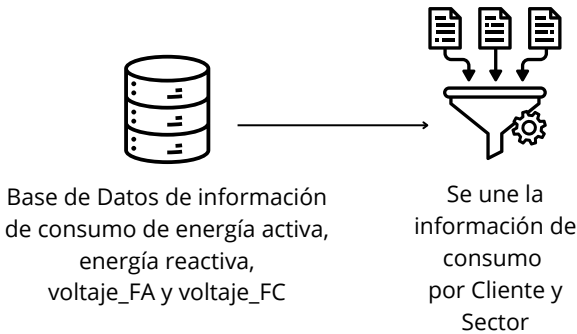


Diagrama Esquemático

Curso Proyecto aplicado analítica de datos

Grupo 11 Lina Marcela Ladino Solis,
Mayra Alejandra Neisa Valero, Juan Felipe Manjarres Mur



Extracción

¿Qué fuentes de datos se deben incluir y como se acceden a ellas?

Se deben incluir:

- Tabla con la medición de 4 variables: energía activa, energía reactiva, voltaje_FA y voltaje_FC. Las mediciones se registran con un intervalo de una hora, las 24 horas del día.
- Tabla con la información del sector al que pertenece cada cliente. Se tienen 7 diferentes sectores, que cubren tratamiento y distribución de agua, cultivos y tratamiento de alimentos, tratamiento y venta de metales.

A esta información se accede por ahora a través del envío del cliente de la información en un archivo plano.

En una siguiente etapa del proyecto, se debe estudiar la integración directa a la base de datos (por ejemplo, por medio de una API) para tener acceso a la misma y poder reportar las anomalías en tiempo real.

¿Cómo especificar la consulta de los datos?

Los datos serán consultados a través de filtros por fecha, cliente, sector económico permitiendo analizar las características de los diferentes comportamientos de consumo.

Transformación Carga

¿Qué características deben tener los datos que requieren los análisis considerados (formato, tipo de dato, balanceo, errores) para producir conclusiones válidas?

Los datos deben tener las siguientes características:

- Formato: Deben estar grabados bajo el formato CSV (Valores Separados por Coma) con los campos de fecha, cliente, sector económico, energía activa, energía reactiva, voltaje_FA y voltaje_FC.
- Tipo de Datos: Los datos deben estar bajo los siguientes tipos:
 - Fecha: Los registros por intervalos de hora deben estar grabados como datos de fecha.
 - Cliente y Sector: deben estar grabados como cadena de texto.
 - Energía activa, energía reactiva, voltaje_FA y voltaje_FC: deben estar grabados como datos numéricos.
- Errores: Los datos deben ser coherentes bajo la lógica de la industria de energía, por lo tanto se eliminan de los datos con errores que en este caso son aquellos que presentan medidas de medición de energía con valores negativos.

¿Que transformaciones son necesarias para llevar los datos a la forma requerida?

Se eliminaran los valores nulos o negativos del consumo energético y se creara una columna para el año, mes, el día de la semana y si se trata de un festivo o no para tener más variables que permitan comprender los diferentes comportamientos de consumo basados con la estacionalidad en el tiempo.

Dependiendo del modelo estimado se puede recurrir a preprocesamientos como estandarización de variables.

Análisis

¿Qué pregunta(s) debe abordar el análisis (y cómo se relaciona con modelos analíticos descriptivos, predictivos o prescriptivos)?

Para realizar el análisis se deben abordar las siguientes preguntas:

1. ¿Cuál es el patrón de consumo de energía de los diferentes clientes?
2. ¿Cómo se deben segmentar los clientes?
3. ¿Cómo se pueden determinar cuales son los consumos anómalos de energía de los clientes no regulados?
4. ¿A que nivel de anomalía se debe notificar al usuario para que conozca el cliente que esta presentando este consumo por fuera de lo normal?

Estas preguntas de análisis se deben desarrollar desde diferentes modelos de análisis de datos:

- Modelos descriptivos: Se debe realizar un análisis descriptivo de los consumo de energía de clientes no regulados que permita determinar los comportamientos de los mismos.
- Modelos predictivos: Se debe realizar un análisis que permita determinar con base en los datos históricos aquellos consumos anómalos.

¿Qué técnicas, modelos o algoritmos puede involucrar (text-mining, econométrico, clasificación, pronóstico, análisis de redes, anomalías, simulación, optimización)?

Con base en lo que se requiere se determina que se deben aplicar modelos no supervisados (ya que no se cuenta con la información de cuando se han presentado anomalías en el pasado) con base en esto se exploraran los siguientes algoritmos eligiendo aquel que se considere más útil:

1. Isolation Forest.
2. K-Means
3. OCSVM.

Uso

¿Quien es el usuario final del artefacto y como afecta esto el diseño?

El usuario final del producto es el área operativa de Electro Dunas S.A.A y Grupo Energía de Bogota. Se espera que con su conocimiento pueda interpretar la información recibida y tomar las acciones preventivas o correctivas dependiendo del caso.

El diseño debe ser amigable para los usuarios, con un fácil entendimiento y una interfaz amigable

¿Qué tipo de herramienta o uso supone (alerta, dashboard, reporte, app)?

Se creara un dashboard que estará integrado con la información histórica que se brindo y que mostrara aquellos consumos anómalos que se han presentado en el tiempo.

¿Qué funcionalidades o características debería tener?

Esta herramienta tendrá las siguientes funcionalidades:

- Permitirá filtrar el consumo de energía activa, energía reactiva, voltaje_FA y voltaje_FC por cliente, sector económico, año, mes y día de la semana.
- Visualización de las anomalías que se han presentado históricamente y en tiempo real.
- Alerta o marca específica de las anomalías que superen el umbral.

Esto le permitirá a los usuarios conocer los patrones de comportamiento de consumo y detectar de forma rápida posibles anomalías en este para identificar posibles casos de fraude o fallas en el sistema.

Lógica del Sector

¿A que audiencia esta orientado el producto?

El usuario final del producto es el área operativa de Electro Dunas S.A.A y Grupo de Energía de Bogotá.

¿Qué conceptos, procesos, actores factores o métricas claves están involucrados en el análisis?

Para el proceso de detección de anomalías se da uso de modelos de aprendizaje no supervisado. Estos modelos pueden ser de difícil comparación y evaluación, su desempeño dependerá de las características de los datos. Para su evaluación, se deberá recurrir a análisis gráficos de comparación

Se debe encontrar una medida de anomalía (como un puntaje de criticidad) y en conjunto con el conocimiento del cliente, definir un threshold para reportar las medidas extrañas.

¿Qué requerimientos se deben cumplir y cómo medir o verificar su cumplimiento?

Se deben cumplir los siguientes requerimientos:

- Visualizar datos históricos de las variables para cada cliente.
- Desplegar un resumen descriptivo por cliente de su comportamiento histórico.
- Identificar consumos anómalos que no coincidan con los consumos esperados dado el comportamiento histórico.
- Mostrarle al usuario cuando un cliente presente un comportamiento anómalo y la criticidad de éste de acuerdo con los modelos utilizados.
- Ser amigable con el usuario, con una interfaz clara y comprensible

Estos requerimientos se consideran cumplidos tras las revisiones con el stakeholder o los encargados del curso.

¿Qué externalidades o consideraciones (éticas, organizacionales) tener en cuenta?

Como consideración especial se debe tener en cuenta:

- Que se esta trabajando con información confidencial por lo que se debe mantener el acceso a la información brindada solo a aquel personal al que sea estrictamente necesario.
- Que se busca la detección de anomalías a tiempo real con una conexión directa a base de datos por lo cual es necesario buscar licenciamiento que permita tener esta información.