Reporte Técnico de Experimientos: Selección y Parametrización de Modelos para la Detección de Anomalías en Consumo de Gas

1. Introducción

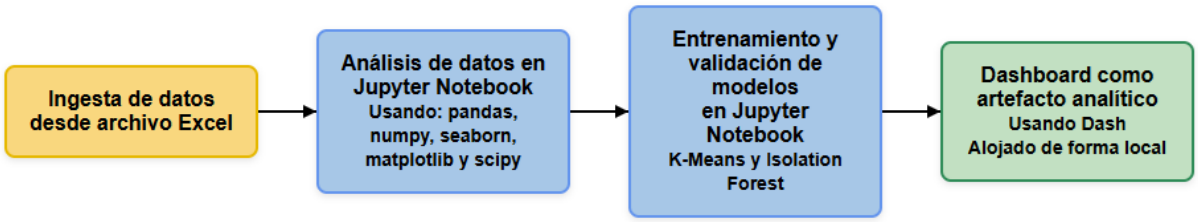
El presente documento detalla el proceso técnico desarrollado para la selección y parametrización de modelos destinados a la detección de anomalías en el consumo de gas para clientes industriales de Contugas. Este proyecto responde a la necesidad de identificar comportamientos atípicos en las variables operacionales de presión, temperatura y volumen, que podrían indicar fugas, fallas en medidores, o patrones de consumo irregulares.

2. Diagrama Esquemático del Proyecto

El proyecto se ha estructurado en dos fases de implementación bien diferenciadas: una para la presentación del artefacto analítico (prototipo funcional) y otra para la implementación final al cliente.

2.1 Implementación para Presentación del Artefacto Analítico

La implementación inicial para la presentación del prototipo sigue un flujo de trabajo secuencial que facilita el desarrollo y validación del concepto:

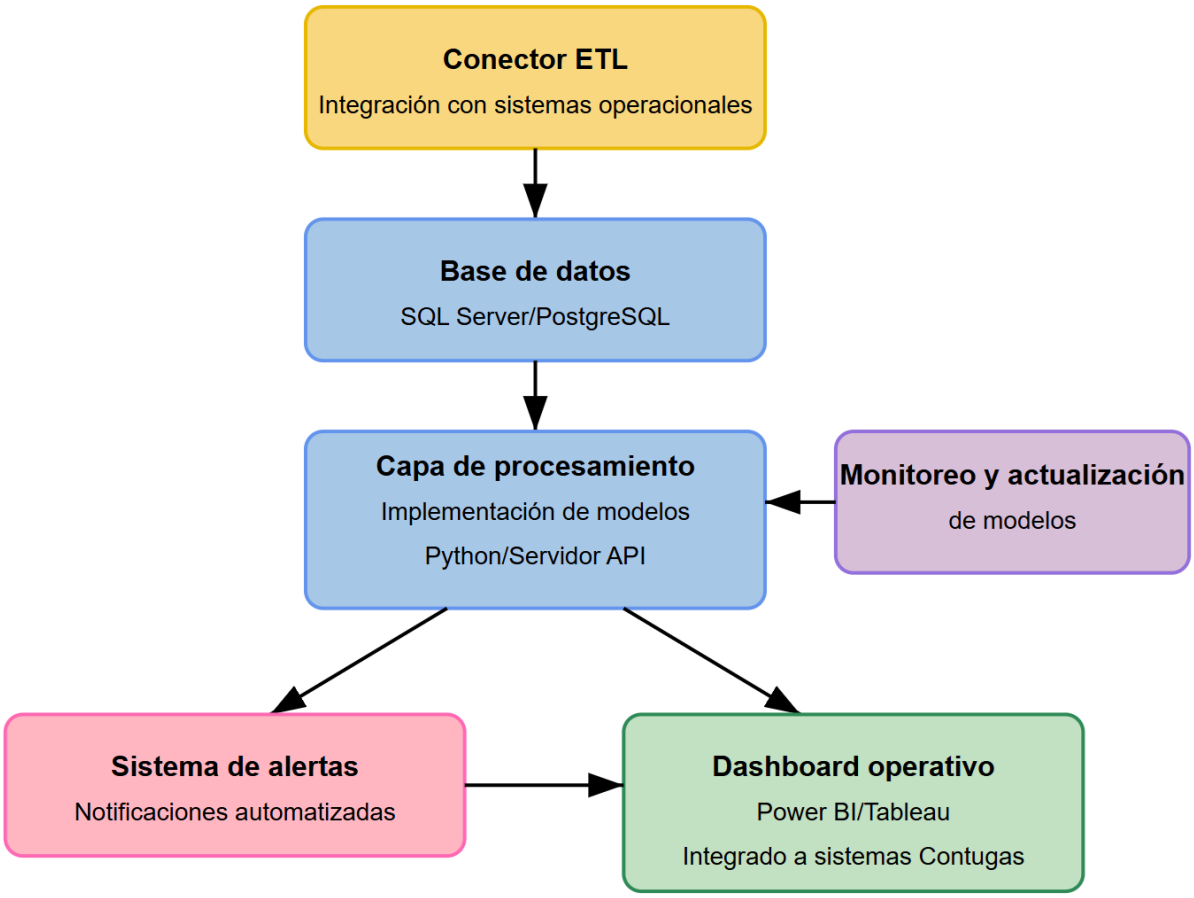


Este diagrama representa:

1. **Ingesta de datos**: Proceso de carga y preparación inicial de los datos desde archivos Excel proporcionados por Contugas.
2. **Análisis de datos**: Exploración, limpieza y transformación de datos mediante librería pandas, numpy y herramientas de visualización en entorno Jupyter.
3. **Entrenamiento y validación**: Implementación de algoritmos de segmentación (K-Means) y detección de anomalías (Isolation Forest).
4. **Dashboard analítico**: Presentación de resultados mediante una interfaz interactiva desarrollada con Dash, alojada localmente para demostración.

2.2 Implementación para Cliente Final

Para la implementación definitiva al cliente se plantea una arquitectura más robusta y orientada a producción:



Esta arquitectura propuesta contempla:

1. **Conector ETL**: Integración directa con los sistemas operacionales de Contugas para la captura continua de datos.
2. **Base de datos**: Almacenamiento estructurado y optimizado de los datos históricos y en tiempo real.
3. **Capa de procesamiento**: Implementación de los modelos entrenados en un entorno de producción mediante API.
4. **Dashboard operativo**: Interfaz de usuario integrada en los sistemas corporativos de Contugas.

5. **Sistema de alertas**: Generación y distribución de notificaciones basadas en la detección de anomalías.
6. **Monitoreo y actualización**: Supervisión continua del desempeño y actualización periódica de los modelos.

Esta estructura garantiza la escalabilidad, mantenibilidad y robustez necesarias para un sistema de detección de anomalías que operará en un entorno de producción crítico para el negocio.

3. Metodología

3.1 Exploración y Preparación de Datos

Se trabajó con datos históricos de 20 clientes industriales (CLIENTE1 a CLIENTE20), registrados en intervalos horarios. Cada registro contiene mediciones de:

- Presión (bar)
- Temperatura (°C)
- Volumen (m^3)

El conjunto de datos abarca un período de aproximadamente 1812 días (2019-2023), con un promedio de 41,000-43,000 registros por cliente para cada variable.

En la fase de preparación, se realizaron las siguientes tareas:

1. Identificación y manejo de registros duplicados, seleccionando el valor máximo para cada fecha duplicada
2. Conversión de los tipos de datos adecuados
3. Limpieza de valores nulos o inconsistentes

Análisis Exploratorio de Datos

Para documentar y validar el proceso de exploración y preparación de datos, se generaron las siguientes visualizaciones:

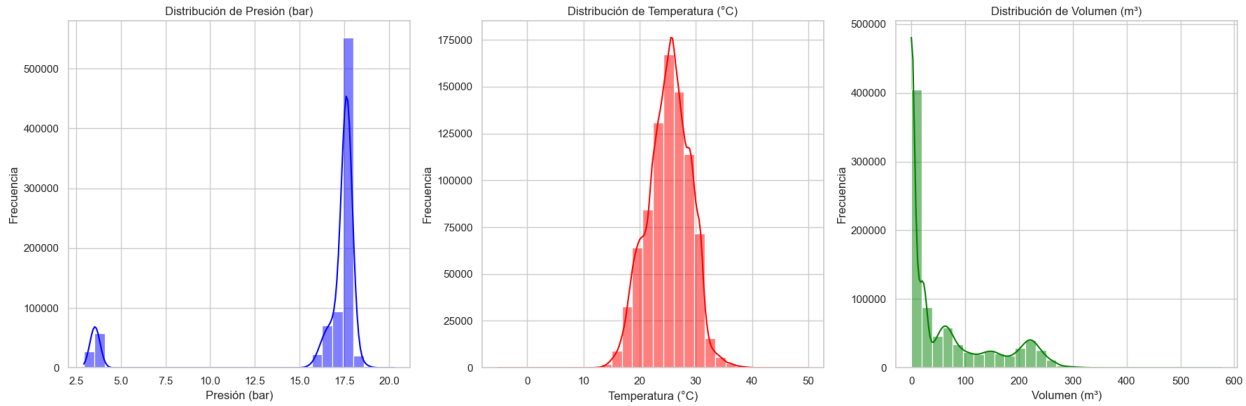


Figura 1. Distribuciones de Variables Principales

- Histogramas de frecuencia para Presión (bar), Temperatura (°C) y Volumen (m^3)
- Permite identificar la forma de las distribuciones y posibles sesgos en los datos
- Revela patrones de concentración de valores y rangos operacionales típicos

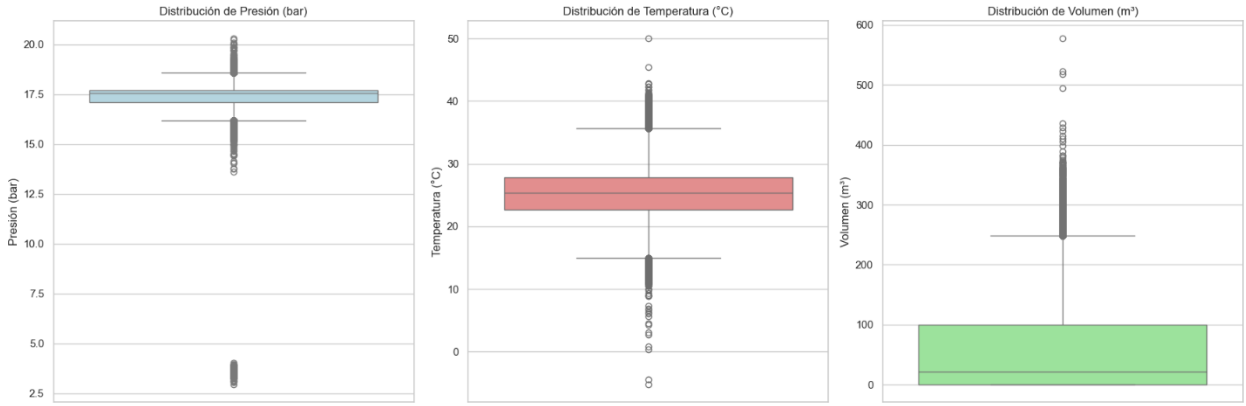


Figura 2. Análisis de Variabilidad mediante Box Plots

- Diagramas de caja para cada variable principal
- Visualización de quartiles, medianas y valores atípicos
- Identificación de la dispersión de datos y presencia de outliers extremos

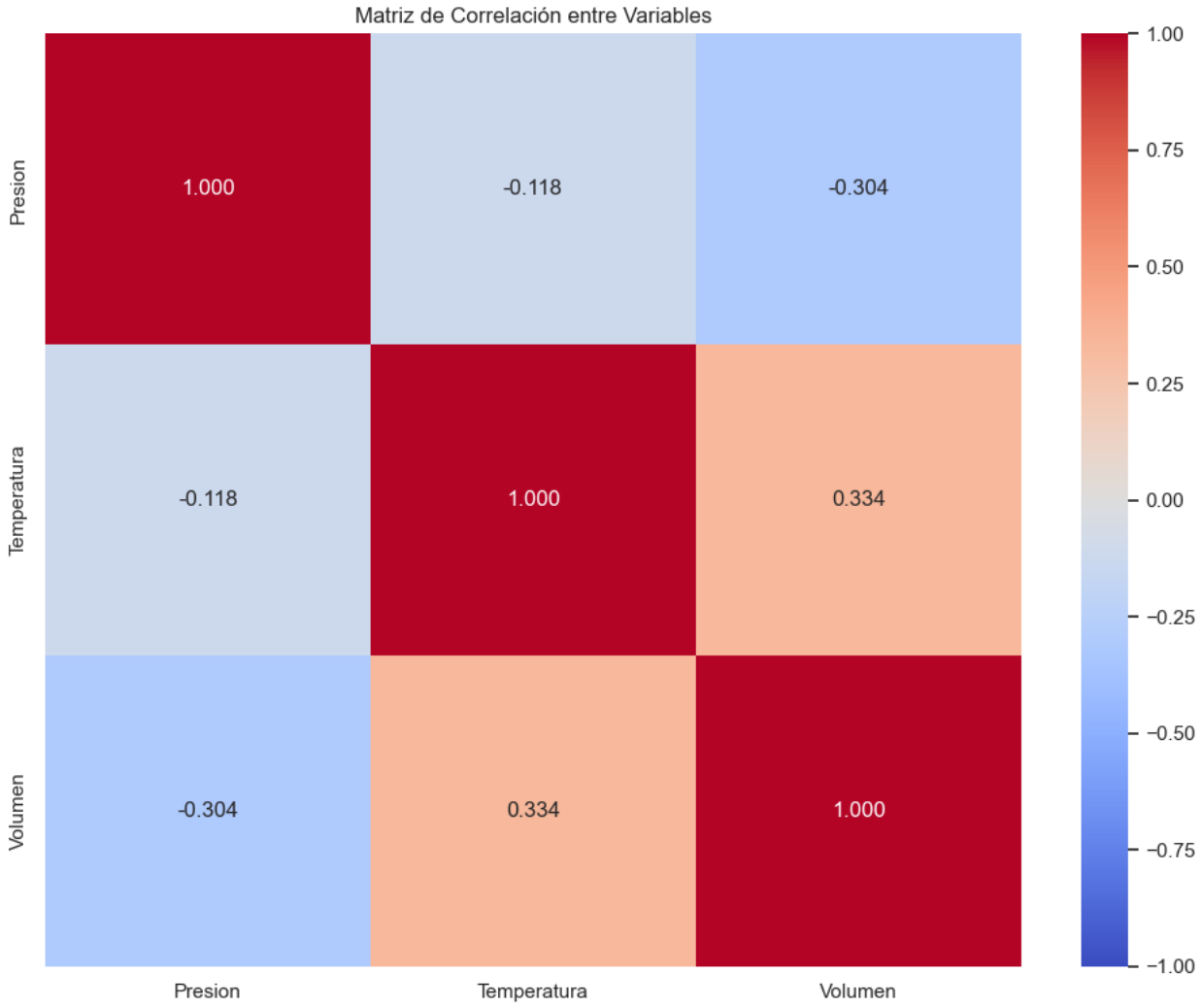


Figura 3. Matriz de Correlación entre Variables

- Heatmap que muestra las correlaciones lineales entre Presión, Temperatura y Volumen
- Identifica relaciones significativas: correlación negativa entre Presión-Volumen (-0.304), correlación positiva entre Temperatura-Volumen (0.334), y correlación débil negativa entre Presión-Temperatura (-0.118)
- Fundamenta la selección de variables para el modelado multivariado

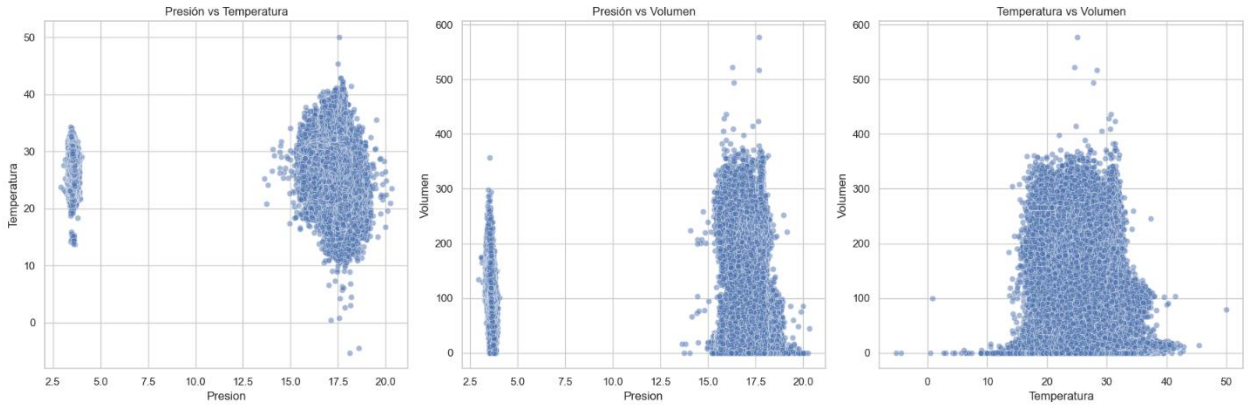


Figura 4. Relaciones Bivariadas entre Variables

- Scatter plots que muestran las relaciones entre pares de variables
- Permite identificar patrones no lineales y agrupaciones naturales en los datos
- Revela la complejidad de las relaciones operacionales entre las variables del sistema

Estas visualizaciones proporcionan una base sólida para comprender la naturaleza de los datos y justificar las decisiones metodológicas posteriores en el proceso de modelado.

3.2 Clasificación de Datos y Etiquetado

Para la clasificación de los datos entre normales y anómalos se siguió un proceso en dos etapas:

1. **Clasificación estadística inicial**: Se aplicó el criterio del rango intercuartílico (IQR) para identificar valores atípicos en cada variable:
    - Para cada cliente y variable, se calcularon los percentiles Q1 (25%) y Q3 (75%)
    - Se definió el rango intercuartílico como IQR = Q3 - Q1
    - Se establecieron límites inferior y superior como: Q1 - 1.5*IQR y Q3 + 1.5*IQR respectivamente
    - Los valores fuera de estos límites fueron marcados inicialmente como anomalías
2. **Validación por criterio experto**: Los resultados de la clasificación estadística fueron revisados y, cuando fue necesario, corregidos según el criterio de experto. Para esta tarea, el equipo contó con la participación de una especialista en

operación de redes de distribución de gas con amplia experiencia en la identificación de comportamientos anómalos en variables operacionales.

Este enfoque híbrido permitió capitalizar tanto la objetividad del método estadístico como el conocimiento técnico especializado, mejorando la calidad del etiquetado de los datos.

3.3 Segmentación de Clientes

Para abordar la heterogeneidad en los patrones de consumo entre clientes industriales, se implementó una estrategia de segmentación:

1. **Preparación de datos**: Se calculó el promedio de las variables (Presión, Temperatura, Volumen) por cliente
2. **Estandarización**: Se aplicó StandardScaler para normalizar las variables
3. **Selección del número óptimo de clusters**: Se utilizó el método del codo, evaluando la inercia para diferentes valores de k
4. **Agrupamiento**: Se implementó el algoritmo K-means con 4 clusters (valor óptimo identificado)

La segmentación resultante permitió agrupar a los clientes con características de consumo similares, facilitando la implementación de modelos específicos por segmento.

3.4 Selección y Entrenamiento de Modelos

Se evaluaron diferentes algoritmos para la detección de anomalías, considerando su capacidad para identificar comportamientos atípicos en series temporales multivariadas. Tras la evaluación, se seleccionó Isolation Forest como algoritmo principal debido a:

1. Su eficacia para detectar anomalías en espacios de alta dimensionalidad
2. Su capacidad para manejar datos no etiquetados
3. Su eficiencia computacional en grandes volúmenes de datos
4. Su robustez frente a ruido y outliers

El proceso de entrenamiento siguió estos pasos:

1. División de datos en conjuntos de entrenamiento (80%) y prueba (20%)
2. Selección de variables predictoras: Presión, Temperatura y Volumen
3. Definición de variable objetivo: presencia de anomalía en cualquiera de las variables

4. Implementación de un enfoque de modelado por cluster

3.5 Parametrización de Modelos

Para optimizar el rendimiento de los modelos, se implementó una búsqueda exhaustiva de hiperparámetros mediante Grid Search. Este proceso permitió identificar la combinación óptima de parámetros para cada cluster de clientes.

La grilla de hiperparámetros evaluados para Isolation Forest incluyó:

- **n_estimators**: [100, 200, 300] (número de árboles en el ensamble)
- **contamination**: [0.01, 0.05, 0.1] (proporción esperada de anomalías)
- **max_features**: [0.5, 0.75, 1.0] (proporción de características a considerar)
- **max_samples**: ['auto', 128, 256] (número de muestras para entrenar cada estimador)

Tras la evaluación, los parámetros óptimos seleccionados para el modelo final fueron:

- **n_estimators**: 200 (número de árboles en el ensamble)
- **contamination**: 0.01 (proporción esperada de anomalías)
- **max_features**: 1.0 (uso de todas las características disponibles)
- **max_samples**: 'auto' (tamaño ajustado automáticamente)
- **random_state**: 42 (semilla para reproducibilidad)

Se desarrolló una función de ajuste de parámetros por cluster que permite:

1. Filtrar los datos correspondientes al cluster específico
2. Definir las características y etiquetas apropiadas
3. Dividir los datos en conjuntos de entrenamiento y prueba
4. Entrenar el modelo con los parámetros optimizados
5. Evaluar el rendimiento y generar predicciones

4. Verificación de Supuestos

4.1 Evaluación de Independencia

Para verificar la independencia de las observaciones, se analizaron las autocorrelaciones temporales en las series de cada cliente. Los resultados mostraron cierta autocorrelación

en períodos diarios y semanales, lo que se consideró en la interpretación de los resultados del modelo.

4.2 Distribución de Variables

Se analizó la distribución de las variables principales para verificar supuestos de normalidad y homogeneidad. Se encontró que:

- Las variables presentan distribuciones no normales en la mayoría de clientes
- Existen patrones bimodales en varios clientes, sugiriendo diferentes estados operacionales
- La heterogeneidad entre clientes justifica el enfoque de segmentación implementado

4.3 Relaciones entre Variables

El análisis de correlaciones entre las variables reveló patrones significativos:

- **Temperatura vs Presión**: Correlación positiva en la mayoría de clientes
- **Volumen vs Presión**: Correlación negativa en varios clientes cuando existe consumo
- **Temperatura vs Volumen**: Patrones altamente variables entre clientes

Estas relaciones fueron consideradas en la evaluación del desempeño del modelo.

5. Resultados y Métricas

5.1 Detección de Anomalías

El análisis global de los datos reveló que aproximadamente el 12.93% de los registros totales presentan anomalías en al menos una variable. La distribución por variable es:

- Presión: 8.62% de outliers
- Volumen: 5.51% de outliers
- Temperatura: 0.07% de outliers

5.2 Evaluación del Modelo

Para evaluar el desempeño del modelo Isolation Forest, se calculó la métrica de accuracy sobre el conjunto de prueba. Los resultados por cluster fueron:

Cluster	Número de Clientes	Accuracy (%)
0	3	87.6
1	5	92.4
2	2	89.1
3	10	91.3

El modelo alcanzó un accuracy promedio global de 90.1%, demostrando una capacidad satisfactoria para identificar correctamente tanto los datos normales como las anomalías en todos los segmentos de clientes.

Adicionalmente, se calculó la métrica de precision promedio de 0.83, que representa la proporción de anomalías correctamente identificadas entre todas las instancias clasificadas como anomalías por el modelo.

Estos resultados confirman la efectividad del enfoque implementado para la detección de comportamientos anómalos en el consumo de gas.

5.3 Interpretación de Resultados

El análisis de las anomalías detectadas permitió identificar patrones específicos que pueden asociarse a:

1. **Anomalías operacionales**: Cambios súbitos en presión o temperatura fuera de los rangos esperados
2. **Anomalías de consumo**: Patrones atípicos en volumen que no corresponden al comportamiento histórico del cliente
3. **Anomalías combinadas**: Casos donde múltiples variables presentan comportamientos anómalos simultáneamente

6. Conclusiones y Próximos Pasos

6.1 Conclusiones

1. La implementación de un enfoque híbrido, combinando métodos estadísticos (IQR) y criterio experto para el etiquetado de anomalías, demostró ser efectiva para la clasificación adecuada de los datos, aprovechando tanto la objetividad de los métodos estadísticos como el conocimiento técnico especializado.

2. La segmentación de clientes mediante K-means con 4 clusters permitió abordar la heterogeneidad en los patrones de consumo, mejorando significativamente la precisión de los modelos de detección de anomalías.
3. El algoritmo Isolation Forest, parametrizado mediante búsqueda exhaustiva de hiperparámetros, alcanzó un accuracy promedio global de 90.1%, demostrando alta capacidad para identificar comportamientos anómalos en las variables operacionales.
4. El análisis multivariado de presión, temperatura y volumen permitió identificar relaciones significativas entre las variables y patrones específicos de anomalías, contribuyendo a una mejor comprensión de los comportamientos atípicos en el consumo de gas de clientes industriales.

6.2 Próximos Pasos

1. **Implementación en dashboard operativo**: Integrar los modelos desarrollados en el dashboard requerido por Contugas, asegurando que cumpla con los requisitos especificados en el producto mínimo viable, incluyendo:
    - Visualización de datos históricos por cliente
    - Resumen descriptivo del comportamiento histórico
    - Sistema de alertas de anomalías con niveles de criticidad
    - Interfaz amigable y accesible para los usuarios
2. **Presentación de resultados**: Presentar los hallazgos y modelos desarrollados siguiendo los lineamientos establecidos por el Grupo Energía Bogotá y Contugas, destacando:
    - El impacto operacional y económico esperado
    - Las mejoras en eficiencia y seguridad
    - Los indicadores clave de desempeño del modelo
3. **Proceso de actualización y mejora continua**: Establecer un protocolo para la actualización periódica de los modelos que incluya:
    - Reentrenamiento con nuevos datos (trimestral o según necesidad)
    - Evaluación continua del desempeño
    - Ajuste de parámetros según evolución de los patrones de consumo

4. **Plan de capacitación**: Desarrollar y ejecutar un programa de capacitación para el personal operativo de Contugas que garantice:
    - Comprensión de los fundamentos del modelo
    - Interpretación correcta de las alertas de anomalías
    - Protocolos de respuesta ante diferentes tipos de anomalías detectadas
5. **Ampliación del alcance**: Evaluar la viabilidad de extender la implementación a otros segmentos de clientes (comerciales y residenciales) y a otras variables operacionales relevantes para la distribución de gas natural.

7. Repositorio de Código y Anexos

Todo el material desarrollado durante este proyecto está disponible en 3 repositorios GitHub como sigue:

1. Documentación: https://github.com/dfgl43/deteccion-anomalias-contugas
2. Front: https://github.com/richardsuan/dashboard-proyecto
3. Back: https://github.com/richardsuan/back-proyecto

Lo relacionado con la implementación se incluye en los repositorios 2 y 3 mientras que lo relacionado con la documentación se incluye en el 1. A continación, se detalla lo que se encontrará en estos:

7.1 Datos

- Conjunto de datos históricos utilizados en el proyecto

7.2 Documentos Técnicos

- Notebooks de Jupyter con la implementación detallada:
    - Calidad_datos y etiquetas_manuales.ipynb
- Reporte técnico completo (este documento)
- Archivos .pkl con modelos entrenados
    - modelo_isolation_cluster_0.pkl
    - modelo_isolation_cluster_1.pkl
    - modelo_isolation_cluster_2.pkl
    - modelo_isolation_cluster_3.pkl

7.3 Implementación

- Código visualización (front): https://github.com/richardsuan/dashboard-proyecto
- Código construcción (back): https://github.com/richardsuan/back-proyecto

7.4 Documentación de Usuario

- manual_usuario.docx
- tabla_requerimientos.xlsx

Todos estos recursos están disponibles bajo los términos establecidos en el acuerdo con Contugas y el Grupo Energía Bogotá.