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

Ortega Luis, Martí; Carrabina Bordoll, Jordi, tut. Estudio y Predicción del Precio de la Electricidad en España. 2025. (Enginyeria de Dades)

This version is available at <https://ddd.uab.cat/record/317339>

under the terms of the  license

Estudi i Predicció del Preu de l'Electricitat a Espanya

Informe final

Martí Ortega Luis

Resum– El preu de l'electricitat a Espanya es determina diàriament i depèn de múltiples factors. L'oferta i la demanda tenen un paper principal, però també hi influeixen diverses condicions meteorològiques o, fins i tot, factors socioeconòmics. Aquest treball se centra en analitzar aquestes variables i construir models predictius capaços d'estimar el preu horari de l'electricitat. Per fer-ho, primer s'ha recopilat una àmplia base de dades històriques sobre tant els preus de la llum com les diferents variables que hi poden afectar. I posteriorment s'han fet servir tècniques d'aprenentatge automàtic com XGBoost, LightGBM i xarxes LSTM per realitzar les prediccions. Els resultats mostren un bon rendiment dels models en prediccions a curt termini, i una acceptable capacitat per identificar patrons i pics dins el mercat elèctric. A més, s'ha desenvolupat un dashboard interactiu per facilitar la consulta i visualització dels resultats.

Paraules clau ETL, Sèries temporals, Predicció Aprenentatge automàtic, Visualització, Interactivitat

Abstract– Electricity prices in Spain change daily depending on multiple factors, with supply and demand playing a key role. Other variables, such as meteorological conditions or even socioeconomic factors, can also influence prices. This project focuses on analyzing these variables and building predictive models to forecast the hourly electricity prices. To achieve this, the first step consisted in building an accurate and extended database containing, not only historical hourly electricity prices, but also all relevant variables. Consequently, machine learning models such as XGBoost, LightGBM, or LSTM networks have been used for electricity price forecasting. These models show consistent performance in short-term price forecasting, along with strong capability to identify patterns and peaks in the electricity market. Moreover, an interactive dashboard has been implemented to provide an easy way to check the dataset and visualize the results.

Index Terms ETL, Time series, Forecasting, Machine learning, Data visualization, Interactivity

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

L'ENERGIA elèctrica és un recurs essencial en qualsevol societat moderna, ja que impulsa tant l'activitat econòmica com la vida quotidiana de la població. Per aquest motiu, comprendre com funciona el preu de l'electricitat és molt important tant per a grans multinacionals com per a consumidors domèstics habituals.

Actualment, a Espanya, com en la majoria de països d'Europa, el preu de l'electricitat es determina diàriament mitjançant un mecanisme d'oferta i demanda.

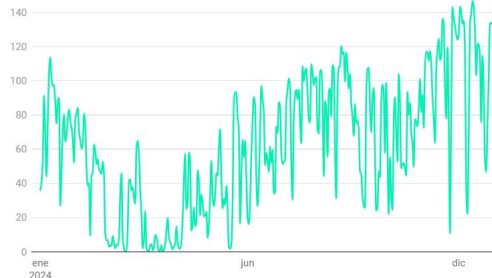


Fig. 1: Preu de l'electricitat (en €/MWh) durant l'any 2024 a Espanya (ESIOS-Red Eléctrica, 2024)

Com es pot apreciar a la figura 1, aquest preu fluctua constantment, això és a causa d'una gran varietat de factors. D'una banda, principalment hi actuen elements com la producció i el consum. D'altra banda, influeixen també aspectes meteorològics, com la temperatura, les precipita-

• E-mail de contacte: Marti.Ortega@autonoma.cat

• Treball tutoritzat per: Jordi Carrabina Bordoll i Irene Gutiérrez del Burgo

Curs 2024/25

cions, les hores de sol o la velocitat del vent, que afecten tant la demanda d'energia com la producció d'electricitat de fonts renovables. A més, factors socioeconòmics com el preu del gas natural, la regulació dels drets d'emissió de diòxid de carboni establerta per la Unió Europea o la inflació poden tenir un paper destacable en el mercat elèctric. També influeix si el dia és laborable o no, ja que els hàbits de consum i els consums industrials solen variar segons el tipus de jornada.

Poder ser capaços de predir el preu de l'electricitat és vital per al sector energètic. La capacitat d'aconseguir fer bons pronòstics pot ser molt útil i facilitar preses de decisions tant per grans empreses com per als consumidors.

2 OBJECTIUS

L'objectiu principal del treball és desenvolupar un model capaç de predir correctament el preu de la llum de forma horària, és a dir, per a cada una de les 24 hores del dia. Per poder fer prediccions fiables, el primer pas és construir una base de dades sòlida, coherent, fiable i de qualitat, a partir de diverses fonts de dades. Per tant, cal integrar dades històriques no només del mercat elèctric, sinó també de variables meteorològiques i factors socioeconòmics que poden influir en el seu comportament.

Pel que fa al desenvolupament dels models de predicció del preu de l'electricitat, es planteja estimar tant el preu de l'hora següent com el preu absolut o relatiu de les pròximes 24 hores.

A més, un altre dels objectius és dissenyar un dashboard interactiu que permeti visualitzar i explorar les dades i prediccions de manera intuïtiva i accessible.

3 METODOLOGIA I DESENVOLUPAMENTS

RECOPILACIÓ DE DADES

Per dur a terme les fases de recopilació de dades, s'ha seguit un procés per garantir que les dades obtingudes són fiables, és a dir, que provenen de fonts oficials o altament reconegudes, de qualitat i totes mantenen el mateix format, per així facilitar la posterior integració entre elles. La metodologia aplicada en aquesta fase inclou els següents passos:

1. Identificació de fonts de dades:

S'identifiquen i es seleccionen fonts de dades fiables i actualitzades per a cada tipus de dades per assegurar que totes aquestes són precises, consistents, i reflecteixen la realitat en tot moment. Això inclou dades històriques sobre el preu de l'electricitat, la producció i la demanda de l'energia, incloent també dades meteorològiques i socioeconòmiques.

2. Obtenció de dades històriques sobre l'energia:

Preu de l'electricitat: Es recopilen dades sobre el preu de l'electricitat a través de la plataforma pública i oficial del mercat majorista d'electricitat a Espanya (OMIE) [10], que ofereix informació pública i detallada de l'evolució dels preus horaris des de l'1 de gener del 2018 fins a la data d'avui. Aquesta plataforma proporciona fitxers diaris amb els preus marginals horaris

de l'electricitat. Cada línia del fitxer recull el preu per hora d'un dia concret, amb valors en €/MWh. Amb la finalitat d'automatitzar la seva recollida, es desenvolupa un script en Python per poder descarregar tots aquests fitxers diaris de forma automàtica fent servir diverses llibreries com `requests` [9] per a la gestió de peticions HTTP, i `BeautifulSoup` [8] per al processament del contingut HTML i la manipulació d'enllaços.

Producció, demanda, importacions i exportacions:

Aquestes dades també es recopilen des de la plataforma del mercat majorista d'electricitat a Espanya (OMIE) [11], que també ofereix dades detallades sobre l'energia total desglossada en el mercat diari, també des de principis del 2018 fins a l'actualitat. De nou, les dades són proporcionades mitjançant fitxers diaris, aquests contenen valors horaris de producció, demanda, importacions i exportacions d'electricitat, amb el desglossament per tipus de tecnologia (com nuclear, hidroelèctrica, cicle combinat, etc.) i per país d'origen o destí quan es tracta d'importacions o exportacions. Donat que el format dels fitxers proporcionats per OMIE és pràcticament igual al dels preus horaris, es pot reutilitzar el mateix script descrit anteriorment per automatitzar la descàrrega d'aquestes dades. Això ha permès agilitzar el procés i evitar la descàrrega manual de milers d'arxius, estalviant així una gran quantitat de temps i esforç.

3. Recopilació de dades meteorològiques:

Les dades meteorològiques s'obtenen del conjunt de dades ERA5, desenvolupat pel *European Centre for Medium-Range Weather Forecasts* (ECMWF) [12] en el marc del programa Copernicus. ERA5 proporciona estimacions horàries de nombroses variables atmosfèriques, de superfície terrestre i d'onatge oceànic des de l'any 1940 fins a l'actualitat, cobrint així més de vuit dècades d'informació climàtica consistent i coherent a escala global. Per descarregar les dades d'interès de forma automatitzada, es programa un script en Python utilitzant l'API de Copernicus i la llibreria `cdsapi`. Aquest permet seleccionar les dates corresponents, és a dir, des del 2018 fins al 2024, delimitar l'àrea geogràfica d'interès, en aquest cas, la península Ibèrica, i seleccionar les variables meteorològiques amb les quals es treballarà. Aquestes variables han estat les següents:

Temperatura a 2 metres: Representa la temperatura de l'aire a una alçada de 2 metres sobre el terra, en graus Kelvin (°K).

La temperatura pot afectar notablement la demanda, ja que temperatures extremes, tant fredes com càlides, incrementen l'ús de sistemes com la calefacció o l'aire acondicionat.

Cobertura total de núvols: Indica el percentatge de la superfície terrestre coberta per núvols en un moment determinat.

Aquesta variable afecta a la producció d'energia a la producció d'energia renovable, ja que, per exemple, si el cel està ennuvolat, disminuirà la generació d'energia solar.

Radiació solar total global: Aquesta variable mesura

la radiació solar total rebuda a la superfície de la Terra, en Joules per metre quadrat (J/m^2).

Precipitació total: Representa la quantitat total de precipitació (pluja, neu, etc.) caiguda en un període de temps determinat, mesurada en metres.

Velocitat del vent en la direcció est-oest: Mesura la velocitat del vent que es mou de l'oest a l'est o viceversa, en metres per segon.

Velocitat del vent en la direcció nord-sud: Mesura la velocitat del vent que es mou del nord al sud o viceversa, també en metres per segon.

4. Recopilació de dades socioeconòmiques:

Preu del gas natural: Les dades s'obtenen de MIB-GAS (Mercado Ibérico del Gas) [16], el mercat de referència per a la compra i venda de gas natural a Espanya i Portugal. Les dades es poden descarregar en fitxers anuals en format CSV o XLSX, que inclouen informació diària sobre el preu del gas natural, en €/MWh.

Preu del petroli Brent per barril: El petroli Brent és un tipus de petroli que es produeix principalment al Mar del Nord i és el tipus de petroli cru més utilitzat com a referència a Europa. Les dades s'obtenen de la plataforma *Investing* [15], una plataforma que proporciona informació financera sobre les matèries primeres, on es poden descarregar dades diàries del preu del petroli Brent en format CSV, mesurades en dòlars per barril (on un barril equival a 159 litres de petroli).

Conversió USD a EUR: Les dades històriques de la conversió de dòlars (USD) a euros (EUR) es poden consultar, novament, a través de la plataforma *Investing* [14]. Aquestes dades inclouen els valors diaris de la taxa de canvi entre el dòlar i l'euro.

Preu dels drets d'emissió de CO₂: Es recullen dades del portal *Sendeco2* [13], que proporciona informació detallada sobre els preus dels drets d'emissió dins del Sistema Europeu de Comerç d'Emissions (EU ETS). Les dades inclouen informació diària sobre el preu dels drets d'emissió en euros per tona de CO₂.

PIB d'Espanya: Es recullen dades sobre el producte interior brut de l'economia espanyola a través de l'Institut Nacional d'Estadística (INE) [17]. Les dades es poden descarregar en fitxers CSV o XLSX que ofereixen informació sobre el PIB de cada un dels trimestres dels anys seleccionats.

Taxa d'atur a Espanya: S'obtenen dades sobre la taxa d'atur espanyola, altre cop a través de l'Institut Nacional d'Estadística (INE)[18], on les dades es poden descarregar en fitxers CSV o XLSX que ofereixen informació sobre la taxa d'atur per a cada un dels quatre trimestres dels anys seleccionats.

5. Integració de les dades:

Un cop recollides totes les dades, s'estableix un format comú per poder integrar-les fàcilment en un CSV final. Com que la finalitat del treball és poder ser capaç de predir el preu de la llum de forma horària, les dades s'han estructurat amb camps de data i hora, de manera que totes les fonts segueixin la mateixa granularitat

temporal. Per a les variables amb valors diaris, com el preu del gas natural, o trimestrals, com el PIB espanyol, s'han generat 24 registres per dia amb la mateixa dada repetida per cada hora, assegurant així la coherència amb altres fonts estructurades per hores.

NETEJA I TRACTAMENT DE DADES

Un cop recollides totes les dades, es procedeix a la fase de processament i neteja, amb l'objectiu de garantir que les dades quedin correctament tractades, consistents i unificades, de manera que estiguin preparades per ser utilitzades en l'entrenament dels diferents models. La metodologia aplicada en aquesta fase inclou els següents passos:

1. Tractament de valors inexistents:

És força comú que certs datasets estiguin incomplets i presentin diversos valors inexistents, fet que pot introduir errors en l'entrenament dels models predictius, i si no es tracten adequadament, aquests valors poden afectar negativament el rendiment i la capacitat de generalització dels models, fent que les prediccions siguin menys fiables. En la nostra base de dades s'identifiquen certes variables amb valors inexistents, com les exportacions, les importacions, el preu tant del gas natural com del petroli Brent, la conversió de USD a EUR o els drets d'emissió de CO₂. Per resoldre aquest problema s'han abordat principalment dues tècniques diferents:

- **Substituir per 0:** En les variables tant d'importacions com d'exportacions, s'observa que la falta d'alguns valors no era provocada per algun error, sino perquè realment haurien de ser 0. Observant les dades, es nota que normalment les exportacions coincideixen bastant amb la diferència entre la producció i la demanda (quan aquesta diferència era positiva), i les importacions passava just al contrari. Per tant, en tots aquells casos on faltaven els valors d'importacions o exportacions, es substitueix per un 0.
- **Interpolació temporal:** És una tècnica que consisteix a estimar els valors que falten a partir dels valors coneguts abans i després del punt amb dades inexistents. Per exemple, si hi ha un valor nul entre dues observacions consecutives, es pot calcular el valor mitjançant una interpolació lineal, assumint una transició suau entre les dades conegudes.

2. **Creació de noves columnes:** Per enriquir el conjunt de dades, s'incorporen noves columnes que puguin aportar informació addicional per a l'entrenament dels models. Per aquest motiu, s'afegeixen variables com si el dia és festiu o no, si es tracta d'un cap de setmana, així com el dia de la setmana. A més, s'incloïen noves *features* relacionades amb el tipus de producció d'energia, com el percentatge d'energia provinent de fonts renovables i el percentatge d'energia tèrmica, amb l'objectiu de capturar millor les dinàmiques que poden influir en el preu de l'electricitat.

ENTRENAMENT DELS MODELS

Un cop completat el tractament i la preparació de les dades, es procedeix a l'entrenament dels diferents models. Els models d'aprenentatge automàtic utilitzats han estat:

- **XGBoostRegressor[19]:** És un model basat en la tècnica de *gradient boosting* que construeix una seqüència d'arbres de decisió on cada arbre intenta corregir els errors dels anteriors. Incorpora tècniques de regularització per evitar el sobreajustament i està altament optimitzat per al rendiment computacional. És especialment útil en problemes de regressió i classificació amb grans volums de dades.
- **LightGBM Regressor (LGBMR)[20]:** També és un model basat en la tècnica de *gradient boosting* que construeix una seqüència d'arbres de decisió on cada arbre intenta corregir els errors dels anteriors. Però LightGBM utilitza una estructura d'arbres basada en fulles (*leaf-wise*) en comptes de nivells (*level-wise*), cosa que el fa més ràpid i eficient en l'ús de memòria, especialment amb grans volums de dades.

El primer pas abans de començar a entrenar els models és realitzar una primera separació de les dades en dues parts:

- X, que conté les característiques d'entrada al model .
- Y, que representa la sortida o variable objectiu del model.

Com que en el nostre cas, l'objectiu és predir el preu de l'hora posterior, aquesta variable objectiu correspon al preu de l'hora següent.

Com en tot procés d'aprenentatge automàtic, el següent pas consisteix en dividir les dades en conjunts d'entrenament i validació. Però, no s'utilitza la tècnica tradicional de divisió aleatòria (80% per a l'entrenament i 20% per al test), ja que, al tractar-se d'una sèrie temporal, això permetria al model "veure"el futur, cosa que invalidaria els resultats. Per aquest motiu, es selecciona el primer 80% del conjunt de dades (des de gener de 2018 fins a principis d'agost de 2023) per a l'entrenament del model, i l'últim 20% (des d'agost de 2023 fins a desembre de 2024) per al test.

Pel que fa al model XGBoost, els resultats inicials respecte al conjunt de test van estar els següents:

MAE	RMSE	R²
7.26	10.36	0.940

TAULA 1: MÈTRIQUES MODEL INICIAL

Però, al tractar-se d'una sèrie temporal, es va contemplar la idea d'afegir lag features (variables temporals retardades, com els preus de la llum d'hores anteriors. S'han provat diferents valors de lags per avaluar el seu impacte en la precisió del model. A la taula següent es poden consultar els resultats obtinguts al fer aquesta anàlisi:

Pel que fa al model basat en *LightGBM Regressor*, s'observa el mateix comportament: les versions entrenades amb *lag features* aconsegueixen tant prediccions més precises com millores en les mètriques d'avaluació. A continuació

Retard	MAE	RMSE	R²
Sense retard	7.26	10.36	0.940
1 hora	6.59	9.76	0.947
Fins a 4 hores	6.49	9.59	0.949
Fins a 12 hores	6.80	9.95	0.946
Fins a 24 hores	6.70	9.73	0.948

TAULA 2: COMPARATIVA DE MÈTRIQUES AL UTILITZAR DIFERENTS VALORS DE lag features (XGB)

Retard	MAE	RMSE	R²
Sense retard	7.52	10.32	0.941
1 hora	6.02	9.21	0.953
Fins a 4 hores	6.02	9.19	9.53
Fins a 12 hores	6.02	9.14	0.954
Fins a 24 hores	5.65	8.78	0.957

TAULA 3: COMPARATIVA DE MÈTRIQUES AL UTILITZAR DIFERENTS VALORS DE lag features (LGBM)

es mostra la taula amb les diferents mètriques obtingudes al aplicar o no aquestes característiques de retard temporal:

Per poder comparar el rendiment d'aquests dos models d'aprenentatge automàtic, es mostra la següent gràfica, on es poden observar les diferències entre les prediccions d'ambdós models i els valors reals de la primera setmana de l'any 2024:

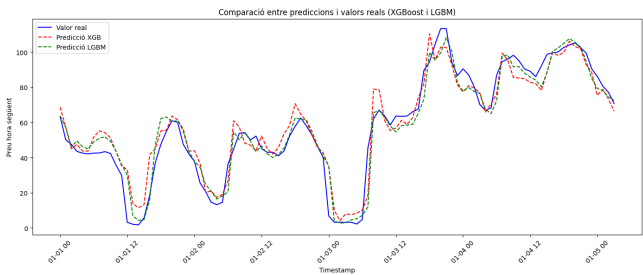


Fig. 2: Comparació de les prediccions dels models XGBoost i LightGBM .

Com es pot observar a la Figura 2 (on l'eix Y representa el valor del preu de la llum, i l'eix X representa el pas del temps), tant el model XGBoost (línia discontinua vermella) com el LightGBM (línia discontinua verda) presenten unes prediccions similars, a més de ser força precises en comparació amb els valors reals (línia blava). Per tant, es podria dir que aquests dos tipus de models tenen un comportament similar i obtenen resultats gairebé igual de bons.

A part d'aquests dos models basats en aprenentatge automàtic, també s'ha començat a treballar amb altres models més complexos basats en xarxes neuronals com:

- **Long Short-Term Memory (LSTM):** És un tipus de xarxa neuronal recurrent especialitzada en dades seqüencials (com sèries temporals). La seva clau està en la capacitat de recordar informació important a llarg termini i descuidar el que no és rellevant i són ideals

per capturar dependències a llarg termini i patrons temporals complexos.

El procediment per entrenar aquest model és molt similar al dels anomenats anteriorment, les dades es divideixen en dos conjunts diferents, el d'entrenament i el de validació, tot respectant la temporalitat de les dades. Per tant, s'utilitza de nou el primer 80% del conjunt de dades (des de gener de 2018 fins a principis d'agost de 2023) per a l'entrenament del model, i l'últim 20% (des d'agost de 2023 fins a desembre de 2024) per al test.

Però al estar treballant amb xarxes neuronals, s'ha de tenir en compte que aquests models més avançats són molt sensibles a l'escala, per tant és crucial primer normalitzar les dades, és a dir, transformar-les perquè totes les variables estiguin en un rang similar. Això es pot dur a terme ja sigui mitjançant estandardització (restant la mitjana i dividint per la desviació estàndard) o escalat min-max (adaptant al rang [0,1]), entre d'altres mètodes... En el nostre cas, s'ha fet servir l'escalat min-max mitjançant la classe `MinMaxScaler` de `sklearn` [22], per tal d'evitar que variables amb valors numèricament més grans dominin el procés d'entrenament.

A més, per tal d'incorporar la informació temporal de manera eficient, s'ha optat per l'ús de finestres lliscants (*sliding windows*). Això vol dir que s'agafen segments de dades consecutius de la mateixa mida, que es van desplaçant pas a pas al llarg del temps. Cada segment o finestra, conté valors anteriors de les variables que el model utilitza per predir el valor següent del Preu de la llum.

Com que la mida de la finestra pot influir en la capacitat del model per captar patrons temporals, s'han provat diferents valors. A la taula següent es poden observar els diferents resultats obtinguts en el conjunt de test.

Window Size	MAE	RMSE	R ²
2	9.1429	12.4366	0.9090
4	7.3834	11.0992	0.9275
8	9.9443	12.6871	0.9053
12	8.8177	12.4561	0.9087
24	8.7135	12.2370	0.9119

TAULA 4: COMPARATIVA DE MÈTRQUES DEPENDENT DEL TAMANY DE LA FINESTRA

Dels resultats es pot observar que la millor finestra és la de mida 4, ja que obté els valors més baixos tant de l'MAE com del RMSE, i el valor més alt de R². Això indica que amb una finestra de només 4 instants temporals el model és capaç de captar prou bé la dinàmica de les dades per fer bones prediccions. En canvi, amb finestres més grans no s'aconsegueixen millores significatives, i fins i tot en alguns casos el rendiment empitjora lleugerament, possiblement per excés d'informació o per soroll acumulat. A la següent gràfica podem observar algunes prediccions de la primera setmana de 2024 d'aquest nou model.

On podem veure com, sense la necessitat d'afegir *lag features*, el model presenta molt bones prediccions. Aquí es pot veure la importància de la *LSTM* i la seva capacitat d'aprendre patrons temporals.

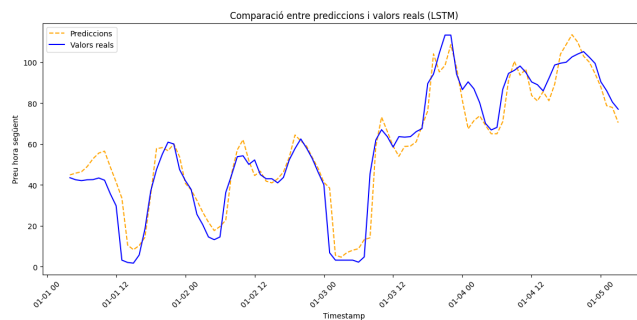


Fig. 3: Comparació de les prediccions de la LSTM (línia taronja discontinua) i els valors reals (línia blava).

Cross Validation TEMPORAL

Per dur a terme una validació més exhaustiva de tots 3 models, s'ha dut a terme un procés de **cross-validation temporal (Time Series Cross-Validation)**, que serveix per avaluar el rendiment d'un model sobre dades temporals, dividint la sèrie en diversos subconjunts (*o folds*) que mantenen l'ordre temporal. Com es pot veure a la següent figura, el procediment és el següent:

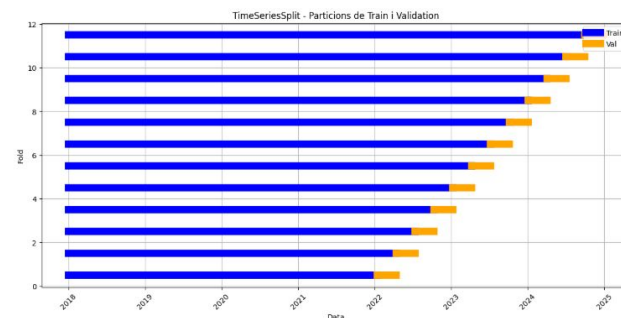


Fig. 4: Visió gràfica dels *Folds* utilitzats per dur a terme el *cross-validation* temporal.

Es divideixen les dades en diversos blocs, que corresponen a cada una de les línies de la gràfica. En cada iteració, s'entrena el model amb un període inicial de dades (la part blava) i es valida amb el període següent (la part taronja). Aquesta finestra d'entrenament s'ampliarà progressivament. D'aquesta manera, es pot observar com el model generalitza en diferents moments del temps, i no només en un únic període fix.

A diferència de l'enfocament anterior (on s'utilitzava sempre el primer 80% de les dades per entrenar i l'últim 20% per validar) aquest mètode permet una avaluació més completa i representativa, ja que la validació es fa en diversos punts temporals i no només al final de la sèrie.

A la següent taula es mostren les mitjanes i desviacions estàndard ($\pm$) de les mètriques obtingudes per cada model mitjançant *cross-validation* temporal.

Com es pot veure a la taula, tots tres models presenten un rendiment similar, amb lleugeres diferències segons la mètrica analitzada. I, tot i que els resultats mitjans són bons, les desviacions estàndard indiquen que el rendiment dels models pot variar depenent del període temporal utilitzat per validar.

Model	MAE	RMSE	R ²
XGBoost	8.72 ± 3.01	13.502 ± 7.04	0.899 ± 0.051
LightGBMR	8.039 ± 2.91	12.56 ± 6.62	0.9125 ± 0.044
LSTM	8.214 ± 2.97	11.739 ± 3.92	0.916 ± 0.049

TAULA 5: COMPARATIVA DE MÈTRIQUES ENTRE MO-DELS

PREDICCIONS A 24 HORES

Ara, l'objectiu principal era predir no el preu de l'hora següent, sinó el de les pròximes 24 hores. Així que, inicialment, es van modificar els tres models anteriors per ara dur a terme la predicció del preu absolut de les pròximes 24 hores, es van valorar diferents valors de retard temporal pels models XGBoost i LightGBM, i mides de finestra per la xarxa LSTM, però no es van obtenir resultats força prometedors, tal i com es pot observar a la següent taula, que mostra les mètriques mitjanes de les prediccions per dia:

Retard	MAE mitjà	RMSE mitjà	R ² mediana
24	16.477 ± 9.630	19.905 ± 10.577	0.132
48	17.164 ± 9.589	20.99 ± 10.569	0.142
72	16.684 ± 9.649	20.709 ± 10.576	0.1638
96	16.789 ± 9.724	20.509 ± 10.698	0.205

TAULA 6: RESULTATS MODEL XGBOOST 24 HORES

Retard	MAE mitjà	RMSE mitjà	R ² mediana
24	16.315 ± 10.386	20.310 ± 11.349	0.172
48	16.458 ± 10.553	19.966 ± 11.625	0.197
72	16.240 ± 10.497	19.863 ± 11.534	0.181
96	16.614 ± 10.409	20.383 ± 11.488	0.177

TAULA 7: RESULTATS MODEL LIGHTGBM 24 HORES

Window Size	MAE mitjà	RMSE mitjà	R ² mediana
24	14.56 ± 9.33	17.37 ± 10.65	0.253
48	14.69 ± 9.86	17.72 ± 11.31	0.264
72	15.28 ± 9.12	18.33 ± 10.39	0.3
96	15.39 ± 9.73	18.58 ± 11.21	0.228

TAULA 8: RESULTATS LSTM 24 HORES

Cal destacar que, al estar calculant les mètriques per dia, s'ha fet servir la mediana del coeficient de determinació (R²) en comptes de la mitjana perquè la mediana és una mesura més robusta davant de valors extrems. El R² pot prendre valors molt negatius en casos en què el model no encerta bé les prediccions per un dia concret, cosa que pot distorsionar la mitjana. En canvi, la mediana proporciona una visió més representativa del comportament típic del model al llarg dels diferents dies, evitant que un petit nombre de dies amb prediccions molt dolentes influeixi desproporcionadament en el resultat global.

Observant els resultats, s'aprecia que els models no prediuen amb gaire precisió el preu absolut de la llum, ja que presenta una alta volatilitat i la impossibilitat de conèixer els valors futurs de variables clau com el gas o la demanda real. Tot i això, entre els tres models comparats, la xarxa LSTM ha estat la que ha ofert els resultats més consistents, amb menors errors (MAE i RMSE) i un coeficient de determinació (R²) més elevat en la majoria de casos. Aquesta millora es deu a la capacitat de les LSTM per captar millor els patrons temporals. Per aquest motiu, s'ha decidit continuar treballant amb aquest model.

Per tant, finalment s'ha optat per calcular-ne el preu relatiu, és a dir, a quina hora del dia de demà el preu serà més car o més barat, independentment del preu en €/MWh exacte. Per obtenir aquest preu relatiu, tant les prediccions com els valors reals s'han normalitzat per tal que es trobin dins d'un rang comú de 0 a 1, cosa que permet comparar les tendències sense dependre de l'escala original. Un clar exemple d'aquest comportament el podem observar a les següents gràfiques, que corresponen al dia 10 de novembre del 2024 i es pot apreciar la diferència entre el comportament respecte al preu absolut i al preu relatiu.

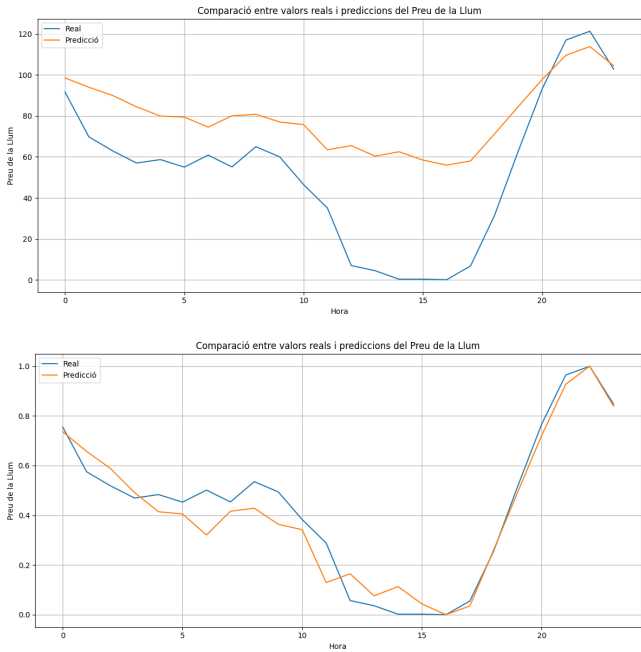


Fig. 5: Comparació entre la predicció i els valors reals dels preus absoluts (a dalt) i relatius (a baix) del dia

A les dues gràfiques, la línia blava representa els valors reals, mentre que la taronja els predits, i podem apreciar com la primera gràfica mostra la predicció del model per preus absoluts, mentre que a la segona, aquests estan normalitzats, de manera que el valor més alt de cada dia pren el valor 1 i el més baix el valor 0. Aquesta normalització permet analitzar el patró diari de variació del preu, centrant-se en la forma de la corba i no en els valors exactes. Així, encara que el model no encerti amb precisió els preus en €/MWh, pot ser útil per identificar a quines hores el preu serà relativament més alt o més baix.

Per poder avaluar quantitativament si el model encerta bé els preus màxims i mínims, s’ha comparat l’hora del dia del preu més alt i més baix entre la predicció i la realitat. S’ha considerat que una predicció és correcta si la posició del màxim o del mínim no es troba molt lluny de la real, amb una tolerància de 2 o 3 hores i s’han obtingut les següents resultats, on, per obtenir l’*accuracy*, s’ha dividit el nombre d’encerts entre el total de dies del conjunt de test.

Marge	Accuracy del Preu Mínim	Accuracy del Preu Màxim
±2 hores	0.67	0.73
±3 hores	0.73	0.73

TAULA 9: ACCURACY DE LA PREDICCIÓ DELS PICS DE PREU AMB MARGE HORARI

Així que, podríem dir que el model és capaç d’identificar amb una precisió raonable les hores dels pics diaris del preu de l’electricitat.

DESENVOLUPAMENT DEL DASHBOARD INTERACTIU

Per tal de millorar l’experiència de l’usuari i facilitar la visualització i interpretació dels diferents resultats, s’ha desenvolupat un dashboard interactiu utilitzant l’eina Streamlit[23]. Aquesta interfície intuïtiva permet a l’usuari accedir a diverses funcionalitats, com ara:

- Explorar i veure característiques sobre les dades utilitzades.
- Visualitzar gràficament correlacions i tendències de les dades.
- Revisar els resultats de les prediccions històriques realitzades.
- Consultar els valors estimats de les prediccions a futur.

A continuació es dona una explicació més detallada de cada una de les seccions que proporciona la interfície:

1. **Consultar les dades:** En aquesta primera secció, s’hi poden diferenciar dos apartats principals, al primer, nomenat exploració de les dades utilitzades, l’usuari pot filtrar les dades per les dates que li interessin. I un cop seleccionat el rang de dates desitjat, es mostra una taula amb totes les dades corresponents dins d’aquest interval. On l’usuari hi pot interactuar, consultar totes les variables, ordenar-les i fins i tot es dona l’opció a descarregar aquestes dades.

El segon apartat correspon a la part on es poden veure i consultar les estadístiques de totes les variables contínues com ara el preu de la llum, la demanda, el preu del gas natural, etc.. Es mostren valors com el mínim i el màxim, que permeten conèixer els límits dels valors observats dins del rang seleccionat anteriorment, la mitjana i la desviació estàndard, que donen

informació sobre la tendència i la variabilitat, i la mediana, que és útil per entendre el valor central de la distribució.

2. **Gràfiques sobre les dades:** Aquesta segona secció està molt relacionada amb l’anterior, ja que també té com a objectiu proporcionar informació rellevant sobre les dades a l’usuari. Però aquesta secció es centra principalment en la dada més important, el preu de la llum. Primer es permet a l’usuari agrupar les dades per diferents escales temporals, com per hora del dia, dia de la setmana, dia del mes, setmana de l’any, mes o any. I visualitzar les tendències mitjançant gràfiques. A més a més, també s’ofereix a l’usuari l’opció d’incloure dades complementàries sobre la demanda i la producció elèctrica, per tal d’analitzar com es relacionen amb el preu de la llum.

Com es pot observar a la següent gràfica, que mostra la tendència del preu de la llum agrupat per dia de la setmana, i la comparació amb la tendència també de la demanda i la producció. Es pot apreciar que hi ha una forta caiguda durant el cap de setmana.

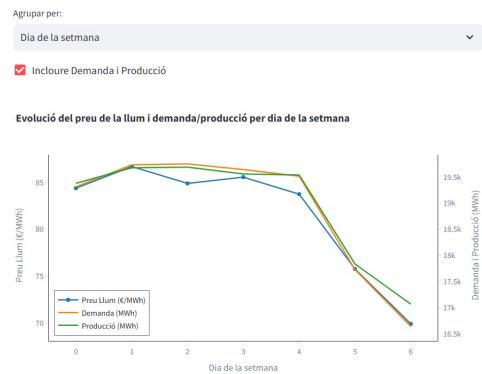


Fig. 6: Comparació de les tendències setmanals del preu (línia blava), la demanda (línia taronja) i la producció (línia verda)

En aquesta secció també es presenta una taula amb algunes correlacions interessants entre el preu de la llum i la resta de les variables. I es pot observar com factors socioeconòmics com el Preu del gas natural, el preu del petroli i els drets d’emissió de CO₂ estan directament relacionats amb el preu de la llum. Mentre que altres variables com el percentatge d’energia renovable produïda ho estan indirectament.

3. **Prediccions i resultats anteriors :** En aquesta tercera secció, l’usuari pot consultar els resultats de les prediccions històriques generades pels diferents models. Un cop seleccionada una data concreta, es mostra una gràfica amb les prediccions i els valors reals del preu de la llum corresponents als set dies anteriors. L’usuari pot escollir quin model vol visualitzar a la gràfica interactiva, i també es mostren les mètriques d’avaluació d’aquell model específic per a la setmana seleccionada.

4. **Prediccions a futur:** En aquesta última secció, l'usuari pot consultar les prediccions a futur, tant el preu absolut com quins seran els preus més alt i més baix de les properes 24 hores. Per fer-ho, primer ha d'indicar el dia i l'hora en què es troba on el que es desitja fer la predicció, i a continuació es mostra un missatge amb el preu actual de la llum, com a punt de referència. Després, l'usuari pot seleccionar l'horitzó de predicció. Si selecciona 1 hora, es mostren les prediccions dels tres models per al preu de la llum de l'hora següent. A més, l'usuari té l'opció de visualitzar els valors reals per comparar-los amb les prediccions. Si es selecciona l'horitzó de 24 hores, es mostren les prediccions de les hores més barates i més cares del proper dia.

A continuació, es proporciona un enllaç a un vídeo de demostració del funcionament del dashboard:

Clica per accedir al vídeo demostratiu del dashboard interactiu

4 CONCLUSIÓ

Per concloure el treball, es podria dir que s'ha dut a terme una anàlisi exhaustiva per a la predicció del preu de l'electricitat a Espanya. S'ha creat una base de dades sòlida que combina tant dades històriques sobre el mercat elèctric espanyol, com dades meteorològiques i socioeconòmiques. Totes netes i processades.

Després, pel que fa a les prediccions, s'han utilitzat tècniques d'aprenentatge automàtic com XGBoost, LightGBM i xarxes LSTM, demostrant un bon rendiment quan es tracta de predir el preu de l'hora següent. Però, tot i que les prediccions a 24 hores no presenten un rendiment tan òptim, s'ha demostrat que el model pot identificar amb encert les hores de preu màxim i mínim del dia següent.

A més, s'ha desenvolupat amb èxit un dashboard interactiu que permet visualitzar i consultar fàcilment dades i prediccions.

En definitiva, aquest treball ha servit per entendre el comportament del mercat elèctric i demostrar que, amb les dades adequades i els models correctes, és possible predir l'evolució del preu. Malgrat les limitacions en horitzons de 24 hores, els resultats obtinguts confirmen que l'enginyeria de dades és una eina molt útil en tot tipus d'àmbits, incloent-hi el mercat energètic.

AGRAÏMENTS

En primer lloc, vull agrair a la meua família pel seu suport incondicional i per estar sempre al meu costat.

També vull agrair especialment als meus dos tutors, Jordi Carrabina i Irene Gutiérrez, ja que, sense la seva ajuda i el seu ampli coneixement en el tema, la realització d'aquest treball no hagués estat possible.

REFERÈNCIES

- [1] Newtral. (2024). *Precio de la luz en 2024*. Disponible a: <https://www.newtral.es/precio-luz-2024/20241230/>.
- [2] Del Río Huertas, J. (2024). *Diseño y desarrollo de una aplicación web para la estimación de precios de la electricidad en España*.
- [3] Endesa. (2022). *Why is electricity expensive in Spain?* Disponible a: <https://www.endesa.com/en/blogs/endesa-s-blog/light/why-electricity-expensive-in-spain>.
- [4] Grigoryan, H. (2021). *Electricity Consumption Prediction using Energy Data, Socio-economic and Weather Indicators. A Case Study of Spain*.
- [5] Naturgy. *¿Cómo funciona el sistema eléctrico español?* Disponible a: <https://www.naturgy.com/como-funciona-el-sistema-electrico-espanol>.
- [6] REE. *El suministro de la electricidad. Un equilibrio entre generación y consumo*.
- [7] Endesa. (2021). *¿De qué depende el precio de la luz?* Disponible a: <https://www.endesa.com/es/blog/blog-de-endesa/luz/de-que-depende-el-precio-de-la-luz>.
- [8] Beautiful Soup Documentation. Recuperat de <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- [9] Requests: HTTP for Humans. Recuperat de <https://requests.readthedocs.io/en/latest/>
- [10] OMIE - Operador del Mercado Ibérico de Energía. *Precios horarios del mercado diario en España*. Recuperat de <https://www.omie.es/es/file-access-list?parents=/Mercado%20Diario/1.%20Precios&dir=Precios%20horarios%20del%20mercado%20diario%20en%20Espa%C3%81a&readdir=marginalpdbc>
- [11] OMIE - Operador del Mercado Ibérico de Energía. *Energía total desglosada en el mercado diario*. Recuperat de https://www.omie.es/es/file-access-list?parents=/Mercado%20Diario/2.%20Programas&dir=Energ%C3%81a%20total%20desglosada%20en%20el%20mercado%20diario&readdir=pdbc_stota
- [12] Copernicus Climate Data Store. *ERA5 reanalysis - single levels*. Recuperat de <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download>
- [13] Sendeco2. *Precios CO2*. Recuperat de <https://www.sendeco2.com/es/precios-co2>
- [14] Investing.com. *EUR/USD Datos históricos*. Recuperat de <https://www.investing.com/currencies/eur-usd-historical-data>

- [15] Investing.com. *Brent Oil Datos históricos*. Recuperat de <https://www.investing.com/commodities/brent-oil-historical-data>
- [16] MIBGAS. *Fichero AGNO 2021*. Recuperat de https://www.mibgas.es/es/file-access?path=AGNO_2021/XLS
- [17] INE - Instituto Nacional de Estadística. *Producto Interior Bruto (PIB)*. Recuperat de https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736164439&menu=ultiDatos&idp=1254735576581
- [18] INE - Instituto Nacional de Estadística. *Tasas de paro por distintos grupos de edad, sexo y comunidad autónoma*. Recuperat de: <https://www.ine.es/jaxiT3/Tabla.htm?t=65334>
- [19] XGBoost Developers, *XGBoost Documentation, Version 3.0.0*. Recuperat de: https://xgboost.readthedocs.io/en/release_3.0.0/
- [20] LightGBM Developers, *LightGBM Documentation (Stable version)*. Recuperat de: <https://lightgbm.readthedocs.io/en/stable/>
- [21] Keras Team, *Keras API Documentation: LSTM layer*. Recuperat de: https://keras.io/api/layers/recurrent_layers/lstm/
- [22] scikit-learn: Machine Learning in Python, *scikit-learn*. Recuperat de: <https://scikit-learn.org/stable/>.
- [23] StreamlitDocs, *Streamlit Documentation*. Recuperat de: <https://docs.streamlit.io/>