UNIVERSITAT OBERTA DE CATALUNYA (UOC)

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (DATA SCIENCE)

TRABAJO FINAL DE MÁSTER

Predicción de consumos eléctricos en Inglaterra, Gales y Escocia a través de datos de medidores inteligentes

---

Autor: Jorge Arias Martín

Tutor: Dr. Sergi Trilles Oliver

Profesor: Dr. Jordi Casas Roma

Madrid, 9 de Junio de 2019

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción de consumos eléctricos en Inglaterra, Gales y Escocia a través de datos de medidores inteligente*
Nombre del autor:	*Jorge Arias Martín*
Nombre del consultor/a:	*Sergi Trilles Oliver*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	06/2019
Titulación::	*Máster Universitario en Ciencia de Datos*
Área del Trabajo Final:	*Minería de datos y Machine Learning*
Idioma del trabajo:	Español
Palabras clave	*Supervised Learning, Tensorflow, Deep Learning*

Dedicado a Beatriz, Álvaro y Mateo, por ser tan extraordinarios

Agradecimientos

Deseo expresar mi especial agradecimiento:

A mi tutor de proyecto el **Dr Sergio Trilles Oliver**, por su dedicación
A Dr. Jordi Casas y el resto de la comunidad educativa de la UOC
A todas las personas que con sus escritos científicos han ayudado al desarrollo del conocimiento y que han sido citadas.

Gracias por estar ahí y creer en la ciencia.

Resumen

Este TFM tiene como objetivo la predicción de los consumos energéticos de las distintas zonas geográficas de Londres, Gales y Escocia, determinando qué variables climáticas o días de semana o festivo determinan los cambios de consumo en la población

Con la nueva normativa europea que obliga a los proveedores de energía, a desplegar medidores inteligentes en todas las viviendas, se pueden obtener información utilizable a la hora de predecir con mayor precisión el consumo energético en distintas áreas geográfica.

Las compañías compran energía en el mercado energético, lo que supone que la energía no consumida al no poder ser almacenada, se perderá, lo que implica la generación de nueva energía a través de los medios tradicionales con el consiguiente perjuicio para el planeta y el calentamiento global, al ser ésta un bien escaso

Esta pérdida energética puede minimizarse si se racionaliza la compra más equilibrada con el consumo que ejercerá la población gracias a predicción obtenida a través de algoritmos.

Esta compra puede predecirse a partir de los consumos realizados en las mismas épocas de históricos, y cruzarse con datos del clima, temperatura, festivos y otras fuentes de datos para predecir con mayor exactitud que consumo se realizará durante el siguiente periodo

Palabras Clave: Predicción de consumos energéticos, LSTM, ARIMA

Abstract

The objective of this TFM is to predict the energy consumption of the different geographical areas of London, Wales and Scotland, determining which climatic variables or weekdays or holidays determine the changes in consumption in the population

With the new European regulation that forces energy suppliers to deploy smart meters in all homes, usable information can be obtained when predicting more accurately the energy consumption in different geographical areas.

Companies buy energy in the energy market, which means that energy not consumed as it can not be stored will be lost, which implies the generation of new energy through traditional means with the consequent damage to the planet and warming global, as this is a scarce resource

This energy loss can be minimized if the more balanced purchase is rationalized with the consumption that the population will exercise thanks to prediction obtained through algorithms.

This purchase can be predicted from the consumptions made in the same times of historical, and cross with data of the climate, temperature, holidays and other sources of data to predict with greater accuracy what consumption will be made during the next period

Keywords: Forecast Energy Consumption, LSTM, ARIMA

Indice General

Agradecimientos 6

Resumen 7

Abstract 8

Prolegómeno 12

Introducción 13

Transporte y distribución 15

Impacto Ambiental de la generación de energía 16

Motivación 18

Objetivos del trabajo 19

Método 20

Enfoque 21

Tecnología utilizada 22

Planificación del Trabajo 23

Estado del arte 24

Investigación 31

Desarrollo de trabajo 32

Estudio de los datos 34

Adaptación de los datos 46

Datos del clima 47

Temperatura 47

Humedad 48

Cielos cubiertos 49

Visibilidad 50

Velocidad del viento 51

Índice UV 52

Rocío 53

Fase Lunar 54

Presión atmosférica 55

Conclusión de las variables 56

El Dataset final tiene las siguientes columnas: 59

Algoritmo ARIMA 60

Algoritmo LSTM 63

Resultados 64

Epílogo 67

Conclusiones 68

Trabajos Futuros 69

Acrónimos 70

Bibliografía 71

Anexos 74

Índice de figuras (Norma APA)

1-Recorrido de la energía. 13
2-Energías Primarias (Recuperado de http://huellasdearquitectura.wordpress.com). 14
3-Red de transporte eléctrico basado en cables (Comprada en https://stock.adobe.com). 15
4-Central de distribución eléctrica de las Rozas, Madrid. 15
5-Huella ecológica de la energía (Comprada en https://stock.adobe.com). 16
6-Planificación de trabajo (Realizado en Openproject). 23
7-Ejemplo componente temporal. 24
8-Ejemplo de ventas estacionarias con proyección tendencial. 25
9-Cálculo de predicción Winters 26.
10-Red Neuronal Feedforward Obtenida de Reserchgate .27
11-Representación matemática de RNN con memoria. 28
12-Estadística del dataset obtenido de acorn_details.csv. 34
13-Primeros registros del fichero acorn_details.csv. 35
14-Items en cada columna del fichero acorn_details.csv. 35
15-Estadística del dataset obtenido de informations_households.csv. 36
16-Detalle de los primeros 10 registros del fichero informations_households. 36
17-Items en cada columna del fichero informations_households.csv. 36
18-Estadística del dataset obtenido de uk_bank_holidays.csv. 37
19-Primeros 10 registros del fichero uk_bank_holidays. 37
20-Items en cada columna del fichero uk_bank_holidays.csv. 37
21-Estadística del dataset obtenido de weather_daily_darksky.csv. 38
22-Primeros 10 registros del fichero weather_daily_darksky. 39
23-Items en cada columna del fichero weather_daily_darksky.csv. 40
24-Estadística del dataset obtenido de weather_hourly_darksky.csv. 41
25-Primeros 10 registros del fichero weather_hourly_darksky. 42
26-Items en cada columna del fichero weather_hourly_darksky.csv. 43
27-Estadística del dataset obtenido de energy.csv. 44
28-Detalle de los primeros 10 registros del fichero energy. 44
29-Ítems en cada columna del fichero energy.csv. 45
30-Primeros datos del dataset energy. 46
31-Gráfica de Inconsistencia de los datos consumidos por los hogares. 46
32-Comparativa entre Energía y Temperatura. 47
33-Comparativa entre Energía y Humedad. 48
34- Comparativa entre Energía y Cielos cubiertos. 49

35-Comparativa entre Energía y Visibilidad. 50
36-Comparativa entre Energía y Velocidad del viento. 51
37-Comparativa entre Energía e Índice UV. 52
38-Comparativa entre Energía y Rocio. 53
39-Comparativa entre Energía y Fase Lunar. 54
40-Comparativa entre Energía y Presión Atmosférica. 55
41-Matriz de correlación. 56
42-Valores obtenidos a la hora de discretizar valores del clima. 57
43-Gráfica de relación entre variables. 58
44-Detalle de Dataset uk_bank_holidays. 58
45-Separación datos de test y entrenamiento. 59
46-Primeros datos de dataset test. 60
47-Representación de Datos de entrenamiento junto a la predicción. 60
48-Primeros datos del dataset de entrenamiento. 61
49-Gráfica de datos de test junto con la predicción. 61
50-Detalle de dataset con las predicciones hasta en 7 días. 62
51-Evolución de entrenamiento. 62
52-Representación de las diferencias entre los datos de entrenamiento y la predicción. 63
53-Representación de los datos de entrenamiento y la predicción con LSTM. 64

Parte I

Prolegómeno

Capítulo 1

Introducción

El proceso desde la producción de energía eléctrica hasta que es consumida por los usuarios en los distintos entornos, industriales o domésticos, siguen el recorrido ; Las empresas generadoras de electricidad, obtienen la energía eléctrica a través de distintos medios como energía primaria (Gómez Expósito, 2000).

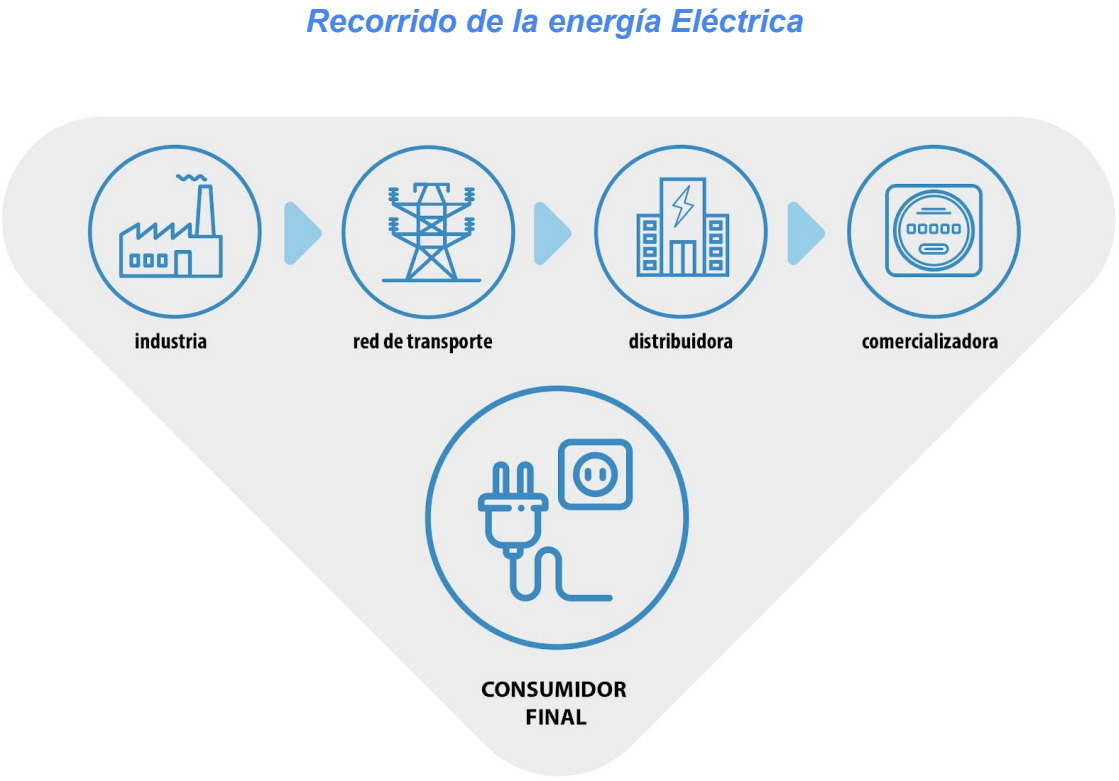

Figura 1. Recorrido de la energía eléctrica.

Posteriormente las empresas dedicadas al transporte, llevan a través de distintos medios la electricidad hasta las localidades donde serán consumidas (Ley 24/2013, de 26 de Diciembre, del Sector Eléctrico, 2013).

Las empresas distribuidoras son las responsables de transportar la energía desde las redes los centros de transformación (instalaciones técnicas que dotan a la energía de condiciones más óptimas para hacerla llegar hasta las redes de uso) hasta los clientes finales (Ley 24/2013, de 26 de Diciembre, del Sector Eléctrico, 2013).

Las comercializadoras son las que se encargan de comprar la energía al mercado pool y venderla al consumidor final a través de sus tarifas. Es la entidad que mantiene la comunicación directa con el cliente final (Cayetano and Ramón, 2010).

Energía primaria

Una fuente de energía primaria es aquella que se encuentra presente en la naturaleza y se transforma para su consumo final, pudiendo tratarse de algunas de las siguientes: Energía Térmica, Energía Nuclear, Energía Hidráulica, Energía Solar, Energía Eólica, Energía Geotérmica, Energía Biomasa, Energía Mareomotriz (Gómez Expósito, 2000) (Figura 2).

Para poder aprovechar el potencial de la energía primaria, se debe de convertir en una fuente de energía secundaria o intermedia (electricidad o combustible) que permita su consumo (Báez-Matos, Rodríguez and Abreu, 2018). A diferencia de otras energía, la energía eléctrica, no puede almacenarse (Nandwani, 2005). Esto supone que la producción, transporte y distribución deba ajustarse a la demanda y deba de realizarse de forma equilibrada.

Figura 2. Energías Primarias como Energía Térmica, Energía Nuclear, Energía Hidráulica, Energía Solar, Energía Eólica, Energía Geotérmica, Energía Biomasa, Energía Mareomotriz.

Transporte y distribución

Una vez la electricidad ha sido generada según la demanda, las centrales eléctricas deben de volcar la energía en la red de transporte, trasladando la electricidad de forma ordenada a la distribuidora asignada a nuestra localidad a través de la infraestructura desplegada recorriendo grandes distancias (Antolin, 1988). Los responsables de este transporte y distribución son las empresas distribuidoras, estando asignadas por zonas geográficas siendo estas las responsables del mantenimiento de la infraestructura (Ley 24/2013, de 26 de Diciembre, del Sector Eléctrico, 2013) (Figura 3).

Figura 3, Red de transporte energético basado en cables aéreos.

Comercializadoras

Las comercializadoras “compran” energía en el mercado para organizar la producción energética de alguna de las fuentes disponibles generando un gran impacto ambiental (Tsoutsos, Frantzeskaki and Gekas, 2005). En el caso de no ser consumida por los clientes de la comercializadora por no haber previsto o haber cambiado las variables que influyen en los hábitos de consumo, la energía simplemente se desperdicia (*El sistema eléctrico español: diversificado, sobredimensionado, aislado… - El Blog de Ignacio Mártil*, 2016) (Figura 4).

Figura 4. Central de distribución eléctrica de Las Rozas, Madrid.

Impacto Ambiental de la generación de energía

En ninguna fase del proceso existe la posibilidad de acumular la energía para ponerla a disposición de la infraestructura energética de manera económica y eficiente (Martín Chicharro, 2016), y esto es debido a la imposibilidad actual de almacenarla en alguno de los puntos del proceso, de forma que permitan el abastecimiento instantáneo en momentos de gran demanda en la red energética. El cálculo para determinar qué cantidad de energía se compra está basado en históricos de años anteriores, una vez superado el consumo de dicha compra, el precio de la cantidad de energía dispuesta tiene un precio superior, así pues, para cambiar esta tendencia es necesario calcular esta compra con la mayor precisión posible (Roos Fraga, 2019).

La compra realizada por la comercializadora, y determinada por el histórico de consumo, se puede afinar aún más si añadimos variables externas como el clima o fiestas locales siendo el objetivo del presente TFM determinar con precisión la previsión de compra y compararla con el consumo que posteriormente se ha realizado. De esta manera la empresa Distribuidora comprará menos energía para ajustar mejor sus ganancias, lo que significa que se desperdicia menos energía y en consecuencia minimizará el impacto ambiental (Figura 5).

Figura 5. Huella ecológica de la energía.

Los numerosos hogares e industrias, como consumidores finales de la energía eléctrica, disponen de medidores inteligentes que permiten informar del consumo a la empresa distribuidora, esto es debido a la normativa europea que ha solicitado a los gobiernos, estos dispositivos como parte de las medidas de mejorar nuestro suministro de energía para hacer frente al cambio climático como indica la Directiva 2009/72/ce del Parlamento Europeo (Parlamento Europeo y Consejo de la Unión Europea, 2009) .

Para estudiar los objetivos se toma las lecturas de los medidores inteligentes de 5.567 hogares de Londres que participaron en el proyecto de Low Carbon Networks entre noviembre de 2011 y febrero de 2014 asociados al consumo eléctrico únicamente, estos datos han sido publicados en el dataset ***Smart meter data from London area*** dentro de la competición de algoritmos de Kaggel https://www.kaggle.com/jeanmidev/smart-meters-in-london

Capítulo 2

Motivación

La investigación basada en los datos (ciencia de datos) es el pilar de los avances de la actualidad, siendo un privilegio estar en el sitio y el momento adecuados para poder empujar en el sentido correcto, que no es otro que el de aportar conocimiento con la ética bien entendida

La generación de energía no ha sido un gran problema hasta que se percibió como este abuso de los recursos naturales está debilitando el planeta que dejamos a nuestros descendientes por culpa del calentamiento global

El conocimiento adquirido durante el Máster en Ciencia de Datos permitirá ayudar a las compañías eléctricas a determinar con mayor precisión el consumo que será requerido, y por lo tanto no se fabricará en exceso la energía que debilita nuestro planeta.

Capítulo 3

Objetivos del trabajo

El objetivo principal de este TFM es validar la propuesta de cálculo para determinar la previsión de consumo eléctrico más ajustado al consumo real durante el periodo de medición, es decir, gracias a las redes neuronales recurrentes y utilizando los datos históricos de consumo y los datos externos de clima así como los datos de consumo por hogar, podremos determinar con mayor precisión el consumo energético de una región que utilizando otras técnicas de aprendizaje automático.

El objetivo secundario será determinar las variables que influyen en mayor o menor medida en el consumo energético y su determinación para mejorar la precisión del cálculo de previsión de consumo.

Capítulo 4

Método

El método utilizado durante el trabajo ha sido iterativo, con el que se han ido completando las fases de los trabajos planificados aprendiendo e innovando en cada uno de ellos buscando el objetivo final que no es otro que el de validar las hipótesis planteadas en los objetivos del TFM

La hipótesis principal para demostrar el que será el objetivo principal del presente TFM "**Mediante la utilización de las redes neuronales y los datos históricos, podemos determinar con una precisión del 99% el consumo energético de una región**" y como Hipótesis secundaria, "**Mediante la utilización de Redes Neuronales podremos determinar qué tres variables externas influyen más en el cálculo de la predicción del consumo energético de una región**".

En segundo lugar, se alcanzarán objetivos parciales para confirmar las hipótesis determinando los pasos previos y las preguntas de investigación oportunas.

Para alcanzar los objetivos voy a centrar los esfuerzos en tres partes, los **datos**, los **algoritmos** y el **resultado** obtenido.

Capítulo 5

Enfoque

La Estrategia que se aplicará en este análisis se basará en la combinación de la regresión lineal con las redes neuronales para evaluar el resultado comparado con los resultados reales.

Por otro lado, se comparará de forma iterativa los resultados obtenidos al aplicar diversos algoritmos de los más simples a los más complejos y combinados para acercarse a la menor tasa de error posible.

En primer lugar se realizará un análisis para estudiar con detenimiento el contenido de los dataset y decidir qué enfoque utilizar.

En segundo lugar se realizará la fase de diseño, implementación y evaluación, comenzando por la selección del lenguaje a utilizar y realizando la codificación de la solución propuesta.

Una vez diseñada la implementación, se procederá a la obtención de datos de las distintas fuentes de datos para añadirlas como variables que puedan determinar una menor tasa de error en el resultado esperado.

La preparación de los distintos dataset será determinante para la validación de las distintas hipótesis puesto que la granularidad en la que viene el dataset, y la distribución de la estacionalidad, puede modificar el resultado en función del periodo de tiempo que se esté estimando.

Capítulo 6

Tecnología utilizada

Como lenguaje principal con la que se han realizado los cálculos ha sido el Python a través de la herramienta **Google Colaboratory**.

Google Colab es un servicio en la nube, que nos provee de un Jupyter Notebook como editar al que podemos acceder con un navegador web sin importar la ubicación o la arquitectura tecnológica desde la que se acceda.

Tiene como grandes ventajas:

- Posibilidad de activar una GPU.
- Podemos compartir el código fácilmente.
- Está basado en Jupyter Notebook y nos resultará un entorno ya conocido.
- Podemos crear libros en Python 2 ó en 3.
- Tiene preinstaladas las librerías comunes usadas en data science y la posibilidad de instalar otras que necesitemos.
- Al enlazar con nuestra cuenta de Google Drive, podemos leer desde ahí archivos csv de entrada ó guardar imágenes de salida, etc.

Capítulo 7

Planificación del Trabajo

La planificación de trabajo está orientada a cumplir con las fechas establecidas en el calendario de entregas indicado en el portal de la UOC.

En la gráfica de Gantt (Figura 6) se muestran las actividades planificadas desde el comienzo del TFM.

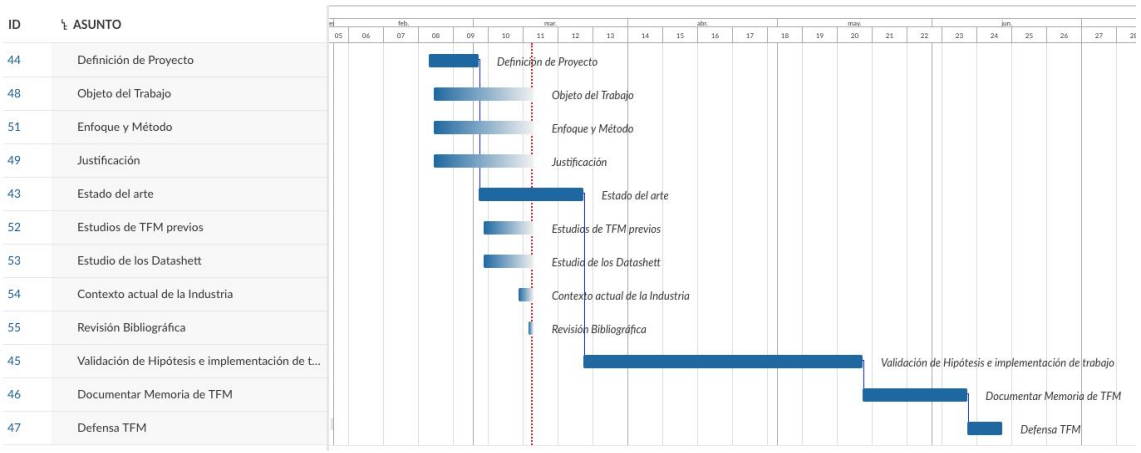

Figura 6. Calendario de entregas del Máster de Ciencia de Datos de la UOC 2019.

Capítulo 8

Estado del arte

El análisis del estado del arte se agrupa en dos tipos, el primero de ellos trata el problema desde un punto de vista estadístico, el segundo de estos puntos de vista, se realiza con algoritmos de tipo Data Science.

Métodos Estadísticos

ARIMA

El modelos paramétrico ARIMA (Autoregressive Integrated Moving Average Model) trata de obtener la representación de la serie de términos de la interrelación temporal de sus elementos (Pilar, Casimiro and Casimiro, no date). Este tipo de modelos que caracterizan las series como sumas o diferencias, ponderadas o no, de variables aleatorias o de las series resultantes, fue propuesto por Yule y Slutzky durante 1927 (Yule, 1927).

Fue la base de los procesos de mediciones móviles y auto regresivos presentados tras la publicación en 1970 del libro de Box-Jenkins sobre modelos ARIMA (Asteriou and Hall, 2011) (Figura 7).

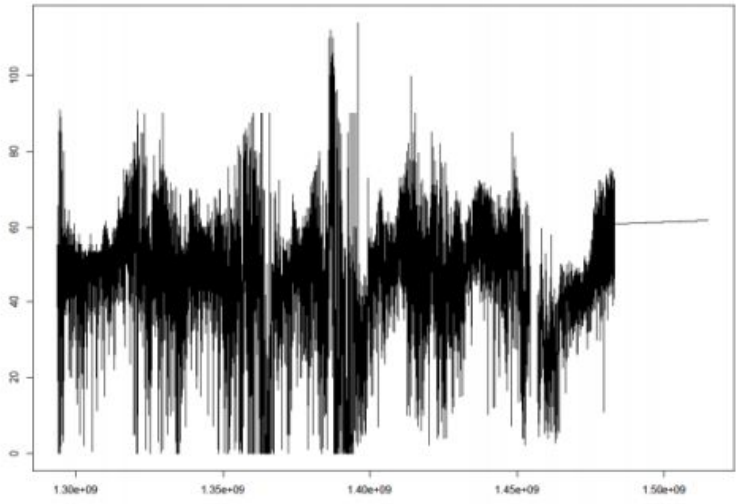

Figura 7. Ejemplo de representación con alto componente estacional.

El instrumento fundamental a la hora de analizar las propiedades de una serie temporal en términos de interrelación temporal de sus observaciones, es decir, el grado de asociación lineal que existe entre observaciones separadas K periodos.

Estos coeficiente de autocorrelación proporcionan mucha información sobre cómo están relacionadas entre sí las distintas observaciones de una serie temporal, lo que ayudará a construir el modelo apropiado para los datos (Peña, 2019).

Este tipo de modelo de regresión estadística, predice una línea con un amplio umbral de valores, lo que impide la precisión necesaria. (Bianco, Manca and Nardini, 2009).

Método Winters

En una serie temporal, pueden aparecer un componente estacionario y una tendencia que explica su comportamiento, esto es debido a la alta estacionalidad, este comportamiento consiste en que cada cierto número de periodos T en la demanda de cada uno de los T periodos, se repite sistemáticamente un aumento o decremento sobre la demanda explicable como estable o con tendencia específica para cada uno de estos periodos.

Lo habitual es que exista una estacionalidad mensual que se repite anualmente (Figura 8).

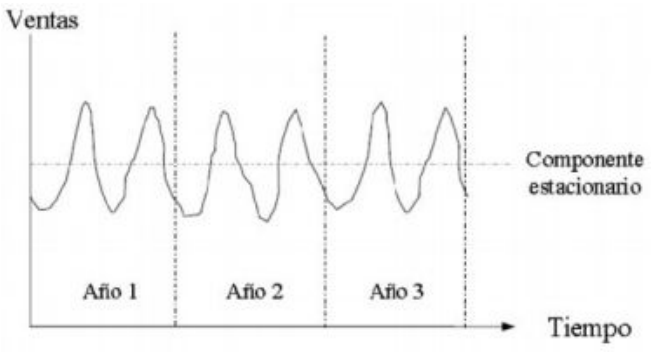

Figura 8. Análisis de ventas con marcado componente estacional.

El método Winters se utiliza para la estimación de series temporales que posean una tendencia lineal y una variación estacional multiplicativa, rectificando los componentes de la serie mediante alisado exponencial (Figura 9).

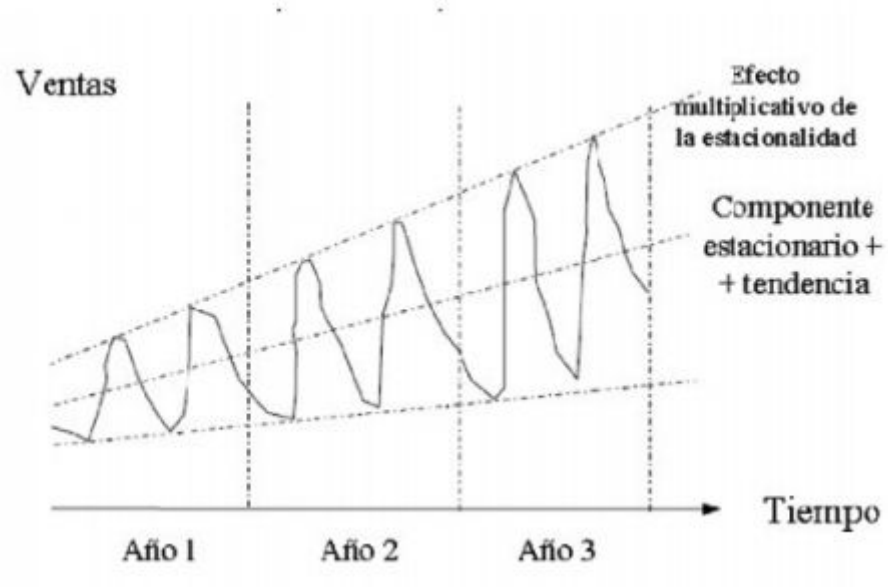

Figura 9. Ejemplo de ventas estacionarias con predicción de tendencia

La principal dificultad que presenta ese tipo de resultados estacionales es que si utilizamos frecuencias de 356 X 24 para calcular series multi-anuales, se pierden el efecto de los cambios de precio por horas, si se utiliza una frecuencia de 24 horas, se pierde el efecto sobre días, semanas y meses.

Método Winters Multi Estacionalidad

En el caso de que la estacionalidad de las series, respondan a múltiples variables temporales, se considera que el patrón periódico es un ciclo. Las series de tipo MSTS (Múltiple Seasonality time Series) con 3 modos (diario, semanal y anual) utilizan un elevado consumo computacional, lo que hace este tipo de métodos poco eficaces.

Métodos Machine Learning

Este método utiliza la Red Neuronal Artificial (RNA), Una Red Neuronal Artificial es "Un grafo dirigido y no lineal con arcos ponderados, capaz de almacenar patrones cambiando los pesos de los arcos, y capaz de recordar patrones a partir de entradas incompletas y desconocidas" (Gale Group., 1928).

La arquitectura de redes neuronales más ampliamente utilizada es la que se conoce como el nombre de Perceptrón Multicapa, la cual se caracteriza por el hecho de que sus neuronas se agrupan en capas por niveles Cada una de estas capas está constituida por un conjunto de neuronas. Hay tres tipos de capas diferentes: la capa de entrada, las capas ocultas y la capa de salida, como se observa en la Figura 10.

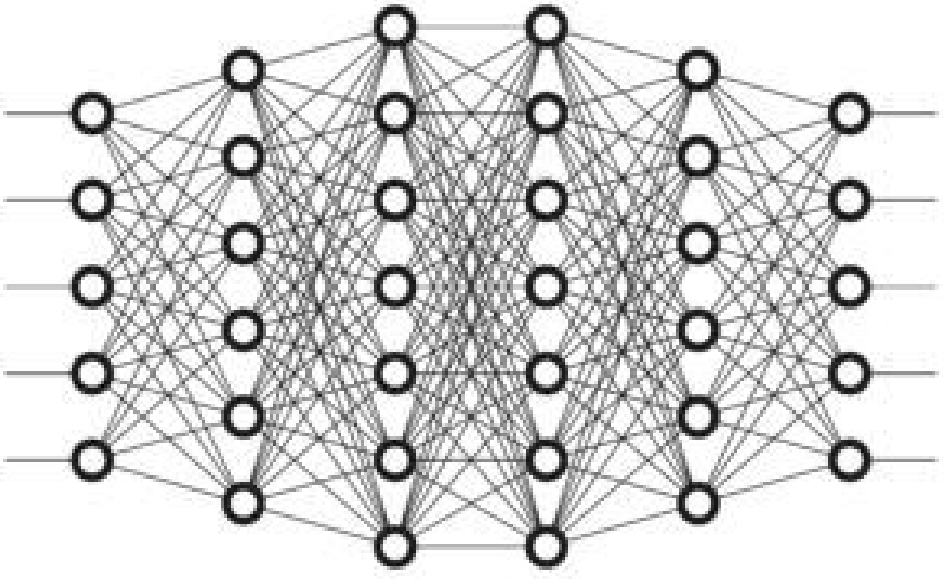

Figura 10, Ejemplo de Red Neuronal

Este tipo de modelos requiere los datos históricos así como otros datos externos como son los datos demográficos y de temperatura, que permite mayor precisión de la predicción (Veit *et al.*, 2014).

Como avance de las Redes Neuronales, está disponible las Redes Neuronales Recurrentes (RNN). Las RNN son un tipo de red neuronal artificial diseñada para reconocer patrones en secuencias de datos, como texto, genomas, escritura a mano, palabra hablada o datos de series de tiempo numéricas que emanan de sensores, mercados bursátiles y agencias gubernamentales. Estos algoritmos toman en cuenta el tiempo y la secuencia, tienen una dimensión temporal.

Las RNN reciben su nombre por la forma en que canalizan la información a través de una serie de operaciones matemáticas realizadas en los nodos de la red. Uno alimenta la información directamente (nunca toca dos veces un nodo determinado), mientras que el otro la recorre a través de un bucle, y este último se llama recurrente.

Las RNN toman como su entrada no solo el ejemplo de entrada actual que ven, sino también lo que han percibido previamente en el tiempo.

Las RNN se distinguen de las redes de avance hacia adelante por ese circuito de retroalimentación conectado a sus decisiones pasadas, ingiriendo sus propias salidas momento tras momento como entrada. Se suele decir que las redes recurrentes tienen memoria así que agregar memoria a las redes neuronales tiene de propósito de utilizar la información en la secuencia misma, ya que las redes de avance no pueden.

La información secuencial se conserva en el estado oculto de la red recurrente, que logra abarcar muchos pasos de tiempo a medida que pasa en cascada para afectar el procesamiento de cada nuevo ejemplo. Está encontrando correlaciones entre eventos separados por muchos momentos, y estas correlaciones se llaman "dependencias a largo plazo", porque un evento en el tiempo depende de uno o más eventos que vinieron antes. Una forma de pensar sobre las RNN es esta: son una forma de compartir ponderaciones a lo largo del tiempo.

$$\mathbf{h}_t = \phi\left(W\mathbf{x}_t + U\mathbf{h}_{t-1}\right)$$

Figura 11. Representación matemática de la memoria hacia delante

A mediados de los años 90, los investigadores alemanes Sepp Hochreiter y Juergen Schmidhuber (Hochreiter and Schmidhuber, 1997), propusieron una variación de las RNN con las llamadas unidades de memoria a corto plazo largo, o LSTM (Long Short-Term Memory), como una solución al problema del gradiente de fuga.

Las LSTM ayudan a preservar el error que se puede propagar a lo largo del tiempo y las capas. Al mantener un error más constante, permiten que las RNN continúen aprendiendo a lo largo de muchos pasos de tiempo (más de 1000), lo que abre un canal para vincular causas y efectos de forma remota.

Este es uno de los desafíos centrales para el aprendizaje automático y la IA (Inteligencia Artificial), ya que los algoritmos se enfrentan con frecuencia a entornos donde las señales de recompensa son escasas y retrasadas, como la vida misma.

Las LSTM contienen información fuera del flujo normal de la RNN en una celda cerrada. La información se puede almacenar, escribir o leer desde una celda, como los datos en la memoria de una computadora. La celda toma decisiones sobre qué almacenar y cuándo permitir lecturas, escrituras y borrados, a través de puertas que se abren y cierran. Sin embargo, a diferencia del almacenamiento digital en las computadoras, estas puertas son analógicas, implementadas con la multiplicación de elementos por sigmoides, que están todos en el rango de 0-1. El análogo tiene la ventaja sobre lo digital de ser diferenciable y, por lo tanto, adecuado para la propagación hacia atrás.

Esas puertas actúan sobre las señales que reciben, y, de forma similar a los nodos de la red neuronal, bloquean o transmiten información en función de su fuerza e importación, que filtran con sus propios conjuntos de pesos. Esos pesos, como los pesos que modulan los estados de entrada y ocultos, se ajustan a través del proceso de aprendizaje de las RNN. Es decir, las celdas aprenden cuándo permitir que los datos ingresen, salgan o se eliminen a través del proceso iterativo de realizar conjeturas, errores de propagación hacia atrás y ajustar los pesos a través del gradiente de pendiente.

Parte II

Investigación

Capítulo 8

Desarrollo de trabajo

Con el propósito de predecir el consumo energético de una población, se ha estudiado los datasets disponibles en la competición Kaggle en la siguiente dirección web:
https://www.kaggle.com/rheajgurung/energy-consumption-forecast/data

- **informations_households.csv**: Archivo que contiene toda la información sobre los hogares en el panel (grupo ACORN, su tarifa) y en qué archivo *block.csv.gz* se almacenan sus datos.
- **daily_dataset.zip**: archivo Zip que contiene los archivos de bloque con la información diaria como el número de medidas, mínimo, máximo, promedio, mediana, suma y estándar.
- **acorn_details.csv**: Los detalles de los grupos ACORN y su perfil de las personas en el grupo provienen de esta hoja de cálculo xlsx. Las tres primeras columnas son los atributos estudiados, el ACORN-X es el índice del atributo. A escala nacional, el índice es 100 si para una columna el valor es 150 significa que hay 1.5 veces más personas con este atributo en el grupo ACORN que en la escala nacional.
- **weather_daily_darksky.csv**: que contiene los datos diarios de la api de *darksky*.
- **weather_hourly_darksky.csv**: que contiene los datos por hora de la api de *darksky*.
- **uk_bank_holidays**: Festividades en UK

Los datos proporcionado han sido almacenados en Google drive para su posterior procesamiento por parte de la herramientas Google Colab.

Con los datos obtenidos, se ha buscado que variables son las más relacionadas con los datos de consumo y se ha creado un dataset para generar un modelo.

El siguiente paso ha sido generar un modelo predictivo basado en Redes neuronales, obteniendo un modelo para comparar la eficiencia con los datos de los consumos reales proporcionados en el dataset.

Como último paso, se ha estudiado la tasa de acierto del modelo obtenido así como mejorar la precisión.

Capítulo 9

Estudio de los datos

Para poder enfocar el problema, es preciso hacer un estudio de los 8 ficheros proporcionados en la URL:
https://www.kaggle.com/rheajgurung/energy-consumption-forecast/data

Fichero: acorn_details.csv

Codificación: CP1252

Tamaño:826 registros X 20 Columnas..

Estadística de los registros del fichero acorn_details.csv (Figura 12).

	ACORN-A	ACORN-B	ACORN-C	ACORN-D	ACORN-E
count	826.000000	826.000000	826.000000	826.000000	826.000000
mean	131.313495	110.860256	100.080789	136.857507	117.894757
std	201.448212	42.464050	30.099529	97.740794	35.768807
min	12.000000	0.957011	0.281968	2.000000	21.000000
25%	87.000000	94.000000	86.000000	93.092150	99.000000
50%	104.000000	107.000000	100.000000	121.000000	117.000000
75%	128.000000	122.000000	113.000000	154.000000	135.000000
max	3795.000000	419.000000	272.000000	1159.034650	286.000000

	ACORN-F	ACORN-G	ACORN-H	ACORN-I	ACORN-J
count	826.000000	826.000000	826.000000	826.000000	826.000000
mean	95.574535	101.444276	97.298915	87.028545	104.216563
std	33.636661	21.798994	18.229234	30.337794	19.924033
min	0.000000	0.791419	1.155448	6.363259	16.050708
25%	81.000000	94.138076	91.000000	70.000000	97.000000
50%	98.000000	102.000000	99.000000	88.000000	105.000000
75%	108.000000	109.000000	105.000000	101.750000	115.000000
max	462.000000	295.000000	192.000000	410.000000	197.000000

	ACORN-K	ACORN-L	ACORN-M	ACORN-N	ACORN-O
count	826.000000	826.000000	826.000000	826.000000	826.000000
mean	127.482911	93.724209	91.410277	79.912379	95.579335
std	97.428159	22.177041	22.909602	33.995192	25.935770
min	17.000000	0.393546	0.714857	2.000000	11.000000
25%	85.000000	86.000000	82.000000	60.253502	86.000000
50%	109.000000	95.000000	93.000000	74.000000	96.000000
75%	144.000000	102.000000	101.000000	93.158386	104.000000
max	1821.000000	280.000000	161.000000	295.000000	252.000000

	ACORN-P	ACORN-Q
count	826.000000	826.000000
mean	100.141309	90.855423
std	37.210288	37.634017
min	9.000000	1.000000
25%	82.250000	71.250000
50%	96.000000	87.000000
75%	109.000000	101.000000
max	389.000000	326.000000

figura 12. Estadística del dataset obtenido de acorn_details.csv

Detalle de los primeros registros del fichero acorn_details.csv (Figura 13).

	MAIN CATEGORIES	CATEGORIES	REFERENCE	ACORN-A	ACORN-B	ACORN-C	ACORN-D	ACORN-E
0	POPULATION	Age	Age 0-4	77.0	83.0	72.0	100.0	120.0
1	POPULATION	Age	Age 5-17	117.0	109.0	87.0	69.0	94.0
2	POPULATION	Age	Age 18-24	64.0	73.0	67.0	107.0	100.0
3	POPULATION	Age	Age 25-34	52.0	63.0	62.0	197.0	151.0
4	POPULATION	Age	Age 35-49	102.0	105.0	91.0	124.0	118.0
5	POPULATION	Age	Age 50-64	124.0	121.0	120.0	72.0	82.0
6	POPULATION	Age	Aged 65-74	125.0	120.0	152.0	55.0	61.0
7	POPULATION	Age	Aged 75 plus	112.0	103.0	157.0	49.0	57.0
8	POPULATION	Geography	England	107.0	101.0	103.0	114.0	106.0
9	POPULATION	Geography	Northern Ireland	30.0	95.0	45.0	2.0	49.0

Figura 13. Detalles de los 10 primeros registros del fichero acorn_details.csv

Observamos que no existen registros nulos al contar todos los registros de cada columna (Figura 14).

```
MAIN CATEGORIES    826
CATEGORIES         826
REFERENCE          826
ACORN-A            826
ACORN-B            826
ACORN-C            826
ACORN-D            826
ACORN-E            826
ACORN-F            826
ACORN-G            826
ACORN-H            826
ACORN-I            826
ACORN-J            826
ACORN-K            826
ACORN-L            826
ACORN-M            826
ACORN-N            826
ACORN-O            826
ACORN-P            826
ACORN-Q            826
```

Figura 14, Número de ítems en cada columna del fichero acorn_details.csv

Fichero: informations_households.csv

Codificación: cp1252

Tamaño:5566 registros X 5 Columnas.

Estadística de los registros del fichero informations_households.csv (Figura 15).

	LCLid	stdorToU	Acorn	Acorn_grouped	file
count	5566	5566	5566	5566	5566
unique	5566	2	19	5	112
top	MAC004415	Std	ACORN-E	Affluent	block_42
freq	1	4443	1567	2192	50

figura 15. Estadística del dataset obtenido de informations_households.csv

Detalle de los primeros registros del fichero acorn_details.csv (Figura 16).

	LCLid	stdorToU	Acorn	Acorn_grouped	file
0	MAC005492	ToU	ACORN-	ACORN-	block_0
1	MAC001074	ToU	ACORN-	ACORN-	block_0
2	MAC000002	Std	ACORN-A	Affluent	block_0
3	MAC003613	Std	ACORN-A	Affluent	block_0
4	MAC003597	Std	ACORN-A	Affluent	block_0
5	MAC003579	Std	ACORN-A	Affluent	block_0
6	MAC003566	Std	ACORN-A	Affluent	block_0
7	MAC003557	Std	ACORN-A	Affluent	block_0
8	MAC003553	Std	ACORN-A	Affluent	block_0
9	MAC003482	Std	ACORN-A	Affluent	block_0

Figura 16. Detalle de los primeros 10 registros del fichero informations_households

Observamos que no existen registros nulos al contar todos los registros de cada columna (Figura 17).

LCLid	5566
stdorToU	5566
Acorn	5566
Acorn_grouped	5566
file	5566

Figura 17, Número de ítems en cada columna del fichero informations_households.csv

Fichero: uk_bank_holidays.csv

Codificación: cp1252

Tamaño:25 registros X 2 Columnas

Estadística de los registros del fichero uk_bank_holidays.csv (Figura 18).

	Bank holidays	Type
count	25	25
unique	25	11
top	2013-01-04	Christmas Day
freq	1	3

figura 18. Estadística del dataset obtenido de uk_bank_holidays.csv

Detalle de los primeros registros del fichero uk_bank_holidays.csv (Figura 19).

	Bank holidays	Type
0	2012-12-26	Boxing Day
1	2012-12-25	Christmas Day
2	2012-08-27	Summer bank holiday
3	2012-05-06	Queen?s Diamond Jubilee (extra bank holiday)
4	2012-04-06	Spring bank holiday (substitute day)
5	2012-07-05	Early May bank holiday
6	2012-09-04	Easter Monday
7	2012-06-04	Good Friday
8	2012-02-01	New Year?s Day (substitute day)
9	2013-12-26	Boxing Day

Figura 19. Detalle de los primeros 10 registros del fichero uk_bank_holidays

Observamos que no existen registros nulos al contar todos los registros de cada columna (Figura 20).

Bank holidays	25
Type	25

Figura 20, Número de ítems en cada columna del fichero uk_bank_holidays.csv

Fichero: weather_daily_darksky.csv

Codificación: cp1252

Tamaño:882 registros X 32 Columnas

Estadística de los registros del fichero weather_daily_darksky.csv (Figura 21).

	temperatureMax	windBearing	dewPoint	cloudCover	windSpeed	\
count	882.000000	882.000000	882.000000	881.000000	882.000000	
mean	13.660113	195.702948	6.530034	0.477605	3.581803	
std	6.182744	89.340783	4.830875	0.193514	1.694007	
min	-0.060000	0.000000	-7.840000	0.000000	0.200000	
25%	9.502500	120.500000	3.180000	0.350000	2.370000	
50%	12.625000	219.000000	6.380000	0.470000	3.440000	
75%	17.920000	255.000000	10.057500	0.600000	4.577500	
max	32.400000	359.000000	17.770000	1.000000	9.960000	

	pressure	apparentTemperatureHigh	visibility	humidity	\
count	882.000000	882.000000	882.000000	882.000000	
mean	1014.127540	12.723866	11.167143	0.781871	
std	11.073038	7.279168	2.466109	0.095348	
min	979.250000	-6.460000	1.480000	0.430000	
25%	1007.435000	7.032500	10.327500	0.720000	
50%	1014.615000	12.470000	11.970000	0.790000	
75%	1021.755000	17.910000	12.830000	0.860000	
max	1040.920000	32.420000	15.340000	0.980000	

	apparentTemperatureLow	apparentTemperatureMax	uvIndex	\
count	882.000000	882.000000	881.000000	
mean	6.085045	12.929467	2.542565	
std	6.031967	7.105426	1.832985	
min	-8.880000	-4.110000	0.000000	
25%	1.522500	7.332500	1.000000	
50%	5.315000	12.625000	2.000000	
75%	11.467500	17.920000	4.000000	
max	20.540000	32.420000	7.000000	

	temperatureLow	temperatureMin	temperatureHigh	\
count	882.000000	882.000000	882.000000	
mean	7.709841	7.414161	13.542392	
std	4.871004	4.888852	6.260196	
min	-5.640000	-5.640000	-0.810000	
25%	3.990000	3.705000	9.212500	
50%	7.540000	7.100000	12.470000	
75%	11.467500	11.277500	17.910000	
max	20.540000	20.540000	32.400000	

figura 21. Estadística del dataset obtenido de weather_daily_darksky.csv

Detalle de los primeros registros del fichero weather_daily_darksky.csv (Figura 22).

	temperatureMax	temperatureMaxTime	windBearing	icon	\
0	11.96	2011-11-11 23:00:00	123	fog	
1	8.59	2011-12-11 14:00:00	198	partly-cloudy-day	
2	10.33	2011-12-27 02:00:00	225	partly-cloudy-day	
3	8.07	2011-12-02 23:00:00	232	wind	
4	8.22	2011-12-24 23:00:00	252	partly-cloudy-night	
5	7.97	2011-12-15 14:00:00	234	wind	
6	13.19	2011-11-19 14:00:00	117	fog	
7	8.32	2011-11-16 23:00:00	117	fog	
8	9.82	2011-12-12 23:00:00	221	wind	
9	9.71	2011-11-20 14:00:00	115	fog	

	dewPoint	temperatureMinTime	cloudCover	windSpeed	pressure	\
0	9.40	2011-11-11 07:00:00	0.79	3.88	1016.08	
1	4.49	2011-12-11 01:00:00	0.56	3.94	1007.71	
2	5.47	2011-12-27 23:00:00	0.85	3.54	1032.76	
3	3.69	2011-12-02 07:00:00	0.32	3.00	1012.12	
4	2.79	2011-12-24 07:00:00	0.37	4.46	1028.17	
5	2.41	2011-12-15 00:00:00	0.42	4.71	996.75	
6	8.12	2011-11-19 23:00:00	0.26	2.37	1016.80	
7	5.58	2011-11-16 07:00:00	0.81	2.36	1017.40	
8	4.10	2011-12-12 07:00:00	0.38	5.02	1002.47	
9	6.62	2011-11-20 08:00:00	0.41	1.24	1018.82	

	apparentTemperatureMinTime	...	temperatureHigh	sunriseTime	\
0	2011-11-11 07:00:00	...	10.87	2011-11-11 07:12:14	
1	2011-12-11 02:00:00	...	8.59	2011-12-11 07:57:02	
2	2011-12-27 22:00:00	...	10.33	2011-12-27 08:07:06	
3	2011-12-02 07:00:00	...	7.36	2011-12-02 07:46:09	
4	2011-12-24 07:00:00	...	7.93	2011-12-24 08:06:15	
5	2011-12-15 00:00:00	...	7.97	2011-12-15 08:00:46	
6	2011-11-19 08:00:00	...	13.19	2011-11-19 07:26:03	
7	2011-11-16 04:00:00	...	8.18	2011-11-16 07:20:57	
8	2011-12-12 08:00:00	...	8.53	2011-12-12 07:58:02	
9	2011-11-20 08:00:00	...	9.71	2011-11-20 07:27:43	

	temperatureHighTime	uvIndexTime	\
0	2011-11-11 19:00:00	2011-11-11 11:00:00	
1	2011-12-11 14:00:00	2011-12-11 12:00:00	
2	2011-12-27 14:00:00	2011-12-27 00:00:00	
3	2011-12-02 12:00:00	2011-12-02 10:00:00	
4	2011-12-24 15:00:00	2011-12-24 13:00:00	
5	2011-12-15 14:00:00	2011-12-15 11:00:00	
6	2011-11-19 14:00:00	2011-11-19 10:00:00	
7	2011-11-16 14:00:00	2011-11-16 11:00:00	
8	2011-12-12 19:00:00	2011-12-12 11:00:00	
9	2011-11-20 14:00:00	2011-11-20 10:00:00	

Figura 22. Detalle de los primeros 10 registros del fichero weather_daily_darksky

Observamos que si existen registros nulos al contar todos los registros de cada columna (Figura 23).

```
[10 rows x 32 columns]
temperatureMax                882
temperatureMaxTime            882
windBearing                   882
icon                          882
dewPoint                      882
temperatureMinTime            882
cloudCover                    881
windSpeed                     882
pressure                      882
apparentTemperatureMinTime    882
apparentTemperatureHigh       882
precipType                    882
visibility                    882
humidity                      882
apparentTemperatureHighTime   882
apparentTemperatureLow        882
apparentTemperatureMax        882
uvIndex                       881
time                          882
sunsetTime                    882
temperatureLow                882
temperatureMin                882
temperatureHigh               882
sunriseTime                   882
temperatureHighTime           882
uvIndexTime                   881
summary                       882
temperatureLowTime            882
apparentTemperatureMin        882
apparentTemperatureMaxTime    882
apparentTemperatureLowTime    882
moonPhase                     882
```

Figura 23, Número de ítems en cada columna del fichero weather_daily_darksky.csv

Fichero: weather_hourly_darksky.csv

Codificación: cp1252

Tamaño:21165 registros X 12 Columnas

Estadística de los registros del fichero weather_hourly_darksky.csv (Figura 24).

	visibility	windBearing	temperature	dewPoint	pressure	\
count	21165.000000	21165.000000	21165.000000	21165.000000	21152.000000	
mean	11.166485	195.685897	10.471486	6.530501	1014.125153	
std	3.099337	90.629453	5.781904	5.041965	11.388337	
min	0.180000	0.000000	-5.640000	-9.980000	975.740000	
25%	10.120000	121.000000	6.470000	2.820000	1007.430000	
50%	12.260000	217.000000	9.930000	6.570000	1014.780000	
75%	13.080000	256.000000	14.310000	10.330000	1022.050000	
max	16.090000	359.000000	32.400000	19.880000	1043.320000	

	apparentTemperature	windSpeed	humidity
count	21165.000000	21165.000000	21165.000000
mean	9.230338	3.905215	0.781829
std	6.940919	2.026854	0.140369
min	-8.880000	0.040000	0.230000
25%	3.900000	2.420000	0.700000
50%	9.360000	3.680000	0.810000
75%	14.320000	5.070000	0.890000
max	32.420000	14.800000	1.000000

figura 24. Estadística del dataset obtenido de weather_hourly_darksky.csv

Detalle de los primeros registros del fichero weather_hourly_darksky.csv (Figura 25).

	visibility	windBearing	temperature	time	dewPoint	\
0	5.97	104	10.24	2011-11-11 00:00:00	8.86	
1	4.88	99	9.76	2011-11-11 01:00:00	8.83	
2	3.70	98	9.46	2011-11-11 02:00:00	8.79	
3	3.12	99	9.23	2011-11-11 03:00:00	8.63	
4	1.85	111	9.26	2011-11-11 04:00:00	9.21	
5	1.96	115	9.33	2011-11-11 05:00:00	8.87	
6	1.30	118	9.31	2011-11-11 06:00:00	8.82	
7	1.22	114	8.85	2011-11-11 07:00:00	8.69	
8	1.40	120	9.13	2011-11-11 08:00:00	8.75	
9	1.38	121	9.23	2011-11-11 09:00:00	8.70	

	pressure	apparentTemperature	windSpeed	precipType	icon	\
0	1016.76	10.24	2.77	rain	partly-cloudy-night	
1	1016.63	8.24	2.95	rain	partly-cloudy-night	
2	1016.36	7.76	3.17	rain	partly-cloudy-night	
3	1016.28	7.44	3.25	rain	fog	
4	1015.98	7.24	3.70	rain	fog	
5	1015.91	7.19	3.97	rain	fog	
6	1015.70	7.10	4.10	rain	fog	
7	1016.08	6.48	4.23	rain	fog	
8	1016.33	6.84	4.20	rain	fog	
9	1016.57	7.07	3.96	rain	fog	

	humidity	summary
0	0.91	Partly Cloudy
1	0.94	Partly Cloudy
2	0.96	Partly Cloudy
3	0.96	Foggy
4	1.00	Foggy
5	0.97	Foggy
6	0.97	Foggy
7	0.99	Foggy
8	0.97	Foggy
9	0.97	Foggy

Figura 25. Detalle de los primeros 10 registros del fichero weather_hourly_darksky

Observamos que no existen registros nulos al contar todos los registros de cada columna (Figura 26).

visibility	21165
windBearing	21165
temperature	21165
time	21165
dewPoint	21165
pressure	21152
apparentTemperature	21165
windSpeed	21165
precipType	21165
icon	21165
humidity	21165
summary	21165

Figura 26, Número de ítems en cada columna del fichero weather_hourly_darksky.csv

Fichero: daily_dataset.zip

Codificación: cp1252

Tamaño:3536007 registros X 4 Columnas

Estadística de los registros del fichero daily_dataset.zip convertido en fichero tras descomprimir los 111 ficheros que lo integran en el fichero energy.csv (Figura 27).

	Unnamed: 0	energy_sum
count	3.536007e+06	3.535977e+06
mean	1.576990e+04	1.020674e+01
std	9.194901e+03	9.307041e+00
min	0.000000e+00	0.000000e+00
25%	7.823000e+03	4.698000e+00
50%	1.569700e+04	7.852000e+00
75%	2.359000e+04	1.264100e+01
max	3.616700e+04	3.325560e+02

figura 27. Estadística del dataset obtenido de energy.csv

Detalle de los primeros registros del fichero energy.csv (Figura 28)

	Unnamed: 0	day	LCLid	energy_sum
0	0	2012-10-12	MAC000002	7.098
1	1	2012-10-13	MAC000002	11.087
2	2	2012-10-14	MAC000002	13.223
3	3	2012-10-15	MAC000002	10.257
4	4	2012-10-16	MAC000002	9.769
5	5	2012-10-17	MAC000002	10.885
6	6	2012-10-18	MAC000002	10.751
7	7	2012-10-19	MAC000002	8.431
8	8	2012-10-20	MAC000002	17.378
9	9	2012-10-21	MAC000002	24.490

Figura 28. Detalle de los primeros 10 registros del fichero energy

Observamos que no existen registros nulos al contar todos los registros de cada columna (Figura 29).

```
Unnamed: 0     3536007
day            3536007
LCLid          3536007
energy_sum     3535977
```

Figura 29. Número de ítems en cada columna del fichero energy.csv

Adaptación de los datos

Se modifica el dataset ***energy*** ordenando el contenido por la columna que indica las fechas, esta labor se realiza para determinar el número de hogares que generaron los datos, puesto que la implementación del sistema de medida fue paulatino y podría arrojar datos confusos de consumos en los primero días de la instalación (Figura 30).

day	
2011-11-23	13
2011-11-24	25
2011-11-25	32
2011-11-26	41

Figura 30. Primeros datos del dataset ***energy***

Comprobamos la inconsistencia de los datos por hogar, por lo tanto no pueden utilizarse este dato para realizar la predicción, así que se utilizará la suma de la energía global para calcular la predicción y de esta forma normalizamos los datos de consumo energético. (Figura 31).

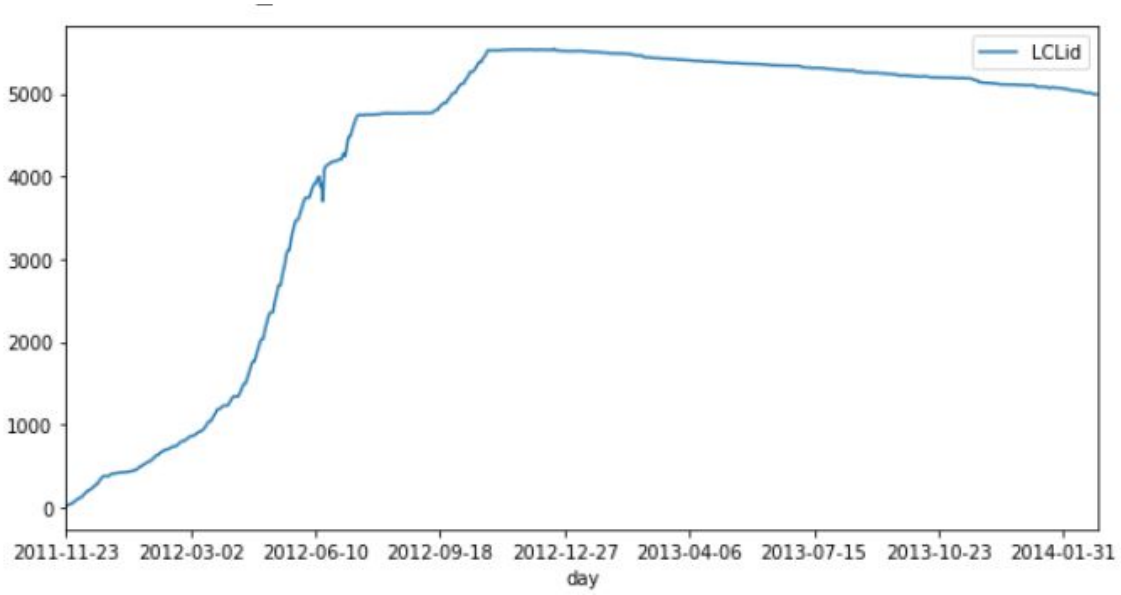

Figura 31. Inconsistencia de los datos consumidos por los hogares.

Datos del clima

Tras cargar los datos del clima denominado ***weather_energy,*** comenzamos a comparar gráficamente algunas de las variables con el consumo energético del dataset ***energy.***

Temperatura

Los picos en los valores aparecen coincidentes con los canales en el otro. Esto confirma la intuición de las compañías eléctricas de que con temperaturas bajas, es probable que aumente el consumo de energía a través de aparatos calentadores.

La gráfica muestra la temperatura máxima y mínima con los colores naranja y rosa respectivamente, junto con el consume eléctrico en color azul (Figura 32).

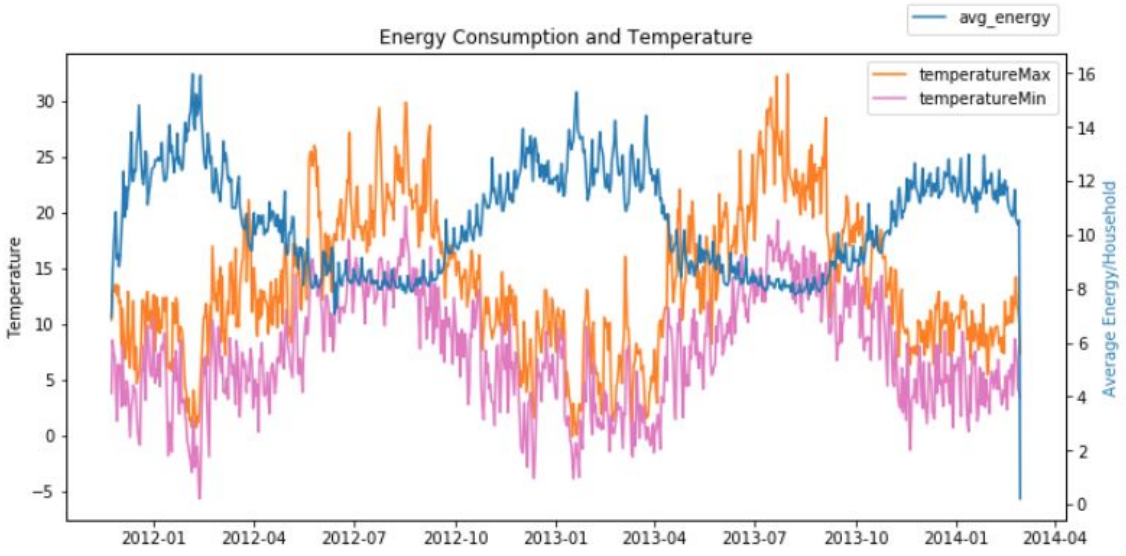

Figura 32. Comparativa entre Energía y Temperatura.

Humedad

La humedad y el consumo medio de energía parecen tener la misma tendencia de correlación positiva que la *temperatura* (Figura 33).

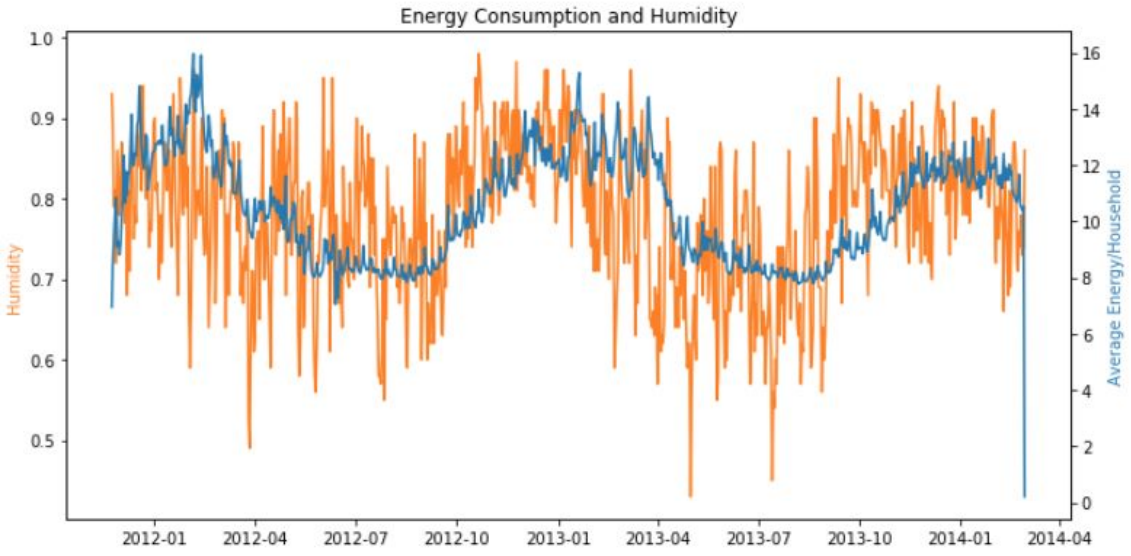

Figura 33. Comparativa entre Energía y Humedad.

Cielos cubiertos

El valor de la cobertura de las nubes parece seguir el mismo patrón que el consumo de energía; mostrando multicolinealidad con la humedad (Figura 34).

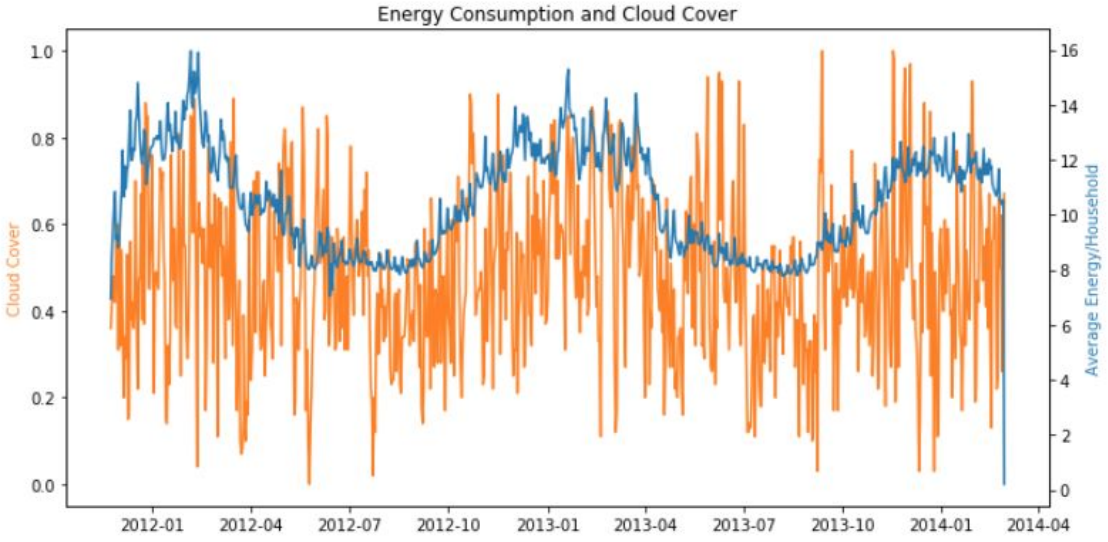

Figura 34. Comparativa entre Energía y Cielos cubiertos.

Visibilidad

El factor de visibilidad no parece afectar el consumo de energía en absoluto, ya que la visibilidad es probablemente un factor en el exterior, es poco probable que su aumento o disminución afecte el consumo de energía dentro de un hogar, mostrando multicolinealidad con la humedad (Figura 35).

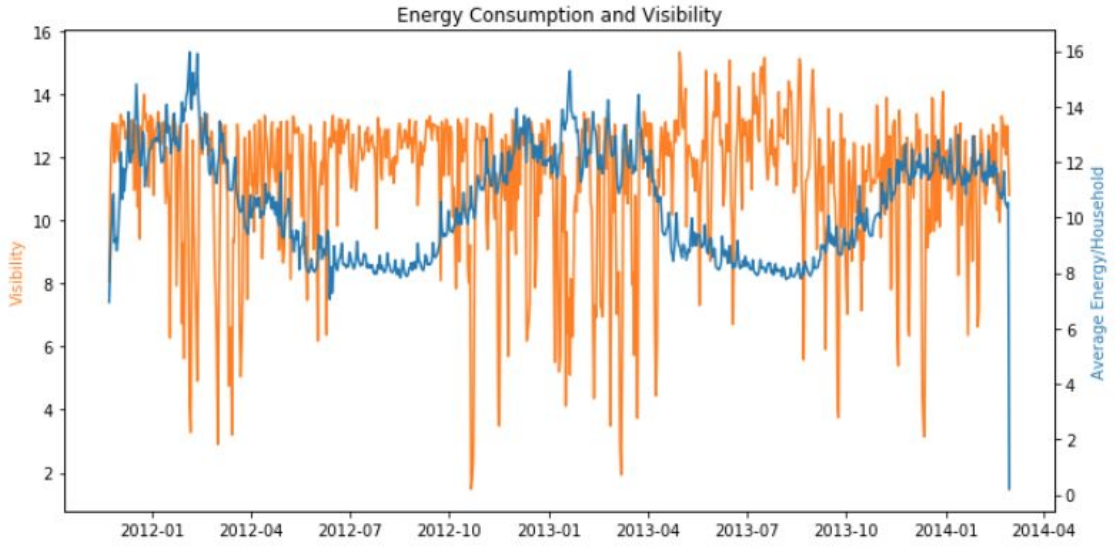

Figura 35. Comparativa entre Energía y Visibilidad.

Velocidad del viento

Al igual que la visibilidad, la velocidad del viento parece ser un factor al aire libre que no afecta el consumo de energía como tal no mostrando correlación y sin mostrar multicolinealidad con otros factores (Figura 36).

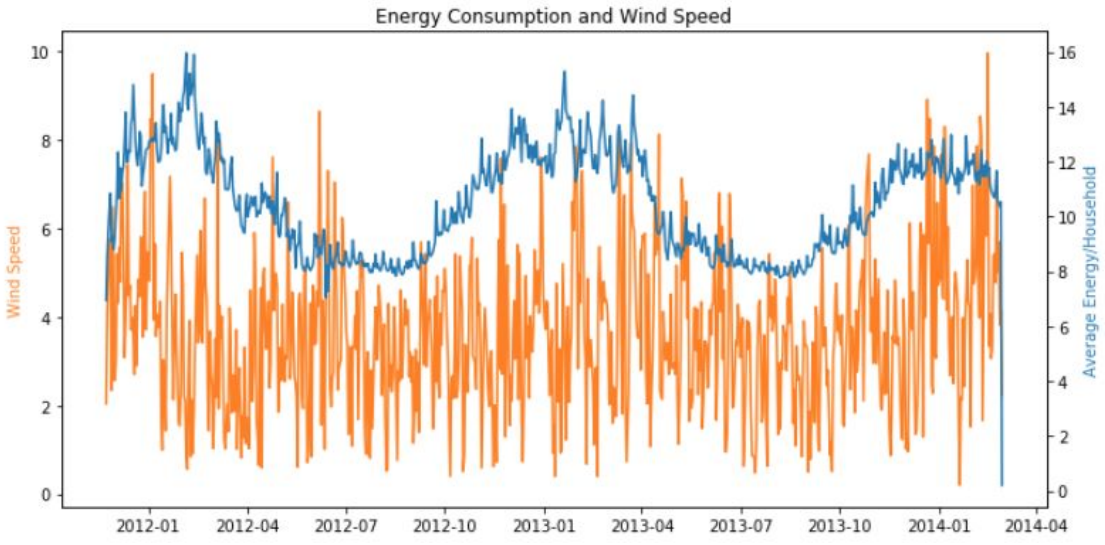

Figura 36. Comparativa entre Energía y Velocidad del viento.

Índice UV

El índice UV tiene una relación inversa con el consumo de energía, teniendo multicolinealidad con la temperatura como en el caso de la variable Rocio(Figura 37).

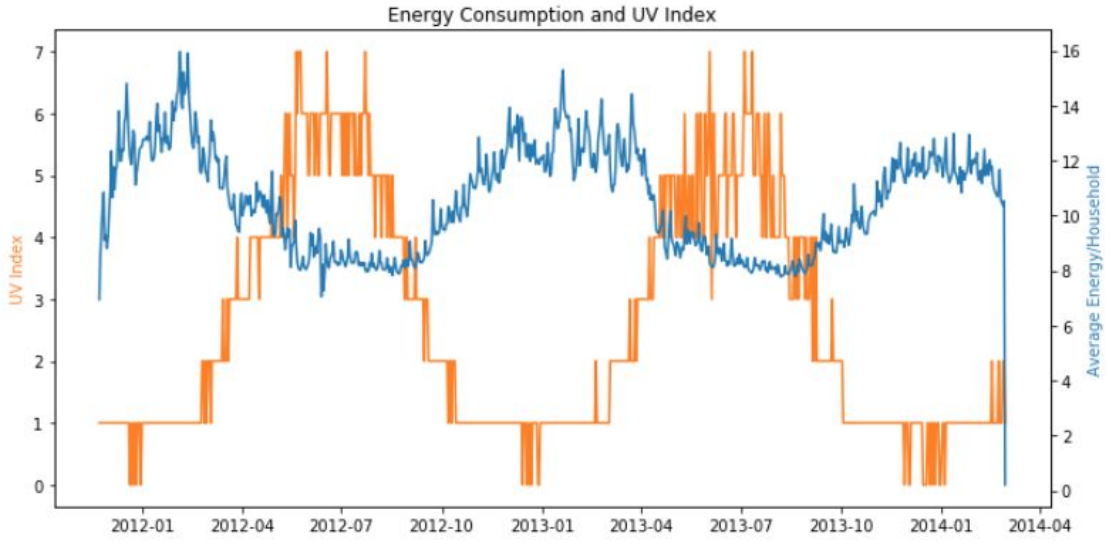

Figura 37. Comparativa entre Energía e Índice UV.

Rocío

El punto de rocío es una función de la humedad y la temperatura, por lo que muestra una relación similar con el consumo de energía, considerando multicolinealidad con la temperatura como en el caso del **Índice UV** (Figura 38).

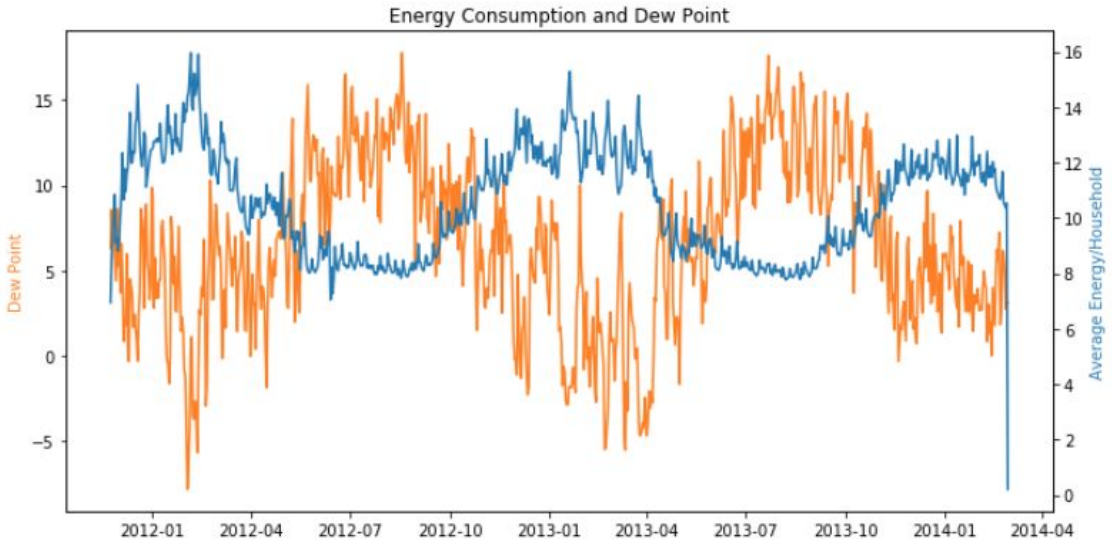

Figura 38. Comparativa entre Energía y Rocio.

Fase Lunar

La fase lunar no parece que tenga repercusión el consumo energético. (Figura 39).

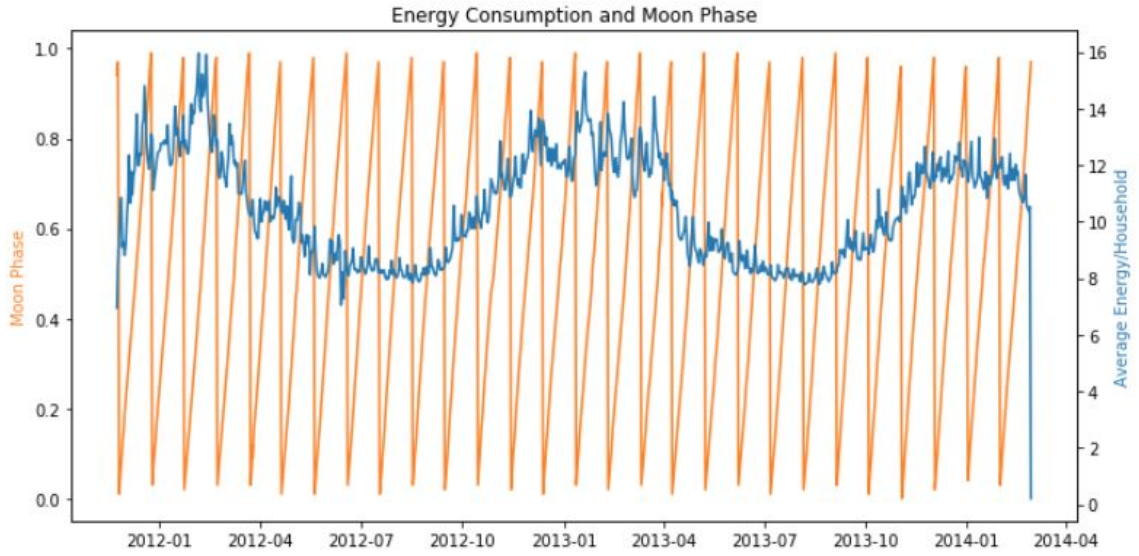

Figura 39. Comparativa entre Energía y Fase Lunar.

Presión atmosférica

La presión atmosférica no parece que tenga repercusión el consumo energético (Figura 40).

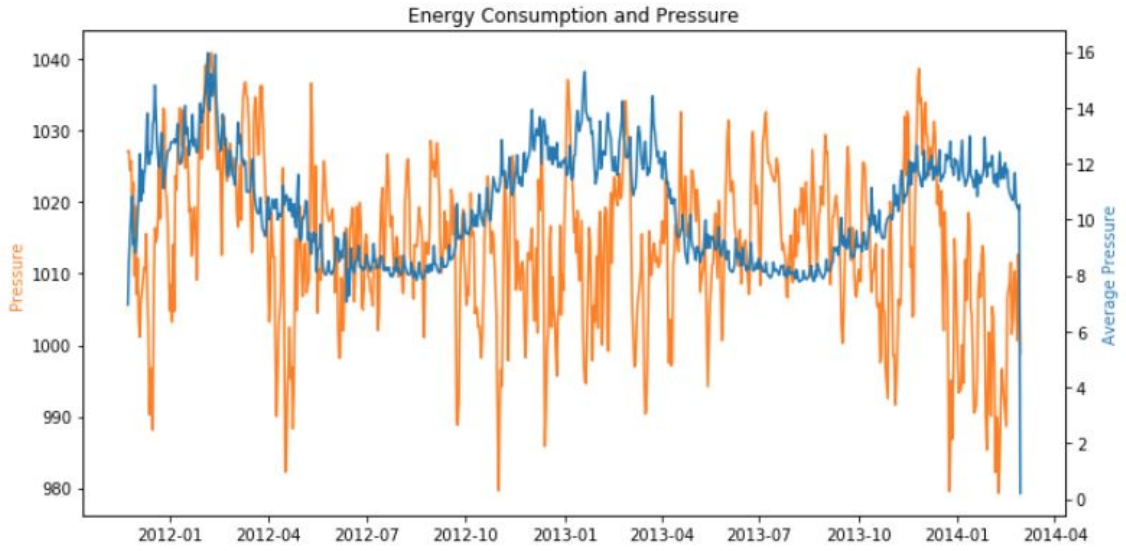

Figura 40. Comparativa entre Energía y Presión Atmosférica.

Conclusión de las variables

La correlación entre las variables son discretizadas para obtener las más significativas (> +0.7 y < -0.7) (Figura 40):

- La energía tiene una alta correlación positiva con la humedad y una alta correlación negativa con la temperatura.
- El punto de rocío y el índice UV, muestra una multicolinealidad con la temperatura, por lo que se descartan.
- La cobertura de nubes y la visibilidad muestran una multicolinealidad con humedad, por lo que se descartan.
- La presión y la fase lunar tienen una correlación mínima con la energía, por lo que se descartan.
- La velocidad del viento tiene una baja correlación con la energía, pero no muestra multicolinealidad.

	avg_energy	temperatureMax	dewPoint	cloudCover	windSpeed	pressure	visibility	humidity	uvIndex	moonPhase
avg_energy	1.000000	-0.846965	-0.755901	0.241779	0.149624	-0.028851	-0.246404	0.361237	-0.733171	-0.031716
temperatureMax	-0.846965	1.000000	0.865038	-0.333409	-0.153602	0.118933	0.259108	-0.404899	0.696497	0.003636
dewPoint	-0.755901	0.865038	1.000000	-0.025207	-0.092212	-0.028121	0.042633	0.055514	0.486692	-0.008239
cloudCover	0.241779	-0.333409	-0.025207	1.000000	0.170235	-0.101079	-0.330177	0.480056	-0.248695	-0.062126
windSpeed	0.149624	-0.153602	-0.092212	0.170235	1.000000	-0.344354	0.281088	-0.042391	-0.152634	-0.023273
pressure	-0.028851	0.118933	-0.028121	-0.101079	-0.344354	1.000000	-0.012508	-0.250941	0.100774	0.038462
visibility	-0.246404	0.259108	0.042633	-0.330177	0.281088	-0.012508	1.000000	-0.578130	0.240485	0.062813
humidity	0.361237	-0.404899	0.055514	0.480056	-0.042391	-0.250941	-0.578130	1.000000	-0.533919	-0.013997
uvIndex	-0.733171	0.696497	0.486692	-0.248695	-0.152634	0.100774	0.240485	-0.533919	1.000000	0.012833
moonPhase	-0.031716	0.003636	-0.008239	-0.062126	-0.023273	0.038462	0.062813	-0.013997	0.012833	1.000000

Figura 41. Matriz de correlación.

Comprobamos que la información climática dispone de múltiples variables, sin ser útiles algunas de ellas, comprobamos cuántos pueden ser los grupos en los que pueden ser divididos los valores agrupados. Utilizando la curva del codo o Elbow Curve, donde comprobamos que pueden ser 3 grupos o clusters (Figura 42).

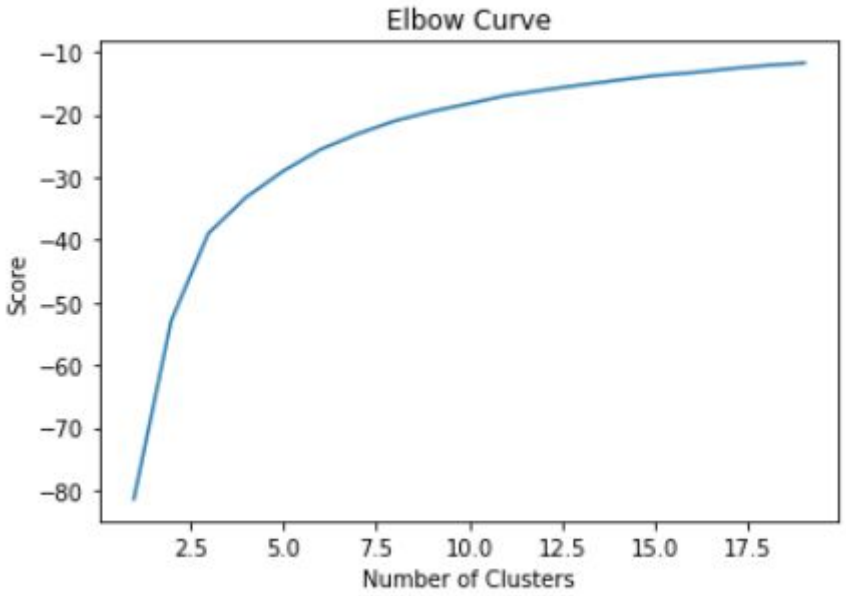

Figura 42. Número de valores obtenidos a la hora de discretizar valores del clima.

Las variables **temperatureMax**, **humidity** y **windSpeed** son discretizadas como las referidas al clima que pueden agruparse en una sola entre los valores 0 y 2, comprobando que realmente todos los grupos obtenidos están discretizados para el clima que hubo para cada día (Figura 43) y son añadidas al dataset ***weather_energy*** donde se agrupan los valores de clima y los de energía consumida en el periodo de tiempo disponible en el dataset.

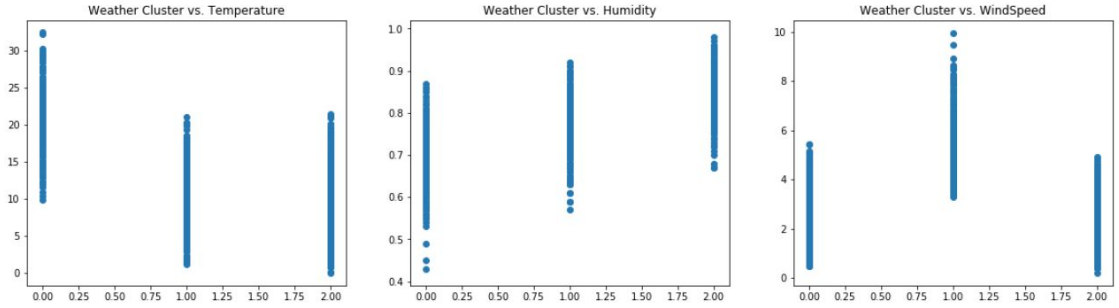

Figura 43. Grafica de relacion entre variables.

Cargamos los días Festivos para determinar si son valores que influyen a la hora de determinar la predicción del consumo de energía y lo añadimos al dataset ***weather_energy*** (Figura 44).

	Bank holidays	Type
0	2012-12-26	Boxing Day
1	2012-12-25	Christmas Day
2	2012-08-27	Summer bank holiday
3	2012-05-06	Queen?s Diamond Jubilee (extra bank holiday)

Figura 44. Detalle de Dataset uk_bank_holidays

Capítulo 10

Datos de entrenamiento

Obtenemos unos datos de entrenamiento y otros con los que validamos el modelo, el entrenamiento son toda la muestra menos los últimos 30 días, y el test de validación (dataset *training*) los últimos 30 días (Figura 45).

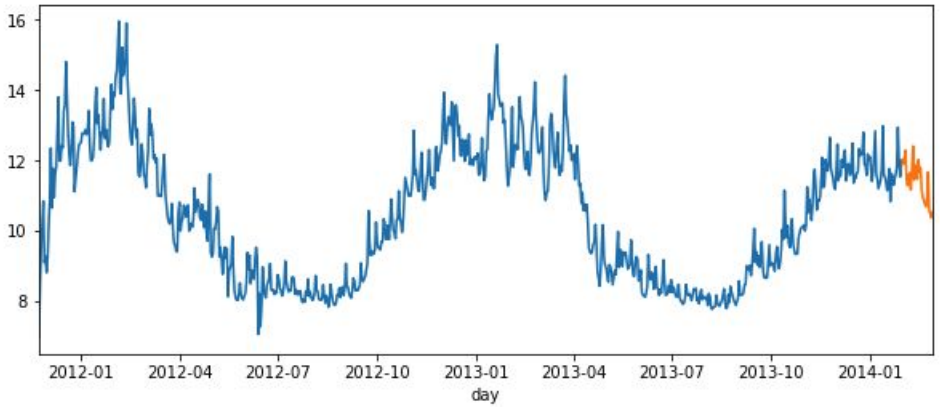

Figura 45, separación datos de test y entrenamiento.

Vemos las primeras filas del dataset *training* para comprobar que es correcto (Figura 46).

day	avg_energy	weather_cluster	holiday_ind
2014-01-30	11.886982	2	0
2014-01-31	12.051321	2	0
2014-02-01	11.921217	0	0
2014-02-02	12.291726	0	0
2014-02-03	11.471760	0	0

Figura 46. Primeros datos de dataset *test.*

El Dataset final tiene las siguientes columnas:

day: Día del registro de los datos (Año-mes-día)
avg_energy: Energía acumulada consumida por todos los dispositivos
weather_cluster: Cluster asignado de las variables climáticas (entre 1 y 3)
holiday_ind: Si fué un día festivo (Valor 1) o no (Valor 0)

Algoritmo ARIMA

Para encontrar patrones en variaciones y regresiones estadísticas, se utiliza el modelo *Media móvil integrada autorregresiva* o ARIMA (acrónimo del inglés autoregressive integrated moving average). Se utiliza para encontrar patrones que puedan determinar una predicción o estimación futura basados en datos pasados.

El modelo ARIMA se representa con la siguiente ecuación donde *d* corresponde a las d diferencias que son necesarias para convertir la serie original en estacionaria, $\phi_1, \ldots, \phi_p$ son los parámetros pertenecientes a la parte *autorregresiva* del modelo, $\theta_1, \ldots, \theta_q$ los parámetros pertenecientes a la parte *medias móviles* del modelo, ϕ_0 es una constante, y ε_t es el término de error (llamado también *innovación* o *perturbación estocástica* esta última asociada más para modelos econométricos uniecuacionales o multiecuacionales) (Geurts, Box and Jenkins, 2006).

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Aunque el método puede manejar datos con una tendencia, no admite series de tiempo con un componente estacional por lo que es necesario utilizar una extensión que admite el modelado directo del componente estacional de la serie se llama *Promedio Móvil Integrado Autorregresivo Estacional* o SARIMAX (acrónimo en inglés de Seasonal Autoregressive Integrated Moving Average,).

Una vez entrenado el modelo con la función SARIMAX (statsmodels.tsa.statespace.sarimax.SARIMAX) siguiendo el procedimiento descrito por Jason (Jason Brownlee, 2010), representamos el resultado de la predicción de consumo de energía junto con los datos de consumo energético de entrenamiento, viendo similitud visual (Figura 47).

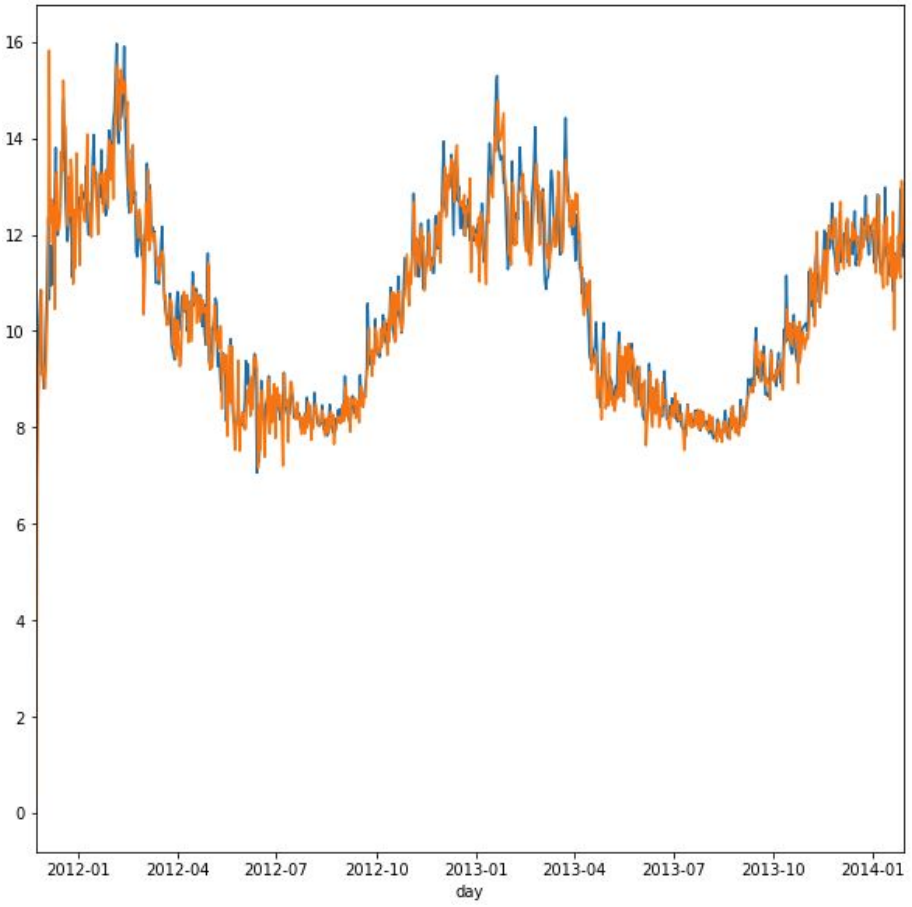

Figura 47. Representación de Datos de entrenamiento junto a la predicción.

Ahora añadimos una nueva columna a los datos de entrenamiento (Figura 48).

day	avg_energy	weather_cluster	holiday_ind	predicted
2014-02-23	11.673756	0	0	11.558238
2014-02-24	10.586235	0	0	10.709800
2014-02-25	10.476498	0	0	11.448047
2014-02-26	10.375366	0	0	11.871417
2014-02-27	10.537250	0	0	11.486039

Figura 48. Primeros datos del dataset de entrenamiento junto con la nueva columna predicted.

Para comprobar la precisión de la predicción, pasamos el modelo con los datos de test de los últimos 30 días obteniendo los siguientes resultados (Figura 49) donde encontramos que la media de error entre la predicción y el resultado real es de 0.5 unidades de medida de el valor avg_energy.

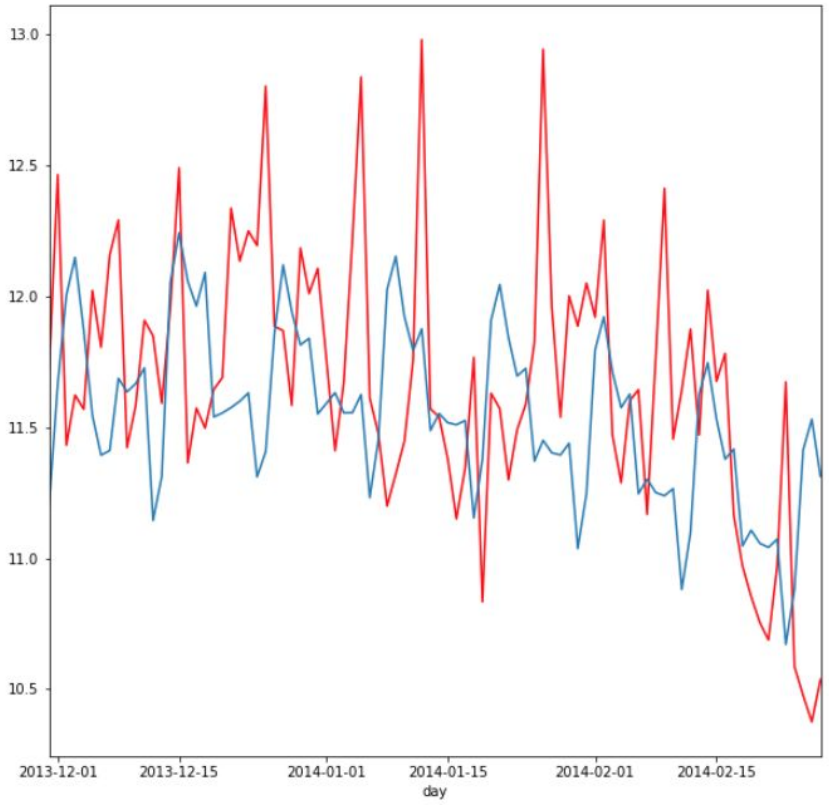

Figura 49. Gráfica de datos de test junto con la predicción.

Algoritmo LSTM

La memoria a largo plazo Long short-term memory o LSTM (acronimo del ingles Long short-term memory) es una arquitectura de red neuronal recurrente artificial o RNN (acrónimo del inglés recurrent neural network) utilizada en el campo del aprendizaje profundo, así el problema supervisado se convierte con los datos multivariados a un marco de datos supervisado.

Para convertir los datos en multivariable, se sigue el procedimiento descrito por Jason Brownlee (Jason Brownlee, 2010) para convertir los valores en la variable t y guardar los valores predecidos hasta siete días (Figura 50).

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
7	6.952693	8.536480	9.499782	10.267707	10.850805	9.103382	9.274873	8.813513
8	8.536480	9.499782	10.267707	10.850805	9.103382	9.274873	8.813513	9.227707
9	9.499782	10.267707	10.850805	9.103382	9.274873	8.813513	9.227707	10.145910

Figura 50. Detalle de dataset con las predicciones hasta en 7 días.

Una vez Normalizado y modelado el dataset, obtenemos la siguiente gráfica de evolución del entrenamiento donde se observa que a partir de la época 20, la función de coste tiende a 0 (Figura 51).

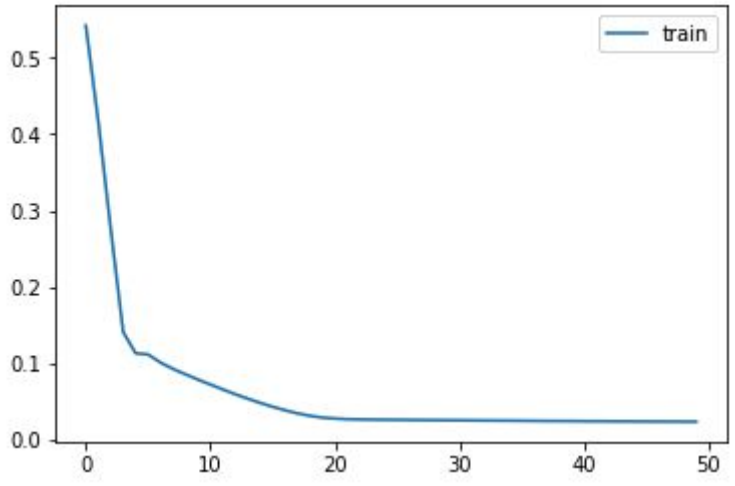

Figura 51. Evolución de entrenamiento.

Capítulo 11

Resultados

El resultado inicial del modelo ARIMA, tiene grandes diferencias entre la predicción y la energía consumida en el mismo periodo, lo que implica que el modelo obtenido, no tiene la precisión requerida.

Si representamos las diferencias entre la energía consumida y la predicción obtenida con el modelo ARIMA, observamos que a nivel acumulado hace inviable este modelo (Figura 52).

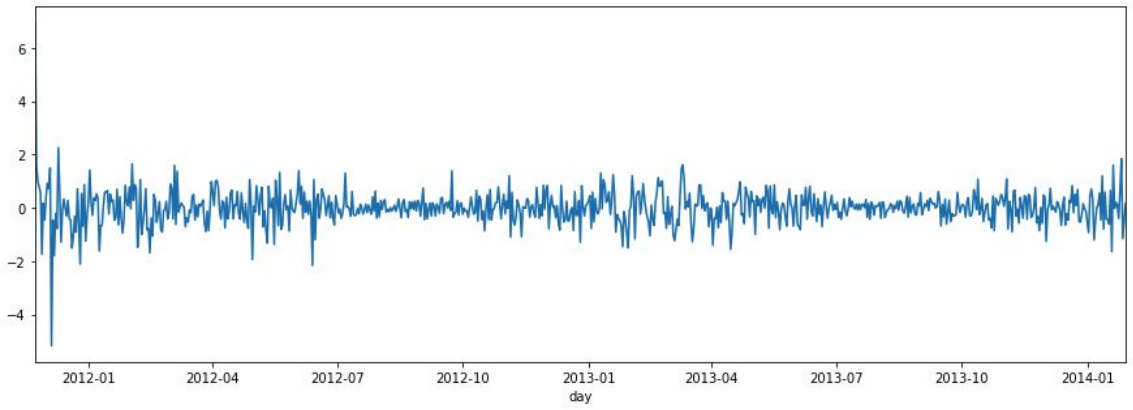

Figura 52. Representación de las diferencias entre los datos de entrenamiento y la predicción.

Las medidas de precisión de pronóstico denominadas Error Medio Absoluto o MAE (acrónimo en inglés de Mean Absolute Error) y Porcentaje de error medio absoluto o MAPE (acrónimo en inglés de Mean Absolute Percentage Error) ofrecen los siguientes resultados

```
MAE: 0.5857525774802094
MAPE: 5.243071233773702
```

El promedio entre la diferencia absoluta entre los valores predichos y el valor observado (MAE) y utilizando la misma unidad que el valor observado es de 0.58 con un desvío (MAPE) del 5.24% sobre el valor real.

El resultado con LSTM

Una vez generado el modelo, vemos que el test RMSE de da como resultado RMSE: 1.179, esto indica que la desviación entre el valor predicho y el valor real de media tiene un valor de 1.179

En la representación gráfica en la que se superponen los valores de test junto con la predicción, podemos observar que la coincidencia es cercana, correspondiente al Raíz del error cuadrático medio de 1,179 (Figura 53).

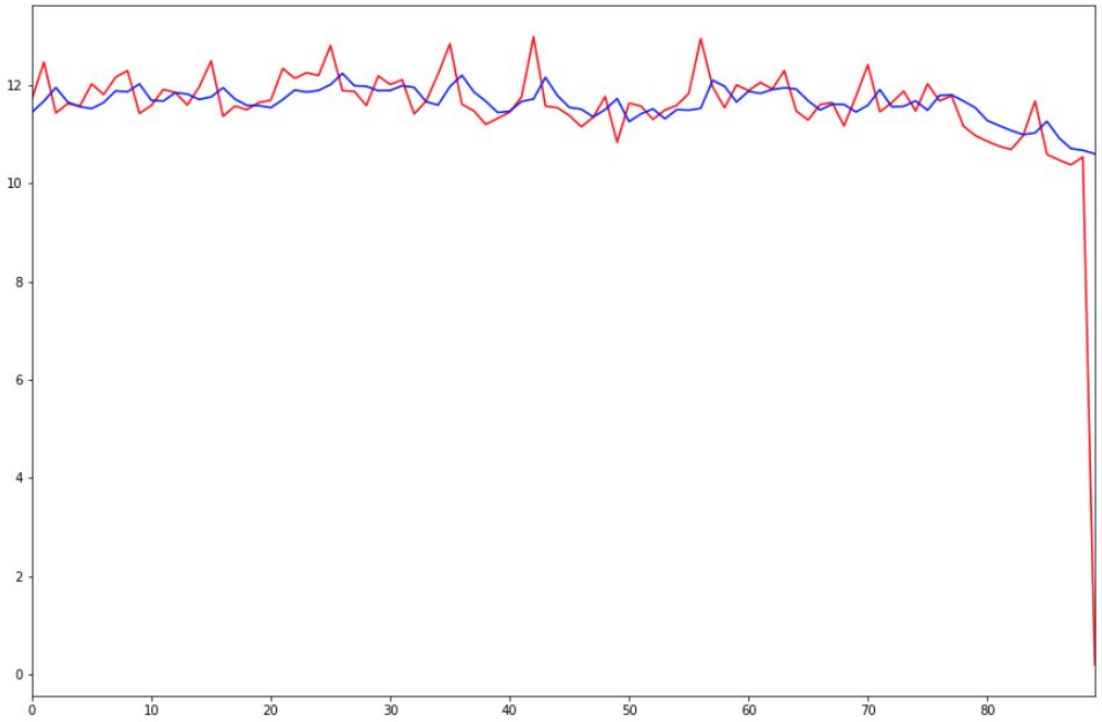

Figura 53. Representación de los datos de entrenamiento y la predicción con LSTM.

Parte III

Epílogo

Capítulo 12

Conclusiones

Como conclusión a la primera hipótesis planteada, queda demostrado que es posible determinar que variables externas son las que más influyen a la hora de crear un modelo matemático que permita realizar una predicción.

Los datos proporcionados en los ficheros de origen han sido procesados para compararlos con la energía acumulada consumida, esto nos ha permitido comparar de forma visual los datos en búsqueda de patrones de correlación, y han sido corroborados con la matriz de correlación

Las variables que más influyen son la temperatura máxima, la humedad relativa, y la velocidad del viento, tal y como hemos podido comprobar en el apartado de conclusión de las variables.

La segunda hipótesis, relativa a la tasa de precisión de la predicción, no ha podido ser verificada, puesto que hay una desviación un poco mayor que 1 respecto a los datos reales desde los predichos utilizando modelos basados en LSTM

.

Capítulo 13

Trabajos Futuros

Utilizar alguna técnica de regularización para evitar el overfitting como por ejemplo Dropout, que consiste en que para cada etapa de entrenamiento, los nodos individuales se eliminan de la red con probabilidad 1-p o se mantienen con probabilidad p, de modo que queda una red reducida; Los bordes entrantes y salientes de un nodo descartado también son eliminados, de esta forma se evita el sobreentrenamiento

También es posible evaluar otras arquitecturas de redes recurrentes basadas en celdas de Unidad Recurrente Cerrada o GRU (acrónimo del inglés Gated Recurrent Unit) siendo una variación mejorada de LSTM

Para resolver el problema de la degradación de la desaparición de un RNN estándar, se puede actualizar y restablecer la puerta. Básicamente, estos son dos vectores que deciden qué información se debe pasar a la salida. Lo especial de ellos es que pueden capacitarse para mantener la información desde hace mucho tiempo, sin purga de datos a través del tiempo ni eliminar la información que es irrelevante para la predicción.

Acrónimos

TFM: Trabajo Final de Máster.

UOC: Universitat Oberta de Catalunya.

GPU:Unidad de procesamiento gráfico.

Jupyter Notebook: Entorno de trabajo interactivo que permite desarrollar código en Python de manera dinámica, a la vez que integrar en un mismo documento tanto bloques de código como texto, gráficas o imágenes.

ARIMA: Autoregressive Integrated Moving Average Model.

MSTS: Múltiple Seasonality time Series.

RNA: Red Neuronal Artificial.

ACORN: Se trata de una clasificación de consumidores que segmenta la población del Reino Unido a nivel de código postal en 6 categorías, 18 grupos y 62 tipos. https://www.caci.co.uk/products/product/acorn

UK. United Kingdom.

CP1252: Windows-1252 o CP-1252 es una codificación de caracteres del alfabeto latino, usada por defecto cuando unicode no se usa en los componentes oficiales de Microsoft Windows en inglés y en algunos lenguajes occidentales.

Índice IV: es un indicador de la intensidad de radiación ultravioleta proveniente del Sol en la superficie terrestre.

RNN: Red Neuronal Recurrente.

LSTM: Long Short-Term Memory Units.

RMSE: Raíz del error cuadrático medio.

Bibliografía

Antolin, F. (1988) 'Electricidad y crecimiento economico. Los inicios de la electricidad en España', *Revista de Historia Económica / Journal of Iberian and Latin American Economic History*. Cambridge University Press, 6(03), pp. 635–655. doi: 10.1017/S0212610900000938.

Asteriou, D. and Hall, S. G. (2011) *ARIMA Models and the Box–Jenkins Methodology*, *Applied Econometrics (Second ed.). Palgrave MacMillan.* Available at: https://autobox.com/makridakis.pdf (Accessed: 17 March 2019).

Báez-Matos, J. F., Rodríguez, R. A. J.- and Abreu, L. V.- (2018) 'CONFIGURACIONES DE LAS REDES ELÉCTRICAS DE DISTRIBUCIÓN PRIMARIA QUE DETERIORAN SU EFICIENCIA ENERGÉTICA (Original)', *Redel. Revista granmense de Desarrollo Local*, 2(3), pp. 69–82. Available at: http://revistas.udg.co.cu/index.php/redel/article/view/117 (Accessed: 19 May 2019).

Bianco, V., Manca, O. and Nardini, S. (2009) *Electricity consumption forecasting in Italy using linear regression models*, *Energy*. doi: 10.1016/j.energy.2009.06.034.

Cayetano, E. M. de G. U. de M. and Ramón, G. M. de C. del T. U. de E. (2010) *AGUA Y ENERGÍA: PRODUCCIÓN HIDROELÉCTRICA EN ESPAÑA*. Secretariado de Publicaciones, Universidad de Alicante. Available at: http://rua.ua.es/dspace/handle/10045/17169

El sistema eléctrico español: diversificado, sobredimensionado, aislado… - El Blog de Ignacio Mártil (2016). Available at: https://blogs.cdecomunicacion.es/ignacio/2016/05/17/el-sistema-electrico-espanol-diversificado-sobredimensionado-aislado/ (Accessed: 19 May 2019).

Ley 24/2013, de 26 de Diciembre, del Sector Eléctrico (2019) 'Ley 24/2013, de 26 de Diciembre, del Sector Eléctrico', *Boletín Oficial de Estado*, pp. 105198–105294. doi: Ley 24/2013, de 26 de diciembre, del Sector Eléctrico.

Gale Group., Y. P. G. J. G. C. M. M. H. M. (1928) *Tecnologia Quimica.*, *Tecnología Química*. Universidad de la Habana, Departamento de Actividades Culturales. Available at: https://www.redalyc.org/html/4455/445543748008/ (Accessed: 24 March 2019).

Geurts, M., Box, G. E. P. and Jenkins, G. M. (2006) 'Time Series Analysis:

Forecasting and Control', *Journal of Marketing Research*, 14(2), p. 269. doi: 10.2307/3150485.

Gómez Expósito, A. (2000) 'Análisis y operación de sistemas de energía eléctrica.', p. 793. Available at: https://www.worldcat.org/title/analisis-y-operacion-de-sistemas-de-energia-electrica/oclc/932806716&referer=brief_results.

Martín Chicharro, G. J. (2016) 'PPT: Sistemas De Almacenamiento De Energía', *Presentation*, pp. 1–133. Available at: https://uvadoc.uva.es/bitstream/10324/18325/1/TFG-P-432.pdf (Accessed: 23 March 2019).

Nandwani, S. (2005) *Energía solar. Conceptos básicos y su utilización*, *Universidad Nacional, Heredia (Costa Rica). Jun*. Available at: http://www.catalogosolar.mx/download/Energia_Solar_Conceptos_Basicos.pdf (Accessed: 19 May 2019).

Parlamento Europeo y Consejo de la Unión Europea (2009) 'DIRECTIVA 2009/72/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJOOPEO Y DEL CONSEJO de 13 de julio de 2009', 2008. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:211:0055:0093:ES:PDF.

Yule, G. U. (1927) 'On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 226(636–646), pp. 267–298. doi: 10.1098/rsta.1927.0007.

Jason Brownlee (2010) 'Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras', Econometric Reviews, 29(5–6), pp. 594–621. doi: 10.1080/07474938.2010.481556.

Peña, D. (2019) *Estadística, modelos y métodos / Daniel Peña Sánchez de Rivera*, *SERBIULA (sistema Librum 2.0)*.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', Neural Computation, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Pilar, M., Casimiro, G. and Casimiro, P. G. (no date) *Análisis de series temporales: Modelos ARIMA*. Available at: https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf (Accessed: 17 March 2019).

Roos Fraga, J. (2019) 'Game theoretic approach to define optimal strategies of users in a smart grid'. E.T.S.I. Industriales (UPM). Available at: http://oa.upm.es/54238/ (Accessed: 19 May 2019).

Tsoutsos, T., Frantzeskaki, N. and Gekas, V. (2005) ‘Environmental impacts from the solar energy technologies’, *Energy Policy*. Elsevier, 33(3), pp. 289–296. doi: 10.1016/S0301-4215(03)00241-6.

Universidad de Alicante. Instituto Universitario de Geografía., C. and García Marín, R. (2010) *Anales de la Universidad de Alicante. Investigaciones geográficas*. Secretariado de Publicaciones, Universidad de Alicante. Available at: http://rua.ua.es/dspace/handle/10045/17169 (Accessed: 19 May 2019).

Veit, A. *et al.* (2014) ‘Household Electricity Demand Forecasting -- Benchmarking State-of-the-Art Methods’. Available at: https://arxiv.org/pdf/1404.0200.pdf (Accessed: 17 March 2019).

Anexos

1-Fichero en formato Jupyter Notebook con el análisis de los documentos, la generación de modelos y análisis predictivo, también incluido experimento de predicción. TFM_CIENCIA DE DATOS_JORGE_ARIAS_2019_ANEXO1.ipynb