Machine Learning Techniques for Egg Production Prediction

Valdinéia Lemke[1], Jefferson O. Andrade[1], Karin S. Komati[1]

[1]*Graduate program in Applied Computing (PPComp)*
Instituto Federal do Espírito Santo (IFES), Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, 29166-630, Serra-ES, Brazil
valdilemke@gmail.com, {jefferson.andrade, kkomati}@ifes.edu.br

Abstract. Predicting egg production in poultry farming is a complex task due to the multitude of influencing factors such as temperature, nutrition, and environmental conditions. This study aims to evaluate the performance of various machine learning models in forecasting egg production using multivariate time series data. The dataset comprises records of the Hy-Line chicken breed, divided into four batches, with attributes including age, maximum and minimum temperature, feed and water consumption, and daily production percentage. The study employs a sliding window technique to capture temporal patterns and evaluates models including Ridge Regression, Random Forest, XGBoost, and MLP (Multi-Layer Perceptron). The models were trained on three batches and tested on the fourth, with performance measured using Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The results indicate that Ridge Regression, with a window size of 7 days, provided the most accurate predictions, achieving an MSE of 19.74 and a MAPE of 3.81%. This study demonstrates the effectiveness of machine learning techniques and the sliding window approach in improving the accuracy of egg production forecasts, offering valuable insights for poultry farm management and optimization.

Keywords: Sliding Window, Hy-Line breed, Poultry farming.

1 Introduction

Poultry farming has played a significant role in food production worldwide due to its ability to provide high-quality protein. Poultry farming is crucial for food security, offering an essential protein source [1]. In 2021, global table egg production reached 87.60 million tonnes, marking a significant increase of 26.78% compared to 2010 [2]. Projections indicate that production will reach 95 million tonnes by 2030, with a predicted 9% increase in global consumption compared to 2021. Brazil ranks as the seventh-largest egg producer in the world, accounting for approximately 3.15% of global production.

Predicting egg production is a complex challenge due to the multiple variables involved, such as temperature, nutrition, and environmental conditions. Machine learning models have proven effective in forecasting egg production, enabling the analysis of large datasets and the identification of complex patterns influencing production [3]. The use of multivariate time series, which consider the interdependence between multiple variables over time, offers a more comprehensive approach to modeling and predicting dynamic phenomena such as egg production. These series allow for the incorporation of factors like temperature, humidity, feed and water consumption, among others, improving the accuracy of predictions.

The work of Bumanis et al. [4] uses a dataset with daily records containing egg production, temperature, humidity, CO_2, NH_3, feed, and water consumption. Different window sizes (1, 2, 3, 5, 7, and 14 days) were tested to determine the optimal window size for each model, including LSTM, CNN, Random Forest, and XGBoost. The machine learning models are trained with 90% of the data from one batch, validated with 10% of the same batch, and tested with another batch from the Lochman brown breed. The evaluation metrics are Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Percentage Error (RMSPE).

This work aims to continue the research of Bumanis et al. [4] by incorporating their methodology, previously used machine learning algorithms (Random Forest and XGBoost), and introducing additional ones (Ridge Regression and Multi-Layer Perceptron – MLP). Some differences include using the Hy-Line breed for the dataset in this work, training and validating the models with data from three batches, and testing with a fourth batch. The evaluation metrics are MSE and MAPE. Also, our dataset did not include some attributes such as humidity, CO_2, and NH_3.

The structure of the article is as follows: Section 2 explores related works. Section 3 describes the experimental methodology conducted. Section 4 details the results obtained and their analysis. Finally, Section 5 concludes the article with final considerations and suggestions for future work.

2 Related Work

The work of Ahmad [5] evaluates various mathematical, statistical, and artificial intelligence models to forecast egg production in commercial layers. The study uses data collected from a comparative layer trial on 22 commercial strains at the Poultry Research Farms, Auburn University, and simulated data generated using mean and standard deviation of egg production. The study compares three neural network architectures—back-propagation-3, Ward-5, and the general regression neural network (GRNN)—against traditional models like linear regression and the Gompertz nonlinear model. The GRNN model demonstrated superior performance, achieving an R^2 of 0.715, significantly higher than other models. The study highlights that brown-shelled strains consumed more feed (114 g/bird/day) and produced more eggs (89.78% egg production) than white-shelled strains (105 g/bird/day feed consumption and 86.89% egg production). Regression analysis indicated a positive correlation between feed consumption and egg production, with a prediction equation of 4.0428 + 0.7663(feed consumption). Despite initial overprediction, the GRNN accurately predicted egg production phases, suggesting its practical application in commercial farm management for efficient and accurate egg production forecasting.

The article of Gonzalez-Mora et al. [6] examines the impact of environmental control strategies (ECSs) on hen-day egg production (HDEP) and daily egg cleanliness (EGC) in cage-free aviary housing systems. Utilizing Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), the study evaluated various ECSs, including reduced litter surface area, heated floor with oil sprinkling, and litter absorbent with oil sprinkling. The study found that ECSs did not disrupt egg production, with HDEP averaging 97.6% and EGC 87.0%. The Random Forest model, with a 14-day window, effectively predicted HDEP fluctuations, achieving a RMSE of 0.176% and 0.368%, and R^2 of 0.94 and 0.78 for training and testing datasets, respectively. Temperature emerged as the dominant factor influencing egg production, followed by hen's age and relative humidity. A scenario analysis indicated that a 5% increase in temperature could negatively impact egg yield, highlighting the significance of maintaining optimal environmental conditions for maximizing productivity in cage-free aviary systems.

The work of Magemo et al. [7] explores the application of various machine learning algorithms to forecast egg production, aiming to enhance both farm-level and national agricultural economics. The study evaluates four machine learning models: Artificial Neural Network (ANN), Fuzzy Logic, Random Forest, and Support Vector Machine (SVM), each using different sets of input features. The ANN, despite its popularity and high mean feature value, struggles to extract core features due to its limited dataset size, impacting model stability. Fuzzy Logic, employing numerous features, also suffers from small dataset limitations, yielding high prediction accuracy but with a notable relative error of 0.11744. Random Forest and SVM, though efficient with fewer features and datasets, demonstrate varying success, with SVM achieving a high accuracy of 98%. However, these methods lack comprehensive feature utilization, reducing model robustness.

Bumanis et al. [4] investigates the application of machine learning models to predict hen egg production with limited data. Comparing traditional non-linear models, such as the Modified Compartmental Model, with machine learning models like LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), XGB Regressor, and Random Forest Regressor (RF). They conclude that the LSTM, RF, and XGB models overall showed the best performances. The LSTM model achieved the best performance with a two-day sliding window, presenting a MAPE of 5.39% and a Root Mean Squared Percentage Error (RMSPE) of 7.75%. In contrast, the Modified Compartmental Model had a MAPE of 9.13% and an RMSPE of 14.81%. These results indicate that machine learning models are effective in adapting to variations in egg production data, providing more accurate forecasts compared to traditional non-linear models.

3 Materials and methods

3.1 Database

The raising of laying hens involves three main phases: brooding, rearing, and laying [3]. During the brooding phase (0 to 10 weeks), chicks receive intensive care, including temperature control and balanced nutrition. In the rearing phase (10 to 17 weeks), the birds develop rapidly and require adequate nutrition. In the laying phase (from the 17th week onward), hens reach sexual maturity and begin laying eggs, necessitating specific feeding and management conditions. To ensure satisfactory development of pullets in all phases, it is crucial to provide high-quality feed that meets their nutritional needs. These needs are influenced by a variety of factors, including breed, age, temperature, environment, health status, and rearing systems [8].

The dataset used in this study comprises 1761 records of the Hy-Line breed, divided into four distinct batches: H1, H2, H3, and H4. These batches were housed at different periods, allowing for the analysis of temporal and environmental variables on egg production. The age of the birds considered in the analysis ranges from 23 to 85 weeks (161 to 601 days), covering a period of 63 weeks. The dataset was provided by the Nater Coop cooperative, located in the municipality of Santa Maria de Jetibá, ES, which operates an avian condominium. The dataset includes the following attributes:

- Batch: Identifying code of the bird batch.
- Age in Days: Age of the birds in days.
- Max Temperature: Maximum temperature recorded on the egg collection day.
- Min Temperature: Minimum temperature recorded on the egg collection day.
- Feed Consumption (Day): Amount of feed consumed, in grams, per bird on the egg collection day.
- Water Consumption (Day): Amount of water consumed, in milliliters, per bird on the egg collection day.
- % Production: Daily production percentage, calculated as the number of eggs divided by the number of live birds.

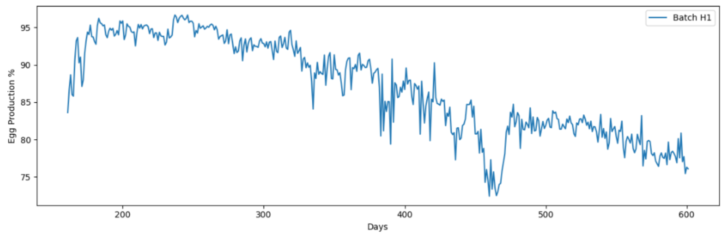

Figure 1. H1.

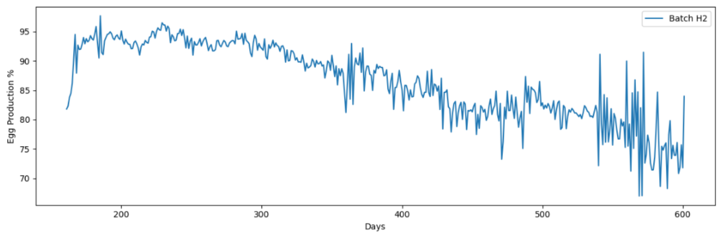

Figure 2. H2.

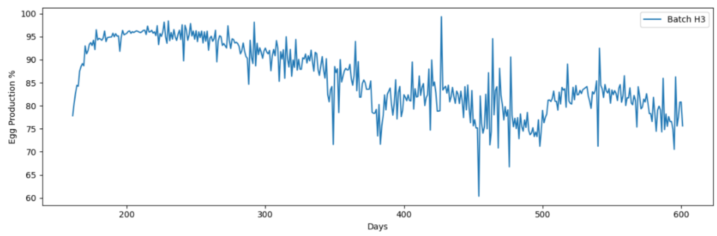

Figure 3. H3.

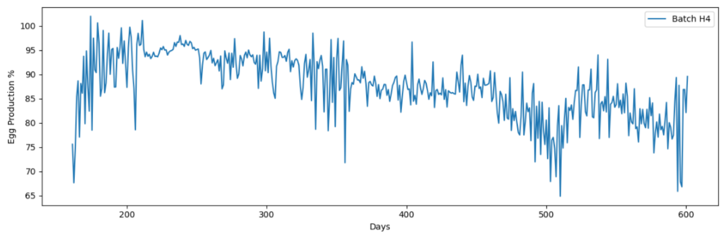

Figure 4. H4.

The target variable in this study is the production percentage (% Production), while the other variables, except for the batch code, were used as predictors in the machine learning model. The graphs showing the production percentage over the days for the four batches are presented in Figure 1, Figure 2, Figure 3, and Figure 4. The Y-axis represents the production percentage, and the X-axis represents the age of the birds in days. The curves of the graphs for the four batches (H1, H2, H3, and H4) show notable differences, although all exhibit a general decreasing trend over time. Batch H1 starts with high and relatively stable production but experiences significant fluctuations and a sharp decline around 400 days. Batch H2 maintains consistent high production for a longer period, showing a gradual decrease and instabilities approaching 500 days. Batch H3 follows a similar pattern, with a strong start and a more pronounced and irregular decline, especially after 300 days. Finally, Batch H4, while also starting with high production, displays greater variability over time, with frequent peaks and troughs, and a less uniform decrease compared to the other batches. These differences can be attributed to variations in environmental conditions, management practices, and other factors specific to each batch.

3.2 Methods

The sliding window technique is widely used in time series modeling to create training and testing datasets. This technique involves creating fixed-size temporal windows that move along the time series, capturing subsequences of the original data. Each window consists of a set of historical data used to predict future values. For example, when using a window size of 7, data from the past seven days are used to predict the value for the next day. The sliding window technique is effective in capturing temporal patterns and trends, enabling machine learning models to make more accurate predictions, especially when working with multivariate time series where multiple interdependent variables influence the target variable [9].

By applying sliding windows, it is possible to include relevant temporal information from various variables, enhancing the predictive capability of the models. Additionally, this technique helps maintain the temporal continuity of the data, which is crucial for accurately forecasting future events. The choice of window size is a critical aspect and should be determined based on the nature of the data and the periodicity of the expected variations. Smaller window sizes can capture short-term variations, while larger sizes may be more effective for capturing long-term trends.

In this study, the sliding window technique was used to create the training and testing datasets, allowing for a robust evaluation of the machine learning models applied to egg production forecasting. Similar to the work of Bumanis et al. [4], different window sizes (1, 2, 3, 5, 7, and 14 days) were tested to determine the optimal window size for each model, with predictions made one day ahead.

In predicting the percentage of egg production, different machine learning models are applied to optimize prediction accuracy. For the purpose of this research, the following models were selected:

- **Ridge Regression** - This is a form of linear regression that includes a regularization term (L2 regularization), which helps to prevent overfitting by penalizing large coefficients. Ridge Regression is particularly useful in situations where there is multicollinearity among the predictor variables, as it shrinks the coefficients of correlated predictors towards each other, distributing the impact more evenly and improving the model's stability and generalization [10].
- **Random Forest** - This is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction of the individual trees. Random Forests are known for their robustness and accuracy, as they reduce overfitting by averaging multiple trees. Each tree is trained on a bootstrapped sample of the data, and at each split in the tree, a random subset of features is considered, which decorrelates the trees and enhances the model's ability to generalize to unseen data [11].
- **XGBBoost** - The Extreme Gradient Boosting is a powerful and efficient implementation of gradient boosting algorithms. It is designed for speed and performance and can handle large datasets with high dimensionality. XGBoost works by sequentially adding models to correct errors made by previous models, focusing on hard-to-predict cases, and combining the strengths of these "weak" models to form a "strong" learner. It includes several regularization terms (L1 and L2) to control model complexity and prevent overfitting, making it robust and reliable for a wide range of predictive tasks [12].
- **MLP** - The MLP Regressor is a type of feedforward artificial neural network that consists of multiple layers of neurons (perceptrons), including input, hidden, and output layers. Each neuron in one layer connects to every neuron in the next layer, with each connection having an associated weight. The MLP uses non-linear activation functions to capture complex relationships and patterns in the data. It is trained using backpropagation, where the error is propagated backward through the network to update the weights, minimizing the difference between the predicted and actual values. This makes the MLP Regressor highly effective for modeling non-linear relationships and interactions between variables [13].

The data were divided into training and test sets. The H1, H2, and H3 batches were used for training the models, while the H4 batch was reserved for testing. This division results in approximately 75% of the data for training and 25% for testing

3.3 Metrics

In this study, we used two metrics: Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The formulas for these metrics are presented in Table 1. The MSE measures the average of the squared differences between predicted and actual values, more severely penalizing larger errors and being sensitive to outliers. It is particularly useful for identifying the variability of prediction errors. The MAPE calculates the average of the absolute differences between predicted and actual values, expressed as a percentage of the actual values. It enables the comparison of model accuracy across different scales by interpreting errors in relative terms. Note that, in Table 1: y_i means the observed value; $\hat{y}_i$ means the predicted value; and n is the number of records.

Table 1. Criteria used for model evaluation.

Criteria	**Equation**
Mean Squared Error	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Mean Absolute Percentage Error	$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left\|\frac{y_i - \hat{y}_i}{y_i}\right\|$

4 Results

Table 2 presents the results obtained by each machine learning model (columns) for different window sizes (first column) and metrics (second column). The cells with a gray background indicate the best (lowest) values per row, and the bold markings highlight the best values in the table for both MSE and MAPE. The Ridge Regression model presented the best results in almost all windows, except for the window of size 2. The best performance in terms of prediction accuracy, considering all windows, was also achieved with Ridge Regression, with an MSE value of 19.740301 and a MAPE of 0.038132 in the window of size 7. This result suggests that the Ridge Regression model is effective in predicting the percentage of egg production when using a window of size 7.

Table 2. Metric values by model and different window sizes

Windows Size	**Error Metric**	**MLP**	**RF**	**Ridge**	**XGBoost**
1	MSE	24.6131	25.1327	24.3887	27.7447
1	MAPE	0.0427	0.0439	0.0416	0.0458
2	MSE	23.0443	22.7263	23.4308	24.8256
2	MAPE	0.0406	0.0412	0.0408	0.0437
3	MSE	21.6610	21.4344	21.3102	23.5427
3	MAPE	0.0402	0.0400	0.0398	0.0422
5	MSE	21.7703	22.4404	21.2953	24.4543
5	MAPE	0.0401	0.0413	0.0393	0.0433
7	MSE	20.9387	22.6219	**19.7403**	23.5526
7	MAPE	0.0393	0.0409	**0.0381**	0.0417
14	MSE	25.0185	24.3415	23.1969	27.2132
14	MAPE	0.0439	0.0433	0.0410	0.0457

Table 3. The best results of model evaluation

Model	**MSE**	**MAPE**	**Window Size**
MLP	20.939	0.039	7
RF	21.434	0.040	3
Ridge	19.704	0.038	7
XGBoost	23.553	0.041	7

The windows of size 1 and 14 presented the worst results across all models. The XGBRegressor, although generally robust, showed the worst performance. The analysis of the results indicates that window size is a critical factor that significantly influences the models' performance. In particular, windows of size 7 seem to provide a good balance between the amount of historical data used and the accuracy of the predictions. Table 3 shows the best result of the MAPE metric (third column) for each model (rows). The MLP and XGBoost also achieved their best performance with the window of size 7, while the best result for Random Forest was obtained with the window of size 3. Figure 5 presents the graph showing the actual (blue) and predicted (orange) values of the percentage of egg production for Ridge Regression, considering a window size of 7, and Figure 6 shows the graph for XGBoost. Next to each graph, an enlarged image is presented to facilitate the visualization of the section where the XGBoost prediction failed to follow the trend of the curve.

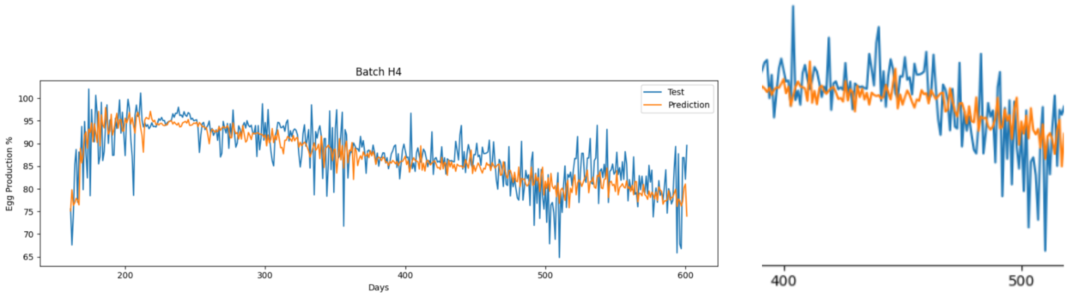

Figure 5. Best result for window size of 7 days: Ridge Regression (MAPE = 3.9%).

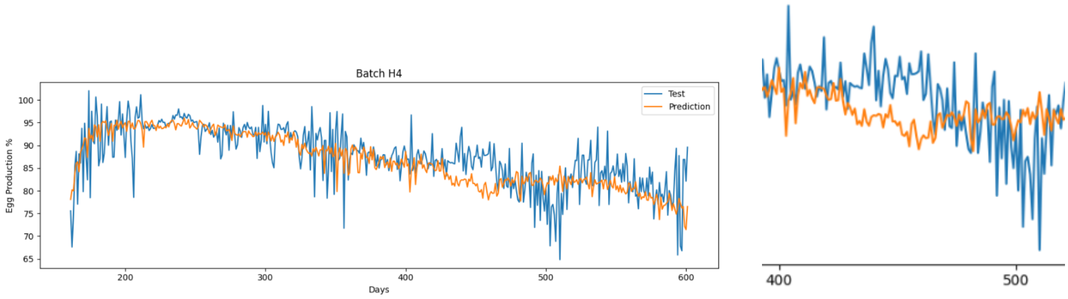

Figure 6. Worst result for window size of 7 days: XGBoost (MAPE = 4.17%).

Compared to the study by Bumanis et al. [4], where the best result was obtained with the LSTM model using a window size of 2, resulting in a MAPE of 5.390%, our study stood out, achieving the best MAPE of 3.81% with the Ridge Regression model using a window size of 7. The worst result of our study was a MAPE of 4.583%, still better than the best result of the compared study. Comparing the algorithms common to both studies, the Random Forest in our study obtained the best result with a window size of 3, presenting a MAPE of 4.00%. In the study by Bumanis et al. [4], the best result for the Random Forest was with a window size of 5, resulting in a MAPE of 6.077%. For the XGBoost algorithm, our study showed the best result with a window size of 7, achieving a MAPE of 4.10%, while in the study by Bumanis et al. [4], the best result for XGBoost was with a window size of 14, presenting a MAPE of 6.114%. This difference in results may be associated with the size of the dataset and the predictive variables used in each study. We were able to confirm the validity of the methodology used in the study by Bumanis et al. [4].

5 Conclusions

Similar to the study by Bumanis et al. [4], this study demonstrated the effectiveness of using machine learning models in scenarios with limited data sets. The contributions of this study are significant for predicting egg production in poultry farms, including small-scale producers who may not have records of many variables or a large dataset. The application of machine learning techniques with the sliding window methodology provides a robust tool for improving the management and optimization of egg production, enabling more informed and efficient decisions.

The Ridge Regression model stood out by achieving the best results in terms of MSE and MAPE, especially with a window size of 7 (one week). This performance suggests that Ridge Regression is effective in handling multicollinearity among predictor variables and provides stable and accurate predictions. On the other hand, models like XGBoost and Random Forest, although generally effective in many applications, may require finer hyperparameter tuning to improve their performance in this specific context. Furthermore, the analysis of different window sizes showed that a window size of 7 offers a good balance between the amount of historical data used and the accuracy of the predictions. Smaller and larger window sizes showed inferior performance, highlighting the importance of carefully selecting the window size to adequately capture the temporal dynamics of the data. The analysis revealed that the choice of machine learning model and window size are crucial factors that significantly influence predictive performance. However, it is important to emphasize that obtaining a larger dataset and more predictor variables could potentially alter the results, possibly further improving model accuracy.

For future research, it is recommended to explore hyperparameter tuning and the inclusion of additional variables in the model, as well as the application of advanced feature selection techniques. Such enhancements can potentially further improve the performance of predictive models. It is also intended to apply these techniques to other chicken breeds. Additionally, the implementation of AutoML techniques using genetic algorithms can help identify the most effective model configurations for different data contexts.

Acknowledgements. The authors would like to thank FAPES/UnAC (No. FAPES 1228/2022 P 2022-CD0RQ, No. SIAFEM 2022-CD0RQ) for the financial support provided through the Sistema UniversidaES; CAPES/FAPES for the PDPG (Graduate Development Program, Strategic Partnerships in the States) – SIAFEM nº 2021-2S6CD, nº FAPES 132/2021; FAPES for project nº 1023/2022 P:2022-8TZV6 and CNPq for project nº407742/2022-0. Professor Komati thanks CNPq for the DT-2 grant (nº 302726/2023-3).

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] G. Gržinić, A. Piotrowicz-Cieślak, A. Klimkowicz-Pawlas, R. L. Górny, A. Ławniczek-Wałczyk, L. Piechowicz, E. Olkowska, M. Potrykus, M. Tankiewicz, M. Krupka, and others. Intensive poultry farming: A review of the impact on the environment and human health. *Science of The Total Environment*, vol. 858, pp. 160014, 2023.
[2] FAO. Faostat. food and agriculture data. Access date: 25 nov. 2023, 2023.
[3] A. Cohen and S. Horovitz. Machine learning based egg supply forecasting for sorting and grading. *Journal of Applied Agricultural Technology*, vol. 10, n. 2, pp. 52, 2023.
[4] N. Bumanis, A. Kviesis, L. Paura, I. Arhipova, and M. Adjutovs. Hen egg production forecasting: Capabilities of machine learning models in scenarios with limited data sets. *Applied Sciences*, vol. 13, n. 7607, 2023.
[5] H. Ahmad. Egg production forecasting: Determining efficient modeling approaches. *Journal of Applied Poultry Research*, vol. 20, n. 4, pp. 463–473, 2011.
[6] A. F. Gonzalez-Mora, A. N. Rousseau, A. D. Larios, S. Godbout, and S. Fournel. Assessing environmental control strategies in cage-free aviary housing systems: Egg production analysis and random forest modeling. *Computers and Electronics in Agriculture*, vol. 196, pp. 106854, 2022.
[7] A. Magemo, L. Minga, and S. L. Kusyama. Analysis of machine learning technique to predict eggs production in poultry farms. *MUST Journal of Research and Development (MJRD*, vol. 2, pp. 393–404, 2022.
[8] S. Leeson and J. D. Summers. *Commercial poultry nutrition*. Nottingham university press, 2009.
[9] T. Cinto, A. Gradvohl, G. Coelho, and A. Da Silva. Solar flare forecasting using time series and extreme gradient boosting ensembles. *Solar Physics*, vol. 295, n. 7, pp. 93, 2020.
[10] S. Liu and E. Dobriban. Ridge regression: Structure, cross-validation, and sketching, 2020.
[11] P. Probst, M. N. Wright, and A.-L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, n. 3, pp. e1301, 2019.
[12] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016.
[13] M. Mahmudul, A. Mia, S. Biswas, M. Urmi, and A. Siddique. An algorithm for training multilayer perceptron (mlp) for image reconstruction using neural network without overfitting. *International Journal of Scientific & Technology Research*, vol. 4, 2015.