137. Machine learning to predict somatic cell count at the subsequent test-day record in the Italian Mediterranean Buffaloes

T. Bobbo[1,2*], R. Matera[3], G. Pedota[4], J. Ramirez-Diaz[1,5], A. Manunza[1], A. Stella[1], A. Cotticelli[3], G. Neglia[3] and S. Biffani[1]

[1]Institute of Agricultural Biology and Biotechnology, National Research Council, Via Edoardo Bassini 15, 20133 Milan, Italy; [2]Department of Agricultural and Environmental Sciences, University of Milan, Via Celoria 2, 20133 Milan, Italy; [3]Department of Veterinary Medicine and Animal Production, Federico II University, Via Federico Delpino 1, 80137 Naples, Italy; [4]Associazione Regionale Allevatori della Basilicata, Via Dell'Edilizia, 85100 Potenza, Italy; [5]Department of Animal Science, Food and Technology, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy; tania.bobbo@unimi.it

Abstract

Collection of a large amount of data during the routine milk recording procedures makes it possible to train machine learning algorithms for predicting traits of interest. In this study, we applied two different machine learning methods (General Linear Models and Random Forest) to predict the udder health status of Italian Mediterranean Buffaloes at the subsequent monthly test-day record, using milk data collected at the previous test-day. Prediction accuracies of both methods were above 73%. According to different metrics, General Linear Models had the best performance in predicting the udder health classes (high or low somatic cell count, based on a threshold of 200,000 cells/ml). Findings of this study highlight the promising role of machine learning algorithms to improve the surveillance of subclinical mastitis, helping farmers in the preventive identification of animals that would possibly have high somatic cell count the subsequent month. Such information may also be exploited for breeding purposes.

Introduction

Subclinical mastitis, i.e. the inflammatory condition of the mammary gland with no clinical signs, is a costly health issues of the dairy Mediterranean Buffaloes (*Bubalus bubalis*) (Alterisio *et al.*, 2021). Efforts have been made to improve udder health, exploring also novel indicators of mammary gland inflammation based on traditional somatic cell count (scc) (Costa *et al.*, 2021). However, there is still need to improve mastitis detection, management and selection. Collection of a large amount of data in the frame of the routine milk recording procedures and by automatic milking recording systems makes it possible to exploit such information to train machine learning algorithms for predicting specific traits of interest. Machine learning analyses have already been applied in different areas of dairy research to support farmers in decision making (Cockburn, 2020). Previous studies have applied machine learning methods to predict subclinical mastitis in dairy cows, some based on the presence of high scc in milk (Bobbo *et al.*, 2021; Ebrahimi *et al.*, 2019), others based on the presence of mastitis pathogens (Hyde *et al.*, 2020; Sharif *et al.*, 2018). Nevertheless, only few studies reported predictions based on data collected at different time points (Anglart *et al.*, 2020; Bobbo *et al.*, 2021). Following the approach reported by Bobbo *et al.* (2021), in the present study we exploit information collected during the routine milk recording to predict the presence of high or low scc level in milk collected the subsequent month in Italian Mediterranean Buffaloes.

Materials & methods

Data collection and editing. Test-day (TD) records were provided by the Italian Breeders Association (Rome, Italy). Data included information about herd, animals (ID, stage of lactation and parity), daily milk production, milk composition, scc, β-hydroxybutyrate (bhb), electrical conductivity, and milk coagulation properties [rennet coagulation time (r) and curd firmness 30 minutes after rennet addition (a30)]. The dataset was edited to include animals that calved in the years 2019 and 2020, with at least two TD within

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_137, © T. Bobbo *et al.* 2022

lactation, and with less than 360 days in milk (dim). In addition, only consecutive TD records with an interval lower than six weeks were selected. Among milk traits, outliers beyond four standard deviations were considered as missing values and only full records were considered. Average daily milk production (milk_htd) and scc (scc_htd) of contemporary groups, i.e. animals sampled in the same herd and day, were also determined. Finally, the two scc-related traits were log-transformed to somatic cell score (scs and scs_htd). A total of 18 features were considered: parity (from 1 to ≥6), stage of lactation (dim30:12 classes of 30 d each), year and month of calving (ymc: 24 levels), year and month of sampling (yms: 19 levels), daily milk production, fat, protein, casein, lactose, pH, urea, scs, bhb, electrical conductivity (cond), milk_htd, scs_htd, and the two milk coagulation properties. Udder health status at the subsequent TD (outcome) was coded as a binary trait: low (≤200,000 cells/ml) or high (>200,000 cells/ml) scc. After editing, the dataset included 3,939 records of 1,035 buffaloes in six herds. In each record, information of two subsequent TD was provided: 18 features of the previous TD and outcome at the subsequent TD. The prevalence of subclinical mastitis (SCC>200,000 cells/ml) was 41.2%.

Data processing and models building. Machine learning analyses were performed following Bobbo *et al.* (2021). Prediction models were developed to identify the best udder health prediction method [Generalized Linear Models (GLM) or Random Forest (RF)] using information recorded at the previous TD. The dataset was randomly split into two subsets: 80% of the data (3,152 records) to train and test the models, and 20% of the data as a validation set. To automatically select the most predictive features, we applied a recursive feature selection using a 10-fold cross validation (cv) repeated 10 times with RF method. A stratified 10-fold cv repeated 10 times was then applied to train and test the models. The train/test dataset was randomly divided into 10 subsets; prediction models were trained on 9 subsets and tested on the last one. Each 10-fold cv was repeated 10 time, for a total of 100 iterations. Data standardization was performed within cv. Data analysis was performed using Caret v.6.0-86 (Kuhn, 2021) and Tidyverse v. 1.3.1 (Wickham *et al.*, 2019) packages of R software v.4.1.2 (R Core Team, 2021).

Evaluation of methods performance on testing and validation set. Sensitivity, specificity, accuracy, Cohen's kappa value, F1 score and area under the receiver operating characteristic curve (AUC) were used to compare the predictive performance of the models on the testing set. Feature importance based on the prediction model with the greatest accuracy was then assessed. Finally, predictive ability of both models on the validation set were estimated and compared using the same metrics reported above.

Results

Results of the recursive feature selection applied before model training suggested that the most parsimonious model (with the highest accuracy of prediction using the lowest number of features) was the one that included all the features of interest (Figure 1a).

Evaluation of methods performance in predicting udder health classes (low or high scc) on testing and validation sets was based on several metrics (Table 1). Whereas similar results were obtained in the testing set, GLM showed a slightly better performance in the validation set in comparison to RF. Nevertheless, in both methods, accuracy was above 73%, F1 above 78%, and AUC above 80%.

According to considered metrics, GLM was the best method in predicting udder health classes. Feature importance using GLM for predicting udder health status on testing set suggested that scs-derived traits were the most important features, followed by electrical conductivity and milk production (Figure 1b).

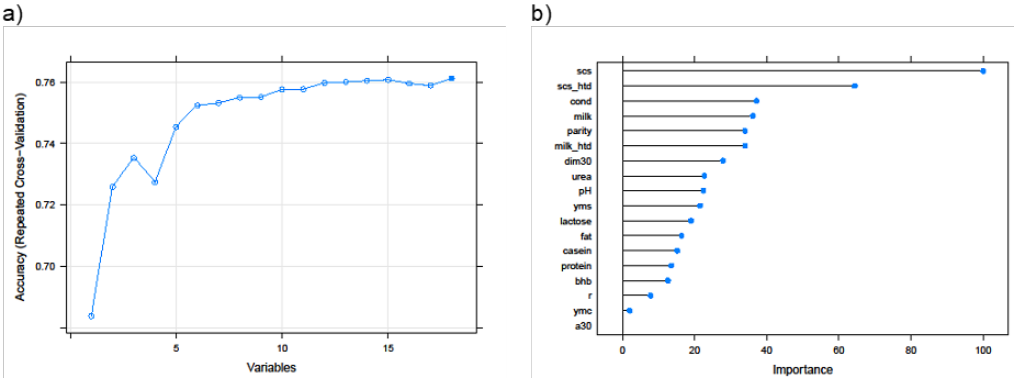

Figure 1. (a) Results of the recursive feature selection incorporating one to 18 features. (b) Plot of the feature importance showing the ranking for the prediction of udder health status in the testing set. Evaluated features, using Generalized Linear Models, are: individual somatic cell score and of contemporary group (scs and scs_htd), electrical conductivity (cond), milk production, parity, milk production of contemporary group (milk_htd), stage of lactation (dim30), urea, pH, year and month of sampling (yms), lactose, fat, casein, protein, β-hydroxybutyrate (bhb), rennet coagulation time (r), year and month of calving (ymc), and curd firmness 30 minutes after rennet addition (a30).

Table 1. Methods performance metrics [Sensitivity (Se), Specificity (Sp), Accuracy (Acc), Kappa value, F1 score, and area under the receiver operating characteristic curve (AUC)] on testing and validation sets.

Method	Testing set						Validation set					
	Se	Sp	Acc	Kappa	F1	AUC	Se	Sp	Acc	Kappa	F1	AUC
GLM[1]	0.84	0.62	0.75	0.47	0.80	0.82	0.84	0.62	0.75	0.47	0.80	0.81
RF[2]	0.83	0.64	0.75	0.47	0.79	0.82	0.81	0.63	0.73	0.44	0.78	0.80

[1] Generalized linear models.

[2] Random forest.

Discussion

In the present study, we applied two different machine learning methods (one linear and one based on decision trees) to predict the udder health status of Italian Mediterranean Buffaloes at the subsequent TD record, using already available milk data collected at the previous TD. The comparison of the performance of the two methods on both testing and external validation sets were based on several metrics. Besides accuracy, which is one of the most used metrics in classification problems, we included also F1 score and AUC, that have the advantage of being independent from outcome rate. Although GLM was the method with the best performance, similar results were obtained using RF. Findings of this study are in accordance with the results of Bobbo *et al.* (2021), who reported linear methods, RF and neural network to be the best methods for predicting the udder health status of Italian Holstein cattle. The scc-traits recorded at the previous TD were expected to be the most important features for making the prediction. The important role of electrical conductivity confirmed results of previous studies (Ebrahimi *et al.*, 2019, Ebrahimie *et al.*, 2018). In conclusion, machine learning methods applied to the large amount of data currently available represent a promising tool to improve the surveillance of subclinical mastitis, helping farmers in the preventive identification of animals that would possibly experience mammary gland issues. Further studies are required for a practical implementation of these methods in the management supporting systems as well as in the use of complex phenotypes in genetic and genomic evaluations.

Acknowledgements

This research was funded by Italian Ministry of Agriculture (MIPAAF – DISR 07). Project: 'BIG' Prot. N. 0215513 11/05/2021

References

Alterisio M.C., Ciaramella P., and Guccione J. (2021) Vet. Sci. 8(10):204. https://doi.org/10.3390/vetsci8100204

Anglart D., Hallén-Sandgren C., Emanuelson U., and Rönnegård L. (2020) J. Dairy Sci. 103(9):8433-8442. https://doi.org/10.3168/jds.2020-18320

Bobbo T., Biffani S., Taccioli C., Penasa M., and Cassandro, M. (2021) Sci. Rep. 11(1):1-10. | https://doi.org/10.1038/s41598-021-93056-4

Cockburn M. (2020) Animals 10(9):1690. https://doi.org/10.3390/ani10091690

Costa A., De Marchi M., Neglia G., Campanile G., and Penasa M. (2021) Ital. J. Anim. Sci. 20(1):548-558. https://doi.org/10.1080/1828051X.2021.1899856

Ebrahimi M., Mohammadi-Dehcheshmeh M., Ebrahimie E., and Petrovski K.R. (2019) Comput. Biol. Med. 114:103456. https://doi.org/10.1016/j.compbiomed.2019.103456

Ebrahimie E., Ebrahimi F., Ebrahimi M., Tomlinson S., and Petrovski K.R. (2018) J. Dairy Res. 85:193–200. https://doi.org/10.1017/S0022029918000249

Hyde R.M., Down P.M., Bradley A.J., Breen J.E., Hudson C., *et al.* (2020) Sci. Rep. 10(1):1-8. https://doi.org/10.1038/s41598-020-61126-8

Kuhn M. (2021). caret: Classification and Regression Training. R package version 6.0-90. Available at: https://CRAN.R-project.org/package=caret

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org

Sharif S., Pakdel A., Ebrahimi M., Reecy J.M., Fazeli Farsani S., *et al.* (2018) PloS one 13(2):e0191227. https://doi.org/10.1371/journal.pone.0191227

Wickham H., Averick M., Bryan J., Chang W., McGowan L.D.A., *et al.* (2019). J. Open Source Softw. 4(43): 1686, https://doi.org/10.21105/joss.01686