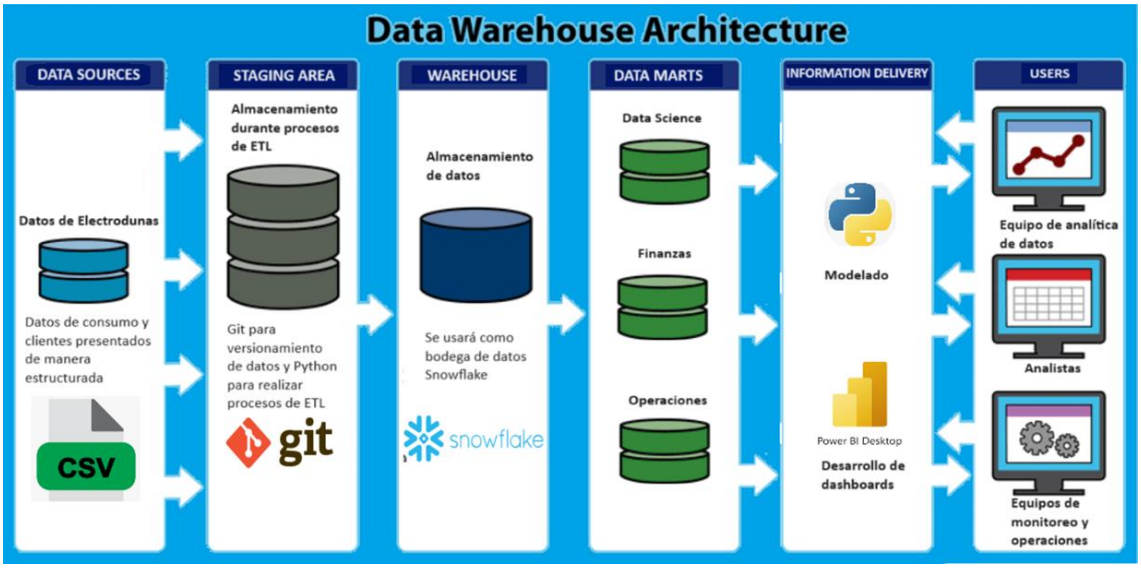
Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive

Diagrama Esquemático Propuesto

DIAGRAMA PROPUESTO SEMANA 1:

Inicialmente el equipo de trabajo propone una arquitectura simplificada, teniendo en cuenta el alcance del proyecto y las herramientas disponibles.

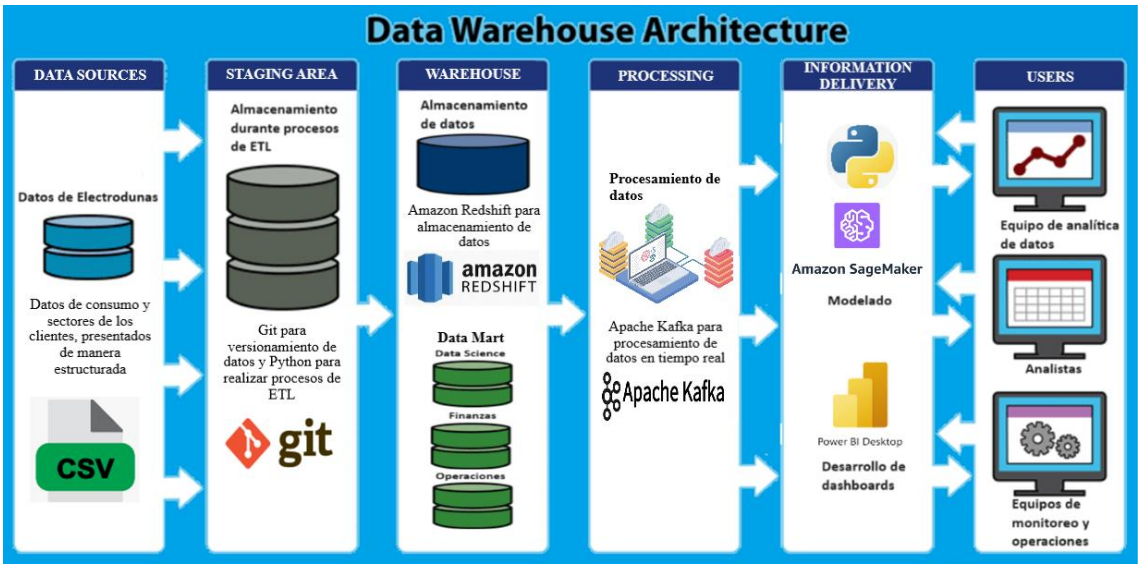


- **Data Sources**. Data de consumo por cliente, incluyendo Energía Activa, Energía Reactiva, Voltaje_FA y Voltaje_FC. Además, data de sectores económicos por cada cliente.
- **Staging Area**. Extraer archivos CSV desde Github o desde una base de datos si lo dispone la empresa. Versionamiento de datos utilizando git y procesamiento en área de staging de Python
- **Warehouse**. Transformaciones de datos utilizando Python, incluyendo técnicas de limpieza de datos (datos faltantes, inconsistencias, valores duplicados y datos atípicos), también agregar columna de mes y hora. Almacenamiento de datos en Bodega Snowflake.
- **Data Marts**. Sistema de almacenamiento de datos que contiene información específica de la unidad de negocio de una organización.
- **Information Delivery**. Modelado de data en Python y desarrollo del modelo. Publicación de resultados en un Dashboard que permita analizar y extraer información de la data.
- **Users**. Equipo de analítica de datos, analistas de pérdidas y equipo de monitoreo y operaciones.

DIAGRAMA PROPUESTO FINAL:

Por otra parte, el equipo propone una arquitectura más robusta que posiblemente puede implementar Electrodunas en un MVP2 considerando gran cantidad de data, clientes, lecturas y demás. Asimismo, un esquema que permite tomar ventaja de las herramientas disponibles en el mercado que facilitan el modelaje.

Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive



- **Data Sources**. Data de consumo por cliente, incluyendo Energía Activa, Energía Reactiva, Voltaje_FA y Voltaje_FC. Además, data de sectores económicos por cada cliente.
- **Staging Area**. Extraer archivos CSV desde Github o desde una base de datos si lo dispone la empresa. Versionamiento de datos utilizando git y procesamiento en área de staging de Python
- **Warehouse**.
  - Uso de Amazon Redshift para Almacenamiento de Datos pues permite almacenar grandes volúmenes de datos históricos y realizar consultas rápidas y eficientes. Ejemplo: Migrar la base de datos existente a Amazon Redshift para manejar datos de miles de clientes y años de datos históricos de consumo energético. Además, transformaciones de datos utilizando Python, incluyendo técnicas de limpieza de datos (datos faltantes, inconsistencias, valores duplicados y datos atípicos), también agregar columna de mes y hora. Almacenamiento de datos en Bodega.
  - En esta parte de la arquitectura, están también los **Data Marts** o el sistema de almacenamiento de datos que contiene información específica de la unidad de negocio de una organización.
- **Processing.** Implementación de Apache Kafka para Procesamiento en Tiempo Real que facilita el procesamiento y análisis de datos de consumo en tiempo real, permitiendo detectar anomalías casi instantáneamente. Ejemplo: Utilizar Apache Kafka para transmitir datos de medidores inteligentes en tiempo real y alimentar un sistema de detección de anomalías basado en machine learning.
- **Information Delivery**.
  - Modelado de data en Python y desarrollo del modelo. Publicación de resultados en un Dashboard que permita analizar y extraer información de la data. Despliegue de Modelos en AWS SageMaker porque simplifica el entrenamiento y despliegue de modelos de machine learning a gran escala, con capacidad para ajustar automáticamente los recursos necesarios. Ejemplo: Entrenar y desplegar modelos de detección de anomalías utilizando AWS SageMaker, permitiendo manejar grandes volúmenes de datos de consumo energético.
  - Escalabilidad de Visualizaciones con Power BI Premium que ofrece capacidades de almacenamiento y procesamiento adicionales en la nube, mejorando el rendimiento y la capacidad de los dashboards de Power BI. Ejemplo: Actualizar a Power BI Premium para manejar visualizaciones y análisis interactivos de datos de consumo energético de miles de clientes.
- **Users**. Equipo de analítica de datos, analistas de pérdidas y equipo de monitoreo y operaciones.

Aspecto	Requerimiento	Prueba Prevista	Criterio o métrica de evaluación y rangos deseados	Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA
Negocio				
R1	Obtener una comprensión más completa de los comportamientos de los clientes no regulados o libres.	Análisis descriptivo detallado de los clientes.	100% de los clientes deben estar cargados con su respectivo resumen descriptivo del comportamiento histórico.	100% de los clientes estan cargados, hay un resumen descriptivo del comportamiento histórico en el primer Dashboard y en la serie de tiempo es posible visualizar las cuatro variables estudiadas.
R2	Identificar anomalías asociadas a pérdidas no técnicas.	Análisis de clústeres.	Reconocer clústeres con comportamientos anómalos asociados a pérdidas no técnicas.	Dashboard con clústeres con comportamientos anómalos asociados a pérdidas no técnicas, teniendo en cuenta la regla creada que relaciona la Energía Activa y Reactiva.
R3	Comparar información de los clientes por sector que permita identificar tendencias.	Análisis descriptivo de los clientes por sector.	Presentar resumen descriptivo por sector.	Resumen descriptivo del comportamiento histórico en el primer Dashboard y en la serie de tiempo visualización de las cuatro variables estudiadas, con segmentación por sector.
Desempeño				
R4	Análisis de clústeres para identificar tendencias en las agrupaciones realizadas por el algoritmo y determinar qué factores influyen en la presencia de comportamientos anómalos.	Métrica de desempeño análisis de clústeres.	Coeficiente de Silhouette > 0.4. Esto sugiere que los clústeres están relativamente bien definidos y separados entre sí.	Coeficiente de Silhouette > 0.4. Esto sugiere que los clústeres están relativamente bien definidos y separados entre sí.
R5	Seleccionar mejor algoritmo que permita identificar anomalías.	Métricas de desempeño.	Coeficiente de Silhouette más cercano a 1 para algoritmos como PCA, LOF y DBSCAN.	Selección de modelo de aprendizaje no supervisado Isolation Forest, comparando Coeficiente de Silhoutte con otros modelos como Local Outlier Factor.
R6	Desarrollar un modelo de aprendizaje no supervisado que permita identificar anomalías.	Métricas basadas en la densidad.	Métrica de Densidad Local y Densidad Local Promedio.	Coeficiente de Silhouette más cercano a 1, desarrollo de algoritmo calibrado Isolation Forest para detección de anomalías.
Funcional				
R7	Construir un dashboard amigable con el usuario, con una interfaz clara y comprensible.	Demo a usuarios.	Dashboard que cumple con requerimientos.	Dashboard amigable y presentación en demo para los usuarios.
R8	Requiere licencia de Power BI.	Revisión.	Cumple.	Licencia Power BI, cumple.
R9	Cargar el dashboard rápidamente.	Pruebas de Performance.	Cargar front end en menos de 10 segundos.	Cargar front end en menos de 10 segundos.
R10	Aplicar filtros de fecha, hora, cliente y sector económico para visualizar subgrupos específicos.	Revisión.	Cumple.	Filtros disponibles para segmentar por clientes y sectores principalmente, cumple.

Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive

Universidad de los Andes

Explicación Tabla de Requerimientos

Requerimientos del Negocio:

- **R1. Obtener una comprensión más completa de los comportamientos de los clientes no regulados o libres.**
  - **Prueba Prevista.** Análisis descriptivo detallado de los clientes.
  - **Criterio o métrica de evaluación y rangos deseados.** 100% de los clientes deben estar cargados con su respectivo resumen descriptivo del comportamiento histórico.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** 100% de los clientes estan cargados, hay un resumen descriptivo del comportamiento histórico en el primer Dashboard y en la serie de tiempo es posible visualizar las cuatro variables estudiadas.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** La solución tiene cargados los 30 clientes, con toda la data asociada. En el Dashboard 1 hay también una tabla con el resumen descriptivo de consumos y el Dashboard 2 permite visualizar las cuatro variables estudiadas: Energía Activa, Energía Reactiva, Voltaje_FA y Voltaje_FC como una serie de tiempo.
  - **¿Qué ajustes fueron implementados para lograrlo?** Carga de toda la data, creación de visualizaciones que permiten observar el comportamiento histórico de cada variable de estudio y además, filtros que se pueden aplicar para segmentar la información.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** Si el cliente considera que es importante visualizar más detalles, sin tener en cuenta clústeres, es posible crear una tabla adicional con valores específicos de las cuatro variables de estudio relacionados con cuartiles para analizar de otra manera el comportamiento de clientes y sectores. Asimismo, es posible agregar diagramas de cajas y bigotes o frecuencias.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Agregar una hoja adicional con detalles específicos del análisis descriptivo como media, mediana, valor mínimo y valor máximo para cada cliente, y visualizarlo con información de lecturas con anomalías para facilitar el trabajo de los analistas de pérdidas.

- **R2. Identificar anomalías asociadas a pérdidas no técnicas.**
  - **Prueba Prevista.** Análisis de clústeres.
  - **Criterio o métrica de evaluación y rangos deseados.** Reconocer clústeres con comportamientos anómalos asociados a pérdidas no técnicas.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Dashboard con clústeres con comportamientos anómalos asociados a pérdidas no técnicas, teniendo en cuenta la regla creada que relaciona la Energía Activa y Reactiva.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** Definición de regla para anomalías considerando lecturas no típicas, específicamente casos donde la energía reactiva es significativamente mayor que la energía activa pues esperamos que la energía activa sea mayor o, al menos, no mucho menor que la energía reactiva. Después, desarrollo de un modelo de clustering para la segmentación de los clientes, elegido entre diferentes propuestas y generación de visualizaciones que permiten reconocer diferentes clústeres con sus características específicas.
  - **¿Qué ajustes fueron implementados para lograrlo?** Creación de regla para reconocer posibles anomalías teniendo en cuenta una relación entre la energía activa y reactiva.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** Esta regla es definida teniendo en cuenta el análisis descriptivo de la data y la bibliografía asociada al caso de estudio, por lo tanto, el analista de pérdidas debe analizar el prototipo para obtener insights.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** El modelo de clustering y la regla propuesta, no define con exactitud una anomalía pero si una posible lectura anómala que debe ser estudiada, es importante que el equipo de analítica reciba data adicional con una variable especifica de anomalías que permita desarrollar un

modelo más robusto.

- **R3. Comparar información de los clientes por sector que permita identificar tendencias.**

  - **Prueba Prevista.** Análisis descriptivo de los clientes por sector.
  - **Criterio o métrica de evaluación y rangos deseados.** Presentar resumen descriptivo por sector.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Resumen descriptivo del comportamiento histórico en el primer Dashboard y en la serie de tiempo visualización de las cuatro variables estudiadas, con segmentación por sector.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** En el Dashboard 1 hay un análisis descriptivo con información de clústeres definidos teniendo en cuenta la regla que relaciona Energía Activa y Reactiva planteada anteriormente, asimismo presentación de serie de tiempo de las cuatro variables de estudio. Todo esto, con la posibilidad de la segmentación por sector por medio de filtros que permite un análisis diferencial para identificar tendencias.
  - **¿Qué ajustes fueron implementados para lograrlo?** Principalmente, aplicación de filtros que permiten la segmentación de la data.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** Es posible segmentar todas las visualizaciones por sectores y además por clientes dentro de ese sector, sin embargo, si el usuario desea un análisis diferente con anomalías reconocidas, es necesario obtener la data adicional.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Agregar label de anomalías con data adicional para hacer un análisis especifico de anomalías reales por sector para definir nuevas reglas en el modelo.

Requerimientos de Desempeño:

- **R4. Análisis de clústeres para identificar tendencias en las agrupaciones realizadas por el algoritmo y determinar qué factores influyen en la presencia de comportamientos anómalos.**

  - **Prueba Prevista.** Métrica de desempeño análisis de clústeres.
  - **Criterio o métrica de evaluación y rangos deseados.** Coeficiente de Silhouette > 0.4.
  - Esto sugiere que los clústeres están relativamente bien definidos y separados entre sí.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Coeficiente de Silhouette > 0.4. Esto sugiere que los clústeres están relativamente bien definidos y separados entre sí.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** Desarrollo de diferentes modelos de clustering como DBSCAN y KMeans para segmentar los datos y detectar posibles anomalías. Para empezar con este modelo, definimos una regla de aplicación que considera la información relevante que encontramos durante el análisis de la base de datos y que posiblemente representa una anomalía relacionando la Energía Activa y Reactiva. El análisis de resultados para seleccionar el mejor algoritmo de clustering fue realizado comparando los diferentes valores del Coeficiente de Silhoutte, en este caso, KMeans resulta en un coeficiente igual a 0.6.
  - **¿Qué ajustes fueron implementados para lograrlo?** El modelo fue calibrado y el resultado sugiere la segmentación de la data en 4 clústeres.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** En caso de que el cliente sugiera una regla o agrupación diferente, el equipo deberá desarrollar un nuevo modelo.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Incluir los labels de las diferentes lecturas para crear un modelo que permita agrupar teniendo en cuenta esas anomalías ya probadas.

- **R5. Seleccionar mejor algoritmo que permita identificar anomalías.**

  - **Prueba Prevista.** Métricas de desempeño.
  - **Criterio o métrica de evaluación y rangos deseados.** Coeficiente de Silhouette más

cercano a 1 para algoritmos como PCA, LOF y DBSCAN.

- **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Selección de modelo de aprendizaje no supervisado Isolation Forest, comparando Coeficiente de Silhoutte con otros modelos como Local Outlier Factor.
- **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** Desarrollo y calibración de diferentes modelos para detectar anomalías, incluyendo PCA, Local outlier Factor, DBSCAN, One Class SVM y también Isolation Forest. Después, análisis de resultados comparando Índice Davies-Bouldin y Coeficiente de Silhoutte.
- **¿Qué ajustes fueron implementados para lograrlo?** Investigación de diferentes modelos para detectar anomalías en una serie de tiempo considerando outliers, además, desarrollo, calibración, implementación y evaluación de los diferentes algoritmos para seleccionar el mejor para la data de estudio.
- **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** Si las posibles anomalías detectadas nos son positivos verdaderos, es necesario recalibrar el modelo con los labels reales.
- **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Incluir labels reales que debe proporcionar Electrodunas, para calibrar el modelo.

- **R6. Desarrollar un modelo de aprendizaje no supervisado que permita identificar anomalías.**

  - **Prueba Prevista.** Métricas basadas en la densidad.
  - **Criterio o métrica de evaluación y rangos deseados.** Métrica de Densidad Local y Densidad Local Promedio.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Coeficiente de Silhouette más cercano a 1, desarrollo de algoritmo calibrado Isolation Forest para detección de anomalías.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** Las métricas arrojadas por los algoritmos son consideradas aceptables para el clustering, por KMeans el coeficiente de silueta resultó en 0.6 y para el algoritmo de Isolation Forest 0.37, lo que indica una separación equitativa entre los grupos.
  - **¿Qué ajustes fueron implementados para lograrlo?** Fue necesario calibrar hiperparámetros en ambos algoritmos, como el número de clústeres, el método de inicialización y número de estimadores.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** La presencia de outliers o falta de data que diferencie los grupos son posibles causas de un mal agrupamiento, por lo que esto deberá ser tratado en la etapa de procesamiento de datos y socializado con el cliente, ya que mediciones incorrectas afectarán el desempeño de los algoritmos.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Recolectar más datos de los clientes hará que los algoritmos tengan más información y por ende brinden mejores resultados. También recomendamos aclaración en cuanto a lo que es considerado una anomalía, ya que los criterios pueden ser distintos entre ambas partes.

Requerimientos Funcionales:

- **R7. Construir un dashboard amigable con el usuario, con una interfaz clara y comprensible.**

  - **Prueba Prevista.** Demo a usuarios.
  - **Criterio o métrica de evaluación y rangos deseados.** Dashboard que cumple con requerimientos.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Dashboard amigable y presentación en demo para los usuarios.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** La selección de gráficos e indicadores fue pensada para un público no experto en el tema, el cual fácilmente podrá interpretar los resultados del modelo y utilizar el dashboard de una manera simple e intuitiva.
  - **¿Qué ajustes fueron implementados para lograrlo?** Fue planteado un modelo relacional que permite mostrar la data de manera sencilla y seleccionamos gráficas de uso común, como gráficas de dispersión y líneas temporales de fácil interpretación.

- **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** El equipo deberá replantear el diseño del dashboard en caso de que el cliente no lo considere amigable y práctico para operación.
- **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Sugerimos al cliente brindar constante feedback de la herramienta para poder generar mejoras acorde a sus necesidades.

- **R8. Requiere licencia de Power BI.**

  - **Prueba Prevista.** Revisión.
  - **Criterio o métrica de evaluación y rangos deseados.** Cumple.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Licencia Power BI, cumple.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** La licencia de Power BI es un requerimiento que el cliente debe adquirir, su precio oscila entre 10USD y 20 USD mensuales por usuario.
  - **¿Qué ajustes fueron implementados para lograrlo?** En este caso, la Universidad proporciona esta licencia, por lo que no fue requirido ningún proceso externo para desarrollar el proyecto.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** En caso de que Power BI no cumpla con los requerimientos del cliente, es indisppensable evaluar otra herramienta de visualización, como Looker Studio o Tableau
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Contabilizar el número de usuarios que usaran la aplicación para no incurrir en sobrecostos y definir si Power BI cumple con los requerimientos operacionales.

- **R9. Cargar el dashboard rápidamente.**

  - **Prueba Prevista.** Pruebas de Performance.
  - **Criterio o métrica de evaluación y rangos deseados.** Cargar front end en menos de 10 segundos.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Cargar front end en menos de 10 segundos.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** El dashboard fue disponibilizado en el siguiente link, donde su tiempo de carga una vez logueado es inferior a 10 segundos.
  - **¿Qué ajustes fueron implementados para lograrlo?** La implementación de gráficas funcionales y el correcto planteamiento del modelo relacional hace que los tiempos de carga y actualización sean reducidos, por lo que el correcto planteamiento de esto es la clave.
  - **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** Es posible que el requerimiento no sea cumplido debido a factores externos, como conexión a internet o velocidad de los dispositivos donde es ejecutado el dashboard, por lo que esto deberá aclararse en el momento de su disponibilización.
  - **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** En caso de contemplarse la inclusión de más datos o elementos en el tablero, el cliente deberá considerar la afectación de tiempos de carga y funcionamiento.

- **R10. Aplicar filtros de fecha, hora, cliente y sector económico para visualizar subgrupos específicos.**

  - **Prueba Prevista.** Revisión.
  - **Criterio o métrica de evaluación y rangos deseados.** Cumple.
  - **Criterio o métrica de evaluación y rangos deseados, SOLUCIÓN ENTREGADA.** Filtros disponibles para segmentar por clientes, fecha y sectores principalmente, cumple.
  - **¿Cómo se satisfacen cada uno de los requerimientos y métricas de evaluación definidos?** Fueron integraron slicers en el tablero que permiten la segmentación según los requerimientos definidos.
  - **¿Qué ajustes fueron implementados para lograrlo?** La herramienta de Power BI cuenta

con la funcionalidad de agregar filtros en el tablero, por lo que únicamente fue necesario plantear un modelo relacional de datos de la manera correcta para obtener un filtrado preciso.

- **¿Cómo justificar los casos en que no se satisfaga algún requerimiento?** En caso de que la segmentación no cumpla con los requerimientos, el equipo deberá indagar si las gráficas utilizadas son las correctas, o si el requerimiento del cliente es realizable en el tablero, ya que las herramientas de visualización tienen sus limitaciones.
- **¿Qué acciones correctivas se proponen al cliente para esta o próximas iteraciones?** Recolectar más campos de información de los clientes permitirá realizar un filtrado más preciso y llegar a conclusiones más asertivas.

Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive

Universidad de los Andes

Reporte técnico de selección e implementación de modelos

ANÁLISIS DE DATOS

Características de calidad de los datos

Totalidad de los datos:

La base de datos está conformada por 30 (treinta) archivos que contienen información de clientes, fecha de datos, valores históricos de energía activa entregada (kWh), energía reactiva entregada (kVarh), voltaje y uno adicional con sector de clientes. Estos archivos fueron consolidados, y durante la exploración de datos, identificamos 463,425 registros distribuidos en 7 columnas, Ilustración 1.

	Fecha	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC	NumeroCliente	Sector Económico:
0	2021-01-01 00:00:00	0.357841	0.282788	455.139171	510.561002	Cliente 1	Elaboración de cacao y chocolate y de producto...
1	2021-01-01 01:00:00	0.372264	0.431377	469.978787	469.917178	Cliente 1	Elaboración de cacao y chocolate y de producto...
2	2021-01-01 02:00:00	1.044687	0.338626	468.721120	546.949147	Cliente 1	Elaboración de cacao y chocolate y de producto...
3	2021-01-01 03:00:00	0.566425	0.495791	452.329255	444.122989	Cliente 1	Elaboración de cacao y chocolate y de producto...
4	2021-01-01 04:00:00	1.080556	0.472018	513.477596	535.463719	Cliente 1	Elaboración de cacao y chocolate y de producto...
...	...	...	...	...	...	...	...
463420	2022-04-21 10:00:00	0.960105	0.473234	1273.150602	1027.084539	Cliente 30	Venta al por mayor de metales y minerales meta...
463421	2022-04-21 11:00:00	0.624300	0.699936	1063.524968	1205.829819	Cliente 30	Venta al por mayor de metales y minerales meta...
463422	2022-04-21 12:00:00	0.985633	0.123560	1207.284283	1127.893714	Cliente 30	Venta al por mayor de metales y minerales meta...
463423	2022-04-21 13:00:00	0.710436	0.399262	1205.012971	1090.835898	Cliente 30	Venta al por mayor de metales y minerales meta...
463424	2022-04-21 14:00:00	1.169279	0.013737	1145.134963	1129.472067	Cliente 30	Venta al por mayor de metales y minerales meta...

463425 rows × 7 columns

Ilustración 1. Cantidad de filas y columnas. Electrodunas

Realizamos una evaluación de la totalidad de los datos que incluyó la preparación de estos y la identificación de posibles valores faltantes. **Es importante destacar que, tras esta evaluación, no fueron encontrados datos nulos.** La data comprende fechas entre 2021-01-01 hasta 2023-04-01 y cuenta con información de las 24 horas del día en la mayor parte del rango temporal, contiene información de 30 clientes y 7 sectores económicos.

Consistencia:

La energía activa promedio, Ilustración 2, que representa la cantidad de energía eléctrica real entregada en un período de tiempo específico, es igual a 1.47 kWh y, por otra parte, la energía reactiva que indica la cantidad de energía intercambiada entre la fuente de energía y una carga sin realizar trabajo útil, a 0.87 kVARh. La variable energía activa está dentro del rango -1.33 y 14.62 kWh, lo cual podría indicar la presencia de datos atípicos o errores en la medición que merecen una atención especial, la desviación estándar es alta, sugiriendo una dispersión significativa en los datos. En el caso de la energía reactiva, identificamos un valor mínimo de 0 y un valor máximo de 11.14 kVARh. De nuevo la desviación estándar es elevada, indicando otra vez una dispersión significativa.

Los voltajes muestran comportamientos similares, con un rango entre 0 y 4399, desviación estándar parecida para FA y FC.

Es importante mencionar que no encontramos valores duplicados en el dataset, y la presencia de outliers fue revisada en los algoritmos y los procedimientos empleados más adelante.

	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC
count	463425.000000	463425.000000	463425.000000	463425.000000
mean	1.472050	0.873086	1420.188470	1438.515836
std	1.718780	1.158846	766.299118	746.447449
min	-1.329018	0.000000	0.031000	0.031000
25%	0.242788	0.112832	719.462874	748.148254
50%	0.810771	0.380650	1625.493463	1634.696089
75%	1.992488	1.222834	2037.276385	2040.554497
max	14.622644	11.135141	4266.229746	4399.038932

Ilustración 2. Estadísticas descriptivas

Para visualizar la distribución cuantitativa de los datos de una manera que facilita la comparación entre las variables, realizamos boxplots. En este caso, podemos confirmar algunos outliers o valores atípicos que son exactamente puntos aislados en la distribución para la energía activa y reactiva.

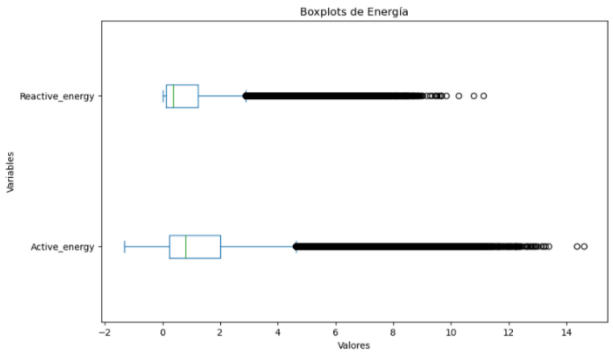


Ilustración 3. Boxplot Energía

Claridad:

Consideramos que los datos son claros, pues su estructura tabular permite el fácil manejo e integración. No encontramos inconsistencias en la presentación de los datos, como por ejemplo nombres ilegibles o unidades de medida diferentes entre las observaciones/variables. Los datos suministrados explican el contexto del cliente y su problemática. De igual manera, no fue necesario realizar algún tipo de estandarización, Electrodunas entregó también las definiciones de todas las variables para contextualizar la información apropiadamente.

Formato:

La columna fecha está en formato datetime. Las energías y voltajes son variables numéricas en formato float, específicamente, la energía activa representa kWh, la energía reactiva kVARh y los voltajes en sus dos fases voltios. Las variables NumeroCliente y Sector Económico son variables categóricas en un tipo de dato string con información cualitativa del cliente, es decir, los datos son coherentes para el proyecto.

Concordancia con el problema de negocio:

A nivel general, los datos proporcionados por Electrodunas son confiables y contienen información relevante para el caso de estudio. El suministro de la información en horas permite tener una granularidad mayor pues tenemos detalles más específicos sobre el consumo que ayudaría a la precisión en los análisis de los datos, por eso abordar la problemática con esta data sería factible para obtener resultados satisfactorios en el desarrollo del proyecto.

Limpieza de datos

Ahora, trabajamos en la limpieza de datos para que sea posible utilizarlos en el proyecto porque deben cumplir con los tres aspectos fundamentales: exactos, procesables y ágiles. Todo esto, para asegurar la calidad de los datos y minimizar el riesgo de tener data poco precisa o errónea en los entregables a la compañía. Para cumplir con el propósito, evaluamos:

- **Datos faltantes**: En el análisis preliminar, encontramos que no hay valores faltantes en la base de datos.
- **Inconsistencias**: Previamente evaluamos la calidad de los datos, en la parte de consistencia, estudiamos la validez, integridad y definición en su estructura. De manera general, consideramos que los formatos y tipos de datos son correctos.
- **Valores duplicados:** En la base de datos combinada, no hay información duplicada que pueda sesgar el análisis de la información proporcionada por la empresa.
- **Datos atípicos:** En cuanto a outliers o valores que pueden afectar la distribución de los datos porque son distantes de los demás, encontramos que la variable energía activa tiene un porcentaje de valores negativos de 0.1%, entonces, esto es posiblemente una anomalía que más adelante será considerada.

	Fecha	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC	NumeroCliente	Sector Económico:
291228	2021-11-09 15:00:00	-0.381904	0.382493	2077.430688	2093.655684	Cliente 17	Captación, tratamiento y distribución de agua
291229	2021-11-09 16:00:00	-0.408740	0.469523	2036.334551	2093.205889	Cliente 17	Captación, tratamiento y distribución de agua
291237	2021-11-10 00:00:00	-0.187803	0.392220	1950.083873	1951.344805	Cliente 17	Captación, tratamiento y distribución de agua
291238	2021-11-10 01:00:00	-0.455174	0.459796	1931.903407	1858.665907	Cliente 17	Captación, tratamiento y distribución de agua
291239	2021-11-10 02:00:00	-0.338941	0.390148	1910.706034	2064.042938	Cliente 17	Captación, tratamiento y distribución de agua
...	...	...	...	...	...	...	...
354069	2022-05-09 12:00:00	-0.457582	0.200000	2069.593516	2077.851435	Cliente 20	Captación, tratamiento y distribución de agua
354070	2022-05-09 13:00:00	-0.458964	1.484933	2078.795633	2073.401393	Cliente 20	Captación, tratamiento y distribución de agua
354071	2022-05-09 14:00:00	-0.427294	0.231000	2071.330458	2079.444732	Cliente 20	Captación, tratamiento y distribución de agua
354072	2022-05-09 15:00:00	-0.455775	0.200000	2039.538550	2088.843514	Cliente 20	Captación, tratamiento y distribución de agua
354073	2022-05-09 16:00:00	-0.420409	0.495809	2057.675348	1993.268293	Cliente 20	Captación, tratamiento y distribución de agua

505 rows × 7 columns

Ilustración 4. Valores negativos Energía Activa

Finalmente, creamos la columna mes y hora, para tener una mayor granularidad en la data pues será posible realizar análisis sobre estas variables para identificar comportamientos de interés para el caso de estudio, Ilustración 5.

	Fecha	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC	NumeroCliente	Sector Económico:	Mes	Hora
0	2021-01-01 00:00:00	0.357841	0.282788	455.139171	510.561002	Cliente 1	Elaboración de cacao y chocolate y de producto...	1	0
1	2021-01-01 01:00:00	0.372264	0.431377	469.978787	469.917178	Cliente 1	Elaboración de cacao y chocolate y de producto...	1	1
2	2021-01-01 02:00:00	1.044687	0.338626	468.721120	546.949147	Cliente 1	Elaboración de cacao y chocolate y de producto...	1	2
3	2021-01-01 03:00:00	0.566425	0.495791	452.329255	444.122989	Cliente 1	Elaboración de cacao y chocolate y de producto...	1	3
4	2021-01-01 04:00:00	1.080556	0.472018	513.477596	535.463719	Cliente 1	Elaboración de cacao y chocolate y de producto...	1	4
...	...	...	...	...	...	...	...	...	...
463420	2022-04-21 10:00:00	0.960105	0.473234	1273.150602	1027.084539	Cliente 30	Venta al por mayor de metales y minerales meta...	4	10
463421	2022-04-21 11:00:00	0.624300	0.699936	1063.524968	1205.829819	Cliente 30	Venta al por mayor de metales y minerales meta...	4	11
463422	2022-04-21 12:00:00	0.985633	0.123560	1207.284283	1127.893714	Cliente 30	Venta al por mayor de metales y minerales meta...	4	12
463423	2022-04-21 13:00:00	0.710436	0.399262	1205.012971	1090.835898	Cliente 30	Venta al por mayor de metales y minerales meta...	4	13
463424	2022-04-21 14:00:00	1.169279	0.013737	1145.134963	1129.472067	Cliente 30	Venta al por mayor de metales y minerales meta...	4	14

462920 rows × 9 columns

Ilustración 5. Cantidad de filas y columnas dataset

Entendimiento de los datos

Dentro del análisis realizado, evaluamos las siguientes técnicas para el entendimiento de los datos:

- **Filtración**: Filtramos las observaciones que tenían energía activa con valores menores a 0. Esto no tiene sentido físicamente pues la energía activa es la energía consumida en un tiempo específico.
- **Imputación**: No fue necesario imputar valores faltantes ya que los datos suministrados presentaron total completitud en sus valores.
- **Reducción de dimensión:** Dependiendo del tipo de modelo utilizado, una reducción de dimensión puede ser requerida, sobre todo en las variables de los voltajes ya que contienen una correlación de 95%. Sin embargo, las dimensiones de la data son relativamente pequeñas para pensar en una reducción dimensional considerable.
- **Extracción de variables explicativas o features:** En caso de tener una variable de respuesta, intentaríamos evaluar el problema mediante técnicas de aprendizaje supervisado. Sin embargo, al

carecer de esta vamos a abordar el problema mediante técnicas de aprendizaje no supervisado. Por lo que la inclusión de variables dummies y la estandarización o normalización del resto de variables dependerá de los requerimientos del algoritmo implementado.

- **Descripción de estadísticas básicas:** En la evaluación anterior de **Características de calidad de los datos**, analizamos las estadísticas básicas de la data para entender el comportamiento y distribución de la misma.
- **Segmentación de datos para encontrar patrones:** Esta técnica es considerada fundamental y la base del desarrollo del modelo, pues lo que buscamos es identificar tendencias en los consumos de los clientes (o clientes entre sí) para identificar posibles comportamientos anómalos y responder a los requerimientos de Electrodunas.
- **Determinación de las variables más importantes:** El equipo considera que la data es limitada y, por lo tanto, muy importante en la evaluación. Encontrar relaciones entre las variables podría ayudar en la implementación del proyecto, por eso, después del entendimiento de los datos decidimos que todas las variables son indispensables.
- **Clasificación o agrupación:** Es posible clasificar los clientes según su sector económico para obtener un mejor entendimiento de los consumos y voltajes, pues esta variable podría dar indicios de ubicaciones de los clientes y las regulaciones en su sector.

Mediante estadística básica, evaluamos información fundamental como valores mínimos y máximos, así como la desviación estándar para la toma de decisiones e identificación de posibles datos atípicos para garantizar la fiabilidad y precisión de los resultados.

Al examinar la distribución del sector económico entre los clientes de Electrodunas en la Ilustración 6, observamos que el 21,3% está vinculado a la elaboración de cacao, chocolate y productos de confitería, así como al cultivo de árboles frutales y nueces. Además, un significativo 21% está involucrado en actividades relacionadas con la captación, tratamiento y distribución de agua.

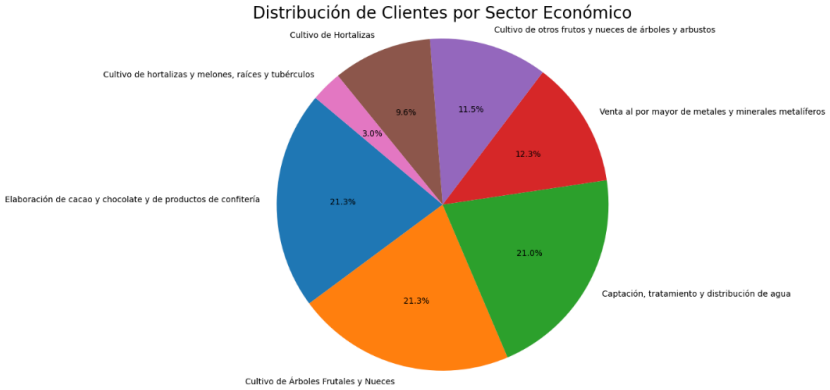


Ilustración 6. Clientes por sector económico

En el histograma de Energía Activa de la Ilustración 7 podemos ver que los consumos están centrados por debajo de los 4 kWh.

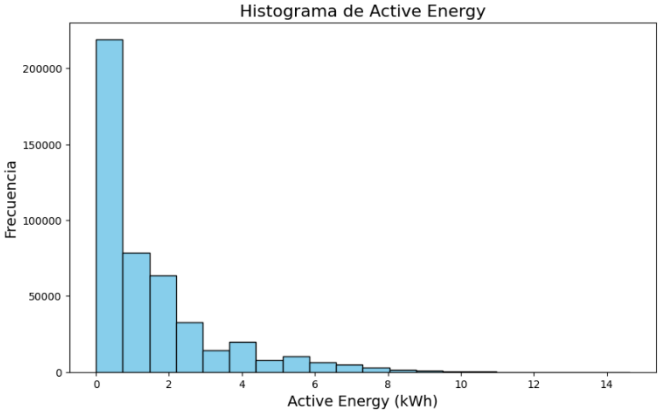


Ilustración 7. Histograma Energía Activa

Al verificar las variables de consumo de energía, encontramos diferencias en el comportamiento mensual de la energía activa, Ilustración 8. El primer y último trimestre tienen valores más altos de energía activa promedio.

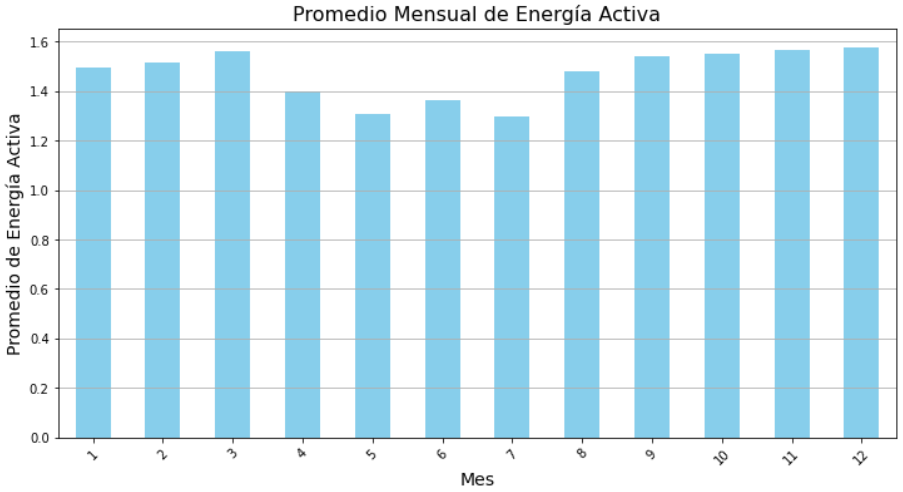


Ilustración 8. Promedio mensual Energía Activa

Asimismo, es posible visualizar la energía activa promedio mensual por sectores económicos. El sector de Elaboración de cacao, chocolate y productos de confitería, y el Cultivo de Árboles Frutales y Nueces, tienen consumos más altos de energía activa y seguramente por épocas de producción hay mayores consumos en ciertos meses (primer y último trimestre como mencionamos anteriormente, Ilustración 9.

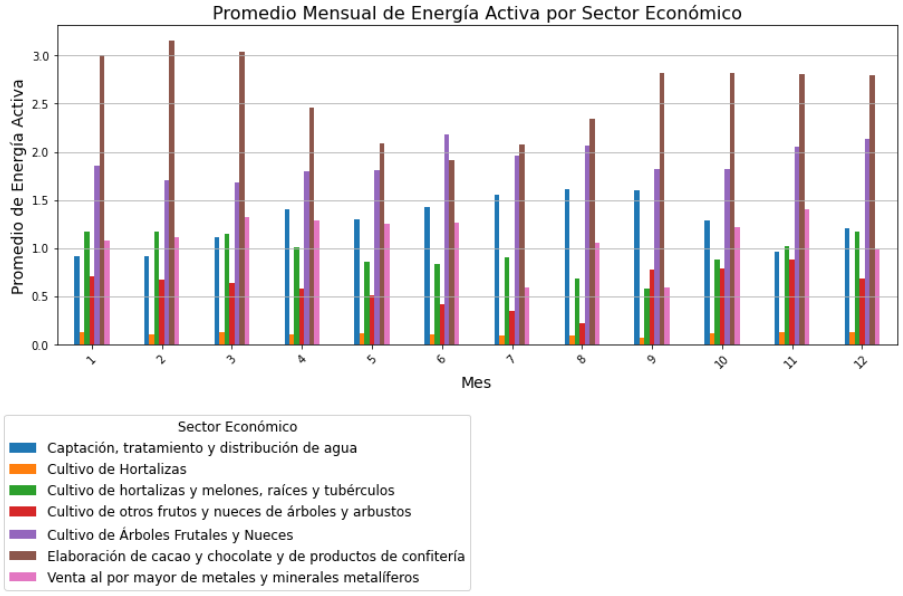


Ilustración 9. Promedio mensual Energía Activa por Sector Económico

Para la energía reactiva, el promedio es mayor en el mes de septiembre, Ilustración 10.

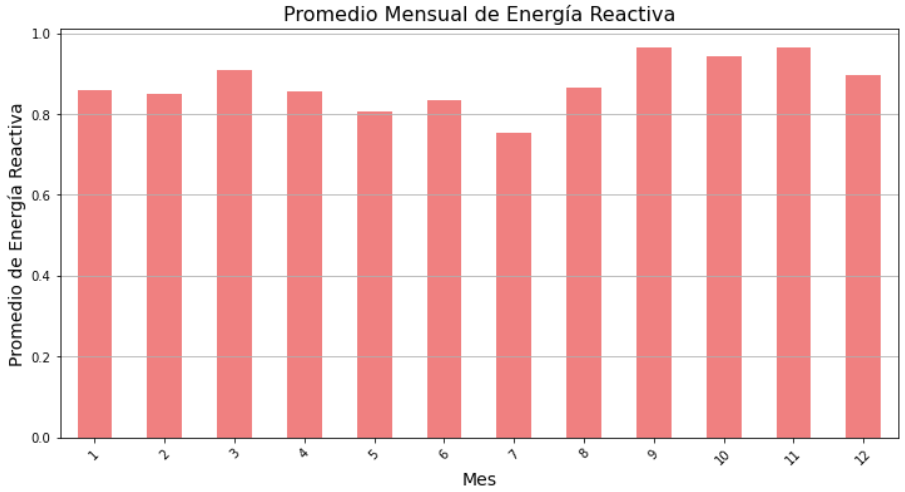


Ilustración 10. Promedio mensual Energía Reactiva

En la Ilustración 11, encontramos relaciones entre los voltajes FA Y FC del 95%, teniendo en cuenta que tenemos un sistema bifásico con fases A y C.

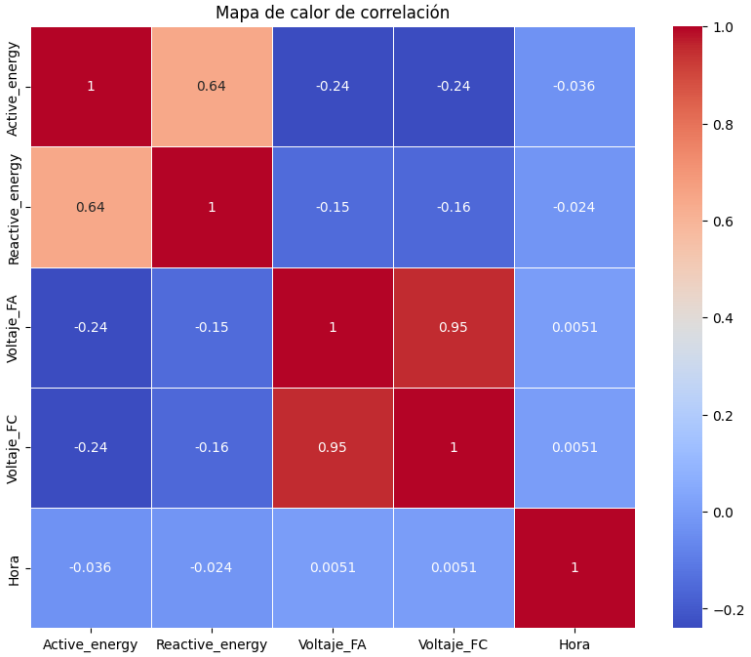


Ilustración 11. Mapa de calor de correlación

Después de estudiar la data, podemos concluir que hay información variada que permite estudiar a los clientes no regulados y también, estudiar patrones de comportamiento y anomalías asociadas a perdidas no técnicas. Hay algunos outliers y en la evaluación de modelos, hay un análisis más profundo pues seguramente contienen información de comportamientos anómalos. Además, encontramos que un enfoque por sectores puede arrojar información importante en la resolución del problema de analytics:

El desarrollo de la implementación de analítica de datos para la detección de anomalías para la toma de decisiones en Electrodunas, permitirá la identificación de posibles clientes con comportamientos anormales, además contribuirá a la toma de decisiones para la reducción en las pérdidas no técnicas, aumento de los ingresos de la empresa y eficiencia en los costos. Al final, la compañía tendrá un Dashboard con un análisis del comportamiento de los clientes con datos históricos de las variables para cada cliente, resumen descriptivo por cliente, consumos posiblemente anómalos que no coincidan con los consumos esperados dado el comportamiento histórico, en una interfaz amigable para los usuarios.

Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive

Universidad de los Andes

PREPARACIÓN DE LOS DATOS

La preparación adecuada de los datos es fundamental antes de aplicar cualquier modelo de machine learning, y en particular para la detección de anomalías, por varias razones:

1. Conversión de Fechas.

   - Indexación Temporal: Las series de tiempo requieren que las fechas sean interpretadas correctamente como índices temporales para realizar análisis en función del tiempo. Esto permite operaciones como agrupación por periodos (horas, días, meses), cálculos de tendencias o estacionalidades, y alineación con otras series de tiempo.
   - Funcionalidad: Muchas bibliotecas de análisis de datos y machine learning (como pandas y scikit-learn) requieren que las columnas de fecha estén en formatos específicos (como objetos datetime de Python) para manejarlas adecuadamente, facilitando así el filtrado, la visualización y el modelado.

2. Normalización de Características Numéricas.

   - Escalado Uniforme: Algoritmos como PCA y muchos métodos de clustering, incluidos DBSCAN y técnicas basadas en vecinos más cercanos como LOF, asumen que todas las características numéricas contribuyen equitativamente. Sin normalización, una característica con valores numéricos grandes dominaría las otras características, sesgando los resultados del modelo.
   - Mejora del Rendimiento: La normalización ayuda a mejorar la convergencia de muchos algoritmos de machine learning, reduciendo la posibilidad de que el modelo sea inestable o converja a un mínimo local no óptimo.
   - Comparabilidad: Permite comparar diferentes variables en un mismo plano de análisis. Por ejemplo, al comparar el consumo de energía activa (medido en kWh) con el voltaje (en voltios), la normalización escala estos valores a un rango común, haciendo sus variaciones comparables en términos relativos.

3. Generalización de los Modelos.

   - Reducción de Sesgos: La normalización asegura que el modelo no sea injustamente sesgado hacia características con rangos de valores más amplios.
   - Adaptabilidad: Un modelo entrenado sobre datos normalizados generalmente es más adaptable y robusto, lo que es crucial cuando los datos futuros pueden tener características numéricas que varíen en magnitudes diferentes a las del conjunto de datos de entrenamiento.

Tiene sentido normalizar los datos incluso en el sector de detección de anomalías en clientes de energía eléctrica, especialmente cuando utilizamos técnicas de aprendizaje automático que son sensibles a la escala de los datos. La normalización ayuda a asegurar que cada característica contribuya equitativamente al análisis, sin que una sobresalga injustamente debido a su escala.

Visualización de Datos Normalizados en Power BI.

Aunque la normalización es útil para el análisis y la modelación, para la visualización en un dashboard, especialmente en un contexto empresarial donde los datos crudos tienen un significado concreto (como kWh en el consumo de energía), es importante mostrar los datos en su forma original o en una forma que sea intuitiva para los usuarios finales:

1. Almacenar Ambas Versiones de los Datos: Al preparar los datos para el análisis, podemos considerar almacenar tanto los valores originales como los normalizados. De esta manera, podemos utilizar los datos normalizados para la detección de anomalías y los datos originales para la visualización.
2. Transformación Dinámica en Power BI: También podemos utilizar DAX (Data Analysis Expressions) en Power BI para revertir la normalización si es necesario. Por ejemplo, si hemos aplicado una normalización estándar (restar la media y dividir por la desviación estándar), es posible almacenar la media y la desviación estándar de cada característica y usar estas para transformar los valores normalizados de vuelta a su escala original mediante una fórmula en Power BI.
3. Visualización de Anomalías con Datos Originales: Configurar las visualizaciones en el dashboard

para utilizar los valores originales y mostrar métricas como el consumo de energía, mientras que las etiquetas o colores en el gráfico pueden reflejar el resultado del análisis de anomalías (por ejemplo, puntos coloreados de manera diferente si es detectada una anomalía).

4. Detalles al Hacer Clic: Una técnica útil es configurar las visualizaciones para que muestren más detalles al hacer clic en un punto específico. Por ejemplo, al seleccionar un punto de datos en el tiempo que ha sido identificado como una anomalía y de esta manera proporcionar un contexto completo al usuario.

Entender el Origen de los Valores Negativos

En el contexto de la detección de anomalías, no siempre es necesario eliminar los valores negativos directamente, ya que estos pueden ser comportamientos anómalos o errores en los datos que son precisamente lo que el modelo busca identificar.

- **Errores de Medición o de Datos:** Si los valores negativos son el resultado de errores de medición o de entrada de datos, corregir estos errores o eliminar las entradas incorrectas puede ser necesario para mantener la integridad del análisis. Sin embargo, después del análisis realizado, esto no será aplicado en este caso.
- **Comportamientos Anómalos Reales:** Si los valores negativos reflejan un comportamiento anómalo legítimo, como el robo de energía (donde el medidor podría mostrar consumos negativos debido a manipulaciones), entonces estos datos son valiosos para el modelo y deben ser conservados. En el problema planteado y específicamente después del estudio previo de la información disponible, esta será la elección pues toda la data es significativa.

Agregar columnas adicionales

Agregar columnas adicionales como el año, mes, día de la semana y hora puede ser muy útil tanto para la elaboración de modelos de machine learning como para la visualización, especialmente en un contexto de análisis de series temporales como el de la detección de anomalías en el consumo de energía eléctrica.

1. Mejora de Modelos de Machine Learning

   - **Patrones Estacionales y Tendencias:** Muchos comportamientos de consumo están fuertemente influenciados por patrones estacionales o diarios. Por ejemplo, el consumo energético puede variar significativamente entre días laborables y fines de semana, o entre diferentes meses del año debido a condiciones climáticas. Identificar estas tendencias puede mejorar significativamente la precisión de los modelos.
   - **Feature Engineering:** Al desglosar la fecha y hora en componentes más granulares (como hora del día o día de la semana), podemos capturar efectivamente estas variabilidades en los modelos. Esto puede ayudar a ajustar las predicciones del modelo para reflejar patrones de uso típicos o atípicos.
   - **Detección de Anomalías Contextual:** En la detección de anomalías, entender el contexto temporal puede ser crucial. Un patrón de consumo que es normal durante el día puede ser anómalo durante la noche, y viceversa.

2. Visualización de Datos

   - **Interpretación Mejorada:** Las visualizaciones que incluyen desgloses por año, mes, día de la semana, hora, pueden ayudar a los usuarios a interpretar mejor los datos. Por ejemplo, permitirían identificar si las anomalías tienden a ocurrir en ciertos momentos específicos.

   - **Interacción del Usuario:** En herramientas como Power BI, que fue elegida para construir el dashboard, permitir a los usuarios filtrar o desglosar datos basados en estos atributos temporales puede mejorar significativamente la experiencia del usuario, permitiéndoles explorar los datos de manera más intuitiva y significativa.

3. Consideraciones Finales

   - **Selección de Características:** Aunque añadir características puede ser muy beneficioso, es

importante evaluar su impacto real en el modelo. Técnicas como el análisis de importancia de características o la eliminación recursiva de características pueden ayudar a determinar si realmente mejoran el modelo. Sin embargo, hay pocas variables y por eso todas son importantes para el modelo.

- **Overfitting:** Debemos tener cuidado para evitar el overfitting, especialmente cuando se añaden muchas características nuevas. Hay que asegurar que el modelo generalice bien los datos no vistos.

Tratamiento de los Datos:

Después del análisis anterior, el tratamiento de los datos definido previo a la implementación de modelos, comprende:

1. Normalizar los datos numéricos.
2. Convertir la columna Fecha a tipo datetime.
3. Agregar columnas adicionales: Año, Mes, Día, Hora, Día_semana (0 a 6, domingo).
4. Calcular semana del mes.
5. Establecer fecha como índice.

```python
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import StandardScaler

# Normalizar los datos numéricos
scaler = StandardScaler()
df[['Active_energy_N', 'Reactive_energy_N', 'Voltaje_FA_N', 'Voltaje_FC_N']] = scaler.fit_transform(
    df[['Active_energy', 'Reactive_energy', 'Voltaje_FA', 'Voltaje_FC']])

# Convertir la columna 'Fecha' a datetime
df['Fecha'] = pd.to_datetime(df['Fecha'])

# Columnas adicionales
df['Año'] = df['Fecha'].dt.year
df['Mes'] = df['Fecha'].dt.month
df['Día'] = df['Fecha'].dt.day
df['Hora'] = df['Fecha'].dt.hour
df['Día_semana'] = df['Fecha'].dt.dayofweek  # 0 es lunes, 6 es domingo

# Calcular la semana del mes
# Se calcula dividiendo el día del mes por 7 y redondeando hacia arriba
df['Semana_del_mes'] = ((df['Fecha'].dt.day - 1) // 7 + 1)

# Establecer la fecha como índice
df.set_index('Fecha', inplace=True)

df.sample(10)
```

Ilustración 12. Preparación de los datos 1

Para la implementación de clústeres, también creamos variables dummies para el sector económico de los clientes. El dataframe df_combined, incluye los sectores económicos y todas las demás variables.

```python
# Data seleccionada
X=combined_df[["Active_energy","Reactive_energy","Voltaje_FA","Voltaje_FC"]]
# Dummies sector económico
dummies = pd.get_dummies(combined_df["Sector Económico:"], prefix="Sector",drop_first=True)
# Concatenar los dummies al DataFrame original
X = pd.concat([X, dummies], axis=1)
X.head()
```

Ilustración 13. Preparación de los datos 2

IMPLEMENTACIÓN Y PRUEBAS DE MODELOS

Análisis de Componentes Principales (PCA)

Realizar un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos es una excelente opción, especialmente cuando se trabaja con aprendizaje no supervisado y no hay etiquetas de salida. PCA puede ayudar a visualizar los datos de alta dimensionalidad en un espacio de dimensiones reducidas y a identificar patrones o estructuras subyacentes que podrían no ser evidentes en el espacio de alta dimensión.

- **Selección de Características:** Antes de aplicar PCA, es importante seleccionar las características numéricas adecuadas. Las características temporales como año, mes, día, etc., suelen ser categóricas

y podrían necesitar ser transformadas adecuadamente o excluidas del PCA si no se codifican de una manera que tenga sentido para el análisis.

Aplicación de PCA: biblioteca scikit-learn para aplicar PCA

Las gráficas de PCA, Ilustración 14, Ilustración 15, Ilustración 16, para los años completos 2021 y 2022 son muy similares, esto puede indicar varios aspectos interesantes sobre los datos y el consumo de energía entre esos períodos:

Consistencia en los Patrones de Consumo: La similitud en las gráficas de PCA sugiere que los patrones de consumo de energía, al menos en términos de las variables analizadas (energía activa, energía reactiva, y voltajes), han sido bastante consistentes a lo largo de esos años. Esto puede deberse a:

- **Estabilidad Operativa:** No hay cambios significativos en las operaciones o en los perfiles de consumo de los clientes.
- **Ausencia de Eventos Disruptivos:** No ha ocurrido ningún evento significativo (como cambios en la regulación, grandes avances tecnológicos en la eficiencia energética, o cambios económicos drásticos) que pudiera alterar los patrones de consumo de energía.

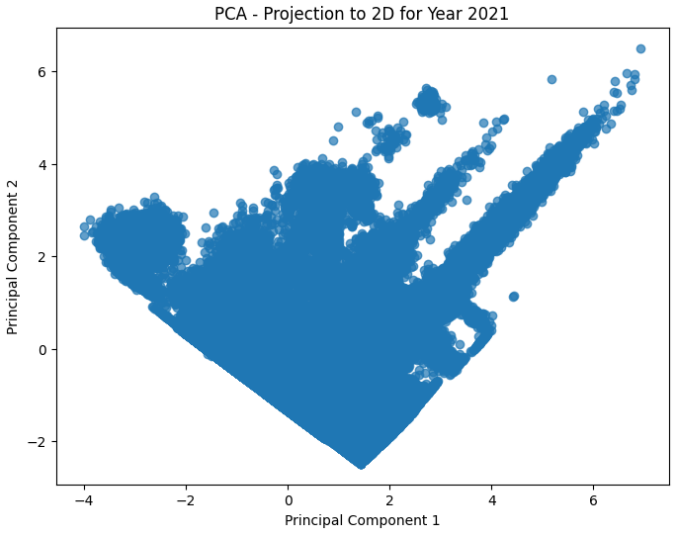


Ilustración 14. PCA 2021

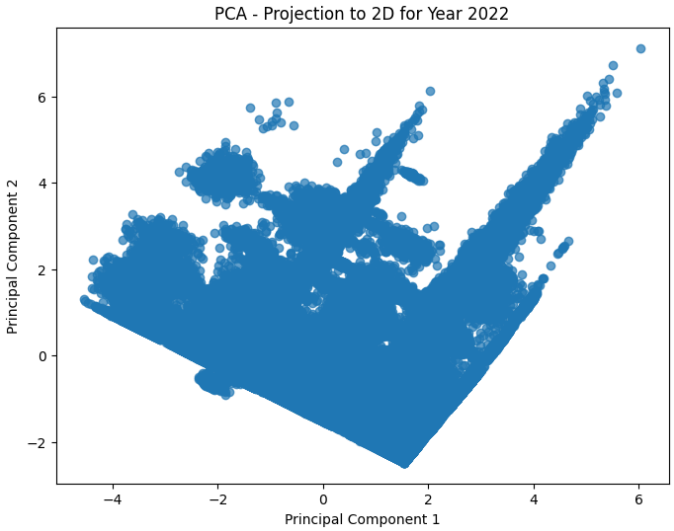


Ilustración 15. PCA 2022

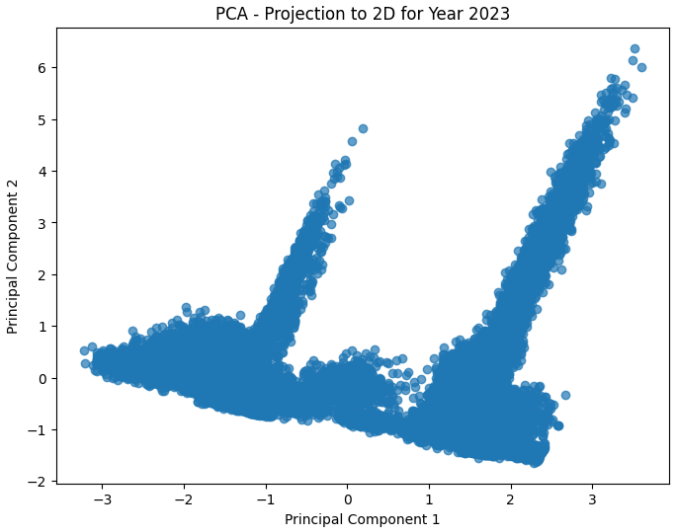


Ilustración 16. PCA 2023

A continuación, realizamos un PCA global para aplicar el gráfico del codo y evaluar la varianza explicada como estrategia, para obtener una visión general de la estructura de los datos a lo largo de todo el período disponible. Este enfoque ayuda a identificar el número óptimo de componentes principales necesarios para capturar la mayor parte de la información relevante en el conjunto completo de datos, sin estar limitado a un análisis anual.

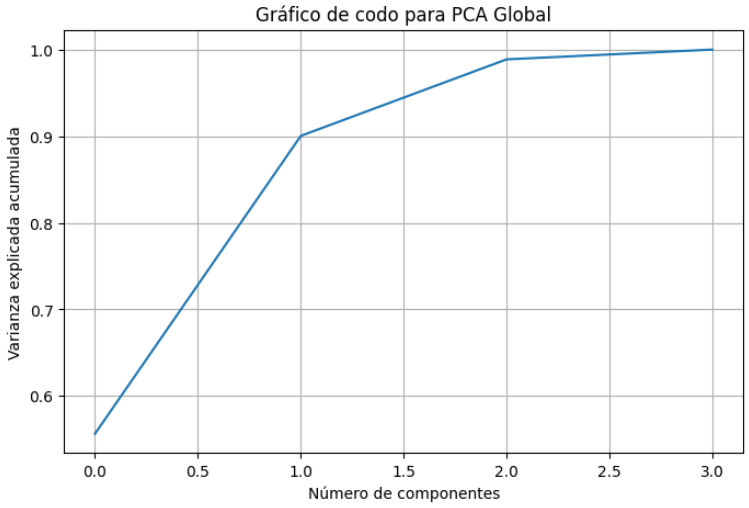


Ilustración 17. Gráfico de codo

Interpretación de Resultados:

- Gráfico de Codo: Visualizamos cómo la varianza explicada acumulada se incrementa con cada componente adicional para ayudar a identificar un punto donde los beneficios de añadir más componentes disminuyen, lo que indica el número óptimo de componentes a retener.
- Evaluación de Varianza Explicada: Usamos un umbral, del 90%, para determinar cuántos componentes son necesarios para capturar la mayoría de la varianza en los datos, lo cual es útil para decisiones sobre la dimensionalidad en aplicaciones subsiguientes.

Dos componentes principales son suficientes para explicar al menos el 90% de la varianza en los datos, esto indica que estas dos dimensiones capturan la gran mayoría de la información relevante en las características de energía y voltaje.

Dado que dos componentes explican la mayor parte de la varianza:

	Active_energy	Reactive_energy	Voltaje_FA	Voltaje_FC
PC1	0.422670	0.368965	-0.584787	-0.585867
PC2	0.552572	0.619041	0.396152	0.393085

Ilustración 18. Variables PCA

- **Componente Principal 1 (PC1):** Este componente probablemente captura la mayor variabilidad en tus datos. Sería útil examinar los vectores de este componente para entender qué variables tienen más peso, lo que puede indicar qué característica(s) son las más influyentes. Los valores positivos para Active_energy y Reactive_energy indican que estas características contribuyen de manera positiva al primer componente principal, cuanto mayor es el valor, más influye esa característica en la dirección del componente. Los valores negativos para Voltaje_FA y Voltaje_FC sugieren que estas variables contribuyen en una dirección opuesta en comparación con Active_energy y Reactive_energy, lo cual podría indicar que cuando la energía activa y reactiva son altas, los voltajes tienden a ser bajos, o viceversa y esto refleja una característica física o técnica del sistema de energía analizado.
- **Componente Principal 2 (PC2):** Este componente añade información adicional que no está capturada por el primer componente. Su interpretación también puede proporcionar insights sobre las dinámicas de los datos que son ortogonales (independientes) al primer componente. Active_energy y Reactive_energy también tienen cargas positivas significativas en el segundo componente principal, pero aquí, Reactive_energy tiene una influencia un poco más fuerte que en PC1, por otro lado, Voltaje_FA y Voltaje_FC también contribuyen positivamente a PC2, lo cual es opuesto a su contribución en PC1. Esto indica que PC2 podría estar capturando una variación en los datos donde aumentos en voltajes coinciden con aumentos en la energía activa y reactiva.

El hecho de que los voltajes tengan una fuerte carga negativa en PC1 y positiva en PC2 sugiere que existen diferentes modos de variabilidad en los datos. En algunos casos, altos niveles de energía están asociados con bajos voltajes y en otros, altos niveles de energía van con altos voltajes.

LOF y DBSCAN

Ahora aplicamos LOF (Local Outlier Factor) y DBSCAN (Density-Based Spatial Clustering of Applications with Noise) sobre los resultados de un PCA para la detección de anomalías como estrategia. Posteriormente, utilizamos el coeficiente de silueta para evaluar la calidad de las agrupaciones y la eficacia de la detección de anomalías.

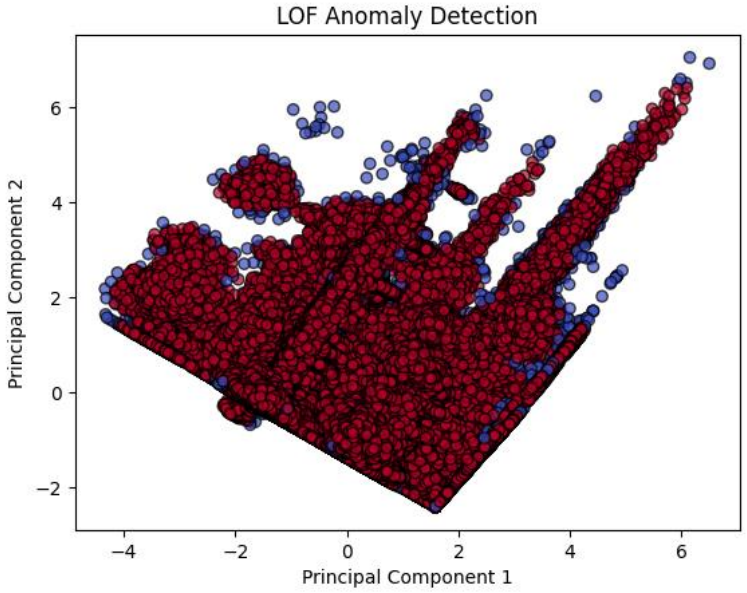


Ilustración 19. Local Outlier Factor

En el caso del modelo DSCAN, descartamos su aplicación por el costo computacional implicado que dificulta el análisis.

Isolation Forest

También, desarrollamos un algoritmo para Isolation Forest que es un algoritmo eficiente para la detección de anomalías, especialmente en conjuntos de datos de alta dimensión. Funciona aislando observaciones al seleccionar aleatoriamente una característica y luego seleccionar aleatoriamente un valor de división entre los valores máximo y mínimo de la característica seleccionada y es utilizado en el desarrollo de modelos de aprendizaje no supervisado, donde no hay variable de respuesta.

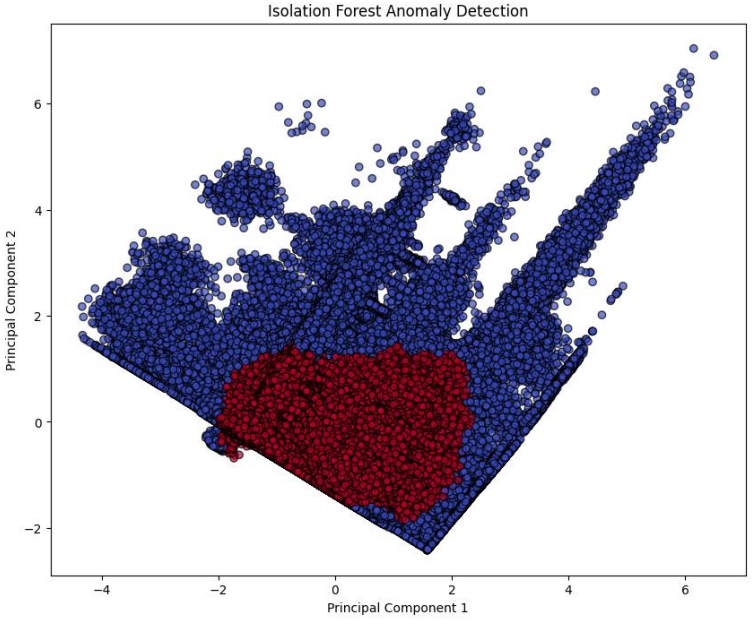


Ilustración 20. Isolation Forest

One-Class SVM

El SVM de una clase es otro método efectivo para la detección de anomalías. Está diseñado para identificar outliers en un conjunto de datos al tratar de encontrar una frontera de decisión que separe las observaciones más "normales" de las menos frecuentes.

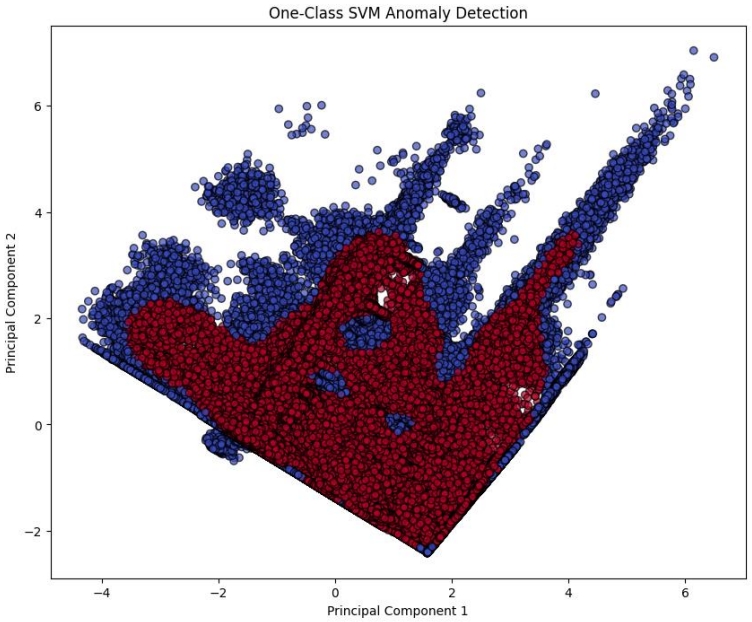


Ilustración 21. One-Class SVM

Clustering

Finalmente, teniendo en cuenta el análisis anterior, también evaluamos un algoritmo de clustering para segmentar los datos y detectar posibles anomalías. Para empezar con este modelo, definimos una regla de aplicación que considera la información relevante que encontramos durante el análisis de la base de datos y que posiblemente representa una anomalía:

Definición de regla de anomalía.

- Propósito: Esta condición asegura que la energía activa no sea cero antes de realizar cualquier cálculo adicional. Esto es importante para evitar la división por cero y asegurar que solo se consideren los registros con consumo energético activo para la evaluación de anomalías.
- Condición Compuesta:
  ((row['Active_energy'] - row['Reactive_energy']) / row['Active_energy']) < -0.9 or row['Active_energy'] < 0.

  Parte 1: ((row['Active_energy'] - row['Reactive_energy']) / row['Active_energy']) < -0.9

  - Cálculo: Esta expresión calcula la diferencia relativa entre la energía activa y la energía reactiva como una fracción de la energía activa.
  - Propósito: Esta condición busca casos donde la energía reactiva es significativamente mayor que la energía activa (específicamente, cuando la energía reactiva es al menos 90% mayor que la energía activa). Esta situación podría indicar un comportamiento anómalo en el consumo de energía, ya que normalmente esperaríamos que la energía activa sea mayor o, al menos, no mucho menor que la energía reactiva.

  Parte 2: row['Active_energy'] < 0

  - Propósito: Esta condición verifica si la energía activa es negativa, lo cual podría ser inusual o indicativo de un error de medición o de una situación anómala, como un posible fraude o mal funcionamiento del sistema de medición.

Resultado Final:

- 'Si': Si las condiciones anteriores se cumplen (es decir, si hay consumo de energía activa no nulo y la energía reactiva supera significativamente a la activa según el criterio dado, o si la energía activa es negativa), la fila se etiqueta como 'Si', indicando una posible anomalía.
- 'No': Si ninguna de las condiciones se cumple, la fila se etiqueta como 'No', indicando que no se detectó anomalía en esa entrada.

Con esta regla, generamos las siguientes etiquetas:

```
No    414128
Si     49297
```

Ilustración 22. Etiquetas anomalías

También, validamos el etiquetado con algoritmos de detección de anomalías como LocalOutlierFactor que funciona con algoritmos que tienen en cuenta las observaciones vecinas para la clasificación, los resultados están en la Ilustración 23:

```python
from sklearn.neighbors import LocalOutlierFactor
# Seleccionar las columnas relevantes para la detección de anomalías
columnas_interesantes = ['Active_energy', 'Reactive_energy', 'Voltaje_FA', 'Voltaje_FC']

# Entrenar el modelo LOF utilizando las columnas seleccionadas
lof = LocalOutlierFactor(n_neighbors=20, contamination=0.1)
etiquetas_anomalias = lof.fit_predict(combined_df[columnas_interesantes])

# Agregar las etiquetas de anomalía al DataFrame
combined_df['Anomalía'] = etiquetas_anomalias

# Cambiar el valor de -1 a 'Anomalía' y 1 a 'Normal'
combined_df['Anomalía'] = combined_df['Anomalía'].apply(lambda x: 'Si' if x == -1 else 'No')
```

```python
combined_df["Anomalía"].value_counts()
```

```
No    417082
Si     46343
Name: Anomalía, dtype: int64
```

```python
# Calcular el porcentaje de igualdad entre las columnas 'Anomalía_vecinos' y 'Anomalía_regla'
porcentaje_igualdad = (combined_df['Anomalía_vecinos'] == combined_df['Anomalía_regla']).mean() * 100

print("Porcentaje de igualdad entre las columnas 'Anomalía_vecinos' y 'Anomalía_regla': {:.2f}%".format(porcentaje_igualdad))
```

```
Porcentaje de igualdad entre las columnas 'Anomalía_vecinos' y 'Anomalía_regla': 82.22%
```

Ilustración 23. Regla de anomalía

Obteniendo un resultado similar al de la regla definida anteriormente, con un porcentaje de coincidencia de 82.22%, decidimos usar la regla para detectar posibles anomalías planteada anteriormente y de esta manera crear los labels.

La función está diseñada para identificar entradas potencialmente anómalas en el dataset, basándose en características específicas del consumo de energía activa y reactiva. Esta identificación preliminar de anomalías puede ser un paso importante en el análisis de datos, especialmente en entornos donde la precisión en el consumo energético es crítica, como en la gestión y facturación de servicios de energía.

Bajo este supuesto, procedimos a desarrollar algoritmos de clustering que intentaran agrupar observaciones similares. El resumen de los modelos desarrollados:

- **Clustering por DBSCAN.** El clustering por DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering basado en la densidad de los puntos en el espacio de características para agrupar los datos. Este no requiere especificar el número de clústeres y puede encontrarlos de formas arbitrarias en los datos.

  Los parámetros variados en este algoritmo fueron EPS (determina la distancia máxima entre dos puntos para que se consideren vecinos directos) y min_samples (número mínimo de puntos que deben encontrarse dentro del radio EPS). Estos fueron calibrados y definidos como 2 y 20 respectivamente, por la cantidad de posibles fraudes agrupados.

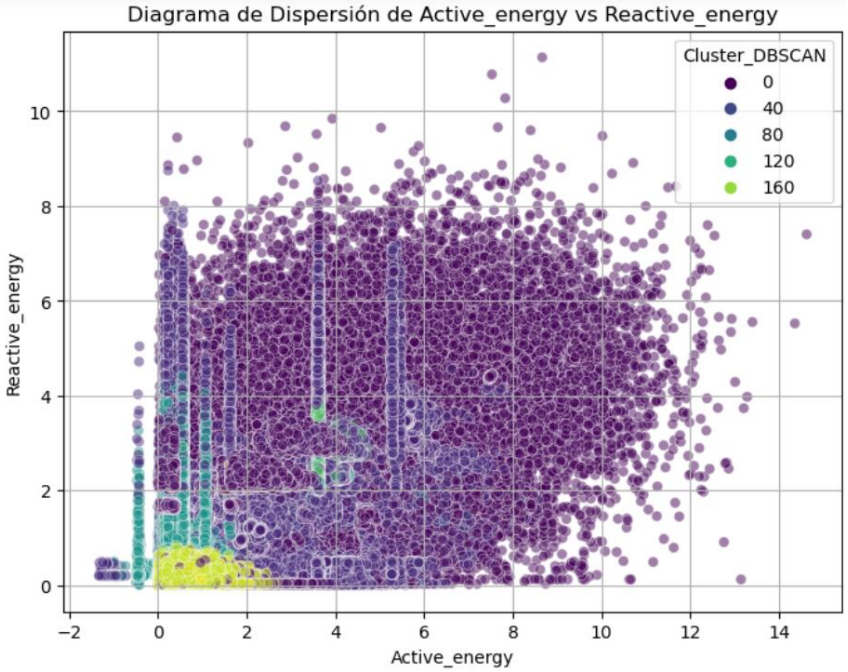


Ilustración 24. Clústeres DBSCAN

	Cluster_DBSCAN	Anomalia_Si	Anomalia_No	Total	Porcentaje_Si	Porcentaje_No
92	164	29.0	55.0	84.0	34.523810	65.476190
69	95	13.0	26.0	39.0	33.333333	66.666667
66	92	198.0	423.0	621.0	31.884058	68.115942
16	24	969.0	2144.0	3113.0	31.127530	68.872470
82	154	48.0	110.0	158.0	30.379747	69.620253
71	97	9138.0	21203.0	30341.0	30.117663	69.882337
97	169	31.0	73.0	104.0	29.807692	70.192308
95	167	8.0	19.0	27.0	29.629630	70.370370
41	61	3144.0	7506.0	10650.0	29.521127	70.478873
99	171	23.0	59.0	82.0	28.048780	71.951220

Ilustración 25. Resultados Clústeres DBSCAN

Debido a la naturaleza del algoritmo, no podemos pre-definir clústeres. En este caso el algoritmo identificó 191 clústeres, la mayoría de las anomalías las agrupó en el clúster -1 (el cual corresponde a no clasificado, según el algoritmo) por lo que esa solución resultó descartada por el equipo.

- **Clustering por K-Means.** El clustering por K-means es un algoritmo popular utilizado en el análisis de datos y aprendizaje automático para agrupar un conjunto de datos en un número predeterminado de clústeres mediante distancias. Funciona dividiendo el conjunto de datos en K grupos basados en características similares.

  En este caso, realizamos una leve calibración del modelo iterando con el número de clústeres, desde 2 hasta 10 siendo 4 el número de clústeres con mejor resultado.

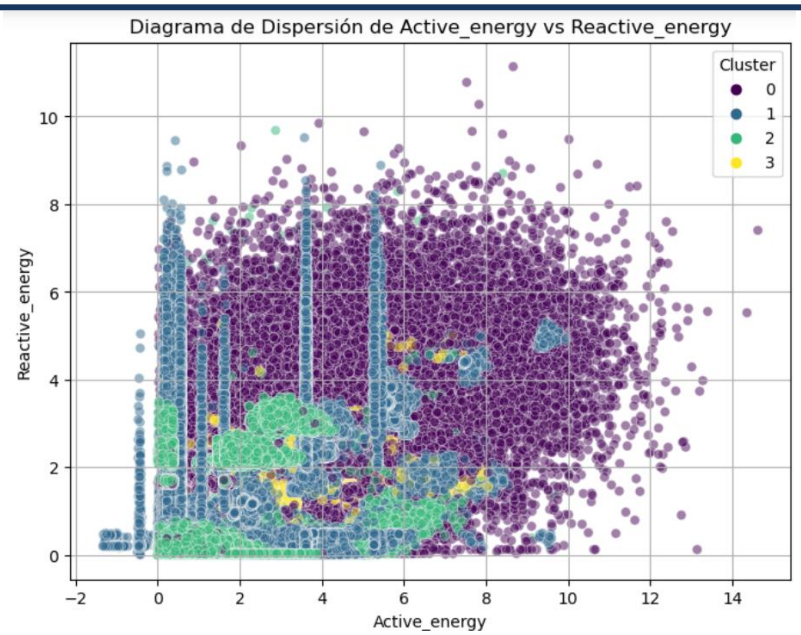


Ilustración 26. Clústeres KMeans

	Cluster	Anomalía_Si	Anomalía_No	Total	Porcentaje_Si	Porcentaje_No
1	1	39441	182122	221563	17.801257	82.198743
2	2	3915	89560	93475	4.188286	95.811714
0	0	5485	129168	134653	4.073433	95.926567
3	3	456	13278	13734	3.320227	96.679773

Ilustración 27. Resultados Clústeres KMeans

Esta solución logra agrupar en el clúster 1 el 17.8% de las anomalías supuestas. Problemas similares han sido abordados en la literatura, Yajure Ramírez, C. A. (2022) ha encontrado una tasa de detección entre el 20% y el 30%, por lo que el valor lo consideramos aceptable e informativo para la compañía. En azul evidenciamos el agrupamiento del clúster de interés, algunas observaciones presentan valores de energía activa negativos, o valores de energía activa muy pequeños comparados con los valores de energía reactiva, consistente con lo que definimos en la regla de detección de anomalías.

En base a este resultado, logramos resumir los fraudes por sector, como podemos apreciar en la Ilustración 28. Además, por cliente y hora:

```
combined_df[(combined_df['Anomalía'] == 'Si') & (combined_df['Cluster'] == 1)].groupby('Sector Económico:').size()
```

```
Sector Económico:
Captación, tratamiento y distribución de agua            17682
Cultivo de Hortalizas                                    11005
Cultivo de hortalizas y melones, raíces y tubérculos      1912
Cultivo de otros frutos y nueces de árboles y arbustos    7900
Cultivo de Árboles Frutales y Nueces                       942
dtype: int64
```

Ilustración 28. Posibles anomalías por sector

Evidenciamos en la data:

- El cliente 20 presenta más anomalías.
- El sector de captación, tratamiento y distribución de agua es el sector con más anomalías.
- Las anomalías a lo largo de las horas de la noche parecen aumentar.

- **Otros algoritmos de clustering testeados:**

  Dentro de las opciones que consideramos, estuvo clustering por K-Medoides y clustering jerárquico, pero estos no presentaron resultados debido al alto poder computacional demandado por la cantidad de observaciones.

ANÁLISIS DE RESULTADOS

Procesamiento de datos:

1. Normalizar los datos numéricos.
2. Convertir la columna Fecha a tipo datetime.
3. Agregar columnas adicionales: Año, Mes, Día, Hora, Día_semana (0 a 6, domingo).
4. Calcular semana del mes.
5. Establecer fecha como índice.

Para el caso del clustering, con el dataframe original, determinamos variables dummies para el sector.

Selección de modelos:

Primera métrica, Coeficiente de Silueta:

```
from sklearn.metrics import silhouette_score

# Asumiendo que las etiquetas son -1 para anomalías y 1 para normales,
# y que necesitamos convertirlas a 0 y 1 respectivamente para que silhouette_score funcione correctamente
lof_labels_adj = (lof_labels == 1).astype(int)
iso_labels_adj = (iso_labels == 1).astype(int)
oc_svm_labels_adj = (oc_svm_labels == 1).astype(int)

# Calcular el coeficiente de silueta
silhouette_lof = silhouette_score(X_pca, lof_labels_adj)
silhouette_iso = silhouette_score(X_pca, iso_labels_adj)
silhouette_svm = silhouette_score(X_pca, oc_svm_labels_adj)

print(f'Coeficiente de Silueta para LOF: {silhouette_lof}')
print(f'Coeficiente de Silueta para Isolation Forest: {silhouette_iso}')
print(f'Coeficiente de Silueta para One-Class SVM: {silhouette_svm}')

```

```
Coeficiente de Silueta para LOF: 0.13336181281446494
Coeficiente de Silueta para Isolation Forest: 0.37256751881882155
Coeficiente de Silueta para One-Class SVM: 0.3777069241171725
```

Ilustración 29. Coeficiente de silueta modelos de detección de anomalías

Los resultados del coeficiente de silueta obtenido para cada uno de los métodos de detección de anomalías indican la calidad de las agrupaciones formadas, en el sentido de la correcta separación de los grupos de datos normales y anómalos:

- **LOF (Local Outlier Factor):** Un coeficiente de silueta de 0.13 sugiere una separación relativamente pobre entre las anomalías y los puntos normales. Esto puede deberse a la naturaleza del algoritmo, que se basa en la densidad local y puede no ser tan efectivo si los patrones de densidad no son uniformes o si el parámetro n_neighbors no está bien ajustado.
- **Isolation Forest:** Con un coeficiente de silueta de 0.37, Isolation Forest muestra una mejor capacidad para distinguir entre datos normales y anómalos. Este método es eficaz para manejar grandes volúmenes de datos y tiende a funcionar bien con diversas distribuciones de datos.
- **One-Class SVM:** Tiene un coeficiente de silueta similar al de Isolation Forest, de 0.38, lo que indica una buena separación de las anomalías respecto a los datos normales. El One-Class SVM es particularmente útil cuando los datos tienen una frontera de decisión clara, aunque la selección de parámetros como gamma y nu es crucial para su rendimiento.
- **Clustering KMeans:** El coeficiente de silueta de 0.60 sugiere una buena clasificación teniendo en cuenta la regla creada.

Segunda métrica Índice Davies-Bouldin:

Dado que trabajamos con algoritmos de detección de anomalías en un entorno de aprendizaje no supervisado, una métrica interesante y relevante para complementar el coeficiente de silueta es el Índice Davies-Bouldin. Esta métrica es útil para evaluar la calidad de la separación entre los clústeres formados por tu modelo. Un valor bajo del índice Davies-Bouldin sugiere que los clústeres están bien separados y que los miembros de cada cluster están cerca uno del otro, lo cual es ideal en la detección de anomalías.

¿Por qué el Índice Davies-Bouldin? El Índice Davies-Bouldin puede proporcionar una visión de cómo los modelos están separando y agrupando los puntos de datos normales y anómalos sin necesidad de etiquetas previas. Además, dado que cada modelo puede tener una tendencia diferente a agrupar los datos, este índice facilita visualizar el modelo que produce clústeres más definidos y compactos.

```
from sklearn.metrics import davies_bouldin_score

db_index_lof = davies_bouldin_score(X_pca, lof_labels)
db_index_iso = davies_bouldin_score(X_pca, iso_labels)
db_index_oc_svm = davies_bouldin_score(X_pca, oc_svm_labels)

print(f'Índice Davies-Bouldin para LOF: {db_index_lof}')
print(f'Índice Davies-Bouldin para Isolation Forest: {db_index_iso}')
print(f'Índice Davies-Bouldin para One-Class SVM: {db_index_oc_svm}')
```

```
Índice Davies-Bouldin para LOF: 21.712807342162947
Índice Davies-Bouldin para Isolation Forest: 2.086616841552757
Índice Davies-Bouldin para One-Class SVM: 2.7994666350550763
```

Ilustración 30. Índice Davies-Bouldin modelos de detección de anomalías

- **LOF (Local Outlier Factor):** Índice Davies-Bouldin: 21.7128. Este valor es considerablemente más alto en comparación con los otros dos modelos. Un valor alto indica que los clústeres formados por LOF no están bien definidos ni separados. Las anomalías no están claramente diferenciadas de los puntos normales, posiblemente debido a una densidad de puntos más uniforme o a la configuración del número de vecinos (n_neighbors). Esto podría sugerir que el ajuste de los parámetros, o la propia naturaleza de los datos, no es ideal para LOF en este caso.
- **Isolation Forest:** Índice Davies-Bouldin: 2.0866. Este resultado es mucho más bajo en comparación con LOF, lo que sugiere que Isolation Forest está haciendo un mejor trabajo al separar grupos de datos normales de las anomalías. Esto indica que los clústeres son más compactos y están mejor separados, lo que es preferible en la detección de anomalías. Isolation Forest es conocido por su eficacia en conjuntos de datos de gran tamaño y diversas distribuciones, lo que podría estar contribuyendo a su buen desempeño aquí.
- **One-Class SVM:** Índice Davies-Bouldin: 2.7995. El valor para One-Class SVM también indica una buena separación y compactación de los clústeres, aunque es ligeramente menos efectivo que Isolation Forest en este aspecto. Este modelo tiende a funcionar bien cuando los datos tienen una frontera de decisión clara. Tomando en cuenta tanto el tiempo de ejecución como las métricas del coeficiente de silueta y el índice Davies-Bouldin, aquí está una conclusión integral sobre el rendimiento de los tres modelos de detección de anomalías que has evaluado: LOF (Local Outlier Factor), Isolation Forest y One-Class SVM.
- **Clustering KMeans:** Índice Davies-Bouldin: 5.5. Sugiere una clasificación que se puede mejorar y complementar con los algoritmos de detención de anomalías.

<u>Evaluación de Rendimiento:</u>

- **LOF (Local Outlier Factor):**
  Tiempo de Ejecución: Es rápido y eficiente.
  Coeficiente de Silueta: 0.1334, significativamente más bajo que los otros dos modelos, indicando una pobre separación.
  Índice Davies-Bouldin: 21.7128, muy alto, lo que sugiere una mala separación y compactación de los clústeres.

Conclusión: LOF parece ser menos adecuado para este conjunto de datos en particular, dado su bajo rendimiento en las métricas de agrupación.

- **Isolation Forest:**
  Tiempo de Ejecución: Es rápido y eficiente, especialmente adecuado para grandes volúmenes de datos.
  Coeficiente de Silueta: 0.3726, indicando una razonable separación entre grupos.
  Índice Davies-Bouldin: 2.0866, sugiriendo una buena separación y compactación de los clústeres.
  Conclusión: Isolation Forest ofrece un excelente equilibrio entre eficiencia de tiempo y calidad de agrupación de los datos. Es robusto, escalable y proporciona un rendimiento consistente, lo que lo hace adecuado para aplicaciones prácticas donde el tiempo y la precisión son críticos.
- **One-Class SVM:**
  Tiempo de Ejecución: Significativamente más alto, lo que puede ser un problema.
  Coeficiente de Silueta: 0.3777, ligeramente mejor que Isolation Forest, indicando una buena separación.
  Índice Davies-Bouldin: 2.7995, bueno, pero no tan eficiente como Isolation Forest en términos de compactación y separación de clústeres.
  Conclusión: Aunque One-Class SVM muestra una buena capacidad para identificar y separar anomalías, su alto costo computacional y sensibilidad a la selección de parámetros pueden limitar su uso práctico en entornos como el caso de aplicación.

Conclusión de las métricas:

Basados en el análisis comprensivo de tiempo de ejecución y las métricas evaluadas, Isolation Forest emerge como el modelo más equilibrado y recomendado para la detección de anomalías en este escenario. Ofrece un buen rendimiento en cuanto a la identificación de grupos claramente definidos y separados de datos normales y anómalos, y lo hace de manera eficiente, lo que lo hace ideal para aplicaciones en tiempo real o en grandes conjuntos de datos.

Por otra parte, aplicaremos la regla creada para establecer el label de posibles anomalías y esto permitirá visualizar los clústeres en el dashboard para dar otra herramienta al analista de datos. De hecho, para esta segmentación de clientes obtuvimos un coeficientee de silueta de 0.60. y un Índice Davies-Bouldin de 5.5, lo que sugiere que el clustering puede mejorarse para que sea más compacto.

```
from sklearn.metrics import silhouette_score

# Calcular el coeficiente de silueta
coeficiente_silueta = silhouette_score(X, etiquetas_clusters)

# Mostrar el coeficiente de silueta
print("Coeficiente de Silueta:", coeficiente_silueta)
```

```
Coeficiente de Silueta: 0.6027635894435924
```

Ilustración 31. Coeficiente de silueta clústeres KMeans

IMPLEMENTACIONES PENDIENTES

- **Visualización interactiva para segmentación de comportamientos anómalos**

  Basándose en la solución aceptada implementando K-Means, optamos por desarrollar una visualización interactiva que permita segmentar las observaciones por cliente, sector económico y anomalías utilizando la librería ipywidgets. La integración de esta característica permitirá realizar análisis exploratorios de las observaciones considerando los valores de energía activa, energía reactiva y la categorización de fraude de la observación. Esta visualización interactiva será implementada en el dashboard de PowerBI mediante la ejecución de scripts de python en la herramienta.

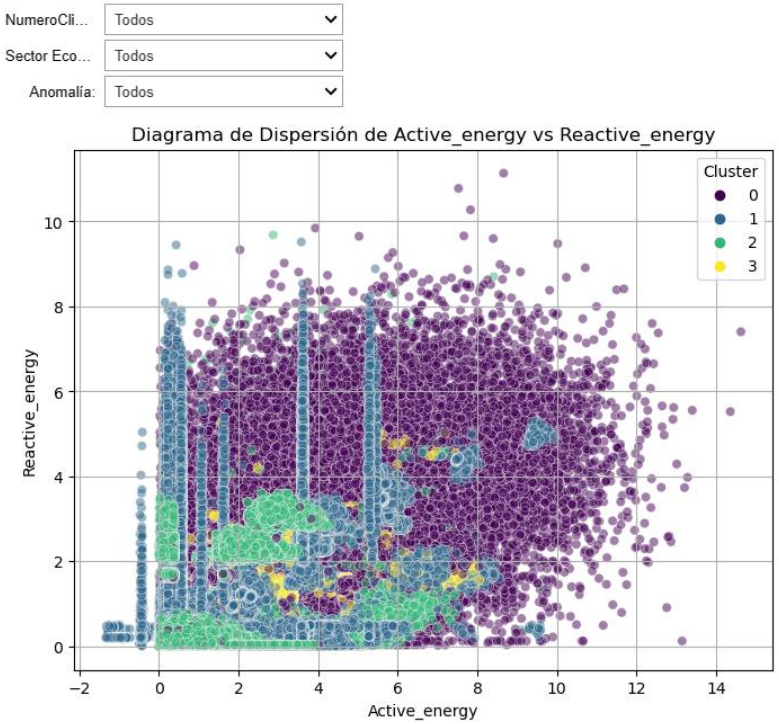


Ilustración 32. Interactividad

Continuando con el desarrollo del clustering, buscaremos mejorar el agrupamiento realizado por el algoritmo, ya sea realizando clustering por cada cliente o por sectores, buscando así la mayor agrupación de anomalías.

- **Visualización descriptiva de valores atípicos**

  Esta gráfica será interactiva y permitirá identificar tendencias de las variables de energía y voltaje a lo largo de la serie de tiempo. Este algoritmo estará basado en Isolation Forest, básicamente en la idea de que las anomalías son puntos que son pocos y distintos en comparación con la mayoría de los otros puntos en el conjunto de datos. Utiliza una estrategia de particionar y conquistar para aislar las anomalías de manera eficiente utilizando árboles de decisión.

  Aunque el algoritmo debe ser refinado para disminuir la sensibilidad de los outliers, es una herramienta útil para la detección de los mismos y será incluida en el dashboard mediante la ejecución de scripts de python en la herramienta.

  Además, será integrado con los clústeres previamente determinados para permitir un análisis más profundo al analista de la empresa. Esto será implementado, al poner un color a la anomalía detectada por Isolation Forest teniendo en cuenta el clúster de la observación.

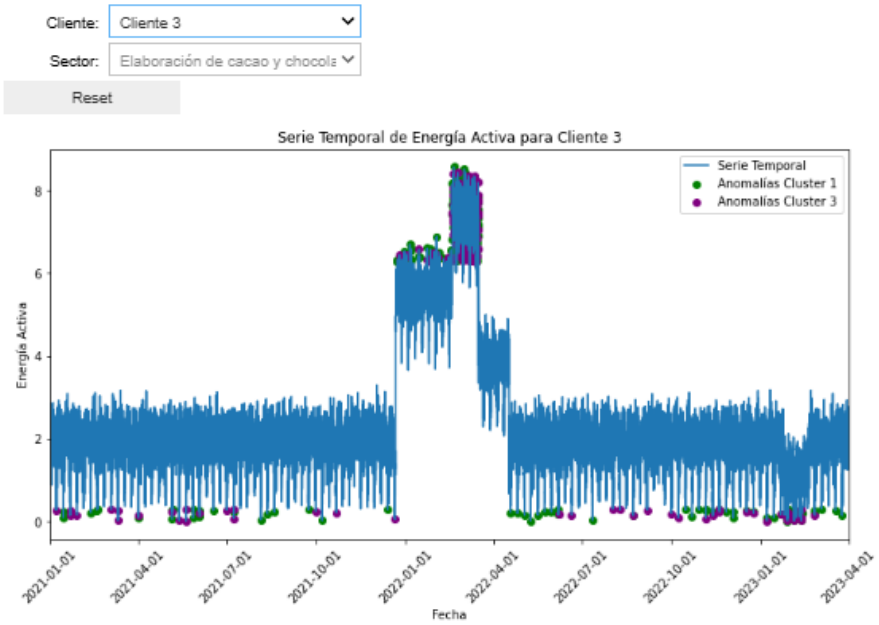


Ilustración 33. Anomalías en una serie de tiempo

CONCLUSIONES

Mediante la realización de análisis estadísticos e interacciones con los datos, hemos podido identificar una posible tendencia de fraude que es de utilidad para la compañía interesada. Dentro de los hallazgos encontrados resaltamos lo siguiente:

- Comportamientos anómalos en energía activa y reactiva: La energía activa es la parte de la energía eléctrica que se convierte en trabajo útil y la cual es facturada, mientras que la energía reactiva es la parte de la energía eléctrica que oscila entre la fuente de alimentación y las cargas en el sistema sin realizar ningún trabajo útil. Por lo que un consumo negativo de energía activa o una diferencia muy grande entre las dos energías puede considerarse como un comportamiento anómalo.
- Algoritmos con mejor desempeño: El clustering por K-Means resultó ser el más útil debido a su manera de agrupación. Según los resultados obtenidos, identificamos cierta tendencia en el clúster 1 relacionada con las anomalías descritas en el tópico anterior.
- Los valores promedio de energía activa y reactiva por hora para todos los clientes parecen seguir cierta estacionalidad, donde hay subidas y bajadas abruptas dependiendo de la hora (pudiéndose atribuir a paradas de planta, cambios de turno, horario de almuerzo etc.).
- El algoritmo Isolation Forest permitirá generar una visualización de anomalías teniendo en cuenta outliers. Esta herramienta será de utilidad para el analista de datos, pues podrá evaluar diferentes filtros.
- Mediante la consecución del proyecto, ahondamos en mayor medida la problemática mediante herramientas interactivas para el análisis de tendencias y comportamientos atípicos en los datos.
- Elegimos entonces el algoritmo de clustering por KMeans después de analizar el Coeficiente de Silueta y el Índice Davies-Bouldin para clasificar las observaciones por clústeres según la regla definida por el equipo de trabajo que relación Energía Activa y Reactiva y completamos con el modelo de Isolation Forest para detectar las posibles anomalías considerando los cambios abruptos en las lecturas de los datos. Sin embargo, es importante mencionar que vamos a trabajar en mejorar los clústeres.
- Sobre el prototipo fachada, vamos a permitir que el usuario pueda comparar la data y los resultados, por clientes y por sectores. El trabajo incluye también integrar la visualización de la serie de tiempo en el front end para responder la pregunta de negocio y facilitar el análisis de los datos.
- Asimismo, el grupo está trabajando en manuales técnicos que proporcionen información sobre el manejo de la herramienta elaborada, mejoras en los modelos y visualizaciones apropiadas con colorimetría de la empresa.

BIBLIOGRAFÍA

Yajure Ramírez, C. A. (2022). Uso de algoritmos de Machine Learning para analizar los datos de energía

eléctrica facturada en la Ciudad de Buenos Aires durante el período 2010 – 2021. Ciencia, Ingenierías y Aplicaciones, 5(2), 7–37. https://doi.org/10.22206/cyap.2022.v5i2.pp7-37.

Calentura R. Yeison F. Algoritmo de agrupación y clasificación para la detección de clientes sospechosos en contribuir a pérdidas no técnicas de energía en una empresa comercializadora eléctrica en Colombia. Trabajo final de maestría como requisito parcial para optar al título de: Magíster en Ingeniería de Sistemas y Computación. Universidad Nacional de Colombia. 2022 https://repositorio.unal.edu.co/bitstream/handle/unal/81555/1121860755.2022.pdf?sequence=1&isAllowed=y

Antmann, P. (2009). Reducing Technical and Non-Technical Losses in the Power Sector. URL: https://openknowledge.worldbank.org/handle/10986/20786.

Targosz, R. (13 de julio de 2009). Electricity theft - a complex problem. URL: http://www.leonardo-energy.org: http://www.leonardo-energy.org/resources/460/electricity-theft-a-complex-problem-581307167ced1

K. Sridharan and N. N. Schulz, "Outage management through amr systems using an intelligent datafilter, "Power Delivery, IEEE Transactions on, vol. 16, pp. 669–675, 2001.

E. Gontijo, A. Delaiba, E. Mazina, J. E. Cabral, J. O. P. Pinto et al. ,"Fraud identification in electricity company customers using decision tree," in Systems, Manand Cybernetics, 2004 IEEE International Conference on, 2004.

https://repositorio.utp.edu.co/server/api/core/bitstreams/77fbeabd-e0dc-4fd4-80dd-21b01162ee17/content

https://biblus.us.es/bibing/proyectos/abreproy/70923/fichero/TFM_MSEE_Modelo+basado+en+mineria+de+datos+para..._+Franna+Quezada+Mateo.pdf

Trejos, R (mayo de 2014). "Metodología para la detección de perdidas no técnicas en sistemas de distribución utilizando métodos de minería de datos" URL: https://repositorio.utp.edu.co/server/api/core/bitstreams/c93646cb-19e3-44a6-82bc-3d3167e32504/content

Giraldo, A (enero de 2018). "Desarrollo y aplicación de la metodología bagging y ADABOOST para la detección de pérdidas no técnicas en el sistema de distribución de la empresa de energía de Pereira S.A. ESP" URL: https://repositorio.utp.edu.co/server/api/core/bitstreams/77fbeabd-e0dc-4fd4-80dd-21b01162ee17/content

Kim, J.; Jung, Y.-A. Methods of Pre-Clustering and Generating Time Series Images for Detecting Anomalies in Electric Power Usage Data. Electronics 2022, 11, 3315. https://doi.org/10.3390/ electronics11203315