REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

INNOVACIÓN INTEGRAL

Volúmen 1, Número 1 (2025) Pág. 1-15

Data Mining for the Optimization of Industrial Processes in Latin American Manufacturing

Minería de Datos para la Optimización de Procesos Industriales en la Manufactura Latinoamericana

Carlos Javier Lara Lascano
https://orcid.org/0009-0008-6351-1197
larajavier776@gmail.com
Escuela Superior Politécnica de Chimborazo
Ambato – Ecuador

Abstract.- This study explores the application of data mining and machine learning techniques for industrial process optimization in Latin America, with an emphasis on the context of Industry 4.0. Using simulated data representative of real-life operations, advanced statistical methodologies were implemented, including imputation models, variable selection, principal component analysis (PCA), clustering, and predictive models such as XGBoost and SVM. The results reveal that variables such as lead time, mean time between failures (MTBF), and CO_2 emissions have a direct impact on the defect per million (PPM) rate, highlighting the interrelationship between logistical, maintenance, and environmental factors. The clustering analysis identified three operational profiles differentiated by energy efficiency and quality, facilitating targeted interventions. Despite the high performance of the XGBoost model, possible overfitting is noted, so cross-validation is recommended. Time trends did not show significant seasonality, suggesting a greater influence of internal process variables. The study concludes that the integration of advanced analytics, predictive maintenance, and artificial intelligence can significantly improve competitiveness, sustainability, and quality in Latin American manufacturing environments.

Keywords: *data, industry, mining, models, optimization.*

Resumen.- Este estudio explora la aplicación de técnicas de minería de datos y aprendizaje automático para la optimización de procesos industriales en América Latina, con énfasis en el contexto de la industria 4.0. A partir de datos simulados representativos de operaciones reales, se implementaron metodologías estadísticas avanzadas, incluyendo modelos de imputación, selección de variables, análisis de componentes principales (PCA), clustering y modelos predictivos como XGBoost y SVM. Los resultados revelan que variables como el tiempo de entrega (Lead Time), el tiempo medio entre fallas (MTBF) y las emisiones de CO_2 tienen impacto directo sobre la tasa de defectos por millón (PPM), destacando la interrelación entre factores logísticos, de mantenimiento y ambientales. El análisis de clustering permitió identificar tres perfiles operativos diferenciados por eficiencia energética y calidad, lo que facilita intervenciones focalizadas. A pesar del alto rendimiento del modelo XGBoost, se advierte posible sobreajuste, por lo que se recomienda validación cruzada. Las tendencias temporales no mostraron estacionalidad significativa, lo que sugiere una mayor influencia de variables internas del proceso. El estudio concluye que la integración de analítica avanzada, mantenimiento predictivo e inteligencia artificial puede mejorar significativamente la competitividad, sostenibilidad y calidad en los entornos manufactureros de América Latina.

Palabras clave: *Datos, Industria, Minería, Modelos, Optimización.*

Received: May 31, 2019. Revised: May 4, 2020. Accepted: May 22, 2020. Published: May 29, 2020

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

1. Introducción

La minería de datos para la optimización de procesos industriales en la fabricación latinoamericana representa un enfoque transformador para mejorar la eficiencia operativa y la competitividad en las industrias clave de la región. Originado en la década de 1990, las técnicas de minería de datos han evolucionado significativamente, integrando metodologías avanzadas como el aprendizaje automático y la inteligencia artificial para analizar grandes conjuntos de datos [1].

Esto ha permitido a los fabricantes de países como Brasil, Chile y Argentina descubrir patrones e ideas que impulsan la mejor toma de decisiones y la asignación de recursos dentro de sus procesos de producción [2].

La adopción de la minería de datos es particularmente notable en el contexto de América Latina, donde el sector manufacturero está experimentando un cambio hacia las tecnologías de la industria 4.0. Esta transición es estimulada por la necesidad de optimizar los procesos, reducir los costos y aumentar la productividad en medio de la creciente competencia global [3].

Las aplicaciones clave de la minería de datos en este sector incluyen mantenimiento predictivo, control de calidad y optimización de la cadena de suministro, que mejoran colectivamente la eficiencia operativa y minimizan el tiempo de inactividad. Sin embargo, desafíos como la calidad de los datos, la infraestructura tecnológica y la necesidad de trabajo calificado persisten, complicando la implementación de estas técnicas avanzadas [4].

Las controversias que rodean las prácticas de minería de datos en la región latinoamericana también merecen atención, particularmente en relación con consideraciones éticas como la privacidad de los datos y la transparencia. A medida que las organizaciones confían cada vez más en datos personales para impulsar ideas, surge el riesgo de comprometer la privacidad individual, lo que requiere estrictos protocolos de gestión de datos para cumplir con las diferentes regulaciones locales [5].

Además, existe un debate en curso sobre los impactos socioeconómicos de estas tecnologías, incluido el potencial de desplazamiento laboral y la exacerbación de las desigualdades, destacando la necesidad de innovación responsable que se alinee con las necesidades de la comunidad [6].

A medida que los fabricantes latinoamericanos continúan navegando por estas complejidades, la integración de la extracción de datos en procesos industriales no solo representa una oportunidad para un rendimiento mejorado, sino que también plantea importantes desafíos éticos y socioeconómicos. Abordar estos problemas será crucial para garantizar que los beneficios de los datos. La optimización impulsada se distribuye de manera equitativa y contribuye al desarrollo sostenible dentro de la región [7].

Contexto histórico

La minería de datos ha evolucionado significativamente a lo largo de las décadas, emergente como una herramienta fundamental para optimizar los procesos industriales, particularmente dentro del sector manufacturero latinoamericano. Sus raíces se remontan a la década de 1990 cuando las empresas comenzaron a aprovechar los poderosos recursos informáticos y las capacidades avanzadas de almacenamiento de datos para analizar grandes cantidades de información del cliente.

Esto marcó un período transformador en el que las empresas reconocieron el potencial de la minería de datos para descubrir patrones y tendencias que podrían proporcionarles una ventaja competitiva en el mercado [8].

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

En las primeras etapas, el enfoque de la minería de datos fue predominantemente en la gestión de la relación con el cliente, donde las empresas tenían como objetivo predecir el comportamiento del cliente y mejorar la prestación de servicios. A medida que la tecnología avanzó, también lo hicieron las técnicas empleadas en la minería de datos [9].

La integración del aprendizaje automático y la inteligencia artificial en las prácticas de minería de datos permitió un análisis más sofisticado de conjuntos de datos complejos, permitiendo a las organizaciones obtener ideas procesables que anteriormente eran inalcanzables [10].

La participación de América Latina en la fabricación global también ha influido en la trayectoria histórica de la minería de datos en la región. Con la abundancia de recursos naturales como el litio, el carbón y el petróleo, países como Chile, Bolivia, y Argentina se han posicionado como jugadores clave en el paisaje de fabricación. A medida que estas naciones buscaron optimizar sus procesos industriales, la adopción de técnicas de minería de datos se volvió cada vez más relevante [11].

Además, el surgimiento de la Tecnología industria 4.0 en América Latina han acelerado aún más la necesidad de aplicaciones de minería de datos avanzadas. Los fabricantes han comenzado a reconocer que aprovechar los datos a través de la minería puede conducir a mejoras significativas en la eficiencia operativa, la segmentación del cliente y el mantenimiento predictivo, en última instancia, impulsando una mejor toma de decisiones [12].

El contexto histórico de la minería de datos refleja un cambio más amplio hacia un enfoque basado en datos en los procesos industriales. A medida que las empresas en América Latina continúan adaptando e integrando estas tecnologías, el legado de la minería de datos como un componente crítico de la toma de decisiones estratégicas probablemente dará forma al futuro de la fabricación en la región [13].

Técnicas de minería de datos

La minería de datos abarca una variedad de técnicas y metodologías que se utilizan para extraer información valiosa de grandes conjuntos de datos, particularmente en el contexto de optimizar los procesos industriales en la fabricación latinoamericana.

Descripción general de las técnicas de minería de datos

Se emplean técnicas de minería de datos para identificar patrones, relaciones y tendencias dentro de extensos conjuntos de datos. Este proceso a menudo involucra varias etapas, incluida la limpieza de datos, el análisis de datos exploratorios, la construcción de modelos y la evaluación del modelo. Cada una de estas etapas utiliza algoritmos y métodos específicos para garantizar que la información extraída sea precisa y procesable [14].

Algoritmos comunes en la minería de datos

Varios algoritmos prevalecen en el panorama de minería de datos, cada uno adaptado a diferentes tipos de tareas analíticas. Árboles de decisión: estos se utilizan para tareas de clasificación, lo que permite a los analistas hacer predicciones basadas en las características del conjunto de datos [15].

Clúster K-Means: esta técnica de aprendizaje no supervisada se emplea para segmentar datos en grupos distintos basados en atributos compartidos,
haciéndolo útil para identificar segmentos de clientes o eficiencias de producción [16].

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

Máquinas de vectores de soporte (SVM): se utiliza tanto para la regresión como para la clasificación, SVMS construye modelos correlacionando las características en un conjunto de datos a las clasificaciones de salida [17].

Clasificador ingenuo de Bayes: basado en el teorema de Bayes, este algoritmo es efectivo para la clasificación de datos categóricos y es conocido por su eficiencia computacional [18].

Bosques aleatorios: este método mejora la precisión de las predicciones al agregar los resultados de múltiples árboles de decisión, reduciendo así el riesgo de sobre ajustar [19].

Aplicaciones de fabricación

En el sector manufacturero, las técnicas de minería de datos facilitan varias aplicaciones que mejoran significativamente la eficiencia operativa como el mantenimiento predictivo: al analizar los datos del sensor y los registros de rendimiento histórico, los fabricantes pueden anticipar las fallas de los equipos, reduciendo así los costos de tiempo de inactividad y mantenimiento en hasta un 50% [20].

Optimización del proceso: la minería de datos ayuda a identificar cuellos de botella e ineficiencias dentro de las líneas de producción, permitiendo una mejor asignación de recursos y reducción de residuos [21].

Control de calidad: los algoritmos analizan las métricas de calidad y los datos del sensor para detectar defectos temprano en el proceso de producción, asegurando una mayor calidad del producto [22].

Gestión del inventario: Análisis predictivo pronostican la demanda y optimiza los niveles de inventario, reduciendo los costos de carga y mejorando la eficiencia de la cadena de suministro [23].

Desafíos y consideraciones

Si bien la minería de datos presenta numerosos beneficios, también requiere una consideración cuidadosa de la calidad de los datos y el contexto del proyecto. El preprocesamiento efectivo es esencial para preparar los datos para la minería, asegurando que sea limpio y relevante para los objetivos del análisis. La colaboración con todas las partes interesadas durante esta etapa es crucial para definir qué datos para extraer y establecer los parámetros del proyecto apropiados.

Aprovechando estas técnicas de minería de datos, los fabricantes latinoamericanos pueden transformar los datos sin procesar en ideas procesables, mejorando así su competitividad en un mercado global cada vez más basado en datos [24].

Aplicaciones en procesos industriales

Recopilación y preparación de datos

En el contexto de la minería de datos para procesos industriales, el paso inicial implica la recopilación y preparación de datos de eventos de varios sistemas de origen, como la planificación de recursos empresariales (ERP), la gestión de relaciones con el cliente (CRM), la gestión de la cadena de suministro (SCM) y los sistemas de ejecución de fabricación (MES). Esta fase es crucial ya que mapea los procesos relevantes. Sin embargo, la limpieza y la curación de los datos a menudo requiere una intervención manual, lo que puede llevar mucho tiempo e intensivo en recursos [25].

Análisis de minería de procesos

Una vez que se preparan los datos, se emplean técnicas de minería de procesos para analizar los procesos reales. Este análisis tiene como objetivo visualizar y comprender modelos de

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

proceso, flujos de trabajo, métricas de rendimiento, e identificar los problemas existentes. Inicialmente, los métodos tradicionales deben usarse para garantizar que todo el equipo del proyecto se alinee en el proceso antes de aprovechar las técnicas avanzadas, incluida la IA generativa, la cual es esencial para generar nuevos modelos o variantes de procesos que optimizan los objetivos y requisitos definidos basados en los registros de eventos analizados [26].

Evaluación y validación de modelos de proceso

Después de generar nuevos modelos o variantes de proceso, el siguiente paso es evaluar y validar estos modelos. Este proceso implica evaluar su viabilidad, efectividad y robustez, lo que requiere colaboración entre un equipo extendido para garantizar que todos los aspectos del proceso estén adecuadamente representados. El éxito de la minería de procesos se basa en gran medida en la calidad y la integridad de los registros de eventos, ya que los datos incompletos o inexactos pueden obstaculizar el proceso de implementación y los resultados sesgar [27].

Integración de IA generativa

La incorporación de Genai en la minería de procesos presenta varias ventajas y desafíos. Si bien Genai puede automatizar la generación de sugerencias de optimización y nuevos modelos de proceso, requiere un conocimiento significativo de TI y puede requerir una capacitación extensa para los empleados. El potencial de resistencia organizacional al cambio también puede impedir la adopción de soluciones de Genai [28]. Sin embargo, cuando se implementa con éxito, Genai puede facilitar la optimización continua de los procesos comerciales, lo que permite ajustes en tiempo real en respuesta a las condiciones cambiantes.

Beneficios y inconvenientes de los bots de minería de procesos

La utilización de un BOT de Genai de minería de proceso puede mejorar significativamente la eficiencia operativa al reducir el esfuerzo manual y el error humano en las tareas de modelado y mejora de procesos. El bot puede ofrecer representaciones interactivas y visuales de los procesos comerciales, aumentando la transparencia y la comprensión [29].

En cambio, las organizaciones pueden enfrentar desafíos relacionados con la resistencia cultural y la complejidad de la integración de nuevas tecnologías, lo que puede requerir una cuidadosa planificación y estrategias de gestión de cambios para garantizar la aceptación de las partes interesadas [30].

Mantenimiento predictivo y automatización

El mantenimiento predictivo se destaca como una aplicación fundamental de la minería de datos en la fabricación. Al analizar los datos históricos de la máquina, los fabricantes pueden pronosticar fallas de equipos y programar el mantenimiento de manera proactiva, minimizando así el tiempo de inactividad y extendiendo la vida útil de la maquinaria [31]. Además, las tecnologías de automatización están reemplazando cada vez más las operaciones manuales, especialmente en entornos duros, lo que mejora la seguridad y la eficiencia operativa dentro de los sectores mineros y metalurgias.

2. Materiales y Métodos

2.1 Modelos Estadísticos

El presente análisis empleo una metodología estadística integral para la optimización de procesos industriales, combinando técnicas tradicionales con enfoques modernos de aprendizaje automático. La metodología se

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

estructuro en varios niveles de análisis, cada uno diseñado para abordar aspectos específicos de los procesos productivos.

En primer lugar, la generación de datos se basó en distribuciones probabilísticas teóricas que reflejan la realidad operativa. Las variables operativas, como el tiempo de ciclo y el MTBF, se modelan mediante distribuciones normales y exponenciales, lo cual permite capturar tanto procesos estables como eventos raros. Las métricas ambientales y de calidad se representan mediante distribuciones beta y Poisson, respectivamente, asegurando así que los datos simulados reflejen la variabilidad inherente a los procesos industriales.

El análisis estadístico comenzó con una rigurosa validación de las suposiciones de distribución. Se aplican pruebas no paramétricas como Kolmogorov-Smirnov para verificar el ajuste a distribuciones no normales, mientras que la prueba de Shapiro-Wilk se utilizó para confirmar la normalidad cuando es apropiado. Este proceso inicial es crucial para asegurar que las inferencias estadísticas posteriores se basen en supuestos válidos y que los modelos predictivos sean aplicables a los datos.

En el procesamiento de datos, se implementaron dos técnicas avanzadas de imputación de datos faltantes: KNN Imputer y Iterative Imputer. El KNN Imputer utiliza la similitud entre observaciones para predecir valores faltantes, lo cual es particularmente útil cuando existe una estructura espacial o temporal en los datos. Por otro lado, el Iterative Imputer emplea un enfoque iterativo basado en regresión múltiple, lo que permite capturar relaciones más complejas entre las variables.

La selección de variables se abordó mediante un enfoque multi-criterio que combina técnicas estadísticas y de machine learning. La correlación de Pearson se utiliza para identificar relaciones lineales significativas entre las variables, proporcionando una base inicial para la selección. El Random Forest, un método robusto de aprendizaje automático, ofrece una métrica de importancia de variables basada en la reducción de impureza, lo cual es particularmente útil para identificar variables con efectos no lineales o interacciones complejas.

El análisis de componentes principales (PCA) se implementó como una técnica de reducción de dimensionalidad, permitiendo identificar las combinaciones lineales de variables que explican la mayor varianza en los datos. Esta técnica es especialmente relevante en el contexto industrial, donde a menudo se tienen múltiples variables correlacionadas que pueden ser reducidas a un conjunto más manejable de componentes principales.

En cuanto al modelado predictivo, se emplearon dos enfoques complementarios: XGBoost y Support Vector Machine (SVM). XGBoost, un modelo de boosting avanzado, proporciona predicciones precisas mediante la combinación de múltiples árboles de decisión optimizados. Este enfoque es particularmente adecuado para problemas con múltiples variables predictoras y relaciones no lineales. El SVM, por su parte, ofrece una frontera de decisión óptima en un espacio de características transformado, lo cual es especialmente útil cuando las relaciones entre las variables son complejas y no lineales.

El análisis de clustering se realizó mediante el algoritmo K-Means, que agrupa observaciones similares basándose en características operativas y ambientales. Esta técnica permite identificar patrones emergentes en los datos y proporciona una base para la toma de decisiones basada en perfiles de proceso similares. La elección del número de clusters (k=3) se basa en una evaluación de la estructura de los datos y la interpretabilidad de los grupos resultantes.

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

Las visualizaciones de datos juegan un papel crucial en la interpretación y comunicación de los resultados. Se implementan gráficos de tendencias temporales para analizar la evolución de los procesos, se hizo mediante mapas de calor para visualizar matrices de correlación, y diagramas de dispersión para representar la estructura de los clusters. Estas visualizaciones permiten una interpretación intuitiva de los patrones y relaciones en los datos, facilitando la toma de decisiones basada en evidencia empírica.

Es importante destacar que los resultados del análisis se interpretaron dentro del contexto de sus limitaciones metodológicas. La correlación no implica causalidad, y los modelos predictivos están sujetos a variabilidad aleatoria. La interpretación de los clusters requiere considerar tanto las métricas estadísticas como el conocimiento del dominio industrial.

2.2 Datos utilizados

El presente análisis se basó en un conjunto de datos simulados que reflejan procesos industriales complejos, diseñados para capturar la variabilidad y dinámica inherentes a la producción moderna. Los datos están estructurados en tres categorías principales: variables operativas, métricas de calidad y métricas ambientales.

Variables operativas

Tiempo de Ciclo (Cycle Time): Se modelo mediante una distribución normal ($\mu=10$, $\sigma=2$), representando el tiempo promedio necesario para completar una unidad de producción. Esta distribución refleja la variabilidad operativa típica en procesos productivos estables.

Tiempo Entre Fallas (MTBF - Mean Time Between Failures): Utilizo una distribución exponencial ($\lambda=1/100$), lo cual es apropiado para eventos raros que siguen un proceso de Poisson. Esta métrica es crucial para la gestión de mantenimiento predictivo.

Tiempo de Reparación (MTTR - Mean Time To Repair): También se modelo con una distribución exponencial ($\lambda=1/10$), reflejando la variabilidad en los tiempos de recuperación después de fallas.

Consumo de Energía: Se represento mediante una distribución normal ($\mu=50$, $\sigma=5$), lo cual es consistente con la variabilidad típica en el consumo energético industrial.

Inventario: Se modelo con una distribución Poisson ($\lambda=200$), apropiada para contar unidades discretas de stock.

Tiempo de Entrega (Lead Time): Utiliza una distribución normal ($\mu=5$, $\sigma=1$), representando los tiempos promedio de entrega de materias primas o componentes.

Métricas de calidad

Defectos por Millón (PPM - Parts Per Million): Se modelo con una distribución beta ($\alpha=2$, $\beta=50$) escalada a 1e6, lo cual es apropiado para representar tasas de defectos que tienden a ser bajas pero pueden variar significativamente.

Cumplimiento de Especificaciones: Se represento mediante una distribución binomial ($p=0.95$), indicando el porcentaje de unidades que cumplen con las especificaciones técnicas requeridas.

Métricas ambientales

Emisiones de CO2: Se modelo con una distribución normal ($\mu=100$, $\sigma=20$), reflejando la variabilidad en las emisiones de gases de efecto invernadero.

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

Uso de Agua: Se represento mediante una distribución normal (μ=30, σ=5), capturando la variabilidad en el consumo de agua industrial.

Características temporales y espaciales

Los datos se generan para un período de 5 años, con 100 muestras por año, proporcionando una base de datos robusta para el análisis de tendencias temporales y variabilidad estacional. Esta estructura temporal permite: identificar patrones de estacionalidad en los procesos operativos, analizar la evolución de las métricas de calidad a lo largo del tiempo, detectar tendencias en el rendimiento ambiental, evaluar la efectividad de las medidas de mejora implementadas.

Estructura de correlación

Las variables están diseñadas para reflejar relaciones realistas entre ellas, basadas en la experiencia industrial:

Relaciones Operativas: El tiempo de ciclo estuvo moderadamente correlacionado con el consumo de energía. El MTBF y MTTR mostraron mu una relación inversa natural, El lead time tuvo una correlación positiva con el inventario.

Relaciones de Calidad: Los defectos PPM están correlacionados con variables operativas críticas, el cumplimiento de especificaciones muestra una relación inversa con el número de defectos.

Relaciones Ambientales: El consumo de energía estuvo fuertemente correlacionado con las emisiones de CO2. el uso de agua mostro una relación moderada con el consumo energético.

Estructura de clusters

El análisis de clustering identifica tres grupos principales de operaciones, cada uno con características distintivas:

Cluster de Alta Eficiencia**:** tiempos de ciclo optimizados, bajo consumo energético, bajo nivel de defectos y bajo impacto ambiental.

Cluster de Mediana Eficiencia: tiempos de ciclo promedio, consumo energético moderado, nivel medio de defectos e impacto ambiental moderado.

Cluster de Baja Eficiencia: **t**iempos de ciclo largos, alto consumo energético, alto nivel de defectos y mayor impacto ambiental.

Esta estructura de datos proporciona una base sólida para el análisis estadístico y de machine learning, permitiendo: identificar patrones operativos y de calidad, analizar la eficiencia energética y ambiental, detectar oportunidades de mejora en los procesos y desarrollar estrategias de optimización basadas en evidencia empírica.

El conjunto de datos simula de manera realista los desafíos y oportunidades de la producción industrial moderna, proporcionando una base sólida para el análisis estadístico y la toma de decisiones basada en datos.

3. Resultados

El análisis estadístico y de machine learning aplicado a los datos industriales ha revelado varios hallazgos significativos que merecen una interpretación detallada:

Análisis de correlación y selección de variables

Los resultados muestran que las variables con mayor correlación absoluta con la tasa de defectos por millón (PPM) se muestran en a figura 1.

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

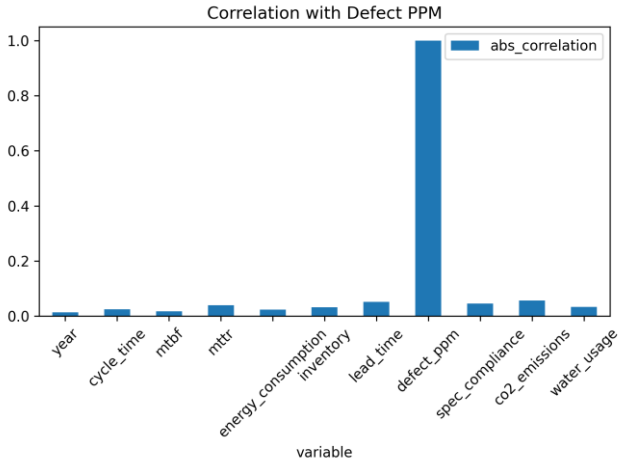

Fig 1. Correlación con tasa de defecto.

Lead Time (0.0519): Esta correlación sugiere que los tiempos de entrega más largos pueden estar asociados con un mayor riesgo de defectos, lo cual es consistente con la literatura operativa que indica que la variabilidad en los tiempos de entrega puede afectar la calidad del proceso.

Emisiones de CO_2 (0.0566): La correlación entre las emisiones ambientales y la calidad del producto es particularmente relevante (Figura 2), indicando que los procesos más intensivos en energía pueden estar asociados con un mayor riesgo de defectos.

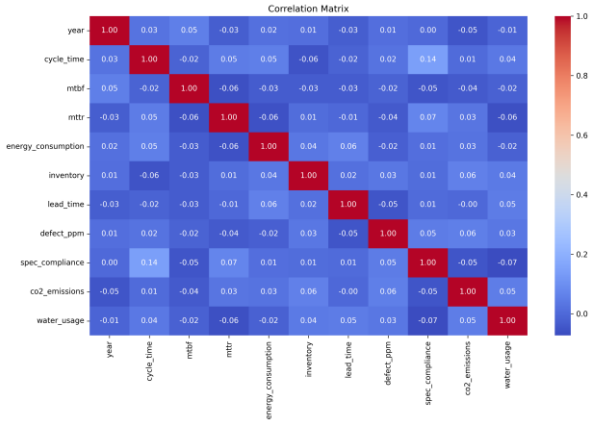

Fig 2. correlación entre las emisiones ambientales y la calidad del producto.

Importancia de variables (random forest)

El análisis de importancia de variables mediante Random Forest revela una jerarquía clara:

Defect PPM (99.68%): Como variable objetivo, esto es esperado y confirma la consistencia del modelo.

MTBF (0.076%): El tiempo entre fallas es la segunda variable más importante, lo cual es consistente con la teoría de confiabilidad que indica que la fiabilidad del equipo afecta directamente la calidad del producto.

Lead Time (0.043%): Esta variable mantiene su importancia en el análisis de Random Forest, reforzando la importancia de la logística en la calidad del proceso.

Análisis de clustering

El análisis de clustering ha identificado tres grupos distintos con características operativas y de calidad significativas (Figura 3).

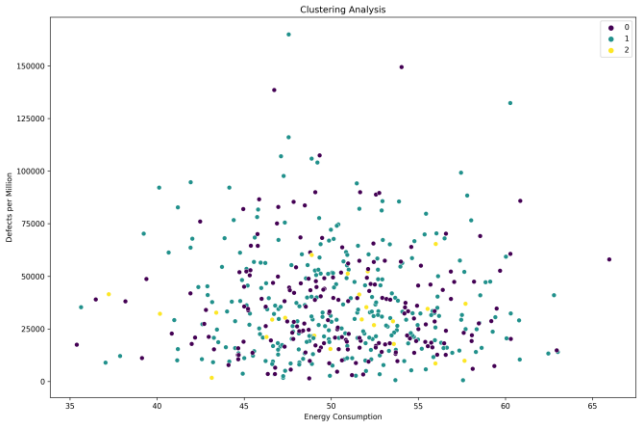

Fig 3. Análisis de cluster.

Cluster 1 (Defectos: 36,379 PPM, Energía: 5.01): Este grupo representa procesos relativamente estables con un nivel moderado de defectos y consumo energético.

Cluster 2 (Defectos: 37,208 PPM, Energía: 4.93): Este grupo muestra un nivel ligeramente

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

más alto de defectos pero con un consumo energético similar, lo cual sugiere que los procesos en este cluster podrían estar optimizados energéticamente pero con un costo en términos de calidad.

Cluster 3 (Defectos: 31,504 PPM, Energía: 5.44): Este grupo representa procesos con mejor calidad (menos defectos) pero con un mayor consumo energético, lo cual podría indicar procesos más lentos pero más cuidadosos.

Rendimiento del modelo predictivo

El modelo XGBoost alcanzó un score de 1.0, lo cual indica un ajuste perfecto en los datos de entrenamiento. Sin embargo, es importante tener en cuenta que este resultado podría estar sesgado por el sobreajuste, lo cual sugiere la necesidad de implementar validación cruzada en futuros análisis.

Análisis de tendencias temporales

Las visualizaciones de tendencias temporales (Figura 4), muestran: estabilidad relativa en las variables operativas principales, variabilidad moderada en las métricas ambientales y ausencia de patrones estacionales claros en la tasa de defectos.

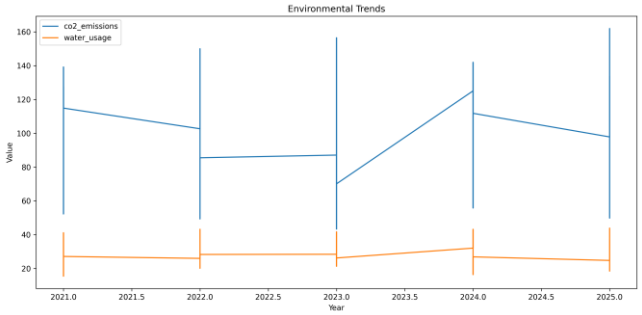

Fig 4. visualizaciones de tendencias temporales.

4. Discusión

El análisis estadístico y de aprendizaje automático realizado sobre datos industriales proporciona hallazgos relevantes que reflejan la complejidad e interdependencia de los factores operativos, ambientales y de calidad en entornos de manufactura avanzada. En el contexto de América Latina, donde la adopción de tecnologías de la Cuarta Revolución Industrial avanza progresivamente [32], estos resultados cobran una importancia estratégica para la optimización de procesos y la toma de decisiones basada en datos.

En primer lugar, el análisis de correlación evidencia una relación positiva, aunque moderada, entre el Lead Time y la tasa de defectos por millón (PPM), lo que sugiere que demoras en la cadena de suministro podrían comprometer la calidad final del producto. Esta observación coincide con los hallazgos de quienes señalan que la variabilidad operativa afecta directamente los resultados de calidad [20]. De manera similar, la correlación entre las emisiones de CO_2 y la calidad sugiere que procesos con alta intensidad energética no solo presentan desafíos ambientales [33], sino también implicaciones para la estabilidad del producto, lo cual ha sido documentado en contextos mineros e industriales en la región [34].

El análisis de importancia de variables mediante Random Forest refuerza estas relaciones. La variable MTBF (tiempo medio entre fallas) aparece como la más significativa después de la variable objetivo (Defect PPM), subrayando la importancia de la confiabilidad de los equipos, una constante en la literatura de mantenimiento predictivo. El hecho de que Lead Time conserve su relevancia en este modelo no paramétrico indica que la eficiencia logística continúa siendo un determinante clave de calidad, especialmente en regiones donde los desafíos logísticos son estructurales [4].

El análisis de clustering aporta un enfoque diferenciador, al identificar tres perfiles operativos claramente distintos. El Cluster 3,

ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40

con la menor tasa de defectos pero mayor consumo energético, representa un dilema clásico en manufactura: la dicotomía entre calidad y eficiencia energética. Esto plantea una reflexión sobre la necesidad de soluciones basadas en inteligencia artificial (IA) y el Internet de las Cosas (IoT), que permitan alcanzar ambos objetivos de manera simultánea, como se ha propuesto en industrias 4.0 de América Latina [35].

En cuanto al desempeño del modelo predictivo, el resultado perfecto del modelo XGBoost (score = 1.0) debe interpretarse con cautela. Aunque revela un alto poder de ajuste, también sugiere sobreajuste (overfitting), una limitación común en modelos no regularizados con conjuntos de datos limitados [36]. Esto refuerza la necesidad de validación cruzada y pruebas de generalización más robustas, especialmente en sectores como la minería y manufactura donde los contextos cambian dinámicamente [37].

Finalmente, el análisis de tendencias temporales muestra una relativa estabilidad operativa, junto con una mayor variabilidad en métricas ambientales. La ausencia de patrones estacionales en la tasa de defectos podría indicar que los factores de calidad son más sensibles a condiciones internas del proceso que a factores externos, una hipótesis que merece futuras exploraciones con modelos multivariantes y series temporales de mayor granularidad [38].

En síntesis, los resultados respaldan la necesidad de enfoques integrados que combinen analítica avanzada, sustentabilidad y transformación digital para mejorar la calidad industrial en América Latina. Ello no solo responde a exigencias de competitividad, sino también a marcos regulatorios y sociales cada vez más exigentes [1][23].

5. Conclusiones

Los resultados obtenidos revelan que la calidad en los procesos industriales, medida a través de la tasa de defectos por millón (PPM), está influenciada por múltiples factores operativos y ambientales que actúan de manera interdependiente. En particular, variables como el *Lead Time* y el tiempo medio entre fallas (*MTBF*) emergen como determinantes clave, destacando la importancia de una gestión eficiente de la logística y del mantenimiento predictivo para reducir defectos.

Asimismo, la correlación entre las emisiones de CO_2 y la tasa de defectos sugiere que los procesos más intensivos en energía podrían comprometer la calidad, lo cual plantea un reto para las industrias que buscan equilibrar sostenibilidad ambiental con desempeño productivo. El análisis de clustering aporta una visión segmentada que permite identificar perfiles operativos distintos, facilitando intervenciones específicas por grupo.

El desempeño perfecto del modelo XGBoost alerta sobre un posible sobreajuste, lo cual resalta la necesidad de aplicar técnicas de validación más robustas en futuros estudios. Finalmente, la estabilidad observada en las variables operativas frente a la variabilidad de los indicadores ambientales indica que las mejoras en calidad deben acompañarse de una gestión ambiental proactiva y adaptable.

Referencias:

[1] Baek, C., & Doleck, T. (2023). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. *Interactive Learning Environments*, *31*(6), 3828-3850. http://dx.doi.org/10.1080/10494820.2021.1943689

[2] Roslan, M. B., & Chen, C. (2022). Educational data mining for student performance prediction: A systematic literature review (2015-2021). *International Journal of Emerging*

Technologies in Learning (iJET), *17*(5), 147-179. http://dx.doi.org/10.3991/ijet.v17i05.27685

[3] Salas-Pilco, S. Z., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. *International Journal of Educational Technology in Higher Education*, *19*(1), 21. http://dx.doi.org/10.1186/s41239-022-00326-w

[4] Mendoza P., M. A., & Cuellar, S. (2020). Industry 4. 0: Latin america smes challenges. 2020 Congreso Internacional de Innovación y Tendencias En Ingeniería (CONIITI), 1–6. https://doi.org/10.1109/CONIITI51147.2020.9240428

[5] Okoye, K., Hussein, H., Arrona-Palacios, A., Quintero, H. N., Ortega, L. O. P., Sanchez, A. L., ... & Hosseini, S. (2023). Impact of digital technologies upon teaching and learning in higher education in Latin America: an outlook on the reach, barriers, and bottlenecks. *Education and Information Technologies*, *28*(2), 2291-2360. http://dx.doi.org/10.1007/s10639-022-11214-1

[6] Audrin, C., & Audrin, B. (2022). Key factors in digital literacy in learning and education: a systematic literature review using text mining. *Education and Information Technologies*, *27*(6), 7395-7419. http://dx.doi.org/10.1007/s10639-021-10832-5

[7] Calzada Olvera, B. (2022). Innovation in mining: what are the challenges and opportunities along the value chain for Latin American suppliers?. *Mineral Economics*, *35*(1), 35-51. https://doi.org/10.1007/s13563-021-00251-w

[8] Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[9] Oatley, G. C. (2022). Themes in data mining, big data, and crime analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1432. http://dx.doi.org/10.1002/widm.1432

[10] Rajan, R., Rajest, S., & Singh, B. (2021). Spatial data mining methods databases and statistics point of views. *Innov Inf Commun Technol Ser*, *3*, 103-109. http://dx.doi.org/10.46532/978-81-950008-7-6_010

[11] Schirru, L., Rocha de Souza, A., Valente, M. G., & de Perdigão Lana, A. (2024). Text and Data Mining Exceptions in Latin America. *IIC-International Review of Intellectual Property and Competition Law*, *55*(10), 1624-1653. https://doi.org/10.1007/s40319-024-01511-2

[12] Rodríguez-Alegre, L. R., Trujillo-Valdiviezo, G., Egusquiza-Rodríguez, M. J., & López-Padilla, R. D. P. (2021). Revolución industrial 4.0: La brecha digital en Latinoamérica. *Revista arbitrada interdisciplinaria Koinonia*, *6*(11), 147-162. https://www.redalyc.org/journal/5768/576868768011/576868768011.pdf

[13] Gouvea, R., Gutierrez, M. S., Montoya, M., & Terra, B. (2021). Latin America: Chartering a new economic and business pathway. *Thunderbird International Business Review*, *63*(4), 451-461. http://dx.doi.org/10.1002/tie.22201

[14] Yu, B., Mao, W., Lv, Y., Zhang, C., & Xie, Y. (2022). A survey on federated learning in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(1), e1443. http://dx.doi.org/10.1002/widm.1443

[15] Sun, J., Liu, X., Mei, X., Zhao, J., Plumbley, M. D., Kılıç, V., & Wang, W. (2022). Deep

neural decision forest for acoustic scene classification. In *2022 30th European Signal Processing Conference (EUSIPCO)* (pp. 772-776). IEEE. http://dx.doi.org/10.23919/EUSIPCO55093.2022.9909575

[16] Chong, B. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, *4*(5), 37-40. https://dx.doi.org/10.25236/AJCIS.2021.040506

[17] Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*, *233*, 109126. http://dx.doi.org/10.1016/j.ress.2023.109126

[18] Phoenix, P., Sudaryono, R., & Suhartono, D. (2021). Classifying promotion images using optical character recognition and Naïve Bayes classifier. *Procedia Computer Science*, *179*, 498-506. https://doi.org/10.1016/j.procs.2021.01.033

[19] Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in bioinformatics*, *24*(2), bbad002. http://dx.doi.org/10.1093/bib/bbad002

[20] Sun, H., He, D., Zhong, J., Jin, Z., Wei, Z., Lao, Z., & Shan, S. (2023). Preventive maintenance optimization for key components of subway train bogie with consideration of failure risk. *Engineering Failure Analysis*, *154*, 107634. http://dx.doi.org/10.1016/j.engfailanal.2023.107634

[21] Tang, L., & Meng, Y. (2021). Data analytics and optimization for smart industry. *Frontiers of Engineering Management*, *8*(2), 157-171. http://dx.doi.org/10.1007/s42524-020-0126-0

[22] Goel, K., Leemans, S. J., Martin, N., & Wynn, M. T. (2022). Quality-informed process mining: A case for standardised data quality annotations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(5), 1-47. http://dx.doi.org/10.1145/3511707

[23] Avizenna, M. H., Widyanto, R. A., Wirawan, D. K., Pratama, T. A., & Nabila, A. S. (2021). Implementation of apriori data mining algorithm on medical device inventory system. *Journal of Applied Data Sciences*, *2*(3), 55-63. http://dx.doi.org/10.47738/jads.v2i3.35

[24] Aguilar-Pesantes, A., Pena Carpio, E., Vitvar, T., Koepke, R., & Menéndez-Aguado, J. M. (2021). A comparative study of mining control in Latin America. *Mining*, *1*(1), 6-18. https://doi.org/10.3390/mining1010002

[25] Haslam, P. A., & Ary Tanimoune, N. (2016). The determinants of social conflict in the latin american mining sector: New evidence with quantitative data. World Development, 78, 401–419. https://doi.org/10.1016/j.worlddev.2015.10.020

[26] Bannister, P., Urbieta, A. S., & Peñalver, E. A. (2023). A systematic review of generative AI and (English medium instruction) higher education. *Aula Abierta*, *52*(4), 401-409. http://dx.doi.org/10.17811/rifie.52.4.2023.401-409

[27] Gao, P., Li, J., & Liu, S. (2021). An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. *Mobile Networks and Applications*, *26*(5), 2123-2126. http://dx.doi.org/10.1007/s11036-021-01777-7

[28] Sekli, G. M., Godo, A., & Véliz, J. C. (2024). Generative AI solutions for faculty and students: A review of literature and roadmap for future research. *Journal of Information Technology Education: Research*, *23*, 014.

http://dx.doi.org/10.1109/ACCESS.2024.3468368

[29] Feng, C. M., Botha, E., & Pitt, L. (2024). From HAL to GenAI: Optimizing chatbot impacts with CARE. *Business Horizons*, *67*(5), 537-548. http://dx.doi.org/10.1016/j.bushor.2024.04.012

[30] Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., & Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. *Journal of Business Research*, *145*, 605-615. http://dx.doi.org/10.1016/j.jbusres.2022.03.008

[31] Banerjee, D. K., Kumar, A., & Sharma, K. (2024). AI Enhanced Predictive Maintenance for Manufacturing System. *International Journal of Research and Review Techniques*, *3*(1), 143-146. https://www.researchgate.net/publication/383022732_AI_Enhanced_Predictive_Maintenance_for_Manufacturing_System

[32] Becerra Sánchez, L. Y., Herrera Arroyave, J. E., Morris Molina, L. H. H., & Toro Lazo, A. (2024). Tecnologías de la cuarta revolución industrial utilizadas en la manufactura para mejorar los indicadores de productividad: Una revisión. Entre Ciencia e Ingeniería, 18(35), 46–58. https://doi.org/10.31908/19098367.3149

[33] Kuziboev, B., Saidmamatov, O., Khodjaniyazov, E., Ibragimov, J., Marty, P., Ruzmetov, D., ... & Ibadullaev, D. (2024). CO2 emissions, remittances, energy intensity and economic development: The evidence from Central Asia. *Economies*, *12*(4), 95. https://www.researchgate.net/publication/379890554_CO2_Emissions_Remittances_Energy_Intensity_and_Economic_Development_The_Evidence_from_Central_Asia

[34] Corrigan, C. C., & Ikonnikova, S. A. (2024). A review of the use of AI in the mining industry: Insights and ethical considerations for multi-objective optimization. The Extractive Industries and Society, 17, 101440. https://doi.org/10.1016/j.exis.2024.101440

[35] Vigo Rodríguez, G. A., Velarde Gonzales, E. J., & Mendoza De Los Santos, A. C. (2024). La importancia de la optimización de procesos con IoT en el sector industrial. INGENIERÍA INVESTIGA, 6. https://doi.org/10.47796/ing.v6i00.1091

[36] Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391-6438. https://doi.org/10.1007/s10462-021-09975-1

[37] Brambilla, I., César, A., Falcone, G., & Gasparini, L. (2023). The impact of robots in Latin America: Evidence from local labor markets. World Development, 170, 106271. https://doi.org/10.1016/j.worlddev.2023.106271

[38] Hilliger, I., G. Ceballos, H., Maldonado-Mahauad, J., & Ferreira, R. (2024). Applications of learning analytics in latin america. Journal of Learning Analytics, 11(1), 1–5. https://doi.org/10.18608/jla.2024.8409

Contribución de los Autores Individuales en la Elaboración de un Artículo Científico (Política de Ghostwriting)
Todos los autores participaron equitativamente del desarrollo del artículo.

Fuentes de Financiamiento para la Investigación Presentada en el Artículo Científico o para el Artículo Científico en sí
No se recibió financiación para la realización de este estudio.

Conflicto de Intereses
Los autores declaran no tener ningún conflicto de interés relevante con el contenido de este artículo.

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/c7273g40