# UNIVERSIDAD DE SANTIAGO DE CHILE

**FACULTAD DE FACULTAD**

**Departamento de Departamento**



## Laboratorio N°2


**Nicolás Alarcón L. - Héctor Pérez M. - Pedro Silva A.**


Profesor Cátedra: Max Chacón

Profesor Laboratorio: Felipe Bello

Santiago – Chile

2020

# TABLA DE CONTENIDO

# ÍNDICE DE TABLAS

# ÍNDICE DE ILUSTRACIONES

# CAPÍTULO 1. INTRODUCCIÓN

## 1.1 ANTECEDENTES Y MOTIVACIÓN

La base de datos de hongos "*Mushrooms*", obtenida del repositorio de bases de datos de la *University of California, Irvine*, corresponde a muestras que fueron donadas por el Dr. Jeff Schlimmer, y fueron recopiladas por *The Audubon Society Field Guide to North American Mushrooms* (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf. Esta base de datos ha sido ampliamente estudiada, con mucha información disponible en la web y también muy utilizada con fines académicos, ya que constituye un muy buen ejemplo del poder de los algoritmos de clasificación aplicado a casos reales, como lo es en este caso, predecir si un hongo es venenoso o comestible en función de sus atributos físicos.

En la instancia anterior de este trabajo se realizó un análisis exploratorio a la base de datos a través de histogramas y análisis de correlación. Esta experiencia realizada fue fundamental para comprender el comportamiento de los algoritmos de agrupamiento utilizados.

## 1.2 DESCRIPCIÓN DEL PROBLEMA

El *dataset* Mushrooms tiene un total de 8124 registros de hongos y 23 atributos físicos como tamaño, color y textura del hongo, la densidad de individuos de la misma especie, presencia de magulladuras, entre otros. El problema es encontrar la forma de diferenciar aquellos hongos que son comestibles y aquellos que no lo son.

## 1.3 SOLUCIÓN PROPUESTA

Considerando la primera instancia de este trabajo donde se realizó un análisis exploratorio, se propone realizar un análisis a través de un algoritmo de agrupamiento de datos, específicamente el método *k-means*. Para esto es necesario pre-procesar la base de datos, para que cumpla con el formato de entrada para el algoritmo. Además se propone utilizar métodos que permitan encontrar los parámetros óptimos de agrupamiento, para obtener mejores resultados y análisis.

## 1.4 OBJETIVOS Y ALCANCE DEL PROYECTO

### 1.4.1 Objetivo general

Extraer conocimiento del problema utilizando algoritmo de agrupamiento como *K-Means* y *PAM*, justificando los parámetros utilizados para realizar un análisis de los resultados y comparar con literatura encontrada.

### 1.4.2 Objetivos específicos

1. Realizar preprocesamiento de la base de datos para realizar agrupamiento utilizando algoritmo *K-Means* y *PAM*.
2. Definir parámetros de funcionamiento del algoritmo a través de un método a elección.
3. Agrupar y realizar el análisis de centroides e identificar características de cada clase.
4. Comprar resultados con literatura existente
5. Analizar los resultados obtenidos por los algoritmos utilizados.
6. Concluir respecto a los resultados obtenidos.

### 1.4.3 Alcances

Se considera como principal alcance, el hecho de que los autores de este documento no posen grandes conocimientos en micología, aun así existe bastante información del manejo de esta base de datos, ya que es ampliamente utilizada en la docencia de ciencia de datos. Para suplir esta debilidad, se ha incorporado la ayuda de personas con conocimiento sobre el reino funji.

## 1.5 METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

### 1.5.1 Metodología

En primera instancia se debe realizar un preprocesamiento a la base de datos para que pueda ser ingresada como input al algoritmo de agrupamiento. El *dataset* solo presenta variables categóricas, pero éstas deben ser ingresadas como variables numéricas, lo que requiere procesamiento de los datos.

Luego de que los datos se encuentren correctamente procesados, es necesario obtener el número óptimo de *clusters*, para esto se utilizarán los métodos de *Silhouette*.

A partir de la cantidad de *clusters* óptimos del punto anterior, se realiza el agrupamiento utilizando el algoritmo *K-Means* y *PAM* para ambos números de *clusters* para su posterior análisis de centroide y caracterización de cada una de las clases.

Finalmente, los resultados serán comparados con literatura existente relacionada a este *dataset*.

### 1.5.2 Herramientas de desarrollo

La base de datos será estudiada y analizada con el software de código abierto Rstudio (R), programa especializado en el análisis estadísticos de bases de datos y ampliamente utilizado en el mundo por su facilidad para el tratado de información.

# CAPÍTULO 2. MARCO TEÓRICO

## 2.1 CLUSTERING

El análisis de *clustering* se basa en encontrar elementos de similares características en un grupo de datos o diferencias entre los grupos de datos. Es un proceso ampliamente utilizado en minería de datos y *machine learning*, con aplicaciones en diversas áreas como ecología, economía, ciencias de la tierra, inteligencia de mercados, bioinformática, y muchos más. Entre los algoritmos más utilizados se encuentra *K-Means*, *Mean-Shift* y *Self-Organizing Maps* (SOM) (Rousseeuw, 1990).

## 2.2 EVALUACIÓN DE NÚMERO DE *CLUSTERS*

### 2.2.1 Metodo *Silhouette*

*Silhoutte* presenta una medida representable gráficamente la similaridad de un registro con respecto a su grupo y ayuda a interpretar una calidad relativa de los *clusters* y la configuración obtenida. Este método sirve para evaluar la validez de un grupo y permite seleccionar un número apropiado de *clusters* (Rousseuw, 1986).

## 2.3 MÉTODOS DE *CLUSTERING*

### 2.3.1 *K-Means*

*K-Means* es probablemente uno de los algoritmos de agrupamiento más estudiados, por su fácil entendimiento y aplicación. Principalmente, separa un conjunto de datos en k grupos o clases, para esto, define centroides de grupos y va iterando esta definición de centroides hasta que forme grupos con distancias similares desde el centro del grupo hacia el límite (Seif, 2018). Para esto, el algoritmo sigue los siguientes pasos:

1. Se selecciona el número de grupos o clases*K-means*.

2. Se define un centroide para cada uno de los grupos.
3. Se calcula la distancia desde cada registro hacia todos los centroides y se asigna cada registro al centroide más cercano.
4. Se vuelve a calcular el centroide de cada grupo nuevo.
5. Iterar paso 3 y 4 hasta la convergencia del algoritmo, es decir, los centroides se mantienen sin modificaciones y los datos no cambian de grupo.

Finalmente los grupos deben ser analizados, para encontrar patrones y caracterizar cada clase.

### 2.3.2 *Partition Around Medoids*

Los algoritmos de *clustering K-medoids*, parecido a *K-means*, y prueba varios métodos para seleccionar el *medoids* inicial. Este algoritmo calcula una matriz de distancia y luego la utiliza para encontrar nuevos medoids para cada iteración. ”Entre los algoritmos de agrupamiento para K-medoids, el algoritmo de partición alrededor de medoids (PAM), es conocido por ser uno de los más poderosos” (Park, 2009).

## 2.4 DISTANCIAS

En orden de medir la similitud o regularidad acerca de registros de un *dataset*, la distancia métrica juega un rol muy importante. Es necesaria para identificar en que modo los datos están interrelacionados, las similitudes o diferencias entre cada uno y que medidas son consideradas para sus comparaciones (Singh, 2013). Al momento de realizar la elección del tipo de distancia a utilizar debemos tener en cuenta los tipos de variables que contiene nuestro *dataset*, por ejemplo, para un *dataset* que tiene únicamente variables cuantitativas es recomendable utilizar la distancia de Mahalanobis debido a que toma en cuenta las correlaciones las variables, cuando el dataset tiene únicamente variables binarias es recomendable utilizar la distancia de Jaccard y al tener variables mixtas una opción es la distancia Gower.

### 2.4.1 Distancia Gower

Se define la distancia de Gower como $d_{ij}^2 = 1 - s_{ij}$ (Grané, .) donde,

$$s_{ij} = \frac{\sum_{i=1}^{p_1}(1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{2.1}$$

Además: $p_1$ es el número de variables cuantitativas continuas,

$p_2$ es el número de variables binarias,

$p_3$ es el número de variables cualitativas(no binarias)

$a$ es el número de coincidencias (1,1) en las variables binarias,

$d$ es el número de coincidencias (0,0) en las variables binarias,

$\alpha$ es el número de coincidencias en las variables cualitativas (no binarias),

$G_h$ es el número de similaridad de Gower

# CAPÍTULO 3. DESARROLLO

## 3.1 PREPROCESAMIENTO

- En primer lugar fueron modificadas el valor de las 18 variables por otros valores con nombre mas representativos.
- La variable ”veil_type” solo posee observaciones de una categoría, es decir, esta no nos aporta nada relevante por lo tanto se decidió eliminar esa columna.
- La variable ”stalk_root” contenía valores perdidos dentro de su categorías, con el fin de no sesgar se decidió eliminar toda las muestras con valor ”*missing*” quedando el dataset de 8124 observaciones en 5644 observaciones.
- Por último, se utilizó el método de ”*One Hot Encoding*” el cual permite transformar las variables categóricas en variables binarias, esto debido a que muchos algoritmos no pueden funcionar directamente con los datos categóricos.

## 3.2 *CLUSTER*

### 3.2.1 Criterio de proximidad

Para el cálculo de todas las diferencias (distancias) de pares entre observaciones en el conjunto de datos, se utilizó la métrica *Gower* descrita anteriormente. Esta métrica permite que las variables originales puedan ser de tipos mixtos,es decir, se pueden utilizar diferentes tipos en el mismo conjunto de datos, tales como binarias, categóricas, numéricas, entre otras.

### 3.2.2 Numero de *Clusters*

Para encontrar el número óptimo de *clusters*, se utilizó el método *Silhouette* o Silueta para la función *K-Means* y para la función *PAM*. No se utilizó el método Gap debido al costo computacional asociado, ya que primero se crea una matriz de similitud, expandiendo a su forma binaria cada categoría de cada atributo, lo cual hace que las iteraciones realizadas por *Gap* sea

demasiada carga para el nivel de recursos con los que se cuenta. Al aplicar Silueta se obtuvo como número óptimo 2 de grupos para ambos tipos de *clusters*. Los resultados se muestran gráficamente en las siguientes Figuras:

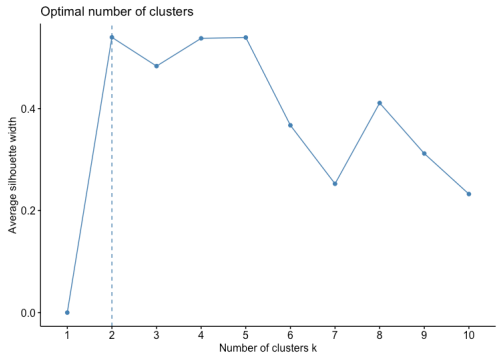


Figura 3.1: Gráfico de resultado de *clustering* obtenido de *Silhouette* para la función *K-Means*.

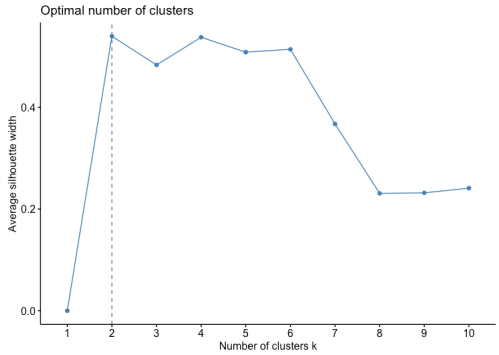


Figura 3.2: Gráfico de resultado de *clustering* obtenido de *Silhouette* para la función *PAM*.

### 3.2.3 *Clusters* generados

Al utilizar el método *Silhouette* este nos indicó que para *K-Means* y *PAM* el número de grupos óptimos para agrupar era de 2, sin embargo, de manera exploratoria y con el fin de ver cómo se comportan ambos algoritmos en torno a diferentes grupos, se agruparon de 2, 3, 4 y 5 grupos utilizando ambos métodos.

#### *3.2.3.1 Clusters generados con 2 grupos*

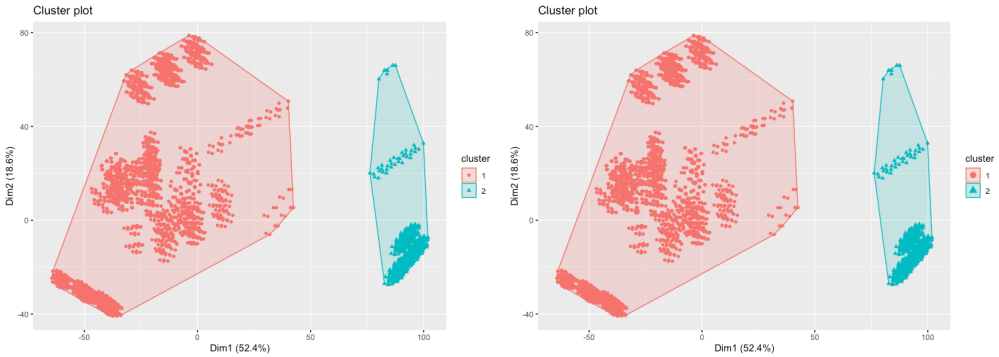


Figura 3.3: *PAM* y *K-Means* respectivamente con 2 grupos.

| **Clase\Grupo** | 1 | 2 |
|---|---|---|
| poisonous | 816 | 1340 |
| edible | 3488 | 0 |

Tabla 3.1: Tabla clase Type vs 2 grupos, utilizando el método de *clusterinig PAM*.

| **Clase\Grupo** | 1 | 2 |
|---|---|---|
| poisonous | 816 | 1340 |
| edible | 3488 | 0 |

Tabla 3.2: Tabla clase Type vs 2 grupos, utilizando el método de *clustering K-Means*.


Figura 3.4: *PAM* y *K-Means* respectivamente con 3 grupos.

| **Clase\Grupo** | 1 | 2 | 3 |
|---|---|---|---|
| poisonous | 820 | 0 | 1336 |
| edible | 1758 | 1730 | 0 |

Tabla 3.3: Tabla clase Type vs 3 grupos, utilizando el método de *clustering PAM*.

| **Clase\Grupo** | 1 | 2 | 3 |
|---|---|---|---|
| poisonous | 0 | 820 | 1336 |
| edible | 1728 | 1760 | 0 |

Tabla 3.4: Tabla clase Type vs 3 grupos, utilizando el método de *clustering K-Means*.

### *3.2.3.2 Clusters generados con 3 grupos*

### *3.2.3.3 Clusters generados con 4 grupos*

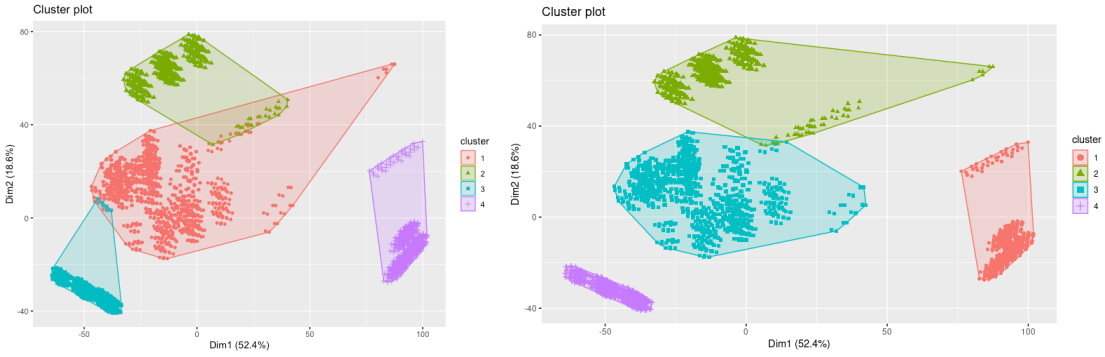


Figura 3.5: *PAM* y *K-Means* respectivamente con 4 grupos.

| Clase\Grupo | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| poisonous | 824 | 0 | 0 | 1332 |
| edible | 950 | 794 | 1744 | 0 |

Tabla 3.5: Tabla clase Type vs 4 grupos, utilizando el método de *clustering PAM*.

| Clase\Grupo | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| poisonous | 1332 | 8 | 816 | 0 |
| edible | 0 | 810 | 950 | 1728 |

Tabla 3.6: Tabla clase Type vs 4 grupos, utilizando el método de *clustering K-Means*.

#### *3.2.3.4 Clusters generados con 5 grupos*

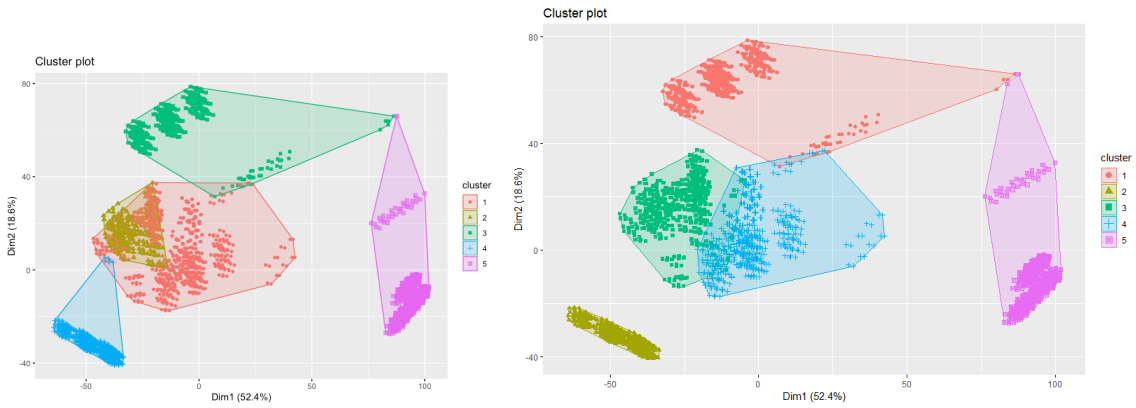


Figura 3.6: PAM y Kmeans respectivamente con 5 grupos

| Clase\Grupo | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| poisonous | 786 | 30 | 6 | 0 | 1334 |
| edible | 241 | 707 | 802 | 1738 | 0 |

Tabla 3.7: Tabla clase Type vs 5 grupos, utilizando el método de *clustering PAM*.

| Clase\Grupo | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| poisonous | 5 | 0 | 268 | 548 | 1335 |
| edible | 803 | 1728 | 928 | 29 | 0 |

Tabla 3.8: Tabla clase Type vs 5 grupos, utilizando el método de *clustering K-Means*.

# CAPÍTULO 4. ANÁLISIS DE LOS RESULTADOS

Considerando que la métrica *gower* nos entregó como resultado que 2 era el número de *clusters* óptimo y que para ambas funciones (*K-means* y *PAM*) nos entregaban los mismos resultados, se decide realizar en análisis de los resultados con aquellos obtenidos de la función *PAM* para 2 grupos.

Como primer paso para analizar las 21 variables con las que contamos, se decidió como grupo explorar a simple vista cuales de ellas presentaban una muestra clara de agrupamiento. Apoyados por la intuición que nos entregó el test de Chi-Cuadrado realizado en la experiencia anterior, confirma que variables como ”odor”, ”spore_print_color” o ”ring_type” son variables que presentan una distribución relevantes en el agrupamiento realizado.

También nos apoyamos en en la publicado por Huan Liu (2001), quien indica en orden mayor a menor importancia los atributos de la base de datos. Un resumen de las primeras cinco variables son: ”odor”, ”spore_print_color”, ”gill_size”, ”ring_type” y ”bruises”. Esto coincide con los resultados analizados visualmente por el equipo, lo cual nos da la confianza de seguir analizando estas variables y algunas más que consideramos relevantes mencionar.

Es importante destacar que el resto de gráficos para las variables restantes se encuentran en el Anexo.

## 4.1 VARIABLES

1. **Type**: Observando el gráfico de barra descrito en la Figura 4.1, el 81% del grupo 1 es comestible y el 19% es venenoso, en cambio la totalidad del grupo 2 es venenoso. Eventualmente si un nuevo hongo es agregado al grupo 1, este no nos permitiría saber con certeza si este hongo es comestible o no, sin embargo si este entra en el grupo 2 tiene una gran probabilidad según este método de agrupación que este hongo no sea comestible.
2. **Bruises**: Observando el gráfico de barra descrito en la Figura 4.2 la totalidad del grupo 2 presentaba moretones, en cambio solo el 26% del grupo 1 contenía moretones. Se podría inferir una relación entre la presencia de moretones con los hongo venenosos, esta relación es justificada al observar que la totalidad del grupo 2, el cual representa el grupo de los hongos venenosos presentaba moretones. Por otro lado el 20% del grupo 1 es venoso y este podría relacionarse con el 26% del grupo 1 que contenía moretones. Sin embargo, estas relaciones deben estudiarse con técnicas de inferencia para determinar si están correlacionadas estas variables o no.

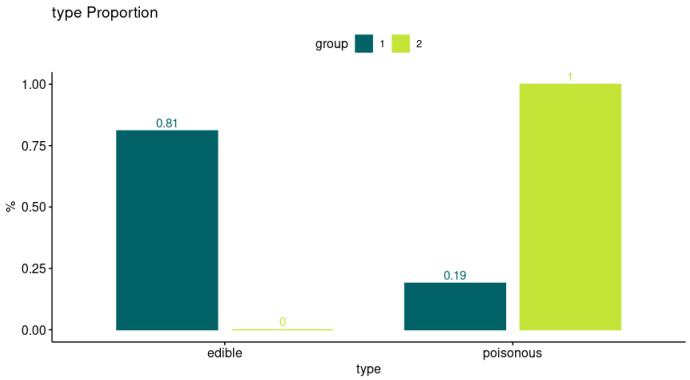


Figura 4.1: Gráfico de barras de porcentaje de grupos para la variable type.

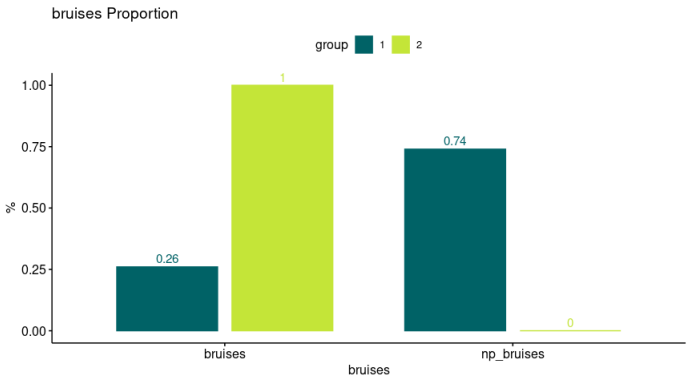


Figura 4.2: Gráfico de barras de porcentaje de grupos para la variable bruises.

3. **Cap_color**: Observando el gráfico de barra descrito en la Figura 4.3, el grupo 2 está compuesto casi en su totalidad por dos colores, gris y amarillo con un 48% y 49% respectivamente del total, por otro lado el grupo 1 se distribuye en 7 colores. A simple vista se podría inferir una relación entre el color gris y amarillo con los hongos venosos, además al relacionar estos dos colores con el grupo 1 se puede observar que el 24% del grupo 1 tiene color gris y menos del 1% tiene color amarillo, este porcentaje podría estar relacionado con las muestras venenosas del grupo 1 con el color amarillo. Sin embargo, estas relaciones deben estudiarse con técnicas de inferencia para determina si están correlacionadas estas variables o no.

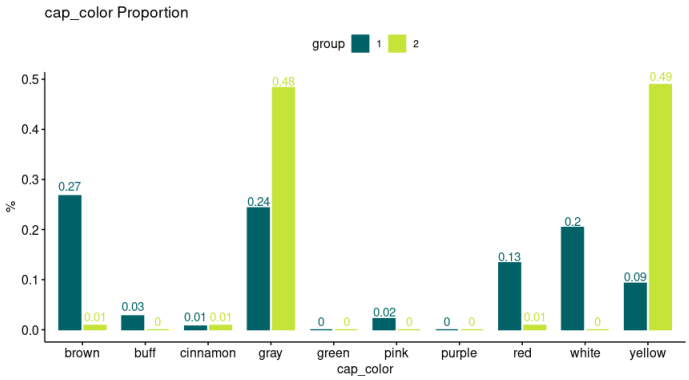


Figura 4.3: Gráfico de barras de porcentaje de grupos para la variable cap_color.

4. **Odor**: Observando el gráfico de barra descrito en la Figura 4.4, el 93% del grupo 2 presentaba mal olor, en cambio el grupo 1 en su mayoría no presentaba olor. A simple vista se podría inferir una correlación entre el mal olor y hongos venenosos. Sin embargo para confirmar lo anterior, estas variables se deben someter a mas técnicas de correlación.

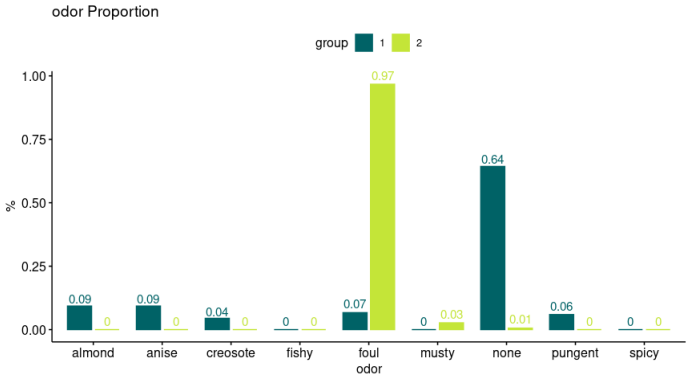


Figura 4.4: Gráfico de barras de porcentaje de grupos para la variable odor.

5. **Ring_type**: Observando el gráfico de barra descrito en la Figura 4.5 el 97% de los registro del grupo 2 corresponden a los anillos del tipo largo, en cambio el grupo 1 en su mayoría correspondían a los anillos del tipo colgante. A simple vista se podría inferir una correlación entre el anillo tipo largo y los hongos venenosos. Sin embargo para confirmar lo anterior, estas variables se deben someter a más técnicas de correlación.

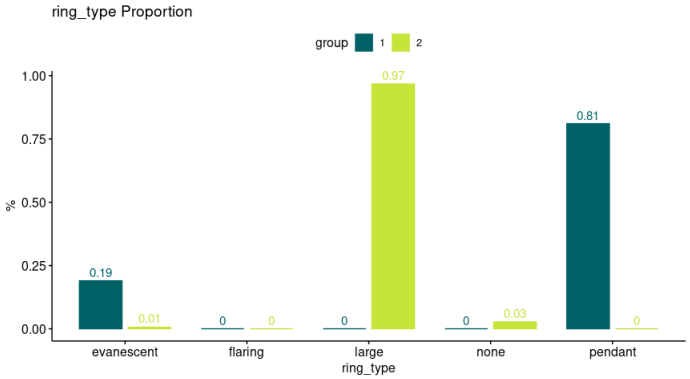


Figura 4.5: Gráfico de barras de porcentaje de grupos para la variable ring_type.

6. **Spore_print_color**: Observando el gráfico de barra descrito en la Figura 4.6 El 97% del grupo 2 tiene esporas de color chocolate y un 3% de color blanco. El 88% del grupo 1 tiene esporas de color negro o café con un 43% y 45% respectivamente. A simple vista se podría inferir una correlación entre el color chocolate y los hongos venenosos. Sin embargo para confirmar lo anterior, estas variables se deben someter a más técnicas de correlación.

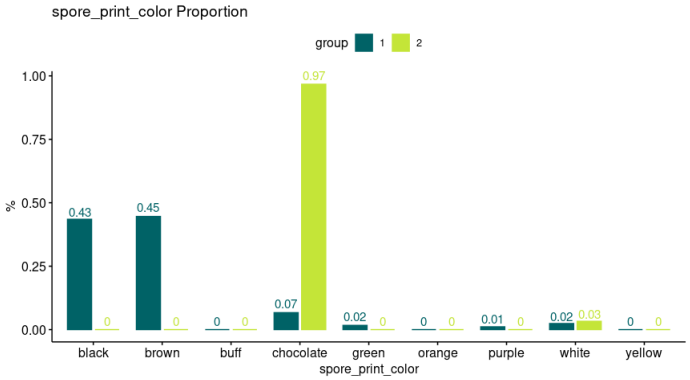


Figura 4.6: Gráfico de barras de porcentaje de grupos para la variable spore_print_color.

7. **Stalk_shape**: Observando el gráfico de barra descrito en la Figura 4.7 la totalidad del grupo 2 presentaba tallos largos, en cambio solo el 33% del grupo 1 presentaba un tallos largos, siendo el resto tallos afilados. Se podría inferir una relación entre la presencia de tallo largo con los hongo venenosos, esta relación es justificada al observar que la totalidad del grupo 2

el cual representa el grupo de los hongos venenosos presentaba moretones, por otro lado el 20% del grupo 1 es venoso y este podría relacionarse con el 33% del grupo 1 que contenía moretones. Sin embargo, estas relaciones deben estudiarse con técnicas de inferencia para determinar si están correlacionadas estas variables o no.

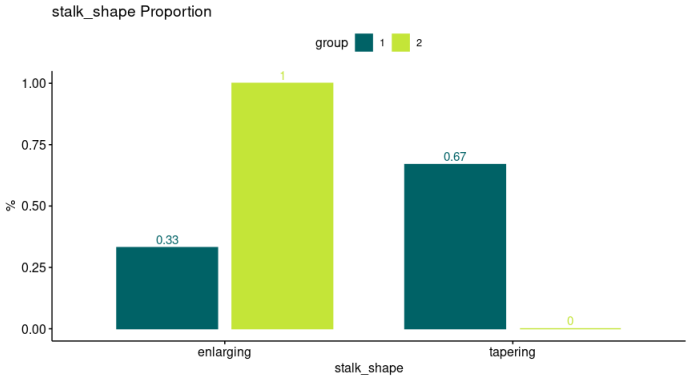


Figura 4.7: Gráfico de barras de porcentaje de grupos para la variable stalk_shape.

8. **Stalk_surface_above_ring**: Observando el gráfico de barra descrito en la Figura 4.8 El 99% del grupo 2 presentaba una superficie sedosa, por otro lado el 87% del grupo 1 presentaba una superficie suave. A simple vista se podría inferir una correlación entre la superficie sedosa y los hongos venenosos. Sin embargo, para confirmar lo anterior, estas variables se deben someter a más técnicas de correlación.

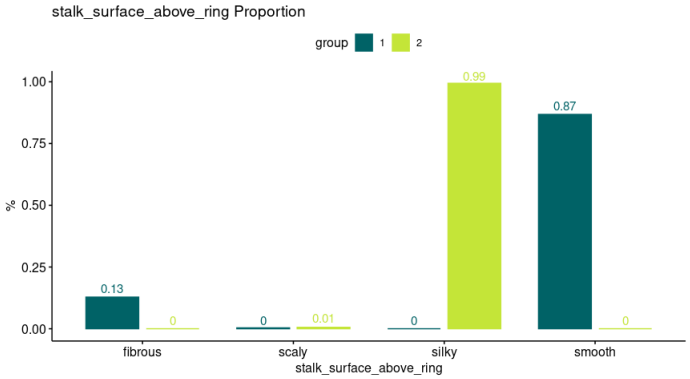


Figura 4.8: Gráfico de barras de porcentaje de grupos para la variable stalk_surface_above_ring.

9. **Stalk_surface_below_ring**: Observando el gráfico de barra descrito en la Figura 4.9 el 97% del grupo 2 presentaba una superficie de tallo bajo el anillo sedosa, por otro lado el 82% del grupo 1 presentaba una superficie bajo el anillo suave. A simple vista se podría inferir una correlación entre la superficie bajo el anillo sedosa y los hongos venenosos. Sin embargo, para confirmar lo anterior, estas variables se deben someter a más técnicas de correlación.

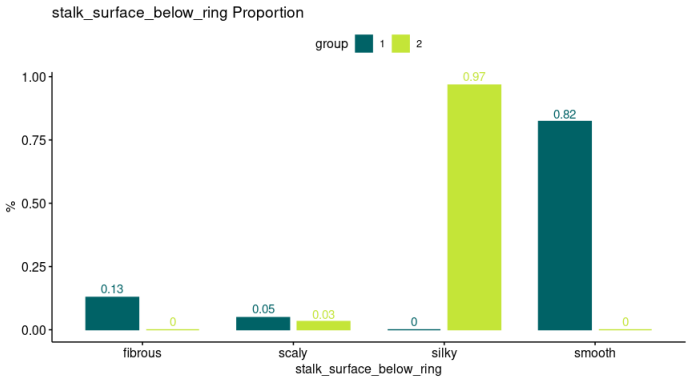


Figura 4.9: Gráfico de barras de porcentaje de grupos para la variable stalk_surface_below_ring.

# CAPÍTULO 5. CONCLUSIONES

Pese a que el *dataset* utilizado puede considerarse pequeño, aún así los métodos utilizados como el método de las siluetas presentaba una alta carga en recursos computacionales para su cálculo, llegando a demorar alrededor de 2 horas aproximadamente. Por otro lado, el método *PAM* y *K-Means* también destacaron por su alta demora en entregar los resultados. Esto habla especialmente de la complejidad algorítmica de los métodos, ya que si utilizáramos un *dataset* con más datos y más categorías el tiempo de espera de los resultados crecería considerablemente.

Con respecto a los resultados, por un lado es bastante lógico pensar que si bien se posee dos clases sobre las cuales decidir, no significa que debemos ajustar los grupos escogidos a este número de clases, es por ello que se utiliza un método de siluetas para escoger el número óptimo de grupos a realizar, apoyándonos así en algo concreto y no solo en nuestra intuición. Por otro lado, mientras se realizaba la experiencia, nos encontramos con varios artículos o guías en *Internet* que describían un proceso de agrupamiento llegando a utilizar alrededor 24 grupos, aunque estos artículos nos basaban su decisión en un modelo matemático sino que al ser un *dataset* compuesto por alrededor de 23 especies de hongos diferentes dentro de la misma familia, como lo describen los autores del *dataset*.

Las variables que fueron escogidas para su análisis demuestran una relevancia en los grupos formados. El más importante de todos y sobre el cual se basa el análisis del resto es en la variable de la clase ”type”, quien nos indica que en el grupo dos se obtuvo un grupo considerablemente marcado por pertenecer a la categoría de los venenosos, lo cual es un buen indicio de que el método utilizado está agrupando correctamente. Aunque el otro grupo tuvo un distribución considerable en ambas clases, lo cual no lo hace tan confiable como el otro grupo.

Según el análisis realizado a los gráficos de barra las variables que podrían tener una correlación son ”odor”, ”ring_type”, ”spore_print_color”, ”stalk_surface_above_ring” y ”stalk_surface_below_ring”. Estas variables coinciden con el análisis de Chi-Cuadrado realizado en el laboratorio anterior.

Durante la investigación se estudiaron otras herramientas de visualización de grupos como ”t-SNE” sin embargo por motivos de tiempo se decidió profundizar en método conocimos como lo son *K-Means* y *PAM*. Sin embargo, es posible considerarlas para experiencias posteriores.

Finalmente es importante recordar que la intoxicación por hongos puede tener diversas consecuencias que pueden afectar gravemente la salud e incluso provocar la muerte. No existe una correlación unitaria entre los atributos físicos de los hongos en general y sus niveles de toxinas, por lo que no es recomendable ingerirlos sin estar 100% seguro de que son consumibles (Konno, 2009).

# REFERENCIAS BIBLIOGRÁFICAS


Grané, A. (.). Distancias estadísticas y escalado multidimensional (análisis de coordenadas principales). Recuperado el Jueves 04 de Junio de 2020, de http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf.

Huan Liu, J. Y., Hongjun Lu (2001). Toward multidatabase mining: identifying relevant databases. urlhttps://ieeexplore.ieee.org/document/940731.

Konno, K. (2009). Xvii. poisonous mushrooms. Food Reviews International, 13:3, 471-487, DOI: 10.1080/87559129709541134.

Park, H.-S. (2009). A simple and fast algorithm for k-medoids clustering. Expert Systems with Applications 36 (2009) 3336–3341.

Rousseeuw, L. K. . P. J. (1990). Finding groups in data. Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). Finding Groups in Data. Wiley Series in Probability and Statistics. doi:10.1002/9780470316801.

Rousseuw, P. J. (1986). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1987) 53-65.

Seif, G. (2018). The 5 clustering algorithms data scientists need to know. Recuperado el Jueves 04 de Junio de 2020, de https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

Singh, A. (2013). K-means with three different distance metrics. International Journal of Computer Applications (0975 – 8887).

# ANEXO A. GRÁFICO DE BARRAS DE VARIABLES

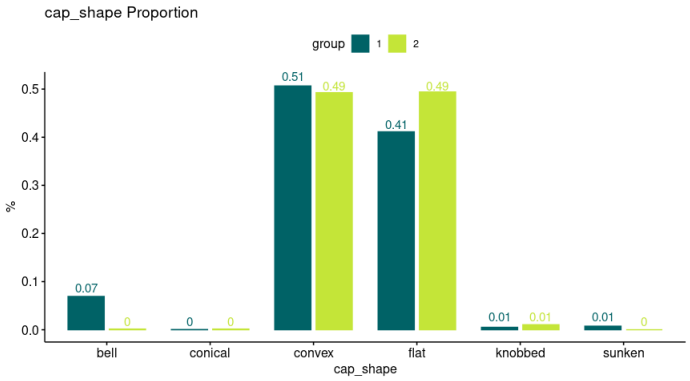


Figura A.1: Gráfico de barras de porcentaje de grupos para la variable cap_shape.

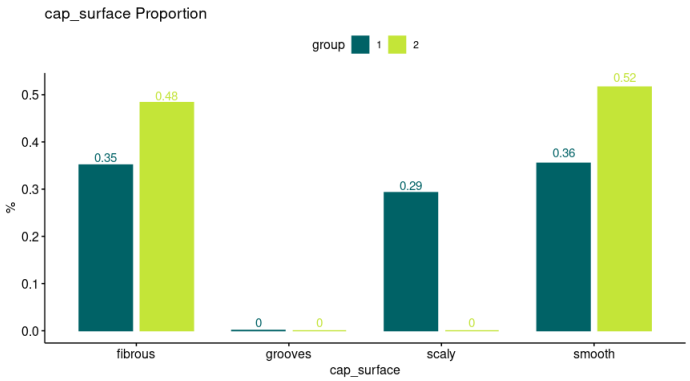


Figura A.2: Gráfico de barras de porcentaje de grupos para la variable cap_surface.

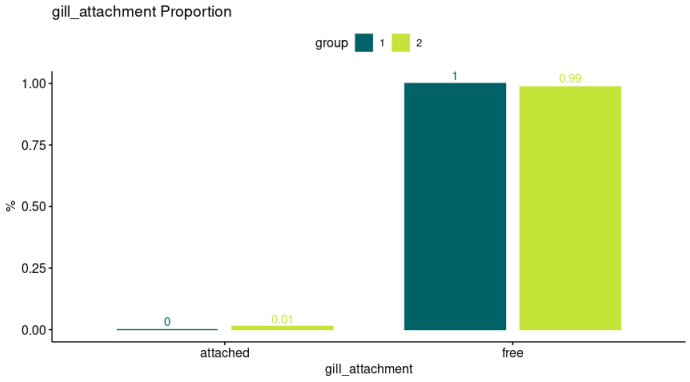


Figura A.3: Gráfico de barras de porcentaje de grupos para la variable gill_attachment.

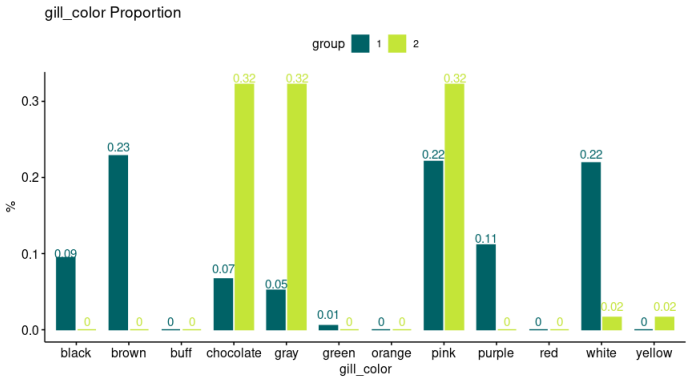


Figura A.4: Gráfico de barras de porcentaje de grupos para la variable gill_color.

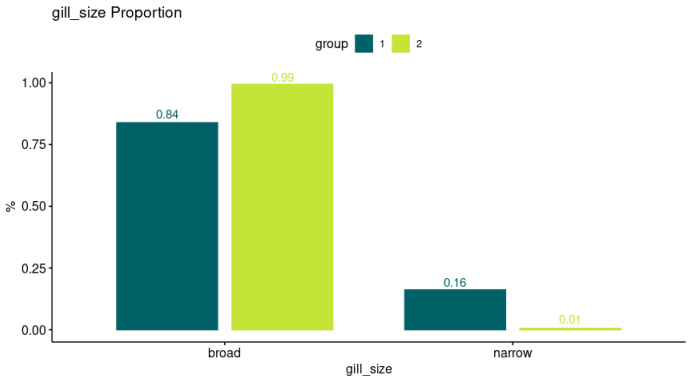


Figura A.5: Gráfico de barras de porcentaje de grupos para la variable gill_size.

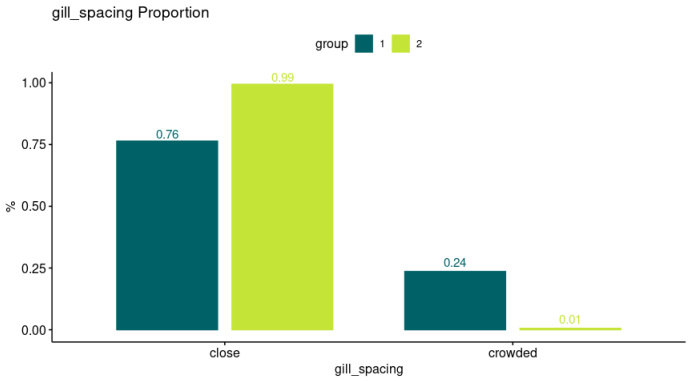


Figura A.6: Gráfico de barras de porcentaje de grupos para la variable gill_spacing.

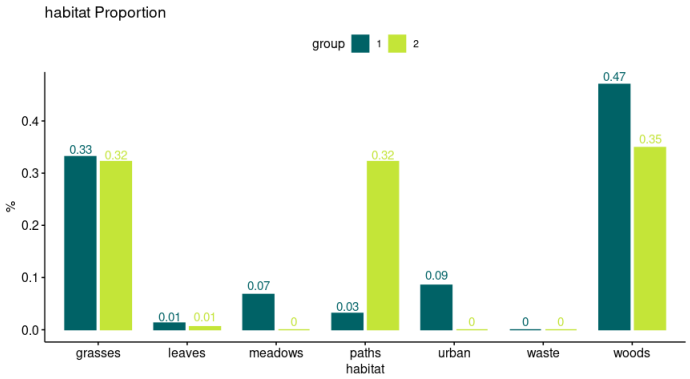


Figura A.7: Gráfico de barras de porcentaje de grupos para la variable habitat.

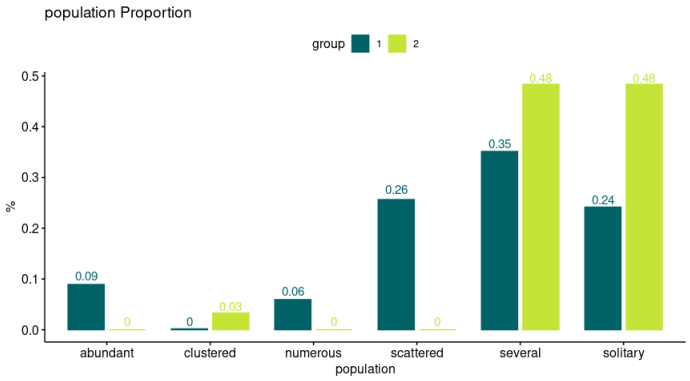


Figura A.8: Gráfico de barras de porcentaje de grupos para la variable population.

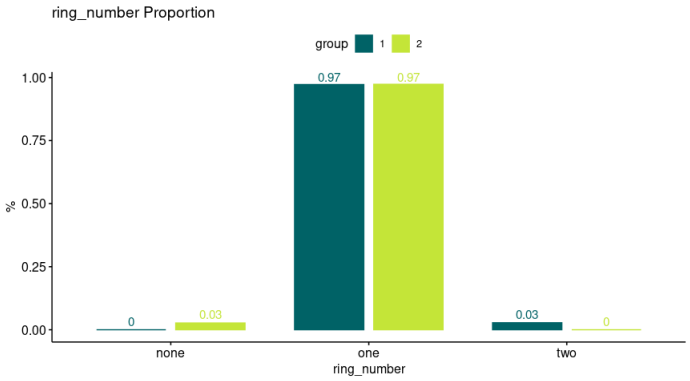


Figura A.9: Gráfico de barras de porcentaje de grupos para la variable ring_number.

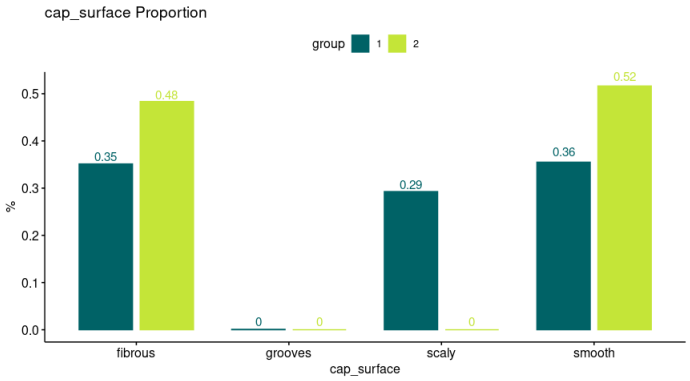


Figura A.10: Gráfico de barras de porcentaje de grupos para la variable cap_surface.

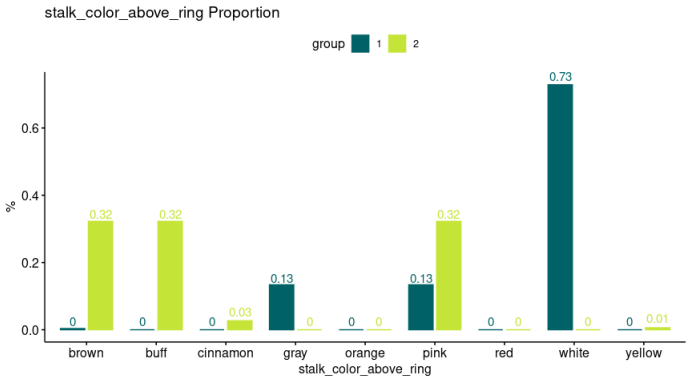


Figura A.11: Gráfico de barras de porcentaje de grupos para la variable stalk_color_above_ring.

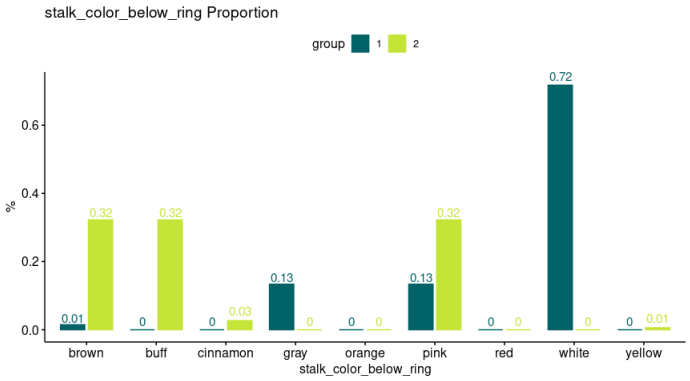


Figura A.12: Gráfico de barras de porcentaje de grupos para la variable stalk_color_below_ring.

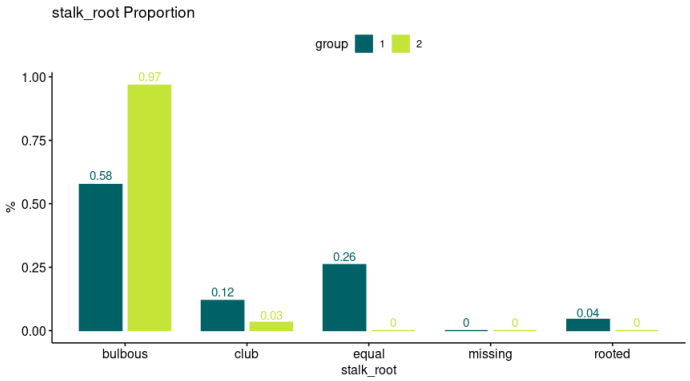


Figura A.13: Gráfico de barras de porcentaje de grupos para la variable stalk_root.

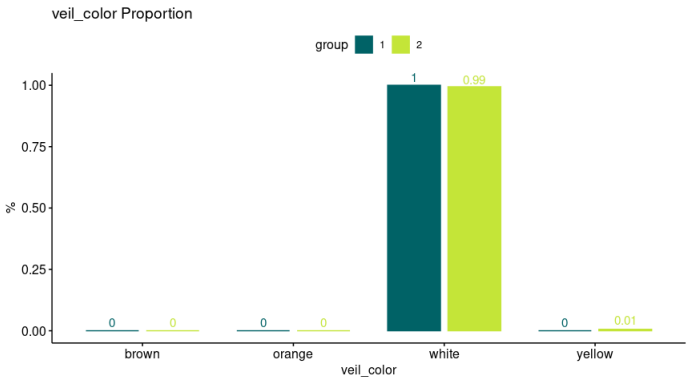


Figura A.14: Gráfico de barras de porcentaje de grupos para la variable veil_color.