Ibn Al-Haitham Journal for Pure and Applied Sciences
Journal homepage: jih.uobaghdad.edu.iq
PISSN: 1609-4042, EISSN: 2521-3407
IHJPAS. 2024, 37(4)

A Modified Multivariate Bayesian Logistic Model with Application to Health Datasets

Azza Mustafa Abd Al Kader Al Kusaem*
Department of Business Administration, College of Administration and Economics, University of Mosul, Mosul, Iraq.
*Corresponding Author.

Received: 26 January 2023 Accepted: 16 March 2023 Published: 20 October 2024
doi.org/10.30526/37.4.3245

Abstract

The application of Bayesian strategies for binary logistic estimation is demonstrated in this article. A modified method of the Bayesian logistic model using the Metropolis-Hasting algorithm is derived and applied to three simulation data sets. We compared the new model with existing classification methods: support vector machine, artificial neural network and regular logistic model. The modified model was used to classify the heart disease dataset. The data came from a database intended for UCI Data Science (https://www.kaggle.com). The clarification accuracy and the time required are checked and compared with other standard methods. It has been shown that the presented model has the best accuracy and efficiency compared to the different classification methods. All calculations were performed with the program R version 4.2.2.

Keywords: Bayesian, classification, heart disease, logistic regression, posterior distribution.

1. Introduction

The Bayesian algorithm, Markov chain Monte Carlo (MCMC), is a widely used technical algorithm used to estimate the parameters of the posterior distribution in the model [1, 2]. The standard estimation of the logistic regression model has some limitations that can be overcome with possible replacement methods. This paper aims to introduce a different approach using Bayesian analysis.

Multinomial logistic regression can predict categorical placement based on multiple independent variables, such as the possibility of belonging to a category on a related variable. The independent variables can be either categorical or continuous. Multinomial logistic regression is often considered an appealing analysis because it does not require homoscedasticity, linearity, or normality [3]. Discriminant function analysis is an excellent substitute for multinomial logistic regression, which requires these assumptions. One of the most common classification methods for health datasets is logistic regression. The maximum likelihood or ordinary least square estimator can be used to estimate the regression coefficients [4, 5].

The dependent variable in logistic regression is either binary or dichotomous, i.e. the logistic regression contains only data coded as "1" (true, yes, healthy, success, pregnant, etc.) or "0" (false, no, sick, failure, not pregnant, etc.) [6]. Many standard classifications such as Support Vector Machine (SVM) [2, 7], Artificial Neural Networks (ANN) [8] and regular logistic regression [9] can be used for classification. However, working with a large dataset leads to inefficiency and time consumption [10, 11]. This paper presents a new Bayesian multivariate model for classifying datasets. The modified model is applied to three simulation datasets and then used to classify the heart disease dataset.

This article is structured as follows: In section two, a general idea for multivariate Bayesian binary logistic linear regression is explained. The Bayesian formulation model is discussed in section three. Simulation studies and experimental results are presented in section four. The real data set is presented in section five. Section six discusses the results.

2. Multivariate Bayesian binary logistic regression model

Binary logistic regression is a special form of regression in which the binary response variable is linked to a discrete or continuous set of explanatory variables. The key point here is that the predicted values of the response variable are modelled based on the mixture of values provided by the predictors in linear regression [12]. Let $X = (X_1, X_2, \dots, X_p)$ be a set of p explanatory variables and Y be a binary response variable; then the binary Logistic regression model can be written as follows,

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta x_i + \cdots + \beta_0 + \beta X_k, \quad (1)$$

Which models the log odds of the probability of "present" as a function of explanatory variables. Assume that $Y_i \sim Bin(n_i, p_i)$, i.e., Y_i is a binary logistic model, then

$$\pi_i = Pro(Y_i = 1 | X_i = x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+ e^{\beta_0+\beta_1 x_i}} \quad (2)$$

The maximum likelihood estimator (MLE) for the parameters, $\{\beta_0, \beta_1\}$, is obtained by finding $(\widehat{\beta_0}, \widehat{\beta_1})$ that maximizes:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i-y_i} = \prod_{i=1}^{n} \frac{e^{\{y_i(\beta_0+\beta_1 x_i)\}}}{1+e^{\beta_0+\beta_1 x_i}} \quad (3)$$

3. Bayesian Formulation

The Bayesian model can be treated like a classification problem. Researchers can infer the individual from the model parameters and the data. From a set of likely divergent opinions about a particular condition, one of two outcomes can be inferred: Does individual j actually have a particular disease state (y = 0 if no and y = 1 if yes) [13, 14]. The Bayesian hierarchical logistic regression model addresses this problem to account for the variability of outcomes stemming from both informants and informative priorities. Bayes' theorem and a generative model can be applied when we have data to calculate the posterior probability distribution of the model parameters using the prior distribution as a function of the predictors (X) and the response (Y). In general, the three critical components associated with parameter estimation in the Bayesian system are the prior distribution, the likelihood function, and the posterior distribution. Bayes' theorem formally combines these three elements:

$$Posterior = Likelihood \times Prior$$

Basically, the above expression means that the knowledge in the sample (reflected in the likelihood function) is combined with data from other sources (summarised in the prior distribution) to obtain the posterior distribution. All available information about the parameters of the model is incorporated into the posterior distribution. [15],[16] deals in detail with the principle of Bayesian analysis.

The likelihood contribution from the i^{th} individual is binomial,

$$L_i = \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \times \left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{(1-y_i)} \quad (4)$$

Since individual subjects are presumed to be separate from one another, the probability function for a data set of n subjects is then the probability distribution.

$$L = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \times \left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{(1-y_i)} \quad (5)$$

By assuming the multivariate normal prior on β; i.e. $\beta_i \sim N(\mu_i, \sigma_i^2)$, we get

$$f(\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2}\left(\frac{\beta_i-\mu_i}{\sigma_i}\right)^2\right\}} \quad (6)$$

As a result, the posterior distribution is calculated by multiplying the probability function by the prior:

$$\text{Posterior} = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \times \left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{(1-y_i)} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2}\left(\frac{\beta_i-\mu_i}{\sigma_i}\right)^2\right\}} \quad (7)$$

We can use the Metropolis-Hasting procedure to sample from the above posterior distribution. The following section will present the algorithm in detail.

3.1 Markov chain Monte Carlo

If a sequence of numbers follows the below graphical model, it is a Markov chain is $P(X5|X4,X3,X2,X1) = P(X5|X4)$. As a result, the probability of reaching a specific state is solely determined by the chain's previous state [10].

Using the full joint density function, the Metropolis-Hasting (MH) algorithm can be used. The MH technique is an iterative method that generates a Markov chain sequence to estimate the posterior distribution's parameters [17,18]. The following steps summarize the Metropolis-Hasting algorithm [19, 20]:

Metropolis-Hasting algorithm

1. Initialize $x^{(0)}$
2. For i = 0 to K−1
   - calculate $u \sim$ uniform distribution[0,1]
   - calculate $X^\star \sim q(X^\star|X^{(i)})$
   - If $u < R\left(X^{(i)}, X^\star\right) = \min\{1, \frac{F(X^\star)q(z^{(i)}|Z^\star)}{F(z^{(i)})q(X^\star|Z^{(i)})}\}$
   - $X^{(i+1)} = X^\star$
   - $else$
   - $X^{(i+1)} = X^{(i)}$

We can carefully choose the proposal distribution q. The MH algorithm assumes a symmetric random walk for the proposal distribution, i.e., $q(x|y) = q(y|x)$. $p(x)$ does not even have to be

the full Bayesian probability but simply is required to be proportionate to it. This is clear since the Bayes denominators will cancel out.

4. Simulation Studies

The posterior distributions for the parameters are applied in this section (Eq. 1 and Eq. 2). Datasets of size 100, 200 and 400 are simulated. First, a set of five variables are simulated uniformly from a uniform distribution U(0 1), and then $\beta_0 = 2.41$, $\beta_1 = -1.11$ $\beta_2 = 4.72$, $\beta_3 = 0.31$, and $\beta_4 = 4.10$ are used. By applying the MH algorithm for 10000 repetitions, the parameters were obtained by calculating the posterior sample mean [21, 22]. In **Figure 1**, the posterior samples are plotted as histograms, and a red line remarks the true values for these parameters. Obviously, the histograms are approximately normal, and the true values are close to the sample's mean. In addition, the model is used to forecast the response for specific different values. The responses are plotted as histograms, **Figure 1**, for all the prediction samples, and the exact value are notified with the blue lines.

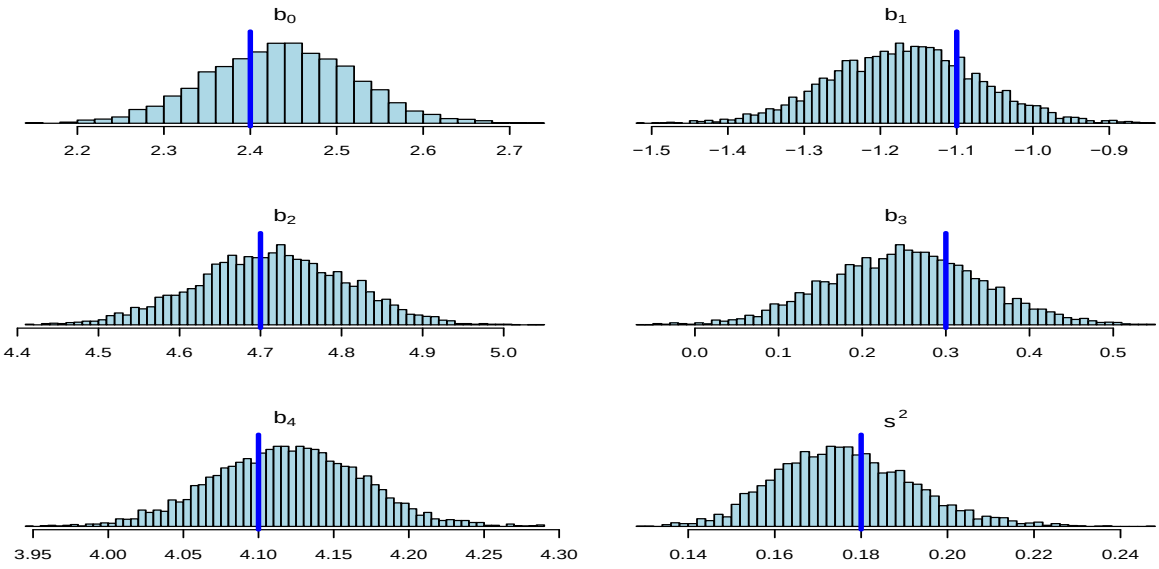

Figure 1. Posterior mean for the parameters.

Table 1. compares the exact parameter values with the classical logistic estimation, and posterior sample mean. Also, 95% credible intervals ($C.I.$)are created for all the parameters. The exact values lie inside the 95% $C.I.$ which implies that the approximation is acceptable [23].

Table 1. Estimation of parameters using classical logistic and posterior distribution.

	True value	Classical Logistic	Posterior mean	95% $C.I.$
$\boldsymbol{\beta_0}$	2.41	2.12	2.45	(2.23 , 2.53)
$\boldsymbol{\beta_1}$	-1.11	-1.43	-1.18	(-1.37 , -1.04)
$\boldsymbol{\beta_2}$	4.72	4.83	4.72	(4.62 , 4.82)
$\boldsymbol{\beta_3}$	0.31	0.21	0.25	(0.12 , 0.42)
$\boldsymbol{\beta_4}$	4.10	4.17	4.12	(4.02 , 4.17)
$\boldsymbol{\sigma^2}$	0.18	0.16	0.17	(0.15 , 0.19)

5. Real Dataset Application

The proposed method has been applied to the real health dataset. We used the heart disease dataset as an application to our work. The dataset has been downloaded from the UCI data science website (https://www.kaggle.com). It contains patients' health history and many medical indicators for each patient. In the following section, a brief description of the heart disease dataset is given.

Heart Disease Dataset Analysis

The description of the variables is given as shown in **Table 2.** The dataset consists of 303 observations and 12 variables: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, slop, ca, and thal. The aim is to forecast whether an individual has heart disease depending on the features or not [24].

Table 2. Descriptive table for the variables

No.	Predicters	Descriptive
1	age	The age in years for patients
2	sex	Male = 1; Female = 0
3	cp	the chest pressure that was felt: Value 1 denotes normal angina, Value 2 denotes atypical angina, Value 3 denotes non-anginal discomfort, and Value 4 denotes asymptomatic
4	trestbps	On admission to the facility, the patient's resting blood pressure was measured in millimeters of mercury (mm Hg).
5	chol	cholesterol levels in milligrams per deciliter
6	fbs	fasting blood sugar is over 120 mg/dl: true = 1; false = 0
7	restecg	Resting electrocardiographic measurement: regular = 0, st-t wave abnormality = 1, potential or definite left ventricular hypertrophy according to Estes' criteria = 3.
8	thalach	Attained optimum heart rate
9	exang	Angina caused by workout: yes = 1; no = 0
10	slop	the slope of the most difficult workout segment: Value 1 indicates an upslope, value 2 indicates a smooth surface, and value 3 indicates a downslope.
11	ca	total of major vessels colored by flourosopy (0 - 3)
12	thal	thalassemia is a form of blood disorder: Standard=3, fixed=6, and reversible=7.

The correlation matrix for predictors in the heart disease dataset are shown in **Figure 2.** All the predictors do not have significant correlations. However, the *slope* has a positive correlation with *thalach*, *age* has negative correlation with *thalach*, and *thalach* has positive and negative correlation with both *slop* and *exang*; respectively. The data set is visualized in **Figure 3.** where it is plotted as a scatter plot. The x-axis and y-axis are chosen to be chol and trestbps, respectively. The chest pain experienced (cp) is used to note the observations for all the individuals.

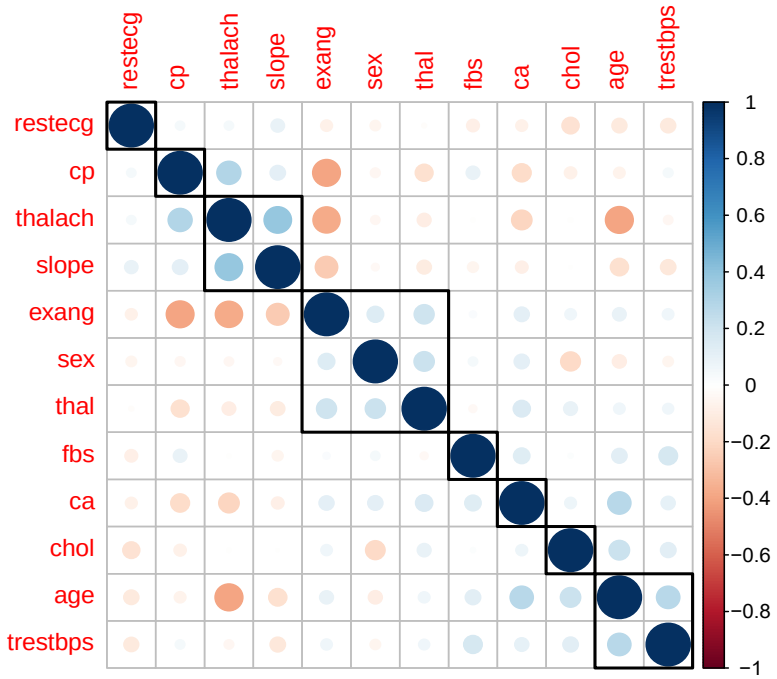

Figure 2. Correlation matrix for predictors in heart disease dataset.

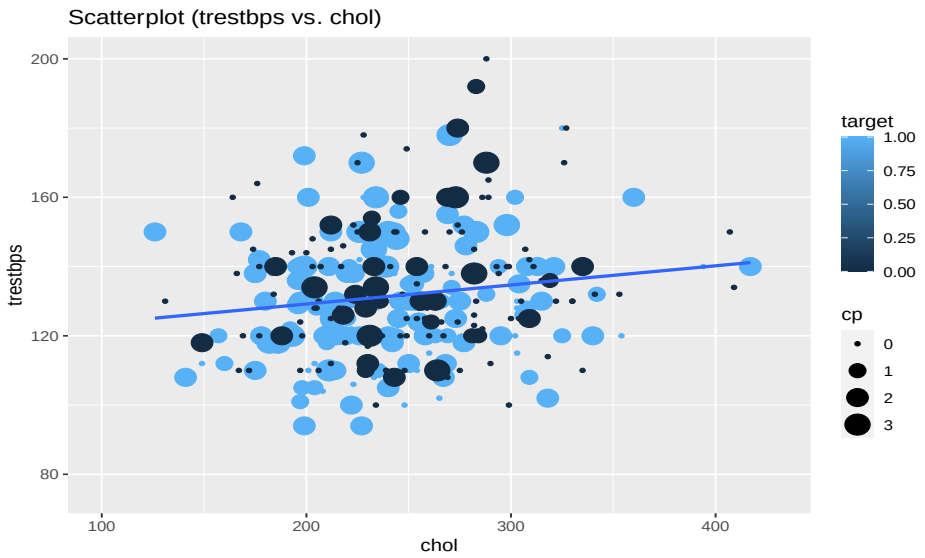

Figure 3. Scatter plot (trestbps vs. chol).

6. Results and discussion

The dataset was split up into training (212 observations) and testing (91 observations) parts. SVM, ANN, classical logistic, and Bayesian logistic classification methods are applied for testing and training datasets. The methods were compared according to their accuracy and consuming time. **Table 3.** compares four methods by calculating the accuracy for both training and testing datasets and the total time-consuming. The best accuracy exists in the Bayesian Logistic method, which is 91.38% in the training dataset and 90.92% in the testing dataset. This indicates the Bayesian Logistic method is the best. Also, Bayesian Logistic consumes less time than other methods.

Table 3. Comparison among SVM, ANN, Classical Logistic, and Bayesian Logistic methods.

Methods	Accuracy		Total time
	Training dataset	Testing dataset	
SVM	88.60%	84.42%	7 seconds
ANN	86.80%	80.55%	9 seconds
Classical Logistic	90.10%	86.39%	6 seconds
Bayesian Logistic	91.38%	90.92%	4 seconds

The accuracy curve for both classical and Bayesian logistic is shown in figure 4. We can see the value of AUC for Bayesian logistics is higher than classical logistic. In general, the best model exists with a higher AUC. For example, the accuracy in the testing dataset in Bayesian Logistic is 90.92 %. The model can discriminate between individuals (patients) with heart disease and no heart disease with an excellent prospect.

In general, the receiver operating characteristic (ROC) curve with a level of 0.7 appears to be very good, and hence the true positives are maximized. The highest quantity of patients with disease is not recognized as well. The higher the AUC, the more the model distinguishes between people who have the disease and others who do not.

Figure 4. shows a comparison between Bayesian and classical logistic classifiers. Overall, the Bayesian logistic classifier performs better than classical logistic. By allowing only a few samples (less than 100), the steep slopes of the recall curves for the point estimate classifiers (top row, red lines) suggest that they better identify patients than the classical logistic classifier. A smaller slope indicates less progress in performance as more patients are included. These trends are also reflected in the overall measurement of balanced accuracy.

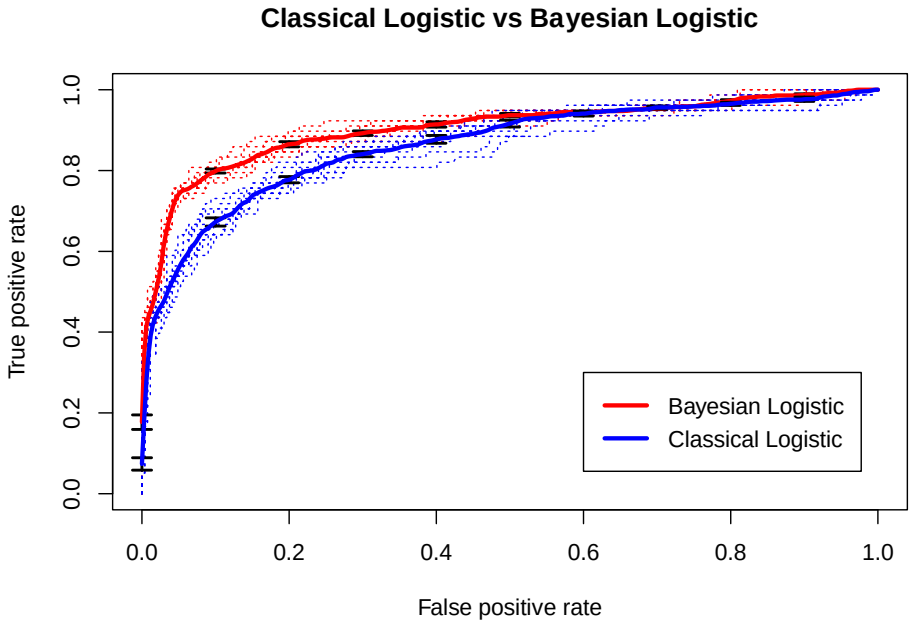

Figure 4. Accuracy curve for both classical and Bayesian Logistic.

7. Conclusion

The Bayesian Markov Chain Monte Carlo (MCMC) technique is presented in this article as an alternative method for estimating the logistic regression model. We derived a Bayesian logistic model by using a particular type of MCMC, the Metropolis-Hastings (MH) algorithm. The MH algorithm was introduced and used to obtain the estimated parameters in the new method. The Bayesian logistic model was used to overcome some limitations of the classical logistic model. The actual data of the heart dataset is used in this paper to classify healthy and non-healthy people. The modified method is compared with the classical logistic, SVM and ANN. It is shown that our method performs better than other methods and takes less time.

Acknowledgment

The authors are very grateful to the executive manager and editorial board members of the College of Administration and Economics

Conflict of Interest

The authors declare that they have no conflicts of interest.

Funding

There is no funding for the article.

References

1. Zhang, L.; Dong, L.; Cheng, S.; Li, W.; Wang, B.; Liu, H.; Chen, K. Efficient reliability assessment method for bridges based on Markov Chain Monte Carlo (MCMC) with Metropolis-Hasting Algorithm (MHA). In *IOP Conference Series: Earth and Environmental Science* **2020**, *580(1)*, 012030. https://doi.org/10.1088/1755-1315/580/1/012030
2. Mahdi GJ, Kalaf BA, Khaleel MA. Enhanced Supervised Principal Component Analysis for Cancer Classification. *Iraqi Journal of Science* **2021**, *62(4)*, 1321-1333. https://doi.org/10.24996/ijs.2021.62.4.28
3. Kaji T; Ročková V. Metropolis–Hastings via Classification. *Journal of the American Statistical Association*. **2023**, *118(544)*, 2533-2547. https://doi.org/10.1080/01621459.2022.2060836
4. Sur, P.; Candès, EJ. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **2019**, *116(29)*, 14516-25. https://doi.org/10.1073/pnas.1907936116
5. Belenguer-Llorens, A.; Sevilla-Salcedo, C.; Desco, M.; Soto-Montenegro, M.L.; Gómez-Verdejo, V. A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data. *Applied Science* **2022,** *12(5)*, 2571. https://doi.org/10.3390/app12052571
6. Jeune, W.; Francelino, M.R.; Souza, E.D.; Fernandes Filho, E.I.; Rocha, G.C. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. *Revista Brasileira de Ciência do Solo* **2018**, *42*, e0170133. https://doi.org/10.1590/18069657rbcs20170133
7. Liu, Y.; Bi, J.W.; Fan, Z.P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. *Information Sciences*. **2017**, *394*, 38-52. https://doi.org/10.1016/j.ins.2017.02.016
8. Dwivedi, AK. Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Computing and Applications* **2018**, *29(12)*, 1545-54. https://doi.org/10.1007/s00521-016-2701-1
9. Algamal, Z. An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. *Electronic Journal of Applied Statistical Analysis* **2017**, *10(1)*, 242-56. https://doi.org/10.1285/i20705948v10n1p242

10. Ding, Y.J.; Wang, Z.C.; Chen, G.; Ren, W.X.; Xin, Y. Markov Chain Monte Carlo-based Bayesian method for nonlinear stochastic model updating. *Journal of Sound and Vibration* **2022**, *520*, 116595. https://doi.org/10.1016/j.jsv.2021.116595
11. Guo, Y.; Li, Z.; Liu, P.; Wu, Y. Modeling correlation and heterogeneity in crash rates by collision types using full Bayesian random parameters multivariate Tobit model. *Accident Analysis & Prevention* **2019**, *128*, 164-74. https://doi.org/10.1016/j.aap.2019.04.001
12. Serrano, BM.; González-Cancelas, N.; Soler-Flores, F. Camarero-Orive, A. Classification and prediction of port variables using Bayesian Networks. *Transport Policy* **2018**, *67*, 57-66. https://doi.org/10.1016/j.tranpol.2017.03.016
13. Algamal, ZY.; Alhamzawi, R.; Ali, HT. Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression. *Computers in biology and medicine* **2018**, *97*, 145-52. https://doi.org/10.1016/j.compbiomed.2018.04.002
14. Mahdi GJM, Mohammed NJ, Al-Sharea ZI. Regression shrinkage and selection variables via an adaptive elastic net model. In *Journal of Physics: Conference Series* **2021,** *1879(3)*, 032014. https://doi.org/10.1088/1742-6596/1879/3/032014
15. Rai A, Chatterjee S, Nag A. A novel hybrid machine learning approach for reliable bridge condition assessment. *Structure and Infrastructure Engineering* **2021**, *17(4)*, 489-502. https://doi.org/10.1080/15732479.2020.1838954
16. Wu, C., Jin, Z.; Sun, J. Structural reliability assessment using adaptive surrogate models and Markov Chain Monte Carlo simulation. *Engineering Structures* **2020,** *222*, 111088. https://doi.org/10.1016/j.engstruct.2020.111088
17. Liang QQ, Li Y, Li AQ, Li J. Application of Bayesian network and support vector machine to bridge safety assessment. *Structural Safety* **2019**, *78*, 52-62. https://doi.org/10.1016/j.strusafe.2018.12.003
18. Huang, H.; Han, Y.; Sun, J. Reliability-based design optimization of steel-concrete composite bridges considering load and resistance factors. *Journal of Bridge Engineering* **2020**, *25(3)*, 04019137. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001513
19. Wang, S.; Li, Q.; Ma, J. A Bayesian approach to reliability assessment for corroded steel bridges using structural health monitoring data. *Journal of Constructional Steel Research* **2021**, *180*, 106618. https://doi.org/10.1016/j.jcsr.2021.106618
20. Lee, K.; Yang. M.; Suh, MW. Reliability analysis of bridges using a stochastic finite element method and surrogate models. *Computers and Structures* **2020**, *238*, 106306. https://doi.org/10.1016/j.compstruc.2020.106306
21. Wu, Y.; Zhang, S.; Cai, CS. Structural reliability analysis of long-span bridges under heavy traffic load and strong wind. *Journal of Structual Engineering* **2020**, *146(9)*, 04020182. https://doi.org/10.1061/(ASCE)ST.1943-541X.0002724
22. Li, S.; Liang, X.; Zhang, Z. A novel hybrid method for reliability assessment of existing bridges under heavy traffic loads. *Advances in Mechanical Engineering* **2021**, *13(2)*, 1687814021993540. https://doi.org/10.1177/1687814021993540
23. Zhang, Q.; Liu, H.; Xu, Z. Reliability-based maintenance optimization for deteriorating bridges using a multi-objective approach. *Automation Construction* **2020,** *117*, 103258. https://doi.org/10.1016/j.autcon.2020.103258
24. Yuen, KK.; Wong, SS.; Yeung, TY. Reliability assessment of bridge structures using non-parametric Bayesian method with Gaussian process. *Reliability Engineering & System Safety* **2021,** *206*, 107279. https://doi.org/10.1016/j.ress.2020.107279