agregando la variable "class_type" con la idea de contrastar los resultados de esta clase la cual es incluida en el dataset y los resultados obtenidos por medio del proceso de clustering

Análisis general de clusters

Para nuestro análisis también hemos utilizado el gráfico de agrupación de similitudes según morfología por medio del método "fviz_cluster" del paquete "FactoExtra", que permite visualizar de forma simplificada, caracterizaciones y agrupaciones de cada uno de los clusters. Algo realmente útil para tal fin, es poder analizar las similitudes entre especies por medio de la diagramación de Venn, como también incluso la posibilidad de analizar miembros en particular de cada uno de los clusters.

Para tal efecto, se presenta a continuación el siguiente grafico en conjunto con la caracterización de cada uno de los clusters según especie a la cual pertenecen los centroides:

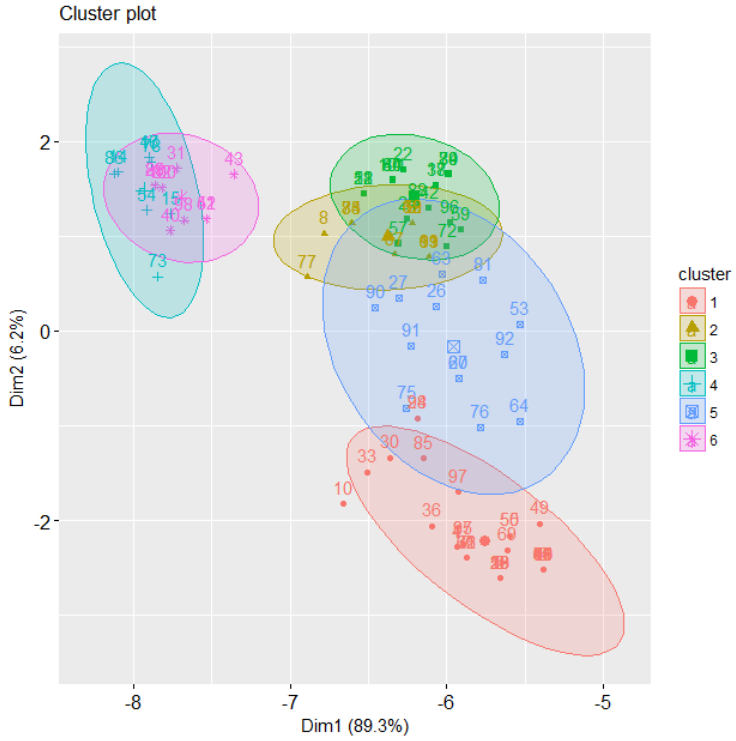

Ilustración 4. Diagrama de Venn para "Hamming" y 6 clusters identificando individuos y excluyendo variable "class_type".

Como es posible observar tanto en la ilustración 4 como en la ilustración 5, existen algunas leves diferencias en cuanto a la diferencia de sus clusters. Es de esperar que la inclusión de la variable "class_type" repercuta en cuan acertado son los clusters. Las mayores diferencias que se pueden aprecias entre estas dos ilustraciones es la disminución de la capacidad de agrupar mamíferos tales como ornitorrinco (identificación 64) o el lobo marino (identificador 76) dentro del grupo de los mamíferos. Estos dos casos mencionados se analizarán mas afondo en las próximas secciones de este documento.

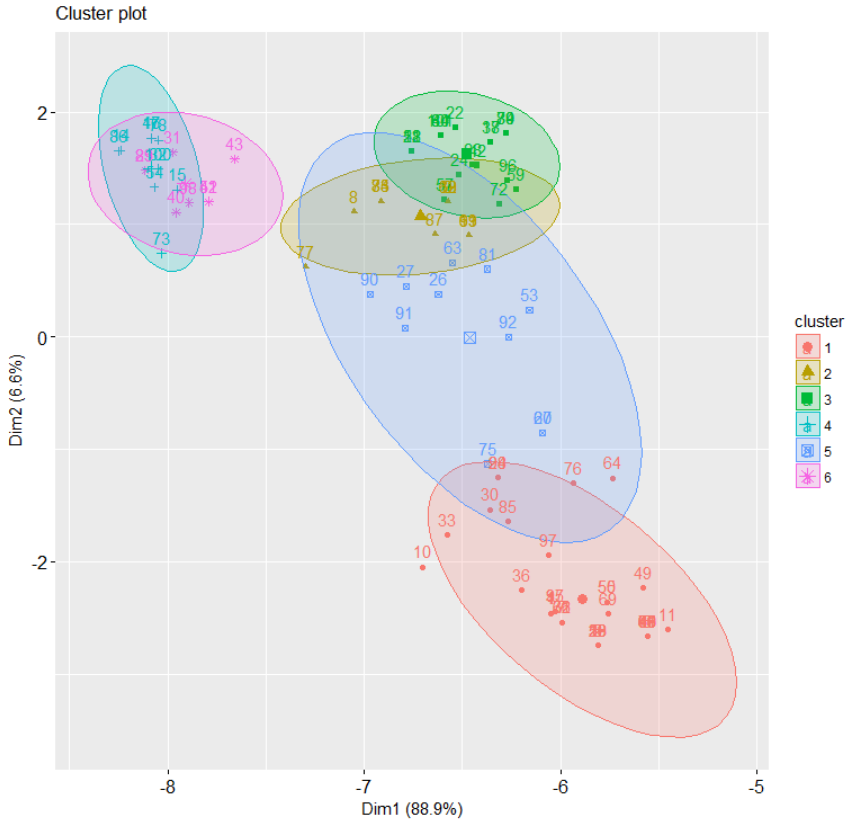

Ilustración 5. Diagrama de Venn para "Hamming" y 6 clusters identificando individuos e incluyendo variable "class_type".

Tabla 6. Clusters según tipos de animales para métrica "Hamming" y K-Medias con 6 K y excluyendo variable "class_type".

Clúster	Cantidad Miembros	ID clase	Nombre de Clase
[1]	38	1	Mamífero
[2]	14	4	Pez
[3]	20	2	Ave
[4]	10	7	Invertebrado
[5]	11	5	Anfibio
[6]	8	6	Insecto

Por medio de la ilustración 4 es posible inferir que la componente "Dim1" corresponde a la variable "Columna Vertebral" y "Amamanta", ya que al extremo izquierdo se encuentran grupos de animales tales como "invertebrados" e "insectos", los cuales no poseen "columna vertebral" y además no dan de "amamantar". Por otra parte, en el extremo derecho están los grupos de animales tales como "mamíferos", "reptiles", "aves" los cuales si poseen columna vertebral. Además, el grupo de animales que se encuentran más a la derecha de la gráfica, están los "mamíferos", los cuales son los únicos que "amamantan".

Por otro lado, la componente "Dim2" corresponde a las variables "Huevo" y "Pelaje", ya que en el extremo inferior se encuentra el clúster correspondiente a los "mamíferos", los cuales son los únicos que no nacen por medio de "huevos". Además, a medida que se asciende en el eje "Dim2", los animales presentes en los clusters poseen menos probabilidades de poseer "pelaje".

Por medio de la gráfica anterior, es posible inferir que los "mamíferos" y los "anfibios" no tienen gran cantidad de similitudes. Por tal motivo se puede suponer que las características en común son:

- Son vertebrados.
- Pulmonados.

Por otro lado, es posible encontrar algunas similitudes que no son comunes para todo el grupo de los “mamíferos”, los cuales son:

- Son acuáticos (mamíferos marinos).
- Nacen por medio de huevos (ornitorrinco).

También es posible observar que el grupo que tiene menor afinidad con el clúster vecino es el grupo de los “mamíferos”, debido a que como se mencionó anteriormente, no comparten muchas variables en común con las otras especies y además, las variables que se aplican sobre ellos son categóricas al momento de agruparlos. Tal es el caso de la variable “amamanta”, la cual solo aplica para la especie “mamífero”.

Además de lo anteriormente mencionado, es muy interesante los estudios realizados por James Fobes y Cynthia Smock (1981) que señala que hay una relación morfológicas muy poco estudiada hasta ese momento, la cual es la capacidad sensorial auditiva que presentan tanto los anfibios como mamíferos acuáticos. Este estudio indica que la capacidad auditiva presente en mamíferos acuáticos es muy similar a la de los anfibios, lo cual es comparable a la capacidad que tiene un animal terrestre de escuchar bajo el agua, en comparación a la capacidad auditiva en un medio no acuático. La capacidad sensorial auditiva tanto en medio acuático como no acuático, se desarrollaron en estas dos especies, lo cual es un indicio de la adaptación al medio en el cual se desenvuelven. Es posible ver esto reflejado en la unión entre los clúster mamífero y anfibio en la ilustración 4. Uno ejemplo de lo que se menciona es el mamífero “león marino” (identificación 76) el cual se encuentra justamente en la unión de estos dos clúster mencionados.

Además es posible inferir que el clúster correspondiente a los “peces” tiene gran afinidad con la agrupación de “anfibios”. Esto es posible comprenderlo al analizar las similitudes que tienen estas dos especies, las cuales algunas de ellas son:

- Nacen por medio de huevos.
- No amamantan
- Son acuáticos, etc.

Por tal razón, se observa que casi en su totalidad, el grupo de los "peces" está casi contenido dentro del grupo de los "anfibios".

Randall(1981) sugiere que muchas de las características en común que poseen estas dos especies, son debido a que la secuencia de evolución fue bastante directa desde los peces a los anfibios, ya que desde que el primer pez anfibio que se aventuró en ambiente no acuático (rhipidistian crossopterygians) hace 350 millones de año, señalo un punto de quiebre entre estas dos especies, lo cual sin embargo represento un punto de unión entre estas dos especies las que es posible ver incluso en la actualidad. Es por tal razón que estos dos clusters están muy unidos entre sí, pero por otro lado, muy bien definidos entre ellos.

También es posible inferir que el clúster correspondiente a la clase "ave" tiene cierta afinidad con las agrupaciones de los "anfibios" y "peces", ya que tiene algunas características en común tales como:

- Nacen por medio de huevos
- No poseen pelaje

Por otro lado, hay una variable la cual como ya se ha visto en laboratorios anteriores, es completamente concluyente en la clasificación de las "aves", la cual es "posee plumas". De esta forma, es posible inferir que la variable que hace posible la agrupación de las "aves", es tanto la capacidad de volar (variable "aéreo") como la capacidad de nacer por medio de "huevos".

Investigadores tales como Hakan Tegelstrom y Hans Rytman (1981) proponen que la similitud cromosómica entre las aves, peses y anfibios son grandes en comparación a tales especies con los mamíferos, lo que fundamenta similitudes morfológicas entre los reptiles, aves y anfibios, y porque estos fueron asimilados por los clusters de estas dos últimas

especies. Ante esto se ve fundamentada la unión minoritaria entre los clusters de las especies "mamíferos" y "anfibios".

Al analizar los clusters formados en el extremo superior izquierdo de la gráfica, es posible visualizar que los grupos correspondientes a las especies "invertebrados" e "insectos voladores", los cuales tienen gran afinidad entre sí. Esto es debido a que es posible encontrar algunas variables las cuales son comunes (vistas en laboratorio 1) entre estas dos especies, las cuales son:

- No poseen columna vertebral
- No poseen pelaje
- No poseen plumas
- No poseen dientes, etc.

Sin embargo, algunas de las variables que si se pueden identificar, las cuales permiten la agrupación de algunas de las especies presentes en el dataset son:

- Capacidad de volar
- Acuático (para algunos casos. No excluyente)

Tal como menciona Baran Mandal (2012), la especie "insecto" es una subcategoría de invertebrados, con lo cual se entiende porque el clúster "invertebrados" está muy bien relacionado con el clúster "insectos". Otra característica muy interesante que presentan los insectos, lo cual es mencionado por James Gould (1986), es la transversalidad de una característica que se aplica a la gran mayoría de insectos, es la capacidad de realizar mapas marcando rutas por medo de feromonas, lo cual es una de las grandes diferencias entre "insectos" y otros animales pertenecientes al grupo de los "invertebrados".

De esta forma, con la ayuda de la gráfica es posible inferir relaciones entre especies, las cuales son reflejadas por similitudes morfológicas e incluso transversales para todos los seres vivos presentes en el dataset "Zoo".

Algo muy particular y lo cual es pertinente mencionar y analizar, es el animal "tortuga" (identificación 91) el cual se puede visualizar en el grupo de los anfibios (clúster 5) en ilustración 4. Este animal cumple algunas características en común con los peces. Los investigadores Bickler y Buck (2006) sugieren semejanzas muy particulares entre estas dos especies, en concreto con el pez "carpa" el cual se muestra muy próximo a la tortuga (identificador "carpa" 8) en la ilustración 4. Las semejanzas entre estas dos especies son los mecanismos de supervivencia en ambos casos, los cuales incluyen grandes reservas de glucógeno y una disminución drástica del metabolismo.

Sin embargo, es probable que surja la interrogante, ¿las citas bibliográficas presentadas durante este análisis pueden tener alguna valides en contraste con las variables encontradas en el dataset "Zoo"? Esta pregunta se puede entender desde la perspectiva de semejanzas entre especies que el algoritmo K-Medias pudo encontrar, incluso cuando tales características morfológicas no se encontraban en el dataset "Zoo". Ante tal pregunta, es necesario recordar al gran científico inglés Charles Darwin, quien dice:

> (…) Que esta estrecha semejanza en tantos puntos de estructura tiene que explicarse, de conformidad con las opiniones expuestas por mí, por herencia de un antepasado en común (Darwin, 1859, p. 40).

Es decir, cada una de las especies presentes tienen un antepasado en común, pero a su vez, cada una de las especies tienen sus propios antepasados. Tal es el caso del animal "rhipidistian crossopterygians", el cual es considerado el primer anfibio que dio la pauta evolutiva y también de características morfológicas que tendrían sus especies sucesoras. Por tal motivo, es probable referirse a características morfológicas que son comunes entre especies en particular, e incluso transversales para un conjunto de ellas.

Por otra parte, en este documento se ha mencionado en algunas ocasiones que los centroides representan a los miembros de cada clúster. Es por eso que a cada clúster le fue asignado un tipo de animal en específico. Debido a tal inferencia, es posible corroborarlo a través del análisis de la ilustración 4, la cual muestra a animales agrupados en clúster, los que están identificados de forma individual. Esto ayuda en gran medida a corroborar la hipótesis

anteriormente planteada. Por medio de la siguiente tabla (muestra de miembros por clúster), es posible determinar que animales son agrupados en cada clúster:

Tabla 7. Clusters según tipos de animales para métrica "Hamming" y K-Medias con 6 K.

Clúster	ID Miembros	Nombre Miembro	Especie del Miembro
[1]	97	Wallaby	Mamífero
[1]	11	Leopardo	Mamífero
[1]	76	Lobo de Mar	Mamífero
[2]	8	Carpa	Pez
[2]	87	Mantarraya	Pez
[2]	77	Serpiente Marina	Reptil
[3]	72	Ñandú	Ave
[3]	96	Buitre	Ave
[3]	59	Pingüino	Ave
[4]	73	Escorpión	Invertebrado
[4]	78	Medusa	Invertebrado
[4]	15	Cangrejo	Invertebrado
[5]	90	Sapo	Anfibio
[5]	27	Rana	Anfibio
[5]	63	Serpiente Crótalo	Reptil
[6]	31	Mosquito	Insecto
[6]	40	Abeja	Insecto
[6]	43	Mariquita	Insecto

Tal como fue mencionado anteriormente, la hipótesis de asignar una especie de animal en particular a los clúster según tipo de centroides fue correcta, lo cual es corroborado por medio de las muestras de animales de cada clúster. Por otra parte, es posible ver que fueron encontrados dos tipos de "reptiles" dentro de las muestras. Ante esto es posible decir que las serpientes fueron distribuidas tanto al grupo de "peces", como de "anfibios" debido a grandes similitudes mencionadas con anterioridad.

Contraste de conocimiento de clusters con reglas de asociación

Es posible visualizar en la ilustración 4 que en su mayoría, los animales pertenecientes al clúster 1 (mamíferos) están contenidos dentro de su grupo, sin tener ningún rasgo en común con el clúster vecino el cuál es el 5 (anfibios). Esto fue posible corroborarlo en el análisis de reglas de asociación, ya que las reglas con un mayor lift representaban a "mamíferos". Este es el caso de la siguiente regla:

- {Da de amamantar} => {mamífero}

Lo cual se puede interpretar como la gran afinidad entre dar de "amamantar" y la clase "mamífero", la que se ve representada como en su mayoría, animales que no tienen ninguna relación con la clase "anfibios" ya que esta especie no da de "amamantar".

Por otra parte, es posible observar algunos de los "mamíferos" que se encuentran en la unión de la clase "mamíferos" y "anfibios", lo que se interpreta como animales que cumplen algunas similitudes con este último grupo pero que son "mamíferos". Un ejemplo de aquello son los siguientes animales:

- León marino (identificador 76)
- Ornitorrinco (identificador 64)

En el caso del león marino, cumple algunas semejanzas con los "anfibios", el cual es vivir en ambiente acuático (con un bajo lift para los "mamíferos" según diagrama de agrupación de reglas). En el caso del ornitorrinco, la gran diferencia con el centroide del clúster 1, es que este animal pone huevos, lo cual es inusual para esta especie. Esto se puede corroborar en la siguiente regla:

- {dan de amamantar, pone huevos} => { mamífero }

Por otra parte, se tiene al grupo de los “reptiles”, de los que no se obtuvieron reglas concluyentes o predominantes para su clasificación. Sin embargo, es posible observar a algunos de ellos en la ilustración 6. Tal como se dijo anteriormente en este documento, los animales de tipo “reptil” fueron absorbidos o asimilados por el clúster 5 (anfibios). Esto se ve reflejado en los siguientes animales:

- Tortuga (identificador 91)
- Serpiente crótalo (identificador 63)

Como se puede observar, estas dos especies, las cuales están identificadas como “reptiles” según el dataset, fueron asimiladas por el clúster con centroide de la especie “anfibio”, ya que se puede inferir que existe gran número de similitudes entre “anfibios” y “reptiles”.

Contraste de K-Medias con agrupación jerárquica

Durante el desarrollo del presente laboratorio, surgió la curiosidad de comparar los resultados obtenidos por medio del algoritmo K-Medias con la agrupación jerárquica de árbol (aun cuando no fue solicitado en este laboratorio), considerando 6 clusters y métrica de distancia "Hamming". Para ello se utilizaron los siguientes parámetros para la agrupación jerárquica:

- Agrupación jerárquica
- Corte del eje Y de dendrograma que permita obtener 6 clusters
- Métrica de distancia "Hamming"

Con los parámetros anteriormente mencionados, se obtuvo el siguiente dendrograma:

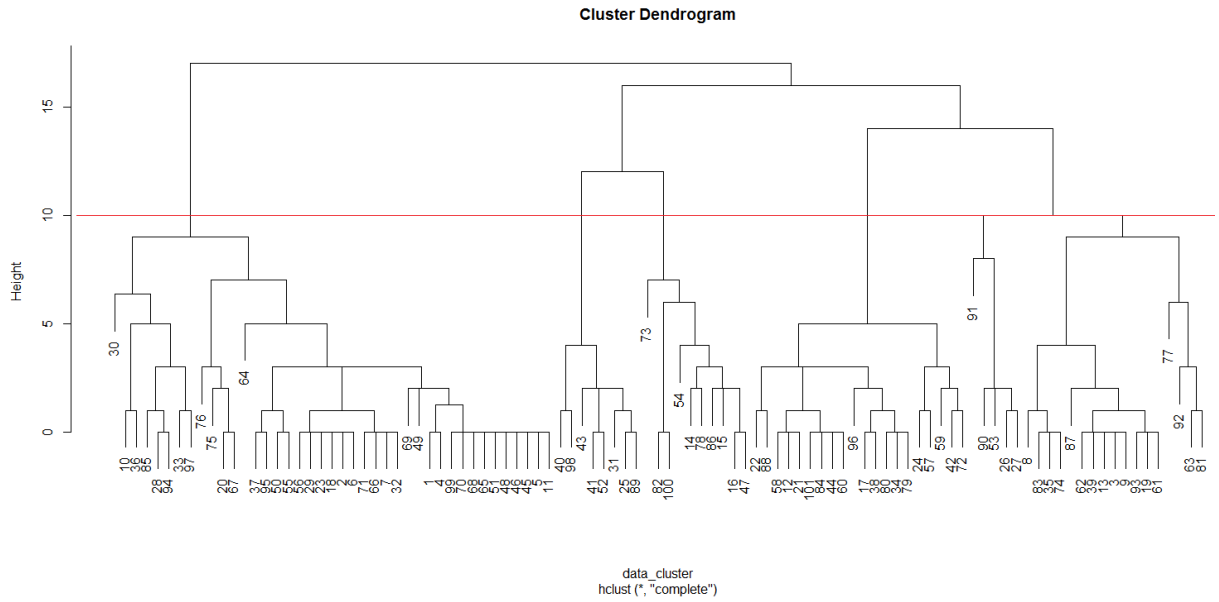

Ilustración 6. Dendrograma para "Hamming" y 6 clusters identificando individuos.

Se ha efectuado un corte del dendrograma en la distancia 10, con lo cual se obtuvieron 6 cluster, lo cual es justamente la cantidad de clusters con los que se efectuó la obtención de conocimiento en este laboratorio, de esta forma es posible compara la efectividad de cada uno de estos algoritmos. Los clusters obtenidos son los siguientes:

Tabla 8. Cantidad de animales por clúster para métrica "Hamming" y agrupamiento jerárquico con 6 K.

Especies/Clúster	[1]	[2]	[3]	[4]	[5]	[6]
Mamíferos	41	0	0	0	0	0
Aves	0	0	20	0	0	0
Reptiles	0	4	0	0	0	1
Peces	0	13	0	0	0	0
Anfibios	0	0	0	0	0	4
Insectos	0	0	0	0	8	0
Invertebrados	0	0	0	10	0	0
Total	***41***	***14***	***20***	***10***	***8***	***5***

Es posible observar que la cantidad de animales agrupados en el clúster de mamíferos coincide a la perfección con esta clase de especímenes disponibles en el dataset "Zoo". Lo mismo sucede con las aves, invertebrados e insectos.

En el caso de los reptiles para este algoritmo, se tiene un resultado similar a K-Medias, ya que fueron agrupados en los grupos de los peces y en su minoría (solo uno) con los anfibios.

Para el caso de los peces, la agrupación fue casi perfecta, ya que fue agregado a su grupo la tortuga, la cual pertenece a los reptiles.

Por otra parte, se obtuvieron los siguientes resultados de cuan acertado fue el algoritmo jerárquico en el proceso de clustering según especies de animales:

- Clúster 1: Tiene un coeficiente de validación sobre el 80%, lo cual indica que altamente probable que dicho clúster sea real, hace referencia principalmente a los mamíferos en donde todos tienen las características de amamantar, no nacen de huevos, no poseen alas y poseen espina dorsal.
- Clúster 2: Tiene un coeficiente de validación sobre el 86%, lo cual indica que altamente probable que dicho clúster sea real, hace referencia principalmente a los peces en donde todos tienen las características de nacen de huevos, poseen aletas, son acuáticos, no tienen patas y tienen espina dorsal.
- Clúster 3: Tiene un coeficiente de validación sobre el 62%, lo cual indica que es probable que gran parte de los integrantes del clúster sea real , hace referencia principalmente a las aves en donde todos tienen las características de que nacen de huevos, poseen plumas, tienen espina dorsal, no amamantan, tienen patas y no poseen dientes.
- Clúster 4: Tiene un coeficiente de validación sobre el 99%, lo cual indica que dicho clúster es real, hace referencia principalmente a los invertebrados en donde todos tienen las características de nacen de huevos, no tienen espina dorsal y no poseen dientes ni alas.
- Clúster 5: Tiene un coeficiente de validación sobre el 72%, lo cual indica que dicho clúster es probable gran parte del grupo tengan características de similitud en común, pero no todos hace referencia principalmente a los reptiles y anfibios en donde todos tienen las características de no tener pelos, no vuelan, no posee plumas ni aletas.
- Clúster 6: Tiene un coeficiente de validación sobre el 74%, es un clúster de un solo integrante, hace referencia principalmente a la estrella de mar (cataloga como invertebrado), está muy relacionada con el clúster 5, con la diferencia que no posee columna vertebral y tiene más de 4 patas.

Dado lo anterior, se obtuvo mejores resultados al aplicar sobre el dataset “Zoo” el algoritmo de agrupación jerárquica, que al utilizar K-Medias, ya que los grupos de animales

perfectamente agrupados con el algoritmo jerárquico fueron los mamíferos, aves, invertebrados e insectos, los que constituyen el 78,2% del universo de animales en el dataset. En contraste, las especies perfectamente agrupadas por K-Medias fueron invertebrados, insectos y peces, los que constituyen tan solo el 30,6% del total de animales presentes en el dataset “Zoo”.

Conclusiones

Respecto al objetivo general de este trabajo, que es la búsqueda de agrupaciones que permitan distinguir características comunes de cada grupo y relaciones con su clasificación original, se pudo comprobar que existen 6 cluster y 5 de ellos muy bien definidos, esto es, el cluster 1 (mamíferos) en donde todos sus integrantes tienen las características de amamantar, no nacen de huevos, no poseen alas y poseen espina dorsal; el cluster 2 (peces) en donde todos sus integrantes tienen las características de que nacen de huevos, poseen aletas, son acuáticos, no tienen patas y tienen espina dorsal; el cluster 3 (aves) en donde todos tienen las características que nacen de huevos, poseen plumas, tienen espina dorsal, no amamantan, tienen patas y no poseen dientes; el cluster 4 (Invertebrados) en donde todos tienen las características que nacen de huevos, no tienen espina dorsal y no poseen dientes ni alas; el cluster 5 (anfibios y reptiles) todos hace referencia principalmente a los reptiles y anfibios en donde todos sus miembros tienen las características de no tener pelos, no vuelan, no posee plumas ni aletas; por último el cluster 6 (insectos) es absorbido por otro cluster (aves) principalmente por las características de nacen por huevos, tienen patas, vuelan y no poseen dientes.

Respecto a la comparación con los otros laboratorios, se puede inferir que se destacan las variables con gran preponderancia encontradas en laboratorios anteriores y que también aparecen en los cluster tales como: amamantar, nace por medio de huevos y son dentados, que definen las clasificación para muchas especies. Respecto al laboratorio 2, en donde las reglas más predominantes y significativas dentro del dataset están relacionadas con los "mamíferos", también en la generación de los cluster aparece este grupo como el más predominante.

Una de las grandes cuestiones a resolver al momento de iniciar el proceso de creación de clusters, fue la selección de la métrica de distancia a utilizar con K-Medias. Para tal efecto,

se hizo necesario tomar en cuenta la naturaleza de los datos disponibles en el dataset (variables cualitativas o cuantitativas) y su posterior normalización.

En el caso del dataset “Zoo”, los datos que se encuentran en él, son dicotómicos, por lo que una alternativa factible que fue considerada es métrica “Hamming” la cual se expresa como la diferencia de variables cualitativas entre vectores. Sin embargo, esta posibilidad no fue empleada sin antes haber corroborado el resultado de agrupación del algoritmo K-Medias en conjunto con otras métricas. Es decir, el proceso de justificación fue a través de ensayo y error, determinando el resultado de aplicar cada uno de ellos.

Las métricas utilizadas para este estudio fueron, “Manhattan”, “Euclidea” y finalmente “Hamming” la cual fue aplicada para el análisis y estudio de clusters.

Por otra parte, se siguió el mismo proceso para determinar el número de clusters, ya que el algoritmo K-Medias requiere que la cantidad de clusters sea entregado de antemano. Con este proceso se determinó que la cantidad de clusters óptimos son 6. Y precisamente en este punto se refleja el mayor inconveniente del algoritmo K-Medias en agrupar los animales según su especie, ya que haciendo uso de las posibilidades entregadas por el método “pam”, no fue posible realizar la agrupación de “reptiles”. Sin embargo, finalmente se corrobora que fueron asimilados por el grupo de “anfibios”, lo que de cierta manera revindica al algoritmo, ya que es el grupo que posee mayores similitudes con “reptiles”.

Por último, se pudo comprobar la gran exactitud que puede proveer el algoritmo K-Medias en conjunto a “Hamming” al analizar algunas especies en particular, las cuales se encontraban en las uniones de los grupos. Tal fue el caso del león marino, el cual es uno de los pocos del dataset que son acuáticos, como también el ornitorrinco, el cual fue muy bien representado al incluirlo en los límites del clúster “mamíferos” y el clúster “anfibios” debido a que comparte con esta especie la capacidad de nacer por medio de huevos.

Debido a lo anteriormente mencionado, se cumplen en gran medida la exactitud buscada para el algoritmo K-Medias por medio de la cantidad de clusters y la métrica de distancia, como también una agrupación bastante aproximada a un óptimo global en la agrupación de animales según sus respectivas especies, lo cual fue apoyado por conceptos proveniente de bibliografía referente a las ciencias de la biología.

Durante este laboratorio fue mencionado en muchas ocasiones que los clusters representaban especies en común, lo cual se infirió por la clase de animal a la cual pertenecían los centroides. Esto se pudo ver reflejado en el análisis dispuesto en el diagrama de Venn, el cual proveyó de información para determinar a cuales especies albergaba cada clúster, de esta forma reafirmando que el resultado obtenido del algoritmo K-Medias fue de gran valor para obtener conocimiento del dataset "Zoo".

Finalmente, como una acción impulsada por la curiosidad, se decidió estudiar la diferencia de los resultados obtenidos por medio del algoritmo K-Medias y la agrupación jerárquica, lo cual sorprendentemente indico que este último algoritmo entrego resultados que por mucho, supero la calidad de clustering sobre el dataset "Zoo" realizada por K-Medias.

8 Referencias

1. Tao li ,T.L. (2010). A General Model for Clustering Binary Data. *School of Computer Science, Florida International University*.

2. Guntur, V. (2003). Simultaneous Pattern and Data Clustering Using Modified K-Means Algorithm*. International Journal on Computer Science and Engineering*. Vol. 02, No. 06

3. Universidad de Granada. (2017). Métodos de Análisis Multivariante: Análisis Clúster. *Retrieved 3 May, 2017, from http://wpd.ugr.es/~bioestad/guia-spss/practica-8/.*

4. James, L. & Smock, C.(1981). Sensory capacities of marine mammals. *Psychological Bulletin*, Vol 89(2).

5. Randall D.J., Burggren W.W., Farrell A.P. & Haswell M.S. (1981). The Evolution of Air Breathing in Vertebrates. *Cambridge University Press*, p 133.

6. Hakan T.& Hans R. (1981). The Evolution of Air Breathing in Vertebrates*. Cambridge: Department of General Genetics, University of Uppsala, Sweden*.

7. Baran F. (2012). Invertebrate Zoology*. Bankura, Weste: Department of Zoology, Bankura Chistian College.*

8. Gould L. (1986). The locale map of honey bees: do insects have cognitive maps?*. Science*, vol. 232, p. 861.

9. Bickler P. & Buck L. (2007). Hypoxia Tolerance in Reptiles, Amphibians, and Fishes: Life with Variable Oxygen Availability. *Annual Review of Physiology*.

10. Darwin C. (1859) The Origin of Species by Means of Natural Selection. *London*, p 44.

11. Henning C. (2006). Cluster-wise assessment of cluster stability. *Research Report 271, Dept. of Statistical Science, University College London.*

12. Chavent M. (1998). A monothetic clustering method*, Pattern Recognition Letters*.

13. Yoon S. & Ebert J. (2007). Clustering protein environments for function prediction: finding PROPOSITE motifs in 3D. *BMC Bioinformatics*.

14. Rousseeuw P.(1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

9 Apéndice

Script

```
setwd(choose.dir())

install.packages('cluster', dep = TRUE)
library(cluster)

install.packages('factoextra', dep = TRUE)
library(factoextra)

install.packages('FactoMineR', dep = TRUE)
library(FactoMineR)

install.packages('ggplot2', dep = TRUE)
library(ggplot2)

data_norm  =  read.csv("zoo-cluster - class_type.csv")
data  =  read.csv("zoo-cluster - class_type.csv")

data_norm["animal_name"] <- NULL

#-----------------------------------------------------------------
#modelo con 7 clusters
#Distancia Hamming
data_cluster <- dist(data_norm, method = "binary", diag = FALSE, upper = FALSE, p = 2)

data_matrix <- as.matrix(data_cluster)

data_kmeans <- pam(data_matrix, 7)

data$class_type[data_kmeans$id.med] # centroides con distancia hamming
#-----------------------------------------------------------------
#Distancia manhattan o euclidea
# datos dicotomicos a variables cuantitativas por medio de "multiple correspondence analysis"
for (i in 1:ncol(data_norm)) data_norm[,i]=as.factor(data_norm[,i])

mca1 = MCA(data_norm, graph = FALSE)

data_matrix <- as.matrix(mca1)
#hacer tabla que explique a que especie corresponde cada cluster (para las dos distancias ("hamm
ing" y "manhattan"))
data_kmeans <- pam(mca1$ind$cos2, 7, metric = "manhattan")
```

```
data$class_type[data_kmeans$id.med]
#-----------------------------------------------------------------
#Desaprobar 7 clusters para clasificar dataset
fviz_silhouette(silhouette(data_kmeans))
sil <- silhouette(data_kmeans)
# se seleccionan los animales con altas probabilidades de  pertenecer a otro cluster
neg_sil_index <- which(sil[, 'sil_width'] < 0.2)
sil[neg_sil_index, , drop = FALSE]
#-----------------------------------------------------------------
#modelo con 6 clusters
data_cluster <- dist(data_norm, method = "binary", diag = FALSE, upper = FALSE, p = 2)

data_matrix <- as.matrix(data_cluster)

data_kmeans <- pam(data_matrix, 6)#se ha elegido 6 clusters
data$class_type[data_kmeans$id.med] # centroides con distancia hamming
data[data_kmeans$id.med,23:29]#informacion centroides

#Grafica de siluetas
fviz_silhouette(silhouette(data_kmeans))

#Agrupaciones sin nombres
fviz_cluster(data_kmeans, stand = FALSE, geom = "point", ellipse.type = "norm", show.clust.cent =
TRUE)
# Agrupaciones de animales en particular
fviz_cluster(data_kmeans, stand = FALSE, data = data_norm, ellipse.type = "norm", show.clust.cen
t = TRUE)
#Graficas de agrupacion y silueta
plot(data$class_type, col = data_kmeans$cluster)
plot(data[data_kmeans$id.med,23:29], col = data_kmeans$cluster)
#Informacion global de clusters
summary(data_kmeans)
plot(data_kmeans)
names(data_kmeans)
table(data_kmeans$clustering)

#BONUS-Algoritmo jerarquico
data  =  read.csv("zoo-cluster.csv")
data_norm["animal_name"] <- NULL
data_cluster <- dist(data_norm, method = "manhattan", diag = FALSE, upper = FALSE, p = 2)

clusters <- hclust(data_cluster)
plot(clusters)
```

```
clusterCut <- cutree(clusters, 6) # a distancia 0.78 en el dendrograma
table(clusterCut, data$class_type)
plot(clusterCut, data$class_type)
```