ISSN 1644-0714
ISSN 2300-6145 (online)
DOI:10.21005/asp.2016.15.4.05
www.asp.zut.edu.pl

Acta Sci. Pol. Zootechnica 15(4) 2016, 55–70

CLASSIFICATION OF CALVING DIFFICULTY SCORES USING DIFFERENT TYPES OF DECISION TREES

Daniel Zaborski✉, Witold S. Proskura, Wilhelm Grzesiak

Department of Ruminants Science, West Pomeranian University of Technology, Szczecin, Doktora Judyma 10, 71-466 Szczecin, Poland

Abstract. The aim of this study was to classify the cases of calving difficulty using selected data mining methods [Classification and Regression Trees (CART), Chi-square Automatic Interaction Detector (CHAID) and Quick, Unbiased, Efficient, Statistical Trees (QUEST)] and generalized linear model (GLZ) and to identify their most important predictors. A total of 1699 records of Polish Holstein-Friesian Black-and-White cows were used. Calving difficulty had three categories (easy, moderate and difficult). Percentages of calvings correctly classified by CART, CHAID, QUEST and GLZ, respectively, were as follows: 60.20, 65.31, 68.88 and 66.33% (easy), 71.36, 69.01, 64.79 and 69.01% (moderate) and 0, 0, 0 and 0% (difficult). The most influential predictors of calving difficulty were the rank of the dam's sire, calf sex, calving age, previous calving difficulty, lactation number, daily milk yield and average milk yield of the farm. The tree models and GLZ were of moderate quality. None of them could correctly indicate dystocia.

Key words: cattle, decision support systems, diagnosis, dystocia

INTRODUCTION

Dystocia may be defined as a difficulty at calving caused by prolonged spontaneous parturition as well as prolonged or severe assisted traction [Vanderick et al. 2014]. Its severity is most often assessed on a point scale (e.g. 1 to 5) and a score above 3 usually describes a difficult calving [Mee 2004, Ghavi Hossein-Zadeh 2013]. Dystocia results in many adverse consequences such as: reduced

✉daniel.zaborski@zut.edu.pl

milk, fat and protein yield, decreased fertility, higher mortality rates, more frequent occurrence of retained placenta, mastitis, hypocalcaemia and a higher risk of difficulties at subsequent parturition [Mee 2008, McHugh et al. 2011, Atashi et al. 2012, Azizzadeh et al. 2012, Ribeiro et al. 2013, Alvasen et al. 2014]. Moreover, the squeals of dystocia include higher culling rates, increased labor intensity and reduced animal welfare, all of which are associated with great economic losses [Barrier et al. 2010]. Besides detrimental effects for a cow, dystocia exerts negative influence on calf vigor and viability (injury and pain, asphyxia and acidosis, impaired homeostasis and thermoregulation, behavioral consequences such as difficulties in standing, walking and reaching the udder, longer periods of lying down on the flank). It also increases the risk of calf stillbirth (even a four times higher rate than for eutocial calves), perinatal mortality and morbidity associated with inadequate passive transfer of immunity mainly caused by the low volume of ingested colostrum [Barrier et al. 2012]. This in turn may result in greater susceptibility to diseases such as pneumonia or diarrhea, reduced daily body weight gains, lower milk production during subsequent lactations and an increased culling rate [Murray and Leslie 2013]. Many direct and indirect factors exist that affect the occurrence of dystocia, e.g. feto-pelvic incompatibility, weak labor, incomplete dilation of the cervix, uterine torsion and fetal malpresentation as well as cow body weight and condition at calving, cow age, breed, parity and nutrition during gestation, calf sex and breed, etc. [Barrier 2012, Schuenemann et al. 2013]. Genetic factors are also involved. From among many genes with confirmed or potential effect on dystocia in cattle, the following are frequently mentioned: hedgehog interacting protein (*HHIP*) gene, microRNA mir-1256 (*MIR1256*) gene, sialic acid binding Ig-like lectin 5 (*Siglec-5*) gene, zinc finger protein 827 gene, IGF-binding protein 2 precursor (*IGFBP-2*) gene, IGF binding protein-3 (*IGFBP-3*) gene, non-SMC condensin I complex, subunit G (*NCAPG*) gene; solute carrier family 44, member 5 (*SLC44A5*) gene, secreted phosphoprotein 1 (*SPP1*) gene, integrin-binding sialoprotein (*IBSP*) gene, matrix extracellular phosphoglycoprotein (*MEPE*) gene, the cluster of differentiation 37 (*CD37*) gene, insulin-like growth factor I (*IGF1*) gene and many others [Cole et al. 2011, Macciotta et al. 2014, Purfield et al. 2014]. In the present study, a set of 10 selected predictors was used to classify calving difficulty. Calving age is an important determinant of dystocia, since younger cows usually experience more difficult parturitions irrespective of parity, although lower-parity dams (especially heifers) also tend to have more difficult calvings than older ones. The effect of the cow's sire is responsible for the genetic component of dystocia, whereas the average milk production on the farm represents the influence of environmental factors associated with the husbandry conditions. Moreover, calvings tend to be more difficult during the autumn-winter season under our climatic conditions and in dams giving birth to

male calves. The difficulty of a preceding calving also belongs to the influential predictors of dystocia, with dystotic cows having significantly higher chances of difficulties at subsequent parturition. Finally, some diseases during pregnancy may increase cow's susceptibility to other undesired conditions such as dystocia [Mee 2008]. One method of detecting animals with potential calving problems is to use statistical predictive models, especially from the currently intensively developing field of data mining.

From among various data mining algorithms, decision trees deserve special attention due to their high predictive efficiency and ease of interpretation. In the present study, three different types of decision trees [Classification and Regression Trees (CART), Chi-square Automatic Interaction Detection trees (CHAID) and Quick, Unbiased, Efficient, Statistical Trees (QUEST)] were investigated in terms of their effectiveness in the classification of calving difficulty scores in cows because each type is created using a different algorithm, which affects their applicability to a given problem. However, a more traditional generalized linear model (GLZ) was also used as a reference for the data mining algorithms.

Therefore, the aim of the present study was to classify calving difficulty cases in dairy cows using CART, CHAID, QUEST and GLZ and to identify the most important predictors of calving course.

MATERIAL AND METHODS

A total of 1699 calving records of Polish Holstein-Friesian Black-and-White cows kept on the four farms located in the West Pomeranian Province were analyzed. The records were collected in the years 2002 through 2013. The animals were housed in free-stall barns and fed a total mixed ration. Milking was performed twice daily in herring-bone milking parlors. Each calving record included ten predictors: X_1 – SIRE – the rank of the cow's sire based on the mean calving difficulty scores of his daughters (the smallest rank indicated a sire with the easiest calvings), X_2 – CALA – calving age (in months), X_3 – FARM – the category of the cow's farm based on its mean milk yield determined using the k-means clustering method (below 10.200 kg – Lower or above 10.200 kg – Higher), X_4 – SEX – calf sex (male or female, multiple calvings were excluded due to their low frequency), X_5 – CALS – calving season (autumn-winter from October to March – AW and spring-summer from April to September – SS), X_6 – DMY – the mean daily milk yield for the preceding complete lactation (in kg), X_7 – CIN – preceding calving interval (in days), X_8 – LACT – lactation number, X_9 – PDIF – the difficulty of the preceding calving (easy, moderate or difficult), X_{10} – MAST – udder diseases (mainly mastitis) during pregnancy (diseased vs. healthy). The dependent variable was a calving difficulty score. Calving difficulty was evaluated

on a five-point (before 2006) or six-point (since 2006) scale. Then, the scale was changed to a categorical one with three categories (easy, moderate and difficult). The descriptive statistics for the predictors are given in Table 1.

The entire data set was divided into a training set (L, 1275 records) for preparing the models and a test set (T, 424 records) for their objective verification. Three different algorithms for classification tree construction (CART, CHAID and QUEST) were applied in the present study. The first one generates only binary trees in which parent nodes are always split into two child nodes. The identification of a predictor on which the split in a given parent node is based is performed through an iterative procedure consisting in the verification of all the possible values of each predictor at each split. This procedure also allows determination of the cut-off value for a given predictor to ensure that the resulting child nodes are as pure as possible (contain the most homogeneous groups of calvings) [Speybroeck 2011]. The whole process is iteratively continued until additional splits are no longer possible. However, the constructed classification tree is most often too complex (this phenomenon is known as overfitting) and must be pruned [Moisen 2008]. In contrast to CART, the CHAID algorithm is not limited to binary splits and utilizes the chi-square test to find the best predictor value for each split. Also, pruning is not necessary for such trees since the addition of new nodes stops before the occurrence of overfitting [Chang 2007]. The last algorithm considered in the present study (QUEST) builds binary trees using a quadratic discriminant function analysis (QDA) for determining the best split [Loh and Shih 1997].

The construction of all the three classification tree types involved the equal costs of misclassifications, the minimization of a misclassification error as a stopping criterion and a 10-fold cross-validation to determine the best tree structure with the highest generalization ability. Moreover, in the case of CART and CHAID, the minimal node size was 169 cases, whereas for QUEST, the minimal terminal node size was 5. For CART and QUEST, the *a priori* probabilities for each calving category were estimated from the L set. Finally, CART used the Gini measure of node impurity, CHAID utilized the Bonferroni adjusted p-values of 0.05 for splitting and merging, respectively, and a more accurate exhaustive method to obtain the best tree, whereas QUEST used the p-value of 0.05 for selecting a splitting variable. The last investigated model, GLZ, included the ordinal multinomial distribution of calving difficulty scores and a logit link function. To evaluate its goodness of fit, the deviance statistic was used and the assumptions of its applicability were verified.

After developing the models, their classification quality was evaluated on the L set by calculating the proportions of correctly classified calvings from each category and from all the categories together (the so-called accuracy – Acc). The significance of the differences in these proportions (for $P \leq 0.05$) was verified

Table 1. Descriptive statistics for input and output variables

Tabela 1. Statystyki opisowe zmiennych wejściowych i wyjściowych

Variable/category Kategoria zmiennej	Training set (n_L = 1275) Zbiór uczący (n_L = 1275)		Test set (n_T = 424) Zbiór testowy (n_T = 424)		Total (n_{total} = 1699) Razem (n_{total} = 1699)	
	Continuous variables – Zmienne ciągłe					
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
CALA, months – miesiąc	49.67	12.97	49.43	12.10	49.61	12.76
SIRE, rank – ranga	52.49	26.77	48.90	27.51	51.59	26.99
DMY, kg	33.72	4.97	33.66	5.13	33.70	5.01
CIN, days – dni	407.34	53.78	406.79	55.99	407.20	54.32
	Categorical variables – Zmienne nominalne					
	n	%	n	%	n	%
			FARM			
Lower – Niższa	553	43.37	212	50.00	765	45.03
Higher – Wyższa	722	56.63	212	50.00	934	54.97
			CALS			
Autumn-winter Jesienno-zimowy	583	45.73	200	47.17	783	46.09
Spring-summer Wiosenno-letni	692	54.27	224	52.83	916	53.91
			SEX			
Male – Męska	654	51.29	197	46.46	851	50.09
Female – Żeńska	621	48.71	227	53.54	848	49.91
			LACT			
2nd – Druga	673	52.78	215	50.71	888	52.27
3rd – Trzecia	355	27.84	141	33.25	496	29.19
4th – Czwarta	179	14.04	48	11.32	227	13.36
5th–10th – 5.–10.	68	5.33	20	4.72	88	5.18
			PDIF			
Easy – Łatwe	490	38.43	173	40.80	663	39.02
Moderate – Średnie	514	40.31	173	40.80	687	40.44
Difficult – Trudne	271	21.25	78	18.40	349	20.54
			MAST			
Healthy – Zdrowa	1179	92.47	399	94.10	1578	92.88
Ill – Chora	96	7.53	25	5.90	121	7.12
			DIF – output variable			
Easy – Łatwe	561	44.00	196	46.23	757	44.56
Moderate – Średnie	671	52.63	213	50.24	884	52.03
Difficult – Trudne	43	3.37	15	3.54	58	3.41

CALA: calving age, SIRE: sire's rank, DMY: mean daily milk yield, CIN: preceding calving interval, FARM: category of the farm where the animal was kept based on its mean milk yield (Lower: < 10,200 kg, Higher: ≥ 10,200 kg), CALS: calving season, SEX: calf sex, LACT: lactation number, PDIF: preceding calving difficulty, MAST: udder diseases during pregnancy, DIF: calving difficulty.

CALA: wiek wycielenia, SIRE: ranga buhaja, DMY: średnia dzienna wydajność mleka, CIN: poprzedni okres międzywycieleniowy, FARM: kategoria gospodarstwa na podstawie średniej wydajności mleka (Niższa: < 10.200 kg, Wyższa: ≥ 10.200 kg), CALS: sezon wycielenia, SEX: płeć cielęcia, LACT: kolejna laktacja, PDIF: trudność poprzedniego porodu, MAST: choroby wymienia podczas ciąży, DIF: trudności wycielenia.

with tests for proportions. Next, the most influential predictors of calving difficulty were identified using the "importance analysis" for the trees and the significance of the Wald statistic for GLZ. In the end, all the models were assessed on the independent T set to objectively verify their classification performance. All the computations were performed using Statistica 12 software (StatSoft Inc., Tulsa, OK, USA).

RESULTS

The layouts of the CART, CHAID and QUEST trees are presented in Figure 1 and the estimated GLZ parameters are shown in Table 2. The deviance statistic divided by its degrees of freedom was 0.73, but not all the assumptions of GLZ were fulfilled (the lack of the normal distribution of residuals). Classification results on the L set are given in Table 3. QUEST most precisely classified easy calvings (68.63%) and CART moderate ones (78.99%). These two tree types also exhibited the greatest Acc (67.61% and 65.41% for QUEST and CART, respectively). However, none of the models could correctly indicate difficult calvings. The importance of calving difficulty predictors for the trees is presented in Figure 2 and the significance of the GLZ effects in Table 2. The most influential factor for all decision trees and one of the two significant predictors for GLZ was SIRE. Also, SEX, CALA, PDIF, LACT, DMY and FARM were important in determining the category of calving difficulty, whereas the influence of the remaining factors varied depending on the model. The classification performance on the T set is shown in Table 3. The results observed on the L set were generally confirmed on the T set. However, the only significant difference in the proportions of correctly classified easy calvings existed between CART and QUEST (60.20% and 68.88%, respectively) and, as previously, none of the models could correctly detect even one dystocia case.

DISCUSSION

The first division in CART and QUEST was based on SIRE and in CHAID on PDIF (Fig. 1). Of the other predictors, SEX was used frequently. The CART and CHAID trees constructed by Piwczyński et al. [2013] for classifying calving difficulty in Polish Holstein-Friesian Black-and-White cows used lactation number as the first splitting variable. This variable was also selected by CHAID in the present work, after PDIF and FARM (Fig. 1B). The next divisions in the cited study were based on calf birth weight, gestation length and management system. It should be noted that the first two of the above-mentioned variables used by Piwczyński et al. [2013] are relatively difficult to obtain before parturition (calf

Table 2. Estimated parameters of the generalized linear model for calving difficulty scores

Tabela 2. Oszacowane parametry uogólnionego modelu liniowego dla kategorii trudności porodu

Model term Składnik modelu	Level Kategoria	Estimate Oszacowanie	Standard error Błąd standardowy	Wald statistic Statystyka Walda	P
Intercept 1 – 1. wyraz wolny		1.2897	1.1333	1.2949	0.2551
Intercept 2 – 2. wyraz wolny		5.2943	1.1486	21.2459	0.0000
CALA		0.0046	0.0202	0.0511	0.8212
CIN		0.0005	0.0014	0.1207	0.7283
DMY		–0.0138	0.0138	0.9984	0.3177
SIRE		–0.0238	0.0027	76.5184	0.0000
FARM	Lower – Niższa	–0.0030	0.0636	0.0023	0.9619
CALS	Autumn-winter Jesienno-zimowy	0.0342	0.0598	0.3258	0.5682
LACT	2nd – 2.	–0.3707	0.4221	0.7713	0.3798
LACT	3rd – 3.	0.0934	0.1813	0.2654	0.6064
LACT	4th – 4.	0.2508	0.2026	1.5325	0.2157
PDIF	Moderate – Średni	0.1616	0.0871	3.4407	0.0636
PDIF	Easy – Łatwy	0.0178	0.0940	0.0360	0.8496
SEX	Male – Męska	–0.3554	0.0597	35.3979	0.0000
MAST	Healthy – Zdrowa	–0.2008	0.1160	2.9972	0.0834

CALA: calving age, SIRE: sire's rank, FARM: farm category, CALS: calving season, SEX: calf sex, CIN: preceding calving interval, DMY: mean daily milk yield, LACT: lactation number, PDIF: preceding calving difficulty, MAST: udder diseases.
CALA: wiek wycielenia, SIRE: ranga buhaja, FARM: kategoria gospodarstwa, CALS: sezon wycielenia, SEX: płeć cielęcia, CIN: poprzedni okres międzywycieleniowy, DMY: średnia dzienna wydajność mleka, LACT: kolejna laktacja, PDIF: trudność poprzedniego porodu, MAST: choroby wymienia podczas ciąży.

birth weight is recorded after delivery and precise gestation length can only be estimated). The estimated GLZ parameters in the present study are given in Table 2. The deviance value relative to its degrees of freedom (0.73) confirmed a relatively good quality of the constructed model as the values of about 1.0 are considered to indicate a good fit of the model to the training data [McCullagh and Nelder 1989]. However, since not all the GLZ assumptions were met, its application may not be completely justified from the statistical point of view. The Acc on the L set (approximately 65%) in the present work (Table 3) was comparable to that (61.50%) reported by Piwczyński et al. [2013]. A very unfavorable outcome observed in the present study was a complete inability of the models to correctly indicate difficult calvings (dystocia). However, it could have resulted from a very low frequency of dystocia in cows (less than 4% – Table 1).

Table 3. Proportions of correctly classified calvings on the training and test sets

Tabela 3. Proporcje poprawnie sklasyfikowanych wycieleń na zbiorze uczącym i testowym

Set Zbiór	Calving – Wycielenie			Accuracy Trafność
	Easy – Łatwe	Moderate – Średnie	Difficult – Trudne	
		CART		
Training – Uczący	0.5419[a]	0.7899[c]	0.0000	0.6541[ab]
Test – Testowy	0.6020[a]	0.7136	0.0000	0.6368
		CHAID		
Training – Uczący	0.5668[ab]	0.7496[a]	0.0000	0.6439[b]
Test – Testowy	0.6531	0.6901	0.0000	0.6486
		QUEST		
Training – Uczący	0.6863[c]	0.7109[b]	0.0000	0.6761[a]
Test – Testowy	0.6888[b]	0.6479	0.0000	0.6439
		GLZ		
Training – Uczący	0.6114[b]	0.7168[ab]	0.0000	0.6463[b]
Test – Testowy	0.6633	0.6901	0.0000	0.6533

[a-b]Values marked with different superscripts within a column (and a set) differ significantly ($P \leq 0.05$), CART: classification and regression trees, CHAID: chi-square automatic interaction detection, QUEST: quick, unbiased, efficient, statistical trees, GLZ: generalized linear model.

[a-b]Wartośći oznaczone różnymi indeksami w obrębie kolumny (i zbioru) różnią się istotnie ($P \leq 0,05$), CART: drzewa klasyfikacyjne i regresyjne, CHAID: chi-square automatic interaction detection, QUEST: quick, unbiased, efficient, statistical trees, GLZ: uogólniony model liniowy.

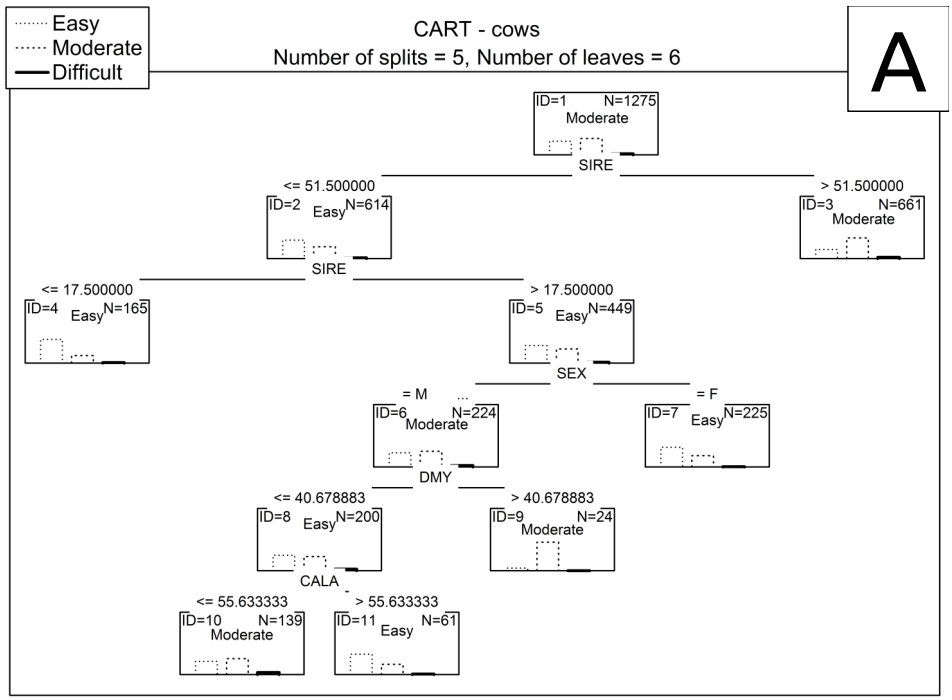
The most influential factor affecting delivery course in cows in the current work was SIRE (Fig. 2). It was based on the mean calving difficulty score for the daughters of the cow's sire and served as a measure of one of the indirect effects influencing calving difficulty [Wautlet et al. 1990]. The next important predictor for all the models was SEX (Fig. 2, Table 2). This effect was especially visible for the CHAID tree (Fig. 1B), where the female sex was usually associated with easy calvings and male sex with moderate ones. The higher incidence of calving assistance and dystocia in dams giving birth to male calves is a well-known fact confirmed by many authors. In a recent study on maternal beef cattle traits in Ireland, McHugh et al. [2014] observed a 2.23 times higher calving assistance risk and a 2.50 times higher dystocia risk for bull calves compared with heifer calves. Similar findings were reported by Gevrekci et al. [2011], who investigated dystocia in US Holsteins, and found that its frequency in bull calves was approximately two times higher compared with heifers (23.29 vs. 13.27% in the first lactation, 11.00 vs. 5.31% in the second lactation and 9.58 vs. 5.37% in the third lactation). Also, the study by Atashi et al. [2012] on the Holstein breed in Iran showed calf sex to be significantly associated with the presence of dystocia with dams giving birth to male calves having a significantly higher probability of its occurrence. However, it should be added that some authors [Bazzi 2010, Gaafar

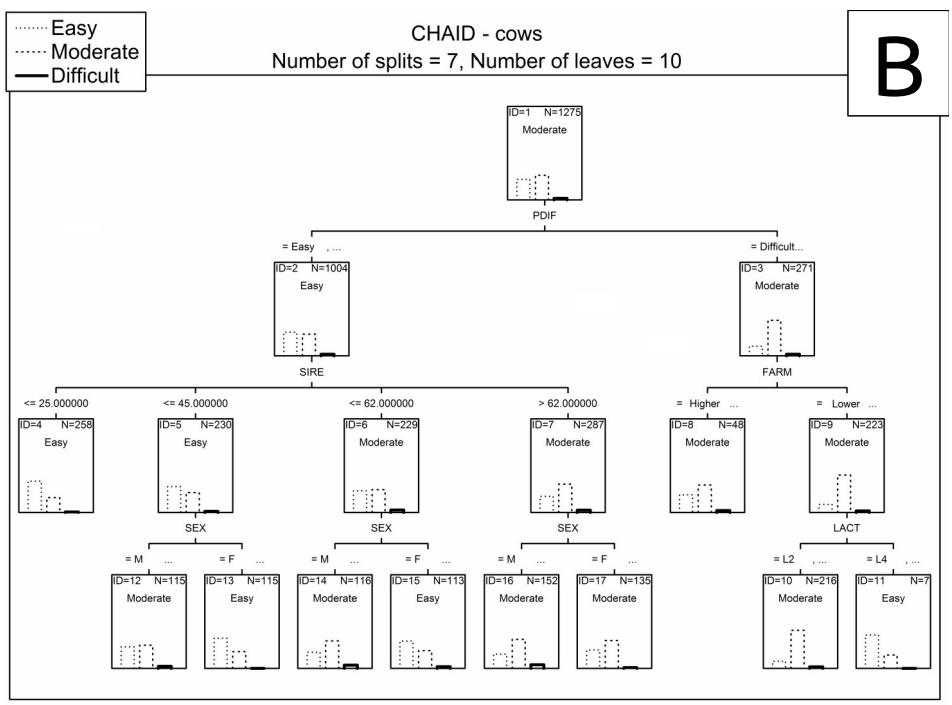
et al. 2011] did not observe any significant effect of calf sex on dystocia frequency. On the other hand, calf sex is also associated with its birth weight and it has been proved in some studies [Johanson and Berger 2003] that birth weight could be even a better predictor of calving difficulty than calf sex as each 1 kg increase in body weight was associated with a 13.0% higher probability of dystocia. Finally, it should be added that because the models in the present study were built mainly for predictive purposes, the sex of the calf must be determined before parturition, which may be problematic in practice. But the latest development in ultrasonography makes it possible to identify calf sex at a relatively early stage of pregnancy (approximately 55 to 60 days post-conception), so its more common use in future could make the determination of calf sex easier.

An important predictor for all the trees but not for GLZ was CALA. It was used once by CART (Fig. 1A) and thrice by QUEST (Fig. 1C). In general, younger cows experience more difficult calvings, irrespective of lactation number [Wautlet et al. 1990], although some authors [Hickey et al. 2007, Bazzi 2010, Yıldız et al. 2011] did not confirm any significant relationship between calving age and difficulty. The next four influential predictors of calving difficulty identified only by the decision trees were PDIF, LACT, DMY and FARM. It has been shown that cows experiencing dystocia had a 1.65 times greater dystocia risk and a 2.90 times greater chance of calving assistance at subsequent parturition compared with the dams without such difficulties [Mee et al. 2011]. The second afore-mentioned variable, i.e. DMY, has not been explicitly proved in the available literature to significantly affect the risk of dystocia so far [Ingvartsen et al. 2003], whereas the third one, i.e. LACT, is closely related to calving age, which has already been discussed above. Finally, FARM was used by the CHAID tree for splitting once (Fig. 1B), but the resulting separation was poor (both child nodes contained mainly moderate calvings). The effect of the remaining predictors included in the analysis was relatively small and differed depending on the type of classification model.

In general, the Acc on the T set in the present work was moderate (Table 3). However, none of the difficult deliveries was correctly detected. On the other hand, the proportion of properly indicated easy and moderate calvings exceeded 60%. The comparison of tree algorithms with GLZ (used as a reference) showed that both model types were characterized by similar classification performance. However, parametric methods such as GLZ require the fulfillment of various assumptions, from which not all were met in the present study. Moreover, a classification tree structure is more easily interpretable, which makes it easier to understand different relationships between the factors affecting calving course.

Two important factors could have affected the relatively low discrimination power of the applied predictors in terms of difficult calvings. One of them was a


Easy
Moderate
Difficult
CART - cows
Number of splits = 5, Number of leaves = 6
A
ID=1 N=1275 Moderate
SIRE
<= 51.500000
> 51.500000
ID=2 Easy N=614
ID=3 N=661 Moderate
SIRE
<= 17.500000
> 17.500000
ID=4 Easy N=165
ID=5 Easy N=449
SEX
= M
= F
ID=6 N=224 Moderate
ID=7 Easy N=225
DMY
<= 40.678883
> 40.678883
ID=8 Easy N=200
ID=9 N=24 Moderate
CALA
<= 55.633333
> 55.633333
ID=10 N=139 Moderate
ID=11 Easy N=61


Easy
Moderate
Difficult
CHAID - cows
Number of splits = 7, Number of leaves = 10
B
ID=1 N=1275 Moderate
PDIF
= Easy , ...
= Difficult...
ID=2 N=1004 Easy
ID=3 N=271 Moderate
SIRE
FARM
<= 25.000000
<= 45.000000
<= 62.000000
> 62.000000
= Higher ...
= Lower ...
ID=4 N=258 Easy
ID=5 N=230 Easy
ID=6 N=229 Moderate
ID=7 N=287 Moderate
ID=8 N=48 Moderate
ID=9 N=223 Moderate
SEX
SEX
SEX
LACT
= M ...
= F ...
= L2 , ...
= L4 , ...
ID=12 N=115 Moderate
ID=13 N=115 Easy
ID=14 N=116 Moderate
ID=15 N=113 Easy
ID=16 N=152 Moderate
ID=17 N=135 Moderate
ID=10 N=216 Moderate
ID=11 N=7 Easy

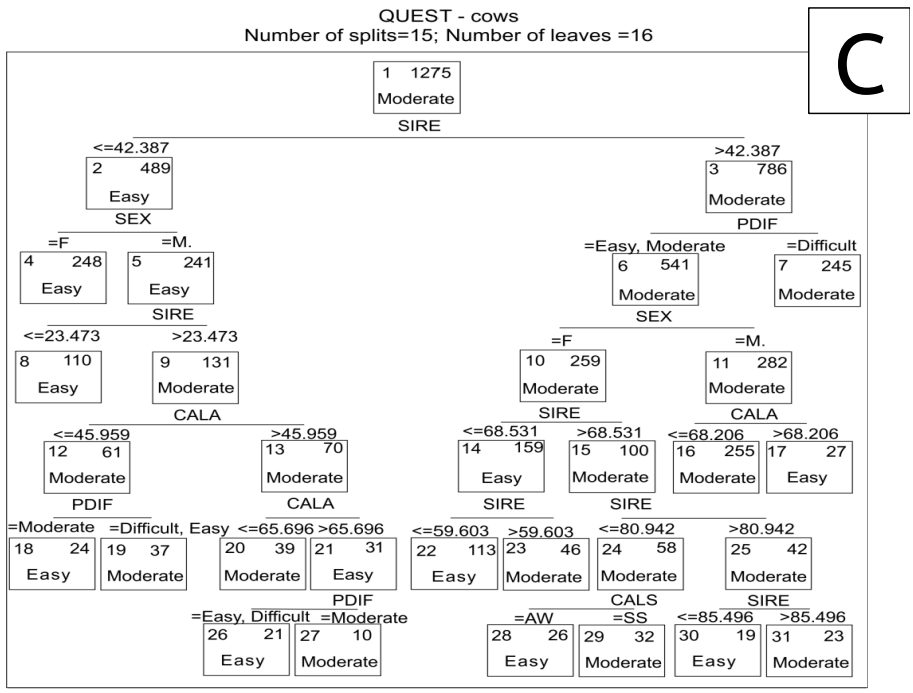


Fig. 1. Models for calving classification: **A** – Classification and Regression Tree (CART), **B** – Chi-square Automatic Interaction Detection (CHAID), **C** – Quick, Unbiased, Efficient, Statistical Trees (QUEST); SIRE: sire's rank, SEX: calf sex, DMY: mean daily milk yield, CALA: calving age; PDIF: preceding calving difficulty, FARM: farm category, LACT: lactation number, CALS: calving season (AW – autumn-winter, SS – spring-summer). In the case of the QUEST trees, the number in the top left-hand corner of each node represents its ID, the number in the top right-hand corner represents its size (the number of calvings in this node) and the label of each node is based on the most numerous calving category in this node

Rys. 1. Modele do klasyfikacji wycieleń: **A** – drzewa klasyfikacyjne i regresyjne (CART), **B** – Chi-square Automatic Interaction Detection (CHAID), **C** – Quick, Unbiased, Efficient, Statistical Trees (QUEST); SIRE: ranga buhaja, SEX: płeć cielęcia (M – buhajek, F – jałówka), DMY: średnia dzienna wydajność mleka, CALA: wiek wycielenia; PDIF: trudność poprzedniego porodu, FARM: kategoria gospodarstwa (Lower – niższa, Higher – wyższa), LACT: kolejna laktacja, CALS: sezon wycielenia (AW – jesienno-zimowy, SS – wiosenno-letni), Easy – wycielenie łatwe, Moderate – wycielenie średnie, Difficult – wycielenie trudne, Number of splits – liczba podziałów, Number of leaves – liczba liści (węzłów końcowych). W przypadku drzew QUEST, liczba w lewym górnym rogu węzła reprezentuje jego kolejny numer, liczba w prawym górnym rogu – liczebność węzła (liczba wycieleń w węźle), zaś etykieta każdego węzła jest określana na podstawie najliczniejszej kategorii wycielenia w tym węźle

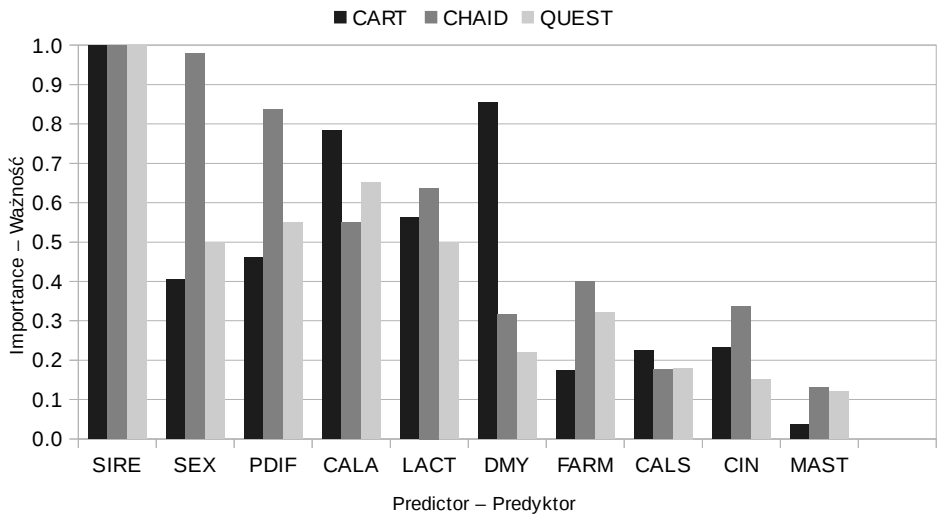


Fig. 2. The importance of individual predictors of calving difficulty; SIRE: sire's rank, SEX: calf sex, PDIF: preceding calving difficulty, CALA: calving age; LACT: lactation number, DMY: mean daily milk yield, FARM: farm category, CALS: calving season; CIN: preceding calving interval, MAST: udder diseases

Rys. 2. Ważność predyktorów trudności wycieleń; SIRE: ranga buhaja, SEX: płeć cielęcia, PDIF: trudność poprzedniego porodu, CALA: wiek wycielenia, LACT: kolejna laktacja, DMY: średnia dzienna wydajność mleka, FARM: kategoria gospodarstwa, CALS: sezon wycielenia, CIN: poprzedni okres międzywycieleniowy, MAST: choroby wymienia podczas ciąży

low occurrence of dystocia in the analyzed dataset (less than 4%), which makes it really hard for the prediction model to correctly identify such cases in practice. The second one could be the categorization of the nominal predictors adopted in the present work. For most of them (SEX, LACT), the division into categories was strictly determined by the structure of the dataset. In the case of FARM, the classification of the four farms according to their average milk production was based on the results obtained using a non-hierarchical k-means clustering method, whereas calving season was split into two categories (autumn-winter and spring summer) with almost equal number of calvings in each, although the division into four categories would also be possible. The MAST variable was categorized based on the frequency of individual diseases occurring during pregnancy, which was highest for mastitis. Finally, classification of the preceding calving difficulty was the same as for the output variable, which in turn resulted from practical reasons (five or six original categories converted to three classes of calving difficulty).

CONCLUSIONS

The tree models (CART, CHAID and QUEST) developed in the present study were of moderate quality. The most important calving category, i.e. dystocia, could not have been correctly indicated by any model. However, it should be added that the final assessment of the model performance should, in general, be based on a simulation analysis in order to determine the real values of population parameters. Therefore, the results obtained in the present study need to be confirmed on a larger sample. The most important factors affecting calving difficulty were sire's rank, calf sex, calving age, previous calving difficulty, lactation number, daily milk yield and farm category. After further improvement of the generated decision trees, they could be potentially applied as a decision aid to a farmer, especially that the created rules are easily interpretable.

REFERENCES

Alvasen, K., Mörk, M.J., Dohoo, I.R., Sandgren, C.H., Thomsen, P.T., Emanuelson, U. (2014). Risk factors associated with on-farm mortality in Swedish dairy cows. Prev. Vet. Med., 117(1), 110–120.

Atashi, H., Abdolmohammadi, A.R., Asaadi, A., Akhlaghi, A., Dadpasand, M., Jafari Ahangari, Y. (2012). Using an incomplete gamma function to quantify the effect of dystocia on the lactation performance of Holstein dairy cows in Iran. J. Dairy Sci., 95(5), 2718–2722.

Azizzadeh, M., Shooroki, H.F., Kamalabadi, A.S., Stevenson, M.A. (2012). Factors affecting calf mortality in Iranian Holstein dairy herds. Prev. Vet. Med., 104(3), 335–340.

Barrier, A.C., Coffey, M.P., Haskell, M.J. (2010). Effect of calving difficulty on the saleable milk yield of UK Holstein Friesian dairy cattle at different stages of lactation. Adv. Anim. Biosci., 1(1), 17.

Barrier, A.C.M. (2012). Effects of a Difficult Calving On the Subsequent Health and Welfare of the Dairy Cows and Calves. University of Edinburgh, Edinburgh.

Bazzi, H. (2010). Evaluation of non-genetic factors affecting birth weight in Sistani cattle. J. Anim. Vet. Adv., 10(23), 3095–3599.

Chang, C.L. (2007). A study of applying data mining to early intervention for developmentally-delayed children. Expert Syst. Appl., 33(2), 407–412.

Cole, J.B., Wiggans, G.R., Ma, L., Sonstegard, T.S., Lawlor Jr, T.J., Crooker, B.A., Van Tassell, C.P., Yang, J., Wang, S., Matukumalli, L.K., Da, Y. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genom., 12(1), 408–425.

Gaafar, H.M.A., Shamiah, S.M., El-Hamd, M.A., Shitta, A.A., El-Din, M.T. (2011). Dystocia in Friesian cows and its effects on postpartum reproductive performance and milk production. Trop. Anim. Health Prod., 43(1), 229–234.

Gevrekci, Y., Akbas, Y., Kizilkaya, K. (2011). Comparison of different models in genetic analysis of dystocia. J. Fac. Vet. Med. Kafkas Univ., 17, 387–392.

Ghavi Hossein-Zadeh, N. (2013). Effect of dystocia on the productive performance and calf stillbirth in Iranian Holsteins. J. Agric. Sci. Technol., 16(1), 69–78.

Hickey, J.M., Keane, M.G., Kenny, D.A., Cromie, A.R., Amer, P.R., Veerkamp, R.F. (2007). Heterogeneity of genetic parameters for calving difficulty in Holstein heifers in Ireland. J. Dairy Sci., 90(8), 3900–3908.

Ingvartsen, K.L., Dewhurst, R.J., Friggens, N.C. (2003). On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. Livest. Prod. Sci., 83(2), 277–308.

Johanson, J.M., Berger, P.J. (2003). Birth weight as a predictor of calving ease and perinatal mortality in Holstein cattle. J. Dairy Sci., 86(11), 3745–3755.

Loh, W.-Y., Shih, Y.-S. (1997). Split selection methods for classification trees. Stat. Sin., 7(4), 815–840.

Macciotta, N.P.P., Dimauro, C., Null, D.J., Gaspa, G., Cellesi, M., Cole, J.B. (2014). Dissection of genomic correlation matrices of US Holsteins using multivariate factor analysis. J. Anim. Breed. Genet., 132(1), 9–20.

McCullagh, P., Nelder, J.A. (1989). Generalized Linear Models. Second Edition. Chapman & Hall/CRC, Boca Raton.

McHugh, N., Cromie, A.R., Evans, R.D., Berry, D.P. (2014). Validation of national genetic evaluations for maternal beef cattle traits using Irish field data. J. Anim. Sci., 92(4), 1423–1432.

McHugh, N., Kearney, J.F., Berry, D.P. (2011). The effect of dystocia on subsequent performance in dairy cows. Moorepark Res. Rep., 15.

Mee, J.F. (2004) Managing the dairy cow at calving time. Vet. Clin. North Am. Food Anim. Pract., 20(3), 521–546.

Mee, J.F. (2008). Prevalence and risk factors for dystocia in dairy cattle: A review. Vet. J., 176(1), 93–101.

Mee, J.F., Berry, D.P., Cromie, A.R. (2011). Risk factors for calving assistance and dystocia in pasture-based Holstein-Friesian heifers and cows in Ireland. Vet. J., 187(2), 189–194.

Moisen, G.G. (2008) Classification and regression trees. In: S.E., Jørgensen, B.D., Fath editors. Encyclopedia of Ecology. Elsevier, Oxford (UK).

Murray, C.F., Leslie K.E. (2013) Newborn calf vitality: Risk factors, characteristics, assessment, resulting outcomes and strategies for improvement. Vet. J., 198(2), 322––328.

Piwczyński, D., Nogalski, Z., Sitkowska, B. (2013). Statistical modeling of calving ease and stillbirths in dairy cattle using the classification tree technique. Livest. Sci., 154(1––3), 19–27.

Purfield, D.C., Bradley, D.G., Kearney, J.F., Berry, D.P. (2014). Genome-wide association study for calving traits in Holstein-Friesian dairy cattle. Animal, 8(2), 224–235.

Ribeiro, E.S., Lima, F.S., Greco, L.F., Bisinotto, R.S., Monteiro, A.P.A., Favoreto, M., Ayres, H., Marsola, R.S., Martinez, N., Thatcher, W.W., Santos, J.E.P. (2013). Prevalence of periparturient diseases and effects on fertility of seasonally calving grazing dairy cows supplemented with concentrates. J. Dairy Sci., 96(9), 5682–5697.

Schuenemann, G.M., Bas, S., Gordon, E., Workman, J.D. (2013). Dairy calving management: Description and assessment of a training program for dairy personnel. J. Dairy Sci., 96(4), 2671–2680.

Speybroeck, N. (2011). Classification and regression trees. Int. J. Public Health, 57(1), 243–246.

Vanderick, S., Troch, T., Gillon, A., Glorieux, G., Gengler, N. (2014). Genetic parameters for direct and maternal calving ease in Walloon dairy cattle based on linear and threshold models. J. Anim. Breed. Genet., 131(6), 513–521.

Wautlet, R.G., Hansen, L.B., Young, C.W., Chester-Jones, H., Marx, G.D. (1990). Calving disorders of primiparous Holsteins from designed selection studies. J. Dairy Sci., 73(9), 2555–2562.

Yıldız, H., Saat, N., Simsek, H. (2011). An investigation on body condition score, body weight, calf weight and hematological profile in crossbred dairy cows suffering from dystocia. Pak. Vet. J., 31, 125–128.

KLASYFIKACJA TRUDNOŚCI WYCIELEŃ Z WYKORZYSTANIEM RÓŻNYCH RODZAJÓW DRZEW DECYZYJNYCH

Streszczenie. Celem pracy była klasyfikacja kategorii trudności wycieleń z wykorzystaniem wybranych metod eksploracji danych [drzewa klasyfikacyjne i regresyjne (CART), *Chi-square Automatic Interaction Detector* (CHAID), *Quick, Unbiased, Efficient, Statistical Trees* (QUEST)] oraz uogólnionego modelu liniowego (GLZ), a także identyfikacja najważniejszych czynników trudności porodu. Przeanalizowano łącznie 1699 rekordów informacyjnych krów rasy polskiej holsztyńsko-fryzyjskiej odmiany czarno-białej. Wyodrębniono trzy kategorie trudności wycielenia (łatwe, średnie i trudne). Odsetek poprawnie zaklasyfikowanych wycieleń przez CART, CHAID, QUEST i GLZ wynosił odpowiednio: 60,20, 65,31, 68,88, 66,33% (wycielenia łatwe), 71,36, 69,01, 64,79%, 69,01% (umiarkowane) oraz 0, 0, 0, 0% (trudne). Najważniejszymi predyktorami trudności porodu były: ranga buhaja – ojca krowy, płeć cielęcia, wiek wycielenia, trudność poprzedniego porodu, kolejna laktacja, dobowa wydajność mleka i kategoria średniej wydajności mleka w gospodarstwie. Drzewa decyzyjne i GLZ charakteryzowały się umiarkowaną jakością. Żaden z modeli nie był w stanie poprawnie wskazać trudnych wycieleń.

Słowa kluczowe: bydło, systemy wspomagania decyzji, diagnoza, trudny poród

Accepted for print: 28.12.2016

This work was supported by the Polish Ministry of Science and Higher Education (grant number 517-01-028-3962/17).

Part of the manuscript was released as a conference abstract in the Proceedings of the Conference "The contribution of the natural sciences to the development of politics of sustainable development of agriculture" held in Szczecin, Poland, June 17–18, 2015.

For citation: Zaborski, D., Proskura, W.S., Grzesiak, W. (2016). Classification of calving difficulty scores using different types of decision trees. Acta Sci. Pol. Zootechnica, 15(4), 55–70. DOI: 10.21005/asp.2016.15.4.05