Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1542 – 1549

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

A study on C.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy

Lakshmi.B.N[a*], Dr.Indumathi.T.S[b], Dr.Nandini Ravi[c]

[a]*Ph.D Scholar, PG Research Centre, VIAT, VTU RRC, Muddenahalli, Chikkaballapura-562101, Karantaka, India*
[b]*Professor and Co-ordinator, PG Research Centre, VIAT, VTU RRC, Muddenahalli, Chikkaballapura-562101, Karnataka, India*
[c]*Dr.Nandini Ravi, MBBS, MD(Obs&Gyn), Dhruva Nursing Home, Hosakote, Bangalore, Karnataka, India*

Abstract

Complication during pregnancy has turned out o be a major problem for women of today's era. Pregnant women must be protected from these complications arising in period of gestation, a stage wherein every woman undergoes many physiological changes, sometimes inducing severe health problems leading to death of both mother and fetus. Technological interventions in the field of medical diagnosis can largely help to find a solution for this problem to protecting pregnant women, thus in turn reducing maternal and fetal mortalities to great extents. Decision Tree Classification method is a popularly used method whose algorithms are best suitable in medical diagnosis. C4.5 Decision Tree algorithm is one of the popular and effectively used classifier for pregnancy data classification in present study. The main aim of this paper is to pinpoint the importance of standardization of parameters selected for data collection in study, compare the results obtained from C4.5 classifier on both un-standardized and standardized datasets and analyse the performance of the C4.5 algorithm in terms of its prediction accuracy when applied on the created database from collected and standardized pregnancy data.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

Keywords: pregnancy; learning models; C4.5 classification algorithm; pregnancy complications; abnormalities; standardized dataset; un-standardized dataset.

1. Introduction

Pregnancy complications leading to increased rates of maternal and fetal mortality have opened a new horizon for researchers to work on and provide a feasible solution to protect pregnant women from possible deaths. Pregnancy a delicate state in each woman's life induces a variety changes in health leading to acute complications sometimes resulting in maternal and fetal mortality. The changes induced by physiological parameters are main reasons for an instance of complication to arise.

* *Lakshmi.B.N Tel: +91-8147949107*
Email-id: keerthisri.20@gmail.com

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.128

Changes in physiological parameters if un-noticed by medical practitioners or pregnant women can aggravate to levels of emergency situations. When an emergency situation is created the health of pregnant woman is complicated and difficult to normalize and in such instances either life of mother or child is lost or both lives are lost. According to a study, each day around 800 women die from different causes of pregnancy and childbirth. Approximately 99% of all these maternal deaths occur in developing countries. A study in 2013 states that 2, 89, 000 pregnant women died during and following pregnancy and childbirth. Almost all of these deaths occurred in low-resource settings and most of these could have been prevented [1][2]. Many of these causes can be prevented or controlled if these pregnant women can be warned before the situation is aggravated to complications and thus protect the pregnant women and their babies from any further complications that may lead to death. Some of these physiological parameters like blood pressure, blood glucose level and weight whose changes during pregnancy can lead to complications can be identified and further complications induced from these changes can be prevented[1]. The aim of the paper is to use different types of databases for classification, to analyse and predict the level of risk in each instance. Here in this paper we have analysed a un-standardized database with all selected parameters for the study and also analysed a database with parameters selected after standardization, a procedure that provides precise, valid, reliable and accurate parameters for study. C4.5 Algorithm is one of the most efficient, powerful and popular algorithm of Decision Tree Induction method used for the task of classification of medical data, hence this algorithm is selected among all the other algorithms of Decision Tree Induction method. This algorithm provided better results when compared with other algorithms and on reviewing many works from researchers it was found that C4.5 outperforms many other classification algorithms when applied on medical data. The second section briefs about pregnancy and the complications induced during pregnancy. The third section presents some related studies. The fourth section provides detailed explanation of analysis algorithm C4.5 decision tree algorithm and its working procedure, used to predict health status of women during pregnancy. The last section presents the results obtained from the analysis on both the standardized data and un-standardized data, compares the results obtained from the analysis and concludes pinpointing how standardization of parameters influence the results obtained from C4.5 algorithm for predicting health status from the collected pregnancy data.

2. Literature Survey

Jay Gholap aim at predicting soil fertility class using decision tree algorithms in data mining and focuses on performance tuning of J48(C4.5) decision tree algorithm with the help of meta technologies such as attribute selection and boosting. [3]. Aniket Ray et al present an ensemble classifier-hierarchical committee of random tress that uses risk factors recorded at birth in order to predict the risk of developing Retinopathy of prematurity. [4]. Karpagavalli S et al present the implementation of 3 supervised learning algorithms C4.5 decision tree classifier, Naive Bayes and Multilayer perceptron in WEKA environment on the pre-operative assessment dataset. The prediction accuracy of classifiers were evaluated using 10 fold cross validation and the results were compared. [5].Mr. Brijain R Patel et al provide focus on various decision tree algorithms, their characteristic, challenges, advantages and disadvantages. [6]. T.Miranda Lakshmi et al compare and analyse the performance of ID3, C4.5 and CART decision tree algorithms using qualitative student data. [7]. Reza Entezari-Maleki et al have compared the efficacy of seven classification methods Decision Tree (DT), k-Nearest Neighbor (k-NN),Logistic Regression (LogR), Naïve Bayes (NB), C4.5, Support Vector Machine (SVM) and Linear Classifier (LC) with regard to the AUC metric. [8]. Sharareh R. Niakan et al have applied and compared machine learning techniques initially to predict the outcome of tuberculosis therapy. After feature analysis six algorithms decision tree (DT), artificial neural network (ANN), logistic regression (LR), radial basis function (RBF), Bayesian networks (BN), and support vector machine (SVM) were developed and validated. [9]. Masud Karim have investigated two data mining techniques naïve bayes and C4.5 decision tree algorithms and are implemented to predict whether a client will subscribe a term deposit. [10].

3. C4.5 Decision Tree Classification Algorithms

Classification a technique most suitable in the area of medical diagnosis for automatic discovery of valid, novel, unknown, useful and understandable information from very big databases consisting of data obtained from several applications. Among all classification methods the best suitable classifier for the present study consisting of medical data is C4.5 decision tree classification algorithm. The decision tree algorithms are powerful, popular, logic based, easily interpretable, straight forward and widely applicable for several problems

in data mining. These algorithms are off-the-shelf algorithms providing very good performance and are easily understandable. C4.5 Decision Tree Classification Algorithm was developed by Ross Quinlan as extension from ID3 algorithm also developed by him. These classifiers construct a decision tree as a learning model from the data samples. The divide and conquer approach is followed for construction of decision tree models using a measure called information gain to select the attribute from the dataset for the tree. Consider a possible test is to be selected with n outcomes which divides the data set L with training samples into subsets $\{L_1, L_2, L_3.....L_n\}$. Distribution of classes in L and its subsets L_i is the only information available for tree construction. Considering P to be any set of samples, then freq (C_i,P) are the total number of samples in P belonging to C_i and |P| is denoted by the number of samples in P. The entropy for the set P is given by

$$\text{Info (P)} = -\sum_{i=0}^{k} \frac{\text{freq Ci,P)}}{|P|} \log_2 \left(\frac{freq(\text{Ci,P})}{|P|} \right) \quad (1)$$

Information content of L can be measured by computing Info (L), the total information content of L can be computed once L is divided with respect to the outcomes of a selected attribute say z. The weighted sum of the entropies of each subets gives the total information content of L.

$$Info_z \text{ (L)} = \sum_{i=0}^{n} \frac{|\text{Li}|}{|L|} Info \text{ (Li)} \quad (2)$$

The gain is given by

$$\text{Gain (z)} = \text{Info (L)} - Info_z \text{ (L)} \quad (3)$$

Gives information gained by dividing L with respect to the test on z. This is done for the selection of attribute z with highest information gain. Three base cases for C4.5 algorithm are considered (1) If all samples in dataset belong to same class a leaf node is created for decision tree choosing that class. (2) If no information gain is provided by any feature/attribute, a decision node is created high up the tree with expected value of the class. (3) If a class of an unseen instance encountered a decision node is created high up the tree with expected value of the class.

4. Methodology

The section of methodology provides details of the techniques and tools used for data collection, selection of respondents and methods used for analysing the collected data instances. The section briefs about the scientific procedure followed for this present work to draw rational, logical and meaningful conclusions.

4.1. Parameter Selection

A set of attributes for the study are identified from the literature survey and suggestions of experts. Extensive literature survey is done to identify the problems faced by pregnant women in terms of changes induced by pregnant women and a set of parameters are selected. These parameters were subjected to further analysis by four doctoral committee members to suggest a set of valid parameters. Based on the suggestions from the committee members a total of twelve parameters are considered as shown in table 1.

Table 1: Pregnant Women attribute and descriptions

Sl. No	Attribute	Description	Data type
1	Age	Age of the pregnant women in years	Numeric
2	Pregnancy Parity	Present pregnancy number	Number
3	History of Pre-eclampsia	Did the pregnant woman experience pre-eclampsia state in previous pregnancies (Yes/No)	Textual
4	History of Gestational Diabetes	Did the pregnant woman experience gestational diabetes state in previous pregnancies (Yes/No)	Textual
5	Mother/Sister with pre-eclampsia	Did the pregnant woman's mother/sister experience pre-eclampsia state any of their pregnancies (Yes/No)	Textual
6	Mother/Sister with	Did the pregnant woman's mother/sister experience gestational	Textual

	Gestational Diabetes	diabetes state any of their pregnancies (Yes/No)	
7	State-	The present health status of the pregnant woman (Hypertensive/Normal/Overweight/Underweight)	Textual
8	Trimester	Present trimester of pregnancy (2/3)	Numeric
9	Present Month	Present month of pregnancy (4/5/6/7/8/9)	Numeric
10	Blood Pressure	Blood Pressure recorded presently in mm/Hg	Numeric
11	Presence of Gestational Diabetes	Blood Sugar Levels recorded presently in mg/dl	Numeric
12	Weight Gain	Present weight gain from the previous month in Kgs	Numeric

These attributes are the input variables identified for the study and the outcome is the risk level in pregnancy represented by tone of the status as Mild Risk, High Risk or No Risk. All the parameters are interlinked, abnormality in one parameter will lead to many severe complications during pregnancy. Hence efficiency of each parameter is important for health of pregnant women.

4.2. Data Collection:

Data collection is carried out in Bangalore district, Karnataka state for present study and three medical centres from Bangalore district are identified for data collection. The required permissions are obtained and consent of pregnant women visiting for monthly consultations are also obtained for collecting data for the study. An interview schedule was developed for collection of that incorporating all the selected attributes of the study. This developed interview schedule was used for data collection from pregnant women who visited the identified medical centres for monthly check up. The population for the study was constituted by 600 pregnant women from who visited the medical centres for check up. Figure 1 shows the interview schedule followed for data collection. Different measuring devices are used for measuring the physiological parameters of the study. The values collected by a medical expert are entered into the dataset.

P_ID / Parameters		P_ID 1	P_ID 2	P_ID 3	P_ID 4	P_ID 5	..... P_ID n
Personal Details							
a	Age						
b	Height						
c	Weight on conceiving						
d	BMI						
e	Present trimester						
f	Present month						
g	Present weight						
Health information							
a.	Health state before pregnancy						
b.	Blood Pressure						
c.	Gestational Diabetic presence						
d.	Weight Gain						
Personal History							
a.	Pregnancy Parity						
b.	History of Pre-eclampsia						
c.	History of Gestational Diabetes						
Family History							
a.	Mother/ Sister with Pre-eclampsia						
b.	Mother/ Sister with Gestational Diabetes						

Fig 1: Interview schedule for data collection

4.3. Pregnancy Dataset Creation:

A pregnancy dataset is constituted by pregnancy data obtained from primary data collection using the interview schedule.

- Un-standardized Pregnancy Dataset: The un-standardized pregnancy dataset is the pregnancy dataset obtained from interview schedule. The dataset is checked for abnormal, error-prone, missing and irrelevant data samples. If irrelevant data samples are identified they are replaced by valid values or removed from the data set. The dataset is then partitioned into training and testing datasets. The partitioning of dataset is done to subject the pregnancy dataset to analysis using C4.5 classification

algorithm. The dataset is randomly divided and around 370 data samples constitute the training data and remaining 230 data samples constitute the test data.

- Standardized Pregnancy Dataset: Standardization is a statistical procedure followed by many researchers in various fields. Standardization of study parameters provides with valid, precise, reliable, meaningful and accurate set of parameters leading to elevated performance by the selected classifier. The procedure of standardization is followed since pregnancy is a delicate medical issue and parameters influencing pregnancy can be better determined by medical experts, who can provide a reliable set of parameters used for the study. A total number of 40 judges who are knowledgeable and experts in the field of gynaecology and obstetrics were selected in random. These judges were requested to give their opinion regarding suitability of the selected broad parameters for standardization. They were requested to use 3 point rating relevancy scale as suggested by Gulford(1954) which consisted of Most Relevant, Relevant and Not Relevant scoring used for responses and scored 2,1 and 0 respectively. The responses received from experts were subjected to analysis of relevancy component of i^{th} component (R_i) by using the formula suggested by Gulford(1954)

$$RC_i = \frac{Total\ score\ of\ all\ the\ judges\ on\ i^{th}\ component}{Maximum\ score\ on\ the\ continuum * Total\ number\ of\ judges} \quad (4)$$

The parameters were subjected to relevancy test and the irrelevant parameters were removed from the list of parameters considered for the study. From a set of twelve parameters selected for the study six parameters were considered to be irrelevant by 40 medical experts and hence removed. A set of six parameters were considered for the study which had a relevancy percentage of more than 70. The remaining parameters of study are as listed along with their ranks in table 2.

Table 2: Ranking order of parameters

Parameter/Attribute	Rank
State	1
Blood Pressure	2
Blood Glucose Level	3
Weight	4
Trimester	5
Month	6

Here standardization is the most appropriate method for parameter selection since problem considered for analysis is a sensitive and delicate medical issue and requires expertise in deciding the parameters for the study. The dataset is randomly divided and around 370 data samples constitute the training data and remaining 230 data samples constitute the test data.

4.4. Data Analysis:

Training and testing datasets are analysed using C4.5 decision tree classification algorithm, the training dataset constituted by 370 data samples is used to construct the decision tree learning model and testing dataset is used to evaluate the learning model constructed.

- C4.5 Classifier on Un-Standardized Dataset: Un-standardized pregnancy dataset is divided into training and testing datasets. Training dataset has 370 data samples with all 12 parametric values and result, the risk-levels corresponding to each data sample. Test dataset has 230 data samples with all 12 parametric values for each sample and no result or risk level corresponding to each data sample. C4.5 classification algorithm is applied on training dataset to build the decision tree learning model for un-standardized dataset. On obtaining the decision tree model test dataset is used to analyse the performance of the learning model in terms of accuracy in prediction of risks during pregnancy from un-standardized pregnancy dataset. The figure 3 shows the application of C4.5 classifier on un-standardized training dataset and evaluation of constructed learning model using un-standardized test dataset.

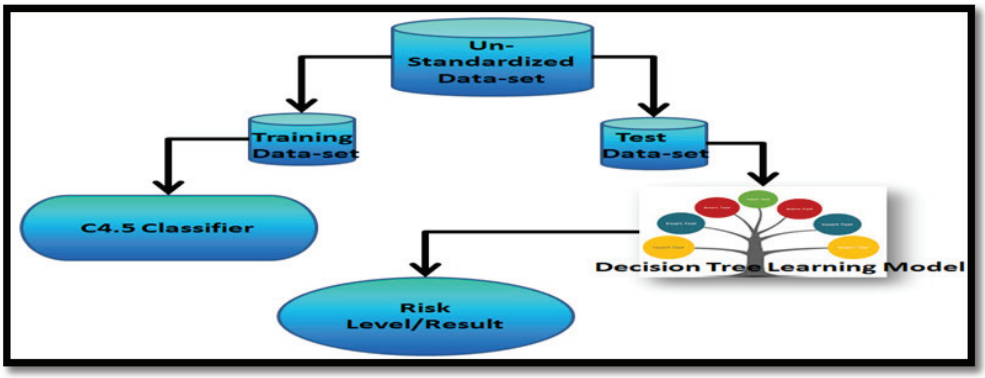


Fig 3: C4.5 Classifier on Un-Standardized Dataset

- C4.5 classifier on Standardized Dataset: After standardizing parameters of study, data samples corresponding to those parameters constitute the standardized pregnancy dataset. Standardized pregnancy dataset is randomly divided into 370 training and 230 testing datasets. Training dataset has 370 data samples with selected 6 parametric values and result, the risk-levels corresponding to each data sample. Test dataset has 230 data samples with selected 6 parametric values for each sample and no result or risk level corresponding to each data sample. C4.5 classification algorithm is applied on training dataset to build the decision tree learning model for standardized dataset. On obtaining the decision tree model test dataset is used to analyse the performance of the learning model in terms of accuracy in prediction of risks during pregnancy using standardized dataset. The figure 4 shows the application of C4.5 classifier on standardized training dataset and evaluation of constructed learning model using standardized test dataset.

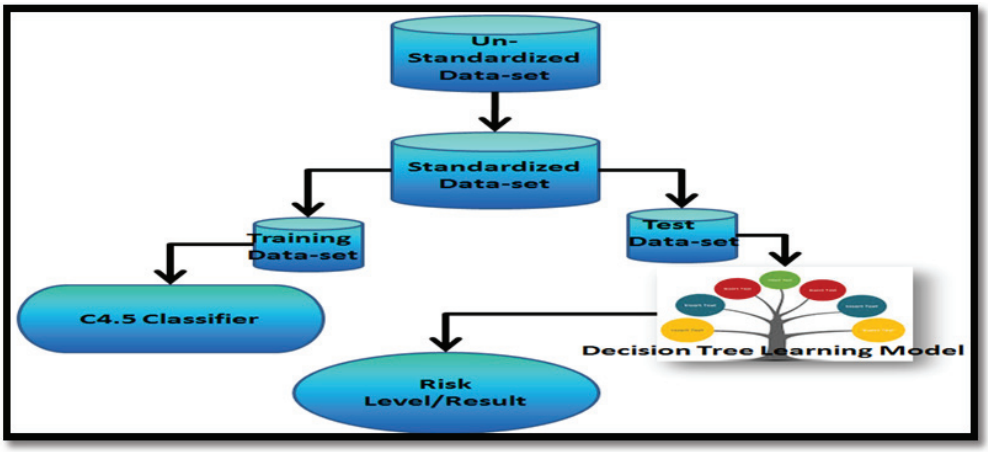
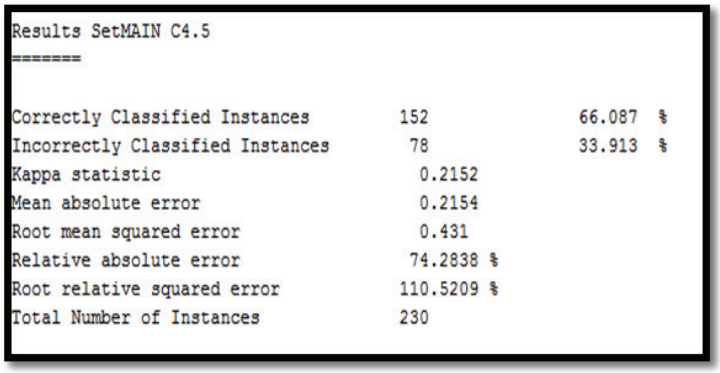


Fig 4:C4.5 on Standardized Dataset

The purpose of this evaluation process is to determine the accuracy of the learning models in prediction and classification when subjected to two different types of databases. The knowledge obtained from this evaluation process helps in identifying the impact of parameters on the accuracy of an algorithm for predicting the risk induced during pregnancy in women.

5. Results

The learning models generated by C4.5 algorithm are easy to understand and require no domain experts. This technique is efficient, powerful and popular in classifying the data and predicting the respective risks induced in pregnancy. The study also highlights a slight improvement obtained in accuracy when standardized parameters are selected for analyzing through C4.5 classifier.

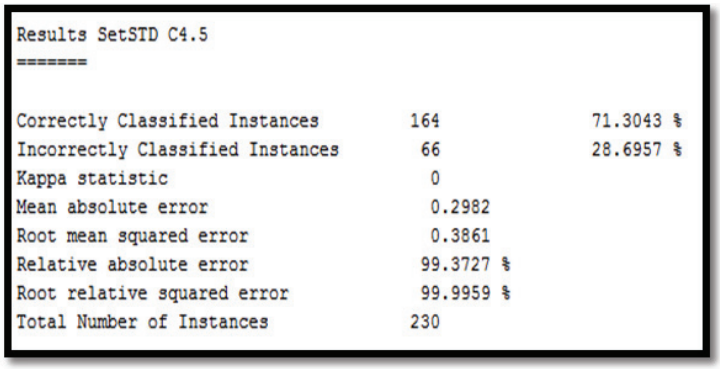


```
Results SetMAIN C4.5
=======

Correctly Classified Instances         152               66.087  %
Incorrectly Classified Instances        78               33.913  %
Kappa statistic                          0.2152
Mean absolute error                      0.2154
Root mean squared error                  0.431
Relative absolute error                 74.2838 %
Root relative squared error            110.5209 %
Total Number of Instances              230
```

Fig 5: Results of C4.5 on unstandardized test data

Figure 5 shows the summary obtained from the analysis of C4.5 on test dataset. From 230 data samples 152 data samples are correctly classified or the corresponding risk levels for data samples are predicted correcly and 78 data samples are incorrectly classified or the risk levels are incorrectly predicted.



```
Results SetSTD C4.5
=======

Correctly Classified Instances         164               71.3043 %
Incorrectly Classified Instances        66               28.6957 %
Kappa statistic                          0
Mean absolute error                      0.2982
Root mean squared error                  0.3861
Relative absolute error                 99.3727 %
Root relative squared error             99.9959 %
Total Number of Instances              230
```

Fig 6: Results of C4.5 on Standardized test data

Figure 5 shows the summary obtained from the analysis of C4.5 on test dataset. From 230 data samples 152 data samples are correctly classified or the corresponding risk levels for data samples are predicted correcly and 78 data samples are incorrectly classified or the risk levels are incorrectly predicted. The extracted summary of analysis for both datasets through learning model generated by C4.5 classifier using WEKA toolkit is as shown in the figures 3 and 4. The accuracy percentage for un-standardized dataset and standardized dataset accounts to 66.087% and 71.3043% respectively and error percentages are 33.913% and 28.6957% respectively.

6. Conclusion

It is evident from the results of analysis on both datasets it can inferred that C4.5 decision tree classifier has greater potential in accuracy for predicting the risk levels during pregnancy when applied on standardized data than on un-standardized data. Figure 7 shows the graph for accuracy and error obtained by C4.5 classifier on both un-standardized and standardized datasets. The graph presents that C4.5 classifier provides better performance with standardization than with un-standardized dataset.

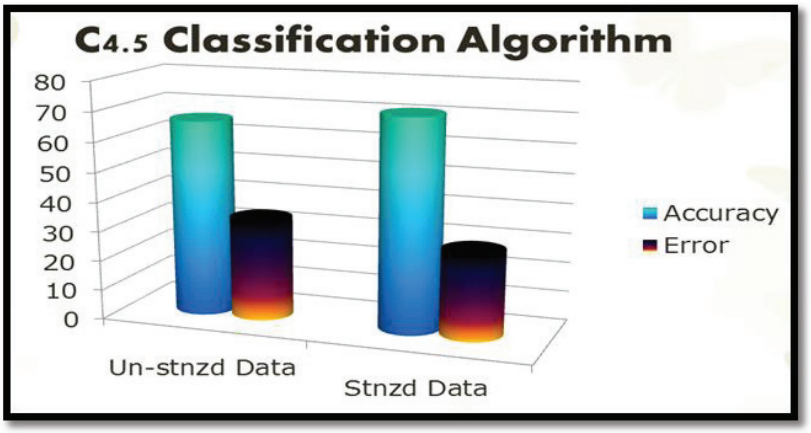


Fig 7: Graph showing the performance of C4.5 classifier on pregnancy data

In this study, the performance of C4.5 classifier is evaluated for accuracy and impact of standardization on accuracy is understood and identified for providing better prediction for risk in pregnancy. The paper provides significance of classification and prediction in mining pregnancy related data. Other classification techniques can also be used for analysing pregnancy data but for this study C4.5 classifier is considered for this problem because of it powerfulness, popularity and efficiency and the delicacy of the pregnancy problem. Fig 7 gives the accuracy and error percentages by C4.5 algorithm when applied on standardized and un-standardized pregnancy data. Thus as a concluding statement C4.5 classifier offers better performance if standardization is followed before analysis as seen from results obtained from both the datasets. Hence, from this study the paper pinpoints the advantages offered by C4.5 decision tree classification algorithm for prediction based health monitoring in pregnant women and importance of standardization. Research is further carried out to improve the accuracy even better after the standardization procedure to provide pregnant women a precise level of risk to give them a safe and healthy pregnancy period.

References

[1]. United Nations Maternal Mortality Estimation Inter-agency Group, consisting of representatives from the World Health Organization (WHO), United Nations Children's Fund (UNICEF), the United Nations Population Fund (UNFPA), United Nations Population Division, the World Bank and Wikipedia.

[2]. World Health Organization (November 2014). "Preterm birth Fact sheet N°363".who.int.Retrieved 6 Mar 2015.

[3]. Applications of Machine Learning in Cancer Prediction and Prognosis by Joseph A. Cruz, David S. Wishart, Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada T6G 2E8 in Cancer Informatics 2006: 2 59– 78

[4]. Decision Trees chapter 9 by Lior Rokach, Department of Industrial Engineerin, Tel-Aviv University and Oded Maimon, Department of Industrial Engineering, Tel-Aviv University

[5]. Machine Learning Approach for Preoperative Anaesthetic Risk Prediction by Karpagavalli S1, Jamuna KS2, and Vijaya MS21,2 GR Govindarajulu School of Applied Computer Technology, PSGR Krishnammal College for Women, Coimbatore, India, Email: karpagam@grgsact.com in International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[6]. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction by Jyoti Soni Ujma Ansari Dipesh Sharma, Student, M.Tech (CSE). Professor Reader, Raipur Institute of Technology, Raipur, Chhattisgarh, India in International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[7]. Research of Knowledge Based Expert System Used in Maternity Diagnosis, Lu Binjie, Ph.D. of Donghua University, IT Department , Renji Hospital affiliated Shanghai Jiaotong University School of Medicine, Shanghai, China, lu-bj@126.com in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010, 978-1-4244-7237-6/$26.00 C 2010 IEEE

[8]. Supervised Machine Learning: A Review of Classification Techniques, S. B. Kotsiantis, Department of Computer Science and Technology, University of Peloponnese, Greece, End of Karaiskaki, 22100 , Tripolis GR. Informatica 31 (2007) 249-268

[9]. Performance Tuning of J48 Algorithm for Prediction of Soil Fertility, Jay Gholap, Dept. of Computer Engineering, College of Engineering, Pune, Maharashtra, India

[10]. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages, Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos, Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", 153 10 Ag. Paraskevi, Athens, Greece