Detección de anomalías en el consumo de gas de clientes industriales de la filial Contugas.

Mauricio Gonzalez Caro

Eva Karina Díaz Gavalo

Juan Felipe Padilla Sepúlveda

Andrés Eduardo Quiñones Ortiz

Universidad de los Andes

Maestría en Inteligencia Analítica de Datos

2025

Introducción

Este reporte tiene como objetivo describir el proceso seguido para la implementación, evaluación y selección definitiva de un modelo analítico que permita detectar anomalías en variables operacionales clave (presión, temperatura y volumen de gas natural) en clientes industriales de Contugas. El propósito final es garantizar una operación eficiente y minimizar los costos operativos derivados de anomalías no detectadas o detectadas tardíamente.

Inicialmente se abordará el análisis exploratorio de los datos (EDA), incluyendo verificación de calidad, detección de faltantes, análisis de distribuciones, identificación visual de posibles anomalías, y generación de variables temporales relevantes. Posteriormente, se realizará el tratamiento previo requerido para los modelos seleccionados, implementando procesos de imputación de valores faltantes, eliminación de duplicados, detección y tratamiento de valores atípicos, escalado de variables, entre otros.

Una vez preparados los datos, se planteará una preselección de modelos no supervisados adecuados al contexto del problema. Luego, se definirá un esquema riguroso de validación, ajustando parámetros e implementando métricas específicas para evaluar y comparar su desempeño, dada la ausencia de etiquetas. Finalmente, se presentarán los resultados obtenidos y se propondrá el modelo más adecuado, identificando componentes pendientes para su completa implementación dentro del prototipo final.

Entendimiento y preparación de los datos

Análisis de calidad de los datos

En aras de entender correctamente los datos y poder escoger el mejor modelo de detección de anomalías, se debe evaluar la calidad de los datos suministrados por Contugas, por lo tanto, se deben reconocer algunas de sus dimensiones de valor, entre estas se encuentran:

- Granularidad: Los datos tienen granularidad horaria (24 registros diarios), facilitando un análisis detallado del consumo y la detección precisa de anomalías, lo que resulta altamente beneficioso para los objetivos analíticos del proyecto.
- Fidelidad y exactitud: Contugas es garante de que los medidores que capturan la información son precisos y deben estar al día con los mantenimientos preventivos para asegurar la confiabilidad de los registros.
- Edad: Se dispone de cinco años completos de información histórica (14/01/2019 al 31/12/2023). Aunque este histórico es robusto, se enfatiza la importancia especial de los registros más recientes para detectar tendencias actuales.

A continuación, se describirán los aspectos relacionados con las características de los datos:

- **Formato:** Se verificaron los tipos de datos de cada columna, confirmando su adecuación. Las variables operativas (volumen, temperatura y presión) están en formato numérico (float), las fechas en formato datetime64 apropiadas para series temporales, y la columna Cliente originalmente en tipo object será convertida a tipo category para optimizar la memoria durante operaciones analíticas.

#	Column	Non-Null Count	Dtype
0	Fecha	847960 non-null	datetime64[ns]
1	Presion	847960 non-null	float64
2	Temperatura	847960 non-null	float64
3	Volumen	847960 non-null	float64
4	Cliente	847960 non-null	object

#	Column	Non-Null Count	Dtype
0	Fecha	847960 non-null	datetime64[ns]
1	Presion	847960 non-null	float64
2	Temperatura	847960 non-null	float64
3	Volumen	847960 non-null	float64
4	Cliente	847960 non-null	category

- **Completitud:** Se esperaba un total de 870,240 registros (43,512 registros por cada uno de los 20 clientes). Sin embargo, se identificaron 847,960 registros existentes y 22,280 faltantes, representando un 2.56% general de datos ausentes. Clientes específicos como el 10 y el 13 tienen mayores porcentajes de faltantes (hasta 5.64%), sugiriendo posibles anomalías operativas según información de los stakeholders. Cabe aclarar que estos faltantes se diferencian claramente de registros con valor cero, los cuales representan mediciones válidas. Los registros faltantes serán gestionados en la fase de limpieza.

Año	Mín.	Máx.	Días	Registros
2019	14/01/2019	31/12/2019	352	8448
2020	01/01/2020	31/12/2020	366	8784
2021	01/01/2021	31/12/2021	365	8760
2022	01/01/2022	31/12/2022	365	8760
2023	01/01/2023	31/12/2023	365	8760
Total			**1813**	**43512**

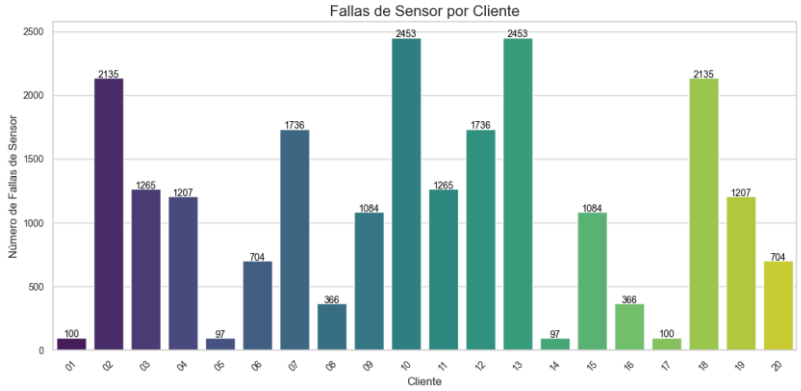
- **Consistencia:** Las unidades empleadas cumplen con lo esperado (temperatura en °C, presión en bar, volumen en m^3), lo que asegura coherencia en su interpretación. Aun así, mediante análisis visual (boxplots) y estadístico (percentiles 1 y 99), se detectaron inconsistencias potenciales:
    - Clientes como 03 y 11 registran presiones por debajo de 5 bar, fuera del rango operativo esperado.
    - Clientes 07 y 12 presentan temperaturas cercanas o inferiores a 0°C, lo que podría indicar errores de medición.
    - Clientes 04 y 09 registran volúmenes próximos a cero con alta frecuencia, pudiendo tratarse de patrones particulares de consumo o fallas de medición.
    - Clientes 06 (volumen) y 07 (temperatura) muestran rangos considerablemente amplios, sugiriendo comportamientos atípicos o problemas de registro.
    - Estas inconsistencias requerirán un análisis posterior para validar si corresponden a condiciones reales o a errores instrumentales.

- **Claridad:** La claridad de los datos está asegurada por su estructura y etiquetado coherente, facilitando la interpretación y análisis posteriores. Las unidades y formatos utilizados (datetime64 y float) cumplen con los estándares técnicos necesarios para el análisis temporal y numérico, simplificando la realización de técnicas estadísticas y transformaciones para modelado de series de tiempo. De igual forma, esto se puede corroborar a través de la implementación de técnicas de análisis estadísticas presentes en este informe y en el repositorio que contiene el pre-procesamiento de datos y análisis exploratorio. Posteriormente se abordarán los resultados asociados al conjunto de datos en general y de manera particular por cliente.

En conclusión, si bien deben ser procesados, los datos son concordantes con el problema del negocio y de acuerdo a lo identificado en esta fase se procederá a realizar la limpieza y/o imputación pertinente para revisar con mayor profundidad las inconsistencias presentadas. En este sentido, a continuación, se describe el proceso detallado de limpieza.

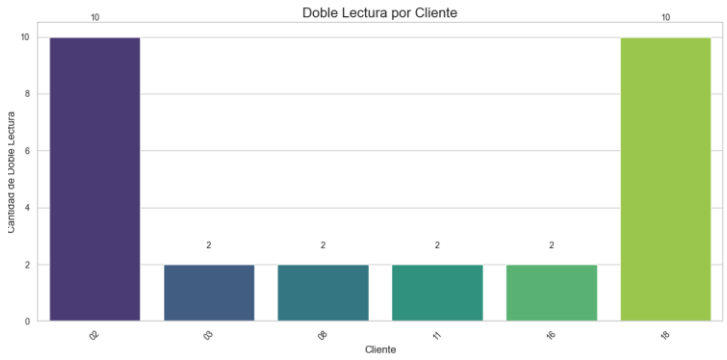
Imputación de Valores Faltantes

Se creó una serie temporal continua por cliente desde el 2019-01-14 00:00:00 hasta el 2023-12-31 23:00:00 (registros cada hora), lo cual permitió identificar valores faltantes para cada cliente. Estos faltantes, causados principalmente por fallas en los sensores según stakeholders. Para tratarlos se decidió aplicar interpolación lineal, la cual es adecuada en este caso porque preserva la tendencia general de los datos sin introducir ruido. Una vez completado el proceso para cada cliente, se unieron los resultados en un único DataFrame, organizando los datos por cliente y fecha y agregando una columna adicional Sensor_Error que identifica las fallas en los sensores.



Eliminación de duplicados

Se identificaron y eliminaron duplicados exactos basados en las columnas Cliente y Fecha. Adicionalmente, se marcaron registros con múltiples lecturas diferentes en la misma hora (Doble_Lectura), preservándolos debido a que reflejan potenciales problemas en sensores. Estos registros se conservarán para entrenar al modelo en el reconocimiento de patrones de falla específicos.



Verificación de Valores Nulos por Variable

Los valores nulos en presión, temperatura y volumen fueron identificados y etiquetados mediante la columna Error_Lectura. Posteriormente, se imputaron utilizando interpolación lineal, garantizando la continuidad de los datos sin introducir sesgos adicionales. Al finalizar este proceso, se aseguró la completitud de todos los registros para su uso en análisis posteriores.

Generación de variables temporales

La extracción de características temporales a partir de la columna Fecha constituye un paso en el análisis de detección de anomalías. Este proceso permite transformar la información para que sea interpretable tanto en análisis exploratorio como en el entrenamiento del modelo.

Las variables temporales nos ayudarán a capturar relaciones entre el tiempo y las variables dependientes (Presión, Temperatura, Volumen). Por ejemplo, ciertos patrones de consumo de gas pueden ser diferentes entre días laborales y fines de semana. Además, al incluir información temporal, el modelo puede detectar anomalías relacionadas con horarios inusuales de consumo o variaciones atípicas en determinados días o meses.

	Fecha	Presion	Temperatura	Volumen	Cliente	Año	Sensor_Error	Double_Reading	Mes	Día	Hora	Día_Semana	Fin_de_Semana
0	2019-01-14 00:00:00	17.732563	28.209354	20.969751	01	2019	0	0	1	14	0	0	0
1	2019-01-14 01:00:00	17.747776	28.518614	17.845739	01	2019	0	0	1	14	1	0	0
2	2019-01-14 02:00:00	17.758916	28.230191	20.975914	01	2019	0	0	1	14	2	0	0
3	2019-01-14 03:00:00	17.727940	27.811509	20.592299	01	2019	0	0	1	14	3	0	0
4	2019-01-14 04:00:00	17.746484	27.795293	21.690626	01	2019	0	0	1	14	4	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
870249	2023-12-31 19:00:00	15.751139	27.460652	204.457549	20	2023	0	0	12	31	19	6	1
870250	2023-12-31 20:00:00	15.614858	27.010382	186.512096	20	2023	0	0	12	31	20	6	1
870251	2023-12-31 21:00:00	15.598944	26.709100	204.456461	20	2023	0	0	12	31	21	6	1
870252	2023-12-31 22:00:00	15.730040	27.266090	203.695596	20	2023	0	0	12	31	22	6	1
870253	2023-12-31 23:00:00	15.624457	27.481288	201.534548	20	2023	0	0	12	31	23	6	1

870254 rows × 13 columns

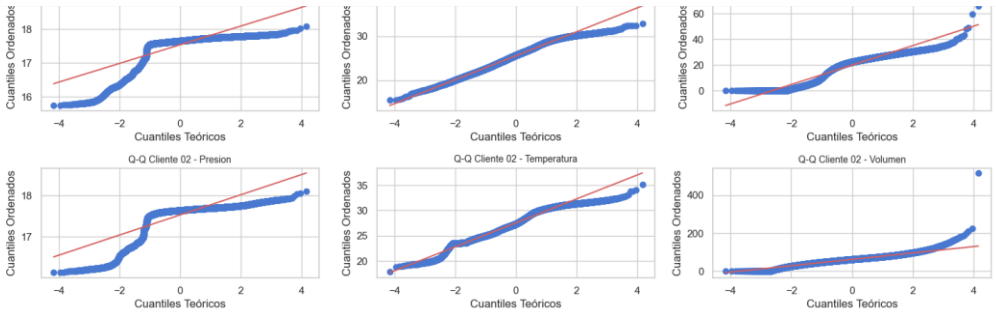
Escalado y estandarización de variables

Según los gráficos Q-Q y los histogramas, las variables de presión y temperatura parecen seguir distribuciones relativamente normales o con ligeras desviaciones. Estas variables son independientes del uso directo y no presentan comportamientos drásticamente sesgados.

Por otro lado, Volumen tiene algunos clientes con distribución extremadamente sesgada con valores atípicos que reflejan el comportamiento de uso. Esto es esperable, ya que el volumen depende de patrones de consumo, lo que puede generar datos altamente no normales y sesgados hacia valores bajos.

Para el proceso de escalado tenemos que garantizar que todas las variables (Presión, Temperatura, Volumen) estén en la misma escala por cliente, permitiendo al modelo considerar sus efectos de manera proporcional. Por esto elegimos Estandarización (Z-score) como técnica de escalado debido a su capacidad para preservar valores atípicos para el caso de la variable volumen, lo que es esencial para la detección de anomalías, y en cuanto a las variables de Presión y Temperatura las ajustara correctamente sin perder sus patrones normales.

Aunque este proceso no es necesario en este punto, nos permite observar en detalle si nuestras variables siguen una distribución normal y poder determinar la mejor técnica para escalar nuestras variables cuando sea necesario introducir los datos al modelo, asegurando que todas las variables contribuyan de manera equitativa, lo que evita que variables con magnitudes mayores (por ejemplo, Volumen) dominen el análisis.



Análisis del comportamiento de los ceros en Volumen

El análisis de registros con valores de volumen igual a cero nos permite comprender los patrones de consumo de gas por cliente y detectar otra posible anomalía. En este contexto, la presencia de ceros puede reflejar comportamientos normales asociados con horarios sin consumo, días no laborables, o horarios no operativos. Sin embargo, un comportamiento irregular o una frecuencia inusualmente alta de ceros o de consumos podría indicar problemas, como fallas en los sensores o anomalías en el suministro de gas.

Mes	1	2	3	4	5	6	7	8	9	10	11	12
Cliente												
01	2.714286	1.142857	4.545455	8.583333	6.666667	2.500000	5.384615	4.384615	1.200000	1.000000	1.000000	1.555556
02	1.000000	1.000000	1.058824	1.071429	1.000000	1.000000	1.000000	1.000000	1.125000	1.000000	1.000000	1.214286
03	1.166667	1.125000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000
04	17.737226	17.776978	17.353333	18.000000	18.080537	17.150685	17.765517	17.370629	17.537415	17.691781	17.903448	17.340000
05	4.718750	5.039370	5.148148	5.114504	5.753968	5.492308	5.143939	4.854839	5.043860	4.794643	5.009259	4.250000
06	3.369231	4.523810	5.049383	5.012658	1.303030	1.612903	1.586957	2.142857	1.984375	1.738462	1.818182	1.352941
07	0.000000	0.000000	11.311111	22.033613	21.958333	23.240000	23.193548	23.329032	23.273333	21.569231	1.000000	0.000000
08	1.100000	1.000000	1.187500	2.666667	1.457143	1.136364	1.458333	2.724138	4.265306	3.204082	1.615385	2.608696
09	21.643939	21.074627	21.758170	22.006803	21.310811	22.550336	22.098039	22.340136	22.333333	22.045752	21.978873	21.219858
10	21.237410	22.029851	21.543046	21.885714	22.607843	21.489796	21.267123	21.527027	22.162963	22.260274	21.462069	21.350993
11	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12	1.000000	1.000000	20.276923	23.783333	23.253425	23.773333	23.800000	23.819355	23.590604	22.309735	1.000000	0.000000
13	20.608696	21.942857	20.935484	21.546667	21.059211	21.130137	21.717241	21.413043	21.492754	21.474026	21.091549	21.046053
14	3.281250	2.970874	3.971963	3.904762	4.233645	4.592233	4.081633	3.505263	3.659341	2.967391	3.126437	2.744444
15	21.485714	21.124088	21.934211	21.598639	21.701299	21.048611	21.859060	21.731544	21.391608	22.331126	21.156028	21.328767
16	1.000000	1.000000	0.000000	2.333333	2.250000	1.000000	1.000000	3.521739	5.266667	2.916667	1.000000	2.111111
17	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	9.000000	8.500000	0.000000	0.000000	0.000000	0.000000
18	1.055556	1.083333	1.062500	1.166667	1.050000	1.090909	1.000000	1.000000	1.111111	1.066667	1.111111	1.047619
19	19.706767	18.926471	19.163265	19.972414	19.851351	19.602740	18.797297	19.097222	19.165517	20.120805	19.761905	19.720779
20	1.390244	3.019608	1.380000	2.929825	1.266667	1.294118	1.966667	2.236842	2.159420	1.966102	1.791667	1.816327

Clientes como 11 y 16 casi no tienen registros en cero, indicando que cualquier ocurrencia de estos valores es una posible anomalía o error del sensor. El cliente 9 muestra un patrón consistente de consumo limitado (87% registros en cero), por lo cual incrementos en estos registros podrían indicar anomalías. Se identificaron

tendencias específicas en ciertos clientes (por ejemplo, cliente 1), relacionados con días y meses particulares, permitiendo definir umbrales personalizados para minimizar falsos positivos.

Correlación entre variables

La correlación entre las variables de presión, temperatura y volumen fue analizada de forma individual para cada cliente, con el objetivo de identificar relaciones significativas entre estas.

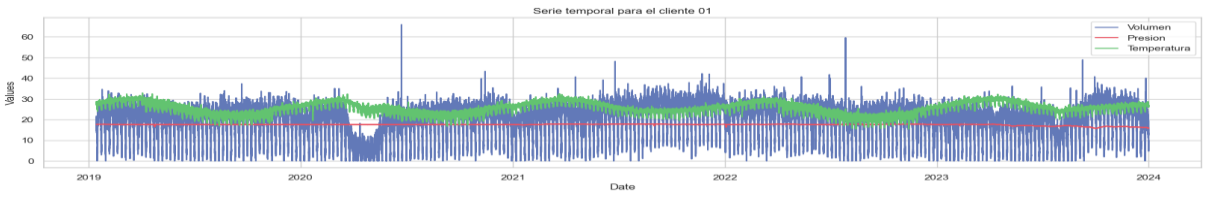
El análisis no reveló correlaciones positivas altas (mayores al 80%) que pudieran representar un problema para la inclusión de las variables en el modelo. Sin embargo, las correlaciones negativas altas observadas entre presión y volumen en dos clientes destacan la necesidad de analizar estas relaciones más a fondo para descartar problemas en los sensores, aunque estas correlaciones no necesariamente representan un problema para el modelo si se entienden y justifican en el contexto del sistema medido.

Estadísticas Descriptivas por cliente

El análisis de métricas básicas por cliente permite entender en detalle las características de los datos y cómo se comportan las variables Presión, Temperatura y Volumen para cada cliente. Este conocimiento nos permite identificar comportamientos normales y posibles desviaciones que puedan indicar anomalías en los datos.

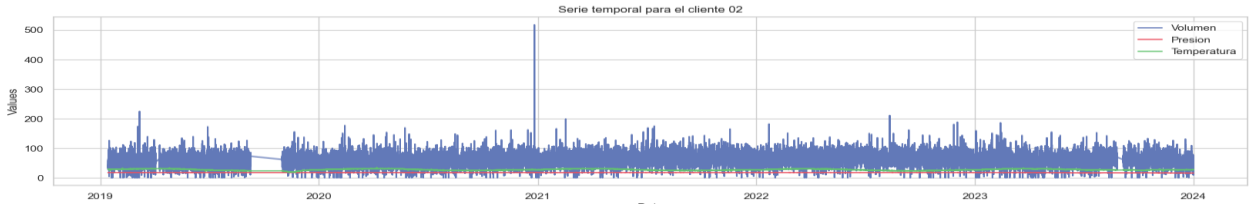
Cliente	Media_Presion	Mediana_Presion	Min_Presion	Max_Presion	Std_Presion	Media_Temperatura	Mediana_Temperatura	Min_Temperatura	Max_Temperatura	Std_Temperatura	Media_Volumen	Mediana_Volumen	Min_Volumen	Max_Volumen	Std_Volumen
1	17.53625213	17.65165391	15.7423367	18.07427385	0.3581138228	25.57531447	25.70483523	15.40180299	32.86911236	2.753574965	19.98746154	22.47499162	0	65.93664367	7.941182018
2	17.52698551	17.6389632	16.12901534	18.10640165	0.317948669	27.55745237	27.36175508	17.88405905	35.20834588	2.411789917	62.01965233	62.41729122	0	517.5648676	17.15896678
3	3.568596265	3.558166502	3.057170922	4.037030329	0.08055458042	26.37374785	26.36959043	14.93303167	34.00923348	2.635774304	117.65269	125.7107764	0	356.7240076	39.74806952
4	17.64248104	17.77983361	15.21499578	20.11293101	0.3876830284	23.25762481	23.69991558	12.71181264	36.89704413	3.023691605	17.46274105	0	0	363.0097759	61.05530001
5	17.48960099	17.61476238	14.48989628	19.0093516	0.3701592997	23.66453051	23.49573648	10.50063757	41.76223215	5.050807379	7.818851123	7.11855607	0	89.24505063	6.409062311
6	17.57060025	17.7074914	13.8100258	19.72086983	0.3955908457	26.37734961	26.34598658	14.15736626	34.18633759	2.721491817	152.8864087	197.8061898	0	366.6563824	85.33262725
7	17.4963032	17.49926935	14.41826092	20.30785175	0.4469026988	23.47364352	23.81295303	0.4334355568	39.9822393	4.949659346	26.47758337	0	0	175.7638578	33.68950794
8	16.70524172	16.56740921	14.99649003	18.95095539	0.5976634836	26.81886683	26.89673523	14.83870285	37.36895856	2.971478273	178.7658053	221.2128445	0	522.7808913	90.76039264
9	17.46679857	17.49667904	16.48647744	18.43046854	0.1568153645	22.065828	22.24626975	11.82010432	31.78396368	2.92792477	12.55790487	0	0	366.01612	49.16531134
10	17.47354564	17.49307259	16.46986398	18.57307889	0.1334072615	23.43801243	23.47012272	14.29103213	35.06972131	2.652476579	12.57862378	0	0	247.0729895	45.48629268
11	3.545151658	3.535421804	2.934872944	3.954039051	0.08242906812	26.33604697	26.22621627	13.73102519	34.35527609	2.71318852	131.7079678	136.2644622	0	298.2595728	34.22126292
12	17.83861354	17.83919038	13.74092216	20.23211047	0.418673319	26.51916288	26.76140178	-5.257899119	50.01985255	5.051278082	32.71658466	0	0	284.4750866	41.69664729
13	17.4936315	17.51000224	16.50507704	18.47104679	0.1177161026	21.53435767	21.49253935	12.24051121	29.75336441	2.861443496	9.909209743	0	0	253.8679162	40.38810436
14	17.51733643	17.63306811	13.61687666	20.02839495	0.3688608445	26.66132599	26.54237859	5.67693284	45.37817643	4.96641776	7.301760533	6.748793379	0	36.79315006	5.25644634
15	17.49543649	17.51432544	16.51449568	18.49323363	0.1388790194	24.0172316	24.10875841	14.71287364	35.66134854	3.117431027	12.46292809	0	0	398.0420267	47.40982952
16	16.70529469	16.61962095	14.73452103	19.44077978	0.4918821844	27.48175338	27.50048629	19.14610127	32.65845563	2.449027805	178.1860625	210.8837973	0	409.8722119	74.59863557
17	17.53389527	17.6535681	15.26570349	18.44537774	0.3797899558	25.84854279	25.76024922	18.44509636	31.9175602	2.361762178	20.56900489	21.75539679	0	48.50483292	5.562752539
18	17.55530945	17.64606968	16.2151046	17.82135868	0.2494967407	27.5577884	27.42364673	16.38078009	33.78936219	2.512794206	61.48135427	61.5697559	0	577.4134248	18.09880785
19	17.6694781	17.786668	16.24748361	18.25861885	0.3223937848	23.23930589	23.76087224	16.18198213	30.64224273	3.024355407	15.74127058	0	0	378.2678031	59.20646127
20	17.58113704	17.68932866	14.84973693	18.57820294	0.4069268782	25.40269963	25.36955456	14.76708132	33.80800577	2.683028454	160.4448343	201.7133671	0	315.8841531	82.91591749

Cliente 01: Muestra un comportamiento estable en términos de presión y temperatura, con promedios de 17.53 y 25.57, respectivamente. Su desviación estándar en ambas variables es baja, indicando un sistema operativo controlado. Sin embargo, su promedio de volumen (19.98) y desviación estándar (7.94) sugieren cierta variabilidad en el consumo, aunque no significativa. Este cliente tiene un porcentaje de ceros bajo (1.41%), lo que refuerza la estabilidad en su operación. Es poco probable que presente anomalías, pero cualquier aumento drástico en ceros podría ser un indicador de problemas.

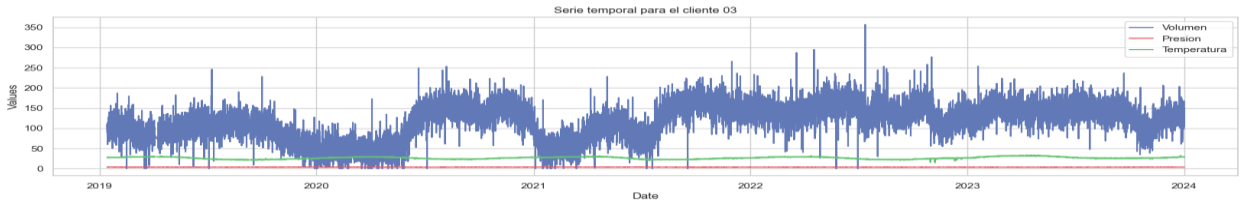


Cliente 02: Tiene un comportamiento regular en presión (promedio de 17.52) y temperatura (27.55), con bajas desviaciones estándar. Su consumo promedio de volumen es 62.01, siendo moderado en comparación con

otros clientes, pero su desviación estándar es relativamente alta (17.15), lo que indica fluctuaciones en el consumo. Con solo un 0.32% de ceros, este cliente tiene patrones operativos estables y controlados con uno que otro outlier.

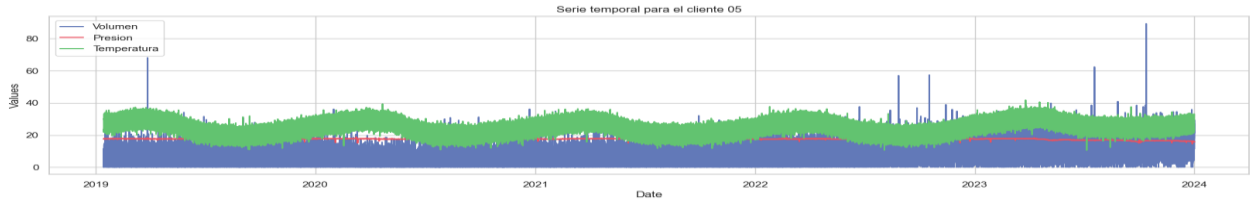


Cliente 03: Presenta un comportamiento único con la presión más baja entre todos los clientes (promedio de 3.56) y una desviación estándar muy baja (0.08), indicando un sistema de baja presión altamente estable. Sin embargo, su consumo de volumen promedio es relativamente alto (117.65) con una desviación significativa (39.74), que oscila entre los meses del año como se ve en la gráfica.

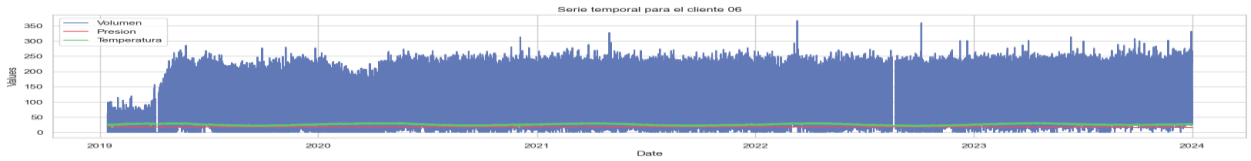


Cliente 04: Este cliente presenta alto porcentaje de ceros (70.50%). Su promedio de consumo descartando los periodos de inactividad es de 59.19 y con una desviación sd alta de 100.82, como se representa en la gráfica, el cliente presenta consumo intermitente pero alto.

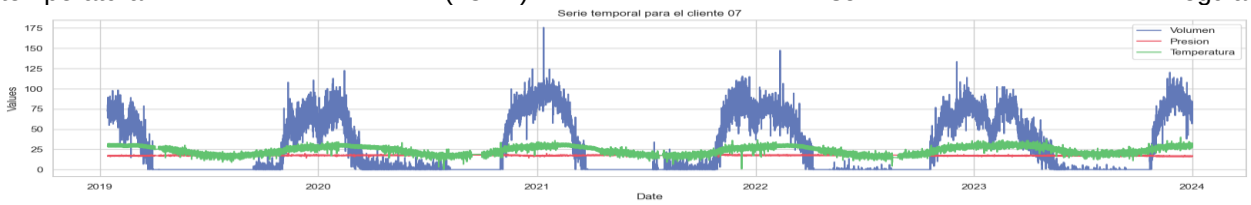
Cliente 05: Posee un promedio de volumen bajo (7.81) con una desviación estándar de 6.40. Su presión y temperatura son moderadas (17.48 y 23.66), con valores extremos notables en temperatura que llegan hasta 41.76, lo que sugiere un sistema que puede operar en condiciones extremas.



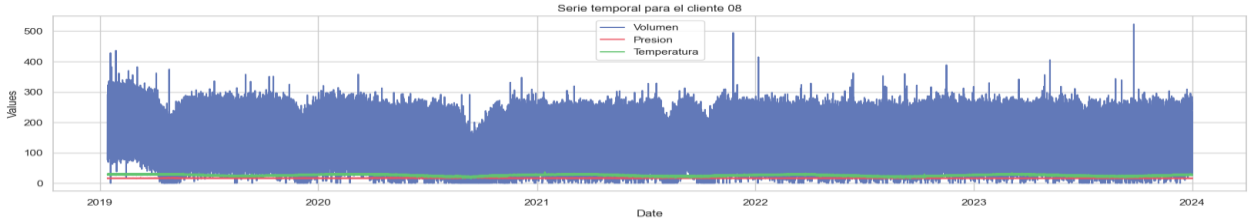
Cliente 06: Uno de los clientes con mayor consumo (152.88 en promedio), pero también con alta variabilidad. La presión (promedio de 17.57) y temperatura (26.37) son estables, aunque también muestran cierta variabilidad. Su porcentaje de ceros (4.76%) no es elevado.



Cliente 07: Tiene un comportamiento distintivo ya que parece operar solo por la temporada de finales y principios de año y con un promedio de volumen bajo (56.96) cuando está en actividad. Su presión (17.49) y temperatura (23.47) son regulares.



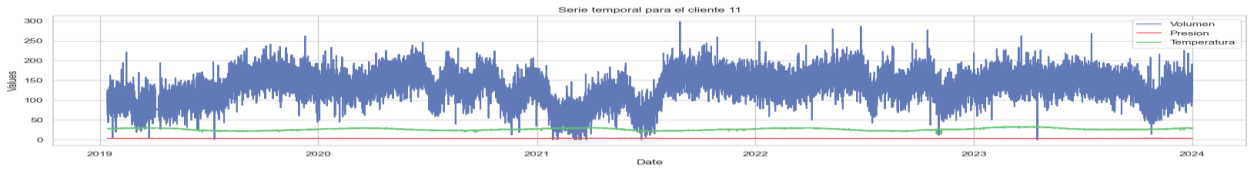
Cliente 08: Otro cliente de alto consumo con un promedio de volumen de 178.76, pero también con alta variabilidad (desviación estándar de 90.76). Su presión promedio (16.70) y temperatura (26.81) son estables. Con un porcentaje de ceros bajo (1.93%), este cliente es altamente operativo, y cualquier incremento en ceros podría ser una señal de alerta.



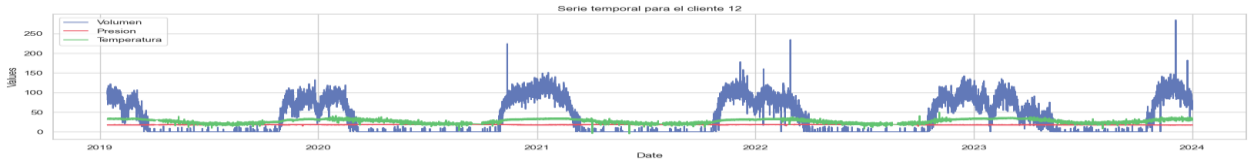
Cliente 09: Presenta un porcentaje de ceros extremadamente alto (87.62%), lo que refleja un comportamiento normal para este cliente. Su promedio de volumen (101.45 en actividad) aunque con una desviación estándar (102.52) lo que representa un comportamiento en picos como se puede observar en el gráfico.

Cliente 10: Patron muy similar al cliente 09 con un alto porcentaje de ceros (86.66%)y un promedio de volumen (94.30). Su presión y temperatura son consistentes y estables, con poca variabilidad.

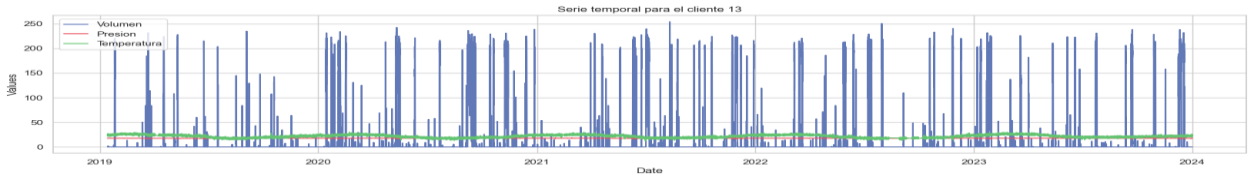
Cliente 11: Posee una presión promedio baja (3.54), con una desviación estándar mínima (0.08). Su volumen promedio (131.70) es moderado, con 0.01% de ceros.



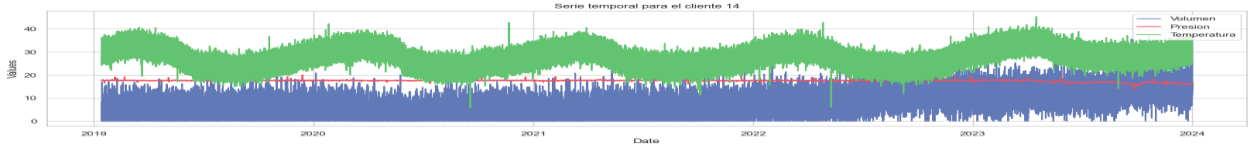
Cliente 12: Con un porcentaje de ceros (56.43%) indica períodos de inactividad que son normales,como se puede apreciar en la gráfica del comportamiento. Su promedio de volumen en actividad es de 75.09 con sd de 28.42.



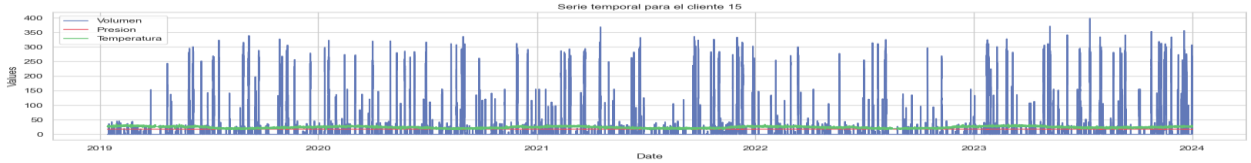
Cliente 13: Este cliente tiene porcentaje de ceros alto (85.60%), sin embargo vemos picos de uso que llegan hasta un máximo de 253.86. Temperatura y presión estables.



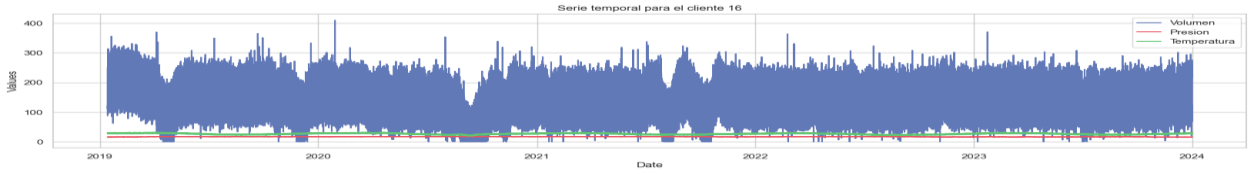
Cliente 14: Tiene promedios bajos en volumen (7.30) y cierta variabilidad (5.25), mientras su temperatura (26.66) tiene una variabilidad de 4.96, una de las más altas en el grupo. Este cliente tiene un porcentaje de ceros moderado (9.74%).



Cliente 15: Similar a los clientes 09, 10 y 13, alto porcentaje de ceros alto (86.88%), con consumos elevados cuando está en actividad con un promedio de 95.05.

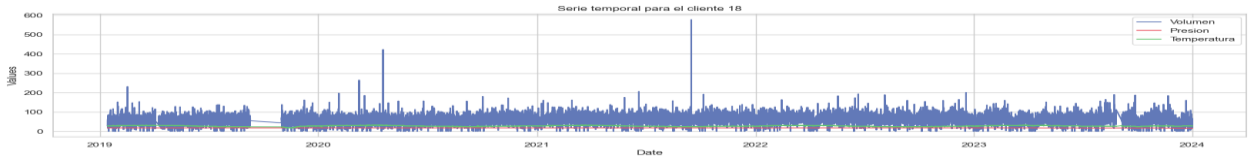


Cliente 16: Tiene un alto promedio de volumen (178.18) y una desviación significativa (74.59), lo que indica operaciones intensivas con alta variabilidad. Su presión (16.70) y temperatura (27.48) son regulares.



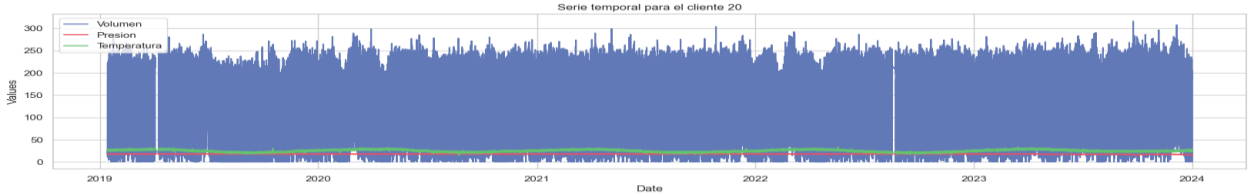
Cliente 17: Posee un comportamiento estable con consumos moderados en volumen (20.56) y una baja desviación estándar (5.56). Su porcentaje de ceros (0.14%) es mínimo, indicando que tiene un sistema altamente operativo.

Cliente 18: Con consumos moderados de volumen (61.48) y baja variabilidad. Su presión (17.55) y temperatura (27.55) son regulares, posee una larga pausa en 2019 pero que no se ha vuelto a repetir. Tiene unos picos de consumo que podrían ser algún tipo de anomalía. Es el cliente con el pico de consumo más alto de todo el cohort.



Cliente 19: Otro cliente con operación en pocas horas al día pero con alto consumo promedio en actividad (71.69). Tiene una presión y temperatura estable.

Cliente 20: Presenta un promedio alto de volumen (160.44) y una desviación significativa (82.91). Su presión y temperatura son estables, pero alta variabilidad en volumen.



Segmentación

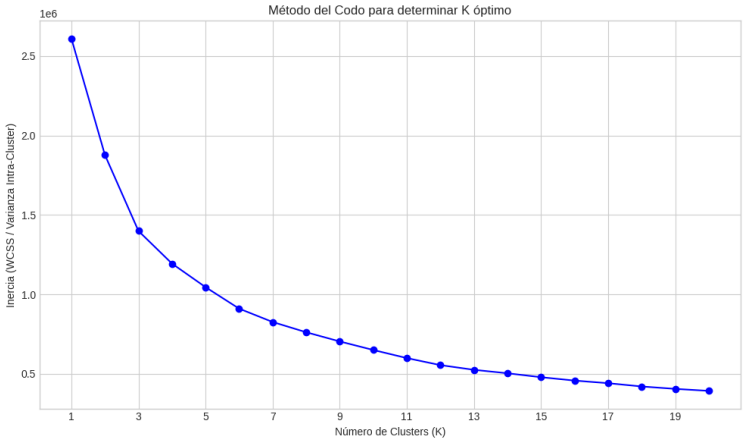
Como paso previo a la selección e implementación de modelos de detección de anomalías, se realizó un análisis de segmentación para identificar patrones comunes entre los 20 clientes industriales de Contugas. El propósito central fue determinar si es posible agrupar clientes con comportamientos operacionales similares en cuanto a presión, temperatura y volumen de gas. Esto permitiría optimizar esfuerzos, reduciendo la necesidad de desarrollar modelos analíticos individuales para cada cliente, simplificando el proceso y optimizando recursos técnicos y operativos.

Para este análisis, se contempló inicialmente la aplicación de dos metodologías estándar: el Método del Codo y el Coeficiente de Silueta. El Método del Codo busca identificar el punto donde añadir más grupos deja de proporcionar una reducción significativa en la varianza intra-cluster (WCSS o Inercia), mientras que el Coeficiente de Silueta evalúa la calidad de la segmentación midiendo qué tan bien separados están los grupos y qué tan cohesivos son internamente. La intención era utilizar ambos enfoques para obtener una perspectiva más robusta sobre el valor K más apropiado.

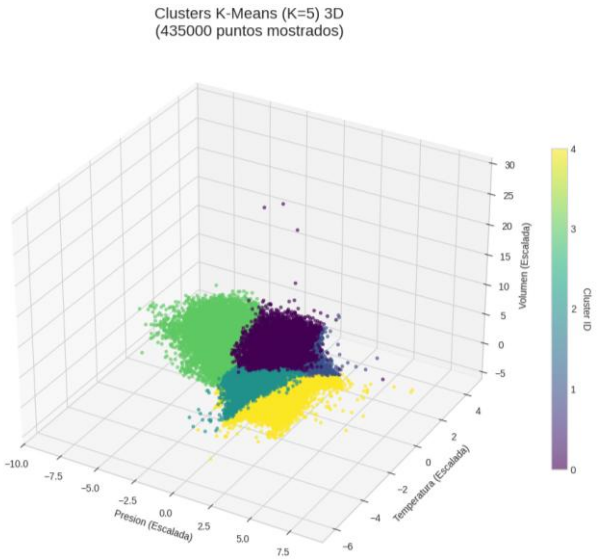
Sin embargo, durante la fase experimental, se encontró una limitación computacional significativa al intentar calcular el Coeficiente de Silueta. El conjunto de datos preprocesado contiene un total de 870,240 registros. El cálculo del Coeficiente de Silueta, que requiere evaluar distancias entre puntos dentro y fuera de los grupos, resultó ser extremadamente intensivo en tiempo al aplicarse repetidamente para cada valor potencial de K (desde 2 hasta 20) sobre este volumen de datos. Debido a que el tiempo requerido para completar esta evaluación excedía los límites prácticos del análisis, se tomó la decisión de prescindir del método de Silueta.

Análisis del Método del Codo para la Selección de K

Para determinar un número apropiado de grupos (K) para agrupar los datos de mediciones (Presión, Temperatura, Volumen), se aplicó el Método del Codo. Este método consiste en calcular la Varianza Intra-Cluster (WCSS o Inercia) para diferentes valores de K (en este caso, de 1 a 20) y buscar un punto de inflexión ("codo") en la gráfica resultante de Inercia vs. K. Como se observa en la siguiente figura, la Inercia disminuye drásticamente a medida que K aumenta de 1 a 4, indicando que los primeros grupos mejoran significativamente la cohesión de los grupos. La tasa de disminución se reduce notablemente a partir de K=4, y el "codo" más pronunciado, donde la adición de nuevos grupos comienza a ofrecer rendimientos decrecientes en la reducción de la varianza, se identifica visualmente alrededor de K=5. Aunque la curva continúa descendiendo suavemente después, K=5 representa el punto de equilibrio más claro entre minimizar la Inercia y evitar una complejidad innecesaria del modelo. Por lo tanto, el análisis mediante el Método del Codo sugiere que K=5 es un número óptimo de grupos para este conjunto de datos.

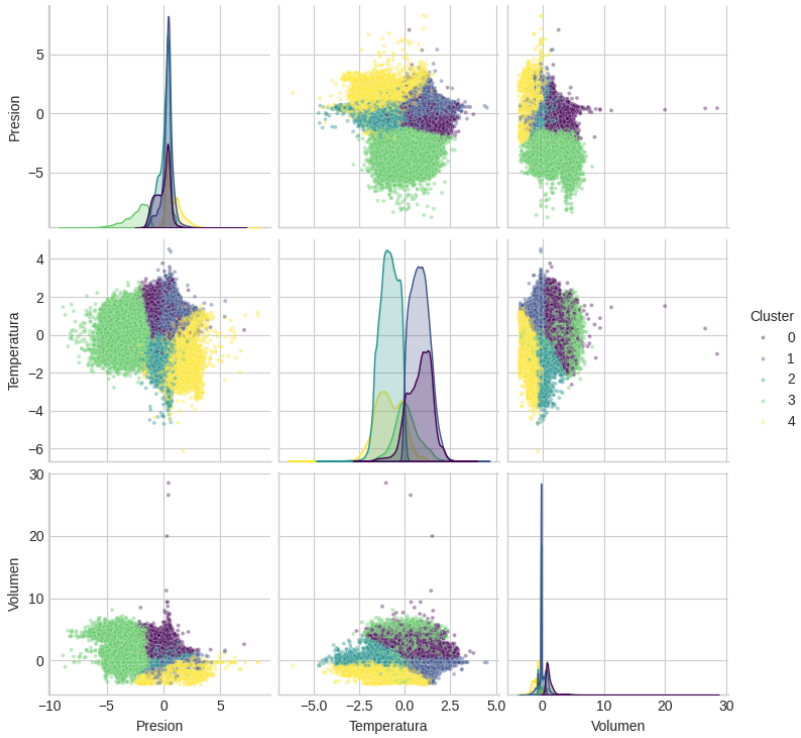


Con el objetivo de identificar patrones operativos distintos en las mediciones de los sensores, se creó una visualización inicial en tres dimensiones, utilizando una muestra representativa de 435,000 puntos (aproximadamente 50% del total), reveló la existencia de cinco grupos razonablemente definidos en el espacio tridimensional. Esta vista espacial destacó inmediatamente la separación del grupo 4 (amarillo), caracterizado por valores notablemente elevados en la dimensión de Volumen, y del grupo 0 (púrpura), que consistentemente agrupaba puntos con bajos valores, particularmente en Presión y Volumen. Los grupos restantes (1, 2 y 3, en tonos azulados y verdosos) mostraron mayor proximidad espacial, sugiriendo perfiles más similares o zonas de transición entre ellos.

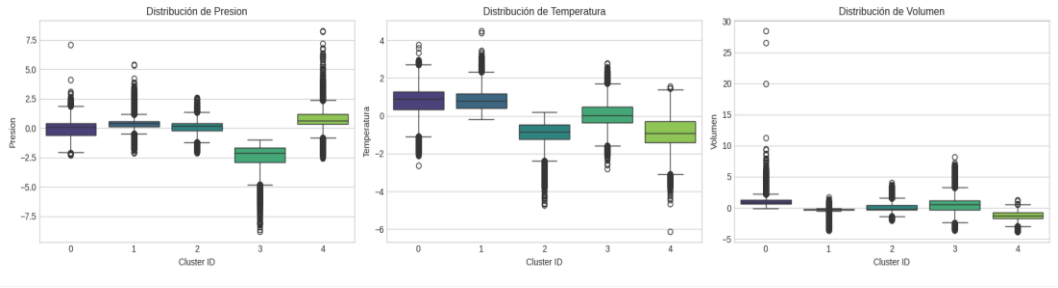


Para detallar estas observaciones, se generó un análisis bidimensional mediante Pair Plots. Esta matriz de gráficos confirmó la clara separación de los grupos 0 y 4, especialmente visible en las proyecciones que involucran la variable Volumen (Presión vs. Volumen y Temperatura vs. Volumen). En estas vistas, el

grupo 4 ocupa una región de alto volumen, mientras que el grupo 0 se concentra en la zona de bajo volumen y baja presión. Sin embargo, el Pair Plot también evidenció de forma más explícita el considerable solapamiento entre los grupos 1, 2 y 3, sobre todo en la proyección de Presión vs. Temperatura. Esta superposición sugiere que, si bien K-Means los asigna a grupos diferentes, estos tres grupos representan perfiles operativos con características más graduales o menos distintivas cuando se consideran solo esas dos variables.

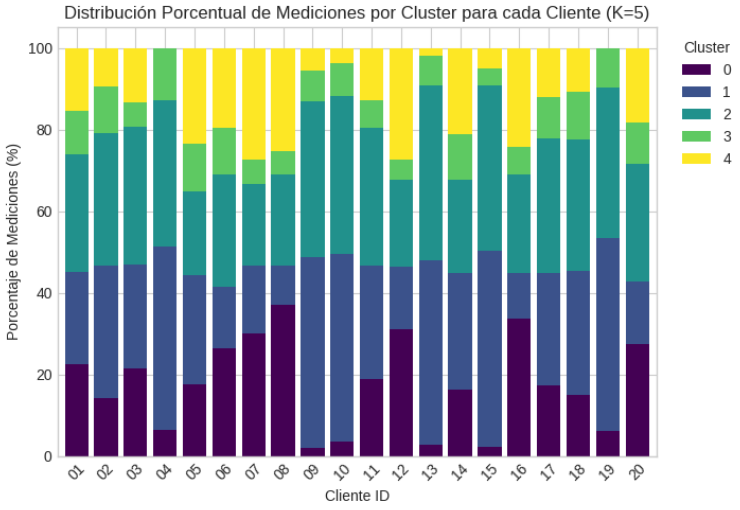


Finalmente, los Box Plots permitieron cuantificar las características distributivas de cada grupo para cada variable. Confirmaron que el grupo 4 no solo posee la mediana de Volumen más alta, sino también la de Temperatura. En contraste, el grupo 0 presenta consistentemente las medianas más bajas en las tres variables estudiadas. Los grupos intermedios (1, 2 y 3) se posicionan entre estos extremos, exhibiendo diferencias más sutiles: se observa una tendencia incremental en la mediana del Volumen del grupo 1 al 3, mientras que los grupos 2 y 3 comparten una mediana de Temperatura muy similar, distinguiéndose más en Presión y Volumen. En conjunto, estas visualizaciones complementarias validan la segmentación en K=5, ilustrando tanto los perfiles claramente diferenciados (grupos 0 y 4) como las relaciones más complejas y las características específicas de los grupos intermedios (grupos 1, 2 y 3).



Análisis de la distribución de clientes bajo el método del CODO:

La siguiente figura muestra cómo se distribuyen porcentualmente las mediciones de cada cliente entre los cinco clusters identificados. Se observa claramente que ningún cliente está asociado exclusivamente a un único cluster; todos presentan perfiles mixtos, indicando que transitan entre diferentes estados operativos (representados por los clusters) a lo largo del tiempo. Si bien algunos clientes muestran mayor afinidad por ciertos clusters, la presencia generalizada de múltiples colores en cada barra subraya un solapamiento significativo en el comportamiento operativo según esta segmentación. Esto sugiere que, aunque K-Means agrupa eficazmente las mediciones instantáneas, los clientes individuales no se ajustan a un único perfil estático.



Esta heterogeneidad intra-cliente y el solapamiento resultante tienen implicaciones importantes para la modelización predictiva. Asignar un perfil dominante a un cliente podría ocultar estados operativos menos frecuentes, pero potencialmente críticos, llevando a posibles errores en predicciones de comportamiento o riesgo. Dada la naturaleza dinámica inherente a las series temporales, un enfoque más robusto podría ser el uso de Dynamic Time Warping (DTW) multivariado. Al calcular la similitud directamente entre las trayectorias temporales completas de los clientes (considerando Presión, Temperatura y Volumen simultáneamente), el clustering basado en distancias DTW podría generar grupos más significativos

basados en patrones dinámicos globales, superando las limitaciones de agrupar puntos de datos individuales aislados en el tiempo.

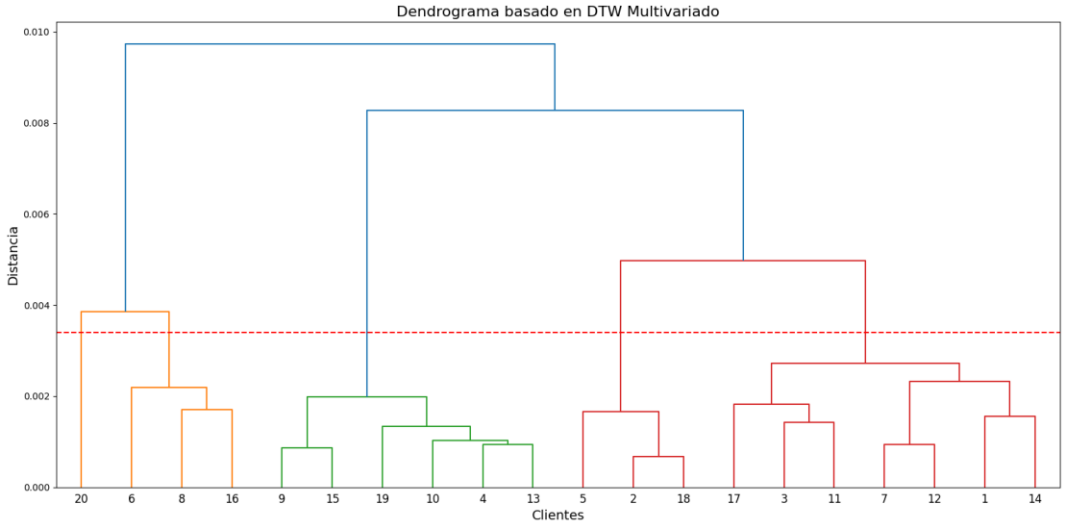
Distancia DTW (Dynamic Time Warping)

Con el propósito de identificar patrones de comportamiento en el consumo de gas entre diferentes clientes industriales, se implementó un algoritmo de agrupamiento jerárquico basado en la distancia DTW (Dynamic Time Warping) multivariada.

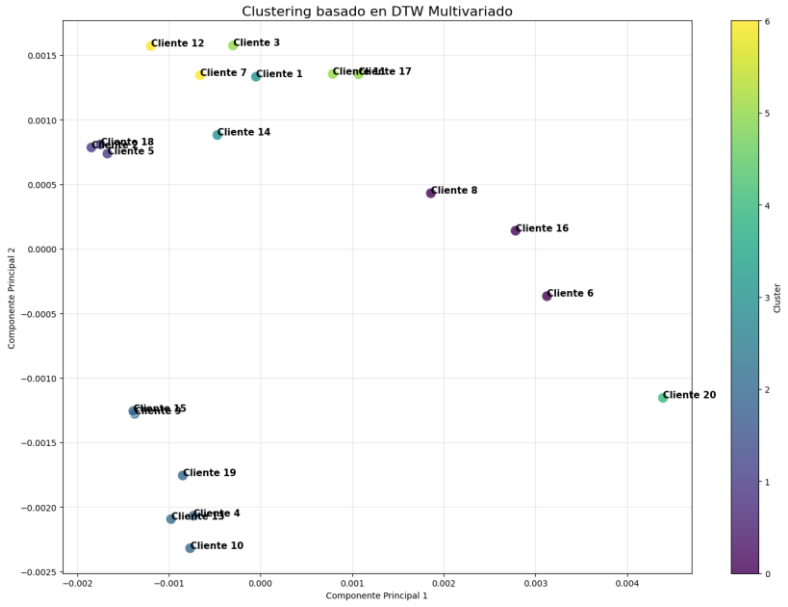
A diferencia de métodos convencionales como K-Means, que asumen una métrica euclidiana y un comportamiento sincrónico entre series, DTW permite alinear dinámicamente dos secuencias para minimizar su distancia incluso cuando existen diferencias en la fase temporal. Esta propiedad nos ayudó a analizar señales multivariadas como volumen, presión y temperatura, donde las fluctuaciones pueden estar desfasadas, pero aun así responder a un mismo patrón operativo.

La matriz de distancias entre clientes fue calculada a partir de una DTW multivariada ponderada, asignando mayor importancia al volumen (80%) y menor peso a presión y temperatura (10% cada una), conforme a la criticidad operacional de cada variable. Las series fueron previamente normalizadas mediante MinMaxScaler para cada cliente y variable, garantizando comparabilidad sin alterar la forma de las señales. Adicionalmente, se aplicó un muestreo proporcional (downsampling) para reducir la carga computacional manteniendo la integridad de los patrones.

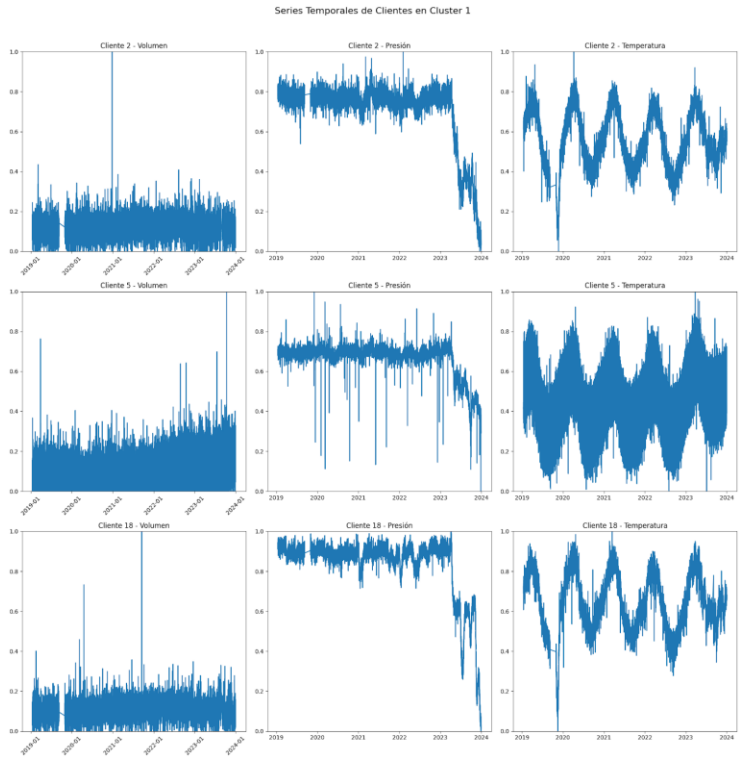
A partir de esta matriz, se aplicó un modelo de Agrupamiento Jerárquico con enlace Ward, que permite construir una jerarquía de similitudes y visualizarla mediante un dendrograma. Este enfoque facilitó la elección del número de clústeres mediante inspección visual del umbral de corte en la estructura del árbol jerárquico.



El proceso permitió identificar siete grupos distintos de clientes industriales, cada uno representando un conjunto de consumidores con comportamientos similares en cuanto a consumo de gas, variaciones estacionales, y estabilidad operacional. La reducción de dimensionalidad mediante PCA sobre la matriz de distancias permitió una visualización clara de la separación entre clústeres, confirmando la efectividad del enfoque DTW jerárquico.

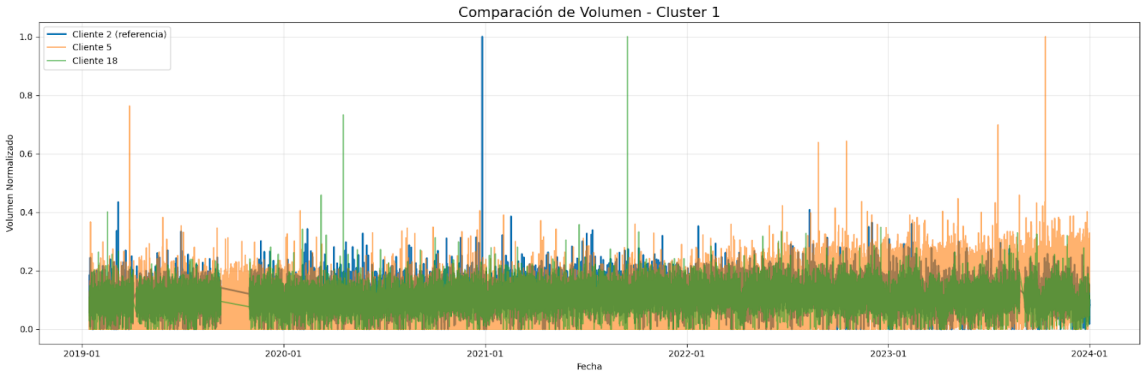


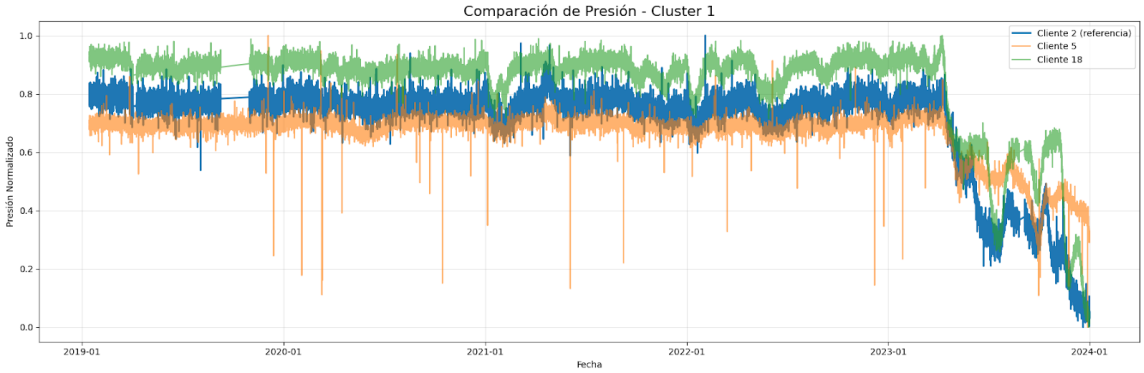
Una vez aplicado el algoritmo, se procedió a examinar la validez cualitativa de los grupos generados. Para ello, se seleccionó el Clúster 1, compuesto por los clientes 2, 5 y 18, y se analizaron sus series temporales individuales y comparativas para las tres variables principales: volumen, presión y temperatura.

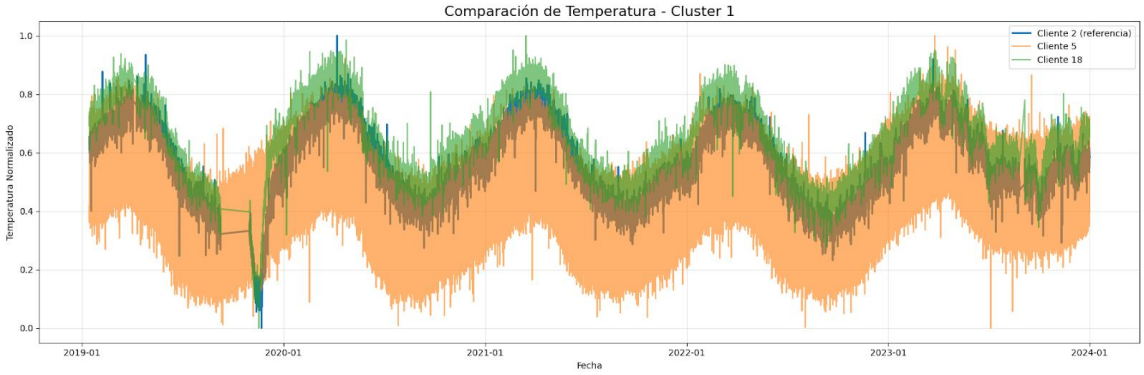


Las gráficas individuales muestran que, a pesar de la variabilidad inherente a cada cliente, existen patrones estacionales y niveles operativos relativamente similares. En particular, las series de temperatura presentan ciclos anuales sincronizados, consistentes con la influencia ambiental esperada. Las curvas de presión exhiben un comportamiento decreciente y homogéneo hacia los últimos años, mientras que el volumen mantiene una banda de variación similar entre los tres clientes, con presencia ocasional de picos que pueden corresponder a eventos operativos o anomalías.

Las comparaciones superpuestas entre clientes (cliente 2 como referencia) refuerzan visualmente la coherencia interna del clúster. Se observan trayectorias similares en la forma general de las series y en los niveles relativos de las variables, lo cual valida la eficacia del enfoque multivariado con ponderación diferenciada por variable (0.8 para volumen y 0.1 para las otras dos). Este análisis evidencia que el agrupamiento no fue arbitrario, sino que refleja afinidades estructurales reales en los perfiles operativos de los clientes.







En consecuencia, el Clúster 1 puede considerarse representativo de un patrón operativo común, lo que justifica el uso posterior de modelos de detección de anomalías entrenados sobre la agregación de sus series, como estrategia eficiente y escalable para aplicaciones industriales.

Detección de Outliers Multivariados

Para establecer una referencia de anomalías que permita evaluar el desempeño de los modelos de detección, se aplicó un enfoque estadístico multivariado basado en la distancia de Mahalanobis. Esta

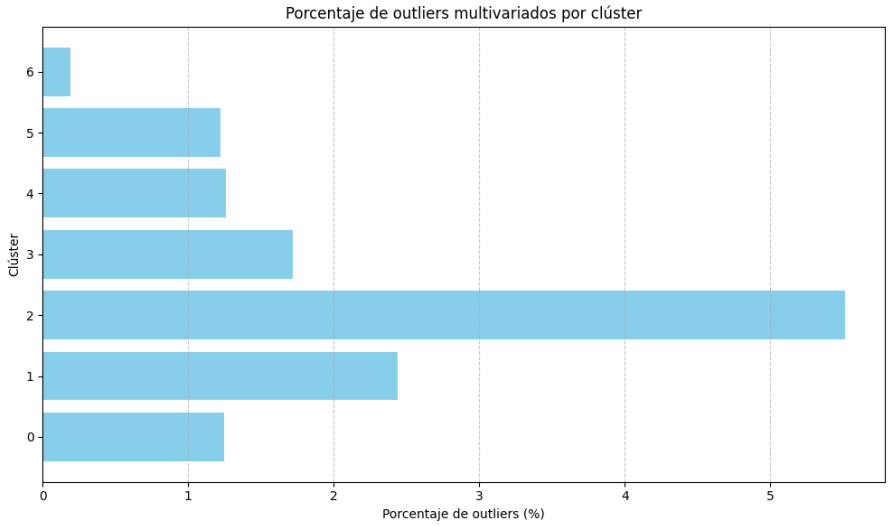
métrica permite identificar observaciones atípicas considerando simultáneamente el comportamiento conjunto de tres variables críticas: volumen, presión y temperatura.

A partir de los datos previamente agrupados en clústeres mediante análisis jerárquico multivariado con DTW, se calcularon los valores atípicos multivariados de forma independiente para cada clúster, con el objetivo de preservar las características estadísticas propias de cada grupo.

Posteriormente, se elaboró un resumen por clúster que incluye el total de observaciones, la cantidad de valores atípicos detectados y el porcentaje relativo de estas anomalías. Este análisis evidenció que la proporción de puntos atípicos varía significativamente entre los diferentes clústeres.

Resumen de outliers multivariados por clúster:

	Cluster	total_registros	total_outliers	porcentaje_outliers
0	0	130536	1636	1.25
1	1	130536	3183	2.44
2	2	261072	14396	5.51
3	3	87024	1500	1.72
4	4	43512	550	1.26
5	5	130536	1592	1.22
6	6	87024	167	0.19



Preselección y Justificación de Modelos

A partir del análisis del contexto técnico de Contugas —detección de anomalías multivariadas en tiempo real, operación en AWS y requisitos de escalabilidad— se plantearon varios modelos candidatos, los cuales fueron evaluados en el clúster número uno.

- ## Local Outlier Factor (LOF)

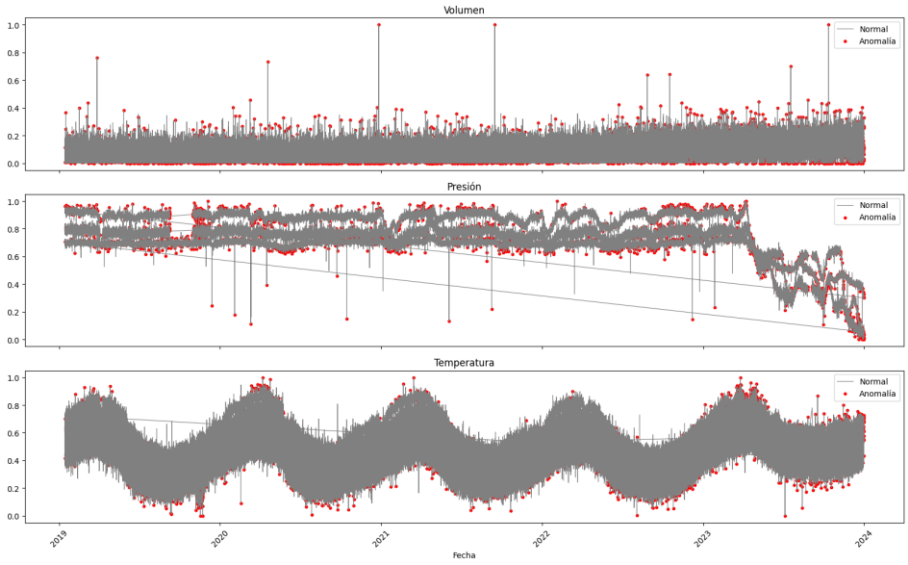
LOF calcula la densidad relativa de cada punto respecto a sus vecinos, lo que permite identificar anomalías locales.

Ventajas:

- Eficaz para detectar desviaciones sutiles en contextos densos.
- Ofrece interpretabilidad a nivel local.

Desventajas:

- Alto costo computacional en escenarios de *streaming*.
- Escalabilidad limitada en entornos en tiempo real.
- Baja adecuación para despliegues en la nube o arquitecturas *serverless*.



Autoencoders

Los autoencoders son redes neuronales que aprenden una representación comprimida de los datos para luego reconstruirla. Las discrepancias entre la entrada y la salida permiten identificar anomalías.

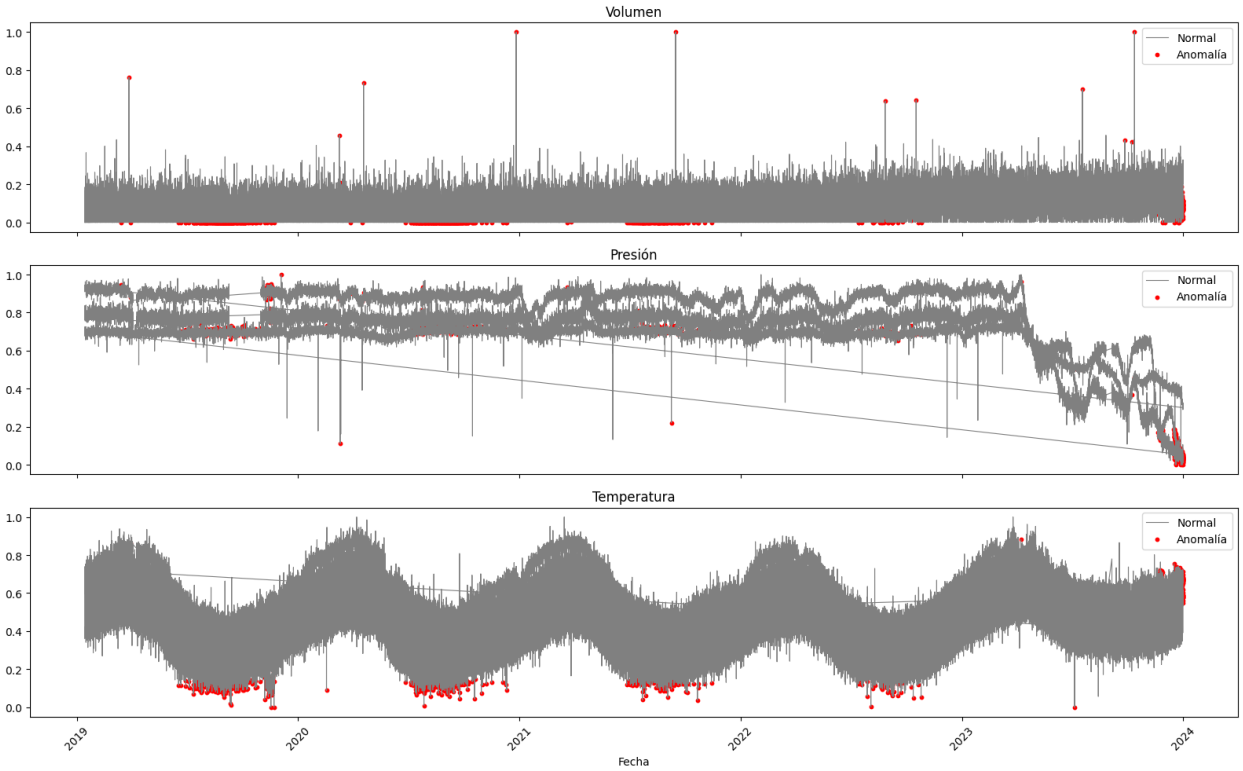
Ventajas:

- Capturan relaciones no lineales complejas entre variables.
- Pueden extenderse con arquitecturas LSTM o mecanismos de atención para modelar dependencias temporales.
- Posibilitan una adaptación segmentada mediante ajustes en la arquitectura o entrenamiento especializado.

Desventajas:

- Alta complejidad en la configuración, entrenamiento y mantenimiento.
- La inferencia en tiempo real requiere contenerización y orquestación (Docker, SageMaker, EKS).

- La definición de umbrales de error puede introducir subjetividad operativa.



Isolation Forest

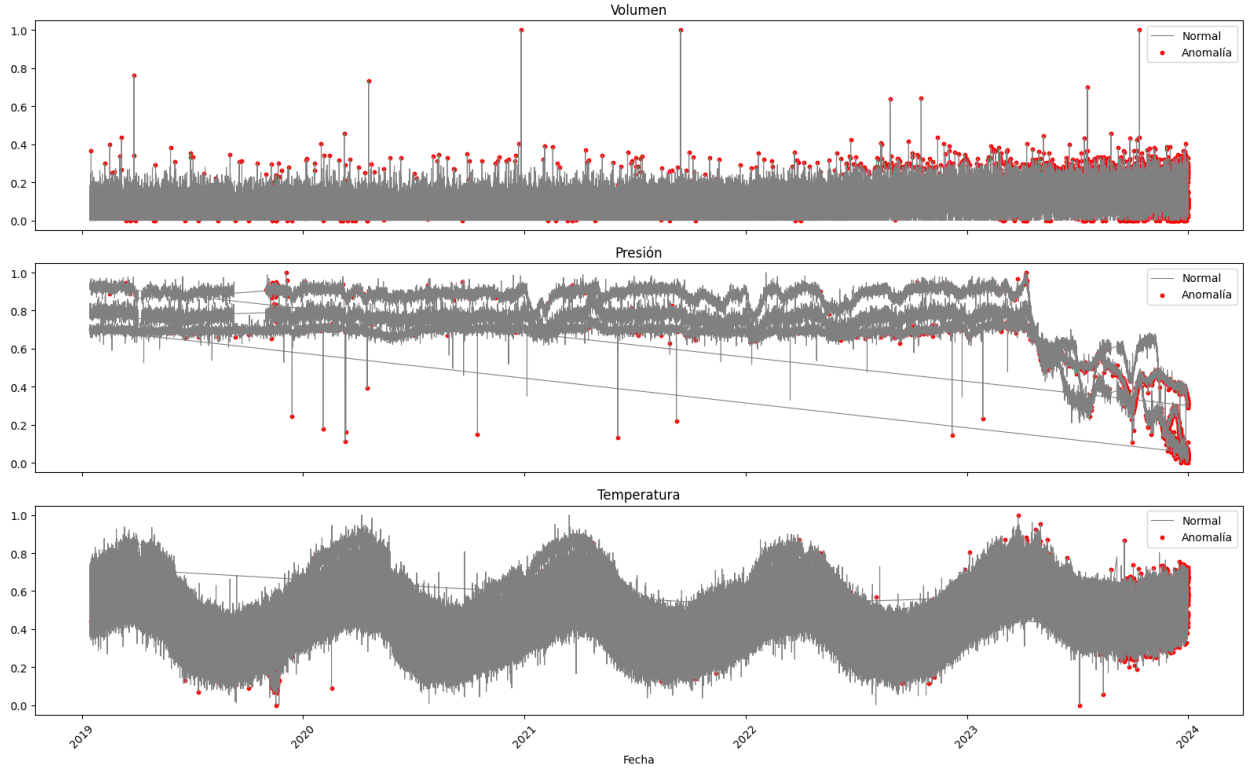
Isolation Forest detecta anomalías separando iterativamente las observaciones mediante cortes aleatorios. Es eficiente, aunque no está diseñado específicamente para series temporales.

Ventajas:

- Alta eficiencia computacional.
- Entrenamiento simple y buena escalabilidad.
- Requiere poca parametrización inicial.

Desventajas:

- No considera relaciones temporales ni secuenciales.
- Menor precisión para detectar anomalías dependientes del tiempo.
- Poco efectivo en contextos multivariados complejos, como el de Contugas.



Random Cut Forest (RCF)

RCF es un algoritmo basado en árboles aleatorios, diseñado específicamente para detectar anomalías en series temporales y espacios multidimensionales.

Ventajas técnicas:

- **Eficiencia en series temporales:** captura fluctuaciones sutiles a lo largo del tiempo, ideal para sensores con frecuencia horaria.
- **Robustez en alta dimensionalidad:** analiza múltiples variables simultáneamente sin comprometer el rendimiento.
- **No requiere etiquetas:** evita la necesidad de anotación manual de datos históricos.
- **Adaptabilidad a patrones cambiantes:** responde eficazmente ante desviaciones contextuales o no estacionarias.
- **Integración nativa con AWS:** permite una operación eficiente y segura mediante servicios como SageMaker, Kinesis y Lambda.

Desventajas:

- Complejidad computacional moderada, dependiente de una adecuada configuración de hiperparámetros (número de árboles, profundidad, tamaño de ventana).

- Requiere calibración precisa, especialmente en entornos con múltiples segmentos operativos.

Este modelo resulta altamente pertinente para el caso de Contugas, donde las condiciones de consumo varían según el cliente, el horario y el clima. Su capacidad para incorporar el tiempo como dimensión de análisis mejora significativamente la detección de patrones anómalos dependientes del contexto.

Comparación Final y Selección del Modelo Principal

```
✅ Total outliers Outliers multivariados: 3183
LOF acertó en: 568
Autoencoder acertó en: 544
Isolation Forest acertó en: 2247

🔍 Comparación completa:
LOF - Total: 3186, Aciertos: 568, Falsos positivos: 2618, Precisión: 0.18
Autoencoder - Total: 1306, Aciertos: 544, Falsos positivos: 762, Precisión: 0.42
Isolation Forest - Total: 3186, Aciertos: 2247, Falsos positivos: 939, Precisión: 0.71
```

Durante la fase de evaluación comparativa, el algoritmo **Isolation Forest** alcanzó una precisión del **71 %** en la detección de *outliers* multivariados en el Clúster 1, superando a **LOF** (18 %) y al **Autoencoder** (42 %). No obstante, la elección final del modelo a desplegar en la solución analítica industrial recayó en **Random Cut Forest (RCF)**, con base en una evaluación integral que consideró no solo las métricas de desempeño, sino también la robustez algorítmica, la escalabilidad operativa y la alineación con la infraestructura tecnológica proyectada para **Contugas**.

RCF representa una evolución conceptual y práctica de Isolation Forest, incorporando mejoras fundamentales. Ambos modelos comparten el principio de identificar anomalías mediante la construcción de árboles que aíslan puntos de datos en espacios de alta dimensión. Sin embargo, RCF introduce la técnica de **cortes aleatorios** (*random cuts*) aplicados a árboles con estructuras más eficientes y balanceadas, lo que le permite detectar *outliers* con mayor precisión en conjuntos de datos complejos y multivariados. A diferencia de Isolation Forest —que opera con árboles binarios estáticos— RCF admite actualizaciones incrementales sin necesidad de reconstruir el modelo por completo, una capacidad crucial en entornos industriales de alto volumen como el de Contugas.

Esta ventaja cobra mayor relevancia al considerar que la solución será implementada sobre la infraestructura de **Amazon Web Services (AWS)**, plataforma en la que RCF ha sido optimizado de forma nativa como parte de sus servicios avanzados de *machine learning* y análisis de series temporales. Su integración directa con servicios como **Amazon SageMaker**, **Amazon Kinesis** y **AWS Lambda** facilita la implementación en entornos productivos, permitiendo el procesamiento en tiempo real de flujos de datos industriales sin requerir reentrenamientos costosos. Esta capacidad de *online learning* y actualización dinámica del modelo posiciona a RCF como una herramienta ideal para escenarios donde la continuidad operativa y la detección temprana de eventos son prioritarias.

Adicionalmente, RCF permite descomponer el *score* de anomalía y atribuirlo a variables específicas del vector de entrada. Esta funcionalidad, ausente en Isolation Forest, representa un valor agregado significativo para la gestión operativa, ya que permite a analistas y operadores identificar rápidamente si un comportamiento anómalo está asociado con la presión, el volumen, la temperatura o una combinación de estas variables.

En conclusión, si bien Isolation Forest constituye una herramienta valiosa y demostró un desempeño destacable, la selección de **Random Cut Forest** como modelo final se fundamenta en su diseño nativo para ambientes en la nube, su eficiencia en el manejo de grandes volúmenes de datos y su mayor capacidad de adaptación al cambio. Estas características lo convierten en la alternativa más robusta, escalable y alineada con las necesidades de Contugas para la detección de anomalías en sus procesos de distribución de gas industrial.

Tipología del Modelo y su Integración

El modelo principal propuesto es *Random Cut Forest* (RCF), un algoritmo de aprendizaje no supervisado desarrollado por Amazon, disponible como servicio administrado en Amazon SageMaker. RCF está diseñado para detectar anomalías en datos multivariados sin necesidad de etiquetas, lo cual resulta especialmente relevante, dado que Contugas no dispone de un historial clasificado de eventos anómalos.

- **Entradas:** presión, temperatura y volumen por cliente, segmentados en siete perfiles operativos.
- **Salidas:** un *anomaly score* en para cada observación, el cual se compara con umbrales calculados a partir de datos históricos.

Este enfoque permite identificar desviaciones significativas en el comportamiento de los sensores, ajustadas al perfil operativo de cada cliente o grupo de clientes.

A nivel operativo, las alertas generadas por los modelos predictivos se integran en un *dashboard* interactivo, desde el cual los operadores pueden:

- Confirmar o descartar eventos anómalos.
- Visualizar la evolución histórica de las variables.
- Priorizar incidentes según el segmento afectado.

Esta validación humana actúa como un componente prescriptivo dentro del proceso de toma de decisiones, al tiempo que permite construir una base de retroalimentación continua para mejorar el desempeño del sistema.

Verificación de Supuestos y Tratamiento Previo de Datos para el Modelo RCF

Random Cut Forest (RCF) es un algoritmo de detección de anomalías basado en árboles aleatorios y cortes espaciales. Su diseño lo hace especialmente robusto frente a supuestos estrictos sobre los datos, como la linealidad, la normalidad o la necesidad de etiquetas supervisadas. No obstante, para garantizar

un rendimiento óptimo y evitar sesgos en la inferencia, se han verificado y tratado aspectos clave relacionados con los tipos de datos y su preparación.

Tipo de datos de entrada

RCF opera sobre vectores numéricos de dimensión fija, por lo que requiere que todas las variables de entrada sean continuas o discretas transformadas a formato numérico. En este caso:

- **Presión, temperatura y volumen** son variables continuas, por lo que son directamente adecuadas para el modelo.
- **No se utilizan variables categóricas ni booleanas** en esta etapa. En caso de incorporarlas en versiones futuras, deberán codificarse mediante técnicas como *target encoding* o *embeddings*, para preservar relaciones numéricas útiles.

Estacionariedad y patrones temporales

RCF no requiere que los datos sean estacionarios, ya que evalúa la inserción de un punto dentro de un bosque sin basarse en suposiciones sobre la distribución o tendencia de los datos. Sin embargo:

- Se **segmentaron los datos** para mejorar la homogeneidad del comportamiento modelado.
- Se incluyó el **timestamp como índice** para asegurar el orden secuencial y preservar las relaciones temporales durante el entrenamiento.

Correlación entre variables

Aunque RCF no asume relaciones lineales entre variables, se verificó la presencia de **correlaciones cruzadas** (por ejemplo, entre presión y volumen), lo que permitió validar la pertinencia del enfoque multivariado. Se observaron patrones consistentes entre sensores, lo que refuerza la necesidad de un modelo que considere simultáneamente todas las dimensiones.

Tratamiento de valores faltantes y duplicados

Se generó un **índice temporal completo** que cubre todo el período de interés (del 14 de enero de 2019 al 31 de diciembre de 2023), con una frecuencia horaria. Esto garantiza que todos los clientes cuenten con conjuntos de datos con la misma granularidad temporal, lo cual es esencial para el funcionamiento de RCF, ya que el modelo depende de secuencias temporales coherentes.

Posteriormente, se interpolaron los **valores faltantes** en las variables clave (presión, temperatura y volumen) utilizando interpolación lineal. Esta operación se realizó de forma individual para cada cliente, asegurando series temporales completas y minimizando la pérdida de patrones relevantes.

En algunos casos, los datos contenían **duplicados**, debido a múltiples registros de un mismo cliente en el mismo momento. Para evitar que estos duplicados afecten el análisis, se agregaron utilizando el promedio

de las variables clave, asegurando así que cada cliente tenga un único registro por *timestamp* y se eviten posibles sesgos en los modelos derivados de la repetición de datos.

Estandarización y escalado

Dado que RCF funciona mejor cuando los datos tienen **magnitudes homogéneas**, se aplicó un escalado estándar a las variables relevantes (presión, temperatura y volumen). Este escalado se realizó de forma **independiente por segmento**, ya que cada uno presenta distintos rangos operativos para las variables. El proceso asegura que todas las variables tengan media cero y desviación estándar uno, lo que facilita la detección de anomalías sin que una variable domine al modelo debido a su escala.

Técnica de “Shingling” para capturar patrones temporales

Con el objetivo de mejorar la detección de anomalías en datos con comportamiento periódico, se incorporó la técnica de **shingling** o segmentación en ventanas deslizantes. Esta técnica consiste en agrupar secuencias consecutivas de puntos de datos en un solo vector de características, transformando así una serie temporal univariada o multivariada en una serie de vectores multivariados de mayor dimensión.

En este caso particular, los datos presentan ciclos operativos diarios, con patrones recurrentes en presión, temperatura y volumen a lo largo del día. Para capturar adecuadamente estas dinámicas temporales, se optó por una **ventana de 24 puntos**, correspondiente a un día completo en datos con frecuencia horaria. Esta elección permite representar de forma compacta la evolución de las variables a lo largo del día, facilitando que el modelo RCF detecte desviaciones en la forma o estructura del comportamiento diario, más allá de simples anomalías puntuales.

La transformación mediante shingling se realizó de la siguiente manera:

- Para cada serie temporal (ya interpolada, deduplicada y estandarizada), se generaron vectores de características concatenando los valores consecutivos de presión, temperatura y volumen durante 24 horas.
- Esto convierte cada observación en un vector de tamaño 24 x n, donde n es el número de variables (en este caso, 3). Los resultados son vectores de 72 dimensiones que representan la "firma temporal" diaria de cada cliente.
- Estos vectores se utilizaron como entrada para el modelo RCF, que ahora puede detectar anomalías a nivel de comportamiento diario completo, en lugar de observar solo instancias individuales.

Este enfoque mejora significativamente la sensibilidad del modelo ante **anomalías contextuales**, como cambios graduales en los patrones diarios o desviaciones persistentes que no serían detectadas si se analizara cada punto en forma aislada.

Reporte de Procesos de Entrenamiento y Validación del Modelo RCF

En el contexto de la implementación del modelo **Random Cut Forest (RCF)** para la detección de anomalías en los datos de consumo de gas de Contugas, se estableció un proceso detallado de **entrenamiento, validación y ajuste de parámetros**. Este procedimiento garantiza que el modelo se adapte adecuadamente a los datos y que sus resultados sean significativos y útiles para la toma de decisiones en tiempo real.

Cabe destacar que dicho proceso se aplica a **cada uno de los segmentos identificados**; sin embargo, con el fin de simplificar el presente informe, se presentan únicamente los resultados correspondientes al **segmento uno** y su modelo asociado.

1. Selección de variables

Para el entrenamiento del modelo se seleccionaron como variables clave **la presión, la temperatura y el volumen** registrados por los sensores, ya que estas métricas representan de forma más precisa el comportamiento del consumo de gas. La elección de estas variables se fundamentó tanto en su **relevancia operativa para la detección de desviaciones** en los sistemas de monitoreo, como en su disponibilidad dentro del conjunto de datos.

Esta selección fue esencial para capturar la **variabilidad inherente a los datos** y para identificar **patrones atípicos o anómalos** que puedan reflejar fallos o comportamientos inusuales en el sistema de distribución de gas. Además, se aplicaron técnicas de **interpolación para el tratamiento de datos faltantes** y **escalado por segmento**, lo cual permitió homogeneizar y hacer comparables las variables. Este paso es crucial para optimizar el rendimiento del modelo RCF.

2. Conjuntos de Entrenamiento y Prueba

El modelo Random Cut Forest (RCF) es un algoritmo de detección de anomalías no supervisado, lo que significa que no requiere un conjunto de datos etiquetado ni una división formal entre las fases de entrenamiento y prueba. En lugar de eso, se utilizó un único conjunto continuo de datos históricos, que incluye registros completos de consumo, presión y temperatura desde el año 2019 hasta finales de 2023.

Esta estrategia se justifica por las siguientes características del modelo:

- **Entrenamiento incremental**: El RCF entrena de manera incremental a medida que procesa los datos, construyendo árboles de corte aleatorios. Por lo tanto, no es necesario definir una separación explícita entre las fases de entrenamiento y evaluación.
- **Ausencia de etiquetas**: Dado que no se disponen de etiquetas para distinguir entre eventos normales y anómalos, no es posible calcular métricas de validación supervisadas tradicionales, como precisión o recall.

- **Evaluación basada en estadísticas**: La evaluación del modelo se realiza a través del análisis estadístico de los scores de anomalía generados y su comportamiento frente a desviaciones esperadas, utilizando umbrales como la regla empírica de tres sigmas.

Al mantener un solo conjunto de datos continuo y bien preprocesado, se garantiza que el modelo capture patrones estacionales, dependencias horarias y diferencias entre segmentos de clientes de manera más representativa. Además, se respeta la secuencia temporal de los eventos, evitando cualquier fuga de información.

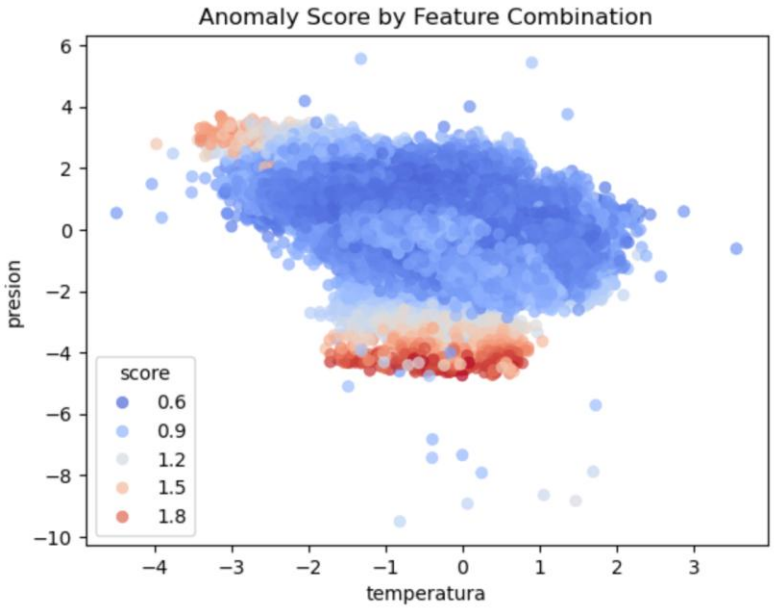
3. Parametrización

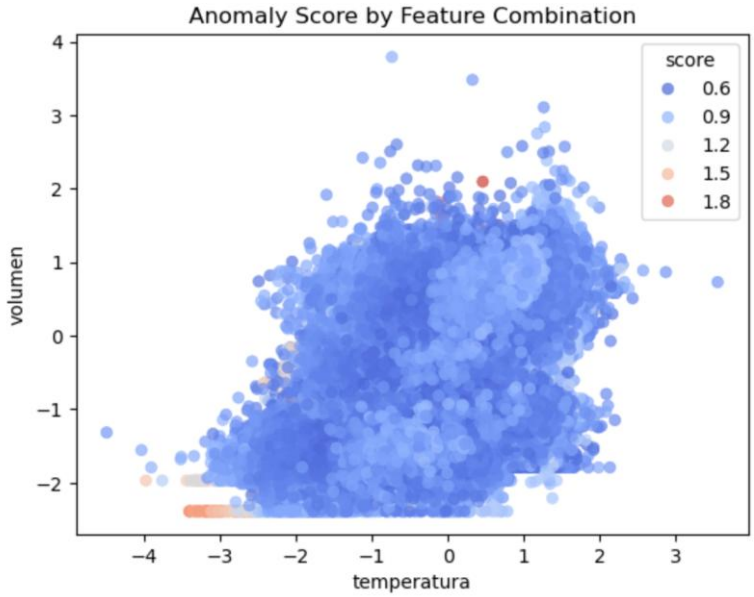
El proceso de entrenamiento del modelo RCF (Random Cut Forest) implicó la selección de varios parámetros clave para optimizar su rendimiento y adaptarlo de manera más eficaz a las condiciones específicas de los datos. Los principales parámetros ajustados fueron los siguientes:

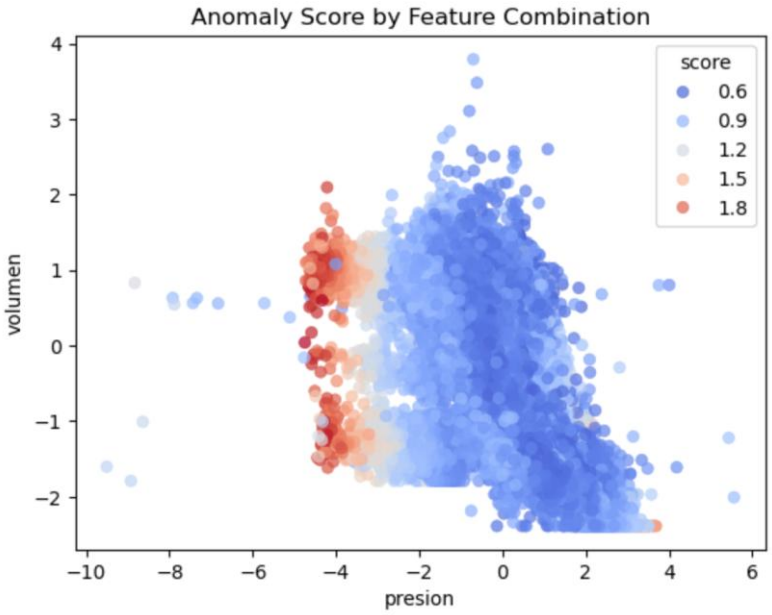
- **Número de árboles (num_trees):** Este parámetro determina cuántos árboles componen el bosque aleatorio. En el caso del RCF, un mayor número de árboles tiende a generar estimaciones más robustas y estables de las puntuaciones de anomalía, ya que mejora la cobertura estadística del espacio de características. Sin embargo, esto también incrementa el tiempo de entrenamiento e inferencia. Para el caso de Contugas, se seleccionó un valor intermedio de 200 árboles, superior al valor predeterminado de 100. Esta elección buscaba mejorar la precisión del modelo sin generar una sobrecarga computacional significativa. Dicha decisión respondió a la necesidad de identificar eventos anómalos con mayor sensibilidad, en un entorno donde los datos pueden presentar patrones estacionales o contextuales sutiles.
- **Tamaño del vector de anomalía (num_samples_per_tree):** Este parámetro define el número de muestras aleatorias del conjunto de datos utilizadas para construir cada árbol. Un valor bajo implica que cada árbol se entrena con menos datos, lo que puede aumentar la sensibilidad a las anomalías locales. Por otro lado, un valor más alto genera árboles más generalistas y estables. En el caso de Contugas, se optó por un valor cercano al obtenido a través de la detección de outliers multivariados por segmento, partiendo de la hipótesis de que las anomalías son eventos raros pero críticos. Esta configuración favorece la detección de outliers sin introducir una variabilidad excesiva en las puntuaciones de anomalía.

4. Definición de Métricas de Evaluación

En un modelo de detección de anomalías no supervisado, no se cuenta con etiquetas reales de anomalías que permitan evaluar el modelo de forma convencional. En este contexto, el modelo RCF (Random Cut Forest) asigna una puntuación de anomalía a cada punto de datos, la cual indica qué tan atípico es un punto en comparación con el comportamiento general del conjunto. Es posible utilizar umbrales sobre estas puntuaciones para determinar qué puntos se consideran anómalos. Para evaluar el rendimiento del modelo, se analiza la distribución de las puntuaciones y se establecen umbrales que optimicen la detección de patrones inusuales.


Anomaly Score by Feature Combination
presion
temperatura
score
0.6
0.9
1.2
1.5
1.8


Anomaly Score by Feature Combination
volumen
temperatura
score
0.6
0.9
1.2
1.5
1.8



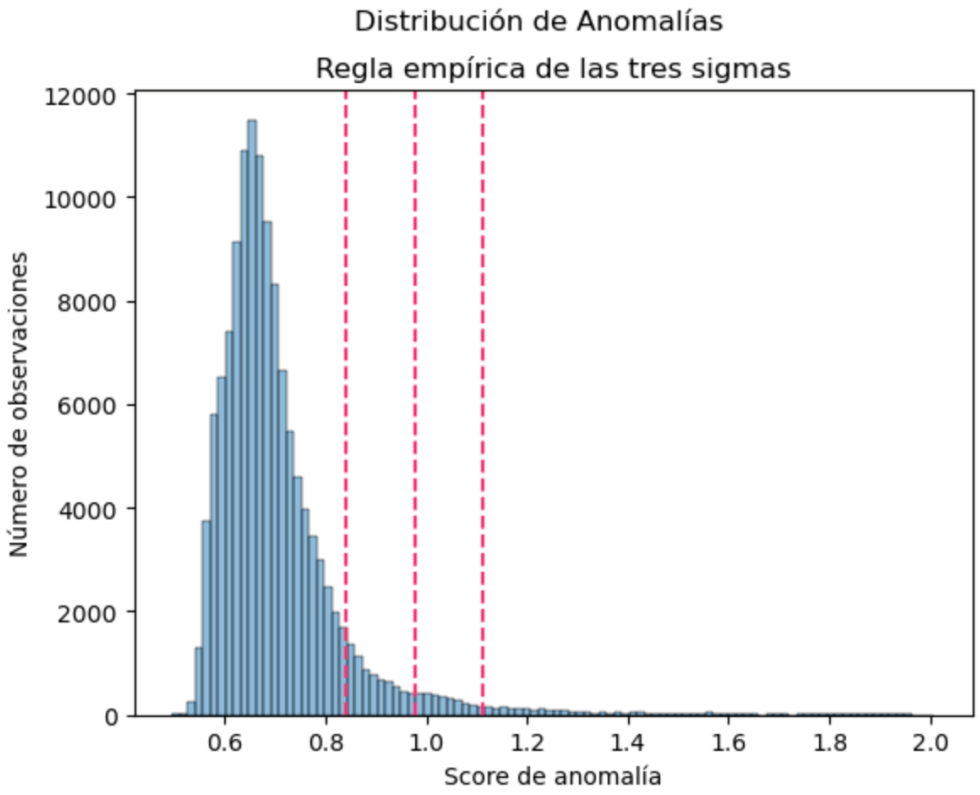
Se presentaron tres gráficas que muestran la puntuación de anomalía (anomaly score) para distintas combinaciones de características. A partir del análisis visual, se puede observar que ciertas combinaciones presentan puntuaciones consistentemente más altas, lo que sugiere una mayor sensibilidad a comportamientos atípicos en esas dimensiones. Estas combinaciones podrían estar capturando interacciones relevantes que no son evidentes cuando se analiza cada característica de forma individual. Por el contrario, otras combinaciones muestran puntuaciones más homogéneas o bajas, lo que indica una menor capacidad para distinguir anomalías. En conjunto, estas gráficas permiten identificar qué combinaciones de variables son más efectivas para la detección de patrones inusuales, lo cual es clave para ajustar umbrales y mejorar la precisión del modelo.

5. Calibración de Umbrales de Anomalía

Dado que no se cuenta con etiquetas que indiquen anomalías reales, se aplicó un criterio estadístico objetivo para determinar cuándo un puntaje de anomalía representa un evento significativo. Para ello, se utilizó la regla empírica de las tres sigmas:

RCF calcula una puntuación de anomalía para cada punto, y se considera anómalo aquel que excede en más de tres desviaciones estándar la media de los puntajes históricos. Esta heurística, conocida como *three-sigma limit*, es una práctica común para identificar valores estadísticamente atípicos, bajo la suposición de una distribución aproximadamente normal.

Este enfoque permite establecer umbrales adaptativos, específicos para cada cliente y contexto, sin necesidad de etiquetas. Su principal ventaja radica en su simplicidad, interpretabilidad y eficacia en entornos no supervisados, donde no se dispone de una "verdad" conocida sobre las anomalías.



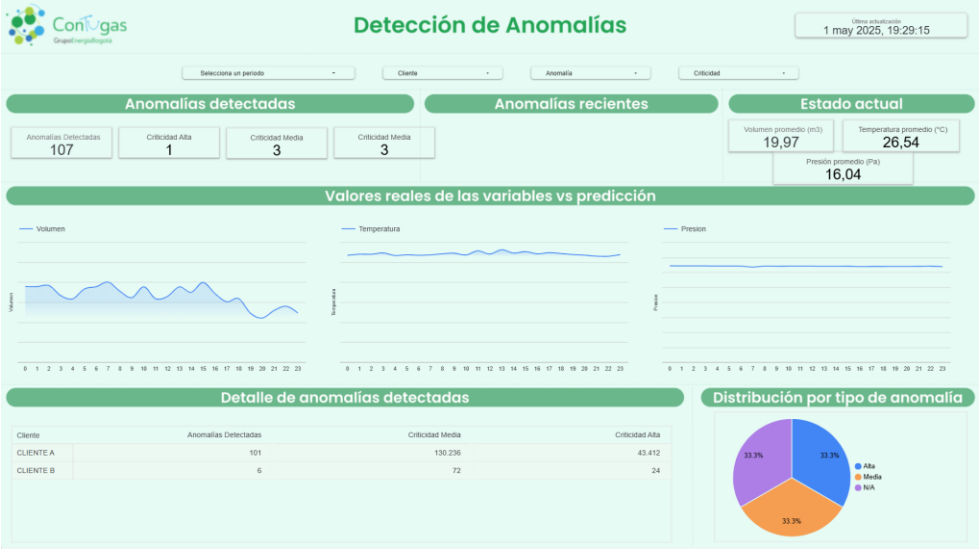
Completitud de la Solución y Próximos Pasos

Con base en los avances alcanzados hasta la fecha —que incluyen el análisis exploratorio, la limpieza avanzada de datos, la generación de variables temporales, la estandarización, la identificación de patrones mediante clustering y la evaluación y selección del modelo analítico— se detallan a continuación los elementos pendientes de desarrollo, junto con un plan de acción para completar el prototipo funcional de detección de anomalías en clientes industriales de Contugas:

Integración del modelo final (RCF): en un flujo automatizado que reciba datos en tiempo casi real, ejecute la predicción de anomalías, clasifique su criticidad y almacene los resultados en Redshift para su posterior visualización. Este pipeline será desarrollado en AWS, utilizando servicios como SageMaker y Lambda, conforme a lo establecido en el diagrama del backend.

Finalización de la interfaz de usuario: que incluirá filtros por fecha, cliente y nivel de criticidad de las anomalías, además de visualizaciones históricas y métricas clave del modelo (por ejemplo, cantidad de anomalías detectadas). Este tablero estará conectado directamente a Redshift y deberá alinearse con los

flujos de trabajo del equipo operativo de Contugas. A continuación, se muestra el estado actual:



Aunque esta sección no estaba contemplada en el mock-up inicial, se decidió incluir un módulo adicional centrado en el análisis descriptivo de los datos, con el fin de proporcionar una herramienta útil para cualquier análisis que realice Contugas. Asimismo, la validación de alertas por parte de los operadores deberá convertirse en un componente persistente del flujo de datos, permitiendo retroalimentar al sistema y, eventualmente, entrenar modelos supervisados o semisupervisados.

Diligenciamiento de la rúbrica de validación: en la cual se especificará:
- Cómo se satisfacen los requerimientos definidos.
- Qué ajustes se realizaron durante la implementación.
- Qué requerimientos (si los hubiera) no se lograron cumplir, y cuál es la propuesta para abordarlos en futuras iteraciones.

Elaboración del manual de usuario y la documentación técnica: que deberá incluir:
- Una descripción del sistema: qué hace, cómo funciona, sus ventajas y limitaciones.
- Instrucciones para interactuar con el sistema (por ejemplo, cómo cargar datos, consultar anomalías, filtrar resultados).
- Nivel de conocimientos técnicos requerido por el usuario final (bajo, dado que el sistema está orientado a equipos operativos).
- Diagrama arquitectónico del backend y frontend.
- Reporte técnico del modelo y su proceso de validación.
- Enlace al repositorio con el código fuente, datos de ejemplo y scripts de despliegue.

Preparación de las diapositivas y guion para la presentación final (duración: 7 minutos): que deberá cubrir:

- El problema abordado y los beneficios esperados para Contugas.
- Una demostración funcional del prototipo.
- Las métricas clave del modelo y los indicadores de negocio que se verán impactados.
- Comparación con el estado actual y valor diferencial de la solución propuesta.
- Un resumen de los costos, riesgos y condiciones necesarias para su adopción.