NOMBRE DEL TRABAJO

ROQUE GONZALES_TESIS_FINAL.pdf

AUTOR

elvis joel roque gonzales

RECUENTO DE PALABRAS

19123 Words

RECUENTO DE CARACTERES

103291 Characters

RECUENTO DE PÁGINAS

114 Pages

TAMAÑO DEL ARCHIVO

4.9MB

FECHA DE ENTREGA

Nov 21, 2023 5:03 AM GMT-5

FECHA DEL INFORME

Nov 21, 2023 5:05 AM GMT-5

16% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base d

- 15% Base de datos de Internet
- 6% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossr
- 0% Base de datos de trabajos entregados

UNTELS | UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR

FORMULARIO DE AUTORIZACIÓN PARA LA PUBLICACIÓN DE TRABAJOS DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL DE LA UNTELS

(Art. 45° de la ley N° 30220 – Ley)

Autorización de la propiedad intelectual del autor para la publicación de tesis en el Repositorio Institucional de la Universidad Nacional Tecnológica de Lima Sur (https://repositorio.untels.edu.pe), de conformidad con el Decreto Legislativo N° 822, sobre la Ley de los Derechos de Autor, Ley N° 30035 del Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, Art. 10° del Rgto. Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales en las universidades – RENATI Res. N° 084-2022-SUNEDU/CD, publicado en El Peruano el 16 de agosto de 2022; y la RCO N° 061-2023-UNTELS del 01 marzo 2023.

TIPO DE TRABAJO DE INVESTIGACIÓN

1). TESIS (x) 2). TRABAJO DE SUFICIENCIA PROFESIONAL ( )

DATOS PERSONALES

Apellidos y Nombres:	ROQUE GONZALES, ELVIS JOEL
D.N.I.:	47688127
Otro Documento:	
Nacionalidad:	PERUANO
Teléfono:	985373707
e-mail:	ELVIS.ROQUE01@GMAIL.COM

DATOS ACADÉMICOS

Pregrado

Facultad:	FACULTAD DE INGENIERÍA Y GESTIÓN
Programa Académico:	75515
Título Profesional otorgado:	INGENIERO ELECTRÓNICO Y TELECOMUNICACIONES

Postgrado

Universidad de Procedencia:	
País:	
Grado Académico otorgado:	

Datos de trabajo de investigación

Título:	"APLICACIÓN DE ALGORITMOS DE AGRUPAMIENTO EN APRENDIZAJE DE MÁQUINA NO SUPERVISADO PARA LA IDENTIFICACIÓN DE PATRONES DE TRÁFICO DE CELDAS LTE EN LA RED NACIONAL MÓVIL DE LA OPERADORA WOM DE CHILE"
Fecha de Sustentación:	04 DICIEMBRE 2023
Calificación:	APROBADO CON DISTINCIÓN
Año de Publicación: 2025	

UNTELS | UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR

AUTORIZACIÓN DE PUBLICACIÓN EN VERSIÓN ELECTRÓNICA

A través de la presente, autorizo la publicación del texto completo de la tesis, en el Repositorio Institucional de la UNTELS especificando los siguientes términos:

Marcar con una X su elección.

1) Usted otorga una licencia especial para publicación de obras en el REPOSITORIO INSTITUCIONAL DE LA UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR.

Si autorizo X No autorizo _____

2) Usted autoriza para que la obra sea puesta a disposición del público conservando los derechos de autor y para ello se elige el siguiente tipo de acceso.

Derechos de autor		
TIPO DE ACCESO	ATRIBUCIONES DE ACCESO	ELECCIÓN
ACCESO ABIERTO 12.1(*)	**info:eu-repo/semantics/openAccess** (Para documentos en acceso abierto)	(X)

3) Si usted dispone de una **PATENTE** puede elegir el tipo de **ACCESO RESTRINGIDO** como derecho de autor y en el marco de confiabilidad dispuesto por los numerales 5.2 y 6.7 de la directiva Nº 004-2016-CONCYTEC DEGC que regula el Repositorio Nacional Digital de CONCYTEC (Se colgará únicamente datos del autor y el resumen del trabajo de investigación).

Derechos de autor		
TIPO DE ACCESO	ATRIBUCIONES DE ACCESO	ELECCIÓN
ACCESO RESTRINGIDO	**info:eu-repo/semantics/restrictedAccess** (Para documentos restringidos)	( )
	info:eu-repo/semantics/embargoedAccess (Para documentos con períodos de embargo. Se debe especificar las fechas de embargo)	( )
	info:eu-repo/semantics/closedAccess (para documentos confidenciales)	( )

(*) http://renati.sunedu.gob.pe

UNTELS | UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR

Rellene la siguiente información si su trabajo de investigación es de acceso restringido:

Atribuciones de acceso restringido:

Motivos de la elección del acceso restringido:

ROBLES GONZALES, ELVIS JOEL

APELLIDOS Y NOMBRES

47768127

DNI

Firma y huella:

Lima, 9 de ENERO del 2025

UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR

FACULTAD DE INGENIERÍA Y GESTIÓN
ESCUELA PROFESIONAL DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES

“APLICACIÓN DE ALGORITMOS DE AGRUPAMIENTO EN APRENDIZAJE DE MÁQUINA NO SUPERVISADO PARA LA IDENTIFICACIÓN DE PATRONES DE TRÁFICO DE CELDAS LTE EN LA RED NACIONAL MÓVIL DE LA OPERADORA WOM DE CHILE”

TESIS
Para optar el Título Profesional de

INGENIERO ELECTRÓNICO Y TELECOMUNICACIONES

PRESENTADO POR EL BACHILLER

ROQUE GONZALES, ELVIS JOEL
ORCID: 0009-0008-7446-9852

ASESOR
CARTAGENA GORDILLO, ALEX
ORCID: 0000-0001-8076-0699

Villa El Salvador
2023

UNIVERSIDAD NACIONAL
TECNOLÓGICA DE LIMA SUR

"Año de la unidad, la paz y el desarrollo"

DECANATO DE LA FACULTAD DE INGENIERÍA Y GESTIÓN

ACTA DE SUSTENTACIÓN DE TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE INGENIERO ELECTRÓNICO Y TELECOMUNICACIONES

En Villa El Salvador, siendo las 13:20 horas del día 4 de diciembre de 2023, en la Facultad de Ingeniería y Gestión, los miembros del Jurado Evaluador, integrado por:

PRESIDENTE: DR. ORLANDO ADRIAN ORTEGA GALICIO	DNI N° 20032665	C.I.P. N° 79878
SECRETARIO: MG. MAX FREDI QUISPE AGUILAR	DNI N° 41618736	C.I.P. N° 138642
VOCAL : MG. JORGE LUIS LÓPEZ CÓRDOVA	DNI N° 09638009	C.I.P. N° 183016
ASESOR : DR. ALEX CARTAGENA GORDILLO	DNI N° 29420194	C.I.P N° 133344

Designados mediante Resolución de Decanato N° 343-2023-UNTELS-R-D de fecha 15 de agosto de 2023 quienes dan inicio a la Sesión Pública de Sustentación y Evaluación de Tesis.

Acto seguido, el aspirante al : Grado de Bachiller ☐ Título Profesional [x]

Don : ELVIS JOEL ROQUE GONZALES identificado con D.N.I. N° 47768127 procedió a la Sustentación de:

Trabajo de investigación ☐ Tesis [x] Trabajo de suficiencia ☐ Artículo científico ☐

Titulada: **"APLICACIÓN DE ALGORITMOS DE AGRUPAMIENTO EN APRENDIZAJE DE MÁQUINA NO SUPERVISADO PARA LA IDENTIFICACIÓN DE PATRONES DE TRÁFICO DE CELDAS LTE EN LA RED NACIONAL MÓVIL DE LA OPERADORA WOM DE CHILE "**.

Aprobado mediante Resolución de Decanato N° 958-2023-UNTELS-R-D de fecha 27 de noviembre de 2023, de conformidad con las disposiciones del Reglamento General de Grados Académicos y Títulos Profesionales vigentes, sustentó y absolvió las interrogantes que le formularon los señores miembros del Jurado Evaluador.

Concluida la Sustentación se procedió a la evaluación y calificación correspondiente, resultando el aspirante APROBADO por ...Distinción... con la nota de: ...Dieciocho... (letras) ...18... (números), de acuerdo al Art. 65° del Reglamento General para optar el Título Profesional.

CALIFICACIÓN		CONDICIÓN	EQUIVALENCIA
NÚMERO	LETRAS		
18	Dieciocho	Aprobado con Distinción	Muy Bueno

Siendo las 14:10 horas del día 4 de diciembre de 2023, se dio por concluido el acto de sustentación, firmando el jurado evaluador el Acta de Sustentación, que obra en el Decanato de la Facultad de Ingeniería y Gestión.

DR. ORLANDO ADRIAN ORTEGA GALICIO
PRESIDENTE

MG. MAX FREDI QUISPE AGUILAR
SECRETARIO

MG. JORGE LUIS LÓPEZ CÓRDOVA
VOCAL

ELVIS JOEL ROQUE GONZALES
BACHILLER

DEDICATORIA

A mi querida familia, por su apoyo incondicional

AGRADECIMIENTOS

Quiero expresar mi profundo agradecimiento a la operadora móvil de Telecomunicaciones WOM,
por brindarme acceso a su valiosa información, lo cual hace posible el desarrollo de la presente tesis.

Agradezco especialmente a Nicolás Carrasco del equipo de analítica de WOM por su orientación experta en diversos aspectos de la investigación. También a Richard Silva del área de despliegue de red, quien generosamente me proporcionó el tiempo necesario para continuar explorando nuevos puntos de mejora.

Mi reconocimiento se extiende al doctor Alex Cartagena por su invaluable asesoramiento en el diseño de los objetivos y la adecuada verificación de los resultados obtenidos durante este estudio.

No puedo pasar por alto el apoyo inquebrantable de mis padres, quienes me alentaron y me respaldaron en los momentos que más los necesitaba.

Asimismo, agradezco a todas las demás personas que estuvieron siempre a mi lado, brindándome aliento y consejos invaluables.

RESUMEN

En la presente tesis, titulada "Aplicación de Algoritmos de Agrupamiento en Aprendizaje de Máquina No Supervisado para la Identificación de Patrones de Tráfico de Celdas LTE en la Red Nacional Móvil de la Operadora WOM de Chile", el autor presentó un enfoque analítico en el procesamiento, análisis y clasificación de la tendencia de la variable "volumen de descarga de datos" de las celdas LTE operativas de la red WOM basado en el uso de algoritmos de aprendizaje automático no supervisado.

La operadora de telecomunicaciones WOM recopiló mediciones de forma horaria a través de las antenas 4G y gestionó los datos mediante la plataforma iMaster MAE, un gestor de redes de acceso. Posteriormente, la operadora procesó los datos utilizando la plataforma PRS de Huawei. Una vez completada esta recopilación, tabuló y proceso los datos, almacenándolos en una base de datos relacional llamada Clickhouse. Luego, el autor extrajo y transformó los datos almacenados en la base de datos utilizando el lenguaje de programación SQL. Además, abordó los datos faltantes mediante algoritmos de interpolación polinómica, y para su análisis, agrupó lo datos de manera semanal, calculando la suma total de la descarga de datos en intervalos de 7 días.

En cuanto a la agrupación de las celdas LTE, el autor comparó dos métodos principales: K-Means y mapas auto organizados (SOM). Tras esta evaluación, concluyó que el método de mapas autoorganizados agrupó de manera más efectiva la tendencia del volumen de descarga de datos de las celdas. Finalmente, presentó los resultados finales y demostró que la tendencia del volumen de descarga de datos se pudo agrupar en un total de 737 grupos.

Por último, para clasificar las celdas, el autor utilizó el método STL, que descompuso las series de tiempo de las tendencias en sus componentes principales. Luego, analizó la tendencia utilizando el método del promedio de media móvil y obtuvo los valores de la tendencia semanal, su magnitud y su dirección. Esto permitió al autor determinar los periodos de tiempo en los que las celdas tienden a tener un mayor volumen de datos y el tipo de patrón que sigue la serie de tiempo.

Palabras clave: K-Means, SOM, Aprendizaje automático, LTE, agrupamiento, base de datos, series de tiempo, tendencia, estacionalidad

ABSTRACT

In the present thesis, titled "Application of Unsupervised Machine Learning Clustering Algorithms for Identifying Traffic Patterns of LTE Cells in the National Mobile Network of Chile's WOM Operator," the author presented an analytical approach to processing, analyzing, and classifying the trend of the variable "data download volume" of all operational LTE cells in the WOM network based on the use of unsupervised machine learning algorithms.

The telecommunications operator WOM collected hourly measurements through 4G antennas and managed the data using the iMaster MAE platform, an access network manager. Subsequently, the operator processed the data using Huawei's PRS platform. Once this data collection was completed, it tabulated and processed the data, storing it in a relational database called Clickhouse. The author then extracted and transformed the data stored in the database using the SQL programming language. In addition, missing data were addressed using polynomial interpolation algorithms, and for analysis purposes, the data was grouped on a weekly basis, calculating the total data download in 7-day intervals.

Regarding the clustering of LTE cells, the author compared two main methods: K-Means and self-organizing maps (SOM). Following this evaluation, it was concluded that the K-Means method effectively clustered the trend of data download volume from the cells. Finally, the author presented the results and demonstrated that the trend of data download volume could be grouped into a total of 50 clusters.

Lastly, to classify the cells, the author used the STL method, which decomposed the time series of trends into their principal components. Subsequently, the trend was analyzed using the moving average method to obtain weekly trend values, their magnitude, and direction. This allowed the author to determine the periods during which cells tended to have higher data volumes and the type of pattern followed by the time series.

Keywords: K-Means, SOM, Machine Learning, LTE, clustering, database, time series, trend, seasonality

INDICE

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Comprender el rendimiento de la red móvil es fundamental para una operadora de telecomunicaciones, ya que esto les permite comparar lo planificado con los resultados reales. La experiencia nos indica que una planificación deficiente puede resultar en problemas como caídas de llamadas, congestión de la red y altos niveles de interferencia, lo que finalmente se traduce en una mala experiencia para el usuario.

La operadora móvil de telecomunicaciones WOM, con estaciones celulares 3G, 4G y 5G distribuidas en todo Chile, brinda a sus usuarios la posibilidad de comunicarse tanto mediante voz como a través de servicios que requieren acceso a Internet, todo ello cumpliendo con rigurosos estándares de calidad de servicio. Para lograrlo, WOM cuenta con diversas áreas encargadas de supervisar y mejorar el rendimiento de la red. En el proceso de proyección del volumen futuro de datos de una estación LTE, el área de planificación de red de WOM utiliza modelos matemáticos basados en datos pasados de celdas cuyas características se asemejan a la celda objetivo. Sin embargo, en algunos casos, el comportamiento futuro supera las proyecciones iniciales, lo que puede afectar la experiencia del usuario.

Muchos de estos problemas de rendimiento de red no se detectan de inmediato debido a la complejidad de analizar un gran volumen de datos provenientes de las mediciones realizadas por las antenas LTE en cada hora. Además, modelar los datos presenta desafíos, ya que incluyen características de series de tiempo con componentes de tendencia y estacionalidad que son difíciles de identificar mediante visualización gráfica. Los algoritmos de aprendizaje automático se presentan como una opción confiable y de bajo costo computacional. En el marco de esta tesis, se han utilizado dos algoritmos de agrupamiento no supervisado en aprendizaje automático para identificar patrones y tendencias inherentes a la red LTE de la operadora WOM.

El Capítulo I abordó el planteamiento del problema, los objetivos, el alcance y la justificación de la tesis. En el Capítulo II, se desarrolló el marco teórico, se expusieron los antecedentes de la investigación y se presentaron las bases teóricas. El Capítulo III se dedicó a la metodología, la implementación y las pruebas realizadas. Finalmente, el Capítulo IV expuso las conclusiones obtenidas.

I. PLANTEAMIENTO DEL PROBLEMA

1.1 Motivación

La presente tesis surgió de la motivación por emplear técnicas de ciencia de datos con el propósito de mejorar el análisis y rendimiento de las redes móviles de comunicaciones. Esta investigación no habría sido posible sin el resultado de años de experiencia dedicados al mantenimiento, operación y optimización de diversas tecnologías que abarcan las redes de acceso celular. Asimismo, se ha contado con la investigación en el uso de innovadoras técnicas de aprendizaje automático que posibilitan la mejora en el análisis y la maximización del rendimiento de estas redes, con el objetivo de hacerlo más accesible para las personas.

1.2 Estado del arte

El análisis del estado del arte en esta tesis abarcó dos grupos principales de investigación. El primer grupo se centró en los trabajos actuales relacionados con el uso de algoritmos de agrupamiento en el contexto del aprendizaje automático no supervisado, aplicados a series de tiempo. El segundo grupo se enfocó en el análisis del comportamiento del tráfico en las redes LTE de telecomunicaciones.

Las series de tiempo se utilizan en diversas industrias y contextos para realizar pronósticos basándose en datos históricos. Su aplicación se extiende a campos como negocios, ingeniería, economía, medicina, entre otros. Por tanto, se ha hecho hincapié en el desarrollo de modelos automáticos que extraigan información valiosa de estos datos. No obstante, las series de tiempo presentan varios desafíos, como ruido, desigualdad e intermitencia según el contexto.

Uno de los modelos de aprendizaje automático más investigados y exitosos en el análisis de series de tiempo es el aprendizaje profundo, que puede manejar la complejidad de estas series. D. Salinas et al. (2020) publicaron un artículo sobre como las redes neuronales recurrentes (RNN) consideran la secuencia natural de las series de tiempo de manera explícita mejorando la eficiencia del aprendizaje automático.

Por otro lado, surge un problema cuando las series de tiempo varían debido a factores ambientales y humanos. X. Zhong et al. (2023) publicaron un artículo donde se aborda este problema utilizando un enfoque basado en modelos DBSCAN incremental y

KNN con autoaprendizaje. Este enfoque busca identificar y ajustar automáticamente el número de grupos cuando las condiciones de trabajo cambian.

Las redes de telecomunicaciones también se benefician de estos avances, ya que los datos recopilados sobre el rendimiento de la red también siguen patrones de series de tiempo. En los últimos años, se han desarrollado técnicas específicas para tratar los datos de las redes móviles, como la identificación de patrones de cobertura basados en la ubicación geográfica de las celdas o en la calidad de la señal y el rendimiento de la red. Además, se ha extendido el uso de técnicas de visualización de datos para ayudar a los operadores de redes a comprender mejor los patrones de cobertura de las celdas LTE.

1.3 Descripción del problema

Las redes móviles de telecomunicaciones generan volúmenes considerables de información. Basado en la información almacenada en la base de datos de la operadora WOM, por cada hora y por cada celda la plataforma de gestión de redes iMAE obtuvo la medición de 150 indicadores de rendimiento, al extender el análisis a un alcance anual se obtuvo que por cada celda se tuvo aproximadamente 1’314’000 mediciones en solo un año. Si extrapolamos este análisis para abarcar todas las celdas de la red nacional de la operadora WOM en Chile, estaríamos tratando con aproximadamente 39,42 billones de registros sin contar los registros adicionales referentes a la información única de cada celda. Esto hace que sea inviable realizar un análisis utilizando métodos convencionales. En la Tabla 1 se muestra un resumen de la información almacenada en la base de datos de WOM considerando solo 30 indicadores de rendimiento en el año 2022

Tabla 1

Cantidad de información relacionado a las celdas LTE que contiene la base de datos de la operadora WOM

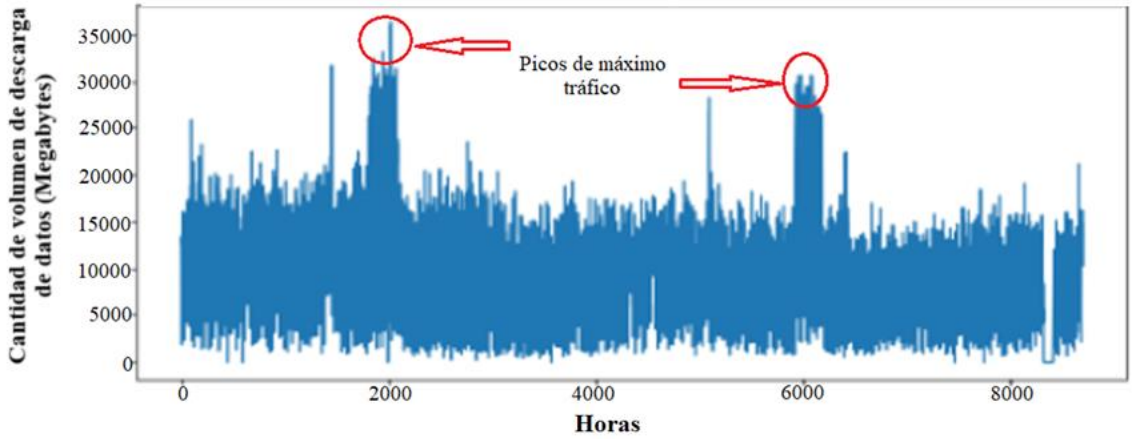
Información	Tamaño de información	Número de filas
Contadores LTE	60 Gigabytes	292505094
Referencia de celdas	4 Megabytes	39351
Lista de KPIs	11 Kilobytes	46

Nota. Elaboración propia basado en la base de datos de WOM. Esta tabla representa la cantidad total de filas que contienen datos pertinentes sobre tráfico, interferencia, cobertura y calidad de señal de la red nacional LTE de la operadora WOM para el año 2022.

Tomando el tráfico de datos de una celda como ejemplo, se puede observar una variación a lo largo del tiempo que muestra un patrón irregular, tal como se ilustra en la Figura 1. En esta figura se representa el volumen total de tráfico de descarga de datos medidos en megabytes (MB) a lo largo de 2022, en intervalos de horas. Los datos corresponden a una celda LTE ubicada en la región del Maule. Como se aprecia en la Figura 1, en ciertos momentos del día el tráfico de descarga de datos presenta picos máximos, seguidos por periodos de tráfico mínimo.

Figura 1

Medición del volumen de tráfico de descarga de datos a lo largo del 2022 para una celda LTE.



Nota. Elaboración propia basado en la base de datos de WOM. Esta figura representa el volumen total de tráfico de descarga de datos, medido en megabytes (MB), a lo largo del año 2022. Los datos corresponden a una celda LTE ubicada en la región del Maule.

La Figura 1 no revela conclusiones significativas respecto al patrón de tráfico de la celda, excepto por los picos máximos y mínimos en intervalos específicos de horas. Si extendiéramos este análisis a las más de 30,000 celdas LTE de la operadora WOM, requeriría un uso considerable de recursos tanto en términos de tiempo como de personal. Además, uno de los problemas principales relacionados con los picos máximos de cobertura es que, cuando no se planifican con anticipación, pueden ocasionar interrupciones en el servicio.

Por otra parte, las señales de telecomunicaciones tomados a lo largo del año presentan características de series de tiempo lo cual conlleva a realizar un análisis

minucioso de sus componentes lo cual añade más complejidad al análisis de la red en general.

Para abordar este problema, se recomendó utilizar técnicas de aprendizaje automático. Estos modelos permiten trabajar con grandes volúmenes de datos y llevar a cabo análisis complejos con menor demanda de recursos computacionales y humanos en comparación con los métodos tradicionales.

1.4 Formulación del problema

1.4.1. Problema general

¿Cómo lograr que la aplicación de algoritmos de aprendizaje de máquina permita identificar patrones de tráfico que poseen cada área de cobertura de las celdas LTE en la red móvil de la operadora WOM a lo largo del 2022?

1.4.2 Problemas específicos

¿Qué técnica de imputación es útil para tratar los datos faltantes del tráfico de las celdas LTE en la red móvil de la operadora WOM?

¿Qué técnica de muestreo es útil para segmentar los datos del tráfico de las celdas LTE en la red móvil de la operadora WOM?

¿Qué método de agrupamiento es útil para identificar patrones de tráfico que poseen cada área de cobertura de las celdas LTE en la red móvil de la operadora WOM a lo largo del tiempo?

¿Qué técnica de determinación de número de grupos es útil para agrupar eficazmente el tráfico de las celdas LTE en la red móvil de la operadora WOM?

¿Qué técnica de selección de datos es útil para tratar los valores atípicos del tráfico de las celdas LTE en la red móvil de la operadora WOM?

1.5. Objetivos:

1.5.1 Objetivo general

Aplicar algoritmos de agrupamiento para la identificación de patrones de tráfico de celdas LTE en la red nacional móvil de la operadora WOM de Chile

1.5.2 Objetivos específicos

- Seleccionar e identificar la técnica de interpolación óptima para tratar los datos faltantes del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de interpolación lineal y el método de interpolación polinómica.
- Seleccionar e identificar la técnica de segmentación más eficaz para separar los datos del tráfico de las celdas LTE entre la separación a nivel de horas y la separación a nivel de semanas
- Seleccionar e identificar la técnica de determinación de número de grupos más eficaz entre el método del codo y el método de silueta para agrupar el tráfico de las celdas LTE en la red móvil de la operadora WOM
- Seleccionar e identificar la técnica más adecuada para tratar los valores atípicos del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de mediana móvil y el método de distancia de la deformidad de tiempo dinámico (DTW).
- Seleccionar e identificar el método de agrupamiento más eficaz entre el método de agrupamiento de series de tiempo (K-Means) y el método de mapas autoorganizadas (SOM) para identificar patrones de tráfico que poseen las celdas LTE en la red móvil y el periodo de tiempo bajo estudio.

1.6 Justificación:

Teórica:

Facilita la adopción de nuevos enfoques para la aplicación de algoritmos de agrupamiento en el análisis de patrones de cobertura, disponibilidad y calidad del tráfico de datos en redes móviles de telecomunicaciones.

Económica:

El análisis e identificación de patrones en el rendimiento de la red permite a las operadoras móviles de telecomunicaciones reducir costos relacionados con el bajo desempeño de la red. Además, facilita la identificación de áreas que requieren una mayor asignación de recursos y de aquellas que cuentan con una capacidad excesiva.

Social:

El análisis e identificación de patrones en el rendimiento de la red posibilita la identificación de áreas geográficas de interés nacional que demandan una mayor disponibilidad de cobertura pero que no están siendo adecuadamente atendidas. Muchas de estas áreas incluyen zonas turísticas, centros hospitalarios, instituciones educativas, reservas naturales y zonas críticas debido a eventos climatológicos recurrentes.

II. MARCO TEÓRICO

2.1 Antecedentes

A continuación, se muestran algunas investigaciones realizadas previamente a esta tesis en el cual se emplearon distintas técnicas de aprendizaje automático para abordar el problema planteado:

Ordoñez (2021), evaluó distintas técnicas de aprendizaje automático para diagnosticar el rendimiento de las celdas 3G tomando como indicadores la exactitud, tasa de error, sensibilidad, especificidad y precisión en los resultados de los modelos. De acuerdo a sus resultados, los modelos más adecuados resultaron ser la técnica de Árbol de Decisión y la Red Neuronal.

Criollo T. et al. (2020), llevaron a cabo un estudio para el desarrollo de una red móvil LTE en la ciudad de Quito realizando un análisis individual de los usuarios utilizando técnicas de Machine Learning para realizar una clasificación del rendimiento final de la red..

Benavides (2021), implementó un modelo predictivo compuesto por tres bloques para detectar celdas LTE con bajo rendimiento en la velocidad de descarga de datos, utilizando un umbral de 3.3 Mbps. El primer bloque constó de un algoritmo clasificador que diferenciaba entre celdas con buen y mal rendimiento. El segundo bloque se enfocó en seleccionar las mejores celdas para el balanceo de carga. Por último, en el tercer bloque aplicó un algoritmo de regresión a cada celda seleccionada.

Gutiérrez (2021), aplicó técnicas de aprendizaje automático para abordar cuestiones relacionadas con el rendimiento de las celdas y la identificación de patrones de uso de las antenas. Entre las distintas técnicas de aprendizaje automático utilizó modelos tales como XGBoost, Random Forest, entre otros. Estos modelos le permitieron generar clasificaciones y predicciones con relación a la cantidad de días en los que, semanalmente, una celda experimenta un bajo rendimiento.

Por último, López (2021) desarrolló un sistema de detección de anomalías en celdas LTE el cual fue dividido en tres etapas: predicción, detección de anomalías y análisis de causa-raíz. Para reducir la dimensionalidad empleó el método conocido como

Bosque Aleatorio y para predecir los recursos utilizados utilizó el método de Bosque de Regresión de Cuantiles.

2.2. Bases teóricas

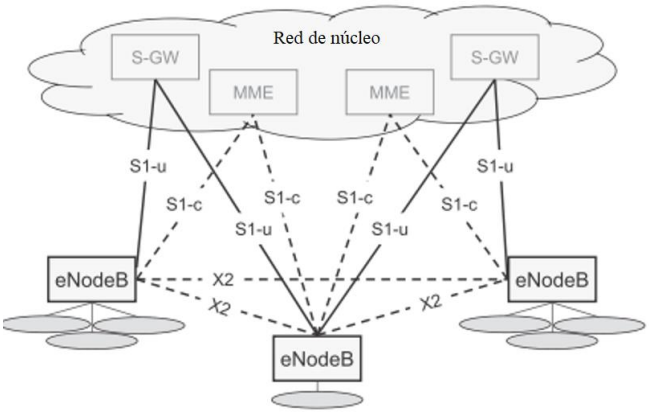
2.2.1 Redes LTE:

La tecnología LTE presenta mejoras notables frente a su predecesora, la red UMTS. Entre ellas se encuentran: Reducción de latencia tanto en el plano de usuario como de control, nuevos esquemas de antenas Múltiple Entrada/Múltiple salida (MIMO) y nuevos esquemas de multiplexación.

Esto se hace posible gracias a la división que existe entre la red de acceso (RAN) y la arquitectura de red central conocida como el núcleo de paquetes evolucionado (EPC). En la Figura 2 se muestran los principales componentes de una red de acceso radio.

Figura 2

Componentes de una red de acceso radio sus interfaces.



Nota. Adaptado de *LTE-Advanced for Mobile Broadband* (p.123), por Dahlman et al. ,2014, Elsevier.

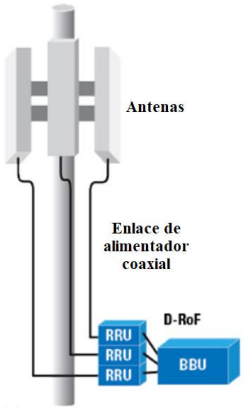
Red de acceso (RAN):

Está encargada de la gestión de la interfaz radio que existe entre la estación base LTE y el equipo móvil. Entre sus principales funcionalidades se encuentra: programación, manejo de recursos de radio, protocolos de retransmisión, codificación y varios esquemas de antena múltiple.

El elemento principal de una red de acceso de radio es la estación base LTE llamada eNodeB. Cabe resaltar que esto es un concepto lógico puesto que la implementación física de un eNodeB requiere la instalación de una antena de 3 sectores, una unidad de procesamiento de banda base en la cual varios cabezales de radio remoto están conectados. Un ejemplo de una estación base eNodeB se muestra en la Figura 3.

Figura 3

Estación base físico

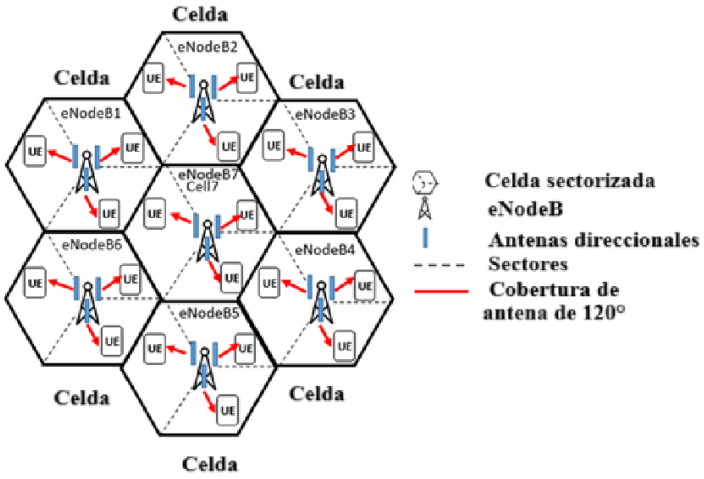


Nota. Esta figura representa la conexión entre antenas y unidades de radio remoto (RRUs) mediante enlaces de cables coaxiales. Posteriormente, las RRUs se conectan a la unidad de banda base (BBU) mediante cables de fibra óptica. Adaptado de *The Evolution of Interconnects in Cellular Networks: From 4G LTE eNodeB to 5G gNB* por Pasternack, I, 2021, Microwave Journal. (https://www.microwavejournal.com/articles/35582-the-evolution-of-interconnects-in-cellular-networks-from-4g-lte-enodeb-to-5g-gnb).

En la Figura 4 se muestra una topología típica de red conformado por un eNodeB y 3 sectores LTE asociados:

Figura 4

Conjunto de celdas conformadas por una estación base y 3 sectores:



Nota. Elaboración propia . Esta figura representa una agrupación de celdas LTE Teniendo como elemento central a la estación física eNodeB.

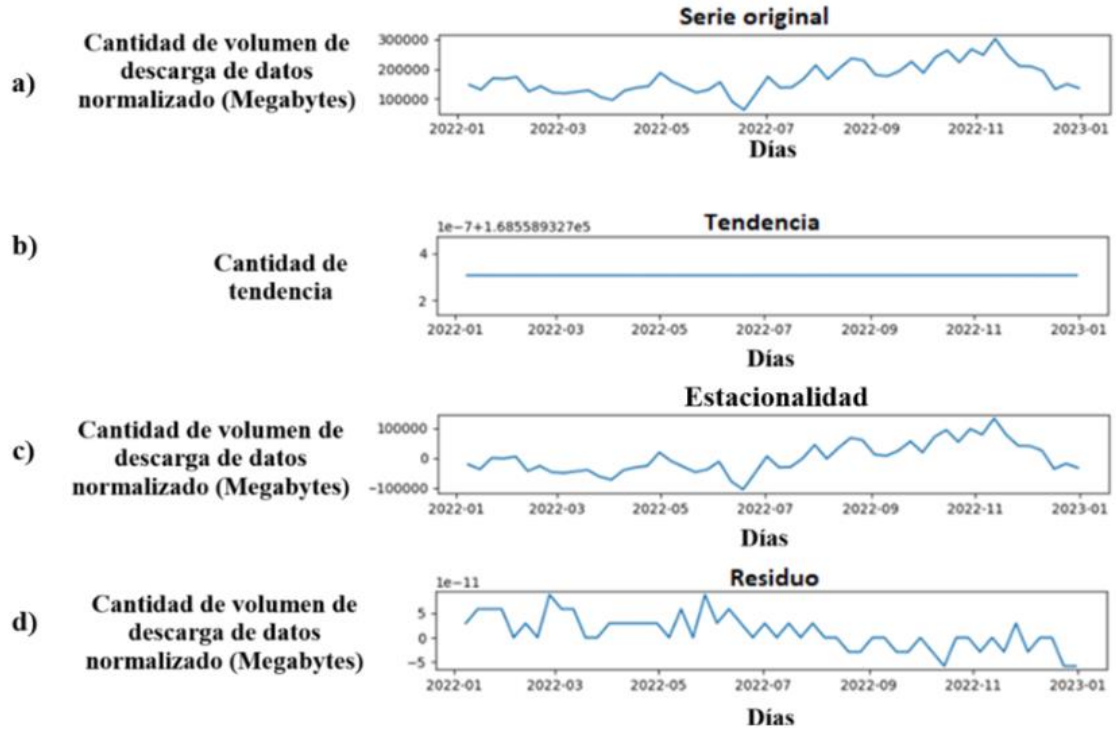
Según Dahlman et al. (2014), varias implementaciones de estaciones base pueden pertenecer a un mismo eNodeB, pero una sola implementación de estación base no puede tener varios eNodeB a la vez. Esto es debido a que cada eNodeB tiene asignado determinados recursos lógicos únicos a una región geográfica en específico.

Series de tiempo:

Según Schaffer, Dobbins, Pearson (2021), una serie temporal se define como una secuencia de puntos de datos espaciados de manera uniforme en el tiempo y dispuestos cronológicamente. Las series temporales típicamente exhiben tres características: la ausencia de tendencia, la presencia de estacionalidad y la existencia de residuos. Al observar las mediciones del volumen de tráfico de descarga de datos por hora en las celdas, se puede apreciar un patrón de serie de tiempo, como se ilustra en la Figura 5.

Figura 5:

Descomposición del volumen de tráfico de descarga de datos a lo largo del 2022 para una celda LTE.



Nota. Elaboración propia basado en la base de datos de WOM. Esta figura ilustra la descomposición del volumen total de tráfico de descarga de datos de una celda LTE a lo largo del año 2022. Se utiliza el método de descomposición STL, y la medición se expresa en megabytes (MB). a) Serie de tiempo original. b) Componente de tendencia de la serie de tiempo. C) Componente de estacionalidad de la serie de tiempo. D) Componente residual de la serie de tiempo.

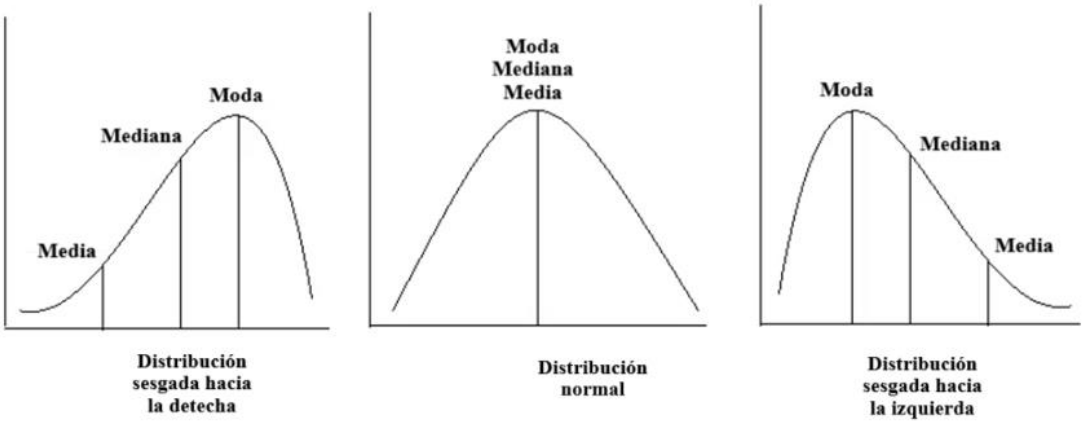
2.2.2. Principales métodos estadísticos:

Media, mediana y moda:

Entre los 3 principales tipos de promedio se encuentran la media, la mediana y la moda. En la Figura 6 se muestra como están representados estas mediciones en diferentes distribuciones estadísticas:

Figura 6:

Representación de la media, mediana y moda



Nota. La gráfica ilustra cómo, mediante el cálculo de la media, mediana y moda, es posible determinar el tipo de sesgo presente en la distribución de los datos.

Adaptado de "*Mean, Median, and Mode in Statistics"* por Tran, N, 2019. https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b

Media:

La suma de todos los datos contenidos en una muestra divido por el tamaño total de la muestra.

Mediana:

La mediana es el dato intermedio que se ubica en el centro de todos los datos de una muestra, cuando estos se encuentran ordenados de menor a mayor. En el caso de que el tamaño de la muestra sea par, la mediana será el promedio de los dos datos centrales.

Moda:

La moda es el dato que presenta la mayor frecuencia, o, dicho de otra manera, el valor que se repite más veces en un conjunto de datos.

Varianza:

La varianza se define como la diferencia al cuadrado entre cada dato del conjunto y la media, dividida entre el número total de datos. Esta medida representa la dispersión

de los datos alrededor de la media, penalizando los casos en los que los datos están considerablemente alejados de la media.

La fórmula que describe la varianza es la siguiente:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n} \quad (1)$$

donde:

x = Cada dato dentro del conjunto de datos

μ = Media del conjunto de datos

n = Tamaño total del conjunto de datos

Desviación estándar:

La desviación estándar es la raíz cuadrada de la varianza. Su valor se expresa en las mismas unidades que la variable, facilitando así su interpretación en comparación con la varianza. Al igual que esta última, la desviación estándar mide la dispersión de los datos alrededor de la media del conjunto. Griffiths (2008).

RMSE:

La raíz cuadrada del error cuadrático medio (RMSE) es una métrica que evalúa la magnitud promedio de los errores entre los valores pronosticados y los reales. Conocida también como la desviación cuadrática media, se utiliza para penalizar los errores grandes, permitiendo así ajustar adecuadamente los parámetros del algoritmo en ejecución. Madrigal (2022).

La fórmula que describe la RMSE es la siguiente:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (2)$$

donde:

n = Número de muestras

yi: Valor real

y sombrero: Valor pronosticado.

2.2.3. Algoritmos de agrupamiento:

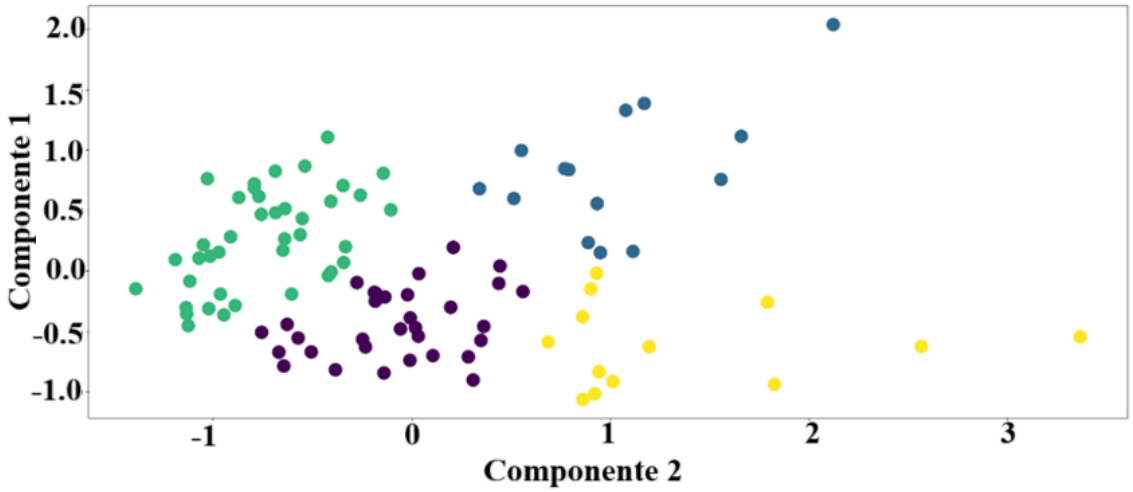
2.2.3.1 K-Means

El método K-Means es un algoritmo de aprendizaje de máquina basado en distancia, útil, por ejemplo, cuando se busca identificar un grupo de consumidores con comportamientos comunes o agrupar documentos según sus similaridades. Según Parsian (2015), el número de grupos se determina mediante el valor de K, que se asigna al algoritmo al ejecutarlo. Al diseñar un algoritmo K-Means, surge la pregunta fundamental: ¿Cómo determinar el valor óptimo de K para representar adecuadamente los grupos en nuestros datos de entrada?

Tomemos, por ejemplo, el caso donde se elige K=4, como se ilustra en la Figura 7. Los puntos verdes representan el primer grupo, los puntos morados representan el segundo grupo, los puntos amarillos representan el tercer grupo, y los puntos azules representan el cuarto grupo.

Figura 7

Agrupamiento con valor de K=4



Nota. Elaboración propia basado en la base de datos de WOM. A modo de ejemplo, se muestra un agrupamiento de datos vinculado a una serie temporal de 100 celdas tomados de la base de datos de WOM. En este escenario, se empleó la técnica de agrupamiento K-Means con un valor de K igual a 4 en conjunto con la técnica de reducción de dimensionalidad PCA, lo que dio como resultado la identificación de 2 componentes principales.

El algoritmo K-Means opera dividiendo el conjunto de datos de entrada en grupos con igual varianza, con el objetivo de minimizar un criterio conocido como inercia o suma de cuadrados intra grupo. En consecuencia, el resultado del algoritmo son grupos cuyos centros, denominados centroides, minimizan el valor de la inercia.

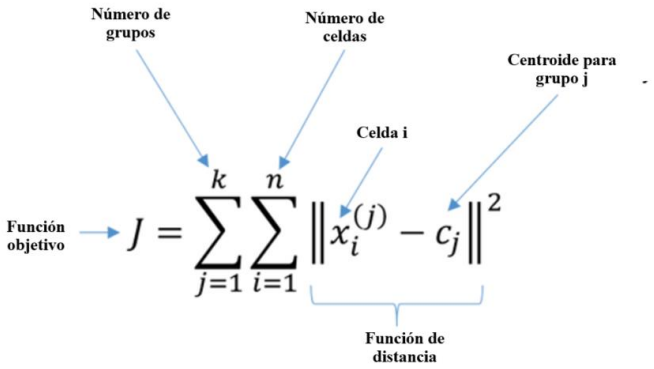
En la fase inicial, el algoritmo selecciona centroides de manera aleatoria. A continuación, evalúa cada punto de entrada midiendo su distancia euclidiana respecto al centroide más cercano y le asigna a ese grupo. Posteriormente, calcula la media de los datos en cada grupo y ajusta el valor del centroide de acuerdo a estos nuevos datos. Este proceso se repite, midiendo las distancias euclidianas hasta que se alcanza la convergencia o se llega al número máximo de iteraciones.

Para determinar el número óptimo de grupos, existen métodos como el método del codo y el método de la silueta.

Método del codo:

Este método calcula la suma de errores cuadráticos (WSS) dentro del grupo para diferentes valores de K y selecciona el valor para el cual estos errores comienzan a disminuir.

La fórmula que describe el cálculo de la inercia es la siguiente:



(3)

donde:

k = Número de grupos

j = Identificador de cada grupo

i = Identificador de cada dato dentro del conjunto de datos

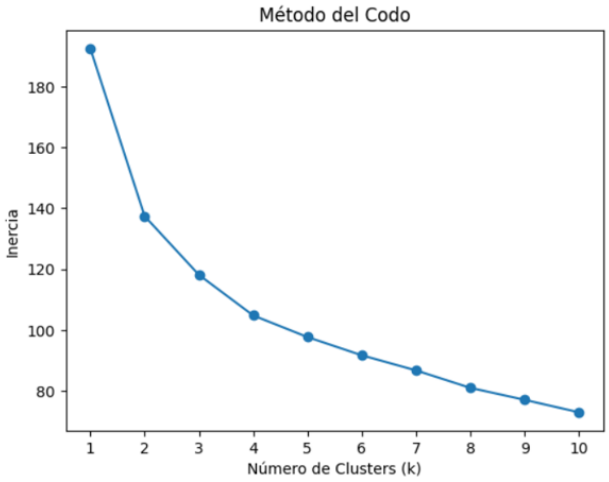
x = Dato individual dentro del conjunto de datos

c = Centroide de cada grupo

En la Figura 8 se muestra la representación gráfica del valor de número de grupos vs la inercia.

Figura 8

Representación gráfica del valor de número de grupos (K) vs la inercia



Nota. Elaboración propia basado en la base de datos de WOM. El gráfico representa la reducción de la inercia a medida que aumenta el número de grupos en K-Means. Según el método del codo, el valor óptimo de K es la porción de la figura con mayor curvatura. Syakur et al. (2018).

Método de silueta:

Este método determina que tan cerca está un punto de un grupo comparado con los puntos de los otros grupos vecinos, su valor varía de -1 a 1. Rousseeuw, P.J (1986).

La fórmula que describe el cálculo del coeficiente de silueta es la siguiente:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{4}$$

donde:

a = Distancia euclidiana promedio entre una celda y todas las celdas dentro del mismo grupo.

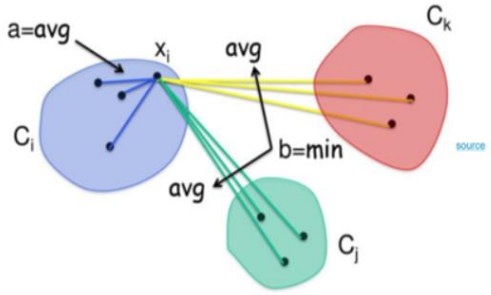
b = Distancia euclidiana promedio entre una celda y todas las celdas dentro del grupo más cercano.

S(x) = Promedio del coeficiente de silueta para todas las celdas agrupadas.

En la Figura 9 se muestra el coeficiente de silueta:

Figura 9:

Coeficiente de silueta



Nota. La gráfica representa la fórmula del coeficiente de silueta donde a es la distancia media a las otras instancias en el mismo grupo y b es la distancia media al grupo más cercano.

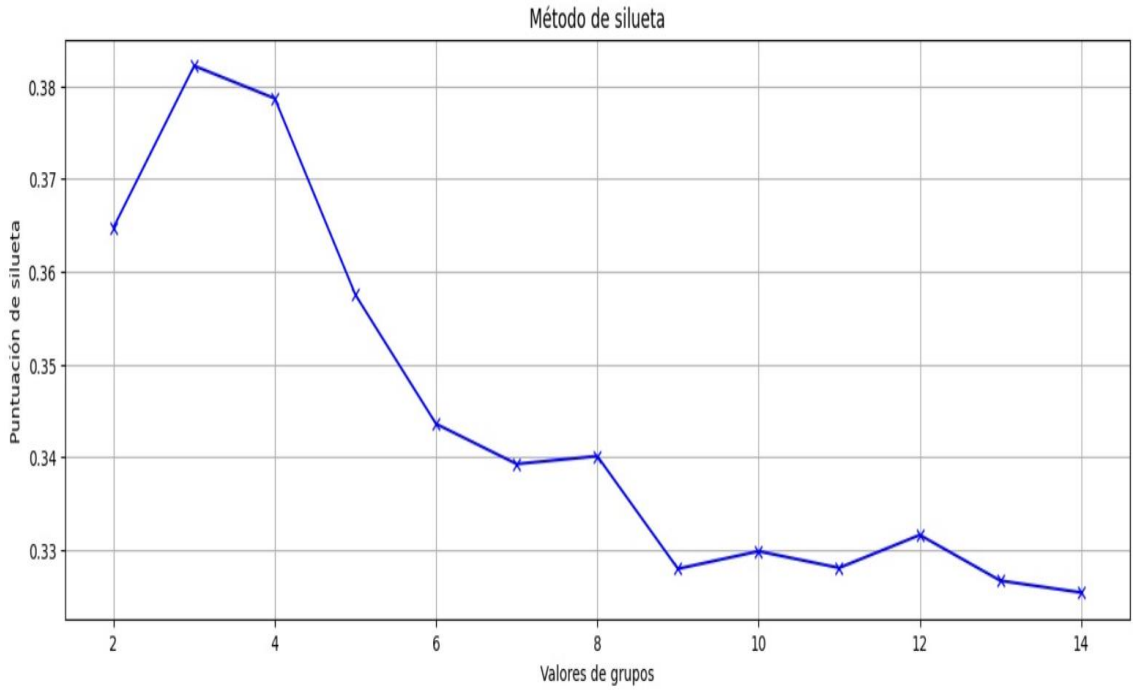
Adaptado de "*How to Evaluate the Performance of Clustering Algorithms Using Silhouette Coefficient"* por Koli, S. 2021. https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03

Cuando el coeficiente b es mucho mayor que a entonces el coeficiente de silueta está más cerca de +1 y por lo tanto la instancia está más cerca que el centro de su propio grupo que del grupo vecino más cercano. Por otra parte, si el coeficiente de b es igual al coeficiente a entonces probablemente el punto se encuentre en el límite de decisión entre su propio grupo y el grupo vecino cercano. Finalmente, si el valor de a es mucho mayor que el coeficiente b entonces probablemente la instancia se encuentra asignado al grupo incorrecto. Rousseeuw, P.J (1986).

En la Figura 10 se muestra la representación de los coeficientes de silueta para distintos valores K:

Figura 10

Representación gráfica del valor de número de grupos (K) vs el coeficiente de silueta



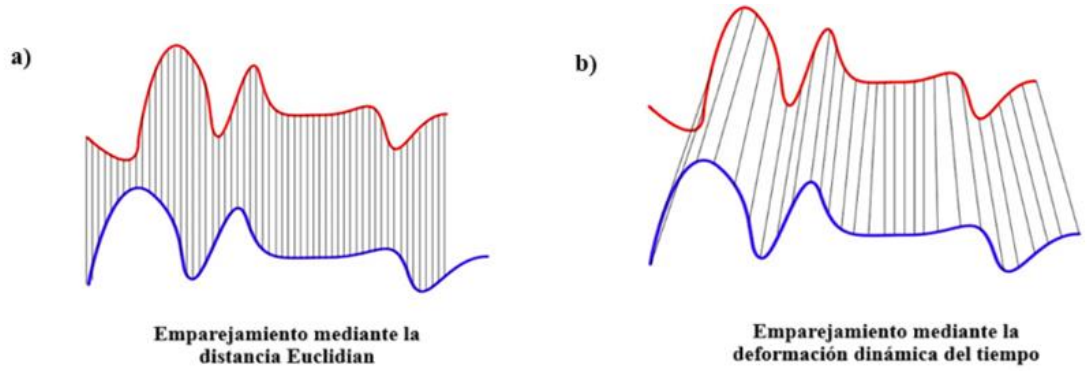
Nota. Elaboración propia basado en la base de datos de WOM. El gráfico representa la puntuación del coeficiente de silueta a medida que aumenta el número de grupos en K-Means. Según el método de la silueta, el valor óptimo de K es la porción de la figura donde la curva alcanza su máximo valor. Rousseeuw, P.J (1986).

Deformación dinámica del tiempo (DTW):

Es un método que se utiliza para comparar la similitud o distancia entre secuencias de 2 tiempos o matrices de diferentes longitudes. A diferencia de la distancia euclidiana que compara entre un punto a otro, la Deformación dinámica del tiempo permite realizar comparaciones de muchos puntos a uno. En la Figura 11 se muestra la comparación entre estos 2 enfoques:

Figura 11

Diferencia entre la identificación de similitudes mediante el método euclidiano y el método de deformación del tiempo.



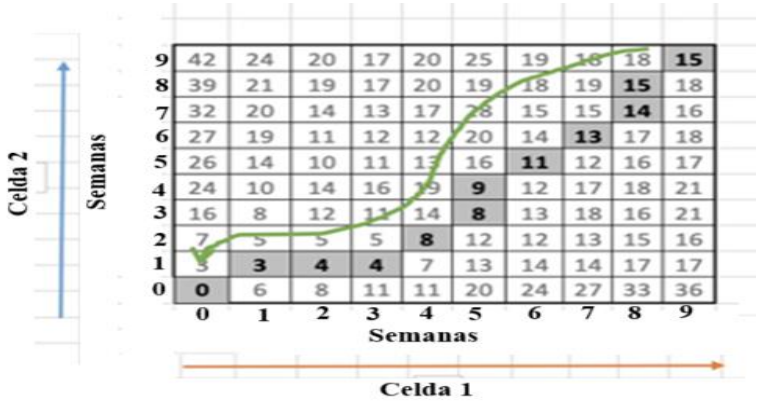
Nota. Esta figura ilustra la diferencia entre el método DTW y la distancia Euclidiana. a) Emparejamiento mediante la distancia Euclidiana. b) Emparejamiento mediante la deformación dinámica del tiempo. Adaptado de "*Time Series Similarity Using Dynamic Time Warping -Explained"* por Mishra, A, 2020. https://medium.com/walmartglobaltech/time-series-similarity-using-dynamic-time-warping-explained-9d09119e48ec

Para comparar dos series de tiempo, DTW utiliza una métrica denominada cálculo de costos, que compara cada dimensión de las series de tiempo y las pondera en una matriz, también conocida como matriz de costos. H. Sakoe et al. (1978).

En la Figura 12 se muestra la matriz de costos y su fórmula asociada:

Figura 12

Generación de matriz de costos



Nota. Curva de matriz de costos el cual se obtiene comparando las dimensiones de cada una de las celdas.
Adaptado de "*Time Series Similarity Using Dynamic Time Warping -Explained"* por Mishra, A, 2020. https://medium.com/walmartglobaltech/time-series-similarity-using-dynamic-time-warping-explained-9d09119e48ec

Finalmente, el resultado se pondera, obteniendo una distancia final que permite comparar el grado de similitud entre unas series de tiempo y otras.

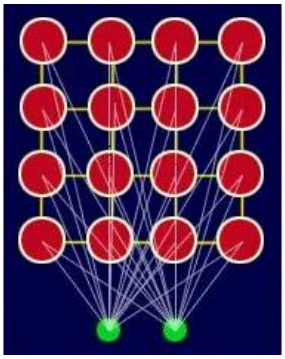
2.2.3.2 Mapas de autoorganización (SOM):

El método de mapa autoorganizado (SOM, por sus siglas en inglés) es un modelo neuronal que forma parte del grupo de redes de aprendizaje competitivo. Su entrenamiento no requiere intervención humana, ya que se lleva a cabo mediante las características de los datos de entrada. Por lo tanto, esta técnica también se clasifica dentro de las técnicas de aprendizaje automático no supervisado. Además, agrupa los datos de entrada similares, lo que lo posiciona como una técnica de agrupación

Para Vellido et al. (2019), esta técnica resulta muy apropiada para visualizar conjuntos de datos de alta dimensionalidad debido a su capacidad para detectar características inherentes al problema y reducir la dimensión de los datos a mapear. El algoritmo consta de una malla de nodos de tamaño n x n conectada a un vector de entrada de n dimensiones, y este arreglo se conoce como una red Kohonen. En la Figura 13 se muestra visualmente esta disposición:

Figura 13

Red de Kohonen 4x4



Nota. Visualización de una red Kohonen compuesta por un vector de entrada de 2 dimensiones y un arreglo de capas de salida de 4x4 dimensiones.

Adaptado de "*Self-Organizing Map (SOM) with Practical Implementation"* por Alí, A, 2019. https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117

Los vectores de entrada representan las características o variables del conjunto de datos, mientras que los nodos de salida constituyen el mapa en sí. Cada nodo tiene una coordenada específica (x, y) y contiene un vector de pesos determinado por la dirección con respecto al vector de entrada. Por lo tanto, si los datos de entrenamiento cuentan con Vn vectores de entrada, los pesos de cada nodo serán de Wn, respectivamente. Cuando los pesos de cada nodo coinciden con el vector de entrada, esa área de la red se optimiza para parecerse más a los datos de la clase a la que pertenece el vector de entrada.

Después de una serie de iteraciones, el algoritmo transforma la distribución inicial de pesos aleatorios en un mapa de zonas estables. Cada zona representa una característica, lo que resulta en un mapa final que muestra las características del espacio de entrada.

Tomemos como ejemplo un conjunto de datos con 3 características y 20,000 datos de entrada. El procedimiento inicia de la siguiente manera: cada nodo tiene un peso inicial de 0. Al extraer los valores de cada característica de la primera fila, el objetivo es encontrar el nodo de salida más cercano a esa fila. Para lograrlo, se recorre cada nodo y se calcula la distancia euclidiana entre el peso de cada nodo y el vector de entrada actual. El nodo con el vector de peso más cercano al vector de entrada se conoce como la unidad de mejor coincidencia (BMU).

La ecuación que describe la distancia euclidiana es la siguiente:

$$\text{Distance} = \sqrt{\sum_{i=0}^{i=n}(X_i - W_i)^2} \tag{5}$$

donde:

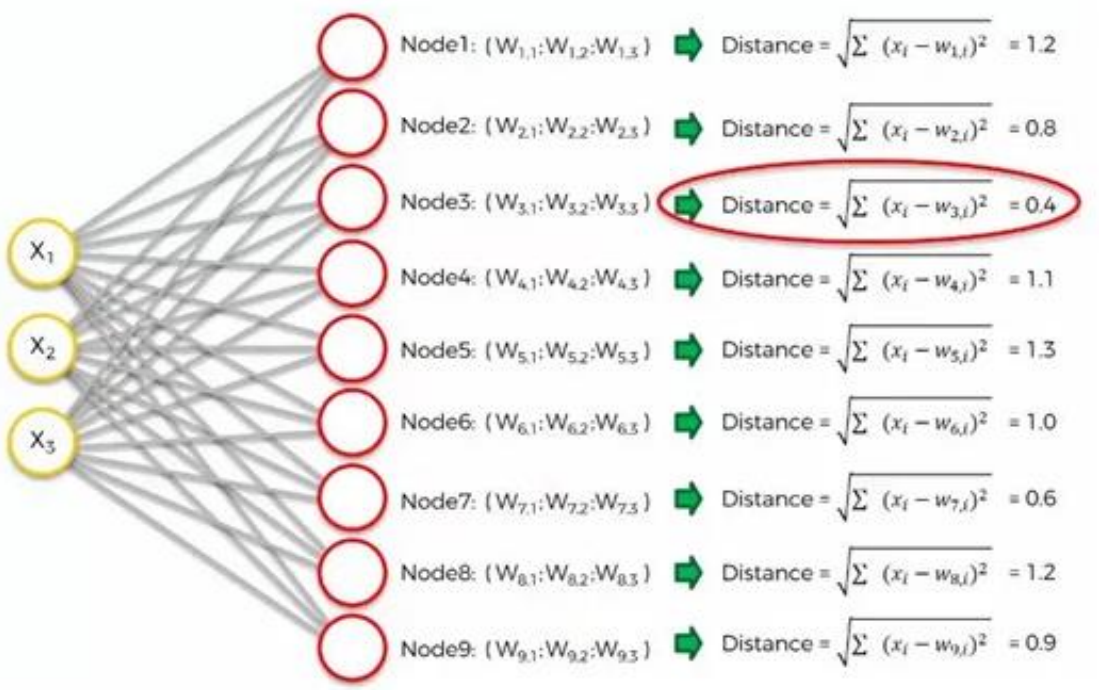
Xi = Vector de entrada actual

Wi = Vector de peso del nodo

En la Figura 14 se muestra de manera gráfica este análisis:

Figura 14

Encontrando el nodo de mejor coincidencia en una red compuesta de un vector de entrada de 3 dimensiones y un mapa de 9 nodos de salida.



Nota. Este gráfico representa la determinación de la unidad de mejor coincidencia (BMU). Aquí se observa que el nodo 3 es la unidad de mejor coincidencia.

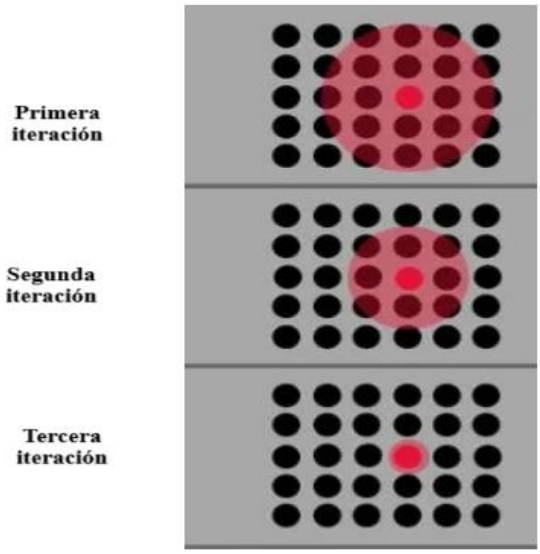
Adaptado de "*Self-Organizing Map (SOM) with Practical Implementation"* por Alí, A, 2019. https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117

Luego de haberse determinado el BMU es necesario determinar la función de vecindad y los nodos que se encuentran contenidos en esta distancia. Para ello es necesario calcular parámetros tales como el radio de la vecindad de tal manera que se pueda demostrar que cada nodo está dentro de la distancia radial. El tamaño de cada vecindad decrece a una tasa determinada por la amplitud de la tasa de aprendizaje. una constante de tiempo y por cada iteración que ocurra:

En la Figura 15 se muestra una representación de las BMU pertenecientes a una vecindad y como esta decrece a medida que se va iterando en la red neuronal.

Figura 15

Red de Kohonen de 6x6



Nota. Representación de una red de Kohonen de 6x6 y las BMUs pertenecientes a una vecindad.

Adaptado de "*Self-Organizing Map (SOM) with Practical Implementation"* por Alí, A, 2019. https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117.

La fórmula de la tasa de aprendizaje viene determinada por la siguiente ecuación:

$$\alpha(t) = \alpha_0 e^{-\frac{t}{\tau_2}} \tag{6}$$

donde:

alfa = Constante de tasa de aprendizaje inicial

t2 = Número de iteraciones/Máximo radio de mapa

Error de cuantización en Mapas autoorganizados:

Esta métrica evalúa la calidad de la red neuronal al comparar la suma de la varianza existente entre cada dato y su unidad de mejor coincidencia (BMU). Es un concepto similar a la inercia en el agrupamiento mediante K-Means. B. Dresp et al. (2018).

Viene determinado por la siguiente fórmula:

$$QE = 1/N \sum_{i=1}^{N} \left\| X_i - \left(BMU_{(i)} \right) \right\| \quad (7)$$

donde:

Xi: Cada unidad dentro del conjunto de datos

BMU: Cada unidad de máxima coincidencia

Producto topográfico:

Esta métrica indica si la dimensión del mapa es apropiada para coincidir con el conjunto de datos o si esto resultará en la violación de la vecindad debido a la distorsión del mapa. La fórmula se expresa mediante la siguiente expresión:

$$\mathrm{TP}(\{\mathbf{m}_k\}) = \frac{1}{K(K-1)} \sum_{j=1}^{K} \sum_{k=1}^{K-1} \log P_3(j,k). \quad (8)$$

donde:

K: Número de grupos

J, k: Iteraciones entre las dimensiones del mapa

P: Producto de los radios definidos por la distancia de un prototipo de mapa j a su k-th vecino más cercano en el mapa y a su vecino k-th más cercano en el espacio de entrada.

III. VARIABLES E HIPÓTESIS

3.1 Operacionalización de las variables

Variable independiente: Volumen de tráfico descargado

Esta variable hace referencia a la cantidad de datos que un usuario descarga desde Internet a través de la red LTE. Esto podría incluir datos de navegación en la red, archivos descargados, transmisión de videos, entre otros. Por lo general, se mide en Megabytes (MB) o Gigabytes (GB).

Operación:

(Volumen total de descarga de datos para en una celda) x 1 hora

Variable dependiente: Patrón de tráfico identificado

Esta variable se refiere al comportamiento del tráfico de las celdas LTE durante el año 2022. Cada celda posee un patrón específico el cual depende de otras variables tales como: ubicación geográfica, implementación de nuevos sectores LTE y aumento de número de usuarios.

3.2. Hipótesis de la investigación:

3.2.1. Hipótesis general:

Se pueden identificar patrones de tráfico de celdas LTE de la red nacional móvil de la operadora WOM de Chile utilizando algoritmos de agrupamiento en aprendizaje de máquina no supervisado a partir de los datos registrados en el 2022.

3.2.2. Hipótesis específicas:

- Es posible identificar los patrones de comportamiento de las celdas LTE con los datos del año 2022

- Se puede utilizar algoritmos de agrupamiento en aprendizaje de máquina no supervisado para explicar el patrón de tráfico de las celdas LTE en una determinada zona geográfica.

IV. METODOLOGÍA

4.1. Descripción de la metodología

Esta investigación fue desarrollada mediante la siguiente metodología:

Recolección de datos:

En esta etapa, se realizaron consultas a la base de datos de WOM, donde estaban almacenados los datos de las celdas. Para ello, se emplearon Código realizados basados en SQL y Python.

Tratamiento de datos faltantes:

En esta etapa, se emplearon principalmente dos técnicas de tratamiento de datos faltantes: interpolación lineal e interpolación polinómica. Se tomaron muestras de 10 celdas al azar y se seleccionaron 8 segmentos con intervalos de tiempo variables también al azar. Posteriormente, se calculó el valor del error cuadrático medio.

Segmentación de datos:

Se utilizaron métodos estadísticos para determinar la mejor forma de agrupar los datos, ya sea por horario, diario, semanal o mensual. Además, se aplicaron métricas estadísticas como el promedio, la mediana o la suma para obtener muestras de cada agrupación de datos.

Agrupamiento de datos:

En esta fase, se determinó el mejor modelo de aprendizaje automático tanto para el método K-Means como para el método de mapas autoorganizados (SOM). Para K-Means, se realizaron pruebas manteniendo las 51 dimensiones y reduciéndolas a 2 mediante la técnica del Análisis de Componentes Principales (PCA). Luego, se aplicaron las técnicas del método del codo y del método de silueta. Para SOM, se buscaron los 3 mejores tamaños de mapa utilizando como métrica el error de cuantización. Posteriormente, después de identificar los mejores tamaños de mapa, se procedió a identificar las mejores combinaciones de dimensiones mediante la métrica del error de producto topográfico. Finalmente, se realizaron iteraciones con distintos valores de tasa de aprendizaje, valor sigma y función de vecindad para encontrar los mejores parámetros de cada red neuronal.

Comparación de modelos:

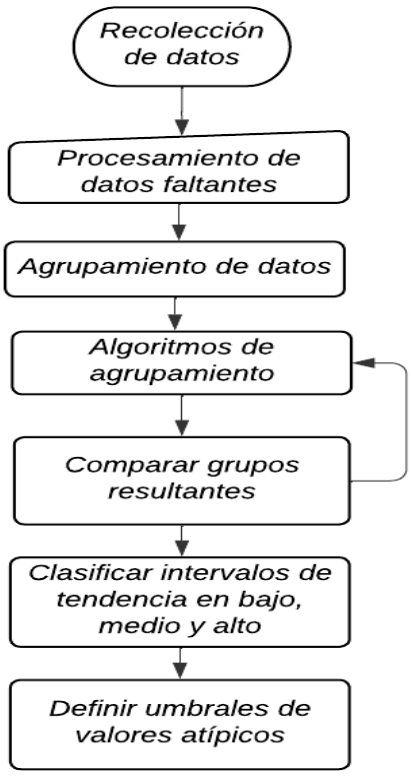
Se realizaron tres evaluaciones para cada modelo en relación con las celdas de la red LTE de WOM: la evaluación de los tipos de tendencia detectados en la red por cada modelo, la evaluación de la cantidad de volumen de descarga de datos detectado por cada modelo y la evaluación de celdas atípicas detectadas por cada modelo.

Interpretación:

En esta etapa, se presentarán los resultados finales y se ofrecerán recomendaciones sobre la aplicación de las técnicas de agrupamiento. La Figura 16 detalla este proceso de manera más exhaustiva.

Figura 16

Diagrama de flujo del diseño de la investigación



Nota. Elaboración propia

4.2. Implementación de la investigación

4.2.1. Herramientas de desarrollo

Para el tratamiento de la información almacenada en la base de datos relacional Clickhouse del operador WOM, se empleó el lenguaje de programación SQL. A través de este lenguaje, fue posible procesar, segmentar, agrupar y transformar las tablas de datos en bruto contenidas en este sistema de información. Posteriormente, utilizando el lenguaje de programación Python, se utilizaron librerías que facilitaron la conexión con la base de datos y la ejecución de consultas con los datos modificados.

Adicionalmente, se emplearon librerías para el manejo de tablas de datos en Python, entre las cuales se incluyen Pandas, que ofrece funciones para el tratamiento de datos estructurados, y Numpy, que proporciona funciones y herramientas matemáticas para trabajar con matrices de datos. Para la generación de interfaces de visualización, se hizo uso de las librerías de Matplotlib.

En la implementación de algoritmos de aprendizaje automático, se utilizaron dos librerías principales: Scikit-learn y tslearn. Scikit-learn proporcionó herramientas útiles para llevar a cabo tareas de agrupamiento, mientras que tslearn resultó útil para el tratamiento de información que contiene características de series de tiempo.

El método de agrupamiento K-Means se llevó a cabo utilizando las librerías de Scikit-learn, mientras que el método de mapas autoorganizados se implementó a través de la librería Minisom, que ofrece herramientas para realizar agrupamientos mediante redes neuronales artificiales.

El entorno de desarrollo interactivo de programación fue la aplicación Jupyter Lab, y la versión del lenguaje Python utilizada fue la 3.10.9.

4.2.2. Pre-procesamiento de datos

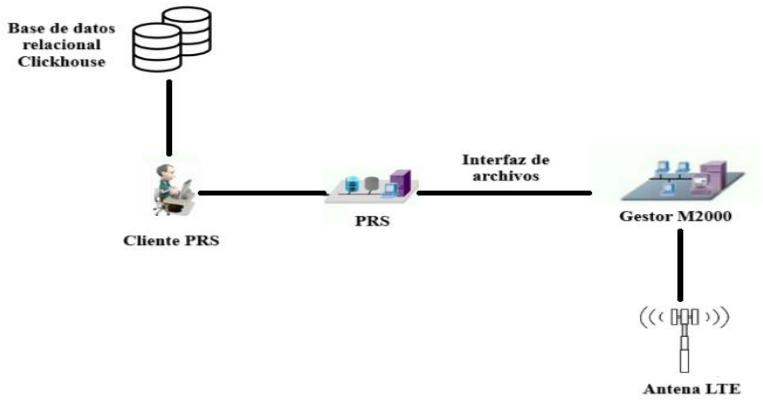
Para la recolección de datos, se empleó la plataforma de administración de redes MAE server de la marca Huawei. Este equipo procesó los datos provenientes del tráfico de usuario entre las antenas LTE y los equipos móviles. CC Huawei iMaster MAE (2020).

Posteriormente, la operadora procesó los datos utilizando la plataforma PRS de Huawei el cual es un sistema especializado que permite a los usuarios gestionar reportes de rendimiento de red de larga data. Huawei PRS documentation (2009).

Una vez completada esta recopilación, se tabularon y procesaron los datos, almacenándolos en una base de datos relacional llamada Clickhouse. En la Figura 17 se observa cómo está desplegada esta solución en una topología de red.

Figura 17

Proceso de recolección de datos



Nota. Adaptado de "*Introduction to PRS Manager*" por Huawei PRS Documentation, 2009. https://carrier.huawei.com/en/products/wireless-network-v3/SubSolution-SingleOSS/iManager-PRS

4.2.3. Variable objetivo

La base de datos de la operadora WOM almacena mediciones de los 109 indicadores principales de rendimiento de la red LTE, los cuales se subdividen en categorías como Disponibilidad de servicio, Accesibilidad, Retinibilidad, Movilidad, Integridad de servicio, Utilización, Disponibilidad y Tráfico. En esta tesis, el enfoque se centró exclusivamente en el indicador de tráfico, específicamente en el indicador de descarga de datos LTE por hora (L_Thrp_bits_DL). Esto se debe a que este indicador no solo considera la velocidad de descarga de datos, sino también la cantidad de datos transferidos, generando así una evaluación más precisa del rendimiento de la red. Hwangnam (2018).

Adicionalmente, este indicador resulta crucial al analizar tendencias, ya que se busca revisar el rendimiento de la red a nivel macro (por semanas) en lugar de realizar análisis detallados (por hora).

4.2.4. Resumen de la base de datos:

La operadora WOM almacenó datos sobre aproximadamente 38,183 celdas LTE operativas durante el año 2022. Estas celdas se distribuyeron en diversas bandas de frecuencia, que incluyen AWS A-F (Canal de descarga: 1700 MHz), EAWS A-J (Canal de descarga: 1700 MHz) y 700 APT (Canal de descarga: 763 MHz). Para proporcionar más detalles, la red LTE de WOM se compone de 19,257 celdas AWS, 14,854 celdas AWS-E y 5,230 celdas 700 APT.

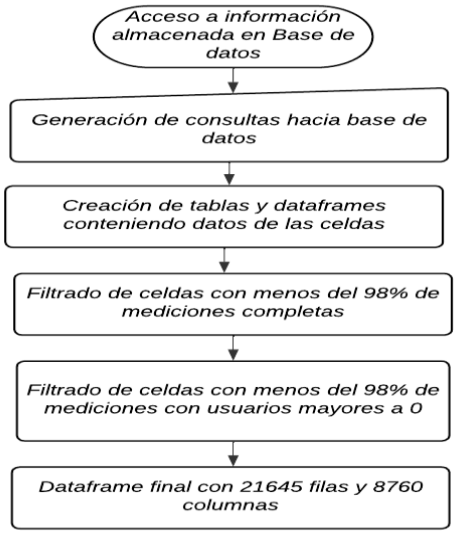
Se eligió la celda, o más específicamente el nombre de la celda, como indicador principal, ya que este atributo posee un carácter único en toda la red e incluye información acerca de la banda de frecuencia, la región y el número de sector. No se seleccionó la estación base LTE o eNodeB como identificador único, ya que una celda es el bloque más básico con el que se construye una red móvil, permitiendo un análisis granular del rendimiento de la red en comparación con un análisis mediante eNodeB. Además, una celda brinda cobertura a un área específica, mientras que un eNodeB es responsable de gestionar los recursos de radio dentro de la celda.

4.2.5. Depuración de datos ambiguos

A continuación, en la Figura 18 se muestra el diagrama de flujo de esta etapa:

Figura 18

Diagrama de flujo de la depuración de datos ambiguos



Nota. Elaboración propia

Cada indicador de rendimiento cuenta con 8760 mediciones a lo largo de un año completo, lo que implica que cada celda LTE poseerá, en el mejor de los casos, 8760 mediciones relacionadas con el indicador de descarga de datos. Sin embargo, este escenario no es aplicable a todas las celdas. Además, existen situaciones en las que ciertas celdas estuvieron temporalmente fuera de servicio debido a problemas de mantenimiento o interrupciones en el servicio. Para abordar estas problemáticas, se aplicaron técnicas de interpolación de datos faltantes. En esta tesis, se ha trabajado con un mínimo de disponibilidad de servicio del 98% en cada celda. Aquellas unidades de estudio cuyos valores fueron inferiores a este límite han sido descartadas (ver Tabla 2).

Tabla 2

Porcentaje de disponibilidad de los datos del 2022 de las celdas y su relación con el número de horas.

Disponibilidad	Horas
100%	8760
98%	8590
95%	8322
90%	7884
80%	7008

Nota. Elaboración propia basado en la base de datos de WOM. Disponibilidad de una celda a lo largo de un año. Como se puede observar en esta tabla, considerando una disponibilidad del 98%, se utilizaron las celdas que tuvieron datos faltantes de aproximadamente 170 horas. No se consideraron las celdas con datos faltantes mayores a este umbral, ya que aumentaba el porcentaje de errores de las técnicas de imputación y hacía menos preciso el algoritmo.

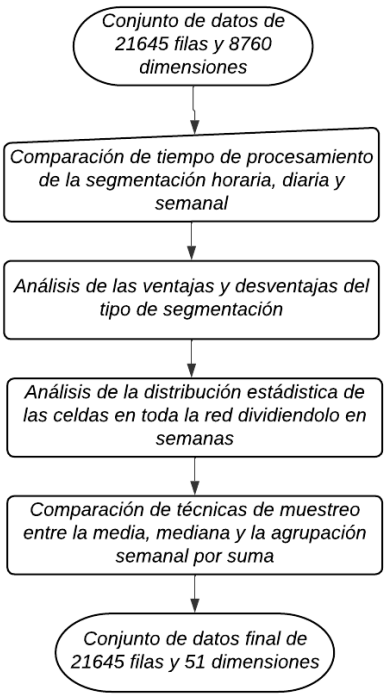
4.2.6. Segmentación y muestreo de datos

En esta etapa, se llevaron a cabo análisis para identificar cuál es la mejor manera de segmentar los datos, considerando opciones como por horas (dato original), por días, semanas y meses. Para ello, se mostraron de manera gráfica las ventajas y desventajas de cada una de estas segmentaciones, y se presentó de manera cuantitativa el tiempo que toma procesar cada elección.

Posteriormente, se realizó un análisis estadístico de las distribuciones de las celdas con el fin de determinar el método óptimo de muestreo, analizando la media, la mediana y agrupando por suma. La Figura 19 ilustra este proceso:

Figura 19

Diagrama de flujo del proceso de segmentación y muestreo de datos.



Nota. Elaboración propia

4.2.7. Normalización de datos:

En esta etapa, los datos fueron normalizados mediante la técnica Min-Max, de modo que la semana con la suma más alta de descarga de datos en todo el año represente el valor de 1, mientras que la semana con la suma más baja represente el valor de 0. Finalmente, el resultado de este proceso fue el conjunto de datos listo para ejecutarse en los algoritmos de agrupamiento.

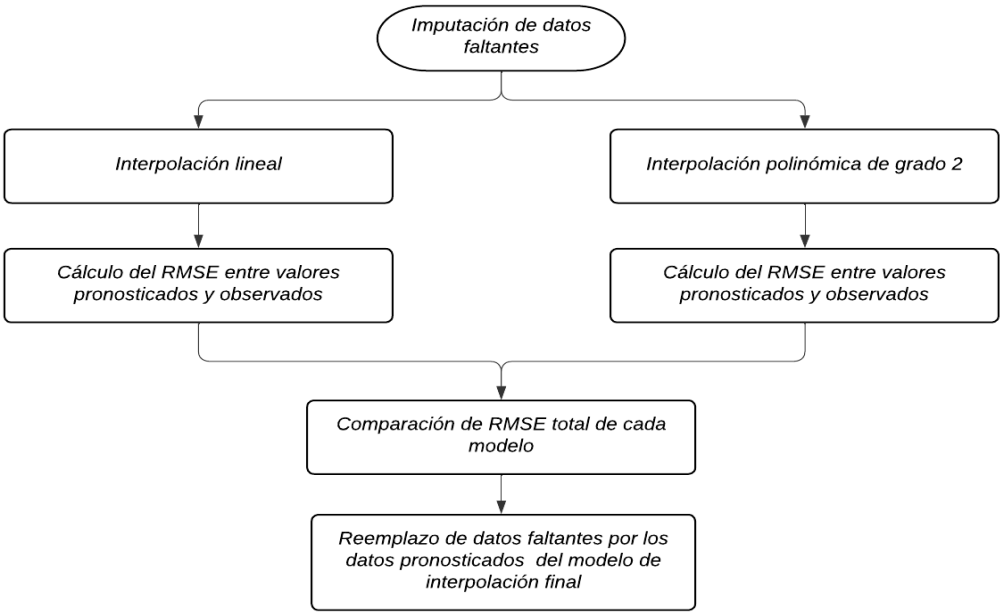
La herramienta de normalización utilizada fue MinMaxScaler, incluida en las bibliotecas de Scikit-learn.

4.2.8. Interpolación de datos faltantes:

A continuación, en la Figura 20 se muestra el diagrama de flujos del desarrollo de esta etapa:

Figura 20

Diagrama de flujo del proceso de la interpolación de datos faltantes:



Nota. Elaboración propia

Para la interpolación, tanto lineal como polinómica, se utilizó la función **interpolate**, incluida en el paquete de la librería Pandas. Esta función permite completar las horas faltantes haciendo una estimación entre los datos ya conocidos.

La función polinómica, por otra parte, permite utilizar distintos grados de polinomios para ajustarse mejor a la serie de tiempo. Para esta investigación, se trabajó con una función polinómica de grado 2, ya que órdenes mayores complicaban el modelo sin aportar mejoras significativas en la reducción del error cuadrático medio (RMSE). Posteriormente, se evaluó el rendimiento de ambos modelos utilizando la métrica del error de raíz cuadrada media (RMSE) para distintos segmentos de tiempo.

4.2.9. Métodos de agrupamiento:

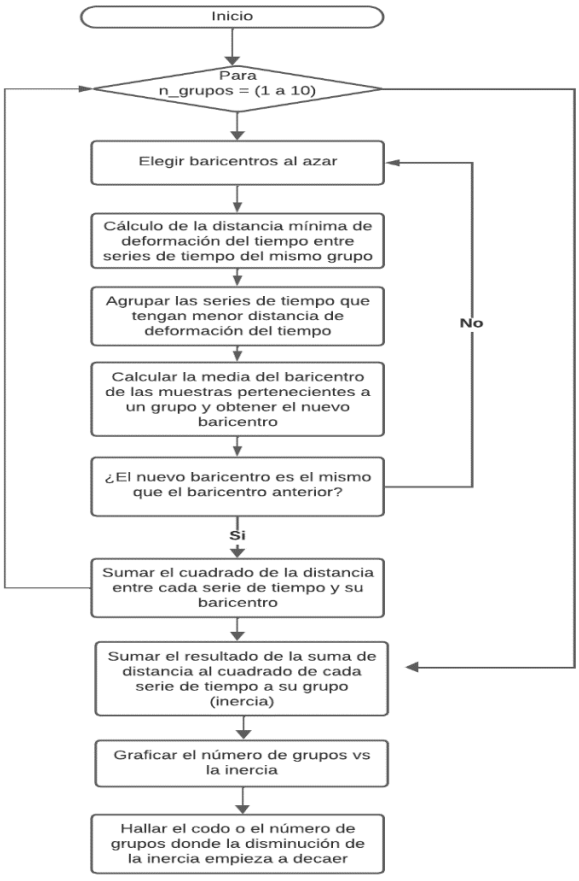
K-Means:

Para encontrar el número óptimo de grupos en K-Means, se utilizó la métrica de la deformación del tiempo (DTW), ya que esta ofrece ventajas superiores al agrupar series de tiempo en comparación con la distancia euclidiana (H. Sakoe, 1978). Su implementación se llevó a cabo mediante la función TimeSeriesKmeans, incluida en las bibliotecas de Tslearn.

Posteriormente, se procedió a calcular el valor de RMSE por cada grupo y se analizaron los resultados utilizando los métodos de codo y de silueta. En la Figura 21 se muestra el diagrama de flujo de este proceso:

Figura 21

Proceso para encontrar el número óptimo de grupos utilizando el método del codo



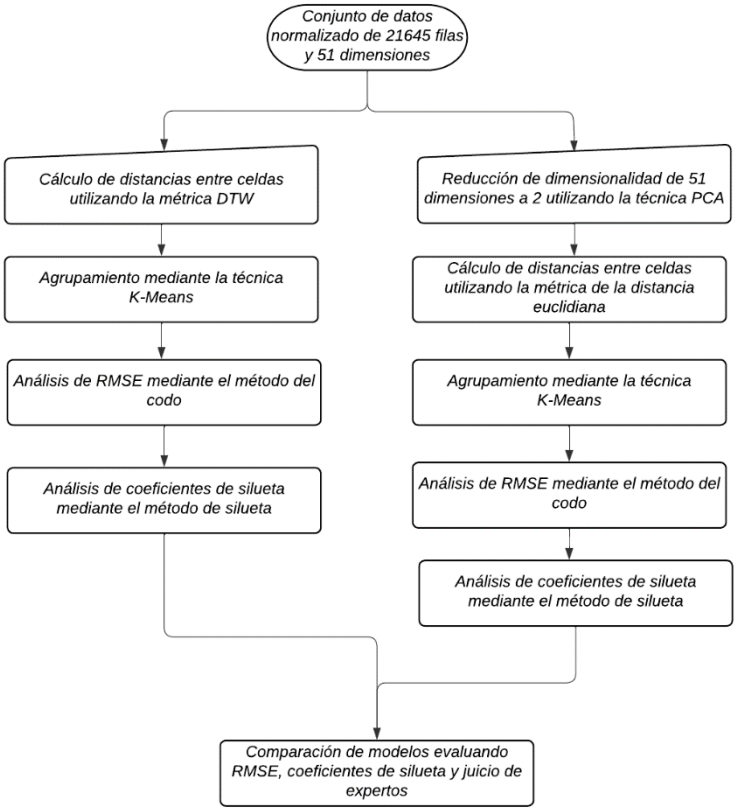
Nota. Elaboración propia

Además, se llevó a cabo otra prueba utilizando la técnica de reducción de dimensiones (PCA) para reducir el conjunto de características de 51 a 2. Se calculó el número de grupos mediante los métodos de codo y de silueta. Finalmente, se compararon ambos modelos: utilizando solo DTW y utilizando solo PCA, encontrándose que ambos tuvieron resultados bastante similares.

La Figura 22 ilustra este proceso en mayor detalle:

Figura 22

Comparación de modelos utilizando DTW y utilizando reducción de la dimensionalidad (PCA)



Nota. Elaboración propia

SOM:

Para encontrar el número óptimo de grupos mediante la técnica de mapas autoorganizados se realizó pruebas con distintos tamaños de mapas y números de neuronas y se halló el error de cuantización asociado a cada combinación. Finalmente, se eligieron 3 combinaciones:

- J. Tian et al (2014) menciona que el número óptimo de grupos debería de ser la raíz cuadrada del total de instancias o filas multiplicado por 5 en este caso con un total de 729 neuronas.

- Por otra parte, existe una "regla de pulgar" que menciona que el valor óptimo de tamaño de mapa de neuronas es de la raíz cuarta del número total de muestras, en este caso correspondería al valor de 147 neuronas que sería de aproximadamente un mapa de 12 x 12 neuronas.

- Finalmente se tiene una elección del número de grupos en el cual el error de cuantización disminuye de manera drástica el cual corresponde a 25 neuronas.

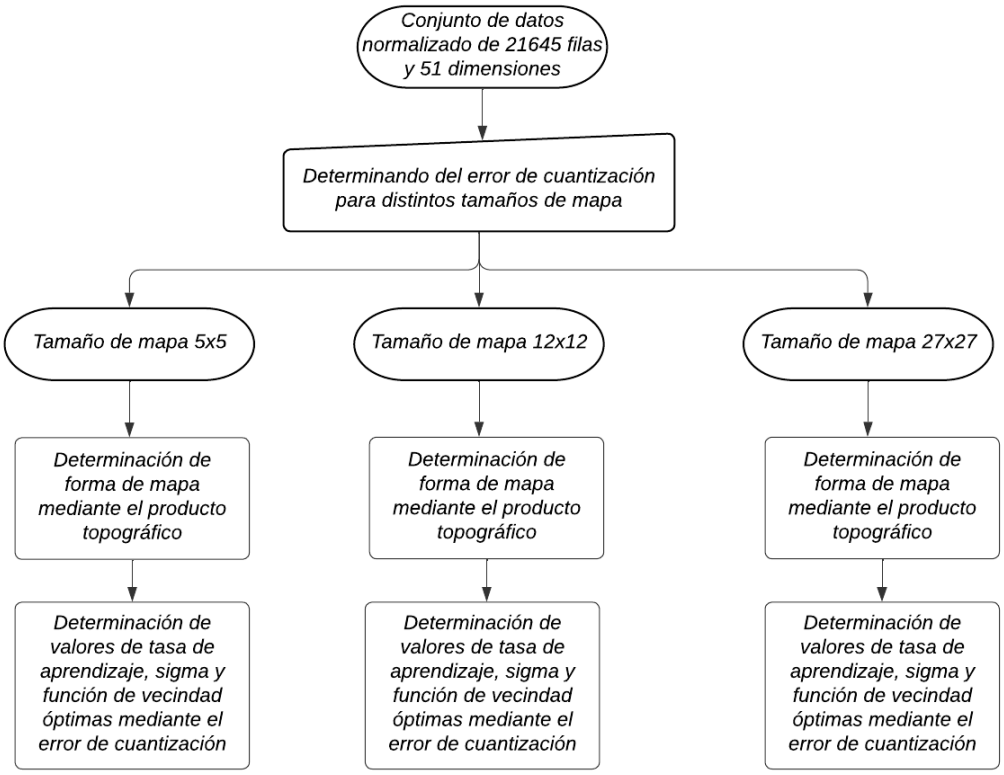
Luego de determinarse las combinaciones que se utilizaran para la comparación con el modelo K-Means se procedió a determinar los tamaños de mapa óptimos para cada combinación que no violen la función de vecindad, para ello se utilizó la métrica del producto topográfico.

Finalmente se realizaron 36 iteraciones en cada combinación de mapa variando los parámetros de tasa de aprendizaje, sigma y función de vecindad con el fin de determinar el modelo con los parámetros óptimos que generen el menor error de cuantización.

La implementación de este modelo fue utilizando las librerías proporcionadas por Minisom. En la Figura 23 se muestra en detalle este proceso:

Figura 23

Diagrama de flujo del modelo SOM



Nota. Elaboración propia

4.2.10. Comparación de rendimiento de modelos

Para evaluar y comparar el rendimiento de cada modelo en la identificación de patrones, se empleó un método extrínseco:

Identificación de patrones de tendencia:

En una primera instancia, se determinó el tipo de agrupamiento más apropiado para analizar los patrones de tendencia. Se realizaron pruebas con distintos intervalos de semanas, evaluando la suma de la varianza total entre las mediciones de cada celda y su centroide en el caso de K-Means o unidad de mejor coincidencia en el caso de SOM.

Posteriormente, se llevó a cabo la categorización del grado de inclinación, calculando la pendiente generada por la regresión lineal. Este valor se convirtió a

porcentaje y se categorizó en función de su magnitud. La Tabla 3 ilustra este proceso de categorización.

Tabla 3

Categorización de umbrales de tendencia.

Valor	Tipo de tendencia
<=-50%	Pendiente negativa muy alta
>-50% y <=-20%	Pendiente negativa alta
>-20% y <=-10%	Pendiente negativa moderada
>-10% y <-5%	Pendiente negativa baja
>=-5% y <=5%	Nula
>5% y <10%	Pendiente positiva baja
>=10% y <20%	Pendiente positiva moderada
>=20% y <50%	Pendiente positiva alta
>=50%	Pendiente positiva muy alta

Nota. Elaboración propia

Finalmente se analizó el tipo de grupos generados por cada modelo K-Means y SOM basados en esta clasificación.

4.3. Pruebas realizadas

En esta fase, se implementaron los Código realizados de los algoritmos de agrupamiento, y se realizaron iteraciones para evaluar los resultados y determinar el mejor modelo de aprendizaje automático. A continuación, se presenta el Código realizado, y en la Tabla 4 se muestra una vista preliminar de los datos generados:

Código realizado:

Importación de librerías importantes:

```
import pandas as pd
from statsmodels.tsa.seasonal import STL
import matplotlib.pyplot as plt
from datetime import datetime
import clickhouse_driver
import clickhouse_connect
import pandas as pd
from clickhouse_driver import connect
from datetime import date
import matplotlib.pyplot as plt
import sklearn
# Native libraries
import os
import sys
import math
# Essential Libraries
import numpy as np
# Preprocessing
from sklearn.preprocessing import MinMaxScaler
# Algorithms
from minisom import MiniSom
from tslearn.barycenters import dtw_barycenter_averaging
from tslearn.clustering import TimeSeriesKMeans
from sklearn.cluster import KMeans

from sklearn.decomposition import PCA
```

Acceso a la base de datos:

```
#Obtenemos la lista de las 10 primeras celdas con menos interrupciones de horas de servicio
conn2 = clickhouse_connect.get_client(
        host = '10.98.***.**',
        port=8123,
        username = '****',
        password = '****')
result = conn2.query_df('SELECT * FROM estival_db.datos_agrupados_2 ORDER BY semana, celda')
```

Tabla 4

Primera vista a los datos de las celdas:

	semana	celda	DL_Traffic_MB
21581	1.00	L28AN0077_1	374992.87
21582	1.00	L28AN0077_2	384496.34
21583	1.00	L28AN0077_3	163504.61
21584	1.00	L28AN0164_1	1129298.21
21585	1.00	L28AN0164_2	456936.47
...	...	...	...
1125471	51.00	L66VA_138_3_MM_3	905593.34
1125472	51.00	L66VA_138_3_MM_4	214343.93
1125473	51.00	L66VA_140_1	839128.28
1125474	51.00	VAL_in_4G_015A_1	80737.43
1125475	51.00	VAL_in_4G_015A_2	159283.79

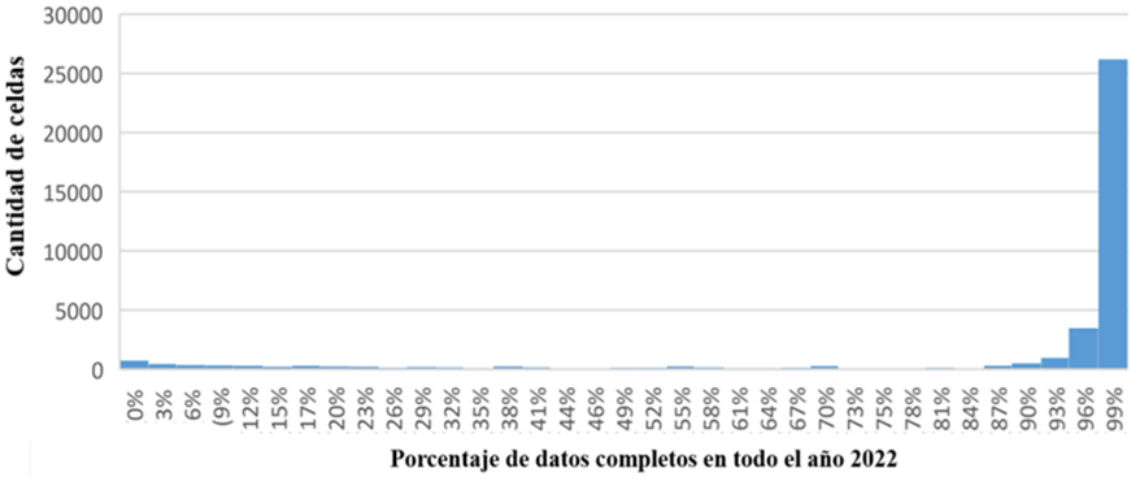
Nota. Elaboración propia basado en la base de datos de WOM.

4.3.1. Objetivo 1: Procesamiento de datos faltantes

En la primera revisión se obtuvo que el número de celdas totales fue de 38183 celdas. En la Figura 24 se muestra de manera gráfica este análisis:

Figura 24

Cantidad de celdas con porcentaje de datos completos en todo el 2022



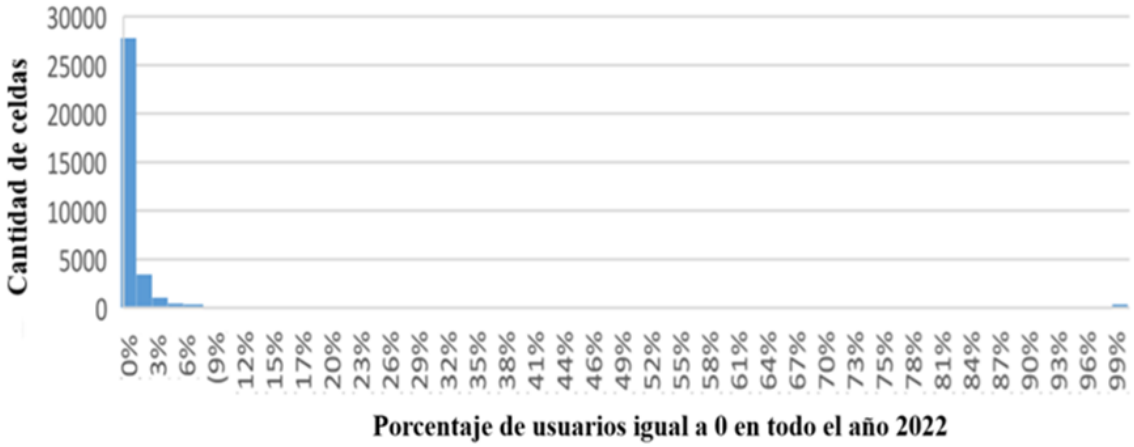
Nota. Elaboración propia basado en la base de datos de WOM.

En un análisis inicial (mediante tabla estadística y diagrama de cajas), se evidencia que las celdas con mediciones completas durante todo el año representan el 71.83% del total de celdas. Además, se realizó un examen específico de las celdas con un número de usuarios igual a cero. Este análisis tiene como objetivo identificar celdas que, aunque estuvieron activas durante todo el año, podrían haber estado apagadas o generando mediciones incorrectas.

En la Figura 25 se muestra de manera gráfica este análisis:

Figura 25

Cantidad de celdas con porcentaje de usuarios igual a 0 en todo el 2022

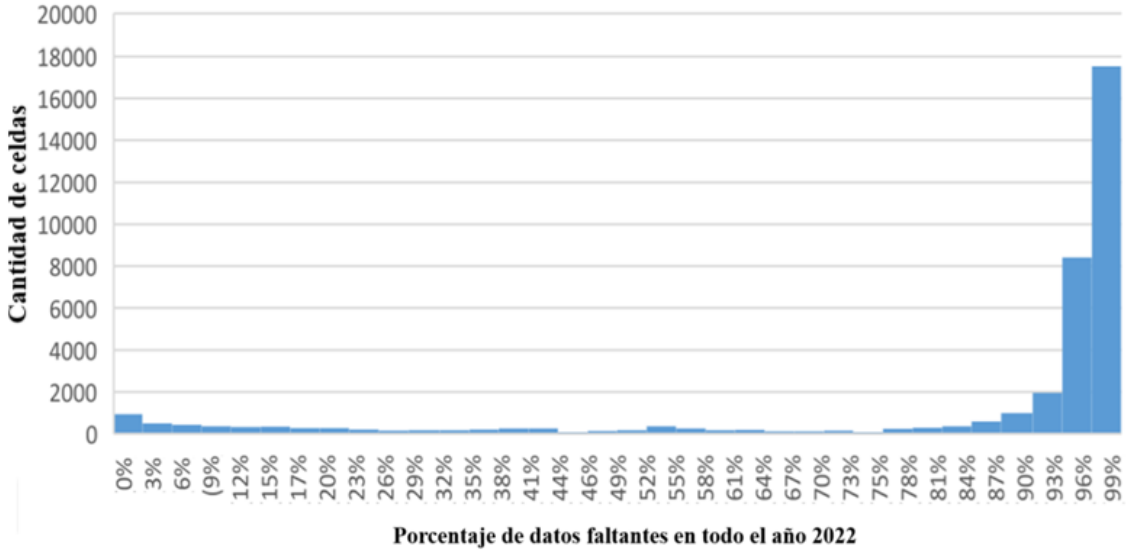


Nota. Elaboración propia basado en la base de datos de WOM.

Finalmente se obtuvo la lista de celdas con porcentaje de datos completo y número de usuarios mayores a 0 en el año 2022. En la Figura 26 se muestra de manera gráfica este análisis:

Figura 26

Cantidad de celdas con porcentaje de datos completos y número de usuarios mayores que 0 en el año 2022.



Nota. Elaboración propia basado en la base de datos de WOM.

4.3.1.1 Técnicas de imputación de datos faltantes:

Se seleccionó aleatoriamente una muestra de 10 sitios que inicialmente contaban con datos completos. Luego, se tomaron 8 segmentos de periodos de tiempo al azar, cada uno con un número variable de horas faltantes, con el objetivo de comparar el rendimiento de las interpolaciones lineales y polinómicas en relación con las horas consideradas.

La Figura 27 ilustra de manera gráfica este análisis:

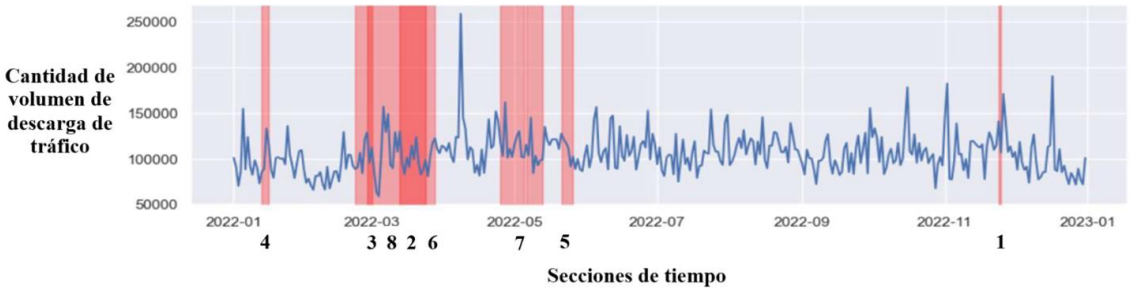
Código realizado ejecutado:

Generación de segmentos de tiempo aleatorios.

```
random.seed(42) #Valor inicial de semilla
ts_missing = []
ts_missing_len = [1,2,3,5,7,10,15,30] #Tamaño de segmentos de datos faltantes en días
for i in range(len(ts_missing_len)):
    start = random.randint(0, len(ts)) #Elegimos periodo de tiempo aleatorios dentro de los datos del 2022
    ts_missing.append((start,start+ts_missing_len[i])) #Tomamos muestra a partir de los periodos de tiempo elegidos

plt.figure(figsize=(12,3))
plt.plot(ts) #Graficamos los segmentos de tiempo
for i,section in enumerate(ts_missing):
    plt.axvspan(ts.index[section[0]],ts.index[section[1]],color='red', alpha=0.3)
    plt.text(ts.index[section[0]],4500,f'Section{i+1}')
plt.show()
```

Figura 27

Cálculo del error de raíz cuadrada media utilizando el método de interpolación lineal para una muestra de 10 celdas.



La Tabla 5 muestra el intervalo de tiempo en días de cada segmento.

Tabla 5

Cantidad de tiempo en días de cada segmento.

Segmentos	Periodo (días)
1	1
2	2
3	3
4	5
5	7
6	10
7	15
8	30

Nota. Elaboración propia

En la Figura 28 y Figura 29 se muestra la aplicación de la interpolación lineal y polinómica respectivamente para una celda.

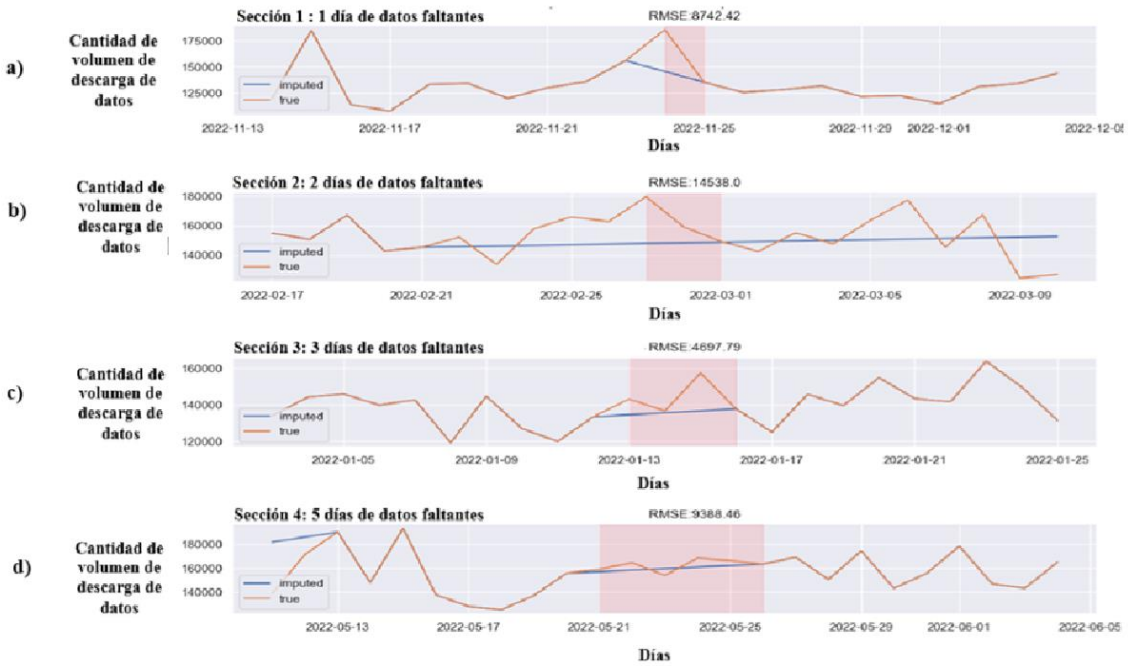
Código realizado ejecutado:

Entrenamiento de datos utilizando la interpolación lineal

```
train['y_mean'] = train['y'].interpolate(method='linear')
train['y_std'] = 0

plot_imputaion(show_ci=False, title='Imputación con interpolación lineal')
```

Figura 28

Cálculo del RMSE aplicando interpolación lineal.



Nota. Elaboración propia basado en la base de datos de WOM. Este gráfico representa el cálculo del RMSE aplicando interpolación lineal para 8 segmentos de tiempo seleccionados aleatoriamente con intervalos de días diferentes en una celda. a) Segmento de 1 día de datos faltantes b) Segmento de 2 días de datos faltantes c) Segmento de 3 días de datos faltantes d) Segmento de 5 días de datos faltantes e) Segmento de 7 días de datos faltantes f) Segmento de 10 días de datos faltantes g) Segmento de 15 días de datos faltantes h) Segmento de 30 días de datos faltantes.

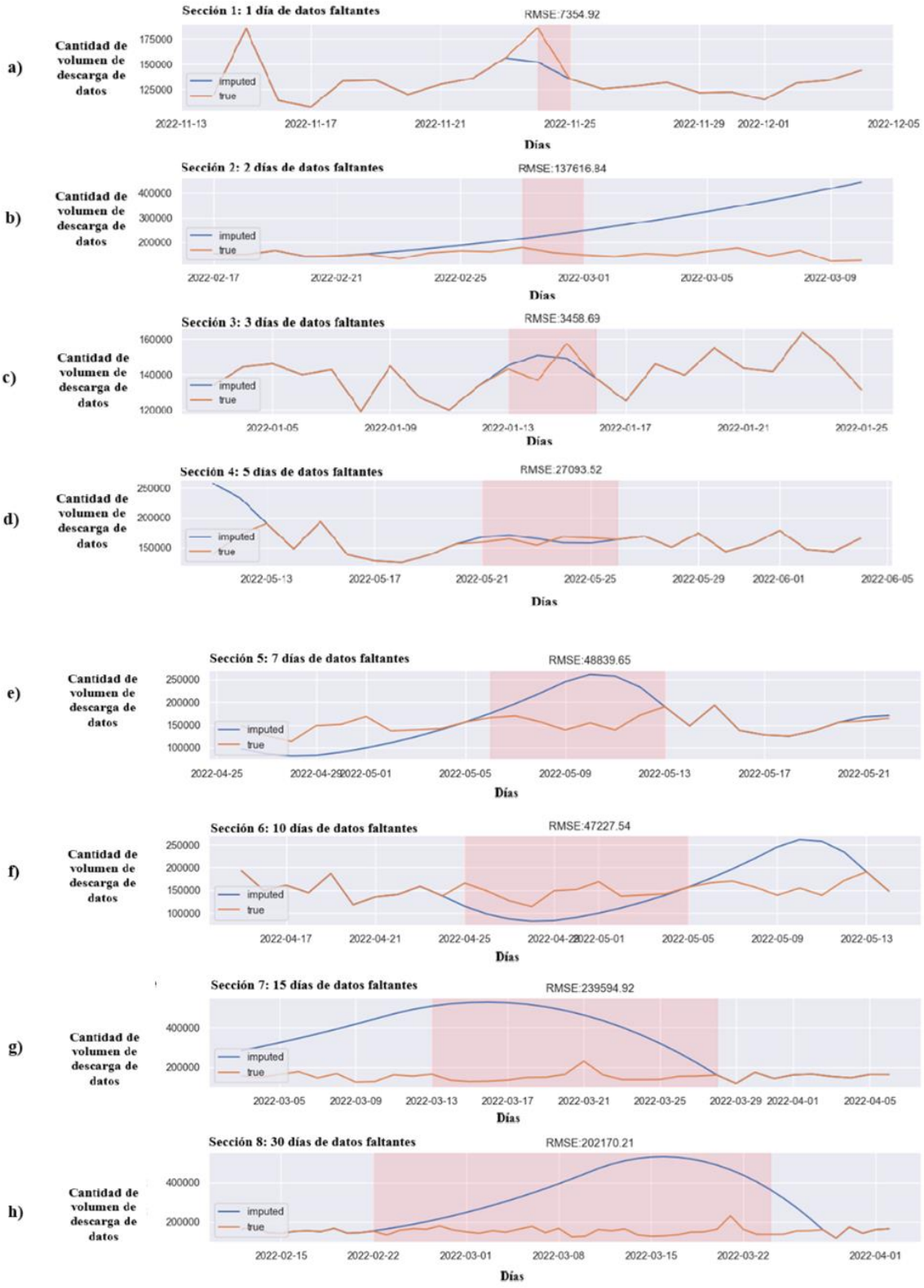
Código realizado ejecutado:

Entrenamiento de datos utilizando la interpolación polinómica

```
train['y_mean'] = train['y'].interpolate(method='polynomial', order=2)
train['y_std'] = 0

plot_imputaion(show_ci=False, title='Imputation with polynomial interpolation')
```

Figura 29

Cálculo del RMSE aplicando interpolación polinómica de grado 2.



Nota. Elaboración propia basado en la base de datos de WOM. Este gráfico representa el cálculo del RMSE aplicando interpolación polinómica de grado 2 para 8 segmentos de tiempo seleccionados aleatoriamente con intervalos de días diferentes en una celda. a) Segmento de 1 día de datos faltantes b) Segmento de 2 días de datos faltantes c) Segmento de 3 días de datos faltantes d) Segmento de 5 días de datos faltantes e) Segmento de 7 días de datos faltantes f) Segmento de 10 días de datos faltantes g) Segmento de 15 días de datos faltantes h) Segmento de 30 días de datos faltantes.

En la Tabla 6 se muestra la comparación del RMSE generado entre el método de interpolación lineal y método de interpolación polinómica de grado 2.

Tabla 6

RMSE utilizando el método lineal y polinómica para una muestra de 10 celdas.

	RMSE Interpolación Lineal	RMSE Interpolación Polinómica
Celda 1	103004	713351
Celda 2	32690	55335
Celda 3	115413	144888
Celda 4	121625	340389
Celda 5	136411	187552
Celda 6	142759	856047
Celda 7	107230	697488
Celda 8	46073	185388
Celda 9	98884	158020
Celda 10	151672	1166828

Nota. Elaboración propia basado en la base de datos de WOM

De los resultados obtenidos se observa que la interpolación lineal tiene menor RMSE comparado a la interpolación polinómica. Por lo tanto, la técnica usada para imputar los datos faltantes fue la técnica de la interpolación lineal.

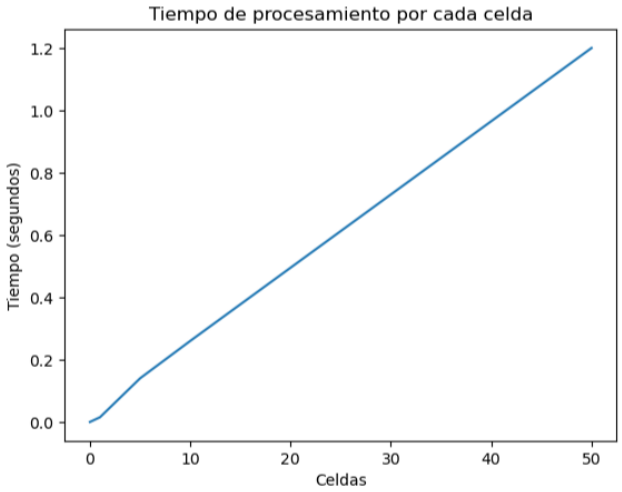
4.3.2. Objetivo 2: Agrupamiento de datos

Tiempo de procesamiento:

Si se quiere procesar la cantidad de mediciones por hora en cada celda entonces se requerirá una gran capacidad de procesamiento y tomará bastante tiempo hacerlo. En la Figura 30 se muestra el tiempo que toma procesar las mediciones anuales en una celda:

Figura 30

Tiempo de procesamiento por cada celda



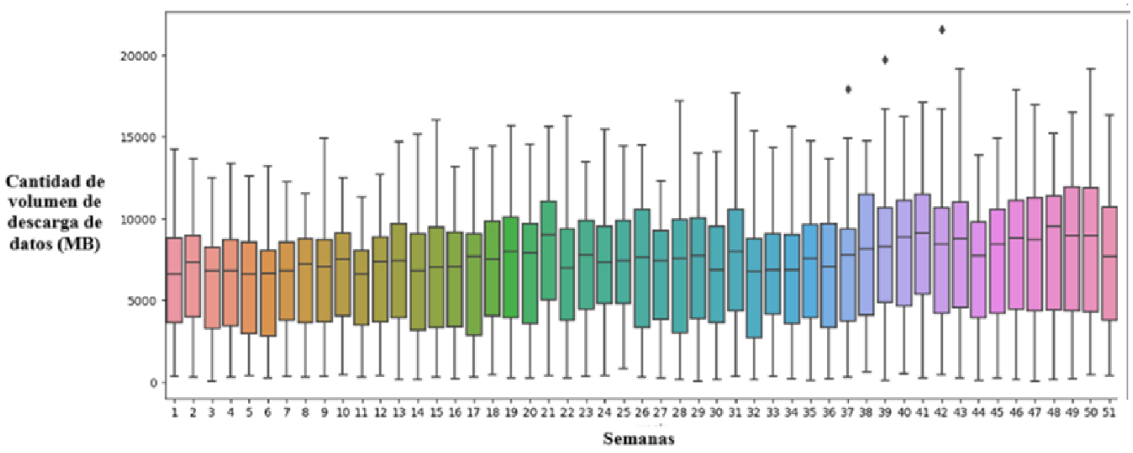
Nota. Elaboración propia basado en la base de datos de WOM

Dispersión de datos semanal:

En los siguientes diagramas de cajas se muestran las distribuciones de las mediciones de la descarga de datos agrupados de manera semanal. En la Figura 31 se muestra de manera gráfica este análisis:

Figura 31

Gráfico de cajas de cada celda de una semana a lo largo del 2022

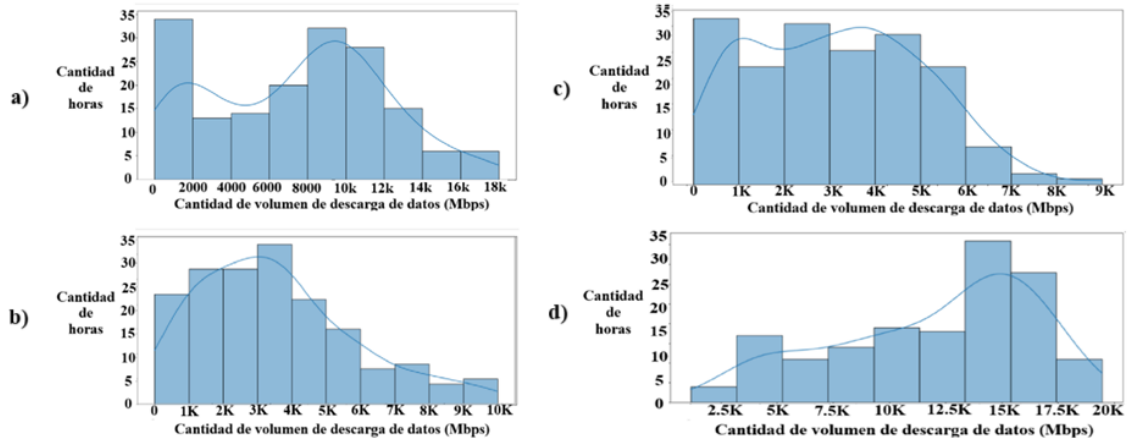


Nota. Elaboración propia basado en la base de datos de WOM

Un análisis exploratorio de algunas semanas tomadas al azar de una celda muestra de manera cercana los tipos de distribución que toman las mediciones de la descarga de datos. En la Figura 32 se muestra este análisis:

Figura 32

Histograma de algunas muestras semanales tomadas al azar de una celda.

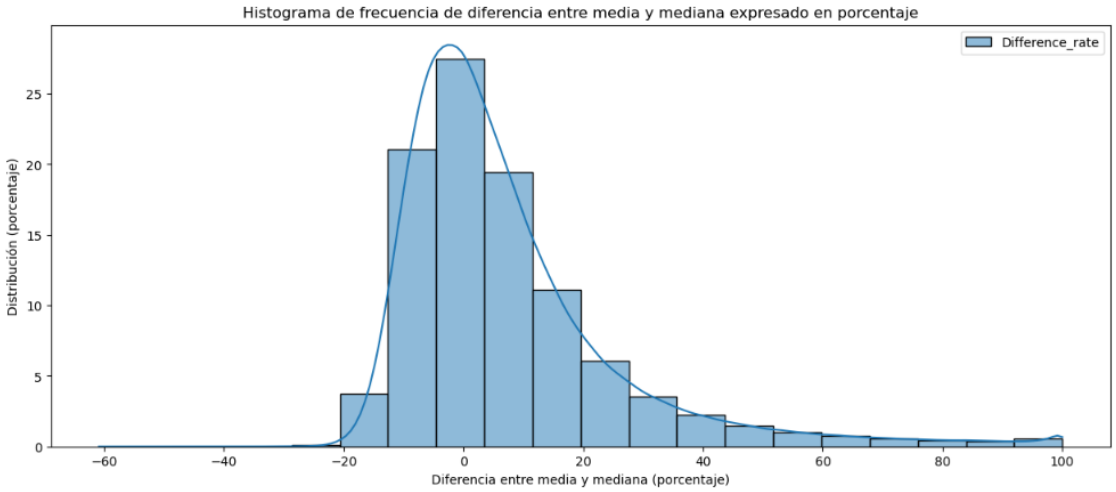


Nota. Elaboración propia basado en la base de datos de WOM. Este gráfico representa las diferentes distribuciones estadísticas de semanas tomadas al azar en una celda. a) Distribución estadística de semana 1 b) Distribución estadística de semana 2 c) Distribución estadística de semana 3 d) Distribución estadística de semana 4.

Del análisis presentado se concluye que cada semana agrupa mediciones de descarga de datos con distintas distribuciones estadísticas. La Figura 33 muestra el histograma de la diferencia porcentual que existe entre la media y la mediana en cada una de las semanas de cada celda que conforman la red LTE. En la Figura 33 se presenta esta distribución.

Figura 33

Porcentaje del número de veces que se repiten los valores de la diferencia entre la media y la mediana.



Nota. Elaboración propia basado en la base de datos de WOM

De la Figura 33 se concluye que aproximadamente el 30% del total de las semanas de las celdas presentan una distribución normal o casi normal con un margen de error del +-4% de diferencia porcentual entre la media y la mediana. También existe un 27% de todas las semanas de la red cuya mediana difiere con la media de datos en más del 12% el valor de la media lo cual significa que hay preponderancia de valor bajos en esas semanas. Finalmente, un 43% del total de semanas de toda la red presenta valores extremadamente altos que aumentan el valor de la media de esas semanas en más del 12%.

Agrupamiento por media y mediana

Por lo tanto, se concluye que la media o la mediana no representa correctamente los datos de las celdas cuando se quiere realizar la agrupación a nivel semanal ya que solo la tercera parte del total de semanas presentan distribución normal mientras que en el resto de las semanas existe presencia de valores extremos que sesgan la media o mediana.

Agrupamiento por suma

La agrupación de datos por suma representa una mejor alternativa que las demás técnicas de agrupación de datos por los siguientes motivos:

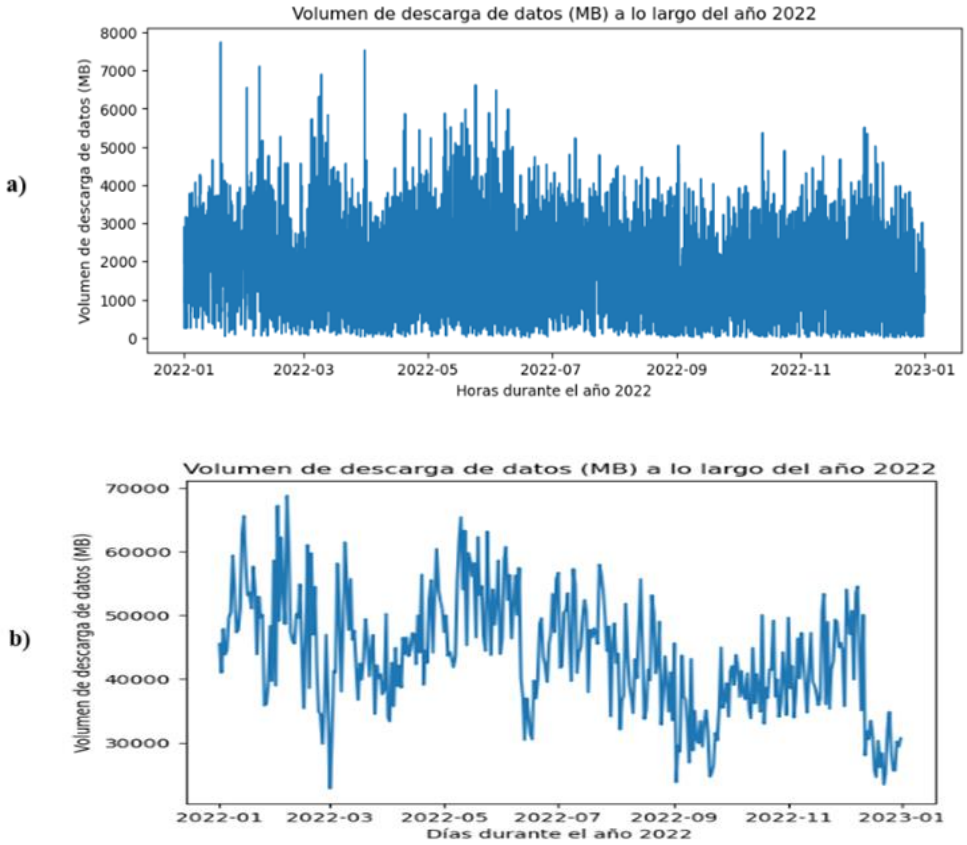
- Debido a que el objetivo de esta tesis es identificar patrones de cada celda, la suma del volumen de la descarga de datos representa una mejor alternativa frente a las otras herramientas estadísticas para analizar las tendencias semanales.
- Se desea observar cuanto volumen de descarga de datos en total ha hecho una celda a lo largo del año. Por lo tanto, la agrupación por suma compara mejor el volumen total de descarga de datos sobre varias semanas.

Granularidad de datos

En la Figura 34 se hace una comparación de la tendencia de descarga de datos utilizando distintas granularidades (horario, diario, semanal y mensual).

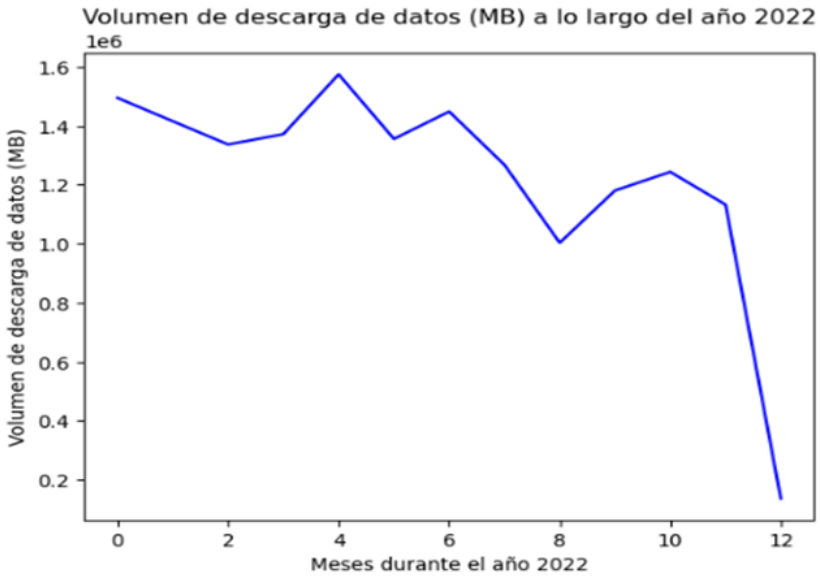
Figura 34

Comparación de diferentes granularidades de datos



c)

d)



Nota. Elaboración propia basado en la base de datos de WOM. Este gráfico representa las diferentes formas de tendencia generadas al usar cada tipo de granularidad. a) Granularidad horaria b) Granularidad diaria c) Granularidad semanal d) Granularidad mensual.

En la Tabla 7 se muestra una tabla comparativa respecto a cada tipo de granularidad:

Tabla 7

Comparación de ventajas y desventajas de cada uno de los tipos de granularidades

Granularidad	Ventajas	Desventajas
Horario	Útil para análisis de comportamiento horario	Demasiado costoso computacionalmente
Diaria	Útil para análisis de comportamiento diario	Costoso computacionalmente
Semanal	Bajo nivel de procesamiento y útil para análisis de tendencia semanal y estacionalidad mensual	Se pierde información de las mediciones obtenidas por hora debido al muestreo semanal.
Mensual	Útil para análisis de ciclicidad	Para realizar análisis de baja granularidad, es necesario utilizar datos recopilados a lo largo de varios años, ya que la información de un solo año resulta poco útil.

Nota. Elaboración propia

En la Tabla 8 se muestra una comparativa del tamaño de datos que se maneja por cada nivel de granularidad:

Tabla 8

Cantidad de registro por cada tipo de granularidad:

Granularidad	Cantidad de registros por celda	Cantidad de registros en total	Cantidad de dimensiones
Horaria	8760	18.9 Millones	8760
Diaria	365	7.9 Millones	365
Semanal	51	1.12 Millones	51
Mensual	12	0.26 Millones	12

Nota. Elaboración propia basado en la base de datos de WOM

En esta tesis se trabajó con la granularidad por semanas debido a los siguientes motivos:

- La granularidad por día es útil cuando se requiere revisar el rendimiento a nivel diario o semanal pero no es óptimo cuando se requiere hacer un análisis anual, por otra parte, para realizar este análisis a nivel macro se requiere tener una capacidad computacional demasiado alta.
- La granularidad por semana es útil cuando se requiere conocer la razón de cambio de la tendencia a nivel semanal o mensual.

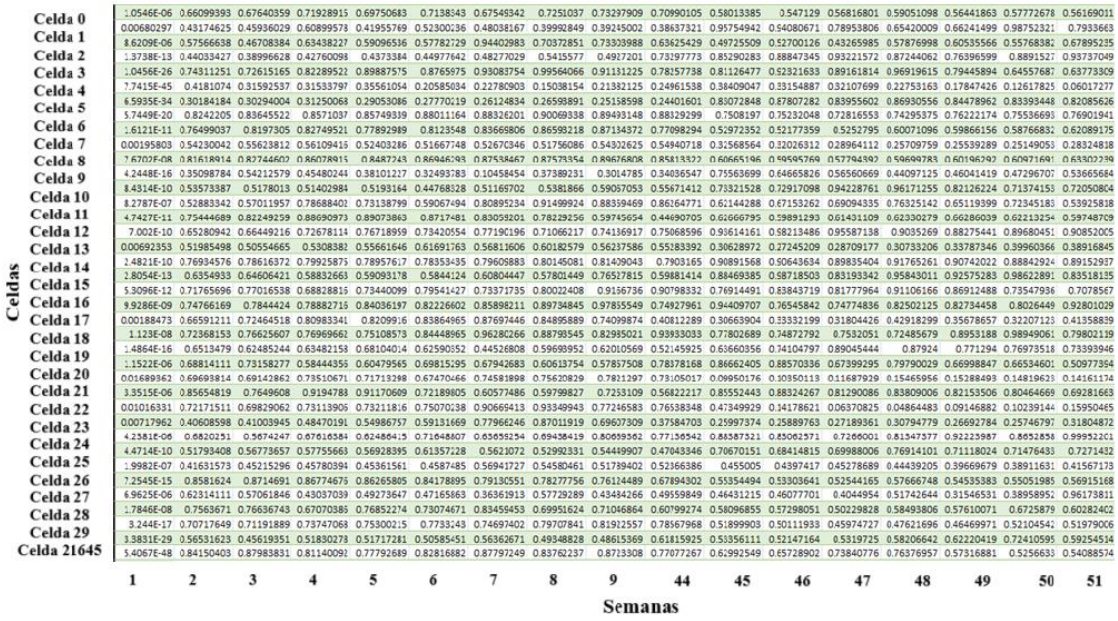
- La granularidad por meses es útil cuando se requiere realizar análisis a mensual o anual, así como también otros factores tales como estacionalidad y ciclicidad.

4.3.3. Algoritmos de agrupamiento

La Figura 35 muestra una vista previa de los datos filtrados y preprocesados que se introdujeron en los algoritmos de agrupamiento:

Figura 35

Conjunto final de datos listo para ser utilizado en los algoritmos de agrupamiento.



Nota. Elaboración propia basado en la base de datos de WOM

4.3.3.1. K-Means:

DTW:

Se utilizo la métrica DTW para calcular la distancia existente entre cada serie de tiempo. En la Figura 36 se muestra un ejemplo de este proceso en el cual se compararon 2 series de tiempo:

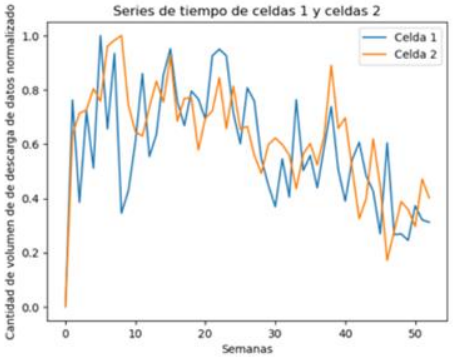
Figura 36

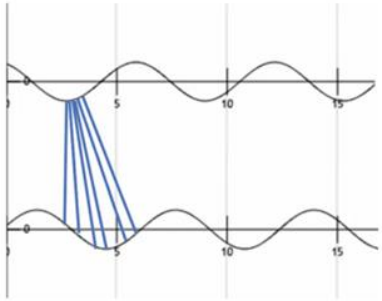
Método de Deformación dinámica del tiempo

a)

Semanas	1	2	3	4	5	6		43	44	45	46	47	48	49	50	51	52	53
Serie 1	0.312318	0.320769	0.372803	0.245431	0.268914	0.26652	....	0.609502	0.428695	0.345616	0.934157	0.656298	1	0.511069	0.721587	0.385939	0.762267	0
Serie 2	0	0.636625	0.713756	0.726064	0.804802	0.758239		0.325197	0.399243	0.619139	0.446262	0.171786	0.279365	0.388366	0.35837	0.29591	0.470906	0.402437

b)





c)

Nota. Elaboración propia. La Figura 36 a muestra el conjunto de valores de 2 series de tiempo diferentes, la Figura 36 b muestra el desfase que existe entre estas 2 variables. La Figura 36 c muestra como es el proceso de cálculo de distancias comparando cada característica de una serie de tiempo con las características de las otras series de tiempo.

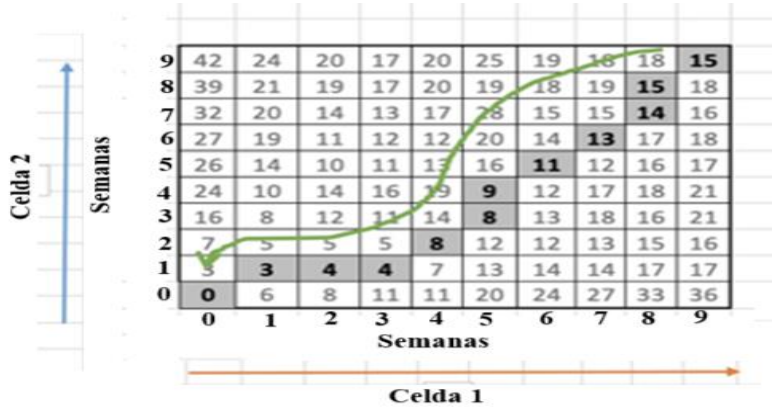
En la Figura 37 se muestra como fue el cálculo de costos y su asociación en la matriz de costos y el valor final de distancia. Este proceso se repitió iterativamente a través del algoritmo implementado en Python:

Código realizado:

```
km = TimeSeriesKMeans(n_clusters=4, metric="dtw")

labels = km.fit_predict(dfs8)
```

Figura 37

Cálculo de matriz de costos



Nota. Elaboración propia basado en la base de datos de WOM. La Figura 37 representa la matriz de costos asociada a la comparación de series de tiempo. El valor de distancia final entre la serie de tiempo 1 y 2 resulto ser de 7.21.

Método del codo

Primeramente, se utilizó la técnica del método del codo manteniendo las 51 dimensiones. En la Figura 38 se muestra de manera gráfica este análisis:

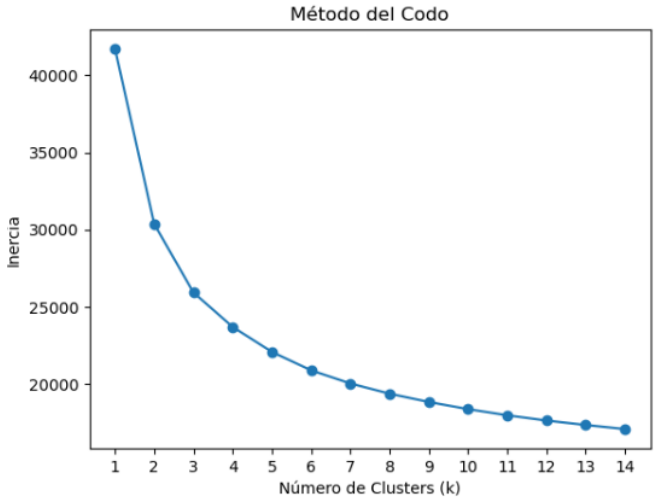
Código realizado:

Código realizado para determinar el número óptimo de grupos utilizando el método del codo

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs

# Inicializar una lista para almacenar los valores de inercia
inertia = []

# Probar diferentes valores de k (número de clusters)
k_values = range(1, 7)
for k in k_values:
    # Aplicar K-means para cada valor de k
    kmeans = TimeSeriesKMeans(n_clusters=k, metric="dtw")
    kmeans.fit(dfs8)
    inertia.append(kmeans.inertia_)

# Graficar los valores de inercia en función de k
plt.plot(k_values, inertia, 'o-')
plt.xlabel('Número de Clusters (k)')
plt.ylabel('Inercia')
plt.title('Método del Codo')
plt.xticks(k_values)
plt.show()
```

Figura 38

Método del codo utilizando 51 dimensiones

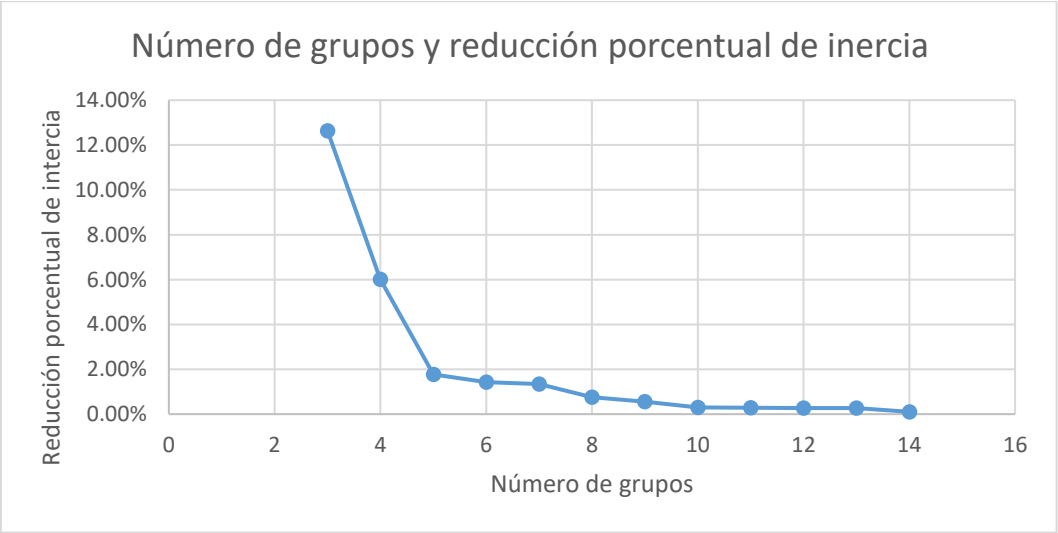


Nota. Elaboración propia.

En la Figura 39 y en la Tabla 9 se muestra de mejor manera como el porcentaje de reducción de inercia disminuyó desde un 6% para 4 grupos hasta un valor de 1.77% para un número de 5 grupos.

Figura 39

Número de grupos y reducción porcentual de la inercia.



Nota. Elaboración propia.

Tabla 9

Reducción de inercia vs número de grupos.

Inercia	K (Grupos)	Disminución de inercia	Porcentaje	Disminución porcentual
41727.71	1			
30359.27	2	11368.44	27.24%	
25922.82	3	4436.45	14.61%	12.63%
23692.11	4	2230.71	8.61%	6.01%
22072.32	5	1619.79	6.84%	1.77%
20878.13	6	1194.19	5.41%	1.43%
20028.71	7	849.42	4.07%	1.34%
19364.46	8	664.25	3.32%	0.75%
18828.82	9	535.64	2.77%	0.55%
18364.98	10	463.84	2.46%	0.30%
17966.11	11	398.87	2.17%	0.29%
17624.74	12	341.37	1.90%	0.27%
17337.33	13	287.41	1.63%	0.27%
17072.61	14	264.72	1.53%	0.10%

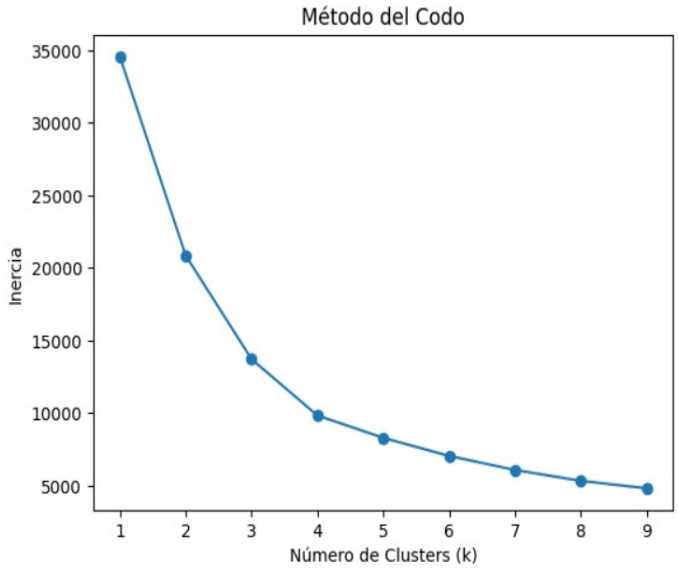
Nota. Elaboración propia

Método del codo utilizando PCA con 2 dimensiones:

En la Figura 40 se muestra el resultado de utilizar el método del codo luego de reducir las 51 dimensiones a solo 2 utilizando el método de PCA(Principal Component Analysis).

Figura 40

Método del codo utilizando PCA



Nota. Elaboración propia

Método de Silueta:

El método de silueta está definido para cada muestra y consta de dos puntuaciones, siendo un coeficiente de silueta más alto indicativo de un modelo con grupos mejor definidos:

a. La distancia media entre una muestra y todos los demás puntos en la misma clase. Esta puntuación mide la cercanía de los puntos en el mismo grupo.

b. La distancia media entre una muestra y todos los puntos en el siguiente grupo más cercano. Esta puntuación mide la distancia entre puntos de diferentes grupos.

La Figura 41 muestra cómo se distribuye el coeficiente de silueta en todos los grupos y cómo cambia el valor de la distancia media a medida que el número de grupos aumenta:

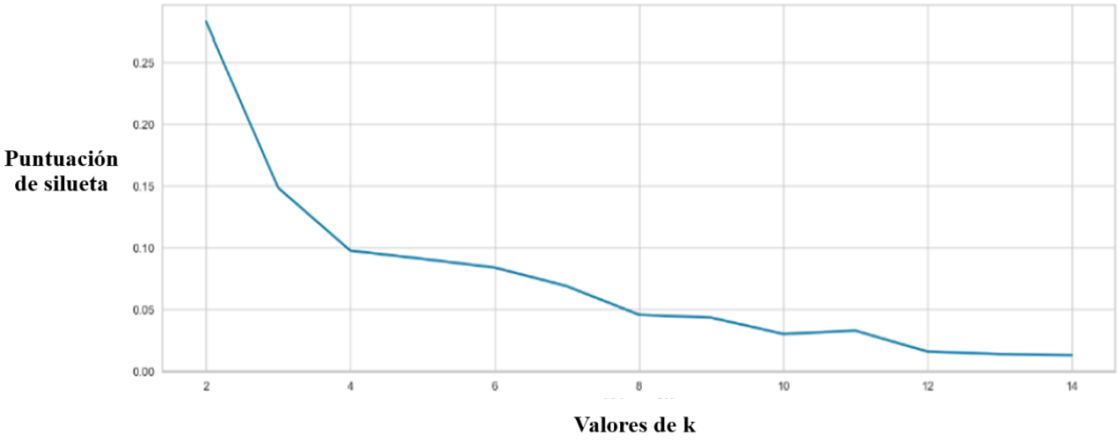
Código realizado:

Determinación de grupos mediante el método de silueta

```
from sklearn.metrics import silhouette_score#For the silhouette method k needs to start from 2
K = range(2,15)
silhouettes = []#Fit the method
for k in K:
    km = TimeSeriesKMeans(n_clusters=k, metric="dtw")
    km.fit_predict(dfs8)
    silhouettes.append(silhouette_score(dfs8, km.labels_))
    #Plot the results
fig = plt.figure(figsize= (15,5))
plt.plot(K, silhouettes, 'bx-')
plt.xlabel('Valores de K')
plt.ylabel('Puntuación de Silueta')
plt.title('Método de Silueta')
plt.grid(True)
plt.show()
```

Figura 41

Número de grupos mediante el método de Silueta



Nota. Elaboración propia

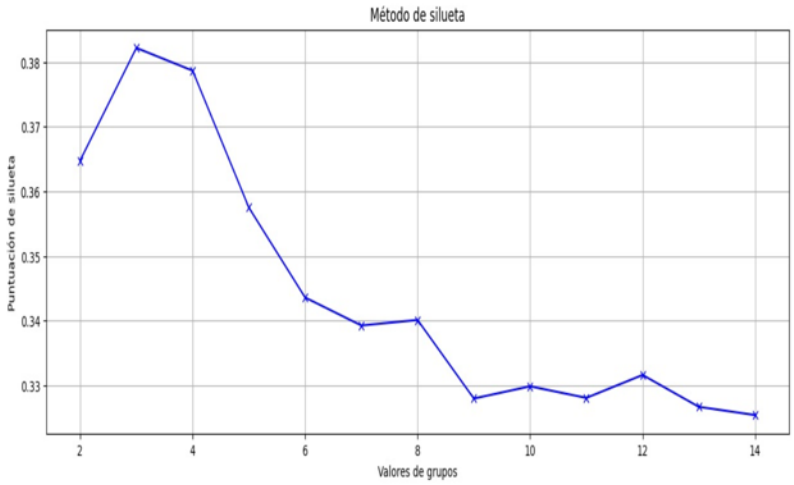
Método de silueta utilizando PCA con 2 dimensiones:

En la Figura 42 se muestra el resultado de utilizar el método de silueta luego de reducir las 51 dimensiones a solo 2 utilizando el método de PCA (Principal Component Analysis).

Figura 42

Método de silueta utilizando PCA

K-Means – Método de silueta (Utilizando PCA=2)



Puntuación de Silueta	Número de grupos
0.365	2
0.382	3
0.379	4
0.358	5
0.344	6
0.339	7
0.340	8
0.327	9
0.330	10
0.328	11
0.329	12
0.325	13
0.325	14

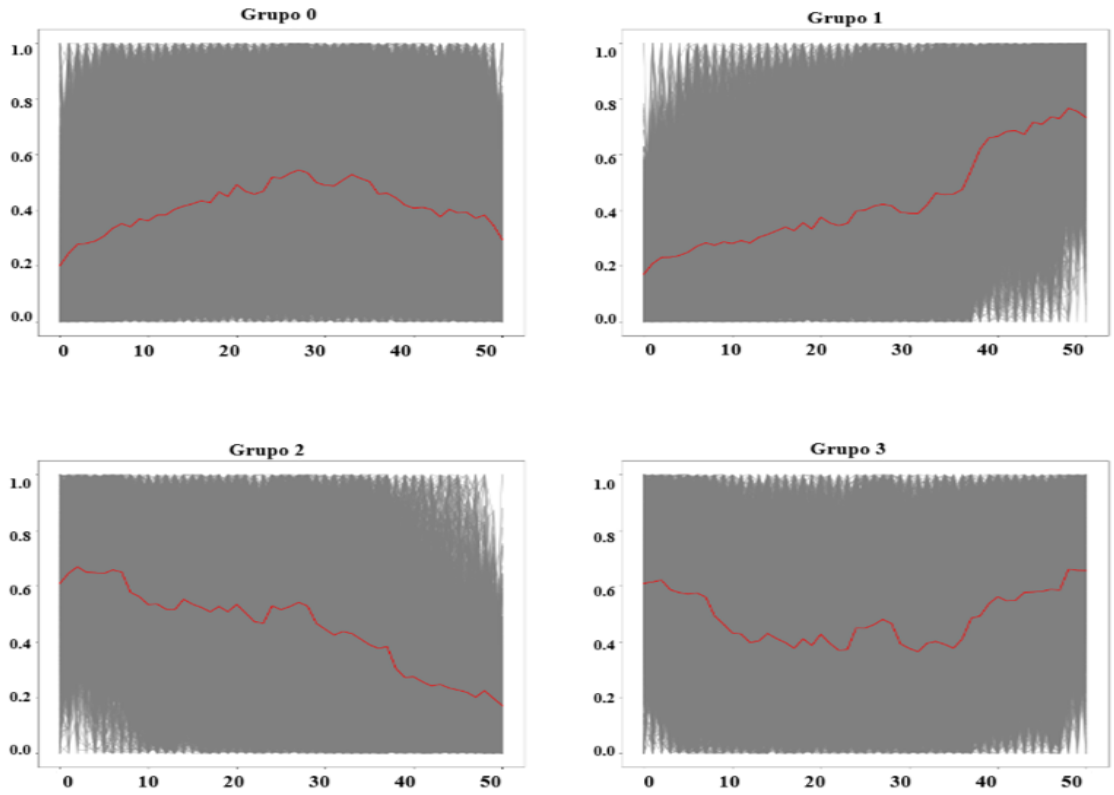
Nota. Elaboración propia

Agrupamiento de celdas con K=4.

De los resultados presentados, se concluyó que el número óptimo de grupos utilizando el algoritmo K-Means varía entre 3 y 4. Es importante destacar que la diferencia entre ambos valores es bastante mínima en comparación con las diferencias observadas con otros números de grupos. Por este motivo, se optará por utilizar el valor de 4 para efectos de pruebas. El resultado muestra una división del conjunto de datos de entrada en 4 grupos principales, y la distribución de las celdas se presenta en la Figura 43:

Figura 43

Agrupamiento de datos K-Means con K=4

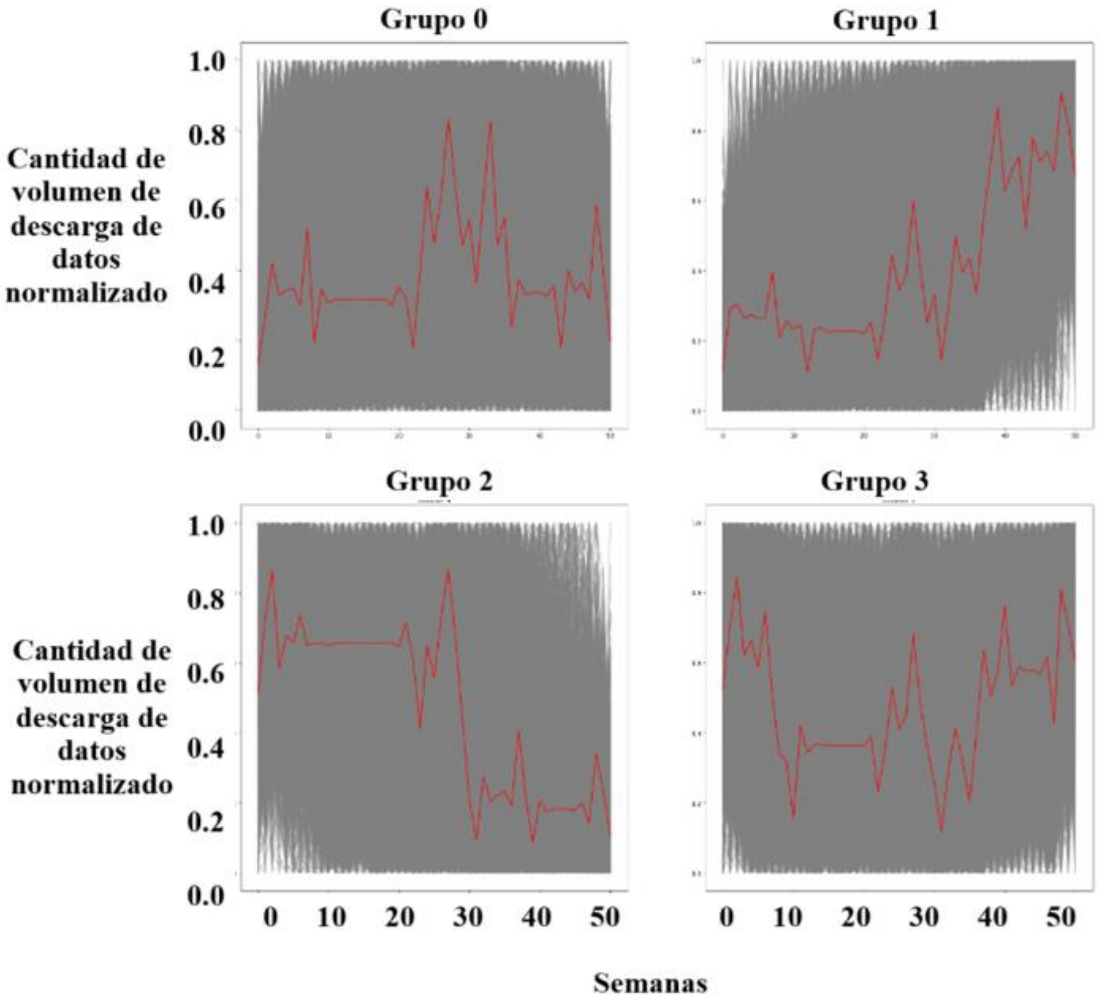


Nota. Elaboración propia.

Por otra parte, se realizó el mismo análisis, pero esta vez utilizando la función de baricentro de promedio DTW con el fin de analizar si mejora la visualización de la línea de tendencia central del grupo. En la Figura 44 se ilustra este enfoque:

Figura 44

Agrupamiento de datos K-Means con K=4 y utilizando la función de baricentro de promedio DTW

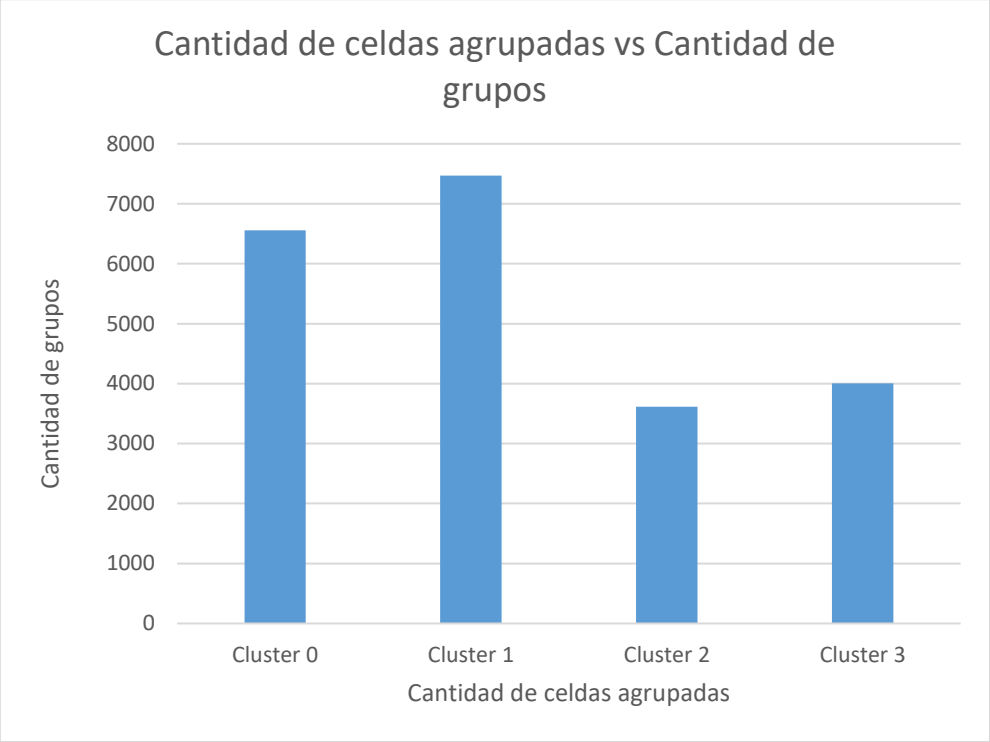


Nota. Elaboración propia

Finalmente se obtuvo la distribución del número de celdas contenidas en cada grupo la cual se muestra en la Figura 45:

Figura 45

Distribución de celdas en cada grupo utilizando K-Means



Nota. Elaboración propia

Identificación de celdas atípicas:

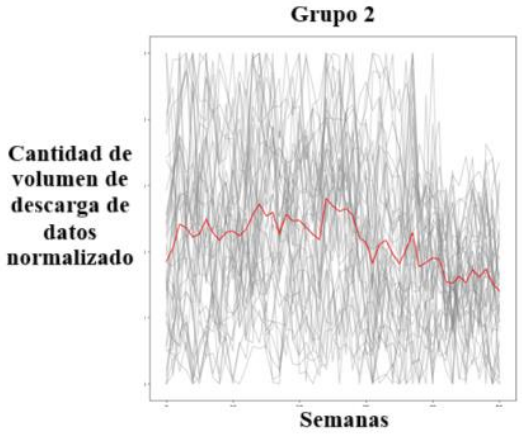
Se identificaron las celdas atípicas en cada grupo mediante dos técnicas principales: la Deformación dinámica del Tiempo (DTW) y la mediana móvil.

Identificación de celdas atípicas utilizando DTW:

Se inició tomando una muestra de 100 celdas del Grupo 2 y calculando la distancia DTW entre cada celda del grupo y su centroide. Este proceso permitió medir el grado de similitud entre cada celda dentro del grupo. La Figura 46 ilustra este procedimiento aplicado al Grupo 2, donde se tomó una muestra de 100 celdas.

Figura 46

Clustering K-Means tomando como muestra 100 celdas del grupo 2

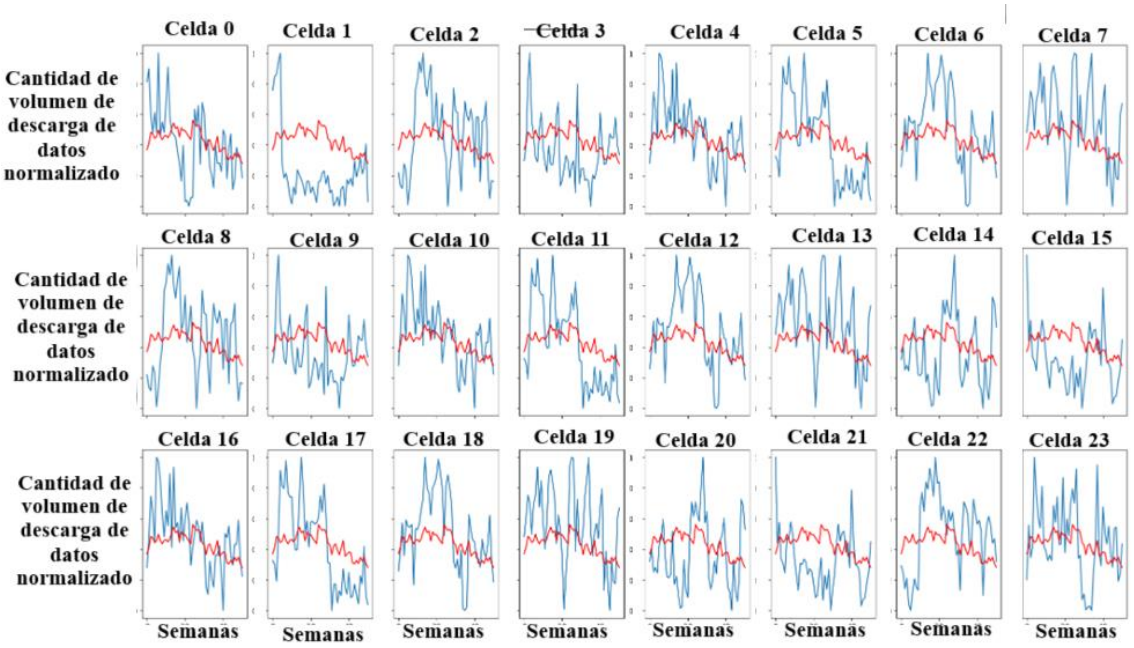


Notas: Elaboración propia

En la Figura 47 se ilustra una comparación entre cada celda respecto a su centroide:

Figura 47

Comparación entre cada celda respecto a su centroide:



Notas: Elaboración propia

Se aplicó la técnica DTW para calcular la distancia entre cada celda y su centroide. Los resultados de esta métrica DTW para las 24 celdas se presentan en la Tabla 10:

Código realizado:

Utilización de la librería Fastdtw para el cálculo del coeficiente DTW.

```
from fastdtw import fastdtw
for i in range(len(cluster_0)):
    dtw_distance, warp_path = fastdtw(Centroide_0, cluster_0[i])
    print(dtw_distance)
```

Tabla 10

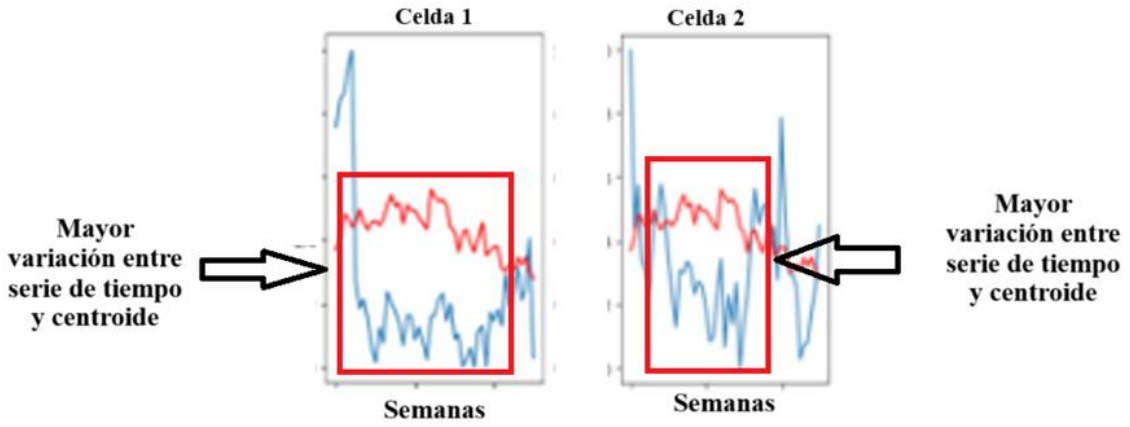
Calculó del coeficiente DTW entre cada celda respecto a su centroide

Celdas	DTW
Celda 0	7.751046596
Celda 1	16.06500372
Celda 2	11.4066624
Celda 3	6.85789941
Celda 4	6.887902323
Celda 5	9.584981927
Celda 6	8.689523808
Celda 7	10.53222487
Celda 8	10.04370927
Celda 9	7.160898872
Celda 10	10.47530837
Celda 11	8.940597003
Celda 12	9.19082199
Celda 13	9.232405951
Celda 14	7.123986896
Celda 15	16.32875232
Celda 16	8.103553996
Celda 17	8.595995806
Celda 18	8.324968985
Celda 19	8.455857426
Celda 20	8.456232455
Celda 21	9.026259229
Celda 22	6.899929289
Celda 23	9.383802768

Nota. Elaboración propia. Se observa que las celdas 1 y 15 presentan los valores más altos de distancia DTW, por lo tanto, se clasifican como celdas atípicas. Un análisis detallado revela una notoria disparidad entre las celdas 1 y 15 en comparación con las demás. La Figura 48 proporciona un enfoque más detallado de este análisis:

Figura 48

Diferencia entre la celda 1 y la celda 15 respecto a su centroide.



Nota. Elaboración propia

Identificación de celdas atípicas utilizando mediana móvil:

Se identificaron las semanas atípicas en cada serie de tiempo mediante la técnica de la mediana móvil. Se llevaron a cabo pruebas utilizando una ventana de 6 semanas. La Figura 49 exhibe el número total de semanas atípicas detectadas en cada celda.

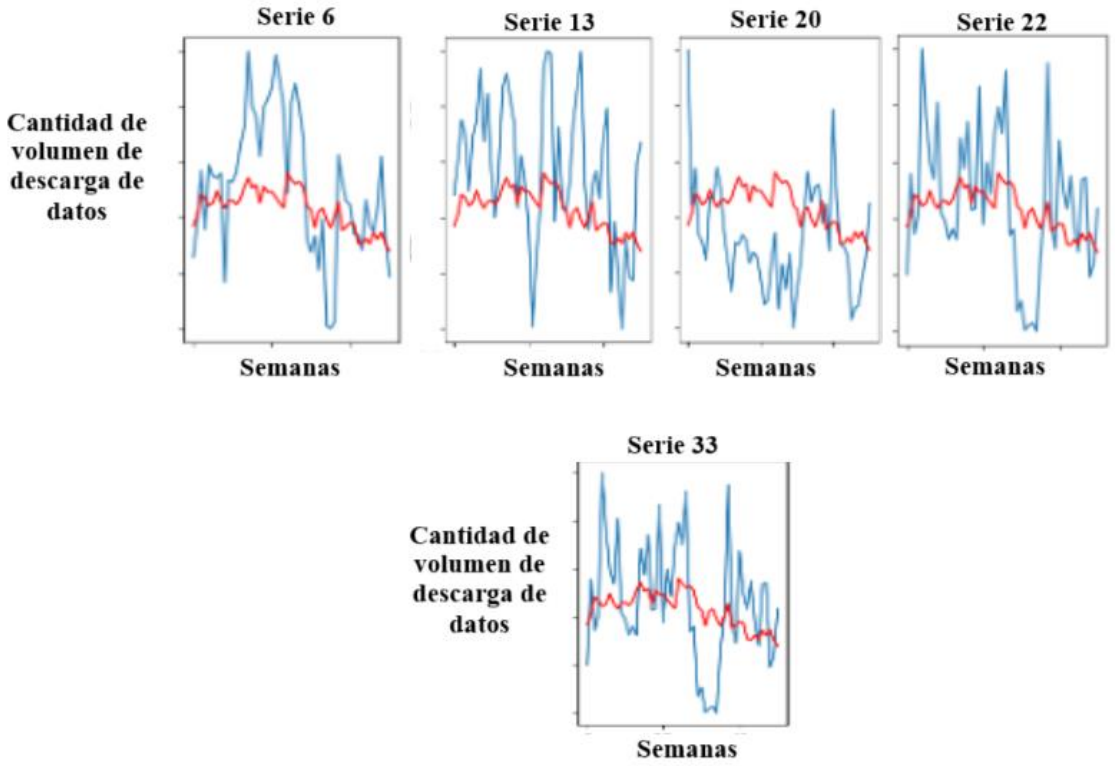
Código realizado:

Detección de semanas atípicas utilizando la mediana móvil

```
import numpy as np
from scipy import stats

# Initialize Analisis_pendientes as an empty list
analisis_atipicos_moving_median = []
analisis_atipicos = []

for m in range(len(cluster_0_df)):
    if(labels[i]==0):
        Centroide_0_high = []  # Use lists to store values for each iteration
        Centroide_0_low = []

        for i in range(0, 45):
                start = i
                end = i + 6
                Centroide_0_high.append(np.median(cluster_0[m][start:end]) + 2*cluster_0[m][start:end].std())
                Centroide_0_low.append(np.median(cluster_0[m][start:end]) - 2*cluster_0[m][start:end].std())

        for i, index in enumerate(cluster_0[m][5:50]):
                cluster_superior = {}  # Initialize an empty dictionary for each cluster

                if index >= Centroide_0_high[i]:
                    cluster_superior["Semana_superior"] = i+5
                    cluster_superior["serie"] = "serie " + str(m)
                    analisis_atipicos.append(cluster_superior)

                elif index <= Centroide_0_low[i]:
                    cluster_superior["Semana_inferior"] = i+5
                    cluster_superior["serie"] = "serie " + str(m)
                    analisis_atipicos.append(cluster_superior)

# Analisis_pendientes is now a list of dictionaries
analisis_atipicos_moving_median = pd.DataFrame(analisis_atipicos)
analisis_atipicos_moving_median
```

Figura 49

Semanas atípicas detectadas en cada celda utilizando la mediana móvil:



Nota. Elaboración propia. Celdas con más de 6 semanas atípicas identificadas: Celda 22, Celda 33, Celda 6, Celda 13 y Celda 20:

Utilizando la mediana móvil se detectaron 3 celdas con más de 7 semanas atípicas y 2 celdas con más de 6 semanas atípicas, comparando estos valores con las semanas atípicas promedio del grupo que es de 4 semanas atípicas.

Finalmente se hizo una comparación entre la puntuación obtenida por DTW y el número de semanas atípicas detectadas obteniéndose que ambos métodos generan resultados bastante diferentes, en la Tabla 11 se observa con más detalle este análisis:

Tabla 11

Comparación entre métricas de mediana móvil (Cantidad de semanas atípicas) y DTW (Puntuación DTW)

Series de tiempo	Semanas atípicas	Puntuación DTW
Serie 22	7	6.899929289
Serie 33	7	7.197150156
Serie 6	7	8.689523808
Serie 13	6	9.232405951
Serie 30	6	8.456232455

Nota. Elaboración propia

El DTW proporciona resultados más precisos al capturar el grado de similitud entre dos series de tiempo. Una puntuación alta en DTW indica que la serie de tiempo es atípica. Por otro lado, el método de la mediana móvil depende de que la celda tenga una distribución normal; de lo contrario, la mediana podría estar sesgada. Además, es importante considerar que ajustar la desviación estándar hasta alcanzar el valor óptimo puede generar un consumo significativo de tiempo y recursos.

4.3.3.2. Mapas autoorganizados (SOM):

A continuación, se presenta un ejemplo del Código realizado para un mapa de 5x5:

Código realizado:

Código para crear una red neuronal 5x5 utilizando la librería Minisom.

```
# file:///C:/Users/lVs_9/Downloads/AnomalyDetectionUsingSelf-OrganizingMaps-BasedK-NearestNeig....pdf
som_x = 5 #Dimension del mapa
som_y = 5 #Dimension del mapa
```

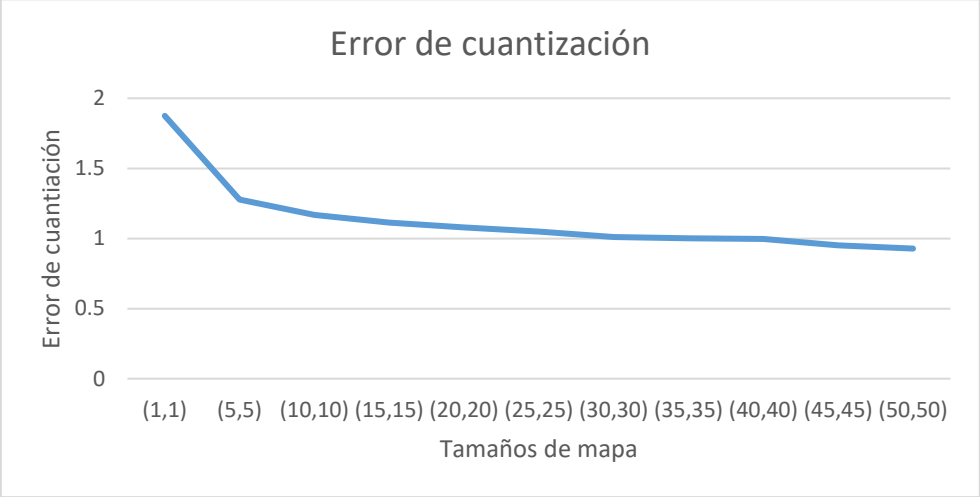
```
som = MiniSom(som_x, som_y,len(dfs8[0]), sigma=0.5, learning_rate = 0.5, neighborhood_function = 'triangle')

som.random_weights_init(dfs8)
som.train(dfs8, 50000)
```

Errores de cuantización:

Se calculó el error de cuantización para distintos tamaños de mapa abarcando desde un mapa de 1x1 (1 grupo) hasta un mapa de 50x50 (2500 grupos). El resultado de este análisis se muestra en la Figura 50.

Figura 50

Error de cuantización para distintos tamaños de mapa



Nota. Elaboración propia

Siguiendo la metodología descrita en la sección anterior y basándonos en los resultados obtenidos, se seleccionaron los mapas de tamaño 5x5 (25 grupos), 12x12 (144 grupos) y 27x27 (729 grupos) para llevar a cabo pruebas de rendimiento.

SOM 5x5

Se utilizó la métrica del error del producto topográfico para determinar la forma del mapa de 25 grupos. Los resultados se muestran de manera gráfica en la Figura 51:

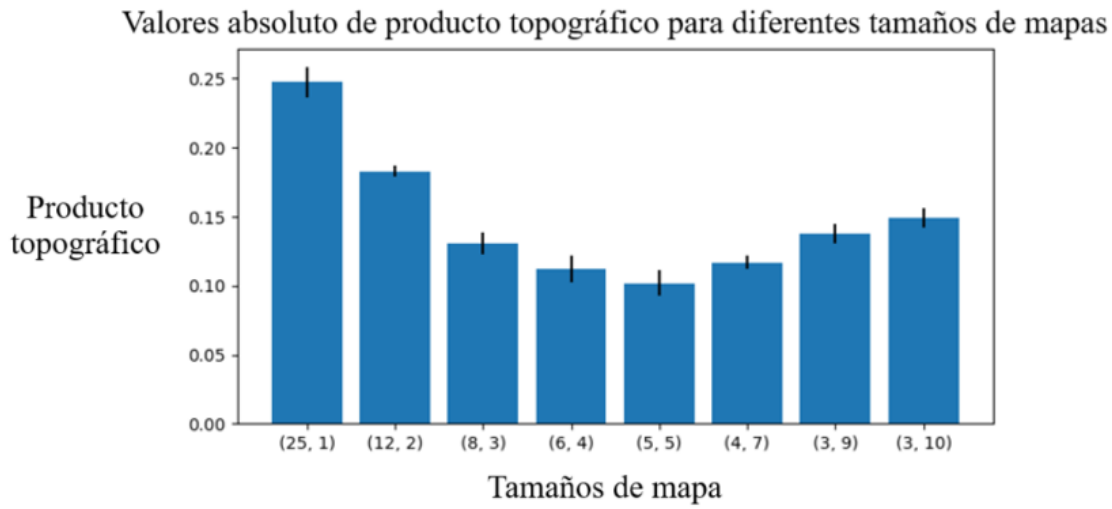
Código realizado:

Código para determinar el producto topográfico de varias combinaciones de 25 nodos.

```
size = len(prueba_dfs8)
n_iter = 10 * size  # 10 epochs
n_runs = 8

tps = np.zeros((len(map_sizes), n_runs))

fig, ax = plt.subplots(2, 4, figsize=(28, 14))

for idx,map_size in enumerate(map_sizes):
    n_units = map_size[0] * map_size[1]
    dist_fun = rectangular_topology_dist(map_size)

    som = MiniSom(map_size[0], map_size[1], prueba_dfs8.shape[-1], sigma=0.3, learning_rate=0.1, random_seed=42)

    best_tp = 1.0
    for run in range(n_runs):
        som.random_weights_init(prueba_dfs8)
        som.train_batch(prueba_dfs8, n_iter)
        prototypes = som.get_weights().reshape(-1, prueba_dfs8.shape[-1])
        tps[idx, run] = topographic_product(dist_fun, prototypes)
        if tps[idx, run] < best_tp:
            best_tp = tps[idx, run]
            best_prototypes = prototypes

    # Visualize map prototypes
    for i in range(n_units):
        for j in range(n_units):
            if dist_fun(i, j) == 1:
                ax[idx//4, idx%4].plot([best_prototypes[i, 0], best_prototypes[j, 0]], [best_prototypes[i, 1],
                                                                                        best_prototypes[j, 1]], 'k-',
                                       linewidth=0.7)
    ax[idx//4, idx%4].scatter(best_prototypes[:, 0], best_prototypes[:, 1], s=20, c='k')
    ax[idx//4, idx%4].set_title('Map size: {}'.format(map_size))
    ax[idx//4, idx%4].set_xlim([0.2, 0.8])
    ax[idx//4, idx%4].set_ylim([0.0, 1.0])
```

Figura 51

Cálculo del error topográfico para distintos mapas múltiplos de 25



Nota. Elaboración propia.

Del resultado se observa que el mapa 5x5 tiene el menor error de producto topográfico.

Para el cálculo de los 3 parámetros principales de la red neuronal se iteró con distintas combinaciones de valor de sigma, tasa de aprendizaje y función de vecindad. El resultado se muestra en la Tabla 12:

Código realizado:

Código para diferentes iteraciones de mapa 5x5

```
# file:///C:/Users/LVs_9/Downloads/AnomalyDetectionUsingSelf-OrganizingMaps-BasedK-NearestNeig....pdf
som_x = 5 #Dimension del mapa
som_y = 5 #Dimension del mapa
```

```
som = MiniSom(som_x, som_y,len(dfs8[0]), sigma=0.5, learning_rate = 0.5, neighborhood_function = 'triangle')

som.random_weights_init(dfs8)
som.train(dfs8, 50000)
```

Tabla 12

Iteración de distintas combinaciones de Sigma, tasa de aprendizaje y función de vecindad.

	Sigma		
Learning Rate-Neighborhood function	0.1	0.5	0.8
0.1-Gaussian	1.278	1.279	1.278
0.1-Mexican_hat	1.29	1.436	1.59
0.1-Bubble	1.277	1.282	1.274
0.1-Triangle	1.412	1.274	1.263
0.5-Gaussian	1.358	1.351	1.357
0.5-Mexican_hat	1.359	1.506	NULL
0.5-Bubble	1.349	1.361	1.346
0.5-Triangle	1.311	1.25	1.26
0.8-Gaussian	1.393	1.393	1.4
0.8-Mexican_hat	1.39	1.58	NULL
0.8-Bubble	1.382	1.401	1.395
0.8-Triangle	1.296	1.264	1.266

Nota. Elaboración propia

Los resultados muestran que los mejores parámetros para una red neuronal SOM de 25 grupos fueron de:

- Forma de mapa = 5x5
- Sigma = 0.5
- Tasa de aprendizaje = 0.5
- Función de vecindad = Triángulo

Finalmente se utilizaron los pesos obtenidos para graficar la red neuronal. En la Figura 52 se muestra la red neuronal conformada por 25 nodos:

Código realizado:

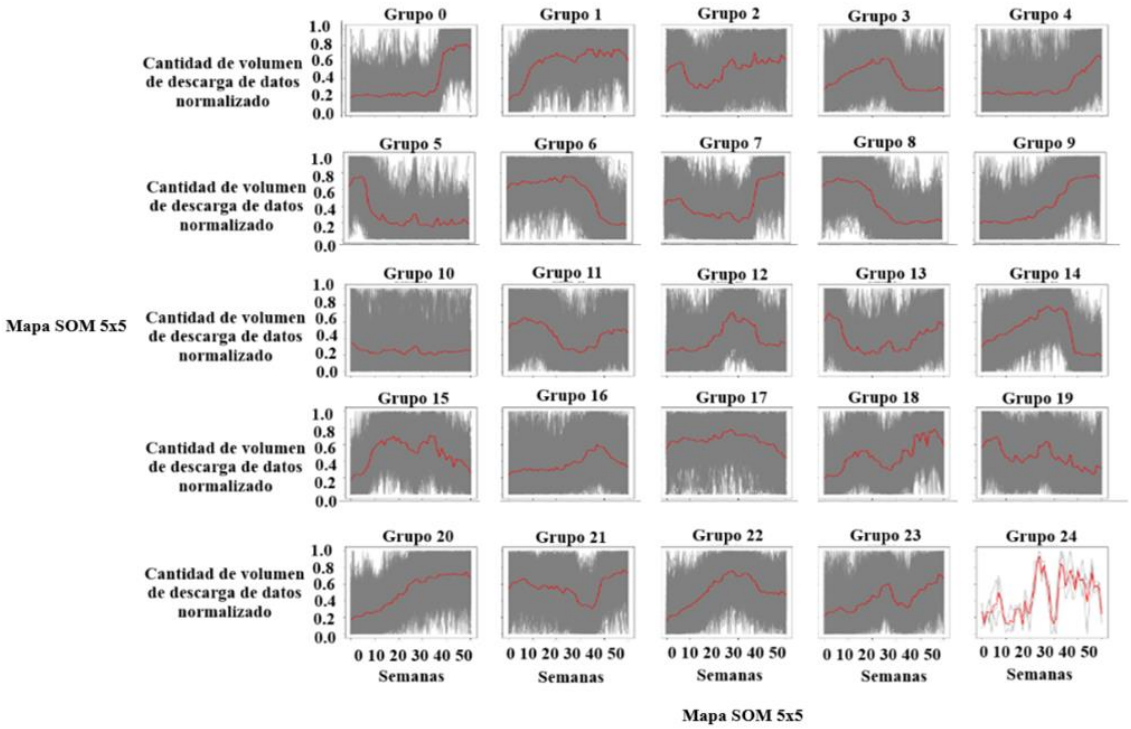
Código para generar gráficas con los pesos obtenidos.

```
# little handy function to plot series
def plot_som_series_averaged_center(som_x, som_y, win_map): #(Dimension, dimension, identificador de listas de numpy)
    fig, axs = plt.subplots(som_x,som_y,figsize=(25,25)) #Crear subgraficos con dimension x dimension
    fig.suptitle('Clusters')
    for x in range(som_x): #Se itera con cada valor de dimension x
        for y in range(som_y): #Se itera con cada valor de dimension y
            cluster = (x,y) #Coordenada del diccionario de lista de numpy
            if cluster in win_map.keys(): #Se itera cada coordenada en la lista de identificadores del diccionario
                for series in win_map[cluster]: #Se itera cada fila numpy dentro de la lista de filas numpy identificadas
                    axs[cluster].plot(series,c="gray",alpha=0.5) #Se grafica la coordenada respectiva en el mapa de neuronas
                axs[cluster].plot(np.average(np.vstack(win_map[cluster]),axis=0),c="red") #Se promedia las filas contenidas dentr
            cluster_number = x*som_y+y+1 #Se obtiene el número del cluster
            axs[cluster].set_title(f"Cluster {cluster_number}")

    plt.show()
```

```
win_map = som.win_map(dfs8)
# Returns the mapping of the winner nodes and inputs

plot_som_series_averaged_center(som_x, som_y, win_map)
```

Figura 52

Red neuronal SOM compuesta de 25 nodos.



Nota. Elaboración propia. Esta figura muestra los grupos generados por la aplicación del algoritmo SOM 5x5. La línea roja en cada nodo representa el promedio de los pesos contenidos en ese grupo.

En la Figura 53 se muestra la distribución de celdas que fueron agrupadas con el método de SOM 5x5:

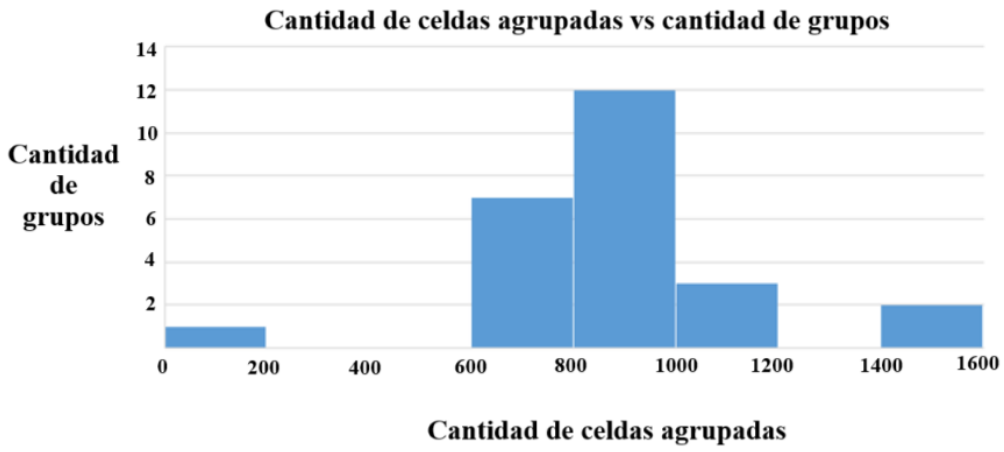
Código realizado:

Código para generar la distribución de celdas por cada neurona.

```
cluster_c = []
cluster_n = []
a = []
b = []
Longitud = []
Lista = []
Avg_cluster = []
for x in range(som_x):
    for y in range(som_y):
        cluster = (x,y)
        if cluster in win_map.keys():
            if len(win_map[cluster])>=0:
                cluster_c.append(len(win_map[cluster]))
                cluster_number = x*som_y+y+1
                a.append(x)
                b.append(y)
                Longitud.append(len(win_map[cluster]))
                cluster_n.append(f"Cluster {cluster_number}")

plt.figure(figsize=(25,5))
plt.title("Cluster Distribution for SOM")
plt.bar(cluster_n,cluster_c)
plt.show()
```

Figura 53

Distribución de celdas que fueron agrupadas con el método de SOM 5x5.



SOM 12x12

Se utilizó la métrica del error del producto topográfico para determinar la forma del mapa de 144 grupos. Los resultados se muestran de manera gráfica en la Figura 54:

Código realizado:

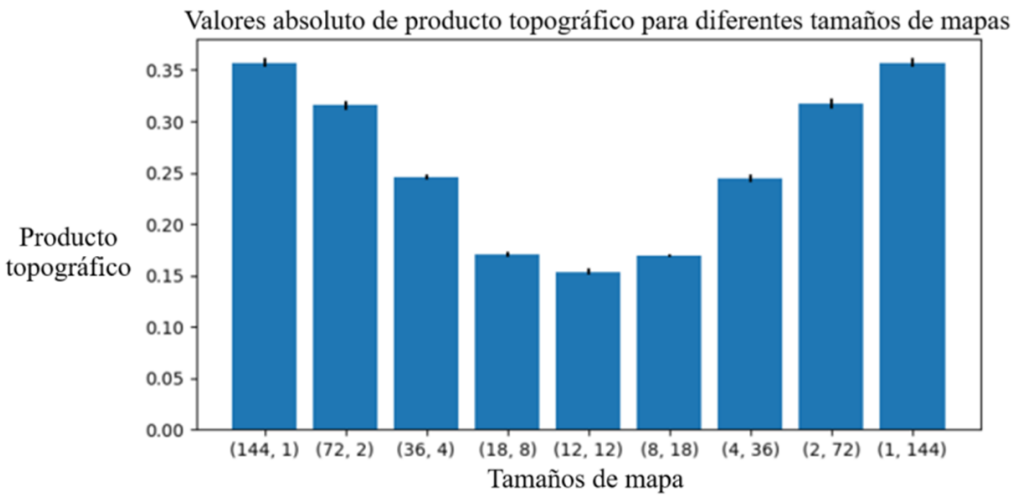
Código para determinar el producto topográfico de varias combinaciones de 144 nodos.

```
#math.ceil(5*math.sqrt(len(dfs8))) = 736
# file:///C:/Users/LVs_9/Downloads/AnomalyDetectionUsingSelf-OrganizingMaps-BasedK-NearestNeig....pdf
som_x = 12 #Dimension del mapa
som_y = 12 #Dimension del mapa
```

```
som = MiniSom(som_x, som_y,len(dfs8[0]), sigma=0.8, learning_rate = 0.1, neighborhood_function = 'gaussian')

som.random_weights_init(dfs8)
som.train(dfs8, 50000)
```

Figura 54

Cálculo del error topográfico para distintos mapas múltiplos de 144



Nota. Elaboración propia

Del resultado se observa que el mapa 12x12 tiene el menor error de producto topográfico.

Para el cálculo de los 3 parámetros principales de la red neuronal se iteró con distintas combinaciones de valor de sigma, tasa de aprendizaje y función de vecindad. El resultado se muestra en la Tabla 13:

Tabla 13

Iteración de distintas combinaciones de Sigma, tasa de aprendizaje y función de vecindad.

	Sigma		
Learning Rate-Neighborhood function	0.1	0.5	0.8
0.1-Gaussian	1.14	1.128	1.124
0.1-Mexican_hat	1.143	1.272	1.368
0.1-Bubble	1.142	1.141	1.14
0.1-Triangle	1.364	1.174	1.148
0.5-Gaussian	1.19	1.184	1.184
0.5-Mexican_hat	1.189	1.338	1.445
0.5-Bubble	1.188	1.188	1.188
0.5-Triangle	1.248	1.139	1.13
0.8-Gaussian	1.232	1.222	1.225
0.8-Mexican_hat	1.232	1.386	1.502
0.8-Bubble	1.226	1.229	1.229
0.8-Triangle	1.227	1.132	1.129

Nota. Elaboración propia

Los resultados muestran que los mejores parámetros para una red neuronal SOM de 144 grupos fueron de:

- Forma de mapa = 12x12
- Sigma = 0.8
- Tasa de aprendizaje = 0.1
- Función de vecindad = Gaussiana

Finalmente se utilizaron los pesos obtenidos para graficar la red neuronal. En la Figura 55 se muestra la red neuronal conformada por 144 nodos:

Código realizado:

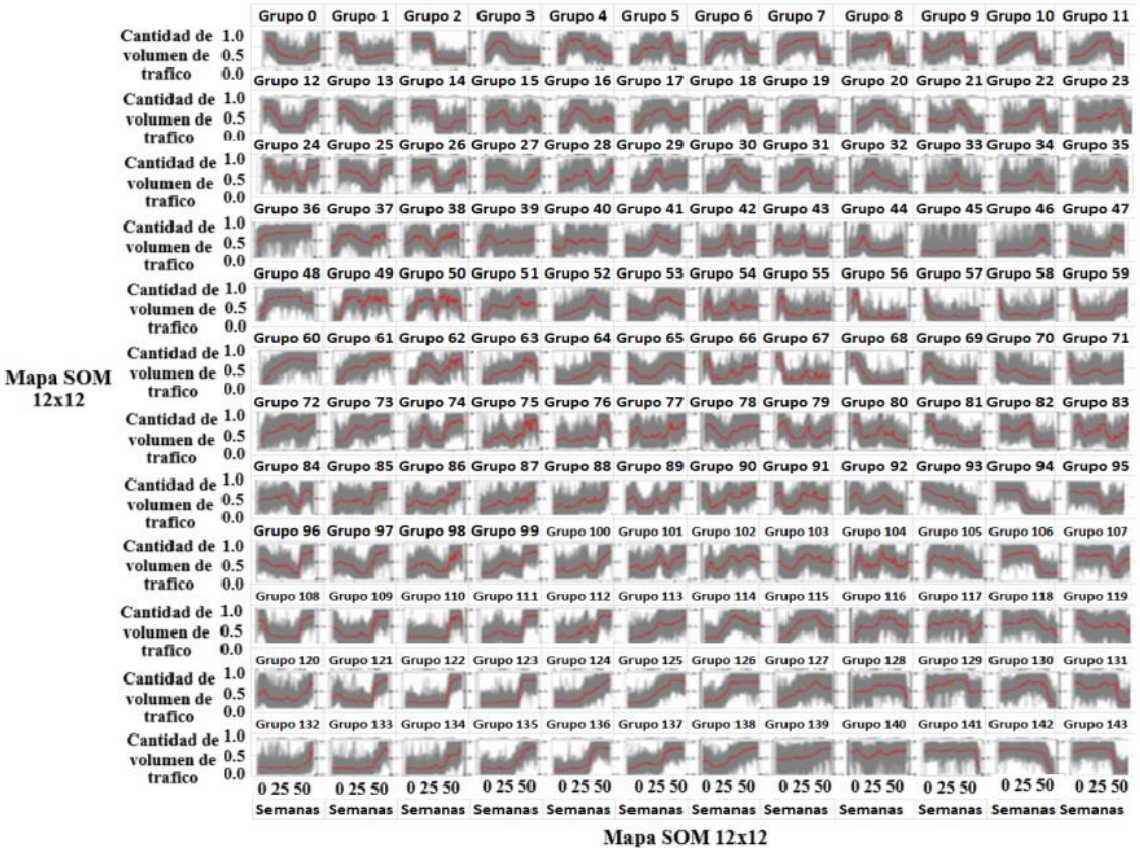
Código para generar gráficas con los pesos obtenidos.

```
# Little handy function to plot series
def plot_som_series_averaged_center(som_x, som_y, win_map): #(Dimension, dimension, identificador de listas de numpy)
    fig, axs = plt.subplots(som_x,som_y,figsize=(25,25)) #Crear subgraficos con dimension x dimension
    fig.suptitle('Clusters')
    for x in range(som_x): #Se itera con cada valor de dimension x
        for y in range(som_y): #Se itera con cada valor de dimension y
            cluster = (x,y) #Coordenada del diccionario de lista de numpy
            if cluster in win_map.keys(): #Se itera cada coordenada en la lista de identificadores del diccionario
                for series in win_map[cluster]: #Se itera cada fila numpy dentro de la lista de filas numpy identificadas
                    axs[cluster].plot(series,c="gray",alpha=0.5) #Se grafica la coordenada respectiva en el mapa de neuronas
                axs[cluster].plot(np.average(np.vstack(win_map[cluster]),axis=0),c="red") #Se promedia las filas contenidas dentr
            cluster_number = x*som_y+y+1 #Se obtiene el número del cluster
            axs[cluster].set_title(f"Cluster {cluster_number}")

    plt.show()
```

```
win_map = som.win_map(dfs8)
# Returns the mapping of the winner nodes and inputs

plot_som_series_averaged_center(som_x, som_y, win_map)
```

Figura 55

Red neuronal compuesta de 144 nodos.



Nota. Elaboración propia. La línea roja en cada nodo representa el promedio de los pesos contenidos en ese grupo.

En la Figura 56 se muestra la distribución de celdas que fueron agrupadas con el método de SOM 12x12:

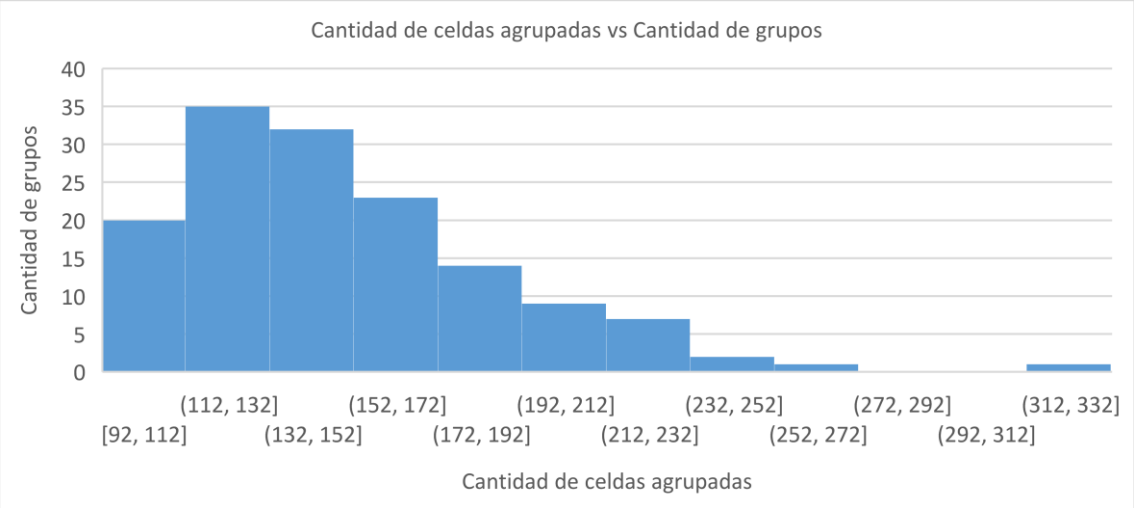
Código realizado:

Código para generar la distribución de celdas por cada neurona.

```
cluster_c = []
cluster_n = []
a = []
b = []
Longitud = []
Lista = []
Avg_cluster = []
for x in range(som_x):
    for y in range(som_y):
        cluster = (x,y)
        if cluster in win_map.keys():
            if len(win_map[cluster])>=0:
                cluster_c.append(len(win_map[cluster]))
                cluster_number = x*som_y+y+1
                a.append(x)
                b.append(y)
                Longitud.append(len(win_map[cluster]))
                cluster_n.append(f"Cluster {cluster_number}")

plt.figure(figsize=(25,5))
plt.title("Cluster Distribution for SOM")
plt.bar(cluster_n,cluster_c)
plt.show()
```

Figura 56

Distribución de celdas que fueron agrupadas con el método de SOM12x12.



Nota. Elaboración propia

SOM 27x27

Se utilizó la métrica del error del producto topográfico para determinar la forma del mapa de 729 grupos. Los resultados se muestran de manera gráfica en la Figura 57:

Código realizado:

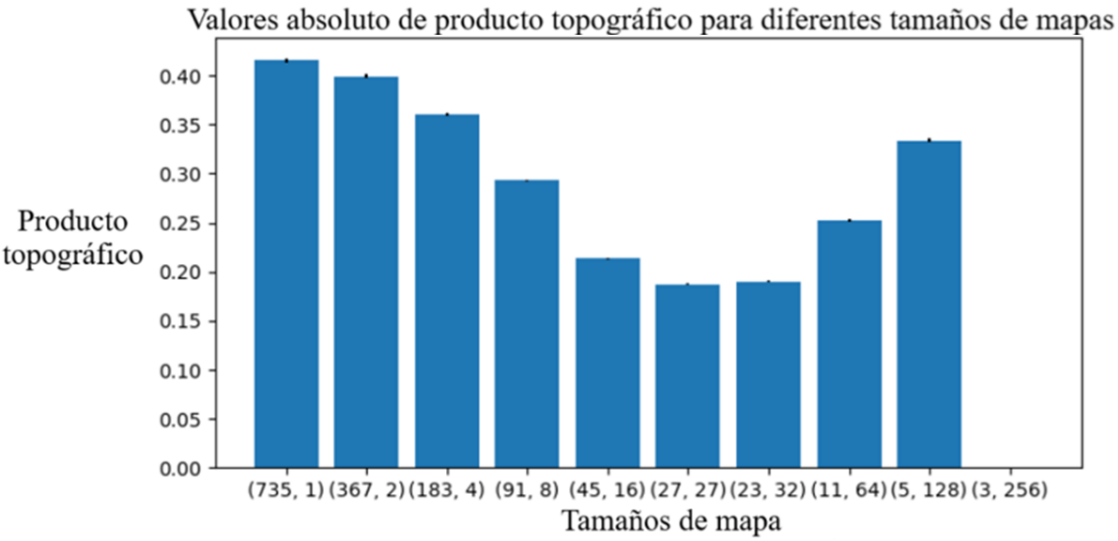
Código para determinar el producto topográfico de varias combinaciones de 729 nodos.

```
#math.ceil(5*math.sqrt(len(dfs8))) = 736
# file:///C:/Users/LVs_9/Downloads/AnomalyDetectionUsingSelf-OrganizingMaps-BasedK-NearestNeig....pdf
som_x = 27 #Dimension del mapa
som_y = 27 #Dimension del mapa
```

```
som = MiniSom(som_x, som_y,len(dfs8[0]), sigma=0.8, learning_rate = 0.8, neighborhood_function = 'triangle')

som.random_weights_init(dfs8)
som.train(dfs8, 50000)
```

Figura 57

Cálculo del error topográfico para distintos mapas múltiplos de 729



Nota. Elaboración propia. Del resultado se observa que el mapa 27x27 tiene el menor error de producto topográfico.

Para el cálculo de los 3 parámetros principales de la red neuronal se iteró con distintas combinaciones de valor de sigma, tasa de aprendizaje y función de vecindad. El resultado se muestra en la Tabla 14:

Tabla 14

Iteración de distintas combinaciones de Sigma, tasa de aprendizaje y función de vecindad.

	Sigma		
Learning Rate-Neighborhood function	0.1	0.5	0.8
0.1-Gaussian	1.04	1.033	1.03
0.1-Mexican_hat	1.04	1.124	1.187
0.1-Bubble	1.034	1.041	1.039
0.1-Triangle	1.21	1.108	1.072
0.5-Gaussian	1.062	1.05	1.043
0.5-Mexican_hat	1.063	1.19	1.271
0.5-Bubble	1.063	1.06	1.061
0.5-Triangle	1.184	1.054	1.034
0.8-Gaussian	1.098	1.088	1.082
0.8-Mexican_hat	1.099	1.227	1.314
0.8-Bubble	1.098	1.098	1.099
0.8-Triangle	1.162	1.042	1.029

Nota. Elaboración propia.

Los resultados muestran que los mejores parámetros para una red neuronal SOM de 729 grupos fueron de:

- Forma de mapa = 27x27
- Sigma = 0.8
- Tasa de aprendizaje = 0.8
- Función de vecindad = Triangulo

Finalmente se utilizaron los pesos obtenidos para graficar la red neuronal. En la Figura 58 se muestra la red neuronal conformada por 729 nodos:

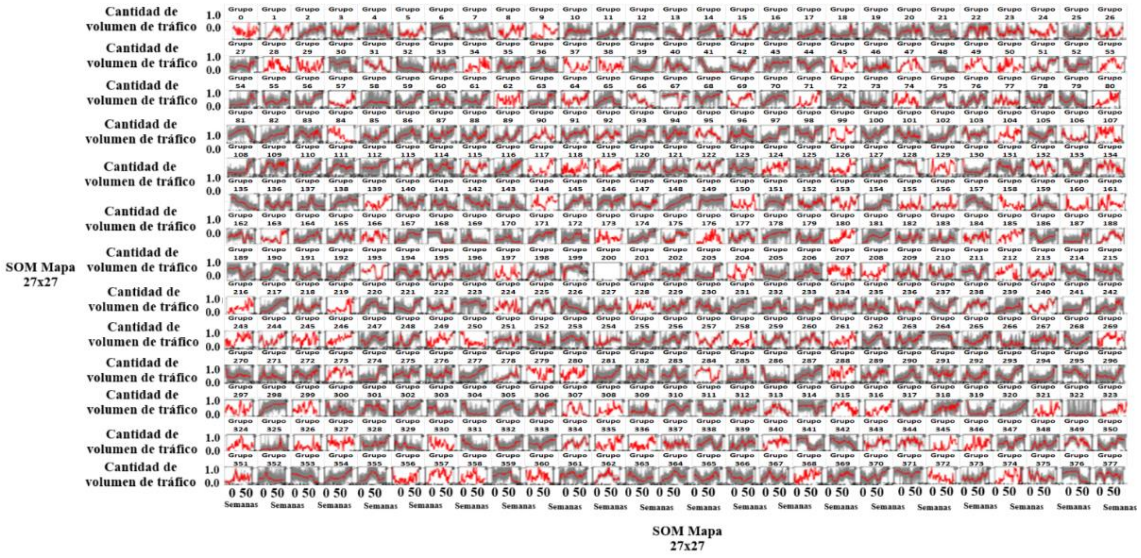
Código realizado:

Código para generar gráficas con los pesos obtenidos.

```
# Little handy function to plot series
def plot_som_series_averaged_center(som_x, som_y, win_map): #(Dimension, dimension, identificador de listas de numpy)
    fig, axs = plt.subplots(som_x,som_y,figsize=(25,25)) #Crear subgraficos con dimension x dimension
    fig.suptitle('Clusters')
    for x in range(som_x): #Se itera con cada valor de dimension x
        for y in range(som_y): #Se itera con cada valor de dimension y
            cluster = (x,y) #Coordenada del diccionario de lista de numpy
            if cluster in win_map.keys(): #Se itera cada coordenada en la lista de identificadores del diccionario
                for series in win_map[cluster]: #Se itera cada fila numpy dentro de la lista de filas numpy identificadas
                    axs[cluster].plot(series,c="gray",alpha=0.5) #Se grafica la coordenada respectiva en el mapa de neuronas
                axs[cluster].plot(np.average(np.vstack(win_map[cluster]),axis=0),c="red") #Se promedia las filas contenidas dentr
            cluster_number = x*som_y+y+1 #Se obtiene el número del cluster
            axs[cluster].set_title(f"Cluster {cluster_number}")

    plt.show()
```

Figura 58

Red neuronal compuesta de 729 nodos.



Nota. Elaboración propia. Vista de la mitad de una red neuronal compuesta de 729 nodos. La línea roja en cada nodo representa el promedio de los pesos contenidos en ese grupo.

En la Figura 59 se muestra la distribución de celdas que fueron agrupadas con el método de SOM 27x27:

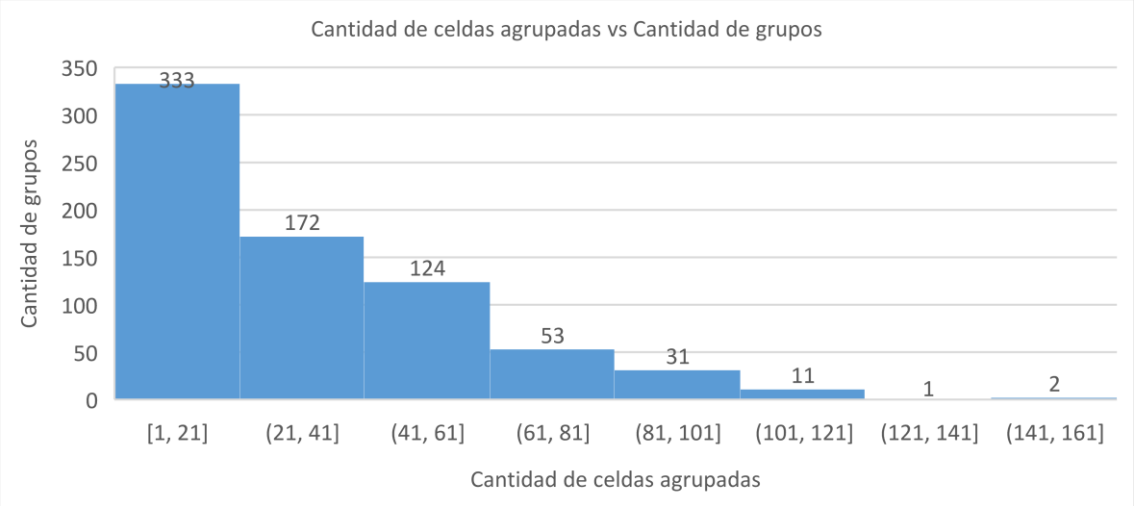
Código realizado:

Código para generar la distribución de celdas por cada neurona.

```
cluster_c = []
cluster_n = []
a = []
b = []
Longitud = []
Lista = []
Avg_cluster = []
for x in range(som_x):
    for y in range(som_y):
        cluster = (x,y)
        if cluster in win_map.keys():
            if len(win_map[cluster])>=0:
                cluster_c.append(len(win_map[cluster]))
                cluster_number = x*som_y+y+1
                a.append(x)
                b.append(y)
                Longitud.append(len(win_map[cluster]))
                cluster_n.append(f"Cluster {cluster_number}")

plt.figure(figsize=(25,5))
plt.title("Cluster Distribution for SOM")
plt.bar(cluster_n,cluster_c)
plt.show()
```

Figura 59

Distribución de celdas que fueron agrupadas con el método de SOM27x27.



Nota. Elaboración propia. Esta figura muestra la distribución estadística de la cantidad de celdas agrupadas en cada grupo.

4.4. Población y muestra

Población:

En el año 2022 se contabilizaron 38183 celdas en total en toda la red LTE

de la operadora WOM.

Muestra:

Para esta tesis se realizó un muestreo subjetivo cuyos parámetros son el número de celdas cuyas horas disponibles fueron mayor al 98% de disponibilidad del año lo cual vino a ser en promedio un total de 8590 horas.

Bajo este concepto se seleccionaron en total una muestra de 21645 celdas.

4.5. Técnicas e instrumentos de recolección de datos

Para esta tesis se realizó una observación experimental analizando como el cambio de una variable independiente afecta a la variable dependiente.

Los datos fueron obtenidos a través del gestor U2020 Huawei el cual colectó de manera horaria la información referente a los indicadores de rendimiento de las celdas de

toda la red WOM de Chile. Estos datos en bruto fueron enviados a una base de datos relacional llamada Clickhouse.

V. DISCUSIÓN DE RESULTADOS

En esta etapa, se discutirán los resultados de las pruebas realizadas tanto para el algoritmo K-Means como para la red neuronal SOM.

Comparación de modelo K-Means y SOM

Se aplicaron los métodos de agrupamiento K-Means y SOM para comprobar su eficacia en la identificación de patrones de tendencia:

Identificación de patrones de tendencia

Para hallar los patrones de tendencia primero se determinó el intervalo óptimo de semanas necesarias para poder analizar las distintas tendencias entre celdas. Para ello se realizó el cálculo de varianzas entre cada celda y su centroide. Este procedimiento se observa en la Tabla 15.

Tabla 15

Análisis de varianza de semanas respecto a sus pendientes

	2 semanas	5 semanas	10 semanas	17 semanas	25 semanas	51 semanas
1	0.90201359	0.018234	0.0020564	0.0003628	8.651E-05	6.27411E-06
2	1.342346932	0.028872	0.0033484	0.0005889	0.0001502	8.97455E-06
3	2.031290831	0.042532	0.0043848	0.000707	0.0001709	1.03986E-05
4	1.757472497	0.039587	0.0043481	0.0007082	0.0001757	1.17126E-05
5	1.437631827	0.030226	0.0031375	0.000511	0.0001079	7.32857E-06
6	1.05289121	0.026562	0.0030711	0.0004862	0.0001288	1.06889E-05
7	1.149519542	0.029976	0.0035073	0.0007779	0.0001608	1.07843E-05
8	1.346287614	0.02851	0.0031701	0.0005581	0.0001321	7.62727E-06
9	1.165851133	0.032492	0.003927	0.000803	0.0002451	1.34023E-05
10	1.145556233	0.025373	0.0026944	0.0005247	0.0001291	7.68979E-06
11	1.650827007	0.037452	0.0040587	0.0007924	0.0001867	1.27755E-05
12	1.871331228	0.040481	0.0047092	0.0008642	0.0002159	1.1914E-05
13	1.664025734	0.034473	0.0034003	0.000648	0.0001447	1.06475E-05
14	1.683014117	0.035198	0.004187	0.000581	0.0001509	9.5334E-06
15	1.275134667	0.027375	0.0031708	0.0006056	0.0001284	8.29945E-06
16	1.623056227	0.036227	0.0039464	0.0006883	0.000178	1.10383E-05
17	1.968047511	0.041958	0.0042064	0.000676	0.0001502	1.17673E-05
18	1.447173226	0.040051	0.0046121	0.0009292	0.0002054	1.48131E-05
19	1.57360579	0.034443	0.0034153	0.0005928	0.000141	1.11568E-05
20	1.920212588	0.04123	0.0043717	0.0007887	0.0001392	1.06764E-05
21	1.352989827	0.028441	0.0030571	0.0005411	0.0001213	7.48967E-06
22	1.666770978	0.035632	0.0043276	0.0008087	0.0002202	1.35916E-05
23	1.513508578	0.031643	0.0034241	0.0005985	0.000122	7.56921E-06
24	1.895416171	0.037526	0.0039643	0.0005728	0.0001384	9.59999E-06
25	0.979531156	0.019588	0.0013887	9.411E-05	1.065E-05	2.21002E-07
Total	37.41550621	0.824081	0.0898847	0.0158094	0.0037403	0.000245974

Nota. Elaboración propia basada en la base de datos de WOM. Esta tabla muestra la suma de las varianzas semanales de las celdas en relación con la tendencia anual media de sus respectivos grupos.

Según la Tabla 15, analizar las tendencias de las celdas cada 5 semanas resulta en una varianza total de 0.82, en contraste con el análisis semanal que da una varianza total de 37. Se optó por la comparación quincenal para diferenciar las tendencias entre semanas de las celdas.

Identificación de patrones de tendencia creciente:

Para considerar a una tendencia como creciente se utilizó como referencia la siguiente combinación de semanas:

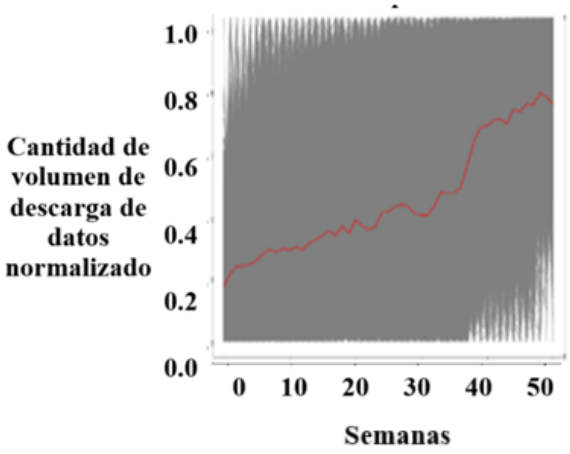
Tendencia creciente = Grupos con tendencia positiva constante (Más de 20 semanas con tendencia positiva en todo el año y ninguna semana con tendencia negativa, el resto de las semanas con tendencia constante

Tendencias crecientes detectadas por K-Means:

Las tendencias crecientes detectadas por K-Means se muestran en la Figura 60.

Figura 60

Tendencias crecientes detectadas por K-Means



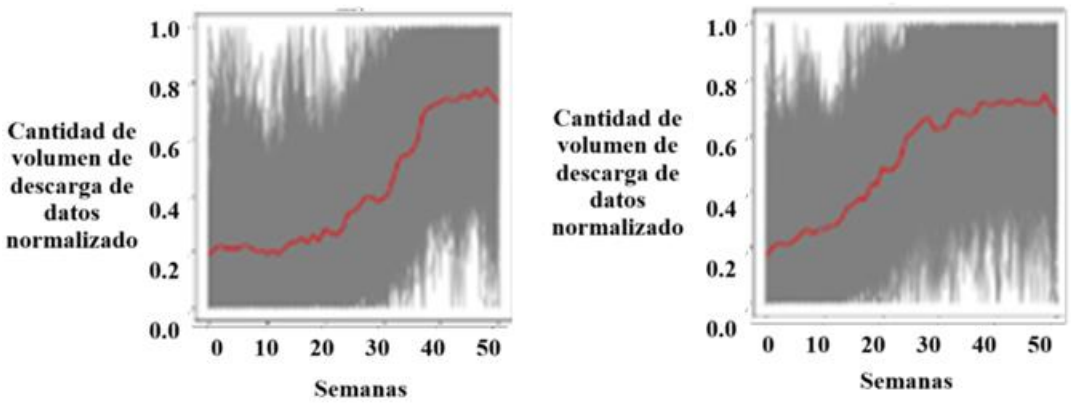
Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón creciente. Según K-Means, las celdas con esta tendencia son más de 7000.

Tendencias crecientes detectadas por SOM con un tamaño de mapa de 5x5:

Las tendencias crecientes detectadas por SOM 5x5 se muestran en la Figura 61.

Figura 61

Tendencias crecientes detectadas por SOM 5x5



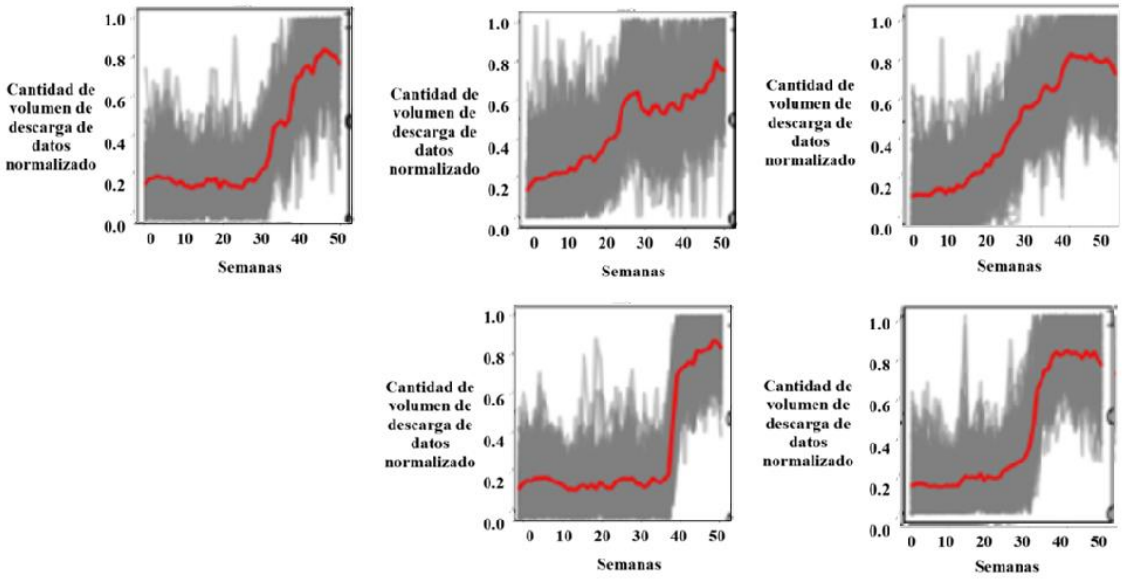
Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón creciente. Según SOM 5x5, las celdas con esta tendencia son más de 1400.

Tendencias crecientes detectadas por SOM con un tamaño de mapa de 12x12:

Las tendencias crecientes detectadas por SOM 12x12 se muestran en la Figura 62.

Figura 62

Tendencias crecientes detectadas por SOM 12x12



Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón creciente. Según SOM 12x12, las celdas con esta tendencia son más de 942.

Detección de patrones de tendencia decreciente:

Para considerar a una tendencia como decreciente se utilizó como referencia la siguiente combinación de semanas:

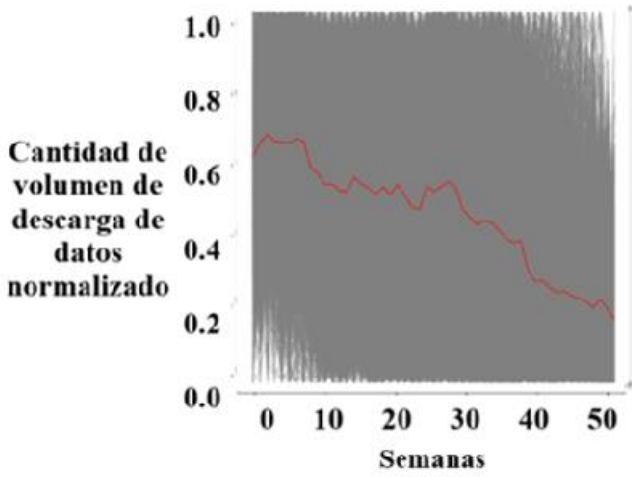
Tendencia decreciente = Celdas con tendencia negativa constante (Más de 15 semanas con tendencia negativa en todo el año y solo una semana positiva, el resto de las semanas constante).

K-Means:

Las tendencias decrecientes detectadas por K-Means se muestran en la Figura 63.

Figura 63

Tendencias decrecientes detectadas por K-Means



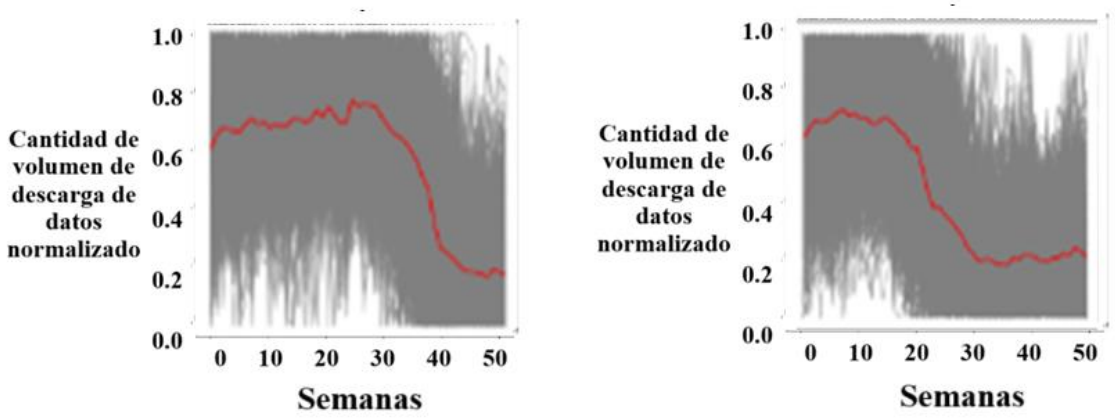
Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón decreciente. Según K-Means, las celdas con esta tendencia son más de 3600.

Tendencias decrecientes detectadas por SOM con un tamaño de mapa de 5x5:

Las tendencias decrecientes detectadas por SOM 5x5 se muestran en la Figura 64.

Figura 64

Tendencias decrecientes detectadas por SOM 5x5



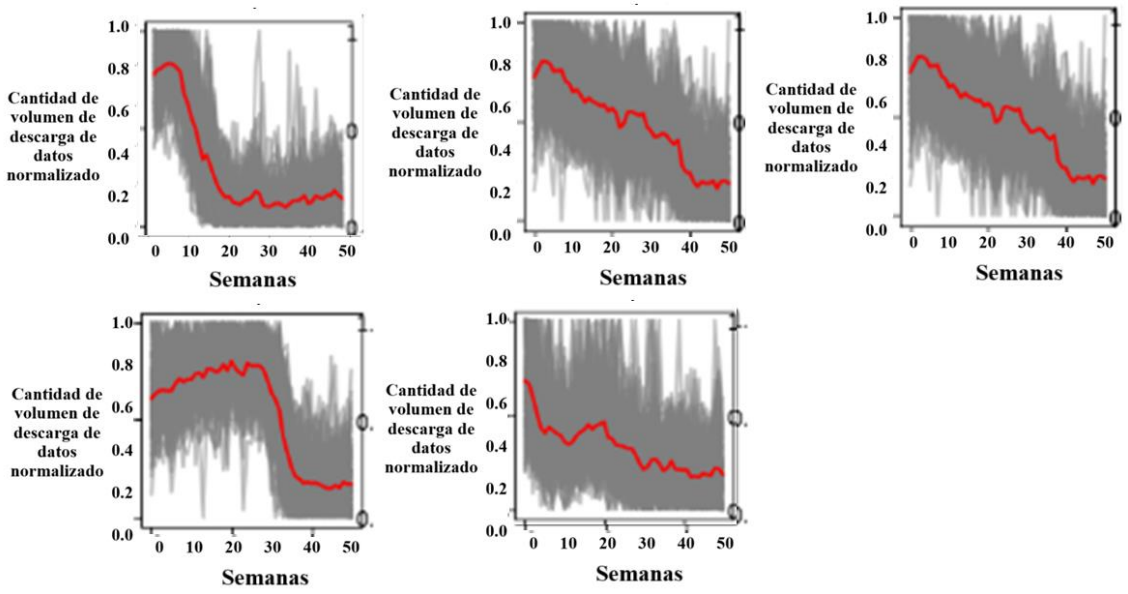
Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón decreciente. Según SOM 5x5, las celdas con esta tendencia son más de 1500.

Tendencias decrecientes detectadas por SOM con un tamaño de mapa de 12x12:

Las tendencias decrecientes detectadas por SOM 12x12 se muestran en la Figura 65.

Figura 65

Tendencias decrecientes detectadas por SOM 12x12



Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón decreciente. Según SOM 12x12, las celdas con esta tendencia son más de 636.

Detección de patrones de tendencia constante:

Para considerar a una tendencia como constante se utilizó como referencia la siguiente combinación de semanas:

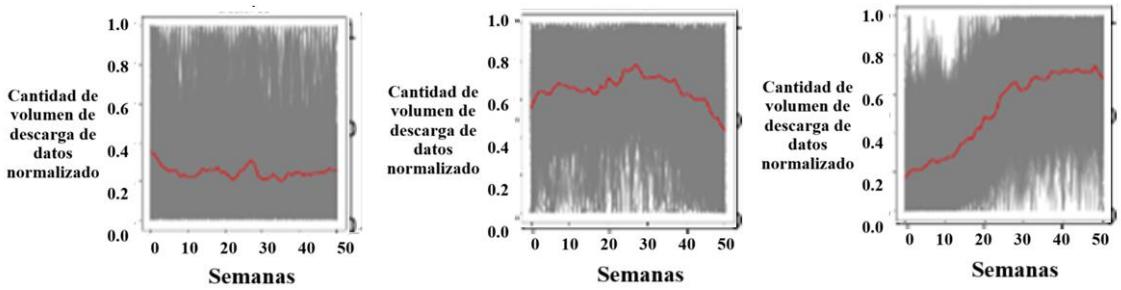
Tendencia constante = Celdas con tendencia cercano a 0 (Más de 30 semanas constantes, sin tendencias positivas muy altas ni muy bajas en todo el año).

Tendencias constantes detectadas por SOM con un tamaño de mapa de 5x5:

Las tendencias constantes detectadas por SOM 5x5 se muestran en la Figura 66.

Figura 66

Tendencias constates detectadas por SOM 5x5



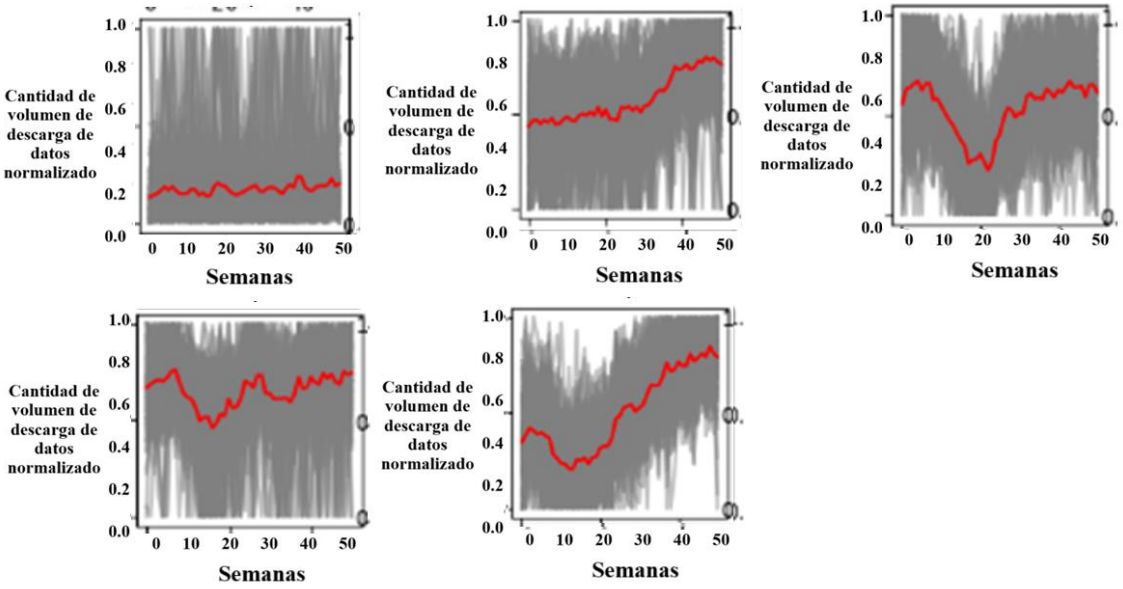
Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón constante. Según SOM 5x5, las celdas con esta tendencia son más de 2465.

Tendencias constantes detectadas por SOM con un tamaño de mapa de 12x12:

Las tendencias constantes detectadas por SOM 12x12 se muestran en la Figura 67.

Figura 67

Tendencias constates detectadas por SOM 12x12



Nota. Elaboración propia. Este gráfico muestra la agrupación de las celdas cuya tendencia tiene un patrón constante. Según SOM 12x12, las celdas con esta tendencia son más de 682.

Finalmente se muestra la tabla comparativa de las otras tendencias identificadas lo cual se muestra en la Tabla 16:

Tabla 16

Comparación entre métodos K-Means y SOM:

	Tendencia creciente	Tendencia decreciente	Tendencia constante	Tendencia de verano	Tendencia de invierno
K-Means	6656	3617	0	0	0
SOM 5x5	1446	1557	2465	1432	122
SOM 12x12	942	636	682	489	627
SOM 27x27	1246	560	92	209	1332

Nota. Elaboración propia. Los modelos SOM 5x5 y SOM 12x12 tuvieron mejores resultados al compararlo con K-Means puesto que identificaron más tipos de patrones.

VI. Conclusiones

Objetivo 1:

Seleccionar e identificar la técnica de interpolación óptima para tratar los datos faltantes del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de interpolación lineal y el método de interpolación polinómica.

Conclusión:

- La interpolación lineal generó un RMSE más bajo en comparación con la interpolación polinómica, y esta diferencia aumentó a medida que se amplió el intervalo de tiempo de los datos faltantes. Por tal motivo, la interpolación lineal resultó una mejor alternativa que la interpolación polinómica grado 2.

Objetivo 2:

Seleccionar e identificar la técnica de segmentación más eficaz para separar los datos del tráfico de las celdas LTE entre la separación a nivel de horas y la separación a nivel de semanas

Conclusión:

- Al agrupar datos semanalmente, se observa que la distribución en una celda puede variar entre patrones como normal, binomial, logarítmica, y sesgada.
- El 65% de la distribución de datos en las celdas muestra variaciones de +-10% entre la media y la mediana.
- Al usar la media o la mediana como valores representativos de las semanas, algunas celdas pueden presentar valores significativamente más altos o bajos debido a la falta de distribución normal. En cambio, la suma total de los datos en una semana ofrece ventajas al no verse afectada por mediciones atípicas, siendo más útil para determinar la tendencia a lo largo del año.
- La granularidad semanal es más eficaz debido a que requiere menos recursos computacionales y reduce el número de dimensiones del conjunto de datos a

diferencia de la granularidad horaria que está más destinada para realizar un análisis de alta granularidad.

Objetivo 3:

Seleccionar e identificar la técnica de determinación de número de grupos más eficaz entre el método del codo y el método de silueta para agrupar el tráfico de las celdas LTE en la red móvil de la operadora WOM

Conclusión:

- Según el método del codo, la reducción de la inercia se vuelve despreciable a partir de 4 grupos.
- Aunque utilizar un mayor número de grupos resultaría en una menor inercia, complicaría el modelo al agrupar un número reducido de celdas cuyos patrones serían casi idénticos a los de otros grupos.
- El método de silueta también indica que a partir de 4 grupos, el coeficiente de silueta se acerca a 0, lo que significa que a mayor número de grupos, mayor es la probabilidad de que una celda esté en el límite de pertenecer a un grupo u otro.

Objetivo 4:

- Seleccionar e identificar la técnica más adecuada para tratar los valores atípicos del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de mediana móvil y el método de distancia de la deformidad de tiempo dinámico (DTW).

Conclusión:

- El DTW proporciona resultados más precisos al capturar el grado de similitud entre dos series de tiempo. Una puntuación alta en DTW indica que la serie de tiempo es atípica. Por otro lado, el método de la mediana móvil depende de que la celda tenga una distribución normal; de lo contrario, la mediana podría estar sesgada. Además, es importante considerar que ajustar la desviación estándar

hasta alcanzar el valor óptimo puede generar un consumo significativo de tiempo y recursos.

-

Objetivo 5:

- Seleccionar e identificar el método de agrupamiento más eficaz entre el método de agrupamiento de series de tiempo (K-Means) y el método de mapas autoorganizadas (SOM) para identificar patrones de tráfico que poseen las celdas LTE en la red móvil y el periodo de tiempo bajo estudio.

Conclusión:

- De acuerdo a los resultados de agrupamiento utilizando K-Means se obtuvieron 4 grupos de tendencias diferentes: Tendencia de mayor volumen de tráfico en los meses medios del año, tendencia de volumen de tráfico creciente a lo largo del año, tendencia de volumen decreciente a lo largo del año y tendencia de mayor volumen de tráfico a inicios y finales del año. Cada grupo contiene en promedio más de 4000 celdas.
- De acuerdo a los resultados de agrupamiento utilizando SOM 5x5 se obtuvieron 25 grupos de tendencias diferentes.
- De acuerdo a los resultados de agrupamiento utilizando SOM 12x12 se obtuvieron 144 grupos de tendencias diferentes.
- De acuerdo a los resultados de agrupamiento utilizando SOM 27x27 se obtuvieron 729 grupos de tendencias diferentes.
- Los modelos SOM 5x5 y SOM 12x12 tuvieron mejores resultados al compararlo con K-Means puesto que identificaron más tipos de patrones.

VII. REFERENCIAS BIBLIOGRÁFICAS

WOM. (2022). *Información confidencial brindada por la empresa WOM.*

D. Salinas, V. Flunkfert, J. Gasthatus y T. Januschowski (2020). DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting*. Recuperada de https://doi.org/10.1016/j.ijforecast.2019.07.001

X. Zhong, L. Chen, Z. Han, J. Zhao y W. Wang (2023). Industrial Time Series Prediction Based on Incremental DBSCAN-KNN with Self-learning Scheme. IEEE Xplore. Recuperada de https://ieeexplore.ieee.org/document/10165826

Ordoñez, M. (2021). Optimización de redes UMTS soportada en machine Learning (Tesis de Maestría, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia). Recuperada de http://hdl.handle.net/11349/27747

Criollo, C. Fuertes, R. Sanmartin, I. (2020). Diseño y optimización de una Red LTE junto técnicas de Machine Learning. (Tesis de Pregrado en Ingeniería Electrónica. Universidad San Francisco de Quito, Quito, Ecuador). Recuperada de http://repositorio.usfq.edu.ec/handle/23000/10107

Benavides. (2021). Predicción de comportamiento en tráfico de red LTE y ajuste de parametrización para maximizar rendimiento de red. (Tesis de Pregrado de Ingeniería Civil Eléctrica de la Universidad de Chile, Santiago de Chile, Chile). Recuperada de

Gutierrez. (2021). Aplicación de técnicas de aprendizaje automático y analítica predictiva para mejorar desempeño de redes 4G LTE. (Tesis de Pregrado de Ingeniería Civil Eléctrica de la Universidad de Chile, Santiago de Chile, Chile). Recuperada de

Lopez. (2021). Anomaly detection and root cause analysis for LTE Radio Base Stations. (Tesis de Maestría en Ciencias de la computación en el instituto de tecnología KTH, Estocolmo, Suecia). Recuperada de: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-231618

Dahlman, E., Parkvall, S., y Skold, J. (2013). 4G: *LTE/LTE-Advanced for Mobile Broadband*. Elsevier.

Parsian, M. (2015). *Data Algoritmo: Recipes for Scaling Up with Hadoop and Spark*. O'Really Media.

Pasternack, I. C. (2021, 7 marzo). The evolution of interconnects in cellular networks: from 4G LTE eNodeB to 5G GNB. 2021-03-06 | *Microwave Journal*. *https://www.microwavejournal.com/articles/35582-the-evolution-of-interconnects-in-cellular-networks-from-4g-lte-enodeb-to-5g-gnb*

Tran, N. (2019). Mean, Median, and Mode in Statistics. Recuperado de: https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b

Griffiths, D. (2008). *First Statistics: A Brain-Friendly Guide*. O'Reilly Media.

Madrigal, E., (2022). Conoce las métricas de precisión más comunes para Modelos de Regresión. Recuperado de: https://www.growupcr.com/post/metricas-precision#:~:text=Ra%C3%ADz%20del%20Error%20Cuadr%C3%A1tico%20Medio%20(RMSE)&text=Es%20conocida%20tambi%C3%A9n%20como%20la,las%20desviaciones%20del%20valor%20real

Scikit-Learn. Clustering. Recuperado de: https://scikit-learn.org/stable/modules/clustering.html#k-means

Rousseeuw, P. (13 Junio 1986). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. North Holland, Paises Bajos.

Vellido, A., Gilbert, K., Angulo, C. y Martín, J. (26-28 de Junio de 2019). *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization.* 13ava WSOM: International Workshop on Self-Organizing Maps. Barcelona, España.

H. Sakoe, S. Chiba (1978). *Dynamic programming algorithm optimization for spoken word recognition*. IEEE Transactions on Acoustics. 10.1109/TASSP.1978.1163055.

Mishra, A., (2020). Time Series Similarity Using Dynamic Time Warping -Explained. Recuperado de: https://medium.com/walmartglobaltech/time-series-similarity-using-dynamic-time-warping-explained-9d09119e48ec

Alí, A., (2019). Self-Organizing Map (SOM) with Practical Implementation. Recuperado de: https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117

Arif, R., (2020). Step by Step to Understanding K-means Clustering and Implementation with sklearn. Recuperado de: https://medium.com/data-folks-indonesia/step-by-step-to-understanding-k-means-clustering-and-implementation-with-sklearn-b55803f519d6

Birgitta Dresp, J. Wandeto, H. Nyongesa. (2018). Using the quantization error from SelfOrganizing Map (SOM) output for fast detection of critical variations in image time series. *Données à la Décision - From Data to Decisions*

Gerón, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Really Media.

CC HUAWEI iMaster MAE (2020). V100R020C10 - Security Target

Huawei PRS Documentation. (2009). PRS: Maximizing MBB network values. https://carrier.huawei.com/en/products/wireless-network-v3/SubSolution-SingleOSS/iManager-PRS

Hwangnam, K., Wonghee, L., Hyunsonn, K., Hwantaee, K y Yang., M. (12 de Julio de 2018). *Protecting Download Traffic from Upload Traffic over Asymmetric Wireless Links.* Wireless Communications and Mobile Computing. Corea, República de Corea.

J. Tian, M. Azarian y M. Pecht (2014). *Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm.* University of Maryland, College Park. Estados Unidos.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means clustering method and Elbow method for identification of the best customer profile cluster. IOP conference series, 336, 012017. https://doi.org/10.1088/1757-899x/336/1/012017

ANEXOS

ANEXO 1: CONTRADO DE AUTORIZACIÓN DE INFORMACIÓN

DocuSign Envelope ID: 2CD8D83D-F32C-47A1-8D02-BCD00D2FA9CF

WOM

AUTORIZACIÓN USO DE INFORMACIÓN.

A 9 de agosto de 2023, entre, Elvis Joel Roque Gonzales, pasaporte peruano Nº pasaporte peruano No 120028919, con domicilio en Sector 3 Grupo 7 Manzana J Lote 07, Villa el Salvador, Lima, Perú, en adelante el "Tesista", por una parte, y por la otra, WOM S.A., RUT Nº 78.921.690-8, representado por Franklin Quijada Campos, cédula de identidad número 11.828.214-0, y por Marcelo Fica Aránguez, cédula de identidad número 13.953.586-3, todos con domicilio en General Mackenna Nº 1369, comuna y ciudad de Santiago, Región Metropolitana, en adelante e indistintamente también denominada "WOM", quienes han convenido lo siguiente:

PRIMERO: Mediante el presente instrumento, WOM autoriza al Tesista para acceder y hacer uso de cierta información de propiedad de la primera consiste únicamente en kpis de red a nivel horario de 2022 (kpis como tráfico, throughput, usuarios, entre otros relacionados directamente con dichos kpis), en adelante el o los "Material(es)", exclusivamente para realizar su tesis de pregrado` con el objeto de obtener el título de `ingeniero en electrónica y telecomunicaciones` en la `Universidad Nacional Tecnológica de Lima Sur`.

SEGUNDO: WOM no garantiza al Tesista la veracidad ni exactitud de los Materiales, por lo que el uso de los mismos no pueden equivaler a una garantía de éxito o resultados futuros. El Tesista libera a WOM, y a sus sociedades relacionadas, a su personal respectivo (de forma individual, cada uno, una "Parte Indemnizada") de y contra cualquier y todas los reclamos, responsabilidades, costos y gastos en que tal Parte Indemnizada pueda incurrir o volverse sujeto a bajo cualquier ley o regulación aplicable, o de otro modo, y relacionado a o que surja de cualquier Material proporcionado por WOM al Tesista.

TERCERO: Todo el Material al que tenga acceso el Tesista es y será de propiedad única y exclusiva de WOM, sin restricción alguna, correspondiéndole a ése último como titular exclusivo, todos los derechos intelectuales sobre ellas, a excepción de la tesis de pregrado referida anteriormente.

El Tesista acepta que por el uso de los Materiales no adquiere ningún derecho de propiedad ni título sobre la misma.

Asimismo, el Tesista declara que, reconoce y acepta que no podrá adquirir ningún derecho de propiedad, usufructo, uso, ni de ningún otro tipo, respecto de marcas de WOM, logotipos, copyrights u otras formas de propiedad intelectual o comercial de WOM o de sus empresas relacionadas.

El Tesista no podrá ceder o transferir los derechos u obligaciones derivados de esta autorización, sin el consentimiento previo y por escrito de WOM.

CUARTO: El Tesista deberá abstenerse de ejecutar cualquier acto que pueda causar o cause perjuicio a los Materiales y a las marcas WOM, y de sus empresas relacionadas y/o que afecte o pueda llegar a afectar el nombre y/o reputación de WOM, y de sus empresas relacionadas.

Asimismo, deberá abstenerse de registrar, depositar y/o solicitar para registro o depósito marcas, lemas comerciales, nombres y/o señas y en general cualquier otro signo, producto, o material que sea objeto de derechos de propiedad industrial, que sea idéntico, similar y/o confundible con las marcas y Materiales de WOM y/o de sus empresas relacionadas.

QUINTO: La presente autorización se concede de forma gratuita.

WOM podrá revocar en cualquier momento la presente autorización, si a su exclusivo juicio causa perjuicio o pudiese causar un perjuicio a la compañía.

SEXTO: Las partes dejan expresa constancia que por el hecho de conceder el uso de los Materiales, el Tesista no tiene ni tendrá la calidad de trabajador dependiente o empleado de WOM.

SÉPTIMO: Cualquier dificultad o controversia que se produzca entre los contratantes respecto de la aplicación, interpretación, duración, validez o ejecución de esta autorización o cualquier otro motivo será sometida al conocimiento de los Tribunales Ordinarios de Justicia de la ciudad y comuna de Santiago, Chile.

La presente autorización se suscribe a través de Docusign.

Elvis Joel Roque Gonzales

Franklin Quijada Campos
p.p. WOM S.A.

Marcelo Fica Aranguez
p.p. WOM S.A.

ANEXO 2: MATRIZ DE CONSISTENCIA

MATRIZ DE CONSISTENCIA

TÍTULO: Aplicación de algoritmos de agrupamiento en aprendizaje de máquina no supervisado para la identificación de patrones de tráfico de celdas LTE en la red nacional móvil de la operadora WOM de Chile **(Ver tabla 12)**

Tabla 12

Matriz de consistencia

<table>
<tr><th>PROBLEMAS</th><th>OBJETIVOS</th><th>HIPÓTESIS</th><th>VARIABLES</th><th>DIMENSIONES</th><th>INDICADORES</th><th>METODOLOGÍA</th></tr>
<tr>
<td rowspan="4">Problema General:
¿Cómo lograr que la aplicación de algoritmos de aprendizaje de máquina permita identificar patrones de tráfico que poseen cada área de cobertura de las celdas LTE en la red móvil de la operadora WOM a lo largo del 2022?
Problemas Específico:
¿Qué método de agrupamiento es útil para identificar patrones de tráfico que poseen cada área de cobertura de las celdas LTE en la red móvil de la operadora WOM a lo largo del tiempo?</td>
<td rowspan="4">Objetivo General:
Aplicar algoritmos de agrupamiento para la identificación de patrones de tráfico de celdas LTE en la red nacional móvil de la operadora WOM de Chile
Objetivos Específicos:
Seleccionar e identificar el método de agrupamiento más eficaz entre el método de agrupamiento de series de tiempo (K-Means) y el método de mapas autoorganizadas (SOM) para identificar patrones de tráfico que poseen las celdas LTE en la red móvil y el periodo de tiempo bajo estudio.
Seleccionar e identificar la</td>
<td rowspan="4">Hipótesis General:
Se pueden identificar patrones de tráfico de celdas LTE de la red nacional móvil de la operadora WOM de Chile utilizando algoritmos de agrupamiento en aprendizaje de máquina no supervisado a partir de los datos registrados en el 2022.
Hipótesis Específica 1
Es posible identificar los patrones de comportamiento de las celdas LTE con los datos del año 2022
Hipótesis Específica 2</td>
<td rowspan="2">Volumen de datos</td>
<td>Tráfico de descarga de datos</td>
<td>• Volumen de descarga de datos medido en Bytes por cada hora en cada celda LTE</td>
<td rowspan="4">Tipo investigación
Investigación aplicada o tecnológica
Nivel de investigación
Nivel experimental
Diseño de investigación
Diseño de investigación experimental
Enfoque de investigación
Enfoque cuantitativo
Técnica
Recolección de datos existentes
Instrumentos
Antenas LTE
Gestor U2020
Población
Las celdas LTE que abarcan la red nacional chilena de la</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td rowspan="2">Patrón de volumen de datos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Patrón del volumen de datos a lo largo de un año de todas las celdas de la red nacional LTE de la operadora WOM</td>
<td>• Identificación y Agrupamiento de celdas en grupos de celdas que tengan un patrón de volumen de datos en común</td>
</tr>
</table>

¿Qué técnica de muestreo es útil para segmentar los datos del tráfico de las celdas LTE en la red móvil de la operadora WOM? ¿Qué técnica de imputación es útil para tratar los datos faltantes del tráfico de las celdas LTE en la red móvil de la operadora WOM? ¿Qué técnica de selección de datos es útil para tratar los valores atípicos del tráfico de las celdas LTE en la red móvil de la operadora WOM? ¿Qué técnica de determinación de número de grupos es útil para agrupar eficazmente el tráfico de las celdas LTE en la red móvil de la operadora WOM?	técnica de segmentación más eficaz para separar los datos del tráfico de las celdas LTE entre la separación a nivel de horas y la separación a nivel de semanas Seleccionar e identificar la técnica de interpolación óptima para tratar los datos faltantes del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de interpolación lineal y el método de interpolación polinómica. Seleccionar e identificar la técnica más adecuada para tratar los valores atípicos del tráfico de las celdas LTE en la red móvil de la operadora WOM entre el método de mediana móvil y el método de distancia de la deformidad de tiempo dinámico (DTW). Seleccionar e identificar la técnica de determinación de número de grupos más eficaz entre el método del codo y el método de silueta para agrupar el tráfico de las celdas LTE en la red móvil de la operadora WOM	Se puede utilizar algoritmos de agrupamiento en aprendizaje de máquina no supervisado para explicar el patrón de tráfico de las celdas LTE en una determinada zona geográfica.				operadora de comunicaciones WOM. **Muestra** Las celdas LTE que abarcan la red nacional chilena de la operadora de comunicaciones WOM que cuenten con 98% de disponibilidad. **Métodos de Análisis de Datos** Análisis cuantitativo de datos Algoritmos de agrupamiento en aprendizaje de máquina no supervisado

ANEXO 3: INSTRUMENTOS DE RECOLECCIÓN DE DATOS

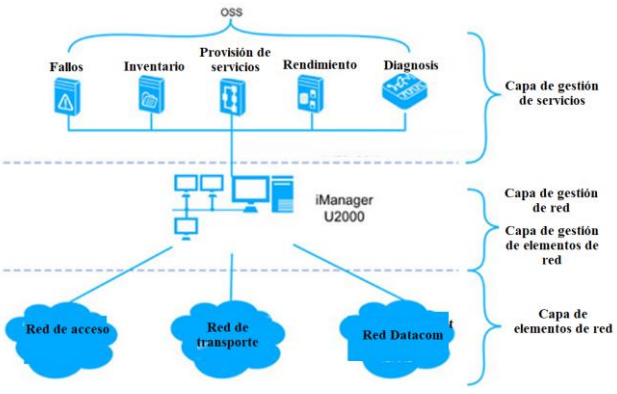
Figura 65

Antenas LTE

Nota. Vista panorámica de 2 antenas, la interfaz principal de comunicación entre una estación base y el equipo celular. Tomado de base de datos de la operadora de telecomunicaciones WOM.

Figura 66

Arquitectura OSS de una red móvil



Nota. Arquitectura OSS de una red móvil, aquí se muestra que el elemento principal de esta topología es el iManager U2000 el cual es una plataforma que se encarga de almacenar toda la información referente al rendimiento de la red.

ANEXO 4: GLOSARIO DE TÉRMINOS

LTE (Long Term Evolution) Es un estándar de comunicaciones móviles de alta velocidad diseñado para teléfonos móviles y dispositivos de datos inalámbricos.

4G Cuarta generación de tecnología de telecomunicaciones móviles que mejora las capacidades de la red 3G, permitiendo velocidades de transferencia de datos mucho más rápidas.

Red de paquetes Una red que transfiere datos divididos en bloques (o paquetes) a través de Internet y otras redes de comunicaciones.

VoLTE (Voz sobre LTE): Es una tecnología que permite realizar llamadas de voz sobre la red LTE.

MIMO (Múltiple entrada/múltiple salida): Es una tecnología que utiliza múltiples antenas en el transmisor y receptor para mejorar la comunicación.

Análisis de datos: Proceso de inspección, limpieza transformación de datos con el objetivo de descubrir información útil, informar conclusiones y apoyar la toma de decisiones.

Big Data: Término que se refiere a conjuntos de datos tan grandes y complejos que requieren métodos avanzados para su procesamiento y análisis.

Minería de datos: Proceso de descubrimiento de patrones en grandes conjuntos de datos utilizando métodos de inteligencia artificial, aprendizaje automático, estadísticas y sistemas de base de datos.

Indicador de rendimiento clave: Indicadores clave de rendimiento que ayudan a una organización a definir y medir el progreso hacia sus objetivos.

Calidad de servicio: Es una medida de la calidad de las condiciones de la red que puede incluir factores como velocidad de transferencia de datos, tasa de error, latencia, etc.

Itinerancia: Capacidad que permite a los usuarios de telefonía móvil hacer y recibir llamadas, enviar y recibir datos, o acceder a otros servicios cuando viajan fuera del área geográfica de su proveedor de red.

Computación en la nube: Tecnología que permite el acceso remoto a software, almacenamiento de archivos y procesamiento de datos a través de Internet en lugar de en un servidor local o una computadora personal.

Internet de las cosas: Concepto que se refiere a la interconexión digital de objetos cotidianos con Internet.

Aprendizaje de máquina: Tipo de inteligencia artificial que permite a los sistemas aprender y mejorar a partir de la experiencia sin ser explícitamente programados.

Modelo Predictivo: Técnica de análisis de datos que utiliza datos históricos para predecir futuros eventos o comportamientos.

Redes Neuronales: Conjunto de algoritmos modelados a partir del cerebro humano, diseñados para reconocer patrones en los datos.

Banda Ancha: Tipo de conexión a Internet de alta velocidad que proporciona una transmisión de datos más rápida en comparación con las conexiones de acceso telefónico tradicionales.

FDD (Duplexación por división de frecuencia): Método que utiliza dos bandas de frecuencia separadas para la transmisión y recepción de datos.

TDD (Duplexación por división de tiempo): Método que utiliza una única banda de frecuencia para la transmisión y recepción de datos, alternando rápidamente entre transmitir y recibir.

Small Cell: Estaciones base de baja potencia que cubren áreas pequeñas y se utilizan para mejorar la cobertura y capacidad de la red en áreas densamente pobladas o de difícil acceso**.**