Journal of
Health & Medical Informatics

Hailemariam et al., J Health Med Informat 2017, 8:3
DOI: 10.4172/2157-7420.1000274

Research Article OMICS International

Application of Data Mining to Predict the Likelihood of Contraceptive Method Use among Women Aged 15-49 Case of 2005 Demographic Health Survey Data Collected by Central Statistics Agency, Addis Ababa, Ethiopia

Tesfahun Hailemariam[1]*, Abraham Gebregiorgis[2], Million Meshesha[3] and Wubgzier Mekonnen[4]

[1]Department of Health Informatics, Hawassa Health Science College, Hawassa, Ethiopia
[2]Department of Health Informatics, College of Health Science, Mekelle University, Mekelle, Ethiopia
[3]Department of Information Science, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
[4]School of Community Health Department, Addis Ababa University, Addis Ababa, Ethiopia

Abstract

Introduction: In Ethiopia a gap between knowledge and use of contraceptive method is observed from many studies. According to the 2005 Ethiopian Demographic Health survey report the knowledge about any modern method among women is 86%, Contraceptive Acceptance Rate is 50.1% whereas the Contraceptive Prevalence Rate is 13.9%.

Methods: In order to find and interpret patterns from data the KDD process model is employed. This has gone through the steps of the process model; data selection and understanding, pre-processing, transformation, data mining, interpretation and evaluation. Decision tree and Naïve Bayes are used for the purpose of classification. The dataset used in this study is the 2005 demographic health survey data collected by central statistics agency. The techniques are tested both on the balanced and unbalanced datasets.

Results: Experimental results show that J48 decision tree performs better than Naïve Bayes. From this model 253 rules are generated. Overall accuracy of 82.85% a true positive (classifying non-user of contraceptive method) 87.3% and a true negative (classification of contraceptive method user) of 74.7% and a precision of 86.3%. One important rule detected was; women who do not know any contraceptive method have no any chance of using contraceptive method. But having knowledge of contraceptive method could not be a guarantee in order to use contraception. Other factors such as Partner occupation, Current marital status, wealth index, type of place were found to be most determinant factors as well.

Conclusion: Data mining techniques have revealed an important socioeconomic, demographic, geographic, reproductive history and knowledge factors associated with contraceptive method use. All concerned parties to strengthen the promotion of contraceptive method knowledge.

Keywords: Data mining; Family planning; Contraceptive method; Decision tree; Naïve bayes

Introduction

Family planning is a mechanism for limiting the size of the family and spacing the pregnancy through the use of either traditional or modern contraceptive method voluntarily. It allows parents to arrange and gain their desired number of children, to space and limit their births. Contraceptive method is one way which enables to attain the goal of family planning. Other services like treatment of infertility is also included under the service of family planning [1,2].

Spacing or limiting the number of pregnancy has an impact on the wellbeing of the mother, the child and on the economy of the family and the country as a whole. Maternal mortality, infant and child mortality, unwanted pregnancy, abortion and post abortion complications can be minimized by an effective use of contraceptive method. Use of methods like condom can as well prevent the transmission of HIV/AIDS and other sexually transmitted diseases [1,2].

In 1993 population policy of Ethiopia was then formulated to narrow the gap between economic development and population growth rate aimed at reducing fertility through utilization of family planning services and promoting socio-economic development to reduce the number of fertility which was 7.7 per each woman at that time to 4 and raise contraceptive prevalence rate to 44 % in the year 2015 [3,4]. The population of Ethiopia was increasing rapidly due to unplanned pregnancy and the low awareness of family planning [5]. In fact one of the main targets of family planning is to limit the growth of population so that the population of the country grows in line with growth of its economy and the available resources.

So many reasons lead for the formulation of the 1993 population policy of Ethiopia. It is shown in many countries that without limiting the growth of population it is difficult to achieve the millennium development goal, which strives to improve the living standard and life with respect to the universal strategy. It is also evident that the population growth has outpaced the economic growth. It is difficult to get developed if the population growth is not controlled. Being an agrarian country, with large population size and poor resource utilization and management the problem has worsened [4,6].

***Corresponding author:** Tesfahun Hailemariam, Department of Health Informatics, Hawassa Health Science College, Hawassa, Ethiopia, Tel: +251934107979; E-mail: tesfahunhailemariam@gmail.com

Received July 09, 2017; **Accepted** July 13, 2017; **Published** June 15, 2017

Citation: Hailemariam T, Gebregiorgis A, Meshesha M, Mekonnen W (2017) Application of Data Mining to Predict the Likelihood of Contraceptive Method Use among Women Aged 15-49 Case of 2005 Demographic Health Survey Data Collected by Central Statistics Agency, Addis Ababa, Ethiopia. J Health Med Informat 8: 274. doi: 10.4172/2157-7420.1000274

Family planning was introduced in Ethiopia during 1948. The services provided were very limited, reached only to large cities and the coverage was small. The coverage of family planning was approximately 17% in 1994 [5]. Modern FP service in Ethiopia is pioneered by The Family Guidance Association of Ethiopia (FGAE), which was established in 1966. Currently there are different health care providers which work in enhancing family planning services. Marie stops, path finder, engender health and governmental health care institutions are some of them.

Family planning enhances efforts to improve family health by limiting the number of family and spacing the pregnancy. However, traditional beliefs, religious barriers and lack of male involvement, low level of woman education with less empowerment, lack of knowledge, less income and other barriers have weakened family planning interventions. It is also confirmed that there is high unmet need for family planning in Ethiopia. When a woman needs to limit her family size or space the pregnancy but do not use any method to do this it is called unmet need [1].

According to EDHS and other studies the gap between the knowledge and use of contraceptive method is high. Women are believed to have high awareness of contraceptive method but researches [6-10] showed that the use of any method is still very low compared with the knowledge they have about any method. The knowledge about any modern method among women is 86%, Contraceptive Acceptance Rate is 50.1% whereas the Contraceptive Prevalence Rate is 13.9%.

Some of the Studies done in Ethiopia so far were used statistical analysis on a limited set of data to assess the factors which contribute to the low utilization of contraceptive method. Studies like that of EDHS on the other hand collect a vast amount of data in a population based and they employed a statistical analysis for analysing the data. While the traditional statistical analysis formulates a hypothesis and test the validity on the data set which are collected for that purpose. On the other hand data mining is applied on large dataset and its intention is not to test a hypothesis rather it tries to discover if there are hidden patterns and relationships from a large amount of data with no prior assumption. It also enables to generate knowledge and predict the likelihood of a certain phenomenon from a previous or historical data using techniques like decision tree, Naive Bayesian method and neural network, etc.

This study is one possible way of showing the application of data mining in health care data. The EDHS data is used for this purpose.

The main reason of this is because there is no organization which has captured much data related to respondent's background. Collecting relevant data that best describes the population with different variables is important during service provision in order to make analysis in the future and create a prediction model. As to the best knowledge of the researcher health care providers do not even ask why they quit and there is no way of knowing why they might stop coming. It is therefore the aim of this study is to apply data mining classification techniques to discover patterns that enable to differentiate the actual and non-actual contraceptive method users.

Towards solving the above-mentioned problem this study attempts to answer the following research questions:

- What are the most determinant attributes of contraceptive method use?
- What classification algorithms best predict the actual and non-actual contraceptive method users?

Methods/Data Mining Modelling

In order to find and interpret patterns from data the KDD process model is employed. This has gone through the steps of the process model; data selection and understanding, pre-processing, transformation, data mining, interpretation and evaluation. KDD is selected for three main reasons because KDD is best suited for academic purpose [9,10]. Reduces the skill required for knowledge discovery to the non-experts. Is independent from any tool and technique so one can use any technique during the study. The steps of KDD are shown in Figure 1.

Two data mining techniques; Decision tree and Naïve bayes algorithm are employed in this study. Sensitivity, specificity, accuracy and precision are used to compare the experiments conducted; number or leaves and size of tree are also considered in comparing the experiments conducted with J48 algorithm.

Data understanding and selection

The Demographic and Health Surveys (DHS) are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition. One of the DHS major topics is family planning which contains information of knowledge and use of contraceptive methods, both modern and traditional [11].

The 2005 EDHS data related to a woman age 15-49 is the target of this study. The women's questioner has 10 sections these are: Respondent's background, Reproduction, Contraception, Pregnancy, delivery, postnatal care and nutrition, Immunization, health, and women's nutrition, Marriage and sexual activity, Fertility preference, Husband's background and woman's work, HIV/AIDS and other

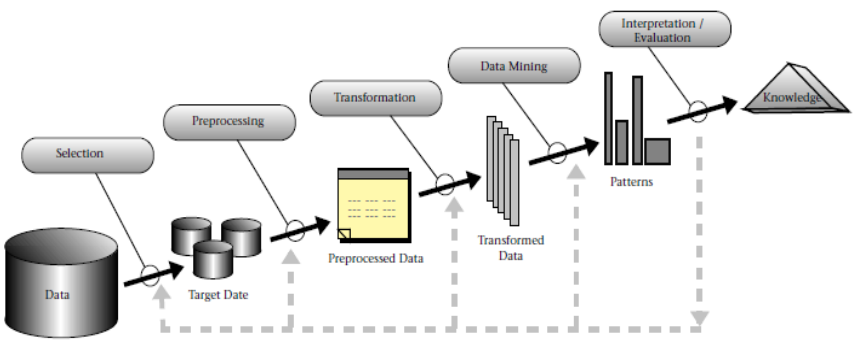

Figure 1: KDD process model steps.

sexually transmitted infections, Harmful traditional practices, Maternal mortality, Attributes which have no data mining value and those which are not relevant to the study subject (contraceptive method use) are removed or not selected.

Therefore the following attributes are selected Respondent's age, region, type of place, religion, marital status, education level, partner's education level, occupation, partner's occupation, number of living children, partner's age, partner approves FP, knowledge of FP, heard FP newspaper last month, heard FP on radio last month, heard FP on television last month, visited by FP worker during last 12 m, discuss FP with partner, wealth index.

Pre-processing

In order to have a good classification, prediction in general to achieve the goal of data mining we have to pre-process our data. Missing values are handled and the data was checked if there is noise, outlier and inconsistency and descriptive statistics was performed on the selected variables.

Transformation

The cardinality of some of the attribute are recoding to a manageable size. Discretization is used to convert those continuous values into discrete values the original attribute values are categorized into a certain interval labelled with a new representation and The value of each selected numeric attributes within the dataset are checked if there is any outlier and the box plot has shown no outlier.

Attribute selection

In order to select the best attributes from this initial collected dataset, the researcher evaluates the information content of the attributes using the select attribute technique of WEKA. WEKA has a built in attribute selection mechanism with a different options of evaluator and method. The dataset is then prepared in a format suitable for the Weka software i.e., in ARFF file format.

Experimentation and evaluation

Different experiments are conducted to investigate. The effect of test validation (10-fold cross validation and percentage split). The accuracy, size of tree, number of leaves of a decision tree by trying with the default value and by changing minimum number object to a different value. The effect of classification with unbalanced records and classification records that made to be balanced using SMOTE. The performance of classification with all variable and selected variable using best first. The effect of pruning on the decision tree accuracy, size of tree, and number of leaves compared with unpruned tree.

Experiment I: The first experimentation is performed with the J48 default parameters.

Experiment II: To make ease the process of generating rule sets or to make it more understandable with this objective, the min Num Obj (minimum number of instances in a leaf) parameter is tried with 5, 10, 15, 20 and 25.

Experiment III: This experiment is performed, by changing the default testing option (the 10-fold cross validation) to the percentage split to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieved a better classification accuracy.

Experiment IV: This experimentation the effect of pruned and un-pruned decision tree was observed using the default parameters of J48 Algorithm.

Experiment V: The fifth experiment is performed by first balancing the original data set using SMOTE. Then an experiment was conducted on this dataset with the J48 default values and adjusted values.

Experiment VI: The effect of attribute selection is also investigated with the default values of J48 algorithm and with adjusted parameter values.

Experiment VII: This experiment is performed using Naïve Bayes with10-fold cross validation.

Experiment VIII: Naïve Bayes also is performed using the percentage split of the default 66%.

Interpretation and evaluation of the discovered Knowledge

In order to select the best model the accuracy TP, TN and precision is considered and compared. In addition to this the number of leaves and size of tree are compared for the experiments conducted using J48 decision tree algorithm (Table 1).

In order to select the best model the accuracy TP, TN and (Figure 2) precision is considered and compared in the following graph. In addition to this the number of leaves and size of tree are compared for the experiments conducted using J48 decision tree algorithm. In this study the best performer was experiment V. From this model a set of rules are extracted simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node. However, the researcher selected best rules that cover most of the data points in the study. After the rule extraction, the researcher turns back to domain experts to discuss up on the generated rules.

Rule 1: If Knowledge of any method=No Then contraceptive method use: No (2494.0). Rule 1 shows that if a woman does not know any method there is a zero chance of using any method. Therefore it is recommended to strengthen the promotion of contraceptive method use.

Rue 2: If Knowledge of any method=Yes and partner occupation agric-employee and Current marital status=never married and Type of place=Rural and Num of living children=no child: Then contraceptive method use: No (2175.0/346.0). Another important observation was that knowledge about a family planning could not be the only reason which makes women to use any method of contraception. Even though a woman knows a method there are other factors which make them either to use or not to use. Knowledge of any method, Partner occupation, Current marital status, Partner's education level, Wealth index, Type of place, FP message, Number of living children, Religion, Respondent age are other factors that determine. Rule 2 shows that a rural women who are not married, have no child and their partner is agric-employee are less likely to use contraceptive method. This classified with an accuracy of 86.27%.

Rule 3: If Knowledge of any method=yes and partner occupation=agric-employee and Current marital status=never married and Type of place=urban: Then contraceptive method use: No 1247.0/17.0). Rule 2 and rule 3 shows that an urban women who are not married their partner is agric-employee are less likely to use contraceptive method. This classified with an accuracy of 98.65%.

Rule 4: If Knowledge of any method=yes and partner occupation=agric-employee and Current marital status=Not Living together then contraceptive method use: No (802.0/47.0).

Exp.	Model	NL	ST	Accuracy	TP rate	TN rate	Precision (Class No.)
I	J48-C 0.25-M 2 Test-mode=10-fold Dataset=Unbalanced Attribute=All	119	155	89.28%	0.97	0.327	91.3
II	J48-C 0.25-M 20 Test-mode=10-fold Dataset=Unbalanced Attribute=All	58	74	89.29%	0.972	0.312	91.12
III	J48-C 0.25-M 2 Test-mode=Split-66% Dataset=Unbalanced Attribute=All	119	155	89.33%	0.961	0.368	92.25
IV	J48-U-M 2 Test-mode=10-fold Dataset=Unbalanced Attribute=All	2073	2529	86.99	0.945	0.317	91.00
V	J48-C 0.25-M 20 Test-mode=10-fold Dataset=Balanced Attribute=All	253	319	82.85%	0.873	0.747	86.3
VI	J48-C 0.25-M 20 Test-mode=10-fold Dataset=Balanced Attribute=Selected 10	208	296	82.00%	0.885	0.701	84.40
VII	weka.classifiers.bayes.NaiveBayes Test mode=10-fold Dataset=Unbalanced Attribute=All	-	-	85.42%	0.897	0.538	93.45
VIII	weka.classifiers.bayes.NaiveBayes Test mode=Split 66% Dataset=Unbalanced Attribute=All	-	-	85.87%	0.897	0.553	93.99

Exp.: Experiment; NL: Number of Leaves; ST: Size of Tree; C: Confidence Level; M: Minimum Number of Instance per Leaf; U: Unpruned; TP: True Positive; TN: True Negative

Table 1: Summery of experimental result of J48 decision tree and naïve bayes.

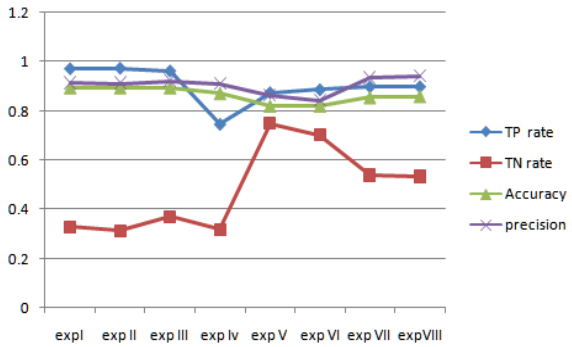

Figure 2: Comparison of the experiments with TP, TN, accuracy and precision.

Rule 5: If Knowledge of any method=yes and partner occupation=Non agric-employee and Current marital status=not living together then contraceptive method use: No (635.0/106.0). Another rule detected in rule 4 and 5 is whether a woman's partner is agric-employee or non agric-employee if they don't live together they are less likely to use contraceptive method.

Rule 6: If Knowledge of any method=yes and partner occupation=Non agric-employee and Current marital status=Living together and Wealth index=rich and Respondent age=20-24 then contraceptive method use: Yes (1102.0/170.0).

Rule 7: If Knowledge of any method=yes and partner occupation=Non agric-employee and Current marital status=Living together and Wealth index=rich and Respondent age=25-29 then contraceptive method use: Yes (1305.0/200.0).

Rule 8: If Knowledge of any method=yes and partner occupation=Non agric-employee and Current marital status=Living together and Wealth index=rich and Respondent age=30-34 then contraceptive method use: Yes (702.0/121.0). In Rule 6-8 it was also observed that women aged 24-34 who are rich and whose partner are a non agric-employee and living together have a high probability of using contraception.

Conclusion

The generated rules shown that knowledge about any CM is the most determinant variable, which is the top splitting variable of the model. Even though a woman knows a method there are other factors which make them either to use or not to use. It observed that rural women who are not married, have no child and their partner is agric-employee are less likely to use CM. It is also shown that an urban women

who are not married and their partner is agric-employee are less likely to use contraceptive method. Another rule detected was whether a woman's partner is agric-employee or non agric-employee if they don't live together they are less likely to use contraceptive method. It was also observed that women aged 24-34 who are rich and whose partner are a non agric-employee and living together have a high probability of using contraception. Data mining techniques have revealed an important socioeconomic, demographic, geographic, reproductive history and knowledge factors associated with contraceptive method use. The variables knowledge about method, partners' occupation, marital status and wealth index were found to be the most determinant attributes of contraceptive method users and non-users.

Recommendation

The researcher recommends the following points based on the outcome of the research. The techniques employed in this study were decision tree and Naïve Bayes algorithm. Even though an encouraging result was obtained, using other types of techniques with a different parameter might perform better therefore it is recommended to test with other types of techniques like Artificial neural network, support vector machine etc. Other important features which can make this study more interesting were not included in the family planning service providers. Data mining can have tremendous potential and benefits if healthcare providers able to capture, store, prepare and mine data. Therefore recording important variable such as partners' occupation, partners' education level etc. while providing service might help for decision making and other purposes.

The data used for this study was the DHS data; as observed family planning service providers collects few variables related to socio-economy, demography, geography, knowledge whereas CSA has collected so many information related to the above variables. It is the belief of the researcher those organizations should look for more variables. The possibility of integrating the discovered knowledge into knowledge based system would be helpful in assisting family planning service provider to identify the actual and non-actual users in priori based on the women socioeconomic, demography, geographic, knowledge and reproductive history etc. This study has attempted to apply DM techniques on contraceptive method data but it could also be applied in other health care data for decision making and other purposes.

Competing Interests

The authors declared that they have no competing interests.

Authors' Contributions

Abraham Gebregiorg is wrote the proposal, participated in data collection, analysed the data and drafted the paper. Dr. Million Mesheshaand Dr. Wubgzier Mekonnen approved the proposal with some revisions, participated in data collection and analysis, commented on the analysis and improved the first draft. All the three authors and Tesfahun Hailemariam revised subsequent drafts of the paper. Tesfahun Hailemariam prepared this manuscript for publication.

Acknowledgment

I am truly thankful to Dr. Million Meshesha, my advisor for his endless support and encouragement. It is really a pleasure to work with such a great person. I also would like to thank Dr. Wubgzier Mekonnen, my advisor for his support and guidance. I sincerely appreciate for the good communication I had with Meseret Ayanew. My sincere gratitude goes to my friends for the discussion I had with them and good comments we interchange.

References

1. http://www.who.int/mediacentre/factsheets/fs35/en/index.html
2. Emoh (2003) Family planning extension package. Ethiopia: Ababa, Ethiopia.
3. http://www.allbusiness.com/society-social/families-children-family/16272042-1.html
4. Onsembe J (2005) Ethiopia situation analysis on population, reproductive health and gender.
5. Ethiopian Society Population of Studies (2005) Levels, trends and determinants life time and desired fertility in Ethiopia: findings from EDHS 2005.
6. Aynalem A (2005) Population policy and projection.
7. Community Supported Agriculture (CSA) (2011) Ethiopia demographic and health survey.
8. Habtamu A (2007) The health extension program of Ethiopia: summery of concepts, progress, achievement and challenge.
9. Fayyad U, Piatetsky G, Smyth P (1996) Advances in knowledge discovery and data mining.
10. Lemaire V, Hue C, Bernier O, Vincent L, Carine H (2010) Correlation analysis in classifiers results for the contraceptive method choice data set.
11. http://www.csa.gov.et/