Interpretación del modelo de agrupación

Una vez obtenidos los grupos, ¿qué interpretación podemos darles? Un aspecto importante en machine learning es la interpretación de los modelos que se generan con sus técnicas, sobre todo cuando se requiere utilizarlos en procesos de toma de decisiones. En el caso de la agrupación, con algoritmos basados en centroides, una vía para interpretar los grupos es examinar los prototipos. Veamos el siguiente ejemplo:

El conjunto de datos está relacionado con las compras anuales que han realizado los clientes de una empresa que vende productos comestibles[1]. El objetivo es descubrir patrones de compra en estos clientes para ofrecer un servicio más personalizado. Una vez aplicado el algoritmo K-medias con un valor de K = 3 (el cual fue determinado aplicando el método de Elbow), se obtuvieron los prototipos que se muestran en la Tabla 1. ¿Qué podemos decir a partir de ellos? Recuerda que el prototipo es el punto medio de los datos de un grupo, por lo que identificando los valores de variables que son diferentes entre los grupos se podría hacer un primer análisis.

Tabla. 1. Prototipos para el caso de estudio.

Variable	Grupo 1	Grupo 2	Grupo 3
Canal	1	2	1
Región	3	3	1
Frescos	14224,64	8892,96	11261,97
Lácteos	3501,31	10888,18	3384,15
Comestibles	3934,80	16606,26	4161,47
Congelados	3894,33	1565,46	3400,78
Limpieza - papel	779,56	7410,46	898,91

[1] https://archive.ics.uci.edu/ml/datasets/wholesale+customers

Variable	Grupo 1	Grupo 2	Grupo 3
Delicatesen	1523,71	1766,79	1120,82

Con base en el diccionario, el cual nos indica: Canal (1: HORECA (Hotel / Restaurante / Café), 2: canal de venta al menor) y Región (1: Lisboa, 2: Oporto, 3: Otros), es posible extraer la siguiente interpretación:

- Grupo 1: están asociados al canal HORECA. Son los clientes que gastan más en productos frescos. Gastan poco en productos de limpieza y papel. Provienen de otras regiones.
- Grupo 2: son minoristas y los que más gastan en productos lácteos, comestibles y productos de limpieza. También son los que menos gastan en congelados.
- Grupo 3: asociados al canal HORECA. Son también grandes compradores de productos frescos, pero provienen de la región de Lisboa. Gastan poco en productos de limpieza.

Otra vía para lograr una interpretación y apoyar el análisis de resultados es hacer un perfilamiento de los grupos. Por ejemplo, se pueden utilizar gráficos, como los histogramas y los diagramas de caja, para visualizar las distribuciones de las variables dentro de cada grupo e identificar aquellas que son importantes para la formación de estos.

Veamos el siguiente caso utilizando el conjunto de datos "Healthy Lifestyle Cities Report 2021"[2], el cual incluye diversas variables relacionadas con la calidad del aire, áreas verdes, calidad del agua potable, disponibilidad de alimentos saludables, accesibilidad a gimnasios y centros de fitness, prevalencia de enfermedades, índice de obesidad e índice de felicidad, entre otros indicadores, Estos factores se combinan para generar una puntuación general que determina el ranking de 43 ciudades en función de su adopción de estilos de vida saludables. El análisis de este conjunto de datos es útil para comprender cómo las políticas públicas, la infraestructura y los servicios en las ciudades pueden influir en la salud y el bienestar de sus residentes, así como para determinar qué factores son más relevantes para la adopción de estilos de vida saludables o identificar estrategias de mejora para ciudades con calificaciones inferiores en términos de promoción de la salud.

[2] https://www.kaggle.com/datasets/prasertk/healthy-lifestyle-cities-report-2021

Al aplicar un algoritmo de agrupación, siguiendo la metodología recomendada, se obtuvieron cuatro grupos. En la Fig. 2 se muestran los diagramas de caja de algunas de las variables para cada uno de estos.

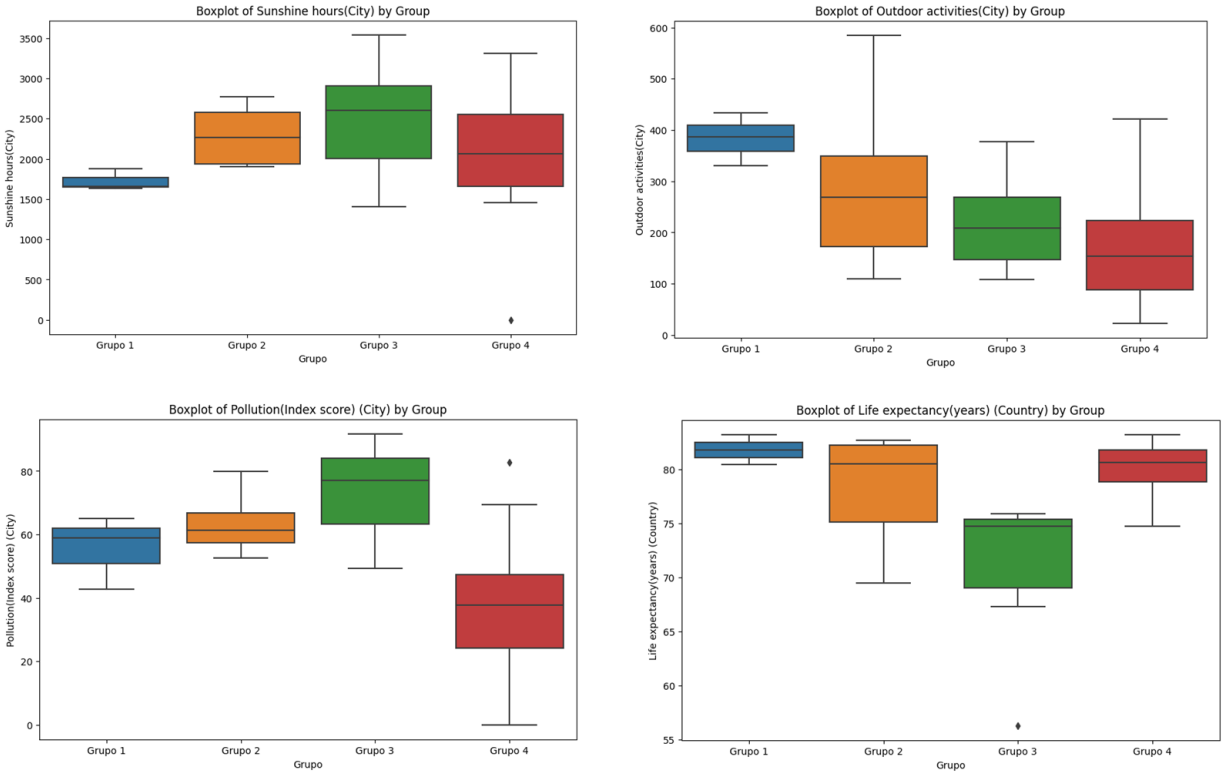

Fig. 2. Diagramas de caja (boxplot) de algunas variables sobre cada grupo.

¿Qué podemos decir de cada grupo? Por ejemplo, en las ciudades del grupo 1, con mejor expectativa de vida, los ciudadanos dedican más tiempo a actividades al aire libre, a pesar de que disponen de menos horas de sol durante el año. Por supuesto, un análisis completo debería incluir todas las variables[3].

[3] Una manera de comunicar los resultados de este tipo de proyectos a los usuarios finales es a través de la construcción de tableros de control o *dashboard*. En este enlace puedes ver un ejemplo utilizando Power BI: https://blog.enterprisedna.co/cluster-analysis-visualization-techniques-in-power-bi/
Una herramienta de visualización para la interpretación de los grupos la encuentras en: "Clustrophile 2: Guided Visual Clustering Analysis" (https://arxiv.org/pdf/1804.03048.pdf)

Otro camino para la interpretación es aplicar un algoritmo supervisado interpretable, como los basados en árboles de decisión. Para esto, una vez que se tienen los grupos, se asigna una etiqueta a cada uno de ellos. Todos los datos tendrán entonces una "clase" que será la etiqueta del grupo al cual pertenecen. Ya con el conjunto de datos anotado es posible entonces aplicar el algoritmo de aprendizaje. En este contexto, estos algoritmos supervisados son utilizados como una herramienta explicativa, no predictiva. Estos dos últimos métodos resultarían más adecuados para los algoritmos que no se basan en prototipos.

Ten en cuenta que, en general, la interpretación de un modelo de agrupación puede ser un proceso complejo y depende en gran medida de las características y el propósito de los datos. La combinación de varias técnicas y herramientas de análisis es fundamental para llegar a conclusiones precisas y significativas.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.