Desarrollo Guía 4 - Gerencia de Proyectos de Analytics
Equipo 21 Contugas

I. Calidad de los datos

Completitud: La base de datos contiene información de Fecha, Presión, Volumen y Temperatura para 20 clientes de Contugas. En total se cuenta con 847960 registros, de los cuales 41059 corresponden al cliente 10 siendo el valor mínimo encontrado, mientras que el cliente con más registros es el del cliente 5 con 43415 registros. No se presentan valores nulos para ninguna de las variables reportadas.

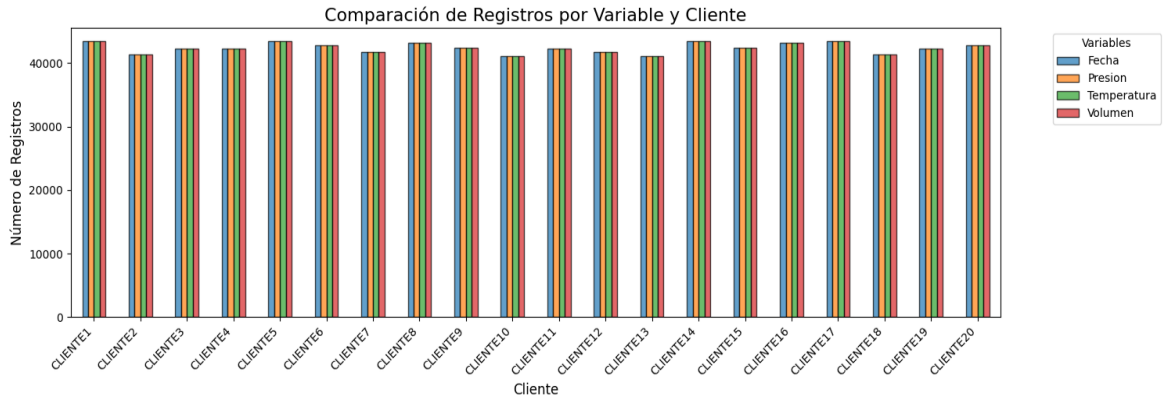


Ilustración 1. Comparación del No de registros por variables

En la ilustración 1, cada barra dentro del grupo de un cliente corresponde a una variable, y su altura representa el número de registros disponibles para esa variable. Como se puede observar todos los datos están completos y equilibrados entre las variables (Fecha, Presión, Temperatura, Volumen) para todos los clientes (CLIENTE1 a CLIENTE20). No se observan discrepancias significativas, lo que sugiere que no hay valores faltantes en las variables para los clientes.
Consistencia: Los formatos de los datos son Fecha (datetime64), Presión (float64) Temperatura (float64) y Volumen (float64) los cuales son adecuados para cada tipo de dato.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 847960 entries, 0 to 847959
Data columns (total 5 columns):
 #   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
dtypes: datetime64[ns](1), float64(3), object(1)
memory usage: 32.3+ MB
```

Ilustración 2. Resumen de la estructura y características del DataFrame formado

Los valores de fechas son consistentes ya que van desde el 14 de enero de 2019 hasta el 31 de diciembre de 2023. Los valores de presión están entre 2.93 y 20.3 bar. El rango del Volumen es 0 a 577.41 m3 de gas. La temperatura mínima reportada es -5.25 °C y la máxima 50.01 °C, en total hay registro de dos datos de temperatura menores a 0°C.

La identificación de datos duplicados se realiza teniendo en cuenta la variable Fecha y se observa que los clientes 2 y 18 tienen de a 10 datos duplicados, mientras que los clientes 3, 8, 11 y 16 tiene 2 duplicados cada uno. El total de duplicados corresponde a un porcentaje muy bajo por lo tanto se decide eliminarlos y dejar el primer dato encontrado en la base de datos.

	Total Columns	Total Rows	Missing Values (Total)	Missing Values (Columns)	Column Data Types	Duplicated Rows
CLIENTE1	4	43412	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE2	4	41382	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE3	4	42248	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE4	4	42305	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE5	4	43415	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE6	4	42808	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE7	4	41776	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE8	4	43147	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE9	4	42428	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE10	4	41059	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE11	4	42248	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE12	4	41776	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE13	4	41059	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE14	4	43415	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE15	4	42428	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE16	4	43147	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE17	4	43412	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE18	4	41382	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE19	4	42305	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0
CLIENTE20	4	42808	0	{'Fecha': 0, 'Presion': 0, 'Temperatura': 0, '...	{'Fecha': datetime64[ns], 'Presion': float64, ...	0

Tabla 1.

De la tabla 1, se muestra que se tienen 4 columnas asociadas a la fecha y los valores de presión, temperatura y volumen por fecha. Esto demuestra consistencia en la estructura de los datos entre los clientes. El número de filas varia ligeramente entre los clientes por ejemplo el cliente 1 tiene 43412 registros y el cliente 20 tiene 42808 registros. Ningún cliente tiene valores faltantes lo cual asegura la completitud de los datos. Todas las columnas estas completas lo cual confirma que no hay pérdida de información en variables. Se muestran los tipos de datos reflejando la consistencia en el formato datos. No hay filas duplicadas lo cual garantiza la unicidad de los registros.

Claridad: Evaluar estadísticas descriptivas, tipos de datos y posibles valores atípicos.

Variable presión: La mayoría de los clientes tienen valores promedio de presión consistentes alrededor de 17.5 con una varianza baja, lo que refleja mediciones estables. El cliente3 y el cliente11: Presentan valores promedio, mínimos y máximos significativamente más bajos que los demás clientes, lo que podría deberse a diferencias en los sistemas de medición o características inherentes al cliente.

El cliente8 y el cliente16 tienen mayor dispersión en sus mediciones, reflejadas en su alta varianza, lo que podría ser relevante para estudios de estabilidad del sistema.

El cliente4, cliente12 y cliente7 presentan valores máximos notablemente altos que podrían indicar eventos excepcionales o anomalías.

Análisis estadístico para la variable: Presion

Cliente	Media	Varianza	Mínimo	Máximo
CLIENTE1	17.535934	0.128386	15.742337	18.074274
CLIENTE10	17.473694	0.018207	16.469864	18.573079
CLIENTE11	3.545512	0.006848	2.934873	3.954039
CLIENTE12	17.833667	0.175134	13.740922	20.232110
CLIENTE13	17.492521	0.014416	16.505077	18.471047
CLIENTE14	17.517128	0.136231	13.616877	20.028395
CLIENTE15	17.495812	0.019406	16.514496	18.493234
CLIENTE16	16.705058	0.242393	14.734521	19.440780
CLIENTE17	17.533599	0.144412	15.265703	18.445378
CLIENTE18	17.554599	0.063612	16.215105	17.821359
CLIENTE19	17.667217	0.105514	16.247484	18.258619
CLIENTE2	17.526440	0.101751	16.129015	18.106402
CLIENTE20	17.579587	0.166682	14.849737	18.578203
CLIENTE3	3.569593	0.006499	3.057171	4.037030
CLIENTE4	17.639754	0.152739	15.214996	20.112931
CLIENTE5	17.489345	0.137192	14.489896	19.009352
CLIENTE6	17.569108	0.157506	13.810026	19.720870
CLIENTE7	17.490812	0.202553	14.418261	20.307852
CLIENTE8	16.705630	0.358331	14.996490	18.950955
CLIENTE9	17.466627	0.024931	16.486477	18.430469

Ilustración 3. Análisis variable presión

Variable temperatura. Existe una amplia variabilidad en las temperaturas promedio, con algunos clientes mostrando temperaturas más consistentes (baja varianza) y otros presentando mayores fluctuaciones.

Valores mínimos como cliente7 (0.43) y cliente12 (-5.26) podrían indicar datos erróneos o problemas en los sensores de medición.

El cliente12 tiene una de las varianzas más altas y un valor mínimo anómalo, pero también alcanza el valor máximo más alto. Los clientes16, cliente17 y cliente2 representan temperaturas promedio más consistentes y dentro de rangos esperados.

Análisis estadístico para la variable: Temperatura

Cliente	Media	Varianza	Mínimo	Máximo
CLIENTE1	25.575853	7.596893	15.401803	32.869112
CLIENTE10	23.500489	6.835477	14.291032	35.069721
CLIENTE11	26.307152	7.263219	13.731025	34.355276
CLIENTE12	26.689551	24.983546	-5.257899	50.019853
CLIENTE13	21.591365	8.116225	12.240511	29.753364
CLIENTE14	26.658520	24.672655	5.676933	45.378176
CLIENTE15	24.026489	9.647350	14.712874	35.661349
CLIENTE16	27.481311	5.988756	19.146101	32.658456
CLIENTE17	25.849342	5.588592	18.445096	31.917560
CLIENTE18	27.668722	5.984329	16.380780	33.789362
CLIENTE19	23.223832	9.142742	16.181982	30.642243
CLIENTE2	27.673040	5.500693	17.884059	35.208346
CLIENTE20	25.407173	7.215496	14.767081	33.808006
CLIENTE3	26.348434	6.859122	14.933032	34.009233
CLIENTE4	23.241102	9.142393	12.711813	36.897044
CLIENTE5	23.665282	25.515723	10.500638	41.762232
CLIENTE6	26.393254	7.410751	14.157366	34.186338
CLIENTE7	23.622548	24.215665	0.433436	39.982239
CLIENTE8	26.814551	8.829751	14.838703	37.368959
CLIENTE9	22.073596	8.495440	11.820104	31.783964

Ilustración 4. Análisis variable temperatura

Variable volumen. La variación en los promedios y las varianzas entre clientes es significativa, lo que sugiere que cada cliente opera bajo diferentes condiciones de uso o consumo.

Los valores mínimos reportados como 0.0 requieren revisión. Aunque puede ser razonable en algunos casos, podrían ser resultado de interrupciones o fallos en la medición o no consumo.

El cliente8 y el cliente16 tienen los más altos valores, pero también presentan alta varianza, lo que indica un comportamiento más dinámico o variable.

El cliente5 y el cliente14 tienen valores más estables (baja varianza y máximos bajos), lo que podría reflejar patrones de operación más consistentes.

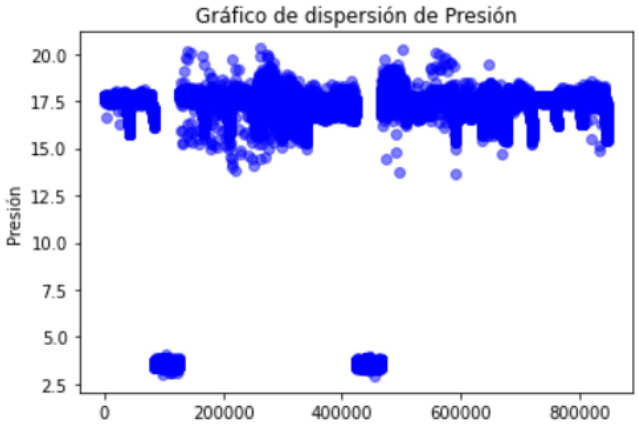
Análisis estadístico para la variable: Volumen

Cliente	Media	Varianza	Mínimo	Máximo
CLIENTE1	19.976401	63.030438	0.0	65.936644
CLIENTE10	11.788209	1935.285491	0.0	247.072990
CLIENTE11	131.648637	1171.383973	0.0	298.259573
CLIENTE12	33.731496	1759.973038	0.0	284.475087
CLIENTE13	10.276088	1696.128327	0.0	253.867916
CLIENTE14	7.301900	27.641921	0.0	36.793150
CLIENTE15	12.553861	2262.489727	0.0	398.042027
CLIENTE16	178.133946	5572.617845	0.0	409.872212
CLIENTE17	20.564425	30.982061	0.0	48.504833
CLIENTE18	61.758087	331.282371	0.0	577.413425
CLIENTE19	15.918062	3537.632623	0.0	378.267803
CLIENTE2	61.819045	306.095017	0.0	517.564868
CLIENTE20	160.816734	6852.477255	0.0	315.884153
CLIENTE3	117.500121	1588.501139	0.0	356.724008
CLIENTE4	17.351570	3715.286297	0.0	363.009776
CLIENTE5	7.817773	41.092777	0.0	89.245051
CLIENTE6	153.799976	7209.145789	0.0	366.656382
CLIENTE7	27.294379	1147.524052	0.0	175.763858
CLIENTE8	178.591825	8253.829039	0.0	522.780891
CLIENTE9	12.734989	2454.195837	0.0	366.016120

II. Técnicas de limpieza de datos.

En general, se observa que la base de datos presenta un alto nivel de calidad en términos de completitud y consistencia. Dado que las variables están completas para cada cliente, no es necesario aplicar técnicas de imputación de datos faltantes. Sin embargo, como se mencionó en la etapa anterior, algunos clientes tienen múltiples registros en la misma fecha. En este caso, se procede a eliminar los registros duplicados, manteniendo el valor máximo de la variable correspondiente a dicha fecha.

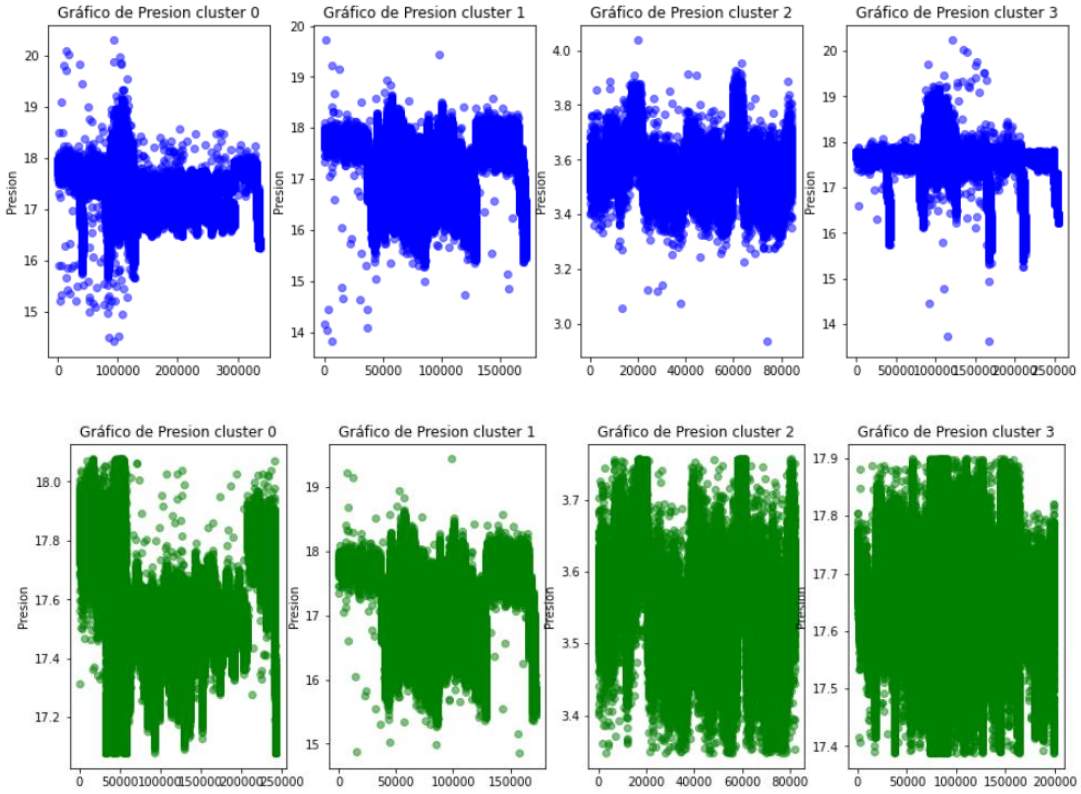
Asimismo, se identificaron dos valores de temperatura negativa, los cuales se eliminan, ya que se encuentran significativamente alejados de la media. En esta etapa, no se identifican valores atípicos (outliers) hasta agrupar a los clientes en clúster. Esto se debe a que, como se observa en la gráfica de presión, algunos clientes muestran un comportamiento muy diferente al de otros, y técnicas convencionales, como el análisis del rango intercuartil, podrían clasificarlos erróneamente como atípicos, eliminándolos de la base de datos.



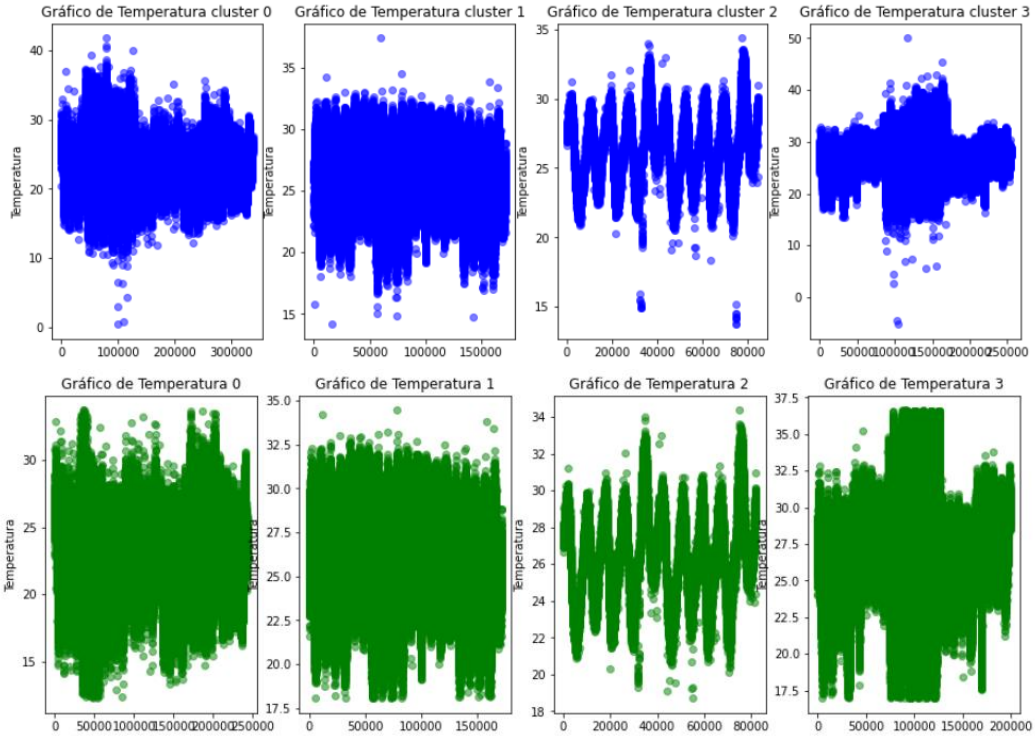
III. Identificación de técnicas para un primer entendimiento de los datos

En esta fase, se lleva a cabo una clusterización de los clientes para facilitar el manejo de los datos y evitar que los análisis futuros deban realizarse de manera individual para cada uno de los 20 clientes. Utilizando el algoritmo KMeans y el método del codo, se identifican 4 clúster. A continuación, se presenta el comportamiento de cada una de las variables por clúster, mostrando que existe una buena agrupación. En color azul se representan los valores originales, incluidos los outliers, mientras que en la gráfica verde se muestran los valores después de la eliminación de estos outliers.

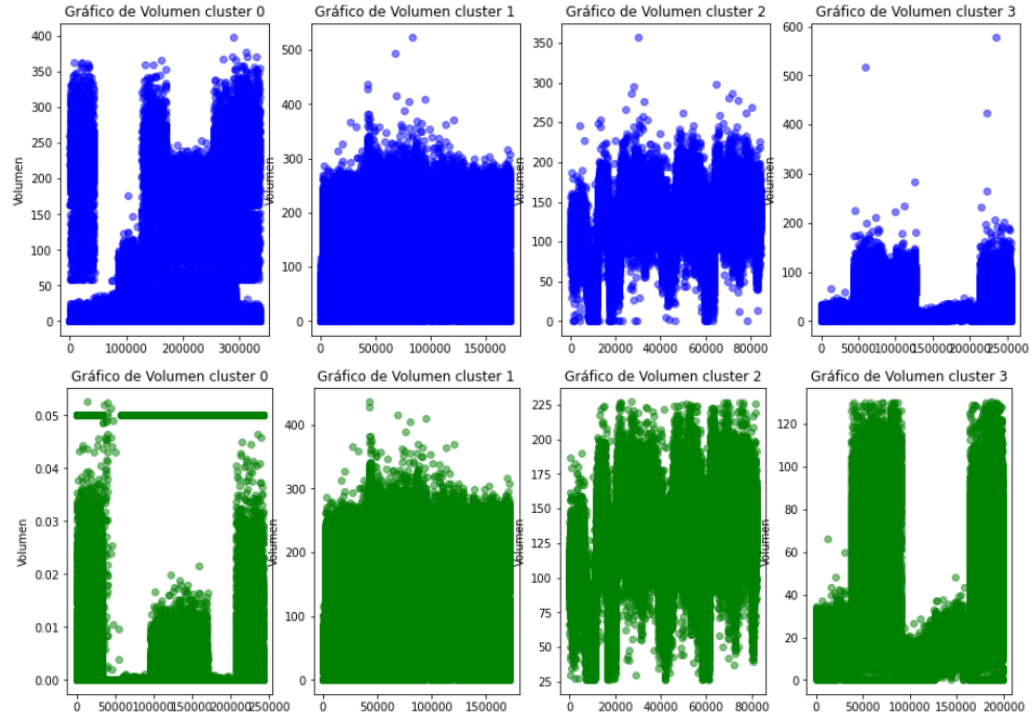
Presión



Temperatura



Volumen



Dado que solo se dispone de tres variables, se decide trabajar con todas ellas sin realizar ninguna extracción de variables ni reducción de dimensionalidad. Esto se debe a que la cantidad de variables es suficiente para el análisis y su eliminación o transformación no aportaría una mejora significativa en los resultados. Además, no se lleva a cabo imputación de datos faltantes, ya que las variables están completas y no se requiere tratamiento adicional de este tipo.

IV. Anexos

Repositorio git: **https://github.com/hugoruizb/Proyecto-Contugas**