EJERCICIO DE CLUSTER

Equipo:

- **Luna Lopez Natalia**
- **Rivero Ronquillo Omar Imanol**
- **Zamorano Cruz Juan Raymundo**

Introducción

Como sabemos el clustering o cluster es una técnica utilizada en minería de datos para identificar de forma automática agrupaciones de elementos de acuerdo a una medida de similitud entre ellos. Esta técnica también se conoce como segmentación.

En este ejemplo utilizaremos el clustering con el nodo K-means, el cual realiza una agrupación nítida que asigna un vector de datos a exactamente un grupo, durante la explicación del ejercicio podremos visualizar su funcionamiento específico y los resultados que arroja.

Objetivo

Crear un modelo de Cluster en Knime utilizando K-means y agrupación jerárquica.

Información sobre el conjunto de datos

Customer Personality Analysis

Autor: Akash Patel

El análisis de la personalidad del cliente es un análisis detallado de los clientes ideales de una empresa. Ayuda a una empresa a comprender mejor a sus clientes y le facilita la modificación de los productos en función de las necesidades, los comportamientos y las preocupaciones específicas de los distintos tipos de clientes.

Diccionario de Datos (Cliente)

Nombre	Significado	Tipo
ID	Identificador de cliente	Numérico
Year_Birth	Año de nacimiento	Numérico
Education	Grado educativo	Categórico
Marital_Status	Estado civil	Categórico
Income	Ingreso anual	Numérico
Kidhome	Número de niños en casa	Numérico
Teenhome	Número de adolescentes	Numérico
Dt_Customer	Fecha de inscripción del cliente	Fecha

Diccionario de Datos (Productos)

Nombre	Significado	Tipo
MntWines	Gasto en vino en los últimos 2 años	Numérico
MntFruits	Gasto en frutas en los últimos 2 años	Numérico
MntMeatProducts	Gasto en carnes en los últimos 2 años	Numérico
MntFishProducts	Gasto en pescado en los últimos 2 años	Numérico
MntSweetProducts	Gasto en dulcería en los últimos 2 años	Numérico
MntGoldProds	Gasto en oro en los últimos 2 años	Numérico

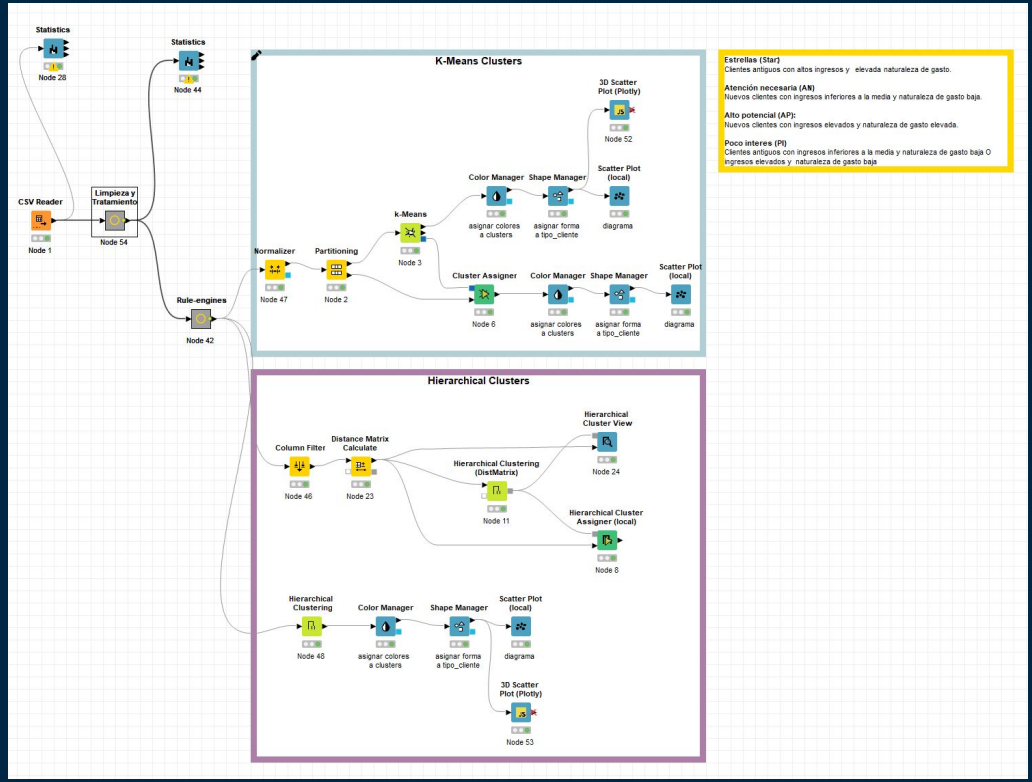
DESARROLLO DEL EJERCICIO

Análisis

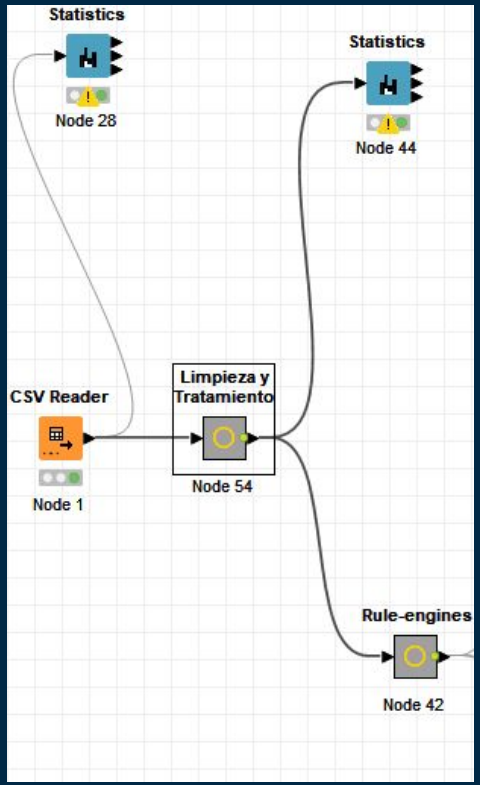
La parte más importante de un análisis de la personalidad del cliente es obtener las respuestas a preguntas como:

- Lo que la gente dice de su producto: lo que da la actitud de los clientes hacia el producto.
- Lo que la gente hace: que revela lo que la gente hace más que lo que dice sobre su producto.

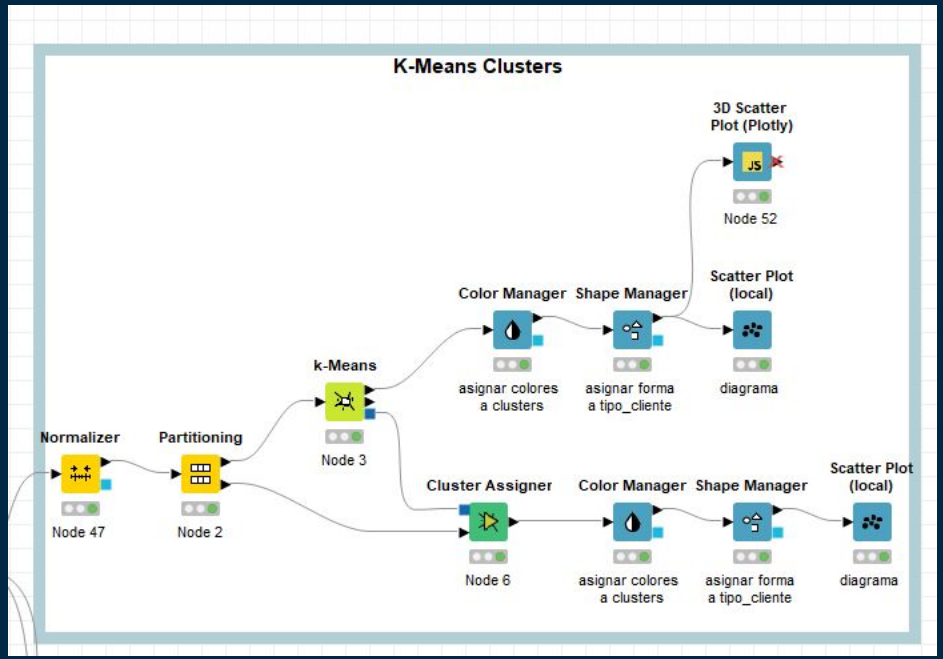
Workflow


Statistics
Node 28
Statistics
Node 44
CSV Reader
Node 1
Limpieza y Tratamiento
Node 54
Rule-engines
Node 42
K-Means Clusters
Normalizer
Node 47
Partitioning
Node 2
k-Means
Node 3
Color Manager
asignar colores a clusters
Shape Manager
asignar forma a tipo_cliente
Scatter Plot (local)
diagrama
3D Scatter Plot (Plotly)
Node 52
Cluster Assigner
Node 6
Color Manager
asignar colores a clusters
Shape Manager
asignar forma a tipo_cliente
Scatter Plot (local)
diagrama
Hierarchical Clusters
Column Filter
Node 46
Distance Matrix Calculate
Node 23
Hierarchical Clustering (DistMatrix)
Node 11
Hierarchical Cluster View
Node 24
Hierarchical Cluster Assigner (local)
Node 8
Hierarchical Clustering
Node 48
Color Manager
asignar colores a clusters
Shape Manager
asignar forma a tipo_cliente
Scatter Plot (local)
diagrama
3D Scatter Plot (Plotly)
Node 53
Estrellas (Star)
Clientes antiguos con altos ingresos y elevada naturaleza de gasto.
Atención necesaria (AN)
Nuevos clientes con ingresos inferiores a la media y naturaleza de gasto baja.
Alto potencial (AP):
Nuevos clientes con ingresos elevados y naturaleza de gasto elevada.
Poco interes (PI)
Clientes antiguos con ingresos inferiores a la media y naturaleza de gasto baja O ingresos elevados y naturaleza de gasto baja

1a Sección – Subida y tratamiento de datos

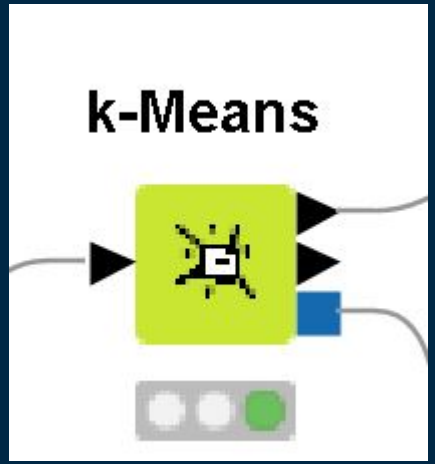


2a Sección - Generación de Clusters con K-Means



El algoritmo de agrupación utiliza la distancia euclidiana en los atributos seleccionados. Los datos no son normalizados por el nodo (si es necesario, debería considerar utilizar el "Normalizador" como un paso de preprocesamiento).

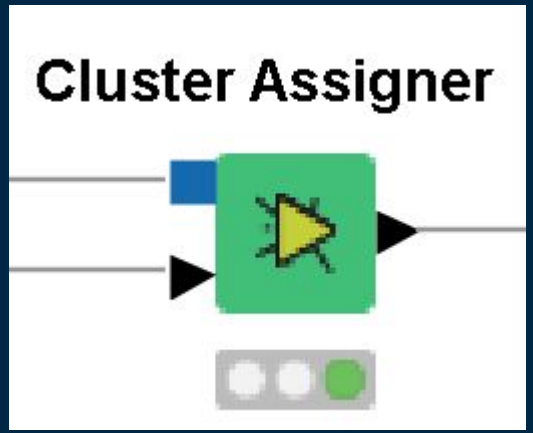
3a Sección - Nodos



Este nodo emite los centros de cluster para un número predefinido de clusters (no hay un número dinámico de clusters). K-means realiza un clustering crispado que asigna un vector de datos exactamente a un cluster. El algoritmo termina cuando las asignaciones de los clusters ya no cambian.

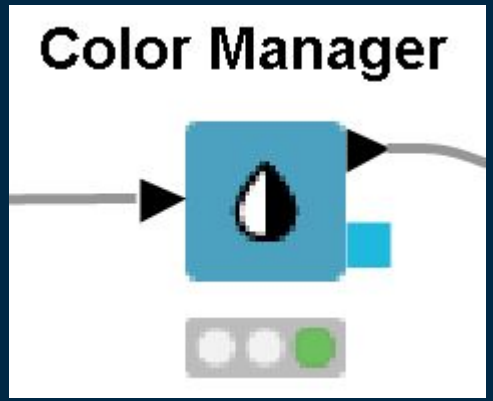
El algoritmo de agrupación utiliza la distancia euclidiana en los atributos seleccionados. Los datos no son normalizados por el nodo (si es necesario, debería considerar utilizar el "Normalizador" como un paso de preprocesamiento)

3a Sección - Nodos



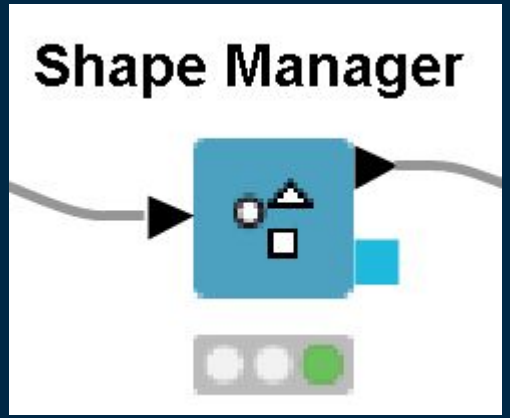
Este nodo asigna los nuevos datos a un conjunto existente de prototipos, que se obtienen, por ejemplo, mediante una agrupación de k-means. Cada punto de datos se asigna a su prototipo más cercano.

3a Sección - Nodos



Se pueden asignar colores para las columnas nominales (los valores posibles tienen que estar disponibles) o numéricas (con límites inferiores y superiores). Si estos límites no están disponibles, se proporciona un '?' como valor mínimo y máximo. Los valores se calculan durante la ejecución. Si se selecciona un atributo de columna, el color puede cambiarse con el selector de color.

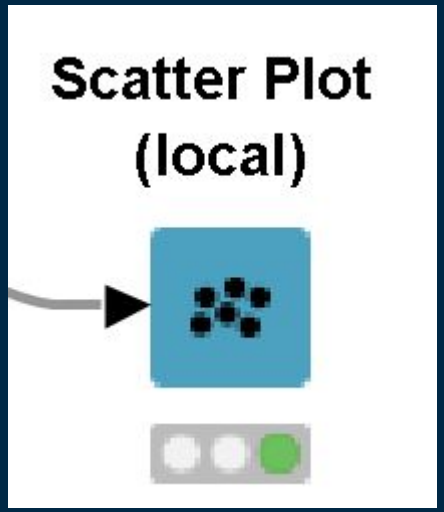
3a Sección - Nodos



Asigna formas (diferentes) para cada valor de atributo de una columna nominal, es decir, para cada valor posible. Las vistas de apoyo muestran entonces los puntos de datos con la forma asociada al valor del atributo correspondiente. Si hay, por ejemplo, un conjunto de datos con dos clases diferentes ("clase1" y "clase2"), la "clase1" puede tener asignado un círculo y la "clase2" un triángulo. Al observar el conjunto de datos, los valores pueden distinguirse fácilmente por su forma.

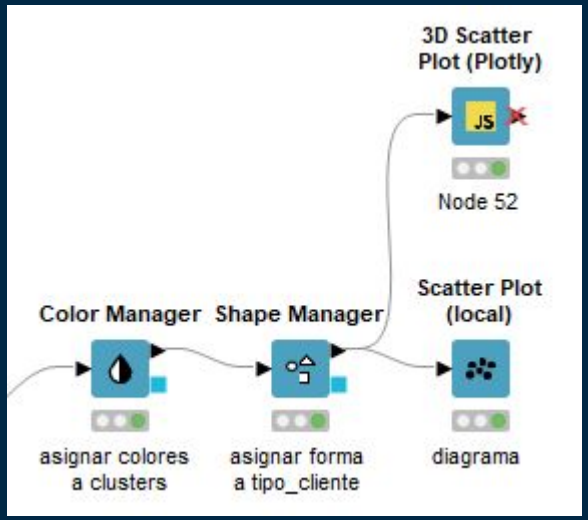
En el diálogo se puede seleccionar la columna nominal con los posibles valores. Los valores posibles aparecen en la columna de la izquierda y la forma puede establecerse en la columna de la derecha de la tabla haciendo clic en ella y seleccionando la forma deseada.

3a Sección - Nodos



Crea un gráfico de dispersión de dos atributos seleccionables. A continuación, cada punto de datos se muestra como un punto en su lugar correspondiente, dependiendo de sus valores de los atributos seleccionados. Los puntos se muestran en el color definido por el Gestor de Color, el tamaño definido por el Gestor de Tamaño y la forma definida por el Gestor de Forma.

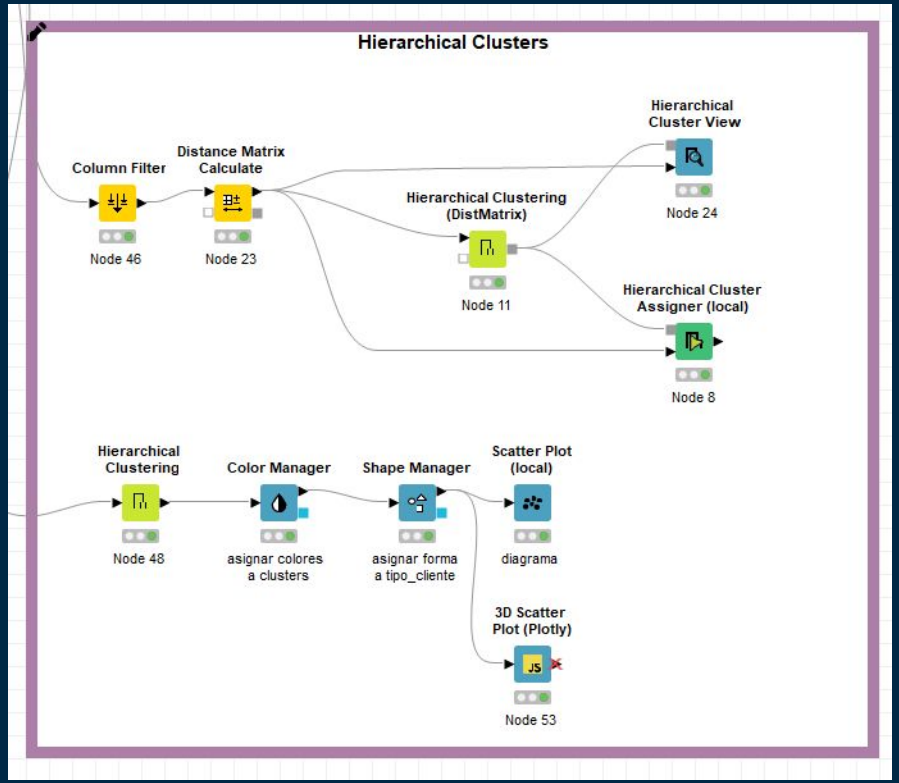
3a Sección – Visualización



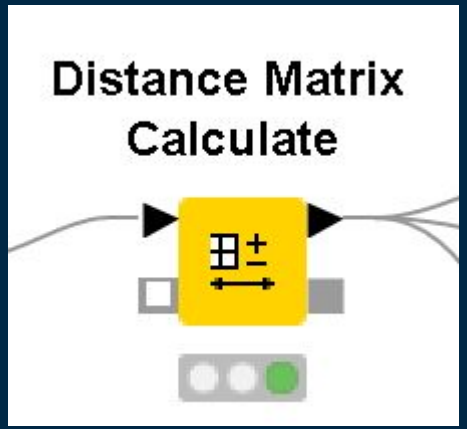
Finalmente, se muestra el uso en conjunto de los nodos mencionados anteriormente. Se puede apreciar el uso del nodo 3D Scatter incluido en las extensión plotly.js para knime.

3a Sección – Generación de Clusters Jerárquicos

El flujo de la parte superior corresponde a la generación del clúster suministrando previamente el cálculo de las distancias euclidianas. Mientras que el de la parte inferior corresponde a la efectuación de un corte sobre los clusters que se generaron.

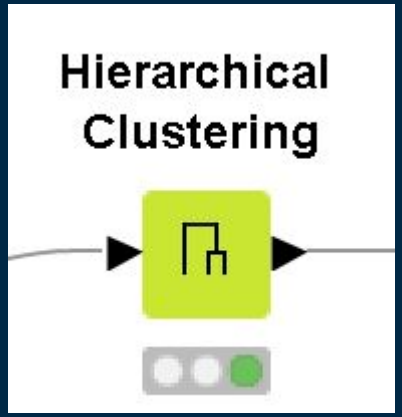


3a Sección - Nodos



Calcula los valores de distancia para todos los pares de filas de la tabla de entrada. El resultado se añade a la tabla de entrada como una sola columna que contiene los valores del vector de distancia.

3a Sección - Nodos



Agrupa jerárquicamente los datos de entrada.

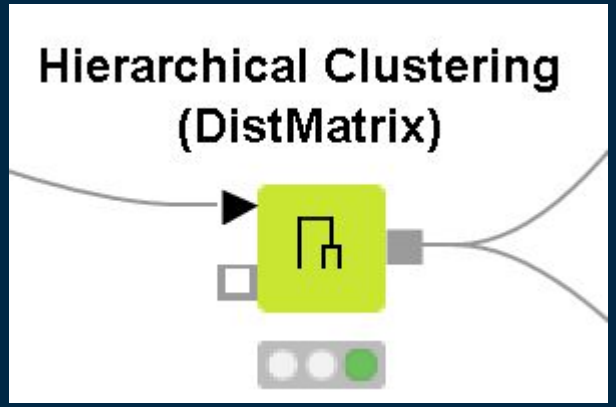
Hay dos métodos para hacer clustering jerárquico:

- Top-down o divisivo, es decir, el algoritmo comienza con todos los puntos de datos en un cluster enorme y los puntos de datos más disímiles se dividen en subclusters hasta que cada cluster está formado exactamente por un punto de datos.

- Bottom-up o aglomerativo, es decir, el algoritmo comienza con cada punto de datos como un único cluster y trata de combinar los más similares en superclusters hasta terminar en un cluster enorme que contiene todos los subclusters.

Este algoritmo funciona de forma aglomerativa.

3a Sección - Nodos



Agrupa jerárquicamente los datos de entrada utilizando una matriz de distancia.

Para determinar la distancia entre clusters hay que definir una medida. Básicamente, existen tres métodos para comparar dos clusters:

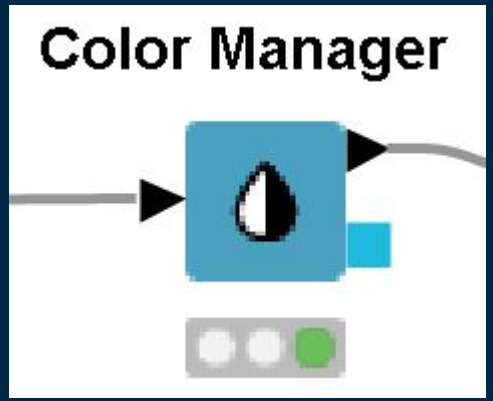
* Single Linkage: define la distancia entre dos clusters c1 y c2 como la distancia mínima entre dos puntos cualesquiera x, y con x en c1 e y en c2.

* Enlace completo: define la distancia entre dos conglomerados c1 y c2 como la distancia máxima entre dos puntos cualesquiera x, y con x en c1 e y en c2.

* Enlace medio: define la distancia entre dos conglomerados c1 y c2 como la distancia media entre todos los puntos de c1 y c2.

La información de distancia utilizada por este nodo se lee de una columna de vector de distancia que debe estar disponible en los datos de entrada o se calcula directamente con el uso de una medida de distancia conectada. Siempre se puede calcular la matriz de distancia utilizando el nodo de cálculo correspondiente.

3a Sección - Nodos



Se pueden asignar colores para las columnas nominales (los valores posibles tienen que estar disponibles) o numéricas (con límites inferiores y superiores). Si estos límites no están disponibles, se proporciona un '?' como valor mínimo y máximo. Los valores se calculan durante la ejecución. Si se selecciona un atributo de columna, el color puede cambiarse con el selector de color.

REFERENCIAS

[1] Conecta Software. (2020, 7 de marzo). *Clustering y Análisis de datos - Conecta Software*. https://conectasoftware.com/analytics/clustering-y-analisis-de-datos/#:~:text=Clustering%20es%20una%20técnica%20utilizada,también%20se%20conoce%20como%20segmentación.

[2] Criado, J. (s. f.). *Análisis de datos*. AnalisisDeDatos.net: Web dedicada a la difusión de técnicas relacionadas con el análisis de datos y su entorno. https://analisisdedatos.net/mineria/tecnicas/clustering/intro.php

[3]KNIME. (s. f.). *k-Means*. KNIME Hub. https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.cluster.kmeans.ClusterNodeFactory2

REFERENCIAS

[4] **Conjunto de datos:**

https://www.kaggle.com/imakash3011/customer-personality-analysis

[5] **Análisis:**

https://thecleverprogrammer.com/2021/02/08/customer-personality-analysis-with-python/