J. Dairy Sci. 102:9409–9421
https://doi.org/10.3168/jds.2019-16295

Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle

E. M. M. van der Heide,[1]* R. F. Veerkamp,[1] M. L. van Pelt,[2] C. Kamphuis,[3] I. Athanasiadis,[3] and B. J. Ducro[1]
[1]Wageningen University and Research Animal Breeding and Genomics, PO Box 338, 6700 AH Wageningen, the Netherlands
[2]Cooperation CRV, Animal Evaluation Unit, PO Box 454, 6800 AL Arnhem, the Netherlands
[3]Wageningen University and Research Information Technology Group, 6706 KN Wageningen, the Netherlands

ABSTRACT

In this study, we compared multiple logistic regression, a linear method, to naive Bayes and random forest, 2 nonlinear machine-learning methods. We used all 3 methods to predict individual survival to second lactation in dairy heifers. The data set used for prediction contained 6,847 heifers born between January 2012 and June 2013, and had known survival outcomes. Each animal had 50 genomic estimated breeding values available at birth and up to 65 phenotypic variables that accumulated over time. Survival was predicted at 5 moments in life: at birth, at 18 mo, at first calving, at 6 wk after first calving, and at 200 d after first calving. The data sets were randomly split into 70% training and 30% testing sets to evaluate model performance for 20-fold validation. The methods were compared for accuracy, sensitivity, specificity, area under the curve (AUC) value, contrasts between groups for the prediction outcomes, and increase in surviving animals in a practical scenario. At birth and 18 mo, all methods had overlapping performance; no method significantly outperformed the other. At first calving, 6 wk after first calving, and 200 d after first calving, random forest and naive Bayes had overlapping performance, and both machine-learning methods outperformed multiple logistic regression. Overall, naive Bayes has the highest average AUC at all decision points up to 200 d after first calving. Random forest had the highest AUC at 200 d after first calving. All methods obtained similar increases in survival in the practical scenario. Despite this, the methods appeared to predict the survival of individual heifers differently. All methods improved over time, but the changes in mean model outcomes for surviving and non-surviving animals differed by method. Furthermore, the correlations of individual predictions between methods ranged from r = 0.417 to r = 0.700; the lowest correlations were at first calving for all methods. In short, all 3 methods were able to predict survival at a population level, because all methods improved survival in a practical scenario. However, depending on the method used, predictions for individual animals were quite different between methods.

Key words: machine learning, naive Bayes, regression, random forest, phenotypic prediction

INTRODUCTION

Machine learning, an invention from the field of computer science originally intended to mimic human intelligence (Michalski et al., 2013), has become a valuable tool for prediction in many fields. Machine-learning methods are versatile, because they can derive a model from available data without previous knowledge of the relations between variables (McQueen et al., 1995; Kotsiantis et al., 2007). Machine-learning methods thrive on large data sets and make fewer assumptions about the data, allowing them to make use of non-normally distributed variables (Gahegan, 2003; Gianola et al., 2011). In dairy science, machine learning has been used successfully to predict a whole range of different traits, such as mastitis (Kamphuis et al., 2010; Ebrahimie et al., 2018), methane production (Zheng et al., 2016), and milk production (Gianola et al., 2011). However, despite the advantages of machine learning, other recent studies have used traditional linear methods to predict disease risk (Moretti et al., 2017), methane production (Engelke et al., 2018), and milk production (Wallén et al., 2018).

One of the reasons for the continued use of more traditional methods such as regression is that despite the potential of machine learning, it has not always proven superior to traditional linear modeling (Cortez et al., 2006; Van Hertem et al., 2014; Hempstalk et al., 2015; Ghafouri-Kesbi et al., 2017). As well, comparisons between machine learning and traditional methods may not be possible in some cases: for example, data sets with missing records, which can be used by some

Received January 14, 2019.
Accepted June 17, 2019.
*Corresponding author: esther.vanderheide@wur.nl

machine-learning methods but not by regression (Bennett, 2001), or data sets that contain video data (Kabra et al., 2013). Furthermore, it is difficult to determine beforehand which method will result in the highest accuracy for a particular prediction problem (White et al., 2018), because in practice many different machine-learning techniques may be suitable for predicting a variable of interest. This challenge results in a trial-and-error approach to finding the best method for each prediction problem (Amrine et al., 2014; Libbrecht and Noble, 2015).

Two machine-learning methods that use very different approaches but are applied competitively in the field of animal science, are naive Bayes and random forest. Naive Bayes is a family of classifiers that implements Bayesian techniques to form a simple network based on previous probabilities (Jensen, 1996). The naive Bayes method relies on independence between the input variables, but it performs surprisingly well even under conditions that might be considered suboptimal for the algorithm (Domingos and Pazzani, 1997; Friedman, 1997). Despite the relative simplicity of its algorithm, naive Bayes is still widely used (Jensen et al., 2016; Drury et al., 2017). Random forest (Breiman, 2001) is another machine-learning method that is successfully implemented in a wide variety of fields, including animal science (Shahinfar et al., 2014; Machado et al., 2015; Brieuc et al., 2018). This regression or classification method makes use of decision trees: a sequence of rules that split the data in a way that most optimally reduces variation. Each tree receives a random subset of training samples, and then the algorithm randomly selects a subset of variables at each split in the tree (Breiman, 2001). These trees, which are relatively poor classifiers individually, are combined into an ensemble of trees called a random forest, which is used for prediction. The prediction results of a random forest are a summation of the prediction outcomes of many individual trees.

The aim of this study was to compare the traditional linear method of regression with the machine-learning methods naive Bayes and random forest. By discovering the advantages and disadvantages of each technique in a dairy cattle case study predicting longevity, we hoped to gain a better understanding of the wide variety of available tools for predicting complex biological traits.

MATERIALS AND METHODS

Data

The data used in this study were identical to data described in a previous study (E. van der Heide, R. Veerkamp, M. van Pelt, C. Kamphuis, and B. J. Ducro, unpublished data). The data set consisted of records on 6,847 heifers born between January 2012 and June 2013, from 463 farms participating in a data-collection program that required the farmer to genotype all female heifers at birth. Each heifer was herd-book-registered and at least 87.5% Holstein. Survival was a binary classifying variable "survival until second calving plus 2 wk," including 2 additional wk to exclude heifers that died or were culled as a direct consequence of second calving. To have a known outcome for survival, all included cows had to be born at least 46 mo before the end of data collection and were not exported abroad. In this data set, 85.8% of heifer calves reached second lactation.

The heifers had records on 50 genomically estimated breeding values, standardized to values between 1 and 10. The genomic breeding values were calculated from the genomic test results by cattle breeding cooperative CRV. These genomic breeding values included only the direct genomic values and did not include performance records. The records also included up to 65 phenotypic variables, including on birth weight and gestation length, insemination records up to second parity, first-parity calving records, and first-parity lactation information (see also Appendix Table A1). Not all animals had records for every variable: some were missing, and some were not collected because the heifer died before the information could be collected. All phenotypic continuous variables were transformed into factors of at least 5 levels, containing a factor level for missing information so that animals with missing records could be used in the regression analysis. At the cost of losing some of the original information, changing the continuous variables into factorial variables allowed us to include all animals in the analyses, regardless of method. From the complete data, we created 5 data sets that contained all information available at 5 distinct moments in the life of a dairy cow. These distinct "decision moments" were points at which new information became available, and when a management decision could be aided by predicting the expected survival of the animal. The decision moments were birth, 18 mo of age, first calving, 6 wk after first calving, and 200 d after first calving. At the first decision moment, genomic information and only limited phenotypic information were available, but at the last decision moment all information was available. Appendix Table A1 shows all available variables and the decision moments at which they were available. We chose the decision moments to investigate the ability of each model to predict survival at various points in the life of a cow. Early prediction was preferred, but because very little information was available early on, this was not feasible. Only animals still alive at the beginning of a decision moment were used in the analysis.

Model and Analysis

The analyses were performed in the statistical program R version 3.3.1 (R Core Team 2016), using the packages “caret” for logistic multiple regression (Kuhn, 2008), “randomForest” for the random forest approach (Liaw and Wiener, 2002), and “naivebayes” for the naive Bayes method (Majka, 2018). Because the randomForest package was unable to make predictions on factorial data exceeding 53 levels, we transformed farm identification number variables (such as birth farm, farm of first calving, and farm of first milking) into sets of dummy variables. No other changes were made to the data, to ensure that the 5 distinct data sets (1 for each distinct decision moment) presented to each model were identical as possible. Our linear method was logistic multiple regression, using the Akaike information criterion in forward stepwise selection to determine the best possible model at each decision moment. The selected regression models included phenotypic variables at the different decision moments, and a set of 6 genomic EBV (**gEBV**) related to longevity, fertility, feet and leg score, conformation score, udder score, and udder health. We did not include production gEBV because the lifespan gEBV was uncorrected for production. Because the Bayesian method we selected was naive Bayes, we did not need to set a priori values for the variables. For the random forest, the number of trees was set to 500 and the number of variables selected at each split was set to the square root of the number of variables available [ranging from 6 to 12 for our data sets, because these values were recommended and gave the highest area under the curve (**AUC**) values]. All 3 methods were trained on a random selection of 70% of each of the 5 data sets, and then validated using the remaining 30%. We repeated this process 20 times for all 5 decision moments and methods.

Unbalanced response variables are challenging to both linear and machine-learning techniques (Kotsiantis and Pintelas, 2003), so we adapted both the regression model and the random forest model to predict the unbalanced response variable survival (85.8% survivors vs. 14.2% non-survivors in the data). For random forest, we tested 3 adaptation methods: stratified sampling, changing the voting rule (or cutoff) of the model, and adding weights to the underrepresented class. Stratified sampling was chosen for further analysis because this method provided the highest AUC value on a single trial run (results not shown). Stratified sampling meant that the model would sample from the training data until it obtained an equal number of samples of both classes. This meant that in a given validation run, not all surviving animals would be used, and non-surviving animals could be included multiple times. For the regression method, the cutoff had to be specified manually. In this case, we chose the random chance of survival of an animal in the data set: 0.858. Animals that received a predicted probability of survival equal or above this cutoff were predicted to be survivors, and animals that scored below this cutoff were predicted to be non-survivors. Naive Bayes has been reported to have issues only with extremely unbalanced data (predictor class of interest occurring in 1% or less of the cases; Domingos and Pazzani, 1997), so no changes were made for this study, because the class imbalance was not extreme.

The performance of the methods was evaluated by measuring the following: contrasts between the mean probabilities of survival for both survival groups; accuracy; sensitivity; and specificity. Contrasts were the differences between the means of the 2 groups, expressed in units of standard deviation, and they allowed us to compare the model outputs for the 2 groups across methods. Accuracy was the proportion of correctly predicted animals; sensitivity was the proportion of surviving heifers correctly predicted to survive; and specificity was the proportion of non-surviving heifers correctly predicted not to survive. However, because survival to second lactation was an unbalanced trait, using accuracy as an indicator of superiority could have been biased. Therefore, we also evaluated the performance of the models by determining the area under the receiver operating characteristic curve (AUC) value using the R package “pROC” (Robin et al., 2011). The AUC metric measured the performance of the methods over the full range of specificities and sensitivities and was not affected by the tradeoff between specificity and accuracy. Finally, we tested a scenario in which only the heifers with scores in the top 50% were kept on a farm at a specific decision moment. We considered this latter evaluation approach an example of how the models could be used in practice as part of a decision-support system.

Because consistency across methods was also of interest, we looked at the correlations between the methods. All heifers from the testing set had 3 predicted probabilities of survival, one for each method. We calculated the Pearson’s r and Spearman’s ρ between all 3 methods for all 5 decision moments (Chok, 2010), and obtained averages for these correlations over the 20 validation runs. We did this because not only was the similarity of the predicted probabilities between the methods important, but also the assigned rank of the heifer in the

Table 1. Contrasts between the mean probability of survival for surviving and non-surviving heifers

Item	Regression	Naive Bayes	Random forest
Birth	0.279	0.327	0.231
18 mo	0.409	0.446	0.331
First calving	0.525	0.435	0.393
6 wk after first calving	0.583	0.554	0.494
200 d after first calving	0.800	0.606	0.747

group; in practice, the decision to cull an animal would be based on rank.

RESULTS

Table 1 shows the contrasts between the average predicted probabilities of surviving and non-surviving heifers. All contrasts were positive and increasing, which means that the average predicted probability of survival was always higher for the surviving group, and increased over time. This indicated that the model could predict survival at least at a population level, and that all model performances increased with additional information, regardless of method. At birth and at 18 mo, naive Bayes showed the highest contrast between the 2 groups; after first calving, regression showed the highest contrasts.

Naive Bayes outperformed regression and random forest for the first 2 decision moments in terms of accuracy (Figure 1). After the second decision moment, regression and naive Bayes alternated in terms of best performance for the third, fourth, and fifth decision moments, respectively. Of the 3 methods, naive Bayes had the highest sensitivity at birth and 18 mo, but the lowest specificity at the first 2 decision moments (Figure 2). Regression had the lowest sensitivity, but also the highest specificity at the first 2 decision moments, and random forest had intermediate scores. There appeared to be a tradeoff between high specificity and high accuracy. Because in practice a smaller proportion of heifers are non-survivors, negative predictions were significantly less likely to be true than positive predictions (i.e., for a random heifer from the data, its odds of surviving were higher than its odds of not surviving). If the model could very accurately predict non-surviving heifers, this would not be a problem. However, because survival is a complex trait and difficult to predict, models that made fewer predictions of non-survival were more likely to be correct by chance. Thus, a model with high specificity made more negative predictions, resulting in a loss of accuracy. The AUC was not biased by this tradeoff, because it considered all possible specificity and sensitivity values (Figure 3). At birth and 18 mo, all methods had overlapping performance; no method significantly outperformed the other. At first calving, 6 wk after first calving, and 200 d after first calving, random forest and naive Bayes had overlapping performance: both machine-learning methods outperformed multiple logistic regression. Overall, naive Bayes had the highest average AUC at all decision moments up to 200 d after first calving. At 200 d after first calving, random forest had the highest AUC. All methods performed significantly better than AUC 0.5 (random chance) at all decision moments, indicating that even at birth it was possible to some extent to predict survival to second lactation. In a practical scenario that selected 50% of the heifer calves with the highest probability of survival (Figure 4), all methods performed similarly. Again, naive Bayes scored highest in the first decision moments, before being outperformed by regression in the fourth and fifth decision moments, but the differences in additional survival realized were marginal. All 3 methods resulted in increased survival compared with a random selection of heifers for every decision moment.

All methods could predict survival and improve over time, but they did not make identical predictions. In all cases, the means of the surviving and non-surviving groups moved apart over time (Figures 5, 6, and 7), although the groups always overlapped. Using regression, the mean predicted probability of survival for surviving heifers increased, but the mean for non-surviving heifers remained stable (Figure 5). We had the opposite finding for naive Bayes; the mean for non-surviving heifers decreased and the mean for surviving heifers remained stable (Figure 6). Naive Bayes also had the largest standard deviation, because it classified cases closer to 0 or 1 than the other methods, making it more sensitive to data partitioning. Random forest had an intermediate approach: the mean model output for surviving heifers increased and the mean for non-surviving heifer decreased slightly (Figure 7). The random forest method was centered on 0.50 because of the stratified sampling, whereas the mean for the other 2 methods was closer to 0.86, random chance of survival. The differences in approach between the methods were also reflected in the correlations between the predictions on the same set of animals. Correlations were always positive, indicating that high scores or ranks in one method also indicated high scores or ranks in another, as ex-

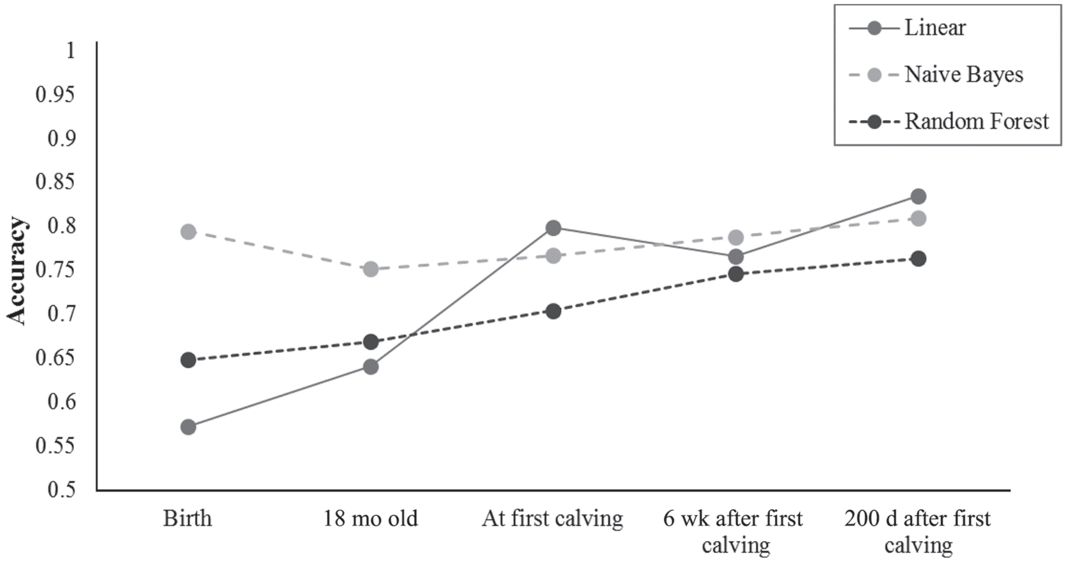

Figure 1. Accuracy of prediction of the regression (linear), naive Bayes, and random forest models.

pected. Spearman's ρ was generally higher than Pearson's r. At birth, the correlations between all methods were moderate to high (Table 2), and ranged from r = 0.653 between naive Bayes and random forest to r = 0.539 between regression and naive Bayes. Correlations decreased toward first calving, and all correlations were lowest at first calving (from r = 0.557 between regression and random forest to r = 0.417 between regression and naive Bayes). Overall, correlations ranged from moderate to high (0.4 to 0.7) and were consistently lowest between naive Bayes and regression. Figure 8 gives an example of the correlations in 1 of the 20 validation runs. Note that naive Bayes had a different distribution of predicted probability of survival than the other 2 methods, favoring predicted probabilities close to 1 or 0, and the other 2 methods favored predicted probabilities closer to their respective mean predicted probabilities of survival. This was in part because of the different methods chosen to deal with the class imbalance issue. Over time, the predicted probabilities moved further apart, with more predictions moving closer to 1 or 0. This was more visible for the lower scores, because we found fewer low scores, and because for regression and random forest, high scores begin to approach 1, but low

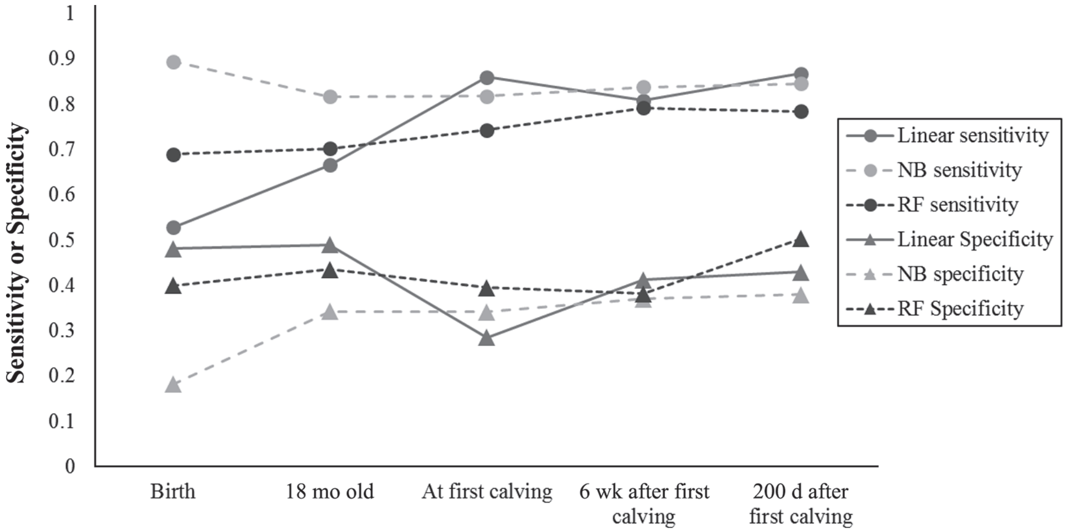

Figure 2. Sensitivity and specificity of the regression (linear), naive Bayes (NB), and random forest (RF) models. This figure shows the balance between sensitivity (lines with circles) and specificity (lines with triangles).

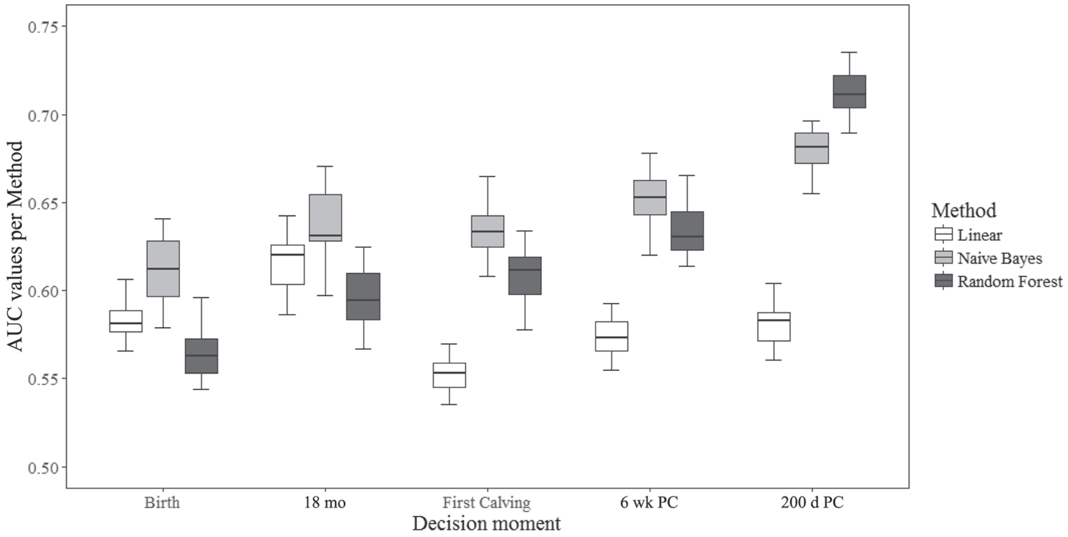

Figure 3. Area under the curve (AUC) of multiple logistic regression (linear), naive Bayes, and random forest methods at the 3 decision moments. Outliers were removed. The box of the boxplot indicates the first quartile, mean, and third quartile borders, and the whiskers show the highest and lowest values found. PC = postcalving.

scores do not approach 0 to the same extent. In general, the trend was that high-scoring heifers from one method also scored high with the other methods, but as we expected because the correlations were medium to high, we also found exceptions, in which a heifer scored very differently between methods.

DISCUSSION

We showed that regression, naive Bayes, and random forest could predict survival to second lactation for dairy cows at a population level. Naive Bayes had the highest AUC value at all decision moments except 200 d after calving, although performance overlapped with random forest at all decision moments. Logistic multiple regression performed similarly to naive Bayes and random forest for the first 2 decision moments, but was outperformed at first calving, 6 wk after first calving, and at 200 d after calving. All methods were significantly different from an AUC of 0.5, but in general AUC values were low; only random forest achieved an average accuracy above 0.7 in the last decision moment. The use of AUC has some limitations as a metric for evaluating methods (Lobo et al., 2008), but in general

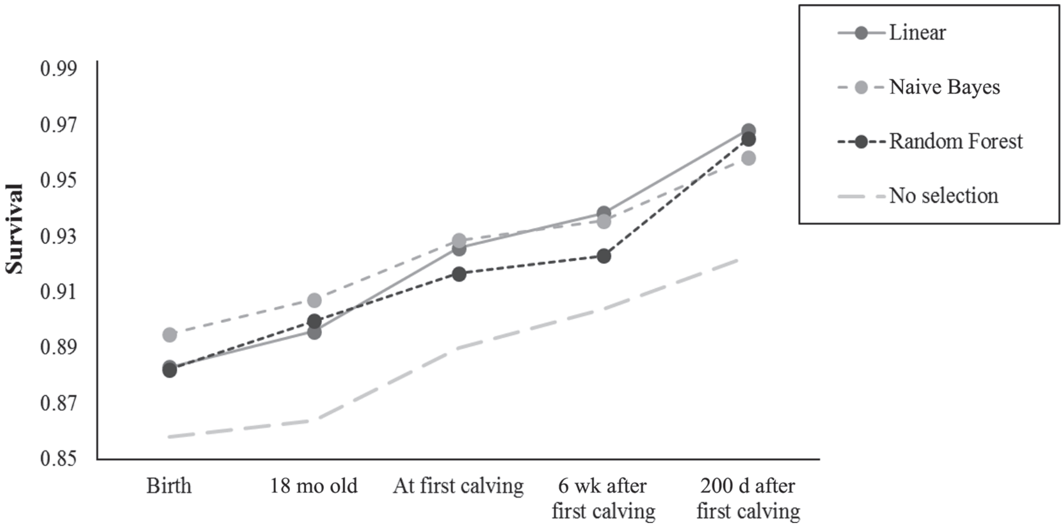

Figure 4. Surviving proportion of the heifer population when selecting 50% of the highest-scoring heifers using regression (linear), naive Bayes, and random forest models.

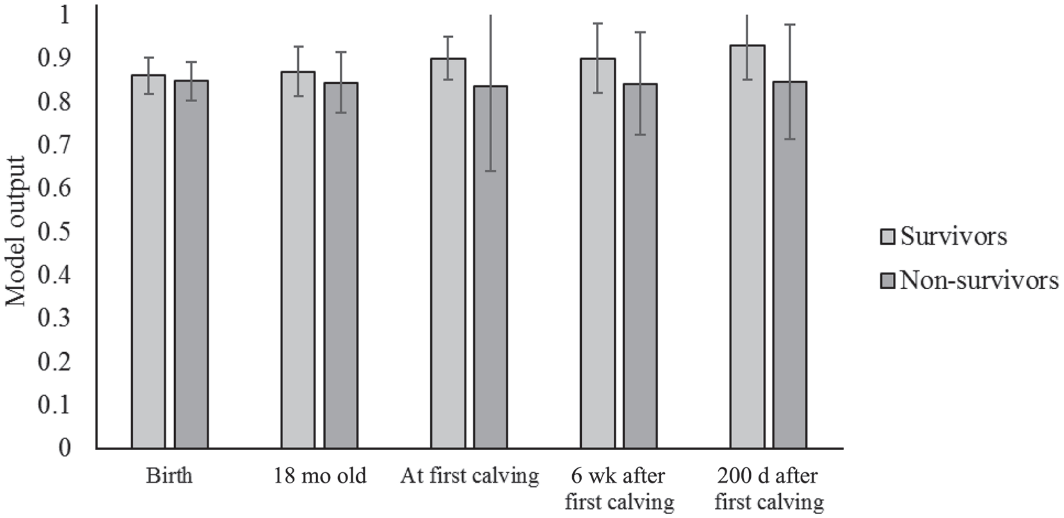

Figure 5. Mean model output of the multiple logistic regression method for surviving and non-surviving animals at all 5 decision moments. Error bars represent standard deviation.

a model with an AUC score of 0.9 indicates a good accuracy, an AUC score between 0.7 and 0.9 indicates moderate accuracy, and an AUC score between 0.5 and 0.7 indicates low accuracy (Akobeng, 2007). In short, although it was possible to determine which method had the best performance at each decision moment, none of these methods was able to accurately predict individual cow survival. Accurate individual predictions are important for the practical application of a model, because a farmer is interested in the accuracy of prediction for a single animal or small group of animals, not the average outcome success for all Dutch Holsteins.

In the literature, cow pregnancy status was similar to cow survival, because they are both complex traits with binary outcomes. Furthermore, survival and fertility are both genetically and phenotypically related (Pritchard et al., 2013), because fertility issues are among the main reasons for culling a cow (Brickell and Wathes, 2011; Zijlstra et al., 2013). Several studies have compared linear and machine-learning methods for the prediction of insemination outcome or cow pregnancy status (Shahinfar et al., 2014; Hempstalk et al., 2015; Fenlon et al., 2016), and the results of these papers were similar to our study: naive Bayes performed well

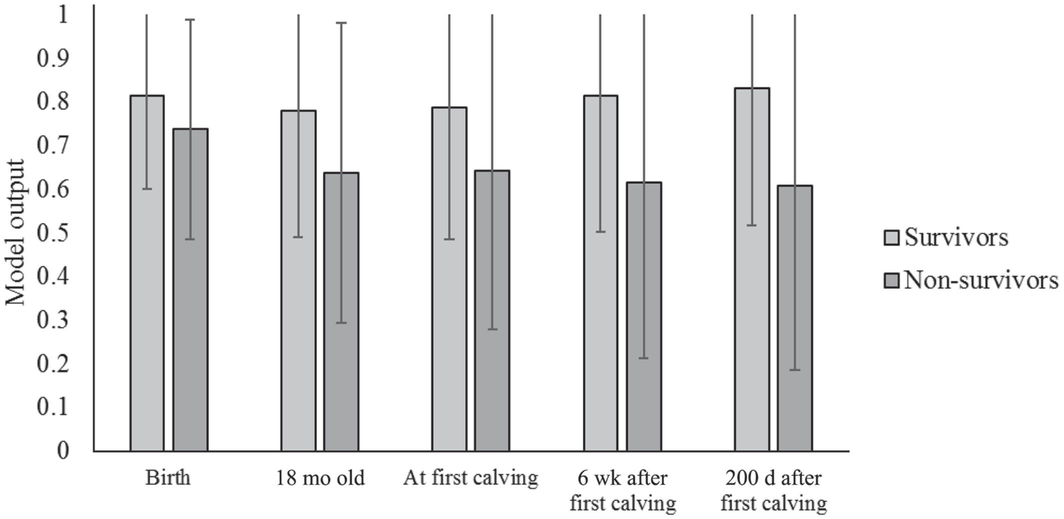

Figure 6. Mean model output of the naive Bayes method for surviving and non-surviving animals at all 5 decision moments. Error bars represent standard deviation. The error bars were exceptionally large because naive Bayes attempts to classify cases closer to 0 or 1 than the other methods.

Table 2. Pearson and Spearman correlation coefficients between the model output of the regression (R), naive Bayes (NB), and random forest (RF) models, averaged over 20 runs

	Pearson correlation			Spearman correlation		
Item	R–NB	R–RF	RF–NB	R–NB	R–RF	RF–NB
Birth	0.539	0.627	0.653	0.564	0.623	0.714
18 mo	0.566	0.666	0.686	0.573	0.692	0.710
First calving	0.417	0.547	0.557	0.429	0.601	0.606
6 wk after first calving	0.560	0.632	0.700	0.532	0.626	0.688
200 d after first calving	0.488	0.694	0.578	0.488	0.732	0.621

when few data were available. Contrary to our results, however, regression outperformed machine-learning methods such as naive Bayes, support vector machines, and random forest in other studies (Hempstalk et al., 2015; Fenlon et al., 2016). Another study that investigated no linear methods found random forest to be superior to Bayesian methods (Shahinfar et al., 2014). When looking closely at these studies, however, the performance of the methods they tested was very similar. Furthermore, none of the methods tested could predict individual pregnancy outcomes well enough to be of use in practice (Hempstalk et al., 2015; Fenlon et al., 2016; Rutten et al., 2016). So although insight was gained into the best methods and mechanisms for predicting a complex trait, ultimately none of the prediction methods used in our study or in previous research is useful for individual prediction of a complex binary trait.

We selected 3 different methods for evaluation: multiple logistic regression, naive Bayes, and random forest. We selected these methods because they each represented large groups of similar methods, but many other methods can predict survival. A linear method that we did not include in our study was survival analysis, a method commonly used for the genetic evaluation of survival traits (Cox and Oates, 1984; Ducrocq et al., 1988). This method is used instead of regression because it can use uncensored records, giving it an advantage (Carlén et al., 2005). Because the data we used in the present study were already censored, survival analysis was not necessary to make optimal use of the data. Another possibility would be to investigate more advanced machine-learning methods such as neural networks. Indeed, neural networks outperformed regression and random forest for the individual prediction of pregnancy status (Fenlon et al., 2017). Neural networks are powerful but complex methods that often require a large number of records to be trained. Although neural networks would have been difficult to apply with our current data set, they may be of use in future research.

Accurate individual prediction of survival to second lactation cannot be achieved by optimizing the choice of prediction method alone; in future research, other ways to improve prediction performance should be considered. Prediction performance may be improved by

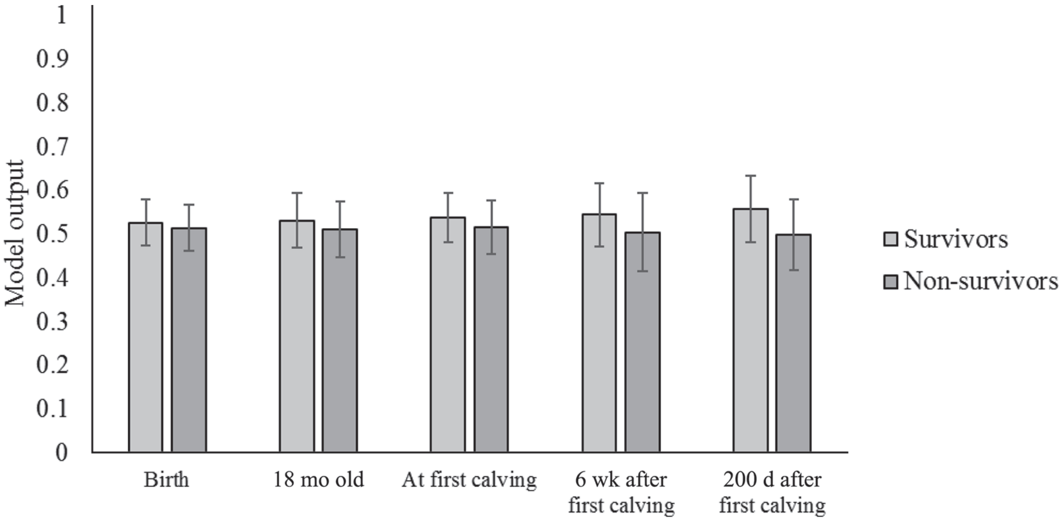

Figure 7. Mean model output of the random forest method for surviving and non-surviving animals at all 5 decision moments. Error bars represent standard deviation. The mean for the random forest model output was set to 0.5 because of stratified sampling.

increasing the number of records. In the case of pregnancy status, results did improve with a large number of additional records (Shahinfar et al., 2014). With over 200,000 records, the AUC reported was 0.76, but the accuracy was still only 72 to 74%, meaning that a full quarter of the animals would be incorrectly classified in practice. In addition to increasing the number of records, increasing the number of available variables may also be necessary to improve accuracy enough for individual prediction. For example, this study lacked some variables known to be relevant for individual survival, such as animal growth, health, housing, and other farm-management factors (Wathes et al., 2008; Brickell and Wathes, 2011). We chose the variables in this study because they were readily available on most Dutch farms. In contrast, information on animal health and growth are often not available and require additional data collection and cost. Finally, although additional records and variables may improve prediction accuracy, additional information will not solve all difficulties. A model can predict a non-surviving animal accurately only when the cause of death is the result of a pattern found in the data. This requirement is problematic; for example, because our study lacked calf health variables,

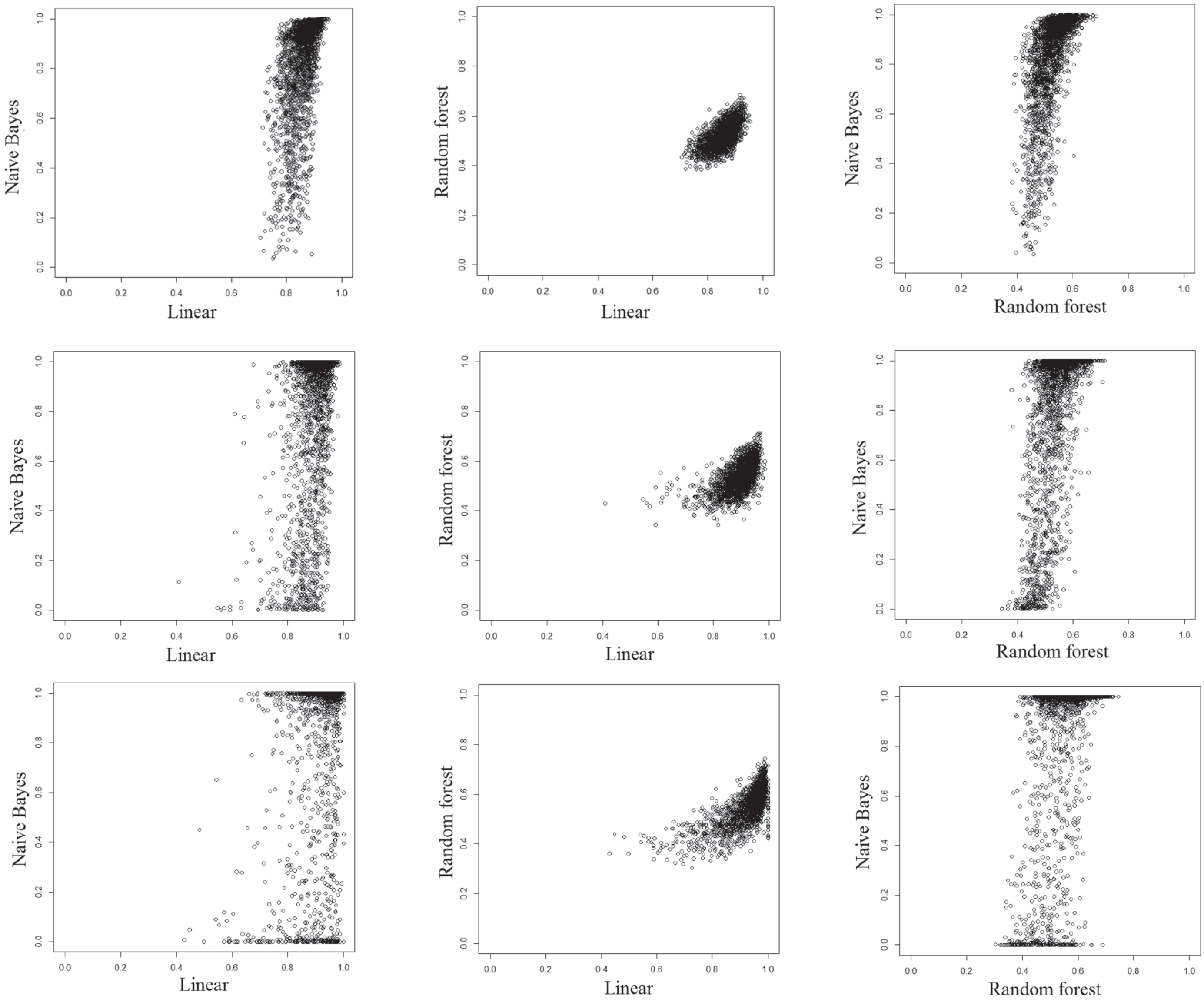

Figure 8. Visualization of the correlations between the 3 methods of 1 of the 20 validation runs. Plotted are the model output values (between 0 and 1) for all 3 methods. The first row depicts correlations at birth, the second row shows correlations at first calving, and the third row shows correlations at 200 d after calving. The first column shows the regression method versus the naive Bayes method, the second column shows the regression method versus the random forest method, and the third column shows the random forest method versus the naive Bayes method.

calf deaths due to illness would be difficult to predict correctly. Furthermore, not all causes of death follow identifiable patterns. Some deaths may be caused by unpredictable accidents (Brickell and Wathes, 2011), or based on individual farmer's decisions, which cannot always be explained by the available information (Hadley et al., 2006; Huijps et al., 2010). Thus although we expect that additional information will increase accuracy, a certain degree of uncertainty will remain.

CONCLUSIONS

All 3 methods (logistic multiple regression, naive Bayes, and random forest) were able to predict survival at a population level. At birth and at 18 mo, all 3 methods reported similar AUC values and increased survival in a practical scenario by similar amounts. Naive Bayes obtained the highest AUC value in all decision moments up to 200 d after calving, but it always overlapped with random forest. At 200 d after first calving, random forest had the highest AUC, but the overlap with naive Bayes persisted. Interestingly, the 3 methods appeared to predict outcomes for individual heifers differently. Correlations between individual predictions for animals were lower than expected, and the models appeared to improve by predicting different groups of animals better. It was possible to choose a "best method" for each moment, but all methods would have resulted in similar improvements in practice.

ACKNOWLEDGMENTS

This work was part of the Breed4Food research program, project "Smart animal breeding with advanced machine learning," with project number 14295, financed by the Netherlands Organization for Scientific Research (NWO), the Dutch Ministry of Economic Affairs (TKI Agri & Food project 12018), and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The data for this study were provided by the cattle improvement cooperative CRV (Arnhem, the Netherlands).

REFERENCES

Akobeng, A. K. 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr. 96:644–647.

Amrine, D. E., B. J. White, and R. L. Larson. 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Comput. Electron. Agric. 105:9–19.

Bennett, D. A. 2001. How can I deal with missing data in my study? Aust. N. Z. J. Public Health 25:464–469.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32.

Brickell, J. S., and D. C. Wathes. 2011. A descriptive study of the survival of Holstein-Friesian heifers through to third calving on English dairy farms. J. Dairy Sci. 94:1831–1838.

Brieuc, M. S., C. D. Waters, D. P. Drinan, and K. A. Naish. 2018. A practical introduction to random forest for genetic association studies in ecology and evolution. Mol. Ecol. Resour. 18: 755–766.

Carlén, E., M. P. Schneider, and E. Strandberg. 2005. Comparison between linear models and survival analysis for genetic evaluation of clinical mastitis in dairy cattle. J. Dairy Sci. 88:797–803.

Chok, N. S. 2010. Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data, University of Pittsburgh, PA.

Cortez, P., M. Portelinha, S. Rodrigues, V. Cadavez, and A. Teixeira. 2006. Lamb meat quality assessment by support vector machines. Neural Process. Lett. 24:41–51.

Cox, D. R., and D. Oates. 1984. Analysis of Survival Data. Chapman and Hall, Stockholm, Sweden.

Domingos, P., and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29:103–130.

Drury, B., J. Valverde-Rebaza, M.-F. Moura, and A. de Andrade Lopes. 2017. A survey of the applications of Bayesian networks in agriculture. Eng. Appl. Artif. Intell. 65:29–42.

Ducrocq, V., R. L. Quaas, E. J. Pollak, and G. Casella. 1988. Length of productive life of dairy cows. 1. Justification of a Weibull model. J. Dairy Sci. 71:3061–3070.

Ebrahimie, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski. 2018. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. Comput. Electron. Agric. 147:6–11.

Engelke, S. W., G. Daş, M. Derno, A. Tuchscherer, W. Berg, B. Kuhla, and C. C. Metges. 2018. Milk fatty acids estimated by mid-infrared spectroscopy and milk yield can predict methane emissions in dairy cows. Agron. Sustain. Dev. 38:27.

Fenlon, C., L. O'Grady, J. Dunnion, L. Shalloo, S. Butler, and M. Doherty. 2016. A comparison of machine learning techniques for predicting insemination outcome in Irish dairy cows. Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland. Teagasc, Carlow, Ireland.

Fenlon, C., L. O'Grady, J. F. Mee, S. T. Butler, M. L. Doherty, and J. Dunnion. 2017. A comparison of 4 predictive models of calving assistance and difficulty in dairy heifers and cows. J. Dairy Sci. 100:9746–9758.

Friedman, J. H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data Min. Knowl. Discov. 1:55–77.

Gahegan, M. 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg)? Int. J. Geogr. Inf. Sci. 17:69–92.

Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar, and A. Nejati-Javaremi. 2017. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. Anim. Prod. Sci. 57:229–236.

Gianola, D., H. Okut, K. A. Weigel, and G. J. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. BMC Genet. 12:87.

Hadley, G. L., C. A. Wolf, and S. B. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. J. Dairy Sci. 89:2286–2296.

Hempstalk, K., S. McParland, and D. Berry. 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. J. Dairy Sci. 98:5262–5273.

Huijps, K., H. Hogeveen, G. Antonides, N. I. Valeeva, T. J. Lam, and A. G. Oude Lansink. 2010. Sub-optimal economic behaviour with respect to mastitis management. Eur. Rev. Agric. Econ. 37:553–568.

Jensen, D. B., H. Hogeveen, and A. De Vries. 2016. Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis. J. Dairy Sci. 99:7344–7361.

Jensen, F. V. 1996. An Introduction to Bayesian Networks. UCL Press, London, UK.

Kabra, M., A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. 2013. JAABA: Interactive machine learning for automatic annotation of animal behavior. Nat. Methods 10:64.

Kamphuis, C., H. Mollenhorst, J. Heesterbeek, and H. Hogeveen. 2010. Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction. J. Dairy Sci. 93:3616–3627.

Kotsiantis, S., and P. Pintelas. 2003. Mixture of expert agents for handling imbalanced data sets. Ann Math Computing Teleinformatics 1:46–55.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. In Emerging Artificial Intelligence Applications in Computer Engineering. IOS Press, Amsterdam, the Netherlands.

Kuhn, M. 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28:1–26.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Libbrecht, M. W., and W. S. Noble. 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16:321.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17:145–151.

Machado, G., M. R. Mendoza, and L. G. Corbellini. 2015. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet. Res. 46:85.

Majka, M. 2018. naivebayes: High performance implementation of the Naive Bayes algorithm. R package version 0.9.2. Accessed Mar. 11, 2018. https://CRAN.project.org/package=naivebayes.

McQueen, R. J., S. R. Garner, C. G. Nevill-Manning, and I. H. Witten. 1995. Applying machine learning to agricultural data. Comput. Electron. Agric. 12:275–293.

Michalski, R. S., J. G. Carbonell, and T. M. Mitchell. 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media, Berlin, Germany.

Moretti, R., S. Biffani, F. Tiezzi, C. Maltecca, S. Chessa, and R. Bozzi. 2017. Rumination time as a potential predictor of common diseases in high-productive Holstein dairy cows. J. Dairy Res. 84:385–390.

Pritchard, T., M. Coffey, R. Mrode, and E. Wall. 2013. Genetic parameters for production, health, fertility and longevity traits in dairy cows. Animal 7:34–46.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.

Rutten, C. J., W. Steeneveld, J. C. M. Vernooij, K. Huijps, M. Nielen, and H. Hogeveen. 2016. A prognostic model to predict the success of artificial insemination in dairy cows based on readily available data. J. Dairy Sci. 99:6764–6779.

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

Van Hertem, T., S. Viazzi, and M. Steensels., EMaltz, A., V. Antler, A. A. Alchanatis, K. Schlageter-Tello, E. C. Lokhorst, Romanini, and C. Bahr. 2014. Automatic lameness detection based on consecutive 3D-video recordings. Biosyst. Eng. 119:108–116.

Wallén, S. E., E. Prestløkken, T. H. E. Meuwissen, S. McParland, and D. P. Berry. 2018. Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. J. Dairy Sci. 101:6232–6243.

Wathes, D., J. Brickell, N. Bourne, A. Swali, and Z. Cheng. 2008. Factors influencing heifer survival and fertility on commercial dairy farms. Animal 2:1135–1143.

White, B. J., D. E. Amrine, and R. L. Larson. 2018. Big data analytics and precision animal agriculture symposium: Data to decisions. J. Anim. Sci. 96:1531–1539.

Zheng, H., H. Wang, and T. Yan. 2016. Modelling enteric methane emissions from milking dairy cows with Bayesian networks. Pages 1635–1640 in Bioinformatics and Biomedicine (BIBM), 2016 IEEE Int. Conf., Shenzhen, China. IEEE, New York, NY.

Zijlstra, J., M. Boer, J. Buiting, K. Colombijn-Van der Wende, and E.-A. Andringa. 2013. Rapport 668: Routekaart Levensduur; Eindrapportage van het project "Verlenging levensduur melkvee". Wageningen UR Livestock Research, Wageningen, the Netherlands.

APPENDIX

Table A1. All 66 phenotypic variables and 50 genomic EBV (gEBV) available and the decision moment in which each variable is available

Phenotypic variable	Decision moment	Continuous
Animal identification number	Birth	No
Year of birth	Birth	No
Birth farm UBN[1]	Birth	No
Month of birth	Birth	No
Birth season	Birth	No
Parity	Birth	Yes
Breed	Birth	No
Holstein, %	Birth	No
Red factor	Birth	No
Calving ease	Birth	No
Gestation duration	Birth	Yes
Birth weight	Birth	Yes
Survival status at 18 mo	18 mo	No
Insemination farm	18 mo	No
Insemination season	18 mo	No
Countable inseminations	18 mo	Yes
Nonreturn status at 18 mo	18 mo	No
No insemination information at 18 mo	18 mo	No
Number of farm movements before 18 mo	18 mo	Yes
Age at first insemination	18 mo	Yes
Type of first insemination	18 mo	No
Number of inseminations	18 mo	Yes
Survival status at 2 yr of age	First calving	No
Raised at a specialty calf-rearing farm	First calving	No
Calving season	First calving	No
Total number of farm movements before calving	First calving	Yes
Age at first calving	First calving	Yes
Calving farm UBN	First calving	No
Calf sex	First calving	No
Calf survival first 24 h	First calving	No
Calving ease, calf	First calving	No
Calf gestation duration	First calving	Yes
Calf birth weight	First calving	Yes
Calf survival, first wk	First calving	No
Calf survival, second wk	First calving	No
Calving records exist	First calving	No
Twins	First calving	No
Milk at 6 wk, kg	6 wk after first calving	Yes
Milk fat percentage at 6 wk	6 wk after first calving	Yes
Milk protein percentage at 6 wk	6 wk after first calving	Yes
Milk SCC at 6 wk	6 wk after first calving	Yes
Milk urea at 6 wk	6 wk after first calving	Yes
Milk lactose percentage at 6 wk	6 wk after first calving	Yes
Cow status indicator at 6 wk	6 wk after first calving	No
Number of negative indications at 6 wk	6 wk after first calving	Yes
Number of days in lactation at 6 wk	6 wk after first calving	Yes
Complete milk measurement available at 6 wk	6 wk after first calving	No
First-parity insemination farm UBN	200 d after first calving	No
First-parity insemination season	200 d after first calving	No
First-parity first insemination type	200 d after first calving	No
Number of inseminations in first parity	200 d after first calving	Yes
Nonreturn status at 200 d after calving	200 d after first calving	No
Age at 200 d after calving	200 d after first calving	Yes
Insemination known in the first parity	200 d after first calving	No
Age at first insemination in the first parity	200 d after first calving	Yes
Number of farm movements at 200 d after calving	200 d after first calving	Yes
Number of known milk testing at 200 d after calving	200 d after first calving	Yes
Milk average kg	200 d after first calving	Yes
Milk average fat percentage	200 d after first calving	Yes
Milk average protein percentage	200 d after first calving	Yes
Milk average SCC	200 d after first calving	Yes
Milk average urea	200 d after first calving	Yes
Milk average lactose percentage	200 d after first calving	Yes
Number of negative indications at 200 d after calving	200 d after first calving	Yes

Continued

Table A1 (Continued). All 66 phenotypic variables and 50 genomic EBV (gEBV) available and the decision moment in which each variable is available

Phenotypic variable	Decision moment	Continuous
Survival status at 200 d after calving	200 d after first calving	No
Number of farm movements in the first parity	200 d after first calving	Yes
gEBV		
NVI, Dutch breeding goal standard	Birth	Yes
Milk, kg	Birth	Yes
Fat, kg	Birth	Yes
Protein, kg	Birth	Yes
Lactose, kg	Birth	Yes
Inet, Dutch production index	Birth	Yes
Cell count	Birth	Yes
Subclinical mastitis	Birth	Yes
Clinical mastitis	Birth	Yes
Udder health	Birth	Yes
Lifespan	Birth	Yes
Lifespan with predictors	Birth	Yes
Birth index	Birth	Yes
Calving ease	Birth	Yes
Post-calving ease	Birth	Yes
Livability calving (maternal)	Birth	Yes
Livability birth (direct)	Birth	Yes
Overall fertility	Birth	Yes
Nonreturn status at 56 d	Birth	Yes
Interval calving to first insemination	Birth	Yes
Calving interval	Birth	Yes
Interval first to last insemination	Birth	Yes
Conception ratio	Birth	Yes
Claw health	Birth	Yes
Calf vitality 3 to 365 d	Birth	Yes
Milking speed	Birth	Yes
Dairy strength	Birth	Yes
Stature	Birth	Yes
Chest width	Birth	Yes
Body depth	Birth	Yes
Angularity	Birth	Yes
Body condition	Birth	Yes
Rump angle	Birth	Yes
Rump width	Birth	Yes
Rear legs hind view	Birth	Yes
Rear leg side view	Birth	Yes
Foot angle	Birth	Yes
Locomotion	Birth	Yes
Fore udder attachment	Birth	Yes
Front teat placement	Birth	Yes
Teat length	Birth	Yes
Udder depth	Birth	Yes
Rear udder height	Birth	Yes
Udder support	Birth	Yes
Rear teat placement	Birth	Yes
Frame	Birth	Yes
Robustness	Birth	Yes
Overall udder score	Birth	Yes
Feet and legs	Birth	Yes
Overall exterior score	Birth	Yes
Milking robot efficiency	Birth	Yes

[1] UBN = unique business number, farm identifier.