Análisis descriptivo para la detección de patrones y anomalías en el consumo de Clientes para Electrodunas

Reporte técnico de selección de modelos, variables y parámetros

Para la solución de analítica solicitada por ElectroDunas, el cliente entregó un archivo por cada cliente tipo CSV con la lectura de cuatro variables hora a hora de energía activa, energía reactiva, potencia activa y potencia reactiva. Adicionalmente, se entregó una tabla con el sector económico al que pertenece cada cliente.

Exploración de datos

Completitud y consistencia

Los archivos CSV entregados por ElectroDunas, se cargaron en un único repositorio de GitHub con el fin de ejecutar el proceso ETL con los siguientes archivos:

- 24 archivos CSV - Datos consumo clientes
- 1 archivo CSV - Datos sector económico

A cada archivo de consumo se verificó la completitud y consistencia. Se verificó el número de filas y columnas, el rango de fechas del consumo, presencia de datos nulo o negativos y si la serie temporal tenía datos faltantes. A continuación se muestra el resultado de los primeros 10 archivos:

	Nombre_archivo	Numero_filas	Numero_columnas	Fecha_minima	Fecha_maxima	Datos_nulos	Valores_negativos	Registros_faltantes
0	DATOSCLIENTE1.csv	19681	5	2021-01-01	2023-04-01 00:00:00	0	0	0.0
1	DATOSCLIENTE10.csv	19681	5	2021-01-01	2023-04-01 00:00:00	0	0	0.0
2	DATOSCLIENTE12.csv	11415	5	2021-01-01	2022-04-21 14:00:00	0	0	0.0
3	DATOSCLIENTE13.csv	14000	5	2021-01-01	2022-08-07 07:00:00	0	0	0.0
4	DATOSCLIENTE14.csv	14000	5	2021-01-01	2022-08-07 07:00:00	0	0	0.0
5	DATOSCLIENTE15.csv	14000	5	2021-01-01	2022-08-07 07:00:00	0	0	0.0
6	DATOSCLIENTE16.csv	19500	5	2021-01-01	2023-03-24 11:00:00	0	0	0.0
7	DATOSCLIENTE17.csv	19500	5	2021-01-01	2023-03-24 11:00:00	0	34	0.0
8	DATOSCLIENTE18.csv	19500	5	2021-01-01	2023-03-24 11:00:00	0	256	0.0
9	DATOSCLIENTE19.csv	19500	5	2021-01-01	2023-03-24 11:00:00	0	0	0.0
10	DATOSCLIENTE20.csv	19500	5	2021-01-01	2023-03-24 11:00:00	0	215	0.0

***Tabla 1**. Análisis exploratorio de los archivos para encontrar diferencias*

Transformaciones

De lo anterior se identifica la primera transformación que consiste en tratar los valores negativos. Como el porcentaje es muy bajo, y el rango de valores es bajo, se decide reemplazar por 0, para mantener la continuidad de la serie.

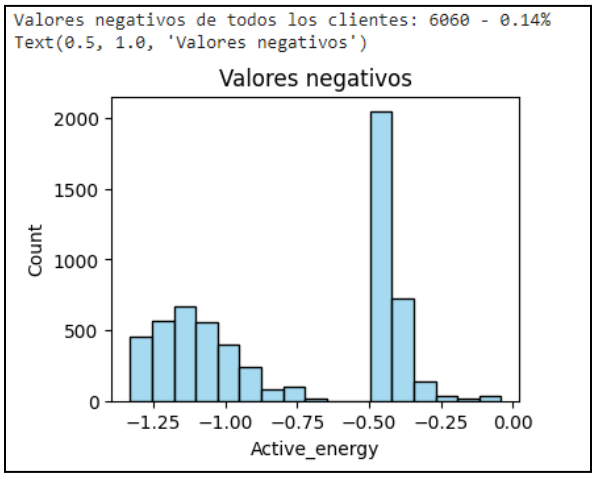

***Figura 1**. Histograma de valores negativos en la variable Active_energy*

Se extrajo el ID de cada cliente y se agregó a un dataset con toda la información de consumo de los clientes, dando como resultado un *dataframe* como se muestra en la siguiente figura:

El tamaño del dataset es: (351020, 6)

	Fecha	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC	Cliente_ID
0	2021-01-01 00:00:00	0.357841	0.282788	455.139171	510.561002	1
1	2021-01-01 01:00:00	0.372264	0.431377	469.978787	469.917178	1
2	2021-01-01 02:00:00	1.044687	0.338626	468.721120	546.949147	1
3	2021-01-01 03:00:00	0.566425	0.495791	452.329255	444.122989	1
4	2021-01-01 04:00:00	1.080556	0.472018	513.477596	535.463719	1

***Tabla 2**. Dataframe con datos de consumo de energía de clientes de Electrodunas*

Normalización de los datos

Los datos de consumo de los clientes tienen un rango muy amplio entre ellos y eso dificulta las comparaciones y la identificación visual de anomalías. Por lo anterior, se decidió normalizar los datos para que las variaciones de los consumos estén en una escala similar. A continuación se muestra la serie de energía activa para dos clientes sin normalizar (izquierda) y normalizada (derecha). Para el análisis descriptivo se utilizarán los datos normalizados.

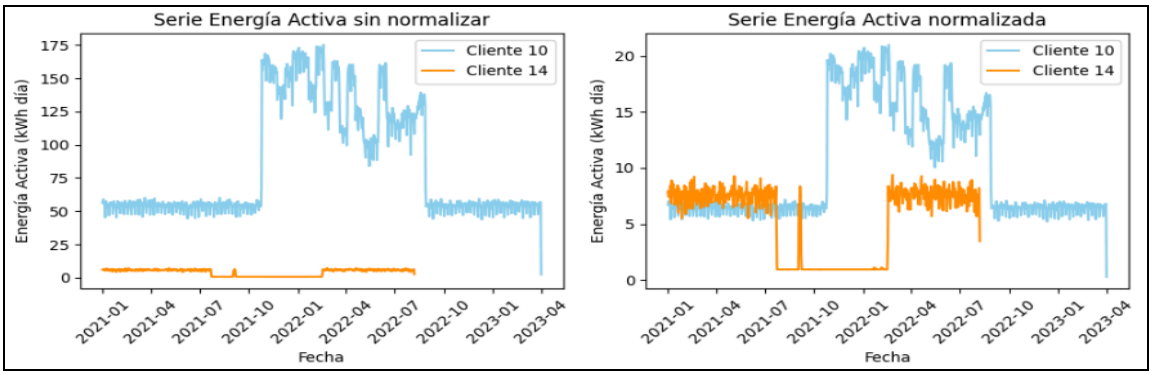

Figura 2. *Comparación de la serie Energía activa para 2 clientes con normalización de datos*

Escala de las series temporales

Antes de iniciar el análisis descriptivo es muy importante resaltar la relevancia de la escala de los datos en los diferentes análisis que se van a realizar. Originalmente los datos de las cuatro variables están en kilovatio hora (KWh) pero estos se pueden agrupar en diferentes escalas para identificar diferentes patrones. A manera de ejemplo se gráfica el consumo de energía activa del Cliente 13 en KWh, KWh promedio día, KWh promedio semana y KWh promedio mes.

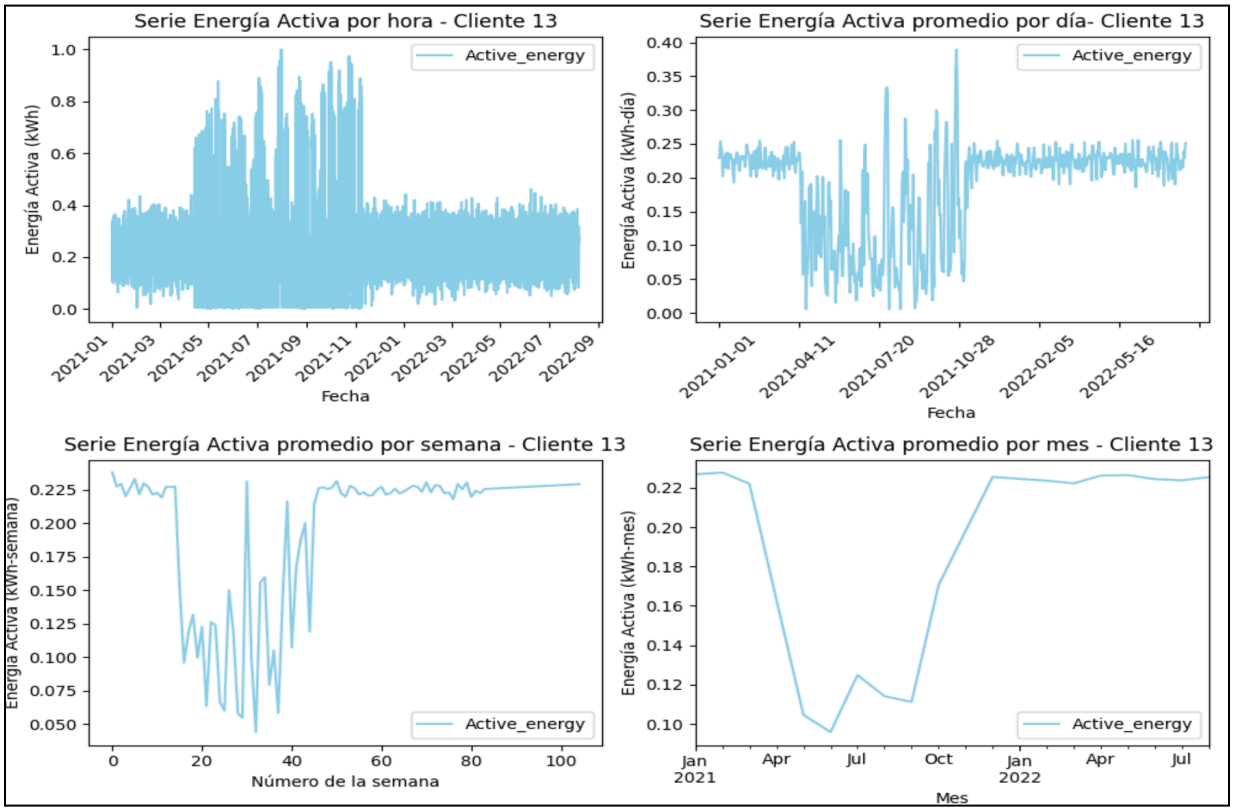

Figura 3. *Series de Energía activa para un mismo cliente en diferente escala de tiempo*

Lo anterior se tendrá en cuenta en el análisis descriptivo y el análisis para la detección de anomalías.

Estadísticas descriptivas

	Fecha	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC	Cliente_ID
count	351020	351020.000000	351020.000000	351020.000000	351020.000000	351020.000000
mean	2021-11-29 16:52:17.769927424	1.428845	0.921105	1461.153534	1460.008100	15.831343
min	2021-01-01 00:00:00	0.000000	0.000000	0.031000	0.031000	1.000000
25%	2021-06-02 08:00:00	0.231000	0.108911	828.115789	823.796854	10.000000
50%	2021-11-01 16:00:00	0.687316	0.346211	1776.599493	1769.524471	17.000000
75%	2022-04-27 05:15:00	1.905677	1.216852	2044.159520	2044.529061	22.000000
max	2023-04-01 00:00:00	14.622644	11.135141	4004.051758	3931.262836	30.000000
std	NaN	1.768029	1.256586	669.931638	669.976550	8.479808

***Tabla 3**. Estadísticas descriptivas del dataframe completo con todos los clientes*

De lo anterior se concluye que la energía activa y reactiva tienen una dispersión alta y la mediana se ubica en el tercer cuartil. Para el voltaje FA y voltaje FC, en el escenario teórico, las estadísticas deberían ser iguales porque es la medición de la fase 1 y fase 2, las cuales deben ser idénticas; no obstante se puede ver que no coinciden los valores máximos exactamente.

Correlación entre variables

A nivel general, hay una fuerte correlación entre las variables de voltaje, lo que sugiere una calidad alta de los voltajes de la red que se entrega a los clientes ya que entre fases se espera que haya el mismo comportamiento en magnitud y dirección.

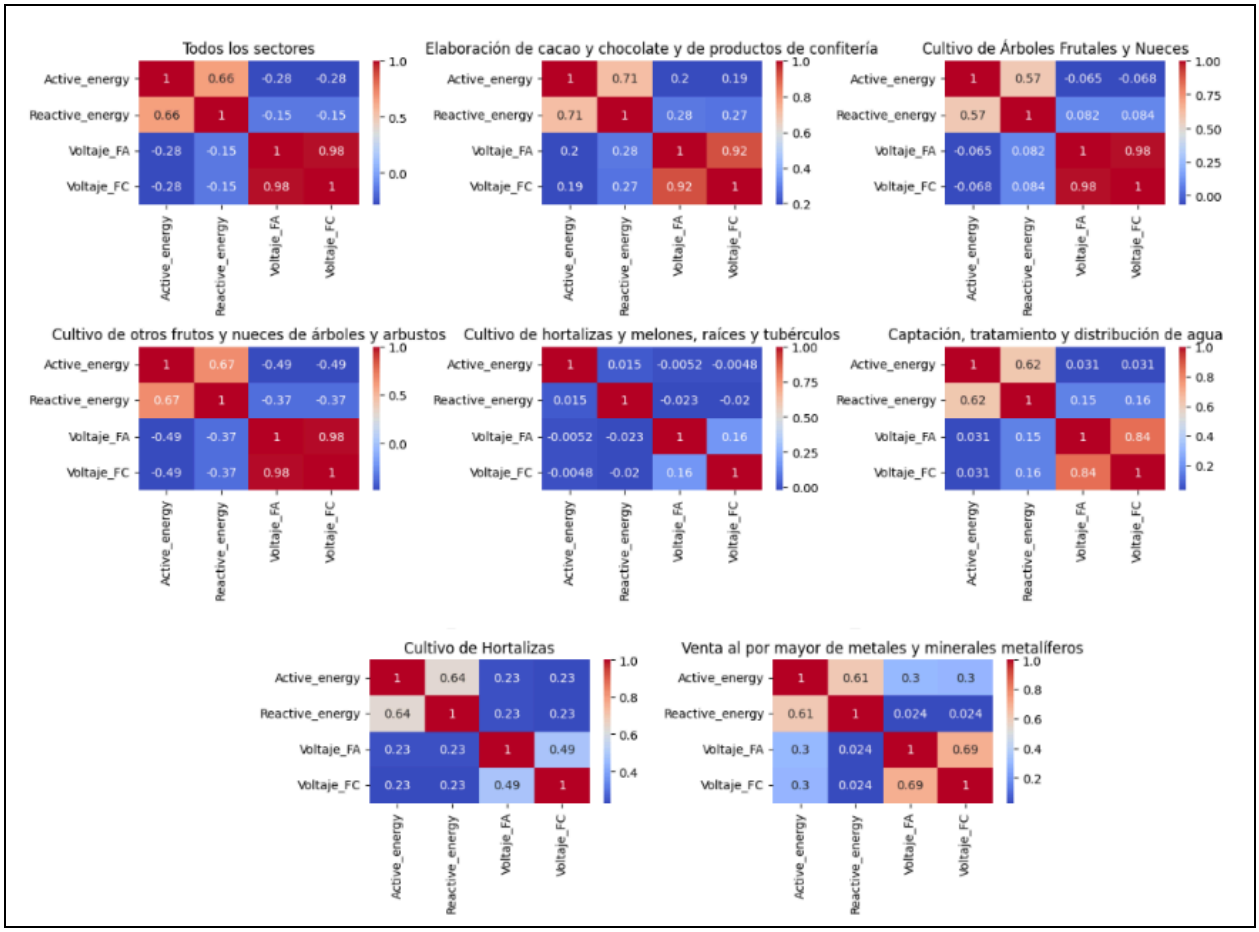

***Figura 4**. Correlaciones entre las cuatro variables por sector económico*

No obstante, se identifican los siguientes sectores donde la correlación de voltajes es muy baja, lo que sugiere que en estos sectores se encontraran anomalías respecto al tema técnico.

- Cultivo de hortalizas y melones, raíces y tubérculos
- Cultivo de Hortalizas
- Captación, tratamiento y distribución de agua
- Venta al por mayor de metales y minerales metalíferos

Análisis descriptivo

Sectores económicos

Los sectores económicos a los cuales pertenecen los clientes son los siguientes:

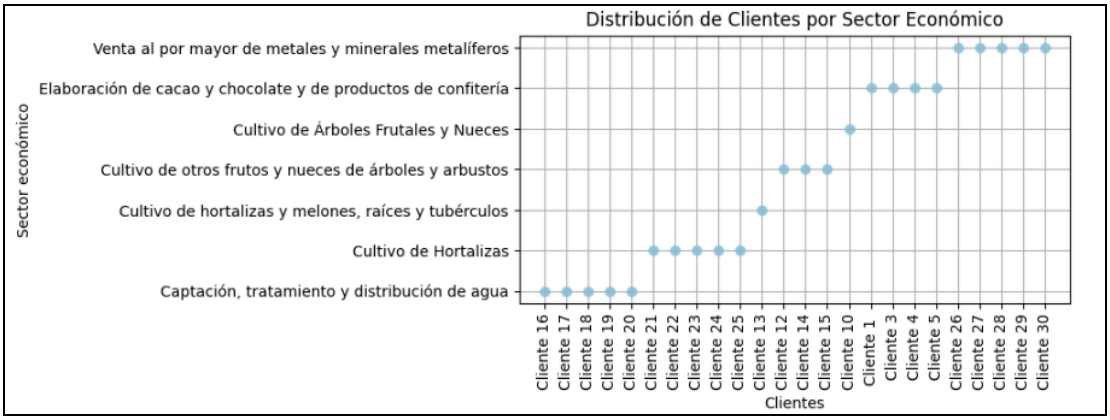

***Figura 5**. Distribución de clientes por sector económico*

El consumo por cada sector acumulado se muestra a continuación. Teniendo en cuenta que las series son diferentes, se analizó el periodo en el que todos los clientes tienen registro de consumo de energía activa.

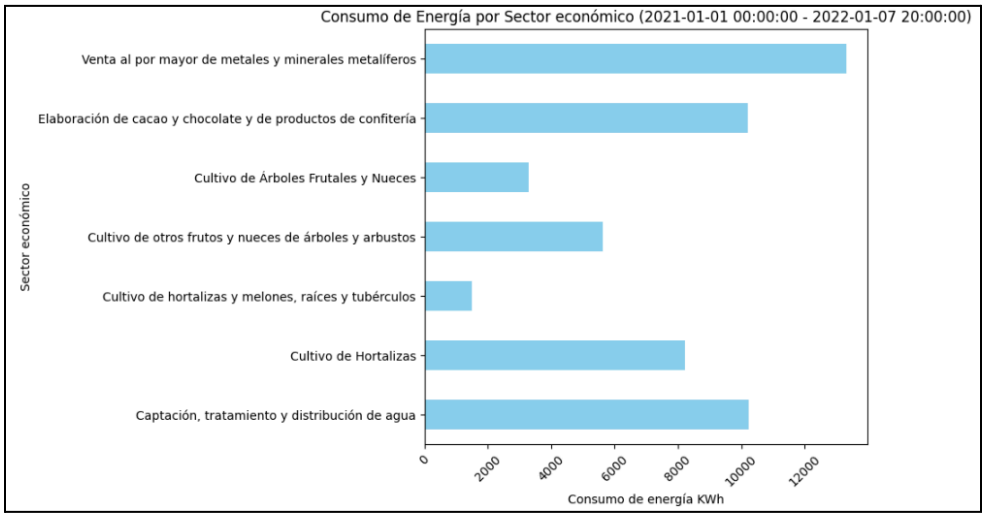

***Figura 6**. Consumo total energía activa entre 2021-01-01 y 2022-01-07 por sector económico*

Cuando se hace un diagrama boxplot de la energía activa consumida en un día, se puede observar cuáles sectores tienen datos atípicos en los dos extremos. Adicionalmente, este gráfico permite ubicar información como el promedio, mediana y los cuartiles de cada sector. Igualmente, se puede identificar el de cada sector y la diferencia de consumos promedio entre los mismos sectores. El consumo promedio diario del sector Cultivo de hortalizas y melones, raíces y tubérculos es cerca de una séptima parte del consumo del sector Venta al por mayor de metales y minerales metalíferos.

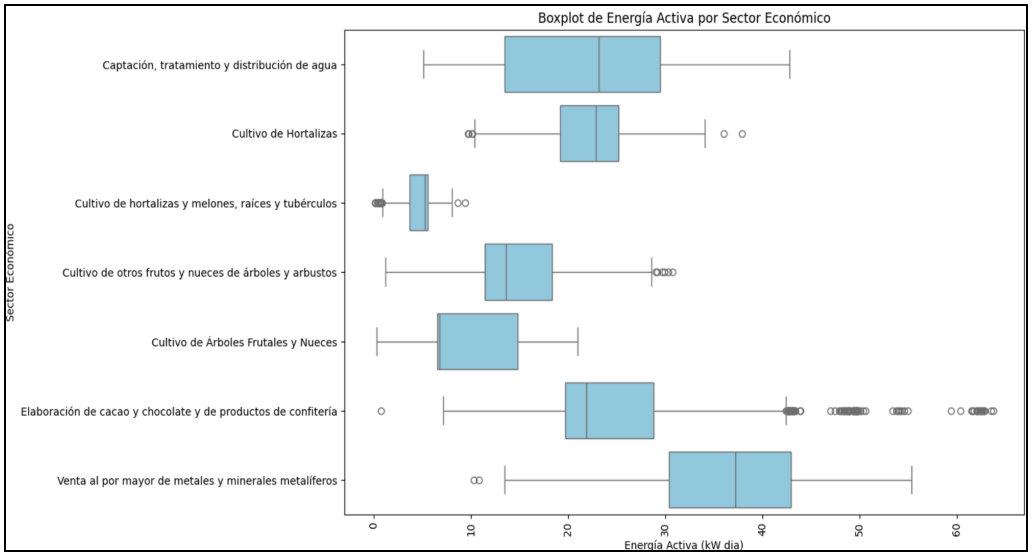

***Figura 7**. Box-plot de la Energia activa por sector económico*

La serie del consumo de energía activa en KWh promedio por día de cada sector se muestra a continuación:

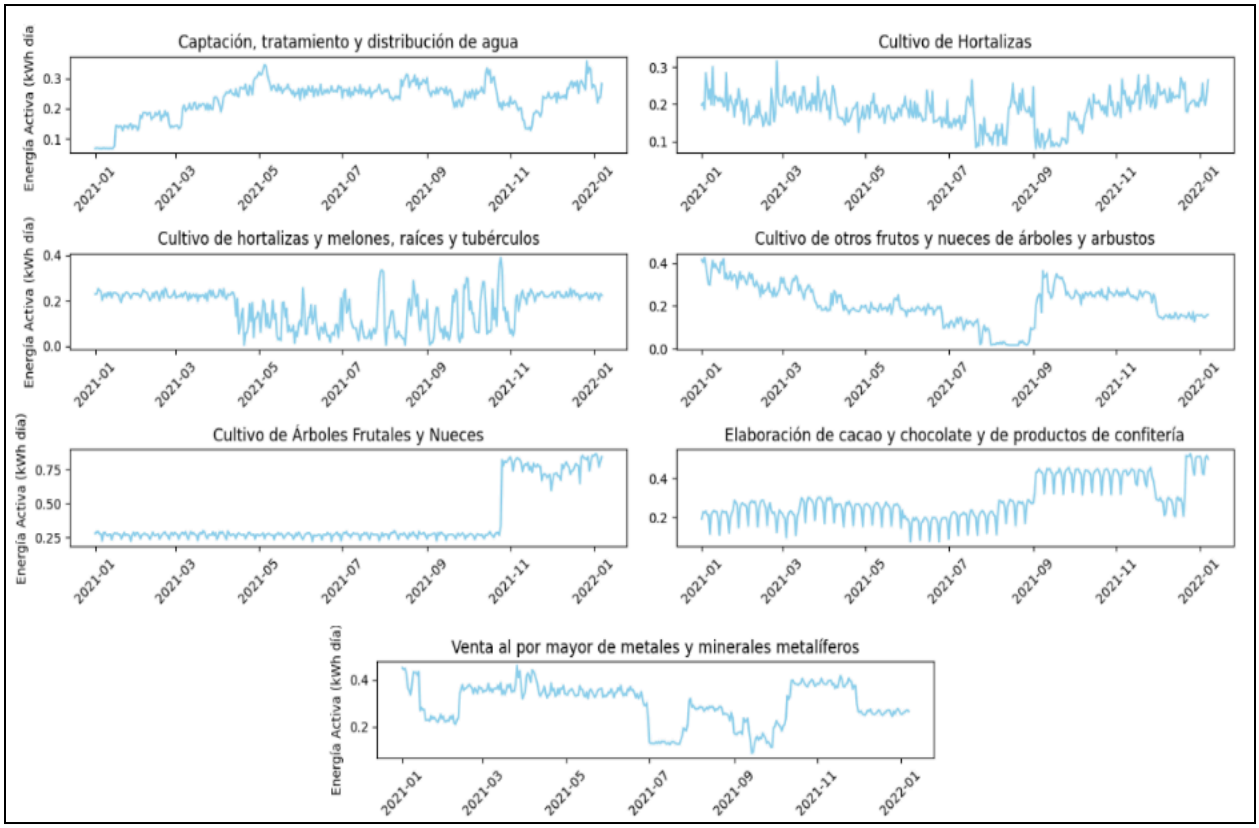

***Figura 8**. Series de Energía activa para cada sector económico*

De las series se puede concluir que no hay un patrón y/o estacionalidad que compartan entre ellas.

Clientes

Los boxplot de energía activa (kWh) y energía reactiva (kVarh) muestran variabilidad en los consumos de los diferentes clientes, esto asociado principalmente a la diferencia de sectores económicos a los que pertenecen y la capacidad instalada de producción, se logra identificar de manera visual que hay extremos en los datos que pueden dar indicios a eventos de consumos anómalos.

Para la energía activa, los clientes 5 y 10 tienen una variación más amplia de los datos entre su primer y tercer cuartil, también los consumos más altos, la mediana sugiere que el primero tiene un comportamiento simétrico. Para la energía reactiva: los clientes 5 y 20 tienen las más amplias variaciones en este ítem y consumos más altos, las medianas sugieren que su comportamiento es simétrico, es de notar que el comportamiento de esta lectura de algunos clientes es superior a la detectada en la energía activa, por lo que hay sospecha de anomalías en el consumo.

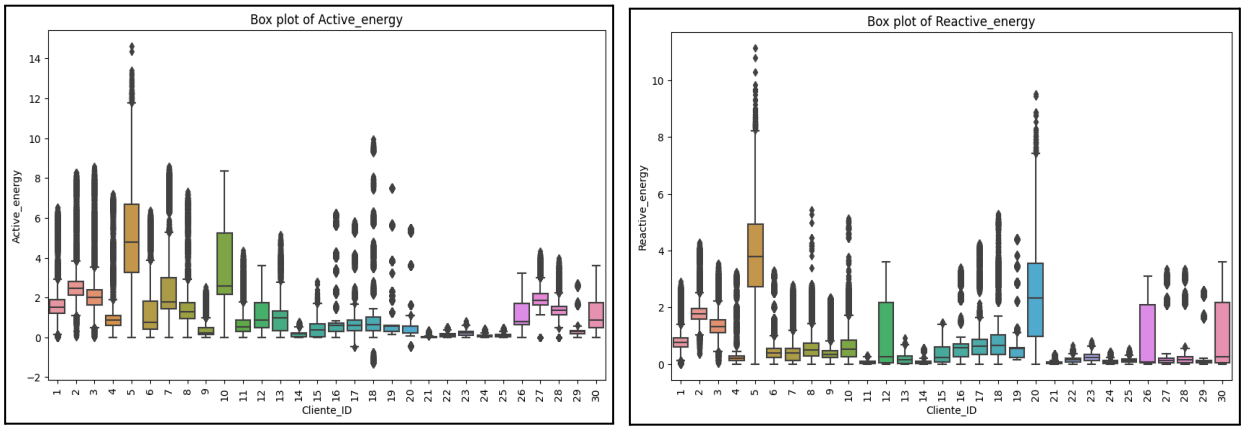

***Figura 9**. Box-plot con la comparación de la energía activa (izq.) y reactiva (der.) por cliente*

Respecto a los voltajes FA y FC (V), estos parecen ser estables para la mayoría de los clientes, sin embargo, se nota una variabilidad significativa en algunos de ellos con valores atípicos, lo que puede indicar problemas de calidad de energía que pueden afectar los equipos de los clientes. Además, clientes con valores extremos muy altos o muy bajos deben ser analizados, ya que pueden experimentar condiciones subestándar de operación o desconexiones, lo cual puede ser anómalo en su consumo.

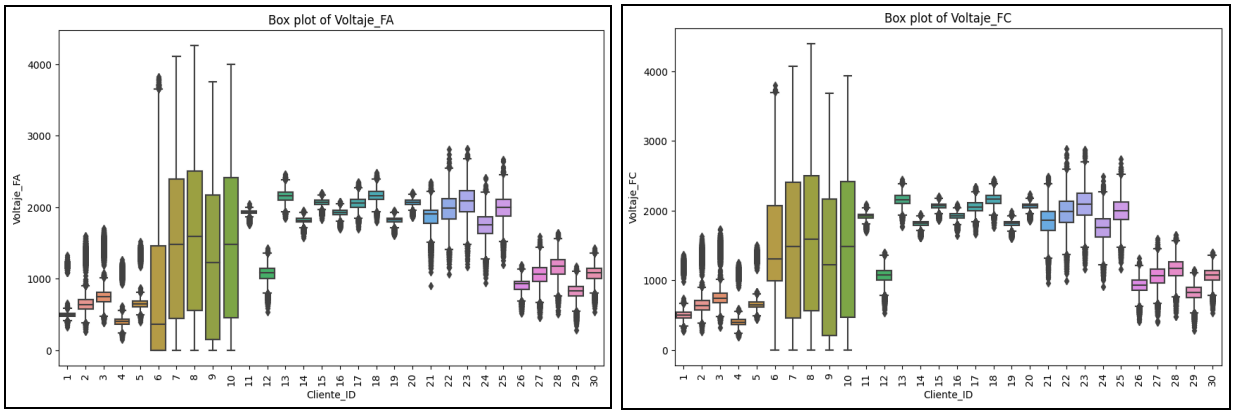

***Figura 10**. Box-plot con la comparación del voltaje FA (izq.) y FC (der.) por cliente*

Dado lo anterior y para comprender mejor la forma en la que se comportan los datos de consumo y lecturas de voltaje de cada cliente, se realizaron gráficas con el acumulado mes a mes del consumo de energía activa y reactiva de cada uno de los clientes, se ajustó un límite correspondiente al 30% de la energía activa para ver gráficamente si alguno de los clientes presenta consumos anómalos. Como se ve en la siguiente gráfica, la energía reactiva es superior en varios meses al límite establecido en los contratos de facturación.

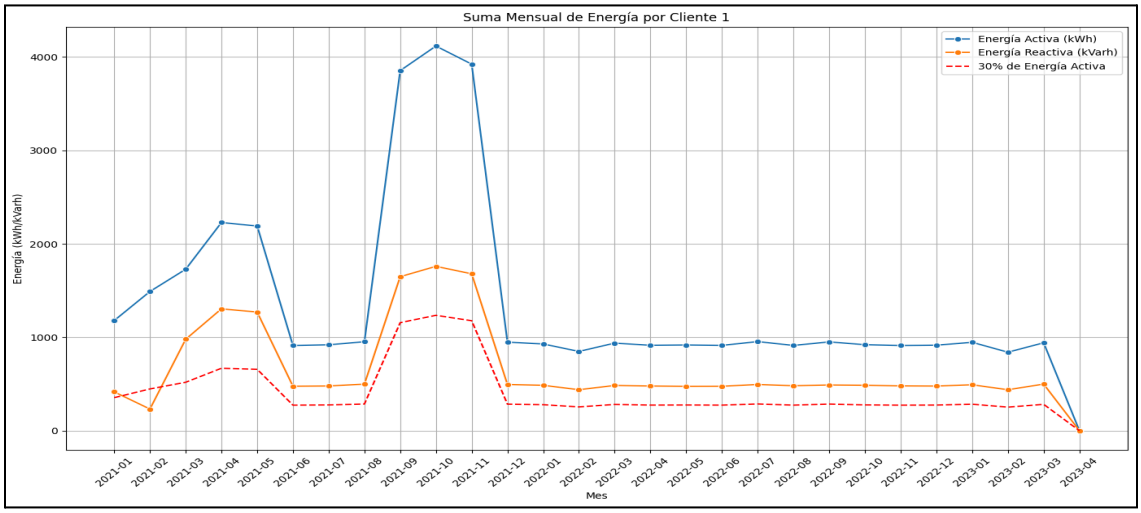

***Figura 11**. Comparación de la energía activa y reactiva con el límite del 30% para la energía reactiva*

Nota: De sobrecostos en la facturación por exceder el 30% de la energía activa con la energía reactiva se tratará en la sección de los modelos con mayor detalle.

De manera similar, se realizó agrupación mes a mes con los voltajes de fase medidos en cada cliente para asegurarnos que los voltajes FA y FC en cada medidor fuera lo más parejo posible. Sin embargo, se encuentra que hay oscilaciones en los niveles de tensión, lo cuál es

extraño no es común pero no está ligado con el consumo del cliente y por ende no se tendrá en cuenta para el modelo de predicción.

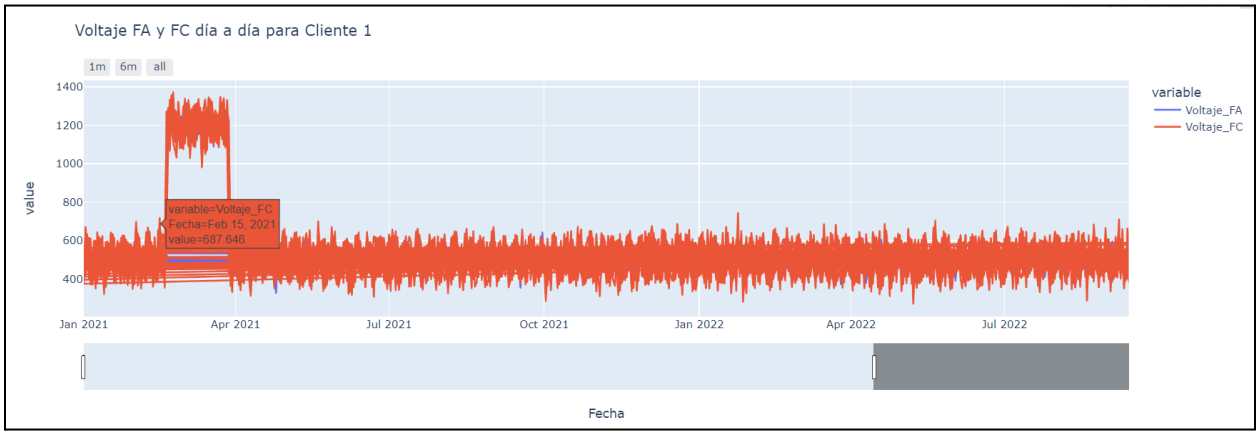

***Figura 12**. Comparación de los voltajes FA y FC para identificar discrepancias entre ellos*

Al analizar si los días festivos tenían influencia sobre los consumos de energía activa para los clientes, contrario a la lógica, hay ciertos patrones que indican que el consumo promedio de algunos clientes aumenta cuando los días son festivos. Por otro lado, cuando se intenta asociar el consumo de energía con la temporada climática se puede observar que aumenta la demanda levemente y puede estar asociado a que en estas temporadas los sistemas de calefacción pueden ser empleados. Sin embargo, no es una diferencia apreciable de manera notoria por ello no es concluyente.

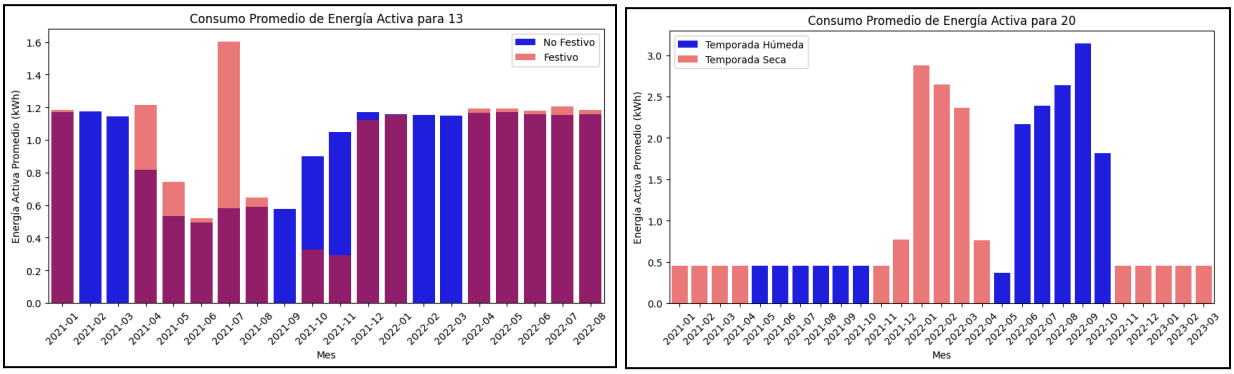

***Figura 13**. Histograma de energía activa para el Cliente 13 (izq.) y Cliente 20 (der.) diferenciando la temporada y los festivos*

Por último, se identificó que los valores más altos de consumo de energía activa se dan en las primeras horas de la mañana y los más bajos en la noche y madrugada, la hora pico de consumo de energía activa es a las 8 am. Se observó un comportamiento de consumo de energía activa similar a lo largo de las semanas, sin embargo los miércoles y los viernes hay un pico leve de esta variable.

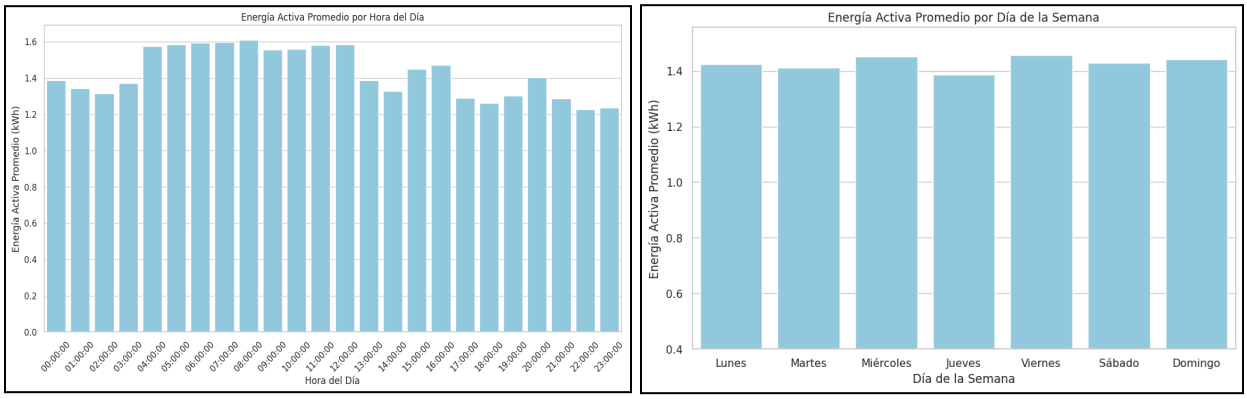

***Figura 14**. Promedio de consumo de energía activa por horas del día y por días a la semana*

ANEXOS

1. Notebook con el análisis exploratorio y descriptivo de la información de los clientes de ElectroDunas:

Nombre: 1. *Análisis exploratorio y descriptivo.* *Link*