ISSN: 1692-7257 - Volumen 2 – Número 48 - 2026

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Predicción de tiempos de operación en recubrimientos orgánicos mediante regresión lineal múltiple y árboles de decisión

Prediction of operating times in organic coatings using multiple linear regression and decision trees

Johan Sebastián Lavacude Galvis[1]
Hugo Fernando Castro Silva[2]
Josué Iván Mesa Mojica[1]

[1] **Universidad Pedagógica y Tecnológica de Colombia, Facultad seccional Sogamoso, Grupo de investigación Observatorio, Escuela de Ingeniería Industrial, Sogamoso, Boyacá, Colombia.**

[2] **Universidad Pedagógica y Tecnológica de Colombia, Facultad seccional Sogamoso, Grupo de investigación GITYD, Escuela de Ingeniería Industrial, Sogamoso, Boyacá, Colombia.**

Recibido: 27 abril 2026. Revisado: 05 junio 2026. Aceptado: 25 junio 2026. Publicado: 09 julio 2026.
Autor de correspondencia: Hugo Fernando Castro Silva

Cómo citar: J. S. Lavacude Galvis, H. F. Castro Silva, and J. I. Mesa Mojica, "Predicción de tiempos de operación en recubrimientos orgánicos mediante regresión lineal múltiple y árboles de decisión", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 48, jul. 2026, art. no. 4500, doi: 10.24054/rcta.v2i48.4500

Esta obra está bajo una licencia internacional Creative Commons Atribución-NoComercial 4.0.

Resumen:

Este artículo presenta el desarrollo de modelos predictivos para estimar tiempos de operación en un proceso industrial de recubrimientos orgánicos, aplicado a la matrícula de nuevos productos en una planta de manufactura. La problemática central radica en que antes de incorporar los nuevos productos a líneas de producción, la organización debe matricular un tiempo estándar preliminar en el sistema de información, aun cuando todavía no se dispone de estudios de métodos y tiempos. Para abordar esta problemática, se consolida una base de datos a partir de registros históricos, mediciones por cronometro y atributos asociados a cada referencia. Para realizar el análisis se definen 3 familias de productos, considerando aspectos fundamentales para su agrupación como lo son: el área superficial de la pieza, expresada en decímetros cuadrados y el número de unidades por ganchera. Seguidamente, se compararon dos técnicas supervisadas: regresión lineal múltiple y árboles de decisión de regresión, siendo necesario definir criterios de depuración, protocolos de entrenamiento y prueba, así como métricas de desempeño, análisis de sensibilidad, diagnóstico de sobreajuste y validación cruzada con tiempos estándar previamente definidos por la organización. Los resultados muestran que los árboles de decisión alcanzan mejores indicadores globales de ajuste que la regresión lineal múltiple en los tres modelos evaluados; sin embargo, se plantea que su uso debe entenderse como una herramienta de apoyo para estimación preliminar y no como sustituto absoluto del estudio de métodos y tiempos.

Palabras clave: árboles de decisión, aprendizaje automático, estudios de tiempos, predicción, recubrimientos orgánicos, regresión lineal múltiple.

Abstract:

This article presents the development of predictive models for estimating operation times in an industrial process of organic coatings, applied to the registration of new products in a manufacturing plant. The central issue lies in the fact that, before incorporating new products into production lines, the organization must record a preliminary standard time in the information system, even though method and time studies are not yet available. To address this challenge, a database was consolidated from historical records, stopwatch measurements, and attributes associated with each reference. For the analysis, three product families were defined, considering fundamental aspects for their grouping, such as the surface area of the piece (expressed in square decimeters) and the number of units per hanger. Subsequently, two supervised techniques were compared: multiple linear regression and regression decision trees. This required the definition of data-cleaning criteria, training and testing protocols, as well as performance metrics, sensitivity analysis, overfitting diagnostics, and cross-validation against standard times previously defined by the organization. The results show that regression decision trees achieve better overall fit indicators than multiple linear regression across the three evaluated models; however, their use should be understood as a support tool for preliminary estimation rather than as an absolute substitute for method and time studies.

Keywords: decision trees, machine learning, multiple linear regression, organic coatings, prediction, time studies.

INTRODUCCIÓN

Una de las actividades más críticas en la ingeniería industrial es la estandarización de procesos, donde resulta esencial la estimación confiable de los tiempos de operación, ya que estos tiempos condicionan la planificación de la producción, la asignación de recursos, las capacidades productivas y la eficiencia de los procesos[1]. No obstante, en organizaciones manufactureras que tienen una gran cantidad de referencias, los estudios tradicionales de tiempos que se basan en observación directa y cronometraje siguen siendo utilizados por su validez técnica, aunque suelen ser costosos y poco escalables cuando se incorporan nuevos productos a las líneas de producción.

En el proceso de recubrimiento orgánicos estudiado, la estimación de tiempos de fabricación para nuevos productos es un requisito fundamental para su matrícula y la asignación de la ruta en el sistema de información. Sin embargo, realizar una estimación de tiempos tradicional antes de la matricula del nuevo producto no es viable, ya que este requerimiento se presenta antes de que este producto se incluya en las líneas de producción. Por ello, la estimación inicial se realiza a partir de analogías entre piezas con formas y características similares, experiencia del ingeniero responsable o criterios no estandarizados.

Sin embargo, la operación de recubrimientos orgánicos depende de características geométricas y productivas de la pieza, así como de condiciones de carga, manipulación, aplicación y secuencia de producción. En la planta estudiada, los ingenieros encargados del análisis de métodos y tiempos enfrentan una variabilidad importante asociada a referencias nuevas, códigos SAP, familias de producto, área superficial y unidades por ganchera. Esta situación limita la capacidad de estimar rápidamente tiempos normales para nuevos productos y genera dependencia de observaciones directas repetitivas.

Dado lo anterior, es fundamental identificar estudios e investigaciones que proponen nuevas alternativas a los métodos tradicionales para estimar tiempos, a continuación, se relacionan una serie de investigaciones relacionadas con el tema. Autores como Çakıt y Dağdeviren compararon algoritmos de machine learning para predecir tiempos estándar en un entorno manufacturero [2]. Backus et al. propusieron un enfoque de minería de datos para predecir tiempos de ciclos en fábricas de semiconductores [3], mientras que Meidan et al. integraron identificación de factores clave y predicción de tiempos de ciclo en manufactura de semiconductores [4]. En una línea similar, Öztürk et al. aplicaron minería de datos para estimar lead time de manufactura [5], y Lingitz et al. demostraron la utilidad de algoritmos de aprendizaje automático para predicción de lead time con datos reales de producción [6].

Rokoss et al. mencionan el uso de técnicas de aprendizaje automático en el campo de la planificación y el control de la producción ofrecen la oportunidad de obtener información valiosa y precisa sobre los procesos de producción [7], quedando demostrado en la investigación de Deepthi et al. en la cual se plantea un modelo de regresión lineal y un modelo de bosque aleatorio para optimizar el proceso [8]. Complementariamente, Flores - Huamán et al. realizaron un análisis de regresión de aprendizaje automático para predecir los tiempos de procesamiento y optimizar la asignación de recursos [9].

En el contexto regional se han realizado aportes relacionados con el uso de la analítica de datos, el aprendizaje autónomo y modelos estadísticos como herramientas de apoyo para los procesos de predicción, clasificación y mejoramiento de sistemas productivos [10], [11]. De igual manera se han documentado modelos de regresión basados en machine learning relacionados con los desafíos de la adopción de tecnologías provenientes de la industria 4.0 [12], [13]. Así mismo, se han desarrollado aplicaciones para soportar la evaluación y toma de decisiones capaces de realizar pronósticos y analítica de datos mediante trabajo supervisado [14], [15]. Otros autores se han enfocado en técnicas estadísticas y de regresión aplicadas a la optimización de procesos [16], [17], y al uso de tecnologías digitales avanzadas, como por ejemplo el caso de gemelos digitales, pare le mejoramiento de sistemas de producción [18]. Estos antecedentes refuerzan la oportunidad de estudiar modelos predictivos aplicados a contextos industriales reales en busca de disminuir la incertidumbre en la estimación de tiempos operativos y apoyar decisiones de planificación y mejoramiento de sistemas de producción.

La revisión de la literatura realizada permitió identificar investigaciones y modelos para predecir tiempos estándar en manufactura, así como la aplicación de aprendizaje autónomo en procesos industriales [19]. Sin embargo, también permitió identificar una oportunidad específica para evaluar modelos interpretables en la predicción de tiempos de operación de recubrimientos orgánicos. En particular dos métodos: primero, la regresión lineal múltiple ya que permite interpretar el efecto marginal de las variables independientes; segundo, los árboles de decisión ya que permiten capturar relaciones no lineales y reglas jerárquicas de decisión [20], [21], [22]. La comparación entre ambos modelos resulta pertinente porque combina aplicabilidad, facilidad de implementación y utilidad práctica para personal técnico de planta.

El objetivo del presente artículo es comparar la capacidad de la regresión lineal múltiple y de los árboles de decisión para estimar tiempos de operación en procesos de recubrimientos orgánicos, a partir de variables claves disponibles antes de que se realice el proceso de matrícula de nuevos productos. La contribución que brinda el estudio realizado es doble: por una parte, se propone una estructura metodológica reproducible para construir, depurar y validar el dataset; por otra, se discuten las implicaciones industriales del modelo como herramienta de apoyo a la planeación de capacidad, programación de operaciones y reducción de incertidumbre en la etapa de desarrollo de nuevos productos.

METODOLOGÍA

Para resolver la problemática planteada, se propone una investigación de estudio aplicado de modelado predictivo en un proceso real de manufactura. La metodología planteada comprendió seis etapas: i) entendimiento del proceso industrial; ii) recolección de datos; iii) limpieza y preparación del dataset; iv) análisis exploratorio y selección de variables; v) entrenamiento y validación de modelos predictivos; y vi) comparación de resultados, análisis de sensibilidad y discusión industrial. A continuación, se describen los principales aspectos desarrollados en cada una de las etapas.

Contexto industrial y proceso analizado

El proceso analizado corresponde a la aplicación de recubrimientos orgánicos sobre piezas industriales, las cuales deben pasar por una secuencia de operaciones en línea. Actualmente, el proceso cuenta con un ERP que asocia un registro de tiempo para cada operación de aplicación y referencia de código SAP. Cabe resaltar, que para el estudio realizado la variable dependiente es el tiempo de operación de recubrimiento, el cual se expresa en segundos / pieza. A su vez, la unidad de análisis es el registro individual de tiempo asociado a una pieza, referencia o carga de producción.

Dado que múltiples referencias o códigos SAP comparten atributos similares, la operación se agrupa en 3 familias de producto con características productivas comparables. Esta segmentación evita mezclar patrones de comportamiento heterogéneos y permite construir modelos por familia según el desempeño observado en la validación experimental.

Para mantener la trazabilidad industrial, cada observación se vinculó con atributos extraídos de los sistemas internos de la organización y de registros de campo: fecha, código SAP, familia de producto, material, tipo o color de pintura, turno, área superficial de la pieza en decímetros cuadrados, unidades por ganchera, colaborador y tiempo observado.

Adquisición, integración y trazabilidad de datos

La información se consolidó a partir de dos fuentes: registros históricos disponibles en la organización de los 2 últimos años (disponibles en el ERP de la organización) y mediciones cronometradas obtenidas durante trabajo de campo. Cada observación se asoció con un código SAP, lo cual permitió cruzar el tiempo de operación con atributos técnicos de la pieza: área superficial, unidades por ganchera, material, color, ruta de proceso, turno y referencia. Para reducir errores de digitación y facilitar la actualización futura, se estructuraron dos tablas relacionales: una tabla de registros individuales de tiempo y una tabla maestra de atributos por referencia.

La Tabla 1, presenta un resumen de la cantidad operaciones y registros para cada modelo planteado, así como el horizonte de tiempo de los registros utilizados para el análisis.

Tabla 1: Caracterización mínima del dataset que debe reportarse en la versión final.

Modelo	Operaciones variables incluidas	Registros iniciales (ERP + mediciones)	Registros excluidos	Registros finales	Periodo
1	FN, PI, B2	3.200	15	3.185	Enero / 2024-Diciembre / 2025
2	FN, PI, B1, B2	4.000	20	3.980	Enero / 2024-Diciembre / 2025
3	PM, B2	2.100	8	2.092	Enero / 2024-Diciembre / 2025

Limpieza de datos y tratamiento de valores atípicos

Los registros iniciales fueron depurados en dos niveles: primer nivel, observación directa, se excluyeron mediciones asociadas a eventos no cíclicos: obstrucción o desconfiguración de pistolas neumáticas, limpieza extraordinaria de boquillas, paradas por infraestructura, mantenimiento del sistema de aire, inicio de turno no estabilizado y ejecución por personal en entrenamiento; segundo nivel, se aplicó depuración estadística sobre la base consolidada aplicado el criterio de rango intercuartílico para cada modelo. Para cada familia, se calcularon el primer cuartil (Q1), el tercer cuartil (Q3) y el rango intercuartílico (IQR = Q3 - Q1). Los límites de aceptación se definieron mediante las ecuaciones 1 y 2.

$$LI = Q1 - 1.5(IQR) \quad (1)$$

$$LS = Q3 + 1.5(IQR) \quad (2)$$

La eliminación definitiva no se realizó de forma automática: cada observación extrema fue contrastada con la trazabilidad del registro para diferenciar errores de captura, paradas no cíclicas y variación operativa real. La Figura 1, ilustra el uso del diagrama de caja para identificar observaciones atípicas.

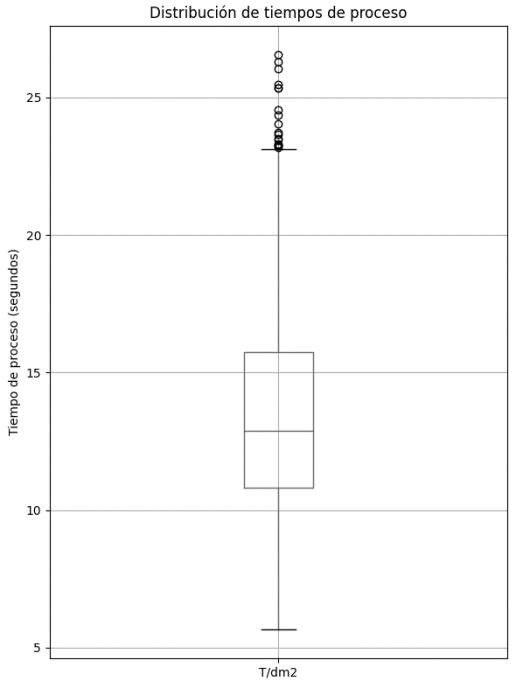

Fig. 1. *Ejemplo de distribución de tiempos normalizados e identificación de valores atípicos.*
Fuente: elaboración propia.

Definición y selección de variables

La selección de variables combinó criterio técnico de proceso y evidencia cuantitativa. En la caracterización se identificaron variables constantes, subjetivas o no disponibles de manera confiable para nuevas referencias; posteriormente, el análisis exploratorio evaluó la relación entre cada predictor y el tiempo de operación[23]. El análisis exploratorio incluyó diagramas de dispersión, matriz de correlación, coeficiente de Pearson para relaciones lineales, coeficiente de Spearman para relaciones monótonas, análisis de dependencia para variables categóricas y comparación de desempeño de modelos con y sin variables candidatas.

En la Figura 2 se relacionan los diagramas que hacen parte del análisis exploratorio, identificando como las variables decímetros cuadrados y unidades por ganchera mostraron una relación más consistente con el tiempo, mientras que color y material presentaron menor capacidad explicativa cuando se analizaron como variables aisladas en el conjunto disponible.

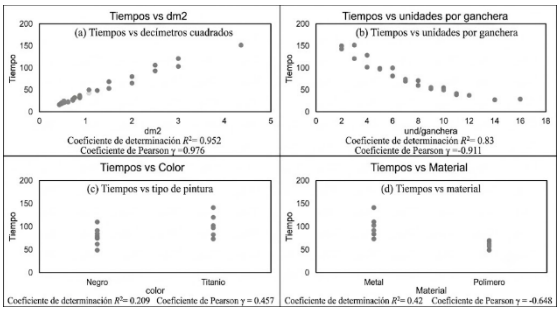

Fig. 2. *Análisis exploratorio de variables candidatas frente al tiempo de operación.*
Fuente: elaboración propia

La decisión final de incluir o excluir variables no se basó únicamente en inspección gráfica, sino en criterios cuantitativos y de interpretabilidad industrial. Dado lo anterior, en la Tabla 2 se consolida el análisis realizado para la selección de las variables que se incluyeron en los modelos planteados.

Modelos predictivos evaluados

En esta etapa se plantean dos técnicas supervisadas a evaluar y comparar: la primera fue la regresión lineal múltiple, seleccionada por su capacidad para cuantificar el efecto marginal de cada predictor y su facilidad de implementación en sistemas productivos; la segunda fue el árbol de decisión de regresión, seleccionado por su capacidad para capturar

segmentaciones no lineales y reglas operativas del tipo “si-entonces”, útiles cuando el comportamiento del proceso cambia por rangos de área o configuración de carga.

Tabla 2: Análisis cuantitativo y técnico para selección de variables

Variable candidata	Tipo	Evidencia cuantitativa disponible	Interpretación técnica	Decisión metodológica
Decímetros cuadrados	Cuantitativa continua	Pearson = 0,976; R^2 exploratorio = 0,952	Mayor superficie exige mayor recorrido de aplicación y mayor tiempo unitario.	Incluir como predictor principal
Unidades por ganchera	Cuantitativa discreta	Pearson = -0,911; R^2 exploratorio = 0,830	Mayor carga por ganchera distribuye el tiempo sobre más unidades y reduce el tiempo unitario.	Incluir como predictor principal
Tipo de pintura / color	Categórica	Pearson codificado = 0,457; R^2 exploratorio = 0,209	Puede afectar el proceso, pero su efecto queda parcialmente absorbido por la ruta/familia y no mostró tendencia suficiente.	Excluir del modelo base; evaluar en futuras versiones con mayor balance muestral
Material	Categórica	Pearson codificado = -0,648; R^2 exploratorio = 0,420	Tiene relación con manejabilidad, pero no fue estable como predictor único en la base disponible.	Excluir del modelo base; mantener para análisis futuro

Regresión lineal múltiple

La regresión lineal múltiple se utilizó como modelo base por su interpretabilidad y facilidad de implementación en ambientes industriales. Para cada observación i, el tiempo de operación y (i) se modeló como una función de las variables predictoras x_1, x_2, ..., x_k, de acuerdo con la ecuación 3.

$$y = b_0 + b_1 x_1 + \ldots + b_k x_k \qquad (3)$$

La variable dependiente es el tiempo de operación asociado al proceso variable de recubrimiento, cabe resaltar que se plantea un modelo de regresión para cada familia, es decir, en total se plantean 3 modelos de regresión. Las variables independientes del modelo base fueron decímetros cuadrados y unidades por ganchera, cuya selección se explicó en secciones anteriores. Para cada modelo, se define la ecuación estimada de la regresión lineal, los coeficientes, errores estándar, intervalos de confianza al 95 % y R^2.

Árbol de decisión de regresión

El árbol de decisión de regresión se empleó por su capacidad para representar relaciones no lineales y segmentar el espacio de predictores mediante reglas jerárquicas interpretables. En cada nodo del árbol, el algoritmo selecciona la variable y el punto de corte que minimizan la impureza del nodo, evaluada mediante el error cuadrático medio, como se expresa en(4).

$$MSE(S) = \frac{1}{|S|} \sum_{i \in s} (y_i - \overline{y}_s)^2 \qquad (4)$$

Para el árbol de decisión se debe reportar criterio de partición, profundidad máxima seleccionada, número mínimo de muestras por división, número de hojas y estrategia de validación. El procesamiento computacional de este modelo se realizó en Python para fines de reproducibilidad, en la Tabla 3 se resume los componentes utilizados, el uso específico en el estudio y la versión utilizada.

Protocolo experimental de entrenamiento, validación y prueba

Para garantizar la capacidad de adaptación del modelo se plantea un modelo experimental que se sustenta una partición hold-out donde el dataset se dividió en un 80% para entrenamiento y un 20% para prueba (el conjunto de prueba permaneció aislado para evaluar la generalización final). Adicionalmente, dentro del conjunto de entrenamiento se aplicó una validación cruzada k-fold (k = 5) para optimizar los hiperparámetros críticos del árbol de decisión: profundidad máxima (max_depth) y el número mínimo de muestras para división (min_samples_split).

Tabla 3: Resumen componentes y usos.

Componente	Uso en el estudio	Versión
Python	Procesamiento general y ejecución de modelos	3.11
pandas	Carga, depuración y combinación de tablas	2.1.1
NumPy	Operaciones numéricas y arreglos	1.26.0
scikit-learn	Regresión, árboles de decisión, partición, validación cruzada y métricas	1.3.1
statsmodels	Pruebas de significancia de la regresión lineal múltiple	0.14.0
matplotlib	Figuras y visualización de resultados	3.8.0

La evaluación de la capacidad de adaptación de los modelos se evaluó a través de métricas estadísticas y operativas. Se definieron cuatro métricas: primera, el coeficiente de determinación (R^2y R^2 ajustado) evalúa la variabilidad explicada por los predictores físicos; segunda, el Error Absoluto Medio (MAE) cuantifica la desviación real en segundos, facilitando la interpretación directa en los talleres; tercera, la Raíz del Error Cuadrático Medio (RMSE) se seleccionó para penalizar desviaciones de gran magnitud y detectar subgrupos mal modelados; cuarta, el Error Porcentual Absoluto Medio (MAPE)

funciona como un indicador adimensional analíticamente comparable entre referencias. En la Tabla 4, se relacionan las fases del protocolo experimental y sus métricas asociadas.

Tabla 4: Protocolo experimental sugerido para la versión final.

Fase	Propósito	Porcentaje / técnica	Salida que debe reportarse
Entrenamiento	Ajustar coeficientes o reglas del modelo	80 % de la base de datos de cada familia de producto	Métricas de ajuste y calibración inicial del modelo
Validación & Prueba	Seleccionar hiperparámetros y profundidad del árbol	-20 % reservado de forma estrictamente aislada -Validación cruzada k = 5 aplicada internamente	Media y desviación estándar de MAE, RMSE y R^2
	Evaluar generalización sobre datos no vistos		Métricas finales de error (MAE, RMSE, MAPE, R^2 de prueba)

RESULTADOS Y DISCUSIÓN

A continuación, se presentan los resultados del desempeño comparativo de los modelos, la validación cruzada, el diagnóstico de sobreajuste y las implicaciones industriales de los resultados.

Desempeño comparativo de modelos

Los resultados muestran que ambas técnicas (regresión lineal múltiple y arboles de decisiones) presenta un alto ajuste en las 3 familias de productos analizadas, sin embargo, cabe resaltar que la técnica de árbol de decisiones obtiene valores superiores de R^2 en los tres modelos evaluados. Estos resultados son coherentes y están alineados con la naturaleza del proceso, ya que la relación de la variable dependiente (tiempo de aplicación del recubrimiento) con las variables independientes (área y unidades por ganchera) no son necesariamente lineales. En la Tablas 5, se consolidad los resultados para cada uno de los modelos, técnicas y métricas utilizadas.

Tabla 5: Consolidado métricas desempeño de modelos y técnicas.

Mod.	Técnica	R^2 prueba	MAE	RMSE	MAPE (%)
1	RLM	0.948	4.25	5.48	5.41
1	AD	0.961	3.72	4.72	4.20
2	RLM	0.887	6.98	9.15	8.40
2	AD	0.898	6.08	7.62	7.10
3	RLM	0.941	4.02	5.12	4.80
3	AD	0.957	3.24	4.10	3.90

Nota: RLM = Regresión lineal múltiple; AD = Árbol de decisión.

Los resultados que se presentan en la Tabla 5, demuestran que ambas técnicas arrojan un ajuste adecuado en las tres familias de productos analizadas. No obstante, la técnica de árboles de decisión obtuvo mejores valores de R^2 y los errores fueron menores sobre datos no vistos, lo cual es consistente con la naturaleza del proceso, considerando que existe una relación no completamente lineal entre el tiempo de aplicación, el área superficial y las unidades por ganchera. En la Tabla 5 se recogen las métricas finales del conjunto de prueba, más relevantes para evaluar la capacidad de generalización.

La Tabla 5, permite evidenciar que la técnica de árbol de decisión arroja un mejor desempeño en los tres modelos evaluados, con mayores R^2 de prueba y menores valores de MAE, RMSE y MAPE frente a la regresión lineal múltiple. En el caso del Modelo 1, el MAPE disminuyó de 5.41 % a 4.20 %; en el caso del Modelo 2, de 8.40 % a 7.10 %; mientras que en el Modelo 3, de 4.80 % a 3.90 %. Esta reducción permite confirmar que el árbol de decisión captura mejor los cambios de comportamiento por rangos de superficie y carga por ganchera, aunque la regresión lineal múltiple sigue siendo útil como línea base interpretable y modelo de contraste técnico.

Para complementar el análisis de desempeño y la validación del ajuste de los modelos, se estimaron los coeficientes estructurales de la regresión lineal múltiple y se verificó su significancia individual. La Tabla 6, presenta únicamente los indicadores estadísticos esenciales: coeficiente estimado, error estándar, p-valor e intervalo de confianza al 95 %.

Tabla 6: Resultados compactos de significancia de la regresión lineal múltiple.

Mod	Predictor	β	Error Estándar	Valor p	IC 95 %
1	Intercepto	35.19	1.15	<0.001	[32.94; 37.44]
1	dm^2	30.22	0.78	<0.001	[28.69; 31.75]
1	und./gan.	-2.51	0.12	<0.001	[-2.75; -2.27]
2	Intercepto	42.85	2.10	<0.001	[38.73; 46.97]
2	dm^2	25.14	1.05	<0.001	[23.08; 27.20]
2	und./gan.	-1.89	0.18	<0.001	[-2.24; -1.54]
3	Intercepto	28.60	0.95	<0.001	[26.74; 30.46]
3	dm^2	34.78	0.65	<0.001	[33.51; 36.05]
3	und./gan.	-3.15	0.09	<0.001	[-3.33; -2.97]

Los resultados que se presentan en la Tabla 6, permiten confirmar que los predictores seleccionados son estadísticamente significativos en los tres modelos considerando que se obtiene un valor $p < 0.001$. Así mismo, el coeficiente positivo de dm^2 indica que el aumento del área superficial incrementa el tiempo de aplicación, mientras que el coeficiente negativo de las unidades por ganchera evidencia una reducción del tiempo unitario cuando aumenta la carga procesada por ciclo. Esta lectura

es coherente con la lógica física del proceso, es decir, piezas con superficie mayor necesitan más recorrido de aplicación, y una mayor cantidad de unidades por ganchera permite distribuir el tiempo de operación entre más piezas.

Validación cruzada con tiempos estándar existentes

Para evaluar la eficiencia de los modelos analíticos planteados en el entorno empresarial para la toma de decisiones, se realiza una validación externa en la que se comparan las estimaciones de cada uno de los modelos frente a los tiempos estándares de productos ya que están matriculados en el ERP de la organización. Los resultados de este análisis se consolidan en la Tabla 7.

Tabla 7: Mejor predicción externa frente a tiempos estándar definidos por la organización.

Mod	Ref.	Est. (s)	Técnica	Pred. (s)	Error (%)
1	456831	79,24	AD	80,49	1,6
1	456992	154,77	AD	148,84	3,8
2	456810	209,23	RLM	201,88	3,5
2	459660	73,74	RLM	66,53	9,8
3	456755	97,37	AD	102,48	5,2
3	456760	34,15	AD	32,60	4,5
3	459654	33,64	AD	32,31	3,9

Nota: Mod= Modelo; Est.= Tiempo estándar en segundos; AD= Árbol de decisión; RLM= Regresión lineal múltiple; Pred= Predicción en segundos; Error= Error porcentual absoluto.

Los resultados de la Tabla 7, evidencian una alta precisión en piezas que tienen geometrías similares, destacando la referencia 456831 del Modelo 1, donde el árbol de decisión registra una desviación marginal de apenas el 1.6% (80.49 s frente a 79.24 s). Este ejemplo demuestra la capacidad que tienen las reglas analíticas para representar la lógica de ingeniería y los tiempos estándares reales estimados por la organización.

Los resultados de la Tabla 7, indican que el árbol de decisión obtuvo el menor error en cinco de las siete referencias evaluadas, mientras que la regresión lineal múltiple presentó mejor desempeño en dos referencias del Modelo 2. En promedio, el error porcentual absoluto de la regresión lineal múltiple fue de 6.43 %, frente a 4.77 % para el árbol de decisión. La referencia 456831 del Modelo 1 ilustra el mejor comportamiento del árbol, con una desviación de 1.6 % respecto al estándar empresarial; sin embargo, los resultados del Modelo 2 indican que la regresión lineal aún puede ser competitiva cuando el comportamiento del proceso es más cercano a una relación lineal o cuando la variabilidad de la familia es menor. En consecuencia, la validación externa confirma la conveniencia de utilizar el árbol de decisión como alternativa principal, sin descartar la regresión lineal múltiple como modelo de contraste y control técnico.

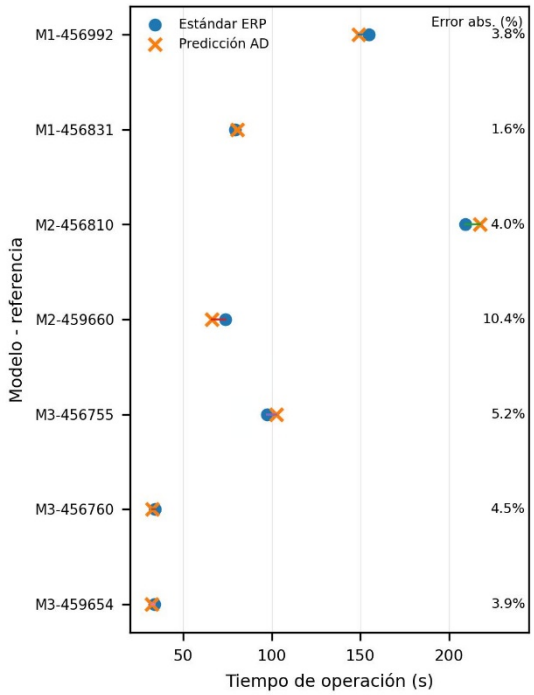

Fig. 3. *Comparación gráfica entre tiempos reales/estándar y tiempos predichos por los árboles de decisión..*
Fuente: elaboración propia

El análisis anterior se complementa con un análisis gráfico de la dispersión y el comportamiento del Error Porcentual Absoluto

Medio (MAPE) frente a los tiempos reales cronometrados para las tres (3) familias de productos, como el que se ilustra en la Figura 3.

En la Figura 3, se observa como la mayoría de las referencias se agrupan alrededor de la línea de concordancia ideal, confirmado visualmente los resultados estadísticos, donde se observaba que los modelos tienen un sesgo controlado y errores muy bajos. No obstante, la representación también expone puntos atípicos correspondientes a geometrías complejas con cavidades profundas, lo que facilita a los analistas de métodos y tiempos la identificación visual inmediata de las piezas que requerirán auditorías de campo complementarias.

Diagnóstico de sobreajuste

Un aspecto de gran importancia a evaluar es el sobreajuste (overfitting) que pueden llegar a tener los modelos de árboles de decisión planteados. Para garantizar y validar que los modelos propuestos no tengan sobreajuste se evaluó el comportamiento del error cuadrático medio (MSE) en función de la complejidad del algoritmo. Está evaluación se realizó mediante la constricción de curvas de aprendizaje, donde se contrastan el rendimiento obtenido en el conjunto de entrenamiento y el error de prueba para diferentes niveles de profundidad del árbol. En la Figura 4 se observan los resultados de la comparación realizada.

Como se observa en la Figura 4, a medida que se incrementa más allá del nivel óptimo, la discrepancia entre ambos errores tiende a estabilizarse. Lo anterior demuestra que los hiperparámetros seleccionados (profundidad máxima fija y el número mínimo de muestras por división) controlan eficazmente la varianza del modelo sin sacrificar el sesgo necesario para una predicción analítica robusta del proceso de recubrimiento.

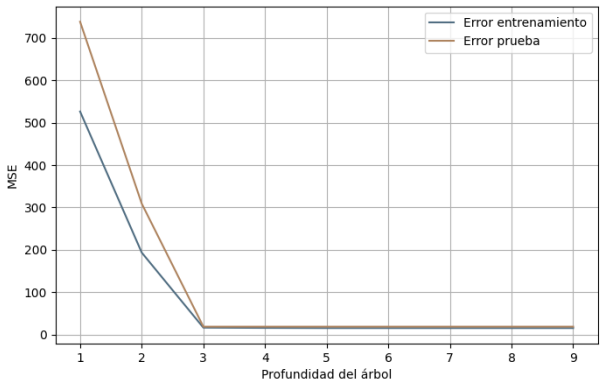

Fig. 4. *Diagnóstico de sobreajuste mediante comparación de error de entrenamiento y prueba por profundidad del árbol.*
Fuente: elaboración propia

Implicaciones industriales de los resultados

Desde la perspectiva de ingeniería industrial, los estudios de métodos y tiempos tradicionales seguirán siendo fundamentales para la estimación de tiempos estándares de producción, sin embargo, el valor agregado de estos modelos radica en apoyar la estimación preliminar y la actualización de tiempos cuando se tiene un gran volumen de datos históricos. A su vez, cabe resaltar que en entornos industriales en el que tiempo es muy valioso, los modelos propuestos permiten reducir el tiempo requerido para generar estimaciones iniciales respecto a técnicas tradicionales.

Sumado a lo anterior, en organizaciones el uso de estos modelos puede contribuir a mejorar al menor cinco (5) procesos de gestión: i) programación de producción, al estimar tiempos esperados para cargas futuras; ii) análisis de capacidad, al convertir características de producto en demanda de tiempo operativo; iii) cotización y costeo, al disminuir incertidumbre en tiempos de nuevos productos; iv) balanceo de cargas, al anticipar diferencias entre familias o referencias; y v) mejora continua, al identificar variables con mayor incidencia sobre el tiempo de operación.

De otro lado, es oportuno resaltar que la comparación con estudios previos confirma que es pertinente la predicción de tiempos de manufactura mediante aprendizaje automático solo cuando se dispone de una gran cantidad de datos históricos[1],[9],[24]. Adicionalmente, la aplicabilidad de los modelos planteados depende de la madurez y estabilidad del proceso, la consistencia de los registros, la actualización periódica ante cambios de producto, tecnología y método de aplicación.

CONCLUSIONES

La investigación define una metodología reproducible y rigurosa para la estimación de tiempos estándar preliminares en procesos de recubrimientos orgánicos para la planta objeto de estudio, permitiendo estimar tiempos de operación para la matrícula de nuevos productos. La robustez metodológica del estudio radica en la estructuración de un dataset cuidadosamente depurado bajo el criterio de rango intercuartílico, el cual integró 3.185 registros para la familia de productos 1, 3.980 registros

para la familia de productos 2 y 2.092 registros para la familia de productos 3. A su vez, el modelo experimental incorporó una partición técnica hold-out con validación cruzada, garantizando la evaluación explícita del desempeño predictivo y la trazabilidad en la depuración estadística de variables operativas sobre datos no vistos en la planta.

En cuanto al modelo de regresión lineal múltiple, este modelo aportó una alta interpretabilidad estadística, validando la significancia de los predictores claves seleccionados (área y unidades por ganchera) mediante coeficientes cuantitativos de correlación. Con relación a estos predictores, los decímetros cuadrados y las unidades por ganchera demostraron una relación consistente con el tiempo operativo, mientras que otras variables del proceso como el color o material no tiene una relación directa con los tiempos de aplicación de los recubrimientos. Sin embargo, es importante mencionar las limitaciones de la regresión lineal, ya que estas no permitieron absorber la complejidad de la variabilidad geométricas de las piezas en el modelo, teniendo en cuenta que se incrementó el error de este modelo respecto al resto de modelos.

En cuanto al algoritmo de árboles de decisión, este demostró una capacidad superior para representar relaciones no lineales y reglas jerárquicas. Los resultados confirmaron la superioridad de esta técnica respecto a la regresión lineal múltiple, alcanzando mejores ajustes en los tres (3) modelos planteados, estando en concordancia con los resultados de otras investigaciones que resaltan la disminución del error con el uso de estas técnicas[25]. De otro lado, al obtener un Error Absoluto Medio (MAE) de solo 3.65 segundos en el Modelo 1, el control del sobreajuste (hiperparámetros) posibilitó que el modelo plasmara reglas comprensibles y precisas para la naturaleza del proceso analizado. A su vez, la validación frente a tiempos estándar preexistentes en el sistema ERP demostró la viabilidad del árbol de decisión como herramienta predictiva para los tiempos de aplicación de recubrimientos en productos nuevos.

Finalmente, el modelo matemático propuesto es una alternativa que proporciona resultados con errores bajos, por lo que se consolida como una herramienta de soporte predictivo para optimizar la programación y planeación de capacidad y el proceso de recubrimientos[26]. Sin embargo, nuevamente se menciona que el modelo planteado no pretende sustituir la observación directa por cronometraje. Como limitación, el desempeño del modelo predictivo propuesto radica en el gran volumen de datos históricos que se tengan, así como de la fiabilidad de su registro y consolidación en bases de datos.

Limitaciones y Trabajo futuro

Como se indicó en la sección anterior, el análisis realizado está restringido por las condiciones observadas en planta y los datos que se habían recolectado anteriormente, por lo cual se sugiere prudencia a la hora de realizar generalizaciones con los resultados de este trabajo investigación. De igual manera, el estudio está sujeto a cambios en el método de trabajo, inclusión de tecnologías, mezcla de productos, experiencia de los operarios y otras variables que exógenas que afectan los procesos productivos.

Los trabajos futuros deben de orientase a ampliar la cantidad de datos con los que se cuenta, incorporar nuevas variables del proceso no registradas actualmente[27], estimar intervalos de predicción y construir mecanismos de monitoreo del error. Adicionalmente, se recomienda implementar una interfaz en la que se logre enlazar los tiempos de producción que se registran en el ERP con el modelo propuesto.

REFERENCIAS

[1] L. J. M. Meléndez, D. A. S. Chávez, and L. E. T. Mata, “El tiempo estándar y su importancia en las cotizaciones de proyectos de manufactura. Un enfoque de gestión,” *NovaRUA*, vol. 14, no. 24, pp. 110–122, Jan. 2022, doi: 10.20983/novarua.2022.24.6.

[2] E. Çakıt and M. Dağdeviren, “Comparative analysis of machine learning algorithms for predicting standard time in a manufacturing environment,” Artificial Intelligence for Engineering Design, Analysis and Manufacturing, vol. 37, e2, 2023, doi: 10.1017/S0890060422000245.

[3] P. Backus, M. Janakiram, S. Mowzoon, G. C. Runger, and A. Bhargava, “Factory cycle-time prediction with a data-mining approach,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 2, pp. 252-258, 2006, doi: 10.1109/TSM.2006.873400.

[4] Y. Meidan, B. Lerner, G. Rabinowitz, and M. Hassoun, “Cycle-Time Key Factor Identification and Prediction in Semiconductor Manufacturing Using Machine Learning and Data Mining,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 237-248, 2011, doi: 10.1109/TSM.2011.2118775.

[5] A. Öztürk, S. Kayalıgil, and N. E. Özdemirel, “Manufacturing lead time estimation using data mining,” *European Journal of Operational Research*, vol. 173, no. 2, pp. 683-700, 2006, doi: 10.1016/j.ejor.2005.03.015.

[6] L. Lingitz, V. Gallina, F. Ansari, D. Gyulai, A. Pfeiffer, W. Sihn, and L. Monostori, “Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer,” *Procedia CIRP*, vol. 72, pp. 1051-1056, 2018, doi: 10.1016/j.procir.2018.03.148.

[7] A. Rokoss, M. Syberg, L. Tomidei, C. Hülsing, J. Deuse, and M. Schmidt, “Case study on delivery time determination using a machine learning approach in small batch production companies,” *Journal of Intelligent Manufacturing*, vol. 35, pp. 3937–3958, 2024, doi: 10.1007/s10845-023-02290-2 (Springer Nature).

[8] Y. P. Deepthi, P. Kalaga, S. K. Sahu, J. J. Jacob, K. P. S., and Q. Ma, “AI-based machine learning prediction for optimization of copper

coating process on graphite powder for green composite fabrication," *International Journal on Interactive Design and Manufacturing*, vol. 19, pp. 4123-4130, 2025, doi: 10.1007/s12008-024-02032-5.

[9] K.-J. Flores-Huamán, A. Escudero-Santana, M.-L. Muñoz-Díaz, and P. Cortés, "Lead-Time Prediction in Wind Tower Manufacturing: A Machine Learning-Based Approach," *Mathematics*, vol. 12, no. 15, art. 2347, 2024, doi: 10.3390/math12152347.

[10] J. J. Paniagua Medina, E. Vargas Rodríguez, and R. Guzmán Cabrera, "Aprendizaje automático y la colección Reuters-21578 en la clasificación de documentos," *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 40, pp. 39–46, Jul. 2022, doi: 10.24054/rcta.v2i40.2344.

[11] A. A. Rosado Gómez, L. Calderón Benavides, and J. A. Parra, "Comparación empírica de dos modelos de aprendizaje automático generados mediante procesos diferentes," *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 1, no. 39, pp. 20–24, 2022, doi: 10.24054/rcta.v1i39.1369.

[12] F. A. Fernández-Gelvez, L. Jaimes-Cerveleón, and L. E. Mendoza, "Modelo de regresión basado en máquinas de aprendizaje utilizando datos estadísticos del café colombiano," *Mundo Fesc*, vol. 13, no. S1, pp. 258–272, 2023, doi: 10.61799/2216-0388.1499.

[13] J. C. Gutiérrez Medina, A. Martínez, and P. Alzate, "La Industria 4.0: Tendencias, barreras y retos en la cuarta revolución industrial," *Mundo Fesc*, vol. 14, no. 30, pp. 439–448, Sep. 2024, doi: 10.61799/2216-0388.1448.

[14] M. P. Brugés-Peláez, C. A. Parra-Ortega, and J. D. Ramón-Valencia, "Forecasting of particulate material concentration using supervised machine learning," *Respuestas*, vol. 29, no. 2, pp. 90–98, May 2024, doi: 10.22463/0122820X.5150.

[15] J. J. Castro-Maldonado, J. A. Patiño-Murillo, and E. Camargo-Casallas, "Aplicación de analítica de datos en la evaluación de los procesos de investigación aplicada y desarrollo experimental para fortalecer las competencias del siglo XXI en una institución de educación no formal," *Respuestas*, vol. 27, no. 2, pp. 6–26, May 2022, doi: 10.22463/0122820X.3541.

[16] L. D. Suárez-Riveros, W. Pineda-Ríos, and I. M. Mendivelso-Ramírez, "Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica," *Eco Matemático*, vol. 12, no. 2, pp. 112–125, Jul. 2021, doi: 10.22463/17948231.3323.

[17] J. R. Vera-Rozo, J. M. Riesco-Ávila, and A. Pardo-García, "Optimización mediante regresión polinomial del rendimiento líquido de la pirólisis de residuos plásticos recolectados en Norte de Santander," *Eco Matemático*, vol. 15, no. 2, pp. 76–82, Jul. 2024, doi: 10.22463/17948231.4999.

[18] A. Bustamante-Limones, C. Rodriguez-Borges, and J. A. Pérez-Rodriguez, "Evaluación del uso de gemelos digitales en los sistemas de producción," AiBi Revista de Investigación, Administración e Ingeniería, vol. 12, no. 3, pp. 195–204, Sep. 2024, doi: 10.15649/2346030X.4382.

[19] D. Arenas Seleey, C. E. Prieto Triana, y D. C. Chacón López, "Ingeniería de requerimientos e inteligencia artificial: una revisión de literatura", *Rev. Colomb. Tecnol. Avanzada*, vol. 1, n.º 39, pp. 101-107, 2022. Disponible: https://doi.org/10.24054/rcta.v1i39.1395

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Boca Raton, FL, USA: Chapman and Hall/CRC, 2017, doi: 10.1201/9781315139470.

[21] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An Introduction to Statistical Learning: with Applications in Python. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-38747-0.

[22] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-6849-3.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.

[24] A. C. Choueiri, D. M. V. Sato, E. E. Scalabrin, and E. A. P. Santos, "An extended model for remaining time prediction in manufacturing systems using process mining," *Journal of Manufacturing Systems*, vol. 56, pp. 188-201, 2020, doi: 10.1016/j.jmsy.2020.06.003.

[25] M. Alnahhal, D. Ahrens, and B. Salah, "Dynamic Lead-Time Forecasting Using Machine Learning in a Make-to-Order Supply Chain," *Applied Sciences*, vol. 11, no. 21, art. 10105, 2021, doi: 10.3390/app112110105.

[26] Y. Li, Z. Fu, X. Yu, Z. Jin, H. Gong, L. Ma, X. Li, and D. Zhang, "Developing an atmospheric aging evaluation model of acrylic coatings: A semi-supervised machine learning algorithm," International Journal of Minerals, Metallurgy and Materials, vol. 31, pp. 1617-1627, 2024, doi: 10.1007/s12613-024-2921-9.

[27] W. Chen et al., "Prediction of coating degradation based on 'Environmental Factors-Physical Property-Corrosion Failure' two-stage machine learning," *npj Materials Degradation*, vol. 9, art. 67, 2025, doi: 10.1038/s41529-025-00614-6.

FINANCIACIÓN

Ninguna

CONFLICTOS DE INTERÉS

No existen

CONTRIBUCIÓN DE AUTORÍA

Redacción - borrador original: Johan Sebastián Lavacude Galvis, Hugo Fernando Castro Silva, Josué Iván Mesa Mojica.

Redacción - revisión y edición: Johan Sebastián Lavacude Galvis, Hugo Fernando Castro Silva, Josué Iván Mesa Mojica.