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

ANÁLISIS DE DATOS: LABORATORIO 1

ANÁLISIS ESTADÍSTICO

PABLO CÁCERES LUZANTO
BENJAMÍN HERNÁNDEZ CORTÉS

Profesor:
Max Chacón
Ayudante:
Adolfo Guzmán

Santiago - Chile
30 de abril de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

En términos generales, la utilización de la Base de Datos *"Zoo"* está dada por la incansable búsqueda de un algoritmo o método de aprendizaje automático por parte de Richard Forsyth (Investigador británico que ha trabajado en los últimos años en las universidades de Warwick, Nottingham, Loughborough, Southampton y Leed).

En 1980, este investigador ideó un software denominado *PC/BEAGLE* (*Biological Evolutionary Algorithm Generating Logical Expressions*), para averiguar si una analogía basada en la evolución Darwiniana, que denominó "Selección Naturalística", podía ser un algoritmo de aprendizaje automático viable. A pesar de su modesto éxito comercial, su software representó un hito en el campo de *Machine Learning*. [1]

A lo largo de los años, el software ha recibido dos actualizaciones. En 1985, decidió escribir su software en Turbo-Pascal bajo MS/DOS. Pero es la última versión del programa la que ha tenido mayor éxito, puesto que es la que más se adapta a la mundo moderno, ya que el código fuente fue escrito en *Python3*.

En cuanto a la motivación por parte de quienes escriben este documento, se encuentra el entender en su totalidad la base de datos donada por el investigador mencionado en párrafos anteriores, tanto sus atributos, como los valores que estos toman en cada una de sus instancias.

Es por todo lo escrito anteriormente, que para empezar a trabajar en dicho contexto, es necesario en primer lugar analizar de manera profunda y minuciosa la base de datos, utilizada por el autor para probar sus métodos, aplicando análisis inferencial y estadístico.

1.2 ORGANIZACIÓN DEL DOCUMENTO

El presente documento está estructurado básicamente en tres grandes capítulos. Uno de ellos corresponde a la Descripción del Problema, en donde se aborda la problemática a tratar y se describe detalladamente la base de datos utilizada para tratar el problema. Por otra parte, se tiene el capítulo de Análisis de los datos, en donde, tal como su nombre lo indica, se analizan los datos mediante técnicas estadísticas ya conocidas (vistas en el ramo de Inferencia y Modelos Estadísticos). Finalmente, se presentan las Conclusiones en donde se expone lo aprendido durante el desarrollo del experimento.

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 análisis estadístico 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. DESCRIPCIÓN DEL PROBLEMA

2.1 DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos en estudio ha sido donada y creada por Richard S. Forsyth en el año 1990. De acuerdo a lo descrito en *UCI Machine Learning Repository*, esta consta de 101 registros de animales de zoológico, centrado en particular en vertebrados, incluso mamíferos, por ende no es representativa de la biosfera en general. En particular, la base de datos consta de 7 clases de animales y 17 atributos entre los que se pueden encontrar del tipo booleano (en su mayoría) y también del tipo numéricas. En la siguiente sección, de manera detallada, se describen estos.

2.2 DESCRIPCIÓN DE CLASES Y VARIABLES

Como se menciona en la sección anterior, la base de datos consta de 7 clases de animales, las cuales están estructuradas de la siguiente forma

Valor	Clase	Instancias	Animales
1	Mamíferos	41	Oso Hormiguero (aardvark), Antílope (antelope), Oso (bear), Jabalí (boar), Búfalo (buffalo), Becerro (calf), Conejillo de Indias (cavy), Guepardo (cheetah), Venado (deer), Delfín (dolphin), Elefante (elephant), Murciélago de la Fruta (fruitbat), Jirafa (giraffe), Niña (girl), Cabra (goat), Gorila (gorilla), Hámster (hamster), Liebre (hare), Leopardo (leopard), León (lion), Lince (lynx), Visón (mink), Topo (mole), Mangosta (mongoose), Zarigüeya (opossum), Antílope Oryx (oryx), Ornitorrinco (platypus), Turón (polecat), Pony (pony), Marsopa (porpoise), Puma (puma), Gato Doméstico (pussycat), Mapache (raccoon), Reno (reindeer), Foca (seal), León Marino (sealion), Ardilla (squirrel), Murciélago (vampire), Ratón de Campo (vole), Macrópodo (wallaby), Lobo (wolf)

Cuadro 2.1: Clasificación de animales según su clase 1/2

Valor	Clase	Instancias	Animales
2	Aves	20	Pollo (chicken), Cuervo (crow), Paloma (dove), Pato (duck), Flamenco (flamingo), Gaviota (gull), Halcón (hawk), Kiwi (kiwi), Alondra (lark), Avestruz (ostrich), Periquito (parakeet), Pingüino (penguin), Faisán (pheasant), Ñandú (rhea), Rayador Americano o Pico de Tijera (skimmer), Skúa (skua), Gorrión (sparrow), Cisne (swan), Buitre (vulture), Reyezuelo (wren).
3	Reptiles	5	Víbora (pitviper), Serpiente de Mar (seasnake), Gusano Lento (slowworm), Tortuga (tortoise), Tuátara (tuatara)
4	Peces	13	Róbalo (bass), Pez Carpa (carp), Bagre (catfish), Pez Cacho (chub), Cazón (dogfish), Eglefino (haddock), Arenque (herring), Pez Lucio (pike), Piraña (piranha), Caballo de Mar (seahorse), Lenguado (sole), stingray, Atún (tuna).
5	Anfibios	4	Rana (frog), Rana (frog), Tritón (newt), Sapo (toad).
6	Insectos	8	Pulga (flea), Mosquito (gnat), Abeja (honeybee), Mosca doméstica (housefly), Mariquita (ladybird), Polilla (moth), Termita (termite), Avispa (wasp).
7	Otros	10	Almeja (clam), Cangrejo (crab), Cigalas (crayfish), Langosta (lobster), Pulpo (octopus), Escorpión (scorpion), Medusa de Caja o Avispa de Mar (seawasp), Babosa (slug), Estrella de Mar (starfish), Gusano (worm).

Cuadro 2.2: Clasificación de animales según su clase 2/2

A continuación, se describen los 17 atributos de los cuales consta la base de datos, con su respectivo tipo y descripción.

Variable (Atributo)	Tipo	Descripción
Nombre (name)	Identificador	Indica el nombre de la especie.
Pelo (hair)	Booleana	Indica si el animal tiene pelo.
Plumas (feathers)	Booleana	Indica si el animal tiene plumas.
Huevos (eggs)	Booleana	Indica si el animal pone huevos.
Leche (milk)	Booleana	Indica si el animal da leche a sus crías.
Aerotransportado (airborne)	Booleana	Indica si el animal tiene la posibilidad de volar.
Acuático (aquatic)	Booleana	Indica si el animal (en algún momento) vive en el agua.
Depredador (predator)	Booleana	Indica si el animal come carne i.e. no es herbívoro.
Dentado (toothed)	Booleana	Indica si el animal tiene dientes.
Columna vertebral (backbone)	Booleana	Indica si el animal tiene espina dorsal.
Respira (breathes)	Booleana	Indica si el animal respira aire.
Venenoso (venomous)	Booleana	Indica si el animal produce veneno.
Aletas (fins)	Booleana	Indica si el animal tiene aletas.
Patas (legs)	Numérica	Indica el número de patas del animal.
Cola (tail)	Booleana	Indica si el animal tiene cola.
Domestico (domestic)	Booleana	Indica si el animal ha sido domesticado.
Tamaño de Gato (catsize)	Booleana	Indica si el animal es al menos tan grande como un gato doméstico.
Tipo (type)	Numérica	Indica la clase a la cual pertenece un animal.

Cuadro 2.3: Descripción de variables

Cabe destacar que todos los atributos al ser booleanos (a excepción de los atributos nombre, patas y tipo) son representados a través de un 1 (presenta un atributo) o 0 (no presenta un atributo).

CAPÍTULO 3. ANÁLISIS DE LOS DATOS

En primer lugar, en el siguiente gráfico, se procede con un análisis general de los datos en donde se pueden observar los distintos porcentajes del total de instancias que presentan los atributos booleanos.

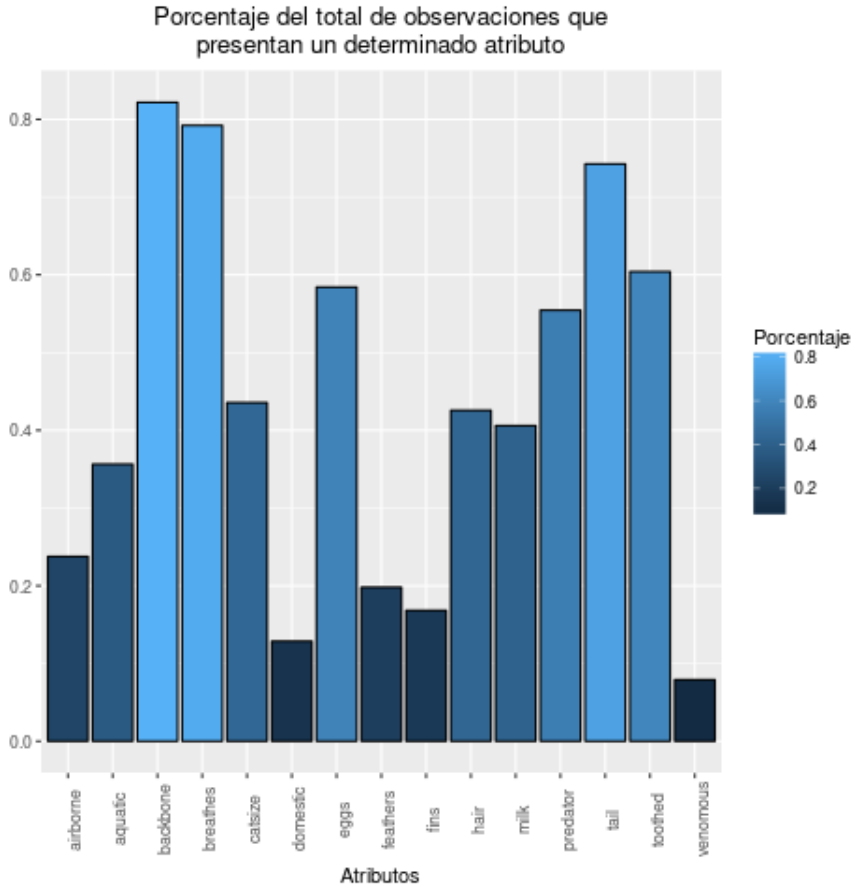

Figura 3.1: Porcentaje total de observaciones que presenta un atributo

Como se aprecia en el gráfico anterior, dentro de los atributos más frecuentes se encuentran las variables que indican presencia de columna vertebral (backbone), cola (tail) y si el animal respira aire (breathes). Esto parece lógico, pues, cabe recordar que la muestra está cargada hacia los animales mamíferos y estos en su mayoría (al menos en la muestra) presentan cola, columna vertebral y respiran aire.

A continuación, yendo hacia lo específico se presenta una serie de figuras que indican el porcentaje de observaciones de cada atributo, pero en esta oportunidad, se hace la distinción para las 7 clases de animales presentes en la base de datos.

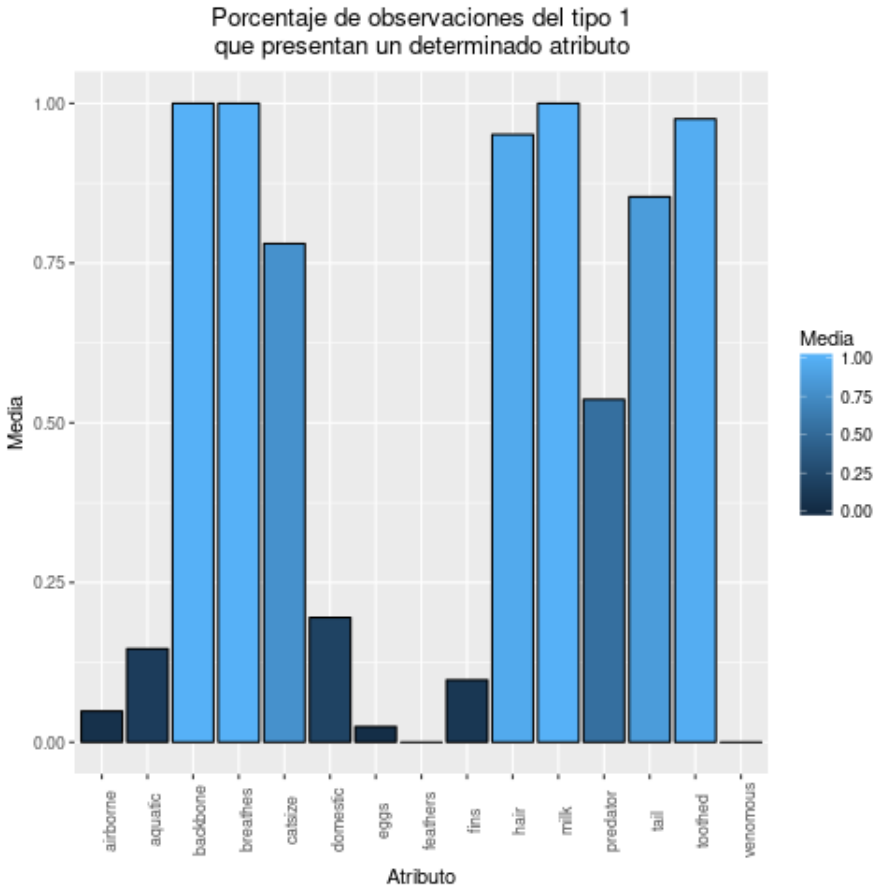

Figura 3.2: Porcentaje total de observaciones que presenta un atributo en Mamíferos

De acuerdo a lo expuesto por el gráfico de arriba, es posible inferir que los Mamíferos, en su totalidad, poseen al menos tres atributos: columna vertebral (backbone), respiran aire (breathes) y dan leche a sus crías (milk). Además, es posible deducir, que en su gran mayoría poseen dientes (toothed) y pelo (hair). Muy por el contrario, es posible concluir que no existen mamíferos que posean plumas (feathers) en la base de datos.

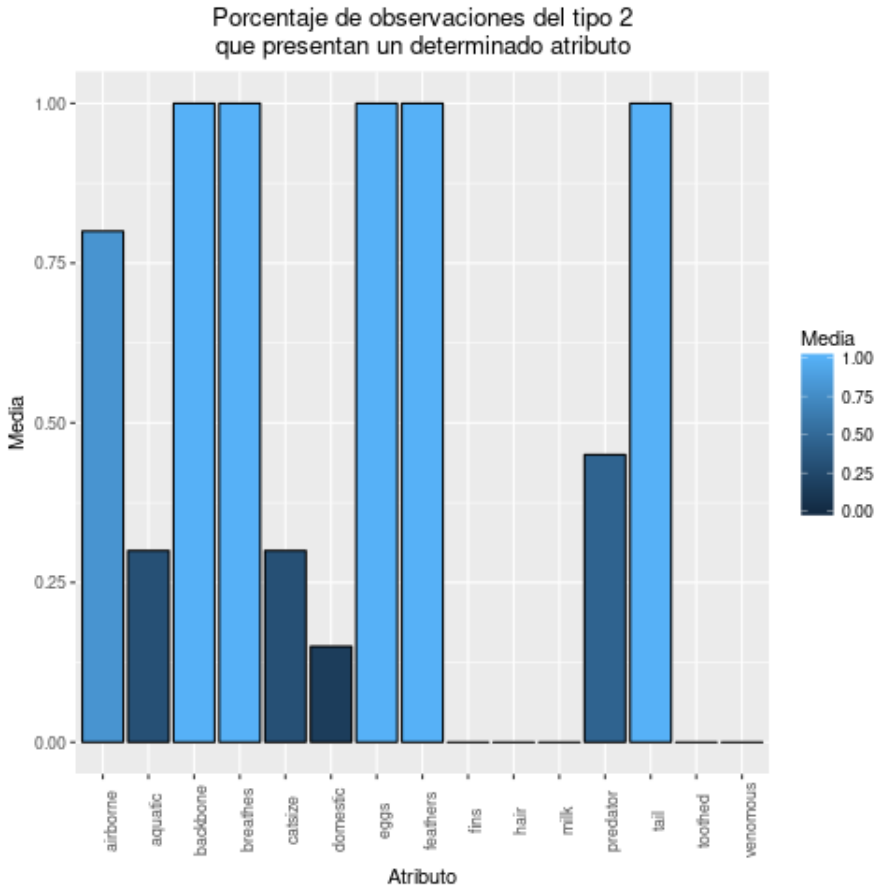

Figura 3.3: Porcentaje total de observaciones que presenta un atributo en Aves

En lo que respecta a la categoría de Aves y de acuerdo a lo que se observa del gráfico 3.3, se tiene que en su totalidad, las aves contienen 5 atributos que las definen como categoría en sí: poseen columna vertebral (backbone), respiran aire (breathes), ponen huevos (eggs), poseen alas (feathers) y tienen cola (tail). Además, observando el gráfico todo cobra sentido, ya que ningún ave posee aletas, ni pelo, ni son dentadas y mucho menos venenosas, lo que viene a confirmar de que se está en presencia de una base de datos verosímil.

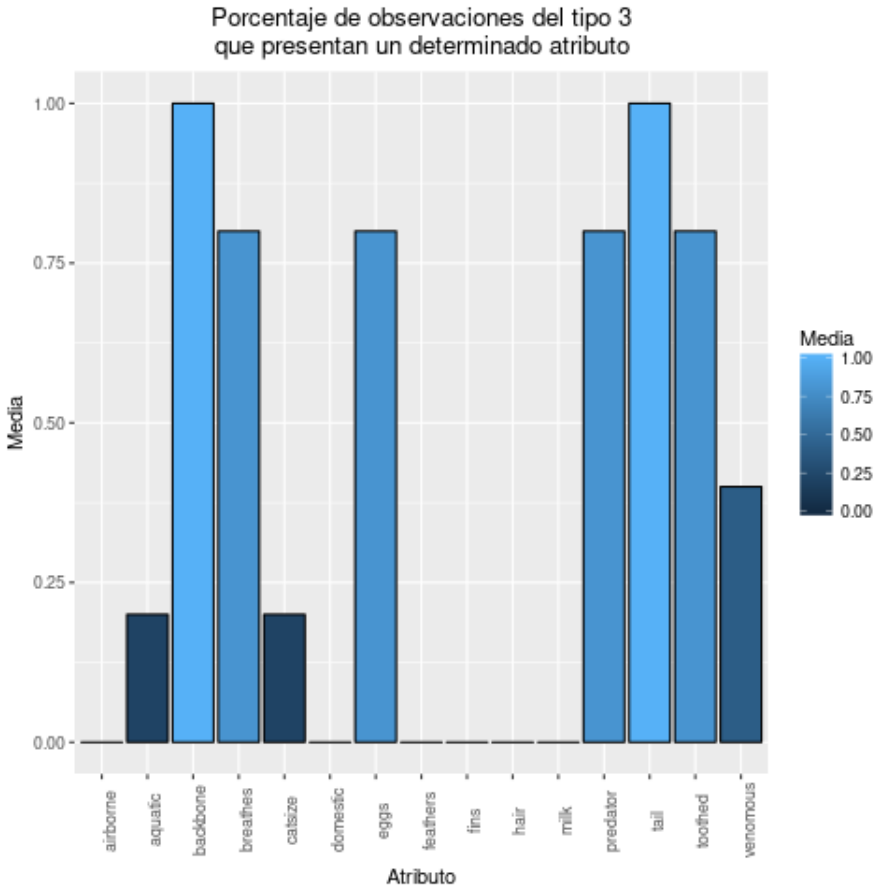

Figura 3.4: Porcentaje total de observaciones que presenta un atributo en Reptiles

En el caso del gráfico 3.4, es posible desprender de este, que básicamente son 2 los atributos que definen en su totalidad a un reptil, a saber: poseen columna vertebral (backbone) y cola (tail). Si, por ejemplo, se quisiera obtener ciertos animales con atributos como los que componen a los reptiles en su totalidad, habría cierto inconveniente para diferenciarlos de los mamiferos u otra especie que contenga estos atributos como componentes principales, por lo que es en ese momento en donde cobra vital importancia los atributos ”secundarios”que componen a esta clase de animales, que son: respiran aire (breathes), ponen huevos (eggs), son depredadores (predator) y poseen dentadura (toothed). De esta forma, se podría diferenciar una especie de otra (principalmente por los atributos de Depredador y Huevos).

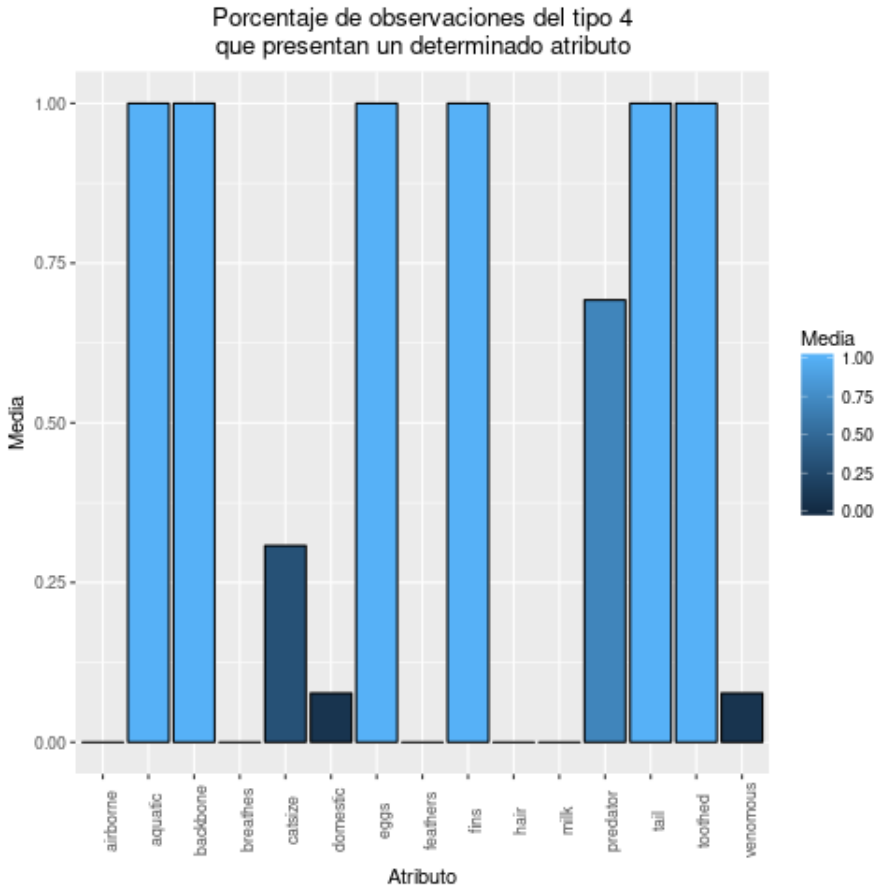

Figura 3.5: Porcentaje total de observaciones que presenta un atributo en Peces

En este caso, de la figura 3.5 se puede desprender que hasta el momento, las especies confinadas a la clase Peces, son las que poseen más atributos (6) que las definen en un ciento por ciento. Dichos atributos corresponden a: Ser acuáticos o vivir en algún momento en el agua (aquatic), poseer columna vertebral (backbone), poner huevos (eggs), tener aletas (fins), tener cola (tail) y ser criaturas dentadas (toothed). Esto facilitaría una posible consulta a la base de datos con el fin de obtener animales de la clase Peces, ya que existen muchos atributos que los definen a sí mismo como especie.

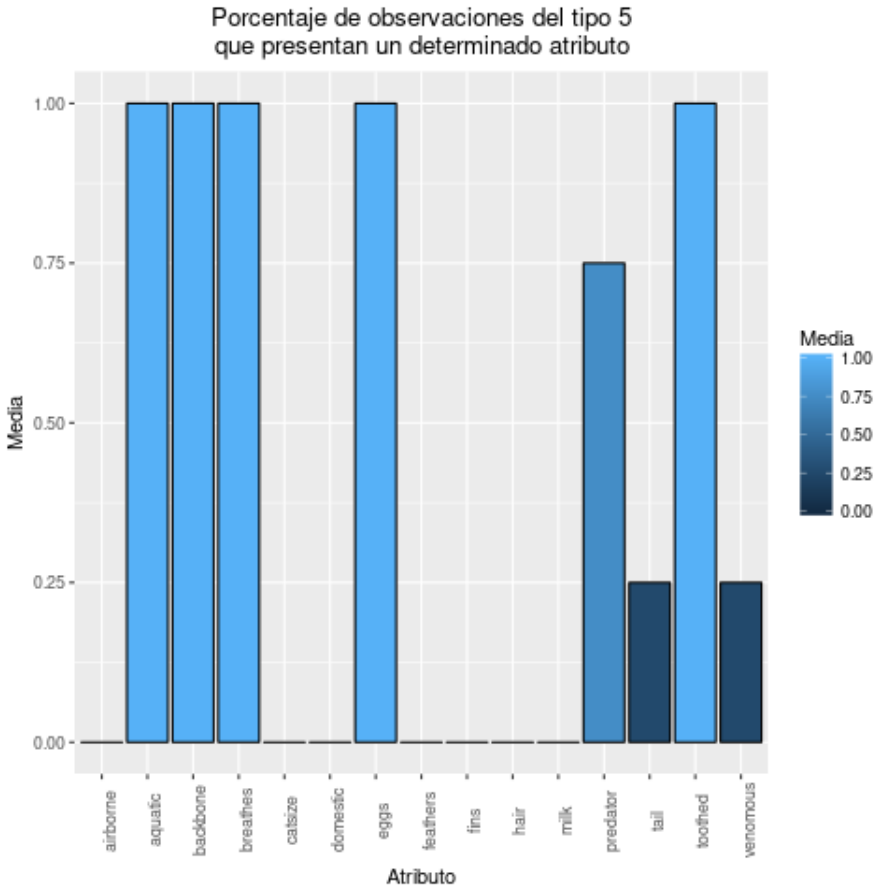

Figura 3.6: Porcentaje total de observaciones que presenta un atributo en Anfibios

De la figura 3.6, es posible desprender que en el caso de los Anfibios, se presentan 5 atributos que definen a los animales de este tipo. Dicho atributos corresponden a: son animales acuáticos, o al menos viven en alguna etapa de sus vidas en el agua (aquatic), poseen columna vertebral (backbone) y respiran aire (breathes), ponen huevos (eggs) y también poseen dentadura (toothed). Lo que los diferencia de la categoría de peces, es que en este caso los anfibios no presentan, en su totalidad ($25\,\%$), cola (tail). Además, en este caso, los anfibios poseen una característica que es nula en el caso de los peces: respirar aire (breathes).

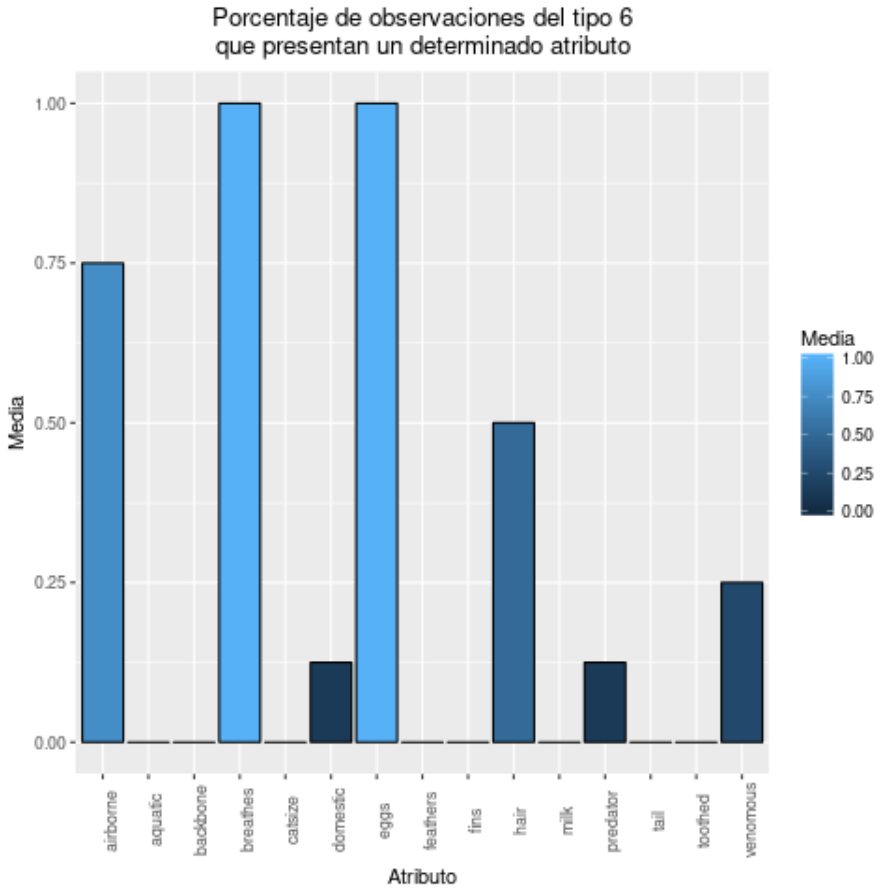

Figura 3.7: Porcentaje total de observaciones que presenta un atributo en Insectos

En el caso de la figura 3.7, resulta que en el caso de los Insectos, solo se tiene dos atributos que definen a esta clase de animales: Respiran aire (breathes) y ponen huevos (eggs). Sin embargo, aquí surge la pregunta de cómo se podrían diferenciar respecto a las aves, que también poseen estos atributos en un $100\,\%$. La respuesta a la interrogante radica principalmente en dos atributos: la presencia de pelo (hair), puesto que en lugar de tener pelo, las aves poseen plumas; y la presencia de veneno (venomous) en algunas especies de insectos, a diferencia de las aves, que ninguna resulta ser venenosa.

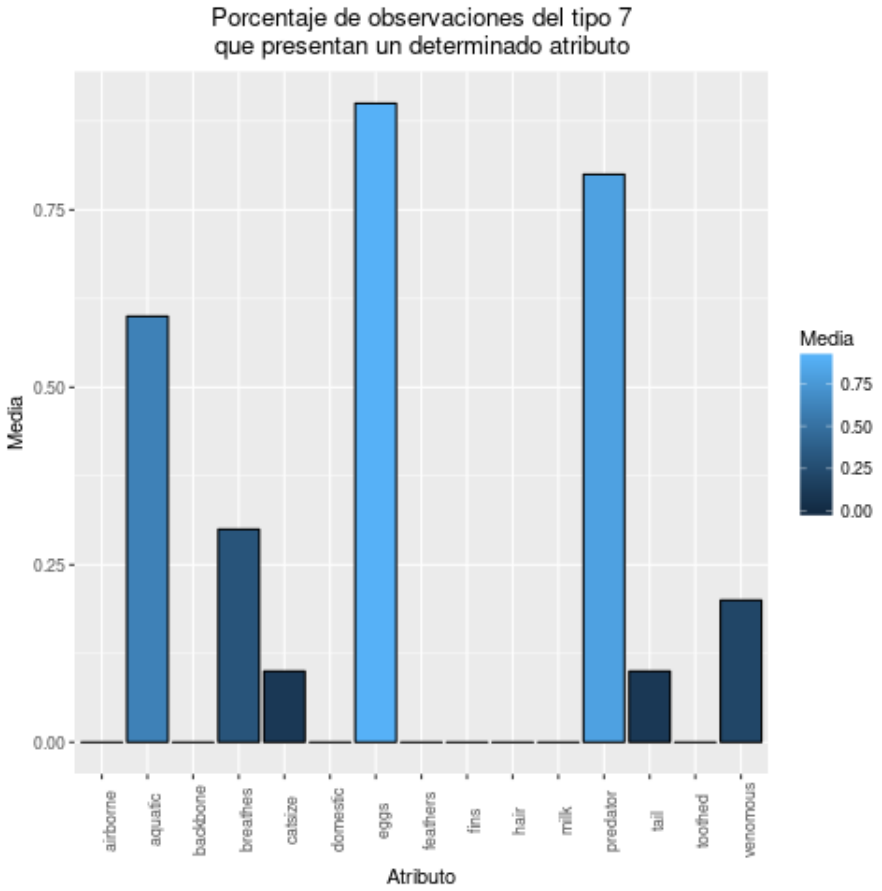

Figura 3.8: Porcentaje total de observaciones que presenta un atributo en animales clasificados como “Otros”

En este caso, no existen atributos que definan en un $100\,\%$ a los animales que están contenidos en esta clasificación, esto se debe principalmente a que en esta clase, existen animales de diferentes tipos, tal como se puede apreciar en la tabla 2.2. Por lo anterior, si se quisieran diferenciar a estos animales de cualquier otro de las clases analizadas anteriormente por sus atributos, quizás sea conveniente diferenciarlos entre sí por los atributos que no poseen, por ejemplo: que no han sido domesticados (domestic), que no poseen plumas (feathers), ni pelos (hair), ni aletas (fins), y también que no poseen columna vertebral (backbone), entre otros atributos que no posee esta clase y que las otras muy probablemente si posean, al menos uno de ellos.

	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	tail	domestic	catsize
hair	1	-0.43	-0.82	0.88	-0.2	-0.47	-0.15	0.49	0.19	0.44	-0.1	-0.28	0.05	0.21	0.46
feathers	-0.43	1	0.42	-0.41	0.66	-0.06	-0.1	-0.61	0.23	0.25	-0.15	-0.22	0.29	0.03	-0.14
eggs	-0.82	0.42	1	-0.94	0.38	0.38	0.01	-0.64	-0.34	-0.38	0.1	0.16	-0.22	-0.16	-0.51
milk	0.88	-0.41	-0.94	1	-0.37	-0.36	-0.03	0.63	0.38	0.42	-0.24	-0.16	0.21	0.16	0.57
airborne	-0.2	0.66	0.38	-0.37	1	-0.17	-0.3	-0.59	-0.1	0.29	0.01	-0.25	0.01	0.06	-0.35
aquatic	-0.47	-0.06	0.38	-0.36	-0.17	1	0.38	0.05	0.02	-0.64	0.09	0.6	-0.03	-0.22	-0.11
predator	-0.15	-0.1	0.01	-0.03	-0.3	0.38	1	0.13	0.05	-0.26	0.12	0.19	0.02	-0.31	0.14
toothed	0.49	-0.61	-0.64	0.63	-0.59	0.05	0.13	1	0.58	-0.07	-0.06	0.36	0.31	0.07	0.34
backbone	0.19	0.23	-0.34	0.38	-0.1	0.02	0.05	0.58	1	0.21	-0.25	0.21	0.73	0.1	0.36
breathes	0.44	0.25	-0.38	0.42	0.29	-0.64	-0.26	-0.07	0.21	1	-0.12	-0.62	0.09	0.12	0.2
venomous	-0.1	-0.15	0.1	-0.24	0.01	0.09	0.12	-0.06	-0.25	-0.12	1	-0.03	-0.16	0	-0.18
fins	-0.28	-0.22	0.16	-0.16	-0.25	0.6	0.19	0.36	0.21	-0.62	-0.03	1	0.2	-0.09	0.03
tail	0.05	0.29	-0.22	0.21	0.01	-0.03	0.02	0.31	0.73	0.09	-0.16	0.2	1	0.02	0.24
domestic	0.21	0.03	-0.16	0.16	0.06	-0.22	-0.31	0.07	0.1	0.12	0	-0.09	0.02	1	0.02
catsize	0.46	-0.14	-0.51	0.57	-0.35	-0.11	0.14	0.34	0.36	0.2	-0.18	0.03	0.24	0.02	1

Figura 3.9: Matriz de correlaciones

Finalmente, se presenta la matriz de correlaciones, en donde se indica que tan fuertemente relacionada están todas y cada una de las variables o atributos entre sí. Las correlaciones se realizan en base al *Coeficiente Phi*, una medida similar al *Coeficiente de correlación de Pearson*; en efecto, la estimación del *Coeficiente de correlación de Pearson* para dos variables binarias, da como resultado el *Coeficiente Phi*. Dicha relación toma valores entre -1 y 1. Si el valor es igual a 1, entonces existe una relación directa entre ambas variables (si aumenta una, la otra lo hará en igual proporción) y en caso de que el valor sea igual a -1 existe una relación indirecta (si disminuye una, la otra aumenta en la misma proporción). Por otro lado, si la relación toma el valor de 0, no existe una relación entre ambas variables.

Por ejemplo, en el caso de los atributos *milk* y *eggs*, el valor de la correlación entre estos es muy cercano a -1 ($-0{,}94$), lo que básicamente indica (al tratarse de atributos booleanos) que si se tiene un animal que pone huevos, existen muchas probabilidades de que ese animal no de leche a sus crías, lo cual sigue toda lógica.

Por otra parte, se tiene una relación igual a 0 entre los atributos *venomous* y *domestic*, lo que indica una relación no determinista entre un atributo y otro. Básicamente, no existe una relación clara o lógica entre estas variables.

Ahora bien, si se observa la relación entre los atributos *hair* y *milk* su valor es muy cercano a 1 ($0{,}88$). De lo anterior, se puede inferir que existe una gran posibilidad de que si en un animal hay presencia de pelo, el mismo animal sea capaz de dar leche a sus crías.

CAPÍTULO 4. CONCLUSIONES

A lo largo de la experiencia, se ha podido corroborar que tanto los atributos como los valores que estos toman en cada una de las instancias de la base de datos tienen sentido respecto a lo que se podría esperar de una base de datos verídica, ya que esta no presenta errores ni datos erróneos. De esta forma, se comprueba lo expuesto en el capítulo introductorio, en donde se indica que la base de datos es ficticia y creada por el investigador para los objetivos allí planteados.

Respecto a los objetivos planteados para la realización de esta experiencia, es posible afirmar que estos han sido completados de manera satisfactoria, ya que se logró entender el propósito de cada una de las variables de la base de datos, y gracias a los análisis realizados su funcionamiento para cada caso analizado.

Finalmente, para un futuro no muy lejano, se espera contar con las herramientas tanto técnicas como teóricas necesarias para relacionar la base de datos en sí con el proceso de agrupamiento de datos y los procesos que vengan por delante en las siguientes experiencias.

CAPÍTULO 5. BIBLIOGRAFÍA

[1] R. Forsyth, *Beagle User Notes*, 2016. dirección: https://www.richardsandesforsyth.net/beagling.html.