Workshop Clustering PCA

Autores: Max Beltrán , Fernando García Catalán
Fecha: 30/05/2025
Github: https://github.com/mr-melenas/unsupervised-ml-mushroom

Introducción

Este repositorio contiene un taller práctico orientado al aprendizaje automático no supervisado, usando técnicas de PCA y Clustering (KMeans), junto con una comparativa con un modelo supervisado (Random Forest).

Usaremos el Mushroom Dataset, un conjunto de datos muy conocido en el ámbito educativo que contiene información sobre diferentes tipos de hongos, incluyendo su clasificación como comestibles o venenosos.

Data Source: Mushroom Dataset - UCI Repository

Data Analysis

Field List

#	name	role	type	demographic
0	poisonous	Target	Categorical	None
1	cap-shape	Feature	Categorical	None
2	cap-surface	Feature	Categorical	None
3	cap-color	Feature	Binary	None
4	bruises	Feature	Categorical	None
5	odor	Feature	Categorical	None
6	gill-attachment	Feature	Categorical	None
7	gill-spacing	Feature	Categorical	None
8	gill-size	Feature	Categorical	None
9	gill-color	Feature	Categorical	None
10	stalk-shape	Feature	Categorical	None
11	stalk-root	Feature	Categorical	None
12	stalk-surface-above-ring	Feature	Categorical	None
13	stalk-surface-below-ring	Feature	Categorical	None
14	stalk-color-above-ring	Feature	Categorical	None
15	stalk-color-below-ring	Feature	Categorical	None
16	veil-type	Feature	Categorical	None
17	veil-color	Feature	Binary	None
18	ring-number	Feature	Categorical	None
19	ring-type	Feature	Categorical	None
20	spore-print-color	Feature	Categorical	None
21	population	Feature	Categorical	None
22	habitat	Feature	Categorical	None

Field Description

Variable Name	Role	Type	Description	Missing Values
poisonous	Target	Categorical		no
cap-shape	Feature	Categorical	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s	no
cap-surface	Feature	Categorical	fibrous=f,grooves=g,scaly=y,smooth=s	no
cap-color	Feature	Binary	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y	no
bruises	Feature	Categorical	bruises=t,no=f	no
odor	Feature	Categorical	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s	no
gill-attachment	Feature	Categorical	attached=a,descending=d,free=f,notched=n	no
gill-spacing	Feature	Categorical	close=c,crowded=w,distant=d	no
gill-size	Feature	Categorical	broad=b,narrow=n	no
gill-color	Feature	Categorical	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y	no
stalk-shape	Feature	Categorical	enlarging=e,tapering=t	no
stalk-root	Feature	Categorical	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?	yes
stalk-surface-above-ring	Feature	Categorical	fibrous=f,scaly=y,silky=k,smooth=s	no
stalk-surface-below-ring	Feature	Categorical	fibrous=f,scaly=y,silky=k,smooth=s	no
stalk-color-above-ring	Feature	Categorical	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y	no
stalk-color-below-ring	Feature	Categorical	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y	no
veil-type	Feature	Binary	partial=p,universal=u	no
veil-color	Feature	Categorical	brown=n,orange=o,white=w,yellow=y	no
ring-number	Feature	Categorical	none=n,one=o,two=t	no
ring-type	Feature	Categorical	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z	no
spore-print-color	Feature	Categorical	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y	no
population	Feature	Categorical	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y	no
habitat	Feature	Categorical	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d	no

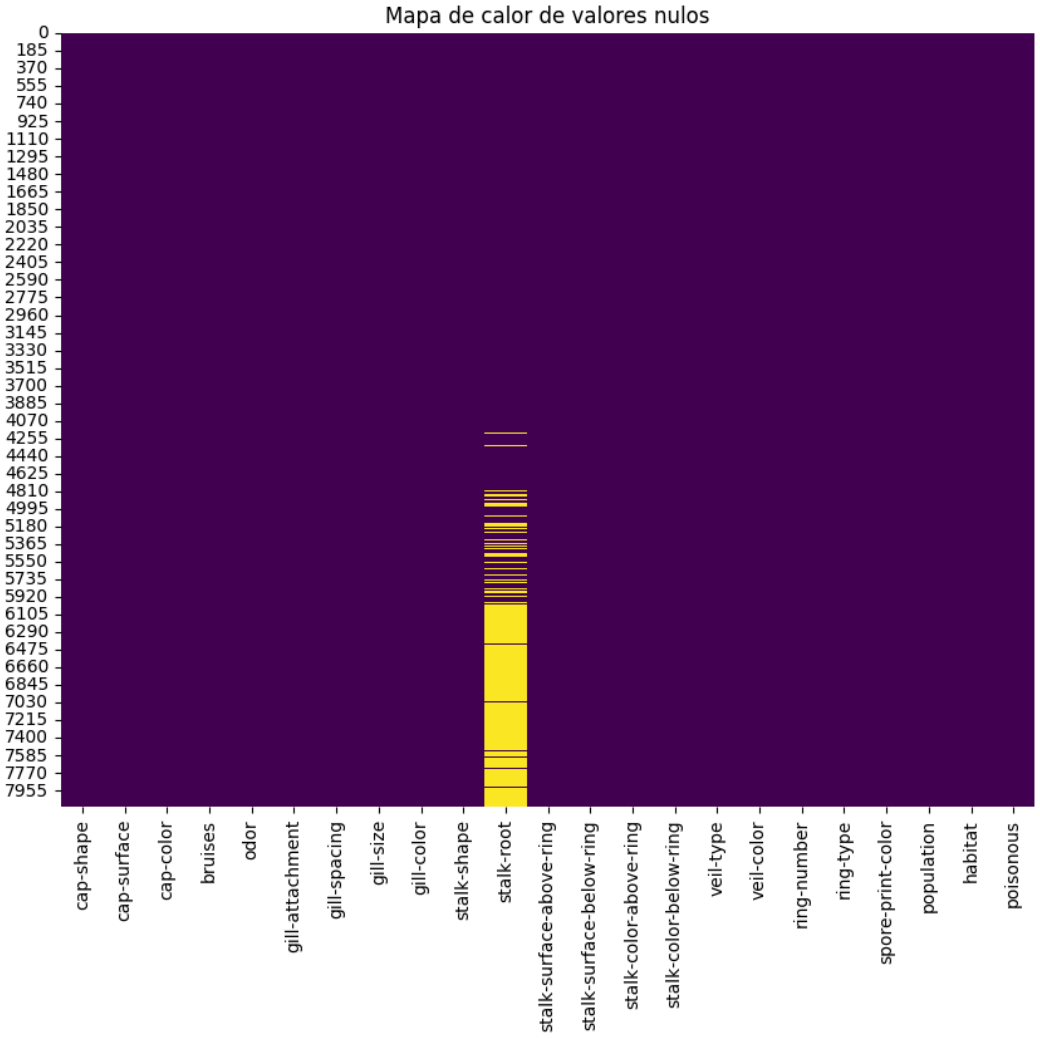
Field Inmmersion

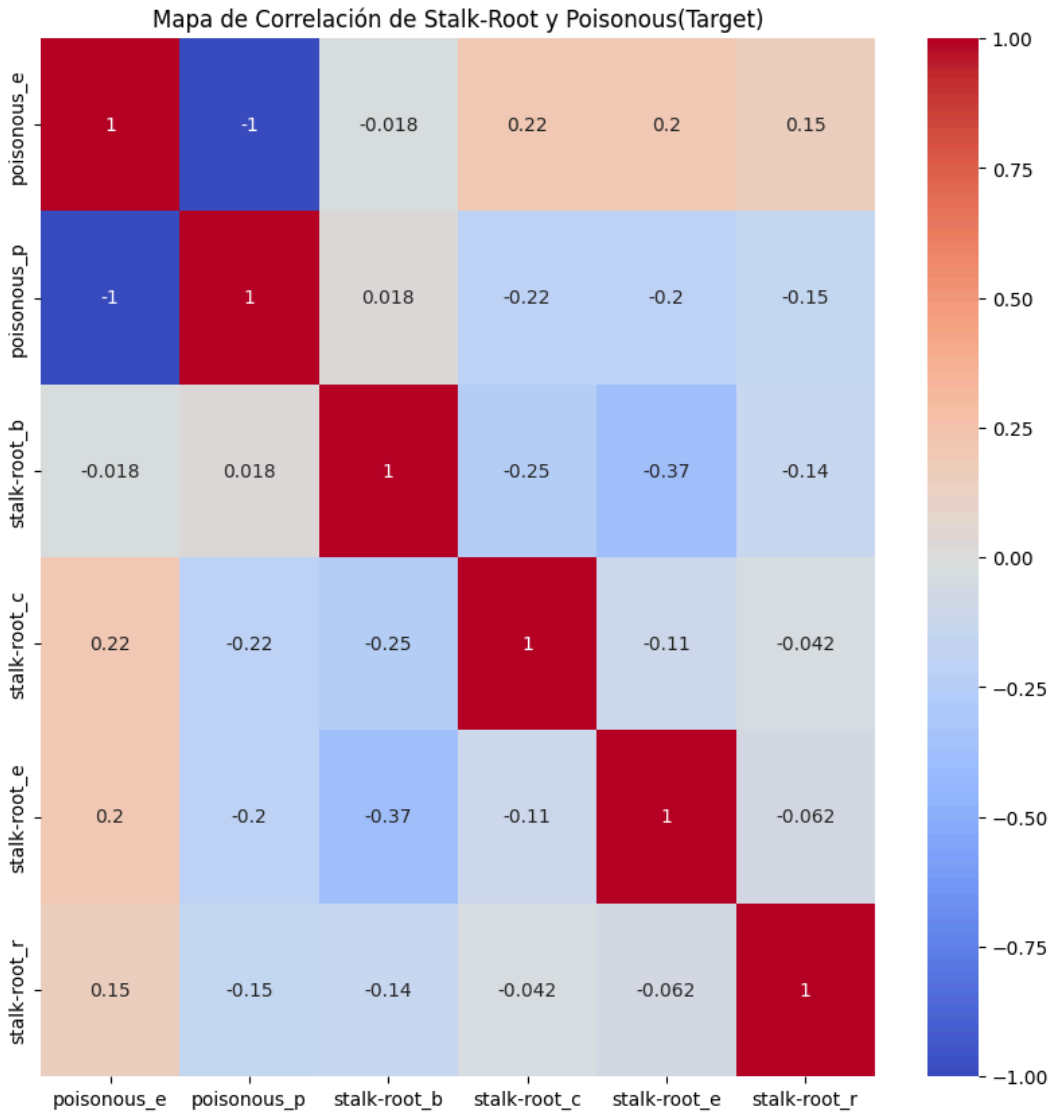
name	count	unique	top	freq
cap-shape	8124	6	x	3656
cap-surface	8124	4	y	3244
cap-color	8124	10	n	2284
bruises	8124	2	f	4748
odor	8124	9	n	3528
gill-attachment	8124	2	f	7914
gill-spacing	8124	2	c	6812
gill-size	8124	2	b	5612
gill-color	8124	12	b	1728
stalk-shape	8124	2	t	4608
stalk-root	5644	4	b	3776
stalk-surface-above-ring	8124	4	s	5176
stalk-surface-below-ring	8124	4	s	4936
stalk-color-above-ring	8124	9	w	4464
stalk-color-below-ring	8124	9	w	4384
veil-type	8124	1	p	8124
veil-color	8124	4	w	7924
ring-number	8124	3	o	7488
ring-type	8124	5	p	3968
spore-print-color	8124	9	w	2388
population	8124	6	v	4040
habitat	8124	7	d	3148
poisonous	8124	2	e	4208

Data Inspection

names	nulls
cap-shape	0
cap-surface	0
cap-color	0
bruises	0
odor	0
gill-attachment	0
gill-spacing	0
gill-size	0
gill-color	0
stalk-shape	0
stalk-root	2480
stalk-surface-above-ring	0
stalk-surface-below-ring	0
stalk-color-above-ring	0
stalk-color-below-ring	0
veil-type	0
veil-color	0
ring-number	0
ring-type	0
spore-print-color	0
population	0
habitat	0
poisonous	0

Null Values


Mapa de calor de valores nulos
0
185
370
555
740
925
1110
1295
1480
1665
1850
2035
2220
2405
2590
2775
2960
3145
3330
3515
3700
3885
4070
4255
4440
4625
4810
4995
5180
5365
5550
5735
5920
6105
6290
6475
6660
6845
7030
7215
7400
7585
7770
7955
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
poisonous



- La inspección de datos arroja la presencia de 2480 valores nulos para el campo ‘stalk-root, lo que representa un 30% del total de los datos.
- Al ser un volumen tan elevado de información, descartamos por completa la opción de eliminar los registros con nulos.

- Contrastamos con un mapa de correlaciones, el bajo impacto de este campo con la variable objetivo e imputamos los campos nulos con la moda.

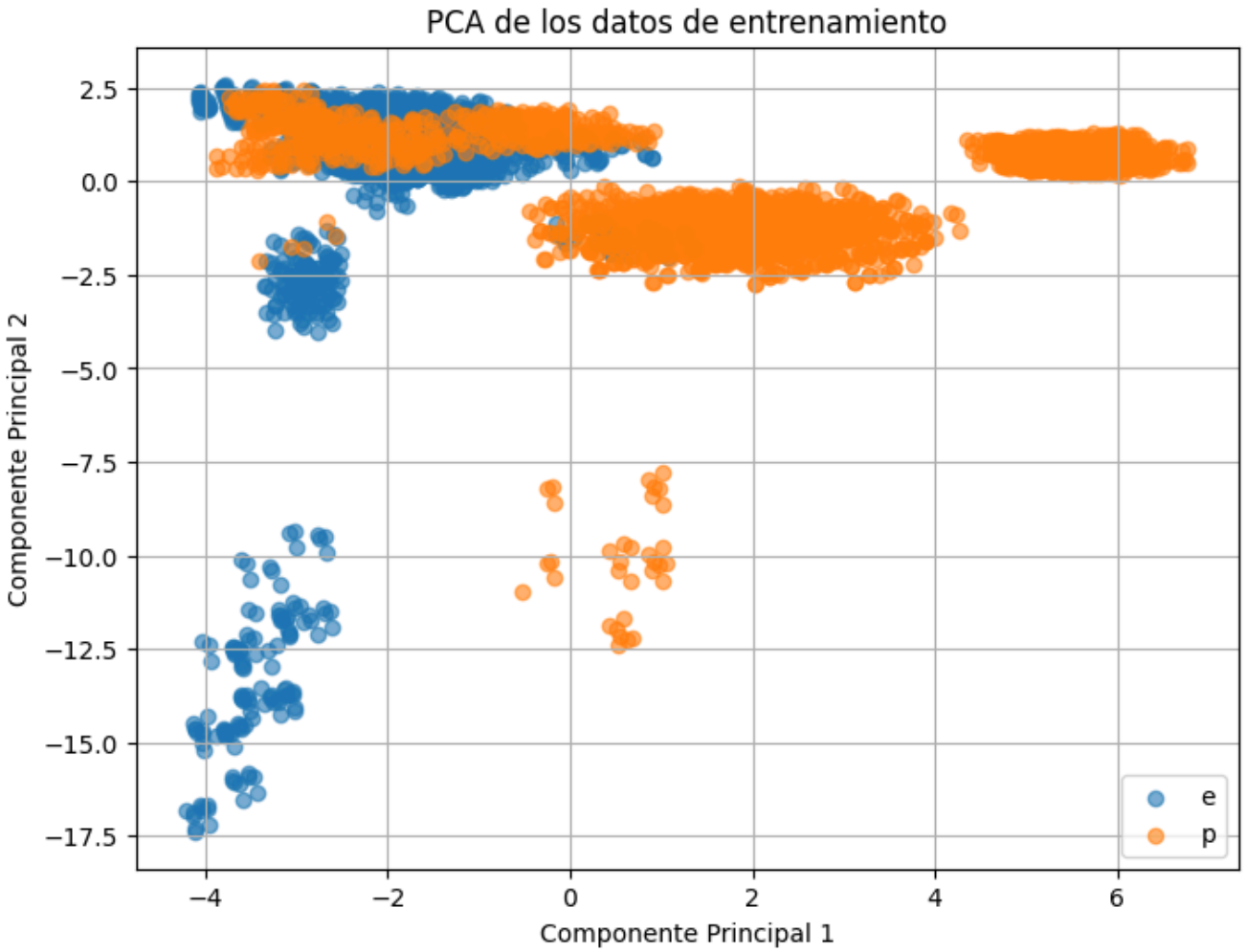
Non Valuable Values

La variable ‘veil-type’ muestra un mismo tipo de valor para todos los registros, por lo que se elimina del dataframe.

Target and Features

Target = ‘poisonous’
Features = Todas menos ‘veil-type’

Principal Component Analysis



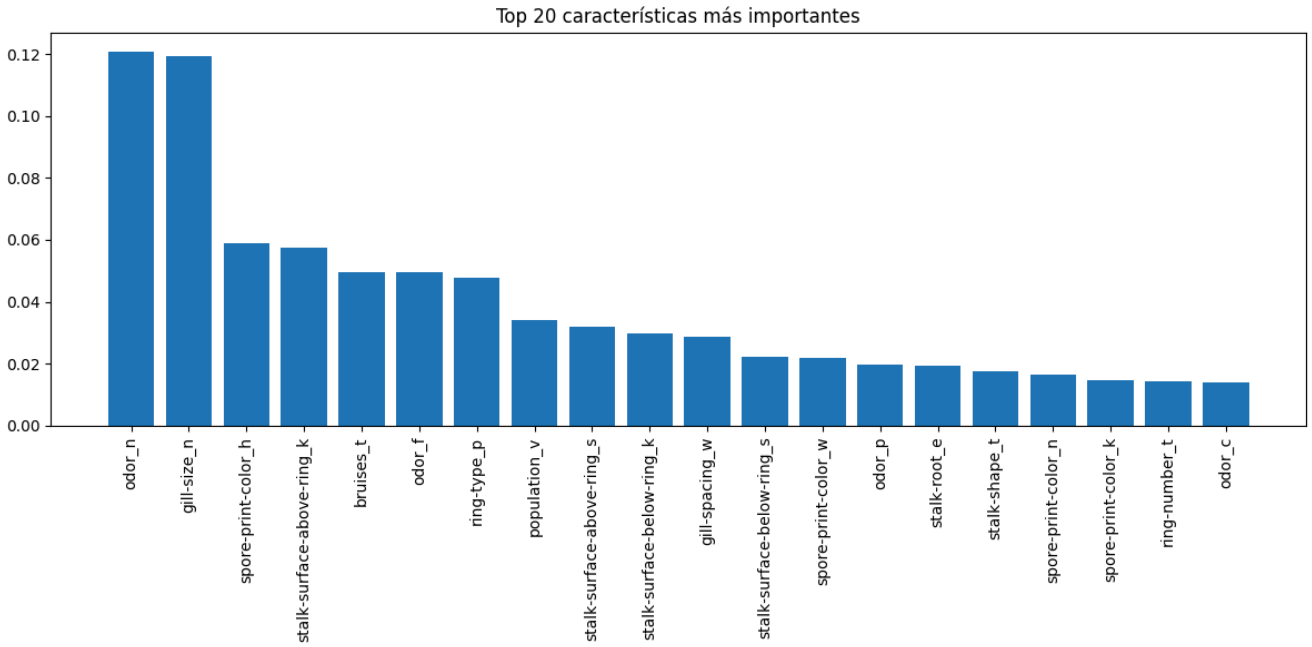
Varianza explicada por cada componente: [0.09446195 0.07177885]
Varianza total explicada por las dos primeras componentes: 0.16624079655030422

El PCA ha logrado reducir la complejidad de los datos al representarlos en un espacio 2D (PC1 y PC2). Esto simplifica la visualización y el análisis

Sin embargo, la PCA muestra **baja varianza explicada**: El valor de 16.62% es relativamente bajo. Esto significa que la representación bidimensional los datos, creada por las dos primeras componentes principales, no captura una gran cantidad de la variabilidad original. En otras palabras, se ha perdido una parte significativa de la información al reducir la dimensionalidad a dos componentes.

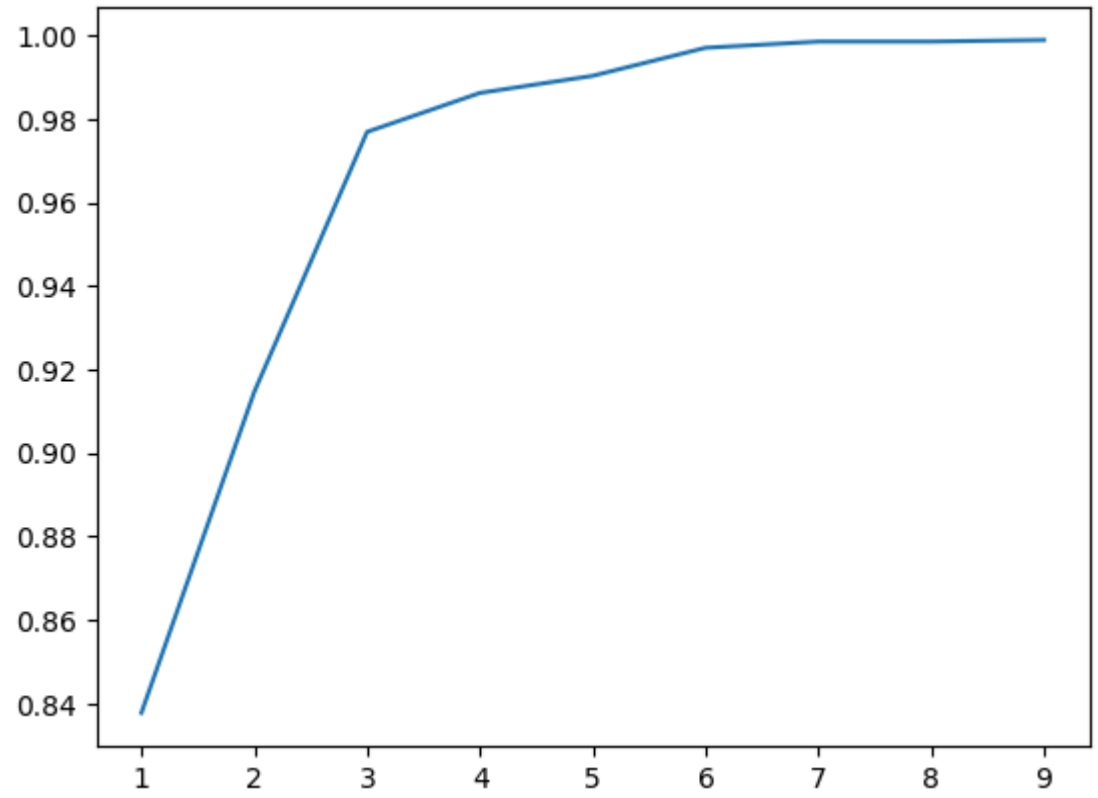
Classification and Importance

Entrenamos un clasificador con *Random Forest* sobre un conjunto de entrenamiento(5443 filas, 94 columnas) y calculamos la importancia de cada característica.



El modelo obtiene una puntuación en **Accuracy** de 1.0

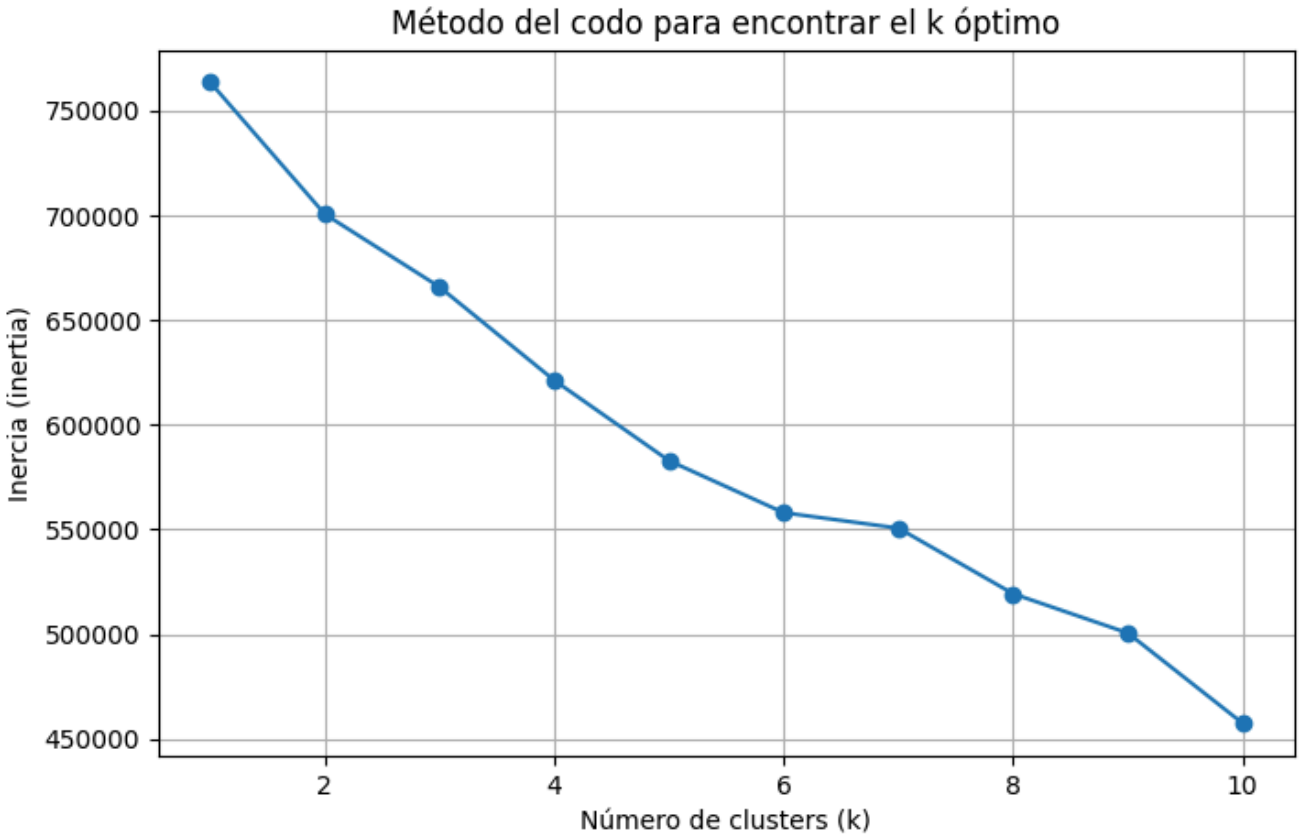
Feature Reduction



Buscamos el mínimo rango de features para obtener una buena accuracy y definimos el valor en 10 features.

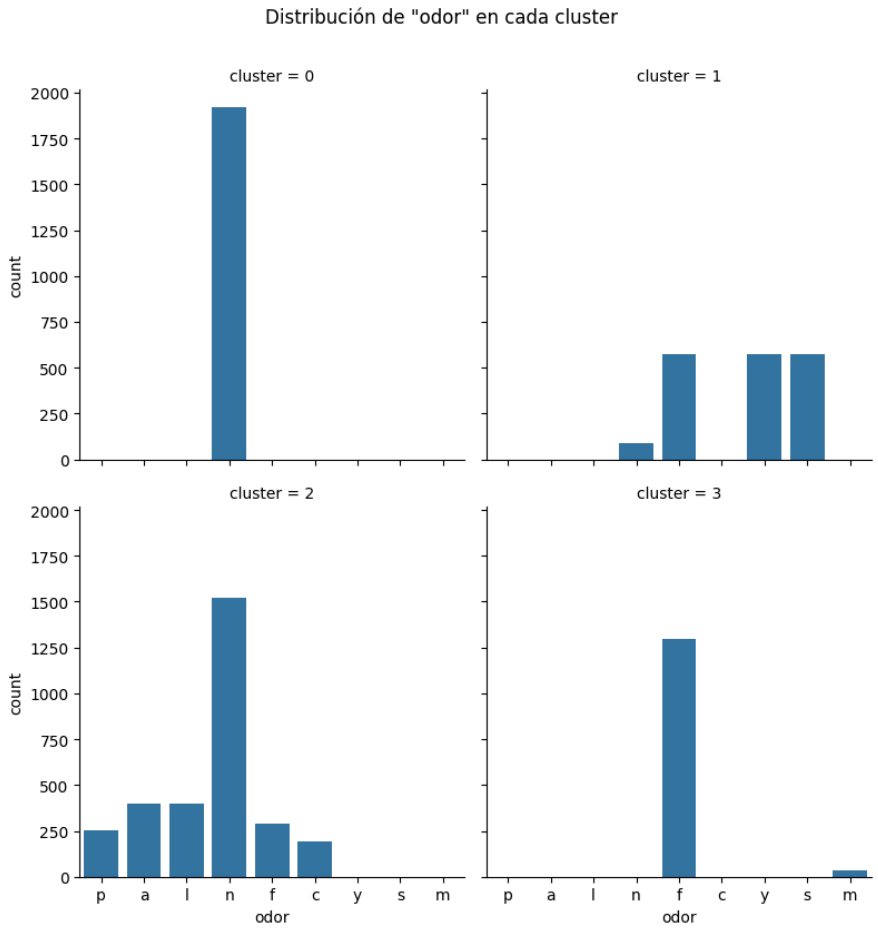
Clustering

Elbow Method



Según los valores aportados, definimos en 4 el número de clusters necesarios.

Cluster distribution

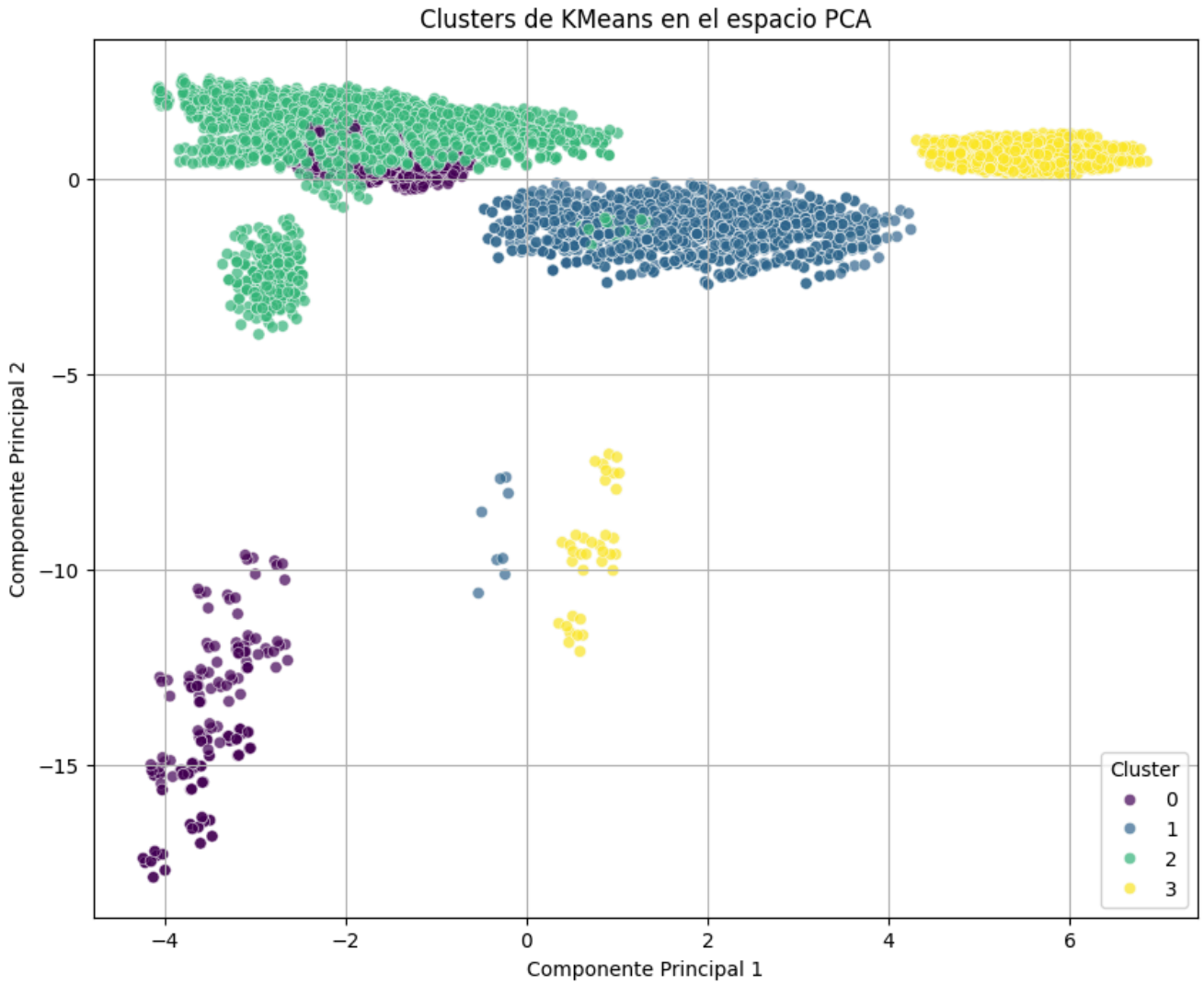


Este gráfico muestra la distribución de los valores de la variable odor dentro de cada cluster identificado por el algoritmo de K-Means

Cluster	Olores dominantes	Interpretación
Cluster 0	Principalmente `"n"`	Muestra homogénea con olor neutro; `"n"` parece ser un fuerte determinante.
Cluster 1	Mezcla de `"c"`, `"y"`, `"s"`	Grupo más heterogéneo; sin un olor claramente dominante.
Cluster 2	`"n"` (mayoría), también `"a"`, `"l"`, `"f"`	Algo más diverso que el cluster 0, aunque sigue predominando el olor neutro `"n"`.
Cluster 3	Dominado por `"f"`, algo de `"m"`	Grupo muy homogéneo, probablemente con olor desagradable como rasgo principal.

Es un grupo bastante homogéneo en torno a ese olor.
Si el olor está relacionado con si un hongo es venenoso (poisonous) o comestible, entonces ciertos olores pueden ser un indicador fuerte de su clasificación

K-Means Clusters PCA



Separación de clusters

El algoritmo K-Means ha identificado **4 clusters distintos** (0, 1, 2, 3) que muestran una separación bastante clara en el espacio reducido de componentes principales.

Características de cada cluster

Cluster 0 (morado): Se concentra en la región inferior izquierda, con valores negativos en ambas componentes principales. Parece ser el más compacto y homogéneo de los grupos.

Cluster 1 (verde): Ocupa principalmente la región superior, extendiéndose tanto hacia valores negativos como positivos del primer componente. Es el cluster más disperso espacialmente.

Cluster 2 (azul): Se localiza en la región central-derecha, con una distribución bastante densa y alargada horizontalmente.

Cluster 3 (amarillo): Aparece en dos zonas principales - una concentración en la parte superior derecha y varios puntos dispersos en la región central-inferior.

Calidad de la segmentación

La separación visual relativamente clara entre grupos sugiere que K-Means ha logrado una segmentación efectiva. Sin embargo, hay algunas observaciones:

- Existe cierto solapamiento entre clusters, especialmente en las zonas de transición
- El cluster 3 (amarillo) muestra cierta dispersión, lo que podría indicar heterogeneidad interna
- La forma alargada de algunos clusters sugiere que los datos podrían beneficiarse de algoritmos que manejen formas no esféricas

Implicaciones para el análisis

Esta visualización indica que tus datos originales probablemente contienen **patrones o subgrupos naturales** que el PCA ha logrado preservar en las dos primeras componentes principales, y que K-Means ha identificado exitosamente estos grupos en el espacio reducido.