Herramienta de Análisis Descriptivo y Detección de Patrones y Anomalías de Consumo para Electrodunas

Reporte técnico de selección de modelos, variables y parámetros

Dando continuidad al análisis exploratorio, este reporte se centrará en el análisis de la energía activa para la identificación de anomalías en los consumos eléctricos de los clientes de Electrodunas, a pesar de contar con otras variables como la energía reactiva y los voltajes, se justifica por varios motivos cruciales. Primero, la energía activa es la que realmente se convierte en trabajo útil (luz, calor, movimiento), siendo por tanto el principal indicador de consumo eficiente y efectivo. Un uso eficiente de la energía activa no solo garantiza una reducción de costos, sino también minimiza el impacto ambiental y mejora la sostenibilidad del sistema.

Por otro lado, aunque la energía reactiva es necesaria para el funcionamiento de ciertos equipos, su gestión inadecuada puede resultar en ineficiencias y penalizaciones económicas, lo que repercute negativamente en la eficiencia general del sistema eléctrico. Además, un buen factor de potencia, que indica una alta eficiencia en el uso de la energía activa respecto a la energía total (potencia aparente), es esencial para reducir el consumo energético y mejorar la eficiencia operativa de las empresas.

Concentrarse en la energía activa permite implementar mejoras directas en el uso de la energía y optimizar las estrategias de gestión energética, lo cual es vital en un contexto donde las fuentes de energía renovable están en crecimiento y la variabilidad en la generación de energía se convierte en un desafío adicional. Esto subraya la importancia de un monitoreo y control precisos sobre el componente de energía que efectivamente realiza trabajo útil, asegurando así la estabilidad y eficiencia del sistema en su conjunto.

En este análisis una anomalía en el consumo de energía se puede entender de dos maneras:

1. Anomalía por incumplimiento de reglas de negociación como clientes no regulados.
Este tipo de anomalías está asociada a la '*Cláusula décima - Facturación y pago*' para clientes no regulados en seguimiento a la Energía Reactiva frente a la Energía Activa. Se presenta cuando:

"*Los consumos de energía inductiva y capacitiva serán medidos en el Punto del Suministro y se facturará este cargo cuando la Energía Reactiva Registrada en el mes de facturación sea mayor al treinta por ciento (30%) de la Energía Activa mensual. El cargo se aplicará a la*

diferencia entre la Energía Reactiva y el treinta por ciento (30%) de la Energía Activa registradas durante el período de facturación mensual. Se aplicará el precio vigente para las tarifas de distribución que publica el OSINERGMIN"

Nota: Este párrafo se toma de los contratos para clientes no regulados publicados abiertamente por Electrodunas. (ejemplo de contrato - cliente no regulado)

La energía reactiva es fundamental en los sistemas de distribución de energía eléctrica, principalmente porque influye en la eficiencia del sistema y en la calidad de la energía entregada a los consumidores. Mantener la energía reactiva en valores inferiores al 30% de la energía activa es importante por varias razones:

a. *Eficiencia Energética:* La energía reactiva, a diferencia de la energía activa, no realiza trabajo real pero es necesaria para el funcionamiento adecuado de equipos que operan con corrientes y campos magnéticos, como motores y transformadores. Si la proporción de energía reactiva es demasiado alta respecto a la energía activa, indica que el sistema no está siendo eficiente. Esto se debe a que se está transportando energía que no contribuye directamente al trabajo útil, lo cual genera pérdidas en forma de calor y, por tanto, un aumento en los costos de energía.
b. *Calidad del Voltaje*: Un alto nivel de energía reactiva puede provocar caídas de voltaje a lo largo de la red eléctrica. Esto afecta la calidad del voltaje recibido por los usuarios finales y puede causar problemas en equipos sensibles y máquinas que requieren un suministro de voltaje estable para operar eficientemente.
c. *Capacidad de Transmisión:* Cuando hay un exceso de energía reactiva, se requiere más corriente para transmitir la misma cantidad de energía activa. Esto puede llevar a una sobrecarga en la capacidad de transmisión de las líneas eléctricas y equipos asociados, incrementando el riesgo de fallos y reduciendo la vida útil de la infraestructura eléctrica.
d. *Costos de Transmisión y Distribución:* Electrodunas pueden incurrir en costos adicionales para gestionar y compensar el exceso de energía reactiva en la red. Esto incluye la instalación de equipos como bancos de condensadores y reguladores de voltaje, que son necesarios para mejorar la eficiencia del sistema y asegurar que el voltaje se mantenga dentro de los parámetros deseados.

Debido a la importancia que tiene el monitoreo y control de la Energía Reactiva (ER) con respecto a la Energía Activa (EA), se procede a realizar la inspección cliente a cliente e identificar estas anomalías en los consumos:

Proceso de Detección y Clasificación de Anomalías

Para detectar y clasificar estas anomalías, se llevaron a cabo los siguientes pasos:

1. Sumar el consumo de la energía activa y la reactiva en cada mes.
2. Calcular el porcentaje que la corriente reactiva representa sobre el total de la corriente activa.
3. Crear una columna denominada 'Anomalía', la cual se cataloga como 1 (presencia de anomalía) si la relación ER/EA es superior al 30%.

Visualización de observaciones Anómalas

Realizando la visualización de las anomalías por cliente se identifica que algunos conservan la calidad de la energía en parámetros normales mientras que otros constantemente incumplen este acuerdo por ejemplificar los hallazgos se ilustra a continuación el cliente 10 (Figura derecha) y el cliente 1 (Figura izquierda).

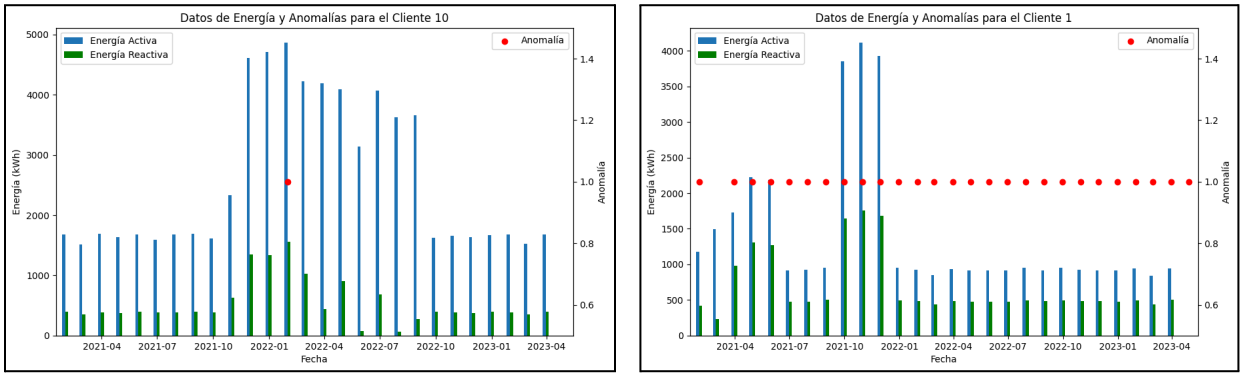


Figura 1*. Anomalías por consumo de ER/EA en los clientes no regulados - Cliente 10 (Izq) y Cliente 1 (Der)*

2. Anomalía en la tendencia:

Una anomalía en la tendencia en este estudio se refiere a cualquier observación de consumo de energía activa que exhiba una desviación significativa respecto a lo que se considera un patrón normal de consumo. Esta definición es válida para diversas metodologías de detección, incluyendo z-score y técnicas avanzadas de aprendizaje automático no supervisado como Isolation Forest, Robust Covariance, Local Outlier Factor (LOF) y PCA. Para esto se consideran dos direcciones principales de desviación:

Anomalías por sobre consumo: Se identifica cuando el consumo de energía activa de una observación supera significativamente el promedio histórico de energía activa.

Anomalías por Subconsumo: Se identifica cuando el consumo de energía activa de una observación es significativamente menor que el promedio histórico.

El proceso de detección de anomalías en Electrodunas se aborda mediante una estrategia de experimentación progresiva, comenzando con el z-score por su claridad y capacidad para destacar desviaciones significativas en el consumo de energía activa. Esta técnica brinda una comprensión inicial de los patrones de consumo.

Se exploran técnicas avanzadas como Isolation Forest, Robust Covariance, Local Outlier Factor (LOF) y PCA, aplicadas a datos horarios, para una detección más precisa de comportamientos atípicos. Además, se implementan análisis con promedios diarios, que no generan mucha información relevante puesto que el volumen de los datos no es alto para generar información comparativa en el tiempo. Con lo anterior, se generan visualizaciones resumidas que facilitan la identificación de tendencias anómalas, útiles para análisis estratégicos a nivel operativo. En la siguiente tabla se resume cada una de las técnicas utilizadas, sus ventajas, desventajas y los supuestos de cada modelo de manera general.

Modelo	Ventajas	Desventajas	Supuestos
Z-score	- Simple y fácil de implementar - Eficiente para detectar outliers globales	- No es efectivo para detectar outliers locales - Sensible a outliers extremos - Requiere conocimiento de la media y la desviación estándar de los datos	- Los datos tienen una distribución normal - Las observaciones son independientes
Isolation Forest	- Rápido y eficiente - Robusto a outliers - Fácil de implementar	- Menos preciso que otros métodos - No proporciona información sobre la estructura de los datos	- Las observaciones son independientes - Los datos tienen una distribución arbitraria
Robust Covariance	- Robusto a outliers - Proporciona información sobre la estructura de los datos	- Más lento que otros métodos - Puede ser sensible a la elección de la ventana de observación	- Las observaciones son independientes - Los datos tienen una distribución gaussiana
Local Outlier Factor (LOF)	- Efectivo para detectar outliers locales	- Más lento que otros métodos - Puede ser sensible a la elección del parámetro k (determina la sensibilidad del algoritmo a los outliers locales)	- Las observaciones son independientes - Los datos tienen una distribución arbitraria
PCA (Análisis de Componentes Principales)	- Reduce la dimensionalidad de los datos - Revela la estructura subyacente de los datos	- No es efectivo para detectar outliers locales - No es robusto a outliers	- Las observaciones son independientes - Los datos tienen una distribución gaussiana

Tabla 1. *Revisión general de los modelos empleados en el desarrollo del estudio.*

Como se puede ver en la **Tabla 1,** para algunos modelos implementados el supuesto más relevante es el de la distribución normal de los datos. Sin embargo, como se tienen datos anómalos en las tendencias analizadas se pueden presentar el incumplimiento de este supuesto. En este contexto se realizaron dos pruebas estadísticas una conocida como Shapiro-Wilk y otra como D'Agnostino's. Los resultados se pueden visualizar a continuación para los primeros 4 clientes:

Cliente	Shapiro-Wilk	D'Agostino's K^2
1	Stat:0.7821, P:0.0	5149.62
3	Stat:0.7601, P:0.0	7529.34
4	Stat:0.6363, P:0.0	9261.80
5	Stat:0.9893, P:0.0	409.44

Tabla 2. *Revisión de distribución por test de normalidad a los clientes para revisar el efecto de las anomalías en los consumos normales de los clientes.*

Para la prueba de Shapiro-Wilk: Todos los clientes tienen estadísticos significativamente menores a 1, indicando fuertes desviaciones de la normalidad, y los p-valores de 0.0 confirman que estas diferencias son estadísticamente significativas.

Para la prueba de D'Agostino's: Los estadísticos son extremadamente altos, lo que refleja asimetría y/o exceso de curtosis en las distribuciones, y nuevamente, los p-valores de 0.0 indican que estas observaciones son altamente significativas.

Peso a esto se realizaron transformaciones de los datos de Energía Activa para buscar el cumplimiento de este supuesto a través de dos métodos una logaritmica y otra de Box Cox, a manera de ilustración se muestran los primeros 5 clientes:

Cliente	Transformación	Shapiro-Wilk	D'Agostino's K^2	Conclusión
1	Original	Stat:0.7821, P:0.0	5149.62	No Normal
1	Logarithmic	Stat:0.8757, P:0.0	5876.34	No Normal
1	Box-Cox	Stat:0.9394, P:0.0	101.97	No Normal
3	Original	Stat:0.7601, P:0.0	7529.34	No Normal
3	Logarithmic	Stat:0.9023, P:0.0	3999.02	No Normal
3	Box-Cox	Stat:0.9392, P:0.0	609.15	No Normal
4	Original	Stat:0.6363, P:0.0	9261.80	No Normal
4	Logarithmic	Stat:0.9080, P:0.0	2313.43	No Normal
4	Box-Cox	Stat:0.9479, P:0.0	19.43	No Normal
5	Original	Stat:0.9893, P:0.0	409.44	No Normal
5	Logarithmic	Stat:0.8553, P:0.0	8567.17	No Normal
5	Box-Cox	Stat:0.9958, P:0.0	165.07	No Normal

Tabla 3. *Revisión de distribución por test de normalidad a los clientes para revisar el efecto de las transformaciones.*

A pesar de los intentos de mitigar las desviaciones de la normalidad mediante transformaciones logarítmicas y de Box-Cox, los resultados no han sido satisfactorios. Las pruebas de normalidad post-transformación continúan mostrando valores significativos de p para la prueba de Shapiro-Wilk, lo que indica que las distribuciones aún difieren significativamente de una normal. Además, los estadísticos de la prueba de D'Agostino's siguen siendo extremadamente altos después de las transformaciones, reflejando persistentes problemas de asimetría y exceso de curtosis en los datos.

Las anomalías desempeñan un papel crucial en la evaluación de la normalidad y en la eficacia de las transformaciones aplicadas a los datos. Debido a su naturaleza extrema, estas observaciones anómalas tienden a distorsionar significativamente la media y la mediana, contribuyendo a distribuciones asimétricas o con colas pesadas. Esta distorsión es un desafío particular porque desvía los datos de la forma de campana que caracteriza a una distribución normal.

Además, los valores extremos influyen en la variabilidad de los datos, aumentando tanto la varianza como la desviación estándar. Esto resulta en estimaciones poco fiables de estos parámetros, que son fundamentales para muchos modelos estadísticos que presuponen homogeneidad en la varianza y normalidad en la distribución. Por ejemplo, cuando se aplican pruebas estadísticas clásicas que asumen una distribución normal, como la prueba t para comparar medias, la presencia de outliers puede llevar a conclusiones erróneas o a la identificación incorrecta de efectos significativos.

Por lo tanto, en presencia de datos anómalos, las transformaciones estándar como las logarítmicas o Box-Cox, aunque útiles, pueden no ser suficientes para corregir completamente las irregularidades en la distribución de los datos. Esto se debe a que estas transformaciones intentan modificar la escala o la forma de la distribución, pero no pueden eliminar el impacto fundamental de las anomalías extremas. Estos resultados sugieren que las transformaciones empleadas no han logrado corregir las características subyacentes de los datos que impiden una distribución normal, destacando la necesidad de considerar métodos alternativos de análisis que no dependan de la normalidad, como técnicas no paramétricas o modelos estadísticos robustos. Sin embargo, se continuó con la implementación de todos los modelos y se realiza un análisis final para la toma de decisión que involucra el mejor modelo para colocar en producción en la identificación de anomalías.

A continuación, se aborda con detalle la implementación de los diferentes métodos para la identificación de dichas anomalías:

2.1 Identificación de Gravedad y Niveles de Criticidad:

En el análisis de consumo energético para Electro Dunas, es fundamental no solo detectar anomalías, sino también clasificar su nivel de criticidad para priorizar intervenciones y monitoreo continuo. Con este fin, se ha implementado un sistema de clasificación basado en el Z score, que permite diferenciar las anomalías detectadas en términos de su gravedad.

2.1.1 Definición de Umbrales de Criticidad

El Z score, que como se comentó previamente, mide cuántas desviaciones estándar una observación se encuentra dentro del promedio, se utiliza como criterio principal para clasificar la criticidad de las anomalías en el consumo de energía. Se establecen tres niveles de criticidad:

Baja: Son las anomalías detectadas por Z-score pero que presentan una puntuación de entre -2,5 a -2,25 y 2,25 a 2,5 desviaciones estándar.

Moderada: Anomalías que presentan una puntuación Z entre - 2,75 a -2,5 y 2,5 a 2,75 desviaciones estándar. Este rango indica que la observación es atípica y se aleja del comportamiento estándar esperado, pero aún no representa una amenaza crítica.

Crítica: Anomalías con una puntuación Z-score mayor a 2,75 y menor a -2,75. Estas se consideran críticas debido a su significativa desviación respecto al promedio, indicando situaciones que podrían implicar pérdidas no técnicas severas, fraudes o fallos en el sistema que requieren una acción inmediata.

A continuación se ilustra un ejemplo de la criticidad de las anomalías en dos de los clientes:

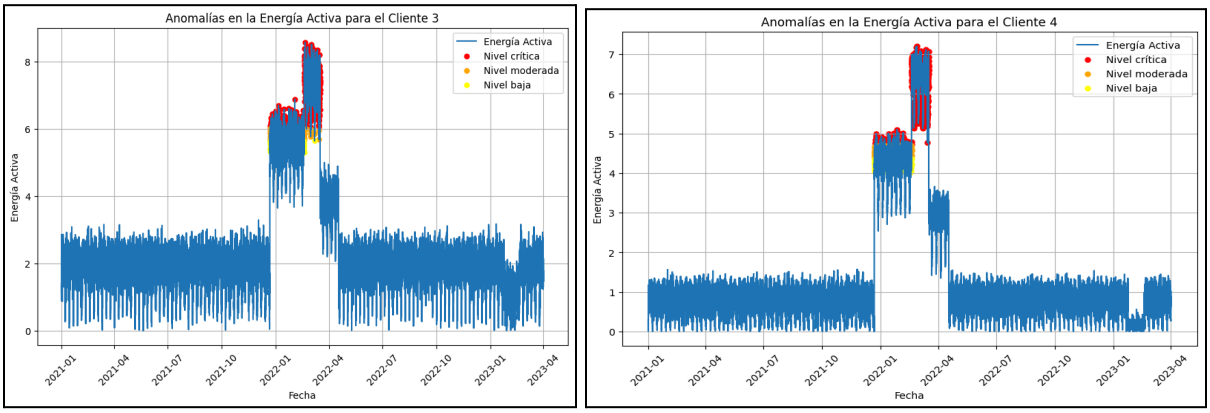


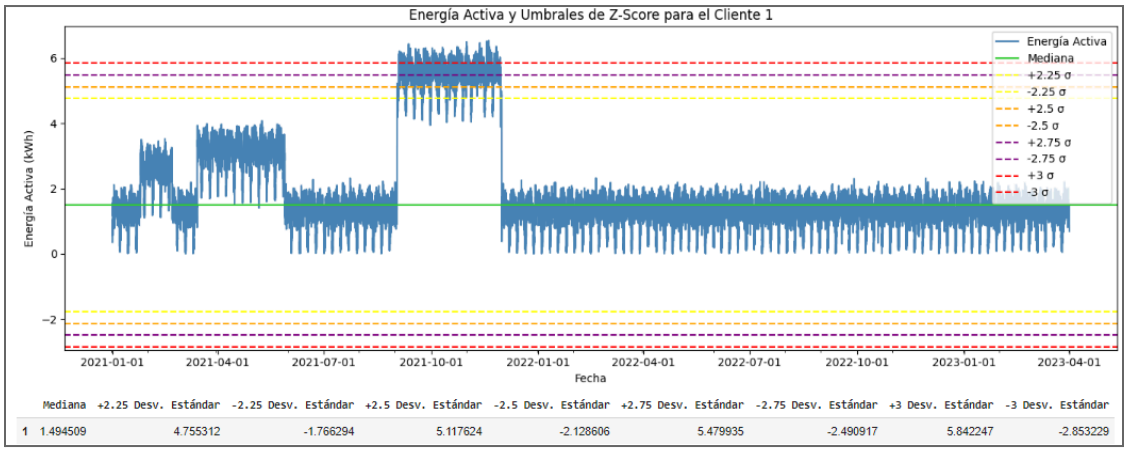
***Figura 2**. Nivel de criticidad en las anomalías de los clientes Cliente 3 (Izq) y Cliente 4 (Der)*

2.2. Primera aproximación a la detección de las anomalías de consumo (Z-Score):

El z-score, también conocido como puntuación estándar, es una herramienta estadística que mide la distancia en desviaciones estándar de un punto de datos respecto al promedio o a la mediana de un conjunto de datos. Esta metodología es particularmente pertinente en la detección de anomalías porque permite identificar valores que son inusualmente altos o bajos en comparación con la norma establecida por el conjunto de datos históricos. En el contexto de Electrodunas, el uso del z-score facilita la identificación rápida de consumos de energía que se desvían significativamente del comportamiento típico del cliente, lo cual puede ser indicativo de errores en la medición, fraude, o averías en el sistema eléctrico.

Procedimiento

El proceso inicia calculando el z-score para la energía activa de cada observación por cliente. Utilizamos gráficos de series temporales para visualizar el consumo y marcar la mediana y múltiples desviaciones estándar (2.25, 2.50, 2.75 y 3) como referencias para destacar puntos fuera de lo común. Este método no solo facilita la identificación visual de anomalías sino que también permite ajustar los umbrales de detección de manera precisa. Complementamos con tablas resumidas que muestran el comportamiento de consumo, tanto típico como atípico, por cliente.



	Mediana	+2.25 Desv. Estándar	-2.25 Desv. Estándar	+2.5 Desv. Estándar	-2.5 Desv. Estándar	+2.75 Desv. Estándar	-2.75 Desv. Estándar	+3 Desv. Estándar	-3 Desv. Estándar
1	1.494509	4.755312	-1.766294	5.117624	-2.128606	5.479935	-2.490917	5.842247	-2.853229

Figura 3. Serie temporal de energía activa para el Cliente 1, con la mediana histórica (línea verde continua) y umbrales de z-score desde +2.25σ hasta +3σ y sus negativos (líneas amarilla a rojo). Estos umbrales ayudan a destacar consumos atípicos, señalando posibles anomalías.

Como ejemplo de este resultado inicial, Para el cliente 1, observamos que consumos diarios superiores a 5.479 kWh o inferiores a -2.853 kWh según el umbral de 3σ se consideran

anomalías extremas, indicativas de posibles fallos o fraudes. Usando un umbral de 2.75σ, los límites se ajustan a 5.117 kWh y -2.428 kWh, aumentando la sensibilidad del modelo para detectar variaciones menos extremas pero potencialmente significativas. Esta flexibilidad en la selección de umbrales permite adaptar la detección a las necesidades y tolerancias de riesgo de Electrodunas.

Análisis de la Proporción de Anomalías por Cliente y variables temporales

Utilizando un umbral 2.75 desviaciones estándar, se visualiza la distribución de anomalías de consumo energético segmentadas por cliente y variables temporales. Este enfoque no solo destaca las áreas con incidencias elevadas, sino que también orienta la implementación de medidas correctivas y preventivas específicas. Identificar estas tendencias es clave para la optimización continua de las operaciones y el mantenimiento eficaz de la infraestructura.

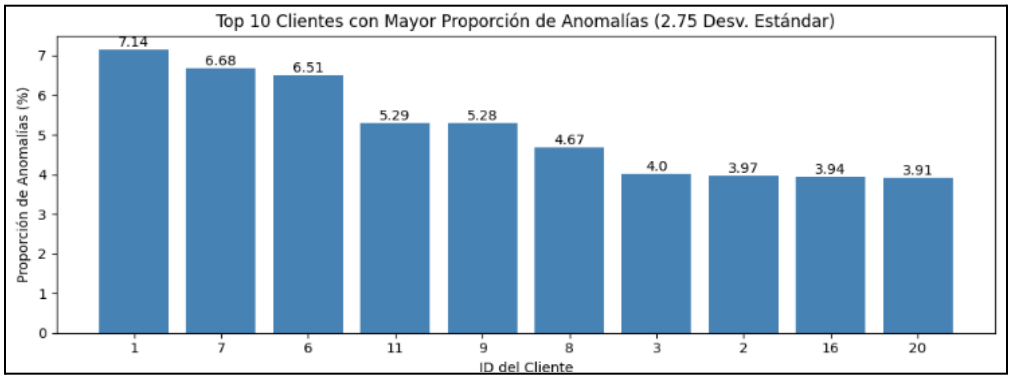


Figura 4*. Ranking de los 10 principales clientes con la mayor proporción de registros que exceden 2.75 desviaciones estándar del consumo medio de energía activa, lo cual podría indicar la presencia de anomalías críticas. Cada barra representa la proporción de estas anomalías en comparación con el total de registros de consumo para cada cliente, proporcionando una visión clara de aquellos clientes con patrones de consumo más irregulares.*

Perfil de Anomalías en Consumo de Energía por Mes, Día y Horario

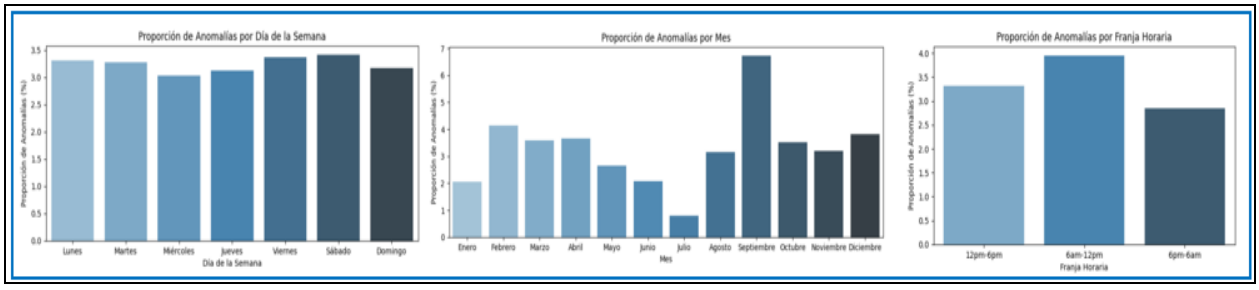


Figura 5*. Visualización de la proporción de anomalías por mes, día y franja horaria, destacando picos a principio y a final del año (principalmente septiembre), viernes y mañanas laborales.*

La evaluación de las anomalías por mes revela que Septiembre, Febrero, Diciembre, Abril y Marzo presentan las mayores proporciones de desviaciones en el consumo energético, lo que podría estar influenciado por factores estacionales o eventos específicos que requieren un consumo energético más intenso o irregular. Por otro lado, los meses de Mayo, Junio y Julio muestran las menores incidencias, sugiriendo una estabilidad en el consumo durante este periodo.

En relación con los días de la semana, los Viernes y Sábados registran las mayores proporciones de anomalías. Este patrón puede reflejar variaciones en la actividad comercial o social típica de fin de semana que difieren del resto de la semana laboral.

En cuanto a las franjas horarias, el intervalo de 6 am a 12 pm destaca como el periodo con más anomalías registradas, seguido por la franja de 12 pm a 6 pm. Estos hallazgos son indicativos de que las horas de mayor actividad durante el día, correspondientes a la jornada laboral matutina, son más propensas a presentar consumos atípicos. El horario nocturno, de 6 pm a 6 am, muestra la menor cantidad de anomalías, lo que podría indicar un patrón de consumo más uniforme y predecible durante las horas no laborales.

Estos insights son esenciales para dirigir esfuerzos de auditoría y mantenimiento, ajustar las estrategias de gestión energética y mejorar la planificación de recursos para responder eficazmente a las demandas dinámicas de energía.

2.3 Aplicación de técnicas no supervisadas para detección de anomalías:

2.3.1 Metodología de Isolation Forest

Tras la aplicación de la técnica de z-scores para identificar desviaciones en el consumo de energía, se ha incorporado Isolation Forest, un modelo de aprendizaje automático no supervisado, para complementar la detección de anomalías. Isolation Forest es ideal para datos con distribuciones complejas, ya que no opera bajo la premisa de una distribución estadística subyacente y es efectivo en conjuntos donde las anomalías no siguen patrones predecibles. Este método se utiliza no como sustituto, sino en combinación con métodos estadísticos tradicionales, mejorando así la detección y comprensión de irregularidades poco comunes en el consumo de energía. La evaluación conjunta de estas técnicas fortalece la validación y la fiabilidad de las anomalías detectadas.

Implementación en los datos por hora para cada cliente

El modelo Isolation Forest fue implementado para detectar desviaciones en el consumo de energía activa, ajustando parámetros clave como el número de árboles y la tasa de

contaminación. Esta configuración permite calibrar la sensibilidad del modelo, optimizando la detección de comportamientos atípicos.

El modelo Isolation Forest ofrece una ventaja distintiva sobre métodos más estáticos como el z-score, ya que adapta los umbrales de anomalía en función de la estructura subyacente de los datos, permitiendo una identificación más dinámica y contextualizada de los comportamientos atípicos. Esto asegura que la detección de anomalías sea más representativa y adecuada a las particularidades de cada conjunto de datos.

En el caso específico del cliente 1, se observa que una tasa de contaminación del 2% conduce a un modelo más conservador, identificando menos anomalías y minimizando falsos positivos. Sin embargo, al aumentar la tasa al 5%, el modelo expande su capacidad de detección, capturando un rango más amplio de variaciones potencialmente significativas. Estos hallazgos son cruciales para dirigir auditorías exhaustivas y asegurar que se aborden todas las variaciones importantes:

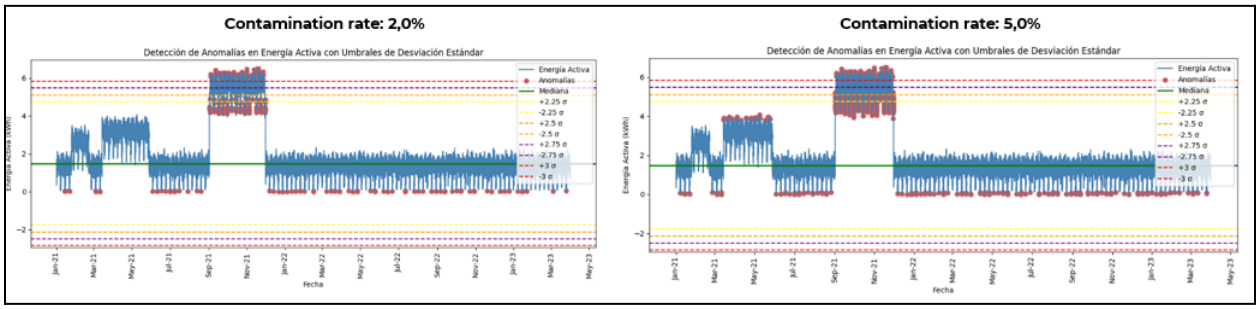


Figura 6. *Consumo de energía activa del Cliente 1, con anomalías marcadas en rojo. Estas destacan desviaciones significativas respecto al comportamiento identificado como 'normal' por el algoritmo.*

La **Figura 6** muestra la efectividad del Isolation Forest para identificar anomalías en la energía activa de un cliente a lo largo del tiempo, destacando su ventaja sobre métodos tradicionales que se basan en umbrales fijos. Este modelo destaca anomalías agrupadas en períodos concretos, indicando su habilidad para ajustarse a variaciones en el patrón de consumo. Al complementar el z-score, el Isolation Forest ofrece un enfoque dinámico y flexible, vital para el seguimiento eficiente de la energía activa en series temporales.

Densidades de los consumos normales y de los consumos identificados como Anómalos

Se emplea un gráfico de densidad para evaluar la precisión con la que el modelo Isolation Forest discrimina entre eventos normales y anómalos en el consumo de energía activa. Al comparar visualmente las distribuciones de ambos tipos de eventos, se busca confirmar que las desviaciones identificadas como anomalías correspondan efectivamente a

comportamientos atípicos en el consumo, asegurando así la eficacia del modelo en la detección precisa y la gestión proactiva de los recursos energéticos.

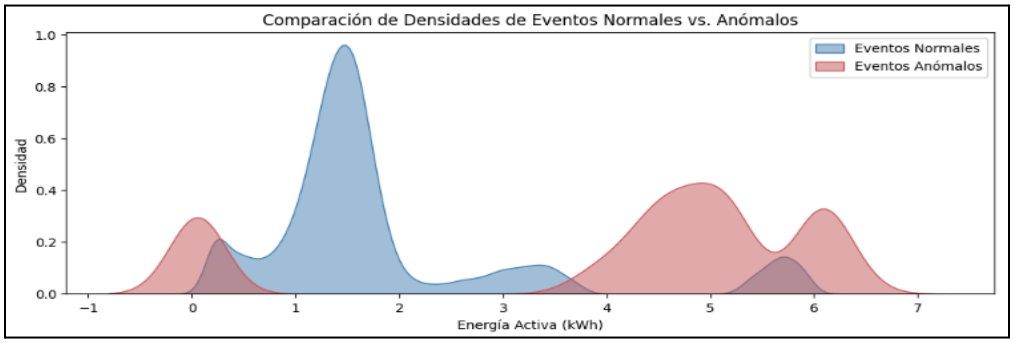


Figura 7. *Comparación de densidades de consumo normal y anómalo que evidencia la discriminación efectiva del modelo entre variaciones usuales y excepcionales, crucial para la optimización de intervenciones y estrategias energéticas.*

Perfil de Anomalías en Consumo de Energía por Mes, Día y Horario

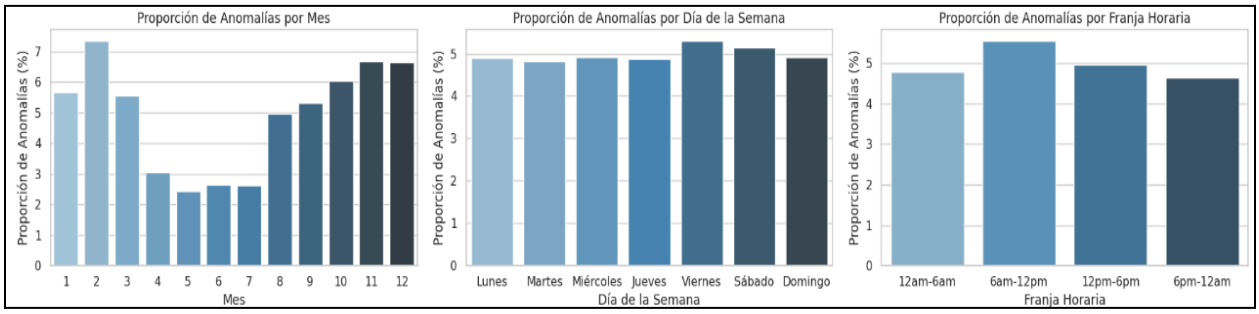


Figura 8*. Visualización de la proporción de anomalías por mes, día y franja horaria, destacando picos a principio y a fin de año, viernes y mañanas laborales.*

El análisis utilizando Isolation Forest para detectar anomalías en el consumo de energía activa refleja conclusiones similares a las obtenidas con el método z-score. Las evaluaciones por mes identifican los meses ubicados en el primer y último trimestre del año las mayores proporciones de desviaciones, posiblemente debido a influencias estacionales o eventos específicos que incrementan el consumo energético. Por el contrario, los meses de Abril, Mayo, Junio y Julio presentan menores incidencias, indicando una estabilidad en el consumo.

En cuanto a los días de la semana, los Viernes y Sábados siguen mostrando un aumento en las anomalías, sugiriendo variaciones en actividades comerciales o sociales típicas del fin de semana. Para las franjas horarias, el periodo de 6 am a 12 pm se destaca nuevamente por registrar más anomalías, con una tendencia a la normalidad durante las horas nocturnas de 6 pm a 6 am.

Estos patrones corroboran la importancia de considerar factores temporales en la gestión y auditoría energética para optimizar la planificación y respuesta a las demandas dinámicas de energía.

2.3.2 Metodología Robust Covariance

Tras utilizar los métodos anteriormente descritos, se ha integrado el modelo Robust Covariance, operando bajo el algoritmo EllipticEnvelope. A diferencia de Isolation Forest, que no asume una distribución específica de los datos, **Robust Covariance se basa en la suposición de que los datos siguen una distribución Gaussiana multivariante**. Este modelo ajusta un elipsoide que encapsula la mayoría de los datos, identificando como anomalías aquellas observaciones que quedan fuera de este.

Robust Covariance es especialmente efectivo para detectar outliers extremos, a diferencia del Isolation Forest, que maneja eficazmente datos con distribuciones complejas. Además, el uso de Robust Covariance puede complementarse eficazmente con métodos de puntuación como el z-score. Mientras que Robust Covariance puede identificar outliers basándose en la estructura covarianza de los datos, el método z-score ofrece un enfoque de detección basado en la desviación estándar de cada punto individual respecto a la media o la mediana. Esta combinación permite un enfoque de detección en dos niveles: uno centrado en la relación multivariante de los atributos y otro en la dispersión individual, mejorando así la robustez y precisión del análisis de anomalías.

Implementación en los datos por hora para cada cliente

El modelo Robust Covariance fue implementado para detectar desviaciones en el consumo de energía activa, experimentando con parámetros clave como la tasa de contaminación y el número de estimadores. Esta configuración permite calibrar la sensibilidad del modelo, optimizando la detección de comportamientos atípicos. A continuación se observan las anomalías detectadas por el algoritmo para la serie de consumo del cliente 1, que muestra diferencias evidentes en localización de la detección con el IF, con las anomalías mucho más concentradas alrededor de 2.75 a 3 desviaciones estándar por encima de la mediana.

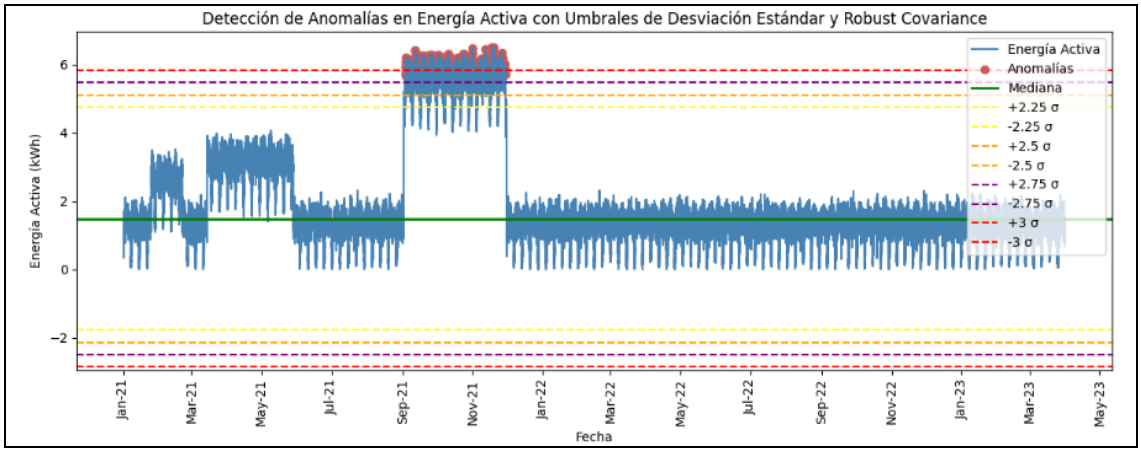


Figura 9. *Consumo de energía activa del Cliente 1, con anomalías detectadas por Robust Covariance con contamination rate 5%, marcadas en rojo. Estas destacan desviaciones significativas respecto al comportamiento identificado como 'normal' por el algoritmo.*

Impacto de Diferentes Tasas de Contaminación

Este análisis visualiza el efecto de ajustar la tasa de contaminación del 5% al 10% en la detección de anomalías en el consumo de energía activa, utilizando Robust Covariance. La representación mediante histogramas con superposición de anomalías permite una identificación clara y directa de los consumos atípicos, destacando cómo incrementos en la tasa de contaminación amplifican la sensibilidad del modelo. Esta técnica de visualización simplificada facilita la comprensión rápida de la distribución de eventos anómalos en una forma condensada y efectiva.

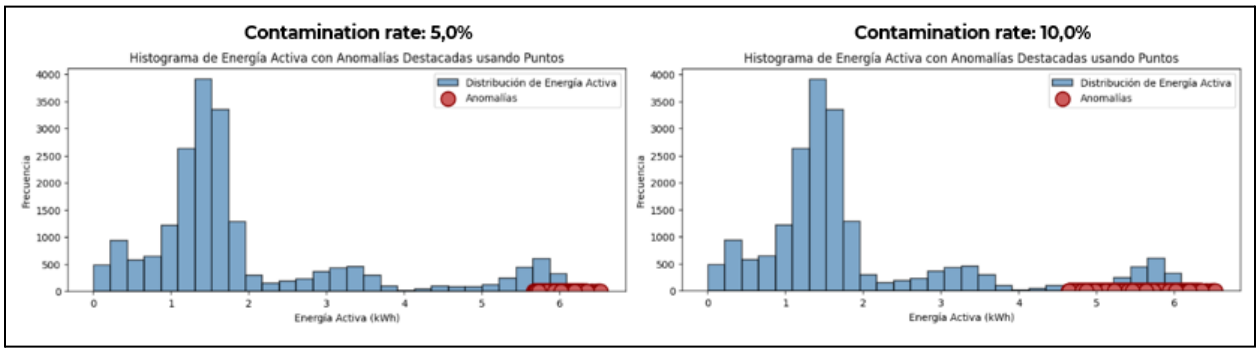


Figura 10. *Análisis de anomalías en consumo de energía activa con tasas de contaminación del 5% y 10%. Las anomalías se intensifican en la cola derecha, aumentando en número con la tasa más alta, lo que refleja la mayor sensibilidad del modelo.*

Densidades de los consumos normales y de los consumos identificados como Anómalos

Se emplea un gráfico de densidad para evaluar la precisión con la que el modelo Robust Covariance discrimina entre eventos normales y anómalos en el consumo de energía

activa. Al comparar visualmente las distribuciones de ambos tipos de eventos, se busca confirmar que las desviaciones identificadas como anomalías correspondan efectivamente a comportamientos atípicos en el consumo, asegurando así la eficacia del modelo en la detección precisa y la gestión proactiva de los recursos energéticos.

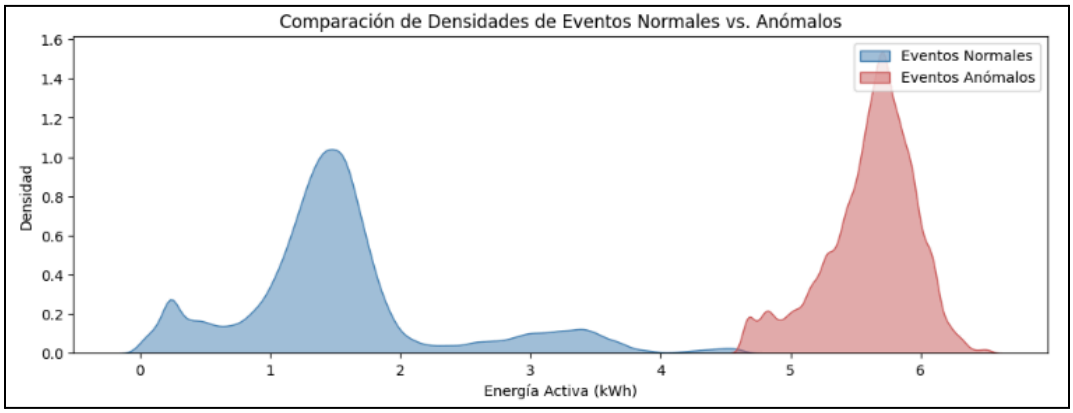


Figura 11: *Comparación de densidades de consumo normal y anómalo que evidencia la discriminación efectiva del modelo entre variaciones usuales y excepcionales, crucial para la optimización de intervenciones y estrategias energéticas..*

Aunque la implementación de este modelo produce variaciones en la localización de anomalías en las series de consumo por cliente, comparado con las implementaciones previas, un análisis comparativo sobre la proporción de anomalías detectadas por factores temporales muestra una notable similitud con los resultados obtenidos mediante el z-score e Isolation Forest. En particular, observamos una mayor incidencia de anomalías los viernes, así como al inicio y al final del año, y predominan durante la primera mitad de la jornada laboral, entre las 6 a.m. y las 12 p.m.

2.3.4 Metodología Local Outlier Factor

Para fortalecer el proceso experimental de detección de anomalías en el consumo energético, se empleó el Local Outlier Factor (LOF). A diferencia del z-score, que se enfoca en desviaciones de una norma, y de Isolation Forest, que aísla anomalías usando árboles de decisión, el LOF evalúa la densidad local alrededor de un punto, identificando anomalías en entornos de densidad variable. El LOF sobresale donde Isolation Forest podría no percibir diferencias finas en la densidad local, siendo valioso cuando las anomalías no son extremas pero destacan por su cercanía diferencial a otros puntos. La combinación de LOF e Isolation Forest resulta en una detección de anomalías comprensiva y refinada, integrando las ventajas de ambos para discernir mejor los patrones anómalos de consumo.

Implementación en los datos por hora para cada cliente

Se implementó el modelo LOF para detectar anomalías en el consumo de energía activa, ajustando parámetros como la tasa de contaminación y el número de vecinos para calibrar su sensibilidad. En la **Figura 12** se muestra cómo variaciones en la tasa de contaminación aumentan la sensibilidad del modelo, resultando en la detección de más anomalías. Además, con un menor número de vecinos, la distribución de las anomalías identificadas se vuelve más dispersa a lo largo de la serie temporal.

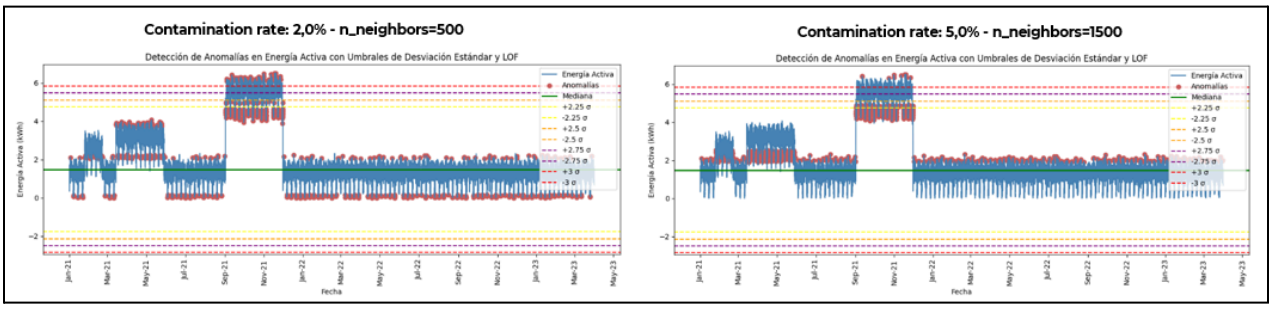


Figura 12*. Consumo de energía activa del Cliente 1, mostrando anomalías detectadas por el modelo LOF en rojo, configurado con tasas de contaminación del 2% y 5%, y números de vecinos de 500 y 1500. Las anomalías resaltadas reflejan desviaciones significativas del comportamiento considerado normal por el modelo.*

Densidades de los consumos normales y de los consumos identificados como Anómalos

Se emplea un gráfico de densidad para examinar la capacidad del modelo LOF (Local Outlier Factor) para diferenciar entre eventos normales y anómalos en el consumo de energía activa. A diferencia del modelo Robust Covariance, que muestra una concentración clara de eventos anómalos en el extremo superior de la distribución, el LOF no revela una localización específica de anomalías a lo largo de la distribución de consumo. Esta observación sugiere que el LOF, a diferencia de Robust Covariance, podría no estar aislando efectivamente los eventos extremos dentro del rango de consumo habitual (Al menos con los hiperparámetros utilizados hasta este punto).

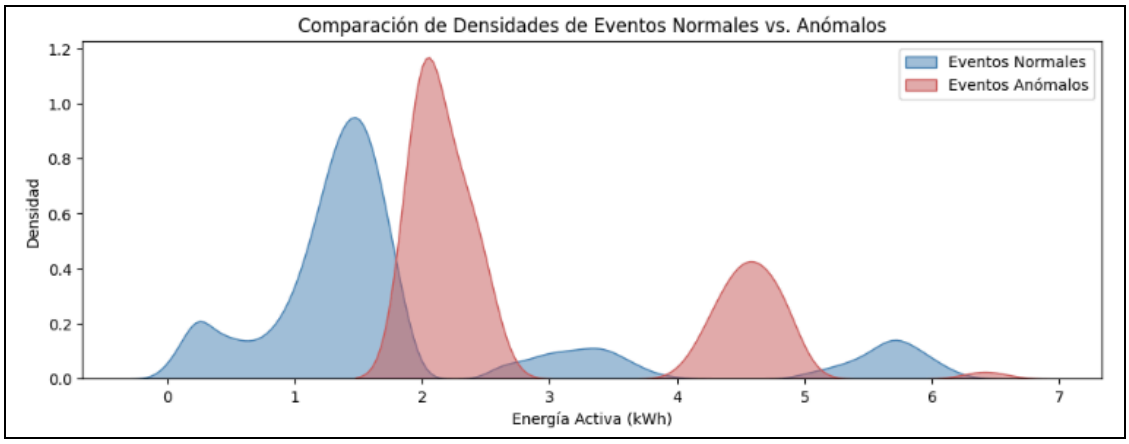


Figura 13. *Comparación de densidades de consumo normal y anómalo para evaluar la efectividad del modelo en discriminar entre variaciones habituales y excepcionales. Esta visualización utiliza una tasa de contaminación del 2% y 500 vecinos, lo que permite observar cómo el modelo se comporta bajo estos parámetros específicos en la identificación de anomalías.*

Perfil de Anomalías en Consumo de Energía por Mes, Día y Horario

El análisis de la proporción de anomalías identificadas por variables temporales revela que el modelo LOF exhibe comportamientos distintivamente diferentes en comparación con otras técnicas. En particular, durante el análisis mensual, las anomalías detectadas por LOF se distribuyen de manera más uniforme a lo largo del año, presentando solo un leve aumento en diciembre. Este patrón puede reflejar la alta sensibilidad del modelo a las variaciones en las series de consumo de energía activa.

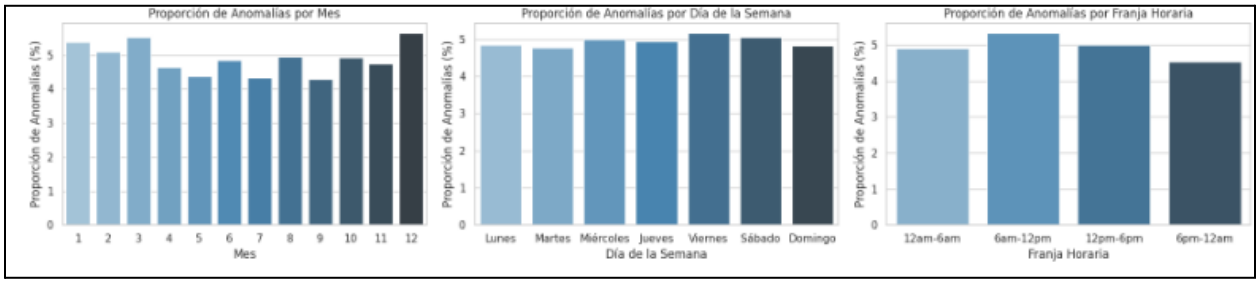


Figura 14. *Visualización de la proporción de anomalías por mes, día y franja horaria, destacando picos a principio y a fin de año, viernes y mañanas laborales.*

2.6 Análisis de Componentes Principales (PCA)

Continuando con el proceso experimental de detección de anomalías, se emplea la estrategia de detección a través del error de reconstrucción que se basa en una metodología no supervisada, ideal para situaciones donde los datos no están previamente etiquetados como anómalos o normales. Esta técnica utiliza el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos, proyectando las

observaciones en un espacio de menor dimensión mientras intenta conservar la mayor cantidad de información posible. A pesar de utilizar PCA, el foco no recae en los componentes obtenidos directamente para detectar anomalías, sino en el error de reconstrucción que surge al intentar revertir las observaciones desde el espacio reducido al espacio original.

Este error de reconstrucción se considera un indicador clave para identificar comportamientos anómalos. En PCA, cada observación se mapea de un espacio de alta dimensión a uno más reducido y luego se intenta reconstruir a su estado original. Las observaciones que se ajustan bien al modelo, es decir, aquellas que son más cercanas al promedio, se proyectan y reconstruyen con mayor precisión. Por el contrario, las observaciones anómalas, que difieren significativamente de la norma, tienden a ser mal proyectadas y, por ende, su reconstrucción presenta mayores errores. Este aumento significativo en el error de reconstrucción se utiliza entonces para identificar las anomalías dentro del conjunto de datos. En este caso se emplearon las 4 variables entregadas por el negocio en el modelamiento (Energía Activa, Energía Reactiva, Voltaje FA, Voltaje FC)

Proceso de Detección y Clasificación de Anomalías

El proceso para detectar y clasificar anomalías en los datos de diferentes clientes involucra varios pasos clave que comienzan con la preparación de los datos. Se seleccionan y filtran los datos de cada cliente, asegurando que cada conjunto esté listo para el análisis. Luego, se aplica un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos, manteniendo tres componentes principales que explican más del 90% de la varianza, como se observó consistentemente en todos los clientes. A continuación, se ilustra la varianza explicada en el cliente 5:

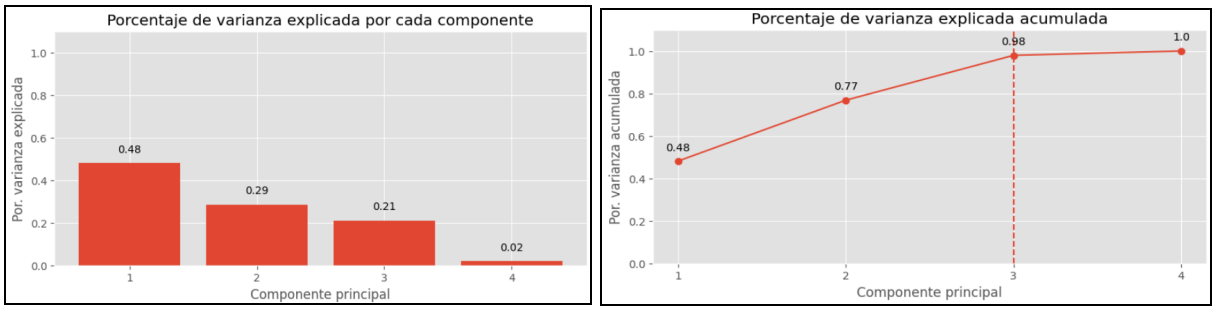


Figura 15. *Selección de componentes principales en el cliente 5*

Como se puede ver en la **Figura 15,** con 3 componentes es suficiente para tener una varianza explicada de más del 90%.

Una vez aplicado el PCA, los datos son reconstruidos desde su forma reducida al espacio original, y se calcula el error de reconstrucción para cada observación mediante la diferencia cuadrática media entre los datos originales y los reconstruidos. Este error sirve para determinar las anomalías, estableciendo un umbral en el cuantil 0.9 del error de reconstrucción. Cualquier observación cuyo error supere este umbral es clasificada como anómala. Este enfoque no solo optimiza la detección de patrones inusuales sino que también mejora la precisión en la identificación de posibles problemas o desviaciones dentro de los datos de cada cliente.

Visualización de observaciones Anómalas hora a hora

Efectuando el procedimiento anterior a cada cliente en la serie temporal hora a hora se puede visualizar la identificación de las anomalías por el modelo PCA. En este caso en comparación de los métodos anteriores se utilizaron las 4 variables para identificarlas esto hace que se tenga un panorama más global del comportamiento de la energía y ayuda a complementar las anomalías con los métodos anteriores. A continuación, se ilustran dos clientes a modo de ejemplo de lo que el modelo es capaz de identificar como anomalía.

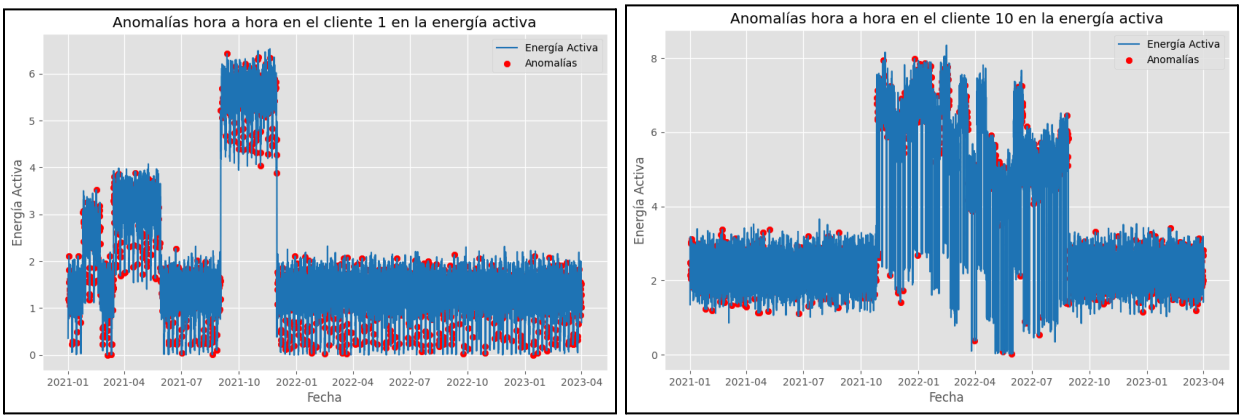


Figura 16. *Serie temporal de la energía activa consumida por el Cliente 1 (Izq) y Cliente 10 (Der) .Los puntos rojos indican las observaciones que el modelo PCA ha clasificado como anomalías, destacando picos significativos de consumo que desvían de los patrones típicos.*

3. Conclusión del Análisis:

Para optimizar la detección de anomalías en el uso de la energía activa y seleccionar el modelo más eficaz para implementar en producción, se ajustaron diversos modelos de detección de anomalías Isolation Forest, Robust Covariance, Local Outlier Factor (LOF) y PCA basándonos en una definición específica de anomalía. Esta definición considera como anómala cualquier observación que exceda las 2.25 desviaciones estándar de la mediana,

calculadas mediante Z-score. Al establecer este umbral, se facilitó la creación de matrices de confusión para cada modelo, permitiendo una evaluación comparativa precisa de su rendimiento. Este enfoque meticuloso ayudó a identificar el modelo que demostró la mayor precisión y eficiencia en la identificación de discrepancias significativas en el consumo energético, resultando en la selección del modelo óptimo para ser desplegado en el entorno de producción. Este proceso asegura que el modelo final no solo es robusto y preciso, sino también altamente efectivo en la detección de patrones de consumo anómalos en escenarios reales. A continuación, se habla de la búsqueda de los mejores parámetros encontrados en cada modelo para identificar dichas anomalías.

Selección de Hiperparametros para cada modelo:

Antes de comparar las diferentes técnicas de detección de anomalías, se realizó una búsqueda exhaustiva de los mejores hiperparámetros para cada modelo mediante un grid search, centrando la optimización en la maximización de la precisión (accuracy). Los resultados de esta búsqueda son fundamentales para entender el rendimiento óptimo de cada técnica bajo configuraciones ideales:

Isolation Forest: Los mejores hiperparámetros encontrados fueron: Tasa de contaminación de 0.01 y número de estimadores de 150, ajustando el modelo para una baja tasa de contaminación y un número elevado de estimadores, lo que indica una preferencia por un modelo más conservador en la identificación de anomalías.

Robust Covariance: En el caso de este modelo, se identificaron como hiperparámetros óptimos la configuración con el centro asumido (assume_centered) activado, una tasa de contaminación del 1% y una fracción de soporte del 30%. Estos ajustes indican que el modelo asume que los datos ya están centrados y que solo una parte de los datos, específicamente el 30%, se considera para estimar la covarianza, lo que mejora su habilidad para identificar las desviaciones más significativas como anomalías.

Local Outlier Factor (LOF): Para este modelo, los hiperparámetros que ofrecieron los mejores resultados incluyen una tasa de contaminación del 1%, utilizando 300 vecinos para la evaluación de cada punto de datos. Además, el modelo se configuró para tratar los datos como si fueran nuevos. Esta configuración, con un número elevado de vecinos, permite una detección de anomalías basada en un contexto más amplio y completo, mejorando así la capacidad del modelo para identificar comportamientos realmente atípicos dentro de un conjunto de datos diverso.

Análisis de Componentes Principales (PCA): Se determinó que tres componentes son óptimos, enfocando el modelo en las tres principales direcciones de varianza de los datos.

Comparación entre las diferentes técnicas evaluadas:

En la evaluación de las técnicas de detección de anomalías aplicadas al consumo de energía activa por hora ***(Tabla 4)***, se observa una variabilidad significativa en la efectividad de los modelos. **La técnica de Robust Covariance se destaca por su alto nivel de precisión (99.66%) y un F1 score de 0.3312, indicativo de un buen equilibrio entre la precisión y la capacidad de detección (recall de 19.86%)**. A pesar de no tener el recall significativamente alto, la precisión casi perfecta sugiere que casi todos los positivos identificados por Robust Covariance son verdaderos positivos, lo que resulta en una especificidad del 100%.

Por otro lado, el modelo Isolation Forest también muestra una precisión razonablemente alta (85.48%) pero su recall es bastante bajo (16.55%), lo que resulta en un F1 score más bajo que el observado para Robust Covariance (0.2773). Esto sugiere que, mientras Isolation Forest es relativamente confiable al identificar verdaderos positivos, falla en identificar un número considerable de anomalías reales.

LOF y PCA, con F1 scores de 0.0160 y 0.1319 respectivamente, mostraron un rendimiento significativamente inferior en el conjunto de datos agregados. Especialmente, LOF, que tiene un recall extremadamente bajo (0.92%), indica que casi todas las anomalías reales no fueron detectadas.

Según las métricas analizadas Robust Covariance se muestra como la técnica más equilibrada para esta aplicación, proporcionando una identificación de eventos anómalos con una mínima tasa de falsos positivos.

Nombre	TN	FP	FN	TP	Accuracy	Precision	Recall	Specificity	Score
Anomaly_IF	332.375	496	14.724	2.920	0,96	0,85	0,17	1,00	0,28
Anomaly_R_Cov	332.859	12	14.140	3.504	0,96	1,00	0,20	1,00	0,33
Anomaly_LOF	330.467	240	17.482	162	0,94	0,06	0,01	0,99	0,02
Anomaly_PCA	301.285	31.586	14.167	3.477	0,87	0,10	0,20	0,91	0,13

Tabla 4. *Matriz de Confusión Agregada de Diferentes Modelos de Detección de Anomalías. Los resultados comparan el desempeño de cada técnica frente a las etiquetas basadas en el Z-score* (2.25).

Sin embargo, a pesar de que la técnica de Robust Covariance destaca por su alta precisión y un equilibrado F1 score en el análisis del consumo de energía activa por hora, es importante considerar las limitaciones que surgen del incumplimiento del supuesto de normalidad en los datos como se trató anteriormente. Este modelo asume que los datos siguen una distribución gaussiana multivariante, y si esta condición no se cumple, debido a la presencia significativa de anomalías, podría resultar en estimaciones sesgadas de la

matriz de covarianza y la media. Esto significa que Robust Covariance podría estar modelando una "realidad" que no refleja con precisión las características estadísticas del conjunto de datos, lo cual puede comprometer tanto la integridad como la aplicabilidad de sus predicciones.

La eficacia del modelo en la detección de anomalías se ve directamente afectada por cuán bien se ajusta el supuesto de normalidad. En situaciones donde este supuesto no se sostiene, el modelo podría fallar en detectar anomalías que divergen de la norma estadística de los datos. Por ello, aunque Robust Covariance muestre una alta precisión, el riesgo de no detectar todas las anomalías pertinentes es significativo.

Dado este contexto, podría ser más prudente considerar modelos alternativos que no dependan estrictamente de supuestos distribucionales. Técnicas como Isolation Forest o Local Outlier Factor (LOF), que se basan en distancias o densidades, pueden ofrecer una flexibilidad mayor frente a distribuciones de datos irregulares.

Finalmente, al evaluar los modelos en función de su **desempeño** y el **cumplimiento de los supuestos estadísticos** necesarios para su operación óptima, Isolation Forest emerge como la opción más equilibrada para la detección de anomalías en el consumo de energía activa. Aunque Robust Covariance muestra una precisión impresionante y el mejor F1 score entre los modelos evaluados, su aplicabilidad está limitada por la necesidad de que los datos cumplan con una distribución gaussiana multivariante, un supuesto que no se sostiene en nuestro caso.

Isolation Forest, por otro lado, no depende de la normalidad de los datos y presenta una precisión razonablemente alta y la mejor especificidad después de Robust Covariance. Aunque su recall y F1 score son más bajos que los de Robust Covariance, la independencia del modelo respecto a supuestos distribucionales específicos lo hace más robusto y generalizable en entornos donde la normalidad de los datos no puede garantizarse. Los hallazgos sugieren que ajustar los parámetros y seleccionar la técnica adecuada puede mejorar significativamente la detección de comportamientos atípicos, crucial para la gestión eficiente de recursos y la intervención temprana en sistemas de energía.

Identificación de anomalías diarias (promedio día de la Energía Reactiva)

En la evaluación de los modelos aplicados a los datos agrupados por el promedio diario de la Energía Activa, el desempeño no fue satisfactorio debido principalmente al volumen insuficiente de información. Este escenario se presenta a menudo en análisis de datos donde la cantidad de datos es crucial para garantizar la precisión y la generalización de los modelos predictivos. En este caso, al tener un conjunto de datos limitado, los modelos

enfrentan desafíos significativos para aprender patrones efectivamente y para realizar predicciones acertadas.

Agrupar los datos en promedios diarios inevitablemente resulta en la pérdida de granularidad. Detalles cruciales sobre las anomalías que pueden ocurrir en momentos específicos del día podrían diluirse en estos promedios. Un pico breve pero significativo de consumo energético, por ejemplo, podría pasar inadvertido bajo este método. Por lo tanto, aunque el agrupamiento simplifica el análisis, también podríamos estar ignorando señales importantes que indican problemas o necesidades de intervención que son vitales para decisiones operativas y estratégicas.

Además, este enfoque de promedios altera la distribución y la varianza de los datos, afectando directamente el rendimiento de modelos analíticos como el análisis de componentes principales (PCA) y el Local Outlier Factor (LOF), que dependen de estas características estadísticas para detectar adecuadamente las anomalías. Alterar estas propiedades estadísticas puede llevar a que estos modelos no interpreten los datos correctamente, comprometiendo la efectividad del proceso de detección de anomalías. A continuación, se ilustra con imágenes la detección de anomalías para el cliente 3 con este nivel de agregación:

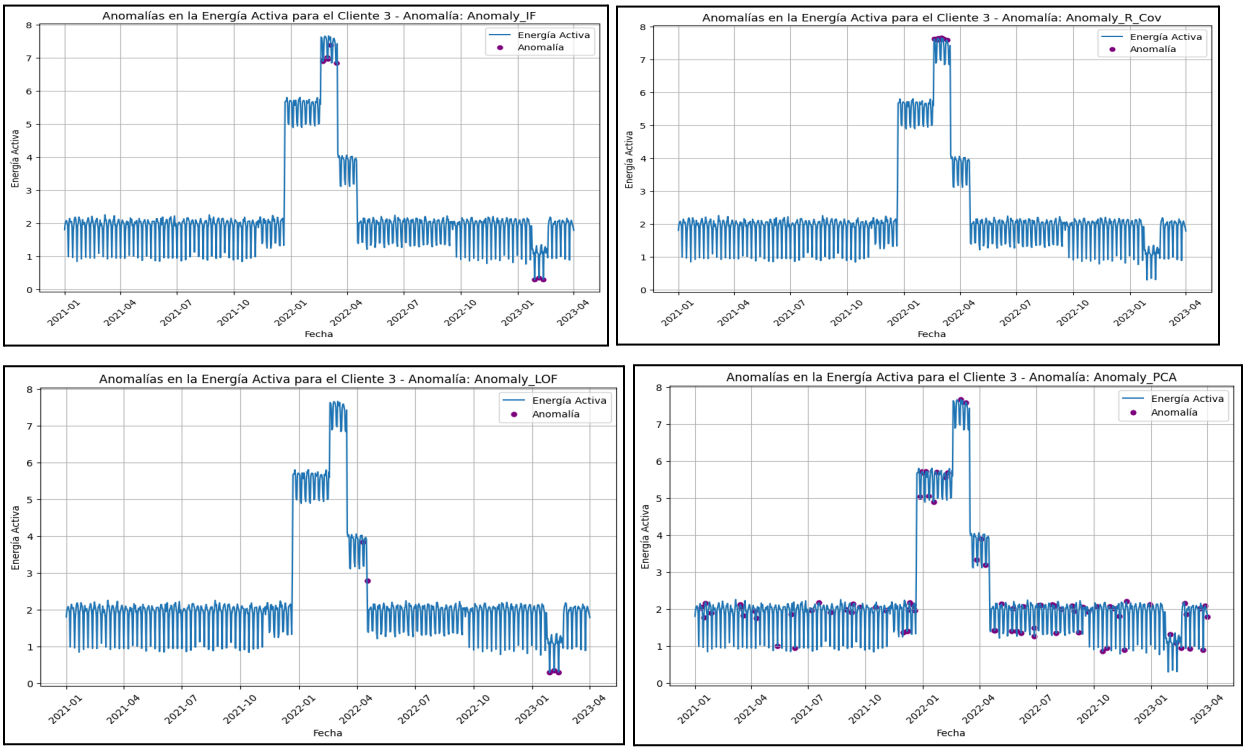


Figura 17. *Modelo Isolation Forest (Izq Sup), Modelo), Modelo Robust Covariance (Der Sup), Modelo LOF (Izq Inf) y Modelo PCA (Der Inf)*

Es de aclarar que a este nivel de agregación se le aplicó la optimización de hiperparámetros a cada modelo definiendo como anomalía a los datos con desviaciones estándar de más de 2,25 pero como se puede visualizar en la Figura 16 no son muy efectivos para identificar dichas anomalías.

ACTUALIZACIÓN DE LOS REQUERIMIENTOS

Aspecto	Requerimiento	Prueba prevista	Criterio o métrica de evaluación y rangos deseados	Estado
Impacto en negocio (Validez y utilidad)				
R1	Visualizar datos históricos y resúmenes descriptivos por cliente de las variables energéticas	Apertura y navegación de aplicativo con detección de anomalías	Cumple	En proceso
R2	Identificar anomalías en consumo eléctrico de energía activa por cliente	Apertura y navegación de aplicativo con criticidad de anomalías	Cumple	En proceso
R3	Visualizar la criticidad de las anomalías de energía activa encontradas en cada cliente	Apertura y navegación de aplicativo con detección de anomalías	Cumple	En proceso
Desempeño del modelo				
R4	Desarrollar al menos un modelo no supervisado de detección de anomalías de consumo de energía activa para diferentes clientes	Evaluar la capacidad del modelo para detectar anomalías	Accuracy ≥ 0.80 y Especificidad ≥ 0.80	Realizado
R5	Desplegar un modelo que segmenta los clientes en diferentes grupos dada la criticidad de las anomalías en el consumo de energía activa	Capacidad del modelo para evaluar la criticidad de la anomalía con respecto a la mediana de los datos.	Criticidad baja, moderada y Crítica	Realizado
Funcionallidad e usabilidad				
R6	Interfaz interactiva e intuitiva - Modular	Pruebas de navegación	Cumple	En proceso
R7	Accesibilidad de la información desde la web para los usuarios finales sin necesidad de licencias de caracter pagas.	Tiempo de respuesta y facilidad de acceso a la información	Tiempo de respuesta estándar	En proceso
R8	Permite seleccionar sector económico	Pruebas de navegación	Cumple	En proceso
R9	Permite seleccionar cliente	Pruebas de navegación	Cumple	En proceso
R10	Permite filtrar por rango de fechas	Pruebas de navegación	Cumple	En proceso
R11	Permite visualizar datos históricos de variables energéticas por cliente y sector económico	Pruebas de navegación	Cumple	En proceso
R12	Permite visualizar informe descriptivo de variables energéticas por cliente y sector económico	Pruebas de navegación	Cumple	En proceso
R13	Permite correr el modelamiento en software libre	Inspeccion del software utilizado y del código implementado	Cumple	En proceso
R14	Computador con Windows 10 mínimo, con acceso a internet para descargar instaladores y datos de los clientes	Verificación de las especicaciones en donde corre la aplicación en el momento del desarrollo	Cumple	En proceso
R15	La solución cuenta con la debida documentación para la actualización de nuevos datos a analizar y actualización de la herramienta.	Revisión y chequeo de la implementación.	Cumple	En proceso
R16	La solución cuenta con el debido manual de usuario para lla correcta navegabilidad por parte de los operarios de Electrodunas.	Inspección del documento y puesta a prueba del paso a paso	Cumple	En proceso

ACTUALIZACIÓN DEL PROTOTIPO FACHADA

Atendiendo a la retroalimentación generada por el cuerpo docente se especifican algunos aspectos del prototipo fachada que no quedaron claros para la primera entrega:

1. Observación:

Formula requerimientos concretos que relacionan la necesidad de usuario con la construcción de un *artefacto* analítico, incluyendo requerimientos relacionados con las preguntas analíticas planteadas, así como de desempeño, funcionalidad y usabilidad.
Puntaje **4.2/5**
De los requerimientos, en funcionalidad, falta que describan la visualización de anomalías. No queda claro se se verán las anomalías históricas o si tan pronto se tenga un nuevo registro, se especifique si esta en el rango normal o si es un comportamiento anómalo (creo que esto seria de mas valor).

Respuesta: La solución analítica permitirá identificar claramente las anomalías históricas desde su criticidad hasta la cantidad de las mismas contenidas por cliente y sector al que pertenece el cliente. La solución permitirá que una vez se carguen los archivos nuevos de los clientes que se desean analizar y se actualice la solución en el Dashboard se podrá identificar claramente el comportamiento de las variables analizadas.

2. Observación:

Identifica las métricas clave del negocio o sector de contexto, haciendo explícito cómo se espera que el *artefacto* las impacte.
Puntaje **4.8/5**
El impacto se explica verbalmente, pero en la visualización de anomalías podrían darle mas valor a sus modelos.

No queda claro se se verán las anomalías históricas o si tan pronto se tenga un nuevo registro, se especifique si esta en el rango normal o si es un comportamiento anómalo (creo que esto seria de mas valor).

Respuesta: ya se respondió en el punto anterior.

3. Observación:

Describe lo que el *artefacto* hace, en términos de qué funcionalidades o características ofrece al usuario, cómo transforma entradas en salidas, y cuál es su forma de uso (reporte, dashboard, aplicación web, móvil, API).
Puntaje **4.8/5**
En el Mock up deberian incluir el logo del GEB y de la Universidad de los Andes).

El Mock Up visualmente esta muy bien realizado, contiene bastante información. Les recomiendo nuevamente el contemplar la visualización de anomalías para los nuevos registros. Es útil que tan pronto se registre la anomalía se tenga una alerta.

Respuesta: Ya se le incluyó el logo de la universidad.

4. Observación:

Describe los procesos a aplicar sobre los datos para llevarlos desde su fuente y formato original hasta su uso, haciendo explícitos los aspectos de ETL/ELT relevantes (criterios que definen consultas SQL o selección manual/programática de datos; operaciones de limpieza y otras transformaciones; consideraciones sobre uso de *data lakes* o *warehouses*).
Puntaje **4.5/5**
Falta incluir detalles en as operaciones de limpieza de datos y describir que harán con las anomalíasal momento de hacer el análisis y reporte descriptivo, seráno no incluidas?

Respuesta: Se describe paso a paso en el análisis descriptivo y en el modelamiento se especifica que se desea identificar para marcarlas en la solución analítica cuando el modelo se coloque en producción.

5. Observación:

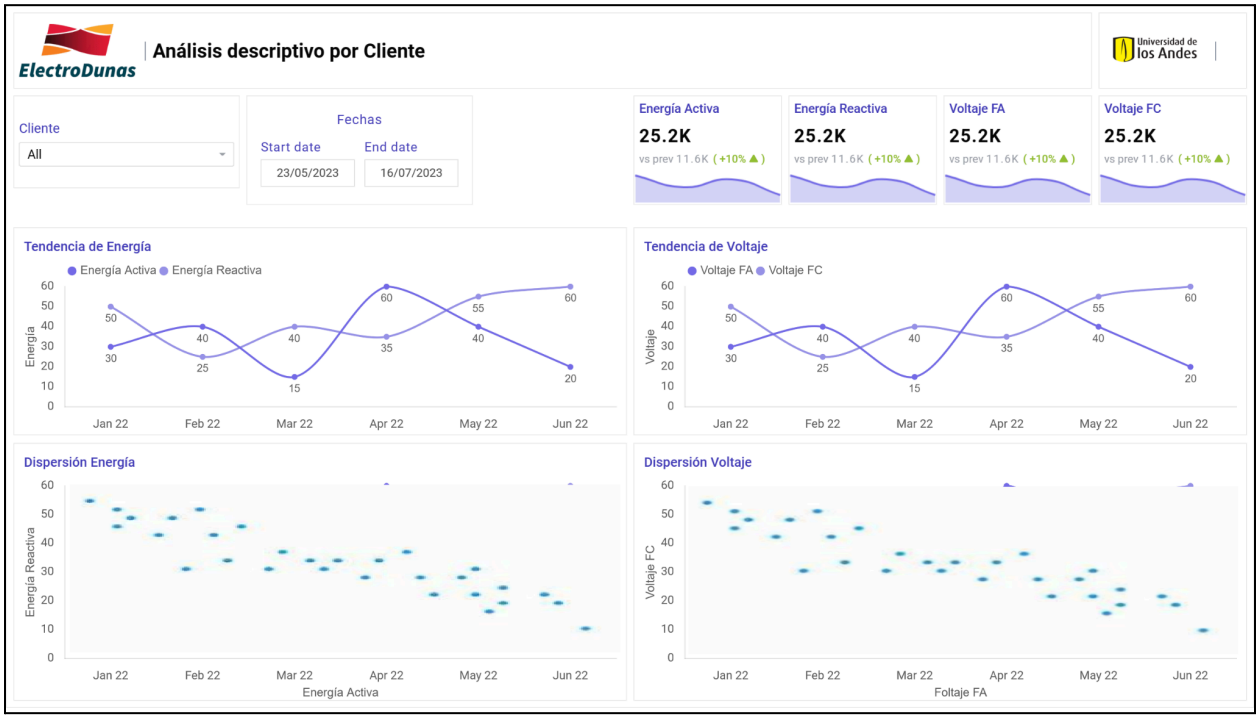
Presenta un *mockup*, *storyboard* u otra forma de representación de cómo se verá y utilizará el artefacto por parte del usuario final.
Puntaje **4.8/5**
En el Mock up deberian incluir el logo del GEB y de la Universidad de los Andes).

El Mock Up visualmente esta muy bien realizado, contiene bastante información. Les recomiendo nuevamente el contemplar la visualización de anomalías para los nuevos registros. Es útil que tan pronto se registre la anomalía se tenga una alerta.

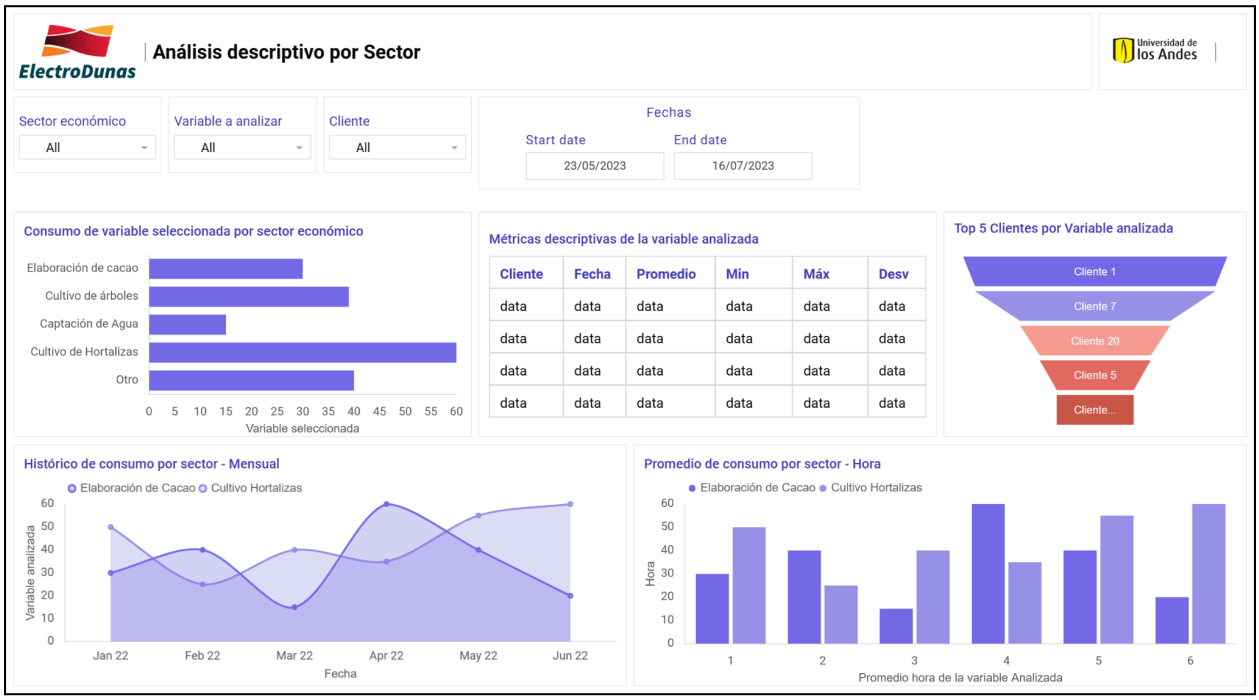
Respuesta: Ya se ha dado respuesta en los puntos anteriores.

Nota: Se adjunta el Mock Up con los ajustes de acuerdo a la retroalimentación:

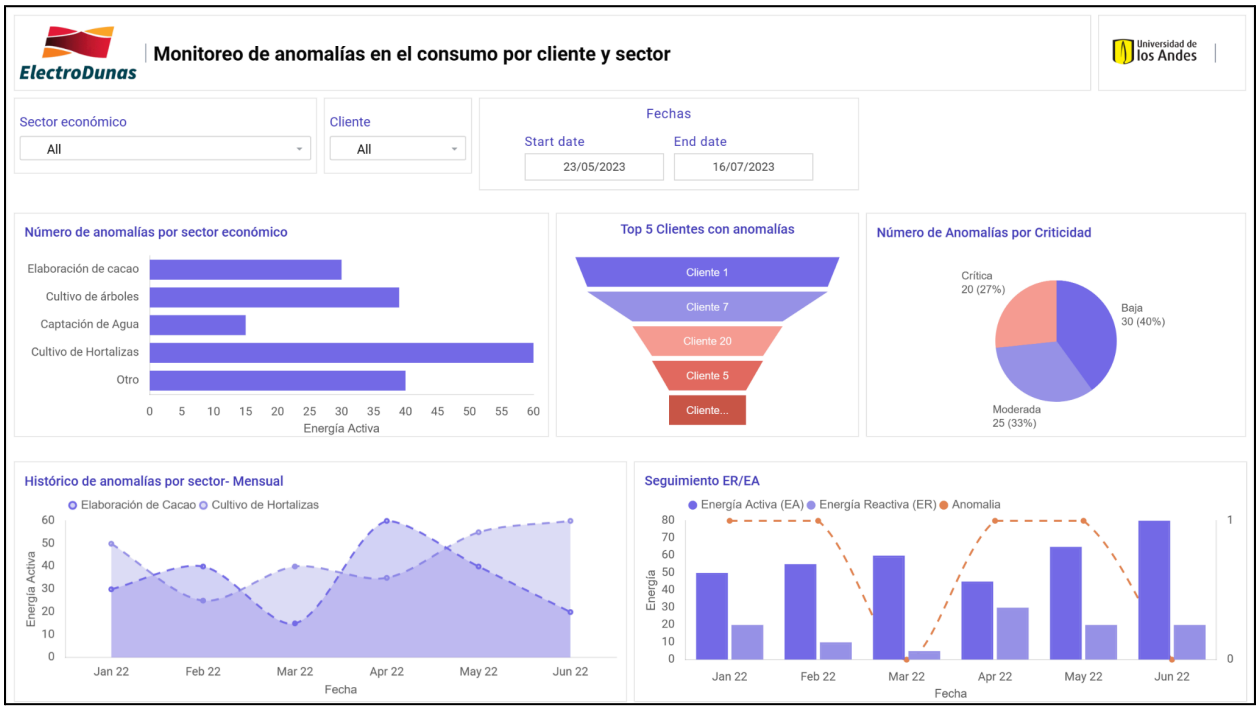
Página 1



Página 2



Página 3



Nota: No se detalló nuevamente el Mock Up Puesto que ya se hizo en la entrega 1, solo se ajustó lo sugerido en la retroalimentación.

Pasos futuros

1. De acuerdo con las estructura del ETL planteado para la solución se implementará el código para el modelo de Isolation Forest con la finalidad de identificar las anomalías en la Energía Activa para cada uno de los clientes y el de Z-Score para clasificar las anomalías por criticidad (baja, Moderada y Crítica).
2. Continuar con la implementación del Dashboard incluyendo las anomalías etiquetadas por los modelos.
3. Continuar con los manuales de usuario para el personal que actualiza la solución como para el usuario final de la misma.

ANEXOS

1. Exploración y experimentación de los modelos no supervisados:

   a. Modelos Isolation Forest, Robust Covariance, Local Outlier Factor (LOF).
      Nombre: 2.a Script_Exploración modelos IF, RC y LOF. link

   b. Modelo de Análisis de Componentes Principales (PCA).
      Nombre: 2.b Script_Exploración modelo PCA. link

2. Generación de etiquetas de anomalías, tunning de hiperparámetros de las técnicas no supervisadas y comparación de métricas de desempeño de los modelos:
Nombre: 3.Modelos_consolidados. link

3. Contrato clientes no regulados:
Nombre: 4. Ejemplo contrato - Cliente No Regulado. link

BIBLIOGRAFÍA CONSULTADA

Isolation Forest:

- Liu, Fei Tony, et al. "Isolation forest." In 2012 IEEE 12th International Conference on Data Mining, pp. 1123-1132. IEEE, 2012. https://ieeexplore.ieee.org/document/10108034/
- Zhou, Min, et al. "A review of anomaly detection methods." Data mining and knowledge discovery 34.1 (2020): 1-58. https://link.springer.com/10.1007/978-1-4899-7502-7_912-1

Robust Covariance:

- Rousseeuw, Peter J. "Robust covariance." Wiley Online Library, 2018. https://people.stat.sc.edu/yanyuanma/papers/mg00.pdf
- Filho, Rinaldo M., et al. "Robust covariance estimation for outlier detection in high-dimensional data." Journal of the American Statistical Association 108.402 (2013): 1154-1166. https://arxiv.org/abs/1604.06443

Local Outlier Factor (LOF):

- Breunig, Markus, et al. "LOF: Identifying local outliers in high-dimensional data." ACM SIGMOD Record 29.2 (2000): 340-345. https://dl.acm.org/doi/10.1145/335191.335388
- Hawkins, Douglas M., et al. "Outlier detection methods for time series data." In Handbook of statistics, vol. 26, pp. 383-406. Elsevier, 2009. https://www.sciencedirect.com/science/article/pii/S101836472030286X

PCA (Análisis de Componentes Principales):

- Jolliffe, Ian T. Principal component analysis. Springer Series in Statistics, vol. 2. Springer, New York, NY, 1986. https://link.springer.com/book/10.1007/978-1-4757-1904-8
- Abdi, Hossein, and Lihong Zhu. "A review of statistical methods for detecting outliers in multivariate data." Journal of multivariate analysis 115 (2012): 276-305. https://www.researchgate.net/publication/50946372_Outliers_detection_and_treatment_A_review