Journal of Dairy Research

cambridge.org/dar

A supervised machine learning approach for the decision-making process on data-based culling in dairy farms

Oscar R. Espinoza Sandoval[1], Juan C. Angeles-Hernandez[2], Agustín Corral-Luna[1], Felipe A. Rodríguez-Almeida[1], Pablo Pinedo[3], Albert De Vries[4], Santiago A. Utsumi[5] and Einar Vargas-Bello-Pérez[1]

[1]Facultad de Zootecnia y Ecología, Universidad Autónoma de Chihuahua, Chihuahua, México; [2]Departamento de Medicina y Zootecnia de Rumiantes, Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Ciudad de México, México; [3]Department of Animal Sciences, Colorado State University, Fort Collins, CO, USA; [4]Department of Animal Sciences, University of Florida, Gainesville, FL, USA and [5]Department of Animal and Range Sciences, New Mexico State University, Las Cruces, NM, USA

Research Article

Cite this article: Espinoza Sandoval OR, Angeles-Hernandez JC, Corral-Luna A, Rodríguez-Almeida FA, Pinedo P, De Vries A, Utsumi SA and Vargas-Bello-Pérez E (2025) A supervised machine learning approach for the decision-making process on data-based culling in dairy farms. *Journal of Dairy Research*, 1–10. https://doi.org/10.1017/S0022029925101416

Received: 5 December 2024
Revised: 26 June 2025
Accepted: 8 July 2025

Keywords:
artificial neural network; culling; dairy cows; machine learning

Corresponding author:
Einar Vargas–Bello–Pérez;
Email: einar.vargasbelloperez@reading.ac.uk

Abstract

This research paper aimed to develop a supervised machine learning (ML) approach that learns and predicts data-based culling from farm information that reflects the criteria of the decisions taken to cull a cow by a farm manager. Data containing the features of milk yield, days in milk, lactation number, pregnancy status, days open and days pregnant were obtained from January to December 2020 from dairy cows on a large dairy farm in northern Mexico. The cows were labelled as those that were data-based culled (*Cull*) and those that were not culled (*Stay*). Six supervised ML algorithms were evaluated in a binary classification including logistic regression (LR), Gaussian naïve Bayes (GNB), k-nearest neighbors (k-NN), support vector machine (SVM), random forest (RF) and multilayer perceptron (MLP). Each model was subjected to hyperparameter optimization using a grid search approach combined with tenfold stratified cross-validation. This ensured that the class imbalance (*Cull* vs. *Stay*) was accounted during model evaluation. The best-performing model for each algorithm was selected on cross-validated accuracy. To evaluate the prediction performance of the ML algorithms on both labels from learned data, the metrics accuracy, precision, recall, F_1-score and the Matthews correlation coefficient (MCC) were employed. Accuracy among all classifiers was >0.90. The poorest prediction performance was observed in GNB (MCC = 0.50) and LR (MCC = 0.72). Conversely, the rest of the classifiers achieved superior prediction performance in learning the specific culling criteria, reaching an MCC score >0.91. Overall, culling criteria can be learned and predicted by ML algorithms and their performance varies among classifiers. This study identified RF as the best performing algorithm, but k-NN, SVM and MLP are possible candidates to be used in on-farm conditions. To increase their reliability, these approaches need to be tested in several farms, under different scenarios and varieties of features.

Culling is a dynamic process in dairy farming, with many factors involved in decision-making, essential to maintain the efficiency and productivity of the herd (Fetrow *et al.*, 2006; Cockram, 2021). At some point, all cows must leave the herd, which results in productive and economic consequences that will impact the farm's profitability (Compton *et al.*, 2017). On most dairy farms, after milk sales, cattle sales represent the second most important income (Calsamiglia *et al.*, 2018).

Two main reasons drive the decision-making in data-based culling. First, economic reasons involving cows that do not justify their place in the dairy farm, mainly due to low milk yield that does not cover the feeding cost, or due to poor reproductive performance (Marshall *et al.*, 2023). Second, replacement pressure, as in some cases, the number of replacement heifers can exceed those required, forcing the culling of the less valuable mature cows (Ferrari *et al.*, 2024). This second reason has gained importance in the past decades due to the adoption at a mass-scale of fertility programmes (Carvalho *et al.*, 2018) that have led to the 'high fertility cycle' as described by Fricke *et al.* (2023). That explains the relationship between high fertility, Body Condition Score and health events during the peripartum. Moreover, the continuous growth in the use of sexed semen has also contributed to this excess of replacement heifers (Berry, 2021). However, to increase the lifespan of dairy cows, it is important to note that each farm will have to find its specific situations for optimal replacement decisions (De Vries and Marcondes, 2020).

Under these scenarios, there is not a universal approach for culling that could be applied to all dairy farms, due to the specific productive and economic characteristics of each system

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

(Jayasundara *et al.*, 2019). The current approaches for culling can be divided into three main categories: (i) the most basic one, based on the intuition of the farmer or the manager, which can be data-based or not (Lehenbauer and Oltjen, 1998); (ii) following the suggestions of herd management software, has according to a programmed module with pre-established parameters; and (iii) based on models that estimate an economic value for each cow considering their productive data (De Vries, 2004; Groenendaal *et al.*, 2004; Sorge *et al.*, 2007; Nielsen *et al.*, 2010; Cabrera, 2012).

Although these approaches are widely used, they are still laborious and complex. In this sense, machine learning (ML) is a subset of artificial intelligence (AI), which learns from data and improves its performance on tasks without being explicitly programmed. Instead of following a fixed set of rules, ML models analyse patterns in data and make predictions or decisions based on experience, presenting a novel alternative to classical approaches (Espinoza-Sandoval *et al.*, 2024).

Some of the advantages of ML include the ability to handle large and complex data with non-linear relationships and without assuming specific data distributions (Brooks, 2024). In dairy science, ML has been successfully used in a wide range of fields, including disease detection, prediction of milk yield and quality and prediction of methane emissions (Slob *et al.*, 2021).

Regarding culling, some work has been done using ML. For example, Lopez-Suarez *et al.* (2018) used decision trees (DT) to predict the milk yield in the first lactation, identifying poorly productive cows in the early stages with an accuracy of 83%. Another approach was presented by Ouseleye *et al.* (2023) who used random forest (RF) to predict the longevity of dairy cows using productive and genetic features with a coefficient of determination of 86%. Adamczyk *et al.* (2016) tested several ML algorithms to identify the culling reasons but with no favourable results. Nonetheless, no scientific report has tried to teach specific criteria of culling to an AI.

Thus, this research paper aimed to develop a supervised ML approach that learns and predicts data-based culling from farm data that reflects the criteria of the decisions taken by a farm manager.

Materials and methods

Ethics statement

The present study only used data obtained from pre-existing databases; thus, approval by an institutional animal care and use committee was not necessary.

Data source

The data were obtained from a commercial dairy farm located in northern México, which is a Holstein-intensive production system with two milkings per day, and an annual average of 1,878 milking cows, excluding dry cows. The dairy farm averaged 31.6 L/cow per day (9,638 L/cow per 305 days), had a 21-day pregnancy rate of 25% and an annual culling rate of 42%. The data were extracted from the dairy management software Dairy Comp 305 (Valley Agricultural Software, Tulare, CA, USA).

Data arrangement and binary classification design

To construct an ML binary classification (Li and Tong, 2020), the common classification of culling into voluntary or involuntary is not useful (Fetrow *et al.*, 2006). So, the culling classification was adapted from Fetrow (1987). The first category is forced culling (non-data-based), which represents cows that have to leave the farm immediately or soon and for which there is no reasonable solution other than culling. This category includes diseases (i.e., contagious mastitis, and lameness), trauma and death. These cows were not considered in this study.

The second category is economic culling (data-based), which is the objective of this study. This category represents those cows selected by the dairy manager at the end of each month during 2020. The applied criteria for each cow considered the features of milk yield (L/d), days in milk (d), lactation number (*n*), pregnancy status (Pregnant/Non-pregnant), days open (d) and days pregnant (d). These cows were labelled as (i) *Cull* ($n = 408$) and represented the first label of the binary classification.

The second label represents those cows that were not selected to be culled and were labelled as (ii) *Stay* ($n = 20{,}800$) these enable a ratio to be determined (*Cull*/*Stay*) 1:50. As described by Krawczyk (2016), an imbalanced dataset represents a common challenge in ML, where the optimal scenario in a binary classification will be a ratio of 1:1, but this is uncommon in real-life scenarios. To reduce this imbalance without losing power in the *Cull* label, an undersampling technique was applied to the *Stay* label to achieve a ratio of (*Cull*/*Stay*) 1:10, by the 'RandomUnderSampler' procedure of the library imbalanced-learn (Lemaître *et al.*, 2017), this represents a ratio closer to that which the algorithms would normally have to deal with, on a typical dairy farm (Liu *et al.*, 2008). Therefore, the final database for the ML binary classification resulted in 408 cows for the label *Cull* and 3993 for the label *Stay*.

Data preprocessing and statical analysis

Descriptive statistics were performed, and a scatterplot matrix was created as described by Emerson *et al.* (2013). This was done to visualize the multivariate data that included categorical and quantitative features from both labels using the function 'ggpairs' from the package 'GGally 2.2.1', in R v4.4.1 statistical software (https://www.R-project.org/). To avoid highly correlated features that can affect the performance of the learning process, a correlation analysis was performed among the six features before the training and testing phase using the library 'seaborn' (Waskom, 2021). A heat-map plot and the corresponding values are shown in Figure S1, where correlations >0.9 were not found, and thus all the features were included in the learning models.

Feature scaling was considered because the features have different ranges, to which some ML algorithms, including k-nearest neighbors (k-NN), support vector machine (SVM) and multilayer perceptron (MLP) are sensitive. Two methods were tested, and the one with the best performance was selected for each algorithm (Ali and Faraj, 2014). The first method is standardization, which removes the mean, and scales to one unit of standard deviation, and is computed as follows:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{x}$$

The second method was max–min normalization, which scales the features in the range 0–1, following the formula:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

ML algorithms and implementation

In the present study, six ML classifier algorithms were compared to represent a range of algorithmic approaches, from linear algorithms to those including ensembles to neural network-based methods. Choosing the correct ones depends on several factors, including dataset size, feature types, data dimensionality, interpretability, linearity and computational constraints. It is a common practice to compare several algorithms that can theoretically perform the desired task (Shahinfar *et al.*, 2014), and from the performance comparison select the best one. As described by Wolpert (1996): 'No single classifier works best across all possible scenarios', according to the 'No Free Lunch' theorem.

The preprocessed database (scaled for k-NN, SVM and MLP) was randomly split into training and testing sets using 'train_test_split' from 'sklearn.model_selection' (Pedregosa *et al.*, 2011), with 70% allocated to train and 30% to test, with a fixed random seed to ensure reproducibility.

To optimize the performance, hyperparameter tuning was conducted using grid search with cross-validation. The GridSearchCV procedure was coupled with a tenfold stratified *k*-fold cross-validation to ensure robust performance estimates across class-balanced folds, accuracy was defined as the scoring metric, and parallel processing enabled. The process involved fitting the base model with all combinations of the parameter grid across the training set defined ahead for each classifier, yielding the best-performing model based on the highest mean cross-validated accuracy. The optimal hyperparameters and corresponding cross-validated score were retrieved, and the best estimator was used for the subsequent evaluation.

To run the supervised ML models, the Google Colaboratory (Bisong, 2019) was used. This platform is a hosted Jupyter Notebook, running Python 3.10.12 (Python Software Foundation). The libraries 'pandas', 'matplotlib' and 'NumPy', were used to handle databases and plot results. The ML classifier algorithms were obtained from the library 'scikit-learn' (Pedregosa *et al.*, 2011) and a brief description and implementation of each one are described below.

Logistic regression

The first algorithm used was logistic regression (LR), a special type of generalized linear model that employs a Bernoulli distribution and the Logit link function, and can be used as a classifier to predict a probability threshold. In ML, it can be used for binary, one-vs-rest or multinomial logistic regression with optional L_1, L_2 or Elastic-Net regularization, which is usually used in ML but not in statistics (Zaidi and Al Luhayb, 2023).

The implementation of LR was performed using the 'LogisticRegression' function of the library 'scikit-learn', with the following hyperparameters defined in the grid search: *penalty* (L_1 and L_2), *C* (0.001, 0.01, 0.1, 1, 10, 100 – with smaller values indicating stronger regularization), *solver* ('liblinear' and 'lbfgs') and *maximum iterations for convergence* (100, 500), ensuring sufficient iterations for solver convergence (Geng *et al.*, 2024).

Gaussian naïve Bayes

The Gaussian naïve Bayes (GNB) algorithm for classification is an implementation of a Bayesian network, that is based on the probabilistic Bayes theorem assuming conditional independence of features, and where the likelihood of the features is assumed to be Gaussian (Kelly and Johnson, 2021). GNB was applied using the function 'GaussianNB' of the library 'scikit-learn' with *var smoothing*. This hyperparameter addresses numerical stability by adding a fraction of the largest variance across all features to the estimated variances of each feature per class. The parameter grid was defined as 1×10^{-11}, 1×10^{-10}, 1×10^{-9}, 1×10^{-8}, 1×10^{-7}, 1×10^{-6}, representing a range of smoothing factors from minimal to moderate adjustment (Wickramasinghe and Kalutarage, 2021).

k-nearest neighbours

The third classifier employed a type of nearest neighbour, k-NN, which belongs to a group of methods commonly described as non-generalizing ML algorithms. Because they remember all their training data, an internal model is not built. The principle behind the k-NN is to find the nearest neighbours (Thirunavukkarasu *et al.*, 2018).

The implementation of k-NN used the function 'KNeighborsClassifier' of the library 'scikit-learn'. The hyperparameter grid was defined as follows: number of neighbours to consider (1, 3, 5, 7, 9, 11 and 15), weighting strategy for neighbours where 'uniform' assigns equal weights and 'distance' weights neighbours inversely proportional to their distance, power parameter for the Minkowski distance metric where Manhattan distance and Euclidean distance, methods for nearest neighbour search (*auto, ball tree, kd tree* and *brute*) (Cunningham and Delany, 2021).

Support vector machine

The basic principle behind SVM is the ability to learn through the construction of an optimal hyperplane in a dimensional space that has the largest distance and maximizes between the margin boundaries 'support vectors' of each label of target features (Cervantes *et al.*, 2020). The implementation was through the SVM function from the 'scikit-learn' library, with the following hyperparameters: kernel function to transform the data (*linear, poly, rbf* and *sigmoid*), which defines the decision boundary's shape; regularization parameter *C* (0.01, 0.1, 1, 10, 100, 1000), which controls the trade-off between maximizing the margin and minimizing classification errors; kernel coefficient (*scale, auto*, 0.001, 0.01, 0.1, 1, 10); degree of the polynomial kernel (2, 3, 4, 5); shrinking heuristic (*True, False*), which reduces the number of support vectors considered during optimization, potentially speeding up training with minimal impact on accuracy. Both options were tested to assess computational efficiency versus performance (Pisner and Schnyer, 2020).

Random forest

RF is made up of an ensemble of several DT that are independent from each other, which allows for increased precision and reduced variance compared to one DT (Parmar *et al.*, 2019). The method of 'bagging' (bootstrap-aggregation) generates many DT by bootstrap samples from the training dataset, in this case, the improvement of RF is that it only uses a subset of the features on each tree-split, and then predicts by the most voted tree.

The implementation was through the function 'RandomForestClassifier' from the 'scikit-learn' library, with the following hyperparameters: number of trees in the forest (50, 100 and 200), maximum depth of each tree (7, 10, 20 and 30) *none*, which allows trees to grow until all leaves are pure – potentially leading to overfitting – smaller depths restrict tree complexity, reducing overfitting but risking underfitting, minimum number of samples required to split an internal node (2, 5, 10 and 20), minimum number of samples required at a leaf node (1, 2, 4 and 8), number of features to consider at each split (*sqrt* and *log2*),

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

weighting strategy for class imbalance (*balanced* and *none*). Also, using bootstrap (*True*) sampling enables bootstrap sampling, creating diverse trees by sampling with replacement, false uses the entire training set for each tree, reducing randomness but potentially increasing overfitting. Both options were tested to assess their impact on ensemble diversity (Biau and Scornet, 2016).

Multilayer perceptron

The last algorithm employed was MLP, a type of artificial neural network with an architecture that consists of an input layer that is equal to the number of scaled features (i.e., milk yield, days in milk), hidden layers and the output layer that receives the values from the last hidden layer following to a transformation into output values (*Stay* or *Cull*) (Gardner and Dorling, 1998).

The implementation was through the function 'MLPClassifier' from the 'scikit-learn' library, with the following hyperparameters *grid*: the activation functions that introduces non-linearity (*tanh* and *relu*), *hidden layer size* was 1 hidden layer with (50, 100 and 200). The optimization algorithm determines how the weights are updated (*lbfgs* and *sgd*), L_2 *regularization* (0.0001, 0.001, 0.01, 0.1 and 1.0) that controls the strength of regularization by adding a penalty to the loss function and the learning rate for 'sgd solver' (*constant, invscaling* and *adaptative*) (Badem *et al.*, 2017). Also, due to the computational demand of the MLP, the hardware accelerator TPU v2 (Google, Mountain View, CA, USA) was employed for the training and testing process.

Model performance evaluation

To evaluate the abilities of the ML algorithms employed on the supervised culling models task, several quantitative measures of performance were computed on the outputs of the testing phase after the models were trained (Luque *et al.*, 2019; Chicco and Jurman, 2020). The first metric is a 2 × 2 confusion matrix, which describes the outcome of the binary classification task, as follows:

$$\text{Confusion Matrix}\,\mathbf{M} = \begin{pmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{pmatrix}$$

where TP is the true positive (*Cull* cows correctly classified), TN is the true negative (*Stay* cows correctly classified), FP is the false positive (actual *Stay* cows wrongly classified as *Cull*) and FN is the false negative (actual *Cull* cows wrongly classified as *Stay*). A normalization was applied to scale the values, typically so that each row sums to 1 (or 100%), making it easier to interpret the model's performance across classes, especially when class imbalances exist.

From this confusion matrix several metrics were computed (Luque *et al.*, 2019). Accuracy is the most commonly used among researchers, represents the fraction of correct classifications of the algorithm, is the ratio of correctly classified positive and negative, for all instances of the dataset, where 0 is the worst value and 1 the best value. It is computed as follows:

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{n}^{+}+\mathrm{n}^{-}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

Precision or positive predictive value is the ratio of true positive to all positive (true and false), so it determines the ability of a classifier to correctly predict positive instances, and ranges from a worst value of 0 to a best value of 1, calculated as follows:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Recall or sensitivity is the ratio of true positive to true positive and false negative. It is the ability of a classifier to find the positive predictions over all the results that are positive, and was estimated as follows:

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

The F_1-score is one of the most used tests to evaluate the performance of ML predictions, and is defined as the harmonic mean of precision and recall. F_1-score ranges from 0 to 1, and has the following shape:

$$\mathrm{F}_1\ \text{score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The alternative measure to the F_1-score is the Matthews correlation coefficient (MCC), which has the advantage of being unaffected by imbalanced datasets (Chicco and Jurman, 2020). It ranges in the interval −1 for perfect misclassification to +1 for perfect classification, and is computed as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\cdot(\mathrm{TP}+\mathrm{FN})\cdot(\mathrm{TN}+\mathrm{FP})\cdot(\mathrm{TN}+\mathrm{FN})}}$$

A model inspection technique (permutation feature importance) was performed to estimate the contribution of each feature to the performance of the ML algorithms by using the procedure of the library scikit-learn permutation importance with the following hyperparameters: the number of repeats = 30, and random state = 42 (Fisher *et al.*, 2019).

Results and discussion

Labels description

Table 1 shows the summary and descriptive statistics of both labels (*Cull* and *Stay*) for quantitative features, while Table 2 shows the summary of the categorical features of the labels (*Cull* and *Stay*). Figure 1 shows generalized pairs plots for exploratory data analysis. In this learning approach, six features were employed per cow, three for productive characteristics and another three for reproductive performance.

The mean of the feature milk yield was lower in the *Cull* label (22.0 ± 5.7 L/d) than in the *Stay* label (31.7 ± 6.3 L/d). This can be explained by the fact that milk yield is one of the most important metrics related to voluntary culling. For days in milk, the mean was higher in the *Cull* label than in the *Stay* label (249.7 ± 93.2 vs 219.5 ± 79.0 d) in the *Stay* label; this trend in the *Cull* label is usually accompanied by a low milk yield and a failure to conceive. The feature lactation number had a similar distribution in cows with both labels. In both labels, most of the cows were in first lactation, 50% in the *Cull* label and 41% in the *Stay* label. The *Stay* label shows the distribution of lactations in the herd; this pattern is commonly observed in herds with high reproductive performance (Hare *et al.*, 2006; Fricke *et al.*, 2023). Even then, it can be supposed that in the *Cull* label, most of the cows would have three lactations or more, and this can be observed in involuntary culling, where the ages play a key role in the incidence of diseases (e.g., Rilanto *et al.*, 2020). It is important to consider that a first lactating cow does not guarantee a continued presence in the herd until the productive and reproductive parameters are met.

The reproductive parameters that comprise the secondary reason of importance are related to voluntary culling according to productive parameters such as milk yield (Weigel *et al.*, 2003). The

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

Table 1. Summary and descriptive statics of both labels (*Cull* and *Stay*), for quantitative features

Label	Feature	*n*	Mean	SD	Minimum	Maximum	95% CI
Cull	Milk yield (L/d)	408	22.0	5.7	0.2	43	21.5–22.6
	Days in milk (d)		249.7	93.2	87	568	240.6–258.8
	Days open (d)		193.4	100.3	55	466	183.7–203.2
	Days pregnant (d)		48.4	61.8	0	184	42.4–54.4
Stay	Milk yield (L/d)	3993	31.7	6.3	8	78	31.5–31.9
	Days in milk (d)		219.5	79.0	70	487	217.0–221.9
	Days open (d)		115.5	55.5	33	296	113.8–117.2
	Days pregnant (d)		98.5	75.6	0	222	96.2–100.9

SD, standard deviation; CI, confidence interval at 95%.

Table 2. Summary of the categorical features of the binary labels (*Cull* and *Stay*)

Label	Feature	*n*	%
Cull	Lactation number		
	1	204	50.00
	2	66	16.18
	3	48	11.76
	4	60	14.71
	5	30	7.35
	Pregnancy status		
	Pregnant	168	41.18
	Non-pregnant	240	58.82
Stay	Lactation		
	1	1668	41.77
	2	864	21.64
	3	780	19.53
	4	438	10.97
	5	243	6.09
	Pregnancy status		
	Pregnant	2850	71.37
	Non-pregnant	1143	28.63

first feature is pregnancy status (pregnant or non-pregnant), and the pregnant cows were higher in the *Stay* label 71% vs 41% in the *Cull* label. These indicate that a confirmed pregnancy does not guarantee permanence in the herd, basically due to other influencing factors. The mean of days of pregnancy were lower in the *Cull* label than in the *Stay* label (48.4 ± 61.8 vs 98.5 ± 75.6 days). This indicates that if a pregnant cow is culled, it would have an early pregnancy, which can be explained by those cows that have too many days in milk and low milk yield. Also, this tendency can be seen in the feature days open which represents the period since calving to conception, where the *Cull* label has a mean of 193.4 ± 100.3 vs *Stay* label with 115.5 ± 55.5 (days). This indicates that for some reason the cows in the *Cull* label experienced more problems with being pregnant, so even if a cow was pregnant, they were too advanced in their lactation curve.

Classification performance

Figure 2 shows the confusion matrix of the best cross-validation model of all the algorithms employed. In Table 3, several metrics were calculated from the confusion matrix and presented as described by Chicco and Jurman (2020). The performance metrics revealed significant variability across the six classifiers, reflecting differences in their underlying assumptions, capacity to model complex decision boundaries and robustness to data characteristics such as feature interactions, noise and class imbalance.

Several factors likely contributed to the variation in performance across classifiers. The first is model complexity and expressiveness of a low complexity (LR, GNB): linear (LR) or independence-based (GNB) models underperform due to limited capacity to capture non-linearities or feature dependencies, yielding lower MCC (0.725, 0.503) and high complexity (k-NN, SVM, RF, MLP): non-parametric (k-NN, RF) or non-linear (SVM with RF, MLP) models excel by adapting to intricate patterns, achieving MCC > 0.91. RF and MLP had high F_1 scores (0.992, 0.983) reflecting their ability to fit the data latent structure. Another factor is the decision boundary flexibility; LR linear boundary and GNB Gaussian assumption constrain adaptability, and missing edge cases (higher FN). k-NN local, RF piecewise, SVM’s kernel-based and MLP’s layered boundaries align better with the data separability, minimizing errors.

Regarding robustness of assumptions: GNB independence violation degrades performance (lowest recall $= 0.459$). LR assumes linearity, underfitting non-linear effects. k-NN, RF and MLP impose fewer assumptions, thriving on the data empirical structure (precision ≥ 0.992). Also, the dataset likely has strong clusterability (k-NN success), non-linear interactions (RF, MLP) and moderate noise (SVM robustness). The variation arises from each algorithm’s capacity to model the data underlying distribution and decision boundary. RF and k-NN outperform due to their flexibility and data-driven nature, while NB and LR lag due to restrictive assumptions. SVM and MLP offer a middle ground, excelling with tuning. These results suggest the culling task benefits from non-linear, high-capacity models, with RF as the top performer due to its ensemble robustness and feature interaction handling. In Figure S2, the learning curves are presented and reveal that RF, k-NN and MLP are the best fits for this

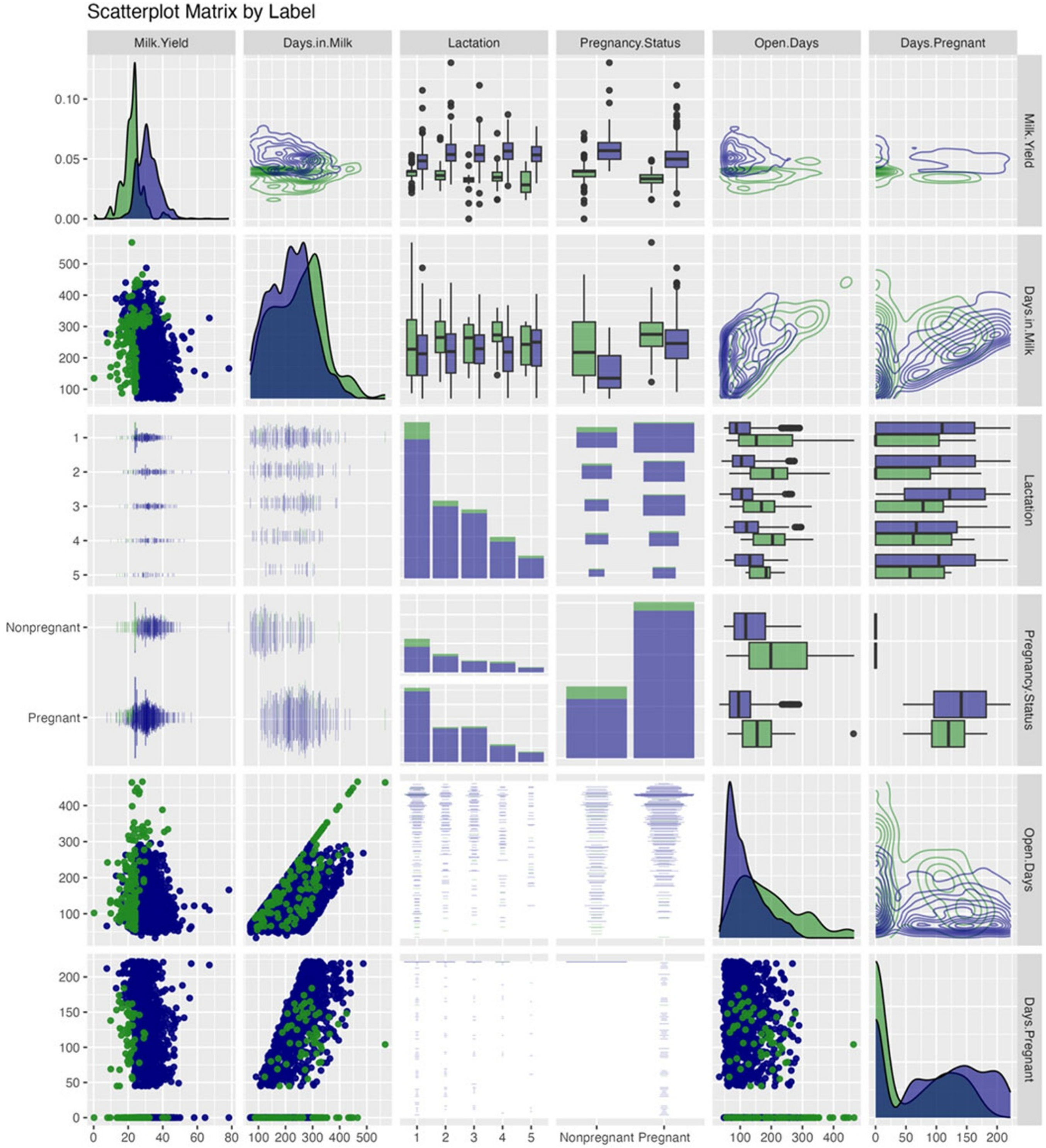

Figure 1. Generalized pairs plot for the six features. *Notes*: The quantitative–quantitative (milk yield, days in milk open days and days of pregnancy) pairs are shown in scatterplots and density plots. The categorical–categorical (pregnancy status and lactation) pairs are shown in mosaic plots. The quantitative-categorical pairs are shown in boxplots and barcode plots. The diagonal (from the left upper to right lower) reflects the marginal distributions of the features. Green = data for the label *Cull* and blue = data for the label *Stay*.

dataset, with RF leading due to its high accuracy and stability, with a cross-validation accuracy of ~0.995 and low variability, making it the most reliable choice for this dataset. The learning curve for the RF model shows that the training samples (ranging from 0 to 2,500) represent the individual data points used to train the model. As the number of training samples increases, the training score remains consistently high at around 0.995, while the cross-validation score improves and stabilizes at approximately 0.985. By the time the model reaches 2,500 samples, it demonstrates robustness to data-related issues,

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

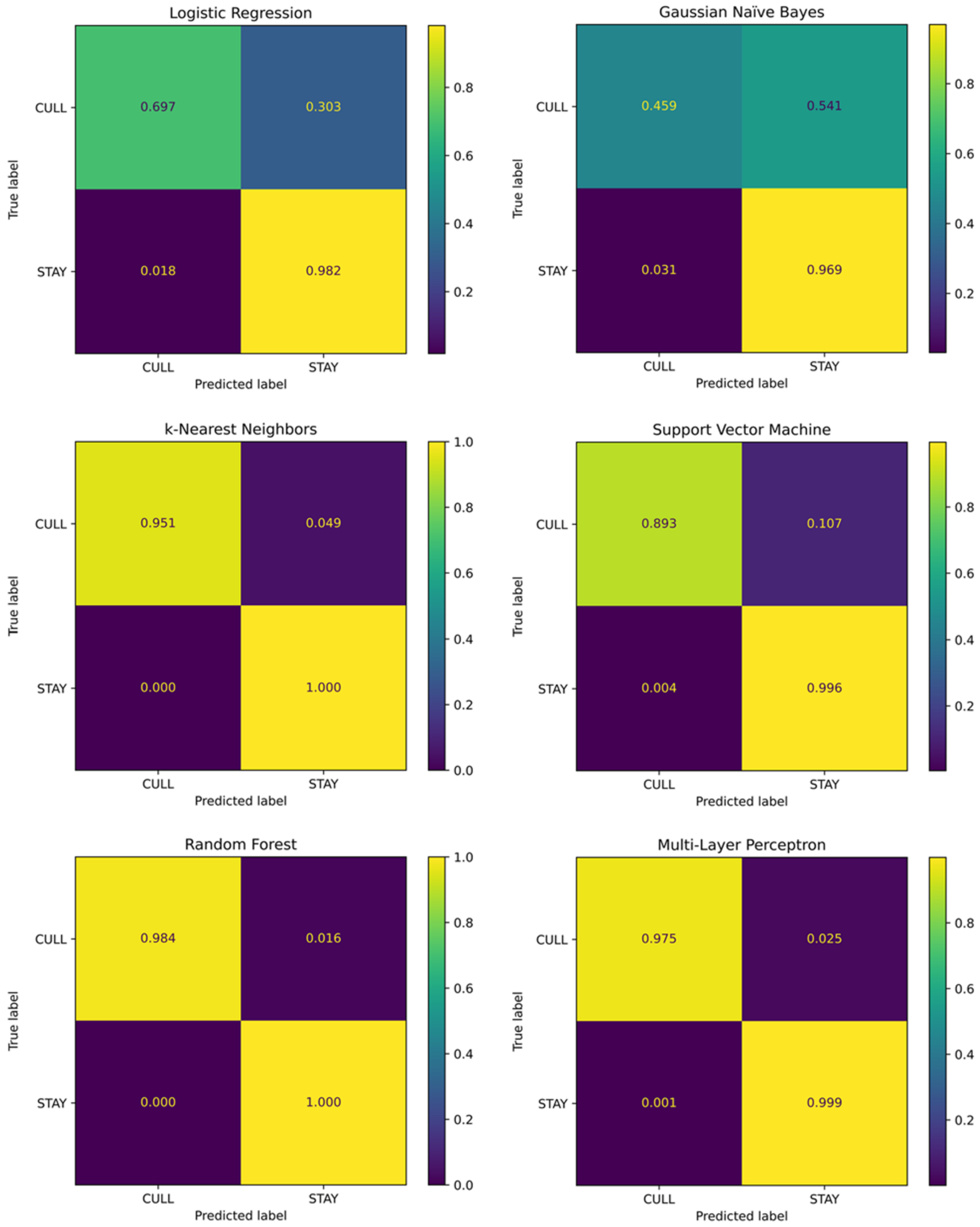

Figure 2. Normalized confusion matrix of the classifier algorithms employed. *Notes*: Each row corresponds to an instance of the true class (*Stay* or *Cull*), while each column corresponds to an instance of the predicted class. As a result, the diagonal elements indicate the proportion of correctly identified classes. Misclassification is reflected in the off-diagonal elements, which represent instances incorrectly assigned to a different class due to confusion.

as the gap between the training and cross-validation scores narrows – indicating minimal overfitting. This suggests that the model generalizes well when provided with a sufficient amount of data, with performance plateauing beyond this point. LR and GNB struggle, likely due to their inability to capture the data complexity.

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

Table 3. Results of the metrics for the evaluation of the prediction performance of the ML classifiers

Algorithm	Accuracy[a]	Precision[a]	Recall[a]	F_1-score[a]	MCC[b]
Logistic regression	0.955	0.794	0.697	0.742	0.725
Gaussian naïve Bayes	0.922	0.602	0.459	0.521	0.503
k-nearest neighbors	0.995	1.000	0.951	0.975	0.973
Support vector machine	0.986	0.956	0.893	0.924	0.917
Random forest	0.998	1.000	0.984	0.992	0.991
Multi-layer perceptron	0.997	0.992	0.975	0.983	0.982

[a]These metrics have a scale of 0–1.
[b]This metric has a scale of −1 to +1.

The feature importance results (Figures S3–S8) highlight that milk yield was the primary driver of culling decisions, followed by reproductive features like days open and days pregnant. Lactation number has minimal impact, and days in milk and pregnancy status vary in importance depending on the model. Overall, RF, k-NN and MLP provided the most reliable insights due to their high performance, while GNB struggles due to its restrictive assumptions.

To summarize, we aimed to transfer the knowledge of culling cows of an expert (farmer or manager) entity (regardless of whether the expert entity was correct or not) to an AI, to prove that specific culling criteria can be learned and, in the future, used as a tool for culling dairy cows in individual dairy farm. It is noteworthy to say that this methodology is for individual herds since each one has its culling criteria. For example, milk yield is one of the most important metrics related to culling, where the level of this production at which the cows are candidates for culling, changes between herds, usually, this is determined by economic estimations like for example income over cost. Also, these criteria are affected by and differs among or according to geographical regions, milk price, feeding costs, replacement costs, climate factors (heat stress) and herd feeding management (Buza *et al.*, 2014).

To our knowledge, a methodology like the one here proposed has not previously been reported in the scientific literature, but other approaches using AI have been used. For example, Lopez-Suarez *et al.* (2018) used DT to predict the milk yield in the first lactation to identify those cows poorly productive in the early stages with an accuracy of 83%. Ouseleye *et al.* (2023) used RF and they were able to predict the longevity of dairy cows using productive and genetic features with a coefficient of determination of 86%. Adamczyk *et al.* (2016) tested several ML algorithms to identify the culling reasons but with unfavourable results.

The number of features that are incorporated into the training and testing process of AI modelling will depend on the data availability. In this study, six features for each label were taken into account because these are the features that the expert entity (farmer or manager) of these farms takes into account for the decision-making in voluntary culling. There may be a case where more or fewer features need to be taken into account, for example, the presence of some disease, where animals that present this disease have more weight in the decision of culling. But *a priori* a common practice in ML is to make a dimensionality reduction by several algorithms, such as principal component analysis and their variations, where only the features that explain the maximum amount of variance are incorporated in the learning process, these have the advantage of increasing the efficacy and reduce the computational time (Bharadiya, 2023).

The feasibility of this methodology applied under on-farm conditions would need an ML-based software system following these conditions as described by Foidl and Felderer (2019): (i) An expert entity (human rationale, i.e., farmer or manager) would be needed to load the information in the interphase as the data source, indicating which cows leave the heard so the system knows that is for voluntary culling. Also, all the information of all the cows would be online, so a stable internet connection would also be needed. (ii) The engineering environment of the pre-training steps includes data integration, data storage, transformation and normalization. (iii) A final step where the selected algorithms are continuously learning from data, the culling criteria of the herd, in the first months would not be precise, but over time as more information is loaded the algorithms would increase their accuracy. (iv) If all the prior steps work correctly, the farmer or the technician would be able to consult in the interface the results of the trained ML algorithms, which would be cows that do not justify their place in the dairy herd based on the criteria learned from data. Even though this approach can look complicated, the more feasible way is that this type of engineering is linked with the daily management software of the farm like Dairy Comp 305 (Valley Agricultural Software, Tulare, CA, USA), AfiFarm (Afimilk, Kibbutz Afikim, Israel) or SmartDairy (Boumatic, Boumatic, Madison, WI, USA) as examples, so the farmer doesn't have to load the information in several platforms.

In conclusion, the learning approach proposed in this research shows that data-based culling criteria can be learned and predicted by AI, with variable performance among ML algorithms. In this study, with the specific features used by this herd, RF is the optimal candidate to be deployed in on-farm conditions. To increase the reliability of these procedures, these approaches need to be tested in several farms, under different scenarios (i.e., geographic and management variability) and varieties of features.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0022029925101416.

Acknowledgements. Oscar R. Espinoza-Sandoval was supported by the PhD scholarship (2022-000018-02NACF-11517) from the National Council of Humanities, Sciences, and Technologies (CONAHCYT-Mexico).

Competing interests. The authors have not stated any conflicts of interest.

References

Adamczyk K, Zaborski D, Grzesiak W, Makulska J and Jagusiak W (2016) Recognition of culling reasons in Polish dairy cows using data mining methods. *Computers and Electronics in Agriculture* **127**, 26–37.

Ali PJM and Faraj RH (2014) Data Normalization and Standardization: a Technical Report. *Machine Learning Technical Reports* **1**, 1–6.

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

Badem H, Basturk A, Caliskan A and Yuksel ME (2017) A new efficient training strategy for deep neural networks by hybridization of artificial bee colony and limited–memory BFGS optimization algorithms. *Neurocomputing* **266**, 506–526.

Berry DP (2021) Invited review: beef-on-dairy—The generation of crossbred beef × dairy cattle. *Journal of Dairy Science* **104**, 3789–3819.

Bharadiya JP (2023) A tutorial on principal component analysis for dimensionality reduction in machine learning. *International Journal of Innovative Science and Research Technology* **8**, 2028–2032.

Biau G and Scornet E (2016) A random forest guided tour. *TEST* **25**, 197–227.

Bisong E (2019) Google Colaboratory. In Bisong E (ed), *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA: Apress, pp. 59–64.

Brooks M (2024) Inside the maths that drives AI. *Nature* **631**, 244–246.

Buza MH, Holden LA, White RA and Ishler VA (2014) Evaluating the effect of ration composition on income over feed cost and milk yield. *Journal of Dairy Science* **97**, 3073–3080.

Cabrera VE (2012) A simple formulation and solution to the replacement problem: a practical tool to assess the economic cow value, the value of a new pregnancy, and the cost of a pregnancy loss. *Journal of Dairy Science* **95**, 4683–4698.

Calsamiglia S, Astiz S, Baucells J and Castillejos L (2018) A stochastic dynamic model of a dairy farm to evaluate the technical and economic performance under different scenarios. *Journal of Dairy Science* **101**, 7517–7530.

Carvalho PD, Santos VG, Giordano JO, Wiltbank MC and Fricke PM (2018) Development of fertility programs to achieve high 21-day pregnancy rates in high-producing dairy cows. *Theriogenology* **114**, 165–172.

Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L and Lopez A (2020) A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* **408**, 189–215.

Chicco D and Jurman G (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics Decision Making* **20**, 16.

Cockram MS (2021) Invited review: the welfare of cull dairy cows. *Applied Animal Science* **37**, 334–352.

Compton CWR, Heuer C, Thomsen PT, Carpenter TE, Phyn CVC and McDougall S (2017) Invited review: a systematic literature review and meta-analysis of mortality and culling in dairy cattle. *Journal of Dairy Science* **100**, 1–16.

Cunningham P and Delany SJ (2021) k-Nearest Neighbor Classifiers – a Tutorial. *ACM Computing Surveys* **54**, 128.

De Vries A (2004) Economics of Delayed Replacement When Cow Performance is Seasonal. *Journal of Dairy Science* **87**, 2947–2958.

De Vries A and Marcondes MI (2020) Review: overview of factors affecting productive lifespan of dairy cows. *Animal* **14**, 155–164.

Emerson JW, Green WA, Schloerke B, Crowley J, Cook D, Hofmann H and Wickham H (2013) The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics* **22**, 79–91.

Espinoza-Sandoval OR, Angeles-Hernandez JC, Gonzalez-Ronquillo M, Ghavipanje N, Zhang N, Bayat AR, Hervás G, Knolif AE, Mele M, Loor JJ, Stergiadis S and Vargas-Bello-Pérez E (2024) Dairy farming in the era of artificial intelligence: Trend or a real game changer? *Journal of Dairy Research* **91**, 139–145.

Ferrari V, Marusi M, Penasa M, van Kaam Jbchm, Finocchiaro R and Cassandro M (2024) A tool to optimise dairy herd replacements combining conventional, sexed, and beef semen. *Italian Journal of Animal Science* **23**, 409–415.

Fetrow J (1987) Culling dairy cows. In Williams E (eds), *Proceedings of the 20th Annual Convention American Association of Bovine Practitioners*. Stillwater, OK, USA. Phoenix, AZ: Frontier Printers, Inc., pp. 102–107.

Fetrow J, Nordlund KV and Norman HD (2006) Invited review: Culling: Nomenclature, definitions, and recommendations. *Journal of Dairy Science* **89**, 1896–1905.

Fisher A, Rudin C and Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* **20**, 1–81.

Foidl H and Felderer M (2019) Risk-based data validation in machine learning-based software systems. In Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE 2019). Association for Computing Machinery, New York, NY, USA, pp. 13–18.

Fricke PM, Wiltbank MC and Pursley JR (2023) The high fertility cycle. *JDS Communications* **4**, 127–131.

Gardner MW and Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* **32**, 2627–2636.

Geng Y, Li Q, Yang G and Qiu W (2024) Logistic Regression. In Geng Y, Li Q, Yang G and Qiu W (eds), *Practical Machine Learning Illustrated with KNIME*. Singapore: Springer, pp. 99–132.

Groenendaal H, Galligan DT and Mulder HA (2004) An economic spreadsheet model to determine optimal breeding and replacement decisions for dairy cows. *Journal of Dairy Science* **87**, 2146–2157.

Hare E, Norman HD and Wright JR (2006) Survival Rates and Productive Herd Life of Dairy Cattle in the United States. *Journal of Dairy Science* **89**, 3713–3720.

Jayasundara S, Worden D, Weersink A, Wright T, VanderZaag A, Gordon R and Wagner-Riddle C (2019) Improving farm profitability also reduces the carbon footprint of milk production in intensive dairy production systems. *Journal of Cleaner Production* **229**, 1018–1028.

Kelly A and Johnson MA (2021) Investigating the Statistical Assumptions of Naïve Bayes Classifiers. *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, pp. 1–6.

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* **5**, 221–232.

Lehenbauer TW and Oltjen JW (1998) Dairy cow culling strategies: Making economical culling decisions. *Journal of Dairy Science* **81**, 264–271.

Lemaître G, Nogueira F and Aridas CK (2017) Imbalanced-learn: a Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **18**, 1–5.

Li JJ and Tong X (2020) Statistical hypothesis testing versus machine learning binary classification: distinctions and guidelines. *Patterns* **1**, 100115.

Liu XY, Wu J and Zhou ZH (2008) Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **39**, 539–550.

Lopez-Suarez M, Armengol E, Calsamiglia S and Castillejos L (2018) Using Decision Trees to Extract Patterns for Dairy Culling Management. In Iliadis L, Maglogiannis I and Plagianakos V (eds), *Artificial Intelligence Applications and Innovations. AIAI 2018. IFIP Advances in Information and Communication Technology*, Vol. 519. Cham: Springer, pp. 231–239.

Luque A, Carrasco A, Martín A and de Las Heras A (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition* **91**, 216–231.

Marshall J, Haley D, Levison L, Kelton DF, Miltenburg C, Roche S and Duffield TF (2023) A survey of practices and attitudes around cull cow management by bovine veterinarians in Ontario, Canada. *Journal of Dairy Science* **106**, 302–311.

Nielsen LR, Jørgensen E, Kristensen AR and Østergaard S (2010) Optimal replacement policies for dairy cows based on daily yield measurements. *Journal of Dairy Science* **93**, 75–92.

Ouseleye A, Rodriguez J and Araujo H (2023) Machine Learning models to optimize dairy cow culling: the case of the experimental farm UniLaSalle Beauvais. *Computer Aided Chemical Engineering* **52**, 1135–1140.

Parmar A, Katariya R and Patel V (2019) A review on random forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P and Baig Z (eds) *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, Vol. 26. Cham: Springer.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press

Pisner DA and Schnyer DM (2020) Support vector machine. In Mechelli A and Vieira S (eds), *Machine Learning*. Cambridge, MA, USA: Academic Press, pp. 101–121.

Python Software Foundation. Python Language Reference, Version 3.10. Available at http://www.python.org.

Rilanto T, Reimus K, Orro T, Emanuelson U, Viltrop A and Mõtus K (2020) Culling reasons and risk factors in Estonian dairy cows. *BMC Veterinary Research* **16**, 173.

Shahinfar S, Page D, Guenther J, Cabrera V, Fricke P and Weigel K (2014) Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *Journal of Dairy Science* **97**, 731–742.

Slob N, Catal C and Kassahun A (2021) Application of machine learning to improve dairy farm management: a systematic literature review. *Preventive Veterinary Medicine* **187**, 105237.

Sorge US, Kelton DF, Lissemore KD, Sears W and Fetrow J (2007) Evaluation of the Dairy Comp 305 Module "Cow Value" in Two Ontario Dairy Herds. *Journal of Dairy Science* **90**, 5784–5797.

Thirunavukkarasu K, Singh AS, Rai P and Gupta S (2018) Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning. *4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1–4.

Waskom ML (2021) Seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021.

Weigel KA, Palmer RW and Caraviello DZ (2003) Investigation of Factors Affecting Voluntary and Involuntary Culling in Expanding Dairy Herds in Wisconsin using Survival Analysis. *Journal of Dairy Science* **86**, 1482–1486.

Wickramasinghe I and Kalutarage H (2021) Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing* **25**, 2277–2293.

Wolpert DH (1996) The existence of a priori distinctions between learning algorithms. *Neural Computation* **8**, 1391–1420.

Zaidi A and Al Luhayb ASM (2023) Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Mathematical Problems in Engineering* **1**, 5525675.

https://doi.org/10.1017/S0022029925101416 Published online by Cambridge University Press