Estudio sobre los hongos: Modelo predictivo para determinar su consumo

Proyecto Final: Data Science

coderhouse

Alumna: Kerlismar Sarays Guarguana

Septiembre 2023

Tabla de Contenido

1.- Descripción del caso de Negocio y Objetivos del modelo

En los últimos años ha crecido exponencialmente un boom en referente a todo lo natural, desde productos de belleza, medicinales, alimenticios, recreacionales, entre otros tantos. Uno de los componentes estrella que destaca dentro de todas las categorías antes mencionadas son los hongos, los cuales al día de hoy y después de más de 40 años de investigación siguen siendo objeto de estudio, por lo complejo, diverso y enigmático que resulta todo sobre ellos.

En este sentido, el caso de estudio es analizar las características de una base de datos sobre características físicas de los hongos y determinar en basea estás cuales son aptas para el consumo humano y cuales no (venenosos).

2.- Tabla de versionado

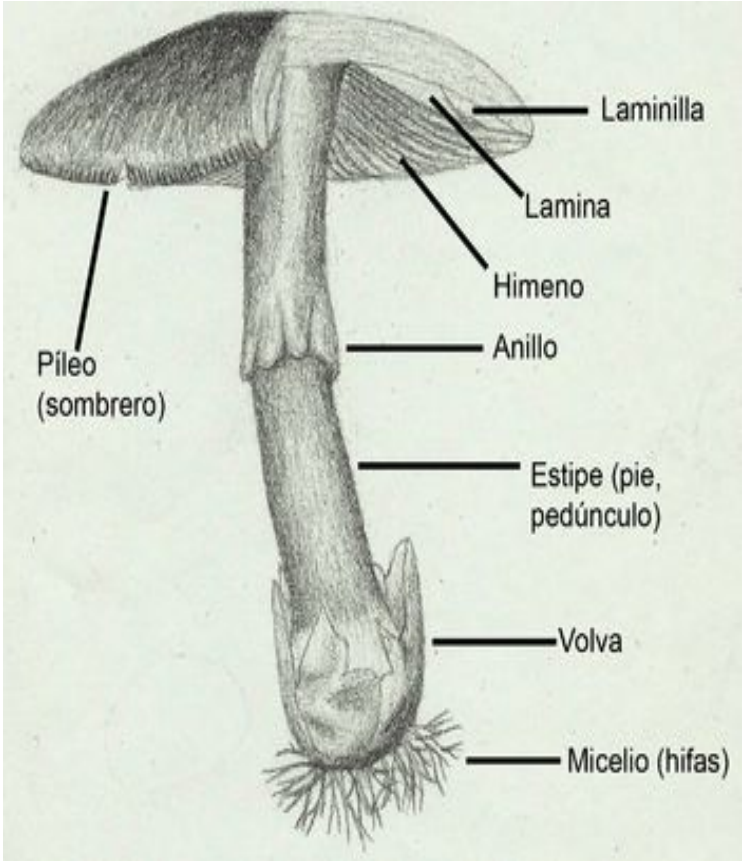
Solo existe una versión

3.- Descripción de los datos

El dataset es un conjunto de datos creado por la UCI (University of California) para el departamento de Machine Learning. Es un dataset público y se encuentra disponible en Kaggle, se puede acceder mediante el siguiente link.

Dentro de las variables que componen el dataset encontramos:

- *class*: clase
- *cap-shape*: forma del sombrero
- *cap-surface*: superficie del sombrero
- *cap-color*: color del sombrero
- *bruises*: manchas
- *odor*: olor
- *gill-attachment*: himenio
- *gill-spacing*: láminas
- *gill-size*: tamaño de las láminas
- *gill-color*: color de las láminas
- *stalk-shape*: pie
- *stalk-root*: raíz
- *stalk-surface-above-ring*:superficie del anillo
- *stalk-surface-below-ring*:superficie debajo del anillo
- *stalk-color-above-ring*:color superficie bajo del anillo
- *stalk-color-below-ring*:color superficie del anillo
- *veil-type*: velo
- *veil-color*: color del velo
- *ring-number*: número de anillos
- *ring-type*: tipo de anillo
- *spore-print-color*: color de espora
- *population*: comunidad alrededor del hongo
- *habitat*: habitat



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   class                     8124 non-null   object
 1   cap-shape                 8124 non-null   object
 2   cap-surface               8124 non-null   object
 3   cap-color                 8124 non-null   object
 4   bruises                   8124 non-null   object
 5   odor                      8124 non-null   object
 6   gill-attachment           8124 non-null   object
 7   gill-spacing              8124 non-null   object
 8   gill-size                 8124 non-null   object
 9   gill-color                8124 non-null   object
 10  stalk-shape               8124 non-null   object
 11  stalk-root                8124 non-null   object
 12  stalk-surface-above-ring  8124 non-null   object
 13  stalk-surface-below-ring  8124 non-null   object
 14  stalk-color-above-ring    8124 non-null   object
 15  stalk-color-below-ring    8124 non-null   object
 16  veil-type                 8124 non-null   object
 17  veil-color                8124 non-null   object
 18  ring-number               8124 non-null   object
 19  ring-type                 8124 non-null   object
 20  spore-print-color         8124 non-null   object
 21  population                8124 non-null   object
 22  habitat                   8124 non-null   object
dtypes: object(23)
memory usage: 1.4+ MB
```

El dataset cuenta con un total de 8124 entradas/registros y 23 variables/columnas, las cuales componen las distintas partes de un hongo, siendo las mismas de tipo categóricas.

```
NAME: class LENGTH: 2 VALUES: ['p' 'e']
NAME: cap-shape LENGTH: 6 VALUES: ['x' 'b' 's' 'f' 'k' 'c']
NAME: cap-surface LENGTH: 4 VALUES: ['s' 'y' 'f' 'g']
NAME: cap-color LENGTH: 10 VALUES: ['n' 'y' 'w' 'g' 'e' 'p' 'b' 'u' 'c' 'r']
NAME: bruises LENGTH: 2 VALUES: ['t' 'f']
NAME: odor LENGTH: 9 VALUES: ['p' 'a' 'l' 'n' 'f' 'c' 'y' 's' 'm']
NAME: gill-attachment LENGTH: 2 VALUES: ['f' 'a']
NAME: gill-spacing LENGTH: 2 VALUES: ['c' 'w']
NAME: gill-size LENGTH: 2 VALUES: ['n' 'b']
NAME: gill-color LENGTH: 12 VALUES: ['k' 'n' 'g' 'p' 'w' 'h' 'u' 'e' 'b' 'r' 'y' 'o']
NAME: stalk-shape LENGTH: 2 VALUES: ['e' 't']
NAME: stalk-root LENGTH: 5 VALUES: ['e' 'c' 'b' 'r' '?']
NAME: stalk-surface-above-ring LENGTH: 4 VALUES: ['s' 'f' 'k' 'y']
NAME: stalk-surface-below-ring LENGTH: 4 VALUES: ['s' 'f' 'y' 'k']
NAME: stalk-color-above-ring LENGTH: 9 VALUES: ['w' 'g' 'p' 'n' 'b' 'e' 'o' 'c' 'y']
NAME: stalk-color-below-ring LENGTH: 9 VALUES: ['w' 'p' 'g' 'b' 'n' 'e' 'y' 'o' 'c']
NAME: veil-type LENGTH: 1 VALUES: ['p']
NAME: veil-color LENGTH: 4 VALUES: ['w' 'n' 'o' 'y']
NAME: ring-number LENGTH: 3 VALUES: ['o' 't' 'n']
NAME: ring-type LENGTH: 5 VALUES: ['p' 'e' 'l' 'f' 'n']
NAME: spore-print-color LENGTH: 9 VALUES: ['k' 'n' 'u' 'h' 'w' 'r' 'o' 'y' 'b']
NAME: population LENGTH: 6 VALUES: ['s' 'n' 'a' 'v' 'y' 'c']
NAME: habitat LENGTH: 7 VALUES: ['u' 'g' 'm' 'd' 'p' 'w' 'l']
```

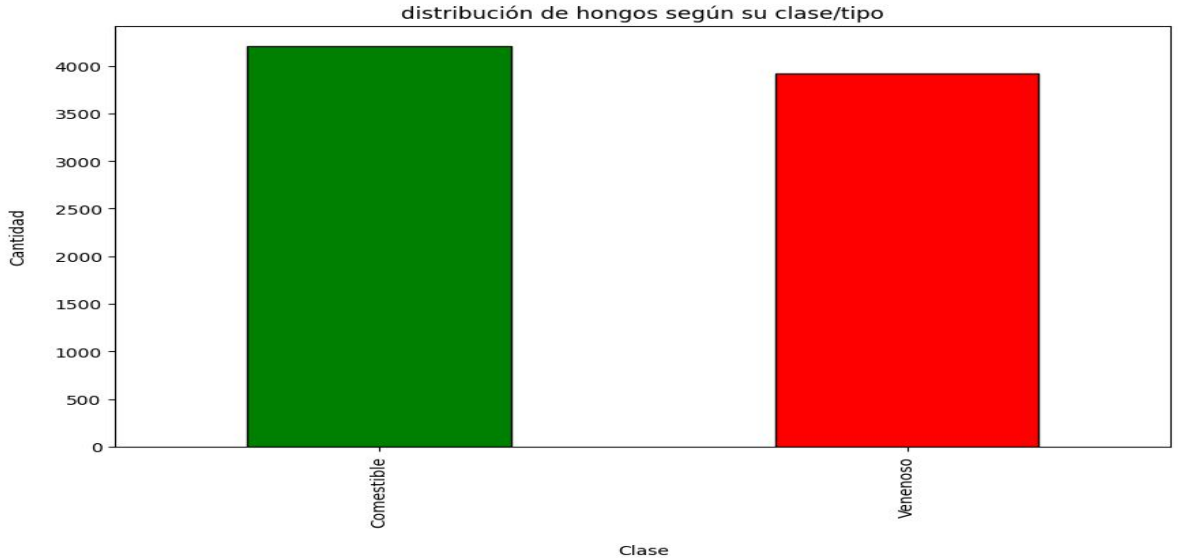
A su vez, cada variable contiene valores únicos, los cuales podemos visualizar cuántos valores y cuales.

3.- EDA: Exploratory Data Analysis

En el análisis exploratorio, se analizaron las características más “visibles” de un hongo. Esto con la intención de obtener un panorama general de lo que la mayoría podría reconocer a simple vista sin tener que manipular el hongo y exponerse a tener contacto con el mismo.

la primera incógnita a estudiar dentro de la temática era: la distribución de los hongos comestibles y venenosos, donde se observa que no hay una diferencia relevante entre uno resultado y otro.

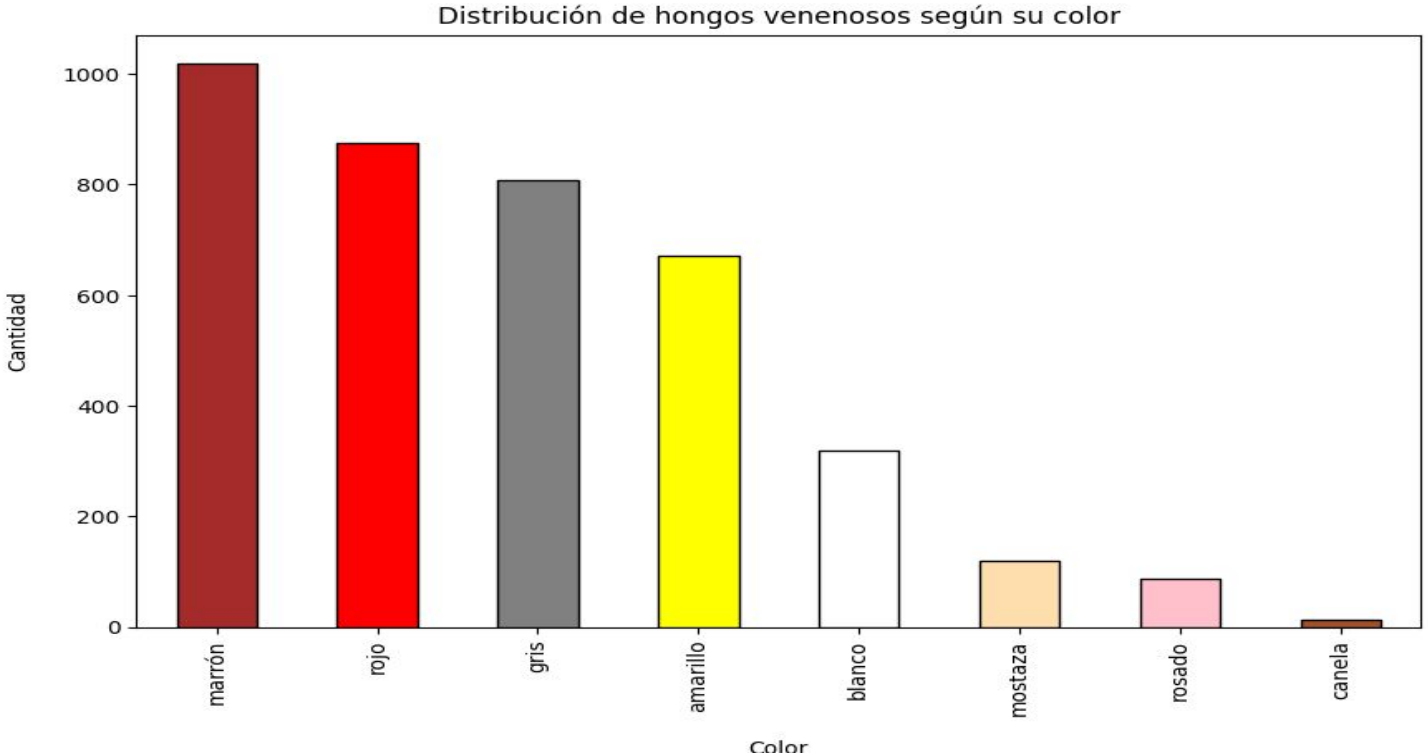
Siendo 4208 hongos de tipo comestible y 3916 de tipo venenoso.



edible (comestibles): 4208
poisonous (venenosos): 3916

Luego de la separación entre hongos comestibles y venenosos, pasamos a cuantificar la distribución de cada clase por el color del sombrero.

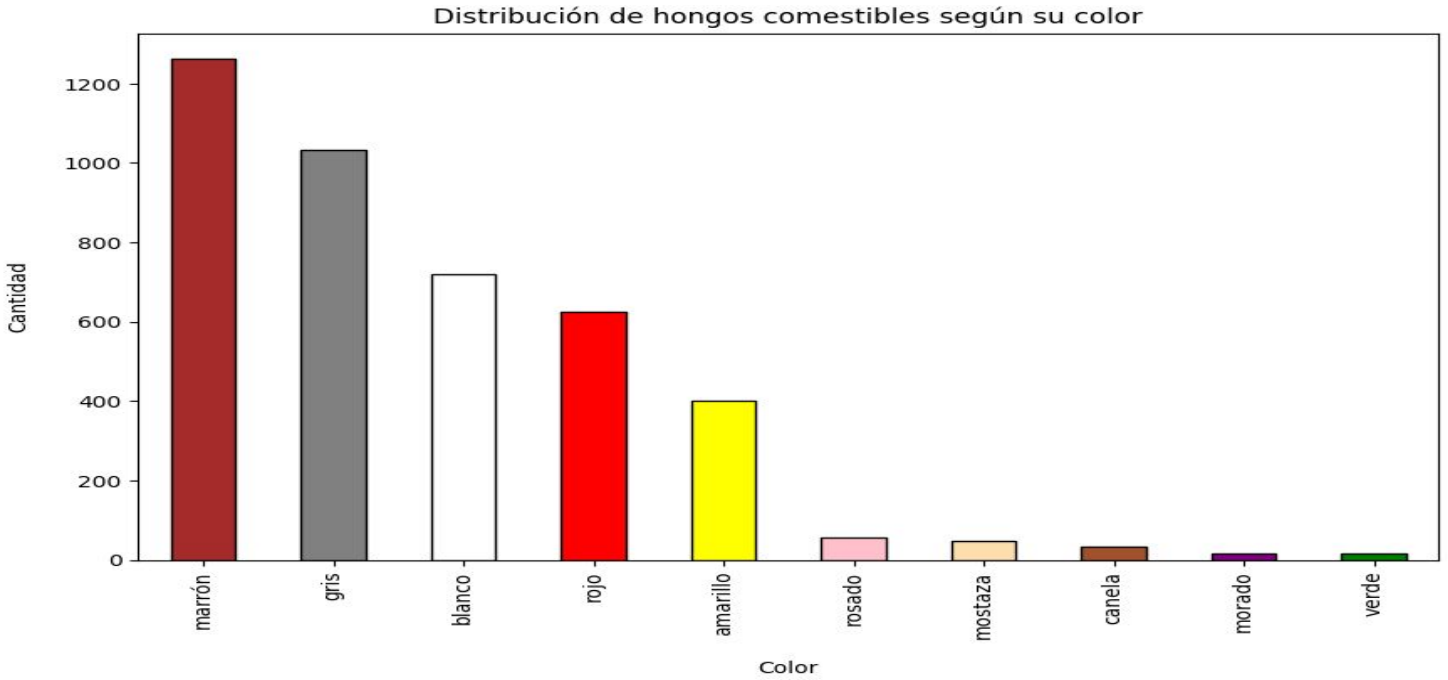
En esta distribución podemos observar que en el caso de los hongos venenosos, los más comunes de ver son los de color marrón (1020), seguidos por los rojos(876) y grises(808).



marrón: 1020
rojo: 876
gris:808
amarrillo:672
blanco: 320
mostaza:120
rosado:88
canela:12

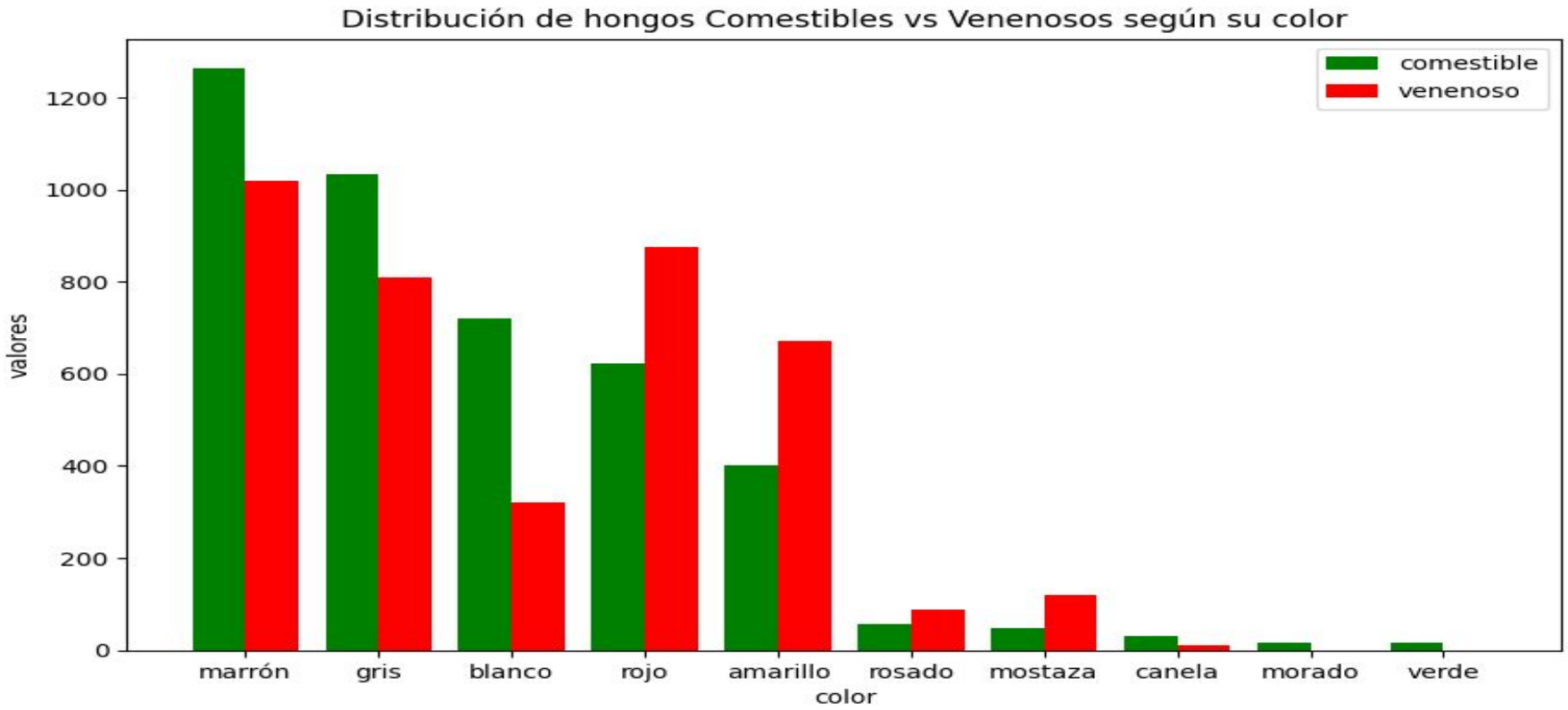
En el caso de los comestibles, observamos que se repite nuevamente como primera opción entre los mas comunes, los de color marrón, seguido por el gris y los blancos.

Por otro lado vemos que se encuentran 2 colores nuevos los cuales son morado y verde, dándonos información importante respecto a características únicas que solo se encuentran dentro de los hongos comestibles que no se ven en los venenosos.



marrón: 1264
gris:1032
blanco:720
rojo:624
amarrillo:400
rosado:56
mostaza:48
canela:32
morado:16
verde:16

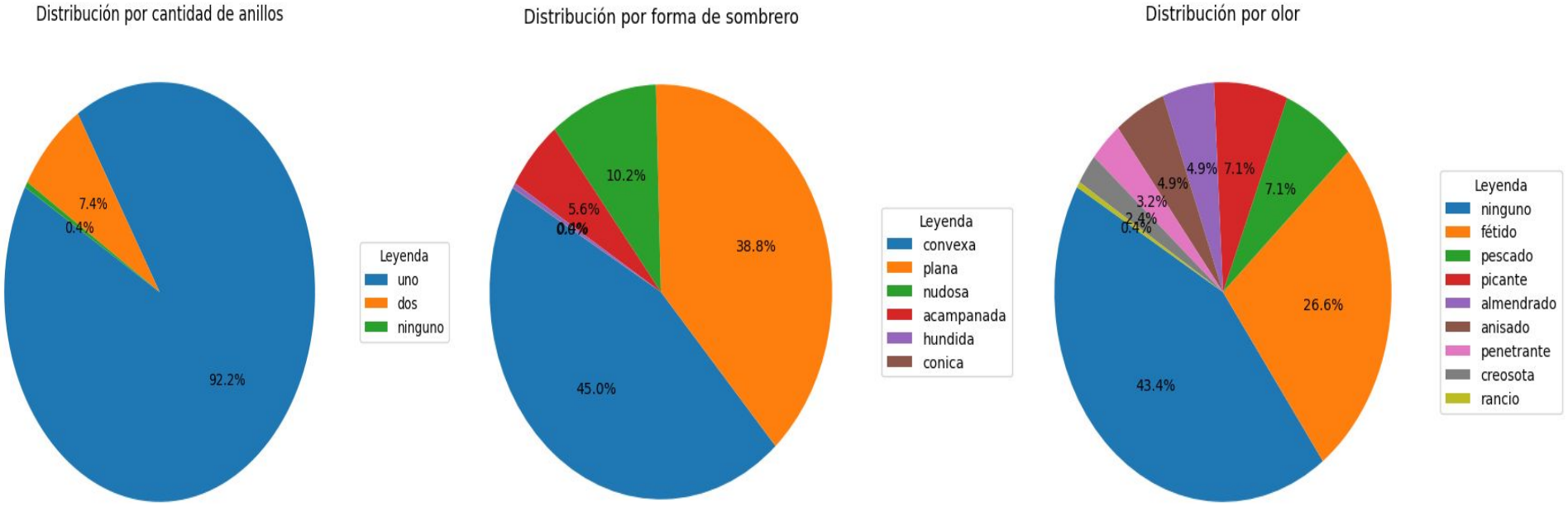
Luego se procede a hacer una comparativa entre ambos gráficos, para tener una visión mas global de ambas características y las distintas distribuciones.



4.- Pie Charts

Aparte de los diagramas de barra, se realizó una serie de pie charts donde se analiza la distribución en % de otra características como: cantidad de anillos, la forma del sombrero y la distribución por olor.

Dentro de lo que podemos observar rápidamente, podemos concluir que la gran mayoría tienen un solo anillo, las formas más comunes son convexa y plana finalizando con que un porcentaje alto no poseen olor alguno seguido por otros con olor fétido.



5.- Entrenamiento de Modelo

Para poder entrenar nuestro modelo, al tener variables de tipo categóricas, es necesario aplicar un método que las transforme en números, en este caso aplicaremos el método get_dummies() lo cual las transformará en 0 y 1.

Luego de transformar nuestros datos categóricos a númericos, separamos la columna target que nos servirá como base para parte de nuestro entrenamiento, en este caso la columna 'class' la cual identifica si es o no comestible un hongo. Luego procedemos a dividir nuestro dataset en 2 (training y test data) los cuales usaremos para aplicar en los distintos modelos a implementar con distintos porcentajes de prueba.

modelos a implementar:

Linear Regression:

Random Forest:

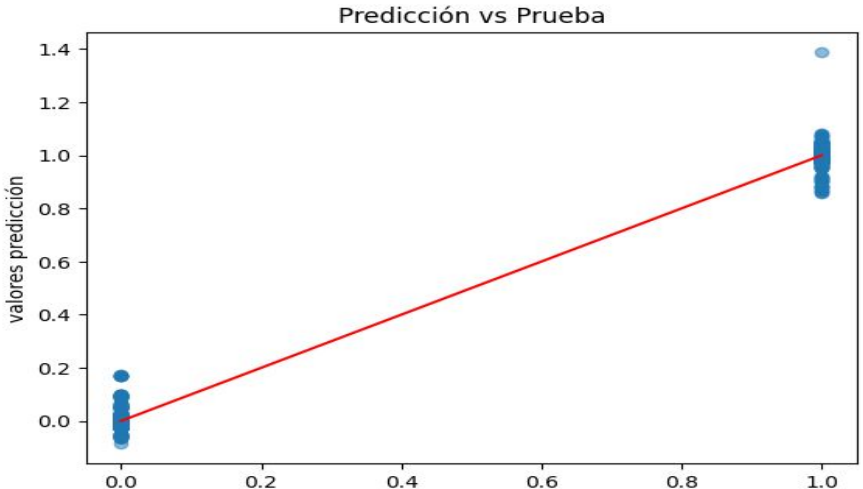
Decision Tree:

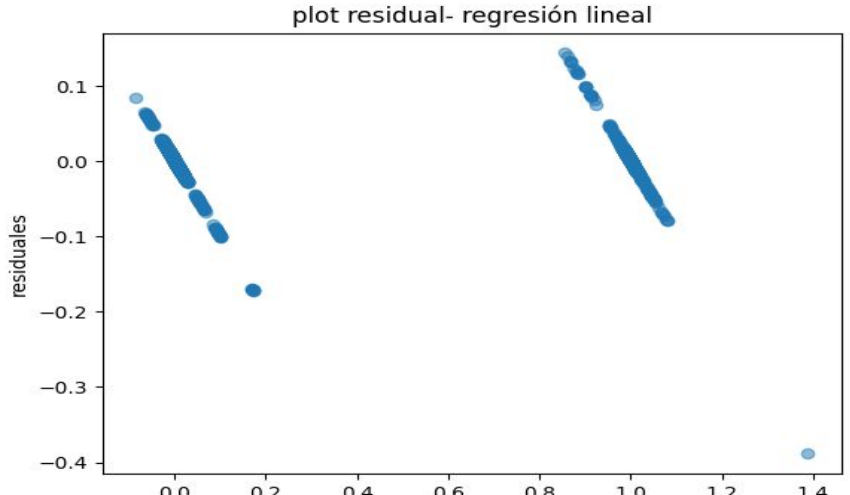
Logistic Classifier:

6.- Métricas de Desempeño y gráficos de Modelos aplicados

Linear regression:

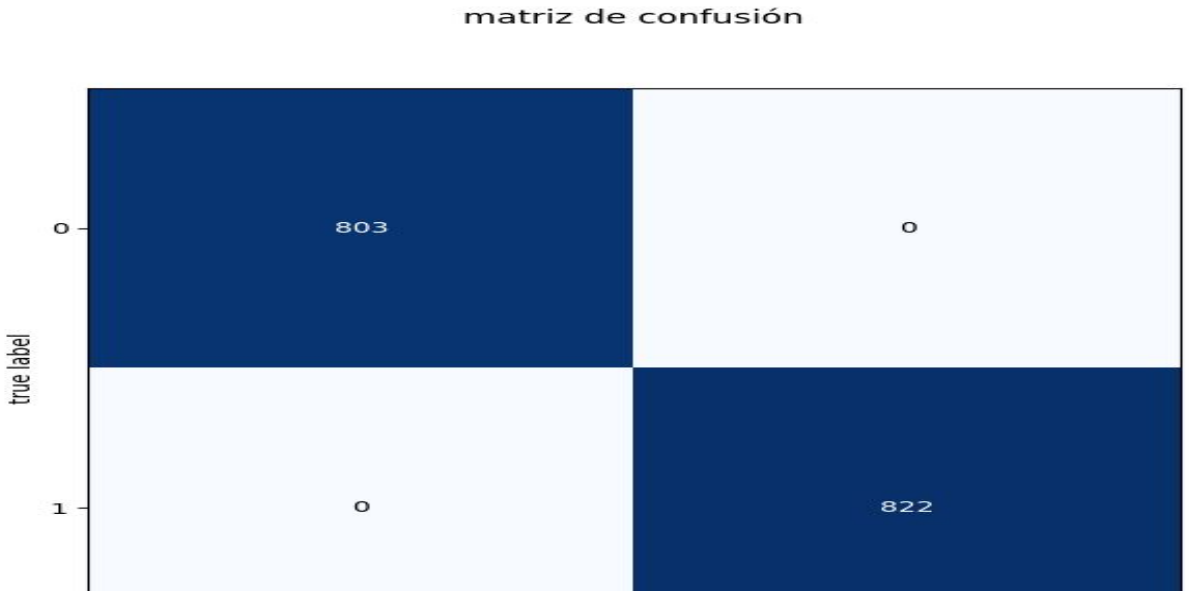
r2 score	0.9968724372399694
Mean squeared score	0.0007807888996142607
Mean absolute score	0.014463003305288461
intercept	140550936339.36368





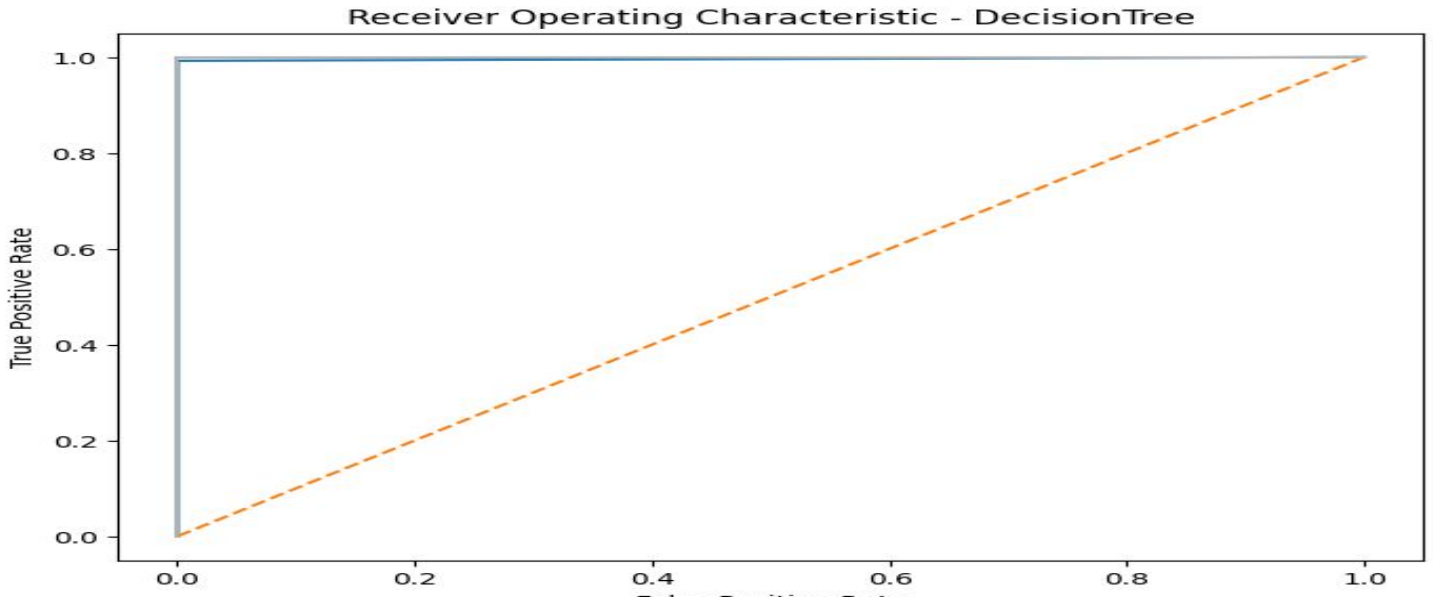
Random Forest:

accuracy	1.0
Precision	1.0
Recall	1.0
f1_score	1.0



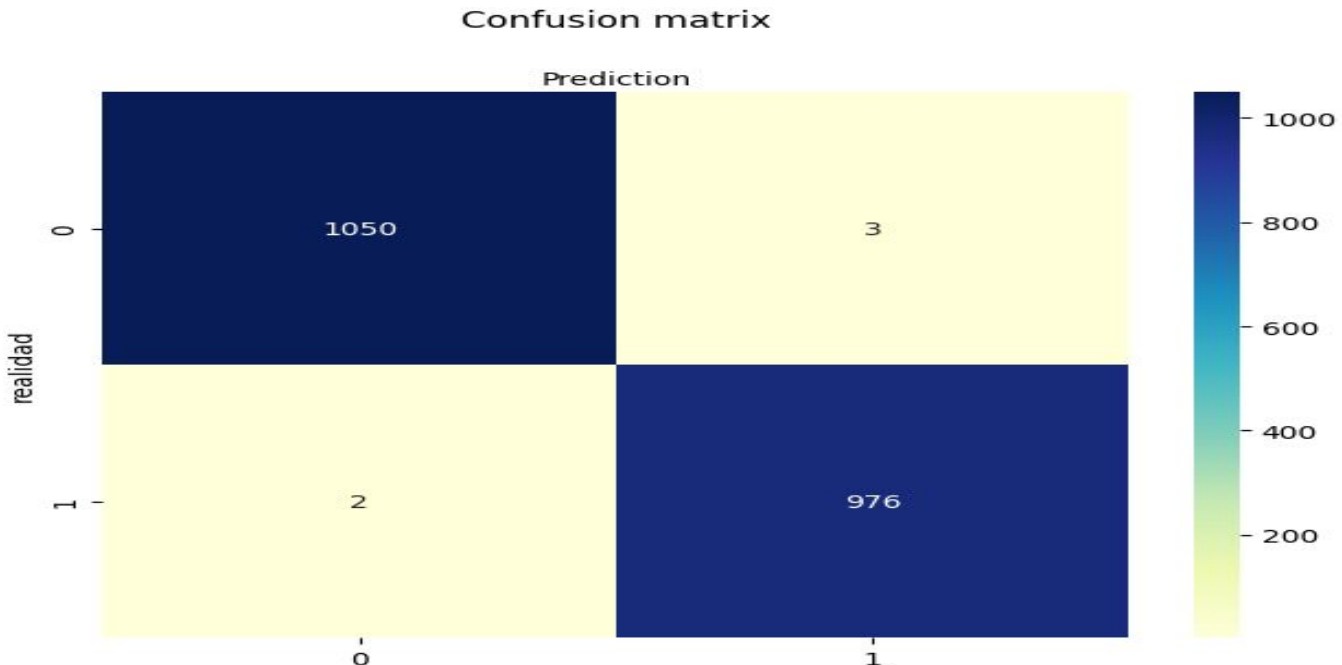
Decision Tree:

tdm.score	0.9963076923076923
Precision	0.9963076923076923
Recall	0.9963076923076923
f1_score	0.9963076923076923



Logistic Classifier:

Precision	0.9975381585425899
f1_score	0.9975381585425899
accuracy	0.9975381585425899



7.- Conclusión

En base a los resultados obtenidos por los diversos modelos entrenados, podemos concluir que todos concuerdan en cuanto a las predicciones arrojadas en cuanto a las características que se deseaban medir (comestibles o venenoso). Parte de esto se debe a que también la predicción se basa en solo 2 resultados.

Por otro lado es importante resaltar que la exploración previa de los datos, nos ayuda a crear mediante pequeños datos una idea o mapa más generalizado de cómo se comportan los datos y anidar pequeñas respuestas que sumadas se convierten en nueva información.