Environmental factors prediction in preterm birth using comparison between logistic regression and decision tree methods: an exploratory analysis

Rakesh Kumar Saroj[a*], Madhu Anand[b] , Neha Kumari[c]

[a] Department of Mathematics and Statistics, SRM University Sikkim/Gangtok -737102, India

[b] Department of Chemistry, Dr. B.R. Ambedkar University, Khandari Campus, Agra-282002, India

[c] School of Information Technology, SRM University Sikkim/Gangtok -737102, India

Abstract:

Objective The main objective of this paper is to compare the performance of logistic regression and decision tree classification methods and to find the significant environment determinants that causes pre-term birth.

Design, setting and population Between 2017 to 2018, 90 pregnant females underwent birth outcome followed by research staff at our institutions, out of those 50 are full-term and 40 are preterm births in this study.

Method Before and after feature selection logistic regression and decision tree classifier model has been compared in this dataset and to evaluate the model accuracy.

Main outcome measures Preforming the accuracy of machine learning classification model and important factors on pre-term birth.

Results: Using chi-square test and find the Area of residence and GSH, MDA, α-HCH, total HCH and total DDT are responsible for the preterm birth. Using the multiple logistic regression, pre term birth was associated with MDA and α-HCH (95% CI 0.04 to 0.48 and 95% CI 0.82 to 0.97). The logistic and decision tree model comparison result shows that logistic regression is better in terms of metrics (precision = 0.92, F1-score = 0.96 and AUROC = 0.97), while decision tree performs the poor (precision = 0.75, F1-score = 0.86 and AUROC = 0.87).

Conclusions The logistic regression is accurate model to predict the pre-term as compare to decision tree method. The variables like α-HCH , total HCH and MDA (Malondialdehyde) are the most influential factors for preterm birth.

Keywords: Machine learning, Pre-term birth, Classifier, Decision tree, Logistic regression

Corresponding Author
Rakesh Kumar Saroj, Assistant Professor
Department of Mathematics and Statistics,
SRM University Sikkim/Gangtok -737102, India
Email: rakesh.saroj@bhu.ac.in
Mob: +91-9454196475

Introduction

Preterm birth is the birth of a baby at before 37 weeks' gestational age, as opposed to the usual about 40 weeks. Preterm birth (PTB) has found through various reason including low socio-economic status, smoking, race and consumption of alcohol (Metzger et al. ,2013). Previous researches have suggested the associations between organochorines and increased risk of abortion, small for gestational age babies, minor malformations, cryptochildism and hypospadias in the infants have been reported (Birnbaum et al., 1994 and Hosie et al.,2000). The studies have suggested the associations between organochorine pesticides and increased risk of preterm births, low birth weight, abortion, small for gestational age, minor malformations, cryptochildism and hypospadias in the infants have been reported (M et al.,2017, Anand, and Taneja,2019). The study suggested that the important factors for PTB are early pregnancy, multiple pregnancies, poor nutrition, use of assisted reproductive technology (ART), excessive physical work and psychological health of maternal (Goldenberg et al.,2008, Muglia and Katz, M, 2010, Klebanoff and Keim, 2011), environmental factors such as lifestyle and chemicals (Kumar,2008, Ashton, and Lee,2009, Landrigan, 2010). Among maternal variables that have been related with increased risk of preterm birth are: young or progressed maternal age, less interval between pregnancy and low body mass index of maternal (Goldenberg et al.,2008). The various machine learning techniques have been used in different fields for prediction and classification model. There are various machine learning prediction and classification models like regression, logistic regression, principal component analysis, decision tree and maximum likelihood method have been used to improve maternal and child health.

The machine learning scientists have worked on interpretable prediction techniques in several places (Jacob and Robert, 2011, Emilio et al.,2015, Emad et al.,2015). The previous research article recommended that if early risk factors are identified through classification methods in acute kidney injury then can be protect the life of patients (Kate et al.,2016). The study has used the data mining method for finds the demographic factors in preterm birth (Goodwin et al., 2001). The binary logistic regression has applied on data after transforming the correlated variables using principal component analysis (PCA) for finds the factors of children death (Khan et al.,2018).

The prediction models can be used to find the important factors in PTB and PTB can be classifying new births as PTB on various risk factors through classification methods. The above cited researches clear indicate that multiple prediction and classification methods were used to find the factors in preterm birth and child mortality but there is lack of proper comparison of classification and prediction methods in the preterm birth at Indian setup in the presence of multiple variables. As well as, there is a need of accurate prediction and classification models to provide the highly accurate results and allows the health researchers to experiment with a various set of aspects.

Logistic regression (LR) is the most accepted parametric model to classify discrete outcome variables using several cofactors. However, the non-parametric decision tree (DT) is favorable when the subjects are described through a predetermined set of characteristics, the outcome variable is binary or multinomial, and disjunctive results are needed (Dietterich,2000). It is very important to predict preterm birth due to its possible for poor long-term consequences. Overall aim of this study to construct the logistic and classification model to predict the high-risk groups of PTB based on several factors and covariates in order to reduce the risk of PTB and compare the predicting performance of LR and DT classification method.

Methods

Data Source

Between 2017 to 2018, 90 pregnant women who had given birth near Agra region. Those women were approached for enrolment for this study at Dr. Bhim Rao Ambedkar University (located at Agra, Uttar Pradesh, India). Out of 90, Forty women delivered <37 weeks and fifty women delivered at 37+ weeks. The details of predictors and outcome variable are given.

Outcome variable

The outcome variable or dependent variable in this study is preterm birth and it is classified between full term birth and pre-term birth.

Explanatory variables

The risk factors considered in this study include variables age, BMI, number of children, lactation duration, addiction, residence, pesticide exposure, drinking water sources; dietary habits baby gender and organ- chlorine pesticides in the placenta of the females.

Logistic Regression

In this regression dependent variables always will be in binary (0/1) or (Yes/No) and independent variable will be any measurement. The equation is given below

$(Y/1-Y)=a+bX_1+cX_2+\ldots\ldots X_n+\epsilon$ (1.1)

$Log(Y/1-Y)=a+bX_1+cX_2+\ldots X_n+\epsilon$ (1.2)

In this model X_1, X_2 and X_n are the predictors/factors/cofactors/independent variables, a is intercept and b,c, are the regression coefficient and ϵ is the residual (error). It represents change in the value of dependent variable corresponding to unit change in the value of independent variable. In this dataset pre-term birth (0) and full-term birth (1) is coded.

Decision Tree

Decision tree is one of the best and mostly used supervised machine learning algorithms that partition the data into subsets. This partitioning process starts with a binary split and continues until no further splits can be made. This machine learning technique covers both classification and regression method. The decision result or model of decisions display tree-like graph that's reason it's known as decision tree. In the decision tree decision nodes, branches, and leaves are mainly three parts.

Statistical Analysis

In this research paper, we used various approaches to achieve the objectives. Firstly, we used the Chi-Square test statistic to measure the significance of association between the outcome variable preterm birth (pre-term or full-term birth) and explanatory variables. Secondly, we used the multivariate analysis using multiple logistic regressions to estimate the odds ratios of outcome variable and significant factors of chi-square. Thirdly, we applied the decision tree model in whole dataset and train datasets because decision tree creates a binary tree and it is very useful in classification problems. The next step is measure to rank of the variables from the data sets through information gain technique. Finally, the dataset is divided into two subsamples. We have used the decision tree and logistic regression both classifier model in the training dataset (70% of cases) and evaluate the model using the test sample (30% of cases). This helps in ensuring that which model is the best in terms of predicting child birth.

Results.

In this study the Table 1 represent the results of the Chi-Square test of association between the preterm birth and the demographic variables. This table result shows that preterm birth is significantly associated with area of residence in demographic variable. The Chi-square test of association between placental organ-chlorine pesticide levels and preterm birth are shown in Table 2 and the preterm birth is significantly associated with GSH (Glutathione), MDA (Malondialdehyde), α-HCH, γ- HCH, total HCH, p,p-DDE (p,p- dichlorodiphenyldichloroethylene), and total DDT (dichlorodiphenyltrichloroethane).The multiple logistic regression uses to find the estimate the odds ratios of outcome variable and factors in the Table 3. This result shows that preterm birth is significantly associated with only γ-HCH and MDA.

The decision tree applies in the train dataset. The decision tree rules and the proportion of the preterm birth in several resultant nodes are shown in Figure 1 for trained dataset. The results of the decision tree showed that α-HCH, total HCH and MDA is the most important variable for the classification of preterm birth in trained dataset. The decision tree starts with root node at the top in figure, it is shows that 55 percent of female give the full delivery. The second node shows that 35% of female's α-HCH value is less than 5.2 and their chance of full-term birth is 100%. The α-HCH value is higher than 5.2 is found in 65% female and their chances of full-term birth is 30%. If the α-HCH value is greater than 5.2 but total HCH value is higher than 76 then, the chances of full-term birth are 0% and which female's total HCH value is less than 76 then those female chance of full-term birth is 55%. If the female's MDA value is greater than 2.2 then chances of full-term birth is 31% and MDA value is less than 2.2 then chances of full-term birth is 89%. The variable rank shows in the Figure 2 using information gain measure. Higher values of information gain indicate those variables are important for preterm birth. The α-HCH, total HCH, MDA, p,p-DDT, β-HCH and p,p-DDE are

the highly ranked variables. In other words, it declares that these variables play important role in preterm birth. The higher rank variables like α-HCH, total HCH and MDA are the most influential factors with respect to association with preterm birth.

We developed the logistic and decision tree classification model for prediction with the help of 70% of training dataset including all variables. The logistics and decision tree models are tested on 30% testing dataset and model evaluation results are given in Table 4. The result shows that logistics classifiers with the better accuracy of predictions compared to decision tree. The accuracy of logistic regression for classifying preterm birth is 0.96, significantly different from the decision tree method. The comparison of the performance of the both classifier models include all variables reveals that logistic regression performs the better in terms of metrics (precision = 0.92, F1-score = 0.96 and AUROC = 0.97), while decision tree performs the poor (precision = 0.75, F1-score = 0.86 and AUROC = 0.87). The Figure 3 shows the receiver operating characteristic (ROC) curve for all variables and it is found that logistic regression model is better than decision tree model. After that we have used the top six ranked variables and apply the both models. The reason behind repeat the same experiment with top six ranked variables are to emphasize how efficient is the use of information gain measures in the data. The results obtained using the top six variables are given in Table 5. The result found that top six variables do not affect the accuracy performance of decision tree method and again the accuracy of logistic regression for classifying preterm birth is 0.78, significantly different from the decision tree method. The performance gain is shown by logistic regression (precision = 0.65, F1-score = 0.79 and AUROC = 0.81) and decision tree performs (precision = 0.56, F1score = 0.69 and AUROC = 0.71). In figure 4, we present ROC curves for both classifiers models with the help of top six ranked variables. It shows that that logistic regression model is again better than decision tree model.

Discussion

While the final objective of specialist classification/prediction machine learning technique development is to predict preterm birth risk, the definition of preterm birth and data needed to analyze preterm birth risk are less amenable to study presently. Already various researches explain that preterm birth is as one of the major causes of neonatal death in across the world and various factors are found associated with its occurrence. Therefore, the purpose of this study is to determine the feasibility of using classification/prediction machine learning to generate expert system (knowledge-base) rules for prediction of preterm birth. The study recommended adding characteristics to improve machine learning classification and suggested that, when selecting characteristics, it is better to error on the side of having too many (Thompson and Thompson,1986). The prior research proposed that women with preeclampsia had higher rates of preterm birth than non-preeclampsia (Derakhshi et al., 2014). The study suggested that women who received with assisted reproductive technology had an increased risk for preterm birth (Dunietz et al.,2015). In this study we use the logistic regression

and decision tree method on preterm birth data and it is found that LR is the most accurate in terms of predicting preterm birth.

With this method, scientists, researchers and practitioners are able to predict and detect the preterm birth of data sets. 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.

Conclusion

This study is limited due to small amount of data because the machine learning techniques give better result in big datasets. This study demonstrated the logistic regression and decision tree method for understanding the preterm birth prediction rule. The result finds that preterm birth is significantly associated with only γ-HCH and MDA in logistic regression and decision tree both methods. Finally, the results revealed that the logistic regression is better accuracy in classifying preterm birth compared to the decision tree method.

Disclosure of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Contribution to authorship

RS Supervision, Conceptualization, Methodology, Writing original draft. MA Data Collection, Laboratory data details, Writing - review & editing. NK Coding Software, Validation.

Details of ethics approval

Ethical approval for this study was granted by the Ethics Board of the Dr. Bhimrao Ambedkar University and S.N. Medical College, Agra (date of approval: 4th December ,2014, reference 2014/08).

Funding

One of the authors (Madhu Anand) gratefully acknowledges Post-Doctoral Fellowship [F.15-1/2016-17PDFWM-2015-17-UTT-36837(SA-II) by the University of Grants Commission (New Delhi), Government of India.

Acknowledgements

The authors thank all current and former Study group members, doctors, and research staff.

They also thank those who contributed to study organization, recruitment, data collection and management. Special thanks are due to the study participants and their families. Finally, they also thank Professor Ajay Taneja for final comment of the manuscript before submission.

References

1. Metzger, M. J., Halperin, A. C., Manhart, L. E., & Hawes, S. E. (2013). Association of maternal smoking during pregnancy with infant hospitalization and mortality due to infectious diseases. The Pediatric infectious disease journal, 32(1), e1–e7. https://doi.org/10.1097/INF.0b013e3182704bb5.
2. Birnbaum, S. C., Kien, N., Martucci, R. W., Gelzleichter, T. R., Witschi, H., Hendrickx, A. G., & Last, J. A. (1994). Nicotine-or epinephrine-induced uteroplacental vasoconstriction and fetal growth in the rat. Toxicology, 94(1-3), 69-80. https://doi.org/10.1016/0300-483X(94)90029-9.
3. Hosie, S., Loff, S., Witt, K., Niessen, K., & Waag, K. L. (2000). Is there a correlation between organochlorine compounds and undescended testes?. European Journal of Pediatric Surgery, 10(05), 304-309.
4. M. Anand, L. Singh, P. Agarwal, R. Saroj & A. Taneja (2019) Pesticides exposure through environment and risk of pre-term birth: a study from Agra city, Drug and Chemical Toxicology, 42:5, 471-477. https://doi.org/10.1080/01480545.2017.1413107.
5. Anand, M., & Taneja, A. (2020). Organochlorine pesticides residue in placenta and their influence on anthropometric measures of infants. Environmental Research, 182, 109106. https://doi.org/10.1016/j.envres.2019.109106.
6. Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. The lancet, 371(9606),75-84. https://doi.org/10.1016/S0140-6736(08)60074-4.
7. Muglia, L. J., & Katz, M. (2010). The enigma of spontaneous preterm birth. New England Journal of Medicine, 362(6), 529-535.
8. Klebanoff, M. A., & Keim, S. A. (2011). Epidemiology: the changing face of preterm birth. Clinics in perinatology, 38(3),339-350. https://doi.org/10.1016/j.clp.2011.06.006.
9. Kumar, S. (2008). Is environmental exposure associated with reproductive health impairments?. Journal of the Turkish-German Gynecological Association, 9(1).
10. Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. Journal of personality assessment, 91(4), 340-345.
11. Landrigan, P. J. (2010). What causes autism? Exploring the environmental contribution. Current opinion in pediatrics, 22(2), 219-225.
12. Goldenberg RL, Culhane JF, Iams JD, Romero R. (2008). Epidemiology and causes of preterm birth. Lancet (London, England)), 371 (9606), 75–84.
13. Jacob Bien, Robert Tibshirani (2011). Prototype selection for interpretable classification. The Annals of Applied Statistics, 5(4),2403–2424.
14. Emilio Carrizosa, Amaya Nogales- G ´omez, and Dolores Romero Morales. (2016).Strongly agree or strongly disagree?: Rating Features in Support Vector Machines,Information Sciences,329:256–273. https://doi.org/10.1016/j.ins.2015.09.031.
15. Emad, A., Varshney, K. R., & Malioutov, D. M. (2015, July). A semiquantitative group testing approach for learning interpretable clinical prediction rules. In Proc. Signal Process. Adapt. Sparse Struct. Repr. Workshop, Cambridge, UK.
16. Kate, R. J., Perez, R. M., Mazumdar, D., Pasupathy, K. S., & Nilakantan, V. (2016). Prediction and detection models for acute kidney injury in hospitalized older adults. BMC medical informatics and decision making, 16(1), 39.
17. Goodwin, L. K., Iannacchione, M. A., Hammond, W. E., Crockett, P., Maher, S., & Schlitz, K. (2001). Data mining methods find demographic predictors of preterm birth. Nursing research, 50(6), 340-345.
18. Khan, R. E. A., Bari, K. M., & Raza, M. A. (2018). Socioeconomic determinants of child mortality: Evidence from Pakistan Demographic and Health Survey.
19. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.
20. Thompson, B., & Thompson, W. (1986). Finding rules in data. Byte, 11(12), 149-158.
21. Derakhshi, B., Esmailnasab, N., Ghaderi, E., & Hemmatpour, S. (2014). Risk factor of preterm labor in the west of iran: a

case-control study. Iranian journal of public health, 43(4), 499.
22. Dunietz, G. L., Holzman, C., McKane, P., Li, C., Boulet, S. L., Todem, D., ... & Diamond, M. P. (2015). Assisted reproductive technology and the risk of preterm birth among primiparas. Fertility and sterility, 103(4), 974-979. https://doi.org/10.1016/j.fertnstert.2015.01.015.

Table 1. Demographic characteristics of the females of preterm and full-term deliveries.

Variables	Pre-term (n=40)	Full-term (n=50)	p-value
Age (Yrs.)	24.25±0.50	25.54±0.55	0.09
Body mass index (kg/m2)	24.57 ± 0.46	24.50 ± 0.39	0.92
Number of children			
One Children	18 (35%)	34 (65%)	0.24
> One Children	22(58%)	16 (42%)	
Duration of lactation	15.1± 1.35	11.5±1.23	0.05
Addiction			
No	20 (50%)	26 (52%)	0.85
Yes	20 (50%)	24 (48%)	
Area of residence*			
Rural	26 (65%)	19 (38%)	0.01
Urban	14 (35%)	31 (62%)	
Exposure to pesticide			
Yes	0 (00.0%)	05 (10%)	0.63
No	40 (100%)	45 (90%)	
Source of drinking water			
Private	15 (38%)	22 (44%)	0.53
Government	25 (62%)	28 (56%)	
Dietary habits			
Vegetarian	29 (73%)	27 (54%)	0.07
Non-vegetarian	11 (27%)	23 (46%)	
Baby Gender			
Female	17 (42%)	18 (36%)	0.34
Male	23 (58%)	32 (64%)	

Table 2. Organochlorine pesticides and biochemical indices (GSH & MDA) in the placenta of the females from pre-term and full- term deliveries.

Name of pesticides	Pre-term (n=40) Mean±S.E.	Full-term (n=50) Mean±S.E.	p-value
GSH*	3.56±0.28	5.09±0.47	0.01
MDA*	3.27±0.17	1.88±0.13	0.00
α-HCH*	39.75±6.28	7.48±1.15	0.00
β-HCH	37.78±17.02	15.67± 2.71	0.16
γ-HCH*	28.65±4.50	11.37±2.64	0.00
δ-HCH	4.67±1.01	4.40±0.67	0.83
Total HCH*	110.87±19.06	38.94.0±4.08	0.00
p,p-DDE*	13.99±1.91	8.99±1.51	0.04
p,p-DDT	4.26±1.83	1.18±0.69	0.09
o,p-DDD	1.44±0.62	0.97±0.28	0.45
Total DDT*	19.70±2.77	11.15±1.83	0.01

Results are in ppb , *p<.05

Table 3. Relationship between pesticide exposure during pregnancy and preterm birth after adjusting for maternal characteristics.

Variables	Coefficients	Odds Ratio	p-value	95% Confidence Interval	
				Lower	Upper
Age	0.19	1.20	0.12	0.95	1.53
GSH	0.12	1.13	0.57	0.74	1.73
MDA*	-1.93	0.14	0.00	0.04	0.48
α-HCH*	-0.11	0.89	0.01	0.82	0.97
γ-HCH	0.01	1.01	0.57	0.97	1.06
Total-HCH	-0.03	0.97	0.18	0.93	1.01
p,p-DDE	0.00	1.00	0.97	0.85	1.17
Total-DDT	-0.05	0.95	0.38	0.85	1.06

Table 4. Evaluation of classification models using all factors.

Performance metrics	Model	
	LR	DT
Accuracy	0.96	0.85
Sensitivity (Recall)	1.00	1.00
Precision (Positive predictive value)	0.92	0.75
F1-score	0.96	0.86
AUROC	0.97	0.87

Table 5. Evaluation of classification models using the important factors.

Performance metrics	Model	
	LR	DT
Accuracy	0.78	0.67
Sensitivity (Recall)	1.0	0.91
Precision (Positive predictive value)	0.65	0.56
F1-score	0.79	0.69
AUROC	0.81	0.71