IMF Business School

MÁSTER EN DATA SCIENCE Y BUSINESS ANALYTICS ONLINE

RESUMEN EJECUTIVO:
“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** que le afectan entre los que están:

- El equilibrio que debe existir entre la generación y la demanda en cada instante, ya que la energía eléctrica no puede ser almacenada en grandes cantidades.
- Las fuentes de producción de energía eléctrica son muy diversas. Se emplean en España hasta 20 tipos diferentes.
- La transición que se está produciendo en los últimos años, de fuentes de energía altamente contaminantes a fuentes renovables.
- Estas fuentes de energía renovables dependen de las condiciones meteorológicas.
- Cada tecnología de producción tiene un coste de producción asociado, ya sea debido a las infraestructuras necesarias o para la propia producción de la energía.
- Existe un coste ambiental como son los derechos de emisión de CO_2.
- La legislación actual del sector aumenta la complejidad en el sector.

Todo esto hace que **la predicción del precio de la electricidad se convierta 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 hoy en día, herramientas de Deep Learnin para predecir el precio de la electricidad a largo plazo.

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

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.

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

Como es lógico e ineludible, hay que emplear los conocimientos adquiridos en el **MODULO I**, sobre los lenguajes de programación **Python** y **R**. Se ha emplearemos R para realizar los trabajos para el procesado de los datos y Python para desarrollar los modelos de Machine Learning.

Los **MODULOS II y II** correspondientes al "**Impacto y valor del Big Data**" y a "**La Ciencia de Datos. Técnicas de análisis, minería y visualización**" respectivamente, son básicos para entender los módulos posteriores y para adquirir los conceptos necesarios para proceder correctamente en la obtención, limpieza, transformación y visualización de los datos.

El **MODULO V: Estadística para Científicos de Datos** es imprescindible para la comprensión del análisis estadístico de los datos en el que se basan los algoritmos de Machine Learning.

El más importante es el **MODULO VI: Aprendizaje automático,** en el se desarrollan los **algoritmos de Aprendizaje Automático**, como son los **problemas de regresión,** y las técnicas de **Deep Learning** como son las **redes neuronales.**

Finalmente, el **MODULO VII** también está relacionado ya que en él se abordan las **técnicas para la toma de decisiones**, y más concretamente lo referente al **Aprendizaje Supervisado** en el que se tratan los **algoritmos para problemas de regresión** y las **redes neuronales.**

3. SISTEMA ELECTRICO ESPAÑOL

3.1 ETAPAS DEL SISTEMA ELÉCTRICO ESPAÑOL

Para que la EE llegue al consumidor existen diferentes fases o etapas que son:

- Centrales eléctricas generadoras.
- Estaciones transformadoras elevadoras.
- Redes de transporte.
- Subestaciones transformadoras reductoras.
- Redes de distribución.
- Centros de transformación.

3.2 TECNOLOGÍAS DE PRODUCCIÓN DE EE

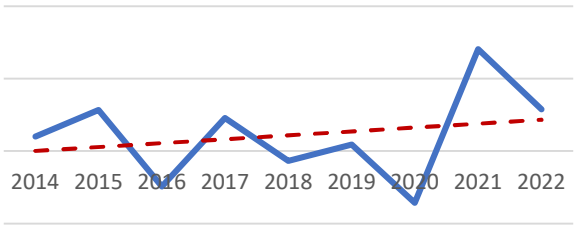
En España se emplean más de 20 tecnologías diferentes para la producción de EE. Estas son:

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

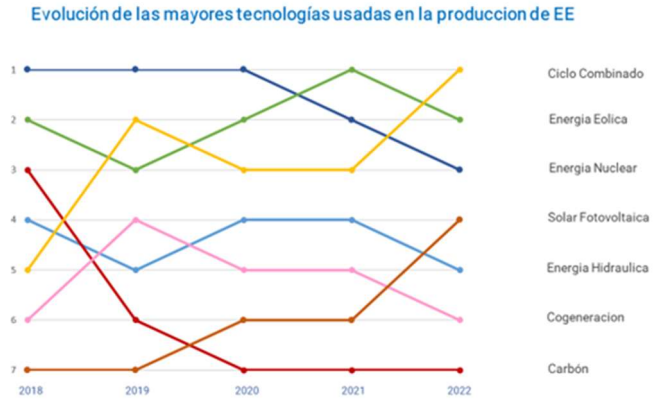
GENERACIÓN NO RENOVABLE	
Turbinación bombeo	Residuos no renovables
Nuclear	Motores diésel
Ciclo combinado	Turbina de vapor
Carbón	Turbina de gas
Cogeneración	Fuel + Gas

3.3 EVOLUCIÓN DE LA PRODUCCIÓN DE EE

La tendencia del consumo de EE en España es ascendente, como se observa en el grafico.



Cada vez es más elevado el empleo de tecnologías renovables.



Las tecnologías más empleadas para la producción son (en orden descendiente en el año 2022): ciclo combinado, producción eólica, producción nuclear, producción fotovoltaica, producción hidráulica, cogeneración y central térmica de carbón.

3.4 AGENTES DEL MERCADO

Los **agentes del mercado** son todas las personas físicas o jurídicas que intervenga en las transacciones económicas que tengan lugar en el mercado de producción de energía eléctrica, comprando o vendiendo electricidad. Según la actual Ley del Sector Eléctrico (LSE) se establecen 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 MERCADO ELECTRICO ESPAÑOL

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).

El mercado de electricidad en España, al igual que en otros países, se organiza en una secuencia de mercados en los que generación y demanda intercambian energía y reservas para distintos plazos.

En los **mercados a plazo** los agentes intercambian contratos con períodos de entrega de distinta duración (anual, trimestral, mensual, etc.), con una antelación de días, semanas, meses e incluso años antes del momento en que la energía sea generada y consumida.

3.6 FORMACIÓN DEL PRECIO EN EL MERCADO MAYORISTA A PLAZO

Los mercados a plazo sirven para que los agentes del mercado puedan realizar sus planes económicos evitando altos riesgos de perdidas, 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)**.**

De acuerdo a la teoría económica, el precio esperado del mercado diario es el coste de oportunidad de los contratos a plazo, por lo que el precio del mercado a plazo refleja el precio del mercado diario esperado a futuro.

4. BUSSINES CASE

La dependencia de la sociedad actual sobre la electricidad, los factores que influyen en su producción y consumo, los costes de las materias primas, los factores climatológicos o la legislación vigente del sector eléctrico, son solo unos pocos factores que afectan en el precio final de la EE.

Estos factores hacen del sector eléctrico un mercado altamente complejo en el que, tal y como indican 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.

La **aplicación principal** de este trabajo es el de **facilitar diferentes herramientas de predicción de precios** que pueden ser útiles para **Productores, Comercializadores y Consumidores**.

4.1 ESTADO DEL CONOCIMIENTO

Según señala Weron, R. (2014), durante los últimos 23 años se han probado una variedad de métodos e ideas para la previsión del precio de la electricidad, con diversos grados de éxito.

A nivel corporativo, **las previsiones de precios de la electricidad se han convertido en un costo fundamental para los mecanismos de toma de decisiones de las empresas energéticas**.

Las previsiones de precios desde unas pocas horas hasta algunos meses se han vuelto de particular interés para los administradores de carteras de energía. Un generador, una empresa de servicios públicos o un gran consumidor industrial que sea capaz de pronosticar los precios mayoristas volátiles con un nivel razonable de precisión puede ajustar su estrategia de licitación y su propio programa de producción o consumo para reducir el riesgo o maximizar las ganancias en el día siguiente.

Muchos de los enfoques actuales, de modelado y pronóstico de precios, son soluciones híbridas, que combinan diferentes técnicas de predicción.

Entre los modelos mas utilizados en los últimos años están:

- **Multi-agente** (simulación multi-agente, equilibrio y teoría de juegos): modelos que simulan un sistema de agentes que interaccionan unos con otros, y construyen el proceso de precios haciendo coincidir la oferta y la demanda en el mercado.
- **Fundamentales o estructurales**: describen la evolución del precio a través de factores físicos y económicos.
- **Reducidos** (cuantitativos, estocásticos): utilizan parámetros estadísticos distribuidos en el tiempo. Este tipo de modelos no buscan tanto predecir el precio

con precisión, sino replicar el comportamiento del precio y su correlación con el precio de otros.

- **Estadísticos** (econométricos, análisis técnico): aplican de forma directa técnicas estadísticas para la previsión del mercado o de la demanda de electricidad en base a datos históricos.
- **Inteligencia computacional** (estadística no lineal): las técnicas aplicadas en estos modelos combinan la capacidad de aprendizaje, la evolución de los datos y la aparente falta de relaciones de los mismos para generar sistemas complejos capaces de adaptarse.

Desde hace pocos años se están desarrollando investigaciones que aplican los modelos tipo **árboles dinámicos**, cuya aplicación en el mercado eléctrico español ha sido escasa hasta el momento, y se presenta como una herramienta con gran potencial, para tener en consideración junto con las técnicas basadas en series temporales que se usan en la actualidad.

4.2 OBJETIVOS A ALCANZAR

El **objetivo principal de este Trabajo Fin de Máster 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.**

Se pretende alcanzar este objetivo partiendo de una serie de **variables**, como son los **precios la las materias primas** que se emplean para la producción de electricidad (gasóleo, gas natural y carbón), la **meteorología** (temperatura, viento y agua embalsamada), **producción diaria por tipo de tecnología**, la **demanda eléctrica** y otros factores como los **derechos de emisión de CO_2** y la situación económica del país con el índice **Ibex-35.** Para ello se han empleando técnicas de **análisis de datos**, **aprendizaje automático** y **Deep Learning** (redes neuronales**)**.

Del articulo de investigación realizado por González C., Mira-McWilliams J. y Juárez I. (2015) Evaluación de variables importantes y pronóstico del precio de la electricidad basado en modelos de árboles de regresión: clasificación y árboles de regresión, Bagging y Random Forests. IET Generation, Transmission & Distribution, 9(11), 1120–1128, se

pueden extraer los **errores medios absolutos en porcentaje (MAPE) en las predicciones de los precios de la electricidad de diversos algoritmos estudiados.**

Según este articulo las técnicas actuales obtienen un error que está entre un 5,76 y un 11%. Por este motivo se considerará **como un buen resultado para los modelos estudiados en este TFM, los valores de MAPE que estén por debajo de ese 11% y posean un porcentaje de precisión alto**.

El presente TFM se han decidido aplicar las metodologías impartidas en el Master como son: **regresión lineal múltiple**, **k vecinos más cercanos**, **árboles de decisión**, **random forest**, **XG Boost** y, por supuesto, las **redes neuronales**.

5. DATOS DE PARTIDA

5.1 INTERVALO TEMPORAL

Se ha establecido como criterio tomar un **horizonte temporal de 10 años**, entre los años **2012 y 2022**, pero para que el estudio no se vea distorsionado por la **anomalía que supuso la pandemia del Covid-19**, **se ha excluido el año 2020 de dicho intervalo**.

5.2 VARIABLES CONSIDERADAS

La **variable dependiente** del estudio serán los datos a predecir, así que corresponde al **precio medio diario de la EE en España**.

Para establecer las **variables independientes** empleadas se ha partido de los **factores que influyen sobre la variable objetivo.** De este modo tendremos las siguientes variables predictoras.

FACTOR	CODIGO	NOMBRE	UNIDADES DE MEDIDA
VARIABLE DEPENDIENTE			
	Pre_elec	Precio medio diario de la EE en España	€/MWh
VARIABLES PREDICTORAS			
Temporal	Dia_sem	Dia de la semana	
Demanda	Dem	Demanda media diaria	MWh

Precios materias primas	Prec_petr	Precio del petroleo	$/barril de Brent
	Prec_gas	Precio del gas natural	£/MillionBtu
	Prec_carb	Precio del carbon	$/tonelada
Producción EE según tecnología	Prod_eol	Producción en parques eólicos	GWh
	Prod_sol	Producción en parques solares	GWh
	Prod_hidr	Producción hidráulica y turbinación por bombeo	GWh
	Prod_ofr	Producción de otras fuentes renovables	GWh
	Prod_nucl	Producción en centrales nucleares	GWh
	Prod_pet	Producción en centrales convencionales de gasóle	GWh
	Prod_gas	Producción en centrales con turbina de gas	GWh
	Prod_carb	Producción en centrales convencionales de carbór	GWh
	Prod_comb	Producción en centrales ciclo combinado	GWh
	Prod_cog	Producción en centrales de cogeneración	GWh
	Prod_noren	Producción de otras fuentes no renovables	GWh
Meteorología	Temp	Temperatura	ºC
	Vel_Vien	Velocidad del viento	m/s
	Res_hidr	Reservas hidraulicas	%
Otros factores	Der_CO2	Precio Derechos de emisión de CO_2	€/tonelada CO2
	Ibex	Situacion socio-económica del pais	Puntos bursátiles
	Int_ee	Intercambio de EE con otros países	GWh

5.3 FORMACION DEL DATASET Y TRATAMIENTO DE LOS DATOS

Lo primero que se ha realizado es **unir todas las variables en un solo dataset**. Para lo cual, se han limpiado los diferentes dataframes originales obtenidos de las fuentes de los datos, y nos hemos quedado solo con los datos necesarios, que no son otros que la fecha (para poder realizar las uniones) y los datos de cada una de las variables.

Una especial mención requiere la variable "**Res_hidr**". Esta variable posee 48.205 registros debido a que existen 100 embalses productores de EE y cada uno posee aproximadamente un registro por semana. Introducir esta variable así no seria viable por lo que tenemos que transformarla para obtener un dato por día. La solución adoptada para ello ha sido obtener el porcentaje de llenado de todos los embalses por semana asignando este valor a todos los días de la semana.

Valores faltantes

Resumen de **datos faltantes (NA´s) en las variables del dataset formado**.

```
##         fecha       dia_sem      Pre_elec           Dem     Prec_petr
##             0             0             0             1          1022
##      Prec_gas     Prec_carb      Prod_eol     Prod_sol     Prod_hidr
##          1111          1094             0             0             0
##      Prod_ofr     Prod_nucl     Prod_pet     Prod_gas     Prod_carb
##             0             0             0             0             0
##     Prod_comb      Prod_cog  Prod_no_ren Temp_min_Mad Temp_max_Mad
##             0             0             0            15            15
##  Temp_min_Bar Temp_max_Bar Temp_min_Val Temp_max_Val Temp_min_Sev
##           426           425             0             8            15
##  Temp_max_Sev Temp_min_Zar Temp_max_Zar Vel_media_Val Vel_media_Alb
##            13             0             0             5            16
## Vel_media_Zar Vel_media_Cor Vel_media_Hue      Res_hidr       Der_CO2
##             1            42             9          3183          1063
##          Ibex        Int_ee
##          1093             0
```

Para las variables que presentan pocos valores faltantes se ha adoptado el criterio general de asignarles el valor del día anterior.

En las variables de los precios de las materias primas, los derechos de emisión de CO_2 o el índice Ibex-35 los datos faltantes se corresponden a los sábados y/o domingos, debido a que esos días no se celebran los mercados donde se negocian sus valores. La solución adoptada ha sido asignar a estos días el valor del viernes anterior.

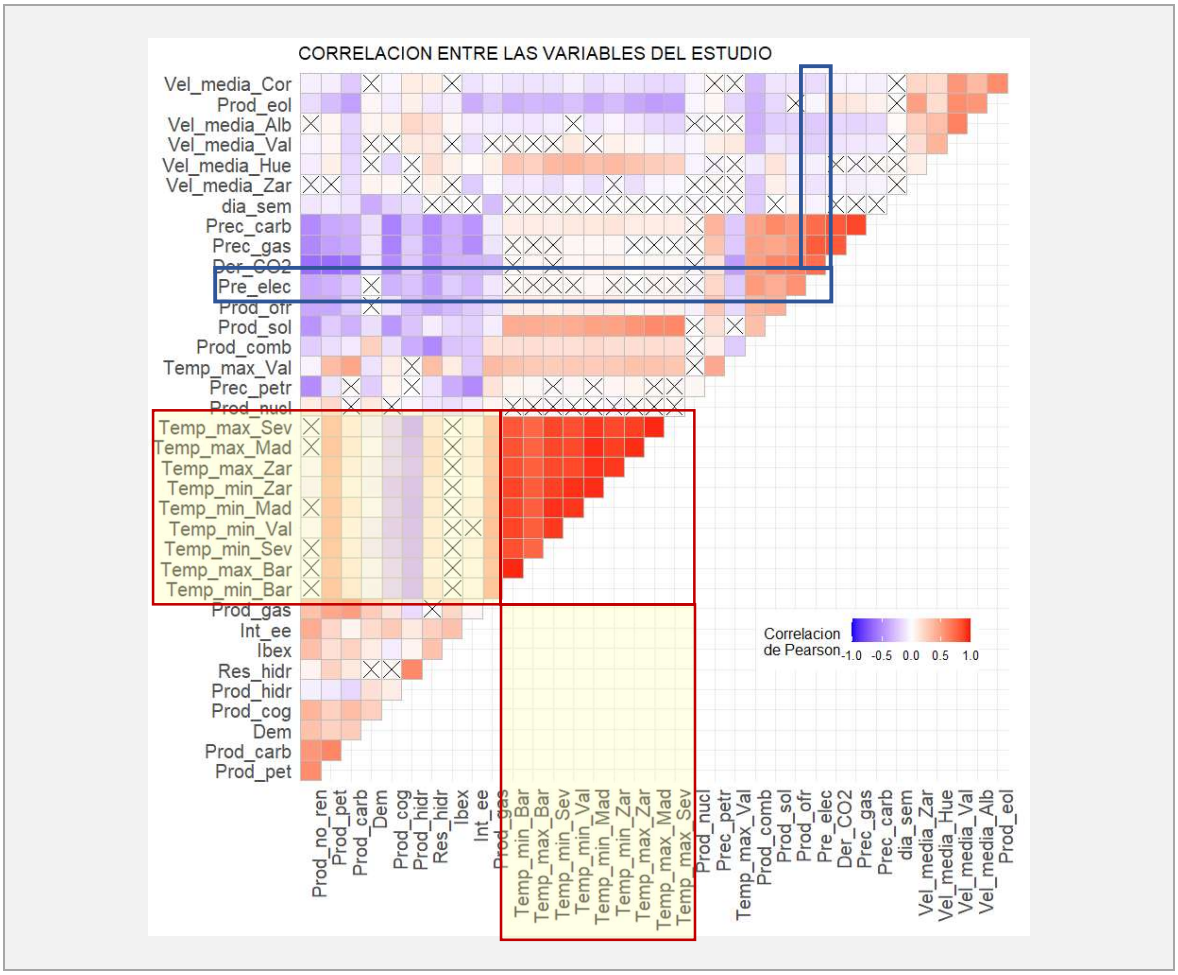
En el caso de la "**Temp_min_Bar**" y "**Temp_max_Bar**" existen valores faltantes debido a que la fuente de datos no los proporciona al hacer la consulta. Existen sobre todo en tres periodos claramente marcados. La solución adoptada ha sido la de tomar para el primer tramo la media de los valores de los dos años posteriores, para el segundo tramo la media de los valores del año anterior y posterior, y para el ultimo tramo la media de los valores de los dos años anteriores.

La variable "**Res_hidr**" posee numerosos datos faltantes debido a que los registros obtenidos son semanales, y solo posee 1 para cada semana. Ya que la variación de los recursos hídricos varia lentamente, la solución adoptada es la de dar el mismo valor a todos los días de la semana.

Correlación entre variables

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 practica es una situación muy improbable. Una solución a este problema sería excluir aquellas variables que presenten una correlación notablemente alta, aunque también se puede incurrir en el error de prescindir de factores importantes en la predicción a realizar.

El método empleado para estudiar la correlación entre variables ha sido el calculo del **Coeficiente de correlación de Pearson**. Si el valor de correlación entre dos variables es superior a 0,95 puntos se puede considerar que esas dos variables están correlacionadas.



Según el cuadro de Correlacion se observa una gran correlacion entre las variables de temperaturas maximás y minimás de las 5 ciudades tomadas. Por este motivo **se ha decidido reducir estas variables en una sola**.

Para esta nueva variable, se asignará en los meses más cálidos, entre abril y septiembre, la temperatura media de las temperaturas máximas de esos meses de las 5 ciudades, y para los otros 6 meses del año, la temperatura media de las temperaturas mínimas.

Dataset final

Con esta ultima sustitución de las variables de temperaturas, por la nueva variable de temperatura obtenemos el **dataframe final,** formado por **28 variables** y **3652 registros**, con el que se va a trabajar en los modelos de predicción.

Tratamiento de los valores outliers

Los **valores extremos** que nos encontramos en nuestro dataframe final pueden influir sobre el funcionamiento de algunos algoritmos. De esta forma sabemos que los modelos de **Regresión lineal múltiple** y **K-vecinos más cercanos (KNN)** son sensibles a estos valores, por lo que habrá que realizar alguna acción sobre ellos.

En este TFM han llevado a cabo **dos acciones** diferentes para conocer cual es la más efectiva han sido:

- **Eliminarlos.**
- **Sustituirlos por valores máximos y mínimos calculados por el método intercuartílico.**

Este método intercuartílico consiste en limitar los valores, tanto por la parte superior como por la inferior, a los valores que superen **± 1,5 veces el rango intercuartílico**, que corresponde a la diferencia entre tercer cuartil y el primero.

6. METODOS Y TECNICAS EMPLEADAS

Como se ha señalado el objetivo de este estudio es el de analizar distintos **modelos de regresión** para comprobar cual es el mejor a la hora de predecir el precio de la EE en España. Los modelos empleados en este trabajo son:

- Modelo de regresión lineal múltiple.
- Modelo k-vecinos más cercanos (KNN).

- Modelo árbol de decisión.
- Modelo Random Forest.
- Modelo XG Boost.
- Modelo red neuronal.

Antes de meternos en el estudio de los modelos, hay que exponer una serie de acciones que hay que llevar a cabo en los datasets empleados. Estas acciones corresponden a los apartados siguientes.

6.1 DIVISION DE LOS DATOS PARA LOS MODELOS

Para poder entrenar y validar los modelos, los datos del dataset deben ser divididos en dos conjuntos, uno llamado de "entrenamiento" o "**train**" y otro de "prueba" o "**test**".

En este estudio se ha recurrido a realizar dos divisiones diferentes, en los conjuntos de datos "train" y "test", para estudiar mas factores que pueden afectar a los precios. Según esto se han establecido las dos siguientes divisiones:

- 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 ESTANDARIZACIÓN DE LOS DATOS

Cuando las **variables predictoras** son numéricas, la escala en la que se miden puede influir en el algoritmo como los basados en **regresión lineal**, **k-vecinos más cercanos** o las **redes neuronales**.

Para evitar esta circunstancia en este tipo de modelos se ha recurrido a la técnica de **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 METRICAS DE VALIDACION DE LOS MODELOS

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, y son:

- **Error medio absoluto (MAE)**: representa la media de la diferencia absoluta entre los valores reales y predichos. Mide el promedio de los residuos.
- **Error cuadrático medio (MSE)**: es una de las métricas más utilizadas en regresión y corresponde a la media de las diferencias entre el valor real y el predicho, elevado al cuadrado. Mide la varianza de los residuos.
- **Raíz cuadrada del error cuadrático medio (RMSE)**: corresponde a la raíz cuadrada de la métrica anterior y mide la desviación estándar de los residuos. La ventaja de esta métrica es que presenta el error en las mismas unidades que la variable objetivo, lo que la hace más fácil de entender.
- **R cuadrado** o **coeficiente de determinación (R^2)**: esta métrica determina la calidad del modelo para replicar los resultados, y la proporción de los resultados que puede explicarse por el modelo.
- **Error en porcentaje medio absoluto (MAPE)**: es una medida relativa que escala esencialmente el MAE para que se muestre en unidades de porcentaje.

En todos los modelos de este estudio se ha obtenido estas cinco métricas, para poder compararlos y poder conocer cuales son los mejores.

7. ANALISIS DE LOS RESULTADOS OBTENIDOS

Una vez desarrollados y entrenados los algoritmos en Python, con los condicionantes anteriormente descritos, pasamos a analizar los resultados de cada uno.

7.1 REGRESION LINEAL MULTIPLE

Para este tipo de algoritmo se han establecido **4 variantes**, que han sido obtenidas al emplear dos divisiones con diferentes proporciones para los datos de "train" y de "test" y como resultado de aplicar el dataset con los valores anormales eliminados o con ellos limitados. Los resultados obtenidos con el conjunto de datos "test" son:

	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

Las dos primeras variantes presentan unos valores muy buenos de error y unos valores aceptables de precisión para un algoritmo de predicción de precios.

Se puede concluir que los modelos de regresión lineal múltiple pueden funcionar bien si se eliminan los valores anormales, y parece que la división del dataset para datos de "entrenamiento" y "test" no tienen mucha influencia.

7.2 K VECINOS MÁS CERCANOS

Este caso es igual que el anterior obteniéndose **4 variantes**. Los resultados obtenidos con el conjunto de datos "test" son:

	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

Este modelo posee las mismas conclusiones que el anterior, con la diferencia de que el porcentaje de precisión aumenta considerablemente.

7.3 ARBOLES DE DECISION

En este tipo de algoritmos solo tendremos 2 variantes debido a que no son sensibles a los valores anormales y por lo tanto solo se ha dependiendo de la proporción en la división de los datos.

	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%

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

7.4 RANDOM FOREST

Como en todos los modelos, se han obtenido 2 variantes derivadas al aplicar diferentes proporciones en la división de los datos para formar los conjuntos de “entrenamiento” y de “test” y a parte de esto, se ha intentado optimizar esas dos variantes a través de validación cruzada, por lo que finalmente tenemos **4 variantes** para este modelo.

	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 por estos cuatro modelos son aceptables, aunque sus resultados sean inferiores a otros. Se puede concluir**, también, **que los modelos random forest trabajan bien en predicción, aun teniendo outliers en los datos**.

7.5 XG BOOST

Este caso es igual que el anterior obteniéndose **4 variantes** para el modelo, derivadas por las mismas razones de división de datos y optimización de los modelos. Los resultados obtenidos son:

	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

En el caso de las **redes neuronales** 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, para ver como funcionan diferentes arquitecturas. Todas estas variantes están duplicadas al haber empleado las dos proporciones diferentes, señaladas en los modelos anteriores, para dividir los datos de "entrenamiento" y "test".

	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_18	**70-30%**	6	**102**	12.22	822.62	28.68	**26.81%**	**75.77%**
RN_2	**80-20%**	3	**37**	11.68	765.27	27.66	**26.33%**	**75.62%**

RN_19	**70-30%**	6	**102**	12.11	828.17	28.78	**26.13%**	**75.60%**
RN_4	**80-20%**	3	**77**	11.15	778.53	27.90	**25.55%**	**75.19%**
RN_7	**80-20%**	6	**67**	11.93	781.40	27.95	**28.11%**	**75.10%**
RN_3	**80-20%**	3	**52**	11.80	791.64	28.14	**25.86%**	**74.78%**
RN_8	**80-20%**	6	**107**	11.61	813.93	28.53	**25.92%**	**74.07%**
RN_13	**80-20%**	30	**237**	11.87	824.83	28.72	**26.38%**	**73.72%**
RN_6	**80-20%**	6	**102**	11.66	828.58	28.79	**27.31%**	**73.60%**
RN_21	**70-30%**	6	**107**	12.54	913.99	30.23	**27.98%**	**73.08%**
RN_5	**80-20%**	6	**102**	12.48	849.82	29.15	**27.87%**	**72.92%**
RN_12	**80-20%**	12	**142**	13.29	884.25	29.74	**32.23%**	**71.83%**
RN_22	**70-30%**	6	**187**	14.43	1034.18	32.16	**26.55%**	**69.54%**
RN_25	**70-30%**	12	**142**	13.14	1037.80	32.21	**26.64%**	**69.43%**
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, presentando una gran diferencia en la precisión con los mejores modelos.

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.

De los datos expuestos en la tabla anterior, **se puede concluir que los modelos de redes neuronales varían mucho según la estructura adoptada de capas y neuronas**, pero que, una vez encontrada una estructura adecuada para los datos, **obtienen buenos resultados con alta precisión y un valor no muy alto de RMSE**.

8. CONCLUSIONES

Se han realizado numerosos modelos modificando diferentes aspectos en ellos y los resultados son los siguientes:

	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%

Esta 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 que mejor se ha comportado con los datos disponibles ha sido claramente las **redes neuronales**, y dentro de ellas **las más simples**, es decir, 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.

Las variantes de los algoritmos **random forest** también **han presentado buenos precisiones**, aunque ligeramente inferiores a las redes neuronales. Sus niveles en los errores son muy similares a los de estas ultimas.

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

Según la **métrica MAPE**, que recordamos que es la métrica empleada en el articulo de investigación realizado por González C., Mira-McWilliams J. y Juárez I. (2015) y que oscilan entre los **5,76** y los **11%**, podemos comprobar que solo **4 de los 44 modelos** creados en este estudio **poseen un MAPE acorde a esos valores, siendo estas las 2 mejores variantes de KNN y las 2 mejores de regresión lineal múltiple.**

	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%

De entre estos 4 modelos claramente los mejores son los KNN ya que poseen una precisión más alta que los RLM.

Otro factor, que en algunas circunstancias puede ser fundamental a la hora de emplear un algoritmo u otro, es el **coste computacional** que requiere.

Se ha comprobado que **los más rápidos son los algoritmos de regresión lineal múltiple**, seguidos de los **arboles 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 emplean entre los **24 segundos y los 5 minutos**.

En cuanto los modelos **random forest**, los simples tardan apenas **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.

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.