Original Article:

Prediction of low birth weight using Random Forest: A comparison with Logistic Regression

Parisa Ahmadi[1], Hamid Alavi Majd[*1], Soheila Khodakarim[2], Leili Tapak[3], Nourossadat Kariman[4], Payam Amini [5],Forough Pazhuheian[1]

[1]Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
[4]Department of Statistics, School of Nursing and Midwifery, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Department of Epidemiology and Reproductive Health, Reproductive Epidemiology Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran

*Corresponding Author: email address: alavimajd@gmail.com (H. Alavi majd)

ABSTRACT

Low birth weight (neonate weighing less than 2500 g) is associated with several maternal and fetal factors, all interrelated with each other [1]. This study is aimed to survey maternal risk factors associated with low birth weight neonates using data mining (Random Forest) to account for interactions between them. We also intended to compare Random Forest with traditional Logistic regression. The dataset used in the present study consisted of 600 volunteer pregnant women. This cross-sectional study was carried out in Milad hospital, Tehran, during 2005-2009. Ten potential risk factors that are commonly associated with low birth weight were selected by using Random Forest technique. Several criteria such as the area under ROC curve were considered in comparing Random Forest with Logistic Regression.According to both criteria, four top rank variables identified by Random Forest were pregnancy age, body mass index during the third three months of pregnancy, mother's age and body mass index during the first three months of pregnancy, respectively. In addition, in terms of different criteria the Random Forest technique outperformed the Logistic regression (area under ROC curve: 93% ; Total Accuracy:95% ; Kappa Coefficient: 66%).The results of the present study showed that using Random Forest improved the prediction of low birth weight compared with Logistic Regression. This is because of the fact that the former accounts for all interactions between covariates. Therefore, this approach is a promising classifier for predicting low birth weight**.**

Keywords: Random Forest; Logistic Regression; Learning Theory; Low Birth Weight

INTRODUCTION

Birth weight is one of the most essential health indicators and survival of newborns and infants. Low birth weight (LBW) is defined as child's birth weight lower than 2500 g [1]. Several maternal and fetal factors have been determined as risk factors of LBW and the incidence rate of LBW have been estimated by several studies in different countries [2]. A low birth weight incidence has been reported about 6.8% for Iran, where 52.3 of these were preterm and 47.8% were the result of intrauterine growth restriction[3]. The prevalence of low birth weight in the United States, Europe, Asia and African regions were 10, 6.4, 18.3, and 14.3 percent based on global survey conducted by UNICEF in 2004[4]. LBW is associated with many socioeconomic factors such as known factors for pre-term delivery and fetal growth retardation which are associated with LBW. They include low maternal food intake and illness, specifically infections. Studies suggest that short maternal stature, very young age, high parity and close birth spacing were all associated factors [5]. There are several socioeconomic factors associated with LBW including residence (urban-rural difference),

mother's age and occupation, birth order, the family's income and many maternal conditions such as nutritional status, mother's educational and health status. Moreover, other factors like low maternal food intake and illness, especially infections are known factors for pre-term delivery and fetal growth retardation related to LBW[6]. Other studies have identified maternal factors related to LBW including demographic variables and medical conditions, such as maternal age, nulliparity, cigarette smoking , short stature , caffeine intake , low or high maternal body mass index (BMI), hypertension and preeclampsia, psychosocial stress, and socioeconomic status, including education [7]. Classification, one of the most important application of statistical methods in various sciences, aims to predict a multicategory response based on some independent variables on subjects [8]. There are several limitations for traditional methods like logistic regression including normality assumption, homogeneity of variances in several groups, linearity, independence, collinearity of covariates associations and interactions [9]. Therefore, using new methods with sufficient prediction accuracy that do not suffer from these limitations are of great interest. Recently, machine learning techniques that provide less prognosis bias and more explicit results has been increasingly receiving much interest in medical diagnosis [10]. Random forest (RF), an extension of classification and regression trees (CART) is a non-parametric and supervised learning group method which has showed a promising performance in several studies [11]. The present study was conducted to determine prevalence of LBW and its associated risk factors in Milad Hospital of Tehran city using RF. We also compared the performance of RF with classical logistic regression in the used data set.

MATERIALS AND METHODS

The data used in this study includes 600 volunteer pregnant women, (1-13 weeks of gestation) referring to Milad Hospital in Tehran and receiving prenatal care from 2009 to 2010. Among 600 birth, 57 (9.5%) were low where α is an intercept and β is a coefficients vector [13]. RF is an "ensemble learning" method in classification problems proposed by Breiman et al to increase classification accuracy weighted. The samples were followed up until the delivery time. In order to collect the data, a checklist was designed by the researcher in three sets. The first set contains demographic logistic and pregnancy data including the duration of folic acid tablet use and medical tests in the first, second and third trimester of pregnancy. The second set contains the information about pregnancy features, medical tests and troubles during pregnancy. And the third set presents the data about pregnancy features, type of childbirth, the baby and common drugs used during pregnancy collected by observing and taking interviews. Moreover, some devices such as blood glucose meter, blood pressure monitor, urine analyzer, mercury sphygmomanometer tools, and meters connected to the scale, adult and baby scales and stethoscope were utilized to measure the variables [12]. In order to evaluate the validity and reliability of tools, content validity and test-retest were applied respectively. The first, second and third sets of the information forms were completed respectively by 30 pregnant women, including 10 women in their first three months of pregnancy, 10 women in their second three months and 10 women before the childbirth. After two weeks, the 30 pregnant women were asked to recomplete the forms. In order to check the reliability and the correlation among the responses of several tests were performed including, McNamara Kappa coefficient, Pearson and Spearman correlation coefficients. The reliability was estimated as 0.8 -1

Data Analysis

Logistic Regression is a very general analytical tool that is utilized in many epidemiological studies to predict a binary response variable through covariates and factors [13]. For binary response variable, Y, a vector of covariates, x, and $\pi(x)$ that represents the probability of success given a specific value of X, the logistic regression model can be presented as follows:

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

as well as to prevent over-fitting issue [14]. In this method, a series of unpruned classification trees using random bootstrap samples of the original data sample is constructed. To produce

one final classification, the outputs of all trees are aggregated and the object belongs to a class There are several advantages for RF over regression, theoretically. First, the RF algorithm can select important variables automatically no matter how many variables are used initially which is different from stepwise variable selection in logistic regression. Second, missing values as well as imbalanced data can be handled automatically by RF[15]. Third, RF works better than logistic regression in large data sets, where the numbers of variables are big. There is also a main disadvantage for RF. Unlike decision trees analysis, it is very complex and the tree structure is in an invisible "black box". Hence, there is an unknown relationship between a particular level of a variable and the outcome [15].

Performance criteria

To compare the discriminative powers of the two models, receiver operating characteristic (ROC) curves and the area under the curves (AUCs) for the data sets were used. Sensitivity, with the majority of predictions given by the trees in the random forest[14] .

specificity, positive predictive value (PPV), negative predictive value (NPV), total accuracy and kappa coefficient were calculated as well. The data was analyzed using statistical R software version 3.2.2.

RESULTS

Among 600 births, there were 57 (9.5%) cases of low weight. The results of t-test and chi-square tests are shown in table 1. Factors such as Mother's age, the number of previous abortions, the number of previous childbirth, gestational age at the time of delivery, body mass index at the first and third trimesters of pregnancy, mother's education and job, child's gender, the use of folic acid iron, calcium or multi-vitamins during pregnancy, developing diabetes or preeclampsia during the third trimester of pregnancy all are significantly associated with LBW.

Table 1. Descriptive statistics of variabels

Characteristic	Value	
	LBW(NO)	LBW(YES)
Mother age	65.1787±11.85	64.28±11.71
The number of weeks of pregnancy	38.65±1.33	34.91±2.93
BMI1 (at the first three months of pregnancy)	24.97±4.36	25.037±4.49
BMI3 (at the third three months of pregnancy)	29.874±4.13	472±4.2929
HB1 (at the first three months of Pregnancy)	12.64±1.01	12.88±1.04
HB3 (at the third three months of pregnancy)	12.51±1.07	12.60±1.19
Delivery Type	0.515±0.604	0.354±0.51
BMI1_CAT		
<19.8	8(0.12)	49(0.09)
19.8-26	32(0.51)	294(0.57)
26-29	10(0.16)	108(0.20)
>29	12(0.19)	86(0.16)
BMI3_CAT		
<19.8	0	1(0.05)
19.8-26	13(0.2)	98(0.8)
26-29	15(0.24)	141(0.26)
>29	34(0.54)	297(0.55)
PREClamsia		
YES	14(0.22)	26(0.048)
NO	48(0.77)	511(0.25)
EDUCATION WOMAN		
Illiterate	0	2(0.003)
School	2(0.03)	14(0.02)
Guidance	35(0.56)	383(0.71)
Collegiate	25(0.4)	138(0.25)
Use folic in pregnancy		
Yes	59(0.95)	517(0.96)

No	3(0.04)	20(0.03)
Baby sex		
Boy	31(0.52)	206(0.38)
Girl	31(0.52)	281(0.52)
Diabetes		
Yes	7(011)	42(0.07)
No	55(0.88)	495(0.92)
Use iron in pregnancy		
Yes	60(0.96)	527(0.98)
No	2(0.03)	10(0.01)
Use calcium in pregnancy		
Yes	37(0.59)	176(0.32)
No	25(0.40)	421(0.78)
Use vitamin in pregnancy		
Yes	52(0.83)	421(0.78)
No	10(0.16)	116(0.21)
Job woman		
Housewife	52(0.83)	449(0.83)
Work at home	0	6(0.01)
Employee	10(0.16)	77(0.14)

Random forest analysis contained 500 trees where 4 variables were considered in each tree resulting in 17 independent variables ($\sqrt{P}$)[16]. K-fold method (k=2) was used to check for the validity of the results and make the results comparable to those from the logistic regression,. The testing set contained one third of the data and the others were used in training set. The learning algorithm was applied on the testing set and the training set was used for the supervised algorithm. The random forest analysis was performed using the package "random Forest" from R software. Moreover, the R package ROCR was used to build the Roc curve and calculating the goodness of fit indexes. According to the Gini coefficient index and permutation importance index, the variables are ordered based on their effect on classification (Figure 1)[17].

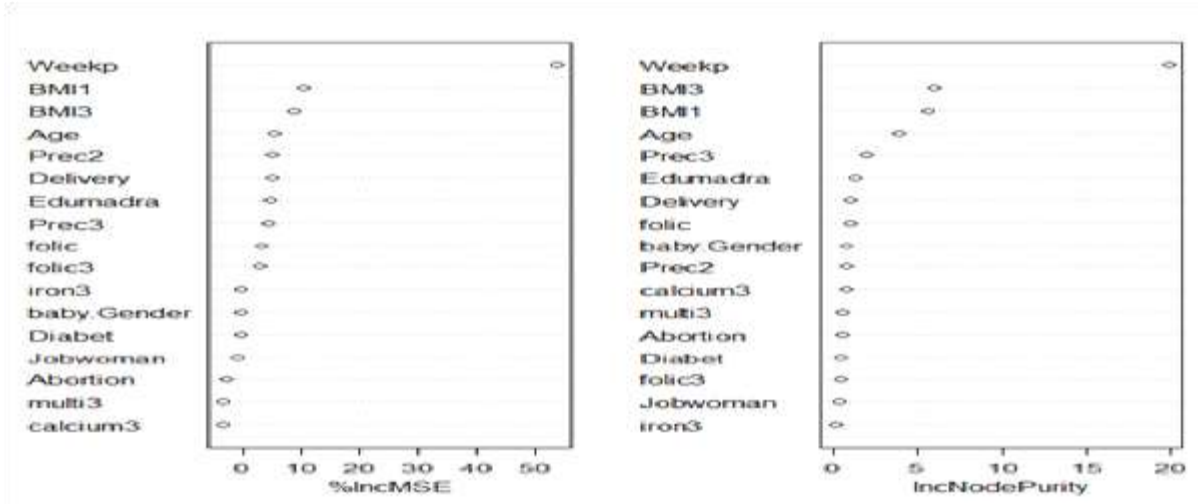

Figure 1. The Gini coefficient index and permutaion importance index in RF

In accordance with the Gini importance index (right side of figure 1), pregnancy age during childbirth, body mass index during the third trimester of pregnancy, mother's age and body mass index during the first trimester of pregnancy are respectively the most important variables affecting LBW. The order of importance according to permutation importance index is as pregnancy age during childbirth, body mass index during the first and third trimester of pregnancy and mother's age.

Marginal effect plots can help assessing the impact of the first two important variables, i.e. the pregnancy age during childbirth, and the body mass index during the third trimester of pregnancy (Figure2). Based on figure 2, one can

find out a decrease in the probability of LBW as the number of weeks of pregnancy increases where before week 32, there is a slight decrease in the probability of LBW and a considerable decrease can be observed after that. A reverse trend is shown for body mass index during the third trimester of pregnancy at week 25 where some increases can be seen after that time.

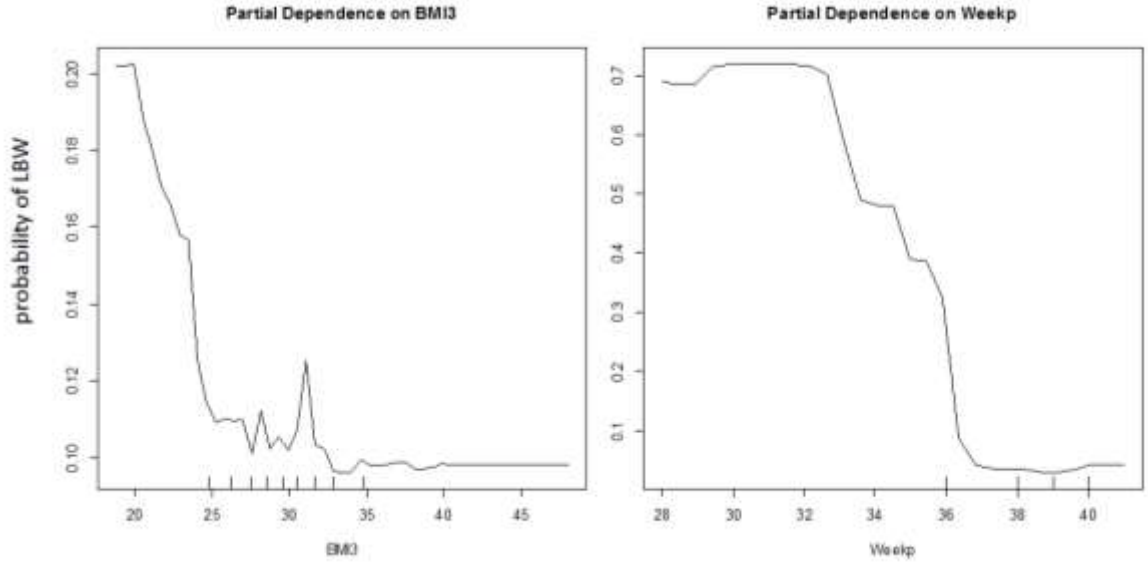

Figure 2. Marginal effect plots assessing the impact of pregnancy age during childbirth and body mass index during the third trimester of pregnancy

These plots can be presented for all other independent variables and their association with the classifier variable can be assessed. Increasing the number of trees reduces the-out-of-bag error. Figure 3 shows that 500 trees can be called enough where after 400 trees the OOB converges to zero[18].

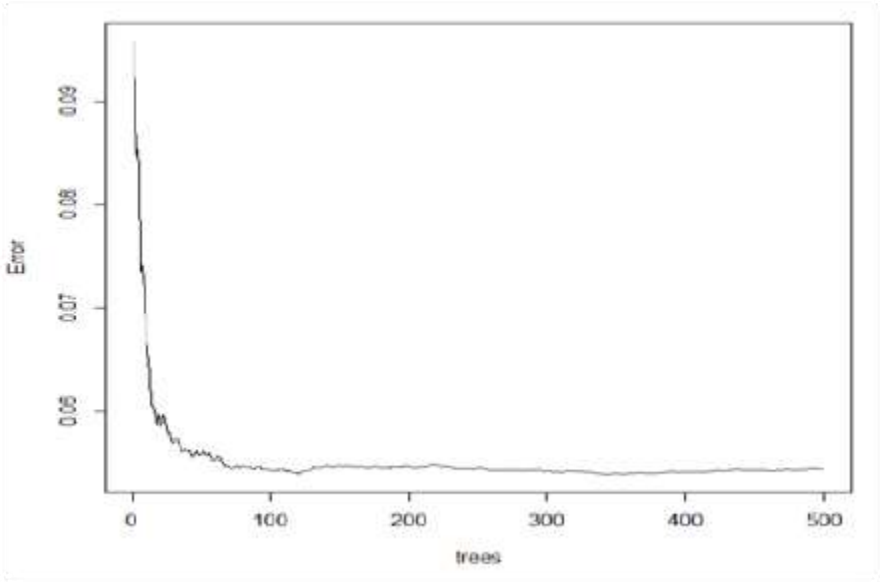

Figure 2. Out-of-bag error versus number of trees

The logistic regression model was performed using R software including the significant variables resulted from the univariate analysis. Entering the variables in this generalized linear regression model was according to the stepwise method and the best model was carried out using

Akaike Information Criterion (AIC). The goodness of fit was checked using Hosmer-Lemeshow test. According to the potential collinearity among independent variables and also data sparsity, the stepwise method was considered to choose the best one among several models with less associated independent variables. After 16 steps, the best model revealed a significant influence of gestational age at the time of delivery and preeclampsia during the third trimester of pregnancy on LBW. Table 2 shows that the odds of LBW are 0.43 as one year increase in age at the time of delivery. In other words, the probability of LBW decreases as age at the time of delivery increases. This ratio was 5 for preeclampsia during the third trimester of pregnancy, showing a booster effect on LBW. The Hosmer-Lemeshow statistic was not significant demonstrating a good fit of logistic regression.

Table2. The results of logistic regression evaluating LBW

Variable	OR (95% CI)	P-value
gestational age at the time of delivery	0.43 (0.36-0.53)	<0.001
preeclampsia during the third trimester of pregnancy	5.02 (1.92-13.04)	<0.001

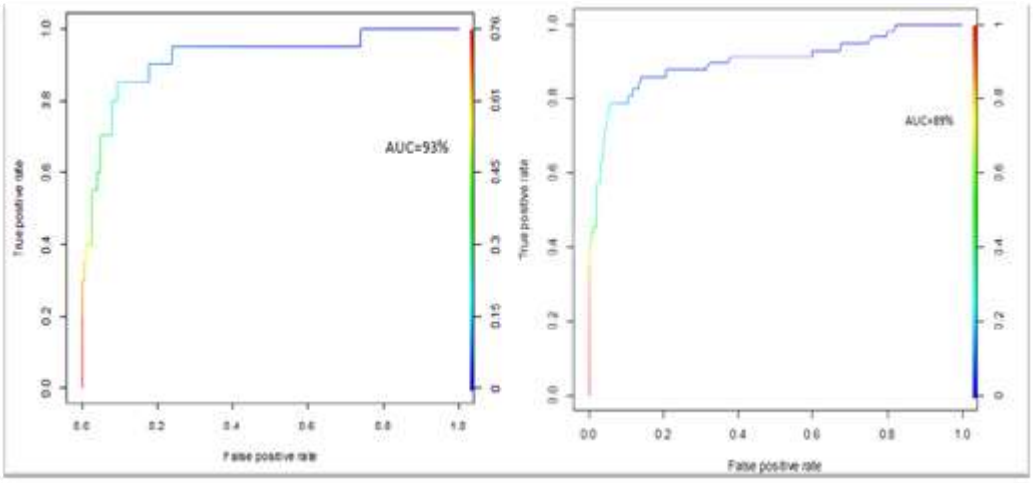

Figure 3. ROC curves for LR and RF methods

Figure 4 presents the ROC curve for LR and RF methods. The AUC for RF and LR methods were 93% and 89%, respectively. This result exposes an outperformance of RF method. Details about the accuracy, sensitivity and specificity for the performed methods are shown in table 3. A better performance was resulted for RF compared with LR.

Table3. Accuracy, sensitivity, specificity and kappa coefficient for LR and RF methods

Goodness of fit indexes	RF	LR
Sensitivity	72%	67%
Specificity	97%	96%
Accuracy	95%	93%
Kappa coefficient	66%	62%

DISCUSSION

LBW is the leading cause of mortality in newborns and infants; they, along with congenital malformations, are the major causes of morbidity[19]. This study was conducted to investigate the risk factors of low birth weight. Two statistical approaches of logistic regression and random forest were utilized. Our results showed that RF outperformed to the logistic regression in the used data set in terms of several criteria. Univariate analysis also revealed factors related to low birth weight, including the age of marriage, mother's age, gestational age, BMI, first pregnancy age, distance between last

pregnancy and recent pregnancy, birth weight of last child, and finally infant weight where they all had obviously a positive effect on weight of new born infants, while unexpected pregnancy had a negative effect and caused LBW. After assessing the LWB risk factors in the study, it was concluded that pregnancy week, BMI3, BMI1 and mother's age were the most important variables. The most important risk factor in all studies and also in our study was the number of weeks of pregnancy. The results showed that the probability of LBW decreases as the number of weeks of pregnancy increases. A slight decrease was observed in the probability before week 32 and a considerable decrease after that which is in concordance with the results of other studies. In the present study, mother's age was identified as an important risk factor for LBW. According to the study conducted by MRCOG et al, mothers aged between 18 and 35 years had the lowest prevalence of LBW in their children and the highest prevalence was observed in mothers younger than 18. Low parity in >35 years old mothers and the appropriate gap between two pregnancies in this group could be the cause of compensation of high age risk (obtained by most studies) as well[20]. accordance with the studies by Gebremedhin and Mirzarahimi , we found a relationship between LBW and body mass index [21, 22]. RF is an appropriate method for datasets with small sample sizes and huge number of variables as well as datasets with collinearity among independent variables and high dimension interactions [23]. Ignoring the order of entering the variables to build tree partitions in regression tree, this method is appropriate for huge sample size with thousands of variable, determining the most important variables in classification, appropriate for data with missing, adjusting classification errors where the number of cases for each dependent variable category are unbalanced [23]. Calculating the proximity measures is performed for each pair of cases for classification, finding outliers and evaluating the data[24]. Logistic regression is a classic and parametric method carrying some limitations such as distribution assumption for the response variable, collinearity among independent variables and missing data in contrast to the modern non-parametric random forest method which is based on machine learning. In order to make the comparison between two performed methods possible, one third of the data was considered as testing set and the random forest analysis was applied and evaluated using the rest of data as training set[25]. All goodness of fit indexes resulted in a weaker performance of logistic regression (AUC, sensitivity, specificity, accuracy and kappa coefficient). Although the logistic regression models didn't use the testing set, the random forest analysis using testing set was better in prediction in comparison with logistic regression using all the dataset. The predicting, although resulting in different powers for two methods was the same in recognizing important affecting variables where both exposed gestational age at the time of delivery had a reverse association with LBW. A restriction caused by the data was an imbalance distribution of the dataset in different categories of response variable (9.5% were low birth weighted) in addition to collinearity. However, a better prediction was achieved from the random forest. Another restriction in logistic regression was the sparsity of the data in classifier categories while random forest was free from this limitation.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

ACKNOWLEDGMENT

We would like to thank the Vice-Chancellor of Health of Shahid Beheshti University of Medical Sciences for their approval and support of this study.

"The authors declared no conflict of interest."

REFERENCES

1.Gebremedhin, M., et al., Maternal associated factors of low birth weight: a hospital based cross sectional mixed study in Tigray, Northern Ethiopia. BMC pregnancy and childbirth, 2015. 15(1): p. 1.

2.Alizadeh, M., H.J. Birami, and S. Moradi, Analysis of trends in birth outcomes and fertility measures in the rural population of east Azerbaijan province, Iran: 2001-2013. 2015.

3.Ota, E., et al., Risk factors and adverse perinatal outcomes among term and preterm infants born small-for-gestational-age: secondary

analyses of the WHO Multi-Country Survey on Maternal and Newborn Health. PLoS One, 2014. 9(8): p. e105155.
4.Palve, S. and A. Shenoy, Study to Assess Sociodemographic Factors Affecting Low Birth Weight Baby in Urban Slum of Mumbai, Maharashtra ,India. International Journal of Scientific Research, 2016. 5(2.(
5.Siza, J., Risk factors associated with low birth weight of neonates among pregnant women attending a referral hospital in northern Tanzania. Tanzania journal of health research, 2008. 10 :(1)p. 1-8.
6.Rajaeefard, A., M. Mohammadi, and A. Choobineh, Preterm delivery risk factors: a prevention strategy in Shiraz, Islamic Republic of Iran. 2007.
7.Hamta, A., et al., Path Analysis of the Risk of Low Birth Weight for Multipara. Iranian Red Crescent Medical Journal, 2013. 15(6): p. 462.
8.Duda, R.O., P.E. Hart, and D.G. Stork, Pattern classification. 2012: John Wiley & Sons.
9.Harrell, F., Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2015: Springer.
10.Cruz, J.A. and D.S. Wishart, Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2006. 2.
11.Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.
12.McGrath ,M.J. and C.N. Scanaill, Body-Worn, Ambient, and Consumer Sensing for Health Applications, in Sensor Technologies. 2013, Springer. p. 181-216.
13.Hosmer Jr, D.W. and S. Lemeshow, Applied logistic regression. 2004: John Wiley & Sons.
14.Svetnik, V., et al ,.Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 2003. 43(6): p. 1947-1958.
15.Geng, M., A COMPARISON OF LOGISTIC REGRESSION TO RANDOM FORESTS FOR EXPLORING DIFFERENCES IN RISK FACTORS ASSOCIATED WITH STAGE AT DIAGNOSIS BETWEEN BLACK AND WHITE COLON CANCER PATIENTS, 2006, University of Pittsburgh.
16.Strobl, C., et al., Conditional variable importance for random forests. BMC bioinformatics, 2008 :(1)9 .p. 1.
17.Boulesteix, A.-L., et al., Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Briefings in Bioinformatics, 2012. 13(3): p. 292-304.
18.Goldstein, B.A., et al., An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics, 2010. 11(1): p. 1.
19.Dollfus, C., et al., Infant mortality: a practical approach to the analysis of the leading causes of death and risk factors. Pediatrics, 1990. 86(2): p. 176-183.
20.Jolly, M., et al., Obstetric risks of pregnancy in women less than 18 years old. Obstetrics & Gynecology, 2000. 96(6): p. 962-966.
21.Neggers, Y., et al., The relationship between maternal and neonatal anthropometric measurements in term newborns. Obstetrics & Gynecology, 1995. 85(2): p. 192-196.
22.Mirzarahimi, M., et al., Prevalence and risk factors for low birth weight in Ardabil, Iran. Iranian Journal of Neonatology IJN, 2013. 4(1): p. 18-23.
23.Breiman, L., C. Chen, and A. Liaw, Using random forest to learn imbalanced data. J. of Machine Learning Research, 2004(666.(
24.Rodriguez-Galiano, V.F., et al., An assessment of the effectiveness of a random forest classifier for land-cover classification . ISPRS Journal of Photogrammetry and Remote Sensing, 2012. 67: p. 93-104.
25.Ruiz, A. and N. Villa, Storms prediction: logistic regression vs random forest for unbalanced data. arXiv preprint arXiv:0804.0650, 2008.