Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques

Senthilkumar D
University College of Engineering, Anna University
Regional Centre,Tiruchirappalli
India
chandsent@yahoo.co.in

Paulraj S
College of Engineering, Anna University
Tamilnadu, India
paulraj@annauniv.edu

Abstract
High rates of maternal mortality, infant mortality, and preterm births, as well as continuing disparities in pregnancy outcomes are an important public health issue in India and worldwide. Health care industries must focus on improving the quality of treatment and continuing care of pregnant women. This study aims comparison of logistic regression with data mining techniques to identify most influenced predictor variables and to develop a decision support system to help the physicians for better decision making in low weight child birth. It is identified that the variables which were highly influenced to predict the low weight child birth are Mother's last weight (pounds) before becoming pregnant, Mothers age, Number of physician visits during the first trimester, Number of previous premature labors. The results of this work have improved prediction accuracy in Datamining techniques when compared to logistic regression.

Key words
Low birth weight, Healthcare, Data mining, Logistic regression, Classifier accuracy.

1. Introduction
Health care is a big concern in India, the land of nearly 1.12 billion people and the second most heavily populated countries in the world [12]. India has relatively poor health outcomes, despite having a well-developed administrative system, good technical skills in many fields, and an extensive network of public health institutions for research, training and diagnostics [48].

The main objective of health care provider is to strengthen the health care system by developing better identify and track disease states and high risk patients, the data mining applications can be developed to better, design the appropriate interventions and reduce the number of hospital admission and claims [51].

Health care industry stores large amounts of data in the form of patient records. It is a key resource for medical research contains hidden information. Data mining methods are used to extract hidden information and relationships and applied to better decision making in many areas like marketing, fraud detection, investment, manufacturing, telecommunication, engineering, medicine, biomedical research [16],[4],[5],[69]. It helps the health care providers for better diagnosis and treatment.

Low birth weight and preterm birth are identified as a leading cause of infant mortality and are at heightened risk. It is a risk factor for brain damage, chronic lung and liver disease, deafness, blindness, epilepsy, learning disabilities, cerebral palsy, mental retardation, neurological disorders, physical disability, respiratory disease, gastrointestinal problems, High blood pressure, Type II diabetes and attention deficit disorder [33],[1],[20],[41], [35].

Birth weight and preterm birth are associated. Preterm birth contributes to 35% of the world's 3.1 million neonatal deaths each year, making it the leading cause of infant mortality [49]. The child mortality rate age <5 years was 63/1000 live births in 2010 (MDG target 38). The maternal mortality ratio was 254/100 000 live births in 2004-6 (MDG target <100). The prevalence of underweight children was 43% in 2005-6 (MDG target 27%). There are 1.8 million child deaths and 68,000 maternal deaths in India each year and 52 million undernourished children [42].

Improving the health status of pregnant women and infants is an important public health issue in India and worldwide needs to be urgently addressed because it leads to short and long term impact of individuals, families, society and health care system [42]. Accurate and non-invasive method of identifying patients who are at risk for preterm birth is greatly needed for both the pregnant woman and health care providers may be used to prevent preterm delivery and provide for improved treatment plans [41], [37], [15].

What is a "low birth weight" birth?

Low birth weight is defined by the World Health Organization as a birth weight less than 2,500 grammes (g)

since below this value birth weight-specific infant mortality begins to rise rapidly (Kramer, 1987). Many factors affect the prevalence of low birth weight, includes Anthropometry/ nutritional status, Micronutrients, Cigarette, smoking, Substance misuse, Work/physical Activity, Prenatal care women, Bacterial vaginosis and multiple birth. Low birth weight babies are at a greater risk of dying during the first year of life and more likely to have psycho-social problems, difficulties at school, and when they become teenagers, lower achievement on intellectual measures, particular arithmetic [35]. Hosmer and Lemeshow used logistic regression to predict low birth weight. Classification accuracy of logistic regression is very good for small to moderate sized datasets [72].

This study aims comparison of logistic regression with data mining techniques to identify important predictor variables and to develop a decision support system to help the physicians for better decision making in low weight child birth.

This paper is organized as follow: Section 2 discusses few studies related to low birth weight child, Section 3 discusses the importance of data mining in healthcare sector in the literature and Section 4 discusses various data mining algorithms used for this study, Section 5 explains the dataset used in the study, Section 6 covers the comparative analysis of these algorithm and Section 7 concludes the paper with future work.

2. Related works

In the recent years, researchers of medical field attempted to predict preterm birth and the associated risk factors. We discuss a few studies here that are relevant to our work. Yavar Naddaf, et al. have previously attempted to predict preterm birth using a very rich dataset resulted poor prediction performance collected by Northern and Central Alberta Perinatal Outreach Program between 1992 and 2003 and suggests that predicting preterm birth is a very challenging problem. The dataset contains maternal and newborn data for 243948 cases, including 21193 preterm cases. There are 244 attributes, containing "maternal demographic information, medical history, such as pre-existing chronic illness, lifestyle information such as smoking and alcohol use, past reproductive history, including previous [preterm] or [small for gestational age] delivery, and history with the current pregnancy such as presence of hypertension or toxaemia" [75].

Rabindra Nath Das, et al. used the dataset collected by 1986 at Baystate Medical Center, Springfield, Massachusetts to identify a causal relationship between the risk factors of Low birth weight using Joint generalized linear log-normal statistical modeling. The traditional, simple, multiple, logistic regression and Log- Gaussian models (with constant variance) not more effective than the joint log-normal models (with non-constant variance) because they better fit the data [62].

Artificial Neural Network (ANN), Clinical Expertise and Multiple Linear Regression (MLR) models is applied for Predicting Extubation Outcome in Preterm Newborns, ANN performed better compared with MLR models and clinical expertise [45]. Mukhopadhyay, et al. applied chi-square test used for categorical variables and independent t test used for normally distributed continuous data and Mann- Whitney U test for skewed data to determine the association between each risk factor and outcome between Mortality and Major Morbidities in Extremely Low Birth Weight Neonates [36]. Ciaran S. Phibbs, et al. used logistic regression to estimate odds ratios for mortality associated with the NICU level of care and annual volume of very-low-birth weight infants [18].

Kleanthis C. Neokleous, et al. applied Neural networks to predict the risk for early spontaneous preterm delivery using various demographic, clinical, and laboratory inputs [38]. Namasivayam Ambalavanan, et al. compared Neural network and multiple logistic regression models in predicting death of extremely low birth weight neonates [52]. Nasreen et al. investigated the independence effect of maternal antepartum depressive and anxiety symptoms on infant low birth weight. It is identified that maternal depressive and anxiety symptoms during pregnancy need to be considered for low birth weight [19]. Yorifuji et al. examined whether socio-economic position and parental characteristics are having the relationship in adverse birth outcomes using logistic regression model [70]. Guillermo Marshall et al. analyzed the associations between in-hospital mortality and prenatal and admission infant characteristics using generalized additive logistic regression model and multiple logistic regression model [18].

Aparajita Dasgupta et al. applied cluster sampling for selecting the sample, univariate analysis for identifying the determinants of LBW and multiple logistic models. Identified that anaemia in pregnancy was significantly associated with LBW [2]. Masaki Ogawa et al. applied univariate analysis to identify the association between the causative determinant and obstertric complications and unconditional logistic regression used for multivariate analysis [46]. Nynke R. van den Broek et al. used multivariate logistic regression for finding the factors independently associated with preterm birth, early and late preterm birth. Women's pregnancy history and identified maternal underweight, malaria and anemia as risk factors for preterm birth ; HIV status does not contribute to the risk of preterm births [56].

3. Importance of Data mining in Health Care

Healthcare sector stores large amount of information about patients and their medical conditions. Medical data analysis plays vital role in decision making and management in healthcare. Large amount of data requires an automated method for analysis. Data mining is the process of selecting, exploring, and modeling large amounts of data to discover unknown patterns or relationships useful to decision making. Data mining has great importance for area of medicine, and it represents a

comprehensive process that demands thorough understanding of the needs of the healthcare organizations. Data mining tools reduce subjectivity and provide better knowledge for medical decision making. Discovered patterns improve the quality of the decision making process in health care [10] [6].

Discovered trends and hidden patterns from data mining significantly enhance our understanding of disease progression, management of the disease and to support medical diagnosis, improving quality of patient care, etc.. Classification analysis is the most commonly used in healthcare applications [68]. Different data mining techniques are applied in clinical decision support, explanatory and confirmative techniques are mostly used in medical data mining [14].

Data mining is a fast evolving technology of high importance for providing prognosis and a deeper understanding of the classification of neurodegenerative diseases, is being adopted in biomedical sciences and research [67]. Biomedicine database existing the phenomenon of "data rich, information poor". Development of database technology has solved the memory and retrieval of substantive data. Knowledge Discovery in Database (KDD) is likely to be of increasing importance in biomedical database can help us with disease diagnosis, treatment, research and decision making, by discovering the rules and mode of medical diagnosis [19].

Clinical databases have accumulated a large amount of data about patients. Knowledge discovery in databases. Data mining techniques have great importance of medical research databases to identify potentially useful relationships and patterns more efficiently to increase in volume of data earlier than using current methods used in decision making in a variety of contexts of healthcare organizations [43]. Medical decision making is highly specialized and challenging due to various factors, especially in the case of rare diseases or diseases that show similar symptoms. The amount of Medical data recorded in hospitals and its significance as an ever-growing source of information is used effectively by extracting valuable information to assist the medical fraternity using Data mining techniques including KNN classification and Neural Network, used to diagnose the most probable disease with the set of similar symptoms [64]. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. Data mining application can have tremendous potential and usefulness to improve the effectiveness of treatment, management of health care; customer relationship management; and detection of fraud and abuse [23].

4. Data mining Algorithms Used in this Study

4.1. Support Vector Machines

Support vector machines (SVM) are introduced by Cortes and effective method for binary classification, regression or ranking function and it is based on statistical learning. It is very popular used by the researcher in health care for classification due to many attractive features, handling complex non-linear data points. Its accuracy is high and less prone to over more fitting than other well-known classifier [23],[13],[22]. It is a good classifier, does not require a priori knowledge, even the input space is very high [65].

4.2. Logistic Regression

Logistic regression is a special case of generalized linear modelling, also called a logistic model or logit model and is extensively used for binary classification method in the medical, social sciences, marketing applications. It is used based on the assumption when the outcome of a situation is not linearly associated to the explanatory variables. It allows probabilistic interpretation; easily we can update the model for the new data, unlike decision trees or SVM and ease of interpretation. It has some drawbacks, not suitable for high-dimensional problems, it is slower than SVM, non-linearities and identifying interaction is difficult. The dependent variable is restricted to discrete number. It accepts large number of explanatory variable, in many situations it is not. The researcher should decide whether to use logistic regression for the classification if the data set is a large size. It is applied to the studies that using between subject design. It may be suitable in the fields of medicine and psychology, in fact; it is not a choice always [3],[[60],[53],[63],[71].

4.3. Neural Network

Neural Networks are a complex non-linear modelling method and able to learn complex relationship between dependent and independent variables without any external assistance based on a model of the neural architecture of the human brain. Neural Network is a successful technique applied for the real world application like accounting and auditing, finance, management, decision making, marketing, production, biology, psychology, handwritten character recognition, pathology, statistics, mathematics, computer science, medical research and many more.

Neural Network is a popular data mining tool because of its predictive power even in the complicated domain compared with statistical techniques. It will handle both categorical and continuous data types and tolerate noisy data. It supports parallelization techniques, which will speed up the computation process. Identifying patterns is very difficult and requires long learning time if the input features are large. Neural Networks are also called as "black boxes" due to its poor transparency to explain the process of neural networks built. Identifying the required number of parameters for modelling neural network is trial and error design like, network topology or structure, number of hidden layers, number of units in each hidden

layer an in output layer [22], [63], [47],[22],[7],[49]. .

Neural network performance can be highly automated, minimizing human involvement. It is very flexible with incomplete, noisy and missing data. It does not make any prior assumption about the distribution of the data or the form of interaction between factors. It can be easily updated with new data. The output of the Neural Network algorithm does not produce an explicit set of rules and it is lacking in classical statistical properties (confidence interval and testing of hypothesis) [61].

4.4. Naïve Bayes

Naïve Bayes Classifier very widely used, simple statistical Bayesian Classifier based on Bayes' theorem with strong class conditional independence assumption for classifying the data. This is an unrealistic assumption for most datasets; however, it leads to a simple prediction framework that gives surprisingly good results in many practical cases. A more descriptive term for the underlying probability model would be "independent feature model". It is also called idiot's Bayes, Simple Bayes and independence Bayes. It depends on the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [74],[58],[50].

Naive Bayes classifier will converge faster than logistic regression, so it requires only less training data. This method is very popular for different applications for several features. It can be trained very efficiently in a supervised learning setting and performance is better in many complex real life problem world situations. It can be used for binary and multi class classification and accurate in providing results (Zhang et al. 2004). The performance of the Bayesian classifiers is better compared to Decision Tree. It is robust, very easy to construct and fast. It does not require any iterative parameter estimation schemes. It can be applied to discrete and continuous attributes; large data set and easy to interpret (not required any skilled person). It may not be the best classifier in any particular application, but it relies on well and robustly. It is good at missing value handling. . Naive Bayes has a less error rate compared to the other classifier, but it is not always true practically and less accurate compared to other classifier. It can't learn interaction between features does not support pruning, contains sharp decision boundaries [74],[58],[50]..

4.5. Random Forest

Random forest is an ensemble classifier, like decision trees, can be used to solve classification and regression problems. It uses the concept of generating multiple random trees with, bootstrap of training dataset, bagging on samples, voting scheme and the features are randomly selected in each decision split, which improves the predictive power and results in higher efficiency. It achieves better results most of the time compared to decision trees. Selection of a random subset of features is an example of the random subspace method. It was founded in 2001 and used in the different number of applications, includes medical research, image processing, etc., [50], [32],[26], [54].

The advantages of random forest are; it does not depend on the data, appropriate for high dimensional data modelling, overcoming the problem of over fitting, eliminates prune the trees. It will generate the most important variable used for classification. It runs efficiently on large databases produce high prediction accuracy. It is good with dealing missing values, outlier and maintain accuracy when a large proportion of the data are missing. The model interpretability and prediction accuracy provided by Random Forest is very unique among popular machine learning methods. It also supports a method for detecting interaction between variables. The main disadvantage is observed to over fit for some datasets with noisy classification/regression tasks. [50], [32],[26],[54].

4.6. Decision Tree

A decision tree is a classifier used in statistics, data mining and machine learning for modeling classification and prediction. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. There are two varieties in decision tree used data mining are classification tree or regression tree. Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. Typically the goal is to find the optimal decision tree by minimizing the generalization error [25],[9],[31],[40],[21],[59]/

Its representation is easy to understand and interpret by non-professional user. It is capable of handling; both nominal and numerical input, requires little data preparation, data sets that may have errors and missing values, efficiency in processing with large datasets. It does not require any domain knowledge or parameter setting, and therefore appropriate for exploratory knowledge discovery. Reliability of the model can be validated using statistical tests. It has some disadvantages such as: most of the algorithms support only discrete values as the target attribute. Perform well if a few highly relevant attributes exist, but less if many complex interaction is present. Classification tree analysis when the predicted outcome is the class to which the data belongs and used only for classifying discrete category (the class). It will construct a model using example of cases and able to predict the class of new example [25],[9],[31],[40],[21],[59].

5. Dataset Description

The original dataset was collected in 1986 at Baystate Medical Center, Springfield, Massachusetts. The data consist of maternal information about 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. There are 11 attributes, containing "ID – identification number, Mothers age in years (AGE), the weight of the subject at her last

menstrual period (LWT), the number of physician visits during the first trimester of pregnancy (FTV), race (RACE), lifestyle information such as smoking (SMOKE), a history of previous preterm delivery (PTL), the presence of uterine irritability (UI), and hypertension (HT)". The infant outcome was low birth weight (yes/no) (LOW). Hosmer and Lemeshow used transformations and stepwise variable selection techniques to fit logistic regression models and there were no significant missing values [28],[24].

6. Results and Analysis

Different 'data mining' algorithms have been applied to the prediction of low birth weight child, including logistic regression, naïve Baye's, random forest, support vector machines (SVM), neural network and classification tree. Performance of data mining algorithms based on accuracy, sensitivity, specificity, area under curve (AUC), F1, precision, recall and the data evaluated using leave-one-out, cross validation, random sampling and test on train data. The accuracy measures of different data mining algorithms with different validations techniques are depicted in the Table 2 and it is shown in Figure 1. The accuracy of a model on a given test set is the percentage of test set that are correctly classified by the classifier. Measures are defined [34],[23],[10],[27], [17] as follows

$$sensitivity = \frac{t\_pos}{pos} \quad (1)$$

$$specificity = \frac{t\_neg}{neg} \quad (2)$$

$$precision = \frac{t\_pos}{t\_pos+f\_pos} \quad (3)$$

where *t_pos* is the number of true positives (*"normal"* tuples that were correctly classified as such), *pos* is the number of positive (*"normal"*) tuples, *t_neg* is the number of true negatives (*"Abnormal"* tuples that were correctly classified as such), *neg* is the number of negative (*"Abnormal"*) tuples, and *f_ pos* is the number of false positives (*"Abnormal"* tuples that were incorrectly labeled as *"normal"*). Accuracy is defined as follows

$$accuracy = sensitivity\frac{pos}{(pos+neg)} + specificity\frac{neg}{(pos+neg)} \quad (4)$$

Precision

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

$$precision = \frac{True\ positive}{(True\ positive+False\ positive)} \quad (5)$$

Recall

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

$$recall = \frac{True\ negative}{(True\ negative+False\ negative)} \quad (6)$$

F-measure is a weighted harmonic mean of precision and recall

$$F = \frac{2*precison*recall}{(precision+recall)} \quad (7)$$

The area under ROC curve (AUC) technique provides a comprehensive assessment of accuracy of a predictor of screening the range of threshold values for the decision making. The larger area, the more accurate the diagnostic test is. AUC of ROC curve can be measured by the following equation, Where t = (1 – specificity) and ROC (t) is sensitivity[34],[33],[10],[27],[17].

$$AUC = \int_0^1 ROC(t)dt \quad (8)$$

Most of the researcher's aim is to identify the most influenced predictors for diagnosis and prediction of diseases. The most influenced predictor is always increasing the predictive accuracy of the model. Generalized cross validation (GCV) developed by Craven and Wahba. Friedman uses the modified form of the generalized cross-validation criterion is used to identify the most influenced predictor, rank the predictor and eliminate insignificant predictor of the model [10], [30],[29],[44],[57],[52]. The rank of the most influenced predictors in the predication of low birth weight has depicted in the Table 3 and it is shown in Figure 2.

Table 2: Predictive performance of various classification methods

Algorithm	Accuracy	Sensitivity	Specificity	AUC	F	Precision	Recall
Logistic regression	0.7407	0.9231	0.3390	0.7724	0.8304	0.7547	0.9231
Naive Bayes	0.7778	0.9000	0.5085	0.8008	0.8478	0.8014	0.9000
Random Forest	0.7090	0.9923	0.0847	0.8420	0.8243	0.7049	0.9923
SVM	0.7407	0.9846	0.2034	0.7738	0.8393	0.7314	0.9846
Neural Network	0.7619	0.9385	0.3729	0.7804	0.8443	0.7673	0.9385
Classification Tree	0.8995	0.9769	0.7288	0.9380	0.9304	0.8881	0.9769

Table 3: Low birth weight risk factors rank

Attribute		Rank
LWT	Mother's last weight (pounds) before becoming pregnant	100.0000
AGE	Mother's age in years	98.0024
FTV	Number of physician visits during the first trimester	45.8563
PTL	Number of previous premature labors	43.1132
HT	Hypertension-high blood pressure	25.4303
RACE	Race	18.2630
UI	Uterine irritability	15.0174
SMOKE	Smoke	7.7369

For prediction of low birth weight infants, it can be seen that, using validation method test on train data the classification tree gives a higher overall prediction accuracy (89.95%), specificity (72.88%) and AUC (93.80%) F-value (93.04%) and Precision (88.81%) comparing other classification methods. It is identified that the area under the curve (AUC) in classification tree is higher than the other methods, indicating that the classification tree excellently determining the low birth weight infants . Sensitivity plays an important role in the correct diagnosis of the disease. Random forest gives a lower overall prediction accuracy 70.9% and higher sensitivity 99.23% than other algorithms. The results of this work have improved prediction accuracy when compared to logistic regression. Mother's last weight (pounds) before becoming pregnant highly influenced to predict the low birth weight, i.e., 100%. The first four variables which were highly influenced are Mother's last weight (pounds) before becoming pregnant, Mothers age in years (98.00%), Number of physician visits during the first trimester (45.86%), Number of previous premature labors (43.11%). Hypertension-high blood pressure, Race, Uterine irritability and Smoke are weakly influenced predictors.

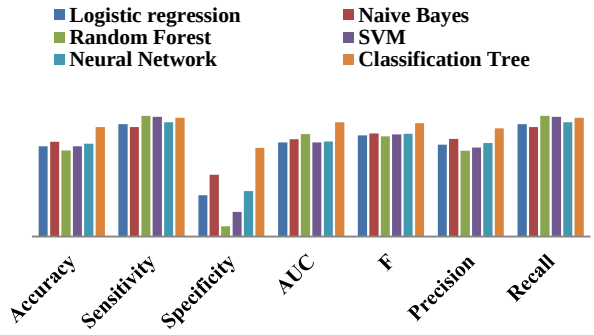

Fig 1. Accuracy measures of classification methods

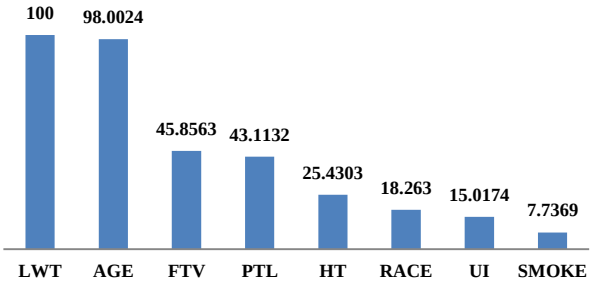

Fig 2. Low birth weight risk factors rank

7. Conclusion and Future works

Low birth weight and preterm birth are identified as a leading cause of infant mortality and are at heightened risk for individuals, families, society. Deeper understanding of the important factors highly associated with low birth weight, different data mining algorithms has been applied to the prediction of low birth weight child, including logistic regression, naïve Baye's, random forest, support vector machines (SVM), neural network and classification tree. The variables which were highly influenced in the predication of low birth weight are Mother's last weight (pounds) before becoming pregnant, Mothers age, Number of physician visits during the first trimester, Number of previous premature labors. Classification tree performed best compared with other algorithms. Hosmer and Lemeshow used transformations and stepwise variable selection techniques to fit logistic regression models. The results of this work have improved prediction accuracy in Datamining techniques when compared to logistic regression. Future work is to improve the accuracy of prediction by using soft computing techniques and to create a powerful tool to assist physicians in their decision making for the prediction of low birth weight.

References

[1].Antoinette Parisi Eaton,Peter C. van Dyck & Kerry P. Nesseier, R.N.,"Low Birth Weight Report and Recommendations, Advisory Committee on Infant Mortality," Final Report to Secretary of the U.S. Department of Health and Human Services, December, 2001,pp.1-32.

[2].Aparajita Dasgupta, Rivu Basu, "Determinants of low birth weight in a Blowk of Hooghly, West Bengal: A multivariate analysis," International Journal of Biological & Medical Research, 2(4),2011, pp.838-842.

[3].Applying Data Mining Techniques in Property/Casualty Insurance, Lijia Guo, Ph.D., ASA, University of Central Florida, pp. 1-25.

[4].Bellazzi R, Zupan B, "Predictive data mining in clinical medicine: current issues and guidelines", Int J Med Inform., 2008;77(2), pp.81-97.

[5].Bellazzi R, Zupan B., "Towards knowledge-based gene expression data mining", J Biomed Inform. 2007;40(6), pp.787-802.

[6].Boris Milovic, Milan Milovic, Prediction and Decision Making in Health Care using Data Mining, International Journal of Public Health Science (IJPHS) Vol. 1, No. 2, December 2012, pp. 69-78

[7].C. Gireesh, V. Punna Rao, " Blood Glucose Prediction Algorithms for Hypoglycemic and/or Hyperglycemic Alerts,", IJCSI - International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, pp.164-168.

[8].Ciaran S. Phibbs, Ph.D., Laurence C. Baker, Ph.D., Aaron B. Caughey, M.D., Ph.D., "Level and Volume of Neonatal Intensive Care and Mortality in Very-Low-Birth-Weight Infants," The new england journal of medicine, 356;21, May 24, 2007, pp.2165-2175.
[9].Classification and Regression Trees,36-350, November 2009, Principles of Data Mining, sections 10.5 and 5.2: Berk, chapter 3.
[10].D. Senthilkumar,, S. Paulraj, "Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines," International Journal of Engineering and Technology (IJET), Volume 5 No 5 Oct-Nov 2013, pp: 3922-3929.
[11].Divya Tomar and Sonali Agarwal, "A survey on Data Mining approaches for Healthcare," International Journal of Bio-Science and Bio-Technology, Vol.5, No.5 (2013), pp. 241-266.
[12].Dr.R.Kavitha, "Health Care Industry in India," International Journal of Scientific and Research Publications, Volume 2, Issue 8, August 2012, pp.1-4.
[13].Dr.S.Santhosh Baboo, Mrs.S.Sasikala, "Multicategory Classification Using Support Vector Machine for Microarray Gene expression Cancer Diagnosis," Global Journal of Computer Science and Technology, Vol.10 Issue 15 (Ver.1.0), December 2010, pp:38-44.
[14].Dursun Delen & Nainish Patil, "Knowledge Extraction from Prostate Cancer Data", Proceedings of the 39th Hawaii International Conference on System Sciences - 2006,pp:1-10.
[15].Elizabeth Vargis, C.Nathan Webb, B.C.Paria, Kelly A.Bennett, Jeff Reese, Ayman Al-Hendy, Anita Mahadevan-Jansen, "Detecting Changes During Pregnancy with Raman Spectroscopy" Biomedical Sciences and Engineering Conference (BSEC), 2011; 04/2011, pp.1-4.
[16].Fayyad U, Piatetsky-Shapiro G, Smyth P., "Data mining and knowledge discovery in databases," Communication ACM, 1996; Volume 39 Issue 11, Nov. 1996, pp.24-26.
[17].Frank Keller, " Evaluation : Connectionist and Statistical Language Processing". Universitat des Saarlandes, pp.1-21.
[18].Guillermo Marshall, et al., "A New Score for Predicting Neonatal Very Low Birth Weight Mortality Risk in the NEOCOSUR South American Network," Journal of Perinatology, 25, 2005, pp.577-582.
[19]. Hashima E Nasreen, Zarina Nahar Kabir, Yvonne Forsell, Maigun Edhborg, "Low birth weight in offspring of women with depressive and anxiety symptoms during pregnancy: results from a population based study in Bangladesh, BMC Public Health, 10:515,2010,pp.1-8.
[20]."Healthy Mothers-Healthy Babies: How to Prevent Low Birth Weight Consensus Statement," Institute of Health Economics Consensus Statements, Volume 2, May 23-25 2007, pp.1-24.
[21].Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva, "An Efficient Classification Approach for Data Mining," International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012, pp.446-448.
[22].Hetal Bhavsar, Amit Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-4, September 2012, pp.74-81.
[23].Hian Chye Koh and Gerald Tan ,"Data Mining Applications in Healthcare," Journal of Healthcare Information Management — Vol. 19, No. 2,pp:65-72
[24].Hosmer, D. W., Jr., S. A. Lemeshow, *and* R. X. Sturdivant, "Applied Logistic Regression," 3rd ed. Hoboken, NJ: Wiley, 2013. .
[25]..http://en.wikipedia.org/wiki/Decision_tree_learning.
[26].http://en.wikipedia.org/wiki/Random_forest. [
[27].http://orange.biolab.si/docs/latest/widgets/rst/evaluate/testlearners
[28].http://www.umass.edu/statdata/statdata/stat-rmult.html
[29].J. H. Friedman and C. B. Roosen, "An introduction to multivariate adaptive regression splines," Statist. Methods Med. Res., vol. 4, 1995, pp. 197–217.
[30].J. H. Friedman, "Multivariate adaptive regression splines (with discussion)," Ann. Statist., vol. 19, 1991, pp. 1–141.
[31].Janez Demšar, "Material for lectures on Data mining,", Kyoto University, Dept. of Health Informatics, July 2010, pp.1-53.
[32].Jehad Ali, Rehanullah Khan, Nasir Ahmad, "Imran Maqsood Random Forests and Decision Trees," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, pp:272-278.
[33].Jenn O'Connor, "Healthy Babies: Efforts to Improve Birth Outcomes and Reduce High Risk Births," NGA Center for Best Practices, Washington, June 2004, pp.1-17.
[34].Jiawei Han and Micheline Kamber, "Data Mining:Concepts and Techniques Second Edition," University of Illinois at Urbana-Champaign, Morgan Kaufmann Publishers, Elsevier.
[35].Julie Bull, Caroline Mulvihill and Robert Quigley , "Prevention of low birth weight: assessing the effectiveness of smoking cessation and nutritional interventions Evidence briefing," 1st Edition – July 2003, , pp.1-56.
[36].Kanya Mukhopadhyay, Deepak Louis, Rama Mahajan, and Praveen Kumar, "Predictors of Mortality and Major Morbidities in Extremely Low Birth Weight Neonates," Indian Pediatrics , Volume 50, December 15, 2013,pp.1119-1123.
[37].Karen P. Tang, Sen H. Hirano, Karen G. Cheng, Gillian R. Hayes, "Balancing Caregiver and Clinician Needs in a Mobile Health Informatics Tool for Preterm Infants," 6th International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health) and Workshops, 2012, pp.1-8.
[38].Kleanthis C. Neokleous, Christos N. Schizas, Costas K. Neokleous, Constantinos S. Pattichis, Panagiotis Anastasopoulos, Kypros H. Nikolaides, "Neural networks to estimate the influence of cervix length on the prediction of spontaneous preterm delivery before 37 weeks," Proceedings of the 5th International Conference on Information Technology and Application in Biomedicine, May 30-31, 2008, pp.423-425.
[39].LI Yun, LI Xiang-sheng,"The Data Mining and Knowledge Discovery in Biomedicine," The 5th

International Conference on Computer Science & Education Hefei, China. August 24–27, 2010,pp:1050-1052.
[40].Lior Rokach &Oded Maimon, "Data Mining and Knowledge Discovery Handbook," (Decision Tree-Chapter 9), pp.165-198.
[41].M. M. Van Dyne, L. K. Woolery, J. mala-Busse, C. Tsatsoulis, "Using Machine Learning and Expert Systems to Predict Preterm Delivery in Pregnant Women," Proceedings of the Tenth Conference on Artificial Intelligence for Application, IEEE, March 1994, pp.344-350.
[42].Mala Rao and David Mant, "Strengthening primary healthcare in India: white paper on opportunities for partnership," Explore how India and the UK can work together on education, professional development, affordable technologies, public-private partnerships, governance, and innovation in primary care in India, BMJ 2012;344:e3151, pp:1-14.
[43].Marjan Khajehei ,Faried Etemady , "Data Mining and Medical Research Studies," Second International Conference on Computational Intelligence, Modelling and Simulation,2010, pp:119-122.
[44].Martian Chronicles: Is MARS better than Neural Networks?- Louise Francis,FCAS, MAAA],pp.269-306.
[45].Martina Mueller, Carol L. Wagner, David J. Annibale, Thomas C. Hulsey, Rebecca G. Knapp, and Jonas S. Almeida , "Predicting Extubation Outcome in Preterm Newborns: A Comparison of Neural Networks with Clinical Expertise and Statistical Modeling," Pediatric Research, Volume. 56, No. 1, 2004, International Pediatric Research Foundation, pp. 11-18.
[46].Masaki Ogawa, Yoshio Matsuda, Eriko Kanda, Jun Konno, Minoru Mitani, Yasuo Makino and Hideo Matsui, "Survival rate of extremely low birth weight infants and its risk factors: Case-Control study in Japan," ISRN Obstetrics and Gynecology, Hindawi Publishing Corporation, Volume 2013,pp.1-6.
[47].Moawia Elfaki Yahia, Murtada El-mukashfi El-taher, "A New Approach for Evaluation of Data Mining Techniques," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010, pp.181-186.
[48].Monica Das Gupta & Manju Rani, "India's Public Health System ::How Well Does It Function at the National Level?," World Bank Policy Research Working Paper 3447, November 2004, pp.1-24.
[49].Mozziyar Etemadi, Philip Chung, J. Alex Heller, Jonathan A. Liu, Larry Rand, Shuvo Roy, "Towards Birth Alert – A Clinical Device Intended for Early Preterm Birth Detection", I*EEE Transactions on Biomedical Engineering*, Volume 60, Issue: 12, Dec. 2013, pp.3484 – 3493.
[50].Mr. Hitesh H. Parmar, Prof. Glory H. Shah, "Experimental and Comparative Analysis of Machine Learning Classifiers," International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October - 2013, pp. 955-963.
[51].N.Sathiya Rani, K.Vimala, Dr.V.Kalaivani, "Health Care Monitoring for the CVD Detection using Soft Computing Techniques", International Journal in Foundations of Computer Science & Technology (IJFCST), Volume 3, No.4, July 2013, pp.21-30.
[52].Namasivayam Ambalavanan, Waldmar A. Carlo, Georgiy Bobashev, Erin Mathias, Bing Liu, Kenneth Poole, Avroy A. Fanaroff, Barbara J. Stoll, Richard Ehrenkranz and Linda L. Wright, " Prediction of Death for Extremely Low Birth Weight Neonates," Pediatrics – Journal of the American Academy of Pediatrics, ,Volume 116 No.6, December 2005, pp.1367-1373.
[53].Nataša Šarlija, Kristina Šorić, Silvija Vlah, Višnja Vojvodić Rosenzweig, "Logistic Regression and Multicriteria Decision Making in Credit Scoring,".
[54].Ned Horning, " Random Forests : An algorithm for image classification and generation of continuous fields data sets," International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences 2010.
[55].Nonlinear Regression: Modern Approaches and Applications – Salford System, 2013.
[56].Nynke R. van den Broek, Rachel Jean-Bapsite, James P.Nelison, "Factors Associated with preterm, early preterm and late preterm birth in Malawi," PLOS ONE, Volume 9, Issue 3, March 2014, pp.1-8.
[57].P. Craven and G. Wahba, "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation," Numer. Math., vol. 31, 1979, pp. 377–403.
[58].P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils," IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009 pp.117- 122.
[59].Patrick Ozer, Dr. I.G. Sprinkhuizen-Kuyper. "Data Mining Algorithms for Classification," Thesis Artificial Intelligence, Radboud University Nijmegen, January 2008, pp.1-27.
[60].Paul Komarek, Andrew Moore, Alain Calvet Alan Frieze, Chair Bob Nicho, "Logistic Regression for Data Mining and High-Dimensional Classification," 2004, pp:1-139, http://repository.cmu.edu/robotics
[61].Portia A. Cerny, "Data mining and Neural Networks from a Commercial Perspective,"pp:1-10
[62].Rabindra Nath Das, Rajkumari Sanatombi Devi, Jinseog Kim, "Mothers' Lifestyle Characteristics Impact on Her Neonates' Low Birth Weight," International Journal of Women's Health and Reproduction Sciences ,Volume 2, No. 4, Summer 2014,pp.229–235.
[63].Raghavendra B.K, S.K. Srivatsa, "Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets," International Journal of Computer Science and Security (IJCSS), Volume (5) : Issue (5) : 2011, pp. 503 -511.
[64].Rahul Isola, Rebeck Carvalho,Mangala Iyer Amiya Kumar Tripathy, "Automated Differential Diagnosis in Medical Systems using Neural Networks, kNN and SOM," 2011 Developments in E-systems Engineering,pp:62-67.
[65].S.Neelamegam, Dr.E.Ramaraj, "Classification algorithm in Data mining: An Overview," International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013, pp.369-374.

[66].Samuel Odei Danso , “An Explloration of Cllassiificatiion Predictiion Techniques iin Data Mining: The Insurance Domain,” September 2006, pp: 1-79.
[67].Sandhya Joshi, L M Patnaik,P L Rrashmi , K R Venugopal,"Classification of Alzheimer’s Disease and Parkinson’s Disease by Using Machine Learning and Neural Network Methods," 2010 Second International Conference on Machine Learning and Computing, pp: 218-222.
[68].Sarojini Balakrishnan,Nickolas Savarimuthu,Ramaraj Narayanaswamy & Rita Samikannu, "SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases", 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008), pp:2628-2633.
[69].Sellappan Palaniappan and Rafiah Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques,” Awang IJCSNS International Journal of Computer Science and Network Security, Volume.8 No.8, August 2008, pp.343-350.
[70].Takashi Yorifuji, Hiroo Naruse, Saori Kashima, Soshi Takao, Takeshi Murakoshi, Hiroyuki Doi, “ Residential proximity to major roads and adverse birth outcomes: a hospital-based study,” Environmental Health, 12:34, 2013, pp.1-11.
[71].The Disadvantages of Logistic Regression, Written by Damon Veria
[72].Trees,Kin-Yee CHAN and Wei-Yin LOH, “LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression,” Journal of Computational and Graphical Statistics, Volume 13, Number 4,2004, pp. 826–852.
[73].Wen Zh, Nancy Zeng, Ning Wang, “Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations” - NESUG 2012–Health Care and Life Sciences,pp.1-9.
[74].XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang ·Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, “Top 10 algorithms in data mining,” Knowl Inf Syst (2008) 14, pp1–37.
[75].Yavar Naddaf, Mojdeh Jalali Heravi and Amit Satsangi., "Predicting Preterm Birth Based on Maternal and Fetal Data".