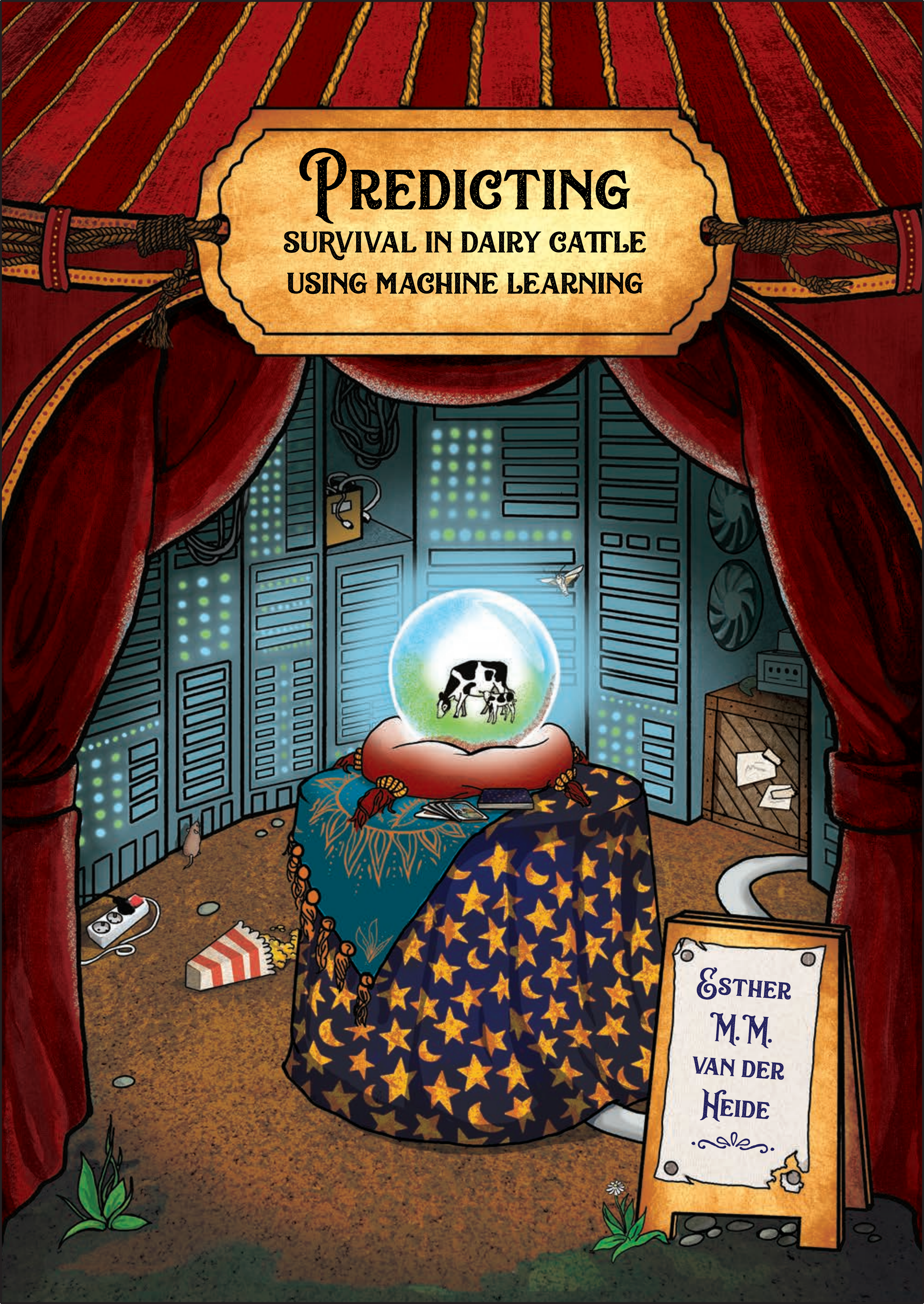
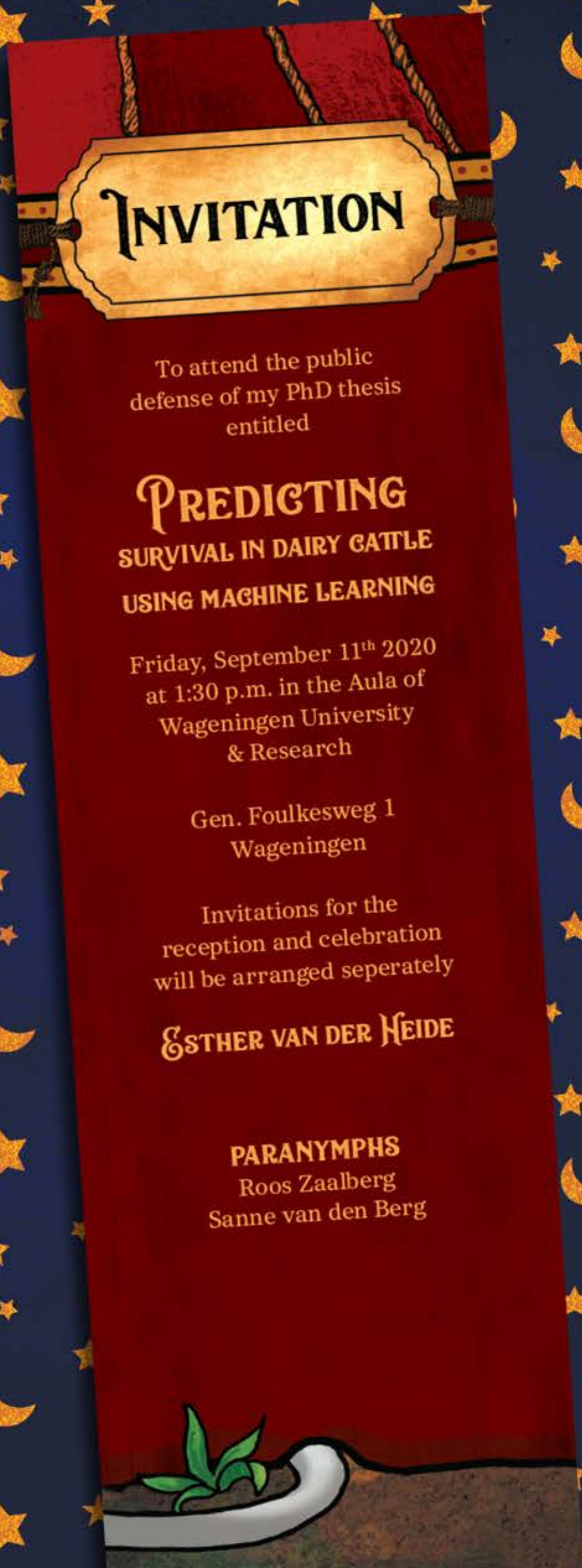
Predicting survival in dairy cattle using machine learning

Esther M.M. van der Heide





INVITATION

To attend the public
defense of my PhD thesis
entitled

PREDICTING SURVIVAL IN DAIRY CATTLE USING MACHINE LEARNING

Friday, September 11th 2020
at 1:30 p.m. in the Aula of
Wageningen University
& Research

Gen. Foulkesweg 1
Wageningen

Invitations for the
reception and celebration
will be arranged seperately

ESTHER VAN DER HEIDE

PARANYMPHS
Roos Zaalberg
Sanne van den Berg

Predicting survival in dairy cattle using machine learning

Esther M.M. van der Heide

Thesis commitee

PROMOTOR

Prof. Dr R.F. Veerkamp
Special Professor Numerical Genetics
Wageningen University & Research

CO-PROMOTORS

Dr B.J. Ducro
Assistant Professor Animal Breeding and Genomics
Wageningen University & Research

Dr C. Kamphuis
Researcher
Wageningen University and Research

OTHER MEMBERS

Prof. Dr H. Hogeveen
Wageningen University & Research

Prof. Dr T.H.E. Meuwissen
Norwegian University of Life Sciences, Ås, Norway

Prof. Dr D.J. de Koning
Swedish University of Agricultural Sciences, Uppsala, Sweden

Prof. Dr D.P. Berry
Teagasc, Cork, Ireland

This research was conducted under the auspices of the
Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Predicting survival in dairy cattle using machine learning

Esther Margaretha Maria van der Heide

THESIS

submitted in fulfillment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday September 11, 2020
at 1.30 p.m. in the Aula

Van der Heide, E.M.M.
Predicting survival in dairy cattle using machine learning

PhD thesis, Wageningen University, the Netherlands (2020)
With references, with summary in English

DOI-link: https://doi.org/10.18174/523729
ISBN: 978-94-6395-427-3

Abstract

VAN DER HEIDE. (2020). Predicting survival in dairy cattle using machine learning.

PhD thesis, Wageningen University, the Netherlands

Although cows can live to be twenty years old, the average lifespan for a dairy cow is only five to six years. Improving the lifespan of dairy cows would have several benefits such as increasing farm profitability and reducing the environmental impact of milk production. However, the complexity of survival makes it difficult to improve this trait in practice. In this thesis, I proposed using phenotypic prediction of survival to select young cows for the dairy herd, improving survival through increased lifespan of selected cows and better heifer management. The aim of this thesis was to investigate if it was possible to predict survival phenotype accurately enough to be of use in selection. I investigated three different methods to predict survival: multiple logistic regression, random forest and naive Bayes. In chapters two to four of this thesis I predicted the survival trait “survival to second lactation” using all three aforementioned methods. In chapter five, I predicted the survival trait “number of parities reached” using only the random forest method. Random forest and naive Bayes proved the best methods for predicting survival to second lactation, although predictive performance overall was low. The correlations between predictions for individual cows were much lower than expected, which indicated that the models predicted individual cows differently. Therefore, in chapter four I investigated if combining the results into an ensemble could improve predictive performance. An ensemble using multiple logistic regression resulted in the largest increase in performance, although none of the explored ensemble methods improved performance consistently across datasets. I further investigated if there was a benefit in including genomic information or a farm-specific effect. In chapter two, I investigated the benefit of combining genomic and phenotypic information. Genomic breeding values especially improved the prediction of survival early in life, with breeding values for fertility and longevity remained informative even after first calving. In chapter five I tested several different methods to include a farm effect and described the advantages and disadvantages of the various approaches. The results of this thesis provide valuable insights in the challenges of predicting survival traits and the suitability of various (machine learning) methods for the prediction of survival in dairy cattle.

Contents

൧

CHAPTER 1

General introduction

1.1 Survival in dairy cattle

In the Netherlands, cows that reach a lifetime production of 10.000 kg of milkfat and protein are publicly celebrated by the herdbook association (Drie, 2009; Zijlstra et al., 2016). After all, it is quite a feat for a cow to survive for the nine or more lactations required; the average cow in the Netherlands only survives up to three or four lactations (Zijlstra et al., 2013; Olechnowicz et al., 2016). Cows with long productive lives are valuable to farmers for several reasons (Van Pelt et al., 2016; De Vries, 2020). For example, farm rearing costs are reduced if cows survive to a higher number of lactations. This reduction in rearing costs is because both the costs for an individual cow get spread out over a longer productive life (Bach, 2011; Boulton et al., 2017) and because fewer replacement heifers have to be reared overall (Mohd Nor et al., 2014). High lactation cows also improve the average production of a farm, both because there is more opportunity to cull cows with a low milk production (Van Arendonk, 1985; De Vries, 2017) and because cows in their third or fourth lactation have a higher milk yield than younger dairy cows (Schutz et al., 1990; Lehmann et al., 2016). There are also environmental and societal benefits to cows with a high number of lactations. Cows with a high number of lactations produce less methane per kg of milk (Bell et al., 2015; Grandl et al., 2019) and indicate good animal welfare on the farm (Ortiz-Pelaez et al., 2008; Santman-Berends et al., 2014; Barkema et al., 2015). Environmental impact and animal welfare are both topics which are increasingly important to policy makers and the general public (Sandgren et al., 2009; Nyman et al., 2011; Doornewaard et al., 2018).

1.2 Definition of survival

The definition of true survival is the total lifespan of a cow, from birth until the end of its life. This can be measured in time (e.g. days or months) but also in functional units, such as the number of calvings or parities a cow has reached or the number of completed lactations. True survival is a combination of production, health and fertility, further influenced by various external factors (De Vries and Marcondes, 2020). Some survival traits do not describe the total lifespan of a cow but whether or not the cow survived to a specific point in time (such as the second lactation) (Veerkamp et al., 2001). The benefit of these binary survival traits is that cows can be included in the analysis even if their full lifespan is not known yet. Other survival traits focus on specific causes of death, rather than survival as a whole (Wright and VanRaden, 2016). An example of such a survival trait is functional survival, which is true survival corrected for milk production (Wright and VanRaden, 2016). By correcting survival for milk production, there is additional emphasis on deaths caused by poor fertility and health problems. Which survival trait is used in a study therefore depends on the purpose of an analysis and available data (Ducrocq, 1988; Holtsmark et al., 2009)

1.3 Mortality and culling

Cows usually die in one of two ways, either on farm due to a severe disease or injury, or because they were sold to slaughter. On-farm deaths are referred to as mortality, and deaths by slaughter as culling (Essl, 1998). Culling can further be broadly split into two groups: voluntary and involuntary culling. Voluntary culling is done to remove low producing cows or because the value of the carcass is worth more than the value as a dairy cow, whereas involuntary culling is done to remove cows that can no longer be productive (Hadley et al., 2006). Which factors contribute most to survival vary over the lifespan of the animal (van Pelt et al., 2012). For young heifer calves, the main causes of death are diseases, like diarrhea (Mee, 2013), and accidents (Brickell and Wathes, 2011; Compton et al., 2017). In the Netherlands, around 5.3 percent of calves kept on as replacement heifers die during their first year (Santman-Berends et al., 2019). In Europe, 12 to 14 percent of female calves do not survive to first calving (Hultgren et al., 2008; Brickell and Wathes, 2011; Raboisson et al., 2013). In this latter group, the main causes of death are also diseases and accidents in the first year, followed by a rise in culling due to infertility later in the rearing period (Wathes et al., 2008). Around two years of age, dairy cows enter their productive life. At this stage, culling becomes more prevalent than mortality (Figure 1). Most cows are culled for a combination of different reasons, for example infertility and low milk production (Fetrow et al., 2006)

1.3.1 INVOLUNTARY CULLING

Infertility is the main cause of involuntary culling in the Netherlands: between 2007 and 2012 20.9% of all cows culled were culled due to fertility problems (Zijlstra et al., 2013; Mohd Nor et al., 2014; Zijlstra et al., 2016). The other most prevalent reasons for involuntary culling are udder health problems (18,5%), lameness (15,0%) and metabolic diseases (10.4%) (Zijlstra et al., 2016; De Vries and Marcondes, 2020). These figures are similar in other countries (Hadley et al., 2006; McConnel et al., 2008; Compton et al., 2017). Many disorders also indirectly increase involuntary culling risk through negative effects on other traits. For example, both lameness (Booth et al., 2004; Dolecheck and Bewley, 2018) and metabolic diseases (Carvalho et al., 2019; Pascottini et al., 2019) are associated with reduced milk production and poorer fertility.

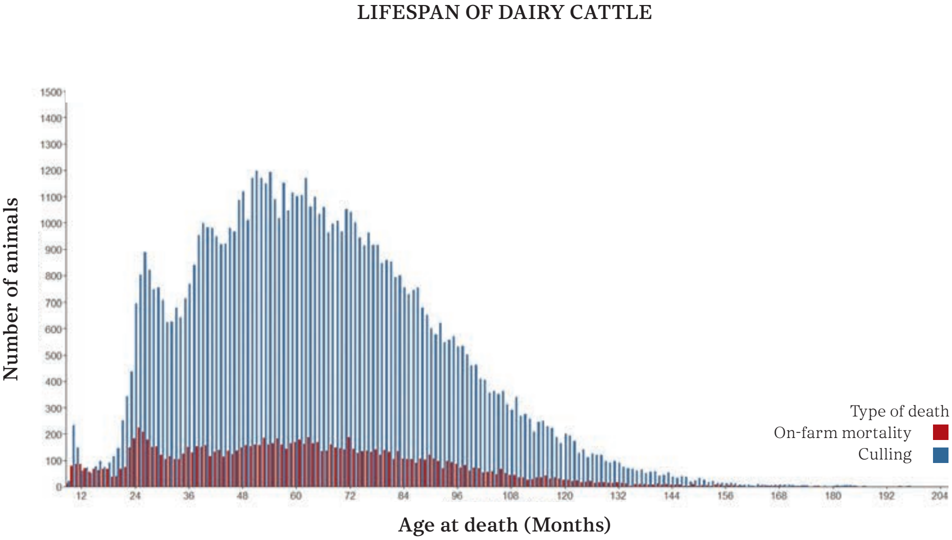


FIGURE 1.1 Age at death in months of female calves born between 2000 and 2007 by type of death. Deaths before day 300 are not shown. Ticks on the x-axis show year-intervals. Culling (the blue bars) includes both voluntary and involuntary culling

1.3.2 VOLUNTARY CULLING

Voluntary culling is the culling of cows to improve farm profitability. The main reasons reported between 2007 and 2013 that could be considered voluntary culling were 'excess' cows (4.8%), low milk production (4.6%) and old age (4.4%) (Zijlstra et al., 2016). There is most likely overlap between these different reasons for culling because only one reason for culling is reported when a cow is slaughtered (Fetrow et al., 2006; Pinedo et al., 2014; van Pelt, 2017). The main reason for a cow to be voluntarily culled is because a replacement heifer is available for which the farmers has higher expectations in terms of productivity, economic profitability or other reasons (De Vries, 2017). Farm management practices and other external factors such as government policy can also influence voluntary culling rates (Beaudeau et al., 1996; Edwards-Jones, 2006; De Vries and Marcondes, 2020). Changes in government policy like the abolishment of the milk quota system (Huettel and Jongeneel, 2011; Läpple and Sirr, 2019) and changes in the nitrogen emission regulations (Doornewaard et al., 2018) both had a strong effect on culling rates.

1.4 IMPROVING SURVIVAL

As survival is the culmination of all aspects required for a profitable, productive dairy cow, attempts have been made to improve this trait through selection (Veerkamp and van Pelt, 2019). By selecting only bulls for breeding that have daughters with long productive lives for breeding, the genetic potential for survival of the next generation is improved (Miglior et al., 2017). The genetic potential of a cow for survival is expressed as the estimated breeding value (EBV), and selection on EBVs for survival is done routinely in many countries (Forabosco et al., 2009). With the introduction of genomic testing, it is possible to directly estimate the genetic potential of a bull for a trait like survival from the genotype (Silva et al., 2014; Misztal and Legarra, 2017). Direct estimation of EBV from a genomic test is especially useful for survival traits because it can take years for a bull to have sufficient daughters with phenotypes for survival. In the Netherlands, survival has been included in the breeding goal since the 1990's. This has resulted in a positive genetic trend for survival in Dutch dairy cows (Van Pelt et al., 2016). However, despite this positive genetic trend, the average lifespan of dairy cows has changed little in recent years (Compton et al., 2017; Doornewaard et al., 2018).

The reason for the lack of phenotypic effect despite the positive genetic trend is the extensive number of external factors that influence survival (De Vries and Marcondes, 2020). This means that the genetics of a cow only explain a small portion of the variation in phenotypic survival (Wray et al., 2013). As survival is influenced by factors other than genetics, combining the genomic information with phenotypic information could result in a more accurate prediction of the phenotype for survival (Kelleher et al., 2015). Accurate prediction of survival phenotype would make it possible to select calves or heifers expected to have a long and productive life. This selection would improve survival both directly through an increase in lifespan of the selected individuals, and indirectly through a reduction in excess heifers reared. Reducing the number of excess heifers would improve survival indirectly because most excess heifers are used to replace older dairy cows early (Overton and Dhuyvetter, 2020). If survival could be estimated early on, farmers could rear a more accurate number of replacement heifers, thereby reducing the number of excess heifers and the number of older cows culled early (Mohd Nor et al., 2015). Phenotypic prediction of survival through the combination of genomic and phenotypic information could be a potential solution to improve dairy cow survival in practice.

1.5 Methods for phenotypic prediction of survival

Phenotypic prediction of a trait can be done using a variety of different methods (Libbrecht and Noble, 2015; Liakos et al., 2018). The recommended methods for the prediction of survival in animal science are regression and the proportional hazard model (Samuels et al., 2010; Lean et al., 2016). Regression is especially versatile because it can predict both binary and continuous survival traits (Wright, 1995), and can be used to estimate the effects of individual variables on the survival trait (Grömping, 2015). Regression is also the ‘traditional’ method for prediction in many fields and is therefore often used to compare novel approaches like machine learning (Dasgupta et al., 2011; Churpek et al., 2016).

Machine learning methods are a newer group of prediction methods which take advantage of the growing amount of data gathered in animal science (Halachmi and Guarino, 2016; Morota et al., 2018). Machine learning methods are data-driven, which means that they learn patterns from the data (Witten et al., 2016; Gianola et al., 2018). This allows them to take advantage of certain aspects of the data, for example non-linearity. Although this gives machine learning advantages over linear methods like regression, these advantages do not necessarily result in improved performance (Gahegan, 2003; Nayeri et al., 2019). This means it is difficult to determine beforehand which method is best suited to predict survival. The only way to determine which prediction method is best suited to a prediction problem is therefore a trial-and- error approach (Kotsiantis et al., 2007). In this thesis, we compare a 'traditional' regression method to two machine learning methods: naive Bayes and random forest to predict survival phenotypes in dairy cows. Both naive Bayes and random forest are representatives of very different groups of machine learning methods and have been used for prediction within the animal science domain. Naive Bayes has for example been used recently for mastitis detection and prediction of calving time in dairy cows (Jensen et al., 2015; Drury et al., 2017; Zehner et al., 2019) and random forest has been recently used to predict calving events and calving problems in dairy cows (Borchers et al., 2017; Zaborski et al., 2019). I will discuss these two methods in more detail in the following sub-chapters.

1.5.1 NAIVE BAYES

Naive Bayes is a machine learning method from a family of classifiers that implements Bayesian methods (Jensen, 1996; Vehtari and Ojanen, 2012). Naive Bayes is a simple approach compared to most other machine learning methods, as it assumes complete independence between the input variables (Friedman et al., 1997). Naive Bayes predicts the range of values for the variables xi given the value for the trait of interest (y). This can be mathematically described as P (xi | y). Research has shown that violating the assumption of independence does not necessarily result in poor model performance (Domingos and Pazzani, 1997; Rish, 2001). Naive Bayes achieves reasonable prediction accuracy in practice despite its simplicity and is considered one of the most efficient machine learning algorithms in terms of computing speed and resource use (Zhang, 2004; Zhang et al., 2017). This makes naive Bayes especially suited for large datasets in a practical setting.

1.5.2 RANDOM FOREST

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). A random forest makes use of decision 'trees'; a sequence of splitting rules which split the data in a way that most optimally reduces variation. Each split divides the data into two 'nodes', which contain a subset of the data (Figure 1.2). The splitting continues until some preset limit is reached. The final nodes in a tree are called 'end nodes', which contain the model's output, for example a class in classification or a numeric value in a regression. The performance of an tree can be quite poor, however, aggregating many of these trees improves the performance drastically (Biau and Scornet, 2016).

A random forest is such a collection of decision trees.

For each tree, the random forest algorithm selects a subset of records and m number of variables from the data (Breiman, 2001). From this subset, the algorithm calculates the optimal split: a value for any of the selected variables that maximally reduces the variation between the remaining two groups. The random forest continues to split the tree until a pre-set condition is reached. This could be a set tree depth (number of splits), desired accuracy of the final nodes or a minimal number of cases per end-node. The purpose of this condition is to avoid over-fitting. This is then repeated n times, where n is a predefined number of trees grown by the algorithm.

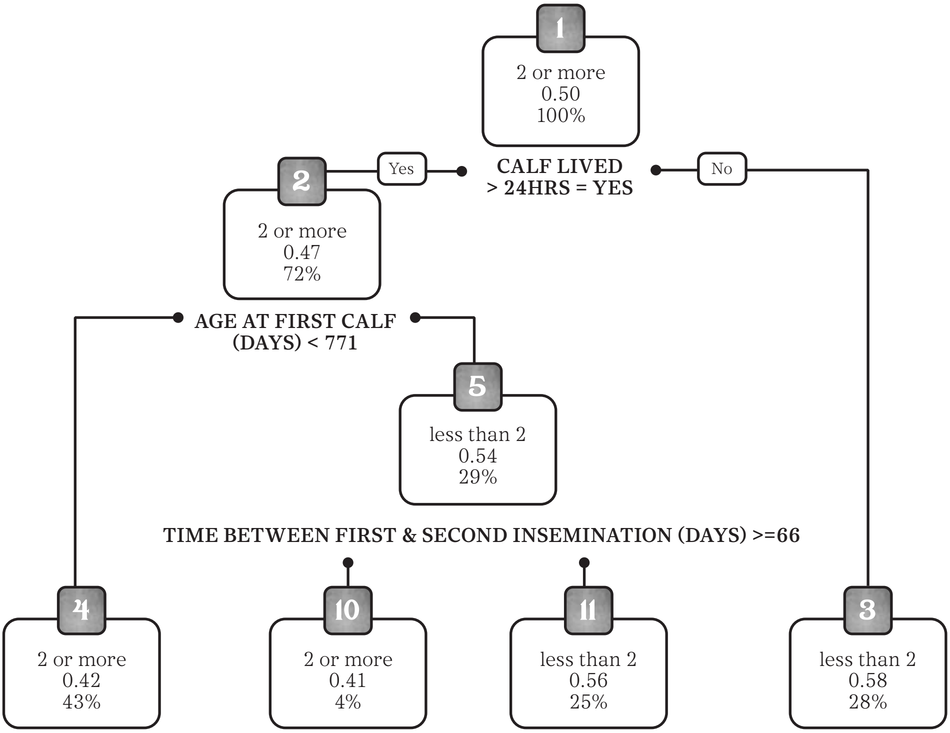


FIGURE 1.2 Schematic representation of a simple decision tree that predicts if a cow will have “less than 2” or “2 or more” lactations based on a subset of the phenotypic data available at first calving. Each node shows the proportion of correctly predicted cows and the percentage of the number of cows in the data that has been assigned to that node. For example, end node ‘3’ on the far right contains 28% of total cases. All these cases are predicted to have less than 2 lactations, and in 58% of the cases this prediction is correct. In node 1, 'no' can also indicate no data was available.

1.6 Aim and outline

The aim of this thesis was to investigate if it was possible to predict survival phenotype accurately enough to be of use in selecting young cows for the production herd. In this introduction, I described the various ways survival could be defined and which factors contribute to survival. I established that the current approach to improve dairy cow survival could be improved by incorporating phenotypic prediction. I then introduced the methods that will be used for phenotypic prediction in this thesis. In the first scientific chapter, **Chapter 2**, I predicted the phenotype survival to second lactation using multiple logistic regression. I further compared three different models: a model including only gEBV, a model including only phenotypic information and a model using both sources of information. This had three purposes: (1) establishing a baseline to compare with other methods, (2) exploring which variables contributed most to accurate phenotypic prediction of survival and (3) investigating the added value of combining phenotypic with genomic information for predicting phenotypic survival. In **Chapter 3**, I followed up on the results of chapter 2 by comparing the results of multiple regression with random forest and naive Bayes for the prediction of survival to second lactation. The results of this chapter formed the basis for **Chapter 4**, where I investigated whether combining the three methods (multiple logistic regression, naive Bayes, and random forest) into an ensemble model improved the prediction of survival to second lactation. Going a step further, in **Chapter 5** I predicted the continuous survival trait number of parities reached, rather than the binary trait survival to second lactation, using the random forest method. In this chapter I also tested different methods to include a farm effect in the models. Finally, in **Chapter 6**, I placed the scientific chapters of this thesis in broader context as I looked at the different steps required to create a model that can be applied in practice. In this chapter I discussed how to ensure the best model is chosen for a prediction problem and how to include practical relevance into the early design steps of modelling.

References

Bach, A. 2011. Associations between several aspects of heifer development and dairy cow survivability to second lactation. Journal of Dairy Science 94(2):1052-1057.

Barkema, H., M. Von Keyserlingk, J. Kastelic, T. Lam, C. Luby, J.-P. Roy, S. LeBlanc, G. Keefe, and D. Kelton. 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. Journal of dairy science 98(11):7426-7445.

Beaudeau, F., J. Van der Ploeg, B. Boileau, H. Seegers, and J. Noordhuizen. 1996. Relationships between culling criteria in dairy herds and farmers' management styles. Preventive Veterinary Medicine 25(3-4):327-342.

Bell, M., P. Garnsworthy, A. Stott, and J. Pryce. 2015. Effects of changing cow production and fitness traits on profit and greenhouse gas emissions of UK dairy systems. The Journal of Agricultural Science 153(1):138-151.

Biau, G., and E. Scornet. 2016. A random forest guided tour. Test 25(2):197-227.

Booth, C., L. Warnick, Y. Gröhn, D. Maizon, C. Guard, and D. Janssen. 2004. Effect of lameness on culling in dairy cows. Journal of dairy science 87(12):4115-4122.

Borchers, M., Y. Chang, K. Proudfoot, B. Wadsworth, A. Stone, and J. Bewley. 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. Journal of dairy science 100(7):5664-5674.

Boulton, A., J. Rushton, and D. Wathes. 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:1-9.

Breiman, L. 2001. Random forests. Machine learning 45(1):5-32.

Brickell, J., and D. Wathes. 2011. A descriptive study of the survival of Holstein-Friesian heifers through to third calving on English dairy farms. Journal of Dairy Science 94(4):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. Molecular ecology resources

Carvalho, M., F. Peñagaricano, J. Santos, T. DeVries, B. McBride, and E. Ribeiro. 2019. Long-term effects of postpartum clinical disease on milk production, reproduction, and culling of dairy cows. Journal of dairy science 102(12):11701-11717.

Churpek, M. M., T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson. 2016. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical care medicine 44(2):368.

Compton, C., C. Heuer, P. T. Thomsen, T. Carpenter, C. Phyn, and S. McDougall. 2017. Invited review: A systematic literature review and meta-analysis of mortality and culling in dairy cattle. Journal of dairy science 100(1):1-16.

Dasgupta, A., Y. V. Sun, I. R. König, J. E. Bailey-Wilson, and J. D. Malley. 2011. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genetic epidemiology 35(S1):S5-S11.

De Vries, A. 2017. Economic trade-offs between genetic improvement and longevity in dairy cattle. Journal of Dairy Science 100(5):4184-4192.

De Vries, A. 2020. Symposium review: Why revisit dairy cattle productive lifespan? Journal of Dairy Science 103(4):3838-3845.

De Vries, A., and M. Marcondes. 2020. Overview of factors affecting productive lifespan of dairy cows. animal 14(S1):s155-s164.

Dolecheck, K., and J. Bewley. 2018. Animal board invited review: Dairy cow lameness expenditures, losses and total cost. animal 12(7):1462-1474.

Domingos, P., and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning 29(2-3):103-130.

Doornewaard, G. J., J. W. Reijs, A. C. G. Beldman, J. H. Jager, and M. W. Hoogeveen. 2018. Sectorrapportage Duurzame Zuivelketen; Prestaties 2017 in perspectief No. Rapport 2018-094. Wageningen Economic Research, Wageningen.

Drie, I. v. 2009. Tientonner nummer duizend meldt zich Veeteelt. p 10-12. CRV Uitgeverij, Arnhem, the Netherlands.

Drury, B., J. Valverde-Rebaza, M.-F. Moura, and A. de Andrade Lopes. 2017. A survey of the applications of Bayesian networks in agriculture. Engineering Applications of Artificial Intelligence 65:29-42.

Ducrocq, V. P. 1988. An analyis of productive life in dairy cattle, Cornell University.

Edwards-Jones, G. 2006. Modelling farmer decision-making: concepts, progress and challenges. Animal science 82(6):783-790.

Essl, A. 1998. Longevity in dairy cattle breeding: a review. Livestock Production Science 57(1):79-89.

Fetrow, J., K. Nordlund, and H. Norman. 2006. Invited review: Culling: Nomenclature, definitions, and recommendations. Journal of dairy science 89(6):1896-1905.

Forabosco, F., J. Jakobsen, and W. Fikse. 2009. International genetic evaluation for direct longevity in dairy bulls. Journal of dairy science 92(5):2338-2347.

Friedman, N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. Machine learning 29(2-3):131-163.

Gahegan, M. 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg)? International Journal of Geographical Information Science 17(1):69-92.

Gianola, D., A. Cecchinato, H. Naya, and C.-C. Schoen. 2018. Prediction of complex traits: robust alternatives to best linear unbiased prediction. Frontiers in Genetics 9:195.

Grandl, F., M. Furger, M. Kreuzer, and M. Zehetmeier. 2019. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. Animal 13(1):198-208.

Grömping, U. 2015. Variable importance in regression models. Wiley Interdisciplinary Reviews: Computational Statistics 7(2):137-152.

Hadley, G., C. Wolf, and S. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. Journal of dairy science 89(6):2286-2296.

Halachmi, I., and M. Guarino. 2016. Precision livestock farming: a 'per animal'approach using advanced monitoring technologies. Animal 10(9):1482-1483.

Holtsmark, M., B. Heringstad, and J. Ødegård. 2009. Predictive abilities of different statistical models for analysis of survival data in dairy cattle. Journal of dairy science 92(11):5730-5738.

Huettel, S., and R. Jongeneel. 2011. How has the EU milk quota affected patterns of herd-size change? European Review of Agricultural Economics 38(4):497-527.

Hultgren, J., C. Svensson, D. O. Maizon, and P. A. Oltenacu. 2008. Rearing conditions, morbidity and breeding performance in dairy heifers in southwest Sweden. Preventive veterinary medicine 87(3):244-260.

Jensen, D. B., H. Hogeveen, and A. De Vries. 2015. Bayesian prediction of mastitis using sensor data routinely collected in dairy herds. In: The 66th Annual Meeting of the European Federation of Animal Science, Warshaw

Jensen, F. V. 1996. An introduction to Bayesian networks. UCL press, London.

Kelleher, M., P. Amer, L. Shalloo, R. Evans, T. Byrne, F. Buckley, and D. P. Berry. 2015. Development of an index to rank dairy females on expected lifetime profit. Journal of dairy science 98(6):4225-4239.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160:3-24.

Läpple, D., and G. Sirr. 2019. Dairy Intensification and Quota Abolition: A Comparative Study of Production in Ireland and the Netherlands. EuroChoices

Lean, I. J., M. C. Lucy, J. P. McNamara, B. J. Bradford, E. Block, J. M. Thomson, J. M. Morton, P. Celi, A. R. Rabiee, and J. E. Santos. 2016. Invited review: Recommendations for reporting intervention studies on reproductive performance in dairy cattle: Improving design, analysis, and interpretation of research on reproduction. Journal of dairy science 99(1):1-17.

Lehmann, J. O., J. Fadel, L. Mogensen, T. Kristensen, C. Gaillard, and E. Kebreab. 2016. Effect of calving interval and parity on milk yield per feeding day in Danish commercial dairy herds. Journal of dairy science 99(1):621-633.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. Machine learning in agriculture: A review. Sensors 18(8):2674.

Libbrecht, M. W., and W. S. Noble. 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16(6):321.

Machado, G., M. R. Mendoza, and L. G. Corbellini. 2015. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Veterinary research 46(1):85.

McConnel, C., J. Lombard, B. Wagner, and F. Garry. 2008. Evaluation of factors associated with increased dairy cow mortality on United States dairy operations. Journal of dairy science 91(4):1423-1432.

Mee, J. F. 2013. Why do so many calves die on modern dairy farms and what can we do about calf welfare in the future? Animals 3(4):1036-1057.

Miglior, F., A. Fleming, F. Malchiodi, L. F. Brito, P. Martin, and C. F. Baes. 2017. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. Journal of dairy science 100(12):10251-10271.

Misztal, I., and A. Legarra. 2017. Invited review: efficient computation strategies in genomic selection. animal 11(5):731-736.

Mohd Nor, N., W. Steeneveld, and H. Hogeveen. 2014. The average culling rate of Dutch dairy herds over the years 2007 to 2010 and its association with herd reproduction, performance and health. Journal of Dairy Research 81(1):1-8.

Mohd Nor, N., W. Steeneveld, M. Mourits, and H. Hogeveen. 2015. The optimal number of heifer calves to be reared as dairy replacements. Journal of dairy science 98(2):861-871.

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. Journal of animal science 96(4):1540-1550.

Nayeri, S., M. Sargolzaei, and D. Tulpan. 2019. A review of traditional and machine learning methods applied to animal breeding. Animal Health Research Reviews 20(1):31-46.

Nyman, A.-K., A. Lindberg, and C. H. Sandgren. 2011. Can pre-collected register data be used to identify dairy herds with good cattle welfare? In: Acta Veterinaria Scandinavica. p S8.

Olechnowicz, J., P. Kneblewski, J. Jaśkowski, and J. Włodarek. 2016. Effect of selected factors on longevity in cattle: a review. J. Anim. Plant Sci 26:1533-1541.

Ortiz-Pelaez, A., D. Pritchard, D. Pfeiffer, E. Jones, P. Honeyman, and J. Mawdsley. 2008. Calf mortality as a welfare indicator on British cattle farms. The Veterinary Journal 176(2):177-181.

Overton, M., and K. Dhuyvetter. 2020. Symposium review: An abundance of replacement heifers: What is the economic impact of raising more than are needed? Journal of Dairy Science

Pascottini, O., M. Probo, S. LeBlanc, G. Opsomer, and M. Hostens. 2019. 96 Association between metabolic diseases and fertility of high-yielding dairy cows in a transition management facility using survival analysis and machine-learning models. Reproduction, Fertility and Development 31(1):174-174.

Pinedo, P., A. Daniels, J. Shumaker, and A. De Vries. 2014. Dynamics of culling for Jersey, Holstein, and Jersey× Holstein crossbred cows in large multibreed dairy herds. Journal of dairy science 97(5):2886-2895.

Raboisson, D., F. Delor, E. Cahuzac, C. Gendre, P. Sans, and G. Allaire. 2013. Perinatal, neonatal, and rearing period mortality of dairy calves and replacement heifers in France. Journal of Dairy Science 96(5):2913-2924.

Rish, I. 2001. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. p 41-46.

Samuels, M. L., J. A. Witmer, and A. Schaffner. 2010. Statistics for the life sciences. Pearson education.

Sandgren, C. H., A. Lindberg, and L. Keeling. 2009. Using a national dairy database to identify herds with poor welfare. Animal Welfare 18(4):523-532.

Santman-Berends, I., M. Buddiger, A. Smolenaars, C. Steuten, C. Roos, A. Van Erp, and G. Van Schaik. 2014. A multidisciplinary approach to determine factors associated with calf rearing practices and calf mortality in dairy herds. Preventive veterinary medicine 117(2):375-387.

Santman-Berends, I., Y. Schukken, and G. van Schaik. 2019. Quantifying calf mortality on dairy farms: Challenges and solutions. Journal of dairy science

Schutz, M. M., L. B. Hansen, G. Steuernagel, and A. Kuck. 1990. Variation of milk, fat, protein, and somatic cells for dairy cattle. Journal of Dairy Science 73(2):484-493.

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. Journal of dairy science 97(2):731-742.

Silva, M. V., D. J. dos Santos, S. A. Boison, A. T. Utsunomiya, A. S. Carmo, T. S. Sonstegard, J. B. Cole, and C. P. Van Tassell. 2014. The development of genomics applied to dairy breeding. Livestock science 166:66-75.

Van Arendonk, J. 1985. Studies on the replacement policies in dairy cattle. II. Optimum policy and influence of changes in production and prices. Livestock Production Science 13(2):101-121.

van Pelt, M. 2017. Genetic improvement of longevity in dairy cows. Wageningen University.

Van Pelt, M., V. Ducrocq, G. De Jong, M. Calus, and R. Veerkamp. 2016. Genetic changes of survival traits over the past 25 yr in Dutch dairy cattle. Journal of dairy science 99(12):9810-9819.

van Pelt, M., H. Eding, P. Vessies, and G. de Jong. 2012. Developing a genetic evaluation for calf survival during rearing in The Netherlands. Interbull Bulletin (46)

Veerkamp, R., S. Brotherstone, and B. Engel. 2001. Analysis of censored survival data using random regression models. Animal Science 72(1):1-10.

Veerkamp, R., and M. van Pelt. 2019. Advances in dairy cattle breeding to improve longevity, Advances in breeding of dairy cattle. Burleigh Dodds Science Publishing Limited. p. 337-354.

Vehtari, A., and J. Ojanen. 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys 6:142-228.

Wathes, D., J. Brickell, N. Bourne, A. Swali, and Z. Cheng. 2008. Factors influencing heifer survival and fertility on commercial dairy farms. Animal 2(8):1135-1143.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Cambridge.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. 2013. Pitfalls of predicting complex traits from SNPs. Nature reviews. Genetics 14(7):507.

Wright, J., and P. VanRaden. 2016. Genetic evaluation of dairy cow livability. Journal of Animal Science 94:178-178.

Wright, R. E. 1995. Logistic regression.

Zaborski, D., W. S. Proskura, W. Grzesiak, J. Różańska-Zawieja, and Z. Sobek. 2019. The comparison between random forest and boosted trees for dystocia detection in dairy cows. Computers and Electronics in Agriculture 163:104856.

Zehner, N., J. J. Niederhauser, M. Schick, and C. Umstatter. 2019. Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. Computers and electronics in agriculture 161:62-71.

Zhang, C., C. Liu, X. Zhang, and G. Almpanidis. 2017. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications 82:128-150.

Zhang, H. 2004. The optimality of naive Bayes. AA 1(2):3.

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.

Zijlstra, J., M. Jiayang, C. Zhijun, and J. van der Fels. 2016. Longevity and culling rate: how to improve?, Wageningen UR Livestock Research, Wageningen, the Netherlands.

II

CHAPTER 2

Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information

E.M.M. van der Heide[1], R. F. Veerkamp[1],
M. L. van Pelt[2], C. Kamphuis[1], B. J. Ducro[1]

Journal of Dairy science (2020) 103.1: 556-571
DOI: S0022030219309580

[1]Wageningen University & Research Animal Breeding and Genomics,
P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]CRV BV, Animal Evaluation Unit,
P.O. box 454,6800 AL Arnhem, the Netherlands

Abstract

Advances in technology and improved data collection have increased the availability of genomic breeding values (gEBV) and phenotypic information on dairy farms. This information could be used for the prediction of complex traits such as survival, which can in turn be used in replacement heifer management. In this study, we investigated which gEBV and phenotypic variables are of use in the prediction of survival. Survival was defined as “survival to second lactation, plus two weeks”, a binary trait. A dataset was obtained of 6847 heifers that were all genotyped at birth. Each heifer had 50 gEBV and up to 62 phenotypic variables that became gradually available over time. Stepwise variable selection on 70% of the data was used to create multiple regression models to predict survival with data available at five ‘decision moments’; distinct points in the life of a heifer at which new phenotypic information becomes available. The remaining 30% of the data were kept apart to investigate predictive performance of the models on independent data. A combination of gEBV and phenotypic variables always resulted in the model with the highest AIC value. The gEBV selected were longevity, feet and leg score, exterior score, udder score and udder health score. Phenotypic variables on fertility, age at first calving and milk quantity were important once available. It was impossible to predict individual survival accurately, but the mean predicted probability of survival of the surviving heifers was always higher than the mean predicted probability of the non-surviving group (difference ranged from 0.014 to 0.028). The model obtained 2.0% to 3.0% more surviving heifers when the highest scoring 50% of heifers were selected compared to randomly selected heifers. Combining phenotypic information and gEBV always resulted in the highest scoring models for the prediction of survival, and especially improved early predictive performance. By selecting the heifers with the highest predicted probability of survival, increased survival could be realized at the population level in practice.

KEY WORDS: dairy cow, survival, longevity, individual prediction

2.1 Introduction

Optimally, not all female calves born on a farm should be kept as replacements heifers to avoid unnecessary costs (Mohd Nor et al., 2012; Mohd Nor et al., 2015). In order to have insight in these costs, economic models have been developed to help farmers make the best choices in terms of heifer management (Mourits et al., 1997; Groenendaal et al., 2004). However, despite the availability of these models, many Dutch farmers keep more than the optimal number of replacement heifers (Mourits et al., 2000; Mohd Nor et al., 2015). Uncertainty about the survival and future performance of replacement heifers and dairy cows is one of the reasons for a farmer to keep on a surplus of replacement heifers (Mohd Nor et al., 2015). Studies in the United Kingdom, France and Sweden show that between 86% and 88% of heifer calves reach their first lactation (Hultgren et al., 2008; Brickell and Wathes, 2011; Raboisson et al., 2013), and between 83% and 96% of all first lactation heifers survive to their second lactation (Dechow and Goodling, 2008; Bach, 2011; Brickell and Wathes, 2011). In the Netherlands, 86.6% of first lactation heifers born between 2009 and 2013 reached their second lactation (Van Pelt et al., 2016b). Because it takes on average around 1,5 lactations to repay the rearing costs of a dairy cow, reducing the number of replacement heifers raised is important for farm profitability (Bach, 2011; Boulton et al., 2017). Surplus heifers may be sold, but this does not always cover the rearing costs (Mohd Nor et al., 2015), or may be used to replace older dairy cows that are voluntarily culled. However, because resilience and longevity of the dairy herd are becoming increasingly important from a societal and welfare point of view (Ortiz-Pelaez et al., 2008; LTO, 2011; Mohd Nor et al., 2014; Barkema et al., 2015), it may become preferable to not cull older dairy cows simply because a younger replacement is available.

Despite the importance of survival, it is rarely included in heifer management models and even when survival is taken into account, there is no consideration for individual differences between heifers (Mohd Nor et al., 2015). Individual prediction of survival in heifers is often not attempted because of the difficulty of predicting phenotypic survival accurately. Survival traits are affected by a combination of production, fertility and health traits (Heise et al., 2016), and environmental factors such as farm management. Furthermore, risk of culling is not constant over time. The complex nature of survival and the fact that true survival can only be measured in retrospect (the animal is culled or has died) has also led to many different definitions of survival traits being used (Essl, 1998; Fetrow et al., 2006). Estimated breeding values (EBV) for at least one survival trait are available in most countries for breeding purposes (Forabosco et al., 2009). However, the heritability of these traits is almost invariably very low (Van Pelt et al., 2015; Heise et al., 2016), making the phenotypic prediction of individual survival based on parent information alone not very accurate.

Combining available genomic breeding values (gEBV) and phenotypic information might yield more accurate predictions for individual survival. Genotyping costs in dairy cattle have reduced to the point where it is economically feasible to obtain gEBV for production animals at birth (Pryce et al., 2012; Weigel et al., 2012; Calus et al., 2015). More phenotypic information on individual animals is also available on farms. The aim of this study is to investigate the possibility of combining phenotypic information and gEBV for the phenotypic prediction of survival. Five distinct moments were chosen for prediction of survival. These "decision moments" were moments in the lifetime of a dairy cow where new phenotypic information becomes available on which a management decision could be made.

2.2 Materials and Method

2.2.1 DATA

A dataset was obtained from cattle improvement cooperative CRV (Arnhem, the Netherlands) to investigate which gEBV and phenotypic variables could serve as predictors for survival. The data was available from animals born on 463 different Dutch and Flemish farms that participate in a 'data plus' program, where additional information is gathered on dairy cows on commercial farms. All animals in this study were genotyped at birth, herd book registered and had at least 87.5% Holstein blood. Survival was defined as the binary trait "survival until second calving, plus 2 weeks", henceforth referred to as 'survival to second lactation'. This point in time was selected due to limitations of our data and because the lactation following second calving is economically significant (Bach, 2011). Two additional weeks were included to avoid counting animals that died during or as a consequence of their second calving. In order to have a known observation for survival, all cows included had to be born at least 46 months prior to the end of data collection (March 2017), and were not exported abroad during this time. Animals exported abroad were excluded as they had unknown dates of death and could not be used in the analysis. Animals sold to other Dutch farms could be used, because records from other farms were available in our dataset. As the cause of death was unknown for all heifers, all deaths were included. This included deaths on farm, involuntary culling and voluntary culling. Twenty-four animals were removed from the dataset due to having only second calving records available. The final dataset consisted of 6847 female cows born between January 2012, and June 2013. Out of these cows, 5872 (85.8%) cows survived until second lactation and 975 (14.2%) cows did not.

Five different datasets were created to predict survival with data available at five different "decision moments" during the life of a cow: at birth, at 18 months of age, at first calving, six weeks post calving and at two hundred days post calving. These moments were selected as they were points in time when more phenotypic information becomes available and where relevant decisions related to survival could be made. This includes not only the decision to cull or not cull an animal, but for example also if an animal should be inseminated again or not. The decision not to inseminate a heifer is effectively an early culling decision, as without getting pregnant it is not possible to enter the next lactation, even if the heifer is kept in the herd for another few months. Apendix 2.A shows all variables, as well as the decision moment in which they become available, and if the variables were originally continuous. Information was cumulative; all records available on the first decision moment were also available during the subsequent decision moments.

Each animal had 50 gEBV, scaled to a value between 0 and 10 where the largest value was set to 10 and the smallest value set to 0. These gEBV are direct genomic values, which did not include any own performance. If regular gEBV were used, own information included would have led to a residual covariance between the gEBV and survival. Using this regular gEBV in a survival prediction would have led to auto-correlation. Because historical gEBV for the various decision moments were not available, only gEBV based solely on genotype were used. Phenotypic records were available on gestation duration and dam parity, herd book status, birth records, calving records, records on moves between different farms, insemination records of the first and second parity and first parity milk records. Calving ease scores were scored 1 to 6, where 1 was an easy birth, 2 was normal, and scores 3 to 6 were considered 'difficult' births, because they denoted long labor and various veterinarian interventions. The milk records at 'six weeks post calving' were of the most recent milk test day record before the end of six weeks. Milk records in the decision moment '200 days post calving' were an average of all milk records available of an animal up to that point. Fertility records were used to determine non-return status at the decision moment '18 months of age' and '200 days post calving'. If a heifer had AI insemination records, but received no AI insemination in the 56 days prior to the two aforementioned decision moments, it was listed as non-return, because it had not received another insemination at least 56 days after the last insemination. Heifers with an insemination in the 56 days prior to the decision moment, no recorded inseminations at that point in time or subject to natural mating were listed as unknown. In these cases the exact insemination or conception date was unknown or it could not otherwise be determined if the animal was pregnant.

Animals with missing records were always included in the datasets. Missing records could have reflected active management decisions (for example not inseminating an animal) or could reflect, for example, a fertility problem, and thus could have been useful in prediction. This was modelled by adding a class "Unknown" to all factorial variables to identify missing values. Continuous variables with missing values were transformed into factors with between 5 and 8 classes, depending on the distribution of individual variables. We chose to select at least five levels to keep sufficient variation within each variable. In total, each animal had 50 gEBV (available at each decision moment), and up to 62 additional phenotypic variables accumulating throughout their lifetime.

2.2.2 MODEL AND ANALYSIS

To determine if survival to second lactation could be predicted, logistic regression models were constructed for each decision moment. Each of the five datasets (one for each decision moment) were split into sets of 70% training and 30% testing data, stratified by survival group to ensure a representative amount of both survival groups in the testing and training sets. Stratified sampling meant that the training and the testing datasets included identical proportions of non-surviving and surviving heifers. We used the statistical program R (R core team, 2016), version 3.3.1, and the package 'caret' (Kuhn, 2008) to select models for each decision moment. Both forward stepwise selection and stepwise selection combining forward and backward stepwise selection were tested on the five training datasets using the following general model:

$$\text{Logit (P)} = \beta_0 + \sum \beta_i \, X_i$$

Where, Logit (P) is the estimated probability of survival, β0 is the population mean and βiXi is the set of predictor variables, consisting of phenotypic variables and gEBV. The Akaike information criterion (AIC) was used to determine the best possible model for each decision moment. The stepwise procedure combining forward and backward selection resulted in the models with the lowest AIC value and thus only the models derived using this procedure were used for analysis. The stepwise procedure was first used on the training sets containing both gEBV and phenotypic variables to identify which variables are significant for the prediction of survival. The models were tested on their corresponding test dataset to get probabilities of survival for the heifers in the test set. By applying the model to the data, predicted probabilities of survival were obtained for each heifer. These were values between 0 and 1, where values close to 1 indicate a high probability to reach the second lactation, and values closer to 0 a lower probability to reach second lactation. To compare and validate the models, the prediction accuracy, specificity, sensitivity, positive predictive value and negative predictive value were calculated. The predicted probability of survival was transformed into a survival prediction of 1 when equal to or above the average probability for survival in our data (0.858), and to a survival prediction of 0 below the average probability of survival. The accuracy was the proportion of correct predictions. The sensitivity was calculated as the true positive outcomes divided by the true positives plus the false negatives, and the specificity as the true negatives divided by the true negatives plus the false positives. The positive predictive values is the true positive value divided by the true positive value plus the false positive value, and similarly the negative predictive value is the true negative value divided by the true negative value plus the false negative value. We also calculated the balanced accuracy and the area under the receiver operating (ROC) curve, the area under the curve (AUC) value. The balanced accuracy is the average of the accuracy for non-surviving heifers and the accuracy of the surviving heifers. The AUC value is the predictive ability of the model including all possible cut-offs and was calculated using the pROC package (Robin et al., 2011). These metrics were selected as survival was an imbalanced trait, with more survivors than non-survivors. Both balanced accuracy and AUC value are more robust against imbalanced predictors than accuracy.

In order to investigate if there was merit in including both gEBV and phenotypic traits, we also selected models for each decision moment using either only gEBVs or only phenotypic information as input for the variable selection. For each decision moment, the three resulting models were tested on a new 70/30 split of training and testing data. The AIC was recorded as indication of model fit and the AUC value calculated using pROC as indication of model performance.

2.3 Results

Figure 1 shows the distribution of age in days at death for heifers in our dataset. This figure indicates that in our dataset most heifers reached 18 months of age, with few early deaths. The number of culled and dead animals increased after 18 months, most due to the availability of phenotypic records on which selective culling could take place.

An overview of the gEBV and phenotypic variables selected using stepwise selection for each decision moment when including only living heifers at a decision moment is shown in Tables 2.1 and 2.2. The coefficients of the selected variables are shown in Appendix 2.B, separated by decision moment. The gEBV for longevity was selected at all decision moments, and was in each case positively associated with survival. While different gEBV were selected at different decision moments, at each decision moment they broadly fell into the same categories: fertility, exterior score, udder score, udder health and feet and leg gEBV. The gEBV for production variables were not prominent among the selected variables, although they were selected at the first three decision moments.

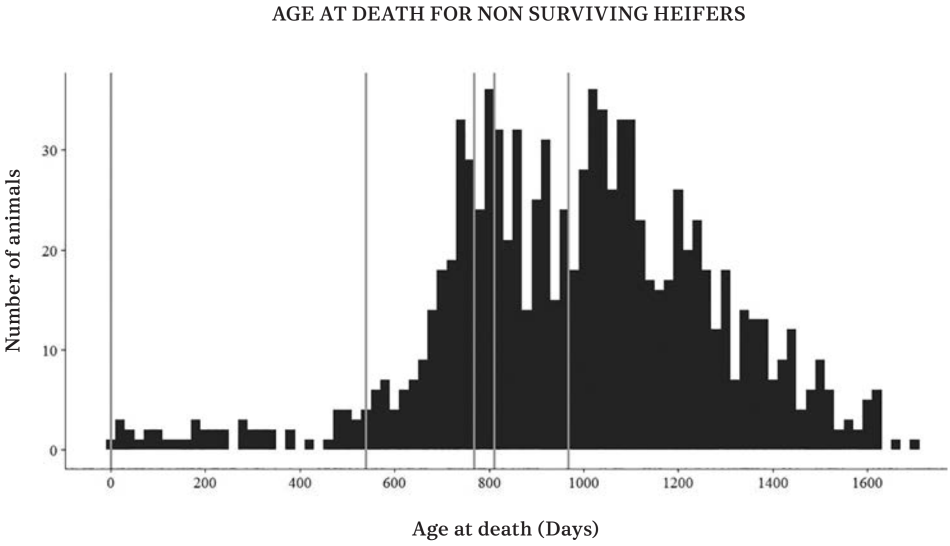


FIGURE 2.1 The distribution of age at death in days for non-surviving heifers. Grey lines indicate the decision moments. The decision moment at first calving was set at the average age at first calving in the dataset.

TABLE 2.1 Estimates for coefficients* of selected gEBV per decision moment

gEBV	Birth	18 months	First calving	6 weeks post calving	200 days post first calving
Longevity	0.128	0.212	0.261	0.242	0.207
Foot angle	-0.115	-0.078	-0.086	-0.085	
Udder depth	-0.102	-0.137	-0.140		-0.106
Frame	0.074				
Kg fat	0.061	0.122			
Non return at 56 days	-0.092			-0.088	
Interval first-last insemination	0.140			-0.213	0.106
Overall fertility		0.145	0.091	0.242	
Milking speed		0.061			
NVI		-0.122			
Overall exterior score		0.143			0.256
Somatic cell count			0.196	0.212	
Udder health			-0.283	-0.289	-0.153
Rear udder height			0.072		
Feet and legs			-0.066		
Chest width			0.099	0.131	
Rear legs hind view				-0.162	
Locomotion				0.125	
Kg Lactose				-0.073	
Stature					-0.094
Udder support					0.074

* For the exact coefficients for each variable, see Appendix 2.B.

At 200 days post calving, several gEBV regarding longevity, exterior score, udder conformation and health and fertility were still selected. The phenotypic variables selected at birth were season of birth and year of birth. Year of birth and season of birth were selected up to the last decision moment, where season of first calving was selected as alternative for season of birth. Phenotypic information new to a decision moments was always selected, with some variables remaining important over the next decision moment(s). Age at first calving, for example, was strongly associated with survival in all decision moments after first calving (Appendix 2.B3 to B5). Earlier calving ages had a more positive association with survival. Phenotypic fertility information appears important in general, as non-return status and number of inseminations were selected in both decision moments when these variables became available (Appendices B2 and B5). At 200 days post calving (abbreviated in the tables to p.c.), both the number of inseminations at 18 months and at 200 days post calving were selected. Phenotypic information on production traits was also important: kg of milk produced at the milk test day closest to 6 week post calving and the average milk production per test milk day at 200 days post calving were both significantly associated with survival (Appendix 2.B4 and B5). Furthermore, the average percentage of protein per test milk day was selected at 200 days post calving. Phenotypic traits on udder health were also selected as number of negative outcomes, which include mastitis and other illnesses, negative indication (Yes/No) at test milking closest to 6 weeks post calving and average cell count per test milk day at 200 days post calving.

In order to determine if adding genotype information has additional value, we compared a model using only genotypes, a model using only phenotypes and the results of the combined model (Table 2.3). The combined model has the highest AUC value at all decision moments, except at first calving where all methods perform equally, and at 200 days post calving, where it performs equal to the model containing only phenotypic variables. The AIC value was always highest for the combined model, as a model containing both gEBV and phenotypic variables was always selected through stepwise selection (Tables 2.1 and 2.2).

On average, the predicted probability for the surviving group was between 0.014 (at birth) and 0.028 (at 200 days post calving) higher than the probability of survival for the non-surviving group (Table 2.4). This means that while there was overlap, there was a difference between the two groups on average.

The accuracy of the models increased from 0.562 to 0.776 in later decision moments (Table 2.5). The AUC values also increased, from 0.578 at birth to 0.648 at 200 days post calving. This means that the model improved at later decision moments. However, the balanced accuracy did not increase, remaining around 0.56 at all decision moments. It appears that the models improved by predicting surviving heifers better. This can be seen from an increase in the sensitivity of the models from 0.566 at birth to 0.816 at 200 days post calving and an increase in positive predictive value from 0.880 to 0.933. This means that both a larger proportion of all surviving heifers in the dataset were predicted correctly as surviving, and a larger

TABLE 2.2 Estimates for coefficients* of selected phenotypic variables per decision moment.

Phenotypic variables	Birth	18 months	First calving	6 weeks post calving	200 days post calving
Season of birth	0.434	0.474	0.423	0.600	
Year of birth	-0.372	-0.268	-0.270	-0.342	-0.239
Number of inseminations at 18 months		1.305			1.553
Coat color			-0.208		
Non return status at 18 months		-0.426			
Season of first calving					0.725
Age in days at first calving			1.700	1.881	2.831
Sex of first calf			0.378	0.272	
Calf survival one week after birth					14.392
Kg of milk produced at milk test day closest to 6 week post calving				2.512	
Total number of transports at 200 days post calving					15.215
Average kg of milk per test milk day at 200 days post calving					1.899
Average cell count (x1000) per test milk day at 200 days post calving					1.032
Average percentage of protein per test milk day at 200 days post calving					1.414
Negative indication at test milking 6 weeks post calving					1.401
Number of negative indications at test milk days before 200 days post calving					-0.788
Number of inseminations at 200 days post calving					1.553
Non return status at 200 days post calving					-1.065

*This table shows the difference in coefficient between the highest and the lowest class for a variable. The coefficient of the reference class is 0. The maximum difference was used because class variables have more than one coefficient. For the exact coefficients for each variable, see Appendix 2.B.

proportion of the heifers predicted to survive actually survive. In contrast, while the model was able to better predict which heifers survived, predicting which heifers did not survive proved more difficult. The negative predictive value did not increase consistently, and was actually lowest at 200 days post calving. Furthermore, the specificity of the model varied inconsistently, ranging from 0.295 to 0.534, meaning that the proportion of non-surviving heifers that is identified did not improve.

Rather than focusing on individual prediction, a situation could be considered where the models were used to select 50% of heifer calves to become replacement

TABLE 2.3 The AUC of the ROC value for a model*

Decision moment	gEBV only	Phenotype only	gEBV and Phenotype
Birth	0.557	0.562	0.584
18 months	0.580	0.594	0.606
At first calving	0.597	0.597	0.596
6 weeks p.c.	0.560	0.646	0.677
200 days p.c.	0.573	0.731	0.731

*AUC: area under the curve, ROC: receiver operator curve, p.c.= post calving. The model in this table includes both gEBV and phenotypic information, only phenotype information and only gEBV.

TABLE 2.4 Predicted probability of survival for the survival and the non-survival heifer groups in the testing set of each decision moment*

Decision moment	Survival = No		Survival = Yes	
	Mean	SD	Mean	SD
Birth	0.847	0.049	0.861	0.044
18 months	0.842	0.075	0.868	0.060
At first calving	0.873	0.073	0.897	0.052
6 weeks p.c.	0.869	0.072	0.896	0.052
200 days p.c.	0.867	0.064	0.896	0.054

*Includes only alive heifers at the start of each decision moment, SD: standard deviation, p.c.= post calving.

heifers. Table 2.6 shows the average probability of survival in a decision moment and the probability of survival of the selected group. For example, in the first decision moment, 85.8% of calves reached second lactation in a random selection, compared to 88.6% of the calves selected through the model. This meant that out of the 1026 selected calves in our testing dataset, 909 of heifers selected by our model would reach second lactation, compared to 880 random heifers. In subsequent decision moments, using a model resulted in up to 3% more surviving heifers compared to random selection.

TABLE 2.5 **Performance metrics of the models at each decision moment.**

Decision moment	Accuracy	Balanced Accuracy	Specificity	Sensitivity	Neg. pred. Value	Pos. pred. Value	AUC
Birth	0.562	0.550	0.534	0.566	0.170	0.880	0.578
18 months	0.634	0.577	0.498	0.656	0.186	0.893	0.606
At first calving	0.756	0.545	0.276	0.813	0.150	0.904	0.594
6 weeks p.c.	0.755	0.567	0.329	0.806	0.168	0.910	0.620
200 days p.c.	0.776	0.555	0.295	0.816	0.117	0.933	0.648

*AUC: area under the curve, p.c.= post calving. The model in this table includes both gEBV and phenotypic information, only phenotype information and only gEBV

TABLE 2.6 **The average chance of survival at a decision moment*,** and the difference in percentage between the two.

Decision moment	Probability of survival (%)		Difference (%)
	Mean	Predicted top 50%	
Birth	85.8	88.6	2.8
18 months	86.4	89.4	3.0
At first calving	89.0	91.4	2.4
6 weeks p.c.	90.4	92.4	2.0
200 days p.c.	92.3	95.3	3.0

*The average chance of survival for the animals in the top 50% predicted probabilities of survival, p.c.= post calving.

2.4 Discussion

Our research showed that it was possible to predict the survival outcome of heifers at population level. Predicting non-surviving animals proved difficult even with large amounts of phenotypes and gEBV available. This was not unexpected because a cow could have been culled for a myriad of reasons that often influence each other, and the decision to cull is time-dependent and based on decisions made by individual farmers (Hadley et al., 2006; Zijlstra et al., 2013). Our dataset also included unpredictable causes of death such as random accidents (Brickell and Wathes, 2011). Prediction may have been improved by including the exact causes of death or culling, for example allowing us to remove deaths caused by accidents, but this information was not available.

The difficulty of predicting survival meant that individual predictions obtained by our models were too inaccurate for the purpose of identifying non-surviving heifers. Because the negative predictive value ranged from 0.12 to 0.17, any heifer predicted as not surviving only had a 12 to 17% chance of not reaching second lactation. However, if applied on a large group of individuals, it was possible to use these models to select the heifers with the highest probability of survival. When 50% of the heifers with the highest probability of survival were selected, 2.0 to 3.0% more heifers reached second lactation compared to random selection. When selecting at birth, this would have resulted in a 2.8% increase in surviving heifers. While a 2.8% increase does not seem like a large improvement, this represented a 15.5% reduction of non-surviving heifers.

The selected variables gave insight on which variables were associated with survival at the various decision moments. The first variables selected were birth season and year of birth. Because these variables were cohort variables, these variables were not animal-specific and did not distinguish between calves born in the same season and year. This meant they could not be used for an individual farmer to distinguish between calves born in the same season or year. Birth season and year are important correction factors at the population level, however, and were selected in all but the last decision moment, where calving season was used instead of season of birth. Production gEBV were surprisingly not selected at every decision moment. This may be explained by the fact that longevity was uncorrected for production, and thus possibly served as a substitute for production gEBV. Phenotypic production traits such as (average) kg of milk produced were selected once available, because there is a known increased risk of culling for animals with low milk production (Hadley et al., 2006). Most of the other associations found were not surprising. The association between feet and leg traits and survival is well known in literature

(Buenger et al., 2001; Caraviello et al., 2004). The gEBV for foot angle was associated with lameness specifically (Wells et al., 1993), which is an important reason for culling (Olechnowicz and Jaskowski, 2011; Zijlstra et al., 2013). Udder conformation traits (Caraviello et al., 2004; Kern et al., 2015) and udder health (Mohd Nor et al., 2014) are known to be strongly associated with survival. Udder and health traits are correlated (Carlström et al., 2016), and udder traits have even been suggested as candidate traits for indirect prediction of longevity (Kern et al., 2015). Udder health phenotype variables were selected in the form of negative indication counts and somatic cell count, which is associated with increased risk of culling (Beaudeau et al., 2000). The exterior score gEBV selected in this study were mostly related to size. Body size has been associated with longevity and efficiency (Getu and Misganaw, 2015; Kern et al., 2015), however the direction of the associations found varied, and some studies find no effect at all (Sewalem et al., 2004). Even when studies find negative associations, culling decisions may be influenced by a (regional) farmer preference for larger cows (Hansen et al., 1999; Caraviello et al., 2004). Lastly, fertility had been reported as one of the most important reason for culling in Dutch Holstein cows (Zijlstra et al., 2013), and there is a well-documented relation between fertility and survival (Pritchard et al., 2013). Age at first calving, which is also an indication of fertility, was always selected when available. A higher age at first calving was associated with higher costs, lower fertility, and higher risk of culling overall (Sewalem et al., 2008; Van Pelt et al., 2016a). Interestingly, despite the availability of two parities worth of phenotypic fertility records, fertility gEBV were still selected by the combined model in the fifth decision moment.

The selection criteria for our data were very stringent because they included only animals genotyped at birth. This was done to avoid a winners' bias; genotyping was more expensive previously, and so only promising heifers and proven cows were genotyped. This means that cows that were genotyped either already had reached second lactation, or were more likely to reach the second lactation than an average cow in the population. Less than 7% of heifers do not reach the second lactation if we included all genotyped cows in our dataset. As we did not include all genotyped heifers, the heifers in our dataset were limited to those born within a period of around a year and a half. This meant it was impossible to investigate true survival or beyond the fourth lactation, since these lactations would require an opportunity group born before the heifers available in our dataset. The limited number of heifers also explains why some variables like 'number of movements' and 'calf-survival at two weeks' were selected. Both variables had very strong associations with survival for some classes (Appendix 2.B5). These classes had less than 10 heifers each, all of which survived by chance. All heifers of a class surviving resulted in a strong association with survival for these variables. In future studies, classes with very few cases could be excluded or merged with other classes, if they prove problematic for the analysis. Increasing the amount of animals in the dataset should also reduce the number of very small classes. In this study, as survival is difficult to predict and information was already limited on some causes of death, we chose not merge or remove classes, resulting in some artifacts. Another effect of the limited number of heifers was that many farms only had a small number of calves. The average number of calves per farm was 15, with many farms having fewer calves and some farms having a much larger number of calves. Farm identification numbers (UBN) could have functioned as a proxy for herd and farm management, which were not included in the data. However, farm UBN's were not selected due to the large number of farms with very few calves. Next to requiring all heifers to be genotyped, this study included only phenotypic information that was readily available. This meant that information gaps still existed at each decision moment. For example between birth and 18 months of age only fertility variables and movements between farms were recorded on a large scale. Variables such as the occurrence of diseases (Svensson and Hultgren, 2008; Heinrichs and Heinrichs, 2011) or simple body(size) measurements of calves (Wathes et al., 2008) could have provided valuable information on the health and development of the animal during the first 18 months of life. Weight or size of the heifer may also be useful for predicting fertility, because while earlier first calving ages seemed preferable, it may be beneficial for some heifers to be inseminated later as early inseminations may have detrimental effects (Hoffman and Funk, 1992; Heinrichs, 1993).

Because a multiple regression was used in this study, the direction and the strength of the association of the variables with survival was only valid in the context of the whole model. It is important in the interpretation that the values cannot be taken individually. For example, gEBV for interval between first and last insemination had both a negative and a positive associations with survival depending on the decision moment... Apendix 2.C shows an example of the differences in direction and strength of associations when some variables available at birth were tested in a single or multiple regression model. The differences between a single and multiple regression model could be explained by high correlations between some variables, because collinearity is known to cause issues with multiple regression (Whittingham et al., 2006; Yoo et al., 2014). Antagonistic relations between multiple gEBV that were both associated with survival (such as fertility and production (Zavadilová and Zink, 2013) could also explain some of the unusual relations found. The combination of various related gEBV could be somewhat mitigated by more stringent selection of the gEBV variables provided. Each decision moment appeared to included one or multiple selected gEBV of several groups; feet and legs, udder, udder health, exterior and fertility. We tested the consistent selection of groups of gEBV by building a model which included gEBV for longevity, overall fertility, udder score, udder health, feet and legs and overall exterior at all decision moments. These 'general' models also included the phenotypic records normally selected at each decision moment. The general models performed identical to the models described in this paper (data not shown).

This study shows that there is merit in the combination of phenotypic information and gEBV for the prediction of survival, because gEBV were always selected in combination with phenotypic information. Accuracies for the combined model were also higher than for models using only gEBV or phenotype information in all but the last decision moment. Literature already shows an increased interest in multiple fields to develop methods to combine phenotypic and genomic information for various purposes (Javed et al., 2014; Blake et al., 2016; Haendel et al., 2016), and has proven valuable for example in disease prognosis in human diseases (Perlee et al., 2013; Javed et al., 2014). In cattle, a recent paper estimating the lifetime profitability of a dairy cow also combines genomically estimated breeding values with a small amount of phenotypic variables in order to obtain more accurate estimates (Kelleher et al., 2015). Because there is interest in combining genotype and phenotypic information, future research could explore the exact benefit of including genotypes for performance measures, as well as investigate other methods compared to multiple regression.

2.5 Conclusion

In this study, genomic information in the form of genomically estimated breeding values was combined with phenotypic information in order to predict survival to second lactation in Holstein dairy cows at five different decision moments. A combination of gEBV and phenotypic information resulted in better models than using only one type of information. The addition of gEBVs improved early prediction especially. A combination of gEBV and phenotypic information also resulted in the best predictive performance up to the last decision moment. While accurate individual prediction of survival outcome could not be achieved, surviving heifers were predicted to have a higher probability of survival than non-surviving heifers on average. By selecting the heifers with the highest predicted probability of survival, increased survival could be realized at the population level in practice.

2.6 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 n°727213. The data for this study was provided by cattle improvement cooperative CRV (Arnhem, the Netherlands).

References

Bach, A. 2011. Associations between several aspects of heifer development and dairy cow survivability to second lactation. Journal of Dairy Science 94(2):1052-1057.

Barkema, H., M. Von Keyserlingk, J. Kastelic, T. Lam, C. Luby, J.-P. Roy, S. LeBlanc, G. Keefe, and D. Kelton. 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. Journal of dairy science 98(11):7426-7445.

Beaudeau, F., H. Seegers, V. Ducrocq, C. Fourichon, and N. Bareille. 2000. Effect of health disorders on culling in dairy cows: a review and a critical discussion. In: Annales de zootechnie. p 293-311.

Blake, V. C., C. Birkett, D. E. Matthews, D. L. Hane, P. Bradbury, and J.-L. Jannink. 2016. The Triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. The plant genome 9(2)

Boulton, A., J. Rushton, and D. Wathes. 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:1-9.

Brickell, J., and D. Wathes. 2011. A descriptive study of the survival of Holstein-Friesian heifers through to third calving on English dairy farms. Journal of Dairy Science 94(4):1831-1838.

Buenger, A., V. Ducrocq, and H. Swalve. 2001. Analysis of survival in dairy cows with supplementary data on type scores and housing systems from a region of Northwest Germany. Journal of dairy science 84(6):1531-1541.

Calus, M., P. Bijma, and R. Veerkamp. 2015. Evaluation of genomic selection for replacement strategies using selection index theory. Journal of dairy science 98(9):6499-6509.

Caraviello, D., K. Weigel, and D. Gianola. 2004. Analysis of the relationship between type traits and functional survival in US Holstein cattle using a Weibull proportional hazards model. Journal of dairy science 87(8):2677-2686.

Carlström, C., E. Strandberg, K. Johansson, G. Pettersson, H. Stålhammar, and J. Philipsson. 2016. Genetic associations of in-line recorded milkability traits and udder conformation with udder health. Acta Agriculturae Scandinavica, Section A—Animal Science 66(2):84-91.

Dechow, C., and R. Goodling. 2008. Mortality, culling by sixty days in milk, and production profiles in high-and low-survival Pennsylvania herds. Journal of dairy science 91(12):4630-4639.

Essl, A. 1998. Longevity in dairy cattle breeding: a review. Livestock Production Science 57(1):79-89.

Fetrow, J., K. Nordlund, and H. Norman. 2006. Invited review: Culling: Nomenclature, definitions, and recommendations. Journal of dairy science 89(6):1896-1905.

Forabosco, F., J. Jakobsen, and W. Fikse. 2009. International genetic evaluation for direct longevity in dairy bulls. Journal of dairy science 92(5):2338-2347.

Getu, A., and G. Misganaw. 2015. The Role of Conformational Traits on Dairy Cattle Production and Their Longevities. Open Access Library Journal 2(3):7.

Groenendaal, H., D. Galligan, and H. Mulder. 2004. An economic spreadsheet model to determine optimal breeding and replacement decisions for dairy cattle. Journal of Dairy Science 87(7):2146-2157.

Hadley, G., C. Wolf, and S. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. Journal of dairy science 89(6):2286-2296.

Haendel, M. A., M. G. Kann, and N. L. Washington. 2016. Innovative approaches to combining genotype, phenotype, epigenetic, and exposure data for precision diagnostics. In: Biocomputing 2016: Proceedings of the Pacific Symposium. p 93-95.

Hansen, L., J. Cole, G. Marx, and A. Seykora. 1999. Productive life and reasons for disposal of Holstein cows selected for large versus small body size. Journal of dairy science 82(4):795-801.

Heinrichs, A. 1993. Raising dairy replacements to meet the needs of the 21st century. Journal of dairy science 76(10):3179-3187.

Heinrichs, A., and B. Heinrichs. 2011. A prospective study of calf factors affecting first-lactation and lifetime milk production and age of cows when removed from the herd. Journal of dairy science 94(1):336-341.

Heise, J., Z. Liu, K. F. Stock, S. Rensing, F. Reinhardt, and H. Simianer. 2016. The genetic structure of longevity in dairy cows. Journal of dairy science 99(2):1253-1265.

Hoffman, P., and D. A. Funk. 1992. Applied Dynamics of Dairy Replacement Growth and Management1. Journal of Dairy Science 75(9):2504-2516.

Hultgren, J., C. Svensson, D. O. Maizon, and P. A. Oltenacu. 2008. Rearing conditions, morbidity and breeding performance in dairy heifers in southwest Sweden. Preventive veterinary medicine 87(3):244-260.

Javed, A., S. Agrawal, and P. C. Ng. 2014. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nature methods 11(9):935.

Kelleher, M., P. Amer, L. Shalloo, R. Evans, T. Byrne, F. Buckley, and D. P. Berry. 2015. Development of an index to rank dairy females on expected lifetime profit. Journal of dairy science 98(6):4225-4239.

Kern, E. L., J. A. Cobuci, C. N. Costa, C. M. McManus, and J. Braccini Neto. 2015. Genetic association between longevity and linear type traits of Holstein cows. Scientia Agricola 72(3):203-209.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of statistical software 28(5):1-26.

LTO. 2011. Melkveehouderij: Midden in de maatschappij – Visie. LTO Nederland Vakgroep Melkveehouderij, Den Haag, The Netherlands.

Mohd Nor, N., W. Steeneveld, and H. Hogeveen. 2014. The average culling rate of Dutch dairy herds over the years 2007 to 2010 and its association with herd reproduction, performance and health. Journal of Dairy Research 81(1):1-8.

Mohd Nor, N., W. Steeneveld, M. Mourits, and H. Hogeveen. 2012. Estimating the costs of rearing young dairy cattle in the Netherlands using a simulation model that accounts for uncertainty related to diseases. Preventive veterinary medicine 106(3):214-224.

Mohd Nor, N., W. Steeneveld, M. Mourits, and H. Hogeveen. 2015. The optimal number of heifer calves to be reared as dairy replacements. Journal of dairy science 98(2):861-871.

Mourits, M., A. Dijkhuizen, R. Huirne, and D. Galligan. 1997. Technical and economic models to support heifer management decisions: basic concepts. Journal of dairy science 80(7):1406-1415.

Mourits, M., H. Van der Fels-Klerx, R. Huirne, and M. Huyben. 2000. Dairy-heifer management in the Netherlands. Preventive veterinary medicine 46(3):197-208.

Olechnowicz, J., and J. M. Jaskowski. 2011. Reasons for culling, culling due to lameness, and economic losses in dairy cows. Medycyna Weterynaryjna 67(9):618-621.

Ortiz-Pelaez, A., D. Pritchard, D. Pfeiffer, E. Jones, P. Honeyman, and J. Mawdsley. 2008. Calf mortality as a welfare indicator on British cattle farms. The Veterinary Journal 176(2):177-181.

Perlee, L. T., A. T. Bansal, K. Gehrs, J. S. Heier, K. Csaky, R. Allikmets, P. Oeth, T. Paladino, D. H. Farkas, and P. L. Rawlings. 2013. Inclusion of genotype with fundus phenotype improves accuracy of predicting choroidal neovascularization and geographic atrophy. Ophthalmology 120(9):1880-1892.

Pritchard, T., M. Coffey, R. Mrode, and E. Wall. 2013. Genetic parameters for production, health, fertility and longevity traits in dairy cows. Animal 7(1):34-46.

Pryce, J., B. Hayes, and M. Goddard. 2012. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Proceedings of the 38th ICAR Session 28

Raboisson, D., F. Delor, E. Cahuzac, C. Gendre, P. Sans, and G. Allaire. 2013. Perinatal, neonatal, and rearing period mortality of dairy calves and replacement heifers in France. Journal of Dairy Science 96(5):2913-2924.

R core team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

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.

Sewalem, A., G. Kistemaker, F. Miglior, and B. Van Doormaal. 2004. Analysis of the relationship between type traits and functional survival in Canadian Holsteins using a Weibull proportional hazards model. Journal of Dairy Science 87(11):3938-3946.

Sewalem, A., F. Miglior, G. J. Kistemaker, P. Sullivan, and B. J. Van Doormaal. 2008. Relationship Between Reproduction Traits and Functional Longevity in Canadian Dairy Cattle. Journal of Dairy Science 91(4):1660-1668. doi: http://dx.doi.org/10.3168/jds.2007-0178

Svensson, C., and J. Hultgren. 2008. Associations between housing, management, and morbidity during rearing and subsequent first-lactation milk production of dairy cows in southwest Sweden. Journal of dairy science 91(4):1510-1518.

Van Pelt, M., G. De Jong, and R. Veerkamp. 2016a. Changes in the genetic level and the effects of age at first calving and milk production on survival during the first lactation over the last 25 years. animal 10(12):2043-2050.

Van Pelt, M., V. Ducrocq, G. De Jong, M. Calus, and R. Veerkamp. 2016b. Genetic changes of survival traits over the past 25 yr in Dutch dairy cattle. Journal of dairy science 99(12):9810-9819.

Van Pelt, M., T. Meuwissen, G. de Jong, and R. Veerkamp. 2015. Genetic analysis of longevity in Dutch dairy cattle using random regression. Journal of dairy science 98(6):4117-4130.

Wathes, D., J. Brickell, N. Bourne, A. Swali, and Z. Cheng. 2008. Factors influencing heifer survival and fertility on commercial dairy farms. animal 2(8):1135-1143.

Weigel, K., P. Hoffman, W. Herring, and T. Lawlor. 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. Journal of dairy science 95(4):2215-2225.

Wells, S., A. Trent, W. Marsh, P. McGovern, and R. Robinson. 1993. Individual cow risk factors for clinical lameness in lactating dairy cows. Preventive veterinary medicine 17(1-2):95-109.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of animal ecology 75(5):1182-1189.

Yoo, W., R. Mayberry, S. Bae, K. Singh, Q. P. He, and J. W. Lillard Jr. 2014. A study of effects of multicollinearity in the multivariable analysis. International journal of applied science and technology 4(5):9.

Zavadilová, L., and V. Zink. 2013. Genetic relationship of functional longevity with female fertility and milk production traits in Czech Holsteins. Czech Journal of Animal Science 58(12):554-565.

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.

2.7 Appendix

TABLE 2.A.1 All variables used in this study, including the decision moment at which each variable became available.

Phenotypic variables	Decision moment	continuous
Animal Identification number	Birth	No
Year of birth	Birth	No
Birth farm UBN	Birth	No
Month of birth	Birth	No
Birth season	Birth	No
Parity dam	Birth	Yes
Breed	Birth	No
Holstein %	Birth	No
Red factor	Birth	No
Calving ease dam	Birth	No
Gestation duration dam	Birth	Yes
Birth weight	Birth	Yes
Insemination farm	18 months	No
Insemination season	18 months	No
Countable inseminations	18 months	Yes
Non return status at 18 months	18 months	No
No insemination information at 18 months	18 months	No
Number of farm movements at 18 months	18 months	Yes
Age at first insemination	18 months	Yes
Type of first insemination	18 months	No
Number of inseminations	18 months	Yes
Raised at a specialty calf-rearing farm	first calving	No
Calving season	first calving	No
Total number of farm movements at 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 hours	first calving	No
Calving ease	first calving	No
Gestation duration	first calving	Yes
Birthweight calf	first calving	Yes
Calf survival first week	first calving	No
Calf survival second week	first calving	No
Twins	first calving	No

TABLE 2.A.1 Continued

Phenotypic variables	Decision moment	continuous
Kg milk at 6 weeks	6 weeks p.c.	Yes
Fat percentage milk at 6 weeks	6 weeks p.c.	Yes
Protein percentage milk at 6 weeks	6 weeks p.c.	Yes
Cell count milk at 6 weeks	6 weeks p.c.	Yes
Urea milk at 6 weeks	6 weeks p.c.	Yes
Lactose percentage milk at 6 weeks	6 weeks p.c.	Yes
Cow status indicator at 6 weeks	6 weeks p.c.	No
Number of negative indications at 6 weeks	6 weeks p.c.	Yes
Number of days in lactation on milk test day	6 weeks p.c.	Yes
Complete milk measurement available at 6 weeks	6 weeks p.c.	No
First parity insemination farm UBN	200 days p.c.	No
First parity insemination season	200 days p.c.	No
First parity first insemination type	200 days p.c.	No
Number of inseminations in first parity	200 days p.c.	Yes
Non return status at 200 days post calving	200 days p.c.	No
Age at 200 days post calving	200 days p.c.	Yes
Insemination known in the first parity	200 days p.c.	No
Age at first insemination in the first parity	200 days p.c.	Yes
Number of farm movements at 200 days post calving	200 days p.c.	Yes
Number of known milk testing’s at 200 days post calving	200 days p.c.	Yes
Average kg of milk	200 days p.c.	Yes
Average fat percentage of milk	200 days p.c.	Yes
Average protein percentage of milk	200 days p.c.	Yes
Average cell count of milk	200 days p.c.	Yes
Average Urea of milk	200 days p.c.	Yes
Average lactose percentage of milk	200 days p.c.	Yes
Number of negative indications at 200 days post calving (mastitis, abortion, other illness, teat disorders)	200 days p.c.	Yes
Number of farm movements in the first parity	200 days p.c.	Yes

gEBV	Decision moment	continuous
"NVI" Dutch breeding goal standard	Birth	Yes
Kg milk	Birth	Yes
Kg fat	Birth	Yes
Kg protein	Birth	Yes
Kg lactose	Birth	Yes
"Inet" Dutch production index	Birth	Yes
Cell count	Birth	Yes

TABLE 2.A.1 Continued

gEBV	Decision moment	continuous
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
Non return status at 56 days	Birth	Yes
Interval calving - first insemination	Birth	Yes
Calving interval	Birth	Yes
Interval first - last insemination	Birth	Yes
Conception ratio	Birth	Yes
Claw health	Birth	Yes
Calf vitality 3 - 365 days	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

TABLE 2.A.1 Continued

gEBV	Decision moment	continuous
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

*gEBV: genomic estimated breeding values, p.c. = post calving

TABLE 2.B.1 Decision moment ‘birth’.*

gEBV and Phenotypic variables selected	coefficients
gEBV longevity	0.128
gEBV foot angle	-0.115
gEBV udder depth	-0.102
gEBV frame	0.074
gEBV interval first-last insemination	0.140
gEBV non return at 56 days	-0.092
gEBV kg fat	0.061
Year of birth (reference = 2012)	
2013	-0.372
Birth season (reference = Fall)	
Spring	0.338
Summer	0.286
Winter	0.434

*Variables selected by stepwise selection. Estimated coefficients are shown for the gEBV and for the individual classes of each phenotypic variable with the reference class in brackets.

TABLE 2.B.2 Decision moment ‘18 months of age’.*

gEBV and Phenotypic variables selected	coefficients
gEBV overall fertility	0.145
gEBV kg fat	0.122
gEBV longevity	0.212
gEBV Milking speed	0.061
gEBV foot angle	-0.078
gEBV udder depth	-0.137
NVI	-0.122
gEBV Overall exterior score	0.143
Birth season (reference = Fall)	
Spring	0.411
Summer	0.474
Winter	0.410
Year of Birth (reference = 2012)	
2013	-0.268
Non-return status at 18 months (reference = Non-return)	
Unknown	-0.426
Number of inseminations at 18 months (reference = 0)	
1	0.206
2	0.317
3	0.096
4	-0.346
5+	-0.988
Unknown	-0.130

*Variables selected by stepwise selection. Estimated coefficients are shown for the gEBV and for the individual classes of each phenotypic variable with the reference class in brackets.

TABLE 2.B.3 Decision moment ‘first calving’.*

gEBV and phenotypic variables selected	coefficients
gEBV longevity	0.261
gEBV udder depth	-0.140
gEBV overall fertility	0.091
gEBV foot angle	-0.086
gEBV somatic cell count	0.196
gEBV udder health	-0.283
gEBV Rear udder height	0.072
gEBV Feet and legs	-0.067
gEBV Chest width	0.099
Year of birth (reference = 2012)	
2013	-0.270
Birth season (reference = Fall)	
Spring	0.423
Summer	0.341
Winter	0.121
Age at first calving in days (reference = >1000)	
< 650	1.531
650-700	1.700
700-750	1.549
750-800	1.417
800-850	1.307
850-900	1.053
900-950	1.005
950-1000	1.110
Sex of calf (reference = Male)	
Female	0.206
Unknown	-0.172
Coat color (reference = Red)	
Black	-0.208

*Variables selected by stepwise selection. Estimated coefficients are shown for the gEBV and for the individual classes of each phenotypic variable with the reference class in brackets.

TABLE 2.B.4 Decision moment 'six weeks post calving'.*

gEBV and phenotypic variables selected	coefficients
gEBV longevity	0.242
gEBV overall fertility	0.242
gEBV rear legs hind view	-0.162
gEBV locomotion	0.124
gEBV Chest width	0.131
gEBV foot angle	-0.085
gEBV somatic cell count	0.212
gEBV udder health	-0.289
gEBV interval first-last insemination	-0.213
gEBV kg lactose	-0.073
gEBV non return at 56 days	-0.088
Year of birth (reference = 2012)	
2013	-0.342
Birth season (reference = Fall)	
Spring	0.459
Summer	0.600
Winter	0.423
Age at first calving in days (reference = >1000)	
< 650	1.520
650-700	1.881
700-750	1.741
750-800	1.660
800-850	1.456
850-900	1.309
900-950	0.849
950-1000	1.097
Sex of calf (reference = Male)	
Female	0.156
Unknown	-0.116
Kg of milk produced at milk test day closest to 6 week post calving (reference = <15)	
15 - 20	0.354
20 - 25	0.825
25 - 30	1.359
30 - 35	1.521
35 - 40	1.688
40+	2.512
Unknown	0.049

*Variables selected by stepwise selection. Estimated coefficients are shown for the gEBV and for the individual classes of each phenotypic variable with the reference class in brackets.

TABLE 2.B.5 Decision moment ‘200 days post calving’. *

gEBV and phenotypic variables selected	coefficients
gEBV longevity	0.207
gEBV udder health	-0.153
gEBV interval first-last insemination	0.106
gEBV leg and feet	-0.247
gEBV overall exterior score	0.256
gEBV stature	-0.094
gEBV udder support	0.074
gEBV udder depth	-0.106
Year of birth (reference = 2012)	
2013	-0.239
Age at first calving in days (reference = >1000)	
< 650	2.831
650-700	1.844
700-750	1.792
750-800	1.958
800-850	1.763
850-900	1.554
900-950	0.835
950-1000	1.385
Calving season (reference = Fall)	
Spring	0.510
Summer	0.725
Winter	0.331
Calf survival at 1 week of age (reference = Alive)	
Died within 1 week	14.250
Dead 24 hours post calving	-0.142
Unknown	0.165
Number of farm moves at 200 days	
1	-0.998
2	13.527
3	12.102
4	14.217
Average kg of milk per test milk day at 200 days post calving (reference = < 20)	
20 – 25	0.335
25 – 30	0.884
30 – 35	1.311
35 – 40	1.060
40+	1.339
Unknown	-0.560

TABLE 2.B.5 Continued

gEBV and phenotypic variables selected	coefficients
Average cell count (x1000) per test milk day at 200 days post calving (reference = < 25)	
25 – 50	-0.145
50 – 75	-0.348
75 – 100	0.448
100 – 125	-0.437
125+	-0.584
Number of inseminations at 200 days post calving (reference = 0)	
1	0.598
2	0.132
3	0.094
4	-0.308
5+	-0.091
Unknown	-0.955
Non-return status at 200 days post calving (reference = non-return)	
Unknown	-1.065
Average percentage of protein per test milk day at 200 days post calving (reference = < 3.0)	
3.00 – 3.25	0.456
3.25 – 3.50	0.784
3.50 – 3.75	0.705
3.75 – 4.00	1.414
4,00+	0.953
Negative indication at test milking 6 weeks post calving (reference = No)	
Yes	1.401
Unknown	0.946
Negative indication count	-0.788
Number of inseminations at 18 months (reference = 0)	
1	0.633
2	0.523
3	0.270
4	0.027
5+	-0.581
Unknown	0.783

*Variables selected by stepwise selection. Estimated coefficients are shown for the gEBV and for the individual classes of each phenotypic variable with the reference class in brackets.

TABLE 2.C.1 Estimated coefficients when using single or multiple regression for the selected variables at Birth

gEBV and Phenotypic variables selected	Multiple regression coefficients	Single regression coefficients
gEBV longevity	0.159	0.153
gEBV foot angle	-0.109	-0.120
gEBV udder depth	-0.068	-0.060
gEBV frame	0.062	0.017
gEBV kg fat	0.063	0.087
gEBV non return at 56 days	-0.086	0.009
gEBV dairy strength	0.057	0.043
gEBV interval first-last insemination	0.218	0.085
gEBV calving interval	-0.106	0.048
Season of birth (reference = fall)		
spring	0.304	0.118
summer	0.231	0.205
winter	0.334	0.119
Year of birth (reference = 2012)		
2013	-0.360	-0.220

III

CHAPTER 3

Comparing regression, naive Bayes and random forest 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[1], I. Athanasiadis[3], B. J. Ducro[1]

Journal of Dairy science (2019) 102.10: 9409-9421
DOI: S0022030219307155

[1]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2] Cooperation CRV, Animal Evaluation Unit, PO Box 454, 6800 AL Arnhem, the Netherlands; [3] Wageningen University & Research Information Technology Group, 6706 KN Wageningen, The Netherlands

Abstract

In this study we compared the linear method multiple logistic regression to the non-linear machine learning methods naive Bayes and random forest. All three methods were used to predict individual survival to second lactation of dairy heifers. The dataset used for prediction contained 6847 heifers born between January 2012 and June 2013, which had a known outcome for survival. Each animal had 50 genomic estimated breeding values available at birth and up to 65 phenotypic variables accumulating over time. Survival was predicted at five moments in life: at birth, at 18 months, at first calving, at six weeks post calving, and at 200 days post calving. The datasets were randomly split in 70% training and 30% testing sets to test model performance for 20-fold validation. The methods were compared on accuracy, sensitivity and 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 months, all methods had overlapping performance, with no method significantly outperforming the other. At first calving, 6 weeks and 200 days post calving random forest and naive Bayes had overlapping performance, with both machine learning methods outperforming multiple logistic regression. Overall, naive Bayes has the highest average AUC at all decision moments up to 200 days past first calving. At 200 days post calving, random forest has the highest AUC. All methods obtained similar increases in survival in the practical scenario. Despite this, the methods appeared to predict individual heifers differently. While all methods improved over time, the changes in mean model outcomes for surviving and non-surviving animals were different per method. Furthermore, Correlations of individual predictions between methods ranged from r = 0.417 to r = 0.700, with the lowest correlations at first calving for all methods. In short, all three methods were able to predict survival on population level as all methods improved survival in a practical scenario. However, depending on the method used, an individual animal could be quite different between methods.

KEY WORDS: machine learning, naive Bayes, regression, random forest, phenotypic prediction

3.1 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 as they are able to derive a model from the available data without prior knowledge of the relations between the variables (McQueen et al., 1995; Kotsiantis et al., 2007). Machine learning methods thrive on large datasets and make less 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 successfully used to predict on a whole range of different traits, e.g., predicting diseases 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, there are also recent papers on similar topics using ‘traditional’ linear methods; for example 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 like regression is that despite the potential of machine learning, they do not always prove superior to ‘traditional’ linear modeling (Cortez et al., 2006; Van Hertem et al., 2014; Hempstalk et al., 2015; Ghafouri-Kesbi et al., 2017). Comparisons between machine learning and traditional methods may also not be possible in some cases, for example due to missing records which can be used by some machine learning methods, but not by regression (Bennett, 2001), or in the case of video data (Kabra et al., 2013). Furthermore, it is difficult to determine beforehand which method will result in for example the highest accuracy for a particular prediction problem (White et al., 2018), as in practice many different machine learning techniques may be suitable for predicting a variable of interest. This results in a trial and error approach for finding the best method for each individual prediction problem (Amrine et al., 2014; Libbrecht and Noble, 2015).

Two machine learning methods that use very different approaches but are both applied competitively in the field of animal science, are naive Bayes and random forests. Naive Bayes is a family of classifiers that implements Bayesian techniques in order to form a simple network based on prior probabilities (Jensen, 1996). The naive Bayes method relies on independence between the input variables, but it performs surprisingly well even under conditions considered suboptimal for the algorithm (Domingos and Pazzani, 1997; Friedman, 1997). Despite the relative simplicity of the algorithm, naive Bayes is still a widely used machine learning method (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 splitting rules which 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 available for selection at each split in the tree (Breiman, 2001). These trees, relatively poor classifiers individually, are then combined into an ensemble of trees called a random forest, which is used for prediction. The prediction results of a random forest are thus a summation of the prediction outcomes of many individual trees.

The aim of this study is to compare the 'traditional' linear method regression to 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 hope to gain knowledge on the wide variety of available tools for the prediction of complex biological traits.

3.2 Materials and Method

3.2.1 DATA

The data used in this study was identical to data described in a previous study (van der Heide et al., submitted). The dataset consisted of records on 6847 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 two weeks', where two additional weeks were included to ensure that a heifer that was only considered to have survived when the heifer did not die or was culled due to the direct consequences of second calving. In order to have had a known outcome for survival, all cows included had to be born at least 46 months prior to the end of data collection and were not exported abroad. In this dataset, 85.8% of the 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 only included the direct genomic values and did not include any own performance records. Furthermore, they had up to 65 phenotypic variables, including records on birthweight and gestation length, insemination records up to second parity, first parity calving records, and first parity lactation information (See also Appendix 3.A). Not all animals had records for every variable, as some were missing, and some were not collected as the heifer died before the variable could be collected. All phenotypic continuous variables were transformed into factors of at least five levels, containing a factor level for missing information to allow animals with missing records to still be used in the regression analysis. At the cost of losing some of the original information, changing the continuous variables into factorial variables does allow us to include all animals in analyses regardless of method.

From the complete data, five datasets were created which contained all information available at five distinct moments in the life of a dairy cow. These distinct 'decision moments' were moments in the life of a dairy cow where new information became available, and where a management decision could be aided by a prediction of the expected survival of the animal. The decision moments were at birth, at 18 months of age, at first calving, at six weeks post calving and at 200 days post calving. At the first decision moment, genomic information and only limited phenotypic information was available, whereas at the last decision moment all information was available. Appendix 3.A shows all available variables and the decision moments at which they are available. The decision moments were chosen to investigate the ability of a model to predict survival at various points in the life of a cow. Early prediction was preferred, but since little information was available early-on, this was not feasible. Only animals still alive at the start of a decision moment were used in the analysis.

3.2.2 MODEL AND ANALYSIS

The analyses were performed in the statistical program R version 3.3.1 (R core team 2016), using the package 'caret' for logistic multiple regression (Kuhn, 2008), the package 'randomForest' for the random forest approach (Liaw and Wiener, 2002) and the package 'naivebayes' to apply the naive Bayes method (Majka, 2018). As the randomForest package was unable to predict on factorial data exceeding 53 levels, farm identification number variables (such as birth farm, farm of first calving and farm of first milking) were transformed into sets of dummy variables. No other changes were made to the data, to keep the five distinct datasets (one for each distinct decision moment) presented to each model as identical as possible. Our linear method was a logistic multiple regression, where the Akaike information criterion (AIC) was used in forward stepwise selection to determine the best possible model at each decision moment. The selected regression models included various phenotypic variables at the different decision moments, and a set of six genomic estimated breeding values (gEBV) which were gEBV related to longevity, fertility, feet and leg score, conformation score, udder score and udder health. Production gEBV were not included as they were not selected, because the lifespan gEBV was uncorrected for production. 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 datasets, as these values were the recommended values and gave the highest AUC values. As the Bayesian machine learning method selected is naive Bayes, it was not necessary to set any a-priori values for the variables. The random forest, naive Bayes and regression methods were trained on a random selection of 70% of each of the five datasets and were then validated on the remaining 30%. This process was repeated twenty times for all five decision moments and methods.

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

The performance of the methods was evaluated by measuring the contrasts between the mean probabilities of survival for both survival groups, the accuracy, sensitivity and specificity. The contrasts are the differences between the means of the two groups, expressed in units of standard deviation. This allowed us to compare the model outputs for the two groups between the methods. The accuracy was the proportion of correctly predicted animals, the sensitivity was the proportion of surviving heifers correctly predicted as surviving, and the specificity was the proportion of non-surviving heifers correctly predicted as not surviving. However, it appears that because the trait of interest 'survival to second lactation' was an unbalanced trait, using the accuracy as an indicator of which method was superior could have been biased. This was why the performance of the models was also evaluated 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 thus not affected by the trade-off between specificity and accuracy. Finally, we tested a scenario where only the top 50% scoring heifers were kept on a farm at a specific decision moment. This latter evaluation approach was considered an example of how the developed models may be used in practice, as part of a decision support system.

As the consistency across methods was also of interest, we also looked at the correlations between the methods. All heifers from the testing set had three predicted probabilities of survival, one for each method. We calculated the Pearson's r and Spearman's p between all three methods for all five decision moments (Chok, 2010), and obtained averages for these correlations over the 20 validation runs. This was done as not only the similarity of the predicted probabilities between the methods was important, but also the assigned rank of the heifer in the group; selection to cull animals would take place on rank in practice.

3.3 Results

Table 3.1 shows the contrasts between the average predicted probabilities of surviving heifers 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 was able to predict survival at least at a population level, and that all model performance increased with additional information regardless of method. At birth and at 18 months, naive Bayes had the highest contrast between the two groups, and after first calving regression had the highest contrasts.

TABLE 3.1 **Contrasts between the mean probability of survival for the surviving heifers and the non-surviving heifers.**

	Regression	Naive Bayes	Random forest
Birth	0.279	0.327	0.231
18 months old	0.409	0.446	0.331
At first calving	0.525	0.435	0.393
6 weeks post calving	0.583	0.554	0.494
200 days post calving	0.800	0.606	0.747

Looking at the accuracy, sensitivity and specificity, naive Bayes outperformed the regression and the random forest methods in the first two decision moment on accuracy (Figure 3.1).

After the second decision moment, the regression model and the naive Bayes model alternatively performed best at the third, fourth and fifth decision moment respectively. Naive Bayes had the highest sensitivity of the three methods at birth and 18 months, but also the lowest specificity at the first two decision moments (Figure 3.2). The regression model had the lowest sensitivity, but also the highest specificity, and random forest had intermediate scores for both decision moments.

There appeared to be a trade-off between high specificity and high accuracy. As a smaller proportion of heifers do not survive in practice, negative predictions were significantly less likely to be true then positive predictions (i.e. for a random heifer from the data, the odds of surviving were higher than the odds of not surviving).

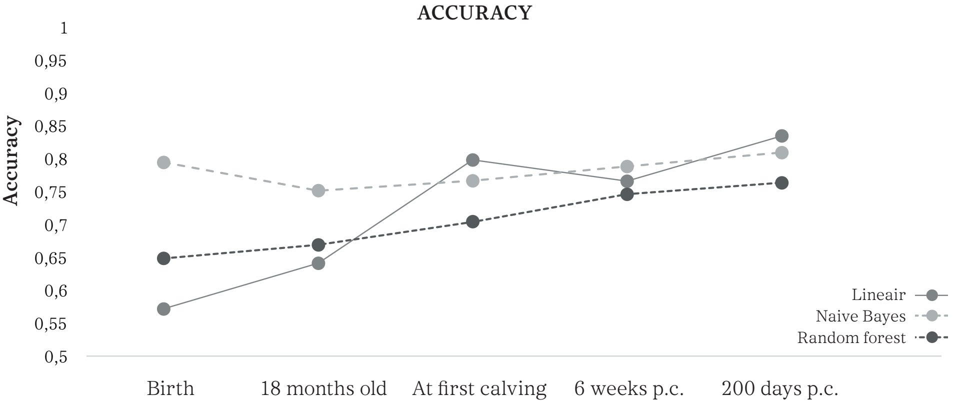


FIGURE 3.1 **Accuracy of prediction of the regression (linear), naive Bayes and random forest models.** p.c. = post calving.

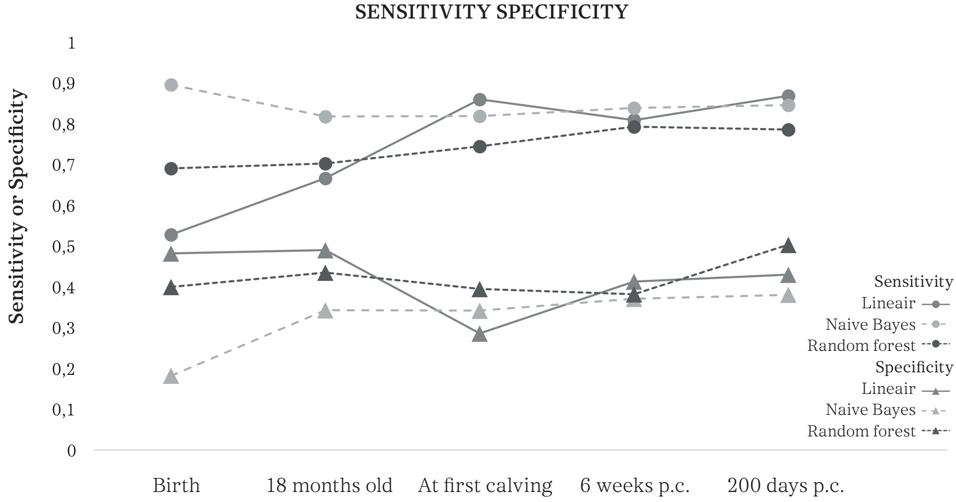


FIGURE 3.2 Sensitivity and Specificity of the regression (Linear), naive Bayes (NB) and random forest models (RF). This figure shows the balance between the sensitivity (lines with circles) and the specificity (lines with triangles), p.c. = post calving.

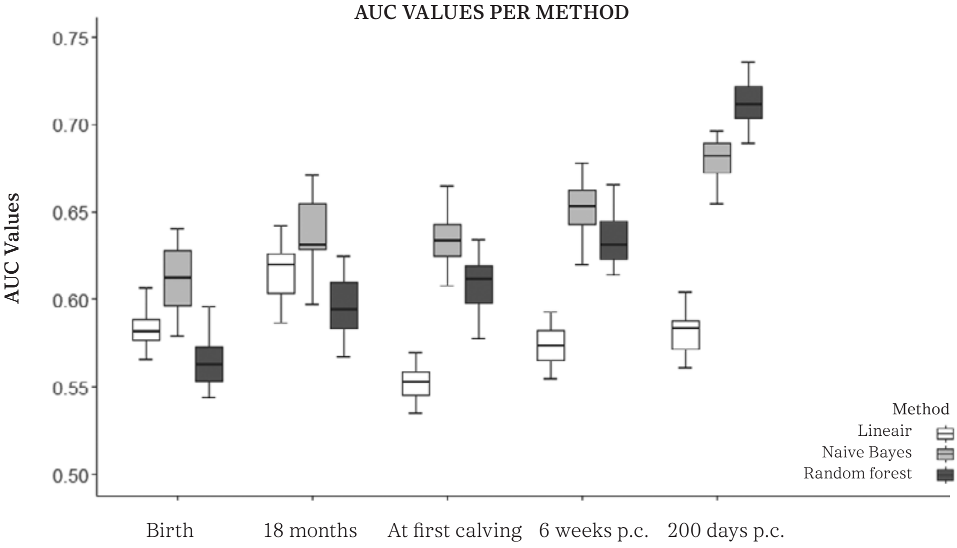


FIGURE 3.3 AUC of the multiple logistic regression (linear), naive Bayes and random forest methods at the three decision moments. Outliers were removed, p.c. = post calving.

If the model was able to very accurately predict non-surviving heifers, this would not be problematic. However, since survival was a complex trait and therefore difficult to predict, models that made less predictions that heifers will not survive were more likely to be correct more often by chance. Thus, a model with a high specificity made more negative predictions, resulting in a loss of accuracy. AUC was not biased by this trade-off, as it considers all possible specificity and sensitivity values (Figure 3.3). At birth and 18 months, all methods had overlapping performance, with no method significantly outperforming the other. At first calving, 6 weeks and 200 days post calving random forest and naive Bayes had overlapping performance, with both machine learning methods outperforming multiple logistic regression. Overall, naive Bayes had the highest average AUC at all decision moments up to 200 days past first calving. At 200 days post calving, random forest had the highest AUC. All methods performed significantly better than an AUC of 0.5 (random chance) at all decision moments, indicating that even at birth it was possible to predict survival to second lactation to some extent.

In a practical scenario where 50% of the heifer calves with the highest probability of survival were selected (Figure 3.4), all methods performed similar. Again, naive Bayes scores 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 three methods resulted in increased survival when compared to a random selection of heifers for every decision moment.

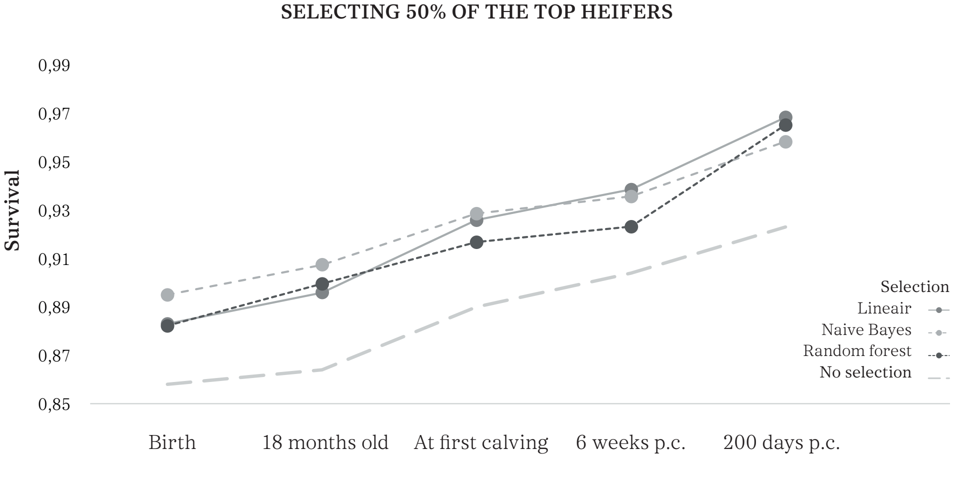


FIGURE 3.4 The surviving proportion of the heifer population when selecting 50% of the highest scoring heifers using regression (linear), naive Bayes and random forest models. p.c. = post calving.

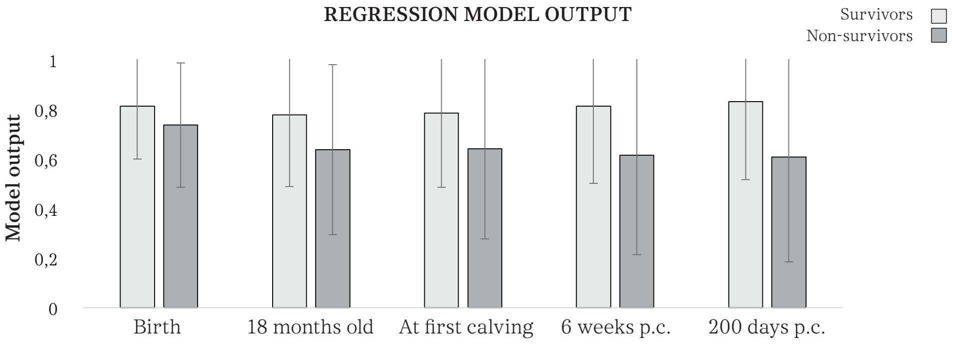


FIGURE 3.5 Mean model output of the multiple logistic regression for surviving and non-surviving animals at all five decision moments. The error bars represent the standard deviation, p.c. = post calving.

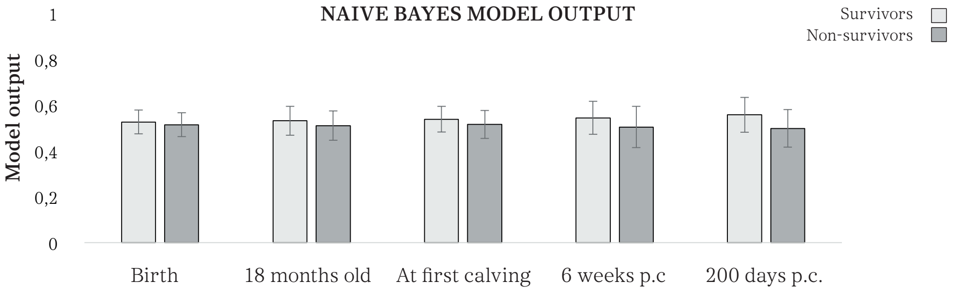


FIGURE 3.6 Mean model output of naive Bayes for both surviving and non-surviving animals at all five decision moments. The error bars represent the standard deviation. The error bars are exceptionally large as naive Bayes attempts to classify cases closer to 0 or 1 than the other methods, p.c. = post calving.

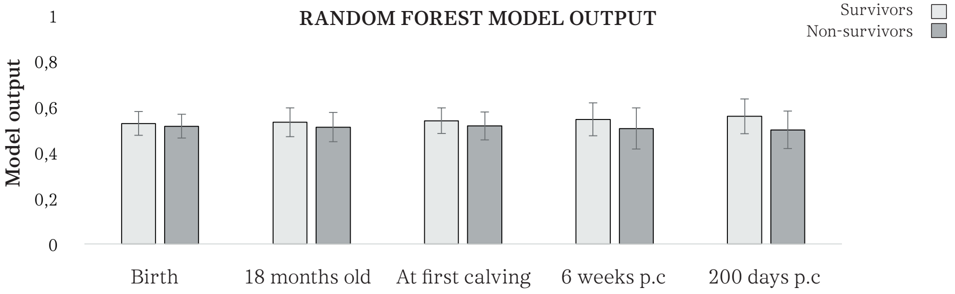


FIGURE 3.7 Mean model output of random forest for both surviving and non-surviving animals at all five decision moments. The error bars represent the standard deviation. The mean for the random forest model output is set to 0.5 due to stratified sampling, p.c. = post calving.

While all methods could predict survival and improve over time, the methods did not make identical predictions. In all cases, the means of the surviving and non-surviving groups moved apart over time (Figures 5 through 7), although there was always overlap between the surviving and non-surviving groups. Using regression, the mean predicted probability of survival for surviving heifer increased, while the non-surviving heifer mean remained stable (Figure 3.5). This was the opposite for naive Bayes; the non-surviving heifer mean decreased while the surviving mean remained stable (Figure 3.6).

Naive Bayes also had the largest standard deviation as 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 heifer increased while the mean for non-surviving heifer decreased slightly (Figure 3.7). The random forest method was centered on 0.50 due to the stratified sampling whereas the mean for the other two methods was closer to 0.86, the random chance of survival.

The difference in approach between the methods was also reflected in the correlations between the predictions of the three methods 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 the other method, as expected. Spearman's *p* was generally higher than Pearson's *r*. At birth, the correlations between all methods were moderate to high (Table 3.2) and ranged from an *r* = 0.653 between naive Bayes and random forest to an *r* = 0.539 between the regression and naive Bayes methods. Correlations decreased towards first calving, with all correlations lowest at first calving (from *r* = 0.557 between regression and random

TABLE 3.2 Pearson's *r* and Spearman's *p* correlation coefficients between the model output of the regression (R) naive Bayes (NB) model and random forest model (RF), averaged over 20 runs.

	Pearson correlation			Spearman correlation		
	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 months old	0.566	0.666	0.686	0.573	0.692	0.710
At first calving	0.417	0.547	0.557	0.429	0.601	0.606
6 weeks post calving	0.560	0.632	0.700	0.532	0.626	0.688
200 days post calving	0.488	0.694	0.578	0.488	0.732	0.621

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 3.8 gives an example of the correlations in one of the 20 validation runs.

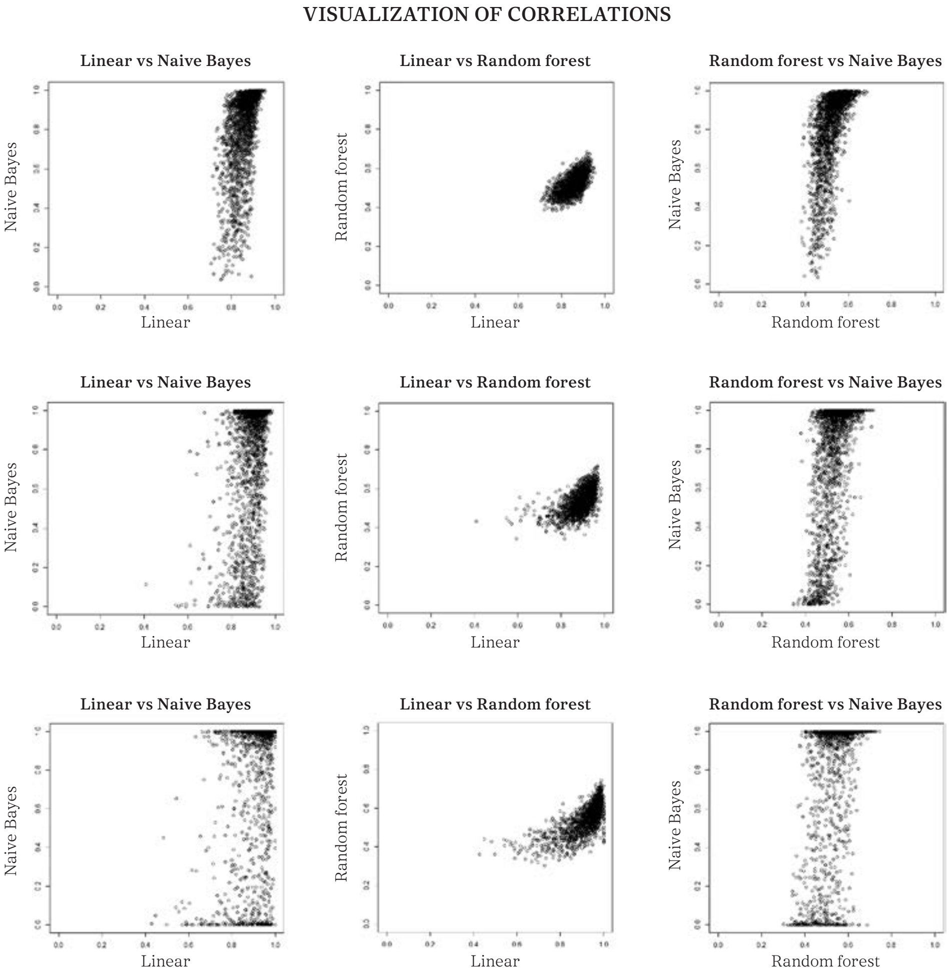


FIGURE 3.8 Visualization of the correlations between the three methods of one of the twenty validation runs. Plotted are the model output values (between 0 and 1) for all three methods. The first row depicts correlations at birth, the second row shows the correlation at first calving, and the third row of images shows the correlation at 200 days post calving. The first column shows the regression method versus the naive Bayes method, the second column the regression method versus the random forest method, and the third column the random forest versus the naive Bayes method.

Note that naive Bayes had a different distribution of predicted probability of survival than the other two methods, favoring predicted probabilities close to 1 or 0, while the other two methods favored predicted probabilities closer to their respective mean predicted probabilities of survival. This is in part due to the different methods chosen to deal with the class imbalance issue. Over time, predicted probabilities moved further apart, with more predictions moving closer to 1 or 0. This is more visible in the figure for the lower scores, as there are fewer low scores, as well as that for regression and random forest high scores begin to approach 1, while low scores do not approach 0 to the same extent. While in general, the trend is that high scoring heifers from one method also score high in the other methods, as expected from the correlations being medium to high, there also appear to be exceptions, where a heifer scores very different between methods.

3.4 Discussion

We showed that regression, naive Bayes and random forest were able to predict survival to second lactation for dairy cows at a population level. Naive Bayes had the highest AUC value at all decision moments except at 200 days post calving, although performance overlapped with random forest in all decision moments. Logistic multiple regression performed similar to naive Bayes and random forest in the first two decision moments, but was outperformed at first calving, 6 weeks post calving and at 200 days after calving. All methods were significantly different from an AUC of 0.5, but in general AUC values were low, with only random forest achieving 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 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 a low accuracy (Akobeng, 2007). In short, while it was possible to determine which method had the best performance at each decision moment, none of these methods were able to accurately predict individual cow survival. Accurate individual predictions on animals are important for practical applications of the model, as a farmer has interest in the accuracy of the prediction for a single or small group of animals, not the average outcome success of all Dutch Holsteins.

A similar case in literature is the cow pregnancy status was similar to cow survival, as they were both complex traits with a binary outcome. Furthermore, survival and fertility are both genetically and phenotypically related (Pritchard et al., 2013), as fertility issues were found to be one of the main reasons for a cow to be culled (Brickell and Wathes, 2011; Zijlstra et al., 2013). There have been several studies comparing 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). The results of these papers were similar to our study; where naive Bayes performed well when little data is available. Contrary to our results however, regression was found to outperform machine learning methods such as naive Bayes, support vector machines and random forest in some other studies (Hempstalk et al., 2015; Fenlon et al., 2016). Another study that did not investigate any linear methods found random forest to be superior to Bayesian methods (Shahinfar et al., 2014). When looking closely at the studies, however, the performance of the tested methods in these papers were very similar. Furthermore, none of the methods tested were able to predict individual pregnancy outcome well enough to be of use in practice (Hempstalk et al., 2015; Fenlon et al., 2016; Rutten et al., 2016). So, while insight was gained in the best method and mechanisms of predicting a complex trait, ultimately none of the prediction methods in our study or in previous research were useful for individual prediction of a complex binary trait.

In our study, we selected three different methods; the linear method multiple logistic regression, naive Bayes and random forest. These methods were selected because they are each representatives of large groups of similar methods, but naturally there are many other methods able to predict survival. A linear method that was not included in our study was the survival analysis, a commonly used method for the genetic evaluations of survival traits (Cox and Oates, 1984; Ducrocq, 1988). This method is used instead of regression because it is able to make use of uncensored records, therefore having an advantage (Carlén et al., 2005). As the data used in this study was 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. 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 amount of records to be trained. While difficult to apply on our current dataset, neural networks may be of use in future research.

Accurate individual prediction of survival to second lactation could not be achieved by optimizing the choice of prediction method alone, so in future research other ways to improve prediction performance should be considered. Prediction performance may be improved by increasing the number of records. In the case of pregnancy status, results did indeed 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%, which would mean that a full quarter of the animals would still 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 the 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). The variables in this study were chosen as they were readily available on most Dutch farms. In contrast, information on for example animal health and growth are often not available and require additional data collection and cost. Finally, while additional records and variables may improve prediction accuracy, additional information will not solve all difficulties. A model can only predict a non-surviving animal accurately when the cause of death is the result of a pattern found in the data. This requirement is problematic, as for example our study lacked health variables on calves, and therefore calves which died 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 may be based on individual farmer's decisions which cannot always be explained by the available information (Hadley et al., 2006; Huijps et al., 2010). Thus, while we expect that additional information will increase the accuracy, a certain degree of uncertainty will remain.

3.5 Conclusion

All three methods (logistic multiple regression, naive Bayes and random forest) were able to predict survival at population level. At birth and at 18 months, all three methods reported similar AUC values and increased survival in a practical scenario by similar amounts. Naive Bayes did obtain the highest AUC value in all decision moments up to 200 days post calving, but there was always overlap with the model performance of random forest. At 200 days post calving, random forest had the highest AUC, although the overlap with naive Bayes persisted. Interestingly, the three methods appeared to predict individual heifers differently. Correlations of individual predictions for animals were lower than expected, and the models appeared to improve by predicting different groups of animals better. While it was possible to choose a 'best' method for each moment, all methods would have resulted in similar improvement in practice.

3.6 Acknowledgements

This work was part of the Breed4Food research programme, project “Smart animal breeding with advanced machine learning” with project number 14295, which was financed by the Netherlands Organisation 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 was provided by cattle improvement cooperative CRV (Arnhem, the Netherlands).

References

Akobeng, A. K. 2007. Understanding diagnostic tests 3: receiver operating characteristic curves. Acta paediatrica 96(5):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. Computers and electronics in agriculture 105:9-19.

Bennett, D. A. 2001. How can I deal with missing data in my study? Australian and New Zealand journal of public health 25(5):464-469.

Breiman, L. 2001. Random forests. Machine learning 45(1):5-32.

Brickell, J., and D. Wathes. 2011. A descriptive study of the survival of Holstein-Friesian heifers through to third calving on English dairy farms. Journal of Dairy Science 94(4):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. Molecular ecology resources

Carlén, E., M. d. P. Schneider, and E. Strandberg. 2005. Comparison between linear models and survival analysis for genetic evaluation of clinical mastitis in dairy cattle. Journal of Dairy Science 88(2):797-803.

Chok, N. S. 2010. Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data, University of Pittsburgh.

Cortez, P., M. Portelinha, S. Rodrigues, V. Cadavez, and A. Teixeira. 2006. Lamb meat quality assessment by support vector machines. Neural Processing Letters 24(1):41-51.

Cox, D. R., and D. Oates. 1984. Analysis of survival data. Chapman and Hall, Stockholm.

Domingos, P., and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning 29(2-3):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. Engineering Applications of Artificial Intelligence 65:29-42.

Ducrocq, V. P. 1988. AN ANALYSIS OF LENGTH OF PRODUCTIVE LIFE IN DAIRY CATTLE.

Ebrahimie, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski. 2018. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. Computers and Electronics in Agriculture 147(C):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. Agronomy for Sustainable Development 38(3):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.

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. Journal of Dairy science 100(12):9746-9758.

Friedman, J. H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery 1(1):55-77.

Gahegan, M. 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg)? International Journal of Geographical Information Science 17(1):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. Animal Production Science 57(2):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 genetics 12(1):87.

Hadley, G., C. Wolf, and S. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. Journal of dairy science 89(6):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. Journal of dairy science 98(8):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. European review of agricultural economics 37(4):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. Journal of dairy science 99(9):7344-7361.

Jensen, F. V. 1996. An introduction to Bayesian networks. UCL press, London.

Kabra, M., A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. Nature methods 10(1):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. Journal of Dairy science 93(8):3616-3627.

Kotsiantis, S., and P. Pintelas. 2003. Mixture of expert agents for handling imbalanced data sets. Annals of Mathematics, Computing & Teleinformatics 1(1):46-55.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160:3-24.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of statistical software 28(5):1-26.

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3):18--22.

Libbrecht, M. W., and W. S. Noble. 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16(6):321.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17(2):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. Veterinary research 46(1):85.

Majka, M. 2018. naivebayes: High Performance Implementation of the Naive Bayes Algorithm.

McQueen, R. J., S. R. Garner, C. G. Nevill-Manning, and I. H. Witten. 1995. Applying machine learning to agricultural data. Computers and electronics in agriculture 12(4):275-293.

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

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. Journal of Dairy Research 84(4):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(1):34-46.

R core team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

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., W. Steeneveld, J. 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. Journal of dairy science 99(8):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. Journal of dairy science 97(2):731-742.

van der Heide, E., R. Veerkamp, M. van Pelt, C. Kamphuis, R. Veerkamp, and B. J. Ducro. submitted. Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information. Journal of Dairy Science

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

Wallén, S., E. Prestløkken, S. McParland, and D. Berry. 2018. Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. Journal of dairy science 101(7):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(8):1135-1143.

White, B., D. Amrine, and R. Larson. 2018. Big data analytics and precision animal agriculture symposium: Data to decisions. Journal of animal science 96(4):1531-1539.

Zheng, H., H. Wang, and T. Yan. 2016. Modelling enteric methane emissions from milking dairy cows with Bayesian networks. In: Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference, Shenzhen China. p 1635-1640.

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.

3.7 Appendix

TABLE 3.A All 66 phenotypic variables and 50 gEBV used in this study*

Phenotypic variables	Decision moment	continuous
Animal Identification number	Birth	No
Year of birth	Birth	No
Birth farm UBN	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 months	18 months	No
Insemination farm	18 months	No
Insemination season	18 months	No
Countable inseminations	18 months	Yes
Non return status at 18 months	18 months	No
No insemination information at 18 months	18 months	No
Number of farm movements prior to 18 months	18 months	Yes
Age at first insemination	18 months	Yes
Type of first insemination	18 months	No
Number of inseminations	18 months	Yes
Survival status at 2 years 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 prior to calving	first calving	Yes
Age at first calving	first calving	Yes
Calving farm UBN	first calving	No
Calf gender	first calving	No
Calf survival first 24 hours	first calving	No
Calving ease calf	first calving	No
gestation duration calf	first calving	Yes
Birthweight calf	first calving	Yes

TABLE 3.A Continued

Phenotypic variables	Decision moment	continuous
Calf survival first week	first calving	No
Calf survival second week	first calving	No
Calving records exist	first calving	No
Twins	first calving	No
Kg milk at 6 weeks	6 weeks p.c.	Yes
Fat percentage milk at 6 weeks	6 weeks p.c.	Yes
Protein percentage milk at 6 weeks	6 weeks p.c.	Yes
Cell count milk at 6 weeks	6 weeks p.c.	Yes
Urea milk at 6 weeks	6 weeks p.c.	Yes
Lactose percentage milk at 6 weeks	6 weeks p.c.	Yes
Cow status indicator at 6 weeks	6 weeks p.c.	No
Number of negative indications at 6 weeks	6 weeks p.c.	Yes
Number of days in lactation at 6 weeks	6 weeks p.c.	Yes
Complete milk measurement available at 6 weeks	6 weeks p.c.	No
First parity insemination farm UBN	200 days p.c.	No
First parity insemination season	200 days p.c.	No
First parity first insemination type	200 days p.c.	No
Number of inseminations in first parity	200 days p.c.	Yes
Non return status at 200 days post calving	200 days p.c.	No
Age at 200 days post calving	200 days p.c.	Yes
Insemination known in the first parity	200 days p.c.	No
Age at first insemination in the first parity	200 days p.c.	Yes
Number of farm movements at 200 days post calving	200 days p.c.	Yes
Number of known milk testing's at 200 days post calving	200 days p.c.	Yes
Average kg of milk	200 days p.c.	Yes
Average fat percentage of milk	200 days p.c.	Yes
Average protein percentage of milk	200 days p.c.	Yes
Average cell count of milk	200 days p.c.	Yes
Average Urea of milk	200 days p.c.	Yes
Average lactose percentage of milk	200 days p.c.	Yes
Number of negative indications at 200 days post calving	200 days p.c.	Yes
Survival status at 200 days post calving	200 days p.c.	No
Number of farm movements in the first parity	200 days p.c.	Yes

TABLE 3.A Continued

gEBV	Decision moment	continuous
"NVI" Dutch breeding goal standard	Birth	Yes
Kg milk	Birth	Yes
Kg fat	Birth	Yes
Kg protein	Birth	Yes
Kg lactose	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
Non return status at 56 days	Birth	Yes
Interval calving - first insemination	Birth	Yes
Calving interval	Birth	Yes
Interval first - last insemination	Birth	Yes
Conception ratio	Birth	Yes
Claw health	Birth	Yes
Calf vitality 3 - 365 days	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

TABLE 3.A Continued

gEBV	Decision moment	continuous
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

*gEBV; genomic estimated breeding values. This table shows the values and variables available and the decision moment in which each variable is available. In the decision moments, "post calving" is abbreviated to p.c..

IV

CHAPTER 4

Improving predictive performance using ensemble methods in a case study on survival in dairy cattle

E.M.M. van der Heide[1], C. Kamphuis[1],
R. F. Veerkamp[1], I. N. Athanasiadis[2], G. Azzopardi[3],
M. L. van Pelt[4] , B. J. Ducro[1]

Submitted for review in
Computers & Electronics in Agriculture

[1] Wageningen University & Research Animal Breeding and Genomics,
P.O. box 338, 6700 AH Wageningen, the Netherlands; [2] Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, P.O. box 47, 6700 AA, Wageningen, The Netherlands;
[3] University of Groningen, Johann Bernoulli Institute of Mathematics and Computer Science,
9747 AG Groningen, The Netherlands; [4] Cooperation CRV, Animal Evaluation Unit,
P.O. box 454, 6800 AL Arnhem, the Netherlands

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 through individual prediction methods. In this study, we investigated if we could improve prediction of cow survival to second lactation by combining the predictions of multiple (weak) methods in an ensemble method. We tested four ensemble methods: voting rule, multiple logistic regression, random forest and naive Bayes. Five performance metrics were calculated for each ensemble method: precision, recall, balanced accuracy, area under the receiver operator curve (AUC), and the gains in surviving animals in a scenario where the best 50% of these animals was selected. We compared the performance metrics of the ensembles against the performance of the constituent models. 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 outcomes resulted in higher performance metrics than using binary outcomes. Recall, AUC and balanced accuracy values improved significantly for naive Bayes and multiple logistic regression ensembles in at least one dataset, although performance metrics did remain low overall. Multiple logistic regression was the best performing ensemble method, naive Bayes the second-best, and the random forest ensemble method resulted in the least significant improvement over the individual methods. The multiple logistic regression ensemble method resulted in equal or better recall, AUC, balanced accuracy and proportion of animals surviving on all datasets.

KEYWORDS: ensemble method, machine learning, survival, dairy cow

4.1 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 for 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 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 the original methods (Kotsiantis et al., 2006b; Witten et al., 2016). They improve prediction by taking advantage of the underlying differences and strengths of the methods the ensemble is built from. 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 are successfully applied in the prediction of survival in cancer patients (Hothorn et al., 2005; Abreu et al., 2013; Leger et al., 2017). This success in other fields makes it an attractive strategy to improve 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. We tested four different ensemble methods; 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 most simple 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 different groups of statistical prediction methods. Selecting these four different ensemble methods resulted in an overview of the possibilities of ensemble methods to improve prediction of survival in dairy cattle.

4.2 Material and Method

4.2.1 DATA

We used five datasets originating from a previous study that predicted survival to second lactation of individual cows (van der Heide et al., 2019). These five datasets consisted of predictions on the test datasets from the previous study, a randomly selected 30% of all available animals, stratified by survival (Figure 4.1). The data used in the current study is therefore the model output from the original study.

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

Prediction outcomes were obtained at five moments in the life of a cow: at birth, at eighteen months of age, at first calving, at six weeks post calving and at two hundred days post calving. Each dataset contained between 2051 (at birth) to 1862 (at 200 days post calving) randomly selected animals (Table 4.1). 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 calculated 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. The mean chance of survival was 0.86 for the regression and naive Bayes, and 0.50 for the random forest, as the latter method had centred survival around 0.5 (see also (van der Heide et al., 2019)).

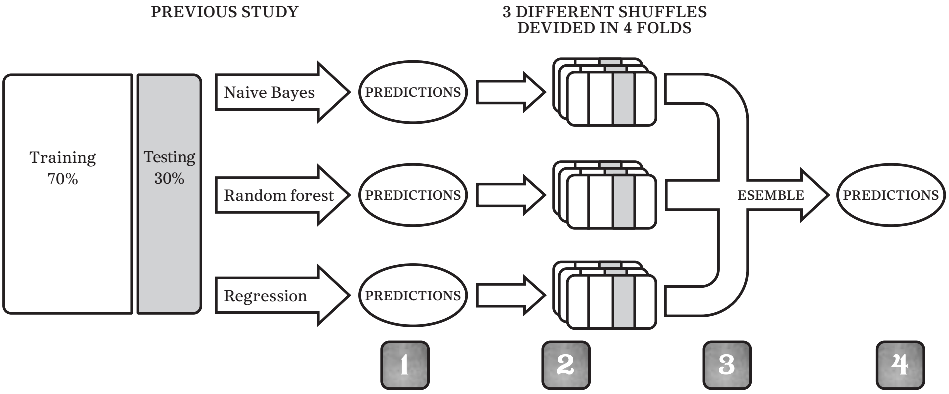


FIGURE 4.1 Schematic depiction of the analysis. The analysis shown was repeated for all datasets: from birth to 200 days post calving, using either continuous or binary outcomes.

TABLE 4.1 Number of animals in each dataset.

Dataset	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 calving	1743	200	1943
200 days post calving	1723	139	1862

4.2.2 MODEL AND ANALYSIS

The datasets were analysed in the statistical program R (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). In order to calculate voting rule, if at least two out of the three original predictions were positive (i.e., animal will survive) was predicted to survive, and any animal with two or more negative predictions was predicted to not survive. No training of the data was required to obtain performance metrics for this method. 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:

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

Where y is the survival status at second calving plus two weeks, X_{i1} through X_{i3} were the predicted outcomes of the three 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 ei. For each of the five datasets (from birth to 200 days post calving) a separate model was created.

For the random forest, we further tested different settings for the hyperparameters of the number of trees, the 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. The number of trees was subsequently set at 200 as there were no significant changes in area under the receiver operating curve (AUC) of balanced accuracy if 150 or more trees were used, regardless of dataset and regardless of if 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.

Figure 4.1 shows a schematic representation of the steps in the analysis. We used four-fold validation to test the random forest, regression and naive Bayes ensemble methods. We repeated this four-fold validation step three times to get a reasonable range of performance metrics to test for significance (Figure 4.1, step 2). The four-fold validation was done for each dataset by splitting each of the three shuffles into four parts, where three of the parts (75%) were used for training the model, and the last part (25%) was used for validation. This was then repeated four times, until every part was used as validation once. The model performance metrics for the regression, naive Bayes and random forest ensemble methods were thus averages of 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. The precision is the proportion of correct non-surviving predictions. Precision is also known as the positive predictive value of the minority class, so the proportion of predictions an animal does not survive that were correct. The recall is equal to the sensitivity of the minority class, which is the proportion of non-surviving animals identified correctly from the total population. Both of these metrics quantify the ability of the ensemble method to identify non-surviving animals correctly. Balanced accuracy and the AUC on the other hand 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 practical measurement of the possible effect of these methods might have in practice. By selecting and only raising heifers with a high prediction outcome, a farmer might reduce the percentage of heifers not reaching second lactation in his herd.

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 results of the ensemble methods and the individual 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.

4.3 Results

First, we will discuss the results of voting rule (the fourth row of results in tables 4.2.A through 4.2.E), as this method proved difficult to compare to the other ensemble methods. The performance metrics for the other ensemble methods are shown in Section 3.2 and Section 3.3, for the datasets with continuous outcomes and binary outcomes respectively (Tables 4.2.A to 4.2.E). In the final section of the results we describe the correlations between the ensemble methods and the individual methods.

TABLE 4.2.A Performance metrics of the predictions at birth.

	Data type	Precision	Recall	Balanced accuracy	ROC	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

[a] significantly outperforms individual method multiple logistic regression,

[b] significantly outperforms individual method random forest,

[c] significantly outperforms individual method naive Bayes,

[x] significantly worse than one or more of the three individual methods.

Note: significance could not be calculated for voting rule results.

TABLE 4.2.B Performance metrics of the predictions at 18 months.*

	Data type	Precision	Recall	Balanced accuracy	ROC	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 4.2.C Performance metrics of the predictions at first calving.*

	Data type	Precision	Recall	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

*for footnotes, see Table 4.2.A.

TABLE 4.2.D Performance metrics of the predictions at 6 weeks p.c.*

	Data type	Precision	Recall	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 4.2.E. Performance metrics of the predictions at 200 days p.c.*

	Data type	Precision	Recall	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	0.706a^{b}	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	0.702a^{c}	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

*for footnotes, see Table 4.2.A. p.c. = post calving.

4.3.1 VOTING RULE

Two metrics could not be calculated for voting rule: the AUC and proportion surviving if the top 50% best animals are selected. This was due to the limited number of possible outcomes and because voting rule does not produce a range of predictions. For example, in all cases more than 50% of the animals received the highest score, and it is not possible to rank animals using voting rule. This means it was not possible to select the highest 50% scoring animals. Voting rule had the highest recall at 18 months of age, calving, 6 weeks post calving and 200 days post calving. However, this came at the cost of having the lowest precision of all methods on those datasets. Voting rule resulted in a lower balanced accuracy (0.001 to 0.018 lower) at the birth, 18 months and first calving. 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.

4.3.2 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 2a and 2b). Furthermore, both also outperformed all three individual methods at recall. At 6 weeks post calving, the regression ensemble method significantly outperformed all three individual methods on balanced accuracy (Table 4.2.d). 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 4.2.a). From 18 months onward, regression improved on at least one individual method at all metrics, and never performed significantly worse than any individual method (Table 4.2.b through 2e). 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 4.2.a through 4.2.e). The random forest 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 methods. For example at 200 days post calving it significantly improved on at least one method in recall and balanced accuracy, but resulted in significantly worse precision, AUC and proportion of heifers surviving than at least one individual method. Random forest performed best at first calving, outperforming at least one individual method on all metrics except AUC.

4.3.3 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 recall, 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 4.2.a). At 18 months, all three methods improved on precision, and regression and naive Bayes also improved on balanced accuracy over the regression individual method (Table 4.2.b). There were no significant differences from the individual methods on the other metrics. From first calving onwards all methods performed equal or better than the individual methods at recall (Tables 2c through 2e). Regression was the only method that did not improve recall over at least one single method at 6 weeks post calving, but was 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 recall. 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 dataset.

4.3.4 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 4.3). 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.

TABLE 4.3 Average correlations between the results of the statistical ensemble methods and the results of the individual methods*

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

* p.c. = post calving

4.4 Discussion

We investigated if the prediction of survival to second lactation in dairy cattle could be improved by using ensemble methods. Ensemble methods have several advantages over using individual methods, such as increased predictive performance and robustness (Woźniak et al., 2014). However, these advantages come at costs such as an increase in model design complexity and less interpretable results (Kotsiantis et al., 2007; Woźniak et al., 2014). The possible advantages of using ensemble methods must thus justify the increase in time and effort required to use these methods. In our study, regression as an ensemble method always resulted in equal or better performance on recall, AUC, balanced accuracy and proportion of surviving animals. It also performed better than at least one individual method on precision from first calving onwards. However, for the other ensemble methods, especially the random forest ensemble, the results were more inconsistent.

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). How much improvement is found in a study may depend on the type of ensemble being tested, as well as the performance metric being used for the comparison between methods. For example, in this study the results show the well-known trade-off between precision and recall (Buckland and Gey, 1994). The random forest single method had the highest precision at birth of any method, significantly outperforming most of the ensemble methods. However, at the same time, it also had the lowest recall, and was significantly outperformed by almost all of the ensemble methods on this metric. Which method is considered better can thus vary depending on which metrics are investigated.

While there appeared to be a benefit of applying the regression ensemble method, performance metrics remained low overall. This meant that while some improvement in prediction could be realized by using an ensemble method, this improvement may only have a small effect in practice. Indeed, looking at the scenario where the top 50% of heifers were selected, none of the ensemble methods significantly outperformed all three individual methods. There may be several reasons why the ensemble methods did not result in a large increase in model performance. A high correlation between the input methods may be one of these. 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 this study, the correlations of the prediction outcomes used as input data were between 0.417 to 0.700 (van der Heide et al., 2019). This was lower than expected as the three methods were trained on the same dataset. However, it is possible that the correlations were still (too) high, limiting the variability among the prediction outcomes. This would in turn have reduced the benefit of applying ensemble methods in this study. Strong correlations between input variables are not only a problem for the effectiveness of the ensemble methods, but it 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. 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). Lastly, it should be noted that survival simply is a very difficult trait to predict. This difficulty is part due to some of the necessary variables missing in original data (van der Heide et al., 2019). For example, there was no information on disease occurrence available, while diseases are an important cause of death in early life (Svensson et al., 2006). So while it was possible to take advantage of the differences between the methods using ensemble methods, there were limitations to increasing model performance by varying the method alone.

Class imbalance is another reason it may be difficult for models to improve the prediction of survival (Stefanowski, 2016). In the case of survival, a majority of 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 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 in this study. This imbalance also made it difficult to compare models based on commonly used performance metrics (Stefanowski, 2016). For example, we did not use the most popular performance metric, accuracy (Hossin and Sulaiman, 2015), because this metric is very sensitive to class imbalance (Kotsiantis et al., 2006a; Ali et al., 2015). By predicting all animals as surviving, an accuracy of 0.86 could be reached, but of course as a model this would not be useful. The imbalance of survival meant that accuracy could be very misleading. Another popular performance metric is the AUC value, which we did report in this study. While more robust, this metric could still be biased, especially in cases with a strong imbalance (Saito and Rehmsmeier, 2015). As all evaluation methods have specific benefits and as drawbacks (Tharwat, 2018), we aimed to provide the reader with a comprehensive selection. In machine learning, choosing the correct performance metric is even more vital, as the performance metric is used to evaluate intermediate steps in determining the final model (Ali et al., 2015; Hossin and Sulaiman, 2015). Selecting a biased performance metric such as accuracy will result in poor performance of the method on the minority class. For example, using the random forest without stratifying for survival would result in the model assigning all animals to the majority class simply because this would result in the highest accuracy. As many machine models use a performance metric as a cost optimization function, it is important to monitor this or adapt the data to avoid problems. Creating or obtaining a more balanced dataset for model development would avoid this problem and possibly even improve model performance.

4.5 Conclusion

Using logistic multiple regression as an ensemble method resulted in equal or better recall, 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. Recall, 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. Class imbalance and lack of underlying variability in the input variables could have resulted in less than optimal results, and are important factors to consider when using an ensemble method. Where multiple prediction models are available, regression can be a useful method to investigate the additional value of using ensemble methods.

4.6 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 n○727213. The data for this study was provided by cattle improvement cooperative CRV (Arnhem, the Netherlands).

References

Abreu, P. H., H. Amaro, D. C. Silva, P. Machado, M. H. Abreu, N. Afonso, and A. Dourado. 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. p 1366-1369.

Ali, A., S. M. Shamsuddin, and A. L. Ralescu. 2015. Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl 7(3):176-204.

Barbareschi, M., S. Del Prete, F. Gargiulo, A. Mazzeo, and C. Sansone. 2015. Decision tree-based multiple classifier systems: An fpga perspective. International Workshop on Multiple Classifier Systems. p 194-205. Springer, Cham.

Barkema, H., M. Von Keyserlingk, J. Kastelic, T. Lam, C. Luby, J.-P. Roy, S. LeBlanc, G. Keefe, and D. Kelton. 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. Journal of dairy science 98(11):7426-7445.

Boulton, A., J. Rushton, and D. Wathes. 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:1-9.

Breiman, L. 2001. Random forests. Machine learning 45(1):5-32.

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.

Buckland, M., and F. Gey. 1994. The relationship between recall and precision. Journal of the American society for information science 45(1):12-19.

Feldwisch-Drentrup, H., B. Schelter, M. Jachan, J. Nawrath, J. Timmer, and A. Schulze-Bonhage. 2010. Joining the benefits: combining epileptic seizure prediction methods. Epilepsia 51(8):1598-1606.

Friedman, N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. Machine learning 29(2-3):131-163.

Gaillard, C., O. Martin, P. Blavy, N. Friggens, J. Sehested, and H. Phuong. 2016. Prediction of the lifetime productive and reproductive performance of Holstein cows managed for different lactation durations, using a model of lifetime nutrient partitioning. Journal of dairy science 99(11):9126-9135.

Grandl, F., M. Furger, M. Kreuzer, and M. Zehetmeier. 2019. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. Animal 13(1):198-208.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications 73:220-239.

Heise, J., Z. Liu, K. F. Stock, S. Rensing, F. Reinhardt, and H. Simianer. 2016. The genetic structure of longevity in dairy cows. Journal of dairy science 99(2):1253-1265.

Hossin, M., and M. Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5(2):1.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan. 2005. Survival ensembles. Biostatistics 7(3):355-373.

Jensen, F. V. 1996. An introduction to Bayesian networks. UCL press, London.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. 2010. Challenges in combining projections from multiple climate models. Journal of Climate 23(10):2739-2758.

Kotsiantis, S., D. Kanellopoulos, and P. Pintelas. 2006a. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30(1):25-36.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160:3-24.

Kotsiantis, S. B., I. D. Zaharakis, and P. E. Pintelas. 2006b. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26(3):159-190.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of statistical software 28(5):1-26.

Larsen, M. L. V., L. J. Pedersen, and D. B. Jensen. 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., A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, and V. Budach. 2017. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Scientific reports 7(1):13206.

Lehmann, J. O., J. Fadel, L. Mogensen, T. Kristensen, C. Gaillard, and E. Kebreab. 2016. Effect of calving interval and parity on milk yield per feeding day in Danish commercial dairy herds. Journal of dairy science 99(1):621-633.

Liaw, A., and M. Wiener. 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., W. Steeneveld, M. Mourits, and H. Hogeveen. 2015. The optimal number of heifer calves to be reared as dairy replacements. Journal of dairy science 98(2):861-871.

Olechnowicz, J., P. Kneblewski, J. Jaśkowski, and J. Włodarek. 2016. Effect of selected factors on longevity in cattle: a review. J. Anim. Plant Sci 26:1533-1541.

Oza, N. C., and K. Tumer. 2008. Classifier ensembles: Select real-world applications. Information Fusion 9(1):4-20.

Pena, M., and H. van den Dool. 2008. Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. Journal of Climate 21(24):6521-6538.

R core team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

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.

Saito, T., and M. Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 10(3):e0118432.

Satopää, V. A., J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. 2014. Combining multiple probability predictions using a simple logit model. International Journal of Forecasting 30(2):344-356.

Seni, G., and J. F. Elder. 2010. Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis lectures on data mining and knowledge discovery 2(1):1-126.

Shahid, M., J. Reneau, H. Chester-Jones, R. Chebel, and M. I. Endres. 2015. Cow-and herd-level risk factors for on-farm mortality in Midwest US dairy herds. Journal of dairy science 98(7):4401-4413.

Sinha, A., H. Chen, D. Danu, T. Kirubarajan, and M. Farooq. 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. p. 333-363.

Svensson, C., and J. Hultgren. 2008. Associations between housing, management, and morbidity during rearing and subsequent first-lactation milk production of dairy cows in southwest Sweden. Journal of dairy science 91(4):1510-1518.

Svensson, C., A. Linder, and S.-O. Olsson. 2006. Mortality in Swedish dairy calves and replacement heifers. Journal of dairy science 89(12):4769-4777.

Tharwat, A. 2018. Classification assessment methods. Applied Computing and Informatics

Tsai, C.-F., and M.-L. Chen. 2010. Credit rating by hybrid machine learning techniques. Applied soft computing 10(2):374-380.

van der Heide, E., R. Veerkamp, M. van Pelt, C. Kamphuis, I. Athanasiadis, and B. Ducro. 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. Journal of dairy science 102(10):9409-9421.

Van Pelt, M., T. Meuwissen, G. de Jong, and R. Veerkamp. 2015. Genetic analysis of longevity in Dutch dairy cattle using random regression. Journal of dairy science 98(6):4117-4130.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Cambridge.

Woźniak, M., M. Graña, and E. Corchado. 2014. A survey of multiple classifier systems as hybrid systems. Information Fusion 16:3-17.

V

CHAPTER 5

Modelling farm-specific effects in the prediction of parity reached in dairy cattle

**E.M.M. van der Heide[1], C. Kamphuis[1], R. F. Veerkamp[1],
M. L. van Pelt[2], B. J. Ducro[1]**

Submitted for review in
Computers & Electronics in Agriculture

[1] Wageningen University & Research Animal Breeding and Genomics,
P.O. Box 338, 6700 AH Wageningen, the Netherlands;
[2] Cooperation CRV, Animal Evaluation Unit, PO Box 454, 6800 AL Arnhem, the Netherlands

Abstract

If the number of parities a dairy cow will reach could be estimated accurately, farmers could optimize their replacement heifer management accordingly. This could increase farm profitability and reduce the early culling of older cows. In this study, we predicted the total number of parities a dairy cow will reach and investigated different approaches to include a farm effect. We obtained data from 40 large Holstein dairy farms in the Netherlands. Parity reached was predicted at first calving (15073 cows) and at second calving (12132 cows). The cows in the data were born between 2000 and 2007, where birthyears 2000 to 2006 were used for training and 2007 was used as validation data. Variables included in the prediction were a range of phenotypic records such as insemination and calving records, milk records when available and estimated breeding values on 162 traits. We tested five different models: (1) the baseline model, containing no farm information, (2) a model containing a farm-ID variable, (3) a model containing nine farm-specific variables (4) a model containing the nine farm-specific variables and 40 farm-average EBV scores and (5) a farm-specific model, where training and validation took place on data from each farm separately. The number of parities reached proved difficult to predict. The highest spearman's rho found was only 0.186 and although mean square errors and mean absolute errors also remained low, this was mainly due to the prediction models predicting conservatively around the mean parity reached. At first calving, the model using both farm variables and farm EBV averages was the best scoring model, and the farm-specific model scored the poorest. At second calving, the farm-specific model was again the worst model and there was little to no difference between the other four models. Accuracies were too low to draw conclusions about the benefit of including a farm specific effect, but farm-specific variables ranked high in variable importance whenever available. Farm-specific models performed the poorest, likely due to the smaller training datasets. Although models accounting for the farm-effect were not significantly better than the baseline model, farm variables ranked high in feature importance scores whenever available, indicating a farm effect does exist.

KEY WORDS: farm effect, survival, prediction, dairy cow, random forest

5.1 Introduction

When a dairy cow is culled or dies, it is replaced in the herd by a young heifer. As it takes on average two years for a replacement heifer to enter production, farmers have to anticipate two years in advance how many heifers will be needed (Heise et al., 2018). Therefore, dairy farmers often rear an excess of replacement heifers to account for unexpected deaths (Mohd Nor et al., 2015b; Overton and Dhuyvetter, 2020). When these excess heifers enter production they are either used to replace an older dairy cow or sold, because most farms maintain a stable dairy herd size (Groenendaal et al., 2004). However, the price of a heifer usually does not cover the rearing costs, resulting in a loss if a heifer is sold (Mohd Nor et al., 2015a). Raising excess heifers therefore often results in an increase in early culling of older cows instead (Overton and Dhuyvetter, 2020).

Culling due to 'excess' cows was reported as 4.6% of all dairy cow culling between 2007 and 2012 in the Netherlands (Zijlstra et al., 2013). The actual number of cows culled for this reason may be even higher, as only one culling reason is routinely reported and most cows are culled for a combination of reasons (Pinedo et al., 2014; van Pelt, 2017). Culling older cows because a younger replacement is available may seem sensible, as young heifers have several advantages over older cows such as improved genetics (Miglior et al., 2017). However, this practice also results in increased rearing costs (Mohd Nor et al., 2015b; Boulton et al., 2017) and fewer cows in the more profitable third and fourth lactations (Lehmann et al., 2016; Grandl et al., 2019). Therefore, culling older cows early to make room for a replacement heifer is not necessarily the best option from an economic point of view (De Vries, 2017).

If farmers could accurately estimate how many parities a young cow is likely to reach, they could anticipate how many replacement heifers will be needed. Unfortunately, accurately predicting which parity an individual cow will reach is very difficult (van der Heide et al., 2020). Therefore, models designed to support heifer management decisions rarely include survival traits tailored to individual cows (Cabrera, 2012; Záhradník and Pokrivčák, 2016). The difficulty of accurately predicting parity reached is due to the complexity of the trait; there are many different reasons a cow may die or be culled (De Vries and Marcondes, 2020). Furthermore, the majority of culling decisions are management decisions, which are not solely based on the biological functioning of the cow (Fetrow et al., 2006). As culling decisions are a combination of cow-specific and farm-specific factors, including a farm effect into a prediction model may be necessary for an accurate prediction of which parity an individual cow will reach. In this study, we investigate if it is possible to accurately predict the total number of parities an individual dairy cow will reach at first and second calving. Additionally, we investigated four approaches to study the added value of including farm effect into the prediction model.

5.2 Material and Method

5.2.1 DATA

We obtained a dataset of dairy cows born between 2000 and 2007 from the cattle improvement cooperative CRV (Arnhem, the Netherlands). All cows were at least 87.5% Holstein, herd book registered, had a known date of death and were not exported abroad during their lives. Cows were further required to have first calving records, at least one insemination recorded prior to first calving and an age at first calving of at minimum 600 days. We selected forty farms with the largest number of cows meeting the criteria. From these forty farms, we created two datasets; one dataset with records available at first calving, and one dataset with records available at second calving. At first calving, the dataset contained 15073 cows and the average number of cows per farm was 377 (St.dev. =133). At second calving, the dataset contained 12132 cows and the average number of cows per farm was 303 (St.dev. =113). At second calving, cows were required to have at least one insemination record in the first parity and completed 305-day lactation records. Not all cows present in the dataset at first calving were also present in the dataset at second calving as cows may have died prior to second calving or did not meet the criteria. If a cow was moved between first and second calving to a dairy farm included in the dataset, the cow would be included in the dataset at second calving. If a cow included at first calving was moved to a farm outside of the dataset, this cow would not be present in the second calving dataset.

The trait of interest in this study was the survival trait ‘parity reached’. Parity reached was calculated as the number of unique calf dates recorded during the lifespan of a cow. For number of cows per parity reached at first and second calving, see Table 5.1.

To predict number of parities reached, 24 phenotypic variables were available at first calving and an additional 44 phenotypic variables were available at second calving. For a full list of variables available, see Appendix 5.A. At first calving (Appendix 5.A.2), the variables available were several birth records, breed of the cow and various insemination and calving records. Insemination records included total number of inseminations before first calving, age at first insemination, season

TABLE 5.1 Number of cows per parity reached at first and second calving.

Parity reached	First calving			Second calving		
	# Training	# Validation	# Total	# Training	# Validation	# Total
1	2170	334	2504.00	-	-	-
2	2518	396	2914.00	2405	386	2791
3	2703	413	3116.00	2615	405	3020
4	2258	378	2636.00	2179	372	2551
5	1541	261	1802.00	1495	254	1749
6	950	174	1124.00	913	173	1086
7	466	80	546.00	448	81	529
8	233	28	261.00	220	30	250
9	99	5	104.00	92	5	97
10	47	0	47.00	44	0	44
11	13	0	13.00	9	0	9
12	5	0	5.00	5	0	5
13	0	0	0.00	0	0	0
14	1	0	1.00	1	0	1

of first insemination and the average time in days between inseminations prior to first calving. For the first three recorded inseminations, number of days between inseminations and type of insemination were available. If the time between inseminations was unknown because a cow had no further inseminations, the variable “time between inseminations’ was set to 0. For example, if a cow had only one insemination prior to first calving, the variables ‘time between first and second insemination’, ‘time between second and third insemination’, and ‘average time between inseminations’ were all set to 0 for that cow. Similarly, variables for second and third insemination type were set to ‘unknown’ for cows without second or third inseminations. Type of insemination was indicated as artificial insemination (AI) by the farmer, AI by a professional, natural mating or pasturing with bull. Calving records included age at first calving in days, the season of calving, the sex of the calf (male, female, stillborn or unknown) and the 24-hour survival of the calf (survived, died or unknown). In the case of twins, the data from the first recorded twin was included in the calving records, as the number of twin-births was less than 2% of the total number of births.

At second calving, additional phenotypic records included first parity insemination records and second parity calving records (Appendix 5.A.3). The same insemination and calving variables were available in this dataset as those described for the first calving dataset. The second calving dataset also included milk information from the first lactation and number of moves during the first parity. Milk records included length of lactation and 305-day averages of milk (kg), protein (%), fat (%) and somatic cell count, as well as the number of times the following conditions were reported during the entire lactation: 3-teat milking, mastitis, cow was sick (other than mastitis), freshly calved or calved early. Specific records were further included for the first two test day milk records available: milk (kg), protein (%), fat (%), cell count, number of days in milk at test day milking and if one of the five listed conditions were reported. This data was included separately since farmers might make cow management decisions based specifically on the first few test milk day records. In 1294 individual test day milk records, either protein (%), fat (%) or both had to be imputed using rfImpute from the randomForest package (Liaw and Wiener, 2002). Other test day milk variables, such as milk production (kg) were always available and used for the imputation. A binary variable was created to indicate if there were imputed lactation records for that cow.

In addition to phenotypic information, each cow had records in 162 estimated breeding values (EBV), including health, fertility and production traits (For a full list, see Appendix 5.A.1). The EBV's were obtained from a national database on august 2017 and are based on a 2015 reference population. We used sire and maternal grandsire EBV's as the EBV's of the cows themselves could not be used. This was because EBV's stored by breeding companies are periodically updated with new phenotypic information. This means that all cow EBV's in this study, with birthyears between 2000 and 2007 were updated with phenotypic information up to and including the parity they reached. In early trials of this study, this resulted in very favorable results as the models were able to extract this information from the EBV values. For sires, which were updated with the EBV of all their daughters, this effect is still present but reduced.

The EBV of a cow was calculated as follows:

$$\text{EBVsire} + (\text{ EBVmaternal\_grandsire} \times 0.5\ )$$

For all cows in the data, we had a total of 2104 sires and 2021 grandsires. In some cases, a limited number of EBV's was missing for either sire or maternal grandsire. These values were imputed using rfImpute based on the mean parity reached by their daughters. No bull (sire or maternal grandsire) was missing more than 20% of their EBV's, and no included EBV had more than 25% missing records in total. A variable was created to indicate if the sire or maternal grandsire had missing EBV's prior to imputation.

5.2.2 ANALYSIS

We used the random forest machine learning method (Breiman, 2001; Qi, 2012) to predict parity reached in dairy cattle at birth and at first calving. The data was analyzed in the statistical program R (Team, 2016), using the packages 'mlr' (Bischl et al., 2016; Probst et al., 2017) to implement 'randomForest'(Liaw and Wiener, 2002). Random forest was applied by first tuning on the training data, which was the data from 2000 to 2006, and then validating on the 2007 data. We used 7-fold validation to determine the optimal parameters for each model from the following ranges: 2 to 200 variables per split, 1 to 100 cows per end-node and 10 to 1200 possible trees. The maximum number of variables per split was set to 100, 150 or 200 for hyperparameter tuning, whichever value did not exceed the total amount of variables in a dataset. We used 7-fold instead of the more commonly used 10 as there was 7 years' worth of data in the dataset used for hyperparameter tuning.

To investigate the effect of farm, we analyzed the data in five different ways. For an overview of the five different set-ups and their main (dis)advantages, see Table 5.2. To establish a baseline, the first set-up ignored farm-effect entirely. This meant the model was trained and validated on all available cows from the training and validation dataset, while no farm ID variable was included. In the second model, a farm-ID variable was included for each cow. This would allow the model to use this farm-ID variable to split the data if this resulted in a better prediction.

TABLE 5.2 Overview of the different models and their advantages.

Model		# training examples	Can be used to predict new farms	Potential to cover all aspects farm effect
1	Baseline	Normal	Yes	No
2	Baseline + Farm ID	Normal	No	Yes
3	Baseline + Farm variables	Normal	Yes[1]	No
4	Baseline + Farm variables & EBV averages	Normal	Yes[12]	No
5	Farm specific	Small but \specific	No	Yes

[1]Although we did not have records outside of the data collection period this data would be obtainable for application in practice. [2]EBV records for farms not in the data can be obtained.

The third set-up did not use a farm-ID variable, but included nine farm-specific variables: mean parity reached for first parity cows in 2000, mean parity reached for first parity cows in 2001, proportion of cow deaths due to culling, average number of cows in the first parity between 2000 and 2006, average percentage of growth (or decline) in number of cows in the first parity per year between 2000 and 2006 and the average percentage of growth (or decline) in number of cows in the first parity per year between 2004 and 2006. These variables could explain a portion of the farm effect, indicating for example if a farm had been expanding or shrinking its herd in recent years. For each cow, we also calculated the number of cows in the same parity and farm the year prior, the number of cows in the same parity and farm two years prior, and the difference between these two variables. For example, for a first parity cow born on farm A in 2002, these variables indicated the number of cows in the first parity on farm A in 2001, the number of cows in the first parity on farm A in 2000 and the difference between these numbers. If a previous year was not included in the dataset, as was the case for cows born in 2000 and 2001, the mean number of cows per year between 2000 and 2006 was used. The three variables indicating number of cows in the previous two years bring the total to nine farm-specific variables included in the third set-up. A benefit of using these variables as opposed to simply using farm ID was that these variables could be useable even for new farms (farms not in the training dataset). When using farm ID, all the phenotypic and genomic variables used in this study would have to be known for a training population for the model to properly fit the farm effect. Using these variables, which were calculated from simple records on cow numbers and deaths which are collected for administrative purposes, farm-specific effects could be fitted even for new farms without training data. The down-side is that this restricts the model to using the variables provided, rather than capturing effects which are not explicitly modelled as is the case when using Farm ID instead. This means that farm effects outside of those explicitly provided may not be captured by this model. The fourth set-up expanded on the model with farm variables by including the average EBV scores per farm for 40 of the 162 EBVs. In this set-up, additional information is provided which could explain farm policies, for example if the farm breeds specifically for high milk production or fertility, which could explain some of the differences in culling policies. The final set-up was used to build separate prediction models per farm. This was done by splitting the dataset into 40 sub-datasets (one for each farm), training and validating the model separately for each farm. For hyperparameter tuning, we used the mean value for the number of trees, variables and number of cases per end-node across farms (rounded to an integer value). For this fifth set-up, the reported results are the mean results across farms.

To evaluate the models, we obtained the mean square error and the mean absolute error. These are the recommended metrics to compare regression models (Bischl et al., 2016). We further provide Spearman's' rho statistic to describe the ranking of the model based on predicted parity reached versus the actual ranking (Rosset et al., 2005). We chose to provide a ranking metric as this was the intended use of the model in practice. We further also investigate the importance scores of each model and compared the ranking of variables between models.

5.3 Results

The best-scoring model on the dataset at first calving was the model using both farm variables and farm EBV averages (Table 5.3). This model had the highest spearman's rho and the lowest MAE and MSE. The worst-scoring model was the farm-specific model, which was also the model with the least amount of data for training. Except for the farm-specific model, all models performed very similarly. At second calving, the results were very similar (Table 5.4). The farm-specific model was again the worst model, and there was little to no difference between the other four models.

TABLE 5.3 Performance metrics for the models predicting at first calving.

	Spearman's rho	Mean absolute error	Mean square error
Baseline model	0.140	1.459	3.136
Baseline + Farm ID	0.147	1.455	3.136
Baseline + Farm variables	0.146	1.465	3.174
Baseline + Farm variables & EBV averages	0.158	1.449	3.099
Farm specific model	0.116	1.475	3.236

TABLE 5.4 Performance metrics for the models predicting at second calving.

	Spearman's rho	Mean absolute error	Mean square error
Baseline model	0.186	1.269	2.385
Baseline + Farm ID	0.181	1.275	2.404
Baseline + Farm variables	0.185	1.275	2.398
Baseline + Farm variables & EBV averages	0.181	1.274	2.405
Farm specific model	0.107	1.312	2.556

As the correlation coefficients were quite low, it is surprising that the MAE and MSE also remained relatively low. The MAE indicates the absolute mean difference between the predicted and actual parity reached and was therefore a useful metric for the practical performance of the model. The MAE was 1.449 and 1.274 for the best models at first and second calving respectively. This meant that if a cow was predicted to reach parity 3, that cow would have actually reached parity 2 or 4 on average. As parity reached is a range of values between 1 and 14 in our dataset with a mean of 4, an average error of less than 2 parities might be useful. Unfortunately, the reason the MAE and MSE remained relatively low is because the models predict conservatively around the mean parity reached (Figures 5.1 and 5.2). These figures show the predicted and actual parity reached of all cows in the validation dataset. At first calving, the models mainly predict cows to reach between the second and fifth parity, with most of predictions between parities three and four. At second calving, the models mainly predict between the third and fifth parity, with most of cases predicted around the fourth parity. Despite small differences in the performance metrics between the different models, all five models had almost identical performance. The only differences between models in the figures were in the extreme parities (Parity reached >7), except for slightly larger standard deviations for the farm-specific models overall. In the extreme parities, the differences between models were likely due to the limited amount of training and testing cases in this group (Table 5.1).

We further investigated the importance scores per model to see if there was added value in including a farm effect in some way. As the values of the feature importance score are scaled differently depending on the model, they cannot be compared directly. However, we could investigate which features ranked highest across models (Table 5.5 and Table 5.6). At both first and second calving, farm specific variables such as ‘Per farm average parity reached in 2001’ ranked in the top five of the twenty-five highest scoring variables at first and second calving, respectively. This indicates that when these variables are available, they are often selected to create splits by the trees in the forest. The variable ‘farm ID’ from the Farm ID model was also the most important variable when this variable was available. Other farm variables of interest were variables related to herd size. The ‘Average number of first parity cows between 2000 and 2006’, ‘Number of cows in first parity last year’, ‘Number of cows in first parity two years ago’ and the ‘difference in number of cows between last year and the year before’ all ranked in the twenty-five highest ranking variables. The farm-average EBV values ranked very low without exception.

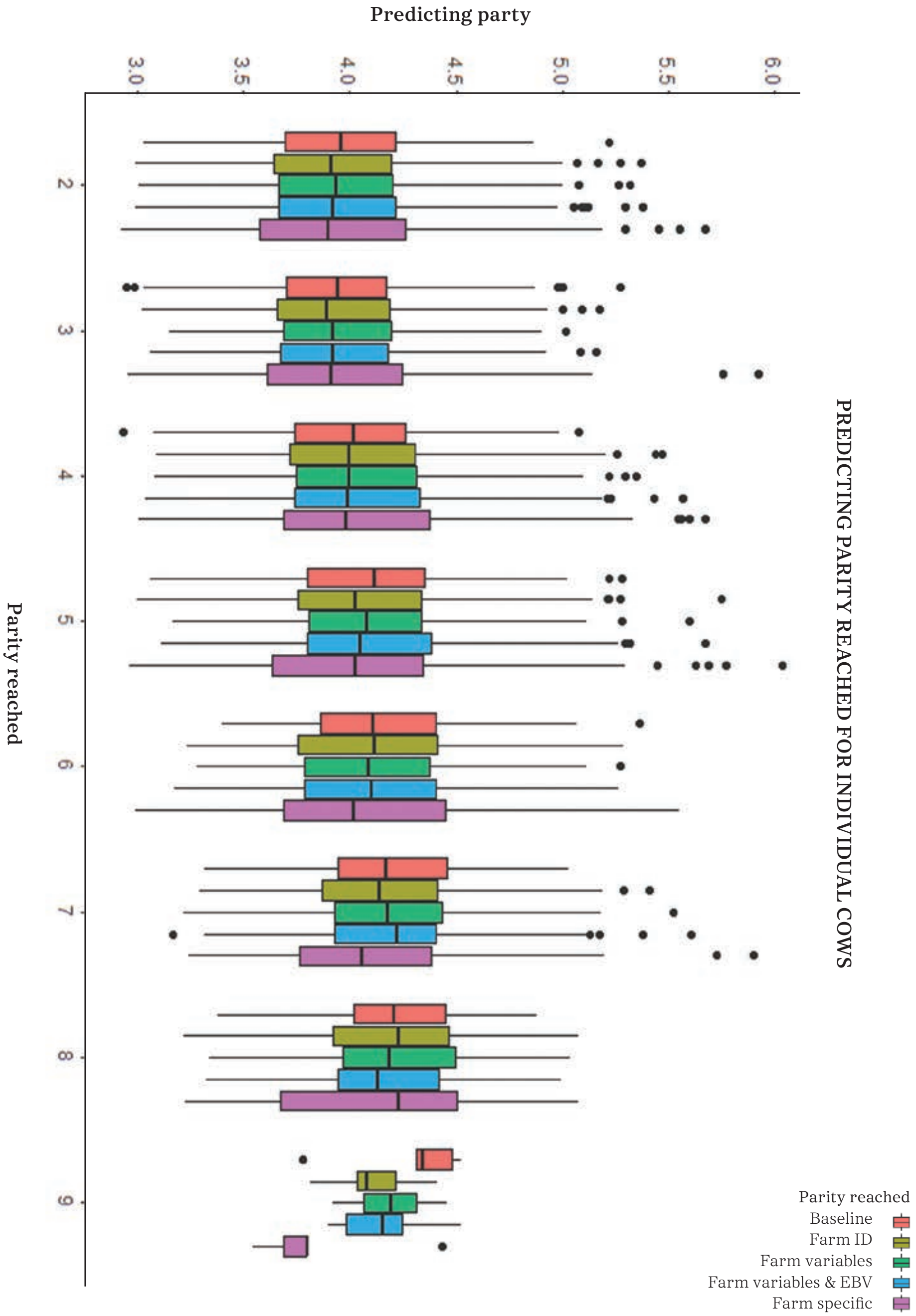


FIGURE 5.1. Predicted vs realized parity reached at first calving. Results from the validation dataset; cows born in 2007.

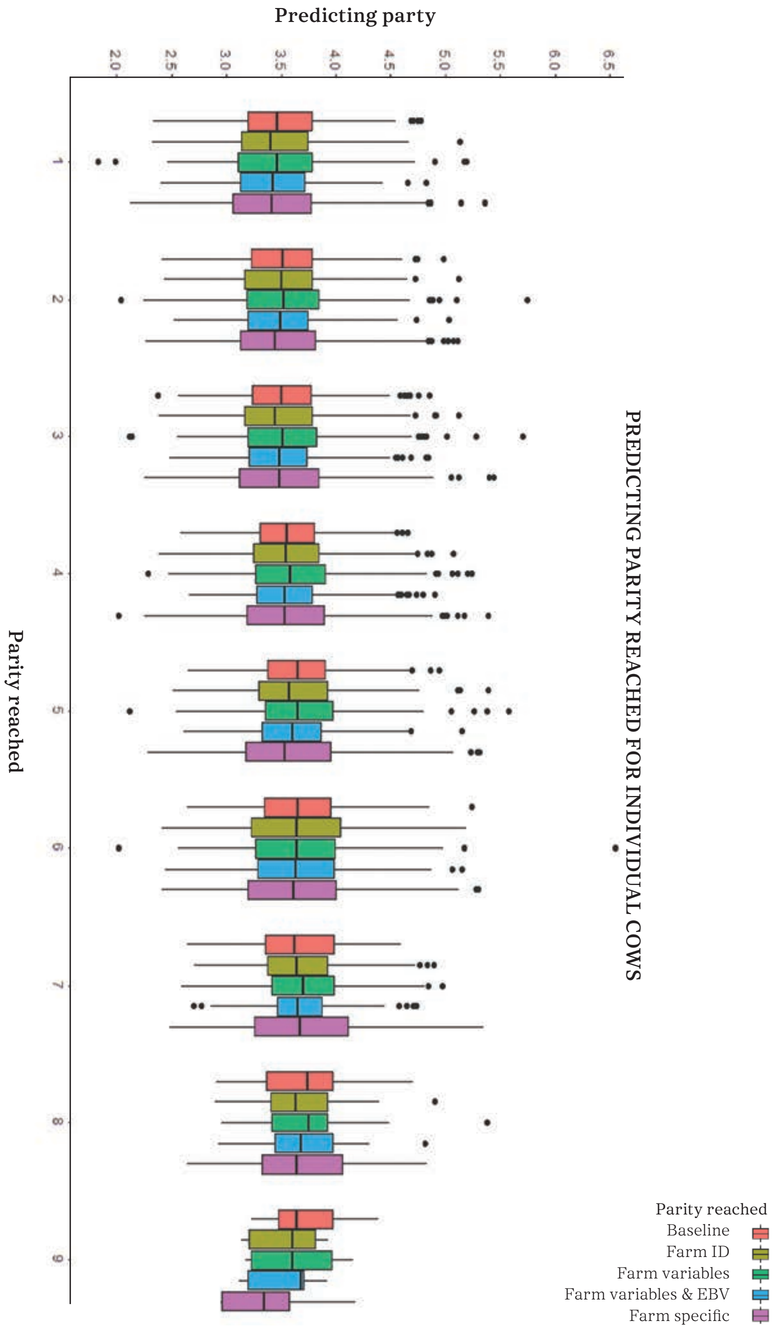


FIGURE 5.2. Predicted vs realized parity reached at second calving. Results from the validation dataset; cows born in 2007.

TABLE 5.5 Twenty-five highest ranked variables based on average rank across models at first calving dataset.

Variable name	Model					
	1	2	3	4	5	Average rank
Farm ID[1]	-	1	-	-	-	1
Age at first calving	1	2	1	1	1	1.2
Per farm average parity reached in 2000[2]	-	-	2	2	-	2
Per farm average parity reached in 2001[2]	-	-	4	3	-	3.5
EBV Lifetime Fat Production	2	3	9	4	3	4.2
Age at first insemination	3	5	3	9	2	4.4
EBV Longevity without predictors	5	6	16	5	4	7.2
Dutch Value Index EBV (NVI)	6	7	19	11	5	9.6
Number of cows in first parity 2 years ago[2]	-	-	6	14	-	10
EBV Longevity with predictors	9	8	22	7	6	10.4
EBV Lifetime Production	7	9	24	6	9	11
EBV Lifetime Milk Production	10	10	27	10	7	12.8
EBV Lifetime Protein Production	8	11	31	8	12	14
EBV Better Life Health (index)	4	4	25	12	28	14.6
Difference in number of first parity cows between last year and the year before[2]	-	-	5	25	-	15
Average number of first parity cows between 2000 and 2006[2]	-	-	15	17	-	16
EBV Direct calving ease	15	15	29	21	8	17.6
Number of cows in first parity last year[2]	-	-	7	29	-	18
EBV Fat score	13	14	30	15	19	18.2
Parity of dam	11	13	14	39	18	19
EBV Fat score	19	18	34	19	11	20.2
EBV Teat length	17	16	28	27	16	20.8
EBV Temperament	12	12	26	20	34	20.8
EBV Saved Feed for Maintenance	32	25	40	18	15	26
EBV Rump Angle	22	24	33	36	17	26.4

[1]Only in Baseline + Farm ID model

[2]Only in Baseline + Farm variables and Baseline + Farm variables and farm EBV models.

TABLE 5.6 Twenty-five highest ranked variables based on average rank across models at second calving dataset.

Variable name	Model					
	1	2	3	4	5	Average rank
Farm ID[1]	-	1	-	-	-	1
Age at second calving	1	2	2	1	1	1.4
Per farm average parity reached in 2001[2]	-	-	1	2	-	1.5
Length of first lactation	2	3	4	3	2	2.8
Per farm average parity reached in 2000[2]	-	-	3	4	-	3.5
Cell count on second test milk day	3	4	5	5	4	4.2
EBV Lifetime Fat Production	4	5	6	6	9	6
Cell count on first test milk day	5	6	8	7	6	6.4
Age at first insemination after first calving	7	7	7	9	3	6.6
First lactation value	6	8	10	8	18	10
Protein % on first test milk day	11	11	15	13	7	11.4
EBV Better Life Health (index)	9	10	14	10	20	12.6
EBV Longevity without predictors	22	9	11	12	10	12.8
Dutch Value Index EBV (NVI)	12	12	13	11	17	13
Protein % on second test milk day	8	15	17	14	13	13.4
Age at first calving	18	14	16	18	5	14.2
EBV Longevity with predictors	28	13	12	15	8	15.2
Fat % on first test milk day	13	16	22	16	23	18
305-day average Milk (kg)	16	19	18	19	19	18.2
Age at first insemination in the first parity	10	20	23	17	27	19.4
305-day average Fat (%)	15	21	25	20	16	19.4
EBV Lifetime Production	24	17	19	26	12	19.6
305-day average Protein (kg)	20	18	20	22	28	21.6
Fat % on second test milk day	17	22	28	21	21	21.8
Milk production (kg) on first test milk day	19	27	31	23	14	22.8

[1]Only in Baseline + Farm ID model, [2]Only in Baseline + Farm variables and Baseline + Farm variables and farm EBV models

Next to farm characteristics, variables that rank highly were almost always related to fertility, production and (udder) health. At first calving, the majority of high ranking variables were EBVs, as no phenotypic information is available on production or health traits. At second calving, phenotypic variables for fertility, production and udder health were available and rank higher than the EBV's instead. The EBV's that remain in the top twenty-five describe production across the lifetime of the cow, for example 'EBV Lifetime fat production' and 'EBV lifetime production', as well as the EBV for longevity.

5.4 Discussion

We investigated if it is possible to accurately predict the total number of parities a dairy cow will reach. We compared four different methods to include a farm effect, while also considering limitations of each of these options. The number of parities reached for an individual cow proved difficult to predict. One of the main reasons why survival traits like parity reached are difficult to predict in general is because the number of new records (cows reaching the end of their lifespan) is limited. Random forest, like all machine learning methods, is highly dependent on the amount of records available to obtain accurate predictions (Zheng and Jin, 2019). The purpose of this study was to investigate different methods to include a farm-effect and therefore only the largest farms in the dataset had enough records per farm. Next to low numbers of records, variables for some causes of death were missing, which could also have contributed to the low prediction accuracy. For example, both lameness and calving problems are important factors in dairy cow survival (De Amicis et al., 2018; Dolecheck and Bewley, 2018), but in the current study no phenotypes were available for lameness and phenotypes for calving ease could not be used due to the large number of missing records for this trait.

The low accuracy across models in this study made it difficult to compare the different strategies to include the farm-effect. We did find differences in the average number of parities reached between farm, which indicated farm-specific culling policies. Farm-specific differences in survival were also found in literature (Boer and Zijlstra, 2013; Armengol and Fraile, 2018). Furthermore, farm variables consistently ranked in the top twenty-five variables when included in the data, indicating a farm-effect. Variables related to herd size, such as 'number of cows in first parity last year' and 'difference in number of first parity cows between last year and the year before' ranked high in the feature importance rankings at both first and second calving, for example. The number of cows in the first parity in any given year is correlated with total herd size of that farm. Herd size is often included as a farm-specific effect or correction factor for farm because this trait is almost always available and has been shown to influence survival (Hadley et al., 2006; Boer and Zijlstra, 2013; De Vries, 2020). The Farm ID variable stood out especially, as it scored over five times higher than the next highest variables at both first and second calving (data not shown).

The high feature importance score of Farm ID may have been caused by the tendency of random forests to prefer to split in factors with large numbers of classes (Strobl et al., 2007; Boulesteix et al., 2012). Farm ID was a factorial variable with 40 classes, ten times more classes than the next largest categorical variables (Appendix 5.A.2 and Appendix 5.A.3). Another possibility for the high feature importance score of Farm ID was that the model used this variable to predict a farm-average value for ‘parity reached’ for most cows. A simple farm-average may have been the best predictor for ‘parity reached’ for many cows given the low predictive ability of the model in general, because the variables ‘average parity reached on a farm in 2000’ and the same variable for 2001 were also ranked highly when available.

An argument against the existence of a (relevant) farm effect is the fact that the baseline model had the best predictive performance at second calving (Table 5.4). One possible explanation why the baseline model performed better than the ‘Farm ID’-, ‘Farm variables’-, and ’Farm variables + EBV’-models is because the baseline model included fewer variables. The addition of variables that do not have much explanatory power can lead to overfitting, although random forests are known to be robust against this problem (Cutler et al., 2012). We chose not to include a variable selection step because (1) the number of variables was much smaller than the number of records whereas variable selection is most often done when the number of variables equals or exceeds the number of records, (2) we were interested in the effect of farm variables and minor farm effects could be removed during a feature selection step and (3) because the large variability in variables in our dataset could have led to bias in common selection procedures (Strobl et al., 2007; Genuer et al., 2010). The addition of uninformative variables could be the cause of the difference in ranking between the ‘Baseline’-, Farm ID’-, ‘Farm variables’-, and ’Farm variables + EBV’-models. However, model performance is also always highly overlapping between these four models and it is extremely unlikely that these differences are statistically significant.

The predictive performance of the models in this study was too low to be useful for selection of cows. In order to be useful in practice, predictive performance must be increased overall. The most obvious solution to improve predictive performance of a machine learning method is to increase the amount of relevant data available (Domingos, 2012). This is especially challenging for survival traits such as number of parities reached, as the number of records per farm is limited. The average farm in the Netherlands had 103 dairy cows in 2018 (Agrimatie, 2018), which means that only a few dozen new records are generated per farm per year. Furthermore, survival traits are not static: survival traits and factors contributing to survival change over time through breeding and changes in the dairy industry (Heise et al., 2016; Van Pelt et al., 2016; Compton et al., 2017). For example, in 2015 the milk quota system was abolished in the Netherlands, which led to many dairy farms growing their herd which reduced culling rates (Huettel and Jongeneel, 2011; Läpple and Sirr, 2019). Then in 2017, this trend reversed with the introduction of new Dutch phosphate regulations which restricted herd size, increasing culling rates in the following years and resulting in the voluntary culling of young cows not normally at risk (for example (Beekman, 2019). This means culling patterns may not be similar between different birthyears, particularly between birthyears that are further apart. Including fluctuations in culling patterns caused by a variety of internal and external factors is a major challenge in predicting a long-term survival trait like number of parities reached.

Due to limitations in obtaining relevant training data, the prediction of survival traits in dairy cows is likely to remain limited to more narrowly defined traits or predictions in the near future (as opposed to predicting events decades into the future). There appears to be a trend for higher accuracies if the moment that is being predicted is closer to when the prediction takes place (Hadley et al., 2006; van der Heide et al., 2019). However, there is a trade-off between prediction accuracy and usefulness of the prediction because culling decisions are often made months prior to a culling event. If the prediction takes place after the decision, the prediction only reflects rather than actually predict farmers decisions (Shmueli, 2010). To investigate what is necessary for (farm-specific) prediction of survival traits to be viable, it may be worthwhile to predict on a different livestock species. Survival traits in sows are similar to survival traits in dairy cows: both traits combine animal-specific factors like production and fertility with farm-characteristics and farm management factors (Soltész et al., 2016; Moeller et al., 2019) and there is evidence that culling policies could be improved through modelling in both species (Bergman et al., 2018). Similar to dairy cows, existing models created to assist replacement policies do not include predictions of survival traits like number of parities reached (Plà, 2007; Hindsborg and Kristensen, 2019). The benefit of predicting on sows instead of dairy cows is that the number of sows per farm is often larger, sow parities are shorter and sow replacement rates tend to be higher than those of dairy cows, resulting in larger and more relevant training datasets (Malanda et al., 2019). Experience on predicting parity reached in sows could then be leveraged to improve the prediction of parity reached in dairy cattle.

5.5 Conclusion

The aim of this study was to investigate if it was possible to accurately predict the total number of parities an individual dairy cow will reach and the role farm-effect plays in this prediction. Accurately predicting the number of parities reached proved very difficult. Performance was overlapping for a baseline model without farm effects and models including a variety of different farm variables. Farm-specific models performed poorest regardless of prediction moment, likely due to their small training sets. Although models accounting for the farm-effect were not significantly better than the baseline model, farm variables ranked high in feature importance scores whenever available, indicating a farm effect does exist.

5.6 Acknowledgements

This work was part of the Breed4Food research program, project 'Smart cow 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 n○727213. The data for this study was provided by cattle improvement cooperative CRV (Arnhem, the Netherlands).

References

Armengol, R., and L. Fraile. 2018. Descriptive study for culling and mortality in five high-producing Spanish dairy cattle farms (2006–2016). Acta Veterinaria Scandinavica 60(1):45.

Beekman, J. 2019. Verlengen levensduur koe levert 56.000 op Mlekvee100Plus. Misset Uitgeverij B.V. , *https://www.melkvee100plus.nl*.

Bergman, P., Y. T. Gröhn, P. Rajala-Schultz, A.-M. Virtala, C. Oliviero, O. Peltoniemi, and M. Heinonen. 2018. Sow removal in commercial herds: Patterns and animal level factors in Finland. Preventive veterinary medicine 159:30-39.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. 2016. mlr: Machine Learning in R. The Journal of Machine Learning Research 17(1):5938-5942.

Boer, H., and J. Zijlstra. 2013. Verschillen tussen bedrijven in levensduur van melkkoeien= Differences between dairy farms in longevity of dairy cows, Wageningen UR Livestock Research.

Boulesteix, A. L., S. Janitza, J. Kruppa, and I. R. König. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6):493-507.

Boulton, A., J. Rushton, and D. Wathes. 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:1-9.

Breiman, L. 2001. Random forests. Machine learning 45(1):5-32.

Cabrera, V. 2012. A simple formulation and solution to the replacement problem: A practical tool to assess the economic cow value, the value of a new pregnancy, and the cost of a pregnancy loss. Journal of dairy science 95(8):4683-4698.

Compton, C., C. Heuer, P. T. Thomsen, T. Carpenter, C. Phyn, and S. McDougall. 2017. Invited review: A systematic literature review and meta-analysis of mortality and culling in dairy cattle. Journal of dairy science 100(1):1-16.

Cutler, A., D. R. Cutler, and J. R. Stevens. 2012. Random forests, Ensemble machine learning. Springer. p. 157-175.

De Amicis, I., M. C. Veronesi, D. Robbe, A. Gloria, and A. Carluccio. 2018. Prevalence, causes, resolution and consequences of bovine dystocia in Italy. Theriogenology 107:104-108.

De Vries, A. 2017. Economic trade-offs between genetic improvement and longevity in dairy cattle. Journal of Dairy Science 100(5):4184-4192.

De Vries, A. 2020. Symposium review: Why revisit dairy cattle productive lifespan? Journal of Dairy Science 103(4):3838-3845.

De Vries, A., and M. Marcondes. 2020. Overview of factors affecting productive lifespan of dairy cows. animal 14(S1):s155-s164.

Dolecheck, K., and J. Bewley. 2018. Animal board invited review: Dairy cow lameness expenditures, losses and total cost. animal 12(7):1462-1474.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM 55(10):78-87.

Fetrow, J., K. Nordlund, and H. Norman. 2006. Invited review: Culling: Nomenclature, definitions, and recommendations. Journal of dairy science 89(6):1896-1905.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot. 2010. Variable selection using random forests. Pattern recognition letters 31(14):2225-2236.

Grandl, F., M. Furger, M. Kreuzer, and M. Zehetmeier. 2019. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. Animal 13(1):198-208.

Groenendaal, H., D. Galligan, and H. Mulder. 2004. An economic spreadsheet model to determine optimal breeding and replacement decisions for dairy cattle. Journal of Dairy Science 87(7):2146-2157.

Hadley, G., C. Wolf, and S. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. Journal of dairy science 89(6):2286-2296.

Heise, J., Z. Liu, K. F. Stock, S. Rensing, F. Reinhardt, and H. Simianer. 2016. The genetic structure of longevity in dairy cows. Journal of dairy science 99(2):1253-1265.

Heise, J., K. F. Stock, F. Reinhardt, N.-T. Ha, and H. Simianer. 2018. Phenotypic and genetic relationships between age at first calving, its component traits, and survival of heifers up to second calving. Journal of dairy science 101(1):425-432.

Hindsborg, J., and A. R. Kristensen. 2019. From data to decision–Implementation of a sow replacement model. Computers and Electronics in Agriculture 165:104970.

Huettel, S., and R. Jongeneel. 2011. How has the EU milk quota affected patterns of herd-size change? European Review of Agricultural Economics 38(4):497-527.

Läpple, D., and G. Sirr. 2019. Dairy Intensification and Quota Abolition: A Comparative Study of Production in Ireland and the Netherlands. EuroChoices

Lehmann, J. O., J. Fadel, L. Mogensen, T. Kristensen, C. Gaillard, and E. Kebreab. 2016. Effect of calving interval and parity on milk yield per feeding day in Danish commercial dairy herds. Journal of dairy science 99(1):621-633.

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3):18--22.

Malanda, J., P. Balogh, and G. N. Dankó. 2019. Sow removal patterns in commercial breed-wean herds of Midwest, USA. Acta Agraria Debreceniensis (2):9-14.

Miglior, F., A. Fleming, F. Malchiodi, L. F. Brito, P. Martin, and C. F. Baes. 2017. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. Journal of dairy science 100(12):10251-10271.

Moeller, G., J. D. Stock, A. K. Johnson, and K. J. Stalder. 2019. A review of aetiology and risk factors affecting sow mortality. CAB Reviews 14(26):1.

Mohd Nor, N., W. Steeneveld, T. Derkman, M. Verbruggen, A. Evers, M. De Haan, and H. Hogeveen. 2015a. The total cost of rearing a heifer on Dutch dairy farms: calculated versus perceived cost. Irish veterinary journal 68(1):29.

Mohd Nor, N., W. Steeneveld, M. Mourits, and H. Hogeveen. 2015b. The optimal number of heifer calves to be reared as dairy replacements. Journal of dairy science 98(2):861-871.

Overton, M., and K. Dhuyvetter. 2020. Symposium review: An abundance of replacement heifers: What is the economic impact of raising more than are needed? Journal of Dairy Science

Pinedo, P., A. Daniels, J. Shumaker, and A. De Vries. 2014. Dynamics of culling for Jersey, Holstein, and Jersey× Holstein crossbred cows in large multibreed dairy herds. Journal of dairy science 97(5):2886-2895.

Plà, L. 2007. Review of mathematical models for sow herd management. Livestock Science 106(2-3):107-119.

Probst, P., Q. Au, G. Casalicchio, C. Stachl, and B. Bischl. 2017. Multilabel classification with R package mlr. arXiv preprint arXiv:1703.08991

Qi, Y. 2012. Random forest for bioinformatics, Ensemble machine learning. Springer. p. 307-323.

R core team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rosset, S., C. Perlich, and B. Zadrozny. 2005. Ranking-based evaluation of regression models. In: Fifth IEEE International Conference on Data Mining (ICDM'05). p 8 pp.

Shmueli, G. 2010. To explain or to predict? Statistical science 25(3):289-310.

Soltész, A., Á. B. Hunyadi, S. Kusza, and P. Balogh. 2016. Survival analysis of sow longevity and lifetime reproductive performance–Review. Acta Agraria Debreceniensis (70):75-80.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8(1):25.

van der Heide, E., R. Veerkamp, M. van Pelt, C. Kamphuis, I. Athanasiadis, and B. Ducro. 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. Journal of dairy science 102(10):9409-9421.

van der Heide, E., R. Veerkamp, M. van Pelt, C. Kamphuis, and B. Ducro. 2020. Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information. Journal of dairy science 103(1):556-571.

van Pelt, M. 2017. Genetic improvement of longevity in dairy cows. Wageningen University.

Van Pelt, M., V. Ducrocq, G. De Jong, M. Calus, and R. Veerkamp. 2016. Genetic changes of survival traits over the past 25 yr in Dutch dairy cattle. Journal of dairy science 99(12):9810-9819.

Záhradník, M., and J. Pokrivčák. 2016. Decision support tool for replacement heifer management: A strategy comparison. Internafional Scienfific Days 2016:" The Agri-Food Value Chain: Challenges for Natural Resources Management and Society:1002-1008.

Zheng, W., and M. Jin. 2019. Effects of Training Data Size and Class Imbalance on the Performance of Classifiers. In: Conference on Artificial Intelligence and Natural Language. p 3-17.

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.

5.7 Appendix

TABLE 5.A.1 Estimated breeding values in the data.*

Variable name	Mean	St.dev.
NVI (Dutch total value index)	-70.22	13.84
Milk yield	-15293	11255
Fat yield	-399.26	326.17
Protein yield	-583.36	276.28
Lactose yield	-867.50	513.30
Fat percentage	45.76	60.69
Protein percentage	1.78	28.21
Lactose percentage	-9.13	4.54
Dutch milk value index	-345.35	173.72
Milk yield Lactation 1	-938.36	1827
Fat yield Lactation 1	-20.80	57.11
Protein yield Lactation 1	-44.71	48.60
Lactose yield Lactation 1	-65.68	82.62
Fat percentage Lactation 1	5.28	10.02
Protein percentage Lactation 1	-0.54	4.16
Lactose percentage Lactation 1	-2.73	1.63
Dutch milk value (index) Lactation 1	-24.37	31.46
Milk yield Lactation 2	-2277	2027
Fat yield Lactation 2	-63.11	55.10
Protein yield Lactation 2	-88.52	48.87
Lactose yield Lactation 2	-131.23	92.32
Fat percentage Lactation 2	6.84	10.69
Protein percentage Lactation 2	0.42	5.06
Lactose percentage Lactation 2	-3.15	1.45
Dutch milk value (index) Lactation 2	-52.87	30.18
Milk yield Lactation 3	-2756	2029
Fat yield Lactation 3	-72.17	57.77
Protein yield Lactation 3	-102.91	49.16
Lactose yield Lactation 3	-155.48	92.71

TABLE 5.A.1 Continued

Variable name	Mean	St.dev.
Fat percentage Lactation 3	7.71	10.26
Protein percentage Lactation 3	0.52	4.85
Lactose percentage	-19.54	9.10
Dutch milk value (index) Lactation 3	-61.36	30.86
Milk yield Lactation 4	-3236	1862
Fat yield Lactation 4	-84.58	54.47
Protein yield Lactation 4	-119.81	45.35
Lactose yield Lactation 4	-179.55	85.22
Fat percentage Lactation 4	8.48	9.78
Protein percentage Lactation 4	0.48	4.72
Dutch milk value (index) Lactation 4	-71.55	28.52
Milk yield Lactation 5	-3862	1732
Fat yield Lactation 5	-100.84	52.68
Protein yield Lactation 5	-140.68	41.90
Lactose yield Lactation 5	-205.53	79.67
Fat percentage Lactation 5	10.35	9.89
Protein percentage Lactation 5	0.76	4.82
Lactose percentage Lactation 2	-3.63	1.69
Dutch milk value (index) Lactation 5	-84.16	26.56
Persistency Overall	15009	76.11
Persistency Lactation 1	15313	64.52
Persistency Lactation 2	15019	71.11
Persistency Lactation 3	14887	71.45
Persistency Lactation 4	14860	73.07
Persistency Lactation 5	14848	69.13
Rate of Maturity	14846	69.05
Longevity without predictors	-113.55	34.52
Longevity with predictors	-113.84	34.48
Stature	14826	90.23
Chest width	14801	72.99
Body depth	14900	85.29

TABLE 5.A.1 Continued

Variable name	Mean	St.dev.
Angularity	14967	77.51
Body Condition score	14781	82.07
Rump Angle	15062	89.25
Rump width	14749	93.72
Rear legs rear view	14907	103.53
Rear leg set side view	15191	72.41
Foot angle	14747	71.06
Locomotion	14920	88.96
Fore udder attachment	14756	90.06
Front teat placement	14853	103.40
Teat length	15178	92.35
Udder depth	14688	82.07
Rear udder height	14847	91.41
Udder support	14847	110.87
Rear teat placement	14924	101.07
Frame overall conformation score	14750	80.61
Dairy strength overall conformation score	14790	68.59
Udder overall conformation score	14679	96.17
Feet and legs overall conformation score	14771	95.04
Overall conformation score	14601	93.89
Birth (index)	29672	175.90
Direct calving ease	14801	95.90
Maternal calving ease	14729	60.63
Milking speed	14916	77.00
Temperament	14857	86.29
Fertility (index)	14767	91.88
Non-Return rate Lactation 56 days	59375	289.69
Interval calving to first insemination	59400	336.11
Calving Interval	59042	361.59
Interval First to Last Insemination	5926	350.83
Conception rate	59130	347.71

TABLE 5.A.1 Continued

Variable name	Mean	St.dev.
Conception rate heifers	14925	97.33
Age first insemination heifers	14837	91.16
Non-Return rate Lactation 56 days Lactation 1	14850	71.78
Interval calving to first insemination Lactation 1	14778	98.65
Calving Interval Lactation 1	14734	101.79
Interval First to Last Insemination Lactation 1	14819	86.84
Conception rate Lactation 1	14802	92.00
Non-Return rate Lactation 56 days Lactation 2	14871	69.78
Interval calving to first insemination Lactation 2	14880	77.30
Calving Interval Lactation 2	14789	82.04
Interval First to Last Insemination Lactation 2	14831	84.11
Conception rate Lactation 2	14791	84.37
Non-Return rate Lactation 56 days Lactation 3	14812	78.51
Interval calving to first insemination Lactation 3	14893	79.28
Calving Interval Lactation 3	14761	88.01
Interval First to Last Insemination Lactation 3	14808	92.32
Conception rate Lactation 3	14759	85.79
Somatic Cell Score	14763	54.87
Somatic Cell Score Lactation 1	14805	59.69
Somatic Cell Score Lactation 2	14793	50.11
Somatic Cell Score Lactation 3	14759	50.47
Somatic Cell Score Lactation 4	14756	59.33
Somatic Cell Score Lactation 5	14761	64.34
Livability index	15013	121.22
Maternal livability index	14819	94.33
Livability in heifers	29835	124.10
Maternal livability in heifers	14818	94.26
Livability in cows	29828	163.15
Maternal livability in cows	14920	73.53
Body weight	14737	67.69
Beef Merit (index)	14993	47.42

TABLE 5.A.1 Continued

Variable name	Mean	St.dev.
Meat content	44918	256.51
Fat score	45277	231.36
Carcass weight	14882	60.53
Meat content Lactation 1	14940	88.73
Fat score Lactation 1	15084	77.97
Growth	29924	85.47
Veal color	15136	92.64
Meat content Lactation 2	15054	107.09
Fat score Lactation 2	15017	97.01
Growth Lactation 1	15002	69.49
Overall Urea	-9.76	32.01
Urea	-42.06	156.68
Urea Lactation 1	-12.73	33.57
Urea Lactation 2	-11.07	31.53
Urea Lactation 3	-6.09	29.65
Urea Lactation 4	-2.30	29.76
Udder health (index)	14755	78.77
Subclinical mastitis	59245	238.73
Clinical mastitis	59101	307.89
Subclinical mastitis Lactation 1	14773	71.57
Clinical mastitis Lactation 1	14771	74.31
Subclinical mastitis Lactation 2	14823	58.97
Clinical mastitis Lactation 2	14838	80.56
Subclinical mastitis Lactation 3	14847	61.85
Clinical mastitis Lactation 3	14735	91.99
Birth (index) Lactation 1	14836	87.95
Calf survival day 3 - 365	15078	72.20
Survival day 3- 14	15034	56.91
Survival 15 - 180	15098	115.59
Lifetime Production (index)	-1199	230.33
Lifetime Milk Production	-5755	1194.18

TABLE 5.A.1 Continued

Variable name	Mean	St.dev.
Lifetime Fat Production	-187.34	35.35
Lifetime Protein Production	-177.36	35.21
Dry matter intake index (with predictors)	-61.10	15.00
Saved Feed for Maintenance	28.74	12.07
Saved Feed Cost for Maintenance	17.33	7.27
Age of first calving	14888	87.72
Better Life Efficiency	-1.13	1.31
Better Life Health	-2.64	0.64

*Mean and standard deviation are derived from the first parity; St.dev.=Standard deviation

TABLE 5.A.2 Phenotypic variables available at first calving. *

Variable name	Factor levels	Mean	St. dev.
Breed**	4[1]		
Dam Parity		2.591	0.230
Birth season	4		
Holstein (100%/87,5%)	2		
Red factor (Yes/No)	2		
Age at first calving (days)		778.119	38.432
Season of first calving	4		
Number of farm-moves prior to first calving		0.487	0.702
Reared at a different farm (Yes/No)	2		
Number of inseminations		1.625	0.214
First insemination type (AI, natural mating, pasture with bull)	3		
First insemination AI type (AI by farmer, AI by professional)	2		
Second insemination type (includes ‘unknown’)	4		
Second insemination AI type (includes ‘unknown’)	3		
Third insemination type (includes ‘unknown’)	4		
Third insemination AI type (includes ‘unknown’)	3		
Time between first and second insemination (days)		17.992	7.432
Time between second and third insemination (days)		6.206	3.653
Average time between inseminations prior to the first parity (days)		18.160	7.981
Season of first insemination	4		
Age at first insemination		473.187	38.449
Gender of first calf (Female/Male/Stillborn/Unknown)	4		
First calf survived for 24 hours (survived, died or unknown)	3		
Twins (Yes/No)	2		

*For factorial variables, the number of classes is shown and for continues variables, the mean and the standard deviation; St.dev.=Standard deviation **Up to 26 different classes of “breed” exist, but only 4 occur in our data

TABLE 5.A.3 Phenotypic variables available at second calving.*

Variable name	Factor levels	Mean	St.dev.
Variables at first calving			
Breed**	4		
Dam Parity		3.943	0.307
Birth season	4		
Holstein (100%/87,5%)	2		
Red factor (Yes/No)	2		
Age at first calving (days)		775.524	37.613
Season of first calving	4		
Number of farm-moves prior to first calving		0.494	0.707
Reared at a different farm (Yes/No)	2		
Number of inseminations		1.608	0.215
First insemination type (AI, natural mating, pasture with bull)	3		
First insemination AI type (AI by farmer, AI by professional)	2		
Second insemination type (includes 'unknown')	4		
Second insemination AI type (includes 'unknown')	3		
Third insemination type (includes 'unknown')	4		
Third insemination AI type (includes 'unknown')	3		
Time between first and second insemination (days)		17.689	8.198
Time between second and third insemination (days)		6.162	3.955
Average time between inseminations prior to the first parity (days)		17.939	8.809
Season of first insemination	4		
Age at first insemination		472.138	38.016
Gender of first calf (Female/Male/Stillborn/Unknown)	4		
First calf survived for 24 hours (survived, died or unknown)	3		
Twins in the first parity (Yes/No)	2		
New variables at second calving			
Number of inseminations in the first parity		2.012	0.289
First insemination type (AI, natural mating, pasture with bull) in the first parity	3		
First insemination AI type (AI by farmer, AI by professional) in the first parity	2		

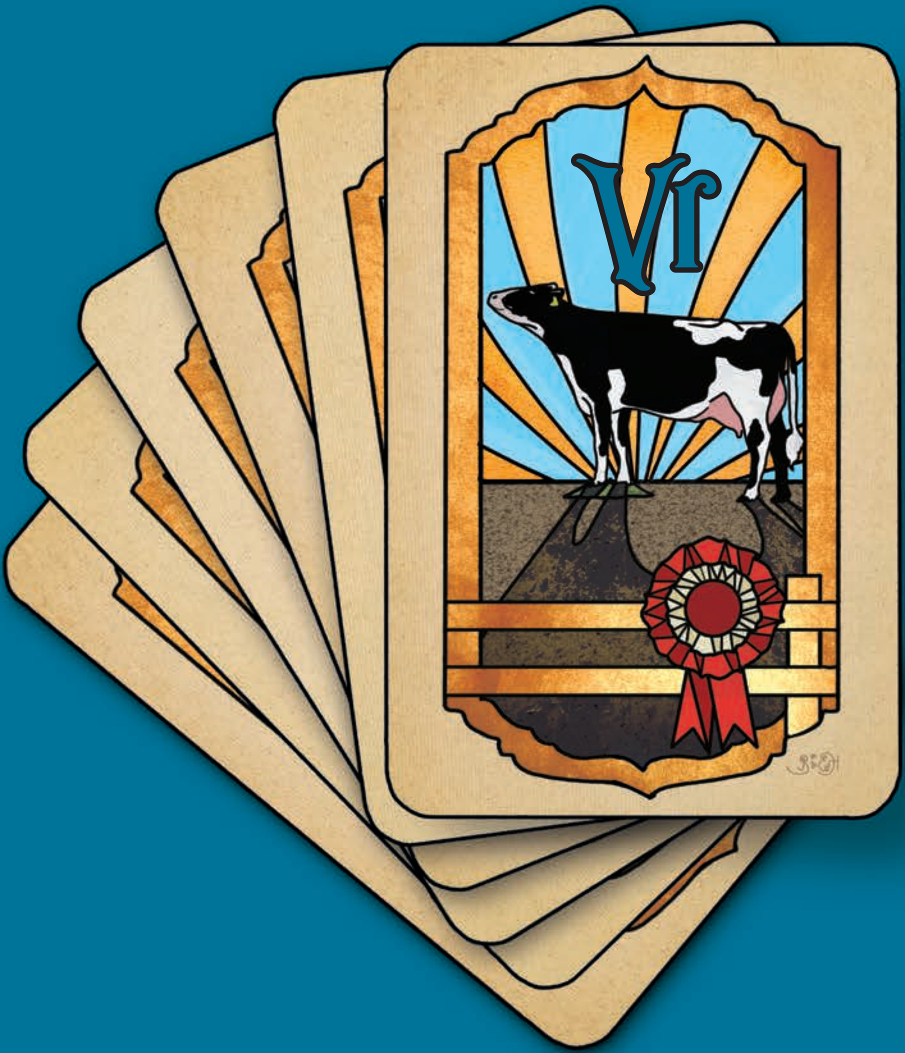
TABLE 5.A.3 Continued

Variable name	Factor levels	Mean	St.dev.
New variables at second calving			
Second insemination type in the first parity	4		
Second insemination AI type in the first parity	3		
Third insemination type in the first parity	4		
Third insemination AI type in the first parity	3		
Time between first and second insemination in the first parity (days)		20.976	3.467
Time between second and third insemination in the first parity (days)		9.668	2.588
Average time between inseminations in the first parity (days)		20.730	3.403
Season of first insemination in the first parity	4		
Age at first insemination in the first parity (days)		861.826	46.598
Lactation length (days)		349.302	16.666
Number of moves between farms in the first parity		0.002	0.005
305-day milk production (kg)		8014	610.397
305-day Fat production (kg)		338.776	23.382
305-day Protein production (kg)		276.917	20.672
Lactation value		103.006	2.091
Milk production (kg) on first test milk day		263.412	16.567
Fat production (%) on first test milk day		455.979	18.336
Protein production (%) on first test milk day		338.650	7.788
Cell count on first test milk day		213.209	53.047
Indication of one of five conditions on first test milk day (3-teat milking, mastitis, cow was sick, fresh, calved early)	5		
Days in milk on first test milk day		19.235	3.067
Milk production (kg) on second test milk day		302.667	19.020
Fat production (%) on second test milk day		403.479	14.536
Protein production (%) on second test milk day		321.567	5.283
Cell count on second test milk day		106.425	28.912
Indication of one of five conditions on second test milk day (3-teat milking, mastitis, cow was sick, fresh, calved early)	5		
Days in milk on second test milk day		53.322	8.614

TABLE 5.A.3 Continued

Variable name	Factor levels	Mean	St.dev.
New variables at second calving			
Test milk day values for fat or protein were imputed (Yes/No)	2		
Total number of failed milkings		0.001	0.002
Total number of 3-teat milkings		0.005	0.018
Total number of milkings while the cow was in heat		0.018	0.029
Total number of milkings with mastitis		0.008	0.023
Total number of milkings where sampling failed		0.049	0.115
Total number of milkings while the cow was sick		0.010	0.012
Total number of milkings after the cow calved recently (max =1)		0.065	0.021
Total number of milkings after the cow calved early (max =1)		0.001	0.003
Age at second calving (days)		1182.216	48.232
Season of second calving	4		
Gender of second calf (Female/Male/Stillborn/Unknown)	4		
Second calf survived for 24 hours (survived, died or unknown)	3		
Twins at second calving	2		

*The first half of the table shows the first parity variables with updated statistics for the second parity dataset. The second half of the table shows new variables available in the second parity dataset. For factorial variables, the number of classes is shown and for continues variables, the mean and the standard deviation; St.dev.=Standard deviation. **Up to 26 different classes of “breed” exist, but only 4 occur in our data


VI

CHAPTER 6

General discussion

6.1 Introduction

Survival traits of dairy cows are important to all aspects of dairy production. For example, improving survival traits increases herd milk yield and farm profitability (De Vries, 2017) while reduces rearing costs on the farm (Mohd Nor et al., 2015; Boulton et al., 2017). Improving survival traits further has a positive effect on the image of the dairy industry in the eyes of consumers and policy makers (Zijlstra et al., 2016; De Vries, 2020) and also reduces the environmental impact of milk production (Bell et al., 2015; Grandl et al., 2019). Because of the importance of survival, attempts to improve this trait are ongoing in many countries (Forabosco et al., 2009). The current approach is improving survival through selective breeding. However, despite a positive genetic trend, this strategy has not resulted in improved survival in practice (Zijlstra et al., 2013; van Pelt, 2017). The lack of improvement in practice is likely due to the large amount of non-genetic factors influencing survival, such as farm management practices and policy changes (De Vries and Marcondes, 2020). In this thesis, I proposed the use of phenotypic prediction as a tool to improve dairy cow survival. By predicting future survival phenotype, farmers could select calves and heifers likely to survive longer. Selecting the best cows through phenotypic prediction would not only directly improve cow survival, but also indirectly by reducing the amount of cows culled early due to excess heifers entering the herd (Overton and Dhuyvetter, 2020).

The aim of this thesis was to investigate if phenotypic prediction of survival in dairy cattle could be done accurately enough to be of use in practice and, secondly, which method was best suited for this prediction problem. Therefore, in **chapter 2**, I investigated if it was possible to predict survival using a traditional method in the animal science domain: multiple logistic regression. The binary trait of interest was survival to second lactation (yes/no). In this chapter, I also investigated which variables contributed most to this prediction, and if there was added value in including estimated breeding values (EBVs). EBVs proved especially useful for predicting survival at a very early stage of life, but also continued to improve prediction throughout the first parity. Although prediction of survival proved possible, the accuracy of prediction for individual cows was low. To study whether the use of a different method might improve accuracy, I compared the results of the method from chapter 2 to two machine learning methods in **chapter 3.** I compared the multiple logistic regression to naive Bayes and random forest. No method proved to be universally better than the others, but each method had different strengths. Multiple regression achieved the highest AUC value on three out of five datasets. Naive Bayes, while not achieving the highest AUC on the most datasets, obtained higher AUC values than the regression for the earliest datasets. The random forest method had the worst performance on overall performance metrics like AUC but scored significantly higher than the other two methods in specificity and negative predictive value. This meant that the random forest model was better at identifying non-surviving cows than the other two methods. Therefore, in **chapter 4**, I investigated if I could harness the different strengths of each individual model to improve prediction by combining the models from chapter 3 into an ensemble. Using multiple regression to combine the outcomes of the three individual methods proved to be the best ensemble method and using an ensemble significantly improved prediction on some, but not all, of the performance metrics. Finally, in **chapter 5**, I attempt phenotypic prediction of true survival in the form of "number of parities reached", rather than a binary trait, and investigate the effect of farm on this prediction. Although accurate prediction of true survival was not possible, this chapter showed several possible strategies of integrating different sources of information for the prediction of survival. By combining the results of this thesis, the first steps were taken in building a model to predict phenotypic survival for use in selecting cows for the production herd. The purpose of this model would be to inform farmers about the future prospects of their cows, allowing them to more accurately estimate how many replacement heifers will be needed, and which young cows to keep in their herds.

The first section of the discussion is dedicated to model evaluation and method selection. I will first describe and review the different performance metrics used in this thesis, then discuss potential problems that can arise during method selection. Lastly in this sub-chapter, I provide an interesting solution to these issues from the data-science domain. As predictive performance in this thesis was low overall regardless of metric, improving the performance of the models will be necessary before they can be of practical use. I therefore explore two strategies that I believe to be the best opportunities to significantly improve predictive performance in section 6.3. In the last section of this discussion, 6.4, I discuss how to move from a scientific study to a practical decision support tool and touch upon how considering the farmer in the method selection and modelling process can help create models with greater practical relevance.

6.2 Method comparison and selection

The main body of this thesis involved the evaluation and comparison of several methods for the prediction of survival. The methods used in this thesis were multiple regression, random forest and naive Bayes. In chapter 4, the results from these three individual methods were combined into ensembles, improving prediction by taking advantage of the differences between methods. Overall, there was no clear consensus on which method (or ensemble) was best for the phenotypic prediction of survival. The most optimal method varied depending on the amount of available phenotypic information and the metric used to evaluate model performance. It is clear that evaluating a prediction model goes beyond simply stating some measure of accuracy and comparing values for this measure between different models (Japkowicz and Shah, 2011; Fawaz et al., 2019). Furthermore, an **accurate** prediction is not necessarily a **useful** prediction, and vice versa. For example, in chapter 2, the multiple logistic regression model was able to predict survival to second lactation with an AUC below 0.7, the threshold for 'decent accuracy' in literature (Akobeng, 2007). However, despite this relatively poor AUC value the model was still able to distinguish between surviving cows and non-surviving cows on average. This means the predictions are not useful for small dairy farms, which only select from a hand-full of calves each time. For those individual calves, the predictions would not be accurate. However, the prediction could be useful for very large farms or for (breeding) companies making decisions on very large groups of calves, as the models were able to distinguish the groups on average. Therefore, depending on the purpose of the model, even models with poor individual accuracy could still be useful for selection at a population level. In the following sections, I will first explore how model performance is defined, then review the performance metrics used in this thesis. Using these performance metrics, I distinguish between what makes a model accurate and what makes a model useful. In the final two sections, I give advice on method selection when no method proves universally better than the other methods and suggest an alternative approach for method selection that avoids some of the pitfalls from manual comparisons as done in this thesis.

6.2.1 DEFINING MODEL PERFORMANCE

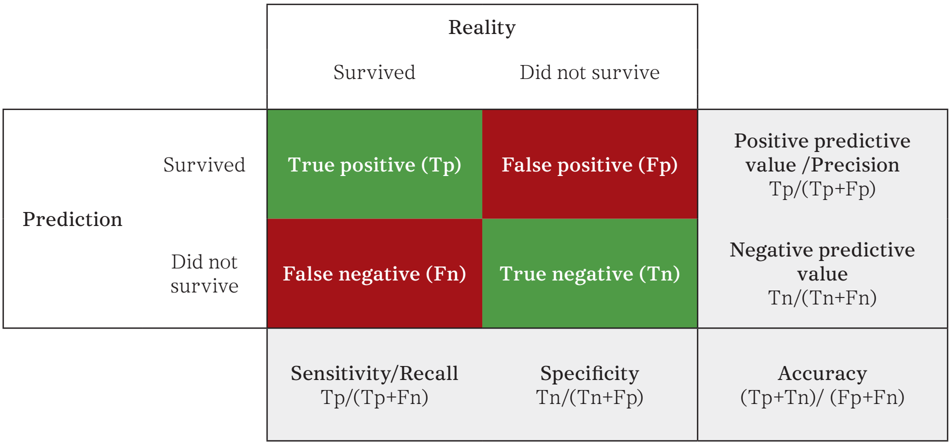
When evaluating or comparing methods, the first step is defining “model accuracy” or “model performance” (Japkowicz and Shah, 2011). Several detailed overviews of metrics to evaluate model performance exist in literature (Baldi et al., 2000; Caruana and Niculescu-Mizil, 2004; Saxena et al., 2008; Ferri et al., 2009; Tharwat, 2018). Because such a wide range of performance metrics is available, the selection of an appropriate metric is challenging. Recommendations on which kind of metric should be used vary from paper to paper (Marcot, 2012; Sileshi, 2014; Chicco, 2017). In precision livestock farming, which includes the use of models that assist farmers in decision-making, the recommended metrics are the “sensitivity, specificity and the overall accuracy” (Norton and Berckmans, 2017). However, accuracy can be very misleading depending on the distribution of the data (Stefanowski, 2016; Akosa, 2017). In this thesis, I saw that when data is very imbalanced, models that optimize accuracy tend to simply predict all cows in the majority class. Sensitivity and specificity are not sensitive to class imbalance, but these metrics are not always a good reflection of the practical relevance of the model. For example, in mastitis detection, a model having specificity of 0.90 is considered quite good from a scientific point of view (Hogeveen et al., 2010; Tharwat, 2018). However, for a farmer, who may have over a hundred cows tested multiple times a day, a specificity of 0.90 would result in dozens of false alerts a day. Achieving a high sensitivity or specificity therefore does not necessarily translate into a model that is useful in practice (Sherlock et al., 2008). In this case, a different metric such as positive predictive value would be better to evaluate the practical usefulness of a model. Which metric is used to evaluate a model therefore depends on characteristics of the data and the intended use of the model.

6.2.2 PERFORMANCE METRICS IN THIS THESIS

A variety of different performance metrics were considered in this thesis. The aim of showing multiple performance metrics was to provide the reader with a comprehensive overview of model performance and avoiding potential bias (Caruana and Niculescu-Mizil, 2004; Jeni et al., 2013; Lever et al., 2016). Most performance metrics measure different aspects of prediction (Ferri et al., 2009). In this thesis, three distinct groups of performance metrics exist (Table 6.1). The first group describes the overall accuracy or performance of the model. The most well-known of the overall performance metrics is accuracy. Other overall performance metrics used in this thesis were balanced accuracy, area under the receiver operator curve (AUC), mean absolute error (mae) and mean squared error (mse), spearman’s rho and the difference in mean predicted probabilities between the two groups.

TABLE 6.1 Performance metrics used in this thesis.

Metric	Chapter 2	Chapter 3	Chapter 4	Chapter 4
Overall performance metrics				
Accuracy	X	X		
Balanced Accuracy	X			
Area under the receiver operator curve	X	X	X	
mean predicted probabilities of surviving and non-surviving cows	X	X		
Mean squared error				X
Mean absolute error				X
Spearman’s rho				X
Class specific performance metrics				
Sensitivity	X	X	X	
Specificity	X	X		
Positive predictive value	X		X	
Negative predictive value	X			
Practical metrics				
Proportion of surviving heifers if 50% of the highest scoring animals were selected		X	X	



		Reality		
		Survived	Did not survive	
Prediction	Survived	True positive (Tp)	False positive (Fp)	Positive predictive value /Precision Tp/(Tp+Fp)
	Did not survive	False negative (Fn)	True negative (Tn)	Negative predictive value Tn/(Tn+Fn)
		Sensitivity/Recall Tp/(Tp+Fn)	Specificity Tn/(Tn+Fp)	Accuracy (Tp+Tn)/ (Fp+Fn)

FIGURE 6.1 Basic confusion matrix showing the results for a binary survival trait. “Survived” is the positive class, and “did not survive” the negative class.

These methods consider the accuracy across all classes, either at a specified cut-off value (accuracy, balanced accuracy and the differences between group averages), or across a range of possible cut-off values (AUC). The second group of performance metrics describes specific aspects of predictive performance, such as the ability of a model to predict one specific class. This group of performance metrics includes the components of a confusion matrix (Figure 6.1): sensitivity, specificity, positive predictive value and negative predictive value. These metrics often require a binary classification problem and depend on a set threshold value to separate the output of the model into different classes. As chapter 5 describes parity reached, which is continuous rather than a class variable, no class-specific metrics are found in this chapter. The third and last group describes model metrics specific to predicting survival that could be used to describe the effect of the model in practice. In the chapters of this thesis, I used 'proportion of surviving heifers if the 50% best scoring heifers were selected' to evaluate the relevance of the methods in practice.

During the writing of this thesis, it became clear that there is no simple answer to which performance metrics should be reported. The selection of metrics depends on the type of trait being investigated, for example binary or continuous, as well as the intended audience of the paper. The intended audience for chapters 2 and 3 of this thesis, for example, were animal scientists. These chapters therefore included metrics which are recommended or commonly used in this field (Norton and Berckmans, 2017). This included accuracy despite the limitations of the metric in our specific prediction problem, because the audience expects this metric when discussing model performance. On the other hand, in chapter 4, the intended audience of the paper were data scientists. In this chapter, the performance metrics included the recall (sensitivity) and positive predictive value (precision), which are common model performance metrics in data science (Powers, 2011). Regardless of who the intended audience of a set of metrics is, I found that in general reporting a variety of different metrics resulted in the most accurate representation of a models' performance. For example, in chapter 3, it would have been possible to select any of the three methods as the 'best' method simply by varying which metric was used for the conclusion. Only by showing several different metrics can the performance of a model be accurately conveyed.

6.2.3 ACCURACY VS USEFULNESS

As long as an overall performance metric is chosen, the most accurate method in a comparison is easy to define: it is the method with the highest value. In an effort to incorporate all aspects of prediction into a single unbiased overall metric, some performance metrics have grown extremely complex (Wagstaff, 2012). While complex performance metrics are useful for the unbiased comparison of different methods, it can be impossible to interpret the practical relevance of these metrics. For example, receiver-operator curves, used to calculate the AUC, indicate the ability of a model across all possible cut-off values. This is excellent for comparing methods, as different methods may have different optimal cut-offs, making them difficult to compare otherwise. However, in practice, cut-offs are required for classification of predictions, usually arising from minimum sensitivity or specificity requirements determined by the practical application of the model. Since the AUC value is a measure of performance value across the entire range of possible cut-offs, this metric reports model performance for cut-off values that will never be used in practice (Muschelli, 2019). Some authors therefore consider metrics assessing prediction **accuracy** for the comparison of methods to be entirely different from metrics that evaluate prediction **usefulness** (Hand, 2012). Perhaps more important than trying to compare methods on abstract performance metrics is to instead describe how well different methods perform on the intended use of the prediction outcome. The method with the best overall metric is not always the best model for a specific job (Lobo et al., 2008; Cawley and Talbot, 2010; Zehner et al., 2019). In this thesis, I chose to use a random forest method in chapter 5 despite poorest overall performance in chapter 4. However, the random forest method also consistently performed (significantly) better than the other methods in specificity and negative predictive value. This means that this method is better at predicting the minority class; cows that do not reach second lactation. This smaller group is also the group of most interest in the prediction problem, as cows that fail to reach second lactation result in the greatest economic loss (Boulton et al., 2017). Therefore, although this method had the worst performance on overall accuracy metrics, it was the most useful method for predicting the non-surviving cows.

6.2.4 METHOD SELECTION WHEN THERE IS NO 'BEST' METHOD

The choice of performance metric is especially important when several methods perform very similarly. For example, in chapters 3 and 4, no method universally outperformed the other methods on the variety of different performance metrics selected. It is not uncommon for multiple methods to be equally (or differently) suitable in a particular prediction problem (Van Wieringen et al., 2009; George et al., 2016), especially on real-world data sets (Sen et al., 2020). In theory, there is no problem with multiple models being equally suitable for a prediction problem. However, in an attempt to provide clarity and direction, many publications still conclude one method to be the most optimal (Hand, 2005, 2006; Boulesteix, 2015). These conclusions are often based on a single performance metric (Lever et al., 2016). When a single performance metric is used to compare methods, the result of the comparison can depend entirely on the metric used (Caruana and Niculescu-Mizil, 2004).

The danger of selecting a 'best' method despite evidence that there is no significant difference between methods is twofold. First, choosing a 'best' method in this situation tends to overestimate the improvement of using more sophisticated methods (Hand, 2006; Boulesteix, 2015). A possible reason for this is publication bias: papers that propose new methods are more likely to be published only if the proposed method performs favorably in model comparisons. This results in a lack of literature showing situations where a new method does not improve over existing methods. Second, the consequence of concluding that one method is better than other methods, despite weak evidence, is that this implies that the outcome of the method comparison also applies in future comparisons. This is unrealistic even for very similar prediction problems, because when methods perform very similarly even minor changes in for example parameter tuning can change the outcome of a comparison (Raeder et al., 2010; Japkowicz and Shah, 2011). Furthermore, (sub-optimal) choices made by researchers in the pre-processing, implementation, optimization or evaluation of a method all affect the outcome of method comparisons (Hand, 2006; Forman and Scholz, 2010; Xu and Goodacre, 2018; Powell et al., 2020). These problems can occur in any statistical process but are especially relevant for machine learning. This is because the increasing availability of free online tools such as "h2o" (Aiello et al., 2015) and "mlr" (Bischl et al., 2016) allow these methods to be used even by novices in the field of computer science (Tarca et al., 2007; Liakos et al., 2018). When researchers use unfamiliar methods on data with high dimensionality, as is often the case in machine learning applications, the likelihood of accidentally influencing model performance increases (Domingos, 2012). It is therefore reasonable that the outcome of a comparison between several methods with similar performance will be different when done by researchers of different expertise levels in computer science.

6.2.5 FUTURE PERSPECTIVES FOR METHOD SELECTION

The consensus may well be that there will never be a most optimal method for all metrics and prediction problems (Jamain and Hand, 2008; Sen et al., 2020). Therefore, testing multiple methods for any prediction problem will likely remain unavoidable. Fortunately, animal science is not the only field where the question of how to choose between all possible prediction methods is the subject of much interest (Zhang et al., 2017). Instead of manually testing many methods as was done in this thesis, the focus is shifting towards developing algorithms that automate method comparison and selection (Guyon et al., 2016; Guyon et al., 2019; Tuggener et al., 2019). Automated method selection algorithms are becoming available to wider scientific use (Kotthoff et al., 2017; Drori et al., 2019; Feurer et al., 2019; Santos et al., 2019). An example of an automated method selection algorithm is the H2O-Automl function of "h2o" (Truong et al., 2019). Most automated method selection algorithms include (a version of) the methods used in this thesis. Furthermore, recent implementations of these algorithms also automatically investigate ensemble methods, which is an automated version of chapter 4 of this thesis (Wistuba et al., 2017). Automated method selection algorithms often include the pre-processing and hyper-parameter tuning step, preventing bias due to differences in user expertise from being introduced into the comparison (Chen et al., 2019). The choice of performance metric for method comparison in these algorithms varies from highly abstract overall metrics to metrics defined by the user. Regardless of the chosen metric, the use of an overall comparison algorithm ensures that the optimization is done on the same metric for every model.

6.3 Improving predictive performance

Unfortunately, the models described in the chapters of this thesis did not achieve sufficient model accuracy to be implemented in practice. For example, in chapter 3, which compared the three different methods, the negative predictive values ranged from 0.12 to 0.17. This meant that if a cow got predicted not to survive to second lactation, there was only a 12 to 17% chance that the prediction was correct. For a farmer, this means that for every 6 or 7 calves indicated as non-survivors by the model only one calf would actually have died or been culled prior to the second lactation. Similarly, in chapter 5, although the mean absolute error indicated that the model was off by between one to two parities on average, there was no clear distinction between groups of cows that die or are culled early and those that reach five lactations or more. The models from this chapter therefore could not be used for accurately ranking the cows for selection in practice. Although it is difficult to estimate the exact accuracy necessary for practical application, it is clear that the accuracy of the models in this thesis needs to be improved before they can be of use.

I selected two strategies which could increase the accuracy of predicting phenotypic survival. The first is increasing the amount of data, and the second is using alternative survival traits in the prediction. Both strategies will be discussed in the following two sub paragraphs. A different approach, investigating alternative methods such as survival analysis or the more advanced neural networks, is also a possibility that is touched upon in the discussions of chapters 2 to 4. However, it is unlikely that choosing an alternative method will improve predictive accuracy more than for example increasing the amount of (relevant) data would (Domingos, 2012). Furthermore, the prime candidate for further investigation is survival analysis. This method was the recommended method for predicting survival traits in literature alongside regression (Lean et al., 2016). In studies comparing survival analysis to other methods the performance of survival analysis is always comparable to methods like those used in this thesis (Kattan, 2003; Van Belle et al., 2011; Mogensen et al., 2012; Gepp and Kumar, 2015). Even in studies where survival analysis was the 'best' method, prediction accuracy of these models only marginally outperformed the other methods. Therefore, changing to this method will likely not result in the increase in accuracy required to allow this model to be used in practice. If method selection were to be revisited, this should be done after the following two strategies are explored.

6.3.1 INCREASING THE AMOUNT OF DATA

The first approach to increase the accuracy of the models in this thesis is by increasing the amount of available relevant data (Domingos, 2012). The available data in the chapters was restricted by design: I explored early prediction and limited the available data to commonly recorded traits. I chose to use only commonly recorded variables for practical reasons. Including variables which may improve prediction accuracy but that were only collected in scientific trials or on research farms would severely limit the application of the models in practice. Early prediction further limits the number of variables available for prediction, as for example milk records only become available after the cow has started lactation. We chose to focus on early prediction because the greatest benefits in improving survival traits in practice are obtained by predicting and selecting cows as early as possible (Overton and Dhuyvetter, 2020).

Increasing the amount of data includes both increasing the number of records (Domingos, 2012; Shahinfar et al., 2014) and variables available. Several variables which could be collected early in life are known to affect future performance of dairy cows. For example, milk production is influenced by calf growth pattern (Van De Stroet et al., 2016; Chuck et al., 2018; Volkmann et al., 2019), calf nutrition (Soberon and Van Amburgh, 2013; Gelsinger et al., 2016) and the occurrence of health events during the rearing period (Aghakeshmiri et al., 2017; Chuck et al., 2018). Data collection on calves is a growing field of science, with recent initiatives to increase data collection on calves such as the innocalf project (Beekman, 2016) and the "COMKALF" project (Goselink, 2018). Information is also increasing from other sources; for example due to increases in variety and availability of sensors, the use of robotics and the use of more advanced software (Wolfert et al., 2017; White et al., 2018; Ferris et al., 2020). Research is already done to create models that include all aspects of a dairy farm, from expected feed crop yields to individual cow management (Lianga et al., 2018; Ferris et al., 2020). These developments are promising for the prediction of survival traits, as survival is influenced by both individual and farm-level factors (Rutten et al., 2013; Zhang et al., 2020).

6.3.2 ALTERNATE SURVIVAL TRAITS

The second approach to improve the accuracy of prediction for survival traits is using a different survival trait as the trait of interest. For example, a survival trait can be both binary (chapters 2 through 4) or continuous (chapter 5). Binary traits are often easier to predict than true survival, because the prediction only attempts to predict one or two years in the future, compared to many years for true survival traits. Although it is difficult to compare performance metrics used for continuous and binary traits directly, true survival (chapter 5) appeared to be more difficult to predict than survival to second lactation (chapters 2 through 4). One of the reasons for this is the difference in length of time between the prediction and the event that is being predicted. In chapter 2, for example, we see that AUC improves greatly between birth (AUC = 0.584) and 200 days post calving (AUC = 0.731). At birth, the time between the moment of prediction and the trait that is being predicted, second lactation, is roughly four years. At 200 days post calving, the time between the moment of prediction and the moment being predicted is only a few months. It is therefore logical that true survival, where the distance between prediction and actual event is often multiple years is even more difficult to predict. For binary traits, data of cows that are still alive can also be used for training through censoring. Next to potentially allowing more cows to be used in the analysis, this also means that more recent training examples can be used. For example, in chapter 2 through 4 of this thesis, the data used dated from 2012 to 2013, whereas in chapter 5 the data dated from 2000 to 2007. The difference is in part due to the fact that records cannot be censored to predict true survival. Using continuous traits also has advantages over binary traits. For example, class imbalance is not problematic in continuous traits, whereas it must be corrected for when using specific methods when predicting binary traits (Stefanowski, 2016). Also, model outputs from models predicting continuous survival traits can be used for ranking of cows at any point in time, which is not always possible using the output from a binary classification, as cows grouped together in one class may be impossible to distinguish amongst themselves.

The survival trait of interest could also be changed by excluding certain causes of death or reasons for culling. Excluding certain causes of death would reduce noise in the trait of interest, which could improve the accuracy of the prediction. In this thesis, I attempted to predict true survival, which included several causes of death that may have been impossible to predict accurately, such as accidents. By excluding causes of death that are difficult to predict, the accuracy could be increased, however, this would also affect the usefulness of the model. For example, if only voluntarily culled cows were included in the trait of interest, the focus of the prediction would shift from predicting survival to predicting farmer decisions. This type of prediction could be useful in certain situations, but it would have a different purpose than this thesis. Going even further, it could be a solution to predict each cause of death individually. By building a model separately for each major cause of death, the noise for each given variable would be reduced while not sacrificing the applicability of the final model. Neural networks, a specific group of machine learning methods, could be especially suited to such an approach because they consider the complex interactions between the inputs. Neural networks have been shown to be a promising method for forecasting complex traits in dairy cows in general (Fenlon et al., 2017), although similar to survival analysis, changing the model alone is not likely to improve the accuracy sufficiently.

Predicting each trait separately would have several benefits. First, it is likely that a model solely built to predict one aspect of survival will have a higher accuracy than models predicting overall survival, due to a reduction in noise. For example, there are models predicting insemination outcome (Shahinfar et al., 2014), disease outcome (Brock et al., 2019) and future milk production (Murphy et al., 2014; Jensen et al., 2016). These models seem to have more accurate forecasting than the models in this thesis. Secondly, a model predicting each cause of death separately would also indicate why the model predicts that a cow is at risk of being culled. By clarifying the reason for culling, the farmer can make a more informed decision: either he can alter the management for that cow to prevent the issue, or he can cull the cow. Lastly, it would be easier to separate incorrect predictions due to accidents and other hard-to-predict events from incorrect prediction due to faulty forecasting. At the same time, although predicting different causes of death separately might reduce noise and improve predictive performance, this would also result in smaller datasets per cause of death. This means for example that farm-specific prediction, such as done in chapter 5, may not be possible for specific causes of death as the number records would be too low. Reducing the number of records would therefore result in a decrease in accuracy for certain causes of death. The main barrier for excluding or modelling causes of death separately however is lacking information on why cows die or are culled. Reasons for culling are often given voluntarily, and even then only a single reason is reported (Pinedo et al., 2014)(personal communication with CRV). Although challenging, collecting detailed information on why cows exit the dairy herd opens up several strategies for improving the accuracy of prediction of survival.

6.4 From prediction to implementation

The reason I investigated phenotypic prediction was not just to investigate the usefulness of various sources of information or find the most suitable method to predict survival, but rather to take the first steps towards a model that could be used for herd management decisions in practice. The development and use of models to support decisions is growing within the dairy industry (Bewley, 2010; Kamphuis and Steeneveld, 2016). The decision support models allow farmers to make more informed decisions using the outcomes of the predictions (Lehenbauer and Oltjen, 1998). The transition from selecting a prediction method to a successful decision support model is perhaps even more complex than the prediction itself. Many decision support models never actually reached the farmer (Rutten et al., 2013; Haine et al., 2017). The step from choosing a prediction method to the implementation of a model in the field also likely exceeds the scope of a single PhD thesis. For example, even in a thesis focusing on implementing a single prediction method to support sow culling decisions (Bono et al., 2012) the implementation of the model did not happen until years later (Hindsborg and Kristensen, 2019). Therefore, even if the accuracy of the models in this thesis were higher, implementation of these models by the end of this thesis was unlikely. However, some parts of this process do deserve to be mentioned, as they are relevant to the method selection and model building steps. In particular, taking into account the needs of the intended users of a model is an important step that is often overlooked (Douthwaite et al., 2001).

6.4.1 PREDICTIONS MUST IMPROVE ON COMMON SENSE

The end-user of a decision support model used in culling and selection of cows is the dairy farmer. For the farmer, a decision support models' usefulness is determined in part by the accuracy of the prediction. However, equally important is that the model must actually provide novel information to aid in the decision-making. Although this step may seem obvious, this aspect is often overlooked in scientific studies. In decision support models that predict on sensor data, the additional benefit of the model is clear because raw sensor data cannot be interpreted by humans (Rue et al., 2020). Without the translation provided by the prediction model, the raw sensor data would not be useful to the farmer. However, when the model predicts a complex trait such as survival, the task of providing additional information is more challenging. For example, many prediction models have been created to help farmers determine the economic value of their cows (Akilli et al., 2016; Cabrera, 2018; Salamone, 2018). The prediction of economic value is similar to survival traits as the predictions are on the individual, the trait is a combination of multiple aspects of milk production (fertility, health and production traits) and the value is used to rank the cow within the herd for selection or culling purposes. Many of these models put a large emphasis on pregnancy status (Kalantari et al., 2010; Nielsen et al., 2010; Cabrera, 2012; Akilli et al., 2016). This means if these models are used for decision support in practice, cows would be ranked almost solely by their pregnancy status. Farmers do not need a mathematical model to tell them ‘do not cull pregnant

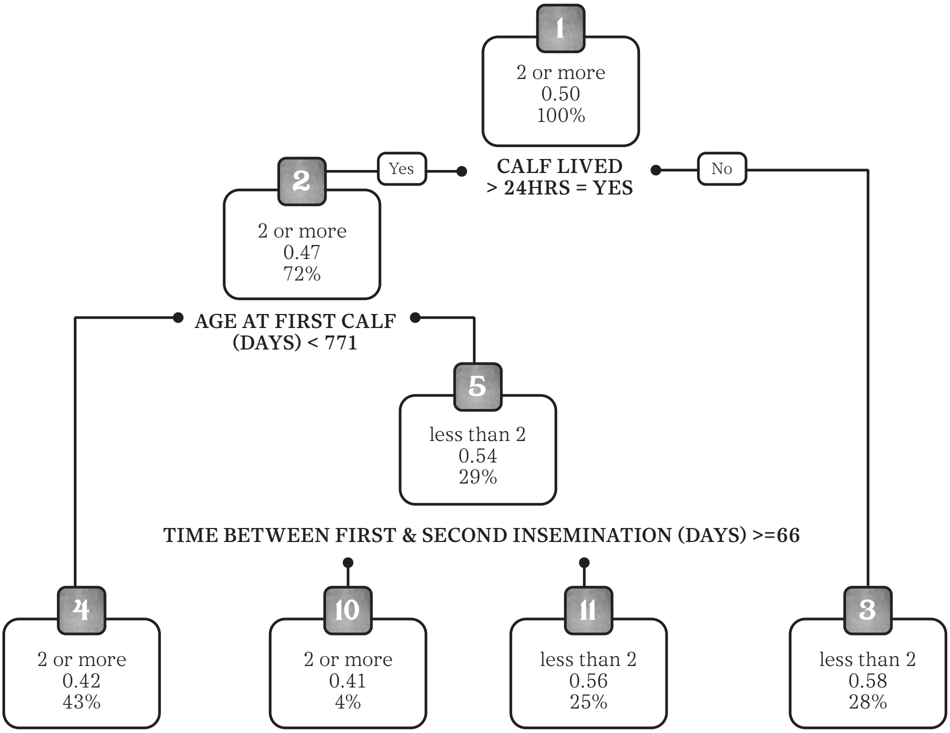


FIGURE 6.2 **A simple decision tree that predicts if a cow will have “less than 2” or “2 or more” lactations based on a subset of the phenotypic data available at first calving.** Each node shows the proportion of correctly predicted cows in the node and the percentage of cows present in that node from the original dataset. In node 1, 'no' can also indicate no data was available.

cows' and to 'cull non-pregnant cows' (Pinedo et al., 2010). Therefore, to be useful, these models must also be accurate at distinguishing the economic value of animals within the pregnancy status groups.

The problem of not providing new information also occurs in this thesis. In chapter 1, I used Figure 6.2 to show an example of a tree used in a random forest. This particular tree makes three splits, including the split "time between first and second insemination >= 66 days". As a cow's estrus cycle is only 18-24 days (Forde et al., 2011), such a long insemination interval only occurs in certain situations. For example, if a cow is inseminated and gets pregnant but the pregnancy fails before day 45. Following the pregnancy failure, the cow then comes into heat again and is inseminated for the second time. Although pregnancy loss is uncommon, as illustrated by the fact that only 2% of the cows in Figure 6.2 are in node 10, this is the most common period of time for early pregnancy failure in dairy cattle (Starbuck et al., 2004). By re-inseminating the cow, the farmer shows his commitment to try to keep the cow in the herd despite the pregnancy failure. Therefore, at least for the group of cows with long insemination intervals, this model does not tell farmers anything other than what they already decided themselves.

Figure 6.2 was only a simple example designed to illustrate how trees work. However, the model in chapter 2 of this thesis also included variables which describe farmer decisions rather than try to predict those decisions. For example, the variable which indicated if insemination records were missing at 18 months was an important predictor for the prediction of survival to second lactation. Cows that do not get inseminated do not get pregnant, and therefore always fail to reach the second lactation. The only reason that this variable was not a perfect predictor for survival to second lactation was that 'missing' insemination records could also indicate cows that still gave birth several months later. This means that the variable included both cows without inseminations and cows that were inseminated but where the insemination was not recorded properly. Including this variable did improve prediction accuracy but adds no additional usefulness for the farmer. The farmer is after all aware if he chose not to inseminate a cow because he has decided to cull it. A useful decision support model must do more than simply predict survival outcomes for cows only after the decision has already been made.

6.4.2 MODEL INTERPRETABILITY

Avoiding uninformative decision support is important, but it is a balancing act. Predictions must also not be too different from the expectations of the farmer (Rudin, 2019). If the user can understand to an extend how the model arrives at its conclusion, the model is perceived as more trustworthy. Interpretability of a model is one of the aspects that should be considered during method selection. Many of the 'traditional' methods used in science are highly interpretable, because scientists also wish to understand how a model arrives at its conclusion. For example, the correlation coefficients from a traditional regression like the method used in chapter 2 of this thesis explain both the magnitude and the direction of the effect of a variable on the trait of interest (Taylor, 1990). In machine learning, interpretability is optional. Machine learning methods can be divided in interpretable 'white box' and un-interpretable 'black box' approaches (Lipton, 2018). In black box approaches, extracting the variable importance is either not possible, or the extracted information is too complex to be interpreted by humans. All methods in this thesis were interpretable, so this was not an issue, but when investigating more advanced deep learning methods, this may play a role. Users prefer interpretable models especially in situations where important decisions are made based on uncertain results (Ribeiro et al., 2016). This is the case in this thesis, where the model supports culling decisions based on prediction of future performance. If a farmer is advised by the prediction model to cull a heifer, this decision cannot be undone or be checked for accuracy after the fact. Interpretable models also allow users to understand how relevant actions influence the predictions, and to act on them (Rudin, 2019). If the multiple logistic regression from chapter 2 was used to predict on a heifer, for example, analyzing the coefficients and the input variables could reveal why the model suspects that heifer will not reach the second lactation. In that case, the farmer might decide to put extra effort in monitoring, resulting in that heifer reaching second lactation despite the negative prediction.

6.5 Concluding remarks

In this thesis, I took the first steps towards building a model that could predict phenotypic survival of dairy cows. I found that all three methods investigated in this thesis had similar overall performance, and that each method had its own strengths and weaknesses. Combining the methods allowed an ensemble method to make use of these differences and showed that there often is no correct answer to the question 'what is the best method?'. Rather than investigating many different methods, it may instead be more useful to focus on choosing the correct trait to investigate, gathering sufficient relevant information (such as EBV and farm characteristics) and selecting a performance metric which can show the ability of the model in practice. This final step, considering the practical use of the model, is a logical addition to the evaluation of prediction models and one that should be considered more often within scientific research. In the future, as method selection is likely to become more and more automated, it is up to the researcher to understand the consequences of choices in trait, available variables and performance metrics on the conclusions and practical applicability of prediction models.

References

Aghakeshmiri, F., M. Azizzadeh, N. Farzaneh, and M. Gorjidooz. 2017. Effects of neonatal diarrhea and other conditions on subsequent productive and reproductive performance of heifer calves. Veterinary research communications 41(2):107-112.

Aiello, S., T. Kraljevic, and P. Maj. 2015. Package 'h2o'. dim 2:12.

Akilli, A., H. Atil, C. Takma, and T. Ayyilmaz. 2016. Fuzzy logic-based decision support system for dairy cattle.

Akobeng, A. K. 2007. Understanding diagnostic tests 3: receiver operating characteristic curves. Acta paediatrica 96(5):644-647.

Akosa, J. 2017. Predictive accuracy: a misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum. p 2-5.

Baldi, P., S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412-424.

Beekman, J. 2016. Dairy Campus Magazine. p 37.

Bell, M., P. Garnsworthy, A. Stott, and J. Pryce. 2015. Effects of changing cow production and fitness traits on profit and greenhouse gas emissions of UK dairy systems. The Journal of Agricultural Science 153(1):138-151.

Bewley, J. 2010. Precision dairy farming: advanced analysis solutions for future profitability. In: Proceedings of the first North American conference on precision dairy management, Toronto, Canada. p 2-5.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. 2016. mlr: Machine Learning in R. The Journal of Machine Learning Research 17(1):5938-5942.

Bono, C., C. Cornou, and A. R. Kristensen. 2012. Dynamic production monitoring in pig herds I: Modeling and monitoring litter size at herd and sow level. Livestock Science 149(3):289-300.

Boulesteix, A.-L. 2015. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS computational biology 11(4)

Boulton, A., J. Rushton, and D. Wathes. 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:1-9.

Brock, J., M. Lange, S. J. More, D. Graham, and H.-H. Thulke. 2019. Reviewing age-structured epidemiological models of cattle diseases tailored to support management decisions: Guidance for the future. Preventive veterinary medicine:104814.

Cabrera, V. 2012. A simple formulation and solution to the replacement problem: A practical tool to assess the economic cow value, the value of a new pregnancy, and the cost of a pregnancy loss. Journal of dairy science 95(8):4683-4698.

Cabrera, V. 2018. Invited review: Helping dairy farmers to improve economic performance utilizing data-driving decision support tools. animal 12(1):134-144.

Caruana, R., and A. Niculescu-Mizil. 2004. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. p 69-78.

Cawley, G. C., and N. L. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11(Jul):2079-2107.

Chen, Y.-W., Q. Song, and X. Hu. 2019. Techniques for Automated Machine Learning. arXiv preprint arXiv:1907.08908

Chicco, D. 2017. Ten quick tips for machine learning in computational biology. BioData mining 10(1):35.

Chuck, G., P. Mansell, M. Stevenson, and M. Izzo. 2018. Early-life events associated with first-lactation performance in pasture-based dairy herds. Journal of dairy science 101(4):3488-3500.

De Vries, A. 2017. Economic trade-offs between genetic improvement and longevity in dairy cattle. Journal of dairy science 100(5):4184-4192.

De Vries, A. 2020. Symposium review: Why revisit dairy cattle productive lifespan? Journal of Dairy Science 103(4):3838-3845.

De Vries, A., and M. Marcondes. 2020. Overview of factors affecting productive lifespan of dairy cows. animal 14(S1):s155-s164.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM 55(10):78-87.

Douthwaite, B., J. Keatinge, and J. Park. 2001. Why promising technologies fail: the neglected role of user innovation during adoption. Research policy 30(5):819-836.

Drori, I., Y. Krishnamurthy, R. Lourenco, R. Rampin, K. Cho, C. Silva, and J. Freire. 2019. Automatic Machine Learning by Pipeline Synthesis using Model-Based Reinforcement Learning and a Grammar. arXiv preprint arXiv:1905.10345

Fawaz, H. I., G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. 2019. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery 33(4):917-963.

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. Journal of Dairy science 100(12):9746-9758.

Ferri, C., J. Hernández-Orallo, and R. Modroiu. 2009. An experimental comparison of performance measures for classification. Pattern Recognition Letters 30(1):27-38.

Ferris, M. C., A. Christensen, and S. R. Wangen. 2020. Symposium review: Dairy Brain—Informing decisions on dairy farms using data analytics. Journal of Dairy Science

Feurer, M., A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter. 2019. Auto-sklearn: efficient and robust automated machine learning, Automated Machine Learning. Springer. p. 113-134.

Forabosco, F., J. Jakobsen, and W. Fikse. 2009. International genetic evaluation for direct longevity in dairy bulls. Journal of dairy science 92(5):2338-2347.

Forde, N., M. Beltman, P. Lonergan, M. Diskin, J. Roche, and M. Crowe. 2011. Oestrous cycles in Bos taurus cattle. Animal reproduction science 124(3-4):163-169.

Forman, G., and M. Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. Acm Sigkdd Explorations Newsletter 12(1):49-57.

Gelsinger, S., A. J. Heinrichs, and C. Jones. 2016. A meta-analysis of the effects of preweaned calf nutrition and growth on first-lactation performance. Journal of dairy science 99(8):6206-6214.

George, N. I., T.-P. Lu, and C.-W. Chang. 2016. Cost-sensitive performance metric for comparing multiple ordinal classifiers. Artificial intelligence research 5(1):135.

Gepp, A., and K. Kumar. 2015. Predicting financial distress: a comparison of survival analysis and decision tree techniques. Procedia Computer Science 54:396-404.

Goselink, R. M. A. 2018. COMKALF: Conceptie op Maat. In: W. U. Wageningen Livestock Research (ed.), melkveefondsprojecten.nl.

Grandl, F., M. Furger, M. Kreuzer, and M. Zehetmeier. 2019. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries.

Animal 13(1):198-208.

Guyon, I., I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macià, B. Ray, L. Romaszko, and M. Sebag. 2016. A brief review of the ChaLearn AutoML challenge: any-time any-dataset learning without human intervention. In: Workshop on Automatic Machine Learning. p 21-30.

Guyon, I., L. Sun-Hosoya, M. Boullé, H. J. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, and M. Sebag. 2019. Analysis of the AutoML challenge series 2015–2018, Automated Machine Learning. Springer. p. 177-219.

Haine, D., R. Cue, A. Sewalem, K. Wade, R. Lacroix, D. Lefebvre, J. Rushton, J. Arsenault, É. Bouchard, and J. Dubuc. 2017. Culling from the actors' perspectives—decision-making criteria for culling in Quebec dairy herds enrolled in a veterinary preventive medicine program. Preventive veterinary medicine 148:1-9.

Hand, D. J. 2005. Supervised classification and tunnel vision. Applied Stochastic Models in Business and Industry 21(2):97-109.

Hand, D. J. 2006. Classifier technology and the illusion of progress. Statistical science:1-14.

Hand, D. J. 2012. Assessing the performance of classification methods. International Statistical Review 80(3):400-414.

Hindsborg, J., and A. R. Kristensen. 2019. From data to decision–Implementation of a sow replacement model. Computers and Electronics in Agriculture 165:104970.

Ho, Y.-C., and D. L. Pepyne. 2002. Simple explanation of the no-free-lunch theorem and its implications. Journal of optimization theory and applications 115(3):549-570.

Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical mastitis—The quest for the perfect alert. Sensors 10(9):7991-8009.

Jamain, A., and D. J. Hand. 2008. Mining supervised classification performance studies: A meta-analytic investigation. Journal of Classification 25(1):87-112.

Japkowicz, N., and M. Shah. 2011. Evaluating learning algorithms: a classification perspective. Cambridge University Press.

Jeni, L. A., J. F. Cohn, and F. De La Torre. 2013. Facing imbalanced data--recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. p 245-251.

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. Journal of dairy science 99(9):7344-7361.

Kalantari, A., H. Mehrabani-Yeganeh, M. Moradi, A. Sanders, and A. De Vries. 2010. Determining the optimum replacement policy for Holstein dairy herds in Iran. Journal of dairy science 93(5):2262-2270.

Kamphuis, C., and W. Steeneveld. 2016. Precision dairy farming 2016. Wageningen Academic Publishers.

Kattan, M. W. 2003. Comparison of Cox regression with other methods for determining prediction models and nomograms. The Journal of urology 170(6):S6-S10.

Kotthoff, L., C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown. 2017. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. The Journal of Machine Learning Research 18(1):826-830.

Lean, I. J., M. C. Lucy, J. P. McNamara, B. J. Bradford, E. Block, J. M. Thomson, J. M. Morton, P. Celi, A. R. Rabiee, and J. E. Santos. 2016. Invited review: Recommendations for reporting intervention studies on reproductive performance in dairy cattle: Improving design, analysis, and interpretation of research on reproduction. Journal of dairy science 99(1):1-17.

Lehenbauer, T. W., and J. W. Oltjen. 1998. Dairy cow culling strategies: making economical culling decisions. Journal of dairy science 81(1):264-271.

Lever, J., M. Krzywinski, and N. Altman. 2016. Points of significance: classification evaluation. Nature Publishing Group.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. Machine learning in agriculture: A review. Sensors 18(8):2674.

Lianga, D., H. Delgadob, and V. Cabrerac. 2018. A virtual dairy farm brain. In: 13th European International Farming Systems Association (IFSA) Symposium, Farming systems: facing uncertainties and enhancing opportunities, 1-5 July 2018, Chania, Crete, Greece. p 1-19.

Lipton, Z. C. 2018. The mythos of model interpretability. Queue 16(3):31-57.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17(2):145-151.

Marcot, B. G. 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. Ecological modelling 230:50-62.

Mogensen, U. B., H. Ishwaran, and T. A. Gerds. 2012. Evaluating random forests for survival analysis using prediction error curves. Journal of statistical software 50(11):1.

Mohd Nor, N., W. Steeneveld, T. Derkman, M. Verbruggen, A. Evers, M. De Haan, and H. Hogeveen. 2015. The total cost of rearing a heifer on Dutch dairy farms: calculated versus perceived cost. Irish veterinary journal 68(1):29.

Murphy, M. D., M. J. O'Mahony, L. Shalloo, P. French, and J. Upton. 2014. Comparison of modelling techniques for milk-production forecasting. Journal of dairy science 97(6):3352-3363.

Muschelli, J. 2019. ROC and AUC with a binary predictor: a potentially misleading metric. Journal of Classification:1-13.

Nielsen, L. R., E. Jørgensen, A. R. Kristensen, and S. Østergaard. 2010. Optimal replacement policies for dairy cows based on daily yield measurements. Journal of dairy science 93(1):75-92.

Norton, T., and D. Berckmans. 2017. Developing precision livestock farming tools for precision dairy farming. Animal Frontiers 7(1):18-23.

Overton, M., and K. Dhuyvetter. 2020. Symposium review: An abundance of replacement heifers: What is the economic impact of raising more than are needed? Journal of Dairy Science

Pinedo, P., A. Daniels, J. Shumaker, and A. De Vries. 2014. Dynamics of culling for Jersey, Holstein, and Jersey× Holstein crossbred cows in large multibreed dairy herds. Journal of dairy science 97(5):2886-2895.

Pinedo, P., A. De Vries, and D. Webb. 2010. Dynamics of culling risk with disposal codes reported by Dairy Herd Improvement dairy herds. Journal of dairy science 93(5):2250-2261.

Powell, M., M. Hosseini, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble. 2020. I tried a bunch of things: the dangers of unexpected overfitting in classification. bioRxiv:078816. doi: 10.1101/078816

Powers, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Raeder, T., T. R. Hoens, and N. V. Chawla. 2010. Consequences of variability in classifier performance estimates. In: 2010 IEEE International Conference on Data Mining. p 421-430.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. p 1135-1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5):206-215.

Rue, B. D., C. Eastwood, J. Edwards, and S. Cuthbert. 2020. New Zealand dairy farmers preference investments in automation technology over decision-support technology. Animal Production Science 60(1):133-137.

Rutten, C. J., A. Velthuis, W. Steeneveld, and H. Hogeveen. 2013. Invited review: Sensors to support health management on dairy farms. Journal of dairy science 96(4):1928-1952.

Salamone, M. 2018. Adaptation and upgrading of a culling decision model to big data scale, Ghent University.

Santos, A., S. Castelo, C. Felix, J. P. Ono, B. Yu, S. R. Hong, C. T. Silva, E. Bertini, and J. Freire. 2019. Visus: An interactive system for automatic machine learning model building and curation. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. p 1-7.

Saxena, A., J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher. 2008. Metrics for evaluating performance of prognostic techniques. In: 2008 International Conference on Prognostics and Health Management. p 1-17.

Sen, P. C., M. Hajra, and M. Ghosh. 2020. Supervised Classification Algorithms in Machine Learning: A Survey and Review, Emerging Technology in Modelling and Graphics. Springer. p. 99-111.

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. Journal of dairy science 97(2):731-742.

Sherlock, R., H. Hogeveen, G. Mein, and M. D. Rasmussen. 2008. Performance evaluation of systems for automated monitoring of udder health: Analytical issues and guidelines, Mastitis Control—From Science to Practice. Wageningen Academic Publishers, Wageningen, the Netherlands. p. 275-282.

Sileshi, G. W. 2014. A critical review of forest biomass estimation models, common mistakes and corrective measures. Forest Ecology and Management 329:237-254.

Soberon, F., and M. Van Amburgh. 2013. Lactation Biology Symposium: The effect of nutrient intake from milk or milk replacer of preweaned dairy calves on lactation milk yield as adults: a meta-analysis of current data. Journal of Animal Science 91(2):706-712.

Starbuck, M. J., R. A. Dailey, and E. K. Inskeep. 2004. Factors affecting retention of early pregnancy in dairy cattle. Animal reproduction science 84(1-2):27-39.

Stefanowski, J. 2016. Dealing with data difficulty factors while learning from imbalanced data, Challenges in computational statistics and data mining. Springer. p. 333-363.

Tarca, A. L., V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici. 2007. Machine learning and its applications to biology. PLoS computational biology 3(6)

Taylor, R. 1990. Interpretation of the correlation coefficient: a basic review. Journal of diagnostic medical sonography 6(1):35-39.

Tharwat, A. 2018. Classification assessment methods. Applied Computing and Informatics

Truong, A., A. Walters, J. Goodsitt, K. Hines, B. Bruss, and R. Farivar. 2019. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. arXiv preprint arXiv:1908.05557

Tuggener, L., M. Amirian, K. Rombach, S. Lörwald, A. Varlet, C. Westermann, and T. Stadelmann. 2019. Automated machine learning in practice: state of the art and recent results. In: 2019 6th Swiss Conference on Data Science (SDS). p 31-36.

Van Belle, V., K. Pelckmans, S. Van Huffel, and J. A. Suykens. 2011. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial intelligence in medicine 53(2):107-118.

Van De Stroet, D., J. C. Díaz, K. Stalder, A. J. Heinrichs, and C. D. Dechow. 2016. Association of calf growth traits with production characteristics in dairy cattle. Journal of dairy science 99(10):8347-8355.

van Pelt, M. 2017. Genetic improvement of longevity in dairy cows. Wageningen University.

Van Wieringen, W. N., D. Kun, R. Hampel, and A.-L. Boulesteix. 2009. Survival prediction using gene expression data: a review and comparison. Computational statistics & data analysis 53(5):1590-1603.

Volkmann, N., N. Kemper, and A. Römer. 2019. Impacts of prepubertal rearIng IntensIty and calf health on fIrst-lactatIon yIeld and lIfetIme performance. Annals of animal science 19(1):201-214.

Wagstaff, K. 2012. Machine learning that matters. arXiv preprint arXiv:1206.4656

White, B., D. Amrine, and R. Larson. 2018. Big data analytics and precision animal agriculture symposium: Data to decisions. Journal of animal science 96(4):1531-1539.

Wistuba, M., N. Schilling, and L. Schmidt-Thieme. 2017. Automatic Frankensteining: Creating complex ensembles autonomously. In: Proceedings of the 2017 SIAM International Conference on Data Mining. p 741-749.

Wolfert, S., L. Ge, C. Verdouw, and M.-J. Bogaardt. 2017. Big data in smart farming–a review. Agricultural Systems 153:69-80.

Xu, Y., and R. Goodacre. 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of Analysis and Testing 2(3):249-262.

Zehner, N., J. J. Niederhauser, M. Schick, and C. Umstatter. 2019. Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. Computers and electronics in agriculture 161:62-71.

Zhang, C., C. Liu, X. Zhang, and G. Almpanidis. 2017. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications 82:128-150.

Zhang, F., J. Upton, L. Shalloo, P. Shine, and M. D. Murphy. 2020. Effect of introducing weather parameters on the accuracy of milk production forecast models. Information Processing in Agriculture 7(1):120-138.

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.

Zijlstra, J., M. Jiayang, C. Zhijun, and J. van der Fels. 2016. Longevity and culling rate: how to improve?, Wageningen UR Livestock Research, Wageningen, the Netherlands.

Summary

Although cows can live to be twenty years old, the average lifespan for a dairy cow is only five to six years. Improving the lifespan of dairy cows would have several benefits such as increasing farm profitability and reducing the environmental impact of milk production. However, the complexity of survival makes it difficult to improve this trait in practice. In this thesis, I proposed using phenotypic prediction of survival to select young cows for the dairy herd, improving survival through increased lifespan of selected cows and better heifer management. The aim of this thesis was to investigate if it was possible to predict survival phenotype accurately enough to be of use in selection.

In **Chapter 1**, I first established the importance of survival traits in dairy cows and the value of improving this trait for the various dairy industry stakeholders. I explained different definitions of survival traits and described the various reasons why cows are culled or die before reaching their full potential. I then describe how the current approach for improving survival - selection using breeding values - ignores the phenotypic and environmental factors which influence cow survival and therefore does not result in increased survival in practice. Therefore, selection on breeding values alone has not been sufficient to improve survival in practice. I therefore proposed phenotypic prediction, which combines genomic and phenotypic information, as a potential solution to improve cow survival. In the last section of this first chapter, I introduced the methods that were used in this thesis

In **Chapter 2**, I investigated if it was possible to predict survival using multiple logistic regression. I further compared three different models: a model including only gEBV, a model including only phenotypic information and a model using both sources of information. This had three purposes: (1) establishing a baseline to compare with other methods, (2) exploring which variables contributed most to accurate phenotypic prediction of survival and (3) investigating the added value of

combining phenotypic with genomic information for predicting phenotypic survival. In this chapter, survival was defined as the binary trait "survival to second lactation, plus two weeks". Survival to second lactation was predicted at five distinct 'decision moments'; moments in the life of a cow at which new phenotypic information becomes available and prediction has added value Combining phenotypic and genomic information had added value at all five moments, as at all decision moments the most optimal model included both types of information. Genomic information proved especially valuable early in life, when little phenotypic information is available. The models were also able to distinguish between surviving and non-surviving cows on average. This meant that in a practical scenario when the highest scoring 50% of heifers were selected, 2.0% to 3.0% more cows would survive when using this model compared to a randomly selected 50%. The results from a practical scenario show that it would be possible to improve survival of dairy cows through selection. However, accurately predict survival for individual cows was not possible using this method.

In **Chapter 3**, I compared the linear method multiple logistic regression to two non-linear machine learning methods naive Bayes and random forest. These three methods were selected because they take very different approaches to predict a trait, which could lead to differences in prediction accuracy. No method proved to be universally better than the others, as all methods obtained similar increases in survival in the practical scenario. Overall, naive Bayes has the highest average AUC at all decision moments except at 200 days post calving, where random forest outperformed the other methods. There were big differences in how the methods predicted individual cows, however. The correlations of prediction outcomes between methods were lower than expected, ranging from r = 0.417 to r = 0.700. This meant that the prediction for an individual cow could be different depending on the method used.

In **Chapter 4**, I investigated if I could harness the differences between the methods from Chapter 3 by combining the prediction outcomes of individual methods into an ensemble. Multiple logistic regression was the best performing ensemble method, naive Bayes the second-best, and the random forest ensemble method resulted in the least significant improvement over the individual methods. The multiple logistic regression ensemble method resulted in equal or better recall, AUC, balanced accuracy and proportion of animals surviving on all datasets. Recall, 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 regression proved a useful method to investigate the additional value of ensemble methods.

In the previous chapters, I examined the ability of various models to predict a binary survival trait. In **Chapter 5**, I attempted the phenotypic prediction of true survival and investigated the effect of farm on this prediction. We tested five different models to investigate the farm-effect: (1) a baseline model, (2) a model including a farm-ID variable, (3) a model including 9 descriptive variables on herd size, farm growth and average parity reached on that farm, (4) a model including the 9 descriptive variables and 40 farm-average EBV variables and finally (5) a model which trained and validated on each farm separately. Overall, rank correlations between true parity reached and the predicted parity reached were low, with models predicting conservatively around the mean. At first calving, the best model was the model that included farm-variables, amongst others herd size and average parity reached on a farm in 2000 and 2001. At second calving, the best model was the baseline model, although the models which included farm-variables or farm ID performed almost identically. Furthermore, when farm variables were included, these variables were consistently ranked high in the feature importance scores, indicating the existence of a farm effect.

Finally, in **Chapter 6**, I discuss different aspects of method selection and place the results of this thesis in the context of building a model to support selection decisions. I describe the various performance metrics used for method comparison in the thesis and how these metrics describe different aspects of prediction. As predictive performance was low for the models in this thesis, I suggest two different strategies for improving predictive performance: collecting more data early in life and changing the definition of the trait of interest. I conclude the discussion by describing the difficulty of moving from a model in a scientific paper to a decision support model in practice and highlight how a focus on practicality can improve the design process of prediction models. The results of this thesis provide valuable insights in the challenges of predicting survival traits and the suitability of various (machine learning) methods for the prediction of survival traits in dairy cattle.

Curriculum vitae

About the author

Esther Margaretha Maria van der Heide was born on September 15th, 1992 in Amsterdam, the Netherlands. She obtained her high school degree (VWO) from the Mgr. Frencken College in Oosterhout in 2010. She then went on to obtain a BSc Biology from the Vrije Universiteit (VU) in Amsterdam, during which she completed her minor thesis at the Wageningen University. This lead her to pursue a MSc degree in Animal science at Wageningen University and Research, majoring in Animal Breeding and Genetics and obtaining a minor in Marketing and Management. Her major thesis was completed in Athens, GA, the United states of America, under supervision of D.A.L. Lourenco at the University of Georgia. The results of the major thesis were published as “Sexual dimorphism in livestock species selected for economically important traits”. In 2016, Esther was accepted as a PhD candidate at the Animal Breeding and Genomics department of Wageningen University. This PhD was part of the SMARTBREED project, a collaboration project between Wageningen University and research, the University of Groningen, government funding agency NWO (formerly known as STW) and four commercial breeding companies through the breed4Food consortium. During this PhD, she investigated both classical linear methods and novel machine learning methods to predict survival in dairy cattle. The results of that research are presented in this thesis.

Publications

van der Heide, E. M. M., Veerkamp, R. F., Kamphuis, C., Azzopardi, G., Athanasiadis, I., van Pelt, M. L., & Ducro, B. J.. Improving predictive performance using ensemble methods in a case study on survival in dairy cattle. Submitted for review in Computers and Electronics in Agriculture.

van der Heide, E.M.M., Veerkamp, R. F., Kamphuis, C., van Pelt, M. L. & Ducro, B. J. (2020). Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information. Journal of dairy science, 103(1), 556-571.

van der Heide, E.M.M., Veerkamp, R. F., Kamphuis, C., Athanasiadis, I., van Pelt, M. L. & Ducro, B. J. (2019). Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. Journal of dairy science, 102(10), 9409-9421.

van der Heide, E.M.M., D. A. L. Lourenco, C. Y. Chen, W. O. Herring, R. L. Sapp, D. W. Moser, S. Tsuruta, Y. Masuda, B. J. Ducro, I. Misztal. (2016). Sexual dimorphism in livestock species selected for economically important traits. Journal of animal science, 94(9), 3684-3692.

Other publications and abstracts

van der Heide, E. M. M., Veerkamp, R. F., Kamphuis, C., van Pelt, M. L., & Ducro, B. J. Modelling farm-specific effects in the prediction of parity reached in dairy cattle. Thesis chapter, to be submitted.

van der Heide, E. M. M., Veerkamp, R. F., Kamphuis, C., van Pelt, M. L., & Ducro, B. J. Comparing multiple regression with two machine learning methods in a case study predicting individual survival to second lactation in Holstein cattle. 2019. ADSA, Cincinnati, OH, United States of America. Poster presentation.

van der Heide, E. M. M., Veerkamp, R. F., van Pelt, M. L., Kamphuis, C. & Ducro, B. J. Combining three prediction methods for the prediction of survival in dairy cattle. 2019. Precision Livestock Farming, Copenhagen, Denmark. Oral Presentation.

van der Heide, E. M. M., Veerkamp, R. F., van Pelt, M. L., Kamphuis, C. & Ducro, B. J. Comparing multiple regression with two machine learning methods in a case study predicting individual survival to second lactation in Holstein cattle. 2019. WIAS Science day. Oral presentation.

van der Heide, E. M. M., Veerkamp, R. F., van Pelt, M. L., Kamphuis, C. & Ducro, B. J. Comparing multiple regression with two machine learning methods in a case study predicting individual survival to second lactation in Holstein cattle. 2018. FAIR Symposium, Wageningen, the Netherlands. Poster presentation.

van der Heide, E. M. M., Veerkamp, R. F., Kamphuis, C. & Ducro, B. J. (2018). Early prediction of survival to second lactation in Holstein cattle by combining phenotypic and genomic information. 2018. WCGALP, Auckland, New Zealand. Poster presentation.

Training and Supervision Plan

THE BASIC PACKAGE (3 ECTS)

WIAS Introduction day	2016
Philosophy of Science and Ethics	2016
Essential Skills (Frank Little)	2017

DISCIPLINARY COMPETENCES (16.1 ECTS)

Research proposal	2016
Stanford Machine learning course *(online)*	2017
Getting started with ASREML	2017
Design of breeding programs with genomic selection	2017
Hadoop parallel processing workshop	2019
Precision Livestock Farming workshop	2018

QUANTITATIVE DISCUSSION GROUP (QDG)

Organisation	2016 - 2019
Attendance	2016 - 2020

PROFESSIONAL COMPETENCES (9.5 ECTS)

Interpersonal communication for PhD candidates	2016
Project and Time Management	2017
Scientific writing	2019
Career Perspectives	2019
Last Stretch of the PhD Programme	2020

WIAS ASSOCIATED PHD STUDENTS (WAPS) COUNCIL

Education Commission	2017-2018
Wageningen PhD council representative	2018-2019
WAPS Chair	2019

PRESENTATION SKILLS (4 ECTS)

WCGALP, Poster, Auckland, New Zealand	2018
Wageningen PhD Symposium, Poster, Wageningen, the Netherlands. *First prize winner (Poster)*	2018
FAIR Symposium, Poster, Wageningen, the Netherlands	2018
WIAS Science day, Oral, Wageningen, the Netherlands	2019
Precision Livestock Farming, Oral, Copenhagen, Denmark	2019
ADSA, Poster, Cincinnati, USA	2019

TEACHING COMPETENCES (2.6 ECTS)

Teaching Assistant - YAS 20806	2017 - 2018
Teaching Assistant – ABG 20306	2018

TOTAL ETCS: 35.2

T

Acknowledgements

Although this thesis only has one author, I would have never been able to write all of this without the support (emotional and otherwise) and assistance from my family, (co)promotors, colleagues and friends. I will try my best to give credit where credit is due in this section, but those who know me know that my memory for names is not the best ^_^'.

First, I would like to thank the people who made this project possible: everyone involved in the smartbreed project, my promotor Roel Veerkamp and project leader/co-promotor/supervisor Bart Ducro, as well as the breeding companies of Breed4food and the University for setting up the project and supporting me throughout. In particular, I would also like to thank Claudia Kamphuis, who joined my project midway through and found the time (and later fortunately the money) to support me with her enthusiasm first and foremost, but also by making the supervision process smoother and her tireless reviewing of my research and text. The reader of this thesis ought to thank my supervision team too, because without them I probably would have used thirty-five different words to say "cow", using different terms for each paragraph in my thesis.

Second is all the different colleagues that I've met through my PhD. As much as I maybe am not one to join coffee breaks or Friday drinks, your friendship, support and distraction when things got rough pulled me through these four years. For most of the four years-and-some-months that I spend on this thesis, I was in PhD room 1, together with an ever-changing rotation of different PhD candidates. I'd like to thank these friends and colleagues who made days of writing a little bit more bearable, or provided answers (and snippets of code) when I couldn't find the answer on my own. Sabine, Marieke, Malou, Ibrahim, Shuwen, Benan, Xaofei and all the others, thanks so much for the fun and support! Outside of the little world that was PhD room/ corridor 1, I would like to thank the PhD's I met through the courses, conferences and different groups and councils I participated in. From Denmark: Dan for the long emails and inspiring talks on just about whatever, Mona for your enthusiasm and tireless work for the PLF (still the best conferences from my career) and Roos, my partner in crime and friend during the conference and the few months you spend at ABG. Closer to home, the QDG members (Ibrahim and Malou, thanks so much for taking over!), the WAPS and the WPC members. From the WAPS, I'd like to thank in particular Chiara who was an organized chair and asked me to design the WIAS Science Day logo for 2018. I'd also like Eline for taking over after me. WAPS representatives, even if it doesn't always feel that way, you're a valuable resource and voice for all WIAS PhDs! I wish you good luck and that one day you can finally recruit one PhD from each chair group 😉. From the WPC, our tireless leader Bart, the other Marieke (I loved doing that housing questionnaire) and Job, who taught me so much about university politics. I feel like I grew so much during my time discussing PhD issues at a higher level, in a way that isn't shown in this thesis.

In particular, I would like to thank Fatma, Sanne and Lisette, without whom I probably never would have finished. Fatma for being my colleague, my neighbor, and especially my friend, dragging me to social events, always introducing me to different people, and bringing me cake throughout the quarantine days so I stood no chance against those corona kilo's 😉. By the time you finish your PhD I'm sure you will actually know all the Turkish people in Wageningen, and I know from experience that those friends and acquaintances will be a great help along the way. Sanne, for being my friend, helping me keep QDG afloat, remembering the names of all the PhD's for me and for being someone I could complain to about the problems we had during our PhDs together. I still admire you for the strength you showed in the choices you made. I think knowing there were other viable options was one of the reasons why I managed to pull through in the end. At the same time, I hope my next job will also be one I can just leave behind at work after 5pm 😊.

Lisette, who was always up for a chat, knew exactly what I had to do or who I should ask for stuff and also when to not ask and just give me a hug. You were an indispensable part of ABG, without whom all the office plants would have died and all the foreign PhD students would get lost in transit somewhere. You're forever my hero for letting me arrange that thing at the end of my PhD that I probably shouldn't spell out and I still feel bad you caught flak for it. I wish you luck with the things you want to do in life and happiness with your family 😊.

Last, but definitely not least, I'd like to thank my family, without whom I wouldn't be here and who supported me throughout. Papa, voor je steun en omdat je mij bent voorgegaan op dit pad, ook al was het allemaal heel anders in 1993. Sommige van die levenslessen neem ik de rest van mijn leven mee ('Het boekje maakt niet uit, niemand gaat dat toch lezen' 😉). Mama, omdat je je zo zorgen hebt gemaakt terwijl het uiteindelijk toch allemaal gewoon goed is gekomen (behalve dan dat gamen, dat doe ik nog steeds). Mijn kleine grote broer, omdat ik altijd met je kan praten, lachen en dingen doen, en omdat je deze halve thesis al gelezen hebt om hem voor mij te proof-readen. Opa en Oma van der Heide, voor jullie support tijdens mijn opleiding, liefde en omdat jullie grote voorbeelden voor mij zijn. De dag van mijn defense wordt nog bijzonderder omdat ik hem mag delen met Opa. Ik weet hoe trots hij zou zijn als hij er zelf bij had kunnen zijn.

Lieve familie, jullie zijn de fundering waarop ik mijn leven kan bouwen, omdat ik weet dat als ik ooit wankel sta ik die kracht en stabiliteit heb om op terug te vallen. Bedankt ♡

A

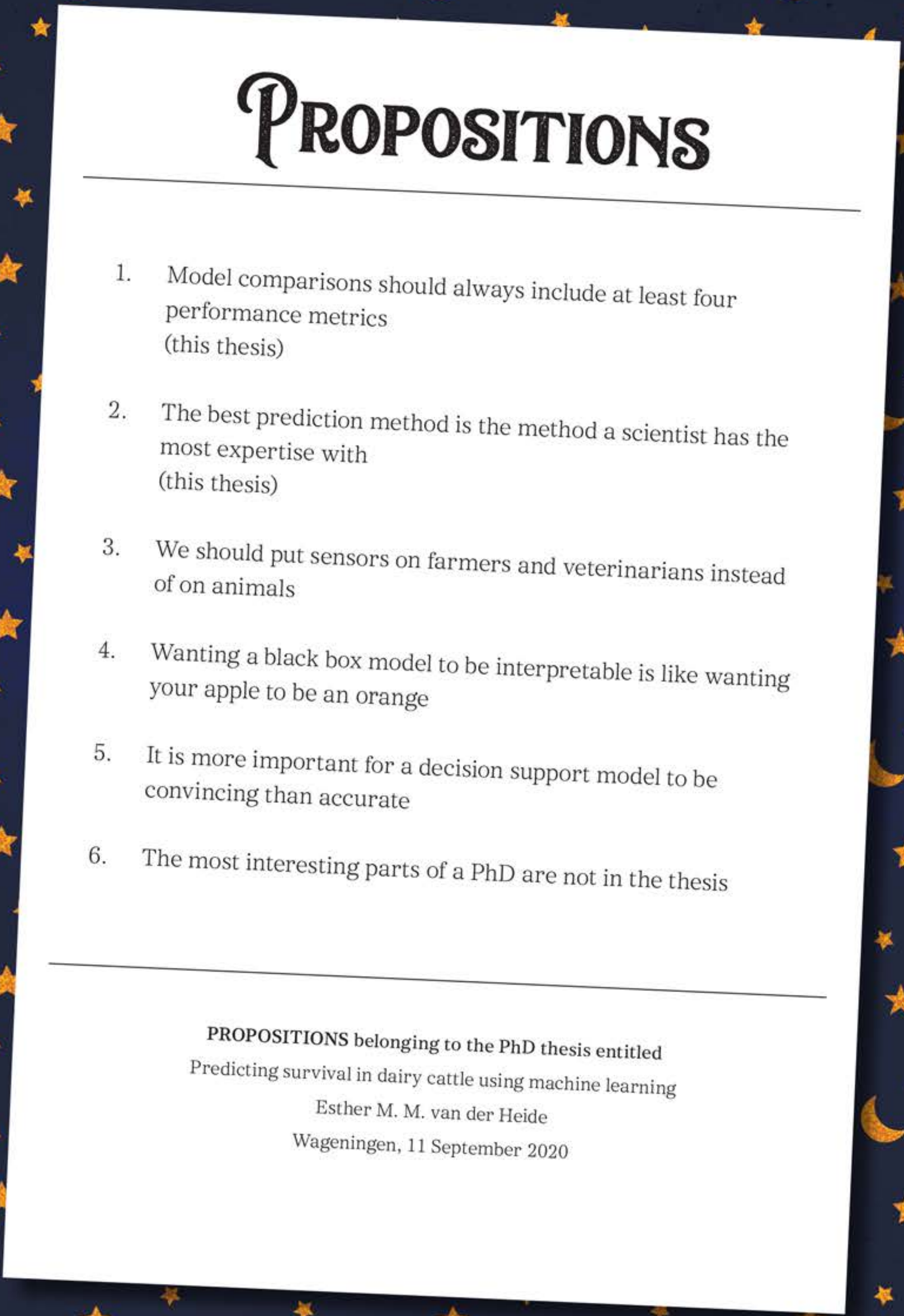
COLOPHON

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 727213.

Author: E.M.M. van der Heide
Concept cover: E.M.M. van der Heide
Bookdesign & illustration: isontwerp.nl, Eindhoven
Print: Digiforce, Vianen

DOI-link: https://doi.org/10.18174/523729
ISBN: 978-94-6395-427-3

The full page illustrations of this book were digitally drawn and designed by Ilse Schrauwers, based on pencil sketches by Esther van der Heide.



Propositions

1. Model comparisons should always include at least four performance metrics
(this thesis)

2. The best prediction method is the method a scientist has the most expertise with
(this thesis)

3. We should put sensors on farmers and veterinarians instead of on animals

4. Wanting a black box model to be interpretable is like wanting your apple to be an orange

5. It is more important for a decision support model to be convincing than accurate

6. The most interesting parts of a PhD are not in the thesis

PROPOSITIONS belonging to the PhD thesis entitled

Predicting survival in dairy cattle using machine learning

Esther M. M. van der Heide

Wageningen, 11 September 2020