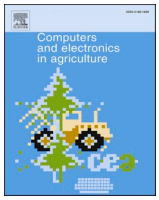
Computers and Electronics in Agriculture 230 (2025) 109860

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



A viable data driven method for the assessment of the productivity level of dairy cows in future lactations

Marco Bovo [a,*], Miki Agrusti [a], Laura Ozella [b], Claudio Forte [b], Daniele Torreggiani [a], Patrizia Tassinari [a]

[a] Department of Agricultural and Food Sciences - University of Bologna, 40127 Bologna (BO), Italy
[b] Department of Veterinary Sciences - University of Turin, 10095 Grugliasco (TO), Italy

ARTICLE INFO

Keywords:
Dairy cow
Lactation curve
Machine learning
Milk yield
KNN
SVM

ABSTRACT

The challenge of increasing the economic and environmental sustainability of the dairy cattle sector involves several aspects and among these, the milk production throughout the career of a cow, is perhaps the parameter that all farmers would like to know for a more efficient planning of entries and exits. In fact, if on one hand numerous researchers are studying the problem selecting most efficient animals, on the other hand, few studies have focused on the definition of tools forecasting the productivity of dairy cows in the future lactations starting from the data collected in past lactations. This aspect has a particular importance in the first years of a cow since, as well known, first lactation usually has lower production than subsequent lactations. For a farmer it is important to know, as soon as possible, if a specific animal will have on a long term, lactations with high, medium or low milk productivity. The current availability of large dataset collected by automatic milking systems or by electronic milking parlors, paves the way for application of big data approaches based on machine learning algorithms with classification learner representing one of the most promising data-driven tools.

In this study, firstly two supervised learning methods, i.e., Super Vector Machine and K-Nearest Neighbors, have been applied to a large dataset of 720 complete lactations, with the object to train machine learning tools able to classify and separate first and second lactation. The two classification algorithms have been applied to the raw dataset and after the application of a dimensionality reduction method. Four different dimensionality reduction methods (i.e., ISOMAP, UMAP, MDS and t-SNE) have been tested to evaluate the most efficient for this application. Finally, the same two classification algorithms have been used for the attribution of the productivity level of the second lactation starting from data of the first lactation. The two classification methods reached very encouraging accuracy values, ranging from 70% to 73%, indicating that selected predictors despite their simplicity look very promising and entail for the definition of enhanced future models. In fact, the method is particularly interesting for practical applications, as it represents a viable support-to-decision tool for selecting the most productive animals.

Interpretive summary

The need to increase the sustainability of the animal production sector makes essential to search for new tools to support farmers and vets, and in this the modern numerical techniques based on artificial intelligence are becoming increasingly popular. Especially in the dairy sector, the increase in sustainability is closely linked to the selection of the cows capable of guaranteeing high milk production. In this work a new approach based on machine learning models is proposed for the identification of the most productive animals and for the assessment of the long-term milk productivity class of a specific cow.

1. Introduction

The challenges of the sustainability in the dairy cattle sector involve milk yield and milk quality levels, cow health and wellbeing, efficient resource use, and emissions reduction. Due to the effects on milk production and quality, which have an impact on how effectively natural resources are used, animal welfare is, at the end, directly linked to sustainability, and as widely demonstrated, increasing animal welfare

* Corresponding author.
E-mail address: marco.bovo@unibo.it (M. Bovo).

https://doi.org/10.1016/j.compag.2024.109860
Received 31 January 2024; Received in revised form 24 August 2024; Accepted 18 December 2024
Available online 28 December 2024

usually increase the milk yield (Allen et al., 2015; Kino et al., 2019). To this regard, in the recent years, several steps forward have been made to increase the production per lactation of the individual animal by working on the genetic selection, on feeding, increasing animal welfare and the quality of the housing environment (Chamberlain et al., 2022; Zhou et al., 2022). Many of these actions are related to daily management decisions that the farmer, nowadays, can undertake with the help of commercial management tools or decision-support systems often associated with sensors or technologies that allow real-time monitoring of the production and health status of the individual animal. In fact, following the Precision Livestock Farming (PLF) approach (Berckmans, 2014; Tullo, Finzi and Guarino, 2019; Lovarelli, Bacenetti and Guarino, 2020), in technological farms, data concerning different parameters of behavior and activity of cows, animal health and welfare are collected from different sensors (e.g., individual cow data recording system, activity tags such as pedometers or neck collars, ear tags for rumination monitoring, automatic concentrate feeders), and used for the daily management of the herd (Bovo et al., 2021). Furthermore, the growing widespread of automatic milking systems (AMSs) and electronics milking parlors (EMPs) provide farmers and technicians with continuous series of detailed data useful to assess health conditions and evaluate parameters connected to the milk quality and quantity (Ozella et al., 2023). But, while most of the recent studies investigated models primarily focusing on the prediction of the daily milk yield for the running lactation period (Jones, 1997; Ji et al., 2022) one of the still open matters involves the prediction of cow productivity in future lactation periods (Rebuli et al., 2023a). This aspect is particularly important in the first years of life of cows, because as is well known, the first lactation usually has a lower production compared to subsequent lactations (Masía et al., 2020), and for a farmer it is important to know, as soon as possible if, compared to the other animals in the herd, a specific animal will have, on a long term, high, medium or low milk productivity (Arulnathan et al., 2020). Then, for the dairy sector, one of the next big challenges will be the development of a system/tool able to classify the future productivity of a cow based on what the cow produced in the past or is currently producing. With reference to this, nowadays, the availability of large dataset collected by AMSs and EMPs, make possible the application of big data approaches (Fuentes et al., 2020) and especially those based on machine learning algorithms(Dulhare, Ahmad and Ahmad, 2020). For these research problems, a classification learner can represent one of the most promising numerical tools (Frades and Matthiesen, 2010; Everitt et al., 2011). Actually, the problem could be divided into two closely related aspects. The first is related to the identification of the features that characterize the lactation number of an animal, classifying first lactation separated from the following lactations. All this allows to estimate the value of a metric providing a measure of the distance between the two clusters (i.e., the cluster of the daily milk yield time series of the first lactation and the cluster of the time series of the lactation periods two and higher). Instead, the second research aim is related to the classification of the productivity level (e.g., low, medium or high) of a cow, in future lactations, starting from the knowledge of the productivity features of its first lactation.

In the present paper these two still open questions have been approached starting from the assumption that first and second lactations have milk yield trends considerably different. So, two supervised learning methods, i.e., the Super Vector Machine (SVM) and the K-Nearest Neighbors (KNN), have been applied to a large dataset with the object to train models for the classification between first and second lactation curves. The two classification algorithms have been applied to the raw dataset and also after the application of a dimensionality reduction method. Four different dimensionality reduction methods (i.e., ISOMAP, UMAP, MDS and t-SNE) have been tested to evaluate the most efficient for this application. Finally, for those cows having available data from first and second lactation curve, the two learning methods have been used for the attribution of a productivity class (i.e., low, medium or high) to the second lactation of a specific cow starting from the data available from the first lactation of the same cow.

2. Materials and methods

This section is structured in the following way. First, the description of the main characteristics of the dataset used in the study is introduced. Then, a focus on data cleaning and data filling process is provided. Moreover, the dimensionality reduction methods used in the work are described. The last section provides the details of the two classification methods which constitute the basis for the data-driven method proposed in the paper for the attribution of future cow productivity class.

2.1. Dataset description

The dataset used in the work was gathered from March 2020 to May 2022 in 13 farms (labelled as in Table 1) located in the Po Valley region, in northern Italy. The 13 farms are equipped with a Merlin AMS (Fullwood Packo, England) that collects data on daily milk yield (DMY) and milk quality (i.e., fat, protein and lactose) for each cow. The size of the 13 herds is similar and the farms have about 60 milking cows each one. In order to uniform the dataset length of the different cows, the lactation period was assumed at the maximum of 305 days in milk (DIM). Therefore, for cows having a long lactation, only the first 305 days have been considered. Moreover, the lactation data has been considered valid for the analyses only if it contains data for at least 250 days. In the two-year time span under consideration, the farms provided data of the valid lactations summarized in Table 1.

In total, the number of unique animals in the dataset is 683 and the number of lactations is 720 (i.e., 465 first lactations and 255 s lactations). The average DMY values, for the two lactation numbers, are also provided in Table 1 for each farm.

2.2. Data cleaning and data filling

The time series of the DMY are sometimes irregular in length and can contain missing values due to missing or wrong records. In the work, all the lactation having more than 20 consecutive missing values, out of the 305 days of the conventional lactation, have been excluded from the analysis. When time series still contained days with no data, the empty cells were filled with the values obtained by the Wood model (Wood, 1967) that best fitted the available data for the single lactation.

2.3. Dimensionality reduction methods

As already cited, dimensionality reduction algorithms are now even widely used in machine learning in order to make easier the work of classifiers and regressors (Velliangiri, Alagumuthukrishnan and Thankumar Joseph, 2019; Ahmad and Nassif, 2022).

Table 1
Statistical summary of the daily milk yield (DMY) for the 13 farms.

Farm ID	N. of first lactations	N. of second lactations	Average DMY of first lactations	Average DMY of second lactations
BS03	20	15	38.10	40.41
BS13	58	34	29.81	35.38
BS14	59	43	33.86	40.84
CR04	45	38	31.85	38.60
CR09	33	18	37.65	45.75
CR12	19	8	33.40	36.50
MN01	30	14	28.76	39.49
MN05	38	16	37.36	46.25
MN07	39	20	35.20	46.45
MN08	3	3	39.86	41.68
MN10	26	10	28.00	32.61
MN15	59	20	33.63	38.40
VR11	36	16	32.40	36.10

They can be categorized into two main groups: the first seeking to preserve the pairwise distance structure amongst all the data samples and the second that favor the preservation of local distances over global distance (McInnes, Healy and Melville, 2020). In the following of this subsection the dimensionality reduction methods used in the paper will be introduced and shortly described. The four dimensionality reduction methods considered here have been selected because of their intrinsic advantages related efficiency and scalability. They can handle large datasets and are generally applicable to various types of data, including images, text, and sensor recordings.

2.3.1. MDS

Multidimensional Scaling (MDS) algorithm (Mead, 1992)attempts to find an embedding from the initial feature vectors in the high-dimensional space such that distances (e.g., the Euclidean distance or more generally a metric or an arbitrary distance function) are preserved in a low-dimensional space. The MDS algorithm works by minimizing an objective function, called the strain or stress function, based on the discrepancy of these distances (Mignotte, 2011). Its goal is to find the multi-dimensional data projections in the lower-dimensional space (R2 or R3) to maintain the similarity or inconsistency of the data. It optimally maps the object's proximity index to the distance between the multidimensional spatial points and visualizes the data so that users can test structured assumptions or discover hidden patterns in the data (Jia et al., 2022). The original algorithm was introduced first time in 1994 (Cox T.F. and Cox M.A.A., 2000) and then improved.

2.3.2. ISOMAP

The Isometric feature Mapping (Isomap) method starts from the assumption that only the geodesic distance reflects the true geometry of the underlying manifold (Tenenbaum et al., 2000). Isomap is a dimensionality reduction technique that aims to preserve the intrinsic structure and geometry of high-dimensional data in a lower-dimensional space. It achieves this by considering the pairwise distances between data points and constructing a graph representation of the data. The algorithm starts by constructing a neighborhood graph, where each data point is connected to its nearest neighbors. Then, it computes the geodesic distances between all pairs of points on this graph. These geodesic distances represent the actual distances along the manifold, taking into account the underlying structure of the data. Next, Isomap uses classical MDS to find a low-dimensional embedding that best preserves the pairwise geodesic distances. MDS aims to find a configuration of points in a lower-dimensional space that mimics the pairwise distances as closely as possible. The result is a low-dimensional representation of the data points, where the distances between points roughly correspond to their geodesic distances on the underlying manifold. This embedding allows for visualizing and analyzing the data in a more interpretable manner, as well as potentially improving the performance of downstream machine learning algorithms.

2.3.3. UMAP

As for other k-neighbor graph-based algorithms, Uniform Manifold Approximation and Projection (UMAP) is set up in two phases. In the first phase a particular weighted k-neighbor graph is constructed. In the second phase a low dimensional layout of this graph is computed. The differences between all algorithms in this class amount to specific details in how the graph is constructed and how the layout is computed (McInnes, Healy and Melville, 2020b).

2.3.4. t-SNE

The t-distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. The cost function used by t-SNE differs from the one used in the original Stochastic Neighbor Embedding (SNE) (Hinton and Roweis, 2002) in two aspects: (i) it uses a symmetrized version of the SNE cost function with simpler gradients that was briefly introduced by Cook et al. (2007) and (ii) it uses a Student's t-distribution rather than Gaussian distribution to compute the similarity between two points in the low-dimensional space. t-SNE employs a heavy-tailed distribution in the low-dimensional space to alleviate both the crowding problem and the optimization problems of SNE method (van der Maaten and Hinton, 2008).

2.4. Classification methods

In this section the two machine learning algorithms used for the lactation classification are described. For both classifiers the k-fold cross-validation procedure was followed in order to keep a stable training-test configuration and to have the best estimation of the classification performance (Witten, Frank and Hall, 2011). The dataset of each cow was sampled with a cross-validation procedure, a resampling procedure evaluating machine learning models on a limited data sample. In particular, the k-fold cross-validation procedure was considered by adopting a k value equal to 20 (so adopting a 20-fold cross-validation procedure). In this procedure, the dataset was divided into 20 equal parts (i.e., groups) and the training/testing process has run 20 times each time with a group used as test, the holdout group, and the others 19 groups used to train the model. In the study, the train and test values are randomly selected by the extraction algorithm. The accuracy metric of each prediction was used to evaluate the performance of each model (Bovo et al., 2024). The confusion matrix is the tool we have used to calculate the global accuracy of the classification methods. The accuracy value of each model is computed as the sum of the observations along the main diagonal of the confusion matrix divided by the total value of observations in the matrix (i.e. the sum of all the cells of the matrix).

2.4.1. SVM

The support-vector machine (SVM) method developed by Vapnik (Cortes and Vapnik, 1995) is based on statistical learning theory. At the first approximation SVM finds a separating line (or hyperplane) between data of different classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes. According to the SVM algorithm, it finds the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called margin. The goal of the algorithm is to maximize the margin in order to define the optimal hyperplane (i.e. the hyperplane for which the margin is maximum). If data are clearly not linearly separable it's impossible to draw a straight line that classify the data and then SVM can convert original data to linearly separable data with a nonlinear transformation of the original space. In its most simple type, SVM doesn't support multiclass classification natively. It supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems.

2.4.2. KNN

The k-nearest neighbors (KNN) algorithm (Cover and Hart, 1967) is a supervised machine learning algorithm used to solve both classification and regression problems. The algorithm assumes that similar things exist in close proximity. In other words, similar things are sufficiently near to each other. The main steps followed by KNN are the following:

Initialize K to your chosen number of neighbors;
For each example in the data, it calculates the distance between the query example and the current example from the data and then adds the distance and the index of the example to an ordered collection;
Sort the ordered collection of distances and indices from smallest to largest by the distances;

Pick the first K entries from the sorted collection;
Get the labels of the selected K entries;
If regression, return the mean of the K labels, otherwise if classification, return the mode of the K labels.

2.4.3. Parameters assumed for the classification analyses

The analyses have been realized in MATLAB environment (Matlab R2023b, 2023). Two standard functions implemented in MATLAB have been used for train and test the two classification models. For applying the SVM classificatory algorithm the *Optimizable SVM* function has been used whereas for the application of the KNN algorithm the *Optimizable KNN* function has been adopted. In order to allow the full reproducibility of the analyses the set of parameters and options assumed for the analyses are summarized in Table 2.

2.5. Lactation curve modelling and labelling

Preliminarily to the classification analyses and with the aim to compare the trends of the different lactation numbers, the concept of lactation curve, i.e., the relation between DMY and DIM, must be defined. The lactation curve was assumed, in accordance with the Wood's model (Wood, 1967):

$$DMY(DIM) = \mathrm{a}\cdot e^{-\mathrm{b}\cdot DIM}\cdot DIM^{c} \quad (1)$$

where: DMY is the daily milk yield at a lactation stage equal to the DIM value; a, b and c are the three parameters that control the shape of the lactation curve: a is the scaling factor that controls the production at beginning of lactation and the peak production; b and c influence respectively the post peak behavior and the final slope of the lactation curve. Collected data were stratified by lactation number (assumed equal to parity number) and the model in Eq.(1) was used as fitting model. Then, an average lactation curve was defined for lactation number one and number two (see Fig. 1). As shown in Fig. 1, the main difference in lactation curves is between the model of primiparous (i.e. parity equal to 1) and the models of cows with parity equal to two. Primiparous cows (i.e., the red line) generally present a lower peak production with a flatter post-peak behavior. The curve of lactation number 2 shows a similar behavior with a more evident production peak at about 50–60 days in milk.

Table 2
Set of parameters and options assumed for the classification analyses.

Parameter / Option	SVM	KNN
Kernel scale	0.001–1000	–
Kernel function	Gaussian, Linear, Quadratic	–
Standardize data	True	True
Multiclass coding	One-vs-One	–
Box constraint level	0.001–1000	–
PCA	Disabled	Disabled
Optimizer	Bayesian optimization	Bayesian optimization
Acquisition function	Expected improvement per second plus	Expected improvement per second plus
Max n. of iterations	30	100
Training time limit	False	False
Number of neighbors	–	1–20
Distance metric	–	City block, Chebyshev, Correlation, Cosine, Euclidean, Hamming, Jaccard, Spearman
Distance weight	–	Equal, Inverse, Squared inverse

Indeed, first parity cows show a lower peak in production, but maintain an almost constant DMY until the end of the lactation period (i.e., 305 days). On the other hand, the curves of second lactations have a higher peak DMY but a steeper decrease in the second stage of lactation (for DIM value between 50 and 200 days). In general, the total milk yield (TMY) for lactation increases after the first one. In the light of this, it is possible to confirm, as expected, that the first lactation curve of a cow is considerable different from the second and the following. This result is well known in literature, but was a fundamental preliminary check prior to the application of the classification models to the available dataset for the study.

Moreover, as already described above, we used the values of the Wood curve that best fitted the available data of the single lactation to fill the possible empty cells.

Finally, for the purpose of the classification analyses each lactation has been labelled with a productivity class based on the TMY calculated out of the 305 days starting from the data collected by the AMS. Three classes have been created for lactation number 1 and 2 identifying three different productivity levels, i.e., high production (HP), medium production (MP) and low production (LP). The threshold values (in kg) selected for the three classes of the first lactations are (5600,9800], (9800,11400) and [11400,15500) respectively for LP, MP and HP classes. The threshold values selected for the three classes of the second lactations are (6100,12000], (12000,13500) and [13500,20500) respectively for LP, MP and HP classes.

3. Results and discussion

3.1. Performance of the classification first/second lactation

For the purpose of the first research problem, the two classificatory algorithms i.e., SVM and KNN were applied to the 305-day time series of the 720 lactation curves in the dataset. The two confusion matrices are equal to [402, 63; 72, 183] and [444, 21; 63, 192] respectively for the SVM and KNN algorithms. The obtained classification accuracies are equal to 0.79 ± 0.03 and 0.88 ± 0.05 respectively for the SVM and KNN algorithms. From the analysis of the results, it emerges that KNN has higher accuracy in the classification. In fact, 585 lactations out of 720 were correctly labelled in comparison to the 507 of the SVM method. The good results of the classification procedure confirm the existence of some intrinsic traits in the time series trends and used for the classification. In fact, for example, Fig. 2 shows the weights attributed by the SVM method to the milk production of the different DIM. It is worth noting that weights range from 0.11 to 0.008 (see Fig. 2a) confirming that for the classification task some particular days during the lactation period are much more important than others (see Fig. 2b) and then are much more influencing in the classification process performed by the two algorithms. This as a whole means that, in the days with highest weights, the trends of first lactation curves have the most significant differences from the trends of second lactation curves.

In addition, in order to facilitate the classification task, a dimensionality reduction method has been applied to the lactation curves data. The performance of four different methods has been tested. The effects of the application of the dimensionality reduction methods can be seen in Fig. 3, that shows the scatter plot of the lactation curves in a 2-dimension space. In the figure, a dot represents a lactation curve. In the scatter plot, the two dimensions (i.e., the two axes) are obtained as nonlinear combination of the original dimensions (i.e., the original features).

As Fig. 3 displays, the dots of the first lactation (with label "1″ and blue colored) are rather well separated by those of the second lactation (with label "2" and orange colored). This effect can be appreciated for every one of the four dimensionality reduction methods considered in the study. As a result, it is expected that the performances of the classification methods increase after the application of the dimensionality reduction methods. Table 3 collects the classification accuracies of the two algorithms without and with the application of a dimensionality

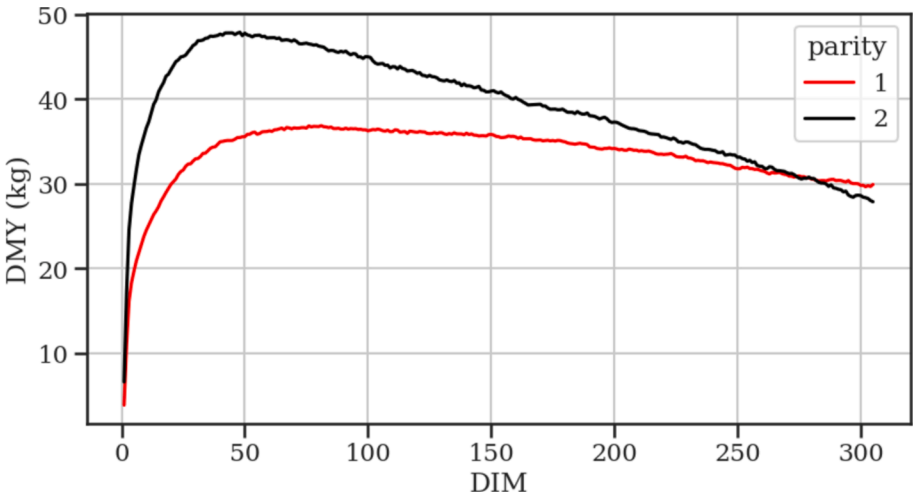


Fig. 1. Best fitting curve for the different lactation numbers. The red line indicates the first lactation. The black line indicates the second lactations.

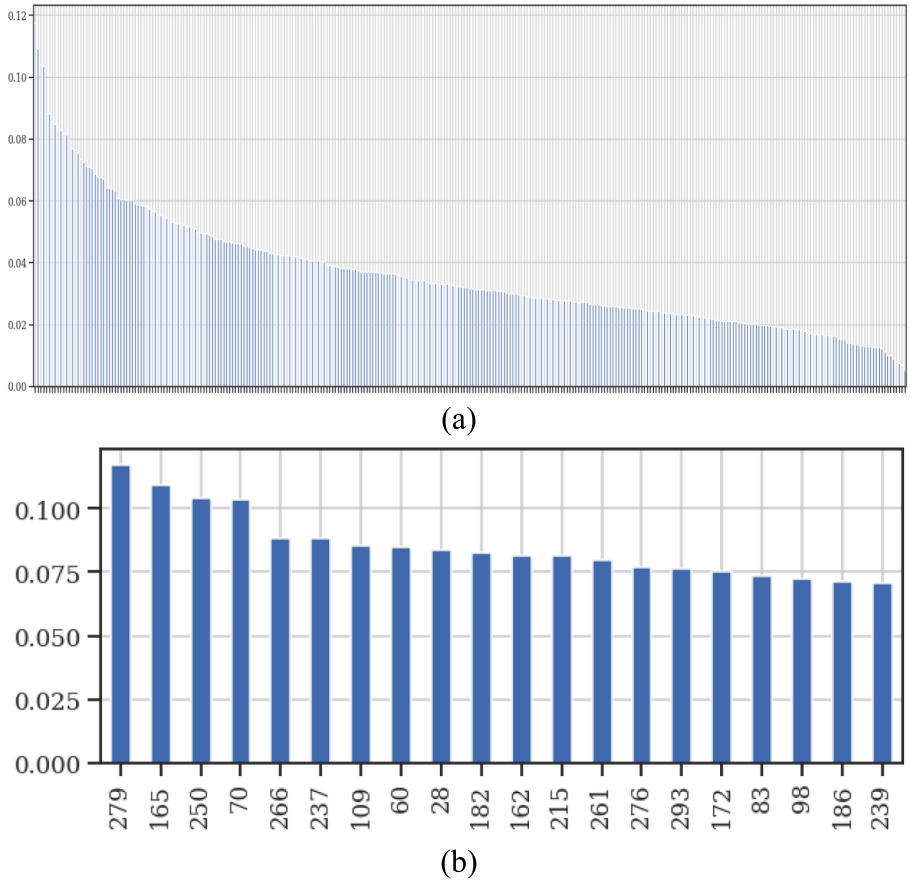


Fig. 2. Feature importance histogram graph (with DIM in the X-axis and with feature importance value in the Y-axis). (a) Weights of the 305 days in ascending order; (b) Detail of the 20 most important days out of the 305 values of DIM for the classification task.

reduction method. After the application of a dimensionality reduction method, SVM and KNN showed similar accuracies and provided good accuracy values in the range 87 %–89 %. Table 3 shows that accuracy of SVM has a significant improvement with the application of a dimensionality reduction method whereas, on the other hand, the accuracy of KNN is practically unaltered.

Figs. 4 and 5 show, in a scatter plot, the results of the classification process of SVM and KNN methods, respectively. In the figures, the two different dot colors, i.e., black and white, indicate correct/wrong classification of the specific lactation curve. Moreover, in Fig. 4 is visible the linear separation of the 2-dimension hyperplane into the two different regions that best separate first from second lactations. It is evident that the separating line is different for the four different subcases since the four different dimensionality reduction methods combine in a different way the original dimensions of the dataset and then project the lactation curve points in different way in the hyperplane. In analogous way can be seen the scatter plot graphs showed in Fig. 5 but with reference to the KNN classification method. In this case it

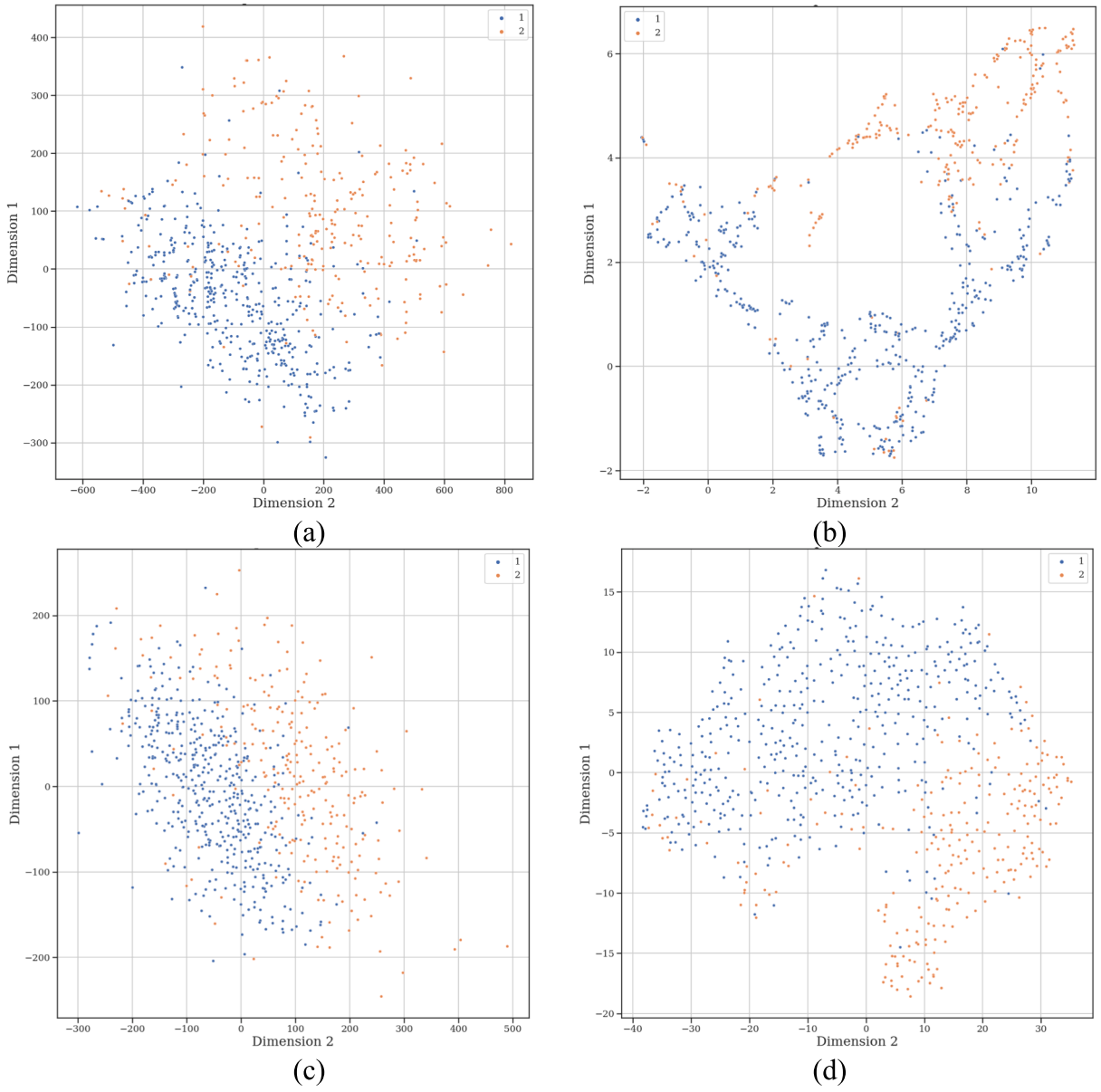


Fig. 3. Scatter plot of the lactation curves in a 2-dimension space after the application of (a) ISOMAP method, (b) UMAP method, (c) MDS method and (d) t-SNE method.

Table 3
Accuracy values of the classification methods SVM and KNN without and with the application of a dimensionality reduction method.

Classification method	Application to raw data (%)	After application of ISOMAP (%)	After application of UMAP (%)	After application of MDS (%)	After application of t-SNE (%)
SVM	79.0 ± 0.03	88.6 ± 5.8	89.9 ± 2.8	89.1 ± 1.6	89.1 ± 1.5
KNN	88.0 ± 0.05	88.1 ± 3.5	88.0 ± 3.2	88.0 ± 3.5	87.3 ± 2.9

is possible to see that the two regions that best separate first and second lactations are separated by strongly nonlinear jagged curves without explicit mathematical equation.

Finally, it is worth noting that for first lactation group the accuracies provided by KNN method in case of application to the raw data are: 88 %, 94 % and 95 % respectively for productivity classes 1, 2 and 3. On the other hand, for the second lactation, the percentage accuracies are: 90 %, 56 % and 87 % respectively for productivity classes 1, 2 and 3. This means probably that intrinsic features of the second lactation curves of the cows with medium productivity are poorly distinguishable and entail high probability of a wrong labelling in the classification procedure. Similar conclusions can be depicted by the analysis of the outcomes of the SVM method.

3.2. *Simplified model for the assessment of the milk yield of the lactation*

By the aggregate analysis of the data of the lactation curves, it emerges that DIM values have very different importance in the attribution of the lactation number class. On the other hand, the different DIM values (considered as a single feature by the classification methods)

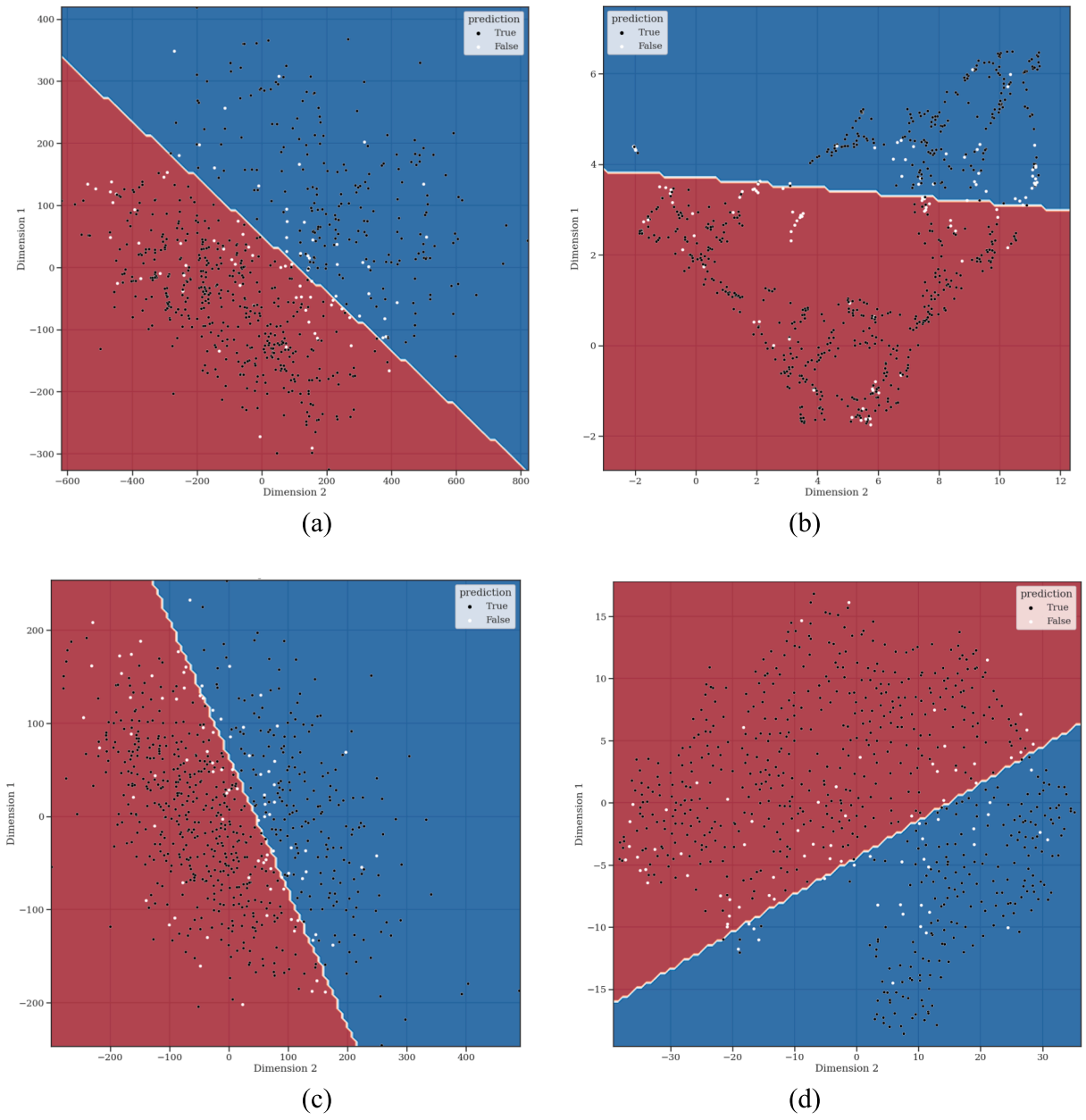


Fig. 4. Classification plot of the lactation curves in a 2-dimension space a for the SVM method after the application of (a) ISOMAP method, (b) UMAP method, (c) MDS method and (d) t-SNE method.

can be combined in order to provide new features (i.e., dimensions) useful for the classification. One of the more evident aspects, clearly highlighted by the lactation curves, is the different TMY per lactation. In Fig. 6, it is possible to see the boxplot of the TMY values, in order to evaluate the differences between the first and the second lactation dataset.

The detailed analysis of the outcomes of the classification methods supports the idea that TMY of a lactation could be assessed by the knowledge of few features of the same lactation. So, the authors considered the following equation:

$$TMY \approx \overline{TMY} = a_0 + a_1 \cdot y_{peak}^{a_2} + a_3 \cdot t_{80\%}^{a_4} \quad (2)$$

where: TMY is the measured total milk yield value, $\overline{TMY}$ is an assessment of TMY, a_0-a_4 are parameters to be evaluated by regression analysis, y_{peak} (kg) is the peak value of the daily milk yield recorded during the lactation and $t_{80\%}$ (days) is the number of days in which the daily milk yield results higher than 80 % of y_{peak}.

The regression model has been calibrated for the first and the second lactations in the dataset and the parameter values reported in Table 4 have been obtained via least squares regression.

For the first lactation the coefficient of determination (R^2) is equal to 0.9586 and the Root Mean Squared Error (RMSE) is equal to 364 kg. For the second lactation, R^2 is 0.9163 and RMSE is equal to 669 kg. Despite its simplicity, the model proposed here seems to be very accurate in predicting TMY for both first and second lactations (see Fig. 7). The general findings reported in this section will be used as input for the next step, i.e., the assessing of the productivity class of future lactations as described in the next sub-section.

3.3. Assessment of the productivity class of future lactations

As already discussed in the previous sections, in the dairy sector, one of the main challenges is the definition of a model evaluating the future

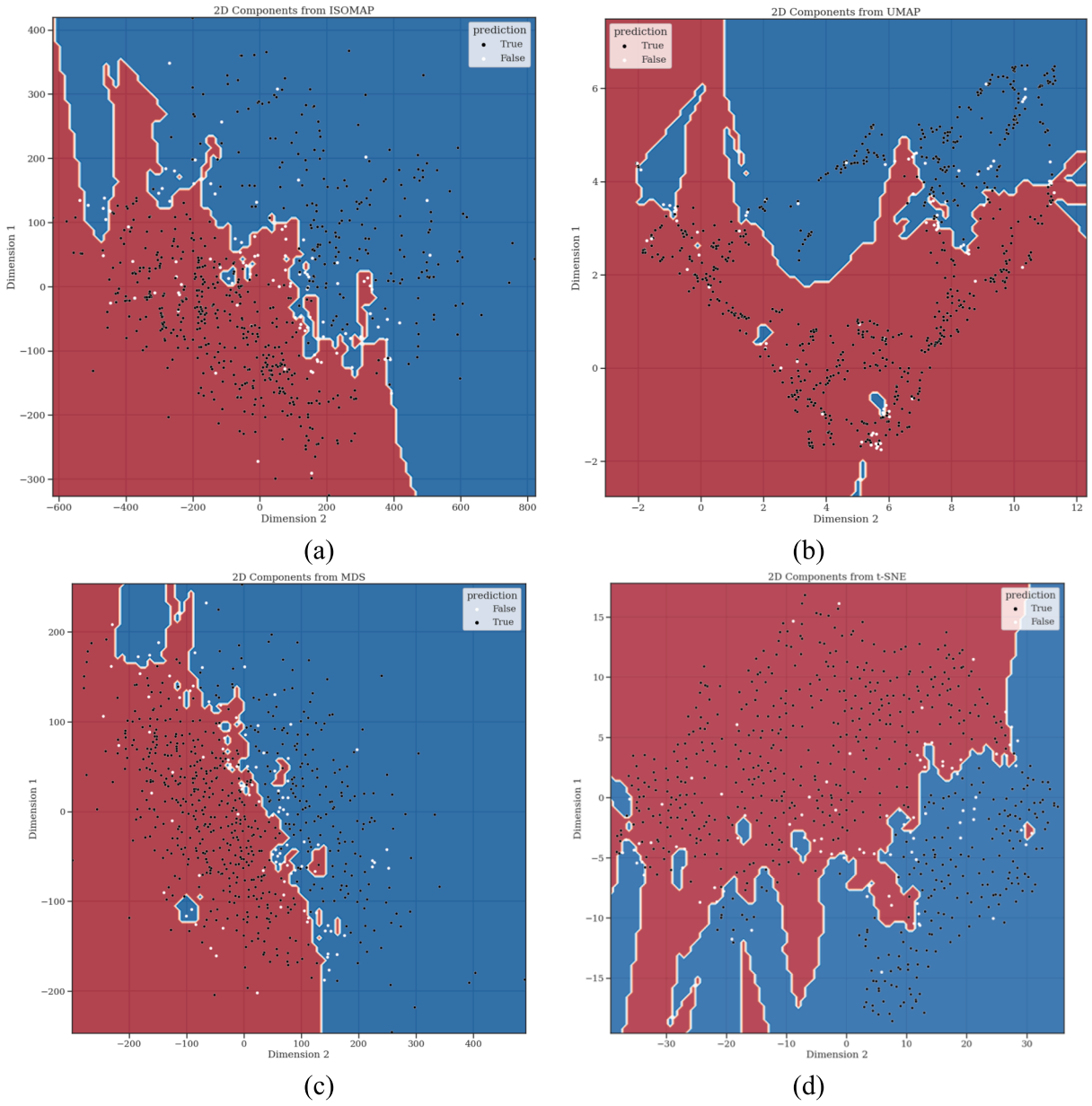


Fig. 5. Classification plot of the lactation curves in a 2-dimension space a for the KNN method after the application of (a) ISOMAP method, (b) UMAP method, (c) MDS method and (d) t-SNE method.

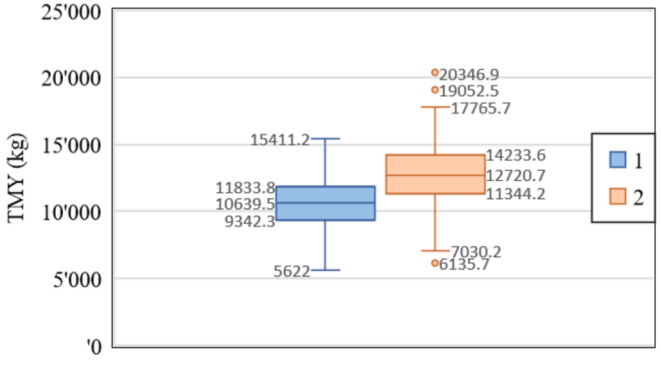


Fig. 6. Boxplot of the total milk yield (TMY) values for first (1) and second (2) lactation.

Table 4
Parameter values obtained for the model in Eq. (2) for first and second lactation.

Parameter	First lactation		Second lactation	
	Estimate	pValue	Estimate	pValue
a_0 (kg)	−1951.850	0.0319	−3597.144	0.0729
a_1 (−)	266.873	0.0163	175.537	0.1271
a_2 (−)	0.979	8.152e-26	1.060	7.9918e-14
a_3 (kg/days)	29.126	0.0275	366.757	0.3717
a_4 (−)	0.839	1.6361e-26	0.511	0.0027

productivity of a cow (in terms of milk yield) based on milk yield of previous lactations. So, in this work, for those cows having available data from first and second lactation curve, two learning methods have been used for the attribution of the productivity class (i.e., high production (HP), medium production (MP), low production (LP)) of the second lactation of a cow starting from data of the first lactation. The dataset contains the data of first and second lactation of 37 different

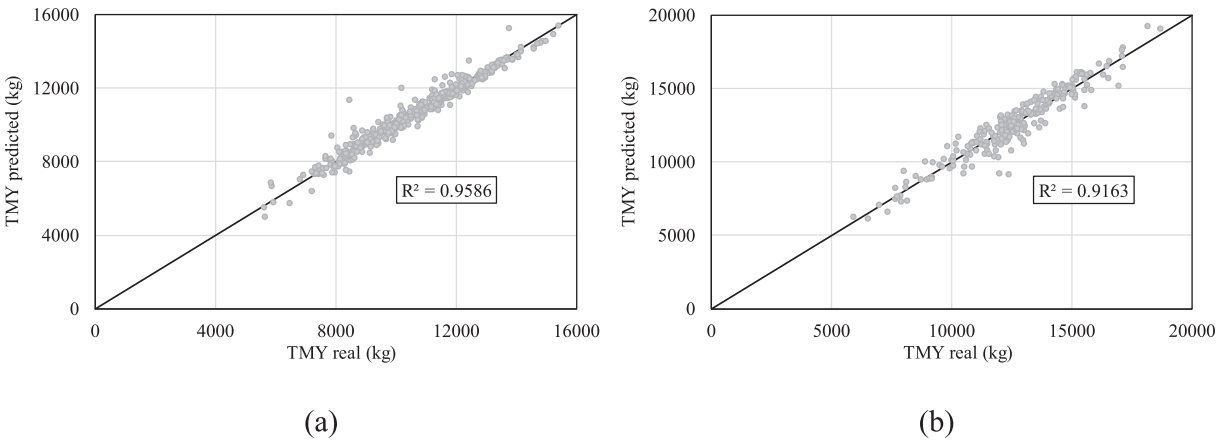


Fig. 7. Comparison between total milk yield (TMY) values real and predicted by Eq.(2) for (a) first lactation and (b) second lactation.

cows. In the first lactation, 13 has HP class, 9 has MP class and 15 LP class. The tools selected to approach this task are the classification methods used before, i.e., SVM and KNN algorithms. It is worth to note that only a partial group of cows maintains the same productivity class moving from first to second lactation. In fact, as shown in Fig. 8, 24 out of 37 lactations (i.e., 65 %) maintained the same label whereas 35 % of the lactations have moved to another productivity class. Further future research will investigate on a larger dataset the stability of this percentage, but the value is in accordance with the outcomes in (Rebuli et al., 2023b) confirming as an important group of primiparous cows has not stable production level moving from the first to the subsequent lactations.

The input features for the two classification models were y_{peak}, $t_{80\%}$ and the productivity class of each first lactation. The vector of the productivity class attributed to the second lactations was used for the training and the test of the models. The output of the models is the vector of the predicted productivity class of each second lactation. The most efficient way to show the results of the two methods is by means of the confusion matrix. They are reported in Fig. 9 for the two classification algorithms.

The classification methods reached an accuracy of 73 % (i.e. (10 + 17)/37 × 100) and 70 % (i.e. (10 + 16)/37 × 100) respectively for SVM and KNN. In the opinion of the authors these first values are very encouraging and indicate that the selected features, despite their simplicity, look very promising and entail further studies and development. It seems worth noting that the method could become particularly interesting for practical applications, as it represents a viable support-to-decision tool that farmers can adopt for the selection of the most productive animals to be kept in the herd compared to those to be discarded.

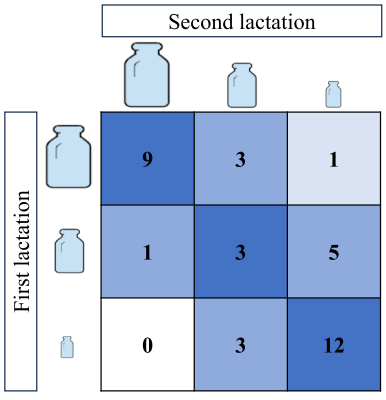


Fig. 8. Matrix of the frequencies for high, medium and low productivity lactations for first and second lactation number.

Although the classification methods proposed here provided good accuracy values, they could be further improved so as to minimize the probabilities of discarding animals with high yields or vice versa keeping some animals with low production in the herd. The current classification algorithm uses a mathematical model based on a very simple equation but it has been demonstrated in the paper to be sufficiently reliable. Among the current limitations of the study are the fact that the model has been trained and tested on a limited sample of lactations and furthermore it makes use of a fully data-driven algorithm (as in the black box modeling approach) in which no information or correction can be entered by the user. Moreover, the dataset contains lactations with same duration, i.e., 305 days, and the further improvement of the models will have to remove this limitation so to become applicable to more general cases.

So, future research steps will have to prove the ability of the algorithms to well perform on larger and diversified dataset, with the further objective of increasing the prediction accuracy. Moreover, the authors are working on the development of classification models in which also selected data, if available, can be entered. Those data must be related to a specific animal (e.g., genetics data, health data, feeding data, milk quality data) in order to have a more reliable but at the same time easy to control tool (as in the grey box modeling approach).

4. Conclusion

In this paper two supervised learning methods have been trained and tested with the main aim to properly recognize and label the first and the second lactation curves of dairy cows. The two classification algorithms have been applied to the raw dataset and then after the application of four different dimensionality reduction methods. Finally, for those cows having available data from first and second lactation curve, the two classification methods have been used for the attribution of the productivity class (i.e., low, medium, high) of the second lactation of a cow starting from its data on the first lactation. The classification methods reached accuracy values ranging from 70 % to 73 %. These values seem very encouraging and entail further studies for the definition of enhanced models useful as decision support tools for farmers for early selection of the most productive animals to keep in the herd.

CRediT authorship contribution statement

Marco Bovo: Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Miki Agrusti:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Laura Ozella:** Writing – review & editing, Validation, Methodology, Conceptualization. **Claudio Forte:** Writing –

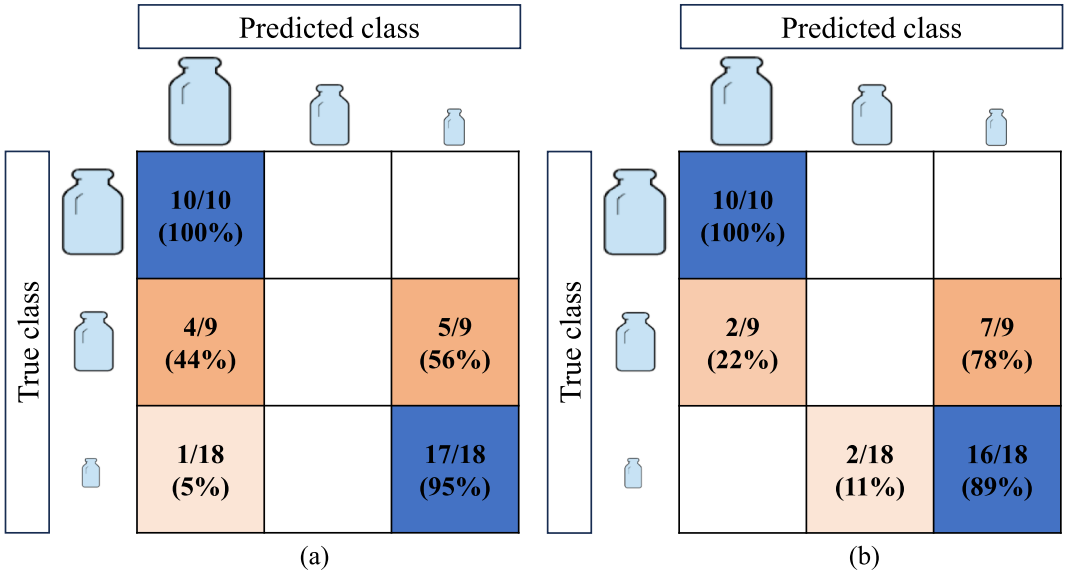


Fig. 9. Confusion matrices for (a) SVM algorithm and (b) KNN algorithm.

review & editing, Validation, Methodology, Conceptualization. **Daniele Torreggiani:** Writing – review & editing, Methodology, Conceptualization. **Patrizia Tassinari:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Funding

Part of this study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ahmad, N., Nassif, A.B., 2022. 'Dimensionality Reduction: Challenges and Solutions', ITM Web of Conferences [Preprint]. Available at: https://api.semanticscholar.org/CorpusID:247463581.

Allen, J.D., et al., 2015. 'Effect of core body temperature, time of day, and climate conditions on behavioral patterns of lactating dairy cows experiencing mild to moderate heat stress'. J. Dairy Sci., 98(1), pp. 118–127. Available at: Doi: 10.3168/jds.2013-7704.

Arulnathan, V., et al., 2020. Farm-level decision support tools: A review of methodological choices and their consistency with principles of sustainability assessment. J. Clean. Prod. 256. https://doi.org/10.1016/j.jclepro.2020.120410, 120410.

Berckmans, D., 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. Available at: Rev. Sci. Tech. Off. Int. Epiz 33 (1), 189–196. https://doi.org/10.20506/rst.33.1.2273.

Bovo, M., et al., 2021. Random forest modelling of milk yield of dairy cows under heat stress conditions. Animals. https://doi.org/10.3390/ani11051305.

Bovo, M., et al., 2024. DAIRY CHAOS: data driven approach identifying daiRY cows affected by HeAt lOad stress. Comput. Electron. Agric. 218. https://doi.org/10.1016/j.compag.2024.108729.

Chamberlain, A.T., et al., 2022. The relationship between on-farm environmental conditions inside and outside cow sheds during the summer in England: can Temperature Humidity Index be predicted from outside conditions? Animal - Open Space 1 (1). https://doi.org/10.1016/j.anopes.2022.100019, 100019.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/BF00994018. Available at:

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13 (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964. Available at:

Cox, T.F., Cox, M.A.A., 2000. Multidimensional Scaling. CRC Press, New York.

Dulhare, U.N., Ahmad, K., Ahmad, K.A.B., 2020. Machine learning and big data: concepts, algorithms, tools and applications. John Wiley & sons.

Everitt, B., et al. (2011) 'Cluster analysis'.

Frades, I., Matthiesen, R., 2010. 'Overview on Techniques in Cluster Analysis', in R. Matthiesen (ed.) *Bioinformatics Methods in Clinical Research*. Totowa, NJ: Humana Press, pp. 81–107. Available at: Doi: 10.1007/978-1-60327-194-3_5.

Fuentes, S., *et al.* (2020) 'Artificial intelligence applied to a robotic dairy farm to model milk productivity and quality based on cow data and daily environmental parameters', Sensors 2020, Vol. 20, Page 2975, 20(10), p. 2975. Available at: Doi: 10.3390/S20102975.

Hinton, G.E., Roweis, S.T., 2002. 'Stochastic Neighbor Embedding', in *Neural Information Processing Systems*. Available at. https://api.semanticscholar.org/CorpusID:20240.

Ji, B., et al., 2022. 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. Biosyst. Eng. 216, 186–197. https://doi.org/10.1016/j.biosystemseng.2022.02.013. Available at:

Jia, W., et al., 2022. Feature dimensionality reduction: a review. Complex & Intelligent Systems 8 (3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x. Available at:

Jones, T., 1997. Empirical bayes prediction of 305-day milk production. J. Dairy Sci. 80 (6), 1060–1075. https://doi.org/10.3168/jds.S0022-0302(97)76031-4. Available at:

Kino, E., et al., 2019. Exploration of factors determining milk production by Holstein cows raised on a dairy farm in a temperate climate area. Tropical Animal Health and Production 51 (3), 529–536. https://doi.org/10.1007/s11250-018-1720-6. Available at:

Lovarelli, D., Bacenetti, J., Guarino, M., 2020. A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production? J. Cleaner Prod. 262. https://doi.org/10.1016/j.jclepro.2020.121409. Available at:

Masía, F.M., et al., 2020. Modeling variability of the lactation curves of cows in automated milking systems. J. Dairy Sci. 103 (9), 8189–8196. https://doi.org/10.3168/jds.2019-17962. Available at:

McInnes, L., Healy, J., Melville, J., 2020a. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'.

McInnes, L., Healy, J., Melville, J., 2020b. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'.

Mead, A., 1992. Review of the development of multidimensional scaling methods. J. Royal Stat. Soc. Ser. D: The Stat. 41 (1), 27–39. https://doi.org/10.2307/2348634. Available at:

Mignotte, M., 2011. MDS-based multiresolution nonlinear dimensionality reduction model for color image segmentation. IEEE Trans. Neural Networks 22 (3), 447–460. https://doi.org/10.1109/TNN.2010.2101614. Available at:

Ozella, L., et al., 2023. A literature review of modeling approaches applied to data collected in automatic milking systems, 13(12), Available at: Animals https://doi.org/10.3390/ani13121916.

Rebuli, K.B., et al., 2023a. 'Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation periods'. Comput. Electron. Agric., 211, p. 108002. Available at: Doi: 10.1016/j.compag.2023.108002.

Rebuli, K.B., et al., 2023b. 'Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation periods'. Comput. Electron. Agric., 211, p. 108002. Available at: Doi: 10.1016/j.compag.2023.108002.

Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction, 290(5500) Available at: Science https://doi.org/10.1126/science.290.5500.2319.

Tullo, E., Finzi, A., Guarino, M., 2019. 'Review: environmental impact of livestock farming and precision livestock farming as a mitigation strategy. Available at: Sci. Total Environ. [Preprint] https://doi.org/10.1016/j.scitotenv.2018.10.018.

van der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE. Available at: J. Mach. Learning Res. 9 (86), 2579–2605 http://jmlr.org/papers/v9/vandermaaten08a.html.

Velliangiri, S., Alagumuthukrishnan, S., Thankumar Joseph, S.I., 2019. A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. 165, 104–111. https://doi.org/10.1016/j.procs.2020.01.079. Available at:

Witten, I.H., Frank, E., Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Wood, P.D.P., 1967. Algebraic model of the lactation curve in cattle. Nature 216 (5111), 164–165. https://doi.org/10.1038/216164a0. Available at:

Zhou, M., et al., 2022. Effects of increasing air temperature on physiological and productive responses of dairy cows at different relative humidity and air velocity levels. J. Dairy Sci. 105 (2), 1701–1716. https://doi.org/10.3168/jds.2021-21164. Available at: