Informe taller aprendizaje No supervisado

1. Introducción:

- En este proyecto se analiza el conjunto de datos de hongos de la web UCI: https://archive.ics.uci.edu/dataset/73/mushroom con 8123 datos, y nuestro objetivo es clasificar si un hongo es comestible o venenoso a partir de sus características morfológicas.
- Vamos a aplicar técnicas de preprocesamiento, reducción de dimensionalidad , clasificación supervisada y clustering no supervisado para explorar y modelar los datos como se nos solicita en el ejercicio.

2. Carga y exploración de datos.

- Hemos cargado el dataset original mencionado antes, con sus 8124 instancias y 23 columnas, todas ellas eran categóricas, hemos renombrado las columnas como en la documentación que venía incluida en la carpeta descargada, class, será nuestra variable objetivo , ya que tiene dos opciones, si es venenosa o comestible, vimos también los siguientes datos edible(comestible): 4208 (51.8%) y poisonous(venenosa): 3916 (48.2%), lo que quiere decir que están bastante balanceadas las posibilidades.

3. Limpieza y preprocesamiento.

- Hemos buscado valores nulos y valores especiales como '?', se detectó que la columna stalk-root contenía muchos valores '?', que hemos convertido a nulos para tratarlos como tal , posteriormente hemos reemplazado con la categoría 'unknow' ya que consideramos que no debemos eliminarlos para no perder información relevante ya que es una cantidad considerables: 2480.
- Además hemos comprobado que la columna veil-type sólo tenía una categoría, es decir, no aportaba nada para nuestra clasificación, así que decidimos eliminarla por no aportar variabilidad.

4. Codificación de variables.

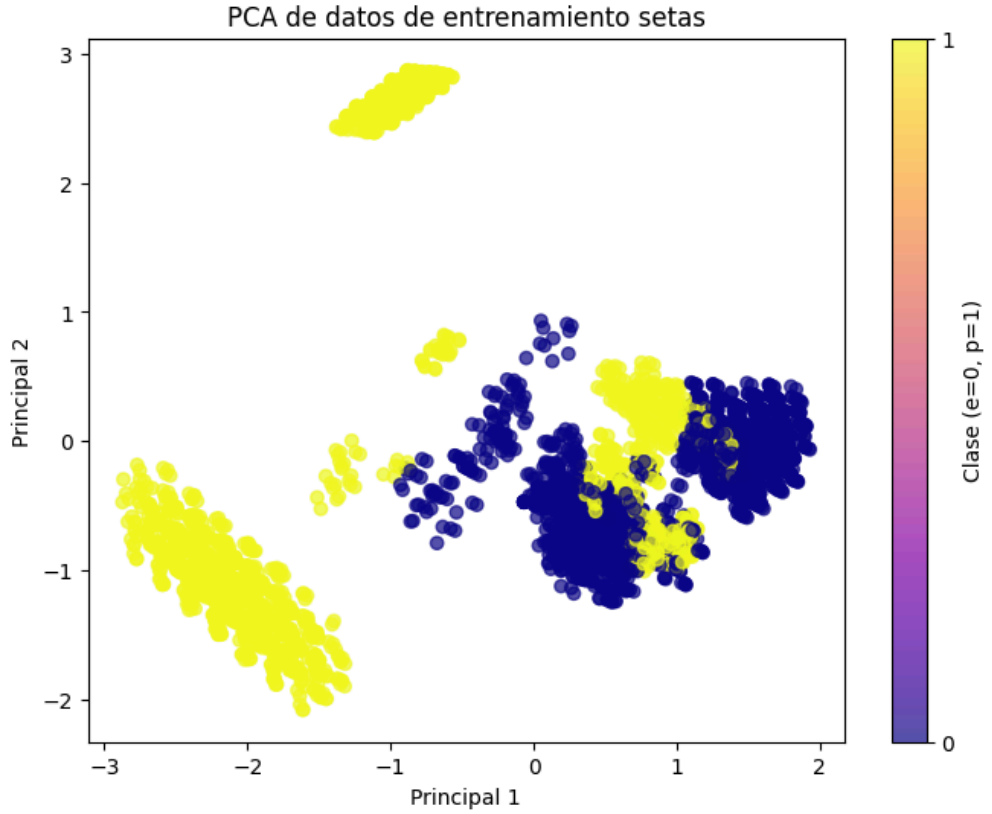
- Como nuestras variables son todas categóricas , hemos utilizado OneHotEncoder para convertirlas en variables numéricas, ya que no parecen tantas columnas, y para obtener un mejor orden, también hemos separado la variable objetivo que es 'class' como hemos dicho antes.

5. División del Dataset.

- Con nuestro dataset ya limpio y codificado, hemos procedido a dividir en conjuntos de entrenamiento y prueba utilizando train_test_split, reservando un 33% de los datos para la prueba, elegimos 33 para una mejor estimación del rendimiento.

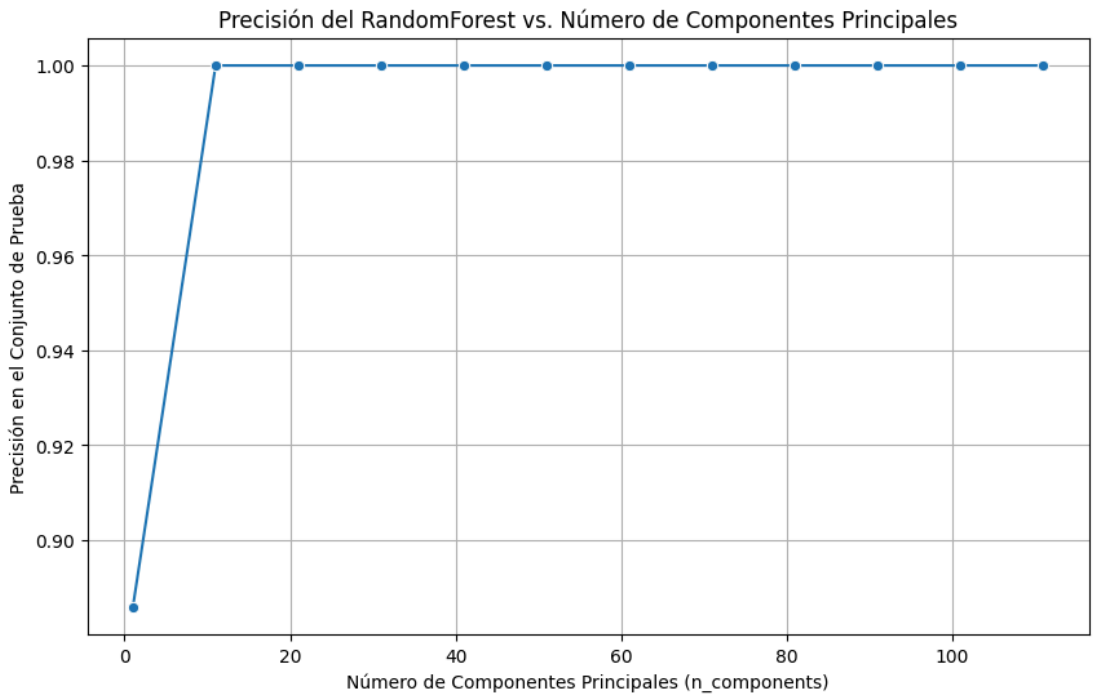
6. Reducción de dimensionalidad(PCA).

- Hemos aplicado PCA para reducir la dimensionalidad y facilitar la visualización, hemos seleccionado dos componentes principales, que nos permite representar los datos en el siguiente scatterplot, en el que observamos que las clases están separadas la mayoría aunque no perfectamente.



7. Clasificación supervisada.

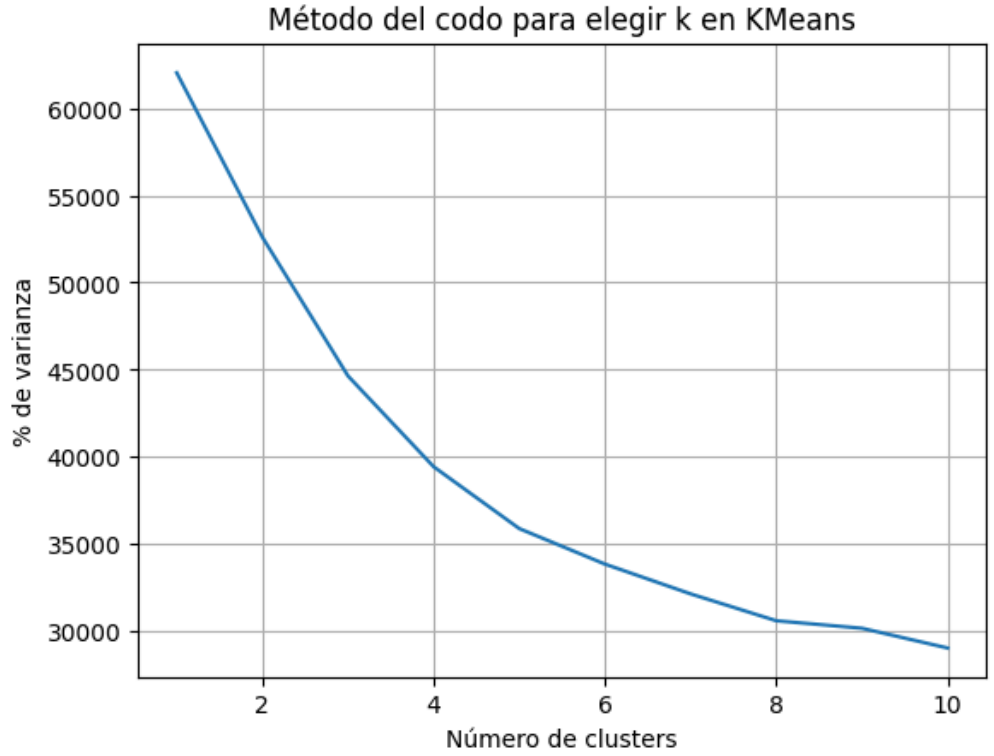
- Hemos entrenado un Random Forest sobre los datos transformados con PCA, el modelo ha alcanzado la precisión del 96%, lo que indica que las características permiten distinguir muy bien entre si un hongo es comestible o no, aunque también hemos probado sin el PCA aplicado y el accuracy es perfecto, un 100%.
- También hemos creado un bucle en el que empezamos en el 1 y vamos de 10 en 10 probando con las variables, con Random Forest y obtuvimos que con 11, aproxiamadamenteo el 10% de las features ya alcanzamos el score perfecto:



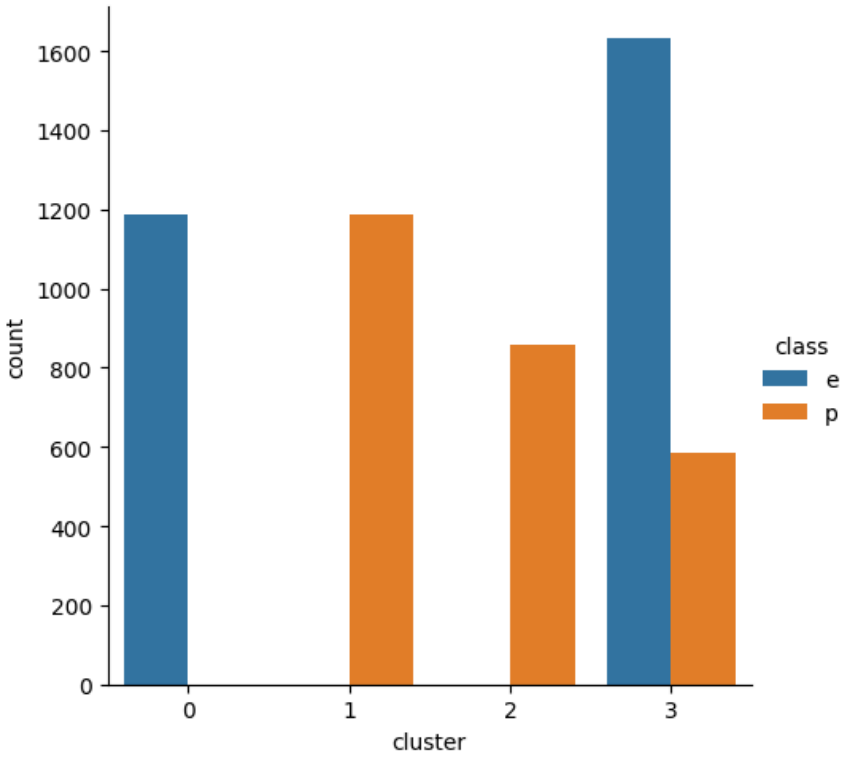
-

8. Clustering no supervisado (KMeans).

- Hemos utilizado el método del codo para ver el número óptimo de clusters, y como se puede ver en el siguiente gráfico, nos da k=4:

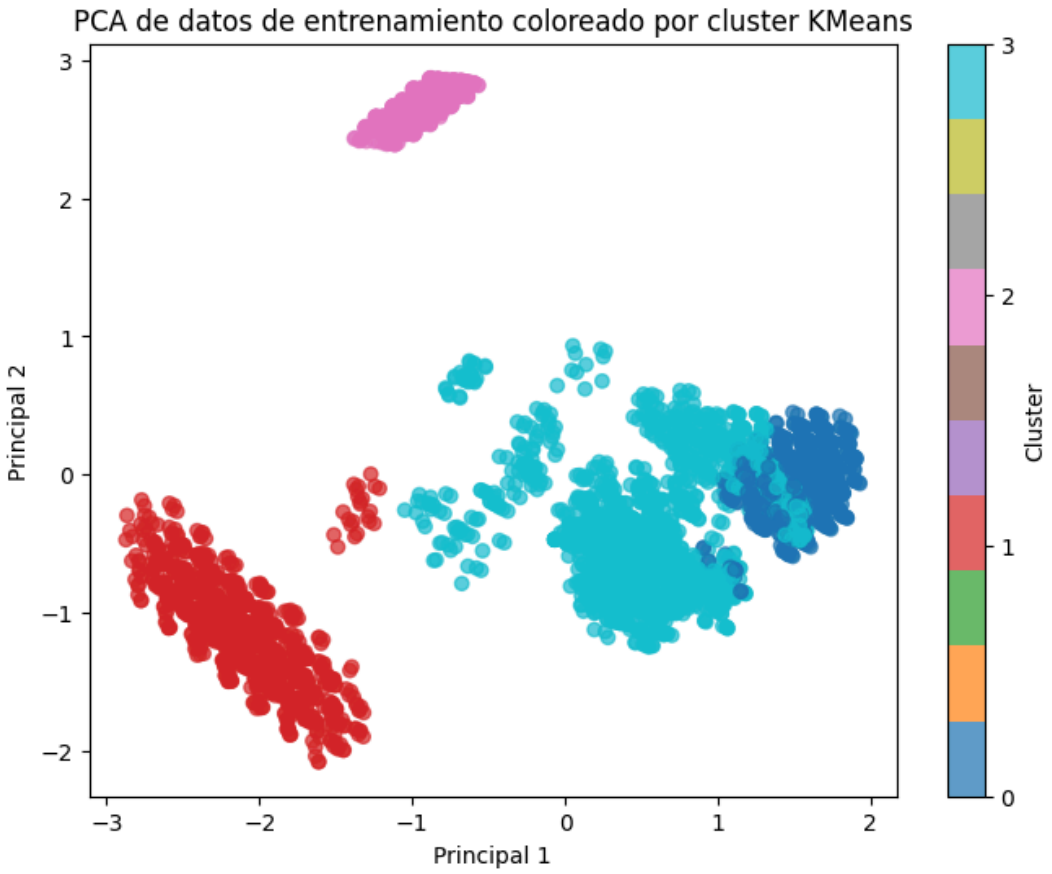


- Luego hemos entrenado con KMeans y hemos pintado los resultados usando catplot:



9. Visualización y comparación.

- Se ha representado un scatterplot como el de antes, de los datos en el espacio PCA, coloreando con los colores propios de KMeans, y observamos los clusters bien definidos , aunque hay algunas zonas de solapamiento, esto nos demuestra que incluso sin etiquetas KMeans puede ayudarnos a identificars grupos naturales en los datos.



10. Conclusiones.

- Este dataset de hongos es bastante separable, tanto con métodos supervisados como con los no supervisados.
- Random Forest logra una precisión muy alta, demostrando la calidad de los datos.
- KMeans, sin usar las etiquetas, también consigue una muy buena agrupación de los datos.
- Si no hubiésemos tenido las etiquetas, el clustering sería una herramienta muy útil para nuestro objetivo de clasificar los hongos ya que los clusters encontrados por KMeans se parecen bastante a las clases reales, eso nos indica que las características de los hongos contienen bastante información para distinguir los tipos.
- Como mejoras a futuro, podríamos añadir ajuste de hiperparámetros o explorar más visualizaciones.

Autores: Nhoeli Salazar / Omar Lengua