NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Impact of different lag-days on predicting one-week and two-weeks cow milk yield using Automated Milking Systems data

Patrícia Alexandra dos Santos Morais

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Impact of different lag-days on predicting one-week and two-weeks cow milk yield using Automated Milking Systems data

by

Patrícia Alexandra dos Santos Morais

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Leonardo Vanneschi, Full Professor, Nova Information Management School, Universidade Nova de Lisboa, Portugal

Co-supervised by

Karina Brotto Rebuli, PhD Student, University of Turin, Italy

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, July 2024]

ACKOWLEDGEMENTS

I would like to express my gratitude to my Supervisor, Professor Leonardo Vanneschi, for his encouragement throughout this journey. His feedback and motivation, especially during the meetings, made me believe that I could achieve my goals and succeed. A special thank you to Karina Rebuli, who demonstrated exceptional support and dedication every single day. Throughout the last few months, Karina was unequivocally my main support, consistently showing her willingness and availability to address my countless questions.

To my friend Marta, I am profoundly grateful for your incredible support, you helped me in every possible way. Your help and patience during all our two-hour calls meant everything to me. To my friends Joana, Margarida, Kiko, and Gui, thank you for your constant efforts to be there for me, even during the times when I didn't have the time to be with you. And also, to Silvana, that even far supported me every single day. To my co-workers, especially Carolina, thank you for sharing every single step of this journey with me. Our mutual motivation and support were essential for me. A big thank you to my friend Samuel, who constantly encouraged me with his advice and never let me give up.

I also want to express my deepest gratitude to my mom. Her daily efforts to ensure I had the time to dedicate to my thesis, including preparing meals for me when I came home tired from work, were invaluable. Her dedication in supporting me in every way every day, made this accomplishment possible. Finally, to my dearest boyfriend, Miguel, thank you for dealing with my bad moods and having the patience to listen to me stress. Without your invaluable help in running the models, sometimes even in the middle of the night, this work would not have been possible. Your support meant everything to me.

ABSTRACT

This study takes advantage of predictive modeling in dairy farming by addressing key gaps identified in the literature, focusing on forecasting one-week and two-weeks-ahead average milk yield using data from Automated Milking Systems (AMSs). By examining the temporal impacts, more specifically how it affects the predictions, and primiparous and multiparous variability, robust models for milk production were developed using data from multiple farms. Three different machine learning methods were employed: Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANNs), and Genetic Programming (GP). XGBoost and ANN, were chosen based on their established effectiveness, while GP introduced a novel approach in the field. The analysis considered seven different lag periods (1 to 7 days) of data to evaluate the importance of historical data in prediction accuracy for both one-week and two-weeks-ahead predictions. The data was also divided by the lactation group (primiparous and multiparous), resulting in 28 different datasets. The findings indicated strong predictive performance, particularly for one-week-ahead forecasts. As expected, the performance decreased for two-weeks-ahead predictions, underscoring the challenges of longer-term forecasting. XGBoost demonstrated the best predictive performance, achieving, overall, the lowest RMSE values on unseen data, while GP showed the weakest performance and ANN yielded variable results. The study revealed that increased lag data significantly enhanced XGBoost's predictions, whereas ANN and GP showed more variable impacts from lag periods. Despite XGBoost's overall superior accuracy, it exhibited substantial overfitting when trained on primiparous data and tested on multiparous cows. On the contrary, ANN and GP showed a better generalization. The GP approach, though presenting lower predictive ability, enabled the identification of the most important features for predictions. The milk yield, number of days in lactation, and number of milking failures from the previous days were particularly important across all models. The unexpected prominence of the number of milking failures suggests its critical role in forecasting milk yield, offering new insights for farm management and animal welfare.

KEYWORDS

AMS; Milk Yield; Machine Learning; Genetic Programming

Sustainable Development Goals (SDG):

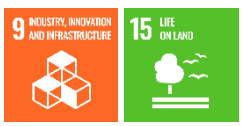

TABLE OF CONTENTS

STATEMENT OF INTEGRITY ii
ACKOWLEDGEMENTS iii
ABSTRACT iv
LIST OF FIGURES vii
LIST OF TABLES viii
LIST OF ABBREVIATIONS AND ACRONYMS xi
1. INTRODUCTION 1
2. THEORETICAL BACKGROUND 3
2.1. Artificial Neural Networks 3
2.2. Extreme Gradient Boosting 6
2.3. Genetic Programming 8
3. LITERATURE REVIEW 12
3.1. Animal welfare and productivity in modern dairy farming 12
3.1.1. Precision Livestock Farming 12
3.1.2. Automated Milking Systems 13
3.2. Predictive modeling in Dairy Farming 14
4. METHODOLOGY 18
4.1. Data Exploration 18
4.2. Data Preparation 19
4.2.1. Daily registers Dataset 19
4.2.1.1. Data filtering 19
4.2.1.2. Feature engineering 21
4.2.2. Panel Dataset 22
4.3. Modeling 25
4.3.1. Temporal and Lactation Modeling 25
4.3.2. Models 25
4.3.2.1. XGBoost 25
4.3.2.2. Artificial Neural Networks 26
4.3.2.3. Genetic Programming 26
4.4. Evaluation 26
5. EXPERIMENTAL STUDY 27

5.1. Experimental Settings 27
5.1.1. XGBoost setup 27
5.1.2. ANN setup 28
5.1.3. GP setup 29
5.2. One week ahead Milk Yield prediction 30
5.2.1. XGBoost 30
5.2.2. ANN 32
5.2.3. GP 34
5.2.4. Comparison between different lags' models 36
5.2.5. Predict lactations_2_3_4_5 target with models trained with first lactation data 40
5.3. Two weeks ahead Milk Yield prediction 42
5.3.1. XGBoost 42
5.3.2. ANN 44
5.3.3. GP 46
5.3.4. Comparison between different lags' models 48
5.3.5. Predict lactations_2_3_4_5 target with models trained with first lactation data 52
5.4. Comparison between different models 54
5.5. Feature importance assessed with GP trees 57
6. CONCLUSION 60
BIBLIOGRAPHICAL REFERENCES 63
APPENDIX A 67

LIST OF FIGURES

Figure 2.1 - Example of a single Neuron Perceptron (Source: Vanneschi & Silva, 2023, p.163). 4
Figure 2.2 - Example of a Feedforward, multilayer ANN (Source: Vanneschi & Silva, 2023, p.176) 5
Figure 2.3 - Example of ensemble of classifiers (source: Ferreira & Figueiredo, 2012, p.2) 7
Figure 2.4 - Graphic description of an evolutionary algorithm, including GP, driven by the concepts of Darwin's Theory of Evolution. (Source: Leonardo Vanneschi, Genetic Programming: Introduction, Motivations and Applications. Lecture Notes. 2020 (cit. on pp. 9, 15)) 9
Figure 2.5 - An example of standard GP crossover (Source: : Vanneschi & Silva, 2023, p.212) 11
Figure 2.6 - An example of standard GP mutation (Source: : Vanneschi & Silva, 2023, p.213) 11
Figure 4.1 - Example of the creation of a register in the panel dataset using 3 lag days 24
Figure A. 1 - GP evolution through 50 generation for both week predictions and lactation groups 71
Figure A. 2 - Features that significantly contributed trees' models grouped by both week predictions and lactation groups 72
Figure A. 3 - Numerical variables distribution before preprocessing (1) 73
Figure A. 4 - Numerical variables distribution before preprocessing (2) 74
Figure A. 5 - Numerical variables distribution before preprocessing (3) 75
Figure A. 6 - Categorical variables frequency before preprocessing 76
Figure A. 7 - Variables distribution after preprocessing (1) 77
Figure A. 8 - Variables distribution after preprocessing (2) 78

LIST OF TABLES

Table 4.1- Available features in the initial dataset 18
Table 4.2 - Number of unique cows and registers in each farm 20
Table 4.3 - Number of registers, unique cows and farms within each lactation 21
Table 4.4 - Available features in the final timeseries dataset 22
Table 5.1 - Parameters' intervals and values used in XGBoost models for the one-week-ahead predictions for both lactation groups 28
Table 5.2 - Parameters' intervals and values used in XGBoost models for the two-weeks-ahead predictions for both lactation groups 28
Table 5.3 - Parameters' intervals and values used in ANN models for the one-week-ahead predictions for both lactation groups 29
Table 5.4 - Parameters' intervals and values used in ANN models for the two-weeks-aheadmpredictions for both lactation groups 29
Table 5.5 - Parameters settled for GP for both weeks and lactation groups predictions 29
Table 5.6 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using XGBoost model 31
Table 5.7 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using XGBoost model 32
Table 5.8 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using ANN model 33
Table 5.9 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using ANN model 34
Table 5.10 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using GP 35
Table 5.11 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using GP 36
Table 5.12 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using XGBoost models for one-week-ahead predictions. * indicates statistical significant differences. 37
Table 5.13 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using XGBoost models for one-week-ahead predictions. * indicates statistical significant differences 37
Table 5.14 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using ANN models for one-week-ahead predictions. * indicates statistical significant differences 38

Table 5.15 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using ANN models for one-week-ahead predictions. * indicates statistical significant differences 38
Table 5.16 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using GP for one-week-ahead predictions. * indicates statistical significant differences 39
Table 5.17 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using GP for one-week-ahead predictions. * indicates statistical significant differences 40
Table 5.18 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using XGBoost model 41
Table 5.19 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using ANN model 41
Table 5.20 - Median $R2$ and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using GP 42
Table 5.21 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using XGBoost model 43
Table 5.22 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using XGBoost model 44
Table 5.23 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using ANN model 45
Table 5.24 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using ANN model 46
Table 5.25 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using GP 47
Table 5.26 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using GP 48
Table 5.27 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using XGBoost for two-weeks-ahead predictions. * indicates statistical significant differences 49
Table 5.28 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using XGBoost for two-weeks-ahead predictions. * indicates statistical significant differences 49
Table 5.29 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using ANN for two-weeks-ahead predictions. * indicates statistical significant differences 50

Table 5.30 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using ANN for two-weeks-ahead predictions. * indicates statistical significant differences 50
Table 5.31 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using GP for two-weeks-ahead predictions. * indicates statistical significant differences..... 51
Table 5.32 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using GP for two-weeks-ahead predictions. * indicates statistical significant differences 51
Table 5.33 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using XGBoost model 53
Table 5.34 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using ANN model 54
Table 5.35 - Median $R2$ and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using GP 54
Table 5.36 - p-value of the Wilcoxon test between RMSE results from XGBoost modeling compared with ANN for the 28 datasets with $H0$: RMSE_xgb $=$ RMSE_ann vs $H1$: RMSE_xgb $-$ RMSE_ann < 0. * indicates statistical significant differences. 56
Table 5.37 - p-value of the Wilcoxon test between RMSE results from XGBoost modeling compared with GP for the 28 datasets with $H0$: RMSE_xgb $=$ RMSE_GP vs $H1$: RMSE_xgb $-$ RMSE_GP < 0. * indicates statistical significant differences. 56
Table 5.38 - p-value of the Wilcoxon test between RMSE results from ANN modeling compared with GP for the 28 datasets with $H0$: RMSE_ann $=$ RMSE_GP vs $H1$: RMSE_ann $-$ RMSE_GP < 0. * indicates statistical significant differences. 57
Table 5.39 - Number of times each feature was present in the top 5 of features that most contribute to the tree model 59
Table A. 1 - Detailed description of every feature in the initial dataset 67
Table A. 2 - Number of observations in each target week and lactation group considered ... 70

LIST OF ABBREVIATIONS AND ACRONYMS

AMS	Automated Milking System
XGBoost	eXtreme Gradient Boosting
ANNs	Artificial Neural Networks
GP	Genetic Programming
MY	Milk Yield
ML	Machine Learning
CART	Classification and Regression Trees
RF	Random Forest
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MLP	Multi-layer Perceptron
RHH	Ramped Half-and-Half
PLF	Precision Livestock Farming
DMY	Daily Milk Yield
MF	Milking Frequency
DLM	Dynamic Linear Model
SCC	Somatic Cells Count
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
NARX	Nonlinear Autoregressive Exogenous
DIM	Days In Milking
THI	Temperature-Humidity Index
TPOT	Tree-based Pipeline Optimization

1. INTRODUCTION

The dairy farming industry has been at the front row of technological advancements, aiming to balance the increasing demands for productivity with the essential need for animal welfare and reduction of environmental impacts. This dual imperative is driven by consumer awareness and regulatory pressures advocating for sustainable and ethical farming practices. Previous studies, such as Villettaz Robichaud et al. (2019) and Becker et al. (2020), have shown that enhancing animal welfare can positively impact productivity. However, challenges remain, as maximizing Milk Yield (MY) often leads to health issues in cows, such as clinical mastitis and lameness (Koeck et al., 2014). Thus, there is a crucial need for approaches that optimize productivity without compromising animals' welfare.

In this context, predictive modeling has emerged as a critical tool for improving decision-making processes in dairy farming. In order to optimize farm management practices, predictive models have been used in the past decades to forecast milk yield. Traditional approaches like the Wood's model (Wood, 1967) have been foundational in the milk industry, but may be limited by their theoretical assumptions (Jensen et al., 2018). Machine Learning (ML) approaches, e.g. decision trees and neural networks, have been used in the past few decades and offer more robust and flexible solutions, capable of handling complex and large datasets (Grzesiak et al., 2021; Fuentes et al., 2020).

Recent studies have demonstrated the effectiveness of decision tree-based methods, including CART, Random Forest (RF), and XGBoost, in predicting and optimizing dairy farm productivity (Piwczyński et al., 2020; Ji et al., 2022). Specifically, Ji et al. (2022) adopted a more detailed prediction approach, focusing on daily forecast rather than predicting an entire lactation's milk yield. With the methodology used by Ji et al. (2022), it is possible to address temporal sensitivity, enabling the identification of daily variations and patterns in milk production. Such detailed predictions are crucial for managing the high variability in daily milk yields and for making timely adjustments to farming practices.

This study focuses on predicting the milk yield for an entire week, specifically, for one and two- weeks-ahead rather than daily predictions. By predicting the average milk production for one-week and two-weeks-ahead, this study aims to provide reliable forecasts that account for the inherent temporal fluctuations in milk yield. To determine the amount of past information needed for accurate predictions, different lag periods are considered, from 1 day until 7 days before the prediction interval, further compared. It is hypothesized that the further the target variable is from the present, the more historical data will be required to make accurate predictions.

To create the panel data, where each individual can be represented in multiple points across time, the average values regarding temporal variables are calculated with the different lag days. Accounting as well for the different production patterns between primiparous and multiparous cows, these two lactation groups are differentiated.

Therefore, incorporating the seven different lag days, two lactation groups (primiparous and multiparous), and two target intervals (one-week and two-weeks-ahead), a total of 28 distinct datasets were analyzed, resulting in 28 models. These datasets are used to develop and tune predictive models, including XGBoost, Artificial Neural Networks (ANNs), and Genetic Programming (GP). XGBoost and ANN are chosen due to their proven effectiveness in previous studies for predicting milk yield (Ji et al., 2022; Fuentes et al., 2020), while GP is included as a different approach to explore its potential in this context. The performance of the three different modeling approaches, XGBoost, ANN, and GP, is then compared to determine which method produces the most accurate and reliable results for all 28 datasets.

Additionally, to evaluate the robustness of the models, predictions are performed using data from primiparous cows to predict the milk yield of multiparous cows. This cross-validation helps to understand the models' ability to generalize across different lactation groups.

Finally, the interpretability of the models in the GP approach is evaluated to identify the features that significantly contribute to milk yield predictions. This analysis allows the understanding of the underlying factors contributing to model's accuracy and provides actionable insights for improving dairy farming practices.

This study aims to significantly advance the field of predictive modeling in dairy farming by addressing key gaps identified in the literature. By focusing on broader prediction intervals and incorporating a detailed analysis of temporal sensitivity and lactation variability, the research seeks to develop robust models for forecasting milk production. Using data from multiple farms enhances the robustness of the analysis. This approach is expected to provide insights, facilitating more personalized, effective, and sustainable dairy cow management strategies. By detecting abnormal milk yield patterns within individual cows, farmers will be able to identify and treat diseases such as mastitis more effectively.

Ultimately, the findings will contribute to optimizing milk productivity and detecting patterns, thereby improving information on cow health (Elgersma et al., 2018) and aligning with the broader goals of modern agriculture.

In the following chapter it is presented an overview of the machine learning methods employed in the study. Chapter 3 discusses the use of predictive models previously used with varied approaches in the field of dairy farming. In chapter 4 the AMSs' data is described and the methodology is detailed. In chapter 5 the experimental study is addressed as well as its results and lastly, in chapter 6 the conclusions from the study are presented.

2. THEORETICAL BACKGROUND

The study incorporates three different predictive models to take advantage of each one's strengths as well as to explore their performance in this specific data and study since that, based on the No Free Lunch Theorem, the best model is problem-specific.

The 3 models used, XGBoost, ANN and GP will be detailed in the next points.

2.1. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network (ANN) is a ML method that tries to mimic the behavior of the human brain, having a set of neurons communicating with each other (Vanneschi & Silva, 2023, p. 161).

One example of a simple ANN is a perceptron, composed of one single layer. This layer can contain one or several neurons that are directly connected to each input through weights and directly connected to the output. Therefore, the flow of information in a perceptron is unidirectional, being also called *Feedforward ANN*.

In a perceptron, the input units are vectors composed of feature values. Each input unit corresponds to a distinct feature in the dataset, with the total number of features matching the number of input units. These input vectors are fed into the ANN where they connect to the neurons through links known as synapses, similar to connections in biological systems, and these synapses are defined by their weights. The influence of each input is determined by their associated weights.

Considering a single-layer perceptron, the number of weights equals the number of input features. This arrangement is illustrated in Figure 2.1. Therefore, each weight quantifies the "importance" of its corresponding input feature. Additionally, this neuron is linked to a bias unit with its value always set to 1 through a threshold, denoted by θ. The bias along with the threshold enables a better generalization on the data.

The output of the perceptron is then computed using an activation function, calculated as in equation *(1)*, where x_i represents each one of the *n* input vectors and w_i represents each one of the associated weights. This function processes the weighted sum of the inputs and the bias's threshold to produce the final output of the network.

$$y = f\left(\sum_{i=1}^{n} w_i x_i + \theta\right) \qquad (1)$$

During the training process, the perceptron adjusts the weights of its connections along with the bias's threshold using the perceptron learning rule, where is evaluated if the predicted output is smaller or bigger than the real one. If the output is not the same as expected, the weights, including the threshold, are readjusted according to the output calculated by the

perceptron. The magnitude of this change is dependent on the set learning rate, η, that can represent a constant or be dynamic. These changes are performed in order to optimize the model's performance.

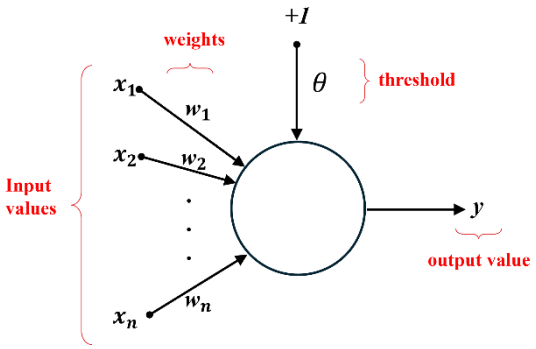

Figure 2.1 - Example of a single Neuron Perceptron (Source: Vanneschi & Silva, 2023, p.163).

Being a single-layer perceptron it is only able to capture linear relationships between the input features and the output variable. If the relationship in the data is nonlinear, the single-layer perceptron's simplicity becomes a limitation, not being able to model it accurately.

In cases where the underlying data relationships are complex and nonlinear, more sophisticated models are necessary. To correctly classify nonlinearly separable problems, multi-layer perceptrons (MLP) with non-linear activation functions are better suited.

In an MLP, the data is processed through multiple layers of interconnected neurons, having at least one layer of hidden neurons whose outputs are not directly connected to the output of the network. An MLP, with an example represented in Figure 2.2, is, as a single layer perceptron, a *Feedforward ANN*, where the neurons in each layer receive as input the outputs of the previous layers along with the bias unit (equal to 1) connected to its threshold. The information's forward flow results in the MLP's inputs propagation through the entire network to generate the output predictions. This implies the network to not contain any connections between neurons from the same layer, not having any cycles that can loop back to the same neuron or neurons in previous layers.

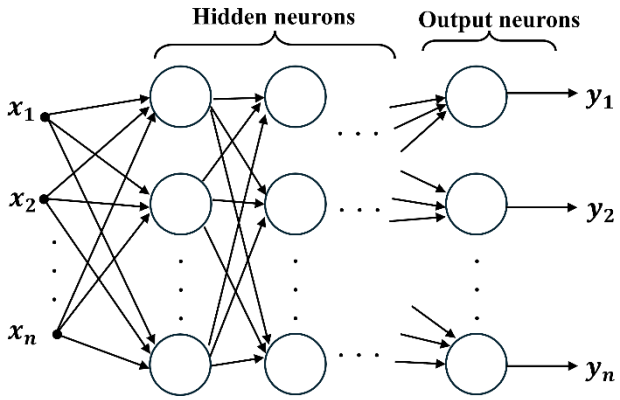

Figure 2.2 - Example of a Feedforward, multilayer ANN (Source: Vanneschi & Silva, 2023, p.176)

Feedforward MLPs are usually trained using the Backpropagation learning rule, which is an extension of the simpler Delta Rule, originally developed for single-layer ANNs. Understanding the Delta Rule (Haykin, 1998) is crucial to understanding Backpropagation, which applies to networks with multiple layers.

The *Delta Rule* is similar to the perceptron learning rule, but with a crucial difference. While the perceptron learning rule only evaluates if the calculated output is smaller or bigger than the target, the *Delta Rule* uses an error to readjust the weights accordingly. In the *Delta Rule* the output's calculation is performed by computing a weighted sum of the inputs, weighted using the weights of the synapses, and possibly followed by an activation function. Then the difference between the calculated output and the target is evaluated using an error metric that quantifies the distance between these values, such as the mean squared error or the absolute error. The calculated error is then used as the foundation for its learning process.

In the process of finding the vector of weights that minimizes the error function, a method called gradient descent is used (Haykin, 1998). The weight adjustment for a synapse can be expressed as a simple function of the calculated and expected outputs for a given observation, as well as the input entering the network through that synapse.

Backpropagation extends the Delta Rule to feedforward MLPs. In this method, the errors of the output neurons are propagated backwards to the hidden neurons. The errors of the hidden neurons are calculated as the sum of all the errors of the neurons in the subsequent layer to which it is directly connected. While in the output neurons, the error is calculated using the Delta Rule.

Therefore, the learning process in Backpropagation involves two main steps: the forward step and the backward step. During the forward step, inputs are propagated through the network to calculate the output. In the backward step, weights are adjusted to reduce the error by propagating the errors from the output layer back through the hidden layers.

Critical to the effectiveness of Backpropagation, therefore to the performance of the model, are various parameters such as the number of layers, the number of neurons in each layer,

and the learning rate. The number of layers and neurons determines the network's capacity to model complex data, taking the risk of over or underfitting the data if too few or too many are chosen. The learning rate controls the speed of learning of the ANN, with a small value of η possibly implicating a slow training and a too big value of η provoking an instability of the convergence during training. These parameters can be tuned to ensure the network's performance is optimized for the specific problem.

2.2. EXTREME GRADIENT BOOSTING

XGBoost, short for Extreme Gradient Boosting, is a decision tree-based model that has gained extreme popularity due to its great performance on a wide range of machine learning tasks. As in other decision tree-based algorithms, it works by combining different trees (sub-models) with each vertex representing a feature and each edge representing a possible value of that same feature. This boosting method uses gradient descent, regularization, and optimization for handling large-scale and sparse datasets, making it a robust solution for complex predictive modeling problems (Vanneschi & Silva, 2023).

Decision tree-based models are crucial algorithms of machine learning, particularly valued for their simplicity and interpretability. These models are trees that split the data into subsets based on the values of the input features. A decision tree starts at the root and makes a series of binary decisions to split the data at each node, the path from the root to a leaf node represents a set of rules for performing a prediction.

In regression trees, which are suitable for the problem addressed in this study, an impurity measure is used to quantify the quality of a split. The impurity measure must reflect the distance between the values predicted by the tree and the expected values, for example, the Mean Squared Error (MSE). In each terminal node of the decision tree, the target value is predicted by calculating the mean value of all the samples that fall into that node (Vanneschi & Silva, 2023, p. 158).

Despite their usefulness and ability to solve many problems, these simple decision trees have limitations. They can easily overfit to the training data, eventually having a high variance, meaning they may give different results if the training data is slightly changed. To overcome these issues, ensemble methods such as boosting are used, where multiple trees are combined to improve performance.

Boosting is an ensemble technique that combines the predictions of a sequence of simple (week) learning algorithms with different associated weights, Figure 2.3. By training these weak models sequentially, each one attempts to correct the errors of its previous algorithm, enabling each one to pay more attention to the predictive errors made by the previous models (Ferreira & Figueiredo, 2012; Vanneschi & Silva, 2023, p. 285). Therefore, the overall model improves iteratively as each new tree focuses on the mistakes made by the preceding trees.

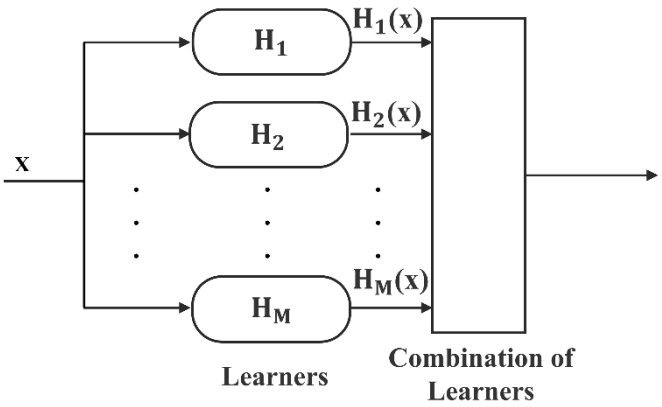

Figure 2.3 - Example of ensemble of classifiers (source: Ferreira & Figueiredo, 2012, p.2)

XGBoost is an implementation of the gradient boosting framework, where the objective function consists of two parts: a loss function and a regularization term that penalizes the complexity of the model while still capturing the necessary patterns in the data. This regularization term helps to prevent overfitting, ensuring that the model remains as simple as possible.

The objective function is presented in equation *(2)* with its two main components. The first term, a differentiable convex loss function, evaluates the difference between the prediction $\hat{y}_i$ and the target y_i. The second term, denoted as Ω, imposes a penalty on the complexity of the model, specifically the regression tree functions, the number of leaves and the size of the leaf weights. This additional regularization term plays a crucial role in smoothing the final learned weights, thereby helping to prevent overfitting (Chen & Guestrin, 2016).

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \qquad (2)$$

At each iteration, when a new tree is added to the ensemble, the gradient descent algorithm is implemented to minimize the loss. Essentially, the new tree attempts to correct the most significant errors made by the current ensemble. Gradient descent is used to update the model parameters by moving them in the direction that reduces the loss function the most.

Other important technique used in XGBoost is shrinkage, similar to a learning rate, it scales the weights of newly added trees by a factor η after each boosting step. This approach reduces the impact of each individual tree and allows room for subsequent trees to improve the model (Chen & Guestrin, 2016). To further enhance performance, XGBoost uses another method called tree pruning, where it allows the trees to grow greedily and then prunes back to the point where adding new nodes no longer improves the model significantly. This helps in keeping the trees simple also avoiding overfitting. In addition to these advantages, XGBoost can handle missing values due to its parallel processing, where it can learn the best way to handle missing values by training both routes of the split and selecting the best option.

Despite these strengths, XGBoost has some drawbacks. It may not perform well on small datasets, though this is not an issue for this study. Additionally, XGBoost has many hyperparameters that require careful tuning to optimize performance.

In summary, XGBoost builds on the foundation of decision trees and gradient boosting, introducing several enhancements that make it highly efficient. By sequentially building trees that correct the errors of the previous ones and employing advanced optimization techniques, XGBoost achieves great performances on a wide range of machine learning tasks. This makes it a popular choice for tackling complex problems with large and sparse datasets.

2.3. GENETIC PROGRAMMING

Genetic programming (GP) is an ML technique, more specifically an evolutionary algorithm inspired by Darwin's theory of evolution to find optimal solutions within a potential solution space (Vanneschi & Silva, 2023, pp. 205-206).

In GP, the main idea is to generate and evolve potential solutions (individuals). The algorithm starts by generating a population in which each individual represents a computer program on its own and the population undergoes evolutionary processes to identify the most effective solution to solve the specific problem at hand. At each evolutionary step, a new generation of solutions is created by offsprings that undergo one genetic operator, crossover, mutation, or replication. The offsprings are, then, selected according to their fitness, their ability to solve the problem.

The steps of GP are described below (Vanneschi & Silva, 2023, pp. 205-206):

1. Generate an initial population of computer programs (or individuals).
2. Iteratively perform the following steps until the termination criterion has been satisfied:
   2.1. Execute each program in the population and assign it a fitness value according to how well it solves the problem.
   2.2. Create a new population by applying the following operations:
      2.2.1. Probabilistically select a set of computer programs to be reproduced, based on their fitness (selection).
      2.2.2. Copy some of the selected individuals, without modifying them, into the new population (replication)
      2.2.3. Create new computer programs by genetically recombining randomly chosen parts of two selected individuals (crossover).
      2.2.4. Create new computer programs by substituting randomly chosen parts of some selected individuals with new randomly generated ones (mutation).
      2.2.5. Between the old and new programs, select the ones that will form the new population (survival).
3. The best computer program that appears in any generation is designated as the result of the GP process at that generation. This result may be a solution (or an approximate solution) to the problem.

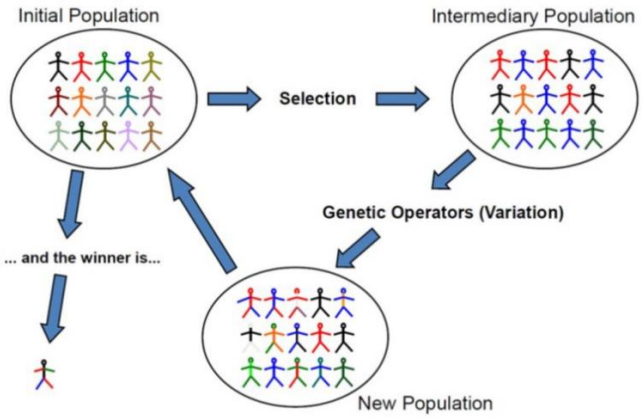

Figure 2.4 - Graphic description of an evolutionary algorithm, including GP, driven by the concepts of Darwin's Theory of Evolution. (Source: Leonardo Vanneschi, Genetic Programming: Introduction, Motivations and Applications. Lecture Notes. 2020 (cit. on pp. 9, 15))

In the previously described algorithm, the individuals are represented as trees. The tree is constructed using nodes, which can either be functions or terminal values. Nodes are selected from two sets: the Function Set (F) and the Terminal Set (T). Function nodes perform operations like addition, subtraction, multiplication, and division, as well as mathematical functions like sine, cosine, and logarithm. Terminal nodes can be constant values from a predefined set (Sakallioglu, 2023), or vectors of the dataset features. Although trees can have different structures and sizes, their grow is usually restricted by setting a maximum depth preventing the trees from growing indefinitely. The structure of the individual or tree is known as genotype while the effectiveness or fitness of the individual in solving the problem is known as phenotype.

The GP algorithm starts by creating trees to form its initial population and the individuals can be created using different initialization methods. The most common method is the Ramped Half-and-Half (RHH), first presented by Koza (1992), that is a combination of two other methods, the grow and full algorithm. In the grow technique the nodes can be either functions or terminals at each level, creating trees of varying shapes and sizes, and potentially less than the maximum depth. In contrast, the full technique, trees are grown to the maximum depth, filling each node with functions until the last level, which is filled with terminals. This results in trees that are uniform in depth. The RHH method considers greater population diversity. A maximum depth parameter is set, and the population is divided into two groups of trees with depths from 2 to the maximum initial depth parameter. For each depth group, half of the trees are initialized with the full technique and half with the grow technique.

The fitness of each individual measures its ability to solve the problem, with those possessing higher fitness being more likely to pass on their genes to future generations in the evolutionary process. For example, the fitness function of a GP implementation can be the Root Mean Squared Error (RMSE), where the lower its value the better is the solution able to solve the problem.

The selection process is where the parents are selected to undergo the genetic operator to produce offsprings. This process is done depending on the phenotype of the individual,

meaning its fitness, while genetic operators depend on the genotype. Various selection methods can be used in Genetic Programming (GP), such as fitness proportionate selection (roulette wheel), ranking selection, and tournament selection. In this study, tournament selection is used, which is explained in detail below.

Tournament selection is a method that combines randomness and determinism. To select a parent, a subset of the population, called a tournament, is randomly created with a predetermined size k. Within this subset, the fitness of each individual is evaluated, and the individual with the highest fitness is selected as a parent. This deterministic step ensures that the best individual in the subset is chosen. The tournament size, k, determines the selection pressure, which is the likelihood of the best individuals being chosen. A smaller tournament size results in lower selection pressure, allowing for more diversity as individuals with lower fitness may also be selected. A larger tournament size increases selection pressure, favoring the best individuals. While it might seem advantageous to always select the best individual, maintaining genetic diversity is crucial for the evolutionary process. By keeping the tournament size small, selection pressure is reduced, ensuring that less fit individuals are occasionally chosen to maintain diversity in the population.

The most used genetic operators are crossover and mutation, which, as mentioned before, are performed based on the trees' genotype. Crossover simulates a sexual recombination between two individuals (parents) that creates a variation in the population by producing an offspring. Each offspring inherits part of its genetic structure from one parent and part from the other, similar to how living organisms inherit chromosomes. Mutation is asexual as it only needs one parent to produce the offspring and it is usually performed by altering part of the tree (Sakallioglu, 2023).

The genetic operators used in the study are the standard crossover, also called swap crossover, and the standard mutation, also known as the subtree or point mutation. The swap crossover begins by, after selecting two random parents, selecting one random point, called the crossover point, at each parent. Then, the subtrees formed from those nodes until the terminal nodes are swapped between the two parents, creating two offsprings as in Figure 2.5. The point mutation begins by choosing a mutation point with a uniform probability. Then the subtree formed from that node until the terminal is removed and replaced with a new randomly generated subtree as in Figure 2.6. (Vanneschi & Silva, 2023, pp. 212-213)

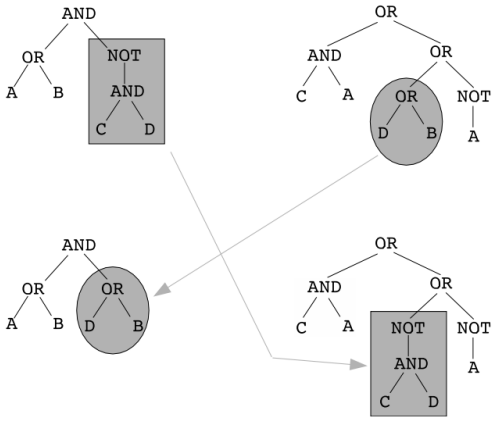

Figure 2.5 - An example of standard GP crossover (Source: Vanneschi & Silva, 2023, p.212)

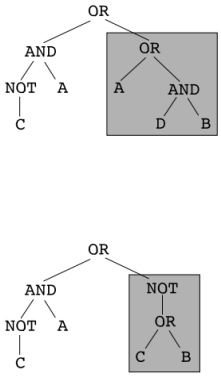

Figure 2.6 - An example of standard GP mutation (Source: Vanneschi & Silva, 2023, p.213)

Both genetic operators are performed based on a set probability, called crossover and mutation probability.

After generating new offsprings through the genetic operators, the previous population can either be entirely replaced by the new one, or a selection process can decide which individuals from both the old and new populations will survive. To ensure that the best individuals remain in the population through the evolution, it is common to use a strategy called elitism, where the top x individuals from both the existing and new populations remain unchanged to the next generation, ensuring their survival. This parameter x specifies the size of the elite group, setting it to 0 meaning not keeping any individual.

3. LITERATURE REVIEW

3.1. ANIMAL WELFARE AND PRODUCTIVITY IN MODERN DAIRY FARMING

Modern agriculture is facing an increasing urgency to address animal welfare, driven by consumers and governments. Consumers are becoming more involved in the origin and production process of foodstuffs. Corresponding to this, governments are pressuring sustainable farming practices and prioritizing ethical animal treatment (Dawkins, 2017; Jacobs & Siegford, 2012). Consequently, farms are being challenged to adjust their practices while maintaining productivity. Even though these objectives seem to be conflicting, there are studies demonstrating that animal welfare and productivity can be adjusted to grow together. For example, Villettaz Robichaud et al. (2019) discussed the potential associations between increased on-farm animal welfare and improved farm productivity, indicating that motivating farmers to change their practices to enhance animal welfare could result in an increased productivity. Within the same scope, Becker et al. (2020) showed that cows under stressful environments, such as heat stress, experience a significant decrease in Milk Yield (MY), milk quality, and fertility, which all lead to a decrease in profit for dairy producers.

In spite of this, overproduction is also a reality in dairy farming, when short-term heightening MY comes at the cost of the animals' welfare. The attempt to enhance milk productivity may inadvertently compromise the prevalence of health problems, including clinical mastitis, cystic ovaries, and lameness (Koeck et al., 2014).

In this scenario, it becomes crucial to rethink livestock production, addressing the increase in productivity as well as a better understanding of the factors that may also increase animal welfare, while decreasing the impact on the environment and maintaining business sustainability.

3.1.1. Precision Livestock Farming

Precision Livestock Farming (PLF) became, in the last few decades, a pivotal technology to enhance livestock management in modern farming by enabling the automated monitoring of animals on commercial farms. These systems can track individual animal behavior, including rumination, estrus, location, or rumen pH (Cockburn, 2020), facilitating continuous, fully automated data collection.

The data can be collected using powerful and affordable tools such as cameras, microphones, sensors (temperature sensors, skin conductivity sensors and glucose sensors) and other devices (Berckmans, 2014). Integrated with the data collection, analysis can then be translated into key indicators to promote the economic, social, and environmental sustainability of livestock farms (Vranken & Berckmans, 2017). These solutions can also significantly contribute to animal welfare improvement while enhancing farm productivity, as well as underpin new strategies for controlling livestock diseases.

Although there is increasing interest in PLF, the considerably high investment costs, coupled with the complexities of data analysis and interpretation can limit its adoption and acceptance (Fuertes, 2023). For example, (Bianchi et al., 2022) showed that the receptivity of adopting these new technologies in Italian dairy farms is still limited as farmers show reluctance to invest in precision systems. One of the prime PLF implementations was Automated Milking Systems (John et al., 2016), which provides Daily Milk Yield (DMY) data and has completely changed the milking paradigm.

3.1.2. Automated Milking Systems

Automated Milking Systems (AMSs) have emerged as a solution in contemporary farming influencing several dimensions on dairy cow health and welfare, including feeding practices, milking routines, and cow's monitoring. These systems promote gathering large amounts of data at an individual level at each milking event, with insights of physiological, behavioral, and productivity indicators in real-time (Masía et al., 2020; Tullo et al., 2019).

In Western Europe, during the 1970s, a significant shift occurred within the dairy industry as institutions began to develop automatic milking processes. Its implementation was a strategic response mainly driven by the increasing labor costs. The first experimental introduction of milking robots was conducted in 1986, and the first commercial installation of AMS was carried out in 1992 in the Netherlands by the company "Lely Industries N V". Since then, its adoption has been increasing significantly. In 2017, the number of operated AMSs amounted to more than 35,000 worldwide (Sharipov et al., 2021).

Automatic sensors collect diverse information about the cows, including health, milk production, reproductive status, and feed intake. With this approach, studies can be performed with new features that provide valuable information to dairy farmers while maintaining a non-invasive method of collecting data. AMS have the potential to increase milk production without depending on human labor, and simultaneously improve cow's welfare by allowing voluntary milking. Compared to cows milked in conventional methods, cows in AMS have more freedom to control their daily activities and have more opportunities to interact with their environment (Jacobs & Siegford, 2012).

AMS, alongside its associated management software, has proven to be valuable for managing dairy cow health through automated health-monitoring systems, which combine parameters including rumination time and physical activity to identify cows with metabolic and digestive disorders (Stangaferro et al., 2016). Preventing diseases such as mastitis, with milk quality and cow data, by detecting small changes within each individual cow, can also have a great impact on farms since it is one of the most important diseases in the dairy industry due to its subsequent reductions in milk production and quality (Naqvi et al., 2022).

3.2. PREDICTIVE MODELING IN DAIRY FARMING

Considering the fast technological transformation that occurred in the past few decades in dairy farms' management and operation, the development of predictive models that provide valuable information for decision-making processes on farms is a turning point. Predictive models can have multiple purposes such as the assessment of cow welfare and management efficiency, helping farmers make more informed business decisions. This is corroborated by Ji et al. (2022), who highlighted the importance of predicting DMY. This kind of modeling allows real-time identification of cows that diverge from the conventional production pattern due to health conditions, enabling advanced interventions. Similarly, the prediction of Milking Frequency (MF) emerges as a vital element in optimizing traffic system management in AMS (Ji et al., 2022).

The financial success of farms with AMS is highly related to MY, which may be affected by several other components, such as milking frequency and duration, number of cows and milk flow rate, type of barn, traffic type and the configuration and accessibility of the feeding zones (Piwczyński et al., 2020). Therefore, it is in farms' interest to integrate and invest in MY predictive models.

In the literature, mathematical models, such as the Wood's model (Wood, 1967), have been a great pillar when it comes to prediction within the dairy industry in conventional and automated milking systems. Jensen et al. (2018) developed a dynamic linear model (DLM), that implements a Wood's function, combined with a second-degree polynomial to forecast the cow's individual milk yields per milking, as they are milked in milking robots. The study aimed to predict the milk production of individual cows under the assumption of they having healthy udders based on the Somatic Cells Count (SCC) level. They were able to detect that the SCC level affected the standardized forecast errors of the DLM predictions, which suggests that the developed model could be used as part of a mastitis detection system, an important indicator of cow's health.

Regardless of the good performance of mathematical and statistical models, and of their relatively elementary implementation, they have a few disadvantages. For example, the restricted number of input features and the single dependency function across the entire dataset underlying theoretical assumptions (normality, linearity, data independence, etc.) that sometimes may not hold (Kominakis et al., 2002). Therefore, to overcome these limitations, Machine Learning (ML) approaches can be used to produce MY models.

Often, the results obtained from these mathematical and statistical models closely match with results from more complex models, such as ML models. Grzesiak et al. (2021) compared the effectiveness of three models in predicting lactational MY, mathematical and statistical, that include the Wood's model and the seasonal auto-regressive integrated moving average (SARIMA), and a ML model, the nonlinear autoregressive exogenous (NARX) artificial neural networks. They concluded that while SARIMA predictions were slightly more accurate in the

initial periods of lactation, the NARX and Wood's models were better at predicting the latter periods of lactation. Overall, there were no significant differences in predictive accuracy between the models. These results might be affected by the data collection method, which was based on test-day. The same conclusions may not hold in farms operated by AMSs since the milking process is very different and, consequently, the milk production can vary significantly. For example, in the former, milk production can be up to 65kg per day, while in the latter the daily milk production is lower. In a farm managed with conventional milking system, Wu et al. (2023) registered daily milk yields averaging 16.4kg in the morning and 15.3kg in the evening, with standard deviations of 5.6 and 5.4 kg, respectively.

While some studies may question the advantages of ML techniques due to their similar results compared with mathematical and statistical models (Grzesiak et al., 2021), ML is currently one of the most promising solutions in the research field of PLF. The adoption of ML in dairy farming can significantly enhance the sector's sustainability and efficiency due to its ability to recognize patterns in large amounts of data as in data collected by AMS. ML algorithms have already been applied to predict several cows' related variables such as fertility.

For example, Kliś et al. (2021) used Classification and Regression Trees (CART) algorithm to predict cow's fertility and the most important cows' characteristics related to it using data recorded during the periparturient period. MY is, as well, one of the most predicted features in this context.

Fuentes et al. (2020) leveraged Artificial Neural Networks (ANNs) to model milk productivity and quality (fat and protein) and feed intake. The study was performed on Holstein-Friesian cows in a farm managed with AMS in Australia. Variables incorporated included cow-specific data, such as days in milking (DIM), lactation number, milking frequency, as well as the environmental factors wind speed and direction and temperature-humidity index (THI). The research was conducted by creating two different predictive models, one for cows with similar heat tolerance and another independent of that condition. Both implemented models delivered high accuracy results in predicting all the key aspects of dairy production referred previously.

It is relevant to consider that it is not only important to address the prediction of MY, but it can also be beneficial to determine the factors that may have an impact on it. Therefore, despite the ability of ML techniques, like ANNs to predict milk productivity and quality accurately, their inherent lack of interpretability makes it challenging to understand which variables are influencing the predictions.

To tackle the lack of explainability of ANNs, decision trees-based approaches are a viable alternative. These offer a more interpretable alternative for analyzing inherent factors related to predictions and possibly providing actionable insights for farm management.

In the literature exist several studies that take advantage of the effectiveness of decision tree algorithms and their extensions in predicting and optimizing dairy farms productivity (Bovo et al., 2021; Kliś, Piwczyński, et al., 2021; Piwczyński et al., 2020). Classification and Regression Trees (CART), mentioned before, Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) are examples of decision trees algorithms that can offer insights into the decision-making process making them particularly valuable for farm management applications.

Piwczyński et al. (2020) provides a key understanding of how decision trees can be useful within the context of dairy farming managed with AMSs. Focusing on a population of nearly four thousand Holstein-Friesian cows, the study was able to determine how variables such as milking frequency, which was the one that showed the biggest influence, lactation number, month of milking, and the type of lying stall significantly impact DMY. The research demonstrated the capacity of decision trees to establish the most influential factors affecting milk production, providing guidance on how to optimize dairy operations.

Moved by the potential of decision trees, using the CART algorithm as well, for analyzing complex data, Kliś, Piwczyński, et al. (2021) explored the prediction of lactation MY of Polish Holstein–Friesian cows based on information recorded by the AMS in the periparturient period. It is important to note that the study's primary scope was to understand the key factors that influenced the MY, with the aim of improving dairy farm management using these insights and not to forecast future milk productivity. The periparturient period used was divided into the second and first week before calving and 1–4, 5–7, 8–14, 15–21 and 22–28 days of lactation. Their findings indicate that the most important factors responsible for predicting lactation yield were survival to the next calving, milking time per visit and number of milkings per day (22–28 days of lactation), milking speed (8–14 days of lactation), milking time per visit (5–7 days of lactation) and protein content of colostrum. This validates the decision tree's utility in identifying key variables that drive lactation performance.

Expanding the application of decision tree-based algorithms, Bovo et al. (2021) implemented a Random Forest-based model to assess how environmental conditions, specifically temperature-humidity index (THI), affect the DMY at a single cow level. The model used bagging to ensemble and combine several classification and regression trees, considered weak learners, reducing, this way, the variance and making it quite robust to outliers. By analyzing environmental conditions and their relationship with daily milk production the model was able to successfully identify heat stress impacts, detecting a drop in the cow's MY due to extreme hot conditions, providing a reliable tool that has the potential to develop strategies to improve decision support for farmers increasing both MY and animal welfare.

Still within the decision trees framework, Ji et al. (2022) showcase the application of XGBoost regression model to predict DMY, composition (fat and protein content) and XGBoost classification model for frequency of individual cow milking during the subsequent 28 days. XGBoost works by building an ensemble of decision trees in a sequential manner, where each

new tree corrects errors made by the previous ones. This approach, due to its daily implementation, led to having, for each of the four variables mentioned previously (MY, milk fat, milk protein and milking frequency), 28 XGBOOST models built to predict 28 days' values. Taking environmental data into consideration as well and with behavior, health, and productivity data from 80 cows in a robotic dairy farm, the models were able to accurately predict the dairy variables from the +1st to the +28th day. This study underscores the advanced capabilities of decision tree extensions in precision dairy farming, providing a robust model for future productivity.

In summary, decision tree-based methods, including CART, RF, and XGBoost, represent a significant advancement in the context of dairy farming operations. These studies collectively highlight the versatility, interpretability, and effectiveness of decision trees in addressing various challenges in dairy production, from optimizing MY to managing environmental stress impacts.

Considering the research progress, it is evident the importance of the accurate forecasting of dairy MY of individual cows and its benefits in dairy cow management. However, there is still room to improve the development and application of more robust ML methodologies (Ozella et al., 2023) and there are additional aspects that still need to be addressed. For example, milking characteristics may change as lactation progresses (Edwards et al., 2014) and MY can have strong seasonality (Fuentes et al., 2020), which should be considered when modeling milk productivity. Therefore, a more detailed prediction is particularly important to tackle the previous limitations. It is also important to address possible bias contemplating different farms and lactation cycles and potentially separate the analysis for primiparous and multiparous as their milking characteristics are substantially different, likely due to differences in cisternal capacity (Edwards et al., 2014).

Lined with the previous considerations, the scope of the thesis is to delve into these identified gaps and challenges. It aims to refine the predictive accuracy and identify influencing features of dairy cow MY integrating ML techniques. A more detailed prediction will undertake milking characteristics, data from multiple farms will enhance the robustness of data and a separate analysis will differentiate milk production of primiparous and multiparous cows. Therefore, it is expected that targeted insights can be identified leading to more personalized effective and sustainable dairy cow management strategies. Ultimately, it will contribute to milk productivity optimization and pattern detection enhancing, consequently, information about cow’s health (Elgersma et al., 2018).

4. METHODOLOGY

In this chapter, it is outlined the methodological approach used in this study, detailing each underlying step. It begins by describing the data obtained from the Automated Milking System (AMS), highlighting the transformations and preparations applied to create the input data for the models. Data exploration and preparation are crucial in studies related to forecasting, necessitating an in-depth understanding of the data and particularly in this study, its biological context. Following this, it is described the modeling approach and the metrics used for evaluation.

The study aims to forecast milk productivity and to understand the impact of historical data on predictions. Additionally, it seeks to identify the most important variables for forecasting, providing valuable insights for farmers. The steps involved in this process are detailed in the subsequent sections.

4.1. DATA EXPLORATION

The initial dataset contained 5074688 registers encompassing daily records covering 16 different farms in Italy between February of 2017 and August of 2023. Each entry represented a daily record of an individual cow with detailed information about its milking process. The 49 variables presented in the initial data are listed in Table 4.1. Variables prefixed with '**' were removed upfront based on two criteria:

- Variables with different metric specifications (avg, min, max) were reduced to only their 'avg' variants.
- Variables with a high number of null values were excluded.

Table 4.1- Available features in the initial dataset

Variables' names	
farm	maximum_milk_flow_duration_of_the_teats_avg_sec
animal_id	**maximum_milk_flow_duration_of_the_teats_max_sec
date_key_wall	**maximum_milk_flow_duration_of_the_teats_min_sec
lactation_number	duration_first_teat_con_till_last_teat_disconnected_avg_sec
days_in_lactation	**duration_first_teat_con_till_last_teat_disconnected_max_sec
animal_date_of_birth	**duration_first_teat_con_till_last_teat_disconnected_min_sec
pregnancy_length_in_days	maximum_milk_conductivity
animal_production_status	average_milk_speed_avg
animal_reproduction_status	**average_milk_speed_max
**average_body_weight_kg	**average_milk_speed_min
milk_production_kg	milk_yield_per_box_time_avg_kg_per_min
milkings	**milk_yield_per_box_time_max_kg_per_min

milking_failures	**milk_yield_per_box_time_min_kg_per_min
milking_refusals	**teat_milk_analysis_result_watery_flag
milk_fat_pct	**teat_milk_analysis_result_mastitis_flag
milk_lactose_pct	**teat_milk_analysis_result_blood_flag
milk_protein_pct	**teat_milk_analysis_result_colostrum_flag
**milk_somatic_cells_count_kcells_per_ml	**teat_milk_analysis_result_abnormal_flag
milking_box_time_avg_sec	**animal_sick_chance
**milking_box_time_max_sec	**eating_minutes
**milking_box_time_min_sec	**isk_individual_standardized_milk_day_production
activity_avg	**animal_max_heat_chance
**activity_max	**milking_visit_time_since_last_successful_milking_max_sec
**activity_min	**milking_visit_time_since_last_successful_milking_min_sec
rumination_minutes	

The initial distribution of each variable is illustrated in the annexes from Figure A. 3 to Figure A. 6. These figures include boxplots for numerical variables and barplots for categorical variables. Both date type variables, 'date_key_wall' and 'animal_date_of_birth,' and the id of each cow, 'animal_id', were excluded from these plots.

Having analyzed the available features and established their significance within the context of the dairy industry, it was possible to move forward to the phase of data preparation.

4.2. DATA PREPARATION

The data preparation phase covers all the steps involved in the cleaning and composition of the final dataset. This stage is one of the most important as it is where the raw data is transformed to ensure its quality when fed into the models. Its quality directly influences the accuracy and reliability of the outputs.

4.2.1. Daily registers Dataset

From the original data, as mentioned before, twenty-five of the existing variables, those prefixed with '**' in Table 4.1, were excluded upfront. Then, a unique cow id was created across all the dataset since the original one referred to the cow within each farm.

4.2.1.1. Data filtering

All data regarding the DryOff, the period where cows are not lactating, was excluded, as for the purpose of the study only data where cows are lactating is considered, consequently the feature "animal_production_status" was deleted. Furthermore, to ensure that the lactation records are complete, lactations not starting on day 0 or 1 were deleted. This ensures that

partial records of lactations are not included, assuring that only complete and consistent lactation cycles are considered for analysis.

Lactations greater than five were excluded from the dataset, which is supported by the number of observations being usually reduced for higher lactation numbers, as fewer cows reach it. By limiting the dataset to the first five lactations, the study ensures, as well, more direct comparisons with the literature as frequently only the first few lactation cycles are considered (Marumo et al., 2022).

As seen before, the relationship between milk production levels and cow health is also very relevant, having abnormally low values of milk yield indicating possible health or stress issues (Becker et al., 2020), or even recording errors. Thus, only registers with a daily milk production greater than 10kg/day were kept in the dataset. Records with null values for milk production were also removed upfront since that is the target variable for this study's analysis.

After removing the previous observations, the dataset ended up having some day gaps within each cows' lactations which meant not having the whole lactation data. To ensure the integrity of the lactations, were only kept the ones without any day gaps and that started on day 1. Excluding day 0 removes all possible incomplete records, as one cow can have been recorded since the beginning of the day and other could have only entered the registration at the end. Following this logic, the last register of each cows' lactation was also removed.

Lastly, all cows' lactations containing any null value were deleted, not having, this way, to input data. These filtering criteria ensures that the data used for analysis is reliable, relevant, and consistent.

The final dataset is composed of 736866 registers reflecting the daily records of each cow. These records are subdivided into 3397 groups of unique (cow id, lactation number) combinations. Within this dataset there are 2701 unique cows from 14 different farms, still ensuring data from multiple farms enhancing its robustness. Nevertheless, Table 4.2 reveals that some farms have a significantly larger number of cows and registrations compared to others.

Table 4.2 - Number of unique cows and registers in each farm

Farm_id	unique_cows	registers
1	170	41510
6	83	21167
7	17	2006
10	467	85756
19	104	26155

31	114	16595
33	271	66977
37	75	18330
38	333	104451
39	110	34478
40	198	69384
42	298	110012
45	111	22971
52	350	117074
Total	2701	736866

Although there are fewer registers and cows in some farms, Table 4.3 shows that all 14 farms provide information across all lactation numbers. Table 4.3 also provides insights on the number of registers and unique cows, within each lactation number. As the lactation number increases, the number of cows reaching the higher lactation stages decline, going from 1024 in lactation 1 to 185 in lactation 5. Despite this, lactation number 2 has a highest number of registers compared with the first lactation.

Table 4.3 - Number of registers, unique cows and farms within each lactation

lactation_number	unique_cows	registers	unique_farms
1	1024	229400	14
2	1077	238749	14
3	692	144326	14
4	419	89095	14
5	185	35296	14

A more detailed analysis on the final time-series data is present in Figure A. 7 and Figure A. 8, including the additional variables created that are discussed in the following chapter. The distribution of each variable is represented by a boxplot being the variable 'farm' not represented as it will not be used to the prediction models as well as the 'date_key_wall' and 'cow_id' features.

4.2.1.2. Feature engineering

To provide the models with more and possibly useful information, three features were created from the original ones. This process can significantly improve performance by transforming raw data into informative features.

The first feature created was the cow age in months, transforming date values from the cows' date of birth into a continuous standardized variable. The cow age may have a great impact

on the milk yield's prediction, as different stages of cows' life are associated with distinct production characteristics (Turiello et al., 2020).

The second feature created was the delivery age in months, associated with each cow's lactation. As the previous feature created, this one can have an impact on milk yield as even the last days' milk production in a specific lactation can be influenced by the age of the cow when giving birth.

The third and last one was the pregnancy length, that was set to 0 if the cow is not pregnant. The reproduction status can also have an impact on the milk produced since it affects the cow's hormones and consequently their ability to produce milk. As Penasa et al. (2016) demonstrated, the milk yield of pregnant cows can differ from nonpregnant ones from the fourth month of pregnancy as milk production decreases significantly.

4.2.2. Panel Dataset

Having all data cleaned with complete lactation cycles for several different cows, the dataset was transformed from a time series type of data into a format that could be analyzed as a regression problem. A panel dataset was created containing observations on cows' data across the different time periods. The 23 features in the timeseries dataset are presented in Table 4.4.

Table 4.4 - Available features in the final timeseries dataset

Variables' names	
farm	milking_box_time_avg_sec
cow_id	maximum_milk_flow_duration_of_the_teats_avg_sec
date_key_wall	duration_first_teat_con_till_last_teat_disconnected_avg_sec
§ lactation_number	maximum_milk_conductivity
§ days_in_lactation	average_milk_speed_avg
milk_production_kg	milk_yield_per_box_time_avg_kg_per_min
milkings	rumination_minutes
milking_failures	activity_avg
milking_refusals	§ cow_age_months
milk_fat_pct	§ delivery_age_months
milk_lactose_pct	§ pregnancy_length
milk_protein_pct	

The data was transformed focusing on the prediction of the average milk production of the week immediately following the training period as well as for two weeks after the training

data period. The previous choice was settled upon the results from Ji et al. (2022), where the decrease in milk yield's predictions become more evident from the fifteenth day onward.

Different datasets were created categorizing data by lactation cycle due to known variations in milk production across different lactations. For instance, data from all cows in their first lactation were compiled into one dataset and data from the subsequent lactations were compiled into other separate dataset, having the latter one more feature regarding the lactation number.

For each lactation group considered, subsets based on different temporal windows (1 week, 6 days, 5 days, etc.) were created on the features present in the dataset. Features like milk yield and number of milkings were managed as temporal data, having its average calculated based on the corresponding number of previous days considered. Features such as lactation number and delivery age months don't have a temporal component, having only one entry representing them despite the number of previous days considered for the prediction. Regardless of the temporal component in features like days in lactation and cow age that follow a fixed sequence overtime, they are treated as non-temporal features, having only one register independently of the number of previous days. The features in Table 4.4 prefixed with '§' represent the ones with one single data point entry independently of the number of lag data considered.

And with this approach, two different target variables were created, the immediately following week's milk production average and the average milk production of two-weeks-ahead. By exploring several different temporal windows, it is possible to get insights on the time frame needed to accurately predict each target variable.

Figure 4.1 and Figure 4.2 contain tables illustrating the process of assembling the datasets regarding multiparous cows' data. The target variable used in the example is the average milk production of the immediately following week. To facilitate the interpretation of the example, a 3-day lag was used and only 4 features were considered, 'milk_production_kg', 'milkings', 'lactation_number' and 'days_in_lactation'.

In each figure, the top table represents the time series data, while the lower table shows the corresponding panel data, having in Figure 4.1 the creation of one entry and in Figure 4.2 the subsequent one. In this approach, each target value is calculated within each cow_id and lactation_number, with a longer a cow's lactation period, corresponding to more records in the dataset.

For the panel dataset, the average values of 'milk_production_kg' and 'milkings' over the previous 3 days were calculated, while for 'lactation_number' and 'days_in_lactation' only the value of the previous day was considered, as previously described.

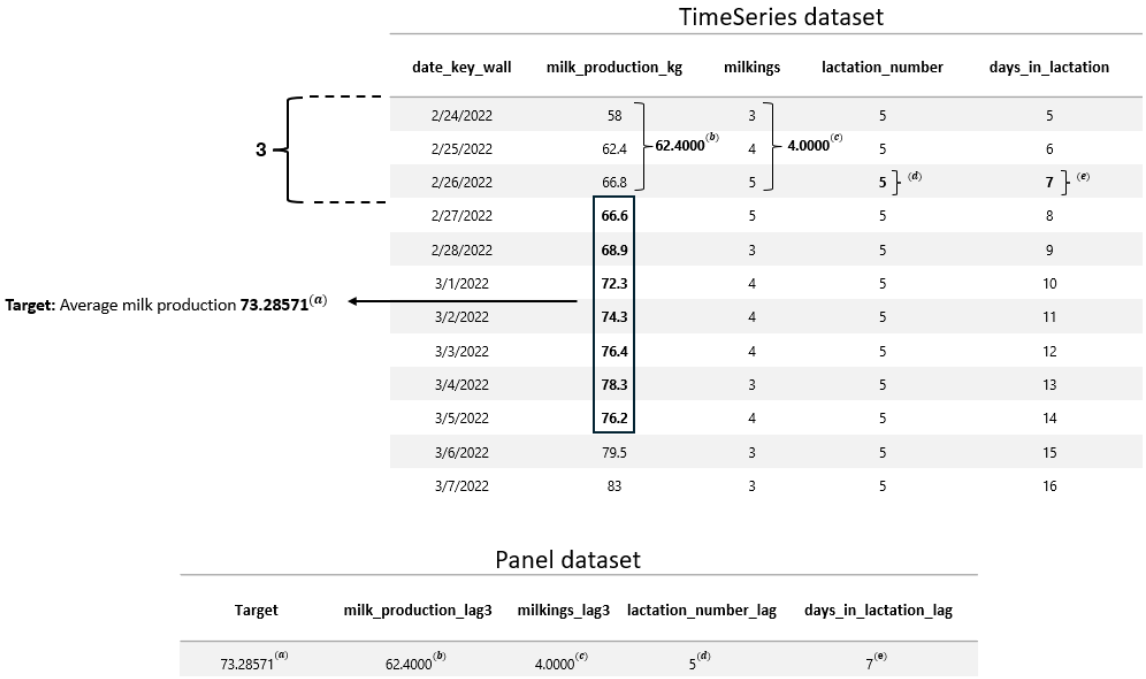

Figure 4.1 - Example of the creation of a register in the panel dataset using 3 lag days

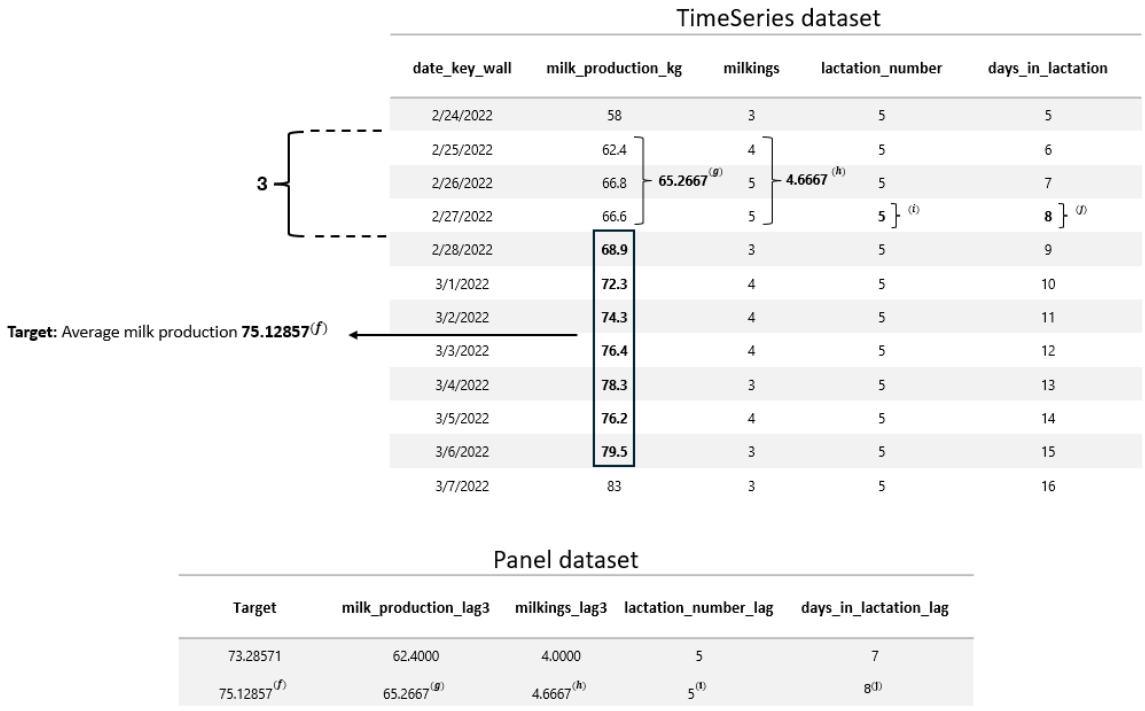

Figure 4.2 - Example of the creation of the subsequent register in the panel dataset using 3 lag days

Having this, were created 28 different datasets regarding the 2 different lactation groups, primiparous and multiparous, the 2 target features, one-week and two-weeks-ahead average milk production and the 7 different number of lag days considered. In each of the 28 datasets 20 features are present in the multiparous cows' data, all present in Table 4.4, except the date, cow_id and farm. In the primiparous cows' data are present 19 features since the lactation_number is constant throughout the datasets. The number of registers in each target week and lactation group considered across all the 7 lag days is detailed in the annexes in Table A. 2.

4.3. MODELING

This phase of the study addresses the development and tuning of various predictive models to forecast milk production. Reflecting the identified gaps in the literature review, the models will assess the impact of temporal sensitivity and lactation cycles on milk production acknowledging the variability in milk yield across different lactation stages and the importance of temporal precision in predictions.

4.3.1. Temporal and Lactation Modeling

To capture the temporal dynamics of milk production, as mentioned before, different models were developed for different time windows: 1 week, 6 days, 5 days, 4 days, 3 days, 2 days and 1 day prior to the first target week. Thus, each model was applied to all 28 different datasets. This approach permits the identification of how many days can influence the next week and next two-weeks average milk yield, providing insights about temporal sensitivity.

Having individual models for each lactation group aligns with the literature review, where it is clear the importance of different lactation cycles, being the main difference between the first and the remaining lactations. This approach not only allows the capturing of distinct production patterns between the first and subsequent lactation cycles but also to validate the similarity between the production in more advanced lactation cycles.

4.3.2. Models

The study incorporated some different predictive models to take advantage of each one's strengths as well as to explore their performance in this specific data and study since that, based on the No Free Lunch Theorem, the best model is problem specific.

4.3.2.1. XGBoost

In the literature, as mentioned before, the effectiveness of tree-based models like XGBoost at predicting milk yield has been demonstrated, therefore it has a great potential to provide valuable results.

4.3.2.2. Artificial Neural Networks

Due to the capabilities of ANNs to capture complex patterns in data in several different fields, including non-linear relationships in dairy data (Fuentes et al., 2020), it was a crucial ML model to implement with the data accordingly prepared.

4.3.2.3. Genetic Programming

Genetic programming (GP) provides a different approach from the other ones presented by enabling the search across a wide range of possible solutions to find the optimal one.

4.4. EVALUATION

To evaluate and compare the models, different aspects were assessed. Different metrics, Coefficient of determination (R^2) and Root Mean Squared Error (RMSE), were used to evaluate each model's performance.

The R^2 measures the proportion of the total variance that is explained by the model and being a standardized metric allows to easily compare the performance of different models. An R^2 value of 1 indicates a good fit of the model while a value close to 0 indicates that it fails to accurately predict the target variable. It is calculated using the following formula:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (3)$$

The RMSE measure provides information on the magnitude of the error in the model's predictions. It is expressed in the same unit as the target variable, being useful to detect large errors and their weight. The formula for calculating the RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (4)$$

Using the two metrics enabled the different models to be evaluated thoroughly and the comparison of their performance across various time windows and lactation cycles.

Cross validation, in particular K-fold, was used to enhance more reliable and general results of the performance of models. By averaging the performance metrics from the different train-test splits K-fold minimizes the selection bias from only one trial of the split.

5. EXPERIMENTAL STUDY

In this chapter are presented the settings and results regarding the predictive modeling for milk yield using three distinct models, XGBoost, ANN and GP.

Twenty-eight different predictive models were performed for each model, representing 28 unique datasets. The different datasets account for varying lag periods (1 to 7), different lactation periods (primiparous and multiparous), as well as the one-week-ahead and two-weeks-ahead prediction intervals.

The chapter begins with the description of the used settings in each approach. It is followed by the one-week-ahead prediction outcomes for each one of the three models, along with a comparative analysis of the performance across the different lag intervals. Following this, the same analytical process was replicated for the two-weeks-ahead predictions. In addition, a comparison among the three performed models is presented.

5.1. EXPERIMENTAL SETTINGS

For each one of the three approaches described below, a prior split was performed where 20% of each dataset was used to perform the tests. Of the remaining 80%, 20% was used for validation and 80% for training.

5.1.1. XGBoost setup

To implement the XGBoost model, the XGBRegressor from the xgboost library in Python was utilized and the models were applied to each of the 28 datasets. Initially, a randomized search with 30 iterations was conducted, focusing on the negative mean squared error to identify the optimal parameters. Following this, a grid search was performed using narrower intervals for the parameters identified in the randomized search. Specifically, the parameter intervals for alpha and learning_rate were refined based on the values from the randomized search, adjusted to [0.75*parameter, parameter, 1.25*parameter]. The other parameters were used directly from the randomized search to avoid excessive computational demands. The grid search was executed with 5 different splits, where 80% of the set was used for training and 20% for validation. All the non-referred parameters were used as the default ones.

The results of both the randomized and grid searches are detailed in Table 5.1 and Table 5.2, which include the parameters searched, the intervals used for the randomized search, and the results for each of the 28 different models. The results are divided into two different tables where in Table 5.1 are the ones regarding one-week-ahead predictions and in Table 5.2 are the ones from two-weeks-ahead predictions.

Table 5.1 - Parameters' intervals and values used in XGBoost models for the one-week-ahead predictions for both lactation groups

	Intervals	Lactation_1							Lactation_2_3_4_5						
		Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
alpha	[2,5,10]	0.375	0.625	0.1	0.125	0.075	0.075	0.375	0.075	0.375	0.075	0.625	0.1	0.125	0.625
learning_rate	[20,200,500]	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
max_depth	[0,.001,0.01,0.1]	5	5	5	5	5	5	5	5	5	5	5	5	5	5
n_estimators	[0.1,0.5,0.9]	200	200	200	200	200	200	200	200	200	200	200	200	200	200

Table 5.2 - Parameters' intervals and values used in XGBoost models for the two-weeks-ahead predictions for both lactation groups

	Intervals	Lactation_1							Lactation_2_3_4_5						
		Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
alpha	[2,5,10]	0.625	0.625	0.1	0.625	0.125	0.075	0.125	0.375	0.625	0.1	0.1	0.1	0.625	0.125
learning_rate	[20,200,500]	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
max_depth	[0,.001,0.01,0.1]	5	5	5	5	5	5	5	5	5	5	5	5	5	5
n_estimators	[0.1,0.5,0.9]	200	200	200	200	200	200	200	200	200	200	200	200	200	200

5.1.2. ANN setup

The same approach applied to XGBoost was also used for the ANN models. The MLPRegressor from the sklearn.neural_network library was employed to perform predictions across all 28 subsets. Initially, as previously explained, a randomized search with 30 iterations was conducted, focusing on the negative mean squared error to identify the optimal parameters. Subsequently, a grid search was performed with narrower parameter intervals based on the results from the randomized search. Specifically, the intervals for alpha and learning_rate_init were refined using the interval [0.75*parameter, parameter, 1.25*parameter]. The grid search was executed, as well, using 5 different splits, with 80% of the set allocated for training and 20% for validation.

The other parameter identified in the randomized search, hidden_layer_sizes, was directly utilized in the grid search directly from the randomized seach. Additional parameters were manually configured: the activation function was set to 'relu', the solver to 'adam', and the maximum number of iterations (max_iter) to 100. All other parameters were left in their default settings.

In Table 5.3 and Table 5.4 are presented the result parameters from both sequentially performed search methods for all the 28 subsets, as well as the intervals used for each

parameter in the randomized search. In Table 5.3 are the results from one-week-ahead predictions and in Table 5.4 are the results from the two-weeks-ahead predictions.

Table 5.3 - Parameters' intervals and values used in ANN models for the one-week-ahead predictions for both lactation groups

	Intervals	Lactation_1							Lactation_2_3_4_5						
		Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
alpha	[0.001, 0.01]	0.00075	0.001	0.00125	0.00075	0.00075	0.00075	0.001	0.00075	0.00125	0.001	0.00125	0.00075	0.00075	0.00075
hidden_layer_sizes	[(100,), (100,50)]	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)
learning_rate_init	[0.01, 0.1]	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 5.4 - Parameters' intervals and values used in ANN models for the two-weeks-aheadmpredictions for both lactation groups

	Intervals	Lactation_1							Lactation_2_3_4_5						
		Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
alpha	[0.001, 0.01]	0.0075	0.00075	0.0075	0.00125	0.001	0.001	0.00075	0.00075	0.01	0.001	0.001	0.00125	0.001	0.001
hidden_layer_sizes	[(100,), (100,50)]	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)	(100,50)
learning_rate_init	[0.01, 0.1]	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

5.1.3. GP setup

To determine the optimal parameters for the evolution of GP, a few tests were conducted with either the mutation or crossover probabilities set to 1, 0 and 0.5. These experiments revealed a fast convergence to an optimal point. Consequently, and having the fast convergence into consideration, the chosen parameters are detailed in Table 5.5.

Table 5.5 - Parameters settled for GP for both weeks and lactation groups predictions

Parameter	Value
population size	40
number of generations	50
selection pressure	0.05
mutation probability	0.5
crossover probability	0.5
elitism	True

The population evolution using these parameters is shown in the annexes in Figure A. 1, covering the four different groups for each week predicted and both lactation groups. The plots display the median RMSE value across 30 runs, along with the 15th and 85th quantiles over the 50 generations. It is evident that most groups achieve a fast convergence to an optimal point, often within the first 10 generations.

5.2. ONE WEEK AHEAD MILK YIELD PREDICTION

5.2.1. XGBoost

For the one-week-ahead milk yield prediction using the XGBoost model, the performance metrics for lactation 1 are summarized in Table 5.6. The table presents the median metrics from a series of 30 modeling runs from different random splits. After the best models' hyperparameters for each dataset were found using 80% of each dataset, with the remaining 20% a repeated k-fold cross-validation with 5 splits was repeated 6 times to estimate the prediction and generalization performances of the model.

The improvement of R^2 throughout the different lags for training data is noticeable. Considering only the information about the previous day, lag 1, the R^2 is 0.9004 while considering information about the previous week, lag 7, the R^2 is 0.9183. The same happens for test data where the R^2 improved from 0.8935 for lag 1 to 0.9131 for lag 7. The RMSE scores decreased from 2.1421 to 1.9401 in training data and from 2.2161 to 2.0000 in test data. Although generally the training performance is slightly better than on test, the proximity between them indicates that the models can generalize the patterns rather than merely memorizing the training data.

Additionally, it is evident that both metrics indicate a progressively stronger predictive capability with the addition of more lag information. However, it is also evident that with each additional lag there is a slower improvement in the models' performance. This suggests that while adding more historical data continues to enhance model performance, the rate of improvement slows down, indicating that as we move further back in time, the data may become less relevant to the current prediction, or the incremental information gained from adding another lag becomes less significant compared to the most recent data.

Table 5.6 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using XGBoost model

	R^2 Lactation_1 train	R^2 Lactation_1 test	RMSE Lactation_1 train	RMSE Lactation_1 test
Lag_1	0.900380	0.893538	2.142133	2.216111
Lag_2	0.909092	0.902770	2.045058	2.118398
Lag_3	0.912956	0.907196	2.001052	2.068227
Lag_4	0.915340	0.909712	1.974913	2.038265
Lag_5	0.916720	0.911330	1.957689	2.019317
Lag_6	0.917755	0.912364	1.946160	2.007129
Lag_7	0.918261	0.912992	1.940100	1.999977

In predictions for lactations 2, 3, 4, and 5, it is also possible to notice an improvement with increasing lag days (Table 5.7). The test R^2 scores went from 0.9365 at lag 1 to 0.9471 at lag 7 in training data while in test data went from 0.9336 to 0.9448. The RMSE test scores showed a corresponding reduction, decreasing from 2.8965 to 2.6434 in training data and from 2.9624 to 2.6972 in test data, indicating robustness as well across different lactation groups when provided with more historical context.

The results from the combined data of lactations 2 to 5 also indicate a great ability to generalize, notably better than those from primiparous cows' data. This probably happens due to a more varied dataset providing the model with more patterns to learn from. When considering lactations 2, 3, 4 and 5, there are more diverse patterns present in data as well as more registers, allowing the XGboost algorithm to build its decision trees with a wider range of patterns, leading to its robustness.

As in the results regarding the first lactation, although there is a continuous improvement in the modes' performances, there is a decrease in the improvement rate as it is added more lag data. In the case of multiparous cows' data, this rate of improvement is even less pronounced when compared to the first lactation data, possibly due to its better starting performance.

Table 5.7 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using XGBoost model

	R^2 Lactation_2_3_4_5 train	R^2 Lactation_2_3_4_5 test	RMSE Lactation_2_3_4_5 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.936470	0.933589	2.896464	2.962430
Lag_2	0.942439	0.939859	2.757402	2.820201
Lag_3	0.944838	0.942467	2.697749	2.755516
Lag_4	0.946099	0.943629	2.666996	2.725874
Lag_5	0.946685	0.944354	2.652805	2.710220
Lag_6	0.947032	0.944793	2.644344	2.700221
Lag_7	0.947057	0.944822	2.643382	2.697203

5.2.2. ANN

The ANN model for lactation 1, as seen in Table 5.8, uses the same test setup as the XGBoost model. Employing a repeated k-fold strategy, there is an increased R^2 training and test scores from 0.8939 to 0.9146 and from 0.8878 to 0.9108 respectively. This progression throughout the lags demonstrates the ANN model's capability to enhance its predictions as more historical data is incorporated, accounting as well with a diminishing rate of improvement. RMSE training and test scores similarly decreased from 2.2101 at lag 1 to 1.9826 at lag 7 in training data and from 2.2726 to 2.0263 in test data, consistent with previous results that indicate that the models become more precise with extended lag days.

The results also indicate a great generalization ability of the models, considering the metrics' proximity between training and test. It also suggests a difference between the ANN and the XGboost model's performance with data from the first lactation. The results appear to be slightly worse when using ANN. This discrepancy can be related to the more constrained search space during the hyperparameter tuning of the ANN, particularly the multi-layer perceptron (MLP), that was set smaller due to the high computational demand and consequently the extend time for model training.

Table 5.8 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using ANN model

	R^2 Lactation_1 train	R^2 Lactation_1 test	RMSE Lactation_1 train	RMSE Lactation_1 test
Lag_1	0.893870	0.887831	2.210078	2.272609
Lag_2	0.902530	0.897531	2.117875	2.169354
Lag_3	0.907522	0.903773	2.064141	2.107280
Lag_4	0.911143	0.906727	2.021790	2.071723
Lag_5	0.913107	0.908579	1.999538	2.060303
Lag_6	0.913644	0.909718	1.992896	2.040851
Lag_7	0.914567	0.910781	1.982624	2.026285

For the combined lactation groups 2, 3, 4, and 5, Table 5.9 shows that the ANN model's training and test R^2 scores did not always improve with the increase in the number of lags. For training data, the R^2 improved from 0.9272 at lag 1 to 0.9416 at lag 3 and then slightly decreased in lag 4 to 0.9414. Then, it increased until lag 7, having the highest R^2 for training of 0.9443. In test data, there is a similar behavior, decreasing the performance from lag 3 to lag 4. The best result was also obtained in the 7th lag, with an R^2 of 0.9426. The RMSE scores follow the same pattern as in the R^2 metric, increasing the error from lag 3 to lag 4. The best results are observed in lag 7 where the RMSE score is 2.7102 for training data and 2.7470 for test.

It is noticeable that the rate of improvement in this set of lactation groups appears to stagnate slightly quicker than in the case of ANN predictions for the first lactation. This might be indicative that, as the model begins with higher performance metrics, the room for further enhancement is reduced compared to the first lactation results.

When comparing the ANN model's performance to the XGBoost for data from multiparous cows, it is observed that the ANN model initially starts with worse predictions with a smaller number of lag days. However, as more historical data is added, more specifically considering the past 7 days, the ANN model converges closely with the XGBoost model's performance. This is despite a noticeable decrease at lag 4, after which the ANN model's performance increases again. The similarity in performance at lag 7 suggests that the ANN model effectively compensates for its initially smaller results by leveraging the broader context provided by the additional lag data.

Table 5.9 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using ANN model

	R^2 Lactation_2_3_4_5 train	R^2 Lactation_2_3_4_5 test	RMSE Lactation_2_3_4_5 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.927222	0.926206	3.101166	3.113784
Lag_2	0.938464	0.937065	2.849210	2.878183
Lag_3	0.941597	0.939925	2.775626	2.812992
Lag_4	0.941408	0.939791	2.780109	2.818732
Lag_5	0.943860	0.941957	2.722435	2.775821
Lag_6	0.944021	0.942181	2.717720	2.765053
Lag_7	0.944303	0.942609	2.710176	2.747034

5.2.3. GP

For one-week-ahead average milk predictions using GP, the results from primiparous data are presented in Table 5.10 using the same cross validation strategy mentioned previously. It is showcased an increase of the R^2 value for training data starting at 0.8584 for lag 1 and reaching a peak at 0.8719 for lag 3. However, this peak is followed by a slight decline within the further lag information, with R^2 values dropping to 0.8562 by lag 7. The test data values follow a similar pattern, starting at 0.8589 for lag 1, peaking at 0.8723 for lag 3, and then declining to 0.8567 by lag 7. The RMSE values support these observations. For training data, RMSE decreases from 2.5448 at lag 1 to 2.4223 at lag 3, indicating improved model accuracy. However, beyond lag 3, RMSE values increase again, reaching 2.5675 at lag 7. The test data shows a similar trend, with RMSE decreasing from 2.5500 at lag 1 to 2.4298 at lag 3, before increasing to 2.5642 at lag 7.

The results suggest that while the GP model initially benefits from the inclusion of additional lag data, its performance does not consistently improve beyond a certain point. Additionally, the metrics' proximity between training and test sets demonstrates the GP model's strong generalization ability, indicating it performs similarly on unseen data.

However, it is noticeable that the GP model generally exhibits smaller R^2 values and larger RMSE values compared to the ANN and XGBoost models for primiparous cows' data. The smaller prediction capacity as well as the decrease from the third lag further might be related to the simplicity of the genetic operations used (like mutation and crossover), which might not always find the optimal way to incorporate more lag data into the model.

Table 5.10 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 1 predictions using GP

	R^2 Lactation_1 train	R^2 Lactation_1 test	RMSE Lactation_1 train	RMSE Lactation_1 test
Lag_1	0.858369	0.858858	2.54475	2.54995
Lag_2	0.869364	0.869479	2.44200	2.45275
Lag_3	0.871888	0.872324	2.42225	2.42975
Lag_4	0.871017	0.871239	2.43140	2.43370
Lag_5	0.867645	0.868298	2.46020	2.46185
Lag_6	0.862918	0.863365	2.50415	2.50715
Lag_7	0.856232	0.856727	2.56750	2.56415

For the combined lactation groups 2, 3, 4, and 5, Table 5.11 illustrates the performance of the GP model. The R^2 values for training data demonstrate a higher level of predictive capacity compared to lactation 1 predictions, starting at 0.9138 for lag 1 and increasing to a peak of 0.9206 at lag 3. Test data follows a similar pattern, with values starting at 0.9141 for lag 1 and peaking at 0.9209 at lag 3. Similarly to what was showcased previously for the lactation 1 data, the R^2 values for both training and test data decrease until 0.9071 and 0.9073 respectively at lag 7. The RMSE values follow the corresponding pattern. For training data decrease from 3.3639 at lag 1 to 3.2311 at lag 3, indicating improved model accuracy. This trend continues for test data, with RMSE values decreasing from 3.3679 at lag 1 to 3.2364 at lag 3. Beyond lag 3, RMSE values show a slight increase, suggesting that the GP model diminishes its predictive accuracy with additional lag data from the 3rd lag forward.

As mentioned before, these findings indicate that although the GP model shows initial gains from incorporating more lag data, its performance reaches a peak after a certain threshold. Furthermore, as the results from primiparous cows' data, the close alignment of metrics between the training and test sets highlights the GP model's robust generalization capability, suggesting it maintains consistent performance even with unseen data.

Aligning with what has been showcased in the previous models, when comparing the results from primiparous with these from multiparous cows' data it is showed that although the R^2 values for the latest are quite superior compared with the ones from lactation 1 predictions, the RMSE values are also superior. The combination of higher variability in milk yield, complexity in lactation patterns, and overall, a higher sample diversity can result in higher R^2 values along with higher RMSE values for predictions involving cows from lactations 2, 3, 4, and 5.

Table 5.11 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions using GP

	R^2 Lactation_2_3_4_5 train	R^2 Lactation_2_3_4_5 test	RMSE Lactation_2_3_4_5 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.913840	0.914123	3.36390	3.36785
Lag_2	0.919905	0.920068	3.24380	3.25435
Lag_3	0.920567	0.920907	3.23110	3.23635
Lag_4	0.919143	0.919339	3.25645	3.26355
Lag_5	0.916295	0.916346	3.31525	3.32140
Lag_6	0.912357	0.912481	3.39095	3.39975
Lag_7	0.907082	0.907286	3.49200	3.50160

5.2.4. Comparison between different lags' models

A statistical comparison between the different lags for lactation 1 and the combined lactation groups 2, 3, 4, and 5 for each model is detailed in the subsequent analysis. The comparison was conducted across the 30 runs for each lag to verify if there were significant changes between the models' performance when varying the historical data considered.

Initially was performed a Friedman test to the RMSE results from the test set where most lag results showcased a significant difference from all others. Nevertheless it was imperative to understand the behavior of these differences and how the various lags differentiated from one another. Therefore, a post hoc Wilcoxon's test was applied to the RMSE results from the test set as well, revealing mostly significant differences across the lags for the three modeling approaches. The test was performed using the Bonferroni correction, adjusting the p-values, due to the multiple comparisons (Bonferroni, 1936).

The results, confirm the speculation from the previous insights, that for XGBoost and ANN most of the models' ability to predict milk yield varies depending on the number of lag days included in the analysis, highlighting the impact of historical data depth on predictive performance.

For the XGBoost model, as shown in Table 5.12 and Table 5.13, the analysis of lactation 1 data consistently showcased highly significant p-values, indicating the models' performance varied significantly across different lags. In the case of lactation 2, 3, 4, and 5 data, all comparisons between the model results from different lags also demonstrated highly significant p-values, except for the comparison between lag 6 and lag 7, which did not show a significant difference. This lack of significant difference between lag 6 and lag 7 suggests that adding more lag data beyond a certain point does not necessarily contribute to improved model performance for these lactations, possibly due to diminish value of additional historical data.

Table 5.12 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using XGBoost models for one-week-ahead predictions. * indicates statistical significant differences.

Week1 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

Table 5.13 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using XGBoost models for one-week-ahead predictions. * indicates statistical significant differences

Week1 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---	1.0000
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	1.0000	---

The results presented in Table 5.14 and Table 5.15 reveal the insights from the same comparisons between different lag models' performance for one-week-ahead milk production as presented before but for ANN. Table 5.14, which focuses on Lactation 1 predictions, shows that the majority of comparisons between different lag periods result in highly significant p-values, indicating substantial variations in the model's performance across different lag information. However, there are a few exceptions. The comparison between lag 5 and lag 6, indicates no significant difference in performance between considering 5 or 6 past days into the data. Similarly, lag 6 and 7 and lag 4 and 5 also show no significant difference. These findings suggest that while the ANN model's performance significantly varies with different lag periods, there are specific instances where the performance remains consistent. Particularly between consecutive later lags, the results indicate that the importance of increasing historical data decreases around the 4th and 5th lag day.

Table 5.14 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using ANN models for one-week-ahead predictions. * indicates statistical significant differences

Week1 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	0.1053	<0.05	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	0.1053	---	1.0000	<0.01*
Lag_6	<0.01*	<0.01*	<0.01*	<0.05*	1.0000	---	0.1214
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	0.1214	---

In Table 5.15, which examines predictions for Lactations 2, 3, 4, and 5, a similar pattern emerges but with more exceptions. Most comparisons show significant differences in performance, indicating the model's responsiveness to the amount of lag data used. For instance, comparisons between lag 1 and models with more historical data consistently show significant differences, demonstrating the impact of additional lag data on model performance. However, certain comparisons do not exhibit significant differences. Specifically, the comparisons between a model using a particular amount of lag data and the models using the next two subsequent lags often do not show significant differences. This indicates that while the inclusion of additional lag data initially improves the model's performance, there is a threshold beyond which further increases in lag data do not lead to substantial performance gains. These exceptions suggest a peak effect, where the benefit of adding more historical data does not benefit after a certain point.

Table 5.15 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using ANN models for one-week-ahead predictions. * indicates statistical significant differences

Week1 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	0.0847	0.2376	<0.01	<0.01	<0.01
Lag_3	<0.01*	0.0847	---	<0.01	0.4641	0.2869	<0.01
Lag_4	<0.01*	0.2376	1.0000	---	<0.01	0.9431	<0.05
Lag_5	<0.01*	<0.05*	0.4641	1.0000	---	<0.01	0.6871
Lag_6	<0.01*	<0.05*	0.2869	0.9431	1.0000	---	1.0000
Lag_7	<0.01*	<0.05*	<0.01*	<0.05*	0.6871	1.0000	---

The post hoc Wilcoxon test results, presented in Table 5.16 and Table 5.17, are a little distinct from the results for the XGBoost and ANN models.

In Table 5.16, which focuses on Lactation 1 predictions using GP, most comparisons show highly significant p-values, indicating notable differences in model performance across different lags. However, the exception does not follow the pattern from the previous model. The exceptions in the previous models were presented between very close lag information considered, while in this case the comparison between lag 2 and lag 4 is the only one showing a non-significant p-value. Given that GP individuals are simple trees, they might not be effectively capturing the additional information provided by more lag data. The simplicity of these trees can lead to a situation where the added complexity from extra lag data does not translate into improved model performance. Furthermore, the simple genetic operators used the GP implementation may not be as effective in combining and integrating the added lag data.

Table 5.16 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using GP for one-week-ahead predictions. * indicates statistical significant differences

Week1 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.05*	0.1397	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.05*	<0.01*	---	<0.05*	<0.05*	<0.05*
Lag_5	<0.01*	0.1397	<0.01*	<0.05*	---	<0.01*	<0.05*
Lag_6	<0.01*	<0.01*	<0.01*	<0.05*	<0.01*	---	<0.01*
Lag_7	<0.05*	<0.01*	<0.01*	<0.05*	<0.05*	<0.01*	---

In Table 5.17, which reflects the comparisons between different lag information for Lactations 2, 3, 4, and 5, a similar pattern is observed. Most comparisons exhibit significant differences, except for the results from the models using lag 1 and lag 6 data, as well as those using lag 2 and lag 4 data, which show no significant differences. Unlike the patterns observed in previous models, these non-significant differences are not between adjacent lags. This suggests that the GP model may not be effectively capturing the additional information provided by these specific lags. This observation aligns with the findings from Lactation 1, further indicating that the GP model's simplicity might limit its ability to leverage extended lag data comprehensively.

Table 5.17 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using GP for one-week-ahead predictions. * indicates statistical significant differences

Week1 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	0.2530	<0.01
Lag_2	<0.01*	---	<0.01	1.0000	<0.05*	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.05*	<0.01	<0.01	<0.01
Lag_4	<0.01*	1.0000	<0.05*	---	<0.01*	<0.01*	<0.01*
Lag_5	<0.01*	<0.05*	<0.01*	<0.01*	---	<0.01*	<0.01*
Lag_6	0.2530	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01*
Lag_7	<0.05*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

Overall, these results highlight that while the GP approach benefits from the inclusion of additional lag data, its performance does not consistently improve with every increase in lag.

5.2.5. Predict lactations_2_3_4_5 target with models trained with first lactation data

To assess the ability of models' generalization for predicting the milk yield on other lactations' data, the models trained on lactation 1 data were used to predict the milk yield for subsequent lactations, using the XGBoost model, ANN, and GP. Models were trained with the same Lactation 1 dataset and tested on data from multiparous cows.

In Table 5.18, it is seen that the XGBoost model's performance generally increases as more lags are added in the first lactation training data, as seen previously. However, this pattern is not followed in the test set. Although there is an overall slight increase in the model's performance, starting with an R^2 value of 0.7980 at lag 1 and 0.8326 at lag 7, it is not reflected across all lags. From lag 2 to 3, 4 to 5 and 6 to 7, the R^2 slightly decreases. The same pattern is seen in the RMSE values, where the overall values of the errors are much superior in the test data.

A similar behavior regarding the models' overall performance is seen in Table 5.19, now regarding the ANN models. The train data from the first lactation show increased performance as more lags are added while in test data the models' performance decreases from lag 2 to 6, having the best R^2 of 0.9061 at lag 2. The same pattern is followed in the RMSE values but with them being slightly smaller than the ones in the XGBoost results.

Table 5.18 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using XGBoost model

	R^2 lactation 1 train	R^2 Lactation_2_3_4_5 test	RMSE lactation 1 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.900380	0.798038	2.142133	5.162976
Lag_2	0.909092	0.819564	2.045058	4.880065
Lag_3	0.912956	0.816978	2.001052	4.914920
Lag_4	0.915340	0.823433	1.974913	4.827475
Lag_5	0.916720	0.820112	1.957689	4.872661
Lag_6	0.917755	0.832670	1.946160	4.699500
Lag_7	0.918261	0.832552	1.940100	4.701107

Table 5.19 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using ANN model

	R^2 lactation 1 train	R^2 Lactation_2_3_4_5 test	RMSE lactation 1 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.893870	0.877925	2.210078	4.014013
Lag_2	0.902530	0.906106	2.117875	3.520315
Lag_3	0.907522	0.905840	2.064141	3.525260
Lag_4	0.911143	0.902714	2.021790	3.583343
Lag_5	0.913107	0.891499	1.999538	3.784245
Lag_6	0.913644	0.890691	1.992896	3.798328
Lag_7	0.914567	0.893125	1.982624	3.755804

In Table 5.20, the GP model exhibits a different trend compared to XGBoost and ANN. As seen previously, the GP model's performance does not show a consistent pattern of improvement with additional lag data in the training set. The same is observed in the test set when using multiparous data. The R^2 values for the test set show fluctuations, starting at 0.9138 for lag 1, increasing slightly for lag 2 to 0.9199 and presenting the highest R^2 value of 0.9209 at lag 3 before dropping.

Concerning the general ability of the models to adapt to multiparous cows' data based only on primiparous data training, both XGBoost and ANN models exhibit overfitting across all lags. Although prior results show XGBoost's superior performance for predicting one-week-ahead average milk yield, it is clear here that ANN offers better generalization for multiparous cow data, as indicated by its closer training and test R^2 values. Furthermore, despite the fluctuations in predictions across lags, the GP model demonstrates a notable generalization

capability, producing a better performance when predicting multiparous cows' data than in the primiparous results, likely due to the simplicity of the tree models.

Table 5.20 - Median R^2 and RMSE metrics results for one-week-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using GP

	R^2 lactation 1 train	R^2 Lactation_2_3_4_5 test	RMSE lactation 1 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.858369	0.913835	2.54475	3.372324
Lag_2	0.869364	0.919908	2.44200	3.251326
Lag_3	0.871888	0.920547	2.42225	3.238320
Lag_4	0.871017	0.919063	2.43140	3.268422
Lag_5	0.867645	0.916249	2.46020	3.324753
Lag_6	0.862918	0.912300	2.50415	3.402237
Lag_7	0.856232	0.906972	2.56750	3.504063

The models trained exclusively on lactation 1 data show significant challenges when predicting milk production for other lactations. This difficulty reveals the inherent struggle in generalizing to multiparous cows' data, even with ANNs and GP demonstrating better generalization capabilities. This may indicate that the division between primiparous and multiparous data is relevant when predicting milk yield for models with a low generalization capacity as the distinct characteristics of these groups significantly impact model effectiveness.

5.3. TWO WEEKS AHEAD MILK YIELD PREDICTION

5.3.1. XGBoost

For two-weeks-ahead milk yield prediction using the XGBoost model, the performance metrics for lactation 1 are represented in Table 5.21. As in the previous results, the table contains the median metrics of 30 runs using a 5-fold cross validation with 6 repetitions.

The results demonstrate the improvement of R^2 values across the lags' datasets for both training and test data. When only the prior day's information is considered, lag 1, an R^2 value of 0.8611 was achieved, while considering data encompassing the previous week got an R^2 of 0.8904. For test data, the R^2 value incremented from 0.8513 for lag 1 to 0.8830 for lag 7. The RMSE scores had a reduction, descending from 2.5254 to 2.2418 for the training data and from 2.6150 to 2.3168 for the test data. The proximity between the training and validation performances suggests a good generalization capability of the models.

Nevertheless, it is also noticeable that the improvement rate in the models' performance diminishes with each subsequent lag. This trend suggests that with the importance of the added historical data decreases. Although the improvement rate is decreasing, compared with

one-week-ahead predictions, its decrease is lower, indicating that the incorporation of a broader historical data scope becomes increasingly more important for further predictions.

Furthermore, it is crucial to acknowledge the worse results in predicting capacity compared to one-week-ahead predictions considering first lactation data. This aligns with expectations since a greater extension of the forecasting horizon typically leads to a decrease in predictive performance.

Table 5.21 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using XGBoost model

	R^2 Lactation_1 train	R^2 Lacation_1 test	RMSE Lacation_1 train	RMSE Lacation_1 test
Lag_1	0.861120	0.851260	2.524548	2.615049
Lag_2	0.873062	0.864211	2.412856	2.500466
Lag_3	0.879237	0.870559	2.356068	2.435120
Lag_4	0.883316	0.875393	2.313222	2.391792
Lag_5	0.886575	0.878749	2.282374	2.360548
Lag_6	0.888955	0.881319	2.258334	2.335269
Lag_7	0.890423	0.882941	2.241783	2.316806

For the combined Lactation groups 2, 3, 4, and 5, Table 5.22 presents the results from the 30 runs of the XGBoost models. The R^2 values increase across the lag datasets for both the training and test datasets. Initiated with lag 1, with an R^2 value of 0.9043 that improves to 0.9210 in lag 7 in training data. In test data, the R^2 starts from 0.8996 in lag 1 to 0.9174 in lag 7. Similarly, the RMSE scores exhibit a decline across both datasets, indicating enhanced forecast accuracy with the extension of historical context.

Despite improved model performance with extended historical data, there is a deceleration in the rate of performance gains with each added lag, as seen in the previous analysis. This deceleration is more significant compared with the first lactation's predictions. Its pattern is more similar to the decrease of the improvement rate in one-week-ahead predictions for multiparous cows' data using the XGBoost.

Nevertheless, the general performance of the models, as in primiparous data, is inferior compared with one-week-ahead predictions. The best R^2 performance (lag 7) for two-weeks-ahead predictions is worse than the worst R^2 value (lag 1) for one-week-ahead predictions, for both train and test. This indicates that the predictions for two-weeks-ahead are more demanding, suggesting that the addition of more lags can be beneficial, especially for long-term predictions.

Table 5.22 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using XGBoost model

	R^2 Lactation_2_3_4_5 train	R^2 Lacation_2_3_4_5 test	RMSE Lacation_2_3_4_5 train	RMSE Lacation_2_3_4_5 test
Lag_1	0.904349	0.899591	3.569138	3.651638
Lag_2	0.911942	0.907562	3.425392	3.503572
Lag_3	0.915599	0.911679	3.352812	3.430385
Lag_4	0.917949	0.914047	3.305538	3.381395
Lag_5	0.919508	0.915736	3.275315	3.350148
Lag_6	0.920463	0.916711	3.256267	3.327394
Lag_7	0.920977	0.917432	3.243987	3.316333

5.3.2. ANN

For the ANN model the two-weeks-ahead milk yield predictions in Table 5.23 contain the performance metrics for lactation 1. In the training dataset, the R^2 values start at 0.8511 for lag 1, reflecting the prediction power when utilizing the immediate past day's data. This value incrementally escalates to 0.8925 for lag 7. A similar trend is observed in the test data, where the R^2 value rises from 0.8431 for lag 1 to 0.8858 for lag 7, underscoring an improvement in predictive capability as more historical data is used. The RMSE scores for both the training and validation datasets mirror the R^2 trend, with the scores for the training data showing a decrease from 2.6154 to 2.2219 and the test data from 2.6709 to 2.2893.

While the ANN model demonstrates reinforced predictive performance with the accumulation of historical data, the magnitude of performance improvement shows a diminishing for each subsequent lag. However, comparing the ANN results to the previously mentioned XGBoost results for lactations 2, 3, 4, and 5, the ANN model's deceleration in performance improvement appears more moderate. This could suggest a different responsiveness to the volume of historical data within the ANN model. The different sensitivity to historical data is also noticeable when comparing the results with XGBoost from lactation 1 data, where the predictions in lag 1 start slightly better but in lag 7 are below than the ANN's results.

Despite this, it is essential to note that, as in the XGBoost models, the ANN model's overall performance for two-weeks-ahead predictions does not surpass the one-week-ahead predictions. The best R^2 performance for two-weeks-ahead forecasts at lag 7 for the ANN model remains slightly below the lowest R^2 achieved at lag 1 for the one-week-ahead forecasts, for both training and validation datasets. This emphasizes the challenge in extending the predictive horizon and reaffirms the increased relevance of historical data for longer-term forecasts.

Table 5.23 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using ANN model

	R^2 Lactation_1 train	R^2 Lacation_1 test	RMSE Lacation_1 train	RMSE Lacation_1 test
Lag_1	0.851093	0.843149	2.615352	2.679049
Lag_2	0.869357	0.860670	2.448269	2.530873
Lag_3	0.873350	0.866283	2.409679	2.472895
Lag_4	0.882083	0.873406	2.329317	2.408838
Lag_5	0.885959	0.879561	2.288693	2.345323
Lag_6	0.889342	0.881830	2.252604	2.322474
Lag_7	0.892536	0.885822	2.221882	2.289273

For the ANN model in two-weeks-ahead milk yield predictions across combined lactations 2, 3, 4, and 5, Table 5.24 displays the performance. In the training set for these combined lactations, R^2 values initiate at 0.9007 for lag 1, demonstrating the model's predictive capacity when considering merely the prior day's data, increasing to 0.9201 for lag 7. The test data has a correspondent behavior, with R^2 values climbing from 0.8963 at lag 1 to 0.9164 at lag 7. Mirroring the trend observed in R^2 values, the RMSE metrics across both training and test sets suffers a decrease, from 3.6376 to 3.2639 in the training set, and from 3.7110 to 3.3374 in the test set.

As in all previous results, there is a consistent enhancement of predictive strength with additional historical data inclusion, but the incremental improvement in performance becomes less pronounced with each added lag.

In comparison to the XGBoost model for lactations 2, 3, 4, and 5, the ANN model demonstrates a similar responsiveness to historical data, being slightly different due to the little increase in improvement rate from lag 2 to lag 3.

It is important to note that, as seen in all results from two-weeks-ahead forecasts, the ANN model's two-weeks-ahead predictions, despite their incremental improvements through the lags, do not outperform the one-week-ahead predictions. The highest R^2 values achieved for the two-weeks-ahead forecast in lag 7 do not exceed the lowest R^2 values from one-week-ahead forecast in lag 1, within both the training and test data. This trend demonstrates one more time the challenges associated with lengthening the forecasting horizon, where historical data becomes even more important.

Table 5.24 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using ANN model

	R^2 Lactation_2_3_4_5 train	R^2 Lacation_2_3_4_5 test	RMSE Lacation_2_3_4_5 train	RMSE Lacation_2_3_4_5 test
Lag_1	0.900653	0.896259	3.637644	3.719599
Lag_2	0.906425	0.903595	3.530825	3.581741
Lag_3	0.913444	0.909403	3.395444	3.473062
Lag_4	0.916467	0.912669	3.335544	3.415636
Lag_5	0.918405	0.914614	3.297940	3.375350
Lag_6	0.919280	0.915390	3.280621	3.354541
Lag_7	0.920097	0.916413	3.263928	3.337402

5.3.3. GP

For the GP model of two-weeks-ahead milk yield predictions, Table 5.25 presents the performance metrics for lactation 1. In the training dataset, the R^2 values start at 0.7611 for lag 1, reflecting the prediction power when utilizing the immediate past day's data. This value incrementally increases to 0.7767 for lag 3, then slightly decreases until the 7th lag to 0.7544 similary to what happens in one-week-ahead predictions. The same trend is observed in the test data, where the R^2 value rises from 0.7614 for lag 1 to 0.7776 for lag 3, before decreasing to 0.7553 for lag 7. The RMSE scores for both the training and test datasets reflect the R^2 trend, with the scores for the training data showing a decrease from 3.3012 to 3.1920 and the test data from 3.3069 to 3.1964 at lag 3, then increasing to 3.3476 for training and 3.3551 for test data at lag 7.

The GP model demonstrates a modest improvement in predictive performance with the inclusion of more historical data. However, the GP performance does not consistently improve beyond lag 3, highlighting a potential limitation in capturing additional information from extended historical data. Despite this, the GP model's metrics suggest a reasonable generalization ability, as the metrics for the training and test sets remain relatively close, indicating consistent performance on unseen data.

Table 5.25 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 1 predictions using GP

	R^2 Lactation_1 train	R^2 Lacation_1 test	RMSE Lacation_1 train	RMSE Lacation_1 test
Lag_1	0.761140	0.761436	3.30115	3.30685
Lag_2	0.773589	0.774118	3.21365	3.21820
Lag_3	0.776689	0.777598	3.19190	3.19635
Lag_4	0.774945	0.776522	3.20215	3.20520
Lag_5	0.770188	0.771239	3.23650	3.23935
Lag_6	0.763867	0.765020	3.28225	3.28570
Lag_7	0.754429	0.755253	3.34760	3.35505

In Table 5.26, which focuses on the two-weeks-ahead predictions for the combined lactations 2, 3, 4, and 5, a similar pattern is observed. The R^2 values in the training set start at 0.8415 for lag 1, increasing to 0.8483 at lag 3, and then gradually decreasing to 0.8302 at lag 7. The test data follows a corresponding trend, with R^2 values rising from 0.8414 at lag 1 to 0.8485 at lag 3, then decreasing to 0.8306 at lag 7. The RMSE scores mirror this trend as well, with the training data showing a decrease from 4.5846 at lag 1 to 4.48345 at lag 3, followed by an increase to 4.7432 at lag 7. The test data similarly shows a decrease in RMSE from 4.5954 at lag 1 to 4.4933 at lag 3, and then an increase to 4.7544 at lag 7.

These results highlight that, while the GP model benefits initially from additional historical data, its performance stabilizes or even diminishes beyond certain lags, specifically after lag 3. This suggests that the GP model, while effective, may struggle to leverage extended lag data as effectively as other models like XGBoost or ANN. The relatively close RMSE values between training and test sets indicate good generalization ability, yet the overall predictive performance for two-weeks-ahead forecasts remains lower than one-week-ahead predictions.

Table 5.26 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions using GP

	R^2 Lacation_2_3_4_5 train	R^2 Lacation_2_3_4_5 test	RMSE Lacation_2_3_4_5 train	RMSE Lacation_2_3_4_5 test
Lag_1	0.841454	0.841383	4.58460	4.59535
Lag_2	0.847786	0.847778	4.49015	4.50040
Lag_3	0.848264	0.848464	4.48345	4.49330
Lag_4	0.846073	0.846903	4.51570	4.51800
Lag_5	0.842237	0.842989	4.57210	4.57825
Lag_6	0.837104	0.837346	4.64590	4.64985
Lag_7	0.830185	0.830630	4.74320	4.75440

Once again, it is observed that although the R^2 values are higher when considering data from multiparous cows, the RMSE errors are also higher. This can be related to inherent variability and scale differences between primiparous and multiparous cows. Multiparous cows, those that have given birth more than once, generally exhibit more stable and consistent milk production patterns compared to primiparous cows, which are in their first lactation. This stability leads to higher R^2 values as the models can more accurately capture the relationship between input features and milk yield. However, multiparous cows also tend to produce higher amounts of milk on average, increasing the absolute errors in prediction, as reflected in the higher RMSE values (Pfeilsticker et al., 1996).

5.3.4. Comparison between different lags' models

Similarly to what was conducted upon the results from the one-week-ahead predictions, were performed statistical comparisons between different lag information for lactation 1 and the combined lactation groups 2, 3, 4, and 5 for the three approaches, XGBoost, ANN, and GP to understand if historical data provides better predictive capacity.

The post hoc Wilcoxon test with the Bonferroni correction, conducted on the 30 runs of the RMSE results from the test set, shows, most of the comparisons, significant differences in performance across lags.

For the XGBoost model, the post hoc Wilcoxon test results for lactation 1 predictions, shown in Table 5.27, and the results for lactations 2, 3, 4 and 5, shown in Table 5.28, revealed that all comparisons between different lags yielded highly significant p-values. This suggests substantial variations in models' depending on the number of lags considered, reinforcing the importance of historical data depth.

Table 5.27 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using XGBoost for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01*	<0.01*	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01*	<0.01*
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01*
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

Table 5.28 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using XGBoost for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.01*
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

The ANN model's two-weeks-ahead predictions for lactation 1, detailed in Table 5.29, also showed significant differences across all lag comparisons. However, there was a decrease in the significance between the results from lag 6 and lag 5. For lactation 2, 3, 4, and 5 predictions, shown in Table 5.30, the ANN model exhibited significant differences across most lag comparisons although there were some exceptions. Notably, the comparisons between lag 6 and lag 5, as well as between lag 7 and lag 5 and 6, did not show significant differences. This suggests that the ANN model's performance improvements peak after a certain point, where additional historical data does not further enhance the model's predictive performance.

Compared to the results from the one-week-ahead predictions, these findings indicate that a greater amount of historical data is crucial for accurately predicting two-weeks-ahead. Compared to XGBoost, the importance of lag data appears to be somewhat less important when using the ANN model.

Table 5.29 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using ANN for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01*	<0.01*	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	<0.05*	<0.01*
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	<0.05*	---	<0.01*
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

Table 5.30 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using ANN for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_3	<0.01*	<0.01*	---	<0.01	<0.01	<0.01	<0.01
Lag_4	<0.01*	<0.01*	<0.01*	---	<0.01	<0.01	<0.01
Lag_5	<0.01*	<0.01*	<0.01*	<0.01*	---	1.0000	0.2086
Lag_6	<0.01*	<0.01*	<0.01*	<0.01*	1.0000	---	1.0000
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	0.2086	1.0000	---

The GP model's predictions for lactation 1, as indicated in Table 5.31, revealed significant differences across most lag comparisons, with some exceptions as well. The model's performance, when considering only one day of past data, differed significantly from all other models that included additional lag data. The test results between each lag and its subsequent lags were often non-significant, with notable differences emerging primarily when compared to the 7th lag data.

Table 5.31 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 1 predictions using GP for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac1	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Lag_2	<0.01*	---	<0.05*	0.9924	1.0000	0.1710	<0.01
Lag_3	<0.01*	<0.05*	---	0.0980	<0.01	0.0168	<0.01
Lag_4	<0.01*	0.9924	0.0980	---	0.1599	0.2869	<0.01
Lag_5	<0.01*	1.0000	<0.01*	0.1599	---	0.2374	<0.01*
Lag_6	<0.01*	0.1710	0.0168	0.2869	0.2374	---	<0.01*
Lag_7	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	---

For lactation 2, 3, 4, and 5 predictions using GP, shown in Table 5.32, significant differences were observed across most lag comparisons. However, exceptions were noted between lag 1 and both lag 6 and lag 7, where differences were not significant. Additionally, no significant differences were found between lag 2 and its two subsequent lags, as well as between lag 3 and its subsequent lag, lag 4.

These non-significant differences that occur between non-adjacent lags, further support the idea that the GP model may struggle to capture the incremental information provided by more extended lag data. This pattern suggests that the GP model's structure and its simple operators may limit its ability to harness the full potential of additional lag data for predictive improvement.

Table 5.32 - Post hoc Wilcoxon test p-values for different lags comparisons for Lactation 2, 3, 4 and 5 predictions using GP for two-weeks-ahead predictions. * indicates statistical significant differences

Week2 Lac 2_3_4_5	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Lag_1	---	<0.01	<0.01	<0.01	<0.05*	1.0000	0.5826
Lag_2	<0.01*	---	1.0000	1.0000	<0.01	<0.01	<0.01
Lag_3	<0.01*	1.0000	---	1.0000	<0.01	<0.01	<0.01
Lag_4	<0.01*	1.0000	1.0000	---	<0.01	<0.01	<0.01
Lag_5	<0.05*	<0.01*	<0.01*	<0.01*	---	<0.05*	<0.01*
Lag_6	1.0000	<0.01*	<0.01*	<0.01*	<0.05*	---	<0.05*
Lag_7	0.5826	<0.01*	<0.01*	<0.01*	<0.01*	<0.05*	---

Considering these results, we see that for two-weeks-ahead predictions, the inclusion of more historical data continues to play, in most cases, a critical role in enhancing the predictive accuracy of the models.

Compared to the one-week-ahead predictions presented in Table 5.12 to Table 5.17, the results for two-weeks-ahead predictions using XGBoost models show very similar patterns, indicating consistent pattern between different lags. However, for the ANN model, the two-weeks-ahead predictions exhibit more significant p-values, highlighting the increased importance of including additional lag data when predicting two-weeks-ahead. In contrast, the GP model shows less significant differences between lags for two-weeks-ahead predictions compared to one-week-ahead predictions, for both primiparous and multiparous data. As mentioned before, this suggests that the GP model's simpler structure may limit its ability to effectively utilize the added complexity from more extended lag data, thereby reducing its performance gains over longer prediction horizons.

Moreover, the results reflect fundamental differences in model's performances throughout the lags and their ability to handle data from primiparous versus multiparous cows. The XGBoost model, as an ensemble model, demonstrates a strong ability to integrate new data inputs, similar to the ANN model, although to a slightly lesser extent. This analysis not only confirms the statistical significance of the depth of historical data in most cases but also offers insights into model behavior across different temporal frameworks. Consistent with one-week-ahead predictions, the importance of incorporating additional historical data using more lag days was evident, leading to better and more reliable predictions.

5.3.5. Predict lactations_2_3_4_5 target with models trained with first lactation data

Following the same approach used in one-week-ahead predictions, it was evaluated the ability to generalize models trained on lactation 1 data to predict the milk production from further lactations' data. Therefore, XGBoost, ANN, and GP were trained using primiparous cows' data and then tested on datasets from multiparous cows.

The XGBoost model, in Table 5.33, demonstrates a gradual improvement in predictive performance as more lags are added as seen previously. The R^2 value for the test set generally increases from 0.7151 at lag 1 to 0.7705 at lag 5. However, lags 6 and 7 show a slight decline, falling to 0.7619 and 0.7457 respectively. The RMSE values for the test set follow the same pattern decreasing from 6.1601 at lag 1 to 5.5284 at lag 5 and then increasing at lag 6 and 7.

Similarly, in Table 5.34, the ANN model follows the trend of improved predictive performance with additional lags in training data. However, the highest R^2 value for the test set of 0.8031 is observed when considering data only from the previous day, at lag 1. When predicting multiparous cows' milk yield, there is not an overall improvement of models' performance as more lags' data is added. Actually, the performances are relatively stable across all lags, indicating consistent predictive performance.

In Table 5.35, the GP model presents an interesting contrast. The R^2 values for the test set peak at 0.8481 for lag 3, with a subsequent decline observed as more lags are added. The

RMSE values follow a similar pattern, decreasing to 4.4988 at lag 3 and then increasing at later lags. This behavior suggests that the GP model benefits from some historical data but may struggle with integrating extensive lag data effectively, possibly due to the simplicity of the genetic programming trees and operators used. This aligns with the results observed before.

As with one-week-ahead predictions, all three approaches exhibit a decline in generalization capability for two-weeks-ahead predictions. For XGBoost, overfitting is more pronounced, with greater disparities between training and test results compared to the ANN models, showing the latter's better ability to adapt to multiparous cows' data. Both models exhibit greater overfitting in two-weeks-ahead predictions compared to one-week-ahead predictions, with more pronounced differences in R^2 values between training and test data.

In contrast, the GP approach appears to generalize better as seen in the results for one-week-ahead predictions, being its R^2 values better in the test data. The discrepancy between the increase in the R^2 values and also the increase in RMSE values may be related to the tendency of multiparous cows to produce higher amounts of milk on average, increasing the absolute errors in prediction (Pfeilsticker et al., 1996).

The two-weeks-ahead prediction in test set show, as in training data, a reduced predictive performance in two-weeks-ahead compared to one-week-ahead predictions with higher RMSE values and slightly lower R^2 values for the two-week predictions.

Predicting two-weeks-ahead milk average production proves to be more challenging than one-week-ahead predictions, as seen in the reduced model performances. This once again emphasizes the importance of a separate analysis to predict milk yield for primiparous and multiparous cows, recognizing the distinct patterns and challenges inherent in each group's data. The results highlight the varying sensitivity of different models to the amount of historical data, with GP although with a lower performance than XGBoost and ANN, showing better adaptation capabilities.

Table 5.33 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using XGBoost model

	R^2 Lactation_1 train	**R^2 Lactation_2_3_4_5 test**	**RMSE Lactation_1 train**	**RMSE Lactation_2_3_4_5 test**
Lag_1	0.861120	0.715101	2.524548	6.160145
Lag_2	0.873062	0.738211	2.412856	5.905023
Lag_3	0.879237	0.747672	2.356068	5.797209
Lag_4	0.883316	0.764678	2.313222	5.598569
Lag_5	0.886575	0.770543	2.282374	5.528373
Lag_6	0.888955	0.761940	2.258334	5.631052
Lag_7	0.890423	0.745747	2.241783	5.819399

Table 5.34 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using ANN model

	R^2 Lactation_1 train	R^2 Lactation_2_3_4_5 test	RMSE Lactation_1 train	RMSE Lacation_2_3_4_5 test
Lag_1	0.851093	0.803120	2.615352	5.120908
Lag_2	0.869357	0.771842	2.448269	5.512667
Lag_3	0.873350	0.793288	2.409679	5.247221
Lag_4	0.882083	0.782780	2.329317	5.378931
Lag_5	0.885959	0.792916	2.288693	5.251938
Lag_6	0.889342	0.797590	2.252604	5.192287
Lag_7	0.892536	0.775217	2.221882	5.471688

Table 5.35 - Median R^2 and RMSE metrics results for two-weeks-ahead milk production for Lactation 2_3_4_5 predictions trained with Lactation 1 data using GP

	R^2 Lactation_1 train	R^2 Lactation_2_3_4_5 test	RMSE Lactation_1 train	RMSE Lactation_2_3_4_5 test
Lag_1	0.761140	0.841295	3.30115	4.597709
Lag_2	0.773589	0.847668	3.21365	4.504457
Lag_3	0.776689	0.848053	3.19190	4.498765
Lag_4	0.774945	0.845887	3.20215	4.530705
Lag_5	0.770188	0.842062	3.23650	4.586583
Lag_6	0.763867	0.836815	3.28225	4.662148
Lag_7	0.754429	0.829973	3.34760	4.758888

5.4. COMPARISON BETWEEN DIFFERENT MODELS

To compare the performance between the two different implemented models a Wilcoxon test was performed for each one of the 28 different datasets. The 30 runs were used to compare the predictive performance of the three different approached two by two for one-week and two-weeks-ahead average milk yield predictions. There were used unilateral tests with the alternative hypothesis testing whether the RMSE values for one approach are better than the results from another. There were conducted three different comparisons with the following alternative hypothesis:

- XGBoosts' RMSE values are smaller (better) than those for ANN (H_1: RMSE_xgb – RMSE_ann < 0) (Table 5.36)
- XGBoosts' RMSE values are smaller (better) than those for GP (H_1: RMSE_xgb – RMSE_GP < 0) (Table 5.37)

- ANNs' RMSE values are smaller (better) than those for GP (H_1: RMSE_ann – RMSE_GP $<$ 0) (Table 5.38)

The results from the 28 performed tests, present for both lactation 1 and the lactations 2, 3, 4, and 5, for the one-week-ahead predictions, extremely low p-values for all lags across all comparisons. This indicates a statistically significant result where the XGBoost models have smaller RMSE values than the ANN and GP, meaning that XGBoost performs better in terms of RMSE across all lags for the prediction of the first week.

Regarding the two-weeks-ahead predictions the results are not the same for all lags in the distinct lactation groups only when comparing XGBoost with ANN. Using a significance level of 0.01, for lactation 1, the p-values are significant from lag 1 to 4, suggesting that XGBoost continues to have a smaller RMSE and thus performs better at these lags. However, at lags 5, 6 and 7, the p-values are not significant. This indicates that as more historical data is added the more pronounced is the non-rejection of the null hypothesis, which implies that there is no statistical evidence to affirm that XGBoost has better performance than ANN at these lags for the two-weeks-ahead predictions. Compared to one-week-ahead predictions, when predicting two-weeks-ahead values with the XGBoost models the results are quite different. While in most cases still outperform the ANN, it showcases a proximity between the two approaches performances when considering more lag data, suggesting that the ANN becomes better when integrating more extensive historical data.

Moreover, these results lead to the conclusion that for the first week predictions, XGBoost significantly outperforms ANN and GP across all lags. For second week, XGBoost generally maintains better performance at earlier lags, but this advantage diminishes as moving to later lags, where the difference between the XGBoost and ANN performance is no longer statistically significant for lactation 1 in lag 5 to 7.

These findings suggest that XGBoost may be particularly well-suited for short-term predictions or scenarios involving smaller windows of historical data. Combining multiple weak learners, XGBoost may be able to capture more complex patterns in the data. In contrast, ANN and GP appear to be less effective for this specific problem, potentially due to a limited hyperparameter search space and, in the case of GP, its inherent simplicity.

Table 5.36 - p-value of the Wilcoxon test between RMSE results from XGBoost modeling compared with ANN for the 28 datasets with H_0: RMSE_xgb = RMSE_ann vs H_1: RMSE_xgb – RMSE_ann < 0. * indicates statistical significant differences.

		Lactation 1	Lacation_2_3_4_5
Week 1	Lag_1	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	<0.01*	<0.01*
	Lag_6	<0.01*	<0.01*
	Lag_7	<0.01*	<0.01*
Week 2	Lag_1	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	0.6039	<0.01*
	Lag_6	0.9080	<0.01*
	Lag_7	1.0000	<0.05*

Table 5.37 - p-value of the Wilcoxon test between RMSE results from XGBoost modeling compared with GP for the 28 datasets with H_0: RMSE_xgb = RMSE_GP vs H_1: RMSE_xgb – RMSE_GP < 0. * indicates statistical significant differences.

		Lactation 1	Lacation_2_3_4_5
Week 1	Lag_1	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	<0.01*	<0.01*
	Lag_6	<0.01*	<0.01*
	Lag_7	<0.01*	<0.01*
Week 2	Lag_1	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	<0.01*	<0.01*

Lag_6	<0.01*	<0.01*
Lag_7	<0.01*	<0.01*

Table 5.38 - p-value of the Wilcoxon test between RMSE results from ANN modeling compared with GP for the 28 datasets with H_0: RMSE_ann = RMSE_GP vs H_1: RMSE_ann − RMSE_GP < 0.
* indicates statistical significant differences.

		Lactation 1	**Lacation_2_3_4_5**
Week 1	**Lag_1**	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	<0.01*	<0.01*
	Lag_6	<0.01*	<0.01*
	Lag_7	<0.01*	<0.01*
Week 2	**Lag_1**	<0.01*	<0.01*
	Lag_2	<0.01*	<0.01*
	Lag_3	<0.01*	<0.01*
	Lag_4	<0.01*	<0.01*
	Lag_5	<0.01*	<0.01*
	Lag_6	<0.01*	<0.01*
	Lag_7	<0.01*	<0.01*

In further investigations, the splitting of the data by each lactation stage and the application of distinct models to each category seems to be a path to follow. This differentiation can allow the identification of unique patterns inherent to each lactation period and undertake possible differences between lactations' milk production.

5.5. FEATURE IMPORTANCE ASSESSED WITH GP TREES

Using the final trees from the GP model in the 50th generation, there were identified the features that significantly contributed to each dataset's tree model. By analyzing the top 5 features used in the trees across 28 datasets for all 30 runs, it was constructed a comprehensive table that highlights the frequency of each feature's appearance in the top 5.

To simplify the interpretation, the features were grouped independently of the number of lags considered for the different datasets and the prediction week. In the annexes, a bar chart is provided where the features are categorized for each of the four groups, reflecting the different lactation groups and weeks predicted in Figure A. 2.

The Table 5.39 reveals insightful patterns about the importance of various features in predicting milk yield. As expected, ‘milk production kg’ emerged as the most significant feature, appearing in the top 5 for all 28 datasets. This indicates it plays a critical role in the prediction models, underscoring its strong influence on the outcomes. Closely following is ‘days in lactation’, that for one-week-ahead predictions refers to the day of lactation immediately preceding the prediction interval and for two-weeks-ahead predictions refers to the day preceding the prediction interval by one week. This feature appeared in the top 5 for 24 datasets, underscoring the importance of the cows' lactation period in accurately predicting milk yield. This matches the result from Piwczyński et al. (2020), where the month of milking revealed to be one of the most important factors responsible for monthly milk yield. The ‘milking failures’ feature was the third most frequent feature, appearing in the top 5 for 16 datasets. This suggests that the number of milking failures is a substantial factor affecting milk production. This is a finding that, to the best of my knowledge, has not been emphasized to this extent in previous studies, highlighting the critical role that milking failures play in predicting milk yield, and offering new insights for farmers. By understanding this relationship, farmers can focus on managing and reducing milking failures, potentially boosting overall milk production and possibly reducing the occurrence of diseases.

Features ‘milkings’ and ‘milk protein pct’ were both featured in the top 5 for 10 datasets, indicating their moderate but notable importance in the prediction model aligning with previous studies that show milking frequency as an important feature to predict milk yield, (Piwczyński et al., 2020; Kliś, Piwczyński, et al., 2021). Other features such as ‘milk lactose pct’, ‘average milk speed avg’ also present in the results from Kliś, Piwczyński, et al. (2021), ‘pregnancy length lag’ and ‘milk yield per box time avg kg’ appeared in the top 5 for a smaller number of datasets, ranging from 5 to 7 occurrences. These features, while not as dominant, still play a significant role in the prediction models, contributing to the overall accuracy and reliability of the forecasts.

Less frequently mentioned features include ‘delivery age months’, ‘lactation number’, ‘cow age months’, ‘milk fat pct’ and ‘maximum milk conductivity’ each appearing 4 times in the top 5. This suggests that these factors have a more limited but still relevant impact on milk yield predictions. Finally, features such as ‘maximum milk flow duration of the teats avg sec’, ‘activity avg’, ‘milking box time avg sec’, ‘duration first teat con till last teat disconnection’, ‘milking refusals’, and ‘rumination minutes’ appeared in the top 5 only once or twice, indicating a minimal significance on influencing on the prediction outcomes.

In summary, the analysis of the top features extracted from the GP model trees underscores the paramount importance of certain factors like milk production and days in lactation, while also highlighting the diverse range of other influential features. The number of milking failures emerged interestingly as well as one of the most important features. These insights can guide further optimization of prediction models and enhance our understanding of the key drivers of milk yield in dairy cows.

Table 5.39 - Number of times each feature was present in the top 5 of features that most contribute to the tree model

Feature name	Number of times the feature is in the top 5
milk_production_kg	28
days_in_lactation	24
milking_failures	16
milkings	10
milk_protein_pct	10
milk_lactose_pct	7
average_milk_speed_avg	7
pregnancy_length_lag	5
milk_yield_per_box_time_avg_kg	5
delivery_age_months	4
lactation_number	4
cow_age_months	4
milk_fat_pct	4
maximum_milk_conductivity	4
maximum_milk_flow_duration_of_the_teats_avg_sec	2
activity_avg	2
milking_box_time_avg_sec	1
duration_first_teat_con_till_last_teat_disconnection	1
milking_refusals	1
rumination_minutes	1

6. CONCLUSION

This study significantly advances the field of predictive modeling in dairy farming by addressing key gaps identified in the literature. This research focused on predicting the milk yield weekly, together with the analysis of impact of temporal data and lactation variability to develop robust models for milk production. Using data from multiple farms enhanced the robustness of the analysis, providing valuable insights that facilitate more personalized, effective, and sustainable dairy cow management strategies.

The analysis explored the prediction of average milk yield for one-week and two-weeks-ahead using three distinct approaches: Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANNs), and Genetic Programming (GP). XGBoost and ANN were chosen based on their established effectiveness in previous studies, while GP introduced a novel approach in the field of cow milk yield prediction.

To assess the importance of historical data for performing better predictions, the study considered the data of various lag periods ranging from 1 day to 7 days before the prediction interval. In alignment with the literature, the analysis was conducted on two groups: primiparous (first-time calving) and multiparous (multiple-calving) cows. This resulted in 28 different datasets, encompassing four primary categories (primiparous and multiparous cows for one-week and two-week predictions), with each category comprising seven datasets corresponding to different lag periods.

The findings from these 28 datasets generally indicated strong predictive performance with no signs of overfitting, underscoring the models' effectiveness in forecasting the average milk yield. As anticipated, the predictive accuracy of the models decreased for two-weeks-ahead forecasts, reflected by lower R^2 values and higher RMSE scores. This decline highlights the inherent challenges of making longer-term predictions in this context.

Among the models used, XGBoost provided in general the best predictions, as evidenced by the lowest RMSE values, highlighting its superior forecasting capability. This result aligns with expectations, given XGBoost's ensemble nature, which typically enhances predictive performance. In contrast, GP exhibited the weakest predictive capacity while ANN performed moderately well, but its results were more variable compared to XGBoost.

An examination of the impact of different lag periods revealed that XGBoost benefited significantly from increased lag data, showing marked improvements in predictive accuracy. This indicates that XGBoost can effectively utilize additional historical information to enhance its predictions. In contrast, ANN and GP displayed more variable results. ANN showed differences only between more distant lags. GP's performance was generally less sensitive to lag variations, indicating a less robust approach to handling temporal data.

Despite XGBoost's overall dominance in predicting average milk yield for both lactation groups, it demonstrated substantial overfitting when models trained on primiparous cows' data were applied to multiparous cows. The cross-validation using data from primiparous cows to predict the milk yield of multiparous cows helped evaluate the models' robustness and generalizability demonstrating a considerably more pronounced overfitting in XGBoost compared to the results observed in ANN and GP models. ANN exhibited minimal overfitting, while GP showed a good generalization capability, yielding superior results on unseen data. This shows that there are underlying patterns in the milking process that are the same for these two groups of cows. Therefore, both ANN and GP appear to be capturing “more general” patterns, while XGBoost, although being highly accurate, seems to capture the idiosyncrasies of each lactation group (primiparous and multiparous).

Although the GP approach demonstrated a smaller prediction capacity, possibly due to the limited use of the most common genetic operators, it showed a fast convergence. This rapid convergence might indicate a relatively low complexity of the patterns in the data.

Nevertheless, using the GP approach provided a deeper understanding of the problem's complexity and highlighted the most important features for accurate predictions. Across all 28 different predictions, the milk production data from the previous day (for lag one data) or the average of the preceding days (for subsequent lags) consistently emerged as the most crucial feature. This consistent prominence underscores the direct impact of recent milk production trends on future yields.

Following milk production, the days in lactation were identified as the second most significant feature. This reaffirms the importance of the lactation stage in predicting milk yield, aligning with existing studies (Kliś, Piwczyński, et al., 2021, where it is shown that productivity may change throughout cows’ lactation periods).

Unexpectedly, the third most important feature was the number of milking failures, which had not been previously highlighted to this extent in similar studies. This suggests that milking failures play a critical role in forecasting milk yield, providing new insights for farmers. Understanding this relationship can help in managing and reducing milking failures, thereby potentially improving overall milk production. Additionally, these findings offer valuable practical insights for farmers as reducing milking failures directly enhances animal welfare. Minimizing stress and discomfort for cows may lead to healthier and more productive livestock.

The number of milkings and the milk protein percentage were also identified as significant features, though to a lesser extent. Other features also helped in forecasting milk yield, but the ones mentioned previously held the most importance.

For future work, it may be important to better fine-tune the ANN and XGBoost models as well as testing the GP with additional different settings, possibly improving the models' predictive and generalization capacity. For example, to improve generalization capacity to multiparous data predictions trained with primiparous data, could be added and tested different regularization techniques to the models. By, for instance, adding a penalty term to the loss function the importance of the models' specific features or coefficients can be reassigned avoiding overfitting. In order to improve generalization, it would also be relevant to add data from more farms, even outside Italy, considering more varied data.

Not only the deepening of the already used machine learning approaches is important but also the use of different ones. The development and application of other algorithms that can generate human-interpretable models would allow to add more robustness in understanding more about the underlying factors and features that impact milking activities. This would also permit a comparison with the already established important features. A hybrid approach would also be a potential path to follow. Tree-based Pipeline Optimization (TPOT) combines the GP approach to efficiently search over an extensive range of supervised regression models, their hyperparameters as well as different data transformations retrieving the best pipeline (Olson et al., 2016). With the TPOT approach, it would be possible, taking advantage of GP, to automatically design and optimize the machine learning pipeline. Nevertheless, TPOT requires a very high computational power considering the number of different combinations performed and the amount of data used.

Additionally, and regarding specifically the results from XGBoost, in both one-week and two-weeks-ahead predictions it was possible to acknowledge the increase in prediction accuracy as more lag days were added. Furthermore, it would be pertinent to consider additional previous lag days, above the 7 days boundary used, to understand if there exists a threshold where the additional past information no longer improves the predictions.

Finally, concerning the distinguished result from the feature importance, a key focus of future investigation would also be to understand the relationship between milking failures and milk production. Exploring this connection could provide valuable information for improving both milk yield and animal welfare.

BIBLIOGRAPHICAL REFERENCES

Becker, C. A., Collier, R. J., & Stone, A. E. (2020). Invited review: Physiological and behavioral effects of heat stress in dairy cows. *Journal of Dairy Science*, *103*(8), 6751–6770. https://doi.org/10.3168/jds.2019-17929

Berckmans, D. (2014). Precision livestock farming technologies for welfare management in intensive livestock systems. *Revue Scientifique Et Technique (International Office of Epizootics)*, *33*(1), 189–196. https://doi.org/10.20506/rst.33.1.2273

Bianchi, M. C., Bava, L., Sandrucci, A., Tangorra, F. M., Tamburini, A., Gislon, G., & Zucali, M. (2022). Diffusion of precision livestock farming technologies in dairy cattle farms. *Animal*, *16*(11), 100650. https://doi.org/10.1016/j.animal.2022.100650

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità.* Seeber.

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random Forest Modelling of Milk Yield of Dairy Cows under Heat Stress Conditions. *Animals*, *11*(5), 1305. https://doi.org/10.3390/ani11051305

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cockburn, M. (2020). Review: Application and Prospective Discussion of Machine Learning for the Management of Dairy Farms. *Animals: An Open Access Journal from MDPI*, *10*(9), 1690. https://doi.org/10.3390/ani10091690

Dawkins, M. S. (2017). Animal welfare and efficient farming: Is conflict inevitable? *Animal Production Science*, *57*(2), 201. https://doi.org/10.1071/AN15383

Edwards, J. P., Jago, J. G., & Lopez-Villalobos, N. (2014). Analysis of milking characteristics in New Zealand dairy cows. *Journal of Dairy Science*, *97*(1), 259–269. https://doi.org/10.3168/jds.2013-7051

Elgersma, G. G., De Jong, G., Van Der Linde, R., & Mulder, H. A. (2018). Fluctuations in milk yield are heritable and can be used as a resilience indicator to breed healthy cows. *Journal of Dairy Science*, *101*(2), 1240–1250. https://doi.org/10.3168/jds.2017-13270

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Boosting Algorithms: A Review of Methods, Theory, and Applications. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp. 35–85). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

Fuentes, S., Gonzalez Viejo, C., Cullen, B., Tongson, E., Chauhan, S., & Dunshea, F. (2020). Artificial Intelligence Applied to a Robotic Dairy Farm to Model Milk Productivity and Quality

based on Cow Data and Daily Environmental Parameters. *Sensors*, *20*. https://doi.org/10.3390/s20102975

Fuertes, E. (2023). *Impact of Climatic Conditions on the milk production of dairy cows in an automatic milking system*. Thesis submitted to Università di Torino. https://sia.unito.it/studenti/intesi/Ricerca_tesi_libera/ricerca_tesi_dettaglio.asp?id_upload=254427&cdl_tesi=&cdl=&matricola=996648

Grzesiak, W., Zaborski, D., Szatkowska, I., & Królaczyk, K. (2021). Lactation milk yield prediction in primiparous cows on a farm using the seasonal auto-regressive integrated moving average model, nonlinear autoregressive exogenous artificial neural networks and Wood's model. *Animal Bioscience*, *34*(4), 770–782. https://doi.org/10.5713/ajas.19.0939

Haykin, S. S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

Jacobs, J. A., & Siegford, J. M. (2012). Invited review: The impact of automatic milking systems on dairy cow management, behavior, health, and welfare. *Journal of Dairy Science*, *95*(5), 2227–2247. https://doi.org/10.3168/jds.2011-4943

Jensen, D., Voort, M., & Hogeveen, H. (2018). Dynamic forecasting of individual cow milk yield in automatic milking systems. *Journal of Dairy Science*, *101*. https://doi.org/10.3168/jds.2017-14134

Ji, B., Banhazi, T., Phillips, C. J. C., Wang, C., & Li, B. (2022). A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. *Biosystems Engineering*, *216*, 186–197. https://doi.org/10.1016/j.biosystemseng.2022.02.013

John, A. J., Clark, C. E. F., Freeman, M. J., Kerrisk, K. L., Garcia, S. C., & Halachmi, I. (2016). Review: Milking robot utilization, a successful precision livestock farming evolution. *Animal*, *10*(9), 1484–1492. https://doi.org/10.1017/S1751731116000495

Kliś, P., Piwczyński, D., Sawa, A., & Sitkowska, B. (2021). Prediction of Lactational Milk Yield of Cows Based on Data Recorded by AMS during the Periparturient Period. *Animals*, *11*(2), 383. https://doi.org/10.3390/ani11020383

Kliś, P., Sawa, A., Piwczyński, D., Sitkowska, B., & Bogucki, M. (2021). Prediction of cow's fertility based on data recorded by automatic milking system during the periparturient period. *Reproduction in Domestic Animals = Zuchthygiene*, *56*(9), 1227–1234. https://doi.org/10.1111/rda.13981

Koeck, A., Loker, S., Miglior, F., Kelton, D. F., Jamrozik, J., & Schenkel, F. S. (2014). Genetic relationships of clinical mastitis, cystic ovaries, and lameness with milk yield and somatic cell score in first-lactation Canadian Holsteins. *Journal of Dairy Science*, *97*(9), 5806–5813. https://doi.org/10.3168/jds.2013-7785

Kominakis, A. P., Abas, Z., Maltaris, I., & Rogdakis, E. (2002). A preliminary study of the application of artificial neural networks to prediction of milk yield in dairy sheep. *Computers and Electronics in Agriculture*, *35*(1), 35–48. https://doi.org/10.1016/S0168-1699(02)00051-0

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

Marumo, J. L., Lusseau, D., Speakman, J. R., Mackie, M., & Hambly, C. (2022). Influence of environmental factors and parity on milk yield dynamics in barn-housed dairy cattle. *Journal of Dairy Science*, *105*(2), 1225–1241. https://doi.org/10.3168/jds.2021-20698

Masía, F. M., Lyons, N. A., Piccardi, M., Balzarini, M., Hovey, R. C., & Garcia, S. C. (2020). Modeling variability of the lactation curves of cows in automated milking systems. *Journal of Dairy Science*, *103*(9), 8189–8196. https://doi.org/10.3168/jds.2019-17962

Naqvi, S. A., King, M. T. M., Matson, R. D., DeVries, T. J., Deardon, R., & Barkema, H. W. (2022). Mastitis detection with recurrent neural networks in farms using automated milking systems. *Computers and Electronics in Agriculture*, *192*, 106618. https://doi.org/10.1016/j.compag.2021.106618

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 485–492. https://doi.org/10.1145/2908812.2908918

Ozella, L., Brotto Rebuli, K., Forte, C., & Giacobini, M. (2023). A Literature Review of Modeling Approaches Applied to Data Collected in Automatic Milking Systems. *Animals: An Open Access Journal from MDPI*, *13*(12), 1916. https://doi.org/10.3390/ani13121916

Penasa, M., De Marchi, M., & Cassandro, M. (2016). *Short communication:* Effects of pregnancy on milk yield, composition traits, and coagulation properties of Holstein cows. *Journal of Dairy Science*, *99*(6), 4864–4869. https://doi.org/10.3168/jds.2015-10168

Pfeilsticker, H.-U., Bruckmaier, R. M., & Blum, J. W. (1996). Cisternal milk in the dairy cow during lactation and after preceding teat stimulation. *Journal of Dairy Research*, *63*(4), 509–515. https://doi.org/10.1017/S0022029900032040

Piwczyński, D., Sitkowska, B., Kolenda, M., Brzozowski, M., Aerts, J., & Schork, P. (2020). Forecasting the milk yield of cows on farms equipped with automatic milking system with the use of decision trees. *Animal Science Journal*, *91*. https://doi.org/10.1111/asj.13414

Sakallioglu, B. (2023). *A Study of Geometric Semantic Genetic Programming with Linear Scaling*. Submitted to NOVA Information Management School http://hdl.handle.net/10362/152098.

Sharipov, D. R., Yakimov, O. A., Gainullina, M. K., Kashaeva, A. R., & Kamaldinov, I. N. (2021). Development of automatic milking systems and their classification. *IOP Conference Series: Earth and Environmental Science*, *659*(1), 012080. https://doi.org/10.1088/1755-1315/659/1/012080

Stangaferro, M. L., Wijma, R., Caixeta, L. S., Al-Abri, M. A., & Giordano, J. O. (2016). Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part III. Metritis. *Journal of Dairy Science*, *99*(9), 7422–7433. https://doi.org/10.3168/jds.2016-11352

Tullo, E., Finzi, A., & Guarino, M. (2019). Review: Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy. *Science of The Total Environment*, *650*, 2751–2760. https://doi.org/10.1016/j.scitotenv.2018.10.018

Turiello, M. P., Vissio, C., Heinrichs, A. J., Issaly, L. C., & Larriestra, A. (2020). Impact of age at first calving on performance and economics in commercial dairy herds in Argentina. *Livestock Science*, *240*, 104108. https://doi.org/10.1016/j.livsci.2020.104108

Vanneschi, L., & Silva, S. (2023). *Lectures on Intelligent Systems*. Springer Nature.

Villettaz Robichaud, M., Rushen, J., de Passillé, A. M., Vasseur, E., Orsel, K., & Pellerin, D. (2019). Associations between on-farm animal welfare indicators and productivity and profitability on Canadian dairies: I. On freestall farms. *Journal of Dairy Science*, *102*(5), 4341–4351. https://doi.org/10.3168/jds.2018-14817

Vranken, E., & Berckmans, D. (2017). Precision livestock farming for pigs. *Animal Frontiers*, *7*(1), 32–37. https://doi.org/10.2527/af.2017.0106

Wood, P. D. P. (1967). Algebraic Model of the Lactation Curve in Cattle. *Nature*, *216*(5111), 164–165. https://doi.org/10.1038/216164a0

Wu, X.-L., Wiggans, G. R., Norman, H. D., Enzenauer, H. A., Miles, A. M., Van Tassell, C. P., Baldwin, R. L., Burchard, J., & Dürr, J. (2023). Estimating test-day milk yields by modeling proportional daily yields: Going beyond linearity. *Journal of Dairy Science*, *106*(12), 8979–9005. https://doi.org/10.3168/jds.2023-23479

APPENDIX A

Table A. 1 - Detailed description of every feature in the initial dataset

Variable name	Unity	Definition
Farm	-	Id of the farm
animal_id	-	Id of the cow within each farm
date_key_wall	-	Date of the recorded data
lactation_number	-	Number of the lactation period
days_in_lactation	days	Number of days within the lactation in that date, for that cow
animal_date_of_birth	-	Date of birth of the cow
pregnancy_length_in_days	days	Number of pregnancy days if the cow is pregnant in that date
animal_production_status	-	Production status of the cow on that day (Dryoff or InLactation)
animal_reproduction_status	-	Reproduction status of the cow on that day (ie. Pregnant, Inseminated, open, …)
average_body_weight_kg	Kg/day	Average body weight in kg
milk_production_kg	Kg/day	Kg of milk produced by day
Milkings	milkings/day	Number of successful milkings by day
milking_failures	milkings/day	Number of milkings failed by day
milking_refusals	milkings/day	Number of milkings refused by the robot for insufficient time interval
milk_fat_pct	%/day	100*fat_kg/milk_kg
milk_lactose_pct	%/day	100*lactose_kg/milk_kg
milk_protein_pct	%/day	100*protein_kg/milk_kg
milk_somatic_cells_count_kcells_per_ml	quantity/ml/day	Number of somatic cells in each ml of milk by day
milking_box_time_avg_sec	sec/day	Average number of seconds spent inside the AMS box in that day
milking_box_time_max_sec	sec/day	Maximum number of seconds spent inside the AMS box in that day

milking_box_time_min_sec	sec/day	Minimum number of seconds spent inside the AMS box in that day
activity_avg	min/day	Average number of minutes spent not resting, ruminating or sleeping
activity_max	min/day	Maximum number of minutes spent not resting, ruminating or sleeping
activity_min	min/day	Minimum number of minutes spent not resting, ruminating or sleeping
maximum_milk_flow_duration_of_the_teats_avg_sec	sec/day	Average number of seconds of the maximum duration of milk flow
maximum_milk_flow_duration_of_the_teats_max_sec	sec/day	Maximum number of seconds of the maximum duration of milk flow
maximum_milk_flow_duration_of_the_teats_min_sec	sec/day	Minimum number of seconds of the maximum duration of milk flow
duration_first_teat_con_till_last_teat_disconnected_avg_sec	sec/day	Average number of seconds from the connection of the first teat until the disconnection of the last teat within the milking
duration_first_teat_con_till_last_teat_disconnected_max_sec	sec/day	Maximum number of seconds from the connection of the first teat until the disconnection of the last teat within the milking
duration_first_teat_con_till_last_teat_disconnected_min_sec	sec/day	Minimum number of seconds from the connection of the first teat until the disconnection of the last teat within the milking
maximum_milk_conductivity		Maximum milk conductivity
average_milk_speed_avg	kg/min	Average of the average number of kg of milk produced by minute in each milking while the cow is being milked
average_milk_speed_max	kg/min	Average of the maximum number of

		kg of milk produced by minute in each milking while the cow is being milked
average_milk_speed_min	kg/min	Average of the minimum number of kg of milk produced by minute in each milking while the cow is being milked
milk_yield_per_box_time_avg_kg_per_min	kg/min	Average of the number of kg of milk produced by minute in each milking while the cow is the AMS
milk_yield_per_box_time_max_kg_per_min	kg/min	Average of the average number of kg of milk produced by minute in each milking while the cow is the AMS
milk_yield_per_box_time_min_kg_per_min	kg/min	Average of the maximum number of kg of milk produced by minute in each milking while the cow is the AMS
teat_milk_analysis_result_watery_flag	kg/min	Average of the minimum number of kg of milk produced by minute in each milking while the cow is the AMS
teat_milk_analysis_result_mastitis_flag	-	Flag indicating the presence of mastitis
teat_milk_analysis_result_blood_flag	-	Flag indicating the result of blood analysis
teat_milk_analysis_result_colostrum_flag	-	Flag indicating the result of colostrum analysis
teat_milk_analysis_result_abnormal_flag	-	Flag indicating the presence of abnormalities
animal_sick_chance	%/day	Chance percentage of the cow being sick
rumination_minutes	min/day	Number of minutes where the cow was rumination by day
eating_minutes	min/day	Number of minutes where the cow was eating by day
isk_individual_standardized_milk_day_production	-	Metric used to compare milk yield between cows
animal_max_heat_chance	%/day	Maximum chance percentage of the cow being too hot
milking_visit_time_since_last_successful_milking_max_sec	sec	Maximum number of seconds between two

		consecutives successful milkings in that day
milking_visit_time_since_last_successful_milking_min_sec	sec	Minimum number of seconds between two consecutives successful milkings in that day

Table A. 2 - Number of observations in each target week and lactation group considered

	Lactation 1	**Lacation_2_3_4_5**
Week 1	215289	481003
Week 2	208420	468273

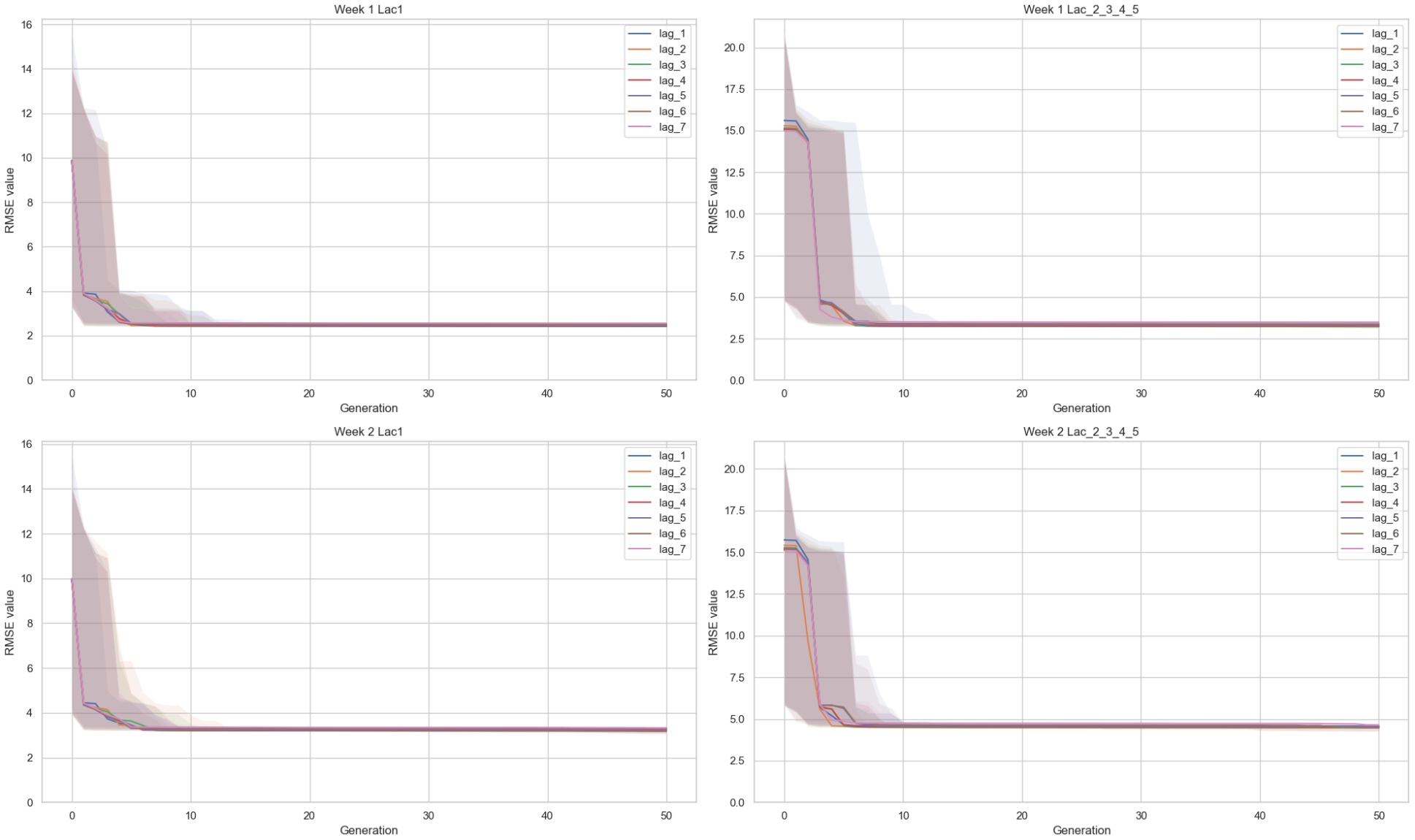

Figure A. 1 - GP evolution through 50 generation for both week predictions and lactation groups

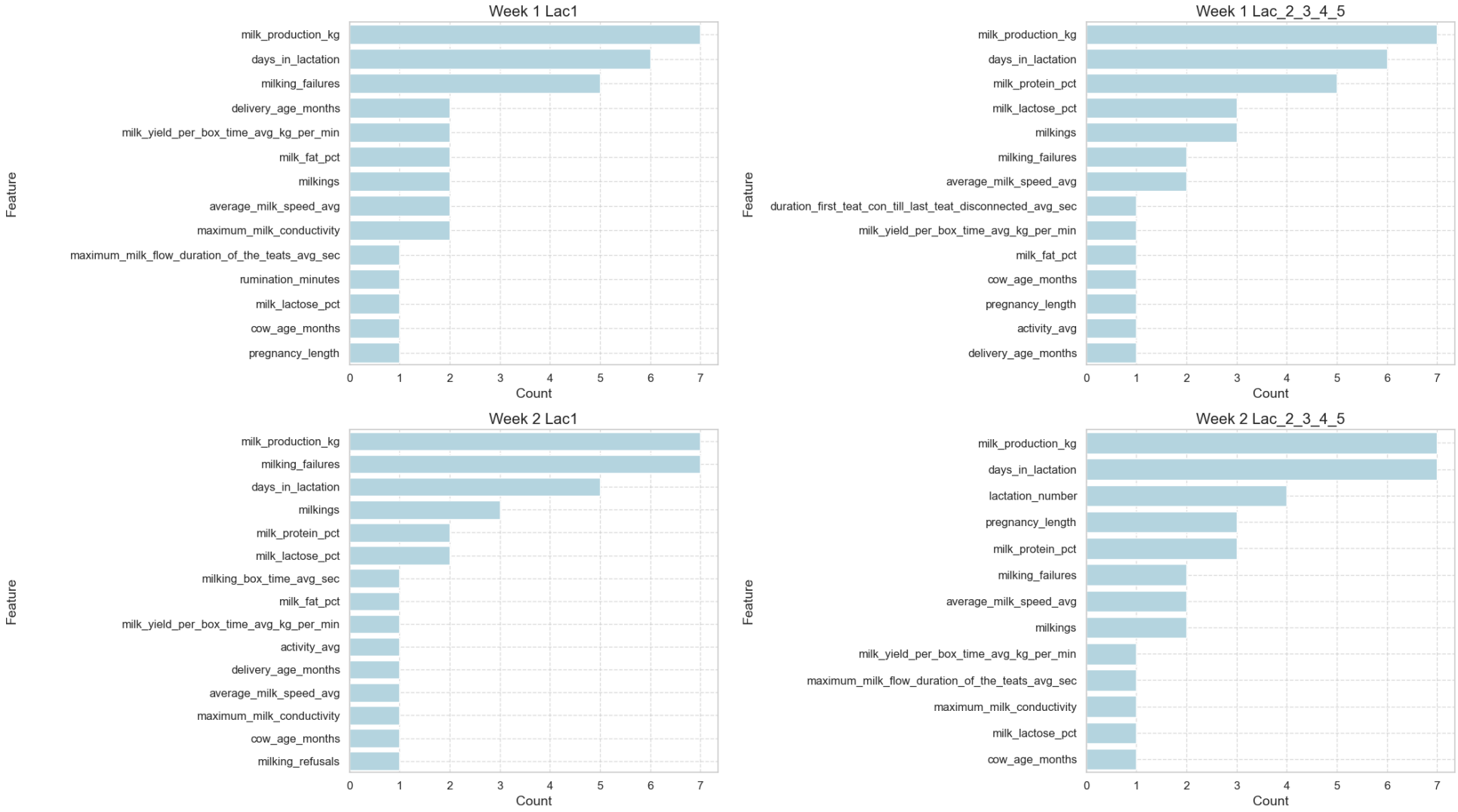

Figure A. 2 - Features that significantly contributed trees' models grouped by both week predictions and lactation groups

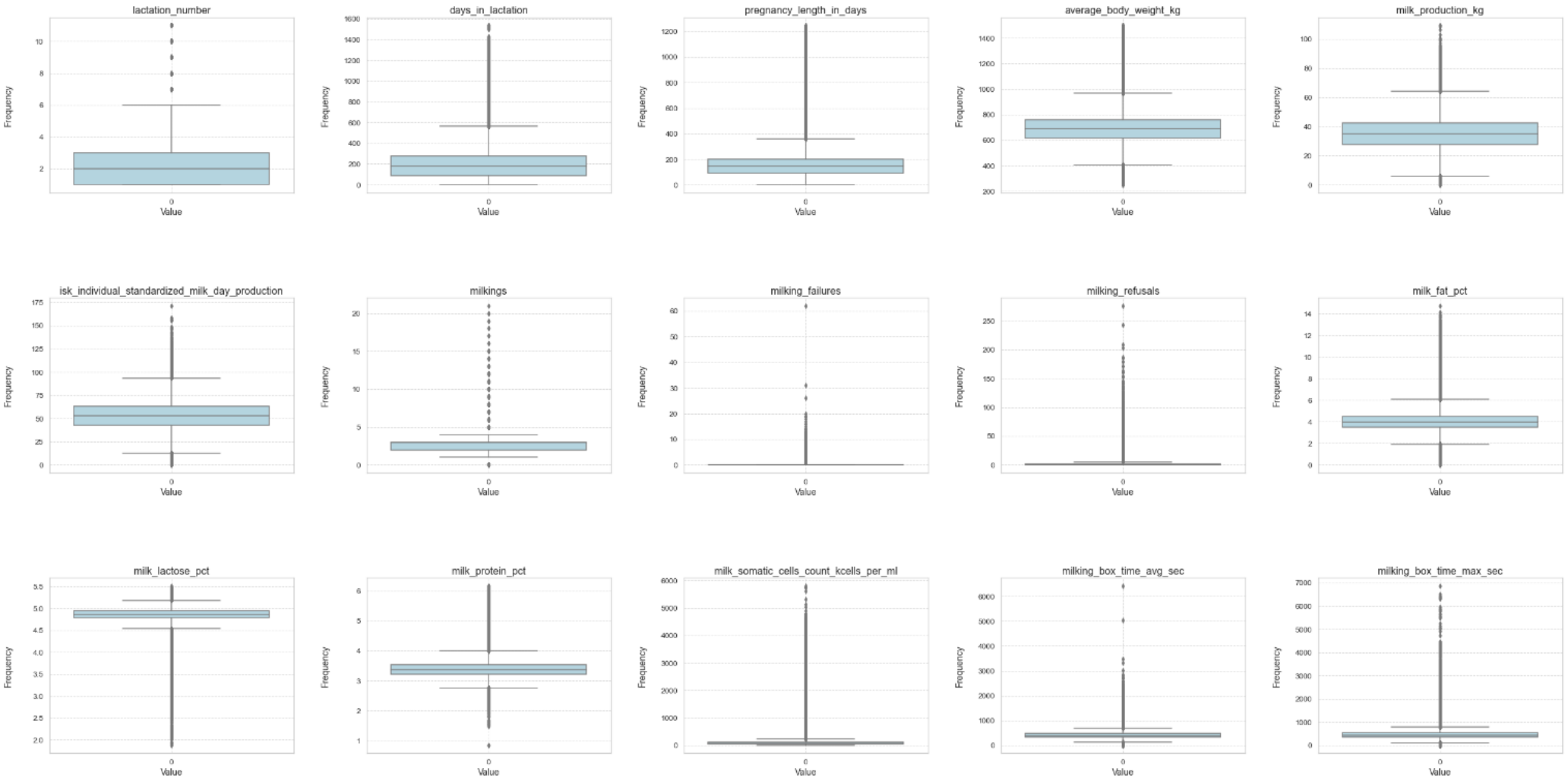

Figure A. 3 - Numerical variables distribution before preprocessing (1)

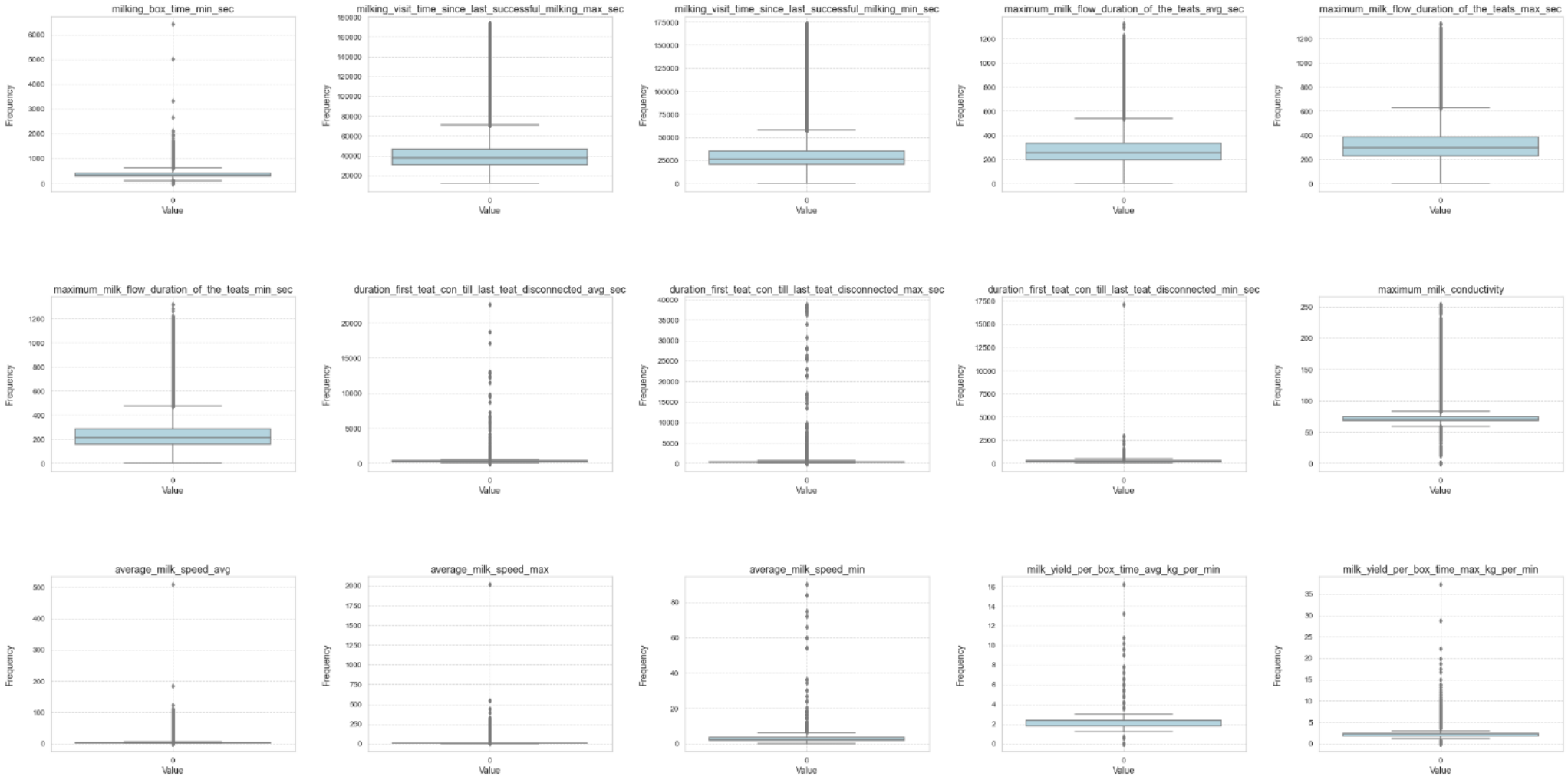

Figure A. 4 - Numerical variables distribution before preprocessing (2)

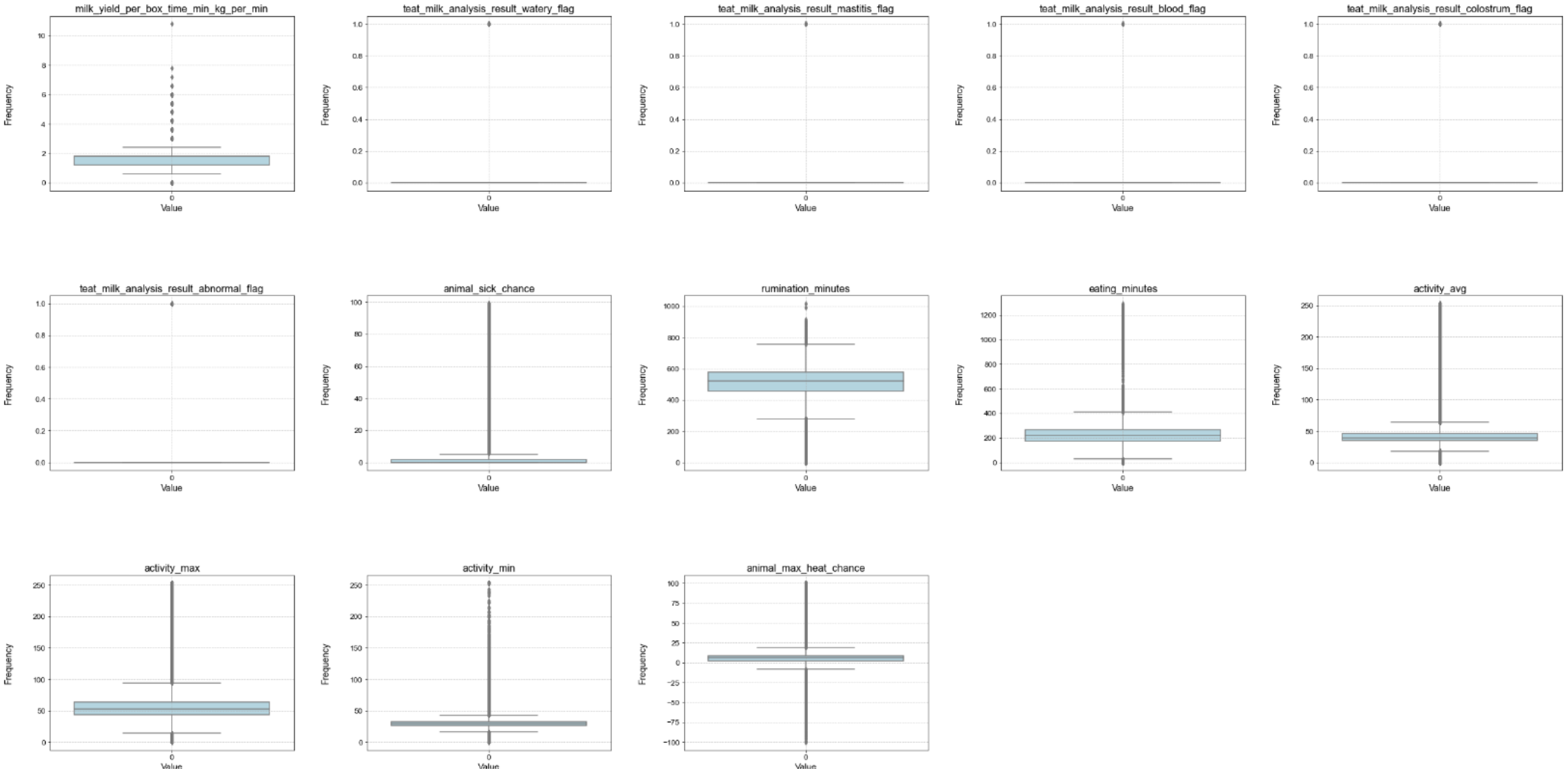

Figure A. 5 - Numerical variables distribution before preprocessing (3)

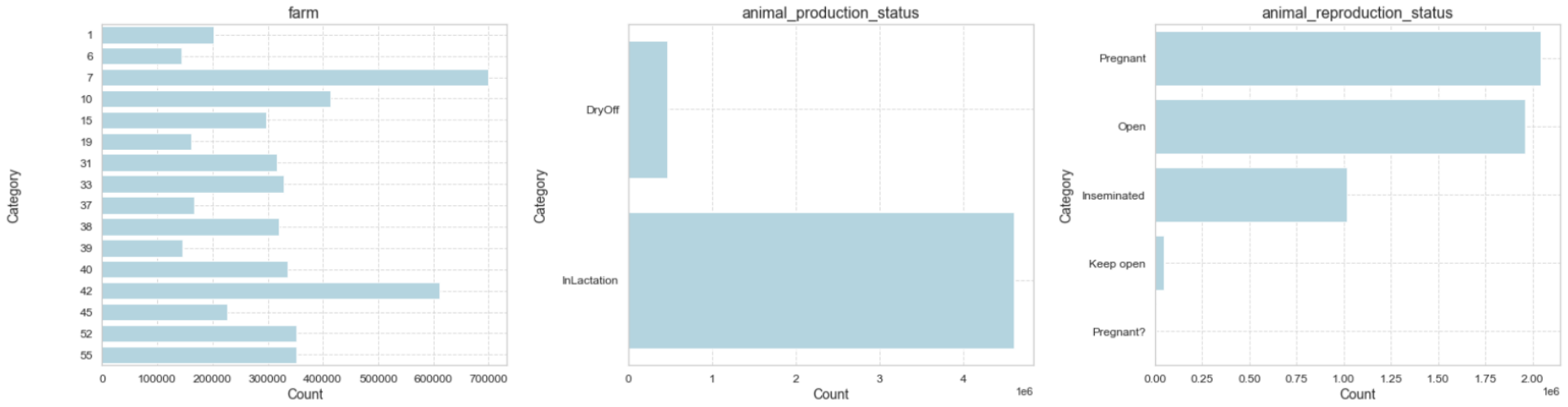

Figure A. 6 - Categorical variables frequency before preprocessing

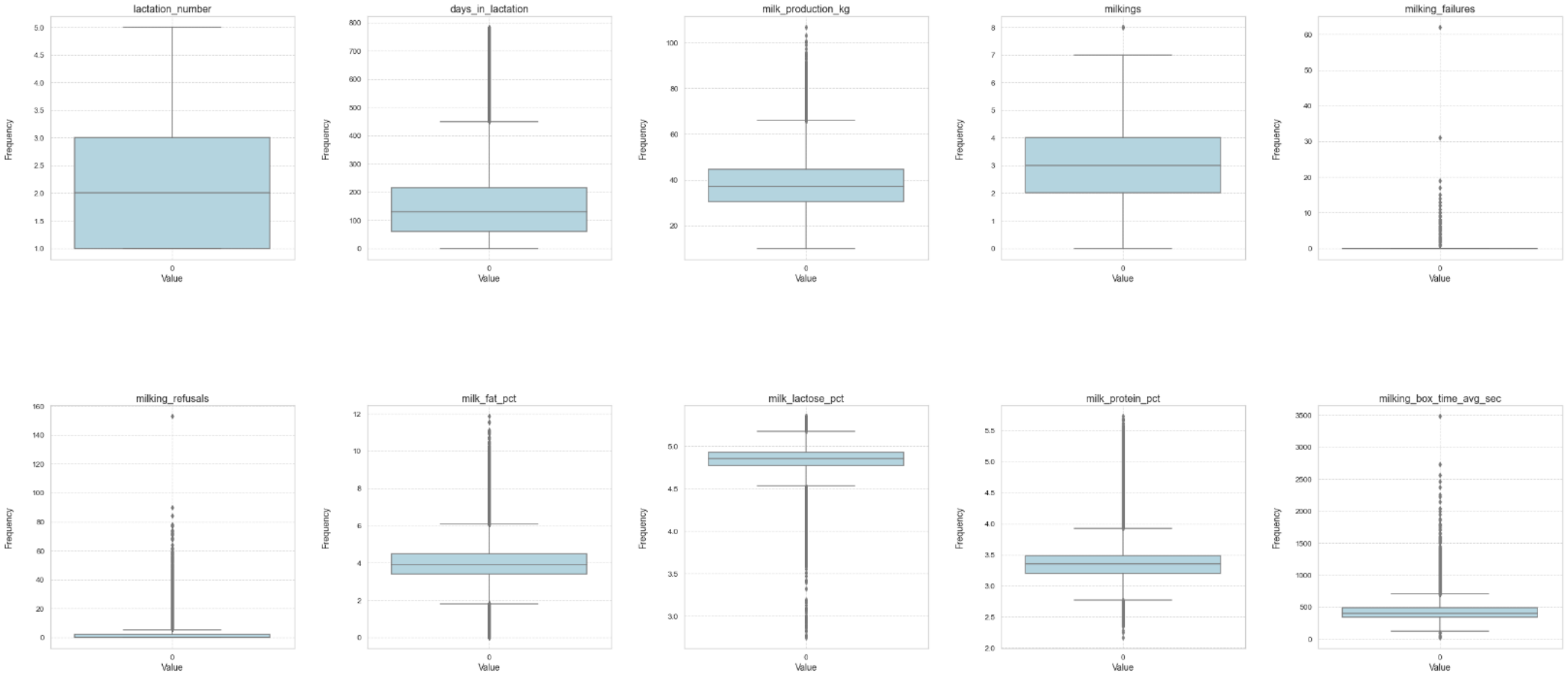

Figure A. 7 - Variables distribution after preprocessing (1)

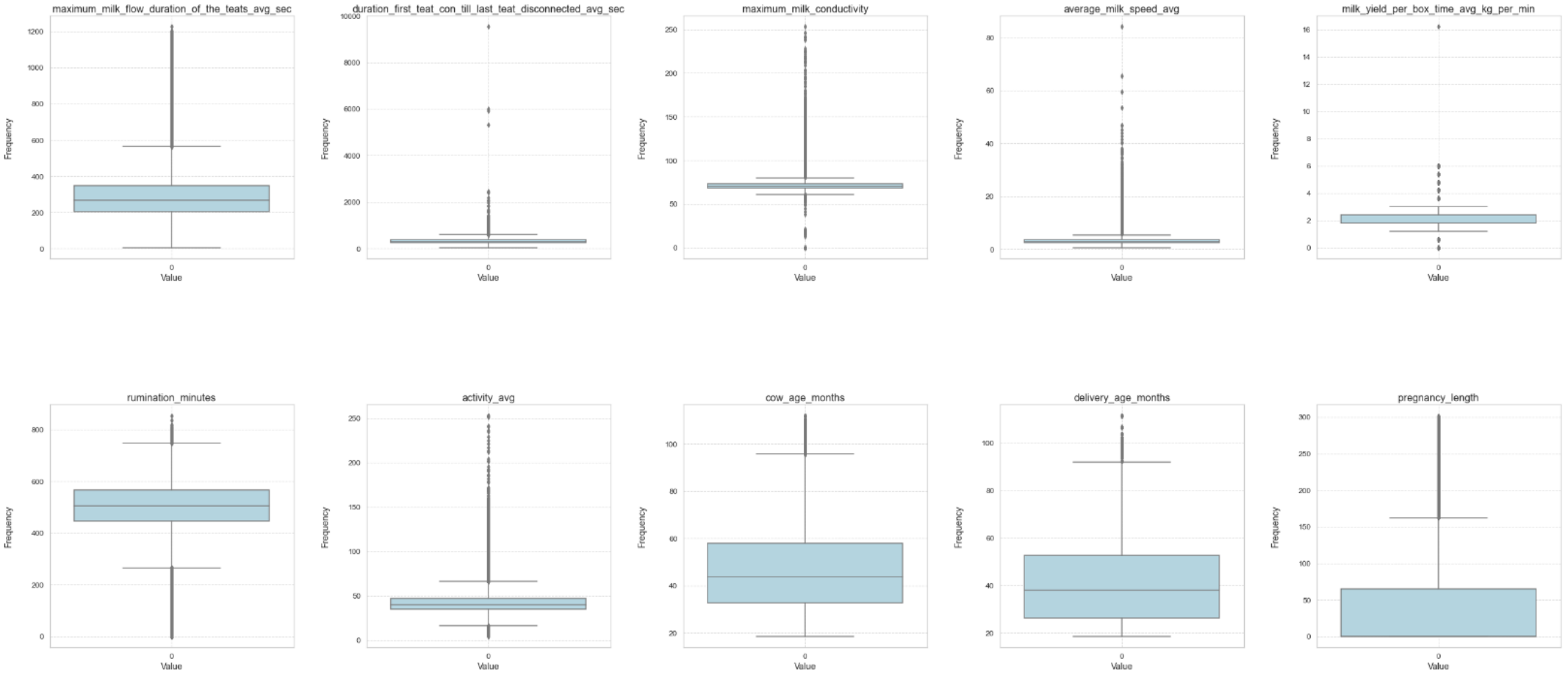

Figure A. 8 - Variables distribution after preprocessing (2)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa