eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Grado en Ingeniería en Tecnología de Telecomunicación

Telekomunikazio Teknologiaren Ingeniaritzako Gradua

TRABAJO FIN DE GRADO

COMPARATIVA DE MODELOS SUPERVISADOS PARA LA PREDICCIÓN DEL CONSUMO DE ENERGÍA ELÉCTRICA

Alumno/Alumna: Martín, Andrés, Beatriz

Director/Directora: Espinosa, Acereda, Koldo

Curso: 2017-2018

Fecha: junio, 2018

RESUMEN

En el presente proyecto se analizarán, procesarán y correlacionarán datos climatológicos, meteorológicos y de consumo energético, con el objetivo de intentar predecir consumos energéticos a partir de las previsiones meteorológicas.

Se estudiarán de manera exhaustiva los conceptos fundamentales de estadística y de *Machine Learning*. La metodología utilizada está basada en el lenguaje *R* y se hará uso de librerías de código abierto para técnicas de minería de datos y *Aprendizaje Automático* supervisado (predictivo), el cual predice un valor desconocido, que en este caso será la energía eléctrica que se estima consumir, a partir de un conjunto de datos conocidos previamente. También se usarán técnicas para manejar y formatear los datos que se obtengan de repositorios de datos abiertos (*Open Data*).

Se realizará una comparativa entre 5 modelos basados en técnicas supervisadas de minería de datos, a partir de los resultados obtenidos de la predicción del consumo energético y comparando estos con datos reales de consumo.

Palabras clave: NoSQL, modelo de datos, análisis de datos, procesos ETL, minería de datos, Machine Learning, Open Data, R, consumo energético, climatología y meteorología.

ABSTRACT

In the present project, climatological, meteorological and energy consumption data will be analyzed, processed and correlated in order to try to predict energy consumption based on weather forecasts.

The fundamental concepts of statistics and Machine Learning will be studied exhaustively. The methodology used is based on the R language and will use open source libraries for data mining techniques and supervised (predictive) automatic learning, which predicts an unknown value that in this case will be the electrical energy consumed, based on a set of previously known data. Techniques will also be used to manage and format the data obtained from open data repositories (Open Data).

A comparison will be made between 5 models based on supervised data mining techniques based on the results obtained from the prediction of energy consumption and comparing these with real consumption data.

Keywords: NoSQL, data model, data analysis, ETL processes, data mining, Machine Learning, Open Data, energy consumption, climatology and meteorology.

LABURPENA

Proiektu honetan, klimatologiaren, meteorologiaren eta energia-kontsumoaren datuak aztertu, prozesatu eta korrelazionatuko dira, datu klimatologikoen aurreikuspenen arabera energia-kontsumoa aurreikusteko.

Estatistika eta *Machine Learning*-aren oinarrizko kontzeptuak aztertuko dira. Erabilitako metodologia R hizkuntzan oinarritzen da eta kode irekiko liburutegiak erabiliko dira datu-meatzaritzako tekniketarako eta aurreikusitako ikasketa automatikorako, balio ezezagun bat aurreikusten duena, kasu honetan energia elektrikoaren kontsumoa izango baita. datu ezagunen multzo batetik. Teknikak ere erabiliko dira datu irekiko datu biltegietatik (Open Data) lortutako datuak kudeatzeko eta formateatzeko.

Datuen meatzaritzako gainbegiratutako tekniketan oinarritutako 5 ereduen arteko konparaketa egingo da, kontsumo energetikoaren aurreikuspenetik lortutako emaitzetan oinarrituta eta kontsumo errealeko datuekin alderatuz.

Hitz gakoak: NoSQL, datu-eredua, datuen analisia, ETL prozesuak, datuen meatzaritza, Makina ikaskuntza, Open Data, R, energia kontsumoa, klimatologia eta meteorologia.

ÍNDICE

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

1. INTRODUCCIÓN

Hoy en día, el ser humano tiene la necesidad de analizar y comprender todo aquello que le rodea. Hasta hace unos años se han realizado dichos análisis de información y datos mediante bases de datos relacionales, o incluso manualmente.

Pero ahora, gracias a Big Data y al *Machine* Learning junto con las existentes técnicas de minería de datos, se pueden realizar análisis de cantidades ingentes de datos e información de cualquier tipo en tiempos relativamente pequeños.

El almacenamiento de información en distintos formatos es cada vez más barato y sencillo. Se genera gran cantidad de datos y hay que intentar sacar partido a estos volúmenes de información para la toma decisiones. Es imprescindible convertir los grandes volúmenes de datos existentes en experiencia, conocimiento y sabiduría para tomar dichas decisiones.

La minería de datos es un conjunto de técnicas agrupadas con el fin de crear mecanismos de reconocimiento de patrones, clasificación y predicción. [1]

La clave para realizar predicciones está en detectar las variables predictoras que más influyen en nuestra predicción, para cambiarlas y hacer que el futuro cambie hacia un estado mucho mejor del que sería si no estuviéramos usando un Big Data con capacidad de análisis predictivo [2][3]. Con este proceso lo que hacemos no es simplemente predecir de forma pasiva el futuro, si no que usamos los resultados obtenidos para construir un futuro más provechoso para nuestros intereses.

En el documento se tratarán los siguientes contenidos:

- Introducción: Definición del contenido junto con una breve introducción y explicación sobre el tipo de análisis que se realizará para realizar el cruce de datos históricos para producir una predicción.
- Contexto y motivación: Se describirá brevemente el contexto en el que se realiza el estudio y la motivación personal para la realización de esta investigación.
- Objetivos y alcance del trabajo: Descripción de los principales objetivos de la investigación y su alcance, habiendo entendido exactamente lo que se quiere conseguir con el proyecto y redefiniendo los aspectos relevantes de éste.
- Beneficios que aporta el trabajo: En este apartado se analizarán los distintos beneficios que se podrían obtener del proyecto.
- Marco teórico: Se explicarán los aspectos más importantes y relevantes a tener en cuenta a la hora de hacer el estudio. De esta forma, al lector se le hará más intuitivo entender los diferentes términos existentes en la investigación.
- Descripción de la solución: Se describirán detalladamente los modelos utilizados para realizar las predicciones. Se desarrollarán los algoritmos y la creación de los scripts en lenguaje R que dan como resultado la predicción del consumo energético. Se explicarán los resultados obtenidos.
- Conclusiones: En este punto se hablará sobre las conclusiones obtenidas una vez terminado el estudio con unos resultados.
- Plan de trabajo: Contiene la descripción de la planificación del proyecto. Incluye la descripción de tareas, fases, y diagrama de Gantt.
- Aspectos económicos: Se describirá el presupuesto del proyecto.
- Bibliografía: Relación de fuentes que se han consultado para la realización del proyecto.

2. CONTEXTO Y MOTIVACIÓN

El procesamiento y análisis de información se enfoca en el desarrollo de técnicas de almacenamiento y análisis de datos para permitir el aprovechamiento de grandes cantidades de estos, con la extracción de información relevante y útil para la toma de decisiones.

Hoy en día, el gran desarrollo de la analítica avanzada es la facilidad con la que grandes ejecutivos de las compañías, científicos de datos o analistas de datos, pueden hacer predicciones y entender el futuro de sus equipos. Uno de los aspectos más importantes de este impulso en la accesibilidad es el uso de las APIs. Compañías de sectores, como pueden ser el financiero o energético, usan interfaces de programación para construir modelos predictivos y extraer valor de los datos para sacar conclusiones de estos y tomar decisiones, como, por ejemplo, conocer cómo se puede aumentar la productividad y el rendimiento, predecir el comportamiento y opiniones de los clientes, para así poder ajustar ofertas y precios, o prevenir y detectar el fraude.

El análisis predictivo ayuda a un grupo de sectores como los siguientes: [4]

Aeroespacial. Gracias al análisis predictivo, se realizan supervisiones del estado del motor de las aeronaves con el fin de aumentar el tiempo productivo de éstas y de reducir los costes de mantenimiento. De hecho, un fabricante de motores ha creado una aplicación de análisis en tiempo real para predecir el rendimiento de los subsistemas en relación con el combustible, el despegue y el buen estado de mecánico y los controles.

Automoción. Se utiliza en el ámbito de los vehículos autónomos para analizar los datos de los sensores de los vehículos conectados, y así crear algoritmos de asistencia a la conducción.

Medicina. Es utilizada básicamente en dispositivos médicos empleando algoritmos de detección de patrones para detectar el asma y EPOC (Enfermedad pulmonar obstructiva crónica).

Servicios Financieros. Establecimientos financieros utilizan herramientas cuantitativas y técnicas de aprendizaje automático para predecir el riesgo crediticio y el fraude.

Maquinaria y automatización industriales. Son utilizados para predecir fallos de la maquinaria y para realizar un mantenimiento predictivo. De esta manera, se reducen los tiempos de inactividad y se minimizan los posibles residuos.

Producción energética. Se realizan predicciones del precio y la demanda de la electricidad. Para todo ello, las aplicaciones de predicción desarrollan un cruce de datos históricos meteorológicos como los del consumo eléctrico, y tratan de predecir datos sobre el precio y el consumo eléctrico.

De acuerdo a varios estudios realizados, el rendimiento energético en distintos lugares geográficos o edificios está influenciado por algunos factores, como pueden ser la estructura del edificio y sus características, las condiciones meteorológicas, operación y componentes (ventilación y climatización del aire, calefacción). La correspondencia entre estos factores puede proporcionar amplias posibilidades para abordar distintas técnicas de predicción. [5]

Los aspectos anteriormente mencionados y los resultados obtenidos de una búsqueda bibliográfica extensa, inducen a explorar y extraer conocimiento de estos datos que permita, además del análisis y comprensión de estos, la toma de decisiones y la posible predicción de acontecimientos asociados a dichas variables.

Para el desarrollo de cualquier actividad, el sector eléctrico es esencial en casi todo el mundo y, es que, se hace cargo de dar un servicio necesario y básico a los usuarios.

A lo largo de los últimos años, debido a la inestabilidad del precio, al cambio climático y a la complejidad de almacenar y ahorrar energía, las técnicas de predicción sobre el consumo eléctrico y el precio se han ido desarrollando y mejorando. El sector de la energía eléctrica en edificios ha sido identificado como uno de los sectores clave para abordar el cambio climático. La calefacción y el aire acondicionado son los principales responsables de las emisiones de gas de efecto invernadero en los edificios. En la UE, los edificios son responsables del 30% de todas las emisiones, es decir, es equivalente a unos 842 millones de toneladas de CO2 cada año.

Con relación a lo anterior, el tratamiento de datos para predecir el consumo energético me llamó mucho la atención desde el primer momento, ya que la idea me parece muy buena porque dicho estudio perseguiría beneficios para el medioambiente, así como ahorros para el usuario.

3. OBJETIVOS Y ALCANCE DEL TRABAJO

Con el presente trabajo de investigación se tiene como objetivo principal hacer un análisis y estudio comparativo de 5 enfoques distintos sobre técnicas supervisadas de minería de datos utilizadas para la predicción. Mediante un análisis y un cruce de datos meteorológicos con los de consumo energético, se tratará de predecir el valor de dicho consumo en una determinada situación para una fecha concreta. Se tendrá en cuenta y a nuestra disposición un histórico de datos del consumo energético, datos históricos de meteorología y datos sobre las previsiones meteorológicas. Se hará una comparativa cruzada de los distintos métodos teniendo en cuenta los resultados obtenidos con cada uno de ellos y se decidirá cuál es el más adecuado y acertado para la realización de la predicción.

Para todo ello se hará un análisis predictivo, el cual agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos, que analiza los datos actuales e históricos reales para hacer predicciones sobre el futuro o acontecimientos no conocidos.

Se explicará qué es el Open Data estudiando en profundidad de dónde proceden los datos y para qué sirve. También se explicará en qué consiste la minería de datos y las diferentes técnicas que existen dentro del análisis predictivo, los modelos aplicables y sus posibles aplicaciones.

Por último, decir que la presente investigación no pretende estudiar las diferentes alternativas existentes sobre técnicas de minería de datos con el fin de obtener una predicción sobre el consumo eléctrico, si no que se analizarán distintos modelos que pueden utilizarse para la obtención de dicha predicción y, mediante una comparativa de resultados, se elegirá el más idóneo para poder hacer una aceptable predicción.

4. BENEFICIOS QUE APORTA EL TRABAJO

4.1. BENEFICIOS TÉCNICOS

La realización de este análisis y comparativa de modelos predictivos junto con su implementación en código podría servir de ayuda para la creación de futuras aplicaciones reales. Se podrían realizar aplicaciones web con las que los usuarios podrían registrar, por ejemplo, los edificios que tengan a su disposición, y una vez hayan introducido un histórico de datos de consumo energético, y gracias a los datos históricos de meteorología, podrán predecir el consumo energético que va a tener el edificio para los próximos días gracias a las previsiones meteorológicas. La aplicación podría proporcionar también la posibilidad de visualizar gráficos del consumo que ha tenido el edificio en cualquier día. Gracias a poder visualizar en qué momentos del día tiene mayor consumo, el usuario podría corregirlo para que, sin cambiar las condiciones del edificio, busque las horas en el que sea más barato el uso de ésta y tener un mayor ahorro.

4.2. BENEFICIOS SOCIO-ECONÓMICOS

La ventaja de poder obtener la predicción del consumo energético es que, en el caso de que cuando se obtengan los valores reales de consumo y sea un valor bastante dispar a lo que se había predicho, el usuario podrá detectar que no está habiendo eficiencia energética en el caso de edificios o también detectar que hay una anomalía en la caldera (por ejemplo). Por otro lado, en el caso en el que se esté haciendo un plan de eficiencia energética, estas previsiones, basadas en el estado anterior del edificio, se podrían utilizar como la marca a bajar.

Por ejemplo, conocer el comportamiento y poder predecir el posible consumo que se podría tener en un determinado día, es importante para favorecer el impacto socio-económico que implica dicho sector eléctrico.

Dependiendo de qué resultados se obtienen de las predicciones, el usuario de dicha energía podrá optar a tarifas del precio de la luz más reducidas y ahorrar mucha más energía. De esta manera, para desarrollos posteriores, cabría la posibilidad de realizar una aplicación en la que muestre al usuario que consumos energéticos ha tenido con el fin de ahorrar energía y dinero.

5. MARCO TEÓRICO

A lo largo de este apartado se realizará una explicación teórica sobre el Open Data, la minería de datos y su importancia junto con el análisis predictivo y los métodos existentes para realizar predicciones.

5.1. OPEN DATA

Significa datos abiertos. Estos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos al requerimiento de atribución y de compartirse de la misma forma en la que aparecen. [6]

La información debe estar accesible y disponible en todo momento a un coste razonable de reproducción en una forma adecuada y modificable.

Por otro lado, los datos deberán estar dotados de términos que permitan ser reutilizados y redistribuidos, además de poder ser integrados con otros conjuntos de datos.

Asimismo, toda esta información debe poder ser utilizada, reutilizada y distribuida por cualquier persona sin ningún tipo de discriminación.

En todos los contextos, Open suele significar la posibilidad de acceder a "algo" para usarlo, modificarlo y compartirlo sin ningún tipo de restricción.

Lo único que se suele exigir es indicar la autoría y mantener su cualidad de abierto. Algunos adjetivos que suelen acompañar al termino abierto respecto al Open Data son: libre, estándar, accesible y reutilizable.

Para caracterizar el Open Data, se suele emplear el sistema de 5 estrellas que *Tim Berners-Lee* en julio de 2006 en su propuesta *Linked Data*. El sistema de 5 estrellas de *Tim Berners-Lee* es acumulativo: un nivel superior incluye a los niveles inferiores. [7]

- Una estrella significa que los datos están disponibles en Internet en cualquier formato como, por ejemplo, PDF siempre que sea con licencia abierta.
- Dos estrellas significan que los datos están disponibles en Internet de manera estructurada como, por ejemplo, en un archivo EXCEL en vez de en PDF.
- Tres estrellas significan que los datos están disponibles en Internet de manera estructurada y en formato no propietario como, por ejemplo, en formato CVS en vez de en EXCEL.

- Cuatro estrellas significan que, además de todo lo anterior, se emplean los estándares establecidos por el *W3C* (El Consorcio WWW, en inglés: World Wide Web Consortium), como RDF y SPARQL, y se usa una URI para identificar los datos y sus propiedades, de manera que las personas las puedan utilizar para sus publicaciones.
- Cinco estrellas incluyen todo lo anterior y además significa que los datos están vinculados a otros datos, creando una red de datos que proporcionan un contexto.

En la *Ilustración 1* se puede observar el modelo de las 5 estrellas propuesto por *Tim Berners-Lee*.

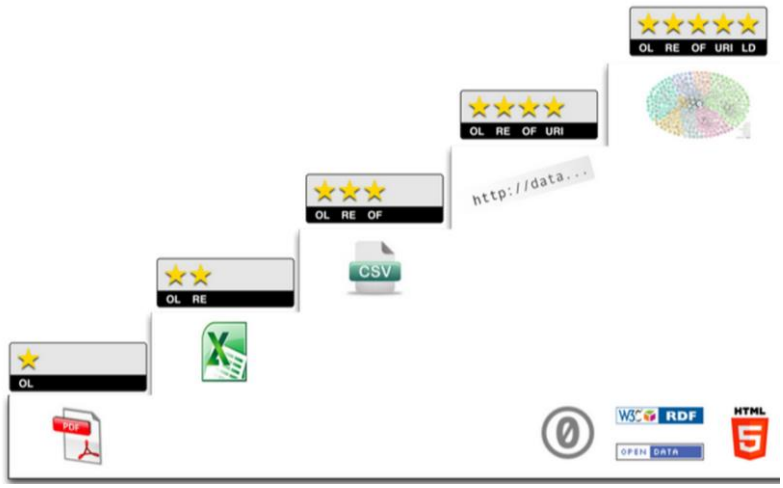

Ilustración 1. Sistema de 5 estrellas de Tim Berners-Lee

Para la obtención de los datos históricos y predicciones del tiempo meteorológico utilizados en este proyecto para su posterior análisis, visité la página web OpenWeatherMap [8], donde conseguí toda la información que necesitaba para después poder utilizarla para hacer la predicción del consumo eléctrico.

Sobre esta página web decir que son una empresa de TI con experiencia práctica en Big Data y tecnologías geoespaciales. Proporcionan una plataforma geoespacial global la cual es asequible para los usuarios y les permite operar con datos como imágenes satélites, datos meteorológicos y fuentes de datos similares sin esfuerzo.

Lo que ofrecen son condiciones actuales y pronósticos para más de 200,000 ciudades y cualquier ubicación geográfica.

Así mismo, proporcionan al usuario información histórica, mapas interactivos del tiempo y mapas de satélites, datos brutos de más de 40,000 estaciones meteorológicas y un API simple y clara para poder utilizar dicha información.

5.2. MINERÍA DE DATOS

La minería de datos tiene como objetivo obtener información útil a través del análisis de los datos. El conocimiento que se obtiene, puede ser en forma de relaciones, patrones o reglas inferidos de los datos y desconocidos, o bien, en forma de una descripción más concisa. [9]

Los modelos obtenidos a partir de los datos y de las técnicas de minería empleadas, pueden dar como resultado uno de los dos tipos de modelos existentes: predictivos o descriptivos. Con los primeros, se intenta determinar valores desconocidos de variables de interés utilizando para ello otras variables o atributos de la base de datos.

Mientras tanto, los modelos descriptivos, inspeccionan las propiedades del conjunto de datos para describir patrones que explican o resumen al mismo conjunto, no para predecir nuevos valores.

Las técnicas de la minería de datos pueden ser predictivas o descriptivas.

Dentro de las técnicas predictivas se encuentran la clasificación y la regresión, mientras que el agrupamiento (clustering), reglas de asociación, reglas de asociación secuenciales y las correlaciones son tareas descriptivas.

- **Técnicas de Minería de Datos**

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

Los algoritmos supervisados o predictivos, como su nombre lo indica, predicen un dato desconocido a partir de un conjunto de datos conocidos previamente llamados descriptivos.

A partir de datos con etiqueta conocida se induce un modelo que relaciona dicha etiqueta con los atributos descriptivos. Tal relación sirve para realizar la predicción en datos cuya etiqueta es desconocida. Los algoritmos supervisados requieren de una fase de entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Por otro lado, los algoritmos no supervisados se caracterizan por descubrir modelos o características significativas a partir únicamente de los datos de entrada. Estos algoritmos realizan tareas descriptivas como el descubrimiento de patrones y tendencias en los datos actuales. El descubrimiento de estos patrones sirve para llevar a cabo acciones y obtener un beneficio científico o de negocio de ellas. [10]

1) Aprendizaje supervisado o predictivo

Predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. Dadas unas variables de entrada, los algoritmos trabajan con datos "etiquetados" para intentar encontrar una función que les asigne una etiqueta de salida adecuada. Primeramente, el algoritmo es "entrenado" con unos datos históricos, de manera que "aprenderá" a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

De esta manera, se dice que el aprendizaje es supervisado porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo que determina el resultado que debería generar el sistema a partir de una entrada determinada. Este agente externo controla la salida del sistema y en el caso de que no coincida con la deseada, se producirán las modificaciones necesarias a los parámetros con el fin de aproximar la salida obtenida a la deseada.

Suele usarse en problemas de clasificación, como identificar diagnósticos o detección de fraude. También se utiliza en problemas de regresión, como pueden ser predicciones meteorológicas, predicciones de consumo o expectativa de vida. Los casos de clasificación son de tipo categórico y los casos de regresión, en cambio, la variable a predecir es de tipo numérico.

2) Aprendizaje no supervisado o del descubrimiento del conocimiento

Es un método de *Aprendizaje Automático* donde un modelo se ajusta a unas observaciones. Con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento obtenido de esa información se utiliza para llevar a cabo acciones y obtener un beneficio de ellas.

A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado no hay un conocimiento a priori. Cuando no se dispone de datos "etiquetados" para el entrenamiento, estamos hablando sobre el aprendizaje no supervisado.

No existen datos de salida que correspondan con un determinado dato de entrada, solo son conocidos los datos de entrada. Por lo que solo se podrá describir la estructura de los datos y así poder encontrar algún tipo de organización o relación que simplifique el análisis.

Este tipo de aprendizaje suele utilizarse en problemas de *clustering*, *profiling* o agrupamientos de co-ocurrencia. Dichas tareas son las que implican tareas que buscan agrupamientos basados en similitudes o reducción de datos. [11] [12]

Los algoritmos más frecuentes utilizados en ambos aprendizajes se muestran en la *Tabla 1*.

SUPERVISADOS	NO SUPERVISADOS
Inducción neuronal	Segmentación
Regresión	Agrupamiento (clustering)
Árboles de decisión	Detección de desviaciones
Series Temporales	Reglas de asociación
	Patrones secuenciales

Tabla 1. Algoritmos de minería de datos

- **Procesos ETL: Extracción, Transformación y Carga**

Estos procesos, dentro de la minería de datos, se encargan, en primer lugar, de la extracción, que es cuando se establece la conexión con las fuentes para recuperar los datos.

Luego tenemos la transformación, que consiste en la aplicación de diversas operaciones para llevar a cabo la migración de los datos. Por ejemplo, realizar una unión "join", un filtrado, una limpieza, etc.

Y finalmente, se cargan los datos según el modelo de los datos común definido en el repositorio Big Data.

Para realizar el proceso de ETL, se requiere tener la posibilidad de contar con diferentes fuentes de datos y poder hacer un filtrado de detección y de corrección, garantizando así la consistencia e integridad de los datos.

De acuerdo con lo anterior, y de manera previa a la aplicación de los métodos y algoritmos predictivos, los datos han sido sometidos a un proceso de limpieza y análisis exploratorio, por medio del cual se normalizaron para lograr estandarizar las diferentes unidades en que se tenía cada variable.

5.3. ANÁLISIS PREDICTIVO

Se pueden definir diferentes tipos de análisis de datos en función de cuál sea nuestro objetivo a realizar. Si queremos saber "cómo actuar" el análisis que se debe hacer es un Análisis Prescriptivo. Por otro lado, si lo que queremos es saber "por qué ha sucedido" estaremos ante un Análisis de tipo Diagnostico. Por otra parte, si queremos saber "qué hacer para que suceda algo" tendremos que utilizar un Análisis Descriptivo. Y, por último, si lo que queremos saber es "qué sucederá" lo más adecuado es el Análisis Predictivo, que es el que se va a utilizar en el presente trabajo.

El Análisis Predictivo es una subdisciplina del análisis de datos que engloba una variedad de técnicas estadísticas de aprendizaje automático, modelización y minería de datos, para desarrollar modelos predictivos que permiten aprovechar patrones de comportamiento encontrados en datos históricos y actuales para identificar eventos futuros, riesgos y oportunidades. [1] [13]

Es una de las herramientas que forman parte del conocido *Bussiness Intelligence*.

Es un área de la minería de datos y su principal objetivo consiste en la extracción de información existente en los datos y su utilización para predecir patrones y tendencias de comportamiento. Identifica relaciones entre variables en eventos pasados, para luego estudiar dichas relaciones y predecir posibles resultados en futuras decisiones.

El análisis predictivo ha recibido mucha atención en los últimos años debido a los avances en la tecnología que lo respalda, especialmente en las áreas de Big Data y aprendizaje automático.

Para poder realizar un análisis predictivo es necesario disponer de una gran cantidad de datos históricos y actuales para poder establecer patrones de comportamiento, de tal manera que se induzca conocimiento.

5.4. MÉTODOS PREDICTIVOS. FUNDAMENTOS TEÓRICOS

En este apartado, se explicarán los métodos principales de predicción existentes y se describirán en profundidad cada uno de los utilizados en este trabajo.

5.4.1. CONTEXTO DE LOS MÉTODOS PREDICTIVOS EXISTENTES

Las técnicas que hoy en día son empleadas se han visto intensamente influenciadas por la aparición de equipos informáticos y a su vez de software capaces de trabajar de una manera más rápida y eficiente.

Generalmente, las técnicas predictivas se clasifican en tres grupos.

1) Técnicas de regresión

Es un proceso estadístico para estimar las relaciones que existen entre variables. Estas técnicas son el pilar de la analítica predictiva. Permiten identificar los dominios más importantes y significativos sobre la variable dependiente sobre las independientes y así analizar su comportamiento para poder realizar una predicción. Los más destacados son los siguientes:

- Análisis de regresión lineal

  El análisis de regresión estudia la relación entre dos variables cuantitativas. Puede ser simple o múltiple. Para obtener la predicción de la energía, en este trabajo se utilizarán ambos modelos.

  En general lo que importa es:

  - Investigar si existe una asociación entre la variable dependiente y las independientes testeando la hipótesis de independencia estadística.
  - Estudiar la fuerza de la asociación, a través del coeficiente de correlación el cual mide dicha asociación.
  - Estudiar la forma de la relación. Usando los datos propondremos un modelo para la relación y a partir de ella será posible predecir el valor de una variable a partir de la otra.

- Análisis de duración o de supervivencia

  Se refiere al análisis de datos en los que se recoge el periodo de tiempo que transcurre desde un punto de partida fijado hasta el momento en que acontece un suceso particular. La no-normalidad y censura de los datos de supervivencia crean una dificultad al intentar analizarlos con modelos estadísticos convencionales como la represión lineal múltiple.

- Árboles de regresión y clasificación

  Es una técnica de aprendizaje de árboles de decisión no paramétrica que crean árboles de clasificación o regresión, dependiendo de si la variable dependiente es numérica o categórica. Para obtener la predicción del consumo energético se hará uso del modelo *Random Forest* el cual es una técnica de árboles de regresión.

2) Técnicas de aprendizaje computacional

Estas técnicas utilizan la capacidad de los equipos informáticos para desarrollar técnicas y trabajar con cantidades ingentes de datos. En ciertas aplicaciones, con solo predecir directamente la variable dependiente sin tener en cuenta las relaciones implícitas entre las variables restantes es suficiente. En otros casos, dichas relaciones pueden ser bastante trabajosas y la forma matemática de las dependencias es desconocida. Los modelos que usan pueden ser no lineales. Como técnicas más importantes tenemos:

- Redes neuronales

  Estudian las relaciones y conexión de las neuronas en el cerebro y modelan dicho comportamiento para que los ordenadores sean capaces de reproducirlo. Pueden ser aplicadas a problemas de clasificación, predicción o control. Son utilizadas cuando se desconoce la naturaleza exacta de la relación entre los valores de las variables de salida y entrada.

- K- vecinos más cercanos (K-NN)

  K vecinos más próximos es un método de clasificación supervisada, es decir, a partir de un conjunto de datos inicial su objetivo es clasificar correctamente todas las nuevas instancias. Pertenece a la clase de métodos estadísticos de reconocimiento de patrones.

  En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras. Una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento y el signo de este punto será el que determine la clasificación de la muestra.

  En este presente trabajo se va a usar esta técnica para obtener la predicción del consumo eléctrico y se analizarán los resultandos comparándolos con los obtenidos en otros métodos.

3) Técnicas de simulación

La simulación trata de un modelo numérico el cual representa la estructura de un proceso dinámico. Modelan el comportamiento de las variables del sistema llevando a cabo una simulación para representar la conducta del proceso tratando de buscar un algoritmo matemático que dé solución al modelo.

5.4.2. TÉCNICAS DE REGRESIÓN

Los métodos de regresión lineal pueden ser simples o múltiples. Esto depende del número de variables explicativas (variables independientes) que se usen para estudiar y predecir la variable final (variable dependiente). Es el más utilizado a la hora de predecir los valores de una variable cuantitativa y continua a partir de los valores de otra variable o de otras variables (en el caso de regresión lineal múltiple) también cuantitativas.

En el caso de estudio del presente trabajo, la variable dependiente a predecir será *energiaRLS* para el modelo de Regresión Lineal Simple y *energiaRLM* para el modelo de Regresión Lineal Múltiple. También se hará la predicción y la comparativa con otro tipo de regresión llamada Regresión Lineal Robusta. Para este caso la variable se llamará *energiaRobusta*. Finalmente, implementaremos el modelo de Random Forest, donde para este caso la variable se llamará *energiaRandomF*. Durante este apartado se hará referencia a ella de distintas maneras como variable dependiente, final o explicada.

Como ya se ha mencionado anteriormente, para predecir esta variable se utilizarán cinco modelos de regresión y, tanto las variables explicativas como la variable final quedan definidas en el apartado 6.1.2.

5.4.2.1. MODELO DE REGRESIÓN LINEAL SIMPLE

Es el modelo de regresión más simple ya que la variable dependiente solo depende de una única variable regresora. De forma matemática se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X + \varepsilon_i$$

Donde:

- Y es el valor de la variable dependiente que se quiere predecir.
- β_0 es el término independiente y representa el valor de la variable de salida cuando la variable regresora es cero.
- β_1 representa la pendiente de la recta el cual indica el incremento que experimenta la media de la variable independiente cuando X aumenta en una unidad.
- X es el valor de la variable independiente.
- ε representa el error aleatorio cometido por el modelo.
- $\varepsilon_i \rightarrow \mathbf{N}(0, \sigma^2)$ donde σ es la varianza.

En la *Ilustración 2* se puede observar una expresión gráfica de un modelo de regresión lineal simple. [14]

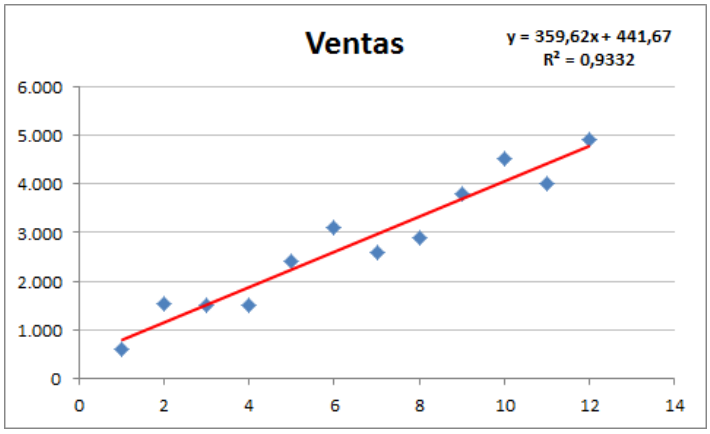

Ilustración 2. Ejemplo de representación gráfica de una regresión lineal simple

En ella se puede ver cómo se realiza el ajuste de los datos sobre ventas mediante regresión lineal simple. De hecho, los parámetros $\boldsymbol{\beta_0}$ (441,67) y $\boldsymbol{\beta_1}$ (359,62) son los parámetros de la recta $y = f(x)$ que mejor ajustan la relación entre las variables a estudiar.

El razonamiento matemático para obtener los parámetros adecuados partiendo de los datos proporcionados se explica a continuación.

Para estimar los parámetros $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ y $\boldsymbol{\sigma^2}$ existen dos técnicas de estimación: mediante el método de máxima verosimilitud (MV) y el método de los mínimos cuadrados (MCO):

$$Max \left\{\frac{1}{(2\Pi)^n \sigma^n} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right]\right\}$$

$$Min \left\{(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

Existe un problema en este tipo de modelo de regresión lineal, y es que simplifica mucho el estudio y es poco objetivo, ya que la mayoría de variables independientes que se estudian no dependen solo de una única variable.

5.4.2.2. MODELO DE REGRESIÓN LINEAL MÚLTIPLE

La regresión lineal múltiple (por sus siglas en inglés, MLR) trata de ajustar modelos lineales entre una variable dependiente a predecir y más de una variable independiente.

Para obtener la predicción de la energía eléctrica, es necesario recurrir a modelos en los que se pueda predecir el comportamiento de la variable dependiente teniendo en cuenta el gran número de variables independientes que influyen en ella. El modelo de regresión lineal múltiple se ajusta bastante a esto último.

La fórmula matemática para expresar este modelo es similar a la de la regresión lineal simple, lo único que varía es que ahora se tendrán *n* variables independientes (regresoras) para *n* observaciones de la variable final dependiente. Se muestra a continuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i \qquad (1)$$

En este caso los parámetros a estimar serán β_0, $\beta_1, \beta_2, \ldots, \beta_n$ y σ^2 y normalmente para estimarlos el método más utilizado es el de los mínimos cuadrados.

Para *n* observaciones de la variable dependiente, podemos plantear el modelo en forma matricial de la siguiente forma: [15]

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La expresión (1) se puede escribir de la siguiente manera asignando la notación a las matrices respectivas:

$$Y = X\beta + \varepsilon$$

Aplicando el método de los mínimos cuadrados para estimar el vector β, se obtiene como resultado el siguiente estimador:

$$\beta = (X^t X)^{-1} X^t Y$$

donde X^t es la matriz transpuesta de X^t.

En la *Ilustración 3* se muestra un ejemplo gráfico para dos variables independientes. En este caso, al tener dos variables independientes (*x,z*) en vez de una como en el caso anterior (*x*), el resultado pasa de ser una recta de ajuste a una plano de ajuste el cual está definido por la expresión $y = f(x, z)$ cuyos parámetros estimados serán $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ y $\boldsymbol{\beta_2}$.

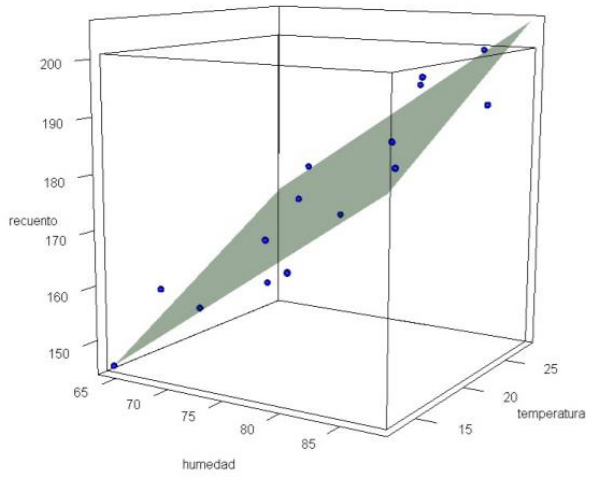

Ilustración 3. Ejemplo de representación gráfica de regresión lineal múltiple con 2 variables regresoras independientes [15]

Por último, decir que este tipo de regresión plantea el problema de que imposibilita representar gráficamente los valores de la variable dependiente para más de dos variables independientes.

5.4.2.3. MODELO DE REGRESIÓN LINEAL ROBUSTA

Es una forma de análisis de la regresión que fue diseñada para evitar algunas limitaciones de los métodos paramétricos y no paramétricos. En general, el análisis de regresión busca encontrar la relación entre una variable dependiente con una o más variables independientes. Algunos métodos que se utilizan para estudiar la regresión, como el de los mínimos cuadrados, pueden dar resultados engañosos para suposiciones que no son ciertas y, por tanto, se dice que el método de los mínimos cuadrados no es robusto al incumplimiento de los supuestos. Los métodos de regresión robusta fueron diseñados para no ser muy afectados por violaciones de los supuestos. Amortigua el efecto de las observaciones que serían muy influyentes si se usaran los mínimos cuadrados. [16] [17]

Gracias a que los métodos de regresión robusta tienen propiedades de robustez y ofrecen mejores soluciones a problemas de regresión con valores no típicos, estos métodos han logrado mucha importancia y difusión a partir del año 1973.

Autores como *Huber* (1973), *Denby, y Larsen* (1977), *Seber* (1977), *McKean y Hettmansperger* (1978), *Hurdle* (1981), *Lachan* (1985), *Staudte, y Sheather* (1990), *Wilcox*, (2005), *Montanari* (2008), *Aelst, Willems y Zamar* (2013) han desarrollado importantes aportes a la estadística y métodos robustos.

Algunos de los métodos de regresión robusta existentes son el de M-Regresión (Máxima Verosimilitud) propuesto por *J. Huber* en 1973 y el de L-Regresión (Combinación lineal de estadísticos de orden).

Por otro lado, existen varios procedimientos de estimación como M, S, MM, L… En este trabajo se utilizará el modelo de regresión robusta sin parámetro y con el parámetro MM, por tanto, solo se explicará este último ya que los demás no son objeto de estudio.

Estimadores MM

Fue propuesto por Yohai en 1987. Este tipo de estimador es un tipo especial del estimador M y es considerado como una generalización de los estimadores M (Máxima Verosimilitud) dado que, en el proceso de estimación, estos son obtenidos después de aplicar consecutivamente el estimador M en las dos últimas etapas del proceso.

El objetivo del modelo de estimación MM es obtener simultáneamente un estimador de punto de quiebre alto que mantenga una alta eficiencia. Tiene las siguientes propiedades: es muy eficiente cuando la distribución del error es normal y su punto de quiebre es de 0,5.

5.4.2.4. MODELO DE ÁRBOLES DE REGRESIÓN: RANDOM FOREST

Hay varios modelos de árboles y cada modelo empleado determinará la manera de elegir la variable a predecir. En este proyecto, como la variable a predecir es la energía eléctrica consumida y es una variable cuantitativa, es decir, es una variable numérica, el modelo que se ha elegido para ello es el de árboles de regresión.

Cuando la variable a predecir es numérica hablamos de *árboles de regresión*, mientras que las variables categóricas se analizan usando *árboles de clasificación*. En ambos casos, el funcionamiento es relativamente parecido: para hacer una predicción para una determinada observación, se utilizará la media (o la moda) de las observaciones que se encuentran en la misma región del espacio multidimensional de predictores. Las reglas que se utilizan para dividir el espacio de predictores pueden ser representadas en forma de árbol; de ahí el nombre de estos métodos. [18]

Un árbol de regresión es un árbol de decisión cuyas hojas predicen una cantidad numérica. Dicho valor numérico es calculado como la media del valor para la variable clase de rodos los ejemplos que han llegado a esas hojas durante el proceso de construcción del árbol. Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por un conjunto de variables que devuelve una respuesta que es tomada a partir de las entradas.

La evaluación de un ejemplo es similar a los árboles de decisión. Durante el proceso de predicción es posible utilizar un suavizado de los valores del ejemplo que va a ser tratado. De esta manera evitarían posibles discontinuidades presentes en los datos.

Los árboles de decisión y regresión conforman el grande grupo de los modelos de aprendizaje supervisado. Estos deben ser entrenados con información que contiene históricos de los datos y los resultados obtenidos en consecuencia de dichos datos históricos para poder realizar predicciones.

En general, la metodología para construir árboles de clasificación y árboles de regresión es la misma. La diferencia radica en la escogencia de la función impureza para dividir un nodo y en la estimación del costo-complejidad para podar el árbol.

A continuación, se mencionarán algunas ventajas y desventajas sobre los árboles de regresión:

Ventajas

- Las reglas de asignación son simples y legibles, por tanto, la interpretación de resultados es directa e intuitiva.
- Es robusto frente a datos atípicos u observaciones mal etiquetadas.
- Es válido para cualquier naturaleza de las variables explicativas: continuas, binarias nominales u ordinales.
- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- Es computacionalmente rápido

Desventajas:

- Las reglas de asignación son bastantes sensibles a pequeñas perturbaciones en los datos.
- Existe una relativa dificultad para elegir el árbol óptimo.
- La ausencia de una función global de las variables causa la pérdida de la representación geométrica.
- Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

Los tres modelos de árboles de decisión más importantes son los modelos *CART*, los modelos *Random Forest* y el modelo *Bagging*. Pero, como en este presente trabajo se ha decidido implementar el modelo de Random Forest, únicamente se explicará este último.

Random Forest

Conocidos en castellano como "Bosques Aleatorios", es una combinación de árboles predictores. Cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

El algoritmo fue desarrollado por *Leo Breiman* y *Adele Cutler*. El método combina la idea de *Bagging* de *Breiman* y la selección aleatoria de atributos, introducida independientemente por *Ho, Amit* y *Geman* para construir una colección de árboles de decisión con variación controlada.

Para la creación de un determinado número de árboles elegidos por el usuario de la técnica, *Random Forest* sigue el siguiente proceso:

Dado un conjunto de datos de entrenamiento N, se selecciona un numero X de datos de forma aleatoria y con reposición para cada uno de los árboles que se van a crear.

A continuación, se crea un árbol de decisión con cada grupo de datos. Si se tiene un número total de *m* variables predictoras, a la hora de crear los árboles se eligen aleatoriamente *p* variables entre todas las predictoras. Estas, serán seleccionadas antes de evaluar la división del nodo que generalmente es evaluado siguiendo el criterio de homogeneización.

Los árboles obtenidos serán todos distintos por la aleatoriedad en la selección de los datos que emplea y las variables predictoras. De esta manera, se crean tantos modelos distintos como árboles cada árbol crece en profundidad sin podar.

Una vez han sido creados, para realizar la predicción, éste se determina mediante el cálculo de la media de los valores de predicción por todos los árboles del modelo o mediante el nodo más votado según el tipo de árbol (de clasificación o regresión).

Posteriormente, se debe realizar la validación del modelo. Para ello se introduce el conjunto de datos iniciales y la determinación del error que posee el modelo, es decir, el *MSE (Mean Squared Error* en inglés*) o ECM (Error Cuadrático Medio* en castellano*)*. El usuario, para realizar el proceso de ajuste, buscará el menor *MSE* posible con un tiempo de computación razonable.

En cada árbol creado para ajustar el modelo, cada vez se utilizan alrededor de 2/3 de los datos de entrenamiento originales. El 1/3 restante se consideran observaciones *out og bag (OOB)*, que se utilizan para estimar el error de cada modelo.

5.4.3. TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Como se ha mencionado anteriormente, son técnicas que trabajan con grandes cantidades de datos y cuyo objetivo es desarrollar técnicas que permitan a las maquinas "aprender". Dentro de los tipos de aprendizaje automático y los diferentes modelos predictivos, en este presente trabajo nos centraremos en el aprendizaje supervisado y en el algoritmo de *k-NN (k-Nearest Neighbors),* dónde la variable dependiente a predecir se llamará *energiaKNN*.

5.4.3.1. REGRESIÓN POR K-VECINOS MÁS CERCANOS (K-NN)

El modelo de clasificación o predicción por proximidad más general es el modelo de los *k* vecinos más cercanos. El método consiste en comparar la variable a predecir con los datos o casos existentes del problema en cuestión, recuperando los *k* casos más próximos. Es un método no paramétrico usado para las tareas de clasificación y regresión. La diferencia entre clasificación y regresión reside en el tipo de la variable final. En ambos casos, las variables de entrada son un número determinado de observaciones del conjunto de entrenamiento. Para realizar la predicción de los valores de la variable final se evalúa una medida de similitud entre dos observaciones dadas.

Después de obtener esta medida de similitud, es usada para calcular cada uno de los datos con una observación nueva y se selecciona un grupo de *k* distancias más cercanas y se le asignará un el mismo peso a cada *k* vecino más próximo. A diferencia de otros métodos, éste es universal y asintóticamente convergente.

Generalmente, la mejor elección de *k* depende principalmente de los datos. Los valores grandes de *k* reducen el efecto de ruido en la clasificación, pero crean límites entre clases que tienen un cierto parecido. Un buen *k* puede ser seleccionado mediante una optimización de su uso.

Un ejemplo de representación se muestra en la *Ilustración 4.*

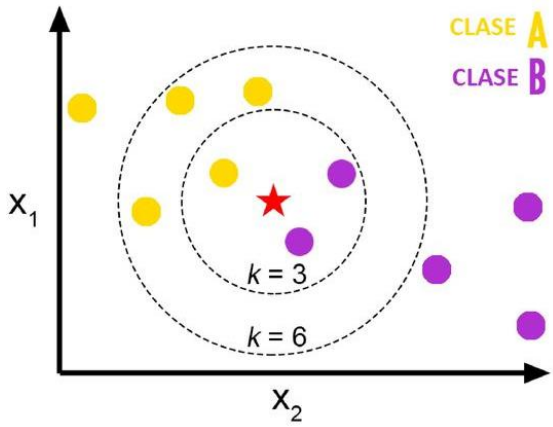

Ilustración 4. Ejemplo de representación gráfica del modelo k-NN para k=3 y k=6 [19]

Algunos de los inconvenientes que tiene este modelo son que es muy sensible a los atributos irrelevantes y la imprecación de la dimensionalidad. Es también muy sensible al ruido y relativamente lento si hay muchos datos de entrenamiento ya que tiene que guardar todos ellos. [20]

5.4.4. ANÁLISIS DE LAS TÉCNICAS DE REGRESIÓN Y SELECCIÓN DEL MODELO DE REGRESIÓN

Una vez evaluados los modelos de regresión y habiendo obtenido los residuos, se deberán comprobar si las hipótesis utilizadas para llevarlo a cabo no son contradictorias con los datos a utilizar.

Si estas hipótesis son apropiadas, se podrán usar dichos modelos para conseguir predicciones y poder lograr el principal objetivo de este proyecto que es comparar dichos métodos y así tomar decisiones basándonos en los resultados de la predicción.

De esta manera, la interfaz que se utilizará es *R-Studio*, que contiene diversidad de funciones e instrucciones junto con gráficos para conseguir las comprobaciones y comparaciones apropiadas.

Por otro lado, cuando se tiene un gran conjunto de variables explicativas o independientes, surge el problema de saber seleccionar el conjunto de las mismas que proporciona el mejor modelo de regresión y que sea el más adecuado para explicar la variable dependiente.

Si tenemos un número de variables explicativas *k,* el número posible de modelos a los que podremos optar es 2^k. En general, cuando el número de variables es alto, el modelo de regresión suele presentar mejores predicciones.

El objetivo, por tanto, será buscar y obtener un subconjunto de dichas variables que sea los suficientemente explicativo y no muy complejo.

5.4.5. ELABORACIÓN DE LOS DISTINTOS MODELOS DE REGRESIÓN

Existen varias estrategias de regresión para la selección de las variables regresoras cuando se tiene un amplio número de estas. [21]

En general, si se incluyen cada vez más variables en un modelo de regresión, el ajuste de los datos mejora, aumenta la cantidad de parámetros a estimar, pero disminuye su precisión individual y por tanto la de la función de regresión estimada.

En otras palabras, debemos seleccionar un subconjunto de variables entre todas las variables candidatas a ser explicativas de la variable dependiente, un subconjunto que resulte suficientemente explicativo.

A continuación, se presentan los procedimientos para seleccionar las variables regresoras los cuales tratarán de seleccionar las variables más explicativas y evitarán aquellas variables innecesarias que empeoran la predicción.

- Eliminación progresiva

  En este tipo, se empieza con una regresión que incluye todas las variables explicativas disponibles y se van eliminando de una en una según su capacidad explicativa. Para ello, se calcula el coeficiente de correlación parcial de cada variable independiente con la dependiente y primeramente se eliminará de la regresión aquella variable que presente el menor coeficiente de correlación con la variable a predecir.

- Introducción progresiva

  Como se puede deducir, este método es el opuesto al anterior. Se empieza con una única variable y se van añadiendo el resto a la regresión una a una. La primera variable a añadir suele ser la que presente un coeficiente de correlación más alto con la variable dependiente. Después, se calcula la regresión simple entre ambas y los coeficientes de correlación parcial con el resto de las variables independientes y la variable respuesta eliminando el efecto de la primera variable. A continuación, se introducirá la variable que tenga un coeficiente de correlación parcial con la respuesta más alto.

- Regresión paso a paso (Stepwise Regression)

  Este último método es una combinación de los procedimientos anteriores. Parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa examinando si todas las variables introducidas en el modelo deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo.

6. DESCRIPCIÓN DE LA SOLUCIÓN

A lo largo de este apartado se explicarán con detalle las distintas fases que se han llevado a cabo para la obtención de los valores futuros energéticos, así como la implementación de los modelos elegidos.

Estas fases consisten en:

- Carga de los ficheros de los datos con el que se hará el estudio.
- Transformación y adecuación de los datos.
- Creación de nuevas tablas para realizar el entrenamiento de los datos y comprobación de las predicciones.
- Calcular la correlación entre las variables independientes y la variable dependiente a predecir.
- Creación de los distintos modelos predictivos.

6.1. IMPLEMENTACIÓN DE LOS MODELOS Y PREDICCIONES EN R

Para alcanzar el objetivo de este proyecto, que es realizar una comparativa y estudiar cuál de los modelos da los mejores resultados, en primer lugar, los modelos estadísticos explicados en el apartado 5.5 serán implementados en el entorno *R-Studio* para obtener la predicción del consumo eléctrico. Posteriormente, se compararán los resultados conseguidos y se analizará la calidad de estos modelos quedándonos con aquel que proporcione una predicción más fiable o con aquel que se asemeje a los datos predictivos reales originales, es decir, con aquel modelo que arroje menor probabilidad de fallo.

6.1.1. ENTORNO Y LENGUAJE R

El lenguaje de programación elegido para el desarrollo del estudio ha sido R. Este dispone de una interfaz gráfica, la cual permite abstraer la complejidad matemática del análisis predictivo, poniéndolo a disposición de organizaciones y empresas.

R es un lenguaje de programación con un enfoque al análisis estadístico. Es un potente lenguaje orientado a objetos y es uno de los más utilizados en investigación y análisis estadístico y muy popular en el campo de la minería de datos, la investigación biomédica y las matemáticas financieras.

Gracias a la posibilidad de cargar diferentes librerías y paquetes, contribuye a poder utilizar con facilidad funciones de cálculo y gráficas para representar los datos.

R-Studio es un entorno integrado de desarrollo (IDE) para el lenguaje de programación *R*. Como se muestra en la *Ilustración 5*, su interfaz incluye un editor de código fuente desde el cual se puede ejecutar el código *R* directamente, una consola, un panel con las pestañas Historial (Histoy) y Espacio de Trabajo (Workspace), un panel con pestañas para Ficheros (Files), Gráficos (Plots), Librerías (Packages) y Ayuda (Help) y múltiples herramientas para el trazado, gestión y depuración del espacio de trabajo.

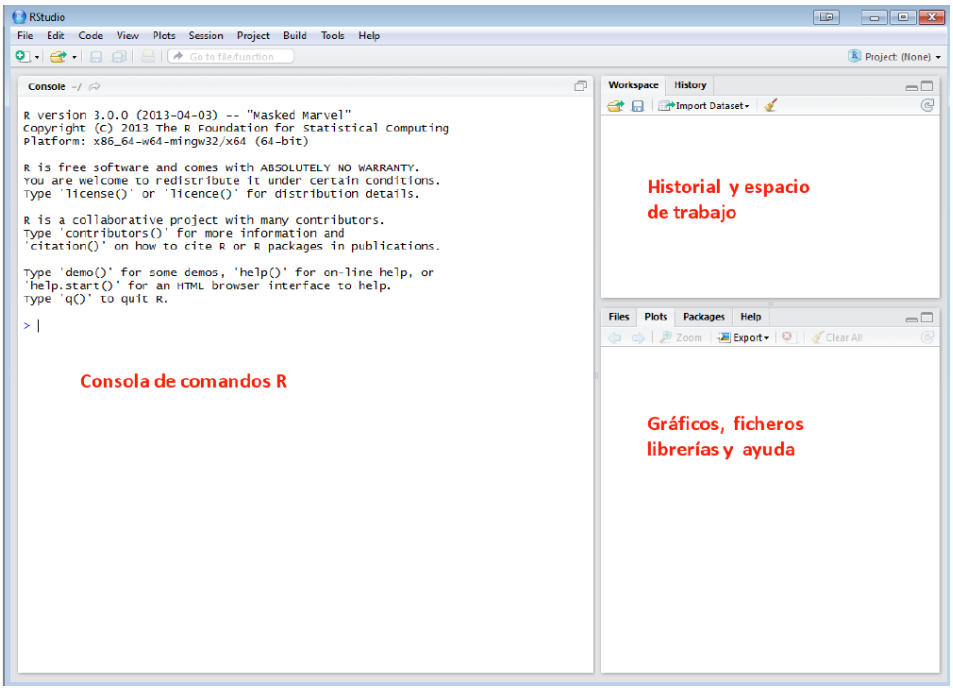

Ilustración 5. Entorno de R-Studio

Tanto *R (R Project)* y *R-Studio* son softwares *Open Source* y gratuitos. Para poder utilizar *R-Studio*, previamente ha sido instalado *R*.

Ofrece un resaltado de sintaxis, finalización de código e identificación inteligente, la gestión de múltiples directorios de trabajo usando proyectos y muchísimas más ventajas.

6.1.2. PREPARACIÓN Y ADECUACIÓN DE LOS DATOS ANALIZADOS

Los ficheros de datos

Para conseguir el objetivo del presente trabajo, ha sido necesario obtener una extensa fuente de datos que permitan el estudio y análisis sobre ellos. Dentro del conjunto de datos a utilizar, los datos meteorológicos, como ya se mencionó en el apartado 5.2 sobre *Open Data*, fueron obtenidos de la página web *OpenWeatherMap* [13] de forma gratuita ya que esta página web proporciona datos abiertos a cualquier usuario sin ningún tipo de coste. Los datos históricos del consumo eléctrico fueron proporcionados por una empresa la cual no puede ser revelada por confidencialidad.

Primeramente, antes de crear los modelos estadísticos predictivos, es preciso preparar correctamente los datos, leyendo y ejecutando en *R-Studio* aquellos datos que nos interesen y dándoles un formato correcto.

Los datos están recogidos en dos ficheros Excel con formato *.csv* (del inglés comma-separated values, es decir, archivo de datos separados por comas) llamados "*consumoElectrico.csv*" y "*weatherData.csv*". Los datos de ambos archivos se organizan en una sola hoja lo que facilita el código *R* para su lectura.

El fichero "*consumoElectrico.csv*" contiene información sobre el consumo eléctrico cada media hora desde el año 2011 al 2012 clasificada por fecha según el formato *dd/mm/aaaa*. La información se corresponde a distintos aspectos relacionados con la energía eléctrica, la cual se representa en distintas columnas indicando el nombre en su correspondiente cabecera. Las variables que forman este fichero quedan explicadas un poco más adelante. La variable realmente significativa y la que requiere interés de esta tabla es la llamada *"Energy.Consumed.kWh"* la cuál luego cambiaremos de nombre por *"energyConsum_kWh"* y será la variable a predecir.

Por otro lado, el fichero "*weatherData.csv*" contiene la información meteorológica obtenida cada día en los años 2011 y 2012 clasificada por el mismo formato de fecha que en el otro fichero. La información se corresponde de distintas variables que definen el tiempo meteorológico para cada día, como pueden ser la temperatura máxima y mínima obtenida para un determinado día, la presión del aire o la velocidad del viento.

A continuación, se nombrarán las diferentes variables que conforman dichos ficheros con los que partimos y que posteriormente se van a utilizar para la creación de los distintos modelos predictivos:

- Fichero "*consumoElectrico.csv*"

Variables cuantitativas

- *energyConsum_kWh*: energía eléctrica consumida media en kW/h. El nombre de esta variable lo cambiaremos posteriormente por "*Energia (kW/h)*".
- *volumeConsumed_m3:* para poder calcular el consumo de gas natural en nuestros hogares hay que mirarlo en el contador instalado, que lo mide en volumen (m^3).
- *volumeConsumed_FT3:* ft^3 (pie cúbico) es la medida de líquidos utilizada en EE.UU. El pie cúbico es una unidad de volumen, equivalente al volumen de un cubo de un pie de lado. Equivale a 0.02831 m^3.
- *InherentCarbon_kgCO2:* emisiones de dióxido de carbono en kg por kW/h.

Variables cuantitativas

- *Date*: fecha con formato *dd/mm/aaaa* junto con la hora (los datos están distribuidos cada media hora). Posteriormente se eliminará la hora de esta columna para tener en ambas tablas la fecha por día únicamente.

- Fichero "*weatherData.csv*"

Variables cuantitativas

- *MAX.øC*: Temperatura máxima registrada para un determinado día expresada en grados Celsius. El nombre de esta variable lo cambiaremos posteriormente por "*temp MAX (C°)".*
- *MIN.øC:* Temperatura mínima registrada para un determinado día expresada en grados Celsius. El nombre de esta variable lo cambiaremos posteriormente por "*temp MIN (C°)".*
- *AVG.øC:* Temperatura media registrada para un determinado día expresada en grados Celsius. El nombre de esta variable lo cambiaremos posteriormente por "*Temp MEDIA (Cº)".*
- *Av.Dew.øC:* Temperatura media de rocío expresada en grados Celsius. El nombre de esta variable lo cambiaremos posteriormente por "*Temp de Rocio (Cº)".*

- *Av.Wet.Bulb.øC:* Temperatura húmeda del aire expresada en grados Celsius. El nombre de esta variable lo cambiaremos posteriormente por "*Temp Húmeda de Aire (Cº)".*
- *Rain.mm:* uno de los parámetros que caracterizan la lluvia es la altura o profundidad, que se define como la altura que tendría en agua precipitada sobre un m^2 de superficie horizontal impermeable, si la totalidad del agua precipitada no se escurriera. Esta dimensión es la que se mide en pluviómetros. Generalmente se expresa en *mm* (1 mm de agua sobre 1 m^2 equivale a un litro). Esta variable expresa esto mismo. El nombre de esta variable lo cambiaremos posteriormente por "*Lluvia (mm)*".
- *Av.Baro.mb:* media por día de la presión atmosférica expresada en milibares. El nombre de esta variable lo cambiaremos posteriormente por "*Lluvia (mm)*".
- *Av.Wind.mph:* media por día de la velocidad del viento expresada en millas por hora. El nombre de esta variable lo cambiaremos posteriormente por "*Viento (mph)*".
- *Dir.ø*: dirección del viento expresada en grados de acimut. El nombre de esta variable lo cambiaremos posteriormente por "*Dirección del viento (Grados)*".
- *Max.Gust.mph:* indican las ráfagas de viento expresadas en millas por hora. El nombre de esta variable lo cambiaremos posteriormente por "*Ráfagas viento (mph)*".
- *Max.Gust.Dirø:* indica la dirección de las ráfagas de viento expresada en grados de acimut. El nombre de esta variable lo cambiaremos posteriormente por "Dirección *Ráfagas viento (Grados)*".

Variables cuantitativas

- *DAY*: fecha con formato *dd/mm/aaaa*.

Habrá que extraer, estudiar y transformar estos datos para que en ambos CSV haya una correspondencia entre ellos, es decir, para poder trabajar con toda esta información en cada CSV debe de haber el mismo tipo de dato y con el mismo formato.

Después de estudiar y dar formato a dichas tablas, crearemos una tabla llamada *"TablaConsumo_Weather"* juntando ambas para luego dividirla en dos partes a partir de una determinada fecha, creando una "*tablaEntrenamiento*" y otra "*tablaResultado*". Por un lado, en la primera tabla, tendremos los datos históricos del consumo eléctrico y del tiempo meteorológico y en la segunda tabla tendremos las predicciones reales del consumo y del tiempo meteorológico.

Los datos de la primera tabla los usaremos de entrenamiento para, una vez realizada la predicción del consumo con cada modelo, poder compararlos con las predicciones reales del consumo de la segunda tabla.

Código en R-Studio

Lo primero que tendremos que hacer es establecer un directorio de trabajo utilizando el comando ***setwd()***. A continuación, guardaremos la información y todos los datos de ambos CSV en un *data frame* respectivamente, leyendo directamente del fichero CSV con el comando ***read.csv()***.

Los datos históricos del consumo eléctrico se guardarán en el *data frame* "*gas2011_2012"* y los datos históricos meteorológicos se guardarán en el *data frame "weather2011_2012"*. Todo esto se muestra en la *Ilustración 6.*

En un principio, la información está comprendida desde el 29-11-2011 hasta el 31-03-2012. Puesto que se trata de predecir el comportamiento de una variable dependiente según otras variables independientes o explicativas, lo lógico es usar datos reales del pasado de esa variable para un día determinado, y con ellos se intentará predecir los datos que se podrían ocurrir en el futuro que todavía desconocemos.

```
# Establecemos el directorio de trabajo
setwd("C:/Users/bmartinandres/Desktop/B_R_Nuevo")

gas2011_2012 = read.csv("C:/Users/bmartinandres/Desktop/B_R_Nuevo/consumoElectrico.csv", header=T, dec=".",sep=";")
weather2011_2012 = read.csv("C:/Users/bmartinandres/Desktop/B_R_Nuevo/weatherData.csv", header=T, dec=".",sep=";")
```

Ilustración 6. Carga de datos en R-Studio

Puesto que en la tabla "*weather2011_2012*" no aparecen los datos por horas y en la tabla "*gas2011_2012"* si, debemos eliminar la hora en esta última para poder tratar los datos de la misma forma. Es decir, tenemos que adaptar los datos a un mismo formato para luego poder procesarlos de una manera adecuada. Para ello será necesario cargar la librería ***reshape*** y hacer uso del comando ***sapply()***.

Los datos tienen que estar en un formato adecuado para poder hacer el posterior análisis. Por ello, pasaremos los datos que están en modo "factor" a "numérico", es decir, cambiaremos las comas por puntos. Para ello será necesario cargar la librería ***plyr*** y hacer uso del comando ***as.numeric()***.

Posteriormente, como ya se ha mencionado antes, los datos de la tabla "*gas2011_2012"* están distribuidos cada media hora por día. Como en la tabla *"weather2011_2012"* los datos están distribuidos por día, se hará la suma de todos los datos de cada columna de la tabla "*gas2011_2012"* para tener información acerca del consumo por día. Para ello haremos uso del comando ***ddply()***.

Se eliminarán filas de los *data frames* que contengan valores vacíos (N/A) ya que no aportan ningún tipo de información.

Todo lo mencionado anteriormente se muestra en la *Ilustración 7.*

```
# Quitamos la hora de la columna Date de gas2011_2012
library(reshape)
gas2011_2012$Date = sapply(strsplit(as.character(gas2011_2012$Date), " "), "[", 1)

# Eliminamos los valores que no nos interesan de weather2011_2012
weather2011_2012$Max.Gust.Dirø = NULL
weather2011_2012$ Dir.ø = NULL
weather2011_2012$Max.Gust.mph = NULL
weather2011_2012$AVG.øC = NULL
weather2011_2012$Av.Wet.Bulb.øC = NULL
weather2011_2012$Av.Dew.øC = NULL

# Mostramos los nombres de las columnas de gas2011_2012
colnames(gas2011_2012)

# Pasamos de factor a numerico (cambiar las "," por ".")
library(plyr)
gas2011_2012$Energy.Consumed.kWh = as.numeric(sub(",", ".", gas2011_2012$Energy.Consumed.kWh, fixed = TRUE))
gas2011_2012$Volume.Consumed.m3 = as.numeric(sub(",", ".", gas2011_2012$Volume.Consumed.m3, fixed = TRUE))
gas2011_2012$Volume.Consumed.ft3 = as.numeric(sub(",", ".", gas2011_2012$Volume.Consumed.ft3, fixed = TRUE))
gas2011_2012$Inherent.Carbon.kgCO2 = as.numeric(sub(",", ".", gas2011_2012$Inherent.Carbon.kgCO2, fixed = TRUE))

# En gas2011_2012, hacemos la suma de las demás columnas para tener el consumo de un dia
# (Los datos son cada media hora)
gas2011_2012 = ddply(gas2011_2012,.(Date),summarize,energyConsum_kWh=sum(Energy.Consumed.kWh),
                     volumeConsumed_m3=sum(Volume.Consumed.m3),
                     volumeConsumed_FT3=sum(Volume.Consumed.ft3),InherentCarbon_KgCO2=sum(Inherent.Carbon.kgCO2))

# Eliminamos las filas de gas2011_2012 que tienen valores vacios (no nos sirven para nada)
gas2011_2012 = gas2011_2012[complete.cases(gas2011_2012),]
```

Ilustración 7. Adecuación de los datos en R-Studio

Seguidamente, se creará una nueva tabla llamada *"TablaConsumo_Weather"* la cual contendrá los datos históricos del consumo y del tiempo meteorológico. Para ello agruparemos las tablas "*gas2011_2012*" y "*weather2011_2012*" por día utilizando el comando ***merge().*** En la *Ilustración 8* se muestra lo anterior.

```
# Creamos una nueva tabla Consumo juntando las de gas2011_2012 weather2011_2012 por dia
TablaConsumo_Weather = merge(gas2011_2012,weather2011_2012, by.x="Date", by.y="DAY")
```

Ilustración 8. Creación de la tabla "TablaConsumo_Weather" en R-Studio

En este punto ya tendremos todos los datos necesarios agrupados en una misma tabla. Para poder hacer la predicción del consumo energético tendremos que partir de unos datos históricos tanto del consumo como de los meteorológicos. Estos serán los datos a entrenar. Asimismo, tendremos que tener a nuestra disposición datos sobre las predicciones reales sobre el consumo energético y el tiempo meteorológico para, después de haber realizado la predicción con los datos de entrenamiento, poder comparar estos con los datos predictivos reales.

Para todo esto, se dividirá la *"TablaConsumo_Weather"* en dos tablas a partir de una fecha que será 2012-02-01. Los datos correspondientes a una fecha menor o igual a 2012-02-01 se guardarán en el *data frame* "*tablaEntrenamiento*" y aquellos datos a partir de la fecha 2012-02-01 se guardarán en el *data frame "tablaResultado".*

La "*tablaEntrenamiento*" contendrá los datos históricos sobre el consumo eléctrico y el tiempo meteorológico con los que se aplicarán los distintos modelos predictivos. La *"tablaResultado"*, en cambio, contendrá las predicciones reales sobre el consumo y el tiempo meteorológico. Se utilizará esta tabla para comparar los resultados obtenidos en cada modelo sobre la predicción del consumo.

En primer lugar, se tuvo que cambiar el formato de la columna *"Date"* de la "*TablaConsumo_Weather*" para poder dividir la tabla por fecha. Se añadió una nueva columna a esta tabla con el nuevo formato llamada *"newDate".* Para ello se hizo uso de los comandos ***strptime()*** y ***format()***. A continuación, se prosiguió a dividir esta tabla en las dos tablas anteriormente mencionadas con el comando ***subset().***

Todo lo anterior se muestra en la *Ilustración 9*.

```
## Cambio del formato en la fecha para poder dividir la tabla por fecha.
# Añadimos una nuevo columna "newDate" con el nuevo formato
TablaConsumo_Weather$newDate = strptime(as.character(TablaConsumo_Weather$Date), "%d/%m/%Y")
format(TablaConsumo_Weather$newDate, "%Y/%m/%d")
# Ordenamos la tabla por fecha
TablaConsumo_Weather = TablaConsumo_Weather[order(as.Date(TablaConsumo_Weather$newDate, format="%d/%m/%Y")),]

# Dividimos la TablaConsumo_Weather en dos tablas: "tablaEntrenamiento" y "tablaResultado"

##  tablaEntrenamiento: datos correspondientes a fechas <=  2012-02-01 de la "TablaConsumo_Weather"  (HISTORICOS REALES)
#Con los datos de la tablaEntrenamiento utilizaremos los algoritmos para obtener la prediccion del consumo
tablaEntrenamiento = subset(TablaConsumo_Weather, newDate <= "2012-02-01",select=energyConsum_kWh:newDate)

## tablaResultado: datos correspondientes a fechas >  2012-02-01 de la "TablaConsumo_Weather"  (PREDICCIONES REALES)
#Se utilizará esta tabla para comparar los resultados obtenidos en la predicción
tablaResultado = subset(TablaConsumo_Weather, newDate > "2012-02-01",select=energyConsum_kWh:newDate)
```

Ilustración 9. Creación de las tablas "tablaEntrenamiento" y "tablaResultado en R-Studio

En la "*tablaEntrenamiento*" tenemos información sobre datos históricos de consumo eléctrico y datos históricos del tiempo meteorológico. Queremos relacionar la variable "*energyConsum_kWh*" (energía consumida) con las distintas variables meteorológicas. Para ello calcularemos la correlación entre dicha variable con las demás para saber qué porcentaje tienen en común y determinar la relación lineal entre ellas. Para ello se hará uso del comando ***cor()*** como se muestra en la *Ilustración 10.*

```
# Calculamos la correlacion entre la energia consumida y varias vbls para ver que porcentage tienen en comun
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_MIN)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_MAX)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_Media)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_deRocio)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_Humeda_Aire)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Presion)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Viento)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Dirección_Viento)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Ráfagas_Viento)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Dirección_RáfagasViento)
cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Lluvia)
```

Ilustración 10. Cálculo del coeficiente de correlación

De esta forma, como se muestra en la *Ilustración 11,* veremos que:

```
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_MIN)
[1] -0.6412217
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_MAX)
[1] -0.547002
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_Media)
[1] -0.6204626
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_deRocio)
[1] -0.5728405
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Temp_Humeda_Aire)
[1] -0.6417423
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Presion)
[1] -0.007287728
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Viento)
[1] -0.09955878
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Dirección_Viento)
[1] -0.4333135
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Ráfagas_Viento)
[1] -0.1114612
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Dirección_RáfagasViento)
[1] -0.3017723
> cor(tablaEntrenamiento$energyConsum_kWh, tablaEntrenamiento$Lluvia)
[1] 0.1564537
```

Ilustración 11. Resultado obtenido sobre los coeficientes de correlación

El coeficiente de correlación obtenido entre la variable "*Temp_MIN*" y la variable "*energyConsum_kWh*" es de -0.641, es decir, estas dos variables tienen una relación lineal del 64,1% (la temperatura mínima influye bastante en el consumo).

El coeficiente de correlación obtenido entre la variable "*Temp_MAX*" y la variable "*energyConsum_kWh*" es de -0.547, es decir, estas dos variables tienen una relación lineal del 54,7% (la temperatura máxima influye bastante en el consumo).

El coeficiente de correlación obtenido entre la variable "*Temp_Media*" y la variable "*energyConsum_kWh*" es de -0.621, es decir, estas dos variables tienen una relación lineal del 62,1% (la temperatura media influye bastante en el consumo).

El coeficiente de correlación obtenido entre la variable "*Temp_deRocio*" y la variable "*energyConsum_kWh*" es de -0.573, es decir, estas dos variables tienen una relación lineal del 57,3% (la temperatura de rocío influye bastante en el consumo).

El coeficiente de correlación obtenido entre la variable "*Temp_Humeda_Aire*" y la variable "*energyConsum_kWh*" es de -0.642, es decir, estas dos variables tienen una relación lineal del 64,2% (la temperatura húmeda del aire influye bastante en el consumo).

El coeficiente de correlación obtenido entre la variable "*Presion*" y la variable "*energyConsum_kWh*" es de -0.007, es decir, estas dos variables tienen una relación lineal de menos de un 1% (la presión apenas influye en el consumo eléctrico).

El coeficiente de correlación obtenido entre la variable "*Viento*" y la variable "*energyConsum_kWh*" es de -0.099, es decir, estas dos variables tienen una relación lineal de menos de un 1% (la velocidad del viento apenas influye en el consumo eléctrico).

El coeficiente de correlación obtenido entre la variable "Direccion_*Viento*" y la variable "*energyConsum_kWh*" es de -0.433, es decir, estas dos variables tienen una relación lineal de menos de un 43,3% (la dirección del viento influye en el consumo eléctrico).

El coeficiente de correlación obtenido entre la variable "Rafagas_*Viento*" y la variable "*energyConsum_kWh*" es de -0.111, es decir, estas dos variables tienen una relación lineal de menos de un 11,1% (las ráfagas de viento influyen poco en el consumo eléctrico).

El coeficiente de correlación obtenido entre la variable "Direccion_Rafagas*Viento*" y la variable "*energyConsum_kWh*" es de -0.302, es decir, estas dos variables tienen una relación lineal de menos de un 30,2% (la dirección de las ráfagas de viento influye en el consumo eléctrico).

El coeficiente de correlación obtenido entre la variable "*Lluvia*" y la variable "*energyConsum_kWh*" es de -0.156, es decir, estas dos variables tienen una relación lineal del 15,6% (la precipitación apenas influye en el consumo eléctrico).

En la *Tabla 2* se muestran dichos coeficientes en tanto porciento de mayor a menor. Los marcados en color rojo indican aquellas variables con mayor relación con la variable dependiente.

Variables explicativas	Coef. Correlación (%)
Temp_Húmeda_Aire	**64,2**
Temp_MIN	**64,1**
Temp_Media	**62,1**
Temp_deRocio	**57,3**
Temp_MAX	**54,7**
Dirección_Viento	43,3
Dirección_RáfagasViento	30,2
Lluvia	15,6
Ráfagas_Viento	11,1
Viento	< 1
Presión	< 1

Tabla 2. Coeficientes de correlación entre la vble dependiente con las variables explicativas

En este momento ya se tienen preparados todos los datos necesarios para poder crear los modelos explicados en el apartado 5.5.2 y 5.5.3 y poder elegir cuál es el mejor observando cuál de ellos nos da mayor probabilidad de acierto en la predicción.

En el siguiente apartado se expondrán cada uno de los modelos junto a los resultados obtenidos.

6.1.3. REGRESIÓN LINEAL SIMPLE

Empezaremos por analizar el modelo de regresión lineal simple, donde en el apartado 5.5.2.1 quedó explicado de forma teórica en qué consistía.

Antes de crear el modelo, es necesario definir la fórmula explicada en el fundamento teórico que relaciona la variable respuesta o dependiente con la variable explicativa o independiente:

$$Y = \beta_0 + \beta_1 X + \varepsilon_i$$

La variable independiente elegida para este caso ha sido la variable "*Temp_Húmeda_Aire*", la cual presenta un coeficiente de correlación con la variable dependiente de un 64,2% (es mayor coeficiente de correlación de los obtenidos en el apartado anterior).

En el código *R*, haremos uso del comando ***lm()***, que representa a la obtención de un modelo lineal con la siguiente sintaxis:

```
lm(formula, data, subset, weights, na.action, method = "qr", model =
TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts =
NULL, offset, ...)
```

En este caso, al tratarse del modelo de regresión lineal simple, utilizaremos la fórmula descrita anteriormente empleando una única variable explicativa; la ya mencionada "*Temp_Húmeda_Aire*".

Seguidamente, se realizará la predicción del consumo utilizando la función ***predict()*** que se define de la siguiente manera:

```
predict (object, ...)
```

Este comando calcula los valores predichos para datos nuevos `(...)` de un modelo ya ajustado `(objetct)`, que este último será el modelo lineal simple.

A continuación, en la *ilustración 12*, se muestra el código implementado en R:

```
################## REGRESION LINEAL SIMPLE #################
modeloEnergiaSimple = lm(energyConsum_kWh ~ Temp_Humeda_Aire , data=tablaEntrenamiento)
# Analizar modelo
summary(modeloEnergiaSimple)

# Predecimos los consumos de energia
tablaResultado$energiaRLS = predict(modeloEnergiaSimple, tablaResultado)

# Gráfico realidad VS predicción
plot(tablaResultado$energiaRLS, tablaResultado$energyConsum_kWh, main = "Predicción energía VS Realidad_RLS",
     xlab = "Predicción Energía", ylab = "Realidad Energía")
```

Ilustración 12. Código R del modelo de regresión lineal simple

La ejecución del modelo se muestra a continuación

```
Call:
lm(formula = energyConsum_kWh ~ Temp_Humeda_Aire, data = tablaEn
trenamiento)

Residuals:
    Min      1Q  Median      3Q     Max
-389.00  -90.22   13.61  121.30  323.43

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1112.346     39.060  28.478  < 2e-16 ***
Temp_Humeda_Aire  -46.567      7.011  -6.642 8.38e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.3 on 63 degrees of freedom
Multiple R-squared:  0.4118,    Adjusted R-squared:  0.4025
F-statistic: 44.11 on 1 and 63 DF,  p-value: 8.382e-09
```

Se puede observar que el coeficiente de determinación múltiple (`Multiple R-squared`) R^2 es 0.4118, lo que significa que la recta de regresión explica el 41,18% de la variabilidad del modelo. Además, para el contraste F, este tiene un valor de 44.11 con un p-valor < 0.001, lo que significa que el modelo se ajusta significativamente a los datos. Para acabar, la estimación de la varianza residual (`Residual standard error`) es de 174.43.

En la *Ilustración 13* se muestra un gráfico de dispersión que representa en cada uno de sus cuadrantes, los valores de la variable explicada "*energyConsum_kWh*" según la predicción obtenida *"energía_RLS"* en el eje X y los datos en la realidad en el eje Y.

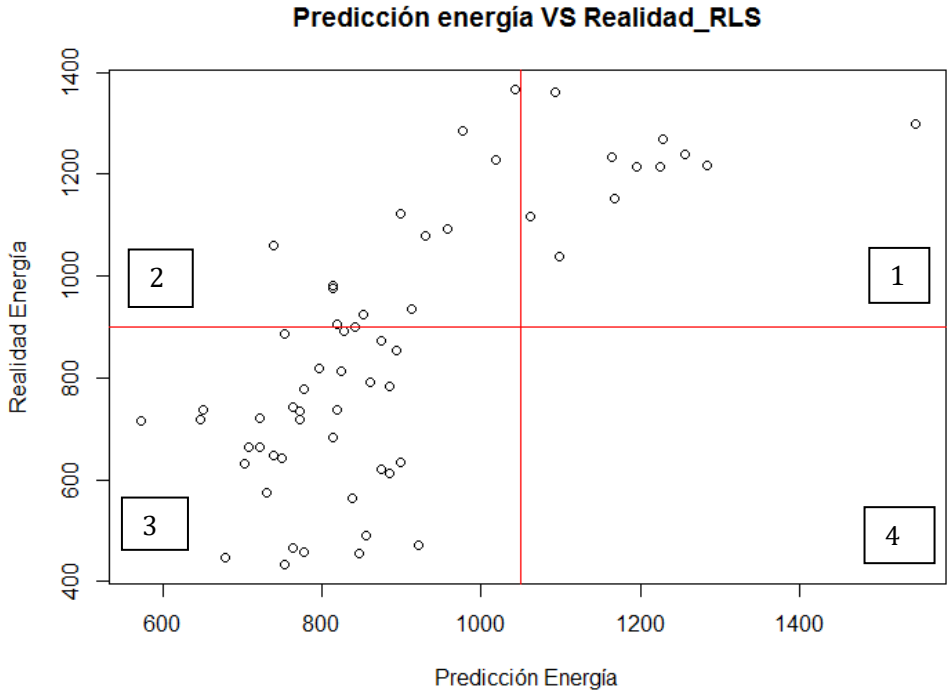

Ilustración 13. Predicción energía VS datos reales modelo RLS

Las líneas en rojo separan cada uno de los cuadrantes. Podemos observar que el modelo está a un nivel adecuado de acierto ya que la mayor parte de la nube de puntos pertenece o al cuadrante 1 o al cuadrante 3 (realidad y predicción positivos, realidad y predicción negativos respectivamente). En este caso la mayor parte de la nube de puntos pertenece al cuadrante 3.

6.1.4. REGRESIÓN LINEAL MÚLTIPLE

Para analizar el modelo de regresión lineal simple, donde en el apartado 5.5.2.2 quedó explicado en qué consistía, antes es necesario definir la fórmula explicada en el fundamento teórico que relaciona la variable respuesta o dependiente con las variables explicativas o independientes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$$

Para la creación del modelo, primeramente, se utilizarán todas las variables explicativas y posteriormente se hará el estudio con aquellas que resulten más significativas para el comportamiento de la variable final a predecir. Para ambos casos, se hará uso del mismo comando que para el modelo de regresión lineal simple, ***lm()*** y se utilizará el mismo método ***predict()*** para predecir los valores energéticos.

- **Utilizando todas las variables independientes o explicativas**

  Se emplearán todas las variables independientes de la tabla "we*ather2011_2012*" anteriormente definidas en el apartado 6.1.2.

A continuación, en la *ilustración 14*, se muestra el código implementado en R:

```
################### REGRESIÓN LINEAL MULTIPLE  con TODAS LAS VARIABLES ##################
modeloEnergiaRLM_TodasLasVbls = lm(energyConsum_kWh ~ Temp_MIN + Temp_MAX + Temp_Media + Temp_deRocio + Temp_Humeda_Aire +
                                     Presion + Viento + Dirección_Viento + Ráfagas_Viento + Dirección_RáfagasViento
                                   + Lluvia, data=tablaEntrenamiento)

# Analizar modelo
summary(modeloEnergiaRLM_TodasLasVbls)

# Predecimos los consumos de energia
tablaResultado$energiaRLM_TodasLasVbls = predict(modeloEnergiaRLM_TodasLasVbls, tablaResultado)

# Gráfico realidad VS predicción
plot(tablaResultado$energiaRLM_TodasLasVbls, tablaResultado$energyConsum_kWh, main = "Predicción energía VS Realidad_RLM_TodasLasVbls",
     xlab = "Predicción Energía", ylab = "Realidad Energía")
```

Ilustración 14. Código R del modelo de regresión lineal múltiple con todas las variables

La ejecución del modelo se muestra a continuación: (2)

```
Call:
lm(formula = energyConsum_kWh ~ Temp_MIN + Temp_MAX + Temp_Media
 +
    Temp_deRocio + Temp_Humeda_Aire + Presion + Viento + Direcci
ón_Viento +
    Ráfagas_Viento + Dirección_RáfagasViento + Lluvia, data = ta
blaEntrenamiento)

Residuals:
    Min      1Q  Median      3Q     Max
-327.81  -90.23   22.63   95.23  294.37

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1355.2649  2198.2149   0.617   0.5402
Temp_MIN                  16.0767    20.3903   0.788   0.4339
Temp_MAX                  47.9496    25.5784   1.875   0.0664 .
Temp_Media                 2.9610    24.1374   0.123   0.9028
Temp_deRocio              67.6267    36.7737   1.839   0.0715 .
Temp_Humeda_Aire        -180.0205    71.3680  -2.522   0.0147 *
Presion                   -0.2508     2.1363  -0.117   0.9070
Viento                   -30.9726    26.3508  -1.175   0.2451
Dirección_Viento          -0.6775     0.4255  -1.592   0.1173
Ráfagas_Viento            15.6684     7.6384   2.051   0.0452 *
Dirección_RáfagasViento   -0.2093     0.4086  -0.512   0.6107
Lluvia                    15.3981    13.6793   1.126   0.2654
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 158.7 on 53 degrees of freedom
Multiple R-squared:   0.59,     Adjusted R-squared:  0.5049
F-statistic: 6.934 on 11 and 53 DF,  p-value: 4.313e-07
```

Se puede observar que el coeficiente de determinación múltiple (`Multiple R-squared`) R^2 es 0.59, lo que significa que la recta de regresión explica el 59% de la variabilidad del modelo y, por tanto, el ajuste es mejor que en el obtenido en el modelo de regresión lineal simple ya que el valor de R^2 es más próximo a 1, que es donde se lograría un ajuste lineal perfecto. Además, para el contraste F, este tiene un valor de 6,934 con un p-valor < 0,05, lo que significa que el modelo se ajusta mejor a los datos. Para acabar, la estimación de la varianza residual (`Residual standard error`) es de 158,7, menor que en el caso anterior.

En la *Ilustración 15* se muestra un gráfico de dispersión que representa en cada uno de sus cuadrantes, los valores de la variable explicada "*energyConsum_kWh*" según la predicción obtenida *"energiaRLM_TodasLasVbls"* en el eje X y los datos en la realidad en el eje Y.

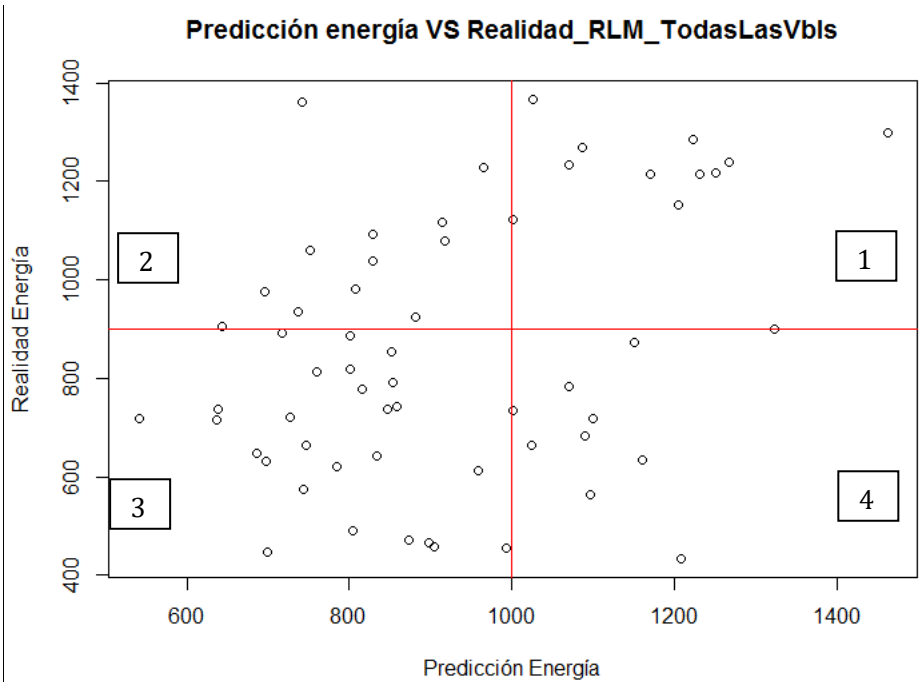

Ilustración 15. Predicción energía VS datos reales modelo RLM con todas las variables

Podemos observar que el modelo está a un nivel adecuado de acierto ya que la mayor parte de la nube de puntos pertenece o al cuadrante 1 o al cuadrante 3 (realidad y predicción positivos, realidad y predicción negativos respectivamente). En este caso la mayor parte de la nube de puntos pertenece al cuadrante 3 pero, respecto al modelo de regresión lineal simple, se observan menos puntos en este cuadrante.

- **Utilizando las variables independientes más significativas**

Como se ha explicado en el apartado 5.5.4, los modelos de regresión lineal determinan qué variables son más significativas en relación al comportamiento de la variable dependiente a predecir.

Por todo esto, se volverá a aplicar el modelo de regresión anterior a aquellas variables que el modelo inicial ha deducido que son más significantes. Esta significación depende del estadístico *t* y en *R* es identificada de la siguiente manera:

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Para la realización de este modelo, de (2) se ha decidido elegir como más significativas aquellas variables con `Pr(>|t|)` señaladas con un * (asterisco) y con un `.` (punto).

En la *Ilustración 16* se muestra el código del nuevo modelo:

```
################## REGRESION LINEAL MULTIPLE con variables MÁS SIGNIFICATIVAS ##################
modeloEnergiaRLM_VariablesMasSignif = lm(energyConsum_kWh ~ Temp_MAX + Temp_deRocio + Temp_Humeda_Aire + Ráfagas_Viento,
                                         data=tablaEntrenamiento)

# Analizar modelo
summary(modeloEnergiaRLM_VariablesMasSignif)

# Predecimos los consumos de energia
tablaResultado$energiaRLM_VariablesMasSignif = predict(modeloEnergiaRLM_VariablesMasSignif, tablaResultado)

# Gráfico realidad VS predicción
plot(tablaResultado$energiaRLM_VariablesMasSignif, tablaResultado$energyConsum_kWh,
     main = "Predicción energía VS Realidad_RLM_VariablesMasSignif", xlab = "Predicción Energía", ylab = "Realidad Energía")
```

Ilustración 16. Código R del modelo de regresión lineal múltiple con variables más significativas

La ejecución del modelo se muestra a continuación:

```
Call:
lm(formula = energyConsum_kWh ~ Temp_MAX + Temp_deRocio + Temp_H
umeda_Aire +
    Ráfagas_Viento, data = tablaEntrenamiento)

Residuals:
    Min      1Q  Median      3Q     Max
-442.88  -67.12   35.24   99.44  289.36

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       975.187     81.555  11.957  < 2e-16 ***
Temp_MAX           40.420     19.726   2.049  0.04484 *
Temp_deRocio       85.770     25.678   3.340  0.00144 **
Temp_Humeda_Aire -178.045     37.806  -4.709 1.51e-05 ***
Ráfagas_Viento      9.909      3.696   2.681  0.00946 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 159.7 on 60 degrees of freedom
Multiple R-squared:  0.5297,    Adjusted R-squared:  0.4983
F-statistic: 16.89 on 4 and 60 DF,  p-value: 2.513e-09
```

Se puede observar que el coeficiente de determinación múltiple (`Multiple R-squared`) R^2 es 0.5297, lo que significa que la recta de regresión explica el 52,97% de la variabilidad del modelo (algo menos respecto al anterior). Además, para el contraste F, este tiene un valor de 16,89 con un p-valor < 0,001. Para acabar, la estimación de la varianza residual (`Residual standard error`) es 159,7. Decir que, respecto al modelo anterior, este ha empeorado, pero no de forma considerable.

6.1.5. REGRESIÓN LINEAL ROBUSTA

Continuaremos analizando el modelo de regresión lineal robusta, donde en el apartado 5.5.2.3 quedó explicado en qué consistía.

Los métodos de regresión robusta son aquellos que tienen propiedades de robustez y ofrecen mejores soluciones a problemas de regresión con valores no típicos. Por esto mismo se ha decidido implementar el modelo y observar los resultados.

En el código R, primeramente, tendremos que cargar las librerías necesarias utilizando el comando ***library(rlm)*** y ***library(MASS)***. A continuación, utilizaremos el método ***rlm()***, que representa a la obtención de un modelo lineal robusto con la siguiente sintaxis:

```
rlm(x, y, weights, ..., w = rep(1, nrow(x)), init = "ls", psi =
psi.huber, scale.est = c("MAD", "Huber", "proposal 2"), k2 = 1.345,
method = c("M", "MM"), wt.method = c("inv.var", "case"), maxit =
20, acc = 1e-4, test.vec = "resid", lqs.control = NULL))
```

En este caso, al tratarse de un método muy similar al descrito anteriormente en el modelo de regresión lineal múltiple, se utilizarán las mismas variables explicativas que en el caso de uso de las variables más significativas.

Seguidamente, como en los casos anteriores, se realizará la predicción objetivo de este trabajo utilizando la función ***predict()*** que se define de la siguiente manera:

```
predict (object, ...)
```

A continuación, en la *Ilustración 17*, se muestra el código implementado en R:

```
################## REGRESION ROBUSTA ##################
library(rlm)

library(MASS) #Libreria para poder utilizar la de rlm

modeloEnergiaRobust = rlm(energyConsum_kWh ~ Temp_MAX + Temp_deRocio + Temp_Humeda_Aire + Ráfagas_Viento,
                          data=tablaEntrenamiento)

# Analizar modelo
summary(modeloEnergiaRobust)

# Predecimos los consumos de energia
tablaResultado$energiaRobusta = predict(modeloEnergiaRobust, tablaResultado)

# Gráfico realidad VS predicción
plot(tablaResultado$energiaRobusta, tablaResultado$energyConsum_kWh, main = "Predicción energía VS Realidad_RLR",
     xlab = "Predicción Energía", ylab = "Realidad Energía")
```

Ilustración 17. Código R del modelo de regresión lineal robusta

La ejecución del modelo se muestra a continuación:

```
Call: rlm(formula = energyConsum_kWh ~ Temp_MAX + Temp_deRocio +
 Temp_Humeda_Aire +
    Ráfagas_Viento, data = tablaEntrenamiento)
Residuals:
    Min      1Q  Median      3Q     Max
-466.52  -84.54   16.94   88.64  291.03

Coefficients:
                  Value     Std. Error t value
(Intercept)      1009.8660   77.3551    13.0549
Temp_MAX           37.6388   18.7100     2.0117
Temp_deRocio       90.2897   24.3552     3.7072
Temp_Humeda_Aire -184.0320   35.8590    -5.1321
Ráfagas_Viento     11.1865    3.5057     3.1909

Residual standard error: 126 on 60 degrees of freedom
```

Para este modelo podemos observar que la estimación de la varianza residual (Residual standard error) es 126, el cual es mucho menor que el de los modelos anteriores.

En la *Ilustración 18* se muestra un gráfico de dispersión que representa en cada uno de sus cuadrantes, los valores de la variable explicada "*energyConsum_kWh*" según la predicción obtenida *"energiaRobusta"* en el eje X y los datos en la realidad en el eje Y.

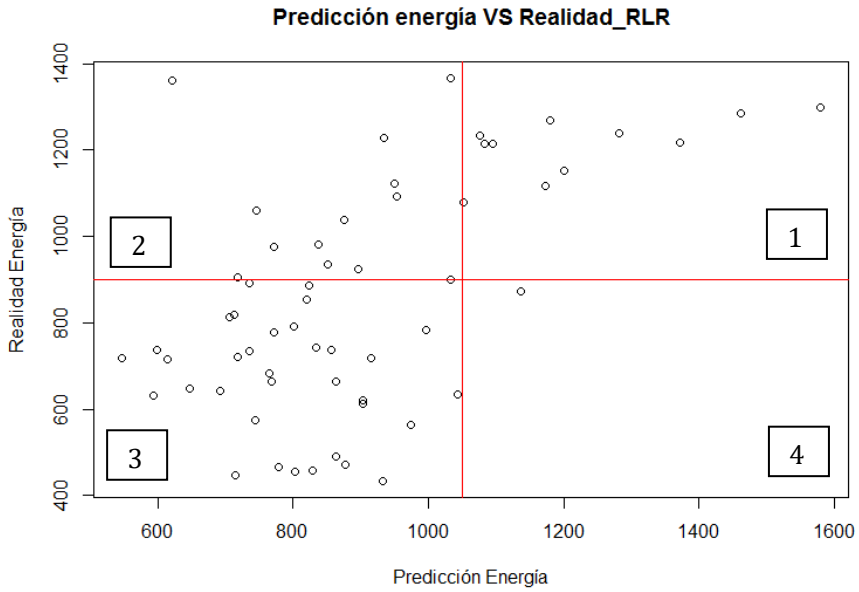

Ilustración 18. Predicción energía VS datos reales modelo RLR

Podemos ver que el modelo es muy parecido al anterior y está a un nivel adecuado de acierto ya que la mayor parte de la nube de puntos pertenece al cuadrante 3.

- **Regresión lineal robusta para estimar por MM**

El objetivo de realizar esta estimación es obtener simultáneamente un estimador de punto de quiebre alto que mantenga una alta eficiencia.

Para ello, en *R* se hará uso de la función ***rlm()*** indicándole con ***method = 'MM'***, como se muestra en la *Ilustración 19,* que se va a hacer una estimación por MM.

```
################## REGRESION ROBUSTA (Con parámetro MM) ##################
modeloEnergiaRobustMM = rlm(energyConsum_kWh ~ Temp_MAX + Temp_deRocio + Temp_Humeda_Aire + Ráfagas_Viento,
                            data=tablaEntrenamiento, method = "MM")
```

Ilustración 19. Código R del modelo de regresión lineal robusta con parámetro MM

La ejecución del modelo se muestra a continuación:

```
Call: rlm(formula = energyConsum_kWh ~ Temp_MAX + Temp_deRocio +
 Temp_Humeda_Aire +
    Ráfagas_Viento, data = tablaEntrenamiento, method = "MM")
Residuals:
    Min      1Q  Median      3Q     Max
-468.86  -85.60   16.16   85.88  290.08

Coefficients:
                 Value     Std. Error t value
(Intercept)      1013.3672   81.5759    12.4224
Temp_MAX           37.4606   19.7309     1.8986
Temp_deRocio       92.0556   25.6841     3.5841
Temp_Humeda_Aire -185.5152   37.8156    -4.9058
Ráfagas_Viento     11.1853    3.6970     3.0255

Residual standard error: 130.8 on 60 degrees of freedom
```

Realizada la estimación para este modelo, podemos observar que la estimación de la varianza residual (`Residual standard error`) es 130,8, la cual ha aumentado respecto al anterior. Por tanto, los resultados predictivos son peores.

6.1.6. RANDOM FOREST

El siguiente método a analizar es el ya explicado en el apartado 5.5.3.4, el método de árboles regresión, *Random Forest.*

En el código R, primeramente, tendremos que cargar el paquete ***library(randomForest)*** el cual incluye el método ***randomForest()*** para poder implementar este modelo. Este método crea un extenso número de árboles de decisión a partir de unos datos iniciales. Con estos datos se calcula el error *MSE* del modelo hasta reducirlo. Luego, cada nueva observación se provee en cada uno de los árboles y el resultado más usual se usará como salida consiguiendo predecir los valores de la variable en cuestión.

Como se ha dicho anteriormente, utilizaremos el método ***randomForest()***, que presenta la siguiente sintaxis:

```
randomForest(x, y, importance, ntree=500, mtry, replace,
classwt=NULL, cutoff, strata, maxnodes, localImp, nPerm=1,
proximity, oob.prox=proximity, norm.votes=TRUE,…)
```

Seguidamente, como en los casos anteriores, se realizará la predicción del consumo utilizando la función ***predict()***.

A continuación, en la *ilustración 20*, se muestra el código implementado en R:

```
################## RANDOM FOREST #################
library(randomForest)

modeloRandomF = randomForest(energyConsum_kWh ~ Temp_MAX + Temp_deRocio + Temp_Humeda_Aire + Ráfagas_Viento,
                             data=tablaEntrenamiento, importance =T, mtry = 3, replace = T, na.action = na.exclude)

# Analizar modelo
print(modeloRandomF)

# Predecimos los consumos de energia
tablaResultado$energiaRandomF = predict(modeloRandomF, tablaResultado)

# Residuo = dato real - prediccion
residuo = tablaResultado$energyConsum_kWh - tablaResultado$energiaRandom
# Estimación varianza del error
estimacion = sqrt(mean(residuo^2))

modeloRandomF$importance
#Gráfico  que muestra la importancia de las variables regresion
varImpPlot(modeloRandomF)
plot(modeloRandomF)
```

Ilustración 20. Código R del modelo de Random Forest

La ejecución del modelo se muestra a continuación:

```
> modeloRandomF$importance
                   %IncMSE IncNodePurity
Temp_MAX         11585.805      671499.0
Temp_deRocio      9349.379      706222.2
Temp_Humeda_Aire 29961.538     1179733.9
Ráfagas_Viento    3466.714      431708.8
```

Esta técnica permite determinar la importancia e influencia que tienen todas las variables independientes o explicativas sobre la variable dependiente a predecir. En la ejecución anterior se muestra dicha importancia medida según dos clasificaciones:

- Según el incremento del error MSE (`%IncMSE`) del modelo cuando la variable en cuestión es intercambiada. Representa un descenso en la precisión de las predicciones cuando la variable es excluida del modelo y es sustituida por otra.
- Según el incremento de pureza de nodos (`IncNodePurity`). Aquellas variables que resultan más útiles logran mayores aumentos en la pureza del nodo.

La *Ilustración 21* muestra de forma gráfica la importancia de las variables independientes después de ejecutar el comando ***varImplot(modeloRandomF)***.

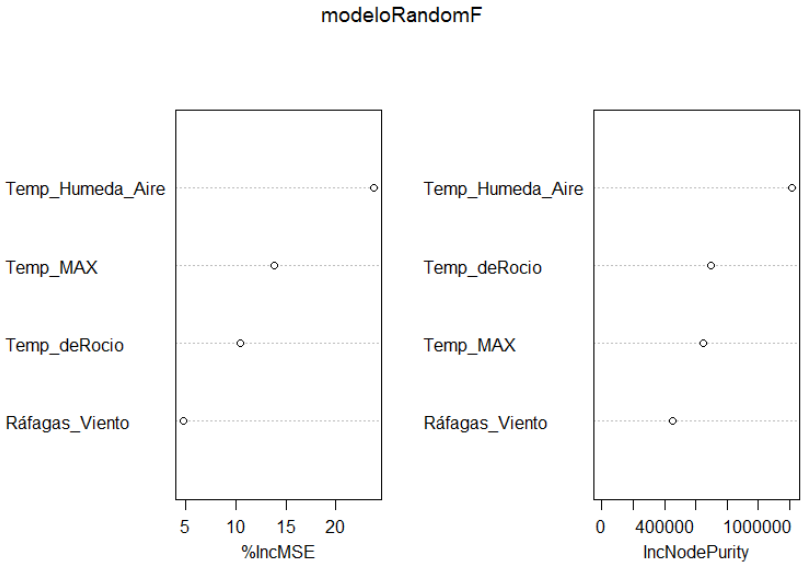

Ilustración 21. Importancia de las variables

Ambos métodos muestran resultados parecidos a la hora de medir la importancia de dichas variables. Las variables que influyen significativamente en el comportamiento del consumo eléctrico son la *Temp_Húmeda_Aire* (temperatura húmeda del aire diaria), la *Temp_MAX* (temperatura máxima diaria) y *Temp_deRocío* (temperatura de rocío diaria). Si estas variables son reemplazadas por otras variables, el error del MSE del modelo se verá aumentado y la pureza del nodo disminuirá considerablemente.

```
> print(modeloRandomF)

Call:
 randomForest(formula = energyConsum_kWh ~ Temp_MAX + Temp_deRocio
 +      Temp_Humeda_Aire + Ráfagas_Viento, data = tablaEntrenamien
to,      importance = T, mtry = 3, replace = T, na.action = na.exc
lude)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 34270.32
                    % Var explained: 31.54
```

Se puede observar que el porcentaje de variabilidad explicado por este método es de un 31,54% y que, frente al modelo de regresión lineal, hay una mejora de más del 20%.

Finalmente, en la *Ilustración 22*, se muestra gráficamente el *error OOB* del modelo utilizando el comando ***plot(modeloRandomF)***.

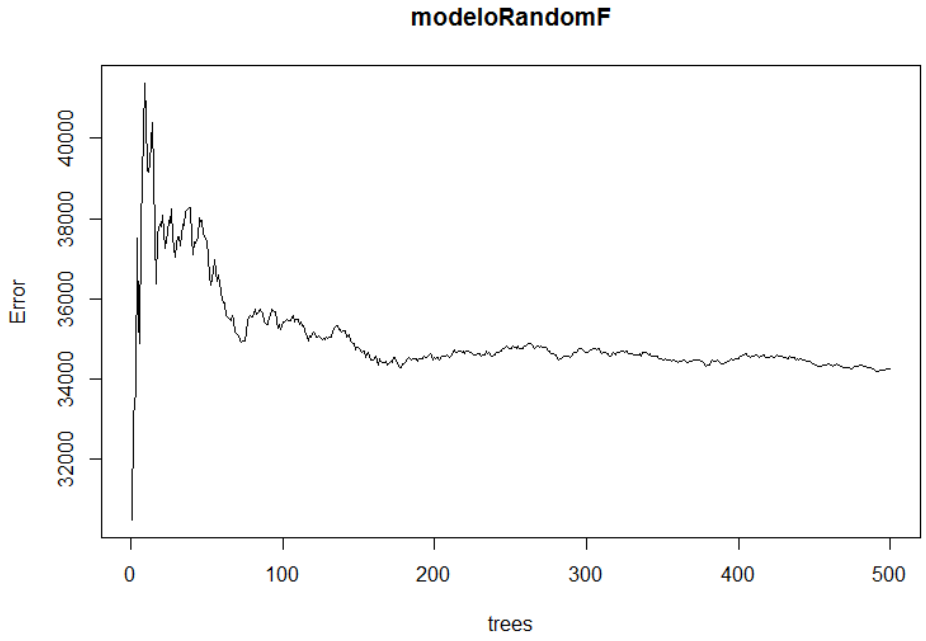

Ilustración 22. Gráfica del error OOB

Se puede observar que el error se estabiliza a partir de los 200 árboles. Lo que significa que no se obtienen resultados mejores usando los 500 árboles que por defecto utiliza el propio software de *R-Studio*.

6.1.7. REGRESIÓN POR K-NN

Finalmente, concluiremos con la implementación del método de regresión por k vecinos más próximos, dónde su fundamento teórico ya fue explicado en el apartado 5.5.3.1.

En el código R, primeramente, tendremos que cargar los paquetes ***library(kknn)*** y ***library(class)*** el cual incluye el método ***train.kknn()*** para poder implementar este modelo. Este permite realizar el entrenamiento del algoritmo *kknn* a través de la validación cruzada, la cual es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición de datos de entrenamiento y prueba.

Como se ha dicho anteriormente, utilizaremos el método ***train.kknn(),*** que presenta la siguiente sintaxis:

```
train.kknn(formula, data, kmax = 11, ks = NULL, distance = 2,
kernel = "optimal", ykernel = NULL, scale = TRUE, contrasts =
c('unordered' = "contr.dummy", ordered = "contr.ordinal"), ...)
cv.kknn(formula, data, kcv = 10, ...)
```

Seguidamente, como en los casos anteriores, se realizará la predicción del consumo utilizando la función ***predict()***.

A continuación, en la *ilustración 23*, se muestra el código implementado en R:

```
################## KNN ##################

#Aplicamos el método Knn
modeloEnergiaKNN5 = train.kknn(energyConsum_kWh ~ Temp_MAX  + Temp_deRocio + Temp_Humeda_Aire + Ráfagas_Viento,
                               data=tablaEntrenamiento, K =5)
# Analizar modelo
summary(modeloEnergiaKNN5)

# Predecimos los consumos de energia
tablaResultado$energiaKNN5 = predict(modeloEnergiaKNN5, tablaResultado)

# Gráfico realidad VS predicción
plot(tablaResultado$energiaKNN5, tablaResultado$energyConsum_kWh, main = "Predicción energía VS Realidad_K-NN",
     xlab = "Predicción Energía", ylab = "Realidad Energía")
```

Ilustración 23. Código R del modelo de K-NN

La ejecución del modelo se muestra a continuación:

```
Call:
train.kknn(formula = energyConsum_kWh ~ Temp_MAX + Temp_deRocio
+     Temp_Humeda_Aire + Ráfagas_Viento, data = tablaEntrenamien
to,     K = 5)

Type of response variable: continuous
minimal mean absolute error: 145.4423
Minimal mean squared error: 32207.8
Best kernel: optimal
Best k: 7
```

Realizada la estimación para este modelo, podemos observar que el valor de la desviación media absoluta (`minimal mean absolute error`) es 145,44. Esta medida indica la precisión de predicción de un método de pronóstico.

Por otro lado, la estimación del error cuadrático medio mínimo (`Minimal mean squared error`) es de 32207,8. Este minimiza el error cuadrático medio, que es una medida común de la calidad del estimador de los valores ajustados de una variable dependiente.

En la *Ilustración 24* se muestra un gráfico que relaciona los valores de la variable explicada "*energyConsum_kWh*" según la predicción obtenida *"energiaKNN5"* en el eje X y los datos en la realidad en el eje Y.

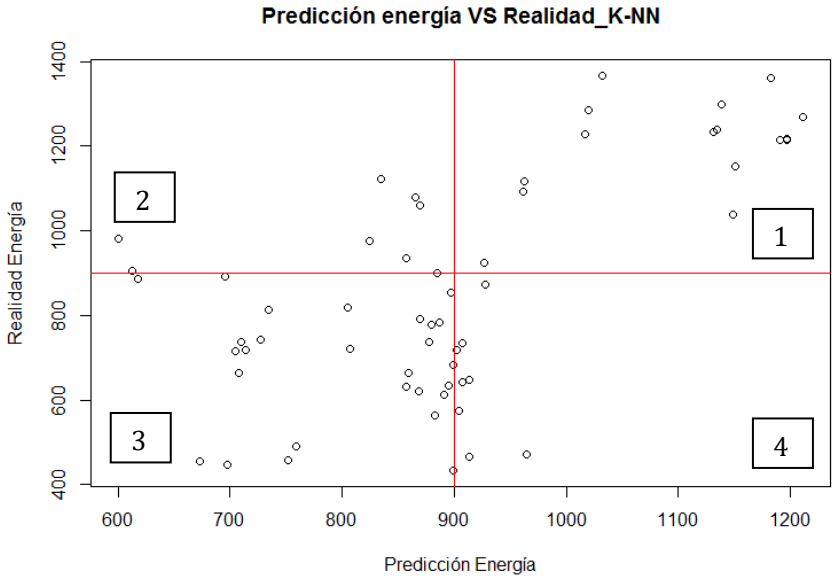

Ilustración 24. Predicción energía VS datos reales modelo K-NN

6.2. ELECCIÓN DEL MEJOR MODELO DE PREDICCIÓN

Después de haber ejecutado todos los modelos predictivos (Regresión lineal simple, Regresión lineal múltiple con todas las variables explicativas y con las variables más significativas, Regresión lineal Robusta con y sin parámetro MM, Random Forest y K-NN), el objetivo principal de este trabajo consistía en hacer una comparativa y quedarnos con aquel método que arroje los mejores resultados sobre el consumo eléctrico que se tendrá en una determinada fecha del futuro.

La *Tabla 3* contiene parte de los datos de la variable "*Energia (kW/h)*" real y la variable resultado correspondiente obtenida de las predicciones de cada modelo sobre la energía resultado.

REALES	PREDICHOS						
Energia (kW/h)	Energia_RLS (kW/h)	Energia_RLM_ todasVbls (kW/h)	Energia_RLM_ VblsMasSig (kW/h)	Energia_RLR	Energia_RLM_MM	Energia_RF (kW/h)	Energia_KNN (kW/h)
1214.58	1224.11	1230.91	1066.74	1094.66	1090.35	1148.01	1196.19
1268.3	1228.76	1086.63	1149.35	1179.39	1176.61	1166.07	1211.26
1217.7	1284.64	1250.71	1323.22	1371.15	1370.30	1082.24	1196.19
1360.08	1093.72	742.03	609.50	620.60	613.36	1154.97	1181.84
1366.35	1042.50	1026.21	1002.39	1032.76	1035.77	1102.27	1031.98
1151.5	1168.23	1204.02	1167.86	1200.98	1201.34	1190.59	1150.64
1214.53	1196.17	1169.38	1051.09	1083.39	1081.55	1081.21	1190.36
1233.51	1163.57	1070.83	1038.83	1075.54	1076.36	1084.76	1130.95
1239.82	1256.70	1267.24	1243.59	1281.85	1282.01	1070.90	1134.18
1299.76	1545.42	1461.84	1523.79	1579.86	1579.31	1062.58	1137.99
1116.85	1061.12	913.67	1140.98	1172.90	1175.34	940.00	962.42
1078.95	930.74	916.90	1025.32	1051.62	1053.44	747.16	865.43
934.62	912.11	737.49	839.03	850.72	849.14	840.07	857.03
906.34	818.97	642.73	717.67	718.95	717.01	700.05	612.45
890.63	828.29	717.81	735.02	735.30	733.85	840.42	695.14

Tabla 3. Resultados obtenidos para cada modelo

A continuación, se mostrarán unos gráficos que representan todos los datos anteriores:

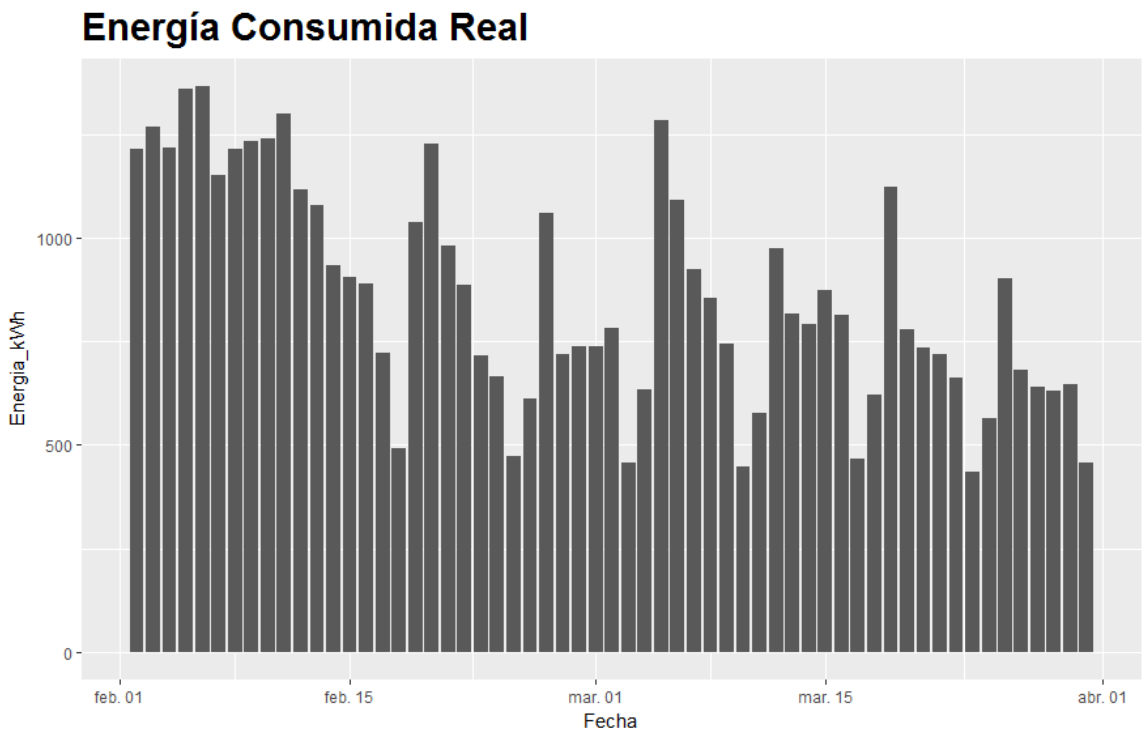

Ilustración 25. Gráfico de la energía real consumida

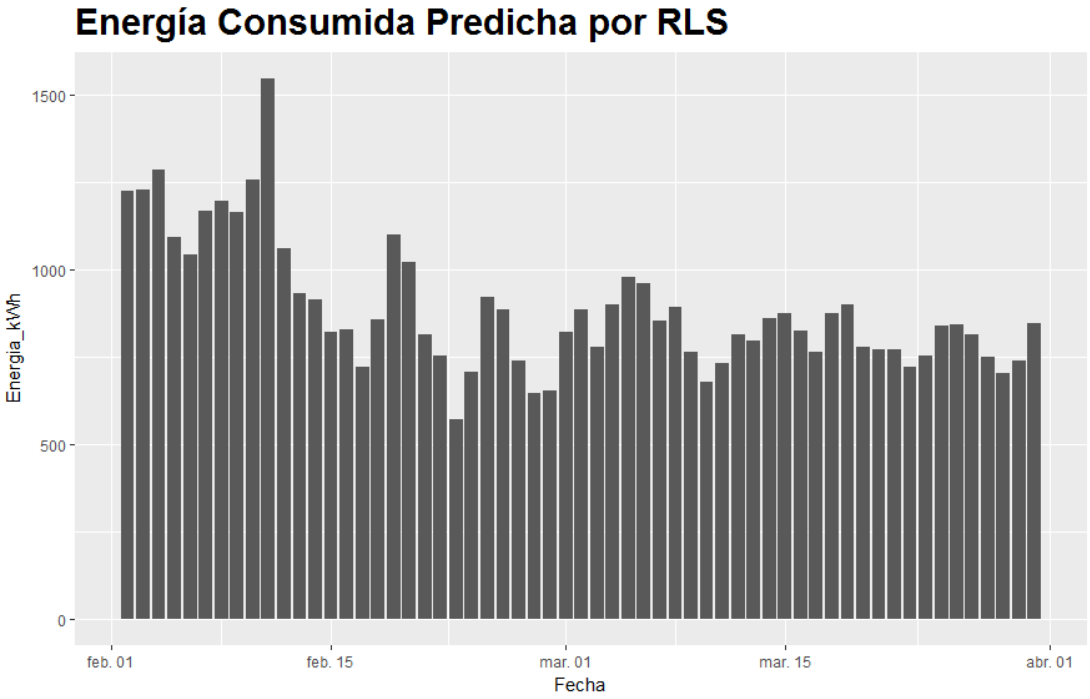

Ilustración 26. Energía predicha por Regresión Lineal Simple

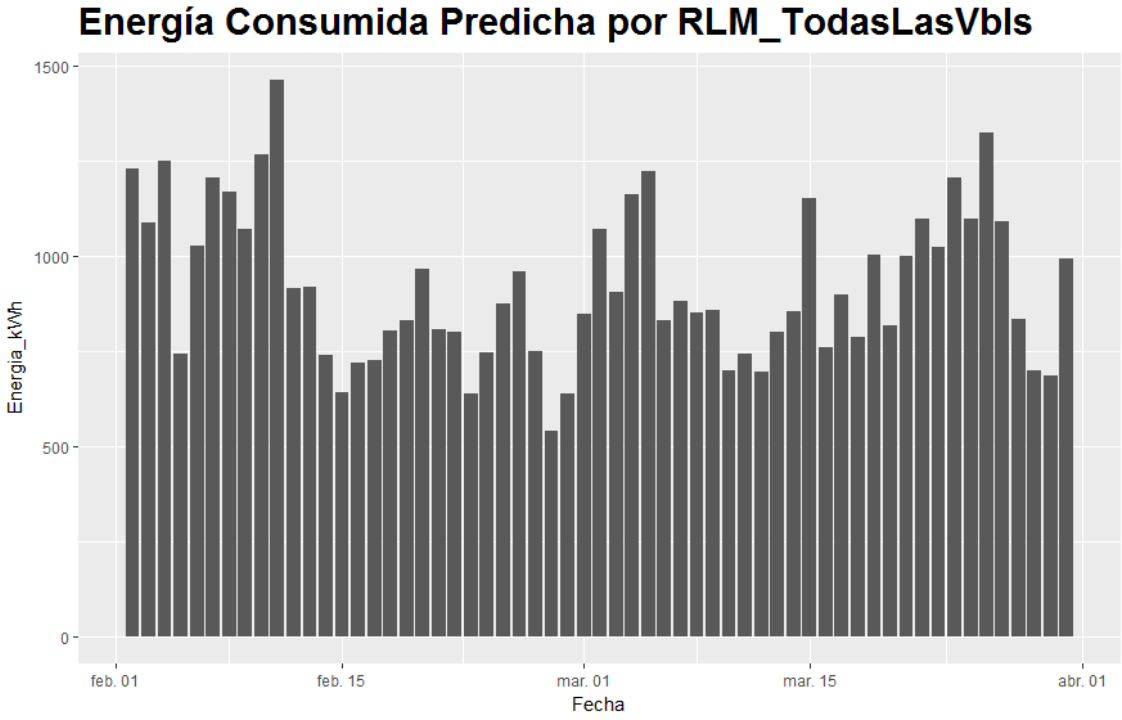

Ilustración 27. Energía predicha por Regresión Lineal Múltiple con todas las variables

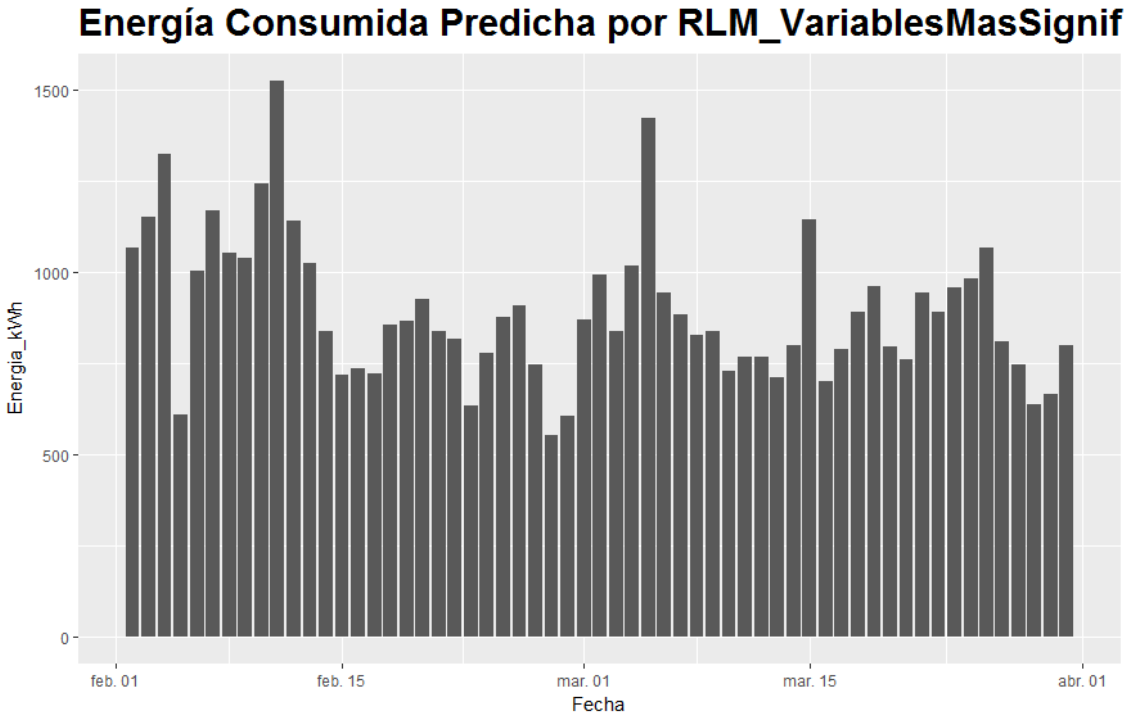

Ilustración 28. Energía predicha por Regresión Lineal Múltiple con variables más significativas

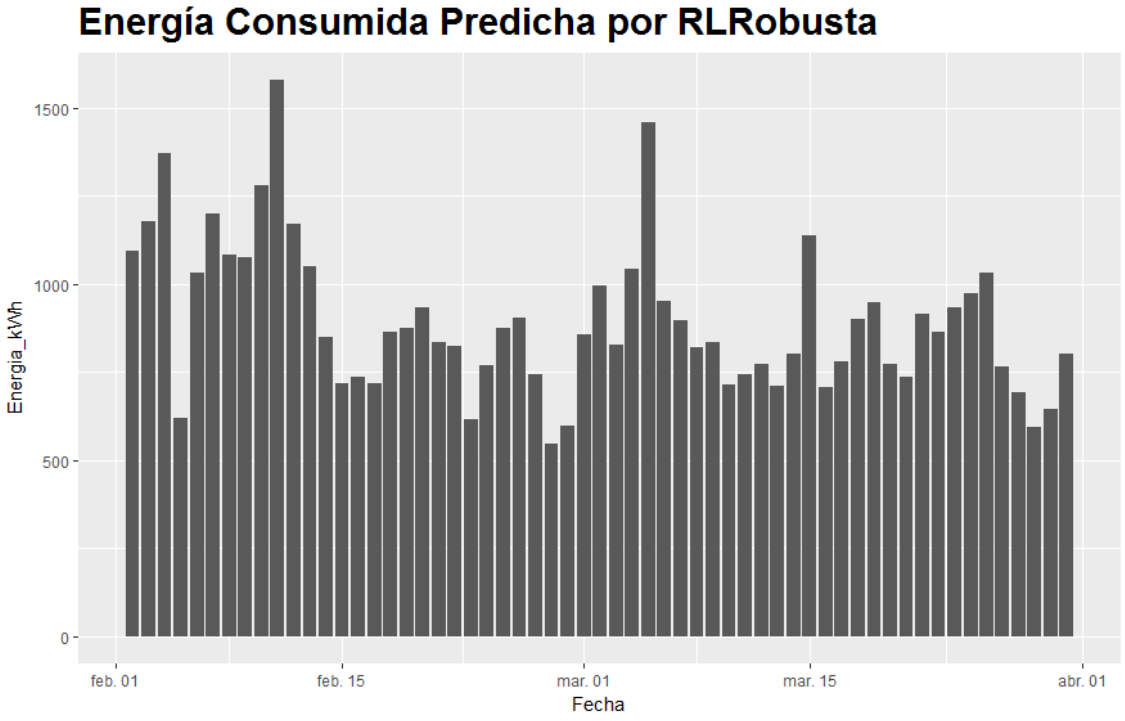

Ilustración 29. Energía predicha por Regresión Lineal Robusta

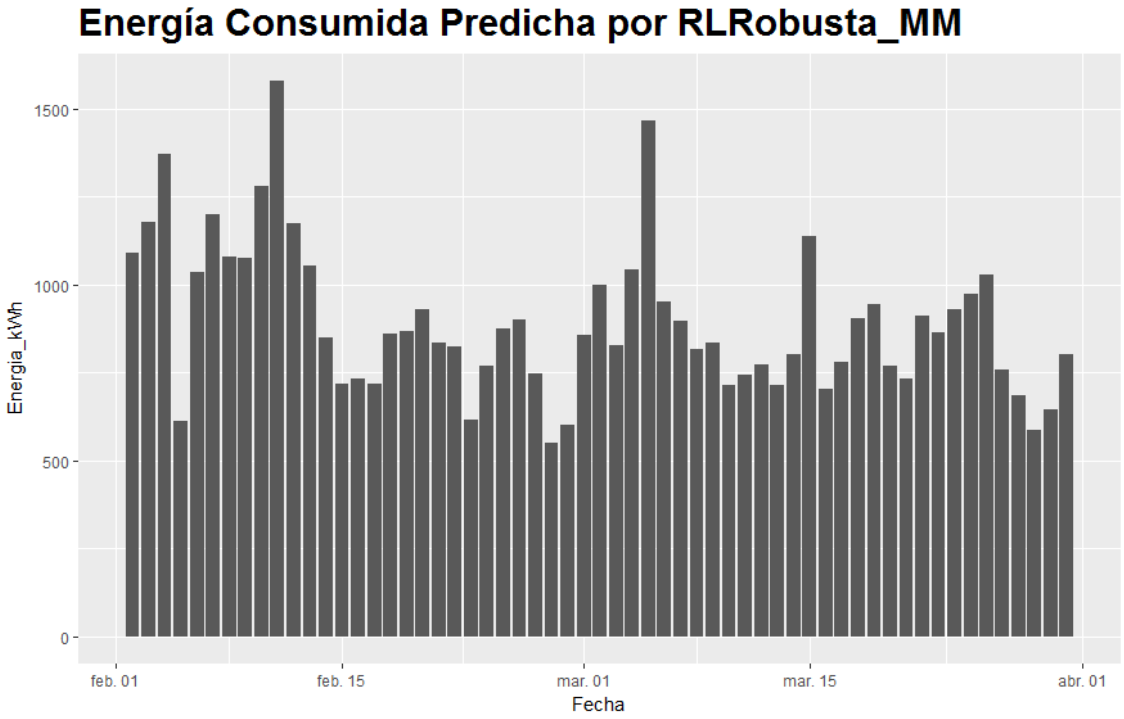

Ilustración 30. Energía predicha por Regresión Lineal Robusta con parámetro MM

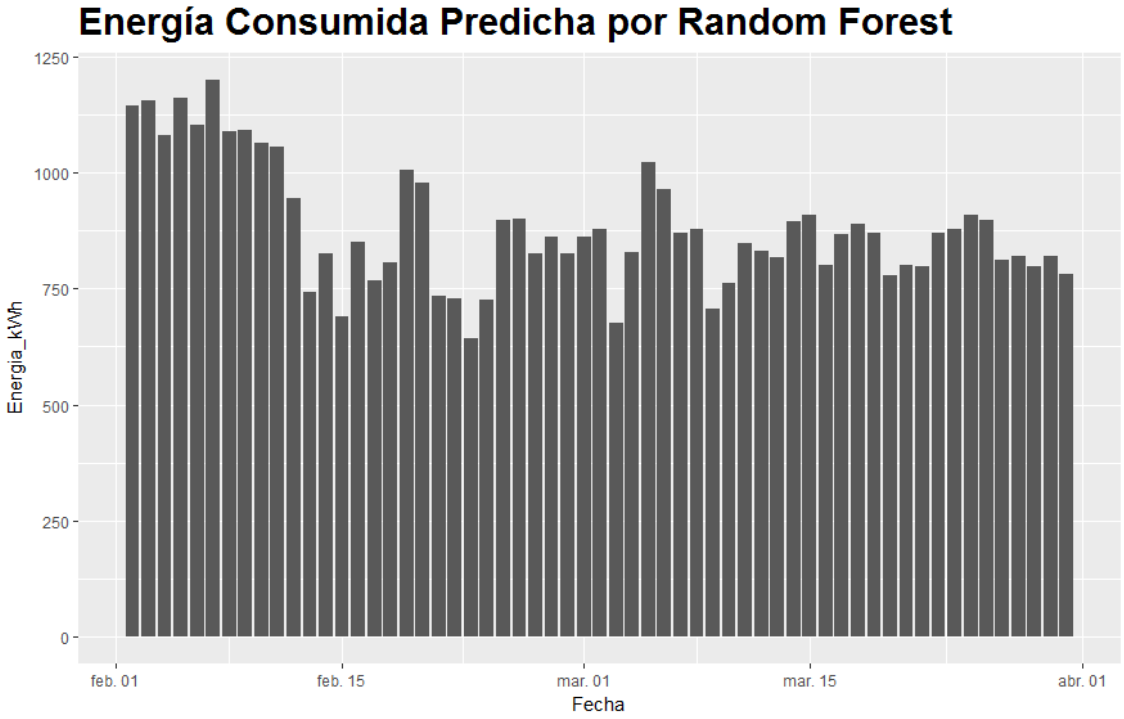

Ilustración 31. Energía predicha por Random Forest

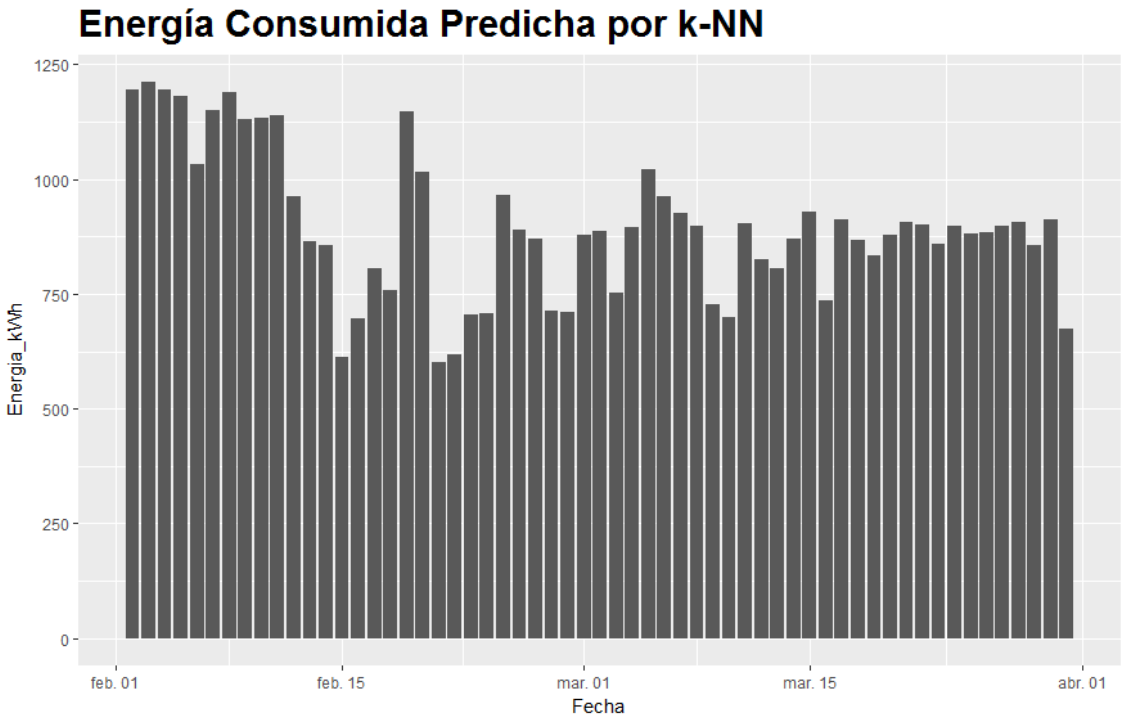

Ilustración 32. Energía predicha por k-NN

Quizás solo con observar los resultados obtenidos en la tabla o gráficos anteriores no podríamos determinar cuál de los modelos es el mejor. Por tanto, se hará el siguiente cálculo para los distintos modelos y así obtener el porcentaje de fallo de cada uno:

Se restarán ambas columnas (la energía real y la predicha) y se guardará el resultado en otra columna. Seguidamente, se calculará la media de todos los valores obtenidos en esa resta y aquel método que tenga la menor media será el modelo más óptimo.

El la *Ilustración 33* se muestra el código implementado en R:

```
tablaResultado$restaRLS = tablaResultado$energyConsum_kwh - tablaResultado$energiaRLS
p_RLS = mean(tablaResultado[["restaRLS"]])

tablaResultado$restaRLM_TodasLasVbls = tablaResultado$energyConsum_kwh - tablaResultado$energiaRLM_TodasLasVbls
p_RLM_TodasLasVbls = mean(tablaResultado[["restaRLM_TodasLasVbls"]])

tablaResultado$restaRLM_VariablesMasSignif = tablaResultado$energyConsum_kwh -
  tablaResultado$energiaRLM_VariablesMasSignif
p_RLM_VariablesMasSignif = mean(tablaResultado[["restaRLM_VariablesMasSignif"]])

tablaResultado$restaRobusta = tablaResultado$energyConsum_kwh - tablaResultado$energiaRobusta
p_Robusta = mean(tablaResultado[["restaRobusta"]])

tablaResultado$restaRobustMM = tablaResultado$energyConsum_kwh - tablaResultado$energiaRobustMM
p_RobustMM = mean(tablaResultado[["restaRobustMM"]])

tablaResultado$restaRandomF = tablaResultado$energyConsum_kwh - tablaResultado$energiaRandomF
p_RandomF = mean(tablaResultado[["restaRandomF"]])

tablaResultado$restaKNN5 = tablaResultado$energyConsum_kwh - tablaResultado$energiaKNN5
p_KNN5 = mean(tablaResultado[["restaKNN5"]])
```

Ilustración 33. Código en R para obtener la probabilidad de error de cada modelo

A continuación, en la *Tabla 4*, se muestra la información obtenida del código anterior donde cada modelo lleva asociado las probabilidades de acierto y fallo correspondientes.

MODELOS PREDICTIVOS	P FALLO (%)	P ACIERTO (%)
Regresión Lineal Simple	29,02656	70,97344
Regresión Lineal Múltiple con todas las variables explicativas	62,72843	37,27157
Regresión Lineal Múltiple con variables más significativas	33,60286	66,39714
Regresión Lineal Robusta	35,58481	64,41519
Regresión Lineal Robusta con parámetro MM	34,56758	65,43242
Random Forest	**24,05095**	**75,94905**
K-NN	36,58994	63,41006

Tabla 4. Tabla comparativa de los modelos

Finalmente, observando los porcentajes de la *Tabla 4*, salta a la vista que el mejor modelo predictivo elaborado para predecir el valor de la energía consumida para un determinado día es el de ***Random Forest*** de entre todos los modelos expuestos por tener menor porcentaje de fallos.

7. CONCLUSIONES

La realización del presente documento de investigación ha tenido un objetivo claro, que es la realización de una comparativa de modelos estadísticos predictivos para la obtención de predicciones sobre el consumo eléctrico partiendo de observaciones y datos históricos de dicho consumo a lo largo de un cierto periodo de tiempo, algo superior a 6 meses, y parámetros del tiempo meteorológico de ese periodo. Con ello se obtiene un patrón de cómo los parámetros meteorológicos influyen sobre el consumo, en unas ciertas condiciones: las que tenía en el periodo "patrón". Si estas condiciones no han variado, se obtiene la predicción del consumo para una fecha posterior conociendo y teniendo las predicciones meteorológicas para esa fecha, y en concreto el valor previsto para los parámetros que son utilizados. Para comparar estas predicciones del consumo, se han tenido en cuenta también datos sobre las predicciones reales de dicho consumo para esa determinada fecha.

Este objetivo se ha conseguido utilizando el lenguaje de programación *R,* ejecutando en el software *R-Studio* todos estos datos mencionados anteriormente utilizando distintas librerías y paquetes junto con los métodos adecuados para la creación de los distintos modelos.

Los modelos implementados han sido: *Regresión Lineal Simple, Regresión Lineal Múltiple* empleando todas las variables explicativas, *Regresión Lineal Múltiple* empleando las variables más significativas, *Regresión Lineal Robusta, Regresión Lineal Robusta* con parámetro *MM*, *Random Forest* y Regresión por *k-Vecinos más Próximos (K-NN)*, descritos todos ellos matemáticamente en los apartados 5.5.2 y 5.5.3.

Apoyándonos en los resultados descritos en el apartado 6, en la *Tabla 5* han sido ordenados de forma decreciente los modelos utilizados anteriormente para hallar la predicción según su porcentaje de fallo.

MODELOS PREDICTIVOS	P FALLO (%)	P ACIERTO (%)
Random Forest	**24,05095**	**75,94905**
Regresión Lineal Simple	29,02656	70,97344
Regresión Lineal Múltiple con variables más significativas	33,60286	66,39714
Regresión Lineal Robusta con parámetro MM	34,56758	65,43242
Regresión Lineal Robusta	35,58481	64,41519
K-NN	36,58994	63,41006
Regresión Lineal Múltiple con todas las variables explicativas	62,72843	37,27157

Tabla 5. Comparación modelos predictivos según porcentajes de acierto y fallo

Como se puede observar en la *Tabla 5*, fue la técnica de *Random Forest* en donde se obtuvieron los mejores resultados con un porcentaje de acierto de casi el 76%, seguida de la técnica de *Regresión Lineal Simple* con un porcentaje de acierto de casi el 71%, lo cual es un poco sorprendente ya que se estimaban mejores resultados para el modelo de *Regresión Lineal Múltiple con variables más significativas* puesto que este último utiliza un mayor número de variables explicativas para predecir el resultado.

Por otro lado, decir que el modelo que arroja peores resultados es el de *Regresión Lineal Múltiple con todas las variables explicativas*. Esto podría ser por un problema de *Overfitting*, Esto se produce cuando el modelo es excesivamente complejo y el problema es que recoge demasiados detalles de los datos de entrenamiento, pero por lo general no se repiten en los datos de test, por ende, no será un buen modelo.

Los modelos de predicción, además de reflejar el comportamiento de la variable final, estiman aquellas variables de entrada que tienen un impacto mayor sobre la variable a predecir estudiada e influyen a la hora de construir el modelo predictivo y calcular los resultados. Todos los modelos coinciden en que las variables más significativas y con un peso mayor en el comportamiento del consumo eléctrico son: la temperatura máxima obtenida por día, la temperatura de rocío, la temperatura húmeda del aire y la velocidad de las ráfagas de viento.

Por todo ello, puede llegar a ser de gran ayuda la utilización de estos modelos ya que de esta forma podrá saberse qué variables son más importantes a la hora de realizar una predicción y obtener buenos resultados.

En conclusión, si se logra alcanzar el objetivo de este proyecto logrando dar respuesta al comportamiento del consumo eléctrico y con ello disminuir los sobrecostes relacionados con el consumo, se estará ayudando, por un lado, a la sociedad en general a consumir menos energía de la que debe llevando esto a un ahorro económico y, por otro lado, a disminuir el efecto de cambio climático, protegiendo el medio ambiente con un uso más eficiente de los recursos.

8. PLAN DE TRABAJO

En este apartado se explicarán las tareas y fases a lo largo de la elaboración del proyecto y la distribución del tiempo de estas representadas en un diagrama de Gantt.

8.1. DESCRIPCIÓN DE TAREAS, FASES.

Para la realización del proyecto se han distinguido 4 fases o tareas:

- Estudio y documentación

En esta fase se investigó sobre cómo se realizan análisis en los datos y dónde pueden almacenarse y acceder a ellos. A continuación, se hizo un extenso estudio sobre la minería de datos y en qué consiste, analizando los distintos tipos de análisis que existen. Finalmente, se realizó una búsqueda y comprensión teórica de los modelos predictivos existentes.

- Aprendizaje del lenguaje R

Se realizó un pequeño curso introductorio a este lenguaje para familiarizarse con él. Se aprendieron las técnicas de tratamiento de datos y estadística descriptiva con *R*, así como el manejo de los datos de entrada del proyecto. Finalmente, se realizaron pruebas iniciales con la interfaz R-Studio.

- Creación de los modelos predictivos

Se elaboraron los 5 modelos con sus variantes: modelo de Regresión Lineal Simple, Regresión Lineal Múltiple, Regresión Lineal Robusta, Random Forest y KNN. Para finalizar esta tarea se realizaron los análisis de los resultados y conclusiones.

- Redacción de la memoria

La última tarea consistió en la elaboración de la memoria con todos los apartados correspondientes.

8.2. PLANIFICACIÓN Y DIAGRAMA DE GANTT

Todas las tareas anteriormente mencionadas se distribuirán temporalmente en la *Tabla 6*, donde a finales de enero de 2018 se comenzó la búsqueda activa de información de este trabajo.

En el Diagrama de Gantt se representará, con fecha de inicio y fin, la duración de cada una de las tareas realizadas para la elaboración del trabajo.

TAREA	INICIO	FIN	DURACIÓN
Estudio y documentación	**mar 23/01/18**	**lun 26/02/18**	**35 días**
1. Open Data	mar 23/01/18	jue 25/01/18	3 días
2. Minería de Datos y Análisis Predictivo	lun 29/01/18	jue 01/02/18	4 días
3. Comprensión teórica modelos predictivos	lun 12/02/18	lun 26/02/18	11 días
Aprendizaje del lenguaje R	**jue 01/03/18**	**lun 02/04/18**	**35 días**
4. Curso introductorio al lenguaje R	jue 01/03/18	vie 16/03/18	12 días
5. Estadística descriptiva y tratamiento de datos en R	lun 19/03/18	jue 22/03/18	4 días
6. Manejo de los datos de entrada del proyecto	lun 26/03/18	mié 28/03/18	3 días
7. Pruebas iniciales con R-Studio	jue 29/03/18	lun 02/04/18	3 días
Creación de modelos predictivos	**mié 04/04/18**	**vie 11/05/18**	**38 días**
8. Modelo Regresión Lineal Simple	mié 04/04/18	jue 04/04/18	2 días
9. Modelo Regresión Lineal Múltiple	lun 09/04/18	mar 10/04/18	2 días
10. Modelo Regresión Lineal Robusta	mié 18/04/18	jue 19/04/18	2 días
11. Modelo Random Forest	mié 25/04/18	mié 02/05/18	6 días
12. Modelo K-NN	lun 07/05/18	vie 11/05/18	5 días
Redacción de la memoria	**lun 14/05/18**	**mié 20/06/18**	**37 días**
13. Resumen, Introducción, Contexto, Objetivos y Beneficios	lun 14/05/18	mié 16/05/18	3 días
14. Bases de datos NoSQL, Open Data, Minería de Datos y Análisis Predictivo	mié 16/05/18	lun 21/05/18	4 días
15. Métodos predictivos. Explicación teórica	lun 21/05/18	mar 29/05/18	7 días
16. Implementación de modelos en R	mar 29/05/18	vie 08/06/18	11 días
17. Resultados y Conclusiones	vie 08/06/18	lun 11/06/18	4 días
18. Plan de Trabajo y Aspectos Económicos	lun 11/06/18	mar 12/06/18	2 días
19. Reuniones	jue 14/06/18	mié 20/06/18	5 días
TOTAL:	**mar 23/01/18**	**mié 20/06/18**	**145 días**

Tabla 6. Distribución de tareas del proyecto

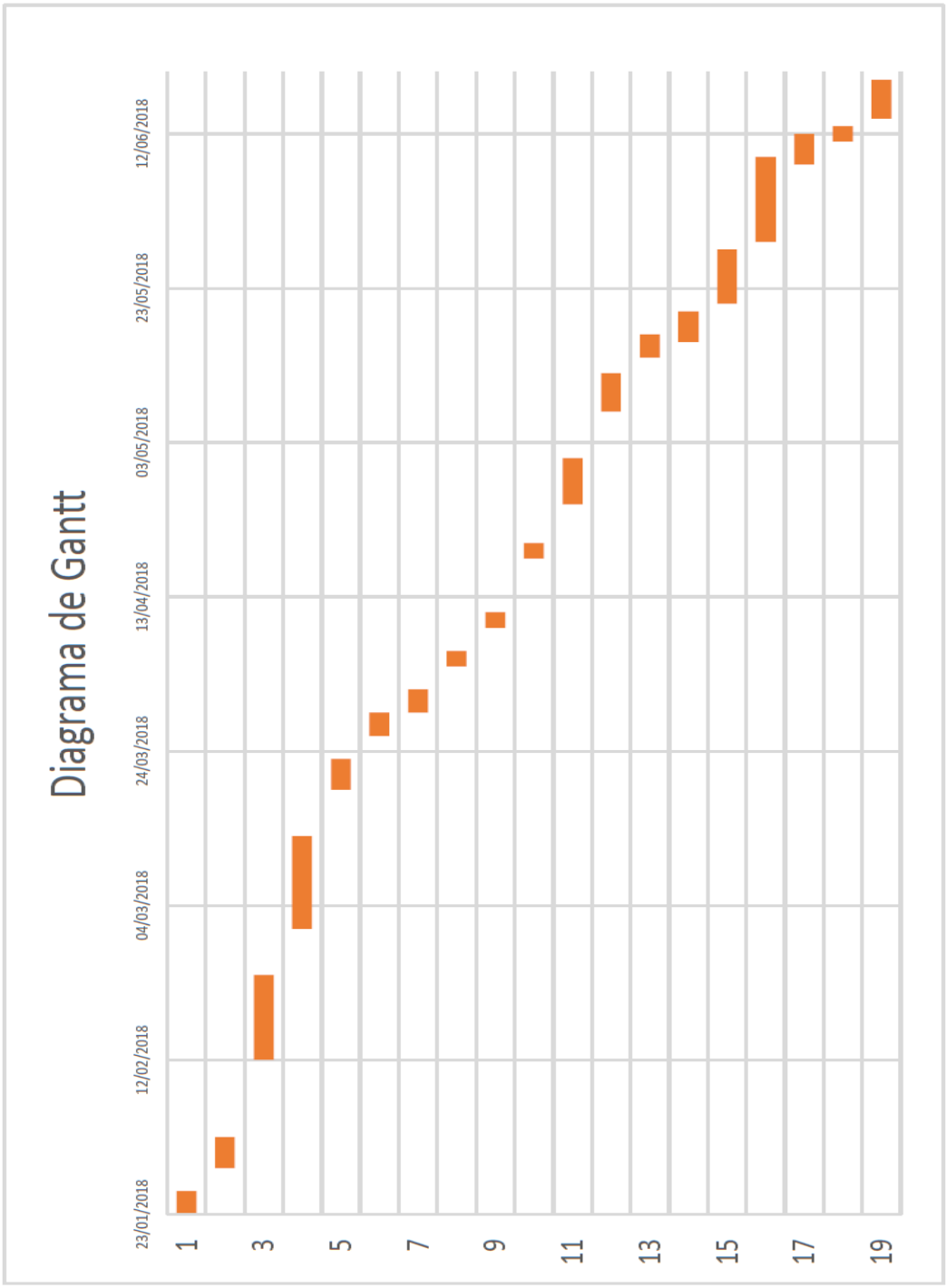

Ilustración 34. Diagrama de Gantt de las tareas del proyecto

9. ASPECTOS ECONÓMICOS

En este apartado se procederá a desglosar y explicar los análisis económicos en el presente proyecto.

9.1. DESCRIPCIÓN DEL PRESUPUESTO

En la preparación del presupuesto para este proyecto de fin de grado, dados los pocos recursos necesarios durante la fase de elaboración, se tendrán únicamente en cuenta los recursos humanos y los materiales. En las *Tabla 7* y *Tabla 8* se muestran la estructura del presupuesto y los importes asociados a cada uno de los recursos.

- **Recursos materiales**

NOMBRE	DESCRIPCIÓN	PRECIO	UNIDADES	IMPORTE TOTAL (€)
Ordenador portátil	HP 840 Elitebook	460 €	1	460 €
Licencia softwate R	Libre	0 €	1	0 €
Entorno desarrollo R-Studio	Libre	0€	1	0 €
Curso programación R	Curso conceptos básicos	70 €		70 €
Material papelería	Cuadernos, bolígrafos, hojas…	10 €		10 €
TOTAL				**540 €**

Tabla 7. Presupuesto Recursos Materiales

- **Recursos humanos**

DESCRIPCIÓN	PRECIO/HORA (€)	HORAS	IMPORTE TOTAL (€)
Estudio y documentación	11	105	1155 €
Aprendizaje del lenguaje R	11	105	1155 €
Creación de modelos predictivos	11	114	1254 €
Redacción de la memoria	11	111	1221 €
Tutorías con el tutor	30	15	450 €
Revisión del trabajo por el tutor	30	7	210 €
TOTAL		**457**	**5.445 €**

Tabla 8. Presupuesto Recursos Humanos

En la *Tabla 9* se muestra la suma de los gastos totales de los recursos humanos y los recursos materiales.

NOMBRE	IMPORTE TOTAL (€)
Recursos materiales	540 €
Recursos humanos	5.445 €
TOTAL	**5.985 €**

Tabla 9. Presupuesto Total

10. BIBIOGRAFIA

[1] Rodríguez Suárez, Yuniet, Díaz Amador, Anolandy, Herramientas de Minería de Datos. Revista Cubana de Ciencias Informáticas [en linea] 2009, 3 (Julio-Diciembre) : [Fecha de consulta: 30 de mayo de 2018] Disponible en: <http://www.redalyc.org/articulo.oa?id=378343637009> ISSN 1994-1536

[2] W. W. Eckerson, «Extending the Value of Your Data Warehousing Investment,» The Data Warehouse Institute, 2007.

[3] R. A. Moreno, «Análisis predictivo: con Big Data el futuro no se predice, se cambia,» 2016.

[4] MATHWORKS, «https://es.mathworks.com/discovery/predictive-analytics.html» [En línea].

[5] ISOVER, «https://www.isover.es/sostenibilidad/la-edificacion-sector-clave» 2015. [En línea].

[6] OPEN DATA HANDBOOK, «http://opendatahandbook.org/guide/es/what-is-open-data/» [En línea].

[7] 5STARDATA, «http://5stardata.info/es/» [En línea].

[8] OPEN WEATHER MAP, «https://openweathermap.org/» [En línea].

[9] Hayati, M.; Mohebi, Z.: Application of Artificial Neural Networks for Temperature Forecasting. http://waset.org/publications/8486/application-of-artificial-neural-networks-for-temperature-forecasting. (2007). [En línea]

[10] M. B. R. V. M. V. B. J. A. P. C. M. Coronado Arjona, «Estudio comparativo de técnicas de minería de datos para la predicción de rutas de huracanes,» TECNOLOGÍA EDUCATIVA REVISTA CONAIC, 2017.

[11] Hernández Leal, E. J., Duque Méndez, N. D. y Moreno Cadavid, J. M. (2016). Generación de pronósticos para la precipitación diaria en una serie de tiempo de datos meteorológicos. Ingenio Magno, 7(1), 144-155.

[12] P. Recuero, «LUCA,» November 2017. [En línea]. Available: http://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html.

[13] Nyce, Charles (2007), Predictive Analytics White Paper, American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

[14] GESTIÓN DE OPERACIONES, «https://www.gestiondeoperaciones.net/proyeccion-de-demanda/como-utilizar-una-regresion-lineal-para-realizar-un-pronostico-de-demanda/» [En línea].

[15] Regresión Lineal Múltiple, «http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140128_RegresionMultiple.pdf» [En línea].

[16] Maronna, R.; D. Martin and V. Yohai (2006). Robust Statistics: Theory and Methods. Wiley.

[17] Regresión Robusta, «https://es.wikipedia.org/wiki/Regresi%C3%B3n_robusta» [En línea].

[18] UAM, «http://www.uam.es/personal_pdi/ciencias/jspinill/CFCUAM2014/Trees-CFCUAM2014.html » [En línea].

[19] SLIDEPLAYER, «http://slideplayer.es/slide/11895856/ » [En línea].

[20] KNN y prototipos, «http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/KNNyPrototipos.pdf» [En línea].

[21] WEB_UAE, «http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_lineal_multiple_3.pdf» [En línea].

[22] Eckerson, Wayne (10 de mayo de 2007), Extending the Value of Your Data Warehousing Investment, The Data Warehouse Institute

[23] Wikipedia, 2007. [En línea]. Available: https://es.wikipedia.org/wiki/An%C3%A1lisis_predictivo.

[24] P. C. A. Charles Nyce, «Predictive Analytics White Paper,» American Institute for CPCU, 2007.

[25] Arahal, M. R., Berenguel Soria, M., & Rodríguez Díaz, F. (2006). *Técnicas de predicción con aplicaciones en ingeniería Universidad de Sevilla.*

[26] González, W. J. (2003). *Racionalidad, historicidad y predicción en herbert A. simon Netbiblo.*

[27] Pérez López, C. (2016). *Técnicas avanzadas de predicción. Garceta.*

[28] Valencia Delfa, J. L., & Díaz Llanos, F. J. (2004). *Métodos de predicción en situaciones límite. La Muralla.*