Data Mining Techniques to Analyze the Reason for Home Birth in Bangladesh

Fahim Jawad, Tawsif Ur Rahman Choudhury, Ahmad Najeeb, Mohammed Faisal,
Fariha Nusrat, Rubaiya Chamon Shamita, Rashedur M Rahman
Department of Electrical and Computer Engineering, North South University
Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh
fahim.jawad.016@gmail.com, tawsifrchoudhury@gmail.com, najeeb_03@hotmail.com, mohammed.faislam@gmail.com,
farihamimfariha@gmail.com, shamita35@gmail.com, rashedur.rahman@northsouth.edu

***Abstract*— Data Mining is the process of finding pattern or useful information from large volume of data. The goal of this paper is to find the reason behind the unusual high birth rate by applying data mining techniques, e.g., decision tree, neural network, Bayes Classifier, Ripper and Support Vector Machine. The datasets were collected from the baseline survey conducted by the maternal neonatal and child health programme by ICDDR,B. If we could find the reason(s), high birth rate at home could be avoided in future. Giving birth at home is very dangerous as many complications may arise during pregnancy as well during birth. From the opinions of experts and professionals, it could be said that the risk of mortality of new born during home birth is quite alarming and birth at hospital/clinics seemed to be the safest place to protect the health and well-being of the woman and her baby.**

Keywords—Home birth; Place of delivery, Data mining, Decision Tree, Neural Network, Probabilistic Neural Network.

I. INTRODUCTION

ICDDR,B [1] is the international center for diarrheal disease research Bangladesh. It is an international health research organization located in Bangladesh. It is dedicated to save lives through translation of research into treatment, training and policy advocacy. It also addresses some of the most critical health concerns facing the world today.

The baseline survey we used for the research was conducted on mothers in the Matlab, Bangladesh region, which is the primary rural field site for ICDDR,B and the world's longest running health project. The survey is an analysis of current situation to identify the starting points for a program or project.

The survey was carried out from 1st. November 2005 to 30th. September 2006 under ICDDR,B. Maternal Neonatal and Child Health programme, commonly known as MNCH. This programme's main purpose is to improve the quality and availability of all maternal, newborn and child health services. Ensuring optimal health for girls of reproductive age, improving the health and nutrition of mothers-to-be, and providing quality reproductive health services including ante- and post-natal care are pivotal to ensuring safe motherhood.

Maternal, newborn and child health is an area in which aid can make a huge difference. By training midwives to prevent deaths during childbirth, increasing access to life-saving vaccines to help providing better nutrition to reduce child deaths and stunting, MNCH is one of the highest impact areas of effective development aid.

Many of the non-government and government organizations are also carrying out MNCH programs both locally and overseas, e.g., Australia, Canada, Pakistan and etc.

This paper is organized as follows: the literature review related to existing research by ICDDR,B is described in Section II. Section III describes the methods used in this research in detail. The results and analysis of results is also discussed here. Finally conclusion and future work is presented in Section IV.

II. LITRETATURE REVIEW

Decision Tree models generated using decision tree induction algorithms are very comprehensive to the end-users. Hospital Surveillance data has been historically used for early detection of emerging epidemics for example influenza or Cholera etc. using time series analysis. Historical data of surveillance system is used to detect recognizable wave patterns or cycles in the earlier epidemic attacks to forecast future probable attacks.

In [11] the authors have used Decision Tree induction algorithm to generate decision tree models from hospital surveillance data to classify hospital patients according to their physical conditions and personal disease history on admission to hospital. Decision tree models are generated using ICDDR,B hospital surveillance data. From the decision tree generated based on earlier cases stored in the surveillance data decision rules are generated. These rules are used to classify patients into three classes according to their criticality: High, Mid, Low so that hospital can take prudent actions for the patients. Different preprocessing and formatting activities have been carried out on the data to make it ready for the model building. Different decision tree models have been generated to find out an optimized model which can classify new patients more accurately i.e., the prediction accuracy is higher. Different optimization techniques have been employed. Lastly performance of different decision tree models have been measured and compared using different performance metrics.

978-1-4799-8676-7/15/$31.00 copyright 2015 IEEE
SNPD 2015, June 1-3 2015, Takamatsu, Japan

The heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Mesthe et al.[15] used data mining algorithms, e.g., J48, Bayes Net, and Naive Bayes, Simple Cart, and REPTREE algorithm to classify and develop a model to diagnose heart attacks in the patient data set from medical practitioners. The objective of the research is to predict possible heart attacks from the patient dataset using data mining techniques and determine the model that gives the highest percentage of correct predictions for the diagnoses.

The main objective of the project [16] was to develop a prototype intelligent Heart Disease Prediction System (IHDPS) using three data mining modelling techniques, namely, Decision Tree, Naïve Bayes and Neural Network. IHDPS can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support system cannot. By providing effective treatments, it also helps to reduce treatment costs.

While the recent advances in neonatal medicine has greatly increased the chance of survival of infants born after 20 weeks of gestation, these infants still frequently suffer from lifelong handicaps, and their care can exceed a million dollars during the first year of life as cited in [17]. As a first step for preventing preterm birth, decision support tools are needed to help doctors predict preterm birth. A number of popular classification algorithms are applied to the dataset for predicting preterm birth, and their prediction performance is compared with the associative classifier. The applied classifiers are Logistic Regression, Naive Bayes, C4.5 decision trees, Support Vector Machines, and Neural Networks. They implemented the algorithms with default parameters provided in Weka, an open source data-mining package [13]. For Naive Bayes and C4.5 Decision Trees, the numerical attributes are recoded into a set of categorical attributes. Three-fold cross validation is used to evaluate the performance of each classifier.

The survey [18] has reviewed standard algorithms that are well known in research community and has discussed the criterion for these algorithms, which are classification, regression, segmentation, association, and sequence analysis. These data mining classifications are subsets of standard algorithms and are used by data miner software vendors for their data analysis services. Depressive symptoms are common amongst pregnant women following anemia and it could predict subsequent maternal mortality and morbidity and fetal abnormalities.

III. METHOD OF ANALYSIS

A. *Choosing the Class Attribute:*

The data we acquired was in .sav or in SPSS format. We used the software SPSS as well as KNIME to handle our dataset. In SPSS we went through most of the attributes to get an appropriate class for our analysis and through rigorous searching we saw an outlier, 58.5% of the baby born were in home and it seemed like a perfect classification problem for our research. It is also represented in Table I.

TABLE I. BIRTH AT DIFFERENT PLACES

	Frequency	Percent	Cumulative %
01	2752	58.5	58.5
02	66	1.4	59.9
03	410	8.7	68.6
04	780	16.6	85.2
05	58	1.2	86.5
06	36	0.8	87.2
07	12	0.3	87.5
08	342	7.3	94.7
09	54	1.1	95.9
10	7	0.1	96.0
11	106	2.3	98.3
12	44	0.9	99.3
98	35	0.7	100
Total	4703	100	

In Table I, 01 represent Home Birth which has a frequency of 2752 whereas in other places like 03(MATLAB SUB-CENTER) and 04(MATLAB HOSPITAL) seem to be very low.

B. *Preprocessing:*

In the dataset we collected, there are about 5500 participants. The survey questionnaire was a very detailed set; it contains a total of 549 attributes. The various fields that are asked from the participants are: general information, previous birth information, pregnancy, delivery, and postpartum care for mother, immediate newborn care, and newborn care, perception of local facility, attitude and perception, etc.

As our main objective was to focus on exploring the unusually high rate of HOME BIRTH and the several factors leading to the decision, we finally concentrated on the following data sets: checkups during pregnancy, who did checkups, plans for birth during pregnancy, complications and seek treatment, type of treatment received, planned for transport, saved money, parent's education background, mother's perception (knowledge), perception of local facility, and attitude, etc.

We trimmed data's in two ways: *Vertically* and *Horizontally*

Vertically we considered the factors after various omissions from the original data set. Firstly, we omitted the tuples which had missing data. Doing that our total number of tuples has been declined in size from 5248 to 4703.

Horizontally we focused on the sections which were irrelevant to birth of the child such as:

i) Postpartum care for mother
ii) Immediate newborn care
iii) Newborn care

All these sections deal with care of the mother or child after the delivery of the child.

Postpartum care for mother:
Postpartum care for mother asks the individual about, complications she faced after child birth, the person who checked her health checkup after delivery. We omitted this section which comprised of questions Q401 to Q406. With this our total attributes are decreased from 549 to 531.

Immediate newborn care:
Immediate newborn care is about, health problems of the baby after birth, basic cares about the baby, etc. We trimmed Q501 to Q527 which were all in this section. With this our total attributes are further decreased from 531 to 446.

Newborn care:
New born care and first month is about: was the baby with mother or relative, was there any checkup performed on baby, did the baby have any problems. Q601 to Q631 were trimmed from this section. With this our attributes are further decreased from 446 to 311.

The general information of the participants was also removed from the attributes list for example: name, id, date, etc. With this we were left with attributes156 from 311.

156 attributes seem to be huge but according to our dataset, the question's answers were provided with their own attribute field, for example Q302 question has 11 choices and in the dataset it is given as Q302A, Q302B and so on. The answers are filled according to their choices on that attribute.
Therefore there are many questions in the dataset that take up more than 10 attribute spaces.
Considering those 10 attributes as 1 (with different values) we actually have an attribute count of 49.

C. Classification Methods

We used SPSS to filter out the attributes that were unnecessary to our objective. SPSS did not have direct access to any data mining algorithm thus we decided to use software for our data mining purpose, KNIME.

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.

For every analysis we used the color RED as Home Birth and the color GREEN as Other.

At first we partitioned the dataset according to 60-40 percentile, 60% used as the training set and the rest 40% as test set. According to Figure 1.0 the 60% training set goes to each of the classifiers learner node and after learning the 40% is used for the Predictor node. As shown in Figure 1, we created our data mining nodes. For our analysis we used 5 classifiers,

- Decision Tree
- Naïve Bayes
- Probabilistic Neural Network
- Multilayer Perceptron
- Support Vector Machine

1. Decision Tree:

KNIME **decision tree learner node** uses C4.5 [2] which is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 [5] can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Pre-pruning a decision tree involves using a 'termination condition' to decide when it is desirable to terminate some of the branches prematurely as the tree is generated. Post-pruning[3,4] a decision tree implies that we begin by generating the tree to its full length completely and then adjust it with the aim of improving the classification accuracy on unseen instances.

For our dataset, the decision tree was given few specific attributes such as, Quality Measure with respect to Gini Index and Pruning method as Minimum Description Length (MDL) and we got a Confusion matrix as shown in Table II.

We found out that the factors or the attributes that played the most important roles in determining the Home Birth Class are:

- Birth delivery Kit available
- Receiving Checkup during last Pregnancy
- How many times Checkup during pregnancy
- Make plans for pregnancy
- Where they planned to give birth
- Who they selected as their birth attendant

For our decision tree we have correctly classified 88.2% of the data and the remaining 11.8% are incorrectly classified.

2. Naïve Bayes:

Naive Bayes is a simple technique for constructing classifiers. Naïve Bayes model assigns class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Figure 2 shows a part of Naïve Bayes learner view in KNIME and also the Confusion Matrix is shown in Table II. More details of this technique could be found elsewhere [6,8].

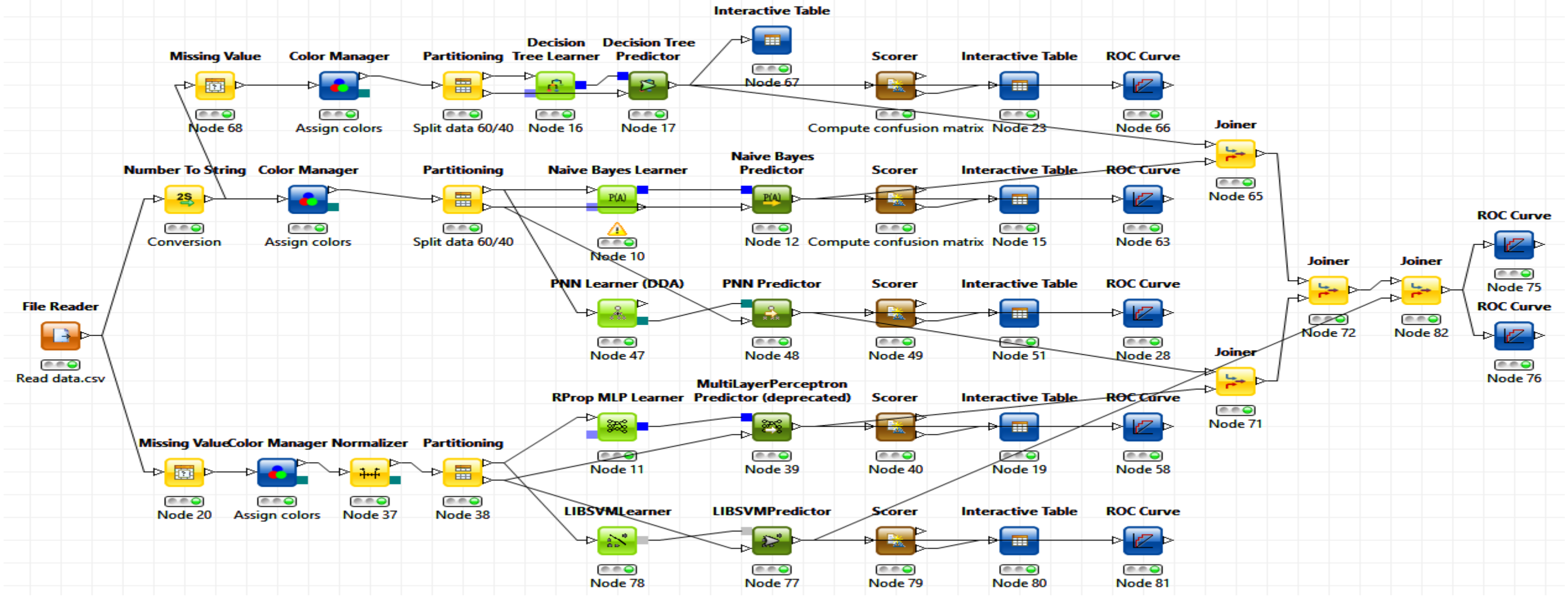
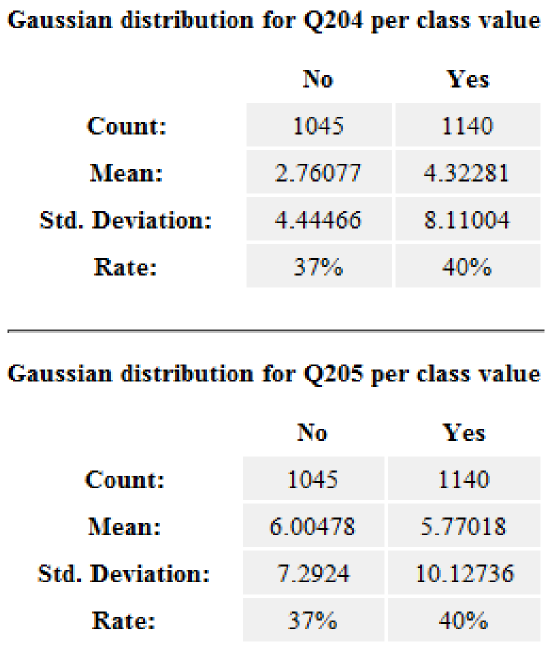


Figure 1. KNIME Data Mining Node Structure

TABLE II. ACCURACY OF DIFFERENT CLASSIFIERS

Classifier	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's Kappa
Decision Tree	0.898	0.798	0.898	0.873	0.845	0.882	0.75
Naïve Bayes	0.873	0.983	0.873	0.987	0.924	0.935	0.868
Probabilistic Neural Network	0.735	0.731	0.735	0.599	0.733	0.68	0.335
Multilayer Perceptron	0.919	0.9	0.919	0.859	0.91	0.894	0.782
Support Vector Machine	0.884	0.906	0.884	0.873	0.895	0.879	0.754



Gaussian distribution for Q204 per class value

	No	Yes
Count:	1045	1140
Mean:	2.76077	4.32281
Std. Deviation:	4.44466	8.11004
Rate:	37%	40%

Gaussian distribution for Q205 per class value

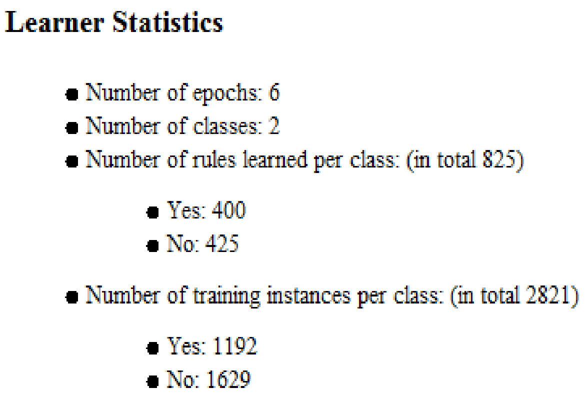
	No	Yes
Count:	1045	1140
Mean:	6.00478	5.77018
Std. Deviation:	7.2924	10.12736
Rate:	37%	40%

Figure 2. Representation of Naïve Bayes in KNIME

3. Probabilistic Neural Network:

Probabilistic Neural Network (PNN) is trained based on the DDA (Dynamic Decay Adjustment) method on data using Constructive Training of Probabilistic Neural Networks as the underlying algorithm. More about this algorithm could be found [7, 13, and 14].

The PNN learner statistics from KNIME is as shown below,



Learner Statistics

- Number of epochs: 6
- Number of classes: 2
- Number of rules learned per class: (in total 825)
  - Yes: 400
  - No: 425
- Number of training instances per class: (in total 2821)
  - Yes: 1192
  - No: 1629

4. Multilayer Perceptron Neural Network:

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer is fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. For our MPNN we used 50 learning iterations, 1 hidden layer and 10 hidden neurons per layer using our homebirth attribute as class column.

The Confusion Matrix of MLP is as shown in Table II.

5. Support Vector Machine:

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

We used RBF (radial basis function) with Sigma = 0.5 as our kernel in SVM with an overlapping penalty of 1, which basically determines the penalty assigned to each point that is misclassified.

The Confusion Matrix of SVM is as shown in Table II.

D. Measuring the performance of a Classifier

We measure the performance of the classifiers with respect to different performance metrics. The performance is represented in Table II. The confusion matrices that we acquired also had the performance of the classifier which was made by the Scorer node and represented by the Interactive table as shown in the Figure 1.
According to that we formed the ROC curve for each classifier and joined them to determine the best classifier.

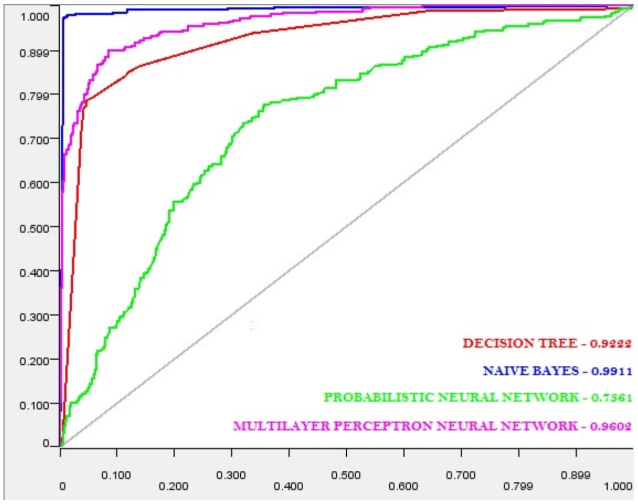


Figure 3. ROC Curve for Classifiers

Figure 3 represents the ROC curves for the classifiers. From here we can determine that Naïve Bayes classified the dataset most accurately with an accuracy of 93.5%.

IV. Conclusion

This paper has analyzed the Matlab Maternal Neonatal and Child Health Programme's baseline survey. After analyzing we saw that there was an anomaly of high home birth. According to MONDAY, Feb. 3, 2014 (HealthDay News)[10] -- The number of pregnant women who elect to deliver their baby at home is increasing, but home delivery can lead to problems.

The risk of a baby dying is nearly four times higher when delivered by a midwife at home than by a midwife in a hospital, according to a new study.

Even though in our dataset the death rate of new born was insignificant it is still a concern as a child may lose their life due to carelessness of the parents or not knowing the actual process of home birth.

Normally, the reasons for home birth are as follows [10]:

-A desire to give birth in a familiar, relaxing environment surrounded by people of your choice
-A desire to wear your own clothes, take a shower or bath, eat, drink and move around freely during labor
-A desire to control your labor position or other aspects of the birthing process
-A desire to give birth without medical intervention, such as pain medication
-Cultural or religious norms or concerns
-A history of fast labor

From our findings we can also observe that the desire to control the time of delivery is the most important part for the mother and as such from our analysis we saw that most mothers had a birth delivery kit as well as a birth attendant of their choice.

Labor position was also an important factor for home birth.

But there are situations where a home birth is not recommended,
-Have diabetes, chronic hypertension, a seizure disorder or any chronic medical condition
-Previously had a C-section
-Develop a pregnancy complication, such as preeclampsia
-Are pregnant with multiples or your baby doesn't settle into a position that allows for a headfirst delivery
-Less than 37 weeks or more than 41 weeks pregnant.

As the region where the survey was conducted has a literacy rate less than 70%, everyone there will not be aware of the problems they might face during home birth. Thus, the main goal from this baseline survey is to initiate a campaign about the dangers of home birth among the local people and also raise awareness about the precautions necessary for home birth.

REFERENCES

[1] ICDDR,B, Available at: http://www.icddrb.org/who-we-are/our-mission/about-us (Accessed: 11 January 2015).

[2] Data Mining , Available at: http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-7.html (Accessed: 12 January 2015).

[3] Mansour Y., Pessimistic decision tree pruning based on tree size, Proceedings of the 14th International Conference on Machine Learning (1997): 195-201, Weblink: http://citeseer.ist.psu.edu/76752.html.

[4] Ignizio J.P. An Introduction to Expert Systems: The Development and Implementation of Rule Based Expert Systems. – McGraw – Hill, Inc., 1991, 402 p.

[5] C4.5 Algorithm, Available at: http://en.wikipedia.org/wiki/C4.5_algorithm (Accessed: 12 January 2015).

[6] Geisser, Seymour. Predictive Inference. New York, NY: Chapman and Hall, 1993.

[7] Kohavi, Ron . "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (San Mateo, CA: Morgan Kaufmann), 1995.

[8] Devijver, Pierre A.; Kittler, Josef. Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall, 1982.

[9] "Newbie question: Confused about train, validation and test data!". Retrieved 2013-11-14.

[10] Mayo Clinic Staff Home birth: Know the pros and cons, Available at: http://www.mayoclinic.org/healthy-living/labor-and-delivery/in-depth/home-birth/art-20046878 (Accessed: 12 January 2015), 2014.

[11] Rashedur M Rahman, Fazle Rabbi Hassan, “Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data”. Expert Syst. Appl. 38(9): 11421-11436 (2011) .

[12] A. Aziz, N. Ismail, and F. Ahmad, “Mining Student’s Academic Performance”, Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.

[13] Feldman W.E. and Wood B., The economic impact of high risk pregnancies. Journal of Health Care Finance, 24 64-71, 1997.

[14] Goodwin, L. and Maher, S. Data mining for preterm birth prediction. In Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1 (Como, Italy). J. Carroll, E. Damiani, H. Haddad, and D. Oppenheim, Eds. SAC '00. ACM, New York, NY, 46-51, 2000

[15] Hlaudi Daniel Masethe, Mosima Anna Masethe, “Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[16] Ms. Ishtake S.H , Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013, pp. 94-101.

[17] Yavar Naddaf, Mojdeh Jalali Heravi and Amit Satsangi, “Predicting Preterm Birth Based on Maternal and Fetal Data”, http://yavar.naddaf.name/downloads/Predicting%20Preterm%20Birth%20Based%20on%20Maternal%20and%20Fetal%20Data.pdf.

[18] Fayyaz Ahmed, Adibah Sitara, “Exploration of Co-Relation between Depression and Anaemia in Pregnant Women using Knowledge Discovery and Data Mining Algorithms and Tools”, Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI), September Edition, 2012.