R. Bras. Zootec., 48:e20170263, 2019
https://doi.org/10.1590/rbz4820170263

Animal production systems and agribusiness
Full-length research article

Revista Brasileira de Zootecnia

Brazilian Journal of Animal Science
ISSN 1806-9290
www.rbz.org.br

Determination of factors affecting dairy cattle: a case study of Ardahan province using data mining algorithms

Koksal Karadas[1*], Avni Birinci[2]

[1] Igdir University, Agricultural Faculty, Department of Agricultural Economics, Igdir, Turkey.
[2] Ataturk University, Agricultural Faculty, Department of Agricultural Economics, Erzurum, Turkey.

***Corresponding author:**
kkaradas2002@gmail.com

Received: November 23, 2017
Accepted: May 1, 2018

How to cite: Karadas, K. and Birinci, A. 2019. Determination of factors affecting dairy cattle: a case study of Ardahan province using data mining algorithms. Revista Brasileira de Zootecnia 48:e20170263. https://doi.org/10.1590/rbz4820170263

ABSTRACT - This study was conducted to compare predictive performances of different data-mining algorithms for determining factors influencing the average daily milk yield at dairy cattle enterprises of Ardahan province, located in the Eastern Anatolia region of Turkey. The algorithms employed in the present study were Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), Exhaustive Chi-Square Automatic Interaction Detector (Exhaustive CHAID), Multivariate Adaptive Regression Splines (MARS), and Multilayer Perceptron (MLP). The MARS algorithm outperformed the other algorithms in the study. Visual results of CART revealed that the culture-breed cows with a lactation length greater than 237.500 days had the highest milk yield (10.64 kg/day). Culture-breed cows calving earlier than the 4th month gave the highest yield of approximately 10 kg/day in the regression tree of CHAID. The Exhaustive CHAID results were almost the same as the structure of the CHAID. The use of MARS may provide an opportunity to detect factors affecting milk production (breed, feed supply, type of milking, mastitis control, cow year group, and lactation length) and their interactions. Moreover, the MARS algorithm may be useful in making an accurate decision about increasing milk yield per cow.

Keywords: milk yield, production economics, statistical model

Introduction

A sufficient and balanced diet is very important for a healthy lifestyle to maintain growth and development and combat diseases. In this scenario, with its considerable vitamin C and iron contents, milk is an important source of macro- and micronutrients (Black et al., 2002). The animal production sector within which milk is produced is important for the national development, since it creates jobs in the rural sector and increases value-added per unit investment (Koseman and Seker, 2015).

Turkey holds 2.05% of the world cattle population (5,609,240 heads) and produces 2.59% of all cow milk (16,998,850 tons) in the world (FAO, 2016). The world average cow milk yield is 2,394 kg, while the average yield in Europe and Turkey is 5,834 and 3,030 kg, respectively. Although the average milk yield in Turkey is above the world average, it is far below that of Europe. To increase the average milk yield in Turkey, investigating the effects of factors such as breed, time of birth, lactation period, feed supply, mastitis control before milking, etc., is an important measure. The Eastern Anatolia region of Turkey has a great potential for dairy farming, where it has pasture and grassland area relative to other regions in the country; however, the milk yield of that region is lower than any other region of the country.

The economic analysis for determining factors affecting daily milk yield per cow was modeled by several algorithms such as the Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Network (ANN) in the literature (Dogan, 2003, Grzesiak et al., 2003; Mundan et al., 2006; Millogo et al., 2008; Mundan et al., 2009; Ozkan and Gunes, 2011; Aygul and Ozkutuk, 2012; Takma et al., 2012; Kulekci, 2013; Shahinfar et al., 2014; Yavuz and Kaygisiz, 2015; Cetin and Mikail, 2016). The present study aimed to identify factors affecting milk yield on dairy farms in Ardahan province of Eastern Anatolia region and compare these factors by using different data-mining and neural network algorithms. In the eastern region of Turkey, there is no study to compare these algorithms leading to advantages and disadvantages in terms of milk yield production. Thus, the region is untouched in terms of the subject and, therefore, the present study will make a contribution to the literature in this perspective.

Material and Methods

In the selection of the enterprises to be included in the sample under investigation, data of enterprises registered at the Farmer Recording System of the Ardahan Directorate of Provincial Food and Agriculture and Livestock in Ardahan province (41°06'47"N latitude, 42°49'15"E longitude, and 1807 m asl) of Eastern Anatolia Region, Turkey, were used. These enterprises owned a total of 656 head of cattle in 2014. The simple random sampling method, based on the number of animals registered in the farmer recording system, was used in the selection of enterprises to be included in the questionnaire study, as recommended by Karadas et al. (2017). The method used to determine the sample size is shown in Equation 1:

$$n = \frac{NS^2t^2}{(N-1)d^2 + S^2t^2} \quad (1)$$

in which n = sample size (sampled farms); N = population size (656 farms); S^2 = variance of the population (4.55^2); d^2 = acceptable error proportion (0.67); and t = critical value (1.65).

$$n = \frac{656 \times 4.55^2 \times 1.65^2}{(656-1)0.67^2 + 4.55^2 \times 1.65^2} = 105$$

The above sample size was completed by the questionnaire.

Definitions of the variables used in the study and their abbreviations are as follows:

Average milk yield per cow (AMYPC) is a continuous target variable (kg); breed (native, crossbred, and culture) is a nominal input variable; and year group (Y = 1, 2, 3, and 4) is an ordinal input variable (in which Y = 1 consisted of cows of two, three, and four years of age; Y = 2 consisted of cows of five, six, and seven years of age; Y = 3 consisted of cows of eight, nine, and ten years of age; and Y = 4 consisted of cows of 11 years or older).

Type of milking (TOM; finger milking, mechanical milking, and finger-mechanical milking) is a nominal input variable; birth month (BM; February (2), March (3), April (4), and May (5)) is a nominal input variable; mastitis control before milking (MASC; yes and no) is a binary input variable; other diseases (OD; anthrax, foot and mouth disease, etc. except for mastitis) is a nominal input variable; cost of veterinary and drugs (COVAD) is a continuous input variable (Turkish Lira); feed supply (FS) is a continuous (input) variable (kg); and lactation length (LL) is a continuous input variable.

In recent years, the CART, CHAID, and ANN algorithms have become increasingly popular in different animal science fields in the scope of predictive modeling for a continuous target variable. Among those, CART produces binary splitting recursively until homogenous subsets are derived from a learning sample data set. The CART algorithm specified for finding interactions for continuous, nominal, and ordinal variables works effectively even in situations where there are missing data on predictors (Kovalchuk et al., 2017). The algorithm procedure contains tree-building and pruning. All predictors are considered for the best split in the tree-building process. As a splitting criterion, the least squares deviation (LSD) method is

implemented for a continuous dependent variable, as it was also used in the present study. However, both CHAID algorithms that make Bonferroni adjustment produce multi-splitting nodes recursively for deriving a great number of homogenous subsets from the learning sample set. The General Linear Model (GLM), based on the least-squares method, has been used in numerous previous studies. Multilayer Perceptron (MLP), as an ANN type having three layers (input, hidden, and output) and resembling the human brain, was specified for training (80%) and testing (20%) sets with one hidden layer, hyperbolic tangent (activation function), and an identity activation function in respect of output layer.

MARS is a nonparametric regression technique using a divide and conquer strategy in which the training data sets are divided into separate piecewise linear segments (splines) of various gradients (slope). The splines are linked smoothly to each other and basis functions, as piecewise curves, identifying linear and non-linear effects. The connection points between the pieces are nominated "knots". The candidate knots were generally inserted at random locations within the range of each predictor. MARS enables one to obtain basis functions by considering all possible candidate knots and interactions among predictors through a stepwise procedure. In the description of a pair of the basis functions, the forward procedure sets up the candidate knots at random locations within the range of each predictor. At each stage, it stimulates the knots and their pairs of basis functions aiming at minimizing the differences between the predicted and actual values in a dependent variable. Until the complicated model is produced, the procedure of involving the basis functions persists. The redundant functions insignificantly contributing to the MARS are removed from the prediction equation by the backward procedure in the MARS. The MARS model can be expressed with the following equation:

$$\hat{y} = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{km}\left(X_{v(k,m)}\right) \quad (2)$$

in which $\hat{y}$ = predicted value of AMY; β_0 = a constant; $h_{km}\left(X_{v(k,m)}\right)$ = basis function, in which $v(k, m)$ is an index of the predictor employed in the m-th component of the k-th product; and K_m = parameter regulating the order of interaction.

The maximum number of basis functions in the current analysis was 100, and the four-order interactions were adopted for the predictive performance of the MARS algorithm. After constructing the most complex MARS model, the basis functions that did not contribute significantly to the quality of the model performance were eliminated in the process of the so-called pruning, based on the following generalized cross-validation error (GCV) (3):

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(y_i - y_{ip})^2}{\left[1 - \frac{M(\lambda)}{n}\right]^2}, \quad (3)$$

in which n = number of training cases; y_i = observed value of average milk yield per cow (AMYPC); y_{ip} = predicted value of AMYPC; and $M(\lambda)$ = penalty function for the complexity of the model including λ terms.

Some model evaluation criteria (R^2, $R^2_{ADJUSTED}$, SD_{RATIO}, CV (%), RMSE, RAE, MAPE, MAD, and Pearson correlation coefficient (r) between actual and predicted lactation milk yield values) were calculated to estimate the predictive performances of GLM, CART, CHAID, Exhaustive CHAID, MLP, and MARS. The following are model evaluation criteria to compare their predictive performances of the algorithms:

Pearson coefficient between actual and predicted values in a target variable,

Coefficient of determination:

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}\right] \quad (4)$$

Adjusted coefficient of determination:

$$R^2_{ADJUSTED} = \left[1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}\right] \quad (5)$$

Coefficient of variation (%):

$$CV\,(\%) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}}{\bar{Y}} * 100 \quad (6)$$

Standard deviation ratio:

$$SD_{RATIO} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} \quad (7)$$

Relative approximation error:

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}Y_i^2}} \quad (8)$$

Root mean-square error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \quad (9)$$

Mean absolute deviation:

$$MAD = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \quad (10)$$

Mean absolute percentage error:

$$MAPE = \frac{\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i}}{n} * 100 \quad Y_i \neq 0 \quad (11)$$

in which Y_i = actual or observed AMYPC (kg/day) of the *i*-th enterprise; $\hat{Y}_i$ = predicted AMYPC (kg/day) value of the *i*-th enterprise; $\bar{Y}$ = average of the actual AMYPC values of the enterprises; ε_i = residual value of the *i*-th enterprise; $\bar{\varepsilon}$ = average of the residual values; *k* = number of input variables in the model; and n = number of enterprises. The residual value of each enterprise is expressed as $\varepsilon_i = Y_i - \hat{Y}_i$.

However, *k* is the number of terms here for only the MARS model. The cross-validation value for CART and CHAID tree-based algorithms was set as 10 (Ali et al., 2015; Karadas et al., 2017; Kovalchuk et al., 2017). Minimum enterprise numbers for parent and child nodes were specified for producing the best predictive accuracy in CART (10:5) and both CHAID (16:8) algorithms (Eyduran et al., 2017) (See Tedeschi (2006) for the equations defined and discussed here).

Except for MARS, the other algorithms were analyzed using the IBM SPSS version 23 program. MARS was performed using the STATISTICA version 8 package program (trial version).

Results

The MARS algorithm showed a much better fit to the studied data and produced a recommendable solution when compared with other approaches (Table 1). The correlation coefficient between actual and predicted AMYPC values for the MARS algorithm (0.985) was significantly different from the correlation coefficients estimated for other approaches (P<0.01). The SD_{RATIO} estimate of the MARS algorithm was much lower when compared with the estimates for other approaches. Moreover, correlation coefficients of training and testing sets of MLP as a type of ANN were not statistically different from each other. Interaction effects were not included in GLM as a result of the great number of the subgroups obtained by main effects. The GLM reflected that BREED, BM, MASC, FS, and LL were significant predictors (P<0.01) (data not shown). The significance order for the BREED factor was Culture breed (6.393 kg/day) > Crossbred (5.75 kg/day) > Native breed (5.077 kg/day) (P<0.05). Enterprises with MASC had higher yields compared with those without it (6.166 vs. 5.314 kg/day).

Based on the regression tree drawn to reveal input variables affecting AMY per cow (Figure 1), Node 0 had 5.377 (kg/day) in AMYPC (at the top of the regression tree structure; Figure 1), and it was split by means of breed factor into three smaller subsets: Nodes 1-3 (Adj. P-value = 0.000, F = 137.397, df1 = 2, and df2 = 102). Node 1 represents enterprises rearing culture cattle breeds, while Node 2 was the subset of those rearing crossbred cattle, and the subsets of those rearing native cattle breeds was shown by Node 3. Unsurprisingly, the significance order in AMYPC was Culture cattle breed (8.459 kg/day) > Crossbred cattle (5.386 kg/day) > Native breed (3.725 kg/day). The AMYPC for the subset of culture cattle breeds was affected by the BM input variable (Adj. P-value = 0.002, F = 16.055, df1 = 1, and df2 = 15) (Figure 1). Thus, two small subsets, Nodes 4 and 5, were obtained by this separation. The enterprise subset of BM ≤ 4th month in culture cattle breed was represented by Node 4 (9.875 kg/day). On the other hand, the enterprise subset with BM later than the 4th month in same cattle breed group was represented by Node 5 (7.2 kg/day). We point out that the BM cut-off value produced by CHAID algorithm for culture cattle breeds reared in the province of Ardahan was determined to be the 4th month. The AMYPC of those rearing crossbred cattle among all the enterprises was significantly affected by MASC (Adj. P-value = 0.000, F = 51.020, df1 = 1, and df2 = 54).

Node 2 was partitioned by the MASC factor into two smaller subsets (Nodes 6 and 7). Node 6 (the subset of the enterprises performing mastitis control within enterprises having crossbred cattle) was higher in terms of AMYPC than Node 7 (the subset of the enterprises not performing mastitis control within enterprises having crossbred cattle) (6.215 vs. 5.135 kg/day) (Visual CHAID diagram in Figure 1). Node 6 was a terminal subset that could not be partitioned into smaller subsets at a subsequent stage, whereas Node 7 was divided into two smaller terminal subsets numbered 10 and 11 with respect to the LL input variable (4.950 vs. 5.673 kg/day: Adj. P-value = 0.000, F = 27.997, df = 1, and df2 = 41). Node 10 was the subset of enterprises rearing crossbred cattle that had LL ≤ 210 days with no mastitis

Table 1 - Results of goodness of fit criteria

Algorithm		r	R^2	$R^2_{ADJUSTED}$	RMSE	RAE	CV (%)	SD_{RATIO}	MAD	MAPE
CHAID		0.885c	0.783	0.774	0.840	0.148	15.69	0.466	0.508	8.77
Exhaustive CHAID		0.881c	0.776	0.769	0.853	0.151	15.94	0.473	0.548	9.96
CART		0.929bc	0.863	0.859	0.669	0.118	12.50	0.371	0.496	9.93
ANN	Training[x]	0.957b	0.916	0.906	0.615	0.106	11.26	0.330	0.441	8.35
	Testing[x]	0.945	0.893	0.829	0.441	0.088	9.46	0.295	0.382	9.12
GLM		0.944b	0.892	0.873	0.593	0.105	6.61	0.329	0.444	8.53
MARS		0.985a	0.970	0.962	0.313	0.055	5.85	0.174	0.231	4.62

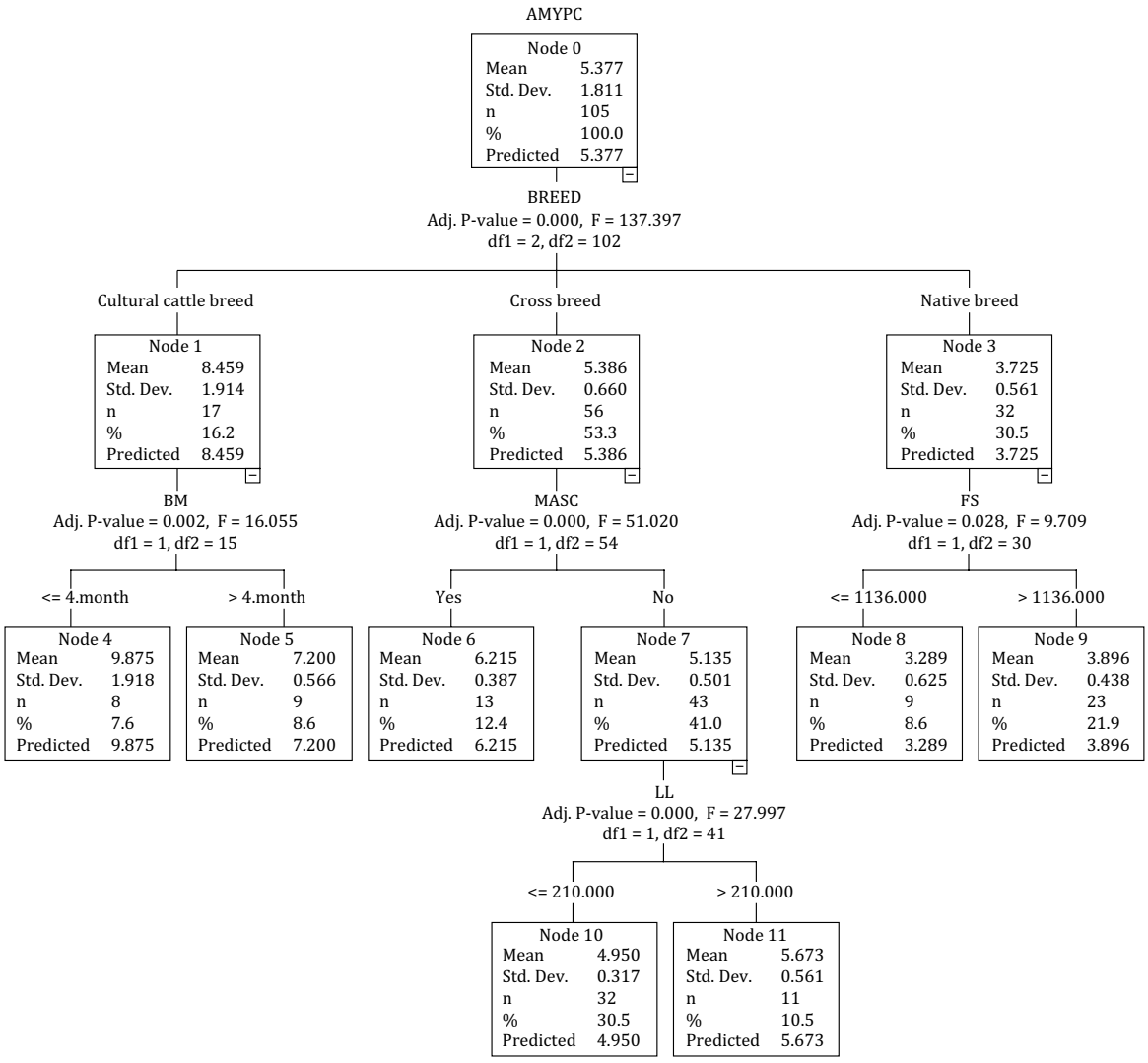
CART - Classification and Regression Tree; CHAID - Chi-Square Automatic Interaction Detector; MARS - Multivariate Adaptive Regression Splines; ANN - Artificial Neural Network; GLM - general linear model; r - Pearson correlation coefficient; R^2 - coefficient of determination; $R^2_{ADJUSTED}$ - Adjusted coefficient of determination; RMSE - root mean-square error; RAE - relative absolute error; CV - coefficient of variation; SD - standard deviation ratio; MAD - mean absolute deviation; MAPE - mean absolute percentage error.
x - The difference between training and testing sets was non-significant.
a,b - The difference between pairs of algorithms having different letters in correlation coefficient was significant.

control, while Node 11 was the subset of enterprises rearing crossbred cattle that had a LL longer than 210 days and performed mastitis control.

Node 3 represented the subset of the enterprises rearing native cattle breeds and was influenced and split by the FS input factor (Adj. P-value = 0.028, F = 9.709, df = 1, and df2 = 30). Thus, two smaller terminal subsets (Nodes 8 and 9) were obtained (3.289 vs. 3.896 (kg/day) in AMYPC). Node 8 was the subset of enterprises feeding native cattle breeds with a FS ≤ 1.136 kg. Node 9 was defined as the subset of enterprises feeding native cattle breeds with a FS > 1.136 kg (Figure 1). Node 8 gave the lowest AMYPC in the diagram of CHAID tree-based algorithm.

When Figures 1 and 2 were elaborated together, the structures of the CHAID and Exhaustive CHAID (Figure 2) algorithms were found to be almost similar; that is, Nodes 8 and 9 were available in the CHAID algorithm, but not in the Exhaustive CHAID algorithm. For both CHAID algorithms, the highest AMYPC was produced by Node 4.

The regression tree constructed by CART algorithm in the description of input variables affecting AMYPC (Figure 3) shows that the AMYPC of 5.377 (kg/day) was obtained by Node 0 found at the top of the regression-tree structure. Node 0 was partitioned by the breed factor into two smaller subsets labeled Node 1 (the subset of the enterprises rearing only culture cattle breeds) and Node 2 (the subset of the enterprises rearing crossbred and native cattle breeds) (8.459 vs. 4.782 kg/day). Node 1



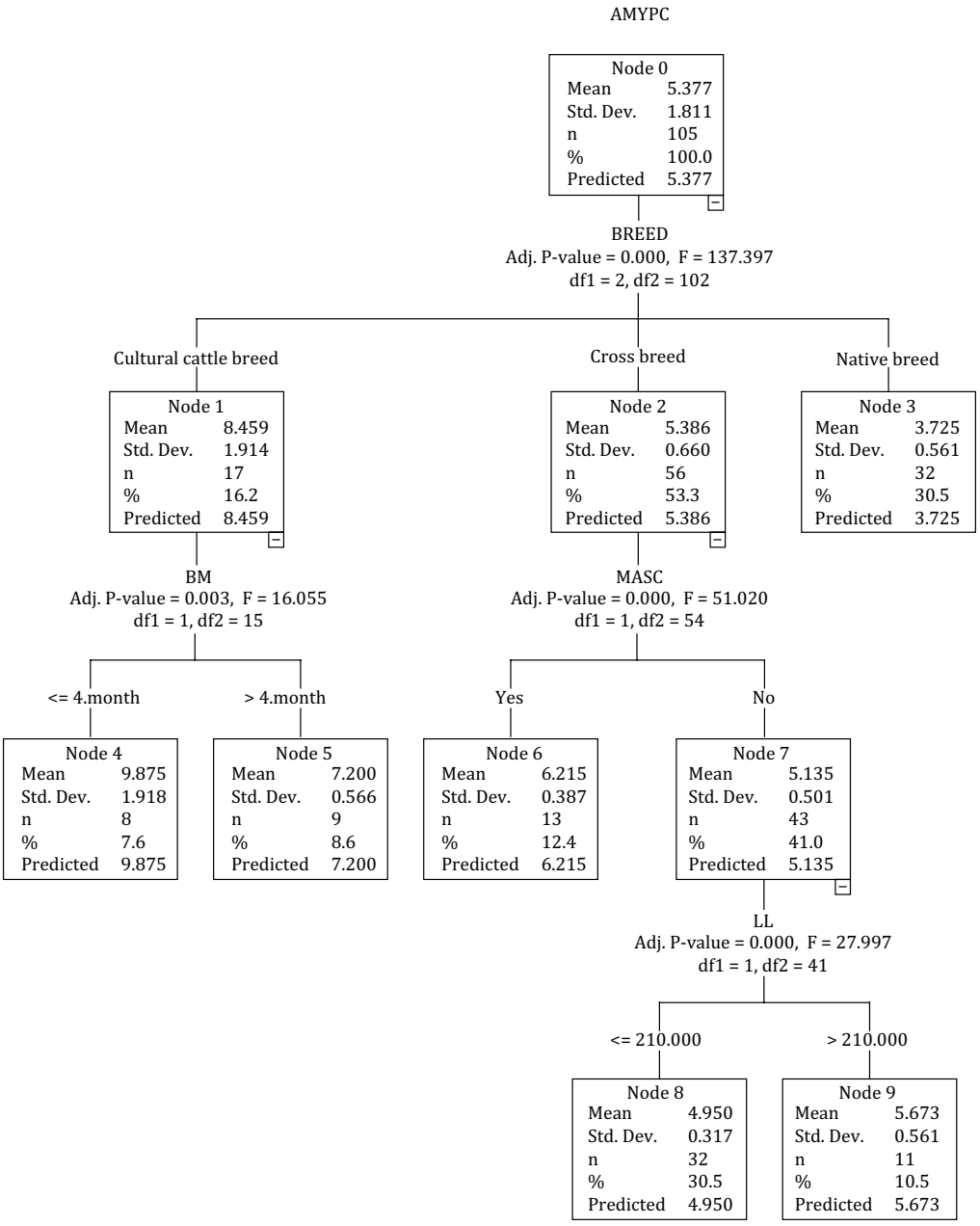
AMYPC - average milk yield per cow; BM - birth month; MASC - mastitis control before milking; FS - feed supply; LL - lactation length.

Figure 1 - Decision-tree diagram constructed by the CHAID algorithm.

is comprised of two smaller subsets (Nodes 3 and 4) as a result of variability in the LL input variable (7.550 vs. 10.640 kg/day). The subset of enterprises rearing culture cattle breeds with LL ≤ 237.5 days was assigned to Node 3. The subset of enterprises rearing culture cattle breeds with LL > 237.5 days was described as Node 4. A cut-off value of 237.500 should be noted for the LL input variable in the subset of the enterprises rearing culture cattle breeds.

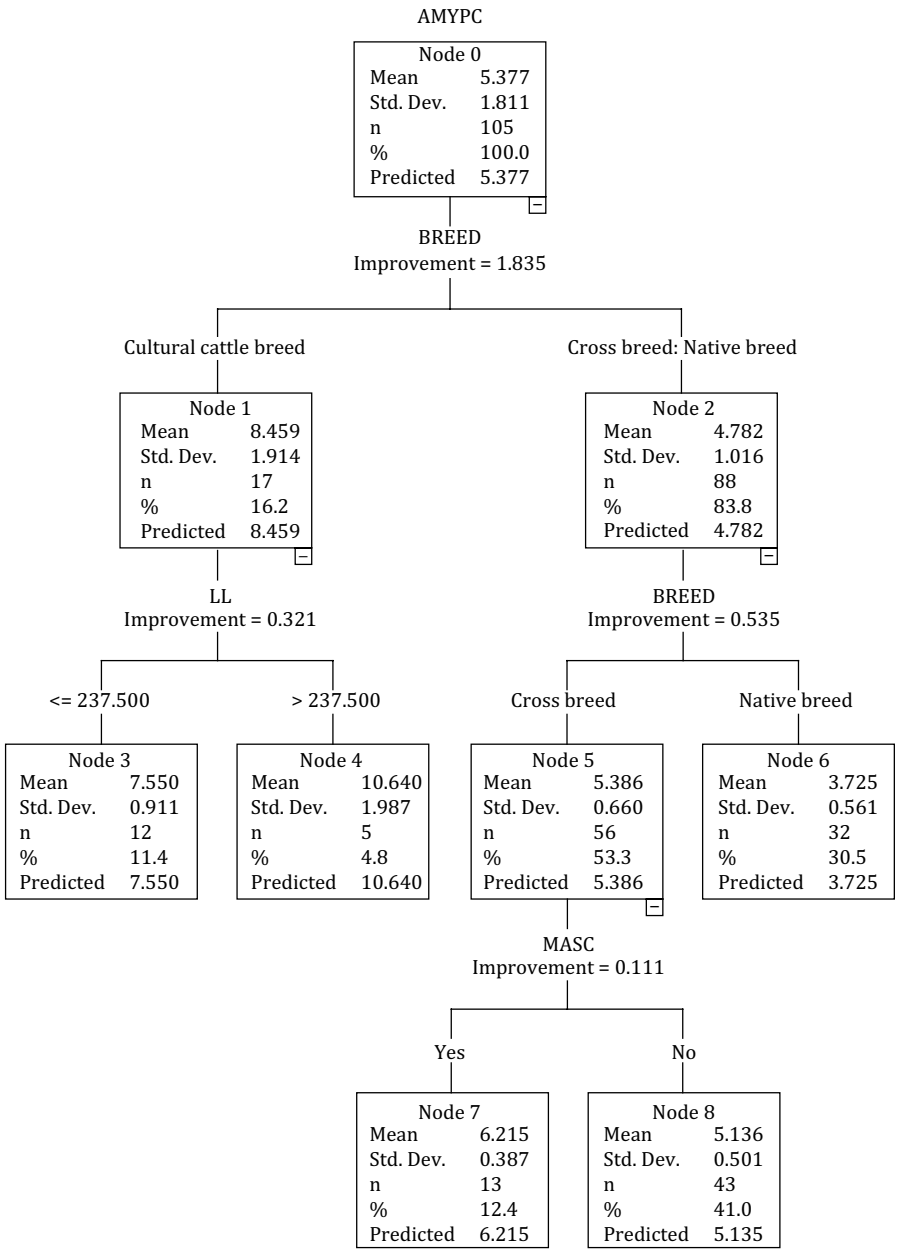
Node 2 was exposed to a new separation that produced Node 5 (the subset of enterprises rearing crossbred cattle) and Node 6 (the subset of enterprises rearing crossbred and native cattle breeds) in the CART tree-based diagram (5.386 vs. 3.725 kg/day).

The subset of enterprises rearing only crossbred cattle and performing mastitis control was Node 7, which was found to be higher in terms of AMYPC when compared with Node 8, the subset of enterprises rearing only crossbred cattle and not performing mastitis control (6.215 vs. 5.135 kg/day). The highest AMYPC was taken by Node 4 in the CART diagram, but the lowest was obtained by Node 6.



AMYPC - average milk yield per cow; BM - birth month; MASC - mastitis control before milking; LL - lactation length.

Figure 2 - Decision-tree diagram constructed by the Exhaustive CHAID algorithm.



AMYPC - average milk yield per cow; MASC - mastitis control before milking; LL - lactation length.

Figure 3 - Decision-tree diagram constructed by the CART algorithm.

The prediction equation obtained by the MARS data-mining algorithm can be written as follows:

AMYPC = 6.30032 – 2.05443*max(0; BREED_1) – 1.22326*max(0; BREED_2)*max(0; MASC_2) – 0.00063*max(0; 1556-FS) – 0.00211*max(1556-FS)*max(0; Y_4)+ 1.97081*max(0; BREED_2)*max(0; Y_4) + 0.24472*max(0; LL-215)*max(0; TOM_2) – 0.00014*max(0; FS-1566)*max(0; LL-215)*max(0; TOM_2) + 0.17223*max(0; LL-215)*max(0; BM_4)*max(0; MASC_2) – 0.01062*max(0; 215-LL)*max(0; BM_4) + 0.51734*max(0; LL-215)*max(0; Y_1) – 0.22682*max(0; LL-215)*max(0; TOM_2)*max(0; BM_4) – 0.00027*max(0; FS-130)*max(0; LL-215)*max(0; MASC_2) – 0.00017*max(0; 2314-FS)*max(0; LL-215)*max(0; BM_3) – 0.11400*max(0; LL-215)*max(0; TOM_2)*max(0; BM_3) + 0.37889*max(0; LL-215)*max(0; BREED_2)*max(0; MASC_2) + 0.00085*max(0; FS-1556)*max(0; BM_2) – 0.000171*max(0; FS-1556)*max(0; TOM_2)*max(0; BM_2) – 0.00268*max(0; FS-1556)*max(0; AG_2) + 0.00006 *max(0; FS-1556)*max(0; LL-160)*max(0; Y_2) + 0.00127*max(0; FS-2477)*max(0; BREED_2)*max(0; MASC_2) – 0.02327*max(0; 188-LL)*max(0; BREED_1)*max(0; TOM_1)

For example, if a prediction equation is desired for BREED = 1, FS = 1556, LL = 215, Y = 4, and BM = 4, the above equation turns into AMYPC = 6.30032 – 2.05443*max(0; BREED_1) = 4.246 kg, in which max(0; BREED_1) = 1, otherwise 0. If a prediction equation is desired for BREED = 2, FS = 1800, LL = 255, Y = 3, BM = 2, MASC = 1, and TOM = 1; however, the prediction equation turns into AMYPC = 6.30032 + 0.00085*max(0; FS-1556)*max(0; BM_2) = 6.50823 kg, in which max(0; FS-1556) = max(0;1800-1556) = 244 if FS = 1800 and max(0; BM_2) = 1 if BM = 2, otherwise 0.

Discussion

The present subset AMYPC values of the CHAID data-mining algorithm (Figure 1) were lower than those recorded by Turki et al. (2012), who reported that the average daily milk yield was affected by the type of feed in native (7.70 to 12.90 kg/day) and crossbred (9.20 to 17.5 kg/day) dairy cows in Sudan and emphasized the effects of the type and amount of feed given per cow in terms of milk components. Gunduz and Dagdeviren (2011) reported that the cost of milk production was the most important factor, and intake of concentrate feed significantly affected milk production. In our study, the amount of concentrate feed consumed had a significant impact on native breed cattle only (Figure 1). This finding was in line with the results of Kulekci (2013). However, Kulekci (2013) found an AMYPC (6.31 kg) lower than those reported in the present study but did not distinguish between cow breeds. As it is well-known, the breed is an important source of variation in milk yield.

The current milk yield averages of subgroups obtained in the Exhaustive CHAID algorithm were lower than those found by Turki et al. (2012), who reported that AMYPC was affected by the type of feed in native (7.70 to 12.90 kg/day) and crossbred (9.20 to 17.5 kg/day) dairy cows reared in Sudan. Kulekci (2013) found that there were significant factors (veterinary medication costs, concentrate feed costs, other costs, etc.) influencing AMYPC (kg/day) in dairy cattle production in Erzurum province of Turkey, a result which was not consistent with the results of the CHAID and Exhaustive CHAID algorithms (Figures 1 and 2, respectively), which is stemming from different factors included in the model and from the use of various statistical approaches. Aytekin et al. (2016) stated that LL and AMYPC are much better predictors of lactation milk yield in the first lactation of Black-White cows. In our study, LL affected only crossbred cows by the CHAID and Exhaustive CHAID algorithms (Figures 1 and 2, respectively); however, it influenced only culture breed cows by the CART algorithm (Figure 3). This may be attributed to various statistical approaches used. Nath et al. (2016) recorded AMYPC values of 7.73 L for the Jersey cross, 12.9 L for the Holstein-Fresian cross, 5.51 L for the Sahiwal cross, and 4.1 L for the Red Sindhi cross breeds in Bangladesh, and these results were in agreement with the findings of this study.

When comparing the six algorithms among themselves, the MARS algorithm seems more powerful to predict the AMYPC dependent variable in terms of statistics conducted (Table 1). Type of milking was an effective factor in the MARS algorithm, in contrast to the CART and both CHAID algorithms, which means that significant predictors in these three algorithms masked the effect of TOM on AMYPC. The present finding echoed with findings of Kulekci (2013), whilst Kulekci (2013) found COVAD to be a significant predictor of AMYPC.

A more convenient and understandable prediction equation can be produced for the MARS algorithm. MARS might provide enterprises with helpful information in organizing optimum management conditions in practice for enhancing AMYPC.

However, the applications of data-mining algorithms used in the present study are very limited in the literature. In this respect, the present results could not be decisively compared with those reported in previous studies.

Conclusions

Breed is the most significant factor in milk yield according to all constructed decision-tree diagrams, followed by birth month, mastitis control before milking, feed supply, and lactation length. Among the

applied algorithms, the Multivariate Adaptive Regression Splines (MARS) outperformed the others. Thus, the application of that algorithm is advised for deciding significant predictors affecting the average milk yield per cow. However, our study lacks for not including all regions in the study. The feature study, thus, may cover the whole regions in the country to perform the analysis for better prediction of the milk yield in the country.

Acknowledgments

The authors acknowledge the scientific support from Assoc. Prof. Dr. Ecevit Eyduran in statistical analyses and express their sincere gratitude for his help.

References

Ali, M.; Eyduran, E.; Tariq, M. M.; Tirink, C.; Abbas, F.; Bajwa, M. A.; Baloch, M. H.; Nizamani, A. H.; Waheed, A.; Awan, M. A.; Shah, S. H.; Ahmad, Z. and Jan, S. 2015. Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai sheep. Pakistan Journal of Zoology 47:1579-1585.

Aygul, H. and Ozkutuk, K. 2012. The structure of dairy cattle and fattening enterprises in Malatya AVKAE Dergisi 2:7-11.

Aytekin, I.; Mammadova, M. M.; Altay, Y.; Topuz, D. and Keskin, I. 2016. Determination of the factors effecting lactation milk yield of Holstein Friesian Cows by the path analysis. Selcuk Journal Agriculture Food Science 30:44-48.

Black, R. E.; Williams, S. M.; Jones, I. E. and Goulding, A. 2002. Children who avoid drinking cow milk have low dietary calcium intakes and poor bone health. American Journal of Clinical Nutrition 76:675-680. https://doi.org/10.1093/ajcn/76.3.675

Cetin, F. A. and Mikail, N. 2016. Data mining aplications in livestock. Turk Journal Agriculture Research 3:79-88.

Dogan, I. 2003. Investigation of the factors which are affecting the milk yield in Holstein by CHAID analysis. Ankara Universitesi Veteriner Fakultesi Dergisi 50:65-70.

Eyduran, E.; Zaborski, D.; Waheed, A.; Celik, S.; Karadas, K. and Grzesiak, W. 2017. Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the Indigenous Beetal Goat of Pakistan. Pakistan Journal Zoology 49:257-265.

FAO - Food and Agriculture Organization of the United Nations. 2016. FAOSTAT. Available at: <http://www.fao.org/faostat/en/#data/QL>. Accessed on: Dec. 26, 2016.

Grzesiak, W.; Lacroix, R.; Wójcik, J. and Blaszczyk, P. 2003. A comparison of neural network and multiple regression predicitions for 305-day lactation yield using partial lactation records. Canadian Journal of Animal Science 83:307-310. https://doi.org/10.4141/A02-002

Gunduz, O. and Dagdeviren, M. 2011. Bafra ilcesinde sut maliyetinin belirlenmesi ve uretimi etkileyen faktorlerin fonksiyonel analizi. YYU J Agriculture Science 21:104-111.

Karadas, K.; Tariq, M.; Tariq, M. M. and Eyduran, E. 2017. Measuring predictive performance of data mining and artificial neural network algorithms for predicting lactation milk yield in Indigenous Akkaraman Sheep. Pakistan Journal Zoology 49:1-7.

Koseman, A. and Seker, I. 2015. Current status of cattle, sheep and goat breeding in Turkey Van Veterinary Journal 26:111-117.

Kovalchuk, I. Y.; Mukhitdinova, Z.; Turdiyev, T.; Gulnara Madiyeva, G.; Akin, M.; Eyduran, E. and Reed, B. M. 2017. Modeling some mineral nutrient requirements for micropropagated wild apricot shoot cultures. Plant Cell, Tissue and Organ Culture 129:325-335. https://doi.org/10.1007/s11240-017-1180-0

Kulekci, M. 2013. Efficiency analysis of dairy cattle farms: Case study in Erzurum. Ataturk Univ Journal of the Agricultural Faculty 44:103-109.

Millogo, V.; Ouédraogo, G. A.; Agenäs, S. and Svennersten-Sjaunja, K. 2008. Survey on dairy cattle milk production and milk quality problems in Peri-Urban areas in Burkina Faso. African Journal of Agricultural Research 3:215-224.

Mundan, D.; Yerturk, M.; Avci, M.; Karabulut, O. and Bozkaya, F. 2006. Siyah alaca ineklerde laktasyon veriminin hesaplanmasinda kullanilan farkli yontemler ve kontrol periyotlarinin karsilastirilmasi. F.U. Saglik Bil Dergisi 20:173-177.

Mundan, D.; Karabulut, O. and Sehar, O. 2009. Holstayn ineklerde laktasyon sut verimini tahmin eden en iyi dogrusal regresyon modelinin belirlenmesi. Firat Universitesi Saglik Bilimleri Veteriner Dergisi 23:129-134.

Nath, S. K.; Bhowmik D. K.; Rokonuzzaman, M.; Afrin, K.; Dash, A. K. and Alam, R. 2016. Production performance of different cross breeds of milch cow in Mithapukur Upazila, Rangpur, Bangladesh. International Journal of Advanced Multidisciplinary Research 3:29-33.

Ozkan, M. and Gunes, H. 2011. Effects of some factors on milk yield characteristics of Sıimmental cows on commercial farms in Kayseri. Journal of the Faculty of Veterinary Medicine, Istanbul University 37:81-88.

Shahinfar, S.; Page, D.; Guenther, J.; Cabrera, V.; Fricke, P. and Weigel, K. 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algoritms. Journal of Dairy Science 97:731-742. https://doi.org/10.3168/jds.2013-6693

Takma, C.; Atil, H. and Aksakal, V. 2012. Coklu dogrusal regresyon ve yapay sinir agi modellerinin laktasyon sut verimine uyum yeteneklerinin karsilastirilmasi. Kafkas Universitesi Veteriner Fakultesi Dergisi 18:941-944. https://doi.org/10.9775/kvfd.2012.6764

Tedeschi, L. O. 2006. Assessment of the adequacy of mathematical models. Agricultural Systems 89:225-247. https://doi.org/10.1016/j.agsy.2005.11.004

Turki, Y. I.; Maehip, A. M.; Muna, E. K.; Miriam, S. A.; Omer, M. E. and Hammed, M. E. 2012. Effect of feeding systems on milk yield and composition of local and cross breed dairy cows. International Journal of Science and Technology 2:5-9.

Yavuz, S. and Kaygisiz, A. 2015. The relationship between some body and udder measurements with somatic cell count in Holstein cows. KSU J Nat Science 18:9-18.