Gradus Vol 8, No 2 (2021)1-8
ISSN 2064-8014

Random Forest Regression models for Lactation and Successful Insemination in Holstein Friesian cows. 1. Mathematical aspects

*Lillian Oluoch[1] *; László Stachó [1], László Viharos[1], Andor Viharos[2], Edit Mikó[3]*

[1]Bolyai Institute, University of Szeged, Aradi vértanúk tere 1., Szeged, Hungary.
[2]Department of Manufacturing Science and Engineering,Budapest University of Technology and Economics, Budapest, Hungary.
[3]Institute of Animal Sciences and Wildlife Management,University of Szeged, H-6800 Hódmezovásárhely, Hungary.
https://doi.org/10.47833/2021.2.AGR.001

Keywords:
random forest regression (RFR),
high performance regressors,
models for lactation of dairy cows,
factor prediction for insemination,
reproductive management.

Article history:
Received 12 August 2021
Revised 28 August 2021
Accepted 06 September 2021

Abstract

The random forest regression (RFR) method is applied to study dairy cows' economical breeding and milk production to overcome well-known difficulties in establishing reliable models based on large data sets. Regarding the features of RFR, there are several positive experiences in various areas of applications supporting that with RFR, one can achieve reliable model predictions for industrial production of any product providing a useful base for decisions. In this study, a data set compiled over a decade for about 80,000 cows, was analyzed through RFR. Ranking of production control parameters is obtained, and the most important explanatory variables are found by computing the variances of the target variable on the sets created during the training phases of the RFR. Predictions are made for milk production and the calves' conception with high accuracy on given data, and simulations are used to investigate prediction accuracy. This study is primarily concerned with the mathematical aspects of a forthcoming article that focuses on the agricultural aspects. The results will be compared with models based on factor analysis and linear regression for future mathematical research plans.

1 Introduction

Reproductive management is a key factor in economic dairy production, and poor practice can cause considerable economic loss, mainly because of decreased milk yield per cow per lactation and decreased number of calves per year per cow. However, it is also associated with reduced conception rates [1]. The conception rate is determined by heat detection, the choice of the first insemination time after calving, the induction of ovulation, and the ovulation synchronization program. As we would like to have the best conception rate, it is also worth noting some environmental features and management practices that would directly affect insemination, thus having adverse effects on reproduction performance. The efficiency, accuracy, and timing of artificial insemination (AI) remain a major challenge to improving many dairy farms' reproductive and economic efficiencies [2, 3]. Various studies have shown that regression models are of great importance in addressing the conception rate issues. Some of them have been used in prediction of the optimal time of insemination

*Corresponding author. Tel.: +36 203731406
E-mail address: oluoch@math.u-szeged.hu

[4]. The probability of conception was analyzed using a logistic procedure that uses the maximum likelihood method to fit linear logistic regression [5]. However, for the logistic regression, the target variables are assumed to be independent and single-valued, yet some data are categorical. Because of these drawbacks, other methods like machine learning procedures can address such problems. Various machine learning algorithms, including Bayesian networks, decision trees, and in particular random forest algorithms, have been employed for such tasks. Bayesian networks are mainly suited for small and incomplete data [6] with challenges in discretizing continuous variables and implementing recursive feedback loops [7]. Decision trees, as well as RFRs, can be used for classification and regression too. RFR algorithm has been widely utilized due to its ability to accommodate complex relationships. RFR calculations can be trivially parallelized to be done on multiple cores of the same central processing unit. Additionally, the RFR algorithm involves very few statistical assumptions, and its hyperparameters can be used to reduce overfitting. The performance of RFR can be explained by the power of ensemble methods to generate high-performance regressors by training a collection of individual regressors. RFR was considered in a study of predicting pregnant versus non-pregnant cows at the time of insemination, and it proved to be significantly better than other machine learning techniques [8]. Random forest was also used in an attempt to predict conception outcome in dairy cows [8].
On the other hand, mathematical models for lactation are not new either. Models of lactation curves were early referenced by Brody [9], but due to the limitations of the computers and computational difficulties experienced, the early models were based on simple logarithmic transformations of exponentials, polynomials, and other linear functions [10]. Another study on Mathematical Modeling of Lactation Curves, gave an overview of the parametric models used to fit lactation curves in dairy cattle by considering linear and non-linear functions [11]. Machine learning approaches have also proved to be essential in the lactation study. Different models based on machine learning in both non-autoregressive and autoregressive cases have been investigated in forecast models of production of dairy cows [12], to find the best performance for both cases with the random forest algorithm. Regression trees have been used in the past to analyze different factors affecting lactation. Studies on effects of the dry period, the lactation parity, the farm, the calving season, the cow's age, the year of calving, and the calving interval have been performed by several authors [13, 14].
Although previous studies have used other machine learning-based models (including RFR) to predict lactation and successful insemination, the proposed study will adopt the RFR technique for the same but with different variables. Our purpose is to investigate how the large collection of data gathered over the last decade used in milk production factories throughout Europe could effectively be analyzed. Therefore, this study aims to apply a random forest regression model to predict lactation influencing factors and the success of insemination (SI) and the choice of the time of insemination attempts.

2 Materials and methods

A large data set was obtained for this analysis. However, some data were not useful due to some missing information; hence were omitted in the study. All the data editing and analyses were conducted using Python, in which *pandas* were used for data preparation and scikit-learn as an open-source machine learning library. We also used the open-source *fastai* library developed at Stanford University(https://github.com /fastai/fastai).

This study obtained a dataset from the Agricultural Department of the University of Szeged, which was collected from three major livestock farms concerning 21 different 82564 Holstein Friesian cows' parameters. In this case, we considered the variables subdivided into the following three groups:

V1. *Geneology*: (i) Settlement (ST); (ii) Cow ID (ID); (iii) Father (FT); (iv)Calving Number (CN); (v) Calving Date (CDT); (vi) Sex of 1st Calf (SX1); (vii) Sex of 2nd Calf (SX2).

V2. *Insemination and calving*: (i) First Date of Insemination after Calving (FDFB); (ii) Time between calving and first insemination (DFI); (iii) Date of successful insemination after calving (PSIS);

(iv) Time between calving and successful insemination (SFAC); (v) Number of unsuccessful inseminations after calving (NUIC); (vi) How many inseminations were unsuccessful in the previous calving (IPAC); (vii) Days open (Number of days to successful insemination in previous lactation, PCIS); (viii) Age in months of conception at heifer (UMP); (ix) Age in months at first calving (MFCA).

V3. *Lactation*: (i) Milk yield in previous lactation (AMPL); (ii) 305-day Milk yield in previous lactation (AMPLD); (iii) Number of days milking previous lactation (DPLM); (iv) Dry days in previous lactation (PLD); (v) Calving interval (DC).

Notice that the data structure was not intended to be the subject of deep mathematical analysis. Furthermore, it can be seen that there are parameters values with a high correlation. On the other hand, it would have been more useful to access genuine-time series in detail instead of accumulated parameters like the calving interval. It is also interesting that the current data do not involve finer details of production volumes of lactation. Studies concerning lactation and BLUP index [15] will be carried out in a separate paper to be submitted in an agricultural journal. Herein we concentrated on the describing the use of RFR techniques, and we focus the following problems:

P1. Estimate the 305-day milk yield in the previous lactation based on the variables except for V3(ii).

P2. Estimate the number of open days based on the variables except for V2(iv).

P3. Provide a weighted ranking of the variables' impact in answer to the above two questions.

We built respective random forests consisting of 1000 decision trees each for the targets of the problems P1 and P2. The trees were constructed using the familiar CART algorithm for the hyperparameters (https: //doi.org/ 10.1023 /A:1010933404324). Each tree was trained on a bootstrap set. The number of samples drawn from the processed training set to form these bootstrap sets were equal to the number of samples in the processed training set. The number of samples drawn was a hyperparameter and the maximum depth of the trees was also a hyperparameter; nevertheless, we did not limit the depth. Figures 1 and 2 depict sample decision trees of depth 3. The depth of the trees used in the models was much bigger. There is one node at the start for each tree, the root node that contains all the samples. For each node that has at least two samples, a split is performed. To split each node, the CART algorithm examines the possible splits of all the features for that specific node, and the best alternative was considered, following the used splitting criterion. Subsequently, the selected feature interval will be split by the feature's selected value, resulting in two new nodes. Consequently, the two new nodes will have fewer samples; moreover, if one or both of these new nodes have at least 2 samples, the splitting process continues. When there is only one sample left in a node, the CART algorithm will not perform the splitting. The minimum number of samples required for splitting is also a hyperparameter and can be changed. We used the default value 2. The trees were grown as long as any stopping rule did not stop the growing process. Pure nodes, where the target variable was identical in all samples, were not split. Nonetheless, we did not use pruning. This way, we got "fully grown and unpruned trees" (https://scikit- learn.org/ stable/modules/generated/ sklearn.ensemble. RandomForestRegressor.html. and https://doi.org/10.1023 /A:1010933404324).

In both cases, the prediction value by means of the random forest is the mean of the prediction values of its individual decision trees.

Given any tree T and one of it nodes n, our procedure fixes a feature $f_{T,n}$ along with a value $v_{T,n}$ in the range of $f_{T,n}$. Furthermore the algorithm fixes a prediction value $p_{T,\ell}$ to every leaf of T. Given a "virtual cow" C with $f_{T,n}$-values $f_{T,n}(C)$ (for each node n of T), our RFR estimate for answering P1. P2 according to tree T is the value $r_T(C) = p_{T,\ell(C)}$ where the leaf $\ell(C)$ is determined as follows: We start at the root of the tree, and if a node n is reached, the decision for continuing to left or right to a next node is done accordingly if we have $f_{T,n}(C) < v_{T,n}$ or $f_{T,n}(C) \geq v_{T,n}$ (cf. Fig 1 and Fig 2). Our steps ended when reaching a leaf which we set to $\ell(C)$.

As for the standard theoretical background on the fact that the predicted value by means of a random forest is the mean of the prediction values of its individual decision trees, we refer to the monograph [16] and Breiman [17]. For the implementation, we used the python machine learning

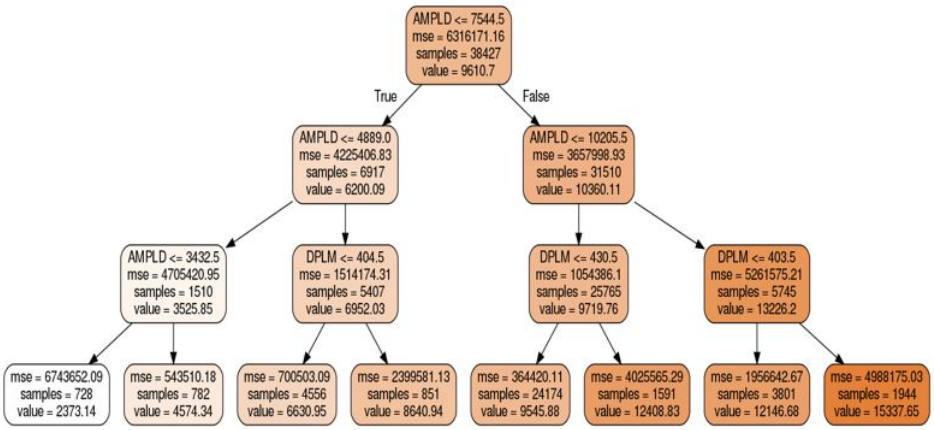

Figure 1. ***Lactation random tree:*** *Sample of unpruned random tree for lactation taken from the data of 82564 Holstein Friesian cows .*

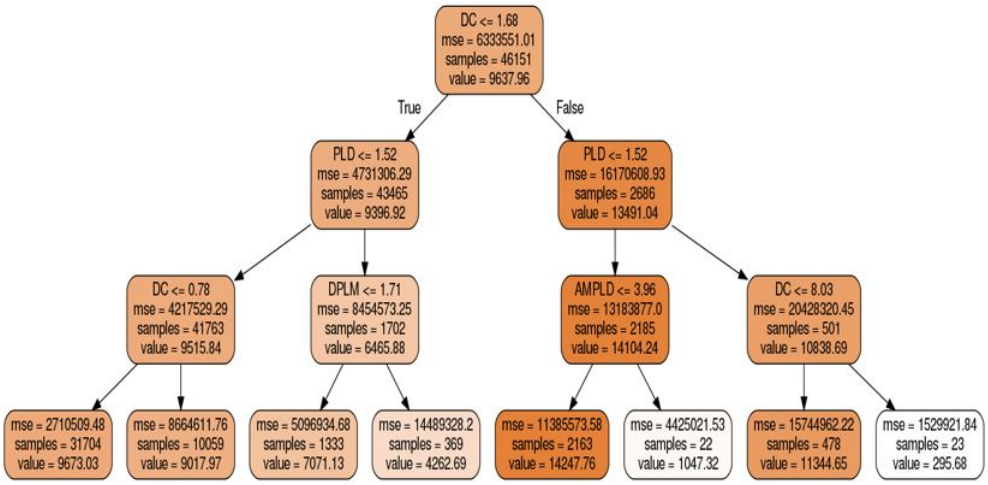

Figure 2. ***Successful insemination random tree:*** *Sample of unpruned random tree for Successful Insemination taken from the data of 82564 Holstein Friesian cows .*

package scikit-learn (https://scikit-learn.org/ stable/). As far as we know, scikit-learn's RandomForestRegressor object is trained using the CART algorithm.

After the training was perfomed, we checked its score on the test set. If the R^2 score of the forest's predictions on the test set is not good, we needed to fine-tune the hyperparameter values of the RFR. When using hyperparameter tuning, we made a grid of the hyperparameter values of the RFR (maximum depth of the trees, minimum samples for splitting, change pruning, etc.). In addition, we needed to split the training set into a real training set and a validation set. We will use the latter to find the best set of hyperparameter. We trained a RFR for each value in the grid and checked their predictive performance using cross-validation. Consequently, the best set of hyperparameters was used for our final model choice. Furthermore, we trained the RFR on the combined training and validation sets, and checked its final predictive performance on the test set, that we did not use at until this final step.

If the model performs well, we need not to perform hyperparameter tuning, because the model is already sufficiently good enough. The original data set consists of 82,563 records, but it contains missing data and invalid data. For this reason, not all records were kept. As for the Lactation Model, 45,461 records were kept, as for the successful insemination model, 82,378 records were kept after the preparation of the data set. In our study, we used 10% of the prepared data set as the test set, and the remaining 90% of the data set as the training set. Actually we ended up having a good model

for the first time with great R^2 scores on the test set. Hence, we did not need hyperparameter tuning and a separate validation set. Of course, hyperparameter tuning could still be done. It could still improve the model by a small margin, but we did not think it was necessary here. It is interesting that the RandomForestRegressor object of scikit-learn automatically calculates the feature importance values of the variables. The sum of these features of importance values is 1. As for the case of artificial insemination, there are two features that stood out with feature importance values of 0.544 and 0.255. For lactation, there are two features with importance values of 0.88 and 0.084. Hence we can investigate the two most important explanatory variables' effect by keeping the remaining variables at their median values.

3 Results and Discussion

Our related program files and the detailed outputs are deposited on the departmental webpage (http://www.mgk.u-szeged.hu/karunkrol/kutatas/cow-article). Although random forest models can handle correlation of variables among the data well, we investigated the dataset in this aspect. Figures 3(a) and 3(b) show the correlation coefficient for lactation analysis and successful insemination analysis, respectively. It was observed that several variables displayed relatively high correlations in the lactation analysis namely, AMPL and AMPLD (0.88), DPLM and PLESI (0.86), PLESI and DC (0.84) DPLM and DC (0.73). However, for SI analysis IPAC and SFAC (0.73) was the only highly correlated case.

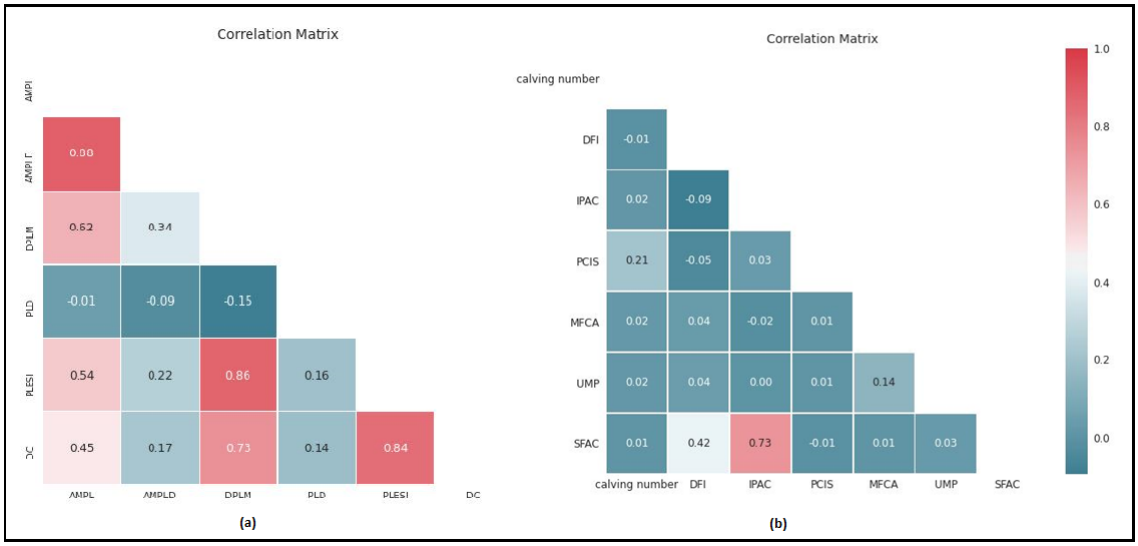

Figure 3. ***The Correlation matrix for Lactation and Successful Insemination*** *Correlation values among various variables for (a) Lactation and (b) Successful Insemination*

First, data separation into the features and targets was done by splitting the data into training and testing sets. It is anticipated that there would be some relationship among all the features and the target value and the model learns this relationship during training. Regarding quantifying the predictive information provided by the variables in the entire random forest, the feature importance of the variables plays a key role. It was evident that AMPLD plays a vital role in the Lactation model with feature importance value of 88.0%, followed by DPLM (8.4%). On the other hand, the most important predictor variable for SI model was IPAC (54.4%) and DFI (25.5%). Figures 4(a) and 4(b) indicate the feature importance of the explanatory variables related to the target variable for lactation and successful inseminations, respectively.

Creating and training the model involves the instantiation of the program object RandomForestRegressor and fitting it on the training data. This results in many expansive trees that form the forest (1000 trees in this case), essential in making and evaluating reasonable predictions. The goodness of fit can be evaluated by means of the R^2 values shown in Table 1.

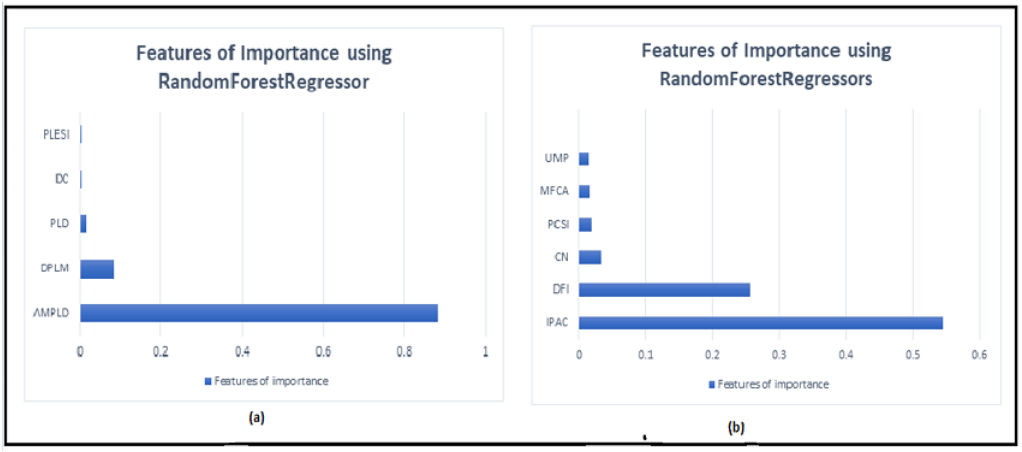

Figure 4. ***Features of importance for Lactation and Successful Insemination.*** *Features importance of variables for (a) Lactation and (b) Successful Insemination in percentage.*

Table 1. R^2 *for training and test set for Lactation and SI*

	Training set	Test set
Lactation	0.998	0.987
SI	0.993	0.948

Since the model performed well on the test set, we did not need to fine-tune the hyperparameter values and use a validation set. The fine-tuning the hyperparameters and checking the model performance on a separate validation set is likely to improve our model's prediction accuracy further. This is a possibility to develop the model. Based on the results from features of importance of the random regressor for both the lactation model and SI model, the model results are summarized in Tables 2 and 3.

Table 2. Features of importance by the random regressor for the Lactation model

Predictors	AMPLD	DPLM	PLD	DC	PLESI
Model Parameters	0.880	0.084	0.015	0.004	0.003

Table 3. Features of importance by the random regressor for SI

Predictors	IPAC	DFI	CN	PCIS	MFCA	UMP
Model Parameters	0.544	0.255	0.034	0.019	0.017	0.016

For both target variables, the two most important explanatory variables have significantly higher feature importance values than the rest of the variables. We tried to visualize the way these affect the targets. We made a data set for each target variable by varying the two most important explanatory variables while keeping the rest of the variables at their median values. We then used trained random forest models to predict the target variables. The results are illustrated in function graph diagrams, as shown in the concluding Figures 5(a) and 5(b). This way, we can know how these two most important variables affect the target variables. The actual effects cannot be visualized in 3D graphs because of the other variables' interactions. These graphs can be analyzed for designing production strategies without specific mathematical preparation by agricultural experts.

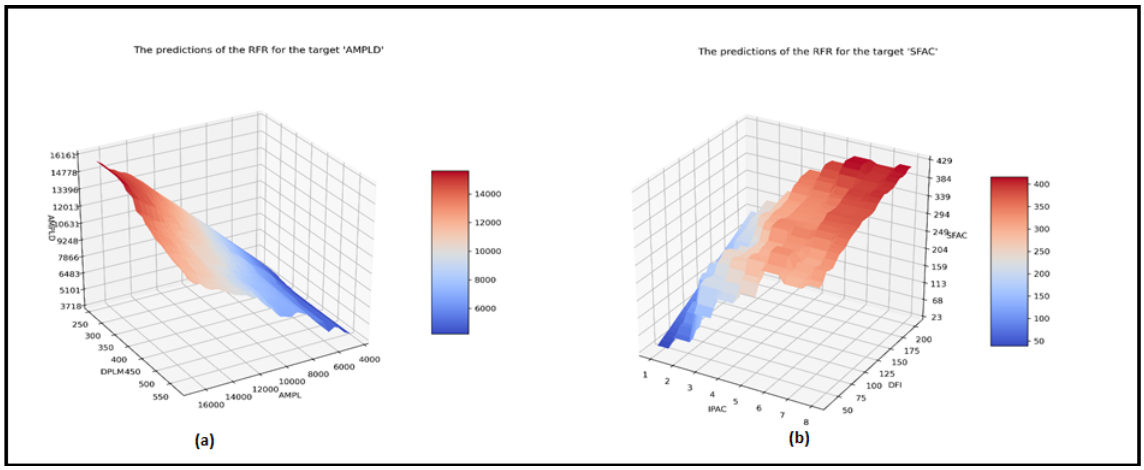

Figure 5. ***RandomForestRegressor predictions for Lactation and Successful Insemination.*** *Diagram for the Predictions of RFR for (a) Lactation and (b) Successful Insemination.*

4 Conclusion

We established an alternative approach to other machine learning-based models concerning problems P1,P2,and P3 by means of random forest regression. The transformed dataset was split into a test size of 10%, and the remaining 90% was used to train the forest. The results indicated that when the target was SI, the prediction of the RFR on the test set and the actual targets had an R^2 $\approx$ 0.948. The important features were the IPAC, DFI, Calving number, PCIS, MFCA, and UMP, with importance scores given in Table 3. When the target was Lactation, the prediction of the RFR on the test set and the actual targets had an $R^2 \approx 0.987$. The important features were AMPLD, DPLM, PLD, DC, and PLESI, with importance scores given in Table 2. Various alternative regression methods were used to analyze this data set; however, all these attempts failed due to the complexity of data and the large sample size. It seems that the RFR is good for practical applications (as for problems P1, P2 and P3).

Acknowledgements

This research was supported by the Ministry of Human Capacities, Hungary grant TUDFO/47138-1/2019-ITM.The research by the fourth author was partially carried out at BME and was supported by the NRDI Fund (TKP2020 IES, Grant No. BME-IE-MISC) based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology.

References

[1] Anzar, M., Farooq, U., Mirza, M.A, et al.: Factors affecting the efficiency of Artificial Insemination in cattle and Buffalo in Punjab, Pakistan. pakista vet journal 2003; 23 (3). http://www.pvj.com.pk

[2] Foote, R.H.: Time of artificial insemination and fertility in dairy cattle. J. Dairy Sci. 1978; 62, 355-358.https://doi.org/10.3168/jds.S0022-0302(79)83248-8

[3] Shaninfar, S., Page, D., Guenther, J., et al.: Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. J. Dairy Sci.2014; 97, 731-742.https://doi.org/10.3168/jds.2013-6693

[4] Mitchell, T.M.: Artificial neural networks Machine Learning, McGraw-Hill International Edition, New York. 1997; 111–112. https://doi.org/10.1002/(SICI)1099-1689(199909)9:3<191::AID-STVR184>3.0.CO;2-E

[5] Dransfield, M.B.G, Pearson,R.E., et al.: Timing of Insemination for Dairy Cows Identified in Estrus by a Radiotelemetric Estrus Detection System. J. Dairy Sci.1998; 81, 1874-1882. DOI: 10.3168/jds.S0022-0302(98)75758-3.

[6] Maatje, K., Loeffler, S.H., and Engel, B.: Optimal time of insemination in cows that show visual signs of estrus by estimating onset of estrus with pedometers. J. Dairy Sci. 1997; 80, 1098–1105. https://doi.org/10.3168/jds.S0022-0302(97)76035-1

[7] Raschka, S., Mirjalili, V.: Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition. 2017.

[8] Uusitalo, L.: Advantages and challenges of Bayesian networks in environmental modeling. Ecol. Modell.2007; 203, 312–318. DOI: 10.1016/j.ecolmodel.2006.11.033

[9] Brody, S., Turner, C.W., Ragsdale, A.C.: The rate of decline of milk secretion with the advance of the period lactation. The Journal of General Physiology. 1923; 5, 442-444. DOI: 10.1085/jgp.5.4.441.

[10] Silvestre, A.M., Petim-Batista, F., Colaco, J.: The accuracy of seven mathematical functions in modeling dairy cattle lactation curves based on test-day records from varying sample schemes. J. Dairy Sci.2006; 89, 1813-1821. https://doi.org/10.3168/jds.S0022-0302(06)72250-0

[11] Mahdi, B., and Naceur, M.: Mathematical Modeling of Lactation Curves: A Review of Parametric Models, IntechOpen.2019; DOI: 10.5772/intechopen.90253.

[12] Thong, N., Fouchereau, R., Frenod, E., Gerard, C., Sincholle, V.: Comparison of forecast models of production of dairy cows combining animal and diet parameters. Computers and Electronics in Agriculture. Elsevier.2020; 170, 1052-1058. DOI: 10.1016/j.compag.2020.105258

[13] Cak, B., Keskin, S., Yilmaz, O.: Regression Tree Analysis for Determining of Affecting Factors to Lactation Milk Yield in Brown Swiss Cattle. Asian Journal of Animal and Veterinary Advances. 2013; 8: 677-682. DOI: 10.3923/ajava.2013.677.682

[14] Mikail, N., Bakir, G.: Regression tree analysis of factors affecting first lactation milk yield of dairy cattle. Applied Ecology and Environmental Research.2019; 17. https://doi.org/10.15666/aeer/$1702_52935303$

[15] Liu, X., Rong, J., Liu, X.: Best linear unbiased prediction for linear combinations in general mixed linear models. Journal of Multivariate Analysis.2008; 99(8). https://doi.org/10.1016/j.jmva.2008.01.004

[16] Gareth, J., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, Springer, New York. 2013.

[17] Breiman, L.: Random Forests Machine Learning.2001; 45, 5-32. https://doi.org/10.1023/A:1010933404324