Machine Learning y políticas públicas-2024

Trabajo práctico N°2

Aprendizaje no supervisado: Clustering

Alumna

Fernández, Silvia silvia2484@gmail.com

Introducción

En el siguiente trabajo práctico se propone implementar un modelo no supervisado utilizando el conjunto de datos Mall Customer. Este es un conjunto de datos comúnmente utilizado para ejercicios y proyectos de Machine Learning, especialmente en el contexto del clustering y la segmentación de clientes.
A partir del modelo no supervisado proporcionado en el notebook del módulo 4, se propone analizar y verificar si el número de clusters es el adecuado al conjunto.

Desarrollo

En un principio el dataset consta de 200 observaciones y 5 variables, las cuales son:

- **Customer ID:** Un identificador único para cada cliente.
- **Gender**: El género del cliente: Male, Female
- **Age**: La edad del cliente.
- **Annual income (K$):** El ingreso anual aproximado del cliente, expresado en miles de dólares.
- **Spending Score (1-100):** El scoring asignado por el centro comercial en función del comportamiento del cliente y la naturaleza del gasto. Este puntaje o scoring está en el rango de 1 a 100.

. Definimos el modelo de la siguiente manera:

- Selección de las características relevantes para el clustering, Annual Income (k$), Spending Score
- Definir el modelo k-means con 5 clusters y ajustar el modelo a los datos
- Obtener las etiquetas predichas
- Agregar los clusters al Data Frame original
- Visualizar los clusters con un scatter plot las divisiones, marcando cada una con un color.

Probamos y analizamos nuestro modelo de la siguiente manera:

- Identificamos los ingresos y gastos por género
- Utilizamos el Método elbow y silhouette para identificar el número óptimo de clusters
- Visualizamos los gráficos el número de cluster definido
- Definimos el tipo de cliente según el cálculo del centroide en cada cluster
- Generamos nuevas observaciones
- Selección de las características relevantes para el clustering, Annual Income (k$), Spending Score.
- Predecir los clusters para las nuevas observaciones
- Agregar los clusters a las nuevas observaciones
- Calcular estadísticas descriptivas para cada cluster tanto en el dataset original cómo en las nuevas observaciones

Resultados

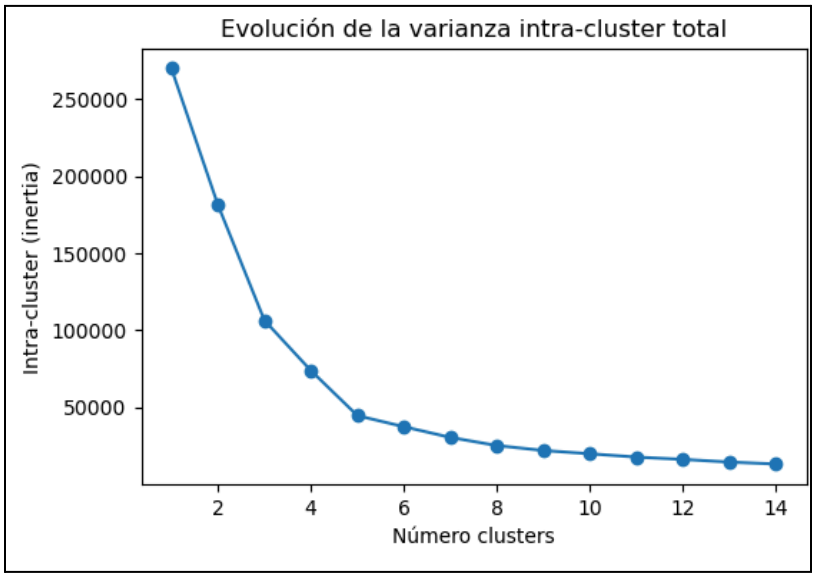
Head del dataset Mall Customer y su correspondiente cluster

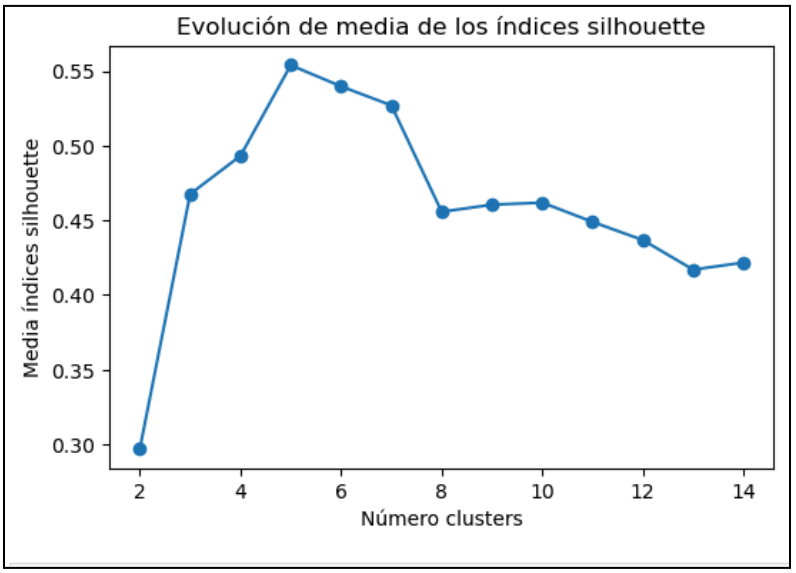
	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)		Cluster
0	1	Male	19	15	39	0	0
1	2	Male	21	15	81	1	2
2	3	Female	20	16	6	2	0
3	4	Female	23	16	77	3	2
4	5	Female	31	17	40	4	0

Ingresos por género

Genre	Annual Income (k$)	Spending Score (1-100)
Female	59.250000	51.526786
Male	62.227273	48.511364

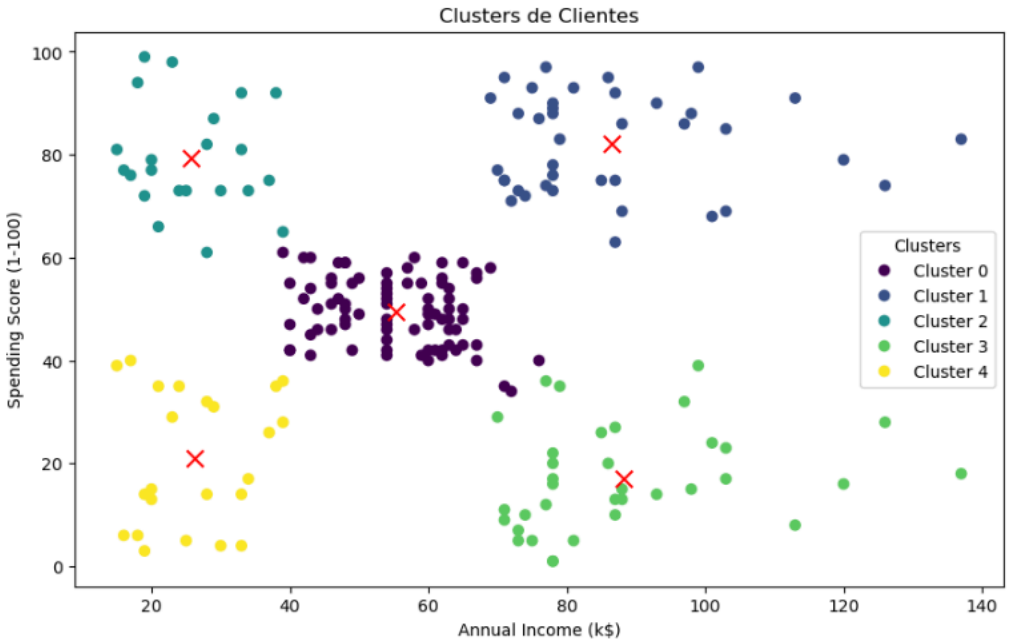
Método elbow y silhouette para identificar el número óptimo de clusters



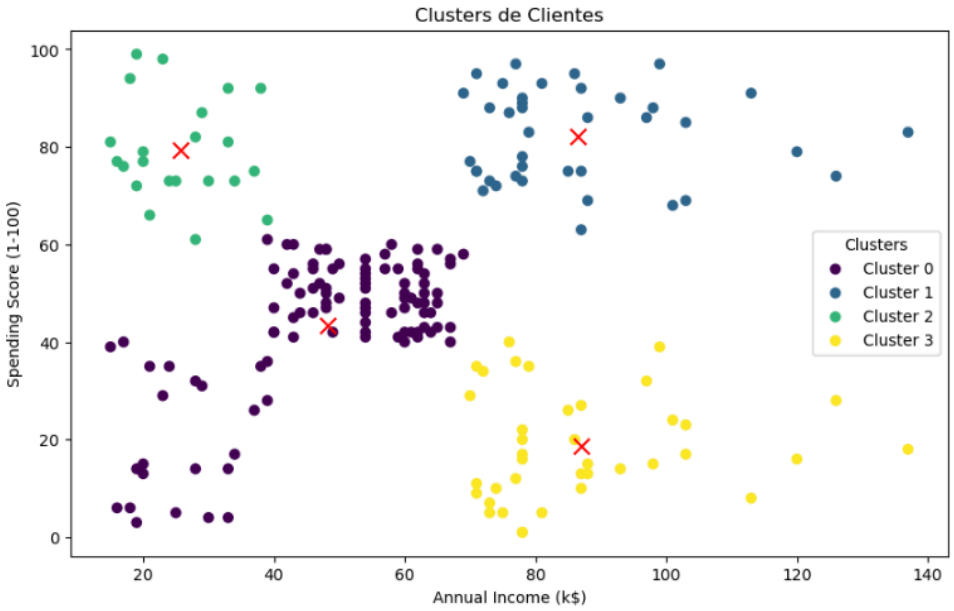


Según estos gráficos con un K = 4 la varianza intra-cluster comienza estabilizarse y en el índice silhouettes, también es el adecuado y más estable para los datos. A continuación graficamos ambos para analizar sus ajustes

Gráficos con K = 5 y K = 4.



Plot con K=5



Plot con k = 4

Cálculo del centroide de cada cluster según ingresos y gastos (K=4)

Cluster	Ingreso	Gasto	Tipo de cliente
0	48.16831683	43.3960396	En general sus ingresos son medios y sus gastos son bajos
1	86.53846154	82.12820513	En general sus ingresos son altos y sus gastos son bajos
2	25.72727273	79.36363636	En general sus ingresos son bajos y sus gastos muy altos
3	87	18.63157895	En general sus ingresos son altos y sus gastos muy bajos

Métricas del dataset original

Cluster	Annual Income (k$)	Spending Score (1-100)
0	55.296296	49.518519
1	86.538462	82.128205
2	25.727273	79.363636
3	88.200000	17.114286
4	26.304348	20.913043

	Annual Income (k$)	Spending Score (1-100)

Head nuevas observaciones y su cluster correspondiente

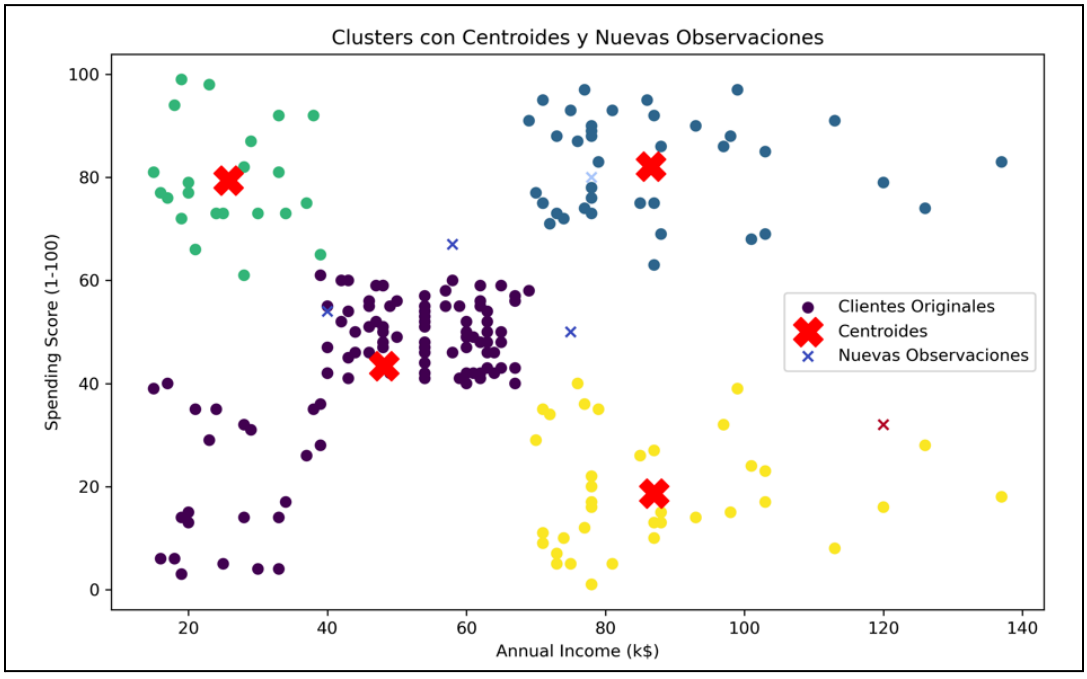
	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	201	Male	23	58	67
1	202	Female	45	120	32
2	203	Female	31	78	80
3	204	Male	22	40	54
4	205	Female	41	75	50

	Cluster
0	0
1	3
2	1
3	0
4	0

Métricas de las nuevas observaciones

Cluster	Annual Income (k$)	Spending Score (1-100)
0	57.666667	57.0
1	78.000000	80.0
3	120.000000	32.0

Gráfico con los centroides calculados, las nuevas observaciones y los clientes originales



Conclusión

En primer lugar, se considera que, en este caso, el género de los clientes no es una variable determinante en la segmentación. Los ingresos y gastos de los mismos son bastante similares, esto sugiere que ser masculino o femenino no definiría una separación clara en los distintos grupos.
Respecto a los gráficos obtenidos de la varianza intra.cluster y silhouette, se puede concluir que un valor K=4 es el más adecuado para la segmentación de la población. Este número permite una división significativa de los datos, creando grupos bien definidos en función de las características de ingresos y gastos.
A partir de los centroides obtenidos, pudimos identificar los diferentes tipos de clientes en función de sus ingresos y gastos. Esto facilitó la definición precisa de los perfiles de los mismos dentro de cada cluster.
Al aplicar el modelo en nuevas observaciones, este ha demostrado ser capaz de separar correctamente los datos, clasificando a las nuevas observaciones en el clúster adecuado. Se concluye entonces, que el modelo es efectivo y puede ser utilizado de manera fiable para segmentar a la población en futuras aplicaciones.

Anexo

Fernandez, Silvia, Jupyter Notebook
AANS - Clustering - TP-Silvia-Fernandez