UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE FACULTAD

Departamento de Departamento

Laboratorio N°1

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.

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. La idea es dividir la base de datos, para entrenar el algoritmo con una parte y evaluar el mismo con la parte restante. De esta manera poder predecir si un hongo es comestible o venenoso.

1.3 SOLUCIÓN PROPUESTA

En este trabajo se realizará un análisis exploratorio de la base de datos, a través de estadística descriptiva e inferencia estadística. La idea es lograr un primer acercamiento a la base de datos, entender su estructura y las relaciones entre sus variables. Finalmente se buscará ayuda de algún experto en el reino funji, para que pueda orientar a los *data scientist* a interpretar de mejor manera los resultados y su visualización.

1.4 OBJETIVOS Y ALCANCE DEL PROYECTO

1.4.1 Objetivo general

Estudiar e interpretar los datos correspondientes a cada base de datos. Para ello es necesario explicar de forma detallada el significado de clases, atributos y sus valores, lo que permitirá obtener el correcto análisis del problema planteado.

1.4.2 Objetivos específicos

1. Explicar de forma detallada los significados de clases, atributos y valores del *dataset*.
2. Analizar base de datos con estadística descriptiva e inferencia estadística.

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

Para la realización del estado del arte, se consultaron las publicaciones que utilizaron este dataset y han sido citadas en el repositorio de la University of California, Irvine (https://archive.ics.uci.edu/). Para el análisis exploratorio de los datos, se utilizarán medidas de centralización (media, moda y mediana), medidas de dispersión y tests de correlación de variables categóricas como el método Chi-cuadrado para probar asociación e independencia.

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.

1.6 ORGANIZACIÓN DEL DOCUMENTO

La idea de este documento es partir con la explicación del contexto en que se diseñó este dataset y la comprensión del problema a solución en función de los atributos. Luego se realizará un análisis exploratorio de los datos, con el fin de caracterizar cada variable en dependiendo si el registro corresponde a un hongo comestible o venenoso. Una vez que se conocen las características explicitas del dataset, se realizará un test de correlación chi-cuadradro de Pearson para variables categóricas, para encontrar la relación entre cada uno de los atributos. En la última instancia de este documento, se analizarán los resultados del proceso de exploración de los datos.

CAPÍTULO 2. DESARROLLO

2.1 DESCRIPCIÓN DE LA BASE DE DATOS

El dataset utilizado es obtenido desde el repositorio de bases de datos de la University of California, Irvine, repositorio con fines académicos en el ámbito de la ciencia de datos. El dataset consta de 23 atributos de 8124 observaciones de campo de hongos, donde se consideran los aspectos físicos de dichos organismos, como el color, textura, presencia de anillos o magulladuras entre otros. El atributo clave de este dataset, es ”type”, variable dummy que da cuenta si el hongo registrado es comestible o es venenoso.

Encontrar patrones o clasificaciones de los hongos que nos permitan identificar si un hongo es comestible o no, es bastante difícil sin realizar inteligencia sobre los datos. Para realizar estos procesos inteligentes es necesario lograr una comprensión del significado de los datos y como se relacionan entre ellos, y eso se logra con la etapa de exploración de datos

2.2 DESCRIPCIÓN DE VARIABLES

Para describir los atributos de la base de datos, primero nombraremos la variable, luego se da un breve detalle sobre que trata y finalmente se detallan los posibles valores que puede tomar. Cabe destacar que parte de la información se encuentra en el archivo adjunto con extensión ”.names” que se encuentra en el repositorio oficial de la base de datos y se mantiene la nomenclatura original tanto de las variables como de las categorías de esta.

A continuación, se presenta la siguiente tabla que describe las variables que posee la base de datos:

Variable	Descripción	Categorías
type	Indica si el hongo es comestible o venenoso	edible (e), poisonous (p)
cap_shape	Forma de sombrero	bell (b), conical (c), convex (x), flat (f), knobbed (k), sunken (s)
cap_surface	Superficie de sombrero	fibrous (f), grooves (g), scaly (y), smooth (s)
cap_color	Color de sombrero	brown (n), buff (b), cinnamon (c), gray (g), green (r), pink (p),purple (u), red (e), white (w), yellow (y)
has_bruises	Presencia de magulladuras	bruises (t), no (f)
odor	Hedor	almond (a), anise (l), creoseto (c), fishy (y), foul (f), musty (m), none (n), pungent (p), spicy (s)
gill_attachment	Láminas (agallas) adjuntas	attached (a), descending (d), free (f), notched (n)
gill_spacing	Espaciado entre láminas	close (c), crowded (w), distant (d)
gill_size	Tamaño de láminas	broad (b), narrow (n)
gill_color	Color de láminas	black (k), brown (n), buff (b), chocolate (h), gray (g), green (r), orange (o), pink (p), purple (u), red (e), white (w), yellow (y)
stalk_shape	Superficie de tallo	enlargin (e), tapering (t)
stalk_root	Tronco en raíz	bulbous (b), club (c), cup (u), equal (e), rhizomophs (z), rooted (r), missing (?)

Tabla 2.1: Descripción de las variables de la base de datos, parte 1.

Variable	Descripción	Categorías
stalk_surface_above_ring	Superficie de tallo sobre el anillo	fibrous (f), scaly (y), silky (k), smooth (s)
stalk_surface_below_ring	Superficie de tallo bajo el anillo	fibrous (f), scaly (y), silky (k), smooth (s)
stalk_color_above_ring	Color de tallo sobre el anillo	brown (n), buff (b), cinnamon (c), gray (g), orange (o),pink (p), red (e), white (w), yellow (y)
stalk_color_below_ring	Color de tallo bajo el anillo	brown (n), buff (b), cinnamon (c), gray (g), orange (o),pink (p), red (e), white (w), yellow (y)
veil_type	Tipo de velo	partial (p), universal (u)
veil_color	Color de velo	brown (n), orange (o), white (w), yellow (y)
ring_number	Número de anillos	none (n), one (o), two (t)
ring_type	Tipo de anillos	cowebby (c), evanescent (e), flaring (f), large (l), none (n), pendant (p), sheathing (s), zone (z)
spore_print_color	Color de espóras	black (k), brown (n), buff (b), chocolate (h), green (r),orange (o), purple (u), white (w), yellow (y)
population	Población	abundant (a), clustered (c), numerous (n), scattered (s), several (v), solitary (y)
habitat	Hábitat	grasses (g), leaves (l), meadows (m), paths (p), urban (u)waste (w), woods (d)

Tabla 2.2: Descripción de las variables de la base de datos, parte 2.

CAPÍTULO 3. ANÁLISIS DE LA BASE DE DATOS

3.1 ANÁLISIS ESTADÍSTICO

Primero analizamos cómo está distribuida la variable dependiente, en este caso, es ”type”.

Variable	Categoria	Cantidad	Porcentaje (%)
type	edible (e)	4208	51,8
	poisonous (p)	3916	48,2

Tabla 3.1: Detalle de la variable dependiente (type) de la base de datos.

En un gráfico de barras, distinguimos que es una base de datos balanceada respecto a la variable dependiente:

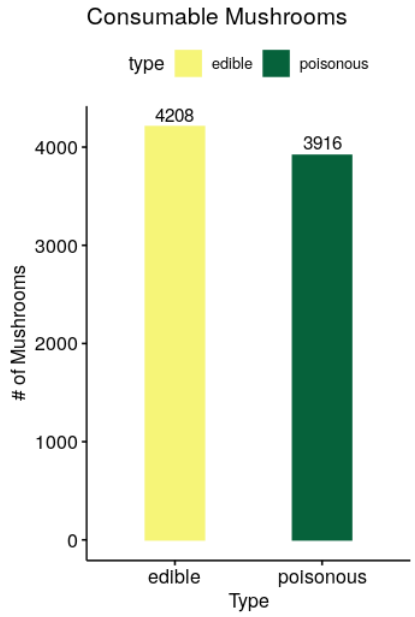

Figura 3.1: Gráfico de barras de la variable ”type”.
Fuente: Elaboración propia, 2020.

En base a esta variable, presentamos los siguientes gráficos de barras, los cuales muestran la distribución en las categorías correspondientes:

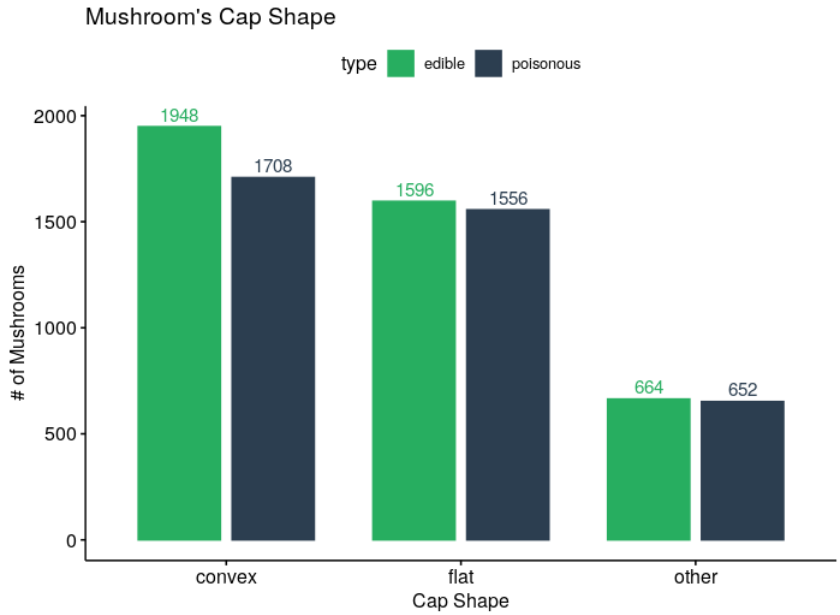

Figura 3.2: Gráfico de barras de la variable ”cap_shape”.
Fuente: Elaboración propia, 2020.

En la Figura 3.2, se muestra la frecuencia de la variable cap_shape, es decir ”forma de sombrero”. Podemos ver que los datos se encuentran balanceados en función de la variable dependiente. La categoría ”other” incluye a las categorías: ”bell”, ”conical”, ”knobbed” y ”sunken”.

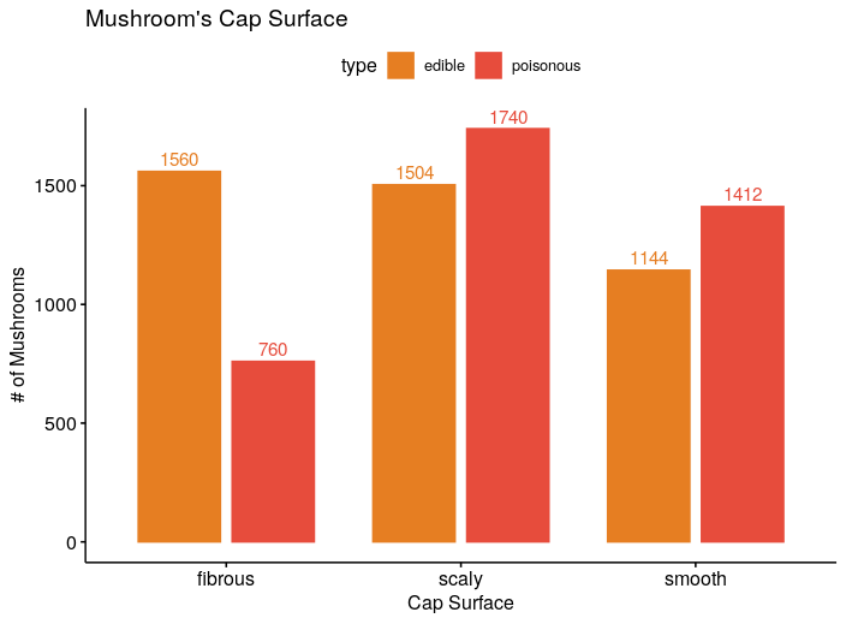

Figura 3.3: Gráfico de barras de la variable ”cap_surface”.
Fuente: Elaboración propia, 2020.

En la Figura 3.3 se muestra la frecuencia de la variable cap_surface o ”superficie de sombrero”. Se eliminaron las cuatro observaciones de la categoría ”grooves”. Esto debido a que no existían más categorías con las cuales reagruparlas.

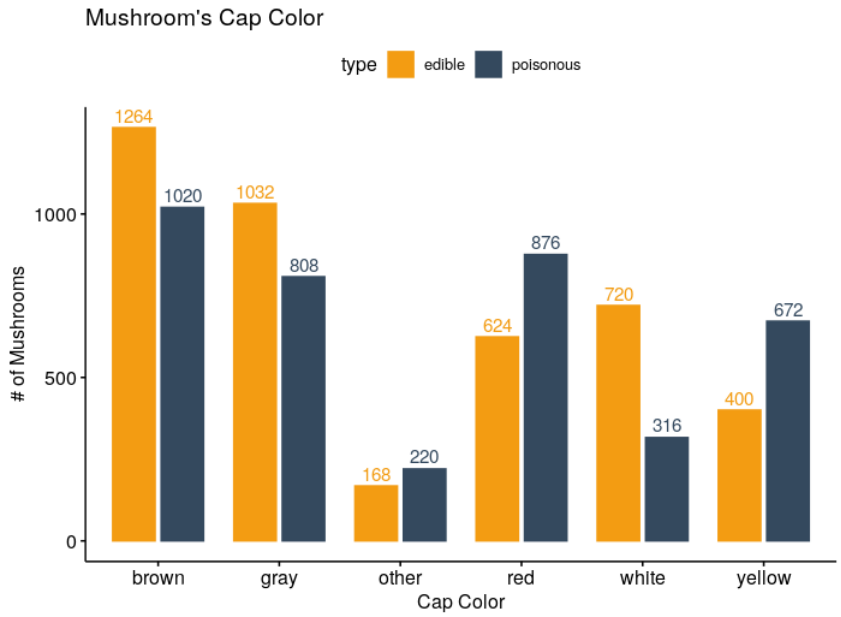

Figura 3.4: Gráfico de barras de la variable ”cap_color”.
Fuente: Elaboración propia, 2020.

En la Figura 3.4 se aprecia un gráfico de frecuencia de la variable cap_color o ”color

de sombrero". Se agruparon en "other" las categorías: "buff", "cinnamon", "green", "pink" y "purple".

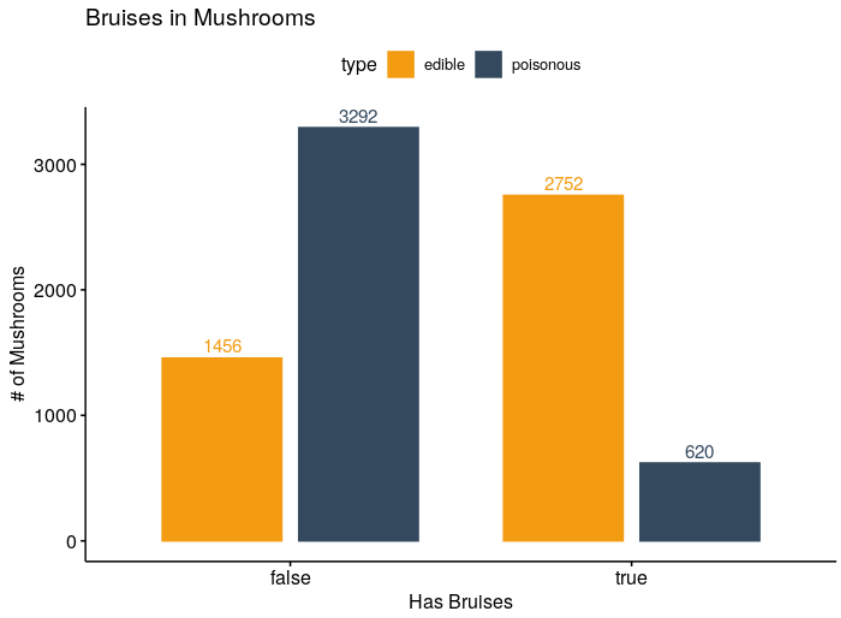

Figura 3.5: Gráfico de barras de la variable "has_bruises".
Fuente: Elaboración propia, 2020.

El gráfico de la Figura 3.5, visualiza la frecuencia de la variable independiente cap_has_bruises o "sombrero tiene moretón o magulladura".

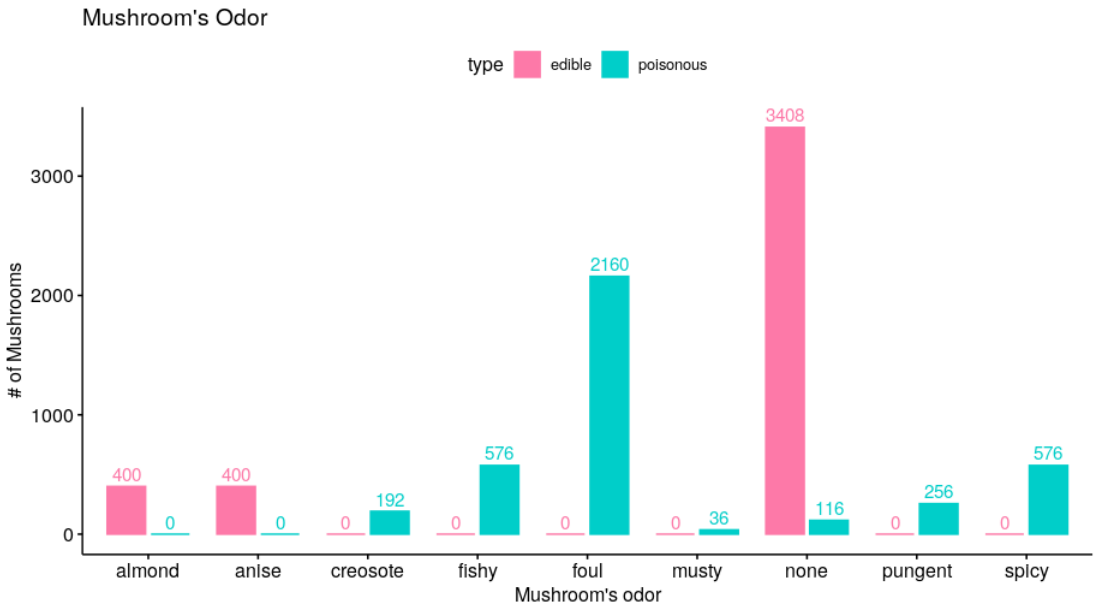

Figura 3.6: Gráfico de barras de la variable "odor".
Fuente: Elaboración propia, 2020.

En la Figura 3.6, se muestra el gráfico de frecuencias de la variable "Olor". No

se realizaron cambios a la distribución de esta variable debido a que cada clase de la variable independiente toma solo uno de los dos posibles valores de la variable dependiente, por lo que se sospecha que la correlación entre ambas variables es muy alta. También se puede inferir (según esta muestra), que el olor del hongo es un rasgo determinante para la clasificación del individuo como comestible o venenoso. Más adelante en el informe se demostrará su importancia a través de los estadísticos adecuados.

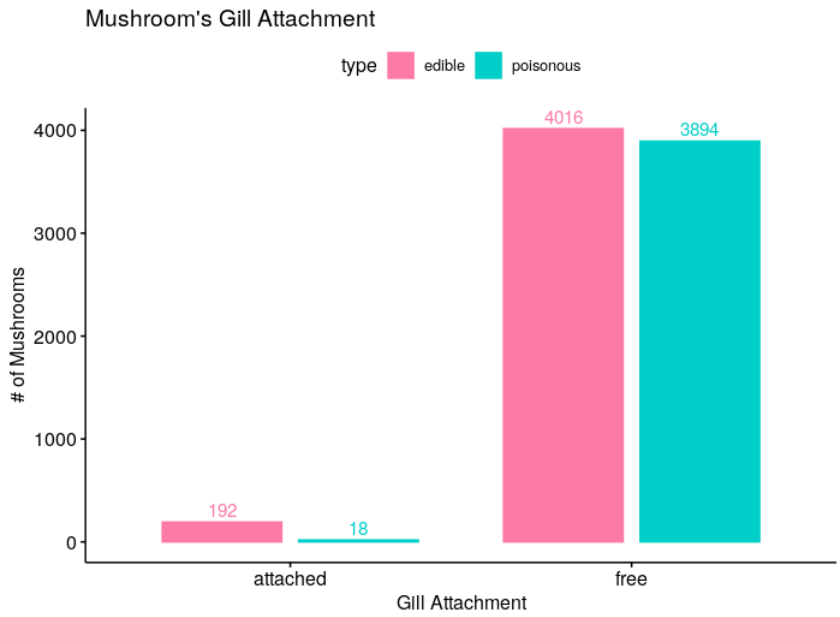

Figura 3.7: Gráfico de barras de la variable ”gill_attachment”.
Fuente: Elaboración propia, 2020.

En la Figura 3.7, se muestra la distribución de la variable ”gill_attachment” o ”Láminas adjuntas”. las categorías ”descending” y ”notched” no poseen observaciones en la base de datos, por lo que se excluyeron del gráfico.

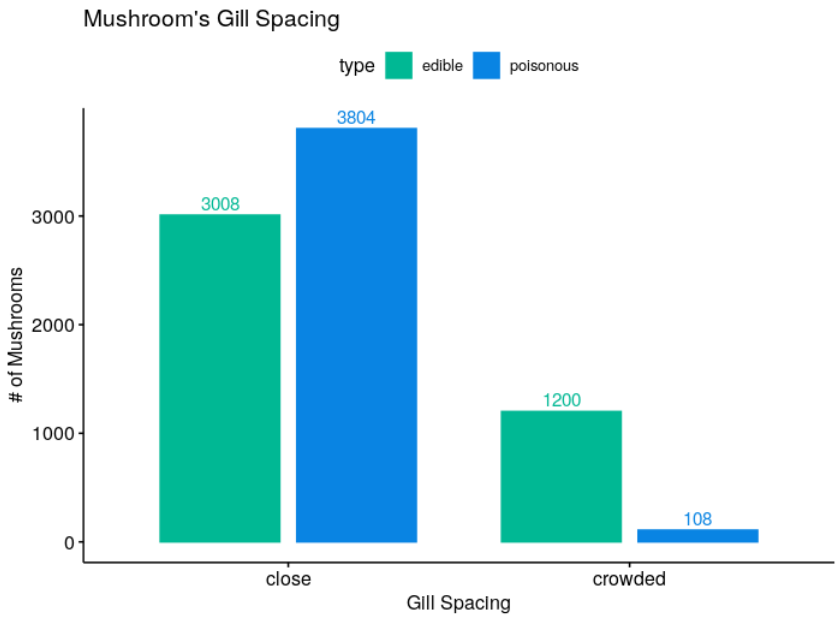

Figura 3.8: Gráfico de barras de la variable ”gill_spacing”.
Fuente: Elaboración propia, 2020.

En la Figura 3.8 se grafican las frecuencias de la variable ”gill_spacing” o ”Espaciado entre láminas”. De este gráfico y según la muestra estudiada, se puede inferir que si el individuo está repleto de láminas, existe una alta probabilidad de que sea comestible. La categoría ”distant” no posee observaciones en la base de datos, por lo que se excluyó del gráfico.

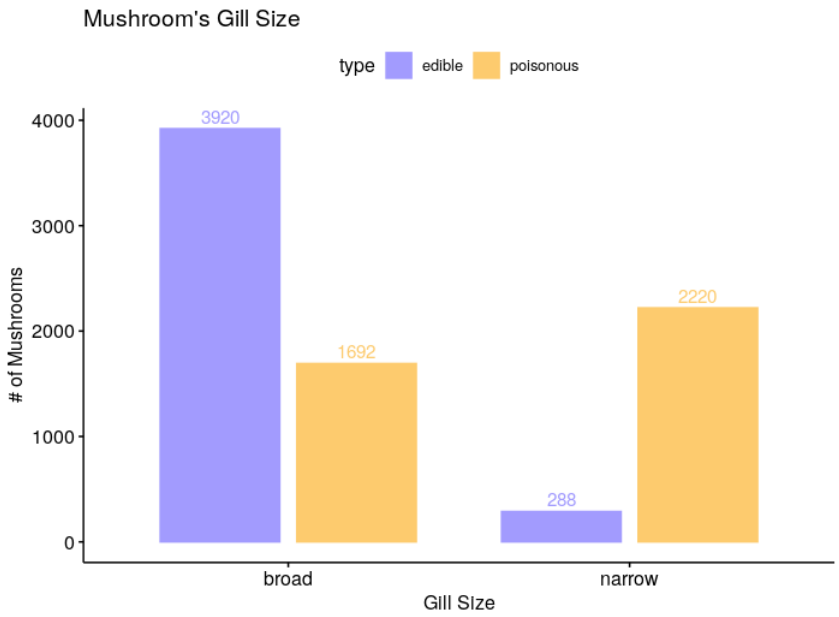

Figura 3.9: Gráfico de barras de la variable ”gill_size”.
Fuente: Elaboración propia, 2020.

La Figura 3.9 muestra la frecuencias de la variable ”gill_size” o tamaño de lámina.

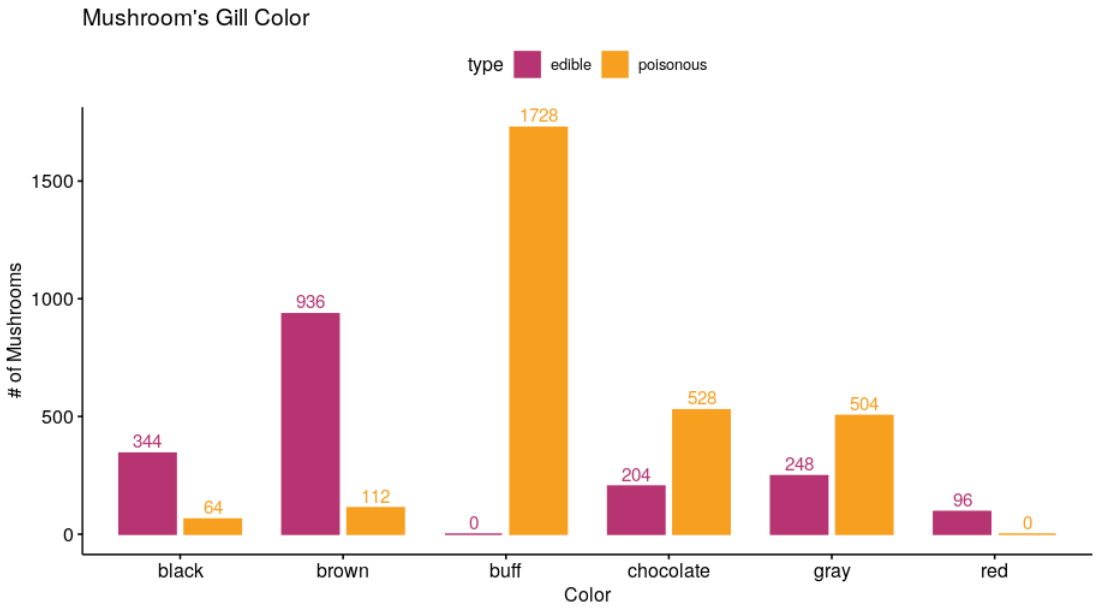

Figura 3.10: Gráfico de barras de la variable ”gill_color”, Parte 1.
Fuente: Elaboración propia, 2020.

La Figura 3.10 muestra la frecuencias de la variable ”gill_color” o color de la lámina. De este gráfico se puede inferir que si la muestra presenta un color de lámina ”beige”, entonces existe una muy alta probabilidad de que sea un individuo tóxico.

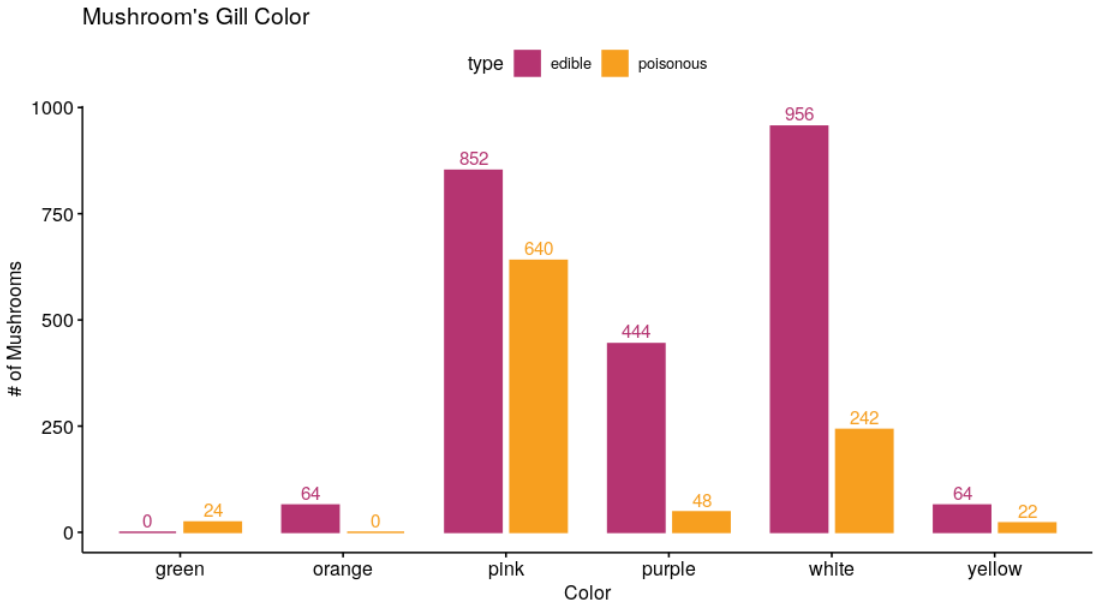

Figura 3.11: Gráfico de barras de la variable ”gill_color”, Parte 2.
Fuente: Elaboración propia, 2020.

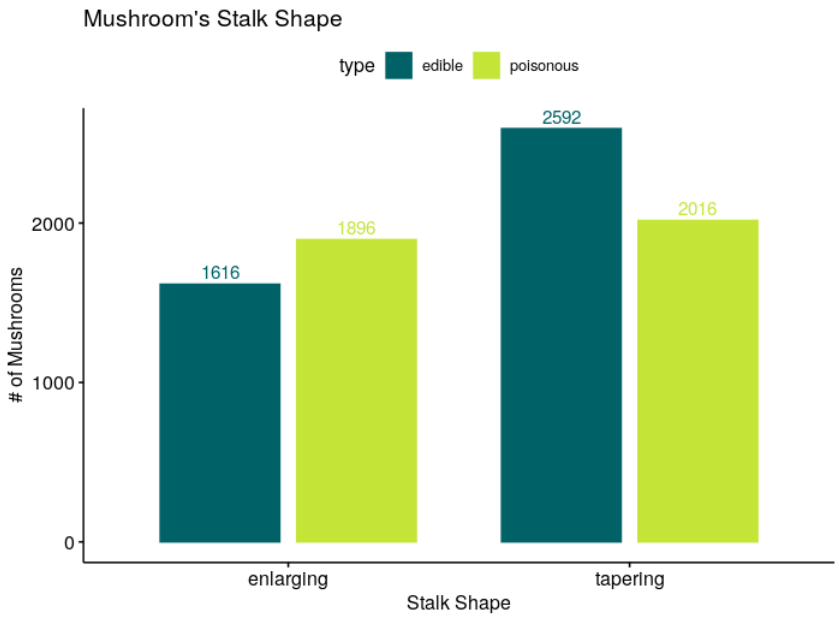

Figura 3.12: Gráfico de barras de la variable ”stalk_shape”.
Fuente: Elaboración propia, 2020.

La Figura 3.12 muestra el gráfico de distribución de la variable ”stalk_shape” o ”Forma del tallo”.

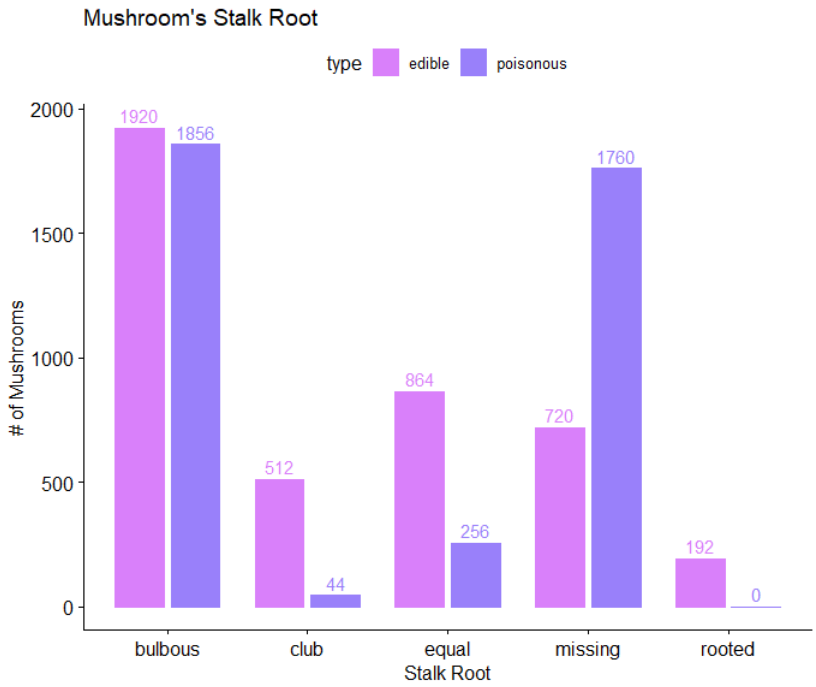

Figura 3.13: Gráfico de barras de la variable ”stalk_root”.
Fuente: Elaboración propia, 2020.

En la Figura 3.13 se puede visualizar las frecuencias de la variable ”stalk_root”

o ”tallo en la raíz”. Para stalk_root, no existían registros para las categorías ”rhizomorphs” y ”cup”. Además, existen 2480 observaciones catalogadas como ”missing”, lo cual constituye aproximadamente el 31% de la base de datos.

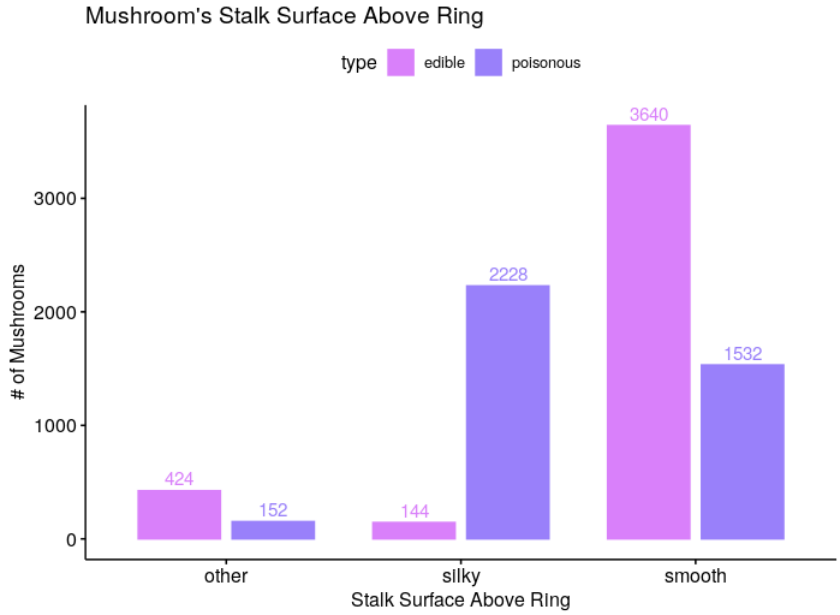

Figura 3.14: Gráfico de barras de la variable ”stalk_surface_above_ring”.
Fuente: Elaboración propia, 2020.

En la Figura 3.14 se muestra el gráfico de frecuencia de la variable stalk_surface_above_ring o ”superficie de tallo sobre anillo”.En la categoría ”other” se agruparon las clases ”fibrous” y ”scaly”.

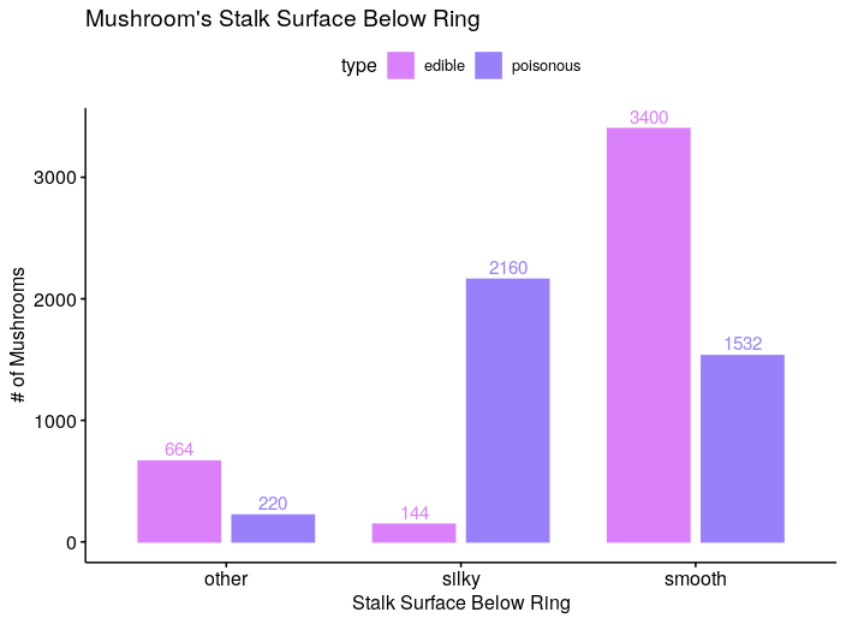

Figura 3.15: Gráfico de barras de la variable ”stalk_surface_below_ring”.
Fuente: Elaboración propia, 2020.

En la Figura 3.15 se muestran las frecuencias de la variable stalk_surface_below_ring o ”superficie de tallo bajo el anillo”. Se agruparon en ”other” las categorías ”fibrous” y ”scaly”.

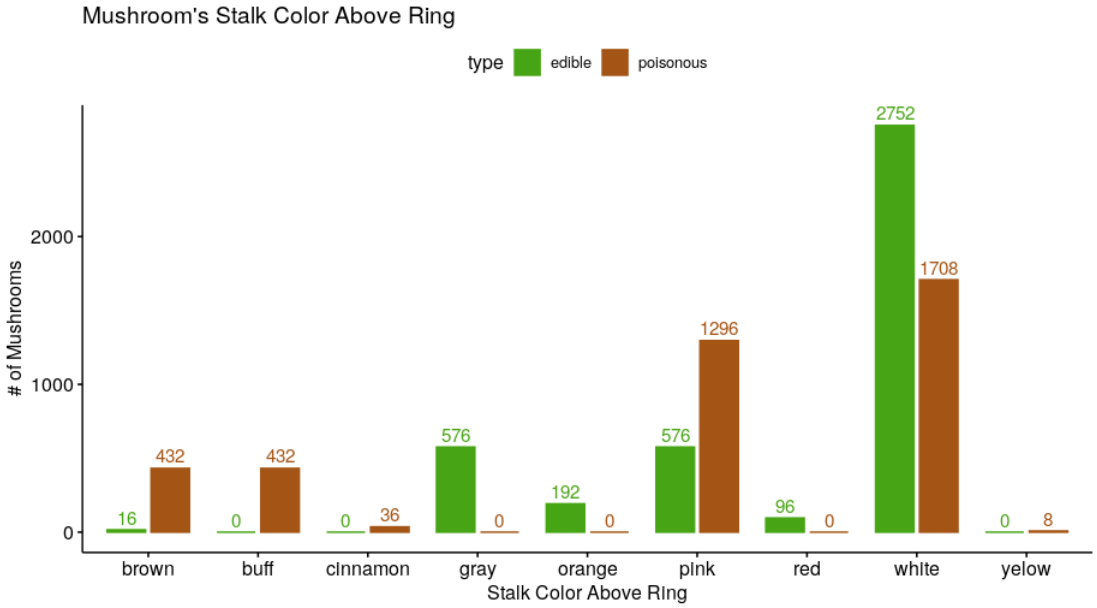

Figura 3.16: Gráfico de barras de la variable ”stalk_color_above_ring”.
Fuente: Elaboración propia, 2020.

En la Figura 3.16 se muestran las frecuencias de la variable stalk_color_above_ring o ”Color de tallo sobre el anillo”.

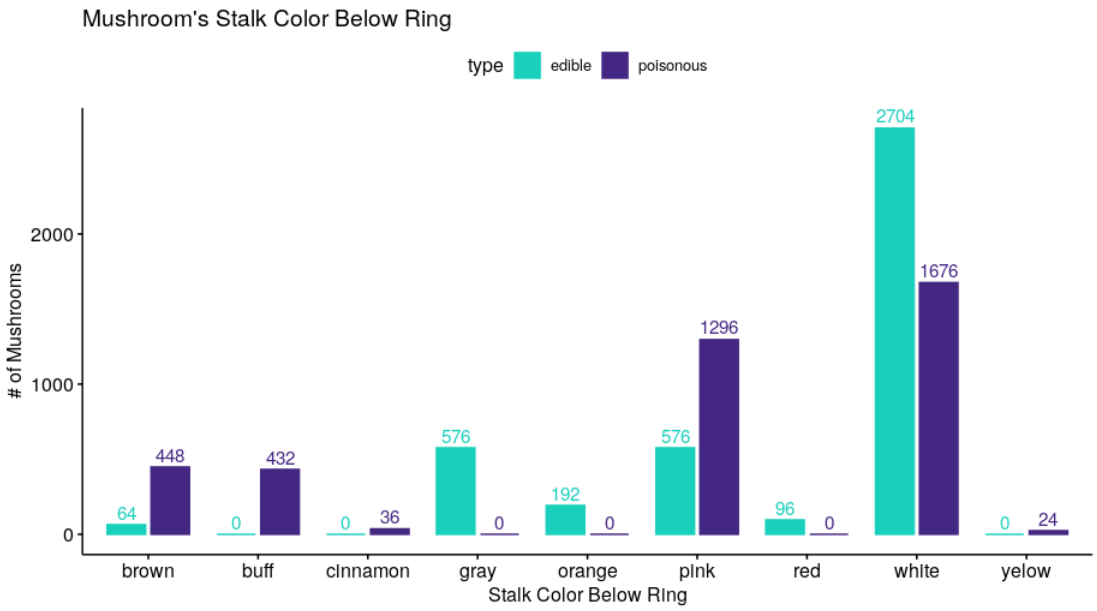

Figura 3.17: Gráfico de barras de la variable ”stalk_color_below_ring”.
Fuente: Elaboración propia, 2020.

En la Figura 3.17 se muestran las frecuencias de la variable stalk_color_below_ring o ”Color de tallo bajo el anillo”.

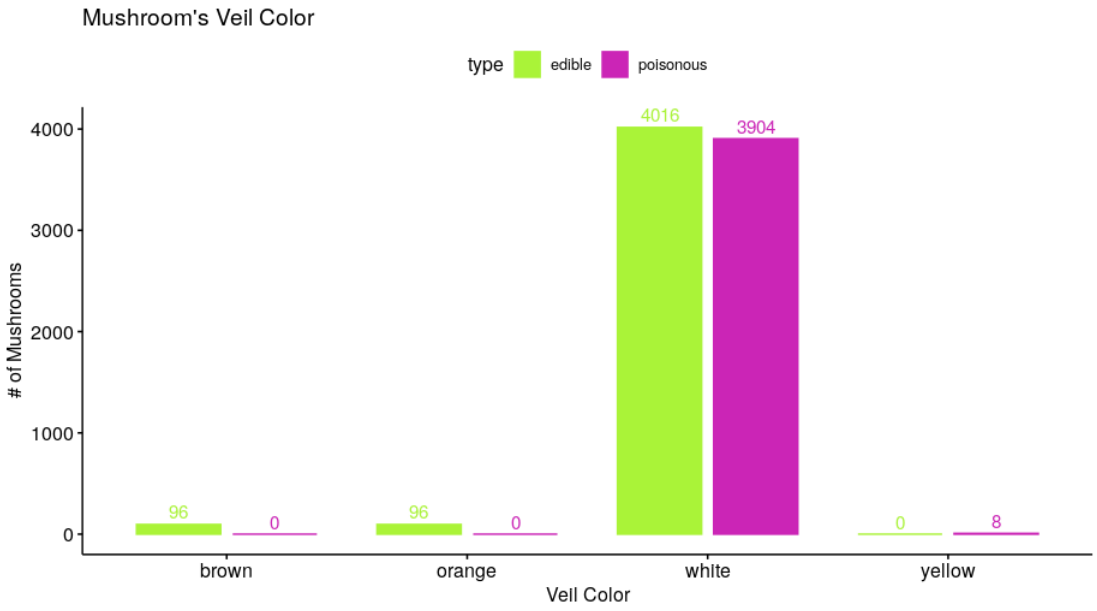

Figura 3.18: Gráfico de barras de la variable ”veil_color”.
Fuente: Elaboración propia, 2020.

En la Figura 3.18 se muestra un gráfico de frecuencias para la variable veil_color o ”Color del velo”. Se puede apreciar que la gran mayoría de los registros corresponden a la clase

”white” y presenta un equilibrio en cuanto a la variable dependiente.

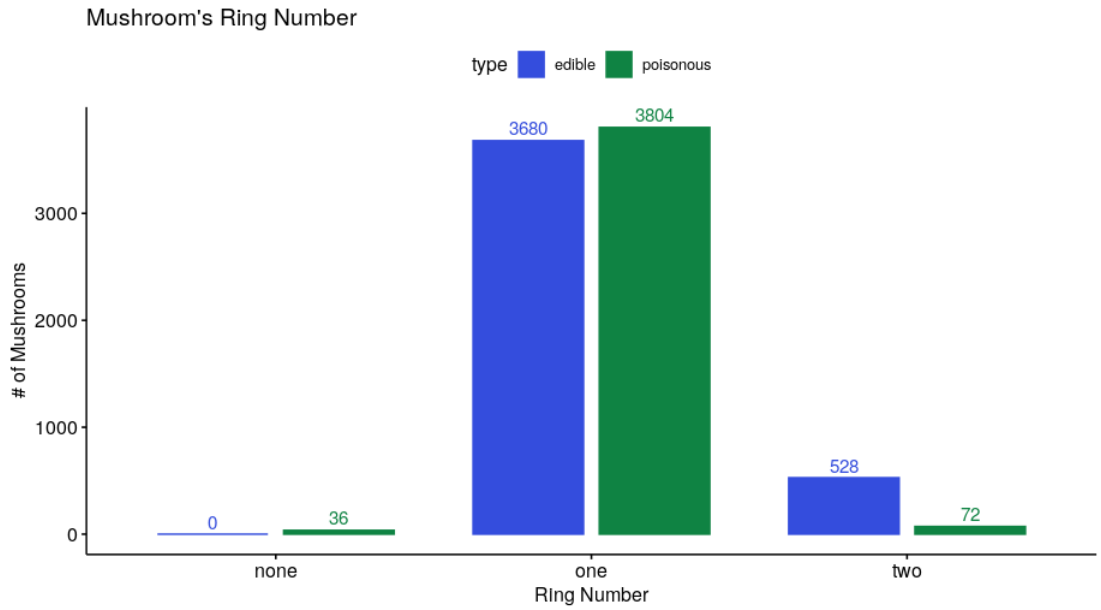

Figura 3.19: Gráfico de barras de la variable ”ring_number”.
Fuente: Elaboración propia, 2020.

La Figura 3.19 muestra un gráfico de la distribución de la variable ring_number o ”Número de anillos”. La gran mayoría de registros asociados a esta variable presentan 1 anillo.

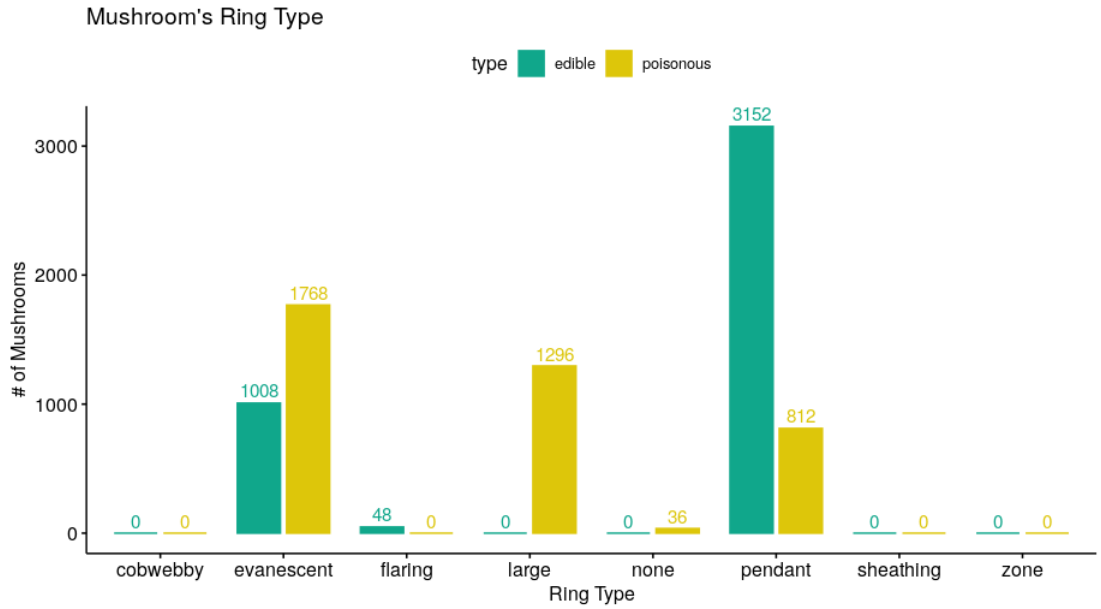

Figura 3.20: Gráfico de barras de la variable ”ring_type”.
Fuente: Elaboración propia, 2020.

En la Figura 3.20 se visualiza un gráfico de la frecuencias de la variable ring_type o

”Tipo de anillos”. Se destaca la categoría ”large”, donde todas las muestras que poseían un anillo largo, se clasificaron como venenosas.

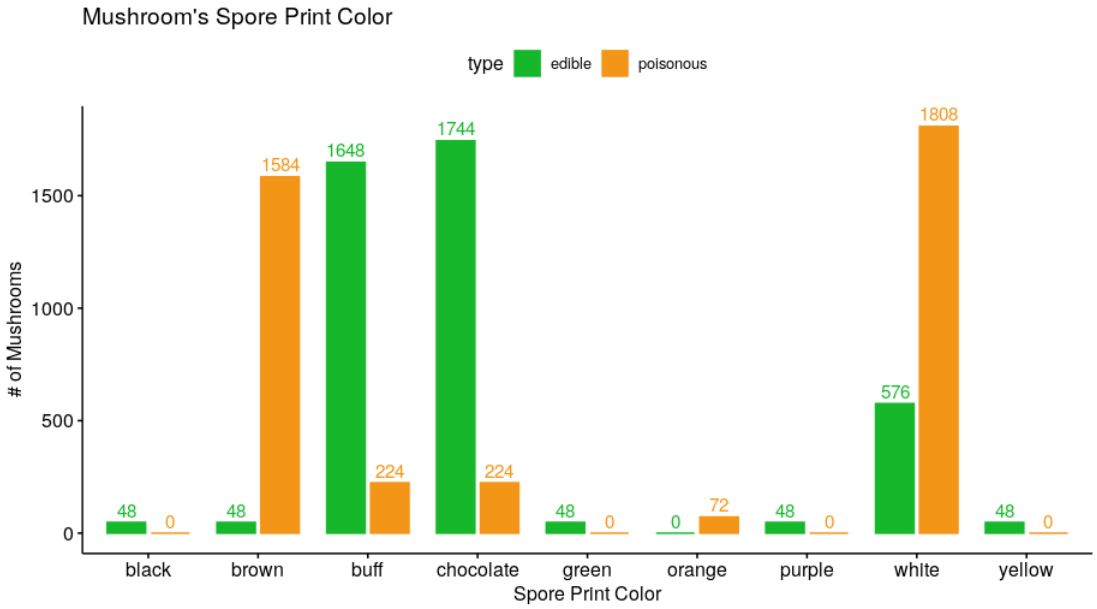

Figura 3.21: Gráfico de barras de la variable ”spore_print_color”.
Fuente: Elaboración propia, 2020.

En la Figura 3.21 se presenta un gráfico de la distribución de la variable sporte_print_color o ”Color de esporas”. Se puede observar que la mayoría de las clases tiene tendencia hacia uno de los dos posibles valores de la variable ”type”, por lo que se puede suponer, en función de esta muestra, que esta variable, tiene una alta correlación con el tipo de hongo y puede ser una variable clave para realizar una futura regresión o clasificación.

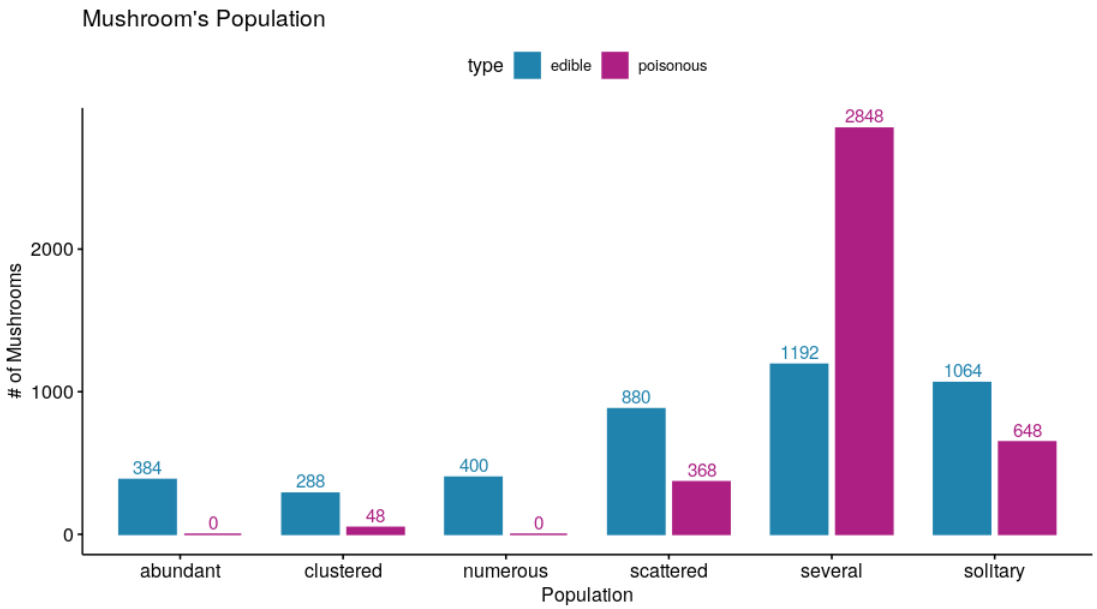

Figura 3.22: Gráfico de barras de la variable ”population”.
Fuente: Elaboración propia, 2020.

La Figura 3.22 muestra las frecuencias de la variable ”population” o ”Población”.

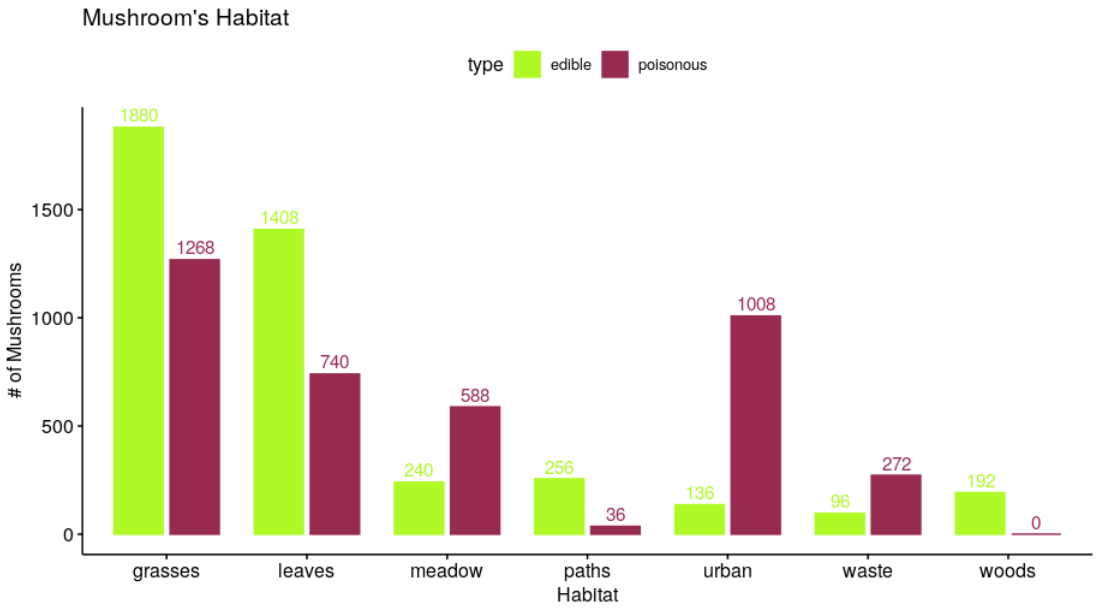

Figura 3.23: Gráfico de barras de la variable ”habitat”.
Fuente: Elaboración propia, 2020.

La Figura 3.23 muestra las frecuencias de la variable ”habitat”.

Por último, la variable veil_type, solo presentó registros de la clase partial (p), por lo

que no fue necesario incluir un gráfico de dicho atributo.

3.2 ANÁLISIS INFERENCIAL

- Prueba de Chi-Cuadrado

  El test de Chi-Cuadrado (χ^2) es un test de hipótesis estadística que asume una hipótesis nula en que las frecuencias observadas de una variable categórica coinciden con la frecuencia esperada para esa variable categórica J. (2019). Este test nos permite entonces conocer la dependencia entre variable rechazando la hipótesis nula, sin embargo, cabe destacar que no nos entrega la fortaleza de la relación, es decir, el valor entregado indica una relación, pero no lo cuantifica.

  A continuación, se presentan los resultados obtenidos de los test, cabe destacar que solo se tomo en cuenta la relación entre la variable dependiente (en este paso ”type”) y el resto de variables, por lo que se excluye la relación consigo misma.

X	**Y**	**Resultado**	**Grados de libertad**	***p-value***
type	odor	7670,7	8	2,20e-16
type	spore_print_color	4598,8	8	2,20e-16
type	gill_color	3771,7	11	2,20e-16
type	ring_type	2962,5	4	2,20e-16
type	stalk_surface_above_ring	2811,5	2	2,20e-16
type	stalk_surface_below_ring	2687,3	2	2,20e-16
type	gill_size	2362,8	1	2,20e-16
type	stalk_color_above_ring	2239,8	8	2,20e-16
type	stalk_color_below_ring	2154,3	8	2,20e-16
xtype	has_bruises	2047,8	1	2,20e-16
type	population	1937,1	5	2,20e-16
type	habitat	1570,9	6	2,20e-16
type	stalk_root	1344,3	4	2,20e-16
type	gill_spacing	993,31	1	2,20e-16
type	cap_shape	490,18	5	2,20e-16
type	cap_color	318,83	5	2,20e-16
type	ring_number	374,32	2	2,20e-16
type	cap_surface	310,75	2	2,20e-16
type	veil_color	191,05	3	2,20e-16
type	gill_attachment	133,82	1	2,20e-16
type	stalk_shape	83,235	1	2,20e-16

Tabla 3.2: Resultado de los test de Chi-Cuadrado realizados.

Los resultados obtenidos, presentados en la Tabla 3.2, indican que todas las variables rechazan la hipótesis nula (las variables comparadas son independientes), pero existen variables las cuales se puede inferir que poseen una mayor relevancia, estas serían ”odor”, ”spore_print_color”, ”gill_color”, ”ring_type”, etc. Las variables que presentaron una menor relevancia fueron ”gill_attachment” y ”stalk_shape”.

Además, destacar que la variable ”veil_type” no se incluyó en los test debido a que no poseía observaciones con más categorías que ”partial”.

- Prueba Cramer's V

  La prueba de Cramers's V ϕ_c es una medida de asociación entre dos variables categóricas entregando un valor entre 0 y 1, donde en 0 no existe asociación entre las variables y 1 cuando existe una completa asociación y sucede cuando una variable está totalmente determinada por otra.

  La matriz de correlación según Cramer's V consta de 253 registros entre las 23 variables. Al ser tantos registros se decidió mostrar sólo las correlaciones que contemplen la variable dependiente ”type”, el cual da cuenta si el hongo registrado es comestible o no. Dichas correlaciones se muestran en la Tabla 3.3.

Y	X	CramersV
type	odor	0.97
type	spore_print_color	0.75
type	gill_color	0.68
type	ring_type	0.60
type	stalk_surface_above_ring	0.59
type	stalk_surface_below_ring	0.57
type	gill_size	0.54
type	stalk_color_above_ring	0.52
type	stalk_color_below_ring	0.51
type	has_bruises	0.50
type	population	0.49
type	habitat	0.44
type	stalk_root	0.41
type	gill_spacing	0.35
type	cap_shape	0.25
type	cap_color	0.22
type	ring_number	0.21
type	cap_surface	0.20
type	veil_color	0.15
type	gill_attachment	0.13
type	stalk_shape	0.10
type	veil_type	0.04

Tabla 3.3: Resultado de los test de Cramer realizados.

De la tabla anterior se puede destacar la correlación existente entre la variable dependiente ”type” y la variable independiente ”odor”, la cual tiene un valor de 0.97, es decir una dependencia casi total entre ambos atributos. Esto era de esperar y fue propuesta en la sección 3.1 de análisis estadístico, donde se muestra el gráfico de la variable ”odor” y se aprecia que para cada clase, la variable dependiente toma solo uno de los dos posibles valores. Un caso parecido pero en menor grado, es el de la variable ”spore print color”, la

cual muestra un comportamiento similar, en cuanto a la tendencia de la variable dependiente a tomar solo uno de los posibles valores, pero en menor medida que la variable odor.

CAPÍTULO 4. CONCLUSIONES

La base de datos ha sido ampliamente estudiada desde su creación. Varios autores estudian la efectividad de alguna técnica de clasificación a través de esta base de datos, lo que nos da una linea base sobre la cual comparar los resultados obtenidos de las técnicas que aplicaremos en las siguientes experiencias de laboratorio.

En el análisis exploratorio, permitió un primer acercamiento a cada variable, incluso la variable ”odor” mostró tempranamente su importancia a través de un gráfico de frecuencias.

Por otro lado, los resultados obtenidos de los test de Chi-Cuadrado como de Cramers V coinciden con la literatura consultada, en la cual la variable ”odor” es la que más influye al momento de decidir si un hongo es comestible o no.

Este informe de exploración de variables será de vital importancia para los siguientes informes de clusterización y regresión utilizando esta base de datos.

REFERENCIAS BIBLIOGRÁFICAS

J., B. (2019). A gentle introduction to the chi-squared test for machine learning. Recuperado el Sábado 16 de Mayo de 2020, de https://machinelearningmastery.com/chi-squared-test-for-machine-learning/.