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

Laboratorio 1: Análisis Estadístico

Integrantes:	Nicolás López
	Alberto Rodríguez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor:	Max Chacón
Ayudante:	Javier Arredondo

20 de Noviembre de 2020

Tabla de contenidos

1. Introducción

Para poder sumergirse en la ciencia del análisis de datos, se debe tener la capacidad de estudiar e interpretar la información de alguna base de datos. Es importante que la información entregada, deben ser los resultados o referencias de un estudio serio. Los también estudios pueden ser realizados mediante técnicas de estadística descriptiva e inferencial.

El objetivo general de esta experiencia es introducirse en lo que es el análisis e interpretación de datos, para nuestro trabajo el set de datos corresponde a las características presentes en hongos comestibles y venenosos. Se trabajará con este set en las siguientes experiencias del curso.

Para poder observar estos datos se utiliza el lenguaje de programación R y en este también se utilizan herramientas estadísticas tales como distribución de probabilidades, matrices de confusión. Todo lo anterior se usa de base para realizar una prueba de hipótesis que nos ayude a determinar cuales hongos son comestibles y cuales son venenosos.

2. Descripción del Problema

El conjunto de datos trabajar corresponde a información respecto a especies de hongos, que según sus atributos se debe determinar si el hongo es comestible o es venenoso.

2.1. Descripción de la base de datos

La base de datos a trabajar corresponde a un conjunto de datos obtenidos a partir de 23 muestras hipotéticas de hongos con branquias de la familia Agaricus y Lepiota. La finalidad de dicha estructura de datos es identificar cada especie como comestible o venenosa. Cabe mencionar que en la clase de venenosa entran los hongos que se categorizan como definitivamente venenosos o de comestibilidad desconocida.

La identificación de cada hongo se entrega gracias a que la base de datos consta de 22 atributos que según los resultados que se obtengan, se indica si el individuo pertenece a una de las dos categorías mencionadas con anterioridad. Sin embargo, la guía establece claramente que no existe una regla simple para declarar la comestibilidad de un hongo.

El set de datos a trabajar consta de 8124 muestras. Por otro lado, las variables con las que se estudian los datos son todas de tipo categóricas, esto quiere decir que no toman valores numéricos, sino que solo valores descriptivos. Cabe destacar, que existe un atributo que puede faltar en ciertas muestras y este pertenece a la raíz del tallo (stalk-root), donde su representación en caso de faltar se representa como faltante (´´?").

2.2. Descripción de clases y variables

Como se mencionaba anteriormente, las clases pertenecen a la categorización final que se le entrega al hongo estudiado, que puede ser comestible o venenoso. Esta identificación se le da a partir de los 22 atributos que contiene la base de datos. A continuación se describe con mayor detalle todos los atributos con los que cuenta el set de datos de hongos.

Nombre del atributo	Descripción	Posibles valores
Forma de tapa	Forma que tiene la copa del hongo, observado en todos sus ángulos	cónica(c), campana(b), convexa(x), plana(f), nudosa(k) y hundida(s)
Superficie de tapa	Textura que presenta la copa del hongo al tacto	fibrosa(f), ranuras(g), escamosa(y) y suave(s)
Color de tapa	Tonalidad que tiene la superficie del ejemplar	marrón(n), beige(b), canela(c), gris(g), verde(r), rosa(p), violeta(u), rojo(e), blanco(w) y amarillo(y)
Moretones	Atributo que indica si el hongo presenta manchas o protuberancias	moretones(t) y no(f)
Olor	Aroma que expele el ejemplar en general	almendra(a), anís(l), creosota(c), pescado(y), sucio(f), mohoso(m), ninguno(n), picante(p)
Branquias	Forma que presentan las agallas que se encuentran bajo la copa del hongo	adjunto(a), descendente(d), libre(f) y con muescas (n)
Espacio entre branquias	Área que hay entre las agallas bajo la copa del hongo	cerca(c), apiñado(w) y distante(d).
Tamaño de branquias	Atributo que indica el porte de las branquias, que puede variar en su ancho o largo	ancho(b) y estrecho(n)

Cuadro 1: Atributos de la base de datos

Nombre del atributo	Descripción	Posibles valores
Forma de tapa	Forma que tiene la copa del hongo, observado en todos sus ángulos	cónica(c), campana(b), convexa(x), plana(f), nudosa(k) y hundida(s)
Color de branquias	Tonalidad que presentan las agallas	negro(k), marrón(n), ante(b), chocolate(h), gris(g), verde(r), naranjo(o), rosa(p), púrpura(u), rojo(e), blanco(w) y amarillo(y).
Forma del tallo	Contextura que tiene el pedúnculo del hongo	agrandado(e) y ahusado(t)
Raíz del tallo	Se obtiene al sacar el ejemplar de su posición, es posible encontrar al ejemplar sin este atributo	bulboso(b), club(c), copa(u), igual(e), rizomorfos(z), enraizado(r) y faltante(?)
Superficie sobre el anillo	Los hongos cuentan con un anillo en el tallo y este atributo muestra la textura sobre este	fibroso(f), escamoso(y), sedoso(k) y liso(s)
Superficie bajo el anillo	Los hongos cuentan con un anillo en el tallo y este atributo muestra la textura bajo este	fibroso(f), escamoso(y), sedoso(k) y liso(s)
Color del tallo sobre el anillo	Tonalidad del pedúnculo sobre la parte del anillo	marrón(n), ante(b), canela(c), gris(g), naranja(o), rosa(p), rojo(e), blanco(w) y amarillo(y)

Cuadro 2: Atributos de la base de datos

Nombre del atributo	Descripción	Posibles valores
Color del tallo bajo del anillo	Tonalidad del pedúnculo bajo la parte del anillo	marrón(n), beige(b), canela(c), gris(g), naranja(o), rosa(p), rojo(e), blanco(w) y amarillo(y)
Tipo de velo	Indica como es el velo del hongo, este se presenta bajo de la copa como una red que cubre el tallo	parcial(p) y universal(u)
Color de velo	Indica la tonalidad del velo mencionado anteriormente	marrón(n), naranja(o), blanco(w) y amarillo(y)
Número de anillos	Cantidad de anillos presentes en el tallo del hongo	ninguno(n), uno(o) y dos(t)
Tipo de anillo	Los anillos del tallo pueden presentar diferentes formas y este atributo indica cual es	telaraña(c), evanescente(e), flameado(f), grande(l), ninguno(n), colgante(p), revestimiento(s) y zona(z)
Color de esporas	Expresa la tonalidad de las esporas que tiene el hongo	negro(k), marrón(n), ante(b), chocolate(h), verde(r), naranja(o), morado(u), blanco(w) y amarillo(y)
Población	Este atributo muestra la cantidad de ejemplares en una misma ubicación	abundante(a), agrupada(c), numerosa(n), dispersa(c), varias(v) y solitaria(y)
Hábitat	Indica el ambiente donde se encuentra ubicado el hongo estudiado	pastos(g), hojas(l), prados(m), caminos(p), urbano(u), residuos(w) y bosques(d)

Cuadro 3: Atributos de la base de datos

3. Análisis Estadístico e Inferencial

Un estudio utilizando redes de retropropagración, extrajo reglas lógicas las cuales permiten determinar si un hongo es venenoso (Wlodzislaw D, 1997).

Estas reglas lógicas se extraen a partir de datos de entrenamiento, es decir, estas reglas son aplicadas una después de la otra a los casos que no fueron reconocidos como venenosos por la regla anterior.

En el siguiente cuadro (Cuadro 4), se muestran los resultados obtenidos en cada regla a partir de los datos de los hongos.

Reglas	**Falsos Negativos**	**Precisión**
R1: odor = $\neg$(almond $\vee anise \vee none$)	120	98.52 %
R2: spore.print.color = green	48	99.41 %
R3 : odor = none $\wedge stalk.surface.below.ring = scaly \wedge$ ($stalk.color.above.ring = brown$)	8	99.90 %
R4: habitat = leaves $\wedge cap.color = white$	0	100 %

Cuadro 4: Resultados de cada regla lógica

Se puede observar que la regla R1 ya es muy precisa porque obtiene la mayoría de los casos con solo verificar el atributo olor ($odor$). Por lo tanto se quiere sabores que valores del atributo olor influyen en cada una de las clases de hongos. Antes de aplicar las reglas podemos analizar el atributo olor y que valores influyen en la clase de hongo.

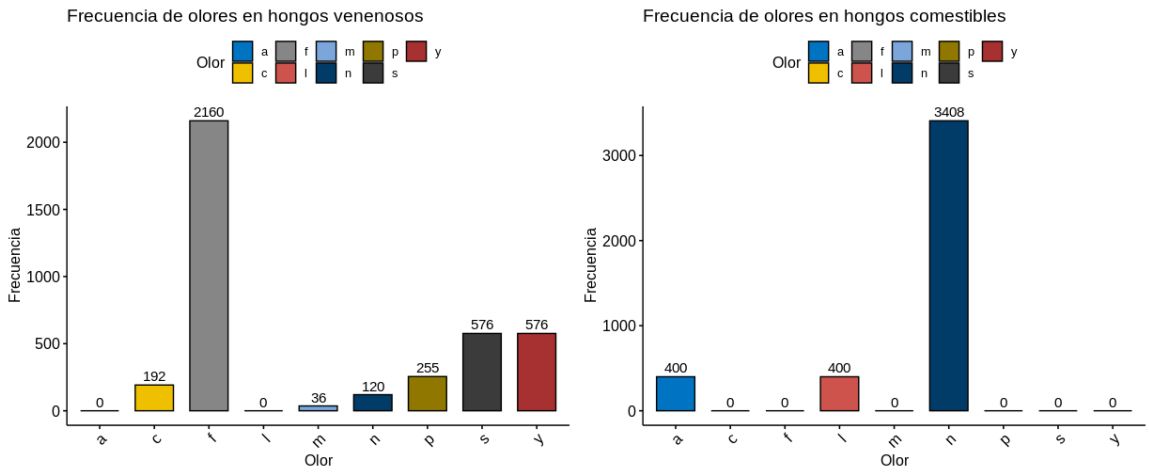

Figura 1: Gráficos de frecuencia de olor para cada clase de hongo

Se puede observar en los gráficos (Figura 1) que existe una clara diferencias entre los olores en cada gráfico. Como gracias a ciertos valores del atributo olor es mas probable que el hongo sea comestible o venenoso.

3.1. Regla R1

Como se dijo anteriormente la regla R1 es muy precisa, por lo que se realiza una matriz de confusión (Cuadro 5) y un conjunto de métricas(Cuadro 6), las cuales son.

1. **Exactitud:** Se refiere a lo cerca que está el resultado de una medición del valor verdadero. En este caso el porcentaje de hongos que se encontraron correctamente.

2. **Precisión:** Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Los hongos no estaban nada dispersos, por lo que la precisión es perfecta.

3. **Sensibilidad:** Tasa de verdaderos positivos (venenosos verdaderos). Es la proporción de casos positivos que fueron correctamente identificados.

4. **Especificidad:** Tasa de verdaderos negativos (comestibles verdaderos). Casos negativos que el modelo ha clasificado correctamente.

	Venenosos	Comestibles
Venenosos	3795	0
Comestibles	120	4208

Cuadro 5: Matriz de confusión

Métrica	**valor**
Exactitud	98.52 %
Precisión	100 %
Sensibilidad	96.93 %
Especificidad	100 %
Índice de éxito	96.93 %

Cuadro 6: Métricas

Viendo las métricas y la matriz de confusión, se puede ver que la regla es muy precisa, pero no se sabe si para otro conjunto de datos pueda arrojar lo mismo, ya que la regla es aplicada al mismo conjunto de datos sobre la cual es entrenada. Puede ser aceptada en el caso de que con la cantidad de datos es suficiente para validarlo, pero no es lo mas recomendado, lo mejor sería tener mas conjuntos de datos independiente del actual para poder ver si se obtienen los mismos resultados.

3.2. Regla R2

Se realiza dos gráficos para ver la frecuencia de cada valor del atributo Color de esporas (*spore-print-color*) dependiendo de que clase de hongo es. (Cuadro 4).

Se puede observar que este atributo no es tan claro como el atributo olor de la regla anterior, existen algunos valores del atributo (como el color blanco) que existe una buena cantidad en ambas clases. Se puede hacer lo mismo que en el caso anterior (matriz de confusión y métricas), pero no es tan importante porque ya la regla R1 abarca casi todos los datos, y no cambia mucho la clasificación.

No se evaluaran las siguientes reglas R, porque como se dijo anteriormente solo abarca una minoría de los datos totales.

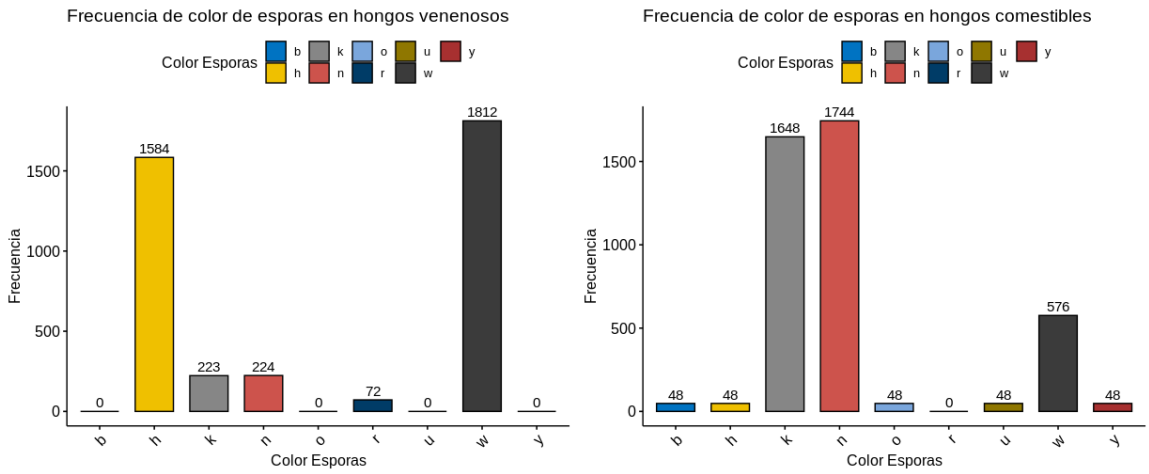

Figura 2: Gráficos de frecuencia del color de esporas para cada clase de hongo

3.3. Análisis del olor de cada hongo

Se quiere verificar si hay una relación significante entre el olor y la clase de hongo. Por lo que el test mas indicado para realizarlo es el test de Fisher, ya que este se utiliza si se quiere estudiar si existe asociación entre 2 variables categóricas, es decir, las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable (Soetewey, 2020).

Para poder usar este test se deben cumplir algunas condiciones (Amat, 2016), las que iremos evaluando.

- **Muestras deben ser independientes:** Se asume que las observaciones son independientes entre si, ya que proviene de un estudio serio.
- **EL tamaño de la muestra es mayor al 10 % de la población:** También se debe asumir ya que son 8124 observaciones, un número bastante grande.
- **Cada observación contribuye únicamente a uno de los niveles:** Se cumple
- **Las frecuencias marginales de columnas y filas tienen que ser fijas:** Se cumple

Las hipótesis serían:

- **H0:** El olor no determina la clase de hongo. Las variables son independientes
- **H1:** El olor determina la clase de un hongo. Las variables no son independientes

El p-valor arrojado por el test de Fisher es de 0.0005, aunque se use un nivel de significancia exigente, por ejemplo 0.01. el p-valor obtenido es menor, por lo que se rechaza H0, en otras palabras, existe una dependencia entre el olor y la clase del hongo, por eso también muestra que para este conjunto de datos la regla R1 es muy precisa.

3.4. Análisis de relación para determinar si un hongo es comestible

Se realiza un análisis para observar gráficamente la relación que tienen dos atributos para determinar si un hongo es venenoso o comestible.

Primero se presenta un caso que tiene que ver con si un ejemplar es comestible y el olor de un hongo. Para realizar este análisis fue necesario reordenar los atributos entregando niveles a cada columna. A continuación se muestra como se distribuyen los datos.

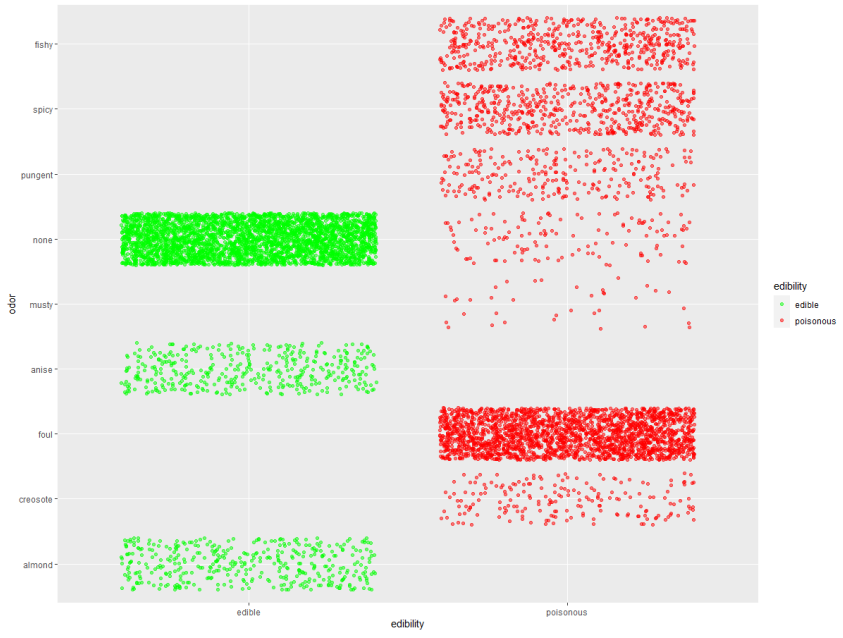

Figura 3: Gráfico de relación entre el olor y si un hongo es comestible

A partir del gráfico de la Figura 3, es posible notar como el olor influye directamente en si un hongo es comestible o venenoso. Ya que los datos se distribuyen marcadamente en cada sector.

Por otro lado, se estudian los atributos de la textura de la copa de un hongo y su color, para ver si estas variables presentan relación con su comestibilidad.

Figura 4: Gráfico de relación entre la textura y color de la copa de un hongo

A partir del gráfico, es posible inferir que los hongos de textura fibrosa en general son comestibles, aunque no se debe confiar en los de color amarillo o gris. Por otra parte, los hongos de textura suave son en su mayoría venenosos, excepto los de color verde y morado.

4. Conclusión

A modo de conclusión, se puede notar que en todo set de datos hay variables que entregan un mejor análisis del estudio y otras que no entregan respuestas claras. A pesar de esto, se necesitan todos los atributos para saber con cuales no contar dependiendo de la investigación que se realiza. En específico, la base de datos de hongos presenta solo dos clases(venenoso o comestible) que se pueden definir a partir de las características del ejemplar, por esto se piensa que su manejo es más accesible y entendible a la hora de trabajar los datos.

A partir del análisis de los atributos es posible detectar que se puede revisar la relación entre todas las combinaciones posibles del estudio. Sin embargo, con las reglas de precisión mencionadas anteriormente, es posible ver las variables que influyen con mayor peso, por esto se busca encontrar la relación entre atributos que aporten una respuesta clara al estudio.

En cuanto al análisis estadístico, se realizó un trabajo exitoso, ya que se pudieron encontrar patrones importantes que sigue la base de datos de hongos. También, destacar que se pudo realizar un estudio más definido en cuanto a las variables y entregar respuestas globales o específicas. Sin embargo, para entrar en conocimiento con el set de datos, se realizó un buen estudio.

Bibliografía

Amat, J. (2016). *Test estadísticos para variables cualitativas: test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran.*

Schlimmer, J. C. (1987). *Concept Acquisition through Representational Adjustment.* PhD thesis, University of California.

Soetewey, A. (2020). *Fisher's exact test in R: independence test for a small sample.*

Wlodzislaw D, Adamczack R, G. K. (1997). *Extraction of crisp logical rules using constructive constrained backpropagation networks.* Proceedings of International Conference on Neural Networks (ICNN'97).