Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 701–706 DOI: 10.15439/2024F9318 ISSN 2300-5963 ACSIS, Vol. 39

Stacking Ensemble Machine Learning Modelling for Milk Yield Prediction Based on Biological Characteristics and Feeding Strategies

Ruiming Xing[a], Baihua Li[a*], Shirin Dora[a], Michael Whittaker[b], and Janette Mathie[b]

[a]Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK
[b]Cattle Information Service, Speir House, Stafford Park 1, Telford, Shropshire, WD3 3BD, UK
[*]Corresponding: b.li@lboro.ac.uk; {r.xing, s.dora}@lboro.ac.uk;{michaelwhittaker, janettemathie}@thecis.co.uk

***Abstract*—Knowing expected milk yield can help dairy farmers in better decision-making and management. The objective of this study was to build and compare predictive models to forecast daily milk yield over a long duration. A machine-learning pipeline was provided and five baseline models as well as a novel stacking model were developed for the prediction of milk yield on the CowNflow dataset using 414 Holstein cattle records collected from 1983 to 2019. Four different feature selection methods were performed to evaluate the essential biological characteristics and feeding-related features which affect milk yield. The results showed that the overall performance of predictive models improved after proper feature selection, with an R^2 value increased to 0.811, and a root mean squared error (RMSE) decreased to 3.627. The stacking model achieved the best performance with an R^2 value of 0.85, a mean absolute error (MAE) of 2.537 and an RMSE of 3.236. This research provides benchmark information for the prediction of milk yield on the CowNflow dataset and identifies useful factors such as dry matter (DM) intake and lactation month in long-term milk yield prediction.**

***Index Terms*—Dairy Cattle, Milk Yield, Machine Learning, Feature Selection**

I. Introduction

FORECASTING milk yield is a matter of great concern for dairy community. It has been shown that global milk demand is expected to grow by 22% between 2018 and 2027 [1]. It is important for dairy farmers to understand the essential factors influencing milk production so that they can deploy optimal management strategies, increase milk yield and reduce their production costs [2]. In this regard, predictive models for milk yield can help them develop better culling strategies and retain high-yielding cows [3].

In the past, livestock management relied more on the collective knowledge of people and their experience to make effective decisions. With the development of technology, dairy farmers are finding more effective management strategies such as using intelligent management systems, and sensors to record the characteristics of their herd and improve the efficiency of dairy production [4]. This has led to an increase in the availability of farm-related data, enabling data-driven management of farming through techniques like Machine Learning (ML).

Several studies have shown their interest in the field of milk prediction. Linear Regression (LR) [5], Multiple Linear Regression (MLR) [6], Random Forest (RF) [7], [8], Support Vector Machine (SVM) [9] and Artificial Neural Network (ANN) [10] have been widely used in the prediction of milk yield. Sharma et al. [11] compared a multiple linear regression and ANN for milk yield prediction in Karan Fries dairy cattle and proved the performance of the ANN model is slightly superior to the regression model. while similar studies carried out in Sahiwal cattle [12] and Karan Fries cattle [13] also showed that ANNs gain good performance. Apart from basic factors such as cow age and lactation, Body weight at calving and the days in milk on the test day are regarded as the variables that are important for ANNs [11]. Other attempts to predict milk production involve a Back Propagation Neural Network (BPNN) optimised using Genetic Algorithm (GA) to analyse the impact of physiological and environmental, which was proposed by Sugiono et al. [14]. It is seen that predictive ML models have been deployed to deal with different scenarios. However, there is not much work being done to find the best-performing model from a machine-learning perspective to compare their performance in the same scenario.

Existing research on milk production has focused on accurately predicting milk yield over short periods. In [12], [13], neural networks are used to predict the milk yield from the first lactation 305-day. In [7], ANNs are deployed to predict milk yield for the first test day of the first lactation period. In addition, the XGBoost algorithm is applied for forecasting the next month's milk yield [15]. The above-mentioned works have not been evaluated for generating long-term predictions. A major reason for the lack of studies on long-term predictions is a lack of a suitable dataset.

Stacking is one of the most popular ensemble ML methods for predicting multiple nodes to build new models and improve model performance. It allows us to train multiple models to solve similar problems and build a new model with better performance based on their combined output [16]. In this study, a novel stacking method was proposed to accurately predict milk yield over a longer duration. Baseline models like LR, SVM, RF regression, AdaBoost and ANN were built and their performance was evaluated and compared. The

TABLE I
ATTRIBUTES RELATED TO COWS AND FEEDING CHARACTERISTICS

Index	Attributes	Num	Data Description	Data Type
1	Cow age (month)	414	Age in month	numerical
2	Body weight (kg)	414	Body weight	numerical
3	Physiological status	414	Two categories: dry, lactating	categorical
4	Lactation month	414	Number of months of lactation	numerical
5	Gestation month	414	Number of months of gestation	numerical
6	Diet type	414	Six categories, about feeding diet type	categorical
7	DM intake (kg/day)	414	Dry matter intake	numerical
8	DM digestibility (g/g)	414	Dry matter digestibility	numerical
9	DMI/100 kg body weight	414	Dry matter intake per 100 kg body weight	numerical
10	OM intake (kg/day)	413	Organic matter intake	numerical
11	Ash intake (kg/day)	413	Ash intake	numerical
12	N intake (g/day)	414	Nitrogen intake	numerical
13	CP intake (g/day)	414	Crude protein intake	numerical
14	Milk production (kg/day)	402	Milk production	numerical

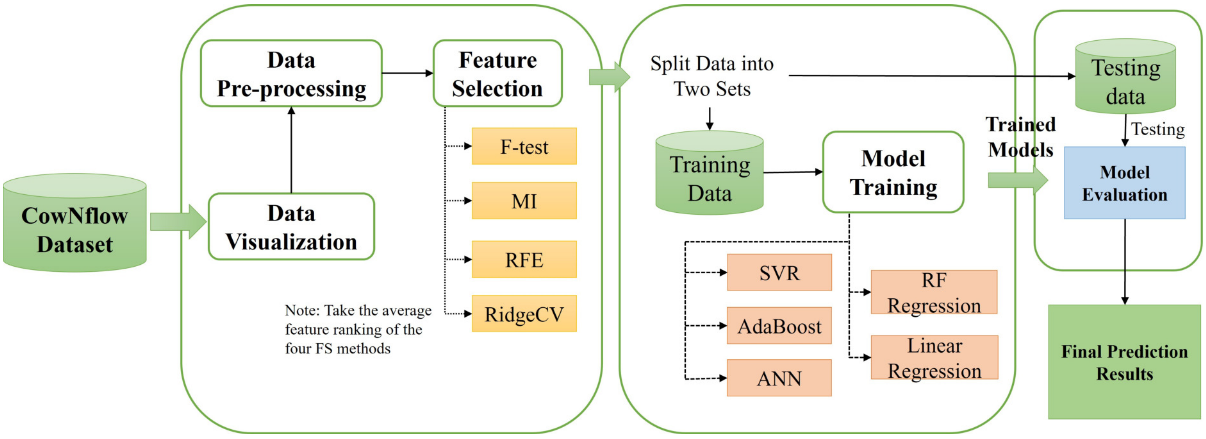

Fig. 1. ML pipeline for milk yield prediction

three main contributions in our study include: (1) Identify useful biological characteristics and feeding-related factors for long-term milk yield prediction. (2) Develop a stacking-based model for long-term milk yield prediction. (3) First build and compare the performance of various ML models on this dataset.

II. Data Description

The data utilised in this study is called the CowNflow dataset [17] which is published by the National Institute for Agriculture, Food and the Environment (INRAE)[1] in France. The dataset has been collected at the experimental dairy farms of INRAE. It reports individual biological measurements from dairy cattle like dry matter (DM) intake, milk yield and feeding attributes like crude protein concentration of each feeding and diet composition. Cows were fed in individual troughs, had free access to water, and were milked twice a day. The dataset contains attributes like cow age, body weight, milk yield, lactation number, feeding types and consumption of diet components. Table I shows the biological characteristics and feeding-related features in the dataset that are considered in this study.

III. Exploratory Data Analysis

Figure 1 shows the pipeline for data analysis utilised in this paper. All the analysis reported in this paper has been carried out using the Python library Scikit-learn (version 1.1.3). Seaborn (version 3.10.6) is used for generating the visualisation.

Based on the dataset, data visualisation and pre-processing will be performed. After understanding the data distribution and cleaning the data, feature selection will be carried out. In feature selection, four different measures are taken into account. The ranking of the importance of features for milk prediction is obtained by averaging the ranking of features in each method. After that, the well-processed data is divided into training set and testing set. The training set is used to train the model and after getting the trained model and the

[1] https://entrepot.recherche.data.gouv.fr/dataverse/inrae

testing set is used for model testing. Finally, the performance of each model will be evaluated and compared.

A. Data Visualisation

The purpose of visualisation is to develop an understanding of the underlying distribution for different features and identify patterns, and trends in the dataset.

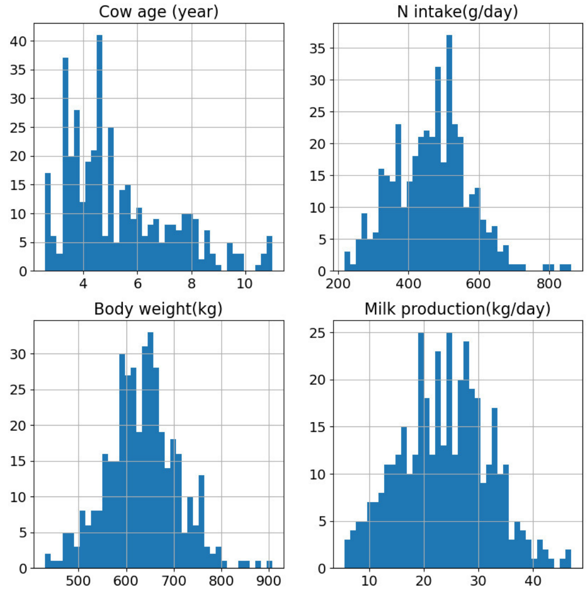

Fig. 2. Histograms of part of cattle features (The vertical axis is the amount of records)

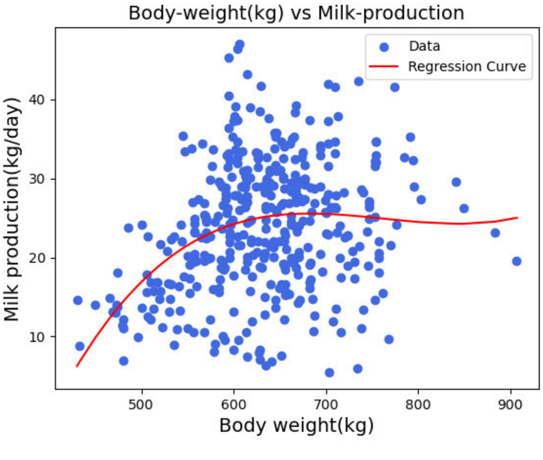

Fig. 3. Distribution of body weight vs. milk production

Figure 2 shows histograms for part of features such as cow age, body weight, and dry matter intake in the dataset. It can be seen that most of the cows in the experiment are between 2 and 10 years old. Histograms of body weight, nitrogen intake and milk production indicate that these features closely follow a Gaussian distribution, which may simplify the modelling process, and reduce the computational resources required for modelling.

The swarm chart (also named scatter plot) can be used to visualise the distribution of the joint distribution of a couple of discrete attributes. Figure 3 illustrates the impact of body weight on milk yield. It indicates that milk production increases as cow weight increases for cows weighing less than 600kg. For cows weighing more than 600kg, the milk yield doesn't exhibit a lot of variation.

B. Pre-processing

In this section, several data-cleaning steps are performed such as cleaning missing data and dealing with outliers to ensure the quality of data for further processes.

The records in which 'physiological status' has the value of 'dry' are not used in this study as these don't contribute towards predicting the milk yield. This resulted in 403 records that are used for further analysis in this paper.

Missing data: Dealing with missing data is important, as it may produce incorrect or biased results if missing data is not addressed properly. There are 3 missing values in each of the features *OM intake*, *Ash intake*, and *milk production*. These values are replaced by the mean values of the respective features.

Outliers: Many learning algorithms are sensitive to the range and distribution of attribute values. The interquartile range (IQR) is a commonly used tool to detect outliers with numerical values. To calculate the IQR, the dataset is divided into rank-ordered even quartiles, denoted by Q1 (lower 25%), Q2 (median 50%) and Q3 (upper 75% quartile), so IQR is the median 50% (Q3 − Q1). The whiskers have an offset length of 1.5*IQR, any data located outside of the whiskers is considered an outlier.

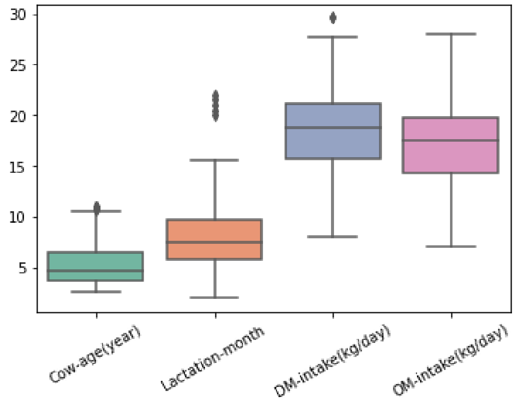

Fig. 4. Outlier detection using IQR box-and-whisker

Figure 4 shows the example box plots for different features in the dataset. The values outside of the whiskers are considered outliers. For example, For feature *DM intake*, the figure illustrates that values over 27 are regarded as outliers. In our task, to prevent loss of data available for training, only those records in which two or more features are identified as outliers are removed.

TABLE II
FEATURE RANKING WITH DIFFERENT FEATURE SELECTION METHODS

Features	F-test	MI	RFE	RidgeCV	Voting Rank
OM intake (kg/day)	1	1	2	2	1.50
DM intake (kg/day)	2	2	4	3	2.75
Lactation month	3	3	1	1	2.00
Diet type	4	7	11	4	6.50
DMI/100 kg body weight	5	8	5	8	6.50
CP intake (g/day)	6	5	9	10	7.50
N intake (g/day)	7	6	8	9	7.50
Gestation month	8	9	6	7	7.50
DM digestibility (g/g)	9	11	3	5	7.00
Body weight (kg)	10	10	12	6	9.50
Cow age (year)	11	4	7	11	8.25
Ash intake(kg/day)	12	12	10	12	11.50

Ranking score:1-12, 1 means most related and 12 represents the least.

C. Feature selection

Since irrelevant, redundant variables can reduce the model's generalisation capability and accuracy, feature selection is an effective step to find the most informative feature set that can have a better impact on the model performance [18]. Before selection. The 'diet type' feature was converted to a numerical feature using CatBoost Encoder [19].

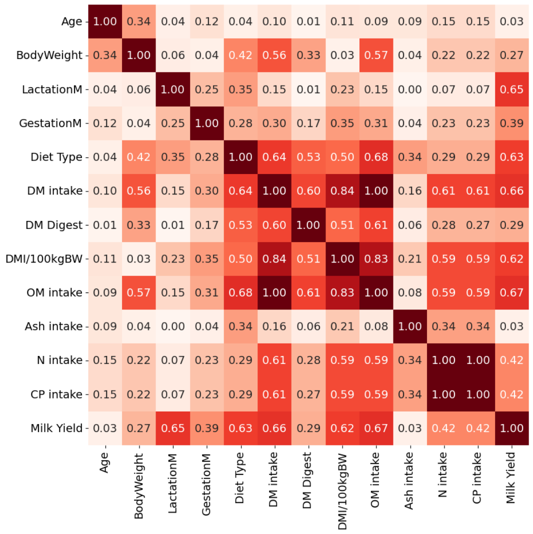

Fig. 5. Heatmap on feature correlation in Pearson coefficients

A correlation heatmap shows the correlation coefficients between a set of variables. It can be especially useful to identify which variables are most strongly correlated with each other and define potential confounding factors. We convert the numbers to absolute values since correlation is shown by numerical magnitude, whereas positive or negative values simply indicate a positive or negative correlation. Then the correlation matrix with Pearson coefficients is shown in Fig. 5.

Four different measurements are applied to rank the features, which are linear regression f-test, Mutual information test, Recursive feature elimination (RFE) and Ridge regression with built-in cross-validation (RidgeCV). **F-test** is a statistical test which provides an f-score by calculating the ratio of variances. The variance of a feature determines how much it impacts the milk yield. If the variance is low, it implies this feature has less importance in predicting milk yield and vice-versa. **Mutual information** (MI) evaluates the gain of each variable in the context of the target variable, and it is predicated on joint probability. It indicates that the higher the mutual information value, the closer the connection between this feature and the target.

In addition to the two filter feature selected methods mentioned previously, a wrapper method **RFE** and an embedded method **RidgeCV** are also adapted to evaluate the correlation. RFE selects features by recursively considering smaller and smaller sets of features, the SVM algorithm and linear kernel were chosen to perform. RidgeCV is normally used in datasets which have multicollinearity. It uses L2 regularisation but performs Leave-One-Out Cross-Validation.

Table II shows the ranks of all the features based on the different methods used for feature selection. It can be observed that Ash-intake has the lowest average rank. Further, there is a difference of 3.25 between the rank of ash intake and the second lowest rank which indicates that ash intake was consistently ranked lower by all the feature selection methods. It can also be seen from Table II that CP intake and N intake have the same average rank. Further, it can be observed from Figure 5 that CP intake and N intake are strongly correlated with a Pearson coefficient of 1. Similarly, OM intake and DM intake are highly correlated with a Pearson coefficient of 1. Based on these observations, the features of ash intake, CP intake and OM intake are not selected for further analysis.

TABLE III
COMPARISON OF LEARNING ALGORITHMS ON THE ORIGINAL AND SELECTED FEATURES

Model	No Feature Selection			With Feature Selection		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Linear Regression	0.776	3.031	3.952	0.776	3.020	3.947
SVR	0.813	2.733	3.613	0.812	2.772	3.619
RF regression	0.805	2.920	3.684	0.804	2.902	3.695
AdaBoost	0.813	2.860	3.601	0.820	2.838	3.542
ANN	0.827	2.670	3.471	**0.841**	2.602	3.330
Stacking	**0.843**	**2.568**	**3.308**	**0.850**	**2.537**	**3.236**
AVERAGE	0.813	2.797	3.605	0.817	2.779	3.562

D. Model Training

After pre-processing and feature extraction, 397 records with 9 features are retained for developing ML models. Records are standardised before performed. The performance of all models is evaluated using hold-out [20] validation framework with 75% and 25% data used for training and testing, respectively.

Five different supervised ML techniques are considered to develop models for predicting milk yield, namely Linear regression (LR), SVM, Random Forest (RF) regression, Adaptive Boosting (AdaBoost) and Artificial Neural Network (ANN). All models are fine-tuned and evaluated to decide the best model. The linear regression algorithm is a basic and relatively common method for generating predictions, which is well understood and can be trained very quickly. For SVM, the kernel is set to 'RBF'. RF regression ensembles multiple decision trees into its final decision. Different numbers of trees are tested to determine that 100 estimators for RF achieved the best performance. AdaBoost is also an ensemble learning algorithm, it aggregates a set of weak classifiers into a strong classifier. We finally set the number of estimators to 100 and the learning rate to 0.5. For the ANN model, an input layer, a dense layer with 100 ReLU-activated neurons, and an output layer with Adam as its optimiser make up the neural network. After feature selection, a stacking method is proposed after building the five baseline models. The weak learners are made up of three best-performed models and the meta learner is set to Ridge regression. The structure of this model is shown in Fig. 6.

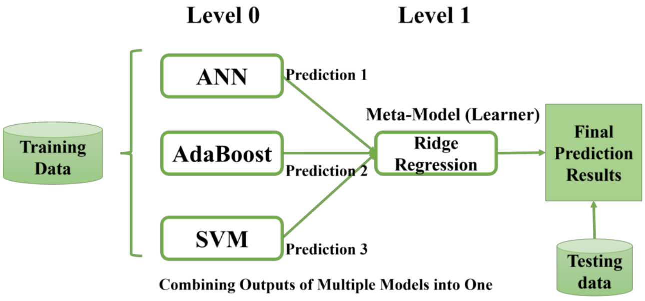

Fig. 6. The structure of stacking model

IV. Experimental Results and Evaluation

In this section, the performance of different models developed in this study is evaluated and compared. The metrics used to measure the performance of the models include R^2, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The R^2 is less than 1 where values close to 1 indicate that the model captures nearly all of the variation in the outcome of the target. MAE calculates the difference between predictive value and actual value for each data sample and takes the average absolute value of all samples. RMSE is similar to MAE but represents the square root of the average of squared errors in the predictions. The mathematical formulas for the different error metrics used in this study are given below:

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

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum |y_i - \hat{y}_i|^2}$$

where y_i represents the actual values, $\hat{y}_i$ is the predicted values $\bar{y}_i$ is the mean of the actual data and n represents the total number of samples. A model that has smaller values of MAE and RMSE represents better performance.

Table III presents the performance of five baseline models on the dataset obtained before and after the feature selection. The table shows that when we remove the ash intake as well as the CP intake and OM intake, the overall performance of the trained models improves, with the average R^2 increasing from 0.807 to 0.811 and the values of both error measurements decreasing. It indicates that feature selection helps improve the performance of most models except SVM and RF regression. The R^2 of the best-performed model ANN increased from 0.827 to 0.840, and RMSE decreased from 0.441 to 0.426. The highest R^2 value obtained by the artificial neural network model was 0.827.

The three best-performed models are used to develop a model using the stacking technique. The performance of the

Stacking model is compared with the ANN model shown in Table III. Compared with the baseline models, the R^2 value of the stacking model improved to 0.85 and values of MAE and RMSE reduced to 2.537, and 3.236 respectively.

V. Conclusion

Milk production has received much attention in dairy farming. In this experiment, an ML pipeline is developed and applied to the CowNflow dataset for predicting milk yield. Four different feature selection methods were performed. 9 features of the original 13 were selected after data preprocessing and feature selection. Five different ML algorithms and a stacking method were utilized. Among the five baseline models, ANN achieved the best performance with a top R^2 value of 0.827 and the lowest RMSE of 3.471 before feature selection. After feature selection. the average values of R^2 for 5 models increased from 0.807 to 0.811, with both error measure matrices reduced. The stacking model had the best performance with an R^2 value of 0.85 and an RMSE value of 3.236.

According to the result, it is indicated that the *Ash intake* doesn't contribute much to the milk yield in long-term prediction. For the feeding factors, CP and OM intake are highly correlated to N and DM intake, respectively, which can be dismissed. The ML pipeline proposed in this study is shown to be efficient and generate good results. In future work, it can be optimised for further analysis and the current results will be a useful benchmark for further model comparison on this dataset.

VI. Acknowledgement

The project is sponsored by Cattle Information Service (CIS) and National Bovine Data Centre (NBDC) in the UK. We also would like to thank UK Research and Innovation (UKRI) and the Engineering and Physical Sciences Research Council (EPSRC) for the support of this work (grant EP/Y00597X/1).

References

[1] OECD, Food, and A. O. of the United Nations, *OECD-FAO Agricultural Outlook 2018-2027*. OECD, 2018.

[2] M. Cockburn, "Review: Application and prospective discussion of machine learning for the management of dairy farms," *Animals*, vol. 10, no. 9, 2020. doi: 10.3390/ani10091690

[3] M. Lopez-Suarez, E. Armengol, S. Calsamiglia, and L. Castillejos, "Using decision trees to extract patterns for dairy culling management," in *Artificial Intelligence Applications and Innovations*. Springer International Publishing, 2018. doi: 10.1007/978-3-319-92007-8_20

[4] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming–a review," *Agricultural systems*, 2017. doi: 10.1016/j.agsy.2017.01.023

[5] A. Saha and S. Bhattacharyya, "Artificial insemination for milk production in india: A statistical insight," *Indian Journal of Animal Sciences*, vol. 90, no. 8, 2020. doi: 10.56093/ijans.v90i8.109314

[6] F. Zhang, J. Upton, L. Shalloo, P. Shine, and M. D. Murphy, "Effect of introducing weather parameters on the accuracy of milk production forecast models," *Information Processing in Agriculture*, vol. 7, no. 1, pp. 120–138, 2020. doi: 10.1016/j.inpa.2019.04.004

[7] G. M. Dallago, D. M. de Figueiredo, P. C. de Resende Andrade, R. A. dos Santos, R. Lacroix, D. E. Santschi, and D. M. Lefebvre, "Predicting first test day milk yield of dairy heifers," *Computers and Electronics in Agriculture*, vol. 166, p. 105032, 2019. doi: 10.1016/j.compag.2019.105032

[8] M. Salamone, I. Adriaens, A. Vervaet, G. Opsomer, H. Atashi, V. Fievez, B. Aernouts, and M. Hostens, "Prediction of first test day milk yield using historical records in dairy cows," *animal*, vol. 16, no. 11, p. 100658, 2022. doi: 10.1016/j.animal.2022.100658

[9] Q. T. Nguyen, R. Fouchereau, E. Frenod, C. Gerard, and V. Sincholle, "Comparison of forecast models of production of dairy cows combining animal and diet parameters," *Computers and Electronics in Agriculture*, vol. 170, p. 105258, 2020. doi: 10.1016/j.compag.2020.105258

[10] H. Radwan, H. El Qaliouby, and E. A. Elfadl, "Classification and prediction of milk yield level for holstein friesian cattle using parametric and non-parametric statistical classification models," *Journal of Advanced Veterinary and Animal Research*, vol. 7, no. 3, 2020. doi: 10.5455/javar.2020.g438

[11] A. K. Sharma, R. Sharma, and H. Kasana, "Prediction of first lactation 305-day milk yield in karan fries dairy cattle using ann modeling," *Applied Soft Computing*, vol. 7, no. 3, 2007. doi: 10.1016/j.asoc.2006.07.002

[12] V. Dongre, R. Gandhi, A. Singh, and A. Ruhil, "Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in sahiwal cattle," *Livestock Science*, vol. 147, no. 1-3, 2012. doi: 10.1016/j.livsci.2012.04.002

[13] D. Njubi, J. Wakhungu, and M. Badamana, "Use of test-day records to predict first lactation 305-day milk yield using artificial neural network in kenyan holstein–friesian dairy cows," *Tropical animal health and production*, vol. 42, 2010. doi: 10.1007/s11250-009-9468-7

[14] S. Sugiono, R. Soenoko, and D. P. Andriani, "Analysis the relationship of physiological, environmental, and cow milk productivity using ai," in *2016 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2016. doi: 10.1109/ICODSE.2016.7936165 pp. 1–6.

[15] B. Ji, T. Banhazi, C. J. Phillips, C. Wang, and B. Li, "A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm," *biosystems engineering*, vol. 216, pp. 186–197, 2022. doi: 10.1016/j.biosystemseng.2022.02.013

[16] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, 2004. doi: 10.1023/B:MACH.0000015881.36452.6e

[17] M. Ferreira, R. Delagarde, and N. Edouard, "Cownflow: A dataset on nitrogen flows and balances in dairy cows fed maize forage or herbage-based diets," *Data in Brief*, 2021. doi: 10.1016/j.dib.2021.107393

[18] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, 2017. doi: 10.1145/3136625

[19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018. doi: 10.48550/arXiv.1706.09516

[20] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018. doi: 10.48550/arXiv.1811.12808