Prediction Model for Breastfeeding practice among Ethiopian Children using Decision tree and Rule induction Algorithms : Data mining

Taddele Weldeslassie Awalom (taddele.weldeslassie@mu.edu.et)
School of Public Health, College of Health Science, Mekelle University https://orcid.org/0000-0003-4923-7511
Mesaud Mohamodbrhan Adem
Tigray National Regional Health Bureau
Mearg Araya
Yekatit 11 Hospital

Research

Keywords: Breastfeeding practice, Data mining, Algorithm, Prediction Model, Ethiopia

Posted Date: August 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-58065/v1

Abstract

Background: Ethiopia have adopted infant feeding guidelines based on World Health Organization's standards to reduce the burden of infant and child mortality due to poor breastfeeding practice. But, breastfeeding practice is still one of challenges affecting infants and child health causing significant amount of deaths(23%-28%) yearly in Ethiopia. Breastfeeding practice is associated with different individual and community specific socio-cultural factors in different countries. Ethiopia is a populated country of communities with a very diverse cultural and societal values administered in nine different regions. Therefore, it is very important to assess breastfeeding practice among the various communities to identify the factors at individual and community level in order to come up with preventive intervention protocols that matches to each particular region. Hence, the study intended to assess patterns of breastfeeding practice among the communities within each specific region and develop predictive model of breastfeeding practice using data mining algorithms in Ethiopia. Different experiments were conducted in four scenarios with two test option (10 cross validation and percentage splits) and different parameter values using J48 and PART algorithms to select best predictive model for developing breastfeeding decision support system using java application programming interface.

Results: About 54.8% (6390) and 3.8% (445) of the mothers have ever and never breastfed their children within the previous five years of the survey respectively while 40.8% (4757) mothers were breastfeeding until the survey date. Both J48 and PART algorithms were able to predict breastfeeding practice with an accuracy of 96.86% and 96.77% respectively. 2316 (96.94%) and 1071 (96.04%) mothers were correctly classified as Normal and poor respectively using PART algorithm with 70-30 percentage-split test option. Only 66 (3.06%) and 43 (3.96%) mothers were misclassified as false positive and false negative respectively.

Conclusions: Almost, half of the mothers with 1-4 births within the five years before the survey have had normal breastfeeding practice. Both J48 and PART algorithms have best fitted to predict breastfeeding practice and can be used to deploy a decision support model of breastfeeding practice as a supporting tool for health practitioners.

Background

Only 34.8% of infants are exclusively breastfed for the first 6 months of life, the majority receiving some other food or fluid in the early months worldwide and complementary foods are often introduced too early or too late and are often nutritionally inadequate and unsafe [1].Breast milk and breastfeeding are among the most advocated natural products and practice respectively worldwide since both of them have beneficial effects to the infant and to the mother. The most significant and obvious benefits of breastfeeding are for the immediate health and survival of the infant. Rates of diarrhea[2], respiratory tract infections, otitis media, and other infections, as well as deaths due to these diseases, are all lower in breastfed than in non-breastfed infants. During the first six months, the rates are lower for exclusively breastfed than in partially breastfed infants[3, 4].

Children who were not breastfed as infants have an increased risk of childhood cancers [2]. The causes of childhood leukemia remain unknown, but there is an association between breastfeeding and reduced risk of childhood leukemia [3]. The beneficial effect of breastfeeding on the incidence and severity of gastrointestinal infections in infancy has been confirmed by multiple clinical trials. In a mature gastrointestinal tract, the intestinal epithelium acts as a barrier to bacteria or toxins. Many of these have nutritional effects on the intestinal mucosa, thereby decreasing the risk of gastrointestinal illness [3].Any breastfeeding is also associated with a small reduction in systolic blood pressure later in life, which may in turn decrease the risk for heart attack and stroke [2].

Alternatives to breast milk, such as animal milk or commercially made formula, carry risks of additional illness and death, particularly in areas where infectious disease levels and the potential for improper preparation and storage practices are high [4]. Breastfeeding can also reduce the severity, duration, and negative nutritional consequences of diarrhea [4].As a study conducted on examining risk factors and protective factors of SIDS through 19 studies, bottle-fed infants had a 2.1 fold increased risk of death from SIDS [5].

Breastfeeding have also a significant advantage to the mothers themselves. Breastfeeding has a dose dependent effect on decreasing risk for breast cancer indicating that breastfeeding for more than 12 months over a woman ‘s life decreases her risk for breast cancer [6]. Women who breastfed for a total of two years or more over their lifetime had a significantly decreased risk of 37% for cardiovascular disease, including stroke, later in life [5].Breastfeeding imposes an increased metabolic burden on mothers that included an increased energy requirement is approximately 480 kcal/d. This metabolic burden may be responsible for reduced blood glucose levels and thus a decreased risk of type 2 diabetes [7].

However, there are lots of factors that contribute to poor consumption and poor practice of breast milk and breastfeeding respectively. Almost all SES factors have a strong association with breastfeeding up to 6 months of infant’s postnatal age. Increasing level of education among women was identified as a factor which plays a role in the adoption of modern ideas, and which usually leads to the abandonment of traditional practices regarding childcare. Increased maternal age and high parity can also lead to breastfeeding of a shorter duration [1, 8, 9].

Data mining is a new generation of computerized technique for extracting previously unknown, valid, and actionable knowledge from enormous database and then using this knowledge to make critical decision[10]. Data mining predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based on the observed data. Rule induction is also a process of extracting useful if/then rules from data based on statistical significance. A decision tree is a tree-shaped structure that visually describes a set of rules that cause a decision to be made [11].

Today, in medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available [12], processing and analyzing the huge health care data by traditional statistical methods has their own difficulties[13]. To overcome this difficulty, Data mining provides the methodology and technology to transform these massive data into useful information for decision-making and problem solving [14]. Hence, the study used data mining predictive algorithms to identify factors associated with breastfeeding practice and develop prediction model with user interface using Java programming to serve as decision support system for health practitioners.

Results

Descriptive statistical summary of Socio-demographic characteristics

The dataset has been described and visualized using SPSS to examine the properties of the dataset relative to the whole records. Simple statistical analysis has been performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions. Hence, the selected attributes used for model building are statistically described in details to understand the dataset during experimentation and increasing the accuracy of the model.

Only 3.8% of the respondents had never breastfed their children until the survey. 40% and 54.8% of the respondents had ever breastfed but were and were not breastfeeding during the time of survey. the attribute Amenorrheic shows the unusual absence of menstruation. 55.1% of the respondents have had Amenorrheic while 44.9% of them have had no Amenorrheic during the time of interview. Majority of the respondents (88.8%) were not pregnant during the study period. About 92% of the children born during the previous five years from the mothers included in the study were alive. Most (85.22%) of the mothers have had one and two birth histories while the rest of them had have 3 and 4 births for the last five years. Only Five of the total respondent's also have greater than four births [Table 1].

About 83% and 17% of the respondents were rural and urban residents respectively. 78.7% of the respondents had have no history of diarrhea within the last recent two weeks before the survey date. Out of the total respondents, 49.2%, 16.1% and 34.7% were poor, middle and high income mothers respectively. 70.3% of the respondents have had an experience of watching television while the rest of them have not practiced watching television before the study. Majority (85.2%) of the mothers included in the study have delivered their labour at home, while only 11.4% and 2% of the mothers have got institutional delivery service at public and private health institutions respectively. 74.7% of the respondents had no fever and 17.9% of them had fever during the time of surveying. Most of the respondents were illiterate (69.9%), 25.1% were primary school attendants, 3.3% were secondary and 1.7 were graduated mothers. About half of the children have had average weight (51.9%), 32.4% less than average, 1.2% greater than the average weight [Table 1].

Table 1
Descriptive statistical summary of Socio-demographic characteristics

Attributes	Values	N	%	Attributes	Values	N	%
Duration of breastfeeding	Ever breastfed	6390	54.8	**Wealth index**	poor	5739	49.2
	Never breastfed	445	3.8		middle	1872	16.1
	Still breastfeeding	4757	40.8		rich	4043	34.7
	Missing	62	.5	**Frequency of Watching TV**	no	8195	70.3
Currently Amenorrheic	No	6422	55.1		yes	3447	29.6
	Yes	5232	44.9		Missing	12	.1
Currently pregnant	No/don't know	10351	88.8	**Place of delivery**	home	9934	85.2
	Yes	1303	11.2		public	1334	11.4
Region	Tigray	1202	10.3		private	237	2.0
	Affar	1130	9.7		others	129	1.1
	Amhara	1294	11.1		Total	11634	99.8
	Oromiya	1761	15.1		Missing	20	.2
	Somali	1027	8.8	**Had fever**	no	8710	74.7
	Benishangul-Gumuz	1020	8.8		yes	2082	17.9
	SNNP	1614	13.8		Total	10792	92.6
	Gambela	851	7.3		Missing	862	7.4
	Harari	659	5.7	**Educational attainment**	Illiterate	8142	69.9
	Addis Ababa	400	3.4		Primary	2930	25.1
	Dire Dawa	696	6.0		Secondary	386	3.3
Child is alive	No	846	7.3		Higher	196	1.7
	Yes	10808	92.7		Total	11654	100.0
Birth in the last five years	1 or 2 births	9926	85.2	**Child weight**	Less than Average	3774	32.4
	3 or 4 births	1723	14.8		Average	6050	51.9
	> 4 births	5	.0		Greater than Average	138	1.2
Type of place of residence	Urban	1986	17.0		Others	447	3.8
	Rural	9668	83.0		Total	10409	89.3
Had **diarrhea**	no	9173	78.7		Missing	1245	10.7
	yes	1620	13.9				
	Missing	861	7.4				

J48 Decision Tree Prediction Model output

In this study, different experiments were conducted altering parameters of the J48 decision tree and PART rule induction algorithm for building the best predictive model. The J48 decision tree algorithm builds decision trees from a set of predefined training dataset using the concept of information entropy and attribute ordering. It uses the fact that each attribute of the data was used to make a decision by splitting the data into smaller subsets.

Table 2
Experimentation result of J48 Algorithms in scenarios one and two

Performance measurements	Experiments												
	Scenario one						Scenario two						
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
Accuracy (%)	96.17	94.92	96.49	95.12	96.77	95.35	96.64	96.45	96.93	96.55	95.44	96.9	96.95
Mean absolute error	0.05	0.07	0.05	0.07	0.04	0.06	0.04	0.05	0.04	0.04	0.06	0.03	0.04
Numbers of leaves	480	280	428	293	454	343	454	408	501	484	408	501	501
Size of tree	555	376	581	396	615	62	615	555	684	657	550	684	684
Time taken	0.26	0.12	0.12	0.11	0.12	0.12	0.26	0.13	0.07	0.07	0.06	0.03	0.04
AV.TP rate	0.96	0.95	0.96	0.95	0.97	0.96	0.96	0.96	0.97	0.97	0.96	0.97	0.97
AV.FP rate	0.04	0.06	0.04	0.05	0.04	0.06	0.04	0.04	0.04	0.04	0.06	0.04	0.03
AV. Precision	0.97	0.96	0.97	0.97	0.97	0.97	0.96	0.98	0.97	0.97	0.96	0.97	0.98
AV.Recall	0.96	0.95	0.96	0.95	0.97	0.96	0.96	0.96	0.97	0.97	0.96	0.97	0.97
AV.ROC area	0.98	0.98	0.99	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
CCI	11209	11062	11246	11086	11277	11113	3829	3372	11297	11253	11123	3839	3390
ICCI	445	592	408	568	377	541	133	124	357	401	531	123	106

Key: CCI: Correctly classified Instance, ICCI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TP: True Positive. FP: False Positive, ROC: Relative Optical character curve.

As we can see in Table 2 the result of each experiment developed model the unpruned experiment have best accuracy more than pruned experiment. As the result Experiment # 13 (building decision tree unpruned with 70 – 30 percentage split) is the best with an accuracy of 96.95%. Experiment # 9 also showed best performance next to experiment # 13 with an accuracy of 96.93%. both experiment #9 and #13 are unpruned experiments. The pruned experiment #5 has also good performance next to the above two experiments and better than all the other pruned experiments with an accuracy of 96.77%. In general, the unpruned experiments had shown good performance than the pruned experiments.

J48 Decision Tree Prediction Model Evaluation

The experiments conducted above have been analyzed and evaluated in terms of classifiers performance values, accuracy, confusion matrix values, TP and FP Rate, number of leaves, and size of the tree generated, ROC curves and execution time. Performance of the classifier on the testing set increased as the confidence factor increased up to about 0.5. Experiment #5 showed an accuracy of 96.77%. At this accuracy correctly and incorrectly classified instance are 11279 and 377 respectively from 11,654 instances [Table 3]. From thirteen different trials experiment #5 is the best model in terms of accuracy and minimized incorrectly classified instances. The Confusion Matrix of Experiment #5 in Table 3 shows the number of instances of each class that are assigned to all possible classes according to the classifier's prediction. The columns represent the predictions, and the rows represent the actual class.

Table 3
summary of confusion matrix for J48

		Predicted Breast feeding practices		
		Positive	**Negative**	**Total**
Actual Breast feeding Practices	**Positive**	7568	7785	7785
	Negative	158	3869	3869
	Total	7726	3928	11654

The confusion matrix in Table 3 shows that 7568 instances were correctly predicted as normal breast feeding practice (True positive). True positive of the actual class of the test instance is Normal breast feeding practice and the classifier correctly predicts the class as Normal breast feeding practice. The numbers of instance which were correctly predicted as poor breastfeeding practice are 3711 instances (True negative). In

this case of true negative the actual class of the test instance is poor breastfeeding practices and the classifier correctly predicts the class as poor breast feeding practices. Therefore, correctly classified instances are the sum of diagonal values of the table, which are 11279 instances correctly classified from 11,654 instances.

In contrast, 158 instances were predicted as a normal breastfeeding practice while they were in fact poor breastfeeding practice (False Positives). A false positive is when the actual class of the test instance is poor breastfeeding practice but the classifier incorrectly predicts the class as normal breast feeding practice. The classifier predicted 217 instances as poor breastfeeding practice (False Negatives). A false negative is when the actual class of the test instance is Normal breast feeding practice but the classifier incorrectly predicts the class as poor breastfeeding practice.

The result in Table 4 has been extracted from Experiment #5 model. True Positive rate shows the percentage of low weight instances whose predicted values of the class attribute are identical with the actual values. FP rate shows the percentage of instances whose predicted values of the class attribute are not identical with the actual values.

Table 4
Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.972	0.041	0.979	0.972	0.976	0.992	NORMAL
	0.959	0.028	0.944	0.959	0.952	0.992	POOR
Weighted Av	.0.968	0.037	0.968	0.968	0.968	0.992	

If we take the first level where 'breast feeding practices = POOR' TP Rate is the ratio of poor breastfeeding cases predicted correctly to the total of positive cases, there were 3711 instances correctly predicted as poor breastfeeding practice, and 3869 instances in all that were poor breastfeeding practice. So the TP Rate (True Positive Rate) of poor breastfeeding practice = 3711/3869 = 0.959. The FP Rate is then the ratio of normal breastfeeding practice of incorrectly predicted as poor breastfeeding practice to the total of normal breastfeeding practice cases. 217 normal breast feeding practice instances were predicted as poor breastfeeding practices and there were 7785 normal poor breastfeeding practices in all. So the FP Rate is 217/7785 = 0.028. We can follow the same method to calculate for 'breast feeding practice = normal' but as we can see from detailed accuracy by class TP Rate and FP Rate of Normal class level are 0.972 and 0.041 respectively. The model performance is good quality because it has high true positive rates with low false positive rates [Table 4].

As can be seen from the detailed accuracy by class output in Table 6, the ROC (Receiver Operating Characteristics) area of this model is highest (0.992). The Area under the ROC area curve of experiment #5 is higher. Higher numbers here indicate the model is the more accurate. The ROC curve is a plot of how the classifier is performed over the entire range of possible choices of cutoff values. Each point on the curve represents the True-Positive Rate plotted on the y-axis and the False-Positive Rate plotted on the x-axis that resulted from a particular cut-off value as shown in *Fig. 1*.

PART Rule Induction Prediction Model output

To build the Rule induction model using PART algorithm, WEKA software package and the same number of datasets were used as an input. The experiments were divided into two scenarios with two test option that are 10-fold cross validation and percentage split evaluator.

Table 5
Experimentation result of PART Algorithms with one and two scenarios

Performance measurements	Experiments												
	Scenario one						Scenario two						
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
Accuracy (%)	96.74	95.45	96.78	95.33	96.86	95.46	96.25	96.86	96.94	96.71	95.93	96.87	97.10
Mean absolute error	0.04	0.06	0.04	0.06	0.04	0.06	0.04	0.04	0.03	0.04	0.04	0.03	0.03
Numbers of leaves	180	150	180	156	191	152	191	191	282	277	262	282	282
Size of tree	0.85	0.38	0.43	0.37	0.49	0.33	0.97	0.94	1.65	1.42	1.40	1.68	1.66
Time taken	0.97	0.96	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.97
AV.TP rate	0.04	0.05	0.04	0.06	0.04	0.06	0.05	0.04	0.04	0.04	0.05	0.04	0.04
AV.FP rate	0098	0.97	0.97	0.96	0.97	0.96	0.97	0.98	0.97	0.97	0.97	0.97	0.98
AV. Precision	0.97	0.96	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.97
AV.Recall	0.99	0.98	0.99	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.98	0.99
AV.ROC area	11274	11124	11279	11110	11282	11126	3814	3386	11298	11271	11180	3838	3387
CCI	380	530	375	544	372	528	148	110	356	383	474	124	109
ICCI	96.74	95.45	96.78	95.33	96.86	95.46	96.25	96.86	96.94	96.71	95.93	96.87	97.10

As shown in Table 5, the registered performance in case of induction rule learner, the unpruned is better than the pruned one. Among the 13 experiments an Experiment #13 (70 – 30 percentage split) registered the best performance of 97.10%. This shows that out of the testing set of 3496 records, 3387 (97.10%) of the records are correctly classified, while 109 (2.9%) of the records are misclassified. Experiment #5 also registered the best performance out of all the experiments using pruned parameter with an accuracy of 96.86%.

PART Rule Induction Prediction Model Evaluation

The resulting confusion matrix shown in Table 6 depicts that out of the total 2382 normal breast feeding practice instances 2316 (96.94%) of them are correctly classified in their respective class, while 66 (3.06%) of the records are incorrectly classified as poor breastfeeding practice. In the other hand, out of the total poor breastfeeding instances 1071 (96.04%) of them are correctly classified as poor breast feeding practices and 43 (3.96%) of the records are misclassified.

Table 6
Confusion matrix of PART algorithm with 70 – 30 percentage-split

		Predicted Breast feeding practices		
		Positive	**Negative**	**Total**
Actual Breast feeding Practices	**Positive**	2316	2382	2382
	Negative	43	1114	1114
Total		2359	1137	3496

J48 And Part Models Accuracy Comparison

The two selected classification models J48 and PART with their respective accuracy, Precision and number of instances correctly classified and misclassified.

Table 7
Performance comparison of selected best models

Types of algorithms	Accuracy (%)	Time taken (sec/)	Correctly classified	Misclassified
J48	96.77	0.97	11277	377
PART	96.86	0.98	3380	110

As shown in Table 7, PART rule induction algorithm classifier outperforms J48 classifier with an accuracy of 96.86% and it was selected as the better classifier for predicting breastfeeding practice.

Evaluation of Discovered Knowledge

About 191 rules/patterns were generated by the PART algorithm from the experiment #5. Consequently, to evaluate the importance of the discovered knowledge/rules, whether they are acceptable/not and whether they go in line with what is already known in the real world practice, domain experts from Mekelle University Ayder Referral Hospital were consulted. Finally, 39 rules generated by the PART algorithm were selected as best rules. Rule 1 – Rule 7listed below were also selected as the most interesting and best rules or discovered knowledge.

Rule 1

If Amenorrheic = “no” AND Birth within 5 Years interval =”one or two” AND Region = “Tigray” AND Watching Television = “no”, Delivery Place = “Home” AND Alive = “Yes” AND Mother Educational Status = “illiterate” AND Weghit of child = “Average” then the child will have poor breast practice (87.0/3.0).

Rule 2

-If Amenorrheic =”no” AND Birth within 5 Years =”one or two” AND Pregnant = “Yes” AND Delivery Place = “Home” AND Fever = “no” AND Diarrhea = “no” then the child will have poor breastfeeding practice (63.0/6.0).

Rule 3

If Amenorrheic= “no” AND Birth within the 5 year interval =“one or two”, Delivery Place=”Home” AND Educational Status of the mother is “illiterate” AND child lives in Amhara, Somali, Tigray, Oromiya, affair, Gamble and Benishangul-Gumuz, then the child will have poor breastfeeding practice.

Rule 7

If Delivery place= “Home”, Television=”yes”, Diarrhea=”no” and Alive= “yes” then child will have a normal breast feeding (120.0).

Rule 6

-If Amenorrheic= “no” AND Birth within 5 years interval= “one or two” AND Diarrhea=”no” AND weight of the child at birth time= ”larger than average”, then child will have Poor breastfeeding practice. (98.0).

Rule 2

If Amenorrheic =”no” AND Birth within 5 year interval=”one or two” AND Delivery Place = “private sector” then child will have poor breastfeeding practices (113.0/7.0).

Rule 3

If Amenorrheic=”no” AND Birth within 5 year interval= “one or two” AND Region=” Addis Ababa” AND Fever=”no”, then the child having poor breastfeeding practice will happen (110.0).

In general, the above rules indicated that, the attributes delivery place, educational status of mother, pregnancy, watching television and the weight of the child at birth time was found to be the most determinate factors for child breastfeeding practice. Whereas, the model assumed that some attributes like region, duration of breastfeeding, amenorrheic, place of residence, number of birth within 5 years’ interval, child Alive, diarrhoea, family wealth status and fever are less determinate factors for breast feeding practice. Finally, we agreed with the general rules that the model produced and findings of the current research.

Use of the Discovered Knowledge

In order to show how to use the discovered knowledge for the domain expert, user interface was designed by using JAVA programming language as an interaction point between the user and the system. WEKA is written in the Java language and contains a Graphical User Interface (GUI) for interacting with data files and producing visual results. It also has a general Application Page Interface (API); WEKA can be embedded like any other library in applications. Hence, Java application was deployed in to the selected predictive model as a decision support system for breastfeeding practice. Accordingly, the outputs of the prediction model were classified as NORMAL and POOR breast feeding practice based on the filled attribute values. You can see a model output predicting breastfeeding practice as NORMAL in *Fig. 2* and a model output predicting breastfeeding practice as POOR *Fig. 3.*

Discussion

As the study result has shown experiment #5, experiment #9 and experiment #13 of the J48 model are the best experiments which had achieved good accuracy 96.77%, 96.93%, and 96.95% respectively. But, when we compare the size and leaves of trees of unpruned J48 model, the number is enormous and complex relative to pruned one. As a result, the algorithms might not reach optimality and generate more generalized decision tree rules and over fitting problem. Besides, such situation has its own impact on classification performance particularly classifying unseen or new instance. Subsequently to solve the problem I have selected pruned scenario that perform better accuracy. Accordingly, experiment #5 (Building pruned decision tree) of 10- fold cross validation selected as the best J48 decision tree model.From the confusion matrix result of the J48 model, experiment #5 predicted 158 instances as a normal breastfeeding practice while they were in fact poor breastfeeding practice (False Positives) and 217 instances as poor breastfeeding practice (False Negatives) while they were in fact normal breastfeeding practice. Therefore, it is possible to say the model was better at predicting poor breastfeeding practice cases than the other experiments.

Furthermore, evaluating the model based on sensitivity and specificity are very significant in decision making. For that reason, the result of the above confusion matrix indicates that the sensitivity of this test was (7568/7785) = 97.21% and the specificity was (3711/3869) = 95.91%. The test indicates that the model appears to be pretty good. Because, based on the evaluation criteria, the classifier correctly classifies child as poor breastfeeding practice who had actually poor breast feeding practices with 95.91% accuracy and classify child as normal breast feeding practice who had actually normal or good breastfeeding practices with 97.21%. As can be seen from the detailed accuracy by class output in Table 6, the ROC (Receiver Operating Characteristics) area of this model is highest (0.992). The larger the area under the ROC curve the more accurate the test. Unpruned methods and techniques have shown increased classification accuracy given an induced decision tree. But the size of the tree is very large and complex to interpret. Hence, the pruned one, experiment #5 was selected.

In the case of PART algorithm, all experiments were also evaluated according to the performance measurement results they attained as in the case of J48 algorithm. And experiment #13 has scored the greatest accuracy than the others. But due to the other performance measurement results, like large number of rules, higher time taken, experiment #5 was selected as working experiment for model building. In general, the two classification models, J48 and PART, with respect to their performance of accuracy, Precision and number of instances correctly classified and misclassified were compared and evaluated. PART rule induction algorithm classifier outperforms J48 classifier with an accuracy of 96.86% and it was the better classifier in predicting breastfeeding practice. While J48 classifier achieved 96.77% accuracy. The better result that was registered in PART rule induction might be due to the linearity of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class. Moreover, in terms of ease and simplicity to users the PART rule induction is more self-explanatory, since; the result is presented in a form of “If-then”. The “If-then” rules can be easily represented in simple human understanding language.

The study showed that if delivery place of child is in home and mother of a child is illiterate; all regions except Addis Ababa the baby will have poor breast feeding practices. Rule1 is a good indicator of this fact. The domain experts were also agreed with this finding. Most of the time, the baby who was delivered at home would not support by health professionals. So mothers might not be consulted regarding to breastfeeding. Similarly,the study showed that a child which is born in the private sector had poor breastfeeding practice. Based on the domain expertise this fact indicates the persons working in the private sector might not have enough skill with regard to breastfeeding or professionals might not properly communicate with mothers on breastfeeding practice.

In the five regions, namely; Amhara, Affar, Dire Dawa and Tigray, if the mothers don’t have television at their home and mother educational status is illiterate then the child will have poor breastfeeding practice. Domain experts agreed with this fact, because the media broadcast advertisements on the benefits of breastfeeding is useful. This study demonstrated that place of delivery and frequency of watching television are determinant factors of breastfeeding practice according to the evaluation of the domain experts and the results from the PART algorithm. This might be due to the information and awareness they gained from health professionals during delivery and from the promotions about breast feeding advantages through mass Medias in Ethiopia.

Conclusions

In this study, attempts have been made to use DM technology with the aim of identifying and predicting breastfeeding practice of child in the healthcare institution. Experimentation was conducted using four scenarios in two test options (10-fold cross validation and percentage split) for each algorithm. J48 and PART algorithm performed 96.77% and 96.86% accuracy respectively. The extracted rules in both algorithms were very effective for predicting breastfeeding practice and PART rule induction algorithm with 70 – 30 percentage split were selected as a predictive model with a better performance than J48.Moreover, the finding of this research indicates that delivery place, mothers' educational status, resident place, child weight, pregnancy and watching television are determinant factors of child breastfeeding practice. In general, the results from this study can contribute towards encouraging and support the decision for healthcare organization and health practitioner.

Methods

Study area and Data Source

The study was conducted in Ethiopia using nationally representative cross-sectional survey data obtained from Ethiopian Demographic Health Survey (EDHS) 2016.The data was taken from the measure demographic health survey repository via official request letter and approval consent letter of measure DHS.The national survey data set contains 928 attributes and to decide on the relevant attributes for this study we have discussed with domain experts in the area and an extensive literature review. In addition of those techniques we used an attribute ranking with the evaluation of information gain. Finally, the following attributes were selected by prioritizing by the WEKA software Information Gain attribute evaluation algorithm, together with their rank and information gain value are listed Table 8.

Table 8
List of candidate attributes ranked according to Information Gain attribute evaluation algorithm.

Rank	Attribute name	Data type	Information gain value
1	Duration breastfeeding	Nominal	0.40558
2	Currently Amenorrheic	Nominal	0.26596
3	Currently Pregnant	Nominal	0.13033
4	Region	Nominal	0.02001
5	Child's Alive	Nominal	0.01767
6	Birth in the last five years	Nominal	0.0131
7	Place of Resident	Nominal	0.01066
8	Had diarrhea recently?	Nominal	0.01
9	Wealth status	Nominal	0.00887
10	watching TV	Nominal	0.00861
11	Place of delivery	Nominal	0.0069
12	Had fever in last two weeks?	Nominal	0.00681
13	Mother's educational status	Nominal	0.00402
14	Child weight	Nominal	0.00392

Data Processing

Usually, a real world database contains incomplete, noisy and inconsistent data and such unclean data may cause confusion in the data mining process. Hence, data was cleaned using SPSS and WEKA (version 3.7.7) data mining tool. Missing values were handled using SPSS preprocessing techniques and replaced with the most frequent (modal) value methods for all categorical variables. Some attributes were discretized to reduce the unlike values of the attribute to obtain knowledge (pattern) and to make the dataset suitable for data mining tools. The original SPSS dataset was then converted in to WEKA acceptable comma separated values (CSV) file format. Then the CSV file format is converted into an ARFF by using WEKA mining software, to take advantage of easier data manipulation and also compatible interaction with WEKA software. Finally, 14 attributes with 11,654 instances that are ready for experimentation process were included in the study.

Experimentations

Two classification algorithms namely J48 and PART induction rule algorithms were selected and deployed through WEKA machine learning software. WEKA 3.7.7 software was used to measure the quality, validity and test of the selected model. For purposes of this study k-fold (10-

folds) cross validation and percentage split test options were used because of their relatively low bias and variations. In 10-fold cross validation, the data were divided in to 10 folds where 9 folds were used as training data whereas the remaining one fold as test data. In the percentage split method, where 70% of the data was used as training and the remaining 30% was used as test data. Accuracy, Precision, Specificity, ROC curve, Recall and confusion matrix standard metrics were also used for evaluation of the results. For both the above methods the following four scenarios has been done with different parameter values of WEKA 3.7.7 software.

- **Scenario 1**: Decision tree with pruning.
- **Scenario 2**: Decision tree without pruning.
- **Scenario 3**: Rule induction with pruning.
- **Scenario 4**: Rule induction without pruning.

Once the modeling tool was chosen based on the performance evaluation criteria established, building model was done with a number of parameters that govern the model generation process [Table 9].

Table 9
Values of parameters used for 13 experiments

Experiments	Parameters			
	Pruned	Confidence factor	(min Numobj)	Test option
Experiment #1	True	0.25	2	10 fold cross validation
Experiment #2	True	0.25	5	10 fold cross validation
Experiment #3	True	0.30	2	10 fold cross validation
Experiment #4	True	0.30	5	10 fold cross validation
Experiment #5	True	0.50	2	10 fold cross validation
Experiment #6	True	0.50	5	10 fold cross validation
Experiment #7	True	0.50	2	66% percentage split
Experiment #8	True	0.50	2	70% percentage split
Experiment #9	False	0.50	2	10 fold cross validation
Experiment #10	False	0.50	3	10 fold cross validation
Experiment #11	False	0.50	5	10 fold cross validation
Experiment #12	False	0.50	2	66% percentage split
Experiment #13	False	0.50	2	70% percentage split

Abbreviations

EDHS
Ethiopia Demographic and Health Survey
ROC
Receiver Operating Characteristics area of the model
SPSS
Statistical Package for Social Sciences
SIDS
Sudden Infant Death Syndrome
WEKA
Waikato Environment for Knowledge Analysis

Declarations

Ethics approval and consent to participate

Ethical clearance for this study was obtained from Mekelle, University College of health Sciences, and Department of public health Ethical Review Board. A permission letter was also obtained from MEASURE DHS online.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analyzed during the current study are available in the measure DHSS repository, at https://dhsprogram.com/data/dataset/Ethiopian_Standard-DHS_2016.cfm. The datasets are also available from the corresponding author up on request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not Applicable

Authors' contributions

TW was the primary author responsible for all implementation activities of the research starting from its conception, design, data collection, analysis, interpretation, and write-up of the manuscript. MM and MA contributed in the data analysis, interpretation, and write-up of the manuscript. All authors read and approved the final manuscript submitted for publication.

Acknowledgements

We would like to thank Mekelle University College of Health Sciences, School of Computing and Tulane University Technical Assistant Project Ethiopia (TUTAPE) for allowing me to join the program, and financial and technical support. We would also like to express our sincere appreciation to SemawFerede (MSc) for his valuable suggestions and comments during the research progress.We are immensely indebted to Dr. Amanuel Hadgu and all pediatrician expert staffs for providing appropriate professional comment and explanation about the problem domain and interesting rules.

Authors' information

TW: BSc. In Computer Science, MSc. In Biostatistics and Health Informatics

MM: BSc. In Computer Science, MSc. In Monitoring and Evaluation

MA: Diploma in Health Information Technology, BSc, in Health Informatics

References

1. UNICEF, *Progress for children-a report card on nutrition*:, in *Ref Type: Report*. 2008: New York.
2. Bener A, et al. Does prolonged breastfeeding reduce the risk for childhood leukemia and lymphomas? Minerva pediatric:. 2008;60:155–61.
3. Papaemmanuil E, et al., *Risk of childhood acute lymphoblastic leukemia. Nature genetics*:. Vol. 41. 2009.
4. Schwarz EB, et al., *Duration of lactation and risk factors for maternal cardiovascular disease*. Obstetrics and gynecology:, 2009: p. 113–974.
5. Lamberti L, et al. Breastfeeding and the risk for diarrhea morbidity and mortality. BMC Public Health:. 2011;11:515.

6. Schubiger G, SCHWARZ U, TONZ O. *baby-friendly hospital initiative: does the use of bottles and pacifiers in the neonatal nursery prevent successful breastfeeding?* 2007, European journal of pediatrics. p. 874–877.
7. Olson DL, et al., *Data Mining Process:Advanced Data Mining Techniques:.* 2008: p. 9–35.
8. Ewa EE, et al., *perceived factors influencing the choice of antenatal care and delivery centers among childbearing women in Ibadan north south-western, Nigeria:.* 2012.
9. Roman SB. *Exclusive breastfeeding practices in rural Haitian women*. 2007, UCHC Graduate School. p. 141.
10. Pal NR, *Advanced techniques in knowledge discovery and data mining.* Springer, 2005.
11. EZEKOWITZ MD, et al. Rationale and design of RE-LY: randomized evaluation of long-term anticoagulant therapy, warfarin, compared with dabigatran. American Heart Journal:. 2009;157:805–10.
12. Eapen AG, *Application of Data mining in Medical Applications.* Citeseer:, 2004.
13. Rogers G. and E. JOYNER, Mining Your Data for Healthcare Quality Improvement. Journal of Healthcare Information Management:. 2005;19(2):65.
14. Koh HC, TAN G. *Data mining applications in healthcare*. Journal of Healthcare Information Management:, 2011. 19(65).

Figures

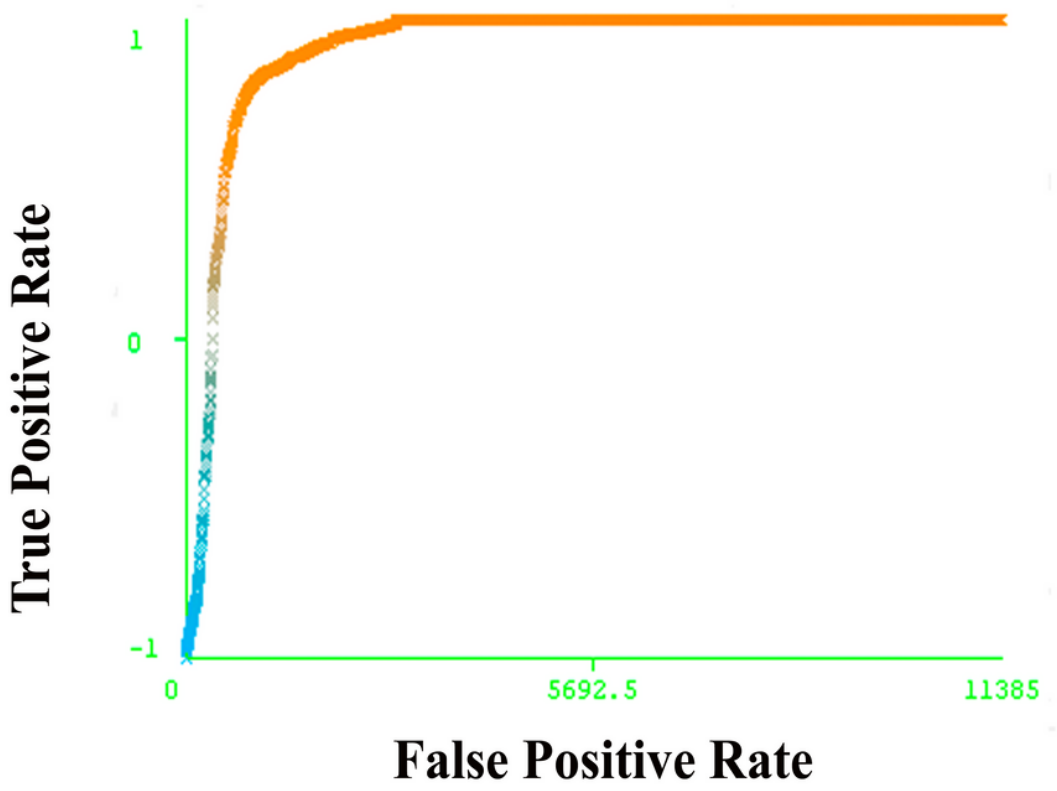

Figure 1: ROC area curve

Figure 1

Figure 1

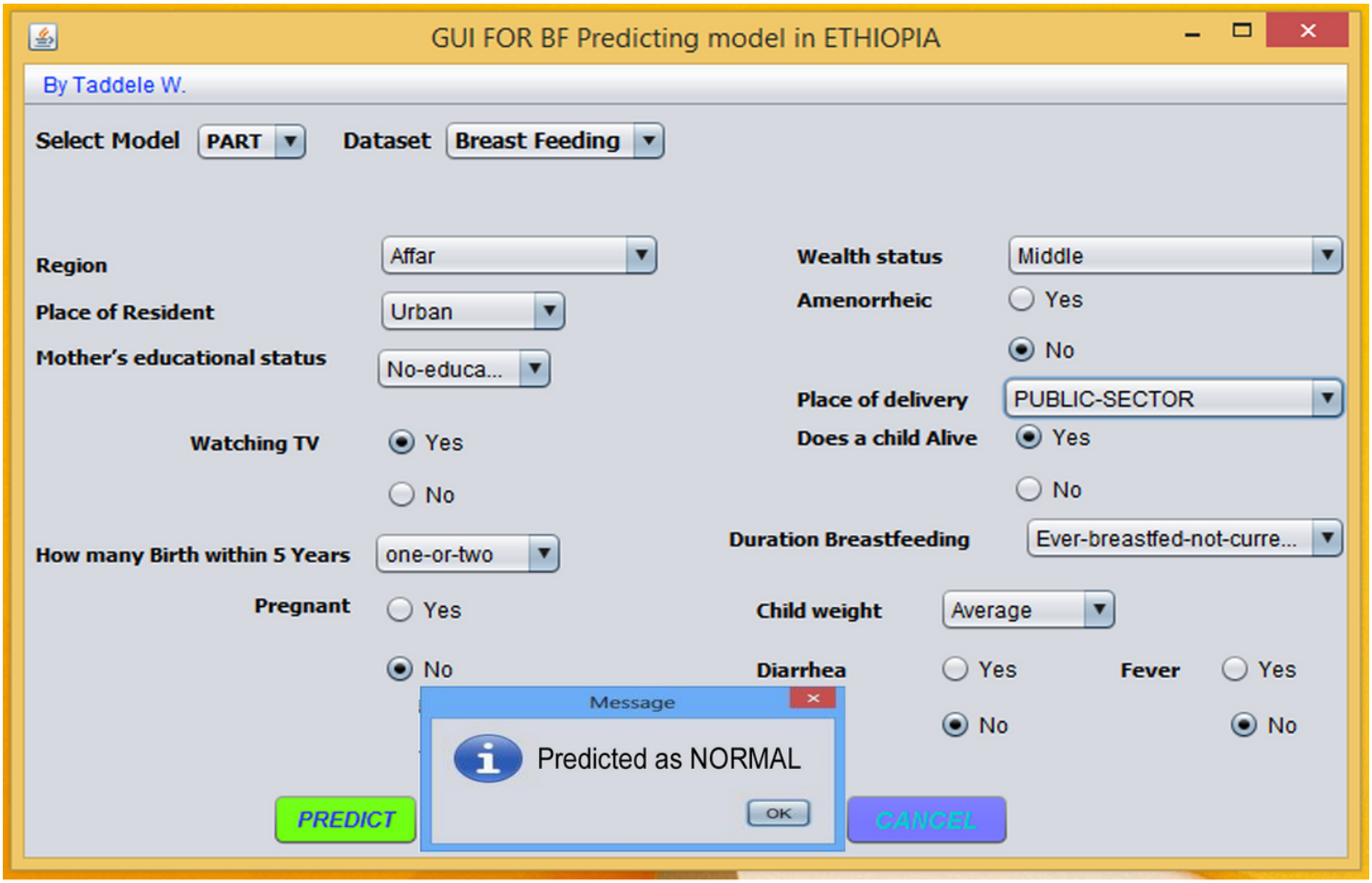

Figure 2:Model output predicting breastfeeding practice as NORMAL

Figure 2

Figure 2

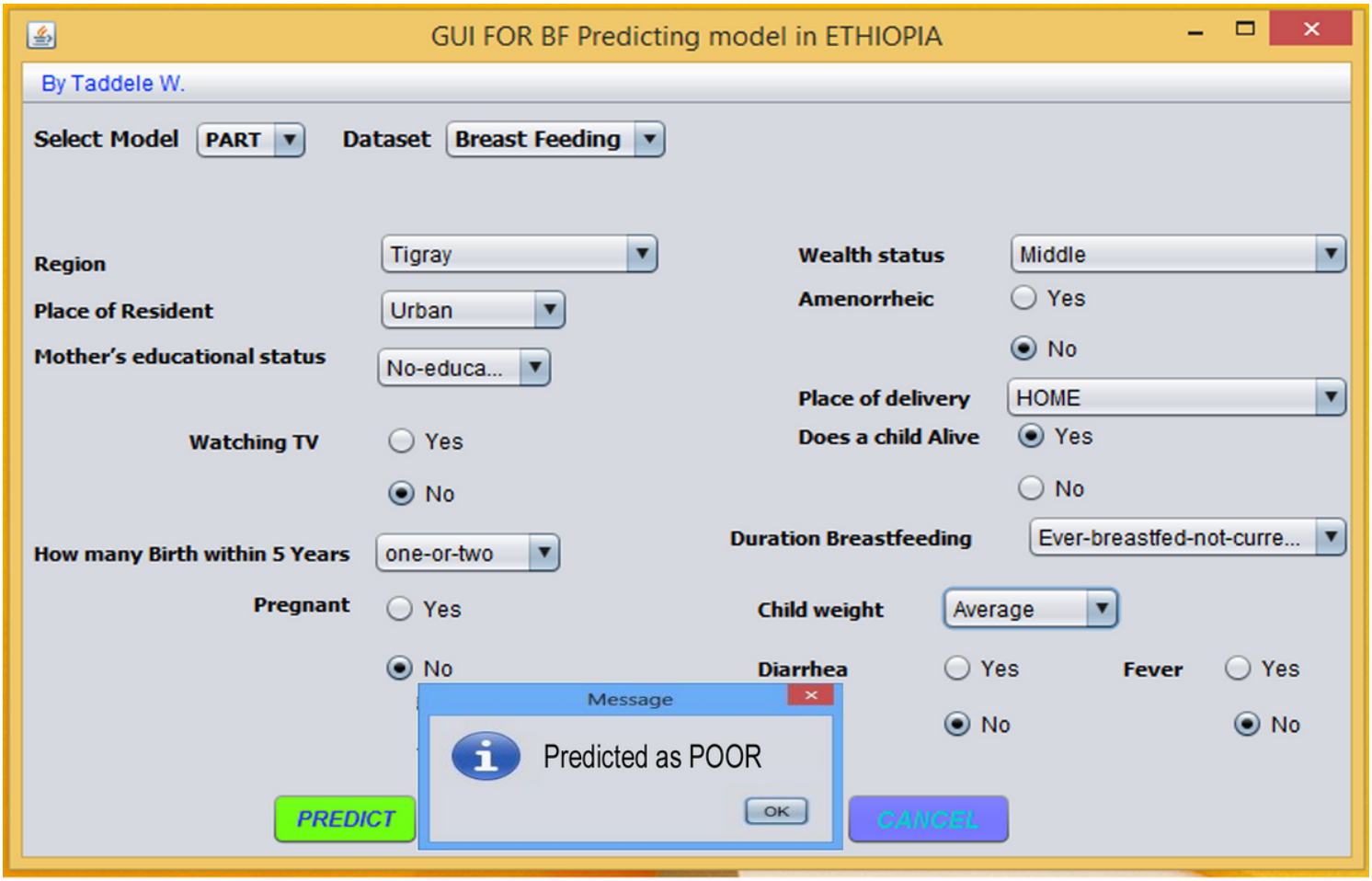

Figure 3:Model output predicting breastfeeding practice as POOR

Figure 3

Figure 3