DIPLOMATURA EN CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS APLICACIONES

MENTORÍA M03-2023

Descifrando el Universo: apariencia de las galaxias

Directora: Ingrid Vanessa Daza Perilla

Grupo 2: Ailín Asís, Joaquín Gamalerio y Pablo Velez.

3er Entregable Primera Parte: Análisis no Supervisado.

Se exploraron el conjunto de datos obtenidos a partir del **S**loan **D**igital **S**ky **S**urvey (https://skyserver.sdss.org/CasJobs/). El mismo contiene información de 92102 galaxias

1- Preparación de los datos

El dataset original consta de 92.102 registros y 14 variables y no posee valores faltantes.
Las variables son: ‘objID’, 'ra', 'dec', 'modelMag_u', 'modelMag_g', 'modelMag_r', 'modelMag_i',
'modelMag_z', 'petroR90_r', 'z', 'Color' ,'elliptical' , 'spiral' y `'uncertain'`.
El preprocesado usado fue hecho anteriormente en el trabajo 2 de la materia "Análisis Exploratorio y Curación de Datos" , a continuación se resumen paso a paso la limpieza y curación implementada.

1.1- Pre-Procesado.

● Se añade una columna con la clase de cada galaxia, asignando la etiqueta ‘I’ (de irregular) a las
galaxias con etiqueta ‘uncertain’.
● Se setea a la columna ‘objID’ como el id del dataframe.
● Se encontraron 34421 objetos con id repetido. Debido a que nuestro dataset es lo suficientemente grande y para evitar conflictos entre datos, decidimos eliminar los duplicados y el
dataset queda con 57681 datos, lo cual consideramos que, a nuestros fines, es una cantidad suficiente para realizar estadística.

1.1.1- KNN sobre valores atípicos.

● Se imputaron usando KNN las columnas que caen fuera de un intervalo definido por una cota inferior y una cota superior.
● La cota inferior elegida para las magnitudes luminicas aparentes (`'modelMag_u'`, `'modelMag_g'`, `'modelMag_r'`,`'modelMag_i'`,

`'modelMag_z'`) y para el tamaño (`'petroR90_r'`) fue de 0, tambien se filtraron outliers de color.
● Las cotas superiores aplicadas a las magnitudes aparentes se eligieron según el criterio;

-'modelMag_u'=22.0,
-'modelMag_g'=22.2,
-'modelMag_r =22.2,
-'modelMag_i'=21.3,
-'modelMag_z'=20.5

Extraidas de https://classic.sdss.org/dr4/

● Se escalaron los datos y se procedió a realizar la imputación utilizando todas las columnas como input.

1.1.2- Eliminación de outliers.

● Por último se eliminaron los outliers con el criterio Q1*2.5 < x < Q3*2.5 y se creó el dataframe df1 (guardado en csv como "galaxias_curadas"), de 56545 datos, es decir, se eliminaron el 1.97% de datos.

2- Clustering

Se usó k-means para encontrar clusters en el dataset. K-means es un algoritmo no supervisado de Clustering, cuyo objetivo es agrupar en **k** clusters disjuntos a cada dato de nuestro dataset. El número de clusters es un parámetro, y el centroide de cada cluster se puede inicializar de forma aleatoria o elegir manualmente si se tienen conocimientos de dominio que permitan hacerlo (entre otras formas). Luego cada dato es asignado al cluster cuyo centroide está más cerca (se busca minimizar la inercia o la suma de cuadrados de cada cluster). Luego se calcula la media geométrica de cada cluster y es asignada como el nuevo centroide. Con este centroide se repiten los pasos anteriores hasta alcanzar un criterio que puede ser de convergencia (los centroides no varían significativamente), o de máximo de iteraciones. Finalmente se eligen los clústeres que hayan minimizado la suma de cuadrados dentro de cada clúster.

Notar que cada vez que se repite el proceso se pueden obtener diferentes resultados ya que el modelo es susceptible a la ubicación inicial de los centroides, por lo cual la convergencia no está garantizada.

El método de las siluetas permite medir, usando el coeficiente de silueta, que tan fuerte es la identificación de un dato y su cluster con respecto a otros clusters.

Los gráficos de silueta mostrados a continuación, muestran el score (**q**) de silueta de cada punto de cada cluster. Mientras más grande el **q**, mejor será nuestro modelo. Además, mientras más uniforme y larga es cada silueta, mejor se identifican los datos con sus respectivos clusters.

Otra característica a tener en cuenta es que **q** pertenece a un intervalo [-1,1], y cuando es calculado para un punto en particular se pueden identificar los siguientes casos puntuales:

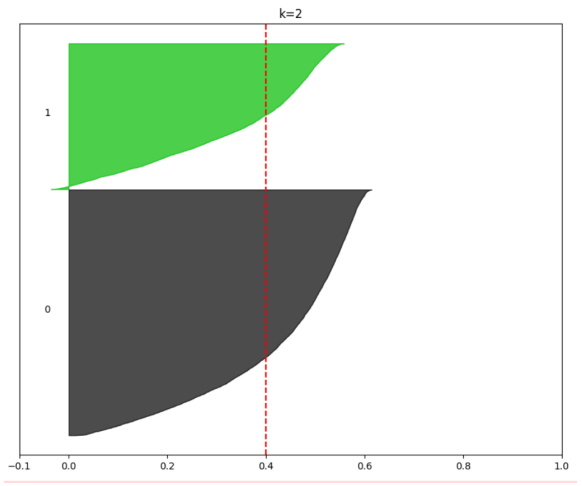
q=1: El punto está fuertemente emparejado con el cluster al que pertenece y nada identificado con su cluster vecino.

q=0: El punto está igual de emparejado con el cluster al que pertenece y con su cluster vecino, pudiéndose dar el caso de que se encuentre en la periferia de ambos clusters.

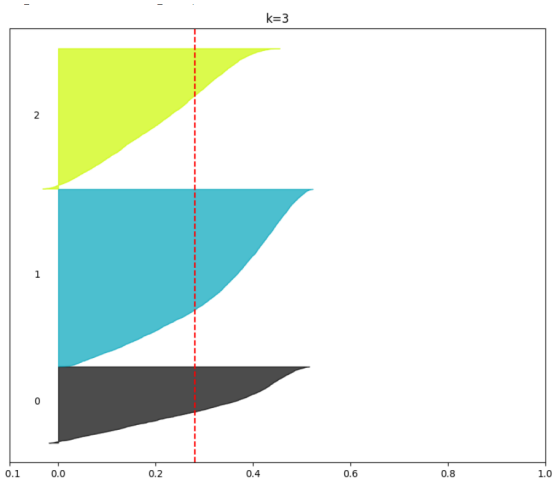
q=-1: El punto está fuertemente desvinculado al cluster al que pertenece, y está fuertemente vinculado a su cluster vecino.

En nuestro caso parece que **k=3** es una buena cantidad de clusters a elegir, con un score de **q=0.2805523577032078**; **k=2** clusters también es una buena opción, con el mejor score de todos **q=0.39953428584614104**, pero no utilizaremos ese **k** porque no dan demasiada información. Se observa que a partir de **k=4** el score disminuye muy lentamente y prácticamente se mantiene constante, eso significa que si bien aumenta **k**, no mejora la métrica de nuestro modelo, es decir el score de silueta no crece, por lo tanto utilizaremos **k=3**.

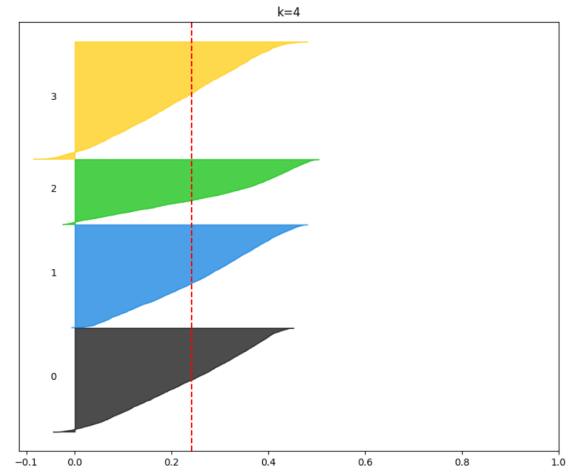
A continuación se presentan los gráficos de siluetas encontrados para diferentes **k**



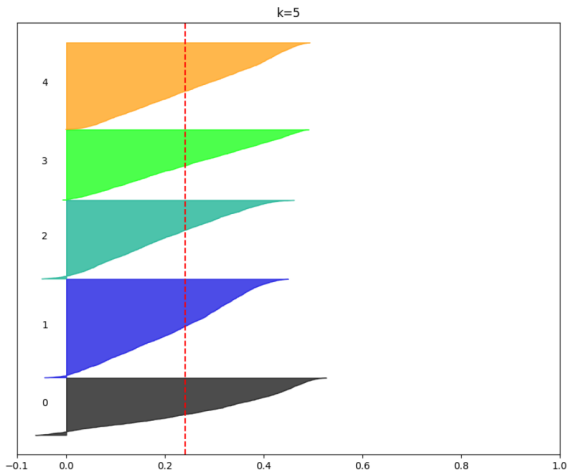
q=0.39953428584614104



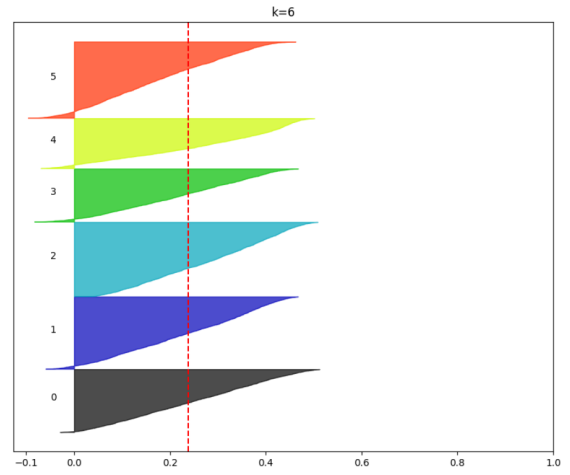
q=0.2805523577032078



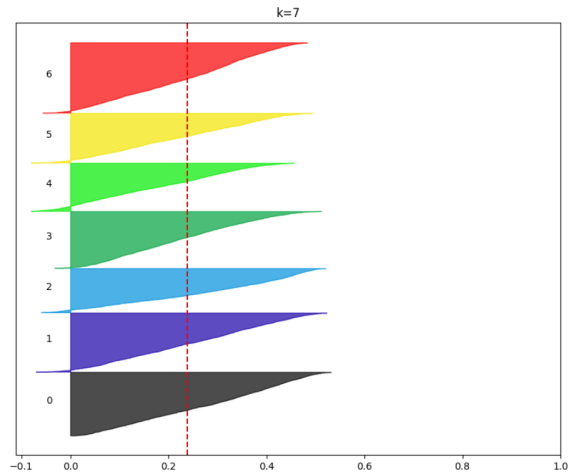
q=0.2412817400535265



q=0.24074437042400673



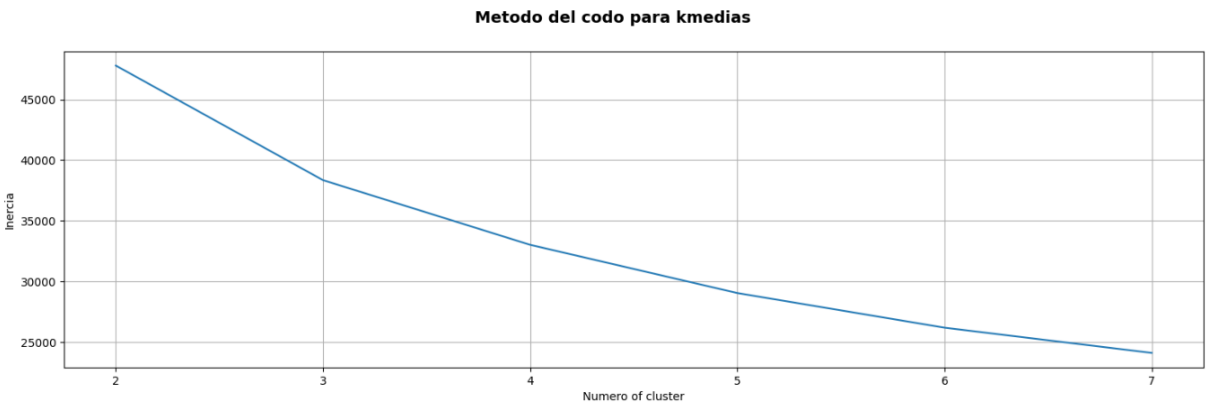
q=0.23801283443608284



q=0.23841961066286071

Otra manera de elegir k es utilizando el método del codo, el cual consiste en graficar la inercia del modelo con respecto a k. En el punto en el cual la curva el ángulo más cerrado se identifica un codo, ese punto o codo corresponde con el k mas optimo, a partir del cual el modelo no presenta (idealmente) una mejora significativa de inercia.

A continuación se presenta el diagrama de codo encontrado.



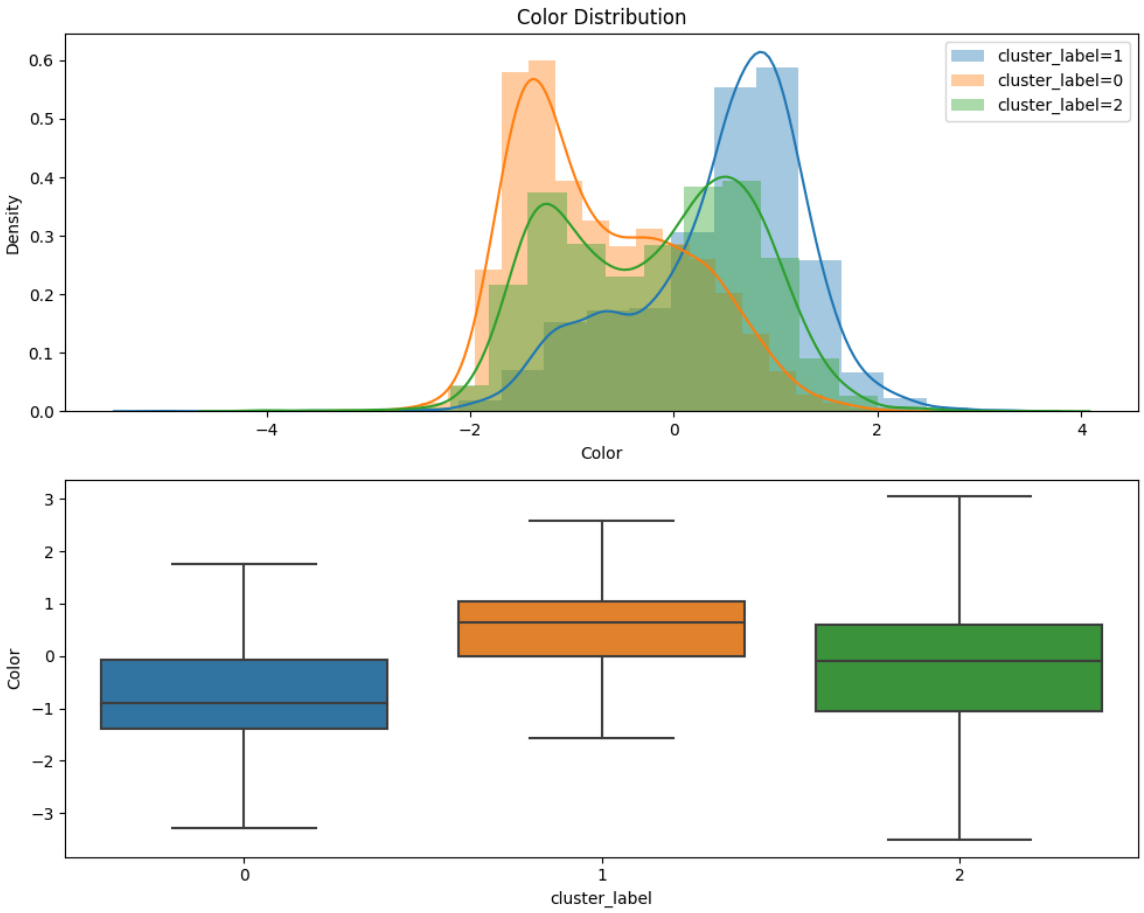
El codo se identifica en **k=3**, en este punto la curva presenta el ángulo con mayor pronunciación, aunque no resalta demasiado entre los otros **k**.

3- Visualización según cluster

3.1- Color

Con respecto al grafico de color, se observan diferencias significativas entre los clusters 0 y 1. También se observa el cluster 2 con bimodalidad, esto puede sugerir que a su vez el cluster 2 puede estar conformado por dos subclusters.

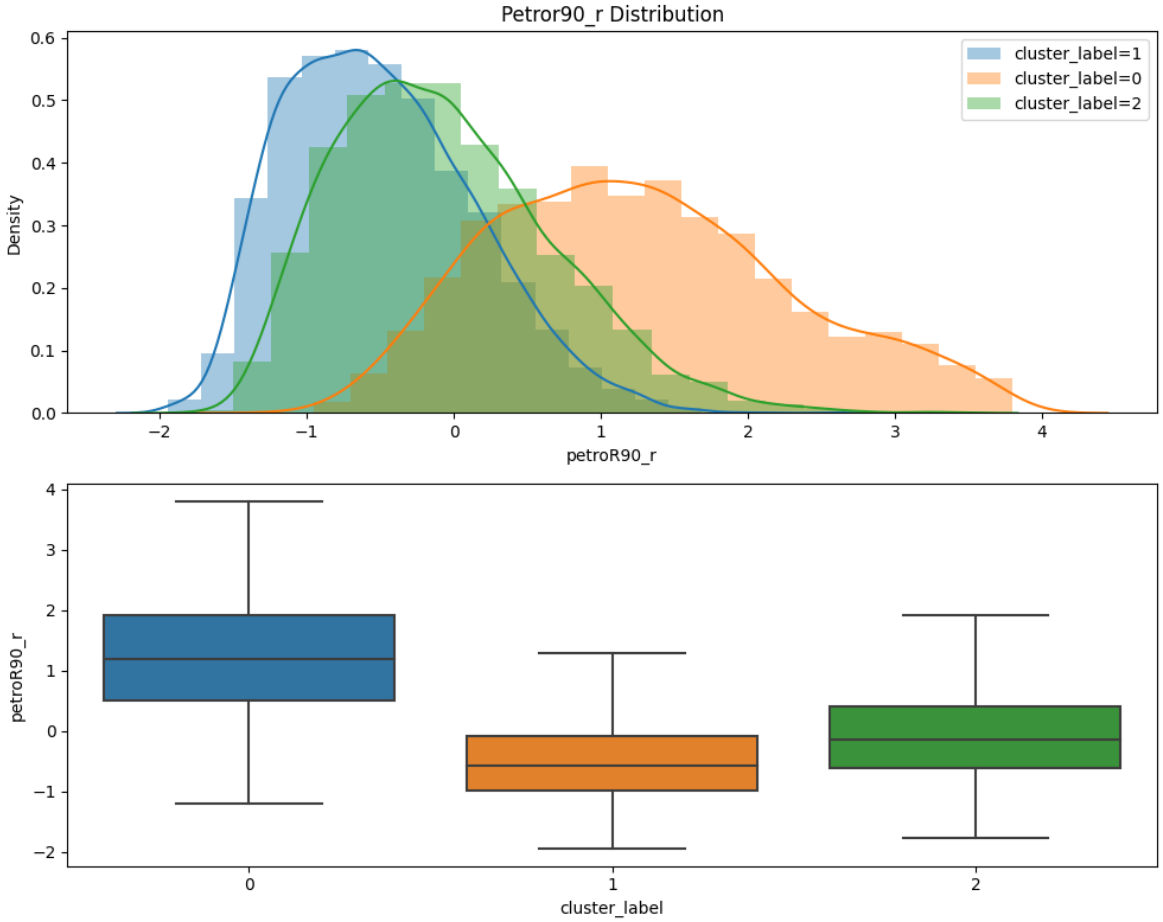
El gráfico es muy similar al obtenido cuando se visualizó la distribución de la clase color distinguiendo por clase de galaxia, hecho en el entregable 1 de la materia.



3.2 - PetroR90_r

Los boxplot muestran nuevamente una clara separacion entre los clusters 0 y 1, intuyendo que el cluster 0 corresponde con las galaxias mas grandes y con indice de color mas rojizo, mientras que el cluster 1 se corresponde con galaxias chicas y azuladas

El cluster 2 contiene galaxias con valores quizás más promedio y más desparramados.



3.3- Magnitudes Aparentes

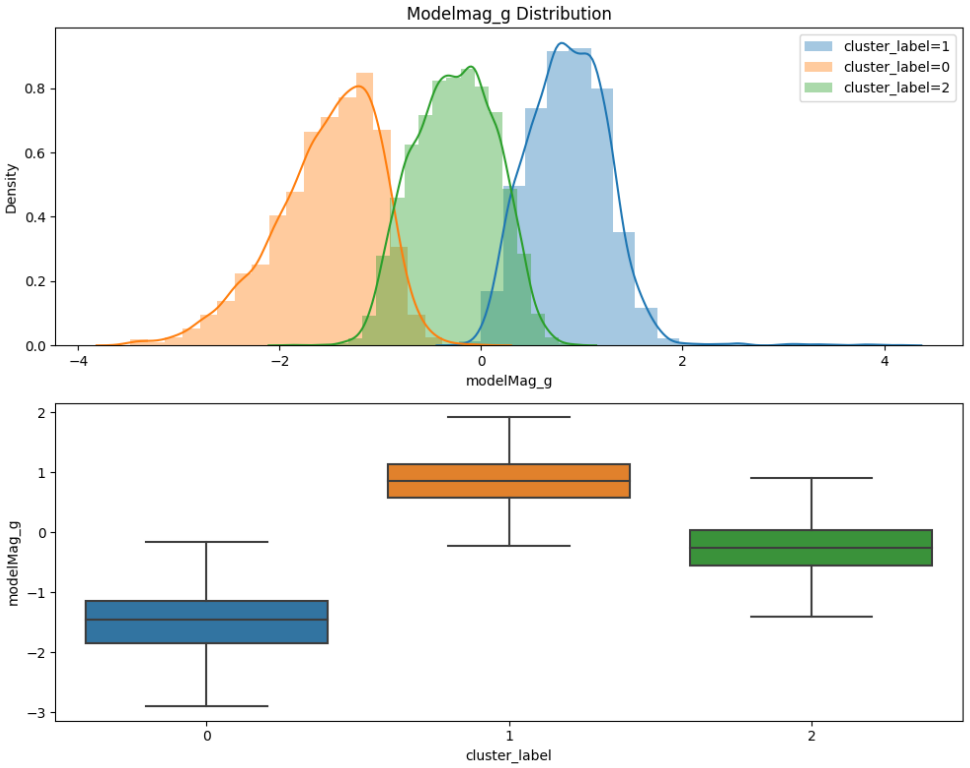
Los boxplot de todas las magnitudes muestran clara separación de clusters, con el cluster 0 correspondiendo a las galaxias más débiles en su correspondiente magnitud, seguidas por el cluster 2 que nuevamente parece tener un valor medio, y por último el cluster 1 con los valores más altos de magnitud, siguiendo nuevamente la tendencia.

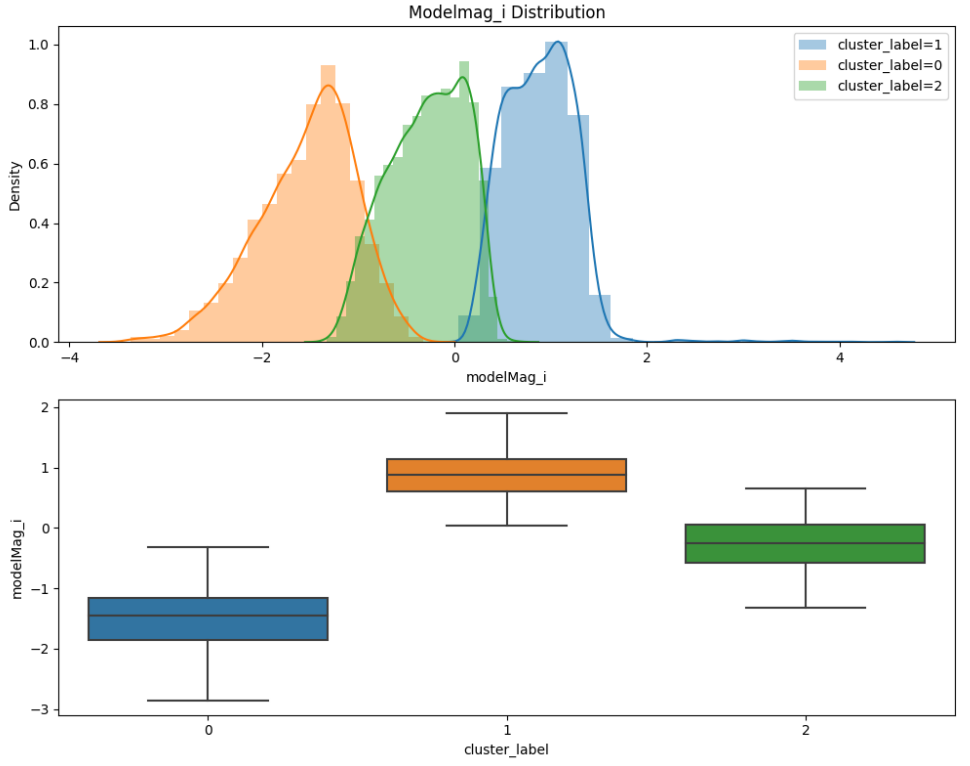
El contenido de cada cluster parece resumirse como sigue:

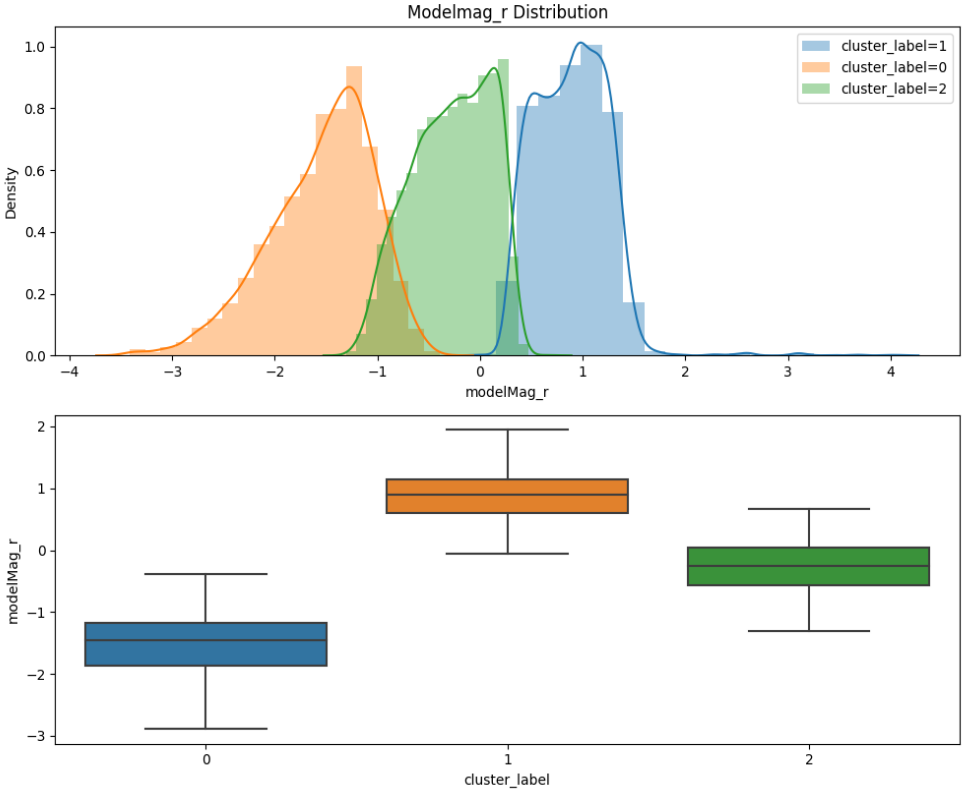
Cluster 0: Galaxias chicas, azuladas y poco brillantes.

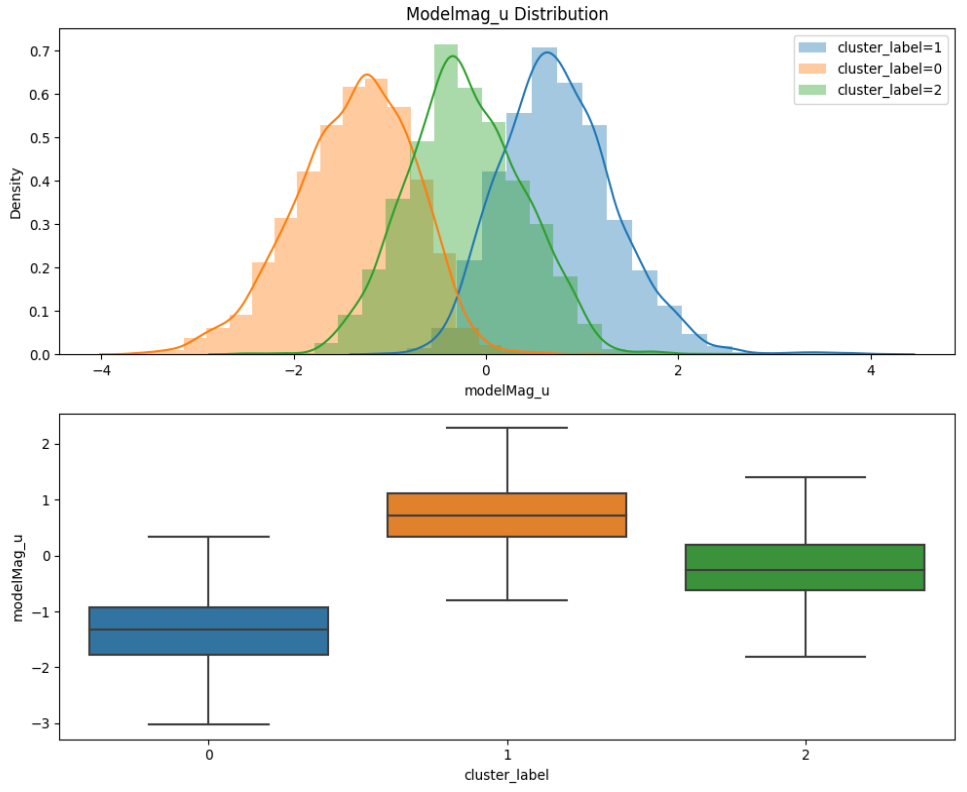
Cluster 1: Galaxias Grandes, rojizas y brillantes.

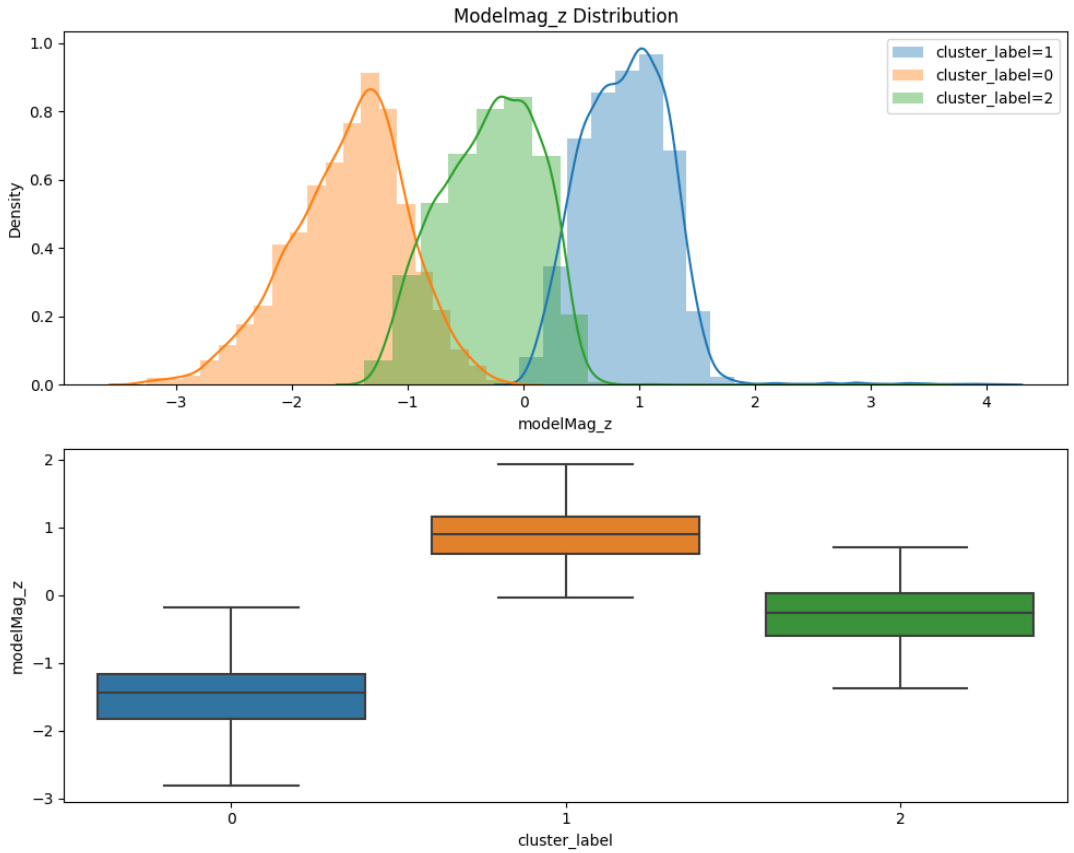
Cluster 2: Galaxias que no entran en los clusters anteriores y tienen valores más bien promedio o valores no-extremos.


Modelmag_g Distribution
cluster_label=1
cluster_label=0
cluster_label=2
Density
modelMag_g
modelMag_g
cluster_label


Modelmag_i Distribution
cluster_label=1
cluster_label=0
cluster_label=2
Density
modelMag_i
modelMag_i
cluster_label


Modelmag_r Distribution
cluster_label=1
cluster_label=0
cluster_label=2
Density
modelMag_r
modelMag_r
cluster_label


Modelmag_u Distribution
cluster_label=1
cluster_label=0
cluster_label=2
Density
modelMag_u
modelMag_u
cluster_label



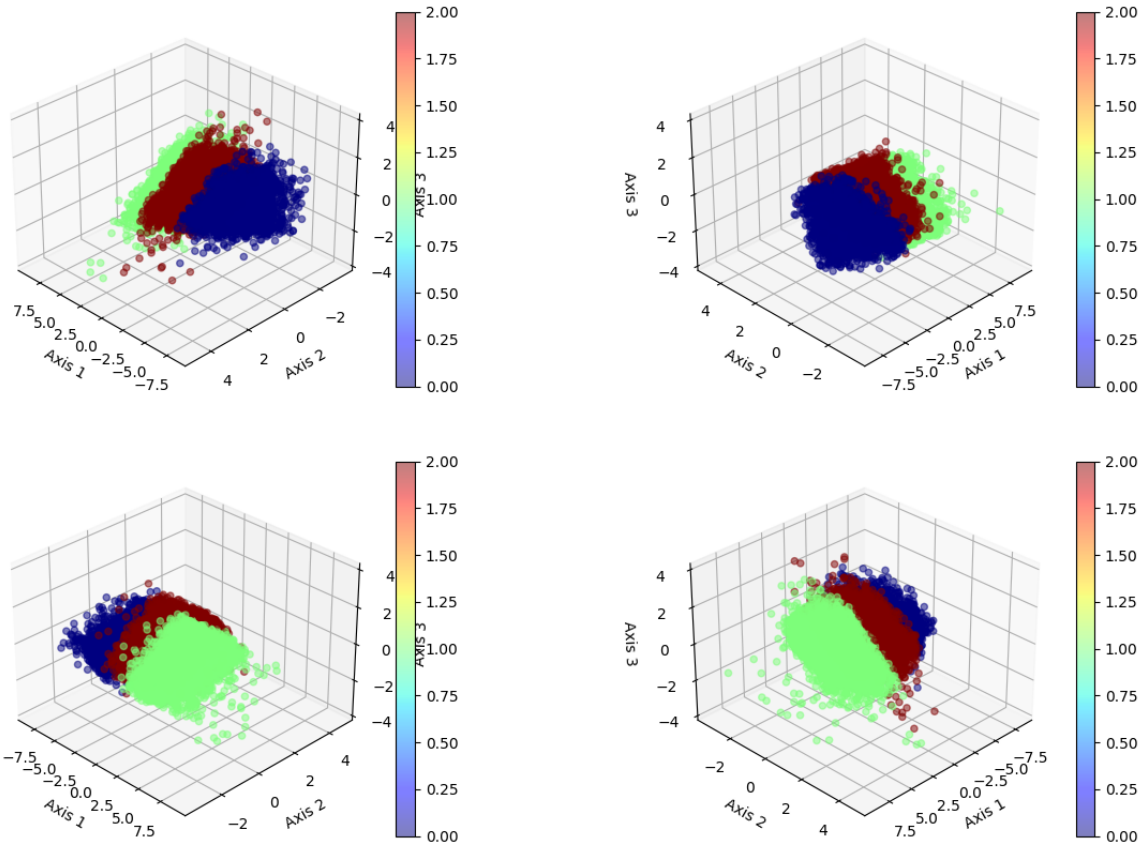
4- Embeddings

Los "embeddings" en el aprendizaje no supervisado son representaciones numéricas de datos que se obtienen al reducir la complejidad de los datos originales y transformarlos en un espacio de características más simples. En el caso de PCA el espacio de características es una proyección lineal del espacio original sobre un espacio de menor dimensión que maximiza la varianza de los datos. T-SNE a diferencia de PCA es método de reducción de dimensionalidad no lineal, que busca respetar lo más posible las distancias entre puntos, no la varianza.

4.1- PCA

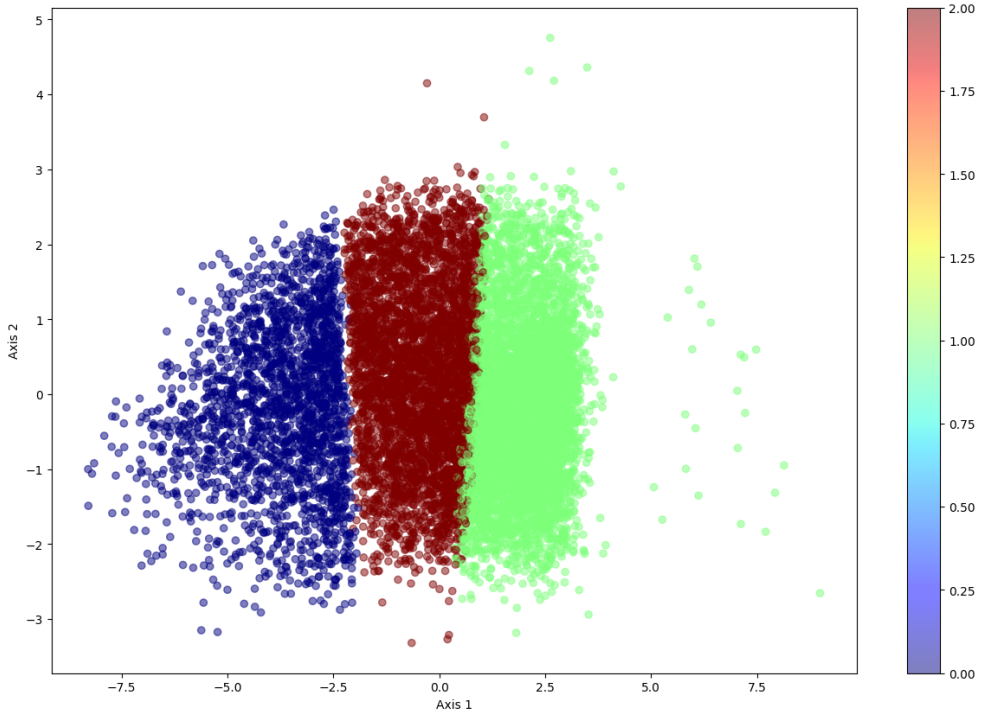
3 componentes

Usando 3 componentes principales se observa la misma tendencia mencionada anteriormente, el cluster 2 parece estar entre medio del cluster 0 y 1. La clusterización parece ser arbitraria vista desde esta proyección, ya que si los clusters no estuvieran coloreados, sería difícil distinguirlos.



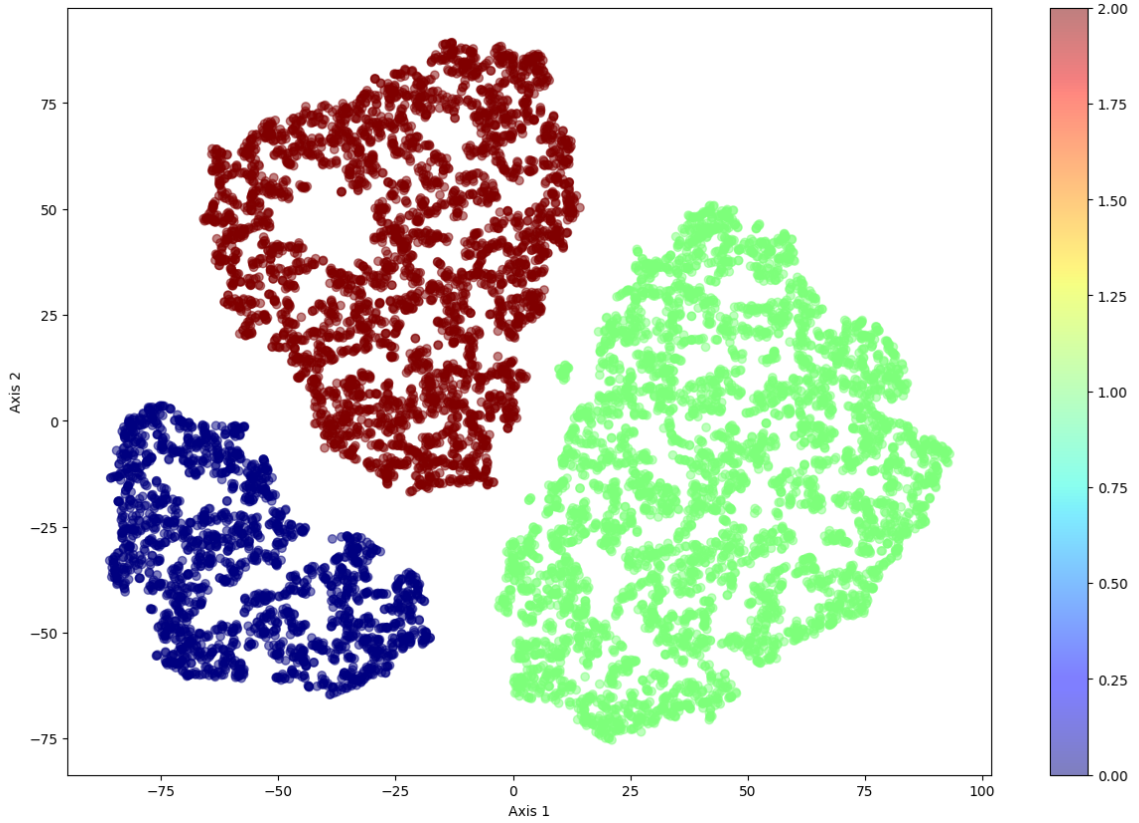
2 componentes

Usando 2 componentes principales se observa nuevamente la tendencia, solo que esta vez se distingue una separación clara entre los cluster 0 y 2.



4.2- T-SNE

En T-SNE se observa una clara separación de clusters, concluyendo que esta representación es la que mejor caracteriza los clusters encontrados y muestra incluso estructura interna en cada cluster, lo que sugiere que se puede realizar más análisis en cada cluster por separado, existiendo así una clusterización jerárquica



5- Conclusiones

Se realizó un aprendizaje no supervisado sobre el *dataset*: "galaxias_curadas", realizando previamente una tarea de visualización y limpieza sobre el dataset original: "galaxias_1.csv".

Se realizó un análisis del score de silueta para determinar el k óptimo de clusterización mediante **k-means**, junto con un análisis de la inercia de cada k conocido como "método del codo". Se encontró, empleando los dos métodos anteriores, que k=3 es el número óptimo de clusters que caracterizan al dataset.

Luego de realizada la clusterización se analizó cada variable numérica por separado distinguiendo por cluster, encontrándose una tendencia que puede resumirse como sigue

Cluster 0: Galaxias chicas, azuladas y poco brillantes.

Cluster 1: Galaxias Grandes, rojizas y brillantes.

Cluster 2: Galaxias que no entran en los clusters anteriores y tienen valores más bien promedio o valores no-extremos.

Por último se realizó una reducción de dimensionalidad con el objetivo de observar las características de cada cluster, encontrándose que la visualización más clara es t-distributed Stochastic Neighbor Embedding (t-sne).