REGRESSION TREE ANALYSIS OF FACTORS AFFECTING FIRST LACTATION MILK YIELD OF DAIRY COWS

MIKAIL, N.1* – BAKIR, G.2

1Department of Animal Science, Faculty of Agriculture, Siirt University, 56100 Siirt, Turkey
2Department of Animal Science, Faculty of Agriculture, Sütçü İmam University, 46100, Kahramanmaraş, Turkey


*Corresponding author

e-mail: naziremikail@siirt.edu.tr; phone: +90-484-212-1111/2818; fax: +90-484-212-1112

(Received 28th Jan 2019; accepted 6th Mar 2019)

Abstract. The objective of this study was to examine the effects of management and environmental factors on first lactation milk yield by means of regression tree method. Regression tree method is useful to determine the effects of several factors on specified depended variables. The independent variables used in the model were factors such as farm, breed, year of calving, season of calving, age at calving and days in milk. Data used in this study were obtained from three state farms. The resulting data set consisted of 754 records from 1st farm, 1120 records from 2nd farm and 324 records from 3rd farm. The average total milk yield and days in milk were 5413.11 ± 2033.18 kg and 324 ± 64 days, respectively. Year of calving and days in milk were important variables affecting first lactation milk yield of dairy cattle, followed by farm and breed (p < 0.01). The determination coefficient of the prediction was found as 78.9%. As a result of the study 14 distinct paths from the root node to the leaves were presented.

Keywords: multiple linear regression, farm, breed, year of calving, days in milk, cow

Introduction

First lactation milk yield is a very important characteristic in all dairy cattle. Factors affecting this trait can be divided into genetical (breed) and environmental, such as farm, year of calving, season of calving, age at calving and days in milk. Understanding the factors that change the environment of the dairy cattle can be used to take advantage of some improvement in milk yield that occur in normal lactation. During first lactation at an average age of 2.5 years cow produces approximately 76% of the milk produced by a mature cow (Nirish, 2010). Milk production is usually less during the summer because of the higher environmental temperatures and the prevalence of green-forage scarcity. Thus the season of calving has got a marked effect on the total production. Cows freshening shortly before winter months produce more total yield than those calving at other times of the year. The increase is probably due to more favorable temperature and more digestible feeds available during the winter. The farmer has no effect on the physiological factors of the cow, but has some effects on environmental factors affecting the cow. Factors that alter the environment of dairy cattle can be used to take advantage of some changes in milk composition and the yield during the lactation (Irshad, 2015).

An alternative approach to nonlinear regression is to sub-divide the space into smaller regions, where the interactions are more manageable. The general model consists of two parts: one is only a recursive section, and the other is a simple model for each cell of the section. The regression trees use the tree to represent the recursive part.
Each terminal node or leaf of the tree represents a cell of the section and has just added a simple pattern applied to it in this cell. This is due to CART’s desirable properties as, automatic handling of the variable selection, variable interaction modeling, local effect modeling, nonlinear relationship modeling. CART also is more robust in the presence of outliers and not affected by monotonic transformations of variables. Not all variables have to be in the same type; some of them can be continuous, some can be discrete, etc. Regression trees have several advantages: there are no complicated calculations and hence the estimation is fast: it is easy to understand what variables are important when estimating. In practice, it is possible to build CART models with dirty data (i.e. missing values, lots of variables, nonlinear relationships, outliers, and numerous local effects).

Many studies have been carried out in the application of regression trees in animal husbandry. Lots of them concerning determination of factors effecting milk yield (Mirtagioglu et al., 2008; Bakır et al., 2010; Topal et al., 2010; Cak et al., 2013; Eyduran et al., 2013) and body weight (Topal et al., 2010) of animals. But also there are some researches concerning the application of regression trees to the different data from animal science (Eyduran et al., 2008, 2016; Koç et al., 2017; Takma et al., 2017)

In the present study we are interested in examining the management and environmental effects affecting cows at first lactation by means of regression tree analysis.

Materials and methods

Materials

Data on total lactation milk production were obtained from Tahirowa, Konuklar and Malya farms. These farms are the state enterprises belonged to General Directorate of Agricultural Enterprises (Fig. 1). Konuklar Farm is located in the 57 km from Konya. The climate is a typical continental climate of Central Anatolia; summers are hot and dry, winters are cold and rainy (GDA, 2016). The average annual rainfall is 322.4 mm, and the altitude is 1050 m. The average annual temperature is 11.6 °C (GDM, 2017). Malya Farm is located 27 km north east of Kırşehir. This farm is located in the Central Anatolia Region, at an altitude of 985 m. Again, the climate is the typical continental climate of Central Anatolia, summers are hot and dry and winters are cold and rainy (GDA, 2016). The average annual rainfall is 378.4 mm. The average annual temperature is 11.4 °C (GDM, 2017). Tahirowa Farm is located in Balıkesir province; the farm is situated at an altitude of 166 m. Summers are hot, winters are rainy. The average annual rainfall is 583.7 mm. The average annual temperature is 14.6 °C (GDM, 2017).

At all three farms the cows were grazed and winter supplement was also provided. In the current paper, 2198 first lactation records were assessed for 1874 Brown-Swiss dairy cattle reared at Konuklar and Malya State Farms, and 324 Holstein dairy cattle reared at Tahirowa State Farm. Records covered the period of calving from 1987 to 2007.

Methods

Regression tree method

Multiple linear regression (Eq. 1) is a way of making quantitative predictions with multiple independent variables $X_i \equiv \{X_1, X_2, \ldots X_n\}$
\[ Y = \beta_0 + \beta^T X + \epsilon, \]  
(Eq.1)

Independent variables have an additive effect on \( Y \), also an interaction \( (Eq. 2) \),
\[ Y' = \beta_0 + \beta^T X + \gamma X X^T + \epsilon \]  
(Eq.2)

Prediction trees use the tree to represent the recursive partition. Each of the terminal nodes, or leaves, of the tree represents a cell of the partition, and has attached to it a simple model which applies in that cell only \( (Fig. 2) \).

\[ Y = \beta_0 + \beta^T X + \epsilon, \]  
(Eq.1)

Independent variables have an additive effect on \( Y \), also an interaction \( (Eq. 2) \),
\[ Y' = \beta_0 + \beta^T X + \gamma X X^T + \epsilon \]  
(Eq.2)

Classification and Regression Tree splits the data into segments that are as homogeneous as possible with respect to the dependent variable \( (Fig. 3) \) \( (Lahmann and Kottner, 2011) \).

A regression tree can be seen as a kind of additive model \( (Hastie and Tibshirani, 1990) \) of the form:

\[ Y = \beta_0 + \beta^T X + \epsilon, \]  
(Eq.1)

Independent variables have an additive effect on \( Y \), also an interaction \( (Eq. 2) \),
\[ Y' = \beta_0 + \beta^T X + \gamma X X^T + \epsilon \]  
(Eq.2)

Classification and Regression Tree splits the data into segments that are as homogeneous as possible with respect to the dependent variable \( (Fig. 3) \) \( (Lahmann and Kottner, 2011) \).

A regression tree can be seen as a kind of additive model \( (Hastie and Tibshirani, 1990) \) of the form:
where \( k_i \) are constants; \( I(.) \) is an indicator function returning 1 if its argument is true and 0 otherwise; \( D_i \) are disjoint partitions of the training data \( D \) such that \( \bigcup_{i=1}^{n} D_i = D \) and \( \bigcap_{i=1}^{n} D_i = \emptyset \).

A terminal node in which all cases have the same value for the dependent variable is a homogeneous, “pure” node. Regression trees (RT) is a method, where the target variable is continuous and tree is used to predict its value.

\begin{equation}
m(x) = \sum_{i=1}^{n} k_i \times I(x \in D_i) \tag{Eq.3}
\end{equation}

**Figure 3.** Classification (left) and regression tree structure (right) for a model with 2 classes

Each path from the root of the tree to a leaf corresponds to a region. Each inner node of the tree is a logical test on a predictor variable. In the particular case of binary trees, there are two possible outcomes of the test, true or false. This means that associated to each partition \( D_i \) we have a path \( p_i \) consisting of a conjunction of a logical tests on the predictor variables. This symbolic representation of the regression function is an important issue when one wants to have a better understanding of the regression surface (Soman et al., 2006).

For all analysis handled in this study IBM SPSS Statistics version 20 was used.

**Goodness of fit**

The coefficient of determination (Eq. 4) was used for determination of goodness of fit of the regression tree model.

\begin{equation}
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{Eq.4}
\end{equation}

**Results and discussion**

There are 2 continuous (days in milk (DIM), calving age (CA) and 4 categorical (farm, breed, season of calving and year of calving) variables used to predict total milk yield (TMY). Table 1 shows descriptive statistics of continuous variables. Table 2 represents the frequency and percentage of categorical variables.

**Table 1.** Descriptive statistics for the continuous variables

<table>
<thead>
<tr>
<th>Traits</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total milk yield (kg)</td>
<td>2198</td>
<td>5413.11</td>
<td>2033.178</td>
<td>1098</td>
<td>12032</td>
<td>37.56%</td>
</tr>
<tr>
<td>Days in milk (days)</td>
<td>2198</td>
<td>324.19</td>
<td>63.899</td>
<td>137</td>
<td>675</td>
<td>19.7%</td>
</tr>
<tr>
<td>Calving age (days)</td>
<td>2198</td>
<td>858.46</td>
<td>90.160</td>
<td>702</td>
<td>1194</td>
<td>10.5%</td>
</tr>
</tbody>
</table>
Table 1 shows that the most variation is in the total milk yield. At the same time, the lowest number of days in milk is 137 days and the lowest calving age is 702 days.

Table 2. Frequency and percentage values for several categorical variables

<table>
<thead>
<tr>
<th>Farm</th>
<th>Frequency</th>
<th>Rate</th>
<th>Season</th>
<th>Frequency</th>
<th>Rate</th>
<th>Breed</th>
<th>Frequency</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konuklar</td>
<td>754</td>
<td>34.3%</td>
<td>Winter</td>
<td>522</td>
<td>23.7%</td>
<td>BS</td>
<td>1874</td>
<td>85.3%</td>
</tr>
<tr>
<td>Malya</td>
<td>1120</td>
<td>51.0%</td>
<td>Spring</td>
<td>656</td>
<td>29.7%</td>
<td>HO</td>
<td>324</td>
<td>14.7%</td>
</tr>
<tr>
<td>Tahirova</td>
<td>324</td>
<td>14.7%</td>
<td>Summer</td>
<td>559</td>
<td>25.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Autumn</td>
<td>461</td>
<td>21.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 2. Frequency and percentage values for several categorical variables

<table>
<thead>
<tr>
<th>YoC</th>
<th>Frequency</th>
<th>Rate</th>
<th>YoC</th>
<th>Frequency</th>
<th>Rate</th>
<th>YoC</th>
<th>Frequency</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>27</td>
<td>1.2%</td>
<td>1995</td>
<td>101</td>
<td>4.6%</td>
<td>2003</td>
<td>162</td>
<td>7.4%</td>
</tr>
<tr>
<td>1988</td>
<td>63</td>
<td>2.9%</td>
<td>1996</td>
<td>104</td>
<td>4.7%</td>
<td>2004</td>
<td>98</td>
<td>4.5%</td>
</tr>
<tr>
<td>1989</td>
<td>52</td>
<td>2.4%</td>
<td>1997</td>
<td>131</td>
<td>6.0%</td>
<td>2005</td>
<td>231</td>
<td>10.5%</td>
</tr>
<tr>
<td>1990</td>
<td>55</td>
<td>2.5%</td>
<td>1998</td>
<td>130</td>
<td>5.9%</td>
<td>2006</td>
<td>134</td>
<td>6.1%</td>
</tr>
<tr>
<td>1991</td>
<td>87</td>
<td>4.0%</td>
<td>1999</td>
<td>141</td>
<td>6.4%</td>
<td>2007</td>
<td>49</td>
<td>2.2%</td>
</tr>
<tr>
<td>1992</td>
<td>84</td>
<td>3.8%</td>
<td>2000</td>
<td>127</td>
<td>5.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>72</td>
<td>3.3%</td>
<td>2001</td>
<td>134</td>
<td>6.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>75</td>
<td>3.4%</td>
<td>2002</td>
<td>141</td>
<td>6.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
<td>Total</td>
<td>2198</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Figure 4 shows the average TMY of different farms according to the calving year.

Figure 4. The average TMY of different farms according to the calving year

The CART algorithm is a process structured as a series of questions that determine what the next problem will be. The result of these questions is a tree-like structure with
endpoint nodes. On each intermediate node, a state goes to the lower left node if the condition is met (Loh, 2011).

The regression tree diagram is depicted in Figure 5.

Figure 5. Regression tree diagram for total milk yield

The node definitions and splitting values for TMY were given in Table 3.

Table 3. Node definitions and splitting values for TMY predictions

<table>
<thead>
<tr>
<th>Node</th>
<th>Predicted mean</th>
<th>SD</th>
<th>N</th>
<th>Percent</th>
<th>Parent node</th>
<th>Independent variable</th>
<th>F value</th>
<th>df1</th>
<th>df2</th>
<th>Split value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5413.11</td>
<td>2033.178</td>
<td>2198</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3238.79</td>
<td>1130.830</td>
<td>440</td>
<td>20.0</td>
<td>0</td>
<td>YoC</td>
<td>55.548**</td>
<td>21</td>
<td>2176</td>
<td>&lt;= 1994</td>
</tr>
<tr>
<td>2</td>
<td>5957.31</td>
<td>1835.555</td>
<td>1758</td>
<td>80.0</td>
<td>0</td>
<td>YoC</td>
<td>55.548**</td>
<td>21</td>
<td>2176</td>
<td>&gt; 1994</td>
</tr>
<tr>
<td>3</td>
<td>2693.45</td>
<td>837.456</td>
<td>223</td>
<td>10.1</td>
<td>1</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&lt;= 307.5</td>
</tr>
<tr>
<td>4</td>
<td>3799.21</td>
<td>1120.546</td>
<td>217</td>
<td>9.9</td>
<td>1</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&gt; 307.5</td>
</tr>
<tr>
<td>5</td>
<td>5172.33</td>
<td>1437.772</td>
<td>1023</td>
<td>46.5</td>
<td>2</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&lt;= 322.5</td>
</tr>
<tr>
<td>6</td>
<td>7049.87</td>
<td>1769.945</td>
<td>735</td>
<td>33.4</td>
<td>2</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&gt; 322.5</td>
</tr>
<tr>
<td>7</td>
<td>2246.47</td>
<td>626.151</td>
<td>76</td>
<td>3.5</td>
<td>3</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&lt;= 271.5</td>
</tr>
<tr>
<td>8</td>
<td>2924.54</td>
<td>840.975</td>
<td>147</td>
<td>6.7</td>
<td>3</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&gt; 271.5</td>
</tr>
<tr>
<td>9</td>
<td>4637.98</td>
<td>1337.806</td>
<td>475</td>
<td>21.6</td>
<td>5</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&lt;= 285.5</td>
</tr>
<tr>
<td>10</td>
<td>5635.50</td>
<td>1359.610</td>
<td>548</td>
<td>24.9</td>
<td>5</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&gt; 285.5</td>
</tr>
<tr>
<td>11</td>
<td>8391.31</td>
<td>1752.235</td>
<td>182</td>
<td>8.3</td>
<td>6</td>
<td>Farm</td>
<td>269.741**</td>
<td>2</td>
<td>2195</td>
<td>Tahirova</td>
</tr>
<tr>
<td>12</td>
<td>6608.38</td>
<td>1539.568</td>
<td>553</td>
<td>25.2</td>
<td>6</td>
<td>Farm</td>
<td>269.741**</td>
<td>2</td>
<td>2195</td>
<td>Malya; Konuklar</td>
</tr>
<tr>
<td>13</td>
<td>3427.19</td>
<td>1333.448</td>
<td>54</td>
<td>2.5</td>
<td>9</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&lt;= 241.5</td>
</tr>
<tr>
<td>14</td>
<td>4793.29</td>
<td>1258.089</td>
<td>421</td>
<td>19.2</td>
<td>9</td>
<td>DiM</td>
<td>3.897**</td>
<td>323</td>
<td>1874</td>
<td>&gt; 241.5</td>
</tr>
<tr>
<td>15</td>
<td>6883.35</td>
<td>1578.813</td>
<td>81</td>
<td>3.7</td>
<td>10</td>
<td>Farm</td>
<td>269.741**</td>
<td>2</td>
<td>2195</td>
<td>Tahirova</td>
</tr>
<tr>
<td>16</td>
<td>5419.06</td>
<td>1193.451</td>
<td>467</td>
<td>21.2</td>
<td>10</td>
<td>Farm</td>
<td>269.741**</td>
<td>2</td>
<td>2195</td>
<td>Malya; Konuklar</td>
</tr>
</tbody>
</table>
At the top of regression tree diagram, Node 0, which gave general descriptive statistics of TMY, was divided into new two child nodes, with respect to YoC factor. The TMY averages of the cows were born in 1994 and early on (Node 1) was estimated as 3238.79 kg and the TMY average of the other cows (Node 2) was estimated as 5957.31 kg. Right side nodes always show high level mean. With an average of 5957.31 kg, Node 2 gave the highest TMY. Numbers (proportions) of lactation records were 440 (20%) for Node 1 and 1758 (80%) for Node 2. The YoC factor yielded a significant influence on TMY of cows whose lactation records were included in Node 1 ($F = 55.548$, $df1 = 21$, $df2 = 2176$, $P < 0.01$). This shows the increase in total milk yield over the years. Afterwards Node 1 was divided into two child nodes (Nodes 3 and 4), depending on DiM factor. Corresponding average values for these two new child nodes were 2693.45 kg and 3799.21 kg, respectively. Numbers of lactation records were established as 223 (10.1%) and 217 (9.9%). While TMY values of cattle with DiM of 307.5 days and less were 2693.45 kg and TMY values of cattle with DiM more than 307.5 days, were found as 3799.21 kg. On the other hand, Node 3 is divided into 2 branches according to DiM factor. While TMY value of 76 cows with DiM 271.5 days and less was found to be 2246.47 kg, TMY value of 147 cows with DiM more than 271.5 days was found as 2924.54 kg. DiM is the first criterion in the TMY classification of records from 1993 and after (Node 2). DiM factor remarkably influenced the TMY of cows whose lactation records were consisted of DiM less than 322.5 days in the Node 2. The TMY averages for these two nodes (Node 5 and Node 6) were found as: 5172.333 kg and 7049.87 kg, respectively. Numbers (proportions) of lactation records were established as 1023 (46.5%) and 735 (33.4%), respectively. According to DiM factor these records are also divided into animal classes which are less than 285.5 days (Node 9) and more than 285.5 days (Node 10). Node 9 is divided into classes with DiM less than 241.5 days (Node 13) and more than 241.5 days (Node 14). The Node 13 consists of 54 records and the average TMY value is 3427.19 kg. The Node 14 was divided into two child nodes: the Tahirova farm with an average TMY of 6138.00 kg and the Malya and Konuklar farms with an average TMY of 4612.06 kg. The records with DiM over 285.5 (Node 10) was divided into two groups. 81 animals with an average TMY value of 6883.35 kg were belonged to Tahirova (Node 15) farm and 467 animals with an average TMY value of 5419.06 kg belonged to the Konuklar and Malya farms (Node 16). The Node 16 was then further classified (Node 23 and Node 24) according to the CA factor that was not statistically significant. Animals with DiM of 322.5 days and more were divided into two child nodes on the basis of farm (Node 11.
and Node 12). 182 animals with an average TMY value of 8391.31 kg were belonged to the Tahirova (Node 11) and 553 animals with an average TMY of 6608.38 kg were belonged to the Malya and Konuklar farms (Node 12). 182 animals (Node 11) in the Tahirova farm were divided into 2 homogeneous classes according to the DiM criteria: 104 animals with an average of 7750.57 kg milk and less than 389.5 DiM (Node 17) and 78 animals with an average of 9245.63 kg of milk and 389.5 and higher DiM (Node 18). The animals in the Malya and Konuklar farms (Node 12) were classified according to the DiM factor. Accordingly, the average TMY value of the 430 animals (Node 19) with a DiM of less than 409.5 days was 6311.57 kg, and the 123 animals (Node 20) with a DiM of more than 409.5 days were 7646.02 kg. Then, Node 19 was divided into two child nodes according to the YoC factor as before 2005 (Node 25) and after 2005 (Node 26).

On account of the fact that Nodes 4, 7, 8, 13, 15, 17, 18, 20, 21, 22, 23, 24, 25 and 26 are terminal nodes, there is not any separation for providing homogeneity in those nodes.

The following algorithms were used for the estimation of the TMY in the regression tree method we applied:

\[
\begin{align*}
R_1 &= (YoC \leq 1994) \cap (DiM > 307.5) \\
R_2 &= (YoC \leq 1994) \cap (DiM \leq 307.5) \cap (DiM \leq 271.5) \\
&= (YoC \leq 1994) \cap (DiM \leq 271.5) \\
R_3 &= (YoC \leq 1994) \cap (DiM \leq 307.5) \cap (DiM > 271.5) \\
&= (YoC \leq 1994) \cap (271.5 < DiM \leq 307.5) \\
R_4 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM \leq 285.5) \cap (DiM \leq 241.5) \\
&= (YoC > 1994) \cap (DiM \leq 241.5) \\
R_5 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM \leq 241.5) \cap (DiM > 241.5) \\
&= (YoC > 1994) \cap (DiM > 241.5) \cap (DiM \leq 265.5) \cap (DiM > 265.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \cap (DiM \leq 265.5) \cap (DiM > 265.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \cap (DiM \leq 265.5) \\
R_6 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (YoC > 1994) \cap (DiM > 241.5) \cap (DiM > 265.5) \\
&= (DiM > 241.5) \cap (DiM > 265.5) \\
R_7 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \cap (DiM > 265.5) \\
&= (DiM > 241.5) \\
R_8 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_9 &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_{10} &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_{11} &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_{12} &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_{13} &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
R_{14} &= (YoC > 1994) \cap (DiM \leq 322.5) \cap (DiM > 285.5) \cap (DiM > 241.5) \\
&= (DiM > 241.5) \\
&= (DiM > 241.5) \\
\end{align*}
\]

\[
\begin{align*}
k_1 &= 3799.2 kg \\
k_2 &= 2246.5 kg \\
k_3 &= 2924.5 kg \\
k_4 &= 3427.2 kg \\
k_5 &= 6138.0 kg \\
k_6 &= 4612.1 kg \\
k_7 &= 6883.4 kg \\
k_8 &= 5157.6 kg \\
k_9 &= 5652.0 kg \\
k_{10} &= 7750.6 kg \\
k_{11} &= 9245.6 kg \\
k_{12} &= 7646.0 kg \\
k_{13} &= 6134.8 kg \\
k_{14} &= 7039.6 kg
\end{align*}
\]
where, $P_i$ shows the algorithm to the terminal Node. TMY obtained from each algorithm is shown with the variable $k_i$.

There are 14 distinct paths from the root node to the leaves. This three divides the input data in 14 different regions. Using the (Eq. 3), we obtain:

$$n_k(x) = \sum_{i=1}^{14} k_i \times I(F_i)$$

Figure 6 shows importance level of the independent variables. According to Figure 6, YoC was significantly identified as the most important factor influencing TMY.

![Figure 6. Importance level of independent variables affecting TMY](image)

Their importance rate was given in Table 4.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Importance</th>
<th>Normalized importance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of Calving</td>
<td>1299236.915</td>
<td>100.0</td>
</tr>
<tr>
<td>Days in Milk</td>
<td>1052872.027</td>
<td>81.0</td>
</tr>
<tr>
<td>Farm</td>
<td>765055.898</td>
<td>58.9</td>
</tr>
<tr>
<td>Breed</td>
<td>741247.200</td>
<td>57.1</td>
</tr>
<tr>
<td>Season</td>
<td>29218.096</td>
<td>2.2</td>
</tr>
<tr>
<td>Calving Age</td>
<td>24202.305</td>
<td>1.9</td>
</tr>
</tbody>
</table>

The $R^2$ of the tree is 78.9%, which is significantly higher than that of a multiple linear regression fit to the same data ($R^2 = 59.1\%$). In current YoC was found most
important factor affecting total milk yield followed by days in milk (81%). Teke and Akdağ (2010); Bayril and Yılmaz (2010) in their study indicated that the effect of calving year on lactation milk yield were significant in Holstein cows. The categorical variable farm, as we expected, is grouped separately Malya and Konuklar (close to each other geographically) and Tahiroya. The other common characteristics of Konuklar and Malya farms are that they have the same breed animal (Holstein). As a result of the analysis, it is expected that the importance rate of both farm and breed characteristics are close (58.9% and 57.1%, respectively). Gürses and Bayraktar (2012) reported that effects of enterprise, calving year were found significant on 305 days milk yield. Season and age at calving was not found so important in the total milk yield prediction (2.2% and 1.9%, respectively). In the study of Topal et al. (2010) season of calving of Swedish Red cattle also was found the least important factor (5.9%) affecting actual milk yield. But there are some other works were claiming the opposite (Bakır and Çetin, 2003; Özçakır and Bakır, 2003; Sehar and Özbeyaz, 2005; Erdem et al., 2007; Akçay et al., 2007; Çilek, 2009).

Conclusion

Due to the fact that CART is not a parametric method, the data used here is not required be belonged to a particular type of distribution. Also, this method can easy determine effects of both continuous and categorical variables on dependent continuous variable.

In regression tree diagram, the year of calving trait was determined to be the most affective factor for total milk yield prediction, followed by days in milk, farm and breed (P < 0.01). Thus, while the average total milk yield in the first lactation was 3238.79 ± 1130.83 until 1993, this yield increased in years and reached 6134.81 ± 1313.56 kg in 2005. This yield has increased to 7039.63 ± 1445.52 kg in Malya and Konuklar farms after 2005. In the Tahiroya farm, we can say from the P11 path that the total milk yield increased after 1994 and reached the average of 9245.63 ± 1580.26 kg. Insignificant input variables such as season and age at calving were excluded from the regression tree.

As a result, the routing of the algorithms leading to the terminal node will help to determine the criteria for the classification in the regression tree and the application of the test data to this algorithm.

REFERENCES


