Inteligencia Artificial

Práctica 2

Laboratorio: Segmentación de gasolineras por precios de combustibles

Carlos Gálvez Reguera

INDICE

INTRODUCCIÓN

En esta práctica he trabajado con un conjunto de datos reales sobre estaciones de servicio en España, con el objetivo de explorar cómo los precios de los carburantes pueden agruparse mediante técnicas de minería de datos. Lejos de limitarme a un análisis descriptivo, he querido comprender qué patrones subyacen en la distribución de precios y si existen agrupaciones significativas que permitan interpretar mejor el mercado de combustibles.

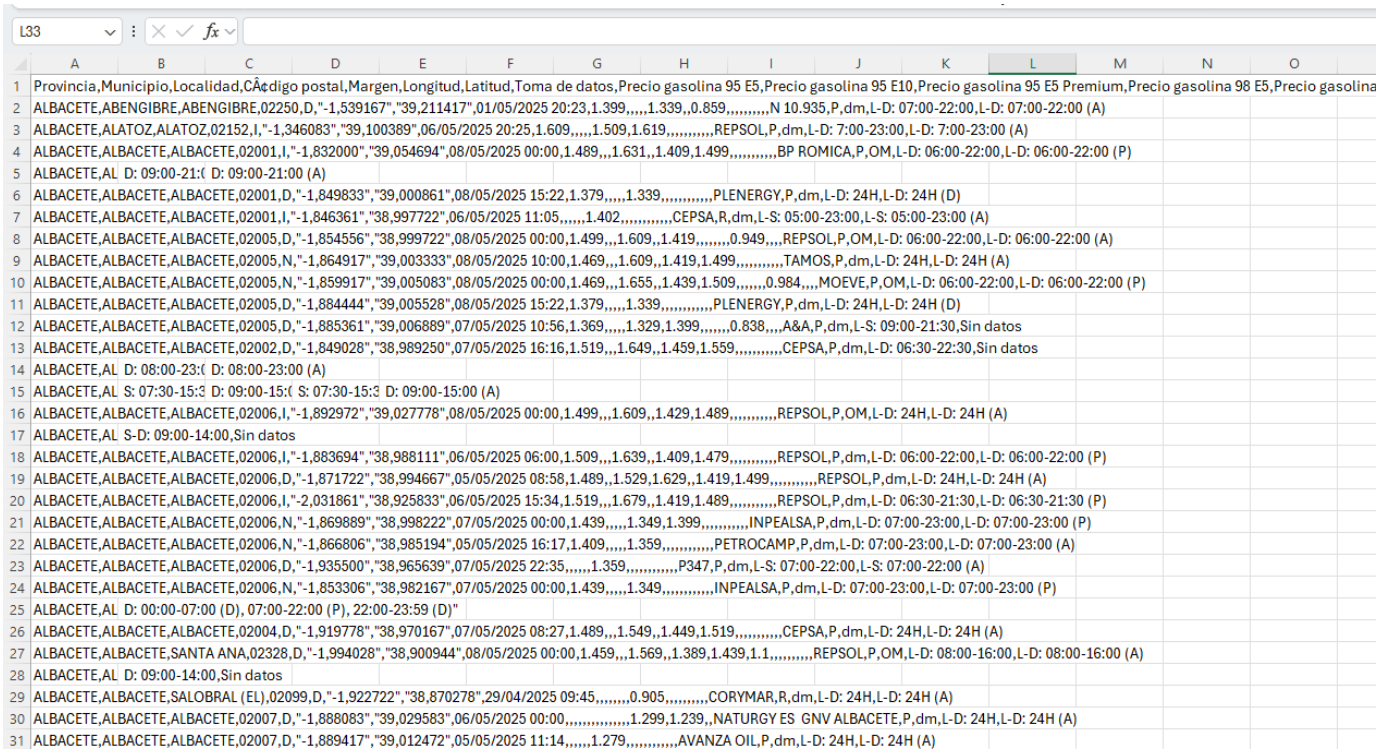
Para ello, he utilizado el software Weka, que permite aplicar algoritmos de aprendizaje automático sobre conjuntos de datos estructurados. La idea no era simplemente ejecutar un clustering, sino reflexionar sobre las decisiones que implica todo el proceso: desde la limpieza del archivo original, pasando por la selección de atributos relevantes, hasta la interpretación crítica de los resultados obtenidos. En definitiva, se trata de entender no solo “qué” resultados arroja el algoritmo, sino también “por qué” los datos se comportan así y qué implicaciones pueden tener.

Esta práctica me ha permitido afianzar conceptos clave del análisis de datos, pero también me ha obligado a tomar decisiones técnicas con criterio, y a pensar más allá de lo que se muestra en pantalla. Al final, la minería de datos no consiste solo en agrupar números, sino en buscar sentido a la información y tomar conciencia del proceso que convierte datos en conocimiento.

1. Contexto y preparación de los datos

El punto de partida de esta práctica fue un archivo con datos abiertos sobre estaciones de servicio en territorio español, publicado por una fuente oficial. Este conjunto incluía una gran variedad de atributos, desde la localización geográfica hasta el tipo de carburante y su precio, pasando por detalles como la marca comercial o el horario de apertura. Su riqueza y volumen lo convierten en una base idónea para experimentar con técnicas de análisis exploratorio y agrupamiento.

Antes de poder trabajar con estos datos en Weka, fue necesario realizar una limpieza cuidadosa. Lo primero fue abrir el archivo en Excel y deshacerse de las filas iniciales, que simplemente ofrecían información introductoria y no formaban parte del contenido útil. Decidí prescindir también de ciertas columnas que no aportaban valor directo al análisis, como la dirección exacta, que además complicaba el tratamiento posterior. Uno de los pasos más tediosos fue la normalización de los textos: quité tildes, corregí símbolos problemáticos y me aseguré de que los precios se interpretaran correctamente como valores numéricos, reemplazando las comas por puntos. Tras estos ajustes, exporté el conjunto en formato CSV, preparado para ser importado sin errores en el entorno de trabajo de Weka.



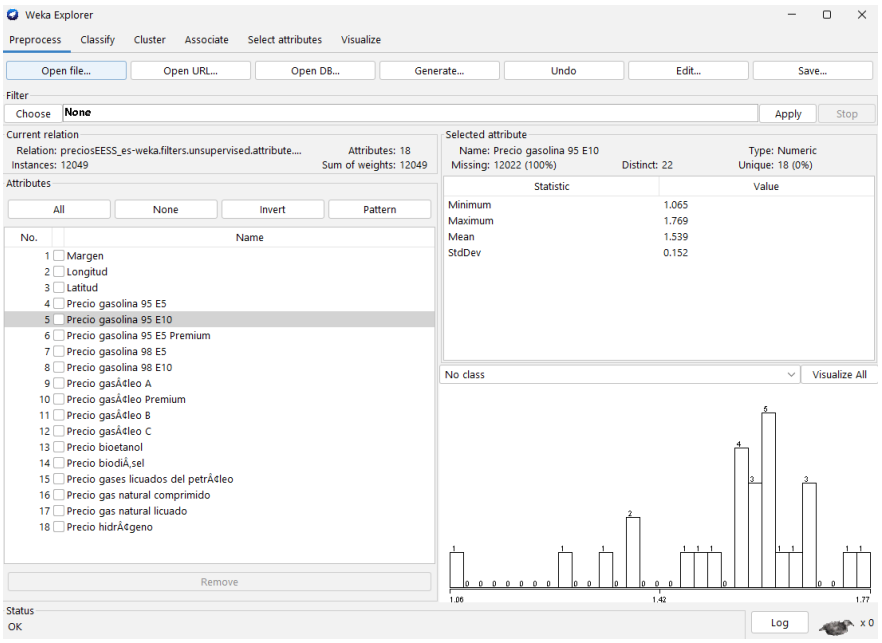
	A
1	Provincia,Municipio,Localidad,CÂ¢digo postal,Margen,Longitud,Latitud,Toma de datos,Precio gasolina 95 E5,Precio gasolina 95 E10,Precio gasolina 95 E5 Premium,Precio gasolina 98 E5,Precio gasolina
2	ALBACETE,ABENGIBRE,ABENGIBRE,02250,D,"-1,539167","39,211417",01/05/2025 20:23,1.399,,,,,1.339,,0.859,,,,,,,,,N 10.935,P,dm,L-D: 07:00-22:00,L-D: 07:00-22:00 (A)
3	ALBACETE,ALATOZ,ALATOZ,02152,I,"-1,346083","39,100389",06/05/2025 20:25,1.609,,,,,1.509,1.619,,,,,,,,,,REPSOL,P,dm,L-D: 7:00-23:00,L-D: 7:00-23:00 (A)
4	ALBACETE,ALBACETE,ALBACETE,02001,I,"-1,832000","39,054694",08/05/2025 00:00,1.489,,,1.631,,1.409,1.499,,,,,,,,,,BP ROMICA,P,OM,L-D: 06:00-22:00,L-D: 06:00-22:00 (P)
5	ALBACETE,AL D: 09:00-21:(D: 09:00-21:00 (A)
6	ALBACETE,ALBACETE,ALBACETE,02001,D,"-1,849833","39,000861",08/05/2025 15:22,1.379,,,,,1.339,,,,,,,,,,,PLENERGY,P,dm,L-D: 24H,L-D: 24H (D)
7	ALBACETE,ALBACETE,ALBACETE,02001,I,"-1,846361","38,997722",06/05/2025 11:05,,,,,,1.402,,,,,,,,,,,,CEPSA,R,dm,L-S: 05:00-23:00,L-S: 05:00-23:00 (A)
8	ALBACETE,ALBACETE,ALBACETE,02005,D,"-1,854556","38,999722",08/05/2025 00:00,1.499,,,1.609,,1.419,,,,,,,0.949,,,,REPSOL,P,OM,L-D: 06:00-22:00,L-D: 06:00-22:00 (A)
9	ALBACETE,ALBACETE,ALBACETE,02005,N,"-1,864917","39,003333",08/05/2025 10:00,1.469,,,1.609,,1.419,1.499,,,,,,,,,,TAMOS,P,dm,L-D: 24H,L-D: 24H (A)
10	ALBACETE,ALBACETE,ALBACETE,02005,N,"-1,859917","39,005083",08/05/2025 00:00,1.469,,,1.655,,1.439,1.509,,,,,,0.984,,,,MOEVE,P,OM,L-D: 06:00-22:00,L-D: 06:00-22:00 (P)
11	ALBACETE,ALBACETE,ALBACETE,02005,D,"-1,884444","39,005528",08/05/2025 15:22,1.379,,,,,1.339,,,,,,,,,,,PLENERGY,P,dm,L-D: 24H,L-D: 24H (D)
12	ALBACETE,ALBACETE,ALBACETE,02005,D,"-1,885361","39,006889",07/05/2025 10:56,1.369,,,,,1.329,1.399,,,,,,0.838,,,,A&A,P,dm,L-S: 09:00-21:30,Sin datos
13	ALBACETE,ALBACETE,ALBACETE,02002,D,"-1,849028","38,989250",07/05/2025 16:16,1.519,,,1.649,,1.459,1.559,,,,,,,,,,CEPSA,P,dm,L-D: 06:30-22:30,Sin datos
14	ALBACETE,AL D: 08:00-23:(D: 08:00-23:00 (A)
15	ALBACETE,AL S: 07:30-15:3 D: 09:00-15:(S: 07:30-15:3 D: 09:00-15:00 (A)
16	ALBACETE,ALBACETE,ALBACETE,02006,I,"-1,892972","39,027778",08/05/2025 00:00,1.499,,,1.609,,1.429,1.489,,,,,,,,,,REPSOL,P,OM,L-D: 24H,L-D: 24H (A)
17	ALBACETE,AL S-D: 09:00-14:00,Sin datos
18	ALBACETE,ALBACETE,ALBACETE,02006,I,"-1,883694","38,988111",06/05/2025 06:00,1.509,,,1.639,,1.409,1.479,,,,,,,,,,REPSOL,P,dm,L-D: 06:00-22:00,L-D: 06:00-22:00 (P)
19	ALBACETE,ALBACETE,ALBACETE,02006,D,"-1,871722","38,994667",05/05/2025 08:58,1.489,,1.529,1.629,,1.419,1.499,,,,,,,,,,REPSOL,P,dm,L-D: 24H,L-D: 24H (A)
20	ALBACETE,ALBACETE,ALBACETE,02006,I,"-2,031861","38,925833",06/05/2025 15:34,1.519,,,1.679,,1.419,1.489,,,,,,,,,,REPSOL,P,dm,L-D: 06:30-21:30,L-D: 06:30-21:30 (P)
21	ALBACETE,ALBACETE,ALBACETE,02006,N,"-1,869889","38,998222",07/05/2025 00:00,1.439,,,,,1.349,1.399,,,,,,,,,,INPEALSA,P,dm,L-D: 07:00-23:00,L-D: 07:00-23:00 (P)
22	ALBACETE,ALBACETE,ALBACETE,02006,N,"-1,866806","38,985194",05/05/2025 16:17,1.409,,,,,1.359,,,,,,,,,,,PETROCAMP,P,dm,L-D: 07:00-23:00,L-D: 07:00-23:00 (A)
23	ALBACETE,ALBACETE,ALBACETE,02006,D,"-1,935500","38,965639",07/05/2025 22:35,,,,,,1.359,,,,,,,,,,,P347,P,dm,L-S: 07:00-22:00,L-S: 07:00-22:00 (A)
24	ALBACETE,ALBACETE,ALBACETE,02006,N,"-1,853306","38,982167",07/05/2025 00:00,1.439,,,,,1.349,,,,,,,,,,,INPEALSA,P,dm,L-D: 07:00-23:00,L-D: 07:00-23:00 (P)
25	ALBACETE,AL D: 00:00-07:00 (D), 07:00-22:00 (P), 22:00-23:59 (D)"
26	ALBACETE,ALBACETE,ALBACETE,02004,D,"-1,919778","38,970167",07/05/2025 08:27,1.489,,,1.549,,1.449,1.519,,,,,,,,,,CEPSA,P,dm,L-D: 24H,L-D: 24H (A)
27	ALBACETE,ALBACETE,SANTA ANA,02328,D,"-1,994028","38,900944",08/05/2025 00:00,1.459,,,1.569,,1.389,1.439,1.1,,,,,,,,,REPSOL,P,OM,L-D: 08:00-16:00,L-D: 08:00-16:00 (A)
28	ALBACETE,AL D: 09:00-14:00,Sin datos
29	ALBACETE,ALBACETE,SALOBRAL (EL),02099,D,"-1,922722","38,870278",29/04/2025 09:45,,,,,,,,0.905,,,,,,,,,CORYMAR,R,dm,L-D: 24H,L-D: 24H (A)
30	ALBACETE,ALBACETE,ALBACETE,02007,D,"-1,888083","39,029583",06/05/2025 00:00,,,,,,,,,,,,,,,1.299,1.239,,NATURGY ES GNV ALBACETE,P,dm,L-D: 24H,L-D: 24H (A)
31	ALBACETE,ALBACETE,ALBACETE,02007,D,"-1,889417","39,012472",05/05/2025 11:14,,,,,,1.279,,,,,,,,,,,AVANZA OIL,P,dm,L-D: 24H,L-D: 24H (A)

Captura de fichero con el formato CSV

2. Aplicación del algoritmo K-Means y primeras observaciones

Una vez preparado el conjunto de datos, comencé a experimentar con el algoritmo de clustering K-Means en Weka. Aunque en un primer momento encontré algunas dificultades técnicas —principalmente relacionadas con el tipo de datos que el algoritmo es capaz de procesar—, pude solventarlas seleccionando exclusivamente aquellas columnas que contenían valores numéricos correspondientes a los precios de los distintos carburantes. El resto de atributos, aunque potencialmente interesantes desde el punto de vista geográfico o comercial, fueron descartados temporalmente para evitar errores en la ejecución.

Con el conjunto ya reducido y depurado, establecí una configuración básica: dos grupos (k = 2) y los parámetros por defecto del algoritmo. Tras varias iteraciones internas, Weka generó una segmentación que repartía las estaciones de servicio en dos grandes conjuntos. El primero agrupaba aproximadamente el 39 % del total de registros, mientras que el segundo comprendía el 61 % restante.



Captura de vista general Weka

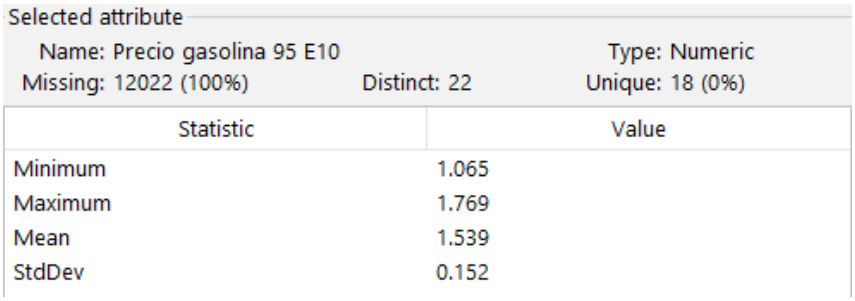
El análisis de los centroides generados por K-Means ofreció resultados reveladores. Algunos carburantes mostraban diferencias significativas entre los dos grupos: la gasolina 95 E5, por ejemplo, presentaba un precio medio claramente más bajo en uno de los clusters. Algo similar ocurría con el gasóleo A, cuyas variaciones apuntaban a la existencia de dos perfiles distintos de estaciones de servicio, probablemente influenciados por factores como la ubicación, la competencia en la zona o el tipo de operador. Aunque ciertos carburantes menos comunes no mostraban grandes diferencias entre grupos, los resultados globales sí reflejaban un patrón consistente: uno de los clusters recogía estaciones con precios más competitivos, mientras que el otro agrupaba aquellas con precios más altos. Esta segmentación, lejos de ser una simple agrupación numérica, abre la puerta a futuras preguntas sobre la lógica económica o geográfica que subyace en el mercado de los combustibles.

3. Provincias y municipios presentes en el conjunto de datos

Al examinar el atributo correspondiente a las provincias en el conjunto de datos, observé que se encuentran representadas las 52 provincias españolas, sin que falte ninguna entrada en esa columna. Esto indica una cobertura territorial completa, lo cual refuerza la representatividad del análisis a nivel nacional.

En cuanto al atributo “Municipio”, el sistema identificó un total de 3.432 municipios distintos. Este dato evidencia la alta granularidad del fichero y refleja la amplia distribución geográfica de las estaciones de servicio. Además, más de 1.500 de esos municipios aparecen una única vez en el conjunto de datos, lo que da cuenta de una dispersión significativa y también de la heterogeneidad en la densidad de estaciones por zona.

4. Precios minimo, máximo y medio de la gasolina 95



Selected attribute

Name: Precio gasolina 95 E10 Type: Numeric

Missing: 12022 (100%) Distinct: 22 Unique: 18 (0%)

Statistic	Value
Minimum	1.065
Maximum	1.769
Mean	1.539
StdDev	0.152

Captura de precios de gasolina 95 e10

Al explorar los diferentes tipos de carburantes presentes en el conjunto de datos, decidí fijarme en el atributo correspondiente a la gasolina 95 E10. Es un tipo de combustible que suelo encontrar habitualmente en las estaciones que visito, por lo que me resultaba familiar y relevante desde una perspectiva personal.

En un primer vistazo, los valores asociados a este carburante parecían estar dentro de un rango esperable: el precio medio se situaba en torno a 1,539 €, con un mínimo registrado de 1,065 € y un máximo que alcanzaba los 1,79 €. Sin embargo, al inspeccionar más a fondo la calidad de los datos, me encontré con una sorpresa bastante llamativa: Weka indicaba que el total de registros para este atributo era nulo. Es decir, aunque el campo existe en el archivo, los 12.022 registros disponibles no contienen valores válidos para esta variable. En la práctica, esto supone que no se dispone de datos reales sobre la gasolina 95 E10 en ninguna estación del fichero.

Esta ausencia completa convierte a este atributo en una variable inutilizable para el análisis, al menos en esta versión del dataset. Resulta un buen recordatorio de la importancia de comprobar la integridad de los datos antes de asumir su utilidad.

5. Selección de atributos con mayor calidad de datos

Antes de proceder al análisis, dediqué un tiempo a revisar qué columnas del conjunto de datos ofrecían una mayor fiabilidad en cuanto a la presencia de valores. La idea era sencilla: si algunos carburantes apenas estaban representados, no tenía sentido incluirlos en un proceso de agrupamiento que se basa precisamente en la comparación numérica.

Revisando los atributos uno a uno en Weka, me centré especialmente en aquellos relacionados con precios. El objetivo era identificar cuáles contaban con la menor cantidad de datos ausentes. Sorprendentemente, el gasóleo B ofrecía una cobertura total, sin ningún valor faltante. Le seguían de cerca el gasóleo A y la gasolina 95 E5, ambos con un número muy reducido de registros incompletos, por debajo del 2 % del total de las más de doce mil estaciones analizadas. La gasolina 98 E5, aunque no perfecta, también presentaba una proporción de valores faltantes lo bastante baja como para ser útil en el análisis.

Estos cuatro atributos fueron finalmente los seleccionados para aplicar el algoritmo de clustering, no solo por su relevancia en términos energéticos, sino porque su consistencia interna aseguraba resultados más fiables y evitaba sesgos derivados de la ausencia masiva de datos.

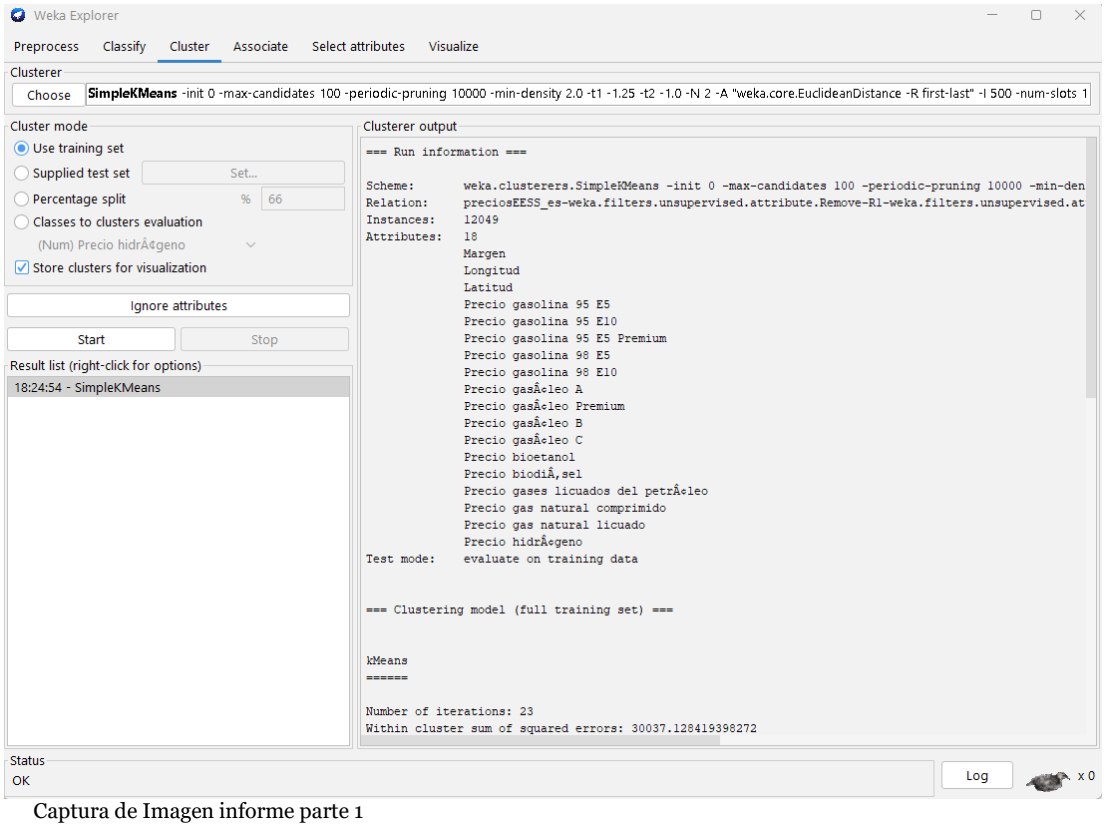
6. Interpretación de los resultados del clustering

Una vez ejecutado el algoritmo K-Means sobre el conjunto depurado —limitado únicamente a los atributos de precio más consistentes—, los resultados no tardaron en revelar una segmentación significativa. El algoritmo agrupó las estaciones de servicio en dos clústeres claramente diferenciados. El primero de ellos, identificado como clúster 0, quedó compuesto por 4.652 instancias. Al analizar los valores promedio de los precios en este grupo, se aprecia que se trata, en términos generales, del segmento más económico. La gasolina 95 E5, por ejemplo, tiene un precio medio inferior a 1,36 €, y el gasóleo A ronda los 1,25 €, lo cual supone una diferencia notable respecto al otro grupo.

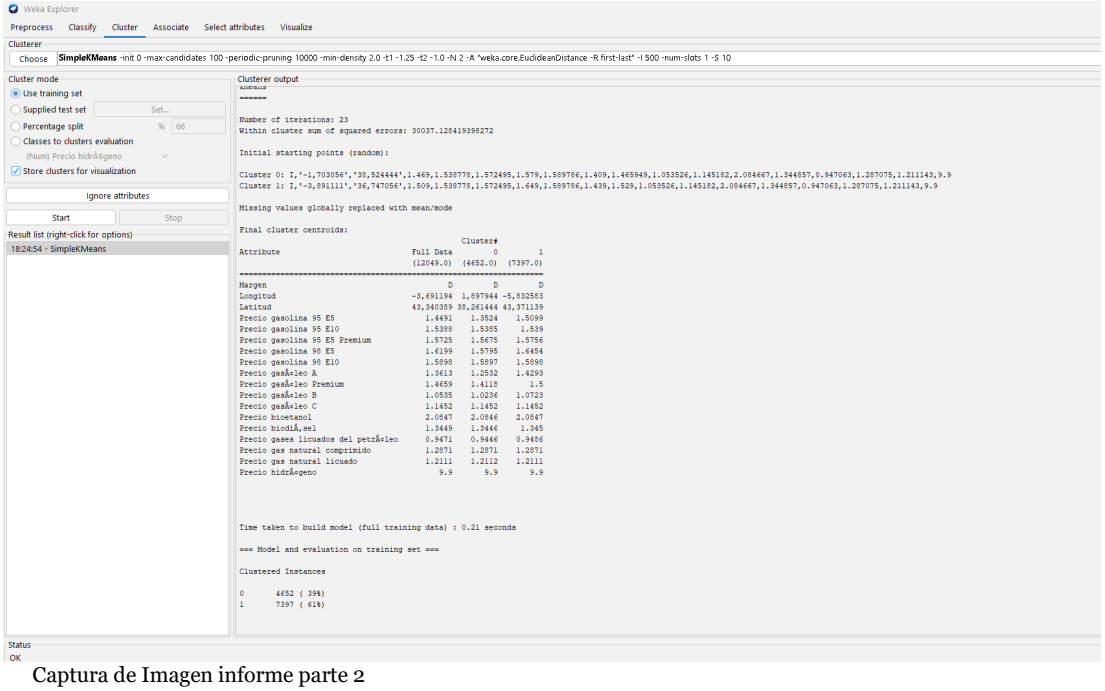
El clúster 1, mucho más poblado con sus 7.397 estaciones, presenta precios sensiblemente más altos en los mismos carburantes: en torno a 1,51 € para la gasolina 95 E5 y alrededor de 1,43 € en el caso del gasóleo A. Esta brecha en los promedios no deja lugar a dudas: estamos ante dos grupos de estaciones de servicio que, aunque no han sido divididos por factores geográficos ni comerciales explícitos, responden a dos perfiles de precio bastante marcados.

Por tanto, aunque no contamos con información contextual como el tipo de operador o la localización exacta, los propios precios permiten intuir una lógica subyacente: hay un grupo de estaciones que

opera con tarifas más competitivas y otro cuyo nivel de precios es claramente más elevado. El clustering no solo ha identificado esta distinción, sino que la ha cuantificado de forma precisa a través de los centroides generados para cada grupo.



Captura de Imagen informe parte 1



Captura de Imagen informe parte 2

7. Visualización e interpretación gráfica del clustering

Tras ejecutar el algoritmo de K-Means en Weka, utilicé la opción que permite almacenar los clústeres para su representación gráfica. Esto me abrió la posibilidad de observar, de manera visual, cómo se habían agrupado las estaciones de servicio según los precios de los carburantes. Lo interesante de este paso no fue solo generar un gráfico, sino explorar activamente qué combinaciones de variables ofrecían una mejor lectura del agrupamiento.

Fui modificando las variables asignadas a los ejes para entender mejor la lógica de segmentación generada por el algoritmo. En uno de los ejemplos, al cruzar el precio de la gasolina 95 E10 con el del gasóleo A, el resultado fue bastante revelador: los puntos se distribuían claramente en dos zonas bien diferenciadas por color, lo que confirmaba visualmente lo que ya sugerían los centroides numéricos.

Esta separación entre grupos, observable en diferentes combinaciones de atributos, reforzaba la idea de que el algoritmo no solo estaba funcionando correctamente desde el punto de vista técnico, sino que también tenía sentido desde una perspectiva interpretativa. Se confirmaba así que el clustering no era un artificio estadístico, sino una herramienta eficaz para identificar dos realidades distintas en el comportamiento de precios de las estaciones: una más ajustada al bolsillo del consumidor, y otra más elevada, probablemente por razones de localización, marca o política comercial.

8. Evaluación del modelo de clasificación

Una vez aplicada la técnica de agrupamiento, decidí comprobar hasta qué punto los grupos generados podían ser reconocidos por un modelo de clasificación supervisada. Para ello, utilicé el algoritmo J48, que permite construir árboles de decisión a partir de un conjunto de datos etiquetado.

El resultado fue bastante satisfactorio: el modelo alcanzó una tasa de acierto cercana al 98 %, lo que sugiere que los clústeres creados previamente siguen una lógica interna sólida y bien diferenciada.

La precisión obtenida no fue uniforme, pero sí alta en ambos grupos. En el clúster identificado como el más económico, el número de aciertos superó las 7.000 instancias correctamente clasificadas, con un margen mínimo de error. En el otro grupo, correspondiente a las estaciones con precios más elevados, también se logró un alto nivel de exactitud, aunque con algunos casos puntuales que se confundieron con el grupo contrario.

La matriz de confusión generada por Weka recoge estos detalles y permite apreciar el equilibrio general del modelo. Aunque siempre existe un pequeño margen de error en este tipo de tareas, lo importante es que los fallos fueron residuales y no comprometen la validez de la clasificación. En conjunto, el árbol J48 confirmó que los patrones identificados en el análisis no eran aleatorios, sino consistentes y replicables.

9. Conclusion

Esta práctica me ha permitido ir más allá de una simple ejecución técnica y acercarme al proceso completo de análisis de datos con una mirada más crítica. Trabajar con un conjunto real de estaciones de servicio en España no solo me ha servido para familiarizarme con herramientas como Weka, sino también para comprender la importancia del preprocesamiento, la limpieza del dataset y la elección adecuada de atributos, elementos clave que a menudo se dan por sentados.

La aplicación del algoritmo K-Means ha sido especialmente reveladora: con una configuración básica, ha logrado identificar diferencias significativas en los precios, segmentando las estaciones en dos perfiles bien definidos. Más adelante, al aplicar el clasificador J48, he podido comprobar que esas agrupaciones no eran aleatorias, sino que respondían a patrones consistentes y reconocibles. Ver que el modelo clasificaba con casi un 98 % de acierto no fue solo un dato técnico, sino una confirmación de que todo el proceso anterior había seguido una lógica coherente.

Más allá de los resultados, lo más valioso ha sido descubrir cómo las decisiones que tomamos durante el análisis —qué columnas conservar, cómo interpretar una diferencia de precio, qué métricas considerar relevantes— son las que realmente dan sentido a los algoritmos. Esta experiencia ha reforzado mi comprensión del análisis de datos y me ha hecho más consciente del papel que juega el criterio humano dentro de un entorno automatizado.