Computers and Electronics in Agriculture 177 (2020) 105675

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Improving predictive performance on survival in dairy cattle using an ensemble learning approach

E.M.M. van der Heide[a,*], C. Kamphuis[a], R.F. Veerkamp[a], I.N. Athanasiadis[c], G. Azzopardi[d], M.L. van Pelt[b], B.J. Ducro[a]

[a] Wageningen University & Research Animal Breeding and Genomics, P.O. box 338, 6700 AH Wageningen, the Netherlands
[b] Cooperation CRV, Animal Evaluation Unit, P.O. box 454, 6800 AL Arnhem, the Netherlands
[c] Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, P.O. box 47, 6700 AA Wageningen, the Netherlands
[d] University of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, 9747 AG Groningen, the Netherlands

ARTICLE INFO

Keywords:
Ensemble
Machine learning
Survival
Dairy cow

ABSTRACT

Cow survival is a complex trait that combines traits like milk production, fertility, health and environmental factors such as farm management. This complexity makes survival difficult to predict accurately. This is probably the reason why few studies attempted to address this problem and no studies are published that use ensemble methods for this purpose. We explored if we could improve prediction of cow survival to second lactation, when predicted at five different moments in a cow's life, by combining the predictions of multiple (weak) methods in an ensemble method. We tested four ensemble methods: majority voting rule, multiple logistic regression, random forest and naive Bayes. Precision, recall, balanced accuracy, area under the curve (AUC) and gains in proportion of surviving cows in a scenario where the best 50% were selected were used to evaluate the ensemble model performance. We also calculated correlations between the ensemble models and obtained McNemar's test statistics. We compared the performance of the ensemble methods against those of the individual methods. We also tested if there was a difference in performance metrics when continuous (from 0 to 1) and binary (0 or 1) prediction outcomes were used. In general, using continuous prediction output resulted in higher performance metrics than binary ones. AUCs for models ranged from 0.561 to 0.731, with generally increasing performance at moments later in life. Precision, AUC and balanced accuracy values improved significantly for the naive Bayes and multiple logistic regression ensembles in at least one data set, although performance metrics did remain low overall. The multiple logistic regression ensemble method resulted in equal or better precision, AUC, balanced accuracy and proportion of animals surviving on all datasets and was significantly different from the other ensembles in three out of five moments. The random forest ensemble method resulted in the least significant improvement over the individual methods.

1. Introduction

Cow survival is important from economic, animal welfare and environmental perspectives. When cows survive to reach a high number of lactations, rearing costs are reduced for individual animals as well as across the herd (Mohd Nor et al., 2015; Boulton et al., 2017). Older cows in their third or fourth lactation also produce more milk than young cows, increasing profit per cow (Lehmann et al., 2016) and reducing environmental impact per litre of milk produced (Grandl et al., 2019). A high farm average of number of lactations reached is also an indication of good farm practices with respect to animal welfare (Barkema et al., 2015). As there are many advantages to cows that live long productive lives, it would be beneficial for farmers to keep only those cows that are likely to thrive in a production environment. Selecting cows that have a high probability to survive to higher lactations would be possible by predicting the ability of a cow to survive early on. However, prediction of survival is often not attempted because survival is a very complex trait, combining cow traits such as milk production, fertility and health (Heise et al., 2016) with environmental factors, such as herd size (Shahid et al., 2015) and other farm management factors (Svensson and Hultgren, 2008; Olechnowicz et al., 2016). Although attempts have been made to predict survival in literature (Van Pelt et al., 2015; Gaillard et al., 2016; van der Heide et al., 2019), the complex nature of survival means the predictive performance of these

* Corresponding author.
E-mail address: esther.vanderheide@wur.nl (E.M.M. van der Heide).

https://doi.org/10.1016/j.compag.2020.105675
Received 18 February 2020; Received in revised form 16 June 2020; Accepted 28 July 2020

models remains low.

The prediction of survival may be improved by combining the predictions of multiple (weak) prediction methods. This approach is known as an ensemble method (Knutti et al., 2010; Woźniak et al., 2014), also referred to as hybrid classifier (Woźniak et al., 2014), decision fusion method (Sinha et al., 2008), or aggregation method (Satopää et al., 2014). Ensemble methods aim to maximize the complementary contribution of various classification models (Kotsiantis et al., 2006; Witten et al., 2016). They improve prediction by taking advantage of the underlying differences and strengths of the involved methods. This gives ensemble methods several advantages over individual methods, such as better performance and more robustness (Seni and Elder, 2010). Due to these advantages, ensemble methods are used extensively in other fields like medicine, finance and meteorology (Feldwisch-Drentrup et al., 2010; Tsai and Chen, 2010; Lavecchia, 2015). In the case of survival, ensemble methods have been successfully applied for the prediction of survival in cancer patients (Hothorn et al., 2005; Abreu et al., 2013; Leger et al., 2017). This success in other fields inspires us to evaluate it in the prediction of survival in dairy cattle.

In this study, we investigated if using an ensemble method could improve prediction of survival to second lactation in dairy cattle (Fig. 1). We did not find any previous studies that used ensemble methods to predict survival traits for individual dairy cows. We tested four different ensemble methods, namely voting rule, random forest (Breiman, 2001), naive Bayes (Jensen, 1996) and multiple logistic regression (hereafter referred to as 'regression'). We selected this combination of methods because they are representatives of different types of ensemble methods. Voting rule is the simplest method but is also the most straightforward and transparent. Furthermore, simple methods are not always outperformed by more complex ensemble methods (Witten et al., 2016). Regression, random forest and naive Bayes were selected as representatives of margin-based, decision tree-based and probability-based prediction methods, respectively. Selecting these four different ensemble methods resulted in an overview of the possibilities of ensemble methods to improve prediction of survival in dairy cattle.

2. Materials and methods

2.1. Data

We used five data sets originating from a previous study that predicted survival to second lactation of individual cows (van der Heide et al., 2019). These five data sets consisted of predictions on test data sets from that previous study, a randomly selected 30% of all available animals, stratified by survival (Fig. 2). The data used in the current study is therefore the output from the methods used in the previous study.

1) The input of the current study: the prediction outcomes of the validation data set of van der Heide et al. (2019). The performance metrics for the individual methods were calculated from this data.
2) The input data sets were randomly shuffled three times, each shuffle being divided into four folds.
3) An ensemble method was applied using four-fold cross validation on each of the three shuffles (except for the voting rule, which was applied directly after step 1).
4) The prediction outcomes were used to calculate the performance metrics of the chosen ensemble method.

Prediction outcomes were obtained from five different datasets reflecting information available at five moments in the life of a cow: at birth, at eighteen months of age, at first calving, at six weeks post first calving and at two hundred days post calving. Each data set contained between 2051 (at birth) to 1862 (at 200 days post calving) randomly selected animals (Table 1). The total number of available animals decreased over time due to the removal of non-surviving animals if they died prior to the next moment in life.

Probabilities of survival were estimated using three methods in the previous study: logistic multiple regression, naive Bayes and random forest. This resulted in three continuous prediction outcomes for each animal, one from each method. In addition to the continuous prediction outcomes, we also created binary prediction outcomes for survival (either 0 or 1). For binary outcomes, animals were predicted to survive (a score of 1) when the animal had a predicted probability of survival equal to or above the observed mean chance of survival. Similarly, an animal was predicted not to survive (a score of 0) if its predicted probability was below the mean chance of survival, which varied between 0.86 and 0.92 for the regression and naive Bayes and set to 0.50 for the random forest (see also van der Heide et al., 2019) (Table 2).

2.2. Model and analysis

The data sets were analysed in the statistical program (R Core Team, 2016), where four different ensemble methods were tested. Voting rule was applied using basic R functions, regression was applied using the 'caret' package (Kuhn, 2008), the random forest was applied using the 'randomForest' package (Liaw and Wiener, 2002) and the naive Bayes was applied using 'naivebayes' (Majka, 2018).

The type of voting rule that we used is the majority voting rule (Zhou, 2012); if at least two out of the three original predictions were positive (i.e., animal will survive) the concerned cow is predicted to survive, and any animal with two or more negative predictions is predicted to not survive. No training of the data was required to obtain performance metrics for this method. All possible combinations of outcomes for the voting rule are shown in Table 3.

For the regression, no interactions between the prediction outcomes from the three individual methods tested significant. The models used for the regression could therefore be described as:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i$$

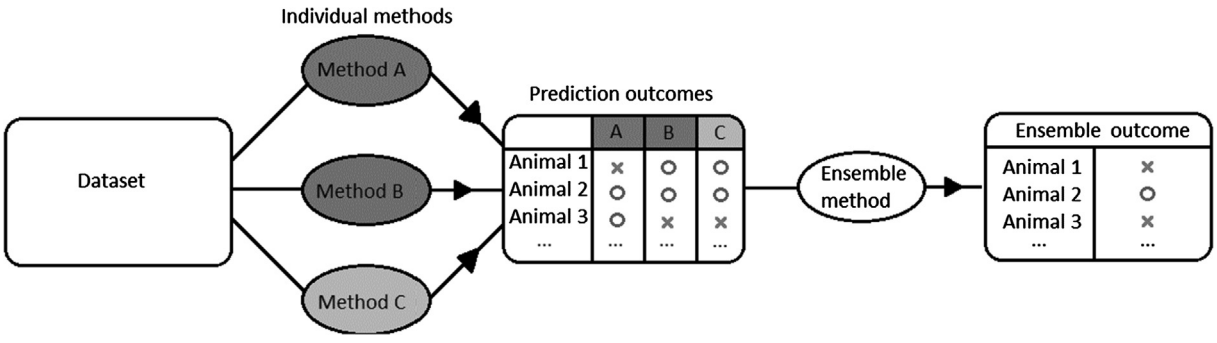

Fig. 1. Ensemble method as applied in this study. Three different methods are used to predict if an animal survives (o) or does not survive (x) based on the same dataset. An ensemble method is then used to aggregate the results in a single prediction.

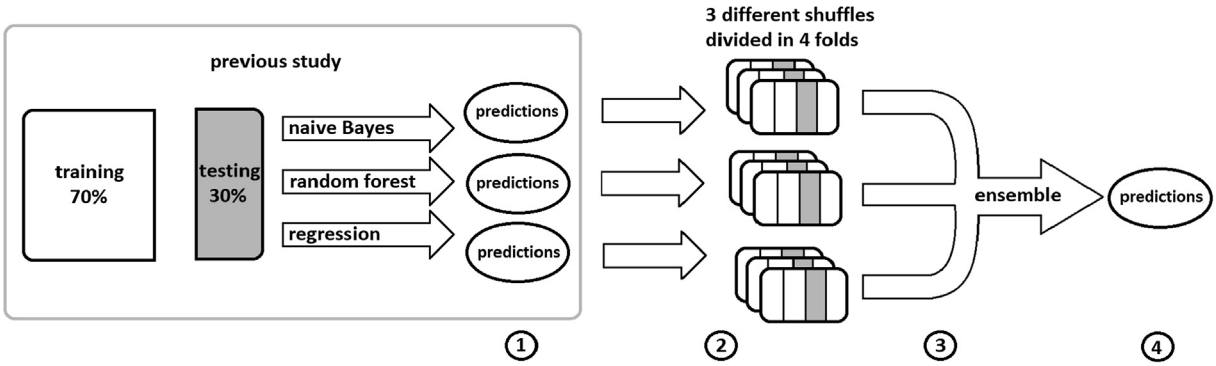

Fig. 2. Schematic depiction of the analysis that was repeated for all data sets: from birth to 200 days post calving, using either continuous or binary outcomes.

Table 1
Distribution of survivors and non-survivors in each data set.

Data set	Survivors	Non-survivors	Total number of animals
Birth	1764	287	2051
18 months of Age	1736	287	2023
First calving	1741	202	1943
6 weeks post first calving	1743	200	1943
200 days post first calving	1723	139	1862

Table 2
Mean and standard deviation of the individual method outcomes used as input in this study.

	Regression		Random Forest		Naive Bayes	
	Mean	St.dev	Mean	St.dev	Mean	St.dev
Birth	0.856	0.041	0.525	0.052	0.798	0.233
18 months	0.866	0.064	0.535	0.062	0.774	0.287
First calving	0.891	0.056	0.532	0.053	0.761	0.315
6 weeks post calving	0.891	0.085	0.535	0.066	0.785	0.330
200 days post calving	0.924	0.083	0.552	0.078	0.815	0.329

Table 3
All possible combinations and outcomes for the majority voting rule.

Individual method outcomes			Voting rule outcome
Multiple linear regression	Random forest	Naive Bayes	
Survives	Survives	Survives	Survives
Survives	Survives	Culled/dies	Survives
Survives	Culled/dies	Survives	Survives
Culled/dies	Survives	Survives	Survives
Survives	Culled/dies	Culled/dies	Culled/dies
Culled/dies	Culled/dies	Survives	Culled/dies
Culled/dies	Survives	Culled/dies	Culled/dies
Culled/dies	Culled/dies	Culled/dies	Culled/dies

where y is the survival status at second calving plus two weeks, X_{i1} through X_{i3} were the predicted outcomes of the three individual methods studied previously (van der Heide et al., 2019), β_1 through β_3 were the regression coefficients for each method and β_0 is the intercept, plus an error term denoted by e_i. For each of the five data sets (from birth to 200 days post calving) a separate model was created.

For the random forest ensemble (Fig. 3), we further tested with different hyperparameters, namely number of trees, number of variables selected at each split and if there was an effect of the seed used. For the number of trees, we tested 1, 5, 10, 50, 100, 150, 200, 250, 300, 500 and 1000 trees in a preliminary study which 500 randomly selected records drawn from the training set of the previous study. The number of trees was subsequently set at 200 as there were no significant

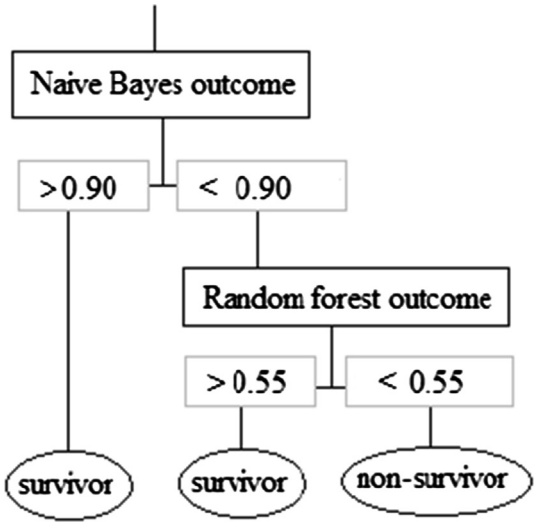

Fig. 3. Example of a decision tree as it might appear in the random forest ensemble method in this study. This tree splits twice, once on the outcomes from the individual naive Bayes and once on the outcomes of the individual random forest method. Values shown for each split are examples. The prediction of this tree is shown in the round end-nodes.

changes in AUC of balanced accuracy if 150 or more trees were used, regardless of the data set and whether binary or continuous input variables were used. A number of variables equal to the square root of the total number of variables was randomly sampled at each split. This is the default setting for this random forest (Liaw and Wiener, 2002). Selecting either 1 (the minimum) or 3 (the maximum) variables per split did not result in significant differences. There were also no significant differences between three randomly selected seeds for randomization. In order to correct for the imbalanced classes for survival, we used class weights set to the proportion of animals in the minority class.

For the naive Bayes classification model, no fine-tuning was required as this method is naturally robust to imbalance and uses no hyperparameters. The prediction using naive Bayes can be described as the predicted range of values for the individual model outcomes given the value for the trait of interest, survival.

Fig. 1 shows a schematic representation of the steps in the analysis. Cross validation was performed to avoid overfitting to the training data (Arlot and Celisse, 2010). We used four-fold cross validation to test the random forest, regression and naive Bayes ensemble methods. We repeated this four-fold cross validation step three times to get a reasonable range of performance metrics to test for significance (Fig. 1, step 2). The four-fold cross validation was done for each data set by splitting each of the three shuffles into four parts, where three of the parts (75%) were used for training the model, and the remaining part (25%) was used for validation. A different seed was used for each shuffle. We used

three x four folds instead of the usual 10-fold cross validation because the number of non-surviving animals in the data set was low. Dividing the data into 10 folds would result in folds with very few non-surviving cases. The model performance metrics for the regression, naive Bayes and random forest ensemble methods are averaged across these twelve validation runs.

Each ensemble method was evaluated by measuring the recall, precision, AUC, balanced accuracy and the proportion of surviving animals when the 50% highest scoring animals were selected. Precision (also known as the positive predictive value) here is the proportion of the correct predictions of non-survivors to the total number of predicted non-survivors. The recall (or sensitivity), which is the proportion of correct predictions of non-surviving cows to the total of non-surviving cows in the entire test set. Both of these metrics quantify the ability of the ensemble method to identify non-surviving animals. As these metrics require a classification, not a probability, animals were divided into the two classes (surviving and non-surviving) using a cut-off determined by the optimal cut-off using the Youden criteria from the receiver-operator curve (Fluss et al., 2005). Balanced accuracy and the AUC are both metrics of overall model performance. The AUC represents the accuracy of the model for all combinations of specificity and sensitivity and was calculated using the R package 'pROC' (Robin et al., 2011). Balanced accuracy is based on the average accuracy from the survivors and non-survivors taken separately (Brodersen et al., 2010). We compared the performance metrics of the ensemble methods to the performance metrics of the individual methods, as published in van der Heide et al, (2019). The proportion of surviving animals when the 50% highest scoring animals were selected is a measurement of the possible effect of these methods in practice. This metric was calculated by selecting the 50% highest scoring animals for a particular method and determining the proportion of animals from that selection that reached second lactation. This mimics how the models could be used in practice; farmers could use the outcomes from the (ensemble) methods to determine which animals to keep (the top 50%) and sell or cull the animals with poor outcomes (the bottom 50%). The percentage of animals kept cannot be too small, as the farmer needs young cows to replace older cows in the herd and was therefore set at 50%.

To determine significance of the ensemble methods' performance metrics compared to the individual methods, we constructed a 95% confidence interval using the mean and standard deviation obtained from the 12 replications for each method. This confidence interval could then be used to determine if the ensemble results were statistically different from the results of the individual methods. We further calculated the correlation between the ensemble methods and the individual methods and performed the McNemar's test to determine statistical significance in performance between methods. The correlations were calculated using the continuous outcomes. The correlations for the voting rule or the binary datasets were not calculated due to the limited number of possible prediction outcomes. We calculated correlations to investigate if any of the statistical ensemble methods resulted in different predictions for individual animals compared to the individual methods. This both shows if there were differences between the individual predictions used as input, as well as if there are still non-random differences between the ensemble methods. The McNemar's test statistic was calculated as

$$X^2 = \frac{(b-c)^2}{b+c}$$

where b is the number of cases where method A correctly predicted survival but method B did not and c is the number of cases where method B predicted survival correctly but method A did not (Table 4). Table 4 also includes an example of McNemar's test on our data McNemar's test statistics were shown only for the data set with raw probabilities. The McNemar's test indicates significant differences between methods in cases that one method classified correctly and the other does not. It does not indicate which of the two tested methods performs significantly better. To determine which prediction method was superior, the paired t-tests of Section 3.2 of the results are more suitable.

Table 4
Schematic representation of McNemar's test as performed in this study. An example of a McNemar's test as performed in this study is shown below in italics at 200 days post calving. The test statistic for this example was 1.291 (p = 0.256).

		Method A	
		Incorrect	Correct
Method B	**Incorrect**	a *294*	b *176*
	Correct	c *199*	d *1193*

3. Results

First, we will discuss the results of voting rule (the fourth row in Tables 5a through 5e), as this method was conceptually difference from the other ensemble methods. The performance metrics for the other ensemble methods are shown in Section 3.2 and Section 3.3, for the data sets with continuous outcomes and binary outcomes respectively (Tables 5a–5e). In the final section of the results we describe the correlations between the ensemble methods and the individual methods.

3.1. Voting rule

Because the voting rule only produces binary output and no probabilities, two metrics could not be calculated for voting rule: the AUC and proportion surviving if the top 50% best animals are selected. Voting rule had the highest precision at 18 months of age, first calving, 6 weeks post calving and 200 days post calving. However, this came at the cost of having the lowest recall of all methods (individual and ensemble) on those datasets. Voting rule resulted in a lower balanced accuracy (0.001 to 0.018 lower) at birth, 18 months and first calving compared to the next worst performing method. The balanced accuracy of the voting rule was lower than at least one of the individual methods at 6 weeks and 200 days post calving.

3.2. Continuous outcomes

Using continuous prediction outcomes, the naive Bayes and regression ensemble methods both significantly outperformed all three individual methods on balanced accuracy at birth and 18 months (Tables 5a and 5b). Furthermore, both outperformed all three individual methods at precision at 18 months. At 6 weeks post first calving, the regression ensemble method significantly outperformed (p less than 0.05) all three individual methods on balanced accuracy (Table 5d). At birth, regression outperformed at least one individual method on all metrics but recall, where it was outperformed by the individual random forest method (Table 5a). From 18 months onward, the regression ensemble improved on at least one individual method at all metrics, and never performed significantly worse than any individual method (Table 5b–5e). Naive Bayes did similarly well, improving on at least one individual method on all performance metrics except for recall at birth and proportion of surviving heifers at 200 days past calving (Tables 5a–5e). The random forest ensemble method never outperformed all three individual methods on any of the metrics in any dataset. It also had less consistent performance than the other two ensemble methods. For example, at 200 days post calving this method significantly improved on at least one individual method in precision and balanced accuracy, but resulted in significantly worse recall, AUC and proportion of heifers surviving than at least one individual method.

Table 5a
Performance metrics of the predictions on the data set gathered at birth.

	Data type	Recall	Precision	Balanced accuracy	AUC	Proportion surviving if best 50% are selected
Individual methods						
Regression		0.446	0.201	0.579	0.599	0.892
Random Forest		0.690	0.159	0.549	0.561	0.881
Naive Bayes		0.432	0.200	0.576	0.598	0.888
Ensemble methods						
Voting rule		0.321	0.168	0.531	NA	NA
Regression	continuous	0.563^{acx}	0.212^{b}	0.603^{abc}	0.606^{b}	0.901^{b}
	binary	0.552^{acx}	0.188^{b}	0.579	0.576^{x}	NA
Random Forest	continuous	0.614^{ac}	0.184^{bx}	0.580^{b}	0.576^{x}	0.890
	binary	0.630^{acx}	0.186^{bx}	0.585^{b}	0.590^{b}	NA
Naive Bayes	continuous	0.583^{acx}	0.196^{b}	0.598^{abc}	0.601^{b}	0.898^{b}
	binary	0.622^{ac}	0.187^{b}	0.582^{b}	0.585^{b}	NA

The random forest ensemble performed best at first calving, outperforming at least one individual method on all metrics except AUC.

3.3. Binary outcomes

In general, using binary outcomes resulted in lower performance metrics than using continuous outcomes, and no ensemble method outperformed all three individual methods using this data type. Despite often improving precision, all three ensemble methods were significantly worse at AUC and precision than at least one individual method from first calving onwards. At birth, naive Bayes scored better than at least one method on all performance metrics available (Table 5a). At 18 months, all three ensemble methods improved on recall, and regression and naive Bayes also improved on balanced accuracy over the regression individual method (Table 5b). There were no significant differences from the individual methods on the other metrics. From first calving onwards, all ensemble methods performed equal or better than the individual methods at precision (Tables 5c–5e). Regression was the only ensemble method that did not improve precision over at least one individual method at 6 weeks post calving but was also the only method to outperform at least one individual method on balanced accuracy at 200 days post calving. At first calving, naive Bayes also improved balanced accuracy in addition to precision. For binary outcomes, the proportion of surviving animals is listed as NA as the animals could not be properly ranked using only binary outcomes. This is due to more than 50% of the animals getting the maximum of 3 positive predictions regardless of the data set.

3.4. Correlations between methods

The naive Bayes and regression ensemble methods resulted in predictions that remained strongly correlated with one or more of the individual methods (Table 6). The regression ensemble method had a correlation of at least 0.692 with the corresponding individual method. Similarly, the naive Bayes ensemble method was correlated at least 0.745 with the naive Bayes individual method. This indicates that both the naive Bayes and regression ensemble methods made similar predictions as their corresponding individual methods. The random forest ensemble method had the lowest correlations with individual methods, ranging from 0.442 to 0.736. The random forest individual method had the lowest correlations with the ensemble methods. This indicates that the random forest ensemble method used all methods to a similar extent and relied least on the results of one individual method out of all the ensemble methods. The highest correlation found overall was 0.970, between the regression ensemble method and the naive Bayes individual method at 18 months of age. The lowest correlations were found at birth, where the regression ensemble method and the naive Bayes ensemble method were both correlated less than 0.5 with the random forest individual method.

3.5. McNemar's test

We further applied the McNemar's test to determine if the differences in classifications between the methods were significant or not (Table 7). For the majority of comparisons between methods there were statistically significant differences, which was as expected as significant differences between ensemble methods in many performance metrics (Tables 5a–5e). At birth, only the comparison between the individual regression and regression ensemble method was not significant, which is surprising as this ensemble did achieve significantly better recall than the individual regression (Table 5a). However, as there is a trade-off between recall (sensitivity) and specificity, it is possible that the specificity of this ensemble was also significantly worse than the individual method, resulting in no net difference.

There were no obvious patterns in non-significant differences in the later moments, although ensembles and individual methods using the same underlying method were more likely to not be statistically different than combinations using different methods. Voting rule was also

Table 5b
Performance metrics of the predictions on the data set gathered at 18 months.

	Data type	Recall	Precision	Balanced accuracy	AUC	Proportion surviving if best 50% are selected
Individual methods						
Regression		0.505	0.211	0.597	0.611	0.897
Random Forest		0.610	0.200	0.604	0.615	0.904
Naive Bayes		0.575	0.223	0.622	0.643	0.904
Ensemble methods						
Voting rule		0.397	0.231	0.589	NA	NA
Regression	continuous	0.554^{ax}	0.250^{abc}	0.638^{abc}	0.643^{ab}	0.908^{a}
	binary	0.588^{a}	0.212	0.613^{a}	0.628	NA
Random Forest	continuous	0.600^{a}	0.221	0.618^{a}	0.623	0.905
	binary	0.566^{a}	0.218	0.612^{a}	0.626	NA
Naive Bayes	continuous	0.572	0.245^{abc}	0.636^{abc}	0.641^{ab}	0.910^{a}
	binary	0.580^{a}	0.214	0.610	0.626	NA

Table 5c
Performance metrics of the predictions on the data set gathered at first calving.

	Data type	Recall	Precision	Balanced accuracy	AUC	Proportion surviving if best 50% are selected
Individual methods						
Regression		0.465	0.152	0.582	0.608	0.920
Random Forest		0.654	0.142	0.597	0.622	0.931
Naive Bayes		0.619	0.175	0.641	0.657	0.939
Ensemble methods						
Voting rule		0.352	0.177	0.581	NA	NA
Regression	continuous	0.688^{a}	0.175^{ab}	0.647^{ab}	0.658^{ab}	0.941^{ab}
	binary	0.509^{x}	0.182^{ab}	0.613^{ax}	0.620^{x}	NA
Random Forest	continuous	0.632^{a}	0.165^{b}	0.628^{a}	0.623	0.934^{a}
	binary	0.516^{x}	0.181^{ab}	0.614^{ax}	0.620^{x}	NA
Naive Bayes	continuous	0.649^{a}	0.174^{ab}	0.643^{ab}	0.655^{ab}	0.939^{a}
	binary	0.622^{x}	0.187^{ab}	0.582^{a}	0.585^{x}	NA

more likely to have little or no difference with the individual prediction methods but was always significantly different from the other ensemble methods. This is likely due to the worse performance of the voting rule compared to the other ensembles (Tables 5a–5e). At birth, all ensembles performed significantly different from each other. In the last decision moment, there was no significant difference between any of the ensembles with the exception of voting rule. In two occasions a pair of methods performed identically, at first calving, between the naive Bayes individual method and voting rule and at 200 days post calving between the multiple regression and naive Bayes ensemble (Table 8). This meant that the diagonal (cases where only one method predicts correctly) is equal for the two methods.

4. Discussion

We investigated if the prediction of survival to second lactation in dairy cattle could be improved by using ensemble methods. In our study, regression as an ensemble method always resulted in equal or better performance on precision, AUC, balanced accuracy and proportion of surviving animals. It also performed better than at least one individual method on the precision metric from first calving onwards. Comparing the results of the current study to other studies predicting survival was difficult, because even though there are many studies studying factors that describe survival (Brickell and Wathes, 2011; De Vries and Marcondes, 2020), few studies actually predict survival (Shmueli, 2010). Furthermore, in studies where survival was predicted, the trait of interest was often continuous, predicting for example lifespan in days (Cruickshank et al., 2002; Caraviello et al., 2004), which makes it difficult to compare results from literature to our own results. Instead, we could compare the results of the current study to studies predicting binary traits that are important components of survival, such as insemination outcome (Pinedo et al., 2010). Insemination outcome is an important component for survival, as fertility problems are the main cause of culling in the Netherlands (Zijlstra et al., 2013; Compton et al., 2017). Therefore, to predict if a cow will survive to the next lactation, a model indirectly also predicts if a cow or heifer will get pregnant in that lactation. Insemination outcome is further similar to survival as it is also an unbalanced binary trait as it is measured in success or failure and the majority of cows requires more than one insemination to become pregnant (Rutten et al., 2016). There are many studies attempting to predict insemination outcome using a variety of different methods. Hempstalk et al. (2015) used a range of different methods including naive Bayes, Bayesian network, logistic regression, support vector machines, partial least squares regression and random forest and reported AUC values ranging from 0.487 to 0.675. Several studies using regression found similar AUCs, ranging from 0.58 to 0.65 on a variety of data sets (Fenlon et al., 2016; Blavy et al., 2018; Toledo-Alvarado et al., 2018; Delhez et al., 2020). Shahinfar et al. (2014) used naive Bayes, Bayesian networks, bagging in combination with random forest amongst other methods on a much larger data set and reported AUC values ranging from 0.61 to 0.76. In our study, AUC values ranged from 0.561 (for the individual random forest method at birth) to 0.713 (for the individual regression at 200 days post calving). This is similar to AUC values found for insemination success, especially at the later moments. AUC values prior to first calving were lower, but this was expected as much less data was available prior to first calving. Higher AUC values than those found in our study did occur in literature as well, for example an AUC of 0.859 for the prediction of insemination success using hormone concentrations in milk (Faustini et al., 2007). The high AUC in this study could be due to the use of variables like hormone concentrations, which were accurate measures for the trait of interest but not routinely collected. The variables at the basis of the ensemble in this study were routinely collected on a large number of farms and therefore no expensive or difficult to measure traits were included, which excluded some very relevant variables.

The goal of this study was to improve the prediction of survival so that it could be used for selection or replacement heifers in practice. While there appeared to be a benefit of applying the regression

Table 5d
Performance metrics of the predictions on the data set gathered at 6 weeks post first calving.

	Data type	Recall	Precision	Balanced accuracy	AUC	Proportion surviving if best 50% are selected
Individual methods						
Regression		0.575	0.214	0.666	0.702	0.944
Random Forest		0.640	0.149	0.611	0.634	0.931
Naive Bayes		0.490	0.219	0.645	0.671	0.935
Ensemble methods						
Voting rule		0.440	0.222	0.631	NA	NA
Regression	continuous	0.616^{c}	0.228^{ab}	0.678^{abc}	0.701^{bc}	0.944^{bc}
	binary	0.507^{x}	0.226	0.651^{bx}	0.658^{x}	NA
Random Forest	continuous	0.555^{cx}	0.240^{b}	0.670^{bc}	0.702^{bc}	0.944^{bc}
	binary	0.545^{x}	0.207^{b}	0.650^{bx}	0.664^{x}	NA
Naive Bayes	continuous	0.624^{c}	0.212^{b}	0.669^{bc}	0.695^{bc}	0.948^{bc}
	binary	0.530^{x}	0.211^{b}	0.650^{bx}	0.664^{x}	NA

Table 5e
Performance metrics of the predictions on the dataset gathered at 200 days post first calving.

	Data type	Recall	Precision	Balanced accuracy	AUC	Proportion surviving if best 50% are selected
Individual methods						
Regression		0.547	0.183	0.675	0.713	0.960
Random Forest		0.770	0.115	0.647	0.687	0.966
Naive Bayes		0.547	0.135	0.632	0.657	0.956
Ensemble methods						
Voting rule		0.425	0.189	0.639	NA	NA
Regression	continuous	0.706[ab]	0.165[b]	0.680[bc]	0.709[c]	0.965[c]
	binary	0.488[x]	0.190[bc]	0.659[c]	0.664[x]	NA
Random Forest	continuous	0.554[x]	0.178[bc]	0.662[c]	0.678[x]	0.954[x]
	binary	0.515[x]	0.186[bc]	0.660[x]	0.672[x]	NA
Naive Bayes	continuous	0.702[ac]	0.166[b]	0.682[bc]	0.704[c]	0.962
	binary	0.516[x]	0.184[bc]	0.662[bcx]	0.673[cx]	NA

Note: significance could not be calculated for voting rule results.
[a] significantly outperforms the individual method multiple logistic regression.
[b] significantly outperforms the individual method random forest.
[c] significantly outperforms the individual method naive Bayes.
[x] significantly worse than one or more of the three individual methods.

Table 6
Average correlations between the results of the three statistical ensemble methods and the results of the three individual methods. The abbreviation p.c. stands for post calving.

Dataset	Ensemble method	Regression individual method	Random Forest individual method	Naive Bayes individual method
Birth	Regression	0.849	0.494	0.801
	Naive Bayes	0.696	0.663	0.890
	Random Forest	0.591	0.442	0.568
18 months	Regression	0.692	0.736	0.970
	Naive Bayes	0.769	0.727	0.867
	Random Forest	0.549	0.626	0.753
First calving	Regression	0.709	0.714	0.911
	Naive Bayes	0.785	0.606	0.749
	Random Forest	0.536	0.603	0.720
6 weeks p.c.	Regression	0.891	0.588	0.759
	Naive Bayes	0.788	0.702	0.806
	Random Forest	0.733	0.602	0.666
200 days p.c.	Regression	0.921	0.736	0.658
	Naive Bayes	0.733	0.602	0.666
	Random Forest	0.718	0.696	0.551

ensemble method, performance metrics remained low overall. In literature, other studies also show only small or inconsistent improvements in predictive performance when using an ensemble method (Knutti et al., 2010; Larsen et al., 2019). Similarly, there are studies where ensemble methods are outperformed by individual methods in certain situations (Barbareschi et al., 2015). Although it is difficult to quantify exactly how good a model must be before it is useful in practice, there are two metrics which are especially relevant: the proportion of surviving heifers if the top 50% scoring heifers is selected and the precision. The proportion of surviving heifers indicates the increase in heifers reaching second lactation when the model is used for selection vs. random selecting 50%, the intended effect of the model in practice. Although some ensembles improved on individual methods for proportion of surviving heifers, none of the ensembles improved consistently over all the individual methods. Precision is important for the practical application of the models in this study because this metric indicated how often the model made a false prediction that an animal would not survive to second lactation. For a farmer, a false prediction for a surviving cow would be more damaging than a false prediction for a non-surviving cow. If the model predicts a cow will not survive, that cow would be sold or slaughtered, resulting in the irrevocable loss of that animal. If the model predicts a cow will survive but in reality, it does not, a farmer would have more opportunities to sell or cull the cow after that prediction moment. In our study, the highest precision found was 0.250. This means that only a quarter of animals predicted to not survive to second lactation would actually fail to do so. Even if the ensembles do significantly improve over the individual models, it is unlikely that farmers would follow the advice to sell a young animal knowing the model is wrong 75% of the time.

There are several possible reasons why the ensemble methods did not result in a large increase in model performance. The correlation between the input data, in this case the output from the three individual models, is an important indicator for the added value of using an ensemble (Woźniak et al., 2014). If the methods in an ensemble are too strongly correlated, combining them does not result in improved predictive ability (Pena and van den Dool, 2008; Knutti et al., 2010). In our current study, the correlations of the prediction outcomes used as input data were between 0.417 and 0.700 (van der Heide et al., 2019). This was lower than expected as the three methods were trained on the same data set. However, it is possible that the correlations were still too high, limiting the variability among the prediction outcomes. In a study predicting lameness in dairy cows, a trait related to survival, even much lower correlations of 0.17 to 0.40 between input methods similarly resulted in only marginal improvement if an ensemble was used, from an AUC between 0.73 and 0.75 to an AUC of 0.76 when using an ensemble (Warner et al., 2020). In any case, significant differences between the input methods did exist as the McNemar's test statistic proved there were differences between all of the input methods in our study except between the naive Bayes and regression models at first calving and 6 weeks post calving (Table 7). Correlations between input variables can also cause additional difficulties when selecting an ensemble method. The naive Bayes ensemble method, for example, assumes independence among the input variables (Friedman et al., 1997). Correlations between input variables could thus have caused underperformance of the ensemble methods. The voting rule may also not be as effective in cases where methods are correlated or where a limited number of models were combined (Oza and Tumer, 2008). Class imbalance in the trait of interest was another potential reason ensembles did not result in sufficient improvement over the individual methods (Stefanowski, 2016). In the case of survival, most animals survive to second lactation (86%), whereas a minority (14%) do not. As there are fewer examples, this minority is more difficult to predict, despite being the class of interest. Although the use of ensemble methods is in fact a popular solution to imbalance problems (Haixiang et al., 2017), an ensemble using only three methods as input may not have been robust

Table 7
p-values of the McNemar's tests between pairs of prediction methods. The rows show the methods per moment in life and the columns indicate the corresponding methods. Values below 0.005 were considered statistically significant. Post calving is abbreviated to 'p.c.'.

		Individual methods			Ensemble methods			
		Naive Bayes	Random Forest	Regression	Voting rule	Naive Bayes	Random Forest	Regression
Birth	Naive Bayes		0.000	0.000	0.000	0.000	0.000	0.000
	Random Forest			0.000	0.000	0.000	0.000	0.000
	Regression				0.000	0.000	0.000	0.239
	Voting rule					0.000	0.000	0.000
	Naive Bayes ensemble						0.000	0.000
	Random Forest ensemble							0.002
	Regression ensemble							
18 months	Naive Bayes		0.000	0.000	0.000	0.000	0.000	0.000
	Random Forest			0.000	0.000	0.000	0.000	0.236
	Regression				0.000	0.428	0.537	0.003
	Voting rule					0.000	0.000	0.000
	Naive Bayes ensemble						0.958	0.000
	Random Forest ensemble							0.000
	Regression ensemble							
First calving	Naive Bayes		0.000	0.622	0.999	0.000	0.000	0.000
	Random Forest			0.000	0.000	0.000	0.000	0.000
	Regression				0.495	0.000	0.000	0.000
	Voting rule					0.000	0.000	0.000
	Naive Bayes ensemble						0.000	0.462
	Random Forest ensemble							0.000
	Regression ensemble							
6 weeks p.c.	Naive Bayes		0.000	0.036	0.123	0.000	0.000	0.000
	Random Forest			0.000	0.000	0.000	0.288	0.721
	Regression				0.171	0.000	0.000	0.000
	Voting rule					0.000	0.000	0.000
	Naive Bayes ensemble						0.000	0.000
	Random Forest ensemble							0.087
	Regression ensemble							
200 days p.c.	Naive Bayes		0.000	0.001	0.116	0.000	0.000	0.000
	Random Forest			0.000	0.000	0.006	0.312	0.006
	Regression				0.001	0.000	0.000	0.000
	Voting rule					0.000	0.000	0.000
	Naive Bayes ensemble						0.231	0.999
	Random Forest ensemble							0.256
	Regression ensemble							

Table 8
Table for McNemar's test for the multiple regression and naive Bayes ensemble methods at 200 days post calving (p value = 0.999).

		Multiple regression ensemble prediction	
		Incorrect	**Correct**
Naive Bayes ensemble prediction	**Incorrect**	435	59
	Correct	58	1310

enough. Furthermore, class imbalance is especially problematic in cases where there are few samples and the classes are difficult to separate (Ali et al., 2015), both of which played a role here. In this study, the individual methods and the ensemble methods were adapted separately to the class imbalance problem. For future studies, (ensemble) algorithms specifically designed to cope with class imbalance such as RUS boosted trees could be considered (Galar et al., 2011). These methods allow for a more systemic approach to coping with the class imbalance problem, potentially increasing the effectiveness of the ensembles.

In general, improving the predictive performance of ensembles is done by increasing the underlying diversity of the ensemble (Dietterich, 2000; Tang et al., 2006; Sagi and Rokach, 2018). It is also possible to change the method which is used to build the ensemble, as there are dozens of different methods which can be used to build an ensemble (Zhou, 2012). However, the benefit of using a different method to build the ensemble is often minimal compared to the benefit of using an ensemble method over an individual method (Džeroski and Ženko, 2004; Berk, 2006). Furthermore, using more complex methods to aggregate the method results in less interpretable results and requires larger training datasets (Ren et al., 2016). There are several approaches to increase the underlying diversity of an ensemble. One approach is to add additional methods to the ensemble. New models included in the ensemble should also be decent stand-alone prediction methods and preferably use a different approach from the methods already in the ensemble (Haixiang et al., 2017). Candidate methods to be included in our ensemble are for example a K-nearest neighbour (Guo et al., 2003) or a neural network approach (Paliwal and Kumar, 2009; Liakos et al., 2018) both of which are different from the methods already included in the ensemble. If no additional methods are used, the diversity of the ensemble can also be increased by obtaining multiple models from the same method and training individual models on different subsets of the data (Ren et al., 2016). Popular ensemble methods such as bootstrap aggregation (bagging), Adaptive boosting (AdaBoost) and random subspace approaches make use of this approach (Freund and Schapire, 1996; Zhou, 2012). In this way itis unlikely to result in major gains in accuracy in our study because the data set was small and unbalanced, which would cause difficulties if it were split into sub-sets. Furthermore, the random forest method, which is also based on this approach (Breiman, 1996, 2001), resulted in the worst performing ensembles.

The last approach to increase the performance of the ensemble method is by increasing the number of available variables. The reason why survival traits are very difficult to predict is the variety of different factors involved. Not all of the different potential causes of death or culling of dairy cows could be predicted using the original data set (van der Heide et al., 2019). For example, there was no information on disease occurrence available in the original data set, which is obviously

an important cause of death for dairy cows (Svensson et al., 2006). So, while it was possible to take advantage of the differences between the methods using ensemble methods, there are likely limitations to increasing model performance by varying the method alone. Increasing the amount of data is more resource intensive than changing the prediction method, but also often the best option to increase predictive accuracy for a particular prediction problem (Domingos, 2012).

5. Conclusion

Using logistic multiple regression as an ensemble method resulted in equal or better precision, AUC, balanced accuracy and improvement in proportion of animals surviving. Naive Bayes was the second-best ensemble method, and the random forest ensemble method resulted in the least significant improvement over the individual methods. Precision, AUC and balanced accuracy values improved significantly over all methods at specific datasets for naive Bayes and logistic multiple regression ensembles, although they remained low overall (AUC's ranged from 0.561 to 0.731, increasing as more variables became available). Where multiple prediction models are available, regression can be a useful method to investigate the additional value of using ensemble methods.

CRediT authorship contribution statement

E.M.M. Heide: Conceptualization, Methodology, Formal analysis, Visualization, Writing - original draft. **C. Kamphuis:** Supervision, Writing - review & editing. **R.F. Veerkamp:** Supervision, Project administration, Writing - review & editing. **I.N. Athanasiadis:** Conceptualization, Methodology, Supervision, Writing - review & editing. **G. Azzopardi:** Supervision, Writing - review & editing. **M.L. Pelt:** Resources. **B.J. Ducro:** Supervision, Project administration, Writing - review & editing.

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.

Acknowledgements

This work was part of the Breed4Food research program, project "Smart animal breeding with advanced machine learning" with project number 14295, which was 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. We also acknowledge GenTORE for funding from European Community's H2020 Framework Programme – GenTORE, under grand agreement no. 727213. The data for this study was provided by cattle improvement cooperative CRV (Arnhem, the Netherlands).

References

Abreu, P.H., Amaro, H., Silva, D.C., Machado, P., Abreu, M.H., Afonso, N., Dourado, A., 2013. Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data. In: Proceedings of the Mediterranean Conference on Medical and Biological Engineering and Computing, pp. 1366–1369.
Ali, A., Shamsuddin, S.M., Ralescu, A.L., 2015. Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl 7 (3), 176–204.
Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79.
Barbareschi, M., Del Prete, S., Gargiulo, F., Mazzeo, A., Sansone, C., 2015. Decision tree-based multiple classifier systems: an FPGA perspective. In: International Workshop on Multiple Classifier Systems. Springer, Cham, pp. 194–205.
Barkema, H., Von Keyserlingk, M., Kastelic, J., Lam, T., Luby, C., Roy, J.-P., LeBlanc, S., Keefe, G., Kelton, D., 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. J. Dairy Sci. 98 (11), 7426–7445.
Berk, R.A., 2006. An introduction to ensemble methods for data analysis. Sociolog. Meth. Res. 34 (3), 263–295.
Blavy, P., Friggens, N., Nielsen, K., Christensen, J., Derks, M., 2018. Estimating probability of insemination success using milk progesterone measurements. J. Dairy Sci. 101 (2), 1648–1660.
Boulton, A.C., Rushton, J., Wathes, D.C., 2017. An empirical analysis of the cost of rearing dairy heifers from birth to first calving and the time taken to repay these costs. Animal 11 (8), 1372–1380. https://doi.org/10.1017/S1751731117000064.
Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Brickell, J., Wathes, D., 2011. A descriptive study of the survival of Holstein-Friesian heifers through to third calving on English dairy farms. J. Dairy Sci. 94 (4), 1831–1838.
Brodersen, K. H., C. S. Ong, K. E. Stephan, and J. M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. p 3121-3124.
Caraviello, D., Weigel, K., Gianola, D., 2004. Prediction of longevity breeding values for US Holstein sires using survival analysis methodology. J. Dairy Sci. 87 (10), 3518–3525.
Compton, C., Heuer, C., Thomsen, P.T., Carpenter, T., Phyn, C., McDougall, S., 2017. Invited review: A systematic literature review and meta-analysis of mortality and culling in dairy cattle. J. Dairy Sci. 100 (1), 1–16.
Cruickshank, J., Weigel, K., Dentine, M., Kirkpatrick, B., 2002. Indirect prediction of herd life in Guernsey dairy cattle. J. Dairy Sci. 85 (5), 1307–1313.
De Vries, A., Marcondes, M., 2020. Overview of factors affecting productive lifespan of dairy cows. Animal 14 (S1), s155–s164.
Delhez, P., Ho, P., Gengler, N., Soyeurt, H., Pryce, J., 2020. Diagnosing the pregnancy status of dairy cows: How useful is milk mid-infrared spectroscopy? J. Dairy Sci.
Dietterich, T. G. 2000. Ensemble methods in machine learning. In: International workshop on multiple classifier systems. p 1-15.
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.
Džeroski, S., Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one? Machine Learning 54 (3), 255–273.
Faustini, M., Battocchio, M., Vigo, D., Prandi, A., Veronesi, M., Comin, A., Cairoli, F., 2007. Pregnancy diagnosis in dairy cows by whey progesterone analysis: An ROC approach. Theriogenology 67 (8), 1386–1392.
Feldwisch-Drentrup, H., Schelter, B., Jachan, M., Nawrath, J., Timmer, J., Schulze-Bonhage, A., 2010. Joining the benefits: combining epileptic seizure prediction methods. Epilepsia 51 (8), 1598–1606.
Fenlon, C., O'Grady, L., Dunnion, J., Shalloo, L., Butler, S., Doherty, M., 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.
Freund, Y., and R. E. Schapire. 1996. Experiments with a new boosting algorithm. In: icml. p 148-156.
Fluss, R., Faraggi, D., Reiser, B., 2005. Estimation of the Youden Index and its associated cutoff point. Biometrical J. Mathematical Meth. Biosci. 47 (4), 458–472.
Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29 (2–3), 131–163.
Gaillard, C., Martin, O., Blavy, P., Friggens, N., Sehested, J., Phuong, H., 2016. Prediction of the lifetime productive and reproductive performance of Holstein cows managed for different lactation durations, using a model of lifetime nutrient partitioning. J. Dairy Sci. 99 (11), 9126–9135.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst Man Cybernet. Part C (Applications Rev.) 42 (4), 463–484.
Grandl, F., Furger, M., Kreuzer, M., Zehetmeier, M., 2019. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. Animal 13 (1), 198–208.
Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN model-based approach in classification. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems, pp. 986–996.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 73, 220–239.
Heise, J., Liu, Z., Stock, K.F., Rensing, S., Reinhardt, F., Simianer, H., 2016. The genetic structure of longevity in dairy cows. J. Dairy Sci. 99 (2), 1253–1265.
Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M.J., 2005. Survival ensembles. Biostatistics 7 (3), 355–373.
Jensen, F.V., 1996. An introduction to Bayesian networks. UCL press, London.
Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A., 2010. Challenges in combining projections from multiple climate models. J. Clim. 23 (10), 2739–2758.
Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., 2006. Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26 (3), 159–190.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.
Larsen, M.L.V., Pedersen, L.J., Jensen, D.B., 2019. Prediction of tail biting events in finisher pigs from automatically recorded sensor data. Animals 9 (7), 458.
Lavecchia, A., 2015. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today 20 (3), 318–331.
Leger, S., Zwanenburg, A., Pilz, K., Lohaus, F., Linge, A., Zöphel, K., Kotzerke, J., Schreiber, A., Tinhofer, I., Budach, V., 2017. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci.

Rep. 7 (1), 13206.
Lehmann, J.O., Fadel, J., Mogensen, L., Kristensen, T., Gaillard, C., Kebreab, E., 2016. Effect of calving interval and parity on milk yield per feeding day in Danish commercial dairy herds. J. Dairy Sci. 99 (1), 621–633.
Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors 18 (8), 2674.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2 (3), 18–22.
Majka, M. 2018. naivebayes: High Performance Implementation of the Naive Bayes Algorithm.
Mohd Nor, N., Steeneveld, W., Mourits, M., Hogeveen, H., 2015. The optimal number of heifer calves to be reared as dairy replacements. J. Dairy Sci. 98 (2), 861–871.
Olechnowicz, J., Kneblewski, P., Jaśkowski, J., Włodarek, J., 2016. Effect of selected factors on longevity in cattle: a review. J. Anim. Plant Sci 26, 1533–1541.
Oza, N.C., Tumer, K., 2008. Classifier ensembles: Select real-world applications. Information Fusion 9 (1), 4–20.
Paliwal, M., Kumar, U.A., 2009. Neural networks and statistical techniques: A review of applications. Expert Syst. Appl. 36 (1), 2–17.
Pena, M., van den Dool, H., 2008. Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. J. Clim. 21 (24), 6521–6538.
Pinedo, P., De Vries, A., Webb, D., 2010. Dynamics of culling risk with disposal codes reported by Dairy Herd Improvement dairy herds. J. Dairy Sci. 93 (5), 2250–2261.
R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
Ren, Y., Zhang, L., Suganthan, P.N., 2016. Ensemble classification and regression-recent developments, applications and future directions. IEEE Comput. Intell. Mag. 11 (1), 41–53.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. 12, 77.
Rutten, C., Steeneveld, W., Vernooij, J., Huijps, K., Nielen, M., Hogeveen, H., 2016. A prognostic model to predict the success of artificial insemination in dairy cows based on readily available data. J. Dairy Sci. 99 (8), 6764–6779.
Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdisciplinary Rev. Data Min. Knowledge Discovery 8 (4), e1249.
Satopää, V.A., Baron, J., Foster, D.P., Mellers, B.A., Tetlock, P.E., Ungar, L.H., 2014. Combining multiple probability predictions using a simple logit model. Int. J. Forecast. 30 (2), 344–356.
Seni, G., Elder, J.F., 2010. Ensemble methods in data mining: improving accuracy through combining predictions. In: Synthesis Lectures on data mining and knowledge discovery, pp. 1–126 1.
Shahid, M., Reneau, J., Chester-Jones, H., Chebel, R., Endres, M.I., 2015. Cow-and herd-level risk factors for on-farm mortality in Midwest US dairy herds. J. Dairy Sci. 98 (7), 4401–4413.
Shmueli, G., 2010. To explain or to predict? Statistical Sci. 25 (3), 289–310.
Sinha, A., Chen, H., Danu, D., Kirubarajan, T., Farooq, M., 2008. Estimation and decision fusion: A survey. Neurocomputing 71 (13–15), 2650–2656.
Stefanowski, J., 2016. Dealing with data difficulty factors while learning from imbalanced data, Challenges in computational statistics and data mining. Springer 333–363.
Svensson, C., Hultgren, J., 2008. Associations between housing, management, and morbidity during rearing and subsequent first-lactation milk production of dairy cows in southwest Sweden. J. Dairy Sci. 91 (4), 1510–1518.
Svensson, C., Linder, A., Olsson, S.-O., 2006. Mortality in Swedish dairy calves and replacement heifers. J. Dairy Sci. 89 (12), 4769–4777.
Tang, E.K., Suganthan, P.N., Yao, X., 2006. An analysis of diversity measures. Machine Learning 65 (1), 247–271.
Toledo-Alvarado, H., Vazquez, A.I., de los Campos, G., Tempelman, R.J., Bittante, G., Cecchinato, A., 2018. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. J. Dairy Sci. 101 (3), 2496–2505.
Tsai, C.-F., Chen, M.-L., 2010. Credit rating by hybrid machine learning techniques. Appl. Soft Comput. 10 (2), 374–380.
van der Heide, E., Veerkamp, R., van Pelt, M., Kamphuis, C., Athanasiadis, I., Ducro, B., 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. J. Dairy Sci. 102 (10), 9409–9421.
Van Pelt, M., Meuwissen, T., de Jong, G., Veerkamp, R., 2015. Genetic analysis of longevity in Dutch dairy cattle using random regression. J. Dairy Sci. 98 (6), 4117–4130.
Warner, D., Vasseur, E., Lefebvre, D.M., Lacroix, R., 2020. A machine learning based decision aid for lameness in dairy herds using farm-based records. Comput. Electron. Agric. 169, 105193.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Cambridge.
Woźniak, M., Graña, M., Corchado, E., 2014. A survey of multiple classifier systems as hybrid systems. Info. Fusion 16, 3–17.
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC Press, Boca Raton.
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.