2024-2025 MEI-IKPD

Minería de datos

Sobre un conjunto de datos de coches

Albert, Carlos, Javier, Juan, Noa, René

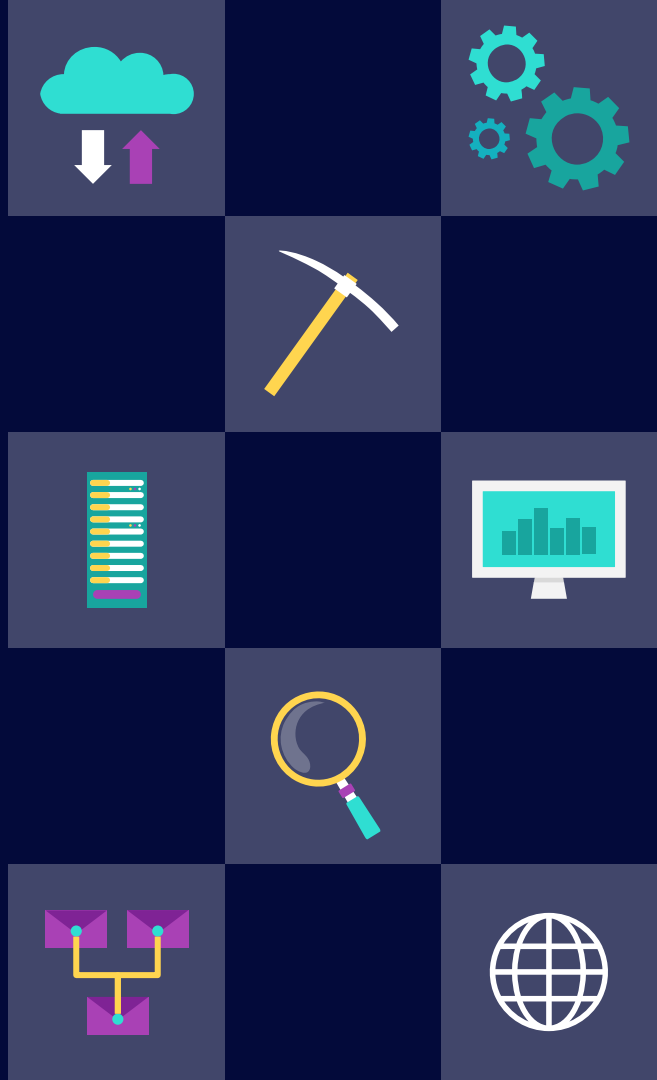


Tabla de contenidos I

Tabla de contenidos II





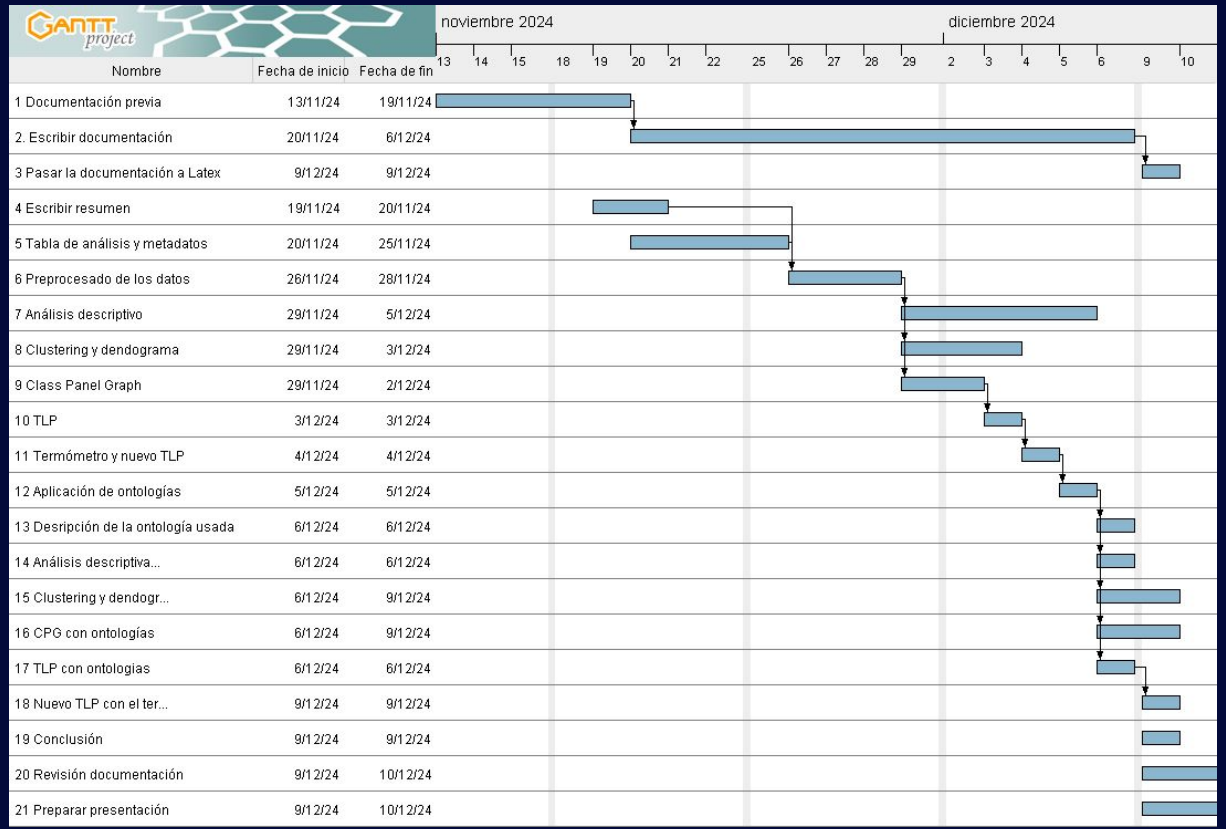



Documentación previa

01

Fuente, descripción y motivación. Planificación.

Diagrama de Gantt



Nombre	Fecha de inicio	Fecha de fin
1 Documentación previa	13/11/24	19/11/24
2. Escribir documentación	20/11/24	6/12/24
3 Pasar la documentación a Latex	9/12/24	9/12/24
4 Escribir resumen	19/11/24	20/11/24
5 Tabla de análisis y metadatos	20/11/24	25/11/24
6 Preprocesado de los datos	26/11/24	28/11/24
7 Análisis descriptivo	29/11/24	5/12/24
8 Clustering y dendograma	29/11/24	3/12/24
9 Class Panel Graph	29/11/24	2/12/24
10 TLP	3/12/24	3/12/24
11 Termómetro y nuevo TLP	4/12/24	4/12/24
12 Aplicación de ontologías	5/12/24	5/12/24
13 Desripción de la ontología usada	6/12/24	6/12/24
14 Análisis descriptiva...	6/12/24	6/12/24
15 Clustering y dendogr...	6/12/24	9/12/24
16 CPG con ontologías	6/12/24	9/12/24
17 TLP con ontologias	6/12/24	6/12/24
18 Nuevo TLP con el ter...	9/12/24	9/12/24
19 Conclusión	9/12/24	9/12/24
20 Revisión documentación	9/12/24	10/12/24
21 Preparar presentación	9/12/24	10/12/24

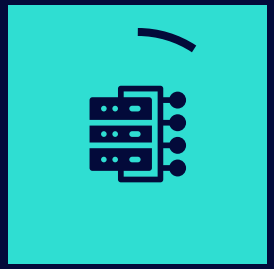
División del trabajo I

Tarea	Tarea principal	Documentación
Documentación previa	Javier Abella	
Escribir documentación	Todos	
Pasar toda la documentación a latex	Javier Abella, Noa Yu Ventura	
Resumen sobre los pasos a seguir	Noa Yu Ventura	
Tabla y análisis de metadatos	René Alonso Cortés	
Preprocesado de los datos	Albert Bausili, Carlos Andrés	Carlos Andrés, Noa Yu Ventura
Análisis descriptivo univariante	Albert Bausili	Juan José
Clustering y dendograma	Albert Bausili	Noa Yu Ventura
Class Panel Graph	Albert Bausili	Noa Yu Ventura
Traffic Light Panel	Albert Bausili	Noa Yu Ventura
Termómetro y nuevos TLP	Albert Bausili, Noa Yu Ventura	Noa Yu Ventura

División del trabajo II

Tarea	Tarea principal	Documentación
Aplicación de ontologías	Carlos Andrés	Juan José, Noa Yu Ventura, Albert Bausili
Descripción de la ontología usada	Noa Yu Ventura	
Análisis descriptivo con ontologías	Albert Bausili	Noa Yu Ventura, Juan José
Clustering y dendograma con ontologías	Albert Bausili	Noa Yu Ventura
CPG con ontologías	Albert Bausili	Noa Yu Ventura
TLP con ontologías	Albert Bausili	Noa Yu Ventura
Termómetro y nuevos TLP con ontologías	Albert Bausili	Noa Yu Ventura
Conclusión	Albert Bausili, Noa Yu Ventura	
Revisión documentación	Javier Abella, Noa Yu Ventura, Albert Bausili	
Preparar presentación	Todos	

Dataset: Car Performance Fuel Efficiency Data



550

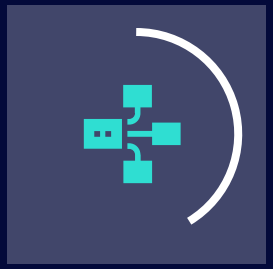
Registros

12

Columnas

6

Numéricas



6

Categóricas

Usos Potenciales: Predicción de consumo, comparación de modelos y análisis de impacto ambiental.

Motivación

¿Qué factores técnicos y de diseño influyen más en la eficiencia de combustible en ciudad y carretera?

Tabla de metadatos

02

Metadatos clave para conocer el conjunto de datos.

Tabla de metadatos

Atributo	Modalidades	Descripción	Tipo (categoría)	Tipo (dato)	Unidad	Missing	Rango	Distintos
city_mpg	N/A	Millas por galón en ciudad	Numérico	int64	mpg	0	11 - 126	31
class	Ej. ['midsize car', 'small sport utility vehicle', ...]	Clase del vehículo	Categórico	string	N/A	0	N/A	13
combination_mpg	N/A	Promedio de millas por galón	Numérico	int64	mpg	0	14 - 112	30
cylinders	N/A	Número de cilindros	Numérico	float64	Unidad	2	3.0 - 12.0	7
displacement	N/A	Cilindrada del motor	Numérico	float64	Litros	2	1.2 - 6.8	29
drive	['fwd', '4wd', 'rwd', 'awd']	Tipo de tracción	Categórico	string	N/A	0	N/A	4
fuel_type	['gas', 'diesel', 'electricity']	Tipo de combustible	Categórico	string	N/A	0	N/A	3
highway_mpg	N/A	Millas por galón en carretera	Numérico	int64	mpg	0	18 - 102	32
make	Ej. ['mazda', 'ford', 'subaru', 'nissan', 'audi', ...]	Marca del vehículo	Categórico	string	N/A	0	N/A	31
model	Ej. ['6', 'cx-5 2wd', 'cx-5 4wd', 'mustang', 'forester awd', ...]	Modelo del vehículo	Categórico	string	N/A	0	N/A	276
transmission	['m' (manual), 'a' (automático)]	Tipo de transmisión	Categórico	string	N/A	0	N/A	2
year	N/A	Año de fabricación	Numérico	string	Años	0	2014 - 2024	11

Cinco categóricas

Cinco numéricas

Tabla de metadatos

Atributo	Modalidades	Descripción	Tipo (categoría)	Tipo (dato)	Unidad	Missing	Rango	Distintos
city_mpg	N/A	Millas por galón en ciudad	Numérico	int64	mpg	0	11 - 126	31
class	Ej. ['midsize car', 'small sport utility vehicle', ...]	Clase del vehículo	Categórico	string	N/A	0	N/A	13
combination_mpg	N/A	Promedio de millas por galón	Numérico	int64	mpg	0	14 - 112	30
cylinders	N/A	Número de cilindros	Numérico	float64	Unidad	2	3.0 - 12.0	7
displacement	N/A	Cilindrada del motor	Numérico	float64	Litros	2	1.2 - 6.8	29
drive	['fwd', '4wd', 'rwd', 'awd']	Tipo de tracción	Categórico	string	N/A	0	N/A	4
fuel_type	['gas', 'diesel', 'electricity']	Tipo de combustible	Categórico	string	N/A	0	N/A	3
highway_mpg	N/A	Millas por galón en carretera	Numérico	int64	mpg	0	18 - 102	32
make	Ej. ['mazda', 'ford', 'subaru', 'nissan', 'audi', ...]	Marca del vehículo	Categórico	string	N/A	0	N/A	31
model	Ej. ['6', 'cx-5 2wd', 'cx-5 4wd', 'mustang', 'forester awd', ...]	Modelo del vehículo	Categórico	string	N/A	0	N/A	276
transmission	['m' (manual), 'a' (automático)]	Tipo de transmisión	Categórico	string	N/A	0	N/A	2
year	N/A	Año de fabricación	Numérico	string	Años	0	2014 - 2024	11

Cinco categóricas

Cinco numéricas








Preprocesado de datos

03

Renombrado, simplificación, capitalización, limpieza y manejo de valores atípicos.

Preprocesado de datos

- Renombrado de columnas (make → Brand)
- Simplificación de valores (range rover evoque → Evoque)
- Capitalización y limpieza de caracteres problemáticos (‘, /, *)
- **Tratamiento de outliers**

Tratamiento de outliers

Valores faltantes

Displacement

Cylinders

Valores extremos

Highway Miles Per Gallon

City Miles Per Gallon

Combined Miles Per Gallon

Valores de otra población

Fuel Type:

- Electric
- Diesel





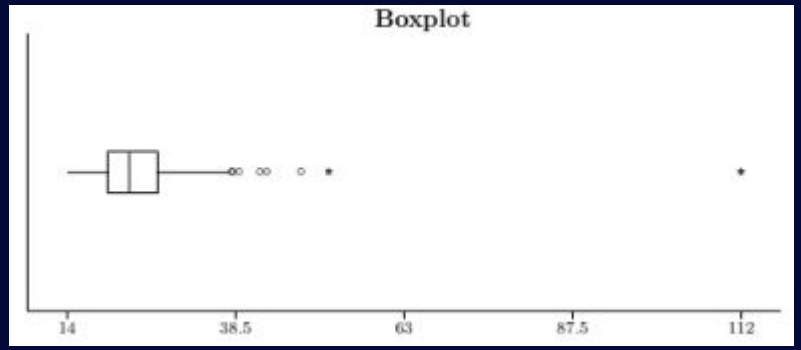



Análisis descriptivo

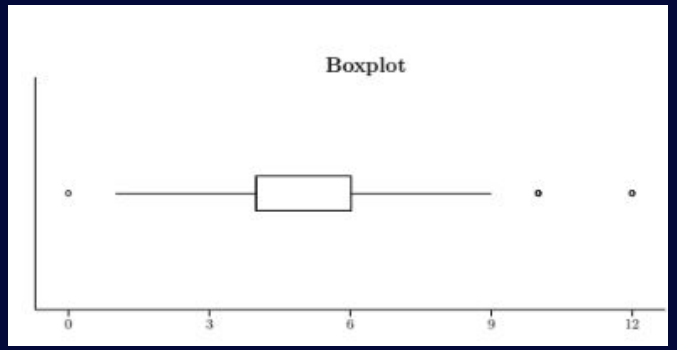
04

Boxplot, histogramas, tablas de frecuencia.

Análisis descriptivo univariante I



Boxplot CombinedMPG



Boxplot Cylinders y Displacement

Análisis descriptivo univariante II

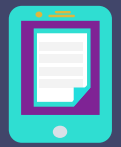
Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Midsize	52	52	0.0952	0.0952
SUV-Small	157	209	0.2875	0.3828
Subcompact	83	292	0.152	0.5348
Large	12	304	0.022	0.5568
Two-Seater	69	373	0.1264	0.6832
Minicompact	21	394	0.0385	0.7216
SUV-Standard	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	

Tabla frecuencias Drive







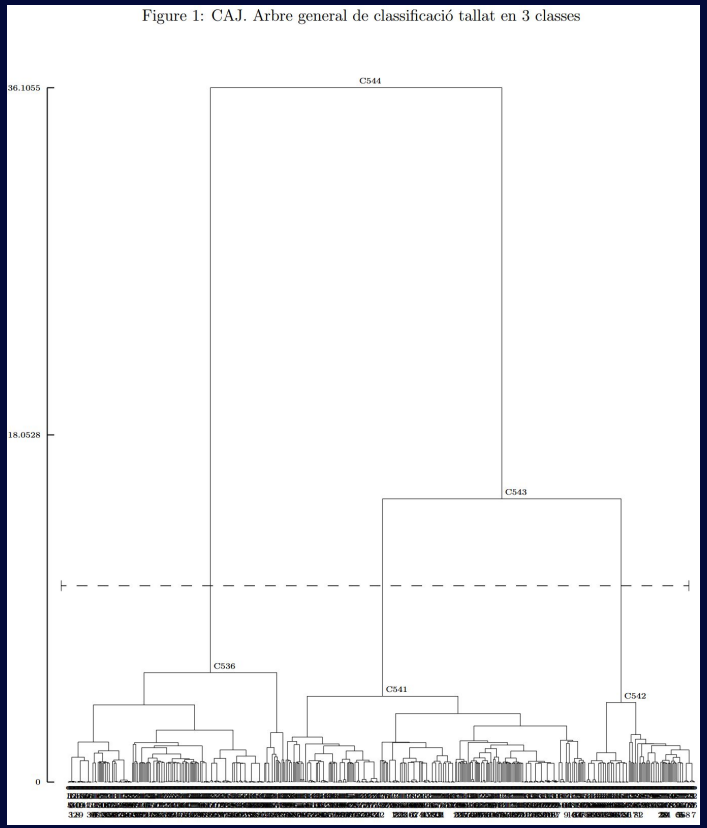
MEI-IKPD 2024-2025

Clustering

05

Obtención dendograma, corte dendograma.

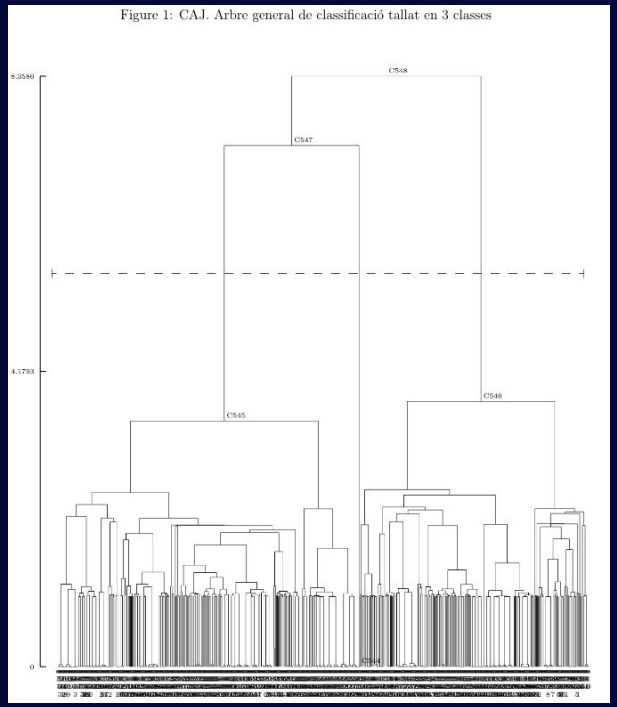
Clustering

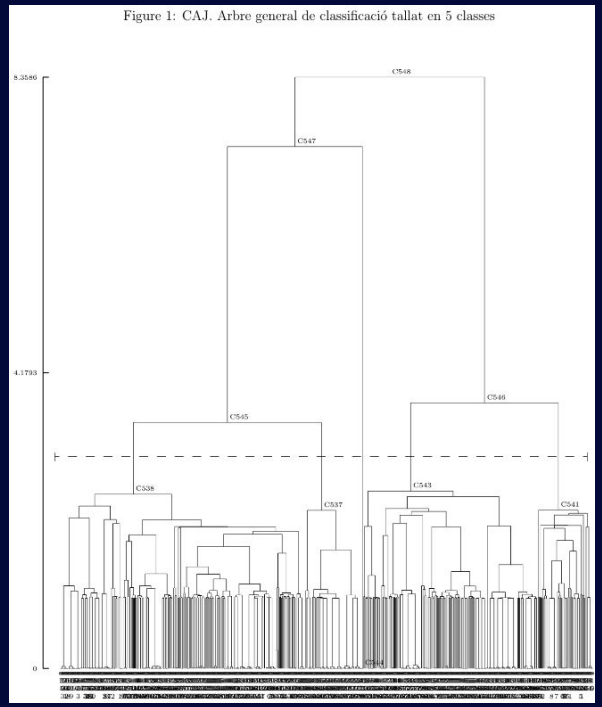


Algoritmo mixto de Gibert

3 clases diferenciadas

Resultado de incluir coches eléctricos y de diesel en nuestro dataset

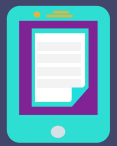










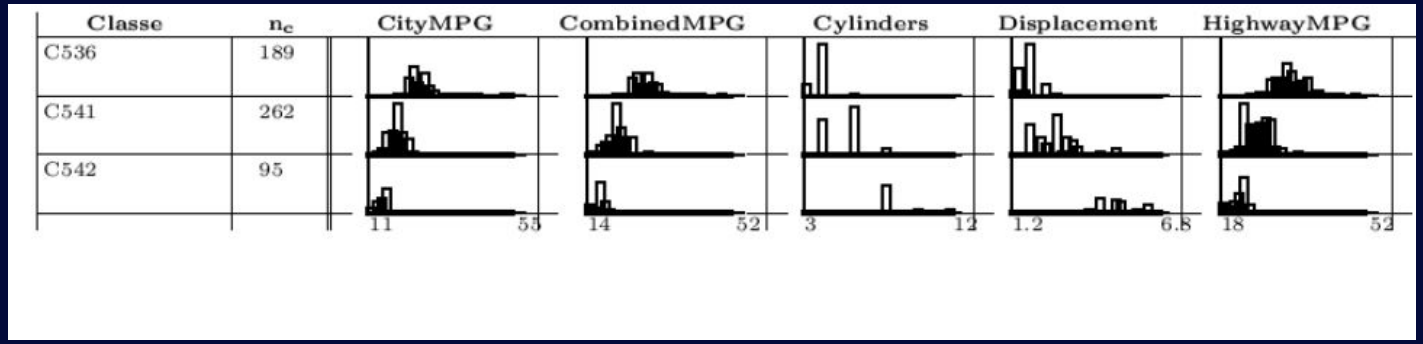



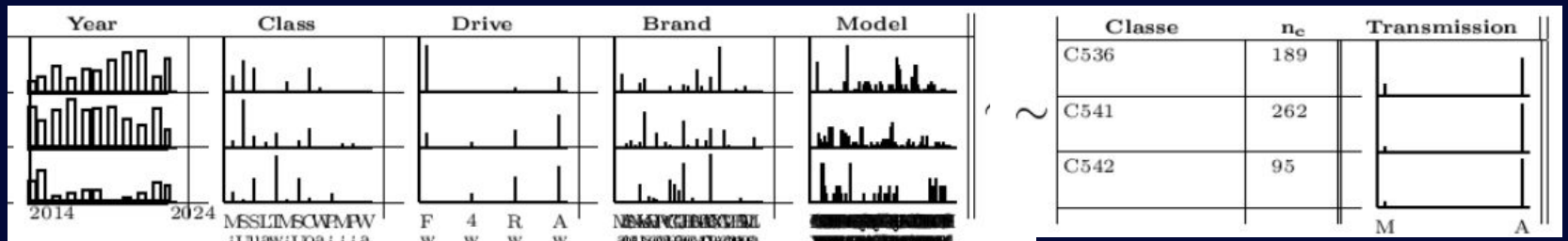
MEI-IKPD 2024-2025

CPG 06

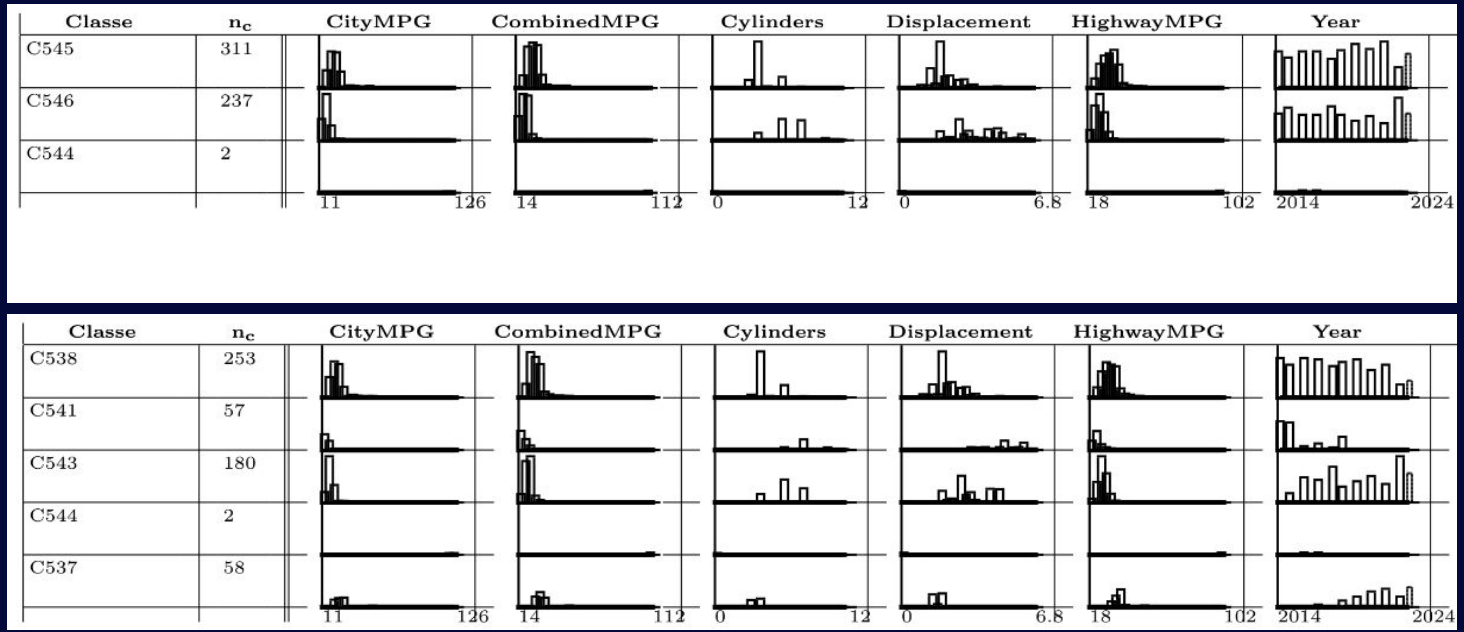
Obtención y análisis del CPG.

Class Panel Graph (CPG)





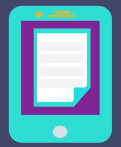
Resultado de incluir coches eléctricos y de diesel en nuestro dataset







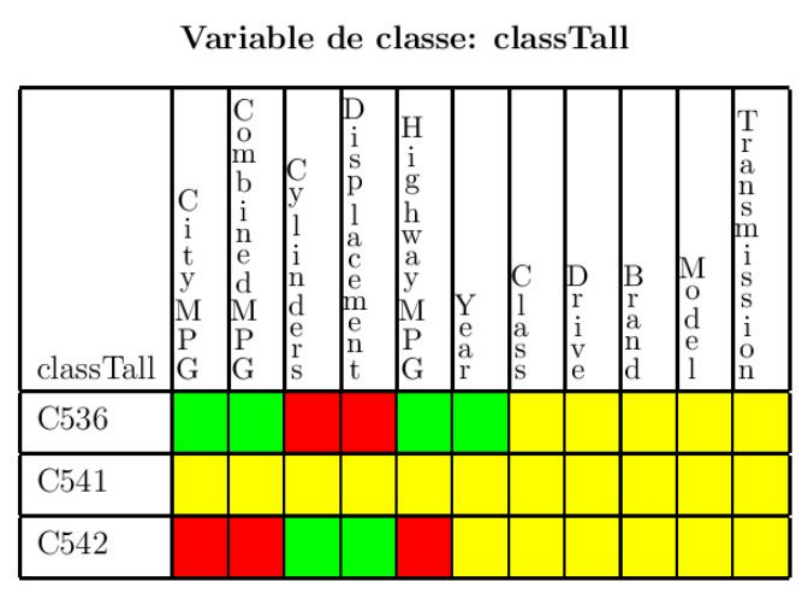


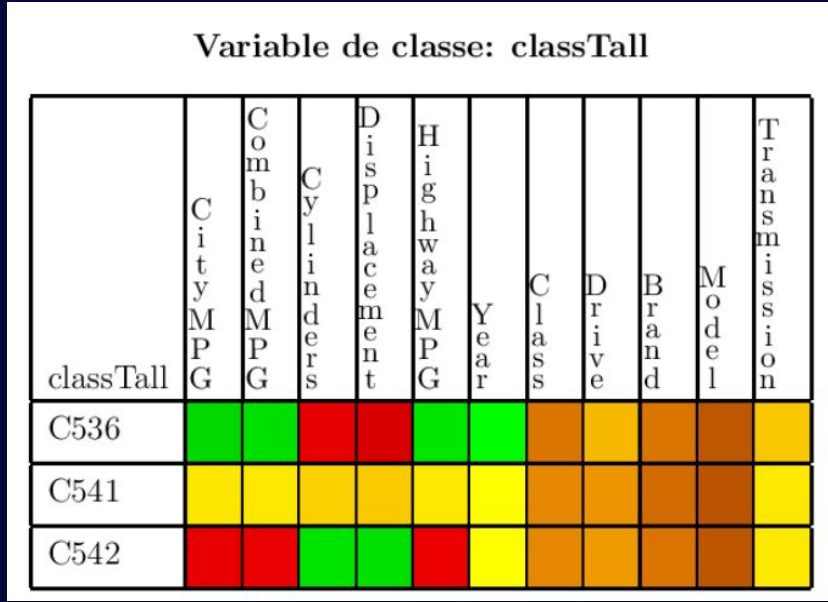



TLP 07

Obtención y análisis de TLP y TLP anotado.

Traffic Light panel

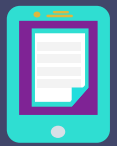










MEI-IKPD 2024-2025

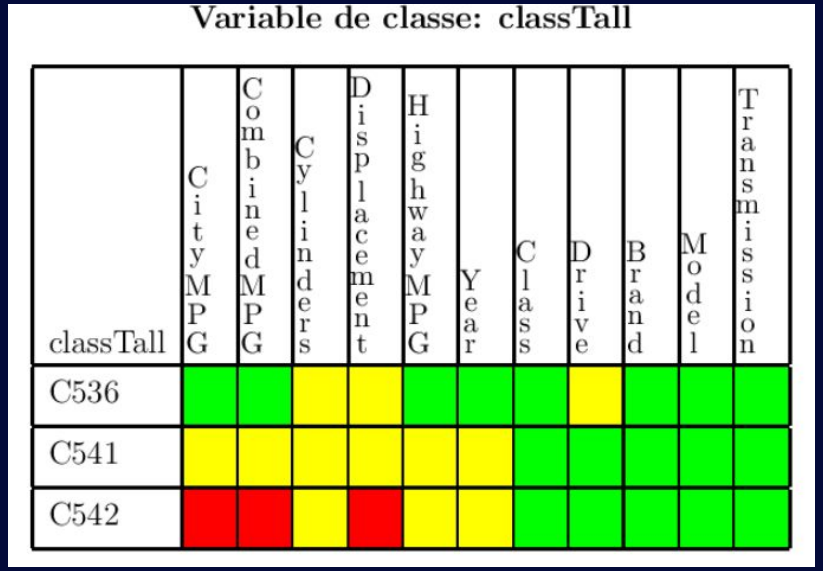
Termómetro

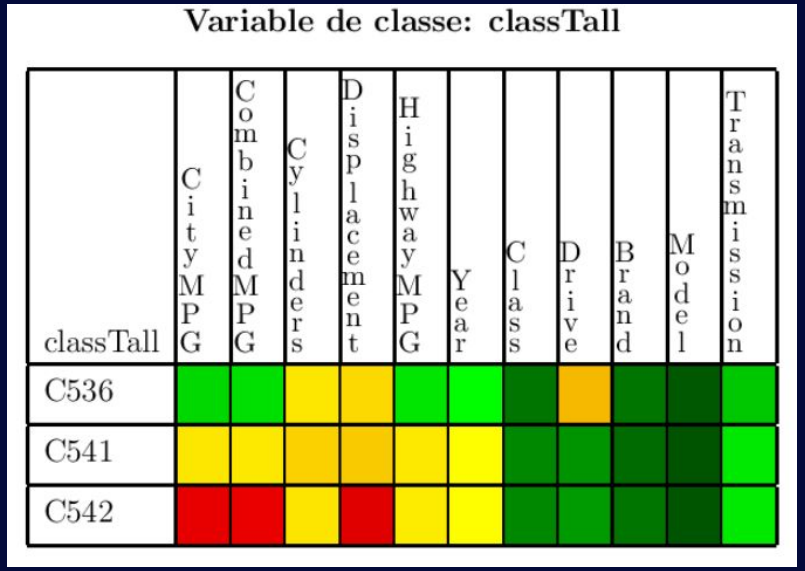
08

Creación y análisis del termómetro y nuevos TLP.

Termómetro usado

Traffic Light panel con termómetro











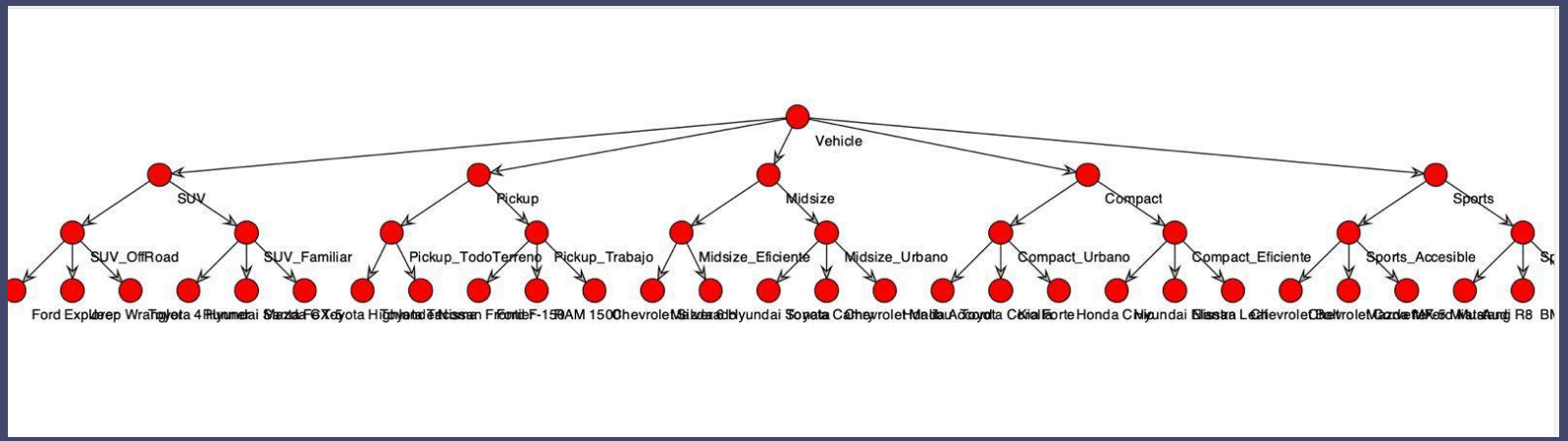

MEI-IKPD 2024-2025

Ontologías

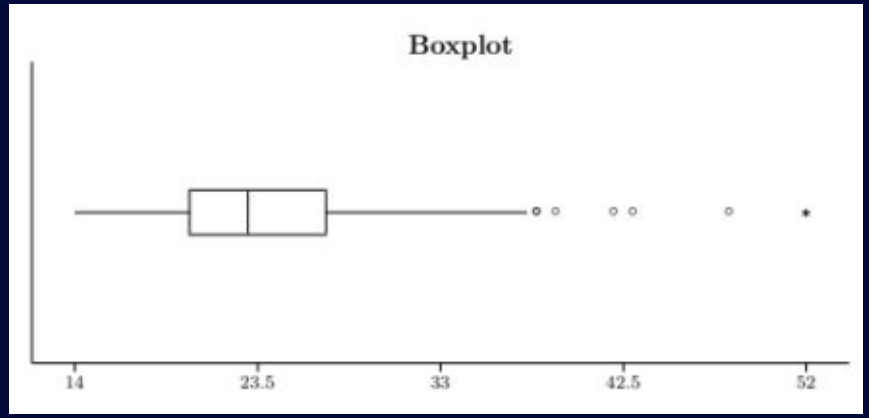
09

Ontología usada, análisis, clustering, CPG, TLP, termómetro. Análisis resultados.

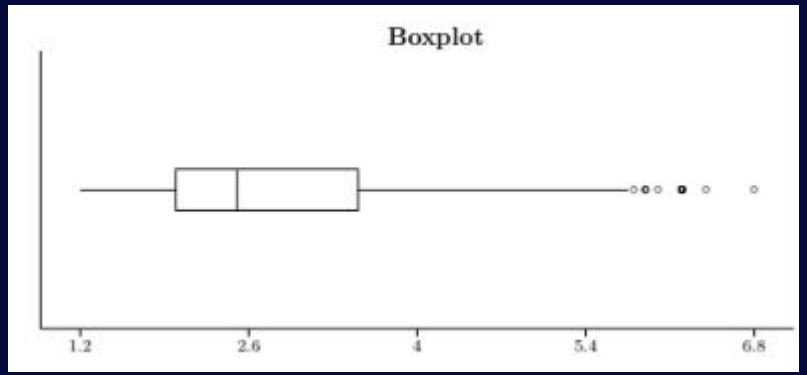
Ontología usada



Análisis descriptivo univariante

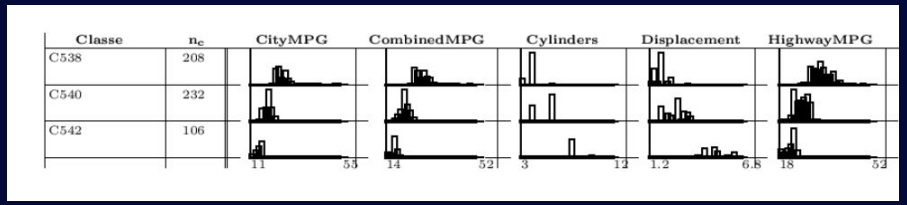


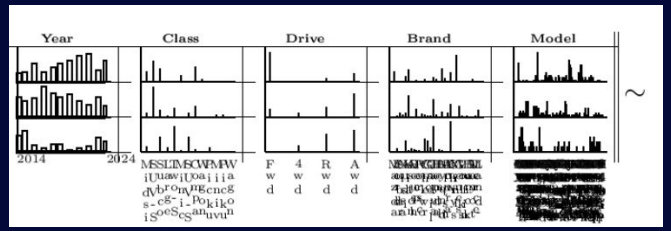
Boxplot CombinedMPG

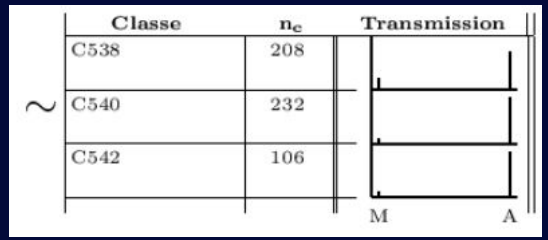


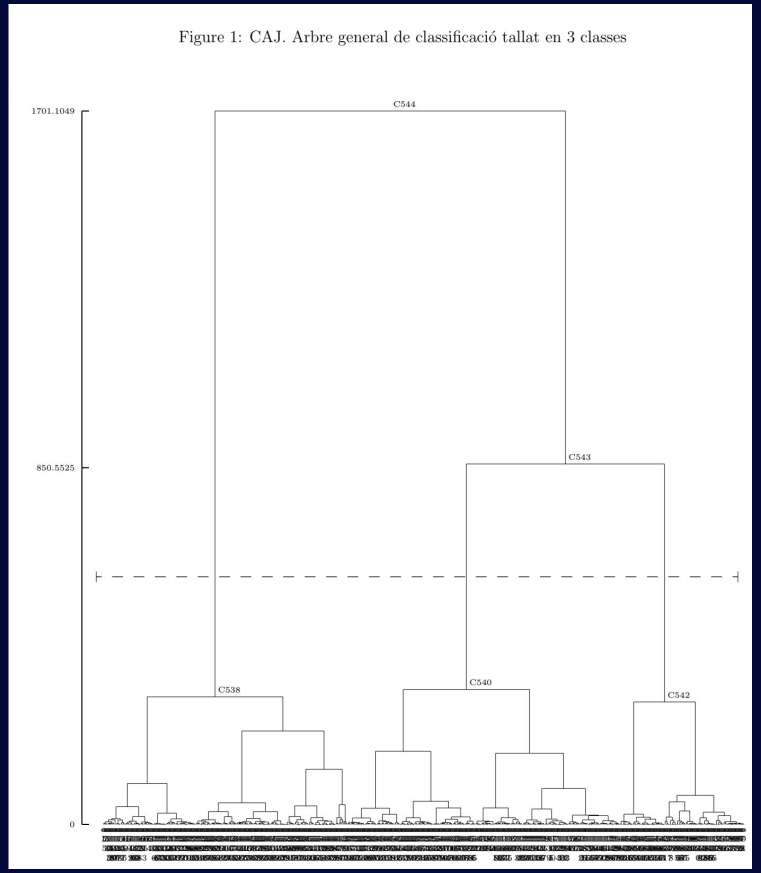
Boxplot Cylinders y Displacement

Clustering y CPG





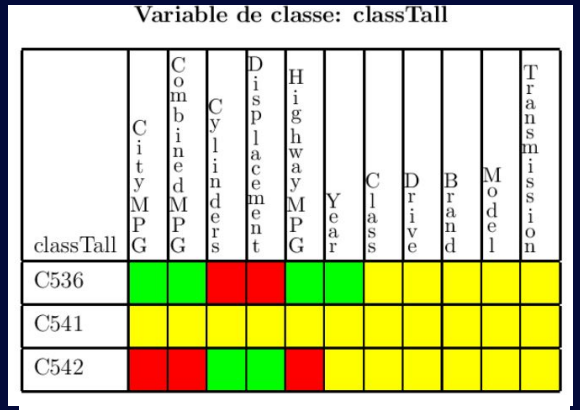


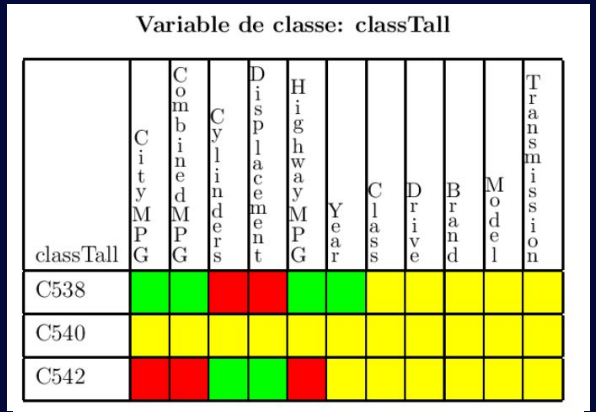


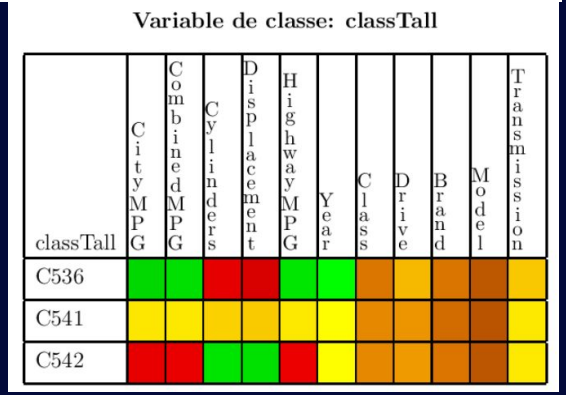
TLPs sin termómetro, sin y con ontologías

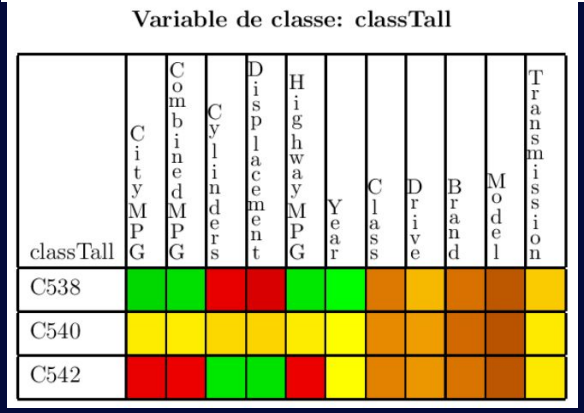
Normal

Annotated









TLPs con termómetro, sin y con ontologías

Normal

Annotated

Variable de classe: classTall

classTall	CityMPG	CombinedMPG	Cylinders	Displacement	HighwayMPG	Year	Class	Drive	Brand	Model	Transmission
C536											
C541											
C542											

Variable de classe: classTall

classTall	CityMPG	CombinedMPG	Cylinders	Displacement	HighwayMPG	Year	Class	Drive	Brand	Model	Transmission
C538											
C540											
C542											

Variable de classe: classTall

classTall	CityMPG	CombinedMPG	Cylinders	Displacement	HighwayMPG	Year	Class	Drive	Brand	Model	Transmission
C536											
C541											
C542											

Variable de classe: classTall

classTall	CityMPG	CombinedMPG	Cylinders	Displacement	HighwayMPG	Year	Class	Drive	Brand	Model	Transmission
C538											
C540											
C542											



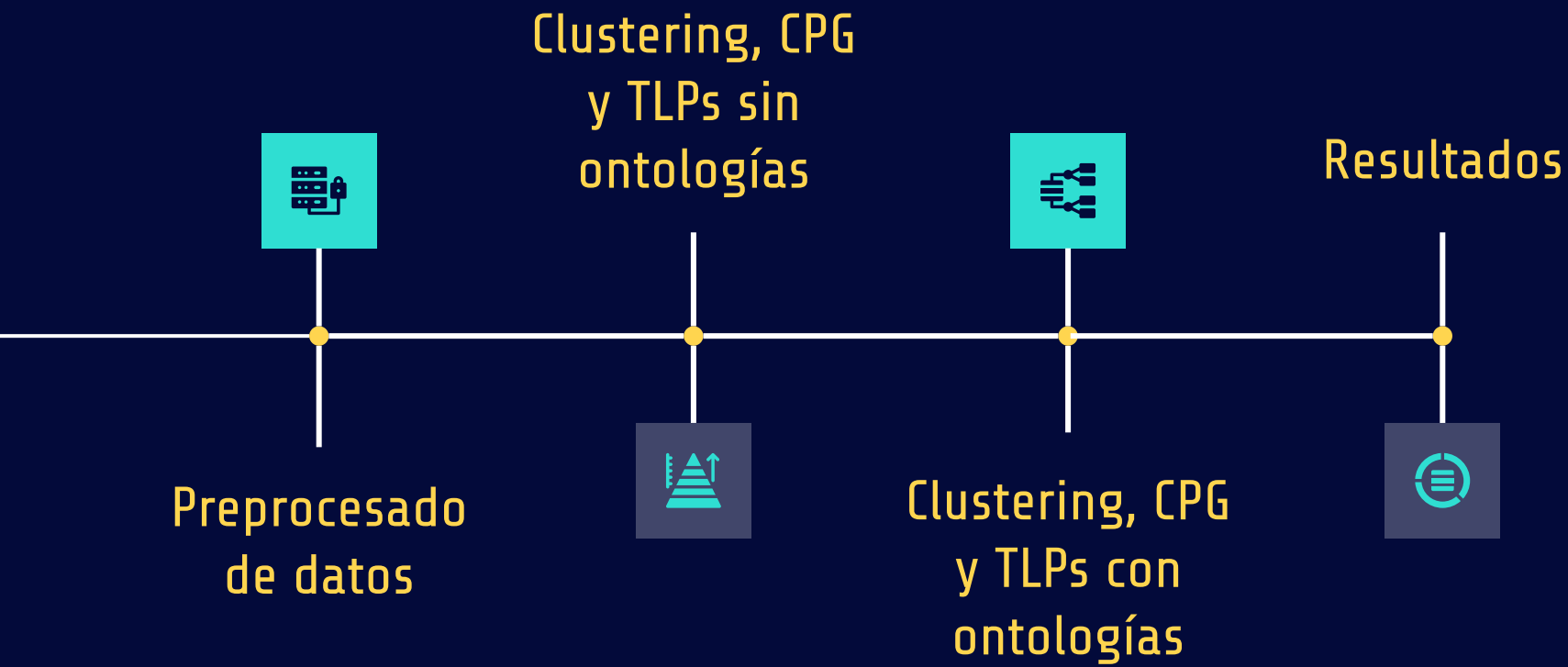




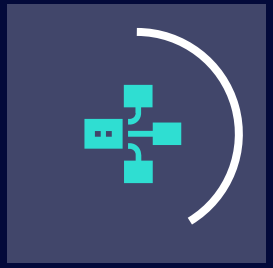

Conclusión 10

Resumen trabajo realizado y conclusiones resultantes.

Resumen


Preprocesado de datos
Clustering, CPG y TLPs sin ontologías
Clustering, CPG y TLPs con ontologías
Resultados

Resultados



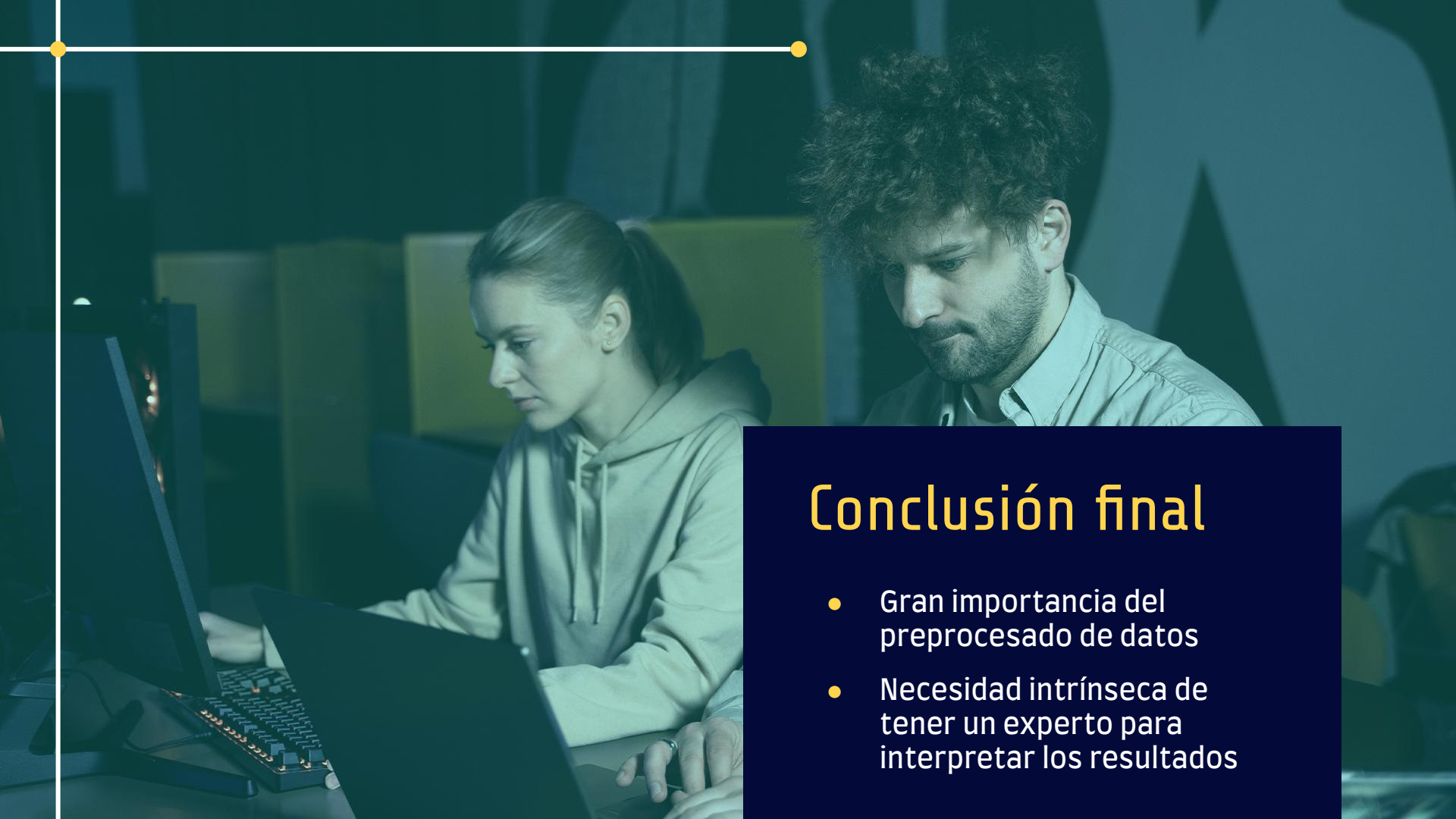
3 clases diferenciadas

HighwayMPG

CityMPG

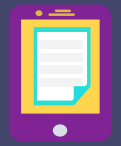
CombinedMPG

- Ontología aplicada no ha sido muy útil
- Datos ya estaban bien distribuidos



Conclusión final

- Gran importancia del preprocesado de datos
- Necesidad intrínseca de tener un experto para interpretar los resultados





¡Gracias!

¿Tenéis alguna pregunta?