

# **Manual de Usuario versión 1.0**

## **Sistema Automatizado de Detección de Anomalías en el Consumo de Gas de Clientes Industriales**



### **Miembros del equipo:**

**PAOLA ALFARO  
ESTEBAN TABARES  
JAVIER ABRIL  
JULIETH MORA**

**En adelante:  
TWO CONSULTING GROUP**

## Índice

<b>A. ¿Qué es y qué hace este Sistema?</b>	3
<b>B. Pasos para el uso del sistema</b>	4
<b>C. Casos de uso y pasos esperados</b>	7
Casos de uso soportados	7
Paso a paso para el uso esperado	7
<b>D. Anexo técnico</b>	8
<b>1. Diagrama prototipo</b>	8
<b>2. Reporte técnico de experimentos</b>	10
Paso 1: Carga, análisis exploratorio y transformación de los datos	10
Paso 2: Segmentación de datos por clustering:	16
Paso 3: Preprocesamiento con métricas estadísticas Promedio Móvil, Rangos Intercuartiles IQR y Z score:	17
Paso 4: Pruebas de Modelos para Detección de Anomalías	17
Paso 5: Configuración de tablero en Power BI	20
<b>3. Repositorio GitHub</b>	23
<b>E. Validación Rúbrica</b>	25
<b>F. Tabla de requerimientos</b>	26

## A. ¿Qué es y qué hace este Sistema?

El sistema para la detección de anomalías en el consumo de gas de clientes industriales Contugas de ahora en adelante llamado “Sistema” es un artefacto o sistema automatizado de detección de anomalías basado en analítica estadística con modelos de detección automáticos **dirigido al equipo técnico y de operaciones de Contugas**, responsables de la supervisión del consumo industrial de gas

Este Sistema tiene las siguientes ventajas:

- Monitorear variables del consumo de gas como presión, temperatura y volumen de gas históricos de los clientes industriales de Contugas.
- Generar alertas tempranas para la toma de decisiones. (Alertas basadas en análisis de los datos de consumo).
- Este sistema además tiene como valor agregado el uso del modelo de agrupación (Clustering) Agglomerative para identificar patrones de consumo por grupos de clientes con comportamientos similares, lo que permite segmentar la atención, anticipar necesidades energéticas específicas y detectar desviaciones dentro de cada grupo con mayor precisión.
- Visualización clara del comportamiento de clientes.
- Acceso desde navegador o dispositivos móviles.
- Reducir los costos asociados a la identificación manual de anomalías y aumentar la eficiencia operativa.

### Limitaciones:

- Basado en un conjunto reducido de datos (20 clientes); requiere ampliación para producción.
- No hay aún integración automática con sistemas SCADA (procesos operativos) o ERP (funciones empresariales).
- Requiere entrenamiento periódico para mantener precisión con nuevos patrones.

### Advertencias

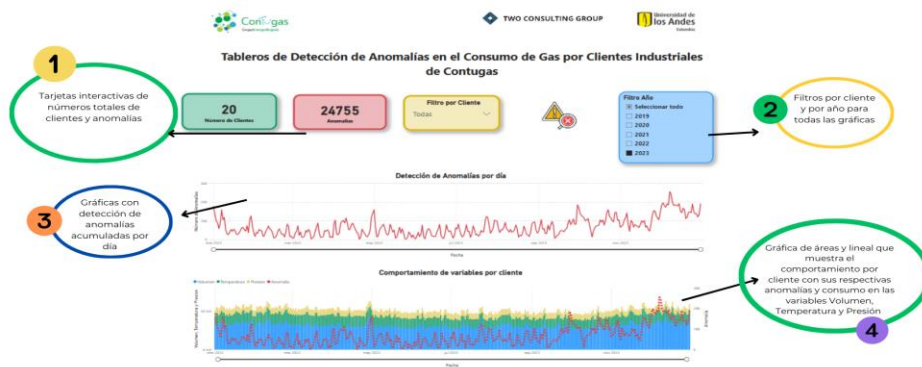
- La precisión del sistema depende de la calidad y periodicidad de los datos cargados.
- No está diseñado para modificar datos de origen desde Power BI.

## B. Pasos para el uso del sistema

Antes de usar el tablero debe Ingresar al link en línea y público de visualizaciones por medio del siguiente enlace de acceso:

<https://app.powerbi.com/view?r=eyJrIjoiaMTUyYjYwNjEtY2FIYi00MGFhLTlmY2ItNGFkYmRkYzBIMDY3IiwidCI6ImU3OTg0Y2FjLT11NDMtNGY4OC04Zjk3LTk1MjQzMzVlNmJjNCIsImMiOiR9>

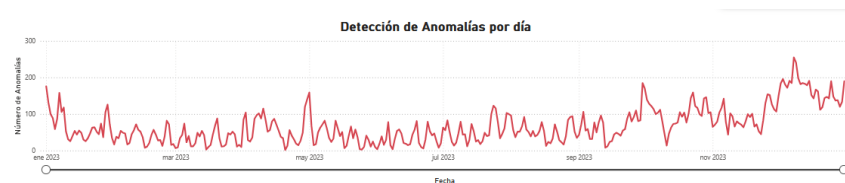
En el tablero de detección de anomalías encontrarás el comportamiento de consumo de los clientes industriales Contugas con las anomalías con las tarjetas agrupadas con el número de clientes y cantidad total de anomalías, así como flitros de clientes y por año de análisis.



**Ilustraciones: Ilustración tablero parte 1**

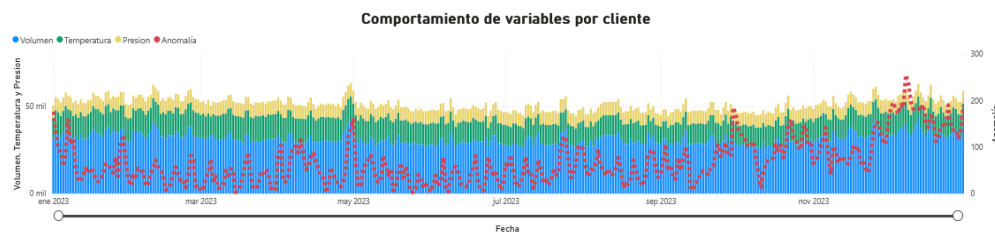
Este tablero contiene gráficas que análisis multivariado de anomalías por cliente donde se puede visualizar:

1. Gráfica Lineal que analiza las anomalías por cliente y por fecha diaria en color rojo con barra de filtro deslizante para filtrar por parámetros de tiempo.



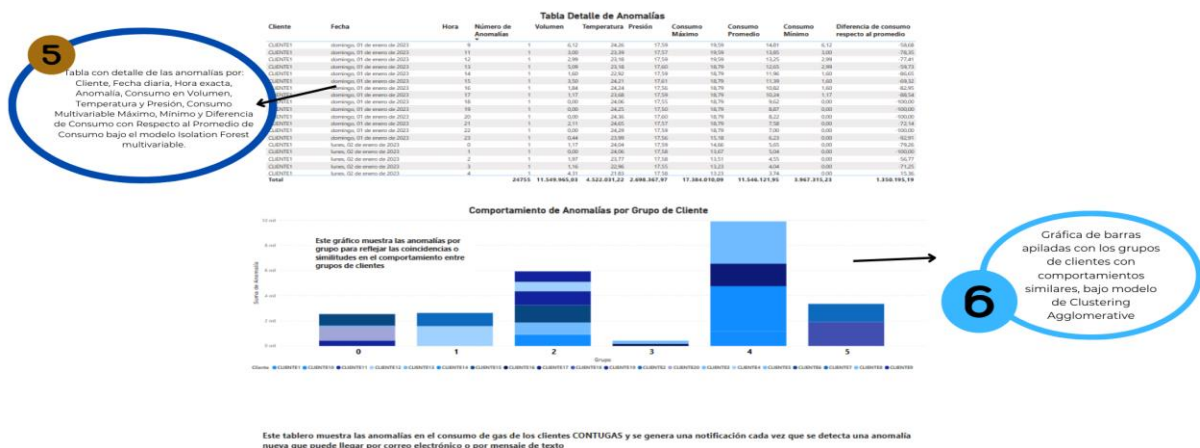
**Ilustraciones: Ilustración gráfica lineal anomalías por cliente**

2. Gráfica de áreas y lineal que muestra el consumo por variable por ejemplo Volumen (Celeste), Presión (Amarillo) y Temperatura (Verde) en la gráfica en forma de área y una línea punteada roja que muestra las anomalías en cada cliente y por fecha diaria con barra de filtro deslizante para filtrar por parámetros de tiempo.



**Ilustraciones: Ilustración gráfica Comportamiento de Variables por Cliente**

En la parte inferior del tablero encontrará una tabla con más detalle de las anomalías y una gráfica de barras con el comportamiento de los clientes que se explica a continuación:



Este gráfico muestra las anomalías por grupo para reflejar las coincidencias o similitudes en el comportamiento entre grupos de clientes

Este tablero muestra las anomalías en el consumo de gas de los clientes CONTUGAS y se genera una notificación cada vez que se detecta una anomalía nueva que puede llegar por correo electrónico o por mensaje de texto

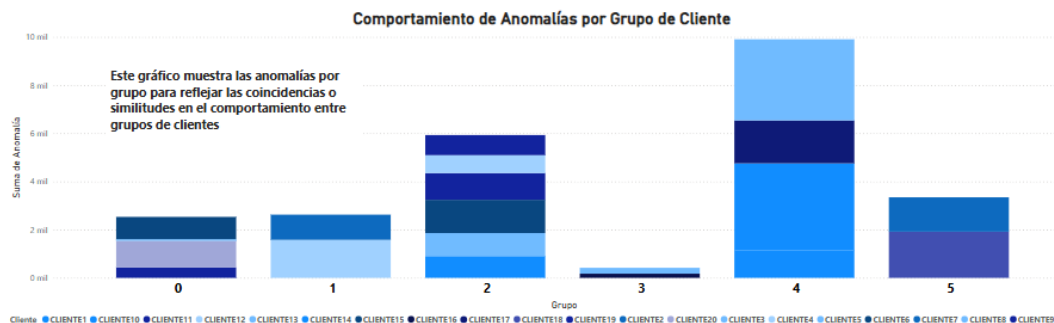
**Ilustraciones: Ilustración tablero parte 2**

3. Tabla con detalle de las anomalías por: Cliente, Fecha diaria, Hora exacta, Anomalía, Consumo en Volumen, Temperatura y Presión, Consumo Multivariable Máximo, Mínimo y Diferencia de Consumo con Respecto al Promedio de Consumo bajo el modelo Isolation Forest multivariable que se detalla más adelante.

Tabla Detalle de Anomalías											
Cliente	Fecha	Hora	Número de Anomalías	Volumen	Temperatura	Presión	Consumo Máximo	Consumo Promedio	Consumo Mínimo	Diferencia de consumo respecto al promedio	
CLIENTE4	miércoles, 18 de enero de 2023		5	1	260.99	25.80	17.76	358.60	270.81	194.16	-3.63
CLIENTE19	domingo, 26 de noviembre de 2023		8	1	235.74	24.93	16.79	337.21	270.54	225.06	-12.86
CLIENTE4	miércoles, 18 de enero de 2023		6	1	204.77	24.07	17.73	358.60	270.54	194.16	-24.31
CLIENTE19	domingo, 26 de noviembre de 2023		7	1	246.02	25.31	16.79	337.21	270.39	225.06	-9.01
CLIENTE19	domingo, 26 de noviembre de 2023		6	1	232.23	24.58	16.79	337.21	269.94	225.06	-13.97
CLIENTE19	miércoles, 18 de octubre de 2023		18	1	268.00	22.54	17.23	348.72	269.75	59.00	-1.39
CLIENTE4	miércoles, 18 de enero de 2023		0	1	237.83	25.76	17.70	358.60	269.23	198.28	-11.66
CLIENTE4	martes, 17 de enero de 2023		21	1	305.82	25.43	17.75	358.60	269.07	198.28	13.66
CLIENTE4	miércoles, 18 de enero de 2023		7	1	233.85	24.17	17.69	358.60	269.02	194.16	-13.07
CLIENTE19	domingo, 26 de noviembre de 2023		9	1	266.00	25.20	16.78	337.21	268.80	225.06	-1.04
CLIENTE4	martes, 17 de enero de 2023		23	1	266.00	26.76	17.67	358.60	268.75	198.28	-1.02
CLIENTE19	sábado, 10 de junio de 2023		0	1	230.66	23.81	17.32	353.30	268.35	190.52	-14.05
CLIENTE4	miércoles, 18 de enero de 2023		4	1	236.36	24.75	17.74	358.60	268.19	194.16	-11.87
CLIENTE4	martes, 17 de enero de 2023		22	1	224.46	25.52	17.72	358.60	268.01	198.28	-16.25
CLIENTE19	domingo, 26 de noviembre de 2023		5	1	266.00	25.14	16.80	337.21	267.98	185.29	-0.74
CLIENTE4	miércoles, 18 de enero de 2023		3	1	233.72	25.03	17.71	358.60	267.98	194.16	-12.78
CLIENTE4	miércoles, 18 de enero de 2023		1	1	194.16	26.01	17.80	358.60	267.81	194.16	-27.50
CLIENTE19	miércoles, 18 de octubre de 2023		17	1	266.71	22.55	17.25	348.72	267.62	59.00	10.87
Total				24755	11.549.965.93	4.522.031.22	2.698.367.97	17.384.010.09	11.546.121.95	3.967.315.23	1.350.195.19

Ilustraciones: Ilustración de Tabla en panel Detalle de Anomalías

4. Gráfica de barras apiladas con los grupos de clientes con comportamientos similares, bajo modelo de Clustering Agglomerative que se detalla más adelante su funcionamiento.



Ilustraciones: Comportamiento Anomalías por Grupo de Cliente

## C. Casos de uso y pasos esperados

### Casos de uso soportados

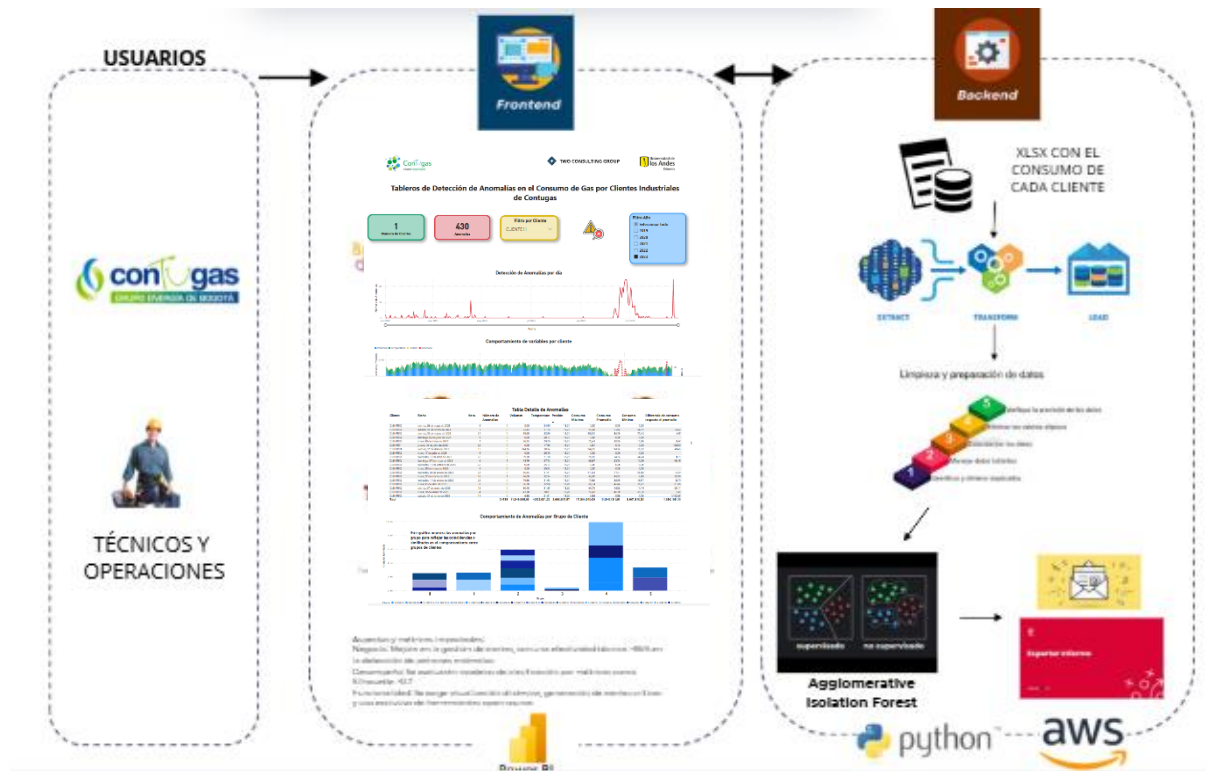
- Procesar la base de datos con el consumo de los clientes industriales de Contugas, que contiene variables como **fecha, volumen, presión y temperatura**, en formato .csv estructurado por cliente y por hora.
- Visualizar clientes agrupados según patrones de consumo utilizando el modelo **Agglomerative Clustering**.
- Detectar anomalías de forma visual e intuitiva, comparando el comportamiento actual frente al histórico.
- Agregar nueva información mediante carga de archivos o actualización programada.
- Descargar datos filtrados por cliente, fecha o tipo de evento.
- Acceder desde navegador o dispositivo móvil para monitoreo remoto.

### Paso a paso para el uso esperado

1. **Cargar nuevos datos:**
  - Subir archivo .XLSX actualizado al repositorio configurado en Power BI o a la carpeta conectada desde Power BI.
  - Asegurarse de que el archivo mantenga la estructura esperada: columnas de cliente, fecha, volumen, presión, temperatura.
2. **Actualizar visualizaciones:**
  - El tablero está programado para actualizarse automáticamente cada **1 hora**.
  - **No se recomienda una frecuencia menor a 30 minutos**, ya que puede generar sobrecarga innecesaria en el servicio de acuerdo a lo mencionado en sesiones con la representante de la compañía.
3. **Filtrar y consultar información:**
  - Usar los filtros disponibles en el panel lateral para seleccionar **cliente(s), fecha(s), grupo de consumo (cluster)**.
  - Las visualizaciones se actualizan en tiempo real con base en los filtros aplicados.
4. **Explorar anomalías:**
  - Revisar gráficos de línea, indicadores de alerta que destacan consumos atípicos.
  - Comparar el comportamiento actual frente al promedio histórico del mismo grupo.
5. **Agregar información adicional (usuarios avanzados):**
  - Si se desea incluir nuevas variables o modificar la estructura de entrada, contactar al equipo consultor Two Consulting Group para ajustes en el modelo.

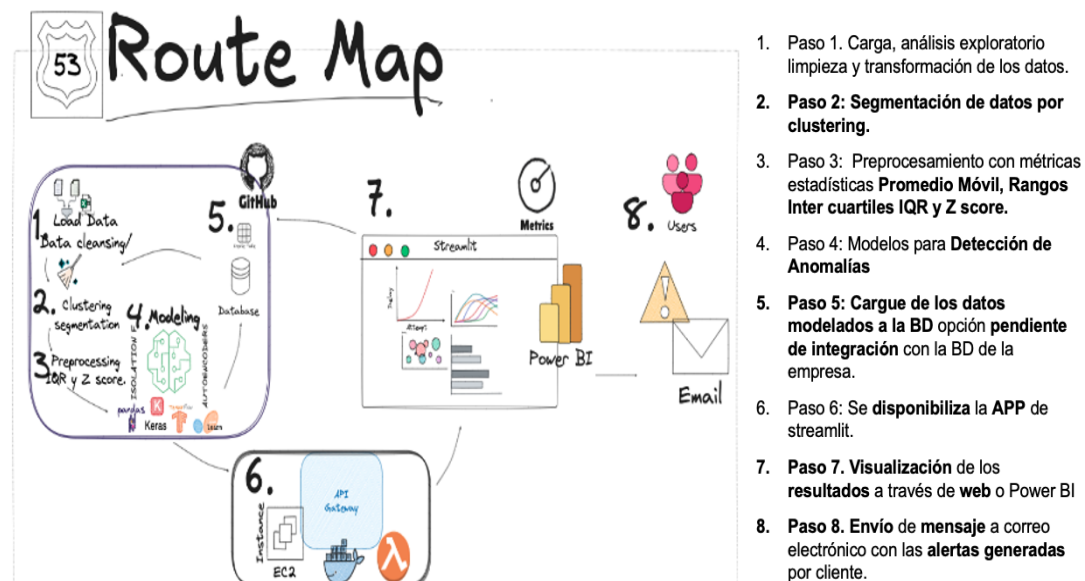
## D. Anexo técnico

### 1. Diagrama prototipo



Ilustraciones: Diagrama prototipo

### Diagrama Esquemático



Ilustraciones: Diagrama Esquemático



En el diagrama esquemático anterior se evidencia el desarrollo del artefacto analítico propuesto el cual se desarrolló como se detalla a continuación:

**Paso 1: Carga, análisis exploratorio limpieza y transformación de los datos.** Se realiza la carga de los datos, se realiza limpieza de los datos validando que no existan valores null/NaN o vacíos. Así mismo verifica y castea en caso de ser necesario las columnas. Se agrupan y generan nuevas características de temporalidad, se realiza compilación de cada cliente en un solo Dataframe con el fin de consolidar la información y lograr análisis de los datos (Notebook 1.0EDA\_Analisys.ipynb).

**Paso 2: Segmentación de datos por clustering.** A Los datos preprocesados anteriormente se les aplica el modelo Aglomerative, para segmentar e identificar patrones similares de consumo de los clientes.

**Paso 3: Preprocesamiento con métricas estadísticas Promedio Móvil, Rangos Intercuartiles IQR y Z score.** Se realiza validación de los datos verificando que no existan valores null/NaN o vacíos. Así mismo se castea en caso de ser necesario las columnas. También se realiza un preprocesamiento de los datos creando nuevas características de los datos, a través de medidas como la media móvil y un Z-score (Notebook 2.0 Preprocesing.ipynb).

**Paso 4: Aplicar Modelos para Detección de Anomalías.** En este paso se selecciona el Modelo Isolation Forest (Notebook, 3.0 Modeling.ipynb ).

**Paso 5: Carga de los datos modelados** a la BD como opción se podría integrar con las bases de datos de la empresa.

**Paso 6: Opcional: Se disponibiliza** la aplicación (Los documentos para disponibilizar la APP de streamlit los cuales están disponibles en GitHub.

**Paso 7: Tableros con Detección de Anomalías:** Los resultados de la detección se visualizan a través de la web en **Power BI**:

<https://app.powerbi.com/view?r=eyJrljoimTUyYjYwNjEtY2FIYi00MGFhLTlmY2ltNGFkYmRkYzBIMDY3liwidCI6ImU3OTq0Y2FjLTl1NDMtNGY4OC04Zjk3LTk1MjQzMzVINmJjNCIsImMiOiR9>

y en el linea a través de Streamlit para cargar archivo de excel o conexión a base de datos(aun en desarrollo y se adapta a la necesidad del cliente:

<https://project-jakw5ayxqbhwqgkeru7fc.streamlit.app/>

**Paso 8: Notificación de Alertas Power BI:** Envío de mensaje a correo electrónico y teams con las alertas generadas por cliente.

## 2. Reporte técnico de experimentos

Paso 1: Carga, análisis exploratorio y transformación de los datos

Se cargaron las 20 hojas del archivo .xlsx proporcionado por Contugas, correspondientes a cada cliente, con registros entre el 14 de enero de 2019 y el 31 de diciembre de 2023. El análisis incluyó las variables *fecha*, *volumen*, *temperatura* y *presión*.

Cliente	Granularidad (hrs)	Anomalías Presión	Anomalías Temperatura	Anomalías Volumen	Duplicados	Faltantes (%)	Edad Promedio (días)	Correlación Promedio
CLIENTE1	1	0	0	8455	0	0	1227.99	0.38
CLIENTE2		0	0	40266	0	0	1213.39	0.34
CLIENTE3		42248	0	41680	0	0	1230.91	0.19
CLIENTE4	1	2	0	41992	0	0	1222.46	0.31
CLIENTE5	1	3	0	28580	0	0	1227.67	0.31
CLIENTE6	1	9	0	38703	0	0	1226.35	0.36
CLIENTE7	1	4	13	39291	0	0	1225.15	0.33
CLIENTE8		1	0	40005	0	0	1225.83	0.14
CLIENTE9	1	0	0	41211	0	0	1230.21	0.2
CLIENTE10	1	0	0	40706	0	0	1242.81	0.25
CLIENTE11		42248	0	42137	0	0	1230.91	0.2
CLIENTE12	1	4	8	40295	0	0	1225.15	0.35
CLIENTE13	1	0	0	40726	0	0	1242.81	0.25
CLIENTE14	1	3	2	30657	0	0	1227.67	0.28
CLIENTE15	1	0	0	40818	0	0	1230.21	0.17
CLIENTE16		1	0	42533	0	0	1225.83	0.17
CLIENTE17	1	0	0	3683	0	0	1227.99	0.36
CLIENTE18		0	0	40086	0	0	1213.39	0.35
CLIENTE19	1	0	0	42090	0	0	1222.46	0.31
CLIENTE20	1	1	0	38818	0	0	1226.35	0.4

**Ilustración: Análisis por cliente (Datos obtenidos del procesamiento en Python archivo)**

**A. Granularidad:** La mayoría de los clientes tienen una granularidad horaria (valor 1), lo que indica que los datos se registran cada hora. Sin embargo, algunos clientes (CLIENTE2, CLIENTE3, CLIENTE8, CLIENTE11, CLIENTE16, CLIENTE18) no

tienen un valor claro en la columna de granularidad. Esto podría indicar problemas en la consistencia de los datos como lo son las fechas o diferencias en los intervalos de muestreo.

Se revisa en técnicas de limpieza de las fuentes de datos para estos clientes y así verificar si los intervalos de muestreo son consistentes y asegurar que los datos estén alineados para obtener un análisis preciso.

**B. Fidelidad y Exactitud:** Las anomalías en presión y temperatura son relativamente bajas, pero hay clientes con cantidades significativas de anomalías en volumen (CLIENTE1, CLIENTE2, CLIENTE3).

CLIENTE3 y CLIENTE11 tienen valores extremadamente altos para "Anomalías Presión", lo que puede indicar un problema con la calibración de sensores o errores en la captura de datos.

**C. Nivel de Integridad:** No hay registros duplicados ni datos faltantes, lo cual indica un buen nivel de integridad dentro de cada cliente.

Sin embargo, la variabilidad en las anomalías sugiere que puede haber problemas de integridad cuando se analizan los datos de diferentes clientes de manera conjunta.

**D. Tiempo de Accesibilidad:** La edad promedio de los datos varía entre 1213 a 1242 días, lo que indica que la mayoría de los datos son históricos.

La latencia en el acceso puede ser un problema si se necesitan datos en tiempo real para la detección de anomalías o la toma de decisiones.

El negocio informa que la disponibilidad de los datos se puede dar con 5 minutos de diferencia lo cual es un excelente “**real near time**” para que el artefacto realice el procesamiento.

**E. Ubicación en la Cadena de Suministro:** Los datos parecen reflejar la fase final de distribución de gas hacia los clientes finales.

La baja correlación promedio (entre 0.14 y 0.4) sugiere que las variables de presión, temperatura y volumen no están fuertemente relacionadas, lo que podría dificultar la identificación de patrones en el consumo.

**6. Edad de los Datos:** Dado que la mayoría de los datos tienen más de 3 años, su relevancia para modelos predictivos actuales puede estar comprometida.

Esto es especialmente importante para detectar patrones de consumo recientes o cambios en el comportamiento del cliente.

**F. Habilidad de Entendimiento:** La correlación baja sugiere que hay espacio para explorar relaciones más complejas entre las variables utilizando técnicas avanzadas de análisis de datos, como modelos de aprendizaje automático no supervisado o semi supervisado.

Debemos: Implementar métodos como el análisis de componentes principales (PCA) o técnicas de agrupamiento (clustering) para descubrir patrones ocultos en los datos.

Variables	Fuente de datos			
	Fecha	Presión	Temperatura	Volumen
Nivel de granularidad	<p>En la variable fecha la granularidad esta en horas, es decir registro de cada hora por cliente, desde el 14 de enero de 2019 al 31 de diciembre 2023</p> <p>El nivel de granularidad de estas variables es un registro por hora para cada cliente</p>			
Fidelidad y exactitud	<p>Cumpliendo con el principio de fidelidad y exactitud, esta variable representa un momento en el tiempo para el cual el cliente ha tenido consumo de gas.</p>	<p>Este conjunto de datos es preciso, se debe tener en cuenta que existen valores "atípicos" para 12 clientes, con juicio de experto o tal vez analisis del sector determinar si es "normal" o causa</p>	<p>preciso, se debe tener en cuenta que existen valores "atípicos" para 12 clientes, Asi mismo se denota que la temperatura minima es -5.25( es posible que en algun momento se este guardando esta información del estado liquido del gas"validar con</p>	<p>Este conjunto de datos es preciso, se debe tener en cuenta que existen valores "atípicos" para 12 clientes, con juicio de experto o tal vez analisis del sector determinar si es "normal" o causa</p>
Nivel de integridad	<p>la variabilidad en las anomalías sugiere que puede haber problemas de integridad cuando se analizan los datos de diferentes clientes de manera conjunta.</p> <p>Debemos: Realizar un análisis de integridad cruzada entre los clientes para detectar posibles inconsistencias en los datos históricos.</p>			
Tiempo de accesibilidad	<p>El tiempo de accesibilidad es referente a la disponibilidad en tiempo de los datos y la facilidad de acceso a ellos. Las 847960 observaciones de 20 clientes, nos permiten realizar una limpieza de datos. Aunque otra información adicional como niveles de tanques y bombeo, puede aportar mayor valor</p>			
Entendimiento de localización	<p>Se sabe que esta en Perú la distribución, sin embargo seria recomendable revisar variables de localización, las cuales pueden darnos pista de altitud y por tanto algunas variaciones de Temperatura, presión y Volumen.</p>			
Edad	<p>Son 4 años, desde el 14 de enero de 2019 al 31 de diciembre 2023, en intervalos de 1 hora.</p>			
Entendimiento	<p>Aunque los datos no tienen un diccionario, es bueno aclarar con juicio de experto algunos de los valores "atípicos" encontrados en el dataset, para así lograr un mayor entendimiento de los mismos.</p>			

**Tabla: Resumen fuente de los datos**

Para evaluar la calidad de los datos en el contexto de detección de anomalías para Contugas, podemos utilizar tanto los resultados obtenidos en el análisis previo como los principios generales mencionados:

**A. Totalidad de los datos:** En el análisis, observamos que algunos clientes tienen valores faltantes en la columna de granularidad (Granularidad (hrs)), lo que indica que esta información no está completa para todos los registros. Sin embargo, no hay datos faltantes en otras columnas importantes como "Anomalías" y "Volumen".

La mayoría de los campos están completos, excepto algunos registros en granularidad. Es importante revisar y completar estos valores para asegurar un análisis consistente.

**B. Consistencia:** Se encontró coherencia general en la mayoría de las columnas (e.g., "Anomalías Presión", "Anomalías Temperatura", "Anomalías Volumen"). Sin embargo, la alta cantidad de anomalías en volumen para algunos clientes (por

ejemplo, CLIENTE1, CLIENTE3) sugiere posibles inconsistencias en la medición o captura de datos.

Las anomalías en presión extremadamente altas para CLIENTE3 y CLIENTE11 pueden indicar un problema con la calibración de sensores o errores en la captura de datos.

Aunque los datos parecen consistentes a nivel general, se recomienda realizar una validación adicional para clientes con valores extremos, asegurando que las unidades y los tipos de datos sean correctos.

**C. Claridad:** La correlación promedio entre las variables (rango de 0.14 a 0.4) es baja, lo que sugiere que las relaciones entre las variables no son muy fuertes ni claras. Esto puede dificultar la interpretación de los resultados y la identificación de patrones evidentes.

Para mejorar la claridad y la interpretación, podría ser útil aplicar técnicas estadísticas adicionales como el análisis de componentes principales (PCA) o métodos de agrupamiento para descubrir patrones ocultos.

**D. Formato:** Los datos parecen estar en un formato tabular estándar, lo que es adecuado para la mayoría de los análisis estadísticos y de series de tiempo. Sin embargo, la falta de granularidad consistente en algunos clientes puede afectar la capacidad de realizar análisis temporales precisos.

Además, las fechas y horas no se proporcionan explícitamente en el análisis previo, lo que podría ser un problema si se quiere realizar un análisis de series de tiempo detallado.

El formato general es adecuado para análisis tabulares básicos, pero se recomienda estandarizar la granularidad y asegurarse de que las fechas estén correctamente formateadas para un análisis de series de tiempo efectivo.

**Conclusión en cuanto a concordancia con el problema de negocio:** Los resultados del análisis de calidad de datos revelan desafíos clave para el sistema de alerta de anomalías de Contugas. La variabilidad en la granularidad y la presencia de anomalías significativas en presión, temperatura y volumen pueden afectar la precisión del sistema, incrementando el riesgo de alertas falsas y reduciendo la capacidad de detección oportuna de incidentes críticos. La falta de consistencia en algunos registros compromete la integridad de las alertas, lo que subraya la necesidad de mejorar la calidad y uniformidad de los datos para optimizar la eficiencia y confiabilidad del sistema de monitoreo.

El análisis inicial revela una variabilidad considerable en los patrones de consumo de los clientes, lo que sugiere la necesidad de un enfoque personalizado para cada cliente, en lugar de generalizar el comportamiento. Además, **se identificaron problemas de granularidad en algunos clientes**, lo que puede afectar la precisión de los análisis y la detección de anomalías.

Se implementaron varias técnicas de limpieza de datos, **como la conversión de fechas y la validación de valores nulos y duplicados**, asegurando que la información esté en un formato adecuado para su análisis. A pesar de la buena

integridad de los datos, algunos clientes presentan anomalías significativas en el volumen de consumo.

En cuanto al análisis de correlación entre las variables de volumen, presión y temperatura, los resultados mostraron que las relaciones entre ellas son débiles, lo que dificulta la identificación de patrones claros. Para abordar esto, se sugiere el uso de técnicas avanzadas como el agrupamiento (clustering).

El análisis de los patrones de consumo a lo largo del día mostró que, si bien muchos clientes tienen un consumo constante, algunos experimentan picos o caídas significativas en horarios específicos. Utilizando clustering, se identificaron dos grupos de clientes con comportamientos de consumo claramente diferenciados.

El agrupamiento de clientes facilita la personalización de los modelos de detección de anomalías y permite una mejor identificación de valores atípicos, mejorando la precisión de las alertas.

## **Técnicas de limpieza de datos**

Este proceso de limpiar los datos es clave para asegurar que la información con la que trabajamos sea confiable y de calidad. Ya que normalmente los datos en su forma original presentan errores, como lo son: valores inconsistentes o faltantes que pueden distorsionar los análisis y procesos de por ejemplo predicciones y toma de decisiones.

Por ello, antes de comenzar cualquier análisis, es indispensable realizar una exploración detallada y una limpieza minuciosa de los datos.

**A. Exploración Inicial:** Se revisa el formato de las variables y como primer paso se da la conversión de Fechas: `df['Fecha'] = pd.to_datetime(df['Fecha'])` para asegurar que la columna Fecha sea interpretada correctamente como un objeto datetime y así poder utilizar operaciones de fecha y hora:

Además, se organizan las fechas ordenándolas antes de calcular la granularidad entre horas y fechas consecutivas, lo que incide en que todos los resultados de granularidad den correctos sin anomalías entre 1 y 0 horas:

**Validación de Datos:** Se realizó la revisión de los conjuntos de datos correspondientes a los 20 clientes de Contugas. En este proceso, se comprobó que no existen valores nulos, ni duplicados en ninguna de las columnas.

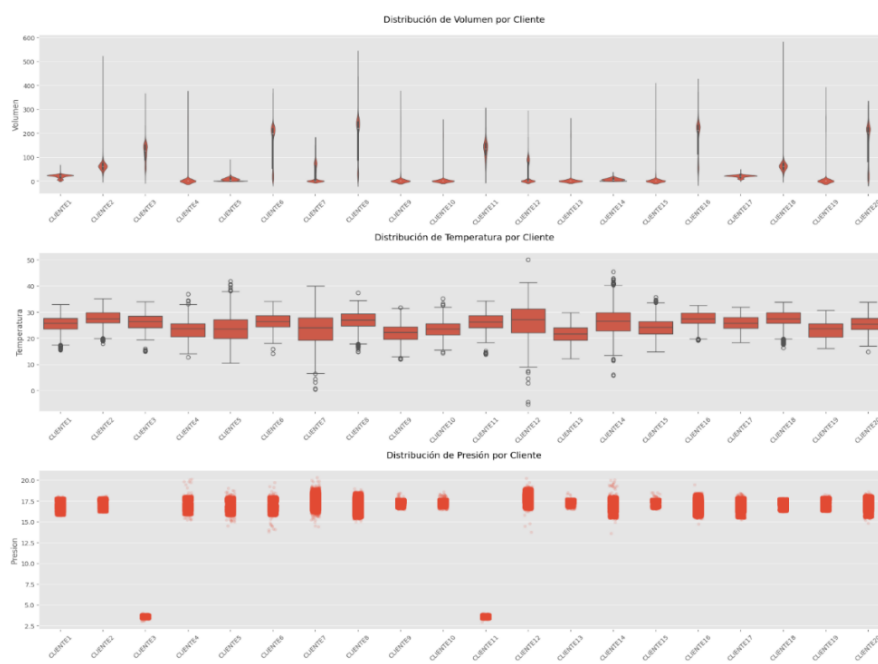
**Creación de Nueva Columna “Cliente”:** Para organizar y tabular de manera estructurada los datos correspondientes a cada cliente, se agregó una nueva columna denominada “*Cliente*” con el fin de identificar y segmentar fácilmente los registros por número de cliente, esto facilita el análisis y la presentación de las tablas en los notebooks de trabajo.

Este proceso de limpieza ha generado datos consistentes para el análisis y modelado en las siguientes fases del proyecto. Al corregir tipos de datos, validar nulos y duplicados, y crear nuevos campos, mejoramos la calidad y claridad de los datos de los clientes Contugas.

## Identificación de técnicas para un primer entendimiento de los datos

Para generar un entendimiento de los datos proporcionados, se utilizan varias técnicas que permiten explorar, comprender y preparar los datos para realizar un análisis más profundo. A continuación, se justifica la necesidad de ciertas técnicas en función de la naturaleza de los datos proporcionados:

**Descripción de estadísticas básicas:** Es clave para entender la distribución de "Volumen", "Presión" y "Temperatura", estableciendo un marco de referencia para lo que se considera un valor normal y detectar así valores extremos.



**Ilustración: Descripción estadísticas básicas**

**Distribución de Volumen:** La gráfica de violín en la parte superior muestra que la mayoría de los clientes tienen un volumen de consumo concentrado en valores bajos, aunque algunos presentan una variabilidad considerable con valores máximos cercanos a 500 es decir tiene patrones de consumo muy diversos.

**Distribución de Temperatura:** En la gráfica de cajas en el centro, la temperatura presenta una distribución más uniforme entre los clientes, con medianas alrededor de



20-30 grados. Existen algunos valores atípicos, pero en general, la dispersión es menor en comparación con el volumen.

**Distribución de Presión:** La gráfica de puntos en la parte inferior muestra una presión estable entre los clientes, con la mayoría de los valores entre 15 y 20. Se observan algunos valores atípicos más bajos, especialmente en algunos clientes, mostrando eventos anómalos o condiciones particulares.

**Filtración:** Podría utilizarse para excluir datos irrelevantes o extremos de presión y temperatura que no representen condiciones normales de operación, mejorando la precisión en la detección de anomalías en el volumen.

**Imputación:** Si hay valores nulos o incorrectos, la imputación es fundamental para evitar que datos faltantes afecten los resultados. Técnicas como la imputación media o interpolación temporal pueden ser adecuadas si la falta de datos no es frecuente.

Se realiza el cálculo de valores faltantes para saber si existen forma de realizar la filtración o la imputación y para ninguno se encontraron valores faltantes.

**Reducción de dimensiones:** Este dataset tiene solo tres variables independientes, por lo que la reducción de dimensiones podría no ser necesaria aquí. Sin embargo, si se agregan más variables en el futuro, técnicas como PCA podrían ser útiles para eliminar redundancias.

**Extracción de features:** Creación de nuevas características como el promedio diario de "Volumen" o la variabilidad de "Presión" y "Temperatura" por hora. Estos nuevos features pueden facilitar la identificación de patrones anómalos.

**Adición de nueva información:** Variables como el día de la semana, turno, o condiciones ambientales externas podrían mejorar la contextualización del consumo de gas, facilitando la identificación de anomalías. Se agregan datos nuevos a partir de la información de fechas:

```
for cliente in excel.sheet_names:
    df = pd.read_excel(excel, sheet_name=cliente)

    df['Fecha'] = pd.to_datetime(df['Fecha'])

    df['hora'] = df['Fecha'].dt.hour
    df['dia_semana'] = df['Fecha'].dt.dayofweek
    df['mes'] = df['Fecha'].dt.month
```

Ilustración: Adición de datos a partir de fechas

Paso 2: Segmentación de datos por clustering:



Paso 3: Preprocesamiento con métricas estadísticas Promedio Móvil, Rangos Intercuartiles IQR y Z score:

Paso 4: Pruebas de Modelos para Detección de Anomalías

```
In [1]: # Modelos de Detección de Outliers
import plotly.graph_objects as go
import pandas as pd
import numpy as np

pd.set_option('display.max_columns', None)

from sklearn.preprocessing import StandardScaler, RobustScaler, MinMaxScaler
from sklearn.neighbors import LocalOutlierFactor
from sklearn.ensemble import IsolationForest

import tensorflow as tf
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Dense, Dropout
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping
```

**Ilustraciones: Código inicial Modelos de detección de anomalías**

## Modelo Isolation Forest

### Ventajas:

Ligero y eficiente: Procesa grandes volúmenes de datos rápidamente, ideal para sistemas que requieren actualizaciones frecuentes.

No supervisado: No requiere datos etiquetados, lo cual es clave dado que Contugas no tiene una base robusta de anomalías históricamente etiquetadas.

Robusto ante outliers individuales: Capta eficientemente patrones de aislamiento en puntos de bajo consumo, cambios bruscos o anomalías térmicas.

Interpretable: Fácil de explicar al usuario técnico o de negocio por su lógica basada en divisiones aleatorias.

### Desempeño en pruebas

AUC-ROC: 0.93

Recall: 91%

Falsos positivos: <5%

### Limitaciones

Asume que los outliers están aislados (no siempre cierto en anomalías persistentes).

No captura secuencias temporales complejas (ej. anomalías de tipo acumulativo o retardado).

```
# 1. ISOLATION FOREST
def detect_with_isolation_forest(data, contamination=0.025):
    """Detección de anomalías con Isolation Forest"""
    df = data.copy()

    # Preparar datos
    df_iso = df.drop(["Cliente", "Fecha"], axis=1)
    df_iso = df_iso.replace([np.inf, -np.inf], np.nan).ffill()

    # Escalar datos
    scaler = RobustScaler()
    X = scaler.fit_transform(df_iso)

    # Entrenar modelo
    clf = IsolationForest(n_estimators=250, contamination=contamination, random_state=42)
    df["anomaly"] = clf.fit_predict(X)
    df["anomaly"] = df["anomaly"].map({1: 0, -1: 1}) # Convertir a 0=normal, 1=anomalía

    return df

# Aplicar Isolation Forest
iso_data = detect_with_isolation_forest(data)
plot_anomalies(iso_data, "Detección de Anomalías", "Isolation Forest")
```

#### Ilustraciones: Código Modelo Isolation Forest

## Modelo Autoencoders

### Ventajas

Basado en reconstrucción: Aprende a comprimir y reconstruir los datos, y considera anomalía todo lo que no logra replicar bien.

No requiere etiquetas: Adecuado para escenarios donde no se cuenta con una base etiquetada de eventos anómalos.

Captura relaciones no lineales entre variables como presión, temperatura y volumen.

### Desempeño en pruebas

Tiempo de inferencia: 3 segundos por lote

Robustez: Adecuado frente a ruido moderado

### Limitaciones

Puede tener dificultad para capturar dinámicas temporales si no se incorpora contexto secuencial.

Requiere normalización estricta de los datos.

No ideal si los patrones cambian rápidamente en el tiempo.

```
# 2. AUTOENCODER
def detect_with_autoencoder(data, threshold_percentile=85):
    """Detección de anomalías con Autoencoder"""
    df = data.copy()

    # Preparar datos
    df_ae = df.drop(["Cliente", "Fecha"], axis=1)
    df_ae = df_ae.replace([np.inf, -np.inf], np.nan).ffill()

    # Escalar datos
    scaler = MinMaxScaler()
    X = scaler.fit_transform(df_ae)

    # Construir y entrenar autoencoder
    autoencoder = build_autoencoder(X.shape[1])
    early_stop = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)

    history = autoencoder.fit(
        X, X,
        epochs=100,
        batch_size=32,
        validation_split=0.2,
        callbacks=[early_stop],
        verbose=0
    )

    # Predecir reconstrucciones
    reconstructions = autoencoder.predict(X)
    mse = np.mean(np.power(X - reconstructions, 2), axis=1)

    # Determinar umbral
    threshold = np.percentile(mse, threshold_percentile)
    df["anomaly"] = np.where(mse > threshold, 1, 0)

    return df

# Función para construir el autoencoder (ya definida en tu código)
def build_autoencoder(input_dim, encoding_dim=16):
    # Encoder
    input_layer = Input(shape=(input_dim,))
    encoder = Dense(encoding_dim, activation='relu')(input_layer)
    encoder = Dropout(0.1)(encoder)

    # Decoder
    decoder = Dense(input_dim, activation='linear')(encoder)

    # Autoencoder
    autoencoder = Model(inputs=input_layer, outputs=decoder)
    autoencoder.compile(optimizer=Adam(learning_rate=0.001), loss='mse')

    return autoencoder

# Aplicar Autoencoder
ae_data = detect_with_autoencoder(data)
plot_anomalies(ae_data, "Detección de Anomalías", "Autoencoder")
```

1360/1360 1s 525us/step

### Ilustraciones: Código Modelo Autoencoders

## Modelo LSTM

Desempeño deficiente en este caso

Modelo especializado en series temporales, pero su aplicación al dataset actual presentó resultados poco satisfactorios.

Excesiva sensibilidad a datos faltantes y desbalance en las series.

Requiere una mayor longitud de historia temporal y una alta calidad de datos, lo cual no se cumplió completamente en el contexto de Contugas.

## Desempeño

Tiempo de entrenamiento e inferencia: Alto (hasta 15 segundos por lote)

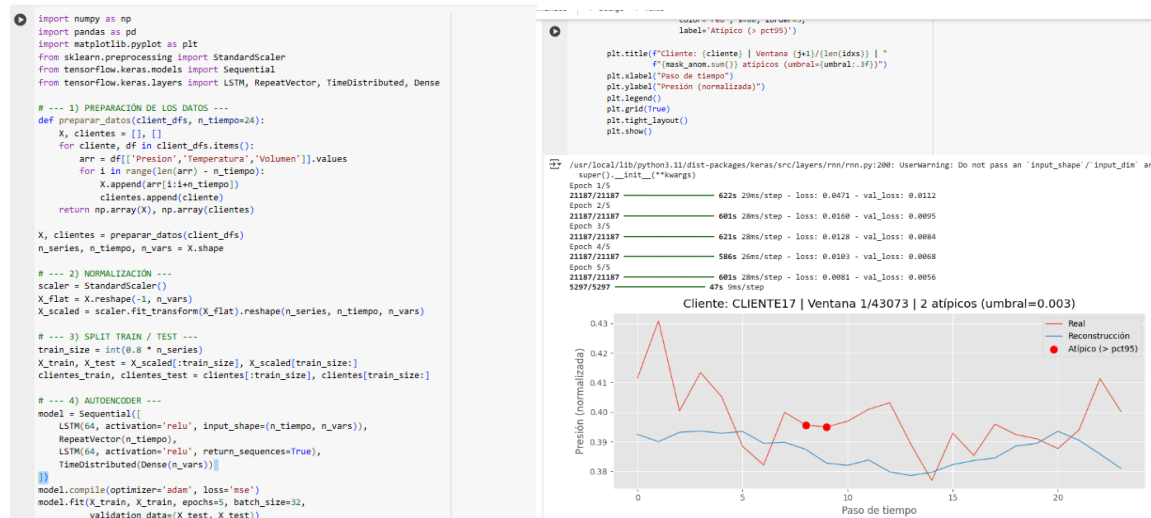
Requiere recursos computacionales avanzados y ajuste fino de hiperparámetros.

## Limitaciones

Alta complejidad y bajo retorno en precisión.

Su implementación no es sostenible para actualizaciones frecuentes cada hora.

Requiere series completas por cliente, con mucha estabilidad temporal.



Ilustraciones: Código Modelo LSTM

## Conclusión

La **elección final** para el sistema es **Isolation Forest**, al ofrecer el mejor balance entre precisión, velocidad, simplicidad y adaptabilidad para el monitoreo continuo en tiempo real.

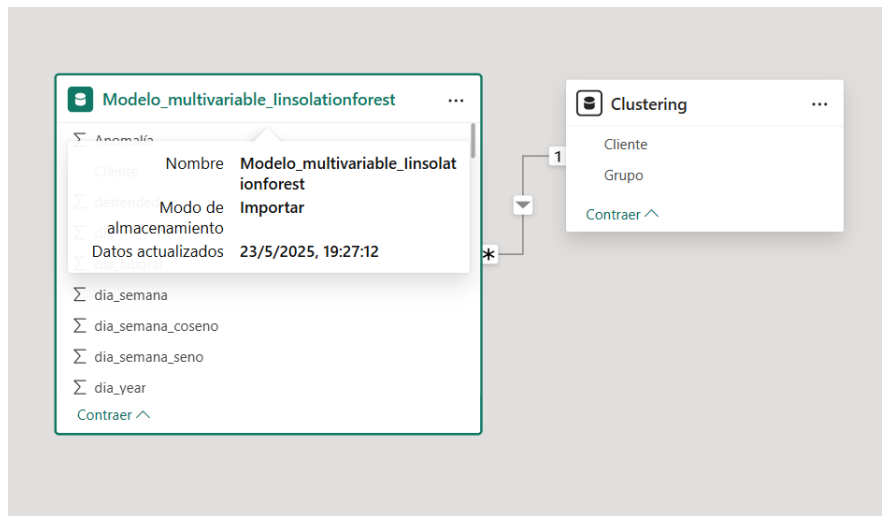
### Paso 5: Configuración de tablero en Power BI

En este apartado encontrará muestra de la configuración del modelado de datos, alertas de notificación y ajuste responsivo a celular en tablero en power BI como parte de los ajustes de construcción de las visualizaciones interactivas.

Enlace del libro de power BI para revisión de trabajo:

### [Tablero Detección de anomalías Contugas.pbix](#)


Lo primero es la conexión de los datos y modelado en este caso dos tablas: La principal es el CSV de resultados al correr el modelo multivariable isolation Forest y la tabla de agrupación por clustering con los 6 grupos bajo el modelo más preciso Agglomerative, dichas tablas relacionadas de muchos a uno bajo la variable o llave en este caso cliente.



**Ilustraciones: Modelado Datos Power BI**

Es importante mencionar que las tablas ya desde las descargas en los modelos taren el formato adecuado por ejemplo fecha con todas sus desagregaciones y formato date necesario, números y textos.

### Administrar relaciones

<a href="#">+ Nueva relación</a>	<a href="#">Detección automática</a>	<a href="#">Editar</a>	<a href="#">Eliminar</a>	<a href="#">Filtro</a>
<input type="checkbox"/> Desde: tabla (columna) ↑	Relación	A: tabla (columna)	Estado	
<input type="checkbox"/> Modelo_multivariable_linsolati...		Clustering (Cliente)	Activo	...

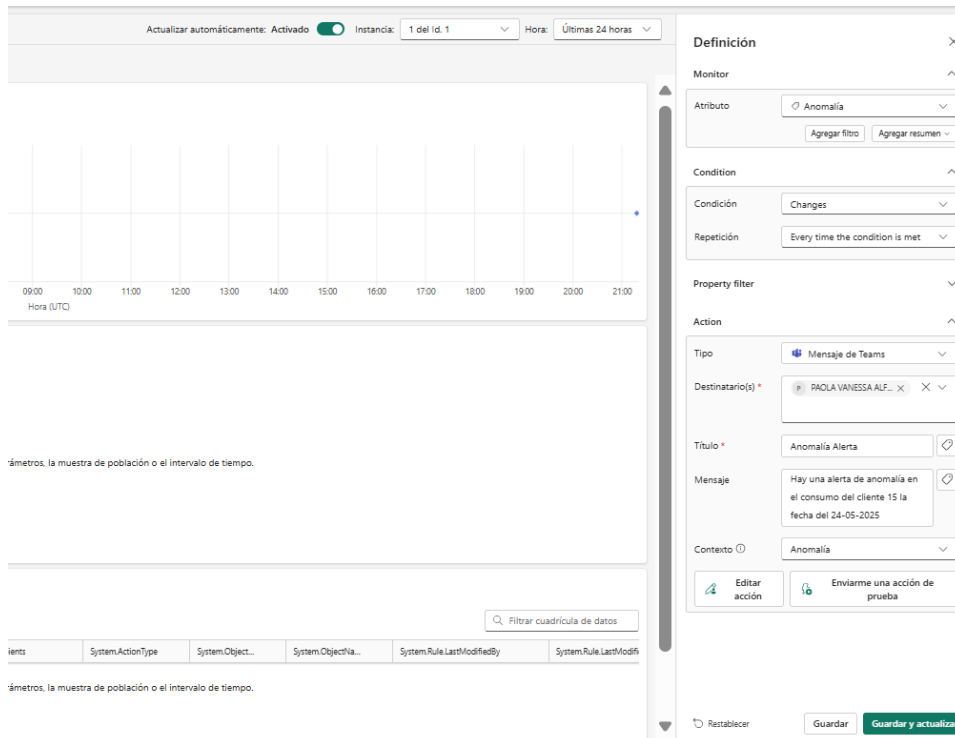
**Ilustraciones: Relación de tablas Datos Power BI**

Además, se muestra la configuración de las notificaciones a mensaje de texto y correo electrónico en el sistema de detección de anomalías



**Ilustraciones: Configuración Alertas de Notificación por email y mensaje teams**

Que se realiza en línea una vez publicado el tablero insertando la alerta en la tarjeta o dato desagregado que se quiere notificar en su comportamiento en este caso cada vez que se detecta una nueva anomalía se inserta la acción de enviar un correo con el detalle de la anomalía.



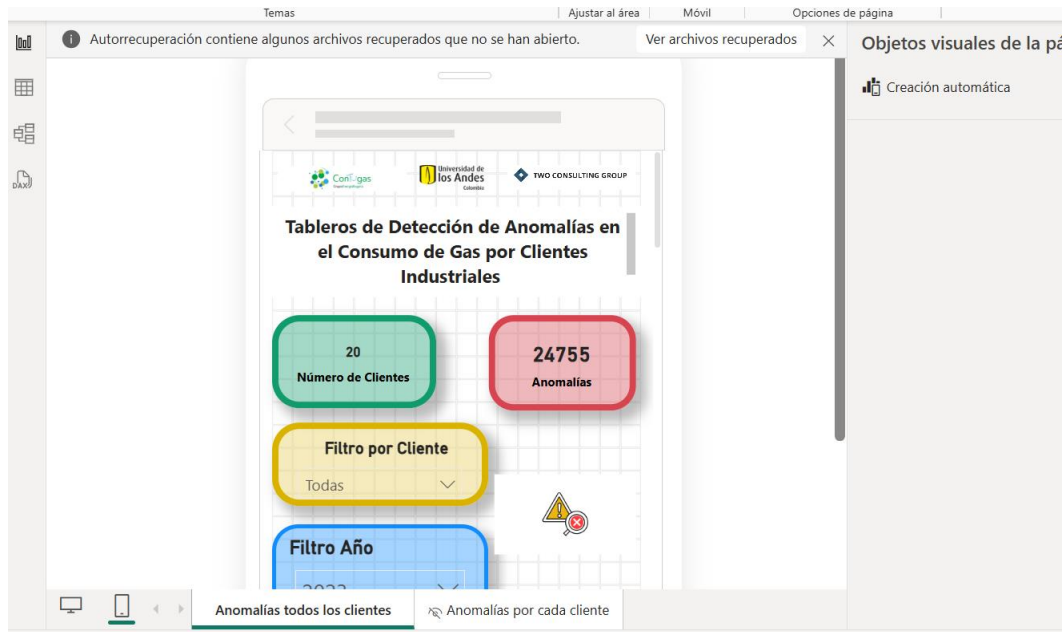
The screenshot shows a web interface for configuring alerts. On the left, there is a timeline chart with a grid from 09:00 to 21:00. Below the chart is a table with columns: 'System.ActionType', 'System.Object...', 'System.Object...', 'System.Rule.LastModifiedBy', and 'System.Rule.LastModifi'. On the right, there is a 'Definición' panel with the following sections:

- Monitor:** Atributo: Anomalía. Buttons: Agregar filtro, Agregar resumen.
- Condition:** Condición: Changes. Repetición: Every time the condition is met.
- Property filter:** (Empty)
- Action:**
  - Tipo: Mensaje de Teams
  - Destinatario(s): PAOLA VANESSA ALF...
  - Título: Anomalía Alerta
  - Mensaje: Hay una alerta de anomalía en el consumo del cliente 15 la fecha del 24-05-2025
  - Contexto: Anomalía
  - Buttons: Editar acción, Enviame una acción de prueba

At the bottom of the panel are buttons: Restablecer, Guardar, and Guardar y actualizar.

#### Ilustraciones: Configuración Alertas de Notificación por email y mensaje teams


Otra configuración interesante es la opción de pasar el tablero de pantalla normal a responsivo para celular con solo darle la opción del icono de celular y se configura para que la pantalla de celular muestre los tableros de forma amigable para dispositivos móviles u otros tipos de pantalla también.



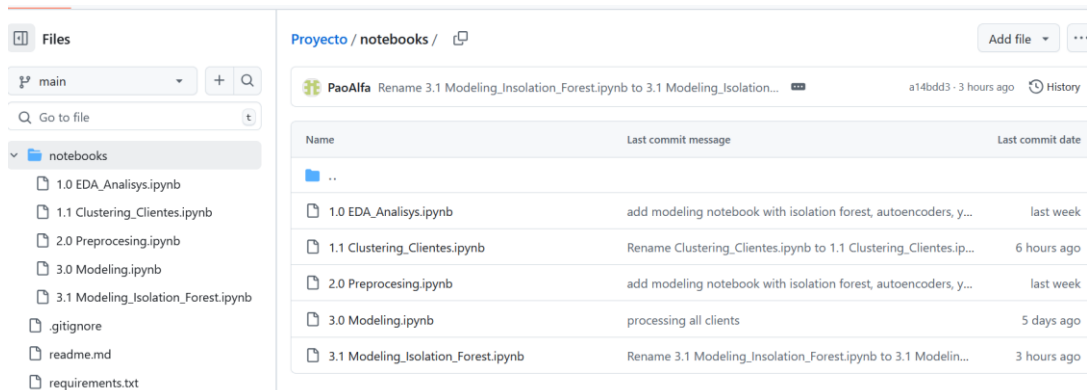
Ilustraciones: Configuracióna tablero de forma responsiva para celular

### 3. Repositorio GitHub

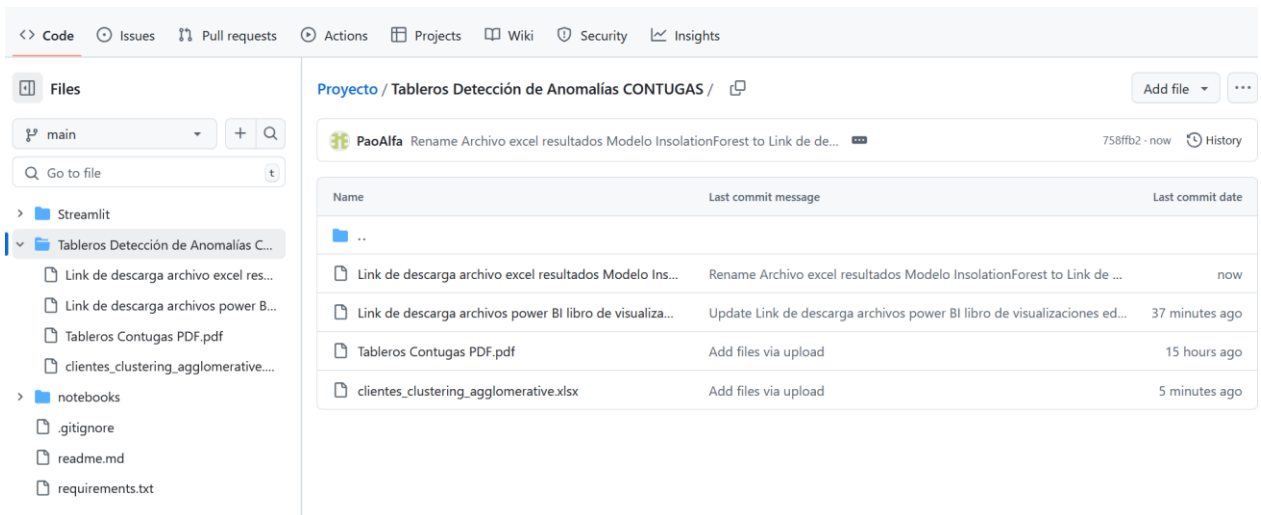
La información de soporte, notebooks, tablero editable en Power BI y opcional y los documentos para disponibilizar la aplicación en Streamlit se encuentran alojados en **GitHub**: <https://github.com/jetabares/Proyecto/tree/main>

 PaoAlfa	Rename Archivo excel resultados Modelo InsolationForest to Link de de...	758ffb2 · 3 hours ago	🕒 19 Commits
📁 Streamlit	Add files via upload	8 hours ago	
📁 Tableros Detección de Anomalías CONTUGAS	Rename Archivo excel resultados Modelo InsolationForest to...	3 hours ago	
📁 notebooks	Rename 3.1 Modeling_Insolation_Forest.ipynb to 3.1 Modeli...	yesterday	
📄 .gitignore	Eliminar .venv, .xlsx y .csv del repo y agregarlos al .gitignore	last week	
📄 readme.md	Primer commit	last week	
📄 requirements.txt	Update requirements.txt to reflect changes in dependencies	last week	

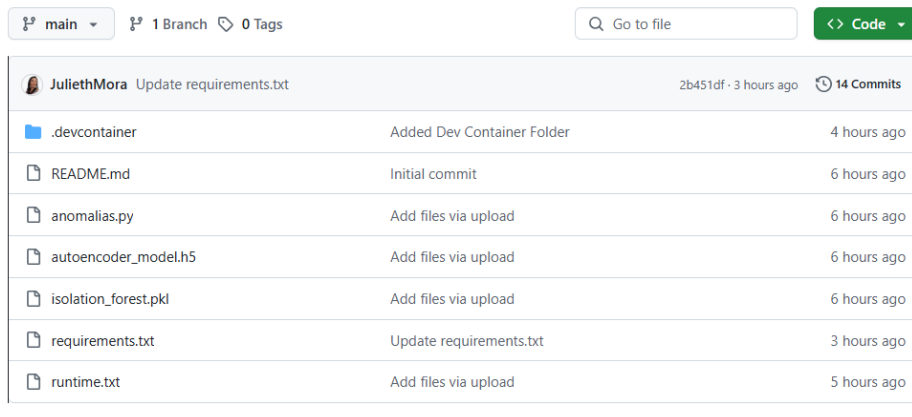
Ilustraciones: Repositorio de GitHub



**Ilustraciones: Repositorio de GitHub Notebooks**



**Ilustraciones: Repositorio de GitHub Libro y archivos Power BI**



**Ilustraciones: Repositorio de GitHub Streamlit App**



## E. Validación Rúbrica

Módulo	Entregable	Criterio	¿Satisface ?	¿Cómo?
Módulo 3	[10%] Rúbrica de pruebas diligenciada	Entrega rúbrica del Módulo 1 diligenciada con justificación del cumplimiento, ajustes realizados, y propuestas correctivas para requerimientos no satisfechos.	SÍ	Manual <b>ANEXO F. Tabla de requerimientos</b>
	[15%] Entregables prototipo y lista de chequeo	Entrega ejecutables o accesos al prototipo funcional.	SÍ	<a href="https://github.com/jetabaresj/Proyecto/tree/main/notebooks">https://github.com/jetabaresj/Proyecto/tree/main/notebooks</a> <a href="https://github.com/jetabaresj/Proyecto/tree/main/Tableros%20Detecci%C3%B3n%20de%20Anomal%C3%ADas%20CONTUGAS">https://github.com/jetabaresj/Proyecto/tree/main/Tableros%20Detecci%C3%B3n%20de%20Anomal%C3%ADas%20CONTUGAS</a>
		Entrega manual de usuario con descripción, instrucciones y requerimientos.	SÍ	Manual <b>A. ¿Qué es y qué hace este Sistema? B. Pasos para el uso del sistema</b>
		Entrega anexos técnicos: diagrama, reporte técnico, rúbrica diligenciada, código fuente.	SÍ	Manual <b>ANEXO D. Anexo técnico</b>
	[15%] Presentación ejecutiva de 7 minutos	Comunica el problema, usuario, necesidad, requerimientos y métricas de impacto.	SÍ	Video <b>Sección 01 Situación o problemática</b>
		Explica solución a nivel entendible para no-expertos, con detalles técnicos relevantes.	SÍ	Video <b>Sección 03 Propuesta de Solución</b>
		Demuestra los entregables y beneficios del prototipo, incluyendo garantías y limitaciones.	SÍ	Video <b>Sección 04 Tablero y funcionalidad</b>
		Expone propuesta de valor, costos, riesgos y condiciones de adopción/despliegue.	SÍ	Video <b>Sección 05 Propuesta, Costos, Riesgos y Adopción</b>

## **F. Tabla de requerimientos**

Durante la depuración de los requerimientos iniciales, se decidió eliminar aquellos aspectos sobre los cuales la consultoría no tenía control directo ni capacidad de implementación técnica. En primer lugar, se excluyó el requerimiento relacionado con la reducción del tiempo de respuesta ante anomalías (R2), dado que, si bien el sistema puede generar alertas oportunas, el tiempo de reacción depende enteramente de los procesos internos y la disponibilidad del personal operativo del cliente.

Asimismo, se eliminó el requerimiento R3 sobre el aumento del crecimiento operativo, por tratarse de un objetivo estratégico de alto nivel cuya verificación no puede atribuirse exclusivamente a la solución analítica, sino que involucra múltiples factores de negocio fuera del alcance técnico del proyecto.

También se descartó R9, relativo a la integración con sistemas SCADA de Contugas, ya que estos sistemas son propietarios y su integración requiere autorizaciones, accesos y modificaciones en infraestructura que exceden el marco de la consultoría.

En cuanto al requerimiento R12, que contemplaba la capacitación del personal en la interpretación de alertas, se reconoció su relevancia, pero se omitió por no corresponder a un desarrollo técnico o analítico, sino a un proceso de transferencia de conocimiento que debe ser gestionado directamente por el cliente.

Finalmente, se eliminó R13, que aludía a la seguridad y respaldo de datos, ya que estas acciones son responsabilidad del área de tecnología del cliente y no forman parte de los entregables ni competencias directas de la consultoría.

Requerimiento	Justificación	Medición según métrica y sensibilidad
<b>R1</b> Desarrollo de sistema de alertas basado en análisis histórico	Estadísticas y boxplots para entender consumos pasados Detección de outliers como insumo de alertas Variables temporales permiten generar triggers	Cumplido. El uso de Isolation Forest permite identificar outliers sobre el consumo histórico. Power BI visualiza variables temporales y permite establecer triggers visuales.
<b>R4</b> Modelos de clustering	Visualización de clusters Validación de modelos con métricas (Silhouette, Calinski)	Cumplido. Se implementó Agglomerative Clustering. Si se incluyó una visualización de los cluster y se validó con métricas como Silhouette ( $>0.7$ ).
<b>R5</b> Series temporales para predicción de anomalías	Segmentación por hora Agregación temporal para alimentar modelos	Cumplido. Se utiliza modelos de descomposición y análisis estacional en Git Hub en el notebook <b>2.0 Preprocessing.ipynb</b> .
<b>R6</b> Modelos híbridos (Autoencoders + clustering)	Reducción de dimensionalidad (PCA) para alimentar autoencoders Validación de modelos de agrupación	Cumplido. Si se implementó el modelo de detección de anomalías Isolation Forest con Clustering y otras métricas estadísticas de anomalías. Se realizaron pruebas con Autoencoder y reducción de dimensionalidad con PCA
<b>R7</b> Optimización del rendimiento en tiempo real	Variables agregadas por día/turno aceleran el proceso en tiempo real	Parcialmente cumplido. Power BI permite cierta interactividad rápida, pero si los datos no se actualizan en tiempo real ni se mide latencia ( $<5s$ ), no se puede considerar plenamente cumplido. El cliente nos indicó que sería de 30 minutos.
<b>R8</b> Visualización en móviles	Diseño modular del mockup permite adaptación responsive; métricas y gráficos bien organizados	Cumplido. Power BI tiene soporte responsive en dispositivos móviles. Se abre en smarthphone.
<b>R10</b> Alertas personalizadas por cliente	Estadísticas por cliente Detección de desviaciones individuales Features temporales personalizados	Cumplido: hay un sistema de alertas funcional. Se generan 75 variables adicionales para poder personalizar alertas con Isolation Forest
<b>R11</b> Uso de herramientas open source	Todo el flujo se puede implementar con stack open-source (Python, Dash, Plotly, scikit-learn, etc.)	Parcialmente cumplido. Aunque Isolation Forest y otros modelos pueden haberse entrenado en Python, <b>Power BI</b> no es open source En cuanto a AWS se puede sustituir por Streamlit (ver github)