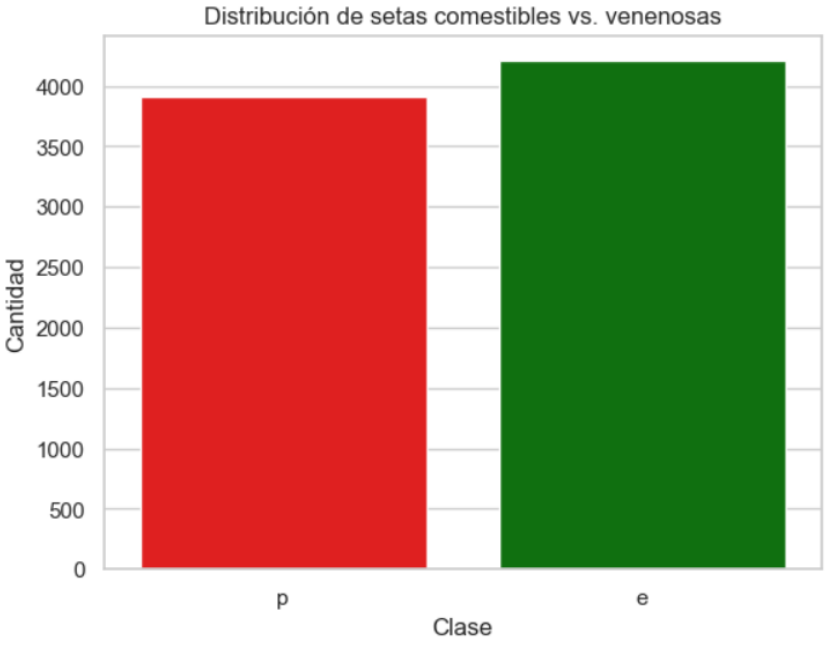
Análisis de Setas - Informe Ejecutivo

Este informe resume un análisis completo del Mushroom Dataset, combinando técnicas de **aprendizaje no supervisado** (PCA + K-Means) y **supervisado** (Random Forest).
El objetivo es explorar patrones ocultos en los datos y clasificar correctamente setas comestibles y venenosas.

EDA

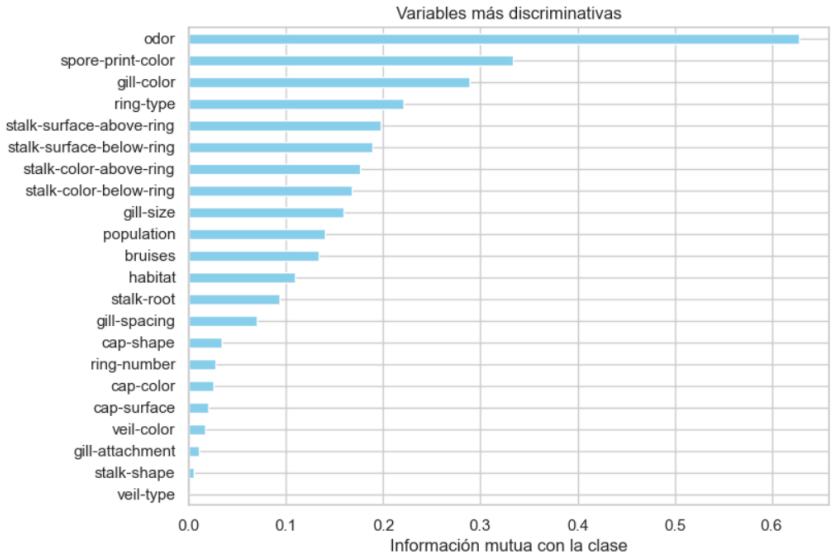
Se exploró el dataset original con 8124 instancias y 21 variables categóricas.



Rojo: Venenoso Verde: Comestible

Se aprecia que la cantidad de setas comestibles y venenosas que hay en el dataset está igualada.

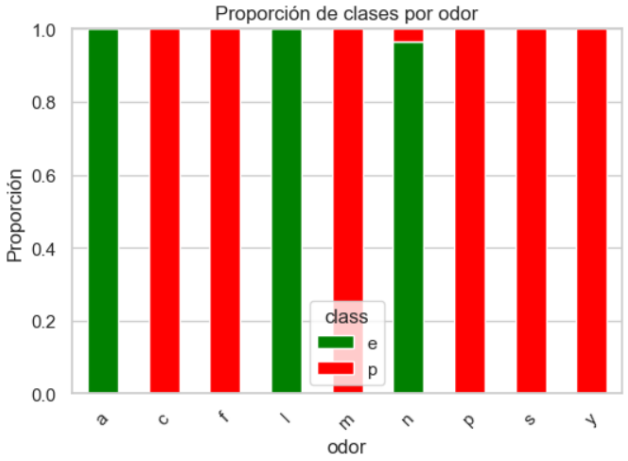
Variables más discriminativas



“Odor”, seguido de “spore-print-color” y “gill-color” son las variables que más relación muestran con la clase”class”.

Por otra parte, veil-type, stalk-shape y gill-attachment son los que menos.

Relación Odor con Class



Al ver que “odor” es el más discriminativo en el ranking anterior, y haciendo este análisis bivariado con “class”, se puede observar que hay tipos de olores 100% venenosos y 100% comestibles.

ETL

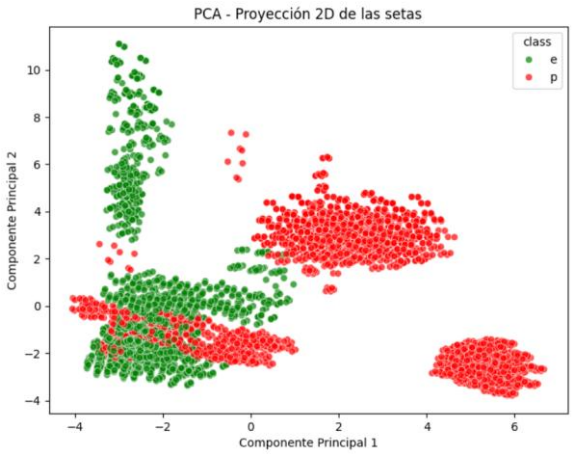
Se eliminan las columnas veil-type y gill-attachment debido a que eran poco informativas

Y se crea la columna has_odor que nos dice si la seta es inolora o no.

```
<class 'pandas.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   class                     8124 non-null   category
 1   cap-shape                 8124 non-null   category
 2   cap-surface               8124 non-null   category
 3   cap-color                 8124 non-null   category
 4   bruises                   8124 non-null   category
 5   odor                      8124 non-null   category
 6   gill-spacing              8124 non-null   category
 7   gill-size                 8124 non-null   category
 8   gill-color                8124 non-null   category
 9   stalk-shape               8124 non-null   category
 10  stalk-root                8124 non-null   category
 11  stalk-surface-above-ring  8124 non-null   category
 12  stalk-surface-below-ring  8124 non-null   category
 13  stalk-color-above-ring    8124 non-null   category
 14  stalk-color-below-ring    8124 non-null   category
 15  veil-color                8124 non-null   category
 16  ring-number               8124 non-null   category
 17  ring-type                 8124 non-null   category
 18  spore-print-color         8124 non-null   category
 19  population                8124 non-null   category
 20  habitat                   8124 non-null   category
 21  has_odor                  8124 non-null   int64
dtypes: category(21), int64(1)
memory usage: 231.6 KB
```

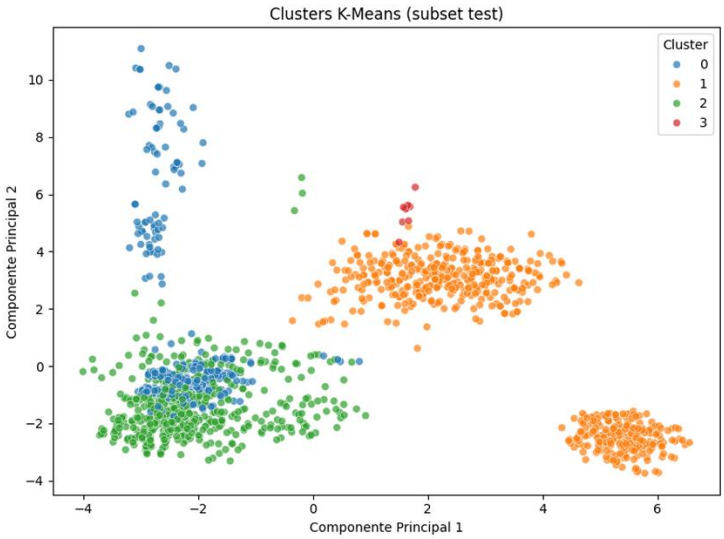
has_odor
1
1
1
1
0

Reducción de Dimensionalidad (PCA)



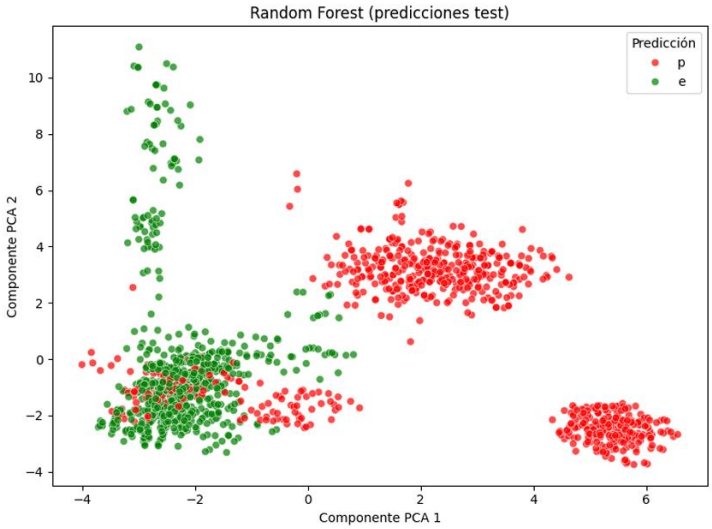
Aplicando PCA reducimos las 95 features a 2 componentes para visualización y a 10 componentes para modelado.
Las primeras 2 componentes muestran que los datos se agrupan naturalmente, sugiriendo estructuras internas.

Clustering con K-Means



Los clusters muestran estructuras internas claras en los datos, aunque no coinciden perfectamente con las clases reales.

Random Forest



Random Forest, usando 10 componentes PCA, logra un accuracy de 0.998 y separa casi perfectamente las setas comestibles de las venenosas.

Conclusión

K-Means es útil para explorar patrones, mientras que Random Forest muestra predicciones muy precisas. PCA permite reducir dimensionalidad sin perder información importante. La combinación de estas técnicas ofrece un flujo de análisis completo y eficiente.