University of Melbourne

Faculty of Sciences

Data Science Project

Big Data in Dairy

Authors:
Harsh Mangla 1418017
Yuqian Xie 1448428
Xuheng Li 1405559
Kaifan Ouyang 1427957

Host:
Kristy Digiacomo
Supervisor:
Mario Andres Munoz Acosta

A capstone project submitted in fulfillment of the requirements
for the degree of Master of Data Science
in the
Faculty of Science

October 25, 2024

Declaration

I certify that this report does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 7668 words in length (excluding text in images, tables, bibliographies and appendices).

Signed: Harsh M

Date: 24 / 10 / 2024

Signed: Yuqian Xue

Date: 24 / 10 / 2024

Signed: Kaifan Ouyang

Date: 24 / 10 / 2024

Signed:

Date: 24 / 10 / 2024

Abstract

This project focused on analyzing milk production data to help optimize dairy farm operations. The main objective was to identify relationships between key variables such as feed consumption, milking speed, and milk yield, and to develop predictive models that could support better decision-making in dairy production. We worked with data from the Herd Daily and Device Daily datasets. To analyze these datasets, we conducted a time series analysis using the SARIMA model to forecast milk production trends over the next decade, accounting for seasonal fluctuations and long-term patterns, along with polynomial regression. Additionally, we applied machine learning techniques like Random Forest and Gradient Boosting to build strong predictive models. This combination of statistical and machine learning methods provided valuable insights for improving efficiency in dairy farming.

To access our project repository, logs, and recordings of all meetings, please refer to the Appendix section.

Acknowledgments

We would like to express our gratitude to our host, Dr. Kristy Digiacomo, for her invaluable guidance and unwavering support throughout our capstone project. Her expertise and encouragement have been vital to the successful completion of this project.

We would also like to sincerely thank our supervisors, Dr. Mario Andres Munoz Acosta and Ms. Fareha (Ph.D. student), for their continuous advice and mentorship. Their insights and constructive feedback have been critical in shaping the direction of our project.

Additionally, we are very grateful to Dr. Joyce Zhang for her advice and support during the meetings. Her advice has significantly enhanced the quality of our work.

Thank you.

Contents

1 Introduction

Dairy farming, particularly in cow production systems, generates vast amounts of data, which is often underutilized and not properly integrated. As the agricultural sector adopts big data analytics and automation, the potential for significant advancements becomes clear. Traditionally, dairy farming has been labor-intensive, but it is now undergoing trans-formative changes with the introduction of technologies such as automated milking systems and advanced data analysis tools. The aim of this project was to explore data from dairy farms, with a focus on connecting different datasets to uncover new patterns and key factors. We also analyzed the critical relationships and drivers that influence production. The ultimate goal is to use these insights to help farmers make more informed decisions and enhance the efficiency of dairy farm operations.

1.1 Lely Horizon

We used Lely Horizon software to extract data. Lely Horizon is a farm management tool that provides detailed reports, helping us make better decisions to improve the efficiency of farm operations.

1.2 Challenges Faced

We faced several challenges with data collection and analysis. One of the main problems was the difficulty in extracting data from the Lely Horizon software. This was because the user interface lacked a clear manual, making it hard to use. As a result, the setup and data download took longer than expected. Even after combining different datasets, the total amount of data we had was less than we thought. Lastly, our data models didn't perform as well as we hoped, mostly due to the small dataset.

1.3 Project Workflow

After collecting and processing the data, and performing exploratory data analysis (EDA) to gain a basic understanding of the dataset, the project moved through three key stages. First, we created detailed installation guides to help future users easily access and work with the data. Next, we focused on data modeling, testing various models, and making adjustments to improve their performance. Finally, we worked on data visualization, creating clear plots to interpret the results. Throughout the process, we maintained regular contact with our host and supervisor and compiled the final report, which summarized the entire project. The complete workflow is illustrated in Figure 1.

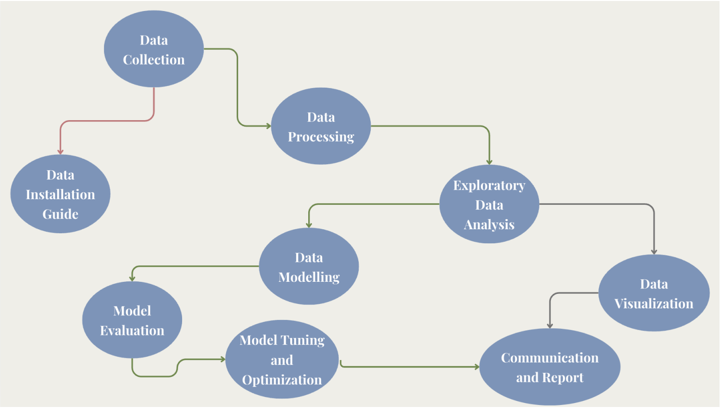


Figure 1: Project Workflow

2 Literature Review

In order to increase milk output and farm management efficiency, big data and automation in dairy farming have been extensively studied. Past and new studies show that data-driven technologies can solve dairy output and decision-making problems.

2.1 Decision Support Systems in Dairy Farming

Ferris et al. [1] and Cabrera et al. [2] created the Dairy Brain decision support system. It integrates farm data using the Agricultural Data Hub. Prescriptive, predictive, and descriptive analytics help dairy farmers choose nutritional groups and detect disease early. Its modular design makes it easy to combine data sources, speeding up production. This project combines data to help dairy farmers make better decisions. This supports that goal.

2.2 Precision Dairy Farming Technologies

Modern technologies like automatic estrous detection and daily milk yield tracking track physiological and behavioral data in precision dairy farming. Wolfert et al. [3] say IoT and cloud computing in smart farming make real-time decision-making easier and provide more future information. This method improves dairy operations by controlling resources and detecting diseases earlier. Precision farming uses data to improve animal health and output, like this project.

2.3 Predictive Models for Milk Yield

Milk yield predictions are now more accurate thanks to predictive models. Liseune et al. [4] This framework was more accurate than existing models, especially early in lactation. They find complex yield patterns using latent representations. This simplifies herd management and helps find diseases early by comparing real yields to predicted yields. Grzesiak et al. [5] found that ANNs predicted daily milk yields better than Wood's lactation curve model. This is because ANNs can model complex, non-linear data relationships. These studies demonstrate the importance of advanced machine learning for milk yield predictions. They also demonstrate the importance of big datasets for accuracy, a problem this project also had.

2.4 Machine Learning and Deep Learning in Dairy Farming

ML is well-known for analyzing large dairy farming datasets. According to Araújo et al. [6] combining cloud computing, IoT, and machine learning enables real-time decision support systems that enhance farming efficiency. Liseune et al. [4] showed that deep learning models, especially neural networks, can make correct predictions about milk yield throughout the lactation cycle. This improves herd management and productivity directly. In the same way, Grzesiak et al. [5] discovered that artificial neural networks (ANNs) did better than traditional regression models, even with smaller datasets. This shows how flexible and accurate they are in dairy farming.

2.5 Health Monitoring and IoT Applications

Health tracking is an important part of dairy farming that can be made a lot better with data driven methods. Andonovic et al. [7] talked about how wireless sensor networks (WSNs) can be used to keep an eye on the health of animals all the time. WSNs are a cheap and real-time way to keep an eye on things. This fits with the project's goal of combining data on health and output to make a full picture of how farms work. Also, IoT-based solutions looked at in other studies have shown to help find health problems early, which leads to better control and better animal welfare.

2.6 Challenges in Data Integration and Modeling

Data-driven approaches could help dairy farming, but it's hard to combine and maintain data quality from different sources. Grzesiak et al. [5] and Liseune et al. [4] emphasize the importance of using large, high-quality datasets to train effective forecasting models. Small datasets make underlying relationships harder to see and model forecasts less accurate. The same data availability and size issues plagued this project, reducing model accuracy. Solve these issues to maximize data-driven dairy farming.

3 Data Preprocessing

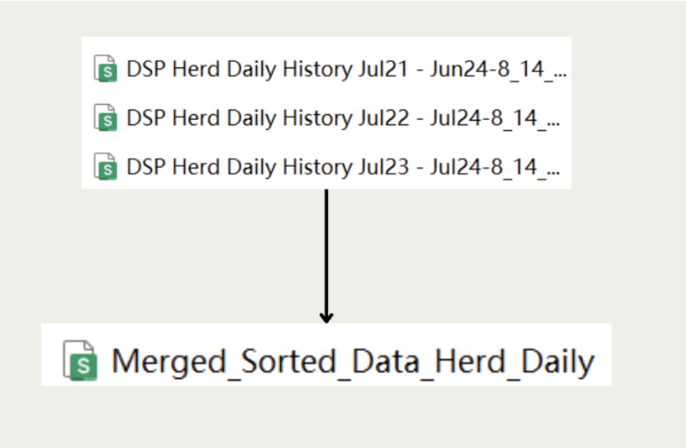
Preprocessing data prepares it for modeling and analysis. It verifies data for processing. Before the research, **Herd Daily Data** and **Device Daily Data** were checked for consistency and accuracy.

3.1 Data Collection and Merging

Multiple Excel files with historical data for both Device Daily Data and Herd Daily History were joined to make a single dataset:

- **Loading and Merging:** For each dataset, a list of file paths was made, and `pd.read_excel()` was used to load the files into pandas DataFrames. With `pd.concat()`, the DataFrames were then joined together to make a single collection.
- **Date Handling:** The Date field was changed to `datetime` format and then sorted to make sure it was in the right order.
- **Review and Output:** After checking to make sure they were correct, the merged datasets were saved as new Excel files so they could be analyzed further.

This process ensured that all available data was integrated into a single, organized dataset, ready for subsequent analysis and modeling. (See Figure 2)



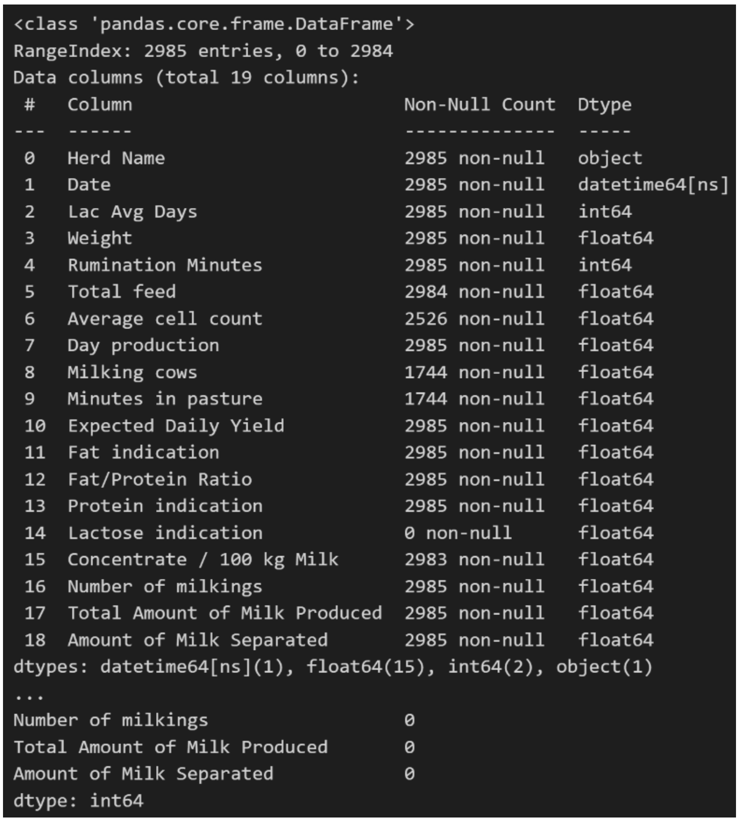
DSP Herd Daily History Jul21 - Jun24-8_14_...
DSP Herd Daily History Jul22 - Jul24-8_14_...
DSP Herd Daily History Jul23 - Jul24-8_14_...

Merged_Sorted_Data_Herd_Daily

Figure 2: Merging and Sorting Herd Daily History Data from Multiple Files

3.2 Data Loading and Initial Inspection

Both datasets were loaded using `pandas`, which made it easier to work with and analyze the data. During the first checks, short descriptions were made using `data.info()` and `data.describe()` to learn about the structure, types, and distribution of values in each dataset. This step helped find important details like *Milk Speed Avg.*, *Milk Duration*, *Milk Exp.*, and *Milk Tot.*. Also, checks for missing numbers ((`data.isnull().sum()`) showed how big the data gaps were, which helped with cleaning up the data that came after. (See Figure 3)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2985 entries, 0 to 2984
Data columns (total 19 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Herd Name                    2985 non-null   object
 1   Date                         2985 non-null   datetime64[ns]
 2   Lac Avg Days                 2985 non-null   int64
 3   Weight                       2985 non-null   float64
 4   Rumination Minutes           2985 non-null   int64
 5   Total feed                   2984 non-null   float64
 6   Average cell count           2526 non-null   float64
 7   Day production               2985 non-null   float64
 8   Milking cows                 1744 non-null   float64
 9   Minutes in pasture           1744 non-null   float64
 10  Expected Daily Yield         2985 non-null   float64
 11  Fat indication               2985 non-null   float64
 12  Fat/Protein Ratio            2985 non-null   float64
 13  Protein indication           2985 non-null   float64
 14  Lactose indication           0 non-null      float64
 15  Concentrate / 100 kg Milk    2983 non-null   float64
 16  Number of milkings           2985 non-null   float64
 17  Total Amount of Milk Produced  2985 non-null   float64
 18  Amount of Milk Separated     2985 non-null   float64
dtypes: datetime64[ns](1), float64(15), int64(2), object(1)
...
Number of milkings               0
Total Amount of Milk Produced    0
Amount of Milk Separated         0
dtype: int64
```

Figure 3: Data Loading and Initial Inspection Summary for Dairy Herd Dataset

3.3 Handling Missing Values

Some important sections, like *Milk Speed Avg.*, *Milk Exp.*, and *Milk Tot.*, were found to be empty. Either the missing numbers were filled in based on what was known about the domain, or records that were missing important data points were thrown away. One way to keep the quality of the dataset good was to get rid of data points that were missing dates or important metrics.

3.4 Data Type Conversions

Before they could be analyzed, entries in some fields, like *Robot Address*, that weren't numbers had to be changed. `pd.to_numeric()` was used with the `errors='coerce'` parameter to change these columns to numeric data types and set any values that didn't work to `NaN`. This change was very important later on for modeling and doing numerical processes.

3.5 Feature Engineering

Some important information, like the *number of feed visits*, was taken out and fixed to make the dataset better for models. For instance, the *Date* field was changed to *datetime* to make it easy to do studies that are based on time. It is now possible to group and combine data over time, which is important for seeing how milk output changes over time.

3.6 Data Normalization and Scaling

Normalization was done to make sure that things like *Milk Speed Avg.* and *Milk Exp.* were all on the same level. This step was meant to reduce the difference between the different traits. It would help machine learning models work better because no feature would be able to get too big and take over.

3.7 Data Integration

After the different datasets were preprocessed, the cleaned data was put together to make a single view that could be used for further research. The *Robot Address* and *Date* were used as shared keys for this integration, which let the Herd Daily and Device Daily datasets be combined. The next steps, exploratory data analysis (EDA) and model development, were based on this shared information.

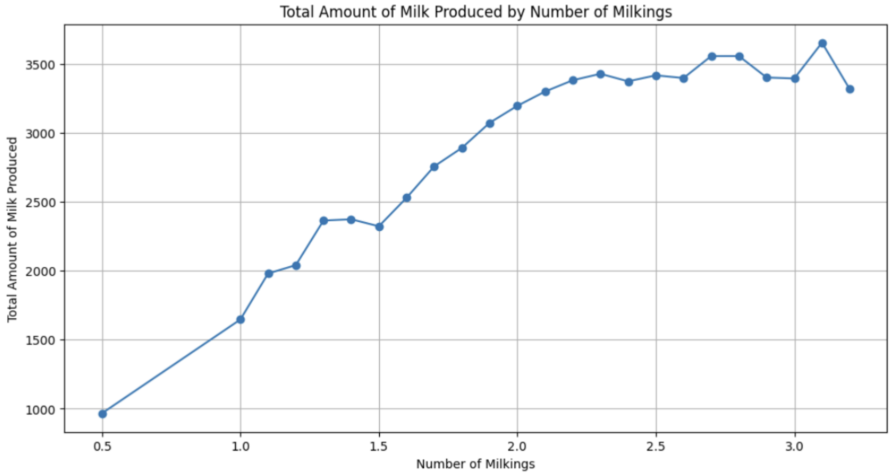
4 Exploratory Data Analysis

After doing necessary data preprocessing steps, we move on forward with exploratory data analysis in search of finding relationships and trends in various fields inside the data. We have two different templates for datasets that we will be working with, one is “Herd Daily” and the other one is “Device Daily”.

4.1 Herd Daily Dataset

4.1.1 Relationship Between Number of Milkings and Total Milk Production

The plot shows that more frequent milking sessions result in better yields, as evidenced by the steady increase in milk output, especially between 0.5 and 1.5 milkings. However, production reaches a plateau after two or three milkings, indicating that the cows are producing as much as they can. Production barely changes after 2.5 milkings, likely due to factors such as cow fatigue or differences in milking efficiency.
To put it briefly, while more frequent milkings result in an initial rise in output, beyond around 2.5 sessions, the amount of milk produced doesn’t really increase. This probably represents the cows’ natural boundaries, so more sessions don’t really add anything. The information points to the possibility that, after a certain period, biological or environmental variables could be reducing milk production.

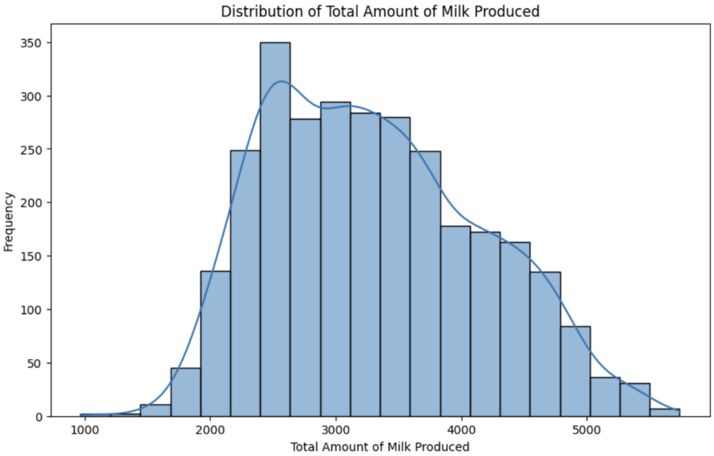


4.1.2 Distribution of Total Milk Production

The histogram illustrates the distribution of the total amount of milk produced, highlighting how often different levels of milk production occur.

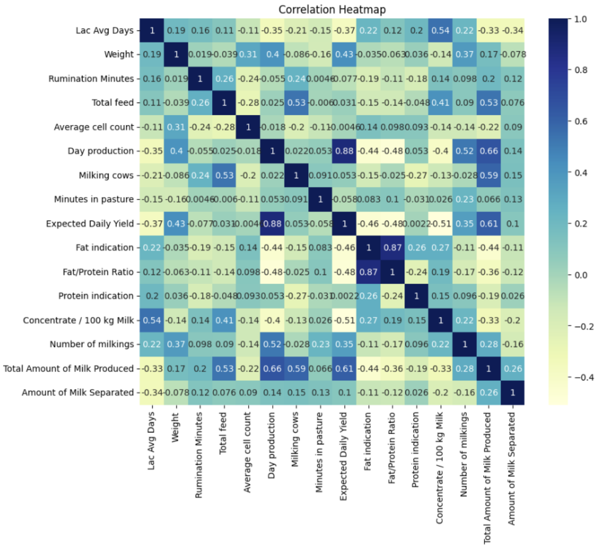
The data follows a uni-modal, bell-shaped distribution with a slight right skew. Most cows or herds fall within the production range of 2500 to 3000 liters, while fewer produce at both lower and higher levels. Although the majority of milk production occurs between 2500 and 4000 liters, the rightward skew suggests that a few instances of significantly higher production exist, possibly reflecting very productive cows or favorable conditions.

In summary, most milk production is concentrated in the 2500 to 3000-liter range, with a small number of outliers producing much more. These anomalies highlight the potential for exceptionally high production under certain circumstances.



4.1.3 Correlation Between Variables

The correlation matrix for the several variables in the dataset is represented by this heatmap, which demonstrates the relationships between the variables. The right-hand color scale goes from -1, which represents a strong negative correlation, to 1, which represents a high positive correlation. The main relationships are broken down as follows:



Many key factors influencing milk production are highlighted in the data. A strong correlation exists between daily output, total milk production, and the estimated daily yield, showing that higher daily production contributes to greater overall milk output and aligns with more accurate yield predictions. Additionally, both the number of milkings and the number of cows milked positively impact total production, suggesting that larger herds and more frequent milking sessions lead to increased output.

Furthermore, there is a positive relationship between the amount of concentrate fed per 100 kg of milk and both cow weight and daily production, indicating that better nutrition results in larger cows and higher milk yields. The significant correlation between improved yield estimates and fat markers, such as the fat-to-protein ratio, underscores the importance of fat content in determining milk quality. On the other hand, lactation days show a negative correlation with both total milk output and daily production, suggesting that milk production tends to decrease as the lactation period extends.

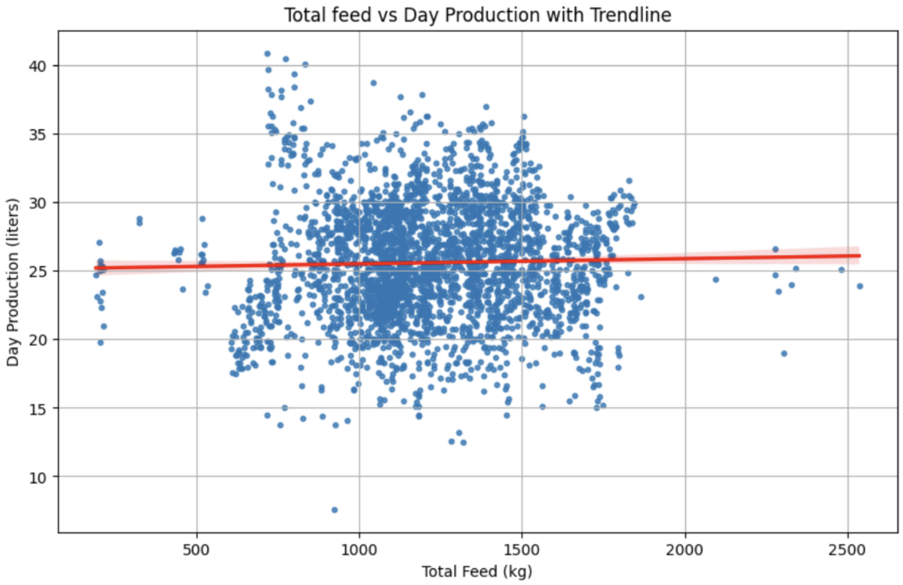
In conclusion, key drivers of milk production include the number of milkings, herd size, and daily output levels. Feed quality and fat content also play a critical role, while extended lactation periods seem to gradually reduce productivity.

4.1.4 Relationship Between Total Feed and Day Production

This scatter-plot illustrates the relationship between Total Feed (kg) on the X-axis and Day Production (liters) on the Y-axis, with a red linear regression trend-line.

The graphic shows that the amount of feed fed and the amount of milk produced each day are not strongly correlated. The majority of the data points, which range from 500 to 2000 kg of feed and daily production levels of 20 to 30 litres, are fairly dispersed. The trend line is almost flat, indicating that daily milk production is not much impacted by changes in feed quantity.

Although there is a modest upward slope to the trend line, indicating a slight positive correlation, this effect is small and probably not significant. All things considered, the evidence points to other variables having a bigger impact on daily milk production than feed quantity.

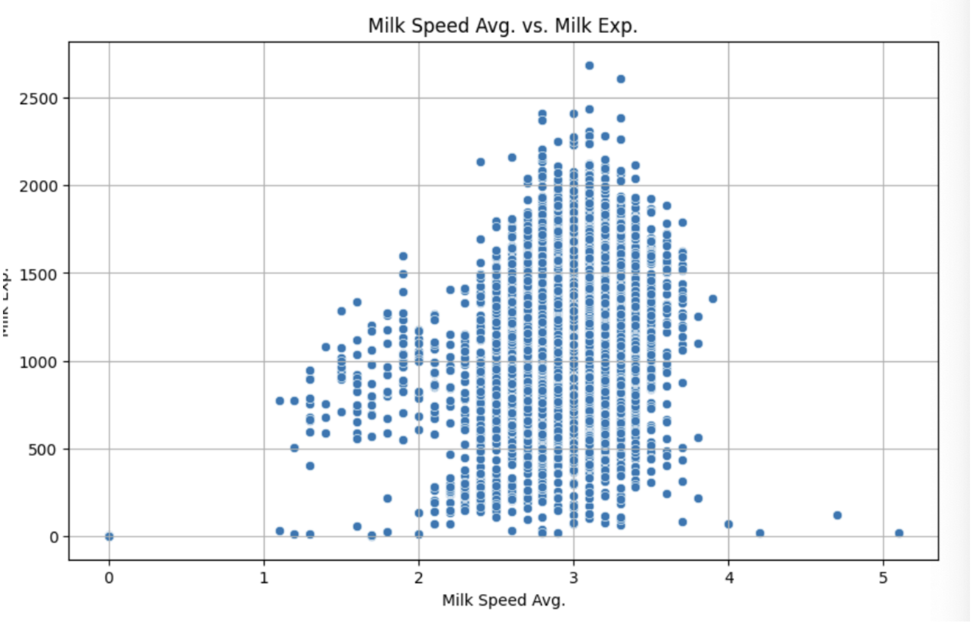


4.2 Device Daily Dataset

4.2.1 Relationship Between Milking Speed and Milk Production

A red linear regression line shows the trend in this scatter plot, which shows the link between daily milk production in litres (on the Y-axis) and feed amount in kilogrammes (on the X-axis). The graphic shows that the amount of milk produced each day and the amount of feed given are not strongly correlated. The majority of the data points fall between the daily production of 20–30 litres of milk and 500–2000 kg of feed. The trend line is almost flat, indicating that daily milk production is not much impacted by changes in feed quantity.

This effect is small and probably not very important, even if the trend line slopes slightly higher, indicating a weak positive association. Overall, the data suggests that other factors play a more important role in determining milk yield, with feed quantity having little to no influence on daily milk production.



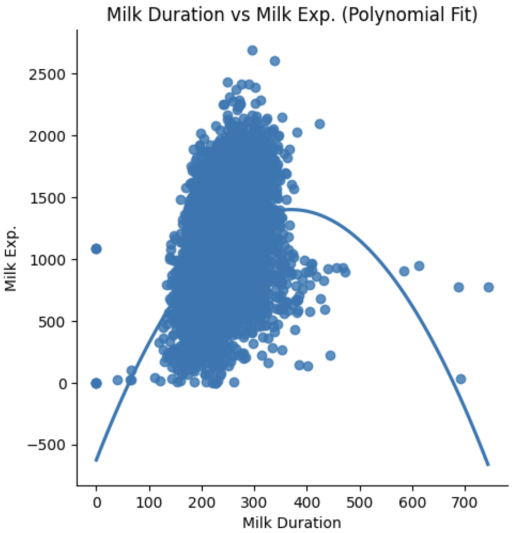
4.2.2 Relationship Between Milking Duration and Expected Milk Production

This plot depicts the relationship between Milk Duration (X-axis) and Milk Exp. (Expected Milk Production) on the Y-axis, with a polynomial regression curve overlaid.

The analysis reveals a second-order polynomial (parabolic) relationship between milking duration and expected milk production. Initially, as milking time increases, milk output also rises, reaching a peak at around 300 seconds. Beyond this point, the curve begins to decline, indicating diminishing returns or even reduced milk production with longer milking sessions.

The results suggest that extending milking sessions increases milk output up to approximately 300 seconds. However, when milking times exceed this threshold, predicted yields begin to decrease, possibly due to inefficiencies or overmilking. Additionally, there are a few outliers at both very short and long durations, where milk production is unexpectedly low or even negative, potentially indicating anomalies or data quality issues.

In summary, the optimal milking duration appears to be around 300 seconds, maximizing milk output. Beyond this point, longer sessions tend to lead to lower yields, likely due to the diminishing benefits of extended milking times.



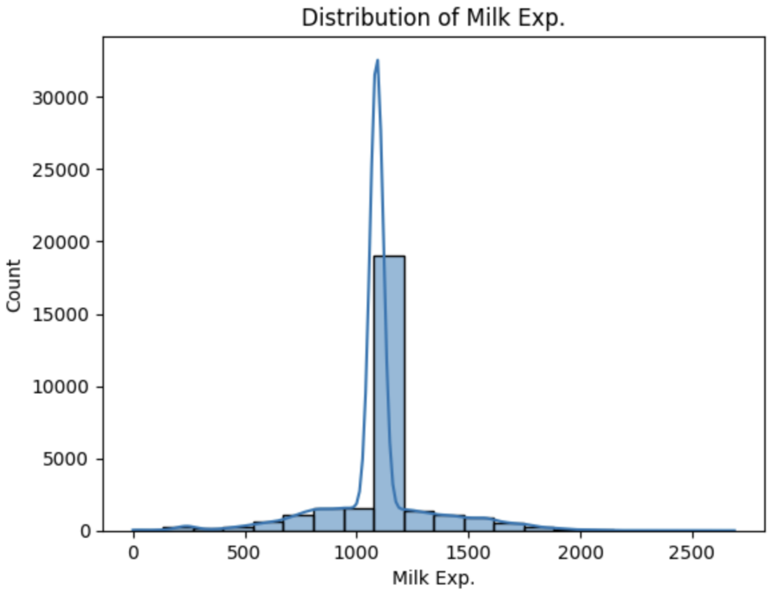
4.2.3 Distribution of Expected Milk Production

This histogram visualizes the distribution of Milk Exp. (Expected Milk Production), highlighting the frequency of different production values.

The distribution is sharply peaked with a narrow spread, indicating that most values are concentrated within a small range. While there are few cases above 1500 liters, the data is right-skewed, with a long tail extending toward higher production values. The majority of milk production values, nearly 30,000 counts, are centered around 1000 liters, suggesting that most expected production levels are closely clustered around this figure.

At the lower and upper ends of the distribution, there are fewer occurrences, with small tails below 500 liters and above 1500 liters, indicating that extremely low or high milk outputs are less common.

In summary, the data shows that the majority of cows or milking sessions yield around 1000 liters of milk, with limited variation beyond this central range. The right-skewed distribution and absence of extreme outliers reinforce the conclusion that predicted milk production is highly concentrated near the 1000-liter mark.

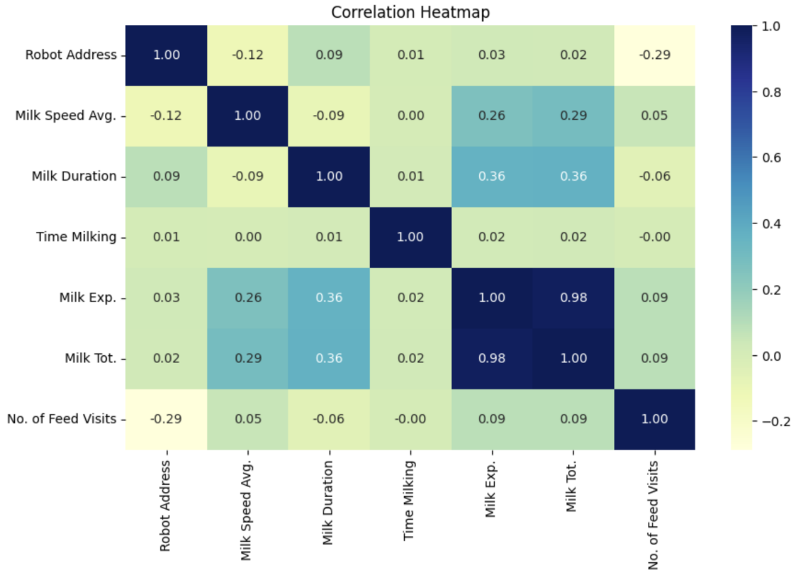


4.2.4 Correlation Analysis of Key Variables

This correlation heatmap illustrates the relationships between several variables in the dataset. The strength and direction of the correlations are represented by the color scale on the right, ranging from -1 (strong negative correlation) to 1 (strong positive correlation).

The analysis reveals a near-perfect positive correlation (0.98) between Milk Exp. (Expected Milk) and Milk Tot. (Total Milk), confirming that these two variables are closely aligned, as expected. Additionally, Milk Duration shows a moderate positive correlation (0.36) with both Milk Tot. and Milk Exp., indicating that longer milking sessions are associated with increased milk production. Similarly, Milk Speed Avg. has a moderate positive correlation (0.26), suggesting that faster milking speeds contribute to a slight but noticeable increase in milk output.

Interestingly, there is almost no significant correlation between milk production and the number of feed visits or milking duration, indicating that these factors have little influence on overall milk volume. Additionally, weak associations were observed between the specific robot used for milking (Robot Address) and other variables, suggesting that the robot assigned to the process has minimal impact on production.



5 Modeling

5.1 Time Series

Time series analysis involves studying data points collected or recorded at specific time intervals to identify patterns, trends, and potential seasonal effects over time. This method is widely used in various fields such as economics, finance, and environmental science to forecast future values based on historical data. The key components of time series data include trend, seasonality, and random fluctuations. A common approach to modeling time series data is the ARIMA (AutoRegressive Integrated Moving Average) model, which combines autoregression, differencing (to make the data stationary), and moving averages for forecasting purposes [8].

In our project, the sample spans from April 29, 2026 to June 30, 2024. We decomposed the time series "Total Amount of Milk Produced" using an additive model with a 365-day period. In this model, the total milk produced on any given day, $Y(t)$, is broken down into three parts: $T(t)$, which represents the overall trend in production over time, such as whether the yearly output is increasing or decreasing; $S(t)$, which captures the seasonal effect, showing the repeating yearly cycle; and $R(t)$, which accounts for the residual variation, or the random noise and unexplained changes in the data. This approach helped us better understand the different components influencing milk production.

Given the time series plot below, it shows a clear cyclical pattern in milk production, with noticeable peaks and troughs recurring throughout the years. There does not appear to be a strong upward or downward trend over the entire period, though periodic fluctuations are evident. Additionally, the seasonal decomposition reveals a repeating pattern, suggesting the influence of seasonal factors on milk production. This is particularly clear in the seasonal component plot, where similar patterns recur annually. (See Figure 4)

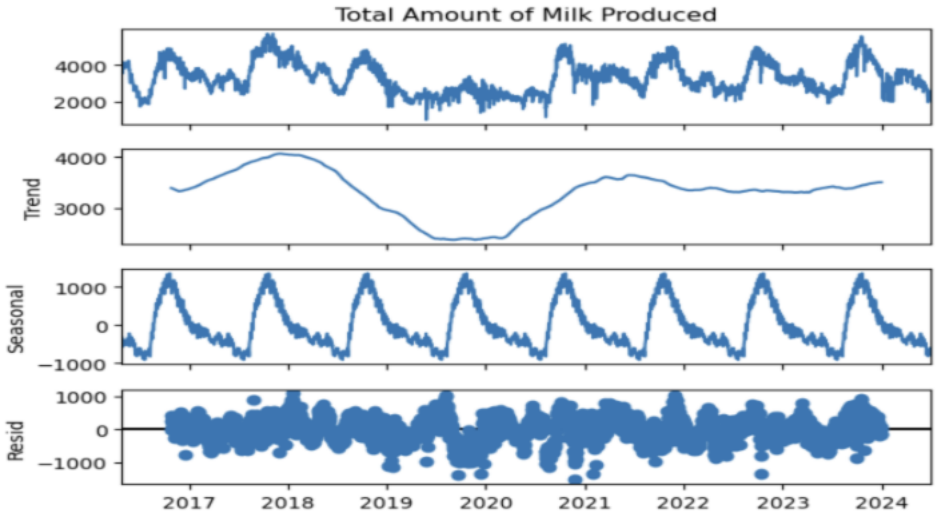


Figure 4: Decomposition of Total Milk Production Time Series

It is important to discuss the concept of stationarity and its relevance to time series analysis. Stationarity refers to a time series whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time. Many models, including ARIMA, rely on the assumption that the series is stationary, as non-stationary data can lead to misleading or unreliable results. To test for stationarity in the time series, we used the **Augmented Dickey-Fuller (ADF)** test, which extends the Dickey-Fuller test by including lagged differences of the series to account for autocorrelation. The ADF test equation is given by:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \ldots + \delta_p \Delta Y_{t-p} + \epsilon_t$$

The test produced a **p-value** of approximately **0.000037**, which is well below the significance level of 0.05. This means that the null hypothesis of non-stationarity is rejected, and the series can be considered **stationary**. In conclusion, the data exhibits both cyclical and seasonal patterns, and since the series is stationary, it is suitable for further time series modeling, such as ARIMA.

The next step was to identify the appropriate **AR** (Auto-Regressive) and **MA** (Moving Average) terms for modeling by examining the ACF and PACF plots. The ACF plot, which shows a slow decay over time, typically suggests the presence of an MA component, with significant lags identified as those falling outside the confidence interval. In this case, the ACF plot indicates an MA order (q) of 1 or 2. Conversely, the PACF plot shows a clear cutoff after lag 1, suggesting the presence of an AR component, with the first spike being significant and the subsequent ones dropping off quickly, indicating an AR order (p) of 1.

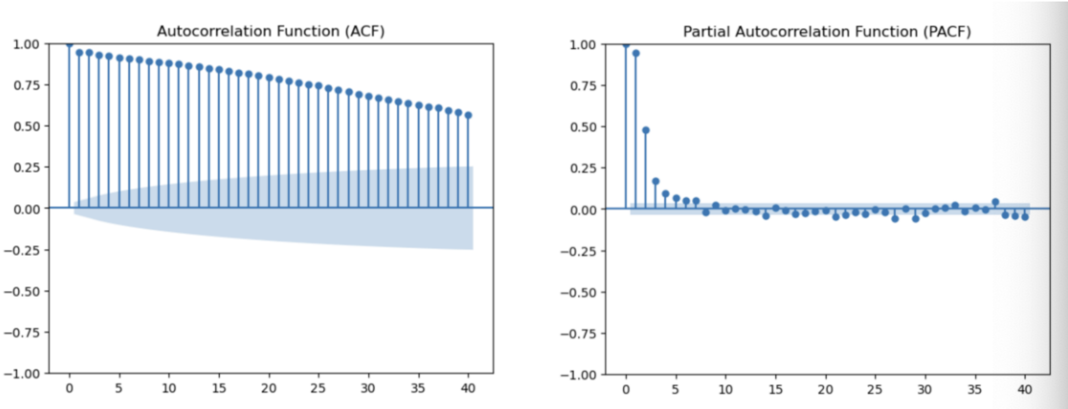


Figure 5: ACF & PACF plot

Based on these observations, we fitted ARIMA models with the orders ARIMA(1,1,1) and ARIMA(1,1,2), and compared their AIC and BIC values. Since we are looking for a model that balances goodness of fit with model complexity, ARIMA(1,1,2) may offer a slightly better fit based on the AIC values.

Model	**AIC Value**	**BIC Value**
ARIMA(1,1,1)	41251.660015583075	41269.66307546705
ARIMA(1,1,2)	41247.19801509631	41271.20209494161

Table 1: Comparison of ARIMA Model AIC and BIC Values

However, when we applied the ARIMA(1,1,2) model and checked the residuals using the ACF and PACF plots, it did not perform well, as the residuals still showed signs of autocorrelation. This indicated that the model failed to capture important patterns, likely due to seasonality. As a result, we switched to a SARIMA model, which is better suited for handling seasonal data. We split our sample into training and validation sets. Based on the comparison with the validation set, our forecasting performance is reasonably good, indicating that we can use the SARIMA(1,1,2)(1,1,1,12) model to forecast the next 10 years of milk production.

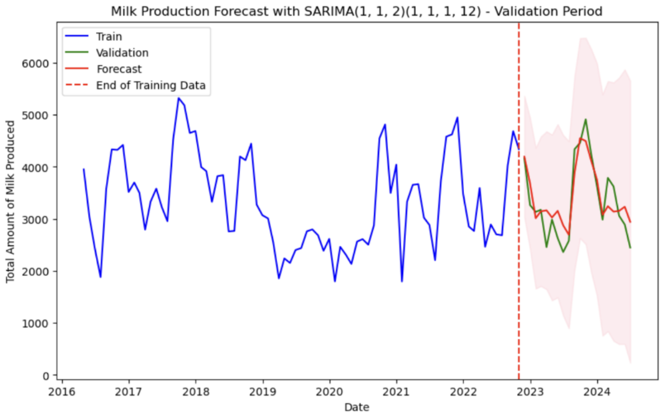


Figure 6: Milk Production Forecast Using SARIMA Model with Validation Period

5.2 High-Order Polynomial Regression

Our main objective for the entire project last semester was to make a prediction, specifically a forecast. However, due to limitations in the dataset, we were unable to achieve this goal.

This semester, we have obtained enough data to pursue our forecasting efforts. After careful consideration, we identified that adjusting the total feed offers a practical and impactful way to either increase production or reduce costs. As a result, we decided to focus on exploring variables with a strong correlation to Total Feed to inform our analysis.

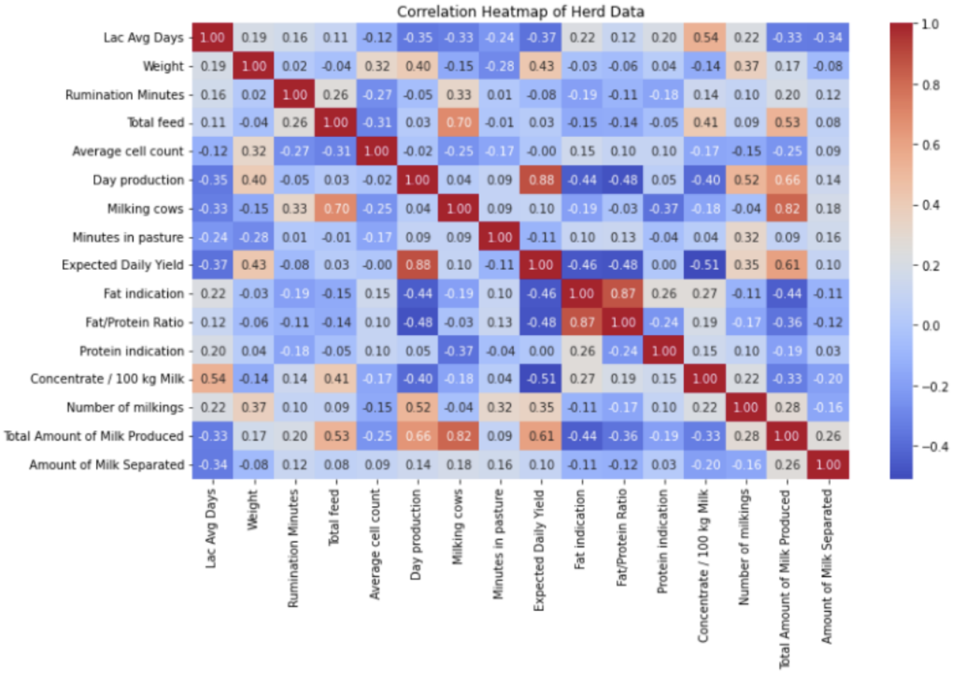


Figure 7: Correlation Heatmap for Herd Daily Dataset

5.2.1 Analysis of Correlation with Total Feed

Using a heatmap, we considered variables with an absolute correlation coefficient greater than 0.50, indicating a high relationship with Total Feed. Two key variables were identified:

1. Number of Milking Cows

The correlation coefficient with Total Feed is 0.697, indicating a significant positive correlation. This suggests a strong relationship between feed consumption and the number of cows milked. As the number of cows milked increases, feed consumption naturally rises. This correlation is intuitive since more milking cows require greater feed demand.

Further analysis: This indicates that feed management is largely dependent on the number of milking cows. Further investigation could examine whether there are significant differences in feed efficiency between herds of different sizes.

2. Total Amount of Milk Produced

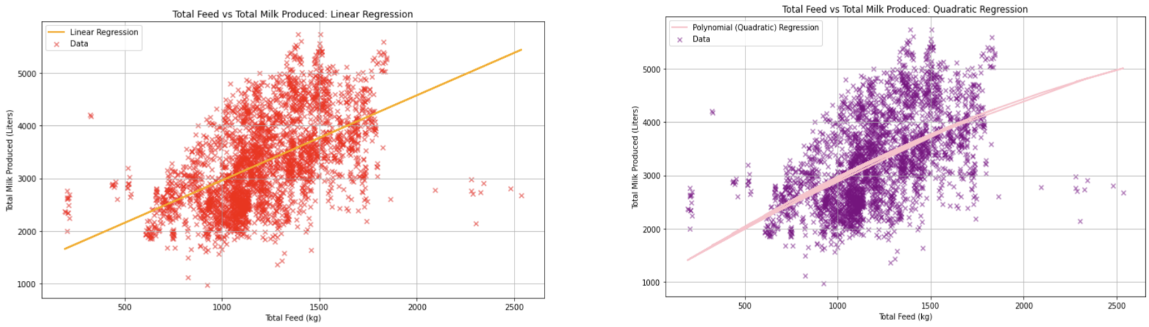
The correlation coefficient with Total Feed is 0.531, showing a moderately positive correlation. This suggests that as feed consumption increases, total milk production also rises to some extent. More feed can support higher milk production, as expected.

Further analysis: It is important to assess whether milk production increases proportionally with feed consumption or if there is a threshold beyond which the rate of milk production growth declines. Such an analysis can help optimize feed management strategies.

Considering the dataset limitations, we decided to focus on the second variable: the relationship between Total Feed and Total Amount of Milk Produced. This approach offers the most manageable and practical way to increase production or reduce costs. By identifying a feed consumption threshold beyond which milk production growth slows, feed management strategies can be optimized. In subsequent steps, we applied multiple polynomial regression models to predict this relationship. The model was evaluated using R^2, residual analysis, and a Q-Q plot to assess its fit and display the results.

5.2.2 Total Feed vs. Total Milk Produced

This single linear regression shows the relationship between total milk production and feed consumption. While total milk production increases as feed consumption increases, the single linear regression model does not capture all the changes in the data well. As shown in the graph, the model struggles to represent the complexity in the data.
Compared to linear regression, the quadratic regression curve provides a better fit for the data, showcasing a non-linear growth pattern. As feed consumption increases, milk production does not always increase proportionally. After a certain point, the rate of growth begins to slow down, suggesting the presence of a feed consumption threshold beyond which milk production growth diminishes.



(a) Total Feed vs Total Milk Produced: Linear Regression

(b) Total Feed vs Total Milk Produced: Quadratic Regression

Figure 8: Comparison between Linear and Quadratic Regression

5.3 Machine Learning Models

In the previous two approaches we used the herd daily dataset and didn't try our approaches on the device daily dataset. In the coming approaches of modeling we try various models on both datasets, herd daily and device daily.

5.3.1 Linear Regression

A simple statistical technique called linear regression (LR) makes the assumption that there is a direct, linear relationship between the input features—the factors influencing the target variable—and the target variable, or what you wish to predict. In order to minimize the discrepancies between the actual and anticipated values, the model fits a straight line. This approach works well when there is a linear relationship between the variables, but it has trouble processing more complicated datasets with intricate feature interactions or non-linear patterns. In our experiment, we began by predicting outcomes such as Milk Exp using LR. The non-linear nature of our data hampered its performance, even though it produced results that were understandable and obvious.[9]

5.3.2 Decision Tree Regressor

A Decision Tree Regressor (DTR) is a model that uses a tree-like structure of decision rules to make predictions based on input features. It works by splitting the data at different points based on feature values, with the goal of reducing prediction errors at each step. The final predictions are made at the tree's leaf nodes.

DTRs have several advantages: they are easy to understand and interpret, and they can handle both numerical and categorical data. However, DTRs also have some drawbacks. They can become too complex and overfit the training data, capturing unnecessary details, which may result in poor performance on new, unseen data.

We used a DTR to capture the non-linear relationships between features such as *Milk Duration*, *Milk Speed Avg.*, and *Milk Exp.*. While it performed better than linear regression, we noticed that it tended to overfit the data. This led us to explore more advanced ensemble models to improve accuracy and generalization [10]

5.3.3 Random Forest Regressor

Random Forest (RF) is a powerful machine learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. By randomly selecting data samples and features for each tree, RF becomes more resilient to noise and variance. This approach lowers the risk of overfitting by averaging the results from multiple trees. However, due to its complexity, the model becomes harder to interpret. We used RF to predict *Milk Exp.* and *Day Production*, and it performed exceptionally well. Its ability to capture complex relationships between features resulted in high accuracy and strong generalization to new data [11].

5.3.4 Gradient Boosting Regressor

The Gradient Boosting Regressor (GBR) builds decision trees one at a time, with each tree correcting the errors of the previous ones. This focused approach allows the model to continuously improve by reducing the remaining errors. GBR is particularly effective for handling complex, non-linear data. However, it can be more prone to overfitting compared to Random Forests, so parameters like the learning rate and tree depth need careful adjustment [12]. In our project, after fine-tuning, the GBR model proved to be the best for predicting *Milk Exp.*. It achieved the highest accuracy and lowest error by effectively capturing complex patterns in the data..

5.3.5 XGBoost

XGBoost (Extreme Gradient Boosting) is an enhanced version of gradient boosting that uses advanced techniques like regularization to prevent overfitting and improve performance. It is highly valued for its speed, efficiency, and ability to handle large datasets. Like traditional Gradient Boosting, XGBoost builds trees one at a time. However, it includes additional features such as regularization and the ability to manage missing data, leading to better overall results. We used both XGBoost and Gradient Boosting to forecast *Milk Exp.* and *Day Production.* The results showed that XGBoost outperformed Gradient Boosting in terms of accuracy while also being more computationally efficient, making it an ideal choice for large-scale prediction tasks [13].

6 Optimization

6.1 Statistical Methods for Optimizing Feed Consumption

The derivative of a quadratic function helps us understand the growth rate of milk production. By finding the point where the derivative equals zero, we can identify the threshold of feed consumption where the growth rate stops increasing and reaches its peak. This threshold marks the point where milk production begins to slow down, meaning that further increases in feed consumption will not lead to proportional gains in milk output. Essentially, it tells us when adding more feed no longer results in significant increases in milk production.
We used the quadratic regression model to determine the feed consumption level that corresponds to maximum milk production, which occurs at the peak of the quadratic curve. This is calculated using the quadratic equation $y = ax^2 + bx + c$, where the derivative $\frac{dy}{dx} = 2ax + b$ is set to zero, giving the optimal threshold as $x = \frac{-b}{2a}$. In our analysis, we applied both quadratic and cubic regression models to predict these thresholds, giving us valuable insights into how feed consumption impacts production. The results were as follows:

Parameter	**Value**
Coefficient a (Quadratic term)	-0.0002591596837366872
Coefficient b (Linear term)	2.2458106269687503
Intercept	982.9924745380172
Optimal Feed Threshold	4332.87 kg

Table 2: Results from the Quadratic Regression Model

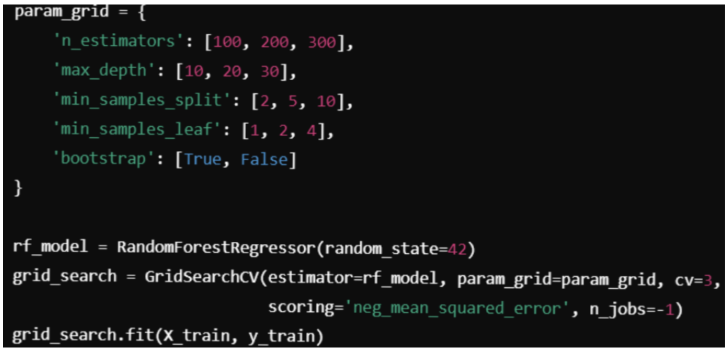
The quadratic regression model reveals that the secondary term (a) is negative, indicating that milk production decreases after feed intake reaches a certain level, while the linear term (b) is positive, suggesting that initial increases in feed consumption raise milk production. The baseline milk production, given by the intercept, is 982 liters, and the optimal feed intake for maximum milk production is 4332.87 kg. For cubic regression, the cubic term is negative, capturing a complex relationship with multiple inflection points, while the quadratic term is positive, implying an initial acceleration in milk production. The linear term is negative, suggesting slower growth beyond a threshold. The baseline milk production, indicated by the intercept, is 3772.81 liters, and the model predicts two thresholds for feed consumption at 1793.31 kg and 557.66 kg, beyond which milk production either peaks or declines. The table below summarizes the key coefficients and predictions from both models.

Parameter	**Quadratic Model**	**Cubic Model**
Coefficient a (Quadratic term)	-0.000259	0.006739
Coefficient b (Linear term)	2.245811	-5.732962
Intercept	982.99	3772.82
Optimal Feed Threshold	4332.87 kg	1793.31 kg, 557.66 kg
Predicted Milk Production at 1000 kg	N/A	2867.60 liters
Predicted Milk Production at 1500 kg	N/A	3886.06 liters
Predicted Milk Production at 2000 kg	N/A	3974.37 liters
Predicted Milk Production at 2500 kg	N/A	1699.36 liters

Table 3: Summary of Key Coefficients and Predictions

6.2 Machine Learning Models Optimization

In this project, we applied **GridSearchCV** to optimize the hyperparameters of the Random Forest and Gradient Boosting models. Our objective was to minimize the Mean Squared Error (MSE) on the training set by identifying the best combination of hyperparameters to enhance model performance. The process involved testing multiple hyperparameter values and leveraging cross-validation to ensure that the models did not overfit and maintained robustness. Figure 9 illustrates the GridSearchCV process for optimizing the models.



```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}

rf_model = RandomForestRegressor(random_state=42)
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=3,
                           scoring='neg_mean_squared_error', n_jobs=-1)
grid_search.fit(X_train, y_train)
```

Figure 9: GridSearchCV implementation for optimizing

To achieve this, we experimented with several key hyperparameters. These included:

- **n_estimators:** Controls the number of decision trees in the model. While increasing the number of trees may improve accuracy, it also raises computational costs. We tested values of 100, 200, and 300.
- **max_depth:** Limits the maximum depth of trees, controlling the complexity of patterns the model can capture. We evaluated depths of 10, 20, and 30.
- **min_samples_split:** Specifies the minimum number of samples required to split an internal node, preventing overly complex models. We explored values of 2, 5, and 10.
- **min_samples_leaf:** Defines the minimum number of samples required for a leaf node, helping to avoid overfitting. We tested values of 1, 2, and 4.
- **bootstrap:** Controls whether bootstrap sampling is used to build each tree. We tried both True and False.

6.2.1 Grid Search and Cross-Validation

To thoroughly evaluate the performance of each hyperparameter combination, we utilized **GridSearchCV** for grid search. Specifically, GridSearchCV explores all possible hyperparameter combinations and performs 3-fold cross-validation (cv=3) for each combination. Cross-validation splits the training data into three subsets, repeatedly training and validating on different subsets, ensuring the model does not overfit to a particular subset. In total, there were $3 \times 3 \times 3 \times 3 \times 2 = 162$ combinations of hyperparameters, meaning GridSearchCV tested 162 different hyperparameter sets.

6.2.2 Hyperparameter Optimization

We used Mean Squared Error (MSE) as the evaluation metric to measure the average squared difference between the predicted values and the actual values, with a lower MSE indicating more accurate predictions. GridSearchCV was employed to optimize the hyperparameters by testing 162 combinations, using negative MSE during cross-validation to ensure consistency. The best combination of hyperparameters was automatically selected based on minimizing the negative MSE, leading to improved model performance. The process of optimization obviously decrease the error of prediction, lead to MSE be lower, and increase fitting ability of model, make it more stable especially in new dataset. After optimized, hyperparameters get balanced between accuracy and efficiency, ensuring model avoid unnecessary complexity and keep best performance in prediction.

7 Evaluation

7.1 Evaluation of Model Fitting

7.1.1 Residual Analysis

We decided to set Residual Analysis to evaluating the fitting effect for cubic and quadratic regression models. The differences between the actual values and predicted values of model called Residuals, as a well-fitting model should have irregular residual even in distributed despite inputs are different. In later, We use the normality of the residual, residual histogram and QQ-plot to evaluate the performance for two regression models. Ideally, the points in a residual plot should be randomly distributed above and below the zero line. However, in both models, the residuals exhibit different patterns:

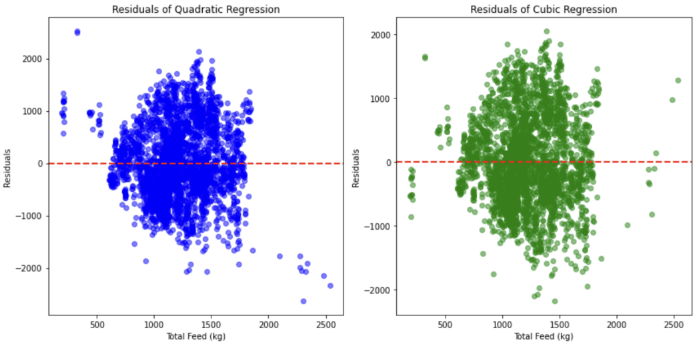


Figure 10: Residual plots for quadratic regression and cubic regression

In the quadratic regression residual plot, there are significant fluctuations in the residuals at lower feed consumption (less than 500 kg) and higher feed consumption (more than 2000 kg). The residuals are more concentrated in the middle range, but at feed consumption levels above 2500 kg, the residuals drop sharply, suggesting that the quadratic regression model struggles to capture the non-linear trend in the data, particularly at higher feed consumption levels. The residuals of the cubic regression model are more evenly distributed across different levels of feed consumption, particularly in the intermediate range (1000 kg to 2000 kg). There is less fluctuation in the residuals at extreme values (above 2000 kg), and overall, the residuals are more randomly distributed, indicating that the cubic regression model may provide a better fit compared to the quadratic model. Although a few residual points deviate, there is no clear pattern, which demonstrates the model's stability at extreme values.

7.1.2 Residual Histogram

The residual histograms for both models reveal insights into the normality of the residual distribution:

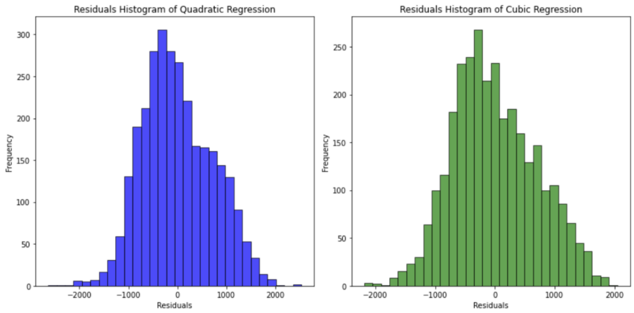


Figure 11: Residual histograms for quadratic regression and cubic regression

The quadratic regression residuals approximate a bell-shaped distribution, indicating that the residuals are close to a normal distribution. The residuals are symmetrically centered around zero, suggesting that the model performs well in most cases. However, there are some extreme residuals (around -2000 and 2000), indicating that the model does not perform as well at very low or very high feed consumption values. The residuals of the cubic regression model display a more symmetrical and smooth distribution. The residuals are concentrated near zero, implying that the model has a slight advantage in fitting the data overall. There are fewer extreme residuals compared to the quadratic regression, and the overall distribution is more centered, which may indicate better generalization.

7.1.3 QQ Plot Analysis

In theory, the residuals of a normally distributed model should lie close to a diagonal line in a QQ plot.

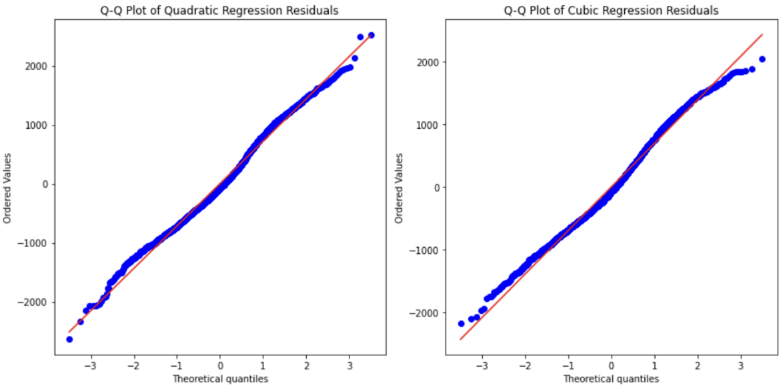


Figure 12: QQ plot for quadratic regression and cubic regression

As figure above, most points of the quadratic regression model are close to the diagonal line, which means the residual is approximately normal distribution, however, there still are some extremes, especially focus on top right and bottom left, that are significantly off the diagonal, indicating there are large residuals at these extremes. This suggests the quadratic model may have some systematic errors or defects at very low or very high levels of feed consumption. For the cubic regression, QQ plot shows most of the points are near the diagonal line, so the normality of the residual is very good. The performance of the cubic regression is better than quadratic model, especially in extreme values, for datasets have more extreme values or stronger nonlinear trends. So cubic regression will be more effective at maintaining the normality of the residuals when dealing with extreme data points.

Conclusion: The residuals of the two regression models show good normality, and the performance of the cubic regression model is slightly better than that of the quadratic regression model, especially at the extreme value. Residual analysis shows that the cubic model captures the nonlinear trend of data better, and its residual distribution is more symmetrical, which makes it more suitable for data modeling with complex nonlinear relationships.

7.2 Evaluation Criteria for Machine Learning Models

We applied the same evaluation process to two datasets, Device Daily and Herd Daily, in order to select and validate the best-performing model. The evaluation process includes model selection, test dataset prediction and through performance metrics for model validity.

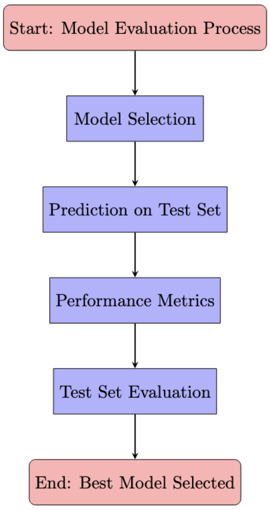


Figure 13: Flowchart illustrating the model evaluation process

Gradient Boosting model was selected, we will through an unknown dataset to test its performance to evaluate its ability of fitting. According this way to verify that the model did overfitting or not, and generalize effectively to new dataset.
Two key performance metrics were used to evaluate the model:
Mean Squared Error (MSE): This value represent the mean squared difference between actual value with prediction. Low value for MSE indicates that predict value are more accurate for this model.
Coefficient of Determination ($\mathbf{R^2}$): This value evaluate the ability of explanation for the model. As R^2 value closer to 1 represent the model has strong explanatory ability.
From the final evaluation for test dataset showed the **Gradient Boosting** has low MSE and high R^2 score, which means this model is pretty good and effective. These results showed strong predictive ability and robustness, verified **Gradient Boosting** model will effectively adapt to new data, provided reliable predictions. During the model evaluation process, it experienced consistently applied different datasets, ensured the accurate, robust, and generalizing ability are well in optimized Gradient Boosting model, so it is suitable for more real-world applications.

8 Results and Discussion

8.1 Time Series Modeling

We used a SARIMA model to forecast total milk production from 2025 to 2035, with a non-seasonal order of $(1, 1, 1)$ and a seasonal order of $(1, 1, 1, 12)$. This setup captures both seasonal trends and fluctuations in milk production, effectively modeling the recurring patterns that occur every 12 months, as seen in the repeating peaks and troughs in the forecast. The red line in the figure shows the predicted values, while the shaded area represents the confidence intervals. The shaded area gets wider over time, indicating more uncertainty in the long-term forecast. Although there are periodic ups and downs, the model successfully captures the seasonality of milk production, suggesting a consistent overall pattern. However, it also shows a slight decline in total milk production toward the end of the forecast period, hinting at a possible downward trend. This suggests that while SARIMA works well for capturing seasonal changes and trends, further investigation is needed to understand what might be causing the decline. Adding more variables or adjusting the model could help improve the accuracy of the forecast and provide deeper insights into milk production trends.. (See Figure 14)

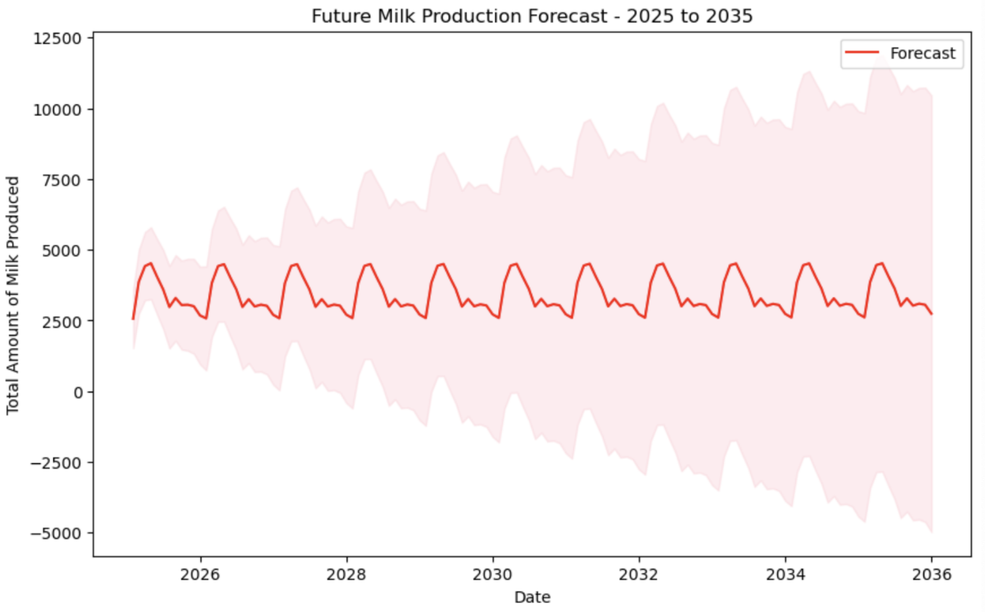


Figure 14: Future Milk Production Forecast (2025 to 2035)

8.2 High-Order Polynomial Regression

In fact, we try several model in polynomial regression, from single to quartic, and evaluate the efficient of fitting result, analyze the relationship between feed consumption with milk production. Based on R^2 scores and MSE to compared different polynomial regression. The table (4) below display these results:

Model	Train R^2 / MSE	Test R^2 / MSE
Quadratic Regression	0.28 / N/A	0.32 / N/A
Cubic Regression	0.32 / 490,940.38	0.35 / 478,832.40
Quartic Regression	0.32 / 490,787.35	0.35 / 478,953.32

Table 4: Performance comparison of polynomial regression models on training and test sets.

In terms of results, the quadratic regression model is almost indistinguishable from the cubic and quartic regressions in both R^2 and MSE. This shows that increasing the polynomial's complexity (e.g., moving from cubic to quartic regression) does not significantly improve the model's fit. This suggests that the data may exhibit a less complex non-linear structure, and polynomial models of degree three or higher may not capture additional meaningful information. In fact, higher-order models may lead to overfitting rather than performance improvements.

Thus, a simpler model like quadratic regression may be more suitable, particularly if higher-order models do not demonstrate clear benefits. Simpler models tend to be more explanatory and can avoid the risk of over fitting.

8.2.1 Comparison of Quadratic and Cubic Regression Models

To further understand the complexity, we compared quadratic and cubic regression models in greater detail. The second-order (quadratic) model captures a single critical point, while the third-order (cubic) model allows for multiple critical points, reflecting more intricate shifts in trends. The quadratic regression form yielded an optimal feed threshold of 4332.87 kg, while the cubic model identified two critical points at 557.66 kg and 1793.31 kg. These points represent which stages that milk production is increasing or decreasing .

As a result, cubic regression obviously provided more detail in capturing complex non-linear relationships aspect, but the quadratic model easier to interpret. All in all, cubic regression provides richer insights from dataset, but quadratic regression looks like simplicity and interpretability, which one is better depends on preference of individual.

8.3 Machine Learning Models

Best Model for Herd Daily Dataset: Random Forest Regressor

The table (5) below compares the performance of several regression models applied to the Herd Daily History dataset, evaluated using Mean Squared Error (MSE) and R^2 score. Random Forest and XGBoost stood out as the best-performing models, achieving low MSE values (0.89 and 0.90, respectively) and high R^2 scores (both 0.95), indicating strong predictive capabilities. In contrast, Linear Regression and Decision Tree Regression demonstrated lower accuracy, while Gradient Boosting performed moderately, with an MSE of 1.36 and an R^2 score of 0.93.

Model	**MSE**	**R^2 Score**
Linear Regression	3.141	0.83
Decision Tree Regression	2.42	0.87
Random Forest	0.89	0.95
Gradient Boosting	1.36	0.93
XGBoost	0.90	0.95

Table 5: Comparison of Regression Models on Herd Daily History Data

Random Forest Regressor (RF) emerged as the best model for the Herd Daily dataset, capturing the complex relationships between features such as Day Production, Total Feed, and Rumination Minutes. A key strength of Random Forest (RF) is its ability to handle complex interactions between herd-level variables without overfitting. It does this by combining predictions from multiple decision trees, which helps model non-linear relationships and reduce noise. The model achieved a Mean Squared Error (MSE) of 0.89 and an R^2 score of 0.95, meaning it explained 95% of the variance in Day Production. This made it the most reliable model for our dataset.

Best Model for Device Daily Dataset: Gradient Boosting Regressor (GBR)

The table (6) shows how different regression models performed on the Device Daily History dataset. Their performance was measured using Mean Squared Error (MSE) and R^2 score. Gradient Boosting and Random Forest performed the best, with low MSE values of 5207.27

and 5403.17, and high R^2 scores of 0.91 and 0.90. XGBoost also did well, with an MSE of 5638.06 and an R^2 score of 0.89. In comparison, Decision Tree Regression showed moderate performance, while Linear Regression had the lowest accuracy, with a high MSE of 41690.78 and an R^2 score of only 0.24.

Model	**MSE**	**R^2 Score**
Linear Regression	41690.78	0.24
Decision Tree	10156.46	0.82
Random Forest	5403.17	0.90
Gradient Boosting	5207.27	0.91
XGBoost	5638.06	0.89

Table 6: Comparison of Regression Models on Device Daily History Data

The Gradient Boosting Regressor (GBR) proved to be the most effective model for the Device Daily dataset. It did a great job capturing non-linear relationships between key factors like Milk Duration, Milk Speed Avg., and Milk Exp. GBR's process, where each new tree corrects the errors of the previous one, made it especially good at modeling these complex interactions. Its flexibility allowed it to find subtle patterns in the data that simpler models, like Linear Regression, couldn't detect.

After tuning important hyperparameters such as the learning rate, number of estimators, and maximum tree depth, GBR achieved the lowest MSE and the highest R^2 score for predicting Milk Exp. Specifically, the model had an MSE of 4863 on the validation set and 4807 on the test set, while explaining over 91% of the variance in Milk Exp. predictions (R^2 values of 0.916 for the test set and 0.91 for the validation set). This high accuracy shows that GBR is excellent at modeling complex relationships in the data, outperforming other models like Random Forest. Overall, GBR was the most reliable model for this task.

9 Conclusions and Recommendations

In this project, we explored several modeling techniques, including time series analysis, polynomial regression, and machine learning models. Time series models like SARIMA worked well for forecasting future trends, capturing seasonal patterns and fluctuations that matched the validation data. Polynomial regression helped explain some patterns using quadratic, cubic, and quartic terms, but adding more terms didn't improve the results much, showing its limits in handling more complex trends. On the other hand, machine learning models like Random Forest and Gradient Boosting performed better when analyzing the Herd Daily dataset, giving strong predictions and capturing the complex relationships between variables. However, the Device Daily dataset had higher error rates, suggesting that more work is needed on feature selection and engineering to improve results.

Looking ahead, we have several recommendations for future work. Adding more data sources could improve the model's accuracy and give a fuller picture of milk production trends. Using real-time monitoring systems would also allow for quicker, better decisions based on up-to-date data, making it easier to make timely adjustments. Exploring more advanced models and extending time series forecasting could help improve predictions for more complex and long-term trends. Lastly, focusing more on feature engineering, especially for datasets like Device Daily, could help reduce error rates and lead to more reliable results in future analyses.

10 Appendix

The source code and the records of all meetings can be accessed at the following link: `https://github.com/SX0818/DS_project`.

References

[1] M. C. Ferris, A. Christensen, and S. R. Wangen, "Symposium review: Dairy brain—informing decisions on dairy farms using data analytics," *Journal of Dairy Science*, vol. 103, no. 4, pp. 3874–3881, 2020.

[2] V. Cabrera, J. Barrientos, L. Fadul, and H. Delgado, "Real-time continuous decision-making using big data," in *Journal of Dairy Science*, vol. 102, pp. 273–273, Elsevier Science Inc Ste 800, 230 Park Ave, New York, NY 10169 USA, 2019.

[3] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – a review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017.

[4] A. Liseune *et al.*, "Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning," *Computers and Electronics in Agriculture*, 2021.

[5] W. Grzesiak *et al.*, "Methods of predicting milk yield in dairy cows-predictive capabilities of wood's lactation curve and artificial neural networks (anns)," *Computers and Electronics in Agriculture*, 2006.

[6] S. Araújo, R. Peres, L. Filipe, A. Manta-Costa, F. Lidon, J. Ramalho, and J. Barata, "Intelligent data-driven decision support for agricultural systems-id3sas," *IEEE Access*, 2023.

[7] I. Andonovic, C. Michie, M. Gilroy, H. Goh, K. Kwong, K. Sasloglou, and T. Wu, "Wireless sensor networks for cattle health monitoring," in *ICT innovations 2009*, pp. 21–31, Springer, 2010.

[8] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

[9] M. Alwadi, A. Alwadi, G. Chetty, and J. Alnaimi, "Smart dairy farming for predicting milk production yield based on deep machine learning," *International Journal of Information Technology*, vol. 16, no. 7, pp. 4181–4190, 2024.

[10] J. Aerts, M. Kolenda, D. Piwczyński, B. Sitkowska, and H. Önder, "Forecasting milking efficiency of dairy cows milked in an automatic milking system using the decision tree technique," *Animals*, vol. 12, no. 8, p. 1040, 2022.

[11] K. S. Themistokleous, N. Sakellariou, and E. Kiossis, "A deep learning algorithm predicts milk yield and production stage of dairy cows utilizing ultrasound echotexture analysis of the mammary gland," *Computers and Electronics in Agriculture*, vol. 198, p. 106992, 2022.

[12] G.-J. Streefland, F. Herrema, and M. Martini, "A gradient boosting model to predict the milk production," *Smart Agricultural Technology*, vol. 6, p. 100302, 2023.

[13] B. Ji, T. Banhazi, C. J. Phillips, C. Wang, and B. Li, "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*, vol. 216, pp. 186–197, 2022.