ORIGINAL ARTICLE

Osong Public Health Res Perspect 2017;8(3):195–200
https://doi.org/10.24171/j.phrp.2017.8.3.06
eISSN 2233-6052

Prevalence and Determinants of Preterm Birth in Tehran, Iran: A Comparison between Logistic Regression and Decision Tree Methods

Payam Amini[a], Saman Maroufizadeh[a], Reza Omani Samani[a], Omid Hamidi[b], Mahdi Sepidarkish[a]

[a]Department of Epidemiology and Reproductive Health, Reproductive Epidemiology Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran
[b]Department of Science, Hamadan University of Technology, Hamadan, Iran

Objectives: Preterm birth (PTB) is a leading cause of neonatal death and the second biggest cause of death in children under five years of age. The objective of this study was to determine the prevalence of PTB and its associated factors using logistic regression and decision tree classification methods.
Methods: This cross-sectional study was conducted on 4,415 pregnant women in Tehran, Iran, from July 6–21, 2015. Data were collected by a researcher-developed questionnaire through interviews with mothers and review of their medical records. To evaluate the accuracy of the logistic regression and decision tree methods, several indices such as sensitivity, specificity, and the area under the curve were used.
Results: The PTB rate was 5.5% in this study. The logistic regression outperformed the decision tree for the classification of PTB based on risk factors. Logistic regression showed that multiple pregnancies, mothers with preeclampsia, and those who conceived with assisted reproductive technology had an increased risk for PTB ($p < 0.05$).
Conclusion: Identifying and training mothers at risk as well as improving prenatal care may reduce the PTB rate. We also recommend that statisticians utilize the logistic regression model for the classification of risk groups for PTB.

Key Words: preterm birth, risk factor, infant, logistic regression, decision tree

Corresponding author:
Reza Omani Samani
E-mail: samani@royaninstitute.org

Received April 10, 2017.
Accepted May 17, 2017.

INTRODUCTION

Preterm birth (PTB) is defined by the World Health Organization as 'any birth before 37 completed weeks of gestation, or fewer than 259 days since the first day of the women's last menstrual period (LMP)'; PTB can also be subdivided on the basis of gestational age: extremely preterm (< 28 weeks), very preterm (28 to < 32 weeks), and moderate to late preterm (32 to < 37 weeks) [1]. It is a leading cause of infant mortality and is the second largest cause of child deaths in children under the age of 5 years [2]. PTB has lifelong impacts on neurodevelopmental functioning, including increased risk of impaired learning, cerebral palsy, and visual disorders. PTB also increases the risk for chronic disease in adulthood [3,4]. An estimated 15 million preterm neonates were born in 2010 worldwide, representing, on average, 11.1% of all live births, ranging from about 5% in several European countries to 18% in some African countries [4].

The rates of PTB have risen in most countries in the past two to three decades, despite advancing knowledge of the possible risk factors and its pathological mechanisms. This trend may be due to the increasing rates of multiple births, greater use of assisted reproductive technology (ART), and more obstetric interventions [4,5]. The known risk factors for PTB include multiple pregnancies, infection, young or advanced maternal age, short interval between pregnancies, low maternal body mass index (BMI) before pregnancy, poor maternal nutrition, use of ART, maternal psychological health, and lifestyle factors such as alcohol consumption, smoking, and excessive physical work [6–8].

Classification methods can be used to classify new births as PTB or non-PTB based on several risk factors and covariates. For instance, acute kidney injury as a predictable outcome can be managed if the early risk factors are identified using classification methods [9]. Several methods have been introduced and evaluated, including data mining (machine learning) techniques [10]. To determine the best classifier methods, several tools are available for accuracy assessment. One technique splits the training and testing sets by using two-thirds of the sample for training. To do so, the training dataset determines the best-fitting model. The testing sample later evaluates the resulting model [11]. Several statistics, such as accuracy, sensitivity, and specificity, can be used to compare the performance of different methods.

Several classification methods can be used to classify a new case into one of the response categories, including logistic regression (LR), decision trees (DTs), artificial neural networks, genetic algorithms, and so on. Among the different classification methods and based on its ease of use and interpretation, LR is the most popular parametric approach to classify discrete response variables using several factors and covariates. However, the nonparametric DT is preferable when the subjects are described through a predetermined set of attributes, the response variable is binary or multinomial, and disjunctive results are needed [12].

It is important to predict PTB due to its potential for adverse long-term consequences. Therefore, this study was conducted to identify the high-risk groups for PTB based on several factors and covariates in order to reduce the risk of PTB. To do so, we performed and compared the results of LR and DT classification methods.

MATERIALS AND METHODS

1. Participants and study design

This cross-sectional study was carried out on 4,419 pregnant women referred to maternity hospitals across Tehran province, affiliated with Tehran University of Medical Sciences, Shahid Beheshti University of Medical Sciences, Iran University of Medical Sciences, or Islamic Azad University between July 6 and 21, 2015.

2. Ethical approval

This study was approved by the Ethics Committee of Royan Institute, Tehran, Iran. The aim and objective of the study and the confidentiality of the data were explained verbally to the women by nurses and midwives prior to their participation. Written informed consent was obtained from all participants before completing the measures.

3. Measures

A checklist was used for data collection, which contained the mother's demographic information, obstetrical data, and newborn's information. The checklists were completed during a direct interview of the mothers and a review of their cases in the delivery room by a nurse or trained obstetrician. The checklists contained information such as mother's age (years), education (academic, non-academic), and occupation (housewife, employed), socio-economic status, BMI (kg/m^2), pregnancy type (wanted, unwanted), type of delivery (natural, Cesarean), preeclampsia (no, yes), history of abortion (no, yes), history of stillbirth (no, yes), history of multiple pregnancies (no, yes), and use of ART (no, yes). These variables were used to classify PTB. The criterion for PTB was gestational age of fewer than 37 weeks of pregnancy after the LMP, while the criterion for preeclampsia was having a blood pressure reading of more than 140/90 mmHg and the presence of an excess of proteins in the urine (proteinuria).

4. Statistical analysis

The dataset was randomly divided into two subsamples. Model fitting was carried out using the training dataset (70% of cases). The resulting models were then assessed using the test sample (30% of cases). The LR and DT methods were fitted to the data.

LR: The most common parametric tool to model binary outcomes is LR. The model can be written as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

In this model, x_i's are the covariates or factors and the α and β_i's are "k+1" regression coefficients stating the measure of the size effect. The response variable can take two values (0 as non-PTB and 1 as PTB). The term, $\frac{\pi}{1-\pi}$, indicates the odds of classifying the response in category one than zero.

DT: One of the most popular and appealing methods in classification and discriminant analysis is DT. This method can be performed on most medical data for the purpose of diagnosis and prediction [13,14]. Subjects can also be identified by DT as high-risk based on their characteristics [15]. A DT contains three main parts: decision nodes, branches, and leaves. The tree starts with a node and extends to the leaf. The risky paths are identified and shown in several nodes. The main characteristic of DT is the graphical display of the choices. This advantage provides alternatives for each decision and possible outcomes and allows comparisons of different alternatives [15].

The accuracy of the classifications was checked by indices such as sensitivity, specificity, diagnostic accuracy, positive predictive value, negative predictive value, and area under the curve (AUC). McNemar's test was used to check for differences in proportions between two performed methods. To evaluate the reliability of the predictions, the Kappa statistic was calculated. All statistical analyses were performed using R version 3.2.3 (http://www.R-project.org). All statistical tests were two-tailed and a *p*-value less than 0.05 was considered statistically significant.

Table 1. Demographic and clinical characteristics of the participants

Variable	PTB (n = 244)	Non-PTB (n = 4,171)	*p*-value
Mother's age (y)	30.51 ± 5.96	29.10 ± 5.31	< 0.001
SES	0.17 ± 2.11	0.022 ± 2.03	0.272
Mother's BMI (kg/m^2)	25.00 ± 4.13	24.99 ± 5.61	0.970
Parity	1.65 ± 0.78	1.65 ± 0.76	0.989
Mother's education			0.035
Non-academic	149 (61.1)	2,820 (67.6)	
Academic	95 (38.9)	1,351 (32.4)	
Mother's occupation			0.314
Housewife	209 (85.7)	3,666 (87.9)	
Employed	35 (14.3)	505 (12.1)	
Type of pregnancy			0.617
Wanted	194 (79.5)	3,369 (80.8)	
Unwanted	50 (20.5)	802 (19.2)	
History of abortion			0.243
No	190 (77.9)	3,373 (80.9)	
Yes	54 (22.1)	798 (19.1)	
History of stillbirth			0.199
No	237 (97.1)	4,101 (98.3)	
Yes	7 (2.9)	70 (1.7)	
Infant sex			0.131
Male	136 (55.7)	2,115 (50.7)	
Female	108 (44.3)	2,056 (49.3)	
Caesarian section			0.022
No	52 (21.3)	1,167 (28.0)	
Yes	192 (78.7)	3,004 (72.0)	
Multiple pregnancy			< 0.001
No	210 (86.1)	4,143 (99.3)	
Yes	34 (13.9)	28 (0.7)	
Preeclampsia			< 0.001
No	198 (81.1)	3,982 (95.5)	
Yes	46 (18.9)	189 (4.5)	
ART			< 0.001
No	197 (80.7)	3,886 (93.2)	
Yes	47 (19.3)	285 (6.8)	

Values are presented as mean ± standard deviation or number (%).
PTB, preterm birth; SES, socioeconomic status; BMI, body mass index; ART, assisted reproductive technology.

RESULTS

Of the 4,415 births included in the study, 244 (5.5%) were PTB. Of the participants, 67.2% were non-academic, 87.8% were housewives, 80.7% had wanted pregnancies, 19.3% had a history of abortion, 1.7% had a history of stillbirth, 72.4% had Caesarian section delivery, 1.4% had multiple pregnancies, 5.3% had preeclampsia, and 7.5% have conceived via ART. The mean age of the mothers was 29.18 ± 5.35 years. Table 1 shows the descriptive characteristics of the mothers as well as the comparison between the PTB and non-PTB groups. The mother's age was significantly higher in the PTB group ($p < 0.05$). Moreover, educated mothers, those who underwent Caesarian section delivery, had multiple pregnancies, had preeclampsia, and had used ART were statistically associated with the incidence of PTB.

The LR model used variable to fit the training data. After a stepwise variable selection, mother's age, BMI, multiple pregnancy, preeclampsia, and ART were identified as significant variables affecting PTB. In order to test the resulting variables, the

Table 2. The results of logistic regression assessing PTB based on mothers' characteristics

Variable	AOR (95% CI)	*p*-value
Age	1.00 (0.96–1.05)	0.889
BMI	0.99 (0.94–1.04)	0.648
Multiple pregnancy	28.63 (10.45–78.42)	< 0.001
Preeclampsia	4.42 (2.12–9.18)	< 0.001
ART	3.23 (1.69–6.19)	< 0.001

PTB, preterm birth; AOR, adjusted odds ratio; CI, confidence interval; BMI, body mass index; ART, assisted reproductive technology.

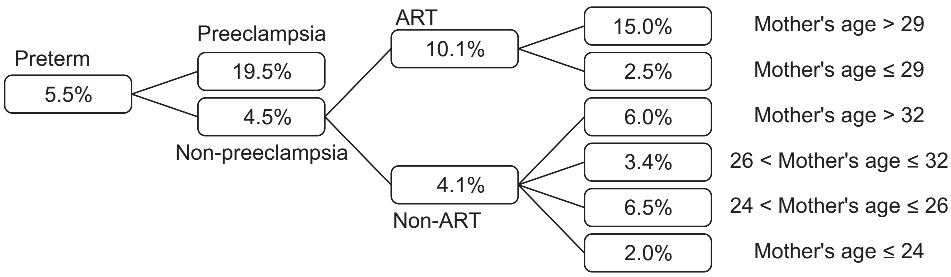

Figure 1. Decision tree results for the evaluation of preterm birth based on mothers' characteristics.
ART, assisted reproductive technology.

Table 3. Accuracy measures of logistic regression and decision tree methods in training and testing subsamples

Model	Training sample		Testing sample	
	LR	DT	LR	DT
Sensitivity	0.46	0.69	0.41	0.57
Specificity	0.83	0.59	0.88	0.59
Positive predictive value	0.13	0.09	0.18	0.08
Negative predictive value	0.96	0.97	0.96	0.96
Accuracy	0.80	0.59	0.85	0.59

LR, logistic regression; DT, decision tree.

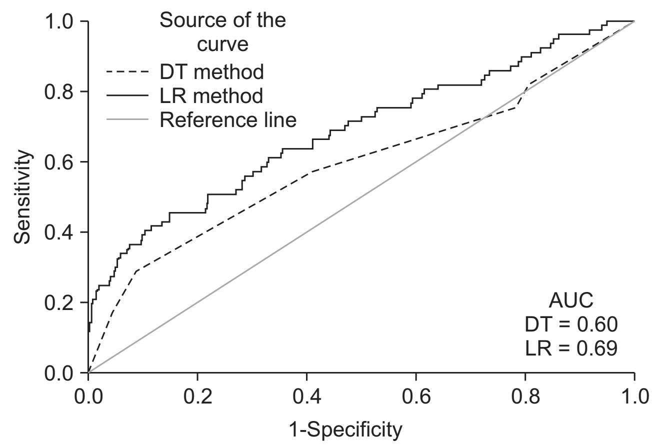

Figure 2. Area under the curve analysis of the logistic regression and decision tree methods.
DT, decision trees; LR, logistic regression; AUC, area under the curve.

LR model was fitted using the testing subsample. The results are shown in Table 2. Mother's age and BMI were not associated with PTB. The findings show that multiple pregnancies were a significant predictor for PTB (odds ratio [OR] = 28.63, 95% confidence interval [CI]: 10.448–78.42). Mothers with preeclampsia (OR = 4.42, 95% CI: 2.124–9.18) and those who conceived using ART (OR = 3.23, 95% CI: 1.69–6.19) had an increased risk for PTB.

The results of the DT method showed that mother's age, preeclampsia, and use of ART were the most important variables for the classification of PTB. The details about the rules and the proportion of PTB in several resultant nodes are shown in Figure 1.

Table 3 shows the comparison of sensitivity, specificity, positive probability value, negative probability value, accuracy, and the AUC for the training and testing sets of classification methods. The LR model resulted in a higher accuracy of predictions compared to that of DT. The accuracy of LR for classifying PTB was 0.85, significantly different from that of the DT method. McNamara's test showed a significant difference in proportions of the two methods ($p < 0.001$). In order to evaluate the association of the method predictions and observed preterm value, kappa statistics were calculated. The kappa coefficients for the association of observed values with the LR and DT predicted values were 0.04 ($p = 0.005$) and 0.18 ($p < 0.001$), respectively, which indicate significant statistical reliabilities. Moreover, the AUC showed significantly higher classification accuracy for LR than for DT (Figure 2).

DISCUSSION

PTB is as one of the leading causes of newborn deaths in every population and several factors are associated with its occurrence. The prevalence of PTB in this study was 5.6%, which is consistent with reported rates in Nordic and some developed countries, but lower than reported rates in some developing countries and poor regions. Based on 184 countries, the global average PTB prevalence rate in 2010 was 11.1% [4]. These differences may be due to geographic and ethnic diversity and nutritional differences.

Multiple pregnancy is a strong predictor for PTB, though the mechanisms may differ from those in women with singleton pregnancies. Nearly 60% of all twins are born preterm. The suggested cause for these PTBs is over-distension of the uterus [16]. Consistent with prior literature, women with preeclampsia had higher rates of PTB than non-preeclampsia [6,17]. We observed that women who conceived with ART had an increased risk for PTB, which is consistent with previous study findings [18,19].

In this study, no statistically significant relationships were observed between PTB and factors such as mother's occupation, socio-economic status, BMI, parity, history of abortion and stillbirth. While a significant correlation was observed between these factors and PTB in some similar studies [20–22], those findings may be due to differences in the characteristics of the studied populations.

To classify births as PTB or non-PTB, we compared two different classification methods. Despite the numerous advantages in utilizing a distribution-free method such as DT, our study showed that the LR model provided more accurate results. The DT method uses an intuitive graphical tool which has been utilized in a number of medical and clinical problems [23,24]. The DT method can accommodate noisy data in addition to providing accurate and precise prediction and classification [25]. However, the parametric LR method requires some distributional assumptions and the results can be interpreted in a more scientific and epidemiologic pattern [26]. The LR model is more accurate in classifying patients when there are low proportions of missing data and outliers [27]. In this study, the LR model performed better than the DT approach. Similar results were reported by Long et al. [28] for classifying patients as having acute cardiac ischemia. In 2011, Chen [29] predicted corporate financial distress based on integration of DT classification and LR. Several conclusions were made according to different situations in the dataset. The authors observed out that in some conditions, the LR or DT methods can outperform. Khemphila and Boonjing [30] classified patients with heart disease using artificial neural network, DT, and LR methods. They found that artificial neural networks had the lowest error rate and the highest accuracy while the DT and LR did not show a considerable difference in performance. The same results were reported by Li et al. [31] in predicting peripheral neuropathy in type 2 diabetes mellitus. In contrast to our study, some studies reported a better performance for the DT method than that for LR such as the study by Samanta et al. [32] who preferred the classifications made by the DT over the LR model. The same results were shown by Safiarian et al. [33] in identifying risk groups for bleeding after coronary artery bypass graft surgery and by Sledjeski et al. [34] in a study using risk assessment to predict recurrent maltreatment. Despite its limitations such as low prevalence of PTB, this study compared two different methods, suggesting LR as the best classifier model, a finding that may help policymakers in determining preterm risk factors.

PTB is a main cause of infant deaths that also results in high costs to national health care systems of the country. Therefore, the rate of PTB needs to be diminished. The results of this study revealed maternal and neonatal factors that contribute to PTB, some of which are changeable and preventable; thus, the implementation of activities such as the identification of mothers at risk, necessary training, and improved prenatal care can reduce premature birth rates.

Moreover, based on our findings, the LR method had a better performance in classifying PTB compared to the DT method.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

This study was supported by the Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran. The authors thank the women for their participation in this study and the staff of the hospitals of Tehran University of Medical Science, Shahid Beheshti University of Medical Sciences, Iran University of Medical Science, and Islamic Azad University for their contributions to this study.

REFERENCES

1. WHO: recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. Modifications recommended by FIGO as amended October 14, 1976. Acta Obstet Gynecol Scand 1977;56:247-53.
2. Liu L, Oza S, Hogan D, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet 2015;385:430-40. https://doi.org/10.1016/S0140-6736(14)61698-6
3. Mwaniki MK, Atieno M, Lawn JE, et al. Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: a systematic review. Lancet 2012;379:445-52. https://doi.org/10.1016/S0140-6736(11)61577-8
4. Blencowe H, Cousens S, Oestergaard MZ, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet 2012;379:2162-72. https://doi.org/10.1016/S0140-6736(12)60820-4
5. Zeitlin J, Szamotulska K, Drewniak N, et al. Preterm birth time trends in Europe: a study of 19 countries. BJOG 2013;120:1356-65.

https://doi.org/10.1111/1471-0528.12281

6. Goldenberg RL, Culhane JF, Iams JD, et al. Epidemiology and causes of preterm birth. Lancet 2008;371:75-84. https://doi.org/10.1016/S0140-6736(08)60074-4
7. Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med 2010;362:529-35. https://doi.org/10.1056/NEJMra0904308
8. Klebanoff MA, Keim SA. Epidemiology: the changing face of preterm birth. Clin Perinatol 2011;38:339-50. https://doi.org/10.1016/j.clp.2011.06.006
9. Kate RJ, Perez RM, Mazumdar D, et al. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 2016;16:39. https://doi.org/10.1186/s12911-016-0277-4
10. Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica 2007;31:249-268.
11. Rencher AC. Methods of multivariate analysis. 2nd ed. New York: John Wiley & Sons; 2003.
12. Dietterich T. Ensemble methods in machine learning. In: Kittler J, Roli F, editors. Proceeding of 1st International Workshop on Multiple Classifier Systems. London: Springer-Verlag; 2000.
13. Rokach L, Maimon O. Data mining with decision trees: theory and applications. Series in Machine Perception and Artificial Intelligence: Volume 69. New Jersey: World Scientific; 2014.
14. Jenhani I, Amor NB, Elouedi Z. Decision trees as possibilistic classifiers. Int J Approx Reason 2008;48:784-807. https://doi.org/10.1016/j.ijar.2007.12.002
15. Detsky AS, Naglie G, Krahn MD, et al. Primer on medical decision analysis: part 2--building a tree. Med Decis Making 1997;17:126-35. https://doi.org/10.1177/0272989X9701700202
16. Mercer BM, Goldenberg RL, Meis PJ, et al. The preterm prediction study: prediction of preterm premature rupture of membranes through clinical findings and ancillary testing. The national institute of child health and human development maternal-fetal medicine units network. Am J Obstet Gynecol 2000;183:738-45. https://doi.org/10.1067/mob.2000.106766
17. Derakhshi B, Esmailnasab N, Ghaderi E, et al. Risk factor of preterm labor in the west of iran: a case-control study. Iran J Public Health 2014;43:499-506.
18. Wang YA, Sullivan EA, Black D, et al. Preterm birth and low birth weight after assisted reproductive technology-related pregnancy in Australia between 1996 and 2000. Fertil Steril 2005;83:1650-8. https://doi.org/10.1016/j.fertnstert.2004.12.033
19. Dunietz GL, Holzman C, McKane P, et al. Assisted reproductive technology and the risk of preterm birth among primiparas. Fertil Steril 2015;103:974-9.e1. https://doi.org/10.1016/j.fertnstert.2015.01.015
20. Morgen CS, Bjørk C, Andersen PK, et al. Socioeconomic position and the risk of preterm birth--a study within the Danish National Birth Cohort. Int J Epidemiol 2008;37:1109-20. https://doi.org/10.1093/ije/dyn112
21. Brett KM, Strogatz DS, Savitz DA. Employment, job strain, and preterm delivery among women in North Carolina. Am J Public Health 1997;87:199-204.
22. Torloni MR, Betrán AP, Daher S, et al. Maternal BMI and preterm birth: a systematic review of the literature with meta-analysis. J Matern Fetal Neonatal Med 2009;22:957-70. https://doi.org/10.3109/14767050903042561
23. Nair SR, Tan LK, Mohd Ramli N, et al. A decision tree for differentiating multiple system atrophy from Parkinson's disease using 3-T MR imaging. Eur Radiol 2013;23:1459-66. https://doi.org/10.1007/s00330-012-2759-9
24. Valero S, Daigre C, Rodríguez-Cintas L, et al. Neuroticism and impulsivity: their hierarchical organization in the personality characterization of drug-dependent patients from a decision tree learning perspective. Compr Psychiatry 2014;55:1227-33. https://doi.org/10.1016/j.comppsych.2014.03.021
25. Li XB. A scalable decision tree system and its application in pattern recognition and intrusion detection. Decis Support Syst 2005;41:112-30. https://doi.org/10.1016/j.dss.2004.06.016
26. Zhu M, Philpotts D, Sparks R, et al. A hybrid approach to combining CART and logistic regression for stock ranking. J Portf Manag 2011;38:100-9. https://doi.org/10.3905/jpm.2011.38.1.100
27. Das U, Maiti T, Pradhan V. Bias correction in logistic regression with missing categorical covariates. J Stat Plan Inference 2010;140:2478-85. https://doi.org/10.1016/j.jspi.2010.02.018
28. Long WJ, Griffith JL, Selker HP, et al. A comparison of logistic regression to decision-tree induction in a medical domain. Comput Biomed Res 1993;26:74-97.
29. Chen MY. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. Expert Syst Appl 2011;38:11261-72. https://doi.org/10.1016/j.eswa.2011.02.173
30. Khemphila A, Boonjing V. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. Paper presented at: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM); 2010 Oct 8–10; Krakow, Poland. IEEE, 2010. pp. 193-8. https://doi.org/10.1109/CISIM.2010.5643666
31. Li CP, Zhi XY, Ma J, et al. Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chin Med J (Engl) 2012;125:851-7.
32. Samanta B, Bird GL, Kuijpers M, et al. Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. Artif Intell Med 2009;46:201-15. https://doi.org/10.1016/j.artmed.2008.12.005
33. Safiarian R, Amini P, Moez EK, et al. Risk group classification for bleeding after coronary artery bypass graft surgery: a comparison of the logistic regression with decision tree models. Türk Göğüs Kalp Damar Cerrahisi Dergisi 2013;21:574-80. https://doi.org/10.5606/tgkdc.dergisi.2013.7680
34. Sledjeski EM, Dierker LC, Brigham R, et al. The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (CART). Prev Sci 2008;9:28-37. https://doi.org/10.1007/s11121-007-0079-0