THE UNIVERSITY OF MELBOURNE

Faculty of Sciences

Big Data in Dairy

May 24, 2024

Project Report

Harsh Mangla (1418017)

Yuqian Xie (1448428)

Xuheng Li (1405559)

Kaifan Ouyang (14279571)

Host:

Kristy Digiacomo

Supervised by:

Mario Andres Munoz Acosta

Contents

1 Introduction

1.1 Background

Dairy milk production systems, which generate a large amount of data on a daily basis, are not well utilized in the field of scientific research because they are not fully integrated with other resources. In recent years, the field of agriculture and animal husbandry is changing dramatically with the rapid development of science and technology, especially the application of big data and automation.

1.2 Problem Difficulty

The core problem of the project is the inefficiency and inconsistency inherent in traditional milking methods. These traditional methods are labor intensive and prone to human error, and often result in less than optimal levels of cow comfort and milk production." The goal of the Big Data in Dairy project is to explore how these issues can be mitigated by combining automated milking systems with big data analytics to improve productivity, consistency and overall farm management.

1.3 Importance of Automation in Dairy Farming

Automation is vital for dairy farms, significantly reducing the amount of manual labor required, which is particularly important in the current context of labor shortages and the physical demands of milking. Automated systems ensure consistency in the milking process, improving milk quality, helping to maintain high standards and meeting regulatory requirements. In addition, automation improves cow welfare by providing a stress-free and precise milking routine, resulting in healthier cows and higher milk yields. In addition, the integration of big data enables real-time monitoring and analytics, enabling farmers to make informed decisions and optimize all aspects of dairy farm operations.

1.4 Data Science Challenges in Dairy Farming

1. Volume of Data: Because of real-time monitoring, the system generates a large amount of data, which requires powerful equipment and complex algorithms to process the data. Although the scale of the data we are currently exposed to is small, it may be necessary to process big data in the future, which will be a big challenge in this area.
2. Data Sparsity: Data tends to be sparse, with missing values, and the distribution of data points is messy and irregular, which negatively impacts the analysis.
3. Degree of Noise: Sensor data is susceptible to noise and outliers, which can reduce reliability. This data needs to be effectively filtered to clean up these outliers.
4. Data Structure: Data is highly heterogeneous and comes from a wide range of sources, like milking machines, animal health monitors, and environmental sensors. Disparate data, integrating trait values from different sources into a coherent analytical framework is a challenge.

5. Real-Time Processing: In order to provide effective automation, data must be processed in real-time to provide instant adjustments. Developing algorithms that can process real-time data streams and provide actionable insights is a key aspect of the project.
6. Predictive Analytics: Building accurate predictive models to predict variables such as milk production, animal health, and operational efficiency requires advanced machine learning techniques and an in-depth understanding of biological, and environmental factors in the farming.

2 Related Work

Exploring the application of automation and big data in dairy farming, there has been a lot of research and discovery. By looking at reputable academic publications and research reports on business systems, we will increase our knowledge of the mission and context. At the same time, it also provides resources and methodologies that can be used to develop or address solutions to identified challenges.

2.1 Decision Support Systems in Dairy Farming

The Dairy Brain, a decision support system created to improve dairy farm management by integrating data from several sources, is described by Ferris et al. (2020)[1]. Through the Agricultural Data Hub, the system uses data management and analytical tools to process data streams from farm activities and other sources. Cow Brain uses prescriptive, predictive, and descriptive analytics to provide useful information to support farm management decisions. Because its modular architecture enables flexible integration of various data types, it is a valuable tool to enhance dairy farming methods and improve overall production and efficiency levels. The Dairy Brain is further explained by Cabrera et al. (2020)[2], who highlighted the system's ability to integrate data and make decisions in real time. By integrating big data analytics, precision agriculture and the Internet of Things, the system provides continuous decision assistance, promoting relevant and useful farm management operations. Use cases include nutritional grouping, early risk detection of clinical mastitis and prediction of the onset of clinical mastitis, highlighting how the system can enhance decision-making and farm management.

2.2 Precision Dairy Farming Technologies

By using cutting-edge technologies to track physiological, behavioural, and production parameters in individual animals, precision dairy farming[3] improves farm management practices. Real-time data collecting for better decision-making is made possible by technologies such as automatic oestrous detection, daily milk yield recording, and milk component monitoring. Maximising animal potential, early disease detection, and medication reduction through preventive health interventions are the main objectives. This method enhances animal health and welfare while cutting expenses and increasing efficiency. In their discussion of the function of big data in smart farming, Wolfert et al. (2017)[4] place special emphasis on the management of farms through the integration of IoT and cloud computing technologies. The review demonstrates how the use of Big Data applications goes beyond primary production to impact the whole food supply chain by providing real-time decision-making, predictive insights, and the redesign of business processes. This is consistent with our project as it demonstrates how big data can improve operational efficiency and decision-making processes in dairy farming by integrating data more effectively and using predictive analytics.

2.3 Health Monitoring Systems

Andonovic et al.(2016)[5] explore the use of wireless sensor networks (WSNs) for continuous monitoring of cattle health, addressing the inconsistent and costly data produced by current monitoring techniques. The proposed approach uses inexpensive, low-power sensor nodes to communicate health data to farm managers in real time. This approach is adaptable to different farm conditions and animal species and can improve timely detection of critical health events. The study highlights the benefits of using WSNs for continuous, real-time health monitoring and shows how it can improve animal welfare, stop the spread of disease, and improve farm management techniques in dairy production.

2.4 Intelligent Data-Driven Decision Support

Araújo et al (2023)[6] proposed the Intelligent Data Driven Agricultural System Decision Support (ID3SAS) approach, which combines cloud computing, machine learning, IoT, and sensor technologies for real-time decision making in agriculture. The system collects and evaluates data from various sensors to maximize crop productivity, minimize water usage, and optimize irrigation schedules. This approach demonstrates the potential of data-driven insights to improve agricultural sustainability and efficiency, while also being in line with the Agriculture 4.0 trend. ID3SAS provides a framework that we can use to integrate cutting-edge technologies to optimize dairy farming operations, improve resource utilization, and enhance overall farm management.

2.5 Internet of Things (IoT) in Dairy Farming

This study[7] emphasises how fixing long-standing issues and enhancing milk production may be accomplished through the use of IoT and data-driven farming practices. Smart dairy farming (SDF) uses artificial intelligence (AI), cloud computing, and cutting-edge sensor technologies to track and monitor several variables in real time, including feed intake, milk production, behaviour, and health of the cows. Along with increasing output, these advances help lessen their negative effects on the environment and increase animal welfare. Continuous data from wearable sensors and Internet of Things (IoT) devices allows for the prompt identification of health problems and efficient resource management.

2.6 Machine Learning Applications in Dairy Farming

The use of machine learning (ML) in dairy farming is reviewed in this review[8], with particular attention to ML's capacity to analyse huge datasets produced by diverse sensors and herd management systems. ML techniques have been used to forecast energy usage, milk production, and other management elements of dairy farms. Even with their promise, ML apps nowadays frequently don't integrate data well enough, which lowers their dependability. The review emphasises that ML algorithms may perform better on bigger, more integrated datasets. In light of our initiative, this study emphasises how crucial it is to use machine learning (ML) to analyse integrated data sources in order to improve decision support systems, animal welfare, and dairy farming production.

2.7 Comparative Performance Analysis of Automatic Milking Systems

The impact of upgrading an Automatic Milking System (AMS) from an earlier model (Classic) to a newer generation model (VMS 300) on a buffalo farm in Southern Italy is examined in this study[9]. The results show that when compared to the Classic model, the VMS 300 greatly improves milk output and quality. Through the promotion of natural milking behaviour and the reduction of stress, AMS enhances animal welfare while decreasing the need for human labour, increasing production, standardising teat washing and disinfection. With the latest AMS, continuous monitoring of milk parameters and animal health condition enables prompt actions, further enhancing farm management. This study supports our effort by demonstrating the advantages of implementing cutting-edge AMS technology to improve milk yield and quality, boost productivity, and optimise dairy farming operations.

The comprehensive literature research and industry-standard systems demonstrate the substantial potential of automation and big data technology in dairy production. The necessity of real-time processing, advanced analytics, and data integration are important topics that will improve farm management and decision-making. The "Big Data in Dairy" initiative intends to create creative solutions that increase production, consistency, and sustainability in dairy farming by expanding on these insights and tackling the stated difficulties.

3 Data Analysis

3.1 Data Overview

The dataset used for this project is derived from the Lely Horizon app, which includes comprehensive data from dairy farms, with a particular focus on the milking process, cow health and production output. The data contain all kinds of parameters like milk production, lactation Day, number of lactation, fat content and other relevant indicators of cow productivity and health.

3.2 Data Cleaning and Processing

Before any analysis is performed, it is critical to clean and preprocess the data to ensure its quality and reliability. The following steps were taken:

3.2.1 Data Loading

The data is loaded from multiple CSV files downloaded from the Lely Horizon app. Our first step is to display the first few rows of the dataset. This gives a clear idea of the structure and content of the data. Here are some of the key columns in the dataset:

Column Name	Description
Cow Number	Unique identifier for each cow
Day Production	Daily milk production in litres
Lactation Days	Number of days into the current lactation cycle
Fat Indication	Percentage of fat in the milk
Protein Indication	Percentage of protein in the milk
SCC Indication	Somatic Cell Count, indicating milk quality and udder health

Table 1: Key columns in the dataset

3.2.2 Handling Missing Values

We examined missing values across the dataset We examined all the collected datasets for missing values (Figure 1), which were found to contain a large number of missing values, and calculated the percentage of missing values for each column and excluded columns with a percentage of missing values greater than 50 percentage. To deal with the data that did not have that much to deal with missing values, we used a forward padding method to ensure that the time series remained complete. We removed columns such as “days of pregnancy” and “fat/protein annotations” that were deemed unnecessary. For missing values in “days of lactation”, we chose to use the median for estimation to minimize the impact of missing values.

	Cow Number	Lac. no	Calving date	Group	Sire	Lactation days	No of Calves	Day production	Reproduction	Last heat	Ins. Date	Since Insemination	Sire.1	Pregnancy Check Date	Days Pregnant	Days To Dry Off	Dry off days	Expected Dry Off Date	Dry off date	Expected calving date
0	AVG	NaN	NaN	NaN	NaN	192.0	NaN	26.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	44.0	NaN	NaN	NaN
1	SUM	NaN	NaN	NaN	NaN	NaN	NaN	2677.7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	1045	9.0	2022-04-11	Dry Cows	REECE	749.0	1.0	NaN	Open Cyclic	2022-05-13	2023-05-19	346.0	AMEN-PP	2024-01-31	NaN	NaN	20.0	NaN	2024-04-09	NaN
3	1134	9.0	2024-02-24	Milking Cows	NaN	65.0	1.0	29.3	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	1177	8.0	2023-08-19	Dry Cows	PAVLICH	254.0	1.0	NaN	Open	NaN	2023-11-10	171.0	AMEN-PP	2024-01-31	NaN	NaN	69.0	NaN	2024-02-20	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
120	2202	1.0	2024-02-15	Milking Heifers	NaN	74.0	1.0	22.9	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
121	2203	1.0	2024-02-20	Milking Heifers	SALOM	69.0	1.0	19.2	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
122	2205	1.0	2024-02-15	Milking Heifers	CONTARDO	74.0	1.0	15.1	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
123	2209	1.0	2024-02-23	Milking Heifers	SALOM	66.0	1.0	18.0	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
124	2219	1.0	2024-02-19	Milking Heifers	RIVERBEND	70.0	1.0	13.3	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

125 rows × 20 columns

Figure 1: NaN values in the Dataset

3.2.3 Outlier Detection and Removal

We used the interquartile spacing (IQR) method to detect outliers in “days of lactation”. Data points exceeding the IQR by a factor of 1.5 are excluded to minimize bias and increase the reliability of the analysis.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Correlation Heatmap Display and Analysis

The correlation heatmap (Figure 2) was created to visualize to illustrate the relationships among various metrics for dairy cows.

High Positive Correlations

- **Lactation Number and Lactation Days**: There is a high positive correlation (0.85) indicating that an increase in the number of lactation cycles is associated with longer lactation periods.
- **Programmed Total and Total Intake**: A strong positive correlation (0.78) suggests that the planned feed amounts closely match the actual intake.

High Negative Correlations

- **Day Production and Lactation Days**: A moderate negative correlation (-0.61) implies that longer lactation periods may lead to a reduction in daily milk production.

They provide farm managers with specific data to develop and adjust feeding plans based on the actual needs and performance of dairy cows. This allows for more efficient management of resources and improved cow productivity and health, resulting in more sustainable and profitable dairy farming.

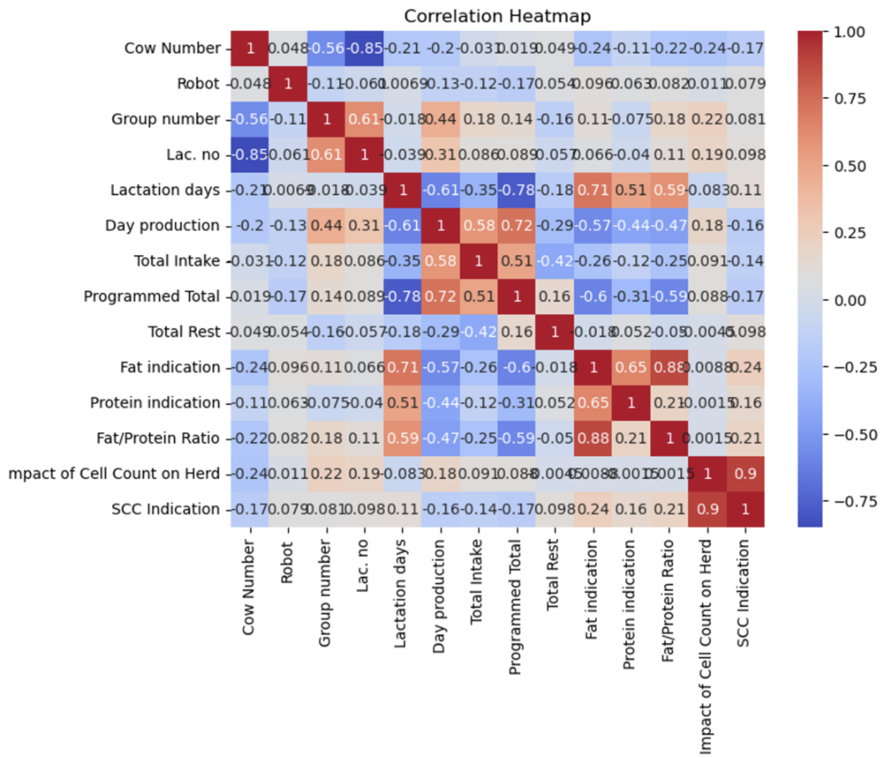

Figure 2: Correlation Heatmap

3.3.2 Daily Milk Production vs. Number of Lactation

This chart (Figure 3) illustrates the relationship between lactation days and daily milk production:

Key Observations

- **Early Peak**: At about 50 days, daily milk production peaks nearly to 45 liters,this showing the highest productivity in whole period.
- **Fluctuations**: Between 100 and 200 days, there are significant fluctuations for milk production, this is important to know which factor lead to this situation.
- **Decline**: After 220 days, milk production is generally in a state of decline, especially after 350 days when it falls below 20 litres.
- **Sparse Data**: There are very few data points between 350 and 400 days, which means that most cows do not continue to lactate for long periods of time.

With this data, we can get a clear picture of these production trends and make adjustments to specify effective lactation management, which will give a crucial image to Day Production.

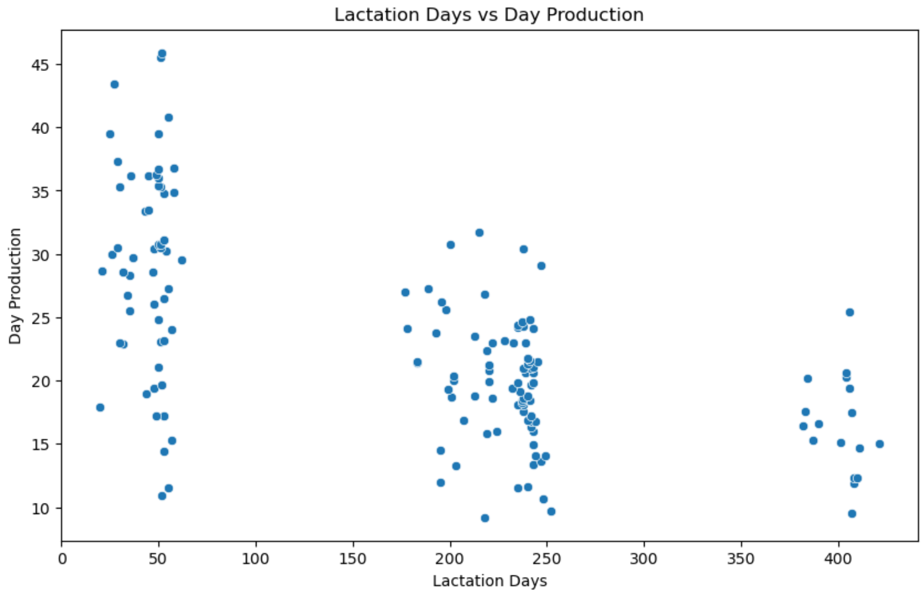

Figure 3: Daily Milk Production vs. Number of Lactation

3.3.3 Average Daily Milk Production vs. Number of Lactations in Dairy Cows

This line chart display the relationship between average daily milk production with the number of lactations:

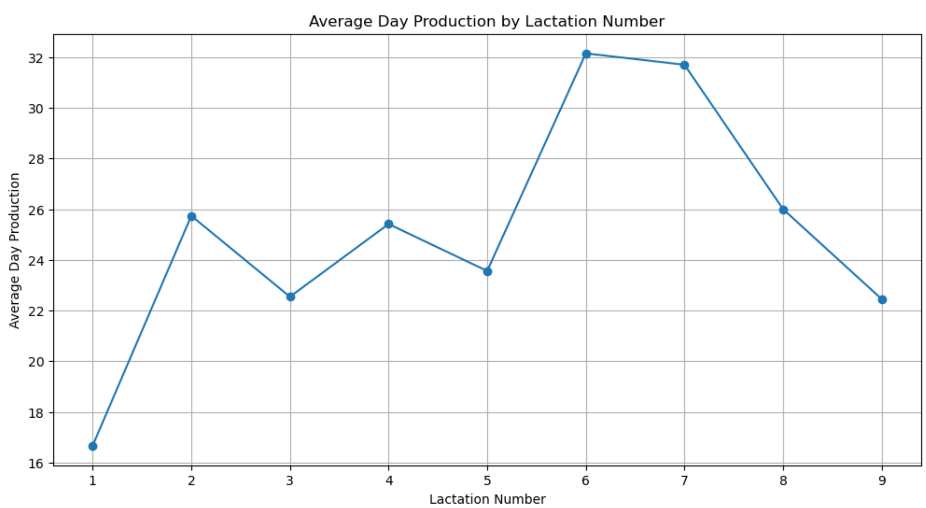

Figure 4: Average Daily Milk Production vs. Number of Lactations

Key Observations

- **Initial Increase**: From the first lactation to the second lactation, and from fifth to the seventh lactation, daily milk production increases dramatically.
- **Mid-Cycle Variability**: There are considerable fluctuations in production from the second to the

sixth, peaking at the sixth lactation. This fluctuation can be influenced by a variety of factors. It is recommended to use control variables to specifically check what the attributes are that cause this fluctuation.

- **Decline in Later Stages**: After the seventh lactation, there was a noticeable and sustained decline in production. This decline may be due to the natural weakening process of the cow, as the cow's productivity naturally and continuously declines once it has exceeded its peak location.

Implications for Management

Understanding the back reason of the wave of average day production will improve yields and increase the benefit .

- **Farm Management**: Adapting strategies to lactation performance can increase productivity.
- **Nutrition and Health Support**: Providing additional nutritional or health support can stabilize yields, especially during peak and up-and-down times.

Allocating resources more effectively based on these insights can help maximize productivity and ensure sustainable dairy farming practices.

4 Solution Approach and Plan for Next Semester

4.1 Data Utilization for Analysis and Prediction

Using a variety of cutting-edge analytical and predictive methods, the data collected from automated milking systems will be used to address the inefficiencies of traditional milking methods and improve overall farm management. Here are the main models and methods that will be used:

a. Predictive Analytics

- **Machine Learning Algorithms:** To predict milk production, identify early disease symptoms, and improve feeding schedules, we will use supervised learning algorithms including random forests, support vector machines (SVM), and gradient enhancers. These models will use data from the past to predict what will happen in the future.
- **Time Series Analysis:** Time-dependent trends for milk production and health indicators will be modeled using techniques such as LSTM(Long short-term Memory) networks and ARIMA(autoregressive Integrated Moving Average).

b. Descriptive and Prescriptive Analytics

- **Descriptive analytics:** To find links between various variables, including milk production, cow health, and environmental conditions, we will use tools such as clustering and correlated heat maps. This will help to understand the underlying patterns and trends in the data.
- **Prescriptive analytics:** Based on predictive models, optimization techniques such as genetics and linear programming will be applied to recommend the best course of action. For example, prescriptive analysis predicts ideal milking intervals or recommends optimal feeding times to maximize productivity while reducing stress on animals.

4.2 Techniques from Related Work

In our literature, key strategies include machine learning methods, iot applications, intelligent data-driven decision support, health monitoring systems, precision dairy farming techniques and decision support systems, among others According to Ferris et al. (2020)[1] and Cabrera et al. (2020)[2], the Dairy Brain uses descriptive, prescriptive, and predictive analytics to combine data from several sources to provide real-time, actionable insights and enhance farm management. Bewley (2010)[**?**] and Wolfert et al. (2017)[4] discuss techniques for monitoring physiological, behavioral, and productive characteristics. These technologies can collect real-time data and improve decision-making by detecting diseases early and implementing preventive care. Furthermore, Andonovic et al. (2016)[5] has researched the application of wireless sensor networks (wsn) in continuous, real-time health monitoring, enhancing the timely detection of health events and improving animal welfare.
The ID3SAS technique, presented by Araújo et al. (2023)[6], combines cloud computing, machine learning, iot and sensing technologies to deliver real-time decisions for operational optimization and efficiency gains. In order to improve productivity and animal welfare, Akbar et al. (2020)[7] emphasis is on the use of artificial intelligence (AI), cloud computing, and sophisticated sensing technologies to monitor variables including cow behavior, health, and milk production. Cockburn (2020)[8] highlights that even if current systems often lack proper data integration, machine learning algorithms have great potential for analyzing large data sets, predicting milk production, and other aspects of farm management.
These studies highlight the importance of integrating data, leveraging real-time analytics, and utilizing sophisticated techniques to optimize dairy farming practices.

4.3 Evaluation of Success

The following criteria will be used to assess the effectiveness of the model adopted:

- **Accuracy of Predictions:** mean absolute error (MAE), root mean square error and other indicators (RMSE) and r squared.
- **Improvement in Milk Yield:** Measurements will be made and compared with baseline data.

- **Animal Welfare and Health:** We will track indicators such as disease incidence to look at the overall health of animals.
- **Operational Efficiency:** Evaluate the farm's operations, labor and resource utilization.

4.4 Potential Hurdles

There are some obstacles to using these technologies:

- **Data Integrity and Quality:** Because data comes from different sources, it is difficult to guarantee the correctness and integrity of data when combining them into a coherent framework.
- **Real-Time Processing:** Creating algorithms that can efficiently analyze data in real time and provide useful insights.
- **Scalability:** Ensure that the solution is suitable for a variety of farm configurations.

5 Timeline

The provided Gantt chart (Figure 5) outlines our project timeline for the next semester, spanning 12 weeks.

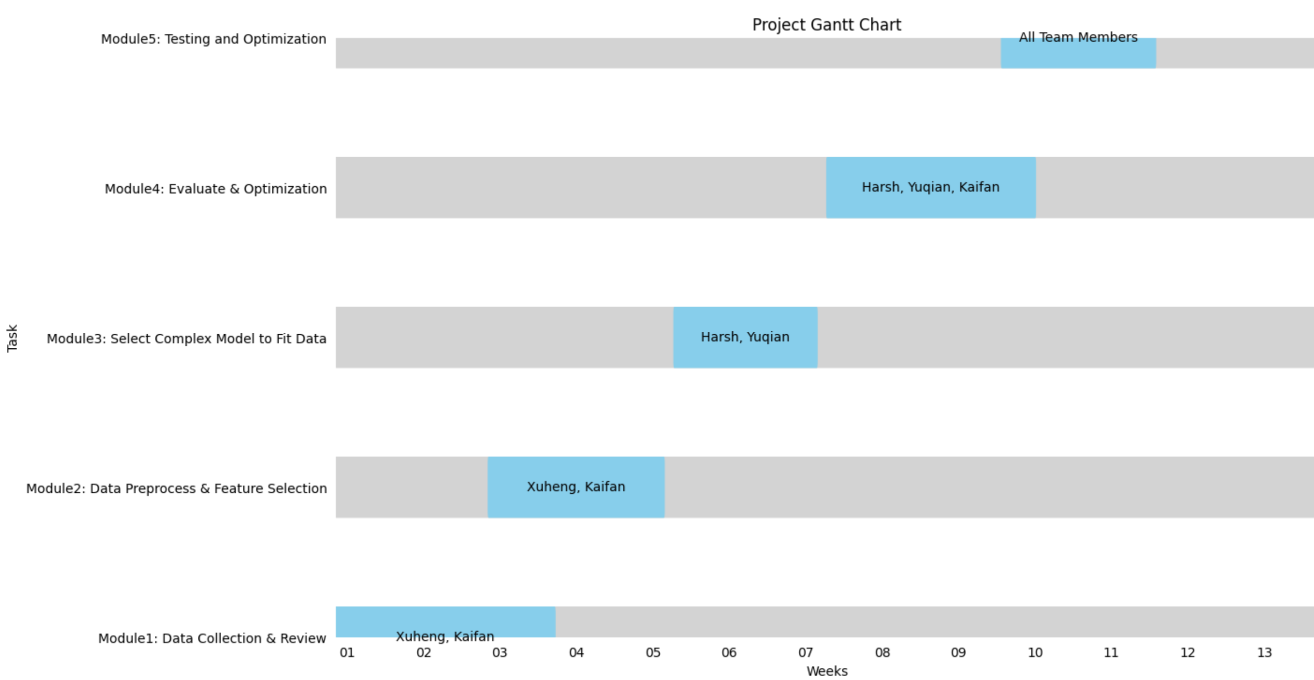

Figure 5: The Gantt chart displaying the second-semester project schedule

6 Conclusion

In the first semester, we undertook an in-depth analysis of dairy data, which provided us with an initial understanding and insight into the whole topic. In the process, we have consulted a large amount of relevant literature and gained a deeper understanding of the main characteristics and problems of dairy farms. These literature reviews not only enrich our theoretical knowledge, but also help us identify specific Challenges in practice.

Through this semester's exploration and research, we have identified several challenges that need to be addressed in the next semester. For example, we found that the existing data set has certain limitations, which will affect our further analysis and optimization of automated milking processes. Therefore, we plan to seek and integrate more diverse and high-quality data in the next semester to overcome the limitations of these questions.

To sum up, through the experiments and research in this semester, we have a profound and unique insight into the whole subject, and have formulated a detailed implementation plan for the next semester. This not only lays a solid foundation for our next stage of research, but also provides a clear direction for our further exploration in the field of dairy big data.

7 Team Member Contributions

Group Member	**Contribution Factor**
Yuqian Xie	0.25
Harsh Mangla	0.25
Xuheng Li	0.25
KAIFAN OUYANG	0.25
Total	**1**

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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030220301454

[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. ELSEVIER SCIENCE INC STE 800, 230 PARK AVE, NEW YORK, NY 10169 USA, 2019, pp. 273–273.

[3] J. Bewley, "Precision dairy farming: Advanced analysis solutions for future profitability," in *The First North American Conference on Precision Dairy Management*, vol. 16, 2010.

[4] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308521X16303754

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

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

[7] M. O. Akbar, M. S. Shahbaz khan, M. J. Ali, A. Hussain, G. Qaiser, M. Pasha, U. Pasha, M. S. Missen, and N. Akhtar, "Iot for development of smart dairy farming," *Journal of Food Quality*, vol. 2020, pp. 1–8, 2020.

[8] M. Cockburn, "Application and prospective discussion of machine learning for the management of dairy farms," *Animals*, vol. 10, no. 9, p. 1690, 2020.

[9] M. T. Verde, R. Matera, F. Bonavolonta, F. Lamonaca, L. Angrisani, C. Fezza, L. Borzacchiello, A. Cotticelli, and G. Neglia, "Comparative performance analysis between two different generations of an automatic milking system," *Acta Imeko*, vol. 12, no. 4, pp. 1–6, 2023.