
TÉCNICAS *BIG DATA* PARA LA PREDICCIÓN DE LA DEMANDA Y PRECIO ELÉCTRICO

LAURA MELGAR GARCÍA
JOSÉ FRANCISCO TORRES MALDONADO
ALICIA TRONCOSO
JOSÉ CRISTÓBAL RIQUELME SANTOS

En la actualidad, la electricidad es un pilar indispensable en el día a día de la sociedad. De hecho, es calificada como uno de los indicadores más importantes para medir el nivel tecnológico e industrial de desarrollo de un país [1]. La concienciación sobre su uso de forma sostenible y responsable está siendo uno de los retos más importantes de las últimas décadas. Concretamente, en términos de energía, en [2] se describe la eficiencia energética como el factor de mitigación principal para rebajar el crecimiento del consumo energético, siendo éste crucial para el desarrollo sostenible.

A partir de la desregularización del mercado eléctrico, la posibilidad de prever o predecir los valores futuros que toman sus variables principales: demanda eléctrica y precio eléctrico, está generando mayor interés. El objetivo principal de obtener previsiones a futuro es aumentar los beneficios y/o reducir los costes. En cuanto a la demanda eléctrica, la previsión de su valor a futuro es muy útil, por ejemplo, para aportar información a los agentes implicados en la cadena de suministro energético donde destaca la generación, el almacenamiento y la distribución de energía eléctrica [2].

Por otro lado, la predicción del precio eléctrico implica tratar un problema de naturaleza volátil, no lineal y sin componente estacional, en la mayoría de los casos. Los productores y consumidores hacen uso de las previsiones de precio eléctrico del día anterior para desarrollar sus propias estrategias. Por ejemplo, si se predice de forma precisa y correcta las horas de precio alto, éstas podrán evitarse y centrarse en las horas en las que se prediga un precio menor [3].

La Agencia Internacional de Energía (*International Energy Agency* o IEA, en inglés) ha presentado recientemente sus previsiones del mercado eléctrico para el próximo año 2024 [4]. Algunas de las conclusiones extraídas para la Unión Europea son que el consumo de electricidad decreció en un 3,2% en 2022 convirtiéndose en la segunda mayor caída desde el 2009 con la crisis global financiera, solo después de la pandemia por Covid-19 en 2020. En el informe, la IEA prevé un repunte del 1,7% en 2024 pero haciendo hincapié en que se trata de un valor muy incierto ligado a la recuperación de la demanda industrial. En cuanto al precio eléctrico, en el documento la Agencia Internacional de Energía destaca los últimos años como los más comunes con valores de precio eléctrico negativos, en todo

el mundo. En el documento no se detallan predicciones de precio eléctrico específicas para el año 2024 en su totalidad, aunque sí se detalla que éste dependerá del gas en los territorios de la Unión Europea, como suele ser habitual.

Los algoritmos de Inteligencia Artificial permiten analizar y extraer información de conjuntos de datos. Una de sus aplicaciones más estudiadas y usadas es la predicción de valores futuros. Existen numerosas técnicas estadísticas tradicionales que permiten realizar predicciones. Sin embargo, los modelos de Inteligencia Artificial aportan ventajas interesantes sobre esas técnicas tradicionales. Algunas de ellas son: la capacidad de gestionar grandes volúmenes de datos haciendo uso de técnicas *Big Data* para reducir el uso de recursos y tiempos de ejecución, la detección de patrones de datos complejos, el manejo de relaciones no lineales complejas entre los datos, la capacidad de adaptar los algoritmos en tiempo real a los cambios que ocurren en los datos, etc.

En este artículo científico se presentan métodos para prever la demanda eléctrica y precio eléctrico a largo plazo (1 año) y corto plazo (siguiente hora, siguientes cuatro horas, etc.) haciendo uso de técnicas *Big Data* y modelos de Inteligencia Artificial.

TÉCNICAS BIG DATA

El gran volumen de datos que nos rodean en la actualidad ha derivado en que uno de los recursos más demandados por las empresas sea el uso de técnicas *Big Data*. Este concepto se refiere a datos no solo de gran volumen (como su propio nombre indica, en inglés), sino también a datos que pueden recibirse de diversas fuentes (datos variados), y que se reciben a gran velocidad. Estos tres términos aúnan la definición de las “tres Vs” tradicionales del *Big Data*. En las últimas décadas, se han ido añadiendo nuevas “Vs” al concepto *Big Data* como son el valor y la veracidad de los datos, entre otros, para concretar cómo debe ser este paradigma.

El auge del *Big Data* ha llevado, a su vez, al aumento del estudio de metodologías que obtienen información de calidad de los datos. Una de las técnicas más en desarrollo en la actualidad para este fin es la Inteligencia Artificial (IA). La IA nació con el objetivo de dotar a las máquinas u ordenadores con las mismas capacidades de inteligencia que los humanos. En este contexto, la definición de inteligencia ha sido ambigua durante varias décadas, en las cuales ha ido tomando distintos enfoques. A día de hoy, el concepto de inteligencia en IA se entiende como la posibilidad de dotar de capacidad de actuación racional a los sistemas. El objetivo final es conseguir que las máquinas aprendan por sí solas.

El aprendizaje automático o *Machine Learning* (ML) es una de las ramas más analizadas de la Inteligencia Artificial. Como se deduce de su nombre, esta técnica se enfoca en aprender automáticamente de experiencias pasadas sin tener que programar tareas específicamente. En otras palabras, estos algoritmos aprenden de los datos observando sus patrones y son capaces de tomar decisiones sin necesidad de que intervenga un ser humano.

A su vez, el aprendizaje profundo o *Deep Learning* (DL) es una rama del *Machine Learning*. El DL se inspira en el comportamiento del cerebro humano implementando redes neuronales con codificación matemática. En el cerebro humano los impulsos nerviosos se reciben a través de las dendritas neuronales, se procesan en el propio soma y se transmiten a la siguiente neurona mediante el axón. El aprendizaje profundo modela esta realidad a través de capas con neuronas o nodos que transmiten información de una capa a otra. Esa propagación de información, que equivaldría a los axones neuronales, se modela con pesos que se procesan en el interior de cada neurona, que equivaldría al soma, mediante una función de activación. Existen múltiples tipos de redes neuronales desde la sencilla red de perceptrón de una capa (*Single Layer Perceptron*, en inglés) hasta redes profundas convolucionales, recurrentes etc.

Uno de los principales desafíos cuando se pretende usar técnicas de aprendizaje profundo, aprendizaje automático e Inteligen-

cia Artificial es conseguir un conjunto de datos o *dataset* apropiado. Actualmente, gran parte de los datos que se generan disponen de una componente temporal, con datos indexados en el tiempo, también conocidos como series temporales de datos. Esta tipología de *datasets* lleva a introducir el concepto de ciudades inteligentes o *Smart Cities* donde se registran los datos de las ciudades mediante sensores o dispositivos electrónicos de alta frecuencia, pudiendo tener acceso a datos diarios, horarios o hasta incluso cada milisegundo. De forma específica una instancia de una serie temporal de datos está formada por una ventana de datos pasados denominada w (de *window*, ventana, en inglés) y un horizonte de predicción h de valores futuros a predecir. Cuando se trabaja con datos que dependen del tiempo es importante recibirlos y mantenerlos cronológicamente ordenados. Por tanto, una serie temporal X_t se define como un conjunto de datos cronológicamente ordenados que siguen la siguiente representación:

$$X_t = \{(x^1, y^1), \dots, (x^T, y^T)\} \quad x^i \in \mathbb{R}^w \quad y^i \in \mathbb{R}^h$$

donde la serie temporal se transforma en T instancias formadas por características y clases. La instancia i se representa como (x^i, y^i) con x^i haciendo referencia a los w datos pasados e y^i haciendo referencia a las h clases futuras a predecir. En este artículo científico estas clases a predecir se refieren a los valores numéricos futuros de la serie temporal.

En este artículo científico se presentan algoritmos de *Machine Learning* y *Deep Learning* para predecir datos de demanda eléctrica y precio eléctrico en España. Los modelos implementados para este fin se presentan en las siguientes secciones.

Algoritmos de *Deep Learning*

Las redes neuronales profundas desarrolladas en este artículo científico se centran en el entorno de aprendizaje supervisado donde el objetivo es predecir una clase o valor numérico real. En esta sección se describen de forma general las principales redes implementadas.

- **Perceptrón multicapa** (*Multi Layer perceptron* o MLP, en inglés). El MLP es una de las redes neuronales más sencillas donde todas las neuronas de la capa i están conectadas con las neuronas de la capa anterior $i-1$, lo que se conoce como red totalmente conectada o *fully connected*. No obstante, las neuronas de una misma capa son independientes e inconexas, es decir, trabajan de forma autónoma. El entrenamiento de esta red se basa en calcular las salidas de cada neurona en su capa usando el algoritmo *feed-forward* (hacia delante), calcular los errores de cada una de ellas y actualizar los pesos que propagan la información entre neuronas usando el algoritmo *back propagation* (hacia atrás). Por tanto, implementa un algoritmo de optimización para encontrar el mínimo, local o global, de la función coste o pérdida definida. La función coste expresa la diferencia entre los valores esperados o reales y los valores predichos por la red. Una descripción completa de esta arquitectura se encuentra en [5].
- **Red convolucional** (*Convolutional Neural Network* o CNN, en inglés). Tradicionalmente las CNN se centran en problemas relacionados con imágenes como clasificación o detección de objetos, aunque actualmente se usan cada vez más para lograr objetivos distintos. Las redes convolucionales están formadas principalmente por capas convolucionales, de ahí su nombre, que constan de un filtro o *kernel* deslizante cuya finalidad es crear un mapa de características. Otro tipo de capas a destacar en estas redes son las capas de agrupación donde se pretende reducir el número de parámetros haciendo uso de otro tipo de filtro o *kernel*. Dependiendo del objetivo a conseguir y de los datos de partida, estas capas se combinan de un modo u otro. De forma general, se pretende que en las primeras capas se identifiquen características simples y generales de los datos de entrada y que en las capas más avanzadas se consiga mayor detalle sobre ellas. En la arquitectura de una red convolucional CNN, las neuronas se conectan solo a una región de la entrada aplicando el concepto de

conectividad local. Además, las neuronas de una misma capa comparten la misma matriz de pesos para la capa de convolución. La última capa en una red CNN siempre estará totalmente conectada o *fully connected*. Para más detalles sobre esta red consultar [6].

- **Redes convolucionales temporales** (*Temporal Convolutional Network* o TCN, en inglés). Las TCN son una adaptación de las redes convolucionales tradicionales específica para las series temporales de datos. Las redes convolucionales temporales usan convoluciones causales dilatadas para aprender las dependencias a largo plazo en series temporales. También pueden realizar las convoluciones en paralelo y usan un solo filtro o *kernel* por capa. Los detalles de esta arquitectura se encuentran en [7].
- **Memoria a corto plazo** (*Long Short Term Memory* o LSTM, en inglés). Las redes neuronales recurrentes (*Recurrent Neural Network* o RNN, en inglés) se aplican sobre series temporales de datos. Una de las RNN más comunes son las de memoria a corto plazo. Las LSTM permiten conservar las entradas durante un largo periodo de tiempo dependiendo de la importancia que tenga cada una de las ellas. Las redes de memoria a corto plazo contienen tres puertas en sus capas ocultas: puerta de entrada, puerta de salida y puerta de olvido. Estas puertas controlan el flujo de información que se necesita para predecir la salida en la red. En [8] se define la red LSTM con más detalle.
- **NDL** (*Neuroevolutionary Deep Learning*, en inglés). Los algoritmos evolutivos, inspirados en el proceso de selección natural han resultado de interés como herramienta en la optimización de modelos Deep Learning. Estos algoritmos utilizan operadores genéticos para evolucionar soluciones a lo largo del tiempo, permitiendo una exploración eficiente de un amplio espacio de posibles configuraciones. Su capacidad para realizar búsquedas y adaptarse dinámicamente a cambios en el entorno los distingue en la optimización de modelos. Esta red neuronal se describe en [9].

Algoritmos de *Machine Learning* tradicional

Los algoritmos tradicionales de aprendizaje automático que se presentan en este artículo científico se basan en modelos de toma de decisión. Ambos algoritmos se encuadran dentro del tipo de aprendizaje supervisado y pueden implementarse para predecir valores numéricos o clases, es decir, problemas de regresión o clasificación, respectivamente. En [10] se definen más en profundidad ambos algoritmos.

- **Árbol de decisión** (*Decision Tree*, en inglés). Este algoritmo de *Machine Learning* es uno de los más sencillos de interpretar debido a que el resultado final que ofrecen es un modelo que se representa como un diagrama de flujo con forma de árbol. Los árboles de decisión siguen un proceso recursivo en el que se parte de un nodo inicial o raíz que se va dividiendo en nodos internos o hijos hasta llegar a los nodos finales u hojas. En cada uno de los nodos internos se pretende dividir los datos de la manera más correcta. Para ello se usan medidas de selección como la ganancia de información o la entropía. De esta manera se decide cuál es el mejor atributo para cada nodo y su rango de valores. Se pretende que la mayoría de las instancias del conjunto de datos se clasifiquen correctamente con el árbol creado.
- **Bosque aleatorio** (*Random Forest*, en inglés). Uno de los desafíos de los árboles de decisión es no llegar a sobre ajustarse (*overfitting*, en inglés) a los datos con los que se entrena. Este riesgo se reduce al trabajar con bosques aleatorios. Un bosque aleatorio contiene varios árboles de decisión formados cada uno de ellos por distintos subconjuntos de atributos y distintos subconjuntos de instancias que se seleccionan de forma aleatoria. La predicción final del algoritmo se realiza calculando la predicción de cada árbol de decisión individualmente y considerando el resultado más repetido (si es un problema de clasificación) o calculando la media de las predicciones (si es un problema de regresión). Además de la reducción del

sobre ajuste debido a la aleatoriedad del algoritmo, los bosques aleatorios son más flexibles que un único árbol de decisión. Sin embargo, un bosque aleatorio es más complicado de interpretar, a simple vista, y requiere mayor tiempo y recursos para modelarse.

Algoritmo de Machine Learning con aprendizaje dinámico en tiempo real

En esta sección se presenta una modificación del algoritmo tradicional K vecinos cercanos (*K-Nearest Neighbors* o KNN, en inglés). Este nuevo algoritmo se denomina: **StreamWNN** donde “*Stream*” hace referencia a su capacidad para recibir datos en tiempo real (*streaming*, en inglés) y “*WNN*” a una versión ponderada del KNN (de *weighted*, en inglés). El *StreamWNN* proporciona predicciones en tiempo real y actualiza su modelo a medida que recibe datos, aplicando un aprendizaje de tipo incremental o dinámico.

Los modelos tradicionales de *Machine Learning* se basan en un aprendizaje estático que usan datos también estáticos. Sin embargo, la actual tendencia de recibir continuamente datos y querer procesarlos a medida que se reciben, no es compatible con el modelado tradicional de ML. En el caso habitual del aprendizaje automático, cuando se quiere incorporar nueva información a un modelo ya entrenado se debe volver a entrenar todo el modelo por completo. Esta necesidad puede darse debido, por ejemplo, a que existan nuevos patrones en los datos actuales que sean totalmente distintos a los patrones de los datos con los que se entrenó el modelo. Este cambio de comportamiento o de patrones en los datos a lo largo del tiempo es común en muchos casos reales. El modelado incremental o dinámico con datos que se reciben de forma continua o en *streaming*, permite actualizar el propio modelo en tiempo real considerando los nuevos patrones que se detecten en los datos.

El *StreamWNN* modela el aprendizaje incremental mediante dos fases diferenciadas: una fase offline seguida por una fase online. En la primera fase, se crea un mode-

lo de *Big Data* distribuido con datos estáticos usando como base la metodología del KNN tradicional, donde cada predicción se calcula considerando el comportamiento de sus K registros más similares, siendo K un valor numérico entero. El modelo offline mapea cada una de las series temporales del 30% más reciente de los datos estáticos con sus K instancias más similares (vecinos más cercanos) del 70% restante de la serie temporal estática. Una vez el modelo offline está validado y creado, se comienzan a recibir datos dinámicos a medida que el sensor o dispositivo electrónico los genera.

En la fase online, se deben cumplir todos los requisitos del paradigma *streaming* [11]. Los 4 requisitos principales de este paradigma son: los datos que se reciben en tiempo real se procesan uno por uno, solo una vez cada uno de ellos y siempre en el orden de llegada; el modelo puede conservar o guardar solamente un conjunto limitado de datos pasados; el modelo debe proporcionar respuestas lo más rápido posible (baja latencia); y, por último, el modelo tiene que adaptarse al concepto de deriva de los datos (*concept drift*, en inglés) y así proporcionar respuestas ajustadas a los nuevos patrones de los datos. En la segunda fase se deben cumplir 3 objetivos: predecir los valores futuros de la serie temporal de forma precisa y rápida considerando el modelo actual; actualizar el modelo incrementalmente, considerando los nuevos patrones que se detecten en los datos que se reciben dinámicamente; y detectar comportamientos poco usuales que se espera que no vuelvan a repetirse y que deben ser revisados de forma individual. El algoritmo se describe por completo en [12].

PREDICCIÓN DE DEMANDA Y PRECIO ELÉCTRICO EN ESPAÑA

La electricidad es esencial para el desarrollo económico e industrial de la sociedad. Uno de los mayores desafíos de la gestión de la electricidad es la dificultad para almacenar grandes cantidades de ella. Es por ello que es necesario un equilibrio entre demanda y oferta, lo que conlleva, además, a precios de mercado altamente volátiles. Estas carac-

terísticas de la demanda y precio eléctrico hacen que su predicción sea necesaria.

Para poder evaluar el comportamiento de los modelos de *Machine Learning* y *Deep Learning* que se implementan, se calculan las siguientes métricas de error:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

, donde y_i es el valor real del registro i , $\hat{y}_i$ es el valor predicho del registro i y n es el número total de instancias en el conjunto de datos que se evalúa. La métrica de error *MAPE* toma su nombre de *Mean Absolute Percentage Error*, en inglés, y se refiere por tanto al error absoluto medio en porcentaje. Por otro lado, la métrica de error *MAE* toma su nombre de *Mean Absolute Error*, en inglés, y se refiere al error absoluto medio cometido. En ambos casos, cuanto menor sea el error *MAPE* o *MAE* cometido, mejores serán las predicciones.

En las siguientes secciones se presenta cómo se ha abordado cada uno de estos problemas de forma independiente implementando las técnicas de *Big Data* e Inteligencia Artificial descritas en la sección anterior.

Predicción de demanda eléctrica española

La predicción de la demanda eléctrica tiene como objetivo principal mejorar la eficiencia de la gestión de la misma. La demanda eléctrica española se registra a lo largo del tiempo y sus valores se publican cada 10 minutos en webs como <https://www.omie.es>. En particular, en este trabajo se aplican algoritmos de predicción univariados donde la única característica de la serie temporal, además del tiempo, que se considera para predecir es la propia demanda eléctrica en megavatios (MW). La experimentación se realiza con 497.832 instancias que abarcan la demanda eléctrica española durante 9 años y 6 meses: desde el 1 de enero de 2007 hasta el 21 de junio de

2016. Este periodo temporal ha sido seleccionado para poder realizar comparaciones entre estudios científicos disponibles que trabajan con dicha ventana temporal.

La serie temporal seleccionada se preprocesa antes de aplicarse a los algoritmos descritos para poder obtener resultados reales y útiles. Por ejemplo, los registros en los días de cambio horario en España se preprocesan para mantener el total de 6 registros cada hora y proporcionar, así, 144 registros para cada día.

Debido al hecho de que este problema aborda un aprendizaje de tipo supervisado y de regresión en el entorno de *Machine Learning*, es esencial dividir los datos en un conjunto de entrenamiento (*training*, en inglés) y conjunto de prueba (*test*, en inglés). La serie temporal se divide por tanto en, aproximadamente, el 30% de los datos temporales más recientes para el conjunto de prueba y restante 70% para el conjunto de entrenamiento. Concretamente se usan los registros de demanda eléctrica entre el 1 de enero de 2007 y el 19 de agosto de 2013 para entrenar los modelos de aprendizaje profundo y aprendizaje automático tradicional y dinámico. Los datos temporales entre el 20 de agosto de 2013 y el 21 de junio de 2016 se usan para testear los modelos obtenidos con registros distintos a los usados durante el entrenamiento, siguiendo la definición de *training-test* de los modelos de ML y DL.

Cada modelo se ejecuta independientemente y se realiza su propio ajuste de parámetros individualmente. Cabe destacar que todas las arquitecturas analizadas usan como ventana de datos pasados (w) 144 registros que equivalen a un día y predicen (h) 24 registros futuros que equivalen a las 4 horas siguientes. Los parámetros óptimos para los modelos de *Deep Learning* y *Machine Learning* tradicional se seleccionan mediante una búsqueda evolutiva, específicamente usando el paquete GA de R [13] y se encuentran detallados en [9]. En el caso del algoritmo de *Machine Learning* dinámico los parámetros se seleccionan mediante una búsqueda en rejilla o *grid search*, en inglés, y se encuentran detallados en [11].

Los resultados obtenidos se muestran en la Tabla 1:

TABLA 1
RESULTADOS PREDICCIÓN DEMANDA
ELÉCTRICA ESPAÑOLA

MODELO	MAPE (%)
Deep Learning	
■ MLP	2,32
■ CNN	1,71
■ LSTM	1,78
■ NDL	1,51
Machine Learning tradicional	
■ Decision Tree	8,86
■ Random Forest	2,22
Machine Learning dinámico	
■ StreamWNN	2,19

Fuente: elaboración propia

Como se puede observar en la Tabla 1 las técnicas usadas de aprendizaje profundo obtienen buenos errores de predicción. Concretamente, la arquitectura NDL es la que proporciona el menor error posible entre todos los modelos representados en dicha Tabla. Es importante tener en cuenta que a pesar de la exactitud que habitualmente proporcionan los algoritmos de *Deep Learning*, el tiempo de ejecución y uso de recursos de éstos es altamente superior a los de las arquitecturas de *Machine Learning*. Dependiendo de la aplicación específica que quiera darse a las predicciones, tendrá más importancia o menos emplear un elevado tiempo en entrenar un modelo de aprendizaje profundo y elevados recursos para conseguir mayor precisión en las predicciones.

En cuanto a los algoritmos de *Machine Learning* tradicional, cabe destacar la gran reducción del error cometido MAPE para la técnica *Random Forest*. Haciendo referencia a la definición del algoritmo realizada previamente, esta mejora en el resultado se debe a la reducción del sobre ajuste a los datos de entrenamiento que realiza el algoritmo de *Decision Tree*.

Por último, el algoritmo de aprendizaje automático dinámico *StreamWNN* obtiene

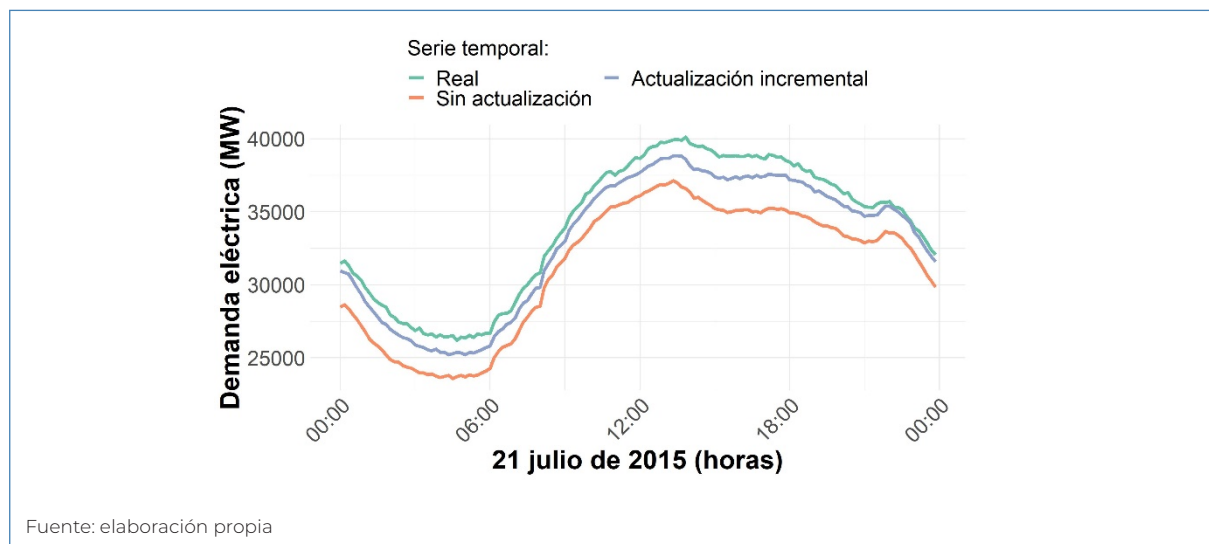
un error de predicción menor que los de aprendizaje automático tradicional, aunque no mejora los resultados obtenidos por los algoritmos de aprendizaje profundo, a excepción de la arquitectura MLP o perceptrón multicapa. El objetivo de este algoritmo es distinto al del resto de los modelos de la Tabla 1, al igual que la forma de recibir y tratar los datos. En este caso, el modelo *StreamWNN* no se mantiene invariable después del entrenamiento, sino que se actualiza de forma continua a medida que recibe nuevas instancias de la serie temporal. Por tanto, el fin principal de este algoritmo es mantener el modelo siempre actualizado a los nuevos patrones de los datos y, sobre todo, proporcionar predicciones rápidas o en tiempo real. En la Figura 1 se muestran los valores reales de la demanda eléctrica del día 21 de julio de 2015 junto con las predicciones de ese día del algoritmo *StreamWNN* sin actualizar de forma dinámica y actualizándolo dinámicamente. El 21 de julio de 2015 fue el peor día predicho del mes de julio de 2015 con este algoritmo. El error en métrica MAPE cometido para el modelo *StreamWNN* sin actualización durante la etapa online es mayor del 8%. Sin embargo, cuando el modelo se actualiza dinámicamente considerando los nuevos patrones detectados, se comete un error del 2,91%.

Todos los algoritmos que se presentan son escalables, siguiendo así los requisitos del paradigma *Big Data*. Es decir, los algoritmos no ven influenciada su capacidad de respuesta al aumentar el número de datos tanto en el conjunto de entrenamiento como en el conjunto de prueba o *test*.

Predicción de precio eléctrico español

El precio eléctrico español se registra de forma horaria en términos de €/megavatio hora (€/MWh) y se publica en webs como <https://www.esios.ree.es>. Debido a la naturaleza volátil de este tipo de dato, se seleccionan 3 periodos temporales con características diferentes para evaluar los modelos de predicción propuestos.

FIGURA 1
PEOR DÍA PREDICHO EN EL MES DE JULIO DE 2015 USANDO STREAMWNN



El primero de los periodos se define como “periodo normal” ya que no se detectó ninguna anomalía ni patrón diferente a los esperados. Este primer experimento presenta la predicción del precio eléctrico español de forma horaria entre el 18 de noviembre de 2019 y el 15 de diciembre de 2019. El segundo de los periodos se califica como “periodo cuarentena” y presenta la predicción entre el 19 de mayo de 2020 y el 14 de junio de 2020. Este periodo toma dicho nombre debido a la pandemia global por Covid-19 que mantuvo en cuarentena a la población española modificando así, muchos de los patrones de consumo domésticos e incluso industriales. El tercer y último periodo que se analiza se denomina “periodo de fraude” puesto que hace referencia a un periodo en el que se alteraron los precios eléctricos en España [1]. Para este último experimento se predicen los registros desde el 5 de diciembre de 2016 hasta 1 de enero de 2017.

En este caso, la serie temporal de precio eléctrico español también se preprocesa. Se consigue así que cada día contenga 24 registros.

Al igual que en el estudio de la demanda eléctrica española, los datos se dividen en conjuntos de entrenamiento y *test* para evaluar los modelos implementados. Los

conjuntos de prueba (conjuntos de *test*) de cada uno de los tres periodos de estudio, son los indicados previamente. Los conjuntos de entrenamiento de cada uno de ellos se remontan a 2 meses previos al conjunto de *test*. Es decir, el periodo normal se entrena con datos desde el 15 de septiembre de 2019 hasta el 17 de noviembre de 2019, el periodo de cuarentena con datos desde el 15 de marzo de 2020 hasta el 18 de mayo de 2020 y el periodo de fraude con datos desde el 1 de octubre de 2016 hasta el 4 de diciembre de 2016.

Los modelos se implementan, entrenan y validan individualmente. Los hiper parámetros óptimos de cada modelo se encuentran aplicando una búsqueda en rejilla. Los valores de cada uno de ellos considerados óptimos y usados finalmente para la construcción de cada modelo se encuentran en [1] para los modelos de aprendizaje profundo y aprendizaje automático tradicional y en [11] para el modelo de aprendizaje automático dinámico.

Los resultados obtenidos por los modelos de aprendizaje profundo y aprendizaje automático tradicional y dinámico se encuentran en la Tabla 2, Tabla 3 y Tabla 4 para los periodos normal, cuarentena y fraude, respectivamente.

TABLA 2
RESULTADOS PREDICCIÓN PRECIO
ELÉCTRICO ESPAÑOL PARA PERIODO
NORMAL

MODELO	MAE (€/MWH)	MAPE (%)
Deep Learning		
■ MLP	4,19	10,10
■ CNN	4,12	10,20
■ LSTM	4,21	9,8
■ TCN	4,99	12,00
Machine Learning tradicional		
■ Decision Tree	4,16	9,5
■ Random Forest	4,15	9,6
Machine Learning dinámico		
■ StreamWNN	4,04	9,35

Fuente: elaboración propia

TABLA 4
RESULTADOS PREDICCIÓN PRECIO
ELÉCTRICO ESPAÑOL PARA PERIODO
FRAUDE

MODELO	MAE (€/MWH)	MAPE (%)
Deep Learning		
■ MLP	4,83	7,20
■ CNN	4,84	7,20
■ LSTM	4,44	6,80
■ TCN	5,60	8,50
Machine Learning tradicional		
■ Decision Tree	4,03	5,90
■ Random Forest	4,88	7,10
Machine Learning dinámico		
■ StreamWNN	3,89	5,77

Fuente: elaboración propia

TABLA 3
RESULTADOS PREDICCIÓN PRECIO
ELÉCTRICO ESPAÑOL PARA PERIODO
CUARENTENA

MODELO	MAE (€/MWH)	MAPE (%)
Deep Learning		
■ MLP	3,64	12,70
■ CNN	4,19	14,80
■ LSTM	3,53	12,60
■ TCN	3,99	13,50
Machine Learning tradicional		
■ Decision Tree	4,06	13,30
■ Random Forest	4,41	14,80
Machine Learning dinámico		
■ StreamWNN	3,38	11,68

Fuente: elaboración propia

Los resultados de la predicción de precio eléctrico muestran que la arquitectura que menores errores comete, tanto en términos de *MAE* como de *MAPE*, es el algoritmo de *Machine Learning* dinámico *StreamWNN*. Este algoritmo aporta la posibilidad de ir actualizando el modelo a medida que recibe datos en *streaming*. Esta característica es de especial relevancia en problemas volátiles como la predicción de precio eléctrico. En las Figuras 2, 3 y 4 se representan los valores reales y predichos por *StreamWNN* con actualización dinámica de los periodos normales, cuarentena y fraude, respectivamente.

En cuanto a los algoritmos de aprendizaje profundo el mejor algoritmo en términos de *MAPE* para todos los periodos temporales es la red de memoria a corto plazo o LSTM. En términos de *MAE* la LSTM es la mejor red profunda para el periodo de cuarentena y de fraude, pero no para el periodo normal donde destaca la red convolucional o CNN.

Sin embargo, cabe destacar que los modelos de aprendizaje automático tradicional

FIGURA 2
PREDICCIÓN PRECIO ELÉCTRICO PERIODO NORMAL CON STREAMWNN

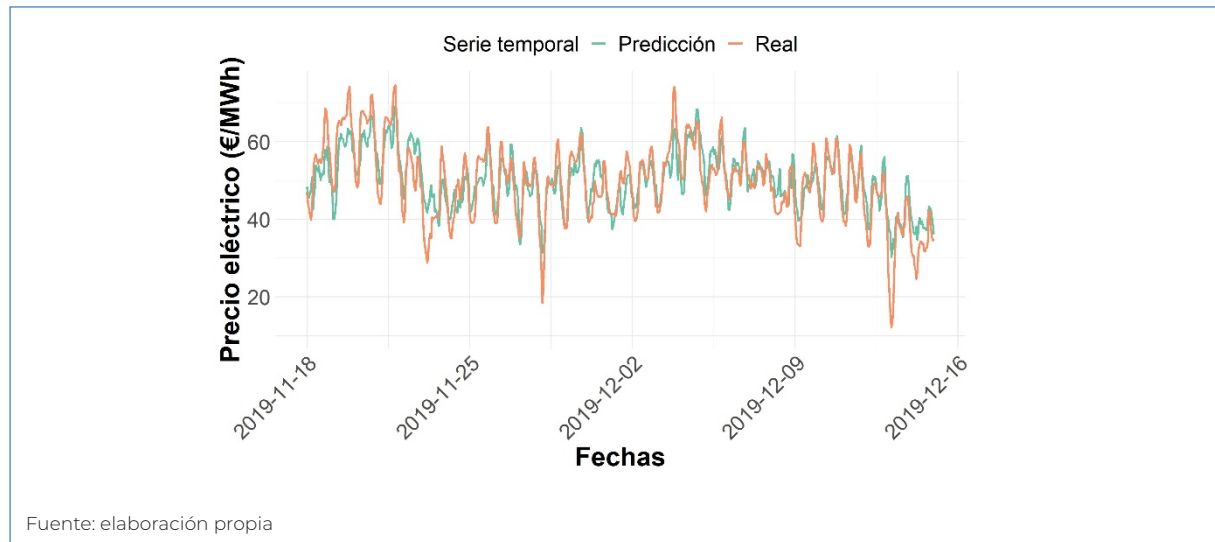
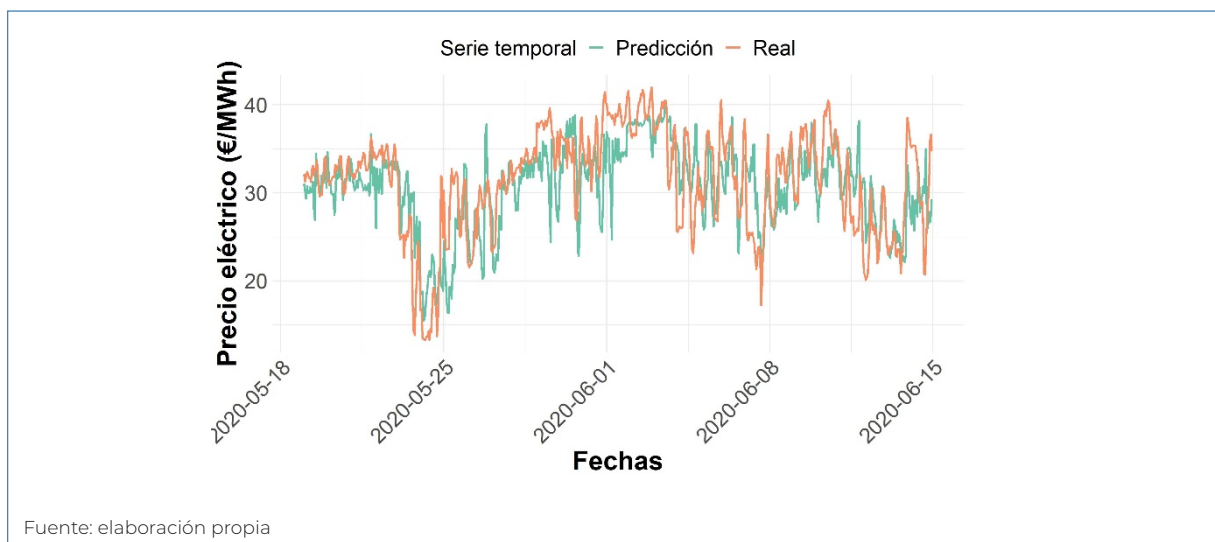


FIGURA 3
PREDICCIÓN PRECIO ELÉCTRICO PERIODO CUARENTENA CON STREAMWNN



implementados (árboles de decisión y bosque aleatorio) obtienen errores muy similares a los obtenidos por las redes profundas. Además, en el caso del periodo de cuarentena y fraude los resultados obtenidos son mejores para los árboles de decisión que para el bosque aleatorio.

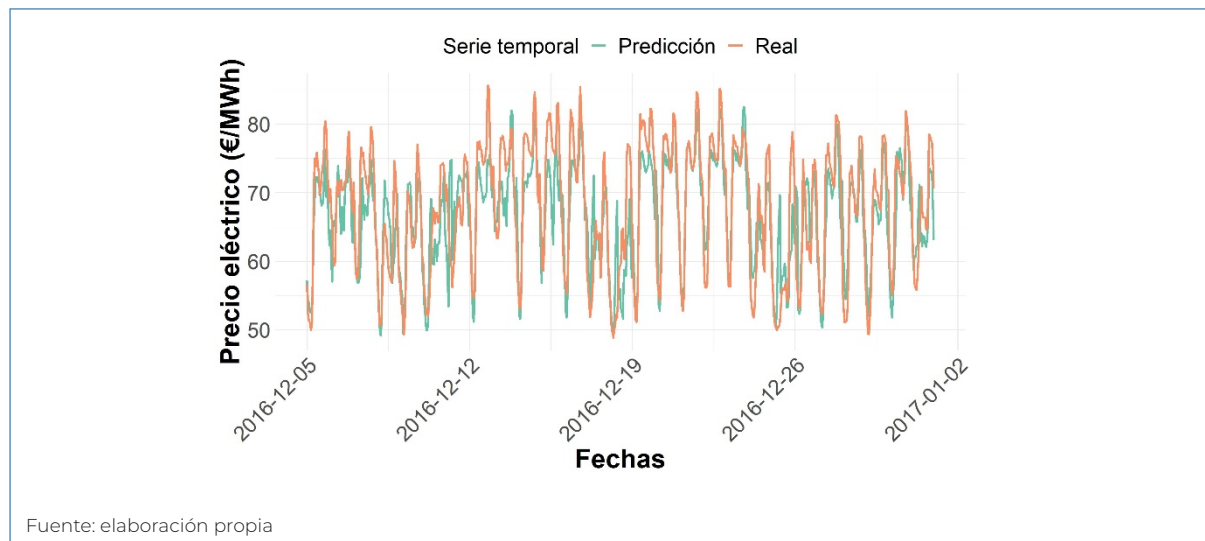
Por otro lado, los algoritmos de aprendizaje profundo conllevan un tiempo de ejecución bastante más elevado para la etapa de selección de hiper parámetros y entrenamiento del modelo, comparado con los

algoritmos de aprendizaje automático tradicional y dinámico.

CONCLUSIONES

En este artículo científico se han presentado técnicas que trabajan en el paradigma *Big Data* para predecir demanda y precio eléctrico español. Debido a las características de ambas series temporales de datos, se ha optado por aplicar modelos de Inteligencia

FIGURA 4
PREDICCIÓN PRECIO ELÉCTRICO PERIODO FRAUDE CON STREAMWNN



Artificial para conseguir predicciones precisas siguiendo un aprendizaje de tipo supervisado para problemas de regresión. Específicamente se han aplicado algoritmos de aprendizaje automático tradicional basados en árboles como los árboles de decisión y bosques aleatorios; el algoritmo *StreamWNN* de aprendizaje automático con actualización dinámica; y distintas arquitecturas de aprendizaje profundo como las redes perceptrón multicapa, redes convoluciones o redes de memoria a corto plazo.

En el problema de predicción de la demanda eléctrica española el modelo que ha proporcionado mejores resultados para el periodo temporal de prueba seleccionado de casi 3 años ha sido la red neuronal NDL. En cuanto al problema de predicción de precio eléctrico, el modelo que comete menor error para los tres periodos temporales seleccionados, con características distintas y de 1 mes cada uno de ellos, ha sido el algoritmo de aprendizaje automático dinámico *StreamWNN*. Es importante destacar que en cada problema se trata un horizonte de predicción distinto: a largo plazo para el caso de demanda eléctrica y a corto plazo para el caso de precio eléctrico.

Los modelos de Inteligencia Artificial que se han implementado tienen peculiaridades distintas que deben tenerse en cuenta. Por un lado, las redes neuronales profundas consumen más recursos y sus tiempos de

ejecución son mayores, pero son óptimas para aplicar a grandes conjuntos de datos y obtienen normalmente respuestas muy precisas. Sin embargo, requieren de algoritmos externos para conseguir entender el funcionamiento del modelo y las predicciones proporcionadas, como por ejemplo haciendo uso de algoritmos de Inteligencia Artificial Explicable (*Explainable Artificial Intelligence* o XAI, en inglés). Sin embargo, los 3 algoritmos de aprendizaje automático implementados (tradicionales y dinámico) son más fáciles de interpretar. Por otro lado, el algoritmo de aprendizaje automático dinámico tiene como objetivo proporcionar predicciones rápidamente y ajustadas a los nuevos patrones que se reciben en los datos, por eso se actualiza dinámicamente. Es por ello que este último no consigue superar los resultados de las arquitecturas de aprendizaje profundo para el caso de la demanda eléctrica, ya que se ha abordado un problema de predicción a largo plazo que no es el óptimo, a priori, para la naturaleza del *StreamWNN* en tiempo real.

REFERENCIAS

- [1] Vega-Márquez, B., Rubio-Escudero, C., Nepomuceno-Chamorro, I.A., Arcos-Vargas, Á. *Use of Deep Learning Architectures for Day-Ahead Electricity Price Forecasting over Different Time Periods in the Spanish Electricity Market*. Appl. Sci. 2021, 11, 6097.

- [2] González-Torres, M., Pérez-Lombard, L., Coronel, J., Maestre, I. *A cross-country review on energy efficiency drivers*. Applied Energy. 2021, 289. 116681.
- [3] Mathaba, T., Xia, X., Zhang, J. *Analysing the economic benefit of electricity price forecast in industrial load scheduling*. Electric Power Systems Research. 2014, 116, 158-165.
- [4] International Energy Agency. *Electricity Market Report Update. Outlook for 2023 and 2024*. 2023.
- [5] Popescu, M., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.E. *Multilayer perceptron and neural networks*. WSEAS Transactions on Circuits and Systems archive. 2009, 8, 579-588.
- [6] Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J. *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects*. IEEE Transactions on Neural Networks and Learning Systems. 2022, 33, 12, 6999-7019.
- [7] Lea, C., Vidal, R., Reiter, A., Hager, G.D. *Temporal Convolutional Networks: A Unified Approach to Action Segmentation*. ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science. 2016, 9915.
- [8] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., Weyrich, M. *A survey on long short-term memory networks for time series prediction*. Procedia CIRP. 2021, 99, 650-655.
- [9] Divina, F., Torres Maldonado, J.F., García-Torres, M., Martínez-Álvarez, F., Troncoso, A. *Hybridizing Deep Learning and Neuroevolution: Application to the Spanish Short-Term Electric Energy Consumption Forecasting*. Appl. Sci. 2020, 10, 5487.
- [10] Bernard, S., Heutte, L., Adam, S. *On the selection of decision trees in Random Forests*. International Joint Conference on Neural Networks. 2009, 302-307.
- [11] Melgar-García, L., Gutiérrez-Avilés, D., Rubio-Escudero, C., Troncoso, A. *A novel distributed forecasting method based on information fusion and incremental Learning for streaming time series*. Information Fusion. 2023, 95, 163-173.
- [12] Melgar-García, L., Gutiérrez-Avilés, D., Rubio-Escudero, C., Troncoso, A. *Identifying novelties and anomalies for incremental learning in streaming time series forecasting*. Engineering Applications of Artificial Intelligence. 2023, 123, Part B, 106326.
- [13] Scrucca, L. *On some extensions to GA Package: Hybrid optimisation, parallelisation and islands evolution*. The R Journal. 2017, 9/1, 187-206.

SOBRE LOS AUTORES

Laura Melgar García obtuvo el Doctorado Internacional en Ciencias de la Computación por la Universidad Pablo de Olavide en marzo de 2023. En su formación académica destaca su estancia de investigación predoctoral en la New York University durante 6 con beca Fulbright. Sus investigaciones se centran en la implementación de modelos de Machine Learning con aprendizaje incremental para datos recibidos en streaming. Actualmente es investigadora postdoctoral FPU y cuenta con 15 artículos científicos en revistas y congresos.

José Francisco Torres Maldonado es doctor en Ciencias de la Computación por la Universidad Pablo de Olavide de Sevilla desde marzo de 2022, obteniendo Premio Extraordinario de doctorado. Actualmente ejerce como Profesor Ayudante Doctor en el Área de Lenguajes y Sistemas Informáticos de misma universidad. Su investigación se centra en el desarrollo de modelos deep learning aplicados a la predicción de series temporales. Cuenta con más de 20 artículos publicados entre revistas de impacto y congresos internacionales.

Alicia Troncoso Lora, Ph.D. es Catedrática de Lenguajes y Sistemas Informáticos en la Universidad Pablo de Olavide (UPO). Con anterioridad, fue profesora en la Universidad de Sevilla hasta su incorporación a la UPO en 2005. Su formación postdoctoral ha sido completada en Estados Unidos, concretamente en Columbia University de Nueva York, la Universidad de Colorado en Boulder y la Universidad de California en San Diego. Sus áreas de investigación de interés incluyen la predicción de series temporales, Machine Learning y Big Data.

José Cristóbal Riquelme Santos es Catedrático de Universidad en la ETS de Ingeniería Informática de la Universidad de Sevilla. Durante más de 25 años ha trabajado en Inteligencia Artificial, especialmente en técnicas de Machine Learning aplicadas a series temporales y datos en streaming. Es autor más de 200 artículos y ponencias en congresos, director de 14 tesis doctorales e investigador principal de más de 25 proyectos y contratos de investigación. José es Licenciado en Matemáticas, especialidad de Estadística e Investigación Operativa y Doctor en Informática. Ambos títulos obtenidos en la Universidad de Sevilla con Premio Extraordinario.