UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

Aplicación de técnicas Boosting al modelado y predicción de la serie de precios de la energía eléctrica

Madrid, febrero 2018

TUTORES:

Jose Manuel Mira McWilliams

María Camino González Fernández

AUTORA:

Sandra Blasco Santos

It is never too late to be

what you might have been.

AGRADECIMIENTOS

Mientras escribo me doy cuenta de que finaliza un largo camino, por el cual he luchado mucho. Conseguir llegar a la meta no ha sido fácil, ha sido una trayectoria llena de baches y altibajos que me han servido para crecer y formarme como persona. Pero si he podido llegar hasta aquí, ha sido no sólo gracias a mi esfuerzo, si no a la ayuda de la gente que me rodea, los cuales sería muy egoísta pasar por alto y no agradecer.

Quiero agradecer a aquellos profesores que gracias a su verdadera vocación por la docencia nos han tendido sus manos, así como todas las herramientas posibles para poder aprender y enfrentarnos a los problemas que se nos pongan en adelante. En especial, quiero dar las gracias a los directores de este proyecto, Camino González y José Mira, por apoyarme y guiarme a lo largo de la recta final, y no dejar en ningún momento de alentarme a seguir adelante. Vuestras palabras tranquilizadoras y vuestro optimismo y confianza han hecho que los malabarismos para compaginar el día a día hayan sido más sencillos.

Gracias, muchas gracias a mi familia. Gracias por sufrir conmigo cuando ha habido decepciones, por alegraros con las buenas noticias y por seguir a mi lado. Gracias por hacer vuestros mis sacrificios en algunas ocasiones, y por ver modificados los planes. Mis triunfos siempre serán también vuestros.

Mis amigos, los de toda la vida, los que pasen los años que pasen siguen ahí, me mantienen y me devuelven la perspectiva. A los que están día a día y a los que no pero que espero que lo estén celebrando conmigo.

A los amigos que me he encontrado a lo largo de estos años. Si hay algo que ha merecido la pena durante el tiempo que he pasado dentro de estas aulas, sois vosotros, tanto los de los primeros años como los de los últimos. Gracias por la comprensión, por el apoyo, por las risas. Gracias por ser rocas a las que agarrarme y ganchos que me levantaban siempre que me he caído. Sin vuestra presencia en mi vida todo habría sido mil veces más complicado. Si hay un tesoro que me llevo y del que estoy orgullosa, es el haberos conocido y tener el privilegio de contaros como parte de mi vida como una segunda familia.

Gracias por todo, os quiero.

ÍNDICE DE CONTENIDO

1- RESUMEN EJECUTIVO

Debido a la fluctuación en el precio de la electricidad española, la capacidad de predecir con exactitud los gastos que se van a realizar cada mes debidos al coste de este tipo de energía resulta una tarea complicada, lo que puede llegar a afectar económicamente en gran medida a la sociedad. A continuación, se muestra un gráfico con las variaciones horarias en el precio de la electricidad durante la primera semana del mes de marzo de 2015.

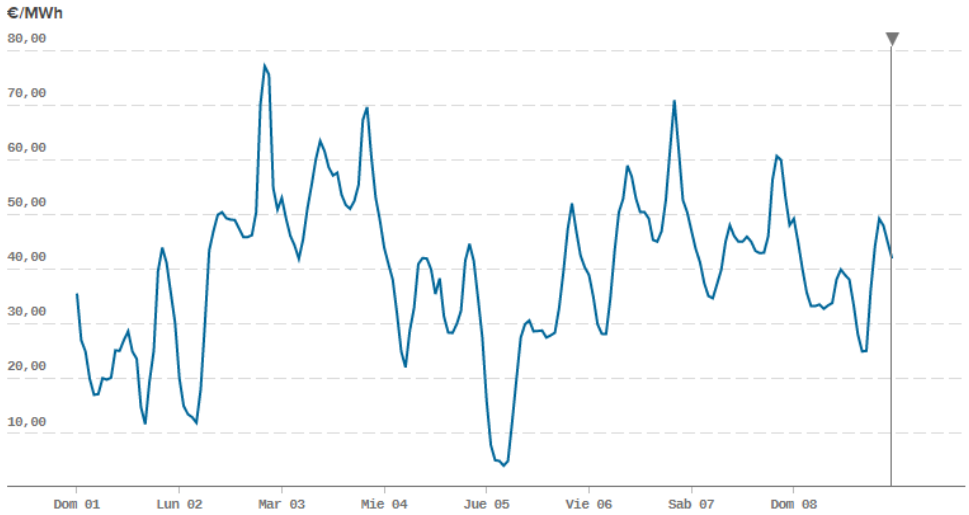

Figura 1: Variación horaria del precio de la electricidad en la primera semana del mes de marzo de 2015.

Pese a la variabilidad debida a factores externos que se puede encontrar en el precio a lo largo del tiempo, se puede encontrar estacionalidad a largo plazo, tal y como se puede ver en la siguiente imagen, que muestra el precio horario durante todo el año 2015.

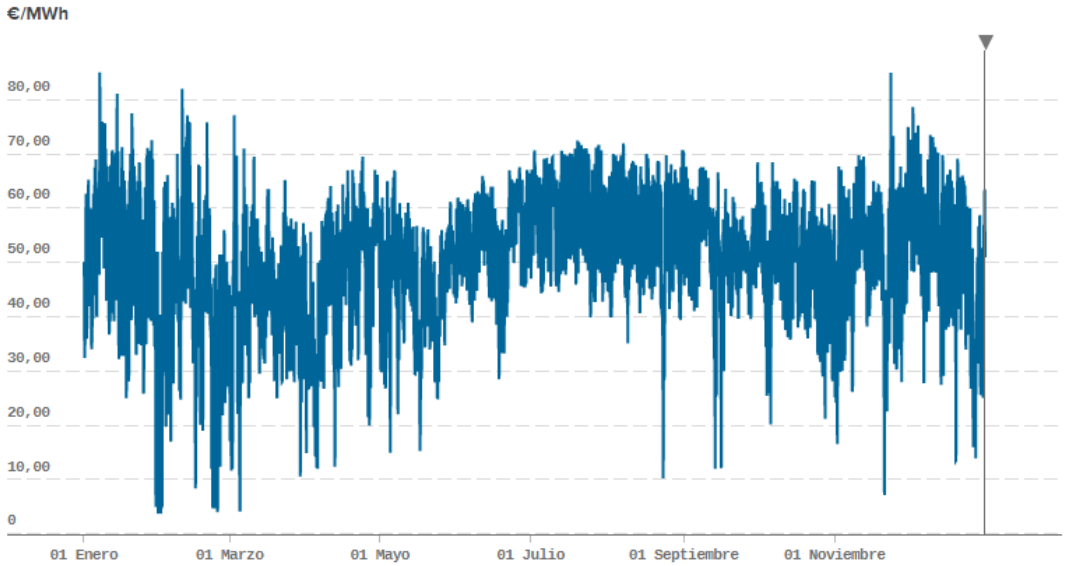

Figura 2: Precio de la energía eléctrica durante el 2015.

En este proyecto se realiza un análisis en el que se estudia el peso de determinados factores externos en la variación del precio de la electricidad. Entre las ventajas de conocer las variables que influyen en el cambio del precio de la electricidad, se encuentra la posibilidad de actuar sobre ellas para estabilizar los precios, o al menos predecirlos con mínimo error.

Otro de los objetos de estudio es el de realizar una predicción estadística lo más ajustada posible, de modo que la predicción realizada se iguale lo máximo posible al precio real de la electricidad, a través de diferentes algoritmos. Conseguir dicha predicción tiene la virtud de permitir a los usuarios tener una mayor posibilidad de planificación de las diferentes actividades, con lo que se optimizan los costes ocasionados.

Para realizar este tipo de estudios, se utilizan diferentes herramientas estadísticas. Con el paso del tiempo, han aumentado la cantidad de técnicas y herramientas capaces de analizar grandes volúmenes de datos. Estos nuevos desarrollos hacen que merezca la pena ser analizados, así como comparados con los resultados obtenidos a través de las técnicas convencionales. Éste es el caso principal por el que se decide utilizar nuevas técnicas en el proyecto. La mejor solución actual para resolver este tipo de problemas cuando el volumen de datos utilizados es muy elevado se trata de procedimientos especiales con una mayor capacidad de análisis e indicados para ello, como es el caso de la Minería de Datos o *Data Mining*.

Dentro del campo de la estadística perteneciendo al *Data Mining*, existen diversas técnicas capaces de realizar los distintos estudios, como son las redes neuronales, clustering o los árboles de decisión. Para el estudio de los precios de la electricidad y su posterior predicción, el proyecto se centrará en la parte del *Data Mining* que utiliza técnicas de los árboles de decisión., los cuales analizan la relevancia de cada una de las variables estudiadas, así como la validez de los resultados. De este modo, se obtienen las mejores conclusiones en cada uno de los diferentes escenarios que se estén analizando.

En concreto, con este proyecto se analizan los resultados obtenidos con el empleo de los algoritmos de Minería de Datos pertenecientes a las técnicas de *boosting*. Esta metodología realiza los árboles de regresión, estudiando las variables que tienen una mayor relevancia en los resultados, de este modo, aquellas ramas que presenten un menor error serán las que tendrán preferencia a la hora de continuar. En primer lugar, se necesitará realizar unas pruebas de entrenamiento con diferentes valores para poder conocer la relevancia de cada variable. Una vez se obtengan estos resultados, se podrá pasar a la fase de predicción del resto de valores. Los algoritmos de *boosting* son recientes, por lo que no se tienen aún muchos estudios en los que se utilice el mismo. De este modo, a la vez que se logra encontrar solución a un problema predictivo actual, se consigue ampliar el conocimiento y la validez del uso de este método para este tipo de situaciones. Como consecuencia de este estudio, se podrá conocer si los algoritmos utilizados en este proyecto son adecuados para este tipo de datos.

Una vez se obtengan los resultados del estudio, y conociendo cuáles son las variables con mayor relevancia en el precio de la energía eléctrica, se podrán analizar las causas, las cuales podrían tener una gran importancia en estudios futuros.

El desarrollo del proyecto ha tenido varias fases. En la primera fase, se buscan todos los datos necesarios para poder realizar los análisis. Este proceso debe se realiza a conciencia, ya que cuanto mayor sea su volumen, mayor es la probabilidad de encontrar la relación entre las diferentes variables. De este modo, serán menores los errores cometidos, aumentando su exactitud. Por otro lado, se evalúan las posibles técnicas a utilizar para minimizar los errores de predicción. Entre todas las técnicas de minería de datos, tal y como se ha comentado anteriormente, se inclinó por probar con el uso de los algoritmos propios del *boosting* que *a priori* no se sabe si son los más adecuados.

En la segunda fase, se realizan una serie de pruebas de entrenamiento. En los mismos, se extraerán una determinada cantidad de datos relativos al precio, así como los valores correspondientes al resto de las variables que se van a analizar. Gracias a esta etapa de ensayo,

se podrá valorar la relevancia de cada una de las variables en el precio de la electricidad. Esta fase es muy importante, ya que gracias a ella se puede tener una primera idea de la validez de los resultados que se van a obtener en la predicción.

Por último, el proyecto tiene una tercera fase, en la que se realiza la predicción con los valores restantes y los resultados de las pruebas de entrenamiento realizadas anteriormente. Con los mismos, se realiza la estimación del precio de la electricidad de un modo ajustado en un periodo temporal de mínimo tres horas. Con este conocimiento, se puede dar la opción a los consumidores de cambiar su hábito de uso en el caso de que lo consideren oportuno.

2- INTRODUCCIÓN

2.1- Evolución del sector.

Las primeras referencias a la aplicación práctica de electricidad en España datan en 1852, cuando el farmacéutico Domenech consigue iluminar su botica, situada en Barcelona. Ese mismo año, se iluminan en Madrid la plaza de la Armería y el Congreso de los Diputados. En 1875, se realiza la instalación de una dinamo en la ciudad de Barcelona, con lo que se consigue la iluminación de las Ramblas, la Boquería, la parte de los altos de Gracia o el emblemático Castillo de Montjuic, con lo que comenzaría al año siguiente la electrificación industrial en España.

Figura 3: Quinqué o lámpara de Argand, artilugio de mechero que sustituyó a las lámparas de aceite hasta la llegada de la electricidad a mediados del siglo XIX.

Sin embargo, no es hasta 1888 que aparece una Real Orden encargada de regular el alumbrado de los teatros, prohibiendo la utilización de gas y limitando el uso de las lámparas de aceite a situaciones de emergencia. El desarrollo de la energía eléctrica fue uno de los factores que ayudó durante la primera mitad del siglo XX a que se produjera una gran evolución en el mundo industrial. Este desarrollo se consiguió gracias a la posibilidad de generar electricidad a larga distancia gracias a la aparición de corriente alterna a principios del siglo XX y que derivó en el desarrollo de centrales hidroeléctricas. Este hecho permitió disponer de más recursos de cara a la industrialización y ayudó a realizar grandes avances a gran escala, de modo que la producción de energía se vio triplicada en 1970.

Gracias al Real Decreto de 1987, se consigue la regulación del sistema de ingresos, con lo que se establecieron una serie de normas que fijaran los precios. Además, con la llegada de la Ley 54/97 aparecida en el BOE[1] de 1997, se conduce a un proceso de liberación dentro del sector del Sistema Eléctrico, con la consecuente eliminación del monopolio existente. En concreto, dicha ley dice lo siguiente: "*La presente Ley se asienta en el convencimiento de que garantizar el suministro eléctrico, su calidad y su coste no requiere de más intervención estatal que la que la propia regulación específica supone. No se considera necesario que el Estado se reserve para sí el ejercicio de ninguna de las actividades que integran el suministro eléctrico. Así, se*

[1] Fuente: BOE: Boletín Oficial del Estado.

abandona la noción de servicio público, tradicional en nuestro ordenamiento pese a su progresiva pérdida de trascendencia en la práctica, sustituyéndola por la expresa garantía del suministro a todos los consumidores demandantes del servicio dentro del territorio nacional."[2] Dicha ley fue modificada en 2013, siendo la vigente la Ley Orgánica 24/2013. En la misma, las principales novedades son las siguientes:

- Todas las unidades de producción deben realizar ofertas al mercado (Art. 23).
- Se establece el PVPC[3] como precio máximo de referencia al que podrán contratar los consumidores de menos de determinada potencia contratada (Art. 17).
- Serán considerados consumidores vulnerables aquellos que cumplan con las características sociales, de consumo y poder adquisitivo que se determinen. En todo caso, se circunscribirá a personas físicas en su vivienda habitual. (Art.45.1).
- Planificación eléctrica: se incorporan herramientas para alinear el nivel de inversiones a la situación del ciclo económico y a los principios de sostenibilidad económica (Art. 4 y 14).

La liberalización del sector tuvo como consecuencia la eliminación del monopolio que existía, por lo que los precios establecidos para el precio de la electricidad hasta entonces tuvieron el aliciente de la competitividad, por lo que se comenzó a intentar implantar el mejor servicio de suministro de electricidad a la sociedad con los mejores precios.

Por último, a continuación se muestra una figura de Barcelona a finales del siglo XIX y una fotografía actual de la ciudad, como muestra de las diferencias existentes gracias a la evolución de la electricidad en los últimos 100 años.

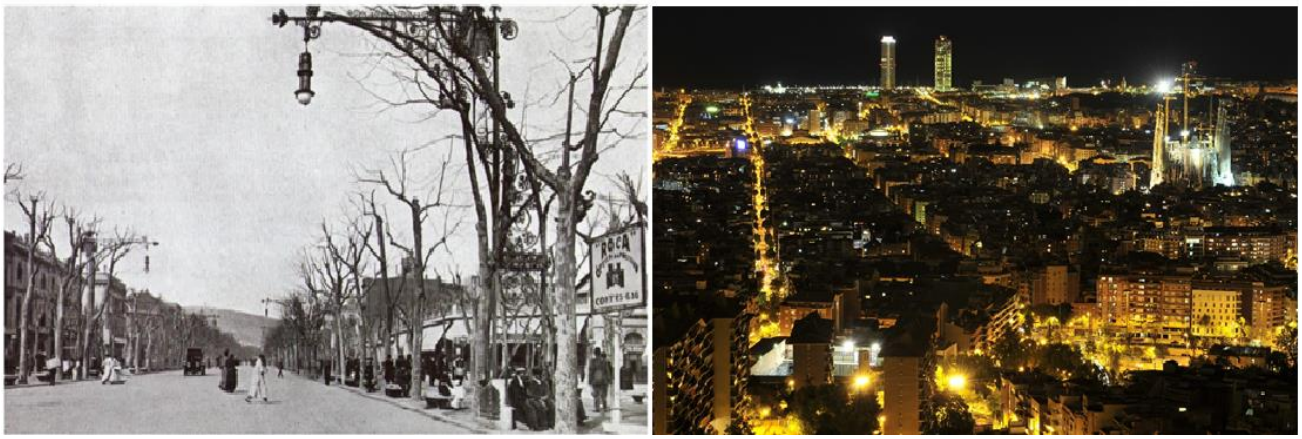

Figura 4: Imagen que contrasta el Paseo de Gracia a finales del siglo XIX y una imagen nocturna actual de la ciudad de Barcelona.

2.2- Situación actual.

La red eléctrica actual en España se establece de modo que las empresas encargadas del suministro eléctrico sean capaces de generar la cantidad que se esté demandando en cada momento, procurando evitar su almacenamiento, el cual no es posible en grandes cantidades. Este hecho hace que sea muy importante procurar prever de antemano cuál va a ser la demanda, ya que la misma depende de muchos factores, como la hora del día o la estación del año. Además, en la actualidad se debe añadir la cantidad de electricidad generada gracias a las

[2] Fuente: Ley Orgánica 24/2013. Boletín Oficial del Estado, núm. 310, de 27 de diciembre de 2013.
[3] PVPC: Precio Voluntario para el pequeño consumidor.

energías renovables, también conocida como generación libre de CO_2. Esto ha supuesto un cambio notable, ya que el único uso de petróleo que ocurría en las primeras fases se ha ido modificando de modo que la generación de electricidad se realiza a través de numerosos métodos. Aun así, la existencia de tantas metodologías para la generación de electricidad (solar, nuclear, eólica, carbón…) hace que sea complejo obtener la cantidad de energía demandada, ya que no se puede predecir de modo exacto la cantidad de electricidad generada mediante métodos renovables. A continuación, se observa un gráfico en el que se puede apreciar las diferentes fuentes de generación eléctrica que se utilizaron en España a lo largo de 2014, así como un esquema característico que explica el ciclo de generación de la energía.

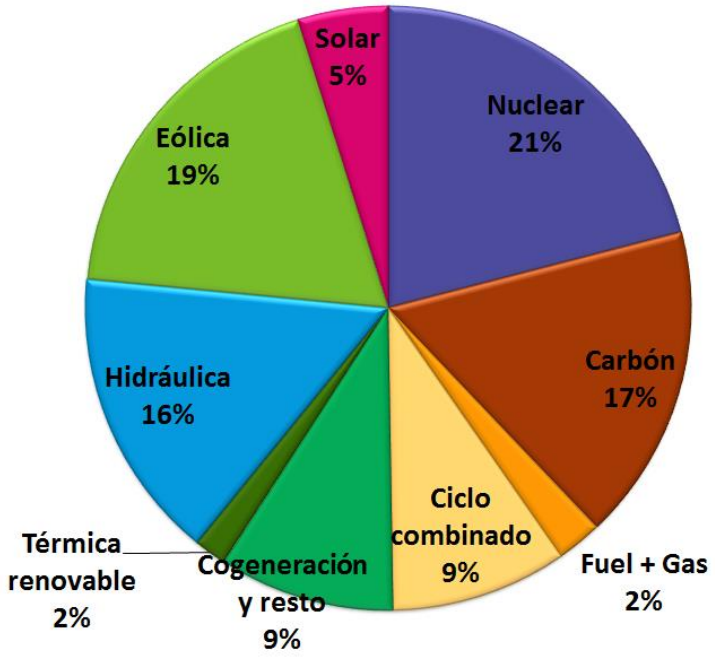

Figura 5: Origen de la generación eléctrica en España (2014). Fuente http://www.unesa.es/.

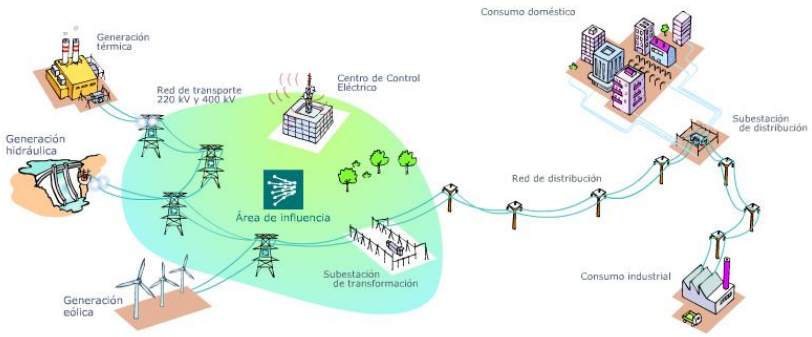

Figura 6: Esquema del proceso de generación y suministro de la electricidad. Fuente www.ree.es.

De este modo, la demanda de energía eléctrica es un factor importante de cara a conocer la cantidad de energía que se debe generar mediante la utilización de carbón o derivados del petróleo. Esto se traslada también a los precios de la electricidad consumida, por lo que la cantidad de energía renovable generada es una variable muy importante de cara a la predicción del precio de la electricidad. Este precio es común para todo el Estado, y varía cada hora en función de diversos factores, entre lo que se encuentran el momento del día, si el mismo es festivo o laborable, la estación del año, la cantidad de agua embalsada, la temperatura ambiente, la demanda o la cantidad de energía que se ha generado a través de métodos libres de CO_2.

Como factor económico, la demanda de energía se presentaba en primer momento como una variable muy importante, ya que *a priori* se estima que el consumo de electricidad en los meses con estaciones más frías requiere una mayor cantidad de energía. Esta primera idea se debe a la menor existencia de horas de luz natural, así como a la necesidad de calentar los espacios en algunos casos.

En la figura 7, sin embargo, no muestra lo mismo. En ella aparecen los datos de la demanda de electricidad en un día laboral en cada una de las estaciones, recogidos cada diez minutos. En la gráfica, se observa que, si bien la demanda parece ligeramente inferior durante el otoño, todas las estaciones mantienen la misma escala de demanda. Esto invalida la premisa de que la época del año sea de algún modo determinante para el precio de la electricidad. Por otro lado, también se observa que a lo largo del día existe una estacionalidad respecto de los picos de demanda de electricidad. Los datos mencionados se obtienen a través de la página oficial de la Red Eléctrica de España https://www.esios.ree.es/es.

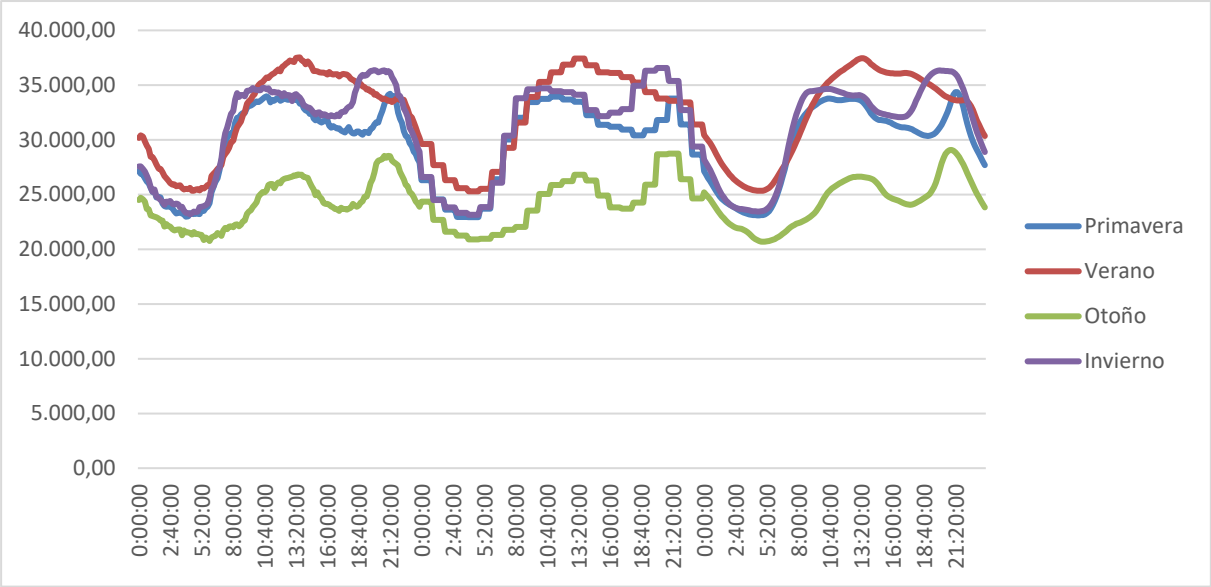

Figura 7: Demanda diaria de electricidad típica en las cuatro estaciones en 2015. Fuente https://www.esios.ree.es/es.

Por otro lado, el creciente uso de las energías renovables ha resultado una gran ventaja de cara a la estabilización del precio de la energía. El aumento de la generación mediante energías renovables ha causado una menor dependencia de los productos derivados del petróleo, cuya utilización había sido fundamental en las centrales eléctricas en sus inicios. Un problema en la utilización del petróleo se encuentra en la variación de su precio, lo que repercute en el precio final de la electricidad. Esto se observa claramente en los últimos años, donde el precio del Brent ha sufrido una disminución en el coste destacados. En la figura 8 se muestra una comparativa entre el precio del barril de Brent y el precio de la electricidad desde enero del 2014 a diciembre de 2016.

Figura 8: Precio del barril de Brent y de la electricidaddesde 2014 a 2016. Fuentes https://www.esios.ree.es/es y https://www.preciopetroleo.net/brent.html.

Con el aumento del uso de otros tipos de energías no sólo se consigue mejorar las condiciones ambientales, si no que se consigue un precio eléctrico mucho más competitivo, por lo que se está procurando incentivar este tipo de fuentes energéticas por delante de las fuentes energéticas convencionales.

Para el establecimiento actual de los precios de la electricidad en la actualidad se realizan subastas gestionadas por la OMIE[4]. Para ello, las empresas generadoras de luz ofrecen su electricidad a un precio determinado. Es entonces cuando la OMIE analiza y gestiona todas las ofertas, escogiendo aquellas que ofrezcan los precios más competitivos y asegurándose de que se cubren todas las necesidades. Una vez se escogen todas las ofertas, se utiliza la energía desde la más barata a la más cara. De este modo, todas las empresas que hayan suministrado energía eléctrica percibirán el coste al mismo precio que la oferta más cara que haya sido utilizada. En la figura 9, se ve una gráfica facilitada desde la página oficial OMIE donde se ve la asignación del precio de la electricidad para una hora dada para el 12 de octubre de 2016[5].

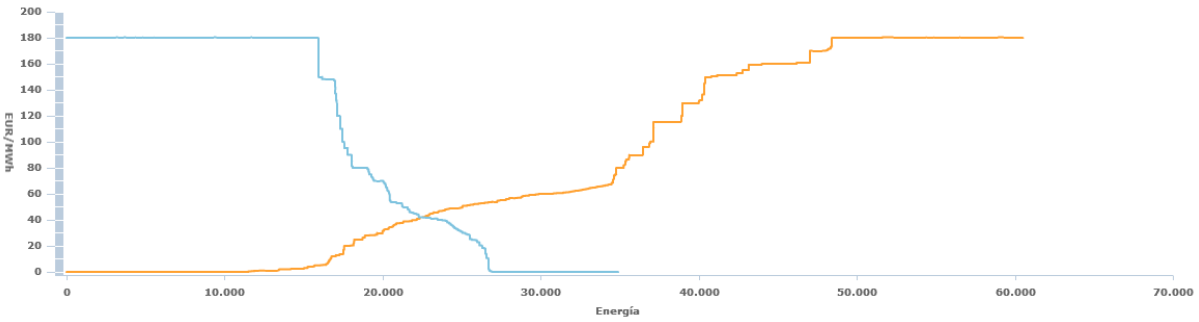

Figura 9: Gráfica en la que se observa la obtención del precio de la electricidad para una hora determinada.

[4] OMIE: Operador del Mercado Ibérico de Energía.
[5] Fuente: http://www.omie.es/inicio

En la imagen, se observa una línea naranja, que representa la oferta de venta, es decir, la generación de energía. En la línea celeste se observa la oferta de compra, que se trata de la demanda de energía para esa hora en concreto. El punto en el que ambas líneas se cortan da lugar al precio marginal, a través del cual más adelante se obtendrá el precio final de la electricidad en esa hora en concreto.

3- OBJETIVOS

El objetivo principal de este proyecto se centra en el estudio de uno de los últimos algoritmos desarrollados para la creación de árboles de predicción de *Data Mining*, de modo que pueda predecir el precio de la energía eléctrica, con un error lo más bajo posible.

Para conseguir dicha predicción, se analizan los diferentes factores que pueden influir en la fluctuación del precio de la electricidad, para con ello poder predecir a corto plazo cuál será el precio del mismo. Para la obtención de los valores, se extrae el historial de datos de las diferentes variables que se espera que pueden ser influyentes, como la temperatura, el precio en tiempos anteriores o el incipiente crecimiento de energías renovables, y con ello analizar las causas de la subida o bajada de los precios.

Por otro lado, existen muchos algoritmos que trabajan realizando árboles de regresión que han sido ampliamente utilizados en el pasado para poder realizar labores de clasificación y predicción. El uso en este proyecto de un algoritmo relativamente nuevo como el *boosting*, permite no sólo analizar el caso que se estudia con este proyecto, si no profundizar más en la validez del algoritmo. Además, permitirá estudiar la validez de *boosting* para proyectos con este tipo de datos, ya que antes de elegir un algoritmo para realizar las distintas predicciones, es conveniente conocer las características más convenientes para cada uno de ellos.

Por último, un objetivo que pretende alcanzarse con el correcto desarrollo del estudio reside en la capacidad de informar al consumidor, de modo que conozca los precios que tendrá la energía eléctrica, y con ello conseguir un coste más bajo con el consumo de la electricidad o realizar una predicción de los gastos realizados. Además, también resulta interesante conocer los factores que pueden hacer que el precio sea más bajo. De este modo el sector podrá incentivarse y llegar a un punto en el precio de la electricidad sea más competitivo tanto para consumidores como para los agentes encargados de la compra y venta de energía, sin por ello inducir a grandes pérdidas económicas por parte del sector.

4- METODOLOGÍA

Existen muchos métodos utilizados para poder realizar previsiones, por lo cual se debe analizar la situación para cada caso en el que se requieran realizar este tipo de actividades, con el objetivo de conseguir los mejores resultados posibles en función del tipo de datos con los que se cuentan para poder realizar la estimación.

Uno de los métodos más utilizados es la simulación de escenarios, en los que, realizada una hipótesis de partida sobre el modelo generador de los datos e incluyendo una limitada variabilidad, el sistema te permite estimar los valores con un elevado nivel de acierto. Esta metodología, a pesar de poder analizar sistemas con grandes volúmenes de datos, deben utilizar los algoritmos sobre una cantidad limitada de variables, ya que cuando los factores que influyen en el análisis son numerosos, el intento de aplicar la simulación se convierte en un proceso lento y con un margen de error mayor[6].

Una metodología que está cobrando mayor relevancia en los últimos años es la llamada Minería de Datos o *Data Mining*[7]. Mediante este método, que contiene numerosas técnicas y algoritmos capaces de resolver los problemas planteados dependiendo la naturaleza de los datos que se estén manejando, se encuentran patrones en los datos almacenados, permitiendo obtener una predicción fiable con el análisis de una base de datos en función de gran cantidad de factores. Debido a la naturaleza de los objetivos que se plantean con el presente proyecto, se considera que el uso de la minería de datos es lo más adecuado, por lo que se pasa al análisis de los algoritmos más adecuados para el estudio.

4.1- Minería de datos.

Una definición válida de minería de datos podría ser la siguiente: "*la exploración y el análisis -por medios automáticos o semiautomáticos- de grandes cantidades de datos con el fin de descubrir patrones con significado*"[8].

En el caso del proyecto, el análisis de los datos recopilados de los precios de la electricidad en España cada hora a lo largo de todo el año 2016 se realiza con el objetivo de encontrar factores externos y ambientales que sean los causantes de crear unos patrones que ayuden a poder realizar predicciones futuras en dicho precio, en función de dichas variables.

Los algoritmos planteados para el proyecto son aquellos encargados de realizar árboles de decisión, los cuales caracterizan las variables y realizan un mapeo de la base de datos inicial. Para ello, se escogen un conjunto de valores de cada una de las variables utilizadas y se configuran árboles, en los que se da importancia a los valores de los factores que muestren una mayor relevancia en los resultados. Dependiendo del tipo de datos de los que se disponen, los árboles pueden ser de clasificación (los datos son de tipo clase) o de regresión (los datos manejados son numéricos). En la figura 10 se muestra una clasificación típica en forma de árbol.

[6] Fuente: Harrington, H. J.; Tumay, K.(1999) *Simulation modeling models*
[7] Fuente: Benjamin Hofner, Andreas Mayr, et al. (2014). *Model-based Boosting in R – A Hands-on Tutorial Using the R Package mboost. Computational Statistics, 29:3-35.*
[8] Fuente: Oded Maimon and Lior Rokach (2010). *Data Mining and Knowledge Discovery Handbook.*

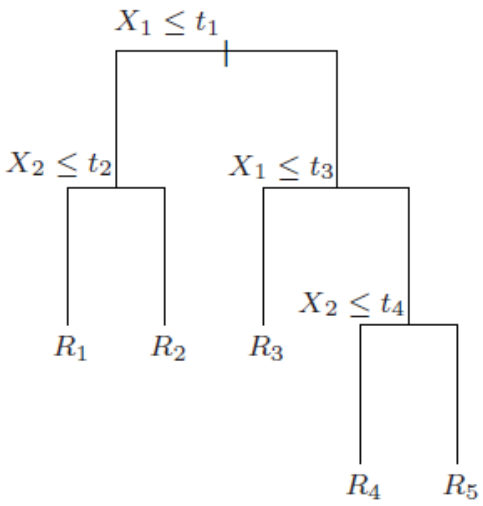

Figura 10: Ejemplo genérico de la configuración de un árbol.

En este proyecto, los datos recopilados darían lugar a un árbol de regresión, de modo que su esquema sería el que sigue a continuación, donde se marca el camino que se seguirá finalmente en naranja.

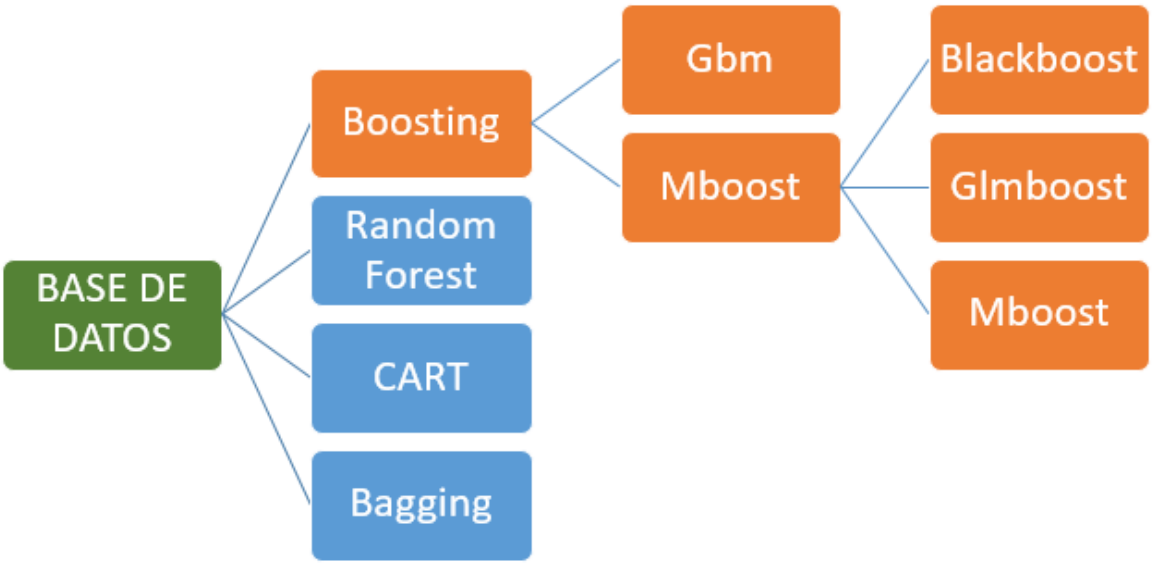

Figura 11: Esquema general del estudio de datos basado en los árboles.

Todos los datos extraídos para la realización del proyecto forman la base de datos, los cuales podrán ser estudiados mediante la metodología de árboles a través de diferentes algoritmos. Al tratarse de datos cuantitativos en los que una variable (el precio de la electricidad) depende de diversos factores, los resultados se analizarán a través de árboles de regresión. Estos árboles se podrán realizar con todos los algoritmos reflejados en la figura 11. La metodología utilizada en este caso será la perteneciente a los algoritmos de *boosting*, pudiendo comparar los resultados obtenidos con los del resto.[9]

[9] Fuente: Bühlmann P (2006). *Boosting for high-dimensional linear models. Ann Stat 34:559-583.*

A continuación, se realizará un repaso más exhaustivo por todos los algoritmos que podrían realizar el estudio planteado en el proyecto.

4.1.1- Modelo CART.

Se define como CART[10] el algoritmo introducido por Breiman, Friedman et al. (1984), de gran popularidad ya que es capaz de realizar tanto árboles de clasificación como de regresión. De este modo, puede clasificar la relevancia de un patrón, ya sea numérico o cualitativo. Las soluciones con este tipo de procedimientos obtienen árboles de gran tamaño, que consiguen una gran profundidad en la búsqueda de resultados, pero que tiene el riesgo de sobreajuste.

El árbol se construye a través de un nodo raíz que contiene todos los datos incluidos en el estudio, el cual se va clasificando mediante segmentación binaria los datos en diferentes subgrupos, los cuales se diferencian entre sí en base a un criterio concreto. Para realizar las divisiones, se escoge un criterio respecto a una única variable, la cual será la que mayor heterogeneidad en los nodos hijos, es decir, la más discriminante. El modo de partición se realiza del siguiente modo:

$$\text{Para } X=(X_1,\ldots,X_P),$$

$$X_n<c \text{ o } X_n\geq c$$

Las distintas ramas se dividen en los casos en los que existe heterogeneidad en la variable respuesta, parándose cuando en un nodo no existe una variable que sea lo suficientemente discriminante, o al no existir un número suficiente de datos para poder continuar con el análisis. Es lo que se denomina nodo terminal.

De este modo, el árbol maximal realizado con CART tiene tantos nodos terminales como datos posea el estudio, lo que hace que se convierta en un método muy largo para casos en los que exista una gran cantidad de datos. Además, el hecho de que existan tantos nodos terminales dificulta la correcta interpretación de los resultados, lo que obliga a realizar podas sobre las ramas en las que se estime que los resultados obtenidos no van a ser los deseados.

Ventajas	Inconvenientes
Capaz de operar con datos numéricos o cualitativos	Modelo demasiado exhaustivo y lento para el tratamiento de gran cantidad de datos
Resulta sencillo interpretar los resultados obtenidos	Debido a la gran cantidad de resultados, resulta necesario podar ramas
Permite operar con gran cantidad de variables	Posibilidad de que pequeños cambios en los datos conlleven grandes variaciones en los árboles

Tabla 1: Ventajas e inconvenientes del método CART.

4.1.2- Modelo *Bagging*.

El *bagging* se trata de un método posterior que con la ayuda de la metodología CART realiza la evaluación de distintos árboles, eligiendo conjuntos distintos de datos por remuestreo. Esta técnica, propuesta por Breiman (1996), procura reducir la variabilidad que pueda aparecer en

[10] CART: Classification And Regression Trees.

los árboles de regresión. Para poder realizar estos árboles, escoge las variables que vaya a utilizar y establece ciertas restricciones para valorar los distintos resultados que se obtengan.

Para poder realizar las predicciones del modo más exacto posible, la técnica del *bagging* dedica una cantidad determinada de datos a realizar un entrenamiento. En él, se generan nuevos datos por remuestreo en diferentes muestras aleatorias y uniformes (bootstrap). Los resultados obtenidos tendrán unos parámetros que permitirán conseguir los mejores resultados con los datos que se utilicen para la realización del modelo. De este modo, se realizan numerosos árboles, los cuales pueden tomar mayor o menor relevancia en los valores finales. Además, gracias a los resultados obtenidos en las etapas de ensayo, se consigue identificar las posibles diferencias imperceptibles que llevaran a mayores errores más tarde.

En los casos en los que existen variaciones en los datos existentes, los mismos pueden llevar a cambios importantes en los resultados. Es conveniente la utilización de los algoritmos de *bagging* en estos casos, ya que debido a la aleatoriedad en la realización de los diferentes árboles de entrenamiento permiten la reducción en los errores cometidos.

En cuanto a la realización del modelo, dado conjunto inicial de datos donde $D=\{(x_i, y_i), i=1,m\}$, donde $x_i \epsilon X$ e $y_i \epsilon Y=\{1,..., k\}$, el sistema se divide en diferentes muestras $w_i(1)=1/m$, $\forall i=1,m$. De este modo, la aplicación del algoritmo se muestra en la figura 12.

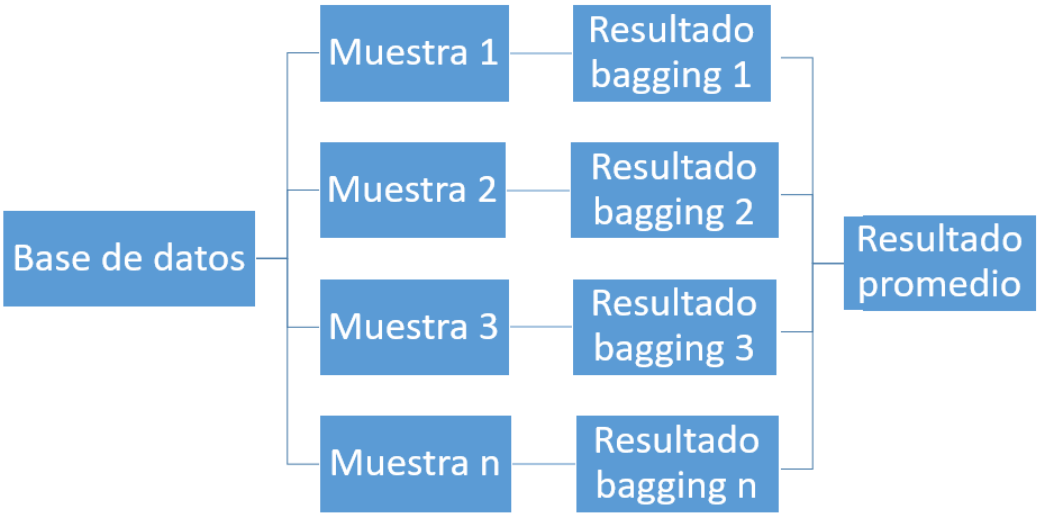

Figura 12: Esquema del proceso de *bagging*.

La figura 13 muestra un ejemplo de los árboles de entrenamiento mencionados anteriormente[11].

[11] Fuente: Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.*

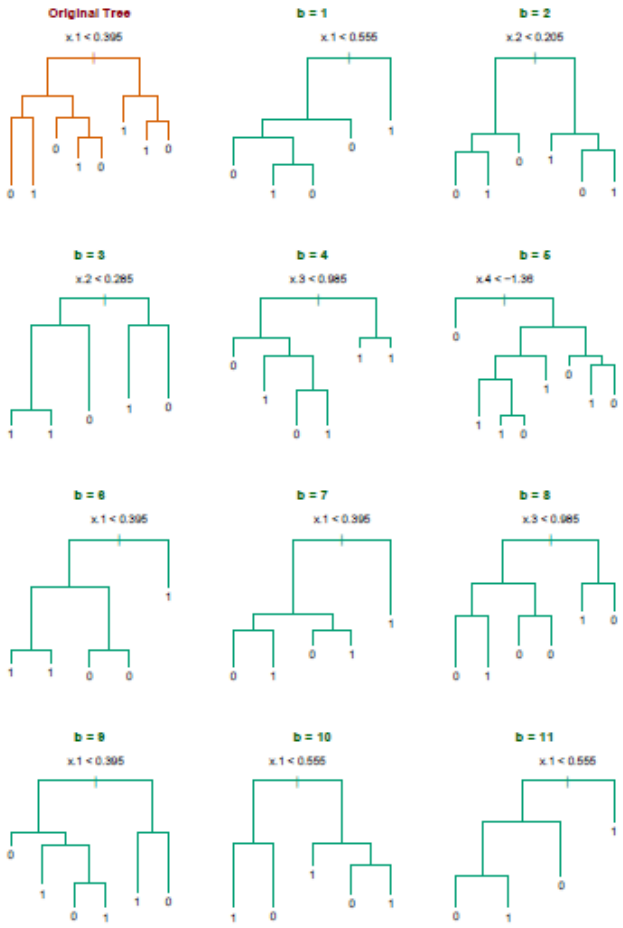

Figura 13: Diferencia entre un árbol que acepte todas las variables y un estudio mediante árboles obtenidos por *bagging*.

La tabla 2 que se presenta a continuación resume las ventajas e inconvenientes del uso de esta técnica.

Ventajas	Inconvenientes
La existencia de árboles de entrenamiento permiten reducir la variabilidad	Puede dejarse combinaciones sin estudiar
No necesita realizar la poda de los árboles	La interpretación de los resultados al tratarse de distintos árboles puede resultar complicada

Tabla 2: Ventajas e inconvenientes de uso de *bagging*.

4.1.3- Modelo *Random Forest*.

Los algoritmos encargados de la metodología *Random Forest* (Breiman, 2001) se desarrollaron como una mejora de la técnica anterior. Se basan en la técnica que estudia los datos pertenecientes a una situación concreta, lo que le convierte en un método adecuado en la búsqueda de factores determinantes. La elección de los árboles a analizar se realiza de un modo completamente aleatorio, por lo que todas las variables que pueden ser causantes de las variaciones de los datos a analizar son utilizadas. Al igual que en el *bagging*, se generan diferentes remuestreos de modo aleatorio, y además se elige la variable de partición de los nodos de entre un subconjunto de todas las variables explicativas. Cuanto mayor sea el número de

árboles realizados a través de este método, mayor será la precisión de los resultados obtenidos, ya que se habrán estudiado muchos más casos.

Para realizar la predicción con cada uno de los árboles realizados con este método, se analiza un subconjunto de todas las variables, realizando diferentes cortes y separaciones en los mismos con el fin de obtener los valores más acertados. Además, al igual que ocurría con el *bagging*, las técnicas de *Random Forest* utilizan diferentes muestras como entrenamiento antes de pasar a obtener los resultados reales que van a ser sujetos a análisis. Una vez se tienen todas las muestras que se van a realizar, se construyen los árboles máximos de cada uno de ellos, sin la realización de podas en cada uno de ellos. La diferencia entre *bagging* y *Random Forest* reside en la elección aleatoria de los predictores posibles, y la elección de la mejor división dentro de las variables incluidas en el mismo (en *bagging* no se realizaba elección de las variables predictivas).

En la tabla 3 se muestra un resumen de las ventajas e inconvenientes que se encuentran en el uso de *Random Forest*.

Ventajas	Inconvenientes
Su estabilización se produce relativamente rápido	Los resultados obtenidos resultan más complicados de interpretar que en los métodos antes mencionados
Poder analizar de modo aleatorio varias variables permite realizar una mejor predicción	No se puede realizar un análisis individual sobre cada árbol generado

Tabla 3: Ventajas e inconvenientes de la aplicación de *Random Forest*.

4.1.4- Modelo *Boosting*.

El algoritmo *boosting* se presenta como la herramienta más potente en la creación de árboles introducida en los últimos años. Los orígenes del *boosting* se remontan al trabajo de Vilian y Kerns (1989), que plantearon la posibilidad de estimular (*boost)* un algoritmo teniendo en cuenta sus errores. Desde que se planteó por primera vez, los algoritmos de *boosting* han seguido desarrollándose y mejorando[12].

El *boosting* parte de la premisa de que el conjunto de diversos clasificadores débiles combinados resulta en un clasificador grande más preciso. En este caso, a pesar de que los datos a analizar se eligen también de un modo aleatorio, el árbol se va construyendo a partir de los errores que van apareciendo en su desarrollo, de manera que el algoritmo prioriza y avanza por las partes que generen menor error, con el objetivo de conseguir los resultados más adecuados a la realidad. Para ello, se realizan diversas iteraciones en el sistema que se esté estudiando de un modo aleatorio. Una vez se tengan todos los resultados, se establecerá una jerarquía en el cual se den más peso a los resultados que hayan implicado un error menor. Esta es la principal referencia en relación a los algoritmos descritos anteriormente. La ilustración 14 indica el flujograma que sigue el algoritmo.

[12] Fuente: Bühlmann P (2006). *Boosting for high-dimensional linear models.*

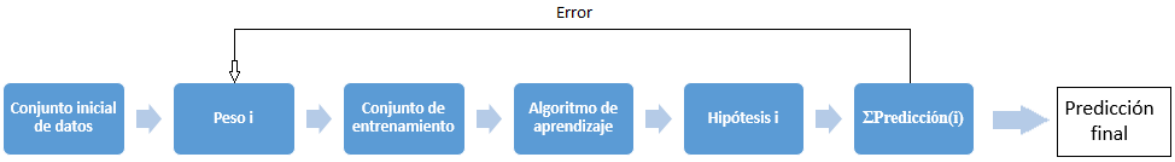

Figura 14: Esquema de funcionamiento del algoritmo *boosting*.

Dentro de este método, existe un algoritmo complementario llamado *AdaBoost* (*adaptive boosting*), y que se ha convertido en el algoritmo más estudiado y popular de la familia. Formulado por Yoav Freund y Robert Schafire (1997), el objetivo de este algoritmo es el de entrenar iterativamente una serie de predictores base, de modo que en cada iteración se tenga en cuenta los errores cometidos y se dé más peso a los resultados que menor error han emitido. Con ello, se podrá construir una predicción óptima en forma de árboles de regresión.

En primer lugar, se definen las entradas, determinando cada una de las variables implicadas. A continuación, se inicializa la distribución de los pesos en cada uno de los datos existentes. Para cada conjunto de valores $(y_1,x_1)\ldots(y_m,x_m)$, donde $y_i \in Y$, $x_i \in -1, +1$.

Cada uno de los datos en este primer momento tiene el mismo error, $D_i=1/m$ para $i=1, \ldots, m$.

Durante la primera iteración, se entrena con el clasificador débil, utilizando el conjunto de pesos de error D_i. Con ello, se obtiene una primera hipótesis h_t: $Y \rightarrow -1, +1$. Se pretende obtener una h_t con un error bajo.

$$\varepsilon_t = Pr_{i\ Di}[h_t(y) \neq y_i]$$

Con los resultados obtenidos en los errores para cada dato, éstos actualizan el peso en los clasificadores anteriores

$$\propto_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

Llegados a este punto, los errores pertenecientes a cada dato pueden actualizarse, de modo que en la siguiente iteración los pesos de los mismos permitan obtener unos mejores valores y una disminución en los errores.

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\propto_t x_i h_t(y_i))}{Z_t}$$

Donde Z_t se trata de un factor de normalización. La hipótesis final se expresaría de la siguiente manera:

$$H(y) = signo\left(\sum_{t=1}^{T} \propto_t h_t(x)\right)$$

Durante la realización de las diferentes operaciones, se modifican los errores existentes en cada uno de los datos, de manera que durante las siguientes iteraciones dichos errores sirvan para la modificación y ajuste de los resultados. Cuanto mayor sea el número de iteraciones realizadas, menor será el error cometido hasta que llegue el punto en el que el mismo se estabilice.

Con ello, las ventajas e inconvenientes que se observan de la utilización del *boosting* se observan en la tabla 4.

Ventajas	Inconvenientes
Prioriza los datos que aportan un menor error, aumentando la predictibilidad de los resultados	Necesita un gran número de árboles antes de ver sus resultados estabilizados, por lo que se trata de una metodología más lenta.
Disminuye la variabilidad resultando de la realización de los árboles	No estudia todas las variables involucradas en el estudio, descartando las que no parecen tener relevancia en un primer momento.

Tabla 4: Ventajas e inconvenientes en la aplicación de *boosting*.

En el gráfico 15[13] se ve una comparación de tres de las metodologías mencionadas anteriormente (*bagging*, *Random Forest* y *boosting*), en el cuál se intuye *a priori* que el algoritmo que aporta un menor error para números de árboles grandes es el *boosting*, siendo el que peor resultado da para menor cantidad de árboles. Ésta se convierte en una de las razones por las que resulta interesante ver lo que sucede en la predicción del precio de la electricidad.

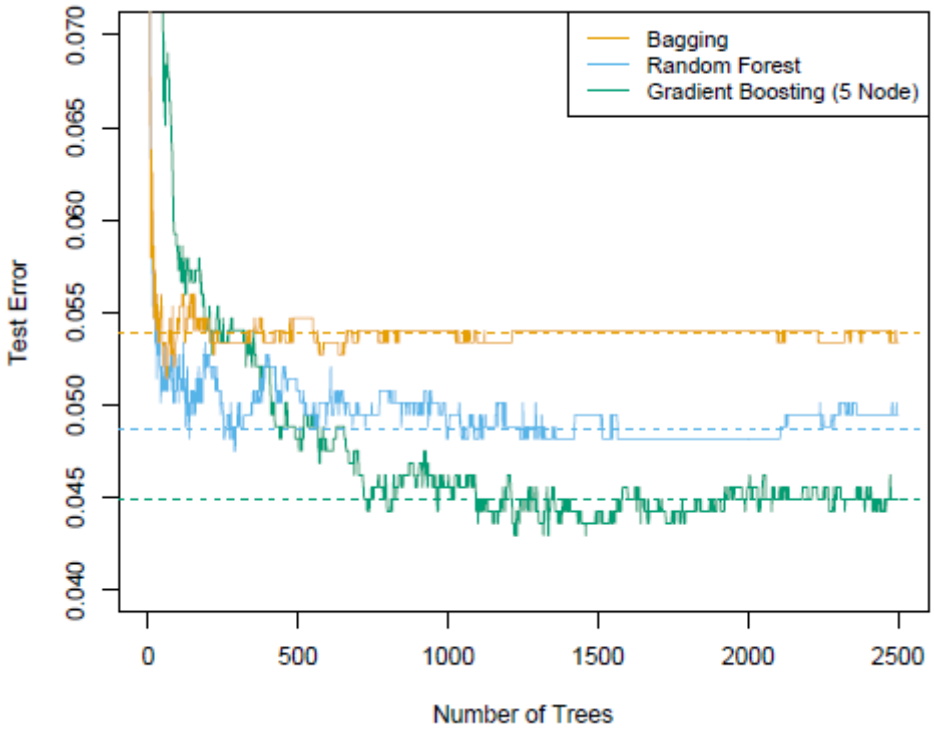

Ilustración 15: Porcentaje de error de los métodos *bagging*, *Random Forest* y *boosting* en estacionalidad.

Una vez analizados todos los modelos anteriores que podrían realizar los análisis pertinentes para la predicción de los datos con el menor error posible, finalmente este proyecto se decanta por la utilización del modelo *boosting*. Esto se debe a la investigación de los valores que arroja el uso de un algoritmo nuevo en el tratamiento de este tipo de datos, lo que resulta de gran interés para la evaluación de su utilidad, así como para la posible valoración de mejoras en el

[13] Fuente: Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.*

mismo. Para ello, se analizarán los resultados obtenidos con diferentes funciones y paquetes del programa estadístico R. Las especificaciones de cada una de las funciones utilizadas se detallarán más adelante en el apartado adecuado.

4.2- Software estadístico empleado: R.

R se trata de un lenguaje libre especializado en la programación estadística realizado y actualizado por la R Foundation for Statistical Computing[14]. Se trata de un GNU[15] con un sistema operativo de software libre, y en el que los usuarios tienen control sobre las tareas de computación del programa.

El sistema fue desarrollado inicialmente en 1993 por Robert Gentleman y Ross Ihaka en la Universidad de Auckland. Sin embargo, no se hizo público para el uso general de la población hasta el año 2000[16].

R proporciona una gran variedad de técnicas tanto gráficas como estadísticas (modelos lineares y no lineares, tests estadísticos clásicos, análisis de series temporales, clasificaciones, clustering…). Una de las mayores ventajas que se encuentran con el uso de R reside precisamente en la facilidad existente a la hora de producir diferentes tipos de gráficos y diagramas, así como la inclusión de símbolos matemáticos y diferentes paquetes que permitan la formulación de múltiples algoritmos y estudios estadísticos. Además, debido a la inclusión individualizada de dichos paquetes, el programa se encuentra en constante desarrollo.

Otra de las ventajas existentes en el uso de R es que, a pesar de tratarse de un servicio gratuito, su capacidad de análisis y cálculo estadístico es amplio, además de permitir la inclusión de nuevas fórmulas o extender las ya existentes para llegar aún más lejos en los resultados.

Para poder facilitar el uso de este sistema, se instala además el programa RStudio. RStudio se trata de un entorno desarrollado especialmente para R, a través del cual se facilita la navegación por el programa, así como añadir nuevas facilidades que permitan mejorar la programación del mismo. A continuación, se ven dos ejemplos de la diferencia entre ambos entornos. En ambos, se incluye un ejemplo en el que se almacena una tabla simple de datos. En ella se ve como con el programa R sólo se tiene una interfaz simple en el que se realiza la programación que vaya a utilizarse. Sin embargo, con R Studio además de la propia ventana de programación se cuenta con la aparición de dos ventanas más. La ventana superior derecha se encarga de la visualización de las diferentes variables que se estén manejando en el estudio, así como un historial de la programación realizada. En la parte inferior derecha, se obtiene un apartado en el que pueden visualizarse las diferentes gráficas definidas, los paquetes disponibles en la biblioteca de R o los archivos con datos que puedan utilizarse como parte de la programación.

[14] R Foundation for Statistical Computing o Fundación para la Programación Estadística en R: organización no gubernamental desde la que actualizan y mejoran el programa de un modo continuo.
[15] GNU: se trata de un sistema operativo tipo Unix de software libre fundado en septiembre de 1983 por Richard M. Stallman. Su objetivo es que los usuarios tengan la libertad de compartir y mejorar el software integrado.
[16] Fuente: http://www.cran.r-proyect.org.

```
R version 3.2.5 (2016-04-14) -- "Very, Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

[Previously saved workspace restored]

> nombre<-c('Juan','Pedro','Maria')
> edad<-c(18,21,19)
> personas<-data.frame(nombre,edad)
> personas
  nombre edad
1   Juan   18
2  Pedro   21
3  Maria   19
>
```

Figura 16: Entorno de R. Una única ventana.

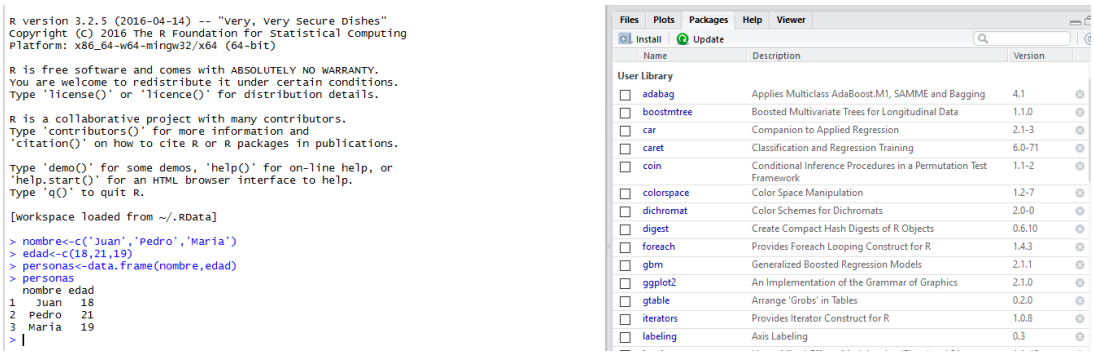

Figura 17: Entorno de R con RStudio. Varias ventanas con atajos y simplificaciones.

Otra de las razones que hacen atractiva la utilización tanto de R como de su complemento RStudio, es que se trata de un programa gratuito gracias a su elaboración altruista. De este modo, y conocidas todas las bondades que ofrece el programa y sus capacidades para manejar grandes volúmenes de datos, la única desventaja que se encuentra para su uso es la necesidad de tener conocimientos de programación y estadística, ya que en muchas ocasiones los resultados no son fácilmente interpretables.

Para poder realizar ciertas operaciones o estudios más detallados, y como ya se ha comentado anteriormente, R cuenta con una serie de paquetes específicos, con diferentes tipos de funciones y desarrollos dependiendo de lo que se quiera analizar. Estos paquetes no son triviales, y deben descargarse previamente. Para la realización del proyecto, los paquetes que se han utilizado se describirán a continuación con más detenimiento.

Para el proyecto que estamos realizando, que utiliza algoritmos de *boosting*, existen fundamentalmente dos paquetes que pueden realizar los análisis que buscamos, *mboost* y *gbm*, y que son los que vamos a utilizar. Ambos paquetes parten de una metodología similar con los mismos pasos que se habían comentado en el punto 4.1.4. Sin embargo, existen unas pequeñas diferencias que harán posible su distinción de cara a los resultados obtenidos para cada una de las estimaciones propuestas, con un mayor o menor error cometido dependiendo del caso.

4.2.1- Paquete *mboost*.

El paquete *mboost*[17] se puede utilizar para estudios de regresión, clasificación y predicción en el tiempo, así como para el estudio de una gran variedad de problemas estadísticos en los que exista una gran cantidad de dimensiones. Se trata de un algoritmo de descenso de gradiente funcional (*boosting*) para optimizar las funciones de riesgo general utilizando estimaciones de mínimos cuadrados de componentes mínimos o árboles de regresión como aprendices de base para ajustar modelos lineales, aditivos e interactivos generalizados a datos potencialmente de alta dimensión

Para poder realizar los análisis, el programa considera una variable de salida *y* y algunas variables de predicción $x_{1,\ldots,}x_p$. De este modo, el programa será capaz de obtener el valor óptimo para la predicción de y en función de x. Con los datos existentes, el paquete es capaz de realizar árboles en los cuales se dan mayor importancia a las variables más significativas, de modo que las futuras predicciones sean lo más exactas posibles. Para ello, descarta las ramas que más error den, priorizando las restantes. A pesar de que el paquete cuenta con distintas funciones con determinadas especificaciones, todas ellas cuentan con unos procedimientos comunes, los cuales se describen a continuación.

1. Se toma una muestra aleatoria con n=p x N casos, se asigna a cada uno de los datos un error inicial idéntico, $w_i = \frac{1}{n}$, donde n es el número de valores existentes (siendo p el número de variables estudiadas y N el número de casos estudiados).
2. Se entrenan los datos existentes, calculando el error existente en cada uno de ellos.
3. Se cuentan e identifican los datos mal clasificados.
4. Se incrementan los pesos en los casos de entrenamiento que el modelo calcula erróneamente.
5. Volver a repetir el punto 2 hasta llegar al número de iteraciones establecidas en un primer momento.
6. Ponderación de los errores existentes en todas las iteraciones.

Dentro de este paquete, existen tres funciones posibles encargadas de hacer el *boosting*, las cuales se utilizan en este proyecto. Éstas son *mboost*, *blackboost* y *glmboost*, las cuales por defecto utilizan con los datos la distribución gaussiana. La diferencia en cada uno de ellos es la siguiente:

- ***Mboost*:** optimiza el peso de los errores a través de muestras aleatorias que utilizan los diferentes datos como bases de aprendizaje a través del análisis en árboles aditivos.
- ***Blackboost*:** optimiza los errores utilizando los árboles de regresión como bases de aprendizaje.
- ***Glmboost*:** optimiza los errores utilizando los modelos lineales como bases de aprendizaje.

4.2.2- Paquete *gbm*.

El paquete *gbm* se trata de una implementación de extensiones del algoritmo *AdaBoost* de Freund y Schapire y la máquina de gradiente de *boosting* de Friedman[18]. Incluye métodos de regresión para mínimos cuadrados, funciones de pérdida valor absoluto, regresión logística,

[17] Fuente: Benjamin Hofner, Andreas Mayr, et al. (2014). *Model-based Boosting in R – A Hands-on Tutorial Using the R Package mboost. Computational Statistics*, 29:3-35.
[18] Fuente: Greg Ridgeway (2007). *Generalized Boosted Models: A guide to the gbm package.*

regresión de Poisson, verosimilitud parcial de riesgos proporcionales de Cox, distribución multinomial, distribución t y *AdaBoost*.

Una de las grandes ventajas que supone la utilización de este paquete sobre los otros que también realizan el *boosting* sobre los diferentes modelos proviene de una mayor tasa de aprendizaje de los modelos con cada iteración.

La metodología principal que se lleva a cabo a través de este paquete es la que se muestra a continuación.

1. Se inicializa $\hat{f}(x)$ como una constante, siendo ésta $\hat{f}(x) = \arg min_{\rho} \sum_{i=1}^{N} \Psi(y_i, \rho)$.
2. Calcular el gradiente negativo como la siguiente respuesta:

$$z_i = -\frac{\partial}{\partial f(\mathbf{x}_i)} \Psi(y_i, f(\mathbf{x}_i)) \Big|_{f(\mathbf{x}_i)=\hat{f}(\mathbf{x}_i)}$$

3. Seleccionar p x N casos de la muestra de manera aleatoria.
4. Realización del árbol con K nodos terminales, utilizando los datos seleccionados en el punto anterior, $g(x) = E(z|x)$.
5. Calcular las predicciones óptimas en los nodos terminales, $\rho_1,\ldots,\rho_k$ del modo siguiente:

$$\rho_k = \arg\min_{\rho} \sum_{\mathbf{x}_i \in S_k} \Psi(y_i, \hat{f}(\mathbf{x}_i) + \rho)$$

Donde S_k es el conjunto de x_i que define el nodo terminal k.

6. Actualizar $\hat{f}(x)$ como $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda\rho_{k(x)}$, donde k(x) es el índice del nodo terminal en el cual una observación que contenga x baja.

Donde:

f̂(x): función de regresión.

$\Psi(y_i,\rho)$: errores cometidos.

Z_i: gradiente de los datos resultado respecto de las variables utilizadas.

5- RESULTADOS Y DISCUSIÓN

Una vez se tiene conocimiento de las características generales del sector, así como las herramientas principales que van ser necesarias para la ejecución del proyecto, existen tres partes fundamentales en la realización de los trabajos de investigación: obtención de la base de datos, realización del modelo y análisis de los resultados.

5.1- Obtención de la base de datos.

Para la correcta realización del modelo, será fundamental contar con datos que se consideren especialmente relevantes en el objeto de estudio. En este caso, las variables que se consideren importantes para la predicción del precio de la energía eléctrica. El modelo contará con varios factores que se consideran importantes para la variación del precio de la electricidad, de los cuales se obtendrán datos a cada hora a lo largo del año 2016, con lo que cada una tendrá un total de 8.784 datos. Las variables que se plantean en este caso se presentan a continuación en la tabla 5.

Variable	Descripción
V1	Precio de la electricidad en t
V2	Precio de la electricidad en t-1
V3	Energía renovable generada
V4	Temperatura máxima
V5	Temperatura mínima
V6	Temperatura media
V7	Hora del año

Tabla 5: Descripción de las distintas variables utilizadas en el análisis.

Una vez se consiguen recopilar todos los datos necesarios respecto a las variables establecidas, se crea una base de datos que más adelante sirva para poder alimentar el programa de R, y con ello poder realizar todos los análisis necesarios para realizar el modelo con los paquetes específicos proporcionados por el programa.

V1	V2	V3	V4	V5	V6	V7
48,55	50,95	12030	157,7	68	112,85	1
40	48,55	12438	157,7	68	112,85	2
33,1	40	11821	157,7	68	112,85	3
28,11	33,1	11101	157,7	68	112,85	4
27,13	28,11	10363	157,7	68	112,85	5
25,24	27,13	10310	157,7	68	112,85	6
19,98	25,24	10483	157,7	68	112,85	7
18,16	19,98	10817	157,7	68	112,85	8
17,73	18,16	11282	157,7	68	112,85	9
19,77	17,73	11721	157,7	68	112,85	10
23,75	19,77	13591	157,7	68	112,85	11
26,03	23,75	15700	157,7	68	112,85	12
27,06	26,03	17058	157,7	68	112,85	13
26,59	27,06	17785	157,7	68	112,85	14
25	26,59	18817	157,7	68	112,85	15
20,06	25	18772	157,7	68	112,85	16
19,43	20,06	18453	157,7	68	112,85	17
24,57	19,43	18049	157,7	68	112,85	18
33,11	24,57	19629	157,7	68	112,85	19
35,34	33,11	20710	157,7	68	112,85	20
33,07	35,34	20579	157,7	68	112,85	21
29,52	33,07	20289	157,7	68	112,85	22
30,1	29,52	19812	157,7	68	112,85	23
24,57	30,1	18712	157,7	68	112,85	24
22,2	24,57	17343	142,2	63,9	103,05	25
16,57	22,2	16467	142,2	63,9	103,05	26
15,35	16,57	16093	142,2	63,9	103,05	27
12,77	15,35	15638	142,2	63,9	103,05	28
11,27	12,77	15055	142,2	63,9	103,05	29
11,91	11,27	15715	142,2	63,9	103,05	30
12,62	11,91	15996	142,2	63,9	103,05	31
14,17	12,62	16307	142,2	63,9	103,05	32
17,73	14,17	16457	142,2	63,9	103,05	33
26,81	17,73	17613	142,2	63,9	103,05	34
35,4	26,81	19694	142,2	63,9	103,05	35
36,1	35,4	20524	142,2	63,9	103,05	36
30,34	36,1	21152	142,2	63,9	103,05	37
31,26	30,34	21362	142,2	63,9	103,05	38
30,99	31,26	21326	142,2	63,9	103,05	39
29	30,99	20444	142,2	63,9	103,05	40
32	29	19606	142,2	63,9	103,05	41
39,99	32	19425	142,2	63,9	103,05	42
45,5	39,99	20570	142,2	63,9	103,05	43
55,1	45,5	20877	142,2	63,9	103,05	44

Figura 18: Muestra de los datos recopilados para la elaboración del modelo.

Cuando se tienen todos los datos clasificados, los que se necesiten para cada uno de los análisis puede ser exportado en formato .txt, de modo que sea más sencilla su introducción en R y su posterior utilización.

Con los datos ya incluidos en R, y con ellos ya dentro de un *data.frame* que contenga la muestra sujeta a análisis, se activan los distintos paquetes que vayan a ser utilizados para la confección del modelo (en este caso los paquetes *gbm* y *mboost*). En la siguiente imagen, se observa una captura de RStudio, con los paquetes mencionados activados y una muestra introducida y lista para ser analizada.

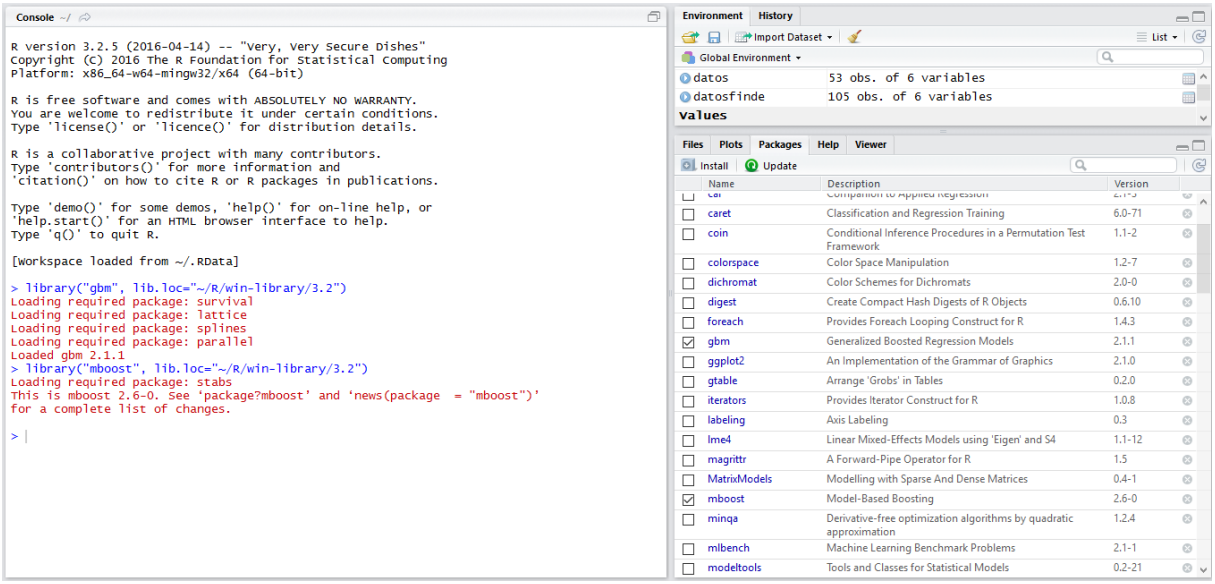

Figura 19: Captura de pantalla con dos muestras de datos distintas dentro del sistema y los paquetes *mboost* y *gbm* activados.

5.2- Estimación del modelo.

Una vez estudiados los paquetes que van a utilizarse, se pasará a la estimación de distintos modelos para la predicción de precios. Para ello, se utilizarán todas las funciones capaces de realizar árboles de clasificación mediante *boosting* incluidos en los paquetes elegidos. Las funciones utilizadas en el paquete *mboost* serán *glmboost*, *blackboost* y *mboost*, mientras que para el paquete *gbm* la única función con *mboos* que se utilizará será la función *gbm*.

Para poder valorar si la utilización de algoritmos propios del *boosting* son útiles de cara a la previsión de precios de electricidad, se establecerán varios escenarios (horizontes de predicción), en los cuales se vaya aumentando la dificultad y horizonte en la previsión. Dado que la programación en cada una de las diferentes funciones utilizadas es similar, la realización de los diferentes escenarios se analizará de manera general, profundizando en los diferentes resultados obtenidos en el apartado correspondiente. Asimismo, el código utilizado para la programación en cada uno de los casos se encontrará en el anexo correspondiente.

Previsión simple a t+1.

En primer lugar, se realizará una primera previsión, a través de la cual se obtendrá el precio estimado de la hora siguiente respecto a la que se está estudiando. Para ello, una vez se han introducido todos los datos que van a utilizarse dentro del sistema, se realiza la aplicación de los algoritmos correspondientes a cada una de las funciones.

Los resultados obtenidos de dichos algoritmos, además de para la realización posterior de la previsión, nos servirá para el análisis de la influencia de cada una de las variables utilizadas dentro de la previsión, lo que permite a su vez conocer la dependencia existente en la estipulación del precio respecto de los otros factores. En la figura 20 puede verse un ejemplo de dicha influencia a través de la utilización de una de las funciones. Como era de esperar, se ve que las variables que tienen una mayor influencia para la predicción de futuros precios son el precio de la hora anterior y la energía renovable generada. La correcta elección de las variables es muy importante, ya que no sólo ayuda a realizar predicciones futuras, si no a al estudio de la importancia de determinadas variables para los diferentes casos. Por ejemplo, pese a que el objeto de estudio de este proyecto consta únicamente de 5 variables fundamentales, se

observa que las más importantes de cara a la estimación del precio de la electricidad son el precio de la hora anterior y la cantidad de energía renovable generada, siendo menos relevante en este caso las temperaturas. En casos en los que la cantidad de variables analizadas sea mayor, será muy importante la realización de un estudio previo que señale cuáles son las variables relevantes para la obtención de los resultados.

```
> coef(resultadoglm)
  (Intercept)            V2            V3            V5
-3.613713e+01  9.679908e-01 -3.038515e-05  1.318839e-03
attr(,"offset")
[1] 37.74081
```

Figura 20: Peso de las variables en la dependencia del precio de la electricidad para el uso de la función *glmboost*.

A partir de aquí, se obtiene la predicción del tiempo t+1, gracias a la utilización de una muestra de entrenamiento con la que se conocerán los resultados obtenidos de la realización del *boosting*, y de una tabla con los valores de las variables pertenecientes a las horas que se quieran predecir. Como ya se verá en mayor profundidad en el siguiente apartado, los valores de error medio en este estudio son aproximadamente del 4%, lo que resulta un resultado muy competitivo. Con estos valores, se podría asegurar que mediante técnicas de *boosting*, la predicción de este tipo de valores resulta igual acertada que mediante otras técnicas que utilizan metodologías diferentes.

Previsión a t+2.

Pese a que la predicción a t+1 devuelve unos valores muy adecuados a los precios finales, se observa que normalmente la predicción se necesita con horizontes más largos, por lo que se debe realizar la predicción en tiempos posteriores. Para ello, y utilizando de nuevo los los mismos datos que con las funciones de *boosting* anteriores, se realizan predicciones a dos horas vistas, con las cuales valorar la validez de los resultados obtenidos respecto de los valores reales. Si estos valores son aceptables, los mismos se podrán evaluar a más largo plazo, lo que sería muy beneficioso para los usuarios.

Previsión a t+3.

Por último, se realizará la previsión para un tiempo t+3. Debido al mayor horizonte del modelo, se espera que los resultados obtenidos posean unos errores mayores que en el primer caso. Sin embargo, es necesario que los mismos sean lo suficientemente bajos como para poder ser considerados de cara a posibles estudios de precios de electricidad, o como para realizar estimaciones de cara a clientes y empresas.

El horizonte de predicción se puede modificar y a tal efecto se incluye un apartado dentro del anexo de las funciones utilizadas, en el cual se plasmará un programa auxiliar en el que estableciendo las horas horizontes a las que se quiera realizar la previsión, la misma sea fácilmente evaluable.

Además de la utilización de las diferentes funciones y tipos de previsiones, con el objetivo de realizar un análisis intensivo del funcionamiento interno de R, se realizaron todas las estimaciones utilizando dos metodologías distintas. En una de ellas, se realizaron dos *data.frames* distintos, uno con los datos que se utilizan en el training, y otro con los datos que

se vayan a utilizar para la predicción. Esta metodología se utiliza en un principio para que ambas tablas de datos sean totalmente independientes. Por el otro lado, se introducen todos los datos con los que se va a trabajar en un único *data.frame*, del cual se extraen dos *subsets* para el entrenamiento y la predicción. Realizando todas las estimaciones con ambos grupos, se comprueba que los resultados son los mismos con ambas metodologías, por lo que la elección de uno u otro no afecta a la realización del proyecto.

Con todas las estimaciones realizadas, los resultados obtenidos se analizarán en el siguiente apartado.

5.3- Análisis de los resultados.

Con los modelos realizados en el apartado anterior, se obtienen numerosos resultados, los cuales conviene estudiar por separado y comparar, de modo que se evalúe cuál de todas las funciones es la óptima para la realización de la predicción en el modelo estudiado.

Para comparar la mayor o menor ajuste de las distintas predicciones realizadas en función de los valores reales de precio para las diferentes horas calculadas, se utilizará el MAPE[19], que mide el tamaño del error absoluto producido en términos porcentuales. La fórmula utilizada para calcular el error es la siguiente:

$$MAPE = \frac{100 \sum_{t=1}^{n} \left| \frac{precio\ real - predicción}{precio\ real} \right|}{n}$$

Para calcular el error máximo y mínimo que se ha producido en cada uno de los casos, se utilizará una fórmula similar a MAPE, pero en el que se calcula el porcentaje de error de cada dato individual.

$$Error\ relativo = \left| \frac{precio\ real - predicción}{precio\ real} \right|$$

La programación utilizada para cada una de las hipótesis propuestas se muestra en el anexo correspondiente, viéndose a continuación los resultados obtenidos en cada caso.

Previsión simple a t+1.

Para la previsión de los precios de la electricidad en t+1, los resultados obtenidos en cada una de las funciones utilizadas se presentan a continuación.

En la figura 21 se presenta una tabla con una muestra de los datos utilizados para la predicción. En él, se muestran de modo individual en verde los datos que se utilizan para la predicción, dando como resultado el precio que debería compararse con el precio sombreado en rosa. De este modo, para realizar la previsión del precio en el tiempo t+1, se utilizarán como datos la previsión de temperaturas y energía renovable generada en t+1, así como el precio de la electricidad en el tiempo t, además de los resultados obtenidos a través del algoritmo *boosting*. Esta selección de datos para cada hora se mantiene en cada una de las funciones utilizadas.

[19] MAPE: Mean Absolute Percentaje Error. Método para la medición del error absoluto en términos porcentuales.

V1	V2	V3	V4	V5	V6	V7
48,55	50,95	12030	157,7	68	112,85	1
40	48,55	12438	157,7	68	112,85	2
33,1	40	11821	157,7	68	112,85	3
28,11	33,1	11101	157,7	68	112,85	4
27,13	28,11	10363	157,7	68	112,85	5
25,24	27,13	10310	157,7	68	112,85	6
19,98	25,24	10483	157,7	68	112,85	7
18,16	19,98	10817	157,7	68	112,85	8
17,73	18,16	11282	157,7	68	112,85	9
19,77	17,73	11721	157,7	68	112,85	10
23,75	19,77	13591	157,7	68	112,85	11
26,03	23,75	15700	157,7	68	112,85	12
27,06	26,03	17058	157,7	68	112,85	13

Figura 21: Datos utilizados para la predicción en t+1.

Mboost.

Los resultados numéricos dados por el MAPE para la función *mboost* demuestran que los errores cometidos son competitivos, resultado que también se aprecia en la gráfica de correlación. Todos estos resultados se muestran a continuación.

Error mínimo	Error máximo	Mediana	MAPE
0,0027%	22,32%	2,78%	4,44%

Tabla 6: Errores obtenidos con la función *mboost* a t+1.

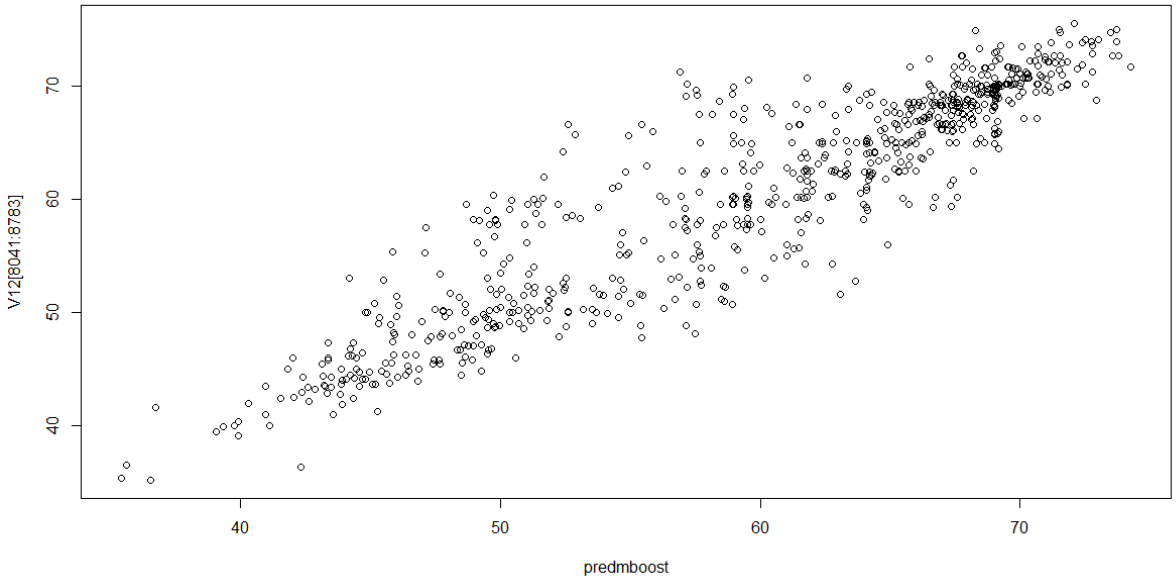

Figura 22: Correlación entre el precio real y el resultado de la predicción en t+1.

En el siguiente gráfico de caja y bigotes se observa que el 75% de los errores cometidos durante la predicción se encuentran entre el 0 y el 5%.

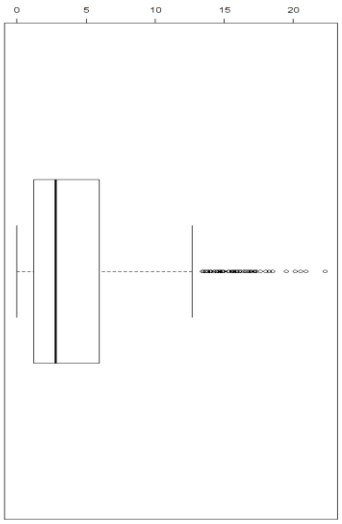

Figura 23: Diagrama de caja y bigotes para los errores de la función *mboost* en t+1.

Blackboost.

En este caso, los errores son similares al caso anterior, con una media de 4,55%. Sin embargo, existe más variabilidad en los resultados, tal y como se ve en la siguiente tabla y en el gráfico de la correlación.

Error mínimo	Error máximo	Mediana	MAPE
0,0022%	21,65%	3,13%	4,55%

Tabla 7: Errores obtenidos con la función *blackboost* a t+1.

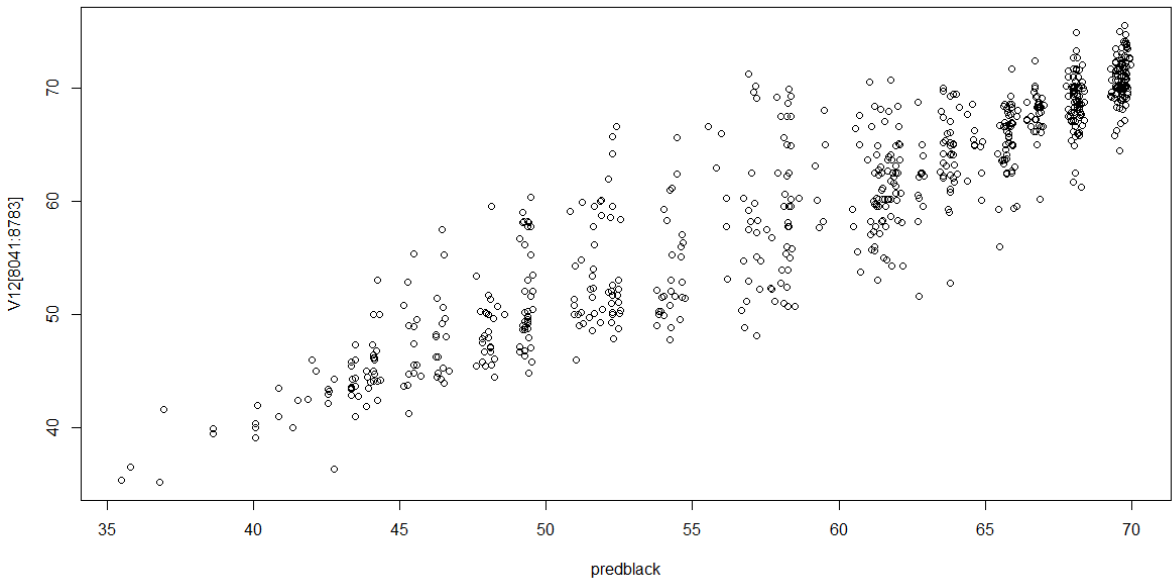

Figura 24: Gráfica de correlación entre los valores de precio reales y los predichos con la función *blackboost* en t+1.

Igual que en caso anterior, con el diagrama de caja y bigotes se muestra que el 75% de los errores cometidos con la función *blackboost* se encuentra de nuevo bajos entre 0 y el 5%.

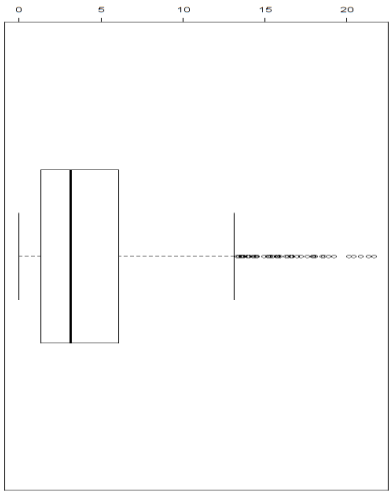

Figura 25: Diagrama de caja y bijotes para los errores de la función *blackboost* en t+1.

Glmboost.

Con la función *glmboost*, al igual que con las anteriores, los errores y variabilidades son similares. Esto se comprueba fácilmente en la gráfica de correlación siguiente.

Error mínimo	Error máximo	Mediana	MAPE
0,002%	22,63%	2,93%	4,47%

Tabla 8: Errores obtenidos con la función *glmboost* a t+1.

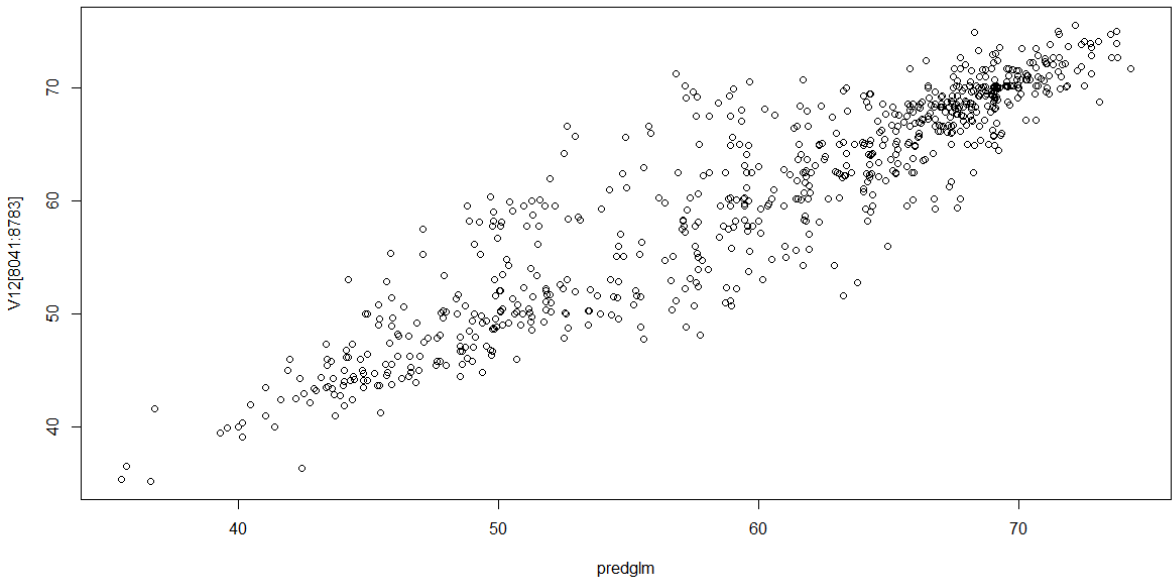

Figura 26: Correlación entre el precio real y el precio predicho con la función *glmboost* en t+1.

En este caso el gráfico de caja y bigotes muestra que la gran parte de los errores se encuentra entre el 0 y el 5%.

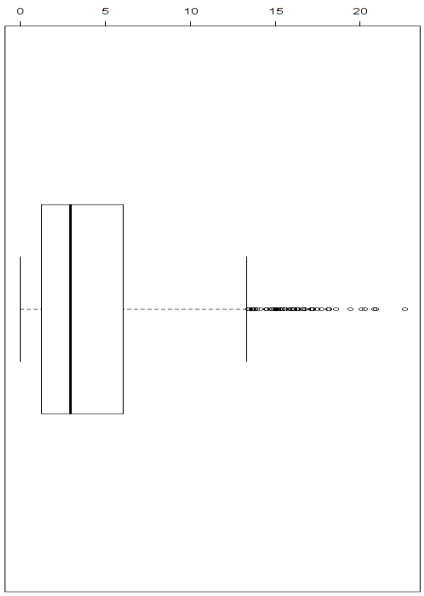

Figura 27: Diagrama de caja y bigotes para los errores de la función *glmboost* en t+1.

***Gbm*.**

En todos los casos, la predicción de precios a través de la función *gbm* resulta poco precisa ya que, si bien simplemente mirando el valor del MAPE ya se comprueba que el error es mucho mayor que en el caso del resto de funciones, analizando la gráfica de la correlación se ve claramente que la misma es inexistente.

Error mínimo	Error máximo	Mediana	MAPE
1,61%	49,03%	38,4%	34,7%

Tabla 9: Errores obtenidos con la función *gbm* a t+1.

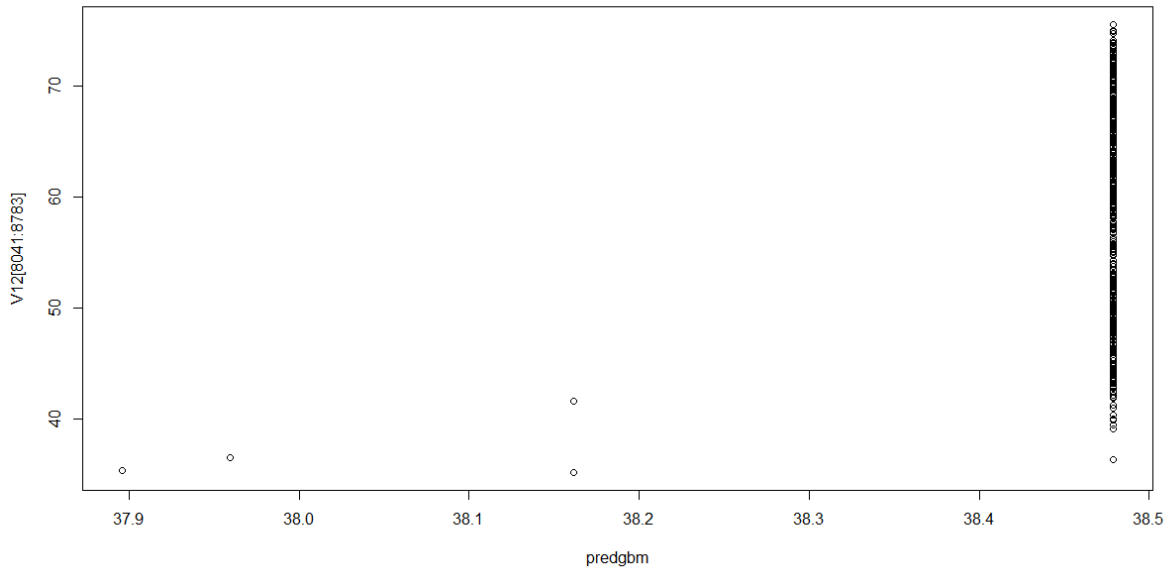

Figura 28: Correlación entre el precio real y el predicho para la función *gbm* en t+1.

En este caso, al contrario que con el resto de funciones utilizadas, se ve a través del gráfico de caja y bigotes que el 75% de errores son mucho mayores, siendo inadecuado su uso para la predicción de precios futuros.

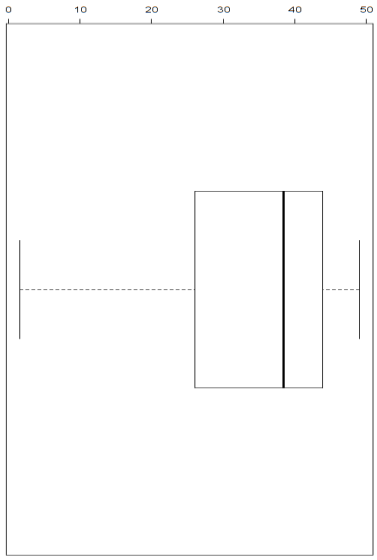

Figura 29: Diagrama de caja y bigotes para la función *gbm* en t+1.

Previsión a t+2.

En la mayoría de las ocasiones resulta más interesante realizar la previsión a tiempos más lejanos. Por ejemplo, en esta ocasión se realizará la previsión para tiempos t+2 horas. En un primer momento, debido a la mayor lejanía de la previsión, se esperarán errores mayores que en el caso anterior.

Al igual que en el caso anterior, en la figura 30 muestra un ejemplo de los datos que se extraerían para la previsión de cada uno de los datos. En este caso, para realizar la predicción del precio en el tiempo 3, cuyo precio final real se muestra en color rosa, se utilizarían los datos sombreados en verde. De este modo, para predecir el precio de la electricidad en el tiempo t+2, se utilizarán las predicciones de temperaturas y energía renovable generada en los tiempos t+2, así como el precio de la electricidad en t y los valores obtenidos tras la utilización de los algoritmos de *boosting*.

V1	V2	V3	V4	V5	V6	V7
48,55	50,95	12030	157,7	68	112,85	1
40	48,55	12438	157,7	68	112,85	2
33,1	40	11821	157,7	68	112,85	3
28,11	33,1	11101	157,7	68	112,85	4
27,13	28,11	10363	157,7	68	112,85	5
25,24	27,13	10310	157,7	68	112,85	6
19,98	25,24	10483	157,7	68	112,85	7
18,16	19,98	10817	157,7	68	112,85	8
17,73	18,16	11282	157,7	68	112,85	9
19,77	17,73	11721	157,7	68	112,85	10
23,75	19,77	13591	157,7	68	112,85	11
26,03	23,75	15700	157,7	68	112,85	12
27,06	26,03	17058	157,7	68	112,85	13

Figura 30: Conjunto de datos seleccionados para la predicción en t+2.

Mboost.

Tal y como se había previsto en un primer momento, los nuevos valores del precio calculados a t+2 suponen un aumento del error cometido, ascendiendo al 8%, tal y como se ve en la tabla 10.

Error mínimo	Error máximo	Mediana	MAPE
0,01%	34,77%	5,46%	8,12%

Tabla 10: Errores obtenidos con la función *mboost* a t+2.

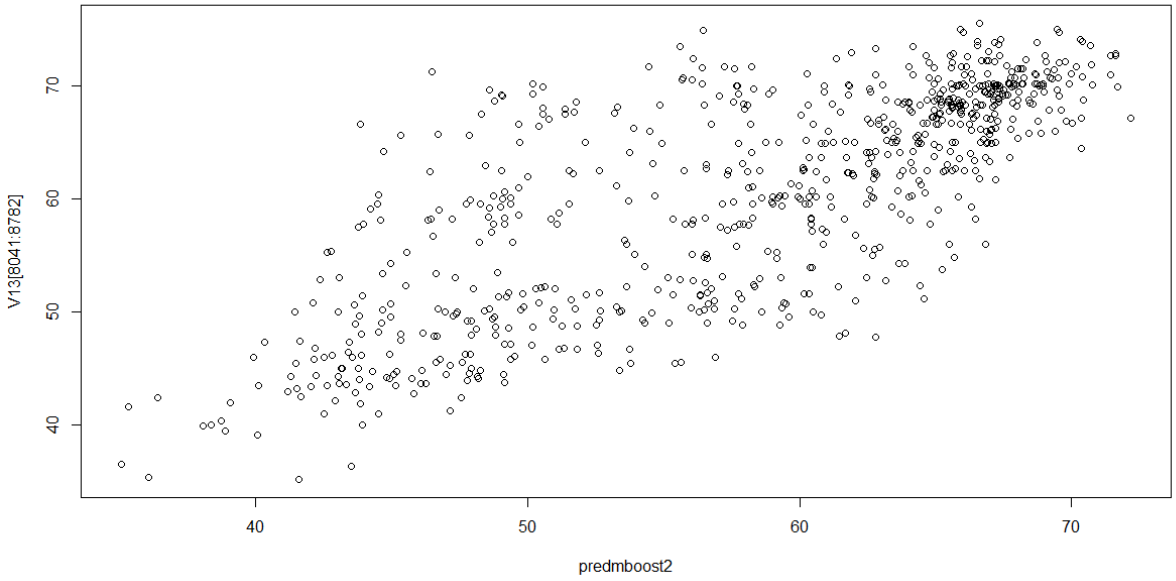

Figura 31: Correlación entre el precio real y el resultado de la predicción con *mboost* en t+2.

En este caso, analizando el gráfico de caja y bigotes que se muestra a continuación, se observa que el 50% de los errores cometidos se encuentran entre el 2 y el 12%, con lo que se afirma y aumento respecto de la previsión a t+1.

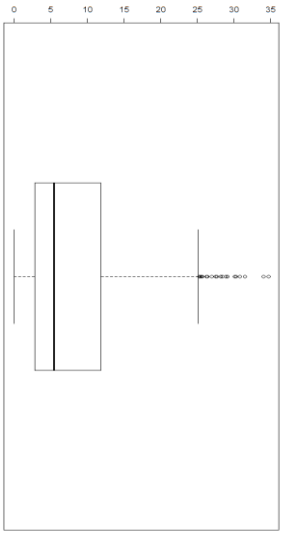

Figura 32: Diagrama de caja y bigotes para la función *mboost* en t+2.

Blackboost.

Con el uso de la función *blackboost* aumenta ligeramente el error cometido, si bien tanto el error como la variabilidad continúan siendo bastante similares. Esto puede comprobarse en la tabla 11.

Error mínimo	Error máximo	Mediana	MAPE
0,018%	37,99%	6,5%	8,55%

Tabla 11: Errores obtenidos con la función *blackboost* a t+2.

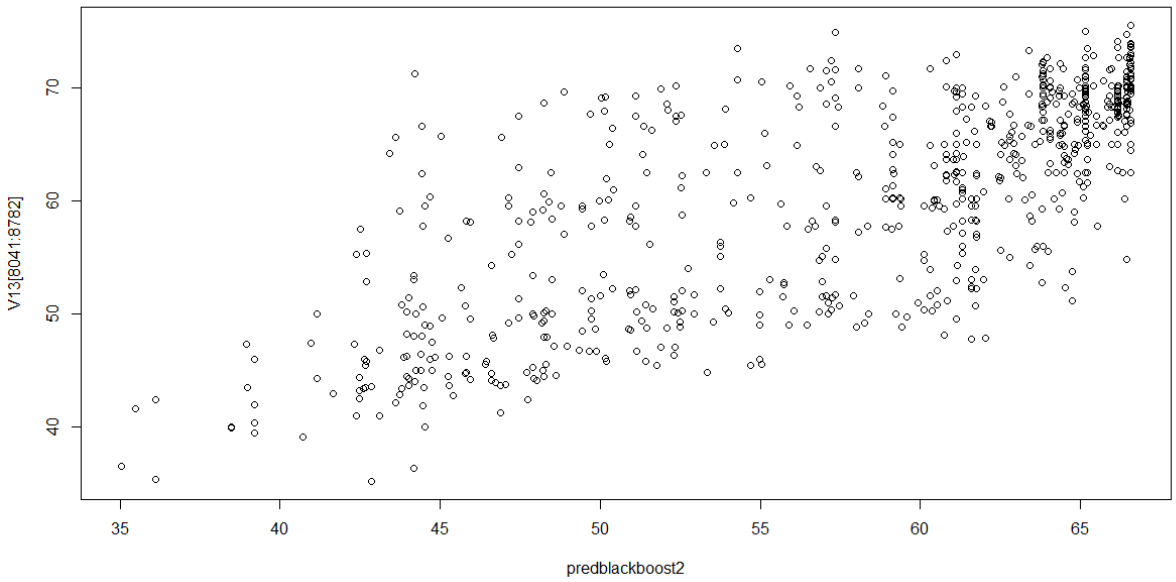

Figura 33: Correlación entre el precio real y el resultado de la predicción con *blackboost* en t+2.

Igual que en el caso de la función *blackboost*, se observa de nuevo en la gráfica de caja y bigotes que el 50% de los errores se encuentra entre el 2 y 12%.

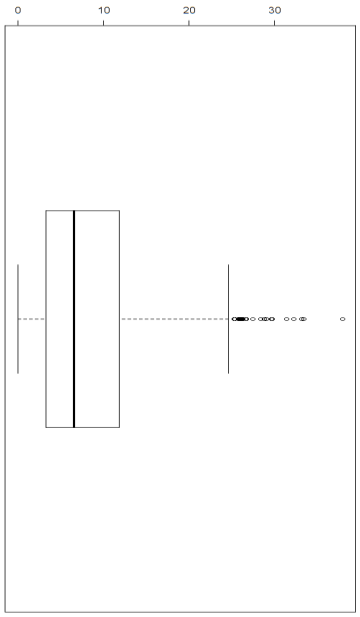

Figura 34: Diagrama de caja y bigotes para la finción *blackboost* en t+2.

Glmboost.

Los resultados obtenidos con la función *glmboost* se observan a numéricamente en la tabla 12, en el cual se ven que los errores cometidos son ligeramente inferiores que en el caso anterior. Estos valores se verán complementados por la figura 35, que muestra la correlación existente en los resultados.

Error mínimo	Error máximo	Mediana	MAPE
0,01%	35,24%	5,58%	8,28%

Tabla 12: Errores obtenidos con la función *glmboost* a t+2.

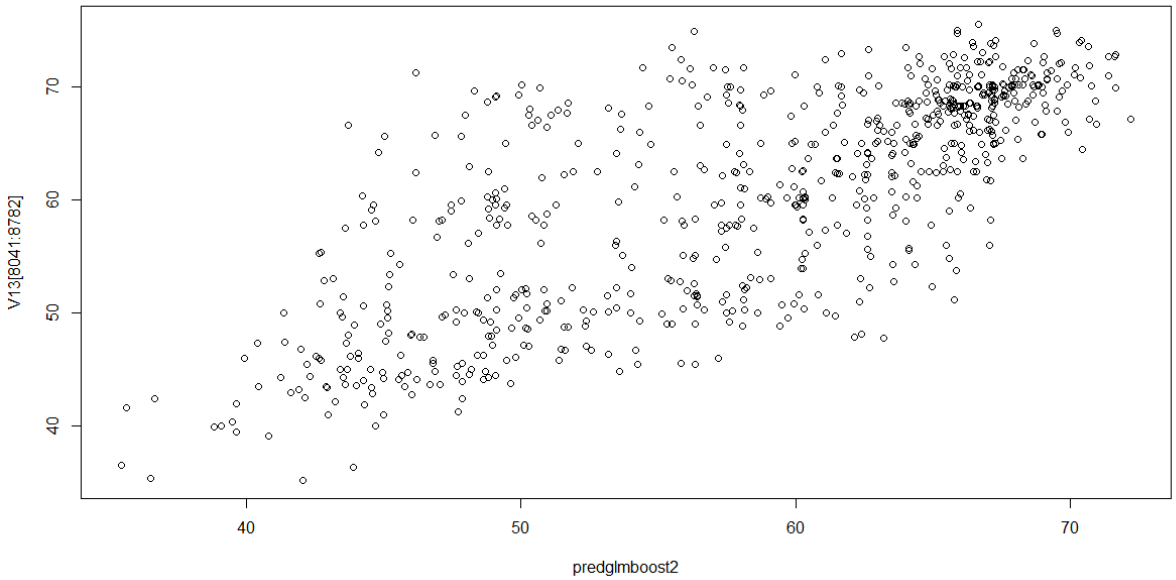

Figura 35: Correlación entre el precio real y el resultado de la predicción con *glmboost* en t+2.

Por último, en esta función del paquete *mboost* se mantiene la tendencia en cuanto al valor del error, de modo que el 50% de los datos tienen un valor entre el 2 y el 12%.

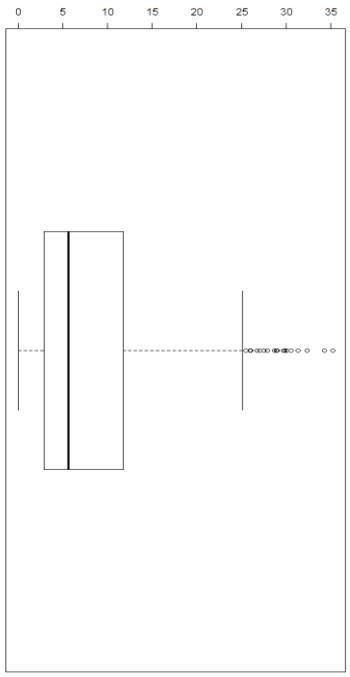

Figura 36: Diagrama de cajas y bigotes para la función *glmboost* en t+2.

***Gbm*.**

Como en el caso t+1, el error cometido con la función *gbm* es grande, de modo que comprobando con los resultados del resto de funciones, el uso de esta función podría ser descartada del estudio.

Error mínimo	Error máximo	Mediana	MAPE
1,74%	49,1%	38,52%	34,79%

Tabla 13: Errores obtenidos con la función *gbm* a t+2.

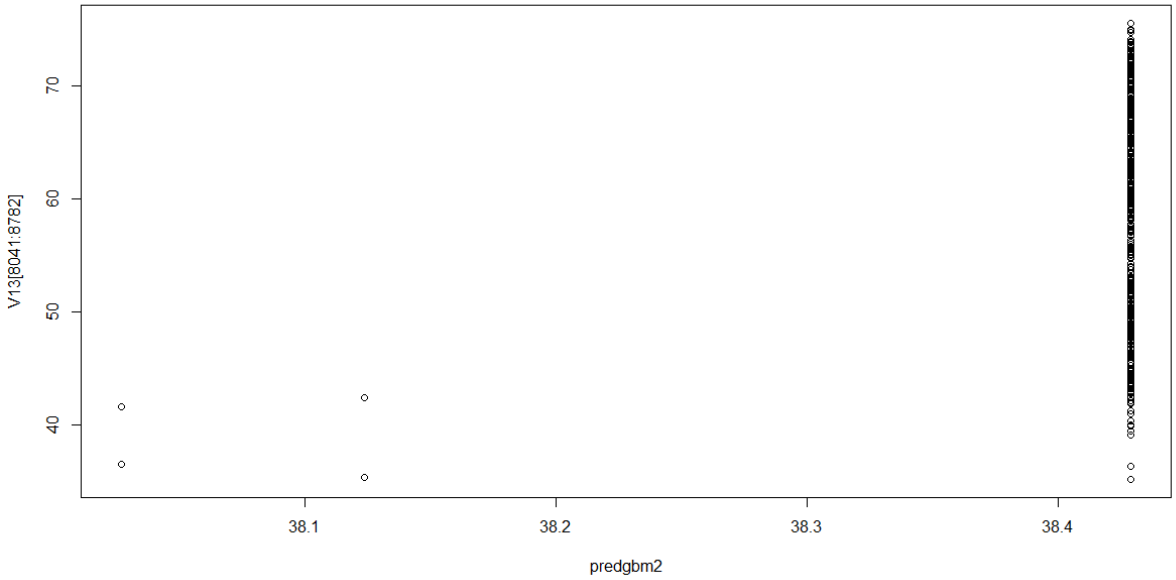

Figura 37: Correlación entre el precio real y el resultado de la predicción con *gbm* en t+2.

La tendencia de los errores en la función propia del paquete *gbm* se mantiene con los valores similares a la predicción de t+1, con la diferencia de que en este caso la mayoría de los datos cometen un error del 38%.

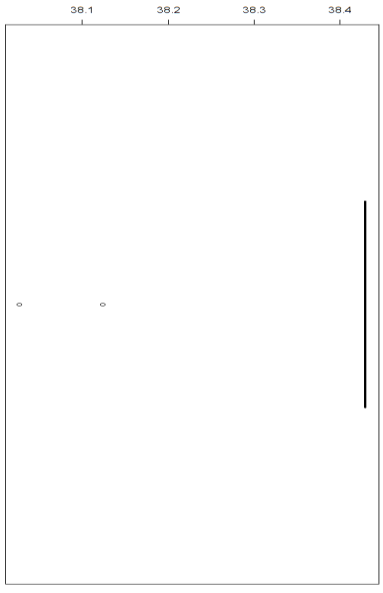

Figura 38: Diagrama de cajas y bigotes para la función *gbm* en t+2.

Previsión a t+3.

Como última previsión del proyecto, se comprobará lo que ocurre con los precios si se calculan a tres horas vista. Igual que en los casos anteriores, se evaluarán los resultados para la utilización de las cuatro funciones elegidas en los casos anteriores.

Antes de nada, y al igual que en los dos casos anteriores, en la figura 39 se muestra la utilización de los diferentes datos de modo gráfico. Así, atendiendo a la muestra, el precio que se obtenga de la previsión se tendrá que comparar con el real sombreado en rosa. Para ello, se utilizarán los resultados del *boosting* realizado, las previsiones de las temperaturas y la energía renovable en t+3 y el precio de la electricidad para t.

V1	V2	V3	V4	V5	V6	V7
48,55	50,95	12030	157,7	68	112,85	1
40	48,55	12438	157,7	68	112,85	2
33,1	40	11821	157,7	68	112,85	3
28,11	33,1	11101	157,7	68	112,85	4
27,13	28,11	10363	157,7	68	112,85	5
25,24	27,13	10310	157,7	68	112,85	6
19,98	25,24	10483	157,7	68	112,85	7
18,16	19,98	10817	157,7	68	112,85	8
17,73	18,16	11282	157,7	68	112,85	9
19,77	17,73	11721	157,7	68	112,85	10
23,75	19,77	13591	157,7	68	112,85	11
26,03	23,75	15700	157,7	68	112,85	12
27,06	26,03	17058	157,7	68	112,85	13

Figura 39: Selección de datos para la previsión en t+3.

Mboost.

Tal y como se preveía antes de realizar el estudio, el error medio que se produce para las predicciones a t+3 son superiores a las obtenidas para tiempos más cortos, aumentando aproximadamente del 8 al 11%. Un resumen de los resultados obtenidos se encuentra en la tabla 14 y figura 40.

Error mínimo	Error máximo	Mediana	MAPE
0,03%	39,32%	8,37%	11,33%

Tabla 14: Errores obtenidos con la función *mboost* a t+3.

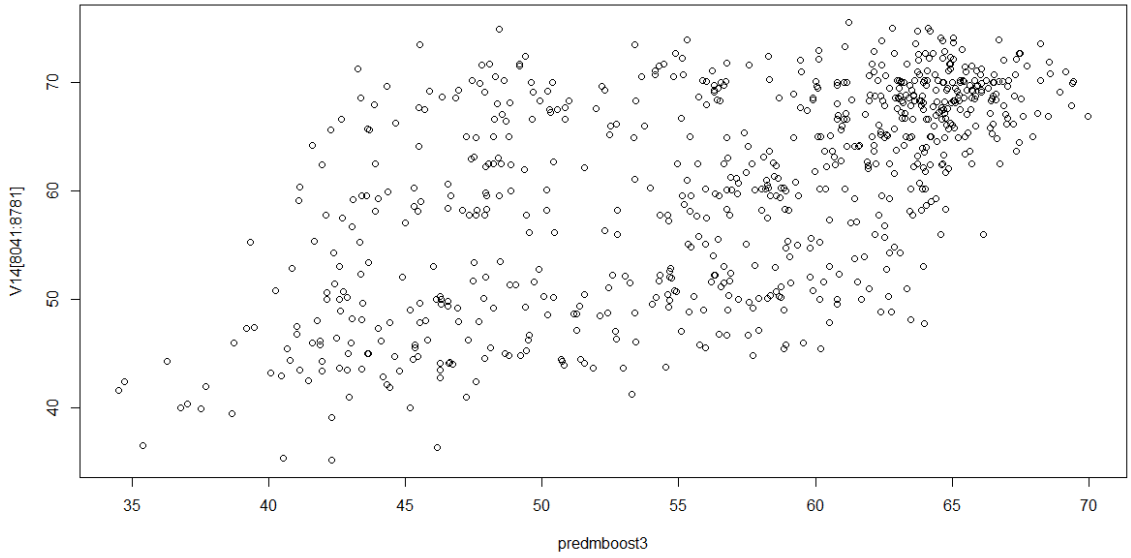

Figura 40: Correlación entre el precio real y el resultado de la predicción con *mboost* en t+3.

Los resultados obtenidos en este caso no muestran un aumento mucho mayor que en el caso de t+2 para la misma función. Si bien el MAPE ha aumentado a un 11,33% en este caso, observando el gráfico de caja y bigotes que se encuentra a continuación se ve que el 50% de los errores se encuentran entre el 5 y el 15%.

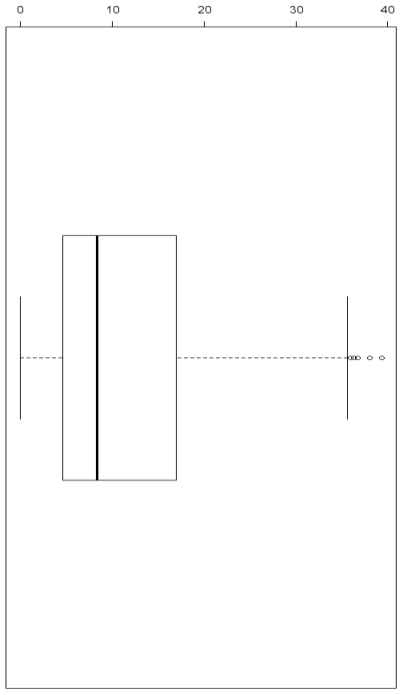

Figura 41: Diagrama de caja y bigotes para la función *mboost* en t+3.

Blackboost.

En este caso, la utilización de la fórmula *blackboost* presenta un ligero aumento de los errores máximo, mínimo y medio respecto de la fórmula *mboost*. Sin embargo, calculando el MAPE, éste resulta ser ligeramente inferior. Dichos resultados se pueden ver en la tabla 15 y la figura 42.

Error mínimo	Error máximo	Mediana	MAPE
0,06%	39,66%	8,46%	10,96%

Tabla 15: Errores obtenidos con la función *blackboost* a t+3.

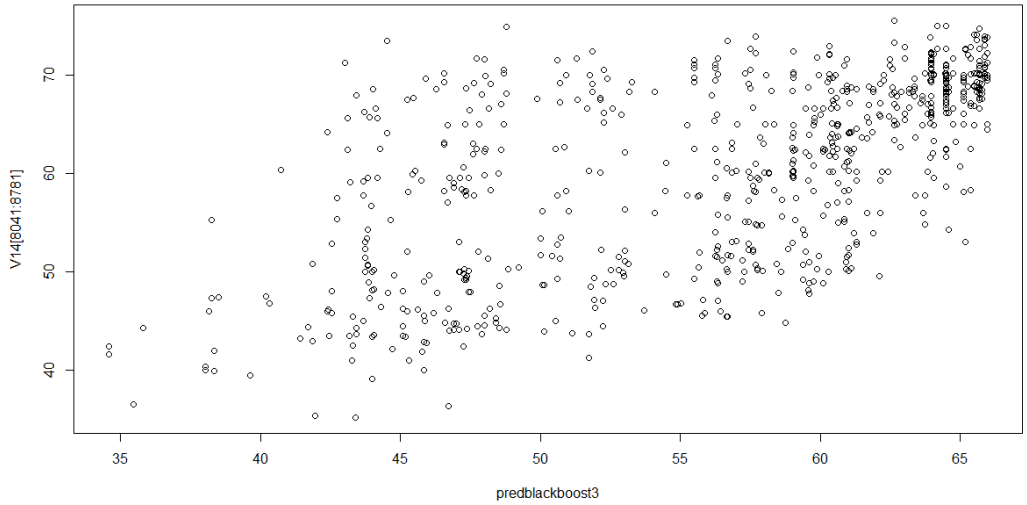

Figura 42: Correlación entre el precio real y el resultado de la predicción con *blackboost* en t+3.

En la siguiente gráfica de caja y bigotes se ve como el 50% de los errores cometidos se encuentra entre el 4 y el 15%.

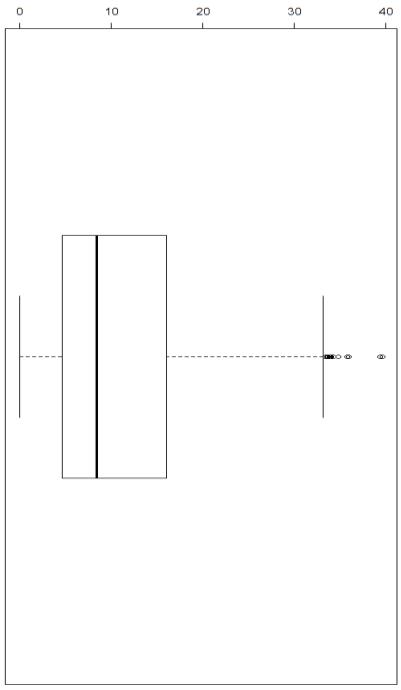

Figura 43: Diagrama de caja y bigotes para la función *blackboost* en t+3.

Glmboost.

La función *glmboost* es la única que consigue un error mínimo del 0%. Sin embargo, se trata de la función dentro del paquete *mboost* que obtiene unos errores ligeramente superiores al resto.

Error mínimo	Error máximo	Mediana	MAPE
0%	39,98%	8,6%	11,56%

Tabla 16: Errores obtenidos con la función *glmboost* a t+3.

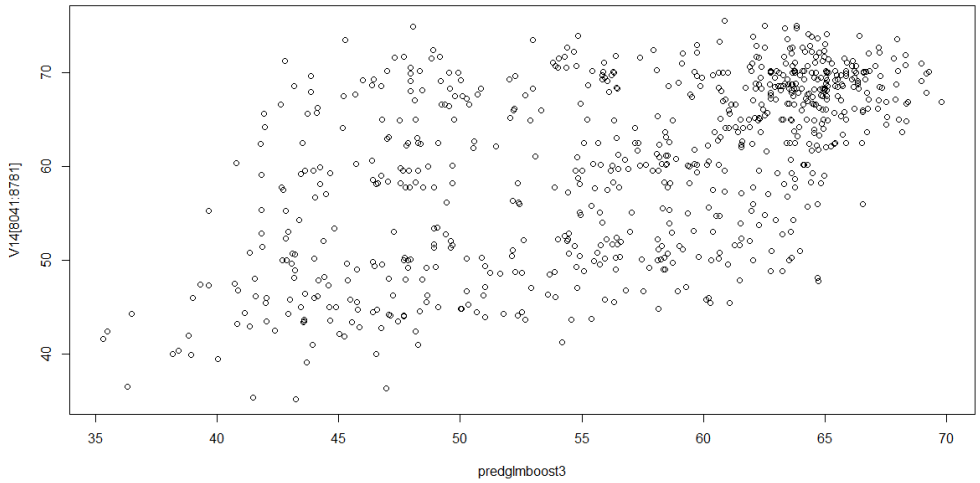

Figura 44: Correlación entre el precio real y el resultado de la predicción con *glmboost* en t+3.

En el diagrama de caja y bigotes muestra que, al igual que en caso de las otras funciones dentro del paquete, el 50% de los errores cometidos dentro de la previsión se encuentran entre el 4 y el 16%.

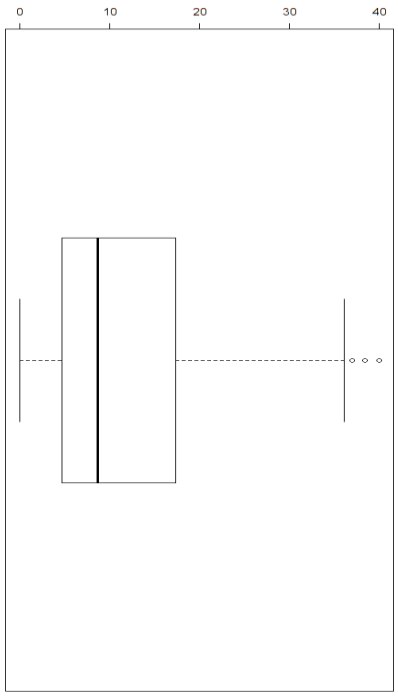

Figura 45: Diagrama de caja y bigotes para la función *glmboost* en t+3.

***Gbm*.**

La función *gbm* vuelve a ser en este caso la que aporta unos errores mayores con un MAPE aproximado del 35,9%, tal y como se ve en el siguiente resumen de resultados.

Error mínimo	Error máximo	Mediana	MAPE
1,86%	49,16%	38,6%	34,89%

Tabla 17: Errores obtenidos con la función *gbm* a t+3.

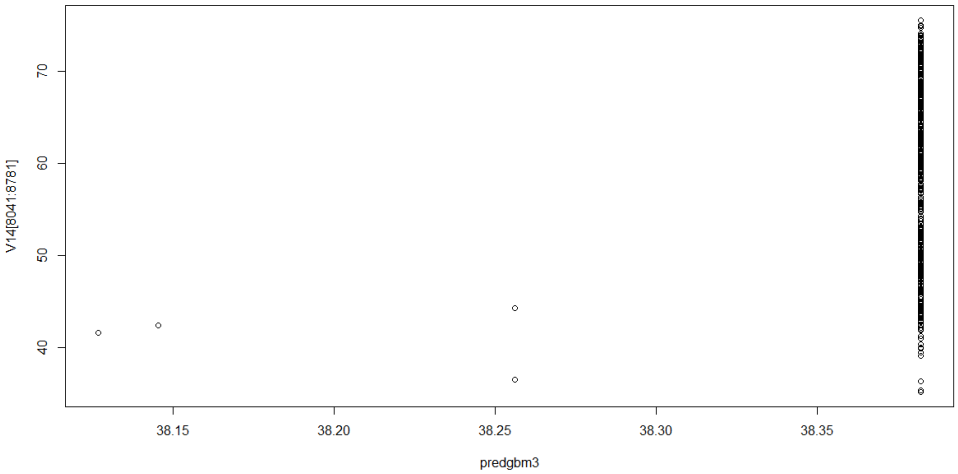

Figura 46: Correlación entre el precio real y el resultado de la predicción con *gbm* en t+3.

En este caso, en la gráfica de caja y bigotes que se muestra a continuación se ve como la gran parte de los errores cometidos se encuentra alrededor de 39%.

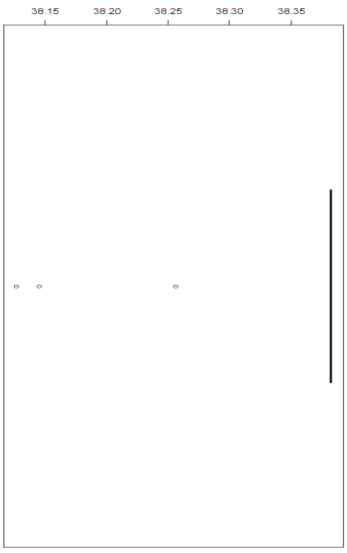

Figura 47: Diagrama de cajas y bigotes para la función *gbm* en t+3.

Como ya se ha comentado anteriormente, uno de los objetivos, aparte de conseguir una predicción de los precios con el menor margen de error posible, consiste también en el estudio de un nuevo modelo de predicción, como es el *boosting*. Desde un primer momento se presentaba la metodología *boosting* como una variación respecto del anterior *Random Forest*, pero sin embargo todavía no existen muchos estudios realizados con el mismo.

En cuanto al modelo *Random Forest*, se han realizado anteriormente estudios que conseguían realizar una predicción en el precio de la electricidad en base a las variables que se consideran determinantes, con un error medio del 2,4% para horizontes de t+1. De este modo, será interesante realizar comparaciones entre los resultados obtenidos entre ambas metodologías, con el objetivo de determinar qué metodología resulta más acertado utilizar en presente proyecto.

5.4- Comparación de los resultados obtenidos.

Una vez realizadas todas las estimaciones pertinentes, se puede entrar a valorar los resultados obtenidos y sacar conclusiones respecto a ellos. Si se analizan dichos resultados desde el plano interno, y analizando únicamente la validez de los algoritmos aportados por la metodología *boosting*, se observa que el uso de las tres funciones incluidas en el paquete *mboost* son las que aportan unos mejores resultados, con un error medio de 4,5% aproximadamente para tiempos t+1, lo que resulta una predicción bastante acertada. Por otro lado, y como se esperaba desde un primer momento, los errores cometidos para t+2 y t+3 aumentan.

Los errores cometidos por la función *gbm*, sin embargo, son mucho mayores que las del paquete *mboost* en todos los casos, pero similares entre sí, con una media del 34,7%. Este valor resulta demasiado alto para la predicción de un valor a corto plazo, lo que haría que su uso no fuera fiable en ningún caso. Debido a la gran diferencia entre los resultados de ambos paquetes, será

interesante realizar estudios posteriores que permitan analizar las causas de dicha diferencia y solventarla si fuera posible.

En la siguiente tabla, se observa el MAPE de estos doce escenarios, con el menor error en cada uno de remarcado. En dicho catálogo, se observa que tanto para t+1 como para t+2 la función que comete el menor error es *mboost*, mientras que para t+3 la tendencia cambia cometiendo un error menor la función *blackboost*.

	t+1	**t+2**	**t+3**
Mboost	**4,44**	**8,12**	11,33
Blackboost	4,55	8,55	**10,96**
Glmboost	4,47	8,28	11,56
Gbm	34,7	34,79	34,89

Tabla 18: Comparación de los resultados para los tres primeros tiempos con cada una de las funciones utilizadas.

En un intento de estudiar lo que sucede a más largo plazo, se realizan las predicciones para tiempos más lejanos. Si estos resultados se visualizan a través de una gráfica, se observa que el error medio cometido por las funciones utilizadas incluidas en el paquete *mboost* tienen una tendencia más o menos uniforme. La misma sufre una subida en los dos primeros tiempos, estabilizándose aproximadamente para el tiempo t+7, de manera que en todos los casos sufren ciertas oscilaciones, pero con unos errores cometidos del 18±2%. En este caso también se observa que la que consigue mejores resultados es la función *blackboost*, seguido por *mboost* y *glmboost* respectivamente. Este dato resulta relevante, ya que asegura una predicción bastante fiable a largo plazo, la cual podría dar una idea de los valores que podrían presentar, los cuales se irían ajustando conforme la fecha que se quiere predecir sea más cercana.

Por otro lado, en la gráfica también es sencillo ver la diferencia con la función *gbm*, la cual en todo caso sufre la misma tendencia en cuanto a la magnitud del error, del 34% aproximadamente, muy alejado del resto de funciones. Esto nos indica que, a la hora de realizar predicciones ajustadas, en ningún caso se elegiría el uso del paquete *gbm*.

Los valores a largo plazo comentados para cada una de las funciones estudiadas, y que reflejan lo comentado anteriormente, se recogen en el siguiente gráfico.

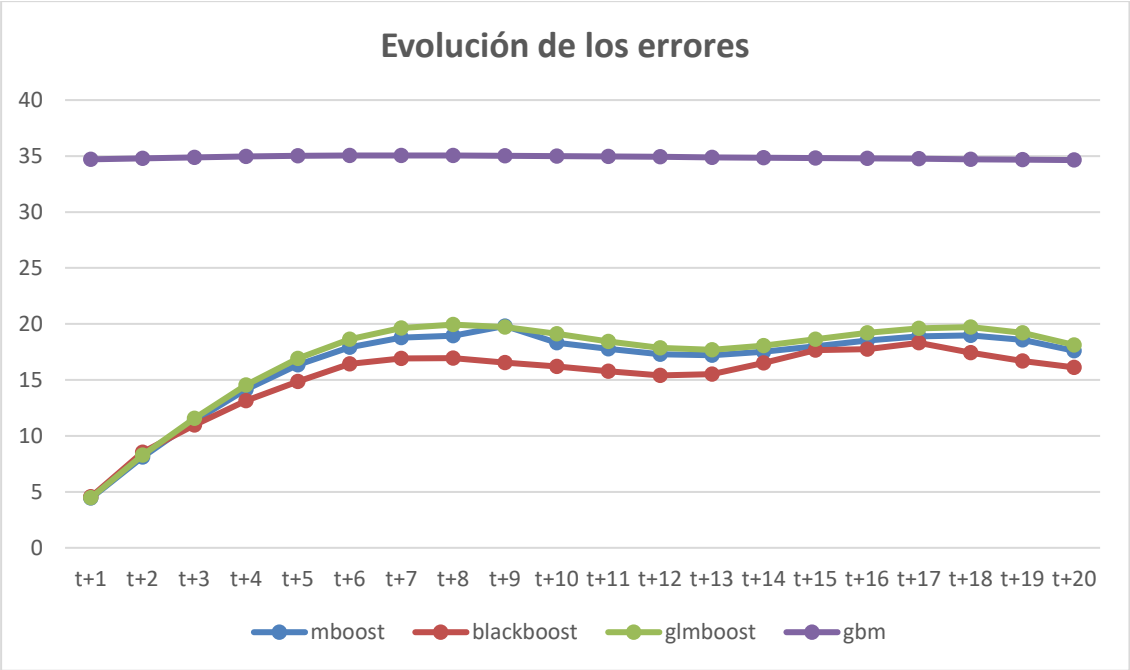

Figura 48: Evolución de los errores para las distintas funciones.

Por otro lado, es interesante comparar los resultados obtenidos con los algoritmos *boosting* con los errores cometidos con otras metodologías de minería de datos en las mismas circunstancias para horizontes t+1. En este caso se comparan los valores medios del MAPE obtenidos con algoritmos de *boosting*, con los obtenidos con CART, *Random Forest* y *bagging*, cuyos valores se observan en la siguiente tabla[20].

	MAPE
Boosting	**4,49**
CART	4,50
Random Forest	**2,36**
Bagging	3,60

Tabla 19: Tabla con los valores del MAPE para los distintos algoritmos.

Si se comparan los valores, se observa que, si bien las predicciones obtenidas en el presente proyecto a través de *boosting* son muy buenos, con unos errores del 4,49%, estos valores se alejan de los sacados gracias al *Random Forest*, que resultan ser los mejores en este caso. De este modo, si se necesita obtener los resultados más acertados, la tendencia en un primer momento sería la de la utilización de *Random Forest*.

[20] Fuente: Isabel Juárez Barrios (2013). *Predicción del precio de la energía eléctrica utilizando modelos de Minería de Datos: árboles de clasificación y regresión, Random Forest y bagging.* Escuela Técnica Superior de Ingeniería Industrial, Madrid, España (Proyecto de fin de carrera).

6- CONCLUSIONES

En este proyecto se ha hecho un estudio del uso del algoritmo *boosting* para la predicción de los precios de la electricidad en base a determinadas variables específicas que se consideran determinantes para la estimación de dicho precio.

Con lo estudiado en el proyecto, se ha valorado si el uso de la metodología *boosting* es adecuado para los casos con este tipo de datos, consiguiendo unos porcentajes de error bajos que puedan ser considerados lo suficientemente precisos como para su utilización en desarrollos posteriores.

En cuanto al procedimiento utilizado para la estimación de los precios, se han utilizado varias funciones dentro del algoritmo *boosting*, lo que permite valorar si todos ellos son igualmente válidos o si existen diferencias significativas entre ellos que determinen que el uso de unas sea más beneficios que otros.

Para la realización de las estimaciones se ha utilizado el software estadístico R, el cual está considerado como uno de los más interesantes debido a la variedad de métodos estadísticos que cubre gracias a la adición de paquetes específicos. Para la realización de este proyecto, se utilizan dos paquetes específicos de *boosting*, *mboost* y *gbm*. El proyecto permitirá conocer si el uso de ambos paquetes es adecuado para el estudio de la base de datos generada.

Los algoritmos de *boosting* comentados existentes en R comentados anteriormente se ha usado para hacer el estudio sobre una base de datos determinada, constituida por: el precio de la electricidad en la hora t, el precio de la electricidad en la hora t+1, la cantidad de energía renovable generada, temperatura máxima, temperatura mínima y temperatura media. Se almacenan los datos pertenecientes a cada una de las variables generados cada hora. Cuando ya se tengan todos los datos, se utilizarán los pertenecientes a los primeros 11 meses para realizar el training con los algoritmos, utilizando los pertenecientes al último mes para la estimación de precios futuros.

Se ha obtenido que para la utilización del *boosting*, el paquete *mboost* y la función *mboost* es el que mejores resultad da para horizontes a corto plazo. Sin embargo, esto cambia para horizontes más lejanos, donde la función que obtiene menores errores en la estimación de precios de la electricidad pasa a ser *blackboost*.

Los errores obtenidos finalmente son similares a los obtenidos por otras técnicas de árboles de decisión como *Random Forest*, en torno al 5 % para horizontes t+1, si bien los errores que se encuentran con el uso de este último algoritmo siguen siendo más bajos. Por otro lado, para el estudio de los errores cometidos para horizontes más lejanos, se observa que los errores cometidos por los algoritmos de *boosting* se estabilizan para t+8, con un error medio para las funciones propias del paquete *mboost* del 18%.

La realización de este proyecto ha conseguido arrojar datos relevantes, como son:

- El estudio de una metodología alternativa que estudia los datos con una mayor profundidad.
- Comparación de la utilización de este tipo de metodologías con otras diferentes, de modo que permita estudiar la validez de uno respecto del otro en función de las variables utilizadas.
- Unos resultados que presentan unos errores de predicción bajos, que aseguran un buen conocimiento de datos futuros.

Entre las desventajas encontradas, sin embargo, se encontrarían la imposibilidad de obtener los árboles de regresión gráficamente con las funciones aportadas por R o el hecho de que los errores sean mayores que con otras metodologías. Estas desventajas serían ámbitos interesantes de cara a futuras investigaciones.

7- LÍNEAS FUTURAS

Una vez analizados los resultados obtenidos en el proyecto, resulta interesante valorar las posibles líneas futuras que podrían seguirse para continuar con el estudio y desarrollo del algoritmo *boosting*.

Para ello, se deberán probar otras situaciones en las que se han utilizado otros métodos, de manera que se pueda analizar en cuáles de ellas los valores que se obtengan sean adecuados con la utilización de los nuevos algoritmos. Además, será necesario continuar el desarrollo de los mismos, de manera que el análisis de las variables y su utilización sea aún más exhaustivo, consiguiendo que las futuras predicciones sean más ajustadas de lo que son en la actualidad.

Por otro lado, atendiendo al estudio de los precios de la electricidad, se deberá seguir estudiando el desarrollo de nuevos algoritmos, así como la valoración de posibles nuevas variables que sean las que influyen en el precio de la energía eléctrica, haciendo que los programas que utilicen las diferentes empresas para realizar las distintas predicciones aporten cada vez soluciones mejores, que permitan a su vez optimizar también otros factores, como el conocimiento de la energía necesitada o la cantidad de energía renovable que se genere.

8- PLANIFICACIÓN TEMPORAL Y PRESUPUESTO

8.1- Planificación temporal.

A lo largo de este apartado se realizará una contabilización de las actividades realizadas para la ejecución del proyecto, así como el tiempo empleado en cada una de ellas. Además, con el objetivo de facilitar la comprensión de la distribución temporal del trabajo, más adelante se pondrá un Diagrama de Gantt que represente la realización del proyecto. Merece la pena indicar que, debido a la inexperiencia con la utilización del modelo de *boosting* con el programa R, fue necesario emplear un tiempo en la realización de diferentes modelos más sencillos del mismo para estudiar su funcionamiento, tal y como se ve en la siguiente tabla.

8.1.1- Actividades realizadas.

Para la realización del proyecto se emplearon 517 horas distribuidas a lo largo de seis actividades principales. La cantidad de tiempo empleada en cada una de esas actividades mencionadas se encuentran reflejados de modo resumido en la siguiente tabla:

Actividad	Duración
Investigación	223 horas
Búsqueda de datos	32 horas
Manejo del programa R	56 horas
Simulaciones con el programa	85 horas
Predicción	61 horas
Redacción de la memoria	60 horas

Tabla 20: Actividades realizadas a lo largo del proyecto.

8.1.2- Diagrama de Gantt.

La distribución del trabajo a lo largo del proyecto se ve reflejado en el siguiente Diagrama de Gantt.

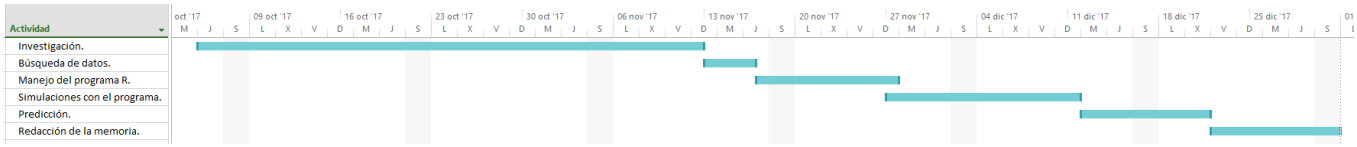

Tabla 21: Diagrama de Gantt del desarrollo del proyecto.

8.2- Presupuesto.

Para la realización del proyecto, se debe en primer lugar establecer un presupuesto en el que se contabilicen los posibles gastos incurridos que va acarrear el mismo. Con ello, se pretende establecer de modo aproximado la cantidad monetaria necesaria para llevar a cabo una investigación con las características dadas. Existen distintos apartados de gastos, siento los mismos: gastos de materiales o de equipamiento (material fungible), gastos de personal y costes indirectos.

8.2.1- Material fungible.

Dentro de este apartado, se encontrarán todos los costes derivados de la compra del distinto material que haya sido necesario para la realización del proyecto, así como la adquisición de los distintos programas informáticos que haya sido utilizados durante las diferentes operaciones realizadas a lo largo de la duración del proyecto.

Asimismo, teniendo en cuenta que algunos de los distintos equipamientos no van a ser utilizados exclusivamente para el proyecto, ya que su vida útil es superior, se tendrán en cuenta únicamente el coste de amortización de cada uno de los elementos derivados de su uso en estos casos. Los costes existentes en este apartado son los siguientes.

Ordenador portátil: para la realización de este proyecto se necesita la compra de un ordenador portátil. Realizando una comparación del mercado y atendiendo a las necesidades de memoria y procesador necesitados para el presente proyecto, se compra un Acer Aspire ES1-571-5074 por 449 €. Teniendo en cuenta que la vida útil de un ordenador de estas características se encuentra alrededor de los 5 años, y que la duración del proyecto es aproximadamente de medio año, el coste del ordenador para la realización del proyecto ha sido de aproximadamente 45 €.

Impresión y encuadernación del proyecto: la impresión del proyecto es responsable de dos costes distintos. Por un lado, la impresión propia del mismo. Teniendo en cuenta la extensión y características (impresión a color) aproximadas que tendrá el documento, se estima que su coste se encontrará en torno a 50 €. Por otro lado, al tratarse de una encuadernación con un formato especial, su coste se estima en 25 €. De este modo, la realización de la copia física del proyecto finalizado tendría un coste de 75 €.

Software estadístico R: se trata de un software libre perteneciente al sistema operativo GNU[21], por lo que su implantación en el ordenador ha sido completamente gratuita (coste de 0 €).

Licencia de Microsoft Office 2016: la compra de la licencia de Microsoft Office 2016 incluye los programas Microsoft Word, Microsoft Excel y Microsoft Project, todos ellos utilizados en la realización del presente proyecto. Su coste anual ha ascendido a 253,90 €. Dado que la utilización del paquete debido al proyecto ha sido del 25%, el coste aplicado al proyecto ha sido de 63,48 €.

8.2.2- Mano de obra.

En base a los costes derivados de la dedicación de horas y conocimiento para la realización del proyecto, conviene destacar la existencia de dos tipos distintos de coste. Por un lado, las horas trabajadas de manera personal, y el coste producido por las horas dedicadas al proyecto por parte de los dos directores del mismo. De este modo, las horas dedicadas de manera personal tendrán un coste de 5 €/hora, mientras que el coste por parte de cada uno de los directores será de 15 €/hora.

8.2.3- Costes indirectos.

Dentro de este apartado se encuentran contabilizados todos aquellos costes que no se pueden contabilizar de un modo sencillo, así como los posibles costes imprevistos que no se puedan tener en cuenta desde un primer momento. Entre los costes incluidos en este apartado se

[21] GNU: Sistema operativo de tipo Unix, formado en su totalidad por software libre.

encuentran los debidos a la electricidad necesaria para alimentar los equipos, así como del uso de internet. Estos costes se contabilizarán como un 10% sobre el coste subtotal del proyecto.

Concepto	Cantidad	Precio unitario	Precio total
Ordenador portátil	1	45 €	45 €
Impresión del proyecto	1	75 €	75 €
Software estadístico R	1	0 €	0 €
Licencia de Microsoft Office 2016	1	63,48 €	63,48 €
Horas de dedicación al proyecto	517	5 €	2.585 €
Horas de dedicación del proyecto por parte de los directores de proyecto	116	15 €	1.740 €
Subtotal			4.508,48 €
Costes indirectos	1	450,85 €	450,85 €
Coste total			4959,33 €

Tabla 22:Presupuesto asociado a la realización del proyecto.

ANEXOS

Anexo: Sentencias de R.

Teniendo en cuenta que la programación para cada una de las funciones es similar, se pondrá en cada uno de los estudios el código de una de las funciones, de manera que de tal modo queda explicada la programación del resto.

Introducción de los datos en el sistema.

Como ya se ha comentado anteriormente, las estimaciones se realizaron a través de dos tipos de datos. Por un lado, la introducción de un único *data.frame*, del cual se extraerán los *subsets* necesarios para hacer las estimaciones. Por el otro lado, con un *data.frame* por cada conjunto de datos que vaya a utilizarse.

Para un único *data.frame*, la introducción de datos es la siguiente:

```
datosTotal<-read.table("DatosTotal.txt")
attach(datosTotal);
```

Para la introducción de dos *data.frames* distintos, la introducción de datos es la siguiente:

```
precio<-read.table("precio.txt")
energia<-read.table("energia.txt")
precioA<-read.table("precioA.txt")
tmax<-read.table("tmax.txt")
tmin<-read.table("tmin.txt")
tmed<-read.table("tmed.txt")
datos<-data.frame(precio,precioA,energia,tmax,tmin,tmed)
preciopred<-read.table("preciopred.txt")
precioApred<-read.table("precioApred.txt")
energiapred<-read.table("energiapred.txt")
tmaxpred<-read.table("tmaxpred.txt")
tminpred<-read.table("tminpred.txt")
tmedpred<-read.table("tmedpred.txt")
datospred<-data.frame(preciopred,precioApred,energiapred,tmaxpred,tminpred,tmedpred)
```

Predicción simple a t+1.

```
V12=V1[2:8784];V22=V1[1:8783];V32=V3[2:8784];V42=V4[2:8784];V52=V5[2:8784];V6
2=V6[2:8784];
datosTotal2=data.frame(V12,V22,V32,V42,V52,V62);
attach(datosTotal2);
training<-datosTotal2[1:8040,];
test=datosTotal2[8041:8783,2:6];
resultadomboost<-mboost(V12~V22+V32+V42+V52+V62,data=training);
predmboost=predict(resultadomboost,test);
errormboost=100*abs((V12[8041:8783]-predmboost)/V12[8041:8783]);
summary(errormboost);
```

Previsión a t+2.

```
V13=V1[3:8784];V23=V1[1:8782];V33=V3[3:8784];V43=V4[3:8784];V53=V5[3:8784];V6
3=V6[3:8784];
datosTotal3=data.frame(V13,V23,V33,V43,V53,V63);
attach(datosTotal3);
training2<-datosTotal3[1:8040,];
test2=datosTotal3[8041:8782,2:6];
resultadomboost2<-mboost(V13~V23+V33+V43+V53+V63,data=training2);
predmboost2=predict(resultadomboost2,test2);
errormboost2=100*abs((V13[8041:8782]-predmboost2)/V13[8041:8782]);
summary(errormboost2);
```

Previsión a t+3.

```
V14=V1[4:8784];V24=V1[1:8781];V34=V3[4:8784];V44=V4[4:8784];V54=V5[4:8784];V6
4=V6[4:8784];
datosTotal4=data.frame(V14,V24,V34,V44,V54,V64);
attach(datosTotal4);
training3<-datosTotal4[1:8040,];
test3=datosTotal4[8041:8781,2:6];
resultadomboost3<-mboost(V14~V24+V34+V44+V54+V64,data=training3);
predmboost3=predict(resultadomboost3,test3);
errormboost3=100*abs((V14[8041:8781]-predmboost3)/V14[8041:8781]);
summary(errormboost3);
```

Previsión a t+n.

Una vez que se conozca a qué tiempo se quiere realizar la previsión, valdría con añadir dicho valor a la variable n. en el siguiente ejemplo se realizará a t+5:

```
n<-5;
V16=V1[(n+1):8784];V26=V1[1:(8784-n)]; V36=V3[(n+1):8784]; V46=V4[(n+1):8784];
V56=V5[(n+1):8784]; V66=V6[(n+1):8784];
datosTotal=data.frame(V16,V26,V36,V46,V56,V66);
attach(datosTotal);
training<-datosTotal[1:8040,];
test=datosTotal[8041:(8784-n),2:6];
resultadomboost<-mboost(V16~V26+V36+V46+V56+V66,data=training);
predmboost=predict(resultadomboost,test);
errormboost=100*abs((V16[8041:(8784-n)]-predmboost)/V16[8041:(8784-n)]);
summary(errormboost);
```

BIBLIOGRAFÍA

1. Benjamin Hofner, Andreas Mayr, et al. (2014). *Model-based Boosting in R – A Hands-on Tutorial Using the R Package mboost. Computational Statistics, 29:3-35.*
2. Bühlmann P (2006). *Boosting for high-dimensional linear models. Ann Stat 34:559-583.*
3. Greg Ridgeway (2007). *Generalized Boosted Models: A guide to the gbm package.*
4. Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.*
5. Harrington, H. J.; Tumay, K. (1999) *Simulation modeling models*.
6. http://www.cran.r-proyect.org/web/views/ (n.d.).
7. https://www.esios.ree.es/es (n.d.).
8. https://www.preciopetroleo.net/brent.html (n.d.).
9. http://www.omie.es/inicio (n.d.).
10. https://www.r-project.org/about.html (n.d.).
11. https://www.ree.es (n.d.).
12. https://www.unesa.es (n.d.).
13. Isabel Juárez Barrios (2013). *Predicción del precio de la energía eléctrica utilizando modelos de Minería de Datos: árboles de clasificación y regresión, Random Forest y bagging*. Escuela Técnica Superior de Ingeniería Industrial, Madrid, España (Proyecto de Fin de Carrera).
14. Ley Orgánica 54/1997. Boletín Oficial del Estado, núm. 285, de 28 de noviembre de 1997.
15. Ley Orgánica 24/2013. Boletín Oficial del Estado, núm. 310, de 27 de diciembre de 2013.
16. Oded Maimon and Lior Rokach (2010). *Data Mining and Knowledge Discovery Handbook*.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS