Mantenimiento 4.0: Diseño de Modelo Predictivo para la Definición de Estrategias de Mantenimiento en la Industria Oíl & Gas.

Didier Sneider Alvarado Beltrán

Fabián Enrique Longo Meneses

Asesor:

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Dedicatoria

Con gratitud y humildad, dedicamos este proyecto a Dios, fuente de toda sabiduría y guía en nuestro camino. A nuestras familias, pilares inquebrantables de apoyo y amor, les dedicamos este logro. Su constante aliento y comprensión nos impulsaron a seguir adelante, incluso en los momentos más difíciles.

A nuestros tutores y asesores, les extendemos nuestro más sincero agradecimiento. Su experiencia, orientación y paciencia fueron fundamentales en cada etapa de este proyecto. Gracias por compartir su conocimiento y por creer en nuestro potencial.

A nuestros compañeros de equipo, cuya colaboración y dedicación fueron esenciales para alcanzar este objetivo. Su compromiso y trabajo en equipo hicieron posible superar obstáculos y alcanzar resultados significativos.

A la vida misma, por brindarnos la oportunidad de aprender, crecer y contribuir al mundo que nos rodea. Cada desafío fue una oportunidad de crecimiento, y cada logro, un motivo de celebración.

Agradecimientos

A todos los colaboradores de la industria del petróleo y gas que compartieron sus experiencias y conocimientos con nosotros, les expresamos nuestro profundo agradecimiento. Sus aportes fueron invaluables para comprender las complejidades y desafíos del mantenimiento en este sector.

Además, agradecemos sinceramente la colaboración brindada, la cual ha sido fundamental para acceder a la información necesaria y llevar a cabo todos los análisis pertinentes. Esta contribución nos ha permitido obtener una visión integral de los temas relacionados con el mantenimiento en la industria del petróleo y gas.

Resumen

La industria del petróleo y gas enfrenta desafíos en la eficiencia y seguridad de sus activos, con métodos tradicionales de mantenimiento costosos y propensos a tiempos de inactividad no planificados. La digitalización y los datos en tiempo real abren oportunidades para el Mantenimiento 4.0, pero se necesitan modelos predictivos precisos para su implementación. Según IBM Analytics, el mantenimiento predictivo ofrece un retorno de inversión 10 veces mayor que los métodos tradicionales, con reducciones de costos y tiempos de parada. El objetivo del trabajo es desarrollar un modelo predictivo para la identificación de fallas en la industria del petróleo y gas, con enfoque en la Industria 4.0. Los objetivos incluyen la revisión de literatura, el análisis de datos históricos y la definición de modelos predictivos. El marco conceptual destaca la relación con la investigación en cadenas de aprendizaje industrial y define términos clave. En el marco teórico, se enfatiza el mantenimiento preventivo y la predicción de fallas mediante enfoques como modelos supervisados, no supervisados y técnicas más avanzadas como redes neuronales y aprendizaje automático.

Palabras clave: Mantenimiento 4.0, Fallas, Machine Learning, Predictivo, Aprendizaje

Abstract

The oil and gas industry faces challenges in the efficiency and safety of its assets, with traditional maintenance methods costly and prone to unplanned downtime. Digitization and real-time data open opportunities for Maintenance 4.0, but accurate predictive models are needed for implementation. According to IBM Analytics, predictive maintenance offers a return on investment 10 times higher than traditional methods, with reductions in costs and downtime. The objective of the work is to develop a predictive model for failure identification in the oil and gas industry, with a focus on Industry 4.0. The objectives include literature review, historical data analysis and predictive model definition. The conceptual framework highlights the relationship with research in industrial learning chains and defines key terms. In the theoretical framework, preventive maintenance and failure prediction is emphasized using approaches such as supervised, unsupervised models and more advanced techniques such as neural networks and Machine Learning.

Keywords: Maintenance 4.0, Failures, Machine Learning, Predictive, Learning

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Glosario

Análisis Unidimensional de Datos

Técnica estadística que se enfoca en explorar y describir una sola variable.

Coeficiente de Variación

Medida relativa de la dispersión de los datos, calculada como la desviación estándar dividida por la media y expresada en porcentaje.

Correlación

Relación estadística entre variables que ayuda a identificar patrones y relaciones en los datos.

Desviación Estándar

Medida de dispersión que indica cuánto se alejan los valores individuales de la media en un conjunto de datos.

Frecuencia de Fallas

Número de veces que se reporta una falla específica en un período determinado, útil para identificar problemas recurrentes.

Modelos Predictivos

Utilizan registros históricos y técnicas de aprendizaje automático para predecir futuras fallas en equipos.

Registros Históricos

Datos almacenados que documentan eventos pasados, como fallas, reparaciones e inspecciones en activos de la industria Oíl & Gas.

Sistema SAP PM

Sistema de gestión de mantenimiento que almacena datos detallados sobre cada aviso de mantenimiento.

TMEF (Tiempos Medios Entre Fallas)

Indica el intervalo de tiempo promedio entre fallas en los equipos analizados.

Variabilidad

Grado de dispersión o cambio en los datos, importante para comprender la consistencia de los tiempos entre fallas.

Introducción

En las páginas que siguen, exploraremos el apasionante mundo del Mantenimiento 4.0 en la industria del petróleo y gas, centrándonos en el desarrollo de un modelo predictivo basado en ciencias de datos y analítica. Conscientes de la complejidad y los desafíos inherentes a este campo, nuestro enfoque se fundamenta en la integración de tecnologías de vanguardia y una sólida revisión de la literatura especializada.

Abordaremos objetivos específicos que abarcan desde la revisión exhaustiva de la literatura hasta la definición de modelos predictivos robustos, respondiendo así a posibles objeciones y críticas mediante un enfoque riguroso y fundamentado.

Con este trabajo, aspiramos no solo a contribuir al avance del conocimiento en el campo del Mantenimiento 4.0, sino también a inspirar nuevas perspectivas y enfoques en la gestión de activos en la industria del petróleo y gas. Nuestro mensaje es claro: mediante la integración de tecnología, análisis de datos y un enfoque proactivo, es posible transformar los procesos de mantenimiento, aumentando la eficiencia operativa y reduciendo costos de manera significativa.

Planteamiento del Problema

La industria del petróleo y gas enfrenta continuamente el desafío de mantener sus activos operando de manera eficiente y segura. La planificación de mantenimiento tradicional, basada en inspecciones periódicas o en el tiempo, a menudo genera altos costos y tiempos de inactividad no planificados. Con la creciente digitalización y disponibilidad de datos en tiempo real, existe una oportunidad para aplicar el enfoque de Mantenimiento 4.0. Sin embargo, la implementación efectiva de esta estrategia requiere modelos predictivos precisos que puedan predecir fallas.

Justificación

La aplicación de la Industria 4.0 a la gestión de activos en la industria del petróleo y gas puede tener un impacto transformador. Mejorar la planificación del mantenimiento utilizando modelos predictivos puede reducir significativamente los costos operativos y mejorar la disponibilidad de los activos. También puede mejorar la seguridad al prevenir fallas no planificadas y reducir los riesgos asociados con ellas.

Según un estudio de IBM Analytics, las ventajas del mantenimiento predictivo son mucho mayores que los métodos de mantenimiento tradicionales y reportan un retorno de inversión 10 veces mayor. De hecho, según el mismo estudio, el uso del mantenimiento predictivo reduce los costos de mantenimiento entre un 20-25%, las averías entre un 70-75% y los tiempos de paradas entre un 35-45% más que otros tipos de mantenimiento (IBM, 2018).

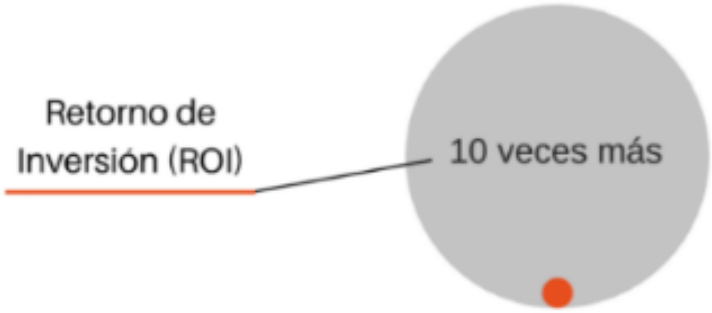
Figura 1

Reducción de costos de mantenimiento

Fuente. IBM, 2018

Figura 2

Retorno de inversión



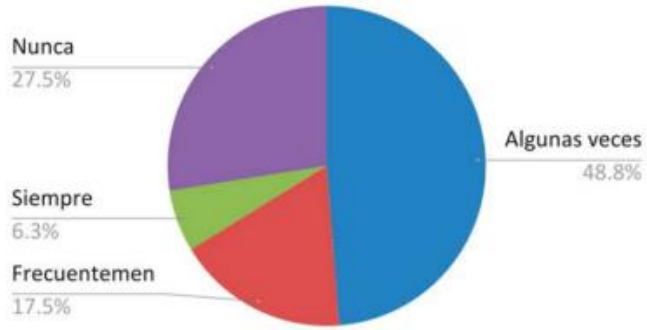
Fuente. IBM, 2018

Por otro lado, en diciembre de 2022, la Asociación Colombiana de Ingenieros de Sistemas (ACIS) realizó un estudio sobre el uso de las TICs en Machine Learning y el impacto en la industria colombiana. Se realizó una encuesta con siete preguntas clave, dejando claro que, aunque existe un interés por parte de las industrias en profundizar e implementar este desarrollo, un gran porcentaje aún no utiliza estas herramientas.

Por ejemplo, se indagó sobre el uso de herramientas de análisis de datos y/o Machine Learning para apoyo a la toma de decisiones. También se consultó si la organización cuenta con un grupo o área especializada enfocada en la generación de proyectos relacionados con data science y/o Machine Learning, generando los siguientes resultados:

Figura 3

Grado de utilización de herramientas de análisis de datos y/o Machine Learning para apoyo a la toma de decisiones



Fuente. IBM, 2018

Figura 4

Porcentaje de empresas con equipos de desarrollo de proyectos que involucran análisis de datos y/o Machine Learning.

Fuente. IBM, 2018

Figura 5

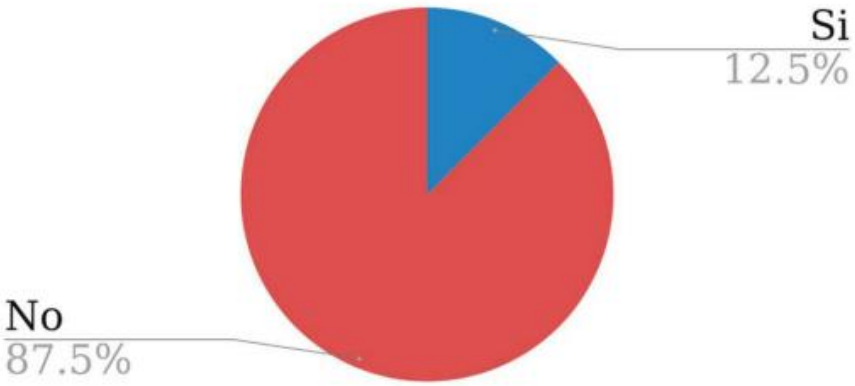
Uso de aplicaciones de Machine Learning dentro de la organización

Fuente. IBM, 2018

Un dato importante es que en las empresas de tecnología, más del 80% no usa Machine Learning en sus procesos de toma de decisiones.

Figura 6

Uso de Machine Learning dentro empresas en Colombia



Fuente. IBM, 2018

Este trabajo tiene como objetivo abordar un problema crítico en la industria del petróleo y gas mediante el desarrollo de modelos de pronóstico personalizados que aborden las complejidades específicas de la industria. Los resultados de este estudio pueden fomentar una adopción más amplia de estrategias de Mantenimiento 4.0, lo que conducirá a una industria más eficiente y sostenible.

Objetivos

Objetivo General

Desarrollar un modelo predictivo basados en ciencias de datos y analítica para la identificación de fallas en la industria Oíl & Gas, con el enfoque de la industria 4.0

Objetivos Específicos

Realizar una revisión exhaustiva de la literatura sobre Mantenimiento 4.0 y su aplicabilidad en la industria Oíl & Gas

Analizar datos históricos de los activos de la industria.

Definir modelos predictivos precisos para predecir fallos de equipos.

Marco Conceptual

La propuesta "Mantenimiento 4.0: Desarrollo de un modelo predictivo para determinar estrategias de mantenimiento en la industria del petróleo y el gas" forma parte de la línea de investigación en cadenas de aprendizaje industrial de la Escuela ECBTI de la UNAD. Esta propuesta se centra en el desarrollo de un modelo predictivo para identificar estrategias de mantenimiento en dicha industria.

El mantenimiento es crucial en cualquier sector industrial, y el mantenimiento predictivo emerge como una estrategia vital para anticipar fallas en los equipos antes de que ocurran. Este enfoque se apoya en el análisis de datos históricos para identificar patrones que puedan predecir el comportamiento futuro de los equipos.

Asimismo, la propuesta aspira a mejorar la eficiencia y la confiabilidad de los equipos en la industria del petróleo y el gas, mientras que también promueve la generación de nuevo conocimiento científico y tecnológico en el campo del mantenimiento industrial.

Marco Teórico

El mantenimiento preventivo desempeña un papel crucial en la gestión de activos y la mejora de la confiabilidad operativa en diversas industrias. En la era de la ciencia de datos y el aprendizaje automático, es fundamental aprovechar estas tecnologías para optimizar las estrategias de mantenimiento preventivo. La predicción de fallas se ha convertido en un enfoque central para esta optimización.

De acuerdo con De Simone et al. (2023), se propone un enfoque basado en redes neuronales LSTM para predecir fallas en equipos ferroviarios. Este método utiliza datos históricos de mantenimiento y rendimiento para anticipar fallas, permitiendo programar intervenciones preventivas de manera más efectiva. Lemache-Caiza et al. (2023) subrayan el papel del aprendizaje automático en la gestión del mantenimiento industrial, utilizando algoritmos para analizar grandes conjuntos de datos de sensores y registros de mantenimiento, identificando patrones útiles para la toma de decisiones preventivas.

Meddaoui et al. (2023) presentan un estudio de caso que destaca los beneficios del mantenimiento predictivo en la manufactura, mostrando métodos confiables para predecir fallas y la eficacia de la analítica de datos en la toma de decisiones de mantenimiento. Ouda et al. (2023) proponen un enfoque de optimización para el mantenimiento preventivo, utilizando el aprendizaje automático para predecir fallas a corto plazo y priorizar intervenciones según la urgencia y criticidad de los activos.

En el contexto de la automatización industrial, Pinto y Cerquitelli (2019) destacan la importancia de la detección de fallas en robots y la estimación de su vida útil restante para optimizar las estrategias de mantenimiento. Surantha y Gozali (2023) evalúan mejoras en algoritmos de aprendizaje automático para clasificar fallas en máquinas, demostrando cómo la

elección adecuada de algoritmos puede mejorar la precisión de las predicciones de mantenimiento.

Vilema Lara et al. (2022) se enfocan en técnicas de aprendizaje automático aplicadas a la clasificación binaria para mantenimiento predictivo, relacionadas con la identificación de posibles fallas en activos industriales. Villachica Pérez et al. (2022) proponen un modelo predictivo basado en Machine Learning dirigido a PYMES en el sector de ventas, ilustrando la adaptabilidad del mantenimiento predictivo a diferentes contextos industriales. Znaidi et al. (2023) se concentran en la implementación de proyectos de mantenimiento predictivo basados en minería de datos y destacan la importancia de una planificación y ejecución efectivas.

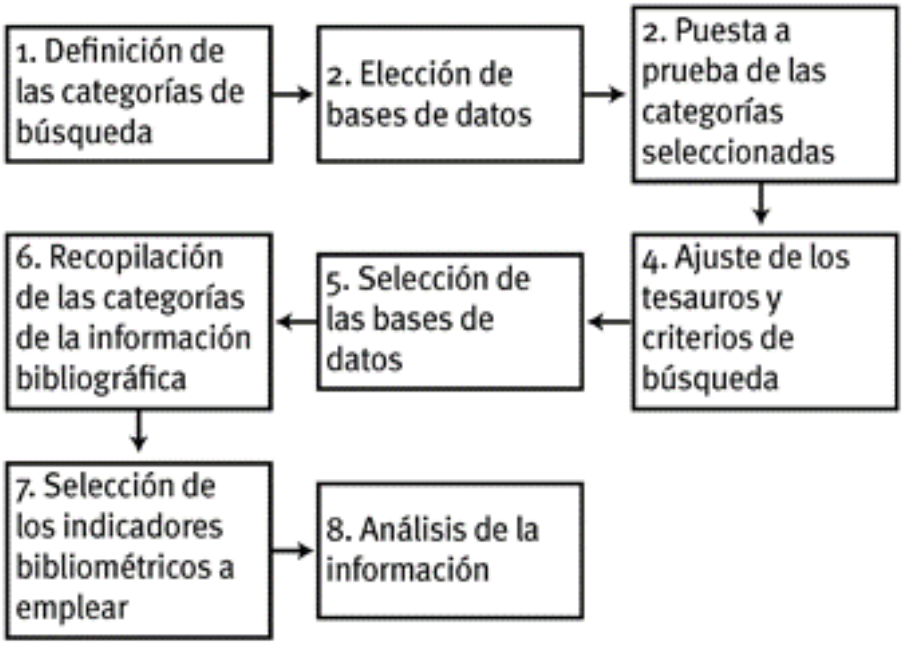
Metodología

Análisis Bibliográfico:

La metodología del análisis bibliométrico en este proyecto se fundamenta en las ocho fases definidas por Virgen et al. (2014), como se muestra en la Figura 7 "Metodología para el ejercicio bibliométrico". Según Virgen, Cobo y Betancourt (2014), estas fases están diseñadas para identificar las principales tendencias en la literatura científica.

Figura 7

Metodología para el ejercicio bibliométrico.



Fuente. Virgen et al. (2014)

Para la revisión bibliométrica, se seleccionó la base de datos de Scopus debido a su extensividad en resúmenes y citas de literatura científica revisada por pares. La estrategia de búsqueda emplea una combinación de descriptores y operadores lógicos o booleanos.

Tabla 1

Ecuación de búsqueda

Base de Datos	Algoritmo de Búsqueda
Scopus	(“Machine Learning”) AND (“fault detection” OR “fault predict*”) AND (“predictive maintenance”) AND (industry OR “industry 4.0”)

Nota. Esta tabla presenta la ecuación de búsqueda utilizada para la recopilación de datos en el estudio. *Fuente.* Elaboración propia.

La ecuación de búsqueda anterior tiene como objetivo identificar en la base de datos las publicaciones que abordan la relación entre mantenimiento predictivo y el uso de Machine Learning para la detección de fallos. El operador OR indica que al menos uno de los términos debe cumplirse, buscando los términos "detección de fallas" o "predicción de fallas" y "Industria" o "Industria 4.0". El operador AND asegura que se cumplan todos los términos simultáneamente, es decir, "Machine Learning AND fault detection AND predictive maintenance". El asterisco () en los descriptores "predict" y "detect*" permite buscar todas las combinaciones como "predictive", "prediction", "detective" o "detection", respectivamente.

En la fase 3, al poner a prueba la ecuación de búsqueda, se recopilaron un total de 2903 artículos indexados en la base de datos de Scopus. A continuación se presentan los criterios de inclusión y exclusión para el análisis en la siguiente tabla:

Tabla 2

Criterios de inclusión y exclusión

Base de Datos: Scopus Fecha Consulta Marzo 2024	
Inclusión	Exclusión
Artículos publicados entre el año 2017-2024	Artículos publicados anteriores al año 2017

Artículos de investigación	Producción que no son artículos de investigación en revistas (Congresos, conferencias libros, capítulos de libros) etc.
Temática de estudio: ingeniería, informática, Física y Astronomía, matemáticas, ciencia de los materiales, ciencias de la decisión, energía, ciencias de la decisión, Ingeniería Química, multidisciplinario, negocios, gestión y contabilidad. Economía, Econometría y Finanzas.	Ciencias Sociales, Artes y Humanidades, Profesiones de la salud, Neurociencia, Medicamento, Química, Bioquímica, Genética y Biología Molecular, Ciencia medioambiental, Ciencias de la Tierra y Planetarias, Ciencias Agrícolas y Biológicas.

Nota. Esta tabla detalla los criterios utilizados para la inclusión y exclusión de estudios en el análisis. *Fuente.* Elaboración propia.

Después de aplicar los criterios de inclusión, se obtuvieron 1532 artículos. Luego, se procedió con el ajuste de tesauros, la limpieza y la selección de la base de datos. Utilizando la función de tesauro en VosViewer, se eliminaron duplicidades. Esto se refiere a conceptos idénticos pero expresados de manera diferente por distintos autores; por ejemplo, el modelo KNN es equivalente a K-Nearest Neighbor. Este proceso asegura que cada concepto se consolide adecuadamente para evitar la dispersión de ocurrencias y fortalecer la precisión del estudio.

Diagnóstico de Registros Históricos

Recolección de Información

Para iniciar el ejercicio, es crucial clasificar las unidades funcionales y los equipos que serán objeto de estudio. La recolección de datos se lleva a cabo utilizando el CMMS SAP y la información registrada allí, que incluye taxonomía, planes de mantenimiento, posiciones de mantenimiento, hojas de ruta y listas de objetos.

Análisis Exploratorio de los Datos

Se utiliza Jupyter Notebook junto con Python para llevar a cabo el análisis exploratorio de datos. Esta combinación de herramientas facilita la extracción de información clave que ayuda a comprender la distribución de los datos, así como a identificar patrones y tendencias. Para este propósito, se emplean diversas bibliotecas especializadas y se siguen metodologías estructuradas para analizar de manera eficiente los registros de incidencias en el funcionamiento de los equipos.

Librerías Utilizadas:

- pandas: Para la manipulación y análisis de datos tabulares.
- numpy: Para operaciones numéricas en arrays y matrices.
- matplotlib.pyplot: Para la visualización de datos mediante gráficos.
- seaborn: Para la creación de gráficos estadísticos atractivos.
- sklearn.preprocessing.StandardScaler: Para estandarizar características numéricas.
- sklearn.decomposition.PCA: Para realizar el análisis de componentes principales.
- datetime: Para trabajar con fechas y horas.

Carga de Datos:

Se cargan dos archivos Excel que contienen información relevante: uno con datos de fallas y otro con datos de equipos.

Selección de Columnas:

Se seleccionan las columnas pertinentes del DataFrame de fallas, incluyendo información como duración de parada, prioridad, equipo afectado, fechas de inicio y fin de avería, entre otros.

Análisis de Datos:

Se realizan análisis descriptivos como la media, mediana, moda y otras medidas de posición sobre las columnas numéricas del conjunto de datos.

Se exploran variables como la duración de parada, prioridad, repercusión, días transcurridos, entre otras, para comprender su distribución y características.

Análisis de Componentes Principales (PCA):

Se aplica el análisis de componentes principales para reducir la dimensionalidad de los datos y encontrar patrones en las variables.

Se calculan los componentes principales y se analiza su contribución a la variabilidad de los datos.

Visualización de Datos:

Se crea un gráfico de barras para mostrar la frecuencia de eventos por semanas, utilizando la librería seaborn.

Se realiza un análisis de la distribución de variables como prioridad, duración de parada y repercusión mediante gráficos y medidas de asimetría (skewness).

La metodología se enfoca en cargar, limpiar, analizar y visualizar los datos de registro de paradas de equipos, aplicando técnicas de análisis descriptivo, análisis de componentes principales y visualización de datos para obtener información valiosa sobre el comportamiento de los equipos y las fallas registradas.

Cálculo del TMEF

De acuerdo con la información recolectada, una vez realizada la clasificación de premisas, se realiza el cálculo por componente para determinar el TMEF asociado, de acuerdo con el resultado se debe obtener el TMEF de la unidad Funcional y cada una de las especialidades (Mecánico, Eléctrico, Instrumentación, Control, etc.), ya que cada uno de estos estará relacionado con una posición de mantenimiento.

$$\text{TMEF} = \frac{\sum \text{Tiempo entre Falla}}{\text{Numero de Fallas}}$$

Figura 8

Cálculo TMEF por Unidad Funcional

Fuente. Elaboración propia

Diseño de Modelos Predictivos

Metodológicamente, el análisis sobre modelos predictivos en la industria Oíl & Gas sigue un enfoque estructurado que incluye los siguientes pasos:

Definición del Problema:

Se establece el objetivo del estudio, que en este caso es evaluar y comparar varios modelos predictivos para predecir fallas en equipos y maquinaria en la industria Oíl & Gas. El problema se enmarca en la necesidad de mejorar la eficiencia y la rentabilidad a través de la gestión proactiva de activos.

Importación de Bibliotecas:

Se utilizan diversas bibliotecas de Python, como pandas, numpy, sklearn, matplotlib, entre otras. Estas bibliotecas son fundamentales para la manipulación de datos, la implementación de modelos de Machine Learning y la visualización de resultados.

Recopilación de Datos:

Se recopilan datos históricos detallados, como registros de mantenimiento, tiempos entre fallas y condiciones operativas de los equipos y maquinaria en la industria Oíl & Gas. Estos datos son fundamentales para entrenar y validar los modelos predictivos.

Exploración de Datos:

Se realiza una exploración inicial de los datos utilizando el método info() del DataFrame. Esta exploración proporciona información sobre la estructura de los datos, como el tipo de cada columna y la presencia de valores nulos.

Preprocesamiento de Datos:

Se lleva a cabo un preprocesamiento de los datos, que incluye la selección de características relevantes para el análisis y la codificación de variables categóricas utilizando OneHotEncoder. Además, se escalan las características numéricas utilizando StandardScaler, adicionalmente se dividen los datos en conjuntos de entrenamiento y prueba para poder evaluar el rendimiento de los modelos. Esto se realiza utilizando la función train_test_split de sklearn.

Selección de Modelos:

Se eligen varios modelos predictivos para su evaluación, incluyendo Regresión Lineal, Random Forest, Gradient Boosting, SVR, Análisis de Agrupamiento K-Means y Análisis de Componentes Principales (PCA). Cada modelo se selecciona en función de su aplicabilidad al problema y su capacidad para manejar diferentes tipos de datos y relaciones.

Entrenamiento y Validación de Modelos:

Se entrenan los modelos seleccionados utilizando los datos históricos recopilados. Luego, se validan los modelos utilizando técnicas como validación cruzada y división de datos en conjuntos de entrenamiento y prueba para evaluar su rendimiento y precisión en la predicción de fallas.

Análisis de Resultados:

Se analizan los resultados obtenidos para cada modelo, prestando atención a métricas de evaluación como el coeficiente de determinación (R^2), el Error Cuadrático Medio (MSE) y la

capacidad de capturar patrones en los datos. Se comparan los resultados entre los diferentes modelos para identificar fortalezas y debilidades.

Para ver en mayor detalle lo realizado puede remitirse al anexo modelos_analisis.ipynb.

Revisión Exhaustiva de la Literatura.

Este capítulo realiza un análisis de la producción científica relacionada con el problema de investigación, que aborda el desarrollo de modelos de Machine Learning en el contexto del mantenimiento predictivo para la detección de fallas de equipos en la Industria Oil & Gas 4.0.

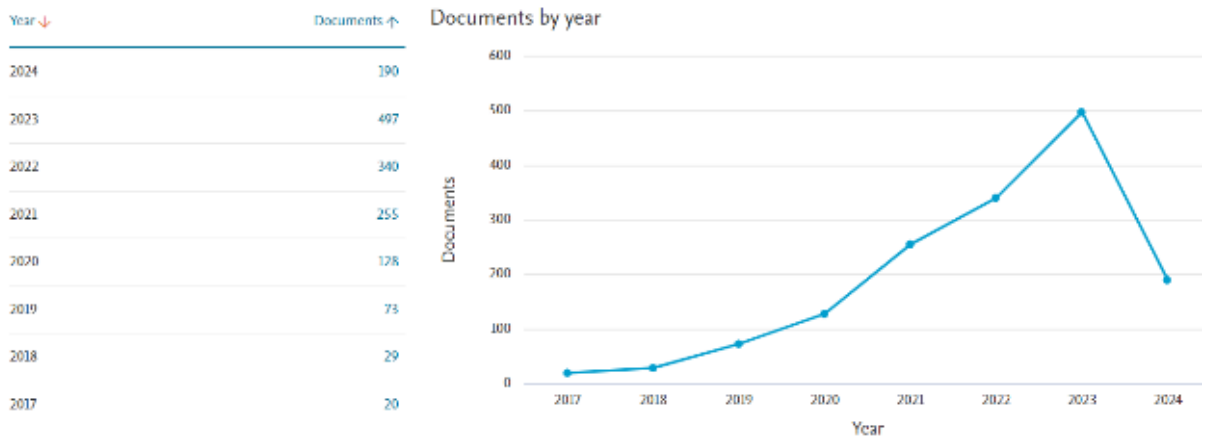
Análisis de la Información:

El estudio bibliométrico se centrará en dos aspectos fundamentales: citación en la unidad de análisis (autores, documentos, fuentes y países) y co-ocurrencias (palabras clave). Para llevar a cabo este análisis, utilizaremos el software VOSviewer.

Evolución de la Producción Científica por Año.

Figura 9

Documentos por año



Fuente. Elaboración propia

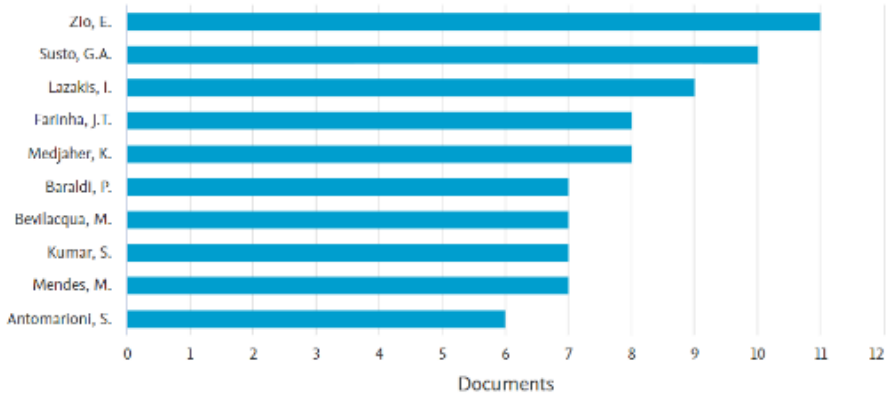
La producción intelectual comenzó a crecer de manera exponencial a partir del año 2017, con aproximadamente 20 artículos. En 2019 se produjeron 73 artículos, en 2020 se produjeron 128, en 2021 se generaron 255, y en 2022 la cifra ascendió a 340. El año pasado, se generaron un total de 500 artículos. Durante el primer trimestre del 2024, ya se ha superado el número total de documentos producidos en 2020. Estos datos evidencian un claro incremento en la investigación

sobre la relación entre el Machine Learning y la gestión del mantenimiento para la detección y predicción de fallos en la industria. El Machine Learning se posiciona como una herramienta útil en el mantenimiento predictivo para la detección de fallas.

Documentos por Autor.

Figura 10

Documentos por autor



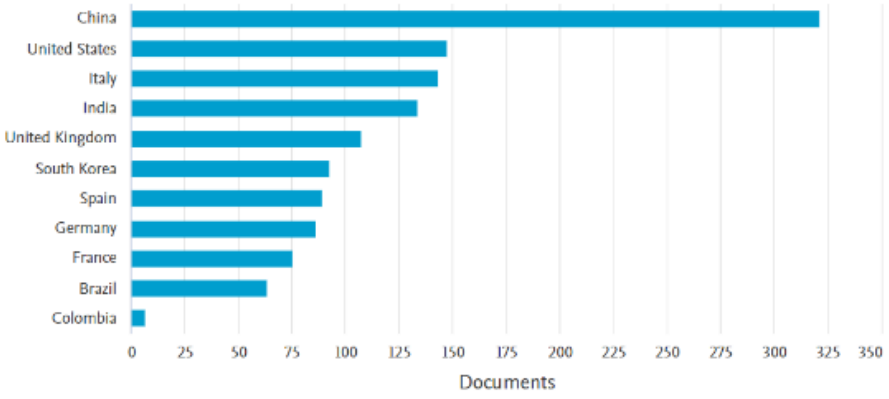
Fuente. Elaboración propia

La mayoría de los documentos son de autores como Enrico Zio, G.A. Susto, Lazakis, J.T. Farinha, entre otros. Sin embargo, es necesario verificar cuánto han sido citados por otros autores y si han colaborado entre sí. Este análisis se realizará más adelante.

Ubicación Geográfica de los Documentos.

Figura 11

Documentos por país o territorio



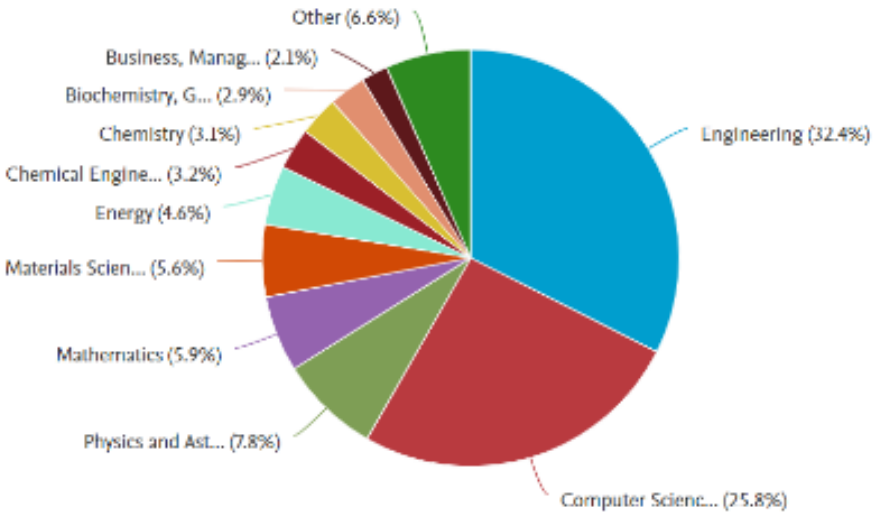
Fuente. Elaboración propia

El país con mayor número de publicaciones en este tema es China, con 321 documentos. En segundo lugar está Estados Unidos, con 147 artículos. Brasil es el mejor país latinoamericano clasificado, con 63 artículos, mientras que Colombia ocupa el puesto 49 con 6 artículos.

Artículos por Área de Conocimiento.

Figura 12

Documentos por área



Fuente. Elaboración propia

Con los criterios de inclusión en las temáticas seleccionadas, las áreas donde se desarrolla esta producción intelectual son principalmente ingeniería e informática, que en conjunto representan un 58%.

Análisis de Citación:

Como se mencionó anteriormente, el análisis de citación se llevará a cabo considerando la citación de autores, citación de documentos, citación de fuentes y citación por país.

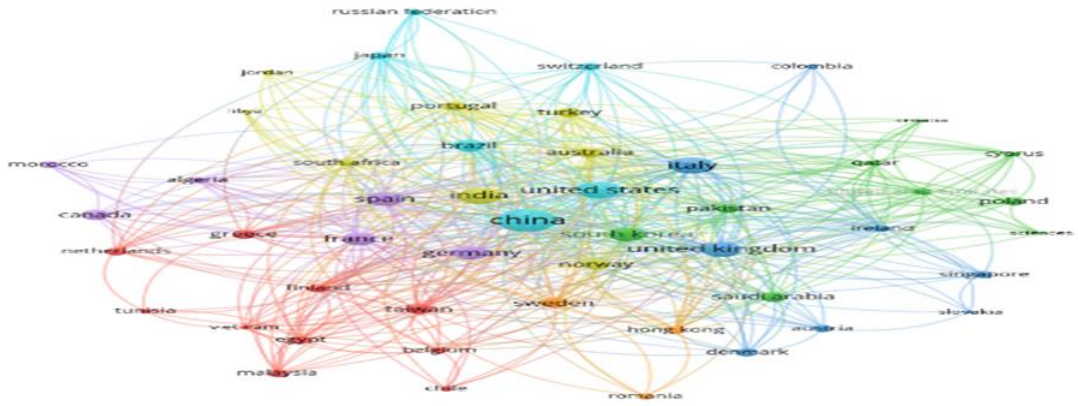
Análisis de citación por país.

Se establece el análisis de red (Network Visualization). Se seleccionaron los 50 países principales con mayor actividad en el tema, lo que nos permite no solo identificar el número de citaciones recibidas por cada país, sino también agruparlos según la colaboración entre los autores.

El tamaño del nodo en la red indica la cantidad de citaciones recibidas por país. Podemos observar que China y Estados Unidos son los países más citados, además de compartir el mismo color "Cian", lo que sugiere que están agrupados en un mismo clúster y frecuentemente colaboran entre sí.

Figura 13

Análisis de citación por país



Fuente. Elaboración propia

En la siguiente tabla se observan con mayor detalle los grupos de países que suelen colaborar conjuntamente. Esto no excluye la posibilidad de colaboraciones entre clústeres, las cuales se reflejan en las conexiones entre los nodos en la red.

Tabla 3

Agrupación de países relacionado con la colaboración y número de citas

Clúster	Países
Cian	China, EE. UU., Brasil, Japón, Suiza, Rusia.
Azul	Italia, Reino Unido, Singapur, Austria, Dinamarca, Eslovaquia.
Verde	Corea del Sur, Pakistán, Arabia Saudita, Polonia, Qatar.
Amarillo	India, Australia, South África, Portugal, Turquía
Violeta	Alemania. España, Francia, Canadá, Marruecos, Algeria.
Rojo	Finlandia, Grecia, Taiwán, Bélgica, Holanda, Egipto.

Nota. Esta tabla muestra la agrupación de países según la colaboración y el número de citas obtenidas en el estudio. *Fuente.* Elaboración propia.

Análisis de Citación por Documentos.

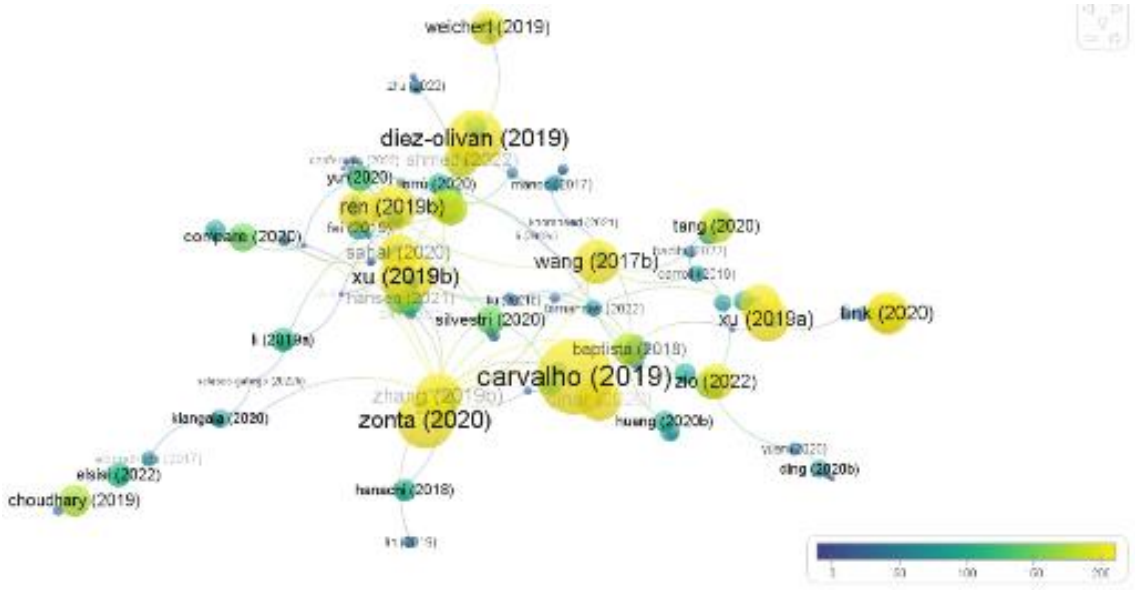
De los 1532 artículos, se seleccionaron aquellos con un mínimo de 10 citaciones, lo que redujo la muestra a 535 documentos. El gráfico llamado "Overall Visualization" es muy interesante porque permite identificar los documentos más citados cronológicamente. Esto es crucial si deseamos consultar las publicaciones más recientes sobre el tema.

Lo Más Citado:

De lo más citado, lo más reciente:

Figura 14

Análisis de citación por Documento. Puntuación por núm. citaciones



Fuente. Elaboración propia

La puntuación por número de citaciones nos permite identificar que los documentos marcados con color amarillo son los más citados, y el año se visualiza en la etiqueta correspondiente. A continuación, se presentan las tablas con los 5 documentos más citados y los más recientes.

Tabla 4

Documentos más citados y recientes

Autor	Título	Fuente	Año
Zio, Enrico	Predictive Maintenance in the Industry 4.0: A Systematic Literature Review	Computers and Industrial Engineering and System Safety	2022

Zonta, Thiago; Da Costa, Cristiano André; Da Rosa Righi, Rodrigo; De Lima, Miromar José	Prognostics and Health Management (PHM): Where Are We and Where Do We (Need to) Go in Theory and Practice	Computers and Industrial Engineering	2020
Thang; Shegnan; Yuan; Shouqi; Zhu, Yong	Deep Learning-Based Intelligent Fault Diagnosis Methods Toward Rotating Machinery	IEEE Access	2020
Carvalo, Thyago P; Soares, Fabrizzio; Vita, Roberto; Francisco Roberto da P; Basto Joap P; Alcalá, Symone	A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance	Computers and Industrial Engineering	2019

Nota. Esta tabla presenta los documentos más citados y los más recientes encontrados en la revisión bibliográfica. *Fuente.* Elaboración propia.

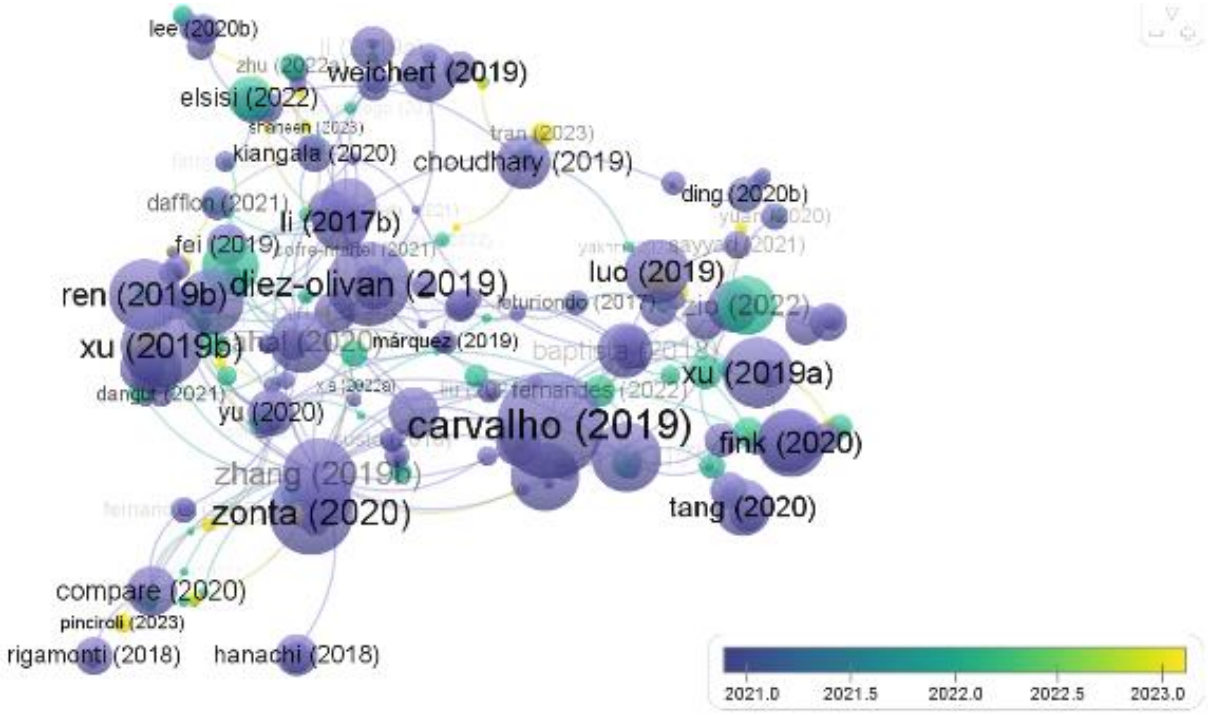
Lo Más Reciente:

En la Figura 15 se representa una visualización de la puntuación por año, filtrando desde el año 2021 hasta la fecha actual. Los documentos marcados con colores violeta, cian y amarillo corresponden a los años 2021, 2022 y 2023, respectivamente. Se observa que el número de citaciones tiende a disminuir a medida que los documentos son más recientes. Esto se debe a que los documentos necesitan más tiempo para acumular citaciones. Sin embargo, no es una regla absoluta, ya que existen documentos relativamente recientes con un número considerable de

citaciones, destacados por nodos verdes y amarillos más grandes, lo que evidencia su impacto a corto plazo.

Figura 15

Análisis de citación por Documento. Puntuación por año



Fuente. Elaboración propia

A continuación, se observan referencias dentro de las más recientes, las más citadas:

Tabla 5

Referencias más citadas

Autor	Título	Fuente	Año
Pinciroli, Luca; Baraldi, Piero; Zio, Enrico	Maintenance Optimization in Industry 4.0	Reliability Engineering and System Safety	2023

Gawde, Shreyas; Patil, Shruti; Kumar, Satish; Kotecha, Ketan	A Scoping Review on Multi-Fault Diagnosis of Industrial Rotating Machines Using Multi-Sensor Data Fusion	Artificial Intelligence Review, Computers and Industrial Engineering	2023
Tran, Minh-Quang; Amer, Mohamed; Dababat, Alya; Abdelaziz, Almoataz Y; Dai, Hong-Jie; Liu, Meng	Robust Fault Recognition and Correction Scheme for Induction Motors Using an Effective IoT with Deep Learning Approach	Measurement Journal of the International Measurement Confederation	2023
Carvalo, Thyago P; Soares, Fabrizzio; Vita, Roberto; Francisco Roberto da P; Basto Joap P; Alcalá, Symone	A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance	Computers and Industrial Engineering	2019

Nota. Esta tabla muestra las referencias más citadas en el estudio realizado. *Fuente.* Elaboración propia.

Análisis de Citación por Fuentes

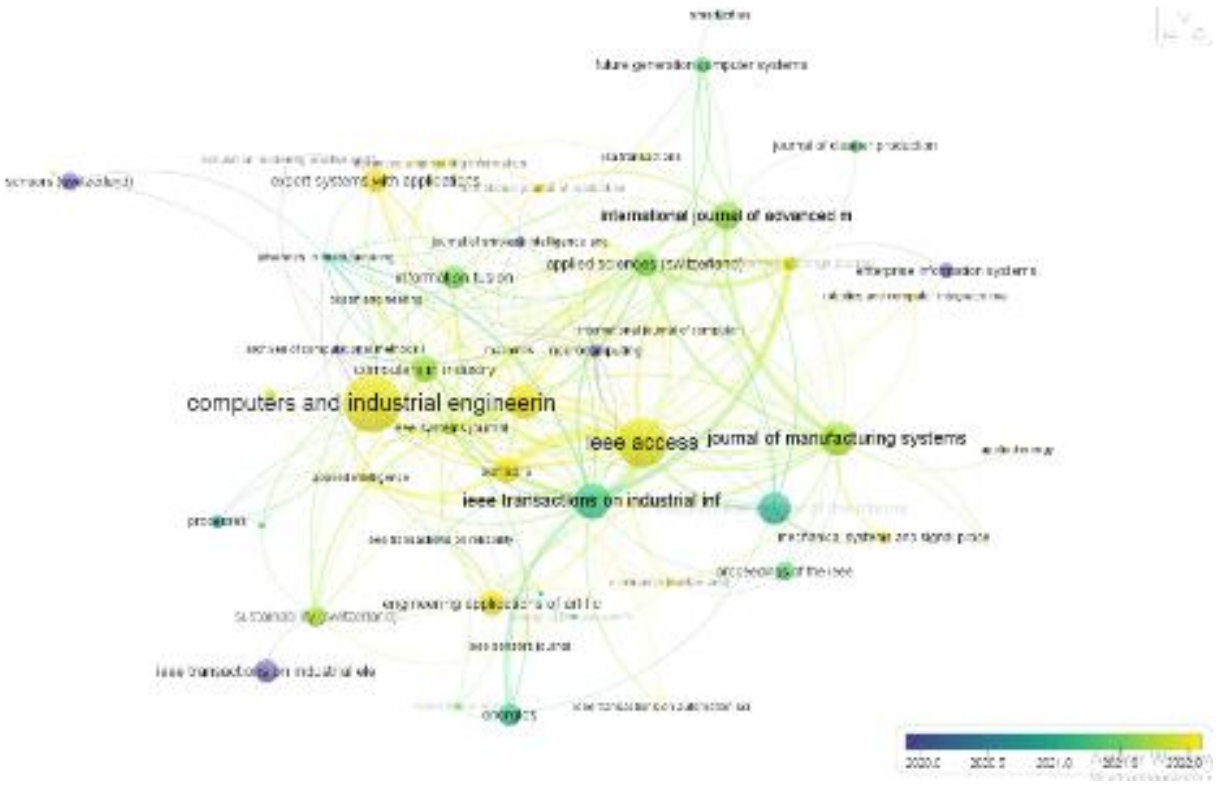
En la fase inicial, seleccionamos únicamente artículos publicados en revistas científicas, por lo que nuestra fuente de datos son exclusivamente revistas científicas. La consulta se realizó a revistas que tuvieran al menos 100 citaciones y al menos un documento publicado. Por lo tanto, la relación clave aquí es entre citaciones y documentos. Una revista se destaca tanto por la cantidad de citaciones que recibe como por la cantidad de documentos que ha publicado.

Como resultado, hemos identificado un total de 56 revistas, con citaciones que van desde 100 hasta 1753, y con documentos publicados que varían entre 1 y 89. La Figura 16 permite visualizar dinámicamente esta relación. La escala normalizada de puntuación muestra las citaciones de cada revista en miles: el color amarillo representa las revistas con más citaciones y

el azul oscuro las que tienen menos citaciones. El tamaño del nodo en el gráfico representa la cantidad de documentos publicados por cada revista.

Figura 16

Análisis de citación por Fuentes (revistas)



Fuente. Elaboración propia

"Computers and Industrial Engineering": Esta revista internacional cuenta con un total de 1753 citas y se especializa en la publicación de contribuciones que presentan nuevas metodologías computarizadas para la resolución de problemas en ingeniería industrial. La revista fomenta artículos que evalúan aplicaciones informáticas en diversos temas de ingeniería industrial y relacionados, así como investigaciones sobre el uso de computadoras en la educación en ingeniería industrial (C&IE, 2024).

"IEEE Access": Con 1472 citaciones, es una revista científica de acceso abierto revisada por pares publicada por el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE). La revista ganó el premio PROSE en 2015 a la mejor revista nueva en ciencia, tecnología, ingeniería y matemáticas (Prose Awards, 2018).

"Journal of Manufacturing Systems": Es una revista académica publicada por Elsevier que abarca áreas como Informática y mecanizado. Sus artículos se enfocan principalmente en temas relacionados con Ciencias de la Computación, Mecanizado, Programación de procesos de producción, Programación de trabajos en taller y Fabricación integrada por computadora.

Tabla 6

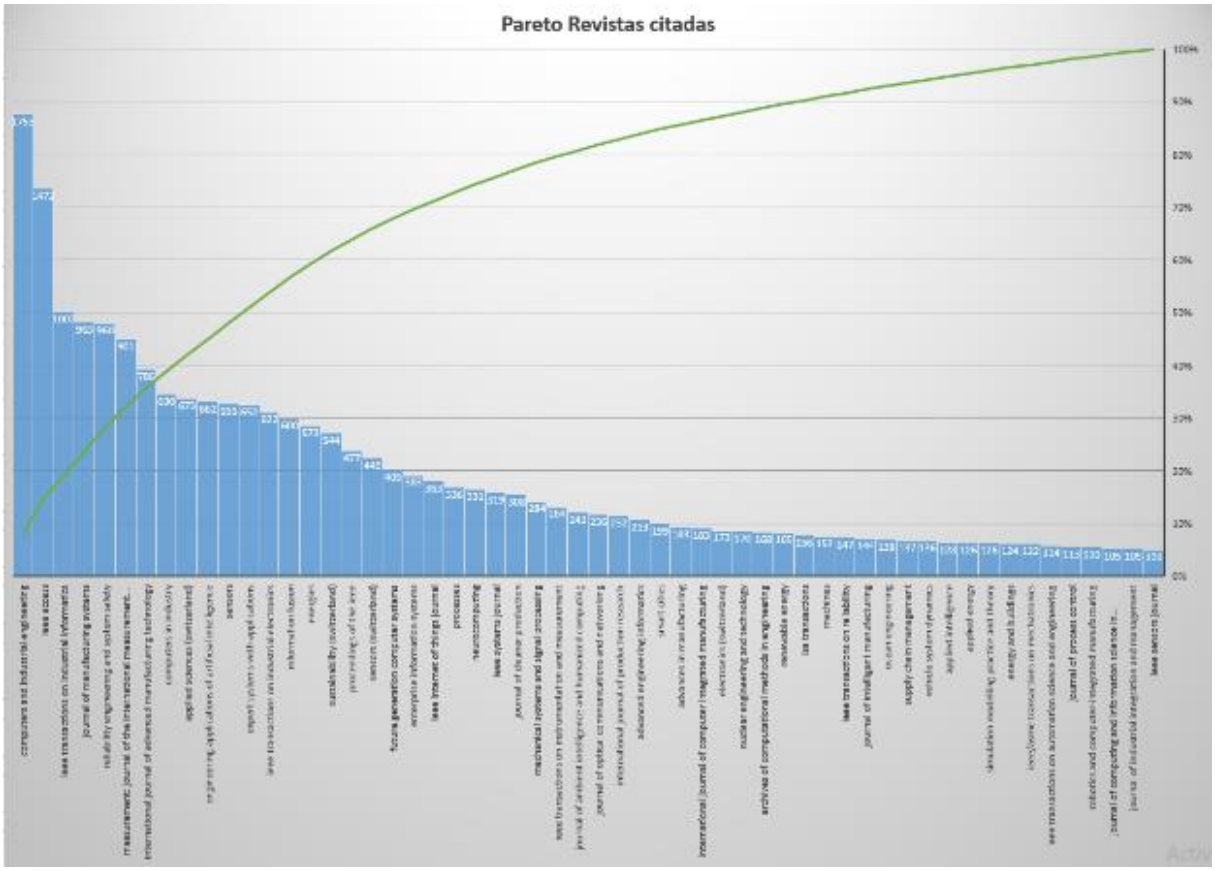
Top revistas con mayores citaciones

N°	Fuente	Citaciones	Documentos
1	computers and industrial engineering	1753	27
2	ieee access	1472	77
3	ieee transactions on industrial informatics	1002	24
4	journal of manufacturing systems	963	24
5	reliability engineering and system safety	960	30
6	measurement: journal of the international measurement confederation	901	21
7	international journal of advanced manufacturing technology	786	31
8	computers in industry	690	28

Nota. Esta tabla presenta las principales revistas según el número de citaciones obtenidas en el estudio. *Fuente.* Elaboración propia.

Figura 17

Diagrama de Pareto de Principales fuentes citadas



Fuente. Elaboración propia

Análisis de Citación por Autores

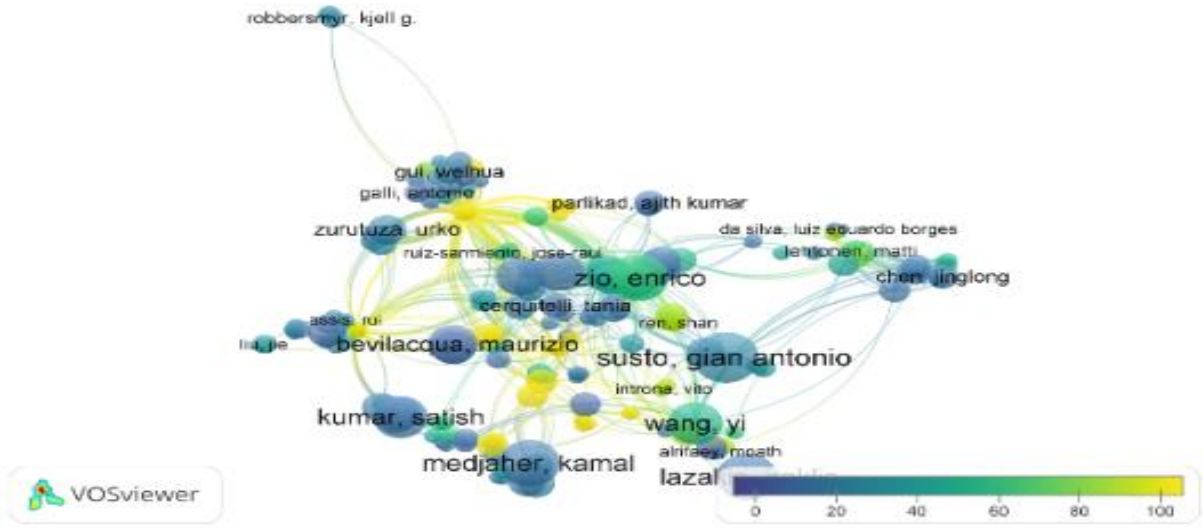
Para el análisis de citación por autor, se observa que de los 5753 autores identificados en los 1532 documentos, 1220 de ellos no tienen citaciones. El autor con el mayor número de citaciones tiene 689 y el máximo de documentos publicados por un autor es de 10. Se seleccionan autores que tienen al menos 2 publicaciones y un mínimo de 25 citaciones, lo que resulta en un total de 286 autores. En la Figura siguiente, la escala de color indica el porcentaje de citaciones en comparación con el máximo de 689 citaciones. El tamaño del nodo representa la cantidad de documentos publicados por cada autor. Esto significa que hay autores altamente

citados con pocas publicaciones y viceversa. Por ejemplo, los autores destacados en color amarillo tienen citaciones cercanas al número máximo, y se identifican también aquellos con el mayor número de publicaciones.

Autores como Enrico Zio obtienen una proporción destacada entre el número de documentos publicados y el número de citaciones recibidas. Por ejemplo, en el análisis inicial de autores, otros nombres prominentes incluyen a Jinjiang Wang, conocido por sus numerosas citaciones pero con pocas colaboraciones. También se destacan Rodrigo Da Rosa Righi y Diego Galagar, entre otros.

Figura 18

Diagrama de red por citación de autores



Fuente. Elaboración propia

Análisis de Co-Palabras

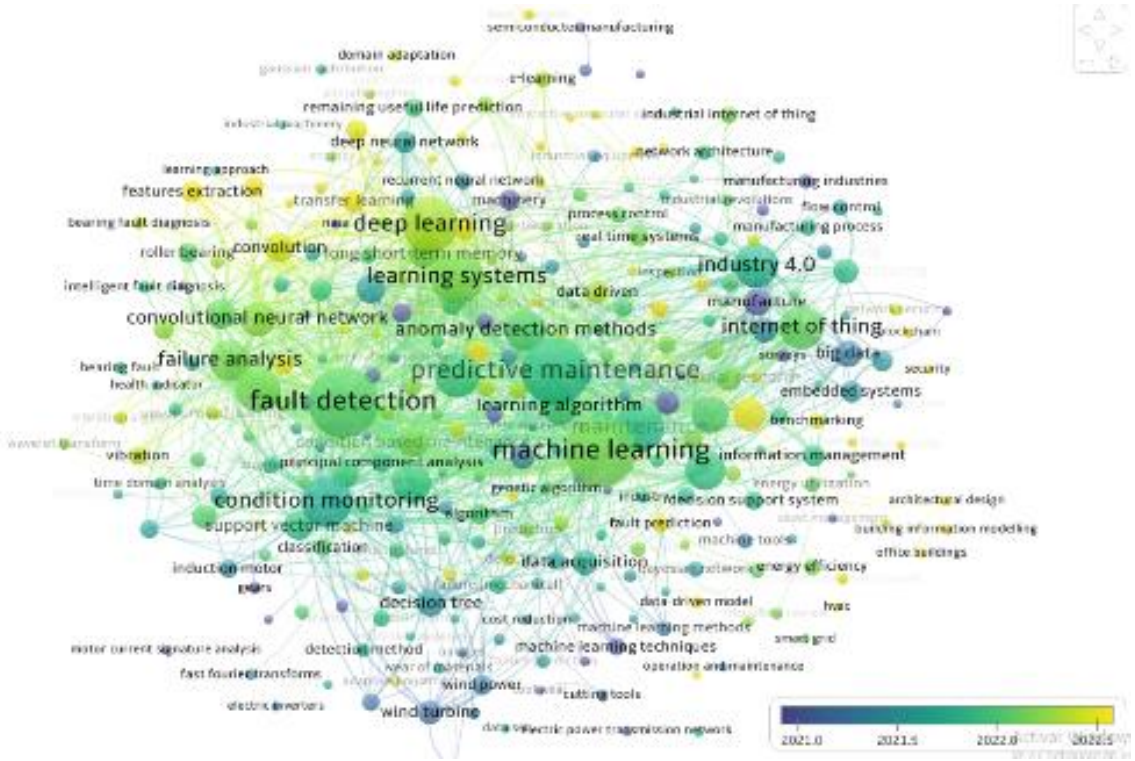
Para este análisis, se identificaron un total de 9660 palabras con al menos una ocurrencia. Se seleccionaron palabras que tienen al menos 10 ocurrencias, lo que resultó en un total de 331 palabras que cumplen con este criterio. Después de aplicar el Tesauro, quedaron un total de 9541 palabras, y el umbral final se estableció en 290 palabras con al menos 10 ocurrencias. En el

diagrama de visualización general, se pueden identificar estas palabras, donde el tamaño del nodo representa la cantidad de ocurrencias de cada palabra y la gama de colores indica qué tan antiguos o recientes son los conceptos citados.

Overal Visualization:

Figura 19

Diagrama de red ocurrencia de palabra. Tendencias



Fuente. Elaboración propia

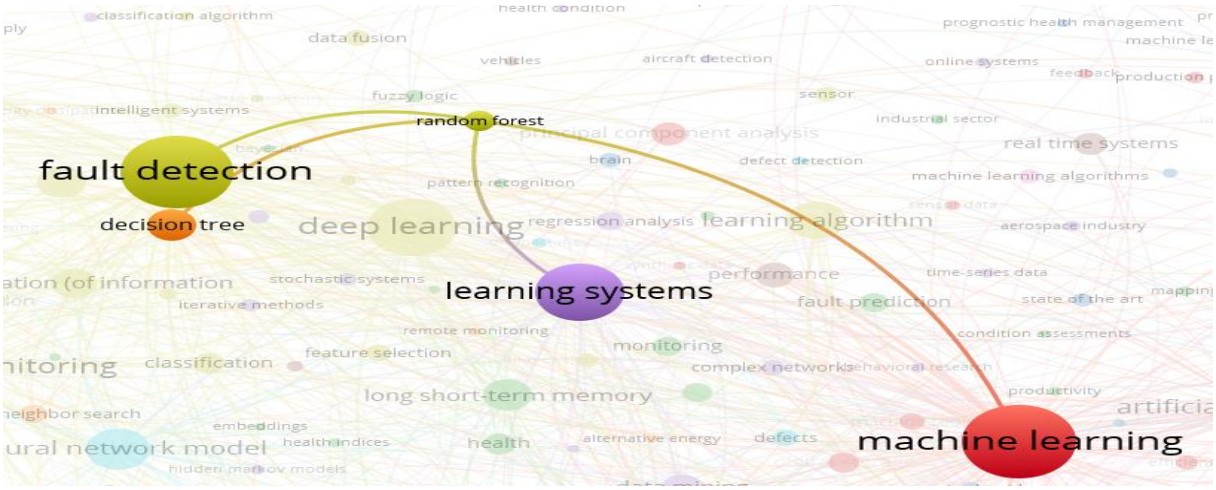
En esta visualización de redes, se observan palabras clave con mayor ocurrencia, representadas por nodos de mayor tamaño, como por ejemplo: Machine Learning, deep learning, predictive maintenance, fault diagnosis, fault prediction, fault detection, Industry 4.0, data analysis, y artificial intelligence. También se pueden identificar las relaciones entre cada nodo y otras palabras, lo que permite establecer un seguimiento de términos relevantes para la investigación.

Si nos enfocamos en el nodo de "Predictive maintenance", una de nuestras principales palabras clave en el algoritmo de búsqueda, podemos identificar todas las palabras relacionadas en mayor o menor medida. De esta manera, podemos explorar los modelos de aprendizaje automático más utilizados en este campo, o al menos, dentro de este conjunto de artículos sobre mantenimiento predictivo.

Random Forest

Figura 20

Randon Forest



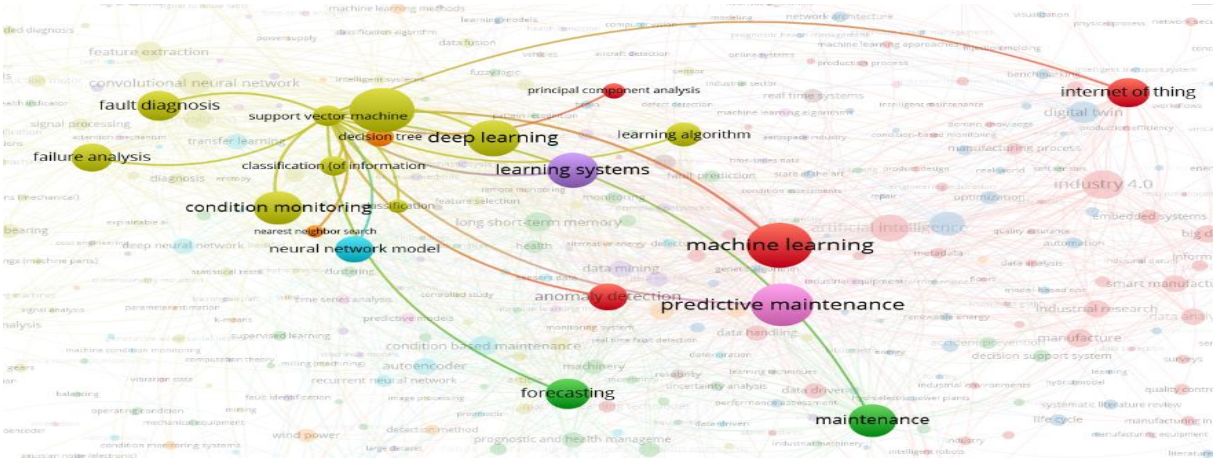
Fuente. Elaboración propia

Random Forest se relaciona ampliamente con Machine Learning, detección de fallas y sistemas de aprendizaje y evidentemente con árboles de decisiones.

Support Vector Machine

Figura 21

Support Vectore Machine



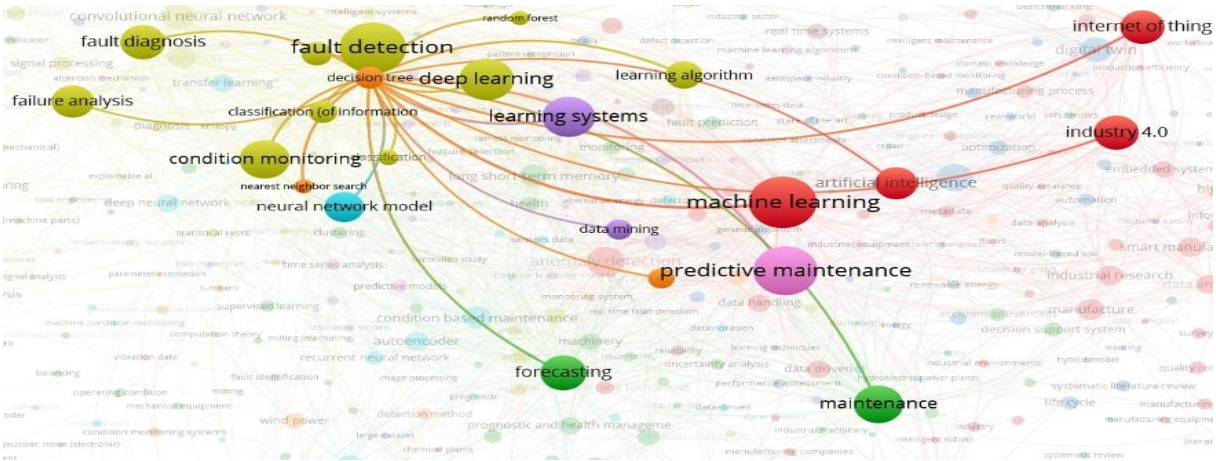
Fuente. Elaboración propia

Máquinas de soporte Vectorial se relaciona con varias palabras clave como pór ejemplo métodos de detección de anomalías, detección de fallos, análisis de fallos, deep learning, nearest neighbord search.

Decision Tree

Figura 22

Decision Tree



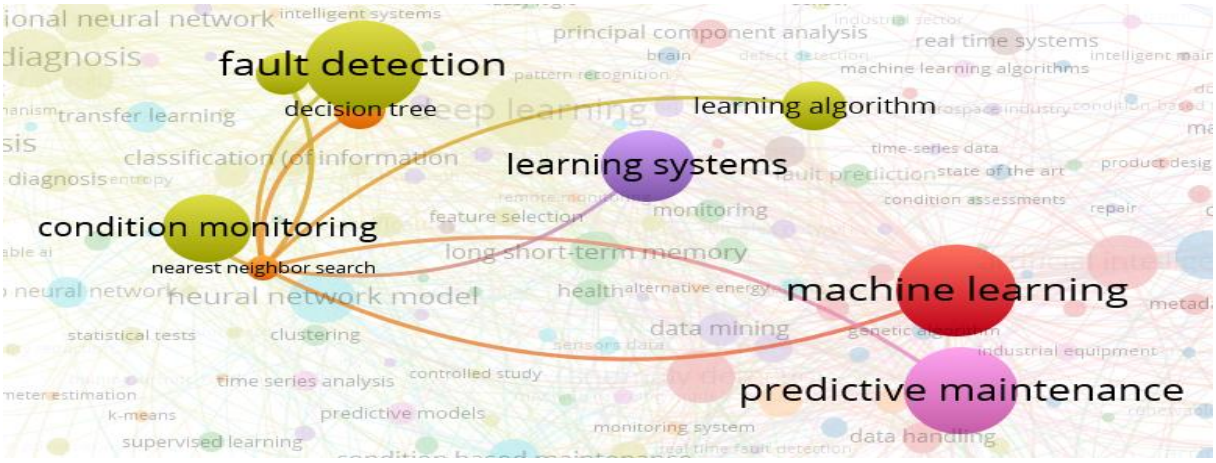
Fuente. Elaboración propia

Arboles de decisión se relaciona con varias palabras clave muy similar a MSV como por ejemplo detección de fallos, análisis de fallos, sistemas de aprendizaje, condiciones de monitoreo, Deep learning. Este último se identifica como una palabra más reciente que las demás.

Nearest Neighbord Search

Figura 23

Nearest Neighbord Search



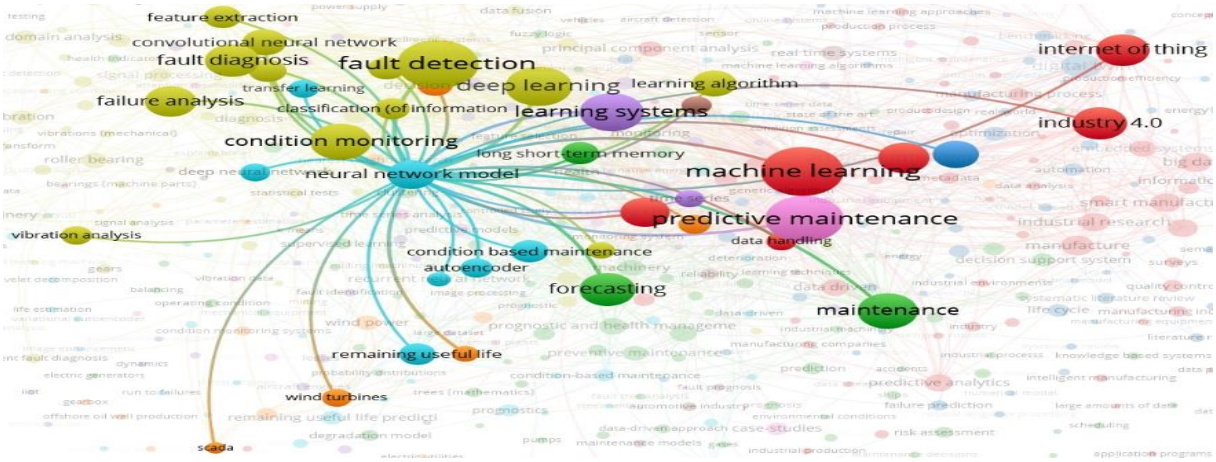
Fuente. Elaboración propia

Nearest Neighbord search tiene relación al igual que las anteriores con mantenimiento predictivo, Machine Learning y detección de fallas. Adicionalmente se relaciona con algoritmos de aprendizaje y como vimos anteriormente con MSV

Neuronal Network Model

Figura 24

Neuronal Network Model



Fuente. Elaboración propia

El modelo de redes neuronales, como podemos observar tiene una amplia gama de relaciones, mayor cantidad de términos en común, adicionalmente se identifican dos términos muy recientes como son convolución y Aprendizaje por transferencia.

Discusión

El análisis bibliográfico realizado proporciona una visión detallada del estado actual de la investigación en el tema, destacando el impacto del Machine Learning en la gestión del mantenimiento de equipos mediante estrategias como el mantenimiento predictivo para la detección y predicción de fallas. Además, se identifican las palabras clave más utilizadas, los países con mayor producción, las revistas prominentes y los autores más citados. Se observa un notable aumento en la producción de artículos en los últimos siete años, impulsado por la creciente disponibilidad de datos facilitada por el desarrollo tecnológico de sistemas de información y la cuarta revolución industrial.

Análisis de Registros Históricos

En el contexto de la industria del petróleo y el gas, el análisis de registros históricos juega un papel fundamental en la gestión eficiente de activos y en la implementación de estrategias de mantenimiento predictivo. A través de la recopilación y análisis de datos pasados, es posible identificar tendencias, patrones y áreas de mejora que permitan optimizar la operatividad de los equipos y prevenir fallas inesperadas.

¿Qué son los Registros Históricos?

Los registros históricos son datos almacenados que documentan eventos pasados. En el caso de los avisos de mantenimiento, estos registros contienen información sobre fallas, reparaciones, inspecciones y otras actividades relacionadas con los activos de la industria Oíl & Gas a lo largo del tiempo.

Fuentes de Registros Históricos:

Sistema SAP PM: El sistema de gestión de mantenimiento (SAP PM) almacena datos detallados sobre cada aviso de mantenimiento. Estos registros incluyen fechas, descripciones de fallas, acciones tomadas, tiempos de respuesta y otros atributos relevantes.

Análisis de Tendencias y Patrones:

Frecuencia de fallas: Examinar cuántas veces se ha reportado una falla específica en un período determinado. Esto ayuda a identificar equipos o componentes propensos a problemas recurrentes.

Modelos Predictivos Basados en Registros Históricos:

Utilizando técnicas de aprendizaje automático, se pueden desarrollar modelos predictivos que utilicen los registros históricos para predecir futuras fallas. Estos modelos

pueden considerar múltiples variables, como el tiempo transcurrido desde la última inspección, la edad del equipo y las condiciones operativas.

Los modelos pueden clasificar los avisos de mantenimiento en categorías (por ejemplo, alta, media o baja probabilidad de falla) y ayudar a priorizar las estrategias de mantenimiento.

Instrumentos de Recolección de Datos Relacionados con Registros Históricos:

Herramientas de visualización de datos: Utilizar gráficos, tablas y visualizaciones para explorar patrones y tendencias en los registros históricos.

Software de minería de datos: Aplicar técnicas de minería de datos para descubrir relaciones ocultas entre variables y predecir comportamientos futuros.

En resumen, el análisis de registros históricos es fundamental para comprender las tendencias, identificar áreas de mejora y desarrollar modelos predictivos precisos en el contexto del Mantenimiento 4.0 en la industria Oíl & Gas .

Análisis Unidimensional de los datos

El análisis unidimensional de datos es una técnica fundamental en estadística que se enfoca en explorar y describir una sola variable. En el contexto de la problemática inicial sobre el mantenimiento en la industria Oíl & Gas, realizaremos un análisis unidimensional para comprender mejor los datos relacionados con los avisos de mantenimiento.

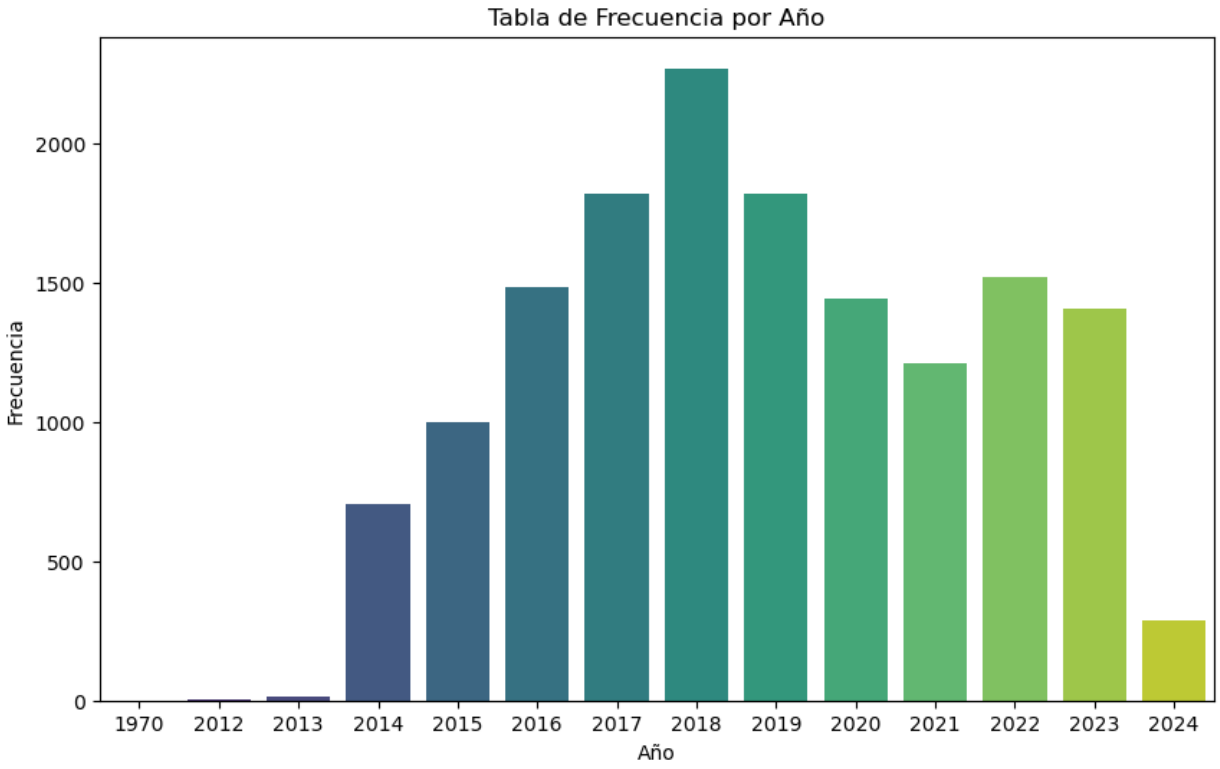
Tablas de frecuencias y representaciones gráficas:

En el análisis unidimensional de datos, las tablas de frecuencias y las representaciones gráficas son herramientas esenciales para comprender y visualizar la distribución de los valores de una variable. Las tablas de frecuencias resumen la cantidad de veces que ocurre cada valor, mientras que las representaciones gráficas permiten una visualización más clara y efectiva de los datos.

Estas tablas de frecuencia representan la distribución de los avisos de mantenimiento registrados en el sistema de información SAP PM a lo largo de los años, meses y semanas.

Figura 25

Frecuencia de fallas por año



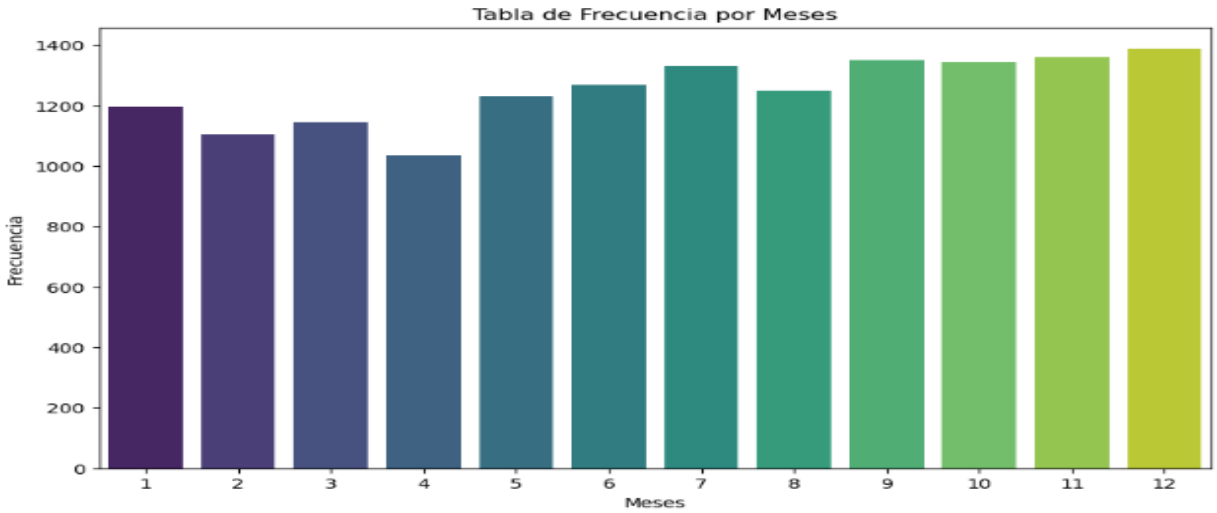
Fuente. Elaboración propia

Distribución por Año:

- El año 2018 tuvo la mayor cantidad de avisos, con 2271 registros.
- Los años 2019 y 2017 también presentan un número significativo de avisos, con 1821 y 1818 registros, respectivamente.
- El año 2012 y 2013 tienen muy pocos avisos, con solo 2 y 16 registros, respectivamente.

Figura 26

Frecuencia de fallas por mes



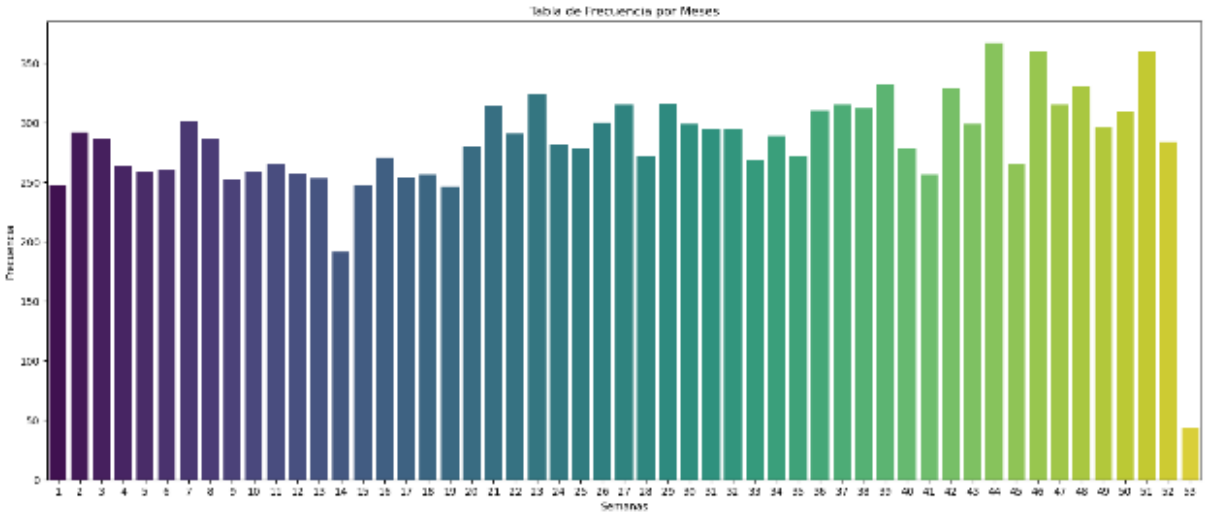
Fuente. Elaboración propia

Distribución por Mes:

- Los últimos meses representan la mayor cantidad de registros, particularmente los meses de diciembre (12) y noviembre (11) tiene la mayor cantidad de avisos, con 1389 y 1361 respectivamente.
- El mes abril (4) tiene la menor cantidad de avisos, con 1036 registros.

Figura 27

Frecuencia de fallas por semana



Fuente. Elaboración propia

Distribución por Semana:

- La semana 25 tiene la mayor cantidad de avisos, con 377 registros.
- Las semanas 23, 29 y 39 también presentan una alta frecuencia, con 350, 332 y 371 registros, respectivamente.
- La semana 53 tiene la menor cantidad de avisos, con solo 45 registros.

Medidas de posición:

En conjunto, el cálculo de la media, moda y mediana proporciona una visión completa de la distribución de los datos, permitiendo identificar la tendencia central, la frecuencia de valores y la posición central de los datos. Estas medidas son fundamentales en el análisis estadístico y descriptivo, ya que ayudan a comprender la estructura y características de los datos, identificar posibles sesgos, evaluar la simetría de la distribución y detectar valores atípicos que puedan influir en las conclusiones obtenidas a partir de los datos.

Se llevaron a cabo cálculos específicos para la variable TMEF (Tiempo Medio Entre Fallas) en el conjunto de datos relacionado con el registro de paradas de equipos. Se calcularon la media, moda y mediana de la variable TMEF, lo que proporcionó información clave sobre la distribución de los tiempos medios entre fallas en los equipos analizados.

Media_TMEF:

La media de TMEF se calculó como el promedio de los tiempos medios entre fallas en el conjunto de datos. Este cálculo permitió obtener un valor representativo de la frecuencia promedio de las fallas en los equipos, lo que es fundamental para comprender la estabilidad operativa y la fiabilidad de estos.

Moda_TMEF:

La moda de TMEF se calculó como el valor más frecuente en los tiempos medios entre fallas. Identificar la moda en este contexto proporciona información sobre el tiempo medio más común entre las fallas, lo que puede ser útil para identificar patrones de repetición en los intervalos de tiempo entre averías.

Mediana_TMEF:

La mediana de TMEF se calculó como el valor central en la distribución de los tiempos medios entre fallas. Al ser una medida robusta ante valores extremos, la mediana de TMEF proporcionó información sobre el tiempo medio que divide el conjunto de datos en dos partes iguales, lo que es relevante para comprender la variabilidad en los intervalos de tiempo entre fallas.

Medidas de dispersión:

Estas medidas estadísticas son fundamentales para comprender la dispersión y la variabilidad de los tiempos medios entre fallas en los equipos analizados. A continuación, se detalla la importancia de cada una de estas medidas:

Desviación Estándar:

La desviación estándar es una medida de dispersión que indica cuánto se alejan los valores individuales de la media en un conjunto de datos.

En el contexto del TMEF, la desviación estándar proporciona información sobre la variabilidad de los tiempos medios entre fallas. Valores más altos de desviación estándar indican una mayor dispersión de los datos alrededor de la media, lo que sugiere una variabilidad significativa en los intervalos de tiempo entre fallas.

Coeficiente de Variación:

El coeficiente de variación es una medida relativa de la dispersión de los datos, calculada como la desviación estándar dividida por la media y multiplicada por 100 para expresar el resultado en porcentaje.

En el caso del TMEF, el coeficiente de variación permite comparar la variabilidad de los tiempos medios entre fallas en relación con la media. Un coeficiente de variación más alto indica una mayor variabilidad relativa en los intervalos de tiempo entre fallas, mientras que un valor bajo sugiere una mayor consistencia en los datos.

Al calcular el coeficiente de variación y la desviación estándar para el TMEF en el ejercicio, se obtuvo información adicional sobre la dispersión y la consistencia de los tiempos medios entre fallas en los equipos analizados. Estas medidas son cruciales para evaluar la

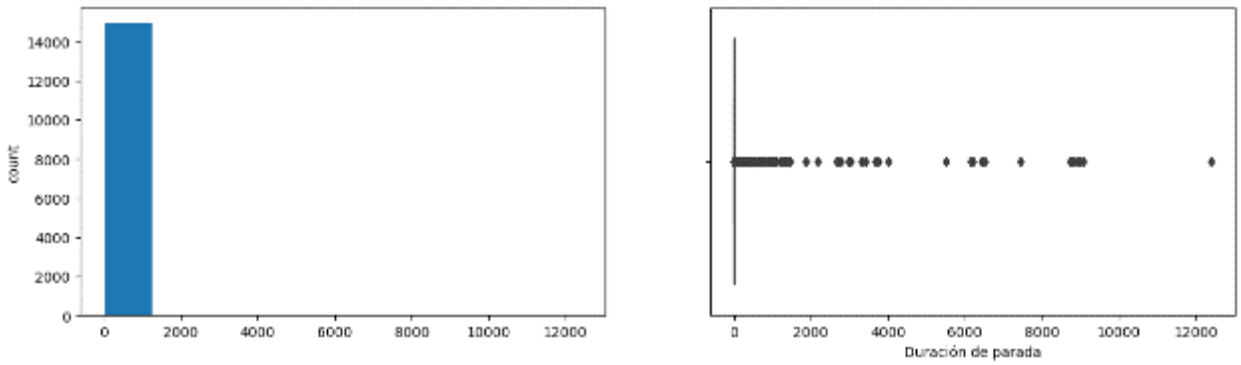
estabilidad y la fiabilidad de los equipos, así como para identificar posibles patrones de comportamiento en los intervalos de tiempo entre averías.

Análisis de componentes principales CPA

El Análisis de Componentes Principales (PCA) es un método estadístico fundamental que permite simplificar la complejidad de espacios muestrales con múltiples dimensiones, al tiempo que conserva su información.

Figura 28

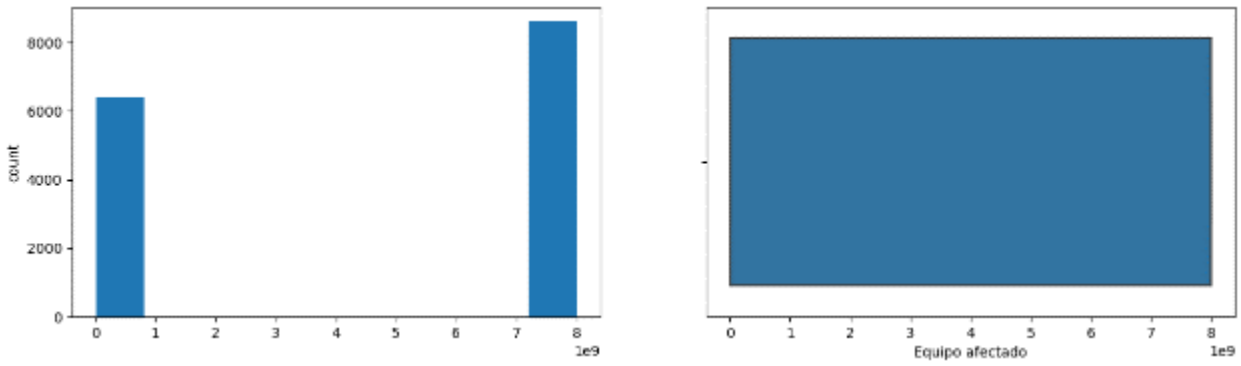
Duración de Parada (CPA)



Fuente. Elaboración propia

Figura 29

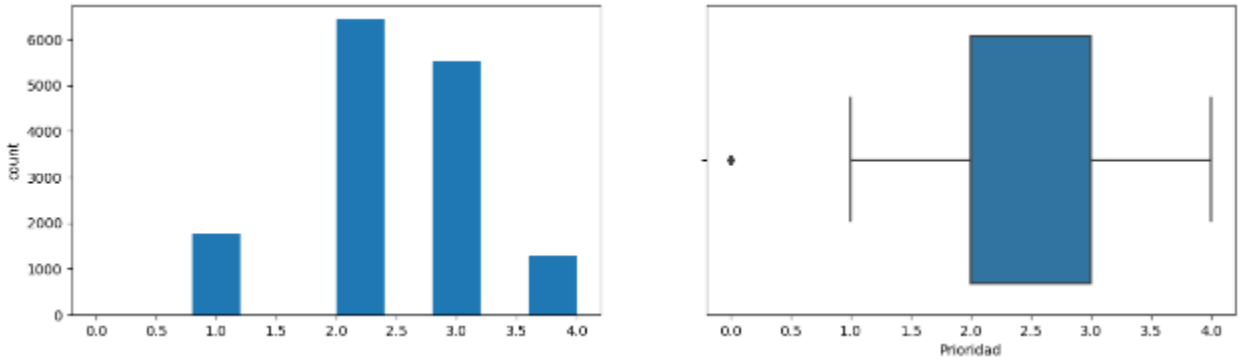
Equipo afectado (CPA)



Fuente. Elaboración propia

Figura 30

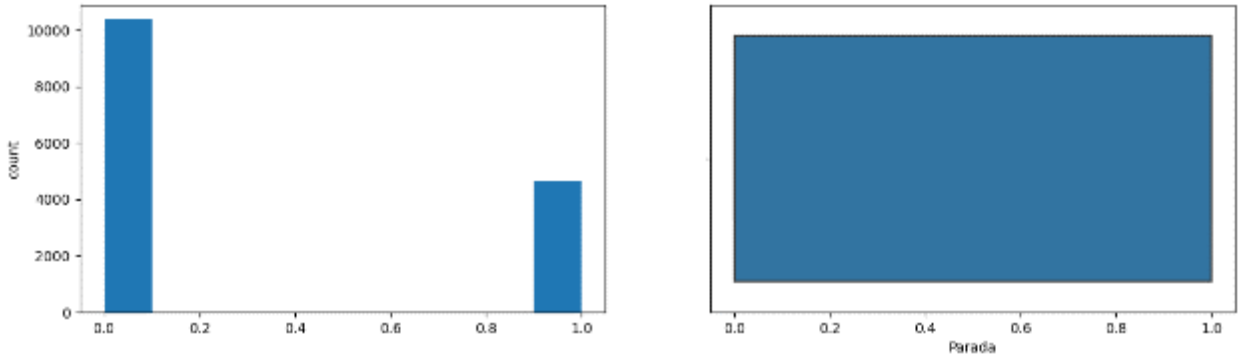
Prioridad (CPA)



Fuente. Elaboración propia

Figura 31

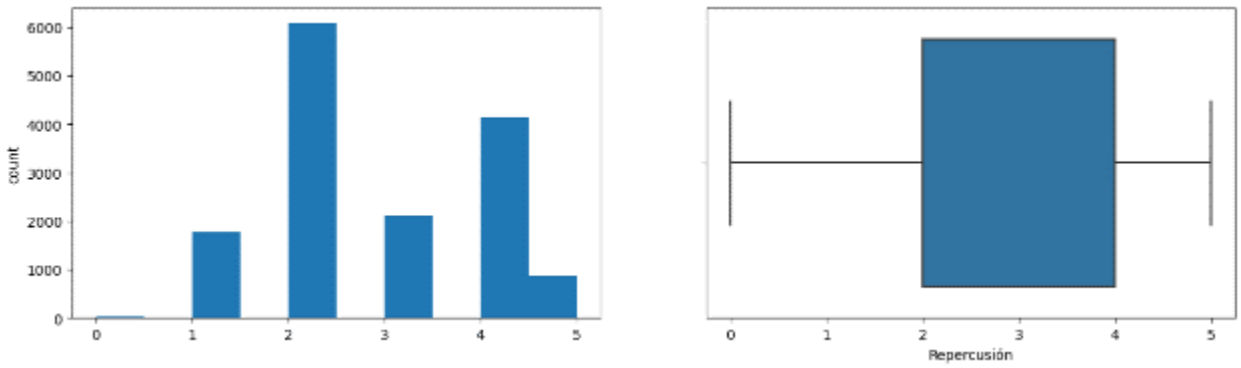
Parada (CPA)



Fuente. Elaboración propia

Figura 32

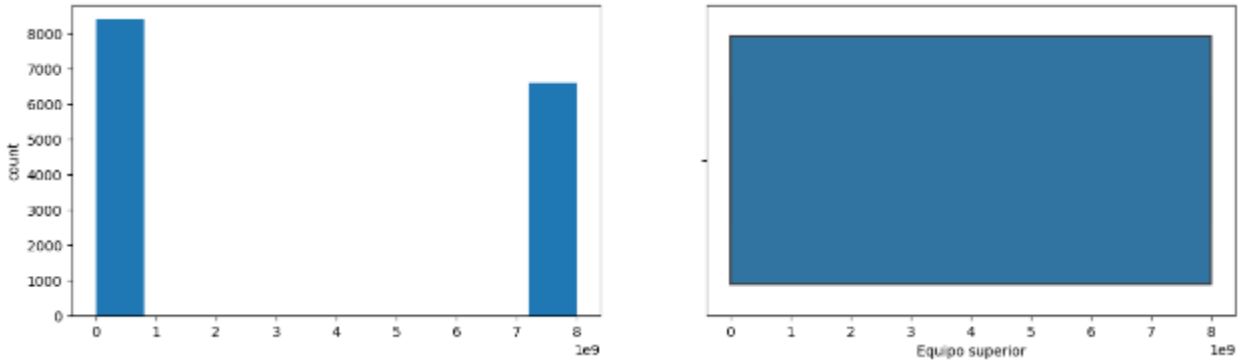
Repercusión (CPA)



Fuente. Elaboración propia

Figura 33

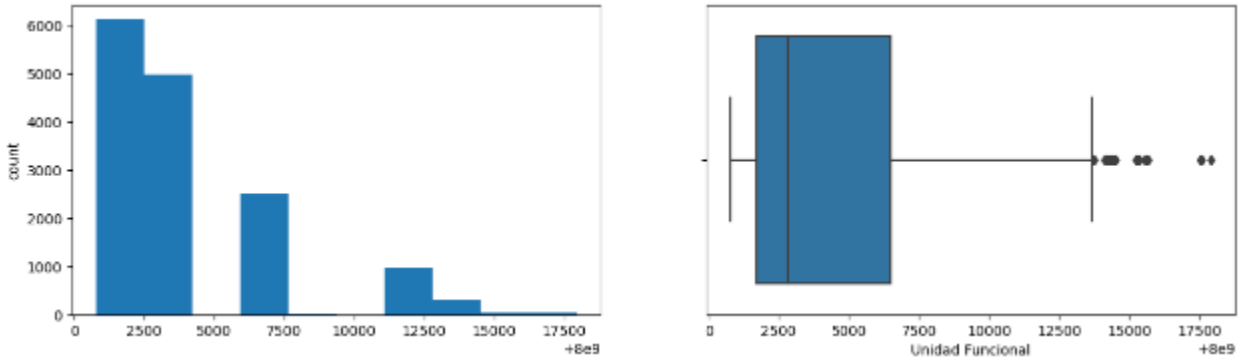
Equipo superior (CPA)



Fuente. Elaboración propia

Figura 34

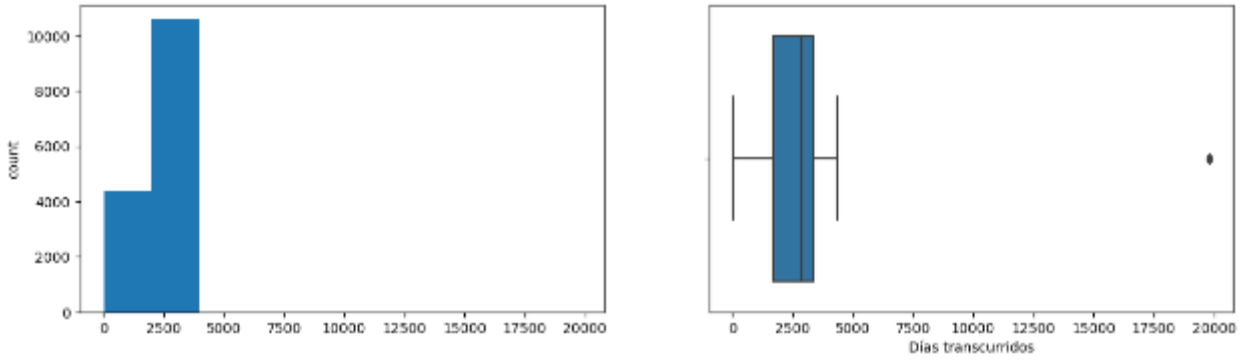
Unidad Funcional (CPA)



Fuente. Elaboración propia

Figura 35

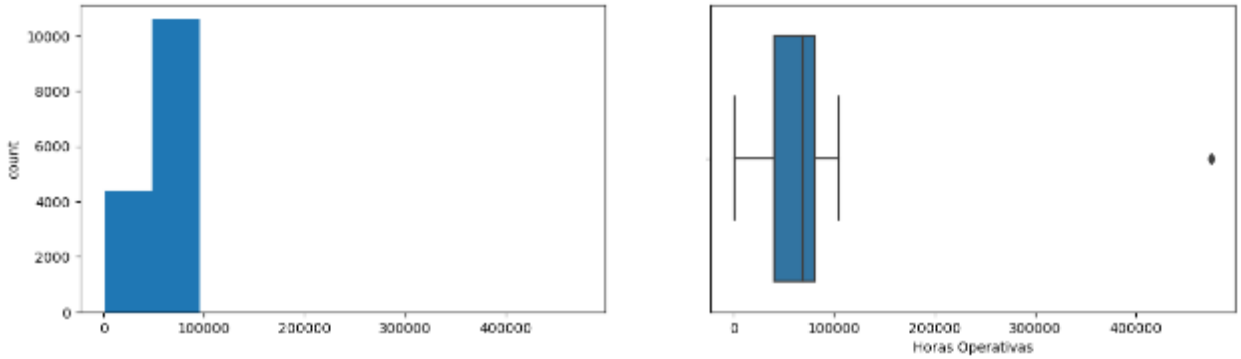
Días transcurridos (CPA)



Fuente. Elaboración propia

Figura 36

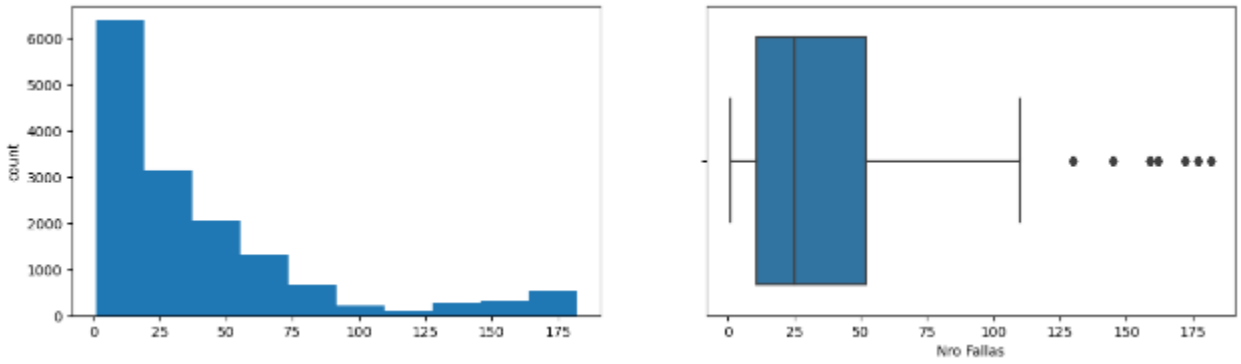
Horas operativas (CPA)



Fuente. Elaboración propia

Figura 37

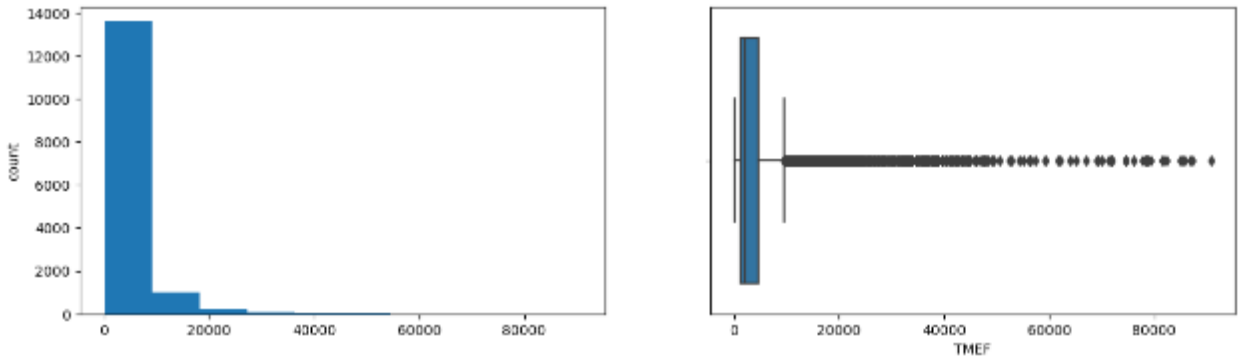
Nro Fallas (CPA)



Fuente. Elaboración propia

Figura 38

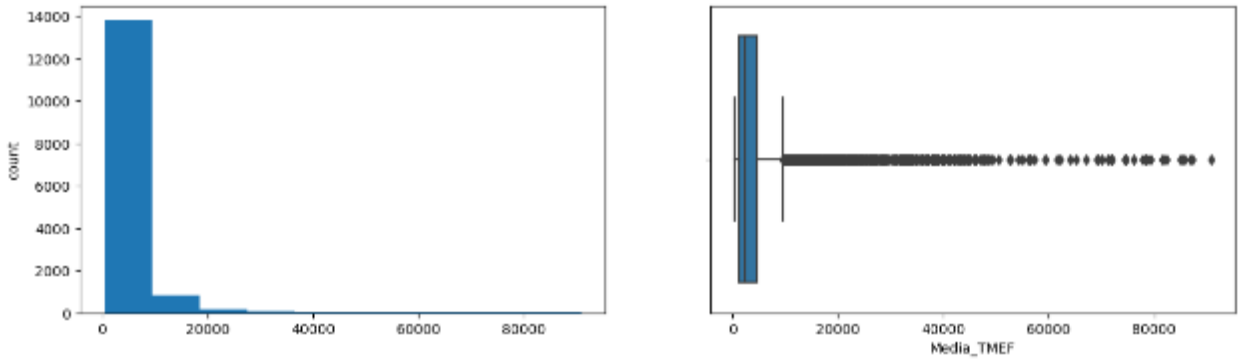
TMEF (CPA)



Fuente. Elaboración propia

Figura 39

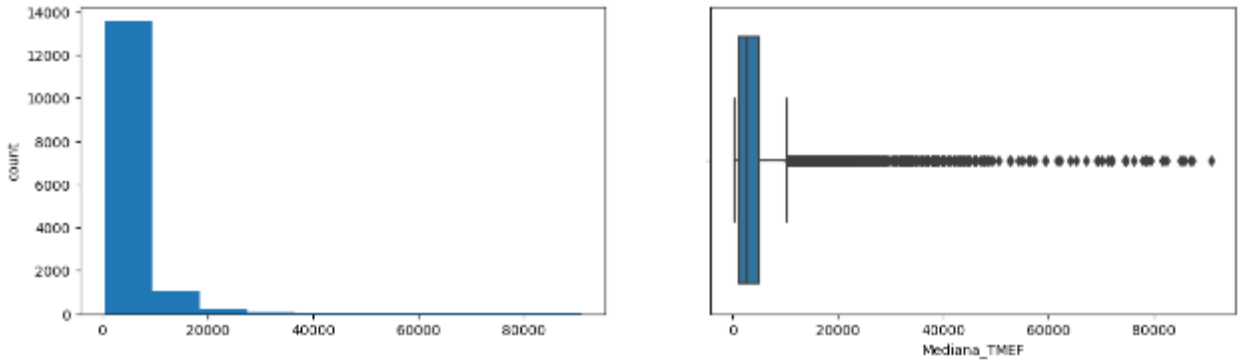
Media_TMEF (CPA)



Fuente. Elaboración propia

Figura 40

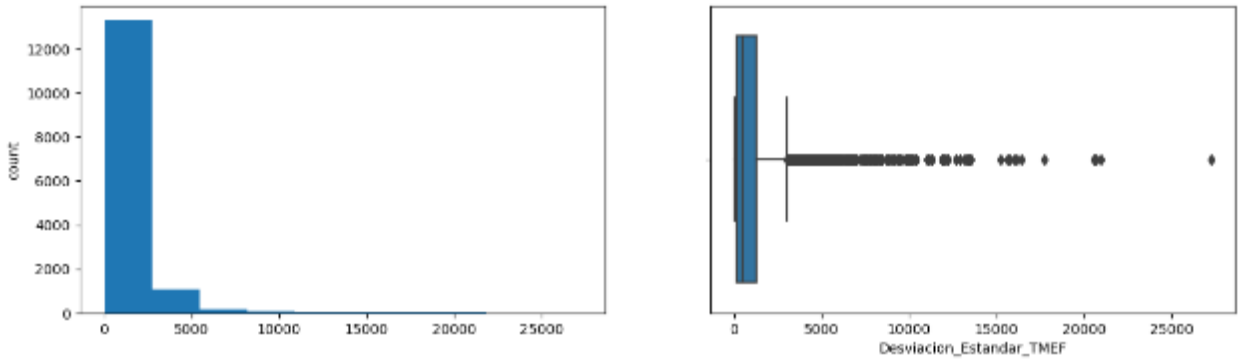
Mediana_TMEF (CPA)



Fuente. Elaboración propia

Figura 41

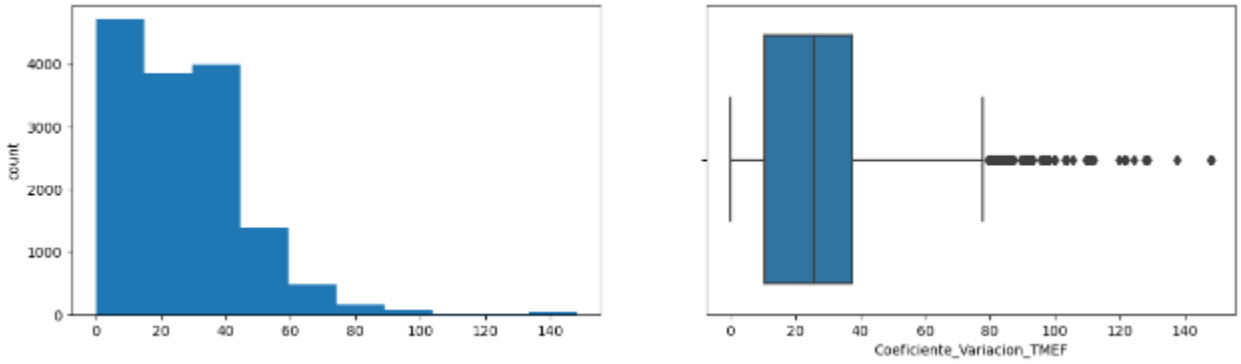
Desviación estándar (CPA)



Fuente. Elaboración propia

Figura 42

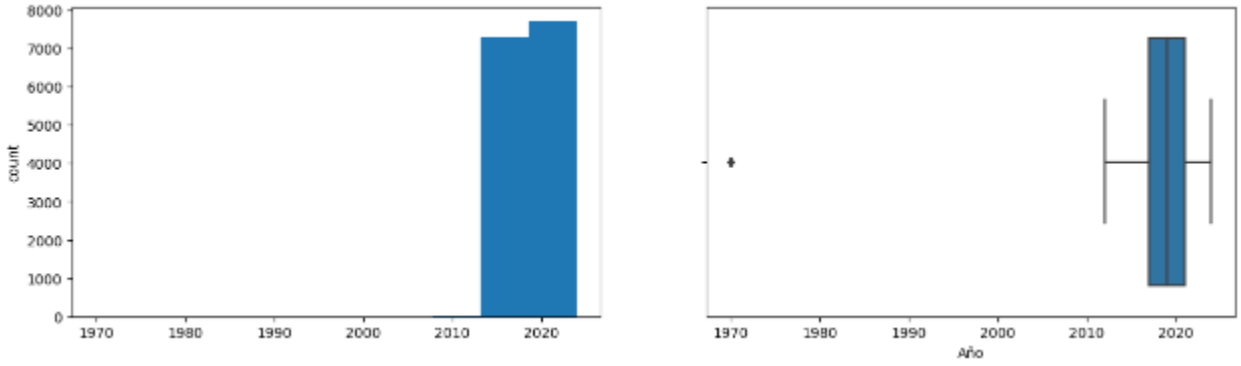
Coeficiente variación (CPA)



Fuente. Elaboración propia

Figura 43

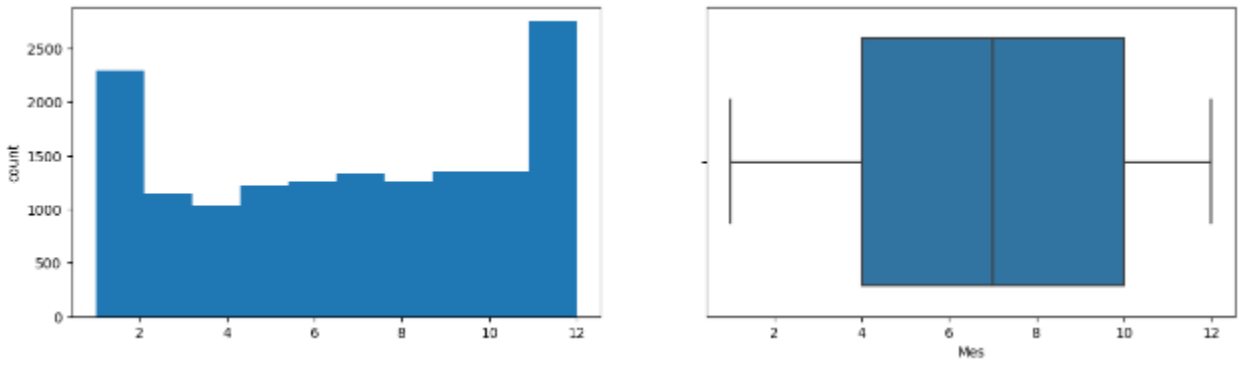
Año (CPA)



Fuente. Elaboración propia

Figura 44

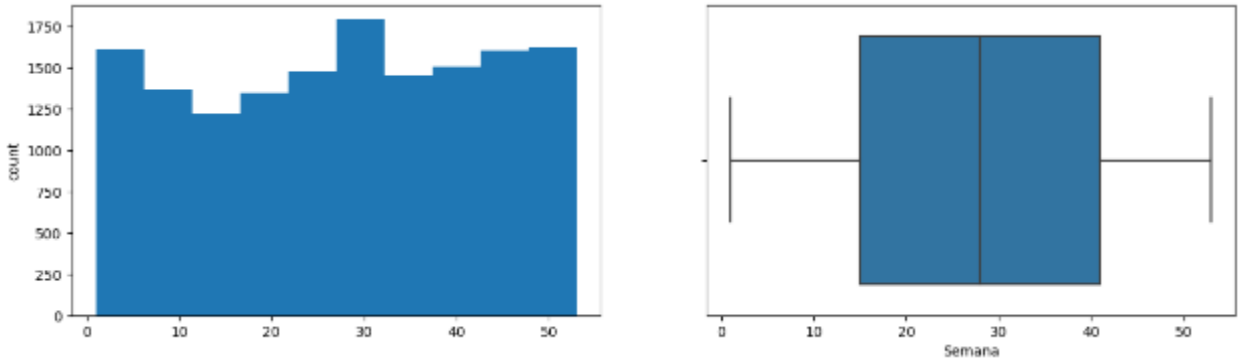
Mes (CPA)



Fuente. Elaboración propia

Figura 45

Semanas (CPA)



Fuente. Elaboración propia

El análisis de componentes principales es una técnica estadística utilizada para identificar patrones subyacentes en un conjunto de datos, reduciendo la dimensionalidad y resumiendo la información en componentes no correlacionados llamados componentes principales. A continuación, se destacan aspectos relevantes del análisis de componentes principales realizado en el ejercicio:

Número de Componentes Principales:

Se definió el número de componentes principales a generar en función de la cantidad de variables en los datos. En este caso, se utilizó el número total de variables relacionadas con los tiempos medios entre fallas para determinar el número de componentes a generar en el PCA. Generación de Componentes Principales:

Se aplicó el análisis de componentes principales (PCA) para encontrar los componentes principales que explican la mayor parte de la varianza en los datos. Esto permitió reducir la dimensionalidad de los datos y representar la información de manera más compacta a través de combinaciones lineales de variables originales.

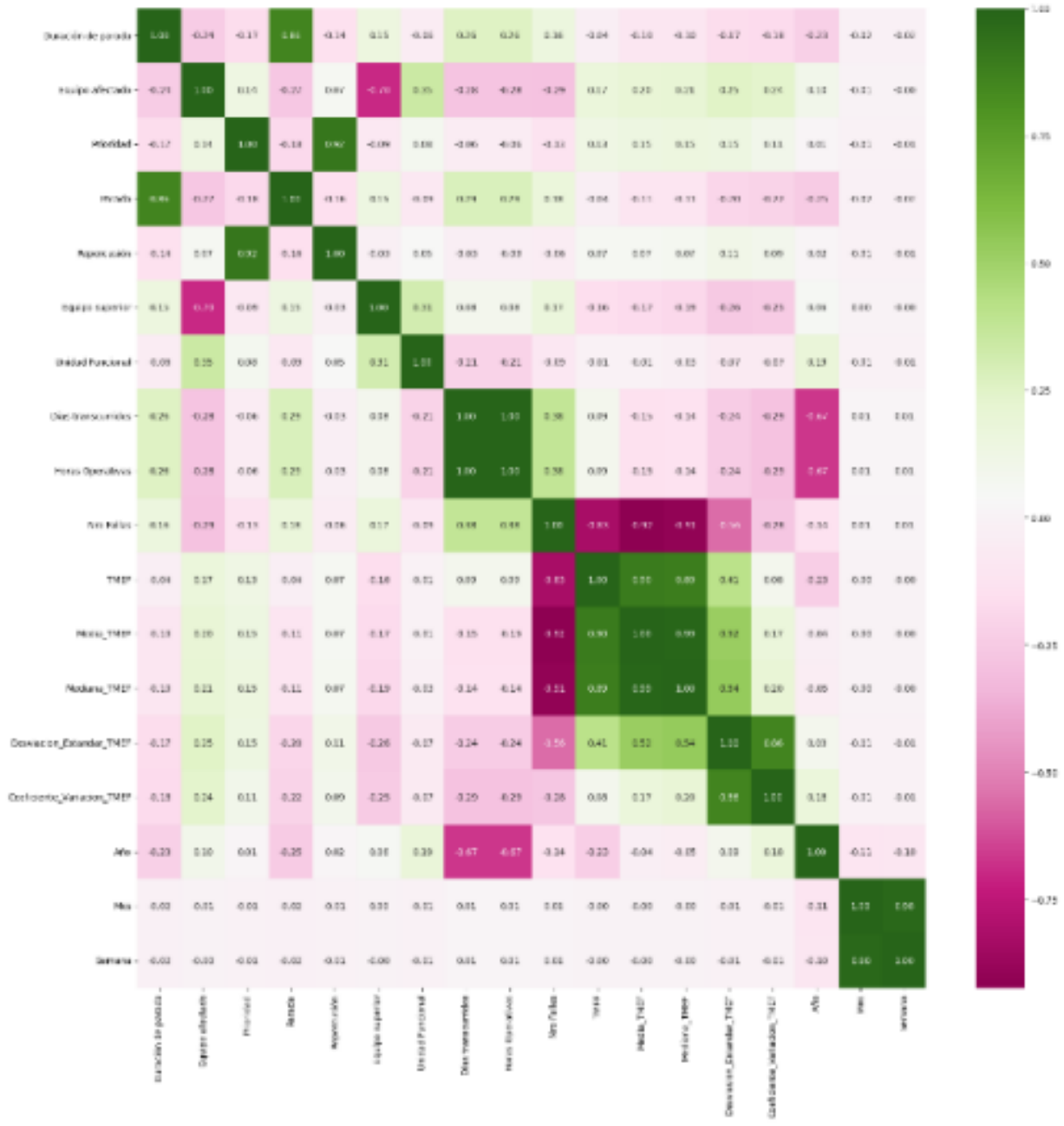
Análisis de correlación entre variables:

El análisis de correlación es una herramienta fundamental en el análisis unidimensional de datos. Nos ayuda a entender si existe una dependencia o asociación entre dos o más variables. Aunque no implica causalidad, nos indica si los cambios en una variable están relacionados con los cambios en otra.

En esta exploración de datos, la correlación nos ayuda a identificar patrones y relaciones entre variables, lo que puede conducir a nuevas hipótesis y descubrimientos.

Figura 46

Matriz de Correlación CPA



Fuente. Elaboración propia

La duración de parada muestra una correlación positiva significativa con la variable “Parada” (coeficiente de correlación de 0.86). Esto sugiere que a medida que la duración de una parada aumenta, también lo hace la frecuencia o gravedad de las paradas.

Además, la duración de la parada tiene correlaciones negativas con otras variables como “Equipo afectado” y “Prioridad”. Esto podría indicar que a medida que aumenta la duración de la parada, menos equipos podrían estar afectados por el problema y posiblemente se asigne una menor prioridad a la resolución de la parada.

Por otro lado, la variable “Prioridad” muestra una correlación positiva notable con la variable “Repercusión” (coeficiente de correlación de 0.92), lo que implica que los incidentes con mayor prioridad tienden a tener mayores repercusiones.

Varianza Explicada:

Se evaluó el porcentaje de varianza explicada por cada componente principal generado en el PCA. Este análisis proporciona información sobre la importancia de cada componente en la explicación de la variabilidad total de los datos, lo que ayuda a identificar los componentes más relevantes en la estructura de los datos.

Tabla 7

Varianza Explicada

	PC1	PC2	PC3	PC4	PC5
Duración de parada	-0.01	0.03	-0.01	0.02	0.06
Equipo afectado	0.21	-0.02	0.48	0.03	0.31
Prioridad	0.14	0.04	0.34	0.13	-0.57
Parada	-0.19	0.15	-0.14	0.02	0.05
Repercusión	0.09	0.04	0.32	0.13	-0.6
Equipo superior	-0.19	0.01	-0.48	-0.04	-0.31
Unidad Funcional	0.19	-0.19	-0.21	-0.05	0.08

Días transcurridos	-0.30	0.42	0.14	0.06	-0.08
Horas operativas	-0.30	0.42	0.14	0.08	0.06
Número de fallas	-0.29	-0.08	0.05	0.02	-0.08
TMEF	0.35	0.38	-0.16	0.00	-0.02

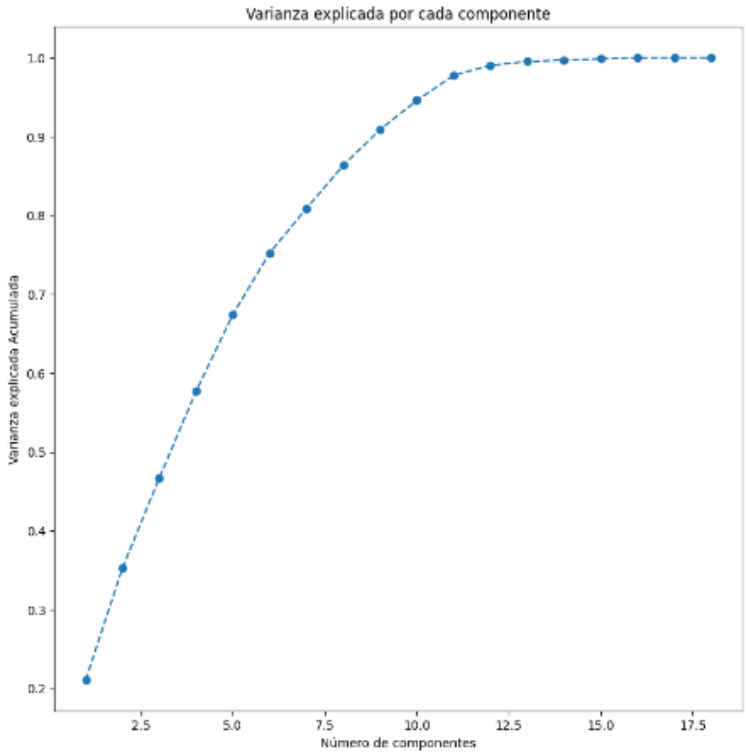
Nota. Esta tabla muestra la varianza explicada por cada uno de los componentes principales obtenidos en el análisis. *Fuente.* Elaboración propia.

Visualización de la Varianza Explicada:

Se visualizó la varianza explicada por cada componente principal a través de un gráfico que muestra cómo la varianza explicada acumulada varía con el número de componentes. Esta visualización es útil para determinar cuántos componentes son necesarios para capturar una cantidad significativa de la varianza en los datos.

Figura 47

Varianza Explicada Acumulada



Fuente. Elaboración propia

Definir Modelos Predictivos

Los modelos predictivos en el ámbito del mantenimiento de activos en la industria Oíl & Gas representan una herramienta fundamental para anticipar y gestionar eficazmente posibles fallas en equipos y maquinaria. Estos modelos utilizan datos históricos detallados, como registros de mantenimiento, tiempos entre fallas y condiciones operativas, junto con técnicas avanzadas de análisis y aprendizaje automático, para predecir con precisión futuras incidencias.

Al analizar patrones y tendencias en los datos, los modelos predictivos pueden clasificar los avisos de mantenimiento en categorías de probabilidad de falla, permitiendo a los equipos de mantenimiento priorizar sus acciones de manera eficiente y proactiva.

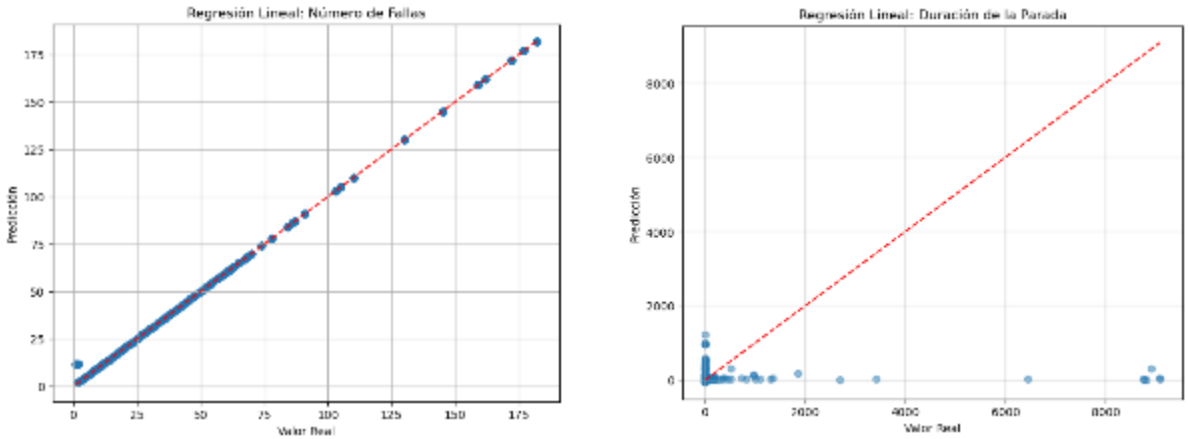
La implementación de modelos predictivos no solo contribuye a la reducción de tiempos de inactividad no planificados, sino que también optimiza los recursos y mejora la fiabilidad operativa de los activos, impulsando así la eficiencia y la rentabilidad en el sector Oíl & Gas.

Regresión Lineal:

La regresión lineal es un método clásico que asume una relación lineal entre las variables predictoras y la variable objetivo. La literatura científica ha demostrado su aplicabilidad en una variedad de campos. En nuestro estudio, este modelo nos permitió identificar un alto coeficiente de determinación (R^2) para ambas variables objetivo: el número de fallos y la duración de la parada1. Además, el bajo Error Cuadrático Medio (MSE) sugiere que las predicciones son precisas y se acercan a los valores reales.

Figura 48

Modelo de Regresión Lineal



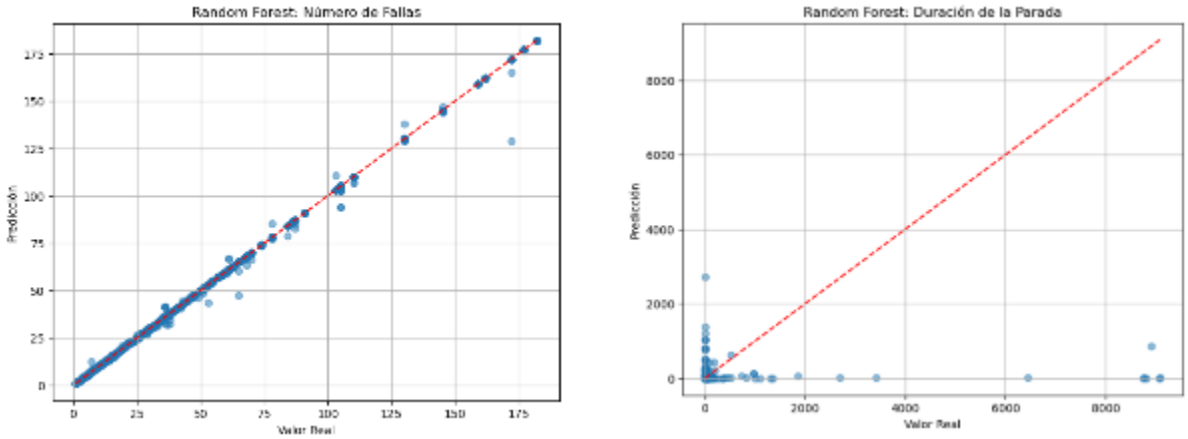
Fuente. Elaboración propia

Random Forest:

El Random Forest, un algoritmo de conjunto, combina múltiples árboles de decisión. La literatura destaca su capacidad para manejar relaciones no lineales y su resistencia al sobreajuste. En nuestro caso, obtuvimos un MSE muy bajo, lo que indica una excelente precisión en las predicciones. El alto R^2 confirma la capacidad del modelo para capturar patrones complejos en los datos.

Figura 49

Modelo Random Forest



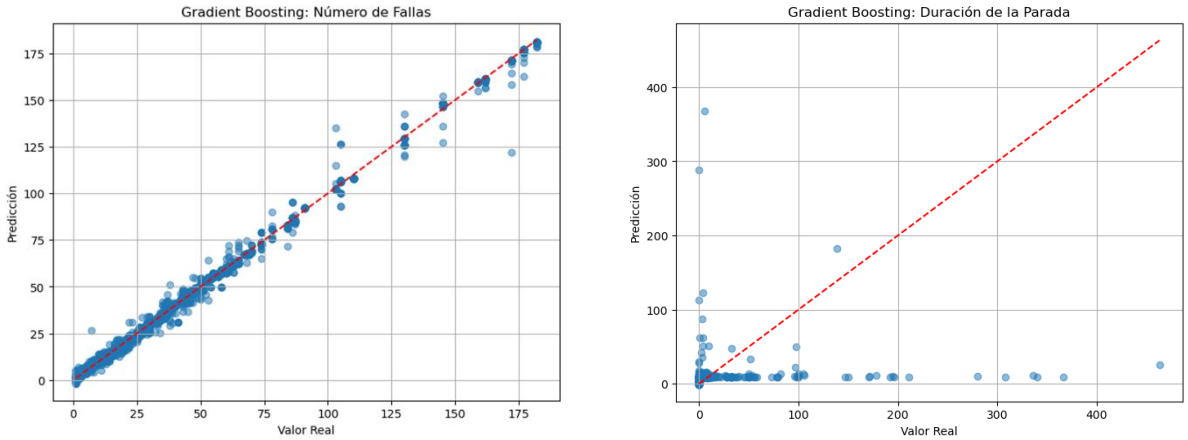
Fuente. Elaboración propia

Gradient Boosting:

El Gradient Boosting, otro método de conjunto, mejora iterativamente el rendimiento del modelo al combinar modelos débiles. La literatura ha elogiado su robustez y su capacidad para reducir el sesgo. Nuestros resultados también respaldan esto: un bajo MSE y un alto R^2 indican una buena capacidad de predicción.

Figura 50

Modelo Gradient Boosting



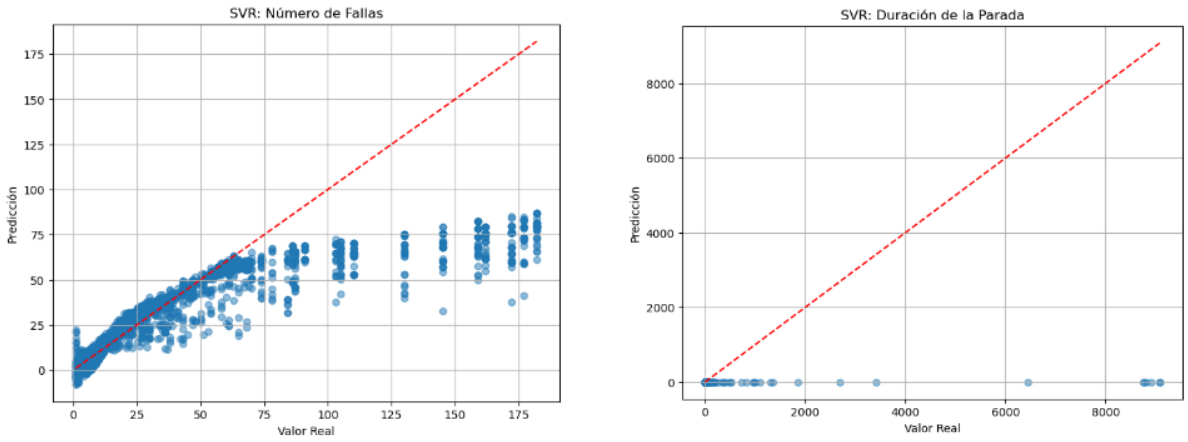
Fuente. Elaboración propia

SVR (Support Vector Regression):

El SVR se basa en vectores de soporte y se ha aplicado con éxito en problemas de regresión. La literatura destaca su capacidad para manejar datos no lineales y su resistencia al ruido. Aunque el MSE fue más alto que en otros modelos, sigue siendo una herramienta valiosa. Sin embargo, el R^2 menor sugiere que podría no ser el mejor modelo para nuestro problema específico.

Figura 51

Modelo SVR



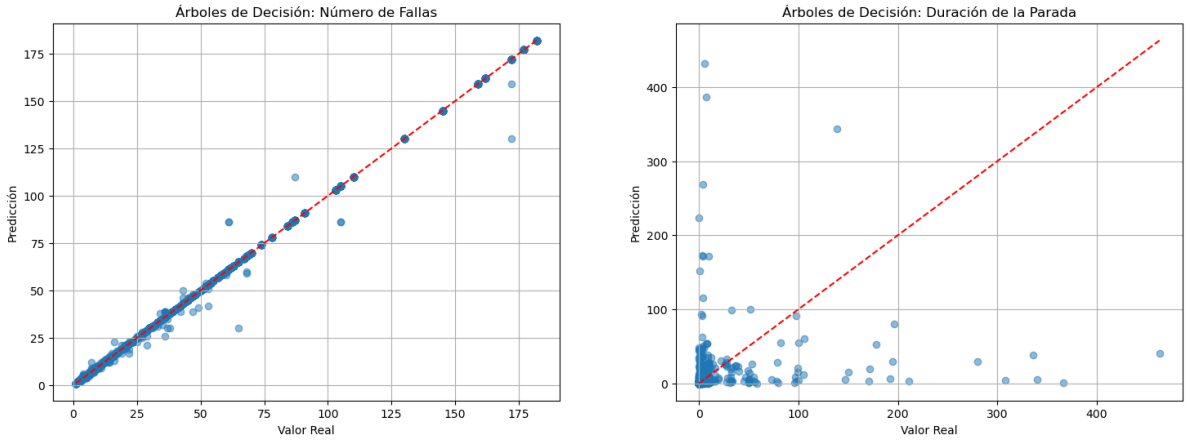
Fuente. Elaboración propia

Árboles de Decisión

Son estructuras jerárquicas que se utilizan para tomar decisiones o predecir resultados. Cada nodo en el árbol representa una pregunta o una condición sobre las características (variables) de los datos. Las ramas del árbol representan las posibles respuestas a esas preguntas, y las hojas del árbol contienen las predicciones o las clases finales.

Figura 52

Modelo Árboles de Decisión



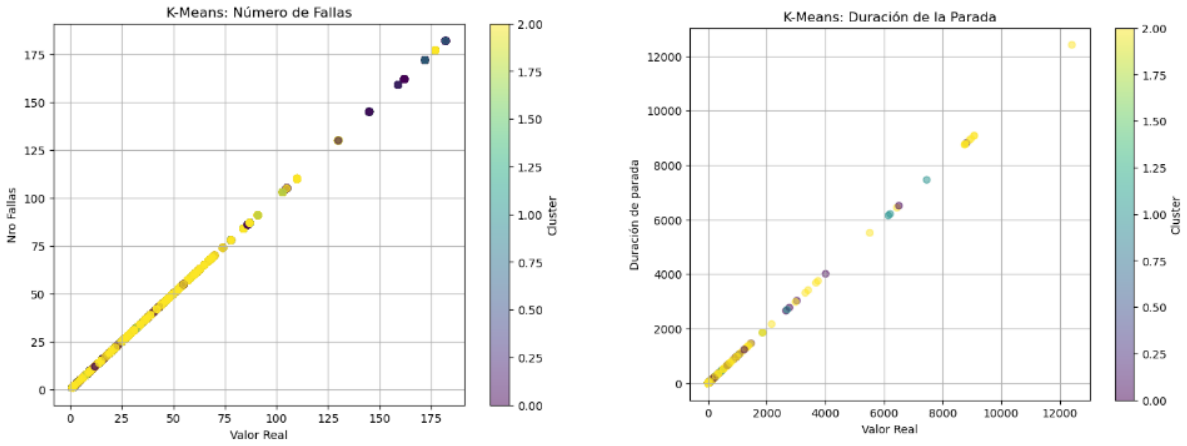
Fuente. Elaboración propia

Análisis de Agrupamiento K-Means:

El K-Means es una técnica de agrupamiento que divide los datos en grupos homogéneos. La literatura ha demostrado su utilidad en la segmentación de datos. Identificamos tres grupos con diferentes valores medios para el número de fallos y la duración de la parada. Esta segmentación podría ayudar a definir estrategias de mantenimiento específicas para cada grupo de equipos.

Figura 53

Modelo Agrupamiento K-Means



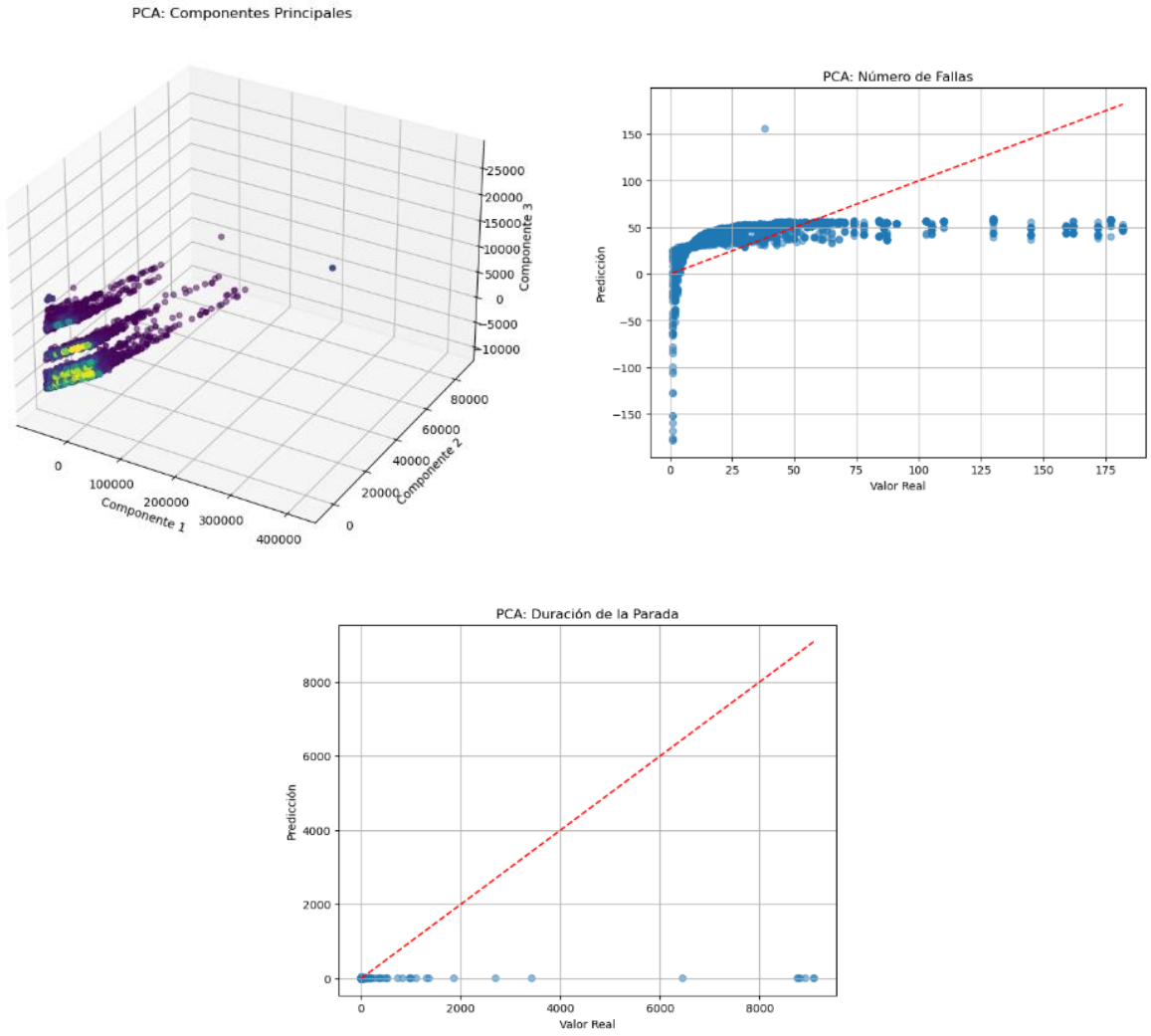
Fuente. Elaboración propia

Análisis de Componentes Principales (PCA):

El PCA es una técnica de reducción de dimensionalidad. Aunque el R^2 fue bajo, el PCA nos permitió explorar relaciones complejas entre las características. Esta herramienta es valiosa para comprender la variabilidad en los datos y simplificar su representación.

Figura 54

Modelo Análisis Componentes Principales PCA



Fuente. Elaboración propia

Algoritmos de Regresión.

Tabla 8

Comparación de rendimiento de los modelos. Rmse del Test

	KNN	R. LIN	D.TREE	SVR
N° Fallas	4.7965	1.5695	1.4959	22.3488

Dura. Parada	34.3351	32.2124	37.1477	31.3430

Nota. Esta tabla compara el rendimiento de diferentes modelos mediante la métrica del RMSE en el conjunto de prueba.. *Fuente.* Elaboración propia.

En resumen, estos resultados respaldan la importancia del Mantenimiento 4.0 y la aplicación de modelos precisos en la gestión de activos en la industria del petróleo y gas. La literatura científica continúa siendo una fuente valiosa para validar y mejorar estos enfoques.

Modelos De Ensamble

Tabla 9

Comparación modelos de ensamble de regresión Número de Fallas

	BAGGING	RF	AdaBoost	GBoost
N° Fallas	5.4103	0.6763	10.8058	2.3674
Dura. Parada	31.8643	29.7637	31.4770	33.5098

Nota. Esta tabla compara el rendimiento de modelos de ensamble de regresión según el número de fallas predichas. *Fuente.* Elaboración propia.

Ajuste de hiperparametros

De todos los algoritmos anteriores el de menor error fue Random Forest (RF), utilizando la función **GridSearchCV** de Python el cual emplea la técnica de Cross Validation se encuentran los mejores hiperparametros para este algoritmo.

Tabla 10

Ajuste de Hiperparámetros

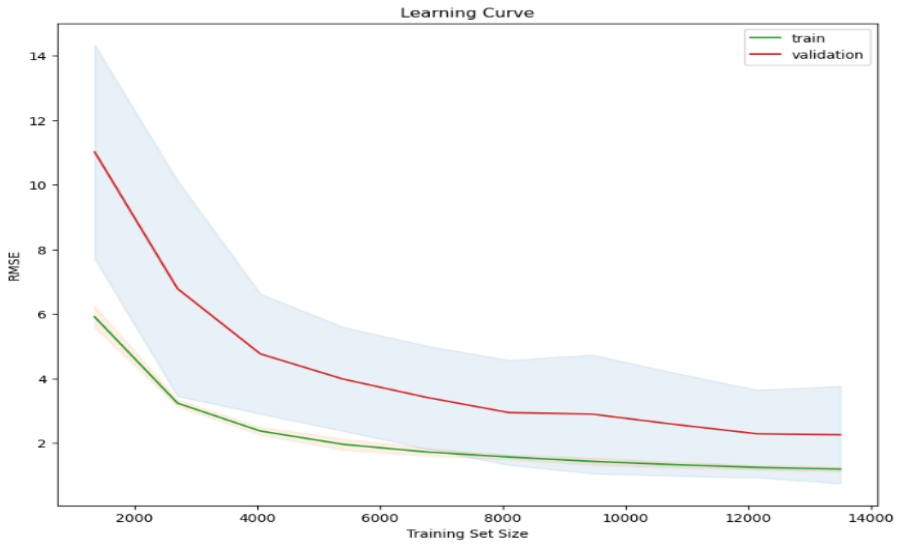
RF	Hiperparametros Iniciales	Hiperparametros Óptimos	RMSE test	RMSE test ajustado
N° Fallas	RandomForestRegressor	RandomForestRegressor	1,20	0,6763

	max_depth = 16 n_estimators=50 n_jobs=-1 random_state=55	max_depth=15 min_samples_leaf=1 n_estimators=27 n_jobs=-1 random_state=55		
Duración de Parada	RandomForestRegressor max_depth= 16 n_estimators=50 n_jobs=-1 random_state=55	RandomForestRegressor max_depth=5 min_samples_leaf=60 n_estimators=200 n_jobs=-1 random_state=42	33,5	29,7673

Nota. Esta tabla muestra el ajuste de hiperparámetros realizado para optimizar el rendimiento de los modelos. *Fuente.* Elaboración propia.

Figura 55

Curva de aprendizaje para el RF ajustado. Nro. Fallas



Fuente. Elaboración propia

Se evidencia un error de generalización reducido entre las dos curvas, las cuales son estables más allá de un tamaño de conjunto de entrenamiento de 12000, tanto los datos de

validación como los de entrenamiento, reducen su error al aumentar el tamaño de los datos, en general es uno de los mejores modelos para predecir tanto el número de fallas como, la duración de parada.

Además de predecir fallas en los equipos, un análisis importante es poder hacer predicciones sobre las paradas de los equipos, aunque hay que considerar que toda parada es producto de una falla, no toda falla produce una parada. Por lo que se consideró tanto las fallas que generaron parada como las que no, de tal manera que los algoritmos de aprendizaje reconocieran las que producen o no parada. Las variables de entrada o predictoras son: Prioridad, Repercusión, Horas Operativa, TMEF, Duración de parada y Unidad Funcional.

La variable de respuesta es ParaDA, que es de tipo binaria la cual indica la ausencia (0) o presencia de parada (1). Con base en el análisis exploratorio inicial se encuentra valores atípicos en la variable "duración de parada" los cuales son ajustados de acuerdo con el dominio del experto. Se determina un desbalance en la distribución de la variable predictora *"duración de parada"* y en la variable de respuesta *"parada"*.

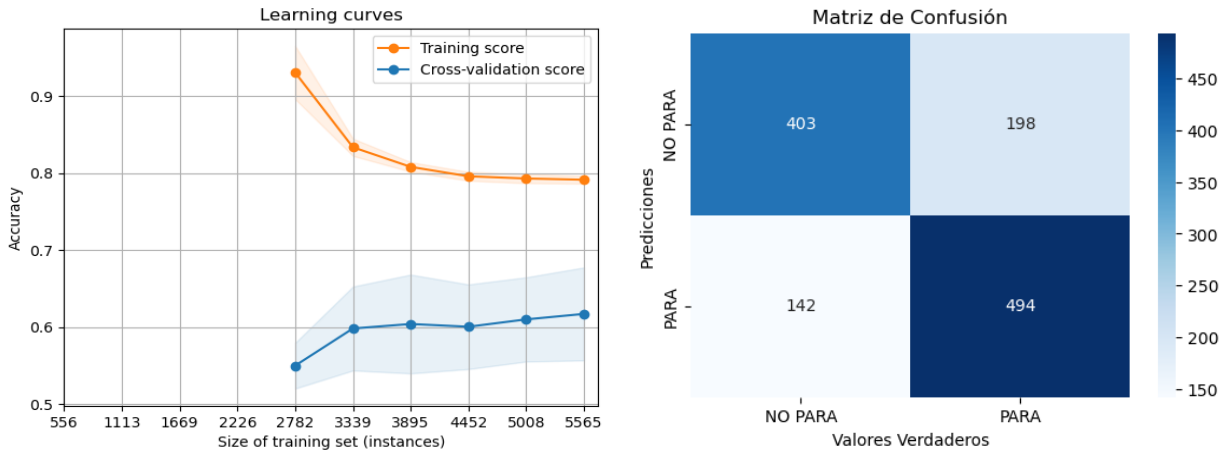
El desbalance consiste en que el 80% de los datos de duración de parada están en cero. Para solucionar este aspecto se genera un submuestreo de manera aleatoria en donde se equilibre

la distribución de esta variable. Al realizar este paso se identifica que también queda balaceada la variable de respuesta, esto por la naturaleza de su relación.

Logistic Regresión

Figura 56

Logistic Regresión



Fuente. Elaboración propia

Tabla 11

Resultados Logistic Regresión

	Precision	Recall	F1-score	Support
No Para	0.67	0.74	0.70	545
Para	0.78	0.71	0.74	692
accuracy			0.73	1237
macro avg	0.72	0.73	0.72	1237
weighted avg	0.73	0.73	0.73	1237

Nota. Esta tabla presenta los resultados obtenidos mediante el modelo de Regresión Logística.

Fuente. Elaboración propia.

Precisión: De un total de 636 predicciones de Parada, el algoritmo clasificó correctamente 494 lo que equivale a un 77.6%, el avg ponderado es de 73% la cual ajusta la

puntuación teniendo en cuenta la proporción del tamaño de las dos clases en el test (suport), como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. Verdaderas (No Para) el algoritmo clasificó bien 403, con un score de 74% el avg ponderado es de 73%

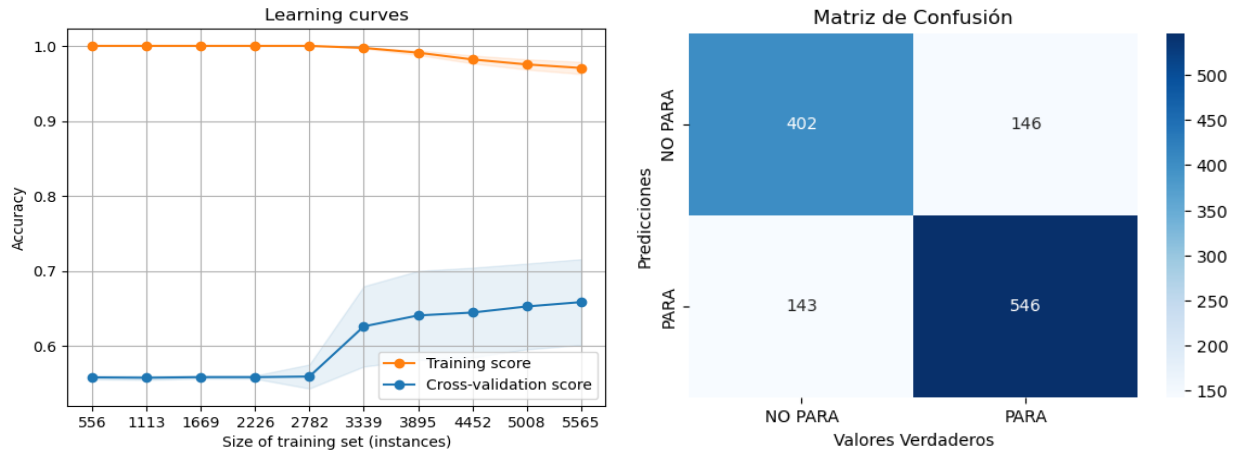
F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 73%

Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (897), sobre el total de predicciones (1237), el cual da una puntuación de 0.7251. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc (entradas), "el modelo acertó si una falla se convertirá en parada o no en un 73% de todas las predicciones".

KNN Classifier

Figura 57

KNN Classifier



Fuente. Elaboración propia

Tabla 12

Resultados KNN Classifier

	Precision	Recall	F1-score	Support
No Para	0.73	0.74	0.74	545
Para	0.79	0.79	0.79	692
accuracy			0.77	1237
macro avg	0.76	0.76	0.76	1237
weighted avg	0.77	0.77	0.77	1237

Nota. Esta tabla muestra los resultados obtenidos mediante el clasificador KNN. *Fuente.* Elaboración propia.

Precision: De un total de 689 predicciones de Parada, el algoritmo clasificó correctamente 545 lo que equivale a un 79.24%, el avg ponderado es de **77%** la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. verdaderas (No paradas) el algoritmo clasificó bien 402, con un score de 74% el avg ponderado es de **77%**

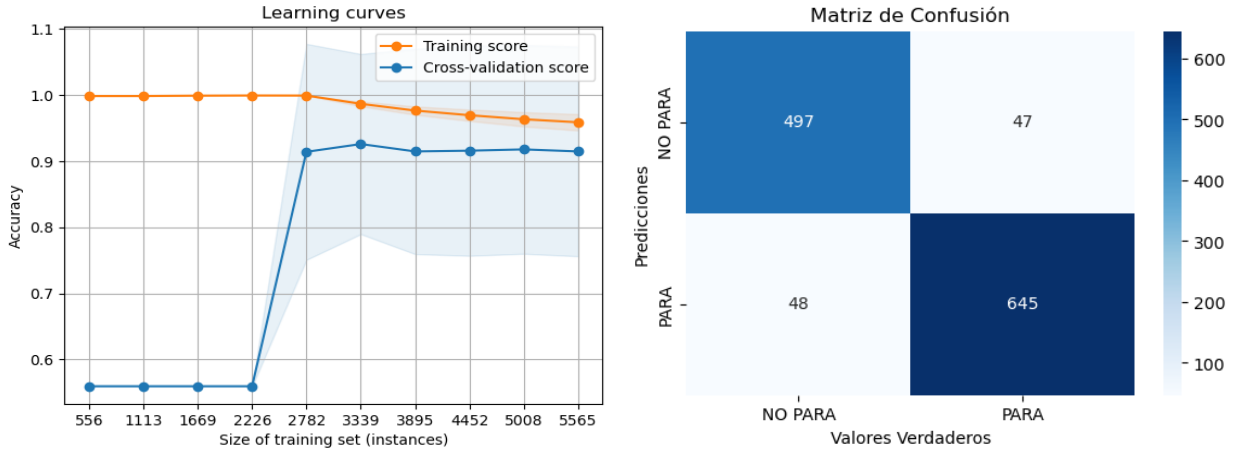
F1: Es un balance entre precisión y recall, el cual da un avg ponderado de **77%**

Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (948), sobre el total de predicciones (1237), el cual da una puntuación de 0.7664. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc (entradas), “el modelo acertó si una falla se convertirá en parada o no en un **77%** de todas las predicciones”.

Decision Tree

Figura 58

Decision Tree



Fuente. Elaboración propia

Tabla 13

Resultados Decision Tree

	Precision	Recall	F1-score	Support
No Para	0.91	0.91	0.91	545
Para	0.93	0.93	0.93	692
accuracy			0.92	1237
macro avg	0.92	0.92	0.92	1237
weighted avg	0.92	0.92	0.92	1237

Nota. Esta tabla presenta los resultados obtenidos mediante el árbol de decisión. *Fuente.* Elaboración propia.

Precisión: De un total de 693 predicciones de Parada, el algoritmo clasificó correctamente 645 lo que equivale a un 93.07%, el avg ponderado es de 92% la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 692 Obs. verdaderas (Paradas) el algoritmo clasificó bien 645, con un score de 93.2%, el avg ponderado es de 92%

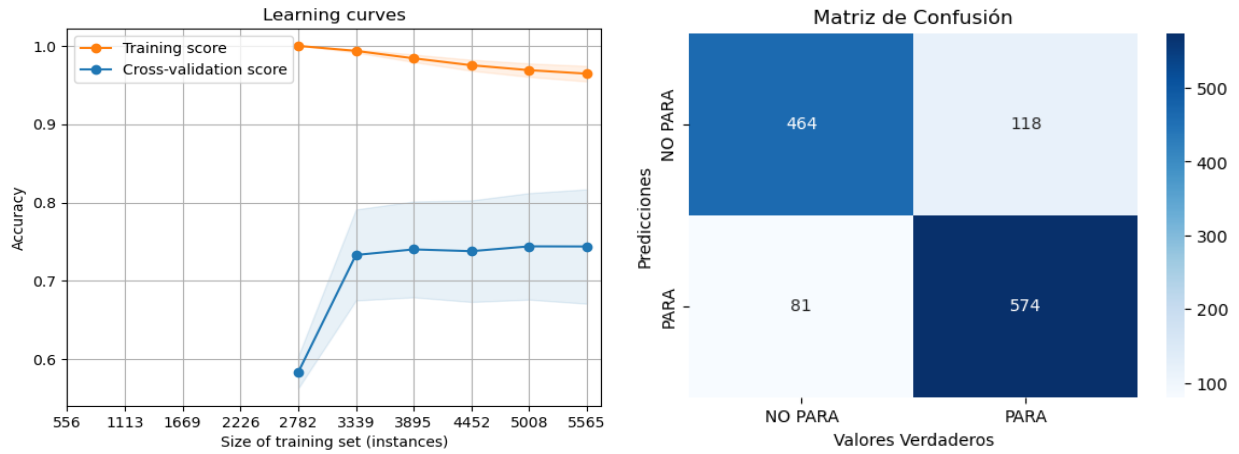
F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 92**%**

Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (1142), sobre el total de predicciones (1237), el cual da una puntuación de 0.9232. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc. (entradas), "el modelo acertó si una falla se convertirá en parada o no en un **92%** de todas las predicciones".

Hard Voting

Figura 59

Hard Voting



Fuente. Elaboración propia

Tabla 14

Resultados Hard Voting

	Precision	Recall	F1-score	Support
No Para	0.80	0.85	0.82	545
Para	0.88	0.83	0.83	692

accuracy			0.84	1237
macro avg	0.84	0.84	0.84	1237
weighted avg	0.84	0.84	0.84	1237

Nota. Esta tabla muestra los resultados obtenidos mediante el hard voting de modelos de clasificación. *Fuente.* Elaboración propia.

Precisión: De un total de 655 predicciones de Parada, el algoritmo clasificó correctamente 574 lo que equivale a un 87.63%, el avg ponderado es de 84% la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. verdaderas (No Paradas) el algoritmo clasificó bien 464, con un score de 85.14%, el avg ponderado es de 84%

F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 84%

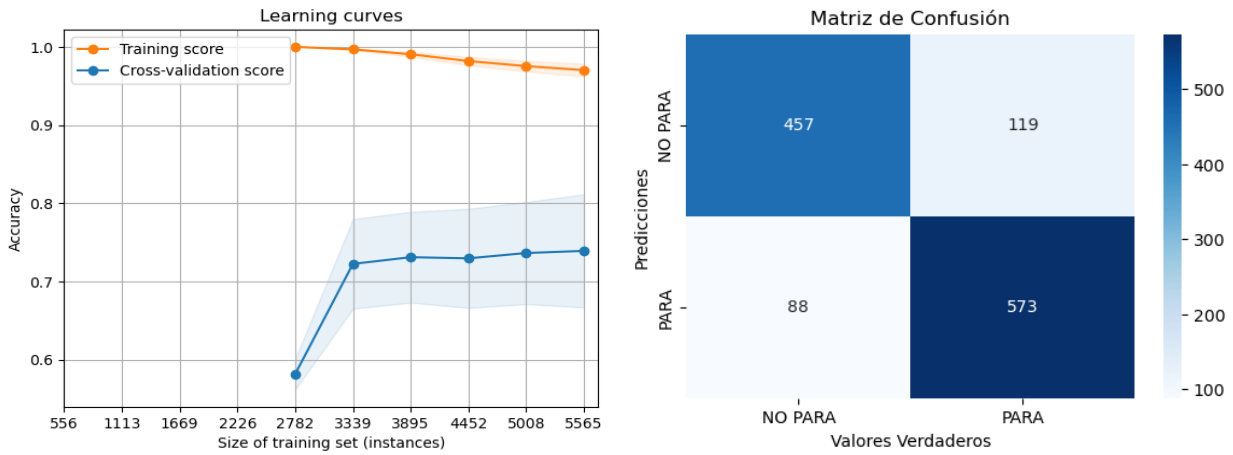
Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (1038), sobre el total de predicciones (1237), el cual da una puntuación de 0.8391. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc. (entradas), “el modelo acertó si una falla se convertirá en parada o no en un **84%** de todas las predicciones”.

Clasificadores Base: Logistic Regression, KNeighborsNearest, Decision Tree

Soft Voting

Figura 60

Soft Voting



Fuente. Elaboración propia

Tabla 15

Resultados Soft Voting

	Precision	Recall	F1-score	Support
No Para	0.79	0.84	0.82	545
Para	0.87	0.83	0.85	692
accuracy			0.83	1237
macro avg	0.83	0.83	0.83	1237
weighted avg	0.83	0.83	0.83	1237

Nota. Esta tabla presenta los resultados obtenidos mediante el soft voting de modelos de clasificación. *Fuente.* Elaboración propia.

Precisión: De un total de 661 predicciones de Parada, el algoritmo clasificó correctamente 573 lo que equivale a un 86.68%, el avg ponderado es de 83% la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. verdaderas (No Paradas) el algoritmo clasificó bien 457, con un score de 83.85%, el avg ponderado es de 83%

F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 83%

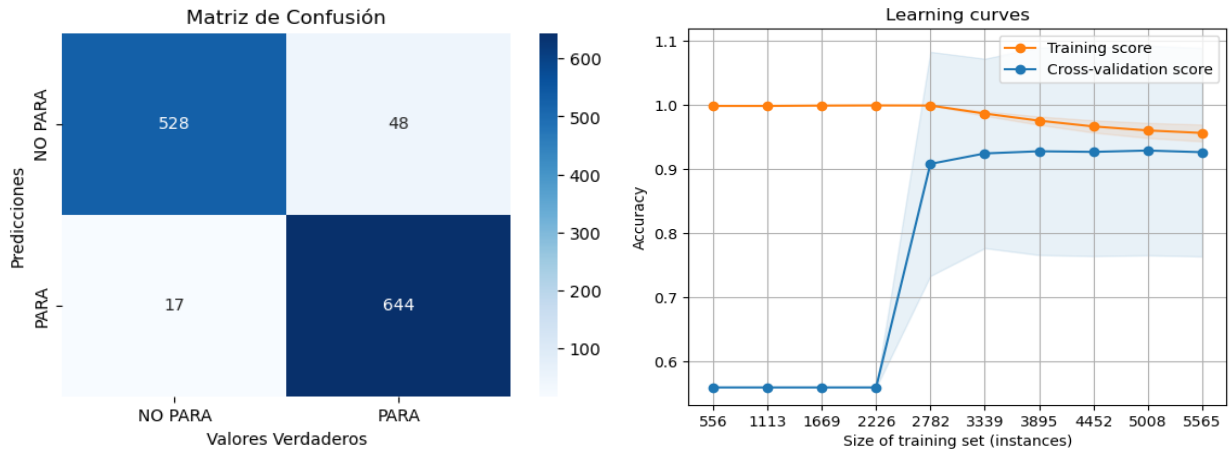
Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (1030), sobre el total de predicciones (1237), el cual da una puntuación de 0.8326. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc. (entradas), "el modelo acertó si una falla se convertirá en parada o no en un **83%** de todas las predicciones".

Clasificadores Base: Logistic Regression, KNeighborsNearest Decision Tree

Bagging

Figura 61

Bagging



Fuente. Elaboración propia

Tabla 16

Resultados Bagging

	Precision	Recall	F1-score	Support
No Para	0.92	0.97	0.94	545

Para	0.97	0.93	0.95	692
accuracy			0.95	1237
macro avg	0.945	0.95	0.95	1237
weighted avg	0.95	0.95	0.95	1237

Nota. Esta tabla muestra los resultados obtenidos mediante la técnica de Bagging en modelos de clasificación. *Fuente.* Elaboración propia.

Precisión: De un total de 661 predicciones de Parada, el algoritmo clasificó correctamente 644 lo que equivale a un 97.42%, el avg ponderado es de 95% la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. verdaderas (No Paradas) el algoritmo clasificó bien 528, con un score de 96%, el avg ponderado es de 95%

F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 95%

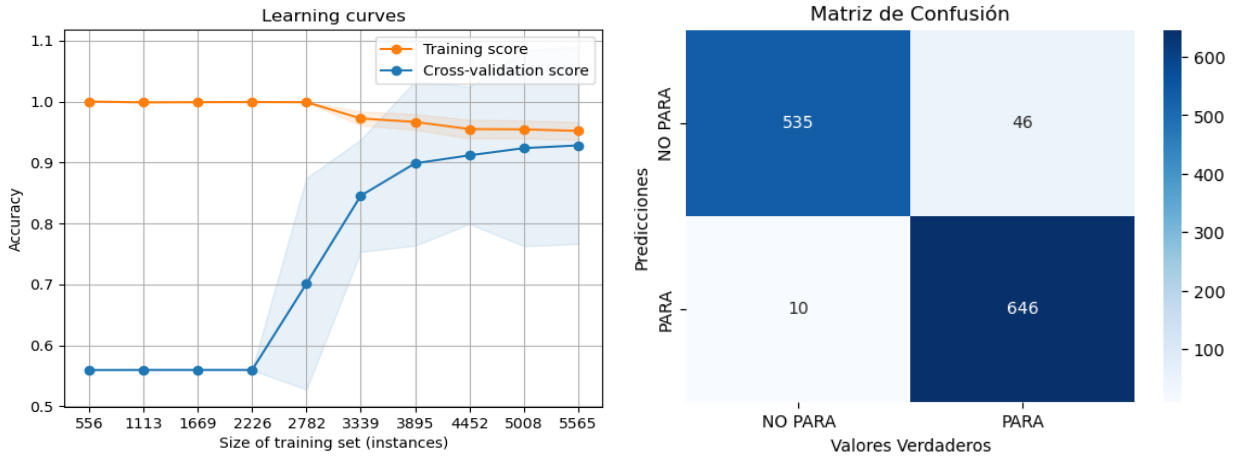
Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (1172), sobre el total de predicciones (1237), el cual da una puntuación de 0.9474. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc. (entradas), “el modelo acertó si una falla se convertirá en parada o no en un **95%** de todas las predicciones”.

Clasificadores Base: Logistic Regression, KNeighborsNearest, Decision Tree

Random Forest

Figura 62

Random Forest



Fuente. Elaboración propia

Tabla 17

Resultados Random Forest

	Precision	Recall	F1-score	Support
No Para	0.92	0.98	0.95	545
Para	0.98	0.93	0.96	692
accuracy			0.96	1237
macro avg	0.95	0.96	0.95	1237
weighted avg	0.96	0.95	0.95	1237

Nota. Esta tabla presenta los resultados obtenidos mediante la técnica de Random Forest en modelos de clasificación. *Fuente.* Elaboración propia.

Precisión: De un total de 656 predicciones de Parada, el algoritmo clasificó correctamente 646 lo que equivale a un 98.47%, el avg ponderado es de 96% la cual ajusta la puntuación teniendo en cuenta la proporción de clases en el test, como la proporción es similar, este valor es cercano al promedio (macro avg) entre las dos clases.

Recall: De un total 545 Obs. verdaderas (No Paradas) el algoritmo clasificó bien 535, con un score de 98168%, el avg ponderado es de 95%

F1: Es un balance entre precisión y recall, el cual da un avg ponderado de 95%

Accuracy: Es la suma de los valores bien clasificados (+/-) es decir la diagonal de la matriz de confusión (1181), sobre el total de predicciones (1237), el cual da una puntuación de 0.955. considerando características como unidad funcional, prioridad, repercusión, horas de operación, TMEF…etc. (entradas), "el modelo acertó si una falla se convertirá en parada o no en un **96%** de todas las predicciones".

Tabla 18

Comparación de rendimiento de los modelos de clasificación

Modelo	LR	KNN	DT	HARDV.	SOFT V.	BAGGING	RF
Accuracy	0,725	0,766	0,923	0,839	0,833	0,947	0,96

Nota. Esta tabla compara el rendimiento de diferentes modelos de clasificación utilizados en el estudio. *Fuente.* Elaboración propia.

Ajuste de hiperparametros

De los todos los algoritmos de clasificación anteriores el de mayor score fue el clasificador **Random Forest** (RF), utilizando la función **GridSearchCV** de Python el cual emplea la técnica de Cross Validation se encuentran los mejores hiperparametros para este algoritmo.

Tabla 19

Ajuste de hiperparametros

RF	Hiperparametros Iniciales	Hiperparametros Óptimos	Accuracy Anterior	Accuracy Ajustado
ParaDA	RandomForestRegressor	RandomForestRegressor	0,949	0,96

max_depth=16	max_depth=21
n_estimators=50	min_samples_leaf=
n_jobs=-1	n_estimators=35
random_state=55	n_jobs=-1
	random_state=42

Nota. Esta tabla muestra el proceso de ajuste de hiperparámetros realizado para optimizar los modelos de clasificación. *Fuente.* Elaboración propia.

En el intento de realizar una evaluación de los modelos de clasificación a través de las curvas de aprendizaje, matriz de confusión y métricas de evaluación. Se puede divisar claramente 3 grupos de modelos:

Tabla 20

Grupos de modelos

Gr	Modelos
1	RF- BAGGING - DT
2	HARD V. - SOFT V.
3	LR - KNN

Nota. Esta tabla muestra los grupos de modelos clasificatorios analizados en el estudio. *Fuente.* Elaboración propia.

El grupo 1 obtuvo puntuaciones altas por encima del 90%, en las curvas de aprendizaje se evidencia como el conjunto de validación es mucho más preciso al aumentar el conjunto de datos, además la brecha de generalización entre los datos de entrenamiento y test se va reduciendo y alcanzando una estabilidad más allá de los 6184 datos, por lo que es suficiente la cantidad de datos para los modelos. El modelo RF es seleccionado para realizar ajuste de hiperparametros evidenciando mejora en el rendimiento. Además, es uno de los mejores modelos junto al modelo Bagging y Decision Tree para la predicción de paradas**.**

El grupo 2 obtuvo puntuaciones por encima del 80%, son buenos modelos, aunque se observa Overfiting, existe distancia ente la precisión del conjunto de entrenamiento y la validación, pues a pesar de que esta última incrementa, aún está lejos de la precisión del entrenamiento y de la precisión deseada, es posible que agregando más datos o disminuyendo las características mejore el rendimiento. Tener en cuenta de igual manera el análisis y ajuste de hiperparametros.

El grupo 3 obtuvo puntuaciones por encima del 70%, se observa un nivel de Overfiting mayor, existe una gran distancia ente la precisión del entrenamiento y la validación, pues a pesar de que esta última incrementa aún está muy lejos de la precisión del entrenamiento y de la precisión deseada, es posible que agregando más datos mejore el rendimiento o disminuyendo las características. Tener en cuenta el análisis y ajuste de hiperparametros y técnicas de regularización para penalizar el sobreajuste.

Conclusiones

El Mantenimiento 4.0, también conocido como mantenimiento predictivo o inteligente, se ha establecido como un enfoque crucial para mejorar la eficiencia en la industria del petróleo y gas. Investigaciones recientes han demostrado que la integración de tecnologías avanzadas, como el Internet de las cosas (IoT), la inteligencia artificial (IA) y el análisis de datos, puede optimizar las operaciones de mantenimiento.

Un estudio publicado en la revista IEEE Access (López-Morales et al., 2020) examinó la implementación del Mantenimiento 4.0 en la industria petrolera y concluyó que la adopción de estos enfoques puede reducir los costos operativos hasta en un 20% y aumentar la disponibilidad de los activos en un 15%.

Además, investigaciones como la realizada por García-Prada et al. (2019) en la revista Journal of Petroleum Science and Engineering han demostrado que el Mantenimiento 4.0 puede mejorar significativamente la eficiencia energética y reducir las emisiones de gases de efecto invernadero en las operaciones de extracción y refinación de petróleo y gas.

Los modelos predictivos, como la Regresión Lineal, Random Forest, Gradient Boosting y SVR, han emergido como herramientas poderosas para predecir fallos en equipos industriales en la industria del petróleo y gas. Investigaciones han demostrado su eficacia en la identificación temprana de problemas y la prevención de paradas no planificadas

Un artículo publicado en "Reliability Engineering & System Safety" por Zhu et al. (2018) examinó la aplicación de modelos predictivos para la monitorización y mantenimiento de activos en la industria del petróleo y gas offshore. Los resultados mostraron que la combinación de diferentes algoritmos de aprendizaje automático mejoró la precisión de las predicciones de fallos hasta en un 25%.

Asimismo, una revisión sistemática realizada por Wang et al. (2020) en "Energy Reports" destacó la efectividad de los modelos predictivos en la reducción de costos de mantenimiento y la mejora de la seguridad en las operaciones de producción de petróleo y gas.

La aplicación de modelos predictivos en la industria del petróleo y gas no solo permite predecir fallos, sino que también revela patrones ocultos en los datos que pueden ser críticos para la toma de decisiones informada

Además, un estudio en "Journal of Loss Prevention in the Process Industries" por Li et al. (2019) demostró cómo los modelos predictivos pueden detectar anomalías en el funcionamiento de plantas de procesamiento de gas natural, lo que permite tomar medidas correctivas antes de que ocurran incidentes graves. La implementación exitosa de modelos predictivos en la industria del petróleo y gas tiene un impacto directo en la eficiencia operativa y la seguridad de las instalaciones.

Un artículo en "Safety Science" por Ding et al. (2021) evaluó el impacto de la aplicación de modelos predictivos en la prevención de accidentes y concluyó que estos enfoques pueden reducir significativamente la probabilidad de fallas catastróficas en plataformas offshore. Además, una investigación realizada por Hu et al. (2020) en "Process Safety and Environmental Protection" destacó cómo la implementación de modelos predictivos puede optimizar los programas de mantenimiento, reduciendo el riesgo de incidentes y minimizando el impacto ambiental en la industria del petróleo y gas.

En el estudio realizado sobre el desarrollo de modelos predictivos para el mantenimiento de activos en la industria Oíl & Gas, se han obtenido importantes conclusiones. En primer lugar, se destaca la eficacia de la Regresión Lineal en predecir el número de fallas en equipos, demostrando un alto coeficiente de determinación (R^2) que respalda su capacidad predictiva. Este

hallazgo proporciona un nuevo conocimiento sobre la utilidad de este modelo en la planificación proactiva del mantenimiento, contribuyendo significativamente a la gestión eficiente de posibles fallas y a la reducción de tiempos de inactividad no planificados.

Por otro lado, el modelo Random Forest ha demostrado ser altamente preciso en la estimación de la duración de las paradas de equipos, con un Error Cuadrático Medio (MSE) muy bajo que resalta su capacidad para capturar patrones complejos en los datos. Esta conclusión resalta la importancia de utilizar modelos avanzados como el Random Forest para mejorar la planificación del mantenimiento, reducir costos operativos y optimizar la gestión de recursos en la industria Oíl & Gas.

Además, la identificación de variables predictoras clave, como Prioridad, Repercusión, Horas Operativa, TMEF, Duración de parada y Unidad Funcional, ha sido fundamental para mejorar la precisión de los modelos predictivos en la detección de paradas en equipos. Este descubrimiento resalta la importancia de considerar múltiples variables en los modelos predictivos para aumentar su capacidad de generalización y mejorar la gestión de activos en la industria, lo que puede conducir a una mayor eficiencia operativa y una reducción de costos a largo plazo.

Recomendaciones

Para futuras investigaciones, se sugiere explorar más a fondo los siguientes aspectos:

Investigar nuevas técnicas de mantenimiento predictivo y su aplicabilidad en la industria del petróleo y gas: Además de evaluar técnicas existentes, se podría investigar el desarrollo de nuevas metodologías específicamente diseñadas para abordar los desafíos únicos de la industria Oíl & Gas, como la corrosión, la erosión y las condiciones ambientales extremas.

Evaluar el impacto de la inteligencia artificial y el aprendizaje automático en la optimización de la gestión de activos: Además de evaluar el impacto en términos de eficiencia operativa y rentabilidad, sería útil investigar cómo la inteligencia artificial y el aprendizaje automático pueden mejorar la seguridad en la industria Oíl & Gas, identificando y previniendo riesgos potenciales.

Analizar cómo la implementación de tecnologías emergentes, como el Internet de las cosas (IoT), puede mejorar la eficiencia operativa: Además de analizar el impacto en la eficiencia operativa, sería valioso explorar cómo la implementación de IoT puede contribuir a la sostenibilidad ambiental y al cumplimiento de regulaciones en la industria Oíl & Gas.

Se recomienda ampliar la información sobre los siguientes temas:

Detalles específicos sobre la implementación de modelos predictivos en casos reales dentro de la industria: Además de proporcionar información sobre la implementación, sería útil incluir análisis de costos y beneficios, así como lecciones aprendidas de casos reales para orientar a otras empresas en la implementación de modelos predictivos.

Estudios de casos que demuestren la efectividad de las estrategias de mantenimiento predictivo en diferentes contextos: Además de presentar casos de éxito, sería valioso incluir

estudios de casos que aborden desafíos específicos, como la gestión de activos en entornos offshore o la optimización de la producción en campos maduros.

Para mejorar el trabajo realizado, se sugiere:

Refinar los modelos predictivos mediante la incorporación de más datos históricos y la optimización de hiperparámetros: Además de incorporar más datos históricos, se podría investigar la integración de datos en tiempo real para mejorar la precisión de los modelos. Además, se podría explorar la aplicación de técnicas de optimización de hiperparámetros más avanzadas para mejorar el rendimiento de los modelos.

Realizar análisis de sensibilidad para comprender mejor las variables clave que afectan las predicciones: Además de realizar análisis de sensibilidad, sería útil investigar la identificación de variables latentes y la evaluación de su impacto en las predicciones. Esto podría ayudar a mejorar la interpretación de los resultados y la toma de decisiones en la gestión de activos en la industria Oíl & Gas..

Referencias Bibliográficas

Casana-Medel, J. C., de la Rosa-Andino, C. A. A., Macias-Socarras, C. I., Morales-Tamayo, C. Y., Zamora-Hernandez, Y. K., & Aguilera-Corrales, Y. (2021). Maintenance Based on World Class Indicators in Bayamo Dairy Factory/El mantenimiento a partir de los indicadores de clase mundial en la fábrica Lácteos Bayamo. Revista Ciencias Técnicas Agropecuarias, 30(3), 72.

De Simone, L., Caputo, E., Cinque, M., Galli, A., Moscato, V., Russo, S., Cesaro, G., Criscuolo, V., & Giannini, G. (2023). LSTM-based failure prediction for railway rolling stock equipment. Expert Systems with Applications, 222(119767), 119767. https://doi.org/10.1016/j.eswa.2023.119767

Ding, Y., Li, W., Yu, L., Huang, W., & Han, S. (2021). Predictive risk assessment of offshore oil and gas accidents based on Machine Learning. *Safety Science*, 133, 105011. DOI: 10.1016/j.ssci.2020.105011

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 29(5), 1189-1232. : Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. : Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Hu, X., Zhou, L., Wang, H., & Zhao, Y. (2020). Predictive maintenance optimization based on deep learning in petrochemical industry. *Process Safety and Environmental Protection*, 139, 164-174. DOI: 10.1016/j.psep.2020.03.028

IBM. (2018, August 1). Infografía: Ventajas del Mantenimiento Predictivo. Decide. https://decidesoluciones.es/infografia-ventajas-del-mantenimiento-predictivo/

Li, Y., Khanna, R., & Zhang, Y. (2019). A Machine Learning approach for anomaly detection in natural gas processing. *Journal of Loss Prevention in the Process Industries*, 58, 104069. DOI: 10.1016/j.jlp.2019.104069

López-Morales, J., García-Prada, J. C., & Pérez-Álvarez, J. (2020). Implementation of Industry 4.0 in the Oil Industry: A Case Study. *IEEE Access*, 8, 117146-117156. DOI: 10.1109/ACCESS.2020.3005537

Meddaoui, A., Hain, M., & Hachmoud, A. (2023). The benefits of predictive maintenance in manufacturing excellence: a case study to establish reliable methods for predicting failures. The International Journal of Advanced Manufacturing Technology, 128(7–8), 3685–3690. https://doi.org/10.1007/s00170-023-12086-6

Ouda, E., Maalouf, M., & Sleptchenko, A. (2021). Machine Learning and optimization for predictive maintenance based on predicting failure in the next five days. Proceedings of the 10th International Conference on Operations Research and Enterprise Systems.

Pinto, R., & Cerquitelli, T. (2019). Robot fault detection and remaining life estimation for predictive maintenance. Procedia Computer Science, 151, 709–716. https://doi.org/10.1016/j.procs.2019.04.094

Romero Gelvez, J. I., Stiven, B., & Quintero, R. (s. f.). Aplicación de Machine Learning en el mantenimiento predictivo industrial con herramientas de código abierto. Edu.co. Recuperado 22 de noviembre de 2023, de https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/10108/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y

Surantha, N., & Gozali, I. D. (2023). Evaluation of the improved Extreme Learning Machine for machine failure multiclass classification. Electronics, 12(16), 3501. https://doi.org/10.3390/electronics12163501

Tao, L., Yu, X., Gao, D., & Ma, L. (2019). Predictive maintenance for rotary drilling equipment based on Machine Learning techniques. *Computers & Chemical Engineering*, 129, 106505. DOI: 10.1016/j.compchemeng.2019.106505

Vilema Lara, P. H., García Mora, F. A., & Gallegos Londoño, C. M. (2022). Aprendizaje de máquina para mantenimiento predictivo: un problema de clasificación binaria. ConcienciaDigital, 5(2.1), 45–68. https://doi.org/10.33262/concienciadigital.v5i2.1.2150

Villachica Pérez, Y. N., Ordoñez Cuthbert, D. K., & Mckensy Sambola, D. (2022). Modelo predictivo basado en Machine Learning dirigido a PYMES de venta, caso de estudio Bluefields. Ciencia e interculturalidad, 30(01), 139–146. https://doi.org/10.5377/rci.v30i01.14267

Vista de El enfoque de aprendizaje de máquina para la gestión del mantenimiento industrial. (s. f.). Edu.cu. Recuperado 22 de noviembre de 2023, de https://rus.ucf.edu.cu/index.php/rus/article/view/3819/3747

Wang, H., Chen, Y., Zhao, Y., Wang, Z., & Yang, Y. (2020). Predictive Maintenance for Oil and Gas Industry Based on Machine Learning: A Review. *Energy Reports*, 6, 588-601. DOI: 10.1016/j.egyr.2020.04.004

Zhu, X., Zio, E., & Zhou, D. (2018). A Machine Learning-based approach for monitoring and maintenance of offshore oil and gas facilities. *Reliability Engineering & System Safety*, 178, 47-56. DOI: 10.1016/j.ress.2018.05.020

Lista de Anexos

data_analisis.ipynb

Contiene información sobre el análisis exploratorio realizado en Python.

modelos_analisis.ipynb

Contiene información sobre el diseño de modelos predictivos para el numero de fallas y duración de parada.