Modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas para mejorar la disponibilidad del servicio de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador

Barrionuevo Gallardo, David Danilo

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios

Ing. Tapia León, Freddy Mauricio PhD.

10 de diciembre del 2022

COPYLEAKS

Tesis David Barrionuevo v4.1 - Solo Cuerpo.docx

Scanned on: 16:52 November 14, 2022 UTC

FREDDY
MAURICIO
TAPIA LEON

1.7%	14	24350
Overall Similarity Score	Results Found	Total Words in Text

Identical Words	59
Words with Minor Changes	5
Paraphrased Words	304
Omitted Words	2690

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Certificación

Certifico que el trabajo de titulación: "**Modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas para mejorar la disponibilidad del servicio de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador**" fue realizado por el señor **David Danilo Barrionuevo Gallardo;** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 10 de diciembre del 2022

Firmado electrónicamente por:
FREDDY MAURICIO TAPIA LEON

.....

PhD. Tapia León, Freddy Mauricio

Director

C.C.: 1714745690

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Responsabilidad de Autoría

Yo **Barrionuevo Gallardo**, **David Danilo**, con cédula de ciudadanía n° 1715577993, declaro que el contenido, ideas y criterios del trabajo de titulación: "**Modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas para mejorar la disponibilidad del servicio de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador**" es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 10 de diciembre del 2022

.....

Barrionuevo Gallardo, David Danilo

C.C.: 1715577993

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Autorización de Publicación

Yo **Barrionuevo Gallardo, David Danilo**, con cédula de ciudadanía n° 1715577993, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas para mejorar la disponibilidad del servicio de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 10 de diciembre del 2022

.....

Barrionuevo Gallardo, David Danilo

C.C.: 1715577993

Dedicatoria

Quiero dedicar mi trabajo de titulación a Dios, quien supo renovar mis fuerzas y darme la sabiduría para concluir con este proyecto.

A mi familia, por la paciencia y el apoyo que me han brindado para que yo pueda ver mis metas alcanzadas, todos mis triunfos se los dedico a ellos, son mi inspiración Vale, Sarahí e Isaac.

A mis padres, Mariano Barrionuevo e Inés Gallardo por sus infinitas bendiciones, oraciones y apoyo gracias por saberme guiar y siempre creer en mí.

A mis padrinos, Ángel Gallardo y Marcia Maya quienes desde el cielo son una inspiración en mi vida.

David Barrionuevo

Agradecimiento

A mi Padre Celestial, quien con su infinito amor y fortaleza me ha sabido guiar y levantar en cada instante de mi vida, toda la gloria y la honra sean para Él.

A mi querida Universidad de la Fuerzas Armadas ESPE, por abrirme las puertas por segunda vez y prepararme como un profesional digno de llevar en alto el nombre de esta noble institución.

A mi tutor de tesis Freddy Tapia, con su vasta experiencia, conocimiento y tiempo, supo direccionarme para plasmar todas mis ideas en este trabajo de titulación.

A mis maestros y compañeros de maestría, de quienes obtuve mucho conocimiento y amistad, el proceso no ha sido sencillo, pero gracias a su dedicación y apoyo he logrado culminar con éxito esta etapa de mi vida.

A mi familia y amigos, que son un soporte en cada momento, gracias por abrirme sus brazos y por estar en los buenos y no tan buenos momentos.

Índice de contenidos

Índice de tablas

Índice de figuras

Resumen

La evolución de las tecnologías de la información y el conocimiento han provocado un cambio disruptivo empresarial y nuevos modelos de negocio, que se basan principalmente en el acceso ágil y seguro a los datos. Por lo tanto, las empresas de telecomunicaciones se ven en la necesidad de brindar servicios de calidad, que garanticen la continuidad, eficiencia y privacidad en la transmisión de la información.

El presente trabajo aborda la problemática de una empresa nacional de telecomunicaciones, que ha detectado bajos índices de disponibilidad e intermitencia en el servicio de datos e internet de sus clientes corporativos, en parte generados por la inadecuada gestión de monitoreo de los enlaces, con alertas y acciones correctivas, que han provocado efectos a la compañía tales como: pérdida de clientes, multas, incremento en los costos de operación, incumplimiento de acuerdos de servicio, disminución de confianza e imagen corporativa, etc. En tal virtud, se propone un modelo predictivo basado en técnicas de minería de datos y aprendizaje automático, que identifique patrones en la información histórica de los equipos de red. Este modelo puede ser empleado para: predecir el comportamiento de los enlaces, detectar posibles cortes, generar alertas tempranas y apoyar proactivamente en la toma de decisiones de la compañía. Se ha utilizado la metodología CRISP-DM para el desarrollo y evaluación del modelo predictivo XG Boost (técnica evolucionada de los Árboles de decisión) en Python Jupyter Notebook. Los resultados son alentadores con una efectividad de hasta un 95.5% de predicción.

Palabras clave: telecomunicaciones, disponibilidad de red, minería de datos, modelo predictivo.

Abstract

The evolution of information and knowledge technologies have caused a disruptive business change and new business models, which are based on agile and secure access to data. Therefore, telecommunications companies need to provide quality services that guarantee continuity, efficiency, and privacy in the transmission of information.

This paper addresses the problem of a national telecommunications company, which has detected low rates of availability and intermittence in the data and internet service of its corporate clients, partly generated by the inadequate management of link monitoring, with alerts and corrective actions, which have caused effects to the company such as: loss of customers, fines, increase in operating costs, breach of service agreements, decrease in trust and corporate image, etc. In this virtue, we have proposed a predictive model based on data mining and machine learning techniques, which identifies patterns in the historical information of the network equipment. This model can be used to: predict link behavior, detect outages, generate early warnings, and proactively support the company's decision-making. The CRISP-DM methodology has been used for the development and evaluation of the predictive model XG Boost (evolved technique of Decision Trees) in Python Jupyter Notebook. The results are encouraging with an effectiveness of up to 95.5% prediction.

Keywords: telecommunications, network availability, data mining, predictive model.

Capítulo I: Introducción

El presente capítulo especifica el problema, la justificación, los objetivos y el alcance del presente proyecto de titulación.

Antecedentes

La información se ha convertido en uno de los activos más valiosos dentro de una organización; por lo tanto, es necesario una adecuada gestión de la información, que garantice: la disponibilidad, confidencialidad, privacidad, integridad y autenticidad de los datos (Suárez, 2019). Las telecomunicaciones, en conjunto con otros componentes de las Tecnologías de la Información y la Comunicación (TIC), permiten a las organizaciones, realizar una adecuada gestión de la información, ayudando a las instituciones a cumplir sus estrategias y metas planteadas (Barbosa & Miño, 2017).

En el Ecuador, debido a la importancia de la disponibilidad y el acceso a la información, se han creado políticas y normativas tanto públicas como privadas, que ayudan a garantizar la continuidad y calidad del servicio de telecomunicaciones (MINTEL, 2020).

Una de las principales empresas proveedoras de telecomunicaciones del país, con sede matriz en la ciudad de Quito y con cobertura a nivel nacional, establece con sus clientes corporativos Acuerdos de Nivel de Servicio o Service Level Agreement (SLA por sus siglas en inglés), que comprometen a la institución a realizar una adecuada gestión de disponibilidad del servicio de telecomunicaciones, que garantice el acceso y transmisión de la información de sus clientes.

En los últimos años, la empresa proveedora de telecomunicaciones ha sufrido constantes problemas de disponibilidad por cortes e intermitencia de los servicios de internet y datos, que provocan incumplimientos de los SLA establecidos con sus clientes corporativos, esto ha generado: multas, terminaciones de contratos, pérdida de confiabilidad y disminución de la imagen corporativa. En la actualidad, la empresa proveedora de telecomunicaciones ha

implementado un sistema de monitoreo para sus principales clientes, el mismo que almacena la información monitoreada y envía alertas basadas en umbrales estáticos, que se activan una vez superados los límites configurados o posterior a un corte del servicio, permitiendo al personal operativo tomar acciones correctivas, mismas que no son suficientes para una adecuada gestión de disponibilidad.

En un análisis de los datos almacenados por el sistema de monitoreo, realizado y socializado por los expertos de la empresa proveedora de telecomunicaciones, se ha detectado que, el comportamiento de los sensores de los dispositivos de red tiene patrones y tendencias de saturación previo a los errores o pérdidas de servicio. Debido a la cantidad de datos de monitoreo recolectados, el análisis únicamente se puede realizar en pocos dispositivos y bajo demanda; por lo tanto, se requiere de un análisis masivo y automático, para detectar tendencias que alerten de forma temprana posibles cortes del servicio de transmisión de datos y permitan al personal operativo de la empresa tomar acciones proactivas.

Problema

Actualmente, las empresas proveedoras de telecomunicaciones del Ecuador están obligadas a cumplir políticas y normativas que garanticen a sus clientes la prestación de servicios de internet y datos con calidad, accesibilidad, continuidad y seguridad (ARCOTEL, 2021). Por esta razón, las empresas proveedoras de servicios de telecomunicaciones, establecen contractualmente con sus clientes corporativos SLA, con altos índices de disponibilidad y continuidad de los servicios.

En reuniones mantenidas con el personal técnico y operativo de la empresa de telecomunicaciones, se ha realizado un análisis de los datos obtenidos entre los años 2019 y 2020 por su software de monitoreo NMS (GrupoCONTEXT S.A., 2022), los cuales presentan un porcentaje de disponibilidad promedio entre el 98.7% y 99.3% para los servicios de internet y datos de los clientes corporativos de la compañía. Considerando que en términos contractuales

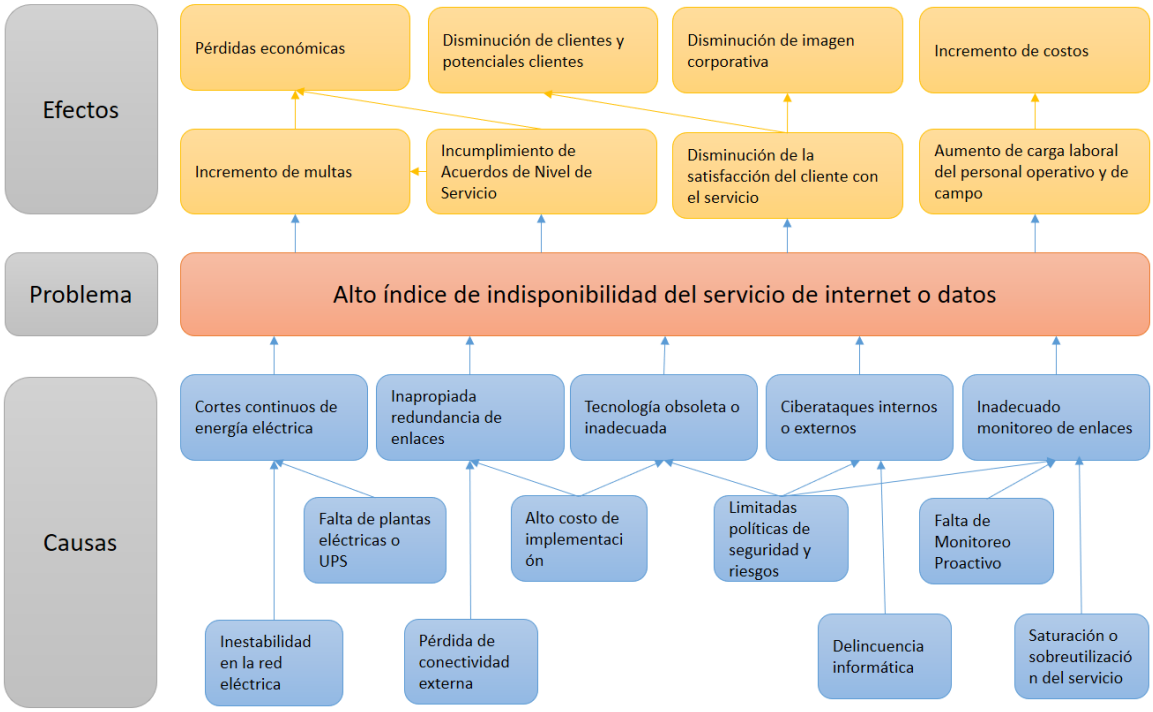
la empresa ofrece un 99.6% de disponibilidad, se puede determinar que existe un alto índice de indisponibilidad del servicio debido a cortes e intermitencias de los enlaces. Esto provoca las siguientes afectaciones: incumplimiento de los SLA; aumento de quejas y reclamos por parte de los clientes (ARCOTEL - DEAC, 2019); incremento de la carga laboral del personal operativo y de campo. Y en los casos de mayor criticidad, se presentan: pérdidas económicas, por multas y devoluciones; pérdida de clientes, actuales y potenciales; aumento en los costos, de operación y mantenimiento.

En vista de la problemática detectada, los directivos de la empresa proveedora de telecomunicaciones, han tomado ciertas acciones, que ayudan a mitigar algunas de las causas que provocan la indisponibilidad de los enlaces. Sin embargo, una de las causas que necesitan ser abordadas, es la inadecuada monitorización de los dispositivos de red, ya que su sistema de monitoreo actual detecta y envía notificaciones posteriores a la presencia de un evento o incidente, estableciendo políticas y acciones correctivas por parte del personal operativo y de campo de la compañía, que no son suficientes para una adecuada gestión de disponibilidad y cumplimiento de los acuerdos de niveles de servicio.

En reuniones realizadas con el personal operativo y administrativo de la empresa, se ha generado un árbol de problema. La Figura 1, permite visualizar y comprender el problema detectado con los clientes corporativos de la empresa de telecomunicaciones, sus efectos y las principales causas que lo provocan. Al existir varias causas encontradas para el problema de indisponibilidad, es necesario especificar que, la presente propuesta se va a enfocar en el inadecuado monitoreo de los enlaces.

Figura 1

Árbol de problema de la investigación



Nota. Presenta los efectos y causas del problema del alto índice de indisponibilidad del servicio de internet y datos.

Justificación

Una adecuada medición en la industria de telecomunicaciones es importante para tomar decisiones tanto del personal operativo como comercial, con la finalidad de garantizar la disponibilidad de los servicios. Sin embargo, esta medición es un tema complejo debido a la gran cantidad de posibilidades de medida dentro de este entorno. En efecto, cada tarjeta, puerto o sensor de cualquier dispositivo de comunicación proporciona una serie de oportunidades de medición que generan gran cantidad de datos, haciendo complejo su análisis dentro del entorno mencionado anteriormente (Lartey, 2017).

Si bien la implementación de un sistema de monitoreo tradicional permite alertar al operador para resolver problemas en la infraestructura de la red, es necesario identificar de

manera proactiva estos problemas, para resolverlos antes de que produzcan efectos notorios para los clientes. Sin embargo, se dedica tiempo a resolver un problema que aún no es real para los usuarios finales y la empresa necesita identificar y asignar los recursos adecuados y automatizados para llevar a cabo dichas tareas (Lartey, 2017).

Considerando los efectos negativos para la empresa proveedora de telecomunicaciones; en la que se basa el presente estudio; tales como: la pérdida de clientes, las multas establecidas, la pérdida de imagen y confianza corporativa y el incremento en los costos de operación, causados por la indisponibilidad del servicio corporativo de internet y datos hacia sus clientes, y con la visión de mejorar la calidad y estabilidad de los enlaces, se justifica la incorporación del modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas, ya que traerá a la compañía de telecomunicaciones una mejor gestión de disponibilidad, al detectar con anterioridad posibles cortes e intermitencias de conexión, esto podría permitir al personal operativo tomar acciones proactivas, minimizando los riesgos y mitigando el inadecuado monitoreo de los enlaces, que representa una de las principales causas de la inestabilidad del servicio.

Objetivos

Objetivo general

Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas, mediante técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos, con la finalidad de mejorar los índices de disponibilidad del servicio corporativo de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador.

Objetivos específicos

- **OE1**: Determinar las técnicas predictivas supervisadas de minería de datos idóneas, que permitan pronosticar y establecer umbrales dinámicos de sensores de red, a través de una revisión preliminar de literatura especializada.
- **OE2**: Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red, mediante el uso de técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos.
- **OE3**: Evaluar el modelo propuesto en una empresa proveedora de telecomunicaciones del Ecuador, por medio de un caso de estudio Benchmark, el cual permita obtener niveles de rendimiento y disponibilidad de los servicios.

Alcance

La Tabla 1, presenta las preguntas de investigación obtenidas del planteamiento de los objetivos específicos.

Tabla 1

Preguntas de investigación por objetivo específico

Objetivos específicos	Preguntas de investigación
OE1: Determinar las técnicas predictivas supervisadas de minería de datos idóneas, que permitan pronosticar y establecer umbrales dinámicos de sensores de red, a través de una revisión preliminar de literatura especializada.	**OE1-RQ1**: ¿Qué soluciones presentan los estudios realizados sobre técnicas de minería de datos de predicción de umbrales dinámicos de sensores de red? **OE1-RQ2**: ¿Determinar cuál es la técnica de minería de datos más adecuada para la predicción de umbrales dinámicos de sensores de red?
OE2: Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red, mediante el uso de técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos.	**OE2-RQ1**: ¿Qué modelo de minería de datos predice umbrales dinámicos de sensores de red, basados en tendencias de comportamiento de dispositivos? **OE2-RQ2**: ¿Cuáles son las aplicaciones y fuentes de datos de la empresa proveedora de telecomunicaciones, que almacenan la información histórica del monitoreo de los sensores de red?
OE3: Evaluar el modelo propuesto en una empresa proveedora de telecomunicaciones del Ecuador, por medio de un caso de estudio Benchmark, el cual permita obtener niveles de rendimiento y disponibilidad de los servicios.	**OE3-RQ1**: ¿Cuáles son los indicadores que permiten medir la mejora de disponibilidad del servicio de internet y datos? **OE3-RQ2**: ¿Cuál es el nivel de confianza de los resultados obtenidos?

Nota. Presenta la relación entre los objetivos y las preguntas de investigación.

Hipótesis

El desarrollo de un modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas permitirá la mejora de los índices de disponibilidad del servicio corporativo de acceso a Internet y datos de una empresa proveedora de telecomunicaciones del Ecuador.

Categorización de variables de la hipótesis

Variable Dependiente (VD). Índices de disponibilidad del servicio corporativo de acceso a Internet y datos.

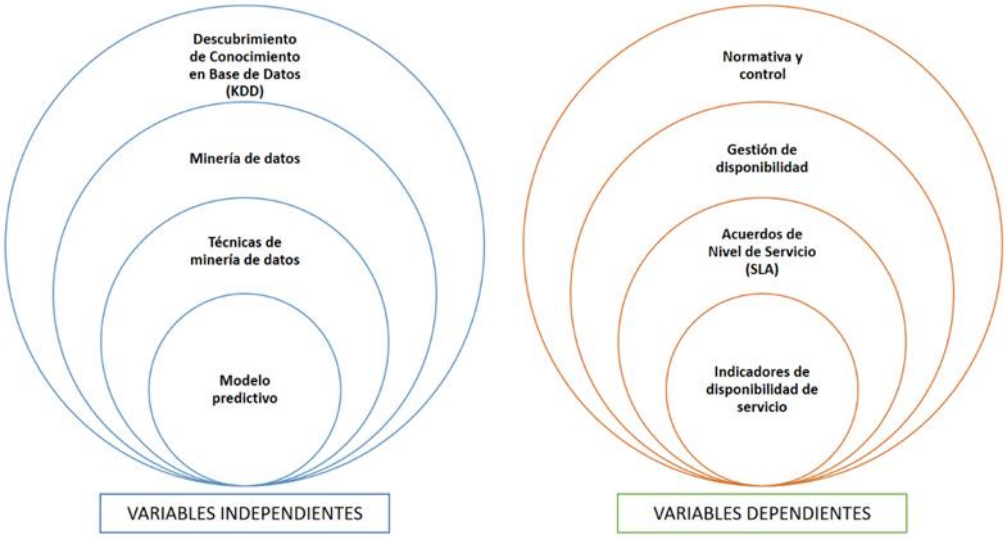
Variable Independiente (VI). Modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas.

Capítulo II: Marco teórico

La Figura 2, permite visualizar de forma jerárquica las diferentes categorías y subcategorías de los conceptos que se van a investigar para argumentar y sustentar el presente proyecto de investigación.

Figura 2

Categorías de estudio de las variables de la hipótesis



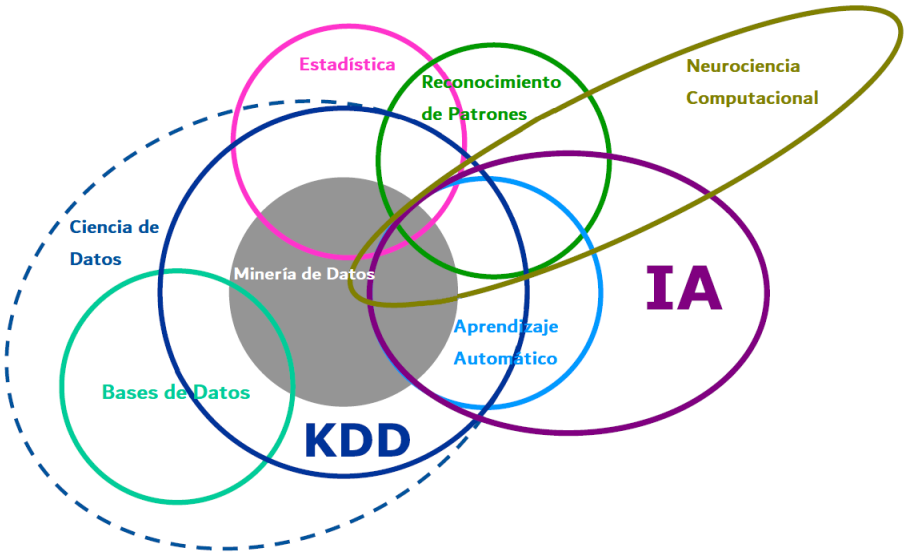
Nota. Presenta las categorías de estudio de las variables dependientes e independientes de la hipótesis planteada.

Descubrimiento de Conocimiento en Bases de datos (KDD)

El Descubrimiento de Conocimiento en Bases de datos o Knowledge Discovery in Databases (KDD por sus siglas en inglés) se define como el proceso de identificar patrones significativos en los datos, los cuales deben ser: válidos, novedosos, útiles y comprensibles (García Gutiérrez, 2016). KDD es el nombre técnico con que se denomina al proceso global de extracción de conocimiento. La Figura 3, muestra como KDD forma parte de los campos multidisciplinares del Big Data Analytics; también, presenta los campos de estudio que contiene y sus intersecciones.

Figura 3

Campos multidisciplinares del Big Data Analytics



Nota. Presenta los campos multidisciplinares que forman parte y contiene KDD. Adaptado de *Machine Learning*, por B. Wujek, 2014.

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido ponderado a nivel de la investigación y de la industria como un tópico clave en la gestión de los sistemas de base de datos; a nivel empresarial se constituye en una oportunidad para incrementar sus ganancias (Timarán, 2009). Fayyad, Piatetsky-Shapiro y Smith, lo definen como "*El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos*" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

La explotación de información puede definirse como una disciplina que provee de herramientas a la Inteligencia de Negocios o Business Intelligence (BI por sus siglas en inglés) para transformar la información en conocimiento, a través de la búsqueda de patrones y regularidades en grandes volúmenes de datos (Britos, 2008). Esta transformación responde a

un conjunto de tareas relacionadas lógicamente, las cuales agregan valor a los datos provistos como entrada (García, 2004).

KDD es básicamente un proceso automático que combina tanto el descubrimiento como el análisis. A partir de los datos se extraen patrones en forma de funciones o reglas que pueden ser analizados posteriormente, lo cual implica: preprocesar los datos, hacer minería de datos y presentar los resultados (Han & Kamber, 2001). KDD tiene diversos dominios de aplicación, como, por ejemplo: en el diagnóstico médico, para realizar relaciones implícitas entre síntomas y enfermedades; en las ciencias sociales, para obtener perfiles de estudiantes con relación a sus características socioeconómicas; en el marketing, para establecer patrones de compra de los clientes; entre muchos otros.

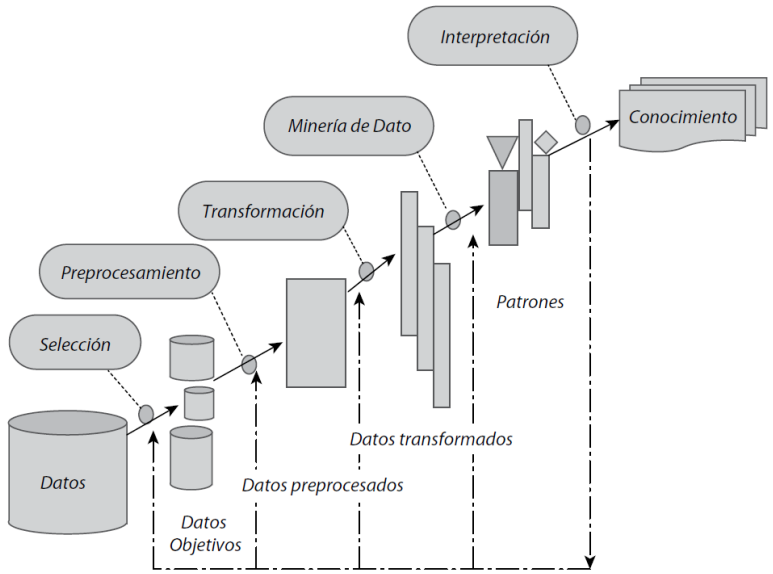
Para Valcárcel, las tareas comunes en el proceso de KDD son: inducción de reglas, clasificación y segmentación (Clustering), reconocimiento de patrones, modelado predictivo, y detección de dependencias. Los datos describen hechos y se presentan organizados en bases de datos, y los patrones son expresiones (modelo aplicable) que describen un subconjunto de esos datos (Valcárcel, 2004).

Etapas del proceso KDD

La Figura 4, muestra el proceso interactivo e iterativo de KDD propuesto por Timarán Pereira en 2016, este involucra pasos en los cuales se requieren la intervención del usuario en la toma de decisiones y abarca las siguientes etapas:

Figura 4

Etapas del proceso KDD



Nota. Tomado de *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, por Pereira, et al., 2016, Ediciones Universidad Cooperativa de Colombia.

Etapa de Selección. Consiste en elegir un conjunto de datos objetivo para realizar el proceso de descubrimiento, estos datos deben estar alineados a las metas del negocio o proyecto (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Etapa de Preprocesamiento (limpieza). En esta etapa se analiza la calidad de los datos. Se eliminan datos ruidosos (valores extremos) y se aplican estrategias para reemplazar datos desconocidos (perdidos y vacíos) con valores por defecto o usando métricas estadísticas como media, moda, mínimo y máximo, entre otras (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Etapa de Transformación (reducción). En esta etapa, dependiendo del objetivo del proceso, se buscan características útiles que permitan representar los datos. Para ello se emplean métodos de reducción de dimensiones, con la finalidad de disminuir el número

efectivo de variables a ser consideradas o para identificar representaciones que no varíen los datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Los métodos de reducción de dimensiones simplifican una tabla de base de datos de forma vertical (registros o filas) u horizontal (columnas o atributos), utilizando técnicas de reducción, tales como: agregaciones, compresión de datos, histogramas, segmentación, discretización (convertir una variable continua en una discreta), muestreo, entre otras (Han & Kamber, 2001).

Etapa de Minería de datos. Esta etapa tiene como objetivo la búsqueda y descubrimiento de patrones inesperados y de interés en los datos, aplicando para ello procesos de descubrimiento tales como: segmentación (Clustering) (Zhang, Ramakrishnan, & Livny, 1996), clasificación (Wang, Iyer, & Scott, 1998), asociaciones (Srikant & Agrawal, 1996), patrones secuenciales (Agrawal & Srikant, 1995), entre otros.

Etapa de Interpretación (evaluación). En esta etapa final de evaluación de datos, se interpretan los patrones antes descubiertos, y se podrían producir iteraciones con las etapas anteriores. Dentro de ella se incluye la visualización y análisis de los patrones extraídos, la exclusión de datos redundantes o irrelevantes, y la traducción de los patrones útiles a fin de que puedan ser comprendidos por los usuarios. Finalmente, se consolida el conocimiento alcanzado, a fin de incorporarlo en otro sistema, documentarlo o reportarlo para posteriores acciones, también para resolver potenciales conflictos con conocimiento previamente descubierto (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Minería de Datos

La Minería de Datos (MD) o Data Mining, también llamada exploración de datos es el núcleo del proceso KDD y es el campo de la estadística y las ciencias de la computación que intenta descubrir patrones y encontrar correlaciones en grandes volúmenes de información (Maimon & Rokach, 2010).

Molina (2000) define a la Minería de Datos como un proceso para extraer conocimiento de bases de datos, siendo su objetivo descubrir situaciones de especial interés, patrones o secuencias en los datos. Se trata de un término genérico que engloba los resultados de procesos de investigación, así como de herramientas empleadas para extraer información útil de grandes bases de datos. A pesar de ser considerada como parte del proceso KDD, en gran parte de la literatura ambos conceptos (MD y KDD) se identifican como iguales. De manera más específica, el término Minería de Datos es comúnmente empleado en el área estadística, por analistas de datos y administradores de sistemas informáticos, mientras que KDD es usado más por especialistas en Inteligencia Artificial. A continuación, varias definiciones de Minería de Datos:

- "*MD es el proceso de extracción y refinamiento de conocimiento útil desde grandes bases de datos*" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).
- "*MD es el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones*" (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1997).
- "*MD es el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos*" (Thuraisingham, 1999).
- "*MD es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos*" (Berry & Linoff, 1997).

Para Valcárcel (2004), la MD es una etapa dentro de todo el proceso de descubrimiento de conocimiento o KDD, el cual busca obtener patrones o modelos a partir de datos recabados. Según la autora, los algoritmos de MD tienen en general 3 componentes:

1. **Modelo**: Contiene parámetros a ser definidos a partir de los datos de entrada.
2. **Criterio de preferencia**: Se emplea para comparar modelos alternativos.
3. **Algoritmo de búsqueda**: Cualquier programa de inteligencia artificial (IA).

Las principales diferencias entre algoritmos de MD se encuentran en el modelo de representación escogido y su función (objetivo perseguido).

Herramientas de Minería de Datos

En la actualidad existen varias alternativas de herramientas y plataformas de software para tareas de Minería de Datos, cada una con diferentes características, opciones y licencias de uso. La Figura 5, presenta las herramientas posicionadas en el Cuadrante Mágico de Gartner para plataformas de Ciencia de Datos y Aprendizaje Automático (Machine Learning).

Figura 5

Plataformas de Data Science y ML en el Cuadrante Mágico de Gartner

Nota. Tomado de *Magic Quadrant Gartner*, por Gartner Inc., 2022.

Se debe considerar que, Gartner en su Cuadrante Mágico incluye herramientas de proveedores tecnológicos que en su mayoría son de uso comercial, sin embargo, existen otras herramientas y lenguajes que por sus características de propósito educativo y/o colaborativo son excluidos.

La Tabla 2, presenta un resumen comparativo de las principales herramientas y lenguajes de software de aprendizaje automático para tareas de minería de datos. Se han seleccionado plataformas de los proveedores que muestra el Cuadrante de Gartner, junto a herramientas de software libre de carácter: comercial, investigativo y educacional.

Tabla 2

Comparación de herramientas de Minería de Datos

Herramienta	Características	Ventajas	Desventajas
Weka	Software libre GPL. Universidad de Waikato. Lenguaje Java.	Técnicas estándares de MD. Interfaz Gráfica. Integración a diferentes fuentes. Funciones de ML con algoritmos de predicción. Pionera en enseñanza e investigación.	Baja potencia en análisis de clústeres. Problemas con grandes volúmenes de datos. Problemas con carga en memoria.
RapidMiner Studio	Software libre y comercial, diferentes versiones de pago. Licencia educativa (1 año). Lenguaje Java.	Software dedicado a Data Science de extremo a extremo. Gráficos potentes. GUI intuitiva. Incluye casos de ejemplo interactivos para su aprendizaje. 500+ operadores. Conectores a diferentes fuentes de datos. Integración con Weka y R. Usado en medianas y grandes empresas. Curva de aprendizaje corto.	Las integraciones a terceros no son excelentes y requieren mejora. Problemas de carga y rendimiento con grandes volúmenes de información.
SAS	Licencia comercial, diferentes versiones de pago. SAS Institute. Lenguaje SAS.	Tecnología de punta en predicciones. GUI interactiva e intuitiva. Alta Escalabilidad. Potente para grandes empresas. Gráficos y tablas potentes. Curva de aprendizaje medio.	Alto costo de licenciamiento. Lenguaje propietario. Aprendizaje propietario. Implementación inicial compleja.
R y R Studio	Software libre. Colaborativo. Lenguaje Java, C++, Javascript	Contiene librerías para Data Science. GUI basada en IDE. Gráficos e Informes intuitivos. Curva de aprendizaje medio.	Ciertas librerías desarrolladas no son estándar. Instrucciones por línea de comandos. Conexión y gestión de paquetes no sencilla.

Herramienta	Características	Ventajas	Desventajas
Anaconda Distribution (Python)	Software libre. Lenguaje Python Suite de código abierto para el desarrollo de Ciencia de datos	Contiene librerías especializadas y entornos de desarrollo para Data Science (Anaconda Navigator, Jupyter, JupyterLab, Spyder y R Studio). Posee paquetes de Big Data para carga y análisis de grandes volúmenes de información (Dask, Numba). GUI intuitiva y sencilla que permite organizar el código por secciones. Contiene paquetes para gráficos. Conectores a diferentes fuentes de datos. Software multiplataforma para varios sistemas operativos. Usado en medianas y grandes empresas. Curva de aprendizaje medio.	Instrucciones por línea de comandos. Dificultad de instalación en ciertas librerías por dependencias.
IBM SPSS Modeler	Licencia comercial. IBM. Licencia educativa (1 mes).	GUI Intuitiva y visualización del proceso. Líder en el análisis predictivo. Módulos para manejo de grandes volúmenes de datos. Gráficos potentes. Incluye casos de ejemplo interactivos para su aprendizaje. Curva de aprendizaje corto.	Alto costo de licenciamiento. Sintaxis limitada.
Orange	Software libre GPL. Universidad de Liubliana. Lenguaje C++ y Python.	Aplicaciones de análisis de datos. Carácterísticas de ML. GUI interactiva Documentación extensa. Curva de aprendizaje corto.	Soporte limitado. Usado por empresas pequeñas.
Knime	Software libre GPL. Universidad de Constanza. Lenguaje Java Uso en el sector farmaceutico y financiero.	1000+ módulos y paquetes. Análisis de datos integrativo. Integración de procedimientos de ML y DM. SW de DM orientado al flujo de datos. Módulos de BI. Se adapta a otros lenguajes como Python. Curva de aprendizaje corto.	Para propósitos específicos se debe incluir código de programación. Implementación no intuitiva.

Nota. Adaptado de *Análisis comparativo de herramientas Open Source para Data Mining sobre datos públicos del Ministerio de Educación de la República del Ecuador*, por S. Páez, 2019.

Posterior a la comparación y análisis realizado, se toma la decisión de usar la herramienta **Anaconda Distribution (Python)** para la generación del modelo predictivo del presente trabajo de titulación. A continuación, se detallan las principales razones para su selección:

- Lenguaje de programación con gran cantidad de librerías especializadas y entornos de desarrollo dedicado a Data Science.
- Contiene paquetes de carga y análisis de grandes volúmenes de información.
- Cuenta con operadores y algoritmos nativos y externos de predicción con ML.

- Posee conectores a diferentes fuentes de información (bases de datos, archivos planos, hojas de cálculo, etc.).
- GUI intuitiva y sencilla que permite ordenar el código y realizar el seguimiento de ejecución del proceso.
- Curva de aprendizaje medio, con documentación extensa y con su propia comunidad.
- Software Libre con Licencia BSD, utilizada en medianas y grandes empresas en diversos ámbitos y aplicaciones.

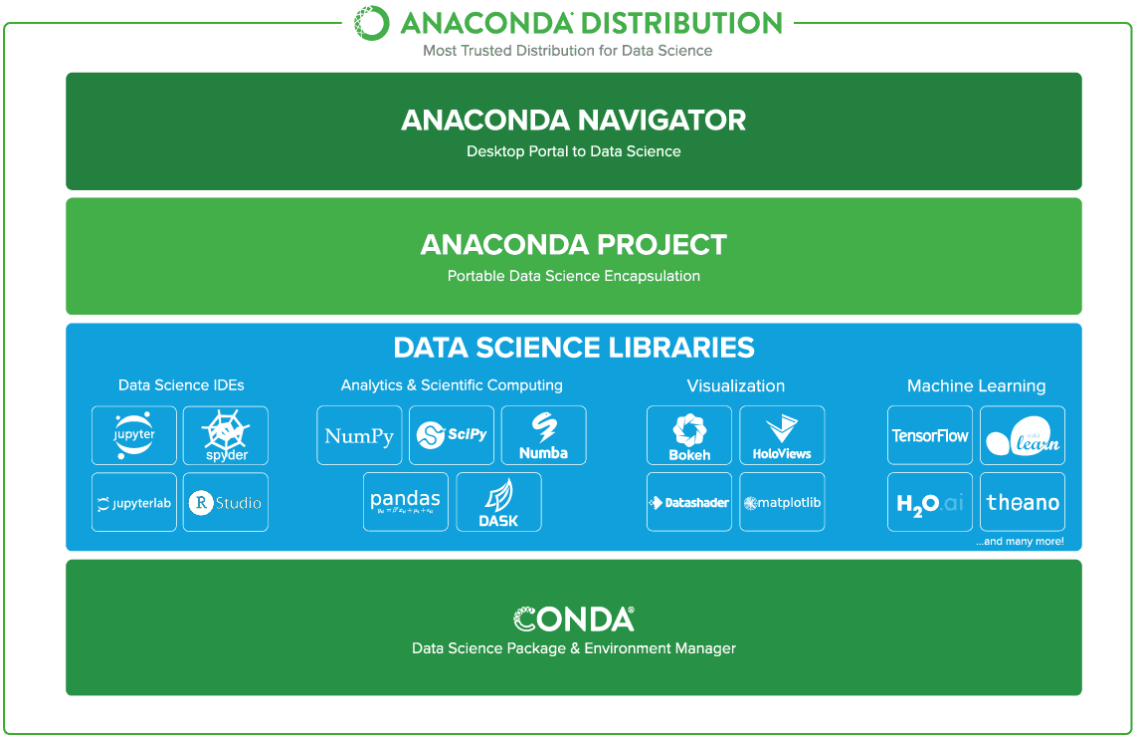
Anaconda Distribution (Python). Es una plataforma de software desarrollada en código abierto Python, cuenta con múltiples aplicaciones, librerías y paquetes especializados para Ciencia de datos. Se puede instalar y compilar en la mayoría de los sistemas operativos y arquitecturas. Entre los principales ámbitos de solución de esta herramienta, se tiene: Minería de datos, Aprendizaje Automático (Machine Learning), Redes Neuronales, Análisis predictivos, Visualización y exploración de datos (Anaconda Inc., 2022).

La Figura 6, muestra la arquitectura de la plataforma Anaconda Distribution que actúa como un gestor de paquetes, agrupándolos en 4 soluciones tecnológicas: Anaconda Navigator, Anaconda Project, Librerías de Ciencia de datos y Conda (AB Internet Networks, 2017). Para el presente proyecto se van a utilizar los siguientes componentes de la plataforma Anaconda:

- **Conda**: Para la gestión de los paquetes de Ciencia de Datos y Machine Learning.
- **IDE Jupyter Notebook**: Para la escritura y estructuración del código Python.
- **Pandas**: Para el manejo y análisis de estructuras de datos.
- **DASK**: Para el manejo de grandes volúmenes de información en paralelo.
- **XGBoost**: Para generar el modelo de Machine Learning basado en árboles de decisiones.
- **Matplotlib**: Librería para generación de gráficos.

Figura 6

Arquitectura Anaconda Distribution (Python)



Nota. Presenta los principales componentes de Anaconda. Tomado de *Desde Linux - Anaconda Distribution*, por AB Internet Networks, 2017.

Técnicas de Minería de Datos

Las técnicas de Minería de Datos dentro de la etapa del proceso KDD, tienen como objetivo la obtención de patrones o modelos en base al tratamiento de datos recopilados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

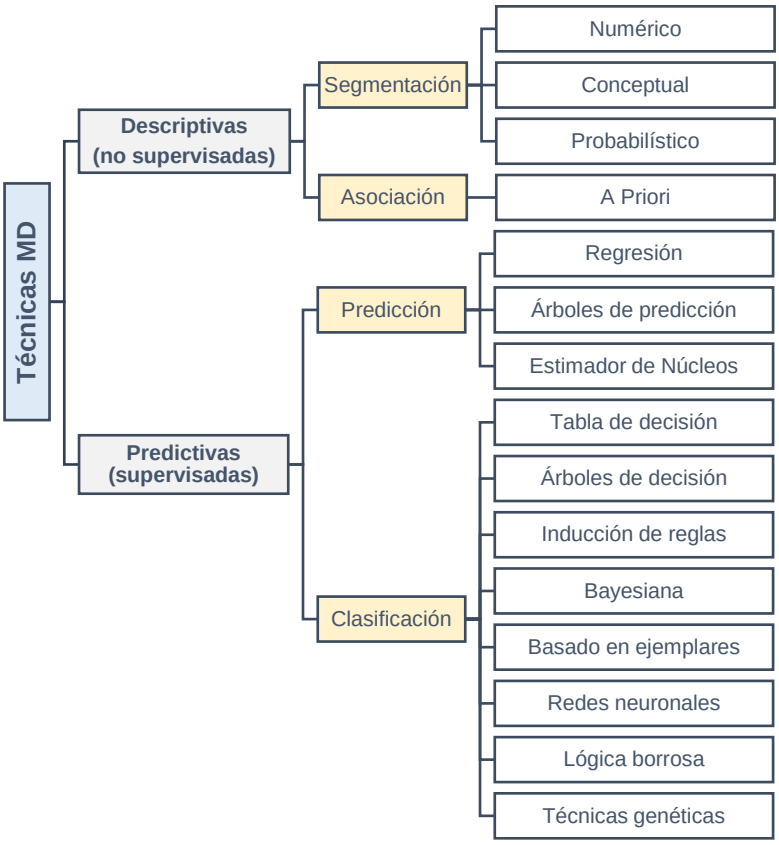
Para García & Molina (2012), una técnica es un enfoque conceptual que señala o indica el proceso de extracción de información en base a los datos, y es implementada generalmente por varios algoritmos. Estos algoritmos representan la forma en que se desarrolla una técnica específica paso a paso, razón por la cual, es primordial conocer a detalle cada algoritmo para identificar cual es la técnica más apropiada para un determinado problema. Es fundamental

comprender las caracteristicas y parámetros de los algoritmos, a fin de preparar los datos a ser analizados.

La Figura 7, representa una categorización basada en el propósito para el cual son empleadas las técnicas de Minería de Datos. Weiss & Indurkhya (2008) indican que, de acuerdo con el objetivo de análisis de los datos las técnicas se clasifican en dos grupos: Técnicas descriptivas (no supervisadas) y Técnicas predictivas (supervisadas)

Figura 7

Clasificación de las Técnicas de Minaría de Datos basada en el propósito



Nota. Adaptado de *Técnicas de Minería de Datos*, por J. García y J.M. Molina, 2012.

Las **técnicas descriptivas** o **no supervisadas** se utilizan cuando no se conocen los datos objetivo y se orientan a describir un conjunto de datos (Villena Román, Crespo García, &

García Rueda, 2012). A estas técnicas también se las denomina como no supervisadas debido a que el aprendizaje se desarrolla sin ninguna indicación por parte del usuario. De manera general, obtienen patrones que resumen las relaciones implícitas en los datos (tendencias, trayectorias, grupos, correlaciones y anomalías). Normalmente son tareas exploratorias que requieren validación y explicación (Universidad Nacional de San Luis, 2018).

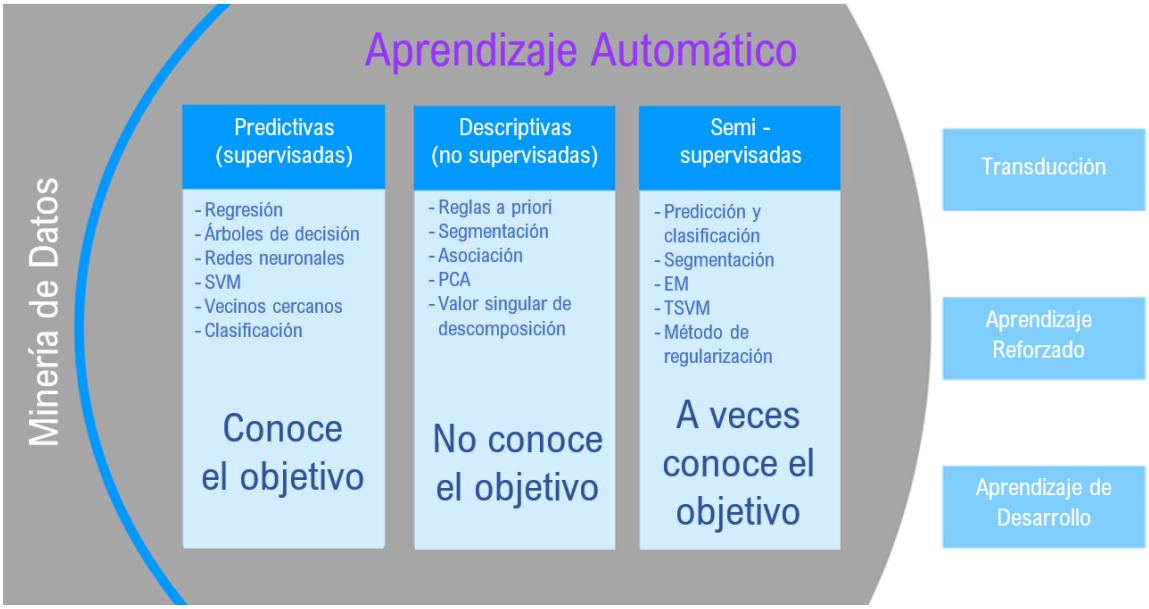
En cuanto a las **técnicas predictivas** o **supervisadas** se orientan a estimar o predecir valores de salida y se utilizan cuando se conoce los resultados objetivo de un conjunto de datos (Villena Román, Crespo García, & García Rueda, 2012). El aprendizaje supervisado construye criterios a partir de un conjunto de casos o ejemplos denominados conjunto de entrenamiento (training set), cuyos resultados son conocidos, con el fin de determinar (predecir) los resultados de nuevos casos en el dominio (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Se debe considerar que, las disciplinas de **Minería de Datos** (MD) y **Aprendizaje Automático** o **Machine Learning** (ML por sus siglas en inglés) comparten las mismas técnicas y modelos. Sin embargo, existen diferencias considerables respecto a su origen y uso, por ejemplo: MD adquiere la experiencia desde grandes volúmenes de bases de datos, mientras que ML incluye otras formas de entrenamiento; MD suele requerir mayor intervención humana que ML; MD incluye técnicas estadísticas que no son utilizadas por ML (Universidad Nacional de San Luis, 2018). Otra perspectiva indica que, “En *Machine Learning mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones, descubriendo información clave dentro de los proyectos de Minería de Datos*” (IBM Cloud Education, 2020).

La Figura 8, muestra de forma gráfica las técnicas compartidas en la intersección de las disciplinas de Minería de Datos y Machine Learning dentro del proceso de Descubrimiento de Conocimientos en Bases de datos (KDD).

Figura 8

Técnicas de Minería de Datos y Aprendizaje Automático en KDD



Nota. Presenta las técnicas comunes entre Minería de Datos y Aprendizaje Automático en KDD. Adaptado de *Machine Learning*, por B. Wujek, 2014.

Considerando el objetivo del presente proyecto de investigación, se analizan a continuación las técnicas del modelo predictivo y sus diferentes algoritmos.

Modelos predictivos

"*El análisis predictivo construye un modelo estadístico que utiliza los datos existentes para predecir datos de los cuales no se dispone*" (Espino Timón, 2017). Para lo cual se cuenta con un conjunto de datos que conoce el resultado de predicción y se lo denomina de entrenamiento (train), también se cuenta con otro conjunto de datos que no se conoce su resultado de predicción y se lo denomina de prueba (test).

Para la consecución del modelo predictivo, este se apoya en enfoques y técnicas que se exponen a continuación:

Regresión

Regresión lineal. Este tipo de regresión es empleado en la realización de procesos de análisis de asociación, es decir, la predicción se efectúa mediante la relación de dos variables: una dependiente y otra independiente, las cuales forman una ecuación lineal que representa de la manera más cercana posible a una determinada serie de datos; con el coeficiente de correlación adecuado, y la correcta ubicación de la recta de regresión (Romero & Ventura, 2010).

Regresión no lineal. Los datos analizados, en ocasiones, podrían no mostrar una dependencia lineal, en cuyo caso la relación se representa según una función polinómica (regresión polinómica) que puede planearse agregando las condiciones polinómicas al modelo lineal base. Mediante procesos de transformación de variables, se transforma el modelo no lineal en uno lineal a fin de resolverse mediante el método de mínimos cuadrados (García Herrero & Molina López, 2012).

Regresión logística. La regresión logística es una variación del algoritmo de red neuronal, cuyo objeto es transformar una función exponencial en lineal. Puede definirse como "*la probabilidad de que algunos eventos ocurran como una función lineal de un conjunto de variables predictoras*" (Westreich, Lessler, & Funk, 2010).

Árboles de predicción

Los árboles de predicción numérica, en general, son similares a los árboles de decisión, con la diferencia de que la clase a predecir es de tipo continua. Al haberse construido el árbol que clasifica las instancias, empieza el proceso de poda, luego del cual se genera para cada nodo hoja una constante (árboles de regresión) o un plano de regresión (árboles de modelos) (García Herrero & Molina López, 2012).

En el entrenamiento de un árbol de predicción numérica, las observaciones se van bifurcando por los diferentes nodos, generando la estructura del árbol hasta llegar a los nodos

terminales. Para predecir una nueva observación, se transita el árbol recorriendo los nodos predictores hasta alcanzar uno de los nodos terminales. La predicción del árbol es la media de las observaciones de los datos de entrenamiento que se encuentran en el mismo nodo terminal (Amat, 2017).

Serie temporal

Las series temporales se emplean en la predicción de tendencias basadas exclusivamente en el conjunto de datos original usado para crear el modelo, en predicción cruzada para crear un modelo general aplicable a múltiples series, en la previsión del tiempo de valores continuos (Medina & Gómez, 2014).

Un ejemplo de aplicación se encuentra en la medicina, para la predicción del tiempo de supervivencia de un paciente con un determinado diagnóstico, acorde a los síntomas presentados. Otro ejemplo lo encontramos en la industria, para la predicción de fallos en los componentes de una determinada máquina. Su aplicación, en general, implica la supervivencia o fracaso respecto a un evento específico (Salvo, Montato, Nunnaria, Neri, & Puglisi, 2013).

Árboles de decisiones

Los árboles de decisión pueden considerarse como la representación del método lógico que se emplea en un modelo jerárquico de aprendizaje supervisado en el cual "*la región local es identificada en una secuencia de divisiones recursivas a través de nodos de decisión con función de prueba (…) lo hace desde un conjunto de entradas- salidas de las muestras*". Generar un árbol de decisión puede considerarse como un método eficiente para la generación de clasificadores de datos (Medina & Gómez, 2014). Se considera un eficiente método no paramétrico (desde la perspectiva estadística) empleado en el tratamiento de datos tanto continuos como discretos, con distintos métodos para la creación del árbol y en el cual pueden incluirse tareas de análisis, la clasificación, la asociación y la regresión.

La creación del árbol se realiza en base a la determinación de las correlaciones existentes entre una entrada y el resultado deseado, luego de lo cual se emplea la ecuación que calcula la obtención de la información a fin de identificar en dicha información un atributo único de mayor puntuación, que separa los resultados en casos (subconjuntos de datos que se analizan en forma recursiva, mientras el árbol pueda dividirse); cada caso contiene una única red bayesiana anterior y una única medida de confianza para dicha red (Medina & Gómez, 2014).

Un árbol puede contener varias bifurcaciones, y su forma y profundidad dependen del método de puntuación (y otros parámetros empleados); un modelo podría además contener varios árboles para diferentes atributos de predicción (Barros, Basgalupp, Carvalho, & Freitas, 2012).

Redes neuronales

Las redes neuronales tienen como utilidad brindar análisis a datos complejos de entrada, o a relaciones complejas entre muchas entradas y pocas salidas. En cuanto a la red, cada uno de los nodos de cada capa oculta se encuentra restringido al mismo atributo de entrada, además "*contiene varias redes dependiendo del número de columnas de entrada y de los estados que contiene cada una de ellas, sus conexiones nodo-objetivo representa reglas de asociación entre entrada y objetivo*" (Medina & Gómez, 2014). Se aplica a la clasificación y reconocimiento de patrones de voz, predicciones del tiempo, mercado financiero, fraudes económicos, entre los principales (Swiderski, Kurek, & Osowski, 2012).

Normativa y Control en el Ecuador

Las empresas proveedoras de internet y datos del Ecuador se encuentran normadas por el Ministerio de Telecomunicaciones y de la Sociedad de la Información, y reguladas por la Agencia de Regulación y Control de las Telecomunicaciones. Estas Instituciones del estado han establecido leyes y reglamentos cuyo objetivo es garantizar la calidad, universalidad,

accesibilidad, continuidad y seguridad de los servicios de comunicación que se proveen a los clientes finales (ARCOTEL, 2019) (MINTEL, 2019).

La Ley Orgánica de Telecomunicaciones (2015) regula las actividades de establecimiento, instalación y explotación de redes, uso y explotación del espectro radioeléctrico, servicios de telecomunicaciones, así como a las personas naturales y jurídicas que las realizan. En su **Artículo 3**, respecto a los objetivos de dicha ley señala: *"11. Establecer el ámbito de control de calidad y los procedimientos de defensa de los usuarios de servicios de telecomunicaciones, las sanciones por la vulneración de estos derechos, la reparación e indemnización por deficiencias, daños o mala calidad de los servicios (…)".* Por otra parte, en el **Artículo 22** sobre los derechos de los abonados, clientes y usuarios determina que éstos tienen derecho a *"A disponer y recibir los servicios de telecomunicaciones contratados de forma continua, regular, eficiente, con calidad y eficacia.* El **Artículo 118** sobre las infracciones de segunda clase, incluye *"11. El incumplimiento de los valores objetivos de los parámetros de calidad contenidos en los títulos habilitantes, planes, normas técnicas y resoluciones emitidas por la Agencia de Regulación y Control de las Telecomunicaciones.".* Finalmente, el **Artículo 132** sobre legitimidad, ejecutividad y medidas correctivas establece que *"se podrá ordenar la reparación de los daños y perjuicios a terceros, tales como (..) la compensación a los abonados, clientes o usuarios por suspensión, interrupción o mala calidad del servicio".*

Gestión de disponibilidad

La gestión de la disponibilidad persigue como objetivo principal asegurar que los servicios de Tecnología de la Información estén disponibles y funcionen de manera correcta cuando los usuarios o clientes deseen utilizarlos (ITIL Foundation, 2011). La gestión de la disponibilidad se convierte entonces en una herramienta que promueve mantener una óptima calidad de servicio.

Dentro de las actividades o responsabilidades a ser consideradas dentro de la Gestión de la Disponibilidad, se encuentran las siguientes (ITIL Foundation, 2011):

- Establecer los requisitos de disponibilidad, en conjunto con los clientes.
- Garantizar el nivel de disponibilidad de servicios establecido.
- Mantener una continua monitorización de la disponibilidad de los sistemas TI.
- Elaborar propuestas de mejora a la infraestructura y servicios TI, con la finalidad de mejorar los niveles de disponibilidad.
- Supervisar el cumplimiento a nivel de proveedores internos y externos.

La disponibilidad depende en gran medida del adecuado o correcto diseño de los servicios y su mantenimiento, así como la calidad de los servicios internos y externos de terceros. Los beneficios obtenidos en base a una adecuada Gestión de la Disponibilidad son (ITIL Foundation, 2011):

- Cumplimiento de los niveles de disponibilidad establecidos.
- Reducción de costos.
- Mayor satisfacción del cliente, al percibir una mejor calidad de servicio.
- Aumento progresivo de los niveles de disponibilidad.
- Reducción del número de incidentes reportados.

Acuerdos de Nivel de Servicio (SLA)

Los Acuerdos de Nivel de Servicio o Service Level Agreement (SLA por sus siglas en inglés) describen los compromisos que el proveedor acuerda con el cliente, respecto a la entrega de servicios de TI. Éstos pueden ser medidos de manera cuantitativa o cualitativa, así como asociarse a una o varias escalas, las cuales describen las acciones a realizarse de no cumplirse los compromisos establecidos (IBM, 2019). Un SLA, por lo tanto, se define como un acuerdo establecido entre un proveedor de servicio TI y un cliente (ITIL Foundation, 2011). Los

SLA se traducen en un documento que suele estar anexo al Contrato de Prestación de Servicios, en este documento se estipulan las condiciones y parámetros que comprometen al prestador de servicio a cumplir en términos de niveles de calidad de servicio frente a sus clientes (Acens Technologies S.A., 2021).

Entre los principales apartados, referentes al servicio que debe contener un documento de SLA, se encuentran: Disponibilidad del servicio, Atención al cliente, Tiempo de respuesta, Mantenimiento, Penalizaciones (Acens Technologies S.A., 2021).

Indicadores de disponibilidad de servicio

En el siglo XIX, el británico Sir William Thomson, Lord Kelvin, escribió "*Lo que no se define no se puede medir, Lo que no se mide, no se puede mejorar. Lo que no se mejora, se degrada siempre*". Hoy en día esta frase sigue vigente, debido a la importancia que tiene la mejora continua en productos y servicios, por ello es imprescindible utilizar métricas adecuadas, estableciendo indicadores correctamente definidos y con objetivos claros (PÁGINA 7 COMUNICACIÓN S.L. MADRID, 2017).

De acuerdo con ITIL (ITIL Foundation, 2011), para realizar una adecuada gestión de disponibilidad y dar cumplimiento con los SLA, es necesario establecer indicadores que permitan medir con precisión las distintas fases del ciclo de vida de la interrupción del servicio. García (2018) establece los siguientes indicadores de disponibilidad:

- Disponibilidad total
- Tiempo medio de reparación (MTTR por sus siglas en inglés)
- Tiempo medio entre fallos (MTBF por sus siglas en inglés)

Disponibilidad total

La disponibilidad total (uptime) se calcula mediante la división del número de horas que un servicio ha estado realmente disponible y el número de horas acordado del servicio dentro de un determinado periodo (mensual, trimestral, semestral o anual) (García Garrido, 2018):

$$Disponibilidad = \frac{Horas\ totales - Horas\ de\ interrupción}{Horas\ totales}$$

Donde, "*Horas totales"* representa el tiempo acordado de servicio, "*Horas de interrupción"* representa el tiempo fuera de servicio. Para (ITIL Foundation, 2011) la Disponibilidad se mide en términos porcentuales y por tanto se multiplica por 100. Por ejemplo, si el servicio es 24/7 (**720 horas** por mes) y en el último mes el servicio se ha interrumpido **10 horas**, el porcentaje de disponibilidad se calcula de la siguiente forma:

$$\%\ \boldsymbol{Disponibilidad} = \frac{720\ h\ - 10\ h}{720\ h}\ x\ 100 = \mathbf{98.61}\%$$

Tiempo medio de reparación (MTTR)

El tiempo medio de reparación o Mid Time To Repair (MTTR por sus siglas en inglés) es el tiempo promedio de duración de una interrupción del servicio (Downtime), e incluye el tiempo de detección, respuesta y resolución (ITIL Foundation, 2011).

$$MTTR = \frac{Horas\ de\ interrupción}{Número\ de\ averías}$$

Por ejemplo, si un sistema falla **5 ocasiones** en un mes, y las fallas resultaron en un total de **10 horas** de tiempo de inactividad, el MTTR se calcula de la siguiente forma:

$$\boldsymbol{MTTR} = \frac{10\ h}{5} = \mathbf{2}\ \boldsymbol{horas}$$

Tiempo medio entre fallos (MTBF)

Según (ITIL Foundation, 2011), el tiempo medio entre fallos o Mid Time Between Failure (MTBF por sus siglas en inglés) es el tiempo promedio durante el cual el servicio se encuentra

disponible sin interrupciones. Se usa para dar un seguimiento de la disponibilidad y la fiabilidad del servicio, cuanto mayor sea esta métrica más fiable será el servicio (Atlassian Software, 2021).

$$MTBF = \frac{Horas\ totales - Horas\ de\ interrupción}{Número\ de\ averías}$$

Por ejemplo, si el servicio es 24/7 (**720 horas** por mes) y en el último mes el servicio se ha interrumpido **10 horas** por **5 ocasiones**, el cálculo de MTBF sería:

$$\boldsymbol{MTBF} = \frac{720\ h - 10\ h}{5} = \boldsymbol{142\ h}$$

Interpretando los indicadores calculados anteriormente, se tiene que, la ***Disponibilidad total*** de 98.61% se debe comparar con la **Disponibilidad acordada del servicio**, por ejemplo, si el objetivo es de 99.99%, entonces se tiene un margen de mejora. En este caso es necesario indagar las otras métricas para descubrir dónde puede estar el problema. Un **Tiempo medio de reparación (MTTR)** de 2 horas no es un valor óptimo, pero es aceptable. Sin embargo, un **Tiempo medio entre fallos (MTBF)** de 142 horas promedio al mes, quiere decir que aproximadamente cada 6 días se tiene un incidente, lo que indica que son muy frecuentes las interrupciones del servicio. Aquí es donde se torna urgente el investigar los datos de los incidentes para identificar tendencias o patrones de cortes recurrentes (Cheng, 2018). Lo ideal es tener un porcentaje de Disponibilidad total igual o superior al acordado, un MTTR bajo y un MTBF elevado (RinconTIC, 2020).

Capítulo III: Metodología

El presente capítulo explica la metodología de investigación general que se va a utilizar para el desarrollo del proyecto de titulación, así también, presenta los métodos y técnicas específicos que permitan contestar las preguntas de investigación planteadas. Adicionalmente, se va a realizar un estudio del estado del arte, con la finalidad de encontrar si existen soluciones formuladas al problema de investigación o en su defecto destacar los aportes del presente trabajo.

Metodología de Investigación

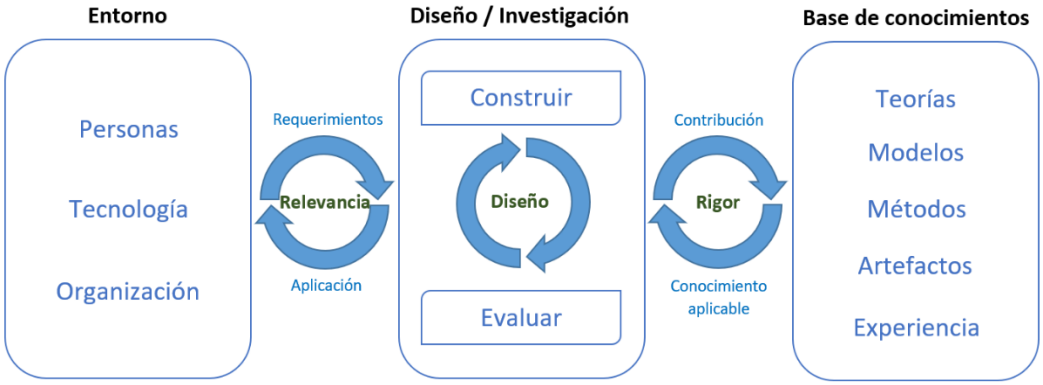
Investigación Científica basada en el Diseño (DSR)

La Investigación Científica basada en el Diseño (Design Science Research - DSR por sus siglas en inglés), contribuye a la solución de problemas relevantes en el mundo real, mediante el diseño novedoso y riguroso de artefactos, realizando aportes importantes en una determinada área del conocimiento (Piirainen, Gonzalez, & Kolsfschoten, 2010). Esta metodología se caracteriza por la construcción de artefactos innovadores y la retroalimentación que proveen para enriquecer los fundamentos teóricos (González & Pomares, 2012).

La Figura 9, muestra el modelo de proceso para DSR, que consiste en tres ciclos de investigación: Relevancia, Rigor y Diseño, estos permiten integrar el **entorno** donde se encuentra el problema y la **base del conocimiento** existente, para construir y evaluar un **artefacto** denominado también solución (Hevner & Chatterjee, Design Research in Information Systems Theory and Practice, 2010) .

Figura 9

Arquitectura general DSR



Nota. Presenta la investigación mediante ciclos de relevancia, rigor y diseño de DSR. Adaptado de *La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería*, por R. González y A. Pomares, 2012.

Se debe considerar que se trata de ciclos de investigación y no fases, por lo tanto, estos ciclos funcionan como engranajes que se conectan entre sí de forma iterativa, esto implica que mientras un ciclo cambia, los otros dos ciclos también cambian. Es decir, mientras se obtienen los requerimientos del entorno (relevancia), se debe sustentar de manera rigurosa por métodos existentes (rigor), aportando en la construcción del artefacto (diseño). Aunque de forma natural se puede iniciar por el ciclo de relevancia, DSR permite iniciar en cualquiera de los tres ciclos, dependiendo del estado de cada proyecto, brindando así flexibilidad, continuidad y evitando estancamientos dentro del proceso (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Ciclo de Relevancia. En este ciclo se identifica el dominio y contexto del problema, permite detectar objetivos organizacionales y oportunidades de mejora. Aquí se ejecuta las investigaciones del estado del arte para determinar la relevancia de abordar la solución del problema. También se formulan las preguntas de investigación y los requerimientos que guiarán el diseño y construcción del artefacto (Tebes, y otros, 2019).

Ciclo de Rigor. Es el encargado de asegurar que el diseño se fundamente en teorías científicas y métodos probados. Si el artefacto diseñado se convierte en solución al problema, se puede incluir en la Base de Conocimiento (Knowledge Base) como aporte a la comunidad investigativa (Hevner, A Three Cycle View of Design Science Research, 2007).

Ciclo de Diseño. Es el corazón del proyecto de investigación en DSR. Este ciclo itera más rápido entre la construcción de un artefacto, su evaluación y la retroalimentación, con la finalidad de refinar aún más el diseño. Evalúa alternativas frente a los requisitos del entorno hasta lograr un resultado que brinde solución al problema (Hevner, A Three Cycle View of Design Science Research, 2007). La evaluación de los artefactos de diseño es demostrable tras un tiempo de uso y previo a su implementación completa se debe validar la utilidad potencial sustentada científicamente (González & Pomares, 2012).

El presente proyecto de titulación plantea la generación de un **modelo predictivo,** que evalúe los sensores de los dispositivos de red de datos en una empresa proveedora de telecomunicaciones del Ecuador, para detectar con antelación posibles problemas de red, alertando al personal operativo para tomar acciones proactivas y mantener disponibles los servicios de internet y datos de sus clientes. Por tal razón, se decide utilizar la metodología de investigación basada en el diseño DSR, siendo el modelo predictivo el artefacto que se va a construir y evaluar (diseño) como solución a las necesidades de disponibilidad de la compañía (relevancia), basada y fundamentada en la teoría, métodos y modelos existentes (rigor).

Para la ejecución de las diferentes etapas del proyecto se va a utilizar la metodología DSR, sin embargo, en cada una de las fases se va a utilizar métodos y técnicas de investigación según los objetivos planteados.

Tabla 3

Métodos de investigación por preguntas de investigación

Objetivos específicos	Preguntas de investigación	Método de investigación
OE1: Determinar las técnicas predictivas supervisadas de minería de datos idóneas, que permitan pronosticar y establecer umbrales dinámicos de sensores de red, a través de una revisión preliminar de literatura especializada.	**OE1-RQ1**: ¿Qué soluciones presentan los estudios realizados sobre técnicas de minería de datos de predicción de umbrales dinámicos de sensores de red? **OE1-RQ2**: ¿Determinar cuál es la técnica de minería de datos más adecuada para la predicción de umbrales dinámicos de sensores de red?	Revisión Preliminar de Literatura
OE2: Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red, mediante el uso de técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos.	**OE2-RQ1**: ¿Qué modelo de minería de datos predice umbrales dinámicos de sensores de red, basados en tendencias de comportamiento de dispositivos? **OE2-RQ2**: ¿Cuáles son las aplicaciones y fuentes de datos de la empresa proveedora de telecomunicaciones, que almacenan la información histórica del monitoreo de los sensores de red?	CRISP-DM
OE3: Evaluar el modelo propuesto en una empresa proveedora de telecomunicaciones del Ecuador, por medio de un caso de estudio Benchmark, el cual permita obtener niveles de rendimiento y disponibilidad de los servicios.	**OE3-RQ1**: ¿Cuáles son los indicadores que permiten medir la mejora de disponibilidad del servicio de internet y datos? **OE3-RQ2**: ¿Cuál es el nivel de confianza de los resultados obtenidos?	Caso de estudio Caso de estudio Benchmark

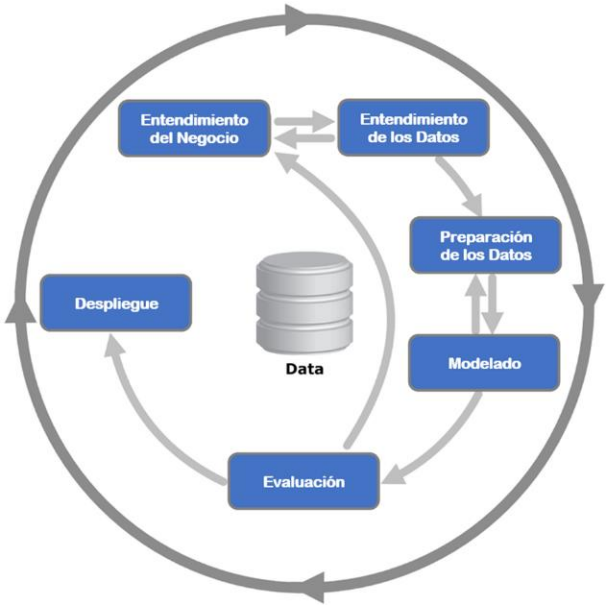
Nota. Presenta los métodos de investigación que se va a utilizar por cada objetivo específico y pregunta de investigación.

Metodología CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM por sus siglas en inglés) es una metodología estándar, probada en proyectos de Minería de Datos (IBM, 2021). La Figura 10, muestra las seis fases que comprende el ciclo de vida CRISP-DM, la secuencia de ejecución no es rígida, puede ser necesario avanzar y retroceder entre las diferentes fases (Chapman, y otros, 2000).

Figura 10

Ciclo de vida CRISP-DM



Nota. Presenta la secuencia e iteración de las fases de CRISP-DM. Adaptado de *CRISP-DM 1.0 Step-by-step data mining guide*, por Chapman, et al., 2000.

Entendimiento del Negocio. Permite comprender los requisitos y objetivos del proyecto desde una visión empresarial o institucional, para convertirlos en objetivos técnicos y generar un plan de proyecto (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016). Las tareas de esta fase son: determinar objetivos del negocio, valoración de la situación, determinar los objetivos de MD y realizar un plan de proyecto (Chapman, y otros, 2000).

Entendimiento de los Datos. Existen dos puntos clave en esta fase: Comprender los datos, su estructura y distribución; Y verificar la calidad de los datos (Vallalta Rueda, 2022). Las principales tareas que se debe efectuar en esta fase son: recolectar datos iniciales, describir, explorar y verificar la calidad de los datos (Chapman, y otros, 2000).

Preparación de los Datos. Es una de las fases más importantes y con frecuencia la más larga en proyectos de MD, se estima que un 50 – 70 % del tiempo y esfuerzo se aplica en la ejecución de esta fase (IBM, 2021). Entre las principales tareas de esta fase se tiene: seleccionar, limpiar, estructurar, integrar y formatear los datos (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Modelado. En esta fase, se seleccionan y aplican varias técnicas de modelado de MD, se calibran valores hasta encontrar resultados óptimos. Las tareas que abarca esta fase son: seleccionar la técnica de modelado, generar el plan de prueba, construir el modelo, evaluar el modelo (Chapman, y otros, 2000).

Evaluación. Permite evaluar el grado de acercamiento del modelo de MD a los objetivos del negocio planteados (Vallalta Rueda, 2022). Al finalizar la evaluación se debe llegar a una decisión sobre el uso del modelo de MD y sus resultados (Chapman, y otros, 2000). Las principales tareas de esta fase son: evaluar los resultados, revisión del proceso y determinar los próximos pasos (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Despliegue. Consiste en utilizar los nuevos conocimientos para implementar las mejoras en la organización (IBM, 2021). Las tareas principales de esta fase son: plan de implantación, plan de monitoreo y mantenimiento, informe final y revisión del proyecto (Chapman, y otros, 2000).

Estado del Arte

En esta sección se va a realizar una revisión sistemática de literatura sobre aspectos especializados, para la resolución del **Objetivo Específico No.1 (OE1)** del presente proyecto de titulación.

Definición del objetivo

Identificar los estudios existentes en la literatura científica relacionados a las preguntas de investigación planteadas, con la finalidad de encontrar soluciones basadas en técnicas de minería de datos para predicción de umbrales dinámicos de sensores de red de datos.

Criterios de inclusión y exclusión

Tanto los criterios de inclusión como de exclusión son los parámetros definidos y limitantes para la ubicación de artículos primarios correspondientes al tema de investigación.

Criterios de inclusión:

- Artículos relacionados a modelos o técnicas predictivas de problemas en redes de datos, que abarquen los campos multidisciplinares de KDD.
- Estudios que describan información de alertas tempranas o alertas proactivas en el monitoreo de sensores de dispositivos de red.
- Artículos que traten sobre proveedores de telecomunicaciones, servicios de internet o datos.

Criterios de exclusión:

- Estudios que únicamente se refieran a la implementación y gestión de sistemas de monitoreo de red sin uso de técnicas predictivas.
- Artículos generales de modelos o técnicas de predicción no relacionados al monitoreo de red de datos.

Grupo de control

La Tabla 4, presenta los estudios de investigación encontrados de acuerdo con los criterios de inclusión y exclusión, con los mismos se realiza un primer análisis de su título, introducción y conclusiones a fin de obtener las palabras claves.

Tabla 4

Grupo de control (GC)

GC	Título	Palabras clave
EC1	Machine Learning Methods for Traffic Prediction in Dynamic Optical Networks with Service Chains	Network traffic, prediction, machine learning
EC2	Improving the Performance of Network Traffic Prediction for Academic Organization by Using Association Rule Mining	Network traffic, prediction, data mining, association rule, prediction model
EC3	Network Quality Operation Prediction Based on Machine Learning Algorithms	Network traffic, prediction, decision tree, machine learning, network performance, data mining, forecasting, prediction model, network monitoring, telecomunication
EC4	Research on Network Traffic Prediction Model Based on Neural Network	Network traffic, prediction, machine learning, prediction model, neural network, computer network, communication
EC5	Network Traffic Prediction using Quantile Regression with linear, Tree, and Deep Learning Models	Machine learning, network traffic, prediction, traffic analysis, regression, decision tree, neural network
EC6	Deep Learning for Proactive Network Monitoring and Security Protection	Deep learning, proactive, forecasting, network monitoring, machine learning, data mining, decision tree, regression, communication
EC7	Automated Optical Networks with Monitoring and Machine Learning	Optical network, communication, network monitoring, machine learning, telecommunication

Nota. Presenta los estudios que conforman el grupo de control y sus palabras claves.

Construcción de la cadena de búsqueda

Para la construcción de la cadena de búsqueda se procedió a enmarcar en un contexto los términos que más se repiten y las palabras clave comunes entre los estudios del grupo de control. La Tabla 5, muestra el conteo de los términos y la totalización de cada uno de ellos.

Tabla 5

Conteo de palabras clave

Contexto	Palabra clave	EC1	EC2	EC3	EC4	EC5	EC6	EC7	# Rep.
Network	network traffic	x	x	x	x	x			5
	network monitoring			x			x	x	3
	communication				x		x	x	3
	telecommunication			x				x	2
	network performance			x					1
	computer network				x				1

Contexto	Palabra clave	EC1	EC2	EC3	EC4	EC5	EC6	EC7	# Rep.
Data Analysis	traffic analysis					x			1
	machine learning	x		x	x	x	x	x	6
	data mining		x	x			x		3
	decision tree			x		x	x		3
	neural network				x	x			2
	regression					x	x		2
	assosiation rule		x						1
	deep learning						x		1
Predictive	prediction	x	x	x	x	x			5
	prediction model		x	x	x				3
	forecasting			x			x		2
	proactive						x		1

Nota. Presenta el conteo de las palabras clave por contexto de estudios del GC.

Con el uso de las palabras clave más frecuentes y conectores lógicos AND y OR por su relación y contexto, se estableció la siguiente cadena de búsqueda:

(("network traffic" OR "network monitoring") AND ("machine learning" OR "data mining") AND ("decision tree" OR "neural network" OR "regression") AND ("prediction model" OR "forecasting" OR "prediction") AND ("communication" OR "telecommunication"))

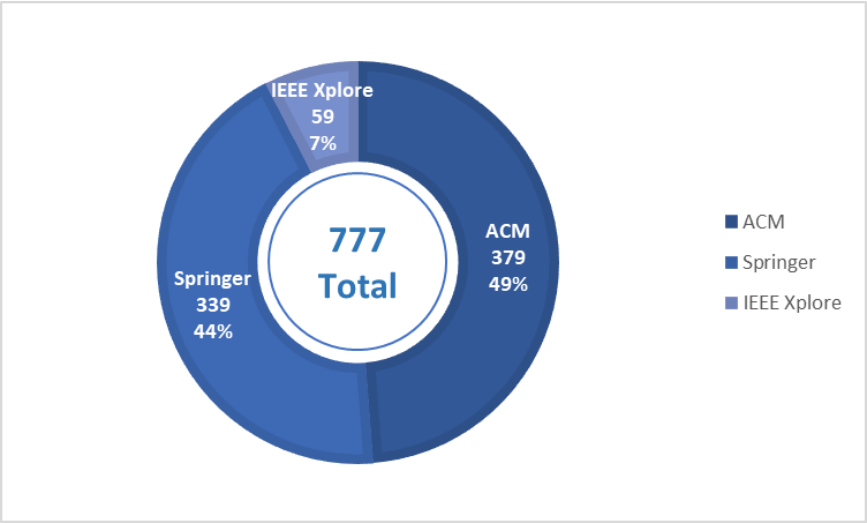
Estudios candidatos

Una vez ingresada la cadena de búsqueda en las bases digitales se obtuvo los siguientes resultados: 339 (Springer), 379 (ACM) y 59 (IEEE Xplore) estudios encontrados a partir del 2011.

La Figura 11, permite visualizar los 777 estudios candidatos y su distribución en cada una de las bases digitales encontradas, siendo ACM y Springer las que mayor cantidad de resultados presenta.

Figura 11

Resumen de estudios candidatos



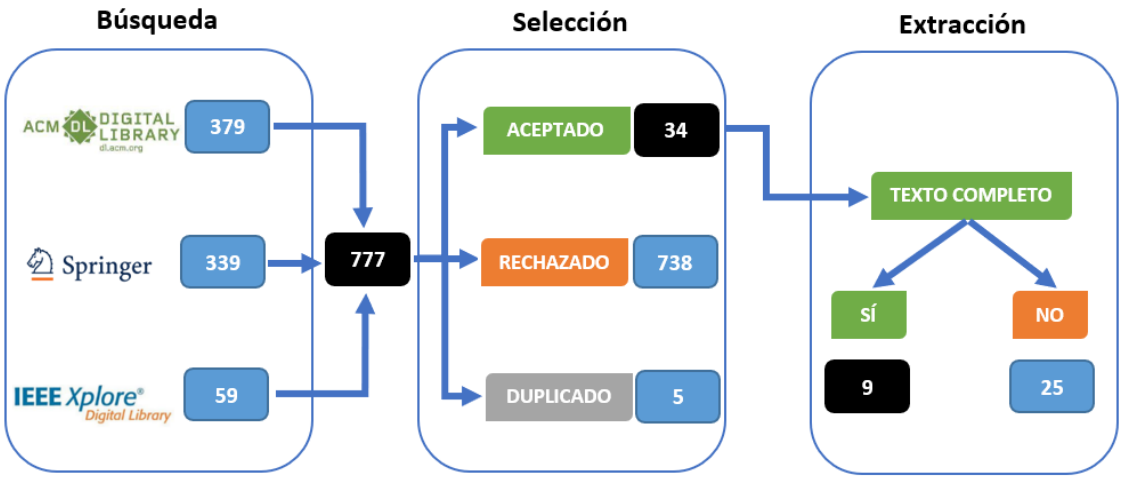
Nota. Estudios obtenidos con la cadena de búsqueda en las bases digitales.

Depuración de estudios candidatos

La Figura 12, presenta el flujo del proceso de depuración de estudios candidatos encontrados en las bases digitales. El objetivo principal es identificar los artículos que no se encuentren relacionados con el tema de investigación o que sean duplicados, para posteriormente seleccionar los estudios primarios.

Figura 12

Proceso de depuración de estudios candidatos



Nota. Depuración de los estudios obtenidos de las bases digitales.

Proceso detallado de depuración de estudios candidatos. Los pasos para este proceso son los siguientes:

a) Búsqueda

a.1. Ingreso de la cadena de búsqueda en las bases digitales para obtener los **estudios candidatos (777)**.

b) Selección

b.1. Exportación de resultados en archivos planos mediante las opciones que prestan los sitios web de las bases digitales.

b.2. Se unifica en una hoja de cálculo los resultados de las bases digitales.

b.3. A través de las opciones que poseen las herramientas de ofimática se procede a encontrar **estudios duplicados (5)**.

b.4. Mediante filtros, cruce de información y búsquedas (automáticas y manuales) en la hoja de cálculo, se procede a depurar los estudios con los criterios de inclusión y exclusión, a estos se identifican como **estudios rechazados (738)**.

b.5. Los **estudios aceptados (34)** son aquellos que cumplen con los criterios de inclusión y exclusión, por lo tanto, se encuentran relacionados al tema de investigación.

c) Extracción

c.1. Con los 34 estudios aceptados, se procede a realizar un primer análisis del título, la introducción y las conclusiones. Con esto se determina que (25) estudios no forman parte de los estudios primarios.

c.2. Finalmente, se obtienen los **estudios primarios (9)**, a los cuales se realiza una lectura de texto completo para la extracción de resultados del estado del arte.

Estudios primarios

A continuación, se realiza un análisis y resumen de los estudios que cumplen con los criterios de búsqueda y apoyan a responder a las preguntas de investigación planteadas.

Tabla 6

Resumen de estudios primarios

#	Título	Autor	Resumen
EP1	Automatic Detection of Network Traffic Anomalies and Changes – ACM 2019	Astha Syal, Alina Lazar, Jinoh Kim, Alex Sim, Kesheng Wu	Se basa en la utilización de un conjunto de mediciones de flujo de TCP recopiladas en múltiples nodos de transferencia de datos, para identificar posibles problemas de red, el estudio propone un clasificador de flujos capturados con algoritmos de clasificación lineal
EP2	DNS Traffic Forecasting Using Deep Neural Networks – Springer 2019	Diego Madariaga, Martín Panza, Javier Bustos	El estudio realiza un análisis de las El estudio realiza un análisis de las consultas que generan los usuarios hacia los servidores DNS, mediante el uso de técnicas de Machine Learning captura patrones del uso de internet para predecir el tráfico DNS y con esto compara los valores esperados con los reales. Para este proposito se propone un modelo de redes neuronales que permita detectar anomalías en los flujos de datos DNS
EP3	Network Quality Operation Prediction Based on Machine Learning Algorithms – IEEE Xplore 2019	A. V. Osin, O. I. Sheluhin	El artículo propone el análisis y predicción de la calidad de una red de datos, mediante el uso de Machine Learning y con la implantación de un modelo de árbol de decisión. Para seleccionar esta técnica, se realiza una comparación de las ventajas y desventajas de los diferentes métodos de predicción

#	Título	Autor	Resumen
EP4	Optimal Network route estimator using prediction algorithms – IEEE Xplore 2020	Shana B, Akarshan Agarwal, Abhisheck Ameta, Akshay Bhati, Abhay Desharpande	Este estudio se basa en la estimación de la mejor ruta de tráfico de red, basado en las observaciones anteriores del tráfico, con la intención de mejorar la calidad de los servicios. Para esto los autores proponen y aplican los algoritmos de ML: Random Forest y XGBoost para entrenar el modelo de predicción, alcanzando la presición máxima el modelo XGBoost
EP5	Network Traffic Prediction Using Recurrent Neural Networks – IEEE Xplore 2018	Nipun Ramakrishnan, Tarun Soni	Se basa en el problema de predecir el tráfico de red y sus características usando las observaciones del tráfico pasado, el estudio propone un modelo de Red Neuronal Recurrente (RNN por sus siglas en inglés) que permita predecir el tráfico de red y tomar acciones proactivas por parte del personal de operaciones
EP6	PreNoc: Neural Network based Predictive Routing for Network-on-Chip Architectures – ACM 2017	Michel Kinsy, Shreeya Khadka, Mihailo Isakov	El estudio plantea un algoritmo basado en redes neuronales de Machine Learning, para presentar información anticipada del estado de la red, con la finalidad de enrutar de manera eficiente el tráfico de red
EP7	Machine learning, Prophet and XGBoost algorithm: Analysis of Traffic Forecasting in Telecom Networks with time series data – IEEE Xplore 2020	Garima Jain, Rajveev Ranjan Prasad	Los autores de este estudio indican que, la previsión del tráfico de red tiene su importancia en fines comerciales, de planificación y soporte. Por lo tanto, proponen un modelo que pronostique el tráfico de red basado en algoritmos Prophet y XGBoost
EP8	Traffic Arrival Prediction for WiFi Network: A Machine Learning Approach – Springer 2020	Ning Wang, Bo Li, Mao Yang, Zhongjiang Yan, Ding Wang	Este trabajo propone la predicción del tráfico de red en redes WiFi, mediante el uso de Machine Learning con un algoritmo de regresión de bosque aleatorio (Random Forest)
EP9	Traffic Data Classification using Machine Learning Algorithms in SDN Networks – IEEE Xplore 2020	Jungmin Kwon, Daeun Jung, Hyunggon Park	El estudio se fundamenta en el monitoreo dinámico y proactivo de las redes de datos, al clasificar automáticamente los datos de tráfico mediante técnicas de Machine Learning como: Random Forest (RF), Linear Discriminant Analysis (LDA) y Deep Neural Network (DNN)

Nota. Presenta los estudios primarios y un extracto principal de su contenido.

Resultados del estado del arte

Con la elaboración del estado del arte para el presente trabajo de titulación y basado en las preguntas de investigación planteadas se tiene:

OE1-RQ1 ¿Qué soluciones presentan los estudios realizados sobre técnicas de minería de datos de predicción de umbrales dinámicos de sensores de red?

Posterior al análisis y búsqueda en los estudios primarios, se determina que las técnicas más utilizadas para predicción de umbrales dinámicos de red son:

- Árboles de decisión
- Redes neuronales
- Clasificación lineal
- Modelos de regresión

OE1-RQ2 ¿Determinar cuál es la técnica de minería de datos más adecuada para la predicción de umbrales dinámicos de sensores de red?

La Tabla 7, muestra el análisis de los estudios primarios, basados en: el problema planteado; las soluciones y técnicas encontradas para su resolución; y los aportes que deja para futuros estudios.

Tabla 7

Análisis de los estudios primarios y sus resultados

#	Problema definido	Solución encontrada	Técnica predictiva	Aporte
EP1	Periodos de tiempo lentos en nodos de una Red Tolerante al Retardo (DTN, por sus siglas en inglés)	El sistema propuesto divide los registros de monitoreo de la red en fragmentos de tamaño fijo y propone un clasificador de flujos de última generación para identificar las ventanas de tiempo lentas.	Clasificación lineal	El método propuesto es capaz de generar modelos para detectar rápidamente grandes intervalos de transferencias de red de bajo rendimiento, para la toma de decisiones de los ingenieros de redes
EP2	Gran cantidad de datos recopilados por los servidores DNS sin análisis y clasificación para detección de posibles problemas	Propone el uso de modelos de Machine Learning para capturar los patrones de uso de Internet para predecir el tráfico DNS	Redes neuronales	Compara la diferencia entre el tráfico DNS esperado y el real, detectando anomalías en los flujos de datos causados por ataques o fallos.
EP3	Predicir la calidad de red informática que los usuarios experimentan en la navegación Web	Uso de Machine Learning mediante la técnica de árboles de decisión para detectar cuando los usuarios experimentan baja calidad de red al acceder a los sitios Web	Árboles de decisión	Utilizando el modelo obtenido, se realizaron experimentos numéricos para predecir los momentos en los que los usuarios tienen problemas para abrir páginas web
EP4	Preveer o estimar la mejor ruta del tráfico de red a partir de observaciones de las	Este estudio se concentra en examinar dos algoritmos de ML Random Forest y XGBoost para	Árboles de decisión (Random	Los resultados de la simulación se han implementado en Python y se ha obtenido que el modelo

#	Problema definido	Solución encontrada	Técnica predictiva	Aporte
	rutas de tráfico pasado para mejorar la calidad de los servicios	analizar y pronosticar la mejor ruta de red.	Forest, XGBoost)	XGBoost es más preciso que Random Forest
EP5	Predicción del tráfico de la red implica predecir las características del tráfico futuro de la red a partir de las observaciones del tráfico pasado	El estudio propone un modelo de Red Neuronal Recurrente (RNN por sus siglas en inglés) que permita predecir el tráfico de red y tomar acciones proactivas por parte del personal de operaciones.	Redes neuronales	Los autores plantean que las arquitecturas RNN demuestran resultados prometedores, superando a los modelos de pronóstico estándar
EP6	Los sistemas informáticos actuales contienen 64 o más núcleos en un chip de CPU, pero las técnicas de comunicación basadas en bus no superan los 16 núcleos	El artículo propone un algoritmo de enrutamiento predictivo basado en redes neuronales para redes en chip que utiliza información anticipada sobre el estado de la red global y la congestión para enrutar de manera eficiente el tráfico de la red	Redes neuronales	Los autores indican que, en la mayoría de los casos de prueba, el enfoque propuesto supera las actuales técnicas de enrutamiento adaptativo y determinista en términos de latencia y rendimiento
EP7	Predecir el tráfico de red por temas comerciales, de planificación y de soporte	Este artículo propone un modelo de predicción de tráfico de red basado en algoritmos de Machine Learning: Prophet y XGBoost	Árboles de decisión (Prophet, XGBoost)	Los resultados indican que el modelo Prophet es preciso con una gran cantidad de datos, en cambio el modelo XGBoost es preciso incluso con menos datos, por lo tanto este último se usa para estimar el tráfico en horas pico.
EP8	Dada la importancia de las redes WiFi, se necesita predecir el tiempo de llegada del tráfico de red	Este documento propone un método de predicción de llegada de tráfico de red basado en Machine Learning mediante el uso de un algoritmo de regresión de bosque aleatorio (random Forest)	Árboles de decisión (Random Forest regression)	Según los autores, los resultados muestran que la precisión de predicción del modelo Random Forest es de alrededor del 95%, superando el flujo de predicción lineal
EP9	Monitorear de manera proactiva la dinámica de la red, analizar los datos de la red y predecir el uso de la red de manera automática	El estudio plantea el monitoreo dinámico y proactivo de las redes de datos, al clasificar automáticamente los datos de tráfico mediante técnicas de Machine Learning como: Random Forest (RF), Linear Discriminant Analysis (LDA) y Deep Neural Network (DNN).	Árboles de decisión (Random Forest), Redes neuronales, Clasificación lineal	A partir de los resultados del experimento con una topología de red simple, los autores indican que, los algoritmos de aprendizaje automático pueden clasificar los datos del tráfico de red. Pero para redes más complejas, se necesitan otros mecanismos.

Nota. Muestra el planteamiento del problema, la solución propuesta y los aportes de los estudios seleccionados.

La Tabla 8, presenta un conteo de las técnicas predictivas en las disciplinas de Data Mining, Machine Learning y Deep Learning utilizadas en los diferentes estudios primarios seleccionados, con la finalidad de contestar la pregunta de investigación planteada.

Tabla 8

Técnicas predictivas utilizadas en los estudios primarios (EP)

Técnica predictiva	EP1	EP2	EP3	EP4	EP5	EP6	EP7	EP8	EC9	# Rep.
Árboles de decisión			x	x			x	x	x	5
Redes neuronales		x			x	x			x	4
Clasificación lineal	x								x	2
Modelos de regresión								x		1

Nota. Presenta el conteo de las técnicas predictivas de los estudios seleccionados.

Mediante el análisis realizado a los estudios seleccionados, se puede visualizar que, los **Árboles de decisión** es la técnica más utilizada por los investigadores para la predicción del tráfico de red y la detección de problemas en los sensores de red. Cabe recalcar que en especial se mencionan los algoritmos de **Random Forest (RF)** y **XGBoost,** los cuales, son una evolución de los Árboles de decisión tradicionales, que se caracterizan por el manejo de: gran cantidad de variables; mayor profundidad de análisis de los datos; y mayor precisión en la estimación (Code Algorithms Pvt. Ltd., 2022). Por estas razones, se toma la decisión de usar esta técnica para el desarrollo de la solución planteada en el presente estudio de investigación, la misma que se va a detallar en el siguiente capítulo.

Capítulo IV: Desarrollo de la solución

Este capítulo presenta paso a paso la elaboración de la solución propuesta (modelo predictivo), mediante la Metodología CRISP-DM (ver sección *Metodología CRISP-DM*).

Fase 1: Entendimiento del negocio

A continuación, se presenta las tareas y actividades que comprenden la fase inicial del proceso de minería de datos, su finalidad es comprender los objetivos y requisitos del proyecto desde un punto de vista de negocio, para convertirlos en objetivos técnicos de Minería de Datos.

Determinar los objetivos del negocio

Contexto. Una de las principales empresas proveedoras de telecomunicaciones del Ecuador con cobertura a escala nacional, ha encontrado problemas en los índices de disponibilidad de sus servicios de internet y datos brindados a sus clientes corporativos. Entre las diferentes causas analizadas por el personal técnico y operativo de la compañía, se ha determinado que existe un inadecuado monitoreo de los enlaces, debido a que no se cuenta con una herramienta proactiva, que permita predecir y alertar tempranamente sobre posibles fallas o cortes en sus servicios de red.

Objetivos del negocio. Basado en los objetivos planteados en el *Capítulo I*, se describen los siguientes objetivos de negocio:

- "*Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red y alertas tempranas, mediante técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos, con la finalidad de mejorar los índices de disponibilidad del servicio corporativo de internet y datos de una empresa proveedora de telecomunicaciones del Ecuador*" (Objetivo general).

- "*Evaluar el modelo propuesto en una empresa proveedora de telecomunicaciones del Ecuador, por medio de un caso de estudio Benchmark, el cual permita obtener niveles de rendimiento y disponibilidad de los servicios"* (Objetivo específico 3, OE3).

Criterios de éxito. Desde el punto de vista empresarial, se establece como criterio de éxito, que los resultados de la evaluación del modelo de MD sean fiables en la detección de posibles cortes de red. Para determinar la fiabilidad del modelo se ha optado por dos tipos de evaluación: La primera, es el uso de ROC - AUC que es una métrica estadística para problemas de clasificación (ver sección *Evaluación de los resultados* de la Fase 5 de CRISP-DM); La segunda, es una comparación Benchmark entre clientes de la empresa de telecomunicaciones a fin de determinar si el modelo se aplica a varios escenarios (ver sección *Proceso de revisión* de la Fase 5 de CRISP-DM).

Evaluación de la situación actual

Desde el año 2017, la empresa proveedora de telecomunicaciones cuenta con un Sistema de Administración de Redes **NMS** (Network Management System por sus siglas en inglés), que se encarga de monitorear los dispositivos de red; enrutadores (routers) y conmutadores (switches); de los clientes corporativos de la compañía. Esta herramienta monitorea sensores como: ping, tráfico, memoria, CPU y temperatura de los equipos de red (GrupoCONTEXT S.A., 2022), al momento de existir un error o corte en los enlaces, se alerta al personal de soporte y operaciones. Sin embargo, estas acciones son correctivas (posteriores a la ocurrencia de un incidente), impactando finalmente a los porcentajes de disponibilidad establecidos en los SLA con los clientes.

NMS (Network Management System): Es una herramienta con licencia comercial distribuida por la empresa ecuatoriana GrupoCONTEXT S.A., dedicada a la administración de redes, que monitorea y almacena la información de elementos de red, permitiendo hacer

diagnósticos Casi en Tiempo Real o NRT (Near Real Time por sus siglas en inglés), entre sus principales características se tiene (GrupoCONTEXT S.A., 2022):

- Monitoreo de red bajo demanda
- Interfaz web de usuario intuitiva
- Arquitectura robusta, escalable y distribuida
- Agentless, es decir, sin la necesidad de agentes instalados para monitoreo
- Base de datos que consolida la información de los monitoreos
- Integración con sistemas de descubrimiento automático y alertas

Inventario de recursos. A continuación, se detallan los recursos disponibles para la ejecución del proceso de Minería de Datos:

- Hardware
    - Laptop Marca: DELL. Modelo: Inspiron 7573
    - Procesador: Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz 1.99 GHz
    - Memoria RAM: 12 GB
    - Almacenamiento: 512 GB SSD
- Software
    - Sistema Operativo: Windows 11 Home
    - Base de datos: SQL Server 2019
    - Programa de MD: RapidMiner Studio 9.10
    - Programa de Ofimática: Microsoft Office 2019
- Fuente de datos
    - Se dispone de datos en SQL Server, de monitoreo de red del software NMS, desde el 01 de marzo al 20 de junio del 2020.

Requisitos, supuestos y restricciones. En esta sección se especifica los requisitos, supuestos y restricciones de presentación y divulgación de la información del presente proyecto. Por motivos de confidencialidad entre la empresa de telecomunicaciones y sus clientes corporativos, los nombres de los clientes, agrupaciones y dispositivos deben ser reservados y no pueden ser divulgados por ningún motivo. Por lo tanto, es necesario establecer nombres de clientes ficticios genéricos, los mismos que, se pueden visualizar más adelante en la sección *Limpieza de datos* de la Fase 3 de CRISP-DM.

Riesgos y contingencias. Para el presente proyecto se prevé que, existen riesgos latentes que pueden impedir la consecución de los objetivos y resultados propuestos. Sin embargo, también existen contingencias que van a ayudar a mitigar el riesgo. La Tabla 9, muestra los riesgos y contingencias encontrados para el proyecto de MD.

Tabla 9

Riesgos y contingencias del proyecto de MD

#	Riesgos	Contingencias
R1	Cambio de software NMS de monitoreo parcial o total dentro de la empresa.	El software NMS se encuentra con licencia activa desde el año 2017 y en febrero del año 2022 se renovó su licencia por un año calendario. El software NMS actualmente genera reportes operativos y gerenciales indispensables para el soporte de los clientes. Se estima que a corto plazo no exista un cambio radical del sistema de monitoreo dentro de la compañía.
R2	Restricción de acceso a los datos de monitoreo.	Se han firmado contratos de confidencialidad de la información entre el autor del proyecto y la compañía de telecomunicaciones vigentes haste el 2022.
R3	Pérdida de información y/o respaldos de los datos de monitoreo	La empresa mantiene su polìtica propia de respaldos del software y la información con manejo de discos externos y/o en la nube.

Nota. Presenta los riesgos potenciales y sus contingencias para mitigarlos.

Costos y Beneficios. Existen costos intrínsecos para la consecución de la solución de MD planteada, los cuales se detallan en la Tabla 10. Sin embargo, estos han sido asumidos por el autor del trabajo de titulación.

Tabla 10

Costos del proyecto de MD

Grupo	Descripción	Cantidad	Costo Unitario	Costo Total
Hardware	Laptop	1	1,200.00	1,200.00
	Impresora	1	300.00	300.00
Artículos de oficina	Resmas de papel	2	4.00	8.00
	Tinta de impresora	4	22.00	88.00
	Perfiles	6	0.50	3.00
Servicios	Internet (mensual)	8	30.00	240.00
	Transporte (mensual)	8	15.00	120.00
	TOTAL			**1,959.00**

Nota. Presenta los costos tangibles e intangibles.

Los beneficios que el proyecto de Minería de Datos va a brindar a la compañía de telecomunicaciones, se listan a continuación:

- La empresa proveedora de servicios de internet y datos, va a contar con una herramienta proactiva de monitoreo de red basada en técnicas fiables de MD.
- Mejorar el soporte del personal operativo, al contar con alertas tempranas de posibles fallas o cortes en el servicio de telecomunicaciones.
- Mejorar los índices de disponibilidad de los enlaces de los clientes corporativos de la compañía.

Determinar los objetivos de Minería de Datos

Basado en los objetivos del presente trabajo de titulación (ver sección *Objetivos* del *Capítulo I*) se plantean los siguientes objetivos de Minería de Datos:

- "*Determinar las técnicas predictivas supervisadas de minería de datos idóneas, que permitan pronosticar y establecer umbrales dinámicos de sensores de red, a través de una revisión preliminar de literatura especializada*" (Objetivo específico 1, OE1).

- “Desarrollar un modelo predictivo de umbrales dinámicos de sensores de red, mediante el uso de técnicas de minería de datos basadas en tendencias de comportamiento de los dispositivos” (Objetivo específico 2, OE2).

Nota: El primer objetivo de MD, fue resuelto en la sección *Resultados del estado del arte* del *Capítulo III*. Sin embargo, en el apartado *Selección de técnicas de modelado* del presente capítulo se detalla a profundidad este tema.

Fase 2: Entendimiento de los datos

La fase 2 de la metodología CRISP-DM se encarga de: entender y recolectar los datos iniciales para la solución del problema de MD. También, permite familiarizarse con la información, averiguar su calidad e identificar las relaciones existentes entre los datos.

Recopilación de datos iniciales

Se ha analizado el funcionamiento de los sistemas y los repositorios de datos que posee la compañía para la información de red y se obtiene la siguiente descripción:

- El software NMS (ver sección *Evaluación de la situación actual* de la Fase 1 de CRISP-DM) monitorea los dispositivos de redes propias y de sus clientes corporativos. Es necesario configurar en la aplicación la dirección IP y la comunidad SNMP (Simple Network Management Protocol), por cada dispositivo. Existiendo la opción de ingreso masivo por rango de IP.
- La información del monitoreo de red recopilada por el software NMS, se almacenan en una base de datos SQL Server 2017 en tablas estructuradas, con valores por minuto de cada sensor de los dispositivos.
- Existe una estructura vertical que contiene la relación entre grupos, dispositivos y sensores de red y para obtener sus características es necesario correlacionar con otras tablas que almacenan estos datos.

- Existe una base de datos que almacena las capturas del monitoreo de red de los últimos tres meses, incluido el monitoreo actual en línea.
- Mediante un proceso nocturno se envía la información estructurada del monitoreo histórico (mayor a tres meses) a otra base de datos, la cual se encuentra configurada en la misma instancia del servidor de SQL Server 2017.

Posterior al análisis inicial de la base de datos, se obtiene las estructuras principales donde se almacena la información necesaria para el proceso de MD.

Recopilación de Entidades: Grupos, Dispositivos y Sensores. En el sistema existen 3 entidades principales de telecomunicaciones, los cuales se mantienen en una única estructura de datos que se relacionan entre sí mediante códigos en cascada (padre, hijo), estos se detallan a continuación:

- **Grupo**: Es un conjunto de dispositivos de red que se relacionan de acuerdo con diversas características como: región, ciudad, empresa, tipo, etc.
- **Dispositivo**: Se refiere a los equipos de red como enrutadores (routers) y conmutadores (switches).
- **Sensor**: Es una interfaz y/o protocolo que pertenece a un equipo de red, el cual se puede configurar para captar mediciones específicas en un tiempo determinado. El sistema maneja sensores como: CPU, ping, memoria, temperatura de los equipos y tráfico (por cada interfaz de red). Un dispositivo puede tener uno o varios sensores de red.

La entidad a nivel de base de datos que gestiona esta información se llama [ESTRUCTURA], esta se apoya con otras entidades para determinar el tipo y sus características, entre las cuales se tiene: [ELEMENTO], [ELEMENTO_DETALLE] y [ELEMENTO_DETALLE_VALOR].

Recopilación de Entidades: Canales. En el software NMS, un sensor tiene uno o varios canales de medición, los cuales dependen del tipo de sensor monitoreado. Por ejemplo, el sensor de ping posee un canal llamado "*Ping Time*", el cual mide el tiempo de latencia en milisegundos; en cambio, el sensor de tráfico posee varios canales tales como: "*Traffic In (volume)*", "*Traffic In (speed)*", "*Traffic Out (volume)*", "*Traffic Out (speed)*", que miden la cantidad de paquetes y la velocidad del tráfico de entrada y salida en un momento determinado. La entidad que almacena las características de estos canales se llama [CANAL] y se relaciona a cada sensor a través de la entidad [ESTRUCTURA_CANAL].

Recopilación de Entidades: Monitoreo. El monitoreo de red en el software NMS, se almacena en la entidad de Base de datos denominada [RESULTADO_ABSOLUTO], la misma que contiene varios atributos, tales como: el **valor** de medición, la **fecha** de captura de la medida y un código que se relaciona con la tabla [ESTRUCTURA_CANAL] para identificar el canal, el sensor y el dispositivo al cual se hace referencia el monitoreo.

Por motivos de presentación de la información, en pantallas y reportes del software NMS, el sistema cuenta con tablas alternas para mejora de rendimiento. Estas entidades se encargan de tener información resumida en diferentes escalas de tiempo, entre las cuales se tiene: [RESULTADO_MINUTO], [RESULTADO_HORA], [RESULTADO_DIA], [RESULTADO_SEMANA]. Adicionalmente, mantiene una tabla histórica de monitoreo [RESULTADO_ABSOLUTO_HIS].

Recopilación de Entidades: Eventos y Umbrales. Los eventos son aquellas novedades que el sistema NMS genera y alerta al personal de soporte de la compañía de telecomunicaciones. Estos eventos, según su severidad se catalogan en diferentes tipos de impacto (menor, alerta o crítica). Entre los principales eventos se tiene: los cortes de conexión, la saturación del enlace, la superación de los rangos o umbrales establecidos, entre otros. Toda esta información se almacena en la entidad de base de datos [EVENTOS].

Los umbrales estáticos son valores establecidos como límites o rangos para cada canal de los sensores de red. Si el valor o medida del monitoreo en un tiempo determinado sobrepasa o está por debajo de los umbrales, se genera un evento con la severidad e impacto correspondientes. La entidad [UMBRAL_CANAL] almacena esta información.

Recopilación de Entidades: Disponibilidad. El sistema NMS, a través del monitoreo y los eventos generados, determina los porcentajes de disponibilidad y almacena en la entidad de base de datos denominada [SENSOR_DISPONIBILIDAD]. Se debe considerar que, esta información se correlaciona entre los canales, sensores y dispositivos para emitir un valor de disponibilidad por cada dispositivo o por grupo de dispositivos.

Descripción de los datos

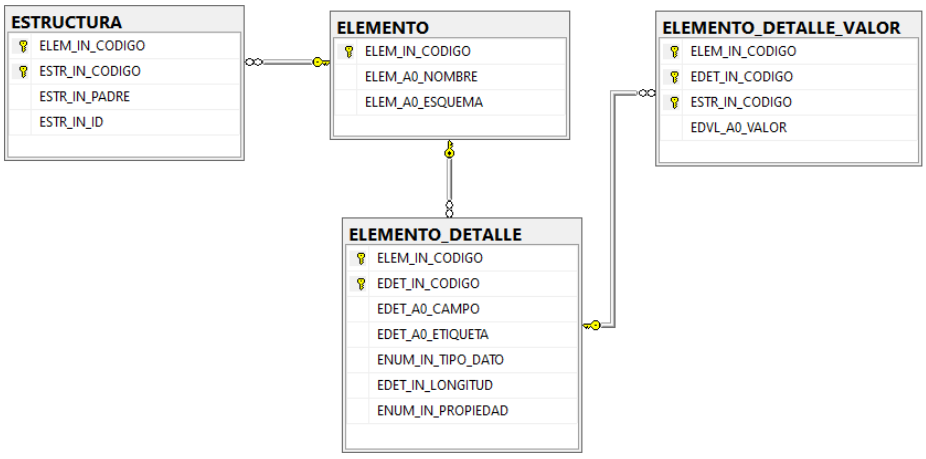
La información original del sistema NMS, se encuentra estructurada y almacenada en una base de datos relacional SQL Server. A continuación, se procede a describir de manera agrupada las entidades y relaciones necesarias para el proceso de MD.

Nota: El Apéndice B. Diccionario de datos, contiene la descripción detallada de cada una de las tablas, campos y relaciones de base de datos del presente apartado.

Descripción de Entidades: Grupos, Dispositivos y Sensores. La Figura 13, muestra el Modelo Entidad Relación (ER) que permite obtener las características y descripciones de los grupos, dispositivos y sensores de red. Las tablas que almacenan esta información son: [ESTRUCTURA], [ELEMENTO], [ELEMENTO_DETALLE] y [ELEMENTO_DETALLE_VALOR], en la siguiente sección, se procede a describir cada una de las entidades y sus campos.

Figura 13

Modelo ER de Grupos, Dispositivos y Sensores de red

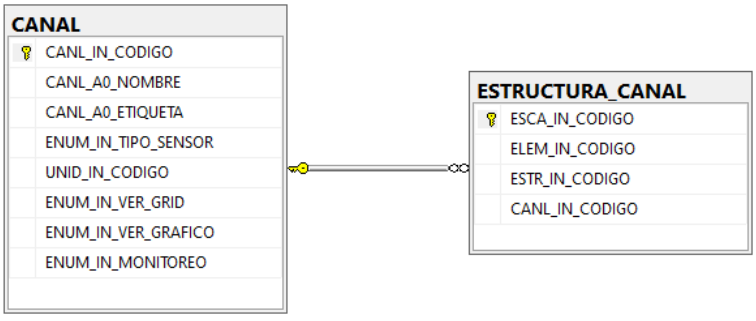


Nota. Modelo obtenido con la herramienta Database Diagram de SQL Server.

Descripción de Entidades: Canales. La Figura 14, presenta el Modelo ER de las estructuras que almacenan las características de los canales de monitoreo de red y su relación con los sensores de red. Las tablas que contienen esta información son: [CANAL] y [ESTRUCTURA_CANAL], en el siguiente apartado, se describe cada una de las entidades y sus campos.

Figura 14

Modelo ER de Canales

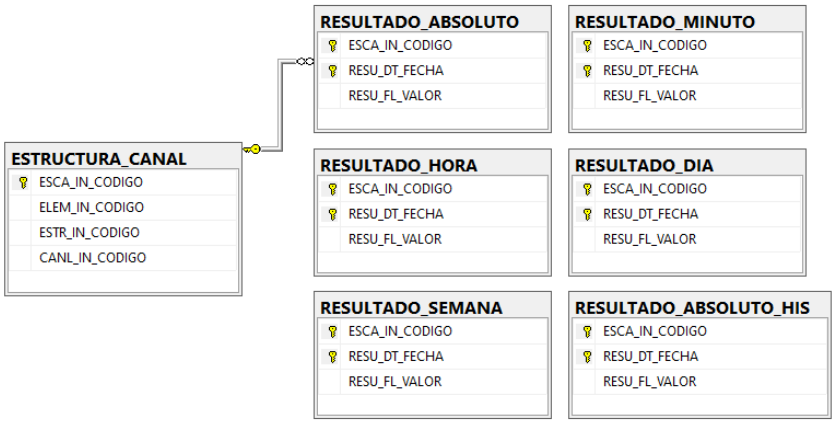


Nota. Modelo obtenido con la herramienta Database Diagram de SQL Server.

Descripción de Entidades: Monitoreo. La Figura 15, presenta el Modelo ER de las tablas que guardan los valores de monitoreo capturados en un tiempo determinado por cada canal y sensor. La tabla principal para esta tarea se llama [RESULTADO_ABSOLUTO] la cual se relaciona con la tabla [ESTRUCTURA_CANAL] (descrita en la sección anterior). Adicionalmente, existen entidades que apoyan en los resúmenes e históricos de información, las cuales son: [RESULTADO_MINUTO], [RESULTADO_HORA], [RESULTADO_DIA], [RESULTADO_SEMANA] y [RESULTADO_ABSOLUTO_HIS], en la siguiente sección, se describe los campos de estas entidades.

Figura 15

Modelo ER de Monitoreo

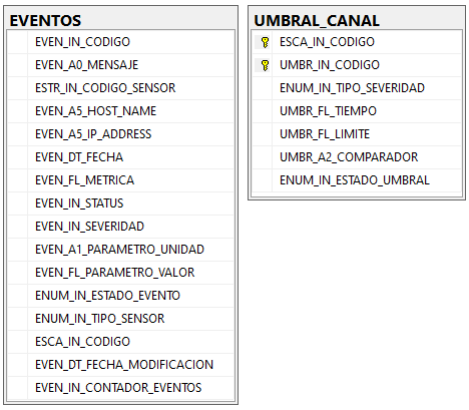


Nota. Obtenido con la herramienta Database Diagram de SQL Server.

Descripción de Entidades: Eventos y Umbrales. La Figura 16, presenta el Modelo Entidad Relación de las tablas que almacenan los eventos y los umbrales estáticos configurados en los dispositivos de red en el sistema NMS. Las entidades que conforman este grupo de estructuras son: [EVENTOS] y [UMBRAL_CANAL]. En el siguiente apartado, se describe los campos de estas entidades.

Figura 16

Modelo ER de Eventos y Umbrales estáticos



Nota. Obtenido con la herramienta Database Diagram de SQL Server.

Descripción de Entidades: Disponibilidad. La Figura 17, presenta el Modelo ER que almacena los valores de disponibilidad por cada canal de los sensores de red, la tabla que lleva a cabo esta tarea se denomina: [SENSOR_DISPONIBILIDAD], la cual se va a describir en el siguiente apartado.

Figura 17

Modelo ER de Disponibilidad

Nota. Modelo obtenido con la herramienta Database Diagram de SQL Server.

Exploración de datos

En esta sección se va a realizar una exploración de los datos almacenados, con la finalidad de descubrir tendencias e información inicial para el proceso de MD. Esta tarea se realiza a través de consultas SQL a las tablas y vistas del sistema NMS y con la ayuda de herramientas de ofimática para la representación gráfica. Es importante recalcar que, por

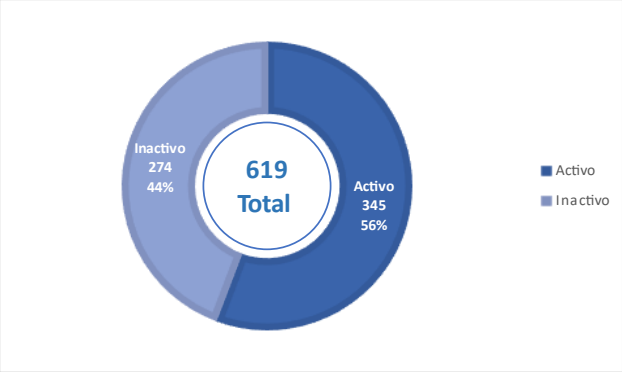
motivos de confidencialidad de los clientes corporativos los nombres de los grupos, dispositivos y sensores han sido modificados con nombres ficticios genéricos.

Nota: El Apéndice C. Scripts de consultas de base de datos, contiene las consultas utilizadas para la obtención de los siguientes resultados.

Exploración de Entidades: Grupos, Dispositivos y Sensores. La Figura 18, presenta la cantidad, el porcentaje y el estado de los dispositivos existentes en el sistema de monitoreo. Se aprecia que existen **345 dispositivos activos** de los cuales se puede obtener información para el modelo de MD. Se descartan 274 dispositivos inactivos, debido a que no presentan datos de monitoreo históricos dentro del período analizado.

Figura 18

Cantidad de dispositivos por estado

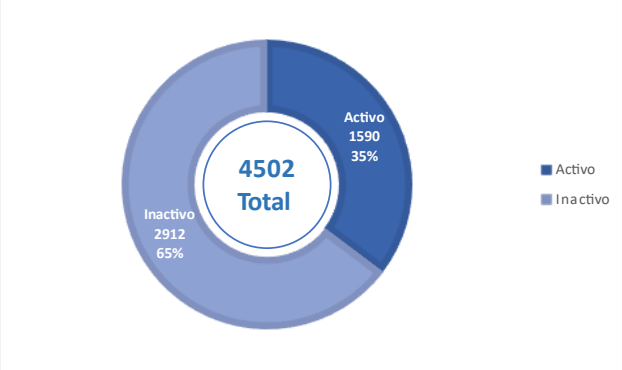


Nota. Muestra la cantidad y porcentaje de dispositivos por estado.

La Figura 19, presenta la cantidad, el porcentaje y el estado de los sensores de red (únicamente de los dispositivos activos) de NMS. Existen **1590 sensores activos** que se van a utilizar para el modelo de MD.

Figura 19

Cantidad de sensores por estado

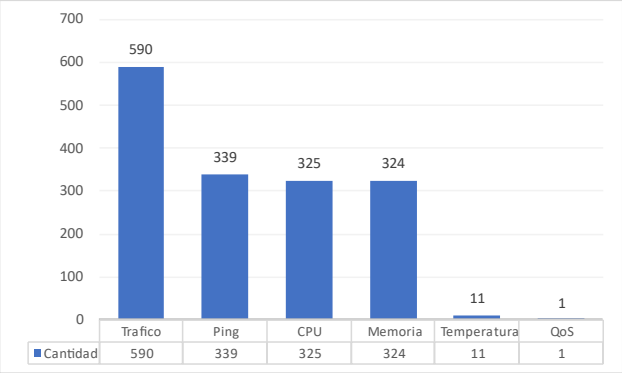


Nota. Sensores obtenidos de los dispositivos activos.

La Figura 20, muestra la distribución en cantidad de los 1590 sensores de red activos por su tipo: 590 de Tráfico, 339 de Ping, 325 de CPU y 324 de Memoria, estos tipos de sensores son los que se van a utilizar en el proceso de MD. Se descartan por su baja representación los sensores de Temperatura y QoS.

Figura 20

Cantidad de sensores activos por tipo

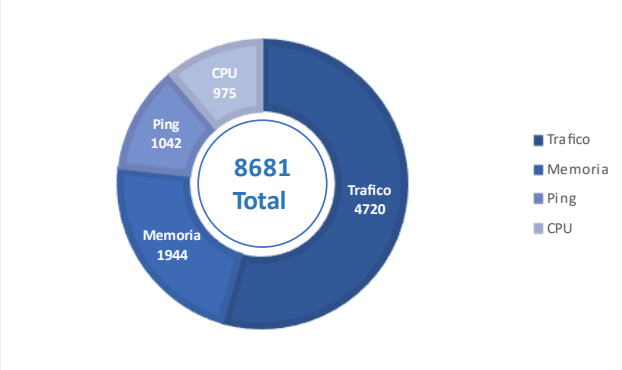


Nota. Sensores obtenidos de los dispositivos activos.

Exploración de Entidades: Canales. La Figura 21, muestra que existen 8681 canales disponibles de monitoreo (descartando los canales de temperatura y QoS) y su distribución por tipo de sensor. Se aprecia que el tipo de sensor Tráfico cuenta con 4720 canales, la Memoria con 1944 canales y finalmente el CPU y el Ping con 1042 y 975 canales respectivamente.

Figura 21

Cantidad de canales por tipo de sensor

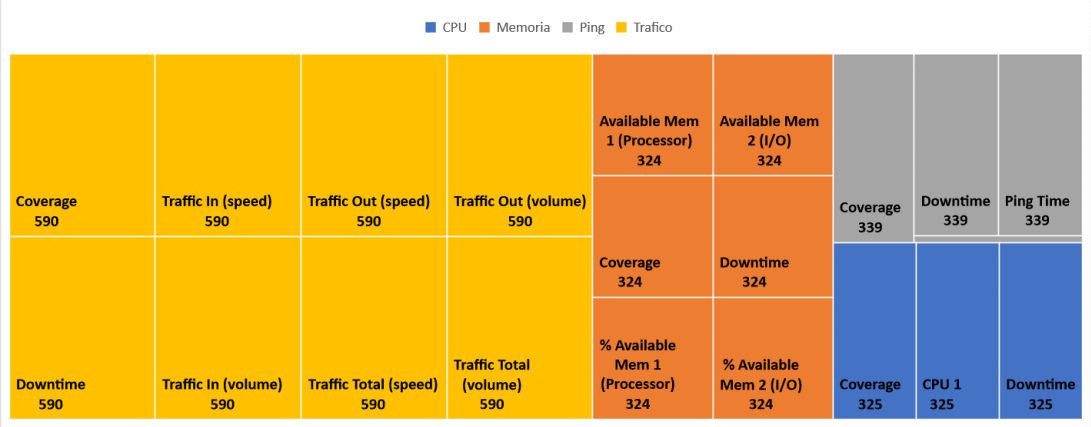


Nota. Se descarta los canales de temperatura y QoS.

La Figura 22, presenta la distribución de los canales agrupados por cada tipo de sensor. Se han considerado únicamente los dispositivos y sensores activos del sistema NMS. Los canales se correlacionan directamente a cada sensor y, por lo tanto, difieren unos de otros por su tipo, unidad de medida y valor.

Figura 22

Cantidad de canales por tipo de sensor detallado



Nota. Se han considerado los dispositivos y sensores activos.

Se ha analizado la información de los canales monitoreados y se detecta que no todos los canales presentan información, esto quiere decir que, existen valores nulos. Por lo tanto, se seleccionan para el modelo de MD, únicamente los canales con valores de monitoreo. La Tabla 11, muestra los canales seleccionados con su descripción y cantidad:

Tabla 11

Canales por sensor seleccionados para el proceso de MD

Sensor	Canal	Descripción	Unidad de medida	Cantidad
Ping	Ping Time	Mide la latencia que existe al enviar un paquete ICMP y su respuesta, desde un dispositivo origen hasta un dispositivo destino	Milisegundos	339
Ping	Downtime	Mide el porcentaje de indisponibilidad del dispositivo.	Porcentaje	339
CPU	CPU	Mide el porcentaje de utilización del CPU del dispositivo	Porcentaje	325
Memoria	% Available Mem_1 (processor)	Mide el porcentaje de Memoria_1 (por operaciones de procesamiento) disponible en el dispositivo	Porcentaje	324
Memoria	% Available Mem_2 (I/O)	Mide el porcentaje de memoria_2 (por operaciones de entrada / salida) disponible en el dispositivo	Porcentaje	324
Tráfico	Traffic In (volume)	Mide el volumen de datos de entrada a las interfaces de red del dispositivo	MByte	590
Tráfico	Traffic In (speed)	Mide la velocidad de entrada de datos a las interfaces de red del dispositivo	Mbit/s o Mbps	590

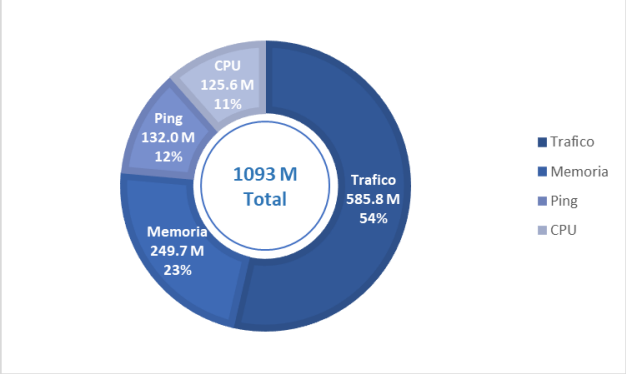
Sensor	Canal	Descripción	Unidad de medida	Cantidad
Tráfico	Traffic Out (volume)	Mide el volumen de datos de salida desde las interfaces de red del dispositivo	MByte	590
Tráfico	Traffic Out (speed)	Mide la velocidad de salida de datos desde las interfaces de red del dispositivo	Mbit/s o Mbps	590

Nota. Estos canales se van a utilizar en el modelo de MD.

Exploración de Entidades: Monitoreo. La Figura 23, muestra la cantidad de registros (1093.0 M) de monitoreo de red entre el 01 de marzo y el 20 de junio del 2020, distribuido por cada tipo de sensor. Se puede notar que, la distribución va acorde a la cantidad de canales por sensor y es así como el tráfico con 4 canales representa el 54 %, la memoria con 2 canales representa el 23% y el Ping y CPU con un canal representan el 12% y 11% respectivamente.

Figura 23

Cantidad de monitoreo por tipo de sensor



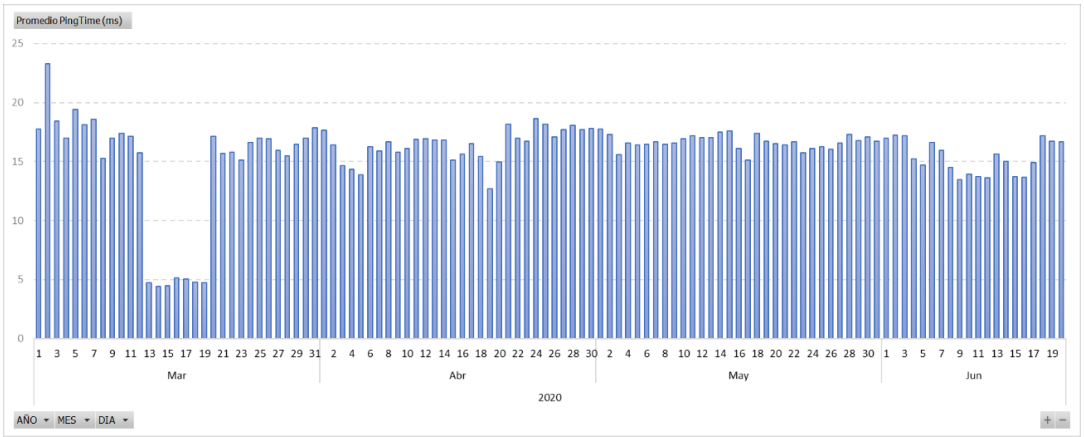
Nota. Valores en millones de registros (M).

La exploración de los datos de monitoreo se va a realizar por tipo de sensor y canales de red, con la finalidad de encontrar patrones y tendencias en el tiempo que ayuden a seleccionar los datos que pueden influir en el modelo predictivo de MD.

La Figura 24, presenta el promedio de los valores del canal PingTime (sensor ping) por día. La gráfica indica que, no existe una variación significativa en el tiempo, por lo tanto, se estima que esta variable no influya en el modelo de MD.

Figura 24

Evolución diaria de monitoreo del canal PingTime

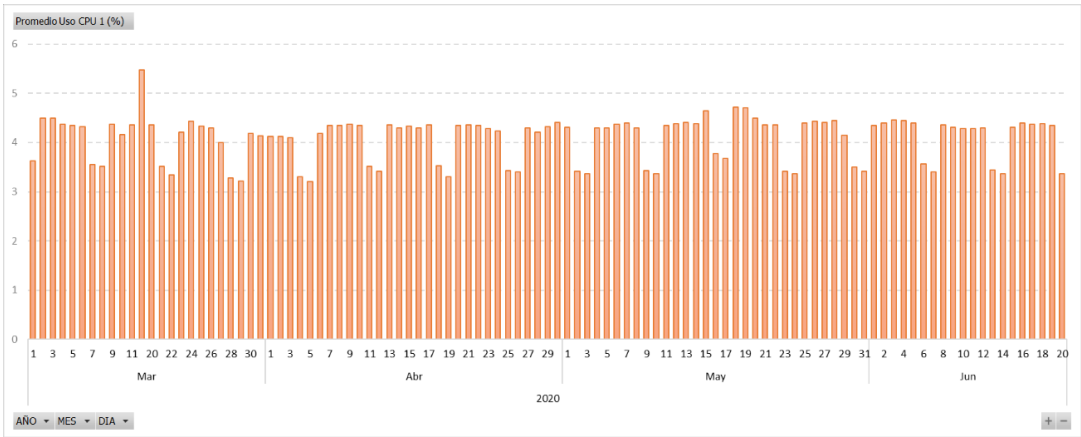


Nota. Valores indican el promedio diario en milisegundos (ms).

La Figura 25, presenta el promedio del porcentaje de utilización del CPU por cada día. Se aprecia que, existe un patrón de variación por día de la semana en los datos. Es decir, los días de la semana laborables sube el porcentaje de uso del CPU y los fines de semana este valor baja. Por esta razón, se estima que esta variable sí podría influir en el modelo de MD.

Figura 25

Evolución diaria de monitoreo del canal CPU

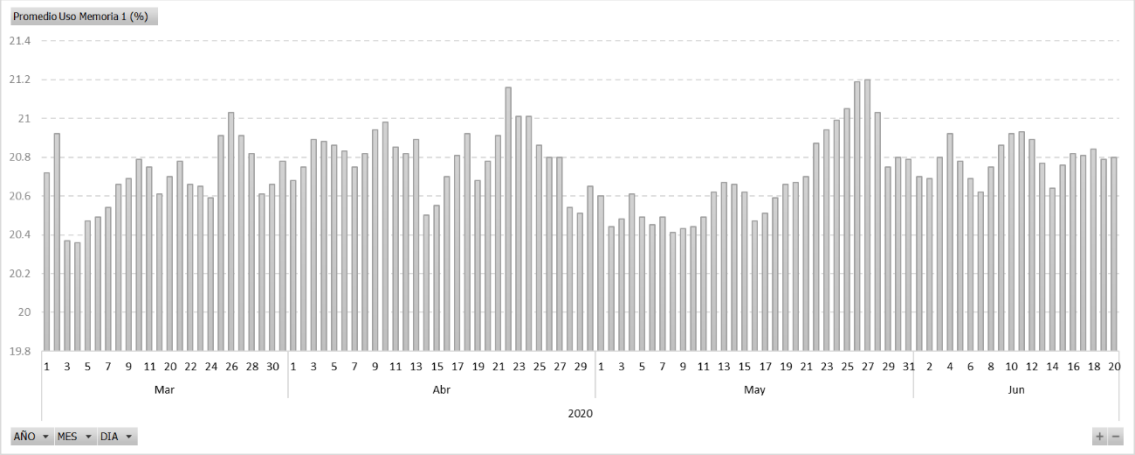


Nota. Valores indican el promedio diario en porcentaje de uso (%).

La Figura 26, muestra el promedio del porcentaje de uso de la Memoria_1 (ver Tabla 11), existe una variación sin un patrón establecido. Sin embargo, se mantiene este canal para realizar pruebas del modelo de MD y establecer si influye o no en los resultados.

Figura 26

Evolución diaria de monitoreo del canal Memoria_1

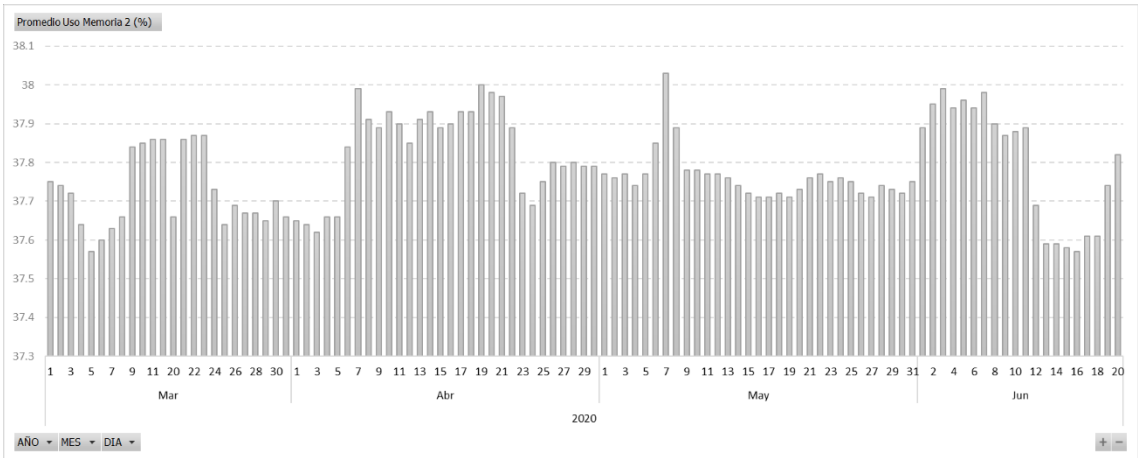


Nota. Valores indican el promedio diario en porcentaje de uso (%).

La Figura 27, muestra el promedio del porcentaje de uso de la Memoria_2 (ver Tabla 11), existe una variación sin un patrón establecido. Sin embargo, se mantiene este canal para realizar pruebas del modelo de MD y establecer si influye o no en los resultados.

Figura 27

Evolución diaria de monitoreo del canal Memoria 2

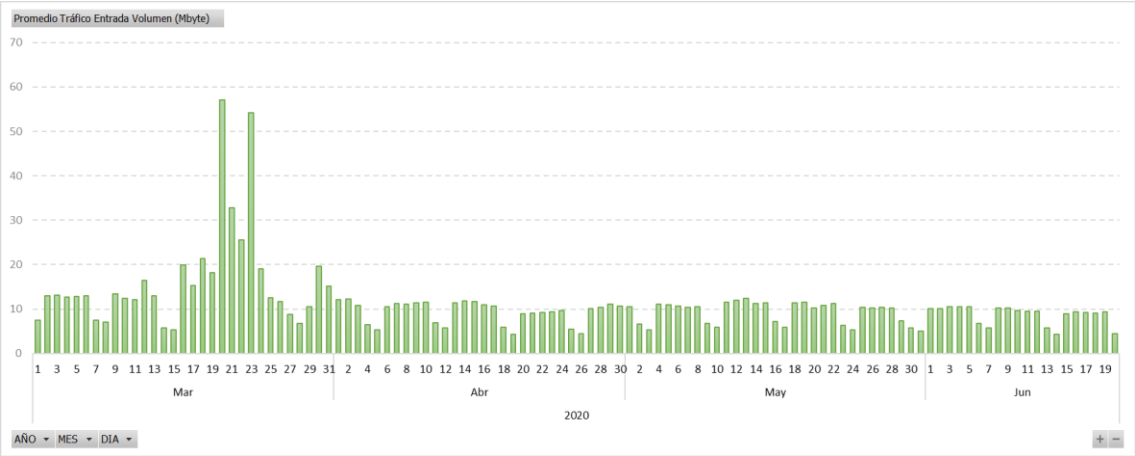


Nota. Valores indican el promedio diario en porcentaje de uso (%).

La Figura 28, presenta el promedio de volumen de tráfico de entrada (MByte) por día. Se aprecia que, existe un patrón de variación por día de la semana en los datos. Es decir, los días de la semana laborables sube el tráfico y los fines de semana los valores bajan. Por esta razón, se estima que esta variable podría influir en el modelo de MD.

Figura 28

Evolución diaria de monitoreo del canal Tráfico de Entrada (volumen)

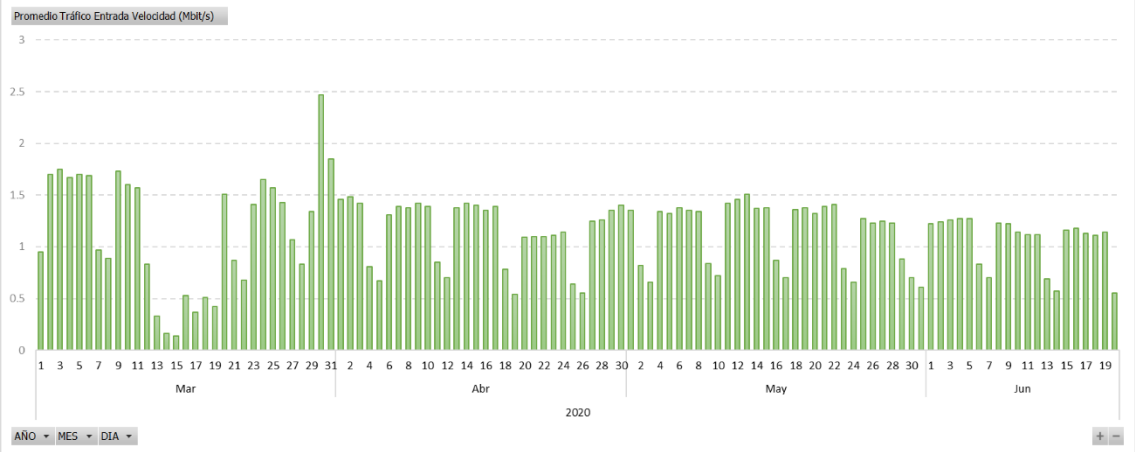


Nota. Valores indican el promedio en Megabytes (MByte).

La Figura 29, presenta el promedio diario de velocidad de tráfico de entrada (Mbit/s o Mbps). Según estos datos, existe un patrón de variación por día de la semana. Es decir, los días laborables suben el tráfico y los fines de semana este valor baja. Por esta razón, se estima que esta variable sí podría influir en el modelo de MD.

Figura 29

Evolución diaria de monitoreo del canal Tráfico de Entrada (velocidad)

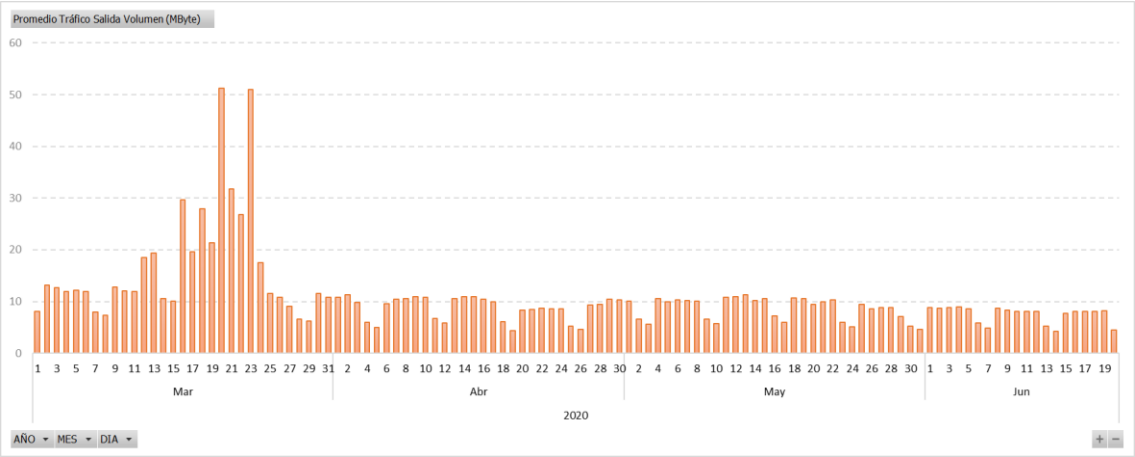


Nota. Valores indican el promedio en Megabits por segundo (Mbit/s).

La Figura 30, presenta el promedio de volumen de tráfico de salida (MByte) por día. Se aprecia que, existe un patrón de variación por día de la semana en los datos. Es decir, los días de la semana laborables sube el tráfico y los fines de semana este valor baja. Por esta razón, se estima que esta variable sí podría influir en el modelo de MD.

Figura 30

Evolución diaria de monitoreo del canal Tráfico de Salida (volumen)

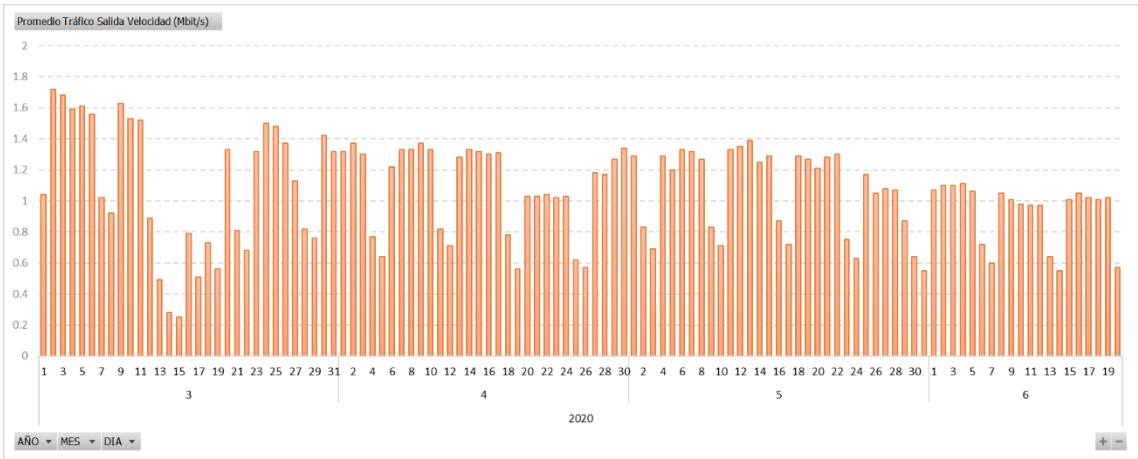


Nota. Valores indican el promedio en Megabytes (MByte).

La Figura 31, presenta el promedio diario de velocidad de tráfico de salida (Mbit/s o Mbps). Existe un patrón de variación por día de la semana en los datos. Es decir, los días laborables suben el tráfico y los fines de semana este valor baja. Por esta razón, se estima que esta variable sí podría influir en el modelo de MD.

Figura 31

Evolución diaria de monitoreo del canal Tráfico de Salida (velocidad)

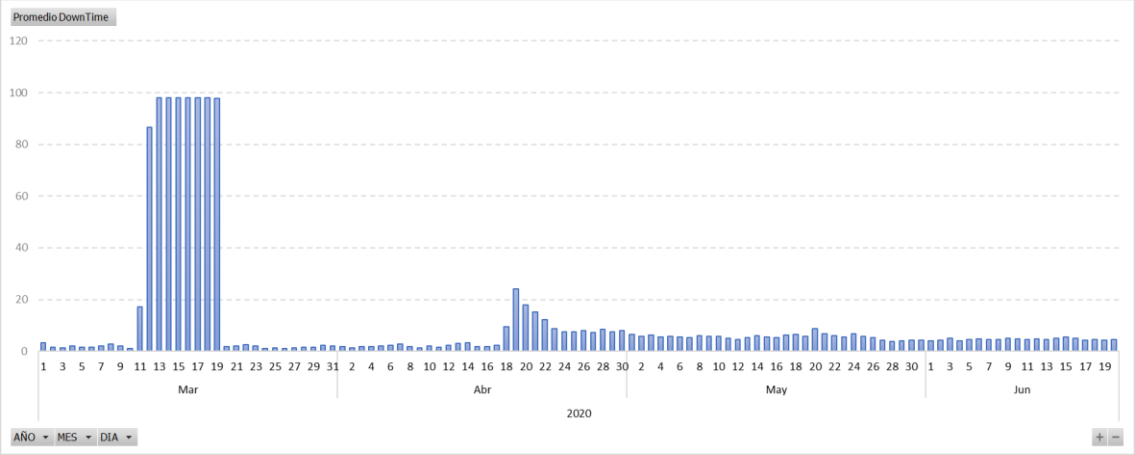


Nota. Valores indican el promedio en Megabits por segundo (Mbit/s).

El promedio diario de Downtime indica el porcentaje de indisponibilidad del dispositivo, y se puede interpretar como un valor contrario a la disponibilidad. Por lo tanto, esta medida forma parte de la variable objetivo del modelo predictivo de MD, el mismo que se detalla más adelante (ver sección *Creación de la variable objetivo "falla"* de la Fase 3 de CRISP-DM). La Figura 32, presenta las mediciones de Downtime entre marzo y junio del 2020. Sin embargo, se aprecia que, entre los días 11 al 20 de marzo del 2020 existen valores atípicos de indisponibilidad, por esta razón, se descarta esta información para el entrenamiento del modelo de MD.

Figura 32

Evolución diaria de monitoreo del canal Downtime



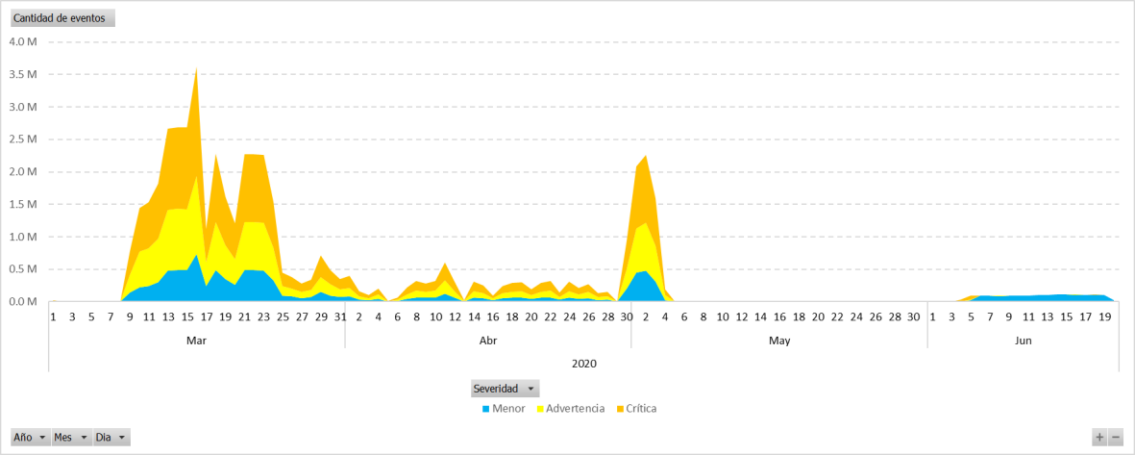
Nota. Valores indican el porcentaje diario de indisponibilidad (%).

Exploración de Entidades: Eventos y Umbrales. La Figura 33, muestra la cantidad diaria de eventos alertados por el sistema NMS (total 50.21 Millones), los cuales son de diferente tipo de impacto: menor, advertencia y crítica. Se puede visualizar la información entre los meses de marzo y junio del año 2020. Sin embargo, existen datos atípicos en la mayor parte del mes de marzo, esto coincide con lo analizado y apreciado en la sección anterior. Adicionalmente, en reuniones mantenidas con el personal de la empresa de telecomunicaciones, esta variación se debe a saturación y carga del sistema en el inicio de la pandemia de la Covid-19. Por lo tanto, se toma la decisión de descartar los datos de marzo del

2020, debido a que esta información no tiene un flujo normal, y puede generar un modelo de MD sesgado con un porcentaje de predicción incorrecto.

Figura 33

Evolución diaria de cantidad de eventos por severidad

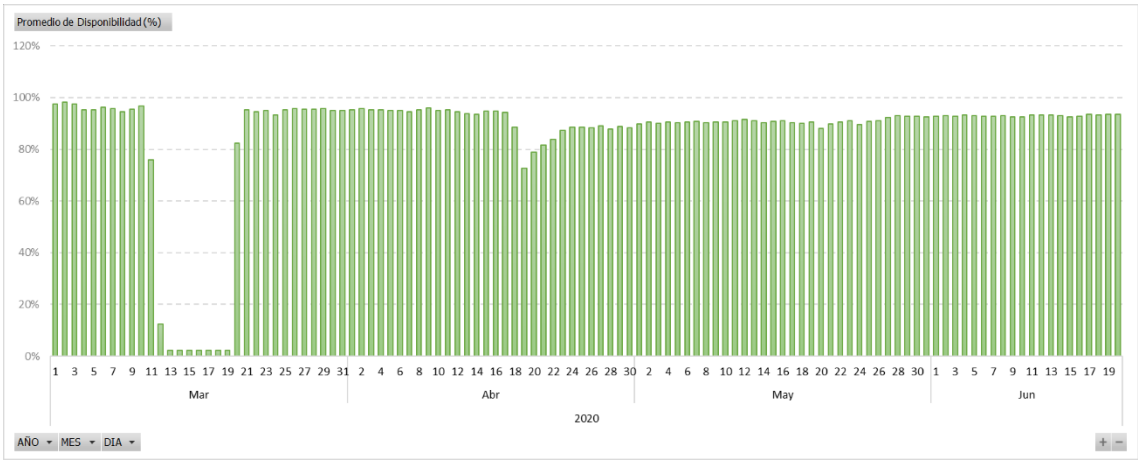


Nota. Valores en millones de registros (M).

Exploración de Entidades: Disponibilidad. La Figura 34, muestra el porcentaje de disponibilidad diario entre marzo y junio del 2020. Se aprecia que, entre los días 11 al 20 de marzo del 2020 existen valores atípicos y por tal razón esta información se va a descartar para el modelo de MD. Adicionalmente, esta información inicial es fundamental para comparar si el modelo de MD efectivamente ayudará en el mejoramiento de este indicador.

Figura 34

Evolución diaria de porcentaje de disponibilidad



Nota. Valores indican el promedio diario de disponibilidad (%).

Verificación de calidad de datos

Posterior a la exploración inicial de los datos se concluye que: Los datos están completos y almacenados en una base de datos SQL Server; Los datos encontrados cubren los requerimientos para obtener los resultados necesarios y cumplir los objetivos del proyecto de MD. Los datos no contienen errores ya que son los ingresados por el software NMS automáticamente; no se encuentran datos fuera de rango lo cual elimina posibles ruidos para el modelo de MD. Existen registros con valores nulos de monitoreo en la tabla [RESULTADO_ABSOLUTO], reflejando que pueden existir pérdidas en el monitoreo y/o cortes de enlace.

En general con la exploración realizada se verifica que los datos son válidos, completos y sin errores para el proceso de MD. Adicionalmente, existe una cantidad extensa de información con monitoreo por minuto para la generación pruebas y evaluación del modelo de MD.

Fase 3: Preparación de los datos

Esta fase de la metodología CRISP-DM permite preparar los datos para que se adecúen a las técnicas de MD. Por lo tanto, se debe seleccionar los subconjuntos de datos que se van a utilizar, realizar una limpieza de estos para mejorar su calidad, construir nuevos datos, integrarlos y formatearlos para ser utilizado por las aplicaciones de modelado de Minería de Datos.

Selección de datos

En función a la recopilación de datos inicial realizada, se procede a seleccionar los datos relevantes alineados a la consecución del modelo predictivo de MD y su posterior evaluación. A continuación, se lista la selección de elementos a nivel de registros realizada:

- Datos de grupos, dispositivos y sensores de red activos en el sistema de monitoreo
- Datos de canales y captura de monitoreos realizados entre los meses de abril y junio del 2020
- Datos de umbrales estáticos activos
- Datos de eventos suscitados entre abril y junio del 2020
- Información de disponibilidad obtenida entre abril y junio del 2020

Se debe considerar que no todos los campos de las entidades seleccionadas son necesarios, debido a que no aportan valor en la consecución de los objetivos de MD y, por lo tanto, no serán considerados en esta selección. A continuación, se presenta las tablas y campos de base de datos, seleccionados para el proceso de MD:

Tabla 12

Tablas y campos seleccionados para el proceso de MD

Tabla	Campo	Observación
EST_ESTRUCTURA_VIEW	GrupoNombre	Nombre de la agrupación de los dispositivos de red
	DispositivoNombre	Nombre del dispositivo de red

Tabla	Campo	Observación
CAN_ESTRUCTURACANAL_VIEW	DispositivoEstado	Estado del dispositivo de red (activo / inactivo)
	SensorNombre	Nombre del sensor de red
	SensorId	Código identificador del sensor de red
	SensorEstado	Estado del sensor de red (activo / inactivo)
	SensorTipo	Tipo de sensor de red (CPU, memoria, ping y tráfico)
	EstructuraCanalId	Código identificador que une el sensor de red con cada canal de monitoreo
	CanalNombre	Nombre del canal de monitoreo
	Unidad	Unidad de medida
	UnidadOperacion	Operación para el cálculo de la unidad
	UnidadOperacionValor	Valor para el cálculo de la unidad
RESULTADO_ABSOLUTO	ESCA_IN_CODIGO	Código identificador que une el sensor de red con cada canal de monitoreo
	RESU_DT_FECHA	Fecha y hora que se toma el valor de la medida del monitoreo
	RESU_FL_VALOR	Valor de la medida del monitoreo del canal y sensor
EVENTOS	EVEN_IN_CODIGO	Código identificador de los eventos
	ESCA_IN_CODIGO	Código identificador que une el sensor de red con cada canal de monitoreo
	EVEN_DT_FECHA	Fecha y hora en que se sucitó el evento de red
	EVEN_FL_METRICA	Valor de la medida del monitoreo del canal y sensor en que se sucitó el evento de red
	EVEN_IN_STATUS	Estado del evento de red (abierto o cerrado)
	EVEN_IN_SEVERIDAD	Severidad de impacto del evento de red (crítica, alerta o menor)
UMBRAL_CANAL	ESCA_IN_CODIGO	Código identificador que une el sensor de red con cada canal de monitoreo
	UMBR_IN_CODIGO	Código que identifica cada umbral estático
	ENUM_IN_TIPO_SEVERIDAD	Código que indica la severidad de impacto de sobrepasar el umbral
	UMBR_FL_TIEMPO	Medida de tiempo que el umbral debe permanecer fuera de rangos normales
	UMBR_FL_LIMITE	Valor de medida límite del umbral
	UMBR_A2_COMPARADOR	Indica la comparación aritmética que se hace entre la medida obtenida por el monitoreo y la medida del umbral definido
	ENUM_IN_ESTADO_UMBRAL	Estado del umbral (activo / inactivo)
SENSOR_DISPONIBILIDAD	ESCA_IN_CODIGO	Código identificador que une el sensor de red con cada canal de monitoreo
	SDIS_DT_FECHA	Fecha y hora para el cálculo de la disponibilidad
	SDIS_FL_VALOR	Valor promedio de la medida del monitoreo del canal y sensor por cada hora del día

Nota. Presenta las estructuras de base de datos seleccionadas.

Limpieza de datos

Para esta tarea es necesario determinar los datos perdidos, erróneos, incoherentes y la protección de estos.

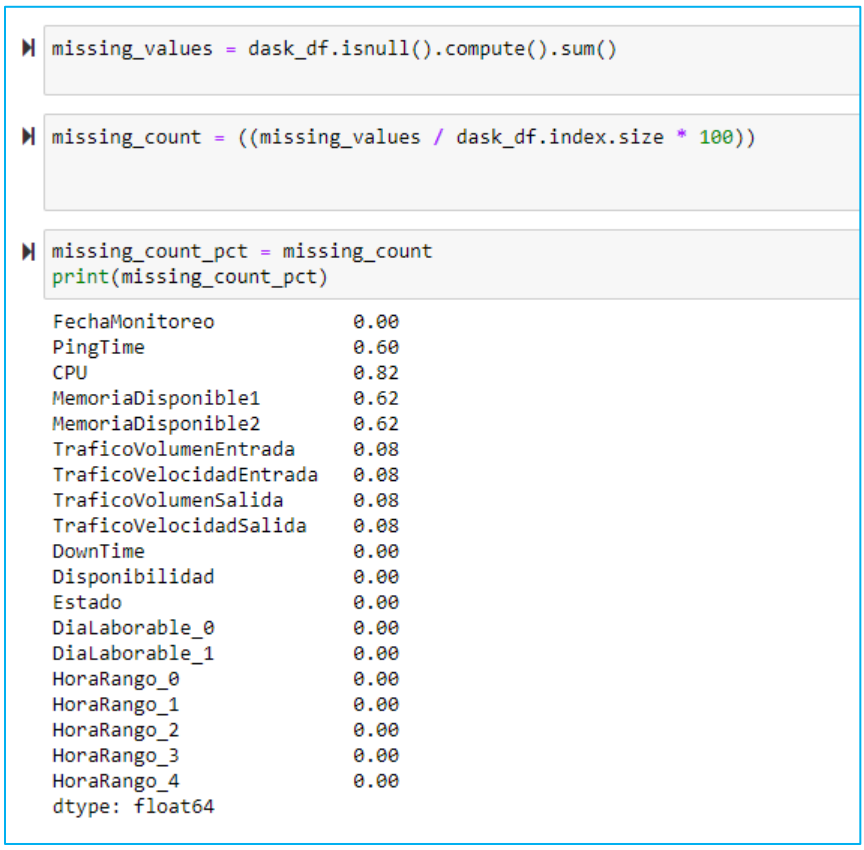
Datos perdidos. Existen datos perdidos de monitoreo de los dispositivos que se inactivaron temporalmente por cuestiones de mantenimiento o decisiones que dependieron del cliente o proveedor y que posteriormente se activaron nuevamente. Adicionalmente, por temas de mantenimiento del propio sistema NMS y/o su proveedor existen datos históricos que se han movido a otras bases de datos a las que no se tiene acceso por parte del autor del proyecto.

Errores de datos. En la exploración realizada a nivel de datos, se ha determinado que el modelo de base de datos contiene: tablas, relaciones de campos, claves primarias, claves foráneas adecuados y normalizados, esto minimiza el riesgo de encontrar datos erróneos. Sin embargo, puede existir datos erróneos debido al ingreso, agrupación de dispositivos por parte de los operadores del sistema NMS, esto no es corregible ya que no se tiene un parámetro que permita detectar o depurar esta información.

Mediante el uso de técnicas de limpieza de datos proporcionadas en el lenguaje Python, se procede a limpiar datos nulos encontrados. La Figura 35, presenta el código para encontrar los valores nulos y perdidos, se imprime en pantalla los porcentajes de nulos en comparación al total de datos por cada columna del Conjunto de datos.

Figura 35

Código Python que presenta los datos nulos



```
missing_values = dask_df.isnull().compute().sum()
```

```
missing_count = ((missing_values / dask_df.index.size * 100))
```

```
missing_count_pct = missing_count
print(missing_count_pct)
```

```
FechaMonitoreo            0.00
PingTime                  0.60
CPU                       0.82
MemoriaDisponible1        0.62
MemoriaDisponible2        0.62
TraficoVolumenEntrada     0.08
TraficoVelocidadEntrada   0.08
TraficoVolumenSalida      0.08
TraficoVelocidadSalida    0.08
DownTime                  0.00
Disponibilidad            0.00
Estado                    0.00
DiaLaborable_0            0.00
DiaLaborable_1            0.00
HoraRango_0               0.00
HoraRango_1               0.00
HoraRango_2               0.00
HoraRango_3               0.00
HoraRango_4               0.00
dtype: float64
```

Nota. Valores medidos en porcentaje. Código Python en Jupyter Notebook.

Por otro lado, La Figura 36, muestra el código Python aplicado para realizar la limpieza de los datos nulos, nótese que posterior a este proceso el porcentaje de estos baja a 0%.

Figura 36

Código Python de limpieza de datos nulos

```
dask_df = dask_df.fillna(dask_df.mean(numeric_only=True))
```

```
missing_values = dask_df.isnull().compute().sum()
missing_count = ((missing_values / dask_df.index.size * 100))
missing_count_pct = missing_count
print(missing_count_pct)
```

```
FechaMonitoreo            0.00
PingTime                  0.00
CPU                       0.00
MemoriaDisponible1        0.00
MemoriaDisponible2        0.00
TraficoVolumenEntrada     0.00
TraficoVelocidadEntrada   0.00
TraficoVolumenSalida      0.00
TraficoVelocidadSalida    0.00
DownTime                  0.00
Disponibilidad            0.00
Estado                    0.00
DiaLaborable_0            0.00
DiaLaborable_1            0.00
HoraRango_0               0.00
HoraRango_1               0.00
HoraRango_2               0.00
HoraRango_3               0.00
HoraRango_4               0.00
dtype: float64
```

Nota. Valores medidos en porcentaje. Código Python en Jupyter Notebook.

Incoherencia de datos. Basado en la exploración realizada, no se ha encontrado datos incoherentes ni en los datos relacionados entre estructuras ni en las mediciones de monitoreo.

Protección de datos. Debido a los convenios de confidencialidad y seguridad que la empresa proveedora de telecomunicaciones mantiene con sus clientes corporativos, se ha cambiado los nombres de grupos, dispositivos y sensores por nombres genéricos, los cuales se han concebido de la siguiente forma:

Tabla 13

Protección y confidencialidad de datos de clientes corporativos

#	Tipo de estructura	Ejemplos de nombres nuevos
C1	Grupos	E_625 E_345 E_6147
C2	Dispositivos	D_178 D_16789 D_37017
C3	Sensores	Ping CPU Memoria Temperatura Trafico (nombre de la interfaz de red fìsica o lógica) (001) FastEthernet0 Traffic (014) WAN Traffic (012) Vlan1 Traffic

Nota. Cambio de nombres de grupos, dispositivos y sensores.

Construcción de nuevos datos

Para la consecución de los objetivos de MD, se procede a construir nuevos datos y visualizaciones.

Creación de vistas de base de datos. Se han creado vistas en la base de datos SQL Server que permiten agrupar y facilitar el acceso a la información, estas se listan a continuación:

- Vista [EST_GRUPO_VIEW]
- Vista [EST_DISPOSITIVO_VIEW]
- Vista [EST_SENSOR_VIEW]
- Vista [EST_ESTRUCTURA_VIEW]
- Vista [CAN_CANAL_VIEW]
- Vista [CAN_ESTRUCTURACANAL_VIEW]

Nota: El Apéndice A. Definición de nuevas estructuras de base de datos, contiene la descripción y scripts de creación de las vistas creadas.

Creación de tablas auxiliares de base de datos. Para la consecución del proceso de MD, es necesario crear tablas auxiliares que faciliten y mejoren el rendimiento de las consultas y filtros en la base de datos. Se han creado las siguientes tablas auxiliares:

- Tabla [MON_MONITOREOFECHAS_TABLE]
- Tabla [EVE_EVENTOS_TABLE]
- Tabla [DIS_DISPONIBILIDAD_TABLE]

Nota: Los Apéndices A, B, C del presente documento, contienen la especificación de los campos, scripts de creación y llenado de datos de las tablas auxiliares.

Creación de estructura horizontal de monitoreo. Considerando que la información de monitoreo se almacena en una estructura vertical en la tabla [RESULTADO_ABSOLUTO], se ha detectado la necesidad de transponer estos datos en una estructura horizontal. Es decir, los valores de monitoreo que actualmente se guardan en una columna por cada canal de red en un tiempo determinado (fecha y hora), se van a almacenar en varias columnas, agrupando los canales de red que pertenecen a un mismo dispositivo.

Para esta tarea se ha creado la tabla [MON_MONITOREO_TABLE]. Con este cambio en la forma de estructurar los datos se pretende que, cada columna de esta tabla se establezca como una entrada (input) al modelo de MD.

La Figura 37, presenta el Modelo Entidad Relación de la estructura horizontal de monitoreo planteada.

Figura 37

Modelo ER de Monitoreo horizontal

Nota. Modelo obtenido con la herramienta Database Diagram de SQL Server.

La Figura 38, presenta los campos y un extracto de la información almacenada en el Conjunto de Datos que se va a utilizar para la generación del modelo de MD (tabla [MON_MONITOREO_TABLE]).

Figura 38

Registros de la tabla [MON_MONITOREO_TABLE]

Results Messages

	FechaMonitoreo	Dispositivo	DiaLaborable	HoraRango	PingTime	CPU	MemoriaDisponible1	MemoriaDisponible2	TraficoVolumenEntrada	TraficoVelocidadEntrada	TraficoVolumenSalida	TraficoVelocidadSalida	DownTime	Disponibilidad	Estado
1	2020-04-01 15:00:00.000	D_61592	1	3	657	1	84.3049	63.5535	0.057	0.0037	0.0522	0.0034	0	1	0
2	2020-04-01 15:00:00.000	D_61601	1	3	13	3	83.8162	63.546	0.0685	0.0075	20.0458	2.1974	0	1	0
3	2020-04-01 15:00:00.000	D_61600	1	3	12	3	85.5351	63.5722	0.4539	0.0373	53.7218	4.4116	0	1	0
4	2020-04-01 15:00:00.000	D_1066	1	3	1	4	83.7782	63.4607	0.3385	0.0226	2.8481	0.1899	0	1	0
5	2020-04-01 15:00:00.000	D_7	1	3	7	6	83.1941	57.0069	401.8678	32.7766	313.7216	25.5	0	1	0
6	2020-04-01 15:00:00.000	D_61578	1	3	6	4	82.7838	56.9236	62.4587	4.8756	63.8619	4.8913	0	1	0
7	2020-04-01 15:00:00.000	D_1894	1	3	1	9	69.654	63.1401	47.2996	3.502	46.4499	3.6143	0	1	0
8	2020-04-01 15:00:00.000	D_61588	1	3	12	5	81.5508	56.1969	0.2068	0.0172	17.4995	1.4594	0	1	0
9	2020-04-01 15:00:00.000	D_61602	1	3	12	3	83.8107	63.5525	0.2088	0.0156	10.2489	0.7668	0	1	0
10	2020-04-01 15:00:00.000	D_3	1	3	1.7	3	77.9674	63.5722	0.0779	0.0056	0.0948	0.0069	0	1	0
11	2020-04-01 15:00:00.000	D_12	1	3	7	1	69.9089	71.2903	0.018	0.0014	0.0408	0.0032	0	1	0
12	2020-04-01 15:00:00.000	D_61546	1	3	10	9	83.6419	63.5458	0.9893	0.0747	9.0493	0.6831	0	1	0
13	2020-04-01 15:00:00.000	D_8	1	3	7	0	71.4773	72.5354	0.2494	0.0222	0.204	0.0215	0	1	0
14	2020-04-01 15:00:00.000	D_61580	1	3	10	2	83.7142	63.546	0.119	0.0096	9.7316	0.7866	0	1	0
15	2020-04-01 15:00:00.000	D_2193	1	3	2	5	82.8301	56.1968	1.3038	0.1255	0.0519	0.005	0	1	0
16	2020-04-01 15:00:00.000	D_61594	1	3	12	4	85.6576	63.545	0.1243	0.0126	32.5263	3.2976	0	1	0
17	2020-04-01 15:00:00.000	D_9	1	3	0	4	69.9582	67.8474	8.8065	0.8339	8.8216	0.8347	0	1	0
18	2020-04-01 15:00:00.000	D_10	1	3	1	3	70.5517	67.9353	0.0211	0.0022	0.0148	0.0016	0	1	0
19	2020-04-01 15:00:00.000	D_1070	1	3	1	7	79.7796	63.1401	107.3335	8.8887	114.0435	9.2837	0	1	0
20	2020-04-01 15:00:00.000	D_61547	1	3	18	5	82.8701	56.9227	61.138	7.0926	49.0346	6.0369	0	1	0

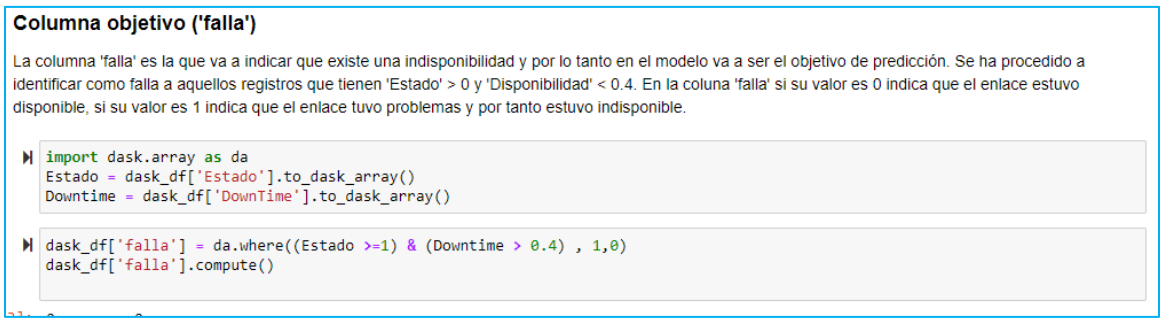
Nota. Datos ejemplos obtenidos en la herramienta SQL Server Management Studio.

Nota: Los Apéndices A, B, C del presente documento, contienen la especificación de los campos, scripts de creación y llenado de datos.

Creación de la variable objetivo “falla”. Para el modelo predictivo de MD es necesario tener una variable objetivo, la misma que permite identificar que se desea predecir. En el Conjunto de datos se procede a crear la columna “*falla*” (variable objetivo), esta indica de manera binaria si existe indisponibilidad (valor 1) o si la disponibilidad es normal (valor 0). La Figura 39, muestra el código Python que permite crear la variable “*falla*” para el Conjunto de datos.

Figura 39

Código Python de creación de la variable objetivo “falla”



Nota. Código Python en Jupyter Notebook.

Integración de datos

Debido a que el sistema NMS, almacena sus datos en un solo repositorio de información no es necesario realizar una integración de diferentes orígenes de datos.

Formato de datos

En esta sección se procede a dar formato a los datos según obtenidos, con la finalidad de llevar esta información a un análisis y modelado de MD.

Los valores recopilados en la tabla [MON_MONITOREO_TABLE] se obtuvieron directamente de la tabla [RESULTADO_ABSOLUTO]. Sin embargo, es necesario cambiar los valores de monitoreo, debido a que NMS en sus procesos de captura y recopilación del monitoreo almacena los datos crudos y sin formato. Al cambiar estos valores se establecen las

unidades de medida utilizadas en el lenguaje técnico y comercial de redes de telecomunicaciones.

Para dar formato por unidad de medida de monitoreo, es necesario hacer los cálculos de transformación basado en las operaciones de la vista [CAN_CANAL_VIEW] con sus campos:

- [Unidad]: Especifica la unidad de medida del monitoreo
- [UnidadOperacion]: Indica la operación aritmética que se debe aplicar a la medida del monitoreo (multiplicación / división).
- [UnidadOperacionValor]: Es el valor que se debe multiplicar o dividir a la medida de monitoreo

La Tabla 14, indica las operaciones que se deben aplicar por cada canal. Se debe considerar que, los canales CPU, Ping y Memoria en el momento de recolección y almacenamiento (por el sistema NMS) toma el valor real de la medida. Por lo tanto, su valor de medida se multiplica por 1 para que no exista ninguna variación en la medida. En cambio, para los canales de Tráfico es necesario realizar las operaciones de división por volumen y velocidad.

Tabla 14

Operaciones aritméticas para valores de monitoreo por canal

Canal Id	Canal Nombre	Sensor Tipo	Unidad	Unidad Operación	Unidad Operación Valor
104	Ping Time	Ping	mseg	MUL	1
154	CPU	CPU	%	MUL	1
206	Percent Available Memory_1 (Processor)	Memoria	%	MUL	1
210	Percent Available Memory_2 (I/O)	Memoria	%	MUL	1
361	Traffic In (volume)	Tráfico	MByte	DIV	1048576
363	Traffic In (speed)	Tráfico	Mbit/s o Mbps	DIV	131072
365	Traffic Out (volume)	Tráfico	MByte	DIV	1048576

Canal Id	Canal Nombre	Sensor Tipo	Unidad	Unidad Operación	Unidad Operación Valor
367	Traffic Out (speed)	Tráfico	Mbit/s o Mbps	DIV	131072

Nota. Cambio de nombres de grupos, dispositivos y sensores.

Nota: El Apéndice C. Scripts de consultas de base de datos, presenta los scripts de actualización de la tabla [MON_MONITOREO_TABLE].

Fase 4: Modelado

En esta fase de la metodología CRISP-DM se genera el modelo de Minería de Datos propuesto. En el caso particular del presente proyecto, se ha planteado obtener un modelo predictivo basado en la tendencia de comportamiento de los dispositivos, sensores y canales de red.

Selección de técnicas de modelado

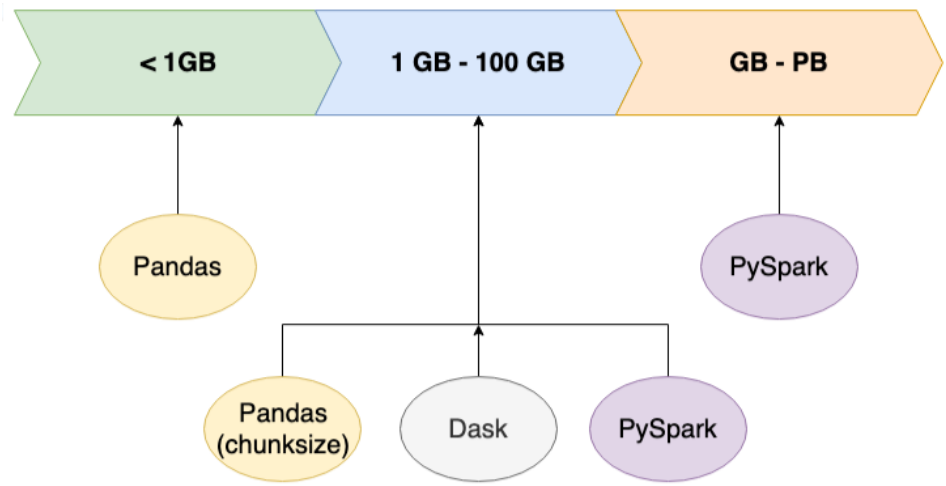
En esta sección se realiza la selección de la técnica que se va a utilizar para el modelo de Minería de Datos (MD). Para el presente proyecto, se ha decidido utilizar **Árboles de decisión**, esto debido al análisis de las técnicas de MD para la predicción de umbrales de sensores de red realizado en el apartado *Resultados del estado del arte* del *Capítulo III*. Considerando que, adicionalmente se ha planteado el uso de la herramienta Anaconda Dsitribution (Python) (ver *Herramientas de Minería de Datos* del *Capítulo II*), es necesario utilizar librerías y paquetes de este lenguaje, que permitan generar el modelo con la técnica escogida y que hagan uso de gran cantidad de datos. A continuación, se especifican las librerías de Python utilizadas:

Dask. Es una librería que permite procesamiento paralelo en Python. Al utilizar gran cantidad de datos, para esto se hace necesario particionar y procesar la información en paquetes más pequeños, haciendo que los recursos de computador sean utilizados eficientemente en paralelo y con gran escalabilidad (Anaconda, Inc. and contributors, 2018). La Figura 40, presenta un comparativo entre librerías para el manejo de información por su

tamaño, se aprecia que Dask es aconsejable para utilizar en datos que tienen entre 1 y 100 GB de tamaño (Zhang A. , 2019). En el presente proyecto se utiliza Dask ya que los datos poseen un tamaño de alrededor de 6 GB con 40 millones de registros:

Figura 40

Comparación de librerías Python por el tamaño de datos

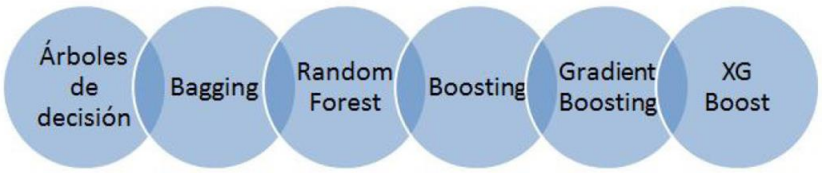


Nota. Tamaños comparados en GB (Gigabytes). Tomado de *Data Driven Investor*, por A. Zhang, 2019.

XGboost. Es una librería de Python que permite utilizar el algoritmo XG Boost (Extreme Gradient Boosting) que es una técnica de aprendizaje supervisado, resultado de una evolución de los Árboles de decisión (Espinoza Zúñiga, 2020).

Figura 41

Evolución de los algoritmos basados en árboles de decisión



Nota. Adaptado de *Application of Random Forest and XGBoost algorithms based on a credit card applications database*, por J. Espinoza, 2020.

Entre las principales características de XG Boost se tiene que: Es un ensamblado en secuencia de árboles de decisión denominado CART (Classification and Regression Trees por sus siglas en inglés), los árboles se agregan secuencialmente y aprenden de los árboles anteriores, corrigiendo el error, esto lo realiza sucesivamente hasta que ya no puede corregir dicho error (gradiente descendente); Utiliza procesamiento paralelo, realiza poda de árboles, maneja valores perdidos (Chen & Guestrin, 2016).

La técnica seleccionada para la generación del modelo de MD es el algoritmo **XG Boost** con las librerías de **Python Dask** y **XGboost,** por las características descritas y su adaptación a los objetivos propuestos en el proyecto de MD.

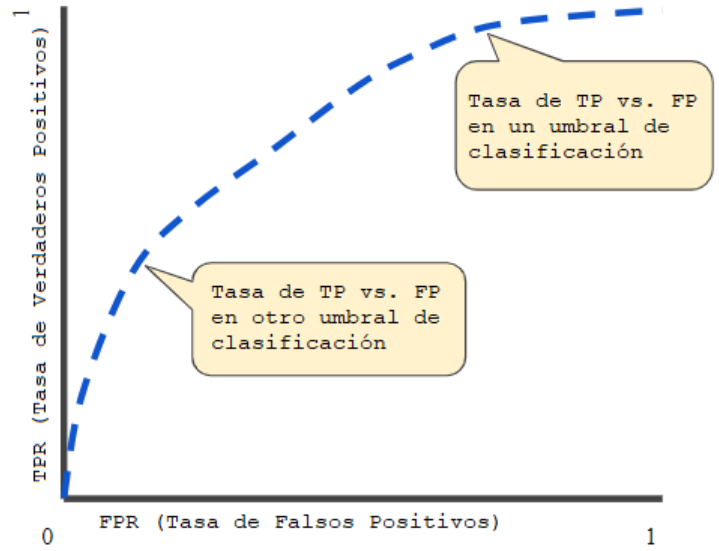
Generación de un diseño de comprobación

Para probar la calidad y validez del modelo se va a utilizar ROC – AUC que es una métrica de evaluación estadística para problemas de clasificación binaria, sus abreviaturas provienen de AUC (Area Under the Curve por sus siglas en inglés) y ROC (Receiver Operating Characteristic por sus siglas en inglés) (Bhandari, 2020).

Curva ROC. Muestra gráficamente la probabilidad de rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Representa dos parámetros: La Taza de Verdaderos Positivos TPR (True Positive Rate por sus siglas en inglés) y La Taza de Falsos Positivos FPR (False Positive Rate por sus siglas en inglés). La Figura 42, presenta un ejemplo para interpretar la Curva ROC, que mide el TPR vs. el FPR en diferentes umbrales de clasificación, el conjunto de estos valores (entre 0 y 1) se muestra como una línea entre los ejes x, y (Google Developers, 2022).

Figura 42

Ejemplo Curva ROC



Nota. TP (Verdadero Positivo) y FP (Falso Positivo). Adaptado de *Machine Learning Clasificación: Curva ROC y AUC*, por Google Developers, 2022.

Con la anterior definición, es posible calcular la sensibilidad y especificidad para cada punto de corte, mediante las siguientes fórmulas:

$$\text{Sensibilidad} = TP / (TP + FN)$$

$$\text{Especificidad} = TN / (FP + TN)$$

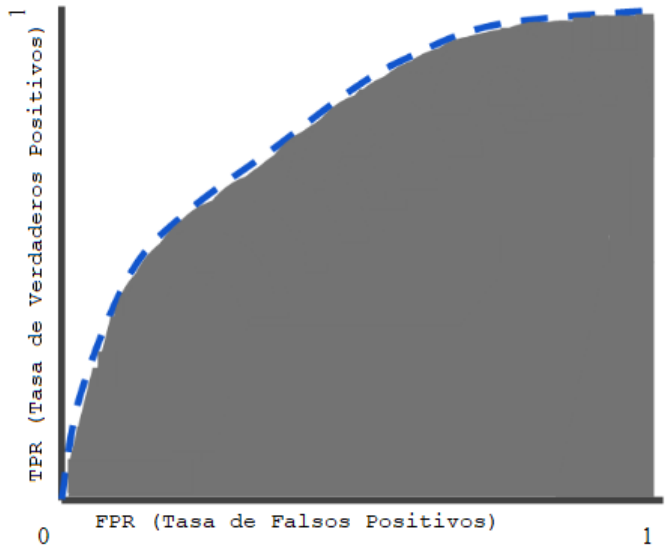
Dónde, TP = Verdaderos Positivos, FN = Falsos Negativos, TN = Verdaderos Negativos y FP = Falsos Positivos.

Esto se realiza, al comparar los valores reales versus los valores obtenidos con el modelo de predicción a través de una matriz de confusión, entre más sensible (mayor número de Verdaderos Positivos TP) y más específico (mayor número de Verdaderos Negativos TN) el modelo se puede interpretar que tiene una probabilidad de acierto mayor (Bhandari, 2020).

AUC. Es toda el Área Bajo la Curva ROC, proporciona una medida del rendimiento en todos los umbrales de clasificación posibles (Bhandari, 2020).

Figura 43

Ejemplo AUC



Nota. Presenta el Área Bajo la Curva ROC. Adaptado de *Machine Learning Clasificación: Curva ROC y AUC*, por Google Developers, 2022.

Se concluye que, “un valor más alto en el eje X indica un mayor número de Falsos Positivos (FP) que Verdaderos Negativos (TN), mientras que un valor más alto en el eje Y indica una mayor cantidad de Verdaderos Positivos (TP) que Falsos Negativos (FN). La elección del umbral depende de la capacidad de equilibrar entre Falsos Positivos (FP) y Falsos Negativos (FN)” (Bhandari, 2020).

Generación de los modelos

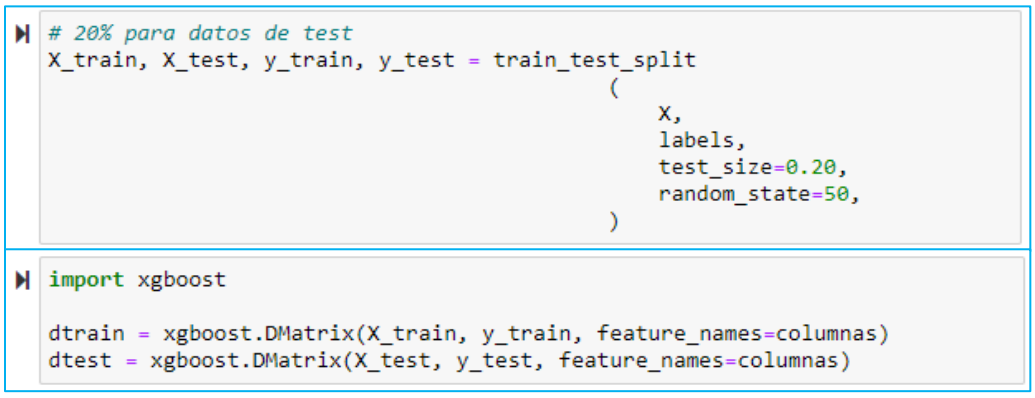
Según lo indicado en el apartado *Selección de técnicas de modelado*, se va a utilizar el algoritmo **XG Boost** que es una técnica de predicción no supervisada, evolucionada de los Árboles de Predicción. El modelo se ha programado en el IDE **Jupyter Notebook** de Anaconda Distribution (Python).

El conjunto de datos total para el modelo de MD se ha obtenido mediante el proceso realizado en las fases previas de CRISP-DM (secciones *Fase 2: Entendimiento de los datos* y *Fase 3: Preparación de los datos*). Estos datos contienen información de monitoreo de red desde el 01 de marzo al 20 de junio del año 2020. Debido a problemas en el software NMS por saturación y carga de información por la COVID-19 formulado por el personal técnico de la compañía proveedora de telecomunicaciones, se procede a eliminar los datos de marzo y de inicios de abril del año 2020, ya que no tienen un patrón normal de recopilación de monitoreo.

Para el Conjunto de Datos de entrenamiento (train) y prueba (test) se ha tomado datos de los dispositivos entre el 15 de abril y el 15 de mayo del año 2020, esto debido al análisis exploratorio realizado en la Fase 3 de CRISP-DM y corroborado en reuniones mantenidas con el personal de la empresa de telecomunicaciones, quienes indican que en este período de tiempo el sistema NMS se encuentra estable. Se ha realizado una partición del 80 % y 20 % para entrenamiento (dtrain) y test (dtest) respectivamente. Es importante mencionar que estos datos se encuentran cargados en memoria utilizando la librería Dask de Python que brinda escalabilidad y paralelismo en el procesamiento.

Figura 44

Partición y carga del Conjunto de Datos (train, test)



```
# 20% para datos de test
X_train, X_test, y_train, y_test = train_test_split
                                                (
                                                    X,
                                                    labels,
                                                    test_size=0.20,
                                                    random_state=50,
                                                )
```

```
import xgboost

dtrain = xgboost.DMatrix(X_train, y_train, feature_names=columnas)
dtest = xgboost.DMatrix(X_test, y_test, feature_names=columnas)
```

Nota. Código Python en Jupyter Notebook.

Para el modelo se ha utilizado la librería de Python **xgboost**, el mismo que se ha calibrado con las siguientes opciones:

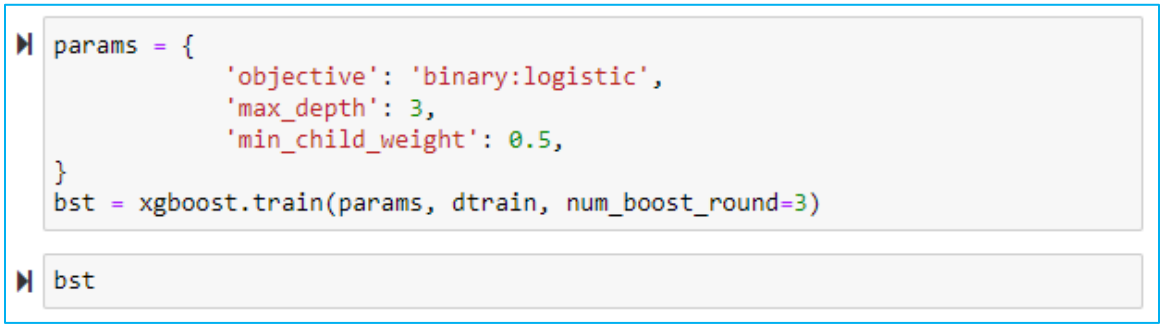
- **objetive: 'binary:logistic'** Indica que se va a utilizar regresión logística para clasificación binaria de probabilidad de salida. Esto debido a que nuestra variable objetivo de predicción "*falla*" es binaria: 1 = Sin disponibilidad, 0 = Disponibilidad normal.
- **max_depth: 3** Indica la profundidad máxima de cada árbol. Un valor de este parámetro puede hacer que el modelo se sobreajuste. Adicionalmente, por la cantidad de datos este valor no puede ser muy alto ya que a mayor profundidad de los árboles se tiene un mayor consumo de recursos de memoria y procesador.
- **min_child_weight: 0.5** Es la mínima suma del peso de la instancia necesaria en un árbol hijo. Evaluando este valor con el paso de la partición del árbol, permite o no continuar con otro árbol.

Las opciones seleccionadas y sus argumentos se basan en lo indicado en la documentación web de la librería XGBoost (xgboost developers, 2022). Es importante conocer que esta calibración se ha realizado en base a los recursos de hardware disponibles para el desarrollo del proyecto de MD (ver sección *Inventario de recursos* de la Fase 1 de CRISP-DM) y la experiencia del autor. Se hicieron diferentes ajustes a estos parámetros que permitieron la obtención de resultados positivos coherentes.

La Figura 45, presenta el código Python de programación y configuración del modelo de MD. Se envían los parámetros configurados (params), el Conjunto de Datos de entrenamiento (dtrain) y el número de bosques aleatorios que se desea generar internamente (num_boost_round=3). Posterior a la ejecución, los resultados se obtendrán en la variable llamada "***bst***" que es el modelo predictivo ya entrenado, y que posteriormente será evaluado.

Figura 45

Programación y configuración del Modelo XG Boost



```
params = {
            'objective': 'binary:logistic',
            'max_depth': 3,
            'min_child_weight': 0.5,
}
bst = xgboost.train(params, dtrain, num_boost_round=3)
```

```
bst
```

Nota. Código Python en Jupyter Notebook.

La Figura 46, presenta la configuración realizada con la librería Dask para la carga de datos (train y test) en memoria y el procesamiento del modelo predictivo XG Boost. Esta configuración, permite optimizar el paralelismo y escalabilidad de los recursos de hardware configurados.

Figura 46

Programación y configuración de Dask

```
from dask_ml.preprocessing import DummyEncoder as DE
from dask.distributed import Client, LocalCluster

cluster = LocalCluster(n_workers=5, threads_per_worker=4)
client = Client(cluster)
client
```

Out[1]:

Client

Client-caa6cb1a-5e50-11ed-977c-dd23da8d2fe1

Connection method: Cluster object **Cluster type:** distributed.LocalCluster

Dashboard: http://127.0.0.1:8787/status

Cluster Info

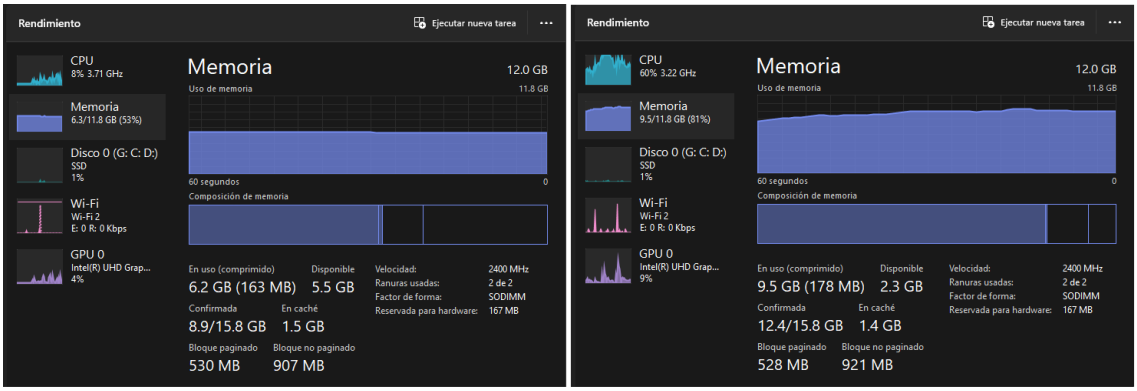
Nota. Código Python en Jupyter Notebook.

La Figura 47, presenta una comparación del uso de memoria y CPU en Windows antes y durante la ejecución de la carga de datos y el procesamiento del modelo. Se aprecia que, de

6.2 GB de memoria y 8 % de CPU (previo a la ejecución) aumenta a 9.5 GB de memoria y 60 % de CPU (durante la ejecución con Dask)

Figura 47

Comparación de uso de recursos con Dask



Nota. Gráficos obtenidos del Administrador de tareas de Windows.

La Figura 48, presenta el Dashboard del estado de CPU, memoria y tareas en tiempo real, durante la carga de datos y ejecución del proceso del modelo de MD, representando así el paralelismo y partición de en paquetes pequeños la información. En la parte superior izquierda, se muestra el uso de 3.78 GB de memoria "*Bytes stored*" particionados en 5 grupos de trabajo (número de workers configurados). El gráfico "*Task Processing*" presenta que se están procesando en paralelo 29 tareas y "*Task Stream*" presenta gráficamente la cantidad y tamaño de los paquetes de datos que se ejecutan en ese instante. Finalmente, en la parte inferior derecha gráfico "*Progress*" se tiene las diferentes barras que muestran el progreso de la ejecución.

Figura 48

Estado de recursos durante la ejecución del modelo con Dask

Nota. Gráficos obtenidos en la herramienta Dashboard de Dask.

Resultados del modelado

Con las Fases 1 al 4 del proceso de CRISP-DM (ver secciones *Fase 1: Entendimiento del negocio*, *Fase 2: Entendimiento de los datos*, *Fase 3: Preparación de los datos* y *Fase 4: Modelado* del presente capítulo) se resuelven las preguntas de investigación planteadas para el "*Objetivo específico 2 (OE2)*" del presente trabajo de titulación:

OE2-RQ1 ¿Qué modelo de minería de datos predice umbrales dinámicos de sensores de red, basados en tendencias de comportamiento de dispositivos?

Basado en la documentación y la realización de las Fases 1 al 4 del proceso CRISP-DM, se concluye que una alternativa para contestar esta pregunta es, utilizar un modelo predictivo basado en el algoritmo **XG Boost (Extreme Gradient Boosting)**, por las siguientes razones:

- Se puede entrenar con datos históricos del monitoreo de los sensores de red.
- Se tiene una variable objetivo de predicción binaria: 1 = indica indisponibilidad, 0 = indica disponibilidad normal.
- Es un modelo evolucionado de los Árboles de decisión tradicionales, que se caracteriza por ser un método gradiente descendente. Es decir, analiza los árboles secuencialmente, corrigiendo los errores encontrados en el árbol predecesor y enviando un árbol mejorado para el siguiente análisis.
- Los resultados y evaluación del modelo (siguiente sección), prevé resultados alentadores de predicción. Con lo cual, se puede determinar umbrales dinámicos que indiquen si la red tiene problemas que generen indisponibilidad.

OE2-RQ2 ¿Cuáles son las aplicaciones y fuentes de datos de la empresa proveedora de telecomunicaciones, que almacenan la información histórica del monitoreo de los sensores de red?

Las Fases 2 y 3 del proceso CRISP-DM nos permiten determinar las aplicaciones y fuentes de datos históricos que tiene la empresa de telecomunicaciones para el monitoreo de equipos de red:

- Aplicaciones: Network Management System (NMS) instalado y configurado en la organización. Software encargado de monitorear y recolectar la información de los sensores y dispositivos de red de clientes corporativos.
- Fuentes de datos: Base de datos SQL Server 2017, que almacena la información actual e histórica de la aplicación NMS.

Fase 5: Evaluación

Esta sección permite evaluar los resultados de Minería de Datos, obtenido en las fases anteriores de CRISP-DM.

Evaluación de los resultados

En términos de Minería de Datos se van a evaluar los resultados con las métricas de ROC – AUC planteadas en la Fase 4 - Modelado (ver sección *Generación de un diseño de comprobación*).

La Figura 49, presenta el código Python utilizado para obtener los resultados de la Curva ROC y el Área AUC. El mismo que, captura los resultados del modelo “bst” en la variable “y_pred” que es la colección (array) que contiene los resultados de predicción, esta se compara con la variable “y_test” que es la colección de los resultados reales de test. Posteriormente, se calculan las Tasas (Ratios) de Falsos Positivos (FPR) y Verdaderos Positivos (TPR), Finalmente, con estos datos se calculan los valores de ROC – AUC y se proceden a graficar.

Figura 49

Programación ROC – AUC

```
In [56]: y_pred = bst.predict(dtest)
         y_pred

Out[56]: array([0.18780446, 0.18780446, 0.18780446, ..., 0.18780446, 0.18780446,
                0.18780446], dtype=float32)

In [57]: gc.collect()

Out[57]: 0

In [58]: from sklearn.metrics import roc_curve
         fpr, tpr, _ = roc_curve(y_test, y_pred)

In [59]: gc.collect()

Out[59]: 200

In [60]: from sklearn.metrics import auc

         fig, ax = plt.subplots(figsize=(6, 6))
         ax.plot(fpr, tpr, lw=3,color='darkorange',
                 label='ROC Curve (area = {:.4f})'.format(auc(fpr, tpr)))
         ax.plot([0, 1], [0, 1], 'k--', lw=2)
         ax.set(
             xlim=(0, 1),
             ylim=(0, 1),
             title="ROC Curve",
             xlabel="False Positive Rate",
             ylabel="True Positive Rate",
         )
         ax.legend();
         plt.show()
```

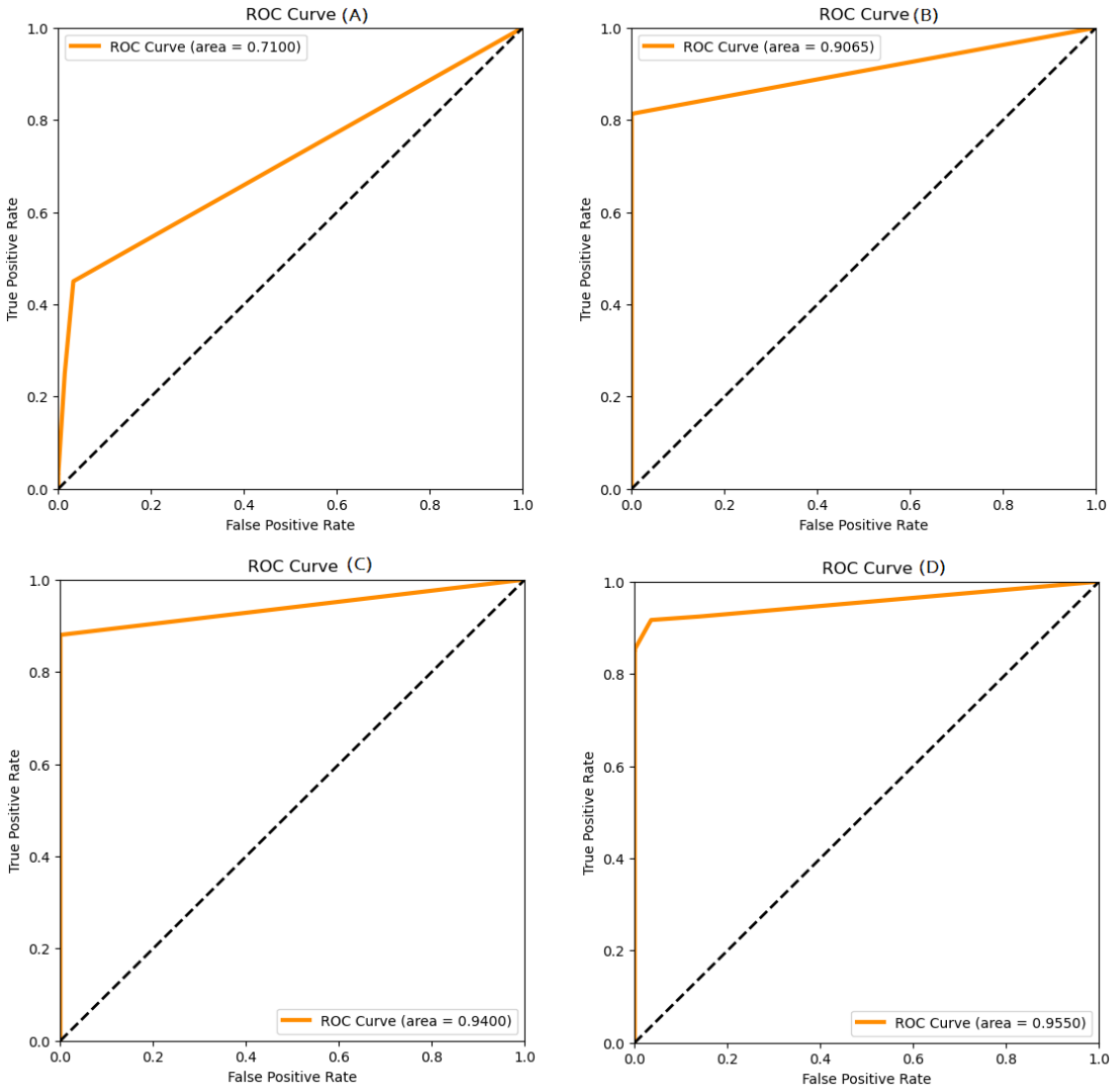
Nota. Código Python en Jupyter Notebook.

Posterior a la ejecución del modelo con diferentes grupos de dispositivos en la herramienta Jupyter Notebook de Python se encuentra la siguiente evidencia:

La Figura 50, presenta los resultados de Curva ROC y AUC de las diferentes ejecuciones realizadas, según lo explicado en la sección *Generación de un diseño de comprobación* de la Fase 4 de CRISP-DM, para tener una mayor probabilidad de acierto, los resultados deben equilibrarse entre ser más sensibles (mayor número de Verdaderos Positivos TP) y más específicos (mayor número de Verdaderos Negativos TN). Es decir, el modelo tiene mayor fiabilidad si la curva ROC se acerca al valor 1 y el Área AUC contiene la mayor cantidad de umbrales de clasificación. En consecuencia, mediante el proceso iterativo CRISP-DM fue factible mejorar los resultados del modelo desde un AUC de 0.71 (A) hasta un AUC de 0.955 (D), lo cual se interpreta en que el modelo final de MD tiene una efectividad del **95.5 %** de predicción.

Figura 50

Resultados ROC – AUC aplicado al Modelo de MD



Nota. Gráficos obtenidos en diferentes ejecuciones y calibraciones del modelo de MD.

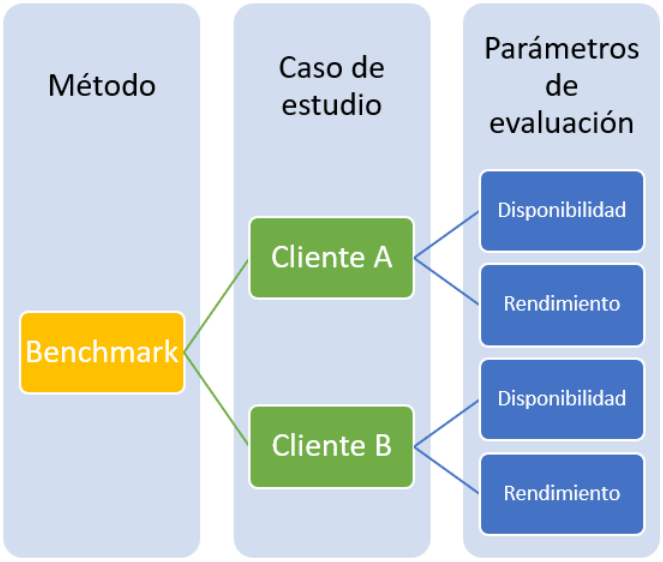
Proceso de revisión

Siguiendo los objetivos planteados de negocio y minería de datos (Fase 1 de CRISP-DM), se hace indispensable tener un proceso de revisión que permita evaluar y comprobar los resultados, para esto se va a realizar una comparación Benchmark a través de un caso de estudio. La Figura 51, presenta el proceso para la revisión de casos de estudio Benchmark,

este se basa en la comparación de dos clientes de la empresa proveedora al aplicar el modelo de MD en sus dispositivos.

Figura 51

Proceso de revisión caso de estudio Benchmark



Nota. Presenta los pasos para realizar el Benchmark.

De manera general el método Benchmark permite la comparación entre dos entidades similares, por lo tanto, se propone realizar una comparación entre un Cliente A y un Cliente B de la compañía de telecomunicaciones. La Tabla 15, presenta una comparación de las características de la empresa y los resultados de aplicación del modelo de predicción.

Tabla 15

Comparación de casos de estudio del Cliente A y B

Característica	Cliente A	Cliente B	Observación
Nombre de empresa	E_434320	E_233536	Nombre ficticio por confidencialidad de la información
Cantidad de dispositivos	3	3	Cantidad de dispositivos monitoreados
Fecha de inicio	15-Abr-2020	15-Abr-2020	Fecha inicio del conjunto de datos de monitoreo de sensores de red
Fecha de fin	15-May-2020	15-May-2020	Fecha fin del conjunto de datos de monitoreo de sensores de red
Cantidad de registros	430,053	435,479	Cantidad total de registros del conjunto de datos

Característica	Cliente A	Cliente B	Observación
Cantidad de registros nulos	8,942	9,822	Se obtiene al filtrar los registros nulos del conjunto de datos total
Porcentaje de registros nulos	2.07%	2.25%	Cantidad de registros nulos / Cantidad de registros
Cantidad de registros con falla	8,384	9,227	Se obtiene al filtrar los registros con estado = 1 y Downtime > 0.4
Porcentaje de registros con falla	1.94%	2.11%	Cantidad de registros con falla / Cantidad de registros
AUC-ROC	70.99%	94.00%	Porcentaje de predicción del modelo obtenido.
Disponibilidad promedio sin aplicación del modelo de MD	98.07%	97.87%	Disponibilidad obtenida del Conjunto de datos
Disponibilidad promedio con simulación de aplicación del modelo de MD	98.88%	99.74%	Se aplica el % de predicción del modelo sobre los datos con falla dividido para 2 (considerando que la mitad de las alertas se pueden solucionar)
Porcentaje de mejora posterior a la aplicación del modelo	0.81%	1.87%	Porcentaje de mejora de la disponibilidad
Tiempo de entrenamiento del modelo de MD	8:55	8:46	Valores en minutos y segundos

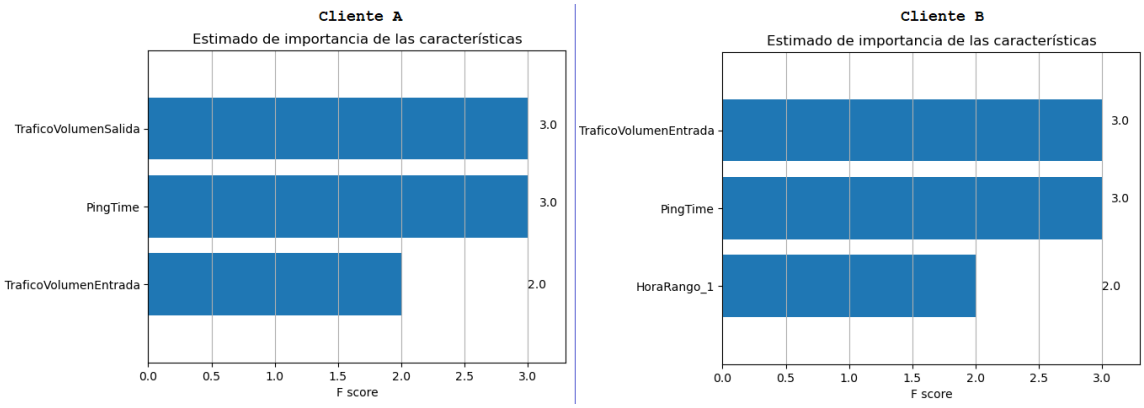
Nota. Valores obtenidos del análisis de los datos y aplicación del modelo de MD.

En la evaluación del modelo de MD se ha realizado un estimado de la importancia de las características, es decir, se obtiene los campos que influyen con mayor peso dentro del modelo de MD. La Figura 52, presenta una comparación de los resultados de estas características, las cuales se detallan a continuación:

- Cliente A
    - “TraficoVolumenSalida” con score de 3
    - “PingTime” con score de 3
    - “TraficoVolumenEntrada” con score de 2
- Cliente B
    - “TraficoVolumenEntrada” con score de 3
    - “PingTime” con score de 3
    - “HoraRango_1” con score de 2

Figura 52

Comparación de características de importancia en el modelo de MD



Nota. Gráficos obtenidos en Python en Jupyter Notebook.

Resultados de la evaluación

Con la Fase 5 del proceso de MD, se resuelven las preguntas de investigación planteadas para el "*Objetivo específico 3 (OE3)*" del presente trabajo de titulación:

OE3-RQ1 ¿Cuáles son los indicadores que permiten medir la mejora de disponibilidad del servicio de internet y datos?

Según el análisis Benchmark realizado en la sección *Proceso de revisión* de la Fase 5 de CRISP-DM, se puede concluir que los indicadores que permiten medir la mejora de disponibilidad del servicio de internet y datos son:

- Cantidad de registros
- Cantidad de registros nulos
- Porcentaje de registros nulos
- Cantidad de registros con falla
- Porcentaje de registros con falla
- AUC-ROC
- Disponibilidad promedio sin aplicación del modelo de MD

- Disponibilidad promedio con simulación de aplicación del modelo de MD
- Porcentaje de mejora posterior a la aplicación del modelo

Adicionalmente, las características de rendimiento que más influyen (score alto) en el modelo de MD para la predicción son:

- Tráfico de Volumen de Salida
- Tráfico de Volumen de Entrada
- Ping Time

OE3-RQ2 ¿Cuál es el nivel de confianza de los resultados obtenidos?

En el apartado *Evaluación de los resultados*, se detalla que el modelo alcanza valores de predicción de hasta el **95.5%**, esto a través de la evaluación de las métricas AUC y ROC. En términos generales es un porcentaje alentador ya que con estos valores y una correcta gestión en la toma de decisiones se puede a incrementar entre 0.5 y 2.0 puntos los porcentajes actuales de disponibilidad (ver Tabla 15
Comparación de casos de estudio del Cliente A y B). Con esto se concluye que el nivel de confianza es relativamente alto.

Fase 6: Despliegue

El presente trabajo de titulación no contempla el despliegue del modelo de MD en la empresa proveedora de telecomunicaciones. Sin embargo, se expone los pasos generales que se deben seguir para su implementación:

1. Exportar el modelo XG Boost de Python ya entrenado y testeado.
2. Obtener información reciente de monitoreo de red, por la gran cantidad de información se sugiere realizar esta tarea por dispositivo o grupo de dispositivos o periodos cortos de entre 3 y 10 minutos.

3. Preparar los datos de monitoreo obtenidos, limpiar y transformar la información con la finalidad de tener un Conjunto de datos con el formato requerido para el modelo (ver Campos de la Entidad [MON_MONITOREO_TABLE] del Apéndice C del presente documento).
4. Ejecutar en Python el modelo XG Boost con los datos de entrada del paso 3.
5. Evaluar los resultados y determinar si estos datos generan una posible indisponibilidad del servicio.

Contribuciones

Investigación adscrita a la Red Sistemas Inteligentes y Expertos Modelos Computacionales Iberoamericanos (SIEMCI), número de proyecto 522RT0130 en Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED).

Conclusiones

Se determinaron las técnicas predictivas supervisadas de Minería de Datos idóneas para pronosticar umbrales de sensores de red. Los resultados de la revisión preliminar de literatura arrojaron que las técnicas más utilizadas son: Árboles de decisión, Redes neuronales, Clasificación lineal y Modelos de regresión. Adicionalmente, se seleccionó el algoritmo XG Boost (técnica evolucionada de los Árboles de decisión) para la construcción y desarrollo del modelo predictivo de MD.

El desarrollo del modelo de predicción se realizó siguiendo los pasos de la metodología CRISP-DM. Se utilizó la plataforma Anaconda Distribution (Python) y el código de programación se escribió en Jupyter Notebook. Se utilizaron dos librerías principales: Dask (para el procesamiento paralelo) y XGBoost (para el modelo, su entrenamiento y evaluación). El modelo fue entrenado y testeado con datos históricos de monitoreo de red de los clientes corporativos de la empresa de telecomunicaciones. La fuente de información se almacena en la base de datos SQL Server 2017 de la aplicación NMS que administra la compañía.

Es factible crear un modelo predictivo de MD de sensores de red basado en la tendencia de monitoreos históricos de los dispositivos. Sin embargo, es necesario utilizar herramientas y técnicas de Minería de Datos y Machine Learning que puedan soportar gran cantidad de datos.

Se encontró que la cantidad de datos históricos de monitoreo, más los recursos de hardware y software disponibles, influyen en el tiempo de procesamiento del modelo de MD en Python. Es decir, a menor número de registros y mayor cantidad de recursos de hardware y software, menor será el tiempo de procesamiento y viceversa.

Se realizó la evaluación del modelo de MD, utilizando las métricas ROC y AUC, las cuales se utilizan ampliamente en ciencia de datos para comprobar modelos predictivos generando fiabilidad en los resultados. El porcentaje de predicción obtenido con valores de

hasta un 95.5% se puede considerar que es fiable y puede ser utilizado en la industria de telecomunicaciones.

El análisis Benchmark realizado entre dos clientes corporativos de la empresa proveedora, permitió revelar aspectos cuantitativos de disponibilidad. Tales como: El modelo de MD tiene un mayor porcentaje de predicción, cuando en los datos históricos existe un mayor porcentaje de fallas y viceversa; Al aplicar el porcentaje de predicción de fallas y considerando que la mitad de estas se puedan solventar proactivamente, se tiene una mejora de disponibilidad de entre 0.81 y 1.87 %.

Recomendaciones

En base a la experiencia alcanzada con la ejecución del presente proyecto se realizan las siguientes recomendaciones, las cuales pueden converger en futuros trabajos de investigación:

A la comunidad tecnológica se propone el uso del algoritmo XG Boost (Gradiente Descendiente), como una alternativa de técnica no supervisada de predicción de Minería de Datos y Machine Learning, que se puede aplicar en nuevos proyectos de diferentes industrias.

A la empresa proveedora de telecomunicaciones se sugiere la implementación (despliegue) del modelo de MD propuesto, debido a que los resultados obtenidos son alentadores, con porcentajes confiables de predicción de fallas, que puedan alertar proactivamente al personal de soporte para su posterior toma de decisiones.

Al considerar que, los datos de monitoreo y dispositivos fueron exclusivos de los clientes corporativos de una empresa de telecomunicaciones del Ecuador, no se puede concluir que el modelo predictivo de MD se puede aplicar a todas las empresas de telecomunicaciones. Por lo tanto, se abre una puerta para nuevos estudios que aborden esta problemática.

En el análisis Benchmark del presente trabajo, se encontraron aspectos cualitativos que mejoran el rendimiento del servicio de internet y datos. Sin embargo, se plantea a la empresa de telecomunicaciones un estudio cuantitativo, que aborde las características principales de los sensores de red que influyen en el rendimiento, tales como: Tráfico de Volumen de Salida, Tráfico de Volumen de Entrada y Ping Time.

Bibliografía

AB Internet Networks. (2017). *Desde Linux - Anaconda Distribution*. Obtenido de https://blog.desdelinux.net/ciencia-de-datos-con-python/

Acens Technologies S.A. (2021). *Acens White Papers.* Obtenido de www.acens.com

Agrawal, R., & Srikant, R. (1995). *Mining Sequential Patterns.* Taipei, República de China: The 11th International Conference on Data Engineering.

Amat, J. (Febrero de 2017). *Árboles de decisión, random, forest, gradient boosting y C5.0*. Obtenido de https://www.cienciadedatos.net/

Anaconda Inc. (2022). *Anaconda Distribution*. Obtenido de https://www.anaconda.com/products/distribution

Anaconda, Inc. and contributors. (2018). *Documentation Dask*. Obtenido de https://docs.dask.org/en/stable/

ARCOTEL - DEAC. (2019). *Requerimientos Ciudadanos - Atención al usuario.* Quito.

ARCOTEL. (2019). *Misión, visión, principios y valores*. Obtenido de Agencia de Regulación y Control de las Telecomunicaciones: https://www.arcotel.gob.ec/mision-vision-principios-y-valores2

ARCOTEL. (2021). *Sitio oficial de la Agencia de Regulación y Control de las Telecomunicaciones*.

Atlassian Software. (2021). *Incident Management*. Obtenido de https://www.atlassian.com/incident-management/kpis/common-metrics

Barbosa, E., & Miño, C. (2017). Aporte decisivo al mundo de los negocios. *Computerword 302 Telecomunicaciones*.

Barros, R., Basgalupp, M., Carvalho, A., & Freitas, A. (2012). A Survey of Evolutionary Algorithms for Decision-Tree Induction. *Systems, Man, and Cybemetics, Part C: Applications and Reviews, 42*, pp. 291-312.

Berry, M., & Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales, and Customer Support.* NY: John Wiley.

Bhandari, A. (2020). *AUC-ROC Curve in Machine Learning Clearly Explained*. Obtenido de https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Britos, P. (2008). *Procesos de explotación de información basados en sistemas inteligentes.* Buenos Aires: Doctoral dissertation, Facultad de Informática de la Universidad Nacional de la Plata.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1997). *Discovering Data Mining From concept to implementation.* Prentice Hall.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.*

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

Cheng, Y. (18 de Noviembre de 2018). *Pager Tree - What is MTTR? Critical Incident Recovery Metrics to Reduce Downtime*. Obtenido de https://pagertree.com/2018/11/20/what-is-mttr/

Code Algorithms Pvt. Ltd. (2022). *Enjoy Algorithms - XG-Boost (Extreme Gradient Boosting) Algorithm in Machine Learning*. Obtenido de https://www.enjoyalgorithms.com/blog/xg-boost-algorithm-in-ml

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners.* CreateSpace Independent Publishing Platform.

Digital Guide IONOS. (2018). *Digital Guide*. Obtenido de https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas

Espino Timón, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.

Espinoza Zúñiga, J. (2020). Application of Random Forest and XGBoost algorithms based on a credit card applications database.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD.* (Vol. 96).

García Garrido, S. (2018). *Renovetec*. Obtenido de http://www.renovetec.com/590-mantenimiento-industrial/110-mantenimiento-industrial/300-indicadores-en-mantenimiento

García González, F. J. (2013). Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Meedio Ambiente (CEAMA).

García Gutiérrez, J. A. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas.

García Herrero, J., & Molina López, J. M. (2012). *Técnicas de Análisis de Datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka.* Madrid: Universidad Carlos III.

García, R. (2004). *Sistemas Autónomos. Aprendizaje Automático.* Editorial Nueva Librería.

Gartner Inc. (2022). *Magic Quadrant Gartner*. Obtenido de https://www.gartner.es/es/metodologias/magic-quadrant

González, R., & Pomares, A. (2012). La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería. *Reunión Nacional ACOFI.* Medellín.

Google Developers. (2022). *Machine Learning Clasificación: Curva ROC y AUC*. Obtenido de https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

GrupoCONTEXT S.A. (2022). *Portarl GrupoCONTEXT - NMS+ (NETWORK MANAGEMENT SYSTEM)*. Obtenido de https://www.grupocontext.com/bsm/nms-network-management-system/

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques.* San Francisco: Morgan Kaufmann Publishers.

Hevner, A. (2007). A Three Cycle View of Design Science Research. *Sandinavian Journal of Information Systems*.

Hevner, A., & Chatterjee, S. (2010). Design Research in Information Systems Theory and Practice.

IBM. (2019). *Acuerdos de Nivel de Servicio (SLA)*. Obtenido de IBM® IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/es/SS2T64/com.ibm.spr.doc/sla_spr/c_sla_application.html

IBM. (2021). *Guía de CRISP-DM de IBM SPSS Modeler*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

IBM Cloud Education. (Julio de 2020). *IBM Cloud Learn Hub*. Obtenido de https://www.ibm.com/cloud/learn/machine-learning?lnk=fle

ITIL Foundation. (2011). *Gestión de la Disponibilidad*. Obtenido de ITIL Foundation Gestión de Servicios TI: http://segenuino.com/itil/diseno_servicios_TI/gestion_disponibilidad/introduccion_objetivos.html

Lartey, F. (2017). Proactive Network and Technical Facilities Monitoring Using Standardized Scorecards. *Fall Technical Forum.* Denver, CO.

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook.

Medina , F., & Gómez, C. (2014). Funcionalidades de la minería de datos. *Revista Ingeniería y Región, 12*, 31-40.

MINTEL. (2019). *Valores / Misión / Visión*. Obtenido de Ministerio de Telecomunicaciones y de la Sociedad de la Información: https://www.telecomunicaciones.gob.ec/valores-mision-vision/

MINTEL. (2020). *Sitio oficial Ministerio de Telecomunicaciones y de la Sociedad de la Información*. Obtenido de https://www.telecomunicaciones.gob.ec/valores-mision-vision/

Páez Juka, S. (2019). *Análisis comparativo de herramientas Open Source para Data Mining sobre datos públicos del Ministerio de Educación de la República del Ecuador.* Quito.

PÁGINA 7 COMUNICACIÓN S.L. MADRID. (2017). *Nueva Tribuna Es.* Obtenido de https://www.nuevatribuna.es/articulo/consumo/no-mide-no-puede-mejorar/20170621154349141055.html

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*.

Peredo Cortes, I. (Octubre de 2020). *RPubs*. Obtenido de https://rpubs.com/IranNash/discretizacion

Pérez, M. (2014). *Minería de datos a través de ejemplos.* Madrid: RC Libros.

Piatesky-Shapiro, G., & Frawley, W. (1991). *Knowledge Discovery in Databases.* Cambridge, MA: AAA/MIT Press.

Piirainen, K., Gonzalez, R., & Kolsfschoten, G. (2010). Quo Vadis, Design Science? – A Survey of Literature. *Global Perspectives on Design Science Research, Lecture Notes in Computer Science*.

RinconTIC. (2020). *RinconTIC MTBF*. Obtenido de https://rincontic.org/2020/04/12/que-es-el-mtbf-y-como-calcularlo/

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*, 601 - 618.

Salvo, R., Montato, P., Nunnaria, G., Neri, M., & Puglisi, G. (2013). Multivariate time series clustering on geophysical data recorded at Mt. Etna from 1996 to 2003. *Journal of Volcanology and Geothermal Research, 251*, 65-74.

Srikant, R., & Agrawal, R. (1996). *Mining quantitative association rules in large relational tables.* Obtenido de http://rakesh.agrawal-family.com/papers/sigmod- 96qassoc.pdf

Suárez, D. (2019). Claves corportativas contra el peligro digital. *DATTA*.

Swiderski, B., Kurek, J., & Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. *Decisión Support Systems, 52*, 538-547.

Tebes, G., Peppino, D., Rivera, B., Becker, P., Papa, F., & Olsina, L. (2019). Especificación del Proceso de Design Science Research: Caso Aplicado a una Ontología de Testing de Software. *7mo Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI).* La Pampa.

Thuraisingham, B. (1999). *Data Mining: Technologies, Techniques, Tools and Trends.* CRC Press.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.* Bogotá: Ediciones Universidad Cooperativa de Colombia. doi:http://dx.doi.org/10.16925/9789587600490

Timarán, R. (2009). Una mirada al descubrimiento de conocimiento en bases de datos. *Revista Ventana Informática*, 39-58.

Universidad Nacional de San Luis. (2018). Aprendizaje Automático y Minería de Datos. Argentina.

Valcárcel, V. (2004). Data Mining y el Descubrimiento del Conocimiento. *Revista de la Facultad de Ingeniería Industrial UNMSM, 7*(2), 83-86.

Vallalta Rueda, J. (2022). *CRISP-DM: una metodología para minería de datos en salud*. Obtenido de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Villena Román, J., Crespo García, R., & García Rueda, J. (2012). Inteligencia en Redes de Comunicaciones. Madrid.

Wang, M., Iyer, B., & Scott, J. (1998). *Scalable Mining for Classification Rules in Relational Databases.* Cardiff, Wales: International Database Engineering and Application Symposium - Ideas.

Weiss, S., & Indurkhya, N. (2008). *Predictive data mining. A practical guide.* San Francisco: Morgan Kaufmann Publishers.

Westreich, D., Lessler, J., & Funk, M. (2010). Propensity score estimation: neural networks, support vector machines, decisión trees (CART), and meta-classifiers as altematives to logistic regression. *Journal o f Clinical Epidemiology, 63*(8), 826-833.

Witten, H., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco, CA: Morgan Kaufmann.

Wujek, B. (2014). *Machine Learning.* Obtenido de SAS Institute Inc.: https://www.sas.com/en_us/insights/analytics/machine-learning.html

xgboost developers. (2022). *XGBoost Documentation*. Obtenido de https://xgboost.readthedocs.io/en/stable/index.html

Zhang, A. (2019). *Data Driven Investor*. Obtenido de https://medium.datadriveninvestor.com/pandas-dask-or-pyspark-what-should-you-choose-for-your-dataset-c0f67e1b1d36

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). *An Efficient Data Clustering Method for Very Large Databases.* Montreal, Canadá: International Conference on Management of Data.

Apéndices