UNIVERSIDAD DE GRANADA

Tratamiento Inteligente de Datos

Práctica 4

Autor

Antonio José Muriel Gálvez

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Granada, 3 de diciembre de 2024

Índice

Ejercicio 1

Enunciado

Vino

El archivo wine.data contiene los datos reales de 178 vinos de una misma región de Italia. Cada instancia está compuesta por trece atributos numéricos más una clase (la primera columna) que determina el nivel de alcohol del vino (tres tipos; 1, 2 y 3), y es la solución al proceso de clustering (por lo que se debe eliminar para realizar la tarea de minería de datos). La descripción de los atributos se encuentra en el fichero wine_names.txt (por algún motivo, falta la descripción de una de las columnas). Se deben realizar las siguientes actividades:

- Realizar un algoritmo de clustering jerárquico para analizar en cuántos clusters diferentes podríamos agrupar los datos.
- Analizar la existencia de outliers y eliminarlos si consideras que existe alguno. Repite el clustering jerárquico y vuelve a analizar el número de clusters a considerar.
- Aplicar el algoritmo k-medias al archivo del punto anterior con el número de clusters elegido. Compara cómo se distribuyen los clusters respecto a las clases de la primera columna.
- Aplicar algún tipo de reducción de dimensionalidad que consideres oportuno (filtrando columnas, correlación, análisis de componentes principales, etc.) y aplica el algoritmo k-medias para tres clusters. De nuevo, compara cómo se distribuyen los clusters respecto a las clases de la primera columna.
- Opcionalmente, aplica el algoritmo DBSCAN basado en la densidad al archivo original, para varios parámetros de radio (epsilon) y puntos mínimos. Compara cómo se distribuyen los clusters respecto a la primera columna. ¿Se pueden identificar los outliers?

Solución KNIME

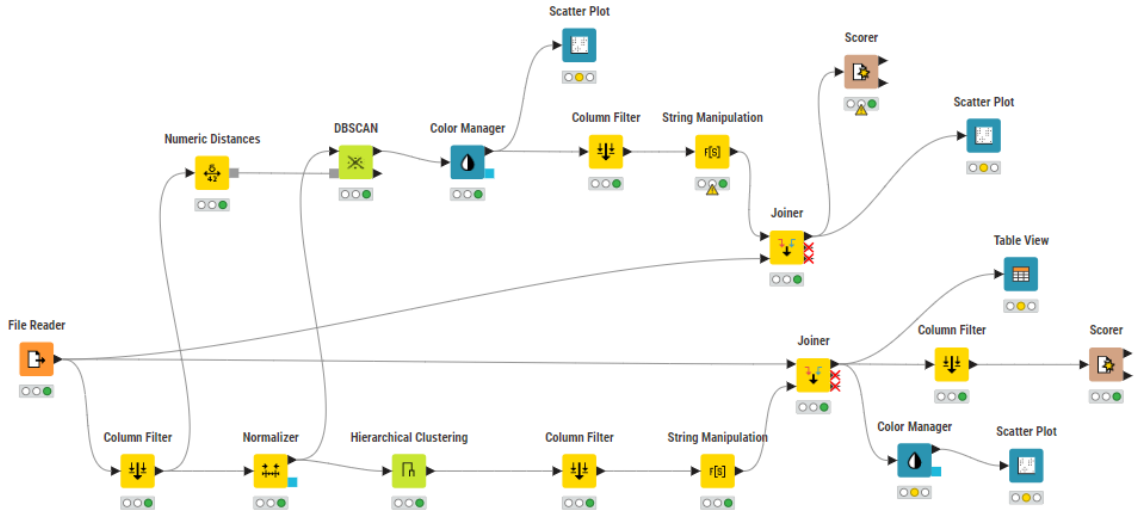

Figura 1: Solución Knime

Nodo Column Filter

La primera columna, que contiene las soluciones o clases, fue eliminada del conjunto de datos. Esta columna no es relevante para la segmentación y podría haber sesgado el análisis.

Nodo Normalizer

Para asegurar que todas las características tuvieran la misma escala, los datos fueron normalizados utilizando el método **Min-Max**. Este proceso transforma cada característica en un rango de [0, 1].

Esto garantiza que todas las características contribuyan de manera equitativa al cálculo de las distancias, evitando que las variables con mayor rango dominen el proceso de clustering.

Nodo Hierarchical Clusterin

El objetivo de este análisis es segmentar un conjunto de datos en 3 clusters utilizando el algoritmo de clustering jerárquico. Como sabemos que el número óptimo de clusters es 3, hemos decidido experimentar con diferentes configuraciones de **Distance Function** y **Linkage Type** para ver cómo afectan a la segmentación de los datos. Las configuraciones que se probaron son las siguientes:

- **Distance Functions**:
  - Euclidean
  - Manhattan
- **Linkage Types**:
  - Single
  - Average
  - Complete

Se utilizó el algoritmo de **Clustering Jerárquico** con las siguientes opciones:

- **Distance Function**: Euclidean o Manhattan, para determinar la medida de distancia entre los puntos de datos.
- **Linkage Type**: Single, Average, Complete, para determinar cómo se calcula la distancia entre clusters.

Se analizó el dendrograma y la distribución de los clusters generados para cada combinación de estos parámetros.

Pruebas con Distance Function: Euclidean

En esta configuración, se utilizó la distancia **Euclidiana** para calcular la proximidad entre los puntos de datos. Se probaron diferentes tipos de enlace:

- **Single Linkage**: En el análisis realizado utilizando el método de enlace "Singleçon la distancia Euclidiana, se observa que la mayoría de las filas han sido agrupadas en el cluster_2, mientras que unas pocas filas han sido asignadas a cluster_0 y cluster_1. Esta distribución sugiere que el uso de la distancia Euclidiana y el tipo de linkage "Single"tiende a agrupar datos similares en un solo cluster, con solo un pequeño número de elementos asignados a clusters separados.

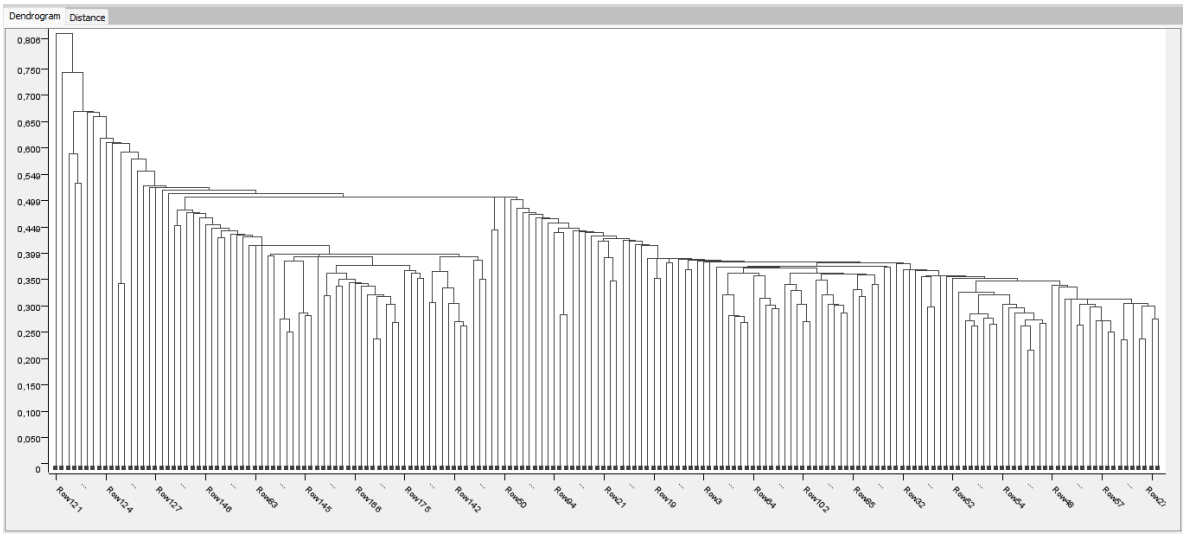

Figura 2: Euclidean - Single - Dendrogram

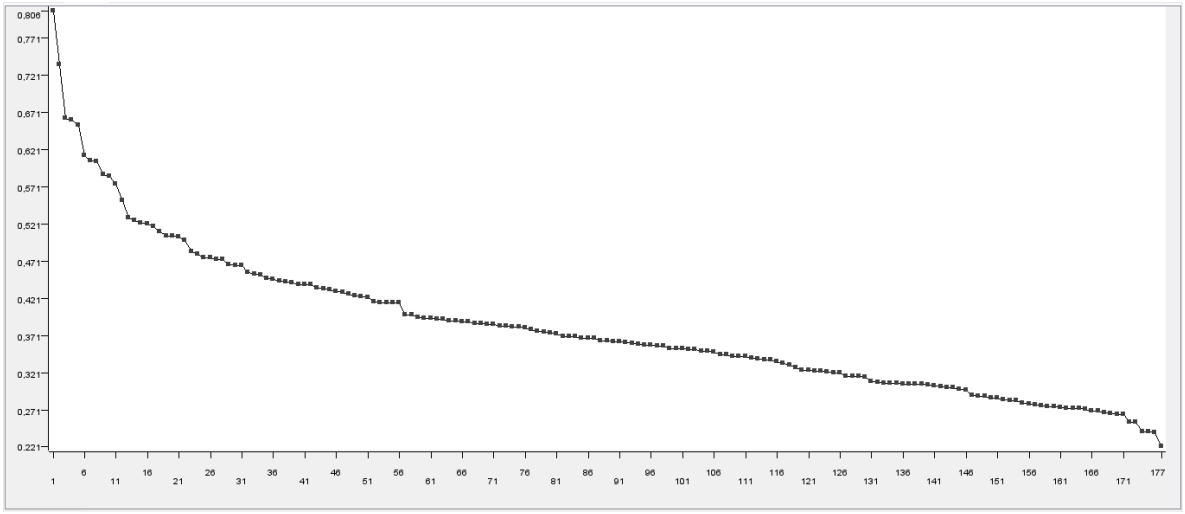

Figura 3: Euclidean - Sinble - Distance

- **Average Linkage**: Muestra una división de los datos en tres clusters, mostrando una mayor homogeneidad dentro de cluster_2 y destacando la variabilidad de cluster_0 y cluster_1.

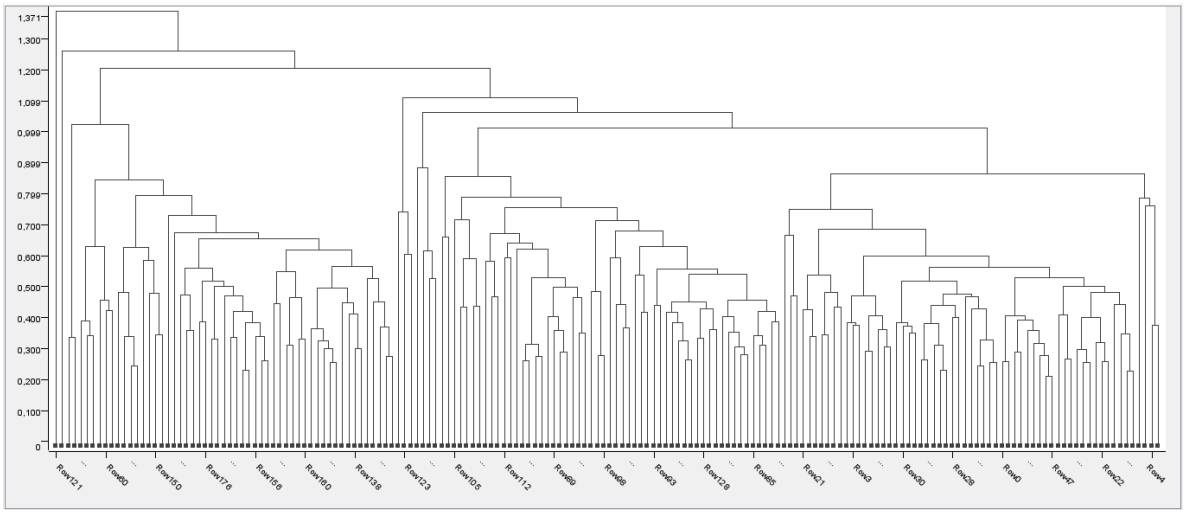

Figura 4: Euclidean - Average - Dendrogram

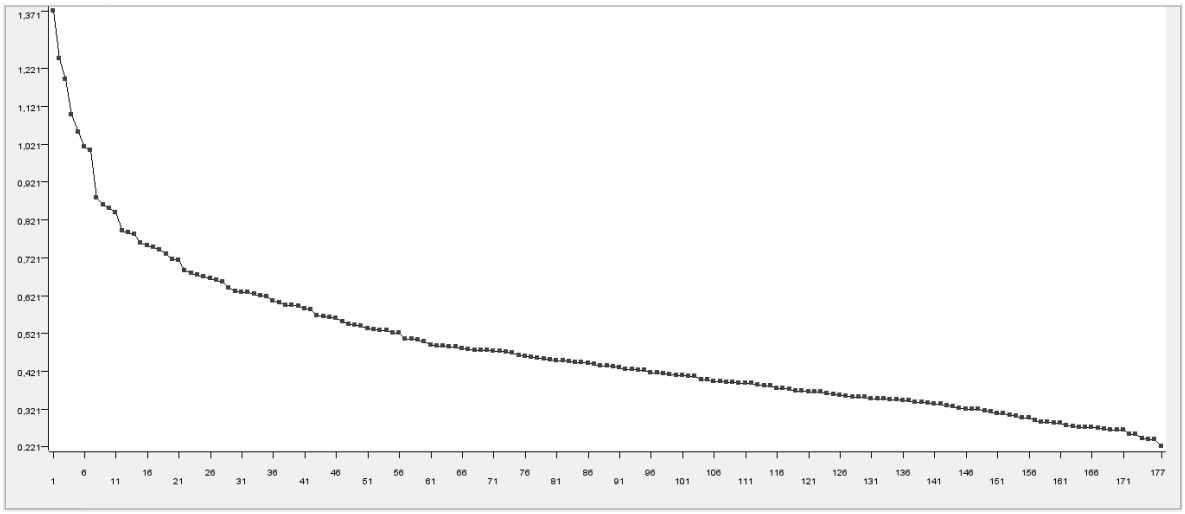

Figura 5: Euclidean - Average - Distance

- **Complete Linkage**: agrupa la mayoría de los datos, con características más homogéneas, mientras que los Cluster 0 y Cluster 1 contienen elementos con características más únicas y dispersas, lo que resalta la capacidad de este método para segmentar los datos de manera más precisa.

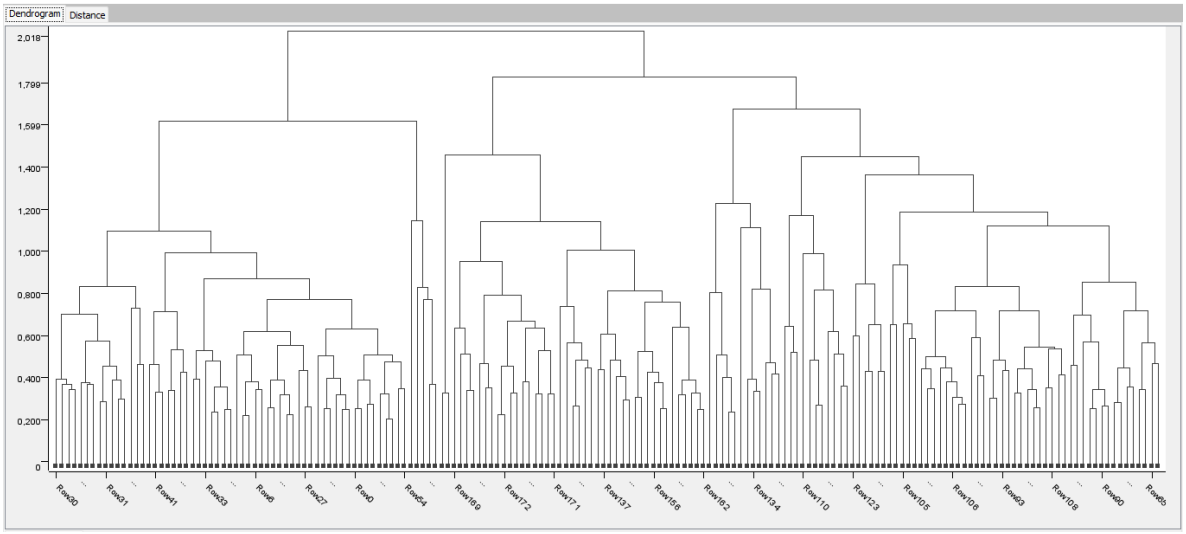

Figura 6: Euclidean - Complete - Dendrogram

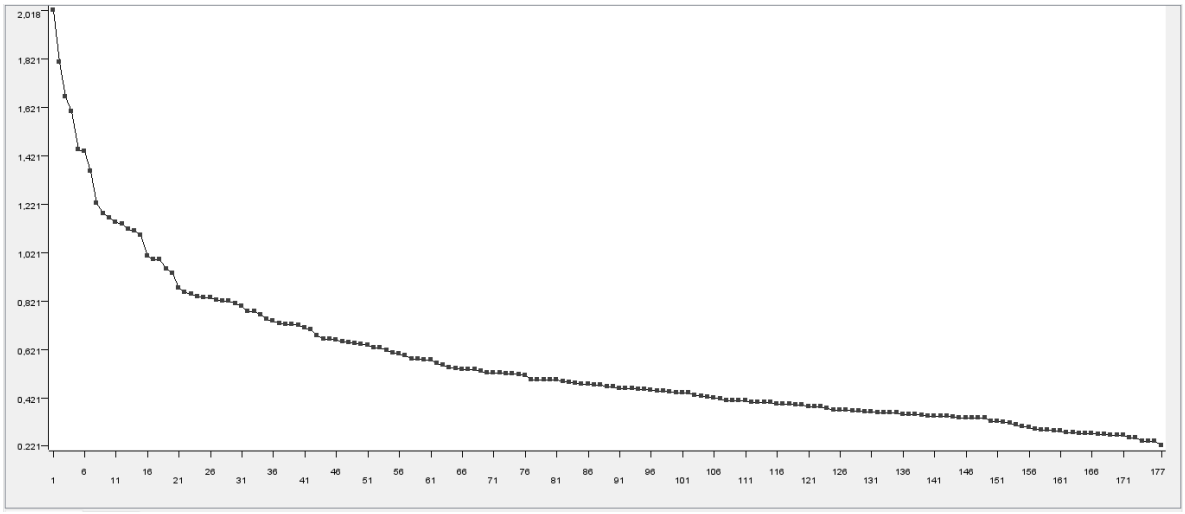

Figura 7: Euclidean - Complete - Distance

Pruebas con Distance Function: Manhattan

Se repitieron las pruebas anteriores, pero esta vez utilizando la distancia **Manhattan**, que calcula la distancia como la suma de las diferencias absolutas entre las coordenadas de los puntos. Los resultados obtenidos fueron:

- **Single Linkage**: Muestra en una segmentación poco efectiva de los datos. Este enfoque ha mostrado ser ineficaz para generar clusters bien definidos, ya que los grupos resultantes presentan una alta dispersión interna. Tiende a formar clusters que incluyen elementos con diferencias significativas entre sí, lo que da lugar a agrupaciones menos coherentes.

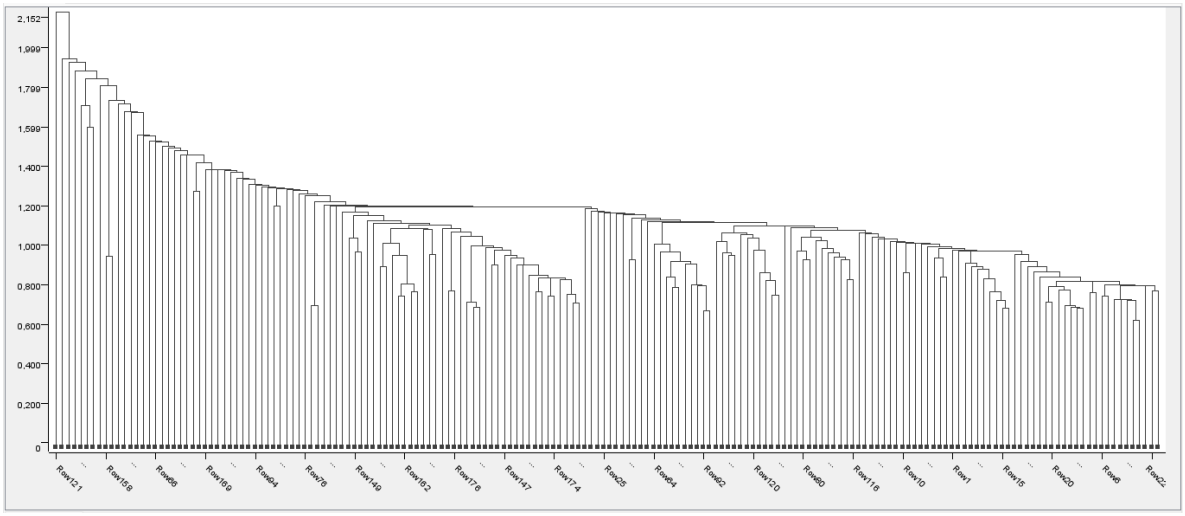

Figura 8: Manhattan - Single - Dendrogram

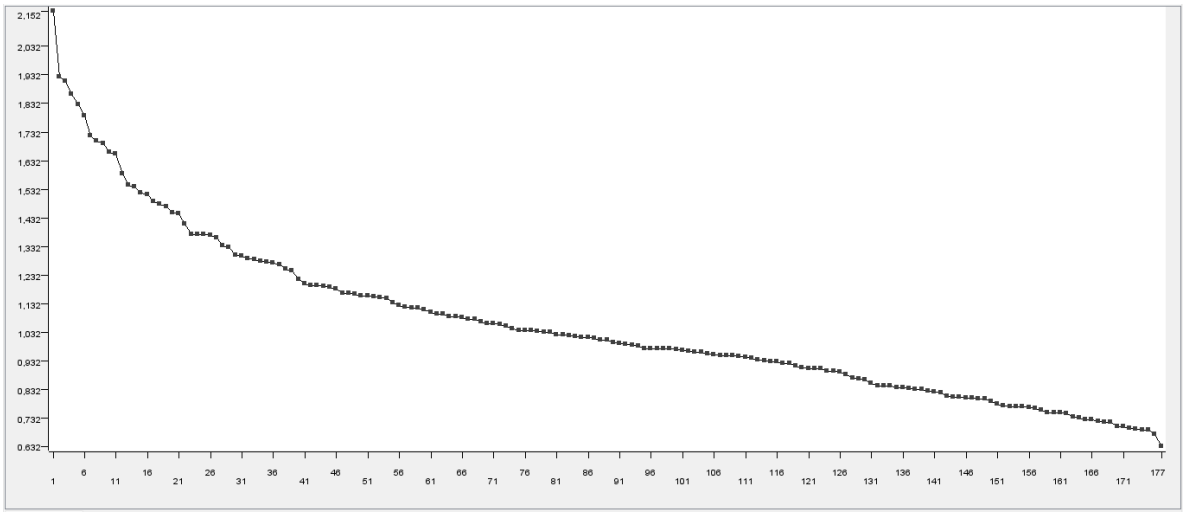

Figura 9: Manhattan - Single - Distance

- **Average Linkage**: Ha agrupado los elementos de forma más adecuada, reflejando relaciones más estrechas entre las características de las filas dentro de cada cluster. Los valores de cohesión interna son más sólidos, lo que indica que los elementos dentro de un mismo cluster son más similares entre sí.

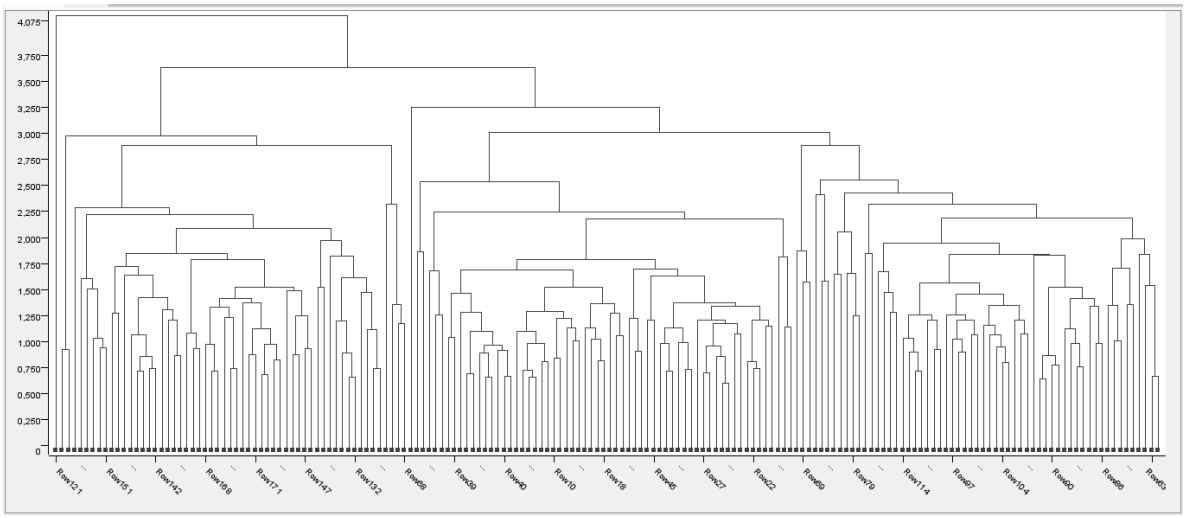

Figura 10: Manhattan - Average - Dendrogram

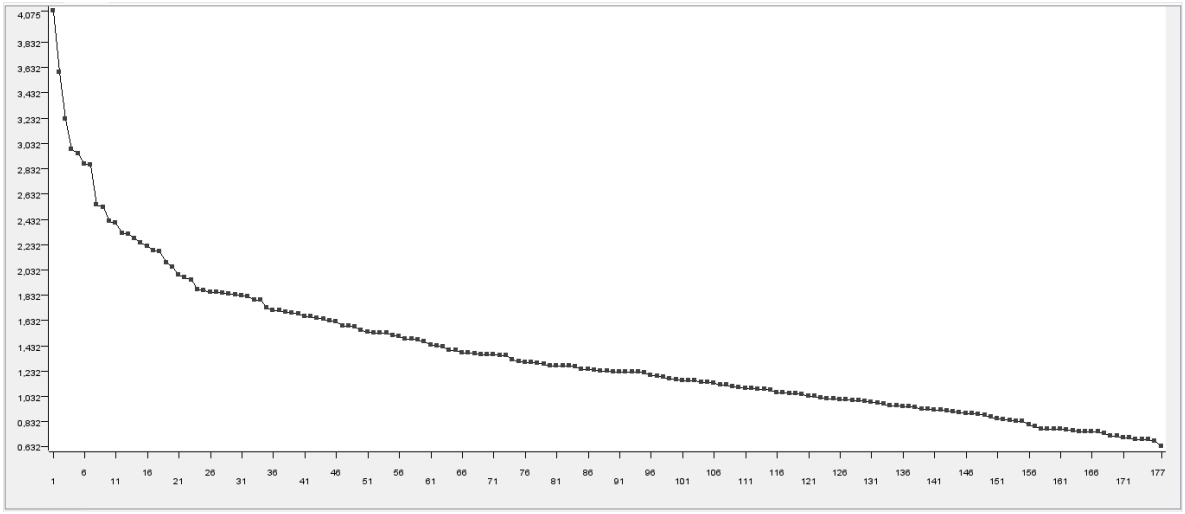

Figura 11: Manhattan - Average - Distance

- **Complete Linkage**: Ha agrupado los elementos de forma más adecuada, reflejando relaciones más estrechas entre las características de las filas dentro de cada cluster. Los valores de cohesión interna son más sólidos, lo que indica que los elementos dentro de un mismo cluster son más similares entre sí.
  Ha resultado ser significativamente más efectivo en la formación de clusters coherentes en comparación con otros enfoques.

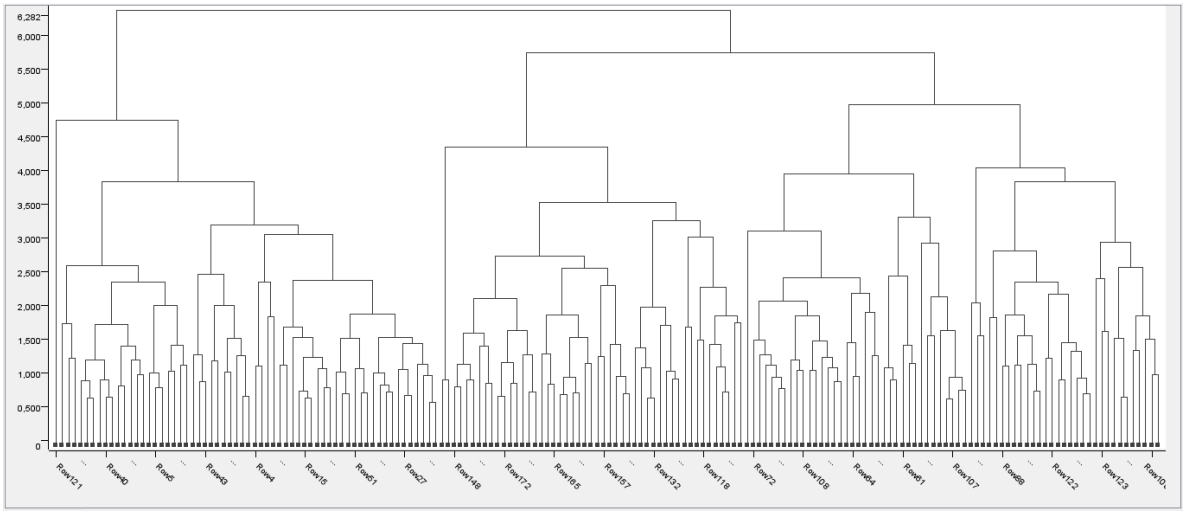

Figura 12: Manhattan - Comlete - Dendrogram

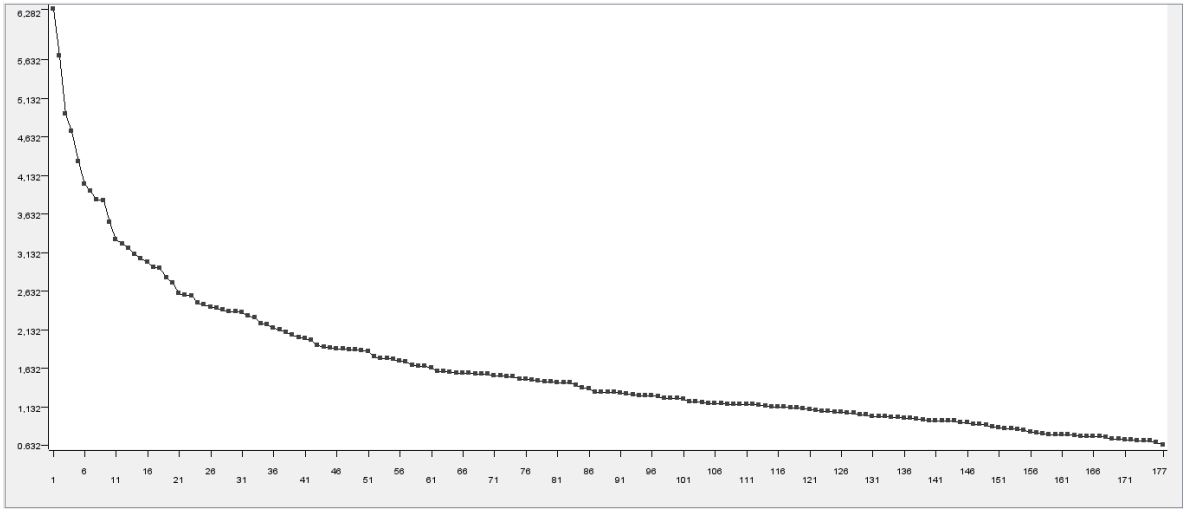

Figura 13: Manhattan - Comlete - Distance

Resultados y Discusión

Se analizaron los dendrogramas generados para cada combinación de parámetros, y se observó que las diferentes configuraciones de **Distance Function** y **Linkage Type** afectan la forma en que los clusters se definen.

A continuación, se analizan los resultados obtenidos utilizando el **Linkage Completo** para cada **Distance Function** (Euclidiana y Manhattan).

Resultados Euclidean

El uso de la distancia Euclidiana con linkage completo produjo una clara segmentación de los datos en 3 clusters. Los resultados fueron consistentes con la intuición sobre la distribución de los datos,

logrando una agrupación coherente y homogénea. **Pruebas con Distance Function: Euclidean**

#	RowID	1 *Number (integer)*	2 *Number (integer)*	3 *Number (integer)*
1	1	59	0	0
2	2	3	66	2
3	3	0	7	41

Figura 14: Euclidean - Complete - Scorer

Resultados Manhattan

De manera similar, la función de distancia Manhattan también generó una segmentación en 3 clusters. Aunque los clusters son igualmente distintos, la forma en que los datos fueron agrupados difiere ligeramente debido a las características de la distancia Manhattan, que se enfoca más en las diferencias absolutas en cada dimensión.

RowID	1 *Number (integer)*	2 *Number (integer)*	3 *Number (integer)*
1	59	0	0
2	3	64	4
3	0	3	45

Figura 15: Manhattan - Comlete - Scorer

Conclusiones

Tras experimentar con las diferentes combinaciones de **Distance Function** y **Linkage Type**, se concluye lo siguiente:

- El método **Average Linkage** con **Euclidean Distance** parece ser el más adecuado para segmentar este conjunto de datos en 3 clusters, ya que produce clusters homogéneos y bien definidos. - **Manhattan Distance** no proporciona resultados tan compactos como la distancia Euclidiana, pero puede ser útil para otros tipos de datos más dispersos. - **Complete Linkage** también produce buenos resultados, aunque en algunos casos puede ser demasiado rígido al fusionar clusters.

Opcional DBSCAN

Para esta parte se ha incluido los nodos **Numeric Distances** y **DBSCAN** con las siguientes configuraciones:

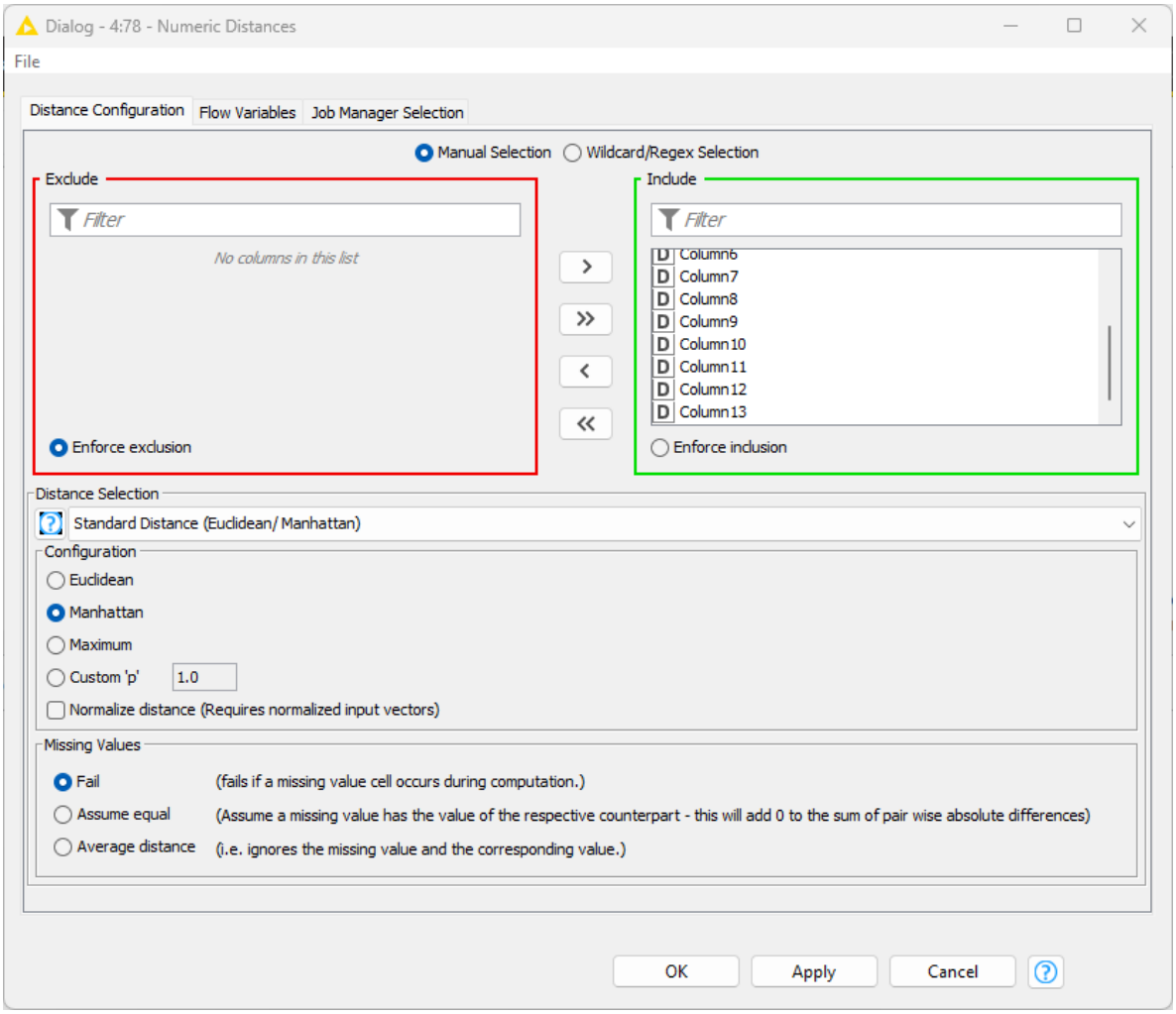

Figura 16: Numeric Distances

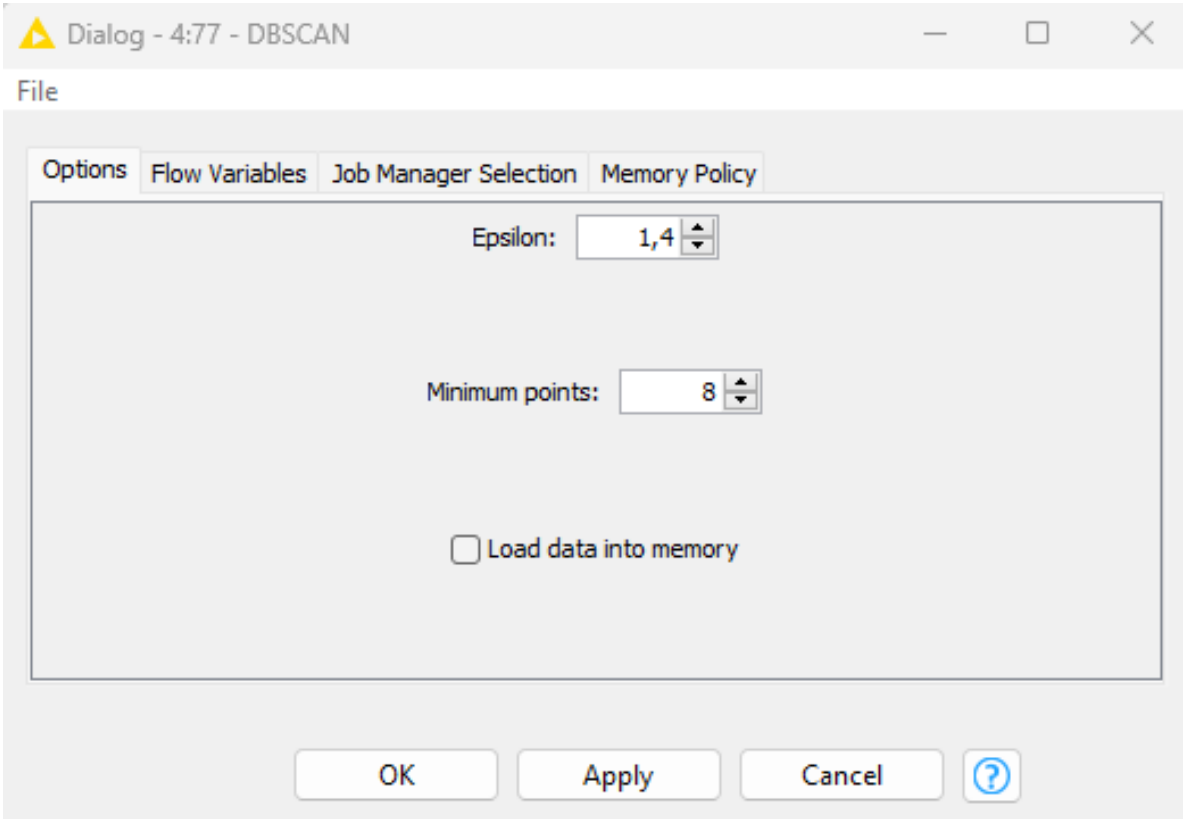

Figura 17: DBSCAN

Al compararlos con la solución de la columna 0 salen los siugientes resultados:

#	RowID	1 Number (integer)	3 Number (integer)	2 Number (integer)	-1 Number (integer)
1	1	56	0	0	3
2	3	0	44	0	4
3	2	44	1	0	26
4	-1	0	0	0	0

Figura 18: Scorer

1. Cluster 1: Este cluster agrupa principalmente instancias de la Clase 1, con 56 instancias correctamente identificadas. Además, hay 3 puntos considerados como outliers asociados al cluster.
2. Cluster 2: Este cluster es mayoritariamente de la Clase 1, con 44 instancias, pero también contiene 1 instancia de la Clase 2. Hay un número significativo de 26 outliers asociados a este cluster.
3. Cluster 3: Este cluster agrupa 44 instancias de la Clase 3 y 4 puntos identificados como outliers.
4. Outliers (-1): No se han identificado 3 puntos del cluster 1, 4 del cluster 3 y 26 del cluster 2 como outliers.

Cluster 1 agrupa bien la Clase 1, con una asignación clara. Cluster 3 también separa correctamente la mayoría de las instancias de la Clase 3. Cluster 2 es una mezcla, agrupando mayoritariamente a la Clase 1, pero incluye un pequeño número de casos de la Clase 2, lo cual sugiere que esta clase presenta una mayor dificultad para separarse completamente.

Ejercicio 2

Enunciado

Este ejercicio se deja como un problema abierto consistente en desarrollar una agrupación/clustering de un dataset de jugadores de baloncesto que jugaron algún partido de la NBA durante la temporada regular 2021-2022. Concretamente, se pretende aprovechar las técnicas de clustering para examinar qué diferentes tipos de jugador podemos encontrarnos en los equipos de la NBA. El dataset se llama NBA_RegularSeason2021_2022.xlsx, y es recomendable realizar una análisis exploratorio y un preprocesamiento de datos para agilizar los algoritmos de clustering.

Solución

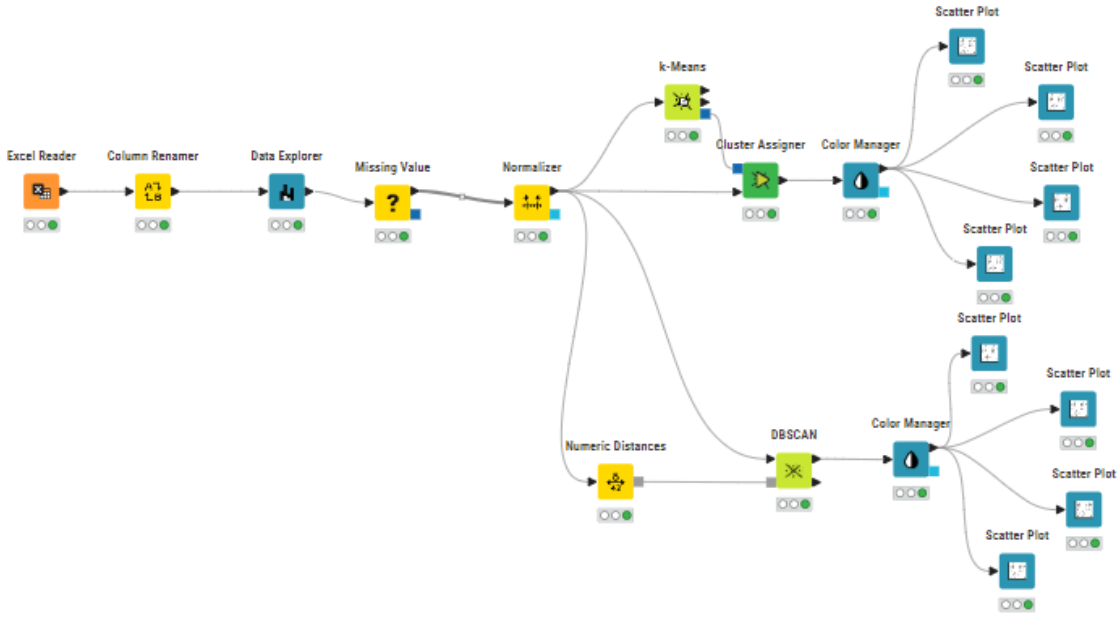

Figura 19: Solución

Nodo 1: Excel Reader

El primer paso es cargar los datos desde el archivo Excel. Esto se hace utilizando el nodo Excel Reader en KNIME

Nodo 2: Column Renamer

Se utiliza para ajustar y limpiar los nombres de las columnas del dataset con el objetivo de facilitar el análisis y evitar problemas de interpretación en los pasos posteriores. En este caso, el nodo se configuró para renombrar columnas y ajustar su formato según las siguientes consideraciones:

- Simplificar los nombres de las columnas eliminando caracteres especiales o espacios en blanco.

- Asegurar consistencia en los nombres para que sean claros y fáciles de usar en el análisis.
- Adaptar las etiquetas de las columnas para que reflejen mejor su contenido.

Nodo 3: Data Explorer

Se utiliza para realizar un análisis exploratorio de los datos. Este paso permite obtener un resumen estadístico detallado de las columnas del dataset, lo que facilita entender la distribución de los datos, identificar valores atípicos, detectar problemas como valores ausentes y evaluar las características clave de las variables. Variables Relevantes para Clustering:

- Rendimiento en Juego: PPG (Puntos por juego) tiene un rango amplio (0 a 30.6) y es una métrica clave para clasificar jugadores por su capacidad ofensiva. Rebounds per game y Assists per game permiten segmentar jugadores según sus roles de reboteador o creador de juego.
- Eficiencia: True Shooting Percentage y Effective Shooting Percentage son útiles para evaluar la eficiencia ofensiva de los jugadores.
- Contribución Defensiva: Blocks per game y Steals per game ayudan a categorizar jugadores defensivos.
- Versatilidad: El Versatility Index mide la capacidad de los jugadores para impactar en múltiples áreas del juego.

Nodo 4: Missing Value

El nodo Missing Value en KNIME se utiliza para gestionar valores faltantes en el dataset, asegurando que los datos estén completos antes de continuar con el análisis. Este paso es crucial, ya que los valores ausentes pueden afectar la precisión de los algoritmos de clustering y otras técnicas analíticas.

Nodo 5: Normalizer

Para escalar los valores numéricos del dataset dentro de un rango específico. En este caso, se aplicó la normalización Min-Max (0-1), que ajusta los valores de cada columna numérica para que estén entre 0 y 1.

La normalización asegura que cada variable contribuya de manera proporcional al análisis y una representación equitativa de todas las características en el análisis.

Nodo 6: K-means

El nodo K-means aplica el algoritmo de agrupación basado en distancias para dividir los datos en 4 clústeres. Este algoritmo particiona los puntos de datos en grupos que minimizan la distancia total al centroide más cercano.

Cada jugador se asigna a uno de los 4 clústeres según sus características. Una nueva columna con el número de clúster se añade al dataset.

Nodo 7: Cluster Assigner

Toma los resultados generados por el nodo K-means (número de clúster y centroides finales) y asigna a cada jugador un número de clúster correspondiente. Esto permite identificar claramente a qué grupo pertenece cada registro del dataset. Añade una columna de clúster al dataset original para facilitar la identificación de cada jugador dentro de su grupo. Permitir la comparación de las características originales de los jugadores con su asignación de clúster. Provee un dataset listo para la visualización y análisis posterior, como gráficos de dispersión o análisis estadístico.

Nodo 8: Color Manager

Para asignar colores específicos a los clústeres creados en el proceso de agrupamiento. Este paso es crucial para visualizar las agrupaciones de manera clara y comprensible, especialmente cuando se trata de gráficos o representaciones visuales como los gráficos de dispersión.

Nodo 9: Scatter Plot

El nodo Scatter Plot en KNIME se utiliza para crear representaciones gráficas de los datos, permitiendo visualizar las relaciones entre diferentes variables numéricas. En este caso, voy a realizar 4 estudios:

- Effective Shooting Percentage vs. Minutes Percentage: Hay una relación positiva débil entre los minutos jugados y la efectividad en el tiro dentro del Cluster 0. El Cluster 1 está claramente delimitado con minutos jugados cercanos a cero. El Cluster 2 muestra mayor dispersión, indicando una amplia diversidad en eficiencia y uso de minutos.
    - Cluster 0 (verde): Jugadores con un alto porcentaje de minutos jugados y efectividad en el tiro (Effective Shooting Percentage). Representan jugadores clave en el equipo.
    - Cluster 1 (rojo): Jugadores con poca participación en minutos y una efectividad baja o nula. Representan jugadores de rol menor o con menor impacto ofensivo.
    - Cluster 2 (naranja): Jugadores con minutos moderados y una efectividad de tiro aceptable. Son jugadores de soporte, no necesariamente titulares.
    - Cluster 3 (morado): Jugadores con minutos moderados a altos y una efectividad variable, pero en general buena.

Figura 20: Scatter Plot - Effective Shooting Percentage-Minutes Percentage

- Points per game vs. Rebounds per game:Existe una correlación positiva moderada entre puntos y rebotes en el Cluster 0, destacando su impacto multidimensional. El Cluster 1 está agrupado cerca del origen, mostrando poca contribución en ambas métricas. El Cluster 3 tiene dispersión en puntos, indicando que algunos jugadores priorizan anotación sobre rebotes.
    - Cluster 0 (verde): Jugadores con alta capacidad anotadora y buena cantidad de rebotes. Representan jugadores versátiles, probablemente jugadores estrella.
    - Cluster 1 (rojo): Jugadores con bajos puntos por juego y rebotes. Son jugadores con un impacto limitado.
    - Cluster 2 (naranja): Jugadores con producción moderada tanto en puntos como en rebotes. Representan jugadores promedio o de rol secundario.
    - Cluster 3 (morado): Jugadores con producción media-alta en puntos y rebotes. Representan jugadores importantes, pero no necesariamente estrellas.

Figura 21: Scatter Plot - Points per game-Rebounds per game

- Points per game vs. Usage Rate: Hay una fuerte correlación positiva entre la tasa de uso (Usage Rate) y puntos por juego en el Cluster 0, destacando a los jugadores con roles centrales en la ofensiva. El Cluster 1 está agrupado cerca del origen, lo que refleja poca participación en la ofensiva. En el Cluster 3, los jugadores tienen tasas de uso variadas, lo que indica roles ofensivos complementarios.
    - Cluster 0 (verde): Jugadores con alta producción de puntos y un alto uso en el equipo (Usage Rate). Representan jugadores clave ofensivos.
    - Cluster 1 (rojo): Jugadores con bajo uso y baja producción de puntos. Son jugadores marginales en el esquema ofensivo.
    - Cluster 2 (naranja): Jugadores con uso y puntos moderados. Representan jugadores de soporte ofensivo.
    - Cluster 3 (morado): Jugadores con producción de puntos media y un uso moderado a alto.

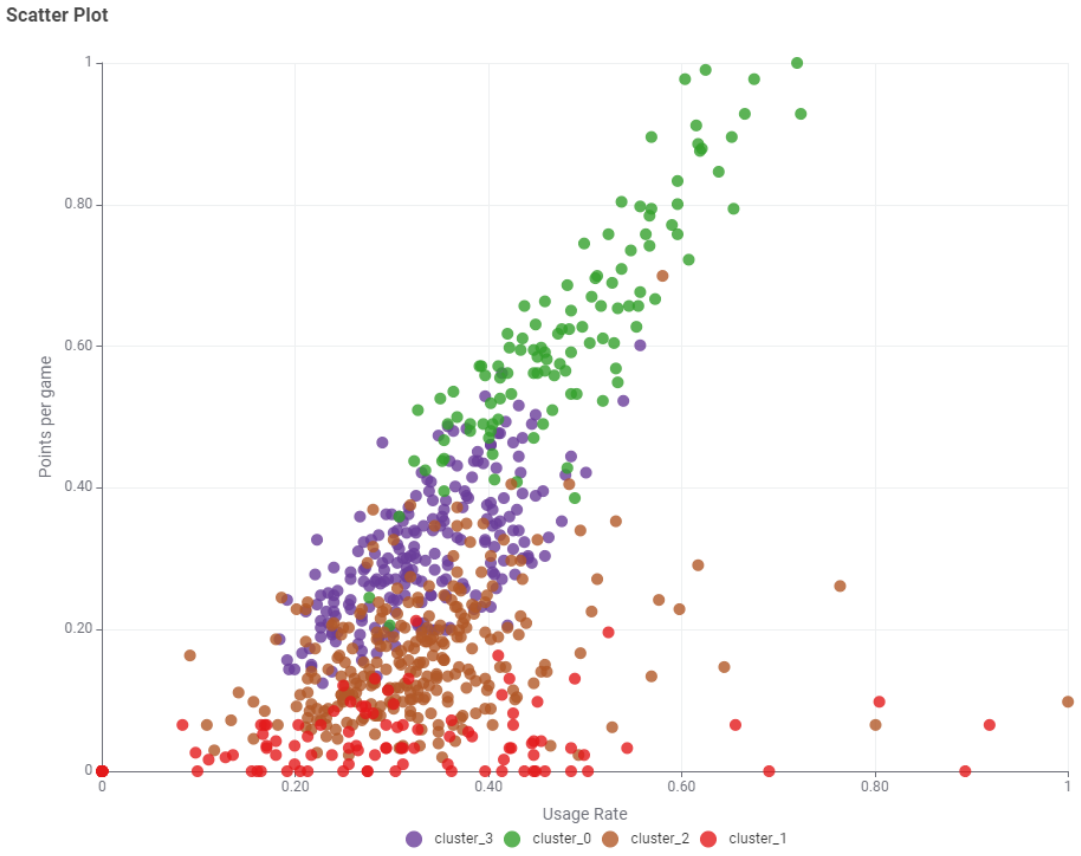

Figura 22: Scatter Plot - Points per game-Usage Rate

- 3P % vs. 3PA: Existe una correlación positiva moderada entre intentos de triples (3PA) y el porcentaje de aciertos (3P %) en el Cluster 0. El Cluster 1 se agrupa cerca del eje (0, 0), indicando poca participación en tiros de tres puntos. Los jugadores del Cluster 2 parecen ser inconsistentes, ya que intentan triples, pero sus tasas de éxito varían significativamente.
    - Cluster 0 (verde): Jugadores con una alta proporción de intentos de triples (3PA) y una eficiencia moderada a alta en porcentaje de triples acertados (3P %). Este grupo representa tiradores frecuentes con buen rendimiento.
    - Cluster 1 (rojo): Jugadores con bajo 3PA (casi nulo) y muy baja eficiencia en triples (3P %). Este grupo puede incluir jugadores que no intentan triples o tienen un rendimiento deficiente.
    - Cluster 2 (naranja): Representa una amplia gama de intentos de triples con eficiencia baja o moderada en 3P %. Puede reflejar jugadores que intentan triples ocasionalmente pero no son especialistas.
    - Cluster 3 (morado): Incluye jugadores con una proporción de intentos de triples moderada y eficiencia aceptable (3P %). Son tiradores complementarios.

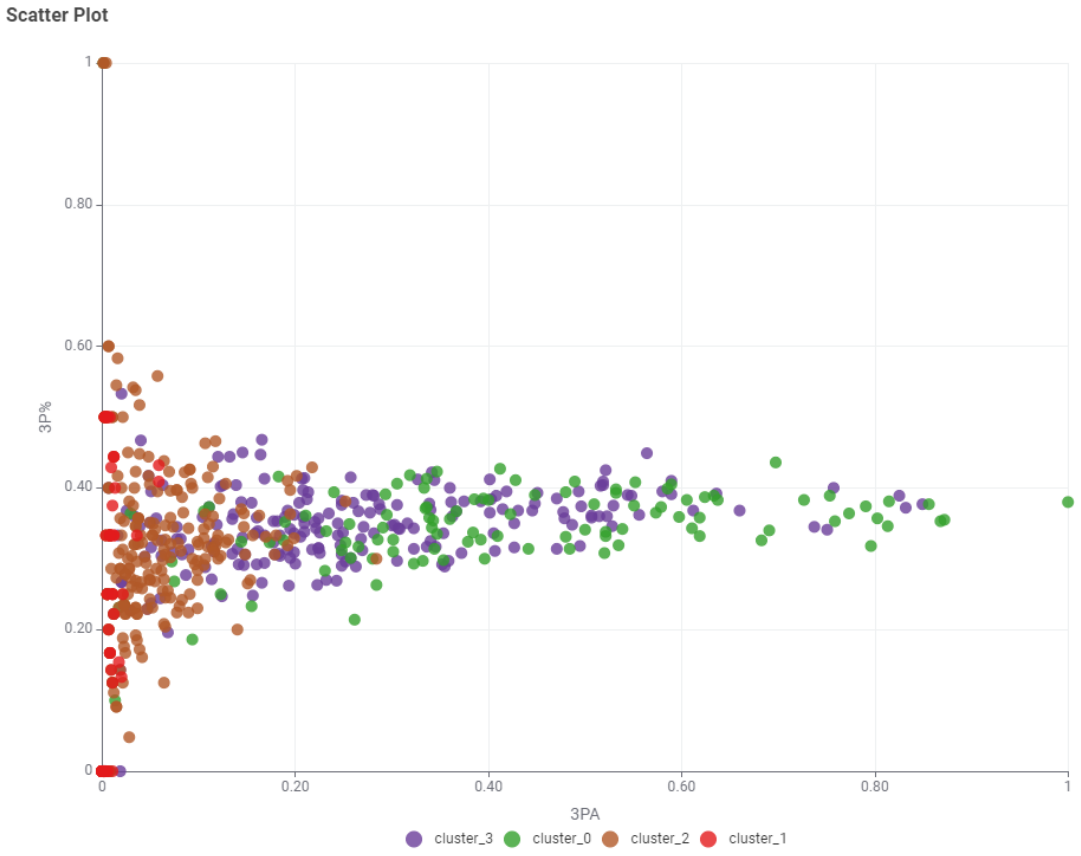

Figura 23: Scatter Plot - 3P %-3PA

Conclusión

Los clusters proporcionan una clasificación clara de los jugadores basados en su desempeño y roles en el equipo. Los Clusters 0 y 3 tienden a representar a jugadores destacados con roles importantes en ofensiva o defensiva. Los Clusters 1 y 2 reflejan jugadores de menor impacto o roles más específicos dentro del equipo. Las correlaciones entre variables como Usage Rate vs. Points per Game y 3PA vs. 3P % destacan las diferencias clave en los tipos de jugadores dentro de cada cluster.