LICENCIATURA EN ESTADÍSTICA

“¿Qué hay debajo del sombrero?”

Trabajo Práctico - Modelos lineales generalizados

Autores: Tomás Anderson - Alejo Vaschetti - Rocio Yael Canteros

Docentes: Gabriela Boggio - Victorio Costa - Guillermina Harvey

16/10/2024

Tabla de contenidos

Introducción 1
Objetivos 1
Análisis descriptivo 1
Selección de variables 4
Analisis de residuos 6
Modelo estimado 6
Razones de odds 7
Predicción y capacidad predictiva 7

Introducción

Se estima que existen entre 2 a 4 millones de hongos en el mundo de los cuales la ciencia ha logrado describir solo una pequeña porción de estos. Los hongos de sombrero a veces pueden ser comestibles y usarse en diversos platos pero otros pueden llegar a ser venenosos. Es esencial poder distinguir entre estos para evitar posibles inconvenientes que pueden ir desde el malestar hasta la muerte.

La estadística nos puede ayudar a determinar cuáles pueden llegar a ser las principales diferencias entre las características de los hongos comestibles y venenosos. Para ello se cuenta con un conjunto de datos de 250 hongos clasificados en 9 variables. Estas son:

- Venenoso o Comestible (Variable Respuesta)
- Diámetro del sombrero (cm)
- Altura del tallo (cm)
- Ancho del tallo (mm)
- Forma del sombrero
- Tipo de lámina
- Color de las láminas
- Color del tallo
- Temporada del año

Objetivos

Se busca poder estimar qué impacto tienen las características de los hongos para su clasificación y cuáles de estas pueden ser señal de cuándo será adecuado el uso de un hongo en la cocina, si es que este no está claro. Para lograr esto, se desea obtener un modelo para predecir si los hongos son comestibles o no.

La metodología considerada para responder a estos propósitos es la construcción de un modelo lineal generalizado.

Análisis descriptivo

Primero, se muestra la distribución marginal de la variable respuesta.

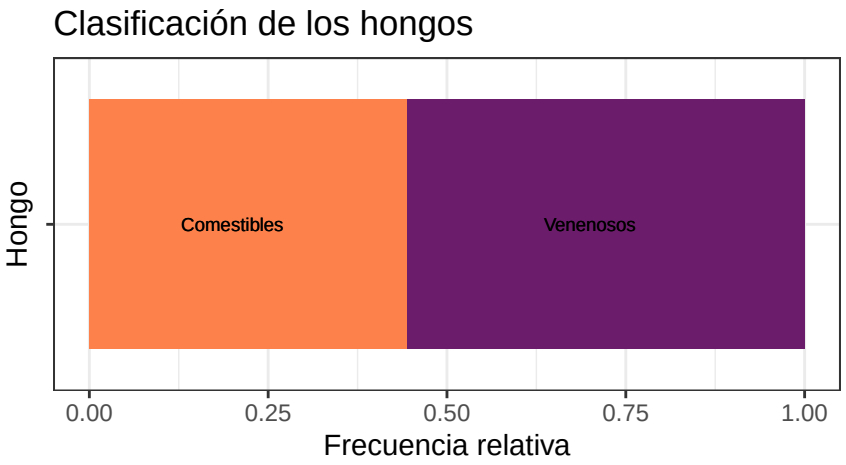

Figura 1: Porcentaje de hongos según el tipo

En la muestra, el 44.4% (111) de los hongos son comestibles mientras que el resto son venenosos (139).

Luego se grafican las variables explicativas según si el hongo es comestible o no.

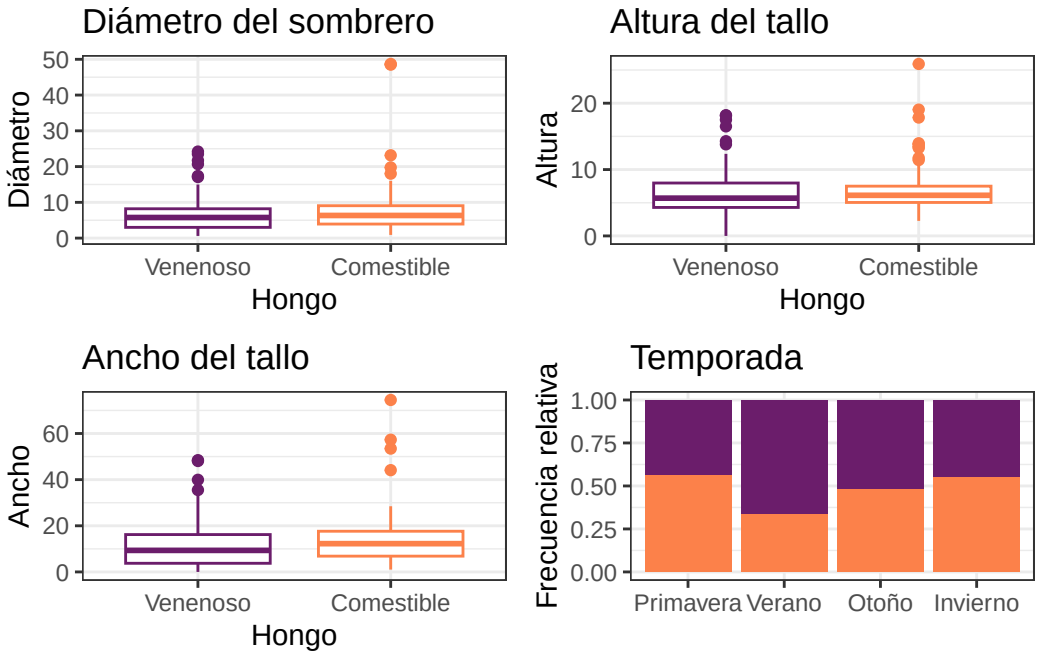

Figura 2: Variables explicativas vs Tipo de hongo

La distribución de la altura y el ancho del tallo junto al diámetro del sombrero parecen ser parecidas para ambos tipos de hongos. Hay algunas observaciones atípicas en dichas variables que podrían afectar negativamente al momento de plantear un modelo. La temporada con menor porcentaje de hongos comestibles es el verano, con un 33.33% (28 de 84).

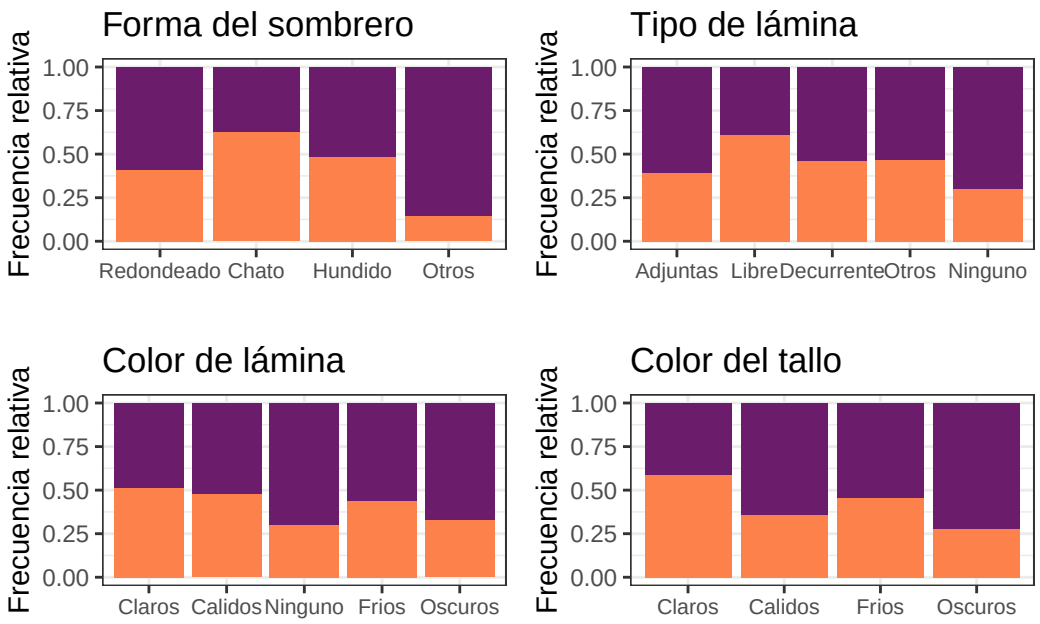

Figura 3: Variables explicativas vs Tipo de hongo

Para la forma del sombrero, la categoría que tiene el mayor porcentaje de hongos comestibles es la *chata* con el 62.5% (35 de 56) y la menor es *otros* con el 14.29% (3 de 21).

Para el tipo de lámina, la categoría que tiene el mayor porcentaje de hongos comestibles es la *libre* con el 60.87% (14 de 23) y la menor es *ninguno* con el 30% (6 de 20).

Para el color de lámina, la categoría que tiene el mayor porcentaje de hongos comestibles es *claros* con el 51.14% (45 de 88) y la menor es *ninguno* con el 30% (6 de 20).

Para el color del tallo, la categoría que tiene el mayor porcentaje de hongos comestibles es *claros* con el 58.77% (67 de 114) y la menor es *oscuros* con el 27.78% (20 de 72).

Selección de variables

Se tienen muchas variables que pueden llegar a considerarse para el análisis, pero se busca aquellas que aporten en la discriminación de los hongos al construir un modelo. Para ello se observa qué variable en forma marginal produce un cambio significativo en la explicación de sí un hongo es comestible o no.

Variables	**gl**	D	**p-value**
Diámetro del sombrero	1	3.277	0.07
Altura del tallo	1	2.158	0.142
Ancho del tallo	1	1.686	0.194
Forma del sombrero	3	17.037	0.001
Tipo de lámina	4	5.258	0.262
Color de las láminas	4	6.328	0.176
Color del tallo	4	19.511	<0.001
Temporada	3	7.14	0.068
Nivel de significación: 0.10			

Tabla 1: Modelos marginales

Las variables significativas de forma marginal son el diámetro del sombrero, la forma del sombrero, el color del tallo y la temporada del año, por lo que se plantea un modelo con todas ellas incluidas y se quitan todas las que no aporten significativamente a la explicación.

Modelo	**gl**	D	**p-value**
Sin diámetro del sombrero	1	4.861	0.027
Sin forma del sombrero	3	13.228	0.004
Sin color de las láminas	4	17.104	0.001
Sin temporada	3	7.645	0.055
Nivel de significación: 0.05			

Tabla 2: Modelos con las variables marginales

Se elimina la temporada del modelo y se prueba incluir las demás variables explicativas que no se incluyeron al principio.

Modelo	**gl**	D	**p-value**
Añadir tipo de lámina	4	3.621	0.46
Añadir color de las láminas	4	4.331	0.363
Añadir altura del tallo	1	0.069	0.793
Añadir ancho del tallo	1	0.002	0.966
Nivel de significación: 0.05			

Tabla 3: Modelos con las variables marginales no significativas

No se incorpora ninguna variable nueva, por lo que se pasa a probar las interacciones.

Modelo	**gl**	D	**p-value**
Diámetro del sombrero x Forma del sombrero	3	7.985	0.066
Diámetro del sombrero x Color de las láminas	8	9.49	0.303
Forma del sombrero x Color de las láminas	4	7.568	0.056
Nivel de significación: 0.05			

Tabla 4: Modelos con las interacciones

Ninguna interacción es significativa por lo que no se incluyen en el modelo.

Una vez elegidas las variables, se prueba si se debe añadir el Diámetro del sombrero del hongo de forma lineal o como un factor. Se divide la variable en cuatro grupos de tamaño parecido y se crean dos variables con ellos, una ordinal y la otra categórica. Se definen dos nuevos modelos con cada una de estas nuevas variables (sin incluir la forma original) y se comparan las deviance de los modelos. No se rechaza (**p-value:0.20**) la hipótesis nula de no linealidad, por lo que se mantiene en el modelo la variable en forma lineal y continua.

Además, para verificar la bondad del ajuste del modelo se utiliza el test de Hosmer-Lemeshow con 10 grupos, resultando en un no rechazo (**p-value:0.93**) de la hipótesis nula de bondad del ajuste.

Analisis de residuos

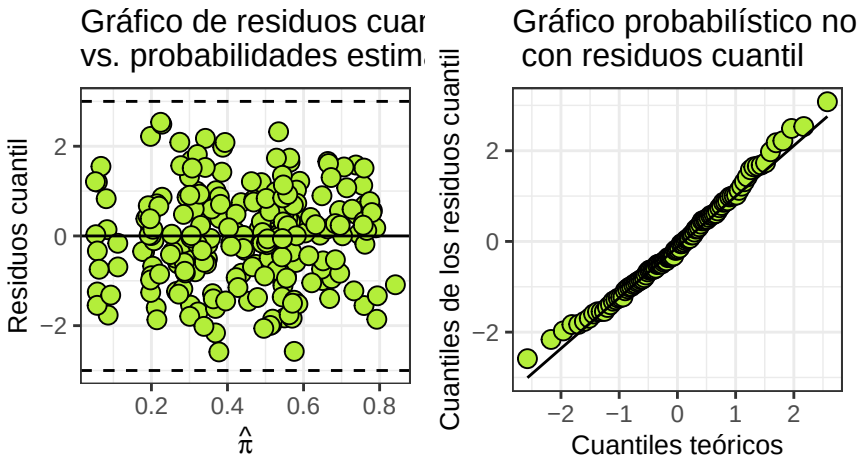

Figura 4: Analisis de residuos del enlace logit

Parecería no haber problemas con los residuos del modelo al usar la función de enlace logit. Para verificar su adecuación se agrega al modelo el predictor lineal estimado elevado al cuadrado($\hat{\eta}^2$) como covariable y se prueba si es significativa o no. Se ajusta el nuevo modelo y se llega a la conclusión de que es adecuado el uso de dicho enlace (**p-value:0.41**).

Modelo estimado

Luego de seleccionar las variables a incluir y la función de enlace, se llega al siguiente modelo:

$$logit(\hat{\pi}_i) = -0.20 + 0.06D_i + 0.80C_i + 0.07H_i - 1.49O_i - 0.85C_i - 0.59F_i - 1.28Osc_i$$

Donde $logit(\pi_i) = \eta_i$

D_i: Diámetro del sombrero del i-ésimo hongo.

C_i: Variable dummy asociada a si el i-ésimo hongo tiene un sombrero chato.

H_i: Variable dummy asociada a si el i-ésimo hongo tiene un sombrero hundido.

O_i: Variable dummy asociada a si el i-ésimo hongo tiene un sombrero de otro tipo.

C_i: Variable dummy asociada a si el i-ésimo hongo tiene un tallo de color cálido.

F_i: Variable dummy asociada a si el i-ésimo hongo tiene un tallo de color frío.

Osc_i: Variable dummy asociada a si el i-ésimo hongo tiene un tallo de color oscuro.

La categoría de referencia del modelo es un hongo con un tallo de color claro y una forma redondeada del sombrero.

Razones de odds

Para dar una idea de como cambian las probabilidades, se calculan las razones de odds para cambios de las categorías de las variables.

Al aumentar en 5cm el diámetro del sombrero del hongo, la chance de que el hongo sea comestible o no aumenta en un 36.36%, al mantener fijas las demás variables del modelo.

La chance de que los hongos con sombreros chatos sean comestibles o no es un 121.78% más grande que esa misma chance para los hongos con forma redondeada del sombrero, al mantener fijas las demás variables del modelo.

La chance de que los hongos con tallo de color oscuro sean comestibles o no es un 72.19% más chica que esa misma chance para los hongos con tallo de color claro, al mantener fijas las demás variables del modelo.

Predicción y capacidad predictiva

Pese a que la probabilidad asociada a la variable *temporada* fue mayor que el nivel de signficación usado para determinar la inclusión de una variable al modelo, algunos expertos aseguran que es importante tener en cuenta la estación del año en la que se desarrolla el hongo a la hora de clasificarlo.

Por ello, se ajusta un modelo que incluya esa variable y se compara su capacidad predictiva con la del modelo que no la incluye a través de las gráficas de las curvas ROC y las predicciones obtenidas con el punto de corte óptimo.

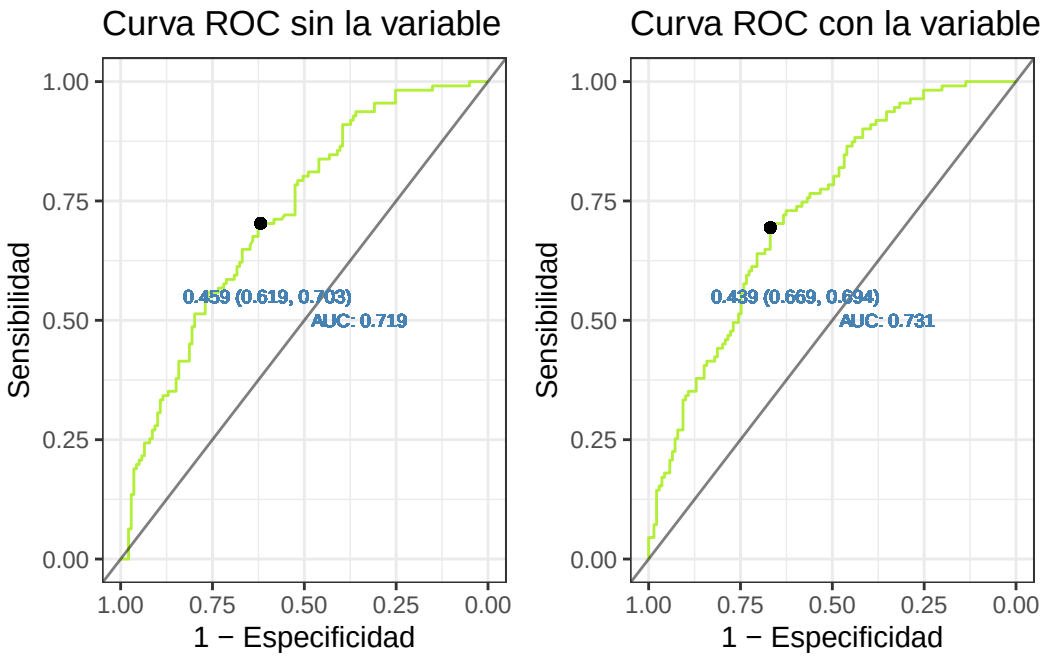

Figura 5: Comparación de curvas ROC

Se tienen entonces las siguientes Métricas:

Modelo	Corte	AUC	Presición Global	Especificidad	Sensibilidad
Sin "Temporada"	0.459	0.719	0.656	0.619	0.703
Con "Temporada"	0.439	0.731	0.668	0.712	0.613

Tabla 5: Métricas de Evaluación para ambos modelos

Bajo el punto de corte del modelo sin la variable (0.459), la matriz de confusión es la siguiente:

	Observado	
Predicho	Venenosos	Comestibles
Venenosos	86	33
Comestibles	53	78

Tabla 6: Matriz de confusión sin temporada

Mientras que bajo el punto de corte del modelo con la variable (0.439), la matriz de confusión es la siguiente:

	Observado	
Predicho	Venenosos	Comestibles
Venenosos	99	43
Comestibles	40	68

Tabla 7: Matriz de confusión con temporada

Como el objetivo del trabajo es encontrar un modelo que encuentre que hongos no son los venenosos, se elige el modelo con la variable temporada incluida, ya que tiene una precisión global y una especificidad mayor.