Contenido

PRÁCTICA 2 – Limpieza y análisis de datos.

1. ¿Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

De Kaggle.com he escogido el dataset de nombre: Hourly energy demand generation and weather (https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)

Este dataset contiene, a intervalos de horas, y durante un período de 4 años, los datos relacionados con el consumo y generación (de diferentes tipos de fuentes) de energía, su coste y la predicción del tiempo de España.

Me ha parecido un interesante dataset con el que trabajar porque es el resultado de unir no solamente cuatro fuentes información sino cuatro tipos diferentes de información, en cuanto a que no es solamente más información/observaciones realizadas, sino variables entre las que buscar relación.

También resulta interesante este dataset porque dispone de:

- un número importante de observaciones (más de 35.000)
- bastantes variables sobre las que poder trabajar (29)
- datos perdidos (en la propia descripción del dataset se identifican varias variables con "Missing values")

En cuanto a los posibles preguntas que se podría intentar responder:

- ¿qué es lo que más influye en el consumo de los diferentes tipos de energía?
- las previsiones que se realizan en cuanto a coste y carga, ¿se pueden considerar buenas previsiones?¿sería posible podría ajustarlas más?

- Ya que la energía solar y eólica son fuentes intermitentes, dependientes del clima y de las horas de sol ¿las previsiones de sol y viento son acertadas?

Fuera del ámbito de esta práctica, se podría analizar el consumo de las diferentes fuentes de energía y así, mejorar su rendimiento, e incluso el agrupar por los diferentes tipos de energías (biomasa, fósil,….)

2. Integración y selección de los datos de interés a analizar.

El dataset está formado por las siguientes variables iniciales (incluyos tanto el nombre original como la traducción en castellano para su mejor comprensión a la hora de la interpretación posterior de los datos) y los he agrupado en función de a qué tipo de información hacen referencia:

1. Time / Fecha y hora: de recogida de la información. En formato aaaa-mm-dd hh:mm:ss + 01:00
2. Generation biomass/ generación de bioenergía (materia orgánica de procedencia vegetal y/o animal)

Energías fósiles:

3. Generation fossil Brown coal/lignite / generación de carbón marrón / lignito
4. Generation fossil coal derived gas / generación de carbón derivado del gas (sintético)
5. Generation fossil gas / generación de gas
6. Generation fossil hard coal / generación de carbón
7. Generation fossil oil / generación de aceite
8. Generation fossil oil sale / generación de petróleo esquisto
9. Generation fossil peat / generación de turba
10. Generation geothermal / generación de energía geotérmica

Energías obtenidas a partir de (la fuerza del) agua:

11. Generation hydro pumped storage aggregated /
12. Generation hydro pumped storage consumption / generación por consumo de almacenamiento por bombeo hidráulico .
13. Generation hydro run-of-river and poundage / generación por caudal del río (creo que se refiere a la energía hidroeléctrica de pasada).
14. Generation hydro water reservoir / generación de energía eléctrica de embalse
15. Generation marine / generación de energía marina (o energía oceánica)

16. Generation nuclear / generación nuclear
17. Generation other / generación de otras fuentes
18. Generation renewable / generación de energía renovable
19. Generation solar / generación solar
20. Generation waste /generación a partir de residuos
21. Generation wind offshore /
22. Generation wind onshore / generación eólica en tierra

Previsiones climatológicas

23. Forecast solar day ahead /previsión solar
24. Forecast wind offshore day ahead / previsión del viento fuera de tierra

25. Forecast wind onshore daya head / previsión del viento en tierra

"Necesidades de energía"

26. Total load forecast / necesidades previstas
27. Total load actual /necesidades reales

Precio

28. Price day ahead: previsión del precio
29. Price actual: precio real

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Los datos tienen tanto elementos vacíos como elementos con valor 0. A continuación describo los elementos vacíos y los nulos y las decisiones que he ido adoptando en función del análisis de cada uno de esos atributos:

- **Elementos vacíos**

He identificados dos de los atributos Generation_hydro_pumped_storage_aggregated y forecast_wind_offshore_eday_ahead que contienen todas las observaciones con nulos, con lo que decido eliminar estos atributos puesto que no aportan nada al contener todo el atributo valores NA.

Al continuar buscando valores NA's se puede ver que hay bastantes atributos en los que el número de valores NA's son 18-19. Analizando estos casos, se ve que coinciden en el mismo registro estos valores nulos, es decir, las observaciones tienen todos estos atributos con valores nulos (que se identifican con los datos relacionados con generación de energía).

Dado que tenemos un elevado número de observaciones, decido eliminarlas. En un análisis más profundo se podría intentar identificar por qué todas estas observaciones tienen todos estos atributos con NA.

Al volver a revisar los NA se pueden detectar casos únicos con valor NA. Para estos atributos y dado el elevado número de observaciones y como no va a influir en el resultado, decido imputar como valor, la media de cada atributo respectivamente.

- **Elementos valor 0**

También se pueden observar varias columnas en las que todos los valores son 0. Concretamente ocurre en 6 columnas y ya que se trata de información relacionada con la generación de energía, decidimos suprimir estas columnas porque no nos van a proporcionar nada de información.

Se han detectado varios atributos em los que hay una única observación con un valor NA. En estos casos, y después de analizar la distribución de los datos y viendo que no hay mucha dispersión en los datos, hemos decidido aplicar la media del valor de ese atributo a la observación con NA.

Ahora ya solamente tenemos como valores nulos, el atributo total load actual, AQUÍ ES DONDE DEBERÍA APLICAR KNN PERO EL ORDENADOR ME BLOQUEA LA INSTALACIÓN DEL PAQUETE.

- **Elementos nulos**

Pasando ahora a analizar los valores 0, veo que existen 16 atributos que contienen valores 0 y estos atributos los podemos distribuir, para tratarlos, en dos grandes grupos:

- atributos que hacen referencia a información relacionada con la generación de energía
- atributos que hacen referencia a información relacionada con la predicción del tiempo.

En el caso del primer grupo, los relacionados con la generación de energía, se puede ver que existen dos casos muy diferentes: atributos en los que las observaciones con valor 0 son solamente 3-4 observaciones y atributos en los que representan un total del 30-35% de los casos.

Tanto en uno como en otro caso, decido dejar los valores 0 tal y como están, es decir considerarlos como datos válidos porque en caso en que hay pocos valores, puede ocurrir que no haya habido realmente generación de ese tipo de energía y si no ha sido así, su repercusión en el estudio no va a ser significativo. Mientras que en el otro caso, son un número muy importante de valores como para que hayan sido todos errores.

En el caso del segundo grupo de valores 0, es decir, el relacionado con la predicción del tiempo, en este caso hay 539 ocurrencias y decido utilizar la función Knn para imputar valores. PERO NO ME HA FUNCIONADO.

Todo este tratamiento de valores nulos y valores 0, ha permitido también trabajar la reducción, tanto en dimensionalidad (eliminación de atributos) como en cantidad (eliminación de observaciones) ya que se ha paso de tener 29 atributos y 35064 observaciones a tener 21 atributos y 35046 observaciones (se ha realizado una reducción de casi un 23%)

3.2 Identificación y tratamiento de valores extremos

En el caso de los outliers, de entre todos los atributos, identifico, en principio, tres atributos que podrían tener outliers (es decir, tres atributos en los que la diferencia entre la media y la mediana es superior al doble) pero después de visualizarlos gráficamente con un boxplot, se observa que sólamente el atributo hydro pumped tiene outliers.

Este atributo tiene un número bastante alto de “posibles outliers” como para despreciarlos o corregirlos, con lo que decidimos que realmente no son outliers, sino que son valores alejados de la media pero que son “posibles”. Después de analizarlo un poco más, creo que se puede afirmar que los valores no son outliers, sino que se trata de un atributo que tiene los valores en un rango muy amplio y con pocas repeticiones.

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quiere analizar/comparar (planificación de los análisis a aplicar)

Del total de variables del dataset, energy_dataset, ya que existe información sobre previsión solar y previsión del viento voy a estudiar los atributos que hacen referencia a ambos:

- Time
- Solar
- Eólica
- Previsión solar
- Previsión viento

- Carga previsión
- Carga real
- Precio previsión
- Precio real

Por otra parte y viendo el elevado número de observaciones que hay, voy a ver si consigo algo trabajando solamente con los datos del mes de enero y para ello creo el dataset datosenero, quedando reducido el número de observaciones a casi 3.000.
Con lo que he vuelto a aplicar reducción, tanto en las dimensiones, como en la cantidad de observaciones.

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Al analizar gráficamente la normalidad de las variables se ve claramente que no siguen una distribución normal, lo que es luego confirmado con la realización del test de Kolmogorov-Smirnov. En este test, en todos los casos p-valor es inferior a 0,05 lo que confirma que los datos no siguen una distribución normal.

Para la comprobación de la normalidad y al utilizar el test de Shapiro-Wilk, con el conjunto de todos los datos me aparecía un error indicándome que la muestra era demasiado grande como para poder utilizar este test. Al hacer el estudio de la comprobación de la normalidad para el dataset que contiene los datos solo del mes de enero, este error h desaparecido, por lo que también aparece en el código.

Repito el proceso de comprobación de la normalidad aplicándolo al conjunto de datosenero (el que tiene solamente los datos de todos los años pero solamente del mes de enero); las distribuciones siguen sin ser normales, pero al ser el número de observaciones inferior a 3.000, puedo aplicar el test de Shapiro-Wilk, que lo único que hace es confirmar la no-nomalidad.

Para comprobar la homocedasticidad y teniendo en cuenta que se trata de variables que no siguen una distribución normal he utilizado el test de Fligner-Killeen.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Para comparar los grupos de datos y viendo que no siguen una distribución normal hay que utilizar pruebas no paramétricas, pero en primer lugar vamos a ver si hay o no dependencias entre ellas.

Como no se cumple la normalidad con las variables que estamos trabajando, vamos a utilizar el test de Spearman.

Se puede ver que existe relación entre los pares de variables:

- Energía solar y la previsión de energía solar
- Energía eólica y la previsión de energía eólica
- Previsión de carga y la carga real

También hay relación entre el precio estimado y el precio real.

He aplicado la regresión lineal entre las variables:

- Energía solar y previsión solar y por el valor R-squared, 0,9868, podemos infererir la relación entre las variables
- Energía solar y Eólica y se puede comprobar por el valor de Multiple R-squared, 0,008 que no están correlacionadas.
- Carga_previsión y carga_real: se puede comprobar que existe una fuerte correlación (R-squared= 0,9929)
- Precio_prevision y precio_real: hay correlación, pero no excesiva correlación (R-Squared= 0,5986)

5. Representación de los resultados a partir de tablas y gráficas.

Si utilizamos un gráfico Q-Q para el análisis de la normalidad, comprobamos lo que ya sabíamos: que no siguen una distribución normal, puesto que los datos no se encuentran alineados a la diagonal.

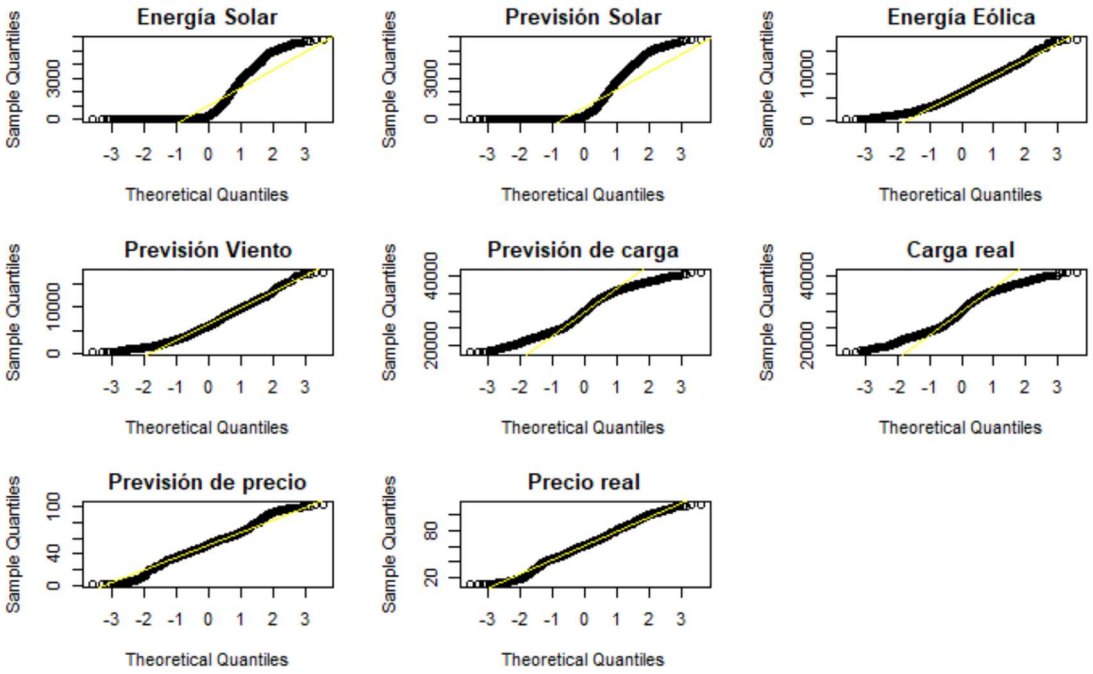
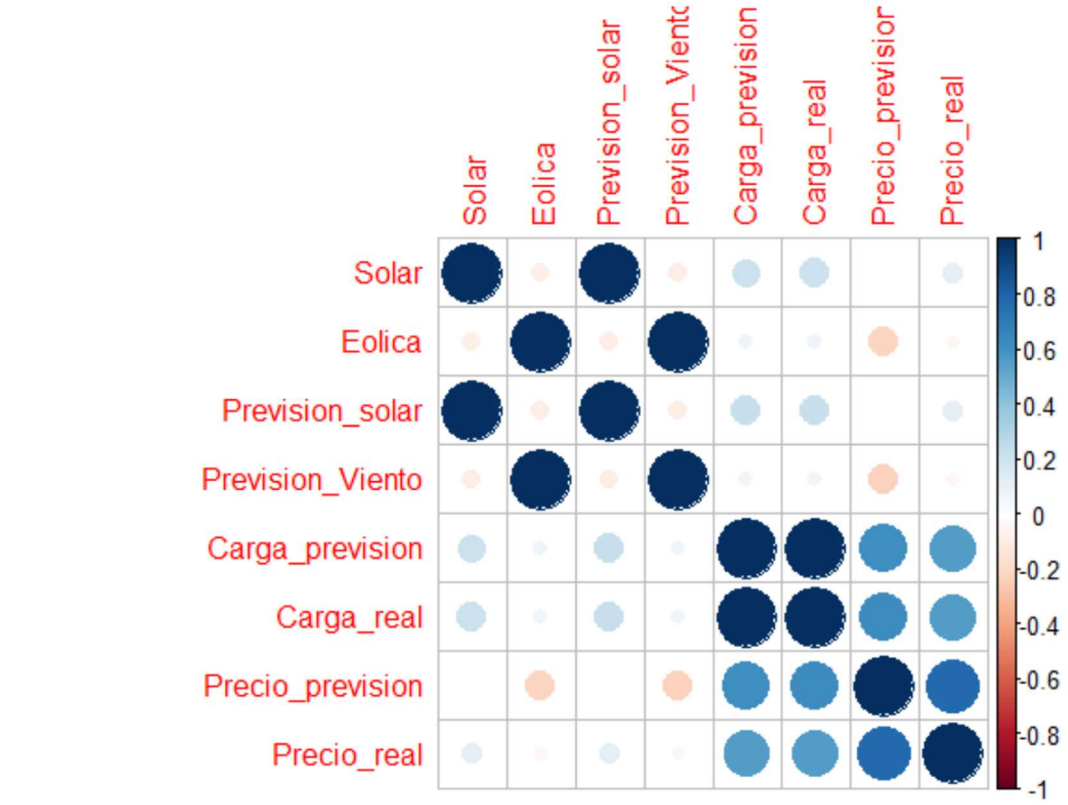
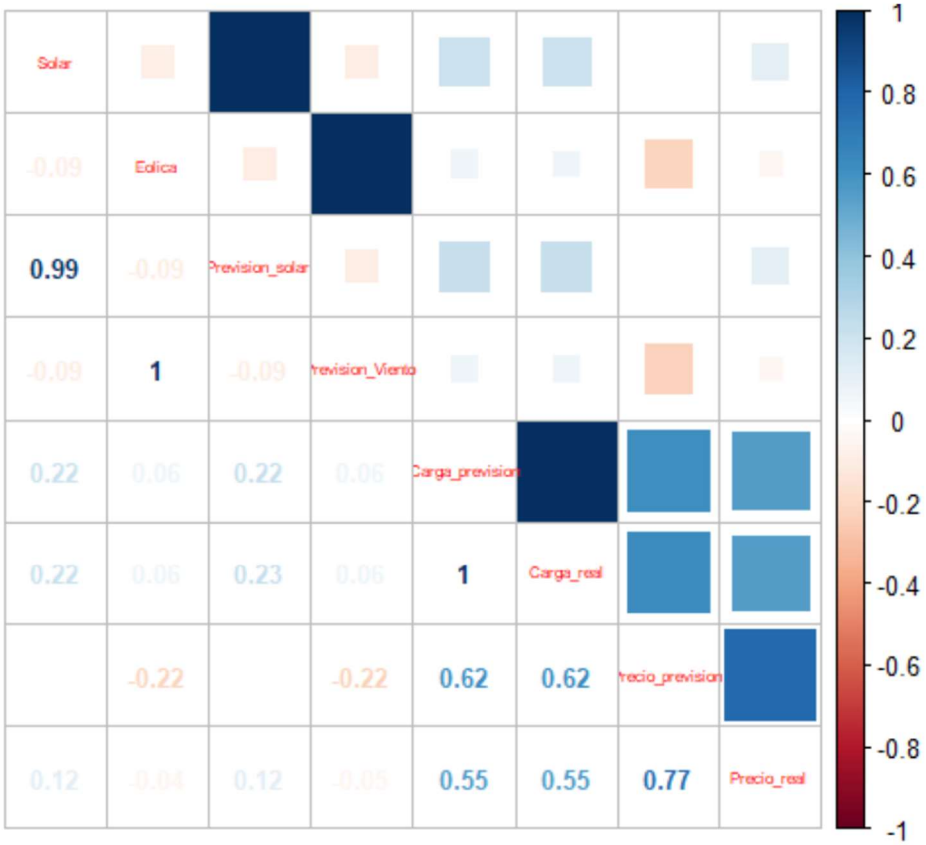


Gráfico de correlaciones entre pares del dataset: en el método de representación de los valores numéricos y gráficos, se puede ver lo que se ha analizado de manera numérica en el apartado anterior: hay correlación positiva entre las variables y sus previsiones correspondientes.

También se ven otras correlaciones no tan fuertes (en sentido positivo) pero que son importantes entre las previsiones de la carga y el precio, tanto la previsión como el precio real. Lo que se puede observar es que hay más relación entre la previsión de carga y la carga real que la relación existente en la previsión del precio y el precio real.


Solar
Eolica
Prevision_solar
Prevision_Viento
Carga_prevision
Carga_real
Precio_prevision
Precio_real
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Solar
Eolica
Prevision_solar
Prevision_Viento
Carga_prevision
Carga_real
Precio_prevision
Precio_real
0.99
1
0.22
0.22
0.06
0.23
0.06
0.06
-0.22
-0.22
0.62
0.62
0.12
0.55
0.55
0.77
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

6. Resolución del problema. A partir de los resultados obtenidos ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Sin lugar a dudas, la primera conclusión es que se debe trabajar más los análisis realizados; por ejemplo, haciendo agrupaciones de datos por meses o estaciones de año principalmente porque el tiempo afecta a los valores de todas aquellas generaciones de energía que dependen de la meteorología.

7. Código: hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

Al código se puede acceder a través del enlace del https://github.com/pborao/PRA2

8. Tabla de contribuciones al trabajo:

Contibuciones	Firma
Investigación previa	PBE
Redacción de las respuestas	PBE
Desarrollo código	PBE