Clustering de series temporales pertenecientes al consumo de productos para la agrupación por patrones.

David Zapata Chaves

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Daniela Serna Buitrago, Especialista (Esp) en Analítica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	Zapata Chaves [1]
Referencia Estilo IEEE (2020)	[1] D. Zapata Chaves, “Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

TABLA DE CONTENIDO

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

DTW.	Dynamic Time Warping
UMB.	Unidad de Medida Base
ERP.	Enterprise Resource Planning

RESUMEN

Lograr identificar los patrones de consumo y comportamiento del mercado se convierte en información relevante para ajustar las estrategias relacionadas a la cadena de suministro. Esto, conduce a una mejor toma de decisiones en los eslabones de compra, producción y distribución, traduciéndose en una rentabilidad y eficiencia.

Por otra parte, la identificación de segmentos específicos de productos, entendiendo sus necesidades y comportamiento, permite a la empresa personalizar ofertas y disponibilidad mejorando indicadores como nivel de servicio, días de inventario y pedido perfecto.

Este proyecto, por medio de la clusterización de series de tiempo, presenta una agrupación del comportamiento temporal de materiales por componentes, buscando diferenciarlos por tendencia, variación estacional, variación cíclica y variación irregular.

La metodología descrita, inicia con la comprensión de los datos, continua con la prepararon y limpieza, para que finalmente estos sean modelados por técnicas de aprendizaje no supervisado cuyo objetivo sea el clustering de datos, como lo es el algoritmo DTW (Dynamic Time Warping). Finalmente, luego de la experimentación, se elige la mejor opción basada en las necesidades del negocio.

Las conclusiones serán de insumo para proyectos del área de Analítica y Demanda de la cadena de suministros de una empresa de Alimentos Cárnicos en Colombia, basados en el pronóstico de consumos de materiales y el entendimiento del comportamiento portafolio.

Palabras clave — **DTW, Series de Tiempo, Segmentos de Productos, Clustering.**

Repositorio GitHub del proyecto: https://github.com/DavidZap/Time-Series-Clustering-DTW---UdeA

I. DESRIPCIÓN DEL PROBLEMA

A. Problema de negocio

Una empresa de producción de alimentos cárnicos de Colombia tiene la necesidad de entender el comportamiento temporal de los materiales que fabrica y comercializa, esto con el objetivo de desarrollar estrategias para cada segmento de productos y optimizar su proceso de pronóstico.

El portafolio es amplio y el nivel detalle que se desea obtener a futuro es un entendimiento por centro de distribución, por lo tanto, la metodología propuesta debe ser eficiente, que permita extraer insights a gran escala y generalizarlos para apoyar la toma de decisiones.

El equipo de demanda, principal cliente del ejercicio se verá altamente beneficiado, dado que les permitirá conseguir perspectivas nuevas para entender el comportamiento de forma masiva de los productos y tendrán un apoyo para definir lineamientos de distribución y solicitud de producción sobre el.

Por otra parte, el equipo de modelación de la cadena de suministros podrá sistematizar patrones comunes en grupos de series, lo que les permitirá aplicar los algoritmos de pronóstico más adecuados a cada particularidad.

B. Aproximación desde la analítica de datos

La clusterización de los productos busca la segmentación de estos por patrones de consumos, buscando así un mayor conocimiento del comportamiento del portafolio.

Para resolver esta necesidad es requerida la información histórica de los consumos, la cual se encuentra espaciado en el tiempo por semanas y la unidad con la que se registra la venta es UMB (Unidad Medida Base) década materia.

Los cluster obtenidos en el presente trabajo, serán de insumo para un posterior modelo de clasificación con el cual se espera definir el portafolio por patrones de consumo y segmentar los productos para un proceso de pronóstico más eficiente.

C. Origen de los datos

Los datos se extraen del ERP (Enterprise Resource Planning) de la empresa. Estos poseen 155 periodos de consumo de materiales, entendiendo un periodo como una semana, comenzado la fecha del 02/03/2020 y terminando el 13/02/2023.

Allí se tiene información de 14 centros de distribución, 77 materiales y 694 combinaciones de Centro – Material. Los materiales usados en este proyecto hacen parte de la línea de "Salchichas".

Como se mencionó anteriormente, el consumo por material esta dado en UMB, es decir si un paquete de salchichas 480gr, el cual tiene 15 unidades, se entendería que su UMB es el paquete de 480 gr, más no las 15 unidades individuales contenidas en el empaque.

D. Métricas de desempeño

Métrica de Machine Learning: Se elige como métrica el valor de la silueta (silhouette score) la cual es una medida para evaluar la calidad de agrupamiento (clustering) de un conjunto de datos.

Lo que se busca, es cuantificar la similitud de una serie de tiempo con su propio grupo en comparación con otros grupos, por medio de una puntuación que está dentro del rango -1 y 1, donde valores más altos indican que las series están bien ajustadas a sus grupos y alejadas de otros grupos.

La elección de esta métrica se basa en la capacidad de evaluar la calidad del clúster de manera más completa, además de su interpretabilidad por parte de equipos que no se encuentran inmiscuidos en la ciencia de datos. [1]

Métrica de negocio: Dado que no hay un baseline con el cual comparar, el equipo de demanda establece su nivel de aceptación por medio de la evaluación visual de las series en cada clúster.

II. DATOS

A. Datos Originales

Los datos obtenidos de SAP se consolidaron en un archivo con formato XLSX. Este posee 4 columnas, 78889 registros, Con un peso final de 9840 kB.

A continuación, se describirán las columnas:

'semana': Fecha del día donde inicia la semana

'Codmat': Concatenación del centro y el código del material.

'Cant': Cantidad consumida en UMB por centro – código material.

'semanacalendario': se registra la semana del año de la siguiente manera, ejem: ‘2020.01’.

El modo de acceso a los datos se encuentra en el repositorio, con el nombre “Pronosticos_ .xlsx”, en la hoja “ConFecha”.

B. Análisis Descriptivo

Al dataframe original se le agregaran las siguientes características, esto con el fin de mejorar el análisis descriptivo:

‘Centro’: Indicativo del lugar de distribución del material.

‘numeroSemana’: Número de la semana del año de la cual se toma el registro.

‘Semana Corregida’: Campo en formato fecha

‘Material’: Código del material

‘year’: Año del campo ‘Semana Corregida’

‘mes’: Mes del campo ‘Semana Corregida’

	semana	Codmat	cant	semanacalendario	centro	material	numeroSemana	Semana Corregida	year	mes
0	2020-03-02	NN13-1000497	1891.000	2020.10	NN13	1000497	10	2020-03-02	2020	3
1	2020-03-02	NN13-1000498	544.333	2020.10	NN13	1000498	10	2020-03-02	2020	3
2	2020-03-02	NN13-1000794	4370.000	2020.10	NN13	1000794	10	2020-03-02	2020	3
3	2020-03-02	NN13-1000814	10.000	2020.10	NN13	1000814	10	2020-03-02	2020	3
4	2020-03-02	NN13-1000820	489.500	2020.10	NN13	1000820	10	2020-03-02	2020	3

Ilustración 1. Datos Originales

```
df_caracteristicas.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78889 entries, 0 to 78888
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   semana            78889 non-null  object
 1   Codmat            78889 non-null  object
 2   cant              78889 non-null  float64
 3   semanacalendario  78889 non-null  object
 4   centro            78889 non-null  object
 5   material          78889 non-null  object
 6   numeroSemana      78889 non-null  object
 7   Semana Corregida  78889 non-null  datetime64[ns]
 8   year              78889 non-null  int64
 9   mes               78889 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(2), object(6)
memory usage: 6.0+ MB
```

Ilustración 2. Información de las variables

Para el análisis de la completitud de las series y sus datos atípicos, se agrupan la suma de cantidades consumidas por material y semana, de esta manera tener una perspectiva global de donde podrían estar los faltantes.

material	1000031	1000497	1000498	1000794	1000812	1000814	1000820	1000821	1000833	1000840	...	1050413	1050702	1051543	1051544
semana															
2020-03-02	4553.695	2347.000	548.333	5981.0	24.0	14861.851	7356.609	4503.815	3573.975	2498.545	...	NaN	NaN	NaN	NaN
2020-03-09	5492.006	1822.000	1095.333	5012.0	24.0	20152.220	8500.055	6022.121	4475.146	3671.445	...	NaN	NaN	NaN	NaN
2020-03-16	3870.255	3676.000	310.000	6980.0	54.0	17591.170	10662.832	2794.617	3850.982	2206.495	...	NaN	NaN	NaN	NaN
2020-03-23	2284.952	3771.000	220.334	8112.0	55.0	12575.380	7451.010	962.701	4442.017	934.211	...	NaN	NaN	NaN	NaN
2020-03-30	4020.223	3540.000	348.333	7770.0	40.0	14922.400	10086.938	1370.585	5302.665	1392.940	...	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-01-16	5917.500	2829.625	1180.000	1768.0	56.0	19041.167	13601.722	5955.619	5770.000	5275.000	...	NaN	NaN	477.00	NaN
2023-01-23	5024.438	2208.000	319.667	1362.0	40.0	16314.000	12550.486	4863.438	4954.501	3554.500	...	NaN	NaN	600.00	NaN
2023-01-30	4574.348	3158.000	484.333	1944.0	40.0	17300.000	13762.514	4977.264	4293.500	2934.500	...	NaN	NaN	532.00	1.0

Ilustración 3.Dataframe con cantidades vendidas por semana

Se procedió a normalizar los datos para entender las diferencias dentro de una misma escala. Se obtuvo el siguiente Box plot.

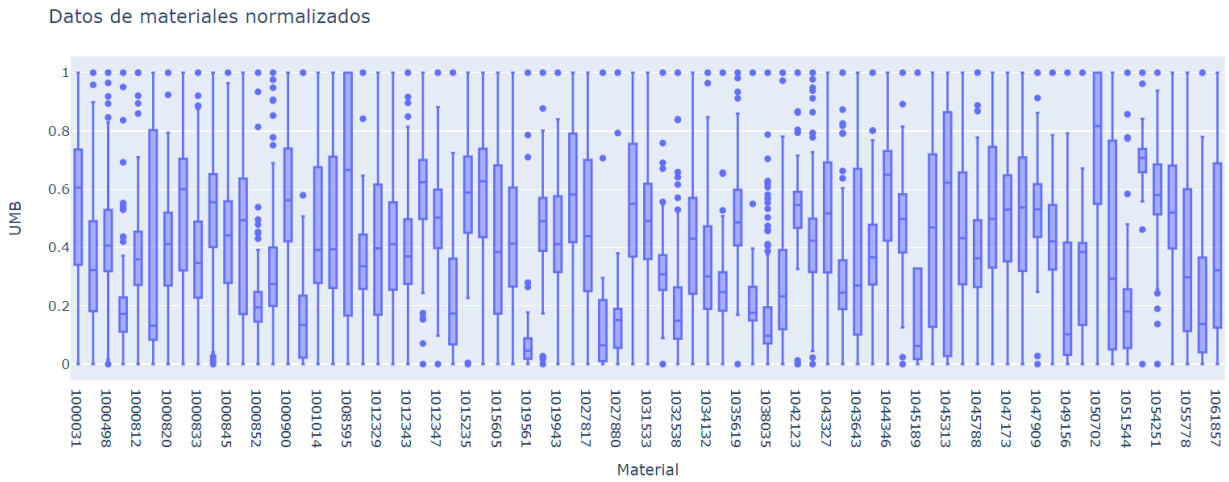

Ilustración 4. Box plot de datos de materiales normalizados

Se encuentran materiales que poseen gran cantidad de atípicos, la mayoría de estos se encuentran luego del percentil 75%. Estos serán tratados luego de la debida imputación de datos.

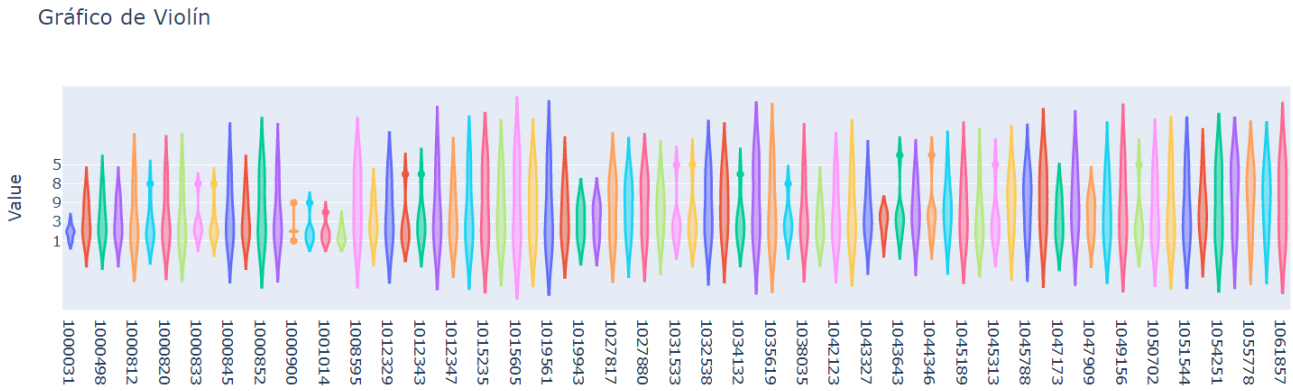

Ilustración 5. Gráfico de Violín por Materiales

En el diagrama de violín se puede observar una simetría en la distribución de la mayoría de los datos, por lo que es un indicio de que el comportamiento de los datos puede asociarse a distribuciones como la normal o la t-student.

Por otro lado, hay materiales que presentan asimetrías en la densidad de su distribución, indicando así sesgos fuertes, como es el caso de los siguientes materiales: ['1000031','1000820','1000833','1000840','1000900','1000911','1001014,'1001311','1012331','1031533','1031534','1034132','1038034','1044346','1043643','1045313','1050413'].

Se procede a realizar el tratamiento de datos atípicos. El análisis de outliers se refiere a la identificación y tratamiento de valores atípicos en un conjunto de datos, que pueden ser errores o valores extremos que afectan negativamente el análisis de los datos.

Es importante realizar este análisis antes de imputar los valores faltantes, ya que los valores atípicos pueden afectar la distribución y las relaciones entre las variables y, por lo tanto, también pueden afectar el método de imputación.[2]

El método utilizado para la detección de datos atípicos fue el Rango Intercuartil (IQR), esta medida se calcula a partir de los cuartiles de los datos y se define como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), es decir, IQR = Q3 - Q1. Luego se calcular el límite inferior Q1 - 1.5 * IQR y el límite superior Q3 + 1.5*IQR.

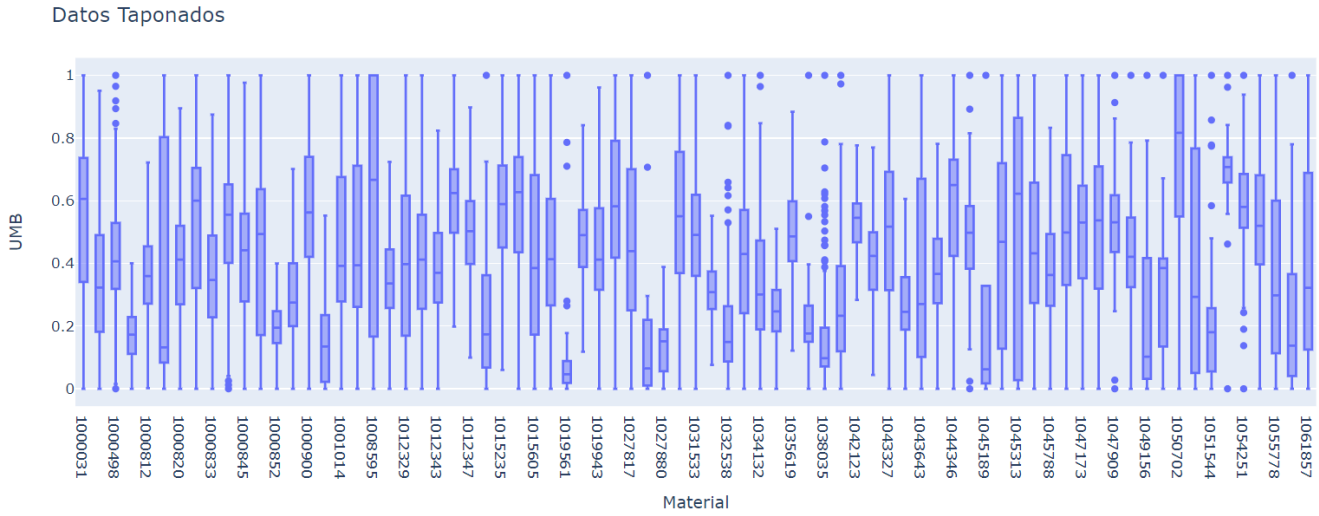

Ilustración 6.Datos de Materiales Taponados

Como resultado, por lo menos 50 de los materiales poseen 1 dato atípico. Con el fin de no eliminar ningún periodo de tiempo, se aplicó el “Método del Taponamiento” o también conocido como “Método del Recorte”. Los valores que se encontraban por debajo del límite inferior o por encima del límite superior, fueron reemplazados por el valor del límite respectivo.

Los materiales que no poseen valores nulos fueron afectados por la imputación de datos atípicos como se puede observar en la Ilustración 6. Más adelante se analizará una posible depuración de materiales con datos incompletos.

Se generar una matriz de dispersión para los primeros 20 materiales y así entender la calidad de los datos y las posibles correlaciones que existan entre ellos. Ilustración 7.

Se evidencian materiales con una calidad de datos aceptable, seguramente por la completitud que estos poseen, además se debe tener en cuenta que los materiales se encuentran agregados por todos los centros, esperando así que esto no afecte en sobremanera la distribución de estos, con excepción del material ‘1000814’. También se puede analizar de forma evidente que el material ‘1008595’ presenta gran cantidad de faltantes.

En el mapa de calor (Ilustración 8), es factible evidenciar correlaciones superiores a 0.6, 0.7 o incluso 0.8, siendo esto un posible indicio de similitud entre las series de tiempo. También se observan correlaciones negativas, como señal de productos totalmente contrarios en su patrón de consumo o también de canibalización en las ventas.

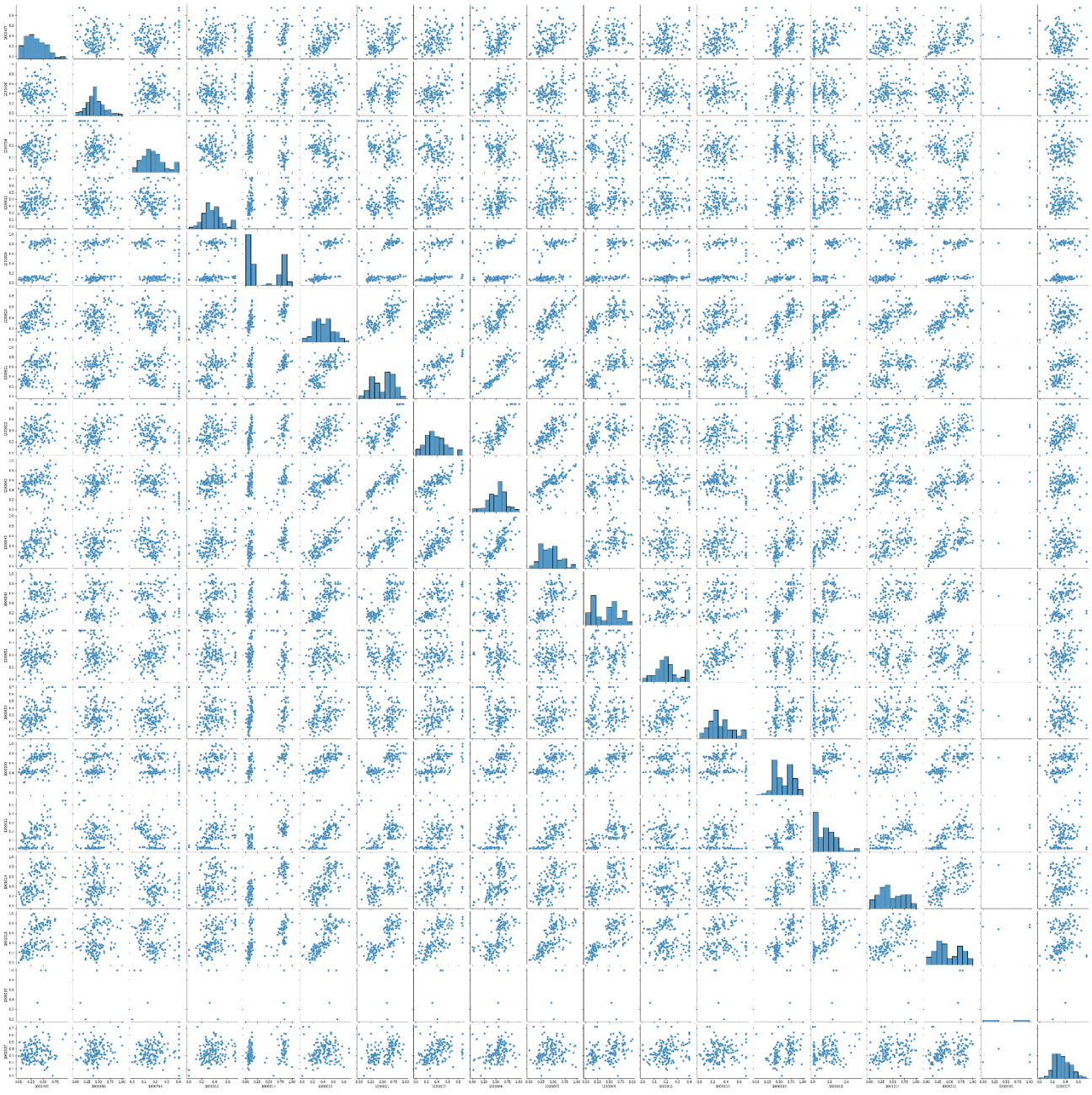

Ilustración 7. Matriz de Correlación.

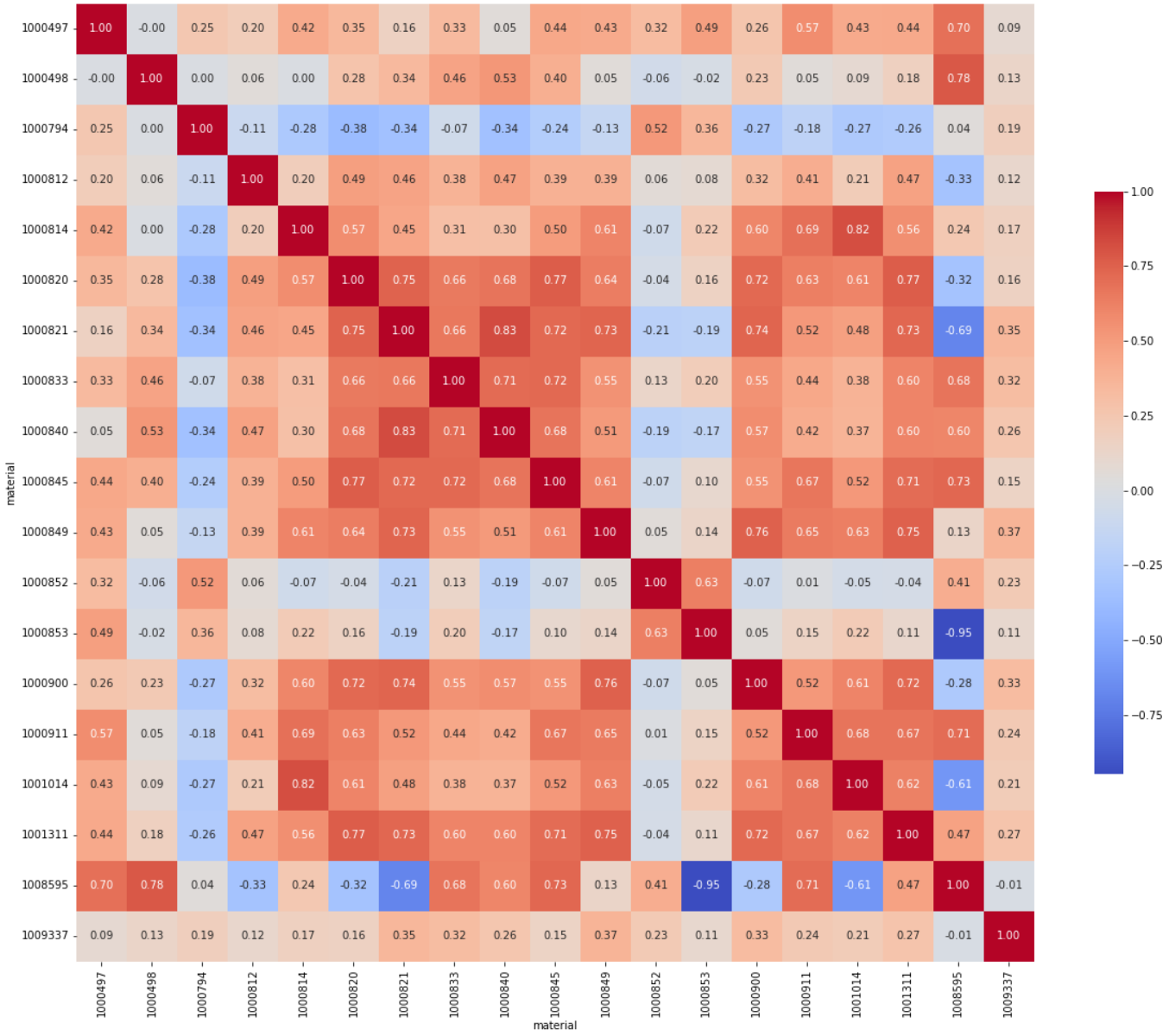

Ilustración 8.Mapa de Calor de las correlaciones de los materiales.

Para la identificación de datos faltantes, se realizó inicialmente un mapa de calor donde se visualizan las franjas de datos faltantes por material.

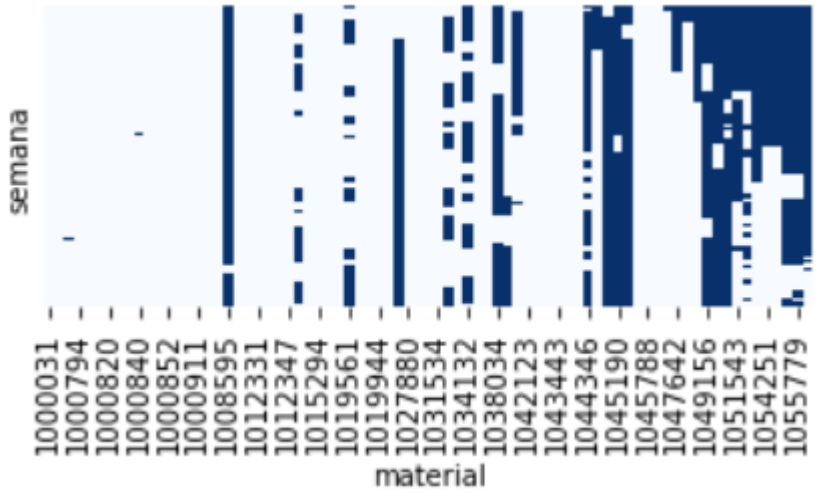

Ilustración 9.Mapa de Calor franjas de datos faltantes por periodo

Del dataframe de materiales agrupados, al menos 31 de estos poseen un dato vacío (40%) y la cantidad de materiales que poseen más del 30% de los periodos vacíos son 24 (31%). Dichos materiales son los siguientes:

['1008595', '1045189', '1045313', '1045190', '1050702', '1050413', '1027839', '1061857', '1049156', '1055779', '1038034', '1044346', '1055778', '1051543', '1034132', '1053497', '1051544', '1032538', '1013553', '1054252', '1054251', '1019561', '1041047', '1048519']

Este mismo análisis se realizó para la combinación Centro-Material, donde se encontró que al menos 378 (54,46%) combinaciones tienen un periodo vacío y la cantidad de combinaciones que poseen más del 30% de los periodos vacíos es de 225 (32,42%).

Al realizar la depuración, se pasa de tener 77 materiales a 53, de los cuales 7 poseen periodos vacíos. Dichos materiales fueron imputados bajo el método de los vecinos más cercanos o K-Nearest Neighbors (KNN). Este algoritmo se basa en la idea de que los objetos que son similares se encuentran cerca unos de otros en el espacio de características. En otras palabras, KNN busca los K objetos más cercanos en la serie de tiempo, y utiliza la mayoría de los consumos más cercanos para predecir el dato faltante.[3]

Para este caso, el método completo los valores faltantes implementado la distancia euclidiana a los vecinos más cercanos e imputando por medio de un promedio uniforme.

III. MODELAMIENTO

A. Pipeline

Las fases por implementar en este proyecto se ajustarán a la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Estándar utilizado en proyectos de minería y análisis de datos, ampliamente utilizada en la industria y en la academia [4]. Consta de 6 pasos los cuales se ejemplifican al desarrollo a continuación:

1. Entendimiento del problema: Etapa en la cual se debe definir el problema a resolver, identificar objetivos y establecer preguntas claves. Etapa realizada con el equipo de planeación de la demanda del área.
2. Recopilación de datos: Identificar las fuentes de datos necesarias, en la cual se debe asegurar que los datos sean verídicos, se encuentren limpios y estructurados de forma adecuada. Para el proyecto la fuente de información fue el ERP SAP.
3. Preparación de datos: Se realizan transformaciones a los datos, incluyendo eliminación de valores atípicos, imputación de valores faltantes y la normalización de estos. Se extraen conclusiones del análisis exploratorio de los datos para su posterior modelado.
4. Modelado: Selección de técnicas de modelado de datos para responder a los objetivos del proyecto definidos en la primera etapa del proyecto. En este caso las técnicas usadas estarán asociadas al aprendizaje no supervisado, específicamente al clustering de datos.
5. Evaluación: Se busca medir los resultados de los experimentos realizados en el apartado de modelación, para la posterior validación en contraste con las preguntas de negocio asociadas.
6. Despliegue: Esta etapa busca presentar resultados y conclusiones a los stakeholders del proyecto, al igual que implementar técnicas para el consumo y uso continuo del modelo. Para el caso del proyecto, no se llegará hasta este ítem.

B. Preprocesamiento

Inicialmente, se aplica un suavizado a las series de tiempo por medio de la aplicación de una media móvil con ventana de dos periodos, es decir se pasa de tener 155 características a 154. Esto con el fin de promediar picos extraordinarios entre periodos.

Luego, para la normalización de la data, se aplican las metodologías de Z-Score Normalization y MinMax Normalization, con el fin de usar los datasets arrojados por cada transformación en los algoritmos de clustering y definir cual tiene mejor desempeño dependiendo del método usado.

C. Aproximación al algoritmo Dynamic Time Warping (DTW):

Este método es usado para encontrar la alineación óptima entre dos series de tiempo, encontrando la secuencia de pares de puntos de tiempo correspondientes que minimiza la distancia total entre ambas series.

Para ejemplificar el algoritmo DTW usaremos a manera de ejemplo los materiales '1031534' y '1032957'.

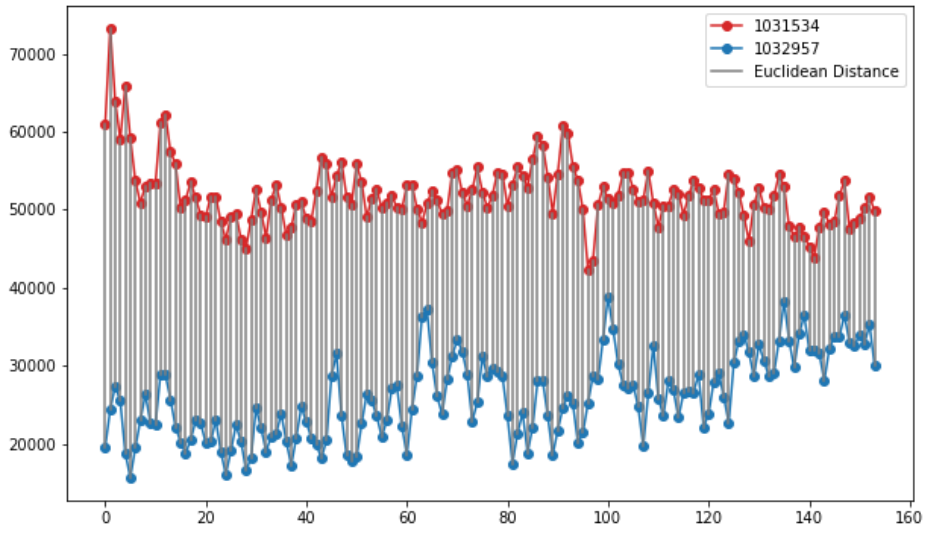

Ilustración 10. Distancia Euclidiana entre dos series de tiempo

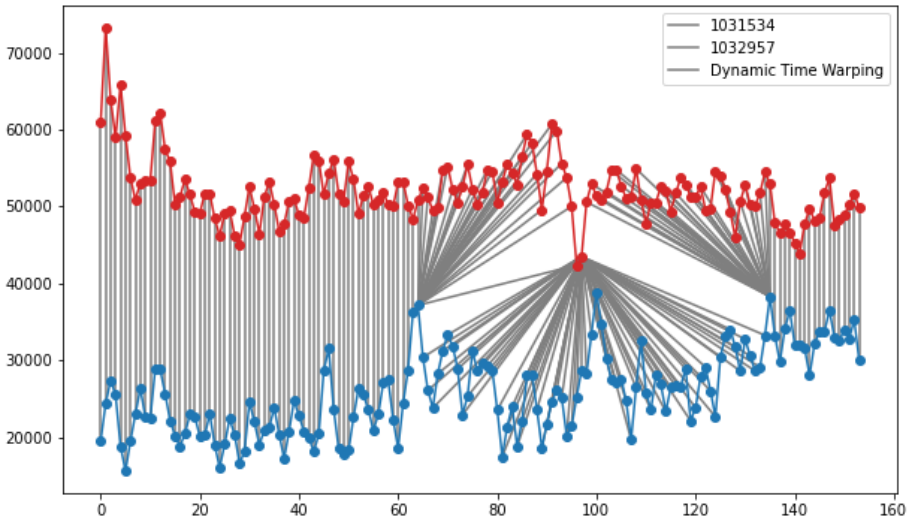

Ilustración 11. DTW entre dos series de tiempo

En la Ilustración 10 encontramos la similitud entre dos series de tiempo utilizando la distancia euclidiana y en la Ilustración 11 utilizando la medida de deformación dinámica del tiempo DTW. Ambos arrojan un valor de la suma de las distancias coincidentes, pero si se analiza a detalle la similitud generada con DTW, allí se encuentran coincidencias con los patrones distintivos de la serie temporal, lo que probablemente resulte en una evaluación de similitud más sólida que cuando es usada la distancia euclidiana.[5]

DTW inicia con la creación de una matriz de distancias entre cada par de puntos, para luego encontrar la ruta de alineación óptima que minimiza la distancia total. La matriz es completada de manera iterativa, calculando las distancias entre cada punto, para más tarde definir qué puntos de cada serie de tiempo corresponden entre sí.

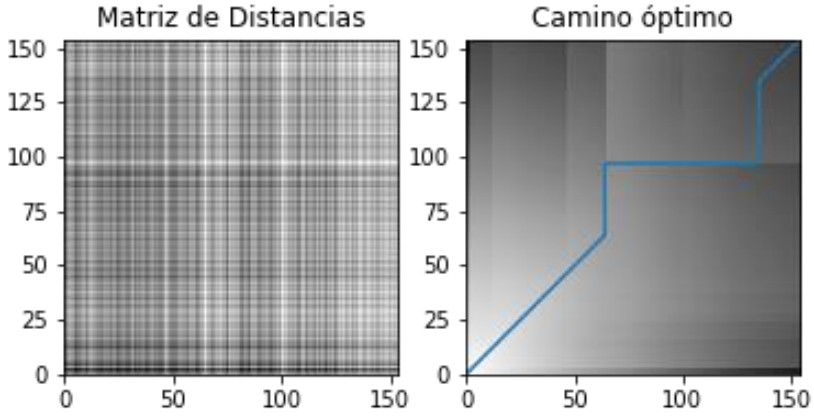

Ilustración 12. Matriz de distancias y camino óptimo entre las series de tiempo de los materiales

El problema de optimización se puede resumir en la siguiente ecuación:

$$DTW(x,y) = \min \pi \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}$$

Donde $\pi = [\pi_0, \dots\dots, \pi_K]$ son los caminos posibles entre la serie de tiempo x y y, para cada i y j. Se podría concluir, que DTW es equivalente a minimizar la distancia euclidiana entre series de tiempo alienadas bajo todas las alineaciones temporales admisibles. La Ilustración 13 exhibe el camino óptimo para un par de series de tiempo a través de la matriz de similitud que almacena los valores $d(x_i, y_j)$. [6]

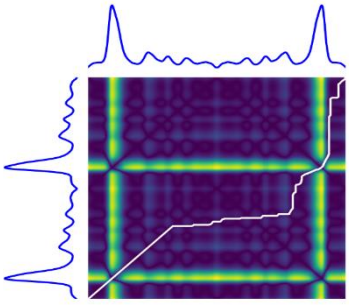

Ilustración 13. Camino óptimo entre dos series DTW

Dentro de la familia de estos algoritmos, también se encuentra el Soft Dynamic Time Warping (SDTW). Este es una variante del DTW que permite incorporar una medida de incertidumbre o variabilidad a los datos de entrada. La matriz de costo utilizada en SDTW, usa una función de costo suave, modelada con una distribución Gaussiana, de esta manera no busca una única ruta óptima cómo lo hace DTW, sino que considera múltiples rutas posibles, permitiendo una comparación más flexible y robusta de series de tiempo que presentan variabilidad temporal, generando así mayor tolerancia al ruido. [7]

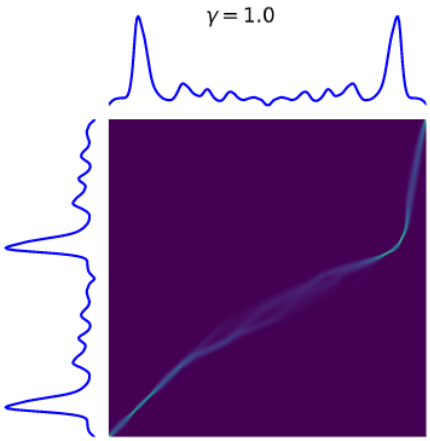

Ilustración 14. Camino óptimo entre dos series SDTW

D. Modelación:

Para la modelación se usará Tslearn, biblioteca de aprendizaje automático de código abierto diseñada específicamente para el análisis y la manipulación de series de tiempo [8], para calcular los cluster por medio de K-means usando las distancias DTW, Soft-DTW y Euclidiana. Los resultados se compararán por medio del valor de la silueta.

Se ejecutarán los algoritmos en paralelo usando los conjuntos escalados previamente bajo las técnicas mencionadas, y así encontrar la combinación de mayor rendimiento. Cada experimento evaluara la posibilidad de realizar clusters en un rango desde 2 hasta 9.

Experimentos 1:

Se uso la distancia DTW, con un tiempo promedio aproximado de generación de cluster y su respectivo valor de la silueta de 2 segundos, para ambos conjuntos de datos escalados.

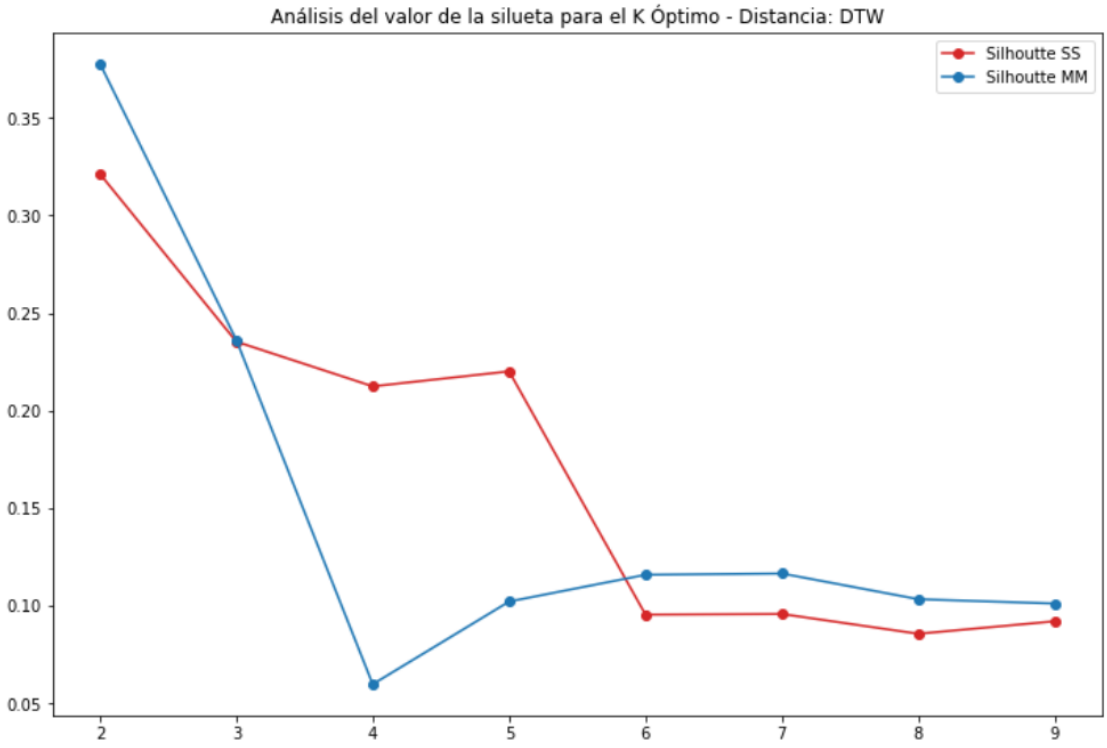

Ilustración 15. Valor de la silueta Distancia DTW

Experimentos 2:

Se uso la distancia SDTW, con un tiempo promedio aproximado de generación de cluster y su respectivo valor de la silueta de 22 segundos para el conjunto de datos escalado bajo el StandardScaler, y con un tiempo promedio aproximado de generación de cluster y su respectivo valor de la silueta de 12 segundos para el conjunto de datos escalado bajo MinMaxScaler.

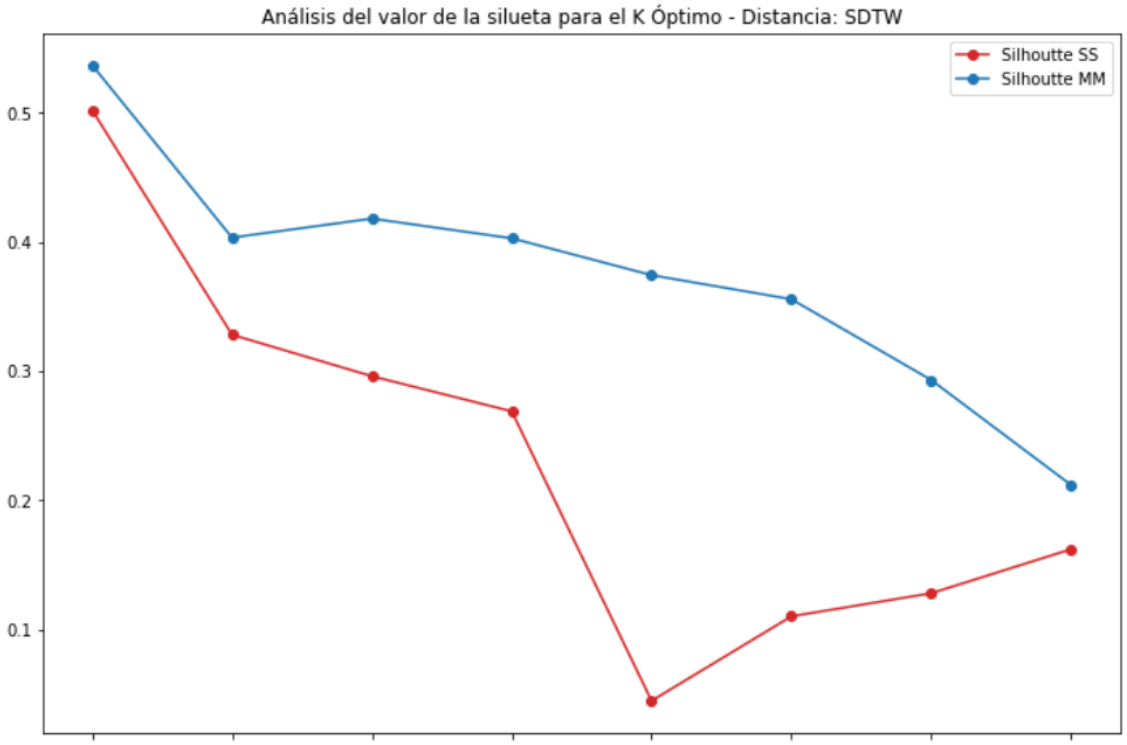

Ilustración 16. Valor de la silueta Distancia SDTW

Experimentos 3:

Se uso la distancia Euclidiana, con un tiempo promedio aproximado de generación de cluster y su respectivo valor de la silueta de 0.025 segundos para ambos conjuntos de datos escalados.

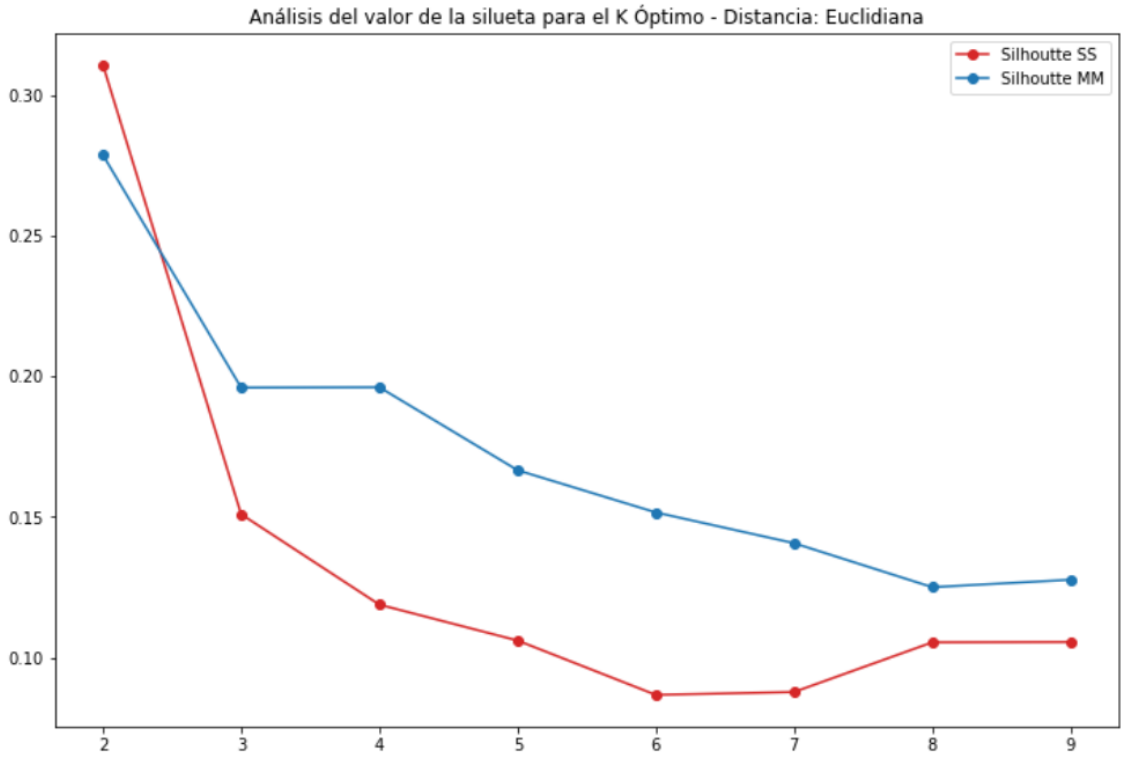

Ilustración 17. Valor de la silueta Distancia Euclidiana

IV. RESULTADOS

Al evaluar los resultados de los tres experimentos, se encuentra que, para todos los casos, el mayor valor de la silueta es para k igual a 2.

Ahora bien, para el experimento 1 (DTW) el valor máximo de silueta alcanzado fue aproximadamente 0.38 para el conjunto de datos MinMaxScaler con k igual 2. Para los demás k el valor de la silueta es por debajo de 0.25 y como observación relevante con k igual a 4 y 5, el desempeño de los conjuntos de datos StandardScaler que el conjunto MinMaxScaler.

En el experimento 2 (SDTW) se observa el mejor desempeño, con un valor de silueta máximo para k igual a 2 de 0.53. En este caso los datos MinMaxScaler siempre están por encima del conjunto StandardScaler. Un análisis enfocado en responder la pregunta de negocio se podría resolver con este modelo ya que el segundo mejor valor de silueta es para k igual 4 (0.41), siendo esta segmentación más atractiva que 2 clúster.

Finalmente, la puntuación máxima más baja se encuentra en el experimento 3 (Euclidiana), con un valor de 0.32 para k igual a 2 con el conjunto de datos StandardScaler. Los demás valores al ser tan bajos no son atractivos para el análisis.

Se decide correr el algoritmo K-means con la técnica de normalización y los parámetros que más alto tengan el valor de la silueta, teniendo en cuenta las preguntas de negocio se desean responder.

Por lo anterior, la distancia elegida fue SDTW con un número de clúster igual a 4, obteniendo los siguientes resultados.

Clúster 1:

El 33.96% de las observaciones se encuentran este segmento. Las características que comparten los productos que hacen parte de esta agrupación es que no solamente poseen una tendencia positiva clara, sino que el incremento luego del periodo 60 fue de manera aditiva, dado que antes el consumo reportado era bajo.

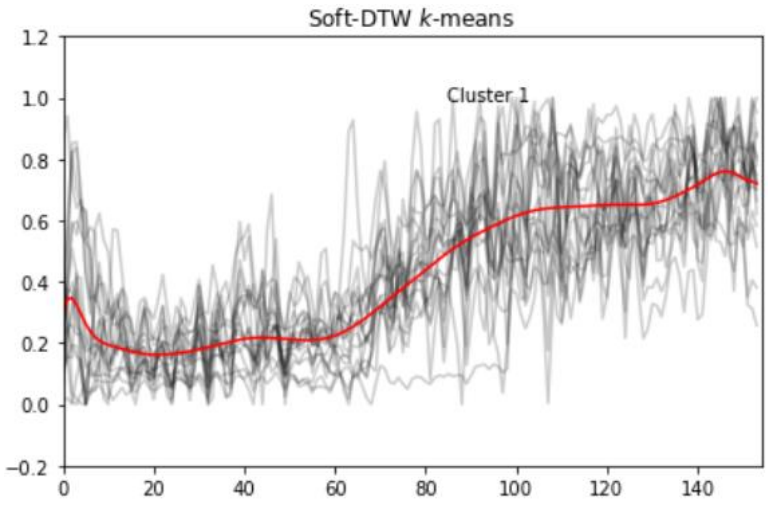

Ilustración 18. Resultado Clúster 1

Clúster 2:

El 24.53% de las observaciones se encuentran este segmento. Las características que comparten los productos que hacen parte de esta agrupación es que poseen una tendencia negativa. Los primeros periodos tienen consumos altos, pero luego disminuyen abruptamente. Algunos de estos muestran señales de crecimiento otra vez, pero la tendencia se conserva.

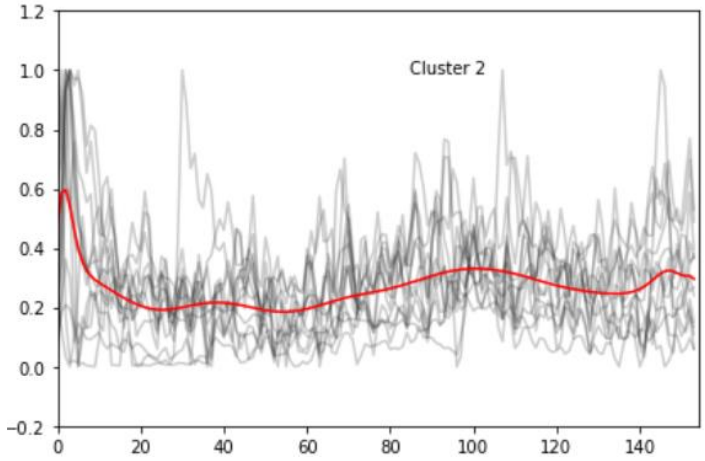

Ilustración 19. Resultado Clúster 2

Clúster 3:

El 28.30% de las observaciones se encuentran este segmento. Las características que comparten los productos que hacen parte de esta agrupación son similares al clúster 1, pero la principal diferencia es que su tendencia positiva ya estaba marcada desde antes del periodo 60, es decir los productos allí ya habían sufrido de un incremento aditivo.

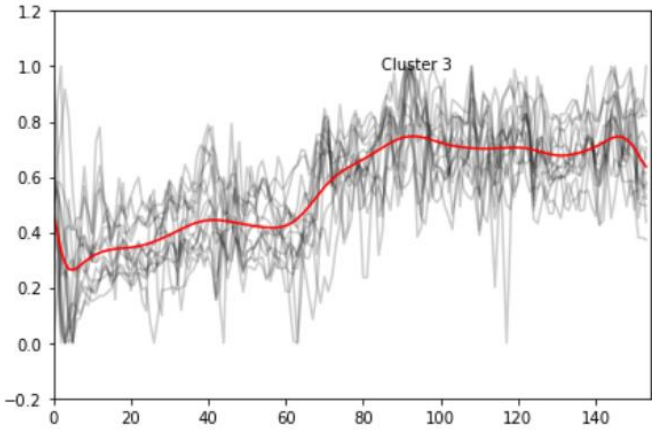

Ilustración 20. Resultado Clúster 3

Clúster 4:

El 13.21% de las observaciones se encuentran este segmento. Las características que comparten los productos que hacen parte de esta agrupación se pueden observar de manera clara, dado que comparten un patrón estacional aproximadamente cada 45 - 50 periodos. No se refleja alguna tendencia.

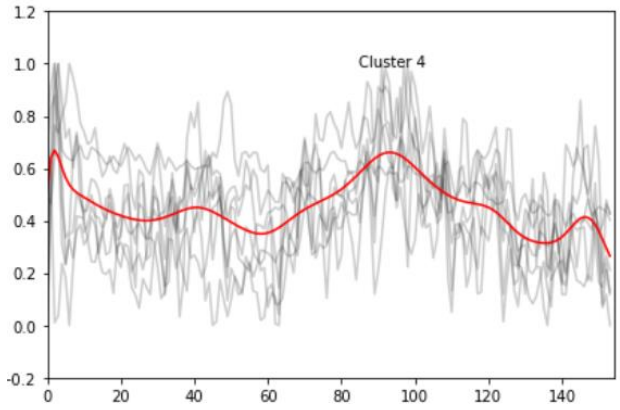

Ilustración 21. Resultado Clúster 4

V. EVALUACIÓN

De acuerdo con el objetivo del proyecto se alcanzó un resultado satisfactorio describiendo una metodología que permita segmentar los productos por su comportamiento temporal.

El equipo de planeación de la demanda puede aprovechar el modelo para extraer insights relevantes de sus productos, y así entender tendencias que a lo largo del tiempo su consumo establece, los cuales no son exclusivas de una referencia, marca, línea, volumen de venta, sino que por medio de este clustering se logran encontrar relaciones que la gestión del portafolio en el día a día se pierden.

Ahora bien, el equipo de modelación de la cadena de suministro puede hacer uso de este clustering por componentes, para desarrollar un modelo de clasificación, el cual sirva de triage para modelo de pronóstico actual, ya que se entenderían los patrones que posee cada serie temporal, actividad que es difícil realizar de manera masiva con el conocimiento actual.

V. CONCLUSIONES

Con la información suministrada se logra construir y diseñar un pipeline que permita la agrupación por patrones del comportamiento temporal de los productos, abarcando desde la extracción, limpieza, imputación de datos faltantes y atípicos hasta la realización de un modelo analítico que responda las preguntas de negocio.

Al momento de clustering es de suma importancia el conocimiento de la naturaleza del problema y que hiperparámetros de distancia se desea aplicar, dado que como se expuso, los más comunes no son los más eficaces en todos los casos.

La aplicación de metodologías usadas en otras áreas del conocimiento como la física de ondas pueden ser totalmente aplicados a industrias económicas tradicionales bajo el contexto adecuado, así es el caso de la matemática que sostiene los algoritmos DTW y SDTW

El algoritmo con mejor desempeño fue el SDTW para las 4 segmentaciones, atendiendo a la métrica de machine learning y las expectativas de negocio. No obstante, se podría realizar un análisis con más información y de manera continua, de esta manera mantener actualizados los insight generados por medio de analítica.

La abstracción que puede adquirir el negocio gracias al modelo desarrollado impacta directamente al entendimiento de sus productos, dado que al realizar cluster de los mismos, se pueden generalizar estrategias ya sean comerciales o de modelación de pronósticos lo que conlleva a reducir el tiempo de generación de información y aumentar la eficacia al tomar decisiones.

Vale la pena explorar la segmentación de datos, no solo teniendo como insumo las características correspondientes a los consumos en cada periodo de tiempo, sino también información como estadísticos básicos y variaciones entre periodos.

BIBLIOGRAFÍA

[1] A. Müller and S. Guido, Introduction to Machine Learning with Python, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2016.

[2] C. K. Enders, Applied missing data analysis. New York: Guilford Press, 2010.

[3] H. Ghaleb, M. A. El-Nasr and M. El-Bakry, "A Review of a Text Classification Technique: K-Nearest Neighbor,", 2016

[4] S. C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," in Proc. Computer Science, vol. 181, pp. 526-534, 2021, doi: 10.1016/j.procs.2021.01.199.

[5] "Dynamic Time Warping (DTW) - tslearn 0.5.0 documentation," tslearn.readthedocs.io. [En línea]. Disponible en: https://tslearn.readthedocs.io/en/stable/user_guide/dtw.html.

[6] C. Tralie and E. Dempsey, "Exact, parallelizable dynamic time warping alignment with linear memory," arXiv:2008.02734 [cs.SD], 2020. [En línea]. Disponible en: https://arxiv.org/abs/2008.02734.

[7] M. Cuturi and M. Blondel, "Soft-DTW: A Differentiable Loss Function for Time-Series," in Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, August 6-11, 2017, pp. 894-903. DOI: 10.1007/978-3-030-81596-8_10. [En línea]. Disponible en: https://arxiv.org/pdf/1703.01541.pdf

[8] R. Fauvel, M. Yger, A. Roy, and N. Courty, "Tslearn: A Machine Learning Toolkit for Time Series Data," Journal of Machine Learning Research, vol. 22, pp. 1-6, 2023.