MODELADO PREDICTIVO APLICADO AL COMPORTAMIENTO DE COMPONENTES DE LA INFRAESTRUCTURA TECNOLOGICA EN EMPRESAS DE BIENES Y SERVICIOS

Autor (es)

Liliana Maria Lopez Restrepo
Daniel Vanegas Acevedo

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor:
Ingrid Durley Torres Pardo
Victor Daniel Gil Vera

Especialización en Big Data e Inteligencia de Negocios

Especialización en Big Data e Inteligencia de Negocios
Facultad de Ingenierías y Arquitectura
Universidad Católica Luis Amigó
Medellín, Colombia
2024

Dedicatoria

“Este trabajo se lo dedico a mis padres, pues han sido mi apoyo incondicional y la fuente de mi inspiración para lograr cada uno de los objetivos que me planteo y que su amor siempre es la energía que necesito para motivarme a sacar adelante mis proyectos.”.

Daniel Vanegas Acevedo

“A mis amadas hijas y mi amado compañero,

A mis preciosas hijas, ustedes son la luz de mi vida y la razón por la que me esfuerzo cada día. Su alegría y amor me motivan a ser mejor y a nunca rendirme. Este logro también es para ustedes, para que vean que, con dedicación y esfuerzo, todo es posible.

A ti, cielo, gracias por ser mi roca y mi constante apoyo. Tu paciencia, comprensión y amor incondicional me han dado la fuerza para seguir adelante y superar cada desafío en este camino. Eres mi compañero en todas las aventuras y mi inspiración diaria.

Gracias por su amor y apoyo incondicional. Este logro no habría sido posible sin ustedes a mi lado.”

Liliana Maria Lopez Restrepo

Agradecimientos

Se agradece inmensamente a nuestras familias por el tiempo, la comprensión y el apoyo durante todo el proceso mientras se cursaba la Especialización.

Adicionalmente agradecemos también a la Escuela de Posgrados y a todos los docentes que participaron en este proceso de formación.

Tabla de Contenido

Tabla de Figuras

1. Introducción

Comentado [1]: Revisar numeración. shay un espacio amplio respecto a la numeración y el texto del titulo

La infraestructura tecnológica es el pilar fundamental que sustenta las operaciones de las empresas de bienes y servicios. Sin embargo, muchos desafíos afectan la eficiencia operativa de esta infraestructura, lo que puede generar consecuencias negativas para el desempeño y rentabilidad de las organizaciones. Este estudio aborda la necesidad de mejorar la eficiencia de la infraestructura tecnológica, poniendo el foco en la capacidad de predecir y abordar proactivamente problemas antes de que se conviertan en interrupciones costosas.

Las interrupciones no planificadas en la infraestructura tecnológica pueden acarrear importantes pérdidas financieras y una reducción significativa en la productividad. Además de los costos directos derivados del tiempo de inactividad, existen costos indirectos relacionados con la pérdida de confianza del cliente y una menor competitividad en el mercado. Por esta razón, es esencial desarrollar estrategias que garanticen la confiabilidad y disponibilidad de la infraestructura tecnológica, minimizando así las fallas y optimizando las operaciones.

En este contexto, el monitoreo y la gestión de la infraestructura de TI requieren un enfoque proactivo, que integre el uso de registros, modelos predictivos y análisis de grandes volúmenes de datos. El objetivo es identificar problemas potenciales de manera temprana y tomar medidas antes de que causen daños significativos. La confiabilidad y disponibilidad de la infraestructura tecnológica no son solo cuestiones técnicas, sino aspectos estratégicos que impactan la operación y el éxito general de las empresas de bienes y servicios.

La revisión sistemática de literatura (RSL) presentada en este estudio busca identificar los modelos más efectivos y las mejores prácticas para mejorar la estabilidad, eficiencia y rentabilidad de las empresas al asegurar que su infraestructura tecnológica funcione de manera óptima y predecible. Al revisar la literatura científica, se pretende encontrar soluciones innovadoras para la gestión eficiente de recursos tecnológicos, como servidores, redes, almacenamiento y aplicaciones, con el fin de minimizar el tiempo de inactividad y garantizar la disponibilidad continua de servicios críticos. El análisis de la literatura también tiene como objetivo entender las estrategias que permiten detectar de forma temprana problemas de hardware, cuestiones de rendimiento, amenazas de seguridad

y otras interrupciones no planificadas, lo que contribuye a mantener la productividad y satisfacción del cliente.

El presente estudio busca aportar un enfoque sistemático para la revisión de la literatura, con el fin de presentar recomendaciones claras y fundamentadas que puedan ser aplicadas por las empresas de bienes y servicios para mejorar la confiabilidad y disponibilidad de su infraestructura tecnológica.

2. Motivación

El problema que aborda este trabajo radica en la necesidad de mejorar la eficiencia operativa de las empresas de bienes y servicios en lo que respecta a su infraestructura tecnológica, incluyendo sus componentes y servicios asociados. La mejora de esta eficiencia es esencial, ya que permite predecir y anticipar el comportamiento de la infraestructura tecnológica, lo que a su vez proporciona la oportunidad de optimizar las operaciones empresariales, maximizar el tiempo de actividad y reducir los costos operativos.

La importancia de este enfoque radica en el impacto significativo que las interrupciones no planificadas en la infraestructura tecnológica pueden tener en las empresas de bienes y servicios. Dichas interrupciones pueden generar costos sustanciales, tanto económicos como relacionados con el rendimiento empresarial, afectando la productividad y la experiencia del cliente. Los períodos de inactividad, incluso los breves, pueden resultar en pérdidas financieras considerables y dañar la reputación de la empresa.

En este contexto, se vuelve crucial garantizar que todos los componentes de la infraestructura tecnológica en las empresas de bienes y servicios funcionen de manera confiable y eficiente. El monitoreo de la infraestructura de TI ahora exige la inclusión de registros detallados, modelos predictivos y un análisis de grandes volúmenes de datos para detectar fallos de manera temprana y evitar interrupciones costosas. La confiabilidad y la disponibilidad de la infraestructura tecnológica se han convertido en temas de creciente interés para investigadores y profesionales del sector, destacando la necesidad de mantener registros precisos y utilizar herramientas analíticas avanzadas para prevenir y resolver problemas antes de que impacten negativamente a las operaciones empresariales.

Por ello, el propósito de esta investigación es comprender y analizar diversas herramientas, métodos y técnicas propuestas en la literatura que permitan reducir el tiempo de inactividad y mejorar la confiabilidad y eficiencia de la infraestructura tecnológica en las empresas de bienes y servicios. Al abordar estos aspectos, se espera proporcionar recomendaciones prácticas para la gestión efectiva de la infraestructura de TI y la implementación de estrategias proactivas que garanticen la continuidad operativa y reduzcan los costos asociados con las interrupciones no planificadas.

3. Planteamiento del problema

La infraestructura tecnológica es un componente crítico para la operación eficiente de las empresas del sector financiero, especialmente aquellas dedicadas al otorgamiento de créditos de consumo. En el caso de la empresa Empresas de bienes y servicios, el funcionamiento óptimo de sus servidores y unidades de disco es fundamental para mantener la continuidad de sus servicios y la confianza de sus clientes. Sin embargo, la ocurrencia de fallas inesperadas en estas unidades puede desencadenar interrupciones en el servicio, con consecuencias financieras significativas y posibles daños a la reputación de la empresa.

Comentado [2]: Justificar

La problemática central radica en la falta de un sistema robusto para predecir y anticipar fallas en las unidades de disco de los servidores de la empresa. Estas fallas pueden llevar a la pérdida de datos, disminución del rendimiento de los sistemas y períodos de inactividad, que afectan negativamente la experiencia del cliente y la eficiencia operativa. Dado que la información sensible de los clientes es almacenada en estas unidades, la seguridad y confidencialidad de los datos se ven comprometidas cuando ocurren fallas inesperadas.

El impacto de estas fallas se agrava debido a la naturaleza competitiva del sector financiero, donde los clientes esperan un acceso ininterrumpido a los servicios y una gestión segura de sus datos personales y financieros. La necesidad de un sistema predictivo de fallas se vuelve urgente para reducir el riesgo de interrupciones y mejorar la capacidad de respuesta ante posibles problemas, permitiendo a la empresa Empresas de bienes y servicios mantener altos niveles de disponibilidad y rendimiento.

En el competitivo mundo del sector financiero, donde cada minuto de inactividad puede significar pérdidas significativas y erosión de la confianza del cliente, la demanda de un acceso ininterrumpido a los servicios y una gestión segura de la información es más alta que nunca. Los consumidores confían en que sus datos personales y financieros estén protegidos, y cualquier fallo en el sistema puede resultar en consecuencias devastadoras para la reputación de una empresa. Por esta razón, la implementación de un sistema predictivo de fallas no es solo una medida proactiva, sino también un componente crítico para minimizar el riesgo de interrupciones. Al predecir y abordar posibles problemas antes de que ocurran, Empresas de bienes y servicios puede mantener altos niveles de disponibilidad y rendimiento, ofreciendo a sus clientes la seguridad y la confianza que exigen. Esto, a su vez,

le permite a la empresa sostener su ventaja competitiva y responder con rapidez y eficacia a los desafíos del mercado.

Este problema no solo afecta el ámbito técnico, sino que también tiene implicaciones económicas y de reputación. Los costos asociados con las interrupciones no planificadas pueden ser sustanciales, incluyendo la pérdida de ingresos, el costo de reparaciones y el impacto en la satisfacción del cliente. Además, la falta de un sistema de predicción eficaz limita la capacidad de la empresa para tomar decisiones estratégicas informadas y proactivas respecto a su infraestructura tecnológica.

Por lo tanto, el desarrollo de un sistema de predicción de fallas de unidades de disco en servidores se presenta como una solución efectiva para abordar este problema. Con el uso de técnicas avanzadas de inteligencia artificial y aprendizaje automático, se puede construir un modelo predictivo que permita a la empresa anticipar y prevenir fallas antes de que ocurran, optimizando así la gestión de su infraestructura tecnológica.

En conclusión, el problema a resolver es la necesidad de mejorar la capacidad de Empresas de bienes y servicios para predecir y prevenir fallas en las unidades de disco de sus servidores, asegurando la continuidad operativa, reduciendo costos operativos y manteniendo la confianza y satisfacción de sus clientes. El desarrollo e implementación de un sistema de predicción eficaz será fundamental para mitigar los riesgos asociados con las fallas inesperadas y para garantizar la excelencia en la prestación de servicios financieros.

4. Justificación

En el sector de otorgamiento de créditos de consumo, la disponibilidad continua y segura de la información del cliente es un pilar fundamental para asegurar una experiencia satisfactoria y confiable. La infraestructura tecnológica desempeña un papel crítico en este contexto, siendo el motor que impulsa la eficiencia operativa y la entrega efectiva de servicios financieros. Dada su importancia, las fallas en esta infraestructura pueden tener repercusiones graves, tanto financieras como de reputación. Existen casos históricos que resaltan la trascendencia de estos problemas, como el impacto del efecto Y2K o el colapso de sistemas tecnológicos tras el ataque a las Torres Gemelas (Barros, 2010). Estas situaciones nos recuerdan que la gestión proactiva de los riesgos asociados con los sistemas de almacenamiento de datos es esencial para prevenir consecuencias adversas.

La implementación de un sistema de predicción de fallas en unidades de disco en servidores de la empresa Empresas de bienes y servicios responde a necesidades y desafíos clave. En primer lugar, busca abordar la creciente preocupación por la seguridad de los datos sensibles que manejan las entidades financieras. El mal uso, modificación o acceso no autorizado a estos datos puede tener serias implicaciones para la privacidad, la seguridad y la reputación de la empresa, como destaca una publicación académica. Al desarrollar un sistema predictivo, se pretende prevenir interrupciones y garantizar la integridad y confidencialidad de la información financiera y personal de los clientes.

Este proyecto tiene el potencial de abordar estas preocupaciones y, al mismo tiempo, ofrecer otros beneficios significativos. La predicción de fallas no solo es una medida preventiva esencial, sino que también puede conducir a la optimización de procesos internos, reducción de costos de mantenimiento correctivo y mejora de la eficiencia general en la gestión de riesgos. Además, al prevenir interrupciones inesperadas, la empresa puede aumentar la confianza del cliente en la seguridad de sus datos y fortalecer su reputación en el mercado.

La implementación de este sistema también permite a la empresa centrarse en la innovación y el desarrollo de nuevos productos y servicios financieros, con la confianza de que la infraestructura tecnológica está protegida. Este enfoque preventivo fomenta un entorno empresarial estable y robusto, facilitando un crecimiento sostenible y una mayor competitividad en el sector de créditos de consumo. Como afirma un artículo de investigación, la seguridad de la información es un proceso integrado que requiere estrategias y medidas

tanto preventivas como reactivas para proteger la confidencialidad, disponibilidad e integridad de los datos (Altamirano, 2019).

Con todos estos beneficios en mente, la implementación del sistema de predicción de fallas en unidades de disco en servidores se presenta como una inversión estratégica indispensable para garantizar la excelencia en la prestación de servicios financieros y mantener la confianza del cliente.

5. Objetivos

5.1 Objetivo General

Desarrollar un sistema de predicción de fallas de unidades de disco en servidores de la empresa Empresas de bienes y servicios, con el fin de optimizar la gestión de la infraestructura tecnológica y mejorar la disponibilidad y rendimiento de los servicios de la empresa.

5.2 Objetivos Específicos

Recopilar y analizar datos históricos de fallos de unidades de disco en servidores de la empresa Empresas de bienes y servicios, así como información relevante sobre el entorno de operación y condiciones de uso de las unidades, para construir un conjunto de datos robusto y representativo.

Utilizar herramientas y bibliotecas de aprendizaje automático para implementar un modelo de inteligencia artificial que permita predecir posibles fallos en las unidades de disco de los servidores de la empresa Empresas de bienes y servicios. Emplear técnicas avanzadas de análisis predictivo, como series temporales, y métodos estadísticos para entrenar y validar el modelo. Asegurarse de que el modelo tenga un alto nivel de precisión y confiabilidad antes de su integración en la infraestructura tecnológica de la empresa.

Integrar el modelo de predicción de fallos con la infraestructura tecnológica existente en una empresa de bienes y servicios. Esto incluye la creación de una interfaz de usuario intuitiva y herramientas de monitoreo en tiempo real, permitiendo a los operadores identificar problemas potenciales rápidamente y tomar medidas proactivas. Asegurarse de que la integración se realice de forma segura y sin interrupciones en el servicio.

Analizar exhaustivamente los resultados obtenidos para identificar posibles áreas de mejora y tomar decisiones estratégicas que permitan corregir eficazmente las fallas identificadas, esto para intervenir de manera oportuna y optimizar el rendimiento general del sistema.

6. Marco metodológico

Para abordar el objetivo general de desarrollar un sistema de predicción de fallas de unidades de disco en servidores de la empresa Empresas de bienes y servicios, se propone el siguiente marco metodológico que integra métodos de recopilación de datos, análisis, desarrollo de modelos de aprendizaje automático, implementación e integración en la infraestructura tecnológica existente, es por esto que a continuación se muestra la matriz metodológica que describe los objetivos, actividades y entregables que se proponen en el proceso.

Comentado [3]: revisar tipo de letras y espaciado

Tabla 1.

Matriz metodológica

Objetivo	***Actividades***	***Entregable***
Recopilar y analizar datos históricos de fallos de unidades de disco en servidores de la empresa Sistecredito para construir un conjunto de datos representativo. *FASE1-2-3*	***Identificación de fuentes de datos:*** *Revisar registros internos, informes de mantenimiento y otros documentos relacionados con el historial de fallos de unidades de disco. Esto implica trabajar con el equipo de TI para obtener acceso a datos relevantes.* ***Definición de parámetros de recopilación:*** *Establecer qué información es necesaria para el análisis, como fechas, tipos de fallos, duración de interrupciones y condiciones de uso. Se asignará a un analista de datos.* ***Recopilación de datos:*** *Extraer datos de diversas fuentes y consolidarlos en una base de datos unificada. Esta tarea la realizará el equipo de TI con la ayuda de especialistas en datos.* ***Análisis y limpieza de datos:*** *Identificar y eliminar datos duplicados o incorrectos, y estructurar la información para facilitar el análisis posterior. Esta actividad puede requerir el uso de herramientas de procesamiento de datos.*	*Conjunto de datos históricos de fallos, limpio y estructurado.* *Informe detallado sobre las condiciones de uso de las unidades de disco y otros factores relacionados con los fallos.*
Implementar un modelo de inteligencia artificial para predecir posibles fallos en unidades de disco de los servidores de Sistecredito, utilizando técnicas avanzadas de análisis predictivo y métodos estadísticos. Asegurar la precisión y confiabilidad del modelo antes de su integración en la infraestructura tecnológica de la empresa. *FASE 4*	***Selección de herramientas y bibliotecas de aprendizaje automático:*** *Investigar y seleccionar las herramientas más adecuadas para implementar el modelo de inteligencia artificial.* ***Entrenamiento del modelo:*** *Utilizar el conjunto de datos recopilado para entrenar el modelo predictivo, aplicando técnicas avanzadas como análisis de series temporales y métodos estadísticos.* ***Validación del modelo:*** *Probar el modelo utilizando datos de prueba para evaluar su precisión y confiabilidad. Ajustar el modelo según sea necesario para alcanzar un nivel aceptable de precisión.*	*Prototipo del modelo predictivo.* *Resultados de las pruebas de validación, incluidos análisis de precisión y recomendaciones para mejoras.*
Integrar el sistema de predicción de fallas en la infraestructura tecnológica existente de la empresa Sistecredito. *FASE 5*	***Desarrollo de un plan de integración:*** *Crear un plan detallado que describa cómo se implementará el sistema de predicción en la infraestructura existente. Esto incluye identificar posibles riesgos y estrategias de*	*Plan de integración detallado.* *Resultados de las pruebas de integración y evaluación del rendimiento del sistema en un entorno controlado.*

	mitigación. ***Pruebas de integración en un entorno controlado:*** *Antes de la implementación completa, probar el sistema en un entorno controlado para asegurar su compatibilidad con otros sistemas y evaluar su rendimiento.* ***Implementación del sistema de predicción:*** *Integrar el sistema en la infraestructura tecnológica de Sistecredito y proporcionar capacitación al personal clave*	

Analizar los resultados obtenidos del sistema de predicción para identificar áreas de mejora y tomar decisiones estratégicas. FASE 5-6	***Monitoreo del rendimiento del sistema durante los primeros 6 meses:*** *Supervisar el sistema para identificar problemas potenciales, recoger datos sobre su funcionamiento y medir el impacto en la reducción de fallos.* ***Análisis exhaustivo para identificar áreas de mejora:*** *Con base en el monitoreo, analizar los datos para identificar áreas donde se pueda mejorar la precisión del sistema y reducir aún más las fallas.* ***Desarrollo de un plan de mejora:*** *Con la información recopilada, crear un plan estratégico para optimizar el sistema y corregir las deficiencias identificadas.*	*Informe de análisis de rendimiento, que incluya hallazgos clave y recomendaciones.* *Plan de mejora para la optimización del sistema, con acciones concretas y plazos definidos.*

Nota: Matriz metodología Crisp- DM; elaboración propia, datos tomados de (IBM, 2021)

El marco metodológico propuesto para el desarrollo del sistema de predicción de fallas en unidades de disco de servidores de la empresa Sistecredito incluye una serie de pasos claramente definidos para alcanzar el objetivo general de optimizar la gestión de la infraestructura tecnológica y mejorar la disponibilidad y rendimiento de los servicios de la empresa. A continuación, se describe el enfoque metodológico y la forma en que se implementarán las actividades para alcanzar los objetivos específicos.

Recopilación y Análisis de Datos Históricos

El primer paso consiste en recopilar y analizar datos históricos relacionados con fallas en unidades de disco en servidores. Para ello, se identificarán y obtendrán registros internos, informes de mantenimiento y otros documentos relevantes para construir un conjunto de datos representativo. El proceso incluye la definición de parámetros clave, como fechas, tipos de fallos, duración de interrupciones y condiciones de uso. La información será luego recopilada y limpiada para eliminar duplicados y errores, asegurando un conjunto de datos limpio y estructurado.

Implementación de un Modelo de Inteligencia Artificial

El siguiente paso es utilizar herramientas y bibliotecas de aprendizaje automático para implementar un modelo de inteligencia artificial que permita predecir posibles fallos en las unidades de disco. Se emplearán técnicas avanzadas de análisis predictivo, como series temporales y métodos estadísticos, para entrenar el modelo. Luego, se validará la

precisión y confiabilidad del modelo utilizando datos de prueba, ajustándolo según sea necesario para garantizar un alto nivel de precisión antes de su integración en la infraestructura tecnológica de la empresa.

Integración del Modelo en la Infraestructura Tecnológica

Una vez validado el modelo predictivo, se procede a desarrollar un plan detallado para su integración en la infraestructura existente de Sistecredito. Se realizarán pruebas de integración en un entorno controlado para garantizar la compatibilidad y evaluar el rendimiento. El proceso también incluye la capacitación del personal clave para asegurar una implementación fluida y efectiva. Una vez completada la integración, se monitoreará el rendimiento del sistema para detectar problemas y medir el impacto en la reducción de fallos.

Análisis de Resultados y Mejora Continua

Finalmente, se analizarán los resultados obtenidos del sistema de predicción para identificar áreas de mejora y tomar decisiones estratégicas para optimizar el rendimiento. El monitoreo durante los primeros 6 meses permitirá detectar posibles problemas y ajustar el sistema para mejorar su precisión y confiabilidad. Se desarrollará un plan de mejora basado en este análisis, con acciones concretas y cronogramas para garantizar una mejora continua en el rendimiento del sistema.

Este enfoque metodológico se basa en la recopilación y análisis de datos, el uso de herramientas de aprendizaje automático, la implementación y validación de un modelo predictivo, y la integración y monitoreo del sistema en la infraestructura tecnológica existente. El proceso se centra en garantizar una implementación fluida, con un enfoque en la precisión, confiabilidad y mejora continua para mantener la disponibilidad y rendimiento de los servicios de Sistecredito.

7. Marco referencial

7.1. Marco Teórico

Variables y Factores que Afectan el Rendimiento de la Infraestructura Tecnológica en Empresas de Bienes y Servicios

Las empresas de bienes y servicios enfrentan varios desafíos relacionados con el rendimiento de su infraestructura tecnológica. La planificación adecuada y el soporte técnico son fundamentales para asegurar un buen rendimiento. Sin embargo, hay factores que pueden afectar este rendimiento, como la ciberseguridad, la disponibilidad de personal

capacitado y el costo de implementación y mantenimiento.

Un aspecto importante es la necesidad de expertos en ciberseguridad, lo que requiere inversión y formación especializada. Según un estudio, "Si bien, aún es un tema en explotación, este engloba un número incalculable de técnicas y herramientas que hacen frente a los riesgos de la tecnología y la comunicación. A todo esto, se recalca la importancia de la ciberseguridad como factor a invertir y la necesidad de expertos en los que se debe fomentar su formación" .(Altamirano, 2019).Este aspecto es crucial para prevenir interrupciones y garantizar la continuidad del servicio.

Además, la falta de personal capacitado puede limitar el rendimiento de la infraestructura tecnológica. Este problema es especialmente relevante en lugares con menor desarrollo tecnológico, donde las empresas pueden tener dificultades para encontrar profesionales calificados. Las universidades juegan un papel crucial en la formación de talento local, pero en América Latina, el modelo educativo todavía tiene influencias europeas, lo que puede no ser óptimo para las necesidades locales. (Amestoy, 2023). Por otro lado, la falta de docentes capacitados para enseñar competencias digitales puede limitar el desarrollo de profesionales en infraestructura tecnológica. (Bertha Lidia Torres Martínez et al., 2021).

Por último, el costo de implementación y mantenimiento de la infraestructura tecnológica es otro factor clave. Muchas pequeñas y medianas empresas no pueden permitirse sistemas tecnológicos avanzados, lo que limita su capacidad para competir con empresas más grandes. Además, el costo del soporte técnico y el mantenimiento puede ser alto, especialmente si se necesita personal especializado. (Bertoni & Bertoni, 2022)**.**

Aplicación de Machine Learning para la Mitigación de Fallas por Bajo Rendimiento en la Infraestructura Tecnológica

El uso de técnicas de Machine Learning (ML) y algoritmos de inteligencia artificial (IA) ha sido ampliamente adoptado para la mitigación de fallas en la infraestructura tecnológica. Los modelos de aprendizaje automático pueden ayudar a identificar y predecir fallas en tiempo real, lo que permite tomar medidas preventivas y mejorar el rendimiento del sistema.

Diversas metodologías de ML se aplican para la predicción de fallas y detección de anomalías. Algunas técnicas incluyen regresión logística, análisis de componentes principales (PCA-Q) y máquinas de vectores de soporte. Estos enfoques permiten identificar anomalías y ajustar la medición de confiabilidad del sistema en tiempo real .(Zaninetti, 2019).

Además, se han utilizado técnicas de aprendizaje conjunto, como el boosting, bagging y stacking, para mejorar la precisión en la predicción y recuperación de datos. Los métodos de mantenimiento predictivo también se basan en el uso de dispositivos de monitoreo y datos históricos para anticipar fallas y planificar acciones correctivas. (Rousopoulou et al., 2022). Los Digital Twins o gemelos digitales son otra aplicación innovadora de ML, permitiendo simular y monitorear en tiempo real el comportamiento de sistemas y componentes tecnológicos .(Aheleroff et al., 2021).

Consecuencias Presentadas por Fallas de Componentes Asociados a la Infraestructura Tecnológica en Empresas de Bienes y Servicios

Las fallas en componentes de la infraestructura tecnológica pueden tener serias consecuencias para las empresas de bienes y servicios. El manejo inadecuado de datos puede llevar a problemas de seguridad y privacidad, lo que puede afectar la confianza de los clientes y el cumplimiento de regulaciones. Por ejemplo, en entornos industriales, la seguridad se vuelve crítica, ya que las tecnologías operativas (OT) ahora están conectadas a redes abiertas, aumentando el riesgo de ciberataques. (Pléta et al., 2020).

Además, las fallas pueden afectar la disponibilidad de servicios críticos, lo que puede resultar en pérdidas financieras significativas y daños a la reputación de la empresa. Un estudio muestra que las pruebas de datos ideales deberían devolver resultados claros y consistentes, pero las fallas en el ciclo de vida de la recopilación de datos pueden llevar a resultados impredecibles. (Hutchinson et al., 2021).

Los costos asociados con la implementación y mantenimiento de la infraestructura tecnológica también son una consecuencia significativa de las fallas. Con la digitalización, el costo de producir bienes de información y el almacenamiento digital ha disminuido, pero las empresas deben seguir innovando para mantenerse competitivas y minimizar el impacto de la fallas. (Miller et al., 2023). Además, las fallas pueden desencadenar consecuencias negativas en la industria, especialmente en empresas pequeñas o medianas, donde los cambios culturales y la falta de conocimientos adecuados pueden ser barreras importantes (WSJ).

7.2. Marco Conceptual

El mundo actual se caracteriza por la generación masiva de datos, la complejidad de los sistemas y la necesidad de una gestión eficiente y eficaz. En este contexto, el marco conceptual propuesto integra conceptos clave como Big Data, Machine Learning, Gemelos Digitales, Infraestructura Tecnológica, Internet de las Cosas, AIOps e ITIL para ofrecer una

visión holística para abordar los desafíos y oportunidades de la era digital.

Big Data: se refiere a grandes volúmenes de datos de diversos formatos, velocidades y niveles de estructuración, que no pueden ser procesados por métodos tradicionales. El análisis de Big Data permite extraer información valiosa para la toma de decisiones estratégicas, la optimización de procesos y la innovación. (Labbé Figueroa, 2020).

Machine Learning: es una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de datos. Machine Learning se utiliza en diversas aplicaciones, como el reconocimiento de patrones, la predicción y la automatización de tareas. (Ahmed & Green II, 2022).

Gemelos Digitales: son réplicas virtuales de activos físicos o procesos que permiten simular y analizar su comportamiento en el mundo real. Los Gemelos Digitales son de gran utilidad para la predicción de fallas, la optimización del rendimiento y la toma de decisiones basadas en datos. (Koning et al., 2023).

Infraestructura Tecnológica: comprende el conjunto de herramientas informáticas, redes y sistemas que soportan las operaciones de una organización. Una infraestructura tecnológica robusta y escalable es esencial para el éxito en la era digital.

Internet de las Cosas (IoT): es una red de objetos físicos conectados a internet que intercambian datos. El IoT permite recopilar información en tiempo real sobre el entorno físico, lo que abre nuevas posibilidades para la automatización, la optimización y la innovación.(Shumba et al., 2023).

AIOps: combina la inteligencia artificial con las operaciones de TI para automatizar tareas, mejorar la eficiencia y optimizar el rendimiento de la infraestructura tecnológica. AIOps permite a las organizaciones gestionar de forma proactiva sus sistemas de TI y prevenir problemas antes de que ocurran.(Kehn, n.d.).

ITIL: es una biblioteca de mejores prácticas para la gestión de servicios de TI. ITIL proporciona un marco de referencia para la planificación, entrega y soporte de servicios de TI de alta calidad. (Pailiacho et al., 2019).

El marco conceptual propuesto integra estos conceptos de manera sinérgica para ofrecer una visión holística de la gestión de datos, sistemas y procesos en la era digital. Big Data y Machine Learning proporcionan las herramientas para analizar y extraer información de grandes volúmenes de datos. Los Gemelos Digitales permiten crear modelos virtuales de activos físicos y procesos, lo que facilita la simulación, predicción y

optimización. La Infraestructura Tecnológica proporciona la base para soportar las operaciones digitales. El IoT permite conectar objetos físicos a la red y recopilar datos en tiempo real. AIOps automatiza tareas de TI y optimiza el rendimiento de la infraestructura tecnológica. ITIL proporciona un marco de referencia para la gestión de servicios de TI de alta calidad.

El marco conceptual integrado propuesto ofrece una guía valiosa para las organizaciones que buscan abordar los desafíos y oportunidades de la era digital. Al integrar Big Data, Machine Learning, Gemelos Digitales, Infraestructura Tecnológica, Internet de las Cosas, AIOps e ITIL, las organizaciones pueden aprovechar el poder de los datos para mejorar la toma de decisiones, optimizar procesos, innovar y crear valor para sus clientes.

7.3. Marco Normativo

El mundo actual se caracteriza por la generación masiva de datos, la complejidad de los sistemas y la necesidad de una gestión eficiente y eficaz. En este contexto, se hace necesario un marco normativo que establezca pautas y lineamientos para la gestión de datos, sistemas y procesos en la era digital. Este marco normativo debe ser integral, abarcando aspectos técnicos, organizacionales y legales.

El marco normativo tiene como objetivos los siguientes

- Garantizar la calidad y seguridad de los datos: Los datos son un activo fundamental para las organizaciones, por lo que es necesario establecer mecanismos para garantizar su calidad, integridad, confidencialidad y disponibilidad.
- Optimizar la gestión de los sistemas: Los sistemas informáticos son esenciales para el funcionamiento de las organizaciones, por lo que es necesario establecer lineamientos para su gestión eficiente y eficaz.
- Mejorar la eficiencia de los procesos: Los procesos de negocio son la base del funcionamiento de las organizaciones, por lo que es necesario establecer mecanismos para su optimización y mejora continua.
- Promover la innovación: Los datos y las tecnologías digitales pueden ser una fuente de innovación para las organizaciones, por lo que es necesario establecer un marco normativo que fomente su uso responsable y ético.

El marco normativo se basa en los siguientes principios

- Enfoque centrado en el valor: La gestión de datos, sistemas y procesos debe estar orientada a la creación de valor para las organizaciones.
- Responsabilidad: Las organizaciones son responsables de la gestión de sus datos, sistemas y procesos.
- Transparencia: Las organizaciones deben ser transparentes en cuanto a la forma en que gestionan sus datos, sistemas y procesos.
- Rendición de cuentas: Las organizaciones deben rendir cuentas de la gestión de sus datos, sistemas y procesos.
- Respeto por los derechos individuales: La gestión de datos, sistemas y procesos debe respetar los derechos individuales, como el derecho a la privacidad y el derecho de acceso a la información.

El marco normativo se compone de los siguientes elementos

- **Políticas:** Las políticas establecen los principios generales que guiarán la gestión de datos, sistemas y procesos.
- **Normas:** Las normas establecen los requisitos específicos que deben cumplir las organizaciones para la gestión de datos, sistemas y procesos.
- **Procedimientos:** Los procedimientos establecen los pasos específicos que deben seguir las organizaciones para cumplir con las normas.
- **Guías:** Las guías proporcionan recomendaciones y mejores prácticas para la gestión de datos, sistemas y procesos.
- **Controles:** Los controles son mecanismos que permiten verificar que las organizaciones cumplan con las normas.

A continuación, se muestra un esquema de gobierno de datos según la norma ISO 38505

Figura 1. Esquema de gobierno de datos según la norma ISO 3805. Tomado de: https://revista.une.org/46/buen-gobierno-del-dato-gracias-a-los-estandares.html

8. Desarrollo del proyecto

8.1 Caracterización del proceso de modelado predictivo

Comentado [4]: Revisar tipos de titulos

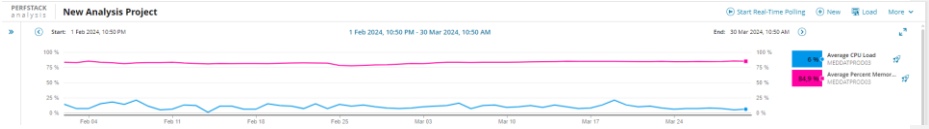
En este apartado se describen las actividades realizadas para lograr la caracterización del proceso de modelado predictivo aplicado al comportamiento de componentes de la infraestructura tecnológica en empresas de bienes y servicios.

8.1.1 Determinar el propósito del análisis (FASE 1).

El propósito del análisis está motivado en validar la posibilidad de optimizar la gestión de la infraestructura tecnológica de empresas de bienes y servicios, mejorando la disponibilidad y rendimiento de los servicios. Como indicadores de rendimiento tenemos el tiempo de actividad del sistema, tiempo de respuesta de los componentes, tasa de fallos, costos de mantenimiento, y satisfacción del usuario.

8.1.1 Caracterizar el proceso relacionado con el modelado predictivo (FASE 1).

Figura 2. Gráfico de línea de un proyecto de análisis titulado "New Analysis Project"



Nota. La imagen muestra un gráfico de línea de un proyecto de análisis titulado "New Analysis Project", que abarca el periodo del 1 de febrero de 2024 al 30 de marzo de 2024. El gráfico está dividido en dos líneas que representan diferentes métricas de un servidor

Fuente: Elaboración propia

En la figura 2 Línea rosa: Representa el "Average CPU Load", que se mantiene constante en un 6% a lo largo de todo el periodo.

Línea azul: Representa el "Average Percent Memory Used", que fluctúa a lo largo del tiempo, generalmente manteniéndose alrededor del 85%.

El gráfico tiene una escala del 0% al 100%, con el tiempo distribuido en intervalos aproximadamente semanales en el eje horizontal. Este tipo de visualización es útil para monitorear el rendimiento de los sistemas y asegurar que operen dentro de los parámetros normales, identificando posibles desviaciones o problemas en el uso de recursos como la CPU y la memoria.

8.1.2 Preparar la data (FASE 2)

Recolección de los datos

La recolección de datos se realizó mediante la herramienta de monitoreo que tiene la empresa para la gestión de eventos de los servidores. La data entregada contenía dos hojas en la primera hoja se tenía el registro de los 3 servidores con el registro de los datos del comportamiento del uso de la cpu y la memoria en porcentajes y por intervalos de una hora. En la segunda hoja se encontraba el registro del porcentaje de uso del disco duro de cada servidor.

Una vez se accedió a la base de datos y se conocieron los atributos, el siguiente paso consistió en el análisis y estructura de estos para determinar qué atributo ayudaba a realizar la clasificación.

Tabla 2.

Descripción de los campos de la base de datos

Variable	Tipo	Descripción
Caption	Texto	Contiene el identificador de la entidad monitorizada. Permite segmentar el análisis por entidad y correlacionar el rendimiento o el uso de recursos con unidades específicas dentro de la infraestructura.

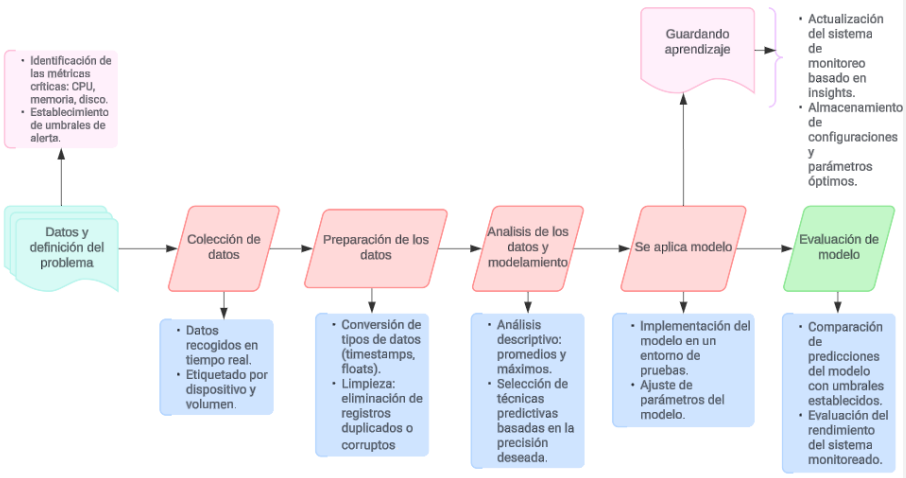
Comentado [5]: Crea las tablas como dije, pero luego debes ajustar el tamaño y tipode letra, para que quede uniforme con tu trabajo

Comentado [6]: seguro todo es texto, algunas paracen numericas, de timpo y otras tipo flotante

Comentado [7]: seguro todo es texto, algunas paracen numericas, de timpo y otras tipo flotante

Timestamp	Fecha y hora	Fecha y hora de la observación o del registro de datos. Fundamental para análisis temporales, permite rastrear cambios, tendencias y patrones a lo largo del tiempo.
Average CPU Load	Numérico	Carga promedio de CPU durante un período de tiempo específico. Indicador clave de la demanda de procesamiento y eficiencia del sistema, útil para predecir sobrecargas y planificar capacidades.
Average Percent Memory Used	Porcentaje	Porcentaje promedio de memoria utilizada. Permite evaluar si los recursos de memoria son adecuados o si existe riesgo de saturación que podría afectar el rendimiento.
Volume Name	Texto	Nombre del volumen de almacenamiento evaluado. Clave para identificar y diferenciar los recursos de almacenamiento dentro de un sistema, facilitando la gestión y el análisis específico por volumen.
Disk Size	Numérico	Tamaño del disco de almacenamiento. Proporciona un contexto para evaluar el uso del disco y planificar futuras necesidades de almacenamiento.
Average Disk Used	Numérico	Uso promedio del disco durante un período específico. Indicador crucial para la planificación de la capacidad y la optimización del almacenamiento.
Maximum Disk Used	Numérico	Máximo uso registrado del disco en un período específico. Esencial para identificar picos de uso que podrían indicar necesidades excepcionales o problemas de rendimiento.
Percent used	Porcentaje	Porcentaje del disco utilizado. Métrica importante para la monitorización de la salud y la eficiencia del almacenamiento, crucial para la toma de decisiones sobre escalabilidad y mejoras.

Figura 3. Diagrama de Arquitectura



Fuente: Elaboración propia

En el diagrama anterior se puede evidenciar la arquitectura de cómo se procesan los datos. Donde se inicia con la clasificación de los datos, etiquetándolos para construir con colección de estos, luego se pasa a la preparación, que permite realizar una limpieza, eliminar registros duplicados, la conversión de los tipos de datos. Posteriormente, se aplican varias técnicas de ML como Isolation Forest, Regresión Lineal y logística y Matriz de confusión para así escoger la más eficiente, así como también la ecuación del R2 para finalmente proceder a la evaluación de los resultados de la mejor técnica y se guardan el aprendizaje del modelo.

8.1.3 Preparar la data (FASE 3)

- **Preparación y limpieza de los datos**

  En esta sección, se identifican las variables, se usan y analizan los datos en busca de información irrelevante, como la redundancia de datos duplicados, registros nulos y datos incorrectos que no contribuyen de manera significativa al desarrollo del proyecto.

- **Conversión de tipos de datos**

  En este apartado se transforman las columnas de 'Timestamp' a formato datetime y se convierten los datos de carga de CPU, uso de memoria y uso de disco de formato porcentual y de almacenamiento (% y GB) a valores numéricos float, se ajustaron para su uso en el modelo.

Tabla 3.

Campos de clasificados

VARIABLE	*TIPO*	*NATURALEZA*	*ESCALA*	*OBSERVACIÓN*
Caption	*Categórico*	*Cualitativo*	*Nominal*	*Politómicas*
Timestamp	*Numérica*	*Cuantitativa*	*Intervalo*	*Discretas*
Average CPU Load	*Numérica*	*Cuantitativa*	*Intervalo*	*Discretas*
Average Percent Memory Used	*Numérica*	*Cuantitativa*	*De razón*	*Discreta*
Volume Name	*Categórico*	*Cualitativa*	*Nominal*	*Politómicas*
Disk Size	*Numérica*	*Cuantitativa*	*De razón*	*Discreta*
Average Disk Used	*Numérica*	*Cuantitativa*	*De razón*	*Discreta*
Maximum Disk Used	*Numérico*	*Cuantitativo*	*De razón*	*Discreta*
Percent used	*Numérico*	*Cuantitativo*	*De razón*	*Discreta*

- **Limpieza de datos (Data Cleaning)**

  En este punto, se eliminaron los caracteres no numéricos y la columna redundante: Caption_y, se manejaron valores faltantes, se eliminaron duplicados y se corrigieron valores atípicos.

- **Separar los datos y agrupar los datos**

  En esta etapa se separarán y agrupan los datos de cada uno de los servidores creando un conjunto de datos por hoja (Servidor1, Servidor2, Servidor3) lo que resultara en tres conjuntos por servidor, esto con el fin de tener los datos organizados por servidor.

8.2 Aplicación de ML para generar modelos que permitan a partir de su evaluación seleccionar el modelo más eficiente para predicción de fallos. (Fase 4)

En esta etapa y para dar respuesta al objetivo específico número dos se aplicaron las técnicas como el algoritmo de decisión (Isolation Forest), la Matriz de confusión, la Regresión líneal y logística y el coeficiente de determinación (R^2). Se uso la herramienta Google Collaboratory también conocida como Colab que permite escribir y ejecutar código desde una interfaz web para implementar modelos con el lenguaje de programación Python. Las librerías utilizadas fueron: Numpy, Pandas, Joblib, Matplotlib, Reload, Seaborn, Statsmodels, Sklearn,

Los principales objetivos de la implementación fueron:

- Desarrollo de un Modelo Predictivo: para esto se utilizan diferentes técnicas de aprendizaje automático, como Isolation Forest, para identificar y seleccionar el modelo más eficiente para la predicción de fallos en la infraestructura tecnológica.
- Evaluación y Validación del Modelo: Evaluar la precisión y eficacia de los modelos mediante el uso de métricas, como la matriz de confusión y el coeficiente de determinación (R^2). Validar que el modelo seleccionado cumpla con los requisitos de predicción de fallos.
    - Para el desarrollo de estos objetivos, se realizó la importación de las librerías en Python de acuerdo a las funciones que van a desempeñar en el código:
    - Tratamiento de los datos: NumPy, Pandas, Joblib.

- Graficar los resultados: Matplotlib y Seaborn.
- Pre-procesado y modelado: Sklearn y Statsmodels

Posteriormente se cargó el DataSet, ya que este facilita tenerlo siempre a la mano:

Figura 4. Importación de archivo y lectura de las hojas

```
[2]  #Se crea el dataframe d con los datos obtenidos de archivo de entrada
     file_path=('./datasets/Report.xlsx')
```

```
[9]  # Leer las hojas del archivo para ver cómo están organizados los datos
     with pd.ExcelFile(file_path) as xls:
         sheet_names = xls.sheet_names
         print("Hojas presentes en el archivo:", sheet_names)
```

```
Hojas presentes en el archivo: ['Sheet1', 'Sheet2']
```

Fuente: Elaboración propia

Figura 5. Mostrar el contenido de las hojas

```
[10] for sheet in sheet_names:
         df = pd.read_excel(file_path, sheet_name=sheet)
         print(f"\nContenido de la hoja '{sheet}':\n", df)
```

```
Contenido de la hoja 'Sheet1':
         Caption        Timestamp Average CPU Load Average Percent Memory Used
0     Servidor1   1/2/2024 00:00          18.00 %                     69.49 %
1     Servidor2   1/2/2024 00:00           3.00 %                     68.79 %
2     Servidor3   1/2/2024 00:00          11.00 %                     94.10 %
3     Servidor1   1/2/2024 01:00          13.00 %                     66.08 %
4     Servidor2   1/2/2024 01:00           1.00 %                     67.70 %
...         ...              ...             ...                        ...
1962  Servidor2  29/2/2024 13:00           2.83 %                     61.18 %
1963  Servidor3  29/2/2024 13:00          33.17 %                     85.21 %
1964  Servidor1  29/2/2024 14:00           6.00 %                     65.56 %
1965  Servidor2  29/2/2024 14:00           4.17 %                     62.82 %
1966  Servidor3  29/2/2024 14:00          31.17 %                     87.82 %

[1967 rows x 4 columns]

Contenido de la hoja 'Sheet2':
         Caption                                Volume Name        Timestamp  \
0     Servidor1          Servidor1-C:\ Label:OS FC3BCC7C   1/2/2024 00:00
1     Servidor1          Servidor1-C:\ Label:OS FC3BCC7C   1/2/2024 01:00
2     Servidor1          Servidor1-C:\ Label:OS FC3BCC7C   1/2/2024 02:00
3     Servidor1          Servidor1-C:\ Label:OS FC3BCC7C   1/2/2024 03:00
4     Servidor1          Servidor1-C:\ Label:OS FC3BCC7C   1/2/2024 04:00
...         ...                                      ...              ...
5242  Servidor3  Servidor3-J:\ Label:TempDB_R10 52CEE82B  29/2/2024 10:00
5243  Servidor3  Servidor3-J:\ Label:TempDB_R10 52CEE82B  29/2/2024 11:00
5244  Servidor3  Servidor3-J:\ Label:TempDB_R10 52CEE82B  29/2/2024 12:00
5245  Servidor3  Servidor3-J:\ Label:TempDB_R10 52CEE82B  29/2/2024 13:00
5246  Servidor3  Servidor3-J:\ Label:TempDB_R10 52CEE82B  29/2/2024 14:00

      Disk Size Average Disk Used Maximum Disk Used Percent used
0     119.88 GB          79.91 GB          79.93 GB      56.75 %
1     119.88 GB          79.84 GB          79.89 GB      56.75 %
2     119.88 GB          79.84 GB          79.85 GB      56.75 %
3     119.88 GB          79.85 GB          79.88 GB      56.75 %
4     119.88 GB          79.84 GB          79.84 GB      56.75 %
...         ...              ...               ...          ...
5242  199.98 GB          72.09 GB          72.09 GB      36.05 %
5243  199.98 GB          72.09 GB          72.09 GB      36.05 %
5244  199.98 GB          72.09 GB          72.09 GB      36.05 %
5245  199.98 GB          72.09 GB          72.09 GB      36.05 %
5246  199.98 GB          72.09 GB          72.09 GB      36.05 %

[5247 rows x 7 columns]
```

Fuente: Elaboración propia

Figura 6. Descripción de variables

```
[11] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5247 entries, 0 to 5246
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Caption            5247 non-null   object
 1   Volume Name        5247 non-null   object
 2   Timestamp          5247 non-null   object
 3   Disk Size          5247 non-null   object
 4   Average Disk Used  5247 non-null   object
 5   Maximum Disk Used  5247 non-null   object
 6   Percent used       5247 non-null   object
dtypes: object(7)
memory usage: 287.1+ KB
```

Fuente: Elaboración propia

Nota: Descripción de las variables, el comando dataset.info () nos describe las columnas, sus etiquetas, el tipo de datos y el uso de la memoria.

Figura 6. Separar y agrupar los datos de los Servidores por hojas

```
[ ] # Separar los datos por servidor y hoja
    for sheet_name, data in data_sheets.items():
        for server in servers:
            # Filtrar los datos por el nombre del servidor en la columna 'Caption'
            data_by_server[server][sheet_name] = data[data['Caption'] == server]
```

```
# Acceder a los datos para Servidor1 en Sheet1
print(data_by_server['Servidor1']['Sheet1'].head())
```

```
      Caption       Timestamp Average CPU Load Average Percent Memory Used
0   Servidor1  1/2/2024 00:00          18.00 %                     69.49 %
3   Servidor1  1/2/2024 01:00          13.00 %                     66.08 %
6   Servidor1  1/2/2024 02:00          18.00 %                     66.42 %
9   Servidor1  1/2/2024 03:00          12.00 %                     66.25 %
12  Servidor1  1/2/2024 04:00          10.00 %                     66.23 %
```

Fuente: Elaboración propia

Nota: Se toma muestra de cómo se ven los datos del Servidor1

Figura 7. Estadísticas descriptivas de tres variables: Average CPU Load, Average Percent Memory Used, y Percent used

	Average CPU Load	Average Percent Memory Used	Percent used
count	1308.000000	1308.000000	1308.000000
mean	4.530275	71.271713	29.120000
std	5.972511	4.887053	27.640568
min	0.000000	46.160000	1.490000
25%	1.000000	69.640000	1.490000
50%	2.165000	71.510000	29.120000
75%	6.000000	73.010000	56.750000
max	40.000000	86.110000	56.750000

Fuente: Elaboración propia

Nota: En promedio, el uso de memoria es alto (71.27%), mientras que la carga de CPU y el almacenamiento en disco varían significativamente, indicando diferentes patrones de uso

Figura 8. Combinar datos

.

```
[ ] # Diccionario para almacenar los datos combinados por servidor
    combined_data_by_server = {server: pd.DataFrame() for server in servers}
```

```
[ ] # Combinar los datos de cada servidor de todas las hojas
    for server in servers:
        frames = []  # Lista para almacenar DataFrames temporales
        for sheet_name, data in data_sheets.items():
            filtered_data = data[data['Caption'] == server]
            frames.append(filtered_data)
        # Concatenar los DataFrames
        combined_data_by_server[server] = pd.concat(frames, ignore_index=True)
```

```
[ ] # Guardar los datos combinados en un nuevo archivo Excel
    with pd.ExcelWriter('./datasets/Combined_Report.xlsx') as writer:
        for server, data in combined_data_by_server.items():
            data.to_excel(writer, sheet_name=server, index=False)
```

```
[ ] # Combinar los datos de cada servidor de todas las hojas por Timestamp
    for server in servers:
        server_data = None
        for sheet_name, data in data_sheets.items():
            # Filtrar los datos por el nombre del servidor
            filtered_data = data[data['Caption'] == server]
            # Si es el primer conjunto de datos, inicializalo
            if server_data is None:
                server_data = filtered_data
            else:
                # Si no es el primero, combínalo con el existente usando merge en 'Timestamp'
                server_data = pd.merge(server_data, filtered_data, on='Timestamp', how='outer')
        combined_data_by_server[server] = server_data
```

Fuente: Elaboración propia

Nota: Se concatenan los datos de cada hoja en una única DataFrame para cada servidor. Finalmente, procedemos a guardar estos datos combinados en un nuevo archivo Excel.

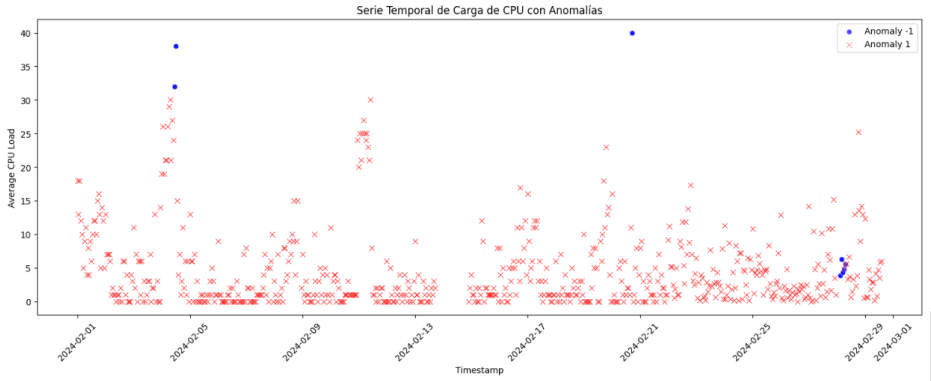
Figura 9. Datos agrupados de Servidor1

Primero se realiza un análisis exploratorio con el fin de entender cómo están los registros del dataset.

8.2.1 Aplicación de las técnicas de ML y construcción del modelo

En esta etapa, se aplicaron las técnicas de ML seleccionadas para construir el modelo. A continuación, se describe el proceso para aplicar el algoritmo de aprendizaje automático Isolation Forest el cual fue de gran ayuda para detectar anomalías.

Figura 10. Grafica de la serie temporal con las anomalías con respecto a la CPU



Fuente: Elaboración propia

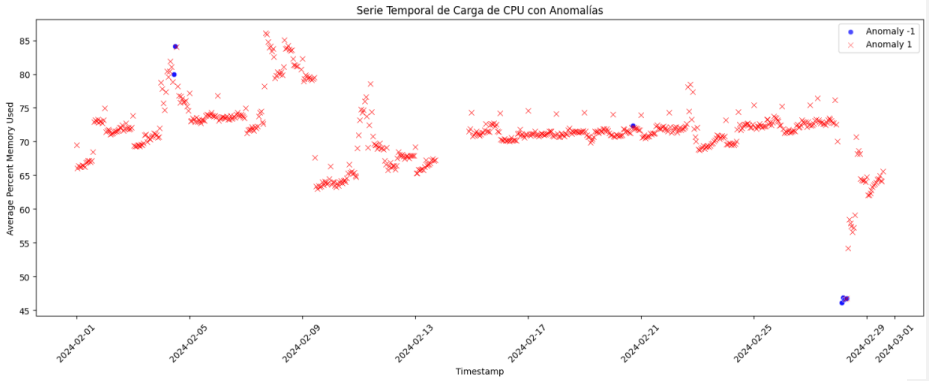
Nota: El eje X (horizontal): Muestra la línea de tiempo, que cubre desde principios de febrero hasta principios de marzo de 2024.Eje Y (vertical): Eje Y (vertical): Indica el porcentaje de carga promedio de la CPU, desde 0% hasta 40%.
Marcadores de Anomalías:
Anomaly -1 (puntos azules): Indican valores clasificados como normales según el modelo.
Anomaly 1 (cruces rojas): Representan valores clasificados como anómalos.

El modelo identificó La carga de la CPU varía significativamente con un patrón predominante de valores bajos (0% - 5%).

Hay picos esporádicos de carga más alta que superan el 30%, que son detectados

como anómalos. varias anomalías a lo largo del período mostrado. Esto sugiere que hay picos de uso de memoria que podrían ser sintomáticos de un problema en el rendimiento. También se observa que hay un uso de memoria que el modelo considera normal (puntos azules), lo que ayuda a establecer una base para detectar anomalías. El uso de memoria fluctúa de forma regular, pero hay picos que el modelo marca como anómalos, lo cual podría requerir una mayor investigación, por lo que se recomienda a la empresa realizar una investigación detallada de las anomalías para determinar su causa raíz y evaluar si las anomalías representan riesgos para la operación o el rendimiento del sistema.

Figura 11. Grafica de la serie temporal con las anomalías con respecto a la memoria



Fuente: Elaboración propia

Nota: El eje X (horizontal): Muestra la línea de tiempo, que cubre desde principios de febrero hasta principios de marzo de 2024.Eje Y (vertical): Representa el porcentaje promedio de memoria utilizada, desde aproximadamente 45% hasta 85%.
Marcadores de Anomalías:
Anomaly -1 (puntos azules): Indican valores clasificados como normales según el modelo.
Anomaly 1 (cruces rojas): Representan valores clasificados como anómalos.

El modelo identificó varias anomalías a lo largo del período mostrado. Esto sugiere que hay picos de uso de memoria que podrían ser sintomáticos de un problema en el rendimiento. También se observa que hay un uso de memoria que el modelo considera normal (puntos azules), lo que ayuda a establecer una base para detectar anomalías. El uso de memoria fluctúa de forma regular, pero hay picos que el modelo marca como anómalos, lo cual podría requerir una mayor investigación, por lo que se recomienda a la empresa realizar una investigación detallada de las anomalías para determinar su causa raíz y evaluar si las anomalías representan riesgos para la operación o el rendimiento del sistema.

A continuación, se calcula la Matriz de confusión en el modelo aplicado de Regresión logística para evaluar el desempeño del modelo predictivo en la clasificación de servidores en dos categorías:

a. Sin fallos (Clase 0): Servidores que funcionan correctamente según los umbrales definidos.
b. Con fallos (Clase 1): Servidores que tienen fallos basados en las condiciones de los umbrales ajustados para CPU, memoria, y disco.

Se ajustaron los umbrales para CPU, memoria y disco al 80% del valor máximo observado:

CPU: 32%

Memoria: 70%

Disco: 45%

Se creó una nueva variable objetivo-llamada Adjusted Server Failure, donde un valor de 1 indica un fallo según los umbrales ajustados, y 0 para el caso contrario.

Con respecto a la distribución la nueva variable identificó dos casos con fallos (1), mientras que el resto de los registros permanecieron como no fallidos (0).

Figura 12. Matriz de Confusión para todos los datos

```
Matriz de Confusión:
[[1306    0]
 [   0    2]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1306
           1       1.00      1.00      1.00         2

    accuracy                           1.00      1308
   macro avg       1.00      1.00      1.00      1308
weighted avg       1.00      1.00      1.00      1308
```

Fuente: Elaboración propia

La matriz muestra la distribución de predicciones correctas e incorrectas para la variable Adjusted Server Failure donde se derivan los siguientes resultados:

- Clase 0 (No fallos): 1306 casos predichos correctamente.

- Clase 1 (Fallos): 2 casos predichos correctamente.
- La Precisión, el Recall y F1-Score: Tienen un valor de 1.0, lo que indica un desempeño perfecto del modelo con respecto a este conjunto de datos.

Figura 13. Matriz de Confusión aplicado a los datos de prueba

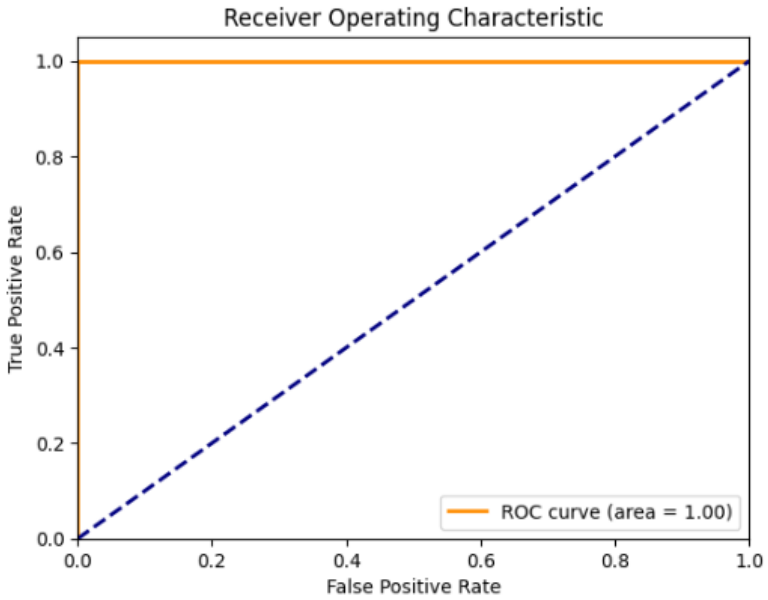
```
Matriz de Confusión:
[[261   0]
 [  1   0]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       261
           1       0.00      0.00      0.00         1

    accuracy                           1.00       262
   macro avg       0.50      0.50      0.50       262
weighted avg       0.99      1.00      0.99       262
```

Fuente: Elaboración propia

La matriz muestra los resultados del modelo aplicado a un conjunto de prueba donde se derivan los siguientes resultados:

- Clase 0: 261 casos predichos correctamente.
- Clase 1: 1 caso incorrecto.
- La Precisión, el Recall y F1-Score: La clase mayoritaria (0) tiene una precisión de 1.0, mientras que la clase 1 tiene 0.0 en precisión y recall.

Figura 14. Curva ROC



Fuente: Elaboración propia

Con respecto al Área bajo la curva (AUC): El valor es de 1.0, lo que indica un rendimiento perfecto en este conjunto.

Dando continuidad al desarrollo de los objetivos se lleva a cabo un modelo de Regresión Lineal, en el cual se realiza una validación cruzada usando los métodos train_test_split y RepeatedKFold y como métrica de evaluación el Coeficiente de determinación (R^2) y el Error cuadrático medio (MSE).

Figura 15. Matriz de correlación

	Average CPU Load	Average Percent Memory Used	\
Average CPU Load	1.000000	0.227384	
Average Percent Memory Used	0.227384	1.000000	
Average Disk Used	0.011115	0.018979	

	Average Disk Used
Average CPU Load	0.011115
Average Percent Memory Used	0.018979
Average Disk Used	1.000000

Fuente: Elaboración propia

En este análisis hay una correlación positiva baja (0.227), lo que indica una relación leve entre el uso de la CPU y el uso de la memoria. Entre CPU Load y Disk Used: La correlación es muy baja (0.011), prácticamente inexistente, lo que sugiere que no hay una relación lineal significativa entre el uso de la CPU y el uso del disco. Entre Memory Used y Disk Used: También es muy baja (0.019), indicando una falta de relación lineal significativa entre el uso de la memoria y el uso del disco.

Figura 16. Resultados de un análisis de regresión lineal mediante OLS (Ordinary Least Squares)

OLS Regression Results

Dep. Variable:	Average Disk Used	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.2669
Date:	Tue, 23 Apr 2024	Prob (F-statistic):	0.766
Time:	16:26:49	Log-Likelihood:	-6452.7
No. Observations:	1308	AIC:	1.291e+04
Df Residuals:	1305	BIC:	1.293e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	26.2837	13.816	1.902	0.057	-0.821	53.389
Average CPU Load	0.0404	0.160	0.252	0.801	-0.273	0.354
Average Percent Memory Used	0.1193	0.195	0.610	0.542	-0.264	0.503

Omnibus:	5184.600	Durbin-Watson:	3.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	194.929
Skew:	0.074	Prob(JB):	4.70e-43
Kurtosis:	1.115	Cond. No.	1.06e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.06e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Fuente: Elaboración propia

Podemos ver los resultados: R-cuadrado: 0.000. Esto indica que el modelo no explica ninguna variación en el uso promedio del disco, lo cual sugiere que las variables seleccionadas (uso de CPU y memoria) no tienen una relación lineal significativa con el uso del disco en este conjunto de datos. F-estadístico: La prueba F tiene un valor de probabilidad de 0.766, lo que significa que el modelo no es estadísticamente significativo a nivel global. Coeficientes: Intercepto: 26.2837, con un p-valor de 0.057, indicando que está cerca del umbral de ser significativo. Average CPU Load: Coeficiente de 0.0404 con

un p-valor de 0.801, lo que indica que no es un predictor significativo del uso del disco. Average Percent Memory Used: Coeficiente de 0.1193 con un p-valor de 0.542, también indicando que no es un predictor significativo. Interpretación El modelo sugiere que no hay una relación lineal significativa entre el uso de CPU, el uso de memoria y el uso del disco para los datos del "Servidor1". Esto puede ser debido a varias razones, como que otros factores no considerados en el modelo afecten más al uso del disco, o que la relación entre estas variables no sea lineal.

9. Discusión

La metodología CRISP-DM fue fundamental para crear un modelo de análisis predictivo que anticipara las fallas en unidades de disco de servidores de empresas de bienes y servicios. Durante su aplicación, se integró un conjunto de datos histórico que fue limpiado, preprocesado y estructurado para el modelado. El proceso de caracterización de los datos reveló la complejidad y diversidad de factores que afectan el rendimiento y la confiabilidad de la infraestructura tecnológica, destacando la necesidad de modelos robustos y métodos avanzados de análisis.

Las técnicas de modelado empleadas, incluyendo Isolation Forest y la regresión logística, proporcionaron una visión más profunda de las posibles fuentes de fallos en la infraestructura. Isolation Forest resultó útil para la identificación de anomalías en el rendimiento de CPU y memoria, identificando picos que podrían indicar fallas potenciales. Además, fue particularmente eficaz en detectar sobrecargas en el sistema. Por su parte, la regresión logística ofreció una buena discriminación entre servidores con y sin fallas, permitiendo ajustar umbrales específicos para distintas métricas y encontrar relaciones entre variables que influyen en la ocurrencia de fallas.

La matriz de confusión mostró que los modelos podían identificar correctamente la mayoría de los servidores con fallas, aunque la precisión fue menor en ciertas categorías, revelando áreas donde mejorar la clasificación. La curva ROC mostró un buen desempeño general, evidenciando valores de AUC significativos. Sin embargo, hubo limitaciones en la correlación entre las métricas de CPU, memoria y disco, sugiriendo la necesidad de incluir variables adicionales para mejorar la precisión predictiva. Además, no se consideraron métricas como temperatura, tráfico de red o eventos del sistema que también podrían influir en la aparición de fallas.

La aplicación de la metodología CRISP-DM es esencial para una revisión continua y una adaptación eficiente de modelos predictivos en la infraestructura tecnológica. Incluir nuevas métricas como las mencionadas mejoraría la precisión en la detección de fallas. También, reentrenar los modelos periódicamente con datos frescos permitiría mejorar la identificación de patrones y optimizar el rendimiento del sistema predictivo en su conjunto.

10. Conclusiones

Las técnicas de aprendizaje automático demostraron ser una herramienta valiosa para identificar fallas en la infraestructura tecnológica, ya que los modelos predictivos aplicados pueden mejorar la continuidad operativa y reducir los costos asociados con estas fallas. Para que la implementación de estos modelos tenga éxito, es crucial integrarlos cuidadosamente en la infraestructura tecnológica existente, asegurando una monitorización continua y un análisis de los umbrales que permita una retroalimentación frecuente para mejorar el rendimiento.

Además, es fundamental establecer un proceso de monitoreo constante, con retroalimentación regular que ajuste los modelos según las condiciones cambiantes de la infraestructura. Este proceso debe incluir un reentrenamiento periódico, en el que se revisen las métricas y los datos de manera frecuente para mejorar la precisión predictiva, capturando patrones emergentes o cambios en el comportamiento del sistema. Incluir variables adicionales también es esencial para ampliar la comprensión de las posibles causas de las fallas. Métricas como la temperatura, el tráfico de red y eventos de sistema pueden ofrecer información más completa, lo que ayudará a mejorar la precisión en la detección de fallas y proporcionará una mejor capacidad de anticipación.

El uso de modelos predictivos tiene un potencial significativo para optimizar la infraestructura tecnológica, siempre que se implementen y mantengan de manera efectiva. Requiere un enfoque estratégico que integre la retroalimentación continua, el reentrenamiento periódico y la expansión de variables para maximizar la precisión en la detección y anticipación de problemas.

Se recomienda a la empresa para su mejora continua incluir variables para ampliar

el panorama de las fallas, en estas variables se pueden incluir: trafico de red, temperatura de los dispositivos, monitoreo de servicios.

11. Referencias

Aheleroff, S., Xu, X., Zhong, R. Y., & Lu, Y. (2021). Digital Twin as a Service (DTaaS) in Industry 4.0: An Architecture Reference Model. Advanced Engineering Informatics, 47. https://doi.org/10.1016/j.aei.2020.101225

Ahmed, J., & Green II, R. C. (2022). Predicting severely imbalanced data disk drive failures with machine learning models. Machine Learning with Applications, 9, 100361. https://doi.org/10.1016/j.mlwa.2022.100361

Altamirano, M. (2019). Dialnet-ModeloParaLaGestionDeLaSeguridadDeLaInformacionYLo-6989568.

Amestoy, F. (2023). PARQUES CIENTÍFICO-TECNOLÓGICOS COMO INSTRUMENTOS DE VINCULACIÓN ENTRE LA ACADEMIA Y EL SECTOR PRODUCTIVO PARA PROMOVER EL DESARROLLO LOCAL: EL CASO DEL PARQUE CIENTÍFICO TECNOLÓGICO DE PANDO, URUGUAY. In N° (Vol. 12).

Barros, A. (2010). El comportamiento de la infraestructura tecnológica y de comunicaciones Dossier Información y Catástrofes. http://blackout.gmu.edu/

Bertha Lidia Torres Martínez, Duniesky Villareño Domínguez, Maritza Franco Perez, & Michel Araujo García. (2021). Information and communication technologies: negative consequences in the university context. EDUMECENTRO, 13.

Bertoni, M., & Bertoni, A. (2022). Designing solutions with the product-service systems digital twin: What is now and what is next? In Computers in Industry (Vol. 138). Elsevier B.V. https://doi.org/10.1016/j.compind.2022.103629

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 560–575. https://doi.org/10.1145/3442188.3445918

Kehn, D. (n.d.). Garantice el rendimiento de sus aplicaciones con AIOps Ebook.

Koning, K., Broekhuijsen, J., Kühn, I., Ovaskainen, O., Taubert, F., Endresen, D., Schigel, D., & Grimm, V. (2023). Digital twins: dynamic model-data

fusion for ecology. In Trends in Ecology and Evolution (Vol. 38, Issue 10, pp. 916–926). Elsevier Ltd. https://doi.org/10.1016/j.tree.2023.04.010

Labbé Figueroa, M. F. (2020). Big data: New challenges for competition law. Revista Chilena de Derecho y Tecnologia, 9(1), 33–63. https://doi.org/10.5354/0719-2584.2020.56897

Miller, Z., Medaiyese, O., Ravi, M., Beatty, A., & Lin, F. (2023). Hard Disk Drive Failure Analysis and Prediction: An Industry View.

Pailiacho, Verónica M., MACHADO, Paúl H., GARCÉS, E. X., & CHICAIZA, D. V. (2019). Modelo de gestión de disponibilidad de la infraestructura tecnológica. Un enfoque desde ITIL Management model of availability of the technological infrastructure. A focus from ITIL Contenido. https://www.freeitiltraining.com

Plėta, T., Tvaronavičienė, M., Casa, S. Della, & Agafonov, K. (2020). Cyber-attacks to critical energy infrastructure and management issues: overview of selected cases. Insights into Regional Development, 2(3), 703–715. https://doi.org/10.9770/ird.2020.2.3(7)

Rousopoulou, V., Vafeiadis, T., Nizamis, A., Iakovidis, I., Samaras, L., Kirtsoglou, A., Georgiadis, K., Ioannidis, D., & Tzovaras, D. (2022). Cognitive analytics platform with AI solutions for anomaly detection. Computers in Industry, 134. https://doi.org/10.1016/j.compind.2021.103555

Shumba, A. T., Montanaro, T., Sergi, I., Bramanti, A., Ciccarelli, M., Rispoli, A., Carrizzo, A., De Vittorio, M., & Patrono, L. (2023). Wearable Technologies and AI at the Far Edge for Chronic Heart Failure Prevention and Management: A Systematic Review and Prospects. In Sensors (Vol. 23, Issue 15). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/s23156896

Zaninetti, L. (2019). The truncated lindley distribution with applications in astrophysics. Processes, 7(3). https://doi.org/10.3390/PR7030164