ITBA
Instituto Tecnológico de Buenos Aires

Trabajo Final

82.05 - Análisis Predictivo

Ian Dalton - 62345

Caso de negocio

Introducción y caso de negocio

Predecir si un hongo es **comestible** o no.

Un hongo puede ser tanto:

- ➔ **Venenoso**
  - Muerte
  - Alucinogenos
  - Dolor de panza
- ➔ **Comestible**

Modelo de Predicción

Objetivo: Predecir si un hongo es **comestible** a partir de las características brindadas por el usuario

Unidad observacional: Características del hongo

Variable Target: “class”

Variables predictoras:

'cap-shape', 'cap-surface', 'cap-color', 'bruises', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'stalk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-print-color', 'population', 'habitat'

Modelo: Clasificación.

Dataset

Información del Dataset

Contiene información de las **características** de hongos, que pueden ser **dimensiones** o **descripciones**.

Cuenta con **8.124** registros y **22** variables.

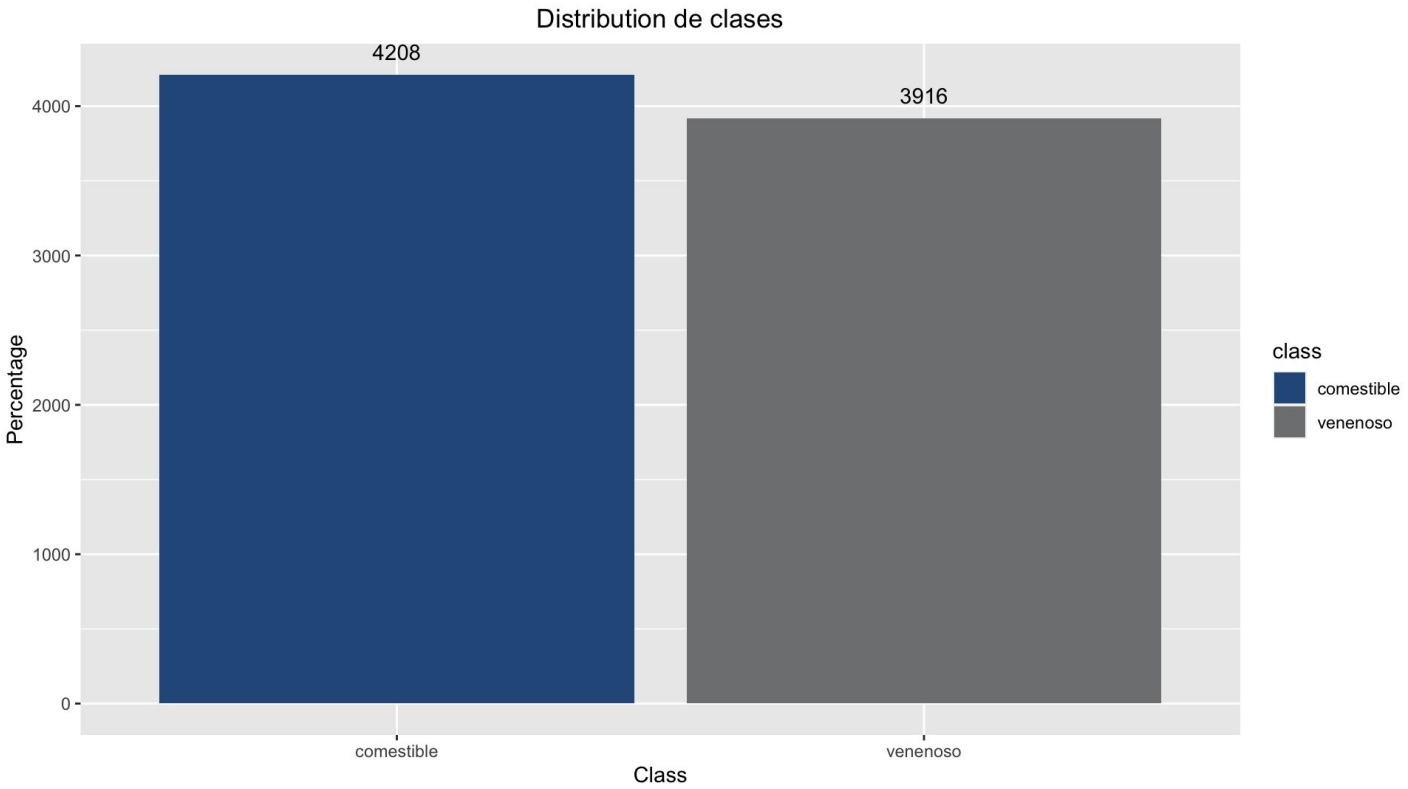
Limpieza de la base

Todas las columnas son categoricas

Frecuencia de actualización: No hay

Análisis exploratorio

Tipos de hongos

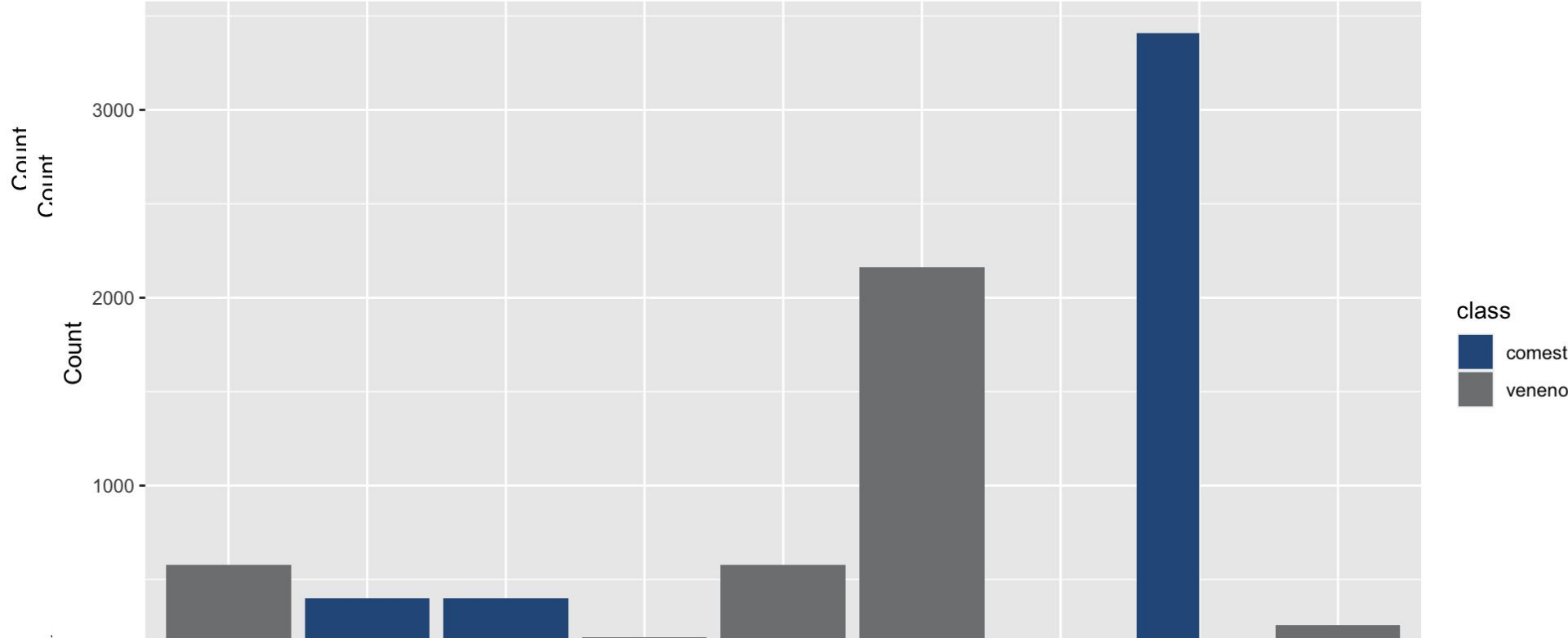


Cantidad de cap.shape

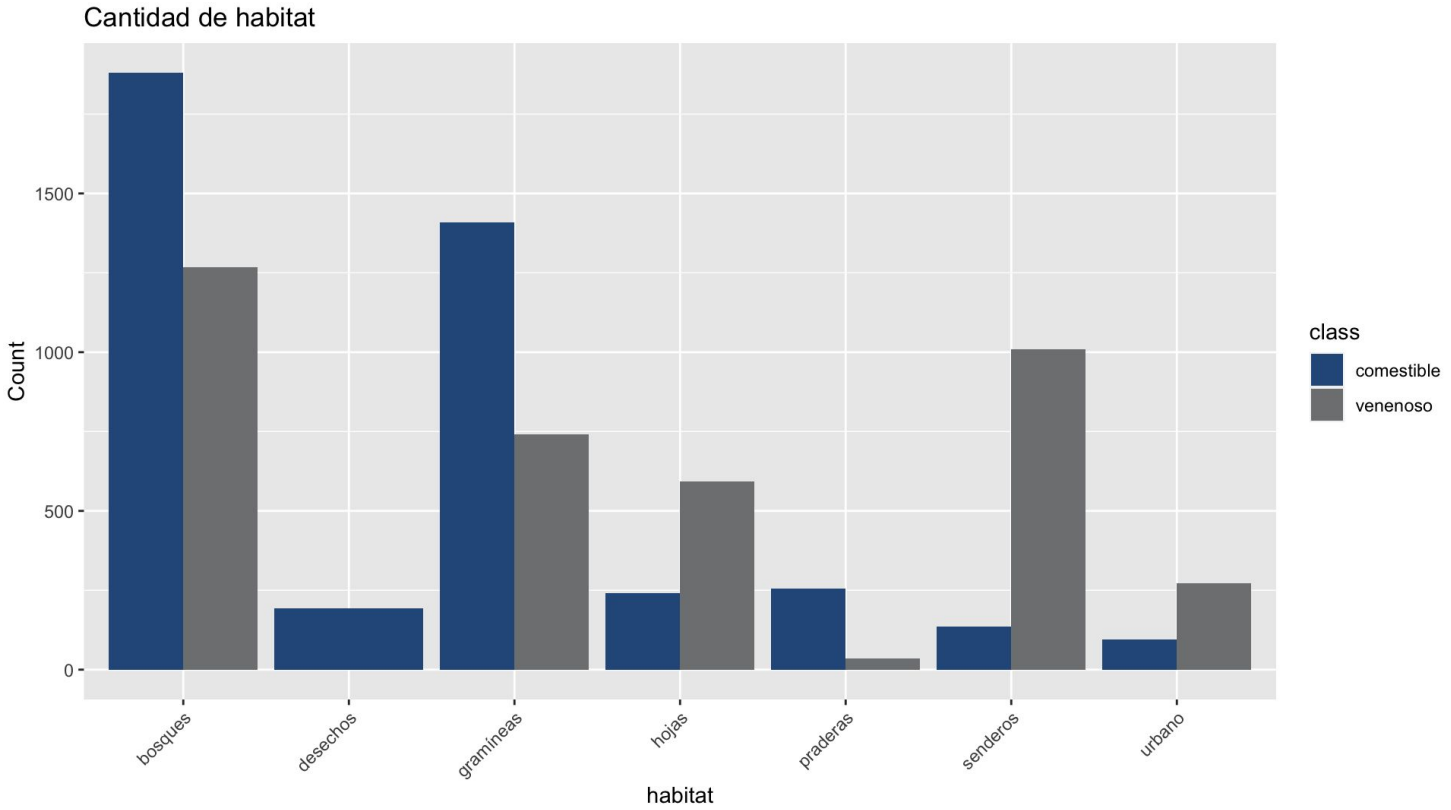
Cantidad de cap.surface

Cantidad de cap.color

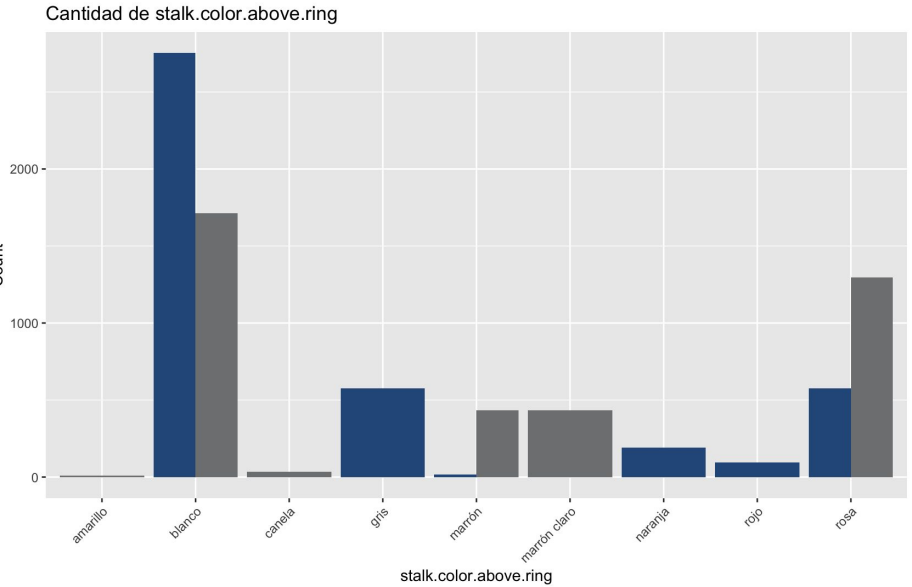
Cantidad de odor

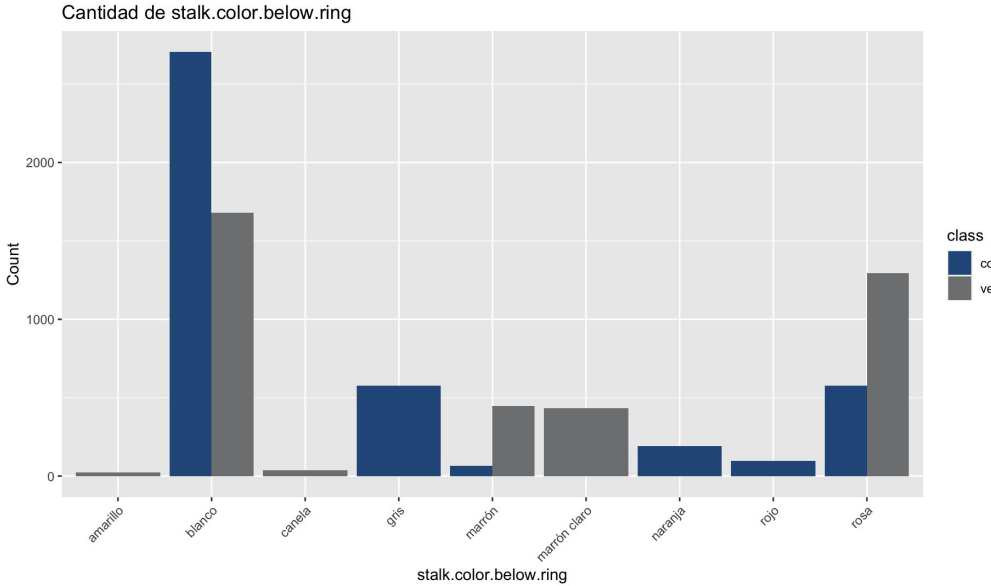


Factores geograficos

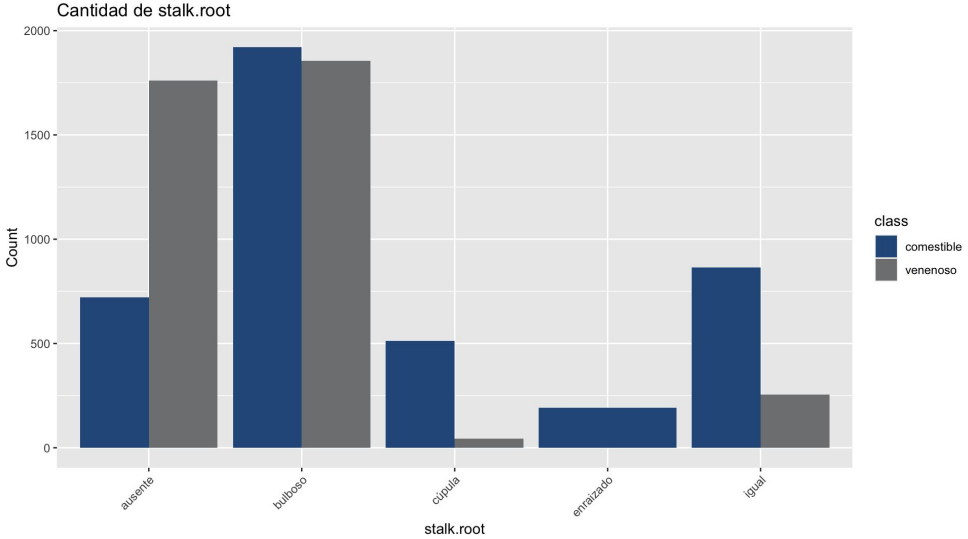


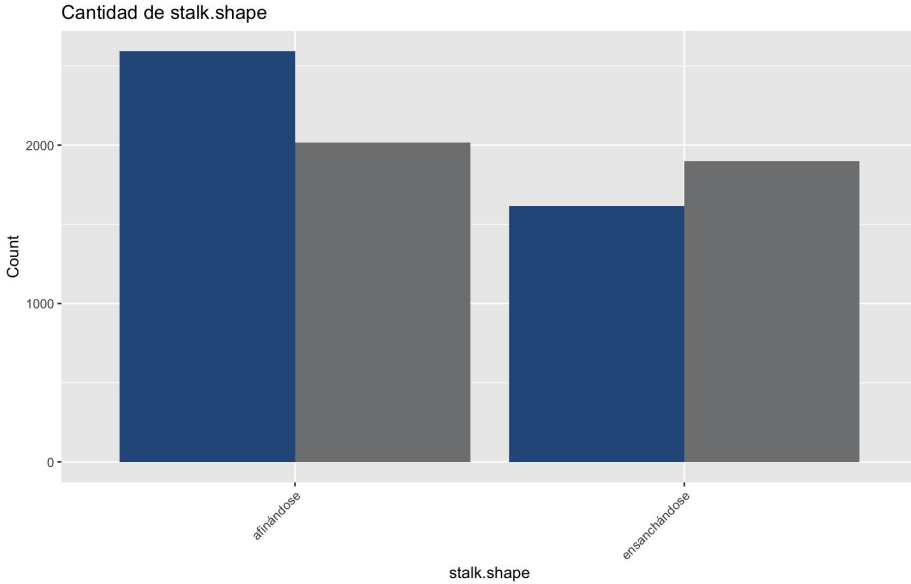
Colores del tallo



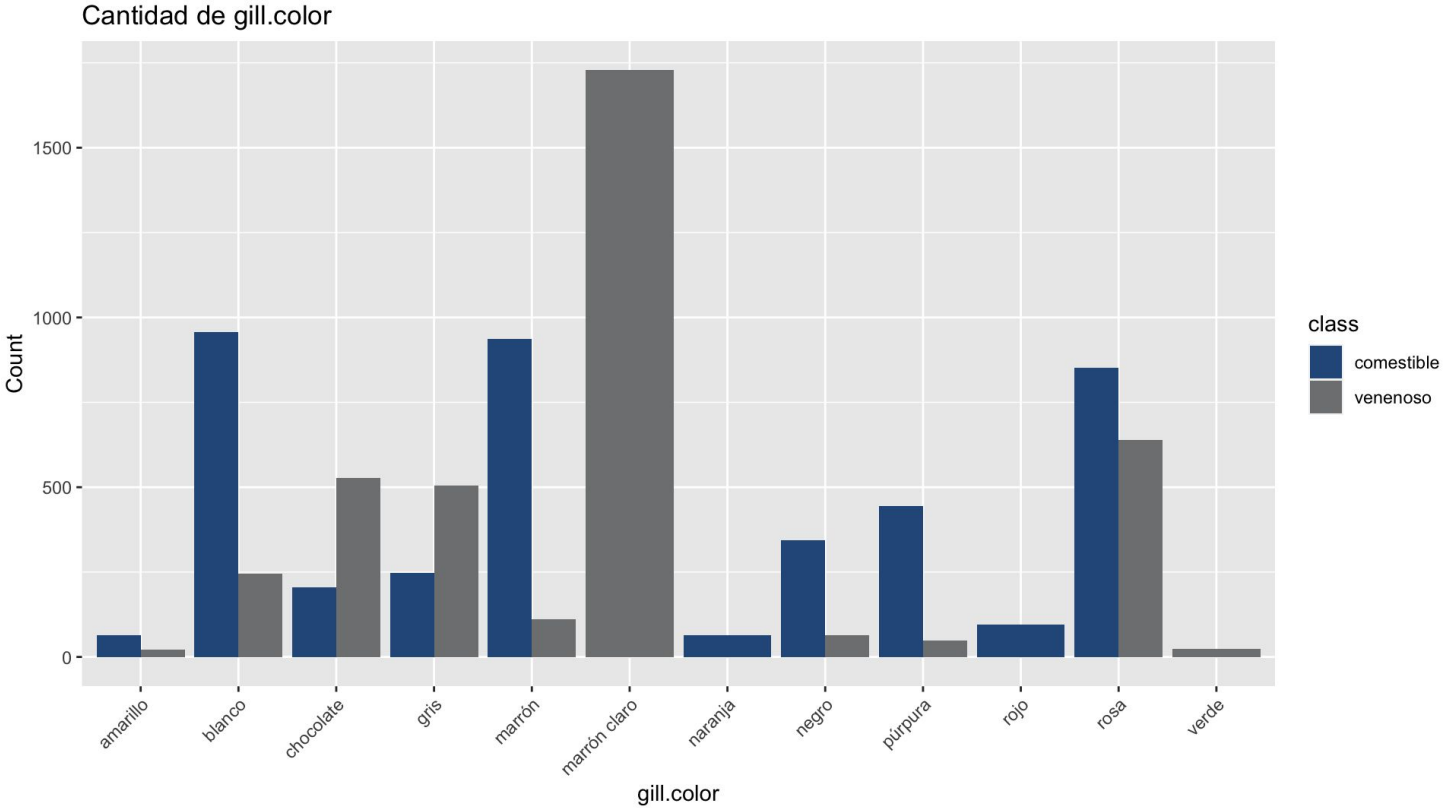


Detalles del tallo

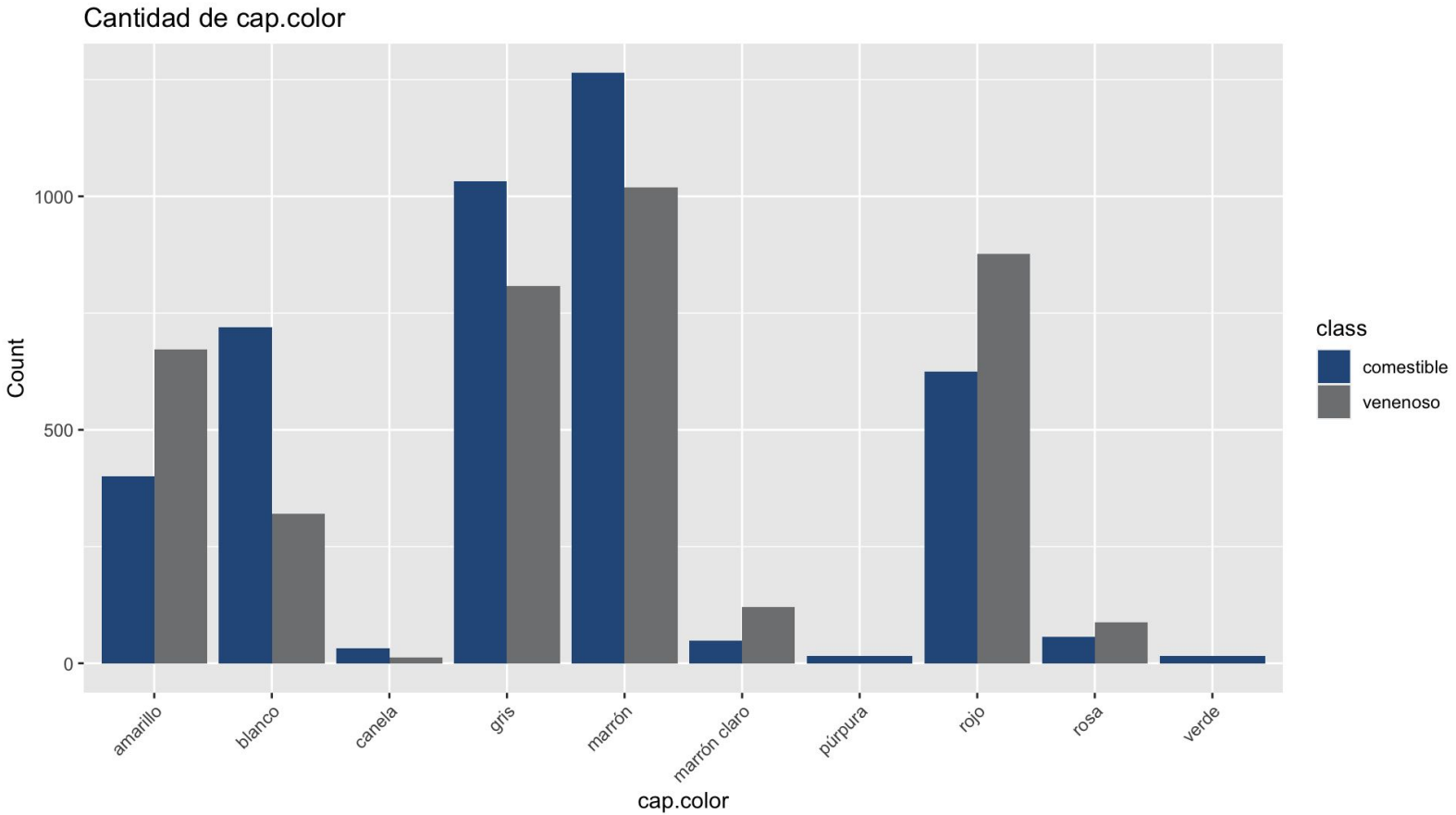




Colores


Cantidad de gill.color
Count
1500
1000
500
0
amarillo
blanco
chocolate
gris
marrón
marrón claro
naranja
negro
púrpura
rojo
rosa
verde
gill.color
class
comestible
venenoso

Colores

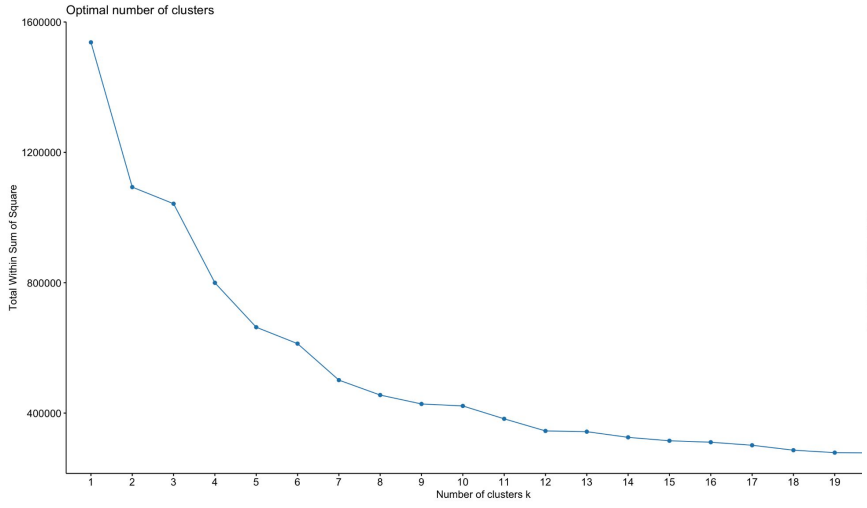

Cantidad de cap.color
Count
1000
500
0
amarillo
blanco
canela
gris
marrón
marrón claro
púrpura
rojo
rosa
verde
cap.color
class
comestible
venenoso

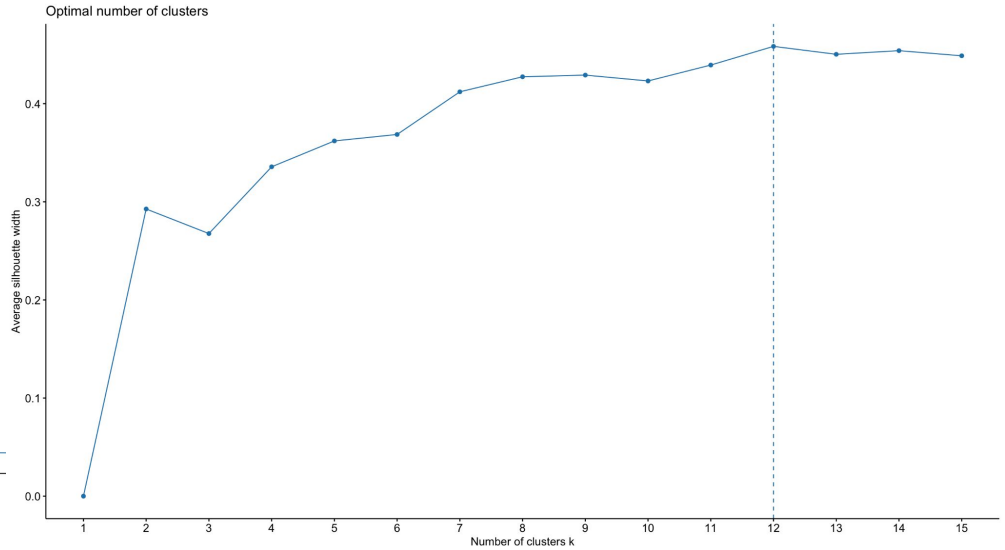
Clustering

Se realizó con el objetivo de comparar y evaluar diferencias entre las diferentes características que puedan ayudar a identificar si el hongo es comestible o no.

Medimos el valor del **estadístico de Hopkins** para evaluar si es conveniente la clusterización, obteniendo un valor cercano al **81,4%**.

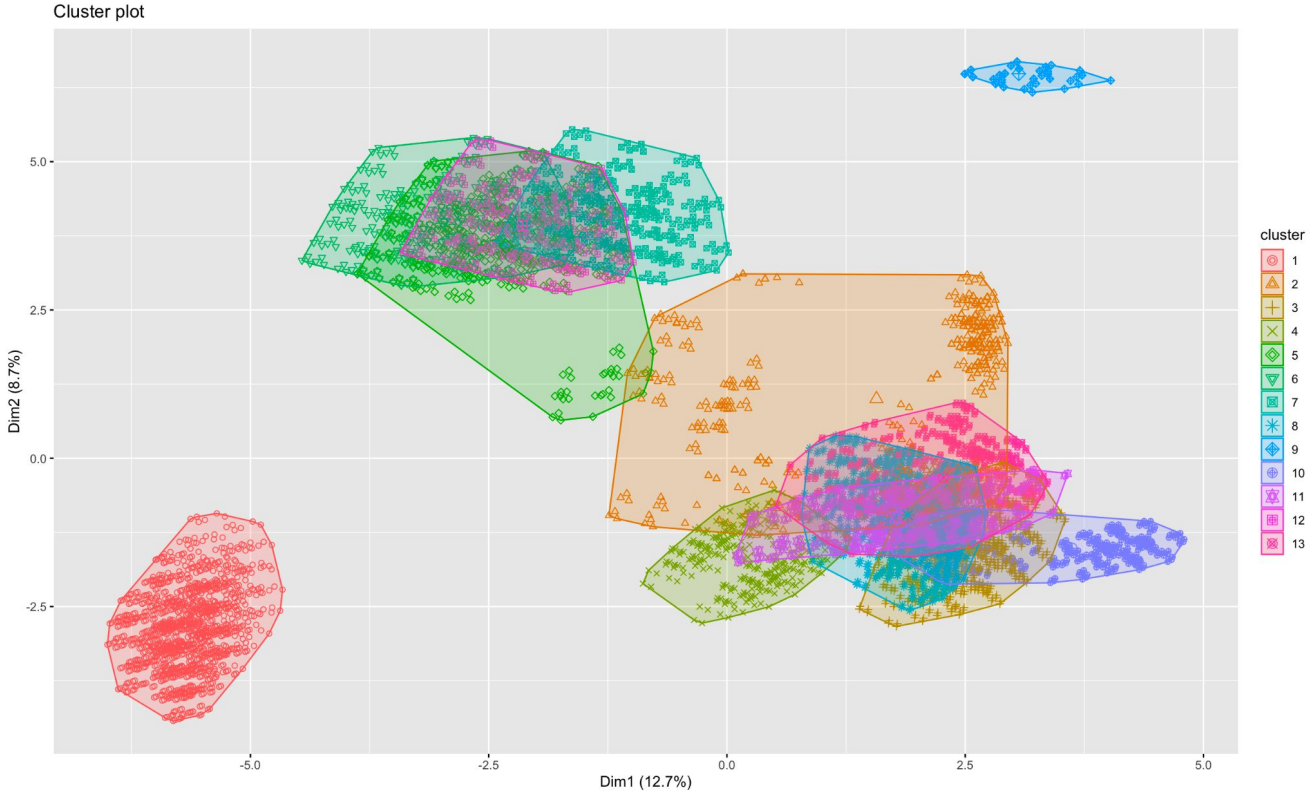
Clustering - Agrupaciones





Número óptimo de clusters: 13

Visualización de Clusters



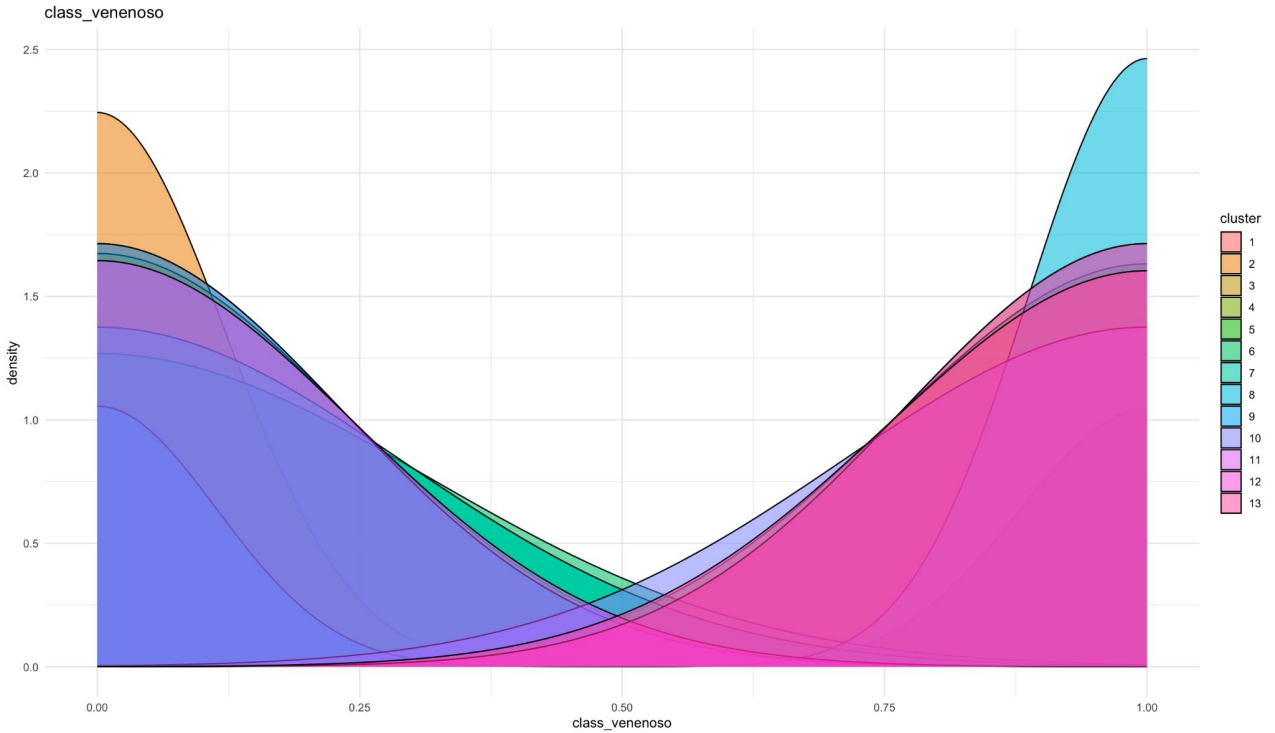
Análisis de clusters

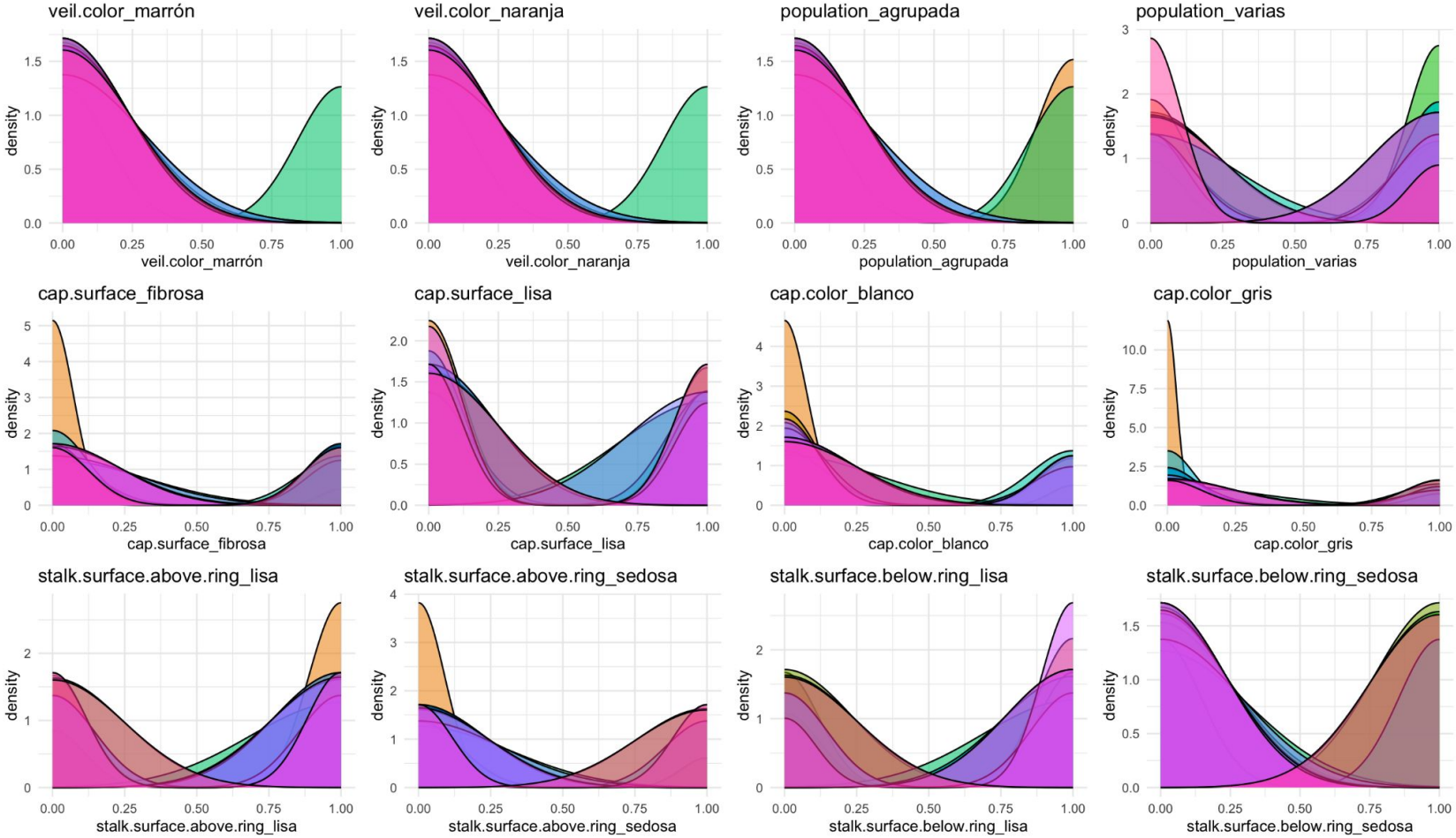
Cluster 2:

- Comestibles
- Se encuentran agrupados
- Tallo liso

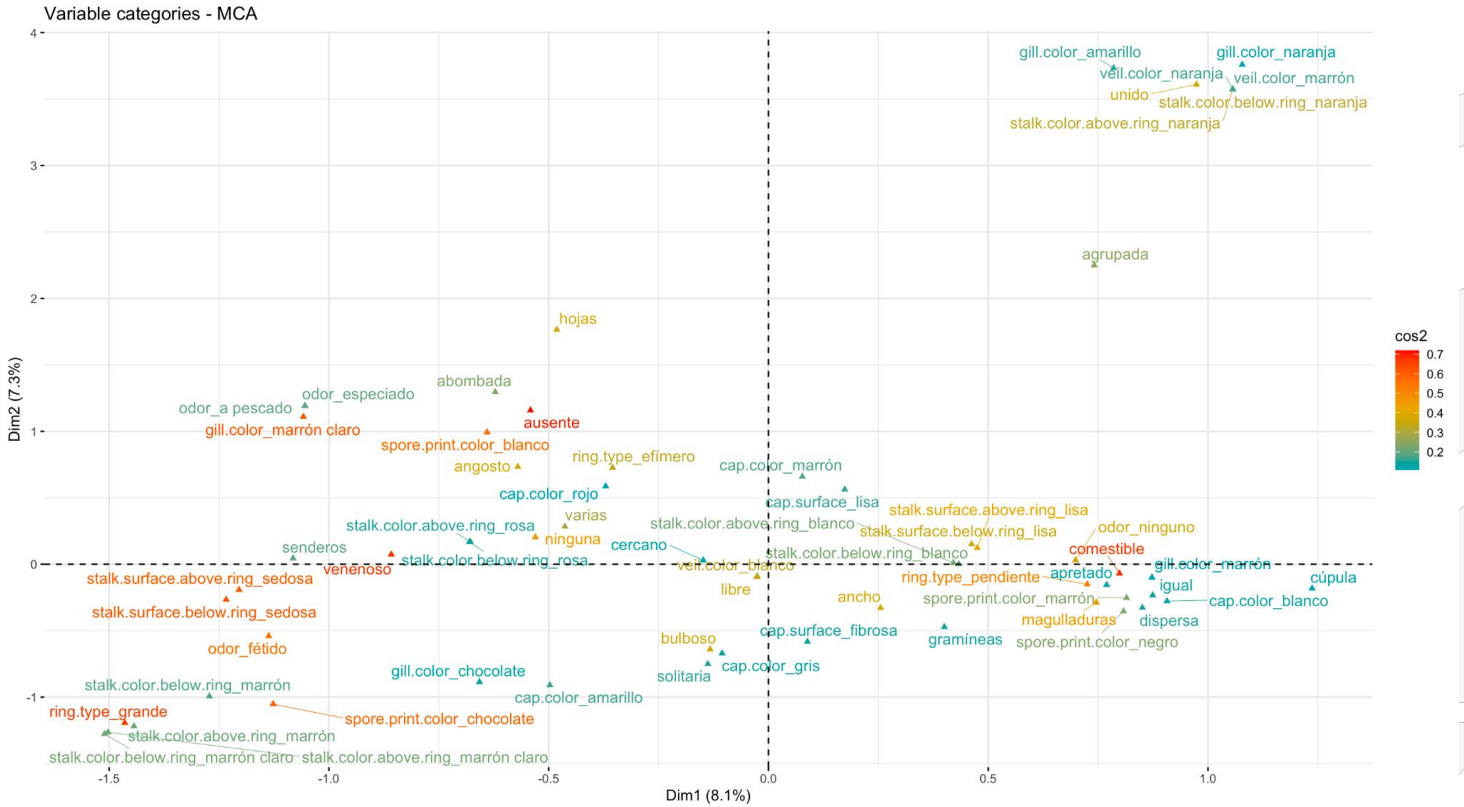
Cluster 8:

- Venenosos
- Cabeza lisa
- No se encuentra en grupos




veil.color_marrón
density
veil.color_marrón
veil.color_naranja
density
veil.color_naranja
population_agrupada
density
population_agrupada
population_varias
density
population_varias
cap.surface_fibrosa
density
cap.surface_fibrosa
cap.surface_lisa
density
cap.surface_lisa
cap.color_blanco
density
cap.color_blanco
cap.color_gris
density
cap.color_gris
stalk.surface.above.ring_lisa
density
stalk.surface.above.ring_lisa
stalk.surface.above.ring_sedosa
density
stalk.surface.above.ring_sedosa
stalk.surface.below.ring_lisa
density
stalk.surface.below.ring_lisa
stalk.surface.below.ring_sedosa
density
stalk.surface.below.ring_sedosa

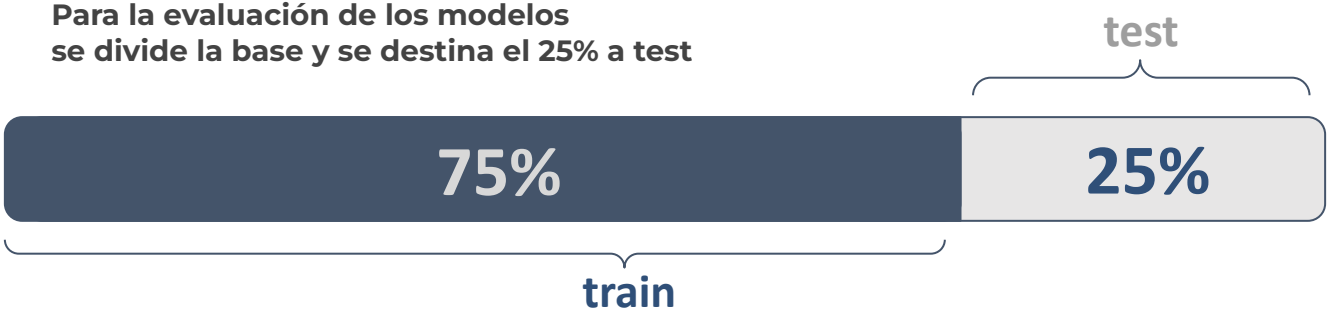
MCA


Variable categories - MCA
Dim2 (7.3%)
Dim1 (8.1%)
cos2
0.7
0.6
0.5
0.4
0.3
0.2
gill.color_amarillo
gill.color_naranja
veil.color_naranja
veil.color_marrón
unido
stalk.color.below.ring_naranja
stalk.color.above.ring_naranja
agrupada
hojas
abombada
odor_especiado
odor_a pescado
ausente
gill.color_marrón claro
spore.print.color_blanco
angosto
ring.type_efímero
cap.color_marrón
cap.color_rojo
cap.surface_lisa
stalk.surface.above.ring_lisa
varias
stalk.color.above.ring_rosa
ninguna
stalk.color.above.ring_blanco
stalk.surface.below.ring_lisa
odor_ninguno
senderos
cercano
comestible
stalk.color.below.ring_rosa
veil.color_blanco
stalk.color.below.ring_blanco
gill.color_marrón
venenoso
stalk.surface.above.ring_sedosa
ring.type_pendiente
apretado
cúpula
libre
igual
ancho
spore.print.color_marrón
cap.color_blanco
stalk.surface.below.ring_sedosa
magulladuras
dispersa
cap.surface_fibrosa
bulboso
gramíneas
spore.print.color_negro
odor_fétido
cap.color_gris
gill.color_chocolate
solitaria
stalk.color.below.ring_marrón
cap.color_amarillo
ring.type_grande
spore.print.color_chocolate
stalk.color.above.ring_marrón
stalk.color.below.ring_marrón claro
stalk.color.above.ring_marrón claro

Evaluación

Partición de datos

Para la evaluación de los modelos
se divide la base y se destina el 25% a test



Modelo Baseline

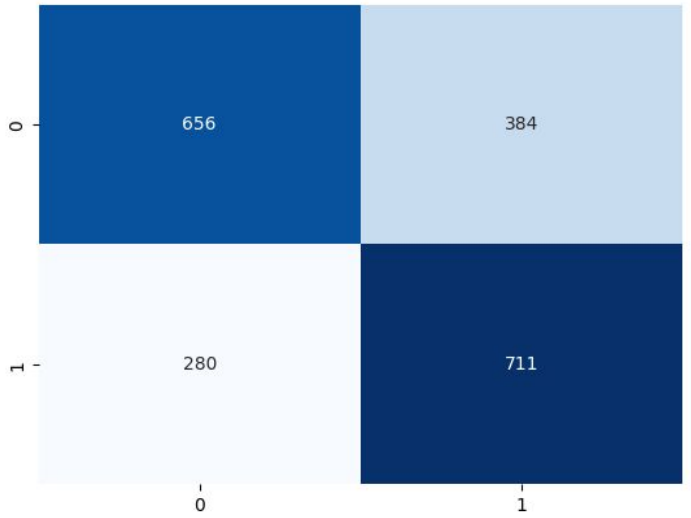
¿Qué modelo use como baseline?

Features:

- Colores
- Forma de la cabeza

Modelo utilizado: LogisticRegression

Rendimiento alcanzado: 0.673



Selección de modelos

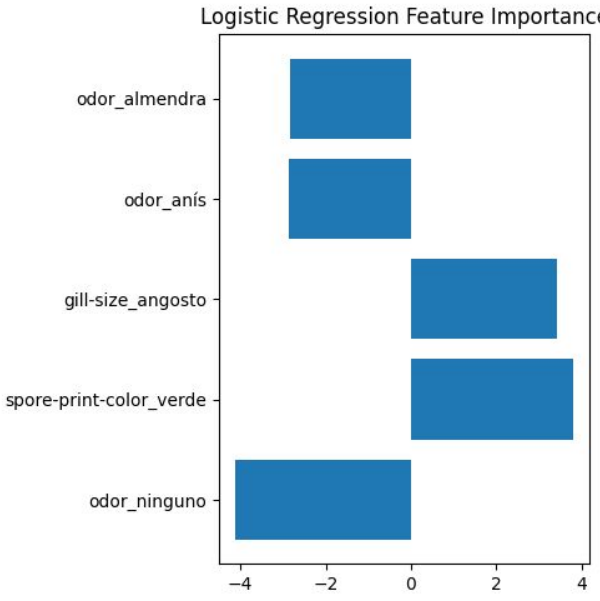
Modelos evaluados

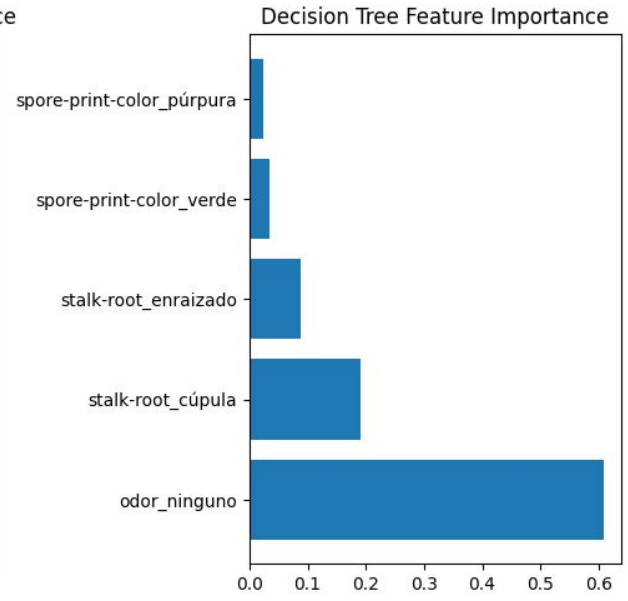
- Arboles
  - XGBoost
  - Decision Tree Classifier
  - Random Forest
- Support Vector Classifier
- KNN
- Logistic Regression

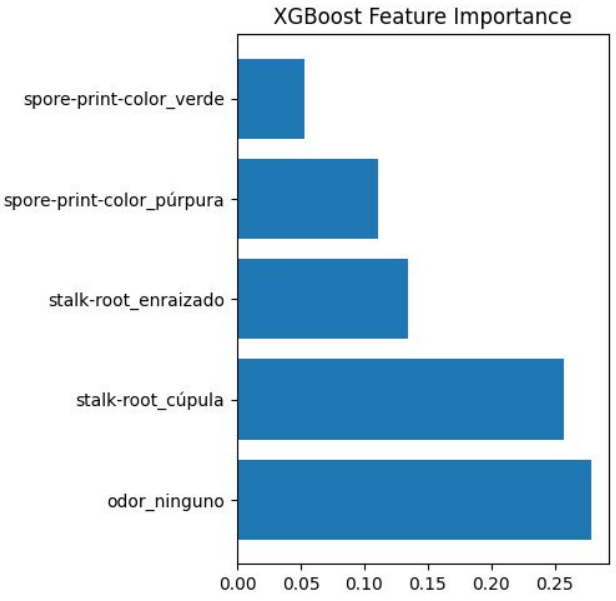
Tratamiento de variables categóricas

- One hot encoding
  - Se eliminó la primer columna
- Flags de diferencia en colores

Feature importance







Modelo Final

Modelo Final

XGBoost

Score: 1.0

Hiper Parámetros

- ➔ Booster Type: 'gbtree'
- ➔ Learning Rate: 0.3
- ➔ n_estimators: 100
- ➔ max_depth: 6
- ➔ min_child_weight: 1

Limitaciones y posibles mejoras

Limitaciones

- ➔ Inexperiencia en modelos predictivos
- ➔ Dataset limitado

Posibles mejoras

- ➔ Aumentar el tamaño del dataset para encontrar edge cases

Conclusiones

Conclusiones

Objetivo: Predecir si un hongo es **comestible** o no.

➔ EDA

- Es muy fácil confundir hongos ya que **comparten** muchas **características**.
- Los comestibles **no** suelen tener olores.
- La forma que tienen los hongos de **producir esporas impacta** bastante en si es comestible o no.

Conclusiones

- **Modelo predictivo**
  - **Herramientas utilizadas**
    - OneHotEncoder
    - Flags en variables importantes
  - **Modelo final:**
    - El **olor** es una variable relevante.

- **Metricas evaluadas**
  - Score
  - Matriz de confusión

ITBA
Instituto Tecnológicc
de Buenos Aires

Gracias!