A study on comparison of various machine learning models for the best prediction of 305 days first lactation milk yield

NAYLA FRAZ
Govind Ballabh Pant University of Agriculture and Technology
B. N. SHAHI
bijendranshahi@gmail.com

Govind Ballabh Pant University of Agriculture and Technology
R. S. BARWAL
Govind Ballabh Pant University of Agriculture and Technology
A. K. GHOSH
Govind Ballabh Pant University of Agriculture and Technology
C. V. SINGH
Govind Ballabh Pant University of Agriculture and Technology
PANKAJ KUMAR
Govind Ballabh Pant University of Agriculture and Technology

Research Article

Keywords: Machine learning models, random forest, gradient boosting regression, extreme gradient boosting, light gradient boosting

Posted Date: June 11th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4484720/v1

Additional Declarations: No competing interests reported.

A study on comparison of various machine learning models for the best prediction of 305 days first lactation milk yield

NAYLA FRAZ[1], B. N. SHAHI[1], R. S. BARWAL[1], A. K. GHOSH[1], C. V. SINGH[1] and PANKAJ KUMAR[2]

[1]Department of Animal Genetics and Breeding , College of Veterinary and Animal Sciences

[2]Department of Soil Water Conservation, College of Technology

G. B. Pant University of Ag. & Tech., Pantnagar, Uttarakhand

ABSTRACT

Machine learning models can be used in dairy industries for the prediction of milk yield in dairy cattle to increase the efficiency of dairy farms and early culling of animals based on 305 days milk yield. Analysis and evaluation of the performances of Multiple linear regression (MLR), Random forest (RF), Gradient boosting regression (GBR), Extreme gradient boosting (XGboost) and Light gradient boosting (lightGBM) were done on the basis of root mean square errors (RMSE) and coefficient of determination (R^2) values. The values of RMSE for MLR, RF, GBR, XGboost and lightGBM for the training period were 478.82, 176.52, 229.65, 271.44 and 214.97 and for the testing period were 469.02, 267.13, 288.10, 338.36 and 293.80, respectively. Similarly, the values of R^2 for the training period were 0.76, 0.92, 0.86, 0.81 and 0.88 and for the testing period were 0.55, 0.85, 0.82, 0.76 and 0.82, respectively. The results obtained suggested that the accuracy and precision of RF, LightGBM, GBR and XGboost models were adequate in predicting first lactation 305 days milk yield, but the best results were obtained by RF in both training and testing period; it outperformed other regression models in predicting first lactation 305 days milk yield. Further, an increase in accuracy and precision can be done by increasing the number of independent variables with a high correlation with the dependent variable and by also increasing the number of observations.

Keywords: Machine learning models, random forest, gradient boosting regression, extreme gradient boosting, light gradient boosting

Machine learning applications are becoming more ubiquitous in dairy farming decision support applications in areas such as feeding, animal husbandry, healthcare, animal behaviour, milking and resource management.

Present address: [1]Department of Animal Genetics and Breeding, College of Veterinary and Animal Sciences,[2]Department of Soil Water Conservation, College of Technology, G. B. Pant University of Ag. & Tech., Pantnagar, Uttarakhand.* Corresponding author email:bijendranshahi@gmail.com

Machine learning models outperform conventional linear models because they can learn from training data and generalise it to unknown test data. The use of software and hardware technologies that support dairy farmers through the automation of on-farm decision-making can help farmers facilitate increased herd sizes without added labour requirements. Conventionally, Multiple Linear Regression (MLR) analysis is being used to fit these prediction models, where the coefficient of determination (R^2) is used as a criterion to evaluate the prediction accuracy of the models. To perform MLR analysis, the data should satisfy certain assumptions, viz., normal distribution, linear association between dependent and independent variables, and absence of multi-collinearity. Therefore, to find a plausible alternative to such assumptions based analytics called parametric linear models, a completely non-parametric statistical computing paradigm, i.e., Machine Learning (ML) models has evolved over the years, which may overcome these constraints. In dairy farms, machine learning has been used effectively in prediction of milk yield (Sharma et al. 2007, Gandhi et al. 2009, 2010, Dongre et al. 2012, Manoj et al. 2014).

Machine learning algorithms and cognate methodologies can provide the necessary prediction accuracy to power these technologies through the ability to self-learn and improve over-time when new data become available. Thus, there has also been an increased prevalence of machine learning algorithms employed through-out the dairy literature. The machine learning models like Random Forests (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting Regression (GBR), and Light Gradient Boosting Machine (lightGBM) machines can be utilized for more accurate predictions vis-à-vis classical MLR analytics.

MATERIAL AND METHODS

Data Collection: For this study, 30 days milk yield, 60 days milk yield , 90 days milk yield, first lactation peak yield (FLPY), first calving interval (FCI), first service period (FSP), days to attend peak yield (DPY), first dry period (FDP), including age at first calving (AFC) and first lactation 305-day milk yield (FL305DMY) were studied using records on 567 daughters from the progeny of 57 sires distributed over a period of 29 years from 1990 to 2019. Data for the present study were collected from the cattle history sheets and daily milk record register of crossbred cattle for various traits maintained at Instructional Dairy Farm, G.B Pant University of Agriculture and Technology, Pantnagar.

For the prediction of first lactation 305 days milk yield, regression models were developed using machine learning techniques using Multiple linear regression (MLR), Random forest (RF), Extreme Gradient boosting package (xgboost), Light gradient boosting (lightGBM) and Gradient boosting regression (GBR) on 567 crossbred cattle data, of which 80% were used during the training and 20% of the overall dataset were used for testing.

Multiple linear regression

For exploring any relationship between small sample sizes of dependent and independent variables, statistical approaches such as regression models are the best instruments (Razi and Athappilly, 2005). Linear regression is one of the most often used linear modelling approaches for examining the relationship between a dependent (response) and many independent (predictors) variables. The dependent variable 'y' is believed to be a function of 'k' independent variables x_1, x_2, x_3,..., x_k in a multiple linear regression model. The following equation can be used to calculate MLR.

$$y = b_o + b_1 x_1 i + \dots\dots\dots + b_k x_k i + e_i$$

where, b_0, b_1,..., and b_k are fitting constants; yi, x_1.....x_k, i are the i^{th} observations of each of the variables y, x_1,..., x_k, respectively; and e_i is a random error term indicating the residual effects on y of variables not explicitly included in the model. e_i can be assumed to be an uncorrelated variable with a zero mean in simple regression models.

Random forest method (RF)

Random forest is one of the most efficacious machine learning methods (Breiman, 2001). It is a part of an ensemble learning classifier which uses a decision tree algorithm in a randomised fashion. This model is capable of both classification and regression tasks. It makes use of CART (classification and regression tree) tools. This method is based on a large number of decision trees in which each decision tree has the space of the variables which is divided into smaller sub-spaces so that each region's data is as uniform as feasible. In this, decision tree structure, the branching point to the two sub-branches is called a node. The first sub-branches i.e. node of the tree is called the root, and the second one is the leaf (Hastie *et al.* 2005). RF breaks variables at each node, chosen from a subset of available data so that the association between trees is reduced. In random forest, each decision tree grows with the help of randomly selected inputs to perform the best division (Breiman, 2001).

These decision trees are generated by using two different sources of randomization. At first, each individual decision tree is trained on a random sample with the same size as the given training set with replacement from the original data. To accomplish so, a subset of the input variables is randomly selected at each node split to find the optimal split.

Gradient boosting regression (GBR)

GBR is a learning algorithm with an integrated model. Gradient boosting uses a tree technique to obtain high accuracy and can also address the problem of over-fitting. A learning technique based on failures combines a number of ineffective learning algorithms. The accuracy of one learning algorithm is not good, however, combining learning algorithms can improve accuracy. Each iteration provides a model, and the algorithm requires 'm' iterations with 'f' weak learners. We use the gradient descent approach to move towards the negative gradient of the loss function in each iteration, which causes the loss function to drop, to minimise the loss function of the model formed by each iteration based on the training set. Finally, the final results are calculated using the weighted total of each stage model.

$$F_m(x) = \sum_{i=1}^{m} \beta_i f_i(x)$$

Extreme gradient boosting package (XGboost):

The Xgboost model is an innovative algorithm suggested by Chen and Guestrin, 2016. Xgboost stands for extreme gradient boosting package. Xgboost is a high-performance Gradient Boosting package that has been built and refined to be versatile, efficient, and portable. This model is based on the concept of "boost," which aims to produce a "strong" learner by integrating all of a group of "weak" learners' predictions using additive training procedures. The main objective functions supported by this boosting package are ranking, classification, and regression (Chen *et al.,* 2017). This model also enables parallelization because it conducts parallelization while determining the best numeration splitting points, resulting in a rapid training speed. When the prediction results are good, the tree building is paused ahead of time, allowing the training pace to be increased. The following is the general function of the prediction at step t:

$$f_i^t = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i)$$

where x_i is the input variable and the learner and predictions at step t are $f_t(xi)$ and $f_i^{(t-1)}$, respectively.

Light gradient boosting method (LightGBM)

LightGBM (Light gradient boosting machine) is a quick and efficient gradient boosting decision tree algorithm or approach designed by Microsoft's 2016 framework (Ma, 2018). The light gradient boosting (lightGBM) model is an effective implementation of the gradient boosting decision tree (GBDT) model (Ke *et al.* 2017), other efficient implementations of this model are xgboost and pGBRT. The (lightGBM) model also handles much more efficiently the classification, regression, and ranking problems in machine learning. GBDT obtains the final answer by trees through ensemble learning i.e., combining multiple decision trees and by adding up or aggregating the results of all the decision trees. Two novel techniques are used in the light gradient boosting ((lightGBM) model to make it more efficient i.e., Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS) in order to deal with a huge number of data instances and features, respectively.

Pre-processing of data using feature selection was done to reduce large number of unwanted traits, as it reduces the time taken to run the models as well as increases the accuracy and precision of results by avoiding over-fitting. The correlations of all the features with the target feature are calculated in this method i.e. it calculates the correlation of each independent feature with that of the target or dependent variable. Features are chosen based on correlation values. A 0.5 threshold will be established for this. A feature is considered for classification if its correlation with the target is greater than 0.5.

The coefficient of determination 'R^2' (Legates and McCabe, 1999) and root mean square error (RMSE) were used to evaluate the quantitative performance of models in this study. In the current study, statistical indices were used to assess the performance of constructed models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_O - Q_P)^2}$$

R^2 and RMSE are commonly used to assess the accuracy and performance of models (Kim and Kim, 2008; Chen *et al.* 2015).

RESULT AND DISCUSSION

Selection of Best Input using Feature Selection

Selection of the best input is an essential part of model development. In this study, the feature selection was used for the selection of the best input for first lactation 305 days milk yield prediction models. It plays a vital role in reducing cost, energy, and time without compromising the results by eliminating the features (variables or traits) as not all features are required for a particular machine learning algorithm. The correlation of all the features with the target variable (FL305DMY), is given in table 1.

Table 1. Correlation of all the features with target variable (FL305DMY)

Features	Correlation values
30 days milk yield (30 DMY)	0.824
60 days milk yield (60 DMY)	0.645
90 days milk yield (90 DMY)	0.640
First lactation peak yield (FLPY)	0.588
First calving interval (FCI)	0.314
First service period (FSP)	0.070
Days to attend peak yield (DPY)	0.022
First dry period (FDP)	0.002
Age at first calving (AFC)	-0.051

So, the features having correlation above 0.5 were selected, which are 30 days milk yield (30 DMY), 60 days milk yield (60 DMY), 90 days milk yield (90 DMY) and First lactation peak yield (FLPY).

Girimal *et al.* (2021) and Arya *et al.* (2020) reported that 305 Days milk yield was significantly impacted by other economically important traits. Similarly, Serdar *et al.* 2021 observed that breed, lactation length, location, and parity were the most crucial variables determining the 305 days milk yield.

Prediction of First Lactation 305 Days Milk Yield using Machine Learning Techniques

In order to learn from data in data-sets, machine learning uses algorithms. They identify patterns, gain insight, make judgments, and assess those judgments. In the present study supervised machine learning was used. So, the data-sets were divided into two groups i.e. Training data- In order to find and understand patterns, a part of the actual data-set was fed into the machine learning model. Testing data- To test unknown data in a machine learning model after it has been constructed (using training data), referred to as testing data, to assess the effectiveness and development of the training of the algorithms and to modify or optimise them for better outcomes. Eighty percent of the data are used as training data while twenty percent as testing data. Usually, training data is larger than testing data. This is to provide the model with as much information as possible for it to identify and learn useful patterns. When the data-sets are supplied to a machine learning algorithm, the programme recognises patterns in the data and draws conclusions.

The qualitative evaluation for first lactation 305 days milk yield was based on the graphical comparison between observed and predicted values. The scatter plot has been plotted between observed and predicted values (Fig 1 to Fig. 10). In table 2, the statistical parameters like root mean square error (RMSE) and coefficient of determination (R^2) were used to evaluate the quantitative performance of the RF model for the prediction of first lactation 305 days milk yield.

MLR model

The multiple linear regression (MLR) technique was used to predict first lactation 305 days milk yield using the best input based on feature selection results. The result showed that there was a large variation between observed and predicted values.

The root mean square error (RMSE) values of MLR model for the training and testing data-set were 469.02 and 478.82 and the values of coefficient of determination (R^2) were 0.76 and 0.55, respectively. The precision of the model was found to be low as a lower R^2 value meant more error in the model. Based on the evaluation of root mean square error (RMSE) and coefficient of determination (R^2) values, it could be concluded that the MLR model lacked in mapping first lactation 305 days milk yield in both accuracy and precision in comparison with other four models. MLR model showed lower precision and accuracy in comparison with other models was also reported by various workers like Ilieva *et al*. 2022.

Random forest (RF)

It was done with the help of PyCaret. It is a low code autoML framework that may be used for both classification and prediction. It showed the agreement of closeness in testing and training dataset results with the best fit line as the data points were less scattered.

For the training and testing data-sets, the root mean square error (RMSE) values for RF model were 176.52 and 267.13, respectively, which expressed its high accuracy, and correspondingly, the coefficient of determination (R^2) values were 0.92 and 0.85, which depicted a stronger linear relationship between observed and predicted dataset. RF was determined as the best model to predict first lactation 305 days milk yield in crossbred cattle based on the study of root mean square error (RMSE) and coefficient of determination (R^2) values. Similar findings were reported by Yordanova *et al.* 2020 in Holstein Friesian cows for their root mean square error (RMSE) values of the RF model was 995.013 while coefficient of determination (R^2) values which was 0.95 which was higher than observed in the present study. Raschia *et al.* 2022 conducted a similar study by constructing machine learning algorithms using RF to find loci that best explained the variation in dairy cattle milk attributes. Sunesh *et al*. 2022 used MLR and Random Forest Model for predicting peak yield in buffaloes.

Gradient boosting regression (GBR)

A positive correlation was found between observed and predicted values in training and testing dataset.The root mean square error (RMSE) values of GBR model for the training and testing data-set were 229.65 and 288.10, respectively. The values of coefficient of determination (R^2) for the testing and training data-sets were 0.86 and 0.82, respectively, which revealed the high precision of the model. Based on the evaluation of root mean square error (RMSE) and coefficient of determination (R^2) values, it could be concluded that the trend predicting first lactation 305 days milk yield for crossbred cattle was satisfactory in GBR model Cai *et al.* 2020 found similar results using GBR model.

Extreme gradient boosting package (XGboost)

The result shows that there was a positive correlation between observed and predicted values of the testing and training dataset. This model gave much better results when compared with MLR, but inferior to GBR and RF.

It was revealed from table 4.9 that for xgboost model, the root mean square error (RMSE) values were 271.44 and 338.36 and the coefficient of determination (R^2) values for the training and testing period were 0.81 and 0.76, respectively. Based on the evaluation of coefficient of determination (R^2) and root mean square error (RMSE) values, it could be concluded that the xgboost model was found to be less precise and accurate in comparison to RF and GBR for prediction of 305 days first lactation milk yield. It is apparent from table 4.9 that it can predict first lactation 305 days milk yield adequately.

Similar study was done by Raschia *et al.* 2022 by constructing machine learning algorithms using xgboost to find loci that best explain the variation in dairy cattle milk attributes.

Light gradient boosting (lightGBM)

It showed that there was a positive correlation between observed and predicted values of testing and training data-sets. For the training and testing datasets, root mean square errors (RMSE) were 214.97 and 293.80, and the coefficients of determination (R^2) were 0.88 and 0.82, respectively. Based on the evaluation of the root mean square errors (RMSE) and coefficient of determination (R^2) values for the prediction of first lactation 305 days milk yield it could be said that the lightGBM model did not perform well as compared to the RF, but its performance was better than the other four models used in this study. Similar work was done by Raschia *et al.* 2022 by constructing machine learning algorithms using lightGBM for SNPs underlying a trait of interest.

Comparative Performance Assessment of Different Machine Learning Models

The comparative results of training and testing data-set sets between the MLR, random forest (RF), extreme gradient boosting package (XGboost), light gradient boosting (lightGBM), and gradient boosting regression (GBR) models in predicting 305 days first lactation milk yield have been presented in table 2.

Among all the developed five models, based on root mean square (RMSE), the models were ranked RF as the highest followed by lightGBM, GBR, xgboost, and MLR for the training data-set and for the testing dataset RF was again highest followed by GBR, lightGBM, xgboost and MLR. Similarly, for the coefficient of determination (R^2) the ranking of models were RF as highest followed by lightGBM, GBR, xgboost, and MLR for the training dataset and RF was highest followed by GBR and lightGBM, xgboost, and MLR for the testing dataset.

The evaluation of the overall performance of multiple linear regression (MLR), random forest (RF), extreme gradient boosting package (xgboost), light gradient boosting (lightGBM), and gradient boosting regression (GBR) for prediction of 305 days first lactation milk yield was conducted for training and testing data-set. It could be concluded from the table that the performance of all the models was not consistent in the training and testing data-set. The MLR model is the simplest among all the other models which were used in the present study, but it was also the model with the least significance. It lagged much behind in mapping first lactation 305 days milk yield for crossbred cattle. xgboost performed well in the training dataset but did not go that well in the testing dataset. The GBR model showed satisfactory performance during the training period and showed a better generalising ability to predict 305 days milk yield. LightGBM slightly performed better than the GBR. The comparative evaluation of performance showed that the RF model outperformed other regression models for predicting 305 days first lactation milk yield in crossbred cattle. The results obtained suggested that the accuracy and precision of RF, lightGBM, GBR and xgboost models were adequate in predicting first lactation 305 days milk yield, but the best results were obtained by RF in both training and testing period, it outperformed other regression models in predicting first lactation 305 days milk yield. So, in the future machine learning models can be used in dairy industries for the prediction of milk yield in dairy cattle to increase the efficiency of dairy farms and early culling of animals based on 305 days milk yield. Further, increase the accuracy and precision can be done by increasing the number of independent variables with high correlation with the dependent variable and by also increasing the number of observations.

The findings of Najubi *et al*. 2010 for the prediction of first lactation 305 days milk yield using test day records through ANN whose R^2 and RMSE values were 0.839 and 423.3, respectively, much more closely resembled the present findings with R^2 but lagged in RMSE values. Its overall accuracy was inferior to all 4 models i.e. RF, xgboost, GBR and lightGBM.

The present investigation's findings closely matched with those that were reported by Gorgulu *et al*. 2012 for ANN models. In this study, the prediction of 305-d milk yield by ANN gave better results that those of MLR, suggesting that ANN can be used as an alternative prediction tool. Similarly, the result of Mundhe *et al*. 2012 for the prediction of first lactation 305 days milk yield using monthly part lactation through ANN models inferred that the R^2 value was 0.89, which also was in close association with the current results. Usman *et al*. 2020 found the value of R^2 as 79.89% for best accuracy for prediction of first lactation 305 days milk yield using ANN models with 16.89% lowest RMSE.

Similarly, Rana *et al*. 2020 concluded that the value of RMSE for ANN model was 121.82 for the prediction of first lactation 305 days milk yield based on bi-monthly test day milk yield which somewhat exceeded the interpretation of the present study. Results of the present study could also be compared with those obtained by other researchers' selective ensembles were derived by Zhou *et al.* 2002 using genetic algorithms.

Table 2. Comparison of different machine learning models

Models	Training		Testing	
	RMSE (kg)	R^2	RMSE (kg)	R^2

MLR	478.82	0.76	469.02	0.55
RF	176.52	0.92	267.13	0.85
GBR	229.65	0.86	288.10	0.82
XGboost	271.44	0.81	338.36	0.76
LightGBM	214.97	0.88	293.80	0.82

The prediction of first lactation 305 days milk yield based on root mean square error (RMSE) were ranked as RF as the highest followed by lightGBM, GBR, xgboost, and MLR for the training data-set and for the testing dataset RF again as highest followed by GBR, lightGBM, xgboost, MLR similarly, for the coefficient of determination (R^2) the ranking of models were RF as highest followed by lightGBM, GBR, xgboost, and MLR for the training dataset and RF as highest followed by GBR and lightGBM, xgboost, and MLR for the testing dataset. RF outperformed other models in both training and testing data-set. The results obtained suggested that the accuracy and precision of RF, LightGBM, GBR and XGboost models were adequate in predicting first lactation 305 days milk yield.

ACKNOWLEDGEMENTS

The present study was M.V.Sc. research work and it was carried out in 2022. The authors are grateful to the Dean, College of Veterinary and Animal Sciences and Joint Director, Instructional Dairy Farm , Pantnagar for providing research facilities and other support for fulfilling the purpose of this study.

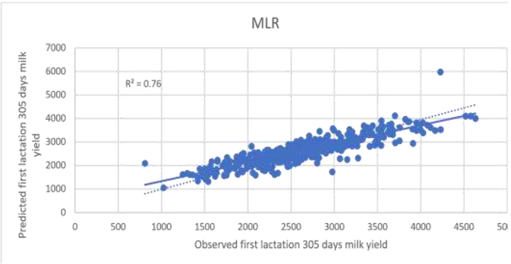

Fig. 1. Scatter plot of first lactation 305 days milk yield using MLR during training period

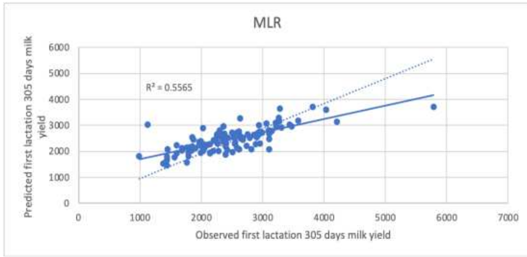

Fig. 2. Scatter plot of first lactation 305 days milk yield using MLR during testing period

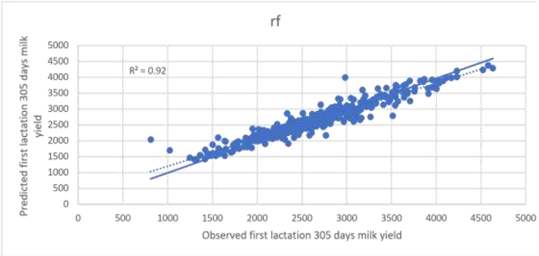

Fig. 3. Scatter plot of first lactation 305 days milk yield using RF during training period

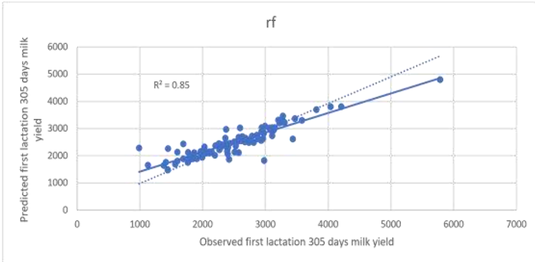

Fig. 4. Scatter plot of first lactation 305 days milk yield using RF during testing period

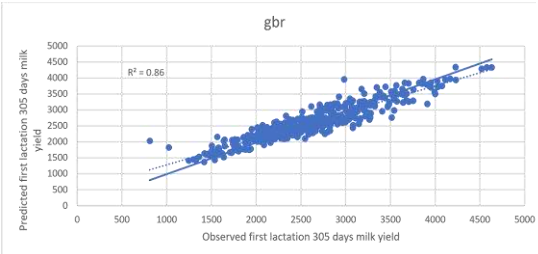

Fig. 5. Scatter plot of first lactation 305 days milk yield using GBR during training period

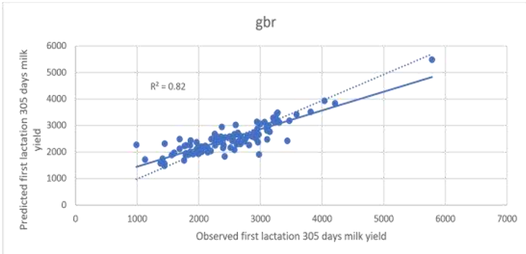

Fig. 6. Scatter plot of first lactation 305 days milk yield using GBR during testing period

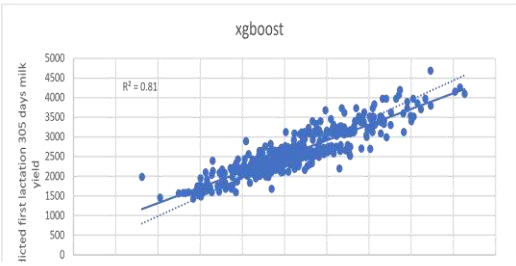

Fig. 7. Scatter plot of first lactation 305 days milk yield using xgboost during training period

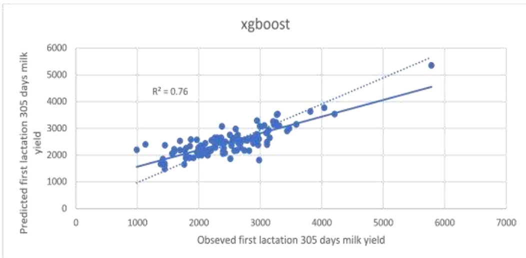

Fig. 8. Scatter plot of first lactation 305 days milk yield using xgboost during testing period

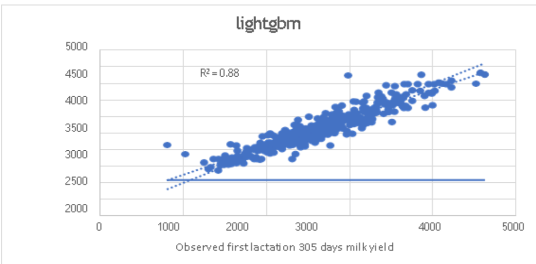

Fig. 9. Scatter plot of first lactation 305 days milk yield using lightGBM during training period

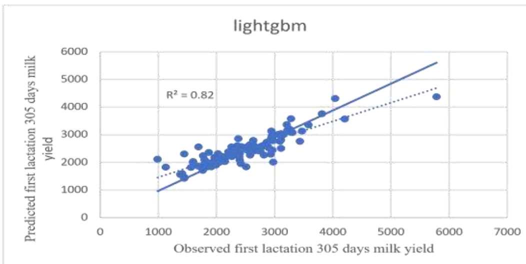

Fig. 10. Scatter plot of first lactation 305 days milk yield using lightGBM during testing period

314 **REFERENCES**

Arya V, Shahi B N, Kumar D, Barwal R S, Kumar S and Gautam L. (2020). Comparison of lactation curve models for fortnightly test day milk yield. *Indian Journal of Animal Science* **90** (3).140-144.

Breiman L.(2001). Random forests. *Machine Learning Sci. Technology* 45(1): 5–32.

Cai J, Xu K, Zhu Y, Hu F and Li L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied Energy* 262: 114566.

Chen J, Li G and Xiao B. (2015). Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature. *Energy Convers Management* 89: 318–329.

Chen T and Guestrin C. 2016. XGBoost: A scalable tree boosting system. *CoRR.,* abs/1603.02754.

Dongre V B, Gandhi R S, Singh A and Ruhil A P. (2012). Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in Sahiwal cattle. *Livestock Science* 147: 192–97.

Gandhi R S, Raja T V, Ruhil A P and Kumar A. (2010). Artificial Neural Network versus Multiple Regression Analysis for prediction of lifetime milk production in Sahiwal cattle. *Journal of Applied Animal Research* 38(2): 233–37.

Girimal D, Kumar D, Shahi B N, Ghosh A K and Kumar S.(2022). Sire evaluation using conventional methods and animal models in Sahiwal cattle. *Indian Journal of Animal Sciences.* **92** (4) : 492-496.

Gorgulu O. 2012. Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks. *South African Journal of Animal Science* 42: 280-287.

Hastie T, Tibshirani R, Friedman J and Franklin J. (2005). The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell., 27*: 83–85.

Ilieva S G, Yordanova A and Kulina H. (2022). Predicting the 305 day milk yield of Holstein-Friesian cows depending on the conformation traits and farm using simplified selective ensembles. *Mathematics* 10: 1254.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y. (2017). LightGBM: A highly efficient gradient boosting decision tree '*In: 31st Conference on Neural Information Processing Systems (**NIPS 2017**)'* at Long Beach. CA, US, during. December 4-9.

Kim S and Kim H S. (2008). Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. *J. Hydrol.* 351: 299–317.

Legates and McCabe Jr G J. (1999). Evaluating the use of goodness of fit measure in hydrological and hydroclimatic model validation. *Water Res*. 35 (1): 233-241.

Maa X, Shaa J, Wang D, YucQian Y and XueqiNiu Y. (2018). Study on a prediction of P2P network loan default based on the machine learning lightGBM and xgboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl*. 31: 24-39.

Manoj M, Gandhi R S, Raja T V, Ruhil A P, Singh A and Gupta A K. (2014). Comparison of artificial neural network and multiple linear regression for prediction of first lactation milk yield using early body weights in Sahiwal cattle. *Indian Journal of Animal Sciences* 84(4): 427–30

Mundhe U T. (2012). Part lactation records for Sahiwal cow evaluation. Thesis, M.V.Sc. NDRI, (Deemed University), Karnal, Haryana.

Njubi D M, Wakhungu J W and Badamana M S. (2010). Use of test-day records to predict first lactation 305-day milk yield using artificial neural network in Kenyan Holstein–Friesian dairy cows; *Trop. Anim. Health Prod. 42*: 639-644.

Rana E, Gupta A, Singh A, Ruhil A, Malhotra R, Yousuf S and Ete G. 2021. Prediction of first lactation 305-day milk yield based on bimonthly test day milk yield records in Murrah buffaloes. *Indian J. Anim. Res.* 55(4): 486-490.

Raschia M A, Rios P J, Maizon D O, Demitrio D and Pol M A. (2022). Methodology for the identification of relevant loci for milk traits in dairy cattle, using machine learning algorithms. *MethodsX.* 9: 101733.

Razi Muhammad, Athappilly Kuriakose.(2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*. 29 (1): 65-74.

Serdar G and Mendes M. (2021). Determining the factors affecting 305-Day milk yield of Dairy cows with regression tree. *J. Food Sci. Technol*. 9: 1154-1158.

Sharma A K, Sharma R K and Kasana H S. (2007). Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modelling. *Applied Soft Computing* 7: 1112–20.

Usman S M, Singh N P, Dutt T, Tiwari R and Kumar A. (2020). Comparative study of artificial neural network algorithms performance for prediction of FL305DMY in crossbred cattle. *J. Entomol. Zool*. 8(5): 516-520.

Yordanova A. and Kulina H. (2020). Random forest models of 305 days milk yield for Holstein cows in Bulgaria; *Application of Mathematics in Technical and Natural Sciences* AIP Conf. Proc. 2302.

Zhou Z H, Wu J and Tang W. (2002). Ensembling neural networks: many could be better than all. *Artificial Intelligence* 137: 239-263.

Sunesh, Balhara A K, Dahiya N K, Himanshu, Singh Rishi Pal and Ruhil A P. (2022). Machine learning algorithms for predicting peak yield in buffaloes using linear traits. *Indian Journal of Animal Sciences* 92 (8): 1013–1019.