**Aivancity | Substrate AI**

**AI in Dairy Farming**

**AI in Dairy Farming**

**AI Clinic Report**

**Group members (PGE 5):**

* Bryan TCHAKOTE
* Daniela WOUELEBAK
* Stella DETIO

*Academic Year 2023-2024*

# Table of Contents

[Table of Contents 2](#_Toc174203851)

[List of Tables 3](#_Toc174203852)

[List of Figures 4](#_Toc174203853)

[I. Problematic 5](#_Toc174203854)

[II. Data 6](#_Toc174203855)

[1. Data extraction 6](#_Toc174203856)

[2. Data processing and analysis 7](#_Toc174203857)

[III. Approaches 13](#_Toc174203858)

[1. Expert algorithm 13](#_Toc174203859)

[2. Machine Learning 17](#_Toc174203860)

[IV. Conclusion 19](#_Toc174203861)

[V. Appendix 20](#_Toc174203862)

[Appendix A 20](#_Toc174203863)

[Appendix B 21](#_Toc174203864)

[Appendix C 22](#_Toc174203865)

# List of Tables

[Table 1 - Number of farms for each species and breed 7](#_Toc174203436)

[Table 2 - Data sample (1) 8](#_Toc174203437)

[Table 3 - Measurement ranges example 8](#_Toc174203438)

[Table 4 - Data sample (1) trigger values 9](#_Toc174203439)

[Table 5 - Availability of measurement ranges per farms (red color = missing) 12](#_Toc174203440)

[Table 6 - Data sample (2) and trigger values 14](#_Toc174203441)

[Table 7 - Data sample (2) target 14](#_Toc174203442)

[Table 8 - Data sample (3) and trigger values 15](#_Toc174203443)

[Table 9 - Data sample (3) target 15](#_Toc174203444)

[Table 10 - Data sample (4) and trigger values 15](#_Toc174203445)

[Table 11 - Botanic pills and related criteria 15](#_Toc174203446)

[Table 12 - Comparison of the sum of ranks per botanic pill 16](#_Toc174203447)

[Table 13 - Data sample (4) target (1) 16](#_Toc174203448)

[Table 14 - Comparison of the number of inadequate parameters 16](#_Toc174203449)

[Table 15 - Data sample (4) target (2) 16](#_Toc174203450)

[Table 16 - Sum of ranks and number of inadequate parameters all equal 16](#_Toc174203451)

[Table 17 - Data sample (4) target (3) 16](#_Toc174203452)

[Table 18 - Target distribution 17](#_Toc174203453)

[Table 19 - Machine learning models features 17](#_Toc174203454)

[Table 20 - Prediction metrics 18](#_Toc174203455)

[Table 21 - Prediction confusion matrixes 18](#_Toc174203456)

# List of Figures

[Figure 1 - Relational diagram of the tables used in our project 6](#_Toc174203343)

[Figure 2 - Trigger values and zones 8](#_Toc174203344)

[Figure 3 - Data recording period per farm 9](#_Toc174203345)

[Figure 4 - Number of samples and date differences per farm 10](#_Toc174203346)

[Figure 5 - Average day difference between measurements 10](#_Toc174203347)

[Figure 6 - Distribution of some parameters per farm 11](#_Toc174203348)

[Figure 7 - Measurement trigger values per farm 11](#_Toc174203349)

[Figure 8 - Distribution of Bacteria trigger value per farm 12](#_Toc174203350)

[Figure 9 - Result assignment process 14](#_Toc174203351)

[Figure 10 - XGBoost feature importance 18](#_Toc174203352)

[Figure 11 - Measurements' distributions per animal species 20](#_Toc174203353)

[Figure 12 - Trigger values’ ranges per farm 21](#_Toc174203354)

[Figure 13 - Trigger values' distributions per farm 23](#_Toc174203355)

# Problematic

The dairy farming industry is continually evolving, driven by the need to enhance milk quality and production efficiency. As consumer demand for high-quality dairy products grows, farmers are increasingly seeking innovative solutions to optimize their operations and improve the nutritional profile of their milk. One promising approach to achieving these goals is the development of intelligent systems that can predict and recommend suitable botanical supplements for dairy animals. These botanical pills are formulated based on various parameters such as milk yield, species, protein content, fat content, lactose levels, and other relevant factors.

This project focuses on leveraging Artificial Intelligence (AI) and the Internet of Things (IoT) to develop a system capable of predicting the most suitable botanical supplement for dairy animals. By analyzing comprehensive data collected from dairy farms, the system aims to provide tailored recommendations that enhance milk quality, improve animal health, and increase overall farm productivity. The specific objectives of this project include:

1. Collecting and analysing data: Gather extensive data from dairy farms, including parameters such as milk quantity, species, protein rate, fat rate, lactose levels, and other relevant factors.
2. Developing predictive models: Utilize AI to create predictive models that analyze the collected data and recommend the most effective botanical supplements.
3. Enhancing milk quality: Implement the system to improve the nutritional quality of milk by optimizing the botanical supplements given to the animals.
4. Supporting farmer decision-making: Provide farmers with actionable insights and recommendations to help them make informed decisions about supplementing their dairy animals' diets.

However, a critical aspect of this project is to determine the necessity of Machine Learning (ML) in achieving these goals. While ML offers powerful tools for predictive analytics, it is essential to explore whether simpler, non-ML approaches can also meet the project's objectives effectively. By evaluating different methodologies, we aim to identify the most practical and efficient solution for predicting suitable botanical supplements in dairy farming.

The central problem this project addresses is the optimization of botanical supplements to improve milk quality, considering the diverse and dynamic factors influencing dairy production. The challenge lies in accurately predicting the right supplement for each animal, based on real-time data and individual characteristics. Through this project, we aim to develop a robust system that empowers dairy farmers with precise recommendations, ultimately leading to better milk quality and enhanced farm productivity.

# Data

## Data extraction

A screenshot of a computer

Description automatically generated

Figure 1 - Relational diagram of the tables used in our project

To extract data for the report, we first establish a connection to the production database using a connection string. The connection string specifies the server's name, database name, and security details such as user ID and password. Once connected, we can access the various tables in the database, those are shown in Figure 1.

The data consists of multiple Excel sheets, each with a specific purpose:

* **ManagementArea**: List of farms for which we are picking date.
* **MilkMeasurements**: Each row represents a milk laboratory analysis for a single farm from a single date.
* **MilkProduction**: Each row represents a single milk quantity got from a group of animals for a single farm on a single date.
* **MedicinalBlend**: Each row represents a single botanic pill.
* **MeasurementType**: Each row represents a single milk parameter that is being used in our analysis.
* **MeasurementRange**: Represents the list of rules of TargetValues (TV) that each farm has decided to apply to each milk parameter.
* **SupplementMatrixValue**: Represents or recommendation matrix. Each row links each botanic pill (SupplementId) with a milk parameter (MeasurementTypeId) indicating that it is recommended when the parameter is either Low / High, and its importance (Rank).
* **AnimalSpecies**: The list of animal species.
* **AnimalBreed**: The list of animal breeds and their corresponding species.
* **AnimalGroup**: The table that links animals with their corresponding farms.

## Data processing and analysis

The data extracted, we then focused on analysis and preprocessing. Following the previous diagram, we applied all necessary merges to gather all relevant information in a single table that’s going to be exploited further.

**MilkMeasurements** is the pivotal table in our work, it’s where all data used to predict the milk quality is stored. Information from other tables is going to be appended to this one. After merging the data with the farms table (**ManagementArea**) to exclude potential measures coming from unknown sources since the intervals used to qualify the criticality level of each parameter depend on the different farms, we deleted some columns from **MilkMeasurements**:

* **NumberOfAnimals**: its value, always set at 0, does not allow us to exploit it, as it is used in the denominator of the calculation used to qualify the criticality level of the milk quantity
* **Lactose**, **Casein**: they are not involved in the recommendation process (moreover, almost all the values (resp. 42 and 84%) are 0, an equaivalent for Null value)
* **Inhibitors**, **MilkQualityValue**, **IsConsolidated**: they are not understandable
* **SampleId**, **LastModified**: the information contained is not relevant

Further, we noticed and corrected some formatting errors in the **FreezingPoint** column: one value was negative, and almost all values (75%) are in range [356, 639] instead of being < 1. We divided the values by 1000 to normalize the latter.

Although this information is not used in the recommendation algorithm, we felt it necessary to add information on the breeds and species of animals producing the milk studied. Data was retrieved from **AnimalBreed**, **AnimalSpecies**, and **AnimalGroup**, and the species names was then converted from Spanish to English. It’s worth noting that animals from two farms were not identified in our database leading to a slight loss of information (179/3529 records or 5% of the records). As we can observe in Table 1, almost all animals in the study are goats of different breeds. Only one breed of sheep and cow are considered.

|  |  |  |
| --- | --- | --- |
| **Species** | **Breed** | **NumberOfFarms** |
| Goat | Florida | 5 |
| Goat | Malagueña | 6 |
| Goat | Murciano-granadina | 3 |
| Goat | Payoya | 3 |
| Sheep | Assaf | 1 |
| Cow | Frisona | 1 |

Table 1 - Number of farms for each species and breed

In what follows, we'll be tackling the measurements that feed directly into the botanical pill recommendation algorithm proposed by our partners. A first operation we need to perform concerns measurement names which are saved in English in the main table, but in Spanish in the **MeasurementType** table that matches the milk characteristics labels to their ids. We then translate Spanish names to English ones. Secondly, we evaluate for each record the number of null values represented by a “0”. Rows with strictly more than 2 missing values were then dropped. After this operation we kept 3148/3350 records (about 95%).

As each measured value is categorized according to intervals defined by the veterinarians of each farm (**MeasurementRange**), we associate each observed value on each parameter with a trigger value: **Danger Zone Low**, **Low**, **Adequate**, **High**, and **Danger Zone High**.

A diagram of value and value

Description automatically generated

Figure 2 - Trigger values and zones

Note again that a value at 0 isn't admissible in the recommendation process, as well as values of the following columns: **Quantity**, **NumberOfAnimals**, **Casein**, and **Lactose**.

To perform this action, we need to melt the actual table to obtain a longer representation of it. In fact, each line of our table represents a milk measurement record with all its characteristics: **Fat**, **Protein**, **ES[[1]](#footnote-1)**, **EQ[[2]](#footnote-2)**, **Bacteria**, **SomaticCellCount**, **Urea**, and **FreezingPoint**. Knowing the farm from which that measurements come from, we now need to map each of the observed values with its corresponding trigger value, such that:

This initial record[[3]](#footnote-3)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ManagementAreaId** | **Fat** | **Protein** | **ES** | **EQ** | **Bacteria** | **SomaticCellCount** | **Urea** | **FreezingPoint** |
| 1000 | 6 | 4 | 9 | 0 | 200 | 100 | 500 | 0 |

Table 2 - Data sample (1)

With these measurement ranges

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ManagementAreaId (or Farm)** | **Danger Low Value** | **Low Value** | **High Value** | **Danger High Value** | **Measurement** |
| 1000 | 3 | 4 | 7 | 8 | Fat |
| 1000 | 3 | 5 | 7 | 10 | Protein |
| 1000 | 6 | 7 | 10 | 12 | ES |
| 1000 | 6 | 8 | 11 | 14 | EQ |
| 1000 | 6 | 10 | 50 | 100 | Bacteria |
| 1000 | 200 | 400 | 1500 | 2500 | SomaticCellCount |
| 1000 | 400 | 450 | 550 | 600 | Urea |
| 1000 | 0.45 | 0.55 | 0.59 | 0.66 | FreezingPoint |

Table 3 - Measurement ranges example

Leads to this melted result

|  |  |  |
| --- | --- | --- |
| **Farm** | **Measurement** | **TriggerValue** |
| 1000 | Fat | Adequate |
| 1000 | Protein | Low |
| 1000 | ES | Adequate |
| 1000 | EQ | *Not included (value = 0)* |
| 1000 | Bacteria | Danger Zone High (**DZ High**) |
| 1000 | SomaticCellCount | Danger Zone Low (**DZ Low**) |
| 1000 | Urea | Adequate |
| 1000 | FreezingPoint | *Not included (value = 0)* |

Table 4 - Data sample (1) trigger values

From the 3148 input records at this step, we come out with 2761 (12% of data loss) rows after merging the data since the **MeasurementRange** table doesn’t provide classification intervals for 12 farms.

This is the first step in the botanical pill recommendation algorithm, if required. But before going into detail about this process, let's highlight a few results from the data analysis.

* **Data recording period**

Figure 3 outlines that the start dates of measurements varied significantly across farms, ranging from March 2022 to July 2023. In contrast, the end dates were more consistent, falling between September and November 2023. This variation in starting periods might have implications on the volume of data collected, we will carry out an in-depth analysis to address this question.

A graph with different colored lines

Description automatically generated

Figure 3 - Data recording period per farm

* **Number of samples per farm and breed**

A graph of different colored bars

Description automatically generated

Figure 4 - Number of samples and date differences per farm

There is a strong correlation between the study period and the number of samples collected across most farms, indicating that longer periods lead to more data collection (Figure 4). However, the data collected for the Payoya goat breed is rather special in that many milk samples were analyzed over relatively short periods of time. It has been verified that in most cases only one milk measurement is made per day; the rule was broken only 13 times.

If we look at the number of days between two successive measurements, we see that while on average this gap is around 2 days for all breeds, it rises to 6 days for Payoya farms (Figure 5).

A graph of a number of red bars

Description automatically generated with medium confidence

Figure 5 - Average day difference between measurements

* **Measurements distributions per farm, species, and breed**

The analysis of milk characteristics reveals significant variability across different farms and breeds for parameters such as **EQ**, **Fat**, **FreezingPoint**, **Protein**, and **Urea**. This suggests that species or breed can indeed impact these milk characteristics. More specifically, we can observe a strong correlation between the values and the species. Farm 1867, which raises sheep, shows very high overall parameter values. In contrast, Farm 1917, which raises cows, exhibits very low overall parameter values. However, some parameters like **Bacteria** and **SomaticCellCount** show uniformity across different breeds and farms, indicating that these specific characteristics might be less influenced by the breed or farm-specific factors. Finally, the **ES** distribution across farms is very atypical: farms 1902, 1901 (all Malagueña, Goat), and 1905 (Murciano-granadina, Goat) have surprisingly high values. It would be prudent to explore the influence of breed on milk parameters to determine if there are underlying genetic or environmental factors contributing to these variations. Relevant graphics are in Appendix A.

A graph of different colored and black lines

Description automatically generated with medium confidenceA graph showing different colored and black lines

Description automatically generated with medium confidence

Figure 6 - Distribution of some parameters per farm

* **Measurement trigger values per farm**

The graphs show significant variations across the different parameters and farms. Some parameters such as **FreezingPoint** and Bacteria show high inter-farm consistency, suggesting uniform control practices. Other parameters such as **Urea**, **Protein** and **Bacteria** show considerable differences between farms and clusters, indicating that a clustered approach to the evaluation and improvement of management practices is required. It's worth noting that one more time farms 1867 and 1917 which have different type of animal species (resp. sheep and cow) stand out in this trigger values also: the first one have higher boundary values for **Protein**, **Fat** and **EQ**, while the second one has lower values for **Urea**, **SomaticCellCount** and **Bacteria**, which aligns with the previous distribution plots. More details can be found in Appendix B.

A graph with blue dots and numbers

Description automatically generatedA graph of a graph with colored lines

Description automatically generated with medium confidence

Figure 7 - Measurement trigger values per farm

Finally, we observe that the measurement ranges of certain params for some farms are missing:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Farm** | **1524** | **1908** | **1898** | **1909** | **1910** | **750** | **1906** | **1903** | **1902** | **1901** | **1907** | **1911** | **1905** | **1912** | **1894** | **1893** | **1892** | **1867** | **1917** |
| **Bacteria** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **ES** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **EQ** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fat** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **FP\*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Protein** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **SCC\*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Urea** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

\*FP: FreezingPoint  
\*\*SCC: SomaticCellCount

Table 5 - Availability of measurement ranges per farms (red color = missing)

* **Distribution of trigger values per farm**

We have examined the distribution of trigger values for each parameter across all farms to analyze in detail the value zones and their distribution.

First, we note that the values for the **Bacteria** parameter are generally in the adequate zone, with a significant portion of the data relatively evenly distributed in the **High** and **Danger Zone High** areas. The **EQ**, **Fat**, and **Protein** parameters mostly have their values concentrated in the **Low** zone, then in the **Adequate** zone, while the **FreezingPoint** is generally **Adequate** and occasionally in the **Low** zone. Along with **Bacteria**, the parameters **SomaticCellCount** and **Urea** often have extreme values, particularly in the **Danger Zone High**. Most of the values for **SomaticCellCount** are in the **High** zone, while the distribution of these values across farms is quite random. A similarly random distribution is also observed for the **ES** parameter, with different profiles depending on the farms.

Overall, we notice that most of the parameters are often problematic (quantities below or above the recommended standards), except for **FreezingPoint**, which mostly has values in the **Adequate** zone. It is also important to highlight that the variables **Bacteria**, **SomaticCellCount**, and **Urea** are very often responsible for human intervention due to their values indicating an extremely concerning situation with excessively high levels (**Danger Zone High**). Appendix C contains other matrixes for further analysis.

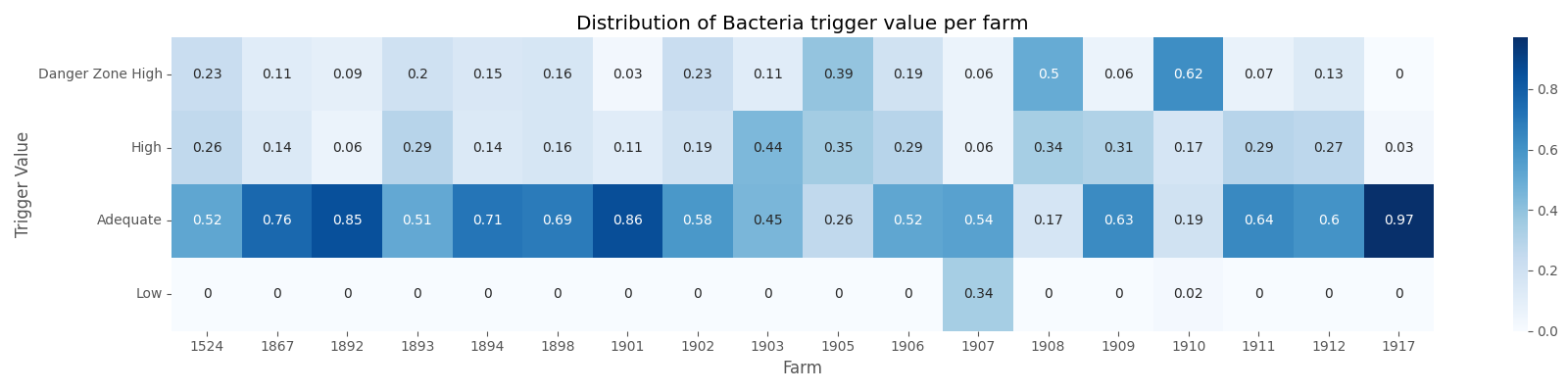


Figure 8 - Distribution of Bacteria trigger value per farm

# Approaches

## Expert algorithm

The following algorithm represents the sequential steps provided by the company. It is used to process and generate the final dataset used for predicting botanic pills efficacy in our machine learning models. The algorithm is structured as follows:

1. **Collection of information**

* List of farms
* List of feeds
  + For each feed, list of feed recommendation allocation rules

E.g.: feed **BP\_X** recommended when Fat is high, and Protein is low

* Milk quality and quality data
  + Milk quality is obtained from the data of the last laboratory analysis
  + The average milk production per animal per day is calculated with a formula including the number of animals, which is always set to zero, which is why this variable is not considered in our solution
* Targets/Critical values for each parameter by farm

1. **Computational matrix creation algorithm**

For each observation of a parameter in a farm sample, a result is computed:

* **Appropriate value:** The parameter is in normal values, no action is required
* **Danger Zone value:** The parameter is in a dangerous range (**much** **higher or lower** than normal), a notification for human intervention is generated
* **Inadequate value:** The parameter is in an inadequate range (**a little** **higher or lower** than normal), we may delve into the recommendation algorithm to assign a botanic pill to solve the problem

**IMPORTANT**: Any parameter whose value is 0 is ignored and is not considered for the generation of recommendations

1. **Result assignment algorithm**

The algorithm compares the results for all parameters of each farm and chooses one of the following cases:

A diagram of a diagram

Description automatically generated

Figure 9 - Result assignment process

1. **All parameters have appropriate values**

*Nothing is done, all parameters are correct.*

Considerate our following melted data obtained after transformations in data preparation:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Farm** | **Measurement** | **Value** | **Danger Low Value** | **Low Value** | **High Value** | **Danger High Value** | **Trigger Value** |
| 1000 | Fat | 5 | 3 | 4 | 7 | 8 | **Adequate** |
| 1000 | Protein | 5.3 | 3 | 5 | 7 | 10 | **Adequate** |
| 1000 | ES | 8 | 6 | 7 | 10 | 12 | **Adequate** |
| 1000 | EQ | 0 | 6 | 8 | 11 | 14 | *Not included* |
| 1000 | Bacteria | 30 | 6 | 10 | 50 | 100 | **Adequate** |
| 1000 | SomaticCellCount | 600 | 200 | 400 | 1500 | 2500 | **Adequate** |
| 1000 | Urea | 455 | 400 | 450 | 550 | 600 | **Adequate** |
| 1000 | FreezingPoint | 0.57 | 0.45 | 0.55 | 0.59 | 0.66 | **Adequate** |

Table 6 - Data sample (2) and trigger values

Then the output is:

|  |  |
| --- | --- |
| **Farm** | **Target** |
| 1000 | **No treatment** |

Table 7 - Data sample (2) target

1. **At least one parameter is in Danger Zone**

*Notification is generated for human intervention.*

Let’s take an example:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Farm** | **Measurement** | **Value** | **Danger Low Value** | **Low Value** | **High Value** | **Danger High Value** | **Trigger Value** |
| 1000 | Fat | 5 | 3 | 4 | 7 | 8 | **Adequate** |
| 1000 | Protein | 2.5 | 3 | 5 | 7 | 10 | **DZ Low** |
| 1000 | ES | 8 | 6 | 7 | 10 | 12 | **Adequate** |
| 1000 | EQ | 0 | 6 | 8 | 11 | 14 | *Not Included* |
| 1000 | Bacteria | 4 | 6 | 10 | 50 | 100 | **DZ Low** |
| 1000 | SomaticCellCount | 600 | 200 | 400 | 1500 | 2500 | **Adequate** |
| 1000 | Urea | 700 | 400 | 450 | 550 | 600 | **DZ High** |
| 1000 | FreezingPoint | 0.57 | 0.45 | 0.55 | 0.59 | 0.66 | **Adequate** |

Table 8 - Data sample (3) and trigger values

Output:

|  |  |
| --- | --- |
| **Farm** | **Target** |
| 1000 | **Veterinarian intervention** |

Table 9 - Data sample (3) target

1. **All other cases (no parameters in Danger Zone, at least one parameter with inappropriate value)**

*Feed allocation algorithm is passed.*

Let’s look at the example below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Farm** | **Measurement** | **Value** | **Danger Low Value** | **Low Value** | **High Value** | **Danger High Value** | **Trigger Value** |
| 1000 | Fat | 5 | 3 | 4 | 7 | 8 | **Adequate** |
| 1000 | Protein | 3.5 | 3 | 5 | 7 | 10 | **Low** |
| 1000 | ES | 8 | 6 | 7 | 10 | 12 | **Adequate** |
| 1000 | EQ | 0 | 6 | 8 | 11 | 14 | *Not Included* |
| 1000 | Bacteria | 70 | 6 | 10 | 50 | 100 | **High** |
| 1000 | SomaticCellCount | 600 | 200 | 400 | 1500 | 2500 | **Adequate** |
| 1000 | Urea | 555 | 400 | 450 | 550 | 600 | **High** |
| 1000 | FreezingPoint | 0.57 | 0.45 | 0.55 | 0.59 | 0.66 | **Adequate** |

Table 10 - Data sample (4) and trigger values

First, we must determine for each feed which parameter assignment rules have been met:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Botanic Pill (BP)** | **Measurement** | **Criteria** | **Trigger Value (inputs)** | **Rank of BP** |
| **BP\_1** | Protein | Low | Low | 3 |
| Bacteria | Low | High | 4 |
| Urea | High | High | 5 |
| **BP\_2** | FreeZingPoint | Low | Adequate | 8 |
| Bacteria | High | High | 9 |
| **BP\_3** | Protein | Low | Low | 7 |
| EQ | Low | Adequate | 1 |
| Urea | High | High | 5 |

Table 11 - Botanic pills and related criteria

The feed with the highest score (sum of ranks of the inadequate parameters) is recommended:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **BP** | **Measurement** | **Criteria** | **Trigger Value** | **Rank of BP** | **Sum of Rank** |
| **BP\_1** | Protein | Low | Low | 3 | **8** |
| Bacteria | Low | High | 4 |
| Urea | High | High | 5 |
| **BP\_2** | FreeZingPoint | Low | Adequate | 8 | **9** |
| Bacteria | High | High | 9 |
| **BP\_3** | Protein | Low | Low | 7 | **12** |
| EQ | Low | Adequate | 1 |
| Urea | High | High | 5 |

Table 12 - Comparison of the sum of ranks per botanic pill

Output:

|  |  |
| --- | --- |
| **Farm** | **Target** |
| 1000 | **BP\_3** |

Table 13 - Data sample (4) target (1)

In case of a tie, the one with the highest number of inappropriate parameters is assigned:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **BP** | **Measurement** | **Criteria** | **Trigger Value** | **Rank of BP** | **Σ of Rank** | **# of Params** |
| **BP\_1** | Protein | Low | Low | 3 | **12** | **3** |
| Bacteria | Low | High | 4 |
| Urea | High | High | 5 |
| **BP\_2** | FreeZingPoint | Low | Adequate | 8 | **9** | **-** |
| Bacteria | High | High | 9 |
| **BP\_3** | Protein | Low | Low | 7 | **12** | **2** |
| EQ | Low | Adequate | 1 |
| Urea | High | High | 5 |

Table 14 - Comparison of the number of inadequate parameters

Output:

|  |  |
| --- | --- |
| **Farm** | **Target** |
| 1000 | **BP\_1** |

Table 15 - Data sample (4) target (2)

**NB**: The company's algorithm ignores ties between feeds with the same inappropriate parameters. We chose to output all tied feeds:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **BP** | **Measurement** | **Criteria** | **Trigger Value** | **Rank of BP** | **Σ of Rank** | **# of Params** |
| **BP\_1** | Protein | Low | Low | 3 | **12** | **2** |
| Bacteria | Low | High | 4 |
| Urea | High | High | 9 |
| **BP\_2** | FreeZingPoint | Low | Adequate | 8 | **9** | **-** |
| Bacteria | High | High | 9 |
| **BP\_3** | Protein | Low | Low | 7 | **12** | **2** |
| EQ | Low | Adequate | 1 |
| Urea | High | High | 5 |

Table 16 - Sum of ranks and number of inadequate parameters all equal

Output:

|  |  |
| --- | --- |
| **Farm** | **Target** |
| 1000 | **[BP\_1, BP\_3]** |

Table 17 - Data sample (4) target (3)

## Machine Learning

The resulting data frame after the algorithm application represents for each row, each milk laboratory analysis, corresponding to a single farm on a specific date. Due to an imbalanced target value (Table 18), we decided to group samples where the predicted class was a given botanic pill as a single **Treatment** value, and to drop the **No treatment** class. Consequently, we have two unique values as target: **Treatment**, and **Veterinarian intervention**.

|  |  |  |
| --- | --- | --- |
| **Target** | | **Count** |
| **Veterinarian intervention** | | **1136** |
| **No treatment** | | **15** |
| **Treatment** | 1 | 1017 |
| 42 | 5 |
| 45 | 1 |
| [38, 39, 40, 42, 46, 47] | 337 |
| [38, 40, 42] | 101 |
| [38, 39, 42] | 78 |
| [39, 42] | 40 |
| [38, 42] | 20 |
| [44, 45] | 6 |
| [38, 39, 40, 41, 42, 43, 44, 45, 46, 47] | 2 |
| [1, 38, 39, 40, 42, 46, 47] | 1 |
| [38, 40, 41, 42, 43] | 1 |
| [38, 39, 42, 43] | 1 |
| ***Total*** | **1610** |

Table 18 - Target distribution

The following features are considered in the database:

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Type | Unit | Description |
| Fat | float | Percentage | Fat proportion |
| Protein | float | Percentage | Protein proportion |
| Dry Extract (ES) | float | Percentage | Residual mass after drying the milk |
| Cheese extract (EQ) | float | Percentage | Residual mass for cheese making |
| Bacteria | int | Colony-forming unit / Milliliter\*1000 | Quantity of bacteria |
| SomaticCellCount | int | Cells / milliliter\*1000 | Number of cells of type “somatic” |
| Urea | float | Milligram per Liter | Quantity of urea |
| FreezingPoint | float | Degree Celsius | Freezing temperature of the milk |
| Breed | object | - | Animal breed |
| Species | object | - | Animal species |
| Target | object | - | Whether to give a treatment, notify a veterinary, and do nothing |

Table 19 - Machine learning models features

We developed three machine learning models: Decision Tree, Random Forest, and XGBoost. The data preparation steps are as follows:

* Drop **Id** and **ManagementAreaId** columns from the transformed data
* Split the data into training (70%) and testing (30%) sets
* Encode the training and testing features splits (X\_train & X\_test):
  + Encode categorical features using **OneHotEncoder**
  + Encode numerical features using **OrdinalEncoder**
* Encode the training and testing labels splits (y\_train & y\_test) using **LabelEncoder**

Here are the results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy** | **Precision (w. avg)** | **Recall (w. avg)** | **F1-score (w. avg)** |
| **Decision Tree** | 91% | 91% | 91% | 91% |
| **Random Forest** | 92% | 92% | 92% | 92% |
| **XGBoost** | 94% | 94% | 94% | 94% |

Table 20 - Prediction metrics

The confusion matrixes (**T: Treatment - VI: Veterinary intervention**):

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Decision Tree** | | |  | **Random Forest** | | |  | **XGBoost** | | |
| **True** | **T** | 443 | 40 | **T** | 461 | 22 | **T** | 464 | 19 |
| **VI** | 37 | 304 | **VI** | 41 | 300 | **VI** | 32 | 309 |
|  | **T** | **VI** |  | **T** | **VI** |  | **T** | **VI** |
|  | **Predicted** | | | | | | | | | | |

Table 21 - Prediction confusion matrixes

\*

Considering overall performance and confusion matrices, **XGBoost** appears to be the best model with an overall **accuracy of 94%**. The **same performance** is observed in **precision**, **recall**, and **f1-score**. Though Random Forest and Decision Tree also demonstrate good performance, XGBoost consistently outperforms them in all metrics and offers the most reliable predictions, it's thus the best performing model for this scenario.

When considering XGBoost for the final model selection, it's crucial to understand the impact of different features on its performance, as illustrated in Figure 10. **Bacteria** and **SomaticCellCount** are the most significant features, **each contributing more than 40% to the model's feature importance**. These variables play a critical role in the model's decision-making process. Urea and ES also make contributions, albeit much smaller compared to the top two features. All other features have negligible importance in the model. The significant reliance on Bacteria and SomaticCellCount highlights these as critical variables impacting the model's predictions.

A graph with red and white text

Description automatically generated

Figure 10 - XGBoost feature importance

# Conclusion

This project demonstrates the potential of using AI in dairy farming to optimize the nutritional quality of milk through recommendations for botanical supplements. By integrating milk measurements, animal species and breeds data, and farm-specific parameters, we developed a system capable of providing farmers insights on the need of a treatment or to a veterinarian intervention.

We gathered and processed a vast amount of data from multiple sources, ensuring that the final data set was comprehensive and representative of different farms and animal breeds. The initial approach utilized an expert-driven algorithm provided by our partner to recommend botanical supplements based on observed milk parameters. This rule-based method provided a clean dataset, the basis of our machine learning experimentations. We explored the use of three ML models - Decision Tree, Random Forest, and XGBoost - to predict the need for treatment or veterinarian intervention. The XGBoost model outperformed the others, achieving a 94% accuracy rate, and proved to be the most reliable in making predictions. Key features driving the model's predictions were identified, with Bacteria and SomaticCellCount being the most influential. While the expert algorithm provided a straightforward approach to recommendations, it lacked the necessary validation from our partner. The absence of a thorough validation process raises concerns about the reliability of the recommendations, particularly in real-world scenarios. Moreover, the challenges were further compounded by the presence of missing data, particularly in the measurement ranges and the identification of certain farms and animals, and the exclusion of important parameters like milk quantity. These gaps in data meant that the algorithm's output could not make full use of all the available information, potentially leading to suboptimal recommendations.

Future work could involve integrating more diverse data sources, such as environmental conditions, feed composition, and animal health records, to further refine the predictive models, and even collect enough data to build a bill recommendation system that predicts the right botanic pills in case a treatment is required. Moreover, while ML has shown great promise, exploring other statistical or heuristic methods could provide additional insights or alternative solutions, especially for farms with limited data. Finally, conducting long-term studies to assess the impact of the recommended botanical supplements on milk quality and farm productivity would provide valuable feedback for continuous improvement of the system.

# Appendix

## Appendix A

A graph showing different colored numbers

Description automatically generated with medium confidenceA graph of different colored and black lines

Description automatically generated with medium confidenceA graph showing different colored and black lines

Description automatically generated with medium confidenceA graph of different colored and black lines

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generatedA graph of different colored and black lines

Description automatically generatedA graph of a number of cells

Description automatically generated with medium confidenceA graph of different colored and black lines

Description automatically generated with medium confidence

Figure 11 - Measurements' distributions per animal species

## Appendix B

A graph with blue and green dots

Description automatically generatedA graph of a diagram

Description automatically generated with medium confidenceA graph with blue and green dots

Description automatically generatedA graph of a fat trigger

Description automatically generatedA graph with blue dots

Description automatically generatedA graph of a protein trigger

Description automatically generatedA graph showing a number of values

Description automatically generatedA graph of a graph with colored lines

Description automatically generated with medium confidence

Figure 12 - Trigger values’ ranges per farm

## Appendix C

A blue squares with black text

Description automatically generatedA screenshot of a graph

Description automatically generatedA blue and white chart

Description automatically generated with medium confidenceA chart with blue squares

Description automatically generatedA blue squares with black text

Description automatically generatedA blue squares with black text

Description automatically generatedA diagram of a farm

Description automatically generated with medium confidence A graph of a number of blue squares

Description automatically generated with medium confidence

Figure 13 - Trigger values' distributions per farm

1. Extracto Seco or Dry Extract [↑](#footnote-ref-1)
2. Extracto Quesero or Chesse Extract [↑](#footnote-ref-2)
3. This tabular data and all following are fake while representive of our dataset. [↑](#footnote-ref-3)