Animal Science Papers and Reports vol. 29 (2011) no. 4, 293-302
Institute of Genetics and Animal Breeding, Jastrzębiec, Poland

Detection of difficult conceptions in dairy cows using selected data mining methods

Wilhelm Grzesiak*, Daniel Zaborski, Piotr Sablik, Renata Pilarczyk

Department of Ruminants Science, West Pomeranian University of Technology,
Doktora Judyma 10, 71-460 Szczecin, Poland

(Received March 9, 2010; accepted September 2, 2011)

In the present study, the detection of difficult conceptions in dairy cows using selected data mining methods – naïve Bayes classifier (NBC) and regression and classifications trees (CART) is presented. The set of 11 diagnostic variables was used, which included, among others, number of lactation, artificial insemination (AI) season, age of inseminated cow, proportion of HF-genes in cow genotype, sex of calf from preceding calving, length of pregnancy, milk, protein and fat yield. Two conception classes were distinguished: the GOOD class, if a cow conceived after one or two AIs and the "POOR" class, if more than two AIs per conception were required. The models were characterized by capability of predicting the membership of conceptions to either class. Correctness of predictions was 83%. CART proved to be more precise in detecting conceptions of the POOR class (sensitivity) compared with predictions by NBC (P≤0.01). Specificity was similar for both classifiers (90% and 93%). Among the variables determining conception class, calving-to-conception interval, calving interval and the difference between the mean body condition and condition at AI were the most significant variables for CART. Utilization of these classifiers, particularly of CART, may help a breeder to appropriately prepare cows for AI, thus contributing to the improved financial results of a herd.

KEY WORDS: conception / problematic cow detection / naïve Bayes classifier / classification and regression trees

Data mining is an advanced method of data analysis enabling "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [Frawley *et al.* 1992]. This information can then be used, among others, for prediction of the membership of objects (e.g. animals) to the distinguished class (e.g.

*corresponding author: wilhelm.grzesiak@zut.edu.pl ill animals, animals in estrus) on the basis of a set of previously accessible diagnostic variables. An example of such a procedure may be the detection of cows with problems at conception. The efficiency of artificial insemination (AI) depends on a number of factors such as: the quality and quantity of bull semen, cow age, cow condition, milk yield and many others [Nebel and McGilliard 1993, Domecq et al. 1997, Jaśkowski and Szenfeld 1999] whereas unsuccessful AIs cause notable economic losses [Sørensen and Østergaard 2003]. Therefore, disregarding the correctness of performing the AI itself or estrus detection, it seems justified to eye animals which can cause problems at AI. Such a possibility is offered by, among others, naïve Bayes classifier (NBC) and regression and classification trees (CART), which belong to the data mining methods.

NBC is a technique utilizing the Bayes' theorem on the conditional probability. Assignment of the object (an individual) to a specific, most probable class is done through the calculation of the probability which is a product of the probabilities estimated on the basis of data. These probabilities are selected depending on the values of the variables of the classified object [Larose 2006]. The variables describing a given individual may be both discrete (e.g. number of lactation) and continuous (e.g. milk yield). This classifier is used in animal breeding, e.g., for classification of the estrus cycle phase [Maldonado-Castillo *et al*. 2007].

The decision tree is a graphical method of decision process support. It is a set of decision nodes connected by means of branches going down from root node to the terminal leaf nodes [Piwczyński 2009]. CARTs are one of the examples of decision trees. In these trees, only two child nodes can branch directly from one decision node. The tree is constructed by repeated divisions of the training dataset (e.g. conceptions records). These divisions are defined by a simple rule based on a single explanatory variable. At each stage, records are divided into two disjoint subsets, each of which is as homogeneous as possible. The procedure of division is then repeated for each subset separately. The manner in which explanatory variables define the division depends on their nature. In the case of categorical variables with two categories (e.g. individual's sex) only one division is possible, and each of the two categories defines a subset. For numeric variables (e.g. age, calving-to-conception interval), the division is made based on the value of such a variable (whether it is greater or smaller than a defined threshold, e.g. calving-to-conception interval ≤135.5 days or >135.5 days). The size of the tree depends on the total number of distinguished subsets [De'ath and Fabricius 2000]. The unquestionable advantages of CARTs are as follows: lack of assumptions concerning the distribution of explanatory variables, possibility of analyzing categorical and continuous variables, insensitivity to outliers, collinearity and unequal variances of the analyzed variables and, finally, the possibility of including interactions among them [Tittonell *et al*. 2008]. Due to these advantages, CARTs are more and more frequently applied in agricultural sciences e.g. to the analysis of an effect of soil parameters and farm management on the yield of maize [Tittonell *et al*. 2008], to the identification of factors influencing the variability of the number of offspring reared by ewes [Piwczyński 2009] or to the analysis of lactation curves in cattle [Pietersma *et al*. 2002].

In various software used for farm management, it would be advisable to apply a method enabling indication of cows with potential difficulties at conception. Hence, the main aim of this study was to verify ability of NBC and CART to detect cows that may have conception problems with respect to their practical application by a breeder.

Material and methods

The study involved 1006 complete records from 594 cows containing data on the number of AIs per conception of the Polish Holstein-Friesian cows, during two production seasons (2007-2009). The cows were kept in a loose barn with an outside run accessible over the whole year. Animals were fed TMR and milked twice a day in a herringbone milking parlor. Cows in which no serious disorders (metritis, ovarian cysts, ketosis, lameness, and clinical mastitis) were found during the period of AI were included in the statistical analyses. The data (1006 records) were randomly divided into training set L (812 records) used to prepare NBC and CART and test set T (194 records) for the verification of model detection abilities. In both sets, the numbers of lactations were proportional. The test set comprised records not included during the preparation of NBC and CART. Selected performance parameters of the cows are presented in Table 1.

Table 1. Characteristics of the animal data sets (standard deviations in parentheses)

Set	n	HF% (%)	AGE (months)	CLVC (days)	CLVI (days)	PREG (days)	IBCS (points)	FCM (kg)	FTPR% (%)
L	812	85.7 (9.5)	38.6 (12.2)	143 (66)	418 (53)	280 (7)	0.12 (0.33)	9965 (2168)	7.06 (0.20)
T	194	85.9 (10.6)	41.01 (12.6)	142 (61)	420 (54)	280 (6)	0.07 (0.32)	10253 (2277)	7.01 (0.20)

L – training set; T – test set; n – number of records; HF% – proportion of HF-genes in cow genotype; AGE – age of inseminated cow; CLVC – calving-to-conception interval; CLVI – calving interval; PREG – length of pregnancy; IBCS – cow body condition index; FCM – 4% fat-corrected milk; FTPR% – mean milk fat and protein content.

Two discrete variables were chosen as predictors: X_1 – LAC – number of lactation prior to conception (from 2 to 4 inclusive); X_2 – SEASON – AI season (autumn-winter and spring-summer), and the following continuous variables: X_3 – HF% – proportion of HF-genes in cow genotype (%); X_4 – AGE – age of inseminated cow (months); X_5 – SEX – sex of the calf born as a result of preceding AI (denoted as a heifer – 0, bull – 0.5, two heifers – 1, heifer and a bull – 1.5, two bulls – 2); X_6 – CLVC - calving-to-conception interval for preceding conception (days); X_7 – CLVI – calving interval for preceding conception (days); X_8 – PREG – length of pregnancy (days);

X_9 – IBCS – cow body condition index calculated as a difference between the average cow condition assessed at calving and during the previous production season and the condition assessed at AI. The condition itself was determined on a 5-point scale, on which emaciated cows were scored 1 and obese ones were scored 5 with an increment of 0.25 point [Ferguson *et al*. 1994]. The scores were then modified so that the classifier correctly interpreted the fact that both too high and too low body condition is disadvantageous to AI. Optimum was set at 3.50 points and, at higher values, the multiple of 0.25 was deducted from 3.50, e.g. if a cow obtained 3.75 points, it was recorded as 3.25 points. Other variables were: X_{10} – FCM – 4% fat-corrected milk calculated according to the formula: FCM = 0.4 MILK+15 FAT (where MILK – actual milk yield during the preceding lactation (kg), FAT – fat yield (kg)); X_{11} – FTPR% - mean milk fat and protein content (%). When selecting the predictors defining the milking capacity of cows, the FCM variable and the sum of milk fat and protein content (FTPR %) were used in order to prevent the occurrence of collinearity.

It was stated on the farm that estrus detection (based on the observation of animals and pedometers readings) was the basis for performing AI by the same experienced inseminator. The included predictors constitute, so to speak, an additional element discriminating the class of conception in a cow (output variable Y).

Taking into account the fact that the optimum number of AIs per conception for the Polish Holstein-Friesian cows is 1.6 [Januś and Borkowska 2006], the division of conceptions into two classes was adopted:

1. The GOOD class – cows conceived after one or two AIs,
2. The POOR class – cows conceived after more than two AIs (from 3 to 13).

The distribution of conceptions according to the class in the training and test sets is presented in Table 2.

Table 2. Distribution of observed conceptions in the sets (L – training set; T – test set)

Set	GOOD	POOR
L	519 (63.9%)	293 (36.1%)
T	134 (69.1%)	60 (30.9%)

The *a priori* probability $P(X_1,..., X_{11}|Y)$ of the membership of conception record to one of the two distinguished classes for NBC, calculated from the previously collected observations, can be presented on the basis of the Bayes' theorem as:

$$P(X_1, ..., X_{11} \mid Y) = \prod_{k=1}^{11} P(X_k \mid Y),$$

where Y is an analyzed output variable (class of conception), $X_1 \ldots X_{11}$ are selected predictors (explanatory variables independent from each other) and $\prod_{k=1}^{11} P(X_k \mid Y)$ denotes the product of distributions of individual predictors for the cases when the output variable Y assumes one of its values (GOOD or POOR). The probability density functions for continuous variables were estimated on the basis of the training data, after verification of the normality of their distribution. New cases, or conception records from the test set, were classified according to the maximum *a posteriori* probability $P(Y \mid X_1,..., X_{11})$:

$$P(Y \mid X_1, ..., X_{11}) = P(Y) \prod_{k=1}^{11} P(X_k \mid Y),$$

where P(Y) denotes the probability of the occurrence of conception record from a given class.

In the preparation of CART, the Gini index (G_I) was applied [Hastie *et al.* 2001] according to the following formula:

$$G_I = \sum_{i=1} p_{mi}(1 - p_{mi}),$$

where p_{mi} – proportion of the observations (conception records) from class *i* (*i* =GOOD or POOR) in a given tree node *m*.

The *a priori* probability was estimated from the training sample assuming equal costs of misclassification. The criterion of algorithm stopping was determined according to the minimum size in the leaf node. The multiple (10-fold) cross-validations were also applied, which consisted in the division of the training set into 10 equal subsets with a random selection of conception records, from which 9 subsets were used to prepare the trees and one subset was used to verify their prognostic abilities. Each time a different subset served as the test set.

The quality of the classification effects was determined by *sensitivity* – conditional true positive probability – P_{TP}, and *specificity* – conditional true negative probability – P_{TN} defined as:

$$P_{TP} = \frac{A}{A + C}\ , \quad P_{TN} = \frac{D}{B + D}$$

where: *A* – number of correctly classified conceptions of the POOR class, *C* – number of conceptions classified as GOOD but in fact belonging to the POOR class, *B* – number of conceptions classified as POOR but in fact belonging to the GOOD class, *D* – number of correctly classified conceptions of the GOOD class (Tab. 3).

Table 3. Observed and predicted conception classes

Observed conception classes	Predicted conception classes		Total
	POOR	GOOD	
POOR	A	C	A+C
GOOD	B	D	B+D
TOTAL	A+B	C+D	A+B+C+D

A measure describing the general probability of error (i.e. the probability of true positive or true negative answers P_G):

$$P_G = \frac{A + D}{A + B + C + D}$$

and the percentage of misclassifications (P_B): P_B=1-P_G were also applied.

When comparing the quality of the analysed models, *AIC* (Akaike information criterion) in a modified form recommended in the case of lower number of observations, when $n/k \leq 40$, by Sugiura [1978] was determined:

$$AIC = \chi^2 + 2k + \frac{2k(k+1)}{n-k-1} \text{ , where}$$

$$\chi^2 = \sum_{i=1}^{2} \frac{(E_i - O_i)^2}{E_i}$$

The G^2 measure, similar to the maximum likelihood χ^2 statistic, was also calculated according to the following formula:

$$G^2 = 2\sum_{i=1}^{2} O_i \cdot \ln\left(\frac{O_i}{E_i}\right) \text{ , where}$$

E_i – number of conception records assigned to class *i* (GOOD or POOR) by the model, O_i – number of conception records in class *i*, *k* – number of model parameters, *n* – number of conception records.

All the calculations were done using Statistica Miner software [2007].

Results and discussion

For the training set, it was found that the percentage of misclassifications (Tab. 4) for NBC (17%) was statistically significantly higher than that for CART (12%). Maldonado-Castillo *et al*. [2007] obtained, in classification of estrus states in cows, just under 14% of misclassifications for NBC, whereas the percentage of misclassifications fell to 0 when using the classification trees.

Sensitivity for NBC (Tab. 4) was statistically significantly lower than that for CART, as opposed to specificity, which was higher in the case of NBC. In a similar study [Grzesiak *et al*. 2009] on the MARS method and classification functions, comparable values of sensitivity and specificity for classification functions and NBC as well as for the MARS method and CART were found. Morrison *et al*. [1985a], analyzing the detection of difficult calvings, obtained sensitivity in the range 0.60 – 0.86 at specificity of 0.81 – 0.94. In the study on mastitis detection in cows, White *et al*. [1986] found lower values of sensitivity and specificity (0.64 and 0.61,

Table 4. Sensitivity (P_{TP}) and specificity (P_{TN}) values, general probability of error (P_G) and percentage of misclassifications for NBC and CART (L – training set; T – test set)

Item	Set	n	P_{TP}	P_{TN}	P_G	Percentage of misclassifications
NBC	*L*	812	0.65^A	0.93^A	0.83^A	17
	T	194	0.72^A	0.90	0.85	15
CART	*L*	812	0.89^A	0.87^A	0.88^A	12
	T	194	0.83^A	0.86	0.90	10

^{A}P≤0.01; NBC – naïve Bayes classifier, CART – classification and regression trees.

Table 5. Qualitative measures of the classifiers – G^2 criterion and *AIC*

Item	NBC	CART
G^2	424.98	253.20
AIC	89.73	36.24

AIC – Akaike information criterion, NBC – naïve Bayes classifier, CART – classification and regression trees.

respectively), whereas Thirunavukkarasu and Kathiravan [2006], studying successful conception in cows and buffaloes by means of logistic regression, obtained very high values of these probabilities exceeding 0.98, which resulted from the appropriate selection of the specific predictors (explanatory variables). It should be emphasized that the correct detection of cows with difficult conception (sensitivity) is definitely more advantageous for a breeder, and classification trees proved to be a much better tool, in the present study. The *AIC* and G^2 quality criteria, which are lower in the case of CART, also favour this classifier (Tab. 5).

The general probability of error (Tab. 4) was statistically significantly higher for CART compared with the value for NBC (0.88 and 0.83), and was similar to results obtained in previous studies (0.86 for MARS and 0.82 for classification functions: [Grzesiak *et al*. 2009]). A very similar value (0.87) was obtained by Morrison *et al*. [1985a], who classified dystocia in beef cattle using classification functions.

The classifiers prepared in this way were used to detect difficult conceptions on the basis of the data inaccessible during their preparation (test set). The prediction of conceptions of the POOR class (P_{TP} - sensitivity) for NBC was statistically significantly lower than that for CART, which in practice was an equivalent of 17 cases (out of 60) which were not considered difficult by NBC, whereas CART did not consider 10 such cases (Tab. 6). Specificity values for NBC and CART did not differ statistically significantly and were similar to those obtained by Morrison *et al*. [1985a]. In another study, Morrison *et al*. [1985b], predicting dystocia in cows using classification

Table 6. Classification matrix for NBC and CART

Observed conception classes	Predicted conception classes			
	POOR		GOOD	
	NBC	CART	NBC	CART
	Training set			
POOR	191	262	102	31
GOOD	38	69	481	450
	Test set			
POOR	43	50	17	10
GOOD	13	19	121	115

NBC – naïve Bayes classifier, CART – classification and regression trees.

functions, obtained higher sensitivity (0.87) than specificity (0.81), however, they used the same data to construct the model and then to verify its predictive performance. The schematic diagram of the classification tree presented in Fig. 1 had three split nodes and four leaf nodes. A key role here was played by the CLVC variable, constituting the main decision node of the tree (CLVC>135.5) and then the next nodes with the CLVI variable (CLVI>508.5) as well as IBCS variable (IBCS>-0.075), which may prove that problems with AIs are also a decrease in condition during the period of AI in

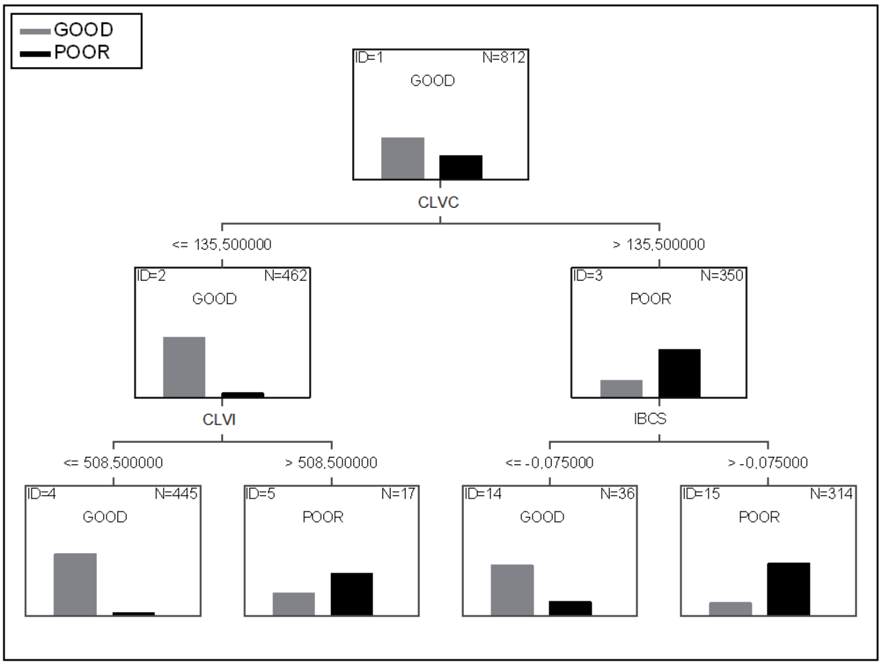

Fig. 1. Diagram of the CART decision tree for conceptions.

relation to the condition during the previous production season [Domecq *et al*. 1997], and not necessarily the lower condition during AI. It was not possible to capture the representation of such a hierarchy of predictors for NBC.

Opinions on the effect of body condition during AI on its effectiveness vary [Grzesiak *et al*. 2010]. On the one hand, it is stated that both too high (over 3.3 points) and too low (below 2.5 points) body condition at AI is associated with problems at conception, which, in the latter case, is additionally supported by the fact that cows after calving suffer more from inactive ovaries. On the other hand, reports exist, in which it is claimed that body condition of cows after calving does not influence the difficulties at AI.

To sum up, the analyses performed showed that the applied classifiers, and CART in particular, exhibited above average capabilities of detecting difficult conceptions in dairy cattle. The CART method, in comparison with NBC, more precisely detected cows with difficult conceptions and more conservatively treated cows which, theoretically, should not have had any problems at AIs. This is an advantage because inconsiderate indication of cows without such problems may appear unreliable under breeding conditions. Since a certain part in the detection of difficult conceptions using CART was played by the body condition index, it may be suggested that breeders assess the condition of cows and compare it with the condition during the production period in order to more precisely detect cows with an appropriate category of conception. The utilization of CART (through the implementation of an appropriate algorithm in on-farm software) may constitute a support for making prophylactic efforts (paying particular attention to the proper time and procedure of AI) directed at cows which may have problems with conception, which in turn may increase fertility in a herd and consequently improve economic effectiveness of a farm.

REFERENCES

1. DE'ATH G., FABRICIUS K.E., 2000 – Classifications and regression trees: a powerful yet simple technique for ecological data analysis. ***Ecology*** 81, 3178-3192.
2. DOMECQ J.J., SKIDMORE A.L., LLOYD J.W., KANEENE J.B., 1997 – Relationship between body condition scores and conception at first artificial insemination in a large dairy herd of high yielding Holstein cows. ***Journal of Dairy Science*** 80, 113-120.
3. FERGUSON J.D., GALLIGAN D.T., THOMSEN N., 1994 – Principal descriptors of body condition in Holstein dairy cattle. ***Journal of Dairy Science*** 77, 2695-2703.
4. FRAWLEY W., PIATETSKY-SHAPIRO G., MATHEUS C., 1992 – Knowledge Discovery in Databases: An Overview. ***AI Magazine*** 13, 57-70.
5. GRZESIAK W., SABLIK P., ZABORSKI D., ŻUKIEWICZ A., DYBUS A., SZATKOWSKA I., 2009 – Zastosowanie metody MARS do klasyfikowania zabiegów inseminacyjnych u bydła mlecznego (Application of MARS method in classification of inseminations of dairy cattle). In Polish, summary in English. ***Roczniki Naukowe Polskiego Towarzystwa Zootechnicznego*** 5, 43-56.
6. GRZESIAK W., ZABORSKI D., SABLIK P., ŻUKIEWICZ A., DYBUS A., SZATKOWSKA I., 2010 – Detection of cows with insemination problems using selected classification models. ***Computers and Electronics in Agriculture*** 74, 265- 273.

7. HASTIE T., TIBSHIRANI R., FRIEDMAN J., 2001 – The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics), 271.
8. JANUŚ E., BORKOWSKA E., 2006 – Wielkość podstawowych wskaźników płodności u krów o różnej wydajności mlecznej (Selected indices of fertility of cows of different milk production). In Polish, Summary in English. ***Annales Universitatis Mariae Curie-Skłodowska Lublin – Polonia.*** *Sectio EE* 24, 33-37.
9. JAŚKOWSKI J.M., SZENFELD J., 1999 – Wpływ ilości i jakości nasienia oraz techniki unasienniania na wyniki zacieleń krów (The influence of the quantity and quality of semen and insemination techniques on results of pregnancies in cows). In Polish, Summary in English. ***Medycyna Weterynaryjna*** 55, 160-162.
10. LAROSE D. T., 2006 – Data Mining Methods and Models. John Wiley & Sons, Inc. Hoboken, 228-252.
11. MALDONADO-CASTILLO I., ERAMIAN M.G., PIERSON R.A., SINGH J., ADAMS G.P., 2007 – Classification of bovine reproductive cycle phase using ultrasound-detected features. Fourth Canadian Conference on Computer and Robot Vision (CRV 2007) 28-30 May 2007, Montreal, Quebec, Canada, pp. 258-265.
12. MORRISON D.G., HUMES P.E., KEITH N.K., GODKE R.A., 1985a – Discriminant analysis for predicting dystocia in beef cattle. I. Comparison with regression analysis. ***Journal of Animal Science*** 60, 608-616.
13. MORRISON D.G., HUMES P.E., KEITH N.K., GODKE R.A., 1985b – Discriminant analysis for predicting dystocia in beef cattle. II. Derivation and validation of a prebreeding prediction model. ***Journal of Animal Science*** 60, 617-621.
14. NEBEL R.L., MCGILLIARD M.L., 1993 – Interactions of high milk yield and reproductive performance in dairy cows. ***Journal of Dairy Science*** 76, 3257-3268.
15. PIETERSMA D., LACROIX R., LEFEBVRE D., WADE K.M., 2002 – Decision-tree induction to interpret lactation curves. ***Canadian Biosystems Engineering/Le génie des biosystemes au Canada*** 44, 7.1-7.13.
16. PIWCZYŃSKI D., 2009 – Using classification trees in statistical analysis of discrete sheep reproduction traits. ***Journal of Central European Agriculture*** 10, 303-310.
17. SØRENSEN J.T., ØSTERGAARD S., 2003 – Economic consequences of postponed first insemination of cows in a dairy cattle herd. ***Livestock Production Science*** 79, 145–153.
18. STATSOFT, INC. (2007). STATISTICA (data analysis software system), version 8.0.
19. SUGIURA N., 1978 – Further analysis of the data by Akaike's information criterion and the finite corrections. ***Communications in Statistics – Theory and Methods*** A7, 13-26.
20. THIRUNAVUKKARASU M., KATHIRAVAN G., 2006 – Predicting the probability of conception in artificially inseminated bovines – A logistic regression analysis. ***Journal of Animal and Veterinary Advances*** 5, 522-527.
21. TITTONELL P., SHEPHERD K.D., VANLAUWE B., GILLER K.E., 2008 – Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—An application of classification and regression tree analysis. ***Agriculture, Ecosystems and Environment*** 123, 137–150.
22. WHITE M.E., GLICKMAN L.T., BARNES-PALLESEN F.D., STEM E.S. 3RD, DINSMORE P., POWERS M.S., POWERS P., SMITH M.C., MONTGOMERY M.E., JASKO D., 1986 – Accuracy of a discriminant analysis model for prediction of coliform mastitis in dairy cows and a comparison with clinical prediction. ***Cornell Veterinarian*** 76, 342-347.