Guía 4 – Anteproyecto

Gerencia de Proyectos de Analytics

Santiago Esteban Forero Hoyos,

César Martínez,

Sofia Salazar Suaza

Camila Alejandra Velasco Ruiz

Noviembre 2024.

UNIVERSIDAD DE LOS ANDES.

DEPARTAMENTO DE INGENIERIA INDUSTRIAL

MAESTRIA EN INTELIGENCIA ANALITICA DE DATOS

APRENDIZAJE NO SUPERVISADO

**Parte I. Características de calidad de los datos**

1. **Disponibilidad y nivel de los datos**

Para la realización del proyecto de mejora en detección de consumos atípicos de Contugas, la empresa cuenta con datos históricos provenientes de sus sistemas de monitoreo, que incluyen mediciones horarias de presión (bar), temperatura (°C) y volumen de gas (m³), que han sido entregados en forma estructurada como tablas individuales de Excel para 20 clientes. Adicionalmente, la empresa mencionó tener más variables características sobre la identificación del cliente, geográficas y propiedades de las tuberías. Sin embargo, éstos no hacen parte del archivo proporcionado. Finalmente, se ha propuesto enriquecer el análisis con datos externos como condiciones meteorológicas (temperatura y humedad) y registros internos de mantenimiento e inspecciones para mejorar la precisión en la detección de irregularidades​. Sin embargo, la accesibilidad de los datos meteorológicos dependerá de la variable geográfica mencionada por la empresa, para poder entender la locación de los clientes y buscar en bases de datos de meteorología la información requerida. De igual forma, se tendría que consultar con los clientes el tema de los registros de mantenimiento y construir una base de datos basada en sus respuestas, lo cual puede ser complicado de obtener. No obstante, por ahora, se asumirá que solo estarán disponibles los datos proporcionados por la empresa en el archivo de Excel.

1. **Dimensiones de valor de los datos**
2. Granularidad: Los datos provistos por Contugas poseen un nivel de detalle alto, con registros hora a hora que permiten un monitoreo preciso de los patrones de consumo, detectando variaciones anómalas en presión, temperatura y volumen para cada cliente.
3. Fidelidad y Exactitud: Los datos, provenientes de sensores automatizados, representan con fidelidad las condiciones reales de consumo y operación, minimizando así sesgos humanos. No obstante, su exactitud puede depender de la calibración y el mantenimiento de los sensores, por lo que sería enriquecedor tener disponibles los registros de mantenimientos como se mencionó anteriormente.
4. Integridad: Los datos son íntegros en la tabla de cada cliente ya que las mediciones horarias están completas y organizadas cronológicamente. A través de los clientes, la integridad puede variar dependiendo de las diferencias en mantenimiento o fallos técnicos en los dispositivos de medición. Por ejemplo, algunos clientes tienen menos registros que otros, aunque todos se mantienen en un promedio de 41,000 datos.
5. Accesibilidad (tiempo de acceso): Los datos están almacenados y estructurados por cliente, lo que permite un acceso rápido para análisis. Sin embargo, la velocidad puede depender del sistema de almacenamiento (local o en la nube) y su optimización a la hora de conectarlos con la aplicación y procesarlos en el modelo. Con respecto a los datos históricos, dado que están presentados en una tabla y no son modificables, el acceso a éstos para el análisis de patrones en el pasado siempre será fácil y directo. De esta forma, lo anteriormente mencionado sobre la optimización de conexión estará más que todo relacionado a la acumulación en tiempo real de los nuevos datos, donde se encuentra el mayor reto en este aspecto, pues el tiempo de respuesta del modelo no debería ser demorado.
6. Edad de los datos: Los datos históricos abarcan desde 2019 hasta 2023, lo que les otorga valor tanto para analizar patrones pasados como para realizar proyecciones futuras, asegurando su utilidad en el tiempo para generar resultados robustos de valor al momento de detectar consumos anómalos.
7. Habilidad de entendimiento: Los datos son altamente robustos en contenido, permitiendo descubrir patrones de consumo atípicos, correlaciones entre variables y eventos anómalos que podrían no ser evidentes sin su análisis. Es decir, no se pueden tomar conclusiones sobre el consumo de un momento específico (determinar si es muy alto o bajo) sin tener en cuenta el consumo en otros momentos y las condiciones detrás de éste (presión y temperatura). No obstante, el tener estos datos por sí solos tampoco es concluyente, pues éstos deben ser procesados por un modelo que considere la correlación y patrones entre las variables para determinar eficientemente si un consumo es atípico o no.
8. Entendimiento de la ubicación dentro de la cadena de suministro: Los datos provienen del último paso en la cadena de suministro del servicio de gas que Contugas le provee a sus clientes. Una vez el gas llega a los clientes, éstos hacen uso de éste y ese es el momento en el que su consumo es registrado junto a su presión y temperatura. Esto, posteriormente, se almacena en una base de datos estructurada para cada cliente bajo una fecha y hora específica. Es importante entender esto, pues la cantidad de gas registrada no depende de cuánto produce Contugas, sino de cuánto de esta producción consume el cliente final.
9. **Características de calidad**
10. **Completitud de los datos:**

Como se puede ver en la siguiente gráfica, no hay ningún valor nulo en ninguna de las variables de los 20 clientes:

*Gráfica 1: completitud de los datos*

A graph of different colored squares

Description automatically generated with medium confidence

1. **Formato de los datos:**

A continuación, se muestra el formato de los datos obtenidos:

*Tabla 1: Formato de variables*

|  |  |
| --- | --- |
| **Variable** | **Formato** |
| Fecha | datetime64[ns] |
| Presión | float64 |
| Temperatura | float64 |
| Volumen | float64 |
| Número\_Cliente | object |

Como se puede ver, la variable de Fecha tiene un formato apropiado (datetime), mientras que las variables continuas como Presión, Temperatura y Volumen tienen un formato float, lo cual tiene sentido con su naturaleza. Por último, el número de cliente tiene un formato object, que significa que, aunque es un número, éste se utiliza más como clasificador que como medida y por lo tanto no es un entero/integer.

1. **Consistencia y claridad de los datos:**

Para definir la consistencia y claridad de los datos se utilizaron las siguientes estadísticas descriptivas:

*Tabla 2: Estadísticas descriptivas*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Presión** | **Temperatura** | **Volumen** |
| **count** | 847960.00 | 847960.00 | 847960.00 |
| **mean** | 16.07 | 25.20 | 62.33 |
| **min** | 2.93 | -5.26 | 0.00 |
| **0.25** | 17.10 | 22.69 | 0.00 |
| **0.5** | 17.57 | 25.38 | 21.77 |
| **0.75** | 17.69 | 27.89 | 99.32 |
| **max** | 20.31 | 50.02 | 577.41 |
| **std** | 4.19 | 3.79 | 80.50 |

Como se puede ver, los todos los datos, al seguir los formatos apropiados, son consistentes con su definición, como se explicó en el punto anterior (formato). Es decir, las fechas son fechas y las medidas son medidas. Sin embargo, en este punto se demuestra que los valores de las variables también son consistentes con su naturaleza. Por ejemplo, en temas de presión, podemos ver que la mínima es de 2.93, mientras que la máxima es de 20.31, con un promedio de 16.07. Si bien esto implica que los valores “normales” deberían estar alrededor de 16.07, una presión mínima de 2.93 no es magníficamente alejada del promedio, por lo que se podría decir que el rango de valores de presión no presente inconsistencias. No obstante, podemos verificar fácilmente el comportamiento de esta variable con un gráfico de dispersión a lo largo del tiempo:

*Gráfica 2: presión del gas a lo largo del tiempo*

A graph with blue lines

Description automatically generated

Como se puede apreciar, hay clientes que mantienen una presión muy baja constantemente mientras que otros la mantienen muy alta. No obstante, los que la mantienen en rangos bajos tienden a ser menos y, por ende, el promedio de presión tiende a los números superiores del rango. Ante esto, podemos ver que los valores son consistentes y se abre una oportunidad para investigar este peculiar comportamiento en la parte 3 de esta guía.

Por otro lado, la temperatura varía de -5.26 C a 50.02 C. Dado que no se tiene conocimiento de las propiedades de la temperatura del gas, procedimos a realizar un gráfico de dispersión de la temperatura del gas a lo largo del tiempo:

*Gráfica 3: temperatura del gas a lo largo del tiempo*

A graph with blue lines

Description automatically generated

Como se puede ver, la temperatura del gas sigue un comportamiento cíclico dependiendo de la época del año manteniendo un promedio de 25.20 C. No obstante, sí hay datos atípicos que son los que podrían terminar explicando los consumos atípicos de los clientes una vez se construya el modelo. Sin embargo, estos datos atípicos se mantienen en un rango razonable. Es decir, no hay datos alarmantes que indiquen que sean errores de medición o que estén midiendo otra variable que no sea temperatura.

Finalmente, el volumen es la variable con mayor rango (0 a 577.41) y dispersión según el siguiente gráfico:

*Gráfica 4: volumen del gas a lo largo del tiempo*

A screen shot of a graph

Description automatically generated

Con un promedio de 62.33, no se puede identificar un patrón específico visualmente como se hizo con las otras variables. En este caso, valdría la pena entrar a ver el consumo individual de cada cliente para encontrar sus patrones específicos. No obstante, hasta ahora, se puede apreciar que, si bien hay datos atípicos, la mayoría de los datos se mantienen en el mismo rango y no hay evidencia para descartar que haya consistencia en éstos. Por otro lado, se podría considerar la limpieza de estos datos atípicos, pero se debería evaluar el impacto que tendría removerlos sobre el desempeño del modelo final.

Teniendo en cuenta el análisis previo, se podría concluir que hay consistencia en los datos y que no se estarán comparando “peras con manzanas” en ningún caso ya que cada variable parece abarcar las medidas de sus datos correctamente.

**Parte II. Técnicas de limpieza de datos**

1. **Razones para limpiar los datos**

Se realiza la validación de los datos contenidos en la dataframe:

1. Columnas y filas con datos vacíos: No existen datos nulos
2. Valores extremos, de acuerdo con el calculo de la media, desviación estándar de cada columna y la definición de un límite superior e inferior basado en la desviación estándar de la media, de acuerdo con esto se identifican los outliers considerando que el resto de los datos presentan una distribución normal.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

De acuerdo con el grafico anterior podemos observar que la variable volumen, presenta una gran cantidad de outliers.

1. **Soluciones de limpieza de datos**

De acuerdo con la validación de datos anterior se procederá a imputar los outliers identificados en las variables volumen evidencia en la gráfica anterior.

***Botplox antes de imputar los outliers***

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

***Botplox despues de imputar los outliers***

Gráfico

Descripción generada automáticamente

De acuerdo con los pasos ejecutados en la limpieza de datos, podemos concluir que son unos datos bastante completos los cuales no requieren grandes esfuerzos de limpieza, sin embargo, se presentan outliers los cuales se imputan para mejorar los resultados de los análisis a realizar

**Parte III. Identificación de técnicas para un primer entendimiento de los datos**

1. **Técnicas para entender los datos**

* **Imputación:** Esta técnica es necesaria ya que se aplica cuando se presentan outliers en las variables, de esta forma no se eliminan los registros, si no se realiza una asignación de acuerdo con los valores superiores inferiores de los datos permitiendo así mejor respuesta en los análisis.
* **Extracción de patrones con transformaciones y normalización:** Realizar una transformación para garantizar que los datos se mantengan dentro de rangos o escalas en el cual no existan datos atípicos. Ayudando así en el análisis de los datos.
* **Recopilar y resumir datos:** Se realiza estadística descriptiva de los datos contenidos en el dataframe generando una consistencia y claridad de los datos, y visualizaciones que permitiendo un entendimiento de negocio más rápido de los datos.
* **Segmentar los datos y encontrar agrupaciones:** De acuerdo con el dataframe entregado, se puede implementar clúster con los clientes contenidos en las diferentes pestañas, esta agrupación permite generar las agrupaciones

***\*El desarrollo de la guía anterior se realiza mediante Python se adjunta enlace de script GITHUB https://github.com/cvelascor1/Guia-\_4.git.***