

The logo consists of a dark red square on the left containing the letters 'IMF' in white, serif, all-caps font. To the right of the square, the words 'Business' and 'School' are stacked vertically in a dark red, serif font.

IMF

**Business
School**

MÁSTER EN DATA SCIENCE Y BUSINESS ANALYTICS ONLINE

**TRABAJO FIN DE MÁSTER: “MODELOS DE MACHINE LEARNING PARA LA
PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL”**

TFM elaborado por: **Iñigo Elorza Barea**

Tutor/a de TFM: **Abel Ángel Soriano Vázquez**

- Madrid a 31 de julio de 2023 -

1. SÍNTESIS DEL TRABAJO

El sector eléctrico en España es un mercado altamente complejo debido a los numerosos **factores** como:

- Equilibrio que debe existir entre la generación y la demanda.
- Fuentes de producción son muy diversas.
- Cada tecnología de producción tiene un coste de producción asociado.
- Transición actual de fuentes de energía altamente contaminantes a fuentes renovables.
- Las fuentes de energía renovables dependen de las condiciones meteorológicas.
- Coste ambiental de derechos de emisiones de CO₂.
- Legislación actual del sector aumenta la complejidad en el sector.

La predicción del precio de la electricidad se ha convertido en un elemento esencial en la toma de decisiones en las compañías del sector.

Gracias la capacidad computacional actual y al desarrollo de software capaz de manejar con efectividad grandes volúmenes de datos, se pueden emplear herramientas de **Deep Learnin** para **predecir el precio de la electricidad a largo plazo**.

Con este trabajo se pretende analizar diferentes modelos predictivos para pronosticar el precio de la energía eléctrica en nuestro país y analizar cual de ellos son los mejores.

2. MODULOS DEL MÁSTER RELACIONADOS CON EL TRABAJO FIN DE MÁSTER

- **MODULO I:** lenguajes de programación Python y R.
- **MODULO II:** Impacto y valor del Big Data y **MODULO III:** La Ciencia de Datos. Técnicas de análisis, minería y visualización, donde se estudian conceptos necesarios para poder operar correctamente con los datos (obtención, limpieza, transformación y visualización).
- **MODULO V:** Estadística para Científicos de Datos, para el análisis estadístico de los datos.
- **MODULO VI:** algoritmos de Aprendizaje Automático, especialmente problemas de regresión y redes neuronales.
- **MODULO VII:** técnicas para la toma de decisiones.

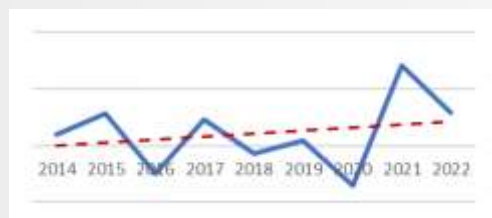
3. SISTEMA ELECTRICO ESPAÑOL

GENERACIÓN NO RENOVABLE	
Turbinación bombeo	Residuos no renovables
Nuclear	Motores diésel
Ciclo combinado	Turbina de vapor
Carbón	Turbina de gas
Cogeneración	Fuel + Gas

GENERACIÓN RENOVABLE
Hidráulica
Eólica
Solar fotovoltaica
Solar térmica
Otras renovables
Residuos renovables
Hidroeléctrica

3.3. EVOLUCION DE LA PRODUCCION DE LA EE

La tendencia del consumo de energía eléctrica en España es ascendente.



Cada vez es más elevado el empleo de tecnologías renovables.

3.4. AGENTES DEL MERCADO

La actual Ley del Sector Eléctrico (LSE) establece los siguientes sujetos participantes en el sector eléctrico:

- Productores
- Transportista
- Distribuidores
- Comercializadores
- Consumidores
- Gestores de cargas del sistema
- Operador del Mercado Ibérico (OMI)
- Operador del sistema

3.5 FORMACION DEL PRECIO EN EL MERCADO MAYORISTA A PLAZO EN ESPAÑA

El **mercado eléctrico** es el conjunto de plataformas de negociación en las que se contrata energía eléctrica para su entrega en diferentes horizontes temporales, que pueden ser **a plazo** (para las próximas semanas, meses, trimestres o años) o **al contado** (para el día siguiente o las horas siguientes).

Los **mercados a plazo** sirven para que los agentes del mercado puedan realizar sus planes económicos evitando altos riesgos de pérdidas, al tener que adquirir la electricidad en el mercado diario a un precio más elevado del que ofreció a sus clientes tiempo atrás.

El **precio se determina por el cruce entre la curva de oferta** (integrada por todas las ofertas que realizan los vendedores) **y la curva de demanda** (integrada por todas las ofertas que realizan los compradores).



4. BUSSINES CASE

Como afirman muchos investigadores como Bunn, D. W. (2004), Eydeland, A., & Wolyniec, K. (2003) o Weron, R. (2006), una herramienta de predicción del precio de la electricidad se ha convertido en una pieza clave en la toma de decisiones en las compañías del sector.

En la documentación consultada se han encontrado pocas referencias a estudios para la predicción de precios de la electricidad a más de tres meses. Por este motivo, la predicción del precio para el mercado a plazo es un territorio desaprovechado y supone una oportunidad para realizar un estudio con el objetivo de aportar nuevos mecanismos en la toma de decisiones para las compañías del sector.

También se ha observado que los diferentes modelos se centran en uno o, a lo sumo, dos factores que afectan al precio. En nuestro caso hemos contado con 22 variables correspondientes a 6 factores diferentes.

FACTOR	NOMBRE	FACTOR	NOMBRE
Temporal	Día de la semana	Producción EE según tecnología	Prod. en parques eólicos
Demanda	Demanda media diaria		Prod. en parques solares
Precios materias primas	Precio del petroleo		Prod. Hidráulica
	Precio del gas natural		Prod. otras f. renovables
	Precio del carbon		Prod. en centrales nucleares
Meteorología	Temperatura		Prod. termicas de gasóleo
	Velocidad del viento		Prod. con turbina de gas
	Reservas hidraulicas		Prod. termica de carbón
Otros factores	Precio Derechos emisión CO ₂		Prod. con ciclo combinado
	Sit. socio-económica del pais		Prod. cogeneración
	Inter. EE con otros países		Prod. otras f. no renovables

OBJETIVOS A ALCANZAR

El objetivo principal es analizar diferentes modelos predictivos para pronosticar el precio de la energía eléctrica en nuestro país, y analizar cual de ellos es el que ofrece mejores resultados, comparándolos con los resultados de modelos estudiados en artículos especializados en el sector.

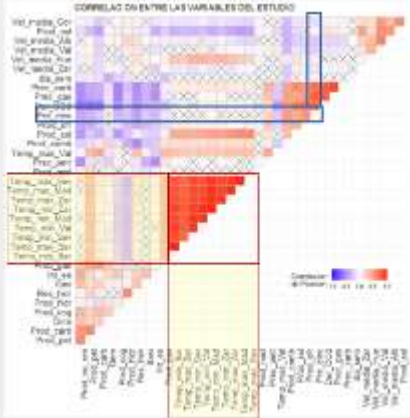
5. DATOS DE PARTIDA

- El **intervalo temporal** considerado ha sido de 10 años, de 2012 a 2022, exceptuando el año 2020 debido a la anomalía del Covid.
- Los datos obtenidos de cada fuente se han unido en un dataset y han sido explorados, y en algunos casos tratados, empleando la herramienta **RStudio**.

Datos faltantes (NA's)

- Las variables que presentaban pocos casos se ha seguido el criterio general de asignarles el valor del día anterior.
- Para las variables que presentan estos valores en días como sábados o domingos (**precio de las materias primas, derechos de emisión CO₂ o Ibex**), se ha tomado el criterio de darles el valor del viernes anterior.
- Para las **temperaturas de Barcelona** se han tomado los valores medios de años anteriores o posteriores según el caso.
- Para las **reservas hidráulicas** que posee 1 solo valor por semana, se ha adoptado la solución de asignar ese valor para todos los días de la semana.

Datos faltantes (NA's)

- El escenario óptimo es aquel en que todas las variables predictoras se correlacionen con la variable de salida, pero no entre sí. Esto en la práctica es una situación muy improbable.
 - El método empleado para estudiar la correlación entre variables ha sido el cálculo del **Coefficiente de correlación de Pearson**.
- 
- Existe una elevada correlación entre las variables de temperatura, por este motivo se ha tomado la decisión de unificar estas 10 variables en una sola.

Dataset final

Para trabajar en los modelos de predicción el **dataframe final** obtenido está formado por **28 variables** y **3652 registros**.

Tratamiento de los valores outliers

Los modelos de **Regresión lineal múltiple** y **K-vecinos más cercanos (KNN)** son sensibles a estos valores.

Se han llevado a cabo dos acciones sobre el dataset final:

- Eliminar estos valores
- Sustituirlos por un valor máximo y mínimo.

6. METODOS Y TECNICAS EMPLEADAS

Los modelos desarrollados se han basado en:

- Algoritmos de regresión lineal múltiple.
- Algoritmos k-vecinos más cercanos.
- Algoritmos de árbol de decisión.
- Algoritmos random forest.
- Algoritmos XG Boost.
- Redes neuronales.

6.1 DIVISION DE LOS DATOS

En este estudio se ha recurrido a realizar dos divisiones diferentes, en los conjuntos de datos “train” y “test”, para estudiar más factores que pueden afectar a los precios.

- 80% para el conjunto “train” y el 20% para el de “test”
- 70% para el conjunto “train” y el 30% para el de “test”

6.2 ESTANDARIZACION LOS DATOS

En algoritmos como la **regresión lineal**, **k-vecinos más cercanos** o las **redes neuronales**, las **variables predictoras numéricas** son necesarias normalizarlas o estandarizarlas.

En nuestro estudio se ha recurrido a la **normalización Z-score** que divide cada variable predictora entre su desviación típica después de haber sido centrada, de esta forma, los datos pasan a tener una distribución normal.

6.3 ESTANDARIZACIÓN DE LOS DATOS

La evaluación de los modelos ayuda a medir su **rendimiento**, es decir, **cuantificar la calidad de las predicciones que efectúa**.

Las métricas de evaluación empleadas en el TFM han sido las más comunes para los modelos de regresión:

- **Error medio absoluto (MAE)**
- **Error cuadrático medio (MSE)**
- **Raíz cuadrada del error cuadrático medio (RMSE)**
- **R cuadrado o coeficiente de determinación (R^2)**
- **Error en porcentaje medio absoluto (MAPE)**

7. ANALISIS DE LOS RESULTADOS OBTENIDOS

7.1 REGRESIÓN LINEAL MÚLTIPLE

Se han establecido **4 variantes** con los siguientes resultados:

	Div. Datos	MAE	MSE	RMSE	MAPE	R^2	Tratamiento de outliers
RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%	Eliminados
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%	Eliminados
RLM_3	80-20%	18.11	1176.19	34.30	39.79%	62.52%	Limitados
RLM_4	70-30%	18.70	1368.13	36.99	42.36%	59.70%	Limitados

Se concluye que los modelos de RLM pueden funcionar bien si se eliminan los valores anormales, y parece que la división del dataset no tienen mucha influencia.

7.2 K VECINOS MÁS CERCANOS

Se han establecido **4 variantes** con los siguientes resultados:

	Div. Datos	MAE	MSE	RMSE	MAPE	R^2	Tratamiento de outliers
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%	Eliminados
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%	Eliminados
KNN_3	80-20%	10.21	773.79	27.82	30.67%	75.35%	Limitados
KNN_4	70-30%	10.88	900.80	30.01	30.00%	73.46%	Limitados

Se concluye que los modelos de KNN pueden funcionar bien si se eliminan los valores anormales. Obtienen muy buenos resultados en precisión.

7.3 ÁRBOLES DE DECISIÓN

Se han establecido **2 variantes** con los siguientes resultados:

	Div. Datos	MAE	MSE	RMSE	MAPE	R^2
AD_1	80-20%	14.36	940.20	30.66	33.24%	70.04%
AD_2	70-30%	11.98	618.20	24.86	33.34%	66.35%

Se concluye en los dos casos son aceptables, pero poseen peores valores que los modelos anteriores.

7.4 RANDOM FOREST

Se han establecido **4 variantes** con los siguientes resultados:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
RF_1	80-20%	11.34	791.17	28.13	28.00%	74.79%
RF_2	70-30%	11.74	873.47	29.55	28.33%	74.27%
RF_1 optimizado	80-20%	9.78	809.44	28.45	26.20%	74.21%
RF_2 optimizado	70-30%	10.00	885.27	29.75	25.92%	73.92%

Las cuatro variantes creadas han mostrado una gran igualdad. Los valores obtenidos son aceptables, aunque sus resultados sean inferiores a otros. También se puede concluir que los modelos random forest trabajan bien en predicción, aun teniendo outliers en los datos.

7.5 XG BOOST

Se han establecido **4 variantes** con los siguientes resultados:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
XGB_2 optimizado	70-30%	12.82	924.92	30.41	29.76%	72.75%
XGB_1	80-20%	11.34	897.33	29.96	28.80%	71.41%
XGB_1 optimizado	80-20%	12.42	900.93	30.02	29.62%	71.29%
XGB_2	70-30%	11.52	1159.21	34.05	25.82%	65.85%

Se concluye que los modelos XG Boost trabajan bien en predicción, aun teniendo outliers en los datos.

7.6 RED NEURONAL

Se ha optado por crear numerosas variantes, modificando tanto el número de capas ocultas dentro de los modelos, como el número de neuronas que forman en cada una de esas capas

	Div. Datos	Nº capas	Nº neuronas	MAE	MSE	RMSE	MAPE	R ²
RN_14	70-30%	3	32	10.35	768.00	27.71	27.68%	77.38%
RN_16	70-30%	3	52	11.51	782.44	27.97	26.58%	76.95%
RN_1	80-20%	3	32	9.62	736.51	27.14	26.52%	76.53%
RN_15	70-30%	3	37	11.62	806.39	28.40	27.05%	76.25%
RN_17	70-30%	3	77	11.86	808.22	28.43	27.32%	76.19%
RN_20	70-30%	6	67	12.28	821.17	28.66	26.37%	75.81%
...
RN_23	70-30%	12	227	13.04	1047.02	32.36	27.31%	69.16%
RN_9	80-20%	6	187	13.01	992.06	31.50	27.31%	68.39%
RN_10	80-20%	12	227	15.20	1118.64	33.45	31.01%	64.36%
RN_11	80-20%	12	402	14.34	1138.28	33.74	27.16%	63.73%
RN_24	70-30%	12	402	14.79	1299.94	36.05	27.20%	61.71%
RN_26	70-30%	30	237	14.76	1402.08	37.44	29.22%	58.70%

De todas estas variantes creadas las **5 que mejor resultados han logrado están formadas por 3 capas** (1 de entrada, 1 oculta y 1 de salida), esto es, las más simples de todas las configuraciones. También se puede observar que, las redes con mayor número de capas y neuronas son las que peor se han comportado.

En cuanto a la división de los datos, claramente el **porcentaje 70-30% obtiene mejores resultados**, habiendo 6 de sus modelos, en las 7 primeras posiciones.

8. CONCLUSIONES

Los resultados obtenidos de todos los modelos creados han sido:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%
RN_14	70-30%	10.35	768.00	27.71	27.68%	77.38%
RN_16	70-30%	11.51	782.44	27.97	26.58%	76.95%
RN_1	80-20%	9.62	736.51	27.14	26.52%	76.53%
RN_15	70-30%	11.62	806.39	28.40	27.05%	76.25%
RN_17	70-30%	11.86	808.22	28.43	27.32%	76.19%
RN_20	70-30%	12.28	821.17	28.66	26.37%	75.81%
RN_18	70-30%	12.22	822.62	28.68	26.81%	75.77%
RN_2	80-20%	11.68	765.27	27.66	26.33%	75.62%
RN_19	70-30%	12.11	828.17	28.78	26.13%	75.60%
KNN_3	80-20%	10.21	773.79	27.82	30.67%	75.35%
RN_4	80-20%	11.15	778.53	27.90	25.55%	75.19%
RN_7	80-20%	11.93	781.40	27.95	28.11%	75.10%
RF_1	80-20%	11.34	791.17	28.13	28.00%	74.79%
RN_3	80-20%	11.80	791.64	28.14	25.86%	74.78%
RF_2	70-30%	11.74	873.47	29.55	28.33%	74.27%
RF_1 optimizado	80-20%	9.78	809.44	28.45	26.20%	74.21%
RN_8	80-20%	11.61	813.93	28.53	25.92%	74.07%
RF_2 optimizado	70-30%	10.00	885.27	29.75	25.92%	73.92%
RN_13	80-20%	11.87	824.83	28.72	26.38%	73.72%
RN_6	80-20%	11.66	828.58	28.79	27.31%	73.60%
KNN_4	70-30%	10.88	900.80	30.01	30.00%	73.46%
RN_21	70-30%	12.54	913.99	30.23	27.98%	73.08%
RN_5	80-20%	12.48	849.82	29.15	27.87%	72.92%
XGB_2 optimizado	70-30%	12.82	924.92	30.41	29.76%	72.75%
RN_12	80-20%	13.29	884.25	29.74	32.23%	71.83%
XGB_1	80-20%	11.34	897.33	29.96	28.80%	71.41%

RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%
XGB_1 optimizado	80-20%	12.42	900.93	30.02	29.62%	71.29%
AD_1	80-20%	14.36	940.20	30.66	33.24%	70.04%
RN_22	70-30%	14.43	1034.18	32.16	26.55%	69.54%
RN_25	70-30%	13.14	1037.80	32.21	26.64%	69.43%
RN_23	70-30%	13.04	1047.02	32.36	27.31%	69.16%
RN_9	80-20%	13.01	992.06	31.50	27.31%	68.39%
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%
AD_2	70-30%	11.98	618.20	24.86	33.34%	66.35%
XGB_2	70-30%	11.52	1159.21	34.05	25.82%	65.85%
RN_10	80-20%	15.20	1118.64	33.45	31.01%	64.36%
RN_11	80-20%	14.34	1138.28	33.74	27.16%	63.73%
RLM_3	80-20%	18.11	1176.19	34.30	39.79%	62.52%
RN_24	70-30%	14.79	1299.94	36.05	27.20%	61.71%
RLM_4	70-30%	18.70	1368.13	36.99	42.36%	59.70%
RN_26	70-30%	14.76	1402.08	37.44	29.22%	58.70%

La clasificación nos revela que los mejores modelos para predecir el precio de la electricidad en España son los k vecinos más cercanos, en los casos en que los outliers han sido eliminados.

El siguiente modelo ha sido claramente las redes neuronales, y dentro de ellas las más simples (las que menos capas y neuronas tienen). Estas han tenido muy buenos valores de predicción, aunque a lo que se refiere a las métricas de errores, poseen valores más elevados que las dos mejores variantes de los modelos de KNN.

Posteriormente están los algoritmos random forest, aunque no alcanzan los mejores resultados en precisión de las redes neuronales.

El resto de modelos, aunque tienen valores que pueden ser aceptables están claramente por debajo de los ya mencionados.

COSTE COMPUTACIONAL

Los modelos más rápidos son los algoritmos de regresión lineal múltiple, seguidos de los árboles de decisión, que invierten en los dos casos menos de medio segundo en realizar los cálculos.

Los k vecinos más cercanos se mueven en una horquilla de entre 1 y 4 segundos.

Las redes neuronales aumentan considerablemente el tiempo tardando entre los 24 segundos y los 5 minutos.

En cuanto los modelos random forest, los simples emplean 25 segundos, pero los optimizados llegan a tener un coste de 8 minutos.

Los modelos basados en XG Boost poseen un coste no muy excesivo presentando una media de 1,7 minutos.

Estos tiempos no son demasiado dramáticos debido a que se ha trabajado con un dataset con poco más de 100.000 datos, pero cuando el dataset tenga millones de datos estos tiempos pueden ser un factor limitante.

8.1 CONCLUSION FINAL

Se ha logrado el objetivo marcado de desarrollar algoritmos capaces de predecir los precios de la energía eléctrica en el mercado español, con valores de error (MAPE) dentro de los varemos establecidos en estudios especializados en la materia.

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%
RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%

Estos modelos han sido 2 algoritmos basados en k vecinos más cercanos y otros 2 basados en regresión lineal múltiple, habiéndose eliminado en los 4 casos los valores anormales, presentando los primeros un nivel de precisión muy alto y los segundos, aceptable.