A comparison of machine learning techniques for predicting insemination outcome in Irish dairy cows

Caroline Fenlon*, Luke O'Grady[†], John Dunnion*, Laurence Shalloo[‡], Stephen Butler[‡], Michael Doherty[†]

*School of Computer Science, University College Dublin, Ireland
[†]School of Veterinary Medicine, University College Dublin, Ireland
[‡]Animal and Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, County Cork, Ireland
caroline.fenlon@ucdconnect.ie,
{luke.ogrady, john.dunnion, michael.doherty}@ucd.ie,
{laurence.shalloo, stephen.butler}@teagasc.ie

Abstract. Reproductive performance has an important effect on economic efficiency in dairy farms with short yearly periods of breeding. The individual factors affecting the outcome of an artificial insemination have been extensively researched in many univariate models. In this study, these factors are analysed in combination to create a comprehensive multivariate model of conception in Irish dairy cows. Logistic regression, Naïve Bayes, Decision Tree learning and Random Forests are trained using 2,723 artificial insemination records from Irish research farms. An additional 4,205 breeding events from commercial dairy farms are used to evaluate and compare the performance of each data mining technique. The models are assessed in terms of both discrimination and calibration ability. The logistic regression model was found to be the most useful model for predicting insemination outcome. This model is proposed as being appropriate for use in decision support and in general simulation of Irish dairy cows.

1 Introduction

Dairy production systems in Ireland are primarily based on seasonal calving patterns. Reproductive performance in these systems has an important impact on economic efficiency. In these pasture-based farms, the aim is to align peak grass availability with peak lactating cow energy demands, by breeding animals during a set time period. Poor reproductive performance results in extended periods of calving, suboptimal utilisation of pastures and increased feed costs.

The individual factors affecting conception have been extensively researched. However, few models have comprehensively examined the factors influencing the outcome of insemination in combination, particularly at the individual breeding event level [19]. Most statistical analysis has focused on identifying important factors in isolation and analysing overall measures of reproductive performance, such as calving to conception interval or the probability of conception during a breeding season [5].

Statistically important factors incorporating both genetic and phenotypic effects (parity, stage of lactation, calving events, measures of energy balance and milk production) were identified as significant in previous analyses of records from Irish herds [2][7]. Binary logistic regression was used to form a predictive model of conception outcome. In this study, the aim was to identify and apply other appropriate machine learning techniques to the problem of predicting insemination outcome. To allow direct comparison of the models, they were all built using the same variables as the previous study.

When evaluating binary predictions, two categories of assessment are possible: discrimination and calibration [21]. Discrimination measures a model's ability to correctly classify cases; i.e. the separation between the successful and unsuccessful outcomes. Evaluations of discrimination depend on a cut-off point to transform the predicted probabilities into outcomes and ignore the raw predictions. Classification tables show the rate of correct class predictions, separated by positive and negative instances. These values can be used to calculate precision and recall [17]. To identify the optimal cut-off point, receiver operating characteristic (ROC) curves are used to plot the false-positive rate against the true-positive.

Calibration compares the predictions to the true proportions of events occuring, i.e. determining if the observed frequency of occurrence is similar to the predicted probability, within groups of records. Reliability measures such as the Hosmer-Lemeshow test [12] are used to test overall goodness-of-fit. Calibration plots [4] allow visual inspection of deviation, with statistical tests for analysis of bias and spread. Analysis of deviances may be used to highlight outlying records or covariate values.

As breeding outcome may be considered both in terms of the probability of occurrence and the binary prediction, the models used were compared using both forms of assessment. Evaluation was carried out on an external dataset of records from typically managed commercial Irish dairy herds.

2 Methods

2.1 Data

The data available for model training were sourced from the centralised database at Teagasc's Animal and Grassland Research and Innovation Centre, Moorepark, Co. Cork. The animals included in the dataset were from the Curtins and Ballydague spring-calving research herds, both of which emulate typical Irish dairy management systems. Additional variables were available in this dataset which were used to find the significant factors in the modelling process. After cleaning, inference and missing value removal, 2,723 artificial insemination service records from 658 lactating cows (1,552 lactations) were available for analysis. Service outcome (i.e. conception or no conception) was recorded as a binary variable and was confirmed by ultrasound pregnancy diagnosis between 30 and 60 days post-service or subsequent calving 282 ± 15 days after conception. 47.88% of the services resulted in conception. The variables analysed were: parity (the number of times the cow has previously calved); log days in milk (days since last calving); inter-service interval; the difficulty of the last calving; body condition score (measure of how fat or thin the cow is), as a second-order polynomial effect due to its non-linear relationship with conception probability; and genetic traits for milk production and calving interval.

Observations within the external testing dataset were recorded on 9 commercial dairy farms involved in a herd fertility consultancy program operated by the School of Veterinary Medicine, University College Dublin (UCD) [20]. 4,205 services from 1,471 cows (2,702 lactations) were available for prediction. The same measurements as in the training set were available. 47.49% of these services were successful.

Descriptive statistics from both datasets are shown in Table 1. All data manipulation, analysis and evaluation were carried out using the R statistical programming language [18] and R libraries.

Variable	Training data mean (SD)	Testing data mean (SD)
Parity	2.48 (1.51)	2.78 (1.74)
Days in milk	91.86 (29.83)	85.60 (28.83)
Calving interval genetic trait	-3.32 (2.68)	-2.90 (2.47)
Milk genetic trait	82.55 (184.91)	169.33 (153.00)
Body condition score at breeding	2.89 (0.31)	2.86 (0.22)

Table 1: Descriptive statistics of Moorepark and UCD School of Veterinary Medicine datasets

2.2 Machine Learning Techniques

Four widely-used methods capable of modelling binary values or probabilities were used to model the outcome of breeding to service.

Logistic Regression. Binary logistic regression [12] (R function `glm` [18]) is a generalisation of simple linear regression designed to model the effect of independent variables on the probability of the modelled outcome occurring. Logistic regression assumes all independent variables are normally distributed and not strongly correlated. Regression analysis allows for interactions between independent variables to be included in the model. Random effects can be incorporated to account for the influence of unmeasurable events or global effects. In this study, a basic logistic regression model without interactions or random effects was built to allow for direct comparison with other models. Logistic regression models predict the probability of the event occurring, which can then be transformed to a binary outcome using a threshold probability.

Naïve Bayes. The implementation of Naïve Bayes used in this study (e1071 library function `naiveBayes` [16]) also makes the assumption that numeric features are normally distributed, but assumes no dependencies between them. If known, a-priori probabilities can be set; in this case, the overall conception rate was used. The Bayes rule calculates the probability of each potential outcome, given the a-priori probabilities and the input values. The outcome with the highest probability is then chosen as the predicted result.

Decision Tree. Tree models are created by recursively splitting the training dataset into subsets based on the value of an attribute. The next node is chosen by finding the attribute that can provide the most information when splitting the set. Cut-off thresholds are generated to discretise numeric variables. Using the `rpart` function (from the R library of the same name [22]) results in probabilistic terminal nodes for binary outcomes.

Random Forest. Random forests (randomForest library function `randomForest` [14]) are an ensemble learning method for Decision Trees. It uses both bootstrapping and random feature selection to train a large number of Decision Trees [23]. In this study, random forests with 100, 250 and 500 trees were built.

2.3 Evaluation

Discrimination analysis. For each of the models, the true and predicted service outcomes (given a threshold probability of 50%) were tabulated in a confusion matrix. From this, precision, recall and F-measure were calculated. The Matthews correlation coefficient was also calculated to show the performance of the models in comparison with a random classifier [15]. It ranges from -1 (completely inaccurate predictions) to +1 (completely accurate predictions), with 0 indicating the same performance as random prediction.

Receiver operating characteristic (ROC) curves were used to assess how performance varied as the discrimination threshold was altered. The plot presents the true positive rate against the false positive rate, allowing the optimal probability or classifier to be interpreted visually or using summary statistics, such as the area under the curve.

Calibration analysis. Each model was used to predict the probability of conception occurring in each row of the test set, using the `predict` function with appropriate arguments.

The Hosmer-Lemeshow test [11] was used to evaluate the overall goodness-of-fit of the models on the testing data. The test (R function `hoslem.test` from the ResourceSelection [13] package) splits the observations (sorted by predicted probability) into 10 equal-sized groups of risk and compares the observed number of events to the mean predicted number of events within each group. The disadvantage of overall goodness-of-fit tests is that they cannot identify more specific cases of poor prediction [6]. For a thorough investigation of capabilities, they should be used in conjunction with the more in-depth tests of calibration described below.

For each set of model predictions, a calibration plot was drawn by grouping the observations into 25 equi-interval bins and plotting the mean predicted probability against the proportion of true events within each group. The data were split into 25 to allow for acceptable-sized groups while still maintaining low within-group probability variation. Bins containing fewer than 20 records were not plotted. Confidence intervals for the proportions of successful inseminations were calculated using the F distribution (`calibration.plot` function of the PresenceAbsence R package [8]).

Binned prediction deviations were visually examined for patterns. 95% of the binned values should lie within two standard deviations of 0 [9]. The absolute group deviances were averaged to find the mean absolute calibration error.

3 Results

All of the variables described were significant at P = 0.05 (using the `drop1` function on the logistic regression model).

3.1 Discrimination

The ROC curve of each of the models is shown in Figure 1. The confusion matrix for each model is in Table 2. Discrimination test results (precision, recall, F-score and Matthews correlation coefficient) are in Table 3. All of the models performed similarly in these tests, with F-scores ranging from 50.01% to 52.03%. All of the models performed better than a random classifier in the Matthews correlation coefficient (range 0.11 to 0.16).

3.2 Calibration

Results of statistical tests carried out to measure calibration and goodness-of-fit are shown in Table 4. These results can be seen visually in the calibration (Figure 2) and deviance plots (Figure 3).

Model		Conceived	Did not conceive
Logistic Regression	Predicted True	895	651
	Predicted False	1102	1557
Nave Bayes	Predicted True	928	745
	Predicted False	1069	1463
Decision Tree	Predicted True	924	774
	Predicted False	1073	1434
Random Forest (100 trees)	Predicted True	981	843
	Predicted False	1016	1365
Random Forest (250 trees)	Predicted True	988	813
	Predicted False	1009	1395
Random Forest (500 trees)	Predicted True	989	830
	Predicted False	1008	1378

Table 2: Confusion matrices for each of the models

Model	Precision	Recall	F-score	Matthews Correlation Coefficient
Logistic Regression	57.89%	44.82%	50.52%	0.16
Naive Bayes	55.47%	46.47%	50.57%	0.13
Decision Tree	54.42%	46.27%	50.01%	0.11
Random Forest (100 trees)	53.78%	49.12%	51.35%	0.11
Random Forest (250 trees)	54.86%	49.47%	52.03%	0.13
Random Forest (500 trees)	54.37%	49.52%	51.83%	0.12

Table 3: Discrimination statistical tests

Model	Hosmer-Lemeshow p-value	Mean absolute calibration error	Deviances within 2 SD
Logistic Regression	0.19	3.48%	100%
Naive Bayes	0.00	4.26%	80%
Decision Tree	1.00	4.66%	94%
Random Forest (100 trees)	0.00	6.40%	63%
Random Forest (250 trees)	0.00	5.96%	64%
Random Forest (500 trees)	0.00	5.99%	68%

Table 4: Calibration statistical tests

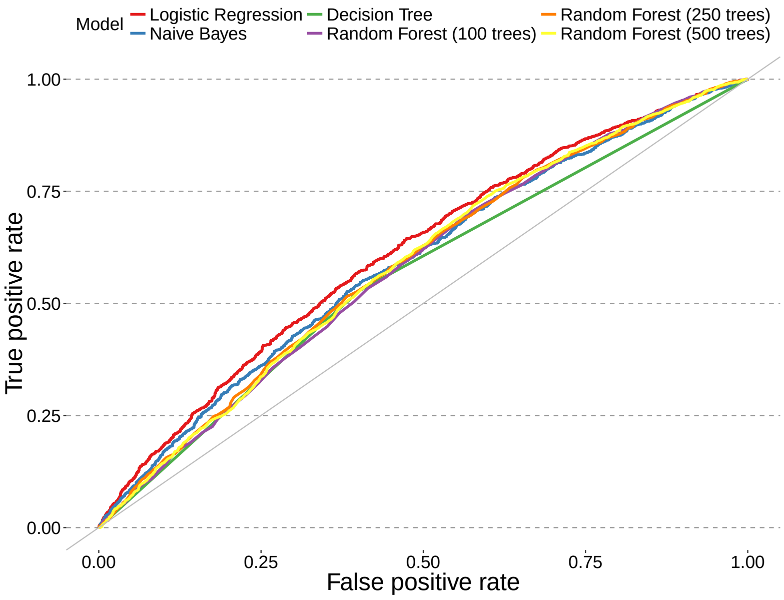


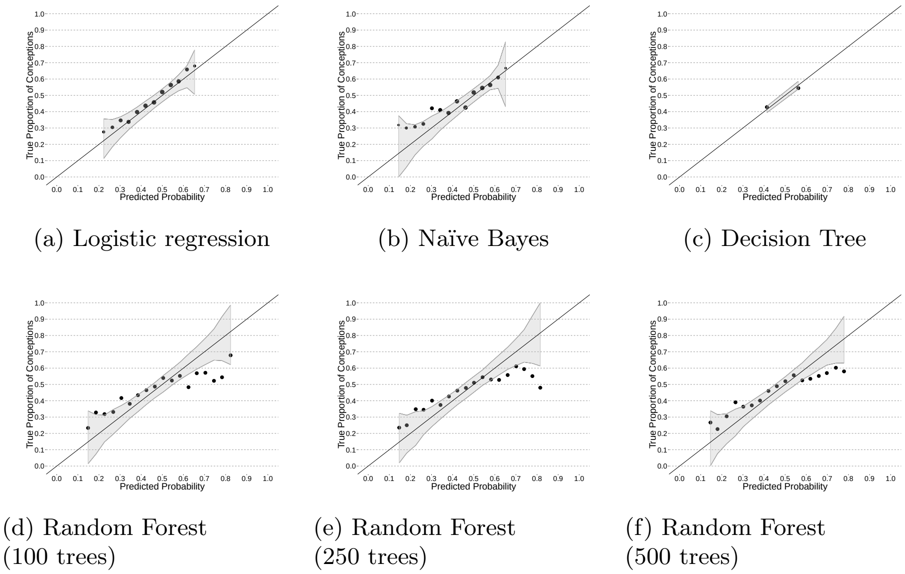
Fig. 1: ROC curves of the four machine learning models

There were no significant differences found between the true and predicted logistic regression and Decision Tree outcomes with the Hosmer-Lemeshow test. The test found significant differences between the true outcomes and the predictions from the Naïve Bayes and all of the Random Forest models.

The models had mean absolute calibration error ranging from 3.48% to 6.40%, with the Random Forest model built with 100 trees having the highest rate of calibration error. The Decision Tree just exceeds the accepted limit of 5% of deviance values outside the two standard deviation limit. The Naïve Bayes and all of the Random Forest models were well above this limit. Some evidence of a deviance pattern is seen in the Naïve Bayes deviance plot, while a very clear pattern is observed for the Random Forest models.

4 Discussion

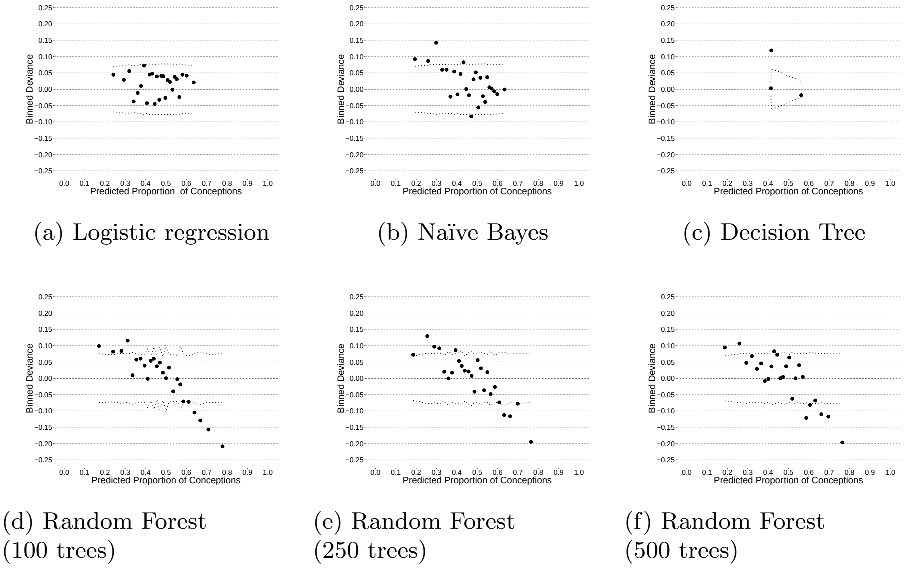
The logistic regression model had the best calibration performance; its calibration error was lowest, along with the most compact deviance spread. The model's F-score was similar to the other models, but it had the highest precision and lowest recall. Its Matthews correlation coefficient was the highest of the models.



(a) Logistic regression (b) Naïve Bayes (c) Decision Tree

(d) Random Forest (100 trees) (e) Random Forest (250 trees) (f) Random Forest (500 trees)

Fig. 2: Calibration plots for the models of dairy cow conception



(a) Logistic regression (b) Naïve Bayes (c) Decision Tree

(d) Random Forest (100 trees) (e) Random Forest (250 trees) (f) Random Forest (500 trees)

Fig. 3: Deviance plots for the models of dairy cow conception

The Naïve Bayes model failed the Hosmer-Lemeshow test of overall goodness-of-fit, and the calibration plot showed some points outside the 95% confidence interval. With 20% of its deviance values outside two standard deviations of 0 and some observation of systematic deviance, it showed poor capability of predicting the probability of conception. This was in spite of discrimination performance comparable to the rest of the models.

The probabilities predicted from the Decision Tree model had a very narrow range; only four distinct probabilities were predicted, resulting in only two probability groups with enough records to display on the calibration plot. This also reduced the number of rows used to calculate the Hosmer-Lemeshow test statistic. Although the discrimination evaluation of the Decision Tree did not differ greatly from the other models, its poor calibration performance makes it an unsuitable choice for predicting the outcome of service.

Because the algorithm continues to create trees until every record is correctly classified, the Random Forests were perfect models of the training data used to build them. Although these models had the best test performance in terms of discrimination, their calibration results were poor. The calibration plots show significant bias, and the distinctly non-random deviance plots indicate that the models are not capturing some important element related to the outcome [10].

Data that are not well separated along different outcomes will be very common in epidemiological applications, where probabilities close to 1 or 0 are uncommon and most in-group probabilities tend to be centred close to 50%. The benefit of modelling these outcomes is to identify events with probabilities outside the norm. This can aid the decision making of farmers and their advisors when selecting the best animals for costly insemination techniques such as sexed semen [3]. Because the probability is the focus, rather than the ultimate outcome, a predictive model with good calibration is key. Thus the logistic regression model is the best model for predicting service outcome. Easily interpretable coefficients or odds ratios may be used to inform farmers about the important risk factors for service outcome.

5 Conclusion

This paper demonstrates a novel application of machine learning algorithms in the context of Irish agriculture. Each technique was trained using data from research herds and tested with an external dataset representing the typical commercial dairy herd in Ireland. The methods implemented all show similar discriminative ability, but logistic regression was found to be the most capable at correctly predicting the probability of conception. Further improvements to the model might be made using regression with ensemble methods such as bagging

[1].

This is, to the authors' knowledge, the first time comprehensive statistical modelling of service outcome in Irish cows has been reported. Having a generalisable predictive model of how various risk factors combine to influence the probability of conception will aid farmers to better understand the performance potential of their animals when making management decisions, such as culling or selection of herd replacements. In addition, the fact that the model is based on easily recordable and obtainable data should further increase the practical utility of the model as a decision support tool. As well as the stand-alone benefits of the model, it is being integrated into a detailed whole-farm model of Irish dairy animals, which will simulate nutrition, reproduction, management and economics in daily time-steps for the entire life of each animal.

Acknowledgements. This research was supported by funding from the Dairy Levy Research Fund. The authors would like to thank Anne Geoghegan in Teagasc, Moorepark and the farmers involved in the UCD School of Veterinary Medicine consultancy programme for assistance in gathering the data used in this study. The suggestions of the AICS reviewers were gratefully received.

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24(421), 123–140 (1996), http://link.springer.com/10.1007/BF00058655
2. Buckley, F., O'Sullivan, K., Mee, J.F., Evans, R.D., Dillon, P.: Relationships among milk yield, body condition, cow weight, and reproduction in spring-calved Holstein-Friesians. Journal of Dairy Science 86(7), 2308–2319 (2003), http://dx.doi.org/10.3168/jds.S0022-0302(03)73823-5
3. Butler, S.T., Hutchinson, I.A., Cromie, A.R., Shalloo, L.: Applications and cost benefits of sexed semen in pasture-based dairy production systems. Animal 8 Suppl 1(s1), 165–72 (2014), http://journals.cambridge.org/abstract{_}S1751731114000664
4. Cohen, I., Goldszmidt, M.: Properties and Benefits of Calibrated Classifiers. In: Proceedings of ECML, pp. 125–148. Springer (2004)
5. Coleman, J., Pierce, K.M., Berry, D.P., Brennan, A., Horan, B.: The influence of genetic selection and feed system on the reproductive performance of spring-calving dairy cows within future pasture-based production systems. Journal of Dairy Science 92(10), 5258–5269 (2009), http://dx.doi.org/10.3168/jds.2009-2108
6. Cox, D., Snell, E.J.: Analysis of Binary Data, Second Edition. CRC Press, Boca Raton (1989), https://books.google.com/books?hl=en{\&}lr={\&}id=0R8J71LCLXsC{\&}pgis=1
7. Cummins, S.B., Lonergan, P., Evans, A.C.O., Berry, D.P., Evans, R.D., Butler, S.T.: Genetic merit for fertility traits in Holstein cows: I. Production characteristics and reproductive efficiency in a pasture-based system. Journal of Dairy Science 95(3), 1310–22 (2012), http://www.ncbi.nlm.nih.gov/pubmed/22365213
8. Freeman, E.A., Moisen, G.: PresenceAbsence: An R Package for Presence Absence Analysis (2008), http://www.jstatsoft.org/v23/i11/paper

9. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge (2006)
10. Harrell, F.E.: rms: Regression Modeling Strategies. R package version 4.3-1 (2015), http://cran.r-project.org/package=rms
11. Hosmer, D.W., Lemeshow, S.: Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods 9(10), 1043–1069 (1980), http://www.tandfonline.com/doi/abs/10.1080/03610928008827941
12. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley, Hoboken (2013)
13. Johnson, C.J., Nielsen, S.E., Merrill, E.H., McDonald, T.L., Boyce, M.S.: Resource Selection Functions Based on Use-Availability Data: Theoretical Motivation and Evaluation Methods. The Journal of Wildlife Management 70(2), 347–357 (2006), http://dx.doi.org/10.2193/0022-541X(2006)70[347:RSFBOU]2.0.CO;2
14. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R News 2(3), 18–22 (2002), http://cran.r-project.org/doc/Rnews/
15. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Structure 405(2), 442–451 (1975)
16. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2015. R package version pp. 1–6
17. Olson, D.L., Delen, D.: Advanced data mining techniques. Springer Science & Business Media, New York (2008)
18. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015), http://www.r-project.org/
19. Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., Weigel, K.: Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. Journal of Dairy Science 97 (2014)
20. Somers, J.R., Huxley, J., Lorenz, I., Doherty, M.L., O'Grady, L.: The effect of Lameness before and during the breeding season on fertility in 10 pasture-based Irish dairy herds. Irish Veterinary Journal 68(14), 1–7 (2015), http://www.irishvetjournal.org/content/68/1/14
21. Tedeschi, L.O.: Assessment of the adequacy of mathematical models. Agricultural Systems 89, 225–247 (2006)
22. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive Partitioning and Regression Trees (2015), http://cran.r-project.org/package=rpart
23. Tin Kam Ho: Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition 1, 278–282 (1995), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=598994