Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

JAS

Application of the Different Machine Learning Algorithms to Predict Dry Matter Intake in Feedlot Cattle

Özgür Koşkan[a], Malik Ergin[a*], Hayati Köknaroğlu[a]

[a]*Department of Animal Science, Faculty of Agriculture, Isparta University of Applied Sciences, 32000, Isparta, TÜRKIYE*

ARTICLE INFO
Research Article
Corresponding Author: Malik Ergin, E-mail: malikergin@isparta.edu.tr
Received: 13 October 2023 / Revised: 06 August 2024 / Accepted: 07 August 2024 / Online: 14 January 2025

Cite this article
Koşkan Ö, Ergin M, Köknaroğlu H (2025). Application of the Different Machine Learning Algorithms to Predict Dry Matter Intake in Feedlot Cattle. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 31(1):91-99. DOI: 10.15832/ankutbd.1375383

ABSTRACT
Due to the development of computing technology and different machine learning models, big data sets have gained importance in animal science as well as in many disciplines. The main objective of this study was to compare different machine learning algorithms to predict daily dry matter intake (DMI) in feedlot cattle. The data consisted of 2660 cattle pens placed on feed between January 1988 and December 1997. Machine learning methods were compared in heifers and steers, with 718 in pens of heifers and 1942 in pens of steers. Initial body weight, days on feed, and average proportion of dietary concentrate were used as independent variables to predict DMI in steers and heifers separately. The multivariate linear regression (LR), random forest (RF), gradient boosting regressor (GBR), and light gradient boosting machine (LGBR) algorithms were compared in terms of several performance metrics (MAE, MAPE, MSE, and RMSE). Results showed that the determination coefficient alone is not a good single criterion. It is recommended that the interpretation of model consistency should also consider MAE, MAPE, MSE, and RMSE values. In the current study, all machine learning algorithms yielded similar and lower performance metrics. However, the LGBR and GBR algorithms, were found to perform slightly better than the other algorithms, especially in heifers. Increasing the number of animals and using different independent variables that are related to the DMI can affect the accuracy of DMI prediction.

Keywords: Bigdata, Feedlot cattle, Machine learning algorithms

1. Introduction

With the advancement of computer and internet technologies in recent years, the amount of data has reached a huge size. Thus, "big data" and "data science" have become the most important subjects in science. Fuzzy logic, artificial neural networks, and machine learning methods have been widely used as computer algorithms that model the dataset of classification or estimation problems (Atalay & Çelik 2017). Especially big data, which is seen as the most valuable information of the future, and data mining, which is the technique of processing this data, have become the most important subject of science. This situation has drawn attention in various fields and emphasized using existing data mining methods. Data mining has also been used in animal husbandry to develop prediction models using artificial intelligence. Asadzadeh et al. (2021) compared seven different machine learning methods to predict the live weight of camels by applying different body measurements. Mikail et al. (2014) predicted daily milk yield in Holstein cattle by using support vector machines and artificial neural network models. Huma & Iqbal (2019) used regression trees, support vector machines, and random forest models to predict live weight in Balochi rams, a Pakistani sheep breed. Mammadova & Keskin (2013) detected subclinical and clinical mastitis using support vector machines in cattle. In this study, we used multivariate linear regression (MLR), random forests (RF), gradient boosting regressor (GBR), and light gradient boosting machine algorithms (LGBM) to predict dry matter intake (DMI) in beef cattle. Since DMI is the basis for the calculation and prediction of nutrient requirements, gain, and profit, DMI must be estimated accurately (Hicks et al. 1990). Combined data from cattle fed high-energy diets and initial weight on feed could be used to predict DMI of cattle (NASEM 2016). Koknaroglu et al. (2017) predicted dry matter intake of steers and heifers in the feedlot by using initial weight. Koskan et al. (2014) predicted dry matter intake of steers and heifers in the feedlot by using categorical and continuous variables. The purpose of this study was to predict dry matter intake of feedlot cattle by using machine learning. In the present study, RF, GBR and LGBR algorithms are tree-based algorithms. These tree-based algorithms have a similar but slightly different mathematical background. In addition, when nonlinear relationships between variables appear, these algorithms can be beneficial for accurate prediction. Therefore, a comparative analysis of these algorithms can be useful for future research.

2. Material and Methods

2.1. Material

Closeout information, which was gathered through the Iowa State University Animal Science Extension Program from Iowa cattle producers using the Iowa State University Feedlot Performance and Cost Monitoring Program, was used to derive data for this study. The following information related to animal performance and management were provided and received on mailed-in data sheets: starting date on feed, end of feeding period date, number of cattle in the pen, sex (1= steer, 2= heifer), facility code (1= confinement, 2= partially open lot, 3= open lot), days on feed, initial pay weight, final pay weight, feed efficiency (FE), average percent concentrate, average daily gain (ADG), percent death loss. To obtain a detailed information about the material, study conducted by Koknaroglu et al. (2005), that examined the factors affecting the performance and profitability of beef cattle should be examined. Since DMI was not provided in the close-out sheets, DMI was generated by computer by using the equation DMI = ADG × FE.

From 1988 through 1997, a total of 405 573 animals were represented in the 2759 pens with an average of 150 cattle per pen. No information was available concerning the age or background of the animals. Average body weight at starting time was 322 kg, and animals were fed for an average of 172 d. The 2759 pens consisted of 2032 pens of steers and 727 pens of heifers. Average percentage concentrate ratio was 81%. Each observation (pen of cattle within close-outs) was accepted as an independent observation, even though some observations were obtained from the same farm.

The study used days on feed (DF), initial weight (IW), and concentrate ratio (PRC) as independent variables to predict DMI. The frequency distribution of DMI is shown in Figure 1. In addition, the descriptive statistics of continuous variables are presented in Table 1. Several machine learning methods used in heifers and steers were compared, with 718 in pens of heifers and 1942 in pens of steers.

Table 1- Descriptive statistics for numerical variables

Variable	*Mean*[a]	*SD*[b]	*CV*[c]
Days on feed (d)	172.49	51.24	29.70
Initial weight (kg)	321.86	60.21	18.71
Concentrate ratio (%)	81.41	7.67	9.43
Dry matter intake (kg)	9.77	1.26	12.92

a, b, c: Arithmetic mean, Standard deviation, Coefficient of variation

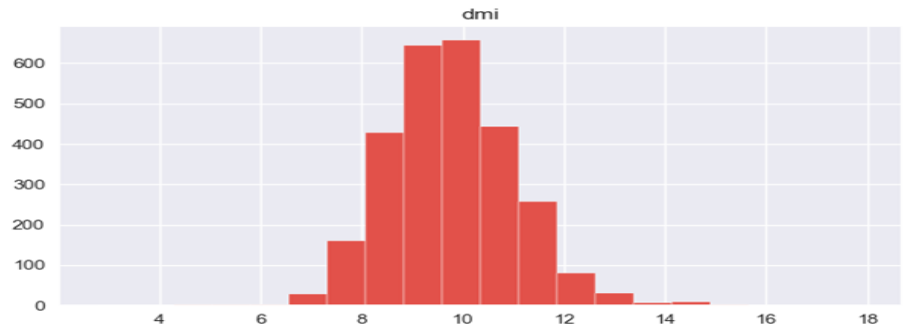

Figure 1- The frequency distribution of dry matter intake (DMI)

2.2. The machine learning algorithms used in the study

2.2.1. Multivariate linear regression (MLR)

Multivariate linear regression is a frequently used and functional algorithm among machine learning algorithms. The multivariate linear regression algorithm best explains the relationship between independent variables and a dependent variable in a linear form. The assumptions of this model include the normal distribution of the data and the elimination of multicollinearity among independent variables. Unlike a simple linear regression model that involves single independent and single dependent variable, an increase in the number of independent variables in a multivariate linear regression model can lead to multicollinearity among these independent variables Equation (1). In general, in a dataset, the desired situation is for each independent variable to have a high correlation with the dependent variable and for the independent variables to have a low correlation with each other (Ray, 2019).

$$DMI_i = \beta_0 + \beta_1 \text{IBW} + \beta_2 \text{DF} + \beta_1 \text{PRCCONC} \ldots\ldots + e_i \quad (1)$$

Where; β_0, Constant; $\beta_1X_1 + \ldots + \beta_nX_n$: coefficient of regression, i: refers to the pen of cattle, IBW, DF, and PRCCONC states the value for i_{th} observation, e: random error term

In the mathematical model given above, β1 means that, the increase in the unit of the dependent variable DMI is equal to the increase in independent variable's (IBW, DF, or PRCCONC) unit.

2.2.2. Random forest (RF)

RF is a supervised machine learning algorithm used in both regression and classification problems when the dependent variable is continuous or categorical, respectively. Its theoretical background is similar to the classification and regression tree (CART). In CART algorithm, single decision tree is used for classification or regression. Therefore, it can cause the overfitting problem on the training dataset. The RF algorithm fixes the major disadvantage of decision trees called overfitting. This is achieved through the combination of multiple classification and regression trees (CARTs), where each tree provides its own prediction, adding diversity to the model (Breiman 2001; Müller & Guido 2016). The basic principle of the RF method is based on the minimization of a function. Random vectors containing observations in dependent and independent variables are defined as *X* and *Y* respectively. It is assumed that the joint distribution, which is a probability distribution of the relationship between these two random vectors, is defined by $P_{XY}(X, Y)$. The main purpose of all these assumptions is to determine an independent function *f(X)* related to the observation in the *X* vector to predict the observation value in the dependent variable. Therefore, this prediction function is determined by a loss function *L(Y, f(X))*, which needs to be minimized. Through this loss function, a penalty technique is applied to measure the distance between *f(X)* and the *Y* vector, penalizing *f(X)* values that are far from *Y*. In the current study, as in the RF method, the least squares method presented in Equation 2 is used to apply the loss function to regression problems:

$$L\big(Y, f(X)\big) = (Y - f(X))^2 \tag{2}$$

To minimize the loss function, the sum of the k basic learners, denoted by $b = [h_1(X), h_2(X), h_3(X),..., h_k(X)]$ in Equation 3, defines the ensemble predictor *f(X)*. This function gives the best prediction of *Y* (Cutler et al. 2012; Bovo et al. 2021).

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} h_i(X) \tag{3}$$

The most advantageous aspect of the RF algorithm is that it can be effectively and easily used in cases of nonlinear relationships between variables. In addition, the RF algorithm is fast for predictions and can handle overfitting problems (Breiman 2015; Çelik & Yılmaz 2023).

In the present study, IW, DF, and PRC variables were considered as independent variables (inputs) in the training dataset. The pen DMI values were considered as dependent (output) in the training dataset. The RF algorithm tries to establish a relationship between the inputs and output variables to predict the pen DMI.

2.2.3. Gradient boosting regressor (GBR)

The gradient boosting algorithm is a tree-based ensemble method developed to enhance predictive performance with respect to the dependent variable in both regression and classification problems. In the boosting method, a series of simple models called weak learners is constructed. These simple models are utilized to correct the errors made by previous models. Similar to RF, it is also formed by combining decision trees, but each one was trained by adjusting the amount of error made by the previous one. When considering the base unit as a decision tree, the final ensemble is indicated as boosted tree (Di Persio & Fraccarolo, 2023). When there are *n* observations in Equation 4, and it is assumed that each observation value of the independent variable x corresponds to a value in the dependent variable y, the GBR algorithm aims to find an estimate $\hat{f}(x)$ which approximates the function $f^*(x)$ that maps observations to the dependent variable.

$$S = \{(x_i, y_i)\}_{i=1}^{n} \tag{4}$$

To achieve this, the algorithm minimizes the expected value of the loss function $L(y, f(x))$. Then, as seen in Equation 5, the additive prediction of the $f^*(x)$ is generated by weighting all obtained functions.

$$f_k(x) = f_{k-1}(x) + p_k h_k(x) \tag{5}$$

Where: p_k, weight of the t^{th} base learner (k = 1, 2, ..., K); h_k, a base learner.

Suppose $L(y_i, a)$ is differentiable loss function, the prediction of the $f^*(x)$ is calculated by an iterative process in Equation 6. Here, in each new iteration, a new tree is constructed that corrects the errors remaining from the predictions of the previous tree.

$$f_0(x) = argmin_a \sum_{i=1}^{n} L(y_i, a) \quad (6)$$

The base learners seek to minimize the expected value of the loss function $L(y_i, a)$ by Equation 7.

$$\left(p_k h_k(x)\right) = argmin_{p,h} \sum_{i=1}^{n} L(y_i, f_{k-1}(x_i) + ph(x_i)) \quad (7)$$

Subsequently, the pseudo-residuals of each observation, which represents the error remaining from the prediction of the previous tree, are calculated according to Equation 8 (Sibindi et al., 2022; Otchere et al., 2022).

$$r_{ti} = \left[\frac{\partial L(y_i, f(x))}{\partial f(x)}\right]_{f(x)=f_{k-1}(x)} \quad (8)$$

There are several advantages such as robustness against non-linear relationships among variables, handling of outliers in the dataset, automatic feature selection for predicting the dependent variable, ability to work with independent variables that have high linear correlations with each other (multicollinearity) and support for various loss functions (Hastie et al. 2009; Ogutu et al. 2011; Hong 2015). The GBR algorithm creates a series of decision trees to estimate the pen DMI in steers and heifers. Each tree contains rules that determine the value of the pen DMI value based on the observations in the input variables (IW, DF, and PRC). The individual predictions of each tree are weighted and combined to optimize the overall prediction accuracy of the model.

2.2.4. Light gradient boosting regressor (LGBR)

Similar to GBR algorithm, the LGBR algorithm uses decision trees in classification and prediction problems. This algorithm has faster training speed and higher performance than many other algorithms when creating models (Chen et al. 2019). Contrary to other tree-based algorithms such as GBR, and XBGR, LGBR algorithms typically grow the tree vertically which is one of the most effective aspects of LGBR for handling large-scale data and variables (Sun et al. 2018). The mathematical background of the LGBR algorithm is similar to GBM but differs in some aspects. A detailed explanation of the calculations for LGBR was given according to Sun et al. (2018). When the training data set is assumed to be as in equation 4, the expected value of the loss function is calculated with Equation 9.

$$\hat{f} = \arg\min_f E_{y,X}\, L\big(y, f(x)\big) \quad (9)$$

The LGBR algorithm integrates *T* regression trees to make its final model using Equation 10.

$$f_T(X) = \sum_{t=1}^{T} f_t(X) \quad (10)$$

Regression trees are characterized by the number of leaves *J* and an index *q* representing the rules of the tree, where the example weight $w_{q(x)}$ applies to the q^{-th} leaf of the regression tree, q ∈{1,2, …, J}. Therefore, in Equation 11, LGBR is additively trained over t steps.

$$\Gamma_t = \sum_{i=1}^{n} L\big(y_i, F_{t-1}(x_i) + f_t(x_i)\big) \quad (11)$$

The objective function is quickly approximated with Newton's approach. After some simplification steps, Equation 11 will be replaced by Equation 12.

$$\Gamma_t \cong \sum_{i=1}^{n} \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) \quad (12)$$

In Equation (X) g_i and h_i states the 1st and 2nd order gradient statistics of the loss function. When *Ij* represents the sample set of leaf *j*, Equation 12 can be explained as Equation X.

$$\Gamma_t = \sum_{j=1}^{j} \left(\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right) \quad (13)$$

In the case of $q(x)$ a tree structure, the optimal scores of the leaf weight for leaf nodes w_j^* and extreme values of Γ_k would be expressed as in Equations 14 and 15.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \quad (14)$$

$$\Gamma_T^* = -\frac{1}{2}\sum_{j=1}^{J} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} \quad (15)$$

The objective function is finally calculated by integrating the split in Equation 16.

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}\right) \quad (16)$$

Where; I_L and I_R state the left and right branches, respectively.

2.3. Evaluation metrics of prediction models

A few performances scores that are frequently used in the literature were used to compare algorithms for predicting capability of pen DMI. The main objective of using these evaluation criteria was to compare the performance of the machine learning models we have used. In this study, evaluation criteria such as mean absolute error (MAD), mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE) were considered (Çelik & Yılmaz 2017). Equations for performance scores of R^2, MAD, MAPE, MSE and RMSE are presented in Equations (3-7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{ip})^2}{\sum_{i=1}^{n}(y_{ip} - \bar{y}_{ip})^2} \quad (3)$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_{ip}\right| \quad (4)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{ip}}{y_i}\right| * 100 \quad (5)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{ip})^2 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{ip})^2} \quad (7)$$

Where: n is the number of animals, p is the number of dependent variables for predicting; DMI, y_i is the actual observation value of DMI of cattle, y_{ip} is the predicted DMI.

K-fold cross validation was applied to all models for optimizing the models. Tested and selected hyperparameters for all algorithms were presented in Table 2. In fact, cross validation is a resampling method. In this method, the dataset is divided into sub-samples that are different from each other and have an equal number of observations. Observations from the training dataset are selected into sub-samples randomly and without replacement. The model in question is trained with k-1 sub-grouped samples. In this approach, during each k iterations, a different data-sample is held out for testing while the remaining k-1 sub-sample is used for training. This process is repeated k times, ensuring that each of the k folds is used exactly once for validation. Ultimately, the results obtained from the k iterations are combined to make a prediction (Refaeilzadeh et al. 2016). In the present study, all statistical analyses were performed using R software version 4.4 (R Core Team, 2024). The RF, GBR and MLR algorithms were executed using the *caret* package (version 6.0.94) that consists of several regression algorithms (Kuhn, 2008). The LGBR algorithm was evaluated using the *LightGBM* package (version 4.3.0) in Python (Ke et al., 2017). Prior to conducting the analyses, all data were divided into training and test sets, and then hyperparameters were randomly searched. For steers, the training and test dataset ratios were determined as 60% and 40%, respectively. In heifers, 70% and 30% of observations were randomly split for the training and test datasets, respectively.

In all training processes, days on feed (DF), initial weight (IW), and the average proportion of dietary concentrate (PRC) were considered as independent variables to predict dry matter intake (DMI). The importance of the predictors was assessed using the *varImp* function from the caret package in R. One of the main objectives of machine learning algorithms is to determine which variable is most important in explaining the variation of the dependent variable (dry matter intake, in the current study). Variable importance uses specific coefficients (gain, weight, cover etc.) to evaluate the relationship between the dependent and independent variables. For instance, in multivariate linear regression, each independent variable is ordered based on correlation coefficients to determine its significance in predicting the DMI variable. This process aids in dimensionality reduction and feature selection, which enhance the model's predictive capability. Determining the key independent variables that account for most of the variance in the predictor variable is essential for developing highly predictive models.

Table 2- Summary of the tested and best hyperparameters of all algorithms

	Steers		*Heifers*	
Algorithms	***Tested Parameters***	***Best Parameters***	***Tested Parameters***	***Best Parameters***
MLR	No need	No need	No need	No need
RF	ntreeTry = 1000 mtry = 1:5	ntreeTry = 1000 mtry = 1	ntreeTry = 500 mtry = 1:5	ntreeTry = 500 mtry = 2
GBM	n.trees = 100:250:500:1000 interaction.depth = 1:3 shrinkage = 0.01:0.05:0.1 n.minobsinnode = 20	n.trees = 500 interaction.depth = 3 shrinkage = 0.01 n.minobsinnode = 20	n.trees = 100:250:500:1000 interaction.depth = 1:3 shrinkage = 0.01:0.05:0.1 n.minobsinnode = 20	n.trees = 200 interaction.depth = 3 shrinkage = 0.01 n.minobsinnode = 10
LGBR	learning_rate = 0.05:0.1 boosting_type = gbdt num_leaves = 2:10 max_depth = 2:8 bagging_freq = 2:6 bagging_fraction = 0.7:0.8 iterations = 25:1000	learning_rate = 0.08 boosting_type = gbdt num_leaves = 3 max_depth = 2 bagging_freq = 6 bagging_fraction = 0.75 iterations = 75	learning_rate = 0.05:0.1 boosting_type = gbdt num_leaves = 2:10 max_depth = 2:8 bagging_freq = 2:6 bagging_fraction = 0.7:0.8 iterations = 25:1000	learning_rate = 0.1 boosting_type = gbdt num_leaves = 2 max_depth = 2 bagging_freq = 2 bagging_fraction = 0.4 iterations = 50

GBR: Gradient Boosting Regressor, LGBR: Light Gradient Boosting Machine, MLR: Multivariate Linear Regression and RF: Random Forests

3. Results and Discussion

The box plot of all variables in the models is provided in Figure 2. When examining the box plot graphic, it is observed that initial weight, which is one of the independent variables, fits the normality, the rest deviate from the normality. Therefore, it is decided that non-linear machine learning algorithms are suitable for this dataset.

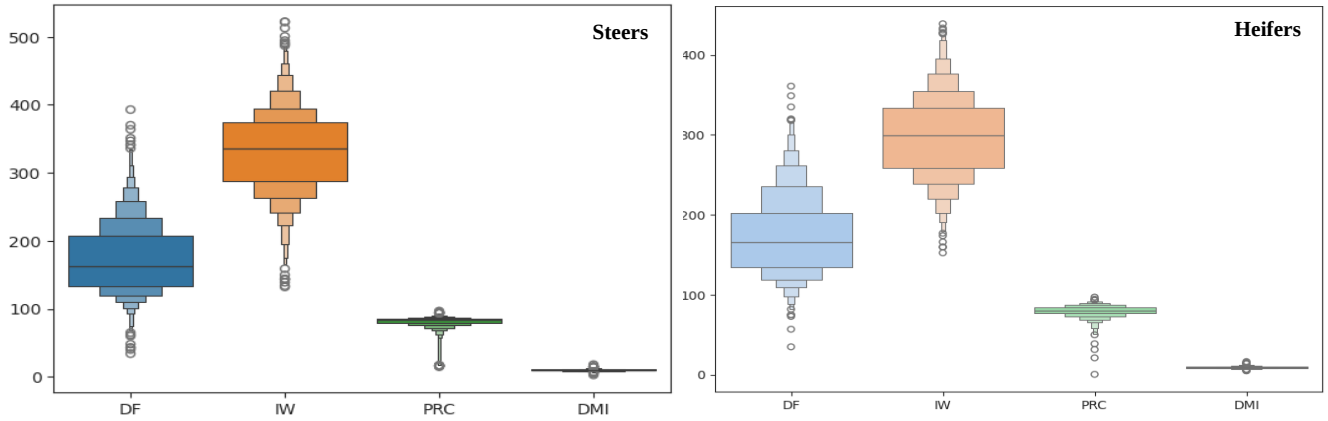

Figure 2- The box plots of the dependent and independent variables in both steers and heifers, DF: days on feed, IW: initial weight (kg), PRC: average proportion of dietary concentrate; DMI: dry matter intake (pen)

In Table 3, the prediction performances of all algorithms on the train and test dataset for predicting DMI in steers and heifers were presented. Various popular performance scores such as R^2, MAD, MAPE, MSE, and RMSE were compared. When looking at the average determination of coefficients (R^2) of the four models, it was seen that GBR and LGBR algorithms explained the variation better than other models. In the test dataset, MAD ranged between 0.64 (RF) and 0.71 (MLR). In addition, MAPE changed between 6.4 (RF) and 7.3 (LR). When assessing the models based on MSE and RMSE, the RF model had the lowest values with 0.72 and 0.85, respectively. According to all performance scores, all algorithms except from MLR, yielded similar predictions on DMI. LGBR and GBR algorithms were considered superior in terms of R^2, while RF performed best in terms of error scores. When training and test datasets were examined simultaneously, it can be said that both LGBR and GBR algorithms resulted quite consistent values in both datasets. The minimal differences between training and test scores suggests that overfitting did not occur in the LGBR and GBR algorithms. Furthermore, it is believed that the LR model, used in this study has a lower impact on predicting daily dry matter intake in cattle than non-linear models.

A test dataset containing 214 observations was used for predicting DMI while 504 observations were used for training in the heifers. For test dataset, the highest average coefficients of determination were found in GBR (R^2 = 0.43) and MLR (R^2 = 0.44). In addition, MAD ranged between 0.69 (RF) and 0.76 (LGBR) in test dataset. MAPE ranged between 7.5 (RF) and 8.4 (LGBR). Regarding MSE and RMSE criteria, the highest values were calculated in LGBR (MSE = 1.15, RMSE = 1.32). The lowest MSE and RMSE values were found as 0.86 and 0.93 for the GBR model, respectively. In all algorithms, it can be observed that the performance scores of all algorithms were not exposed to overfitting and remained consistent based on the relationship between the training and test datasets. Considering all performance scores, the GBR algorithm showed superior results in predicting DMI

in heifers. Statistically, the observed differences between the results of steers and heifers are thought to arise from differences in sample sizes and data splitting.

Table 3- Performances of machine learning algorithms on the training and testing datasets for DMI in steers and heifers

Algorithm	***Steers***									
	Training (n=1225)					***Testing (n=817)***				
	R^2	***MAD***	***MAPE***	***MSE***	***RMSE***	***R^2***	***MAD***	***MAPE***	***MSE***	***RMSE***
RF	0.46	0.65	6.67	0.83	0.91	0.42	0.64	6.4	0.72	0.85
LGBR	0.47	0.64	6.4	0.72	0.85	0.45	0.67	7.0	0.81	0.90
GBR	0.49	0.62	6.35	0.67	0.82	0.45	0.68	6.79	0.86	0.93
MLR	0.40	0.66	6.72	0.83	0.91	0.41	0.71	7.23	0.86	0.93
Algorithm	***Heifers***									
	Training (n=504)					***Testing (n=214)***				
	R^2	***MAD***	***MAPE***	***MSE***	***RMSE***	***R^2***	***MAD***	***MAPE***	***MSE***	***RMSE***
RF	0.44	0.74	7.9	1.01	1.03	0.41	0.69	7.5	0.88	0.94
LGBR	0.43	0.72	7.7	0.99	0.98	0.40	0.76	8.4	1.15	1.32
GBR	0.48	0.71	7.6	1.0	1.0	0.43	0.70	7.73	0.86	0.93
MLR	0.44	0.73	7.7	1.0	1.0	0.44	0.73	7.8	1.0	1.0

n: number of cattle, GBR: Gradient Boosting Regressor, LGBR: Light Gradient Boosting Machine, MLR: Multivariate Linear Regression and RF: Random Forests, R^2: Coefficient of Determination, MAD: Mean Absolute Deviation, MAPE: Mean Absolute Percentage Error, MSE: Mean Squared Error, RMSE: Root Mean Squared Error

The variable importances in steers and heifers were illustrated in Figure 3. Variable importances were achieved using training datasets. Consequently, variable importance plots were generated based on the best algorithm for heifers and steers. GBR was used for heifers, whereas LGBR was used for steers to calculate importances of independent variables. The IW is the most contributed variable for variation in DMI in both steers and heifers. Furthermore, in steers, DF variable is the second most important variable, while in heifers, the PRC variable is the second most important variable.

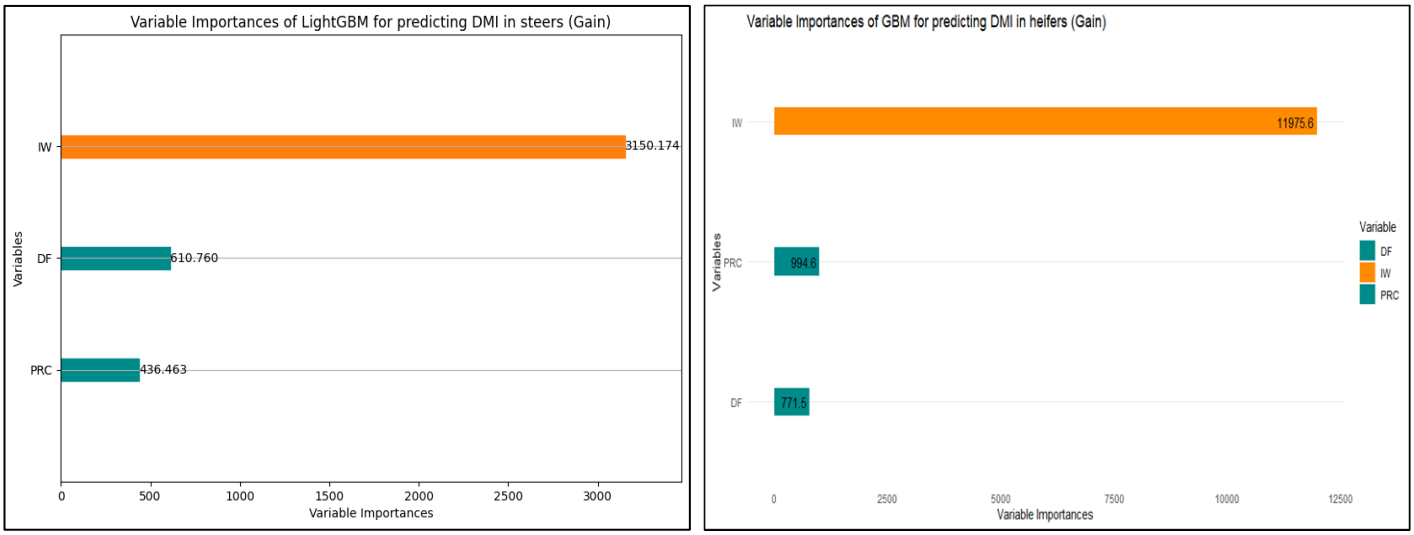

Figure 3- Variable importances of independent variables in steers and heifers for training datasets, DF: days on feed, IW: initial weight (kg), PRC: average proportion of dietary concentrate; DMI: dry matter intake (pen)

In the literature, several DMI predictions were evaluated using different independent variables. Blake et al. (2023) studied repeated measurements ANOVA, repeated measurements random forest regression, and classical random forest regression to predict DMI in 125 bulls and 53 steers. Researchers used different measured variables such as age, sex, full body weight, average daily gain, and climate factors. At the end of the study, the authors stated that the best prediction was performed by the repeated measurements random forest regression model with R2= 0.65 and MSE= 1.09. In the study, the random forest regression model was reported with R2= 0.45 and MSE= 1.71 for predicting DMI. Similarly, the RF model resulted in coefficients of determination of 0.44 and 0.46 in steers and heifers, respectively, in the present study. Authors did not separate results for both bulls and steers. In addition, they used various independent variables for predicting DMI, such as climate factors. These are the main differences between the two studies.

Various technologies have been used for predicting DMI in cattle. For instance, Shadpour et al. (2022) examined different ANN structures to accurately predict weekly DMI in Canadian Holstein cows. Authors used mid-infrared reflectance

spectroscopy (MIRS), weekly average DMI, test-day milk yield, fat yield, protein yield, metabolic body weight, calving traits, country, and herd data. Although we did not use ANN in our study, the results are similar to the findings of the researchers. Additionally, the use of independent variables and a technology such as MIRS also has a significant impact on the results. In addition, Salleh et al. (2023) investigated the use of machine learning algorithms for predicting DMI. They employed partial least squares regression (PLS), support vector machine regression (SVM), and random forest regression (RF) algorithms using MIRS (milk mid-infrared spectra) values. The authors reported that the determination coefficients (R^2) ranged from 0.52 to 0.65. The best determination coefficient was observed in the PLS regression approach with 0.65, followed by 0.62 in RF regression and 0.55 in SVM regression approach. Furthermore, mid-infrared reflectance spectroscopy (MIRS) analysis of milk and near-infrared reflectance spectroscopy (NIRS) analysis of feces from cows were compared in terms of predicting DMI (Lahart et al. 2019). In this comparative study, authors used traditional linear regression and partial least squares regression on the data from 457 cows. In the study where various combinations of the MIRS and NIRS wavelengths with known animal energy sinks and status traits used resulted in the equation with R2= 0.68 and RMSE= 1.52 kg. When compared with the present study, results could change due to independent variables and utilized technologies. In the literature, high determination coefficients are often found in predicting animal body weight, indicating a supportive relationship. Similarly, in predicting DMI, determination coefficients ranged from 0.40 to 0.70, which is consistent with our study.

While a high correlation may seem important, it is crucial to consider the performance of the applied models in revealing the actual relationship. If the amount of variance explained by the independent variables is inherently low, the consistency of the models in revealing the accuracy becomes important. Therefore, the determination coefficient alone should not be the single criteria. The interpretation of model consistency should also consider MAE, MAPE, MSE, and RMSE values. In our study, these values were found to be low and similar.

There is variation in the R^2 for each sex (steers and heifers) of the different machine learning models presented in Table 1 and Table 2. While the best model in steers was RF, the best model in heifers was GBR. The reason for this is that the train and test dataset in machine learning models contains a certain number of individuals. As we mentioned in the method section, to establish a model with the training data set, 70% of the total data set was used and the model was tested for the remaining 30%. For the steers and heifers, 1225 and 504 animals were used for the training data sets, while 817 and 214 animals were used for the test data sets. Therefore, the R^2 of the created model also changes in this direction.

4. Conclusions

The RF, GBR, LGBR, and LR algorithms were used to evaluate model performance in predicting pen DMI of feedlot cattle. In steers, GBR and LGBR algorithms outperformed others with a coefficient of determination of 0.45 in terms of predicting pen DMI, but our results were found lower or similar when compared with other related studies. Compared to other studies conducted on predicting DMI, the model performances in our study differed. The reason for this could be the high variation in the animals' ages, herd management, and body morphological characteristics. Furthermore, the technology such as MIRS could be essential factor in accurately prediction for DMI. In addition, it should be noted that the statistical processes performed on the data may be different for each study. The initial body weight was determined as a key feature for predicting pen DMI in both steers and heifers. Furthermore, while days on feed trait most contributed feature in steers, average proportion of dietary concentrate was significantly important to predict pen DMI in heifers. Therefore, days on feed and average proportion of dietary concentrate variables may contribute to herd management strategies at farm level. This study suggests that ensemble learning techniques can be tried to improve the model performance of DMI predictions. Increasing number of animals and using different independent variables that related to the DMI can affect the accuracy of DMI prediction.

References

Asadzadeh N, Bitaraf D E, Shams H J, Zare M, Khojestehkey S, Abbaasi S & Shafie N (2021). Body weight prediction of dromedary camels using the machine learning models. *Iranian Journal of Applied Animal Science* 11(3): 605-614

Atalay M & Çelik E (2017) Artificial intelligence and machine learning applications in big data analysis. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* 9(22): 155-172

Blake N E, Walker M, Plum S, Hubbart J A, Hatton J, Mata-Padrino D, Holásková I & Wilson M E (2023). Predicting dry matter intake in beef cattle. *Journal of Animal Science* 101: skad269

Bovo M, Agrusti M, Benni S, Torreggiani D & Tassinari P (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. Animals 11(5): 1305

Breiman L & Cutler A (2015). Random forest. Retrieved June 23, 2015, from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Breiman L. (2001). Random forests. In: Blockeel H & Leuven K U (Eds.), Machine Learning, Scientific Research Publishing, New York, pp. 5-32

Celik S & Yılmaz O (2017). Comparison of different data mining algorithms for prediction of body weight from several morphological measurements in dogs. *Journal of Animal and Plant Sciences* 27(1): 57-64

Çelik Ş & Yılmaz O (2023). Investigation of the Relationships between Coat Colour, Sex, and Morphological Characteristics in Donkeys Using Data Mining Algorithms. *Animals* 13(14): 2366. https://doi.org/10.3390/ani13142366

Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, Gao H & Wu J (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. Institute of Electrical and Electronics Engineers 7: 960-968

Cutler A, Cutler D R & Stevens J R (2012). Ensemble Machine Learning: Methods and Applications. In Zhang C. & Ma Y. (Eds.), Random forests (pp. 157–175) Springer

Defalque G, Santos R, Bungenstab D, Echeverria D, Dias A & Defalque C (2024). Machine learning models for dry matter and biomass estimates on cattle grazing systems. Computers and Electronics in Agriculture 216: 108520

Di Persio L & Fraccarolo N (2023). Energy consumption forecasts by gradient boosting regression trees. Mathematics 11(5): 1068

Hastie T, Tibshirani R & Friedman J H (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed., pp. 1-758). Springer

Hicks R B, Owens F N, Gill D R, Oltjen J W & Lake R P (1990). Daily dry matter intake by feedlot cattle: influence of breed and gender. *Journal of Animal Science* 68(1): 245-253

Hong W (2015). Wavelet Gradient Boosting Regression Method Study in Short-Term Load Forecasting. Smart Grid 5: 189–196

Huma Z E & Iqbal F (2019). Predicting the body weight of Balochi sheep using a machine learning approach. *Turkish Journal of Veterinary & Animal Sciences* 43(4): 500-506

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q & Liu T Y (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30: 3146–3154

Koknaroglu H, Demircan V & Yilmaz H (2017). Effect of initial weight on beef cattle performance and profitability. Agronecio 13(1): 26-38

Koknaroglu H, Loy D D, Wilson D E, Hoffman M P & Lawrence J D (2005). Factors affecting beef cattle performance and profitability. The Professional Animal Scientist 21(4): 286-296

Koşkan O, Koknaroglu H, Loy D D & Hoffman M P (2014). Predicting dry matter intake of steers and heifers in the feedlot by using categorical and continuous variables. In: American Society of Animal Science Annual Meeting, 20 – 24 July, Kansas City, Missouri, USA, pp. 721-721

Lahart B, McParland S, Kennedy E, Boland T M, Condon T, Williams M, Galvin N, McCarthy B & Buckley F (2019). Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. *Journal of dairy science* 102(10): 8907-8918

Mammadova N & Keskin I (2013). Application of the support vector machine to predict subclinical mastitis in dairy cattle. *The Scientific World Journal* 2013: 897-906

Mikail N, Keskin I & Altay Y (2014). The use of artificial neural networks and support vector machines methods in milk yield prediction of holstein cows. In: Proceedings of the International Mesopotamia Agriculture Congress, 22 – 25 September, Diyarbakir, 1137 pp

Müller A C & Guido S (2016). Introduction to Machine Learning with Python: A Guide For Data Scientists. O'Reilly Media, USA.

National Academies of Sciences, Engineering, and Medicine (NASEM) (2016). Nutrient Requirements of Beef Cattle, 8th revised edn. Washington, DC: The National Academies Press

Ogutu J O, Piepho H P & Schulz-Streeck T (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC* proceedings 5: 1-5

Otchere D A, Ganat T O A, Ojero J O, Tackie-Otoo B N & Taki M Y (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering* 109: 244-254

PyCaret (2020). An Open Source, Low-Code Machine Learning Library in Python. Retrieved in August, 23, 2023 from https://pycaret.org/

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ray S (2019). A quick review of machine learning algorithms. In: International conference on machine learning, big data, cloud, and parallel computing, 14 – 16 February, Faridabad, India, pp. 35-39

Refaeilzadeh P, Tang L & Liu H (2016). Cross – Validation. Springer, New York.

Salleh S M, Danielsson R & Kronqvist C (2023). Using machine learning methods to predict dry matter intake from milk mid-infrared spectroscopy data on Swedish dairy cattle. *Journal of Dairy Research* 90(1): 5-8

Shadpour S, Chud T C, Hailemariam D, Oliveira H R, Plastow G, Stothard P, Lassen J, Baldwin R, Miglior F, Baes C F Tulpan D & Schenkel F S (2022). Predicting dry matter intake in Canadian Holstein dairy cattle using milk mid-infrared reflectance spectroscopy and other commonly available predictors via artificial neural networks. *Journal of dairy science* 105(10): 8257-8271

Sibindi R, Mwangi R W & Waititu A G (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports 5(4): e12599.

Sun X, Liu M & Sima Z (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Research Letters 32: 101084