Pràctica 1 – Web Scraping

Assignatura: M2.951 – Tipologia i cicle de vida de les dades

Curs 2025-26

Integrants del grup:

Ari Pidevall i Burjalés

Joan Mata Pàrraga

Enllaç web:

https://ev-database.org

Enllaç repositori github:

https://github.com/joan-mata/Practica_1_ Web_Scraping

Enllaç al dataset a Zenodo:

https://zenodo.org/records/17575547

Enllaç al vídeo:

https://drive.google.com/file/d/1LyJqAXbiK4_UrwVPZx_bYDh6VNmMqRlg/view?usp=share_link

1.- Context

Aquesta pràctica té com a objectiu posar en pràctica els coneixements sobre el procés de web scraping, que forma part del cicle de vida de les dades dins de la fase d'obtenció o recol·lecció d'informació. El nostre propòsit ha estat construir un dataset original a partir de la informació disponible en un lloc web real, aplicant criteris de planificació, selecció de fonts, automatització i emmagatzematge responsable.

Hem escollit el domini dels vehicles elèctrics per la seva rellevància actual i per la quantitat de dades tècniques i comercials que es poden trobar en línia. La transició cap a la mobilitat elèctrica és un dels grans reptes industrials i mediambientals del nostre temps, i disposar de dades fiables pot ajudar a analitzar les tendències del mercat, realitzar comparatives de rendiment o estudiar l'evolució tecnològica dels diferents fabricants.

El lloc web EV Database (https://ev-database.org) ha estat seleccionat com a font principal per diverses raons. En primer lloc, ofereix dades públiques i accessibles directament des del navegador, sense necessitat de registre ni autenticació, fet que facilita l'obtenció d'informació per a finalitats acadèmiques. A més, presenta una estructura HTML relativament homogènia i ben organitzada, la qual cosa permet aplicar amb eficàcia tècniques d'extracció de contingut utilitzant les llibreries *Selenium* i *BeautifulSoup*. Finalment, EV Database és una referència popular en portals de comparació de vehicles elèctrics i una de les fonts més citades en anàlisis de mercat, la qual cosa n'assegura la fiabilitat i la rellevància per al nostre projecte. Per tot això, s'ha considerat una font idònia per dur a terme aquesta pràctica, centrant-nos en l'extracció de dades de manera ètica i responsable.

2.- Títol del projecte

“EVTech: Característiques tècniques i comercials de vehicles elèctrics a EV Database”

El títol reflecteix la naturalesa del projecte: un procés d’extracció i estructuració de dades que permet obtenir informació tècnica i econòmica dels principals models de vehicles elèctrics disponibles al mercat europeu.

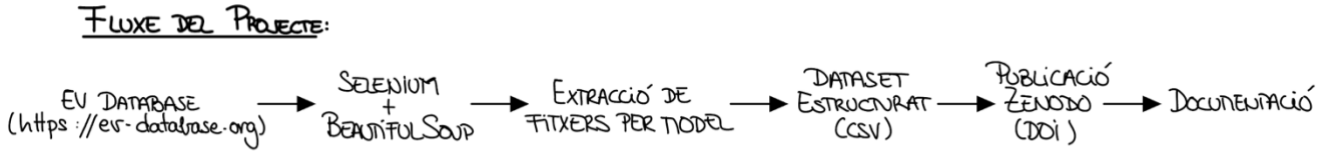
3.- Descripció del dataset

El dataset resultant conté informació tècnica i comercial de diversos models de vehicles elèctrics disponibles al mercat europeu. Cada fila del conjunt de dades representa un vehicle concret, mentre que les columnes descriuen atributs rellevants com la marca, el model, la capacitat útil de la bateria, l'autonomia elèctrica, la potència, la velocitat màxima, el consum energètic, les dimensions o el preu de referència per a un país concret. La informació s'ha extret directament de les fitxes públiques del lloc web EV Database mitjançant un procés automatitzat de web scraping, desenvolupat amb les llibreries Selenium i BeautifulSoup, que permeten accedir a contingut dinàmic i parsejar-lo de manera eficient.

Aquest conjunt de dades pot utilitzar-se per a diverses finalitats analítiques i d'investigació. En primer lloc, permet realitzar anàlisis comparatives de rendiment i preus entre diferents models o marques de vehicles elèctrics, identificant patrons o tendències dins del mercat europeu. També resulta útil per estudiar la relació entre la capacitat de la bateria i l'autonomia efectiva dels vehicles, un aspecte clau per entendre els avanços tecnològics en eficiència energètica. Finalment, el dataset pot servir per analitzar la diversitat de fabricants i la distribució de prestacions (potència, càrrega, velocitat o dimensions), oferint una visió global de l'oferta actual de vehicles elèctrics i dels segments de mercat predominants.

Les dades s'han extret durant la setmana del 4 al 8 de novembre de 2025, i s'han desat en format CSV amb separador de punt i coma (;) per facilitar-ne la interoperabilitat amb eines d'anàlisi posteriors. Tot i que el procés d'extracció es va veure parcialment interromput per un bloqueig temporal del lloc web, les dades obtingudes són representatives i suficients per assolir els objectius de la pràctica i demostrar la viabilitat del mètode implementat.

4.- Representació gràfica



El flux general del projecte segueix un procés estructurat que reflecteix les principals etapes del cicle de vida de les dades, des de la identificació de la font fins a la publicació final del conjunt de dades. En primer lloc, s'ha seleccionat el lloc web **EV Database (https://ev-database.org)** com a font d'informació, ja que conté dades rellevants sobre vehicles elèctrics organitzades en fitxes individuals. A partir d'aquesta font, el nostre *scraper* desenvolupat en **Python** inicia una sessió de navegació automatitzada amb **Selenium**, que permet carregar el contingut dinàmic del web i simular el comportament d'un usuari real. Un cop carregada cada pàgina, la llibreria **BeautifulSoup** analitza el codi HTML i n'extreu la informació pertinent, com ara les dades tècniques i comercials de cada model.

Aquest procés d'extracció converteix el contingut no estructurat del web en un **conjunt de dades tabular**. Les dades s'emmagatzemen progressivament en una estructura interna de Python i, un cop finalitzada la recollida, es guarden en un fitxer **CSV** sota el nom ev_dataset.csv, amb un format interoperable que permet la seva posterior anàlisi. Finalment, aquest dataset s'ha publicat al repositori obert **Zenodo**, assegurant la seva preservació, traçabilitat i disponibilitat pública.

En resum, l'esquema que representa el projecte segueix el flux:

EV Database → Selenium + BeautifulSoup → Dataset estructurat (CSV) → Publicació a Zenodo.

Aquesta seqüència il·lustra com un conjunt de dades web s'ha transformat en un recurs digital obert i reutilitzable mitjançant tècniques de web scraping responsables i documentades.

5.- Contingut del dataset

El dataset inclou variables tant **categòriques** com **numèriques**, amb la finalitat d'oferir una visió global de cada model. Les principals columnes són:

- Brand: marca del vehicle (Tesla, BMW, Volkswagen, etc.).
- **Model**: nom complet del model.
- **Price**: preu de referència per a Alemanya (€).
- **Availability**: disponibilitat comercial (p. ex. "Available to order", "Expected", "Discontinued").
- **Useable Battery**: capacitat útil de la bateria (kWh).
- **Charge Speed**: velocitat de càrrega o potència màxima de càrrega.
- **Acceleration 0–100 km/h**: temps d'acceleració (segons).
- **Top Speed**: velocitat màxima (km/h).
- **Total Power**: potència total del sistema elèctric (kW).
- **Electric Range**: autonomia elèctrica estimada (km).
- **Vehicle Consumption**: consum energètic mitjà (kWh/100 km).
- **Length**, **Width**, **Height**: dimensions del vehicle (mm).
- **Cargo Volume**: volum del maleter (litres).
- **Towing Weight Braked**: capacitat de remolc amb fre (kg).
- **Seats**: nombre de places.

El dataset cobreix informació recollida durant la setmana del **4 al 8 de novembre de 2025**. Les dades corresponen a la versió pública del lloc web durant aquest període.

6.- Propietari i aspectes legals

El propietari de la informació és **EV Database B.V.**, entitat titular del lloc web. Abans de realitzar el procés de *scraping*, s'ha revisat el fitxer robots.txt i s'ha comprovat que les seccions públiques del lloc no presentaven restriccions específiques per a l'accés.

Amb l'objectiu de respectar les normes d'ús i minimitzar l'impacte al servidor, s'han aplicat diverses bones pràctiques:

- Limitació del nombre de vehicles descarregats.
- Introducció d'esperes aleatòries entre peticions (3–7 segons).
- Aturada immediata en detectar pàgines amb missatges de bloqueig ("Request blocked" o "anomalies detected").
- Ús d'un navegador Chrome real en mode *headless* per garantir la compatibilitat amb el contingut dinàmic.
- Documentació transparent del procés i del codi utilitzat.

Durant les proves, el lloc web va bloquejar temporalment les peticions per detectar activitat automatitzada. Davant d'aquesta situació, el procés es va aturar de manera ètica i responsable, mantenint únicament les dades obtingudes fins aquell moment. El projecte s'ha dut a terme exclusivament amb finalitats acadèmiques, sense ús comercial ni redistribució dels continguts originals.

7.- Inspiració i justificació

L'objectiu principal d'aquest treball és demostrar la capacitat de planificar i implementar un procés de web scraping complet, des de la selecció del lloc web fins a la generació d'un conjunt de dades interoperable.

El domini escollit, els vehicles elèctrics, ofereix una gran riquesa de dades i un interès analític evident. A partir d'aquest dataset, es podrien desenvolupar futurs estudis com ara:

- L'anàlisi de la relació entre preu i autonomia.
- La comparació d'eficiència entre marques o tipus de bateria.
- L'evolució de les prestacions dels vehicles al llarg del temps.

Aquesta pràctica ens ha permès entendre millor les dificultats associades al web scraping de llocs dinàmics, com la detecció de comportaments automatitzats, i ha reforçat la importància d'adoptar estratègies ètiques i respectuoses amb els recursos digitals.

8.- Llicència

El **codi font** desenvolupat per a aquest projecte es distribueix sota la **llicència MIT**, que permet la seva reutilització, modificació i adaptació amb reconeixement d'autoria.

El **dataset generat** (ev_dataset.csv) s'ha publicat a **Zenodo** sota la llicència **Creative Commons Attributtion 4.0 International**, que permet la reutilització de les dades sempre que se'n citi l'origen i es comparteixi sota la mateixa llicència.

DOI: https://doi.org/10.5281/zenodo.17575547

9.- Codi font

El codi s'ha implementat en **Python**, i utilitza les llibreries següents:

- selenium per controlar el navegador Chrome.
- webdriver-manager per descarregar automàticament la versió adequada del *driver*.
- beautifulsoup4 per analitzar el contingut HTML.
- requests i pandas per manipular i emmagatzemar les dades.

El programa principal main.py crea una instància de la classe EVDatabaseScraper, executa el mètode scrape() i finalment exporta el resultat a CSV amb data2csv(). El flux d'execució és automàtic i permet afegir un paràmetre limit per controlar el nombre de vehicles descarregats.

Per executar-lo: python main.py

L'script està documentat i disponible al repositori GitHub:

https://github.com/joan-mata/Practica_1_ Web_Scraping

10.- Dataset i publicació

El dataset resultant s'ha desat com a: dataset/ev_dataset.csv i posteriorment s'ha publicat a Zenodo amb la corresponent descripció, etiquetes i llicència oberta.

El fitxer conté capçaleres clares i un conjunt suficient de registres per analitzar tendències bàsiques del mercat de vehicles elèctrics.

11.- Vídeo explicatiu

S'ha enregistrat un vídeo d'una durada aproximada de **8 minuts**, on tots dos membres del grup expliquem:

- El context i la motivació del projecte.
- L'arquitectura del codi i el funcionament de Selenium i BeautifulSoup.
- Les dificultats trobades, especialment el bloqueig del lloc web i les mesures aplicades.
- El resultat final i possibles aplicacions del dataset.

Enllaç al vídeo (Google Drive UOC):

https://drive.google.com/file/d/1LyJqAXbiK4_UrwVPZx_bYDh6VNmMqRlg/view?usp=share_link

També ha sigut pujat a Youtube en ocult per si hi havia algun error amb el link:

https://youtu.be/6xMLZPFqxv8

13.- Conclusions

Aquesta pràctica ens ha permès consolidar els conceptes apresos sobre el cicle de vida de les dades, especialment en la fase d'obtenció. Hem comprovat que el web scraping és una eina potent però que requereix una planificació acurada, coneixements tècnics i una actitud responsable davant les fonts d'informació.

Malgrat les dificultats trobades —com el bloqueig temporal per part del servidor—, s'ha aconseguit desenvolupar un sistema funcional capaç d'extreure i estructurar dades reals. L'experiència ha evidenciat la importància de respectar les bones pràctiques, la legislació vigent i els principis d'ètica digital en qualsevol projecte de ciència de dades.

Contribucions	Signatura
Investigació prèvia	Ari Pidevall, Joan Mata
Redacció de les respostes	Ari Pidevall, Joan Mata
Desenvolupament del codi	Ari Pidevall, Joan Mata
Participació al vídeo	Joan Mata*

*Tenim un correu que l'Ari està exempt de participar que baixa mèdica. Participa en l'elaboració però no grava.