HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

Applications of Artificial Intelligence in Cattle Breeding for Predicting Conformation Scores, Somatic Cell Count, and Casein Content

Doctoral (PhD) thesis

Bence Tarr

Gödöllő

(2025)

Name of doctoral school:	Doctoral School of Mechanical Engineering
Discipline:	Mechanical Engineering
Headed by:	Prof. Dr. Gábor Kalácska professor, DSc Hungarian University of Agriculture and Life Sciences Institute of Technology
Supervisor:	Prof. Dr. István Szabó professor, PhD Hungarian University of Agriculture and Life Sciences Institute of Technology
	Prof. Dr. János Tőzsér professor, DSc Széchenyi István University Albert Kázmér Faculty of Agricultural and Food Sciences

. . . . . .

Approval by the Head of the School

. . . . . .

Approval of the Supervisor

. . . . . .

Approval of the Supervisor

CONTENTS

1. INTRODUCTION, OBJECTIVES

The application of artificial intelligence in livestock farming has become a key prerequisite for further quantitative and qualitative development. Over the past decades, numerous studies have explored the potential uses of AI in animal husbandry. However, the processing of available historical data and the development of predictive and estimation models from such data have not yet received sufficient scientific emphasis.

The focus of my research is to examine existing databases in cattle breeding and to develop estimation models that are practically useful for breeders. Although efforts have been made to analyse and process such data using statistical methods, these solutions have not yet been integrated into everyday practice and have primarily relied on traditional mathematical modelling approaches.

Due to veterinary and quality assurance requirements, the livestock and dairy industries have long-established systems for data collection and analysis. As a result, extensive databases containing large volumes of historical measurements and expert evaluations are available.

One of the most rapidly advancing branches of AI, machine learning, is particularly well-suited to utilizing such datasets to create surprisingly accurate models and algorithms.

Thus, my objective is to demonstrate that models capable of generating reliable predictions for cattle breeders can be developed using already existing data.

I chose to focus on two specific areas:

The estimation of conformation scores in cattle, and
The evaluation of milk quality.

Both services are crucial in animal husbandry. In the case of conformation scoring, the high risk of subjectivity further emphasizes the need for reliable and objective tools.
My research is centered on the processing and utilization of data generated during cattle breeding and milk production. The goal is to develop machine learning-based models and algorithms that contribute to more effective planning and support more efficient breeding practices.
Over the course of my study, I aimed to develop AI-based solutions for three specific problems:

- Estimation of conformation scores in cattle
- Estimation of somatic cell count (SCC) in milk
- Estimation of casein content in milk

In addition to historical data, modern production technologies, such as robotic milking systems, also generate automatically recorded data that are well-suited to training such models.
Therefore, my aim is to process the available datasets and identify machine learning models that can be validated and applied to support predictions in breeding and production processes.
To facilitate my research, I plan to design a custom-developed software system that includes all steps of data processing, preparation, training set selection, training, and validation. This

system could later serve as a foundation for a broadly applicable tool requiring minimal investment, offering breeders a practical solution in the aforementioned areas.

1.1. Review of Literature and Future Research Directions

The vast majority of the articles I reviewed focus on the analysis of image-based data, which is also the area where the most significant advancements have been observed. This is understandable, as such systems have the greatest potential to replace field experts, whose experience and manual labour have traditionally enhanced production efficiency.

Replacing human labour with intelligent systems represents a considerable cost-saving opportunity for farms, which is one of the main reasons why research in this field has gained such popularity.

Nevertheless, it is important to emphasize that the analysis of existing databases and the collection and processing of data from other sensors are also of great value to agricultural operations.

It has also become evident that, based on the available literature and current scientific research, data-driven predictive algorithms and procedures still show limited results in several areas. Therefore, future developments in this domain will be of critical importance.

2. MATERIALS AND METHODS

In line with the research objectives, I developed, tested, and validated three distinct machine learning-based estimation and prediction models:

- Estimation of conformation scores in cattle
- Estimation of somatic cell count (SCC) in milk samples
- Estimation of casein content in milk samples

The underlying hypothesis was that these estimations can be effectively performed using machine learning models trained on existing datasets. The aim was to demonstrate that an automated AI-based system can be developed which, through continuous learning, can deliver increasingly accurate predictions in areas crucial to cattle breeders.

In the first phase of the research, the focus was on identifying available and validated databases suitable for building reliable models.

Several breeding associations in Hungary, such as the Holstein-Friesian Breeders' Association and the Hungarian Simmental Breeders' Association, regularly carry out expert-based conformation scoring using standardized evaluation protocols and certified specialists. For training my model, I used a dataset provided by the *Limousin and Blonde d'Aquitaine Breeders' Association* (Budapest), which included the conformation scores of 325 animals.

For the estimation of milk components, I obtained a dataset from a certified laboratory. This dataset was collected from three different farms over a three-year period (2019–2021). Prior to use, all data were anonymized to ensure the privacy of both the breeders and the laboratory.

Each breeder submitted monthly milk samples for laboratory analysis, resulting in 36 months of data per breeder, with a total of 25,000 measurements. Each measurement included 14 different milk parameters. However, not all parameters were used in every model. The flexibility of the dataset allowed for the development of multiple models depending on the choice of output variable.

My research specifically focused on estimating somatic cell count and casein content, as these are among the most critical indicators for breeders. SCC values provide insight into the health status of dairy cows, while casein levels strongly influence milk quality. Consequently, both models have direct practical relevance, as these parameters can significantly affect milk pricing and marketability.

2.1. Methods

To build and train the models, I employed various machine learning techniques.

In the context of machine learning, algorithms refer to mathematical or statistical models and methods that enable artificial intelligence to learn from data and draw conclusions based on that learning.

The fundamental purpose of these algorithms is to assist in pattern recognition, estimation, and problem solving.
In the following, I provide an overview of the algorithms applied during my research.

2.1.1. Linear regression

The goal of linear regression is to predict a continuous output variable (dependent variable, y) based on a linear combination of input variables (independent variables, X). It is one of the simplest and most widely used regression methods in statistical modelling and machine learning.
The mathematical model of multiple linear regression can be expressed using the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ij} + \varepsilon_i \,, \qquad (2.1)$$

where:

- y_i the output value (dependent variable) for the i-th observation;
- x_{ij} the j-th independent variable (predictor) corresponding to the i-th observation, based on which the prediction is made;
- β_j the coefficient associated with the j-th independent variable (indicating the extent to which each predictor influences the dependent variable);
- β_0 the intercept term, i.e., the predicted value of y when all $x_j = 0$;

- ε_i the error term, representing the deviation of the model from the actual data (commonly assumed to follow a normal distribution, $\varepsilon_i \sim N(0, \sigma^2)$).

2.1.2. Poisson-regression

Poisson regression is a specialized regression model used for modelling discrete, non-negative integer-valued output variables, such as event counts.

A typical example is the conformation scoring of cattle, where the outcomes can only take on a limited number of discrete values.

2.1.3. Evaluation Metrics for Regression Models

R^2 (Coefficient of Determination)

The coefficient of determination (**R^2**) measures how well the independent variables explain the variance of the target (dependent) variable. Its value ranges between 0 and 1:

- $R^2 = 1 \rightarrow$ The model perfectly explains the variance in the data.
- $R^2 = 0 \rightarrow$ The model is no better than a random guess (e.g., predicting the mean).
- $R^2 < 0 \rightarrow$ The model performs worse than simply predicting a constant value.

MSE (Mean Squared Error)

The mean squared error measures the average difference between the predicted values of the model and the actual (true) values.

Adjusted R^2

The adjusted R^2 accounts for the number of predictors in the model. While the standard R^2 always increases when more predictors are added, the adjusted R^2 penalizes unnecessary variables, helping to filter out those that do not improve the model.

2.1.4. Random Forest

Random Forest is an ensemble learning method used for both classification and regression tasks. It is based on decision trees but constructs multiple trees (a "forest") and combines their outputs — by averaging for regression or majority voting for classification — to achieve more robust and accurate results.

The algorithm operates on randomly selected data samples, generated through bootstrap sampling (sampling with replacement) from the original dataset.

2.1.5. Decision Tree Regressor

The Decision Tree Regressor is a tree-based model typically used for predicting continuous target variables. It follows the structure of a traditional decision tree, but instead of predicting categories, it outputs numerical values.

The algorithm recursively splits the dataset into smaller segments, aiming to minimize variance or another error metric within each node. In the resulting tree, the leaf nodes store the mean value of the target variable for the corresponding data subset.

At each step, the algorithm searches for the variable and threshold that result in the best possible split.

2.1.6. Extra Trees Classifier

The Extra Trees Classifier is a decision tree-based algorithm that builds multiple trees on a dataset and combines their individual predictions to make the final decision. Its main goal is to reduce variance (thus avoiding overfitting, similarly to Random Forest) and to offer faster performance on large datasets.

Unlike the Random Forest, which trains each tree on a bootstrap sample, Extra Trees uses the entire dataset for training each tree, further increasing computational efficiency.

2.1.7. Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning algorithm used for both classification and regression tasks. In classification problems, the goal is to separate the training data points into two (or more) distinct classes.

In the case of optimal linear separation, the dividing line (or hyperplane in higher dimensions) is positioned such that it lies at the maximum possible distance from the closest data points of each class. This separation is defined by a margin, or buffer zone, between the classes.

In practice, perfect linear separation is rarely feasible, especially when ensuring a margin for classification robustness. Often, only a narrow margin can be created, or no error-free linear separation is possible at all.

To address this, SVM allows some training points to fall within the margin, thereby significantly increasing the margin width.

The primary goal of SVM is not merely to separate classes by a discrete boundary, but rather to find a hyperplane that offers a better generalization and classification performance than a strict, hard boundary.

2.1.8. Evaluation Metrics for Decision Trees

Accuracy

Measures the proportion of correct predictions out of all samples.

Precision

Indicates how many of the samples predicted as positive were actually positive.

Recall

Measures how many of the actual positive cases were correctly identified by the model.

2.1.9. Traditional Mathematical Methods

The tasks performed using machine learning-based regression models can also be solved using traditional mathematical approaches. In my research, I applied these conventional methods to evaluate the performance of my models.

For this purpose, I used the open-source Python module *statsmodels*, which is designed for the computational implementation of well-validated statistical and mathematical techniques.

In particular, I utilized the *OLS* (Ordinary Least Squares) function from the *statsmodels* library to compare the outcomes of my machine learning models with those produced by traditional mathematical modelling.

2.1.10. Imbalanced Datasets

In real-world scenarios, the number of poor or critical outcomes is typically small — otherwise, existing breeding (or production) practices would be considered unrealistic. Therefore, when processing historical data, it is common to encounter imbalanced datasets.

In my research, the goal is to identify milk samples from sick animals and to predict the onset of disease. As such, the dataset is expected to be highly imbalanced. To develop an effective model, it is essential to balance the dataset; otherwise, the resulting model would be biased and inaccurate.

Two primary data-level approaches are commonly used to balance imbalanced datasets:

Undersampling the overrepresented class and acquiring more data for the minority class.

Oversampling the minority class using techniques such as SMOTE (Synthetic Minority Over-sampling Technique).

Although these methods may slightly reduce overall model accuracy, they generally improve model sensitivity, which is critical in detecting minority-class events such as disease occurrence.

3. RESULTS

3.1. Estimation of Conformation Scores

In the course of my research, I compared the effectiveness of two different machine learning methods for estimating cattle conformation scores: I evaluated the performance of a linear regression model and a Poisson regression model in predicting various phenotypic traits of cattle.

The primary objective was to determine which model yields better results and how these results compare with those of traditional mathematical approaches. To assess model performance, I used Mean Squared Error (MSE) and the Coefficient of Determination (R^2).

Based on the results, I aimed to determine whether an AI-based model could be developed to replicate expert conformation scoring.

The training dataset included conformation scores from 325 animals from a Limousin breeding herd, provided by the *Limousin and Blonde d'Aquitaine Breeders' Association* (Budapest). The animals were offspring of 18 bulls born between 1990 and 1996. Each animal was officially evaluated at 12 months of age at the end of its performance test.

The goal of the research was to design a system that could later be easily reused with data from other breeders. The structure of the prediction system is shown in *Figure 3.1*.

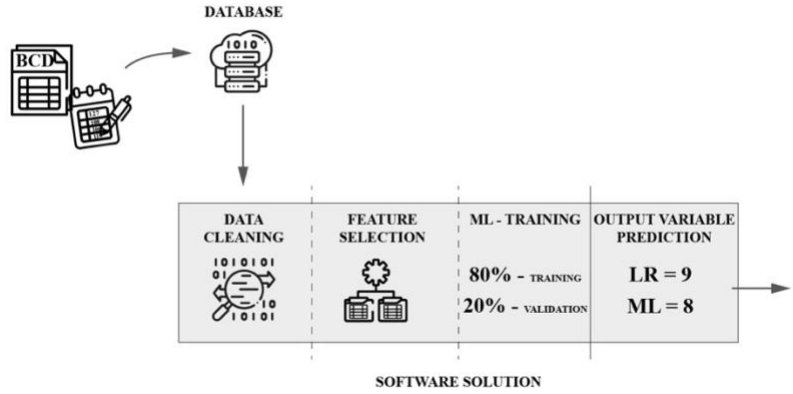

Figure 3.1.: Conceptual Framework of the Estimation Software's Operation

The historical evaluation data stored in the database (*Table 3.1*) were used to train various regression-based artificial intelligence algorithms. The entire system was implemented in Python. Data cleaning and preprocessing were also performed using Python.
For training the models, I developed custom code with the help of the *Scikit-learn* Python library. The dataset was divided into two parts: 90% of the data was used for training the algorithm, and the remaining 10% was reserved for testing and validation after training.

Table 3.1.: Output variables

Dependent variables	**Independent variables**
Muscularity (0–60 pont)	Score for utility value, Score of length, Score for with (0-40 score, or 0-60 score)
Length of the rump (1–9 pont)	Length of the body, Length of the back (1–9 pont)

Muscularity of breast (1–9 pont)	Muscularity of shoulder, Muscularity of back, Muscularity of the round of rump, Muscularity of the width of rump (1–9 pont)
Muscularity of the width of rump (1–9 pont)	Muscularity of shoulder, Muscularity of back, Muscularity of the round of rump, Muscularity of the width of rump (1–9 pont)

Each model was trained 10 times, with the dataset being randomly split into two parts in each iteration. The results were then compared, and for each model, the best-performing outcome was included in the results table.

The trained models were validated using the test dataset, and the results of this evaluation are presented in *Table 3.2.*

Table 3.2.: Results of the Two Model Types (n = 325)

Model	**Dependent variable**	**MSE**	**R^2**	**A-R^2**
Linear regression	Muscularity (0–60)	3.38	0.86	0.83
	Length of the rump (1–9)	0.21	0.93	0.91
	Muscularity of breast (1–9)	0.35	0.86	0.79
	Muscularity of the width of rump (1–9)	0.41	0.77	0.76
Poisson regression	Muscularity (0–60)	4.00	0.81	0.77
	Length of the rump (1–9)	0.23	0.92	0.85

	Muscularity of breast (1–9)	0.39	0.82	0.78
	Muscularity of the width of rump (1–9)	0.49	0.73	0.7

The results clearly show that the R^2 and adjusted R^2 values are very similar — particularly in the case of Score for Muscularity. This indicates that the choice of models was appropriate, and that both approaches are suitable for estimating the conformation scores.

However, for predicting rump length, the linear regression model slightly outperformed the Poisson regression in terms of R^2. Similarly, the estimation of muscularity of the chest yielded comparable results for both models. In contrast, the Poisson regression model produced less accurate predictions for rump width.

The mean squared errors were also more favourable in the linear regression model, which — based on the results — makes it the preferred algorithm for future practical applications. In all four traits analysed, the Poisson model resulted in higher prediction errors.

Although the results showed low variability, they clearly reflect biological and anatomical relationships. It is not surprising that the muscularity score alone was able to predict the values of the other three traits with high accuracy ($R^2 = 0.86$).

Overall, the findings demonstrate that machine learning-based models can reliably predict conformation traits in cattle. The

model's performance, with an R^2 of 86%, is considered very good in terms of practical applicability.

3.2. Estimation of Somatic Cell Count

The aim of this part of the research was to develop a machine learning-based model capable of accurately predicting somatic cell count (SCC) based on other available parameters. To achieve this, the first step involved conducting a statistical analysis of SCC values.

Somatic Cell Count (SCC) is the most commonly used indicator for detecting mastitis (udder inflammation) in dairy cattle.

The statistical analysis of the SCC variable, based on the values stored in the dataset, is presented in *Table 3.3*.

Table 3.3.: Statistical Characteristics of the Somatic Cell Count in the Dataset

No. of samples	26 6686
Min	2000
25%	50 000
50%	150 000
75%	401 000
Max	900 000

- If SCC < 100 000, the cow is likely healthy.
- If SCC > 100 000, but < 300 000, the cow requires special attention.
- If SCC > 300 000, the cow is likely infected.

Az I divided the SCC values into three categories for classification purposes:

0 = Not infected

1 = Possibly infected

2 = Infected

The distribution of these categories in the output variable is presented in *Table 3.4*.

Table 3.4.: Distribution of the Output Variable's Individual Values

Category	**Occurrence**
0	17 500
1	5 200
2	2 600

As expected, the dataset exhibits a highly imbalanced class distribution. This is due to the fact that most cows in production are healthy, and therefore only a small proportion of milk samples contain infected milk. To build a reliable model, it is essential to balance the dataset, as an unbalanced dataset would lead to a biased and inaccurate model.

To address this, I applied a hybrid up sampling technique, combined with increased variance in the input variables.

Next, I needed to select the most suitable algorithm for model development. Since the target variable consists of three categories, it was necessary to use multiclass classification algorithms capable of categorizing data into three or more classes. Based on the literature, I compared the performance of four different

algorithms: *Random Forest, Support Vector Machine (SVM), Decision Tree Regressor, Extra Trees Classifier*.

For model training, the dataset was randomly split into two parts: 90% was used for training, and 10% for testing and validation. The class distribution was preserved across both subsets to ensure representativeness.

I evaluated the model performance separately for each class and then calculated the average accuracy across all categories. Accuracy was defined as the ratio of correctly predicted outcomes to the total number of predictions.

If the model achieved over 80% accuracy, it was considered suitable for practical use. A model with over 90% accuracy could potentially replace laboratory testing in certain scenarios.

The best-performing model used the following input variables: *LNPC, urea, sugar, lactoferrin, fat,* and *protein*. The average accuracy for each model is presented in *Table 3.5*.

Table 3.5.: Comparison of Model Accuracies Achieved by Different Algorithms

ML-algorithm	**LSCC=0**	**LSCC=1**	**LSCC=2**	**Avg.**	**Recall**
Random Forest	0.88	0.86	0.85	0.86	0.75
SVM	0.86	0.85	0.80	0.84	0.72
Decision Tree Regressor	0.83	0.80	0.82	0.82	0.57
Extra Trees Classifier	0.89	0.88	0.86	0.88	0.72

As shown, all selected algorithms achieved accuracy above 80%. The best-performing model was built using the Extra Trees Classifier algorithm.

I chose a tree-based algorithm because such methods are suitable for categorical variables, are not sensitive to normal distribution assumptions, and therefore require less preprocessing of the input data.

3.3. Estimation of Casein Content

The objective of this part of the research was to demonstrate that the casein content in milk samples can be accurately predicted using a machine learning-based model that relies on other milk components as input variables.

The data used in this study were obtained from an accredited laboratory in Hungary and included monthly milk parameter values collected over a three-year period from three different farms.

Since the dependent variable (casein) was continuous and exhibited linear relationships with most input features, I opted to use a regression-based model.

The first step involved data cleaning, addressing missing values and outliers. Normalization of variables was not required, as variables with a large number of outliers were converted into categorical variables. The output variable (casein) remained continuous.

The software was developed in Python using the *Pandas* and *Scikit-learn* libraries. For model building, I applied the *Linear Regression* model from Scikit-learn. Additionally, I used the

widely accepted *statsmodels* module in Python to support the mathematical analysis.

The dataset was randomly split into 90% for training and 10% for testing. To evaluate model performance, I used the *coefficient of determination* (R^2) and the *Root Mean Squared Error of Prediction* (RMSEP).

Each input variable set was tested over 10 training–validation cycles, where the dataset was randomly partitioned for each run. *Table 3.6* presents the best and worst results from these iterations.

Table 3.6.: Results of Estimations Based on Different Input Variables

Input variables	**R^2**	**MRSE**
'Lact days', 'Urea', 'Sugar', 'LF', 'Fat'	0.82/0.78	0.033
' Protein', 'Oil', 'Sugar', 'LF', 'Fat'	0.86/0.84	0.018
'Lact days', 'SNF', 'LF', 'Fat'	0.83/0.76	0.034

The actual and predicted output values were also visualized on a single graph (*Figure 3.2*).

The red line represents the regression line derived from the actual output values, while the blue dots indicate the model's predicted values.

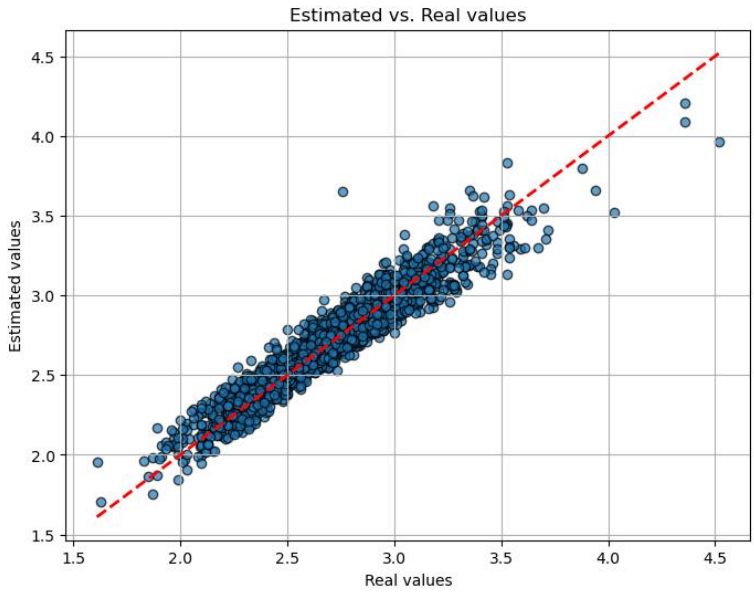

Figure 3.2.: Measured Target Values and Predicted Values Based on the Model

The dataset consisted of 25,000 samples, which were used to train the model for optimal performance. If more data — potentially from other laboratories — becomes available, the model can be further refined using the same system, thereby improving its generalizability and accuracy, since the entire processing pipeline is fully automated.

To identify the optimal model configuration, I also tested various combinations of input variables. Increasing the amount of training data can further enhance the model's precision.

The results show that using *"oil," "sugar," "lactoferrin," and "fat"* as input features yields a model with an R^2 of 0.86 and a low prediction error. This machine learning-based model outperforms traditional methods in estimating casein content.

When applying a traditional mathematical approach, the resulting R^2 was 0.8395, which is slightly lower. Therefore, the machine

learning model, with its 86% prediction accuracy, provides superior performance compared to the 83.95% of the mathematical model.

This level of performance is sufficient for practical applications, such as pricing at milk collection points, where exact casein content does not need to be precisely determined.

The system I developed fully automates the data cleaning, transformation, and prediction processes within a single framework. The model can be easily retrained or fine-tuned with new data — only the input table needs to follow the specified format.

4. CLUSIONS AND SUGGESTIONS

My research findings contribute to the advancement of data-driven approaches in milk quality monitoring, animal health, and breeding value assessment.

Machine learning models — whether used for conformation scoring, early diagnosis of mastitis, or prediction of casein content — are capable of producing objective and reproducible results, thereby reducing errors associated with human subjectivity.

My research confirms that machine learning-based models can deliver results that are equal to or even better than those produced by traditional mathematical methods. Furthermore, AI-based models are expected to perform even better when trained on large-scale datasets, thanks to faster training and the ability to be continuously refined as new data become available.

Thus, the study demonstrates that with the collection of additional data — or the development of truly large-scale datasets — it is feasible to create industrially applicable models and applications using the methods described in this dissertation.

Future research may benefit from incorporating additional, less-explored factors, such as genetic or environmental parameters, to provide an even more comprehensive understanding of animal health and performance.

5. NEW SCIENTIFIC RESULTS

Throughout my research, I uncovered numerous relationships that highlight the potential of utilizing existing cattle breeding databases.

Artificial intelligence, and particularly machine learning, has proven to be an excellent tool for training models that enable breeders and producers to estimate and forecast key parameters related to breeding, product quality, or animal health — even within the framework of a simple application.

5.1. Thesis 1

1. Prediction of Conformation Scores

In my research, I developed and validated a novel model that demonstrates how a machine learning-based regression approach can objectively and accurately reproduce traditional expert-based conformation scoring.

Expert evaluations can be subjective, influenced by individual experience, opinion, or variations in methodology. The results of my study confirm that a machine learning approach is capable of reducing this human variability, thereby providing objective and reproducible results—a significant advancement in evaluation methods used in animal breeding.

The results produced by the trained and validated model are presented in *Table 5.1*.

Table 5.1.: Accuracy of the Best-Performing Model for Scores of Different Body Traits (n = 325)

Model	Dependent variables	MRSE	R^2
Linear regression	Muscularity (0–60)	3,38	0,86
	Length of the rump (1–9)	0,21	0,93
	Muscularity of the breast (1–9)	0,35	0,86
	Muscularity of the width of the rump (1–9)	0,41	0,77
OLS model	Length of the rump	0,33	0,91

Description of the Database Used for Model Development:

Data source: Limousin and Blonde d'Aquitaine Breeders' Association (Budapest), anonymized dataset

Animals included: Offspring of 18 bulls, born between 1990 and 1996

Total number of records: 325 measurements

Description of the Best-Performing Model:

Python module used: *Scikit-learn*

Algorithm applied: *LinearRegression*()

test_size = 0.10: 10% of the data was used for testing

random_state = 10: Ensures reproducibility of the random split by setting the seed of the random number generator

Intercept: –0.079789

Coefficients (weights): [0.22136146; –0.56044067; –0.66474507]

The automated model provides a faster and easily scalable solution for performing conformation evaluations. This enables objective comparisons across larger cattle populations and offers more efficient decision support in the design and optimization of breeding programs.

5.2. Thesis 2

2. Prediction of Somatic Cell Count

In my research, I demonstrated that it is possible to develop a novel machine learning-based model capable of reliably determining whether a cow is healthy or infected based on milk samples — without the need for direct measurement of the somatic cell count.

Using data from 25,000 milk samples, I created a model that predicts SCC values based on other measurable milk parameters. The results confirm that alternative milk parameters can enhance the early detection of mastitis, while the use of machine learning reduces both the cost and time associated with conventional laboratory methods.

Among the algorithms tested, the model trained using the Extra Trees Classifier delivered the best results. These are presented in *Table 5.2*.

Table 5.2.: R^2 Results by Somatic Cell Count Classes and Average R^2 Value

ML-algorithm	**LSCC=0**	**LSCC=1**	**LSCC=2**	**Avg**
Extra Trees Classifier	0.89	0.88	0.86	0.88

I also examined the relative impact of each parameter used in the training process on the final model.
The results are presented in *Figure 5.1.*

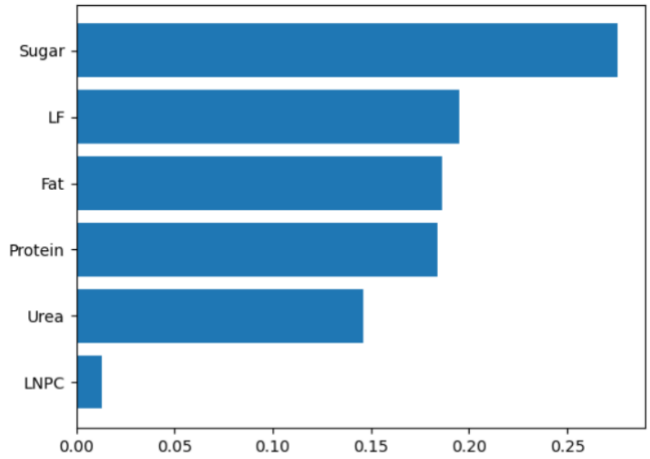

Figure 5.1.: Classification of Milk Samples Based on Somatic Cell Count and Importance of Key Input Variables in Model Development

The results clearly indicate that the information embedded in other milk parameters is sufficient for a machine learning model to reliably classify samples into “healthy” or “infected/mastitis” categories.

Description of the Database Used for Model Development:

Number of contributing breeders: 3

Years covered: 2019–2021

Sampling frequency: Monthly milk quality analysis (36 monthly records per breeder)

Total number of measurements: 25,000

Number of parameters measured: 14

Detailed Model Description

n_estimators = 100 (Defines the number of decision trees built during ensemble learning. The more trees used, the more robust the model becomes due to error averaging—at the cost of increased computational demand.)

test_size = 0.10: 10% of the dataset was used for testing

random_state = 10:

Feature importances: [0.01788029; 0.10898641; 0.13511609; 0.24754355; 0.1710435; 0.15843071]

Number of features used: 6

Target classes: [0, 1, 2]

Maximum depths of decision trees:: [48, 48, 41, 44, 46, 43, 47, 39, 51, 40, 43, 40, 41, 37, 43, 41, 43, 43, 40, 45, 42, 46, 46, 42, 42, 44, 42, 41, 41, 44, 47, 46, 45, 42, 37, 44, 44, 40, 41, 44, 39, 42, 42, 43, 44, 41, 39, 45, 46, 47, 49, 44, 40, 41, 36, 40, 46, 48, 38, 45, 43, 44, 46, 44, 37, 41, 47, 36, 43, 40, 42, 48, 41, 40, 45, 38, 41, 46, 46, 40, 44, 42, 38, 41, 40, 42, 39, 38, 44, 43, 44, 39, 43, 46, 47, 38, 40, 39, 45, 45]

5.3. Thesis 3

3. Estmiation of Casein Content

My research confirmed that a regression-based machine learning model can reliably estimate casein content using only routinely measured milk parameters.

There is no need for direct and costly casein quantification, as relevant information can be extracted from indirect measurement data.

The model not only enables accurate estimation of casein levels, but also helped identify which laboratory measurements are most influential in estimating casein content.

This feature importance analysis provides new insights into which characteristics are most critical from the perspective of milk quality.

Table 5.3.: Best Performing Input Variables for Casein Content Prediction

Input variables	**R^2**	**MRSE**
'Protein', 'Oil', 'Sugar', 'LF', 'Fat'	0.86/0.84	0.018

Accurately estimating casein content is crucial in cheese production and in quality-based milk pricing.

An automated, data-driven approach not only enhances the speed and objectivity of quality control but also contributes to the cost-efficiency of industrial monitoring processes.

These results are particularly relevant in the dairy industry, where product quality is a key factor influencing market value and consumer satisfaction.

Description of the Database Used for Model Development:

Number of contributing breeders: 3

Years covered: 2019–2021

Sampling frequency: Monthly milk quality analysis (36 monthly records per breeder)

Total number of measurements: 25,000

Number of parameters measured: 14

Description of the Best-Performing Model:

Python module used: *Scikit-learn*

Algorithm applied: *LinearRegression*()

test_size = 0.10: 10% of the dataset was used for testing

random_state = 10: Ensures reproducibility of the random split by setting a fixed seed value

Intercept: 0.38

Coefficients (weights): [6.71812578e-01 -3.15429010e-01 -5.58654963e-03 1.19884915e-03 1.60130038e-02]

5.4. Thesis 4

4. Class Imbalance Adjustment in Real Milk Quality Data

The datasets used for model development were highly imbalanced.

During my research, I demonstrated that hybrid data balancing techniques can effectively transform such biased datasets into forms suitable for training machine learning algorithms.

The hybrid approach applied in this study — which combines the inclusion of new relevant data, synthetic data injection (data augmentation), and fine-tuning of input variables — offers a novel solution for handling imbalanced milk quality datasets from an animal health perspective.

Using this method, I was able to reduce the dominance of the overly represented “healthy” class and generate a more balanced and representative dataset.

My research highlighted how data imbalance (particularly the low ratio of infected samples) can distort the learning process of

predictive models and provided concrete techniques for addressing this issue.

6. IMPORTANT PUBLICATIONS RELATED TO THE THESIS

1. Revoly A.; Tarr B.; Szabó I. (2023): A mesterséges intelligencia szerepe az élelmiszerbiztonságban. In: *Biztonságtudományi Szemle,* 5 (3) 141–150. p.

2. Szabó I.; Tarr B.; Revoly A. (2023): Adattechnológiák implementációja mezőgazdasági műszaki folyamatokban. Előadás, MTA Tudományos ülés, Budapest, 2023. 11. 06.

3. Tarr B. et al. (2022): Precíziós eljárások és a mesterséges intelligencia technológia alkalmazása a szarvasmarha-tenyésztésben különös tekintettel a húshasznú szarvasmarhák azonosítására. In: *Animal welfare Etológia és Tartástechnológia (AWETH),* 18 (1) 51–63. p.

4. Tarr B. et al. (2023): Predicting somatic cell count in milk samples using machine learning. In: *12th International Conference on Applied Informatics (ICAI 2023): Abstracts.* 1–2. p.

5. Tarr B.; Szabó I.; Tőzsér J. (2023): Estimation of important milk constituents from milk samples using artificial intelligence. In: *5th International Conference on Biosystems and Food Engineering: Proceedings.*

6. Tarr B.; Szabó I.; Tőzsér J. (2023): Machine learning in cattle breeding. In: *Mechanical Engineering Letters,* 24, 71–84. p.

7. Tarr B.; Szabó I.; Tőzsér J. (2023): Mesterséges intelligencia alkalmazása a szarvasmarha-tenyésztés egyes területein: egyedazonosítás, állategészségügy és állatjóllét: Irodalmi összefoglaló. In: *Magyar Állatorvosok Lapja,* 145 (11) 651–660. p.

8. Tarr B.; Szabó I.; Tőzsér J. (2024): Predicting somatic cell count in milk samples using machine learning. In: *Annales Mathematicae et Informaticae,* 60, 159–168. p.

9. Tarr B.; Szabó I.; Tőzsér J. (2025): Body Conformation Scoring of Cattle, using Machine Learning. In: *Acta Polytechnica Hungarica,* 22 (3) 27–38. p.

10. Tarr B.; Tőzsér J.; Szabó I. (2021): Precíziós gazdálkodás módszerének és a mesterséges intelligencia (MI) eljárásának alkalmazási lehetőségei a szarvasmarha-tenyésztésben. In: *Mezőgazdasági Technika,* 63 (7) 2–5. p.

11. Tőzsér J. et al. (2023): Evaluation of body measurements of Limousin heifers by backward regression analysis in western Hungary. In: *Animal welfare Etológia és Tartástechnológia (AWETH),* 19 (1) 102–117. p.