Minería de datos sobre un conjunto de datos de coches

IKPD

Noa Yu Ventura Vila
Javier Abella Nieto
Albert Bausili Fernández
Carlos Andrés Rodríguez Torres
Juan José Acevedo Serna
René Alonso Cortés López

Q1-Otoño 24/25

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Contents

1 Documentación previa

1.1 Fuente, descripción de los datos y motivación

En esta práctica hemos utilizado un dataset que hemos encontrado en kaggle. El dataset en cuestión nos proporciona información sobre los combustibles y características de diversos modelos de coches y su eficiencia.

www.kaggle.com/datasets/arslaan5/explore-car-performance-fuel-efficiency-data

Entre los datos del dataset podemos encontrar: información del vehículo, sus especificaciones técnicas, unas métricas de eficiencia energética y sus emisiones de CO2. En la figura 1 podemos ver algunos de los campos de las primeras 5 filas del dataset.

	city_mpg	class	combination_mpg	cylinders	displacement	drive	fuel_type	highway_mpg	make	model	transmission	year
0	25	midsize car	29	4.0	2.5	fwd	gas	36	mazda	6	m	2014
1	26	midsize car	30	4.0	2.5	fwd	gas	37	mazda	6	a	2014
2	25	small sport utility vehicle	27	4.0	2.5	fwd	gas	31	mazda	cx-5 2wd	a	2014
3	26	small sport utility vehicle	29	4.0	2.0	fwd	gas	34	mazda	cx-5 2wd	m	2014
4	26	small sport utility vehicle	28	4.0	2.0	fwd	gas	32	mazda	cx-5 2wd	a	2014

Figure 1: Primeras 5 lineas del dataset original

En cuanto a nuestra motivación del uso de este dataset en específico, nosotros hemos elegido este dataset porque ofrece una combinación intrigante y suficiente de variables numéricas, como el consumo de combustible y la potencia del motor, junto con variables categóricas como la marca y el tipo de transmisión, lo que lo hace ideal para realizar análisis exploratorios. Además, el tamaño del conjunto de datos es suficiente para obtener resultados representativos y confiables. Por otro lado, el tema es tan interesante como relevante hoy en día, ya que entender el rendimiento y la eficiencia de los automóviles puede ayudar a abordar cuestiones de sostenibilidad y optimización energética.

1.2 Planificación inicial

1.2.1 Diagrama Gantt

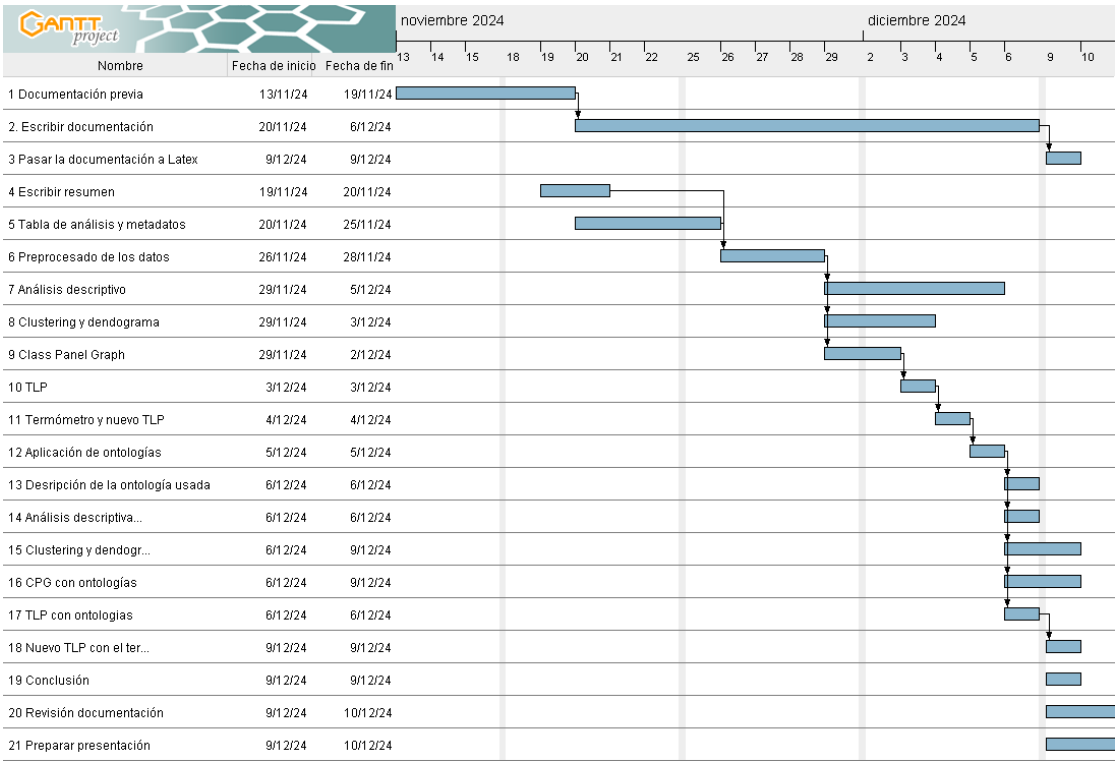

Figure 2: Diagrama de Gantt

1.2.2 División del trabajo

En cuanto a la división del trabajo hemos decidido dividirlo por secciones, siendo cada uno responsable de hacer uno de los apartados del documento y otro responsable de su documentación. Puedea

#tarea	**Tarea principal**	**Documentación**
Tarea 1	Javier Abella	Javier Abella
Tarea 2	Todos	Todos
Tarea 3	Javier Abella, Noa Yu Ventura	Javier Abella, Noa Yu Ventura
Tarea 4	Noa Yu Ventura	Noa Yu Ventura
Tarea 5	René Alonso Cortés	René Alonso Cortés
Tarea 6	Albert Bausili, Carlos Andrés	Carlos Andrés, Noa Yu Ventura

Tarea 7	Albert Bausili	Juan José
Tarea 8	Albert Bausili	Noa Yu Ventura
Tarea 9	Albert Bausili	Noa Yu Ventura
Tarea 10	Albert Bausili	Noa Yu Ventura
Tarea 11	Albert Bausili, Noa Yu Ventura	Noa Yu Ventura
Tarea 12	Carlos Andrés	Juan José, Noa Yu Ventura, Albert Bausili
Tarea 13	Noa Yu Ventura	Noa Yu Ventura
Tarea 14	Albert Bausili	Noa Yu Ventura, Juan José
Tarea 15	Albert Bausili	Noa Yu Ventura
Tarea 16	Albert Bausili	Noa Yu Ventura
Tarea 17	Albert Bausili	Noa Yu Ventura
Tarea 18	Albert Bausili	Noa Yu Ventura
Tarea 19	Noa Yu Ventura, Albert Bausili	Noa Yu Ventura, Albert Bausili
Tarea 20	Javier Abella, Noa Yu Ventura, Albert Bausili	Javier Abella, Noa Yu Ventura, Albert Bausili, Carlos Andrés
Tarea 21	Todos	Todos

2 Tabla de metadatos

El dataset contiene información sobre el rendimiento y características de vehículos. Consta de 550 registros y 12 columnas. Cada fila representa un modelo de vehículo específico, cada columna proporciona información como la eficiencia de combustible, tipo de transmisión, entre otras. Registros: 550 Columnas: 12 Datos faltantes: Valores nulos (cylinders y displacement).

Atributo	Modalidades	Descripción	Tipo (categoría)	Tipo (dato)	Unidad	Missing	Rango
city_mpg	N/A	Millas por galón en ciudad	Numérico	int64	mpg	0	11 - 126
class	Ej. ['midsize car', 'small sport utility vehicle', ...]	Clase del vehículo	Categórico	string	N/A	0	N/A
combination_mpg	N/A	Promedio de millas por galón	Numérico	int64	mpg	0	14 - 112
cylinders	N/A	Número de cilindros	Numérico	float64	Unidad	2	3.0 - 12.0
displacement	N/A	Cilindrada del motor	Numérico	float64	Litros	2	1.2 - 6.8
drive	['fwd', '4wd', 'rwd', 'awd']	Tipo de tracción	Categórico	string	N/A	0	N/A
fuel_type	['gas', 'diesel', 'electricity']	Tipo de combustible	Categórico	string	N/A	0	N/A
highway_mpg	N/A	Millas por galón en carretera	Numérico	int64	mpg	0	18 - 102
make	Ej. ['mazda', 'ford', 'subaru', 'nissan', 'audi', ...]	Marca del vehículo	Categórico	string	N/A	0	N/A
model	Ej. ['6', 'cx-5 2wd', 'cx-5 4wd', 'mustang', 'forester awd', ...]	Modelo del vehículo	Categórico	string	N/A	0	N/A
transmission	['m' (manual), 'a' (automático)]	Tipo de transmisión	Categórico	string	N/A	0	N/A
year	N/A	Año de fabricación	Numérico	string	Años	0	2014 - 2024

Figure 3: Tabla que describe todos los tipos de variables y sus dominios.

Hay seis tipos de variables categóricas y seis tipos de variables numéricas, completando así todos los requerimientos de este proyecto para el tipo de variables usadas en el dataset.

3 Preprocesado de los datos

3.1 Renombrado de columnas

Las columnas originales del dataset fueron renombradas para mejorar la claridad, estandarización y compatibilidad con herramientas de análisis. Esto también evitó problemas asociados a espacios o caracteres especiales en los nombres de las variables. Por ejemplo:

- `Unnamed: 0` → `ID`
- `city_mpg` → Miles per Gallon (City)
- `class` → Vehicle Class
- `combination_mpg` → Miles per Gallon (Combined)
- `cylinders` → Cylinders
- `displacement` → Displacement
- `drive` → Drive
- `fuel_type` → Fuel Type
- `highway_mpg` → Miles per Gallon (Highway)
- `make` → Brand
- `model` → Model
- `transmission` → Transmission
- `year` → Year

3.2 Simplificación de Valores en las Columnas

Los valores textuales en ciertas columnas, como Vehicle Class y Model, fueron simplificados para mejorar su legibilidad y facilitar el análisis posterior.

Por ejemplo:

- `f-type v8 s convertible` → `F-Type V8 Conv`
- `john cooper works convertible` → `Cooper JCW Conv`
- `range rover evoque` → `Evoque`

- `small sport utility vehicle` → `SUV Small`
- `compact car` → `Compact`

Este cambio no solo estandarizó los valores, sino que también redujo la complejidad de las etiquetas al mantener su identidad principal.

3.3 Capitalización y Limpieza de Caracteres Problemáticos

Se capitaliza la primera letra de cada palabra en todas las columnas textuales para mantener consistencia en el formato. Esto fue especialmente útil en nombres de modelos y clases de vehículos. Además, se eliminaron caracteres problemáticos como espacios, apóstrofes, asteriscos, y barras (', *, /), asegurando compatibilidad con herramientas como KLASS.

3.4 Manejo de valores atípicos (outliers)

Se incluyeron únicamente vehículos que usan gasolina (FuelType = gas) para alinear el análisis con el alcance del proyecto. Vehículos eléctricos y diésel fueron excluidos para reducir ruido y enfocarse en la categoría principal de interés, ya que al tener cero cilindros y cero displacement alteran mucho los resultados.

Los outliers de diesel y eléctricos no los hemos eliminado del dataset, hemos escogido una submatriz sin tener en cuenta estos outliers como nuevo input de datos para Klass. El principal motivo por el que hemos decidido no tenerlos en cuenta es porque son outliers de tipo subpoblación, tenemos demasiado pocos datos sobre este tipo de vehículos (dos eléctricos y dos diésel) como para sacar conclusiones claras y significativas. Como hemos visto en clase, el método para tratar este tipo de outliers es simplemente no tenerlos en cuenta, porque son individuos de otras poblaciones. Además hemos comprobado, que reducían drásticamente la claridad del CGP, imposibilitaba sacar un TLP con sentido y al usar el termómetro aumentaban considerablemente la variabilidad.

Otros outliers menos significativos que podrían afectar la calidad del análisis son:

- **Highway, City y Combination MPG:**
  Se detectaron dos vehículos con valores de `Highway_MPG` de **99** y **102**, considerablemente fuera del rango esperado para la eficiencia en carretera. Estos valores se consideran *outliers* del tipo **valor extremo intrínseco**, ya que

superan los límites para vehículos convencionales pero siguen siendo posibles físicamente. Estos datos fueron marcados como atípicos para su documentación, pero tratados como valores válidos para el dataset.

- **Valores Faltantes:**
  En las columnas `Cylinders` y `Displacement` de los coches eléctricos (`FuelType=Eléctrico`) se identificaron valores faltantes (`NA`). Estos valores fueron tratados mediante *imputación*, poniendo un **0** como reemplazo. Esto es debido a que no tiene ningún sentido decir que un coche eléctrico tiene cilindros o centímetros cúbicos.

4 Análisis descriptivo univariante

El análisis descriptivo del dataset brinda información sobre las características de 550 modelos de automóviles. En términos de eficiencia de combustible para conducción urbana (city mpg), la media es de 21.46 millas por galón (mpg), mientras que la mediana se sitúa en 20.0 mpg. Con una desviación estándar de 8.15 mpg y un rango que varía desde 11 mpg hasta un máximo de 126 mpg. En la Figura 3 se observa la distribución de esta variable, se pueden determinar claramente dos atípicos con valores de 126 y 121 mpg correspondientes a los vehículos eléctricos (subcompact car).

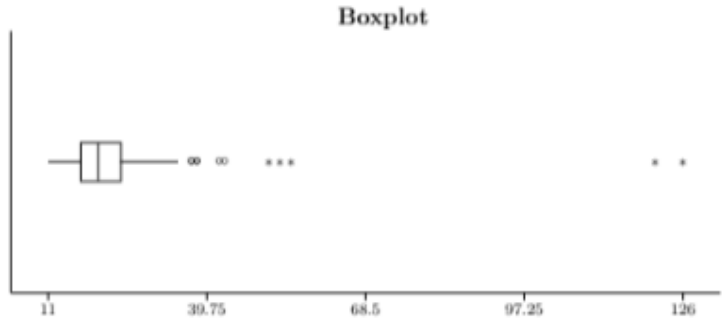

Figure 4: Boxplot que representa la distribución de datos para la variable CityMPG. Útil para identificar outliers.

En cuanto a la eficiencia (**highway mpg**), el promedio es de 28.61 mpg, con una mediana de 28.0 mpg y una moda de 24 mpg. Con una desviación estándar de 6.83 mpg y un rango de 84 mpg, desde 18 mpg hasta un máximo de 102 mpg. En

la Figura 5, se observa de igual manera estos valores atípicos correspondientes a los vehículos eléctricos y se sale de la eficiencia media de los vehículos de combustión.

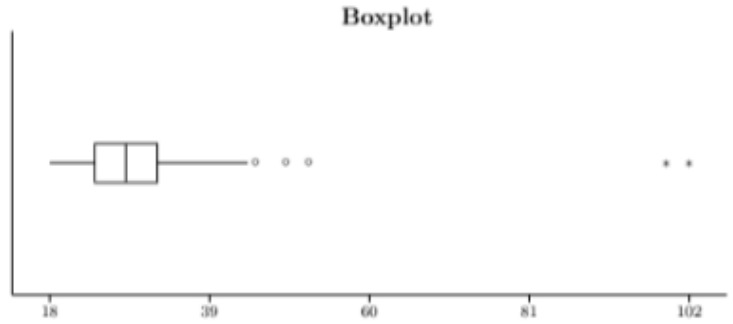

Figure 5: Boxplot que representa la distribución de datos para la variable HighwayMPG. Útil para identificar outliers.

La eficiencia combinada (**combination mpg**), tiene una media de 24.07 mpg, una mediana de 23.0 mpg y una moda de 22 mpg. Los valores oscilan entre un mínimo de 14 mpg y un máximo de 112 mpg, con una desviación estándar de 7.48 mpg. De igual manera existe la presencia de los valores atípicos que se pueden ver en la Figura 5.

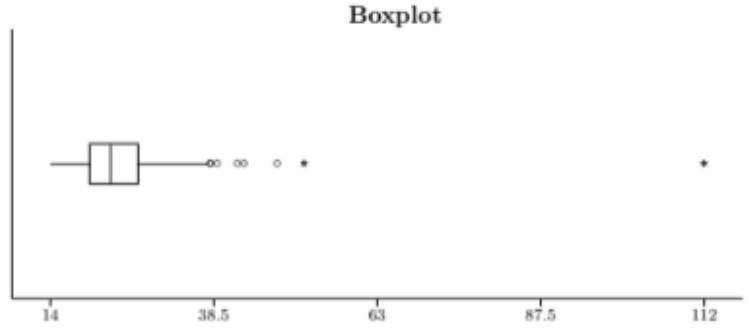

Figure 6: Boxplot que representa la distribución de datos para la variable CombinationMPG. Útil para identificar outliers.

Las especificaciones de los motores (**cylinders and displacement**). Tiene una media de 2.93 litros, con una mediana de 2.5 litros y una moda de 2.0 litros. El rango abarca desde un mínimo de 1.2 litros hasta un máximo de 6.8 litros, con

una desviación estándar de 1.25 litros. En la Figura 7 podemos observar algunos valores atípicos, uno de ellos son los automóviles que no disponen de cilindros, correspondiente a los de motor eléctrico.

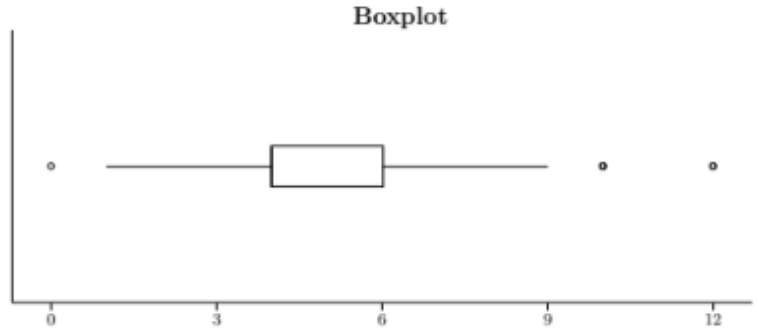

Figure 7: Boxplot que representa la distribución de datos para la variable Cylinders y Displacement. Útil para identificar outliers.

En cuanto al tipo de tracción (drive), se observa que la tracción total (AWD) es la más común, representando 215 vehículos. Le siguen la tracción delantera (FWD) con 178 vehículos, la tracción trasera (RWD) con 115 y, en menor medida, la tracción 4WD con 42 modelos.

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Fwd	178	178	0.3236	0.3236
4wd	42	220	0.0764	0.4
Rwd	115	335	0.2091	0.6091
Awd	215	550	0.3909	1
dades mancants	0	N = 550	0	

Figure 8: Frecuencias para el tipo de tracción que tienen los coches.

Respecto al tipo de combustible (fuel type), el dataset está dominado por vehículos que utilizan gasolina, con 546 modelos en esta categoría. Además, se identificaron 2 modelos que funcionan con diésel y otros 2 eléctricos.

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Gas	546	546	0.9927	0.9927
Diesel	2	548	0.0036	0.9964
Electricity	2	550	0.0036	1
dades mancants	0	N = 550	0	

Figure 9: Frecuencias para el tipo de energía que utiliza el coche.

Entre los fabricantes (**make and models**) más representados, destacan BMW (72 modelos), Jaguar (71 modelos) y Kia (69 modelos). Marcas como Mini y Hyundai también tienen una presencia significativa con 51 y 38 modelos, respectivamente. Por otro lado, fabricantes como Chrysler, Infiniti y Porsche cuentan con una representación mucho más limitada, con solo uno o dos modelos cada uno.

El rango de años de fabricación (**year**) en el dataset va de 2014 a 2024, con una media y una mediana de 2019. El año más frecuente (moda) es 2014. La desviación estándar de 3.17 años.

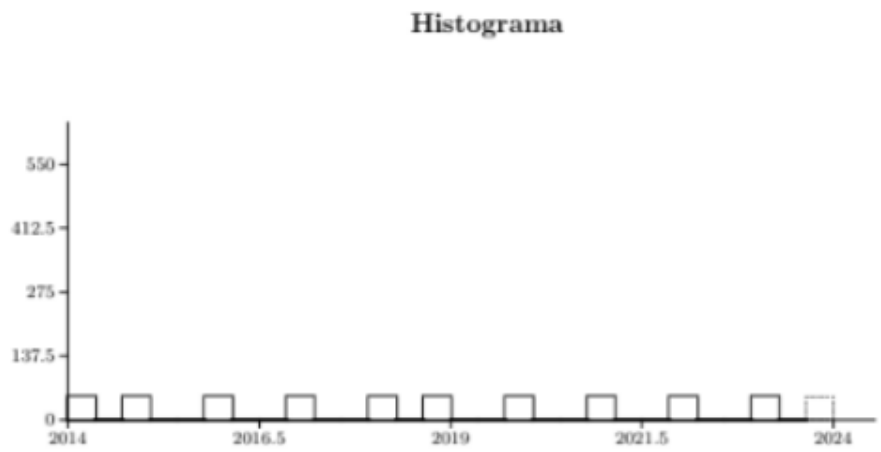

Figure 10: Histograma que representa la distribución de datos para la variable Year.

5 Clustering en Klass

5.1 Dendograma sin corte

Figure 1: CAJ. Arbre general de classificació

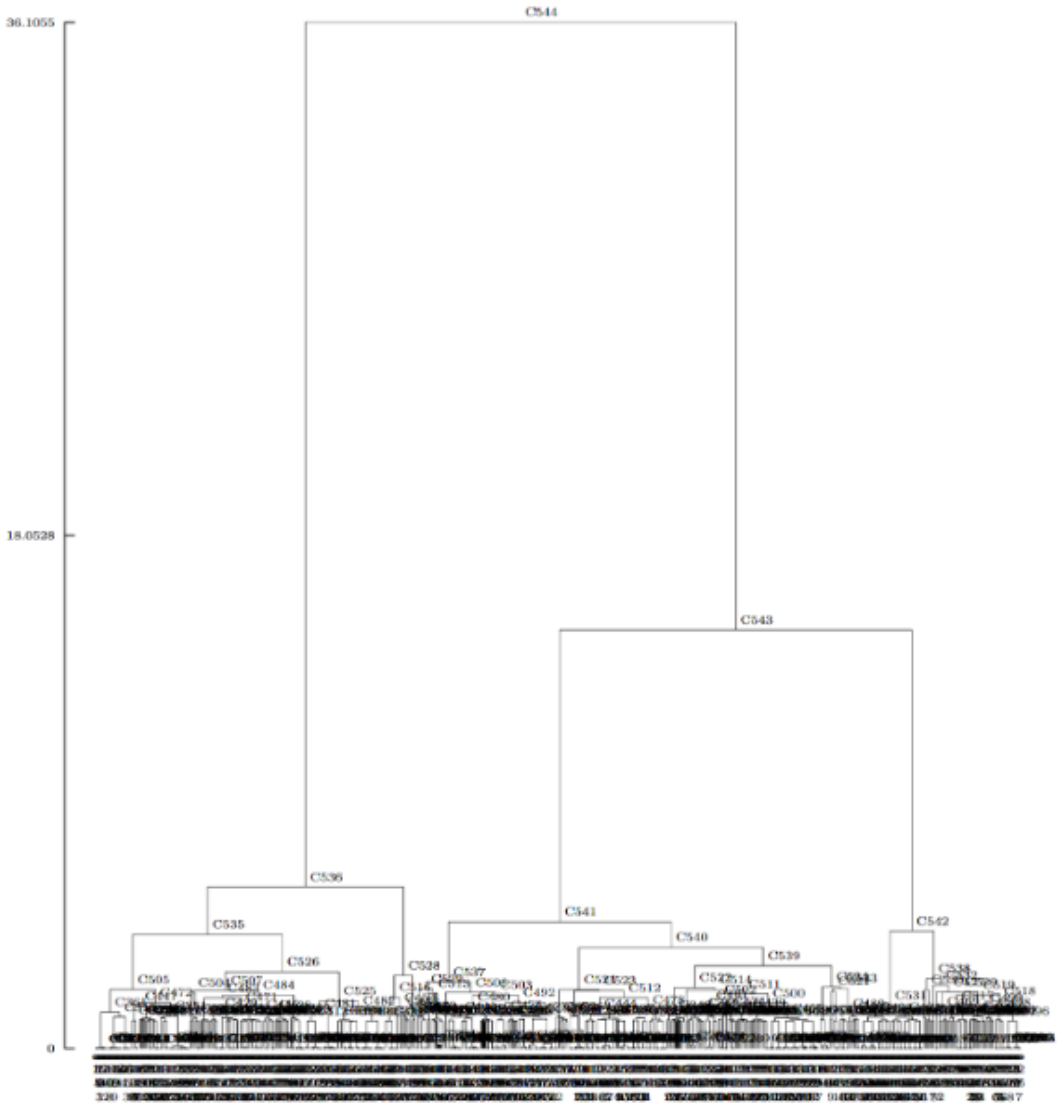

Figure 11: Dendograma resultante de aplicar la estrategia de vecinos recíprocos.

Este es el dendograma resultante de aplicar la estrategia de vecinos recíprocos encadenados usando la distancia mixta de Gibert, que siguiendo el criterio visto en clase el dendograma lo cortamos por debajo de C543, ya que es el trozo más largo sin contar el primero. Esto nos deja con 3 categorías. Podemos afirmar que ha salido correctamente porque las clases salen todas bien definidas y hay una gran diferencia entre ellas.

5.1.1 Dendograma con corte

Hemos decidido cortar por esa rama en específico porque consideramos que la diferencia que hay entre las clases de arriba del corte y de abajo son muy sustanciales, y el gráfico lo representa muy bien. Además, concuerda con lo que nos han enseñado en clase. Por tanto, nos quedamos con 3 clases principales: C536, 541 y 542.

Figure 1: CAJ. Arbre general de classificació tallat en 3 classes

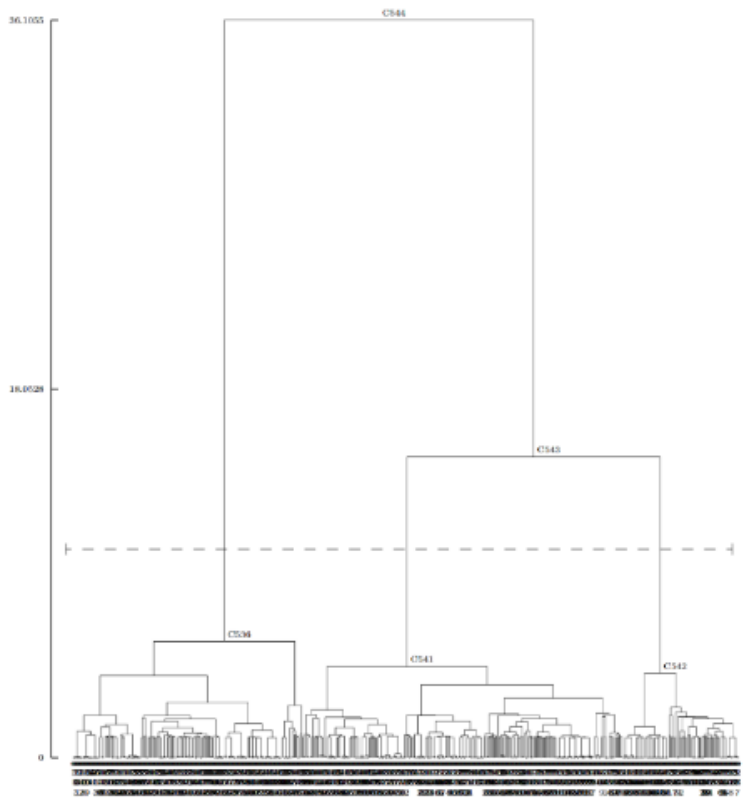

Figure 12: Dendograma con el corte marcado a C543.

6 Class Panel Graph

Podemos ver en la figura 13 el CPG resultante. Cabe destacar que el Klass proporciona un fichero output con más información y con más datos, pero hemos escogido estas figuras porque son las que consideramos más significativas para nuestro análisis.

Como podemos ver, tenemos las 3 clases del apartado anterior en distintos gráficos según ciertos parámetros. Según lo que podemos observar, los que han influenciado más en esta clasificación han sido CityMPG (en la ciudad Miles Per Gallon), CombinedMPG (Combined Miles Per Gallon), Cylinders, Displacement y HighwayMPG (en carretera Miles Per Gallon). Esto lo podemos decir porque estos parámetros se diferencian muy bien entre las tres clases, todos tienen un valor bastante distinguido.

- La clase **C536** tiene el mejor `CityMPG`, y a pesar de tener el peor `HighwayMPG`, consigue que la combinación (`CombinedMPG`) sea mayor que el resto de clases. Al ser la clase que tiene menor número de cilindros y menor capacidad de centímetros cúbicos a priori nos hace pensar que puede ser la mejor clase en cuanto a eficiencia. Otra cosa que la caracteriza y diferencia respecto al resto es su naturaleza por el modo de tracción (`Drive`), es la clase que tiene tracción delantera (`FWD`). Este tipo de tracción se suele usar en coches con potencia baja o media, de modo que nos confirma que es un coche eficiente, ya que es un motor con poca potencia pero te permite recorrer mucha distancia con menos litros de gasolina.

- La clase **C542** es la que tiene tanto peor `City` como `Highway MPG` (y por tanto, la combinada). En contraste, es la que tiene más cilindros y más capacidad en centímetros cúbicos (parámetro `displacement`). Esto nos hace pensar que el antiguo mito de cuántos más cilindros y centímetros cúbicos mejor es en realidad esto: un mito. Con estos datos estamos viendo que en este caso es todo lo contrario. Otro dato interesante sobre esta clase es que son coches muy antiguos o muy nuevos según el parámetro `Year`, de modo que nos hace entender que ahora vuelve a haber una tendencia que existió hace 10-11 años.

- La clase **C541** parece un intermedio entre las anteriores, está entre las otras dos clases en todos los parámetros `MPG`, en el número de cilindros y el desplazamiento. Parece ser el estándar de vehículos de media-alta gama que intenta abarcar todo a un nivel decente pero no sobresale en ninguno.

Los parámetros que no han sido mencionados anteriormente no aportan ninguna

información significativa y no podemos sacar ninguna conclusión de éstos debido a lo similares que son los gráficos comparados con el resto de clases.

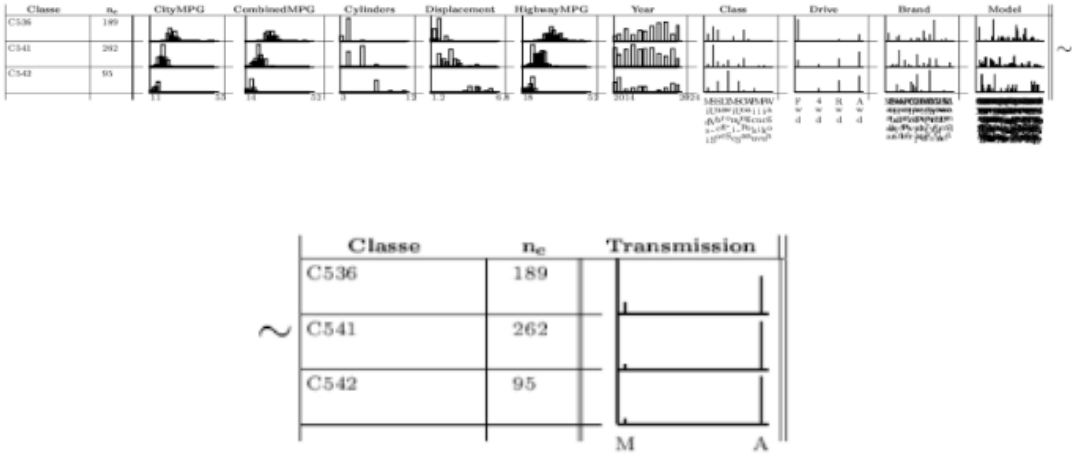

Figure 13: Class Panel Graph con las tres clases definidas anteriormente.

7 Traffic Light Panel

En base al CPG anterior hemos sido capaces de hacer el TLP correspondiente como se ve en la figura 14. Cada gráfico del TLP "normal" lo hemos pintado según nuestro propio criterio y conocimiento, pero evidentemente en un caso real se debería contactar con un experto para que nos dé información aún más concisa, profesional e imparcial.

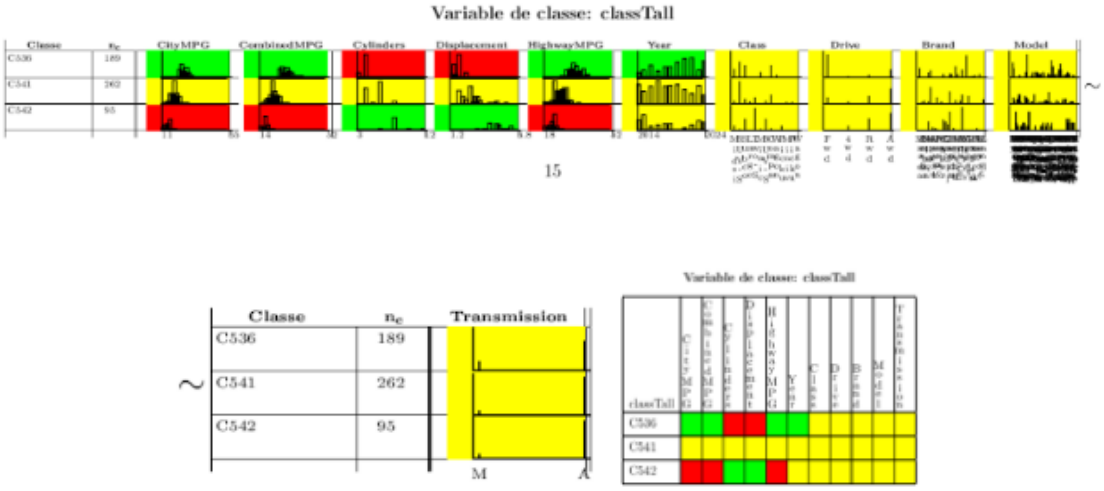

Figure 14: TLP pintado "a ojo" y TLP resumido en forma de mosaico

Criterios para pintar el TLP:

- `CityMPG`, `HighwayMPG`, `CombinedMPG`, Cylinders y Displacement: la clase con más *Miles Per Gallon* (millas por galón) es la que pintamos de verde, la con menos es la roja y la intermedia es la amarilla.
- `Year`: hemos supuesto que **como más nuevo es el coche mejor es**, ya que la tecnología va avanzando muy rápido.
- `Drive`: el tipo de tracción no suele influenciar en cuánto de bueno es el coche objetivamente, ya que siempre va a depender de la situación. Por eso están todos amarillos.
- `Class`, `Brand` y `Model`: todos estos coches varían mucho en estética y funcionalidad, no se puede decir que uno es mejor que otro ya que es estrictamente subjetivo.

- `Transmission`: aunque generalmente pensemos que sea mejor un motor automático, el manual nunca va a dar problemas porque no tiene sistema eléctrico de por medio, y en caso de falla estamos asegurados de tener control sobre él. Por esto mismo también consideramos que es algo subjetivo y depende del uso que le dé el usuario.

En la figura 15 podemos apreciar el mismo TLP pero representando las varianzas (el annotated tiene colores más oscuros si la varianza es más alta, y más claros si es más baja).

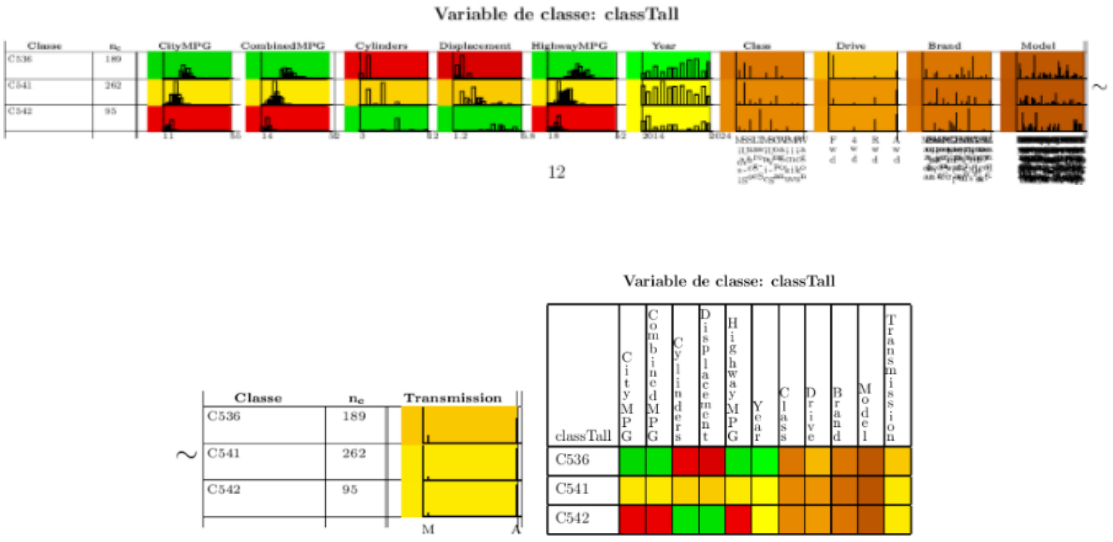

Figure 15: Figuras similares a las de la figura 14 pero con otros colores distintos debido a que ahora la diferencia entre TLPs representa la varianza.

Los colores más oscuros están en la parte derecha del mosaico, y esto nos indica que hay una variancia medio alta. Esto significa que los resultados obtenidos en esos gráficos no son del todo fiables y por lo tanto no han influido tanto en la decisión de poner un objeto en una clase u otra.

8 Termómetro y nuevo Traffic Light Panel

En la figura 17 podemos ver que ahora el TLP tiene otros colores respecto a la 14, y esto es debido a que ahora hemos pintado el TLP con un termómetro. A este termómetro le hemos asignado los siguientes valores que se pueden apreciar en la figura 16. Para determinar estos límites hemos usado todo nuestro conocimiento y criterio propio sobre coches, algunos artículos e incluso experiencias personales con ellos. Tal y como hemos dicho en la sección anterior, en un caso real sería necesario contratar a un experto para que nos guíe más adecuadamente en la interpretación de los datos.

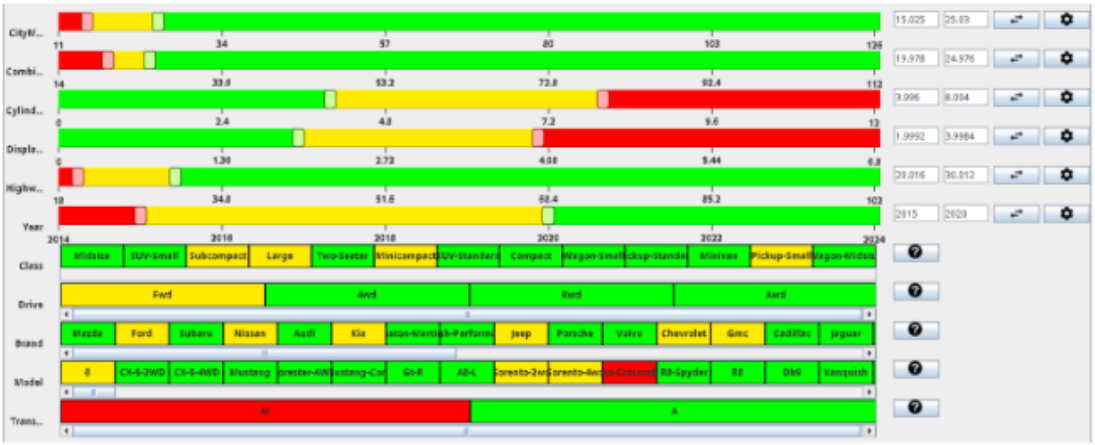

Figure 16: Valores dados al termómetro según nuestro propio conocimiento

Como hemos mencionado anteriormente, aplicando el termómetro obtenemos un nuevo TLP, el cual está representado en la agrupación de la figura 17.

Si comparamos este TLP con termómetro con el anterior TLP sin termómetro podemos notar una diferencia en algunos atributos. Los más notorios son el Displacement y el Cylinder, ya que parece ser que tener más cilindros y más centímetros cúbicos, y por tanto más potencia, no lo hace mejor vehículo. En este nuevo TLP también hay colores más claros y no tan rojos, haciéndonos pensar que al inicio hemos sido muy exigentes y duros con el valor de los parámetros.

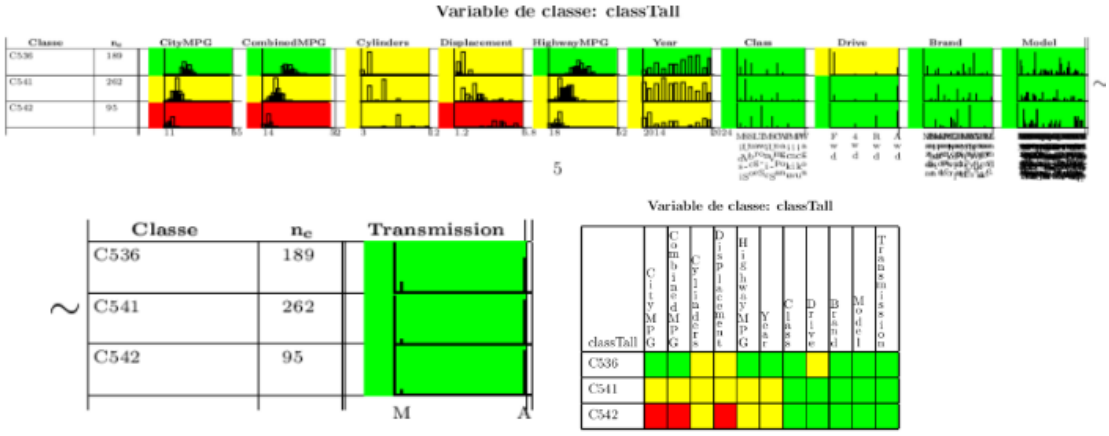

Figure 17: Resultado de aplicar el termómetro de la figura 13 al CPG original

Finalmente nos hace falta ver cuán preciso es el nuevo TLP con termómetro. Como podemos ver, las variables categóricas vuelven a ser de color oscuro indicando una gran variancia y, por lo tanto, volviendo a reforzar el hecho de que las variables categóricas no han ayudado demasiado a la distribución de las clases.

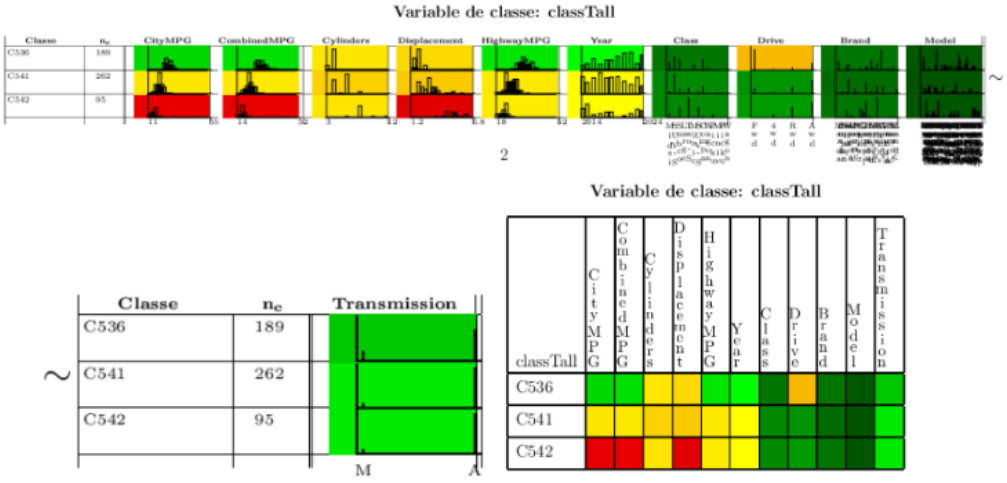

Figure 18: Nuevo TLP con termómetro normal y annotated.

9 Ontologías

Hasta este punto hemos hecho un clustering con datos sin reglas, hemos sacado diversos gráficos que hemos analizado en profundidad y sacado varias hipótesis para poder explicar el comportamiento de estos datos. Estas hipótesis que hemos obtenido son "a priori", queriendo decir que el programa Klass que hemos utilizado tenía total desconocimiento de los datos que le hemos entrado.

Para usar más eficientemente el algoritmo de clustering y obtener resultados con una base más sólida hemos decidido utilizar ontologías, un método de definición de relaciones entre entidades que son reales y se fundamentan en un dominio determinado (en este caso, para una de nuestras variables que hay en los datos).

La metodología que hemos seguido para hacerlo es la misma que hasta ahora, hemos tratado los outliers de la misma manera (para más información ir a la sección 3.4), hemos hecho un clustering con su correspondiente corte y así dando lugar al CPG, de ahí lo hemos pintado para obtener los dos TLPs (el "normal" y el annotated) y con el mismo termómetro que la sección 8 hemos obtenido los dos nuevos TLPs. Esta sección no se centra tanto en el análisis de gráficos como hemos visto en secciones anteriores, sino más en destacar qué diferencias ha habido entre usar las ontologías y no usarlas, y así poder determinar cuán de eficiente han sido las que hemos creado.

9.1 Ontología usada

La ontología que hemos usado se representa en la figura 19. La lógica detrás de ésta está en relacionar distintas variables del modelo de vehículo con el tipo de vehículo que es. Esto es posible debido a que hay una serie de estándares arraigados al sector automovilístico con el objetivo de vender a una audiencia en concreto, y éstos permiten agrupar modelos de coches de distintas marcas en base a una serie de características concretas. Por ejemplo, el Porsche 911 y el BMW Serie 4 son coches conocidos por lo caros que son y entendidos como coches de lujo que están reservados para la alta clase, de hecho, incluso la marca en sí ya tiene esa fama. Así, obtenemos el primer nivel del árbol que relaciona una característica común de un grupo de modelos de coches con el modelo en concreto.

Para el segundo nivel hemos agrupado de nuevo qué tipo de coches son, pero ahora buscando características más generales aún. Por ejemplo, una pick-up es bastante distinta a un coche habitual, ya que no tiene asientos de atrás y se pueden llevar cargas grandes. Esta es una característica muy general y básica de las pick-ups que las distingue del resto de coches. Lo mismo hacemos para el SUV, son

coches que últimamente están muy de moda y son caracterizados por ser coches altos, estilizados y sumamente caros.

Sobre el tercer nivel no hay mucho más a explicar, todos estos coches son un tipo de vehículo, de modo que los juntamos todos en una variable que se llama Vehicle. Ponerle este nombre a la variable, además, nos permite en un futuro ampliar de manera sencilla el árbol para incluir vehículos que no sean coches, ya que sería tan fácil como crear una nueva rama llamada Coche y poner las ramas ya existentes en esta nueva.

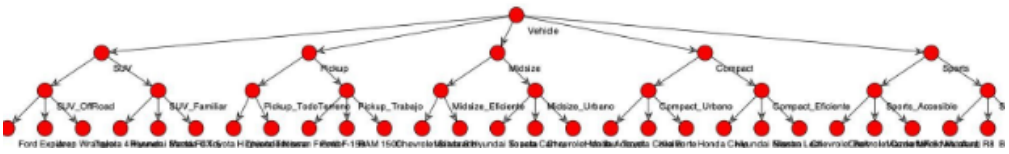

Figure 19: Representación gráfica de la ontología usada correspondiente al fichero CarKlass13-Sub.ont.

9.2 Análisis descriptiva univariante

El análisis descriptivo del dataset teniendo en cuenta la ontología, brinda información sobre las características de 546 modelos de automóviles. En términos de eficiencia de combustible para conducción urbana (city mpg), la media es de 21.07 millas por galón (mpg), mientras que la mediana se sitúa en 20.0 mpg. Con una desviación estándar de 5.33 mpg y un rango que varía desde 11 mpg hasta un máximo de 55 mpg. A diferencia del análisis estadístico anterior, ya no se encuentran los vehículos eléctricos y por lo tanto ahora el máximo de eficiencia en ciudad viene dado por vehículos con combustión.

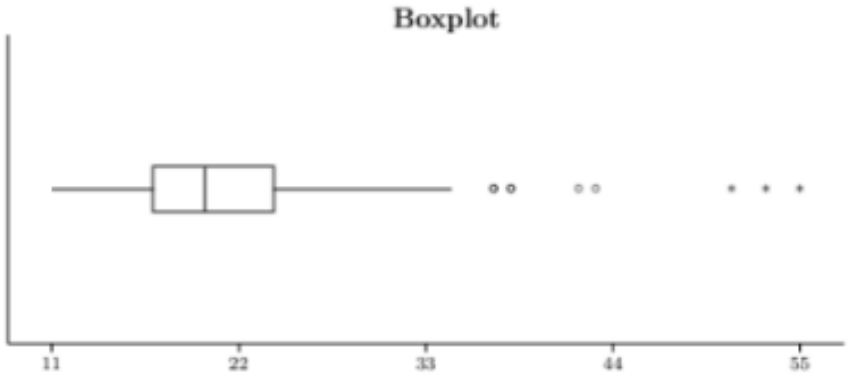

Figure 20: Boxplot de la variable CityMPG.

En cuanto a la eficiencia (highway mpg), el promedio es de 28.31 mpg, con una mediana de 28.0 mpg. Con una desviación estándar de 5.26 mpg y un rango desde 18 mpg hasta un máximo de 52 mpg. En la Figura 21
. podemos observar tres posibles valores atípicos, que a diferencia del análisis anterior, se excluyen los vehículos eléctricos.

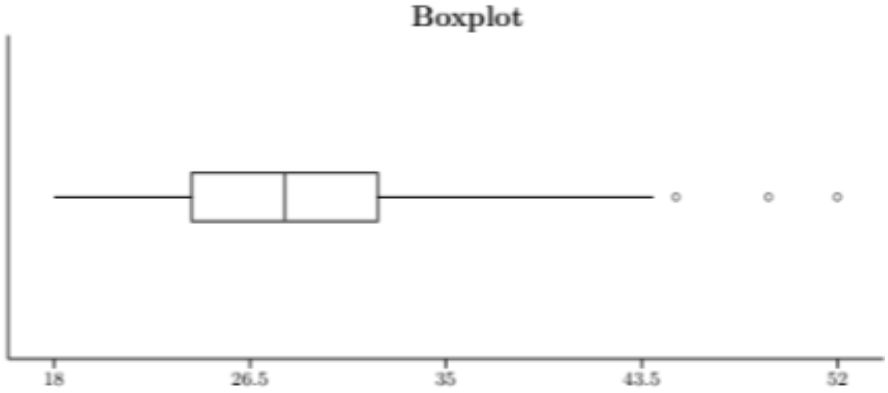

Figure 21: Boxplot de la variable HighwayMPG.

La eficiencia combinada (combination mpg), tiene una media de 23.72 mpg, una mediana de 23.0 mpg. Los valores oscilan entre un mínimo de 14 mpg y un máximo de 52 mpg, con una desviación estándar de 5.26 mpg. De estas tres variables que hemos visto hasta ahora, podemos deducir que no son los coches más eficientes del mundo, ya que el estándar está situado a unos 30-60 MPG (Miles Per Gallon). Los resultados obtenidos tienen sentido, ya que como bien sabemos son coches para las altas clases, significando que no se suelen preocupar por tener

que pagar más gasolina de la habitual, sino tener un coche lo más potente posible y que tenga aspecto moderno.

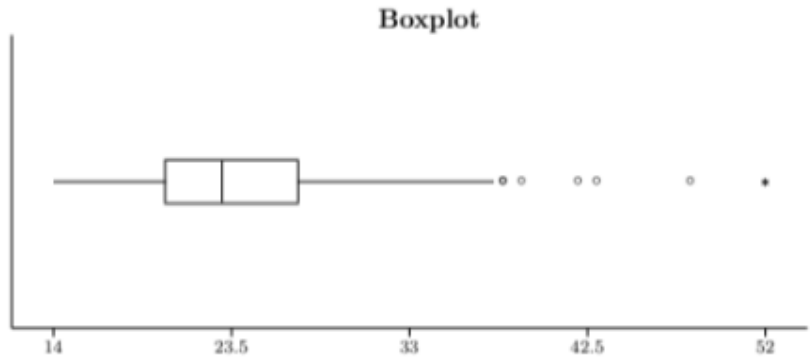

Figure 22: Boxplot de la variable CombinationMPG

Las especificaciones de los motores (**cylinders and displacement**). Tiene una media de 2.93 litros, con una mediana de 2.5 litros. El rango abarca desde un mínimo de 1.2 litros hasta un máximo de 6.8 litros, con una desviación estándar de 1.25 litros.

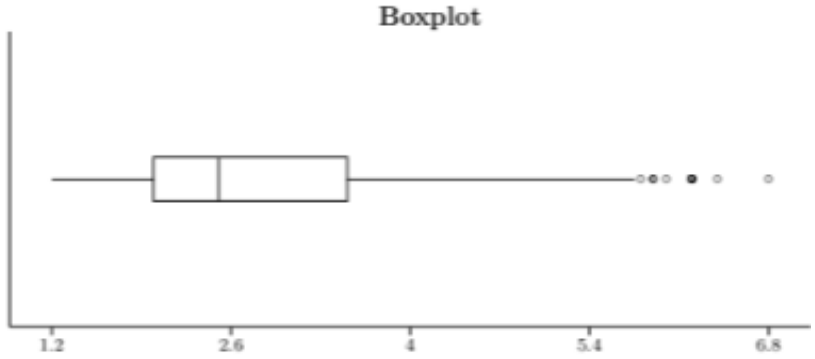

Figure 23: Boxplot de la variable Displacement

En cuanto al tipo de tracción (**drive**), se observa que la tracción total (AWD) es la más común, representando 213 vehículos. Le siguen la tracción delantera (FWD) con 178 vehículos, la tracción trasera (RWD) con 113 y, en menor medida, la tracción 4WD con 42 modelos. De esta variable no podemos sacar mucha información, ya que la eficiencia de esta variable (del tipo de tracción) depende de variables de las que no disponemos actualmente: del motor que usa, el peso del coche, el tipo de rueda, etc.

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Fwd	178	178	0.326	0.326
4wd	42	220	0.0769	0.4029
Rwd	113	333	0.207	0.6099
Awd	213	546	0.3901	1
dades mancants	0	N = 546	0	

Figure 24: Tabla de frecuencia del tipo de tracción (Drive)

Respecto al tipo de combustible (**fuel type**), el dataset está dominado por vehículos que utilizan gasolina, con 546 modelos en esta categoría. Como es el mismo para todos, no lo hemos incluido en la submatriz de datos.

Entre los fabricantes (**make and models**) más representados, destacan BMW (72 modelos), Jaguar (71 modelos) y Kia (69 modelos). Marcas como Mini y Hyundai también tienen una presencia significativa con 51 y 38 modelos, respectivamente. Por otro lado, fabricantes como Chrysler, Infiniti y Porsche cuentan con una representación mucho más limitada, con solo uno o dos modelos cada uno. Adicionalmente se brinda información sobre la clase (class), donde se tiene que 157 están representados por SUV-small, 83 de Compact y otros 83 de Subcompact.

Taula de freqüències [d]				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Midsize	52	52	0.0952	0.0952
SUV-Small	157	209	0.2875	0.3828
Subcompact	83	292	0.152	0.5348
Large	12	304	0.022	0.5568
Two-Seater	69	373	0.1264	0.6832
Minicompact	21	394	0.0385	0.7216
SUV-Standard	34	428	0.0623	0.7839
Compact	83	511	0.152	0.9359
Wagon-Small	11	522	0.0201	0.956
Pickup-Standard	7	529	0.0128	0.9689
Minivan	8	537	0.0147	0.9835
Pickup-Small	7	544	0.0128	0.9963
Wagon-Midsize	2	546	0.0037	1
dades mancants	0	N = 546	0	

Figure 25: Tabla de frecuencia de la clase del vehículo (class).

El rango de años de fabricación (year) en el dataset va de 2014 a 2024, con una media y una mediana de 2019.

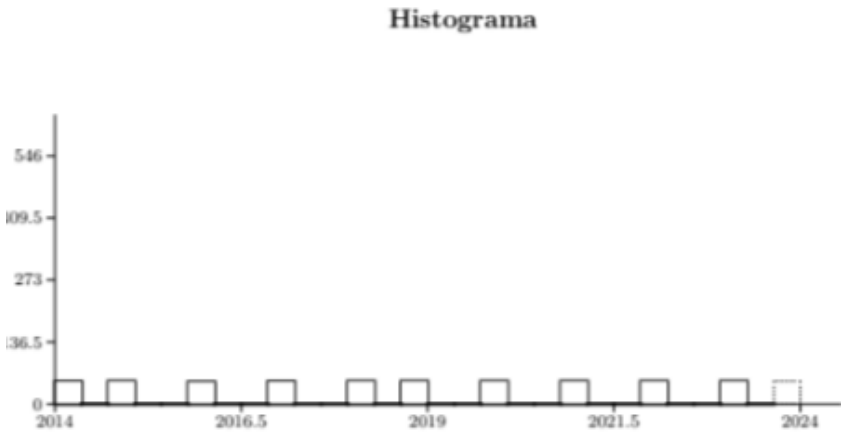

Figure 26: Histograma de la cantidad de modelos por cada año (Year).s

Comparando la descriptiva anterior con la descriptiva actual podemos ver que no hay muchos cambios en las medias, las medianas y las desviaciones estándar. Esto podemos afirmar que es debido a que solamente hemos usado una ontología, pero en un futuro si quisiéramos alargar el proyecto podríamos añadir más, y eso conlleva a tener aún más diferencias entre las dos descriptivas.

9.3 Clustering en Klass

9.3.1 Dendograma sin corte

Figure 1: CAJ. Arbre general de classificació

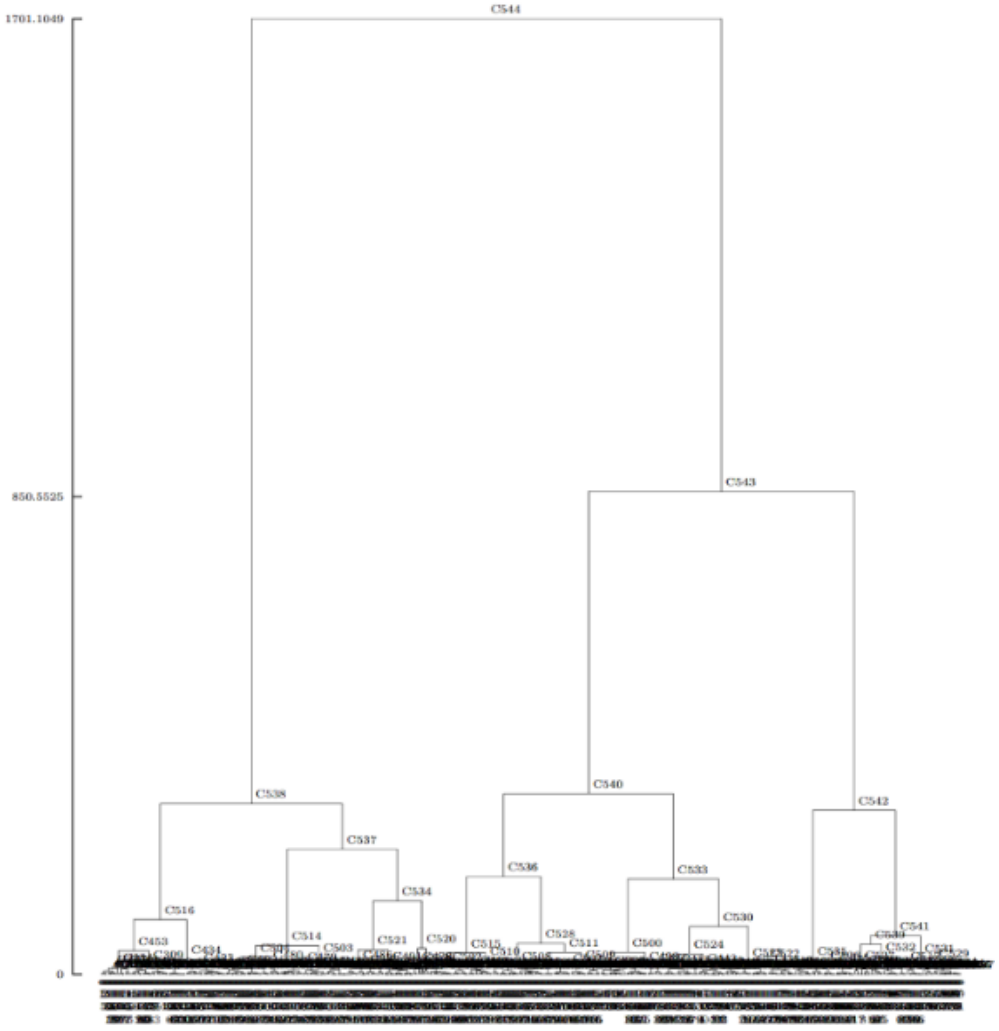

Figure 27: Dendograma resultante de aplicar la estrategia de vecinos recíprocos con ontología.

Al igual como ocurrió con el Dendograma sin considerar las ontologías, en este con ontologías se puede evidenciar que maneja un comportamiento bastante similar por lo que las diferencias son bastante mínimas. Como hemos dicho anteriormente,

esto es debido a que solo hemos usado una sola ontología y la efectividad de ésta no está garantizada. En contraste con el clustering anterior, el agoritmo usado es mixta Gibert generalizada, que permite tener en cuenta ontologías que hayamos definido.

9.3.2 Dendograma con corte

De manera similar al CPG anterior, se decidió seleccionar la clase C543 y se obtuvieron las mismas tres clases principales: C538, 541 y 542.

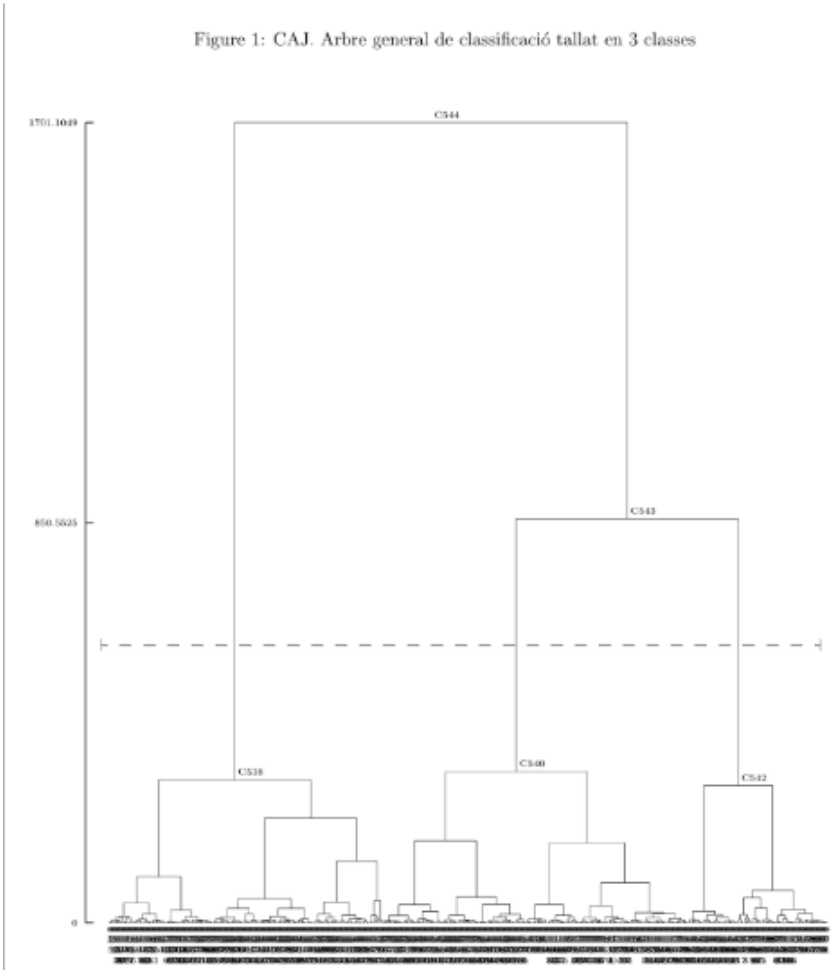

Figure 28: Dendograma con el corte marcado a C543 con ontologías.

9.4 Class Panel Graph

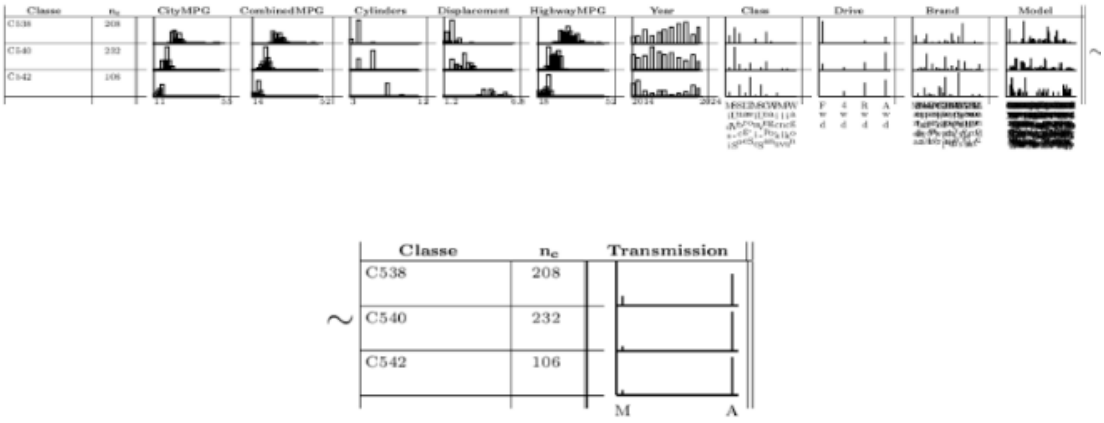

Figure 29: Class Panel Graph resultante de la ontología

El CPG es casi idéntico al anterior de las figura 13, indicando que esta ontología no ha ayudado mucho en la distribución de clases. A su vez, también significa que los datos sin ontologías estaban bien distribuidos y eran claramente diferenciables, el algoritmo es lo suficientemente bueno como para ser capaz de distinguir las clases sin la ayuda de esta ontología. Seguimos podiendo ver diferencias notorias entre las clases, de modo que podemos considerar este clustering como válido para nuestro análisis.

9.5 Traffic Light Panel

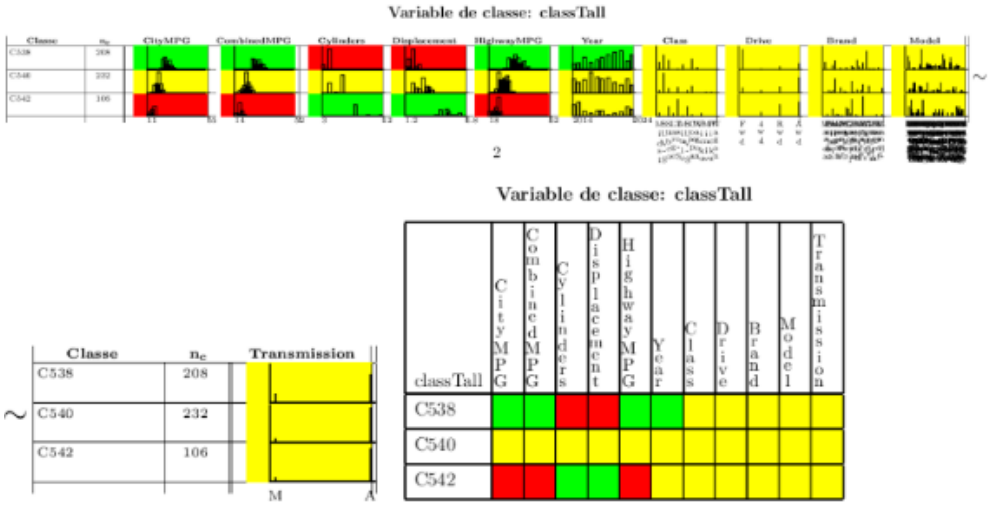

Figure 30: TLP resultante después de aplicar ontologías pintado “a ojo”

En estas figuras podemos apreciar que las ontologías no han ayudado mucho a aportar información porque los dos mosaicos del TLP ”normal” son exactamente iguales, pero nos sirven para verificar que los datos obtenidos anteriormente no son del todo erróneos y que las clases que se han creado están correctamente estructuradas.

En el *annotated* pasa exactamente lo mismo, no hay ninguna diferencia en color respecto al *annotated* sin ontologías.

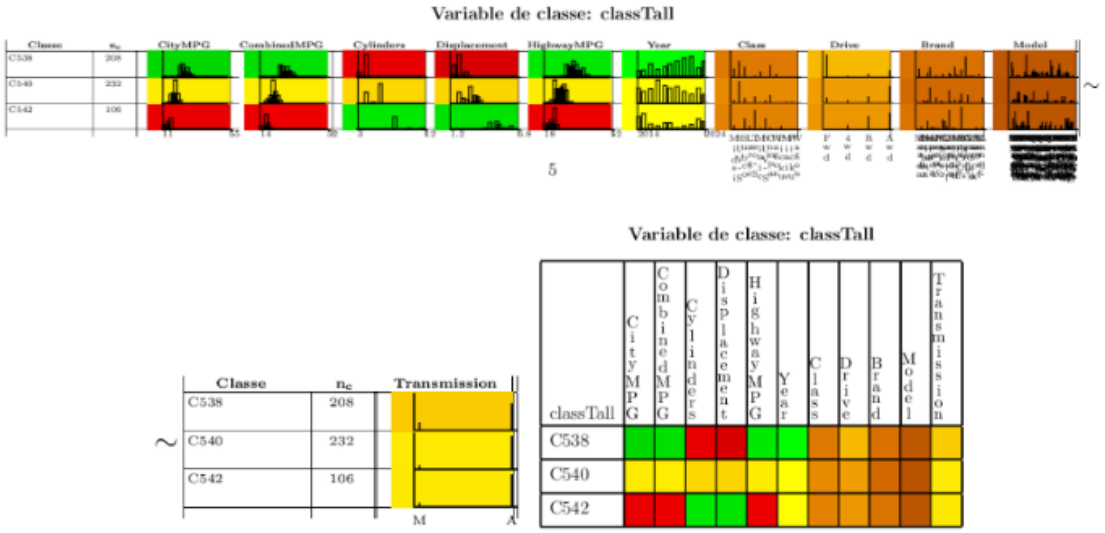

Figure 31: TLP resultante después de aplicar ontologías pintado “a ojo” en modo *annotated*

9.6 Termómetro y nuevo Traffic Light Panel

Como se ha mencionado anteriormente, el termómetro usado es el mismo, en Klass se puede exportar creando el fichero termometre.ter correspondiente a la figura 16. La figura 32 refleja el nuevo TLP usando este termómetro con ontologías.

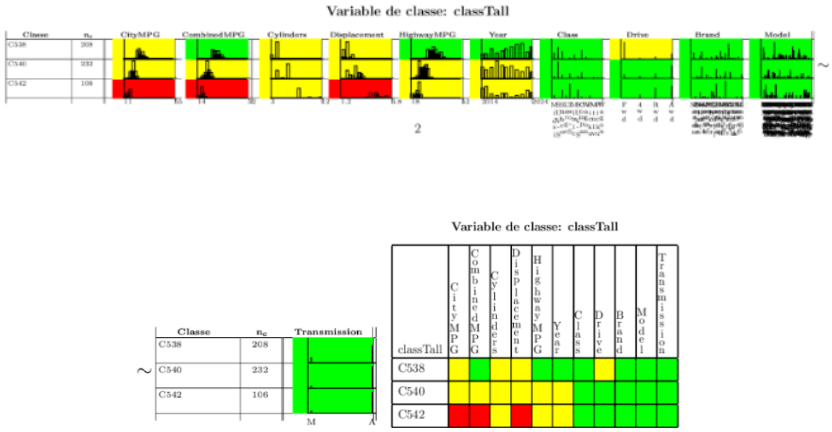

Figure 32: TLP "normal" resultante de aplicar el termómetro al CPG con ontologías

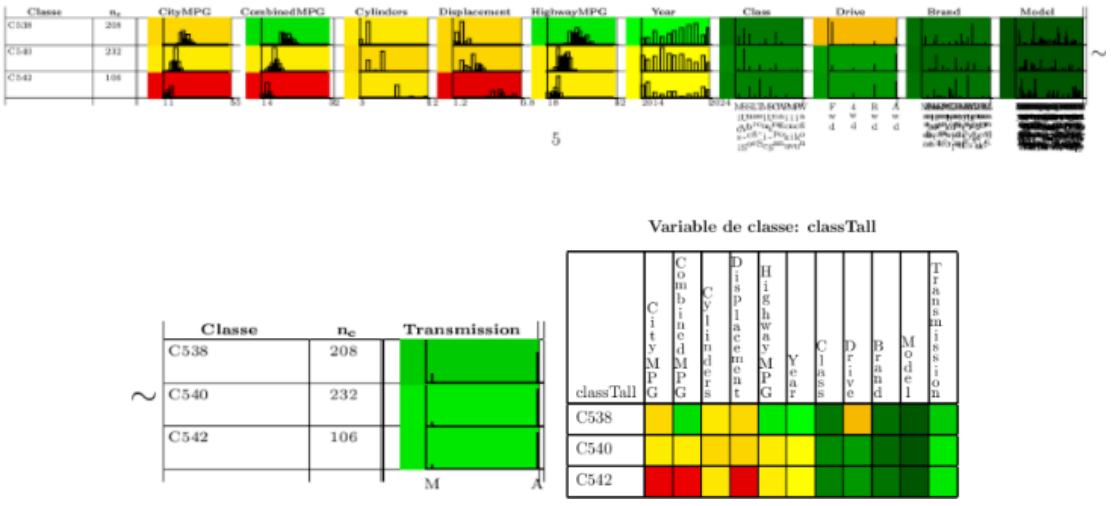

Figure 33: TLP *annotated* resultante de aplicar el termómetro al CPG con ontologías.

Después de aplicar ontologías con termómetro los colores siguen siendo exactamente los mismos tanto para el "normal" como para el *annotated*, lo que nos deja con dos hipótesis:

- Las ontologías no han sido demasiado útiles y no han afectado al *clustering*.
- Los datos ya estaban bien estructurados y no necesitaban de esta ontología para encontrar las **3 clases principales** que hemos obtenido.

10 Conclusión

10.1 Resumen del proyecto

Este proyecto ha empezado con la **selección del dataset**, la cual ha sido difícil debido a los estrictos requisitos que debía cumplir el dataset: mínimo 5 columnas categóricas y 5 numéricas. El dataset que finalmente hemos escogido contiene información sobre la eficiencia del uso de combustible de distintos modelos de coches según distintos parámetros, como por ejemplo las millas por galón y los centímetros cúbicos del motor.

El **preprocesado** de datos lo hemos hecho con un script en R. El script utiliza el método de pasarela para transformar los datos CSV a .obj, .dat y .pro aplicando los cambios descritos en la sección 3.

Posteriormente hemos hecho un clustering sin reglas ni conocimiento previo (**sin ontologías**) usando el algoritmo mixto de Gibert y obtenido un dendograma y un CPG correspondientes. Para tener una visualización más clara de los datos y asegurar su fiabilidad hemos hecho dos TLPs sin termómetro, el "normal" y el annotated. Posteriormente hemos hecho los mismos TLPs pero pintándolos con un termómetro. Para cada gráfico resultante hemos hecho un análisis de los gráficos e intentado explicar el comportamiento de los datos.

Finalmente hemos hecho los mismos pasos anteriores pero aplicando reglas (ontologías) a los datos. En este apartado nos hemos centrado sobretodo en resaltar las diferencias respecto a hacerlo sin ontologías.

10.2 Análisis de resultados

Como resultado del clustering de datos sin reglas hemos obtenido tres clases mayormente distinguidas: C536, C541 y C542. Cada una se diferencia principalmente por el HighwayMPG (Miles Per Gallon en carretera), CityMPG (lo mismo en ciudad), CombinedMPG (lo mismo combinando las dos variables anteriores), Cylinders (número de cilindros) y Displacement (centímetros cúbicos), ya que se pueden apreciar notables diferencias entre ellos, en una clase el gráfico tiene más puntos a la izquierda mientras que otra clase los tiene a la derecha. Para el resto de variables no se puede apreciar ninguna diferencia notable a simple vista, de modo que podríamos considerar que estas variables no han afectado mucho en la decisión del algoritmo para realizar el clustering.

El resultado del clustering con reglas (con ontologías) ha sido distinto, hemos

obtenido tres clases también pero son la C538, C541 y la C542. Igual que anteriormente el corte del dendograma es claro y las diferencias entre las clases están remarcadas. Los TLPs con y sin termómetro con ontologías también dan muy parecido a los sin ontologías, lo que nos lleva a dos posibilidades:

- La ontología aplicada no ha sido demasiado útil y no ha afectado al *clustering*. Se deberían aplicar otras para tener un cambio muy notorio.
- Los datos ya estaban bien estructurados y distribuidos, y no necesitaban de esta ontología para llegar a las clases que hemos obtenido.

10.3 Opinión personal y futuras mejoras

Este proyecto nos ha enseñado la importancia de preprocesar bien todos los datos antes de usarse, ya que de otro modo es muy difícil extraer conclusiones claras sobre los gráficos y que éstos muestren algún patrón significativo. Esto se puede hacer mediante el correcto tratamiento de outliers y valores faltantes y su documentación. Además, hemos notado la intrínseca necesidad de tener un experto a nuestro lado para poder interpretar bien los resultados obtenidos y poder darles un sentido. Para solucionar esto nos hemos informado con información externa (artículos, opiniones y comentarios, experiencia personal, reseñas, vídeos, etc).

La mejora principal que se le puede aplicar a este proyecto es usar otras ontologías para ayudar a definir aún más el clustering resultante y hacer pruebas sin tener en cuenta variables que nos han aportado poca información para la clasificación, como serían el Class, el Drive,el Brand, el Transmission y el Model.