
This is the **published version** of the bachelor thesis:

Marqués Soto, Mauri. *Sistema Intel·ligent d'Optimització de Càrrega per a Vehicles Elèctrics Mitjançant Reinforcement Learning i Machine Learning*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334476>

under the terms of the  license.

Sistema d'Optimització de Càrrega per a Vehicles Elèctrics Mitjançant Reinforcement Learning i Machine Learning

Mauri Marquès Soto

29 de juny de 2026

Resum– Aquest treball presenta el disseny, l'entrenament i la validació d'un agent d'Aprenentatge per Reforç per a l'optimització econòmica de la càrrega de vehicles elèctrics en entorns domèstics. El problema es formula com un Procés de Decisió de Markov on l'agent decideix cada hora si carregar o esperar, minimitzant el cost energètic basat en la tarifa PVPC espanyola mentre garanteix que el vehicle assoleixi el nivell de càrrega objectiu abans de l'hora de sortida prevista. Es comparen múltiples algorismes i estratègies de referència sobre un entorn de simulació realista que incorpora potència domèstica dinàmica i aleatorització del domini. L'agent DQN amb arquitectura profunda seleccionat assoleix un cost mitjà de 5,05 € amb un gap respecte a l'òptim teòric de només el 0,02%, un estalvi del 45,0% respecte a la càrrega immediata sense optimització, i un 100% d'èxit en 1.000 escenaris de validació. Finalment, el model s'integra en un pipeline complet orientat a prova de concepte, amb connexió a APIs de dades de preu en temps real, predicció del comportament de l'usuari mitjançant models de Machine Learning i comunicació directa amb el vehicle via l'API oficial del fabricant.

Paraules clau– Aprenentatge per Reforç, Deep Q-Network, Vehicle Elèctric, Optimització de Càrrega, PVPC, LightGBM, Proof of Concept, Gymnasium.

Abstract– This work presents the design, training and validation of a Reinforcement Learning agent for the economic optimization of electric vehicle charging in domestic environments. The problem is formulated as a Markov Decision Process where the agent decides each hour whether to charge or wait, minimizing energy cost based on the Spanish PVPC tariff while ensuring the vehicle reaches the target charge level before the predicted departure time. Multiple algorithms and baseline strategies are compared on a realistic simulation environment incorporating dynamic household power and domain randomization. The selected DQN agent with a deep architecture achieves an average cost of 5.05 € with a gap from the theoretical optimum of only 0.02%, a 45.0% cost reduction over the immediate-charging baseline, and a 100% success rate across 1,000 validation scenarios. Finally, the model is integrated into a complete proof-of-concept pipeline, including connections to real-time electricity price APIs, user behaviour prediction via Machine Learning models, and direct vehicle communication through the manufacturer's official API.

Keywords– Reinforcement Learning, Deep Q-Network, Electric Vehicle, Charging Optimization, PVPC, LightGBM, Proof of Concept, Gymnasium.



1 INTRODUCCIÓ

LA penetració creixent del vehicle elèctric en l'àmbit domèstic ha generat nous reptes relacionats amb la gestió eficient de la càrrega. En un context on les tarifes elèctriques variables, com la tarifa PVPC espanyola, impliquen diferències significatives de preu al llarg del

- E-mail de contacte: mauri.marques@autonoma.cat
- Menció: Enginyeria de Computadors
- Treball tutoritzat per: Antonio Lozano Bagen
- Curs 2025/26

dia, la decisió de quan carregar el vehicle té un impacte econòmic real per a l'usuari. Malgrat això, és habitual que els conductors connectin el vehicle en arribar a casa per comoditat, sense considerar si aquell moment és especialment car o si existeixen franges horàries més favorables dins del marge disponible.

Davant d'aquest escenari, la planificació intel·ligent de la càrrega representa una oportunitat per reduir costos, adaptar el procés a les necessitats reals de l'usuari i alinear el consum amb situacions més favorables per al sistema elèctric. Aquesta oportunitat, però, implica resoldre un problema de decisió seqüencial sota incertesa, en què intervenen simultàniament factors energètics, temporals i operatius que evolucionen amb el temps: el preu actual i previst de l'energia, la durada disponible de la sessió, la potència accessible a la llar, la potència del carregador, l'estat actual de la bateria i el nivell de càrrega objectiu.

1.1 Marc del projecte

Aquest Treball de Fi de Grau s'emmarca en el desenvolupament d'una prova de concepte real a SEAT S.A., on s'ha construït un sistema complet d'optimització de càrrega intel·ligent per a vehicles CUPRA connectats. El sistema integra un pipeline funcional que combina la lectura de telemetria del vehicle mitjançant l'API oficial del fabricant, la consulta de preus elèctrics en temps real a través de l'API de Red Eléctrica de España, la predicció de l'hora de sortida de l'usuari mitjançant models de Machine Learning, i un agent d'Aprenentatge per Reforç que, a partir de tota aquesta informació, decideix cada hora si el vehicle ha de carregar o esperar.

El present treball se centra en la part nuclear d'aquest sistema: la modelització del problema com un entorn de decisió seqüencial, la comparació rigorosa d'algorismes de control, i l'entrenament i validació de l'agent fins a assolir un rendiment pròxim a l'òptim teòric. La integració en el pipeline complet es descriu posteriorment com a resultat aplicat del projecte.

1.2 Objectius

L'objectiu general d'aquest treball és dissenyar, desenvolupar i validar un sistema basat en Aprenentatge per Reforç per a l'optimització econòmica de la càrrega d'un vehicle elèctric, garantint en tot moment la disponibilitat del vehicle quan l'usuari el necessita. D'aquest objectiu se'n deriven els objectius específics següents:

- Definir formalment el problema de càrrega intel·ligent com un Procés de Decisió de Markov i desenvolupar un entorn de simulació realista basat en Gymnasium.
- Dissenyar una funció de recompensa alineada amb els objectius del problema, prioritzant el compliment de la càrrega requerida i la reducció del cost econòmic.
- Comparar diversos algorismes de RL i estratègies de referència per identificar quin enfocament ofereix millors resultats.
- Entrenar i optimitzar l'agent seleccionat fins a assolir un rendiment pròxim a l'òptim teòric en escenaris estocàstics realistes.

- Validar el model en el context d'una prova de concepte orientada a un cas d'ús real.

2 ESTAT DE L'ART

2.1 Mètodes tradicionals d'optimització de càrrega

Els primers sistemes de gestió intel·ligent de càrrega van adoptar tècniques d'optimització matemàtica clàssiques. La Programació Lineal (LP) i la Programació Lineal Entera Mixta (MILP) permeten obtenir solucions òptimes en escenaris deterministes i faciliten la incorporació de restriccions físiques com els límits de potència o la capacitat de la bateria [1]. El Control Predictiu de Model (MPC) estén aquest enfocament aplicant una finestra d'optimització mòbil que es recalcula periòdicament a mesura que arriben noves dades.

Tot i els seus avantatges teòrics, aquests mètodes presenten limitacions crítiques en entorns reals. La seva efectivitat depèn d'una previsió precisa de variables futures com els preus de l'electricitat, la demanda de la xarxa o el comportament de l'usuari, informació que en la pràctica és incerta o incompleta [2]. A més, la complexitat computacional creix exponencialment amb l'escala del problema, cosa que els fa poc pràctics per a decisions en temps real en entorns domèstics dinàmics [3].

2.2 Reinforcement Learning aplicat a la gestió energètica

L'Aprenentatge per Reforç (RL) planteja el problema de càrrega com un Procés de Decisió de Markov (MDP) on un agent aprèn una política òptima a partir de la interacció amb l'entorn, sense necessitat d'un model matemàtic explícit del sistema. Això el fa especialment adequat per a entorns estocàstics amb preus dinàmics i comportament impredecible de l'usuari [4].

2.2.1 Mètodes basats en valor: DQN

Els algorismes basats en valor, com el Deep Q-Network (DQN), aproximen la funció $Q(s, a)$, que mesura la qualitat esperada de cada acció en cada estat, mitjançant una xarxa neuronal profunda, la qual cosa permet operar sobre espais d'estats continus i d'alta dimensionalitat. L'arquitectura es basa en tècniques d'estabilització com el reaprofitament d'experiències passades i l'ús d'una xarxa auxiliar amb pesos actualitzats periòdicament, cosa que millora la convergència durant l'entrenament. En el context de la càrrega domèstica, DQN és especialment adequat per a espais d'acció discrets i binaris (carregar/no carregar), i treballs recents demostren estalvis superiors al 20% en la factura elèctrica de l'usuari sota tarifes variables [5].

2.2.2 Mètodes basats en política: PPO

El Proximal Policy Optimization (PPO) és un algorisme on-policy que aprèn directament la política $\pi(s)$, és a dir, la regla de decisió que assigna una acció a cada estat, i ofereix bona estabilitat gràcies al mecanisme de *clipping* que limita la magnitud de les actualitzacions. Tot i funcionar

tant per a accions discretes com contínues i ser àmpliament usat en aplicacions de càrrega multi-agent [4], la naturalesa on-policy el fa menys eficient en mostres que DQN en problemes de decisió binària simples.

2.2.3 Mètodes actor-critic: SAC i DDPG

Els mètodes actor-critic avançats, com el Soft Actor-Critic (SAC) i el Deep Deterministic Policy Gradient (DDPG), combinen l'estimació de valor amb l'optimització directa de la política per a espais d'acció continus, com ara la modulació de la potència de càrrega. SAC incorpora un terme d'entropia que afavoreix l'exploració automàtica i representa l'estat de l'art en control continu [6]. No obstant això, en problemes amb acció binària, la complexitat addicional d'aquests mètodes no aporta beneficis respecte a DQN i incrementa el cost computacional d'entrenament.

2.3 Modelatge del comportament de l'usuari

Un dels reptes principals en l'aplicació de RL a la càrrega domèstica és la incertesa sobre l'hora de sortida del vehicle. La majoria de treballs acadèmics assumeixen que l'usuari introdueix manualment la seva hora de sortida en connectar el vehicle, però en entorns industrials reals aquesta interacció és poc freqüent i poc fiable.

La literatura proposa arquitectures híbrides en dues etapes per abordar aquesta limitació: en una primera fase, un model supervisat analitza dades històriques de sessions i prediu l'hora de sortida i l'energia necessària a partir de variables com l'hora d'arribada, el dia de la setmana o l'identificador de l'usuari; en una segona fase, l'agent RL incorpora aquesta predicció al seu vector d'estat. Chung et al. [7] mostren que algorismes d'aprenentatge automàtic supervisat ofereixen bona precisió en entorns residencials, i Frendo et al. [8] confirmen que integrar prediccions de sortida en sistemes de càrrega intel·ligent millora significativament la prioritització i la satisfacció de l'usuari.

2.4 Posicionament del treball

La Taula 1 sintetitza les principals característiques de les famílies d'algorismes estudiades. En aquest treball, l'elecció de DQN respon a tres factors: l'espai d'acció és discret i binari, el problema és monoagent i domèstic, i es prioritza l'eficiència de mostres i la interpretabilitat de la política. La integració d'un model de predicció del comportament de l'usuari basat en LightGBM segueix la línia de les arquitectures híbrides identificades a la literatura, i respon a la necessitat de dotar l'agent d'informació fiable sobre l'horitzó de planificació en un entorn real.

TAULA 1: COMPARATIVA DE FAMÍLIES D'ALGORISMES DE RL PER A CÀRREGA DE VEHICLES ELÈCTRICS

Algorisme	Acció	Off-pol.	Ef. mostres	Binari
DQN	Discreta	Sí	Alta	Sí
PPO	Disc./Cont.	No	Mitjana	Sí
DDPG	Contínua	Sí	Alta	No
SAC	Contínua	Sí	Alta	No

3 FORMULACIÓ DEL PROBLEMA

3.1 Definició del MDP

El problema de càrrega intel·ligent es formalitza com un Procés de Decisió de Markov (MDP) definit per la tupla $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$.

Espai d'estats $\mathcal{S}$ (24 dimensions). El vector d'observació que rep l'agent a cada pas horari t és:

- *Variables base*
- *Variables base [0–6]*: preu normalitzat actual $p_t \in [0, 1]$, temps restant $\frac{T_{dep}-t}{14}$, estat de càrrega actual de la bateria (*State of Charge*, SoC) $s_t \in [0, 1]$, SoC objectiu s^* , ràtio de potència efectiva $\frac{\min(P_{ch}, P_{free})}{P_{ch}}$, potència del carregador normalitzada $\frac{P_{ch}}{22}$ i capacitat de la bateria normalitzada $\frac{C_{bat}}{100}$.
- *Finestra de preus futurs [7–20]*: els 14 preus horaris normalitzats de la sessió.
- *Característiques de planificació [21–23]*: hores de càrrega necessàries normalitzades $h_{need} = \frac{E_{miss}}{P_{ch} \cdot \eta}$, ranking de preu actual $r_{price} \in [0, 1]$ (0 = més car, 1 = més barat respecte als preus restants), i urgència $u = \min\left(\frac{h_{need}}{T_{rem}}, 1.5\right)$.

Espai d'accions $\mathcal{A}$. L'acció és binària: $a_t \in \{0, 1\}$, on 0 correspon a esperar i 1 a carregar a la potència màxima disponible.

Transició d'estat $\mathcal{T}$. La potència efectiva de càrrega en cada pas queda limitada per la potència lliure de la llar:

$$P_{eff} = \min(P_{ch}, \max(0, P_{cont} - P_{home,t})) \quad (1)$$

on P_{cont} és la potència contractada i $P_{home,t}$ el consum domèstic en el pas t . L'actualització del SoC és:

$$s_{t+1} = s_t + \frac{P_{eff} \cdot \eta \cdot \Delta t}{C_{bat}} \quad (2)$$

amb eficiència de càrrega $\eta = 0.95$ i pas temporal $\Delta t = 1$ h.

Factor de descompte: $\gamma = 0.99$.

3.2 La Regla d'Or

La restricció fonamental del sistema és que el vehicle ha d'assolir el nivell de càrrega objectiu s^* abans o en el moment de sortida T_{dep} , independentment de l'estratègia d'optimització emprada. Aquesta premissa, definida en el present treball com la *Regla d'Or*, subordina qualsevol decisió econòmica a la disponibilitat garantida del vehicle. La funció de recompensa reflecteix aquesta prioritat mitjançant penalitzacions creixents en funció de la desviació respecte al target en el moment de finalització de l'episodi.

3.3 Funció de recompensa

La recompensa total combina tres components: el cost monetari real, una senyal de modelatge per tick (*reward shaping*) i una recompensa terminal.

Cost monetari. En cada pas en què s'aplica l'acció de càrrega, es descompta el cost energètic real:

$$r_{cost} = -P_{eff} \cdot \eta \cdot \Delta t \cdot p_t \quad (3)$$

Reward shaping per tick. Per guiar l'aprenentatge cap a una política econòmicament racional, s'afegeix una senyal de forma basada en el ranking de preu i la urgència. Es defineix un llindar de ranking $\rho = 1 - \min\left(\frac{h_{need}}{T_{rem}}, 1\right)$: les hores amb $r_{price} \geq \rho$ es consideren *hores barates necessàries*. La senyal de forma r_{shape} s'aplica segons: +2.0 si es carrega en hora barata necessària; -2.5 si es carrega en hora cara amb marge ($u < 0.6$); -0.5 si es carrega en hora cara sota pressió ($u \in [0.6, 0.9]$); 0 si es carrega forçat per urgència crítica ($u \geq 0.9$); -1.5 si s'omet una hora barata necessària; i 0 si s'espera correctament en hora cara. La recompensa per tick és $r_t = r_{cost} + r_{shape}$.

Recompensa terminal. En finalitzar l'episodi s'aplica un dels règims següents en funció del gap $\delta = s^* - s_{T_{dep}}$: +20 si el target s'ha assolit ($\delta \leq 0$); -10 δ si $\delta \leq 5\%$; -30 δ si $5\% < \delta \leq 8\%$; i $-(50 + E_{miss})$ si $\delta > 8\%$, on $E_{miss} = \delta \cdot C_{bat}$ és l'energia no subministrada en kWh.

3.4 Entorn de simulació

L'entorn es construeix sobre el framework Gymnasium i implementa sessions de càrrega domèstica amb dades reals de preus PVPC i perfils de consum de la llar. Per garantir que l'agent aprengui una política robusta i generalitzable, s'aplica *domain randomization* en cada episodi: la capacitat de la bateria es mostreja uniformement de [30, 100] kWh, la potència del carregador s'escull de {2.3, 3.7, 7.4, 11, 22} kW, la potència contractada de {3.45, 4.6, 5.75, 6.9} kW, el SoC inicial de [0.05, 0.50], i l'hora de sortida de [6, 14] passos. Aquesta aleatorització obliga l'agent a generalitzar sobre la diversitat real d'instal·lacions domèstiques i hàbits d'ús dels usuaris.

L'entorn ha evolucionat de forma incremental al llarg del projecte. La versió V1 modelava un entorn estàtic on la potència domèstica era sempre màxima disponible. La versió V2 va introduir la potència domèstica dinàmica i la randomització del domini. La versió final V2.3 afegeix les tres característiques de planificació [21–23] i el sistema de recompensa per zones de tolerància, que permeten a l'agent operar amb un perfil agressiu en cost però segur en disponibilitat.

4 COMPARATIVA DE MODELS

4.1 Models avaluats i baselines

Per establir un punt de partida objectiu, es defineixen dues estratègies de referència deterministes que simulen comportaments d'usuari sense fer ús d'intel·ligència artificial.

La *Càrrega Immediata* simula el comportament més habitual: s'activa la càrrega des del primer pas de la sessió i es manté fins a assolir el target SoC, ignorant completament el senyal de preu. Garanteix el compliment de la Regla d'Or però representa el pitjor escenari d'optimització econòmica.

L'*Heurística de Preu* implementa una lògica dinàmica: en cada pas es calcula la mitjana dels preus futurs coneguts i només es carrega si el preu actual és inferior o igual a aquesta mitjana. Un mecanisme de seguretat força la càrrega quan

el temps restant és l'estrictament necessari per assolir el target, preservant la Regla d'Or.

Pel que fa als algorismes de RL, la comparativa s'estableix entre PPO i DQN. SAC es descarta en aquesta fase perquè està dissenyat per a espais d'acció continus i la seva aplicació sobre un espai binari no seria natural ni eficient [4]. S'avaluen dues arquitectures de xarxa: DQN estàndard [64,64] i Deep DQN [256,256]. Tots els models es basen en la implementació MlpPolicy de Stable-Baselines3 [9] i s'avaluen sobre un conjunt determinista de 100 episodis amb llavors fixes, garantint que tots els agents s'enfronten exactament a les mateixes condicions inicials.

4.2 Resultats de la comparativa

4.2.1 Entorn estàtic (EVChargingEnv V1)

En l'entorn estàtic, on la potència domèstica és sempre màxima disponible, ambdues baselines assoleixen el 100% d'èxit. La Càrrega Immediata obté un cost mitjà de 4,84 €, mentre que l'Heurística de Preu el redueix fins a 3,70 €, un estalvi del 23,5%.

L'agent PPO, configurat amb els hiperparàmetres per defecte, convergeix cap a un òptim local que mimetitzava exactament la Càrrega Immediata (4,84 €). Això es deu al disseny asimètric de la funció de recompensa: la penalització per incompliment del target és tan severa que l'agent adopta la política més conservadora possible per evitar-la. Per superar aquesta limitació, s'aplica un ajust d'hiperparàmetres que incrementa el coeficient d'entropia a 0,05, forçant l'agent a explorar l'acció d'esperar fins i tot quan ja coneix la política segura. Amb aquest ajust, el PPO baixa el cost a 3,85 €.

L'agent DQN, gràcies al seu mecanisme d'exploració ϵ -greedy natiu, no requereix cap ajust addicional i obté directament el millor resultat: 3,63 €, superant fins i tot l'heurística intel·ligent.

La Taula 2 consolida els resultats de la Fase 1.

TAULA 2: RESULTATS COMPARATIUS A EVCHARGINGENV V1 (100 EPISODIS)

Estratègia	Èxit	Cost	Reward
Càrrega Immediata	100%	4,84 €	5,164
Heurística de Preu	100%	3,70 €	6,303
PPO (sense ajustar)	100%	4,84 €	5,164
PPO (fine-tuned)	100%	3,85 €	6,154
DQN	100%	3,63 €	6,368

4.2.2 Entorn dinàmic (EVChargingEnv V2)

Per avaluar la robustesa dels models en condicions més realistes, es reavaluen les estratègies sobre l'entorn V2, que incorpora potència domèstica dinàmica i *full domain randomization*. La introducció de temps de sessió aleatoris (6–14 passos) fa que en aproximadament el 24% dels episodis sigui físicament impossible assolir el target, establint un límit termodinàmic de viabilitat del 76%.

Sota aquestes condicions, les heurístiques estàtiques col·lapsen: la Càrrega Immediata baixa al 76% d'èxit i l'Heurística de Preu cau al 63%, ja que la seva lògica d'espera

ra ignora la urgència temporal. L'agent DQN Base [64,64] iguala el límit físic del 76% però pateix un clar subajustament: amb només 64 neurones no té capacitat per representar les relacions no lineals entre potència, temps i energia, i acaba mimetitzant la Càrrega Immediata.

L'agent Deep DQN [256,256], entrenat amb una fracció d'exploració del 60%, és l'únic que aconsegueix un comportament diferencial: obté el cost mitjà més baix entre els agents viables (4,888 €) assumint un risc calculat que es tradueix en una taxa d'èxit del 74%, lleugerament per sota del límit físic. La Taula 3 recull els resultats.

TAULA 3: RESULTATS COMPARATIUS A EVCHARGINGENV V2 (100 EPISODIS)

Estratègia	Èxit	Cost	Reward
Càrrega Immediata	76%	5,039 €	-12,082
Heurística de Preu	63%	3,910 €	-19,592
DQN Base [64,64]	76%	4,968 €	-12,028
Deep DQN [256,256]	74%	4,888 €	-13,186

4.3 Selecció del model final

Els resultats de la comparativa permeten extreure conclusions determinants. En l'entorn estàtic, DQN supera tant les heurístiques com PPO gràcies a l'estimació directa del valor $Q(s, a)$ per a cada acció discreta, cosa que el fa especialment eficient en espais d'acció binaris. En l'entorn dinàmic, Deep DQN és l'únic model que va més enllà del comportament conservador i troba un equilibri real entre estalvi econòmic i compliment del servei.

Per aquests motius, l'arquitectura Deep DQN es selecciona com a model de referència per a la fase d'entrenament i optimització. El següent repte consisteix a millorar la qualitat de la política apresada fins a assolir un rendiment pròxim a l'òptim teòric mitjançant un entorn de simulació més ric i un disseny de recompensa més precís.

5 ENTRENAMENT I OPTIMITZACIÓ

5.1 Millores de l'entorn: EVChargingEnv V2.3

Un cop seleccionat Deep DQN com a algorisme de referència, l'esforç s'ha centrat en millorar la qualitat de la política apresada mitjançant dos canvis estructurals respecte a l'entorn V2: l'ampliació de l'espai d'observació i el redisseny de la funció de recompensa. Pel que fa a l'arquitectura de xarxa, un cop enriquit l'espai d'observació amb les característiques de planificació explícites, una xarxa de [128,128] neurones resulta suficient per capturar les relacions del problema sense la capacitat addicional de [256,256], de manera que s'adopta aquesta configuració més lleugera per al model definitiu.

L'espai d'observació passa de 21 a 24 dimensions incorporant tres característiques de planificació explícites: les hores de càrrega necessàries normalitzades (h_{need}), el ranking del preu actual dins la finestra restant (r_{price}) i la urgència (u). Aquestes variables exposen a l'agent la informació de planificació que en V2 havia d'inferir

implícitament, cosa que accelera l'aprenentatge i millora la qualitat de les decisions en situacions de marge ajustat. La formulació matemàtica d'aquestes característiques es detalla a la Secció 3.

La funció de recompensa s'estén amb un *reward shaping* per tick i zones de tolerància terminal, també descrits a la Secció 3, que guien l'agent cap a un perfil agressiu en cost però segur en disponibilitat: l'agent aprèn a carregar en les hores més barates necessàries i a acceptar lleugeres desviacions del target (fins a un 8%) quan fer-ho permet un estalvi econòmic significatiu.

5.2 Protocol de validació

L'avaluació es realitza sobre 1.000 escenaris fixos i factibles generats per *rejection sampling* amb $seed=42$ a partir del conjunt de validació. S'apliquen dos criteris de filtratge: coherència física ($P_{ch} \leq P_{cont}$) i factibilitat del target (existeix almenys una assignació d'hores que permet assolir el SoC objectiu). La taxa d'acceptació del mostreig és del 19,6%, cosa que confirma que una fracció important de l'espai original de configuracions correspon a escenaris físicament irrealitzables. Restringir l'avaluació a escenaris coherents i factibles garanteix que les diferències observades entre polítiques reflecteixen qualitat de decisió i no limitacions de hardware.

Com a cota inferior teòrica s'utilitza un Oracle determinista que coneix tots els preus de l'episodi i carrega en les hores més barates fins a cobrir l'energia necessària, respectant la potència efectiva disponible en cada hora. Cap política pot assolir el target amb un cost estrictament inferior al de l'Oracle.

5.3 Configuració d'entrenament

La Taula 4 mostra els canvis d'hiperparàmetres respecte a V2.2. La resta de paràmetres (taxa d'aprenentatge $\alpha = 10^{-4}$, $\gamma = 0.99$, *batch size* 64, arquitectura [128,128]) es mantenen, ja que el nou espai d'observació no és fonamentalment més difícil, sinó més ric en informació.

TAULA 4: CANVIS D'HIPERPARÀMETRES DE V2.2 A V2.3

Hiperparàmetre	V2.2	V2.3
learning_starts	10.000	20.000
exploration_fraction	0,15	0,25
exploration_final_eps	0,05	0,03
target_update_interval	1.000	2.000
total_timesteps	1,5M	2,5M

L'increment de *learning_starts* i *exploration_fraction* respon a la necessitat d'explorar les subtilitats del nou *reward shaping*: l'agent necessita més experiències diverses abans de convergir per aprofitar correctament les senyals de r_{price} i u . La reducció de *exploration_final_eps* afavoreix una major explotació al final de l'entrenament, on la política ja és prou madura.

5.4 Política apresada

La Fig. 1 mostra el heatmap de la probabilitat de carregar en funció del ranking de preu i la urgència, reconstruït a partir de les senyals internes de l'agent sobre els 1.000 escenaris de validació.

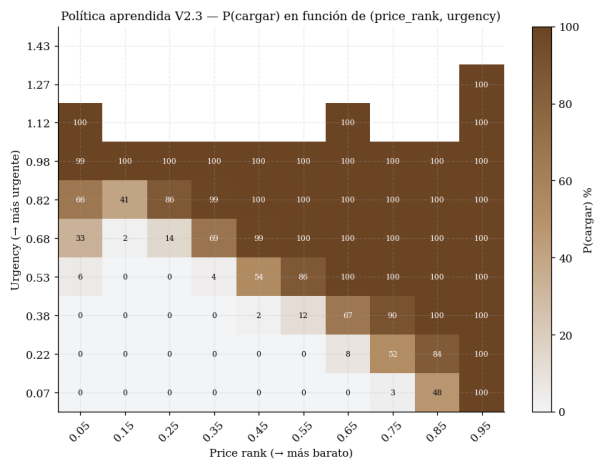


Fig. 1: Política apresada: P(carregar) en funció de price_rank i urgency. Cada cel·la mostra el percentatge de vegades que l'agent activa la càrrega per a aquella combinació de ranking de preu (0 = més car, 1 = més barat) i urgència (0 = lliure, 1 = crític).

El heatmap revela un comportament coherent amb la política òptima teòrica. Quan la urgència és baixa ($u < 0.4$), l'agent carrega gairebé exclusivament en les hores amb ranking superior a 0,85, és a dir, les més barates de la sessió. A mesura que la urgència augmenta, el llindar de preu acceptable es redueix progressivament: per $u \in [0.6, 0.9]$, l'agent accepta carregar en hores de preu mitjà per evitar el risc de no assolir el target. Per $u \geq 0.9$, l'agent carrega independentment del preu, activant el comportament de seguretat forçat pel disseny de la recompensa. Aquest patró confirma que l'agent ha interioritzat la Regla d'Or i l'equilibri entre estalvi econòmic i disponibilitat del vehicle.

6 RESULTATS I AVALUACIÓ

6.1 Rendiment absolut de V2.3

Sobre els 1.000 escenaris de validació factibles, l'agent V2.3 assoleix un rendiment pràcticament òptim en totes les mètriques avaluades. El quadre següent resumeix els resultats principals:

100% d'èxit ($\text{gap} \leq 5\%$ i $\leq 8\%$)
5,05 € cost mitjà per sessió
0,02% gap mitjà respecte al target
+0,22% bretxa màxima respecte a l'Oracle
0 fallades ($\text{gap} > 8\%$) en 1.000 escenaris

El gap màxim observat és de l'1,11%, molt per sota del llindar tolerable del 8%. Aquests resultats indiquen que l'agent ha après a aprofitar les zones de tolerància terminal sense vulnerar la Regla d'Or en cap dels escenaris avaluats.

6.2 Comparativa V2.2 vs V2.3 vs Oracle

La Fig. 2 compara les tres polítiques sobre els mateixos 1.000 escenaris.

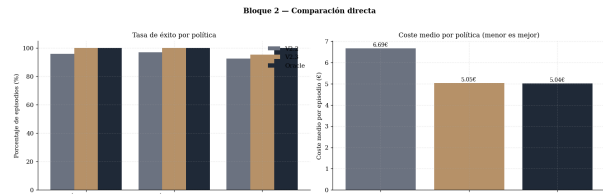


Fig. 2: Comparació directa de V2.2, V2.3 i Oracle. Esquerra: taxes d'èxit per criteri. Dreta: cost mitjà per sessió.

V2.3 millora V2.2 en totes les mètriques: la taxa d'èxit $\leq 5\%$ puja del 95,8% al 100%, el cost mitjà baixa de 6,69 € a 5,05 € (un estalvi del 24,58%), i el gap mitjà es redueix del 0,60% al 0,02%. La bretxa entre V2.3 i l'Oracle és de tan sols 0,22% en cost (5,05 € vs 5,04 €), cosa que situa l'agent a un rendiment pràcticament òptim. En el 71,4% dels episodis, V2.3 obté un cost estrictament inferior al de V2.2.

La Taula 5 consolida les mètriques principals.

TAULA 5: MÈTRIQES COMPARATIVES SOBRE 1.000 ESCENARIS FACTIBLES

Mètrica	V2.2	V2.3	Oracle
Èxit $\leq 5\%$ gap	95,8%	100%	100%
Èxit $\leq 8\%$ gap	96,9%	100%	100%
Target exacte	92,4%	95,3%	100%
Cost mitjà	6,69 €	5,05 €	5,04 €
Gap mitjà	0,60%	0,02%	0,00%

6.3 Diagnòstic de fallades i casos d'estudi

Sobre els 1.000 escenaris factibles, l'agent V2.3 no presenta cap fallada amb $\text{gap} > 8\%$, confirmant que les desviacions residuals es mantenen sempre dins la zona de tolerància acceptable. Les fallades de V2.2 es concentraven en escenaris d'urgència alta on la manca de les característiques de planificació [21–23] impedia a l'agent anticipar correctament el risc de no assolir el target.

L'anàlisi de tres casos d'estudi representatius il·lustra el comportament de l'agent en funció del marge disponible. En el Cas 1 (slack alt), l'agent acumula la càrrega en les últimes hores de la sessió, que coincideixen amb els preus més baixos, però perd algunes oportunitats òptimes identificades per l'Oracle, resultant en una bretxa de cost del +27,9%. En els Casos 2 i 3 (slack mitjà i ajustat), l'agent replica exactament la seqüència de càrrega de l'Oracle (bretxa 0,0%), demostrant que quan la urgència és moderada o alta la política apresada és pràcticament òptima. Els gràfics detallats d'aquests casos s'inclouen a l'Apèndix.

7 SISTEMA COMPLET: PROOF OF CONCEPT

El model entrenat i validat en les seccions anteriors s'integra en un sistema funcional orientat a prova de concepte,

desenvolupat internament a SEAT S.A. El sistema consisteix en un backend FastAPI que connecta el model DQN a vehicles CUPRA reals mitjançant l'API oficial del fabricant, i que opera de forma autònoma prenent decisions de càrrega cada hora.

7.1 Arquitectura general del pipeline

El sistema s'organitza en dues capes. La capa de serveis s'encarrega d'obtenir tota la informació necessària per construir el vector d'observació de 24 dimensions que l'agent necessita: l'estat real del vehicle (SoC, estat de l'endoll, potència de càrrega activa) via l'API del vehicle, els preus horaris de l'electricitat via l'API pública d'ESIOS/REE, i la predicció de l'hora de sortida de l'usuari mitjançant el model de Machine Learning. La capa de control conté el bucle de decisió, que s'executa cada hora per vehicle i aplica l'acció recomanada pel model DQN al vehicle real.

7.2 Connexió amb el vehicle

L'accés a la telemetria i al control del vehicle es realitza a través de l'API oficial del fabricant, que proporciona endpoints per a la lectura de l'estat de càrrega, l'inici de càrrega i l'aturada de càrrega. L'autenticació segueix un flux estàndard d'intercanvi de tokens que garanteix que les peticions estan autoritzades per l'usuari propietari del vehicle. Atès que les ordres són asíncrones, el sistema confirma cada actuació llegint l'estat real del vehicle uns instants després d'enviar la comanda.

7.3 Integració de dades externes i predicció de l'usuari

Els preus horaris de la tarifa PVPC s'obtenen diàriament a través de l'API pública d'ESIOS de Red Eléctrica de España, que publica la corba de preus de les 24 hores del dia següent a les 20:15h. El perfil de consum domèstic s'estima a partir dels coeficients horaris publicats per la REE. Aquestes dues fonts constitueixen la component econòmica i energètica del vector d'observació.

La predicció de l'hora de sortida, necessària per calcular el temps restant i la urgència, s'obté mitjançant dos models LightGBM [10] entrenats sobre l'historial real de sessions de càrrega del vehicle. El primer model prediu l'energia que consumirà el vehicle fins a la propera connexió; el segon prediu les hores de presència previstes, que es tradueixen directament al *departure step* de l'agent. Tots dos models s'entrenen amb un conjunt de 14 features que combinen informació temporal (hora d'endollar codificada sinusoidalment, dia de la setmana, mes), informació de l'estat del vehicle a l'arribada (SoC inicial, dies des de la sessió anterior) i el perfil d'ús habitual de l'usuari (mediana d'energia i presència de sessions anteriors, amb *shrinkage* cap a la mitjana mostral per a usuaris amb poques dades).

La validació es realitza amb *TimeSeriesSplit* per respectar l'ordre cronològic de les sessions i evitar fuga d'informació futura. Els models superen de forma consistent el baseline de perfil d'usuari, que representa el que ja es pot predir sense cap context del dia. La capacitat de la bateria de cada usuari, no disponible directament, s'estima

a partir de l'historial de sessions amb càrrega significativa ($\Delta SoC \geq 20\%$), usant la relació entre l'energia carregada i la variació de SoC.

Aquesta arquitectura de predicció segueix la línia de les arquitectures híbrides identificades a l'estat de l'art (Secció 2), on un model supervisat proveeix informació de context a l'agent de RL, permetent que el sistema operi sense requerir que l'usuari introdueixi manualment l'hora de sortida prevista.

7.4 Cicle de decisió en temps real

En cada iteració horària, el sistema llegeix l'estat real del vehicle via l'API del fabricant, construeix el vector d'observació de 24 dimensions descrit a la Secció 3, i l'envia al model DQN per obtenir l'acció recomanada (carregar o esperar). L'actuació sobre el vehicle només es produeix quan hi ha una transició respecte a l'estat anterior, cosa que minimitza el nombre de comandes enviades al vehicle. El sistema incorpora salvaguardes addicionals com un màxim d'actuacions per dia i un interval mínim entre accions consecutives, i s'atura automàticament en arribar al *departure step* predit.

D'aquesta manera, el model entrenat en un entorn de simulació estocàstica opera sobre dades reals del vehicle, tancant el cicle entre la recerca acadèmica i l'aplicació industrial.

8 CONCLUSIONS

8.1 Conclusions tècniques

Aquest treball ha demostrat la viabilitat de l'Aprenentatge per Reforç per a l'optimització econòmica de la càrrega de vehicles elèctrics en entorns domèstics reals, abordant el problema des de tres angles complementaris.

Pel que fa a la comparativa de models, els resultats confirmen que DQN supera PPO i les estratègies heurístiques en problemes amb espai d'acció binari, tant en condicions estàtiques com en entorns amb potència domèstica dinàmica i aleatorització completa del domini. L'arquitectura profunda resulta determinant per generalitzar sobre la diversitat de configuracions de bateria, carregador i horitzó temporal, superant el subajustament que presenta l'arquitectura base [64,64].

Pel que fa a l'entrenament i optimització, la incorporació de les tres característiques de planificació explícites (hores necessàries, ranking de preu i urgència) i el redisseny de la funció de recompensa amb *reward shaping* per tick i zones de tolerància terminal permeten a l'agent assolir un rendiment pràcticament òptim: 100% d'èxit en 1.000 escenaris factibles, gap mitjà de 0,02% i una bretxa de tan sols 0,22% en cost respecte a l'Oracle determinista. El heatmap de la política apresada confirma que l'agent ha interioritzat correctament la Regla d'Or i el compromís entre estalvi econòmic i disponibilitat del vehicle.

Finalment, la integració del model DQN en un pipeline complet amb connexió a APIs de preus en temps real, predicció del comportament de l'usuari mitjançant LightGBM i comunicació directa amb el vehicle via l'API oficial del fabricant demostra la viabilitat tècnica del sistema com a

prova de concepte industrial, validada sobre un vehicle CUPRA real.

8.2 Treball futur

El sistema desenvolupat ofereix diverses línies naturals d'extensió. En l'àmbit de la seguretat i la mobilitat, es preveu incorporar una capa de reserves mínimes d'energia que garanteixi sempre un nivell de càrrega suficient per cobrir necessitats d'urgència independentment de l'estratègia d'optimització, amb la possibilitat d'integrar serveis de geolocalització per adaptar dinàmicament aquest mínim en funció de la distància a punts d'interès crítics com hospitals o serveis d'emergència.

En l'àmbit energètic, s'identifiquen dues extensions rellevants: la integració amb sistemes de gestió energètica de la llar (HEMS) per coordinar la càrrega del vehicle amb altres dispositius del habitatge, i el suport per a generació solar fotovoltaica i bateria domèstica, que permetria al sistema explotar l'excedent solar i ampliar l'optimització més enllà del preu de la xarxa.

Des del punt de vista de la robustesa i l'escalabilitat, es preveu afegir un mode de funcionament *offline* que mantingui el sistema operatiu en absència de connectivitat a les APIs externes, així com la compatibilitat amb múltiples proveïdors i tarifes elèctriques més enllà del PVPC espanyol, cosa que permetria desplegar el sistema en altres mercats europeus.

Finalment, en la línia de l'agent de RL, queda pendent incorporar criteris de preservació de la salut de la bateria (SoH) com a component addicional de la funció de recompensa, penalitzant comportaments de càrrega potencialment perjudicials per a la durabilitat de la bateria.

ÚS DE LA IA GENERATIVA

Al llarg de l'elaboració d'aquest treball s'ha fet ús de la intel·ligència artificial generativa, concretament de l'eina Claude (Anthropic). La seva utilització s'ha centrat principalment en la millora de la redacció, la correcció d'errades ortogràfiques i gramaticals i l'optimització de l'estil per aconseguir un text més fluid i coherent, tant en la redacció de la memòria com en la documentació tècnica del projecte.

Addicionalment, Claude també s'ha emprat com a suport en la generació i revisió de fragments de codi, especialment en la implementació del pipeline d'integració, el disseny de l'entorn de simulació i la resolució de problemes tècnics puntuals. En aquests casos, l'eina ha actuat com a recurs de consulta i proposta de solucions, sempre revisades i adaptades posteriorment per garantir la correcta integració dins de l'arquitectura definida. En cap cas s'ha emprat per a la creació íntegra d'apartats o de codi, sinó que el contingut, les idees, les decisions tècniques i l'estructura del treball són exclusivament de l'autor, mentre que la IA ha actuat com a recurs de suport al llarg del procés.

AGRAÏMENTS

Vull expressar el meu agraïment, en primer lloc, al meu tutor acadèmic Antonio Lozano Bagen, pel seu acompanyament, orientació i consells al llarg de tot el desenvolupament d'aquest treball, que han estat clau per estructurar-lo i

donar-li coherència. També vull agrair molt especialment al meu tutor d'empresa Alejandro Gorge Planas i a tot l'equip de SEAT S.A. per haver-me donat l'oportunitat de dur a terme aquest projecte en un entorn industrial real, facilitant-me en tot moment els recursos, la informació i el context necessaris per fer-lo possible. Treballar en un projecte d'aquesta envergadura dins d'una empresa com SEAT ha estat una experiència que ha enriquit enormement tant el treball com la meva formació.

Finalment, vull agrair a la meua família i als meus amics el suport incondicional durant tot aquest procés.

REFERÈNCIES

- [1] Z. Moghaddam et al., "Smart charging strategy for electric vehicle charging stations," *IEEE Trans. Transp. Electrific.*, vol. 4, no. 1, pp. 76–88, 2018.
- [2] S. Y. Deb et al., "Energy regulation of EV using MILP for optimal operation of incentive based prosumer microgrid with uncertainty modelling," *Int. J. Electr. Power Energy Syst.*, vol. 132, 2021.
- [3] J. Whitaker, D. Redmond, G. Droge, and J. Gunther, "Scheduling Battery-Electric Bus Charging under Stochasticity using a Receding-Horizon Approach," *arXiv:2408.04087*, 2024.
- [4] M. Lv et al., "Reinforcement Learning for Electric Vehicle Charging Management: Theory and Applications," *Energies*, vol. 18, no. 19, p. 5225, 2025.
- [5] K. Christodoulou et al., "Optimal Scheduling of Electric Vehicle Charging with Deep Reinforcement Learning considering End Users Flexibility," *arXiv:2310.09040*, 2023.
- [6] A. Zare et al., "A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid," *Appl. Energy*, vol. 362, 2024.
- [7] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Appl. Energy*, vol. 254, p. 113732, 2019.
- [8] O. Frendo, N. Gaertner, and H. Stuckenschmidt, "Improving smart charging prioritization by predicting electric vehicle departure time," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6646–6653, 2021.
- [9] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.
- [10] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

APÈNDIX

A.1 Hiperparàmetres del model final

La taula 6 recull la configuració completa de l'agent Deep DQN V2.3 seleccionat, tal com s'instancia sobre la implementació DQN de Stable-Baselines3 [9]. Els valors corresponen a l'entrenament definitiu de 2,5 milions de *timesteps* amb llavor fixa per garantir la reproductibilitat.

TAULA 6: CONFIGURACIÓ COMPLETA DE L'AGENT DEEP DQN V2.3

Paràmetre	Valor
Política	MlpPolicy
Arquitectura de xarxa	[128, 128]
Taxa d'aprenentatge α	10^{-4}
Factor de descompte γ	0,99
Mida del <i>replay buffer</i>	100.000
<i>learning_starts</i>	20.000
<i>batch_size</i>	64
<i>train_freq</i>	4
τ (actualització dura)	1,0
<i>target_update_interval</i>	2.000
<i>exploration_fraction</i>	0,25
<i>exploration_final_eps</i>	0,03
<i>total_timesteps</i>	2.500.000
<i>Protocol d'avaluació durant l'entrenament</i>	
Freqüència d'avaluació	cada 50.000
Episodis per avaluació	100
<i>Checkpoint</i>	cada 250.000

A.2 Corbes d'entrenament

La Fig. 3 mostra l'evolució de la taxa d'èxit i del cost mitjà de l'agent V2.3 al llarg de l'entrenament, registrades cada 50.000 *timesteps* mitjançant un *callback* personalitzat. Les mètriques es calculen sobre escenaris mostrejats aleatòriament durant l'entrenament, incloent-hi configuracions no factibles, que en aquest mostreig limiten la taxa d'èxit màxima observable al voltant del 73,5%, a diferència de l'avaluació final, que es restringeix a escenaris factibles (Secció 6).

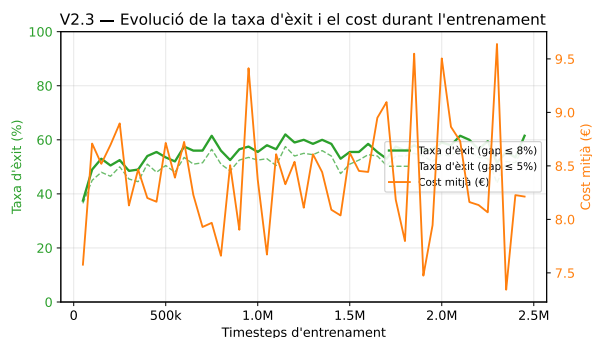


Fig. 3: Evolució de la taxa d'èxit ($\text{gap} \leq 5\%$ i $\leq 8\%$) i del cost mitjà de l'agent V2.3 durant els 2,5 M de *timesteps* d'entrenament.

La Fig. 4 compara la corba de recompensa de validació de les quatre configuracions entrenades al llarg del projecte. Cal interpretar-la amb cautela: el redisseny de la funció de recompensa (*reward shaping* i zones de tolerància) introduït a partir de V2.2 fa que els valors absoluts de recompensa no

siguin directament comparables entre versions amb esquemes de recompensa diferents; la comparació homogènia entre versions es realitza sobre les mètriques d'èxit, cost i gap (Apèndix A.3).

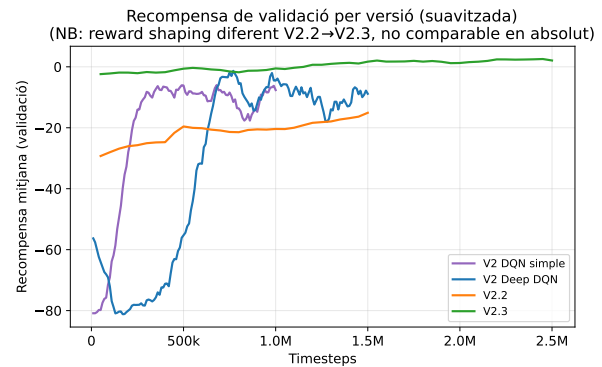


Fig. 4: Recompensa mitjana de validació (suavitzada) per a les quatre configuracions entrenades. Les recompenses de V2.2 i V2.3 no són comparables en valor absolut amb les de V2 a causa del canvi en l'esquema de recompensa.

La Taula 7 resumeix la millor recompensa de validació assolida per cada configuració i el *timestep* on s'assoleix. La columna *Esquema* indica el règim de recompensa de cada configuració: les recompenses només són comparables en valor absolut entre files del mateix esquema.

TAULA 7: RESUM DE LES CORBES DE VALIDACIÓ PER CONFIGURACIÓ

Configuració	Esquema	Millor reward	Timestep
DQN simple [64,64]	V2	7,38	420 k
Deep DQN [256,256]	V2	17,45	1,18 M
DQN V2.2 [128,128]	V2.2	-7,20	1,20 M
DQN V2.3 [128,128]	V2.3	7,50	1,55 M

A.3 Estudi d'ablació incremental

La taula 8 situa els agents de RL respecte a les dues estratègies de referència i a l'Oracle determinista, totes reavaluades sobre el mateix conjunt de 1.000 escenaris factibles i amb l'esquema de cost de V2.3. Cal notar que aquests valors no són directament comparables amb els de la Taula 3, que corresponen a un conjunt diferent de 100 episodis sobre l'entorn V2; aquí les baselines es recalculen sobre el conjunt de validació ampliat per situar-les en igualtat de condicions amb el model final. La progressió V2.2 $\rightarrow$ V2.3 redueix el cost mitjà de 6,69 € a 5,05 € (un estalvi del 24,58%) i situa l'agent a només un 0,22% per sobre del cost de l'Oracle, alhora que eleva la taxa d'èxit fins al 100%.

TAULA 8: ABLACIÓ INCREMENTAL: BASELINES, AGENTS RL I ORACLE, REVALUATS SOBRE 1.000 ESCENARIS FACTIBLES

Política	Èxit ($\leq 8\%$)	Cost mitjà
Càrrega Immediata	73,5%	9,19 €
Heurística de Preu	73,5%	6,34 €
DQN V2.2	96,9%	6,69 €
DQN V2.3	100%	5,05 €
Oracle (òptim)	100%	5,04 €

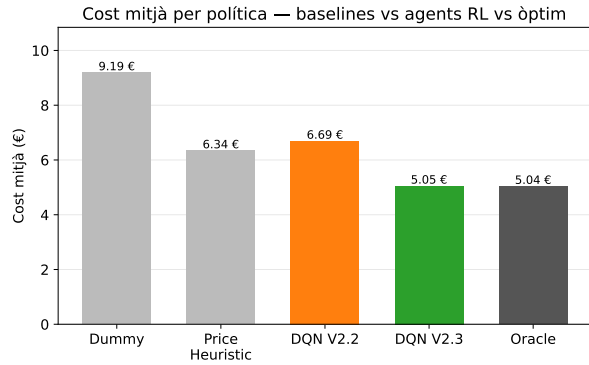


Fig. 5: Cost mitjà per sessió de cada política: de les estratègies de referència als agents de RL, amb l'Oracle com a cota inferior teòrica.

A.4 Casos d'estudi

Per il·lustrar el comportament de la política apresada en funció del marge de maniobra disponible (*slack*, definit com l'energia excedentària que es podria carregar més enllà de la necessària), se seleccionen tres escenaris representatius dels 1.000 de validació, descrits a la taula següent.

TAULA 9: ESCENARIS DELS TRES CASOS D'ESTUDI

	Cas 1 <i>slack alt</i>	Cas 2 <i>mitjà</i>	Cas 3 <i>ajustat</i>
Episodi	#350	#490	#442
SoC inicial	0,47	0,37	0,08
SoC objectiu	0,61	0,57	0,51
Bateria (kWh)	50	62	47
Carregador (kW)	3,7	2,3	2,3
Sortida (pas)	14	12	11
Slack (kWh)	42,0	13,3	3,6
Bretxa vs Oracle	+27,9%	0,0%	0,0%

La Fig. 6 mostra, per a cada cas, el perfil de preus de la sessió, les hores en què l'agent activa la càrrega i l'evolució del SoC fins al target. En el Cas 1, amb un marge molt ampli, l'agent concentra la càrrega en les hores més barates del tram final però perd algunes oportunitats òptimes que sí que aprofita l'Oracle, cosa que es tradueix en una bretxa de cost del +27,9%. En canvi, en els Casos 2 i 3, on el marge és moderat o ajustat, l'agent reproduïx exactament la seqüència de càrrega de l'Oracle (bretxa 0,0%): com menys *slack*, més s'aproxima la política apresada a l'òptim, ja que el marge per prendre decisions subòptimes desapareix.

Bloque 5 — Tres casos de estudio

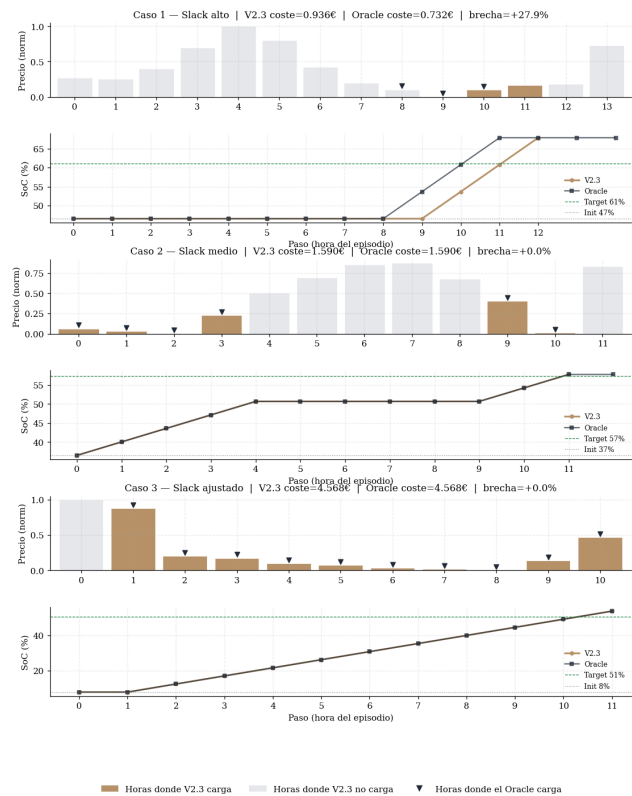


Fig. 6: Tres casos d'estudi (slack alt, mitjà i ajustat). Per a cada cas: corba de preu horari amb les hores de càrrega de l'agent ressaltades (superior) i evolució del SoC fins al target (inferior).

A.5 Validació del model de predicció (LightGBM)

Els dos models LightGBM que alimenten el *departure_step* de l'agent es validen amb *TimeSeriesSplit* sobre l'historial real de sessions. La taula següent mostra que ambdós models superen de forma consistent el baseline de perfil d'usuari, que representa el millor que es pot predir sense context del dia concret, tant en el tall de test com en validació creuada.

TAULA 10: MAE DELS MODELS DE PREDICCIÓ VS BASELINES

Predicció	Model	Baseline
<i>Tall de test</i>		
Energia (kWh)	8,16	8,82
Presència (h)	3,57	4,41
<i>Validació creuada (5 talls temporals)</i>		
Energia (kWh)	9,11	9,54
Presència (h)	3,46	4,22

La figura següent ho representa visualment per a les dues prediccions.

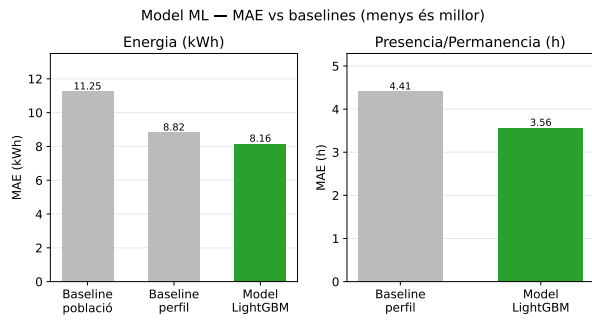


Fig. 7: Error absolut mitjà (MAE) dels models LightGBM d'energia i de presència, comparat amb els baselines (menys és millor).

Finalment, la taula següent il·lustra la capacitat d'adaptació en línia del model: partint d'una estimació poblacional (*cold-start*), la predicció d'energia convergeix cap al perfil real de l'usuari en molt poques sessions observades, reduint l'error de +8,3 kWh inicial a pràcticament zero a partir de la tercera sessió.

TAULA 11: CONVERGÈNCIA DE LA PREDICCIÓ D'ENERGIA PER NOMBRE DE SESSIONS OBSERVADES (ESCENARI D'ADAPTACIÓ)

Sessions	Predicció (kWh)	Error (kWh)
0 (<i>cold-start</i>)	22,34	+8,34
1	18,95	+4,95
3	14,18	+0,18
8	13,69	-0,31
50	11,29	-2,71