Biblioteca INFOTEC

Ciudad de México, a 30 de abril de 2025

VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Análisis de indicadores de calidad en una red de conmutación de paquetes"

Desarrollado por el alumno: **Carlos Alejandro Taboada Sánchez**, bajo la asesoría del **Dr. Mario Graff Guerrero** y el **Dr. José Ortiz Bejar** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que el alumno no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Dr. Juan Antonio Vega Garfias
Subgerente de Innovación Gubernamental

JAVG/jah

C.c.p. Mtra. Analy Mendoza Rosales. – Encargada de la Gerencia de Capital Humano. - Para su conocimiento.
Carlos Alejandro Taboada Sánchez. – Alumno de la Maestría en Ciencia de Datos e Información. – Para su conocimiento.

MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

“Análisis de indicadores de calidad en una red de conmutación de paquetes.”

Tesis de Maestría
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Carlos Alejandro Taboada Sánchez

Asesor:

Doctor Mario Graff Guerrero
Doctor José Ortiz Bejar

Ciudad de México, Abril, 2025.

INFOTEC POSGRADOS

Agradecimientos

Quiero expresar mi más profundo agradecimiento a todos aquellos que me han acompañado y apoyado a lo largo de este trayecto académico.

En primer lugar, a mi esposa Gabriela, cuyo amor, paciencia y comprensión fueron fundamentales para que pudiera completar esta etapa. Gracias por tu apoyo incondicional, por creer en mí en todo momento, y por ser mi mayor fuente de inspiración.

A mi hijo José María, cuya sonrisa y alegría fueron un constante recordatorio de lo que realmente importa en la vida. Tu presencia ha sido mi motivación para seguir adelante incluso en aquellos momentos donde todo parecía muy difícil. Me has recordado con tu fuerza de voluntad que en aquellos instantes de oscuridad siempre hay luz que guía nuestros pasos.

A los docentes y miembros de INFOTEC, quienes me guiaron y brindaron su valioso conocimiento. Gracias por sus enseñanzas y asesorías, por el tiempo dedicado y por fomentar en mí una pasión por la ciencia de datos y el aprendizaje continuo.

Finalmente, agradezco a todos mis compañeros, amigos y familiares quienes de una u otra manera contribuyeron a que este proyecto se hiciera realidad.

A todos, muchas gracias.

Tabla de contenido

Introducción 12

Capítulo 1. Fundamentos de redes móviles basadas en conmutación de paquetes 21

1.1 Desafíos en la evaluación del desempeño de Red 24

1.2 Necesidad de mediciones precisas para evaluar el desempeño 28

1.3 Latencia, Jitter y Pérdida de paquetes: Indicadores clave de desempeño 28

1.3.1 Latencia, definición y relevancia en redes de conmutación de paquetes 29

1.3.2 Jitter como factor en la calidad de servicio 30

1.3.4 La Pérdida de paquetes 31

1.4 Protocolo TWAMP 32

1.4.1 Funcionamiento del protocolo TWAMP 35

Capítulo 2. Metodología de análisis de datos 40

2.1 Minería de datos en redes de conmutación de paquetes 40

2.2 Metodología CRISP-DM en un proyecto de Ciencia de Datos en una red de conmutación de paquetes. 41

2.3 Etapas de la metodología 43

2.3.1 Entendimiento del negocio 44

2.3.2 Comprensión de los datos 46

2.3.3 Preparación de los datos 52

2.3.4 Modelado 52

2.3.5 Evaluación 53

2.3.6 Despliegue 54

Capítulo 3. Análisis de datos 56

3.1 Herramientas de trabajo para el análisis de datos 58

3.2 Detección de Anomalías en la red de conmutación de paquetes 59

3.3 Procedimiento de análisis de agrupaciones 67

Conclusiones 95

Referencias 99

Índice de figuras

Figura 1. Tráfico de GB del servicio móvil de acceso de Internet 13
Figura 2. Componentes TWAMP 33
Figura 3. Componentes TWAMP Cliente / Servidor 34
Figura 4. Sesiones TWAMP establecidas en una red de conmutación de paquetes (4G / 5G)..... 37
Figura 5. Metodología CRISP-DM..... 42
Figura 6. Mapa de Ciudad Juárez, Chihuahua, México..... 46
Figura 7. *Pantalla principal del aplicativo Orange Data Miner*..... 60
Figura 8. Conjunto de nodos asociados al análisis de anomalías en Orange DataMiner..... 61
Figura 9. Muestra de datos TWAMP de los sitios considerados para el análisis.. 62
Figura 10. Correlación Pearson UL_LOSTPKTS vs. el resto de las variables..... 62
Figura 11. Correlación Pearson UL_LOSTPKTS vs. el resto de las variables..... 63
Figura 12. Análisis de serie de tiempo de la latencia máxima del sitio..... 64
Figura 13. Análisis de serie de tiempo de Jitter máximo del sitio..... 65
Figura 14. Correlación de Spearman de la variable UL_JMAX (Jitter Máximo de Carga)..... 66
Figura 15. Mapa Ciudad Juárez con distribución geográfica de sitios 68
Figura 16. Normalización del conjunto de datos..... 72
Figura 17. Validación de las variables NaN del conjunto de datos..... 72
Figura 18. Matriz de correlación del conjunto de datos 73
Figura 19. Visualización de Componentes Principales..... 78
Figura 20. Biplot de las variables respecto a los componentes principales 79
Figura 21. Visualización de Componentes Principales..... 80
Figura 22. Biplot de Componentes Principales..... 80
Figura 23. Visualización Scree Plot, punto de inflexión..... 86
Figura 24. Participación en porcentaje de cada componente principal 87
Figura 25. Número de clusters considerando el método Silhouette..... 88
Figura 26. Número de clústers considerando el método GAP 89
Figura 27. Agrupaciones tomando en consideración distancia euclidiana..... 90
Figura 28. Agrupaciones tomando en consideración distancia Manhattan 91
Figura 29. Dendograma, k=5..... 92
Figura 30. Diagrama de densidad..... 93

Siglas y abreviaturas

3GPP	3rd Generation Partnership Project
ACP	Análisis de Componentes Principales
AD	Anomaly Detection
AIC	Akaike Information Criterion
AMPS	Advanced Mobile Phone System
CRISP-DM	Cross Industry Standard Process for Data Mining
DL	Downlink
E2E	End to End
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
Gbps	Giga bits por segundo
GSM	Global System for Mobile Communications
GSMA	Global System for Mobile Communications Association
IENT	Instituto Europeo de Normas de Telecomunicaciones
IETF	Internet Engineering Task Force
IFT	Instituto Federal de Telecomunicaciones
IoT	Internet of Things
IP	Internet Protocol
ISP	Internet Service Provider
ITU	International Telecommunication Union
LTE	Long Term Evolution
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Erro
MBB	Mobile Broadband
Mbps	Mega bits por segundo
ML	Machine Learning
MME	Mobility Management Entity
MOS	Mean Opinion Score
PCA	Principal Component Analysis igual a ACP
POCID	Prediction of Change in Direction
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RFC	Request for Comments
RMSE	Root Mean Squared Error
RTT	Round Trip Time
SFP	Small Form-factor Pluggable
SGW	Serving Gateway
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TTI	Transmission Time Interval
TTL	Time to Live

TWAMP	Two Way Active Measurement Protocol
UE	User Equipment
UIT	Unión Internacional de Telecomunicaciones
UL	Uplink
URLLC	Ultra-Reliable Low-Latency Communications
UTRAN	UMTS – Terrestrial Radio Access Network
VoIP	Voice over IP
VoLTE	Voice over LTE
VoWiFi	Voice over WiFi

Glosario

***Backhaul*:** Es un enlace establecido entre una estación y un centro de conmutación o de datos.[1]

***Churn*:** Es una métrica que muestra los clientes que dejan de hacer negocios con una empresa o un servicio en particular, también conocido como deserción de clientes.[2] El *Churn*, puede ser provocado por varias causas, una de ellas es la percepción de la calidad del servicio recibido.

***Jitter*:** Se define como una variación en el retraso de los paquetes recibidos. En el lado del envío, los paquetes se envían en un flujo continuo con una separación uniforme entre ellos. Debido a la congestión de la red, una cola de procesamiento inadecuada o errores de configuración, este flujo constante puede volverse irregular o el retraso entre cada paquete puede variar en lugar de permanecer constante.[3]

Latencia: El tiempo que lleva a una señal propagarse a través de un medio de transmisión o dispositivo. Es el retraso en la comunicación de la red. Muestra el tiempo que tardan los datos en transferirse a través de la red de conmutación de paquetes. Las redes con un mayor retraso o retardo tienen una latencia alta, mientras que las que tienen tiempos de respuesta rápidos tienen una latencia baja.[4]

Perdida de Paquetes: (*Packet Loss*): Al conectar una red de conmutación de paquetes, se empiezan a enviar y recibir unidades de datos llamadas paquetes entre los diferentes elementos que conforman la red. Cuando uno o más paquetes no logran viajar entre el enrutador y el dispositivo, se produce una pérdida de paquetes.

---

[1] *Gartner Glossary, Information Technology Glossary*, Gartner, Disponible en internet en: [https://www.gartner.com/en/information-technology/glossary/backhaul]

[2] Curi, Mariana, *Customer Churn in Telecom Segment, Towards* Data Science, 21 de Julio de 2020, Estados Unidos. Disponible en internet en: [https://towardsdatascience.com/customer-churn-in-telecom-segment-5e49356f39e5/]

[3] Cisco, *Understanding Jitter in Packet Voice Networks (Cisco IOS Platforms)*, documento: 18902. Febrero 2006. Disponible en internet en: [https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/18902-jitter-packet-voice.html]

[4] Amazon, AWS, ¿Qué es la latencia de red?. Disponible en Internet en: [https://aws.amazon.com/what-is/Latency/]

Cuanto más lejos tengan que viajar los paquetes, mayores serán las probabilidades de que se pierdan.[5]

***Quality of Service*: (*QoS*):** Un acuerdo negociado entre un usuario y un proveedor de red que proporciona cierto grado de capacidad confiable en red.[6] Se establecen las características mínimas del servicio.

Small Form-Factor Pluggable (SFP): Un puerto SFP es una ranura en un dispositivo de red o computadora en la que se insertan transceivers conectables de factor de forma pequeño (*SFP*). Un transceiver SFP, también conocido como módulo SFP, es simplemente un componente metálico del tamaño de un meñique, intercambiable en caliente, que, cuando se conecta a otro dispositivo mediante un cable, permite la transmisión de datos.[7]

Streaming: Técnica que permite la transmisión continua y unidireccional de datos de audio y/o video a través de internet y, más recientemente, a través de una red móvil. A diferencia de los archivos de audio (por ejemplo, MP3) y de películas (por ejemplo, MPEG) que primero deben descargarse, la transmisión de medios comienza a reproducirse unos segundos después de la solicitud. La transmisión requiere un codificador de transmisión (que convierte la fuente de audio o video en un flujo de datos), un servidor de transmisión que entrega los medios codificados a través de una red y un reproductor de medios cliente que coopera con el servidor para entregar datos ininterrumpidos. Parra compensar las variaciones en la calidad y la latencia de la red, el cliente almacena en búfer unos segundos de audio o video antes de comenzar la transmisión y luego intenta mantenerse a la vanguardia durante la reproducción.[8]

---

[5] Fortinet, Packet Loss Meaning. Disponible en Internet en: [https://www.fortinet.com/resources/cyberglossary/what-is-packet-loss]

[6] *Gartner Glossary, Information Technology Glossary*, Gartner, Disponible en internet en: [https://www.gartner.com/en/information-technology/glossary/qos-quality-of-service]

[7] Daniel, Brett. *What is an SFP (Small Form Factor Pluggable) Port?*. Trenton Systems. Abril 2020. Disponible en intenet en: [https://www.trentonsystems.com/en-us/resource-hub/blog/what-is-an-sfp-port#:~:text=An%20SFP%20port%20is%20a,for%20the%20transmission%20of%20data.]

[8] Gartner, *Gartner Glossary*, Information Technology Glossary, Streaming. Disponible en Internet en: [https://www.gartner.com/en/information-technology/glossary/streaming]

Introducción

En la evolución continua de las tecnologías de comunicaciones móviles, la importancia de garantizar un desempeño confiable y eficiente es crucial para garantizar una buena experiencia de servicio para el usuario final. A partir de la estandarización de redes de cuarta generación (4G), y posteriores, el tráfico de voz y datos de los usuarios móviles se realiza mediante la implementación de redes de conmutación de paquetes (*Packet Switched Networks*). Este tipo de redes son más eficientes pues se permite compartir recursos en el medio de transporte, de esa forma, múltiples flujos de datos en el mismo ancho de banda.

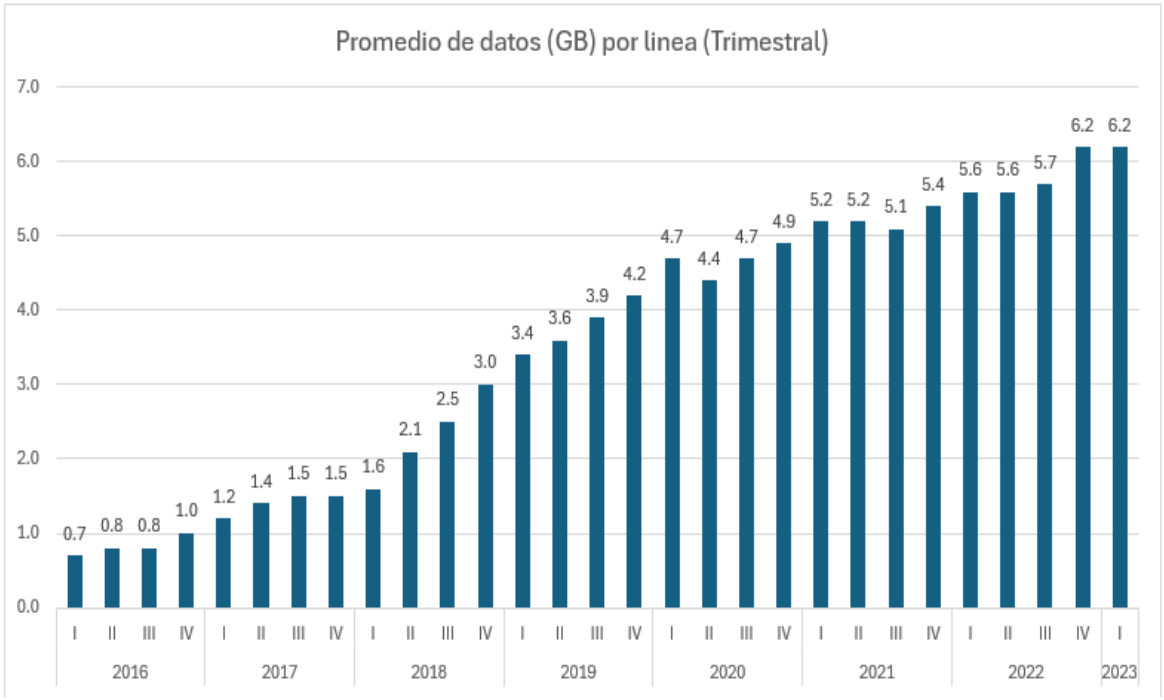
La estandarización de las redes obliga a cada operador móvil a implementar las funciones y herramientas necesarias para cumplir con la prestación del servicio y también a cumplir con los requisitos de calidad de servicio definidos por los organismos reguladores; en México, el Instituto Federal de Telecomunicaciones (IFT) es el encargado de definir y hacer cumplir las mediciones necesarias para que cada usuario del servicio reciba un buen servicio. A nivel internacional también existen organismos de estandarización, por ejemplo: 3GPP y la Unión Internacional de Telecomunicaciones[9] (UIT). Este conjunto de organismos concuerda que la evolución de la demanda de uso de datos por línea móvil se ha incrementado con cada generación de red móvil, lo cual puede ser contrastado en la Figura 1; el incremento de la demanda necesariamente implica un incremento en los recursos necesarios para procesar dicho tráfico.

Dado el incremento de tráfico, es menester que cada operador de redes de conmutación de paquetes, implemente acciones preventivas y/o correctivas para mantener la calidad de servicio para los usuarios finales. Existen indicadores como: Latencia, Jitter y Pérdida de Paquetes (*Packet Loss*) que son considerados para realizar el monitoreo de desempeño y calidad; así como la salud de cada uno de los medios de transmisión que transportan datos en la red.

[9] Union Internacional de Telecomunicaciones (ITU). Quality of Service – Regulation Manual. 2017. Disponible en internet en: [https://www.itu.int/pub/D-PREF-BB.QOS_REG01-2017]

Una regla simple, no escrita, asocia a valores altos en magnitud de los indicadores mencionados implican un incremento en la degradación en el servicio brindado al usuario final, lo cual es sinónimo de disminución en la calidad de la experiencia generando quejas del servicio y en última instancia a falta de atención por parte del operador de servicios un incremento en la tasa de abandono (*Churn*).

Figura 1. *Tráfico de GB del servicio móvil de acceso de Internet*



Nota. El gráfico representa el consumo promedio de datos por cada línea móvil. Elaboración propia, con información del Instituto Federal de Telecomunicaciones (IFT)[10]. Consumo promedio trimestral por línea.

En el contexto de las redes de conmutación de paquetes existen referencias del empleo del protocolo de Medición Activa Bidireccional (TWAMP[11]) definido por el IEFT en la especificación 5357, para la medición de la calidad de servicio: (Kokac & Zaim, 2018), (Chaudhari & Biradar, 2015), (Ekelin *et al.*, 2015), es importante destacar, la forma de implementación de este protocolo y las herramientas para la

[10] Los datos mostrados corresponden al Banco de Información de Telecomunicaciones del Instituto Federal de Telecomunicaciones, se presenta una gráfica similar disponible en internet en: [https://www.ift.org.mx/comunicación-y-medios/comunicados-ift/es/crece-295-el-uso-de-datos-por-linea-en-el-internet-movil-en-los-ultimos-5-anos-comunicado-882023-17]

[11] TWAMP por sus siglas en inglés: *Two-Way Active Measurement Protocol*.

medición son dependientes de cada proveedor de servicio. Existen escenarios donde existe interconexión de redes de operadores y se emplea también este protocolo de común acuerdo para la medición de calidad.

En una red móvil basada en conmutación de paquetes existen diferentes puntos de interconexión de cada uno de sus componentes y funciones. Este trabajo propone el análisis avanzado de las mediciones TWAMP ocupando una metodología orientada a un proyecto de Ciencia de Datos con el objetivo de identificar patrones y detectar anomalías en las observaciones identificando aquellas variables que inciden directamente en el desempeño y calidad, lo cual establece el camino para realizar acciones predictivas y correctivas de la red, así como su optimización.

La detección temprana de problemas relacionados con Latencia, Jitter y Pérdida de paquetes, permite aplicar umbrales de calidad de servicio (*QoS*) y también de umbrales de experiencia de usuario (*QoE*). Un usuario final difícilmente, y sin tener un mecanismo dentro del móvil para medición, percibe variaciones importantes en estas métricas. Sin embargo, el usuario percibe fallas mayormente cuando utiliza aplicaciones en aplicaciones intensivas en datos y cercanas a tiempo real como: *Streaming* de video, *Streaming de audio*, juegos de video, entre otras.

Mientras que la Latencia ayuda a identificar rutas de la red o nodos de esta, con retraso, el Jitter detecta la variabilidad de la Latencia, la cual afecta servicios en tiempo real y la Pérdida de paquetes ayuda a identificar los enlaces que tienen una mayor degradación la implementación de un análisis automático permite monitorear de forma precisa el rendimiento de la red sin la necesidad de tener una revisión uno a uno de los elementos de la red.

Una red móvil basada en conmutación de paquetes, está compuesta por uno o más sitios celulares, estos pueden conectarse a nodos agregadores, a su vez estos agregadores pueden conectarse al Núcleo de Red por lo cual el número de elementos de red interconectados se incrementa para realizar una medición de calidad. Por ejemplo, por cada sitio celular existe un enlace de transporte hacia un nodo agregador o hacia un punto / oficina de conmutación; si consideramos una red con 10,000 sitios celulares cada uno de ellos con un ruteador en sitio y se conectan

a un nodo agregador, implica al menos 10,000 puntos de conexión donde se realizan mediciones para verificar la calidad del transporte. Mediante el empleo de protocolo TWAMP, por cada uno de estos puntos se establece una sesión que mide la calidad de servicio de manera uniforme dentro de la red; La generación de las mediciones TWAMP depende de la implementación realizada por cada operador, en este caso se considera que cada elemento genera una medición por minuto. En el caso más simple, esta interconexión de redes puede ejemplificarse como un grafo conexo donde existe una sesión TWAMP entre cada par de nodos.

Suponiendo el número total de sitios de la red en 10,000 sitios y considerando las mediciones minuto a minuto de cada uno de ellos, serían 1,440 por cada sitio, por lo que por cada día, asumiendo que todos los sitios están operativos, se tiene un total de 14,400,000 eventos TWAMP, en un mes de treinta días, serían 432,000,000 eventos registrados esto implica que debemos considerar y dimensionar la solución para almacenar, procesar y analizar la información generada para posteriormente el proveedor de servicio realice selección de actividad preventiva o correctiva en caso que exista alguna falla en la red que afecte el impacto al usuario final.

En el análisis propuesto en este trabajo, se consideran dos factores en la medición de los enlaces de transporte entre la Red de Acceso y los sitios agregadores. El primer factor, tiene que ver con la capacidad de transmisión del medio de transporte por ejemplo si el enlace punto a punto tiene una capacidad de 1 Mbps, 10 Mbps, 100 Mbps, 1Gbps o son de mayor capacidad; el segundo factor está asociado al desempeño del mismo enlace. Se busca como resultado de este trabajo se cimienten las bases para una medición de la salud de la red, ocupando un análisis de datos, y que permitan al operador de servicio pueda buscar establecer acciones preventivas para desplegar una mayor capacidad de procesamiento de datos de la red.

Con el supuesto previamente descrito de 432,000,000 mediciones TWAMP en un mes, el empleo de análisis manual se descartó para evitar recursos dedicados a procesar y revisar los datos. El análisis manual, inhibe lograr una visión extremo a extremo (*E2E*) de la red. Existe otro factor a considerar en el análisis y es la red móvil al ser un sistema dinámico los patrones de tráfico cambian constantemente,

por ejemplo: Incrementar la capacidad de ancho de banda de los enlaces, un corte de fibra, desvío de tráfico a otra ubicación, falla de energía eléctrica o falla en las baterías de un sitio por vandalismo o robo.

Se ha mencionado que en cada conexión TWAMP se establece entre dos puntos de conexión, en cada uno de los puntos que inician la conexión, normalmente se implementa un dispositivo denominado punto de validación inteligente (*SFP*) que realiza pruebas de desempeño desde el sitio remoto hacia un nodo centralizado sin impactar el tráfico de datos cursado. Al ser un protocolo bidireccional cada una de las mediciones TWAMP contiene variables tanto de carga (*Uplink*) como de descarga (*Downlink*), estas mediciones no nos indican cuales son las variables que inciden en mayor medida en el desempeño de la red móvil simplemente es el estado de un medio de transporte en un momento específico.

Objetivo general

La medición y el análisis del comportamiento de la red de conmutación de paquetes son cruciales para comprender el desempeño de la red. En combinación ambos permiten identificar oportunidades de optimización o de identificar restricciones de capacidad en la red que pueda derivar en un impacto a la experiencia de servicio de los usuarios. Este trabajo propone desarrollar un proyecto de análisis avanzado de datos basado en mediciones de protocolo TWAMP para evaluar y caracterizar la calidad del servicio en redes de transporte de conmutación de paquetes. Con el fin de establecer un marco de referencia base que pueda ser utilizado en diferentes proyectos de datos dentro del proveedor de servicio, se considera el uso de la metodología CRISP-DM.

El empleo de CRISP-DM aplicado a un proyecto de Ciencia de Datos, permite establecer un marco metodológico claro y permite establecer desde el principio las fases por las que el proyecto debe transitar sin traer ambigüedad o incertidumbre en las acciones a realizar en el proyecto de Ciencia de Datos.

Objetivos específicos

1. Análisis y Preprocesamiento de Datos TWAMP

- Diseñar e implementar una propuesta de ingesta y agregación de las mediciones TWAMP para un mercado en particular.
- Desarrollar métodos de preprocesamiento para la limpieza y normalización de datos, e incluir el análisis de datos para identificar datos erróneos y/o faltantes e identificar condiciones atípicas en la red.

2. Desarrollo de Modelos Analíticos

- Implementar estrategias de análisis de datos para la identificación de variables y su relevancia para identificar acciones asociadas a mejorar Latencia, Jitter y Pérdida de paquetes.
- Proponer el ejercicio para la detección de patrones y anomalías en las métricas de red que servirá de base para un modelo predictivo para anticipar degradaciones en la calidad de servicio.
- Mediante la técnica de agrupamiento (*Clustering*) identificar comportamientos similares en términos de Latencia, Jitter y Pérdida de paquetes con comportamientos similares.

3. Evaluación de Métricas de Calidad

- Analizar la correlación entre diferentes métricas TWAMP (Latencia, Jitter y Pérdida de paquetes) y su impacto en la calidad del servicio.
- Validar la factibilidad de establecer umbrales dinámicos para la identificación de problemas de rendimiento.

4. Validación y Optimización

- Realizar pruebas de validación del sistema en diferentes escenarios de red.
- Optimizar los modelos desarrollados para mejorar su precisión y tiempo de respuesta.
- Establecer el marco de referencia para evaluar calidad de servicio ocupando las variables más significativas.

Resultados esperados

1. Una propuesta para analizar los eventos TWAMP en una zona geográfica específica y establecer el marco de referencia para la generalización de las mediciones dentro de la red nacional u ocuparse en otras redes fuera de México.

2. Mediante el uso de la herramienta Orange DataMiner, establecer los pasos para realizar un análisis de serie de tiempo para la detección de anomalías en las variables de Latencia, Jitter y Pérdida de paquetes.

3. Documentación detallada de la metodología y resultados.

Alcance de la investigación

El conjunto de datos contempla los eventos TWAMP generados en el intervalo de tiempo que va del primero de agosto de 2022 (00:00:00) y hasta el 31 de agosto del mismo año (23:59:59). Las mediciones se generan a razón de una por minuto por cada sitio celular por cada hora del día. La zona geográfica considerada en este trabajo corresponde a Ciudad Juárez en Chihuahua, al momento de la captura de los datos existía un universo de 149 sitios conectados con sitios de agregación. A partir del análisis exploratorio de datos preliminar, se consideró utilizar una agregación de tiempo a nivel de hora; esto con el fin de identificar tendencias de comportamiento de tráfico y/o de las fallas o incidencias observadas en ese intervalo de tiempo.

La consideración de la zona geográfica de Ciudad Juárez se debe a que es un mercado fronterizo (México – Estados Unidos); por la misma naturaleza de la interacción de la señal de servicio móvil en ambos países, es común encontrar puntos donde hay interferencia de señal, en ambos lados de la frontera, y que puede dar un falso positivo de una experiencia pobre para un usuario final en cualquiera de los dos países y se asuma de forma errónea que es un problema de calidad debido a valores altos de Latencia, Jitter y Pérdida de paquetes. Otra consideración es que al ser un mercado fronterizo existe infraestructura de transporte que puede ser compartida en ambos países, como los puntos de salida hacia internet a través de los proveedores de servicio (*ISP*), el cual procesa el tráfico generado por los usuarios y cuyo destino es un nodo / host / servicio en Internet, cualquier degradación de los indicadores de Latencia, Jitter y Pérdida de paquetes genera

impacto negativo en servicio recibido por el cliente. Otros criterios, considerados para el uso de esta ubicación tiene que ver con: Número de habitantes, orografía, frecuencias licenciadas, entre otros. Los sitios considerados al momento de la extracción de datos tenían capacidad dual en términos de tecnología al soportar 3G y 4G, actualmente la mayoría de ellos soportan 4G/5G, recordando que 5G también es una tecnología basada en una red de conmutación de paquetes.

Contribuciones

Innovación en el Análisis de Datos de Red

- Desarrollo de un marco de referencia que integra la metodología de gestión de proyecto de ciencias de datos, así como el uso de mediciones de un dominio en particular, telecomunicaciones al incluir eventos TWAMP, superando las limitaciones de los métodos tradicionales de monitoreo.
- Establecer las bases para la automatización de procesos de diagnóstico que tradicionalmente requieren intervención manual
- Es una propuesta de nuevos métodos para la caracterización de la calidad de servicio en redes de transporte. Establecimiento de correlaciones entre diferentes métricas TWAMP y su impacto en la experiencia del usuario.
- Propuesta para un sistema integral que combina análisis cercano a tiempo real con predicciones a largo plazo. Desarrollo de métodos escalables para el procesamiento de grandes volúmenes de datos de red.
- Reducción significativa en el tiempo de detección y resolución de problemas de red. Optimización de recursos basada en predicciones precisas de demanda y comportamiento de la red.
- Mejora en la calidad de servicio a través de la detección temprana de degradaciones.

Capítulo 1

Fundamentos de redes móviles basadas en conmutación de paquetes

Capítulo 1. Fundamentos de redes móviles basadas en conmutación de paquetes

Desde las primeras redes móviles (AMPS, GSM, 2G) y hasta las redes de tercera generación (3G), se ocupó mayormente la conmutación de circuitos como mecanismo para interconexión y transporte de los diferentes elementos de red; si bien los servicios brindados eran de alta confiabilidad, también estaban presentes limitaciones en cuanto a la capacidad para transportar datos del plano de usuario por ejemplo, el caso de GSM permitió la transmisión y recepción de datos con velocidades máximas de 9.6 Kbps[12] en su versión inicial y hasta 14.4 Kbps en versiones posteriores de GSM.

El desarrollo tecnológico de las terminales móviles permitió también la aparición de nuevos servicios con un requerimiento mayor de ancho de banda limitado en redes de conmutación de circuitos. En estas redes, había un límite tecnológico en cuanto a velocidades de carga (*UL*) y descarga (*DL*) de datos. En la búsqueda de lograr mayores anchos de banda, organismos internacionales de estandarización como: 3GPP[13], UIT[14] y GSMA[15] desarrollaron los estándares con la visión y guía de la evolución tecnológica de las redes móviles. Esta visión permitió que los operadores móviles definir un plan de renovación tecnológica hacia redes de mayor capacidad. A partir de la versión ocho de los estándares de redes móviles de 3GPP se consideraron los lineamientos para la evolución de redes móviles a lo que se conoce cuarta generación, donde mayormente el transporte entre los diferentes elementos de red se basa en la conmutación de paquetes con la premisa de ofrecer también servicios con un mayor requerimiento para velocidades transmisión y recepción de

[12] Kbps por siglas en inglés (*Kilobits per second*)
[13] El Proyecto de Asociación de Tercera Generación (3GPP) es un proyecto de colaboración entre un grupo de asociaciones de telecomunicaciones con el objetivo inicial de desarrollar especificaciones aplicables a nivel mundial para sistemas móviles de tercera generación (3G). En fases posteriores se ha encargado de la definición de estándares de redes móviles de cuarta generación (4G), quinta generación (5G) así como la referencia preliminar para redes móviles de sexta generación (6G). [https://www.3gpp.org]
[14] Unión Internacional de Telecomunicaciones, por sus singlas en inglés *ITU*, *International Telecommunications Union*.
[15] Asociación GSM, por sus siglas en inglés *GSMA, GSM Association.*

datos tanto en la red móvil como en los equipos de usuarios. La versión ocho de 3GPP también definió los umbrales de calidad de servicio (Latencia, Jitter y Pérdida de paquetes) en las redes de conmutación de paquetes.[16]

La Unión Internacional de Telecomunicaciones (UIT), considera que las redes de conmutación de paquetes de cuarta generación deben cubrir las siguientes necesidades[17]:

- Velocidades de datos más altas y mayor eficiencia espectral.
- Un sistema optimizado únicamente con conmutación de paquetes.
- Uso de frecuencias licenciadas para garantizar la calidad de los servicios. Evitar interferencias entre operadores de tecnología.
- Experiencia siempre activa, el dispositivo móvil se mantiene siempre conectado (esto permite reducir la latencia del plano de control).
- Establecer criterios aceptados por la industria para la medición y aseguramiento de la calidad en las redes móviles.

En la versión diez de 3GPP, se estableció la transición hacia redes de cuarta generación avanzadas, comúnmente referenciadas como redes de LTE Avanzado, logrando así una mayor capacidad de transmisión y recepción desde un dispositivo móvil. Como resultado de la implementación de estándares de cuarta generación (LTE), el 3GPP considera que se lograron, los siguientes resultados:

- Arquitectura de redes comunes.
- Uso de Frecuencias (flexibilidad)
- Incremento en las tasas de transmisión / recepción de bits
- Reducción de latencia entre los componentes de la red móvil
- Definición de estándares de calidad de la red móvil.

[16] 3GPP, *Study on media handling aspects of Radio Access Network (RAN) delay Budget reporting in Multimedia Telephony Service*, 3GPP. TR 26.910 versión 16.0.0 release: 16. Disponible en internet en: [https://www.etsi.org/deliver/etsi_tr/126900_126999/126910/16.00.00_60/tr_126910v160000p.pdf]

[17] ITU, *4G to 5G networks and standard releases, Training on Traffic engineering and advanced wireless network planning*, 2019. Disponible en internet en: [https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/SiteAssets/Pages/Events/2019/ITU-ASP-CoE-Training-on-/3GPP_4G%20to%205G%20networks%20evolution%20and%20releases.pdf]

- Orientada a conexión (Full-IP), redes de conmutación de paquetes, que busca reducir los problemas de *QoS*, así como normalizar la integración de protocolos a menores costos.

Como resultado de la implementación de estándares de cuarta y/o quinta generación se fomentó el desarrollo de nuevos servicios basados en conmutación de paquetes como por ejemplo *streaming* de video, de audio, juegos de video, realidad aumentada, voz sobre red de paquetes, entre otros. El resultado fue un crecimiento exponencial en el tráfico generado por los usuarios y procesado por la red móvil, este incremento fue superior durante la pandemia de Sars-COVID-2, de acuerdo con la UIT:

> Se estima que las tasas de tráfico mundial de banda ancha móvil alcanzaron los 913 exabytes (EB[18]) en 2022, más del doble del tráfico de 2019 (419 EB). En comparación con 1991 en 2019, se estima que las tarifas de tráfico de banda ancha fija aumentaron a 4378 EB en 2022 (casi cinco veces las del tráfico de banda ancha móvil). Entre 2019 y 2023, el tráfico de banda ancha móvil y fija tuvo un crecimiento promedio anual estimado del 30 por ciento, con una tasa máxima de crecimiento al comienzo de la pandemia de COVID-19 en 2020.[19]

La demanda de nuevos servicios no sólo requiere un mayor ancho de banda para procesar el volumen de tráfico e incrementar la velocidad de transferencia de datos, sino también requiere asegurar la confiabilidad y la calidad de la transferencia de datos. Por tal razón, es necesario contar con indicadores que permitan medir el estado de la transmisión de paquetes en la red. Indicadores como: Latencia, Jitter y Pérdida de paquetes requieren una medición continua de los diferentes puntos de la red para identificar fallas o congestiones de datos.

En 2019, 3GPP publicó la versión quince de los estándares de red, estos estándares incluían los lineamientos para la implementación de redes de quinta generación (5G)[20] también basadas en la conmutación de paquetes. En 2022 se oficializó el

[18] Exabyte. Definición
[19] Unión Internacional de Telecomuicaciones (UIT). Hechos y figuras. Tráfico de Internet. Disponible en internet en: [https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-internet-traffic/]
[20] Reuters, Who was first to launch 5G? Depends who you ask By Kenneth Li and Ju-Min Park April 6, 2019 https://www.reuters.com/article/idUSKCN1RH1V1/

primer despliegue de una red 5G en México[21]. La tecnología de quinta generación ofrece mejoras con respecto a 4G, en particular mejora en la velocidad de carga (UL) y descarga (DL), mayor ancho de banda y una promesa de reducción de latencia. En la actualidad se discute la evolución hacia redes de sexta generación (6G)[22], los organismos de estandarización consideran también como punto de partida el uso de conmutación de paquetes. Garantizar una buena experiencia de los servicios móviles implica necesariamente mantener un desempeño óptimo de las redes móviles mediante la medición continua de indicadores de red.

1.1 Desafíos en la evaluación del desempeño de Red

Una red móvil de conmutación de paquetes para propósitos de análisis de tráfico está conformada por tres dominios: La red de acceso de radio (*RAN*) que el medio entre los sitios celulares y los usuarios finales, la red de transporte o de transmisión, que va de los sitios celulares a otros sitios o a nodos agregadores y el núcleo de la red, que es la interfaz para las redes de conmutación. Cada uno de estos dominios realiza intercambio de paquetes de datos ya sea de plano de control (*Control Plane*) o plano de usuario (*User Plane*).

Para ejemplificar el tráfico de plano de control supongamos el proceso de establecimiento de una llamada de voz, todos los flujos que se procesan antes de que el usuario pueda hablar. Mientras que cuando el usuario ya tiene establecido un canal para la voz, el tráfico subsecuente, es decir la conversación o el intercambio de mensajes voz, corresponde al tráfico de plano de usuario. De la misma forma si suponemos el establecimiento de una sesión sobre https, el tráfico de plano de control corresponde al establecimiento de la sesión, intercambio de certificados, mientras que la información de plano de usuario corresponde al video o audio que el usuario descarga hacia su dispositivo.

[21] Ericsson. Catalina Urita. *How launching the first 5G commercial network in Mexico will transform the region*. Disponible en internet en: [https://www.ericsson.com/en/blog/2022/2/how-launching-the-first-5g-commercial-network-in-mexico-will-transform-the-region]

[22] Nokia, *6G Explained*. Disponible en internet en: [https://www.nokia.com/about-us/newsroom/articles/6g-explained/]

La promesa de las redes de quinta generación (5G) de ofrecer mejores servicios a los usuarios y la comunicación de baja latencia ultra confiable (*URLLC*[23]), es uno de los casos de uso considerados[24]. La comunicación *URLLC* se utiliza en aplicaciones de misión crítica que requieren una conexión garantizada y baja latencia. Por ejemplo, en el uso de operaciones quirúrgicas a distancia o automóviles autónomos. Se estableció en los estándares 3GPP una confiabilidad teórica de red con 99.999% y extremadamente baja Latencia, aproximadamente un milisegundo para la transmisión de datos; es de esperarse que la aparición de una Latencia mayor impida el despliegue de los servicios previamente mencionados por los resultados adversos que podrían generarse.

En las redes móviles existe también una característica importante, la movilidad de los usuarios dentro de la misma red, es decir se asume que los usuarios pueden estar en cualquier punto de la red, siempre que exista cobertura móvil. Mientras más usuarios móviles registrados tenga un operador móvil se puede presentar el escenario adverso de la congestión de datos derivado de la variabilidad en el procesamiento de tráfico. Los indicadores de Latencia, Jitter y Pérdida de Paquetes son de nueva cuenta una forma de entender la salud de la red.

Las redes de conmutación de paquetes también brindan nuevas oportunidades como el uso de Internet de las Cosas (*IoT*), bajo este concepto se permite la conectividad a otro tipo de dispositivos como sensores, elevadores, equipos de rastreo, entre otros; cada dispositivo tiene requisitos específicos de conectividad contribuye al incremento de tráfico de datos en la red.

Con esta diversidad en el uso de datos y la variabilidad en la carga/descarga de tráfico de los dispositivos, la red puede presentar eventos atípicos donde existe una gran aglomeración de usuarios y/o dispositivos por ejemplo de este tipo de eventos son: Las concentraciones masivas de personas (conciertos, manifestaciones, eventos deportivos, etc.), eventos fortuitos como sismos, incendios, entre otros, son

[23] URLLC por sus siglas en Inglés *Ultra-reliable low latency communications*.
[24] Instituto Federal de Telecomunicaciones. Comunicaciones ultra fiables y de baja latencia. Disponible en internet en: [https://sensor5g.ift.org.mx/planeacionTendenciasComunicacionesUltraFiables]

un ejemplo donde hay un incremento de tráfico en la red que es difícil predecir con certeza. Otros eventos como liberación de nuevas aplicaciones o un evento fortuito que hace que los usuarios demanden más tráfico a la red, por ejemplo, la demanda de compra de abarrotes desde los dispositivos móviles durante la pandemia de Sars-COVID-2.[25]

Una mayor demanda de tráfico sobre cuya capacidad está limitada, puede ocasionar congestión de la red, donde no es posible procesar más datos, los datos son descartados y la condición de Pérdida de paquetes se hace presente. La gestión eficiente de esta diversidad de dispositivos, una estimación eficiente de la capacidad, así como la distribución equitativa de recursos son cruciales para mantener un desempeño óptimo de la red. La medición del desempeño de la red de conmutación de paquetes es esencial para asegurar un buen servicio hacia los usuarios finales[26].

Se ha hablado previamente de los conceptos de Latencia, Jitter y Pérdida de paquetes, por lo que, en este capítulo se realizará una revisión de los conceptos de Latencia, Jitter y Pérdida de paquetes como parámetros críticos en una red de conmutación de paquetes, así como su impacto en la calidad de la experiencia del usuario. Adicionalmente, se evaluará la injerencia de estas variables en los eventos de falla de la red (congestión de la red de transporte y variabilidad en la transmisión de UL / DL).

La Latencia representa el tiempo que tarda un paquete de datos en viajar desde su origen hasta su destino e impacta directamente en la velocidad de respuesta de las aplicaciones y servicios. El Instituto Federal de Telecomunicaciones (IFT) define de la siguiente forma a la Latencia:

> Es el tiempo que tardan en entregarse los paquetes de información que son enviados cuando solicitas una página de Internet, un video o la descarga de un archivo, etc. Este tiempo de entrega depende de la

[25] Deloitte, *Digital Consumer Trends* en México 2020. México, 2021. Disponible en internet en: [https://www2.deloitte.com/content/dam/Deloitte/mx/Documents/technology/2021/DigitalCT_2020.pdf]

[26] El buen servicio se asocia cuando menos al cumplimiento de los lineamientos de calidad.

distancia entre el origen y el destino, la congestión de la red y el ancho de banda.[27]

Cuando un usuario utiliza un dispositivo móvil para usar una aplicación de reproducción de video, y si en ese instante en la red existe una Latencia alta, el usuario puede percibir una demora en la reproducción de video. En otros escenarios el usuario puede percibir que el video tiene pausas o las imágenes se muestran borrosas y toma un tiempo completar el proceso de descarga. En un mundo donde la transmisión de contenido multimedia, las videollamadas y los juegos en línea son ya cotidianos una Latencia baja es esencial para garantizar una experiencia del usuario fluida e interactiva.

El Jitter, mide la variabilidad en los tiempos de llegada de los paquetes, puede introducir inconsistencias y perturbaciones en la entrega de datos; por ejemplo, en el caso de la descarga de una imagen o video puede hacer que los paquetes lleguen en desorden y tome más tiempo ordenar el buffer del aplicativo. El Jitter se produce durante la transmisión de datos entre dos elementos de la red de conmutación de paquetes. Otro fenómeno percibido es el de tráfico estancado (*Stalled*), el usuario está tratando de obtener datos de media (imagen, *streaming*, etc.) pero los datos no están disponibles para ser recibidos. El Jitter puede afectar negativamente la calidad de la transmisión de voz y video, generando interrupciones y distorsiones. En entornos de misión crítica, como llamadas de emergencia o transmisiones en vivo, la gestión del Jitter se convierte en un factor determinante para la confiabilidad de la red.

La pérdida de paquetes afecta al rendimiento de una red, implica retrasos y/o lentitud en la carga / descarga de datos y afecta directamente en aplicaciones de tiempo real como llamadas de video. La pérdida de paquetes ocasiona que la comunicación se vea interrumpida.

Medir Latencia, Jitter y la Pérdida de paquetes no es simplemente un ejercicio técnico; es un elemento clave para garantizar que las redes móviles cumplan con las expectativas de los usuarios y también respalden la demanda creciente de uso

[27] Instituto Federal de Telecomunicaciones (IFT). Velocímetros Internet. México. Disponible en internet en: [https://www.ift.org.mx/velocimetros-internet]

de datos de aplicaciones y servicios. La importancia de estas métricas radica en su capacidad para cuantificar y mejorar la calidad de servicio en la red de conmutación de paquetes.

1.2 Necesidad de mediciones precisas para evaluar el desempeño

Dada la complejidad de las redes de conmutación de paquetes y los desafíos en el desempeño de red, la necesidad de mediciones precisas es imprescindible. La evaluación del rendimiento de la red, no se limita a la velocidad de conexión o a la velocidad de transferencia de datos en la transmisión o en la recepción, implica medir y analizar indicadores, en la medida que estos representen de forma precisa el estado de la red se pueden establecer acciones predictivas y/o proactivas de mejora en la misma. Con cada nuevo elemento integrado en la red, sea un ruteador o sitio nuevo, se debe considerar integrar los elementos necesarios para el monitoreo de la calidad.

Existen diversos mecanismos para obtener las mediciones de Latencia, Jitter y Pérdida de paquetes, una de ellas es mediante la implementación de herramientas que soporten el protocolo de Medición Activa Bidireccional (TWAMP[28]), esto permitirá realizar mediciones bidireccionales entre dos puntos determinados de la red. La definición de este protocolo fue responsabilidad del Grupo de Trabajo de Ingeniería de Internet (IETF[29]) detalla las especificaciones necesarias para proporcionar mediciones de extremo a extremo en redes de conmutación de paquetes. El protocolo TWAMP no solo ofrece una metodología estandarizada, sino que también se adapta a las características específicas de las redes 4G, 5G y posteriores, brindando una visión completa y detallada del desempeño de la red.

1.3 Latencia, Jitter y Pérdida de paquetes: Indicadores clave de desempeño

[28] TWAMP por sus siglas en inglés: *Two-Way Active Measurement Protocol*.
[29] IETF por sus siglas en inglés: *Internet Engineering Task Force*

Un indicador clave de rendimiento (KPIS), se refiere a los valores de rendimiento obtenidos de los diferentes componentes de la red. En el caso específico de Latencia, Jitter y Pérdida de paquetes, se analizan para identificar si el servicio brindado está dentro de los parámetros de calidad o no. El entendimiento de estos indicadores claves ayuda en la optimización de las redes de conmutación de paquetes. Cada proveedor de servicio define umbrales propios de evaluación de los indicadores con miras al cumplimiento interno y garantizar que se cumpla con los requerimientos solicitados por los entes regulatorios.

1.3.1 Latencia, definición y relevancia en redes de conmutación de paquetes

Como se ha explicado previamente, Latencia se refiere al tiempo que tarda un paquete de datos en viajar desde el punto de origen hasta su destino sea cualquiera de los diferentes dominios de la red, dentro del ámbito de este trabajo ese origen y destino se encuentran en el dominio de transporte y/o transmisión. En este dominio el tiempo de transmisión puede dividirse en diferentes componentes, como Latencia de transmisión (tiempo que lleva enviar un paquete), Latencia de procesamiento (tiempo que tarda un dispositivo en procesar el paquete) y Latencia de propagación (tiempo que tarda el paquete en viajar a través de la red).

La relevancia de este indicador de desempeño estriba en su impacto en aplicaciones y servicios. Por ejemplo, el impacto de una Latencia elevada en transmisión de Voz sobre IP (*VoIP*), implica por ejemplo que los usuarios escuchen eco por lo que se afecta la calidad de la conversación y la naturalidad de la interacción; otra condición que pueden percibir los usuarios es el audio entrecortado lo que dificulta el entendimiento de la conversación. En aplicaciones que requieren comunicación cercana a tiempo real como los juegos de video, en particular cuando se tienen múltiples jugadores, un valor alto de Latencia implica que un jugador perciba algo diferente del resto y por lo tanto se pierda la cadencia de la interacción. Un ejemplo adicional son las aplicaciones de videoconferencia o transmisión de contenido multimedia, en donde una Latencia alta impacta en la sincronización y la calidad de la reproducción lo cual puede derivar en canales de comunicación con

pérdida de audio o ruidosos inhibiendo la comprensión o elevando el tiempo en la comunicación. Existen otros casos en donde se hace referencia a mantener el valor de Latencia en su mínima expresión:

> Un ejemplo concreto de esta relevancia se aprecia en las operaciones quirúrgicas con 5G, donde una latencia alta puede resultar fatal. Otros usos médicos que pueden verse beneficiados por esta tecnología son la rehabilitación en grupo para pacientes con esclerosis múltiple (combinada con la realidad virtual) o musicoterapia inmersiva con 5G y realidad virtual para tratar enfermedades neurodegenerativas.[30]

Latencia como indicador de desempeño debe tener umbrales establecidos para los diferentes tipos de tecnologías, en una publicación en el sitio de Telefónica se hace referencia estos valores:

> Con el 4G, la latencia es de 200 milisegundos, mientras que con el 5G se puede llegar a reducir drásticamente, hasta un milisegundo. Unas diferencias enormemente significativas si la comparación la llevamos a cabo con el 2G de finales del siglo pasado, donde la latencia podía oscilar entre los 500 milisegundos y un segundo.[31]

El umbral para 5G de 1 milisegundo es un reto ambicioso sobre todo en infraestructura de otras tecnologías de red. Hay un factor adicional en el estudio que se realiza, y es que en redes móviles 4G y 5G, donde la movilidad de los usuarios es una característica distintiva, este indicador es también crítico, pues los usuarios constantemente se mueven dentro de la red.

Cada operador de servicio busca implementar acciones que ayuden a reducir Latencia y minimizar el impacto al servicio de los usuarios finales. El análisis de este indicador permite tomar acciones preventivas y correctivas por parte del operador de la red.

1.3.2 Jitter como factor en la calidad de servicio

[30] Telefónica. Qué es la latencia y por qué es tan importante que sea baja. Grupo Telefónica España. Disponible en internet en: [https://www.telefonica.com/es/sala-comunicacion/blog/que-es-latencia/]
[31] *Ibidem*. Telefónica.

El Jitter representa la variabilidad en los tiempos de llegada de los paquetes de datos a su destino. Mientras que la Latencia mide el tiempo promedio de viaje, el Jitter se encarga de las variaciones en esos tiempos. Valores elevados de Jitter derivan en problemas de calidad de servicio en aplicaciones sensibles a la sincronización, como la transmisión de voz y video en tiempo real. En redes de conmutación de paquetes, el Jitter puede ser originado por diversos factores, por mencionar algunos: La congestión de la red de datos, los cambios en el volumen de tráfico cursado y las interferencias en los medios de transporte.

Cuando el Jitter es grande en magnitud, la calidad del servicio se ve comprometida. En una llamada de Voz sobre IP (*VoIP*), un valor alto de Jitter, puede resultar en interrupciones, pérdida de paquetes y una experiencia de usuario insatisfactoria. Es por tal motivo que la disminución de la variación en los tiempos de llegada de paquetes en una red se convierte en un desafío primario para los operadores de red que buscan ofrecer servicios de alta calidad en un entorno dinámico y diverso.

Para entender la importancia del Jitter, se asume que un usuario está descargando un video de internet, supongamos que la red de conmutación de paquetes presenta congestión, debido a la cual algunos paquetes se retrasan para llegar al usuario final. El retraso entre los primeros tres paquetes (1, 2 y 3) es de 30 milisegundos (ms), el retraso entre los tres paquetes siguientes (4, 5 y 6) es de 40 ms, el retraso de los paquetes (7, 8 y 9) es de 10 ms y así sucesivamente; el receptor de dichos paquetes, en este caso la terminal móvil, debe encargarse de procesar dicha variación asegurando que se pueda aplicar alguna técnica para procesamiento de Jitter.

El Jitter puede ser medido de diferentes formas: Considerando Jitter Pico (Valor máximo), Jitter promedio: el promedio del valor del arribo de los paquetes. Si al medio de transmisión se le aplican reglas de Calidad de Servicio (QoS) puede ayudar a priorizar los paquetes de mayor importancia para aplicaciones sensibles a la demora y esta medida ayudaría a reducir la variación.

1.3.4 La Pérdida de paquetes

Moon (Moon, 2000) establece que la Pérdida de paquetes es un indicador de congestión de la red de conmutación de paquetes. Con este supuesto, cuando un nodo dentro de la red recibe más paquetes de los que puede procesar (o almacenar en buffer), el nodo empieza a descartar paquetes de acuerdo con su política de encolamiento. Un nodo se entera de la congestión cuando detecta una pérdida de paquete. Cabe mencionar que se puede presentar también la Pérdida de paquetes cuando existe algún problema en el medio físico, por ejemplo, un corte de fibra o caída de un enlace.

Es importante destacar que Latencia, Jitter y la Pérdida de paquetes son indicadores que también son considerados por los entes reguladores de la industria de las telecomunicaciones, un ejemplo sería el Instituto Federal de Telecomunicaciones que mantiene los estándares de Calidad en el Servicio Móvil.[32]

1.4 Protocolo TWAMP

El Grupo de Trabajo de Ingeniería de Internet (*IETF*) describe en el documento RFC 5357[33] el protocolo de medición activa bidireccional (*Two Way Active Measurement Protocol*), este protocolo diseñado específicamente para evaluar el rendimiento de las redes basadas en conmutación de paquetes, proporciona una forma de medir bidireccionalmente, tanto la transmisión como la recepción mediante el uso de marcas de tiempo, independientes de la sincronización de tiempo del servidor host para establecer contrastes o identificar donde se encuentran los retrasos, demora o pérdida de paquete entre los puntos de medición.

El RFC 5357 detalla el funcionamiento protocolo TWAMP, que requiere interfaces de red en dos nodos cada nodo con una función definida. Se definen dos modos de operación:

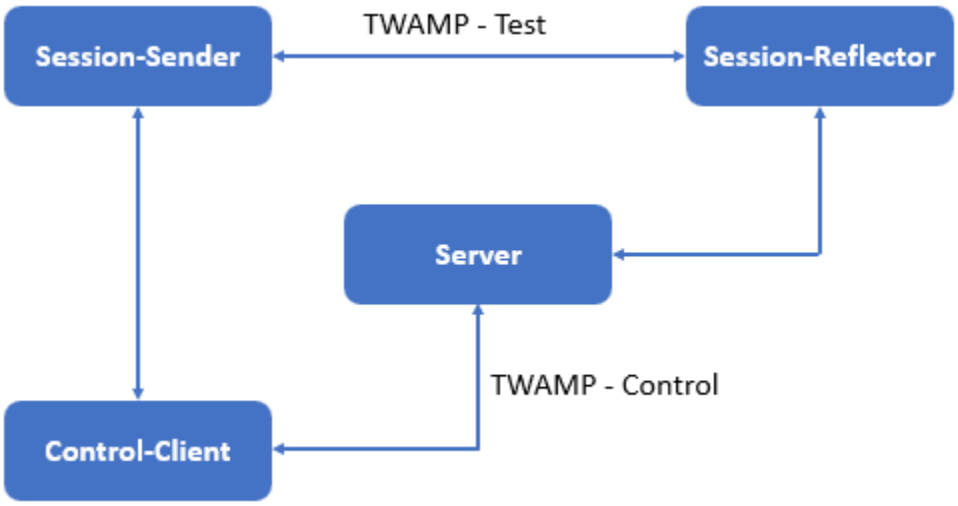
[32] Instituto Federal de Telecomunicaciones. Calidad en el Servicio Móvil. Esta disposición fue publicada en el Diario Oficial de la Federación el 17 de enero de 2018. Disponible en internet en: [https://www.ift.org.mx/usuarios-y-audiencias/calidad-en-el-servicio-movil]

[33] Internet Engineering Task Force. Request for Comments: 5357. A Two-Way Active Measurement Protocol (TWAMP). https://www.rfc-editor.org/rfc/rfc5357.txt

- **TWAMP-Control:** Permite iniciar y finalizar sesiones de prueba. El modo de operación TWAMP-Control se ejecuta entre un elemento Control-Cliente y un elemento Servidor.
- **TWAMP-Test:** Intercambia paquetes de prueba entre dos elementos TWAMP. El modo de operación TWAMP-Test se ejecuta entre un elemento Session-Sender y un elemento Session-Reflector.

La Figura 2 muestra los cuatro componentes del protocolo TWAMP; cada uno de estos componentes se emplea en los modos de operación descritos arriba. TWAMP permite flexibilidad en la implementación, por ejemplo, una opción combina las funciones de Control-Client y Session-Sender en un dispositivo (Conocido como controlador TWAMP o cliente TWAMP) y las funciones de Servidor y Reflector de sesión en el otro dispositivo (Conocido como servidor TWAMP). En este caso, cada dispositivo ejecuta los protocolos TWAMP-Control (Entre Control-Client y Server) y TWAMP-Test (Entre Session-Sender y Session-Reflector).

Figura 2. *Componentes TWAMP*[34]



Nota. Elaboración propia con base en el RFC 5357 (*A Two-Way Active Measurement Protocol - TWAMP*) de IETF.[35]

TWAMP se basa en la idea de la reflexión simétrica, lo que significa que el flujo de tráfico utilizado para la medición es el mismo que el flujo de tráfico que se medirá.

[34] *Ibidem.* IETF RFC 5357
[35] *Ibidem.* IETF RFC 5357

Esta simetría garantiza mediciones precisas, proporcionando una visión holística del desempeño de la red en condiciones reales de operación. El protocolo también establece un método estructurado para enviar paquetes de prueba entre un elemento denominado origen y un elemento denominado destino dentro de una red, registrando y analizando las métricas de rendimiento asociadas con estos paquetes. El protocolo contribuye a la optimización de los parámetros clave que afectan la experiencia del usuario. La calidad del servicio se traduce en la capacidad de proporcionar conexiones rápidas, confiables y consistentes.

La arquitectura cliente-servidor del protocolo TWAMP, tiene los siguientes componentes:[36]

Cliente de TWAMP

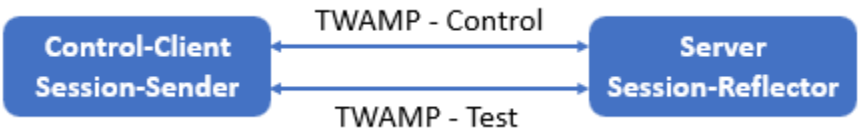
- Control-Client configura, establece y detiene las sesiones de prueba de TWAMP.
- Session-Sender crea paquetes de prueba de TWAMP que se envían al reflector de sesión en el servidor de TWAMP.

Servidor TWAMP

- El reflector de sesión devuelve un paquete de medición cuando se recibe un paquete de prueba, pero no mantiene un registro de dicha información.
- El servidor administra una o más sesiones con el cliente TWAMP y escucha los mensajes de control en un puerto TCP.

La Figura 3 muestra la arquitectura cliente-servidor

Figura 3. *Componentes TWAMP Cliente / Servidor*



Nota. Elaboración propia con base en el RFC 5357 (*A Two-Way Active Measurement Protocol - TWAMP*) de IETF.[37]

[36] *Ibidem.* IETF RFC 5357
[37] *Ibidem.* IETF RFC 5357

1.4.1 Funcionamiento del protocolo TWAMP

En la sección previa se mencionó que TWAMP se basa en la noción de reflexión simétrica. El protocolo utiliza la misma ruta para la transmisión de los paquetes de prueba que para la recepción de estos paquetes. Este enfoque simétrico garantiza mediciones precisas, reflejando el rendimiento de extremo a extremo de la red.

Como se mostró en las Figuras 1 y 2, el funcionamiento inicia con el origen enviando paquetes de prueba al destino. Estos paquetes atraviesan la red de conmutación de paquete y son reflejados de vuelta al origen por el destino. La simetría en la transmisión y reflexión de los paquetes permite medir de forma bidireccional la Latencia, Jitter, Pérdida de paquetes, así como otros parámetros. Además, la estructura simétrica facilita la identificación de cualquier asimetría en la red, lo que permite identificar fallas o desviaciones en tendencia de tráfico.

Un componente clave en el funcionamiento de TWAMP es la capacidad de generar tráfico de prueba sin interrumpir el tráfico operativo normal. Esta característica es esencial para realizar mediciones en entornos de producción sin afectar el transporte de paquetes de tráfico de usuarios. La capacidad de realizar pruebas sin perturbaciones contribuye a la viabilidad práctica de TWAMP como una herramienta de evaluación continua del desempeño.

TWAMP demuestra también una adaptabilidad importante a entornos de conmutación de paquetes, siendo su implementación eficaz tanto en redes fijas como móviles. La flexibilidad para adaptarse a diferentes entornos se deriva de su enfoque basado en estándares y su capacidad para funcionar en conjunción con protocolos de red comunes.

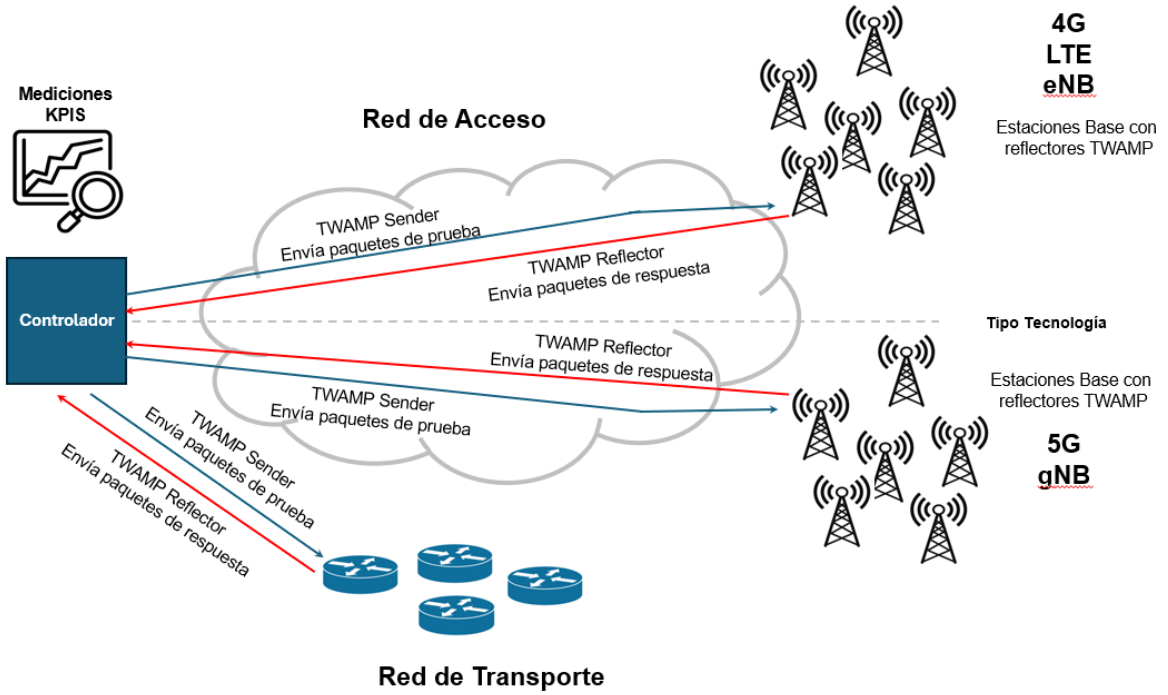
En el caso específico de redes móviles 4G y 5G, TWAMP demuestra su utilidad al proporcionar mediciones precisas en condiciones dinámicas. La movilidad constante de los usuarios y la variabilidad en la carga de tráfico no son obstáculos para TWAMP; por el contrario, este protocolo se adapta a estas condiciones, brindando mediciones relevantes incluso en escenarios de variación continua.

La adaptabilidad de TWAMP a redes de conmutación de paquetes también se manifiesta en su capacidad para operar en entornos heterogéneos. En redes 4G y 5G, donde coexisten diversos tipos de dispositivos y servicios, TWAMP se ajusta a la diversidad de la infraestructura, asegurando las mediciones independientemente de la complejidad de la red. Mediante la implementación de este protocolo, los propios elementos de red ya no necesitan generar ni mantener estadísticas sobre el rendimiento de la red de transporte. Algunas de las implementaciones consideran un sistema de gestión del rendimiento para procesar solo los clientes TWAMP que inician solicitudes de recopilación de estadísticas y obtener fácilmente estadísticas sobre toda la red. De esta forma, las estadísticas de rendimiento de transporte se recopilan de forma rápida y flexible.

TWAMP utiliza paquetes UDP como sondas para recopilar estadísticas sobre la Latencia de ida y vuelta, el Jitter y la tasa de Pérdida de paquetes. Además, separa el control de sesión y la medición del tráfico para brindar alta seguridad. Los dispositivos de red compatibles con este protocolo cooperan entre sí para obtener estadísticas sobre el rendimiento de la red de transporte.

El número de variables obtenidas en las mediciones TWAMP puede ayudar a identificar desviaciones en tendencia de tráfico, como congestión de la red y/o algún evento que pueda causar Latencia o Pérdida de paquetes. Se busca que el análisis descrito en este trabajo pueda ser empleado para otras ubicaciones geográficas, así como otras tecnologías de red de conmutación de paquetes como 5G y 6G. La Figura 4 muestra la perspectiva de medición TWAMP que es considerada para este trabajo.

Figura 4. *Sesiones TWAMP establecidas en una red de conmutación de paquetes (4G / 5G)*



Nota. El gráfico muestra la implementación de protocolo TWAMP tomando en consideración dos subdominios de una red de conmutación de paquetes móvil para las tecnologías 4G-LTE y 5G. El diagrama puede ser extendido también al subdominio de Núcleo de Red.

La entidad eNB (eNodeB o eNB representan el mismo término) gestiona la comunicación de interfaz de la red de acceso de radio con usuarios de la red de cuarta generación. Cada eNB controla una o más celdas, que son áreas geográficas de cobertura por radio. Es decir, los usuarios establecen una conexión hacia la red de acceso, desde el dispositivo móvil hacia los eNB o sitios celulares. Dicho tráfico se transforma para ser enviado por la red de transporte.

En cada sitio celular, se establece una conexión de datos entre diferentes elementos de red y nodos agregadores mediante el empleo del protocolo IP. Existen también nodos destino en el núcleo de red móvil como el MME, elemento de procesamiento de tráfico de plano de control en una red LTE de acuerdo con 3GPP, gestiona el acceso de los usuarios a través de la red de Acceso de 4G (*E-UTRAN*).

El Gateway de servicio o *SGW*, es un elemento de la red de 4G, donde se monitorizan las políticas de conexión de los usuarios hacia la red móvil. Para interconectar ambos extremos se ocupan enlaces de datos, por lo que permite la implementación de monitoreo vía TWAMP entre la red de acceso y el núcleo de red.

Capítulo 2
Metodología de análisis de datos

Capítulo 2. Metodología de análisis de datos

Este trabajo no tiene por objeto explicar cada una de las fases de la metodología CRISP-DM, sino que se centra en su uso para el análisis de datos de las mediciones TWAMP en una red de conmutación de paquetes. CRISP-DM[38] se utilizó inicialmente en proyectos de minería de datos, dada su aplicabilidad, así como la capacidad de interacción entre las diferentes etapas, permitió que en la actualidad se utilice para fundamentar los pasos a seguir en proyectos ahora de ciencia de datos. Saltz (Saltz, 2021) establece por el contrario considera que la metodología CRISP-DM omite aspectos clave del ciclo de vida de un proyecto de ciencia de datos, incluido la forma en que un equipo debe priorizar las tareas, colaborar entre los integrantes del equipo y comunicarse. En el trabajo realizado para este proyecto la metodología CRISP-DM se adecua a las necesidades esperadas de negocio y para un análisis general de la red, se da prioridad a los mercados donde hay más usuarios y/o más tráfico y es ahí donde el supuesto mencionado de Saltz hace mayor sentido.

2.1 Minería de datos en redes de conmutación de paquetes

Una definición de la minería de datos establece que el proceso implica la extracción de información implícita, previamente desconocida y potencialmente útil de los datos. La minería de datos se relaciona con las técnicas y herramientas utilizadas para extraer información útil de grandes volúmenes de datos.[39] El uso de la minería de datos permite transformar la gestión de redes de conmutación de paquetes pasando de una visión reactiva a una visión predictiva y proactiva. En este trabajo se consideran las variables de las mediciones TWAMP, sin embargo, puede ser extensible los procesos de minería de datos a otros dominios de la red.

[38] Data Science PM, *What is CRISP-DM*?, Disponible en internet en: [https://www.datascience-pm.com/crisp-dm-2], establece que CRISP-DM se publicó en 1999.
[39] Universidad Nacional Autónoma de México, Introducción a la minería de datos, Dirección General de Cómputo y Tecnologías de Información y Comunicación. Disponible en: [https://docencia.tic.unam.mx/presenciales/introducción-a-la-mineria-de-datos.html]

En Baldi *et. al.* (Baldi *et. al*, 2010), se establece que uno de los aspectos más críticos en mantener una red (conmutación de paquetes) bajo control es capturar y analizar su tráfico. La complejidad de estas tareas se incrementa debido a que las arquitecturas de red se vuelven más rápidas. Este supuesto puede extenderse a la red de conmutación de paquetes, sin embargo, tenemos una variante en este caso estamos ocupando protocolo TWAMP para realizar la medición de la red.

En Kiang Sing y Huan (Kian Sing y Huan, 2001), se establece el uso potencial de la Minería de Datos en las empresas, como la identificación de nuevas oportunidades de negocio, adaptar los productos ofrecidos o encontrar los clientes más valiosos con el fin de retenerlos, y de esta manera aumentar los ingresos y reducir las pérdidas o costos. Al determinar las características de los buenos clientes, las empresas pueden enfocarse en aquellos de características similares y diseñar productos o servicios acordes a sus necesidades.

Si bien este trabajo se enfoca principalmente en el estudio de mediciones de calidad de la red si podemos generalizar la propuesta anterior de que se busca una nueva oportunidad de negocio, al buscar mejorar la calidad de la experiencia al usuario final. El proceso de minería de datos se ha caracterizado en varias metodologías de los cuales la más extendida es CRISP-DM[40] y su uso se ha propagado en diferentes tipos de proyectos incluidos los de ciencia de datos.

2.2 Metodología CRISP-DM en un proyecto de Ciencia de Datos en una red de conmutación de paquetes.

IBM establece que CRISP-DM es una forma comprobada en la industria para guiar los esfuerzos de minería de datos[41]:

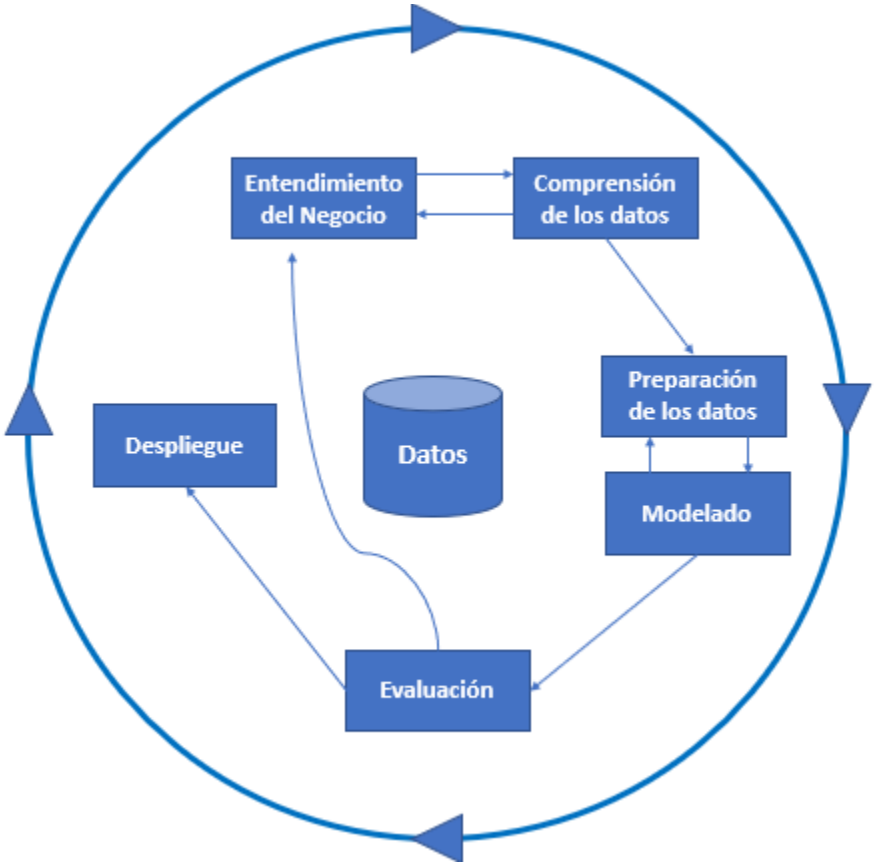
[40] IBM, *CRISP-DM Help Overview*. Disponible en internet en: [https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview].
[41] *Ibidem*. CRISP-DM

- Como metodología, incluye descripciones de las fases típicas de un proyecto, las tareas involucradas en cada fase y una explicación de las relaciones entre estas tareas.[42]
- Tomando en consideración el ámbito en el cual se desarrolló la metodología, esta describe de alto nivel el ciclo de vida de un proyecto de minería de datos.[43]

La implementación de CRISP-DM en un proyecto de ciencia de datos es relevante ya que proporciona al analista de datos / científico de datos una estrategia (mayormente de datos) y un marco metodológico en el cual se puede apoyar. El científico de datos tiene una metodología flexible con múltiples interacciones. La Figura 5 muestra el ciclo de vida de la metodología CRISP-DM, como se puede apreciar es un ciclo iterativo, en cada iteración se profundiza más en el negocio. La metodología es flexible en cuanto a la interacción entre las diferentes etapas, lo cual permite flujos no secuenciales permitiendo regresar y realizar ajustes en las fases previas. Las flechas indican las dependencias más importantes y frecuentes.

Figura 5. *Metodología CRISP-DM*[44]

[42] *Ibidem.* CRISP-DM
[43] *Ibidem.* CRISP-DM
[44] *Ibidem.* CRISP-DM



Nota. La figura representa las diferentes etapas de la metodología, se muestra también la interacción que tienen las diferentes fases entre sí. Elaboración propia con base en las fases CRISP-DM.[45]

2.3 Etapas de la metodología

Se consideran 6 etapas dentro de la metodología, las cuales se enumeran a continuación:

1. Entendimiento del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

[45] *Ibidem.* CRISP-DM

En las siguientes secciones se explicará las acciones realizadas para este proyecto en cada una de las etapas de la metodología.

2.3.1 Entendimiento del negocio

En esta etapa se responden preguntas como: ¿Por qué es necesario utilizar herramientas de minería de datos o ciencia de datos para este tipo de proyectos? ¿Cuál es el objetivo de negocio que pretendemos alcanzar mediante el análisis TWAMP?; ¿Cuál es la situación actual sin este proyecto?; ¿Cuáles son los objetivos del proyecto se alinean a los objetivos de negocio? ¿Por último cual sería el plan de proyecto?

Debido a las dimensiones del conjunto de datos no es factible analizar la información de forma manual, por lo que un proceso de este estilo no es escalable para analizar información de toda la red. El objetivo final es accionar sobre la red, ya sea ampliando capacidad o mejorar calidad de servicio, el proyecto de ciencia de datos es el punto de partida. Una vez seleccionado el proyecto es necesario establecer y detallar las expectativas de negocio de lo contrario la delimitación del problema puede ser ambigua, y los resultados que se pudieran obtener no necesariamente corresponden a los esperados por el negocio.

En este mismo orden de ideas, la ambigüedad en la delimitación del problema y/o del proyecto, implica que no se podría establecer cuando se ha alcanzado el objetivo del proyecto y concluido el trabajo dentro del mismo; por tanto, el éxito o fracaso de la iniciativa propuesta.

CRISP-DM sugiere que antes de iniciar con un proyecto de ciencia de datos, es necesario dedicar tiempo para explorar las expectativas que se tienen del mismo por parte de la organización.[46] Alinear las expectativas del proyecto con los objetivos y requerimientos del negocio, para después convertir este conocimiento en una definición del problema de ciencia de datos. En esta etapa también se plantearon preguntas específicas para delimitar el alcance del proyecto, normalmente se tiene

[46] *Ibidem.* CRISP-DM

una divergencia en entendimiento de las áreas de negocio y las áreas técnicas. Algunas de las preguntas que se ocuparon para delimitar el alcance del proyecto se listan a continuación:

1. ¿Existe algún proceso o procedimiento automatizado para analizar Latencia y Jitter en toda la red, es decir a nivel nacional?
2. ¿Esta solución se basa en estándares de industria de telecomunicaciones que ayuden a mejorar la calidad de servicio y por tanto experiencia de los suscriptores?
3. ¿Con que información y/o datos contamos y tenemos acceso en este momento?
4. ¿Qué datos y/o variables se obtienen de las observaciones de las mediciones de Latencia y Jitter dentro de la red?
5. ¿Todas las variables procesadas son realmente importantes para identificar problemas de Latencia y Jitter?
6. ¿Se puede aplicar alguna técnica de aprendizaje de máquina (*Machine Learning*) para obtener información de los datos y ocuparlo para predecir fallas en la red?

Este trabajo considera que, para realizar un proyecto de ciencia de datos no es necesario contar con un gran volumen de datos que describa toda la red, por el contrario, lo que se pretende es tomar una porción de la red y segmentar el problema para posteriormente generalizar el estudio a toda la red. Esta propuesta también se basa en obtener una eficiencia operativa en términos de costos, bajo el supuesto no desplegar soluciones de infraestructura grandes y de inicio tener costos.

Con esta premisa se seleccionó el mercado (delimitación geográfica) de Ciudad Juárez, Chihuahua. Este mercado cuenta con 149 sitios celulares; esta información es relevante pues nos ayuda a entender la volumetría de eventos generados para monitoreo de la calidad de servicio de la red. Con el objetivo de identificar el área geográfica de estudio, la Figura 6 muestra el mercado de Ciudad Juárez, como se puede observar está delimitado por un polígono punteado que corresponde mayormente a la zona urbana de la ciudad y no se consideran todos los tramos

carreteros donde existen sitios celulares que brindan servicio. En el poniente de la ciudad se encuentra la Sierra de Ciudad Juárez, con una orografía irregular que puede afectar los enlaces de transmisión mientras que el oriente la ciudad es delimitada por la frontera con Estados Unidos.

Figura 6. *Mapa de Ciudad Juárez, Chihuahua, México*

Nota. Se seleccionó el mapa de Ciudad Juárez del Directorio Estadístico Nacional de Unidades Económicas (DENUE) del Instituto Nacional de Estadística y Geografía (INEGI).[47]

2.3.2 Comprensión de los datos

En esta etapa una vez que se verificaron las muestras de protocolo TWAMP para los sitios celulares en este mercado, se realizaron muestras de información para entender y comprender sí se contaba con los insumos para él proyecto. Se tomaron muestras en diferentes horarios del día, en hora pico (intervalo de tiempo donde se presenta un incremento de tráfico en la red), en ocasiones este incremento también incluye un incremento en la movilidad de los suscriptores. Con las muestras

[47] Directorio Estadístico Nacional de Unidades Económicas (DENUE), Instituto Nacional de Estadística y Geografía. La selección del mapa es de forma manual. Disponible en Internet en: [https://www.inegi.org.mx/app/mapa/denue/Default.aspx?idee=6293804]

obtenidas se realizó una descripción y exploración de los datos. Se identificó también que en la red de conmutación de paquetes se tienen diferentes valores de DSCP (*Differentiated Services Code Point*), este último, es un campo dentro del encabezado de los paquetes de protocolo de internet (IP) que se utiliza para identificar y clasificar diferentes tipos de tráfico en una red. Se utiliza para implementar calidad de servicio (*QoS*) y priorización de tráfico en la red. Para el tráfico que me propuse revisar considere DSCP 46 para servicios de telefonía / Voz sobre LTE (*VoLTE*), es decir sobre una red de cuarta generación.

Referente a la colección de datos se identificaron los siguientes supuestos:

- Para cada sitio se toma una medición TWAMP para *DSCP* 46 por minuto.
- Es decir que para cada sitio hay 60 mediciones por hora.
- Y en un periodo de 24 horas se genera un total de 60 x 24 = 1,440 mediciones por día por sitio.
- Con el supuesto que tenemos un total de 149 sitios celulares, el total de mediciones TWAMP para *DSCP* 46 es de (1,440 x 149) = 214,560 por día.
- El periodo de análisis considera 31 días por lo tanto el conjunto de datos debería tener un total de 214,560 x 31 x 2 (2 *DSCP* – 40 / 46) = 13,302,720 registros. En el total de mediciones TWAMP del conjunto de datos se identificaron 16,260,080, es decir 2,957,360 mediciones más.
- Se identificó que las muestras estaban repartidas en dos sitios concentradores (Buenos Aires, BNAS) con 10,074,182; de las cuales 5,037,090 corresponden a DSCP 40 y 5,037,092 a DSCP 46. El segundo concentrador corresponde a Apodaca con 3,092,949 muestras que corresponden a DSCP 40 y 3,092,949 que corresponden a DSCP 46. Estas validaciones son importantes pues ayudan a cerrar la fase de preparación de los datos.

El último paso realizado en esta etapa fue la verificación de la calidad de los datos, entender los valores nulos y/o con un valor en magnitud alto. Con estos pasos completados se estableció considerar información de un mes para entender el

comportamiento de la red, por lo que las mediciones TWAMP corresponde al periodo que va del primero de agosto de 2022 al 31 de agosto de 2022.

Cada registro o medición está conformada por un total de setenta variables con la siguiente estructura y significado:

Variable	Descripción
SESSION_NAME	Identificador de la sesión TWAMP definida para la sesión entre los elementos de transporte
SOURCE_NAME	Nombre del elemento de red de esta sesión de TWAMP
DATETIME	Fecha y hora en la que se registró la medición
UL_MISORDERPKTS	Número de paquetes en sentido origen a destino que se recibieron reordenados (excluidos los duplicados) durante el intervalo de la medición
UL_DUPLICATEPKTS	Número de paquetes en sentido origen a destino que se recibieron duplicados durante el intervalo de la medición
UL_TOOLATEPKTS	Número de paquetes en sentido origen a destino que se recibieron tarde y que ya se contaron como perdidos en el intervalo anterior
UL_LOSTPKTS	Número de paquetes en sentido origen a destino perdidos durante el intervalo de la medición
UL_LOSTBURSTMIN	Duración del período de pérdida más corto en sentido origen a destino durante el intervalo de la medición
UL_LOSTBURSTMAX	Duración del período de pérdida más largo en sentido origen a destino durante el intervalo de la medición
UL_LOSTPERC	Tasa de número de paquetes perdido en sentido origen a destino durante el intervalo de medición
UL_MOS	MOS *(Puntuación media de opinión)* de carga (*Uplink*). Es una medida de calidad de la experiencia. Esta medida se expresa en valores que oscilan de 1.0 a 5.0, donde 5.0 es la calidad más alto.
UL_R	El valor de clasificación en sentido origen a destino (Carga o *Uplink*) (o factor R de carga / *Uplink*) mide la calidad que percibe un suscriptor en una la llamada de voz y se deriva de métricas como el tipo de códec, la latencia, la inestabilidad y la pérdida de paquetes. El valor R se expresa como un valor decimal en el rango de 1.0 a 100.0, donde 100.0 es la calidad más valorada.
UL_TOSMIN	Valor de TOS mínimo recibido en sentido origen a destino durante el intervalo. Valor máximo 255
UL_TOSMAX	Valor de TOS máximo recibido en sentido origen a destino durante el intervalo. Valor máximo 255

UL_TTLMIN	Valor TTL (Tiempo de Vida) mínimo recibido en sentido origen a destino durante el intervalo. Valor máximo 255
UL_TTLMAX	Valor TTL máximo recibido en sentido origen a destino durante el intervalo. Valor máximo 255
UL_DMIN	Demora mínima en sentido origen a destino durante el intervalo, medida en microsegundos
UL_DP95	Percentil 95 de la demora en sentido origen a destino en microsegundos
UL_DPLO	Demora configurada por el usuario en sentido origen a destino (percentil predeterminado 96) durante el intervalo, medida en microsegundos.
UL_DPMI	Demora configurada por el usuario en sentido origen a destino (percentil predeterminado 98) durante el intervalo, medida en microsegundos.
UL_DPHI	Demora configurada por el usuario de origen a destino (percentil predeterminado 99) durante el intervalo, medida en microsegundos.
UL_DMAX	Demora máxima en sentido origen a destino en microsegundos durante el intervalo
UL_DMEAN	Media de la demora en sentido origen a destino en microsegundos durante el intervalo
UL_JMIN	Jitter mínimo en sentido origen a destino durante el intervalo, medido en microsegundos
UL_JP95	Percentil 95 de Jitter en sentido origen a destino en microsegundos
UL_JPLO	Jitter configurado por el usuario en sentido origen a destino (percentil predeterminado 96) durante el intervalo, medida en microsegundos.
UL_JPMI	Jitter configurado por el usuario en sentido origen a destino (percentil predeterminado 98) durante el intervalo, medida en microsegundos.
UL_JPHI	Jitter configurado por el usuario en sentido origen a destino (percentil predeterminado 99) durante el intervalo, medida en microsegundos.
UL_JMAX	Jitter máximo en sentido origen a destino en microsegundos durante el intervalo
UL_JMEAN	Media del Jitter en sentido origen a destino en microsegundos durante el intervalo
UL_DVP95	Percentil 95 de la variación en la demora en sentido origen a destino en microsegundos
UL_DVPLO	Variación en la demora configurada por el usuario en sentido origen a destino (percentil predeterminado 96) durante el intervalo, medida en microsegundos
UL_DVPMI	Variación en la demora configurada por el usuario en sentido origen a destino (percentil predeterminado 98) durante el intervalo, medida en microsegundos

UL_DVPHI	Variación en la demora configurada por el usuario en sentido origen a destino (percentil predeterminado 99) durante el intervalo, medida en microsegundos
UL_DVMAX	Variación en la demora máxima en sentido origen a destino en microsegundos durante el intervalo
UL_DVMEAN	Media de la variación de la demora en sentido origen a destino en microsegundos durante el intervalo
DL_MISORDERPKTS	Número de paquetes en sentido destino a origen que se recibieron reordenados (excluidos los duplicados) durante el intervalo
DL_DUPLICATEPKTS	Número de paquetes en sentido destino a origen que se recibieron duplicados durante el intervalo
DL_TOOLATEPKTS	Número de paquetes en sentido destino a origen que se recibieron tarde y que ya se contaron como perdidos en el intervalo anterior
DL_LOSTPKTS	Número de paquetes en sentido destino a origen perdidos durante el intervalo
DL_LOSTPERIODS	Duración del período de pérdida más corto en sentido destino a origen durante el intervalo
DL_LOSTBURSTMIN	Duración del período de pérdida más largo en sentido destino a origen durante el intervalo
DL_LOSTBURSTMAX	Tasa de número de paquetes perdido en sentido destino a origen durante el intervalo de medición
DL_LOSTPERC	Tasa de número de paquetes perdido en sentido destino a origen durante el intervalo de medición
DL_MOS	*MOS* flotante o MOS de descarga (*Downlink*). La puntuación de opinión media es una medida de calidad de la experiencia. Esta medida se expresa en valores que oscilan de 1.0 a 5.0, donde 5.0 es la calidad más alto.
DL_R	El valor de clasificación de destino a origen (Descarga u Down*link*) (o factor R de descarga / *Downlink*) mide la calidad que percibe un suscriptor en una la llamada de voz y se deriva de métricas como el tipo de códec, la latencia, la inestabilidad y la pérdida de paquetes. El valor R se expresa como un valor decimal en el rango de 1.0 a 100.0, donde 100.0 es la calidad más valorada.
DL_TOSMIN	Valor de TOS mínimo recibido en sentido destino a origen durante el intervalo. Valor máximo 255
DL_TOSMAX	Valor de TOS máximo recibido en sentido destino a origen durante el intervalo. Valor máximo 255
DL_TTLMIN	Valor TTL mínimo recibido en sentido destino a origen durante el intervalo. Valor máximo 255
DL_TTLMAX	Valor TTL máximo recibido en sentido destino a origen durante el intervalo. Valor máximo 255

DL_DMIN	Demora mínima en sentido destino a origen durante el intervalo, medida en microsegundos
DL_DP95	Percentil 95 de la demora en sentido destino a origen en microsegundos
DL_DPLO	Demora configurada por el usuario en sentido destino a origen (percentil predeterminado 96) durante el intervalo, medida en microsegundos.
DL_DPMI	Demora configurada por el usuario en sentido destino a origen (percentil predeterminado 98) durante el intervalo, medida en microsegundos.
DL_DPHI	Demora configurada por el usuario en sentido destino a origen (percentil predeterminado 99) durante el intervalo, medida en microsegundos.
DL_DMAX	Demora máxima en sentido destino a origen en microsegundos durante el intervalo
DL_DMEAN	Media de la demora en sentido destino a origen en microsegundos durante el intervalo
DL_JMIN	Jitter mínimo en sentido destino a origen durante el intervalo, medido en microsegundos
DL_JP95	Percentil 95 de Jitter en sentido destino a origen en microsegundos
DL_JPLO	Jitter configurado por el usuario en sentido destino a origen (percentil predeterminado 96) durante el intervalo, medida en microsegundos.
DL_JPMI	Jitter configurado por el usuario en sentido destino a origen (percentil predeterminado 98) durante el intervalo, medida en microsegundos.
DL_JPHI	Jitter configurado por el usuario en sentido destino a origen (percentil predeterminado 99) durante el intervalo, medida en microsegundos.
DL_JMAX	Jitter máximo en sentido destino a origen en microsegundos durante el intervalo
DL_JMEAN	Media del Jitter en sentido destino a origen en microsegundos durante el intervalo
DL_DVP95	Percentil 95 de la variación en la demora en sentido destino a origen en microsegundos
DL_DVPLO	Variación en la demora configurada por el usuario en sentido destino a origen (percentil predeterminado 96) durante el intervalo, medida en microsegundos
DL_DVPMI	Variación en la demora configurada por el usuario en sentido destino a origen (percentil predeterminado 98) durante el intervalo, medida en microsegundos
DL_DVPHI	Variación en la demora configurada por el usuario en sentido destino a origen (percentil predeterminado 99) durante el intervalo, medida en microsegundos

DL_DVMAX	Variación en la demora máxima en sentido destino a origen en microsegundos durante el intervalo
DL_DVMEAN	Media de la variación de la demora en sentido destino a origen en microsegundos durante el intervalo

2.3.3 Preparación de los datos

En esta etapa ya contamos con el conjunto de eventos TWAMP, la herramienta actual genera archivos por mi minuto para cada conexión, por lo que se realizó una unión de todos los archivos en un solo Dataset. En esta etapa también se realizan actividades como limpieza de los datos, verificación de datos nulos o vacíos, verificar si todos las conexiones generaron datos o durante el proceso se identificaron sitios fuera de servicio por un periodo de tiempo durante el mes en este caso no fue necesario generar variables adicionales, sino que se realizó la agregación de datos a nivel hora, con el fin de hacer comparables los resultados con otros dominios de la red (Red Acceso, Red Core o núcleo de red) y también identificar cambios en las tendencias de las variables de Latencia, Jitter y Pérdida de paquetes. También se identificaron y filtraron las mediciones TWAMP no válidas con valores extremadamente altos, la herramienta ocupa valores máximos de tipo de datos para completar los registros (por ejemplo, para algunas de las variables de tipo entero de 32 bits, se incluyen en las muestras valores máximos de 2,147,483,647, lo cual no es un valor correcto en mediciones de red) y por último se identificaron los valores fuera de rango (*Outliers*) de cada una de las variables.

En esta etapa, se definió la regla de transformación de unidades de tiempo, se aplicó la transformación microsegundos a milisegundos ya que las magnitudes de las variables en microsegundos eran grandes y sin los datos normalizados podrían crear un sesgo en el entendimiento de estos. Por último, se realizó transformación de las variables con valores porcentuales ya que estaban multiplicadas por un factor de 1000.

2.3.4 Modelado

El objetivo de esta fase es seleccionar y aplicar las técnicas de modelado acordes a las características del conjunto de datos. En primer lugar, es importante verificar el modelo propuesto es adecuado para el proyecto de ciencia de datos. Consideré dos opciones para la selección del modelo: El primero para la selección del modelo y realizar análisis de datos de las mediciones TWAMP como serie de tiempo para establecer el proceso de análisis y detección de anomalías, considerando las variables de Latencia, Jitter y Pérdida de paquetes; y el segundo un análisis de datos buscando identificar la importancia de las variables del conjunto de datos para para emplear también un modelo de aprendizaje no supervisado y también identificar la importancia, mediante el análisis de componentes principales, de cada variable dentro del conjunto de datos con miras a incorporar el almacenamiento y su uso futuro; se identificó que el conjunto de datos no estaba separado con una etiqueta específica.

Dentro del análisis de datos se validó si seguían una distribución de tráfico específica o tenían un comportamiento asociado a las horas del día cuando más tráfico (hora pico) se procesa en la red de conmutación de paquetes.

2.3.5 Evaluación

En la fase de evaluación es necesario validar que el o los modelos cumplen con el objetivo de negocio para el cual fueron creados así como el cumplimiento de los criterios de éxito definidos en la etapa de entendimiento de negocio. Se documentaron las secuencias de código en Python / R de tal forma que pudiera repetir el análisis, en caso, que se encontrara un error de código o de la aplicación de algún criterio de evaluación, uno de los objetivos principales del proyecto es buscar un modelo que permita la generalización y aplicación de este en toda la red o en subconjuntos de datos asociados a una o varias regiones geográficas.

En esta etapa también se realizó la comparación en las diferentes horas del día para identificar si se representaba correctamente el comportamiento de los enlaces en Ciudad Juárez. En esta fase también se evaluó si el ejercicio de análisis debía o no continuar, puesto que en las iteraciones iniciales el análisis no reflejaba el comportamiento de la red. Por lo que también se realizaron simulaciones de valores

de las variables. Esta fase también ayudó a identificar los umbrales de servicio y/o que debían ser reportados por impacto en la red. Se documentó el impacto de los modelos y las acciones que pueden tomarse para realizar ajustes ya sea en los datos o en el modelo mismo.

Al final de esta etapa también se aplicó la calificación del proceso empleado para el proyecto de ciencia de datos con miras a identificar áreas de oportunidad que puedan mejorarse para análisis futuros.

2.3.6 Despliegue

De acuerdo con la metodología CRISP-DM, la creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo.[48] En esta etapa consistió en implementar la solución en producción para analizar exclusivamente el tráfico de Ciudad Juárez.

Con la experiencia obtenida el despliegue no es una actividad puntual, por el contrario, se debe asegurar el mantenimiento del modelo, por ejemplo, el impacto del modelo si se deciden agregar más sitios en la zona geográfica o si se realiza una sustitución del medio de transporte de utilizar Microondas a utilizar enlaces de Fibra Óptica, la capacidad de la transmisión es diferente. La implementación del modelo se planificó para iniciar operaciones en el primer día del mes hábil y se colectó información durante todo el mes, la construcción de la serie de tiempo detalla la variabilidad del comportamiento de la red. Uno de los puntos identificados es la escalabilidad del entregable, si bien las condiciones de análisis era un marco cerrado la red es un sistema vivo, por lo que analizar la totalidad de las mediciones TWAMP no es apto para un equipo de las condiciones indicadas en el capítulo 3.

[48] *Ibidem.* CRISP-DM

Capítulo 3
Análisis de Datos

Capítulo 3. Análisis de datos

Dibekulu (Dibekulu, 2020) hace referencia al análisis de datos de la siguiente forma:

> La investigación es un campo científico que ayuda a generar nuevos conocimientos y a resolver problemas existentes. Por lo tanto, el análisis de datos es la parte crucial de la investigación, ya que permite obtener resultados más efectivos. Es un proceso de recopilación, transformación, depuración y modelado de datos con el objetivo de descubrir la información necesaria.

En este trabajo considero que el análisis propuesto se adecua a la definición previa, pues se inició con la recopilación de las muestras TWAMP, la transformación de los datos para posteriormente realizar la depuración y modelado de los mismos. La propuesta de este trabajo es la identificación de aquellas variables con mayor relevancia e impacto en el desempeño y calidad de las redes de conmutación de paquetes, y que a la postre inciden en la experiencia del servicio para el usuario final.

El conjunto de datos contiene un total de setenta variables y no contamos con datos etiquetados. La opción es implementar un modelo de análisis no supervisado para analizar y agrupar el conjunto de las mediciones TWAMP. Dado que el dominio de este problema es el de Telecomunicaciones, de acuerdo con la Telefónica, al explicar la definición de aprendizaje no supervisado:

> Para que las máquinas aprendan hay que enseñarlas. Enseñarlas a pensar y a que piensen por si mismas, sin orientación previa. A esto último se refiere el aprendizaje no supervisado: se trata de que la máquina encuentre por si sola patrones en una base de datos sin ninguna guía previa.[49]

El aprendizaje no supervisado utiliza algoritmos para analizar y agrupar conjuntos de datos no etiquetados.[50] Tiene la capacidad de descubrir similitudes y diferencias

[49] Telefónica Móviles España, ¿Qué es el aprendizaje no supervisado?, Blog. Disponible en Internet en: [https://www.telefonica.com/es/sala-comunicación/blog/que-es-aprendizaje-no-supervisado/]
[50] El término venta cruzada o Cross Selling comprende la venta relacionada de productos o servicios complementarios basados en los intereses del cliente en uno de los productos de su empresa, o bien

en la información.[51] La aplicación del análisis no supervisado sobre el conjunto de datos obtenidos de las mediciones TWAMP se considera de la siguiente forma:

- **Reducción de Dimensionalidad con Análisis de Componentes Principales (ACP):** Esto es mediante la normalización de los valores de los datos posteriormente aplicar Análisis de Componentes Principales (ACP) para reducir la dimensionalidad del conjunto de datos y obtener las variables más importantes o de mayor relevancia del conjunto de variables. Al seleccionar los componentes principales, se identifican aquellos componentes que son representados por una mayor magnitud en la varianza, y estos componentes capturan la dirección en las que los datos exhiben mayor variabilidad.

  Adicional al uso de ACP, se consideró *t-SNE* o *t-distributed Stochastic Neighbor Embedding*, desarrollada en 2008 por Laurens Van Der Maaten y Geoffrey Hinton, esta técnica también permite la reducción de dimensiones de conjuntos de datos, también permite visualizar datos de alta dimensión asignando a cada punto de datos una ubicación en un mapa bidimensional o tridimensional. La razón del uso de *t-SNE* para contrastar los resultados obtenidos por ambos algoritmos;

- **Identificación de patrones con agrupamiento (Clúster).** La aplicación de algoritmos de agrupamiento (*Clustering*) se consideró para identificar patrones y agrupaciones en el conjunto de mediciones TWAMP. Estas agrupaciones pueden representar grupos naturales de observaciones con características similares. A través de técnicas de visualización, como la reducción de dimensionalidad mediante *t-SNE* o ACP, podemos representar estas agrupaciones en un espacio de menor dimensión, lo que facilita la interpretación y la identificación de patrones visualmente.

en la compra de uno de estos. Sales Force. ¿Qué es Cross Selling?. Disponible en internet en: [https://www.salesforce.com/es/learning-centre/sales/cross-selling/]

[51] IBM. ¿Qué es el aprendizaje no supervisado? Disponible en Internet en: [https://www.ibm.com/mx-es/topics/unsupervised-learning]

- **Interpretación conjunta de agrupaciones y componentes principales.** La interpretación de los resultados de agrupamiento y PCA proporciona una visión completa de los datos de mayor relevancia. Por ejemplo, se puede visualizar si las agrupaciones identificadas reflejan patrones específicos en los componentes principales. Además, se pueden identificar variables clave que diferencian agrupaciones o que tienen un impacto significativo en la variabilidad general.

- **Correlación de variables.** Moon (Moon, 2000) en el estudio sobre redes de conmutación de paquetes asociado a Voz sobre IP (*VoIP*), estableció la existencia de correlación entre la Latencia y la Pérdida de paquetes. El supuesto inicial sería que las redes de 4G o posteriores al estar basadas en conmutación de paquetes debería observarse un comportamiento similar, la diferencia radica en que Moon hace referencia al protocolo SNMP (*Simple Network Management Protocol*) y en este trabajo se considera el uso de TWAMP.

3.1 Herramientas de trabajo para el análisis de datos

Con el fin de detallar las capacidades del equipo de cómputo empleado para realizar el análisis de las mediciones TWAMP; se contaron con las siguientes características:

Herramienta	**Versión**
Marca	Asus
Modelo	VivoBook X421FAY_X413FA
Sistema Operativo	Windows 11
Almacenamiento	SSD 512 GB
Memoria	8 GB
CPU	Intel Core i5-10210U CPU @ 1.60GHz / 2.11 GHz

Se consideró *Orange Data Miner*[52], por ser un software de código abierto, que se emplea en la minería de datos y permite visualizar, analizar y modelar datos. Por

[52] Orange es una suite de software integral basada en componentes para aprendizaje automático y minería de datos, desarrollada en el Laboratorio de Bioinformática de la Facultad de Informática y

otro lado, el software empleado para la realización del análisis tiene las siguientes características, se utilizó tanto Python como R:

Herramienta	Versión
R	4.3.2
R Studio	2023.09.1 Build 494
Python	3.10.3
Orange Data Miner	3.36.2

3.2 Detección de Anomalías en la red de conmutación de paquetes

Nugroho *et. al* (Nugroho *et. al*, 2023), establecen que la detección de anomalías es una solución para que los operadores celulares superen la dificultad del control de calidad debido a la proliferación del uso de teléfonos móviles. El sistema de monitoreo de redes de telecomunicaciones con detección de anomalías permite la detección inmediata de problemas antes de que se compliquen.

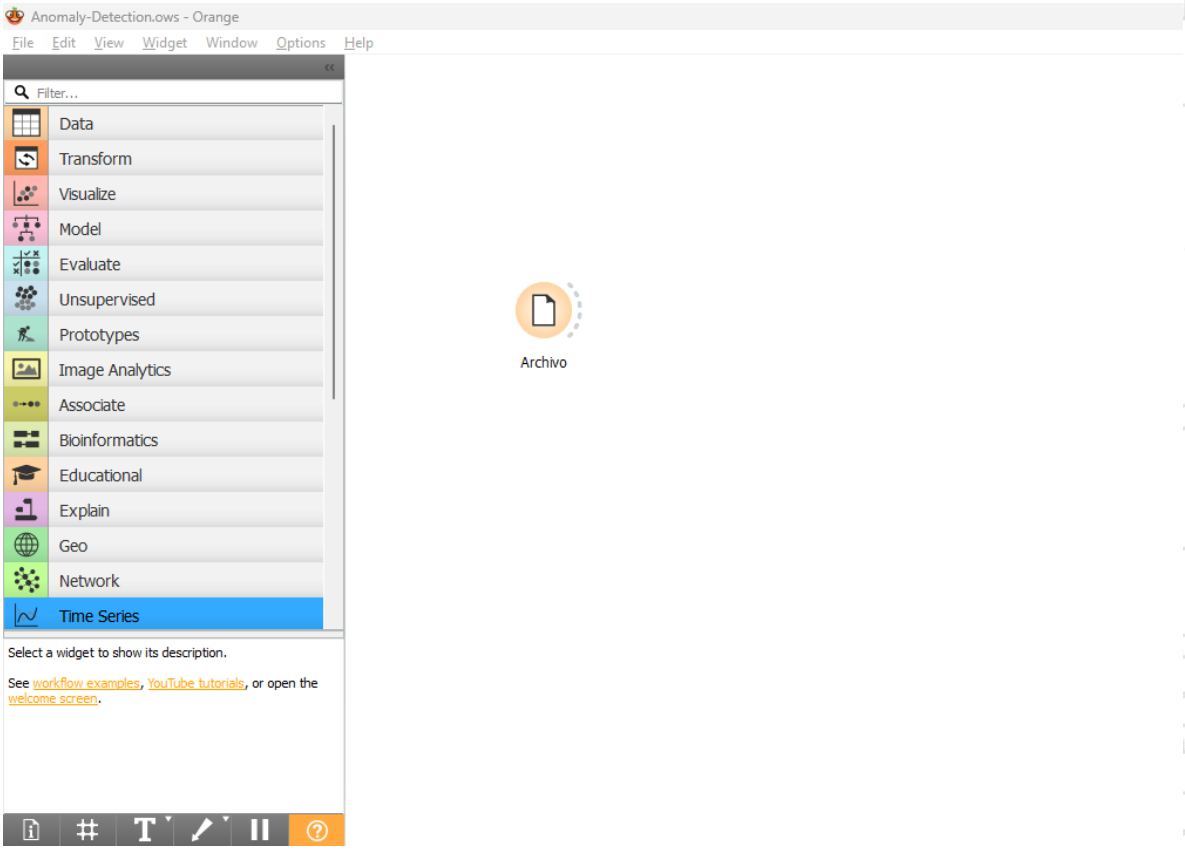
Las mediciones TWAMP se generan en la red de conmutación de paquetes, por lo que la detección inmediata de fallas facilita las acciones reactivas y también evita impactos mayores en el servicio. Un análisis de detección de anomalías sobre los datos de mediciones TWAMP está asociado a la previsión de fallas en los enlaces. Es decir, cada sitio celular requiere de enlaces de transporte de datos para llevar información de cada sitio al resto de puntos remotos de la red. Este análisis tiene como objetivo identificar si existen requerimientos de expansión de capacidad en los enlaces o se debe realizar un mantenimiento preventivo.

En este trabajo se considera el análisis de las variables de Latencia, Jitter y Pérdida de paquetes ya que se puede comprobar la correlación de estas variables y validar el comportamiento de la red, por ejemplo: Si existe latencia alta necesariamente implica la pérdida de paquetes en la red.

Ciencias de la Información de la Universidad de Liubliana (Eslovenia), en colaboración con la comunidad de código abierto. Orange es software libre; puede redistribuirse o modificarse bajo los términos de la Licencia Pública General GNU publicada por la Free Software Foundation; ya sea la versión 3.0 de la Licencia o posterior.

La Figura 7 muestra del lado izquierdo las utilerías disponibles en esta versión de *Orange Data Miner*. Utilerías para carga de datos, visualización, modelos no supervisados, creación de modelos entre otros. Del lado derecho se muestra el punto de partida para el análisis de anomalías.

Figura 7. *Pantalla principal del aplicativo Orange Data Miner*

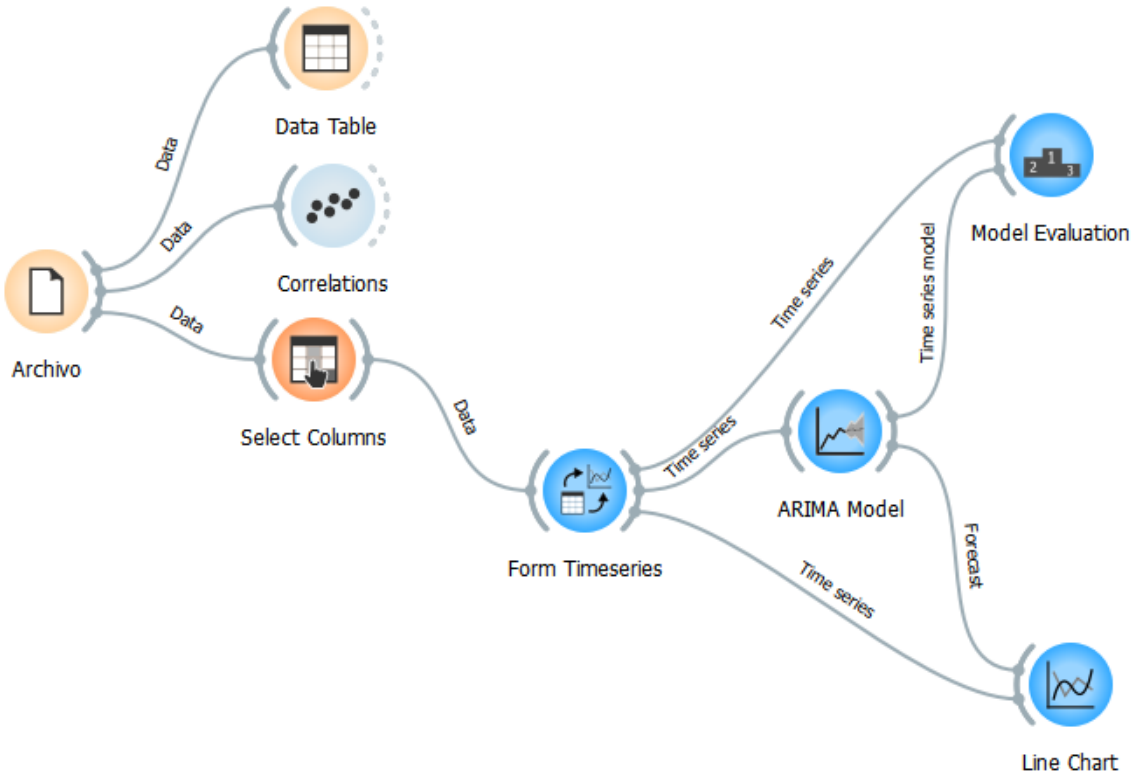


Nota. Se tomó la imagen al iniciar el aplicativo *Orage Data Miner*.

No se considera el detalle de cada paso realizado en este trabajo sin embargo se deja la imagen final del flujo para referencia del modelo, la cual se muestra en la

Figura *8*.

Figura 8. *Conjunto de nodos asociados al análisis de anomalías en Orange DataMiner*

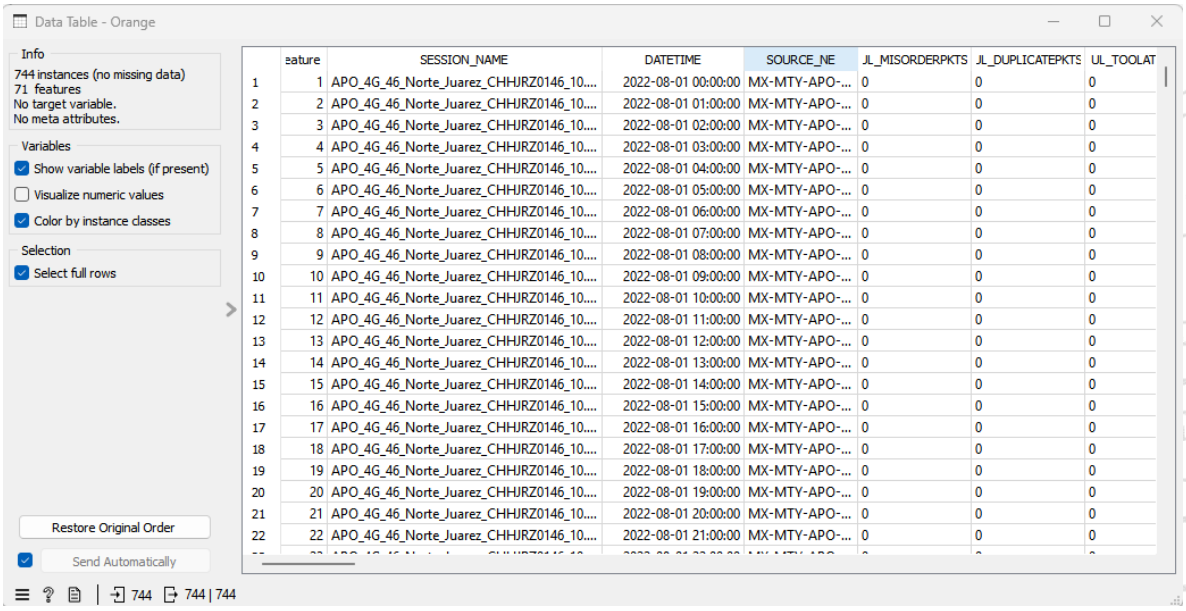


Nota. Se tomó la imagen resultante de la creación de un flujo de datos generado por la herramienta Orange-DataMiner, se genera mediante un grafo dirigido, los nodos en color azul fuerte corresponden a acciones para realizar el análisis de serie de tiempo para identificar anomalías en la red de datos.

Con el fin de ejemplificar el escenario de detección de anomalías se consideró un sitio en específico identificado con el alfanumérico [0146]. Se consideró el uso de información agregada por hora por lo que se tiene un total de 744 observaciones para este sitio, corresponden a 31 días del mes de agosto x 1 medición por hora x 24 horas al día. La

Figura *9* muestra el nodo “Data Table / Tabla de Datos” que muestra la información cargada del archivo que contiene las mediciones, este nodo no utiliza transformación de datos.

Figura 9. *Muestra de datos TWAMP de los sitios considerados para el análisis*

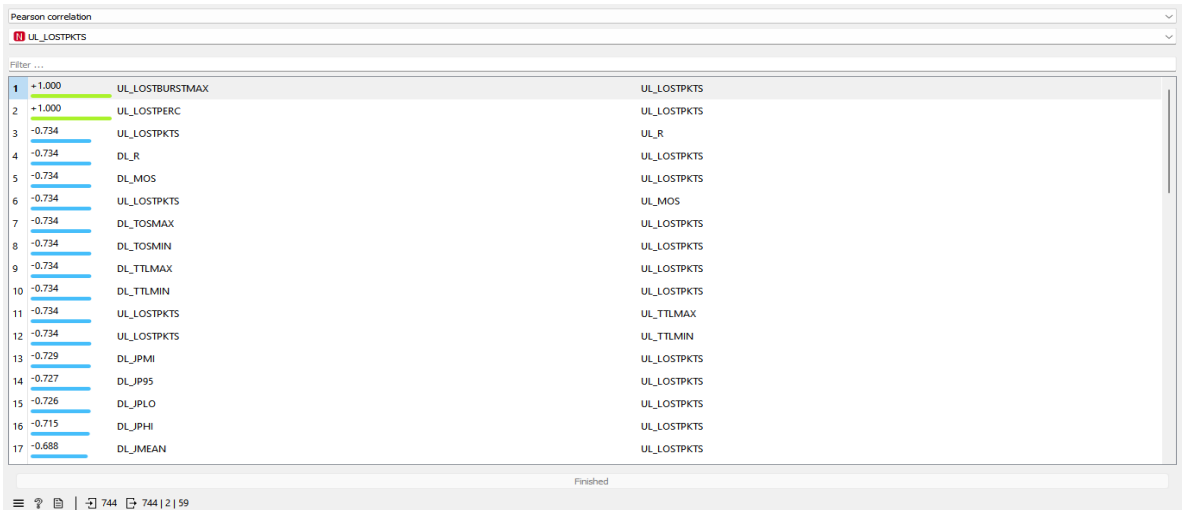


	eature	SESSION_NAME	DATETIME	SOURCE_NE	JL_MISORDERPKTS	JL_DUPLICATEPKTS	UL_TOOLAT
1	1	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 00:00:00	MX-MTY-APO-...	0	0	0
2	2	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 01:00:00	MX-MTY-APO-...	0	0	0
3	3	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 02:00:00	MX-MTY-APO-...	0	0	0
4	4	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 03:00:00	MX-MTY-APO-...	0	0	0
5	5	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 04:00:00	MX-MTY-APO-...	0	0	0
6	6	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 05:00:00	MX-MTY-APO-...	0	0	0
7	7	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 06:00:00	MX-MTY-APO-...	0	0	0
8	8	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 07:00:00	MX-MTY-APO-...	0	0	0
9	9	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 08:00:00	MX-MTY-APO-...	0	0	0
10	10	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 09:00:00	MX-MTY-APO-...	0	0	0
11	11	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 10:00:00	MX-MTY-APO-...	0	0	0
12	12	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 11:00:00	MX-MTY-APO-...	0	0	0
13	13	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 12:00:00	MX-MTY-APO-...	0	0	0
14	14	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 13:00:00	MX-MTY-APO-...	0	0	0
15	15	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 14:00:00	MX-MTY-APO-...	0	0	0
16	16	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 15:00:00	MX-MTY-APO-...	0	0	0
17	17	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 16:00:00	MX-MTY-APO-...	0	0	0
18	18	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 17:00:00	MX-MTY-APO-...	0	0	0
19	19	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 18:00:00	MX-MTY-APO-...	0	0	0
20	20	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 19:00:00	MX-MTY-APO-...	0	0	0
21	21	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 20:00:00	MX-MTY-APO-...	0	0	0
22	22	APO_4G_46_Norte_Juarez_CHHJRZ0146_10....	2022-08-01 21:00:00	MX-MTY-APO-...	0	0	0

Nota. El gráfico muestra la información procesada en el nodo Data Table, los datos se muestran de forma tabular donde cada registro corresponde a una medición TWAMP de los nodos considerados en este trabajo. Elaboración propia, considerando el resultado de *Orange Data Miner*.

Se agrega un nodo de correlación parra entender el conjunto de datos y realizar una validación de Correlación de Pearson sobre todas las variables. La Figura 10 ejemplifica el tipo de información generada tras el cálculo de la correlación. Tomando en consideración la variable UL_LOSTPKTS es decir el número de paquetes perdidos en el Uplink. Esta variable representa los paquetes perdidos en sentido origen a destino.

Figura 10. *Correlación Pearson UL_LOSTPKTS vs. el resto de las variables*

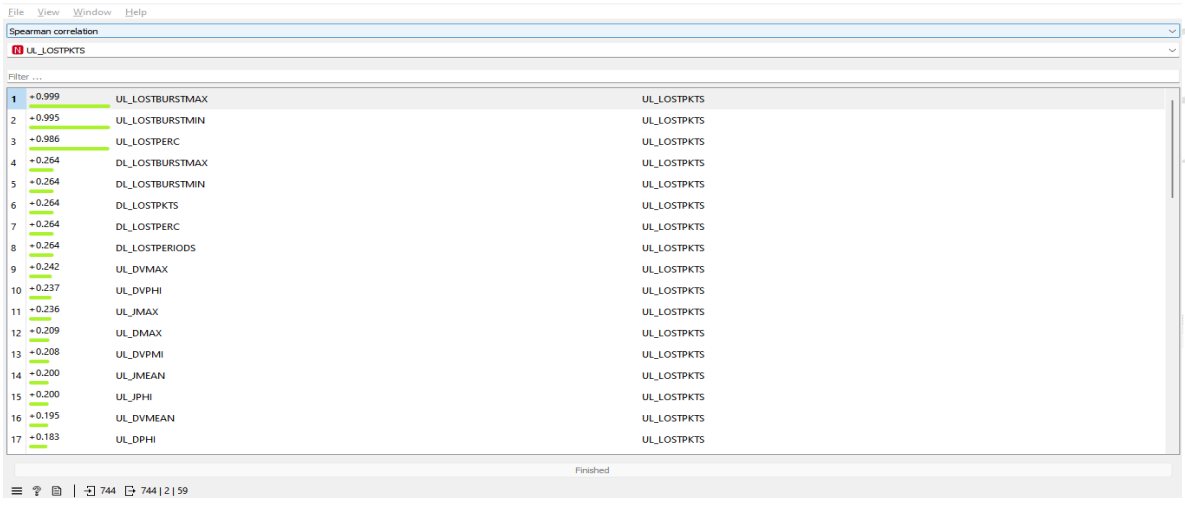


Nota. El gráfico muestra el resultado de ejecutar el cálculo de la Correlación de Pearson sobre el conjunto de datos del Dataset TWAMP. Elaboración propia, considerando el resultado de *Orange Data Miner*.

La imagen muestra una correlación con valor a 1 entre las variables UL_LOSTBURSTMAX y UL_LOSTPERC y UL_LOSTPKTS. Todas estas variables asociadas a la pérdida de paquetes en sentido origen a destino.

La Figura 11 muestra el resultado de calcular la correlación de Spearman tomando en consideración las diferentes variables del conjunto de datos.

Figura 11. *Correlación Pearson UL_LOSTPKTS vs. el resto de las variables*



Nota. El gráfico muestra el resultado de ejecutar el cálculo de la Correlación de Pearson sobre el conjunto de datos del Dataset TWAMP. Elaboración propia, considerando el resultado de Orange DataMiner.

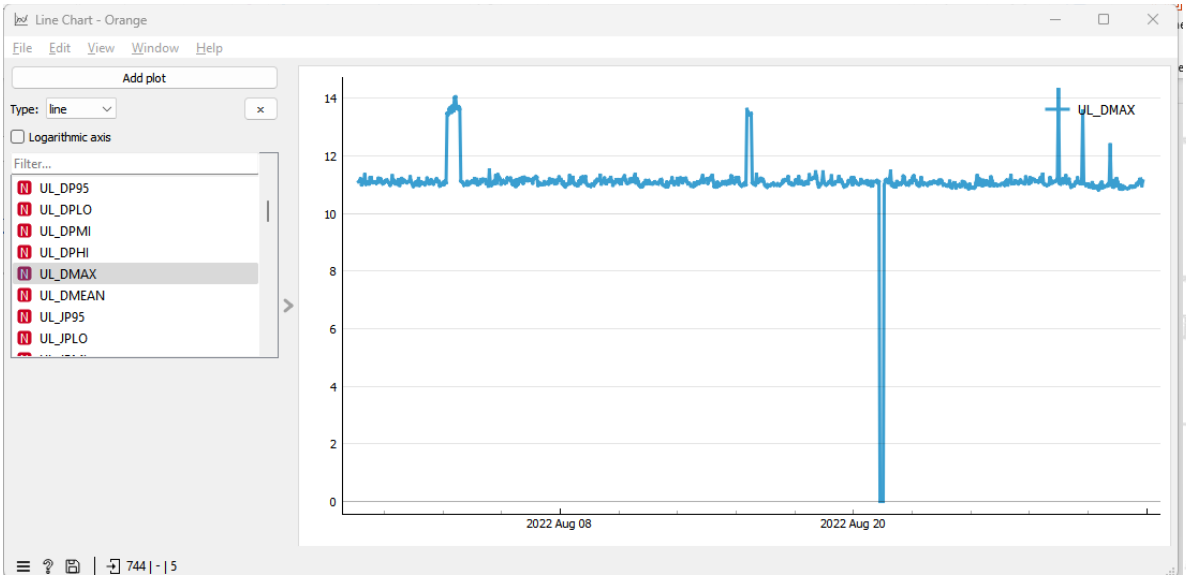
Se aprecia una correlación cercana a 1 entre las variables UL_LOSTBURSTMAX, UL_LOSTBURSTMIN y UL_LOSTPERC y UL_LOSTPKTS. En ambos casos debemos recordar el significado de las variables

UL_LOSTPKTS	Número de paquetes origen a destino perdidos durante el intervalo
UL_LOSTBURSTMIN	Duración del período de pérdida más corto de origen a destino durante el intervalo
UL_LOSTBURSTMAX	Duración del período de pérdida más largo de origen a destino durante el intervalo
UL_LOSTPERC	Tasa de número de paquetes perdido de origen a destino durante el intervalo de medición

La pérdida de paquetes esté asociada con los intervalos de pérdida de paquetes, así como la tasa de número de paquetes o porcentaje de paquetes perdidos. En el análisis de este sitio en particular vemos que existe una correlación débil con la variable de latencia (DL_DMAX)

Si revisamos el comportamiento de la latencia mostrado en la Figura 12 en la serie de tiempo no se aprecia un comportamiento elevado significativamente, se observa un periodo donde la latencia tiende a cero el día 21 de agosto en el intervalo de tiempo, no parece que se deba a una ventana de mantenimiento pues normalmente se ejecutan durante la noche, puede ser que durante ese de tiempo se generó la lectura TWAMP.

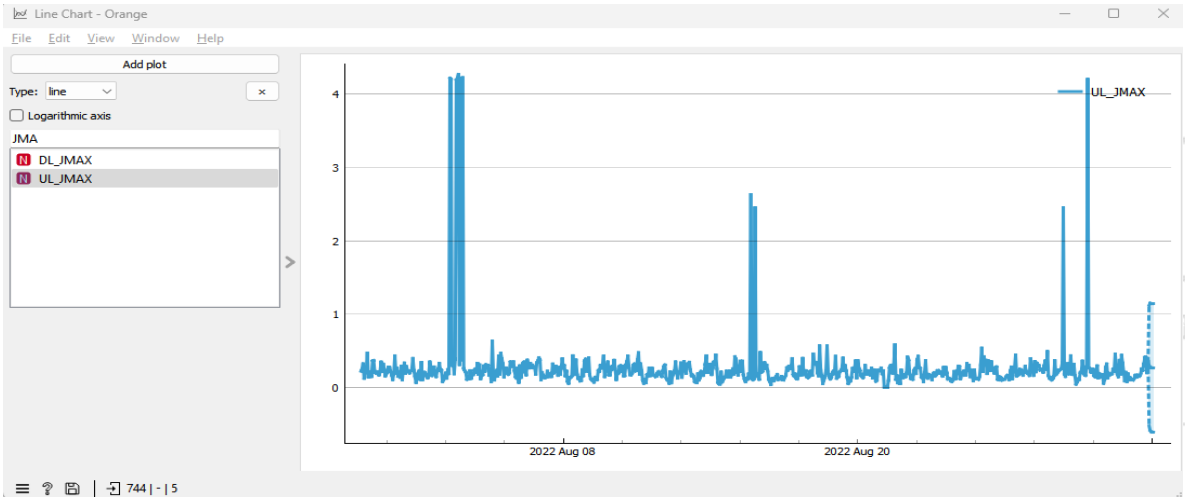
Figura 12. *Análisis de serie de tiempo de la latencia máxima del sitio*



Nota. El gráfico muestra el resultado de ejecutar el análisis de serie de tiempo sobre la variable de Latencia Máxima de Carga sobre el conjunto de datos del Dataset TWAMP. El valor de latencia en cero implica que durante el periodo de tiempo no existió medición TWAMP. Elaboración propia, considerando el resultado de Orange DataMiner.

La latencia oscila entre 12 y 14 milisegundos, dicho valor está dentro del umbral de la red 4G, pero está por encima del esperado en una red 5G donde se esperan latencias no mayores a 1 ms. En la Figura 13 se realiza una validación del Jitter experimentado por el sitio celular en este mismo intervalo de tiempo. Se identifican valores en magnitud cercanos a cero con algunos valores superiores a 4, sin que eso indique un fallo generalizado; los incrementos en Jitter se observan en horarios diurnos del medio día hasta las dos de la tarde, por lo que puede explicarse con un incremento de tráfico experimentado por el sitio, puede ser un incremento en el uso de la red por parte de los usuarios que están cercanos al sitio.

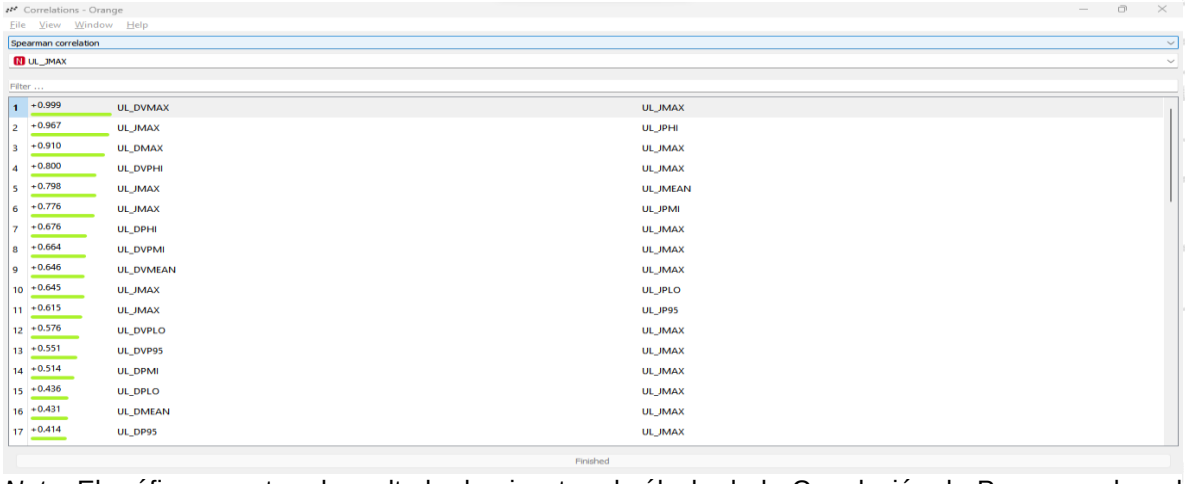
Figura 13. *Análisis de serie de tiempo de Jitter máximo del sitio*



Nota. El gráfico muestra el resultado de ejecutar el análisis de serie de tiempo sobre la variable de Jitter Máximo de Carga sobre el conjunto de datos del Dataset TWAMP. A la izquierda del gráfico se observan periodos continuos de alto Jitter, lo que puede identificar un sitio congestionado. Elaboración propia, considerando el resultado de Orange DataMiner.

Al tomar la variable UL_JMAX y revisar los coeficientes de correlación de Spearman podemos observar que el Jitter está correlacionado con la latencia como se muestra en la Figura 14.

Figura 14. *Correlación de Spearman de la variable UL_JMAX (Jitter Máximo de Carga)*



Nota. El gráfico muestra el resultado de ejecutar el cálculo de la Correlación de Pearson sobre el conjunto de datos del Dataset TWAMP. Elaboración propia, considerando el resultado de Orange DataMiner.

Esto permite entender que en la medida que la Latencia o el Jitter varíen tienen un impacto directo entre ambas.

3.3 Procedimiento de análisis de agrupaciones

A continuación, se muestra la secuencia de código empleado para el análisis de datos de las mediciones TWAMP de los sitios identificados en Ciudad Juárez. Se adjunta el ejercicio de código en R o Python según sea el caso

```
1. Se importan las librerías requeridas para el análisis
suppressWarnings(library(leaflet))
suppressWarnings(library(xtable))
suppressWarnings(library(lubridate))
suppressWarnings(library(dplyr))
suppressWarnings(library(stringr))
suppressWarnings(library(Hmisc))

#Así como para el procesamiento de Componentes Principales (PCA)
suppressWarnings(library(corrr))
suppressWarnings(library(ggcorrplot))
suppressWarnings(library(FactoMineR))
suppressWarnings(library(factoextra))
```

Posteriormente se incluirán librerías específicas en la sección de código que corresponda para evitar conflictos con las librerías cargadas al inen secciones específicas pues entran en conflicto con las librerías previamente indicadas. El siguiente paso es realizar la carga de la totalidad del conjunto de mediciones TWAMP para los sitios de Ciudad Juárez, cada uno de ellos tiene latitud y longitud de forma que se pueden geolocalizar y mostrar en un mapa y ante cualquier indicativo de falla identificar si es un escenario aislado, o se presenta en un más de un sitio.

```
sitios <- read.csv('./Agosto/sitios_cdj.csv')
head(sitios,5)
```

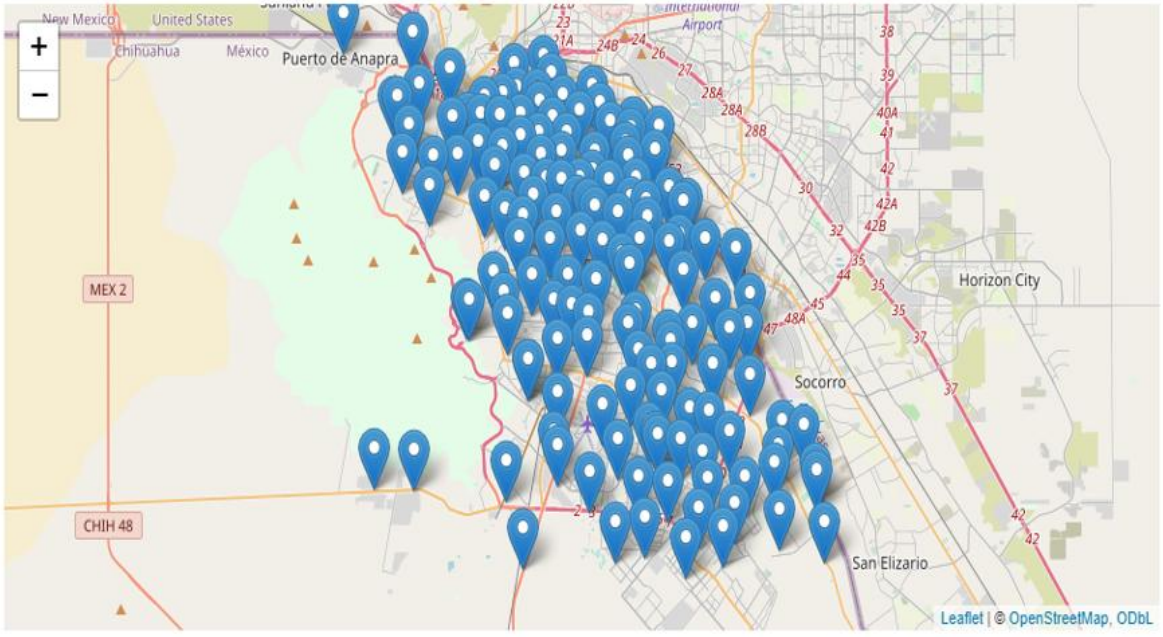
A data.frame: 5 × 3

	Sitios	Latitud	Longitud
	<chr>	<dbl>	<dbl>
1	CHHJRZ0146	31.66056	-106.3722
2	CHHJRZ0147	31.69194	-106.3653
3	CHHJRZ0148	31.70595	-106.3775
4	CHHJRZ0149	31.7x5	-106.4026
5	CHHJRZ0150	31.72273	-106.4069

En la Figura 15 se muestra la dispersión de sitios celulares para Ciudad Juárez, se observa la cercanía que se tiene con la frontera de Estados Unidos, por lo cual un factor adicional que incide en la calidad del servicio percibido por el cliente es la interferencia que se genera entre los sitios celulares en ambos lados de la frontera.

```
leaflet(data = sitios) %>% addTiles() %>%
  addMarkers(~Longitud, ~Latitud, popup = ~as.character(Sitios),
  label = ~as.character(Sitios))
```

Figura 15*. Mapa Ciudad Juárez con distribución geográfica de sitios*



Nota. El gráfico muestra el resultado el nodo “GeoMap” de Orange DataMiner sobre el conjunto de sitios físicos que corresponden a Ciudad Juárez. Elaboración propia, considerando el resultado de Orange DataMiner.

El siguiente paso es realizar la carga de información de las mediciones TWAMP.

```
df_twamp <- read.csv('./Agosto/final.csv')
dim(df_twamp)
     1. 16260080
     2. 70
```

El estimado teórico considerado en el capítulo dos de este documento considera un total de 13,302,720 eventos TWAMP. En el conjunto de datos se observa un total de 5,037,090 eventos para DSCP 40 y 5,037,092 para DSCP 46, esto para el

concentrador de Buenos Aires, mientras que existe un total de 3,092,949 para DSCP 40 y 3,092,949 para DSCP 46 en el concentrador de Apodaca.

En la exploración inicial del conjunto de datos podemos observar variables con diferentes tipos de datos. Por ejemplo, las variables SESSION_NAME y SOURCE_NE son variables de tipo texto, las cuales deberán ser eliminadas del conjunto de análisis. La variable DATETIME, que corresponde a la fecha y hora en la que se llevó a cabo el muestreo TWAMP está presente en el conjunto de datos y es la variable a la cual estaremos buscando agregar en intervalos de una hora. En este punto también se observan sesiones establecidas de los nodos hacia ambos concentradores por eso el valor total del conjunto de datos. Este análisis exploratorio también muestra que la información se envía en intervalos constantes de un minuto. Dirigiendo el análisis de datos para de DSCP 46, se eliminan los valores que no corresponden a ese DSCP del conjunto de datos. Para facilitar el análisis se propone utilizar una agregación por hora y también considerar los valores máximos pues el interés es identificar los horarios donde hay un mayor impacto en la red de conmutación.

```
df_twamp = df_twamp %>%
  filter(str_detect(SESSION_NAME,"4G_46_")) %>%
  mutate(DATETIME=strptime(DATETIME,format="%Y-%m-%d%H:%M:%OS"))
%>%
  mutate(DATETIME = floor_date(DATETIME, unit = "hour")) %>%
  group_by(SESSION_NAME,DATETIME) %>%
  summarise_all(.funs=max,na.rm=T)

  dim(df_twamp)

   1. 136896
   2. 70
```

Como resultado de esta agregación el volumen de muestras disminuyó dejando un total de 136,896 es el total de observaciones del conjunto de datos. Ahora se obtiene este número de los 149 sitios iniciales + 35 sitios que reportan en ambos concentradores. El siguiente paso es realizar la normalización de los datos tanto para subida/carga (Uplink) como para la descarga (*DL*) así como la normalización de las unidades de tiempo, los datos están en microsegundos por lo que las magnitudes de los valores son grandes y se realiza la normalización a milisegundos.

Las variables de porcentajes también son normalizadas ya que se representan en magnitudes grandes.

Normalización de las mediciones Uplink (*UL*)

```
#Normalización de las mediciones entre 1000
#Normalización de (Tiempo), conversión de microsegundos a
#milisegundos
#Normalización de UL_MOS, UL_R y UL_LOSTPERC en este caso tomando
#como denominador el valor de 10,000 de acuerdo a las
#especificaciones de la herramienta.

df_twamp['UL_DMIN'] = df_twamp['UL_DMIN'] / 1000
df_twamp['UL_DP95'] = df_twamp['UL_DP95'] / 1000
df_twamp['UL_DPLO'] = df_twamp['UL_DPLO'] / 1000
df_twamp['UL_DPMI'] = df_twamp['UL_DPMI'] / 1000
df_twamp['UL_DPHI'] = df_twamp['UL_DPHI'] / 1000
df_twamp['UL_DMAX'] = df_twamp['UL_DMAX'] / 1000
df_twamp['UL_DMEAN'] = df_twamp['UL_DMEAN'] / 1000
df_twamp['UL_JMIN'] = df_twamp['UL_JMIN'] / 1000
df_twamp['UL_JP95'] = df_twamp['UL_JP95'] / 1000
df_twamp['UL_JPLO'] = df_twamp['UL_JPLO'] / 1000
df_twamp['UL_JPMI'] = df_twamp['UL_JPMI'] / 1000
df_twamp['UL_JPHI'] = df_twamp['UL_JPHI'] / 1000
df_twamp['UL_JMAX'] = df_twamp['UL_JMAX'] / 1000
df_twamp['UL_JMEAN'] = df_twamp['UL_JMEAN'] / 1000
df_twamp['UL_DVP95'] = df_twamp['UL_DVP95'] / 1000
df_twamp['UL_DVPLO'] = df_twamp['UL_DVPLO'] / 1000
df_twamp['UL_DVPMI'] = df_twamp['UL_DVPMI'] / 1000
df_twamp['UL_DVPHI'] = df_twamp['UL_DVPHI'] / 1000
df_twamp['UL_DVMAX'] = df_twamp['UL_DVMAX'] / 1000
df_twamp['UL_DVMEAN'] = df_twamp['UL_DVMEAN'] / 1000
df_twamp['UL_MOS'] = df_twamp['UL_MOS'] /1000000
df_twamp['UL_R'] = df_twamp['UL_R'] /1000000
df_twamp['UL_LOSTPERC'] = df_twamp['UL_LOSTPERC'] / 10000
```

Normalización de las mediciones Downlink (*DL*)

```
#Normalización de las mediciones entre 1000
#Normalización de (Tiempo), conversión de microsegundos a
#milisegundos
#Normalización de DL_MOS, DL_R y DL_LOSTPERC en este caso tomando
#como denominador el valor de 10,000 de acuerdo a las
#especificaciones de la herramienta.

df_twamp['DL_DMIN'] = df_twamp['DL_DMIN'] / 1000
df_twamp['DL_DP95'] = df_twamp['DL_DP95'] / 1000
df_twamp['DL_DPLO'] = df_twamp['DL_DPLO'] / 1000
df_twamp['DL_DPMI'] = df_twamp['DL_DPMI'] / 1000
df_twamp['DL_DPHI'] = df_twamp['DL_DPHI'] / 1000
df_twamp['DL_DMAX'] = df_twamp['DL_DMAX'] / 1000
```

```
df_twamp['DL_DMEAN'] = df_twamp['DL_DMEAN'] / 1000
df_twamp['DL_JMIN'] = df_twamp['DL_JMIN'] / 1000
df_twamp['DL_JP95'] = df_twamp['DL_JP95'] / 1000
df_twamp['DL_JPLO'] = df_twamp['DL_JPLO'] / 1000
df_twamp['DL_JPMI'] = df_twamp['DL_JPMI'] / 1000
df_twamp['DL_JPHI'] = df_twamp['DL_JPHI'] / 1000
df_twamp['DL_JMAX'] = df_twamp['DL_JMAX'] / 1000
df_twamp['DL_JMEAN'] = df_twamp['DL_JMEAN'] / 1000
df_twamp['DL_DVP95'] = df_twamp['DL_DVP95'] / 1000
df_twamp['DL_DVPLO'] = df_twamp['DL_DVPLO'] / 1000
df_twamp['DL_DVPMI'] = df_twamp['DL_DVPMI'] / 1000
df_twamp['DL_DVPHI'] = df_twamp['DL_DVPHI'] / 1000
df_twamp['DL_DVMAX'] = df_twamp['DL_DVMAX'] / 1000
df_twamp['DL_DVMEAN'] = df_twamp['DL_DVMEAN'] / 1000
df_twamp['DL_MOS'] = df_twamp['DL_MOS'] /1000000
df_twamp['DL_R'] = df_twamp['DL_R'] /1000000
df_twamp['DL_LOSTPERC'] = df_twamp['DL_LOSTPERC'] / 10000
```

Mediante la estadística descriptiva revisamos los datos del conjunto para describir los mismos. Inicialmente observamos que ya existen mediciones donde hay Pérdida de paquetes e incluso el valor máximo observado de pérdida de paquetes es 100% lo cual es un indicador de un sitio con problemas de comunicación y que seguramente está afectando la experiencia del suscriptor.

- UL_LOSTPKTS -> Valor máximo de 596 paquetes perdidos en UpLink
- UL_JMAX -> Valor máximo de Jitter en UpLink de 69.8690
- DL_LOSTPKTS -> Valor máximo de 678 paquetes perdidos en DownLink
- DL_JMAX -> Valor máximo de Jitter en DownLink de 916.0380

El hecho de tener pérdida de paquetes no simétrico es decir para Uplink tenemos un máximo de 596 y para Downlink tenemos un máximo de 678 puede ser indicativo que se están ocupando rutas diferentes para llegar a los sitios celulares, es decir ocupan una ruta para Uplink y otra diferente para Downlink. También puede ser que alguno de los ruteadores intermedios pueda tener congestión. Se eliminaron del conjunto de datos aquellas variables no numéricas y aquellas variables que contienen valores nulos.

```
df_twamp_num = df_twamp %>%
  ungroup %>%
  select(-SESSION_NAME) %>%
  select(-SOURCE_NE) %>%
  select(-DATETIME) %>%
  select(-UL_MISORDERPKTS) %>%
  select(-UL_TOOLATEPKTS) %>%
```

```
  select(-UL_TOSMIN) %>%
  select(-UL_TOSMAX) %>%
  select(-DL_DUPLICATEPKTS)

dim(df_twamp_num)
```

1. 136896
2. 62

El número de muestras se mantiene después de realizar la desagrupación de valores y la eliminación de valores no numéricos. Debido a las magnitudes de cada una de las variables se realiza la aplicación de la función de normalización. La Figura 16 muestra el fragmento de código Python que se utiliza para la normalización del conjunto de datos.

Figura 16. *Normalización del conjunto de datos*

5.2 Normalización de los datos

```
data_normalized <- scale(df_twamp_num)
head(data_normalized)
```

A matrix: 6 × 62 of type

UL_DUPLICATEPKTS	UL_LOSTPKTS	UL_LOSTBURSTMIN	UL_LOSTBURSTMAX	UL_LOSTPERC	UL_MOS	UL_R	UL_TTLMIN	UL_TTLMAX	UL_DMIN
-0.002702742	-0.4329481	-0.4935153	-0.4328811	-0.4329229	0.4246743	0.4246743	0.4478	0.4478	0.3708734
-0.002702742	-0.4329481	-0.4935153	-0.4328811	-0.4329229	0.4246743	0.4246743	0.4478	0.4478	0.3699093
-0.002702742	-0.4329481	-0.4935153	-0.4328811	-0.4329229	0.4246743	0.4246743	0.4478	0.4478	0.3703913
-0.002702742	-0.4282692	1.6529589	-0.4282022	-0.4282220	0.4246743	0.4246743	0.4478	0.4478	0.3696682
-0.002702742	-0.4329481	-0.4935153	-0.4328811	-0.4329229	0.4246743	0.4246743	0.4478	0.4478	0.3699093
-0.002702742	-0.4329481	-0.4935153	-0.4328811	-0.4329229	0.4246743	0.4246743	0.4478	0.4478	0.3696682

Nota. El gráfico muestra el resultado de ejecutar el proceso de normalización y mostrar las primeras filas del resultado. Elaboración propia, considerando el resultado de código Python empleado en el análisis del conjunto de datos.

Para ejecutar el proceso de Análisis de Componentes Principales (ACP) se debe verificar que los valores no sean nulos o no definidos, de lo contrario se presentará una excepción en la ejecución del código. La Figura 17 muestra la validación e identificación de variables NaN, para este ejercicio no existen ese tipo de valores.

Figura 17. *Validación de las variables NaN del conjunto de datos*

```
#Verificar si existe alguna columna como NaN
colSums(is.na(data_normalized))
```

```
UL_DUPLICATEPKTS: 0 UL_LOSTPKTS: 0 UL_LOSTBURSTMIN: 0 UL_LOSTBURSTMAX: 0 UL_LOSTPERC: 0 UL_MOS: 0 UL_R: 0 UL_TTLMIN: 0
UL_TTLMAX: 0 UL_DMIN: 0 UL_DP95: 0 UL_DPLO: 0 UL_DPMI: 0 UL_DPHI: 0 UL_DMAX: 0 UL_DMEAN: 0 UL_JMIN: 0 UL_JP95: 0 UL_JPLO: 0
UL_JPMI: 0 UL_JPHI: 0 UL_JMAX: 0 UL_JMEAN: 0 UL_DVP95: 0 UL_DVPLO: 0 UL_DVPMI: 0 UL_DVPHI: 0 UL_DVMAX: 0 UL_DVMEAN: 0
DL_MISORDERPKTS: 0 DL_TOOLATEPKTS: 0 DL_LOSTPKTS: 0 DL_LOSTPERIODS: 0 DL_LOSTBURSTMIN: 0 DL_LOSTBURSTMAX: 0
DL_LOSTPERC: 0 DL_MOS: 0 DL_R: 0 DL_TOSMIN: 0 DL_TOSMAX: 0 DL_TTLMIN: 0 DL_TTLMAX: 0 DL_DMIN: 0 DL_DP95: 0 DL_DPLO: 0
DL_DPMI: 0 DL_DPHI: 0 DL_DMAX: 0 DL_DMEAN: 0 DL_JMIN: 0 DL_JP95: 0 DL_JPLO: 0 DL_JPMI: 0 DL_JPHI: 0 DL_JMAX: 0 DL_JMEAN: 0
DL_DVP95: 0 DL_DVPLO: 0 DL_DVPMI: 0 DL_DVPHI: 0 DL_DVMAX: 0 DL_DVMEAN: 0
```

```
dim(data_normalized)
```

Nota. El gráfico muestra el resultado de ejecutar el proceso de identificación de variables con observaciones sin valor (NaN) del conjunto de datos. Elaboración propia, considerando el resultado de código Python empleado en el análisis del conjunto de datos.

Se realiza el cálculo de la matriz de correlación. Una matriz de correlación:

> Es un método estadístico para cuantificar y comparar las relaciones entre diferentes variables de un conjunto de datos. Las correlaciones se realizan entre todas las combinaciones de variables y se muestran en una estructura tabular. Cada celda de la matriz contiene el denominado coeficiente de correlación entre las variables definidas en la cada columna y cada fila.[53]

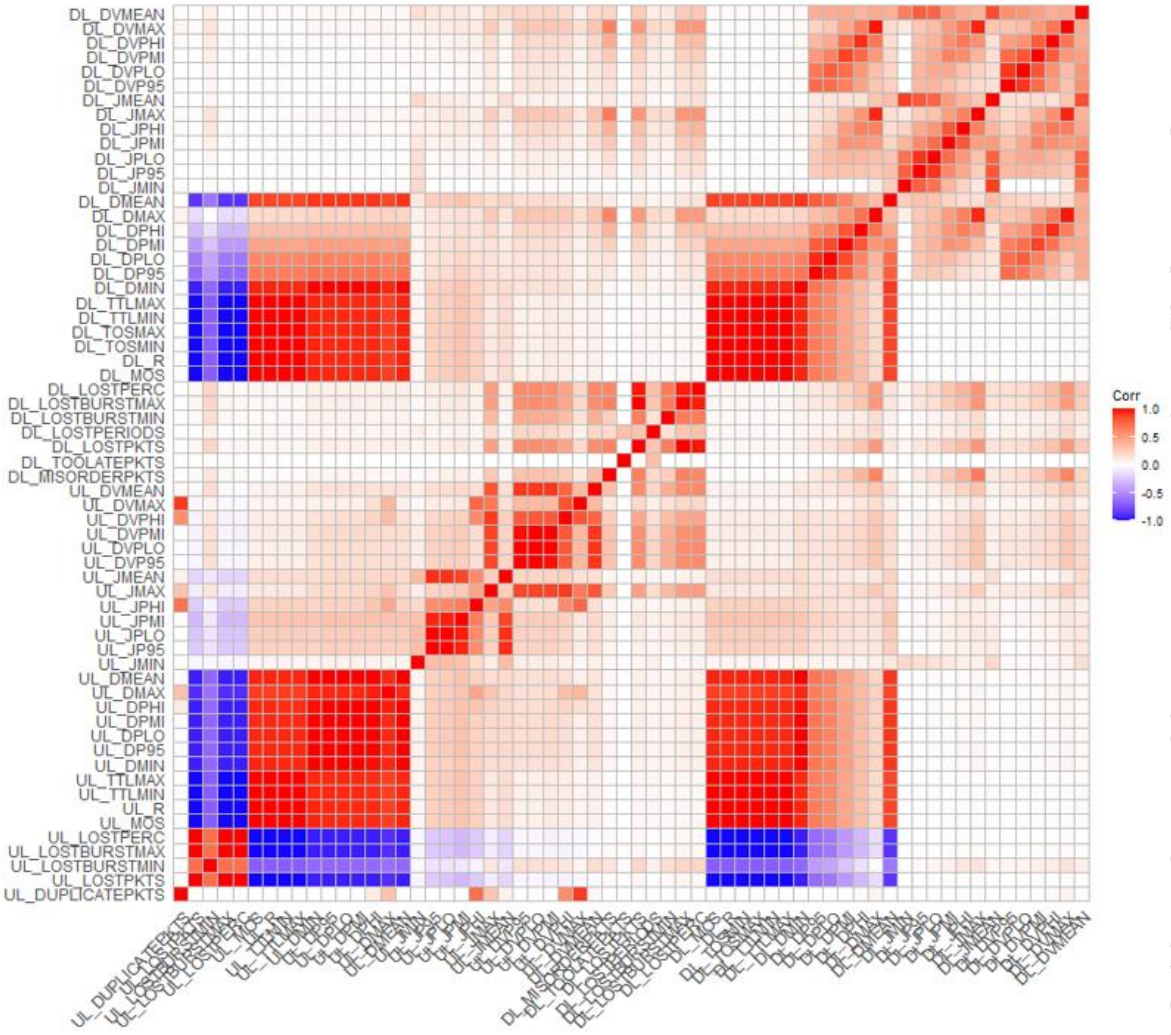
En nuestro caso sería una matriz cuadrada de 62 x 62, recordando que se eliminaron ocho variables del conjunto inicial de 70 variables.

```
png(filename = "corrplot.png", width = 1200, height = 800)
corr_matrix <- cor(data_normalized)
ggcorrplot(corr_matrix)
dev.off()
```

La Figura 18 muestra el resultado de la ejecución de la matriz de correlación podemos observar el comportamiento de las variables. Se observa un patrón diferente en el tráfico de *UL* en comparación con el tráfico de *DL*. En escenarios de tráfico asimétrico, las velocidades de bits de enlace de carga y de descarga pueden diferir y tener una proporción diferente para cada segmento. Es decir, la red móvil no tiene un patrón de tráfico simétrico. Se observa que las variables de descarga tienen un patrón de correlación mayor que el de las variables de carga.

Figura 18. *Matriz de correlación del conjunto de datos*

[53] Lang, Niklas, Demistifying the Correlation Matrix in Dagta Science, Towards Data Science. Disponible en Internet: [https://towardsdatascience.com/demistifying-the-correlation-matrix-in-data-science-6b8a4482b6e2/]. Noviembre 2024.



Nota. El gráfico muestra el resultado de ejecutar la visualización de la correlación de variables del conjunto de datos. Elaboración propia, considerando el resultado de código Python empleado en el análisis del conjunto de datos.

Con la ejecución de la matriz de correlación, el siguiente paso es el cálculo de componentes principales, donde el interés es analizar aquellos componentes que conjuntan una proporción de la varianza.

```
twamp_pca <- princomp(corr_matrix)
summary(twamp_pca)
```

Es un total de 62 componentes principales calculados, se puede apreciar que los primeros ocho explican el 99.12% de la varianza.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3    Comp.4     Comp.5
Standard deviation     2.5183471 0.86372764 0.76348489 0.4669913 0.39622670
Proportion of Variance 0.7579893 0.08916317 0.06966791 0.0260645 0.01876374
Cumulative Proportion  0.7579893 0.84715244 0.91682035 0.9428848 0.96164858
                           Comp.6     Comp.7      Comp.8      Comp.9
Standard deviation     0.35484636 0.30033839 0.177068483 0.153485409
Proportion of Variance 0.01504917 0.01078086 0.003747265 0.002815568
Cumulative Proportion  0.97669775 0.98747862 0.991225884 0.994041452
                          Comp.10     Comp.11      Comp.12      Comp.13
Standard deviation     0.119942521 0.104166488 0.0773111857 0.0655505575
Proportion of Variance 0.001719405 0.001296845 0.0007143591 0.0005135523
Cumulative Proportion  0.995760858 0.997057702 0.9977720613 0.9982856136
                          Comp.14      Comp.15      Comp.16     Comp.17
Standard deviation     0.058874878 0.0564948408 0.0513501518 0.040374020
Proportion of Variance 0.000414278 0.0003814604 0.0003151486 0.000194821
Cumulative Proportion  0.998699892 0.9990813520 0.9993965006 0.999591322
                           Comp.18      Comp.19      Comp.20      Comp.21
Standard deviation     0.0343717009 0.0291395986 2.084922e-02 1.692733e-02
Proportion of Variance 0.0001411998 0.0001014843 5.195309e-05 3.424593e-05
Cumulative Proportion  0.9997325214 0.9998340057 9.998860e-01 9.999202e-01
                           Comp.22      Comp.23      Comp.24      Comp.25
Standard deviation     1.395146e-02 1.247816e-02 1.168894e-02 9.650884e-03
Proportion of Variance 2.326329e-05 1.860943e-05 1.632985e-05 1.113182e-05
Cumulative Proportion  9.999435e-01 9.999621e-01 9.999784e-01 9.999895e-01
                           Comp.26      Comp.27      Comp.28      Comp.29
Standard deviation     5.542274e-03 5.205392e-03 3.929837e-03 2.674751e-03
Proportion of Variance 3.671199e-06 3.238462e-06 1.845784e-06 8.550639e-07
Cumulative Proportion  9.999932e-01 9.999964e-01 9.999983e-01 9.999991e-01
                           Comp.30      Comp.31      Comp.32      Comp.33
Standard deviation     1.963242e-03 1.224721e-03 1.089451e-03 5.821793e-04
Proportion of Variance 4.606588e-07 1.792694e-07 1.418558e-07 4.050843e-08
Cumulative Proportion  9.999996e-01 9.999998e-01 9.999999e-01 1.000000e+00
```

El resultado nos indica que los primeros tres componentes principales explican el 91.68% de la varianza del conjunto de datos, y tomando en consideración los primeros ocho componentes se explica el 99.12% de la varianza. Esto puede ayudar a seleccionar el total de componentes que puede ir de tres y hasta ocho, si se incluyen más componentes el porcentaje de varianza que explican es marginal. Dentro del análisis se realiza una comparación entre componentes principales, en este caso del primer y segundo componente principal. Esto se realiza con la siguiente porción de código:

```
twamp_pca$loadings[, 1:2]
```

A matrix: 62 × 2 of type dbl

	Comp.1	Comp.2
UL_DUPLICATEPKTS	0.0067161941	0.054229879
UL_LOSTPKTS	0.2095635876	-0.002108367
UL_LOSTBURSTMIN	0.1646298397	-0.022381011
UL_LOSTBURSTMAX	0.2095468753	-0.002053691
UL_LOSTPERC	0.2095259519	-0.002079949
UL_MOS	-0.2075254266	-0.006694203
UL_R	-0.2075256044	-0.006695275
UL_TTLMIN	-0.2068396477	-0.006304297
UL_TTLMAX	-0.2068396477	-0.006304297
UL_DMIN	-0.2020076639	-0.013486676
UL_DP95	-0.2013153021	-0.012319041
UL_DPLO	-0.2012999806	-0.012273764
UL_DPMI	-0.2012543642	-0.012051908
UL_DPHI	-0.1996892075	-0.006546073
UL_DMAX	-0.1881548200	0.006723301
UL_DMEAN	-0.2017497322	-0.012658259
UL_JMIN	-0.0006080532	0.017561581

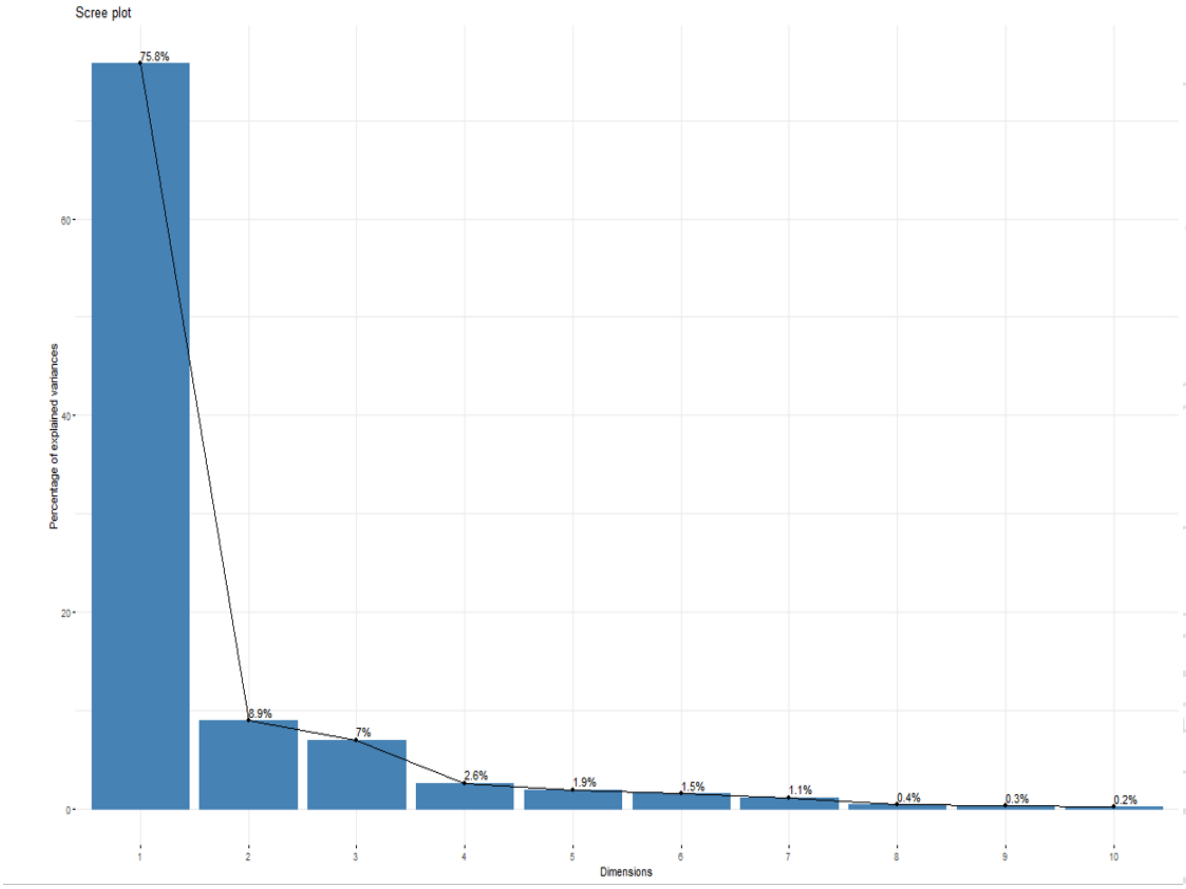
DL_DPLO	-0.11373222	-0.186858909
DL_DPMI	-0.07652259	-0.240900970
DL_DPHI	-0.04113695	-0.252917097
DL_DMAX	-0.00299965	-0.242763522
DL_DMEAN	-0.17537595	-0.089364607
DL_JMIN	0.01825497	-0.070576920
DL_JP95	0.02538414	-0.180306447
DL_JPLO	0.02761933	-0.203300983
DL_JPMI	0.03084640	-0.238738773
DL_JPHI	0.03249524	-0.237232177
DL_JMAX	0.03446420	-0.222481578
DL_JMEAN	0.02502607	-0.128304827
DL_DVP95	0.01491578	-0.214365247
DL_DVPLO	0.01735998	-0.235896994
DL_DVPMI	0.02380611	-0.272278111
DL_DVPHI	0.02809902	-0.267046402
DL_DVMAX	0.03546546	-0.246511770
DL_DVMEAN	0.02937551	-0.212360839

Cabe mencionar que el resultado de la función no debe confundirse con los valores de los Eigenvectores que se calculan posteriormente. Si bien la información numérica ayuda al análisis de los datos, la representación visual ayuda a comprender el comportamiento de los componentes principales. La visualización de componentes principales se realiza mediante la siguiente secuencia de comandos:

```
png(filename = "eigen-01.png", width = 1200, height = 800)
fviz_eig(twamp_pca, addlabels = TRUE)
dev.off()
```

La Figura 19 muestra el resultado de aplicar los comandos previos a fin de desplegar los Componentes Principales identificados por este método.

Figura 19. *Visualización de Componentes Principales*



Nota. El gráfico muestra el resultado de ejecutar la visualización de los Componentes Principales de variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Los primeros tres componentes son los más importantes pues contienen el 91.7% del total de la información total de los datos

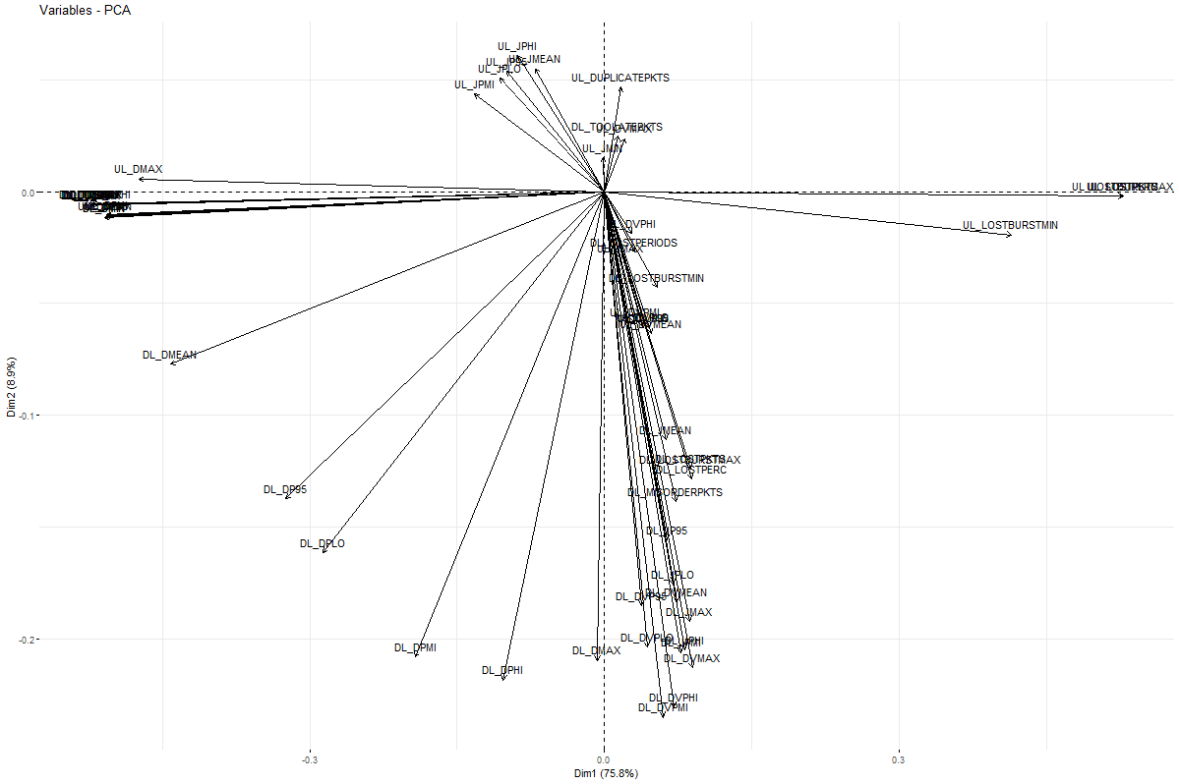
El siguiente paso es la visualización entre similitudes y diferencias de las muestras.

5.5 Visualizar las similitudes y diferencias entre las muestras

```
png(filename = "eigen-02.png", width = 1200, height = 800)
fviz_pca_var(twamp_pca, col.var = "black")
dev.off()
```

Con la visualización Biplot, se puede visualizar las similitudes y diferencias de las muestras, además es posible identificar el impacto de cada atributo en cada uno de los componentes principales. La Figura 20 muestra el Biplot de las variables con respecto a los componentes principales.

Figura 20. *Biplot de las variables respecto a los componentes principales*



Nota. El gráfico muestra el resultado del Biplot sobre los Componentes Principales de variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

1. Todas las variables que se agrupan están correlacionadas positivamente entre sí
2. Cuanto mayor sea la distancia entre la variable y el origen, mejor representada estará esa variable

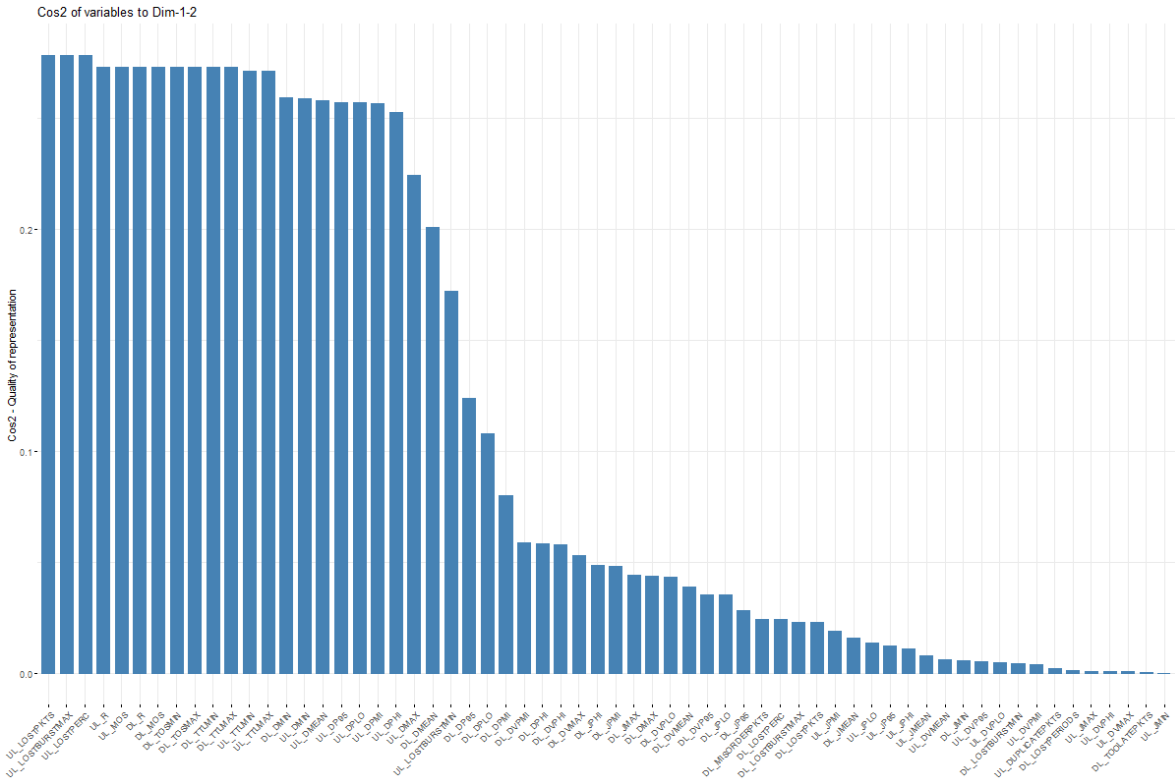
Es de interés conocer como contribuye cada variable del conjunto de datos a los componentes principales, el siguiente ejercicio sólo muestra los dos primeros componentes principales.

5.6 Contribución de cada variable a los dos primeros componentes PC1 / PC2

```
png(filename = "eigen-03.png", width = 1200, height = 800)
fviz_cos2(twamp_pca, choice = "var", axes = 1:2)
dev.off()
```

La Figura 21 muestra el histograma tomando en consideración los dos primeros componentes principales.

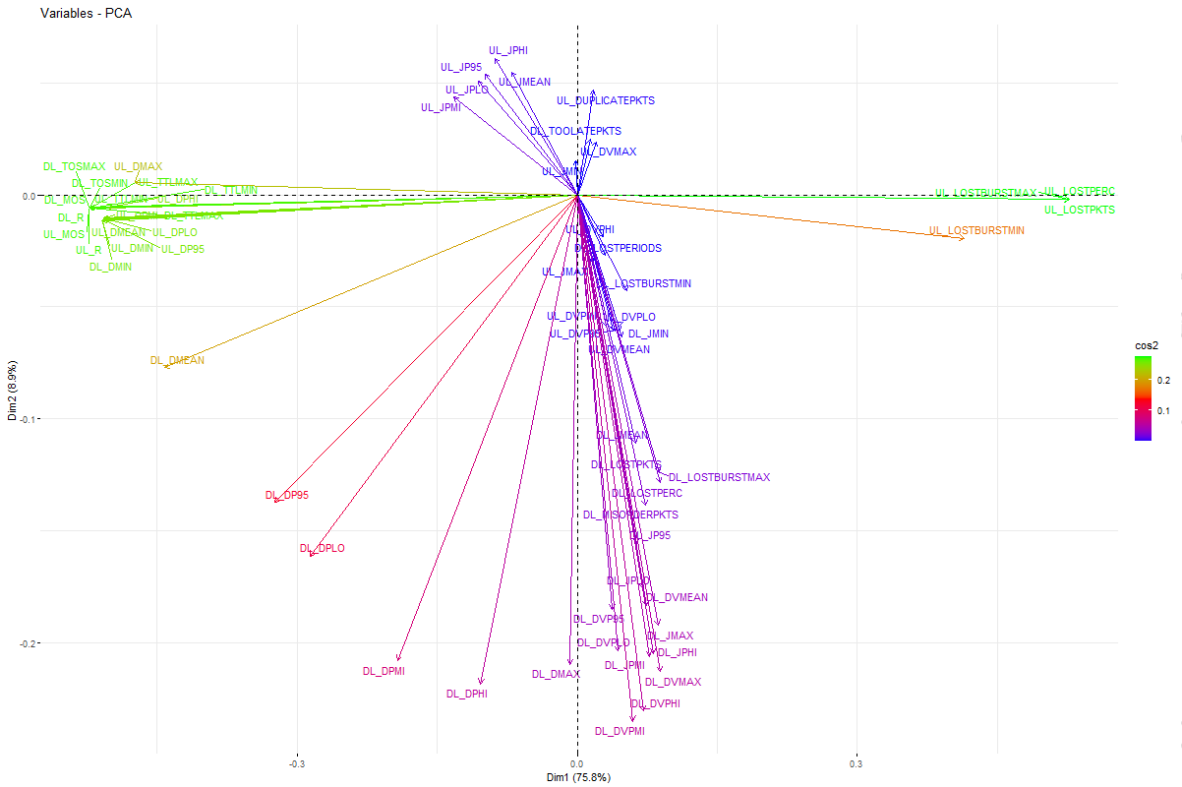
Figura 21. *Visualización de Componentes Principales*



Nota. El gráfico muestra el resultado de ejecutar la visualización de los Componentes Principales de variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

1. Un valor bajo significa que la variable no está perfectamente representada por ese componente.
2. Por otro lado, un valor alto significa una buena representación de la variable en ese componente.
3. UL_LOSTPKTS, UL_LOSTBURSTSMAX, UL_LOSTPERC son las variables que más contribuyen a los PC1 y PC2

Figura 22. *Biplot de Componentes Principales*



Nota. El gráfico muestra el resultado del Biplot sobre los Componentes Principales de variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Otra forma alternativa de mostrar la visualización de componentes principales es la mostrada en la Figura 22 ocupando un Biplot con colores para identificar los componentes principales y su relación entre sí. De esta misma figura se puede inferir:

1. Todas las variables que se agrupan están correlacionadas positivamente entre sí.
2. Mientras mayor sea la distancia entre la variable y el origen estará mejor representada
3. Las variables correlacionadas negativamente se visualizan en lados opuestos del origen del Biplot.

Al decidir cuántos componentes se manifiestan en una situación particular, deberá examinarse cuantos componentes son necesarios incluir para que el porcentaje de variación explicada sea satisfactorio. Kaiser (Kaiser, 1960) propone un criterio para seleccionar el número de componentes principales, el cual consiste en incluir sólo

aquellos componentes cuyos valores propios sean superiores al promedio; la desventaja es que tiende a incluir muy pocos componentes cuando el número de variables es inferior a veinte. La siguiente porción de código realiza el cálculo del criterio de Kaiser.

```
suppressWarnings(library(MASS))
suppressWarnings(library(paran))
#eigen valores
eig.val<-get_eigenvalue(twamp_pca)
eig.val
```

```
A data.frame: 62 × 3
eigenvalue variance.percentcumulative.variance.percent
<dbl> <dbl> <dbl>
Dim.1 6.342072e+00    7.579893e+01    75.79893
Dim.2 7.460254e-01    8.916317e+00    84.71524
Dim.3 5.829092e-01    6.966791e+00    91.68203
Dim.4 2.180808e-01    2.606450e+00    94.28848
Dim.5 1.569956e-01    1.876374e+00    96.16486
Dim.6 1.259159e-01    1.504917e+00    97.66978
Dim.7 9.020315e-02    1.078086e+00    98.74786
Dim.8 3.135325e-02    3.747265e-01    99.12259
Dim.9 2.355777e-02    2.815568e-01    99.40415
Dim.10    1.438621e-02    1.719405e-01    99.57609
Dim.11    1.085066e-02    1.296845e-01    99.70577
Dim.12    5.977019e-03    7.143591e-02    99.77721
Dim.13    4.296876e-03    5.135523e-02    99.82856
Dim.14    3.466251e-03    4.142780e-02    99.86999
Dim.15    3.191667e-03    3.814604e-02    99.90814
Dim.16    2.636838e-03    3.151486e-02    99.93965
Dim.17    1.630061e-03    1.948210e-02    99.95913
Dim.18    1.181414e-03    1.411998e-02    99.97325
Dim.19    8.491162e-04    1.014843e-02    99.98340
Dim.20    4.346899e-04    5.195309e-03    99.98860
Dim.21    2.865346e-04    3.424593e-03    99.99202
Dim.22    1.946432e-04    2.326329e-03    99.99435
Dim.23    1.557045e-04    1.860943e-03    99.99621
Dim.24    1.366313e-04    1.632985e-03    99.99784
Dim.25    9.313956e-05    1.113182e-03    99.99895
Dim.26    3.071680e-05    3.671199e-04    99.99932
Dim.27    2.709611e-05    3.238462e-04    99.99964
Dim.28    1.544362e-05    1.845784e-04    99.99983
Dim.29    7.154292e-06    8.550639e-05    99.99991
Dim.30    3.854317e-06    4.606588e-05    99.99996
⋮    ⋮    ⋮    ⋮
Dim.33    3.389328e-07    4.050843e-06    100
Dim.34    1.258690e-07    1.504356e-06    100
Dim.35    5.150685e-08    6.155976e-07    100
Dim.36    2.296377e-08    2.744575e-07    100
```

Dim.37	1.044728e-08	1.248634e-07	100
Dim.38	1.013250e-08	1.211012e-07	100
Dim.39	6.658222e-09	7.957747e-08	100
Dim.40	3.680985e-09	4.399425e-08	100
Dim.41	1.443153e-09	1.724822e-08	100
Dim.42	1.060837e-09	1.267887e-08	100
Dim.43	4.856331e-10	5.804170e-09	100
Dim.44	2.933492e-10	3.506040e-09	100
Dim.45	2.394358e-10	2.861679e-09	100
Dim.46	1.324763e-11	1.583326e-10	100
Dim.47	1.070167e-11	1.279038e-10	100
Dim.48	1.818613e-12	2.173563e-11	100
Dim.49	1.024625e-12	1.224607e-11	100
Dim.50	8.856057e-14	1.058455e-12	100
Dim.51	3.528024e-14	4.216609e-13	100
Dim.52	1.108050e-14	1.324315e-13	100
Dim.53	9.796392e-15	1.170841e-13	100
Dim.54	4.929281e-15	5.891358e-14	100
Dim.55	3.407566e-15	4.072641e-14	100
Dim.56	5.019326e-16	5.998978e-15	100
Dim.57	0.000000e+00	0.000000e+00	100
Dim.58	0.000000e+00	0.000000e+00	100
Dim.59	0.000000e+00	0.000000e+00	100
Dim.60	0.000000e+00	0.000000e+00	100
Dim.61	0.000000e+00	0.000000e+00	100
Dim.62	0.000000e+00	0.000000e+00	100

Bajo el criterio Kaiser, sólo se debe mantener el primer componente principal, con un valor de seis. De igual forma este criterio muestra de igual forma que los primeros ocho componentes agrupan el 99.12% de la varianza.

De forma alterna se realizó un análisis en paralelo por medio de un *Scree Plot* se visualiza el punto de inflexión. Horn (Horn, 1965) introdujo el análisis paralelo con *Scree Plot*, dicho análisis evalúa gráfica y numéricamente el número de componentes a mantener en el análisis componentes principales. Supone que si los datos fueran aleatorios, observaríamos factores no correlacionados, por lo tanto, el valor propio del análisis de componentes principales sería igual a 1. La técnica de Horn permite comparar los resultados del análisis de componentes principales obtenidos con los resultados del mismo ejercicio usando datos aleatorios. Empleando los criterios de la librería (Paran) de R, los resultados del siguiente gráfico se pueden evaluar de dos maneras:

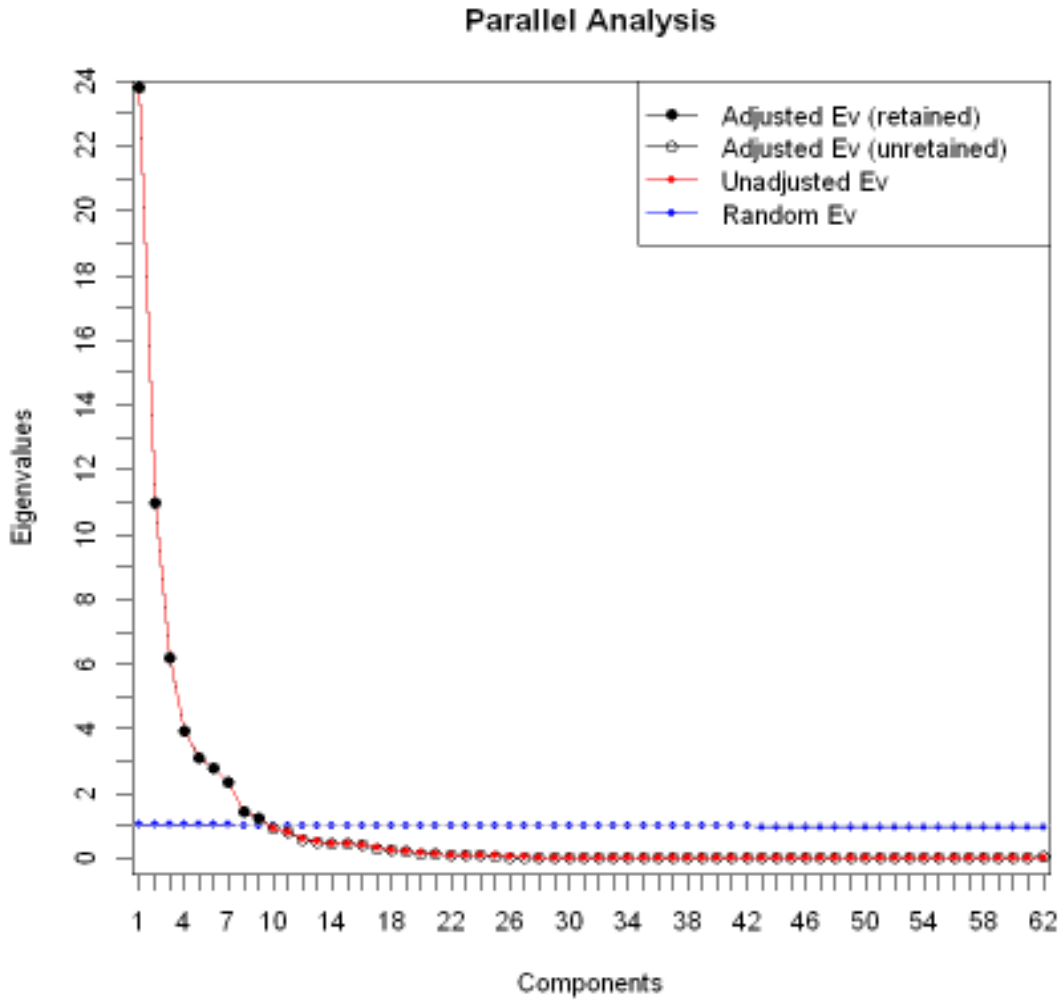
- Mantenemos aquellos factores que tienen un valor propio ajustado superior a 1.
- Mantenemos aquellos factores cuyos valores propios reales de PCA son más altos que los generados aleatoriamente.
- En este ejercicio consideramos únicamente 250 iteraciones debido a la restricción del poder de cómputo del equipo empleado.

```
paran(data_normalized, iterations=250, quietly=FALSE,
    status=FALSE, all=TRUE, cfa=FALSE, graph=TRUE,
    color=TRUE, col=c("black","red","blue"),
    lty=c(1,2,3), lwd=1, legend=TRUE, file="",
    width=640, height=640, grdevice="png", seed=0, mat=NA, n=NA)

Using eigendecomposition of correlation matrix.

Results of Horn's Parallel Analysis for component retention
250 iterations, using the mean estimate

-------------------------------------------------
Component   Adjusted    Unadjusted    Estimated
            Eigenvalue  Eigenvalue    Bias
-------------------------------------------------
1          23.790657   23.831306      0.040649
2          10.990884   11.028654      0.037769
3           6.177764    6.213203      0.035439
4           3.917582    3.951202      0.033619
5           3.098211    3.130111      0.031899
6           2.775487    2.805744      0.030257
7           2.339701    2.368342      0.028640
8           1.426229    1.453404      0.027175
9           1.239352    1.265058      0.025705
10          0.920743    0.945143      0.024399
11          0.801503    0.824549      0.023046
12          0.591408    0.613159      0.021750
13          0.499894    0.520406      0.020512
14          0.450061    0.469330      0.019269
15          0.445410    0.463521      0.018111
16          0.396844    0.413806      0.016961
17          0.304101    0.319877      0.015776
18          0.256557    0.271205      0.014648
19          0.216123    0.229632      0.013509
20          0.151857    0.164197      0.012339
21          0.122193    0.133436      0.011242
22          0.100392    0.110531      0.010138
23          0.089562    0.098567      0.009004
24          0.084285    0.092126      0.007840
25          0.071080    0.077820      0.006740
```

```
26          0.037985    0.043690      0.005705
27          0.036463    0.041169      0.004705
28          0.034048    0.037698      0.003650
29          0.019619    0.022158      0.002539
30          0.016048    0.017459      0.001410
31          0.014954    0.015305      0.000350
32          0.010054    0.009383     -0.00067
33          0.010285    0.008576     -0.00170
34          0.005746    0.002935     -0.00281
35          0.005664    0.001787     -0.00387
36          0.006279    0.001372     -0.00490
37          0.006833    0.000833     -0.00600
38          0.007853    0.000792     -0.00706
39          0.008852    0.000750     -0.00810
40          0.009742    0.000478     -0.00926
41          0.010700    0.000371     -0.01032
42          0.011757    0.000263     -0.01149
43          0.012833    0.000244     -0.01258
44          0.013840    0.000160     -0.01367
45          0.014863    0.000134     -0.01472
46          0.015931    3.306464     -0.01589
47          0.017066    2.780905     -0.01703
48          0.018218    1.181789     -0.01820
49          0.019397    7.987619     -0.01938
50          0.020587    2.954015     -0.02058
51          0.021810    1.614213     -0.02180
52          0.023036    1.479102     -0.02303
53          0.024350    7.779950     -0.02434
54          0.025654    7.546087     -0.02565
55          0.027056    5.418253     -0.02705
56          0.028465    0.000000     -0.02846
57          0.029936    0.000000     -0.02993
58          0.031424    0.000000     -0.03142
59          0.033062    0.000000     -0.03306
60          0.034792   0.000000     -0.03479
61          0.036964   0.000000     -0.03696
62          0.039922   0.000000     -0.03992
-------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to retain.
(9 components retained)
```

Dando como resultado el resultado de la Figura 23, en el cual se tiene un punto de inflexión en el noveno Componente Principal.

Figura 23. *Visualización Scree Plot, punto de inflexión*

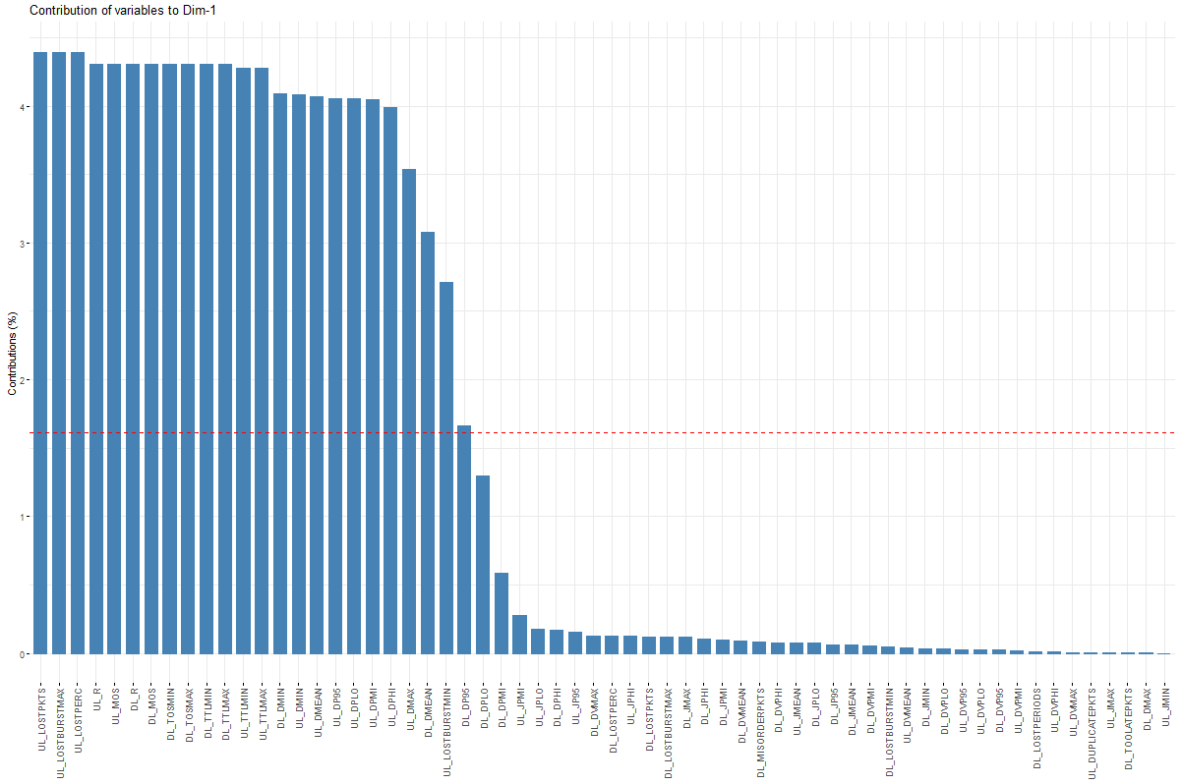


Nota. El gráfico muestra el resultado del Scree Plot o diagrama de sedimentación se utiliza para el análisis de componentes principales que ayuda a determinar cuántos componentes principales mantener. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

La siguiente secuencia de código muestra un histograma con el porcentaje de contribución de las variables referente al PCA y cuyo resultado se muestra en la Figura 24:

```
var<-get_pca_var(twamp_pca)
a<-fviz_contrib(twamp_pca, "var", axes=1, xtickslab.rt=90)
png(filename = "eigen-05.png", width = 1200, height = 800)
plot(a,main = "Variables percentage contribution of first
Principal Components")
dev.off()
```

Figura 24. *Participación en porcentaje de cada componente principal*



Nota. El gráfico muestra el porcentaje de participación de cada componente principal de las variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Continuamos con la agrupación considerando los componentes principales, debido a las limitaciones de cómputo sólo se ocupa el cálculo sobre una muestra de los valores. Un total de 4096 muestras. Finalmente consideramos ocho componentes principales en este ejercicio.

En análisis de agrupaciones, una de las variables más importantes es la definición del número de agrupaciones a considerar. Para evitar el uso de un criterio no subjetivo para la elección de este número, se utilizaron dos métodos en este trabajo Silhouette y GAP.

En este ejercicio, se considera el 10% del total de las muestras para optimizar el procesamiento.

Se realizó la evaluación del estadístico de Hopkins (Hopkins, 1954), el cual es una medida utilizada en análisis de datos en el contexto de la detección de agrupaciones (clustering) en un conjunto de datos. El objetivo primario es evaluar si el conjunto de datos presenta una estructura de agrupación o si los puntos están distribuidos de manera aleatoria en el espacio. La interpretación del estadístico de Hopkins

H **->** 0: Los datos están distribuidos de una manera menos dispersa que aleatoria

H -> 0.5: Indica que los datos están distribuidos aleatoriamente

H -> 1: Sugiere que los datos tienen una estructura de agrupación significativa.

La muestra que tenemos tiene valores de 1 o mayores por lo cual es un buen indicador

```
suppressWarnings(library(cluster))
suppressWarnings(library(hopkins))

twamp_pca<-prcomp(data_normalized,    center=FALSE,    scale.=FALSE,
rank. = 4)
results <- twamp_pca$x
#Se considera el 10% del total de las muestras para optimizar el
procesamiento
hopkins(data_normalized,4160)
```

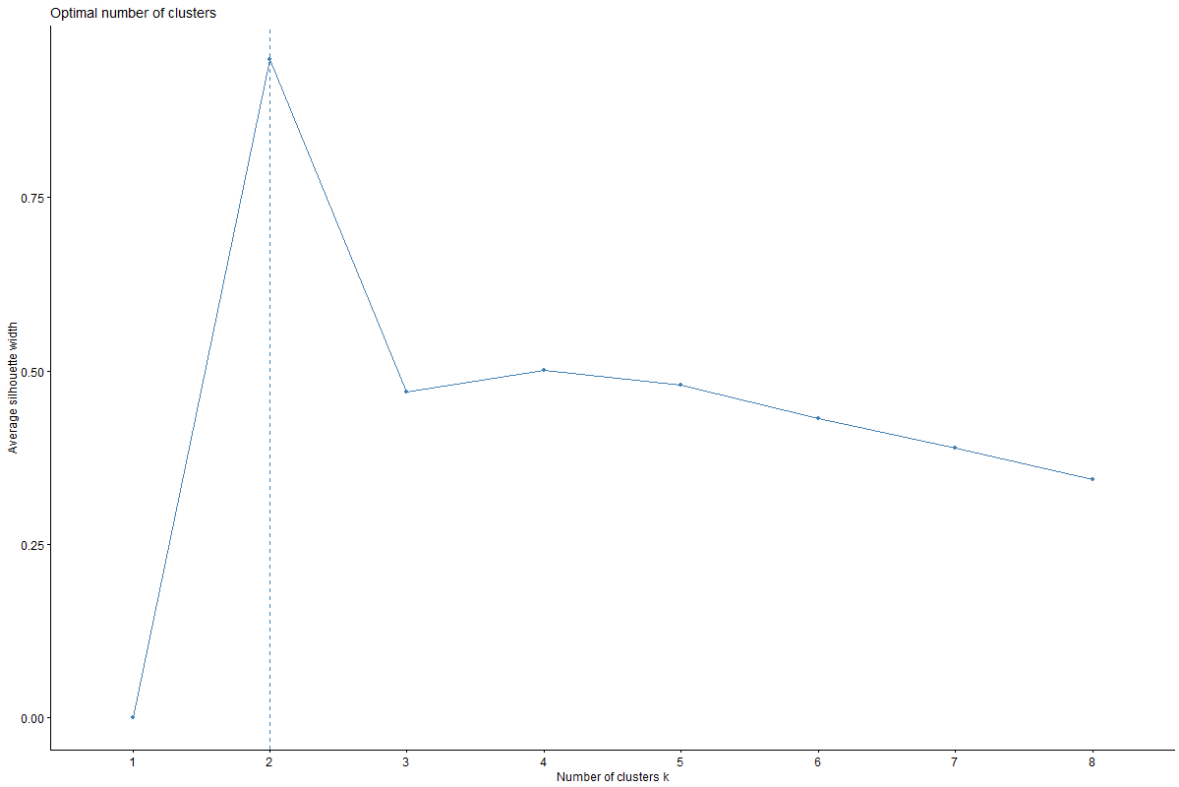
El valor obtenido como resultado de la función hopkins es de 1 por lo cual es un buen indicador.

K-means (Silhoutte)

```
#Se considera una muestra de 1000 muestras
twamp_pca<-prcomp(data_normalized,                center=FALSE,
scale.=FALSE, rank. = 4)
results <- twamp_pca$x
png(filename = "km-s-01.png", width = 1200, height = 800)
fviz_nbclust(results[1:1000,], FUNcluster=kmeans, k.max = 8)
dev.off()
```

El resultado nos indica que el número de clusters es de dos, tal como lo muestra la Figura 25, posiblemente derivado del tipo de variables, es decir generar agrupaciones tomando en consideración las variables de Uplink y en otro las variables de Downlink.

Figura 25. *Número de clusters considerando el método Silhouette*

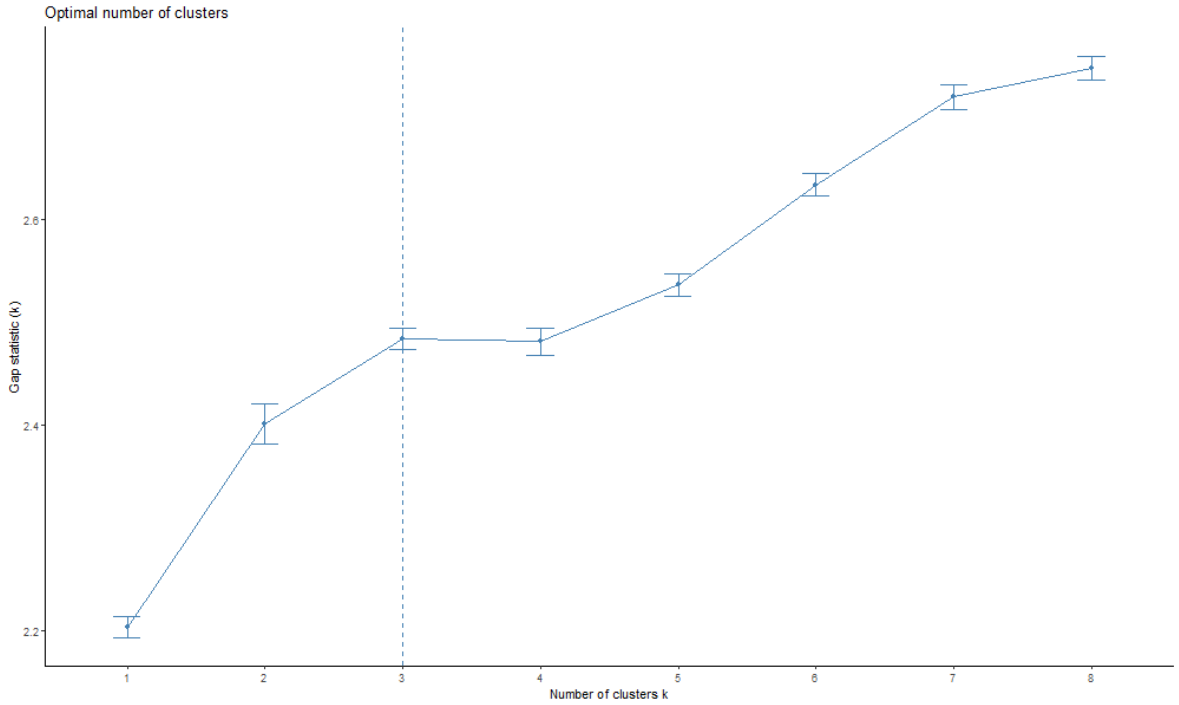


Nota. El gráfico muestra el número de clústers a considerar tomando en consideración el método Silhoutte de las variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Este resultado lo podemos contrastar con el método *K-Means* / GAP, el cual compara la dispersión dentro del conjunto de datos original con la dispersión que se puede esperar en conjuntos de datos generados aleatoriamente. Así, si se busca el punto en el que se maximiza la diferencias entre ambos valores se obtiene se puede estimar la cantidad óptima de clústeres. La Figura 26 muestra que el valor óptimo de las agrupaciones el cual es de tres.

```
#Se considera una muestra de 1000 muestras
png(filename = "km-g-01.png", width = 1000, height = 600)
fviz_nbclust(results[1:1000,],                    FUNcluster=kmeans,
method="gap_stat", k.max = 8)+ theme_classic()
dev.off()
```

Figura 26. *Número de clústers considerando el método GAP*



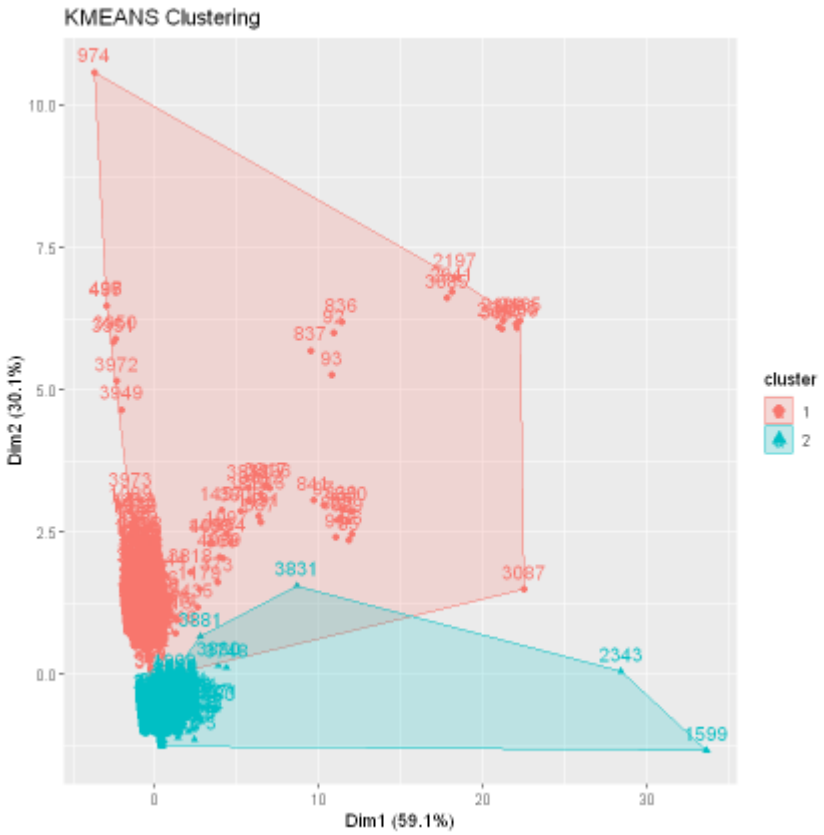
Nota. El gráfico muestra el número de agrupaciones a considerar tomando en consideración el método GAP de las variables del conjunto de datos. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

K-Means separa los datos en agrupaciones excluyentes, tratando de minimizar la distancia entre los puntos de cada grupo y aumentar la distancia entre los grupos. El siguiente paso es calcular la distancia euclidiana. Considero el uso k=2 para identificar el número de agrupaciones.

```
#Ejecución del cálculo de distancia euclidiana.
km_eu_d <- eclust(results[1:4160,],"kmeans",
    hc_metric="eucliden",
    k=2)
```

El resultado de esta instrucción se muestra en la Figura 27, se muestra los datos distribuidos en dos agrupaciones podemos asumir que los datos en este caso la interpretación es que son aquellas agrupaciones donde los valores de Latencia y Jitter son grandes y la otra agrupación corresponde a los valores mínimos de Latencia y Jitter.

Figura 27. *Agrupaciones tomando en consideración distancia euclidiana*

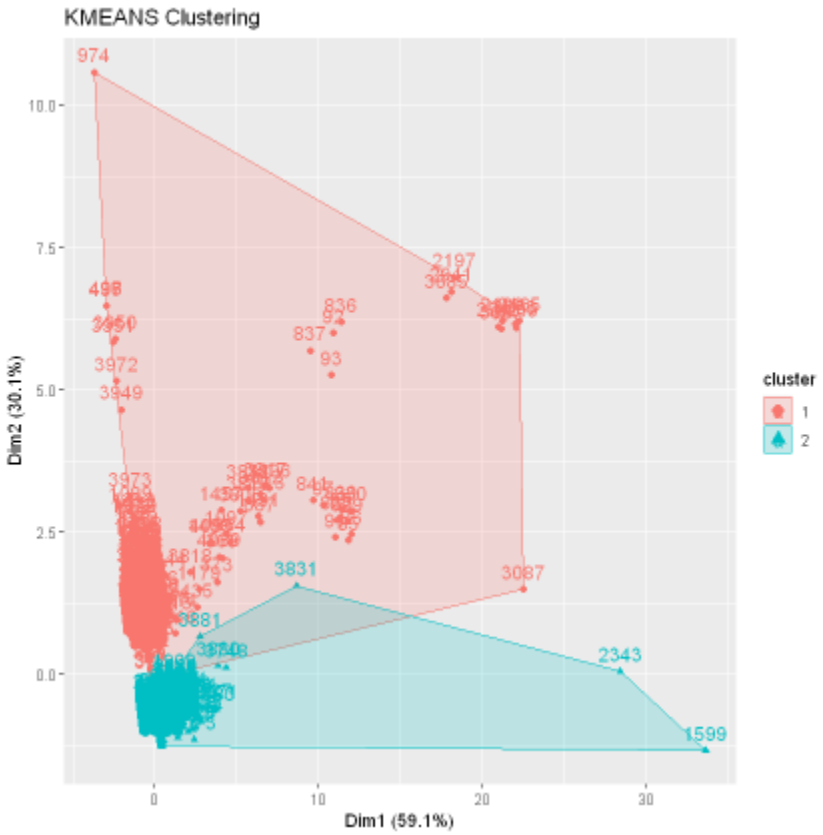


Nota. El gráfico muestra la agrupación de los datos en dos agrupaciones, tomando en consideración la distancia Euclidiana. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Se realiza el mismo ejercicio tomando como validación el uso de la distancia Manhattan para el procedimiento de Cálculo. En la Figura 28 se observa el resultado gráfico, se observa la distribución de datos en dos agrupaciones podemos asumir que los datos en este caso la interpretación es que son aquellas agrupaciones donde los valores de Latencia y Jitter son grandes y la otra agrupación corresponde a los valores mínimos de Latencia y Jitter.

```
#Ejecución del cálculo de distancia Manhattan.
km_manhattan<-eclust(results[1:4160,], "kmeans",
   hc_metric="manhattan",k=2)
```

Figura 28. *Agrupaciones tomando en consideración distancia Manhattan*



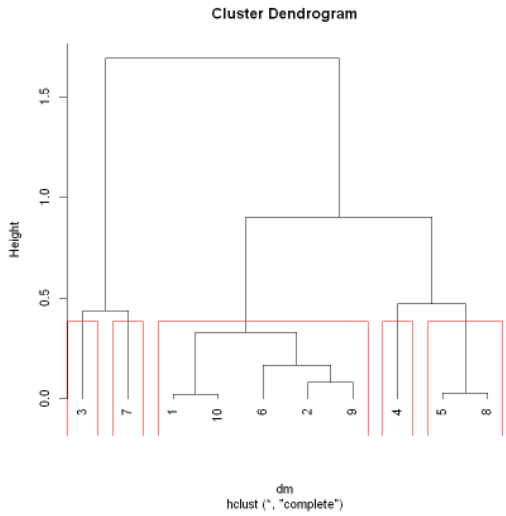
Nota. El gráfico muestra la agrupación de los datos en dos clústers, tomando en consideración la distancia Manhattan. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Se realiza dendograma tomando en consideración cinco agregaciones y el cual se muestra en la Figura 29.

```
dm<-dist(results[1:10,])
hc<-hclust(dm, method="complete")
plot(hc, hang=-1)
rect.hclust(hc, k=5, border="red")
```

Figura 29. *Dendograma, k=5*

```
dm<-dist(results[1:10,])
hc<-hclust(dm, method="complete")
plot(hc, hang=-1)
rect.hclust(hc, k=5, border="red")
```

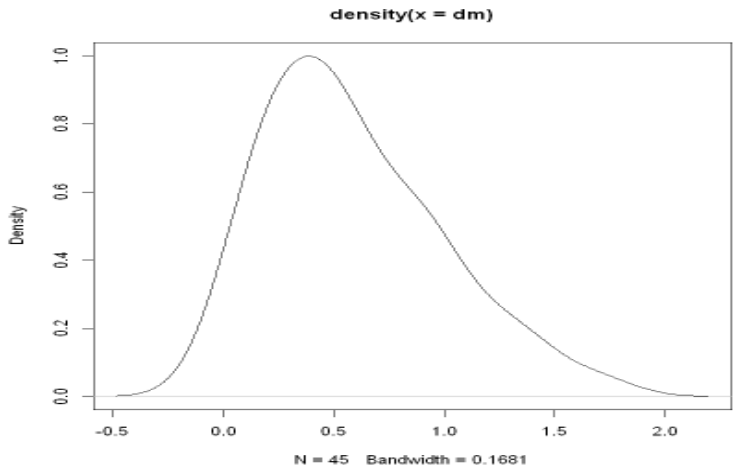


Nota. El gráfico muestra el dendograma con la estructura jerárquica de los datos. Lo cual permite identificar la similitud o distancia entre los componentes principales. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

La Figura 30 muestra el Diagrama de Densidad, que se obtiene al ejecutar el siguiente fragmento de código.

```
plot(density(dm))
```

Figura 30. *Diagrama de densidad*



Nota. El gráfico muestra la distribución de densidad. Elaboración propia, considerando el resultado de código empleado en el análisis del conjunto de datos.

Conclusiones

Conclusiones

A través del presente trabajo se ha mostrado los diferentes componentes para el análisis de eventos de monitoreo de calidad de una red de conmutación de paquetes, los resultados obtenidos se listan a continuación:

- El uso de agrupaciones por hora ayuda a tener un mejor entendimiento del comportamiento del tráfico en cada uno de los nodos, al considerar la agregación de valores máximos en el intervalo de tiempo. Al realizar la exploración de los datos, se identificó que una mayor granularidad en los datos (mediciones con intervalo en segundos) pueden mostrar picos y valles momentáneos que no representan el comportamiento real de la red, mientras que una agregación por hora suaviza estas variaciones, permitiendo identificar tendencias más significativas. Estas agrupaciones por hora, permite obtener valores más estables y representativos del comportamiento de la red y permite detectar patrones diarios, semanales y mensuales más claramente, esto facilita la correlación con eventos naturales, ofertas comerciales y operativos (horarios laborales, mantenimientos). Estas agrupaciones permiten también tener una reducción significativa en el volumen de datos a procesar y favorece tener una menor complejidad en el análisis de datos, pudiendo comparar mercados entre sí.
- Correlación entre métricas TWAMP (Delay, Packet Loss). Considerando las mediciones de Delay y Packet Loss, una correlación positiva media-alta, implica que: Un aumento en el Delay suele anteceder a eventos de Packet Loss. En un entorno de congestión de la red de conmutación de paquetes, ambas métricas se afectan de forma simultánea. Esta congestión puede incrementar la latencia en el tránsito de paquetes de datos.
- El análisis de los datos permite establecer en cualquiera de los mercados que sean analizados: El incremento en la demora (Delay) frecuentemente predicen incrementos en la pérdida de paquetes, la correlación de ambas métricas se intensifica en periodos de alta utilización de red. Por lo que, si

tenemos múltiples mercados con condiciones incrementales de demora, es factible que se encuentren escenarios de congestión de la red.

- Correlación entre métricas TWAMP (Delay, Jitter). Una correlación positiva entre Delay vs. Jitter, implica lo siguiente, Jitter tiende incrementar cuando la demora promedio se incrementa. Un incremento en la magnitud de Jitter implica que la red de conmutación de paquetes se encuentra en condiciones de degradación de servicio. Mientras más fuerte sea la correlación, implica que los enlaces de transporte llegan a su capacidad máxima disponible. El Jitter es un indicador temprano de problemas de demora (Delay) en la red. Sin embargo, un Jitter elevado no necesariamente implica pérdida de paquetes inmediata.
- Existen diferentes factores que influyen en la correlación de métricas de la red. La presencia de congestión en la red de conmutación de paquetes implica un incremento en la correlación entre todas las métricas. La congestión genera patrones que pueden ser predecibles de degradación. Esta condición afecta primero al Jitter, luego al Delay y en último lugar a la Pérdida de Paquetes.
- Otro factor es la capacidad del enlace de transporte, mientras más saturado se encuentre un enlace de transporte en la red de conmutación de paquetes, existe una correlación más fuerte en las variables. La correlación de variables varía también en el porcentaje de utilización.
- Un tercer factor que puede incrementar o disminuir la sensibilidad de la red, es el tipo de tráfico que se cursa en la red. Si es tráfico en tiempo real, este es más sensible a todas las métricas (Delay, Packet Loss y Jitter) y un patrón de degradación mayor; mientras que el tráfico de datos, las correlaciones pueden ser más débiles y las aplicaciones que usan dicha red de conmutación de paquetes tienen una mayor capacidad a la tolerancia de variaciones.
- En el monitoreo de estas variables tenemos algunos límites tempranos que ayudan a la comprensión de la red: El Jitter se considera predictor de los

problemas donde existe demora (Delay), los patrones de la demora deben ser analizados como predecesores de la pérdida de paquetes (Packet Loss).

- El análisis de estas variables permite identificar variaciones en corto plazo y el efecto cascada, que puede presentarse en la red de conmutación de paquetes afectando considerablemente la experiencia del usuario.
- El análisis de serie de tiempo puede ser extendido para la predicción de degradación de la calidad de la red basada en información histórica. Incluso se pueden identificar patrones estacionales de uso y congestión de la red.
- El uso de la metodología CRISP-DM ayudo a dar guía en la forma de trabajo que se siguió durante este trabajo teniendo claro cuales, son las fases esperadas y los resultados de las mismas. Cabe mencionar que si bien CRISP-DM parece similar a la metodología de gestión de proyectos, también puede asumirse como metodología ágil pues al ser iterativa puede ir desarrollando el producto mínimo viable durante todo el proyecto.
- Del análisis exploratorio de datos y el uso de algoritmos de aprendizaje no supervisado podemos identificar las variables y las agrupaciones que tienen mayor impacto en el conjunto de datos. El uso de agrupaciones por hora ayuda a tener un mejor entendimiento del comportamiento del tráfico en cada uno de los nodos, al considerar la agregación de valores máximos en el intervalo de tiempo. Al usar herramientas de minería de datos se modelo el tráfico por sitio lo cual puede ser automatizado para que se realice la revisión de tráfico de forma continua. Lo cual se deja para continuación de este trabajo. Cabe mencionar que para el conjunto de datos es necesario validar que no exista información faltante de lo contrario y posteriormente ajustar los rangos de valores pues las magnitudes generaron en primer lugar una interpretación incorrecta de los datos al tener variables medidas en microsegundos en lugar de milisegundos.
- El uso de aprendizaje no supervisado nos brinda una oportunidad para analizar datos donde no se cuenta con una variable objetivo definida. En el conjunto de datos utilizados existe una gran cantidad de información de la cual sólo tenemos los registros y no tenemos de principio una forma clara de

analizar los datos. Sin embargo, conforme fue avanzando este proyecto se identificaron herramientas como la reducción de dimensionalidad que permitirán explorar conjuntos de datos multivariados.

Referencias

[1] 3GPP, Release 15, TR 21.905: "*Vocabulary for 3GPP Specifications*" V8.

[2] 3GPP, Release 15, TS 23.214: "*Architecture enhancements for control and user plane separation of EPC nodes; Stage 2*". V8.

[3] 3GPP, Release 15, TS 29.281: "*General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)*". Versiones: 8 y 15.

[4] 3GPP, Release 15, TS 29.274: "*3GPP Evolved Packet System. Evolved GPRS Tunnelling Protocol for EPS (GTPv2)*". Versiones: 8 y 15.

[5] 3GPP, Release 15, TS 23.501:"*System Architecture for the 5G System*". Versión 8

[6] 3GPP, Release 15, TS 38.415: "*NG-RAN; PDU Session User Plane Protocol*". Versión 8.

[7] Álvarez Cuevas, Felipe, *et. al*., Voice synchronization in packet switching networks. Revista IEEE Networks, Tomo: 7, Volumen: 5 (Septiembre 1993), pp: 20-25.

[8] Álvarez Martínez, Antonio, Aplicación de Técnicas de Minería de Datos para mejorar el proceso de control de gestión en Entel. Universidad de Chile, Facultad de ciencias físicas y matemáticas, Tesis de Maestría, Santiago de Chile, Chile, 2012.

[9] Azevedo, Ana, Zantos, Manuel Filipe*. KDD, SEMMA and CRISP-DM: a parallel overview*. 2008.

[10] Baldi, Mario, Baralis, Elena, Risso, Fulvio., Data Mining Techniques for Effective *Flow-based Analysis of Multi-Gigabit Network Traffic*. 2010. Instituto Politécnico de Torino, Italia. Disponible en internet en: [https://core.ac.uk/download/pdf/11381392.pdf]

[11] Bolot, Jean-Chrystostome. *End-to-end packet delay and loss behavior in the Internet.* Procedimientos de *SIGCOMM '93*, pp: 289-298, 1993

[12] Chaudhari, S.S., Biradar, R.C*., Survey of Bandwith Estimation Techniques in Communication Networks*, Wireless Personal Communications, Volumen 83, Entrega 2, 2015

[13] Dibekulu Alem, Dawit, An Overview of Data Analysis and Interpretations in Reseach. International Journal of Academic Research in Education and Review. 2020. Volumen 8 (1), pp: 1-27

[14] European Telecommunication Standards Institute, *Speech and multimedia Transmission Quality (STQ); QoS parameters and test scenarios for assessing network capabilities in 5G performance measurements.* ETSI TR 103 702 V1.1.1 (2020-11). Francia

[15] Ekelin, Svante, Johnsson Andreas, Flinta, Christopher, *Scalability and Dimensioning of Network-Capacity Measurement System using Reflecting Servers*, Ericsson, Estocolmo, Suecia. 2015.

[16] GSM Association, *Network Experience Evolution to 5G*, Londres, Reino Unido.

[17] Graves, Alex, *Supervised Sequence Labelling with Recurrent Neural Networks*, Tesis de Doctorado, Facultad de Informática, Universidad Técnica de Munich, Alemania.

[18] Goodfellow, Ian, Bengio, Joshua, Courville, Aaron, Bach, Francis. *Deep Learning*. MIT Press Ltd, Estados Unidos, 2016.

[19] Hadidi, Hammed, Zhang Chaoyun, Patras, Paul. *Deep learning in mobile and wireless networking:a survey*. IEEE Communications, Surveys & Tutorials, 2019.

[20] Haykin, Simon, *Neural Networks, A comprehensive Foundation*. Universidad McMaster, Ontario, Canada. Prentice Hall. 2da edición, p. 842, 1999.

[21] Hewamalage,Hansika, Bergmeir,Christoph, Bandara, Kasun. *Recurrent Neural Networks for Time Series Forecasting: Current status and future directions.* International Journal of Forecasting, Volume 37, Issue 1, 2021, pp 388 - 427.

[22] Hopkins, Brian, Skellam, J.G. *A new Method for determining the Type of Distribution of Plant Individuals, Annals of Botany*, Volumen 18, Abril 1954, pp. 213 – 227.

[23] Ng KianSing, Liu Huan, Customer Retention via Data Mining, Artificial Intelligence Review, Kluwer Academic Publishers. Marzo, 2001

[24] Internet Engineering Task Force (IETF), RFC 768. *User Datagram Protocol.*

[25] Internet Engineering Task Force (IETF), RFC 791. *Internet Protocol.*

[26] Internet Engineering Task Force (IETF), RFC 2581. *TCP Congestion Control.*

[27] Internet Engineering Task Force (IETF), RFC 5841. *Packet Delay Variation Applicability Statement*.

[28] Internet Engineering Task Force (IETF), RFC 6038, Two-Way Active Measurement Protocol (TWAMP) reflects Octets and symmetrical size features.

[29] Internet Engineering Task Force (IETF), RFC 8913. *Two-Way Active Measurement Protocol (TWAMP) YANG Data Model*

[30] Johnsson, Andreas, Meirosu, Catalin.,*Towards Automatic Network Fault Localization in Real Time using Probabilistic Inference*. Ericsson Research,IFIP/IEEE IM2013 Workshop: 6th Intl Workshop on Distributed Autonomous Network Management Systems (DANMS), Kista, Suecia. 2013

[31] Kaiser, Henry F. (1960), *The application of electronic computers to factor Analysis. Educational and Psychological Measuremen*t, 20(1), Pag. 141-151. Estados Unidos, 1960.

[32] Kocak, C., Zaim, K. *Performance measurement of IP networks using Two-Way Active Measurement Protocol*, 2017, 8th International Conference on Information Technology (ICIT), Aman, Jordania, 2017. Pp: 249-254.

[33] Lawrence, E., Michailidis, G. and Nair, V.N. (2003), *Maximum likelihood estimation of internal network link delay distributions using multicast measurements, Proceedings of the Conference on Information Systems and Sciences*, Universidad Johns Hopkins, Estados Unidos, 2003.

[34] Lin Cai Sheng Zou Zhisheng Niu Lu Liu, Yu Cheng. *Deep learning based optimizationin wireless network*. IEEE ICC 2017 Ad-Hoc and Sensor Networking Symposium, 2017.

[35] Michailidis, G., Nair V.N. and Xi, B., *Fast least squares algorithms for estimating and monitoring network link losses based on active probing schemes*. Estados Unidos. 2005

[36] Mohsin Iftikhar, Muhammad Imran. Qazi Emad Ul, Haq Jasneet Kaur, M. Arif Khan. *Machine Learning techniques for 5g and beyond*. IEEE Access, 2021.

[37] Moon, Sue B., *Measurement and Analysis of End-to-End Delay and Loss in the Interne*t. Tesis de Doctorado. Universidad de Massachusetts, Amherst. Estados Unidos, 2000.

[38] Nugroho, Y.N., Harwahyu, R., Sari, R.F., Nikaein, N., Cheng, R., *Performance Evaluation of Anomaly Detection System on Portable LTE Telecommunications Networks Using OpenAirInterface and ELK*. International Journal of Technology, Volumen 14, pp: 549-560, Estados Unidos, 2023.

[39] Ouyang, Ye, Hu Mantian, Huet, Alexis, *et. al*, *Mining Over Air: Wireless Communication* Network Analytics, Springer International Publishing AG, Estados Unidos. 2018

[40] Papagiannaki, K., Taft, N. and Lakhina, A., *A distributed approach to measure traffic matrices*, ACMInternet Measurement Conference Proceedings, Taormina, Italia. 2004

[41] Penkova, T.G.(2017), *Principal Component Analysis and cluster analysis for evaluating the natural and anthropogenic territory safety. International* Conference on Knowledge Based and Intelligent Information and EngineeringSystems, KES2017, 6-8 Septiembre 2017, Marsella, Francia

[42] Ruiz Díaz de Salvioni, Armoa, Alcídes, La importancia de la Minería de Datos como una herramienta estratégica en las Empresas. Ciencia Latina Revista Científica Multidisciplinar, 7(1), 9267-9276. 2023

[43] Salehinejad, Hojjat, Sankar, Sharan, Barfett Joseph, Colak Errol, Valaee Shahrokh, *Recent Advances in Recurrent Neural Networks*, arXiv:1801.01078v3 [cs.NE], 2018.

[44] Salts, J.S., *CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps*, IEEE International Conference on Big Data (Big Data), Florida, Estados Unidos, 2021, pp. 2337-2344. 2021

[45] Shafique, U., Qaiser, H., *A comparative study of data mining process models (KDD, CRISP-DM and SESMA)*. International Journal of Innovation and Scientific Research, 12, pp. 217-222. 2019

[46] Shan Jaffry, *Cellular Traffic Prediction with Recurrent Neural Network*, Dongguan University of Technology, Dongguan, China. 2020

[47] Tebaldi, C. andWest,M, *Bayesian inference of network traffic using link count data (with discussion)*, Journal of the American Statistical Association, 93, 557-576. Estados Unidos, 1998

[48] Torres, Jordi. Introducción al aprendizaje por refuerzo profundo. Teoría y práctica en Python. Universidad Politécnica de Cataluña, Barcelona, España, 2021.

[49] Van der Maaten, Laurens, Hinton Geoffrey, Visualizing Data using t-SNE. 2008, Journal of Machine Learning Research, Volumen 9.

[50] Yang Oliver, Liu Huifang, Jiakun Yantai, Shu Feng, Yu Minfang, *Wireless traffic modeling and prediction using seasonal arima models*. IEICE Transactions on Communications, 2003.

[51] Zappone Alessio, Di Renzo Marco, Debbah Mérouane, "*Wireless networks design in the era of deep learning: Model-based, ai-based, or both?*" arXiv preprint arXiv:1902.02647, 2019.