Predictive Analysis Applied to Milk Cooling Using Regression Models and a Synthetic Dataset

LUIZ CARLOS BRANDÃO JUNIOR[1]
RICARDO RODRIGUES MAGALHÃES[1]

UFLA - Universidade Federal de Lavras
DCC - Departmento de Ciência da Computação
P.O. Box 3037 - Campus da UFLA 37200-000 - Lavras (MG)- Brazil
[1](luiz.junior11@estudante.ufla.br)
[2](ricardorm@ufla.br)

Abstract. Predictive analysis plays a crucial role in optimizing agro-industrial processes, such as milk cooling, which is essential for maintaining its quality. This study investigates the application of multiple regression models to predict critical variables in the milk cooling process, using a synthetic dataset with 10,000 samples. The dataset was structured to reflect key parameters like milk volume and initial temperature, inspired by information from a reference technical document on the numerical simulation of milk cooling [1]. Twenty regression algorithms were trained and evaluated to predict the actual cooling time, heat flux, and simulated cooling time. The results demonstrate that decision tree-based models (e.g., Gradient Boosting, LightGBM, Random Forest) achieved high accuracy ($R^2 > 0.99$) in predicting cooling times and good performance ($R^2 > 0.97$) for heat flux. This study highlights the utility of synthetic datasets for the development and evaluation of predictive models in a controlled environment, providing valuable insights for understanding and potentially optimizing the milk cooling process.

Keywords: Milk Cooling, Machine Learning, Regression Models, Synthetic Data, Predictive Analysis, Surrogate Models.

(Received July 7th, 2025 / Accepted December 19th, 2025)

1 Introduction

The efficient cooling of milk on the farm is a critical step in the dairy production chain, fundamental for preserving its microbiological and physicochemical quality, and for complying with sanitary regulations, such as Normative Instruction No. 62 in Brazil, which recommends cooling to temperatures of up to 7 °C (ideally 4 °C) within 3 hours after milking [1]. Modeling and numerical simulation, as demonstrated by Rezende et al. [1] using the finite element method (FEM) to analyze cooling tanks, offer powerful tools for understanding thermal dynamics and optimizing such systems. These methods, although accurate, can be computationally intensive, especially for extensive parametric analyses or for integration into real-time decision-making systems. In this scenario, predictive analysis based on machine learning (ML) emerges as a promising complementary approach, capable of developing fast and accurate predictive models from data representative of the system [5].

Acquiring large volumes of comprehensive experimental data to train ML models in complex agro-industrial processes like milk cooling faces significant challenges, including high costs, collection time, and the inherent difficulty in systematically covering a vast spectrum of operational conditions. Synthetic datasets, generated from validated simulation models or fundamental physical principles, such as those explored by Rezende et al. [1], constitute a strategic alternative [6]. They allow for the controlled and scalable creation of

datasets that mimic the behavior of real systems, facilitating exploratory research, development, and robust validation of predictive models before their practical implementation. Although synthetic data may not capture all the idiosyncrasies and stochastic noise of experimental data, their use is crucial for accelerating the development cycle and for testing hypotheses in a deterministic and well-characterized environment.

The present study investigates the application of predictive analysis, through a diverse set of regression models, to a detailed synthetic dataset. This dataset was designed to simulate the milk cooling process, incorporating key variables such as milk volume, initial temperature, and resulting in metrics like cooling times and heat fluxes, in line with the parameters and phenomena studied by Rezende et al. [1]. The central research problem is to determine how effectively these ML models can learn the complex relationships inherent to this thermo-fluid-dynamic process from synthetic data, and whether they can serve as efficient "surrogate models" for detailed numerical simulations.

The main objective of this work is, therefore, to evaluate the applicability and accuracy of multiple regression algorithms in predicting critical variables of the milk cooling process (actual cooling time, actual heat flux, and simulated cooling time), using the aforementioned synthetic dataset. The goal is not only to quantify the predictive performance but also to discuss the relevance and limitations of this approach, comparing the insights obtained with established knowledge and the simulation results presented in the study by Rezende et al. [1].

1.1 Theoretical Framework

Predictive analysis uses techniques from statistics, data mining, and machine learning to make predictions about unknown future events. In the context of milk cooling, predicting the time required to reach the target temperature (usually $4\,^{\circ}\mathrm{C}$) is crucial to inhibit microbial growth and preserve the physicochemical properties of the milk [1]. The reference document [1] emphasizes the importance of numerical simulation, such as the finite element method, to model the cooling process and understand the temperature distribution in the tank. This type of simulation generates data that, while not "real" in the direct experimental sense, is grounded in physical principles of heat transfer and fluid dynamics.

Regression models are a pillar of predictive analysis, seeking to establish a functional relationship between input variables (predictors) and a continuous output variable (target). In this study, various models were employed, from simple linear ones (Linear Regression, Ridge, Lasso) to more complex tree-based algorithms (Random Forest, Gradient Boosting, XGBoost, LightGBM), nearest neighbors (KNeighbors), support vector machines (SVR), and neural networks (MLP Regressor). Each of these models has different assumptions and capabilities for capturing linear and non-linear relationships in the data [4].

Synthetic datasets, like the one used in this study, are increasingly relevant. They allow data to be generated in scenarios where real data is scarce, expensive, or sensitive. For studying processes like milk cooling, a synthetic dataset can be constructed by incorporating the physical properties of milk (density, specific heat), tank dimensions, and heat transfer laws, as described in the document [1]. The structure of our synthetic dataset, with columns for volume, initial temperature, calculated characteristics (mass, milk height, heat to be removed), and outcome variables (cooling time, heat flux), was designed to mirror a realistic scenario and allow for the exploration of the relationships governing the process, in line with the parameters used in the numerical simulations of the reference document.

The main objective of this work is, therefore, to evaluate the applicability and accuracy of multiple regression algorithms in predicting critical variables of the milk cooling process (actual cooling time, actual heat flux, and simulated cooling time), using the aforementioned synthetic dataset. The goal is not only to quantify the predictive performance but also to discuss the relevance and limitations of this approach, comparing the insights obtained with established knowledge and the simulation results presented in the study by Rezende et al. [1].

The key contributions of this paper can be summarized as follows:

1. **A comprehensive comparative analysis:** We evaluate and compare the performance of twenty different regression algorithms, identifying tree-based ensemble models as the most effective for predicting milk cooling dynamics with high accuracy ($R^2 > 0.99$ for cooling time).

2. **Demonstration of synthetic data utility:** We highlight the strategic value of using a large-scale, physics-informed synthetic dataset to robustly train and benchmark machine learning models, bypassing the common challenges associated with real-world experimental data acquisition.

3. **Foundation for surrogate modeling:** We establish that the trained machine learning models can

serve as computationally efficient surrogate models for complex and time-consuming Finite Element Method (FEM) simulations, paving the way for applications in real-time process control and optimization.

1.2 Related Work

The literature on process modeling for thermal systems, including milk cooling, is historically dominated by physics-based numerical simulations. Techniques such as the Finite Element Method (FEM) and Computational Fluid Dynamics (CFD) are standard tools for analyzing heat transfer and fluid flow. The work of Rezende et al. [1] is a prime example in our specific domain, using FEM to simulate temperature distribution in a milk cooling tank. Similar approaches have been applied to other complex thermal systems, such as analyzing LPG tanks under fire [8] or modeling the behavior of pressure vessels [9]. While these simulation methods are highly accurate, their primary drawback is the significant computational cost, which limits their use for real-time decision-making or extensive parametric studies.

In parallel, there has been a growing trend of applying Machine Learning (ML) to various domains within agriculture and food science, including crop yield prediction, pest detection, and food quality assessment. The motivation is often to create data-driven models that can make fast and accurate predictions without explicitly solving the underlying physical equations [5]. This shift from explanatory, physics-based modeling to predictive, data-driven modeling represents a significant paradigm change.

Our work is positioned at the intersection of these fields, specifically exploring the use of ML models as computationally efficient *surrogate models* (or metamodels) for complex physical simulations. A surrogate model learns the input-output mapping of a simulation, providing near-instantaneous predictions once trained. This approach has gained traction in engineering and science to accelerate design optimization, uncertainty quantification, and sensitivity analysis, where thousands or millions of model evaluations are required. By training regression models to predict cooling time and heat flux, we aim to create a surrogate for the FEM simulation described by Rezende et al. [1].

A key challenge in developing such data-driven models is the availability of large, high-quality datasets. Experimental data collection is often expensive, time-consuming, and may not cover the full range of operational parameters. To overcome this, our study leverages a *synthetic dataset*. As outlined by Jordon et al. [6], synthetic data generation offers a powerful strategy to create large, perfectly-labeled datasets for training ML models in a controlled manner. Unlike studies that might rely on limited experimental data, our approach uses a dataset generated from established physical principles, allowing us to robustly evaluate the intrinsic ability of different ML algorithms to learn the complex, non-linear dynamics of the milk cooling process. This methodology enables a rigorous comparative analysis that would be difficult to achieve with real-world data alone.

The remainder of this paper is organized as follows. Section 2 details the methodology, describing the synthetic dataset, the suite of regression algorithms, and the performance evaluation metrics used. Section 3 presents and discusses the empirical results, comparing the predictive accuracy of the models across the different target variables. Finally, Section 4 concludes the paper, summarizing the main findings and outlining avenues for future work.

2 Methods

This study employed a quantitative approach based on the analysis of a synthetic dataset to evaluate predictive regression models.

2.1 Synthetic Dataset and Variable Definition

A synthetic dataset containing 10,000 samples was used, each representing a milk cooling cycle. The structure of the dataset includes the following main columns:

- `ID_Amostra`: Unique sample identifier.
- `Volume_Leite_L`: Milk volume in liters.
- `Temperatura_Inicial_C`: Initial milk temperature in Celsius.
- `Altura_Leite_m`: Milk height in the tank, calculated from the volume.
- `Massa_Leite_kg`: Mass of the milk, calculated from volume and density.
- `Delta_T_Resfriamento_C`: Difference between initial and final temperature.
- `Calor_Total_Removido_J`: Total amount of heat to be removed.
- `Tempo_Resfriamento_Real_min`: Actual cooling time in minutes.
- `Fluxo_Calor_Real_Wm2`: Actual heat flux in W/m^2.

- `Erro_Percentual_Simulacao_pct`: Percentage error of the simulation.
- `Tempo_Resfriamento_Simulado_min`: Simulated cooling time in minutes.
- `Temperatura_Final_C`: Final milk temperature (constant at $4\,^{\circ}\mathrm{C}$).

The primary input variables (predictors), defined for all regression models, were `Volume_Leite_L` and `Temperatura_Inicial_C`. The study focused on predicting multiple output variables of interest; however, each regression model was trained and evaluated independently to predict a single output variable at a time (*single-output* approach). The output variables (targets) analyzed individually in this study were: `Tempo_Resfriamento_Real_min`, `Fluxo_Calor_Real_Wm2`, and `Tempo_Resfriamento_Simulado_min`. Thus, for each of these three target variables, a complete set of twenty regression models was fitted and evaluated.

2.2 Regression Models and Evaluation Strategy

The selection of regression algorithms for this study aimed to cover a diverse spectrum of approaches, from classic linear models to complex ensembles and neural networks, allowing for a robust comparative evaluation of their ability to model the relationships present in the synthetic dataset. Twenty distinct algorithms were implemented using the Scikit-learn library [10], along with popular gradient boosting implementations like XGBoost [11] and LightGBM [12].

The set of models included:

- **Linear Models:** Simple Linear Regression, Ridge Regression (L2 regularization), Lasso Regression (L1 regularization, promoting sparsity), ElasticNet (combination of L1 and L2) [7]. These models assume a linear relationship between predictors and the target variable, with different approaches to handle multicollinearity and feature selection. Additionally, Bayesian Ridge, which introduces a Bayesian approach to linear regression, Huber Regressor, robust to outliers, Passive Aggressive Regressor, and SGD Regressor (Stochastic Gradient Descent), suitable for large datasets and online learning, were included. The TheilSen Regressor, another robust estimator, was also considered.
- **Instance-Based Models:** KNeighbors Regressor (KNN), which predicts the target variable's value based on the average of its nearest neighbors in the feature space [13].
- **Decision Tree-Based Models and Ensembles:** Decision Tree Regressor, which recursively partitions the feature space; and its ensemble-based enhancements like Random Forest Regressor (bagging of decision trees) [14], Extra Trees Regressor (Extremely Randomized Trees). Boosting algorithms were also included, such as AdaBoost Regressor (Adaptive Boosting) [15], Gradient Boosting Regressor, XGBoost (Extreme Gradient Boosting), and LightGBM (Light Gradient Boosting Machine), known for their high performance on tabular data by sequentially building trees that correct the errors of previous ones [16].
- **Support Vector Machines (SVM):** SVR (Support Vector Regression) with a linear kernel and with an RBF (Radial Basis Function) kernel, which seek to find a hyperplane that best fits the data, with a margin of tolerance for errors [17].
- **Artificial Neural Networks:** MLP Regressor (Multi-layer Perceptron), a simple feedforward neural network capable of learning complex nonlinear relationships [18].

For each combination of output variable and regression model, the following training and evaluation methodology was applied:

1. **Data Splitting:** The dataset was partitioned into training (80% of samples) and testing (20% of samples) subsets using a random split, ensuring reproducibility through a random seed (‘random_state‘).
2. **Feature Preprocessing:** The input variables were standardized (scaled) using Scikit-learn's ‘StandardScaler‘. This process transforms the data to have a zero mean and unit standard deviation, which is crucial for the optimal performance of many algorithms, such as SVMs, regularized linear models, KNN, and neural networks [7]. The scaler was fitted only on the training data and then applied (transform) to both the training and test data to prevent data leakage. This step was integrated into a Scikit-learn ‘Pipeline‘ to simplify the workflow and ensure correct application.
3. **Training and Prediction:** Each model, encapsulated in the pipeline with the scaler, was trained using the training set (‘X_train‘, ‘y_train‘). Subsequently, the trained models were used to make predictions on the test set (‘X_test‘).

4. **Performance Evaluation Metrics:** The performance of the regression models was quantified using a comprehensive set of metrics, calculated by comparing the predicted values (‘y_pred‘) with the actual values from the test set (‘y_test‘):

    - Coefficient of Determination (R^2): Indicates the proportion of the variance in the dependent variable that is explained by the independent variables. Values closer to 1 indicate a better fit [19].
    - Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values.
    - Mean Squared Error (MSE): The average of the squared differences between predicted and actual values, penalizing larger errors.
    - Root Mean Squared Error (RMSE): The square root of the MSE, expressed in the same unit as the target variable.
    - Median Absolute Error (MedAE): Robust to outliers, represents the median of the absolute errors.
    - Explained Variance Score (EVS): Measures the proportion to which a mathematical model accounts for the dispersion of a given dataset. If the errors have a zero mean, it is equal to R^2.

The results of all metrics were compiled into tables for each output variable. Additionally, the R^2 scores were visualized in horizontal bar charts, sorted in descending order, to facilitate a visual comparison of the predictive power of the different models for each target.

3 Results

The application of regression models to the synthetic dataset revealed significant insights into the predictability of the output variables. Detailed results (metric tables and R^2 plots) were generated and analyzed for each target.

3.1 Prediction of `Tempo_Resfriamento_Real_min`

As visualized in Figure 1, the tree-based ensemble models (Gradient Boosting Regressor, LightGBM Regressor, Random Forest Regressor, XGBoost Regressor) and the KNeighbors Regressor, as well as the MLP Regressor, demonstrated excellent predictive capability, with R^2 scores consistently above 0.99, as seen in Table 1. For example, the Gradient Boosting Regressor achieved an R^2 of 0.995865. This indicates that more than 99.5% of the variability in the actual cooling time can be explained by the milk volume and its initial temperature. Linear models, although showing reasonable R^2 (e.g., Linear Regression with R^2 of 0.863), were significantly outperformed, suggesting predominant non-linear relationships. These high R^2 scores align with the expectation that cooling time is strongly determined by these physical parameters, as would be inferred from a physical simulation model like the one described in [1]. The reference document, by using numerical simulation, also assumes a strong dependence on these parameters.

3.2 Prediction of `Fluxo_Calor_Real_Wm2`

For the prediction of heat flux (Figure 2), the tree-based ensemble models again stood out, with Gradient Boosting (R^2 = 0.973), LightGBM (R^2 = 0.973), and XGBoost (R^2 = 0.971) leading, as also shown in Table 2. However, there was a sharp drop in performance for linear models, with R^2 scores close to 0.04, and negative R^2 for others (e.g., TheilSen Regressor with R^2 = -0.508). This suggests that the relationship between the inputs and heat flux is considerably more complex and non-linear than for cooling time. Heat flux, being a derived rate (energy per unit time and area), is inherently more sensitive to variations and interactions. The ability of tree-based models to capture these complex interactions justifies their better performance. The document [1] does not detail the direct predictability of heat flux, but heat transfer principles would indicate a complex relationship.

3.3 Prediction of `Tempo_Resfriamento_Simulado_min`

The results for predicting the simulated cooling time (Figure 3) were very similar to those for the actual cooling time, with the same tree-based ensemble models and MLP Regressor achieving $R^2 > 0.99$, as can also be seen in Table 3. This was expected, as the simulated time was generated in the synthetic dataset from the actual time with the addition of a small percentage error. The high correlation between the two times ensures that models capable of predicting one will also predict the other with similar effectiveness.

3.4 Further Discussion and Implications

The use of the synthetic dataset in this study transcended the mere generation of data volume; it proved to be a valuable tool for the controlled and comparative evaluation of the intrinsic capabilities of various regression algorithms. By mimicking a complex physical process like milk cooling, based on the principles

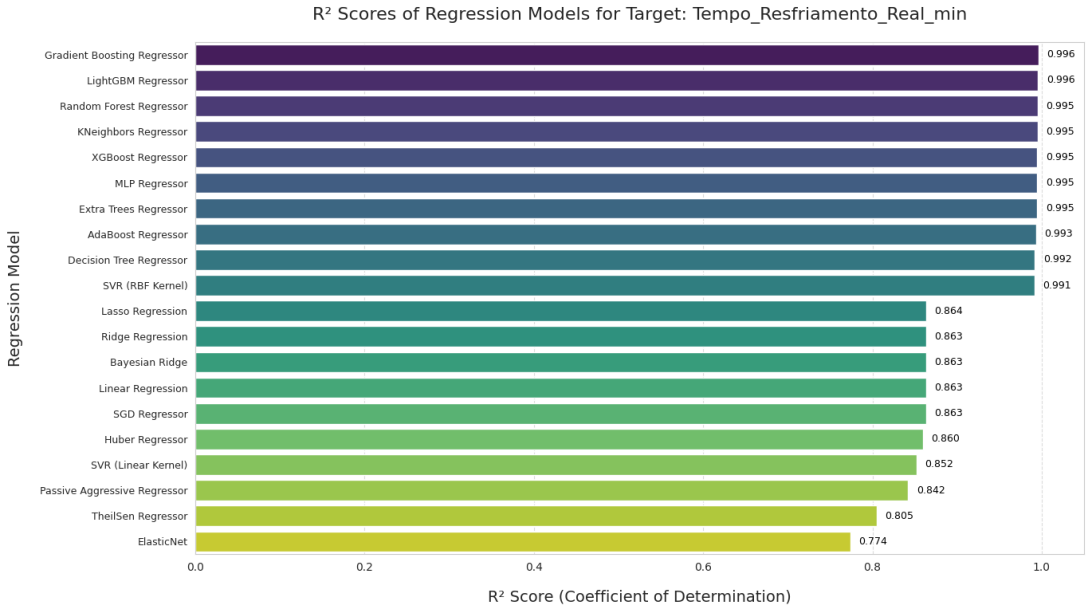

Figure 1: Comparison of R^2 scores in descending order for 'Tempo_Resfriamento_Real_min'

Table 1: Regression Model Results for 'Tempo_Resfriamento_Real_min' (Sorted by R^2 Score)

Model	R^2 Score	MAE	MSE	RMSE	MedAE	EVS
Gradient Boosting Regressor	0.9959	3.006	17.539	4.188	2.131	0.9959
LightGBM Regressor	0.9956	3.063	18.705	4.325	2.155	0.9956
Random Forest Regressor	0.9951	3.255	20.854	4.567	2.313	0.9951
KNeighbors Regressor	0.9950	3.282	21.005	4.583	2.307	0.9951
XGBoost Regressor	0.9949	3.229	21.763	4.665	2.271	0.9949
MLP Regressor	0.9947	3.420	22.420	4.735	2.458	0.9947
Extra Trees Regressor	0.9947	3.373	22.611	4.755	2.376	0.9947
AdaBoost Regressor	0.9931	4.269	29.380	5.420	3.669	0.9932
Decision Tree Regressor	0.9920	4.122	34.101	5.840	2.880	0.9920
SVR (RBF Kernel)	0.9912	3.735	37.208	6.100	2.304	0.9913
Lasso Regression	0.8635	20.351	578.816	24.059	19.851	0.8636
Ridge Regression	0.8633	20.411	579.617	24.075	20.008	0.8634
Bayesian Ridge	0.8633	20.412	579.622	24.075	20.016	0.8634
Linear Regression	0.8633	20.412	579.625	24.075	20.019	0.8634
SGD Regressor	0.8632	20.500	580.272	24.089	20.222	0.8632
Huber Regressor	0.8597	19.962	594.859	24.390	17.931	0.8633
SVR (Linear Kernel)	0.8520	19.862	627.665	25.053	16.298	0.8621
Passive Aggressive Regressor	0.8422	20.119	669.453	25.874	15.171	0.8635
TheilSen Regressor	0.8047	21.095	828.369	28.781	12.164	0.8627
ElasticNet	0.7740	24.019	958.726	30.963	20.325	0.7740

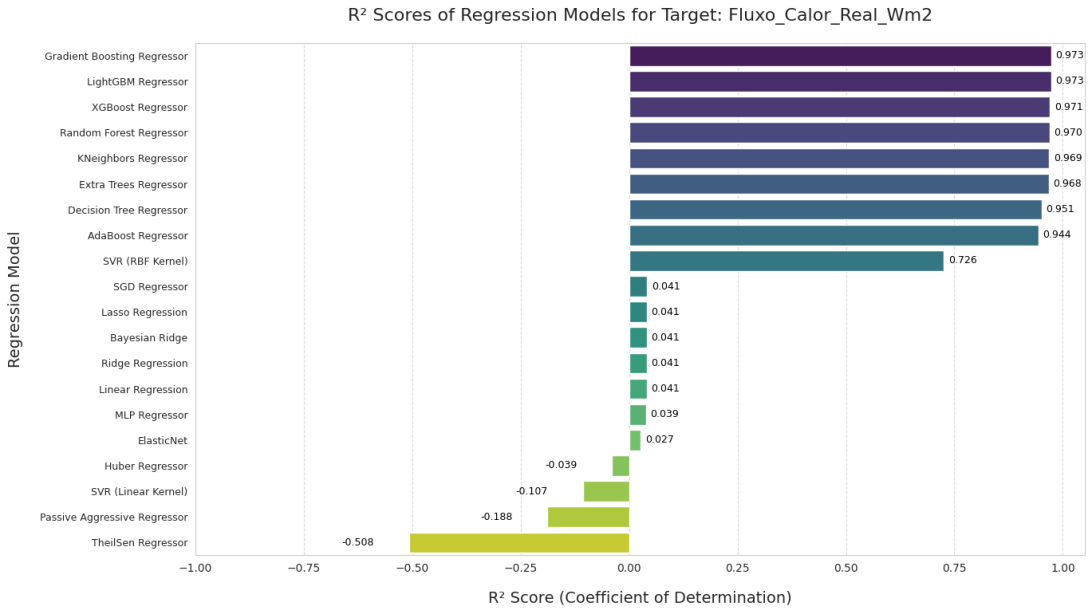

Figure 2: Comparison of R^2 scores in descending order for heat flux prediction (Fluxo_Calor_Real_Wm2)

Table 2: Regression model performance for heat flux prediction (Fluxo_Calor_Real_Wm2) sorted by R^2 score

Model	R^2	MAE	MSE	RMSE	MedAE	EVS
Gradient Boosting	0.9732	85.39	11,996.87	109.53	70.41	0.9732
LightGBM	0.9730	86.29	12,087.87	109.95	72.53	0.9730
XGBoost	0.9706	90.19	13,192.19	114.86	73.86	0.9706
Random Forest	0.9699	91.51	13,501.28	116.20	76.18	0.9699
KNeighbors	0.9687	92.93	14,007.25	118.35	77.83	0.9688
Extra Trees	0.9679	94.23	14,365.44	119.86	79.56	0.9679
Decision Tree	0.9510	116.96	21,945.69	148.14	99.57	0.9510
AdaBoost	0.9437	127.14	25,228.65	158.84	109.81	0.9468
SVR (RBF)	0.7260	246.20	122,797.27	350.42	163.72	0.7318
Linear Models						
SGD Regressor	0.0413	551.14	429,638.28	655.47	537.91	0.0413
Lasso	0.0409	550.25	429,845.63	655.63	534.42	0.0412
Bayesian Ridge	0.0408	550.13	429,856.28	655.63	534.04	0.0411
Ridge	0.0408	550.09	429,860.67	655.64	533.67	0.0411
Linear	0.0408	550.09	429,861.05	655.64	533.64	0.0411
MLP	0.0390	549.82	430,663.02	656.25	533.14	0.0396
ElasticNet	0.0269	557.66	436,116.44	660.39	547.49	0.0272
Huber	-0.0394	535.16	465,837.39	682.52	461.03	-0.0179
SVR (Linear)	-0.1070	534.99	496,107.47	704.35	438.37	-0.0691
Passive Aggressive	-0.1884	534.90	532,610.32	729.80	408.05	-0.1509
TheilSen	-0.5076	568.56	675,634.57	821.97	365.33	-0.1569

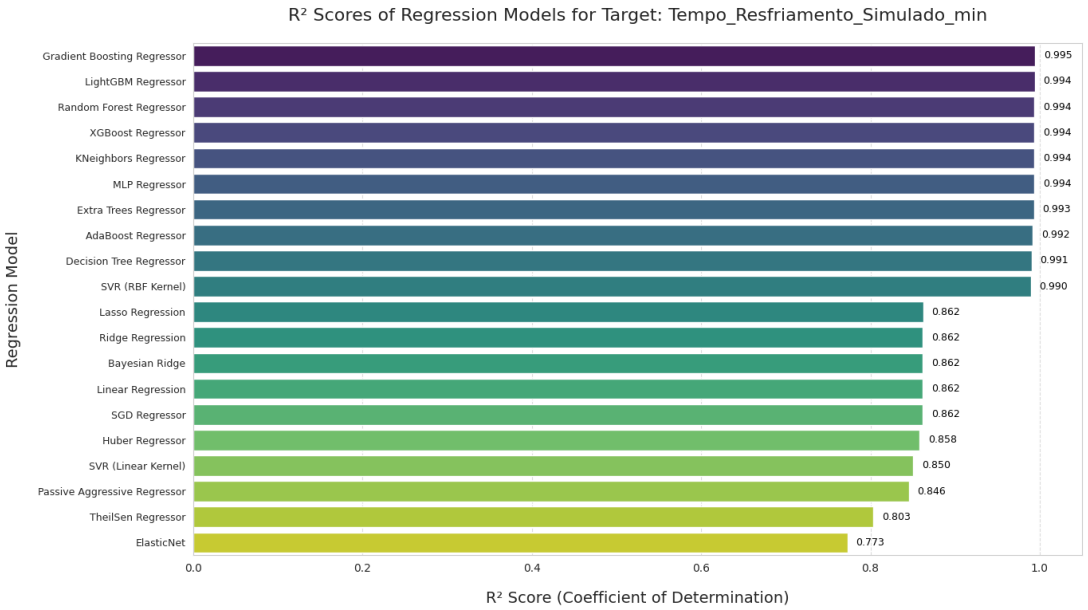

Figure 3: Performance comparison (R^2 scores) for simulated cooling time prediction (Tempo_Resfriamento_Simulado_min)

Table 3: Regression results for simulated cooling time prediction (sorted by R^2 score)

Model	R^2	MAE	MSE	RMSE	MedAE	EVS
Gradient Boosting	0.9948	3.37	22.23	4.72	2.42	0.9948
LightGBM	0.9944	3.44	23.89	4.89	2.41	0.9944
Random Forest	0.9939	3.63	26.02	5.10	2.62	0.9939
XGBoost	0.9938	3.58	26.29	5.13	2.54	0.9938
KNeighbors	0.9938	3.67	26.38	5.14	2.68	0.9938
MLP	0.9937	3.74	26.99	5.20	2.70	0.9937
Extra Trees	0.9934	3.74	27.97	5.29	2.63	0.9934
AdaBoost	0.9919	4.70	34.29	5.86	4.20	0.9920
Decision Tree	0.9906	4.54	40.04	6.33	3.15	0.9906
SVR (RBF)	0.9895	4.14	44.65	6.68	2.56	0.9896
Linear Models						
Lasso	0.8622	20.41	585.99	24.21	20.13	0.8623
Ridge	0.8620	20.48	586.84	24.22	20.22	0.8621
Bayesian Ridge	0.8620	20.48	586.84	24.22	20.22	0.8621
Linear	0.8620	20.48	586.85	24.22	20.22	0.8621
SGD	0.8619	20.56	587.40	24.24	20.44	0.8619
Huber	0.8583	20.01	602.59	24.55	17.83	0.8620
SVR (Linear)	0.8504	19.91	636.32	25.23	16.39	0.8605
Passive Aggressive	0.8456	20.22	656.71	25.63	15.93	0.8613
TheilSen	0.8034	21.15	835.83	28.91	12.29	0.8612
ElasticNet	0.7727	24.03	966.43	31.09	20.07	0.7727

explored by Rezende et al. [1], the synthetic dataset allowed for the isolation of the algorithms' learning capacity from the uncertainties and noise inherent in experimental data. This "digital laboratory" facilitated the systematic investigation of how different model architectures (linear, tree-based, SVMs, neural networks) capture functional relationships of varying complexities, a fundamental step before application in scenarios with real data, which is often scarcer or noisier [6]. Additionally, ML models trained on rich and representative synthetic data have the potential to act as *surrogate models* for more computationally expensive FEM simulations, such as those performed by Rezende et al. [1], enabling rapid parametric exploration and quasi-real-time process optimization.

The high accuracy achieved in predicting `Tempo_Resfriamento_Real_min` and `Tempo_Resfriamento_Simulado_min` (R^2 > 0.99 for the best models) using only milk volume and initial temperature as predictors is of significant practical importance. This predictive capability opens avenues for the development of optimized control systems in cooling tanks, which could dynamically adjust operational parameters to minimize energy consumption while ensuring the milk reaches the target temperature within the regulatory deadline [1]. Additionally, such models can support more efficient logistics planning for milk collection and serve as a basis for early warning systems in case of deviations in the refrigeration system's performance. The fact that tree-based ensemble models (e.g., Gradient Boosting, LightGBM) and the MLP Regressor consistently outperformed linear models reinforces the observation by Rezende et al. [1] about the non-linear and complex nature of the thermal and fluid interactions in the tank.

For the prediction of `Fluxo_Calor_Real_Wm2`, although the best models (R^2 > 0.97) still demonstrate strong predictive ability, the greater dispersion in R^2 scores among the different algorithms and the sharp decline in the performance of linear models (R^2 close to 0.04) are noteworthy. Heat flux, being a derived quantity that intrinsically depends on the rate of temperature change and thus on the cooling time and the total amount of heat to be removed, is inherently more sensitive to non-linear interactions between volume, initial temperature, and cooling time itself (which is also a prediction). This additional complexity, where non-linearities compound, justifies the superiority of tree-based ensemble models, which are capable of modeling these higher-order interactions more effectively than linear models. The exploration of more sophisticated feature engineering, such as explicit interaction terms between input variables or physics-informed features (e.g., a non-dimensional Nusselt number if more system parameters were known), could be investigated to further improve the robustness of the heat flux prediction.

It is crucial to acknowledge the inherent limitations of using synthetic data. The dataset used here, although extensive and well-founded, is an idealized representation. Real-world factors such as variations in milk composition (fat content, solids), the variable efficiency of the agitation system, fouling on heat exchange surfaces, fluctuations in ambient temperature or power grid voltage, and the natural wear and tear of refrigeration equipment were not explicitly modeled. Therefore, the high R^2 scores obtained represent a performance "ceiling" under controlled conditions. The transition of these models to real-world applications (the *sim-to-real gap* challenge) would most likely require a calibration step with actual experimental data and, possibly, the application of domain adaptation or transfer learning techniques [20].

Directions for future work include optimizing the hyperparameters of the most promising models for each target, using techniques such as GridSearchCV or Bayesian Optimization, to further refine their performance. The application of ML model interpretability techniques (e.g., SHAP [21], LIME [22]) could provide deeper insights into how input variables influence predictions, complementing the physical understanding of the system. Another promising avenue would be to enrich the synthetic dataset with variations in other operational parameters (e.g., agitator speed, refrigerant temperature) or with the introduction of different tank geometries, approaching the complexity addressed by advanced FEM simulations even more closely. Finally, integrating these predictive models with real-time sensor data could pave the way for the development of "digital twins" of the milk cooling process, enabling continuous monitoring, diagnosis, and optimization.

4 Conclusion

This study aimed to evaluate the applicability and accuracy of a diverse set of twenty regression models for predicting critical variables in the milk cooling process. To this end, a synthetic dataset with 10,000 samples was used, whose generation was informed by physical principles and the parameters outlined in the numerical simulation study by Rezende et al. [1]. The results obtained offer a clear view of the potential and limitations of predictive analysis based on machine learning in this specific domain.

The main findings confirm the hypothesis that fun-

damental thermodynamic variables, such as cooling time (actual and simulated), are highly predictable (R^2 > 0.99 for the best models, like Gradient Boosting and LightGBM) when using primary predictors like milk volume and its initial temperature. This finding not only validates the internal consistency of the synthetic dataset but also underscores the strong causal determination of these parameters in the physical process, a premise implicit in FEM simulations [1]. The ability of tree-based ensemble models and neural networks to capture the inherent non-linearities of these processes was consistently superior to that of linear models. Additionally, the actual heat flux also proved to be predictable with good performance (R^2 > 0.97), although its derived nature and greater complexity required more robust models to capture its nuances.

In summary, this work demonstrates that predictive analysis, when applied to well-founded synthetic datasets, constitutes a powerful and efficient tool for the development, testing, and comparison of machine learning models in a controlled environment. It complements traditional approaches based on numerical simulation, offering the potential for creating fast and accurate surrogate models, with direct implications for process optimization, energy efficiency, and quality control in the dairy industry. The results reinforce the strong dependence of cooling dynamics on input physical parameters and highlight the remarkable ability of certain ML algorithms to learn and generalize these complex relationships.

Future research should focus on validating these predictive models with real experimental data to quantify the gap between simulation/synthetic and reality (the *sim-to-real gap*) and explore domain adaptation techniques. Expanding the synthetic dataset to include a wider variety of operational parameters (e.g., agitation rate, refrigerant type, tank geometry) and phenomena (e.g., ice formation, variations in milk composition) could lead to even more comprehensive models. Additionally, hyperparameter optimization and the application of model interpretability techniques (XAI) are logical steps to refine performance and increase the confidence in and understanding of the developed models, bringing them closer to practical application in decision support systems in the industry.

References

[1] REZENDE, R. P.; ANDRADE, E. T. de; CORREA, J. L. G.; MAGALHÃES, R. R. Numerical simulation applied to milk cooling. *Revista Engenharia na Agricultura*, Viçosa, MG, v. 29, p. 122-128, Jul. 2021. ISSN 2175-6813. DOI: 10.13083/reveng.v29i1.9527. Available at: `https://www.reveng.ufv.br/reveng/article/view/9527`. Accessed on: [Access Date].

[2] Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, 25(3), 289-310.

[3] Jordon, J., Yoon, J., & van der Schaar, M. (2022). Synthetic data: A conceptual and practical guide. *Patterns*, 3(6), 100516.

[4] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

[5] SHMUELI, G. (2010). To explain or to predict?. *Statistical Science*, 25(3), p. 289-310.

[6] JORDON, J.; YOON, J.; VAN DER SCHAAR, M. (2022). Synthetic data: A conceptual and practical guide. *Patterns*, 3(6), 100516.

[7] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Science & Business Media. (Springer Series in Statistics).

[8] ZHAO, B. (2014). Temperature-coupled field analysis of LPG tank under fire based on wavelet finite element method. *Journal of Thermal Analysis and Calorimetry*, 117(1), p. 413-422.

[9] NIMDUM, P.; PATAMAPROHM, B.; RENARD, J.; VILLALONGA, S. (2015). Experimental method and numerical simulation demonstrate non-linear axial behaviour in composite filament wound pressure vessel due to thermal expansion effect. *International Journal of Hydrogen Energy*, 40(39), p. 13231-13241.

[10] PEDREGOSA, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, p. 2825-2830.

[11] CHEN, T.; GUESTRIN, C. (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, p. 785-794.

[12] KE, G. et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems*

30 (NIPS 2017), Long Beach, CA, USA, p. 3146-3154.

[13] ALTMAN, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), p. 175-185.

[14] BREIMAN, L. (2001). Random forests. *Machine Learning*, 45(1), p. 5-32.

[15] FREUND, Y.; SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), p. 119-139.

[16] FRIEDMAN, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), p. 1189-1232.

[17] DRUCKER, H. et al. (1997). Support vector regression machines. In: *Advances in Neural Information Processing Systems 9 (NIPS 1996)*, Denver, CO, USA, p. 155-161.

[18] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

[19] MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. (2021). *Introduction to Linear Regression Analysis*. 6th ed. Hoboken, NJ: John Wiley & Sons.

[20] PAN, S. J.; YANG, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), p. 1345-1359.

[21] LUNDBERG, S. M.; LEE, S-I. (2017). A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, p. 4765-4774.

[22] RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, p. 1135-1144.