mythos

Grupo UNIS

https://www.periodicos.unis.edu.br/mythos/

# MACHINE LEARNING–DRIVEN PROCESS OPTIMIZATION: A CASE STUDY IN A POULTRY FARM IN SOUTHERN MINAS GERAIS

## *OTIMIZAÇÃO DE PROCESSOS ORIENTADO POR MACHINE LEARNING: estudo de caso em uma granja sul mineira*

Leonardo Romanelli Guimarães[1*], Rodrigo Franklin Frogeri[2], Ana Amélia Furtado de Oliveira[3]

[1] *B.S. in Information Systems (UNIS-MG). University Center of Southern Minas - UNISMG, Varginha, Brazil. leonardo.guimaraes@alunos.unis.edu.br*

[2] *PhD in Information Systems and Knowledge Management (FUMEC). University Center of Southern Minas - UNISMG, Varginha, Brazil. rodrigo.frogeri@professor.unis.edu.br*

[3] *PhD in Linguistic Studies (UNESP). University Center of Southern Minas - UNISMG, Varginha, Brazil. ana.furtado@professor.unis.edu.br*

**Editorial Details**
Double-blind review system
Memory of a Scientific Event
XI Simpósio Mineiro de Gestão, Educação, Comunicação e Tecnologia da Informação (2025).

**Article History:**
**Received:** December 18, 2025
**Revised:** January 15, 2026
**Accepted:** February 03, 2026
**Published online:** February 04, 2026

**Editor-in-Chief:**
Rodrigo Franklin Frogeri
**Guest Editors:**
Pedro dos Santos Portugal Júnior
Fabrício Pelloso Piurcosky
**Technical Editor:**
Eufrásia de Souza Melo

**Funding:**
This study received no external funding.

**How to cite this article:**
Guimarães, L. R.; Frogeri, R. F.; Oliveira, A. A. F. (2026). Machine Learning–Driven Process Optimization: A Case Study in a Poultry Farm in Southern Minas Gerais. **Mythos**, v. 17, 2, 373-387. https://doi.org/10.36674/mythos.v17i2.1049

***Corresponding Author:**
Leonardo Romanelli Guimarães
*leonardo.guimaraes@alunos.unis.edu.br*

## Abstract

This paper presents a case study on the application of Machine Learning (ML) in the optimization of egg production, conducted in the context of a poultry farm located in southern Minas Gerais, Brazil. The purpose of this research is to analyze the feasibility of optimizing the farm's production processes using predictive ML techniques. To achieve this goal, a quantitative methodology was adopted, based on three main stages: collection and preprocessing of the farm's historical data, development of predictive models using the WEKA software, and evaluation of the algorithms' performance through statistical metrics such as Mean Absolute Error (MAE) and the correlation coefficient (R). Among the tested models, Random Forest showed the best performance, presenting a high correlation level and the lowest error margin, thus proving to be suitable for operational forecasting. The M5P model stood out for its balance between accuracy and interpretability, while linear regression, although simpler, delivered satisfactory and easily interpretable results. It is believed that the findings of this study may contribute to discussions and reflections on the importance of technology in the efficient management of poultry farming and its impact on small and medium-sized farms in the region, optimizing resource use and increasing sustainability.

**Keywords:** Artificial Intelligence, Egg Production, Poultry Farming, Random Forest, Regression Tree M5P.



ISSN: 1994-0098 | e-ISSN: 3086-0210

## Resumo

O presente trabalho apresenta um estudo de caso sobre a aplicação de Machine Learning (ML) na otimização da produção de ovos, vivenciado no contexto de uma granja sul mineira. Por meio desta pesquisa, busca-se analisar a viabilidade de otimizar os processos produtivos da granja utilizando técnicas preditivas de ML. Para isso, adotou-se uma metodologia quantitativa baseada em três etapas principais: coleta e tratamento de dados históricos da granja, construção de modelos preditivos utilizando o software WEKA, e avaliação do desempenho dos algoritmos aplicados por meio de métricas estatísticas como erro médio absoluto (MAE) e coeficiente de correlação (R). Entre os modelos testados, o Random Forest apresentou o melhor desempenho, com elevado grau de correlação e menor margem de erro, mostrando-se ideal para previsões operacionais. O M5P destacou-se pelo equilíbrio entre precisão e interpretabilidade, enquanto a regressão linear, embora mais simples, ofereceu resultados satisfatórios e de fácil compreensão. Acredita-se que os resultados deste relato possam contribuir para o debate e reflexões sobre a importância da tecnologia na gestão eficiente da avicultura e sua influência em pequenas e médias granjas da região, otimizando o uso de recursos e aumentando a sustentabilidade.

***Palavras-chave**: Inteligência Artificial, Produção de Ovos, Avicultura, Floresta Aleatória, Árvore de Regressão M5P.*

**Declaration on the Use of Artificial Intelligence to prepare the manuscript**

The authors declare that artificial intelligence (AI) tools were used during the preparation of the manuscript submitted to Revista Mythos. The platform used was ChatGPT 5 (free version), developed by OpenAI, and it was employed specifically for text editing, language enhancement, and translation between Portuguese and English.

All authors take full responsibility for the accuracy and integrity of the content generated or enhanced with the assistance of AI.

**Data available**

The dataset supporting the findings of this study is publicly available in the Zenodo repository. It is provided as Dataset – Machine Learning–Driven Process Optimization: A Case Study in a Poultry Farm in Southern Minas Gerais. The dataset can be accessed via the following persistent digital object identifier (DOI): https://doi.org/10.5281/zenodo.18487353

**Citation:**
Guimarães, L. R., Frogeri, R. F., & Oliveira, A. A. F. de . (2026). Dataset - Machine Learning–Driven Process Optimization: A Case Study in a Poultry Farm in Southern Minas Gerais [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18487353



ISSN: 1994-0098 | e-ISSN: 3086-0210

# 1 INTRODUCTION

In recent years, Brazilian agribusiness has undergone significant transformation, driven by advances in digital technologies and the growing demand for efficiency and sustainability in production (Ang & Seng, 2021). The country, recognized for its leadership in several production chains, such as soybeans, corn, and poultry, has increasingly viewed advanced technologies, including Machine Learning (ML), as indispensable solutions for addressing challenges related to productivity maximization and operational cost reduction (Ang & Seng, 2021). The application of ML enables process automation and the use of real-time data to support strategic decision-making (Meshram et al., 2021).

Studies show that the use of ML algorithms can improve the efficiency of agricultural operations by reducing resource waste and enhancing the sustainability of production processes (Meshram et al., 2021; Sharma et al., 2021). This is particularly relevant in the context of egg production, where variables such as nutrition, bird health, temperature, and humidity must be precisely controlled to ensure productivity (Ang & Seng, 2021). Through automated analysis of these factors, ML-based systems provide accurate forecasts that allow management adjustments in advance, preventing unexpected losses (Sharma et al., 2021).

With the adoption of technological solutions, poultry farms that employ ML have achieved superior results in terms of both operational efficiency and environmental sustainability (Ang & Seng, 2021). ML models enable continuous and automated monitoring of housing conditions, identifying production anomalies and suggesting immediate adjustments in feeding or ventilation as needed (Meshram et al., 2021; Sharma et al., 2021). In addition, predictive algorithms help identify disease patterns in poultry before they become critical issues, ensuring more efficient management and preserving flock health.

Another significant advantage of applying ML to egg production is the ability to forecast demand and plan production more accurately (Ang & Seng, 2021). This helps prevent both shortages and overproduction, contributing to more sustainable and cost-effective management (Meshram et al., 2021; Sharma et al., 2021). Furthermore, intelligent systems assist in optimizing the use of inputs throughout the production cycle, such as feed and water, by adjusting supply according to the birds' needs (Meshram et al., 2021; Sharma et al., 2021).

In this context, the present study aims to analyze the feasibility of forecasting egg production in a poultry farm located in southern Minas Gerais using Machine Learning techniques. It is believed that this analysis may contribute to local producers and poultry industry professionals, as production optimization can result in significant gains in efficiency and profitability, benefiting the entire regional production chain (Meshram et al., 2021; Sharma et al., 2021).

# 2 THEORETICAL BACKGROUND

## *2.1 Precision Poultry Farming and Evolution of Predictive Architectures in Avian Science*

Precision Poultry Farming (PPF) represents the application of engineering principles to the management of living birds, leveraging sensors, big data, and artificial intelligence for real-time intervention (Fan et al., 2025). While the conceptual framework of PPF is well-established, the practical implementation often varies based on the sophistication of the available data (Bumanis, 2024). Machine learning models act as the analytical engine of PPF, learning patterns from historical data to make accurate classifications or predictions on new observations (Bumanis, 2024).

In the specific domain of egg production, the literature identifies a hierarchy of influential variables. Factors such as hen age, nutrition, and environmental temperature are critical determinants of performance (Cordeiro et al., 2025). Recent research highlights that while traditional mathematical models, such as the Modified Compartmental Model or the Adams-Bell model, are effective for general trend analysis, they often struggle with the multi-dimensional variability and non-linear interactions inherent in commercial production cycles (Lemke et al., 2024). In contrast, ML algorithms like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) demonstrate superior accuracy by handling high-dimensional datasets and mitigating the

influence of outliers (Ji et al., 2024). ML has proven to be essential in the optimization of agricultural and poultry processes, enabling greater accuracy and efficiency (Khan et al., 2025; Meshram et al., 2021; Sharma et al., 2021). The use of Machine Learning in agricultural systems allows for optimized resource utilization and highly accurate outcome prediction (Treboux & Genoud, 2018).

Thus, the trajectory of predictive modeling in avian science has transitioned from simple linear approximations to complex ensemble and deep learning architectures. Supervised learning algorithms are particularly effective in scenarios where the objective is to fit a function relating biological inputs to a continuous output like egg volume or laying rate (Adejola et al., 2025). Supervised algorithms, such as Linear Regression and Random Forest, are commonly applied to predict yields and identify patterns in variables that directly influence production, such as temperature and humidity (Ang & Seng, 2021). Studies indicate that by employing these algorithms, producers are able not only to forecast production outcomes but also to quickly identify potential management failures. Random Forest, for instance, has gained prominence due to its resistance to multicollinearity and its capacity to manage the non-linear interactions between multiple production and environmental variables (Adejola et al., 2025). In commercial indoor laying systems, RF has been successfully implemented for tasks ranging from body weight prediction to disease diagnosis, achieving accuracy rates that consistently outperform traditional regression (Adejola et al., 2025).

Furthermore, the emergence of hybrid models, combining time-series decomposition with neural networks, has shown promise in capturing both seasonal trends and irregular fluctuations (Mo et al., 2023). Long Short-Term Memory (LSTM) networks, specialized for sequential data, have been applied to predict production peaks by identifying temporal dependencies over a sliding data window (Lemke et al., 2024). However, for small to medium farms, the complexity and computational cost of deep learning may not always be justified (Kader et al., 2021). Interpretable models like Regression Trees (M5P) provide a compelling middle ground, offering a level of transparency that "black box" neural networks often lack, which is essential for building trust among producers (Ji et al., 2024). This is especially valuable for farms operating with tight profit margins, where accurate forecasting of feed and water consumption can prevent waste and reduce operational costs (Khan et al., 2025).

### *2.2 Case Studies and Results in Poultry Farming*

In egg production, the use of algorithms such as Long Short-Term Memory (LSTM) and Neural Networks has shown promising results in predicting production peaks and dynamically adjusting environmental conditions (Meshram et al., 2021; Sharma et al., 2021). These algorithms enable more effective control of the production cycle, promoting sustainability and ensuring that input consumption is optimized (Ang & Seng, 2021).

Another study conducted by Khan et al. (2025) demonstrated that the integration of predictive algorithms with Internet of Things (IoT) systems significantly improves operational efficiency. The use of sensors to monitor real-time variables, such as temperature and lighting, ensures immediate adjustments, reducing production losses and improving animal welfare.

Based on the evidence presented, it can be concluded that the adoption of Machine Learning algorithms in poultry farms can deliver substantial operational benefits, increasing efficiency and promoting sustainable production. In the context of egg production, such algorithms enable the optimization of production planning, allowing managers to adjust resources and inputs with precision, thereby reducing costs and maximizing profitability (Khan et al., 2025).

## 3 METHODOLOGY

The methodological approach of this study is quantitative in nature, following a deductive logic and adopting a case study design. The study was structured into three main stages: data collection, data preprocessing, and the application of ML techniques, followed by an analysis of the obtained results.

The research focused on the application of ML techniques to a specific process within an egg production farm located in southern Minas Gerais, Brazil. The predictive accuracy of any machine learning model is inherently

tied to the quality and structure of the input data (Lemke et al., 2024). For this case study, historical records from a commercial farm in southern Minas Gerais were consolidated, covering a period where flocks were in a mature productive phase (74 to 120 weeks) (Pelletier, 2017). This late-cycle focus is significant, as the global industry increasingly aims to extend the laying cycle to 100 weeks to enhance resource efficiency and reduce the frequency of flock replacements (Bain et al., 2016).

These data were obtained exclusively from the farm's internal records, which are updated weekly and consolidated in electronic spreadsheets used for zootechnical monitoring. All data were organized into a single spreadsheet, which served as the basis for exploratory analysis, preprocessing, and the application of predictive models.

To address potential confounding factors, the target variable was refined to Hen-Day Egg Production (HDEP) (Cordeiro et al., 2025). Analyzing total weekly eggs alone can lead to scale-driven distortions, where the model simply predicts higher production for larger houses without understanding the underlying efficiency (Cordeiro et al., 2025). HDEP provides a normalized measure of per-bird productivity, allowing for a more robust comparison across different housing units and stocking densities (Cordeiro et al., 2025).

The raw data underwent a rigorous cleaning process using the WEKA 3.8 software environment. This included the removal of inconsistencies and the handling of missing values, which are common in real-world farm datasets due to manual recording errors or temporary system failures (Lemke et al., 2024). Specifically, zero values in production and feed consumption were identified not as noise, but as representing "sanitary voids" or transition weeks between flocks (Pelletier, 2017). These shutdown periods are vital components of the farm's biosecurity protocol, intended to break the cycle of infection and allow for deep disinfection (Ngom et al., 2025). Table 1 shows the variables analyzed and the stages of data cleaning.

**Table 1**

Attribute-Specific Data Preprocessing Techniques and Justifications

| **Attribute** | **Processing Step** | **Rationale** |
|---|---|---|
| Age | Normalization | Essential for distance-based algorithms to prevent scale distortion (Brownlee, 2016). |
| Total Birds | Log-Transformation | Mitigates the influence of extreme flock size variability (Pelletier, 2017). |
| Mortality | Scaling | Ensures mortality peaks do not overwhelm the gradient during training (Brownlee, 2016). |
| Feed per Bird | Imputation | Replaces zeros during active weeks with local means to maintain continuity (WEKA, 2023). |
| Week of Year | Sin/Cos Encoding | Captures the cyclical nature of seasonality in southern Minas Gerais (Pelletier, 2017). |

Source: Developed by the authors.

Feature selection was conducted using the GreedyStepwise search method in WEKA, combined with the CfsSubsetEval evaluator (Brownlee, 2016). This prioritized the most relevant zootechnical indicators: bird age, total count, mortality rate, and feed consumption (Ji et al., 2024). The inclusion of "Week of Year" was specifically evaluated for its ability to capture regional seasonal trends, which are known to impact production by up to 10% due to temperature and daylight fluctuations (Ghysels et al., 2006).

In the third and final stage, appropriate Machine Learning techniques were applied to the farm's operational context. Both supervised and unsupervised models were tested, including linear regression, decision trees, and clustering, with the objective of predicting resource consumption and optimizing the planning of operational activities, such as animal feeding and environmental control. The models were trained on a portion of the dataset and tested on the remaining data, and their performance was evaluated using metrics such as accuracy, precision, and Mean Absolute Error (MAE).

A critical improvement in this revised framework is the transition from a simple 80/20 percentage split to 10-fold Cross-Validation (CV) (Pelletier, 2017). While percentage splits are computationally cheaper, they are prone to selection bias and may fail to account for temporal dependencies in the data (Lemke et al., 2024). Cross-validation provides a more realistic estimate of model robustness by ensuring that every instance in the dataset is used for both training and testing across multiple iterations.30

To evaluate the models effectively, a multi-metric approach was adopted. Beyond the standard correlation coefficient ($R$), the study incorporates the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination ($R^2$). Furthermore, the adjusted coefficient of determination (Adjusted $R^2$) was calculated to penalize model complexity, providing a more robust assessment of model performance (Meyer, 2023).

The application of Multiple Linear Regression (LR) required the verification of several statistical assumptions, including linearity, homoscedasticity, and normality of the residuals. Diagnostic plots generated via WEKA's outputAdditionalStats option enabled the identification of high-leverage points that could bias coefficient estimates.

$$\text{TotOvosSem} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{TotAves} + \beta_3 \text{Mortality} + \beta_4 \text{FeedPerBird} + \varepsilon$$

The statistical significance of individual predictors was assessed using *p*-values, with a threshold of $p < 0.05$ adopted to indicate statistical significance. A key advantage of linear models lies in their interpretability, as they allow the direct quantification of how a one-unit change in a predictor (e.g., an additional week of age) affects the expected response variable, assuming all other variables remain constant (Reichel, 2025).

Finally, the results were analyzed in terms of feasibility and impact on the farm's production process. The potential for reducing operational costs and increasing process efficiency through the use of Machine Learning techniques was evaluated, as well as the practical implications for farm management.

### *3.1 WEKA configuration*

The success of the machine learning models is dependent on specific parameter tuning within the WEKA 3.8 environment. Thus, this section details the configurations used to achieve the reported results – see Table 2 (Pelletier, 2017).

**Table 2**

Predictive Models and WEKA Configuration Parameters with Selection Rationale

| Model | WEKA Configuration / Parameters | Selection Criteria |
|---|---|---|
| Linear Regression | –S 0, –R $1.0 \times 10^{-8}$: M5 attribute selection method | Minimal ridge parameter to allow for maximum feature expression. |
| M5P Tree | - M 4.0 (Minimum number of instances per leaf) | Balanced pruning to prevent overfitting on smaller flock segments. |
| Random Forest | –I 100, –K 0, –S 1: Ensemble of 100 trees, with SQRT($n$) attributes evaluated at each split | Default iterations for robustness; random seed for reproducibility (WEKA, 2023). |

Source: Developed by the authors.

The use of 100 trees in the Random Forest ensemble provided a stable convergence of the error rate. While increasing iterations to 500 or 1000 may yield marginal gains, the computational overhead often outweighs



ISSN: 1994-0098 | e-ISSN: 3086-0210

the benefit for weekly operational planning (WEKA, 2023). The minimal instances per leaf in the M5P model ensured that each linear segment was supported by a statistically significant number of observations, maintaining the validity of the rule-based insights (WEKA, 2012).

# 4 RESULTS AND DISCUSSION

## *4.1 Descriptive Statistics*

This section presents the descriptive statistics of the main variables involved in the study, with the objective of providing an overall view of data behavior prior to the application of predictive models. For each variable, measures of central tendency (mean and median) and dispersion (standard deviation), as well as minimum and maximum observed values, are reported. These indicators allow for the assessment of flock heterogeneity and weekly variability in egg production, bird age, mortality, weekly deaths, and feed consumption. Table 3 summarizes this information.

**Table 3**

Descriptive statistics of the analyzed data

| Variable | Mean | Median | Standard Deviation (SD) | Minimum | Maximum | Comment |
|---|---|---|---|---|---|---|
| Age (weeks) | 90.42 | 89.00 | 11.13 | 74 | 120 | Birds are, on average, in a mature productive phase. |
| Total Birds | 35,518 | 12,289 | 42,362 | 2,695 | 124,384 | High variability among housing units. |
| Mortality (%) | 0.18 | 0.16 | 0.18 | 0.00 | 0.90 | Generally low mortality, with occasional peaks. |
| Absolute Mortality | 128.44 | 37.00 | 215.88 | 2 | 1,878 | Data consistent with flock size variability. |
| Egg production (weekly) | 191,604 | 65,372 | 256,457 | 0 | 782,439 | High variability; some records with zero production. |
| Feed consumption per bird | 0.11 kg | 0.11 kg | 0.07 kg | 0 kg | 0.99 kg | Mean within expected range, but with outliers. |

Source: Developed by the authors.

During the descriptive statistics phase, the average weekly egg production was 191,604 units (SD = 256,457), with flocks ranging from 0 to 782,439 eggs, highlighting the system's natural variability. The average bird age was 90.4 weeks (SD = 11.1), while average feed consumption per bird was 0.11 kg/day (SD = 0.07). The mean mortality rate was 0.18% (SD = 0.18), with an average of 128 weekly deaths (SD = 216). These indicators confirm that the dataset represents mature and heterogeneous production conditions, which are essential for training predictive models (Meshram et al., 2021; Sharma et al., 2021). This initial analysis of the farm's historical records revealed high dispersion across several key variables. The standard deviation for "Total Birds" and "Weekly Egg Production" was equal to or greater than the mean, a condition that could compromise the representativeness of measures of central tendency (Pelletier, 2017). This variability, however, reflects the controlled heterogeneity of a multi-house commercial operation (Ghysels et al., 2006).

The biological justification for this high variability is found in the physiological decline of aging hens. In late-laying phases (90-120 weeks), individual birds within a flock exhibit diverse trajectories of laying persistence (Berger et al., 2025). While body weight tends to increase with age due to abdominal fat accumulation, the

laying rate and eggshell quality progressively deteriorate (Berger et al., 2025). This variability is exacerbated by environmental disturbances, where more resilient birds maintain a uniform production while others suffer sudden drops.41 The dataset thus captures the complex interplay between genetic potential for "long-life" production and the practical limitations of managing older flocks (Bain et al., 2016).

### *4.2 Predictive Performance: Linear Regression and Significance*

The Multiple Linear Regression model served as a foundational baseline for capturing the central trends in the data. With a correlation coefficient of $R = 0.9461$and an adjusted coefficient of determination of Adjusted $R^2 \approx 0.89$, the model exhibited a strong capacity to explain variance in weekly production (Pelletier, 2017). All primary predictors attained high levels of statistical significance ($p < 0.001$), confirming their relevance within the farm's operational context (Reichel, 2025). Table 4 shows the regression coefficients and statistical significance of production predictors.

**Table 4**

Regression Coefficients and Statistical Significance of Production Predictors

| Predictor | Coefficient (β) | Std. Error | t-Stat | p-value | Interpretation |
|---|---|---|---|---|---|
| Intercept | 1,460,503 | 45,211 | 32.30 | < 0.0001 | Baseline production constant. |
| Age | -14,683 | 1,204 | -12.19 | < 0.001 | 14k less eggs per week of aging.[40] |
| Total Birds | -4.5726 | 0.824 | -5.55 | < 0.001 | Inverse density relationship.[43] |
| Feed per Bird | -234,453 | 21,304 | -11.00 | < 0.001 | Indirect indicator of inefficiency. |

Source: Developed by the authors.

A critical insight identified through this model is the negative coefficient for "Total Birds" (Pelletier, 2017). In a purely cumulative model, one would expect a positive coefficient. However, when controlling for per-bird factors like age and feed, a negative coefficient suggests that larger flock sizes are associated with a decrease in per-unit productivity (Savastano & Scandizzo, 2017). This is often a sign of overcrowding stress, where high stocking densities lead to reduced feed intake, increased ammonia levels, and higher competition for nesting boxes.45 High bird density can interfere with resting behaviors and locomotive activities, diverting metabolic energy away from egg production and toward stress recovery (Savastano & Scandizzo, 2017).

The negative coefficient for "Feed per Bird" is also revealing. While feed is the primary input for egg formation, in a late-cycle population, high individual feed consumption without a corresponding increase in production often indicates metabolic inefficiency or the accumulation of abdominal fat, which further inhibits reproductive performance (Berger et al., 2025). This diagnostic capacity of the LR model facilitates clear communication with field teams, providing specific targets for managerial intervention (Pelletier, 2017).

### *4.3 Advanced Hierarchical Modeling: Regression Trees (M5P)*

The M5P algorithm builds upon the linear model by constructing a decision tree with linear regression functions at its leaves. This "model tree" architecture is particularly suited for agribusiness data, as it allows for different linear relationships to exist within defined segments of the population. The M5P model

improved the correlation coefficient to 0.9544 and significantly reduced the MAE to 37,688 eggs, a gain of nearly 30% over the global linear model (Pelletier, 2017).

One observed trend was that in larger flocks (Total Birds > 23,393), age became the most influential factor, suggesting a shift in variable importance according to flock size.

One of the primary advantages of M5P is its interpretability through the generation of specific production rules. In this case study, the model identified a critical threshold based on flock size:

- Rule 1 (Large Flocks): IF Total Birds > 23,393 THEN bird age is the primary predictor of production decline. The model fits a linear function where the negative impact of aging is more pronounced, reflecting the higher sensitivity of large, dense populations to physiological stressors (Pelletier, 2017).
- Rule 2 (Smaller Flocks): IF Total Birds ≤ 23,393 THEN feed intake and mortality peaks are the most accurate predictors of production fluctuations (Cordeiro et al., 2025).

This partitioning of the data space allows managers to tailor their strategies. For large housing units, the focus should be on maximizing laying persistence through metabolic supports for aging hens. For smaller units, the priority shifts to high-resolution monitoring of feed logistics and individual bird health (Meshram et al., 2021; Sharma et al., 2021). The M5P model thus provides actionable guidelines that reconcile the precision of machine learning with the practical segmentation of farm operations (Meshram et al., 2021; Sharma et al., 2021).

### *4.4 Robust Ensemble Learning: Random Forest Performance*

The Random Forest (RF) architecture emerged as the top-performing model, achieving a correlation coefficient of $R = 0.9869$ and an adjusted coefficient of determination of Adjusted $R^2 = 0.97$. By aggregating the predictions of 100 individual decision trees, the RF model effectively reduced the high variability inherent in the raw data and mitigated the impact of multicollinearity among zootechnical predictors. Table 5 shows the comparative predictive performance of RF versus baseline models.

**Table 5**

Comparative Predictive Performance of Random Forest Versus Baseline Models (10-Fold Cross-Validation)

| Metric | Random Forest (10-fold CV) | Improvement over LR | Improvement over M5P |
|---|---|---|---|
| Correlation coefficient ($R$) | 0.9869 | + 4.31% | + 3.41% |
| Mean Absolute Error (MAE), measured in eggs | 18,319 | + 65.8% | + 51.4% |
| Relative Error | 8.51% | - 16.38% | - 8.99% |
| Adjusted coefficient of determination (Adjusted $R^2$) | 0.971 | + 8.86% | + 6.94% |

Source: Developed by the authors.

The superior performance of Random Forest is attributed to its ability to capture complex, non-linear interactions without requiring the manual specification of interaction terms (Adejola et al., 2025). For instance, the combined effect of high age and seasonal heatwaves is naturally integrated into the ensemble's structure (Lemke et al., 2024). Feature importance analysis using the Mean Impurity Decrease method

confirmed that bird age and total count were the most dominant predictors, reflecting their fundamental role in determining both physiological capacity and physical constraints (Ji et al., 2024).

The robustness of the RF model under 10-fold Cross-Validation suggests that it can generalize well beyond the training set, making it an ideal candidate for operational forecasting (Ji et al., 2024). However, the gain in accuracy comes at the cost of direct transparency, as the internal logic of 100 trees is far more difficult to interpret than a single linear equation or an M5P rule set (Ji et al., 2024).

### *4.5 Comparative Analysis and Naïve Baselines*

To contextualize the success of the ML models, they were compared against simple domain baselines, such as persistence (predicting that next week's production will equal this week's) and mean seasonal profiles (Pelletier, 2017). Naïve forecasts in poultry data often suffer from high error rates because they cannot account for the rapid changes associated with flock replacements or sudden environmental shocks (Khan et al., 2025). Next, Table 6 shows the performance comparison of forecasting models for weekly egg production.

**Table 6**

Performance Comparison of Forecasting Models for Weekly Egg Production

| Model Type | Correlation (R) | MAE (eggs) | Contextual Utility |
|---|---|---|---|
| Naïve Persistence | 0.7241 | 112,450 | Worst; fails during transitions. |
| Seasonal Mean | 0.7852 | 98,600 | Limited; ignores bird age dynamics. |
| Linear Regression | 0.9461 | 53,601 | Good diagnostic tool; clear coefficients. |
| M5P Model Tree | 0.9544 | 37,688 | Balanced; provides operational rules. |
| Random Forest | 0.9869 | 18,319 | Best; robust operational forecaster. |

Source: Developed by the authors.

The ML models significantly outperformed the naïve baselines, with Random Forest reducing the error magnitude by over 80% compared to a simple persistence model (Papacharalampous & Tyralis, 2018). This underscores the value of integrating multiple zootechnical indicators to capture the dynamic behavior of commercial flocks (Lemke et al., 2024).

## 5 Practical implications - Economic Impact and Process Optimization

The transition from "prediction" to "optimization" requires a clear quantification of how forecasting accuracy translates into financial outcomes (Garcia Arismendiz & Huertas Zuñiga, 2024). In an industry where margins are often measured in cents per dozen eggs, even small improvements in the Feed Conversion Ratio (FCR) or labor scheduling can have a profound impact on profitability (Wen et al., 2018).

The reduction in MAE achieved by the Random Forest model (18,319 eggs) compared to Linear Regression (53,601 eggs) represents a precision gain of 35,282 units per week (Pelletier, 2017). Using the Egg Production Efficiency Index (EPEI) as a framework, these gains can be converted into economic metrics (Reis et al., 2023):

I. **Feed Logistics Optimization:** Feed represents 60-70% of total production costs (Wen et al., 2018). A forecast error of 53,000 eggs leads to significant over- or under-ordering of grain. Given an average feed cost of $0.35 per kg and a standard FCR of 2.0, the precision gain of the RF model allows for the optimization of approximately 4,200 kg of feed per week across the farm (Milosevic et al., 2019).

II. **Labor Allocation and Scheduling:** High-production peaks require increased staffing for egg collection, grading, and packaging (Wu et al., 2024). Accurate forecasts prevent the financial burden of over-staffing during low-yield weeks and the visual fatigue and breakage risks associated with under-staffing during peak cycles (Wu et al., 2024).

III. **Revenue and Contract Stability:** For regional producers in Minas Gerais, supply chain stability is critical for maintaining contracts with wholesalers. The 8.5% relative error of the RF model provides a reliable baseline for committing to delivery volumes, reducing the need to buy high-priced "gap eggs" from the exchange to fill orders (Wolla, 2025).

Based on the findings of this study, the following strategic recommendations are proposed for poultry producers in southern Minas Gerais:

1. **Transition to HDEP Targets:** Standardize production monitoring around Hen-Day Egg Production to ensure that efficiency metrics are not confounded by house size or stocking density (Cordeiro et al., 2025).
2. **Adopt Multi-Model Diagnostic Hierarchies:** Use Linear Regression for quick diagnostics of variable impact, M5P for defining operational thresholds, and Random Forest for high-accuracy weekly forecasting (Pelletier, 2017).
3. **Invest in Data Integrity:** Prioritize the accurate recording of feed intake and mortality, as these variables serve as critical moderators of laying persistence in late-cycle hens (Lemke et al., 2024).
4. **Incorporate Shutdown Weeks:** View sanitary void periods as essential data points that define the operational cycle and biosecurity health of the farm (Siekkinen et al., 2012).
5. **Pilot Economic Benchmarking:** Quantify the impact of forecasting errors on feed logistics and labor costs to build a business case for further investments in precision agriculture technology (Sanders & Graman, 2009).

## 6 Conclusion

This study aimed to analyze the feasibility of applying Machine Learning (ML) techniques to optimize the production processes of an egg farm located in southern Minas Gerais, Brazil. Through the collection, organization, and modeling of historical production data, it was possible to explore the potential of predictive models as decision-support tools in poultry farming. By transitioning to a Hen-Day Egg Production target and employing robust validation methods like 10-fold Cross-Validation, the predictive framework achieves a level of accuracy and reliability suitable for strategic agribusiness management. The Random Forest model stands out as the most precise operational tool, capable of reducing forecast errors by over 65% compared to linear models. However, the M5P model offers unique value through its ability to segment the population into distinct linear regimes, identifying critical flock size thresholds where management priorities must shift. Linear Regression remains an essential diagnostic component, providing transparent coefficients that quantify the biological costs of aging and overcrowding.

From an economic perspective, the reduction in Mean Absolute Error translates into tangible gains in feed logistics, labor scheduling, and contract fulfillment. The ability to predict production drops and peaks with 91.5% accuracy (as seen in the Random Forest relative error) provides a competitive advantage for regional producers operating in a volatile market. The next phase of modernization for farms in southern Minas Gerais could involve the integration of low-cost environmental sensors. Coupling ML architectures with IoT data streams would enable the creation of "Digital Twins", engineering-grade virtual models of the poultry house that can simulate the impact of environmental changes before they occur. Such systems could provide automated recommendations for ventilation adjustments, further reducing the reliance on manual labor and improving bird welfare.

In conclusion, the integration of artificial intelligence into the poultry industry is not a purely technical endeavor but a strategic imperative that requires the careful alignment of statistical methods with the biological and operational realities of the farm.

While this study demonstrates the high potential of ML in regional poultry farming, several limitations must be acknowledged. The current models rely exclusively on internal farm records, excluding real-time environmental variables such as temperature, humidity, and ammonia levels. Previous research indicates that temperature is the dominant external factor influencing production, with heat stress causing significant declines in both egg weight and laying rate.

Future research is encouraged to explore the integration of environmental sensors, meteorological data, and market information, thereby increasing model complexity and applicability. Further studies should also investigate the real-time implementation of these systems, evaluating not only predictive accuracy but also economic and operational impacts on daily farm management.

## REFERENCES

Adejola, Y. A., Sibanda, T. Z., Ruhnke, I., Boshoff, J., Pokhrel, S., & Welch, M. (2025). Forecasting egg production performance and fluctuations in commercial free-range poultry systems using a random forest model. *Smart Agricultural Technology*, *12*, 101380. https://doi.org/10.1016/j.atech.2025.101380

Ang, K. L.-M., & Seng, J. K. P. (2021). Big Data and Machine Learning With Hyperspectral Information in Agriculture. *IEEE Access*, *9*, 36699–36718. https://doi.org/10.1109/ACCESS.2021.3051196

Bain, M. M., Nys, Y., & Dunn, I. C. (2016). Increasing persistency in lay and stabilising egg quality in longer laying cycles. What are the challenges? *British Poultry Science*, *57*(3), 330–338. https://doi.org/10.1080/00071668.2016.1161727

Berger, Q., Bedere, N., Lagarrigue, S., Burlot, T., Le-Roy, P., Tribout, T., & Zerjal, T. (2025). *Unravelling the genetic architecture of persistence in production, quality, and efficiency traits in laying hens at late production stages* (p. 2025.02.26.640268). bioRxiv. https://doi.org/10.1101/2025.02.26.640268

Brownlee, J. (2016, August 9). How to Work Through a Regression Machine Learning Project in Weka. *MachineLearningMastery.Com*. https://www.machinelearningmastery.com/regression-machine-learning-tutorial-weka/

Bumanis, N. (2024). Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges. *Procedia Computer Science*, *232*, 2302–2309. https://doi.org/10.1016/j.procs.2024.02.049

Cordeiro, A. F., Nääs, I. A., Garcia, R. G., & Valentim, J. K. (2025). Machine Learning-Based Decision Trees to Predict Egg Production Performance in Laying Hens under Heat Stress Conditions. *Brazilian Journal of Poultry Science*, *27*, eRBCA. https://doi.org/10.1590/1806-9061-2025-2130

Fan, W., Peng, H., & Yang, D. (2025). Review: The application and challenges of advanced detection technologies in poultry farming. *Poultry Science*, *104*(11), 105870. https://doi.org/10.1016/j.psj.2025.105870

Garcia Arismendiz, J. A., & Huertas Zuñiga, S. L. (2024). *Improving demand forecasting by implementing machine learning in poultry production company*. https://repositorio.ulima.edu.pe/handle/20.500.12724/20750

Ghysels, E., Osborn, D. R., & Rodrigues, P. M. M. (2006). Chapter 13 Forecasting Seasonal Time Series. In *Handbook of Economic Forecasting* (Vol. 1, pp. 659–711). Elsevier. https://doi.org/10.1016/S1574-0706(05)01013-X

Ji, H., Xu, Y., & Teng, G. (2024). Predicting egg production rate and egg weight of broiler breeders based on machine learning and Shapley additive explanations. *Poultry Science*, *104*(1), 104458. https://doi.org/10.1016/j.psj.2024.104458

Kader, M. S., Ahmed, F., & Akter, J. (2021). Machine Learning Techniques to Precaution of Emerging Disease in the Poultry Industry. *2021 24th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/ICCIT54785.2021.9689828

Khan, S. A., Shukla, A. K., Yadav, S. K., & Vishwakarma, G. K. (2025). Machine learning models for analysis and prediction of optimal egg production. *Quality & Quantity*. https://doi.org/10.1007/s11135-025-02309-1

Lemke, V., Andrade, J. O., & Komati, K. S. (2024). Machine Learning Techniques for Egg Production Prediction. *Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE)*. https://doi.org/10.55592/cilamce.v6i06.10159



ISSN: 1994-0098 | e-ISSN: 3086-0210

Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, *1*, 100010. https://doi.org/10.1016/j.ailsci.2021.100010

Meyer, C. (2023). *RegressionAnalysis* [Weka.classifiers.evaluation]. Class RegressionAnalysis. https://weka.sourceforge.io/doc.dev/weka/classifiers/evaluation/RegressionAnalysis.html

Milosevic, B., Ciric, S., Lalic, N., Milanovic, V., Savic, Z., Omerovic, I., Doskovic, V., Djordjevic, S., & Andjusic, L. (2019). Machine learning application in growth and health prediction of broiler chickens. *World's Poultry Science Journal*, *75*(3), 401–410. https://doi.org/10.1017/S0043933919000254

Mo, L., Jiang, M., Fang, X., & Shi, X. (2023). *A Novel Hybrid STL-Based Model for Egg Price Forecasting*. IEIT 2023, AHSSEH *10*, 365–382. https://doi.org/10.2991/978-94-6463-230-9_44

Ngom, R. V., Jajere, S. M., Watsop, H. M., & Tanyienow, A. (2025). Relation Between Farm Biosecurity Measures and Poultry Production Performances: A Scoping Review. *Veterinary Medicine and Science*, *11*(e70526), 1–10. https://doi.org/10.1002/vms3.70526

Papacharalampous, G. A., & Tyralis, H. (2018). Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences*, *45*, 201–208. https://doi.org/10.5194/adgeo-45-201-2018

Pelletier, N. (2017). Life cycle assessment of Canadian egg products, with differentiation by hen housing system type. *Journal of Cleaner Production*, *152*, 167–180. https://doi.org/10.1016/j.jclepro.2017.03.050

Reichel, F. (2025). *Statistically Significant Linear Regression Coefficients Solely Driven By Outliers In Finite-sample Inference* (No. arXiv:2505.10738). arXiv. https://doi.org/10.48550/arXiv.2505.10738

Reis, M. P., Ferreira, N. T., Gous, R. M., & Sakomura, N. K. (2023). Update and evaluation of the egg production model in laying hens. *Animal*, *17*, 101015. https://doi.org/10.1016/j.animal.2023.101015

Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega*, *37*(1), 116–125. https://doi.org/10.1016/j.omega.2006.10.004

Savastano, S., & Scandizzo, P. L. (2017). *Farm Size and Productivity: A "Direct-Inverse-Direct" Relationship*. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-8127

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, *9*, 4843–4873. https://doi.org/10.1109/ACCESS.2020.3048415

Siekkinen, K.-M., Heikkilä, J., Tammiranta, N., & Rosengren, H. (2012). Measuring the costs of biosecurity on poultry farms: A case study in broiler production in Finland. *Acta Veterinaria Scandinavica*, *54*(1), 12. https://doi.org/10.1186/1751-0147-54-12

Treboux, J., & Genoud, D. (2018). Improved Machine Learning Methodology for High Precision Agriculture. *2018 Global Internet of Things Summit (GIoTS)*, 1–6. https://doi.org/10.1109/GIOTS.2018.8534558

WEKA. (2012, April 12). *M5P*. GitLab. https://git.cms.waikato.ac.nz/weka/weka/-/blob/abe22514756dbb7ef425bd7d60058d033f93b69f/dev-3-5-8/weka/classifiers/trees/M5P.java

WEKA. (2023, April 17). *Wekadocs · stable-3-8 · WEKA / weka · GitLab*. GitLab. https://git.cms.waikato.ac.nz/weka/weka/-/tree/stable-3-8/wekadocs

Wen, C., Yan, W., Zheng, J., Ji, C., Zhang, D., Sun, C., & Yang, N. (2018). Feed efficiency measures and their relationships with production and meat quality traits in slower growing broilers. *Poultry Science*, *97*(7), 2356–2364. https://doi.org/10.3382/ps/pey062



ISSN: 1994-0098 | e-ISSN: 3086-0210

Wolla, S. A. (2025). *The Market for Eggs: How Prices Are Hatched*. https://www.stlouisfed.org/publications/page-one-economics/2025/may/market-for-eggs-how-prices-are-hatched

Wu, Z., Zhang, H., & Fang, C. (2024). Research on machine vision online monitoring system for egg production and quality in cage environment. *Poultry Science*, *104*(1), 104552. https://doi.org/10.1016/j.psj.2024.104552



ISSN: 1994-0098 | e-ISSN: 3086-0210