ISSN: 1692-7257 - Volume 2 – Issue 48 - 2026

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

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

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

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.**

Received: 27 April 2026. Revised: 05 June 2026. Accepted: 25 June 2026. Published: 09 July 2026.
Corresponding Author: Hugo Fernando Castro Silva

How to cite: 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

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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.

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.

INTRODUCTION

One of the most critical activities in industrial engineering is process standardization, where reliable estimation of operating times is essential, since these times condition production planning, resource allocation, productive capacities, and process efficiency[1]. However, in manufacturing organizations that have a large number of references, traditional time studies based on direct observation and timekeeping are still used because of their technical validity, although they tend to be costly and poorly scalable when new products are incorporated into production lines.

In the organic coatings process studied, estimating manufacturing times for new products is a fundamental requirement for their registration and for assigning the route in the information system. However, carrying out a traditional time estimation before the new product is registered is not viable, since this requirement arises before the product is included in the production lines. The initial estimation is therefore based on analogies between parts with similar shapes and characteristics, on the experience of the engineer in charge, or on non-standardized criteria.

However, the organic coatings operation depends on the geometric and productive characteristics of the part, as well as on loading, handling, application, and production sequence conditions. In the plant studied, the engineers in charge of methods and time analysis face significant variability associated with new references, SAP codes, product families, surface area, and units per hanger. This situation limits the ability to quickly estimate normal times for new products and creates dependence on repetitive direct observations.

Given the above, it is essential to identify studies and research that propose new alternatives to traditional methods for estimating times; a series of research studies related to the topic are listed below. Authors such as Çakıt and Dağdeviren compared machine learning algorithms to predict standard times in a manufacturing environment [2]. Backus et al. proposed a data mining approach to predict cycle times in semiconductor factories [3], while Meidan et al. integrated identification of key factors and cycle time prediction in semiconductor manufacturing [4]. Along a similar line, Öztürk et al. applied data mining to estimate manufacturing lead time [5], and Lingitz et al. demonstrated the usefulness of machine learning algorithms for lead time prediction with real production data [6].

Rokoss et al. mention that the use of machine learning techniques in the field of production planning and control offers the opportunity to obtain valuable and accurate information about production processes [7], as demonstrated in the research by Deepthi et al., in which a linear regression model and a random forest model are proposed to optimize the process [8]. Complementarily, Flores-Huamán et al. conducted a machine learning regression analysis to predict processing times and optimize resource allocation [9].

In the regional context, contributions have addressed the use of data analytics, autonomous learning and statistical models as support tools for prediction, classification and improvement of production systems [10], [11]. Regression models based on machine learning have also been documented in relation to the challenges of adopting technologies from Industry 4.0 [12], [13]. Applications have likewise been developed to support evaluation and decision-making, capable of performing forecasts and data analytics through supervised work [14], [15]. Other authors have focused on statistical and regression techniques applied to process optimization [16], [17], and on the use of advanced digital technologies, such as digital twins, for the improvement of production systems [18]. This background reinforces the opportunity to study predictive models applied to real industrial contexts in order to reduce uncertainty in the estimation of operating times and to support planning and improvement decisions for production systems.

The literature review made it possible to identify research and models for predicting standard times in manufacturing, as well as the application of autonomous learning in industrial processes [19]. It also revealed a specific opportunity to evaluate interpretable models in the prediction of operating times of organic coatings. Two methods stand out: first, multiple linear regression, since it allows the marginal effect of the independent variables to be interpreted; second, decision trees, since they allow non-linear relationships and hierarchical decision rules to be captured [20], [21], [22]. Comparing the two models is pertinent because it combines applicability, ease of implementation and practical usefulness for plant technical personnel.

The objective of this article is to compare the capacity of multiple linear regression and decision trees to estimate operating times in organic coatings processes, based on key variables available before the registration of new products process is carried out. The study makes a twofold contribution: on the one hand, it proposes a reproducible methodological structure to build, clean and validate the dataset; on the other, it discusses the industrial implications of the model as a support tool for capacity planning, operations scheduling and reduction of uncertainty in the new product development stage.

METHODOLOGY

To address the problem posed, an applied research study of predictive modeling in a real manufacturing process is proposed. The proposed methodology comprised six stages: i) understanding of the industrial process; ii) data collection; iii) cleaning and preparation of the dataset; iv) exploratory analysis and selection of variables; v) training and validation of

predictive models; and vi) comparison of results, sensitivity analysis and industrial discussion. The main aspects developed in each stage are described below.

Industrial context and process analyzed

The process analyzed corresponds to the application of organic coatings on industrial parts, which must go through a sequence of in-line operations. Currently, the process has an ERP that associates a time record for each application operation and SAP code reference. It should be noted that for the study conducted, the dependent variable is the coating operation time, which is expressed in seconds / part. In turn, the unit of analysis is the individual time record associated with a part, reference or production load.

Since multiple references or SAP codes share similar attributes, the operation is grouped into 3 product families with comparable production characteristics. This segmentation avoids mixing heterogeneous behavioral patterns and allows models to be built per family according to the performance observed in the experimental validation.

To maintain industrial traceability, each observation was linked to attributes extracted from the organization's internal systems and from field records: date, SAP code, product family, material, paint type or color, shift, part surface area in square decimeters, units per hanger, worker and observed time.

Data acquisition, integration and traceability

The information was consolidated from two sources: historical records available in the organization for the last 2 years (available in the organization's ERP) and timed measurements obtained during field work. Each observation was associated with an SAP code, which made it possible to cross-reference the operation time with technical attributes of the part: surface area, units per hanger, material, color, process route, shift and reference. To reduce data entry errors and facilitate future updating, two relational tables were structured: a table of individual time records and a master table of attributes per reference.

Table 1 presents a summary of the number of operations and records for each proposed model, as well as the time horizon of the records used for the analysis.

Table 1: Minimum characterization of the dataset to be reported in the final version.

Model	Variables included	Initial records (ERP + measurements)	Excluded records	Final records	Period
1	FN, PI, B2	3.200	15	3.185	January / 2024-December / 2025
2	FN, PI, B1, B2	4.000	20	3.980	January / 2024-December / 2025
3	PM, B2	2.100	8	2.092	January / 2024-December / 2025

Data cleaning and treatment of outliers

The initial records were cleaned at two levels. At the first level, direct observation, measurements associated with non-cyclic events were excluded: obstruction or misconfiguration of pneumatic guns, extraordinary nozzle cleaning, infrastructure stoppages, air system maintenance, unstabilized start of shift and execution by personnel in training. At the second level, statistical cleaning was applied to the consolidated database using the interquartile range criterion for each model. For each family, the first quartile (Q1), the third quartile (Q3) and the interquartile range (IQR = Q3 - Q1) were calculated. The acceptance limits were defined by equations 1 and 2.

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

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

Final removal was not performed automatically: each extreme observation was checked against the record traceability to differentiate data-entry errors, non-cyclic stoppages and real operational variation. Figure 1 illustrates the use of the box plot to identify atypical observations.

Definition and selection of variables

The selection of variables combined technical process criteria and quantitative evidence. The characterization identified variables that were constant, subjective or not reliably available for new references; subsequently, the exploratory analysis evaluated the relationship between each predictor and the operation time [23]. The exploratory analysis included scatter plots, a correlation matrix, the Pearson coefficient for linear relationships, the Spearman coefficient for monotonic relationships, dependence analysis for categorical variables and a comparison of model performance with and without candidate variables.

Figure 2 shows the plots that form part of the exploratory analysis, identifying how the variables square decimeters and units per hanger showed a more consistent relationship with time, whereas color and material presented lower explanatory capacity when analyzed as isolated variables in the available set.

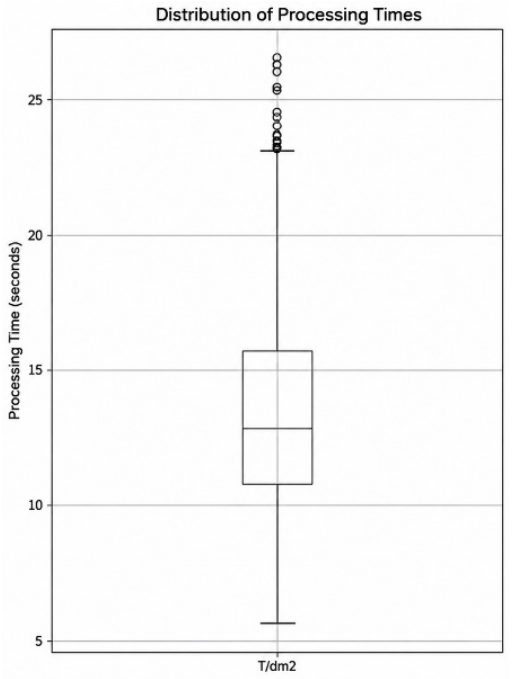

Fig. 1. *Example of the distribution of normalized times and identification of outliers.*
Source: own elaboration.

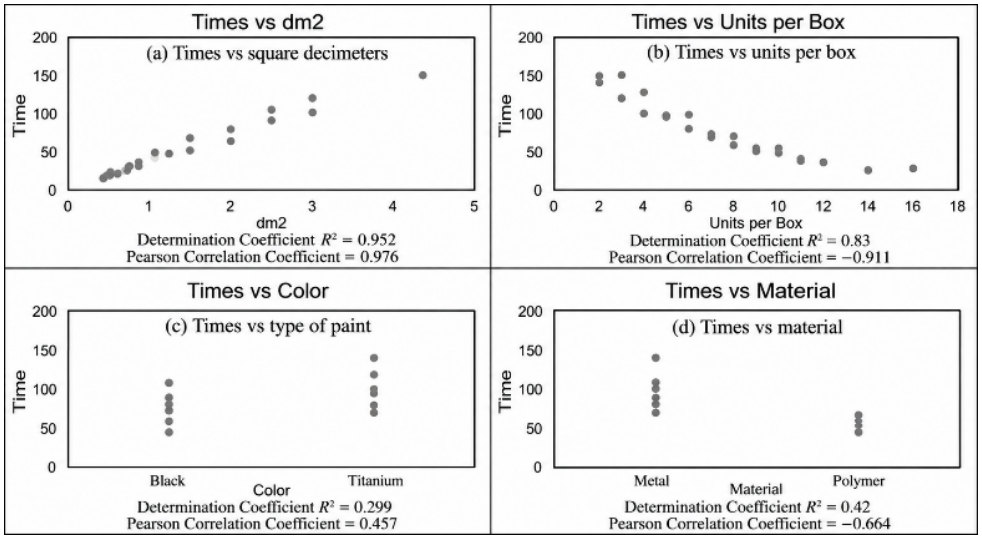

Fig. 2. *Exploratory analysis of candidate variables versus operation time.*
Source: own elaboration

The final decision to include or exclude variables was based not only on graphical inspection but also on quantitative criteria and industrial interpretability. Accordingly, Table 2 consolidates the analysis carried out to select the variables included in the proposed models.

Table 2: Quantitative and technical analysis for variable selection

Candidate variable	Type	Available quantitative evidence	Technical interpretation	Methodological decision
Square decimeters	Continuous quantitative	Pearson = 0.976; exploratory R^2 = 0.952	A larger surface requires a longer application path and a longer unit time.	Include as main predictor
Units per hanger	Discrete quantitative	Pearson = -0.911; exploratory R^2 = 0.830	A higher load per hanger spreads the time over more units and reduces unit time.	Include as main predictor
Paint type / color	Categorical	Coded Pearson = 0.457; exploratory R^2 = 0.209	It may affect the process, but its effect is partially absorbed by route/family and it did not show a sufficient trend.	Exclude from the base model; evaluate in future versions with greater sample balance
Material	Categorical	Coded Pearson = -0.648; exploratory R^2 = 0.420	It is related to handleability, but it was not stable as a single predictor in the available database.	Exclude from the base model; retain for future analysis

Predictive models evaluated

At this stage, two supervised techniques are proposed for evaluation and comparison: the first was multiple linear regression, selected for its capacity to quantify the marginal effect of each predictor and its ease of implementation in production systems; the second was the regression decision tree, selected for its capacity to capture non-linear segmentations and operational "if-then" rules, useful when process behavior changes by area ranges or load configuration.

Multiple linear regression

Multiple linear regression was used as the baseline model because it is interpretable and easy to implement in industrial environments. For each observation i, the operation time y (i) was modeled as a function of the predictor variables x_1, x_2, ..., x_k, according to equation 3.

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

The dependent variable is the operation time associated with the variable coating process. A regression model is proposed for each family, that is, 3 regression models in total. The independent variables of the baseline model were square decimeters and units per hanger, whose selection was explained in previous sections. For each model, the estimated linear regression equation, the coefficients, standard errors, 95% confidence intervals and R^2 are defined.

Regression decision tree

The regression decision tree was used for its capacity to represent non-linear relationships and to segment the predictor space through interpretable hierarchical rules. At each node of the tree, the algorithm selects the variable and the split point that minimize node impurity, evaluated by means of the mean squared error, as expressed in(4).

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

For the decision tree, the splitting criterion, the selected maximum depth, the minimum number of samples per split, the number of leaves and the validation strategy must be reported. This model was processed computationally in Python for reproducibility purposes; Table 3 summarizes the components used, their specific use in the study and the version used.

Table 3: Summary of components and uses.

Component	Use in the study	Version
Python	General processing and model execution	3.11
pandas	Loading, cleaning, and merging tables	2.1.1
NumPy	Numerical operations and arrays	1.26.0
scikit-learn	Regression, decision trees, partitioning, cross-validation, and metrics	1.3.1
statsmodels	Significance tests for multiple linear regression	0.14.0
matplotlib	Figures and visualization of results	3.8.0

Experimental protocol for training, validation, and testing

To ensure the adaptability of the model, an experimental model is proposed that relies on a hold-out partition in which the dataset was split into 80% for training and 20% for testing (the test set remained isolated to evaluate final generalization). In addition, within the training set, k-fold cross-validation (k = 5) was applied to optimize the critical hyperparameters of the decision tree: maximum depth (max_depth) and the minimum number of samples for splitting (min_samples_split).

The adaptability of the models was assessed through statistical and operational metrics. Four metrics were defined: first, the coefficient of determination (R^2 and adjusted R^2) evaluates the variability explained by the physical predictors; second, the mean absolute error (MAE) quantifies the actual deviation in seconds, facilitating direct interpretation in the workshops; third, the root mean square error (RMSE) was selected to penalize large-magnitude deviations and detect poorly modeled subgroups; fourth, the mean absolute percentage error (MAPE) functions as a dimensionless indicator that is analytically comparable between references. Table 4 lists the phases of the experimental protocol and their associated metrics.

Table 4: Suggested experimental protocol for the final version.

Phase	Purpose	Percentage / technique	Output to be reported
Training	Fit model coefficients or rules	80 % of the database for each product family	Fit metrics and initial model calibration
Validation & Test	Select hyperparameters and tree depth	-20 % held out in strict isolation -k = 5 cross-validation applied internally	Mean and standard deviation of MAE, RMSE and R^2
	Assess generalization on unseen data		Final error metrics (MAE, RMSE, MAPE, test R^2)

RESULTS AND DISCUSSION

The following sections present the comparative performance results of the models, the cross-validation, the overfitting diagnostics and the industrial implications of the results.

Comparative performance of models

The results show that both techniques (multiple linear regression and decision trees) present a high fit in the 3 product families analyzed; however, it should be noted that the decision tree technique obtains higher R^2 values in the three models evaluated. These results are coherent and aligned with the nature of the process, since the relationship of the dependent variable (coating application time) with the independent variables (area and units per hanger) is not necessarily linear. Table 5 consolidates the results for each of the models, techniques and metrics used.

Table 5: Consolidated performance metrics of models and techniques.

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

Note: MLR = Multiple linear regression; DT = Decision tree.

The results presented in Table 5 show that both techniques yield an adequate fit for the three product families analyzed. However, the decision trees technique obtained better R^2 values and lower errors on unseen data, which is consistent with the nature of the process, considering that there is a not fully linear relationship among application time, surface area and units per hanger. Table 5 compiles the final test set metrics, the most relevant for evaluating generalization capacity.

Table 5 shows that the decision tree technique yields better performance in the three models evaluated, with higher test R^2 and lower MAE, RMSE and MAPE values compared with multiple linear regression. In the case of Model 1, MAPE decreased from 5.41 % to 4.20 %; in the case of Model 2, from 8.40 % to 7.10 %; while in Model 3, from 4.80 % to 3.90 %. This reduction confirms that the decision tree better captures behavioral changes by surface ranges and load per hanger, although multiple linear regression remains useful as an interpretable baseline and technical contrast model.

To complement the performance analysis and the validation of the model fit, the structural coefficients of the multiple linear regression were estimated and their individual significance was verified. Table 6 presents only the essential statistical indicators: estimated coefficient, standard error, p-value and 95 % confidence interval.

The results presented in Table 6 confirm that the selected predictors are statistically significant in the three models, considering that a p-value < 0.001 is obtained. Likewise, the positive coefficient of dm^2 indicates that an increase in surface area increases the application time, while the negative coefficient of units per hanger evidences a reduction in unit time when the load processed per cycle increases. This reading is coherent with the physical logic of the process, that is, parts with a larger surface require a longer application path, and a larger number of units per hanger allows the operation time to be distributed among more parts.

Table 6: Compact significance results of the multiple linear regression.

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

Cross-validation with existing standard times

To evaluate the efficiency of the analytical models proposed for decision-making in the business environment, an external validation is carried out in which the estimates of each model are compared against the standard times of products, since these are registered in the organization's ERP. The results of this analysis are consolidated in Table 7.

Table 7: Best external prediction versus standard times defined by the organization.

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

Note: Mod= Model; ST= Standard time in seconds; DT= Decision tree; MLR= Multiple linear regression; Pred= Prediction in seconds; Error= Absolute percentage error.

The results in Table 7 show high accuracy for parts with similar geometries, notably reference 456831 of Model 1, where the decision tree records a marginal deviation of only 1.6% (80.49 s versus 79.24 s). This example demonstrates the capacity of analytical rules to represent engineering logic and the actual standard times estimated by the organization.

The results in Table 7 indicate that the decision tree obtained the lowest error in five of the seven references evaluated, whereas multiple linear regression performed better in two references of Model 2. On average, the absolute percentage error of multiple linear regression was 6.43 %, compared with 4.77 % for the decision tree. Reference 456831 of Model 1 illustrates the best performance of the tree, with a deviation of 1.6 % with respect to the company standard; however, the results of Model 2 indicate that linear regression can still be competitive when the process behavior is closer to a linear relationship or when the variability of the family is lower. Consequently, external validation confirms the suitability of using the decision tree as the main alternative, without discarding multiple linear regression as a contrast and technical control model.

The above analysis is complemented by a graphical analysis of the dispersion and the behavior of the Mean Absolute Percentage Error (MAPE) against the actual timed times for the three (3) product families, such as the one illustrated in Figure 3.

In Figure 3, most of the references cluster around the ideal agreement line, visually confirming the statistical results, which showed that the models have a controlled bias and very low errors. Nevertheless, the representation also exposes outlier points corresponding to complex geometries with deep cavities, which facilitates the immediate visual identification by methods and time analysts of the parts that will require complementary field audits.

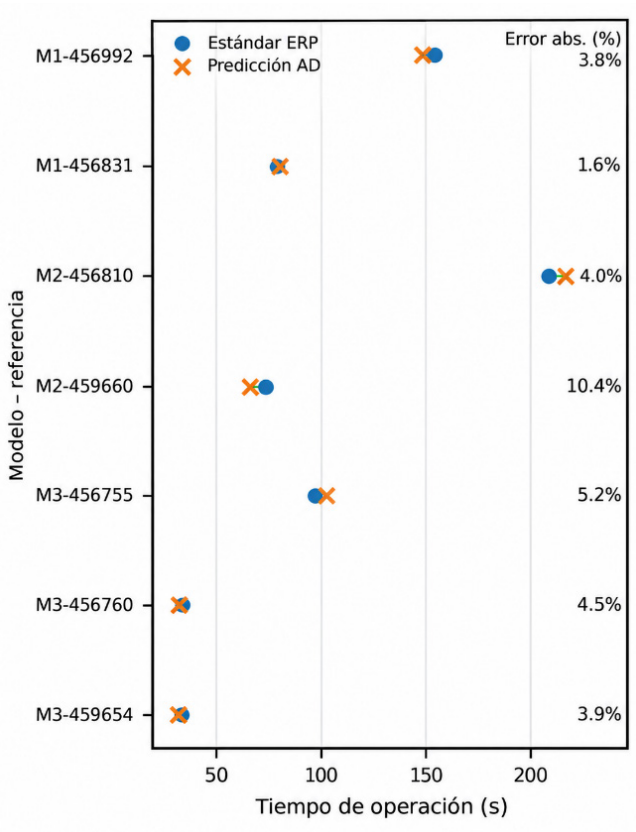

Fig. 3. *Graphical comparison between actual/standard times and times predicted by the decision trees.*
Source: own elaboration

Overfitting diagnostics

An aspect of great importance to evaluate is the overfitting that the proposed decision tree models may reach. To guarantee and validate that the proposed models do not have overfitting, the behavior of the mean squared error (MSE) was evaluated as a function of the complexity of the algorithm. This evaluation was carried out by constructing learning curves, where the performance obtained in the training set and the test error are compared for different levels of tree depth. Figure 4 shows the results of the comparison performed.

As observed in Figure 4, as it increases beyond the optimal level, the discrepancy between both errors tends to stabilize. This demonstrates that the selected hyperparameters (fixed maximum depth and the minimum number of samples per split)

effectively control the variance of the model without sacrificing the bias necessary for a robust analytical prediction of the coating process.

Industrial implications of the results

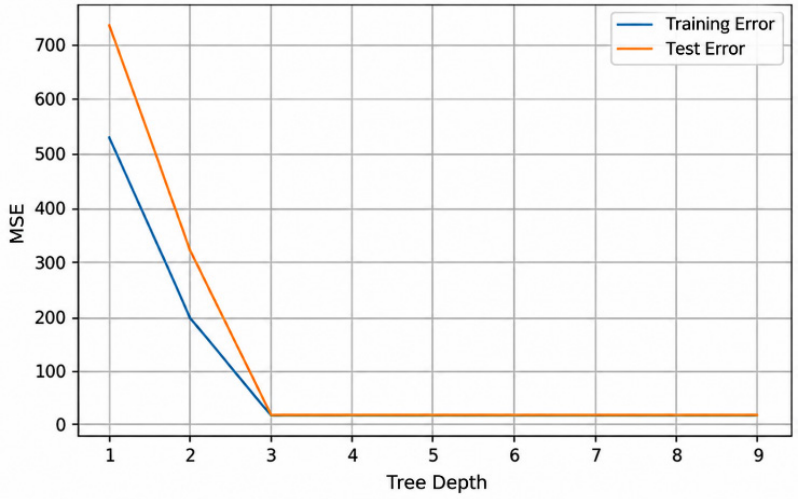

Fig. 4. *Overfitting diagnostics by comparison of training and test error by tree depth.*
Source: own elaboration

From an industrial engineering perspective, traditional method and time studies will continue to be essential for estimating standard production times; however, the added value of these models lies in supporting the preliminary estimation and updating of times when a large volume of historical data is available. In industrial environments where time is highly valuable, the proposed models also make it possible to reduce the time required to generate initial estimates compared with traditional techniques.

In addition to the above, in organizations the use of these models can help improve at least five (5) management processes: i) production scheduling, by estimating expected times for future loads; ii) capacity analysis, by converting product characteristics into operating time demand; iii) quoting and costing, by reducing uncertainty in times for new products; iv) load balancing, by anticipating differences between families or references; and v) continuous improvement, by identifying the variables with the greatest influence on operating time.

On the other hand, the comparison with previous studies confirms that the prediction of manufacturing times through machine learning is appropriate only when a large amount of historical data is available[1],[9],[24]. The applicability of the proposed models also depends on the maturity and stability of the process, the consistency of the records, and periodic updating in response to changes in product, technology, and application method.

CONCLUSIONS

The research defines a reproducible and rigorous methodology for the estimation of preliminary standard times in organic coatings processes for the plant under study, making it possible to estimate operating times for the registration of new products. The methodological robustness of the study lies in the structuring of a carefully refined dataset under the interquartile range criterion, which integrated 3,185 records for product family 1, 3,980 records for product family 2, and 2,092 records for product family 3. The experimental model also incorporated a technical hold-out partition with cross-validation, guaranteeing the explicit evaluation of predictive performance and traceability in the statistical refinement of operational variables on data not seen in the plant.

Regarding the multiple linear regression model, this model provided high statistical interpretability, validating the significance of the key predictors selected (area and units per hanger) through quantitative correlation coefficients. With respect to these predictors, square decimeters and units per hanger showed a consistent relationship with operating time, whereas other process variables such as color or material do not have a direct relationship with coating application times. However, it is important to mention the limitations of linear regression, since it did not allow the complexity of the geometric variability of the parts to be absorbed in the model, considering that the error of this model increased relative to the rest of the models.

The decision trees algorithm showed a superior capacity to represent non-linear relationships and hierarchical rules. The results confirmed the superiority of this technique over multiple linear regression, achieving better fits in all three (3) proposed

models, in agreement with the results of other research that report a reduction of error with the use of these techniques[25]. By obtaining a mean absolute error (MAE) of only 3.65 seconds in Model 1, the control of overfitting (hyperparameters) allowed the model to capture comprehensible and precise rules for the nature of the process analyzed. Validation against standard times pre-existing in the ERP system also showed that the decision tree is a viable predictive tool for coating application times in new products.

Finally, the proposed mathematical model is an alternative that provides results with low errors, and it is therefore consolidated as a predictive support tool to optimize capacity scheduling and planning and the coating process[26]. However, it is again mentioned that the proposed model is not intended to replace direct observation by time study. As a limitation, the performance of the proposed predictive model depends on the large volume of historical data available, as well as on the reliability of their recording and consolidation in databases.

Limitations and Future Work

As indicated in the previous section, the analysis carried out is restricted by the conditions observed in the plant and the data that had been collected previously, so caution is advised when generalizing the results of this research work. Likewise, the study is subject to changes in the work method, the inclusion of technologies, the product mix, the experience of the operators and other exogenous variables that affect the production processes.

Future work should aim to expand the amount of data available, incorporate new process variables not currently recorded[27], estimate prediction intervals and build error monitoring mechanisms. Additionally, an interface should be implemented in which the production times recorded in the ERP can be linked with the proposed model.

REFERENCES

[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.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Johan Sebastián Lavacude Galvis, Hugo Fernando Castro Silva, Josué Iván Mesa Mojica.

Writing–review and editing: Johan Sebastián Lavacude Galvis, Hugo Fernando Castro Silva, Josué Iván Mesa Mojica.