Estudio de modelos de predicción, para la gestión de tiempos correctivos en una empresa Industrial

04/2024

Universidad Internacional de Valencia

Titulación:	Alumno/a:	Convocatoria:
Big Data y Ciencia de Datos	Torres Lombana, Jose Humberto.	Segunda
Curso académico	D.N.I: 9430885	
2023 – 2024	Director/a de TFM: Ortíz Vargas, Walter Andrés	

De:
Planeta Formación y Universidades

Universidad Internacional de Valencia

Índice

Universidad Internacional de Valencia

Índice de cuadros

Índice de figuras

Lista de ecuaciones

Resumen

Este trabajo se centra en la aplicación de técnicas estadísticas como modelos de regresión y algoritmos de aprendizaje automático, con el objetivo de encontrar el mejor modelo y predecir el tiempo requerido para realizar mantenimientos correctivos en una empresa industrial. Generalmente se usan métodos tradicionales, como la observación de patrones históricos, juicio de expertos, análisis de promedios o valores predeterminados donde el margen de error es amplio, lo que afecta la productividad de la empresa. Por tanto, se analizan los históricos de mantenimientos correctivos, con registros de tiempos, fechas, tipos de mantenimiento y otras variables del proceso para llegar a predecir el tiempo que una empresa necesita disponer para atender estos trabajos mencionados, en una frecuencia de tiempo que puede ser mensual, semanal o diaria.

Se entrenan modelos de regresión múltiple, modelos lineales generalizados y algoritmos de aprendizaje automático como Randon Forest, Árbol de Decisión Regresión y Maquina de Vectores Soporte, se hace en tres frecuencias semanal, mensual y diario. Para la evaluación de resultados se utilizan métricas como el error medio absoluto (MAE), error cuadrático medio (MSE) y raíz del error cuadrático medio (RMSE). Finalmente se presentan los resultados donde se elige el mejor modelo, se proponen casos futuros y se dan las recomendaciones para mejorar el ajuste y llegar a predicciones más acertadas.

Palabras clave: Predicción, Mantenimiento, Industria, Técnicas Multivariadas, Evaluación de modelos.

abstract

This work focuses on the application of statistical techniques such as regression models and machine learning algorithms, with the objective of finding the best model and predicting the time required to perform corrective maintenance in an industrial company. Traditional methods are generally used, such as the observation of historical patterns, expert judgment, analysis of averages or predetermined values where the margin of error is wide, which affects the productivity of the company. Therefore, historical records of corrective maintenance are analyzed, with records of times, dates, types of maintenance and other variables of the process to predict the time that a company needs to have available to attend these works mentioned, in a frequency of time that can be monthly, weekly or daily.

Multiple regression models, generalized linear models and automatic learning algorithms such as Randon Forest, Regression Decision Tree and Support Vector Machine are trained in three frequencies: weekly, monthly and daily. For the evaluation of results, metrics such as mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) are used. Finally, the results are presented where the best model is chosen, future cases are proposed and recommendations are given to improve the fit and arrive at more accurate predictions.

Translated with DeepL.com (free version)

Keywords: Prediction, Maintenance, Industry, Multivariate Techniques, Model Evaluation.

1. Introducción

El mantenimiento, frase actividad o acción que día a día es nombrada en todos los ámbitos sociales y todos una u otra vez se han implicado en temas de mantenimiento, cuando el carro, la motocicleta u otro objeto presenta una falla o talvez se requiere un cambio aceite o un ajuste general se incurre en temas de mantenimiento, si no es el carro de igual forma aplica para una propiedad o un electrodoméstico que también necesita ser mantenido para asegurar su vida útil, se define como un conjunto de actividades planificadas y sistemáticas que se realizan para asegurar que los equipos, maquinarias o instalaciones operen de manera óptima y eficiente durante su vida útil. Esto implica tanto acciones preventivas para evitar posibles fallas, como intervenciones correctivas cuando ocurren averías.

El hacer mantenimiento involucra costos, y estos no se limitan únicamente a las reparaciones de equipos averiados, incluyen una amplia gama de actividades, como mantenimiento preventivo, inspecciones, repuestos, mano de obra (tiempos) y paradas no programadas. Según estudios de García (2017), estos costos pueden representar hasta el 40 % del presupuesto operativo de una empresa, lo que resalta su importancia en la gestión financiera y operativa. Ahora el tema de costos es un proceso amplio, pero se destaca la mano de obra o tiempos de mantenimiento, que pueden ser para trabajo correctivo o preventivo, de acuerdo a estudios de Martínez (2020), las organizaciones que implementan estrategias efectivas de mantenimiento logran reducir los tiempos de inactividad, aumentar la vida útil de los activos y mejorar la calidad de sus productos o servicios.

La gestión de costos está relacionada ampliamente con la gestión de tiempo o productivad en el mantenimiento, una gestión eficiente del tiempo en entornos empresariales e industriales es fundamental para optimizar procesos, reducir costos y garantizar entregas puntuales de productos y servicios. Esta premisa subrayada en el PMBOK (Project Management Body of Knowledge), enfatiza la importancia de una administración efectiva del tiempo en proyectos, y propone diferentes técnicas para hacer los cálculos respectivos como juicio de expertos, análisis de la ruta crítica, la estimación paramétrica que implica temas de estadística y la estimación por tres valores donde utilizan una distribución beta para los cálculos de tiempo. Sin embargo, frente a la creciente complejidad de maquinarias y sistemas industriales, los métodos tradicionales basados en la experiencia ya no son suficientes, se necesita abordar técnicas diferentes para calcular los tiempos de mantenimiento basados en datos históricos y estadísticas.

En este contexto, el presente estudio, ubicado en campo del Big Data y la Ciencia de Datos, se orienta hacia la innovación en el mantenimiento industrial mediante la aplicación de técnicas de estadística multivariante y técnicas de Maching learning para contribuir con propuestas de modelos de regresión que sean capaces de predecir el tiempo que se requiere para realizar mantenimientos correctivos en el sector industrial, todo esto analizando datos históricos de fallas con sus respectivos tiempos productivos e improductivos para lograr un acercamiento al número de horas que realmente se necesitan para atender las fallas correctivas en periodo de tiempo que puede ser diario, semanal, mensual o anual. La importancia de transitar hacia métodos cuantitativos avanzados para la predicción y mejora de la gestión del tiempo en el mantenimiento industrial se fundamenta en investigaciones previas. Rojas (1975) en la década de los 70, subrayó la necesidad de contar con sistemas fiables para abordar la complejidad de las maquinarias industriales, destacando la importancia de modelos predictivos para anticipar el funcionamiento futuro de las máquinas. En una línea similar, Morocho Pucuna et al. (2009)

destacaron la relevancia de aplicar modelos predictivos para mejorar la eficiencia y fiabilidad de los procesos industriales. Johnson y Wichern (2007) aportaron al debate, explicando cómo las técnicas de estadística multivariante permiten analizar simultáneamente múltiples variables interdependientes, ofreciendo una comprensión más profunda y precisa que los enfoques univariados tradicionales.

La aplicación de estadística descriptiva para limpieza y visualización de datos hacen parte de los objetivos del proyecto, ayudan a que los datos sean apropiados para los diferentes cálculos propuestos, la aplicación de modelos y algoritmos de Maching learning como árboles de decisión, ramdon Forest y máquinas de vectores SVM serán trascendentes porque al igual que los modelos estadísticos tradicionales como la regresión múltiple y los modelos lineales generalizados permiten trabajar temas de regresión para predecir datos. Con enfoques diferentes, pero con el mismo objetivo se abordan estas técnicas para finalmente identificar por medio de las diferentes métricas de evaluación que algoritmo o qué modelo se adapta mejor a los datos y así encontrar la cantidad de horas con el menor error para realizar mantenimiento correctivo, buscando proponer estrategias y recomendaciones específicas basadas en los resultados de los diferentes análisis propuestos.

En el primer capítulo se detalla el estado del arte y marco teórico que permita referenciar y contextualizar temas de mantenimiento industrial y modelos de regresión clásicos como algoritmos de Maching learning, en el segundo capítulo se detallan el tratamiento de los datos, las técnicas realizadas para el análisis descriptivo y la limpieza de datos, planteamiento de los modelos, comparación de resultados y métricas de evaluación. En el apartado final se realizan las conclusiones al estudio, con lo referente a los principales hallazgos y las limitaciones del proyecto y se proponen estudios futuros para finalizar el trabajo.

2. Objetivos

Contribuir a la mejora de asignación de tiempos correctivos y de emergencia en el proceso de mantenimiento en una empresa del sector industrial.

2.1. Objetivos Específicos:

- Realizar un análisis descriptivo y de limpieza para la preparación de los datos históricos de mantenimiento correctivo.
- Utilizar técnicas de estadística multivariante como modelos de regresión lineal múltiple y modelos lineales generalizados para la predicción de tiempos de Mantenimiento correctivo.
- Utilizar técnicas de Maching Learning en especial los algoritmos de Arboles de decisión, Random Forest y Máquinas de Vectores de Soporte (SVM), para la predicción de tiempos de mantenimiento correctivo.
- Evaluar y comparar los diversos modelos de regresión planteados para predecir tiempos de mantenimiento correctivo, con el fin de elegir el más eficiente para estos casos.
- Proponer estrategias y recomendaciones específicas basadas en los resultados del análisis predictivo para optimizar los procesos de reparación y minimizar los tiempos de inactividad.

3. Estado del Arte y Marco teórico

3.1. El mantenimiento industrial

El mantenimiento industrial se define como un conjunto de técnicas destinadas a conservar equipos e instalaciones en servicio durante el mayor tiempo posible buscando siempre la más alta disponibilidad, confiabilidad y con el máximo rendimiento Según García Garrido (2010), A lo largo de los años, diversas metodologías y técnicas se han desarrollado para mejorar la planificación y ejecución de actividades de mantenimiento industrial.

García en su libro "Organización y gestión Integral de Mantenimiento", también narra que El Mantenimiento nace durante la primera revolución Industrial, periodo que se inició en la segunda mitad del siglo XVIII en Gran Bretaña, unas décadas después se extendió a gran parte de Europa occidental y América Anglosajona y finalmente concluyó entre 1820 y 1840. En los inicios eran los propios operarios quienes realizaban este tipo de tareas de mantenimiento, no había personal dedicado única y exclusivamente a esta actividad. Pero con la aparición de maquinaria más compleja se vio la necesidad de crear un departamento dedicado al mantenimiento dentro de las fábricas, se narra que en los principios las tareas de mantenimiento inicialmente eran correctivas, dedicaban todo el esfuerzo a solucionar fallas presentadas, pero no a prevenirlas. A inicios de la primera guerra mundial y ante todo de la segunda, García Garrido (2010) manifiesta que aparece la palabra Fiabilidad, que se define como la capacidad de un sistema, equipo o máquina para realizar una función o tarea específica durante un período de tiempo determinado y bajo condiciones establecidas.

Según Smith (2018), La fiabilidad en el mantenimiento industrial implica una serie de acciones y estrategias, como el monitoreo continuo del desempeño de los equipos, la implementación de programas de mantenimiento preventivo y predictivo, la gestión eficiente de repuestos y piezas de repuesto, así como la formación adecuada del personal encargado del mantenimiento.

Para esas épocas con estos términos ya definidos y con conceptos más claros del tema, los departamentos de mantenimiento inician el proceso de no solo solucionar las fallas de sus equipos si no actuar para que no se produzcan o prevenirlas García Garrido (2010), de acuerdo a lo anterior la historia narra que nace una nueva figura en los departamentos de mantenimiento y es un personal destinado a estudiar diferentes tareas de mantenimiento para evitar las fallas en sus máquinas, es cuando el autor manifiesta que nacen nuevos conceptos de mantenimiento como: Mantenimiento preventivo, mantenimiento predictivo, mantenimiento proactivo, gestión de mantenimiento asistido por ordenador, mantenimiento basado en fiabilidad RCM entre otros conceptos.

Para abarcar un poco más de historia en estos temas, Carcel-Carrasco (2016) Es su artículo de investigación manifiesta que en los años 1945 se crean y formalizan técnicas para conocer e identificar la probabilidad de fallas en los diferentes componentes, en los años 60 se comienza con la aplicación de las técnicas de fiabilidad permitiendo el cálculo de costos de los fallos y la rentabilidad del mantenimiento. Adicionalmente se inicia el análisis y estudio de las causas y efectos de las incidencias de los equipos industriales objeto básico del mantenimiento, se destaca para esas épocas la necesidad de tener estadísticas e históricas de averías para el análisis y planificación del mantenimiento. De acuerdo a estos descubrimientos y la necesidad de programa gar el mantenimiento aparece la publicación de Darnell y Bert en 1978 donde ellos publican una contribución en la que abogan por el mantenimiento programado como medio de aumentar la productividad y Christer, en 1981 y Boland y Proscan en 1982, insisten

en la misma línea, destacando la incidencia sobre la productividad de una actitud activa y programada en mantenimiento, en contraposición con actitudes pasivas e incontroladas.

Para estas épocas los conceptos de mantenimiento están bien definidos, la norma UNE-EN 13306 que establece principios generales para la creación y el uso de sistemas de clasificación y codificación de objetos técnicos. Proporciona directrices sobre cómo desarrollar sistemas de clasificación y cómo aplicar códigos a los objetos técnicos para facilitar su identificación y gestión. Ellos proponen la siguiente clasificación del mantenimiento. (ver figura 1).

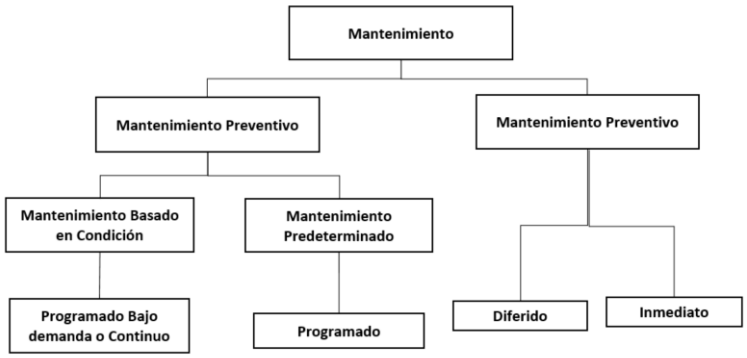

Figura 1: Clasificación general del mantenimiento según UNE-EN 13306. Fuente: UNE EN 13306

Por otra parte, la norma ISO -14224, que es una norma internacional que establece los principios y prácticas para la recopilación y el intercambio de datos de fiabilidad y mantenimiento para equipos industriales. Esta norma proporciona un marco para la gestión de datos de fiabilidad con el objetivo de mejorar la gestión del ciclo de vida de los activos industriales. Dividen en mantenimiento en categorías como:

- a) Aquellas que se realizan para corregir un ítem después de la falla (mantenimiento correctivo).
- b) Aquellas que se realizan para prevenir que un ítem caiga en estado de falla (mantenimiento preventivo); parte de esto pueden ser simplemente los chequeos (inspecciones, pruebas) para verificar la condición y el rendimiento del equipo con el fin de decidir si se requiere un mantenimiento preventivo.

Estas clasificaciones y categorías propuestas por esta norma se pueden evidenciar en la siguiente imagen.

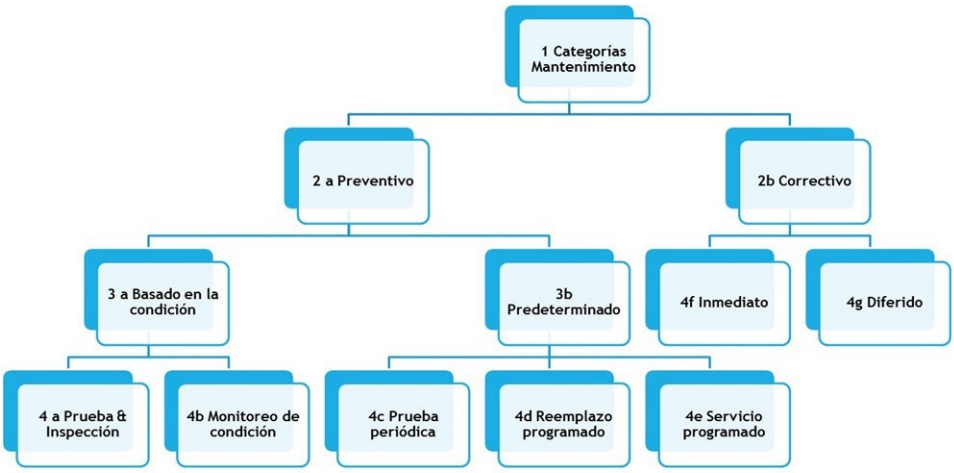

Figura 2: Norma ISO 14224:2016

Haciendo alusión a estos antecedentes históricos que han enmarcado el rumbo del mantenimiento industrial también existen otros puntos de vista interesantes de otros auntores como, F. W. Taylor (1911) aborda la necesidad de aplicar métodos científicos al proceso de trabajo para mejorar la eficiencia y la productividad en las industrias. Taylor propone un enfoque sistemático para analizar y estandarizar los procesos de trabajo, basado en la observación detallada y el estudio riguroso de cada tarea.

Una de las contribuciones clave de Taylor en su trabajo es el desarrollo del estudio de tiempos y movimientos. Este método involucra la descomposición de las tareas en movimientos elementales y la determinación de los tiempos estándar para cada uno de ellos. Al observar y medir repetidamente el tiempo necesario para realizar cada movimiento, los gerentes pueden establecer tiempos promedio y predecir con mayor precisión la duración total de una tarea. En el contexto del mantenimiento industrial, Taylor argumenta que aplicar métodos científicos al estudio de los procesos de mantenimiento puede mejorar significativamente la eficiencia y la efectividad de las operaciones. Al analizar en detalle cada paso del proceso de mantenimiento y estandarizar los métodos de trabajo, las empresas pueden reducir los tiempos de inactividad y aumentar la disponibilidad de los equipos.

Como a porte al proyecto el enfoque de Taylor en los principios de la administración científica proporciona un sólido marco para elevar la eficiencia y efectividad de nuestras operaciones de mantenimiento industrial.

Este enfoque puede conducir a la reducción de costos, un aumento en la disponibilidad de equipos y una mejora en la competitividad de nuestra empresa. Para alcanzar estos objetivos, estamos empleando técnicas de estadística multivariante como herramientas fundamentales en nuestro proceso de optimización.

3.1.1. Planes de mantenimiento Industrial.

Un plan de mantenimiento preventivo, según lo definido la norma ISO 14224, implica un enfoque sistemático y proactivo para mantener la disponibilidad y confiabilidad de los activos industriales. Este tipo de plan se basa en la programación regular de inspecciones, ajustes,

limpiezas, lubricaciones y reemplazos de componentes con el objetivo de prevenir fallas inesperadas. Además, el plan incluye la recolección y el análisis de datos de fiabilidad y mantenimiento para mejorar continuamente las estrategias de mantenimiento y optimizar la vida útil de los equipos (ISO, 2016). Estos planes de mantenimiento preventivo generalmente tienen una hoja de ruta donde se guardar las actividades a realizar, los materiales requeridos y el tiempo necesario para realizar la actividad y de acuerdo a dicha norma todo activo debe tener un plan de mantenimiento ya se preventivo, basado en condición o a falla.

Los planes de mantenimiento industrial son fundamentales en la gestión eficiente de activos en entornos productivos. Estos planes no solo buscan corregir fallas una vez que ocurren, sino que también tienen como objetivo principal prevenir averías y maximizar la disponibilidad de la maquinaria y equipos industriales, Según Martínez (2020) La relevancia de los planes de mantenimiento industrial radica en su capacidad para mejorar la confiabilidad de los equipos y reducir los tiempos improductivos. estos planes permiten realizar un seguimiento sistemático de las condiciones de los activos, anticipando posibles fallas y programando intervenciones preventivas. En un entorno industrial, donde el tiempo de inactividad puede resultar costoso, contar con un plan de mantenimiento adecuado puede marcar la diferencia en términos de eficiencia y rentabilidad.

Gómez Poma, Jhon Angelo en su Artículo, Implementación de un plan de mantenimiento predictivo por análisis vibracional de la centrifugas continuas Broadbent y discontinuas Fives Cail de la empresa Cartavio S.A.A, menciona algunas pautas para implementar planes de mantenimientos y los problemas que causan las intervenciones no planificadas en el proceso productivo. En términos generales proyecto exhibe los resultados alcanzables mediante la implemen- tación de un modelo predictivo para abordar problemas de fallas en máquinas. Se observa que, tras la aplicación de esta técnica Gomez Poma (2022), se logra mejorar la disponibilidad y aumentar la eficiencia del proceso productivo, al mejorar la eficiencia la maquina va a presentar menos fallos lo que indica menos tiempo en intervenciones correctivas.

3.1.2. Estadística y tiempos de mantenimiento industrial

Assis y Marques (2021) en su artículo “A Dynamic Methodology for Setting Up Inspection Time Intervals in Conditional Preventive Maintenance”, manifiestan que uno de los principales problemas en el mantenimiento y especial el basado en condición es el de determinar los intervalos de tiempo de inspección y en proponen un método para determinar los tiempos requeridos utilizando técnicas estadísticas y en especial funciones de probabilidad como la distribución Weibull.

Partes del resumen argumenta, que se presenta un nuevo método para establecer un calendario óptimo para inspeccionar un componente crítico que falla debido al desgaste como lo describe una función de probabilidad de Weibull, Considerando un conjunto de intervalos de inspección, de modo que la confiabilidad entre cada dos inspecciones se mantenga igual o por debajo de un umbral preestablecido, manteniendo al mismo tiempo los costos totales de inspección, producción degradada, consecuencias de fallas y reparación al mínimo, el anterior estudio es un gran aporte al proceso de mantenimiento, se evidencia como aplicando estas distribuciones se puede predecir la probabilidad que el equipo falle en un periodo de tiempo que también se traduce en la confiabilidad de la máquina, es decir que esta pueda dar el rendimiento esperado en un periodo de tiempo sin presentar fallas.

Ellos hacen alusión a la distribución Weibull, de acuerdo a la literatura, es ampliamente usada

en la ingeniería como modelo para la descripción del tiempo de duración de un componente. Esta distribución fue introducida por el científico sueco del mismo nombre, quien demostró que el esfuerzo al que se someten los materiales puede modelarse mediante el empleo de esta distribución. Castañeda Blanco (2004). En el contexto de los Modelos Lineales Generalizados (GLM, por sus siglas en inglés), la distribución Weibull puede ser utilizada como una función de enlace para modelar variables de respuesta que no tienen una distribución normal.

Otra de las investigaciones realizadas donde se usa estadística para temas de mantenimiento, se describe en el artículo "Preventive maintenance models – higher operational reliability" de los autores Legát et al. (2017) Los autores presentan un método para determinar el intervalo óptimo para el mantenimiento periódico preventivo y un parámetro de diagnóstico óptimo para el mantenimiento/reemplazo predictivo. Además, los autores plantean la pregunta: ¿cómo influye el mantenimiento preventivo en la probabilidad de falla y la confiabilidad operativa de los elementos del sistema que han sido sometidos a mantenimiento periódico preventivo? Responden a la pregunta utilizando enfoques informáticos analíticos y de simulación. Los resultados están en forma cuantitativa y dan relaciones entre los intervalos de mantenimiento preventivo y las funciones de confiabilidad. Los ejemplos demuestran la idoneidad del método para objetos de ingeniería típicos que utilizan una distribución de Weibull de tres parámetros. La aplicación del método supone un beneficio sustancial tanto para el fabricante como para el usuario del equipo técnico.

Revisando mas literatura Nardo et al. (2021) en su artículo "Development and implementation of an algorithm for preventive machine maintenance.Engineering Solid Mechanics," -Desarrollo e implementación de un algoritmo para el mantenimiento preventivo de máquinas industriales , citando parte del resumen del articulo manifiestan que el objetivo es desarrollar un modelo de optimización del mantenimiento para mantener un alto nivel de eficiencia y confiabilidad de la maquinaria. El enfoque se basa en el mantenimiento preventivo mediante la sustitución parcial o total de componentes críticos, la atención se centra en una máquina concreta que se ha detenido varias veces, reduciendo su disponibilidad operativa y provocando un elevado coste de no producción. En el estudio utiliza un modelo de Weibull para analizar y optimizar el correcto proceso de mantenimiento de la maquinaria considerada. Luego, los datos de falla se analizan y se programan con el objetivo final de estandarizar los procedimientos de intervención de los operadores para reducir el tiempo de las mismas intervenciones.

En términos resumidos en el artículo anterior lo que buscan es encontrar el tiempo en el que una maquina presenta un fallo, con el fin de realizar la intervención antes de que se presente la avería y así no incurrir en costos por horas hombre correctivas, perdida de producción y repuestos no necesarios, proponen que se debe recopilar los datos históricos de mantenimiento y por medio de distribuciones y técnicas estadísticas tratar de predecir la confiabilidad y disponibilidad del equipo en periodo x de tiempo.

3.1.3. Impacto del mantenimiento en la eficiencia operativa

El mantenimiento en la industria desempeña un papel fundamental al asegurar que los equipos, maquinarias y sistemas funcionen de manera óptima, al funcionar bien la producción se mantiene y no se incurre en pérdidas lo menciona Jardine y Tsang (2013), el mantenimiento efectivo no solo se trata de reparar equipos cuando fallan y en los tiempos efectivos, lo importante de hacer mantenimiento es por medio de planes preventivos y predictivos prevenir averías o fallas antes de que ocurran, esta idea es respaldada por estudios como el de Kumar et

al. (2018), quienes señalan que el mantenimiento preventivo puede reducir significativamente el tiempo de inactividad no planificado y mejorar la eficiencia de las operaciones industriales.

La importante de hacer mantenimiento y en especial el preventivo y predictivo es destacado por varios autores, para citar algunos. Olarte et al. (2010) Resalta la importancia de la planificación del mantenimiento para lograr altos niveles de calidad en cualquier tipo de empresa y proporciona una descripción histórica de los cambios en la implementación del modelo de mantenimiento en la industria, cuando se menciona planificación del mantenimiento se refiere al preventivo, que generalmente tiene un periodo de ejecución o frecuencia, un tiempo de ejecución y acciones a realizar, puntualizar que siempre el objetivo de este es conservar y prevenir que se presenten fallas. Arroyo Vaca y Obando Quito (2022) subraya los beneficios del mantenimiento preventivo, incluido el aumento de la productividad, la reducción de los costos de mantenimiento y la prolongación de la vida útil de la maquinaria.

Para puntualizar de acuerdo a estos el mantenimiento industrial, con técnicas y planeación efectiva juega un papel crítico en la producción y la eficiencia operativa de las empresas industriales. Es crucial que aplique enfoque integral que incluya estrategias preventivas, predictivas y correctivas puede minimizar el tiempo de inactividad, optimizar los recursos, mejorar la calidad del producto y aumentar la rentabilidad. Estas consideraciones son fundamentales para que las empresas mantengan su competitividad en un entorno empresarial cada vez más dinámico y exigente.

3.2. Marco teórico

3.2.1. Estadística Multivariada

La estadística multivariada es una rama de la estadística que estudia la relación entre múltiples variables. Estos métodos permiten analizar la complejidad de los fenómenos reales, donde intervienen diversas variables que interactúan entre sí. Algunas de las técnicas multivariadas más empleadas son el análisis de componentes principales, el análisis factorial, el análisis discriminante y los modelos de ecuaciones estructurales. Estas herramientas son ampliamente utilizadas en campos como la economía, la psicología, la biología y la ingeniería para extraer información valiosa de conjuntos de datos multidimensionales Hair et al. (2018).

Por otro lado, La estadística multivariada ha demostrado ser una herramienta poderosa en diversos campos de investigación. Por ejemplo, en el ámbito de la medicina, los métodos multivariados han sido utilizados para identificar factores de riesgo asociados a enfermedades, clasificar pacientes en grupos de tratamiento y modelar la progresión de dolencias crónicas. En el campo de las ciencias sociales, estas técnicas han permitido analizar la influencia de múltiples variables socioeconómicas en el comportamiento humano Tabachnick y Fidell (2013), para afirmar estos aportes y citando otros artículos anteriores donde varios autores experimentan con estás técnicas para predecir los tiempos efectivos para realizar un mantenimiento , caso de modelos de regresión y modelos de probabilidad como la distribución Weibull.

3.2.2. Técnicas de estadística multivariante.

Las técnicas de estadística multivariante como los modelos de regresión lineal, modelos lineales generalizados, análisis de componentes principales entre otros, son ideales para analizar conjuntos de datos que abarcan múltiples variables simultáneamente. A diferencia de los modelos univariados, que se centran en una sola variable a la vez, estas técnicas exploran las complejas interrelaciones entre varias variables. Esto permite una comprensión más profunda y completa de los datos. El autor Anderson (2009), explica lo útiles que son para identificar interdependencias entre variables que podrían pasarse por alto en análisis univariados. En la investigación moderna, donde las relaciones entre variables son la norma, esta capacidad para descubrir conexiones ocultas es fundamental. Junto a Anderson el autor Hair et al. (2018) resalta los beneficios de estas técnicas, como la capacidad para capturar la complejidad inherente de conjuntos de datos multidimensionales que permite comprender mejor las relaciones subyacentes. No obstante, a pesar de sus ventajas, las técnicas multivariadas presentan limitaciones importantes. La interpretación de los resultados puede ser complicada, especialmente en la gestión de grandes conjuntos de datos con numerosas variables. En situaciones donde la interpretación precisa es crucial, como en estudios médicos, esta complejidad puede representar un desafío significativo. La asunción de linealidad también constituye una limitación relevante. Aunque algunos métodos permiten extensiones no lineales, si las relaciones de las variables no son lineales los modelos podrían dar resultados no confiables al momento de predecir. En nuestro contexto, donde buscamos predecir los tiempos de reparación de máquinas con fallos, la diversidad de técnicas disponibles nos proporciona la capacidad de modelar nuestra variable dependiente mediante múltiples técnicas buscando la que mejor se acople a la naturaleza de nuestros datos.

3.2.3. Modelo de regresión múltiple

El modelo de regresión lineal múltiple es una extensión del modelo de regresión lineal simple que permite explorar la relación entre una variable dependiente y múltiples variables independientes Montgomery et al. (2012). En este modelo, la relación entre las variables se modela mediante un plano o un hiperplano en un espacio de varias dimensiones. Los coeficientes de regresión representan el efecto de cada variable independiente en la variable dependiente Fox (2015). Los modelos de regresión lineal múltiple son ampliamente utilizados en diversas áreas, como la economía, la sociología, la psicología y la epidemiología, entre otros Draper y Smith (2014). Se utilizan para predecir o explicar el valor de una variable dependiente en función de múltiples variables independientes.

la fórmula general para un modelo de regresión lineal es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{1}$$

- Y es la variable dependiente que estamos tratando de predecir.
- β_0 es el intercepto o término constante.
- $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes asociados con las variables predictoras $X_1, X_2, \ldots, X_p$, respectivamente.
- ε es el término de error, que representa la variabilidad no explicada por el modelo.

Los modelos de regresión lineal se basan en múltiples supuestos, como los siguientes:

Linealidad: Se asume que la relación entre las variables predictoras y la variable dependiente es lineal. Esto implica que los cambios en las variables predictoras están asociados con cambios proporcionales en la variable de interés.

Independencia de errores: Se asume que los errores (residuos) de la regresión no están correlacionados entre sí.

Homocedasticidad: La varianza de los errores debe ser constante en todos los niveles de las variables predictoras. Esto significa que la dispersión de los errores es uniforme en toda la gama de valores de las variables predictoras.

Normalidad de errores: Se asume que los errores de la regresión se distribuyen normalmente. Esto significa que los residuos siguen una distribución normal con una media de cero.

Independencia de variables predictoras: Es crucial que las variables predictoras en un modelo de regresión sean independientes unas de otras. Una elevada correlación entre estas variables puede entorpecer la interpretación precisa de los coeficientes, afectando la validez del modelo.

3.2.4. Modelos Lineales Generalizados (GLM)

Los Modelos Lineales Generalizados (GLM) constituyen una poderosa extensión de los modelos lineales tradicionales, permitiendo abordar una amplia gama de tipos de datos y distribuciones de error más allá de la normalidad. A diferencia de los modelos lineales simples que asumen que la variable dependiente es normalmente distribuida, los GLM facilitan la modelización de la relación entre una variable dependiente, que puede seguir diversas distribuciones (como binomial, Poisson, o gamma), y una o más variables independientes a través de una función de enlace.

Esta capacidad para manejar diferentes distribuciones hace que los GLM sean particularmente adecuados para analizar datos donde la variable respuesta es, por ejemplo, una proporción, un conteo, o una medida de tiempo hasta un evento. Esta versatilidad se discute ampliamente en la literatura, con (Nelder & Wedderburn, 1972) siendo uno de los trabajos pioneros que estableció las bases teóricas de los GLM.

Los GLM son útiles en una variedad de campos, incluyendo la biometría, la epidemiología y las ciencias sociales, donde las variables de interés no se ajustan a la distribución normal. Por ejemplo, en estudios de respuesta binaria como la presencia o ausencia de una enfermedad, los datos pueden modelarse efectivamente usando una distribución binomial con una función de enlace logit.

Una de las principales ventajas de los GLM es su capacidad para proporcionar estimaciones y pruebas inferenciales sobre los parámetros del modelo que son interpretables y útiles para la toma de decisiones.

Los GLM ofrecen un marco estadístico robusto y flexible para la modelización de relaciones entre variables, siendo capaces de abordar una amplia gama de situaciones de datos más allá de las limitaciones de los modelos lineales tradicionales. Su aplicación en diversas áreas testimonia su valor en la investigación cuantitativa, como destaca McCullagh y Nelder (1989) en su obra sobre teoría y aplicaciones de los GLM.

La fórmula general de los GLM es la siguiente:

$$g(\mu) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n \tag{2}$$

- $g(\mu)$: Representa la variable dependiente, es decir, la variable que estamos tratando de predecir. En el contexto de la regresión lineal, $g(\mu)$ se estima como una combinación lineal de las variables independientes.
- β_0: Es el intercepto, también conocido como el coeficiente de intersección o término constante. Representa el valor esperado de la variable dependiente cuando todas las variables independientes son iguales a cero.
- $\beta_1, \beta_2, \ldots, \beta_n$: Son los coeficientes de regresión, que representan el cambio esperado en la variable dependiente debido a un cambio unitario en cada una de las variables independientes, manteniendo constantes todas las demás variables independientes.
- $x_1, x_2, \ldots, x_n$: Son las variables independientes o predictores. Cada una de estas variables puede tomar diferentes valores y se utilizan para predecir el valor de la variable dependiente $g(\mu)$.

En la tabla siguiente se presentan diversas distribuciones utilizadas en modelos lineales generalizados (GLM), junto con las fórmulas de las funciones de enlace asociadas. Cada función de enlace tiene sus propias características y es adecuada para diferentes situaciones:

- Identidad: Esta función es apropiada cuando se espera una relación lineal entre las variables predictoras y la variable respuesta. Se utiliza cuando se asume que el efecto de las variables predictoras es aditivo en la escala original de la variable respuesta.
- Logit: Es útil para modelar la probabilidad de éxito en datos binarios. Transforma la probabilidad de éxito en una escala continua y simétrica, lo que facilita la interpretación de los efectos

de las variables predictoras en términos de log-odds.

- Probit: Similar al logit, pero utiliza la función de distribución acumulativa normal inversa (función probit). Se utiliza cuando se prefiere una distribución normal en lugar de una binomial.
- Clog-log: Esta función se emplea para ajustarse a la varianza no constante en datos binarios. Es útil cuando se necesita modelar la probabilidad de éxito en datos con una gran variabilidad en la tasa de éxito.
- Logaritmo: Se utiliza para modelar variables de conteo, como en el caso de la distribución de Poisson. Transforma la media de la variable respuesta en el logaritmo de la media, asegurando que las predicciones sean siempre positivas.
- Inverso: Adecuada para variables con valores positivos y asimétricos, como en la distribución gamma. Transforma la media de la variable respuesta en el inverso de la media.
- Recíproco: Similar al inverso, pero se utiliza cuando se prefiere una distribución más flexible que se ajuste a la varianza no constante en datos de tiempo hasta el evento.

Tabla 1

Distribuciones, fórmulas, funciones de enlace y cuándo se utilizan en GLM

Distribución	Fórmula	Función
Normal	$g(\mu) = \mu$	Identidad
Binomial	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	Logit
Binomial	$g(\mu) = \Phi^{-1}(\mu)$	Probit
Binomial	$g(\mu) = \ln(-\ln(1-\mu))$	Clog-log
Poisson	$g(\mu) = \ln(\mu)$	Logaritmo
Gamma	$g(\mu) = \frac{1}{\mu}$	Inverso
Gamma	$g(\mu) = \ln(\mu)$	Logaritmo
Inversa gaussiana	$g(\mu) = \mu$	Identidad
Inversa gaussiana	$g(\mu) = \frac{1}{\mu}$	Recíproco
Inversa gaussiana	$g(\mu) = \ln(\mu)$	Logaritmo

3.2.5. Modelo de regresión SVR (Máquinas de Vectores de Soporte)

La Regresión mediante Máquinas de Vectores de Soporte (SVR) emerge como una técnica innovadora en el ámbito del aprendizaje automático, evolucionando a partir de los principios establecidos por las Máquinas de Vectores de Soporte (SVM) para enfrentar desafíos específicos asociados con tareas de regresión. Este enfoque se distingue por su capacidad para predecir valores continuos, aplicando un marco que equilibra la precisión predictiva con la complejidad del modelo para asegurar la generalización efectiva a datos no vistos. El desarrollo conceptual y teórico de la SVR se apoya en la teoría de optimización y la teoría estadística del aprendizaje, particularmente en los trabajos de Vapnik y sus colaboradores, quienes han sido fundamentales en la formulación de las SVM y, por extensión, de la SVR Vapnik (1995).

Central para la metodología de la SVR es la implementación de funciones kernel, que facilitan el mapeo no lineal de los datos de entrada a un espacio de alta dimensión donde las relaciones complejas entre variables pueden ser modeladas linealmente. Esta característica es esencial para abordar con éxito datos que presentan patrones no lineales, permitiendo que la SVR

se adapte a una amplia variedad de contextos y tipos de datos. La selección del kernel adecuado (e.g., lineal, polinomial, RBF) y la calibración de sus parámetros son cruciales para el rendimiento del modelo, enfatizando la importancia de una comprensión detallada de la naturaleza del conjunto de datos y el problema específico a resolver Schölkopf y Smola (2002).

La formulación general de la Regresión mediante Máquinas de Vectores de Soporte (SVR) busca encontrar una función $f(x)$ que tenga a lo más un ϵ-desvío de los valores reales y_i para todas las muestras de entrenamiento, y al mismo tiempo sea lo más plana posible. Matemáticamente, esto se traduce en minimizar la siguiente función objetivo:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3}$$

Sujeto a las restricciones:

$$\begin{aligned} y_i - \langle w, x_i\rangle - b &\leq \epsilon + \xi_i, \\ \langle w, x_i\rangle + b - y_i &\leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* &\geq 0, \qquad \text{para todo } i = 1, \ldots, n. \end{aligned}$$

Donde:

- x_i son las características de entrada
- y_i son los valores objetivo
- w es el vector de pesos
- b es el término de sesgo
- ξ_i y ξ_i^* son variables de holgura que miden el grado de desviación permitido para errores por encima y por debajo de ϵ, respectivamente
- C es el parámetro de regularización que establece el balance entre la suavidad de la función $f(x)$ y el grado hasta el cual desviaciones mayores que ϵ son toleradas.

3.2.6. Modelo de Regresión con Árboles de Decisión

La Regresión mediante Árboles de Decisión representa una metodología destacada en el dominio del aprendizaje supervisado, extendiendo los principios fundamentales de los árboles de decisión hacia la predicción de valores continuos. Este enfoque se caracteriza por su habilidad para manejar de manera intuitiva tanto relaciones lineales como no lineales entre las variables, mediante la división del espacio de características en regiones homogéneas. El desarrollo de los árboles de decisión para regresión se nutre de conceptos sólidos de teoría de la información y estadísticas, apoyándose en criterios de división como la reducción de la varianza y el error cuadrático medio para optimizar las decisiones de partición Breiman et al. (1984).

Un aspecto crucial de los árboles de decisión en la regresión es su capacidad para adaptarse automáticamente a las complejidades de los datos, permitiendo una modelización flexible de interacciones entre variables sin requerir especificaciones de modelo predefinidas. La estructura

del árbol de decisión, compuesta por nodos de decisión y hojas, ofrece una representación gráfica que facilita la interpretación de cómo las características influyen en la predicción final. La selección de parámetros como la profundidad máxima del árbol y el mínimo número de muestras requeridas para un nodo, juegan un papel esencial en prevenir el sobreajuste y asegurar una buena generalización del modelo Quinlan (1993).

La formulación general de la regresión con árboles de decisión busca construir un modelo que particione el espacio de características de manera que las predicciones en cada región sean lo más precisas posible, minimizando una función de pérdida predeterminada, típicamente el error cuadrático. Esta meta se consigue mediante la siguiente representación general:

$$\min_{\theta} \sum_{i=1}^{n} L(y_i, f(x_i; \theta)), \tag{4}$$

donde:

- x_i son las características de entrada
- y_i son los valores objetivo
- θ representa los parámetros del modelo, incluyendo las decisiones de división y las predicciones en cada hoja
- L es la función de pérdida, usualmente el error cuadrático medio entre las predicciones y los valores reales.

La construcción de un árbol de decisión implica seleccionar las divisiones que más reduzcan la función de pérdida en el conjunto de datos de entrenamiento, con un proceso iterativo de partición binaria. Las técnicas de poda y validación cruzada se emplean frecuentemente para afinar el modelo y evitar el sobreajuste, manteniendo un equilibrio entre la complejidad del modelo y su capacidad predictiva.

3.2.7. Modelo de Regresión con Random Forest

La regresión con Random Forest es una técnica robusta de aprendizaje automático que construye y combina múltiples árboles de decisión para mejorar la precisión y la capacidad de generalización de las predicciones. A diferencia de un solo árbol de decisión, Random Forest incorpora la sabiduría de la multitud mediante la agregación de los resultados de numerosos árboles para formar la predicción final, un proceso conocido como "bagging" o "Bootstrap Aggregating" Breiman (2001) .

Esta técnica sobresale en su capacidad de manejar grandes conjuntos de datos con múltiples variables de entrada, siendo capaz de capturar interacciones complejas y no lineales sin necesidad de una especificación de modelo detallada. Random Forest es particularmente conocido por su robustez ante el sobreajuste, gracias a la aleatoriedad introducida durante la construcción de los árboles, que incluye la selección de características y muestras.

Al igual que con los árboles individuales, la regresión con Random Forest opera dividiendo el espacio de las características en subespacios, pero mejora la estabilidad y precisión al promediar múltiples árboles, reduciendo la varianza sin aumentar el sesgo.

El desarrollo de un modelo de Random Forest para regresión sigue principios estadísticos sólidos, empleando el promedio de las predicciones de los árboles individuales para minimizar

el error total y proporcionar una estimación más fiable y precisa. En el corazón de Random Forest, la diversidad entre los árboles se fomenta a través de dos mecanismos principales:

Bootstrap de muestras: Cada árbol se entrena con una muestra aleatoria del conjunto de datos (con reemplazo), permitiendo que diferentes árboles aprendan de distintas porciones de los datos. Selección aleatoria de características: En cada división, se selecciona un subconjunto aleatorio de las características disponibles, lo que obliga a los árboles a tomar decisiones basadas en diferentes combinaciones de entradas. Estos procesos se resumen en la siguiente fórmula general de Random Forest para regresión:

$$\min_{\{\theta_k\}} \frac{1}{K} \sum_{k=1}^{K} \sum_{i \in B_k} L(y_i, f(x_i; \theta_k)), \tag{5}$$

donde:

- K es el número de árboles en el bosque.
- B_k es el conjunto de datos bootstrap para el k-ésimo árbol.
- $f(x_i; \theta_k)$ es la predicción del k-ésimo árbol con parámetros θ_k.
- L es la función de pérdida, que continúa siendo comúnmente el error cuadrático medio.

Los parámetros importantes de un Random Forest incluyen el número de árboles (ntree), el número de variables consideradas para dividir en cada nodo (mtry), y el tamaño mínimo de los nodos de hoja. La selección de estos hiperparámetros se realiza típicamente mediante técnicas de búsqueda como la validación cruzada y la búsqueda en cuadrícula o aleatoria.

La interpretación de un modelo de Random Forest puede ser menos directa en comparación con un único árbol de decisión debido a la naturaleza agregada de la predicción. No obstante, los métodos de importancia de las variables proporcionan perspectivas significativas sobre la contribución de cada característica al modelo, ayudando a identificar los predictores más relevantes.

3.3. Maching Learning

El machine Learning o aprendizaje automático, es un subcampo de las ciencias de computación que tienen como finalidad establecer algoritmos para que las computadoras aprendan patrones de datos a gran escala con el fin de extraer información que mejoren los procesos de análisis de los datos, el Machine Learning (ML) se ha consolidado como una herramienta fundamental para extraer conocimiento y generar predicciones acertadas. El Machine Learning, definido por Arthur Samuel en 1959 como “el campo de estudio que da a las computadoras la habilidad de aprender sin ser explícitamente programadas” Samuel (1959), ha evolucionado rápidamente en las últimas décadas, convirtiéndose en un pilar clave para la toma de decisiones informadas en una amplia gama de industrias y campos de investigación.

A diferencia de los enfoques de programación tradicionales, donde los algoritmos se basan en instrucciones y reglas definidas explícitamente, el Machine Learning permite que los sistemas aprendan y mejoren automáticamente a partir de los datos. Esto se logra a través de la aplicación de técnicas como el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo, cada uno de los cuales aborda diferentes tipos de problemas y desafíos Hastie et al. (2009).

3.3.0.1. Aprendizaje Supervisado El aprendizaje supervisado se conoce por su capacidad de extraer conocimiento a partir de datos etiquetados, permitiendo a los algoritmos establecer relaciones entre variables de entrada y variables de salida para generar predicciones y clasificaciones precisas.

En el aprendizaje supervisado, el algoritmo tiene acceso a un conjunto de datos de entrenamiento que contiene observaciones o instancias con valores conocidos para la variable de salida. Esto le permite al modelo "aprender" a reconocer patrones y construir una función que mapee adecuadamente las entradas a las salidas deseadas Hastie et al. (2009), Por citar un ejemplo, en mantenimiento un problema de predicción de tiempos , el algoritmo utilizaría datos históricos de tiempos correctivos, con información sobre variables como total horas, estado del equipo, frecuencia de mantenimientos y tiempos empleados para reparaciones anteriores, para establecer un modelo que permita pronosticar los tiempos a futuro.

Dentro del aprendizaje supervisado, existen dos tipos principales de problemas: regresión y clasificación. Los problemas de regresión buscan predecir una variable numérica continua, como el precio de una acción o la demanda de un producto. Por otro lado, los problemas de clasificación tienen como objetivo asignar instancias a categorías o clases discretas, como determinar si un correo electrónico es spam o no spam Witten et al. (2016). Este se consolida como una herramienta fundamental en el campo del Machine Learning, permitiendo a los investigadores y profesionales generar predicciones y clasificaciones precisas a partir de datos etiquetados. A medida que la disponibilidad y complejidad de los datos continúan aumentando, el aprendizaje supervisado seguirá desempeñando un papel crucial en la toma de decisiones informadas en diversos ámbitos

3.3.0.2. Aprendizaje No Supervisado A diferencia del aprendizaje supervisado, el aprendizaje no supervisado es una rama crucial del machine learning que se enfoca en descubrir patrones y estructuras ocultas en conjuntos de datos sin etiquetar, Según Bishop (2006), el aprendizaje no supervisado implica encontrar una representación útil o una estructura subyacente en los datos sin conocer las salidas esperadas. Esto es especialmente útil en situaciones donde los datos pueden no estar etiquetados o cuando se desea explorar la naturaleza intrínseca de los datos sin ninguna suposición previa sobre las relaciones entre variables.

Algunos de los principales métodos de aprendizaje no supervisado incluyen el análisis de conglomerados (clustering), la reducción de dimensionalidad y la detección de anomalías Witten et al. (2016). El análisis de conglomerados tiene como finalidad agrupar las observaciones en función de sus características, de manera que los elementos dentro de un mismo grupo sean similares entre sí y diferentes a los de otros grupos. Esto resulta útil para segmentación de clientes, identificación de tipos de consumidores, o agrupación de genes con funciones biológicas similares.

Resaltar que una de las principales fortalezas del aprendizaje no supervisado es su capacidad de descubrir patrones y relaciones inesperadas en los datos, lo que puede conducir a nuevos conocimientos y oportunidades de innovación. Además, estas técnicas son especialmente útiles cuando los datos disponibles no cuentan con etiquetas o variables de salida claras, como en el análisis exploratorio de conjuntos de datos complejos. Para citar algunos algoritmos de aprendizaje No Supervisado: K-Means, DBSCAN, Algoritmos de Agrupamiento Jerárquico, Análisis de Componentes Principales (PCA), t-distributed Stochastic Neighbor Embedding,

Gaussian Mixture Models (GMM), Isolation Forest.

3.4. Conceptos y definiciones

3.4.1. R Studio

R destaca por su notable flexibilidad, permitiendo la aplicación de una amplia gama de modelos estadísticos y matemáticos. Desde simples modelos lineales hasta técnicas avanzadas como regresión logística, árboles de decisión, redes neuronales y análisis de series temporales, R ofrece una completa variedad de herramientas para el modelado predictivo y descriptivo.

Una de las principales fortalezas de R reside en su extensa colección de paquetes especializados, los cuales abarcan una amplia diversidad de áreas, desde el análisis estadístico más básico hasta técnicas de modelado más sofisticadas. Estos paquetes son desarrollados y mantenidos tanto por la comunidad de usuarios como por expertos en diversos campos, asegurando una amplia disponibilidad de herramientas para abordar cualquier tarea de modelado que se presente.

Paquetes necesarios

- **DT:** `DT` proporciona una interfaz R para las DataTables de JavaScript, permitiendo la creación de tablas HTML interactivas que pueden ser visualizadas en R Markdown y aplicaciones Shiny. Este paquete es especialmente útil para la presentación de datos de forma dinámica, donde el usuario puede ordenar, filtrar y paginar la información directamente desde la visualización generada.
- **lmtest:** El paquete `lmtest` ofrece una amplia gama de pruebas estadísticas para la evaluación de modelos lineales en R. Incluye herramientas para realizar pruebas de coeficientes, comparaciones de modelos, diagnósticos de residuos, entre otros. `lmtest` es valioso para analistas y estadísticos que buscan validar y comparar modelos lineales mediante pruebas rigurosas y basadas en criterios estadísticos sólidos.
- **car:** `car` (Companion to Applied Regression) está diseñado para complementar el análisis de regresión aplicada en R, proporcionando una serie de funciones y conjuntos de datos útiles para el diagnóstico y la visualización de modelos lineales. Desde análisis de varianza hasta gráficos de influencia y diagnóstico, `car` facilita una comprensión más profunda de los modelos lineales y sus supuestos.
- **lme4:** `lme4` se utiliza para ajustar modelos lineales mixtos y modelos lineales generalizados mixtos en R, ofreciendo una solución robusta para el análisis de datos con estructuras de dependencia complejas o agrupadas. Es ideal para datos jerárquicos, longitudinales o de panel, permitiendo a los investigadores modelar efectos tanto fijos como aleatorios y entender la variabilidad en los datos a múltiples niveles.
- **lubridate:** `lubridate` simplifica el manejo de fechas y horas en R, proporcionando funciones intuitivas para la manipulación, el parseo y el cálculo con objetos de tiempo. Este paquete resuelve muchos de los desafíos comunes al trabajar con datos temporales, facilitando la conversión entre formatos de fecha y hora, la gestión de zonas horarias, y la realización de operaciones aritméticas con fechas.
- **MuMIn:** `MuMIn` es un paquete en R dedicado a la selección y promedio de modelos, basado en criterios de información como el AIC o BIC. Resulta especialmente útil en la comparación

exhaustiva de modelos candidatos y en la síntesis de los resultados a través del promedio de modelos, proporcionando así una visión más holística de los efectos modelados y ayudando a mejorar la precisión predictiva.

- **ggthemes:** `ggthemes` extiende las capacidades de `ggplot2` al ofrecer una variedad de temas y escalas adicionales para la personalización de gráficos en R. Desde estilos inspirados en medios de comunicación y software famosos hasta paletas de colores adaptadas para una visualización de datos más atractiva y profesional, `ggthemes` enriquece la presentación visual de los análisis.
- **rpart:** El paquete `rpart` facilita la implementación de árboles de decisión para clasificación y regresión en R. Utilizando el enfoque CART (Árboles de Clasificación y Regresión), `rpart` permite a los usuarios explorar la estructura de los datos y hacer predicciones basadas en las relaciones identificadas entre las variables. Es ampliamente utilizado tanto en análisis exploratorios como en la construcción de modelos predictivos complejos.
- **e1071:** `e1071` trae a R una serie de algoritmos de aprendizaje automático y estadística, incluyendo máquinas de vectores de soporte (SVM), clasificación Naive Bayes, y análisis de componentes principales (PCA). Desarrollado inicialmente en la Universidad Tecnológica de Viena, este paquete es una herramienta esencial para tareas de clasificación, regresión y reducción de dimensionalidad en análisis de datos avanzados.

3.4.2. Pyhton

Python es un lenguaje de programación de alto nivel, interpretado y de propósito general. Es conocido por su sintaxis clara y legible, lo que lo hace ideal para principiantes y programadores experimentados por igual. Python es ampliamente utilizado en una variedad de campos, incluyendo desarrollo web, ciencia de datos, inteligencia artificial, automatización de tareas, entre otros. Se destaca por su amplia biblioteca estándar y su comunidad activa que contribuye con paquetes y recursos para facilitar el desarrollo de software. Python Software Foundation (2022)

Paquetes y librerías utilizadas

• Sckit-Learn: Scikit-Learn es una biblioteca de aprendizaje automático de código abierto para el lenguaje de programación Python. Proporciona una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, así como herramientas para la preparación, evaluación y visualización de datos. Scikit-Learn es ampliamente utilizado en la comunidad de ciencia de datos y machine learning debido a su facilidad de uso, su documentación detallada y su integración con otras bibliotecas populares de Python, como NumPy, Pandas y Matplotlib

• NumPy: abreviatura de “Numerical Python”, es una biblioteca de código abierto para el lenguaje de programación Python. Proporciona soporte para arrays y matrices multidimensionales, junto con una amplia colección de funciones matemáticas de alto nivel para operar en estas estructuras de datos. NumPy es fundamental en el ecosistema de Python para computación científica y análisis de datos, ya que ofrece una eficiente manipulación de datos numéricos, cálculos numéricos rápidos y funciones para trabajar con arrays de manera eficiente.

• Pandas: es una poderosa biblioteca de código abierto para el lenguaje de programación Python, diseñada principalmente para el análisis y manipulación de datos. Proporciona

estructuras de datos flexibles y eficientes, como DataFrames y Series, que permiten a los usuarios trabajar con datos tabulares de una manera intuitiva y eficaz. Pandas es ampliamente utilizado en la comunidad de ciencia de datos y análisis de datos para tareas como limpieza, manipulación, exploración y análisis de datos. McKinney (2010).

• Matplotlib: es una biblioteca de visualización de datos de código abierto para el lenguaje de programación Python. Proporciona una amplia variedad de herramientas para crear gráficos estáticos, interactivos y animados de manera sencilla y flexible. Matplotlib es ampliamente utilizado en la comunidad de ciencia de datos, análisis de datos, investigación científica y educación debido a su versatilidad y capacidad para producir gráficos de alta calidad en diversos formatos y estilos Hunter (2007).

3.4.3. Variables y Tipos de Variables

En el ámbito de la investigación, el concepto de variables es fundamental para comprender y analizar datos. Las variables son características, propiedades o rasgos que pueden medirse, observarse o manipularse en un estudio. Son elementos clave que los investigadores estudian y comparan para entender las relaciones entre ellos y cómo influyen en un fenómeno o proceso. Una variable es cualquier cantidad que pueda tener más de un valor. En un estudio científico, las variables son las características o atributos que se miden para evaluar los efectos de las condiciones experimentales. Estas pueden ser tan simples como la edad de una persona o tan complejas como el nivel de contaminación en un área determinada.

Las variables se dividen en dos categorías:

a. Variables Independientes: Estas son las variables que un investigador manipula o cambia para observar su efecto sobre otras variables. En un experimento, la variable independiente es la que se controla deliberadamente para ver cómo afecta a la variable dependiente. Por ejemplo, en un estudio sobre el efecto de la luz en el crecimiento de las plantas, la cantidad de luz sería la variable independiente.

a. Variables Dependientes: Son las variables que se observan y miden en respuesta a los cambios en la variable independiente. En el ejemplo de las plantas, el crecimiento de las plantas sería la variable dependiente, ya que se espera que cambie en respuesta a la cantidad de luz que reciben.

3.4.3.1. Tipos de Variables

Las variables pueden clasificarse aún más en diferentes tipos, lo que ayuda a los investigadores a comprender mejor cómo analizar y presentar los datos. Algunas de las clasificaciones comunes son:

a. Variables Categóricas o Nominales: Estas variables representan categorías discretas que no tienen un orden inherente. Ejemplos incluyen el género, el estado civil, el tipo de vehículo (automóvil, camión, motocicleta), etc. Se pueden codificar como números, pero estos números no tienen un significado numérico intrínseco.

b. Variables Ordinales: Las variables ordinales tienen categorías con un orden específico, pero las diferencias entre las categorías no son necesariamente iguales. Por ejemplo, una escala de "satisfacción" que va desde "muy insatisfecho" hasta "muy satisfecho" es una variable ordinal.

c. Variables de Intervalo: Estas variables tienen un orden significativo entre los valores y

las diferencias entre los valores son igualmente significativas. Sin embargo, no hay un punto cero absoluto. Un ejemplo común es la temperatura en grados Celsius o Fahrenheit. En la escala Celsius, por ejemplo, la diferencia entre 20°C y 30°C es la misma que entre 30°C y 40°C.

d. Variables de Razón: Son similares a las variables de intervalo, pero tienen un cero absoluto, lo que significa que el valor cero representa la ausencia completa de la característica que se está midiendo. Por ejemplo, la altura, el peso, el tiempo y la cantidad de dinero son variables de razón. No tener dinero (0) es diferente de tener $10 o $20.

3.4.4. Pruebas de bondad y ajuste

Las pruebas de bondad de ajuste e independencia son ampliamente utilizadas en diversas áreas de la ciencia para llevar a cabo análisis de datos. Una prueba de bondad de ajuste permite evaluar la hipótesis de que una variable aleatoria sigue cierta distribución de probabilidad y se utiliza en situaciones donde se requiere comparar una distribución observada con una teórica o hipotética, compararla con datos históricos o con la distribución conocida de otra población.

Las pruebas más utilizadas para estudiar la independencia y la bondad de ajuste son las chicuadrado de Pearson, el estadístico de Kolmogorov-Smirnov, criterio de información de Akaike (AIC) entre otros debido a su fácil aplicación y a que se encuentran en todos los paquetes estadísticos. Estas pruebas asumen que todas las observaciones son independientes y que están igualmente distribuidas, supuestos que sólo se satisfacen para un muestreo aleatorio simple con reposición y se cumplen aproximadamente en una muestra aleatoria simple sin reposición para una fracción de muestreo pequeña Quintero M (2004).

3.4.5. Distribución de probabilidad.

En teoría de la probabilidad y estadística, la distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra. La distribución de probabilidad está definida sobre el conjunto de todos los sucesos y cada uno de los sucesos es el rango de valores de la variable aleatoria. También puede decirse que tiene una relación estrecha con las distribuciones de frecuencia. De hecho, una distribución de probabilidades puede comprenderse como una frecuencia teórica, ya que describe cómo se espera que varíen los resultados.

La distribución de probabilidad está completamente especificada por la función de distribución, cuyo valor en cada x real es la probabilidad de que la variable aleatoria sea menor o igual que x.

3.4.6. Modelos en mantenimiento Industrial:

En el contexto del mantenimiento industrial, la aplicación de modelos multivariables y de series temporales juega un papel fundamental en la predicción del tiempo necesario para reparar máquinas en un intervalo específico. Estos modelos contribuyen significativamente a la eficiencia y fiabilidad de los procesos industriales.

3.4.7. Optimización de Procesos:

La optimización de tiempos de atención de fallas en redes eléctricas y en el mantenimiento industrial se enfoca en reducir los tiempos de respuesta ante situaciones de emergencia, garantizando la seguridad del personal y la eficacia en la asignación de recursos.

3.4.8. Limitaciones y Consideraciones:

A pesar de las ventajas de las técnicas de estadística multivariante, como la capacidad para identificar interdependencias entre variables, es importante considerar las limitaciones en la interpretación de resultados, especialmente en conjuntos de datos complejos. La asunción de linealidad también puede ser un desafío en ciertos contextos.

3.4.9. Hipótesis nula (H0)

Es la afirmación inicial que se quiere poner a prueba. Generalmente, se establece como la afirmación de que no hay efecto o diferencia entre grupos, o que una población sigue cierta distribución. Se denota como H0.

3.4.10. Hipótesis alternativa (H1)

Es la afirmación opuesta a la hipótesis nula. Se trata de lo que se intenta probar o demostrar con los datos. Puede ser una afirmación de diferencia, efecto o cualquier otra condición diferente a la de la hipótesis nula.

3.4.11. Estadístico de prueba

Es una medida calculada a partir de los datos de la muestra, que se utiliza para tomar una decisión sobre la hipótesis nula. Puede ser una media, una proporción, una diferencia entre medias, entre otros.

3.4.12. Nivel de significancia (alpha)

Es la probabilidad máxima que estamos dispuestos a aceptar de cometer un error tipo I, es decir, rechazar incorrectamente la hipótesis nula cuando es verdadera. Es comúnmente fijado en valores como 0.05 o 0.01.

3.4.13. Regla de decisión

Se basa en comparar el estadístico de prueba con un valor crítico, derivado de la distribución de probabilidad apropiada bajo la hipótesis nula. Si el estadístico de prueba cae en la región de rechazo, se rechaza la hipótesis nula a favor de la hipótesis alternativa.

3.4.14. P-valor

Es la probabilidad, bajo la suposición de que la hipótesis nula es verdadera, de obtener un estadístico de prueba al menos tan extremo como el observado en la muestra. Si el p-valor es menor que el nivel de significancia, se rechaza la hipótesis nula.

3.4.15. Error tipo I y tipo II

El error tipo I ocurre cuando se rechaza incorrectamente la hipótesis nula cuando es verdadera. El error tipo II ocurre cuando se acepta incorrectamente la hipótesis nula cuando es falsa.

3.4.16. Potencia de la prueba

Es la probabilidad de rechazar la hipótesis nula cuando es falsa. Depende del tamaño del efecto, el tamaño de la muestra y el nivel de significancia.

3.4.17. Multicolinealidad

Es una situación en la que dos o más variables predictoras en un modelo de regresión están altamente correlacionadas entre sí. Esto puede causar problemas en la interpretación de los coeficientes del modelo y puede afectar la estabilidad de las estimaciones.

3.4.18. VIF (Factor de Inflación de la Varianza)

Es una medida que cuantifica la gravedad de la multicolinealidad. Un valor de VIF mayor que 10 se considera una indicación de multicolinealidad grave, lo que sugiere que las variables predictoras están muy correlacionadas y pueden estar causando problemas en el modelo.

3.4.19. Criterios de información

Los criterios de información son herramientas estadísticas esenciales en la selección de modelos, diseñadas para evaluar la calidad de un modelo estadístico en términos de su capacidad predictiva, penalizando al mismo tiempo la complejidad del modelo para evitar el sobreajuste. Los tres criterios más destacados en este contexto son el Criterio de Información de Akaike (AIC), el Criterio de Información de Akaike Corregido (AICc), y el Criterio de Información Bayesiano (BIC).

3.4.20. AIC (Criterio de Información de Akaike)

Introducido por Hirotugu Akaike en 1974, el AIC mide la pérdida de información cuando se utiliza un modelo para representar el proceso que generó los datos. Se fundamenta en la teoría de la información.

Fórmula:

$$AIC = 2k - 2\ln(L) \tag{6}$$

donde k es el número de parámetros estimados en el modelo y L es la máxima verosimilitud del modelo.

3.4.21. AICc (Criterio de Información de Akaike Corregido)

Es una versión del AIC que incluye una corrección para tamaños de muestra pequeños. Resulta particularmente relevante cuando el tamaño de la muestra es pequeño en comparación con el número de parámetros estimados.

Fórmula:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} \tag{7}$$

donde n es el tamaño de la muestra, k es el número de parámetros, y L es la máxima verosimilitud.

3.4.22. BIC (Criterio de Información Bayesiano)

También conocido como el Criterio de Schwarz, el BIC es otra medida de selección de modelos que incorpora tanto el número de parámetros en el modelo como el tamaño de la muestra. Fue desarrollado por Gideon Schwarz en 1978.

Fórmula:

$$BIC = \ln(n)k - 2\ln(L) \tag{8}$$

donde n es el tamaño de la muestra, k es el número de parámetros estimados, y L es la máxima verosimilitud.

3.4.23. Error Cuadrático Medio (MSE)

El Error Cuadrático Medio (MSE) es una métrica comúnmente utilizada para evaluar la precisión de un modelo de regresión. Se calcula como el promedio de los cuadrados de las diferencias entre los valores predichos por el modelo y los valores reales. Matemáticamente, se expresa de la siguiente manera:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde: - n es el número de observaciones en el conjunto de datos. - y_i es el valor real de la variable respuesta para la observación i. - $\hat{y}_i$ es el valor predicho por el modelo para la observación i.

El MSE penaliza de manera significativa los errores grandes debido al cuadrado en la fórmula. Cuanto menor sea el valor del MSE, mejor será el ajuste del modelo a los datos.

3.4.24. Raíz del Error Cuadrático Medio (RMSE)

El RMSE es simplemente la raíz cuadrada del MSE. Esta métrica proporciona una medida de la dispersión de los errores en la misma escala que los datos originales, lo que facilita su interpretación. Se calcula de la siguiente manera:

$$RMSE = \sqrt{MSE}$$

Al igual que el MSE, un valor menor de RMSE indica un mejor ajuste del modelo a los datos. El RMSE es especialmente útil para comunicar la precisión del modelo a personas no familiarizadas con las unidades de medida de la variable de interés.

3.4.25. Error Absoluto Medio (MAE)

El Error Absoluto Medio (MAE) es otra métrica de evaluación comúnmente utilizada en modelos de regresión. A diferencia del MSE, el MAE mide el promedio de las diferencias absolutas entre los valores predichos y los valores reales. Matemáticamente, se expresa así:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

El MAE es menos sensible a los valores atípicos en comparación con el MSE, ya que no eleva los errores al cuadrado. Por lo tanto, proporciona una medida más robusta de la precisión del modelo en términos de errores de predicción.

3.4.26. Error Porcentual Absoluto Medio (MAPE)

El Error Porcentual Absoluto Medio (MAPE, por sus siglas en inglés "Mean Absolute Percentage Error") es una medida común utilizada para evaluar la precisión de un modelo de regresión. Mide el porcentaje promedio de los errores en relación con los valores reales.

Matemáticamente, se expresa así:

$$\text{MAPE} = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Donde:

n es el número total de observaciones. y_i es el valor real de la variable dependiente para la observación i. $\hat{y}_i$ es el valor predicho por el modelo para la observación i.

3.4.27. Kernel SVR

El término "kernel" en el contexto de un modelo de Máquina de Soporte Vectorial para Regresión (SVR) se refiere a una función utilizada para transformar el espacio de características de los datos a un nuevo espacio, usualmente de mayor dimensión, donde resulta más sencillo encontrar un hiperplano que se ajuste a los datos en el contexto de regresión. Este concepto permite a los modelos SVR manejar datos que no son linealmente separables o relaciones no lineales entre las características de manera eficiente.

Los kernels más comúnmente utilizados en los modelos SVR incluyen:

1. **Kernel Lineal**: No se realiza ninguna transformación no lineal sobre los datos, útil para relaciones lineales. Se define como:
$$K(x, x\text{t}) = x^T x\text{t}$$

2. **Kernel Polinómico**: Permite modelar relaciones no lineales mediante un polinomio de grado d. Se define como:
$$K(x, x\text{t}) = (1 + x^T x\text{t})^d$$

3. **Kernel RBF (Radial Basis Function) o Gaussiano**: Captura relaciones complejas no lineales y se define como:
$$K(x, x\text{t}) = \exp(-\gamma\|x - x\text{t}\|^2)$$

4. **Kernel Sigmoide**: Transforma los datos de entrada de manera similar a la función sigmoide. Se define como:

$$K(x, x\grave{t}) = \tanh(\alpha x^T x\grave{t} + c)$$

3.4.28. Nodos y hojas de un árbol de decisión

Nodos de decisión (o internos): Estos nodos realizan una pregunta sobre una o más características y dividen el conjunto de datos en dos o más subconjuntos basados en la respuesta a esta pregunta. La pregunta suele tomar la forma de una comparación, como "¿Es el valor de la característica X menor que un cierto umbral?". Cada nodo de decisión tiene dos o más ramas que representan los posibles resultados de la pregunta y conducen a otros nodos o a hojas. La estructura del árbol se construye a partir de estos nodos, comenzando con el nodo raíz en la parte superior, que es el primer punto de decisión.

Nodos hoja (o terminales): Los nodos hoja representan el resultado final del árbol de decisión. En problemas de clasificación, cada hoja está asociada a una clase, que es la predicción del modelo para las instancias que llegan a esa hoja. En problemas de regresión, los nodos hoja suelen contener un valor continuo o el promedio de los valores objetivo de las instancias que llegan a esa hoja, representando la predicción del modelo para esas instancias.

4. Desarrollo del proyecto y resultados

Para el proyecto, "Estudio de modelos de predicción, para la gestión de tiempos correctivos en una empresa Industrial", se hizo un estudio de revisión documental para encontrar información de interés sobre los procesos de mantenimiento industrial y como las técnicas de estadística han contribuido a la mejora de estos, en especial la asignación de tiempos o intervalos de mantenimiento preventivo y correctivo. Diferentes ejemplos verificados donde se usan distribuciones de probabilidad como Weibull para encontrar la confiabilidad de un equipo en un periodo de tiempo, distribuciones como la Beta para calcular las horas en promedio que requiere un proyecto de acuerdo a unos datos históricos como lo describe el pmbok, también se encuentra que muchos autores expresan que el análisis de la información histórica de fallas son importantes para hacer análisis y prevenir que se presenten fallas que afecten la productividad de una planta industrial.

Adicionalmente se estudian diferentes técnicas de estadística multivariante para predecir tiempos como los modelos de regresión múltiple y los modelos lineales generalizados. Para analizar la calidad de estas predicciones se utilizan algoritmos de Maching learning para regresión, como Arboles de decisión, Randon Forest y Máquina de Vectores Soporte (SVM) con el fin de establecer que algoritmo es más potente y genera los mejores resultados de acuerdo a las diferentes métricas de evaluación tales como: error cuadrático medio (MSE), Raíz del error cuadrático medio (RMSE), Error absoluto medio (MAE), MAPE entre otros.

En esta sección, se prestará la metodología a utilizar para la resolución del problema planteado; escribiendo los pasos esenciales mediante la aplicación de técnicas de estadística mutivariante y algortimos de aprendizaje supervizado.

4.1. Metodología

Este proyecto se centra en el estudio de diferentes modelos de predicción para evaluar cuál se adapta mejor a los datos históricos de mantenimiento, previamente recopilados y analizados para estimar el tiempo requerido de reparaciones correctivas en equipos industriales (Motores Eléctricos, Turbinas, Paneles solares, Válvulas, Sistema contra incendios, Alarmas, Aires Acondicionados, Banco de Baterías, Iluminarias, Rodamientos, Instrumentos). Datos históricos, que documentan las horas utilizadas en reparaciones correctivas a lo largo de los años, se constituye como el insumo para el análisis en propuesto. Una base de datos es extensa que enfrenta varios desafíos como la presencia de valores faltantes, la incidencia de datos atípicos y una considerable variabilidad en las horas reportadas de reparaciones.

Ante estos desafíos, se comienza con una limpieza de datos detallada y metódica, proceso crítico implica métodos sofisticados para manejar valores faltantes, técnicas robustas de filtrado para mitigar el impacto de los valores atípicos y estrategias para normalizar la dispersión de los datos. El objetivo principal es desarrollar un modelo que no solo sea el acertado para este caso, sino que también pueda adaptarse a diferentes marcos temporales: diario, semanal y mensual. Esta adaptabilidad asegura que el modelo sea una herramienta versátil para la programación eficiente de las actividades de mantenimiento, mejorando así la planificación de recursos y la estrategia de gestión en las operaciones industriales.

El éxito de este modelo repercutirá directamente en la reducción del tiempo de inactividad de las máquinas y en la optimización del rendimiento y costos de operación. La metodología seleccionada para esta investigación se basa en CRISP-DM (Cross-Industry Standard Process

for Data Mining), que proporciona una estructura sistemática para guiar el desarrollo de proyectos de análisis de datos y modelado predictivo.

CRISP-DM consta de seis fases interconectadas y cíclicas que facilitan el abordaje claro y organizado del proyecto.

Figura 3: Metodología crips dm

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) es un marco estructurado y ampliamente reconocido que guía el proceso de desarrollo de soluciones de minería de datos y modelado de datos. Surgió en la década de 1990 como resultado de un esfuerzo colaborativo de expertos en minería de datos y análisis de datos para estandarizar y formalizar el proceso de minería de datos Chapman et al. (2000).

El desarrollo de CRISP-DM fue liderado por el CRISP-DM Consortium, que incluyó a empresas líderes en la industria, consultoras, universidades y organizaciones de investigación. El objetivo era proporcionar un enfoque común y una estructura estandarizada para proyectos de minería de datos, que pudiera ser aplicable en una amplia variedad de contextos y sectores industriales Shearer (2000).

CRISP-DM consta de seis fases principales, cada una de las cuales aborda aspectos específicos del proceso de modelado de datos:

- Comprensión del negocio (Business Understanding): En esta fase, se busca comprender los objetivos del negocio y los requisitos del proyecto de minería de datos. Se identifican los problemas o las oportunidades que se pretenden abordar, así como los factores críticos de éxito para el proyecto.
- Comprensión de los datos (Data Understanding): Durante esta etapa, se recopilan los datos relevantes para el proyecto y se realiza un análisis exploratorio para comprender su estructura, calidad y significado.
- Preparación de los datos (Data Preparation): En esta fase, se preparan los datos para su uso en el modelado mediante tareas de limpieza, integración, selección y transformación.
- Modelado (Modeling): Durante esta etapa, se seleccionan y desarrollan modelos predictivos o

descriptivos utilizando técnicas de minería de datos y análisis estadístico.

- Evaluación (Evaluation): En esta fase, se evalúan y validan los modelos desarrollados utilizando conjuntos de datos de prueba o técnicas de validación cruzada.
- Despliegue (Deployment): Finalmente, en esta etapa, se implementan los modelos en un entorno operativo y se integran en los sistemas existentes.

CRISP-DM ha sido ampliamente adoptado y utilizado en una variedad de industrias y aplicaciones Pyle (1999), incluyendo banca, telecomunicaciones, comercio electrónico, salud, manufactura, marketing y más.

4.1.1. Comprensión del Negocio

La falta de estimaciones precisas para determinar el tiempo requerido para el mantenimiento correctivo en instalaciones industriales conduce a una serie de repercusiones operativas y estratégicas significativas que afectan el desempeño operacional de una empresa dedicada a estas labores de mantener equipos. Una incorrecta planeación de las horas hombre necesarias para estas tareas pueden desencadenar diferentes problemas y desviaciones que en ultimas afecta económicamente la empresa, porque se emiten programas de mantenimiento y diferentes indicadores para medir la gestión empresarial, lo que por los trabajos mal planeados o tiempos mal asignados llevan a un incumplimiento por falta de recurso técnico lo que son causales para penalizar los contratos de mantenimiento previamente firmados.

El problema radica en las empresas que no tienen un proceso automatizado para definir los tiempos para atender trabajos correctivos y se basan en prácticas como las que propone el pmbok juicio de expertos, donde una persona con un previo conocimiento decide cuanto tiempo disponer para la semana o mes para atender trabajos correctivos, otro proceso que generalmente utilizan es el promedio semanal, mensual o simplemente tienen un porcentaje de las horas totales para entender trabajos correctivos, son argumentos que en términos generales lo explican Jyoti Pareek y Ravi Shankar en su proyecto “Predicting Project Completion Time Using Regression Models: A Case Study in the Manufacturing Industry”, es un artículo presenta un estudio de caso sobre el uso de modelos de regresión para predecir el tiempo de finalización de proyectos en la industria manufacturera y donde concluyen que el uso de técnicas de regresión avanzadas puede mejorar significativamente la capacidad de las organizaciones industriales para estimar con precisión los tiempos de finalización de proyectos, lo que tiene importantes implicaciones en términos de planificación, gestión de riesgos y satisfacción del cliente.

La precisión en los presupuestos de proyectos de mantenimiento es vital para mantener la viabilidad financiera y la sostenibilidad de las operaciones industriales. Una mala predicción en las horas de mantenimiento puede llevar a exceder las estimaciones presupuestarias, afectando negativamente la rentabilidad del proyecto Wang (2017), haciendo énfasis en la anterior, una mala estimación de los tiempos para trabajos correctivos puede ser una afectación grave para la gestión de costos del proyecto y cumplimiento de los indicadores de gestión, y las interrupciones en el mantenimiento preventivo tienen el potencial de afectar adversamente la cadena de suministro y comprometer las relaciones con los clientes Cheng (2019).

Si no se calculan bien los tiempos generalmente se empieza a incurrir en uso de tiempos extras para cumplir con las metas previamente planeadas o se exige que los empleados sean más productivos, mas horas extras mas costos, mas trabajo para el experto en el tema, el bienestar

del empleado y todo el personal de mantenimiento también son áreas de preocupación significativas, donde la incertidumbre y el aumento de la carga de trabajo pueden conducir a un estrés laboral incrementado a un deterioro en la moral del equipo P. Taylor y Schmidt (2020). Además, la urgencia por reanudar la producción no debe comprometer los estándares de seguridad, ya que la seguridad del personal es de suma importancia y debe ser siempre una prioridad González y Martínez (2022). La seguridad es importante y más en empresas industriales donde se tienen equipos que manejan altas presiones y falla puede llegar a causar accidentes de seguridad de procesos considerados catastróficos.

Para abordar estos desafíos y problemas propuestos, es esencial adoptar enfoques basados en datos para mejorar la precisión de las estimaciones de tiempo de mantenimiento correctivo. La implementación de herramientas de análisis predictivo y el uso estratégico de datos históricos son estrategias recomendadas para mejorar la planificación del mantenimiento y reducir los tiempos de inactividad no planificados Morales (2020).

4.1.2. Comprensión de los datos

El conjunto de datos que se empleará en este estudio comprende inicialmente 36,795 observaciones y 11 variables distintas, las cuales relacionan las actividades de reparación industrial realizadas en una empresa del sector de hidrocarburos en Colombia desde el año 2019 hasta el año 2023. Se registran 23132 trabajos preventivos con un total 285,106 horas hombre y 13633 trabajos correctivos con total 206,270 horas hombre. El numero trabajos correctivos corresponden a un 37.05 % y el trabajo preventivo a un 62.95 %, las horas hombre correctivas a un 41,98 % y las preventivas a un 58.02 %. Estos datos fueron suministrados para temas educativos por una empresa del sector y provienen del sistema de manteniendo SAP que utilizan para su respectiva gestión, se aclara que se modifican algunos nombres de las variables y descripciones de equipos por temas de seguridad y demás temas legales.

Seguidamente se presenta un resumen de las variables que conforma esta base de datos antes mencionada:

- Disciplina: La disciplina se refiera al grupo e equipo de trabajo que realiza mantenimientos en la empresa, esta se divide en 7 disciplinas: Pozos, Instrumentos, Mecánica, Mecánica VAL, Electricidad, Línea de Mtto, Mecánica CBM.
- Epocadelanio: Esta variable hace referencia a la época del año en que se ejecuta el mantenimiento, generalmente son dos estaciones Invierno y verano.
- Cumplimiento_Estrategia: Se refiere al cumplimiento del plan de estrategia de mantenimiento preventivo ejecutado en el mes. Cumplimiento_Programa: Se refiere al cumplimiento del programa de mantenimiento ejecutado en el mes.
- Ready_Backlog: Es la carga de laboral que tienen un equipo de trabajo, se mide en semanas y debe estar dentro de un rango de 4 a 6 semanas, un número más alto es un indicador de excesos de trabajo.
- HH_Actv_Gnerales: son las horas hombres que reportan en actividades generales, esta variable cuantitativa se por actividades como charlas de seguridad, celebraciones, reuniones entre otros.
- HH_En_la_Maquina: Es la variable objetivo y hace referencia al total de horas que el técnico necesita para ejecutar un trabajo en la maquina o equipo a reparar.

- Otras_HH: Esta variable se refiere a las horas adicionales que se requieren para hacer actividades propias del mantenimiento como lectura de permisos, transporte a sitio, alistamiento de herramientas entre otros.
- Fe.CreadaOrden: Esta variable indica la fecha cuando se crea el trabajo que puede ser correctivo o preventivo.
- Fe.Ejecutada: Hace referencia a la fecha de ejecución del trabajo en la planta de producción industrial.

En la **tabla 1** a continuación podemos ver las primeras observaciones de nuestra base de datos

Tabla 2

Base de Datos de Mantenimiento de Equipos

Epoca del año	estrategia Mtto	Programa Mtto	Ready back log	HH Actv Gnerales	HH En la Maquina	Otras HH	Tipo mantenimiento	Disciplina
Verano	0.977	0.9459	5.16	0.4	6.0	1.0	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	0.4	2.8	1.0	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	0.3	1.0	0.0	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	1.7	17.7	4.0	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	1.3	15.3	1.2	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	0.9	1.3	4.8	Correctivo	Pozoz
Verano	0.977	0.9459	5.16	1.6	18.0	1.4	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	0.4	3.0	1.8	Correctivo	Pozoz
Verano	0.977	0.9459	5.16	1.6	6.6	3.1	Correctivo	Instrumentos
Verano	0.977	0.9459	5.16	2.6	22.5	5.5	Correctivo	Instrumentos

4.1.3. Preparación de los datos

Antes de proceder con la aplicación de diversas técnicas de estadística multivariante, es importante garantizar la adecuada preparación y limpieza de la base de datos. Esta etapa preliminar incluye la eliminación o imputación de valores vacíos o marcados como NA, la minimización o el tratamiento de valores atípicos en las variables para evitar distorsiones en el análisis, y la evaluación de las correlaciones entre variables para identificar y mitigar posibles problemas de multicolinealidad. Estas acciones son cruciales para asegurar la integridad y la calidad de los datos, sentando así una sólida fundación que permita la implementación efectiva y precisa de las técnicas estadísticas multivariantes. Este proceso no solo facilita un análisis más fiable y exhaustivo, sino que también potencia la capacidad de extraer insights valiosos y significativos de los datos.

4.1.3.1. Eliminando los valores vacíos Si los valores NA no se manejan adecuadamente, pueden introducir sesgos en el análisis de datos y el planteamiento de los modelos, ya sea porque los datos faltan, se ignoran por completo o porque se imputan de manera incorrecta, lo que podría distorsionar las conclusiones y recomendaciones,por tanto es paso crucial en el preprocesamiento de los dato,ignorar o eliminar de manera inadecuada estos valores puede introducir sesgos significativos en los análisis y conducir a conclusiones erróneas.

Afortunadamente en nuestra data los valores **NA** (9) son significativamente bajos, lo que posibilita ignorarlos sin afectar considerablemente nuestros resultados.

4.1.3.2. Valores atípicos Los datos atípicos, conocidos como outliers en inglés, son valores que se encuentran significativamente fuera del rango esperado en una variable del conjunto de datos. En nuestro contexto pueden haber maquinas que conlleven una cantidad significativa de horas mucho mayor a las demás, para analizarlas utilizamos gráficos de caja.

Los gráficos de caja mostrados en la **figura 2** muestran estadísticas de posición clave, como los cuartiles, los valores máximos y mínimos, junto con los posibles valores atípicos. Se nota una gran dispersión de datos en las distintas variables numéricas que estamos analizando.

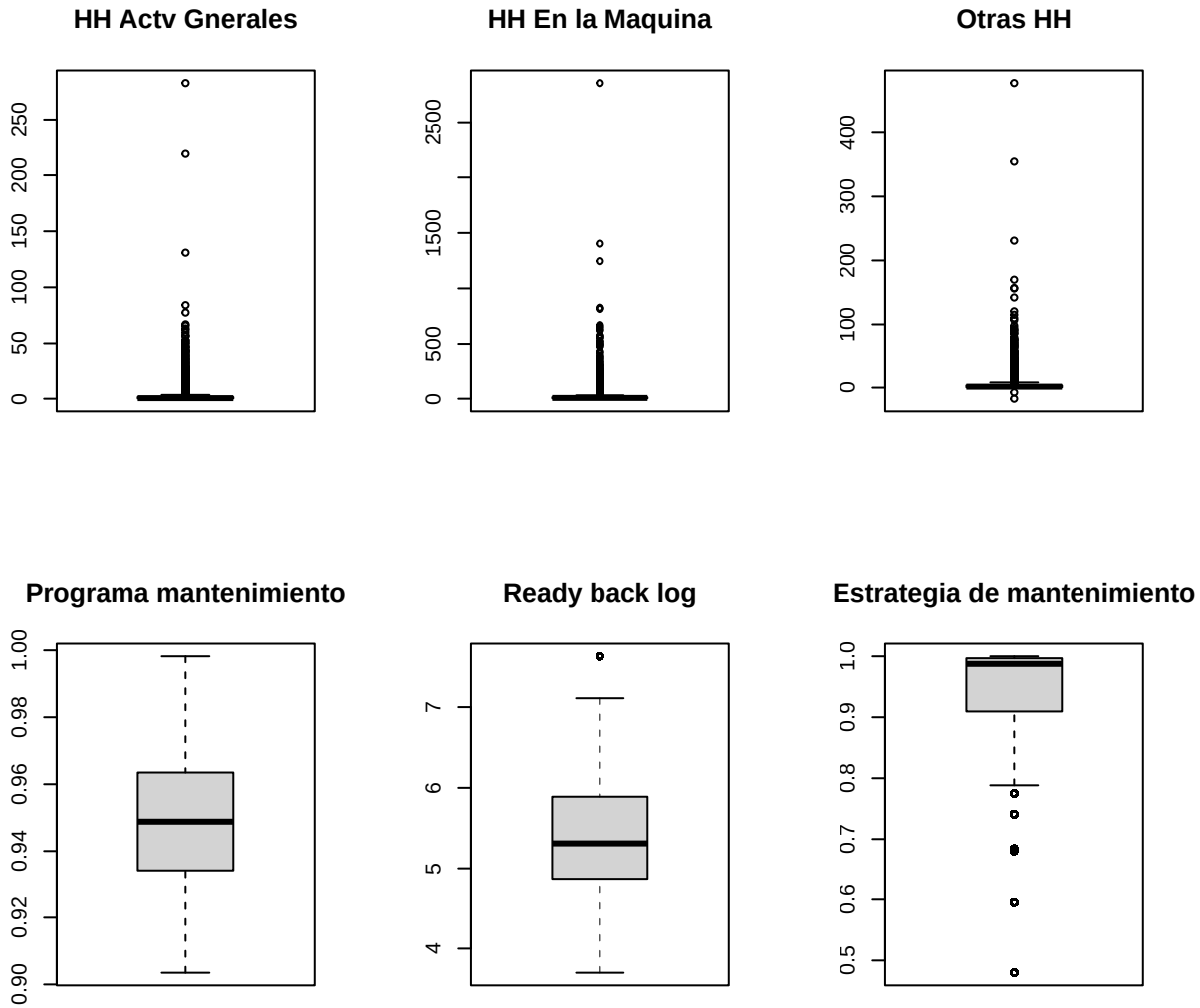

Figura 4: Valores atípicos iniciales

HH Actv Generales: Presenta varios valores que podrían considerarse atípicos, los cuales

están bastante alejados de la mediana.

HH En la Maquina: Hay menos valores atípicos comparados con la variable anterior, pero aún presenta algunos valores que son visiblemente diferentes del resto.

Otras HH: Similarmente, se muestran varios valores atípicos.

Programa Mtto: Los valores atípicos están menos dispersos que en las variables de horas; sin embargo, la mediana está más cerca del cuartil inferior, lo que podría indicar una distribución asimétrica.

Ready back log: Este boxplot también presenta valores atípicos, pero en menor cantidad en comparación con las variables de horas.

Estrategia de mantenimiento: Exhibe una distribución que parece ser más uniforme y simétrica; sin embargo, aún hay presencia de valores atípicos.

Antes de aplicar el filtro de Hampel, es importante reconocer que estos valores atípicos pueden ser indicativos de errores de medición o entrada de datos, pero también podrían representar variaciones reales dentro del proceso de mantenimiento que son importantes para entender la variabilidad del trabajo. El filtro de Hampel reemplazará estos valores atípicos con la mediana, lo cual puede suavizar estas variaciones y presentar una vista más homogénea de los datos, pero también podría ocultar patrones o características importantes. Por ello, siempre se debe proceder con cautela y comprender el contexto de los datos antes de aplicar cualquier método de limpieza de datos.

4.1.3.3. Tratamiento de los valores atípicos El método, conocido como **filtro de Hampel** se basa en el concepto de la mediana absoluta de las desviaciones (MAD, por sus siglas en inglés), que es una medida robusta de la variabilidad de un conjunto de datos. La MAD se calcula como la mediana de las desviaciones absolutas de cada punto de datos con respecto a la mediana del conjunto de datos.

El filtro de Hampel es particularmente útil cuando se trabaja con datos que pueden contener valores atípicos o cuando se requiere una medida robusta de la tendencia central y la variabilidad. Ayuda a mejorar la robustez de los análisis estadísticos al reducir el impacto de los valores extremos en los resultados.

$$\text{Filtro de Hampel: } y_i = \begin{cases} x_i & \text{si } |x_i - \text{mediana}(X)| \leq k \times \text{MAD}(X) \\ \text{mediana}X) & \text{en otro caso} \end{cases} \tag{9}$$

A continuación, se detalla el proceso que describe la fórmula:

- **yi**: Representa el valor de salida para la i-ésima observación después de aplicar el Filtro de Hampel. Este valor puede ser el original (x_i) si no se considera atípico, o puede ser reemplazado por la mediana del conjunto de datos si se identifica como un outlier.
- **xi**: Es el valor original de la i-ésima observación en el conjunto de datos.
- **Mediana(X)**: Es la mediana de todos los valores en el conjunto de datos X. La mediana es usada como medida de tendencia central debido a su robustez frente a valores atípicos, a diferencia de la media.

- **k**: Es un factor de escala que determina el umbral de sensibilidad para identificar a un valor como atípico. Un valor comúnmente usado para k está en el rango de 2.5 a 3.0, aunque puede ajustarse según las necesidades específicas del análisis para controlar la tolerancia frente a las variaciones naturales en los datos.
- **MAD(X)**: La Desviación Mediana Absoluta (MAD) de los valores en el conjunto de datos X. La MAD es una medida de dispersión que, al igual que la mediana, es menos sensible a los valores extremos en comparación con la desviación estándar. Proporciona una estimación de la variabilidad en torno a la mediana.

La condición $x_i - mediana(X) <= kxMAD(X)$ se utiliza para evaluar si un valor x_i es un outlier. Si la diferencia absoluta entre x_i y la mediana de X es mayor que k veces la MAD, se considera que x_i es un outlier y se reemplaza por la mediana del conjunto de datos; de lo contrario, se conserva el valor original x_i.

La razón por la cual un valor se considera atípico si la diferencia absoluta entre x_i y la mediana de X es mayor que k veces la MAD (Desviación Mediana Absoluta) se basa en la premisa de que la mayoría de los datos en un conjunto deberían agruparse cerca de una medida central —en este caso, la mediana— si los datos siguen una distribución relativamente simétrica o incluso si presentan cierto sesgo debido a la presencia de valores atípicos.

La mediana es una medida de tendencia central que es más robusta frente a valores atípicos que, por ejemplo, la media. Esto significa que no se ve tan afectada por valores extremadamente altos o bajos. La MAD, por su parte, es una medida de dispersión que indica cuán dispersos están los datos alrededor de la mediana. Al multiplicar la MAD por un factor de escala k (comúnmente un valor entre 2.5 y 3), se establece un "umbral" que define lo que se considera variabilidad "normal" alrededor de la mediana.

Si la diferencia absoluta entre un valor particular x_i y la mediana excede este umbral (k x MAD(X)), se asume que x_i es tan diferente de la mayoría de los otros valores en el conjunto de datos que no se puede atribuir simplemente a la variabilidad natural de los datos. En cambio, se considera un "outlier" o valor atípico, sugiriendo que podría haber sido generado por un mecanismo diferente al que produjo el resto de los datos o que podría ser el resultado de un error de medición o registro.

Al reemplazar estos valores atípicos por la mediana, se intenta mitigar su impacto en análisis posteriores, evitando que estos valores extremos distorsionen los resultados, como podría ser el caso en la estimación de parámetros de modelos estadísticos o en la construcción de predicciones. La elección de la mediana como valor de reemplazo ayuda a mantener la estructura central de los datos sin introducir un sesgo significativo que podría resultar de utilizar, por ejemplo, la media.

Utilizando el **k** igual a 3 aplicamos el filtro de hampel en todas las variables numéricas de nuestra base de datos:

En la **figura3** se observa una reducción significa en el numero de valores atípicos, lo que indica una buena eficiencia de la técnica en eliminar los valores atípicos en la variable

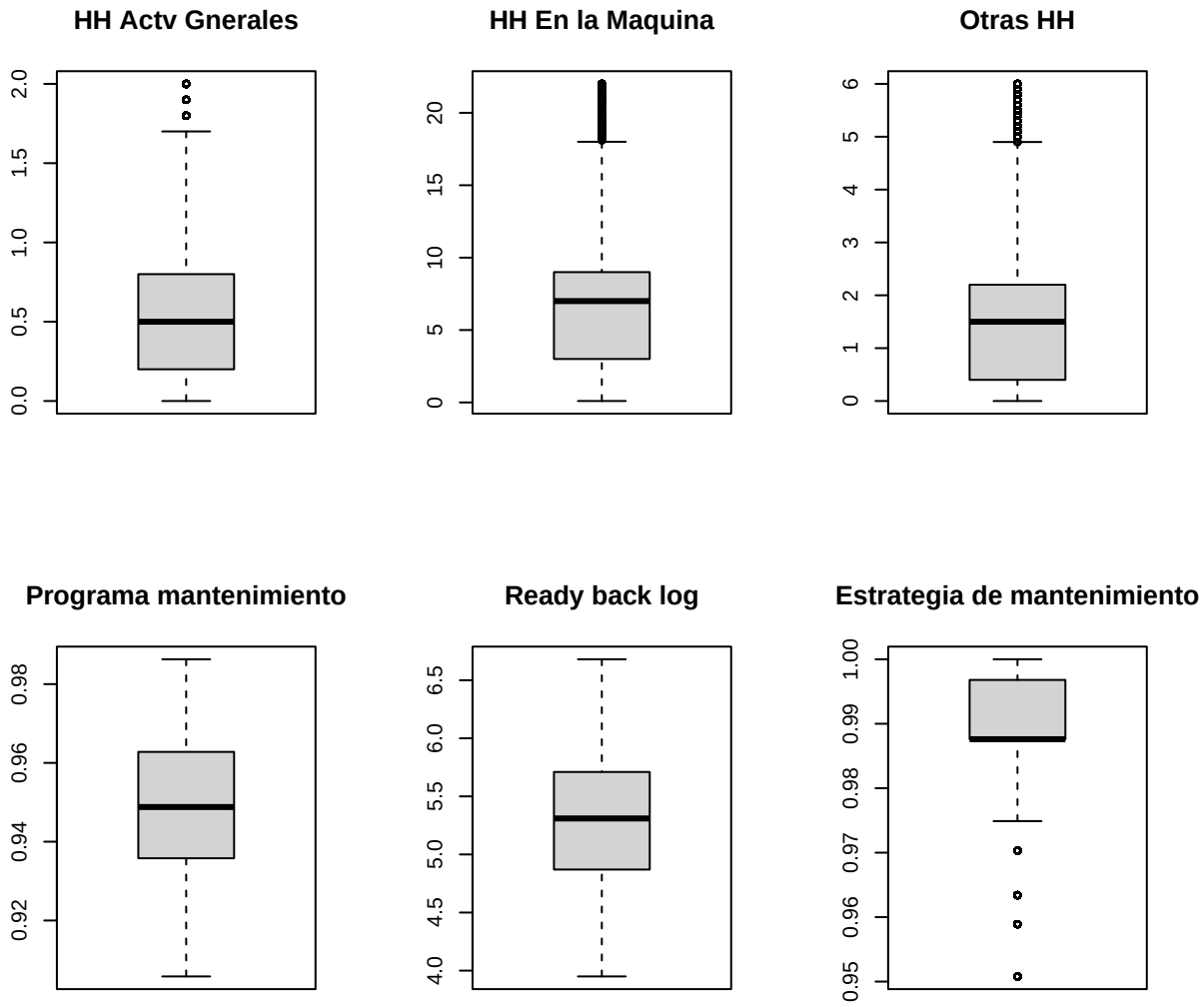

Figura 5: Filtro de hampel aplicado

La principal diferencia visual es una reducción en la cantidad de valores atípicos visibles y una posible disminución en la longitud de los bigotes de los boxplots. Esto implica una apariencia más limpia y posiblemente más simétrica de la distribución de los datos, reflejando una variabilidad que se ajusta más estrechamente a la mayoría de los datos, libre de las distorsiones causadas por valores extremadamente altos o bajos.

4.1.3.4. Análisis descriptivo de los datos En esta sección, se presenta un análisis detallado de los datos utilizados para el proyecto. El objetivo principal es proporcionar una visión general de las características principales de la muestra, así como identificar tendencias y patrones relevantes.

Variables de tipo cuantitativo

Tabla 3

estadísticas descriptivas de las variables numéricas

	estrategia Mtto	Programa Mtto	Ready back log	HH Actv Gnerales	HH En la Maquina	Otras HH
media	0.9895	0.9492	5.3263	0.5469	7.1658	1.5905
mínimo	0.9508	0.9058	3.9500	0.0000	0.1000	0.0000
cuartil 1	0.9876	0.9358	4.8700	0.2000	3.0000	0.4000
mediana	0.9876	0.9488	5.3100	0.5000	7.0000	1.5000
cuartil 3	0.9968	0.9628	5.7100	0.8000	9.0000	2.2000
máximo	1.0000	0.9863	6.6800	2.0000	22.0000	6.0000
varianza	0.0001	0.0004	0.3035	0.2548	24.9059	2.1276
desviación típica	0.0091	0.0199	0.5509	0.5048	4.9906	1.4586
coef.variación	0.0092	0.0210	0.1034	0.9231	0.6964	0.9171
RIC	0.0092	0.0270	0.8400	0.6000	6.0000	1.8000
asimetría	-1.9189	-0.1608	0.2777	1.1477	1.0051	1.0706
curtosis	4.9913	-0.6679	0.0886	0.7937	0.3839	0.6339
moda_1	0.9876	0.9429	5.3100	0.5000	7.0000	0.0000

Al proceder con un análisis descriptivo detallado de los datos relacionados con las actividades de mantenimiento, se establece una base sólida para la construcción y formulación de modelos estadísticos avanzados. Las estadísticas descriptivas revelan diferencias sustanciales en las escalas de medición entre las distintas variables, lo que puede conllevar a desafíos en términos de interpretación y significancia en relación con la variable objetivo.

Un aspecto notable es la variabilidad manifiesta en ciertas variables, como “Programa Mtto” y “Ready back log”, donde se evidencia una dispersión considerable alrededor de la media, tal y como lo indica la magnitud de la varianza. De igual importancia es el reconocimiento de coeficientes de variación prominentes en las variables “HH Actv Generales”, “HH En la Maquina” y “Otras HH”. La alta variación relativa reflejada por estos coeficientes sugiere que la media no representa adecuadamente la distribución de los datos, lo cual cuestiona su utilidad como medida central para estas variables.

Estas características de los datos, como las diferencias de escala y la elevada variabilidad, podrían tener implicaciones significativas en la aplicación de modelos lineales. En particular, los modelos que presuponen homocedasticidad y linealidad en la relación entre predictores y la variable dependiente pueden resultar inadecuados. Por ejemplo, la influencia desmedida de variables con mayor escala puede sesgar los coeficientes estimados, llevando a interpretaciones erróneas de su relevancia en el modelo.

La precisión de las estimaciones de los coeficientes y la fiabilidad de las pruebas estadísticas

subsecuentes también pueden verse afectadas por la alta variabilidad relativa. Los modelos lineales, que son susceptibles a la influencia de valores atípicos y a una distribución anormal de los residuos, podrían no proporcionar inferencias estadísticas confiables bajo estas condiciones.

Estos problemas puede deberse a las técnicas de recolección utilizadas para recolectar los datos, según Gujarati (2009) a medida que mejoren dichas técnicas la presencia de valores atípicos y alta varianza tenderá a disminuirse.

Variables de tipo cualitativo

Época del año

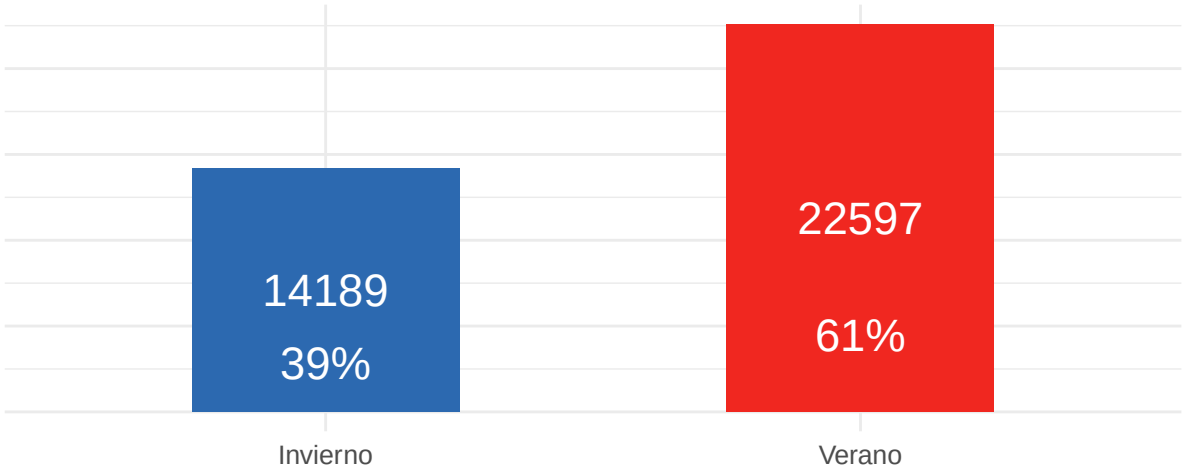

Figura 6: Distribución de reparaciones en las épocas del año

Se nota que la mayoría de las reparaciones se llevaron a cabo durante el verano, representando un 59 % del total. Se sugiere que el invierno podría influir en el número de horas necesarias para completar las reparaciones, lo que potencialmente aumentaría los costos y la complejidad del proceso. De ser así esta variable se destacaría como significativa en nuestra base de datos.

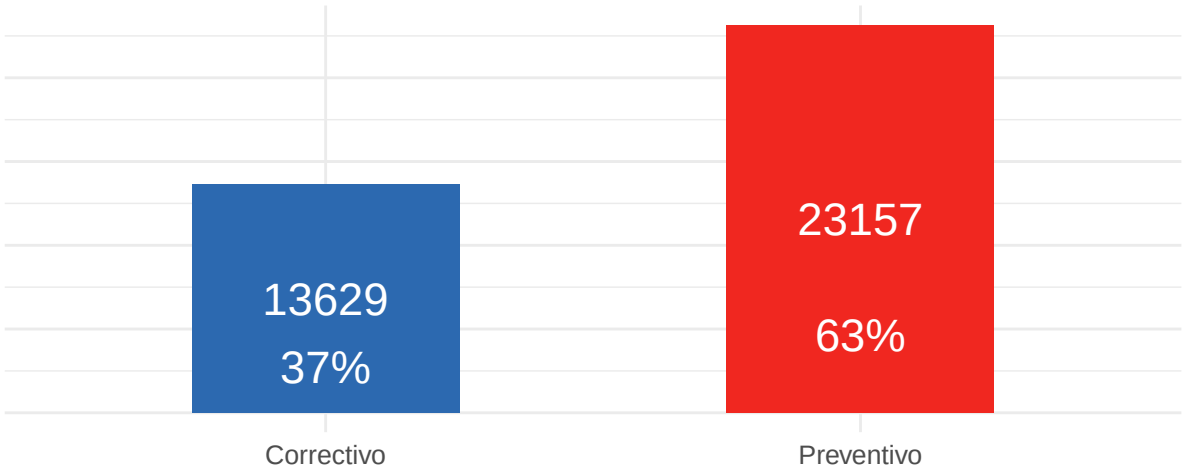

Figura 7: Distribución de reparación según el tipo de mantenimiento

El gráfico de barras ilustra una comparativa entre las actividades de mantenimiento correctivo y preventivo. Los modelos predictivos se centrarán en el mantenimiento correctivo, que representa el 37 % del total de las actividades, con 13,629 eventos registrados. Este enfoque permite especializarse en anticipar las necesidades de mantenimiento correctivo, crucial para mitigar tiempos de inactividad y optimizar recursos, a pesar de ser menos frecuente que el mantenimiento preventivo. Estos modelos tendrán un papel significativo en la mejora de la respuesta a fallos y en la eficiencia operativa general.

Disciplina

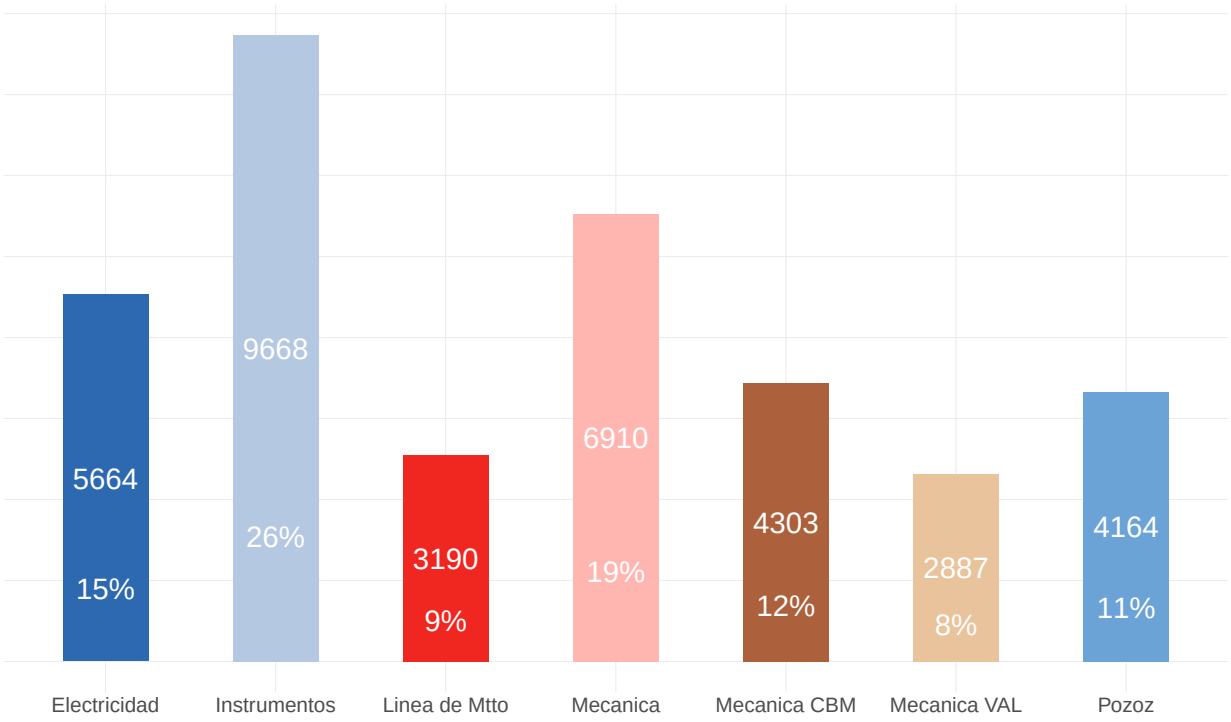

Figura 8: Distribución de reparaciones según disciplina

Instrumentos es la disciplina con el mayor número de reparaciones, con un total de 9,668, lo que representa el 26 % de las reparaciones en todas las disciplinas. Esto indica que es un área significativa de enfoque para el mantenimiento.

Electricidad también muestra una cantidad considerable de reparaciones, 5,664 en total, constituyendo el 15 % del total. Aunque menor que Instrumentos, sigue siendo una parte importante de las actividades de mantenimiento.

Línea de Mtto (Mantenimiento) tiene 3,190 reparaciones, que es el 9 % del total, mostrando que tiene menos reparaciones en comparación con Electricidad e Instrumentos.

Mecánica tiene 6,910 reparaciones, un 19 % del total, lo que la coloca como una disciplina significativa pero con menos incidencias que Instrumentos.

Mecánica CBM (Condition Based Maintenance) cuenta con 4,303 reparaciones, representando el 12 %, sugiriendo que esta práctica proactiva de mantenimiento está razonablemente establecida.

Mecánica VAL (Valor Agregado de Lubricación) tiene 2,887 reparaciones, que es el 8 % del total, lo cual podría indicar que es una disciplina especializada con menos frecuencia de reparaciones.

Pozos, por último, muestra 4,164 reparaciones, equivalentes al 11 % del total, situándose en un punto intermedio en términos de volumen de reparaciones.

Numero de horas en la maquina y su relación con otras variables

Análisis de correlación

Tabla 4

Correlación entre las variables numéricas

	estrategia Mtto	Programa Mtto	Ready back log	HH Actv Gnerales	HH En la Maquina	Otras HH
estrategia Mtto	1.00	0.15	-0.15	-0.02	0.00	-0.03
Programa Mtto	0.15	1.00	-0.35	-0.03	-0.01	-0.03
Ready back log	-0.15	-0.35	1.00	0.00	0.00	0.01
HH Actv Gnerales	-0.02	-0.03	0.00	1.00	0.46	0.50
HH En la Maquina	0.00	-0.01	0.00	0.46	1.00	0.44
Otras HH	-0.03	-0.03	0.01	0.50	0.44	1.00

Cada valor en la matriz representa el coeficiente de correlación entre dos variables específicas. La correlación puede variar en un rango de -1 a 1:

- Un valor de 1 indica una correlación positiva perfecta, lo que significa que las dos variables están perfectamente relacionadas de manera positiva (cuando una aumenta, la otra también aumenta en proporción constante).
- Un valor de -1 indica una correlación negativa perfecta, lo que significa que las dos variables están perfectamente relacionadas de manera negativa (cuando una aumenta, la otra disminuye en proporción constante).
- Un valor de 0 indica que no hay correlación lineal entre las dos variables.

Para la variable dependiente (HH en la maquina) ninguna de las correlaciones con otras variables es significativa en términos de magnitud (cercana a 1 o -1). Esto sugiere que "HH En la Maquina" no está fuertemente correlacionada con ninguna de las otras variables incluidas en la matriz de correlación. Sin embargo, es importante considerar que la ausencia de correlaciones significativas no necesariamente implica una falta de relación entre las variables; podría haber otras formas de relación que no están capturadas por la correlación lineal.

Análisis de densidad por época del año

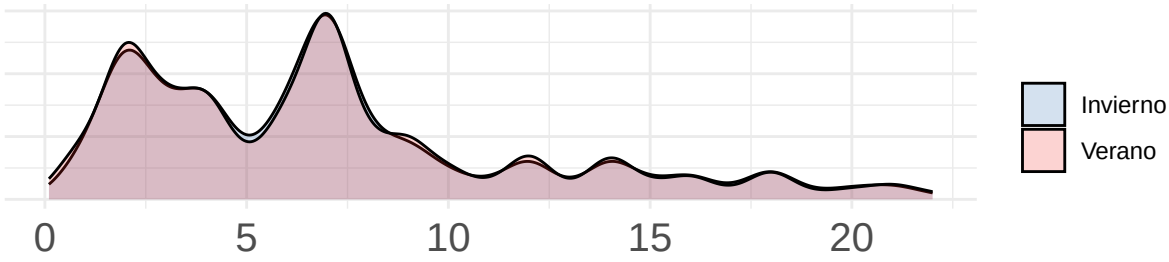

Figura 9: Densidad de horas en la maquina por época del año

No se detectan diferencias significativas al comparar el número de horas requeridas para la reparación de la máquina durante el verano y el invierno. Los datos revelan una variación mínima, indicando una consistencia en el tiempo de reparación independientemente de la estación. Este patrón se mantiene constante incluso cuando el número de horas de reparación es extremadamente alto o bajo. Estos resultados sugieren que las condiciones estacionales no ejercen un impacto significativo en el tiempo necesario para la reparación de la máquina, una conclusión que podría ser confirmada si la variable no resulta significativa en nuestros modelos.

Análisis de densidad por tipo de mantenimiento

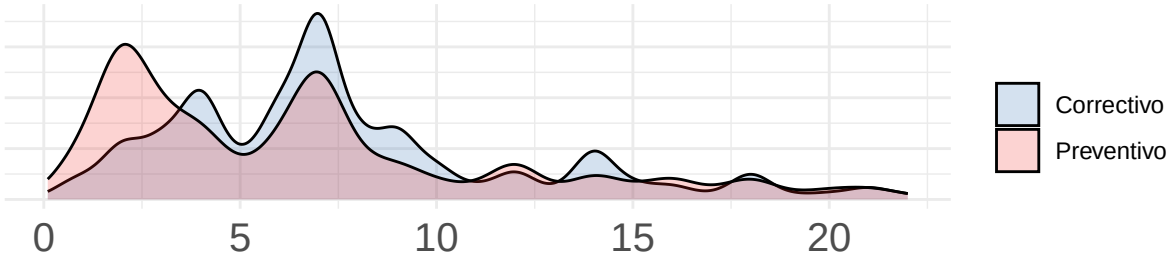

Figura 10: Densidad de horas en la maquina por tipo de mantenimiento

Se evidencia que conforme aumenta el número de horas necesarias para la reparación de la máquina, se observa una clara tendencia en la naturaleza del mantenimiento requerido. Los datos revelan que los mantenimientos correctivos son más frecuentes cuando se necesitan entre 5 y 10 horas para la reparación. Este tipo de mantenimiento reactivo suele ser necesario cuando se detectan problemas más significativos que requieren una intervención inmediata para restaurar el funcionamiento adecuado de la máquina.

Por otro lado, se observa que los mantenimientos preventivos son más comunes cuando las

reparaciones duran entre 0 y 5 horas. Estos mantenimientos planificados se realizan de manera regular y sistemática para evitar la aparición de problemas mayores y mantener el equipo en óptimas condiciones de funcionamiento. Su implementación oportuna puede reducir la necesidad de intervenciones correctivas costosas y prolongadas en el futuro.

Estos hallazgos subrayan la importancia de una gestión efectiva del mantenimiento, donde la combinación adecuada de mantenimiento preventivo y correctivo puede maximizar la disponibilidad y confiabilidad de la maquinaria, al tiempo que se minimizan los costos operativos y de reparación.

Análisis de densidad por disciplina

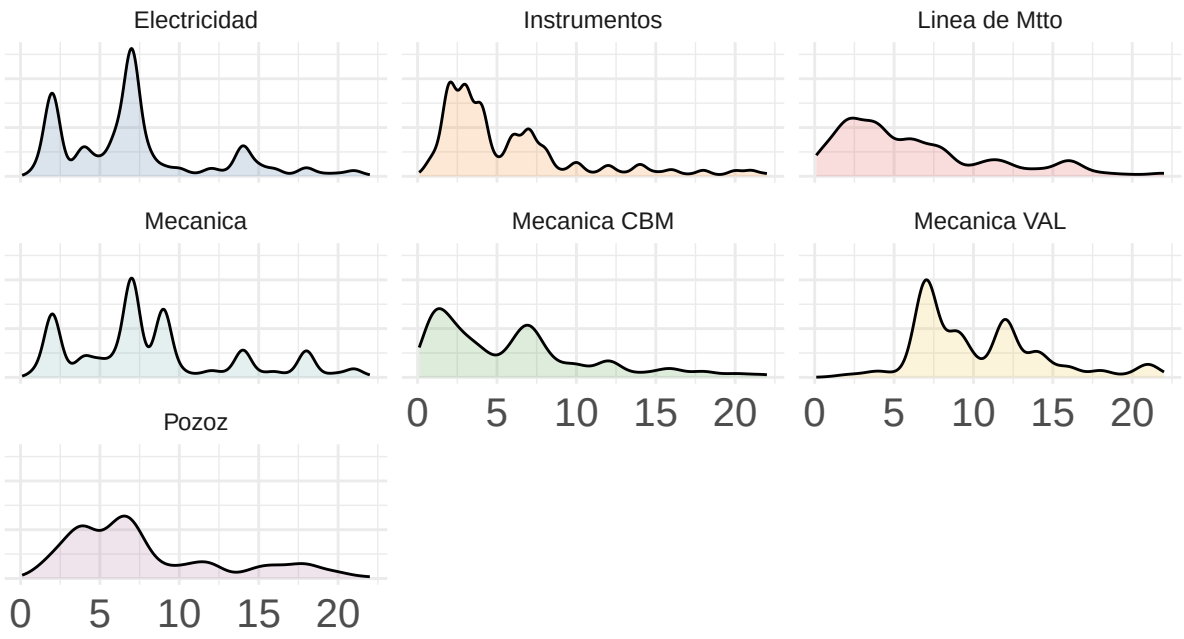

Figura 11: Analisis de dendsidad por disciplina

Se pueden observar distintas densidades según el tipo de disciplina, y es evidente que en la mayoría de los casos las horas de reparación son menores a 10 horas. Sin embargo, se destaca la disciplina VAL, donde los tiempos de reparación son significativamente mayores y se concentran en el rango de 5 a 15 horas. Esto sugiere que esta disciplina tiende a requerir un mayor número de horas para completar las reparaciones en comparación con otras disciplinas.

4.1.4. Modelado

En esta sección, se centra en el análisis de las técnicas de estadística multivariante discutidos previamente el marco teórico. Estos modelos serán aplicados en diversos contextos temporales, incluyendo el análisis diario, semanal y mensual. El objetivo de este espacio es profundizar en la comprensión y explicación de los resultados obtenidos a partir de dichos modelos en cada una de estas escalas temporales. Buscando así obtener una comprensión exhaustiva de sus dinámicas, así como de los impactos que todas las variables de la base de datos pueden tener en el número de horas necesarias para reparar las maquinas.

Cabe resaltar que para estos procesos se usa la herramienta R Studio y también la herramienta Pyhton para ajustar los modelos, los resultados y códigos se pueden consultar en un repositorio de GitHub que en la parte final se da el link.

4.1.4.1. Agrupación de los datos Para poder hacer un análisis de las horas requeridas para mantenimiento correctivo en diferentes periodos de tiempo, fue necesario agrupar la base de datos en un rango mensual, semanal y diario esto con el fin de predecir los tiempos semanales y mensuales, se hace porque en lo general son los periodos que se miden en las empresas dedicadas a estas tareas.

Base de datos agrupada de manera diaria

En esta base de datos, organizamos la información agrupándola por día y también según la temporada del año y la disciplina de reparación. El objetivo es calcular la media de cada variable numérica.

Base de datos agrupada de manera semanal

Inicialmente creamos una variable llamada “semana”, que consiste en la combinación del año extraído de la fecha de ejecución utilizando la función *year*, junto con la semana extraída de la misma fecha mediante la función *week*. Posteriormente, agrupamos los datos por esta variable, la época del año y la disciplina.

Base de datos agrupada de manera mensual

Inicialmente creamos una variable llamada “mes”, que consiste en la combinación del año extraído de la fecha de ejecución utilizando la función *year*, junto con el mes extraído de la misma fecha mediante la función *month*. Posteriormente, agrupamos los datos por esta variable, la época del año y la disciplina.

Tabla 5

Base de Datos agrupada de manera diaria

fecha Ejecutada	Disciplina	Epoca del año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
2019-01-02	Instrumentos	Verano	0.98	0.95	5.16	0.40	4.40	1.00
2019-01-04	Instrumentos	Verano	0.98	0.95	5.16	0.30	1.00	0.00
2019-01-08	Instrumentos	Verano	0.98	0.95	5.16	0.59	6.73	1.54
2019-01-08	Pozoz	Verano	0.98	0.95	5.16	0.65	4.43	2.43
2019-01-09	Instrumentos	Verano	0.98	0.95	5.16	0.49	6.37	1.52
2019-01-09	Mecanica	Verano	0.98	0.95	5.16	0.50	5.00	0.60
2019-01-09	Pozoz	Verano	0.98	0.95	5.16	0.81	4.80	2.70
2019-01-11	Electricidad	Verano	0.98	0.95	5.16	0.90	8.80	1.55
2019-01-11	Instrumentos	Verano	0.98	0.95	5.16	0.83	7.71	2.19
2019-01-11	Mecanica	Verano	0.98	0.95	5.16	0.50	7.00	1.50
2019-01-14	Instrumentos	Verano	0.98	0.95	5.16	0.50	7.00	5.20
2019-01-15	Instrumentos	Verano	0.98	0.95	5.16	0.50	8.50	0.50
2019-01-16	Pozoz	Verano	0.98	0.95	5.16	0.40	1.00	1.60
2019-01-17	Instrumentos	Verano	0.98	0.95	5.16	0.80	3.00	0.60
2019-01-18	Instrumentos	Verano	0.98	0.95	5.16	0.80	2.80	1.10

Tabla 6

Base de Datos agrupada de manera semanal

semana	Disciplina	Epoca del año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
2019-1	Instrumentos	Verano	0.98	0.95	5.16	0.37	3.27	0.67
2019-10	Pozoz	Verano	0.99	0.92	4.67	0.30	1.80	1.90
2019-11	Instrumentos	Verano	0.99	0.92	4.67	0.20	0.50	0.00
2019-12	Instrumentos	Verano	0.99	0.92	4.67	1.05	8.00	1.55
2019-12	Mecanica	Verano	0.99	0.92	4.67	0.70	9.00	1.30
2019-12	Pozoz	Verano	0.99	0.92	4.67	0.60	3.30	2.20
2019-15	Instrumentos	Verano	0.99	0.94	5.65	0.05	6.10	0.50
2019-16	Instrumentos	Verano	0.99	0.94	5.65	0.57	8.33	1.93
2019-17	Instrumentos	Verano	0.99	0.94	5.65	0.50	6.50	0.90
2019-17	Pozoz	Verano	0.99	0.94	5.65	0.50	16.00	1.50
2019-18	Instrumentos	Invierno	0.99	0.94	5.96	0.50	10.70	1.50
2019-18	Mecanica VAL	Invierno	0.99	0.94	5.96	1.50	14.00	2.50
2019-19	Electricidad	Invierno	0.99	0.94	5.96	0.10	2.00	0.35
2019-19	Pozoz	Invierno	0.99	0.94	5.96	0.50	16.00	1.50
2019-2	Electricidad	Verano	0.98	0.95	5.16	0.90	8.80	1.55

Tabla 7

Base de Datos agrupada de manera mensual

mes	Disciplina	Epoca del año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
2019-1	Electricidad	Verano	0.98	0.95	5.16	0.74	10.09	2.39
2019-1	Instrumentos	Verano	0.98	0.95	5.16	0.57	6.65	1.55
2019-1	Mecanica	Verano	0.98	0.95	5.16	0.50	10.83	1.20
2019-1	Mecanica CBM	Verano	0.98	0.95	5.16	0.50	8.00	0.60
2019-1	Pozoz	Verano	0.98	0.95	5.16	0.69	4.51	2.49
2019-10	Electricidad	Verano	0.99	0.93	5.91	0.50	7.00	5.50
2019-10	Instrumentos	Verano	0.99	0.93	5.91	0.41	5.78	1.11
2019-10	Linea de Mtto	Verano	0.99	0.93	5.91	0.71	7.40	2.43
2019-10	Mecanica	Verano	0.99	0.93	5.91	1.45	15.20	1.35
2019-10	Mecanica VAL	Verano	0.99	0.93	5.91	0.00	12.00	0.00
2019-10	Pozoz	Verano	0.99	0.93	5.91	0.73	4.24	2.59
2019-11	Electricidad	Verano	0.99	0.92	5.89	1.80	7.00	5.40
2019-11	Mecanica	Verano	0.99	0.92	5.89	0.83	9.98	2.15
2019-11	Mecanica CBM	Verano	0.99	0.92	5.89	0.35	11.27	1.46
2019-12	Mecanica VAL	Verano	0.99	0.94	5.46	1.60	7.00	1.50

4.1.4.2. Desarrollo de las predicciones Importante tener presente que nos enfrentamos a la tarea de modelar tres variaciones temporales distintas (mensual, semanal y diaria), son las frecuencias de tiempo populares para emitir planes de mantenimiento en el sector industrial, como se explica anteriormente la base de datos se agrupa con estos aspectos para lograr las predicciones en estos intervalos de tiempo. Utilizando los modelos discutidos en el marco teórico, donde el objetivo radica en comprender y anticipar cómo se comporta la variable dependiente con cambios en las variables independientes a lo largo de estas diferentes escalas temporales. Una vez que se ajustan los modelos a cada una de estas variaciones temporales, es

importante evaluar su rendimiento y capacidad explicativa de los datos. Para ello se recurre a una serie de evaluación (Métricas), tales como el error cuadrático medio (MSE), Raíz del error cuadrático medio (RMSE) y Error absoluto medio (MAE) Estos criterios permiten comparar los modelos entre sí y determinar cuál de ellos se ajusta mejor a los datos, teniendo en cuenta la complejidad inherente de cada modelo.

4.1.4.3. Segmentación de datos Con respecto a los datos de entrenamiento y los datos para test, de acuerdo a la literatura los expertos recomiendan el uso de la regla 80/20 lo que puede ser un buen punto de partida, pero todo depende del tamaño y las características del conjunto de datos, por otra parte, también se recomienda el uso de la validación cruzada, que es una técnica más robusta que consiste en dividir los datos en k subconjuntos lo que permite evaluar el modelo de forma más integral, otras técnicas como la estratificación, división aleatoria, tamaño mínimo del conjunto de la prueba, en este último sugieren que el conjunto de prueba debe tener al menos de 50-100 observaciones, dependiendo de la complejidad del modelo, para obtener estimaciones confiables del desempeño. Para este caso de estudio caso se dividen los datos en un 90 % para entrenamiento y un 10 % para realizar las pruebas en todos los modelos entrenados, adicionalmente se hace una práctica en Python donde también se utiliza la validación cruzada.

Modelos a entrenar En este análisis, se exploran diferentes modelos estadísticos para predecir la variable HH_En_la_Maquina utilizando un conjunto de datos de entrenamiento diario llamado (traindiario).

Modelo de regresión múltiple, se inicia con los ajustes del modelo lineal utilizando la función "LM" de R Studio. Este modelo considera una relación lineal entre la variable objetivo y las variables predictoras en el conjunto de datos, después se ven las diferentes métricas y la relación entre variables, donde con la función de resumen se puede verificar si las variables son significativas y afectan al modelo, también se presenta el p-value y el r cuadrado ajustado, que son metricas para evaluar que también se ajustan los datos al modelo. Adicionalmente se aplican los modelos lineales generalizados mediante la función "GLM". Estos modelos permiten modelar relaciones entre variables predictoras y la respuesta que puede no ser lineales, y también pueden manejar variables de respuesta con distribuciones no normales.

Continuando con los casos, se ajusta un modelo de regresión utilizando el método de Regresión de Soporte Vectorial (SVR), que es un algoritmo de Maching learning para temas de regresión y Clasificación, modelo que busca encontrar la mejor función lineal para predecir la variable objetivo "HH_En_la_Maquina", utilizando una función de pérdida epsilon-insensible. se explora también los Bosques Aleatorios "Randon Forest", un método de aprendizaje conjunto que construye múltiples árboles de decisión durante el entrenamiento y produce la predicción promedio de los árboles individuales.

Finalmente, se ajusta un modelo de Árboles de Decisión utilizando el método de partición recursiva (RPART), que divide iterativamente los datos en subconjuntos basados en ciertos criterios, buscando maximizar la homogeneidad dentro de los grupos resultantes.

Todos estos modelos mencionados se ajustan en R-Studio para tener un panorama general de las métricas y elegir cual es más optimo para realizar las predicciones propuestas en el objetivo de este trabajo, por otra parte, se elige la predicción semanal y los modelos más eficientes para hacer un análisis más robusto en Python donde se utiliza la validación cruzada y otras técnicas, toda la información y resultados se anexa en el repositorio de datos en

GitHub.

4.1.4.4. Predicciones diarias Después de dividir los datos en entrenamiento y test se presentan las tablas con los resultados para los dos casos.

Tabla 8

Base de Datos entrenamiento diario

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Mecanica	Verano	0.96	0.96	5.87	0.70	9.50	2.13
Mecanica	Verano	0.99	0.94	5.37	0.65	4.50	0.75
Pozoz	Verano	0.99	0.91	5.93	1.00	7.33	2.02
Instrumentos	Verano	1.00	0.97	6.48	0.30	15.00	2.00
Linea de Mtto	Verano	0.99	0.91	5.93	0.00	0.50	0.00
Electricidad	Invierno	0.99	0.96	5.07	0.50	7.00	1.50
Linea de Mtto	Verano	1.00	0.95	5.04	0.80	15.40	2.78
Mecanica VAL	Invierno	0.98	0.97	5.31	0.50	7.00	1.50
Pozoz	Invierno	0.99	0.96	4.80	0.76	7.80	3.54
Linea de Mtto	Invierno	1.00	0.96	4.64	1.80	10.00	1.50
Mecanica VAL	Verano	1.00	0.98	5.00	0.50	7.71	2.59
Instrumentos	Verano	0.98	0.99	4.47	0.33	5.92	1.18
Mecanica	Invierno	0.99	0.96	4.70	0.72	10.56	1.52
Electricidad	Verano	1.00	0.96	6.25	0.75	15.00	3.30
Mecanica	Invierno	1.00	0.97	4.70	0.50	7.00	1.50

Tabla 9

Base de Datos de prueba diario

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Instrumentos	Verano	0.98	0.95	5.16	0.59	6.73	1.54
Instrumentos	Verano	0.98	0.95	5.16	0.80	2.80	1.10
Mecanica	Verano	0.99	0.94	5.82	1.40	9.00	2.07
Mecanica	Verano	0.99	0.94	5.82	0.80	9.00	1.40
Pozoz	Verano	0.99	0.94	5.65	0.50	16.00	1.50
Pozoz	Invierno	0.99	0.94	5.96	0.50	7.00	1.50
Instrumentos	Invierno	0.99	0.94	5.78	0.46	5.88	1.24
Pozoz	Invierno	0.99	0.95	5.75	0.54	4.20	2.56
Instrumentos	Invierno	0.99	0.91	5.17	0.30	2.00	0.00
Instrumentos	Invierno	0.99	0.94	5.88	0.32	4.88	0.90
Instrumentos	Invierno	0.99	0.94	5.88	0.38	5.50	0.88
Instrumentos	Verano	0.99	0.93	5.91	0.49	8.55	1.60
Electricidad	Verano	0.99	0.94	5.39	0.50	13.30	2.30
Linea de Mtto	Verano	0.99	0.94	5.39	0.58	7.25	2.31
Electricidad	Verano	0.99	0.94	5.39	0.43	6.79	2.05

Modelo elegir

- El modelo Random Forest tiene los valores más bajos de RMSE y MSE, lo que indica que generalmente ha tenido un mejor desempeño en términos de minimizar los errores cuadrados. También tiene el MAE más bajo, lo que sugiere que tiene una buena precisión en términos de errores absolutos.
- Los modelos Lineal y GLM tienen valores idénticos para todas las métricas, lo que podría indicar que tienen un rendimiento muy similar en este conjunto de datos o que podrían estar utilizando la misma fórmula subyacente para la predicción.
- El SVR y el arbol tienen un rendimiento ligeramente mejor que los modelos lineales en términos de RMSE y MSE, pero no superan al Random Forest.

Estos resultados sugieren que el modelo de Random Forest es el más adecuado para los datos y el problema en cuestión, seguido por el modelo SVR, el árbol de decisión y por último, los modelos lineales y GLM que presentan un rendimiento similar entre sí.

Tabla 10

Comparación de los modelos (base de datos diaria)

MODELO	RMSE	MSE	MAE
Lineal	3.4628	11.9911	2.4876
GLM	3.4628	11.9911	2.4876
SVR	3.5079	12.3056	2.4341
ArbolD	3.5014	12.2597	2.5332
Ramdo F	3.2301	10.4337	2.3111

Comparación entre valores reales y predichos

La Tabla 10 presenta una comparativa detallada entre los valores estimados por un modelo de regresión Random Forest y las cifras reales observadas.

El MAE de 2.31 indica que, en promedio, las predicciones del modelo se desvían alrededor de 2.31 horas de los tiempos reales observados. Esta discrepancia se evidencia claramente en los datos, donde las horas estimadas a menudo difieren de las horas reales dedicadas a las actividades. Esta tendencia sugiere una posible sobrevaloración sistemática por parte del modelo, lo que lleva a la asignación excesiva o insuficiente de recursos por parte de la empresa. Esta discrepancia media implica que, en promedio, la empresa podría estar asignando 2.31 horas más o menos de lo necesario para completar las actividades. Esta ineficiencia en el uso de recursos, tanto económicos como humanos, subraya la necesidad de ajustar el modelo para optimizar la planificación y el despliegue de recursos.

Tabla 11

Predichos vs reales (base de datos diaria)

predichos	reales
9.05	6.73
6.38	2.80
12.01	9.00
8.42	9.00
13.16	16.00
13.63	7.00
6.87	5.88
5.51	4.20
4.82	2.00
6.17	4.88

4.1.4.5. Predicción de manera semanal Después de dividir los datos en entrenamiento y test se presentan las tablas con los resultados para los dos casos.

Tabla 12

Base de Datos entrenamiento semanal

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Mecanica VAL	Invierno	1.00	0.97	4.78	0.81	12.64	2.15
Instrumentos	Verano	1.00	0.95	4.84	0.25	3.86	0.40
Pozoz	Invierno	0.99	0.95	5.31	0.78	5.51	1.99
Mecanica CBM	Verano	0.99	0.96	4.87	0.20	8.00	1.00
Pozoz	Invierno	0.99	0.94	5.31	0.97	7.86	3.16
Mecanica	Verano	0.96	0.99	4.06	0.40	8.00	1.35
Electricidad	Invierno	0.99	0.96	4.80	0.68	9.38	2.10
Pozoz	Verano	0.99	0.93	5.31	0.67	5.33	1.63
Pozoz	Verano	1.00	0.96	5.71	0.96	6.20	2.72
Pozoz	Invierno	0.95	0.94	6.20	0.89	7.95	2.07
Electricidad	Verano	0.96	0.96	5.87	0.58	7.83	1.43
Pozoz	Invierno	0.99	0.94	5.96	0.50	16.00	1.50
Instrumentos	Invierno	0.99	0.92	4.87	0.64	9.07	2.77
Pozoz	Invierno	1.00	0.93	4.61	0.58	5.00	1.90
Mecanica CBM	Invierno	0.99	0.95	5.30	0.50	10.00	1.10

Tabla 13

Base de Datos de prueba semanal

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Instrumentos	Verano	0.98	0.95	5.16	0.59	6.73	1.54
Instrumentos	Verano	0.98	0.95	5.16	0.80	2.80	1.10
Mecanica	Verano	0.99	0.94	5.82	1.40	9.00	2.07
Mecanica	Verano	0.99	0.94	5.82	0.80	9.00	1.40
Pozoz	Verano	0.99	0.94	5.65	0.50	16.00	1.50
Pozoz	Invierno	0.99	0.94	5.96	0.50	7.00	1.50
Instrumentos	Invierno	0.99	0.94	5.78	0.46	5.88	1.24
Pozoz	Invierno	0.99	0.95	5.75	0.54	4.20	2.56
Instrumentos	Invierno	0.99	0.91	5.17	0.30	2.00	0.00
Instrumentos	Invierno	0.99	0.94	5.88	0.32	4.88	0.90
Instrumentos	Invierno	0.99	0.94	5.88	0.38	5.50	0.88
Instrumentos	Verano	0.99	0.93	5.91	0.49	8.55	1.60
Electricidad	Verano	0.99	0.94	5.39	0.50	13.30	2.30
Linea de Mtto	Verano	0.99	0.94	5.39	0.58	7.25	2.31
Electricidad	Verano	0.99	0.94	5.39	0.43	6.79	2.05

Modelo a elegir

- El modelo Random Forest, tiene los valores más bajos de RMSE y MSE, lo que indica que generalmente ha tenido un mejor desempeño en términos de minimizar los errores cuadrados. También tiene el MAE más bajo, lo que sugiere que tiene una buena precisión en términos de errores absolutos.
- Los modelos lineales y el lineal Generalizado tienen valores idénticos para todas las métricas, lo que podría indicar que tienen un rendimiento muy similar en este conjunto de datos o que podrían estar utilizando la misma fórmula subyacente para la predicción. El SVR y el Árbol de Decisión tienen un rendimiento ligeramente mejor que los modelos lineales en términos de RMSE y MSE, pero no superan al Random Forest.

Estos resultados sugieren que el modelo de Random Forest es el más adecuado para los datos y el problema en cuestión, seguido por el modelo SVR, el árbol de decisión y, por último, los modelos lineales y GLM que presentan un rendimiento similar entre sí.

Tabla 14

Comparación de los modelos (base de datos semana)

MODELO	RMSE	MSE	MAE
Lineal	3.1227	9.7511	2.2028
GLM	3.1227	9.7511	2.2028
SVR	3.1029	9.6279	2.1338
ArboD	3.0820	9.4985	2.2085
Randon F	2.7535	7.5819	1.9762

Comparación entre valores reales y predichos

La Tabla 14, presenta una comparativa detallada entre los valores estimados por un modelo de regresión Random Forest y las cifras reales observadas.

El MAE de 1.97 indica que, en promedio, las predicciones del modelo se desvían alrededor de 1.97 horas de los tiempos reales observados.

Adicionalmente para este caso se justan los modelos en Python donde los valores de las metricas son similares, en Python se ajustan los hiperparametros y se hacen caclulan las metricas con valiación cruzada. Los resultados y conclusiones estan en el repositorio de datos.

Tabla 15

Predichos vs reales (base de datos semanal)

predichos	reales
5.42	3.27
4.86	0.50
3.95	6.10
12.62	14.00
7.03	6.41
6.42	6.00
5.75	5.53
9.69	16.90
8.62	6.71
8.61	8.00

4.1.4.6. Predicción de manera mensual Después de dividir los datos en entrenamiento y test se presentan las tablas con los resultados para los dos casos

Tabla 16

Base de Datos entrenamiento mensual

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Pozoz	Invierno	0.99	0.92	4.87	0.89	7.51	2.22
Mecanica CBM	Verano	0.99	0.92	5.89	0.35	11.27	1.46
Instrumentos	Invierno	1.00	0.97	4.70	0.44	5.99	0.79
Linea de Mtto	Verano	0.98	0.99	4.47	0.50	14.50	1.43
Electricidad	Invierno	0.95	0.94	6.20	0.84	10.96	2.58
Instrumentos	Verano	0.96	0.99	4.06	0.56	9.83	2.17
Mecanica	Verano	0.99	0.96	4.55	0.83	9.47	2.09
Electricidad	Verano	1.00	0.98	4.75	0.60	8.16	2.23
Mecanica VAL	Invierno	0.98	0.97	5.31	0.69	10.88	1.82
Instrumentos	Verano	1.00	0.94	5.63	0.48	6.39	0.94
Pozoz	Verano	0.99	0.93	5.31	0.79	5.14	2.75
Electricidad	Invierno	0.99	0.93	6.68	0.70	8.06	2.02
Mecanica VAL	Invierno	0.99	0.96	5.07	0.85	10.00	2.61
Mecanica	Invierno	1.00	0.97	4.70	0.70	9.83	1.85
Linea de Mtto	Invierno	0.99	0.95	4.88	0.74	10.39	1.76

Tabla 17

Base de Datos de prueba mensual

Disciplina	Epoca_del_año	Cumplimiento_Estrategia	Cumplimiento_Programa	Ready_Backlog	HH_Actv_Gnerales	HH_En_la_Maquina	Otras_HH
Instrumentos	Verano	0.99	0.94	5.82	0.44	8.33	1.46
Mecanica	Verano	0.99	0.94	5.82	1.36	8.40	1.78
Linea de Mtto	Verano	0.99	0.94	5.37	0.42	5.59	1.92
Pozoz	Verano	0.99	0.94	5.37	0.66	4.24	1.74
Mecanica VAL	Verano	1.00	0.96	5.71	0.72	10.39	2.30
Instrumentos	Verano	0.99	0.94	5.39	0.55	7.57	1.36
Mecanica VAL	Verano	0.99	0.94	5.39	0.74	10.30	2.31
Linea de Mtto	Verano	0.99	0.91	5.93	0.14	2.72	0.50
Mecanica VAL	Invierno	0.99	0.93	6.68	0.82	9.77	2.37
Linea de Mtto	Invierno	0.99	0.95	5.31	0.95	5.75	3.92
Pozoz	Invierno	0.95	0.94	6.20	0.86	7.86	2.24
Mecanica	Verano	1.00	0.97	6.48	0.79	9.79	2.06
Pozoz	Verano	1.00	0.97	6.48	0.71	5.94	2.15
Pozoz	Verano	1.00	0.95	4.84	0.78	4.94	1.82
Instrumentos	Verano	1.00	0.97	5.28	0.29	5.72	0.46

Modelo a elegir

*El modelo Random Forest, tiene los valores más bajos de RMSE y MSE, lo que indica que generalmente ha tenido un mejor desempeño en términos de minimizar los errores cuadrados. También tiene el MAE más bajo, lo que sugiere que tiene una buena precisión en términos de errores absolutos.

- Los modelos lineales y el modelo lineal generalizado tienen valores idénticos para todas las métricas, lo que podría indicar que tienen un rendimiento muy similar en este conjunto de datos o que podrían estar utilizando la misma fórmula subyacente para la predicción. El svr y el arbol tienen un rendimiento ligeramente mejor que los modelos lineales en términos de RMSE y MSE, pero no superan al rf.

Estos resultados sugieren que el modelo de Random Forest es el más adecuado para los datos y el problema en cuestión, seguido por el modelo SVR, el árbol de decisión y, por último, los modelos lineales y GLM que presentan un rendimiento similar entre sí.

Tabla 18

Comparación de los modelos (base de datos mensual)

MODELO	RMSE	MSE	MAE
Lineal	1.8098	3.2752	1.3685
GLM	1.8098	3.2752	1.3685
SVR	1.7908	3.2071	1.3442
Arbold	1.9473	3.7919	1.5777
Random F	1.7572	3.0879	1.3117

Comparación entre valores reales y predichos

La Tabla 18 presenta una comparativa detallada entre los valores estimados por un modelo de regresión Random Forest y las cifras reales observadas.

El MAE de 1.31 indica que, en promedio, las predicciones del modelo se desvían alrededor de 2.38 horas de los tiempos reales observados.

Tabla 19

Predichos vs reales (base de datos mensual)

predichos	reales
7.35	8.33
11.58	8.40
8.44	5.59
6.55	4.24
9.99	10.39
7.43	7.57
10.25	10.30
4.69	2.72
10.62	9.77
9.92	5.75

4.1.5. Validación

Los resultados muestran las métricas de evaluación (RMSE, MSE y MAE) para tres variaciones temporales: diaria, semanal y mensual. Cada variación representa diferentes frecuencias de muestreo de los datos, lo que puede afectar el rendimiento de los modelos de predicción.

Diaria

- **RMSE**: Los valores de RMSE oscilan entre 3.2301 y 3.5079, lo que indica que los modelos tienen un error promedio de predicción diaria en el rango de 3.23 a 3.50 unidades.
- **MSE**: Los valores de MSE varían entre 10.4337 y 12.3056, lo que indica que los modelos tienen un error cuadrático promedio de predicción diaria en el rango de 10.43 a 12.31 unidades al cuadrado.
- **MAE**: Los valores de MAE oscilan entre 2.311 y 2.5332, lo que indica que los modelos tienen un error absoluto promedio de predicción diaria en el rango de 2.31 a 2.53 unidades.

Semanal

- **RMSE**: Los valores de RMSE oscilan entre 2.7535 y 3.1227, lo que indica que los modelos tienen un error promedio de predicción semanal en el rango de 2.75 a 3.12 unidades.
- **MSE**: Los valores de MSE varían entre 7.5819 y 9.7511, lo que indica que los modelos tienen un error cuadrático promedio de predicción semanal en el rango de 7.58 a 9.75 unidades al cuadrado.
- **MAE**: Los valores de MAE oscilan entre 1.9762 y 2.2085, lo que indica que los modelos tienen un error absoluto promedio de predicción semanal en el rango de 1.98 a 2.21 unidades.

Mensual

- **RMSE**: Los valores de RMSE oscilan entre 1.7572 y 1.9473, lo que indica que los modelos tienen un error promedio de predicción mensual en el rango de 1.76 a 1.95 unidades.
- **MSE**: Los valores de MSE varían entre 3.0879 y 3.7919, lo que indica que los modelos tienen un error cuadrático promedio de predicción mensual en el rango de 3.09 a 3.79 unidades al cuadrado.
- **MAE**: Los valores de MAE oscilan entre 1.3117 y 1.5777, lo que indica que los modelos tienen un error absoluto promedio de predicción mensual en el rango de 1.31 a 1.58 unidades.

En general, se observa que los modelos tienden a tener un mejor rendimiento en la predicción a mayor escala temporal (de diaria a mensual), lo que puede deberse a la reducción de la variabilidad en los datos a medida que se agrupan en intervalos de tiempo más amplios. Además, algunos modelos pueden ser más robustos en ciertas escalas temporales que en otras, como se observa en las variaciones en las métricas entre los modelos en cada escala temporal.

5. Conclusiones

La exploración y entendimiento de los datos resulta ser una clave importante a la hora de emprender a realizar un trabajo de esta categoría, cuando se tiene un panorama general del para que o que es lo que se requiere, resulta más sencillo encontrar y plantear soluciones, como este caso que trata del mantenimiento Industrial. El análisis descriptivo de los datos es

fundamental para entender mejor el problema a solucionar y todo lo que implica limpieza de los datos no solo hace que los modelos sean más efectivos a la hora de predecir, también ayudan a entender el problema a solucionar. Resaltar que un buen analista de datos debe comprender muy bien el problema a solucionar, porque el no hacerlo puede generar retrabajos que atrasan el proyecto, por tanto, también es importante contar con personas expertas en el tema, que sirvan de apoyo ante dudas con los datos que se disponen para hacer los análisis, entender los datos y un buen tratamiento para estos es un punto a favor que conlleva a resultados más acertados.

Al aplicar técnicas de estadística multivariante con el objetivo de desarrollar modelos predictivos de los tiempos de mantenimiento correctivo, modelos que se conocen como clásicos, se puede concluir que estos son sencillos de aplicar y sus resultados se pueden interpretar más fácilmente. Claro que para ajustarlos toca cumplir ciertos supuestos como la linealidad, homocedasticidad y normalidad de los residuos, lo cual puede ser desafiante en algunos conjuntos de datos como el planteado en este caso.

Por otra parte, los algoritmos de Maching Learning para Regresión como Random Forest o SVR, tienen una mayor flexibilidad para capturar relaciones no lineales entre las variables predictoras y la variable de respuesta, lo que puede ser útil cuando existen patrones complejos en los datos de mantenimiento, caso que los modelos de regresión lineales tradicionales podrían no ser capaces de modelar adecuadamente, también pueden detectar automáticamente interacciones y efectos no lineales entre las variables, características de aprendizaje automático y adaptabilidad lo que lleva a tener una mayor precisión predictiva.

Todo lo anterior se ve demostrado en los resultados obtenidos, donde las diferentes métricas indican que el mejor modelo para predecir los tiempos correctivos de mantenimiento es el ajustado con el Algoritmo de Random Forest, que no corresponde a los modelos clásicos estudiados.

5.1. Resultados

Al examinar los resultados del modelo de Random Forest, se destaca su desempeño, especialmente en el análisis mensual, donde registra un MAE de 1.31 horas. Este valor más bajo indica una mayor precisión en la predicción del tiempo requerido para realizar mantenimientos correctivos en un periodo prolongado. La consistencia en este rendimiento sugiere que el modelo es especialmente eficaz cuando se consideran períodos extensos.

La ventaja del análisis mensual puede atribuirse a varios factores. En primer lugar, al agrupar datos en un periodo más largo, se pueden eliminar ciertas fluctuaciones o variaciones aleatorias que podrían afectar las predicciones diarias o semanales. Además, al analizar tendencias durante un mes, el modelo puede capturar patrones más estables y significativos en los datos, lo que resulta en predicciones más precisas.

Por lo tanto, al planificar actividades de mantenimiento a largo plazo, el análisis mensual proporcionado por el modelo de Random Forest emerge como la opción más confiable y precisa.

El modelo de Random Forest sobresale entre otros enfoques de modelado, como modelos lineales, GLM (Modelos Lineales Generalizados), SVR (Máquinas de Vectores de Soporte para Regresión) y árboles de decisión, por varias razones:

Robustez frente a datos complejos: Mientras que los modelos lineales y GLM son eficientes bajo ciertas suposiciones sobre la linealidad y la independencia de las variables, los datos en la práctica suelen ser más complejos, exhibiendo relaciones no lineales que podrían limitar la capacidad de estos modelos para capturar patrones subyacentes. En contraste, el modelo de Random Forest, al emplear múltiples árboles de decisión y combinar sus predicciones, puede manejar con mayor eficacia esta complejidad inherente en los datos.

Capacidad para manejar multicolinealidad y características no lineales: En escenarios donde las variables predictoras están altamente correlacionadas o muestran relaciones no lineales con la variable objetivo, los modelos lineales y GLM pueden enfrentar dificultades para capturar estas relaciones de manera efectiva. Por el contrario, los árboles de decisión, incluido el modelo de Random Forest, poseen la capacidad inherente de manejar multicolinealidad y relaciones no lineales sin necesidad de preprocesamiento adicional de datos.

Regularización implícita: A diferencia de los modelos lineales y SVR, que a menudo requieren técnicas de regularización para evitar el sobreajuste, el modelo de Random Forest tiende a ser menos propenso a este fenómeno debido a la naturaleza de ensamblaje de múltiples árboles de decisión. Esto implica que el modelo puede generalizar mejor a datos no vistos, lo que se traduce en predicciones más robustas y confiables.

Manejo efectivo de variables categóricas y no lineales: Los árboles de decisión, incluido Random Forest, son intrínsecamente capaces de manejar variables categóricas sin necesidad de transformaciones adicionales, lo que simplifica el proceso de modelado y reduce la necesidad de preprocesamiento de datos. Además, estos modelos pueden identificar interacciones no lineales entre variables predictoras, permitiendo la captura de patrones más complejos en los datos.

5.2. Dificultades

La predicción del tiempo necesario para reparar máquinas es fundamental en la gestión eficiente del mantenimiento industrial. Sin embargo, este proceso presenta una serie de desafíos que deben abordarse para desarrollar modelos predictivos precisos y útiles.

Uno de los principales desafíos es la variabilidad inherente en el tiempo de reparación, que puede ser influenciada por una multitud de factores, como la complejidad de la falla, la disponibilidad de repuestos y la experiencia del técnico Montgomery et al. (2012). Además, la falta de datos completos sobre reparaciones anteriores puede dificultar la construcción de modelos predictivos robustos.

La relación entre las variables predictoras y el tiempo de reparación también puede ser no lineal, lo que requiere el uso de técnicas avanzadas de modelado, como modelos no lineales y de regresión robustos Gelman et al. (2013). Además, los efectos aleatorios, como la variabilidad entre técnicos o máquinas, pueden introducir una fuente adicional de variabilidad en el modelo.

Para abordar estos desafíos, se pueden emplear una variedad de enfoques estadísticos. Por ejemplo, los modelos mixtos o jerárquicos pueden capturar efectos aleatorios y variabilidad no observada Agresti (2015), mientras que las técnicas de imputación de datos pueden ayudar a manejar la falta de datos completos. Asimismo, los métodos de validación cruzada y las técnicas de selección de modelos pueden utilizarse para evaluar y comparar diferentes modelos y determinar el más adecuado para un conjunto de datos dado.

A pesar de los desafíos, el modelado preciso del tiempo de reparación de máquinas puede tener un impacto significativo en la eficiencia operativa y la planificación del mantenimiento, lo que hace que valga la pena abordar estos desafíos con determinación y utilizando enfoques estadísticos adecuados.

5.3. Trabajos futuros

Existen diferentes investigaciones que buscan como predecir los tiempos necesarios para realizar un mantenimiento correctivo, (Citados en los antecedentes del proyecto), del mismo modo hoy en día el mantenimiento predictivo está tomando más fuerza y consiste en predecir antes de hacer o como se dice en el campo de la industria basado en una condición o condiciones del equipo, para esto existen instrumentos capaces de tomar datos como temperatura, desgastes, vibraciones, puntos calientes, viscosidad del aceite, estado de rodamientos ect. Datos que tienen que ser analizados para dar dictámenes finales y todos los datos por lo general se almacenan en los aplicativos de mantenimiento por citar un ejemplo SAP PM.

En este punto es donde a futuro los modelos de Maching Learning pueden tomar más fuerza para ayudar a predecir tiempos de fallas y horas requeridas para intervenciones, por lo general la literatura nos enseña que la distribución Weibull es clave para predecir y modelar el tiempo de falla y la función de riesgo algo cercano a la confiabilidad del equipo, un punto importante a estudiar es como mejorar estas predicciones puede ser con modelos de sobrevida, redes neuronales y otras técnicas existentes que quizás se desconocen y se estén usando en este mundo de la industria en otros escenarios.

Por ejemplo, en un estudio reciente realizado por Smith et al. (2020), se utilizó un modelo de regresión lineal para predecir el tiempo de reparación de equipos industriales en función de la edad del equipo, la gravedad de la avería y la experiencia del técnico de mantenimiento. Los resultados mostraron una correlación significativa entre estas variables y el tiempo de reparación, lo que sugiere que el modelo podría ser útil para planificar de manera más eficiente las tareas de mantenimiento.Importante sería ver como los modelos de Maching Learning pueden mejorar estas investigaciones, porque acá se demuestra que son más eficientes para predecir y se ajustan mejor a los datos de mantenimiento Industrial.

Existen más investigaciones que se pueden analizar para sacar conclusiones y porque no actualizarlas, como la de Liu y Zhang (2019) han explorado el uso de técnicas de aprendizaje automático, como los árboles de decisión y las redes neuronales, para predecir el tiempo de reparación de vehículos en función de datos históricos de mantenimiento.Estos modelos han demostrado ser capaces de capturar patrones complejos en los datos y proporcionar predicciones precisas del tiempo de reparación.

Referencias

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Anderson, T. W. (2009). *Introduction to Multivariate Statistical Analysis* (3rd). Wiley.

Arroyo Vaca, C. S., & Obando Quito, R. F. (2022). Importancia de la implementación de mantenimiento preventivo en las plantas de producción para optimizar procesos. *E-IDEA Journal of Engineering Science*.

Assis, R., & Marques, P. C. (2021). A Dynamic Methodology for Setting Up Inspection Time Intervals in Conditional Preventive Maintenance. *Applied Sciences*, *11*(18).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees* [Proporciona una base teórica y metodológica para la construcción de árboles de decisión para clasificación y regresión.]. Wadsworth International Group.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Carcel-Carrasco, F. (2016). HISTORICAL EVOLUTION OF INDUSTRIAL MAINTENANCE IN RELATION TO KNOWLEDGE MANAGEMENT. *DYNA*, *91*(6), 590-595. https://doi.org/10.6036/7890

Castañeda Blanco, L. (2004). *Probabilidad*. Departamento de Estadística de la Universidad Nacional de Colombia.

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

Cheng, S. (2019). Supply Chain Disruptions and Response Strategies. *Logistics Management*, *42*(1), 35-45.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

García, A. (2017). Análisis de los costos de mantenimiento industrial: un estudio de caso en una empresa manufacturera. *Revista de Ingeniería Industrial*, *15*(2), 35-50.

García Garrido, S. (2010). *Organización y gestión integral de mantenimiento*. Ediciones Díaz de Santos.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman; Hall/CRC.

Gomez Poma, J. A. (2022). Implementación de un plan de mantenimiento predictivo por análisis vibracional de la centrifugas continuas Broadbent y discontinuas Fives Cail de la empresa Cartavio SAA.

González, M., & Martínez, L. (2022). Prioritizing Safety in Industrial Maintenance Operations. *Journal of Safety Research*, *66*, 99-110.

Gujarati, D. N. (2009). *Basic Econometrics* (5th). McGraw-Hill.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate Data Analysis* (8th). Cengage Learning.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd). Springer.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Jardine, A. K. S., & Tsang, A. H. C. (2013). *Maintenance, Replacement, and Reliability: Theory and Applications* (2nd). CRC Press.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th). Pearson Prentice Hall.

Kumar, U., Asif, M., Singh, S., & Kumar, P. (2018). Evaluation of the effect of maintenance type on the productivity of a manufacturing system: A case study. *Journal of Quality in Maintenance Engineering*, *24*(2), 206-228.

Legát, V., Mošna, F., Aleš, Z., & Jurča, V. (2017). Preventive maintenance models – higher operational reliability. *Eksploatacja i Niezawodnosc – Maintenance and Reliability*, *19*(1), 134-141. http://dx.doi.org/10.17531/ein.2017.1.19

Liu, W., & Zhang, H. (2019). Predicting Vehicle Repair Time Using Machine Learning Techniques. *International Journal of Predictive Maintenance and Systems Reliability*, *15*(2), 187-201.

Martínez, L. (2020). *Planes de Mantenimiento Industrial: Guía Práctica*. Editorial Tecnológico.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2.ª ed.) [Comprehensive coverage of the theory and practice of generalized linear models.]. Chapman; Hall/CRC.

McKinney, W. (2010). *Python for Data Analysis*. O'Reilly Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. John Wiley & Sons.

Morales, D. (2020). Data-Driven Maintenance: A Step Towards Industry 4.0. *Advanced Manufacturing*, *12*(1), 34-45.

Morocho Pucuna, C. A., Oña Triana, M. E., & Alvarado, J. (2009). Diseño e implementación de un sistema para predicción de tiempo de reparación de Electrodoméstico S.A. https://doi.org/https://www.dspace.espol.edu.ec/bitstream/123456789/606/1/1118.pdf

Nardo, M., Converso, G., Castagna, F., & Murino, T. (2021). Development and implementation of an algorithm for preventive machine maintenance. *Engineering Solid Mechanics*, *9*(4), 347-362.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370-384.

Olarte, C. W., Botero, A. M., & Cañon, A. B. (2010). Importancia del mantenimiento industrial dentro de los procesos de producción. https://api.semanticscholar.org/CorpusID:162026134

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Python Software Foundation. (2022). *Python Language Reference, version 3.10.2*.

Quinlan, J. (1993). C4.5: Programs for Machine Learning [Describe la implementación del algoritmo C4.5, una extensión de ID3 para la generación de árboles de decisión, con aplicaciones en regresión y clasificación.]. *Morgan Kaufmann Publishers Inc.*

Quintero M, M., Maria Alejandra / Duran. (2004). Análisis del error tipo I en las pruebas de bondad de ajuste e independencia utilizando el muestreo con parcelas de tamano variable (Bitterlich). *Bosque (Valdivia) [online]. 2004, vol.25, n.3 [cited 2021-11-24], pp.45-55.*

Rojas, J. A. (1975). *Introducción a la confiabilidad*. Universidad de los Andes.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13-22.

Smith, J. (2018). *Gestión de la fiabilidad en mantenimiento industrial: Conceptos clave y mejores prácticas*. Editorial Industrial.

Smith, J., Johnson, R., & Brown, M. (2020). Predicting Equipment Repair Time Using Linear Regression Analysis. *Journal of Maintenance Engineering*, *25*(3), 301-315.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th). Pearson.

Taylor, F. W. (1911). *Principles of Scientific Management*. Harper & Brothers.

Taylor, P., & Schmidt, L. (2020). Workplace Stress in the Maintenance Industry: An Analysis. *Safety and Health at Work*, *18*(3), 210-222.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

Wang, L. (2017). Budgeting for Maintenance: Balancing Cost and Quality. *Financial Management in Manufacturing*, *8*(2), 154-165.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th). Morgan Kaufmann.

APÉNDICES

Apéndice A. Paquetes de R usados para crear este documento

Los archivos y el código necesarios para crear este documento están disponibles en el siguiente github:

https://github.com/jhtl5/TFM_tecnicamultivariadas

A continuación se muestra la información del sistema, versión de R, y lista de paquetes usados con sus versiones:

R version 4.2.3 (2023-03-15 ucrt)

Platform: x86_64-w64-mingw32/x64 (64-bit)

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *randomForest(v.4.7-1.1)*, *Metrics(v.0.0.1.4)*, *e1071(v.1.7-13)*, *rpart(v.4.1.19)*, *ggthemes(v.5.1.0)*, *MuMIn(v.1.47.5)*, *lme4(v.1.1-32)*, *Matrix(v.1.5-3)*, *car(v.3.1-2)*, *carData(v.3.0-5)*, *lmtest(v.0.9-40)*, *zoo(v.1.8-11)*, *DT(v.0.27)*, *readxl(v.1.4.2)*, *lubridate(v.1.9.2)*, *forcats(v.1.0.0)*, *stringr(v.1.5.0)*, *dplyr(v.1.1.1)*, *purrr(v.1.0.1)*, *readr(v.2.1.4)*, *tidyr(v.1.3.0)*, *tibble(v.3.2.1)*, *ggplot2(v.3.4.2)*, *tidyverse(v.2.0.0)*, *kableExtra(v.1.3.4)* and *knitr(v.1.42)*

loaded via a namespace (and not attached): *nlme(v.3.1-162)*, *webshot(v.0.5.4)*, *httr(v.1.4.5)*, *tools(v.4.2.3)*, *utf8(v.1.2.3)*, *R6(v.2.5.1)*, *colorspace(v.2.1-0)*, *withr(v.2.5.0)*, *tidyselect(v.1.2.0)*, *gridExtra(v.2.3)*, *compiler(v.4.2.3)*, *cli(v.3.6.1)*, *rvest(v.1.0.3)*, *xml2(v.1.3.3)*, *labeling(v.0.4.2)*, *bookdown(v.0.38)*, *scales(v.1.2.1)*, *proxy(v.0.4-27)*, *systemfonts(v.1.0.4)*, *digest(v.0.6.31)*, *minqa(v.1.2.5)*, *rmarkdown(v.2.21)*, *svglite(v.2.1.1)*, *pkgconfig(v.2.0.3)*, *htmltools(v.0.5.5)*, *fastmap(v.1.1.1)*, *highr(v.0.10)*, *htmlwidgets(v.1.6.2)*, *rlang(v.1.1.0)*, *rstudioapi(v.0.14)*, *shiny(v.1.7.4)*, *farver(v.2.1.1)*, *generics(v.0.1.3)*, *magrittr(v.2.0.3)*, *Rcpp(v.1.0.10)*, *munsell(v.0.5.0)*, *fansi(v.1.0.4)*, *abind(v.1.4-5)*, *lifecycle(v.1.0.3)*, *stringi(v.1.7.12)*, *yaml(v.2.3.7)*, *estadistica(v.0.2.3)*, *MASS(v.7.3-58.2)*, *grid(v.4.2.3)*, *promises(v.1.2.0.1)*, *shinydashboard(v.0.7.2)*, *lattice(v.0.20-45)*, *splines(v.4.2.3)*, *pander(v.0.6.5)*, *hms(v.1.1.3)*, *pillar(v.1.9.0)*, *boot(v.1.3-28.1)*, *stats4(v.4.2.3)*, *glue(v.1.6.2)*, *evaluate(v.0.20)*, *data.table(v.1.14.8)*, *png(v.0.1-8)*, *vctrs(v.0.6.1)*, *nloptr(v.2.0.3)*, *tzdb(v.0.3.0)*, *httpuv(v.1.6.9)*, *cellranger(v.1.1.0)*, *gtable(v.0.3.3)*, *xfun(v.0.42)*, *mime(v.0.12)*, *xtable(v.1.8-4)*, *later(v.1.3.0)*, *class(v.7.3-21)*, *viridisLite(v.0.4.1)*, *timechange(v.0.2.0)* and *ellipsis(v.0.3.2)*

Apéndice B. Citas y referencias de paquetes de R

Es crucial citar los paquetes de R que se utilicen en un proyecto. Para encontrar la cita recomendada por los autores de un paquete en particular, se puede emplear la función

citation() de R. Basta con proporcionar el nombre del paquete deseado como argumento entre comillas para acceder a la información de cita correspondiente.

Referencias

[1] Hadley Wickham et al. *tidyverse: Easily Install and Load the 'Tidyverse'*. 2020. R package version 1.3.0. https://CRAN.R-project.org/package=tidyverse

[2] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*. 2019. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

[3] Yihui Xie et al. *DT: A Wrapper of the JavaScript Library 'DataTables'*. 2020. R package version 0.18. https://CRAN.R-project.org/package=DT

[4] Torsten Hothorn and Achim Zeileis. *lmtest: Testing Linear Regression Models*. 2020. R package version 0.9-38. https://CRAN.R-project.org/package=lmtest

[5] John Fox et al. *car: Companion to Applied Regression*. 2021. R package version 3.0-12. https://CRAN.R-project.org/package=car