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

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i48.4500

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

Ing. Johan Sebastián Lavacude Galvis[1], PhD. Hugo Fernando Castro Silva[2], MSc. 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.*

Correspondence: *hugofernado.castro@uptc.edu.co*

Received: *april 27, 2026.* ***Accepted:*** *june 25, 2026.* ***Published:*** *july 09, 2026.*

How to cite: *J. S. Lavacude Galvis, H. F. Castro Silva, and J. I. Mesa Mojica, "Prediction of operating times in organic coatings using multiple linear regression and decision trees", RCTA, vol. 2, n.º. 48, pp. 102–112, jul. 2026. Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/4500*

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.

1. INTRODUCTION

One of the most critical activities in industrial engineering is the standardization of processes, where reliable estimation of operating times is essential, since these times condition production planning, resource allocation, production capacities and process efficiency [1]. However, in manufacturing organizations that have a large number of references, traditional time studies that rely on direct observation and timekeeping are still used for their technical validity, although they are often costly and not very scalable when new products are added to production lines.

In the organic coating process studied, the estimation of manufacturing times for new products is a fundamental requirement for their registration and the assignment of the route in the information system. However, making a traditional time estimate before the registration of the new product is not feasible, since this requirement is presented before this product is included in the production lines. Therefore, the initial estimate is made based on analogies between parts with similar shapes and characteristics, experience of the responsible engineer or non-standardized criteria.

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

Given the above, it is essential to identify studies and research that propose new alternatives to traditional methods to estimate times, below, a series of research related to the topic is related. 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 key factor identification and cycle time prediction in semiconductor manufacturing [4]. In a similar vein, 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 of Deepthi et al. in which a linear regression model and a random forest model are proposed to optimize the process [8]. In addition, Flores - Huamán et al. performed a machine learning regression analysis to predict processing times and optimize resource allocation [9].

In the regional context, contributions have been made related to the use of data analytics, autonomous learning, and statistical models as

support tools for the processes of prediction, classification, and improvement of production systems [10], [11]. Similarly, regression models based on machine learning have been documented related to the challenges of adopting technologies from industry 4.0 [12], [13]. Likewise, applications have been developed to support evaluation and decision-making capable of making 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 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 support planning decisions and improvement of production systems.

The literature review carried out allowed the identification of research and models to predict standard times in manufacturing, as well as the application of autonomous learning in industrial processes [19]. However, it also allowed to identify a specific opportunity to evaluate interpretable models in the prediction of operating times of organic coatings. In particular, two methods: first, multiple linear regression since it allows interpreting the marginal effect of the independent variables; second, decision trees since they allow capturing nonlinear relationships and hierarchical decision rules [20], [21], [22]. The comparison between 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 ability of multiple linear regression and decision trees to estimate operating times in organic coating processes, based on key variables available before the registration process of new products is carried out. The contribution provided by the study is twofold: on the one hand, it proposes a reproducible methodological structure to build, debug and validate the dataset; on the other hand, the industrial implications of the model as a tool to support capacity planning, operations scheduling and uncertainty reduction in the development stage of new products are discussed.

2. METHODOLOGY

To solve the problem raised, an applied study of predictive modelling in a real manufacturing process is proposed. The proposed methodology included 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 of the stages are described below.

2.1. Industrial context and process analysed

The process analysed 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 carried out, the dependent variable is the coating operation time, which is expressed in seconds/piece. In turn, the unit of analysis is the individual record of time associated with a part, reference or production load.

Since multiple SKUs or SAP codes share similar attributes, the operation is grouped into 3 product families with comparable production characteristics. This segmentation avoids mixing heterogeneous behaviour patterns and allows models to be built by 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 field records: date, SAP code, product family, material, paint type or color, shift, surface area of the part in square decimetres, units per hook, collaborator, and time observed.

2.2. 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 allowed the operation time to be cross-referenced with technical attributes of the part: surface area, units per hook, material, color, process route, shift and reference. To reduce typing errors and facilitate future updating, two relational tables were structured: a table of individual time records and a master table of attributes by reference.

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

Table 1: *Minimum characterization of the dataset that must be reported in the final version.*

Model	Variable operations included	Initial Registrations (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

2.3. Data cleansing and outlier handling

The initial records were filtered at two levels: first level, direct observation, measurements associated with non-cyclical events were excluded: obstruction or misconfiguration of pneumatic guns, extraordinary cleaning of nozzles, shutdowns due to infrastructure, maintenance of the air system, non-stabilized shift start and execution by personnel in training; At the second level, statistical filtering was applied on the consolidated basis by applying 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 using equations 1 and 2.

$$LI = Q1 - 1.5(IQR) \quad (1)$$
$$LS = Q3 + 1.5(IQR) \quad (2)$$

The final elimination was not carried out automatically: each extreme observation was contrasted with the traceability of the record to differentiate capture errors, non-cyclical stoppages and real operational variation. Figure 1 illustrates the use of the box plot to identify atypical observations.

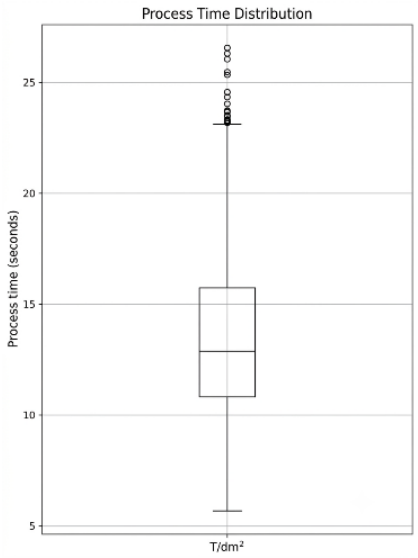

Fig. 1*. Example of normalized time distribution and identification of outliers.*
Source: *Authors.*

2.4. Definition and selection of variables

The selection of variables combined technical process criteria and quantitative evidence. In the characterization, constant, subjective or unreliable variables were identified for new references; Subsequently, the exploratory analysis evaluated the relationship between each predictor and operating time [23]. The exploratory analysis included scatter plots, correlation matrix, Pearson's coefficient for linear relationships, Spearman's coefficient for monotonic relationships, dependency analysis for categorical variables, and comparison of model performance with and without candidate variables.

Figure 2 relates the diagrams that are part of the exploratory analysis, identifying how the variables decimetres squares and units per hook showed a more consistent relationship with time, while color and material showed less explanatory capacity when analysed as isolated variables in the available set.

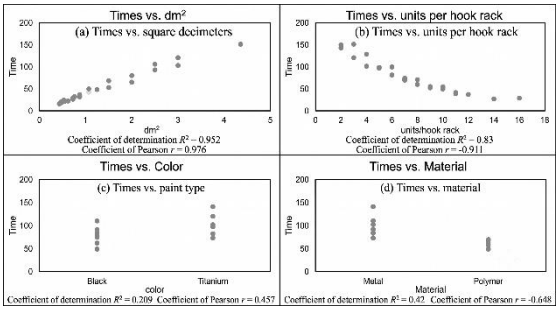

Fig. 2*. Exploratory analysis of candidate variables versus operating time.*
Source: *own elaboration*

The final decision to include or exclude variables was not based solely on graphic inspection, but on quantitative and industrial interpretability criteria. Given the above, Table 2 consolidates the analysis carried out for the selection of the variables that were included in the proposed models.

Table 2: *Quantitative and Technical Analysis for Variable Selection*

Candidate variable	Type	Quantitative evidence available	Technical interpreting	Methodological decision
Square decimeters	Continuous quantitative	Pearson = 0.976; Exploratory R^2 = 0.952	A larger surface area requires a longer application path and a longer unit time.	Include as a primary predictor
Units per hook	Discrete quantitative	Pearson = -0.911; Exploratory R^2 = 0.830	Higher load per hook distributes time over more units and reduces unit time.	Include as a primary predictor
Paint Type / Color	Categorical	Pearson coded = 0.457; Exploratory R^2 = 0.209	It can affect the process, but its effect is partially absorbed by the route/family and did not show sufficient tendency.	Exclude from base model; evaluate in future versions with a higher sample balance
Material	Categorical	Pearson coded = -0.648; Exploratory R^2 = 0.420	It is related to manageability, but it was not stable as the only predictor in the available database.	Exclude from base model; Keep for future analysis

2.5. Predictive models evaluated

In this stage, two supervised techniques are proposed to be evaluated and compared: the first was multiple linear regression, selected for its ability 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 ability to capture nonlinear segmentations and "if-then" operating rules, useful when process behavior changes by area ranges or load configuration.

2.5.1. Multiple linear regression

Multiple linear regression was used as a base model due to its interpretability and ease of implementation in industrial environments. For each observation i, the operating time and (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 operating time associated with the variable coating process, it should be noted that a regression model is proposed for each family, that is, in total 3 regression models are proposed. The independent variables of the base model were square decimeters and units per hook, the selection of which was explained in previous sections. For each model, the estimated equation of linear regression, coefficients, standard errors, 95% confidence intervals, and R^2 are defined.

2.5.2. Regression decision tree

The regression decision tree was used for its ability to represent nonlinear relationships and segment the predictor space using interpretable hierarchical rules. At each node in the tree, the algorithm selects the variable and cut-off point that minimize the node's impurity, as assessed by the mean square error, as expressed in (4).

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

For the decision tree, the partition criteria, maximum depth selected, minimum number of samples per division, number of leaves and validation strategy must be reported. The computational processing of this model was performed in Python for reproducibility purposes, Table 3 summarizes the components used, the specific use in the study and the version used.

Table 3: *Summary of components and uses.*

Component	Use in the studio	Version
Python	General Processing and Model Execution	3.11
pandas	Loading, Debugging, 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
State models	Multiple linear regression significance tests	0.14.0
matplotlib	Figures and Visualization of Results	3.8.0

2.6. Experimental training, validation and testing protocol

To guarantee the adaptability of the model, an experimental model is proposed that is supported by a hold-out partition where the dataset was divided into 80% for training and 20% for test (the test set remained isolated to evaluate the final generalization). Additionally, within the training set, a 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 division (min_samples_split).

The evaluation of the adaptability of the models was evaluated through statistical and operational metrics. Four metrics were defined: first, the coefficient of determination (adjusted R2 and R2) evaluates the variability explained by the physical predictors; second, the Mean Absolute Error (MAE) quantifies the real deviation in seconds, facilitating direct interpretation in workshops; third, the Root Mean Square Error (RMSE) was selected to penalize deviations of great magnitude and detect poorly modelled subgroups; fourth, the Mean Absolute Percentage Error (MAPE) functions as an analytically comparable dimensionless indicator 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	Adjust model coefficients or rules	80% of the database of each product family	Initial model tuning and calibration metrics
Validation & Testing	Select hyperparameters and tree depth	-20% booked in strict isolation -Cross-validation k=5 applied internally	Mean and standard deviation of MAE, RMSE, and R^2
	Evaluate generalization on unseen data		Final Error Metrics (MAE, RMSE, MAPE, Test)R^2

3. RESULTS AND DISCUSSION

The results of the models' comparative performance, cross-validation, overfit diagnosis, and the industrial implications of the results are presented below.

3.1. Comparative Model Performance

The results show that both techniques (multiple linear regression and decision trees) present a high fit in the 3 product families analysed, however, it should be noted that the decision tree technique obtains higher values of R^2 in the three models evaluated. These results are consistent 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 hook) are not necessarily linear. In Table 5, the results for each of the models, techniques and metrics used are consolidated.

Table 5: *Consolidated performance metrics of models and techniques.*

Mod.	Technique	R^2 test	IT IS	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

Note: RLM = Multiple Linear Regression; AD = Decision tree.

The results presented in Table 5 show that both techniques show an adequate fit in the three families of products analyzed. However, the decision tree technique obtained better R^2 values and the errors were lower on unseen data, which is consistent with the nature of the process, considering that there is a non-linear relationship between the application time, the surface area and the units per hook. Table 5 shows the final metrics of the test set, which are most relevant for evaluating generalization capacity.

Table 5 shows that the decision tree technique yields a better performance in the three models evaluated, with higher test R^2 and lower values of MAE, RMSE and MAPE compared to multiple linear regression. In the case of Model 1, ASM 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 allows us to confirm that the decision tree better captures the changes in behaviour by surface ranges and load per hook, although multiple linear regression is still useful as an interpretable baseline and technical contrast model.

To complement the performance analysis and the validation of the fit of the models, 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 dm² indicates that the increase in surface area increases the application time, while the negative coefficient of the units per hook shows a reduction in unit time when the load processed per cycle increases. This reading is consistent with the physical logic of the process, i.e., parts with a larger surface area need more application travel, and a greater number of units per hook allows the operating time to be distributed among more parts.

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

Mod	Predictor	b	Standard Error	p-value	IC 95 %
1	Intercept	35.19	1.15	<0.001	[32.94; 37.44]
1	dm²	30.22	0.78	<0.001	[28.69; 31.75]
1	and./gan.	-2.51	0.12	<0.001	[-2.75; -2.27]
2	Intercept	42.85	2.10	<0.001	[38.73; 46.97]
2	dm²	25.14	1.05	<0.001	[23.08; 27.20]
2	and./gan.	-1.89	0.18	<0.001	[-2.24; -1.54]
3	Intercept	28.60	0.95	<0.001	[26.74; 30.46]
3	dm²	34.78	0.65	<0.001	[33.51; 36.05]
3	and./gan.	-3.15	0.09	<0.001	[-3.33; -2.97]

3.2. Cross-validation with existing standard times

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

Table 7: *Better external prediction vs. standard times defined by the organization.*

Mod	Ref.	Est. (s)	Technique	Before. (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

Note: Mod= Model; Est.= Standard time in seconds; AD= Decision tree; RLM= Multiple linear regression; Pred = Prediction in seconds; Error = Absolute percentage error.

The results of Table 7 show a high precision in parts that have similar geometries, highlighting the 456831 reference of Model 1, where the decision tree registers a marginal deviation of only 1.6% (80.49 s vs. 79.24 s). This example demonstrates the ability of analytical rules to represent engineering logic and the actual standard times estimated by the organization.

The results of Table 7 indicate that the decision tree obtained the lowest error in five of the seven references evaluated, while the multiple linear regression presented better performance in two references of Model 2. On average, the absolute percentage error of the multiple linear regression was 6.43%, compared to 4.77% for the decision tree.

The 456831 reference of Model 1 illustrates the best behaviour of the tree, with a deviation of 1.6% with respect to the business standard; however, the results of Model 2 indicate that linear regression can still be competitive when the process behaviour is closer to a linear relationship or when the variability of the family is lower. Consequently, the external validation confirms the convenience of using the decision tree as the main alternative, without ruling out multiple linear regression as a model of contrast and technical control.

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

In Figure 3, it can be seen how most of the references are grouped around the ideal agreement line, visually confirming the statistical results, where it was observed that the models have a controlled bias and very low errors. However, the rendering also exposes outliers corresponding to complex geometries with deep cavities, making it easier for method and timing analysts to immediately visually identify parts that will require additional field audits.

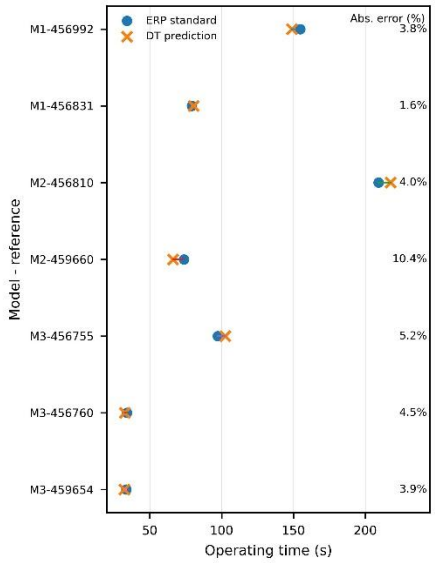

***Fig. 3**. Graphical comparison between real/standard times and times predicted by decision trees.*
***Source:** own elaboration*

3.3. Diagnosing Overfitting

An aspect of great importance to evaluate is the overfitting that the proposed decision tree models can have. To guarantee and validate that the proposed models do not have overfitting, the behavior of the mean square error (MSE) was evaluated according to the complexity of the algorithm. This evaluation was carried out by constricting learning curves, where the performance obtained in the training set and the test error for different levels of tree depth are contrasted. Figure 4 shows the results of the comparison made.

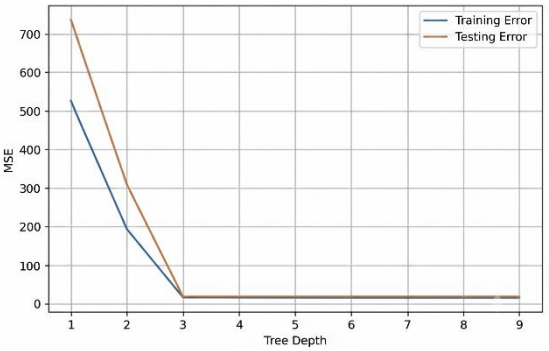

***Fig. 4**. Overfitting diagnosis by training error comparison and tree depth test.*
***Source:** own elaboration*

As can be seen in Figure 4, as it increases beyond the optimal level, the discrepancy between the two errors tends to stabilize. This demonstrates that the selected hyperparameters (fixed maximum depth and minimum number of samples per division) effectively control model variance without sacrificing the bias required for robust analytical prediction of the coating process.

3.4. Industrial implications of the results

From the perspective of industrial engineering, studies of traditional methods and times will continue to be fundamental for the estimation of standard production times, however, the added value of these models lies in supporting the preliminary estimation and updating of times when there is a large volume of historical data. At the same time, it should be noted that in industrial environments where time is very valuable, the proposed models allow reducing the time required to generate initial estimates compared to traditional techniques.

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

On the other hand, it is worth noting that the comparison with previous studies confirms that the prediction of manufacturing times using machine learning is relevant only when a large amount of historical data is available [1], [9], [24]. Additionally, the applicability of the proposed models depends on the maturity and stability of the process, the consistency of the records, and the periodic updating in the face of changes in product, technology and application method.

4. CONCLUSIONS

The research defines a reproducible and rigorous methodology for the estimation of preliminary standard times in organic coating processes for the plant under study, allowing estimating 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. In turn, the experimental model

incorporated a hold-out technical partition with cross-validation, guaranteeing the explicit evaluation of predictive performance and traceability in the statistical purification of operational variables on data not seen in the plant.

Regarding the multiple linear regression model, this model provided a high statistical interpretability, validating the significance of the selected key predictors (area and units per hook) by means of quantitative correlation coefficients. In relation to these predictors, square decimetres and units per hook demonstrated a consistent relationship with operating time, while 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 these did not allow to absorb the complexity of the geometric variability of the parts in the model, taking into account that the error of this model increased compared to the rest of the models.

As for the decision tree algorithm, it demonstrated a superior ability to represent nonlinear relationships and hierarchical rules. The results confirmed the superiority of this technique with respect to multiple linear regression, achieving better fits in the three (3) models proposed, being in agreement with the results of other research that highlight the decrease in error with the use of these techniques [25]. On the other hand, by obtaining a Mean Absolute Error (MAE) of only 3.65 seconds in Model 1, the control of overfitting (hyperparameters) made it possible for the model to capture understandable and precise rules for the nature of the analysed process. In turn, the validation against pre-existing standard times in the ERP system demonstrated the viability of the decision tree as a predictive tool for coating application times in new products.

Finally, the proposed mathematical model is an alternative that provides results with low errors, so it is consolidated as a predictive support tool to optimize capacity scheduling and planning and the coatings process [26]. However, it is again mentioned that the proposed model does not intend to replace direct observation with timekeeping. As a limitation, the performance of the proposed predictive model lies in the large volume of historical data available, as well as the reliability of its registration and consolidation in databases.

5. 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 suggested when making generalizations with the results of this research work. Similarly, the study is subject to changes in the work method, inclusion of technologies, product mix, operator experience and other exogenous variables that affect the production processes.

Future work should be aimed at expanding the amount of data available, incorporating new process variables not currently recorded [27], estimating prediction intervals and building error monitoring mechanisms. Additionally, it is recommended to implement an interface in which it is possible to link the production times that are recorded in the ERP 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.