DEMANDA DE ENERGÍA EN ESPAÑA

Weigandt, Herman
CoderHouse - Data Science

Agosto, 2023

Tabla de contenido

1. Descripción del caso de negocio

El mercado de energía es uno de los mercados más dinámicos de los últimos tiempos por varios aspectos. Por un lado, siempre hablando del mercado mayorista, la energía producida no puede almacenarse por lo que su producción, comercialización y distribución se realiza en tiempo real día a día. Por otra parte, los distintos países y regiones han ido adaptando sus matrices energéticas en consideración de la conciencia ecológica (huella de carbono y energías renovables), la independencia energética y el crecimiento del consumo, como asimismo el cambio de preferencias de los consumidores tanto por los nuevos sistemas de climatización, el uso de dispositivos electrónicos y la movilidad sustentable.

Atendiendo a esto y en especial consideración de que la energía producida no puede acumularse a gran escala, se considera primordial la predicción de la cantidad a producir/cantidad demandada, para evitar desperdicios que llevaran a su encarecimiento, al igual que la subproducción que puede llevar a la sobreexigencia del sistema energético o la falta de su servicio, el cual en muchos casos es de vital importancia para la refrigeración de alimentos u otros casos como el funcionamiento de equipamiento indispensable como lo es el hospitalario.

Sobre lo anterior y destacando la regionalidad del mercado mayorista de energía eléctrica, se aborda en el presente trabajo el mercado español, considerando cuatro regiones y estudiando tanto variables climáticas como de producción y comercialización de la misma.

2. Tabla de versionado.

Sólo existe una versión.

3. Objetivos del modelo.

Con base en todas las variables que presenta el dataset relevado, se propone estudiar: ¿Cuáles son las razones (variables) que determinan la demanda de energía eléctrica?, para de ésta manera predecir la misma a efectos de cubrirla de la manera más eficiente posible.

Partiendo de esta pregunta objetivo, resulta de interés identificar la mínima cantidad de variables que la determinan y así encontrar un modelo de predicción que permita saber con la mayor anticipación posible, el comportamiento de nuestra variable principal.

4. Descripción de los datos

Este conjunto de datos contiene 4 años de consumo eléctrico, generación, precios y datos meteorológicos para España. Los datos de consumo y generación se recuperaron de ENTSOE, un portal público para datos de operadores de servicios de transmisión (TSO). Los precios de liquidación se obtuvieron del TSO español Red Electric España. Los datos meteorológicos se compraron como parte de un proyecto personal de Open Weather API para las 5 ciudades más grandes de España y se publicaron, como dominio público, en el sitio Kaggle y se puede acceder en el siguiente Link.

A continuación se provee de una breve descripción de las variables:

Dataframe clima:

#	Column	Descripción	Non-Null Count	Dtype
0	dt iso	Fecha	178396 non-null	object
1	city_name	Ciudad	178396 non-null	object
2	temp	Temperatura (K°)	178396 non-null	float64
3	temp_min	Temp. minima (K°)	178396 non-null	float64
4	temp_max	Temp. máxima (K°)	178396 non-null	float64
5	pressure	Presión atmosférica	178396 non-null	int64
6	humidity	Humedad	178396 non-null	int64
7	wind_speed	Vel. del viento	178396 non-null	int64

```
 8   wind_deg   Dir. Del viento      178396 non-null  int64
 9   rain_1h Lluvia en la últ hs     178396 non-null  float64
 10  rain_3h Lluvia en las últ. 3hs  178396 non-null  float64
 11  snow_3h Nieve en las últ. 3 hs  178396 non-null  float64
 12  clouds_all  Niebla              178396 non-null  int64
 13  weather_id  Codificador         178396 non-null  int64
 14  weather_main Codificador        178396 non-null  object
 15  weather_description Descripción 178396 non-null  object
 16  weather_icon Ícono              178396 non-null  object
```

Dataframe producción y consumo:

```
#   Column            Descripción          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   time                             35064 non-null  object
 1   generation biomass                    35045 non-null  float64
 2   generation fossil brown coal/lignite       35046 non-null  float64
 3   generation fossil coal-derived gas         35046 non-null  float64
 4   generation fossil gas                 35046 non-null  float64
 5   generation fossil hard coal              35046 non-null  float64
 6   generation fossil oil                 35045 non-null  float64
 7   generation fossil oil shale              35046 non-null  float64
 8   generation fossil peat                35046 non-null  float64
 9   generation geothermal                    35046 non-null  float64
 10  generation hydro pumped storage aggregated   0     non-null  float64
 11  generation hydro pumped storage consumption  35045 non-null  float64
 12  generation hydro run-of-river and poundage   35045 non-null  float64
 13  generation hydro water reservoir          35046 non-null  float64
 14  generation marine                    35045 non-null  float64
 15  generation nuclear                   35047 non-null  float64
 16  generation other                     35046 non-null  float64
 17  generation other renewable               35046 non-null  float64
 18  generation solar                     35046 non-null  float64
 19  generation waste                     35045 non-null  float64
 20  generation wind offshore                35046 non-null  float64
 21  generation wind onshore                  35046 non-null  float64
 22  forecast solar day ahead                 35064 non-null  float64
 23  forecast wind offshore eday ahead          0     non-null  float64
 24  forecast wind onshore day ahead          35064 non-null  float64
 25  total load forecast  Cant. Pronosticada     35064 non-null  float64
 26  total load actual    Cant. Real          35028 non-null  float64
 27  price day ahead                      35064 non-null  float64
 28  price actual                         35064 non-null  float64
```

En principio se realizó un análisis de todas las variables. Se detecto que las columnas “forecast wind offshore day ahead” y “generation hydro pumped storage aggregated” se encuentran vacías y que la columna de cantidad real consumida (26) presenta 36 datos faltantes y algunas otras presentar varios valores iguales a 0, las cuales no se detallan en honor a la brevedad y que las mismas, según se explicará más adelante, no son relevantes.

El dataset contiene 46 variables con 35064 registros para el dataframe de producción y consumo y

178396 registros para el dataframe de clima de lo cual hay que destacar que respecto del clima hay 1 registro por cada ciudad relevada para un mismo momento, mientras que relativo al mercado de energía hay un registro para cada momento y la cantidad consumida no se encuentra discriminada por ciudad.

En el primer relevamiento de las variables, y en vistas de nuestro objetivo se realiza un primer descarte de variables por no considerarse pertinentes y se las excluye del EDA y de los modelos. Las variables excluidas son:

- Las variables del dataframe de producción y consumo, exceptuando las variables de cantidad consumida (total load actual) y de precio (Price actual).

Del total de las 46 variables, si quitamos las mencionadas arriba, nos quedamos con 19 variables, de las cuales se tienen :

- 8 variables categóricas
- 11 variables numéricas (cuantitativas)
  Entre las categóricas hay algunas ordinales y otras nominales.

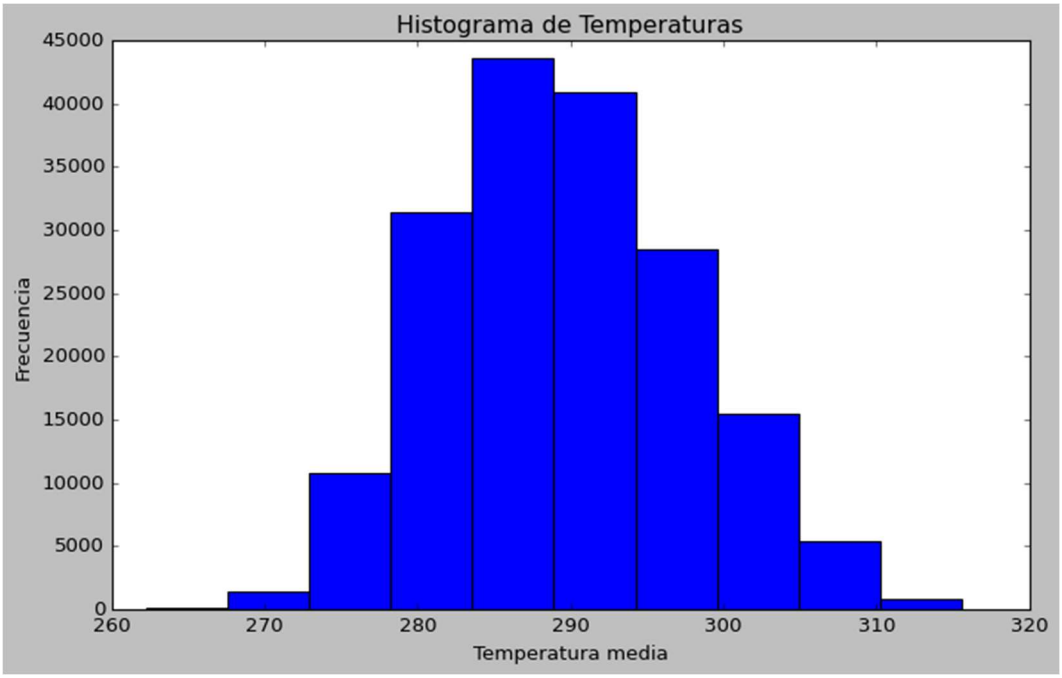
A continuación se detallan los principales resultados del EDA, de las demás variables estudiadas.

5. EDA: Exploratory Data Analysis

Análisis gráfico

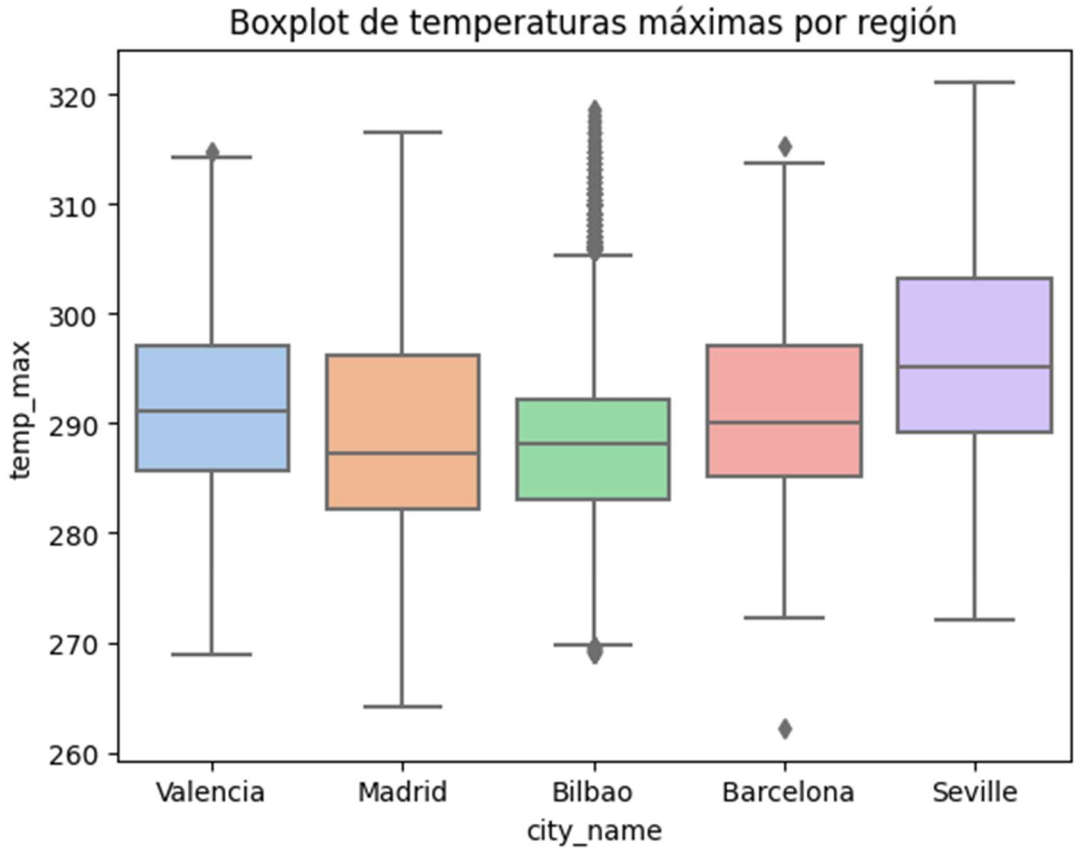
Para iniciar el análisis exploratorio de datos se estudió mediante un histograma la distribución de frecuencias de las temperaturas medias a nivel nacional y luego el comportamiento de éstas solo considerando la ciudad de Sevilla, a efectos de establecer rango de datos y comportamiento general. De ello se obtuvo lo siguiente

- El histograma de temperaturas medias nacionales presenta forma típica de distribución normal (campana de Gauss) con varianza pequeña y por ende gran concentración de valores en la proximidad de la media.



Se observa que la media es aproximadamente 290° K y el rango de valores se encuentra comprendido en el intervalo [260; 315]

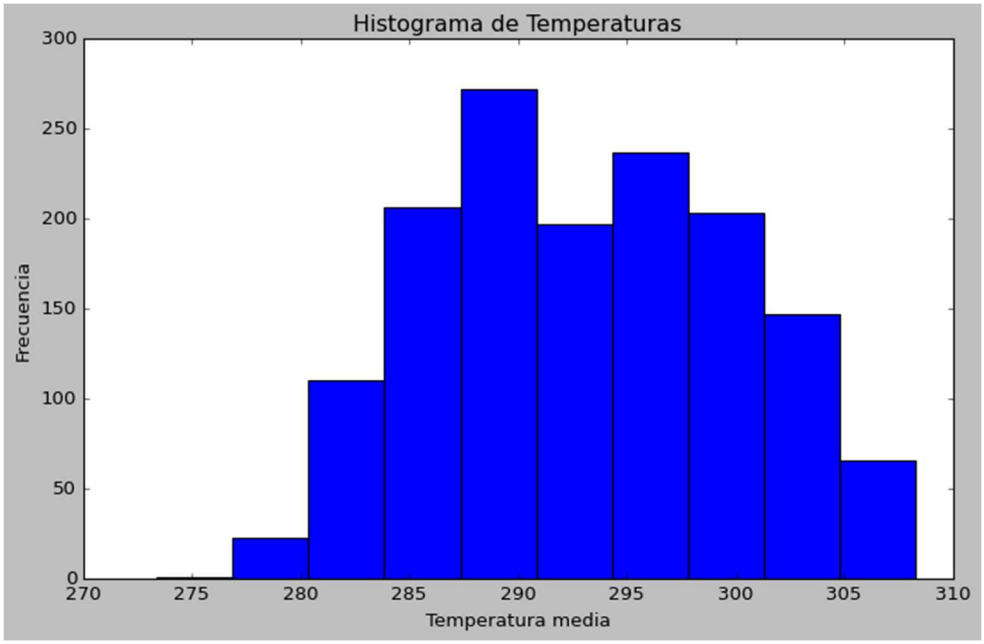
Para un mejor análisis comparativo entre los dos casos particulares en mención se recurre a un boxplot diferenciando las distribuciones por ciudad y considerando las temperaturas máximas para un conocimiento más acabado de los datos (con estudio de mayor número de variables) y tomando como supuesto de apoyo que las temperaturas máximas son las de mayor incidencia en el consumo eléctrico debido a la utilización de aires acondicionados y electrodomésticos afines como son las heladeras. El estudio arroja lo siguiente:



Interpretando el gráfico anterior, puede observarse que, salvando las diferencias en el rango de temperaturas, Sevilla y Madrid presentan la mayor amplitud térmica, teniendo Sevilla en términos generales temperaturas más elevadas. A su vez, Sevilla muestra una media más centrada entre los cuantiles 25% y 75%, como así también una menor distancia entre éstos, lo que se traduce en una concentración más elevada de los valores en términos relativos con la muestra inherente a la ciudad capital.

- Posteriormente, en el caso particular de Sevilla, se extrae que la temperatura media aproximada es de 292° K y las temperaturas con frecuencias mínimamente significativas se encuentran comprendidas entre 272° K y 308° K.

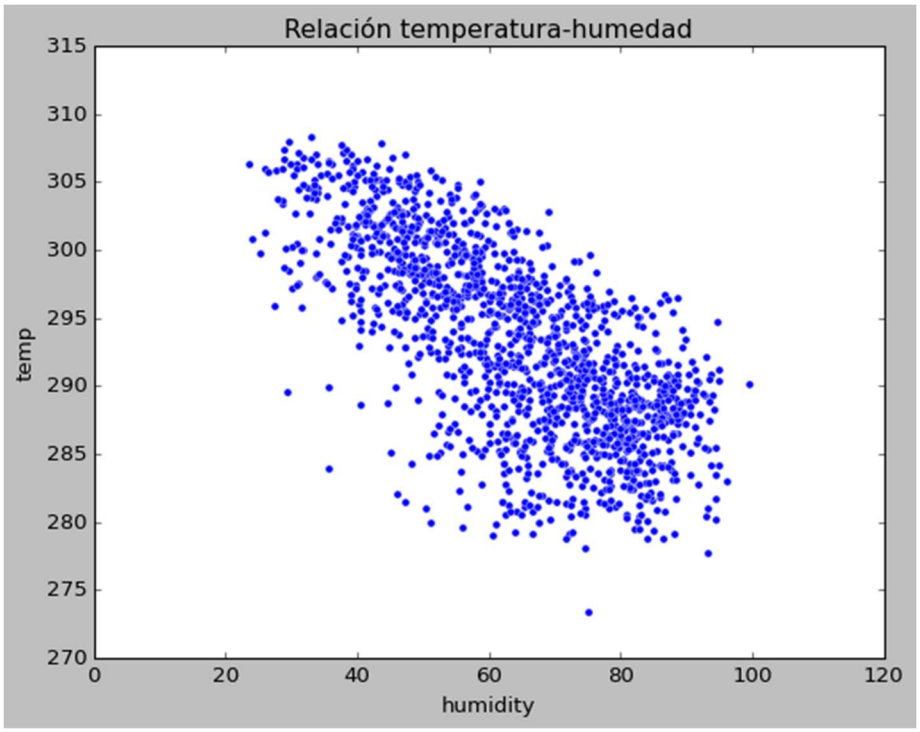
A su vez, se deduce a simple vista que los datos para este caso presentan una varianza mayor que en el histograma de la temperatura media a nivel nacional en su frecuencia y por lo tanto una mayor dispersión de los valores alrededor de la media.



En motivo de lo anterior, sumado al hecho de que las condiciones climáticas vienen detalladas por ciudad, pero los datos de consumo son a nivel nacional; a fin de poder construir un modelo basado en las condiciones climáticas para la predicción del consumo, se realiza el siguiente supuesto fundamental:

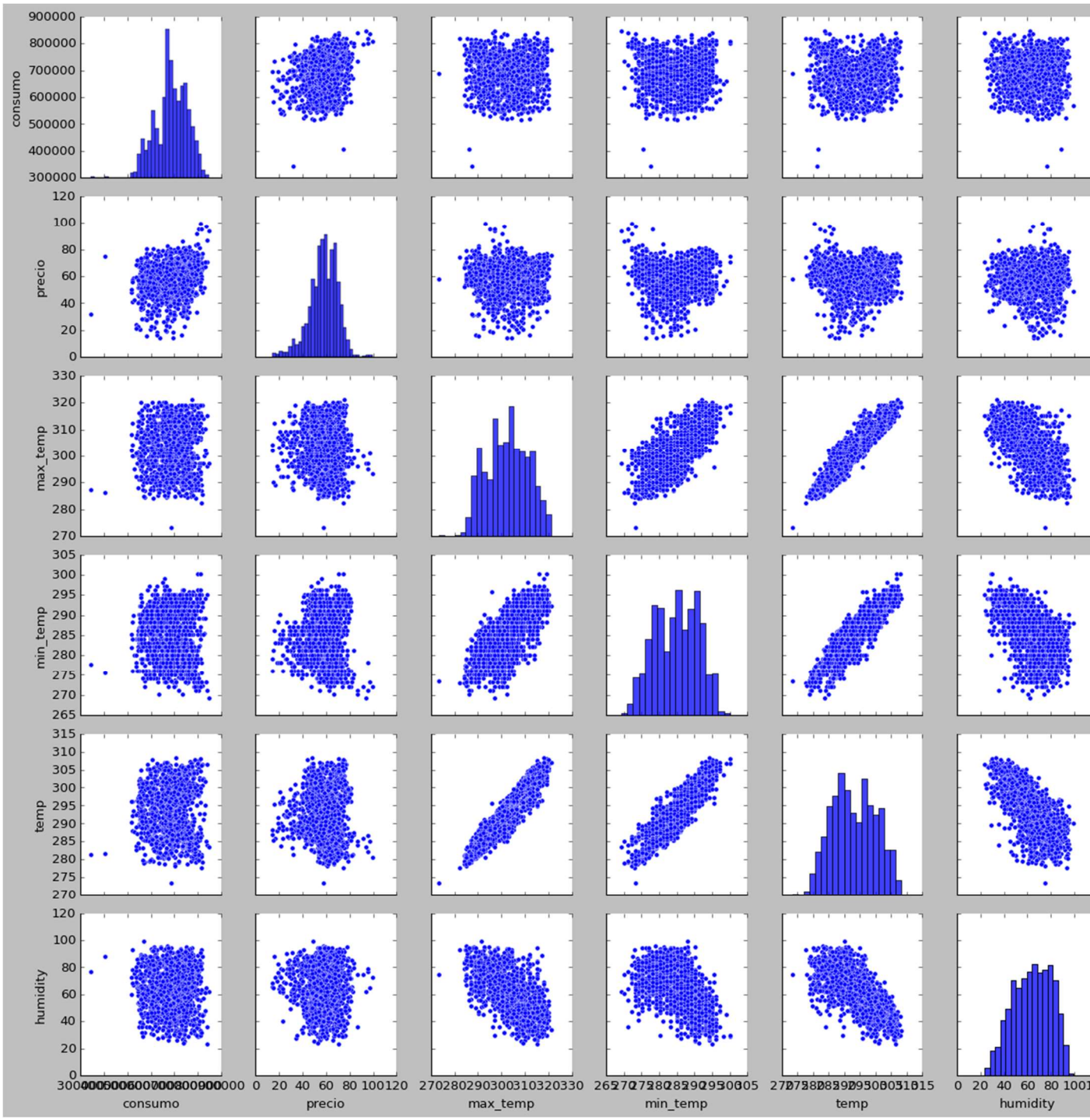
- Por la mayor amplitud térmica, la inexistencia de valores atípicos, su tendencia a temperaturas más altas, con las implicancias antes detalladas y el comportamiento de los valores alrededor de la media; se decide tomar como caso representativo para la construcción del modelo, la ciudad de Sevilla, uniendo estos valores a los de consumo nacional para cada momento temporal.

- Con motivo de profundizar en el estudio gráfico de las variables y en consideración de los conocimientos generales sobre clima, se selecciona como siguiente variable a analizar la humedad y su relación con la temperatura. El resultado se muestra a continuación:



La conclusión es evidente en cuanto a que humedad y temperatura presentan relación inversa entre ellas, con comportamiento lineal. Como conclusión visual se puede agregar que la humedad influye en la amplitud térmica de manera positiva, no así en la temperatura propiamente dicha.

- Ya habiendo analizado las principales variables climáticas a primera vista, procedemos a realizar un estudio bivariado de ellas a nivel general a lo largo del dataframe bajo estudio mediante un gráfico pair.plot de la biblioteca Seaborn, obteniendo lo siguiente:



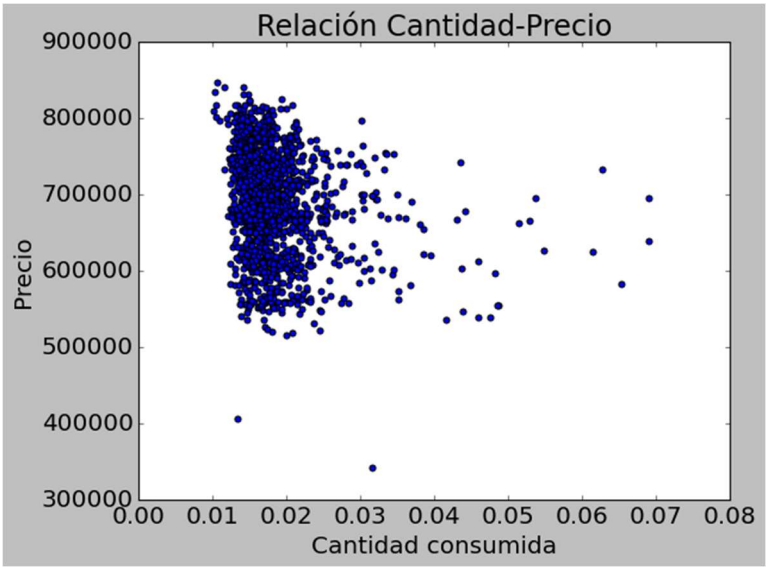
Se extrae del presente examen que:

- Las temperaturas máxima, mínima y simple se encuentran relacionadas entre sí de manera positiva, hecho que se considera una conclusión lógica, si no obvia.
- El consumo de energía no muestra un comportamiento claro respecto de ningunas de las variables del resto del conjunto, al menos en este

análisis bivariado.

- ❖ Lo que podría esperarse como comportamiento de la cantidad consumida en relación al precio (relación inversa), no es tal lo que requerirá un estudio más profundo.

- ➢ Dado lo anterior y recurriendo a la teoría económica, haciendo hincapié en que el consumo de energía presentaría una demanda inelástica debido a que independientemente de su precio, la utilización de los electrodomésticos se mantiene, o incluso es indispensable como es el caso de las heladeras y en menor medida, pero sostenido la utilización de los sistemas eléctricos de climatización; pasa a considerarse la relación de la demanda con el recíproco de su precio (1/Px), lo cual se plasma visualmente con el siguiente gráfico:



Fundamentado esto, se podría decir además que el mercado eléctrico presenta cierta estabilidad, lo que no justificaría grandes variaciones de precios, excepto cambios económicos coyunturales, que pueden considerarse no habituales en economías del primer mundo, aunque no imposible, como evidenció el impacto de la guerra Rusia-Ucrania y adyacente crisis de gas y por extensión energética en todo Europa.

Análisis matemático

Superado el primer pantallazo a nuestros datos, recurriremos a herramientas matemáticas para una elección eficiente de nuestras variables. Para ello elegimos el método PCA (Principal Components Analysis), en el cual, solo tomando variables numéricas, se determina el porcentaje de variabilidad que explica cada una y se ordenan, tomando esta categorización, de mayor a menor.

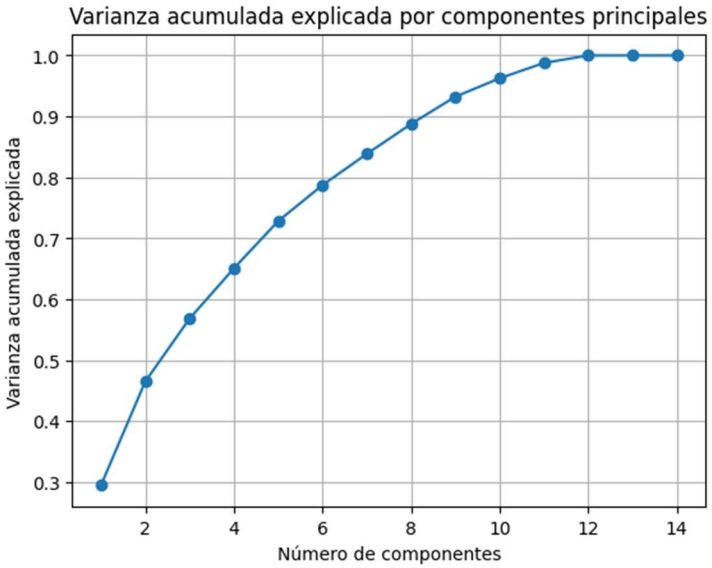
Como criterio general se busca que el modelo desarrollado pueda explicar al menos el 95% de la variabilidad de la variable explicada.

Éste método, además de jerarquizar la importancia de las variables dentro del modelo, busca brindar una herramienta precisa para la reducción de la dimensionalidad, lo que es sumamente deseable, entre otros motivos, por la claridad a la hora de extraer conclusiones.

Para evitar problemas de magnitudes por tener las variables distintas unidades de medida, se estandarizan previo a la aplicación del método (restando a cada valor su media y dividiendo por su desvío estándar).

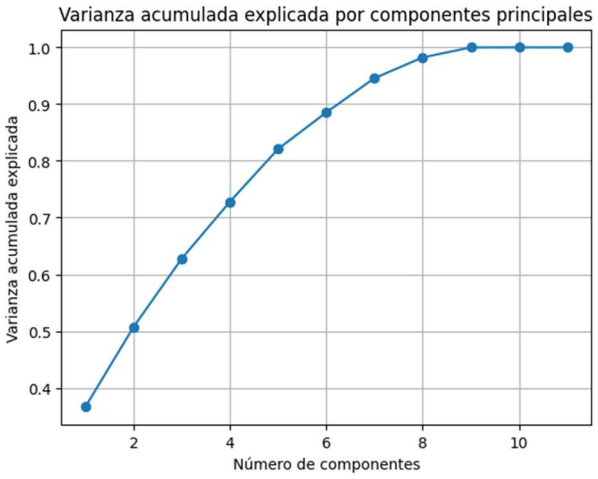
En la primera prueba incluimos todas las variables ya seleccionadas:

Variable	Varianza Explicada
temp	0.29463220578205646
temp_min	0.17021262859307634
temp_max	0.10330780236205588
pressure	0.08243169339701324
humidity	0.07753309313984079
wind_speed	0.059016537669030866
wind_deg	0.05123982066651435
rain_1h	0.04915948209345598
rain_3h	0.04475755859929375
snow_3h	0.03016949972956929
clouds_all	0.025254397980920072
weather_id	0.012219183126557709
total load actual	6,60969E-05
price actual	1,62053E-19



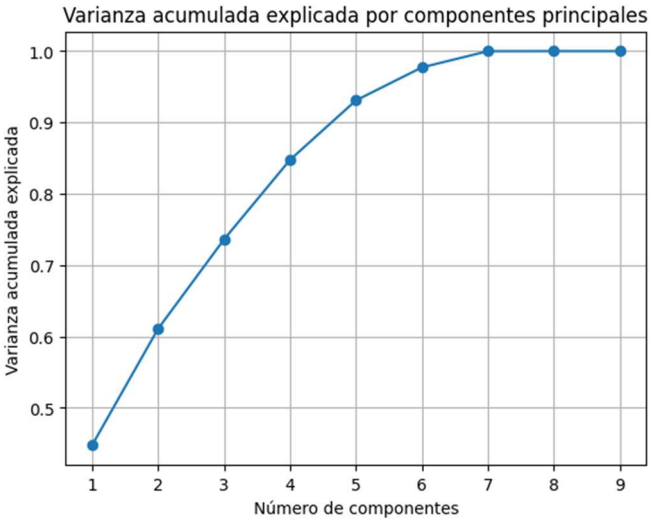
Para lograr una mejor aproximación se eliminan las variables que a pesar de ser numéricas representan categorías (clouds_all y weather_id) como así también la variable relativa a la dirección del viento. Luego se realiza nuevamente el cálculo.

Variable	Variance Explained
temp	0.3662748084502922
temp_min	0.14091433386935465
temp_max	0.12046811111730812
pressure	0.10028664884879483
humidity	0.09304749410973102
wind_speed	0.06438051977718127
rain_1h	0.06041822522067787
rain_3h	0.03646671484909068
snow_3h	0.017657053161306634
total load actual	8,61E-05
price actual	4,86859E-20



Eliminamos ahora el precio por ser el menos relevante (y con fundamento teórico en nuestras conclusiones del análisis gráfico) y la cantidad por ser nuestra variable objetivo y repetimos el test:

Variable	Varianza Explicada
temp	0.4473823782640926
temp_min	0.16320968145362633
temp_max	0.12535691731601104
pressure	0.11154719374128576
humidity	0.08341747410513167
wind_speed	0.046303325994775754
rain_1h	0.02267500733700965
rain_3h	0.00010802178806735523
snow_3h	0.0



En este testeo observamos que las últimas 3 variables no son de relevancia según nuestro criterio. A su vez observamos que de eliminar la 6° variable, no alcanzaríamos el 95% de varianza explicada.

Atento a ello, conservamos como definitivas para nuestro modelo las primeras 6 variables.

Conjetura extra.

A fin de completar el análisis numérico se realiza una hipótesis extra sobre el comportamiento de los datos, la cual se enuncia de la siguiente manera:

- El consumo de energía, más que de la temperatura , dependerá de la diferencia entre ésta y la temperatura de confort, todo ello elevado al cuadrado. Lo anterior se explica en base a qué alejarnos de la temperatura en cuestión, nos llevará a calefaccionar o refrescar nuestro ambiente y ello elevará el consumo de energía. Esto mismo, se toma elevado al cuadrado para convertir en positivos los valores de la diferencia explicada, para temperaturas por debajo de la de confort, la cual es considerada la temperatura de 22° C.

La formula de ello, quedaría expresada así: Dif. Temp. = $(\text{Temp.} - 22)^2$.

Aplicado el método PCA para esta hipótesis, en los casos de 9 variables y de 6 variables (descartando las temperaturas por su correlación con la variable creada), se arrojan los siguientes resultados:

Variable	Varianza Explicada
temp	0.4033840619893744
temp_min	0.14940962878746783
temp_max	0.11254558587664389
pressure	0.11142806098153077
humidity	0.09346176193946029
wind_speed	0.07255827056784468
rain_1h	0.03705468922007599
rain_3h	0.020062684283248488
snow_3h	9,53E-05
diferencia_confort	0.0

Variable	Varianza Explicada
pressure	0.2440362640762379
humidity	0.20188213253248552
wind_speed	0.16717052632579493
rain_1h	0.1624539089013531
rain_3h	0.12702280440861374
snow_3h	0.09743436375551483
diferencia_confort	0.0

Atento a que, en ninguno de los casos, esta nueva variable explica la varianza, se descarta la

hipótesis y se conservan las variables seleccionadas en el paso anterior.

Dataset seleccionado

Luego de aplicar las herramientas antes descriptas y realizado el análisis transmitido, el dataset a modelar es el siguiente:

dt_iso	temp	temp_min	temp_max	pressure	humidity	wind_speed	total load actual
2014-12-31 23:(	0.22500	0.225000000000	0.225000000000	1039	75	1	25385.0
2015-01-01 00:(	0.22500	0.225000000000	0.225000000000	1039	75	1	24382.0
2015-01-01 01:(	0.93600	0.936000000000	0.936000000000	1039	71	3	22734.0
2015-01-01 02:(	0.93600	0.936000000000	0.936000000000	1039	71	3	21286.0
2015-01-01 03:(	0.93600	0.936000000000	0.936000000000	1039	71	3	20264.0

6. Algoritmo Elegido.

A fin de poder aplicar los modelos de ML, se generó una codificación a las variables categóricas, es decir, se les asignó un valor numérico, mediante el algoritmo de *labelencoder*.

Los modelos utilizados para predecir la variable Attrition fueron los siguientes:

1. Árbol de Decisión
2. Random Forest
3. Regresión Logística
4. KNN
5. XGBoost

En un primer lugar se testearon los modelos con el dataset completo, filtrando las variables que se mencionaron en la sección anterior, tomando una división de Train/Test con una relación de 70/30.

Finalmente se utilizó una técnica de optimización de parámetros, sobre tres de los modelos implementados, el bosque aleatorio, y el XGBoost.

Cabe aclarar que originalmente se construyeron los modelos buscando optimizar el resultado de las métricas r^2, MSE y MAE,

Random Forest Regressor:
Gradient Boosting Regressor:

7. Métricas de Desempeño del Modelo.

A continuación se presentan los resultados obtenidos

- **Modelos con dataset considerando la técnica Oversampling**:

Métrica \ Modelo	arbol_1b	arbol_2b	forest_1b	forest_2b	LogRegb	knn_b	XGboost_b
Accuracy	0.646259	0.811791	NaN	NaN	0.673469	NaN	0.786848
Precision	0.176870	0.333330	NaN	NaN	0.221480	NaN	0.296300
Recall	0.426230	0.360660	NaN	NaN	0.540980	NaN	0.393440
ROC_curve	0.617470	0.641780	NaN	NaN	0.641780	NaN	0.620530

- **Modelos optimizados**:

Métrica \ Modelo	arbol_ optimizado
Accuracy	0.82337
Precision	0.22222
Recall	0.03333
ROC_curve	0.66523

8. Iteraciones de Optimización.

No se realizaron iteraciones sobre los modelos optimizados.

9. Métricas finales del Modelo Optimizado.

El modelo considerado óptimo SEGÚN los resultados de las métricas obtenidas, fue la Regresión Logística, para la cual se obtuvo:

Métrica \ Modelo	Regresión Logística optimizada
Accuracy	0.866848
Precision	0.666670
Recall	0.366670
ROC_curve	0.796750

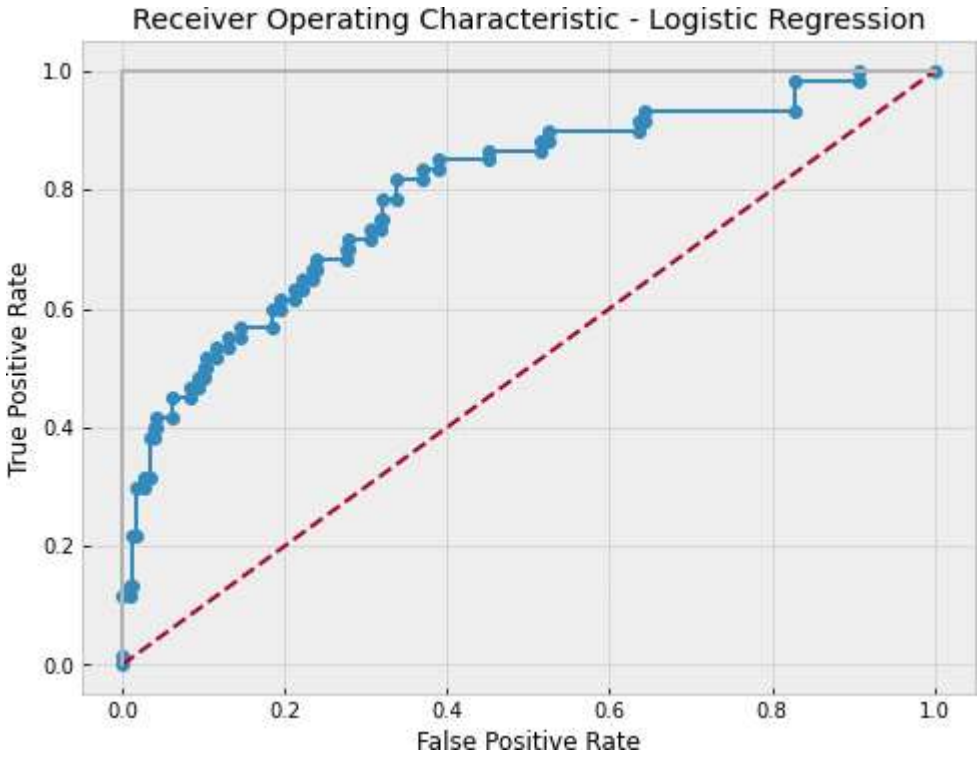


Figura 9.1 **Curva.**

10. Futuras líneas

11. Conclusiones: