ADDIS ABABA UNIVERSITY

SCHOOL OF GRAGUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF HEALTH INFORMATICS

APPLICATION OF DATA MINIG TECHNOLOGY TO IDENTIFAY RISK FACTORS OF ABORTION INCIDENCE AND TO IDENTIFY THEIR ASSOCIATION RULES: THE CASE OF MARIE STOPS INTERNATIONAL ETHIOPIA CENTERS

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY
BINYAM AYELE ZELEKE
January, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF HEALTH INFORMATICS

APPLICATION OF DATA MINIG TECHNOLOGY TO IDENTIFAY RISK FACTORS OF ABORTION OCCURRENCE AND TO IDENTIFY THEIR ASSOCIATION RULES: THE CASE OF MARIE STOPS INTERNATIONAL ETHIOPIA CENTERS

By
BINYAM AYELE ZELEKE

Name and Signature of Members of the Examining Board

Dr Mesfin Addise, Advisor _____

Mr. Getachew Jemaneh, Advisor _____

Dr Gashaw kebede, Internal Examiner _____

Mr. Worku Tefera, External Examiner _____

DECLARATION

The thesis is my original, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

Binyame Ayele Zeleke

January 2013

The thesis has been submitted for examination with our approval as university advisors.

Dr Mesfin Addise

Ato getachew jemaneh

Table of Contents

.

LIST OF ACRONYMS AND ABBREVIATIONS

CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Value
DALY	disability-adjusted life years
EDHS	Ethiopia demographic health survey
FP/RH	Family planning or reproductive health
IQR	Inter Quartile Range
KDD	Knowledge Discovery in Database
KDP	Knowledge Discovery Process
MSI	Marie Stopes International
MSIE	Marie Stopes International Ethiopia
MMR	Maternal Mortality Ratio
MOH	ministry of health
NGO	non-governmental organization
RTI	reproductive tract infection
SPSS	Statistical Package for Social Science
WEKA	Waikato Environment for Knowledge Learning
WHO	World Health Organization

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Background: *In order to fill the gap in evidence based information, and help in programming for the reduction of maternal deaths due to unsafe abortion, Healthcare industry today generates huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, and medical devices. This large amount of data is a key resource to be analyzed and processed to extract hidden information and knowledge. Decision making process at the health care setting needs to be supported with more advanced technology including a computer based information system.*

Objective: *This thesis intends to investigate the potential applicability of data mining technology to identify the major factors that result in abortion and to find their association*

Methods: *A Hybrid Data Mining methodology is followed, which is a six-step knowledge discovery process. The data for this research obtained from MSIE in Addis Abeba, Ethiopia.*

The experiments carried out in this research using association mining algorithm apriori. On MSIE abortion report datasets, descriptive data summarization was taken to gain understanding of the data. Moreover, missing values, outliers data, data integration and transformation were managed at preprocess stage of hybrid process model.

On the basis of subjective (opinions of domain experts) and objective (support and confidence) measures of interestingness, a number of rules having practical relevance or that can add to the current knowledge in the problem domain were identified.

Results: *The results from this study were encouraging, which strengthened the hypothesis that interesting patterns can be generated from MSIE abortion case database by applying one of the data mining techniques: association rule mining. Besides, the results were promising and encouraging especially in the eye of domain experts.*

Conclusion: *The result thus obtained in this study is promising to apply data mining for identifying the risk factors of induced abortion and prevention. To make usable the knowledge extracted in this study, an attempt has made by selecting best association rules.*

Keywords: *Key words: Data mining, Induced abortion, knowledge discovery, association rule, apriori algorithm.*

CHAPTER ONE

INTRODUCTION

1.1 Background

Each year more than half a million maternal deaths occur due to preventable pregnancy related complications. WHO estimates that about 600,000 annual pregnancy-related maternal deaths occur worldwide, of this estimate an average of thirteen percent is due to unsafe abortion (1). WHO defined unsafe abortion as “a procedure for terminating an unwanted pregnancy either by persons lacking the necessary skills or in an environment lacking the minimal medical standards, or both” (1).

The developing world is disproportionately more affected than the developed. It is estimated that annually 2 million to 4.4 million abortions among adolescents occur in developing countries (2). According to hospital records of many developing countries between 38% and 68% of women treated for complications of abortion are under twenty years of age (3). Abortion is known to cause serious short term and longtime negative health consequences including death.

Looking at the Ethiopian context, Ethiopia is one of the countries in sub-Saharan Africa most severely affected by maternal deaths. Ethiopia has the fifth highest number of maternal deaths in the world: One in 27 women die from complications of pregnancy or childbirth annually (3). According to Ethiopian the MOH (4), in Ethiopia about 32% of all maternal deaths are the result of complications related to unsafe abortion. Abortion is the second leading cause of death for women, after tuberculosis (4).

In 2008, an estimated 382,500 induced abortions were performed in Ethiopia, for an annual rate of 23 abortions per 1,000 women aged 15–44. Public hospitals treat an average of 219 post abortion patients per facility per year. By comparison, there are about 50 cases of post abortion care per facility per year at public health centers and 84 at private and NGO facilities.(5)

The factors determining the frequency of induced abortions and the incidence of complications and deaths associated with them vary from place to place. Hence, it is not possible to define a single strategy that can be applied universally to deal with the problem. Noting that the demand for induced abortions is an indication of the frequency of unwanted pregnancies, the prevention of unplanned pregnancies is recognized as an important element in the strategy for reducing the frequency of abortions.

The primary purpose of health information is to improve health care delivery. Specifically, at the primary health care level, reliable and timely data about defined population groups can be used as a basis for effective policy formulation and implementation (6). For instance, data collected about abortion can serve to identify major causes and determinant risk factors of Abortion incidence and to take preventive actions to reduce the rate of maternal mortality and morbidity and to improve maternal survival.

However, in Ethiopia, the capacity to collect, compiles, analyze, interpret, and disseminate timely and accurate health information for decision-making is very poor. But in governmental and non-governmental healthcare service providers there are large volume clinical records i.e. abortion case records in Marie Stopes International Ethiopia (MSIE).

Marie Stopes International (MSI) delivers family planning and sexual healthcare to women throughout the world. In many countries that mean providing contraception, safe abortion and mother and baby care to the poorest and most vulnerable. MSI provides family planning and sexual healthcare to ordinary women in more than 40 countries around the world. (7)

One of the main family planning or reproductive health services providers in Ethiopia is MSIE, partner of MSI, a UK-based, non-governmental, not for profit organization working in the field of population, RH/FP. Its main goal is to prevent unplanned and unwanted pregnancies thereby ensuring the individual's fundamental human right to have “children by choice not by chance”. The ultimate goal of the organization is to improve the reproductive health of women and men. And According to MSIE statistics

on trends of FP and abortion provision, from the total of 2008 safe abortion services, MSIE provided 80,547 which account for 78% of all safe and legal abortion services in the country.(7)

Like MSIE, Ethiopian healthcare industry today generates huge amounts of data about patients, hospitals resources, disease diagnosis, electronic patient records, and medical devices. This large amount of data is a key resource to be analyzed and processed to extract hidden information and knowledge.

Identifying health-related problems of the community is one of the most important steps in the planning of health care interventions. However, appropriate planning, of health programs in turn depends to a large extent on access to timely and accurate information on demographic characteristics, on the occurrence of major health problems, and on associations with underlying factors.

Although the capabilities to collect and store data in large computer databases has increased significantly, the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data elements that are hidden or previously unknown Raghavan(8) on his part states that, health care data bases have accumulated large quantities of information about patients and their medical conditions; there are only few tools to evaluate and analyze the clinical data after it has been captured and stored. The author further stated that evaluation of stored data might lead to the discovery of trends and patterns hidden within the data that could significantly enhances our understanding of the health problem or the disease progression and management.

Recently, to address the problem of identifying useful information and knowledge to support primary healthcare prevention and control activities, health care institutions are employing the data mining approach which uses more flexible models, to discover unanticipated features from large volumes of data stored in clinical databases.

Particularly, in the developed world, data mining technology has enabled health care institutions to identify and search previously unknown, actionable information from large health care databases and to apply it to improve the quality and efficiency of primary health care prevention and control activities. However, to the knowledge of the researcher, there are few studies in Ethiopia has used this state of the art technology to support health care decision-making. (i.e.)Shegaw Anagaw (9) has applied neural network and decision tree models on data set containing 1100 records With the objective of exploring the possible application of data mining technology in health care data of Butajira, by developing a predictive model that could help health care providers to identify children at risk so that they can be treated before the condition escalates into something expensive and potentially fatal. It has proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities.

Data mining and knowledge discovery can also “close the loop” between clinical data mining capture and evidence-based decision support by facilitating the conversion of clinical data into evidence for future decision (10).

Although Wilcox et.al (11) argued that to evaluate and analyze data stored in large clinical data bases, new techniques and methods are needed to search large quantities of data to discover new patterns and relationships hidden in the data. It is due to challenges of searching for knowledge in clinical data bases and our inability to interpret and digest these data as readily as they accumulated, which has created a need for generation of tools and techniques for data base analysis. It is from this consequent that the discipline of knowledge discovery in data base (KDD), which deals with the study of such tools and techniques, has evolved into an important and active area of research (8)

1.2 STATEMENT OF THE PROBLEM

Unsafe abortion is a preventable tragedy and is one of the neglected problems of heath care in developing countries including Ethiopia. Annually, an estimated 25,000 women die of pregnancy and delivery complications in Ethiopia (12). And the maternal mortality ratio (MMR) is 676 deaths per 100,000 live births, one of the highest in the world (13). Unsafe abortion has a significant contribution to this high MMR.

To prevent maternal deaths related to abortion and to meet the development goals of 2015 the government of Ethiopia focus on improving the status of women through education, employment, and health. MOH now actively support and promote availability and distribution of family planning services through government facilities as well as private and marketing channels. Reproductive health care for the various groups of users including adolescents is also being integrated into the health system.

In 2005, Ethiopia expanded its abortion law, which had previously allowed the procedure only to save the life of a woman or protect her physical health. Abortion is now legal in Ethiopia in cases of rape, incest or fetal impairment. In addition, a woman can legally terminate a pregnancy if her life or her child's life is in danger, or if continuing the pregnancy or giving birth endangers her life. A woman may also terminate a pregnancy if she is unable to bring up the child, owing to her status as a minor or to a physical or mental infirmity. (14)

Notwithstanding the new law, almost six in 10 abortions in Ethiopia are unsafe. (14)

Even if there is improvement in FP coverage, the high maternal mortality ratio at 676 deaths per 100,000 live births indicates that access to and quality of emergency obstetric and neonatal care (EmONC) remains a challenge. (15)

The indirect costs of unsafe abortion are substantial, yet more difficult to quantify. They include the loss of productivity from abortion-related morbidity and mortality on women and household members; the effect on children's health and education if their mother dies; the diversion of scarce medical resources for treatment of abortion complications; and secondary infertility, stigma, and other socio-psychological consequences.

Abortion occurrence pattern understanding can be used by almost all sectors of the society at large. Mainly the government, medical sectors and steamed researchers to protect life and to improve the efficiency of operations and to plan a wide range of activities.

The risk factors associated to the problem may be many in number but some variables that make women to be exposed to abortion with high probability are considered to be main risk factors. Most research papers reported so far by health professionals show the occurrence patterns mostly in age and socio-economic group of the women's. However the association patterns of these attributes beyond these attributes have not been investigated.

The underlying research problem that necessitated this research is the existence `of high abortion occurrence and maternal morbidity and mortality at a national level in general and at the Addis Ababa area in particular. Specifically at the service provider clinics like MSIE large amount of clinical data are being collected. but the availability of this data analyzed with the traditional statistical data analysis alone is not enough for health professionals, planners and policy makers to identify major determinants of abortion occurrence and the associations of the those determinants.

On the other hand, the problem is that all those previously done studies on the health were conducted by using a very small proportion of the data bases. (8)Yet, in those studies, data analysis was conducted by using simple statistical techniques, such as regression and verification techniques (8). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world data bases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational data bases (2)

Particularly, in order to identify patterns in abortion occurrence and associated risk factors, it is difficult to conclusively attribute the problem to a certain set of factors since the parameters (attributes or factors) involved are too many. So studying the affinity of the factors rather than simply listing them is of crucial importance. On this point Last and Kandel (16) mentioned that the tools used in research on death causality have

been so far limited to the statistical techniques, like summarization, regression, analysis of variance, etc. however, such methods of data analysis may concentrate on identifying the leading causes of the health problem and listing the association of single factors with certain disease or health problem. No complex models involving more than one factor have been built.

Though MSIE has collected large volume of abortion-related data, including the data of abortion counseled and tested woman, the accumulated data is not as such fully utilized to support abortion occurrence prevention and control activities in the country. So, in such situation, new information technologies like data mining may be appropriate means in discovering hidden patterns in data.

Thus, in this study, the researcher has investigated how the attributes regarding the Socio- demographic characteristics, reproductive and contraceptive history, and health related factors determine abortion occurrence in the selected area by applying the new computational methods of data mining technology. And this study also has examined the association of these frontier factors detailing with what degree of confidence and support they co-exist on a woman that has high probability of seeking abortion.

The ultimate goal of this research is to assess how well a model can be built by using data mining techniques that could perform well in identifying the risk factors of abortion occurrence among the mass based only upon the personal, partner, educational, economical factors, etc., without using diagnostic tests or physical exam findings. And also the research had aimed at finding the degree which the influential factors are associated. This type of model might have application outside of clinical settings to support health workers in taking preventive actions to reduce unwanted pregnancy and abortion related problems as well as to assist health service planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at in the fights against unsafe abortion.

Thus, the result of this data mining process can be used to identify woman at high risk of abortion by looking at the probability and determining power of each attribute.

The hypothesis of this study is that if the existing data in MSIE clinical record is analyzed by using data mining techniques and tools, it can provide knowledge (beyond simply being a data) that can facilitate the improvement of task against unsafe abortion by identifying significant and meaningful patterns among different data elements which are hidden or previously unknown. Specifically, in relation to the problem domain of this study, if the right data included in the MSIE clinical record is selected, organized, cleaned, and analyzed using data mining tools and techniques it might show a result that could help to develop a model that identifies the risk of abortion among woman and to identify major determinants (factors that lead to probability) of abortion occurrence in Ethiopia together with the degree of co-occurrence.

Thus, this study investigate the potential applicability of data mining technology to identify risk factors of abortion incidence based up on clinical abortion case record gathered by Marie Stopes International Ethiopia (MSIE) centers in Addis Ababa. So this study is an attempt that can encourage the experts in the area to exploit the wide range of applications of data mining technology and techniques in various facets of the discipline.

1.3 APPLICATION OF THE STUDY

Even though the primary goal and initiative purpose of this research is academic exercise, the findings of this experimental research can ultimately be used in various areas.

In the first place, the government, MOH, can use the output of this research as a decision support means when setting policies on the issue of abortion incidence prevention and elimination.

Secondly, researchers in domain area can use the result of this study to conduct on abortion and related studies. Like for instance, if the study reveals that employment differences show differences in occurrence pattern, researchers may get some ground to step for further investigation.

MSIE as well as other similar organizations may use the intended model to know deeper about their data and contribution of each attributes of the customer that

determine the probability of abortion occurrence. Beyond simply knowing the factors, it enables also to measure degree of association of the attributes resulting in abortion. This helps to device different strategies for the efforts towards the struggle against abortion and its complications. In doing so they can improve the services they render to save and protect life of the majority.

Therefore, directly or indirectly, the society at large will benefit a lot from the final result of this experimental research.

1.4 Objectives of the Study

1.4.1 General Objective

To investigate the potential applicability of data mining technology to identify the major factors that result in abortion and to find their association rule.

1.4.2 Specific Objectives

In order to achieve the specified general objective, the study will have the following specific objectives:

- Collect data from the sources identified.
- Preprocess the data at hand which involves sampling, significance tests, and data cleaning, including checking the completeness of data records, removing or correcting noise and missing values, etc.
- Identify variables that affect the abortion occurrence by using appropriate data mining tool.
- Find association rules for the factors
- Evaluate and interpret the findings
- Report the result and make recommendations'

1.5 RESEARCH METHODOLOGY

A research methodology is an arrangement of condition for collocation and analysis of data in a manner that aim to address the research problem.

1.5.1 Research design

In order to achieve the objectives of this research, a Hybrid data mining process model were applied that consumes Knowledge Discovery Process (KDD) and Cross-industry Standard Process (CRISP-DM) models.

The methodology applied for this research was adopted from the techniques used by Cios, K., and Kurgan, L. 2005.The underlying reasons for Hybrid Data Mining models selection are

- It provides more general, research-oriented description of the steps
- It does emphasize the iterative aspects of the process, drawing experience form previous models. This enable the principal researcher to iterate when something wrong popup.
- It support academia and industrial data mining projects

The Hybrid DM (see figure 1.1) consists of six-step Knowledge Discovery Process (KDP); i.e. understanding the problem domain, understanding data, preparation of data, data mining, evaluation of the discovered knowledge, use of the discovered knowledge.

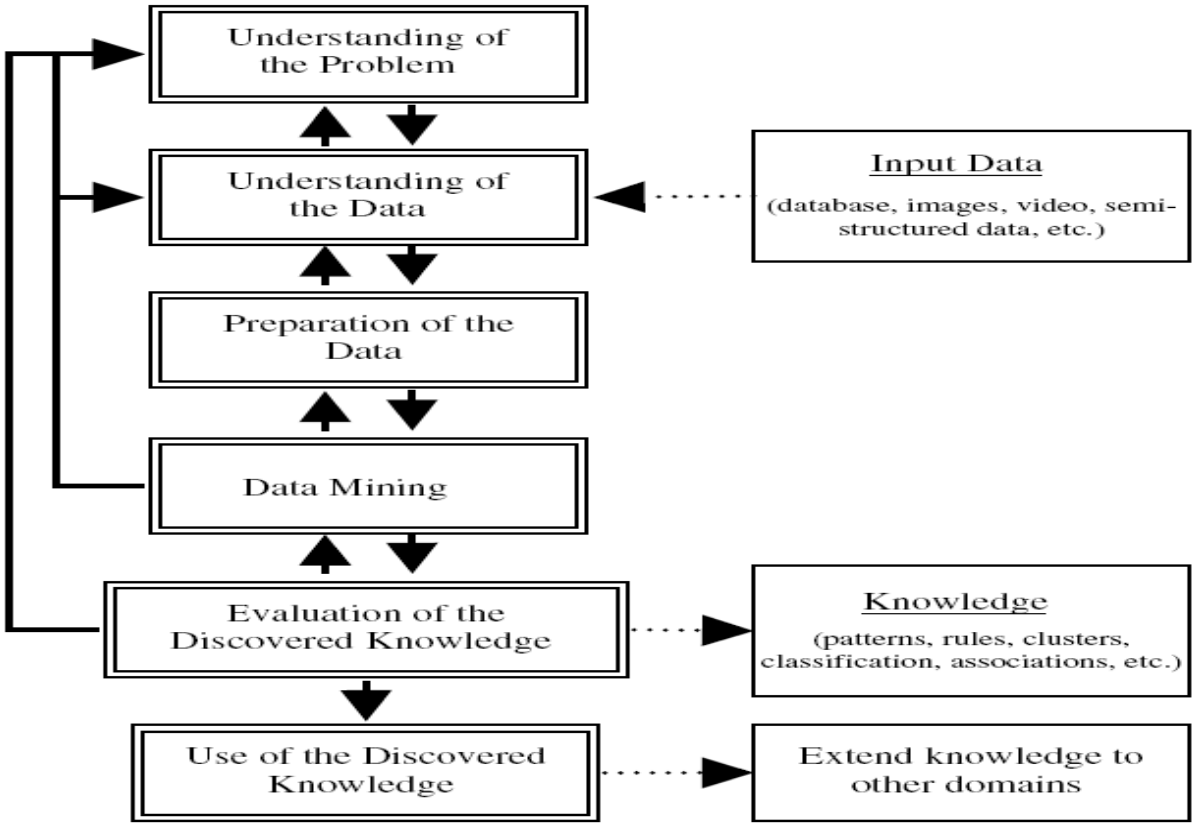

Figure 1.1 Hybrid Process Model

1.5.2 Understanding of the Problem

This stage of hybrid model involves working closely with domain experts to define the problem and corresponding solutions and determine the research goals. It also involves learning about domain-specific terminology, selection of DM tools to be used in the process. Therefore, literatures were consulted to learn about domain specific terminology and DM tools selection as well.

The study area was in Addis Ababa city administration, which is the capital city of Ethiopia. With Average Elevation of 2500m above sea level, the city administration has a geographic and territorial possession of an area of 540 sq. km and a total population of about 3 million [17].

MSIE is a member of the Marie Stopes International (MSI) global partnership and a branch office of MSI in Ethiopia. MSIE was established in May 1990 following a bilateral agreement signed with the Federal Ministry of Health. MSI is a UK-based global, non-profit and non-governmental organization committed to upholding the fundamental rights of women and couples to decide freely and without coercion, the number and spacing of their children. MSI has a global network of country programs in about 40 countries in Africa, Asia, and elsewhere.

MSIE works closely with the Government of Ethiopia and other partners to contribute to the national effort of reducing maternal mortality and morbidity by increasing access to and quality of SRH services. MSIE currently operates 30 clinics that provide comprehensive RH services with a focus on FP and safe abortion in the country. In Addis Ababa, there are six centers, which are managed by MSIE and provide comprehensive obstetric care in addition to all ranges of FP and safe abortion care.

1.5.3 Understanding of the Data

High quality data is a prerequisite for any data mining application. The sources of data for this research are result of abortion case records that were acquired from MSIE centers situated in Addis Abeba, Ethiopia.

Thus, for this research, a total of 5361 dataset utilized. The datasets for this study have a scale of measurement of numeric (2 attributes), and Nominal (5 attributes). This datasets partitioned and used for training the model. Finally, the selected Data was checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

1.5.4 Preparation of the Data

The datasets were undergoing data preparation steps to confirm for completeness, redundancy, missing values and plausibility of attribute. The collected data preprocessed and cleaned in a way to fulfill the requirement of data mining software. After the selection of relevant features/factors of abortion occurrenceattributes based on goal of the study, the next step was data preparation. In this step of data preparation, tasks like handling missing values, handling outliers' data, decoding of inconsistent data, and data reduction were under taken. Feature selection and extraction algorithms process handled to acquire cleaned data. The results are data that meet the specific input requirements for DM tools.

1.5.5 Data mining Techniques

Here the data miner uses the association Data Mining (DM) methods to derive knowledge from the preprocessed data. Among the available algorithms in WEKA machine learning software; apriori, were used in this research. These model were selected in this research due to their popularity in the recently published documents. The subsequent chapters will present a brief introduction to the two association rule algorithms and parameter setting of the model.

Two data mining tools have been used namely, SPSS and WEKA to implement the KDP. WEKA, formally called Waikato Environment for Knowledge Learning developed at the University of Waikato in New Zealand, is open-source data mining software in java. It provides implementations of learning algorithms that can be applied to a given

dataset and analyze its output to learn more about the data, and use learned models to generate association rules on instances [18].

1.5.6 Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, and interpretation of the results by domain experts. Only approved association rules are retained.

As it is often the case, with the application of the learning algorithm, several association rules were discovered from the dataset. Considering the large number of discovered rules, it was imperative to select only those rules that are interesting in relation to the purpose of the research- discovering major risk factors and their association of induced abortion that are relevant to the prevention and/or control of abortion occurrence.

To accomplish this task, two different types of measure of interestingness were adopted, namely objective and subjective.

The objective measures of interestingness of a rule used in the project include confidence and support measures. While Support of an association rule refers to the number of instances the rule predicts correctly, Confidence of an association rule refers to the same number expressed as a proportion of the number of instances that the rule applies to (19).

According to Liu et al. (20), however, although the objective measures of interestingness are useful in many respects, they often fail to capture all the complexities of the pattern discovery process. Accordingly, subjective measures of interestingness were also employed. These measures depend mainly on the user who examines the pattern. The use of subjective measures is considered even more important in many data mining application due to the fact that one can discover a large number of rules that are interesting “objectively” but of little interest to the user (19).Hence, the researcher together with domain experts attempted to determine the subjective interestingness of the discovered regularities based on knowledge about the problem domain.

1.5.7 Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. This last step determines the success of the entire knowledge discovery process. The results of this thesis work will be disseminated to the following stakeholders and to any interested parties. For this reason, interested readers can get access to the research results so as to support the decisions making process, or use it for further research in the area or for any other applicable reasons.

- It will be presented to school of information science and to school of public health.
- A hardcopy of this thesis results will be available to bibliographic library of information science.
- A softcopy of this thesis will upload to Addis Abeba University e-resource official website.
- Maximum effort will be exerted to publish the result in different journals. occurrence

1.6 Ethical Consideration

The research entitled "**APPLICATION OF DATA MINIG TECHNOLOGY TO IDENTIFY RISK FACTORS OF ABORTION OCCURRENCE AND TO IDENTIFY THEIR ASSOCIATION RULES: THE CASE OF MARIE STOPS INTERNATIONAL ETHIOPIA CENTERS** " acquired ethical clearance and approval from the department of Community Health, Faculty of Medicine, Addis Ababa University. The necessary verbal permission was obtained from MSIE Director. In addition, a letter of cooperation was prepared for communicating the concerned officials about the thesis from Faculty of Informatics. Furthermore, during analysis, a study subject's sensitive data was kept anonymous and confidential.

1.7 Scope and limitation of the problem

The scope of this research was limited to the abortion counseled and tested women of MSIE centers, where the required customer data was available. Even though the findings of this study can fairly be adopted as relevant to explore and implement the

potential applicability of data mining tools in other centers of similar service, related to abortion, this research is limited to the case of MSIE.

Furthermore, the study was limited to the abortion case clinical data of MSIE in investigating the contribution of each attribute resulting abortion and finding their association rule. This pattern can be used to understand the existing data.

1.8 Organization of the Research

This thesis contains six chapters. The first chapter deals with the general overview of the study including background, statement of the problem, research objectives and methodology of the research. The second chapter is devoted to literature review of data mining technology, abortion and its impact, and review of related literature.

Chapter Three is about the data mining algorithms that are used to generate association rules based on abortion case clinical records of MSIE. Chapter Four is devoted to Business Understanding, Data understanding and Data preprocessing of the data for generating good quality datasets for generating the association rules task.

Chapter Five reports the experiment of the research. It comprises training, and building the models. Results of the experiment were also analyzed and interpreted. The last, presents chapter six, concluding remarks and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge, data mining has attracted a great deal of attention in the information industry and in society as a whole. This huge volume of data can be accumulated beyond database and data warehouses. The abundance of data, coupled with the need for powerful data analysis tools, has been described as a "data rich but information poor situation".

Now what is the use of all this data? Up to the early 1990's the answer to this was "NOT much". No one was really interested in utilizing data, which was accumulated during the process of daily activities. (21)

As a result, data collected in large data repositories become "data tombs"—data archives that are seldom visited. Consequently, important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In addition, consider expert system technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. (31)

Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Data mining refers to extracting or "mining" knowledge from large amounts of data [22].

In fields of business, science, and engineering a need to understand large, complex, information-rich data sets is common to all. A method to extract useful knowledge hidden in these data and to act on the knowledge is becoming increasingly important in

today's competitive world. The entire process of applying a computer-based methodology for discovering knowledge from data is called data mining [22].

2.1 Multidisciplinary Nature of Data Mining

Han et al [19] mentioned, data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing.

He also, explained that because of the diversity of disciplines contributing to data mining, data mining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems, which may help potential users distinguish between such systems and identify those that best match their needs.

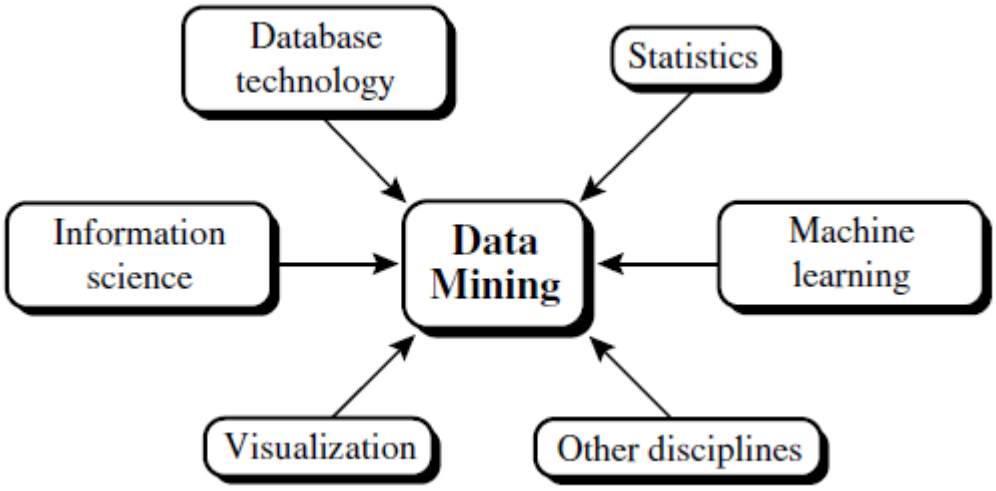

Figure 2.1 Data mining a confluence of multiple disciplines

2.2 Data Mining Techniques

The data mining goals are defined by the intended use of the system. The two high-level primary goals of data mining in practice tend to be prediction and description. Prediction involves using some variables or fields in the data base to predict unknown or future values of other variables of interest. In other words predictive mining tasks perform inference on the current data in order to make prediction. Descriptive mining focuses on finding human-interpretable patterns describing the data (22).

In predictive models, the values or classes we are predicting are called the response, dependent or target variables. The values used to make the prediction are called the predictor or independent variables. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. On the other hand, descriptive techniques are sometimes referred to as unsupervised learning because there is no already known result to guide the algorithms (23).

As Levin and Zehavi (24) stated, descriptive models interrogate the data base to identify patterns and relationships in the data. As Han and Kamber (19) states, users may sometimes have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations of applications. The goals or functions of prediction and description can be achieved using a variety of particular data-mining methods (22).

Clustering algorithms, pattern recognition models, visualization methods, and link analysis are the major members of descriptive models (24).

2.2.1 Descriptive Models

2.2.1.1 Clustering Algorithms

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It is mapping a data item into one of several clusters which are not pre-specified but are determined from the data. Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures (24).

Two Crows Corporation (23) mentioned that the goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. The categories (clusters) can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories. According to Han and Kamber (19), each cluster that is formed can be viewed as a class of objects from which rules can be derived.

Unlike classification, we don't know what the clusters will be when we start, or by which attributes the data will be clustered. In general, the class labels are not present in the training data simply because they are not known to begin with. Consequently, experts' knowledge is required to interpret the clusters (23).

The most common of all automatic clustering algorithms is the K-means algorithm which assigns observations to one of K classes to minimize the within-cluster-sum-of-squares. Another class of models is the self-organizing neural network models.

2.2.1.2 Link Analysis

Link analysis is a descriptive approach to exploring data that can help identify relationships among values in a database. The two most common approaches to link analysis are association discovery and sequence discovery (23).

2.2.1.2.1 Association Discovery

Since its introduction in 1993, the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in knowledge discovery (25).
The aim of association rule mining is to detect interesting associations between items in a database. It was initially proposed in the context of market basket analysis in transaction databases, and has been extended to solve many other problems.

Association mining used to determine which things often go together. Association rule mining finds interesting association or correlation relationships among a large set of data items.

With massive amounts of data continuously being collected and stored in databases, many industries are becoming interested in mining association rules from their databases. For example, the discovery of interesting association relationships among huge amounts of business transaction records can help catalog design, cross-marketing, loss-leader analysis, and other business decision making processes.

Association approaches are most common in market basket analysis and recently, beside market basket analysis association analysis also applicable to other application domains such as bioinformatics, medical diagnosis, web mining and scientific data analysis (19).

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. The rules are given in the form: if item A is part of an event, then X% of the time item B is also part of the event. The rules are written as A=>B, where A is called the antecedent or left-hand side (LHS), and B is called the consequent or right Hand side (RHS) (25).

More formally, association rules are of the form A==>B, that is:

(A1,----,Am → B1,---,Bn), where
Ai (for i e {1,----,m} and Bj (for j e {1,----,n}) are attribute-value pairs.

The associations rule A==>B is interpreted as database tuples that satisfy the condition in A are also likely to satisfy the condition in B. Association rule mining searches for interesting relationships among objects in a given data set.

2.2.1.2.2 Sequence Discovery

Sequence discoveries are association rules with time dimensions. A sequential pattern is an association between sets of items, in which some temporal properties between items in each set and between sets are satisfied. In particular, items in a set have the same temporal reference (24). As Trybula (26) states, sequential patterns are identified in a technique for predicting future activities based on observing trends over a period of time. It is based on the fact that previous activities have the potential for indicating future activities.

Two Crows Corporation (23) point out that association or sequence rules are not really rules, but rather descriptions of relationships in a particular database. There is no formal testing of models on other data to increase the predictive power of these rules. Rather there is an implicit assumption that the past behavior will continue in the future.

2.2.2 Predictive Models

In predictive modeling one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models: classification models, regression models and AI-based models (24).

2.2.2.1 Classification

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from historical database. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. Sometimes an expert classifies a sample of the

database, and this classification is then used to create the model which will be applied to the entire database (23).

2.2.2.2 Regression

Regression uses existing values to forecast what other values will be. This method can be used to define the boundary condition by evaluating the data and determining the boundary through mathematical analysis (26). In the simplest case, regression uses standard statistical techniques such as linear regression. The linear regression method is used for modeling continuous response. Unfortunately, many real world problems are not simply linear projections of previous values. Therefore, more complex techniques, such as logistic regression, decision trees, or neural nets, may be necessary to forecast future values (23).

2.2.2.2.1 Time Series Regression

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time (23).

2.2.2.2.2 AI Based Models

The leading models in this category are Neural Networks (NN) models. NN is a biologically inspired model which tries to mimic the performance of the network or neurons, or nerve cells, in the human brain. Expressed mathematically, a NN model is made up of a collection of processing units (neurons, nodes), connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. A typical NN contains several input nodes connected to one or more output nodes, through an intermediate set of hidden nodes. NN have become of particular interest in data mining because they offer a means for efficiently modeling large and complex problems in which there are hundreds of independent variables that have many interactions.

Pattern-finding mechanism in data mining is data-driven rather than user-driven. The relationships are found inductively by the software itself based on the existing data. It does not require the user or modeler to specify the functional form and interactions. No one model or algorithm can or should be used exclusively. For any given problem, the nature of the data itself will affect the choice of models and algorithms. There is no best model or algorithm. Consequently, we as model developer will need a variety of tools and technologies in order to find the best possible model.

2.3 Data Mining Functionalities

2.3.1 Characterization & Discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. According to Han et al [10], these concepts and class descriptions can be derived using data characterization and data discrimination. Data characterization is a summarization of the general characteristics or features of a target class of data and data discrimination, by comparison of the target class with one or a set of comparative classes.

2.3.2 Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data whose data objects whose class label is known. David Hand [8] argues data mining model is data-driven and the discovery of a highly predictive model should not be taken to mean that there is a causal relationship. For example, an analysis of customer records may show that customers who buy high-quality wines are also more likely to buy designer clothes; in this case clearly one's tendency is not causally related to the other propensity. The derived model may be represented in various forms, such as *classification (IF-THEN) rules*, *decision trees*, *mathematical formulae*, or *neural networks [19].*

In a predictive model, one of the variables is expressed as a function of the others. This permits the value of the *response* variable to be predicted from given values of the *explanatory* or *predictor* variables (19).

2.3.3 Frequent Patterns, Associations, and Correlations

Frequent patterns, as the name implies, are patterns that occur frequently in data. It discusses many kinds of frequent patterns, subsequences and substructures (10). A frequent item set typically refers to a set of items that frequently appear together in a transactional data set. And a frequent occurring subsequence, such as the pattern that customers tend to purchase first a personal computer(PC), followed by a digital camera, and then memory card. Such phenomena described as frequent subsequence.

Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

2.4 Data mining application in health care data

A number of articles published in the health care literature revealed the practical application of data mining techniques in the analysis of health information.

Current trends in medical decision making show awareness of the need to introduce formal reasoning, as well as intelligent data analysis techniques in the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records (27). According to Larvac, machine-learning methods have been applied to a variety of medical domains in order to improve medical decision-making (27).

The reminder of this section attempts to review some empirical studies related to the practical application of data mining technology in the health care sector.

In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficiency of patient care. Data mining can be used to help predict future patient behavior and to

improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today so they do not become the problems of tomorrow (27).

Prather, et. al. (28) conducted a data mining project at Duck University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to prenatal outcomes. The goal of this knowledge discovery effort was to identify factors that will improve the quality and cost effectiveness of prenatal care.

Lavrac et al. (27) combined GIS and data mining to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that “data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making.”

Downs and Wallace (9) have also applied data mining techniques to mine association rules from a pediatric primary care decision support system. According to the authors, the purpose of their study was to apply an unsupervised data-mining algorithm to a database containing data collected at the point of care for clinical decision support.

Shegaw Anagaw (8) has applied neural network and decision tree models on data set containing 1100 records With the objective of exploring the possible application of data mining technology in health care data of Butajira, by developing a predictive model that could help health care providers to identify children at risk so that they can be treated before the condition escalates into something expensive and potentially fatal. It has proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities. He has recommended that the encouraging results obtained from both neural networks and decision trees indicate that data mining is really a technology that should be considered to support child health care prevention

and control activities at the district of Butajira in particular, and at a national level in general. Specifically, such models could be used in settings outside of the hospital to support primary health care prevention activities and health service programs, which are aimed to reduce infant and child mortality (8).

2.5 Abortion and Current Research Outputs

2.5.1 Global overview of abortion

Methods to terminate an unwanted or unintended pregnancy are known to have existed since ancient times. As far back as 5000 years ago, the Chinese Emperor Shen Nung described the use of mercury for inducing abortion (29). A recent publication, (30) lists over 100 traditional methods of inducing abortion, which can be broadly classified into four categories: 1) oral and injectable medicines; 2) vaginal preparations; 3) introduction of a foreign body into the uterus; and 4) trauma to the abdomen. Many of these methods pose serious threats to the woman's life and well-being.

Each year, throughout the world, approximately 210 million women become pregnant and some 130 million of them go on to deliver live-born infants. The remaining 80 million pregnancies (31) end in stillbirth, or spontaneous or induced abortion. Approximately 42 million pregnancies are voluntarily terminated each year – 22 million within the national legal system and 20 million outside it. In the latter case, the abortions are often performed by unskilled providers or in unhygienic conditions, or both.

A term "Abortion" is defined as the termination of pregnancy by the removal or expulsion from the uterus of a fetus or embryo prior to viability. An abortion can occur spontaneously, in which case it is usually called a miscarriage or it can be purposely induced. The term *abortion* most commonly refers to the induced abortion of a human pregnancy (1).

Abortion is a sensitive and contentious issue with religious, moral, cultural, and political dimensions. It is also a public health concern in many parts of the world. More than one-quarter of the world's people live in countries where the procedure is prohibited or permitted only to save the woman's life. Yet, regardless of legal status, abortions still

occur, and nearly half of them are performed by an unskilled practitioner or in less than sanitary conditions, or both (32).

Abortions performed under unsafe conditions claim the lives of tens of thousands of women around the world every year, leave many times that number with chronic and often irreversible health problems, and drain the resources of public health systems. Often, however, controversy overshadows the public health impact (32).

Induced abortions may be performed either within the law, or outside the national legal framework. When induced abortion is performed by qualified persons using correct techniques and insanitary conditions, it is a safe surgical procedure. In the USA, for example, the death rate from induced abortion is now 0.6 per 100 000 procedures, making it as safe as an injection of penicillin (33).

In developing countries, however, the risk of death following unsafe abortion may be several hundred times higher. Spontaneous abortion is rarely fatal and seldom presents complications. The mortality and morbidity risks associated with unsafe induced abortion depend on the facilities and the skill of the abortion provider, the method used and the general health of the woman and the stage of her pregnancy.

After declining substantially between 1995 and 2003, the worldwide abortion rate stalled between 2003 and 2008. Between 1995 and 2003, the abortion rate (the number of abortions per 1,000 women of childbearing age—i.e., those aged 15–44) for the world overall dropped from 35 to 29. It remained virtually unchanged, at 28, in 2008. Nearly half of all abortions worldwide are unsafe, and nearly all unsafe abortions (98%) occur in developing countries (34).

In the developing world, 56% of all abortions are unsafe, compared with just 6% in the developed world. The proportion of abortions worldwide that take place in the developing world increased between 1995 and 2008 from 78% to 86%, in part because the proportion of all women who live in the developing world increased during this period (34).

2.5.2 Global Abortion Circumstances

Of the estimated 42 million induced abortions each year, nearly 20 million are performed in unsafe conditions and/or by unskilled providers and result in the deaths of an estimated 47,000 girls and women. This represents about 13 percent of all pregnancy-related deaths. Almost all unsafe abortions take place in developing countries, and this is where 98 percent of abortion-related deaths occur (21).

A review of the combined impact of mortality and morbidity due to unsafe abortion estimated that, every year, there are 65 000 to 70 000 deaths and close to five million women with temporary or permanent disability due to unsafe abortion. Of these, more than 3 million suffer from the effects of reproductive tract infection (RTI), and almost 1.7 million will develop secondary infertility. Unsafe abortion accounts for 13% of maternal deaths, and 20 % of the total mortality and disability burden due to pregnancy and childbirth, in terms of disability-adjusted life years (DALYs) (35). Altogether some 24 million women currently suffer secondary infertility caused by an unsafe abortion.

The developing world is disproportionately more affected than the developed. It is estimated that annually 2 million to 4.4 million abortions among adolescents occur in developing countries (2). According to hospital records of many developing countries between 38% and 68% of women treated for complications of abortion are under twenty years of age (3). Abortion is known to cause serious short term and longtime negative health consequences including death.

In DALYs, the combined burden of mortality and morbidity per 1000 unsafe abortions is exceptionally high in sub-Saharan Africa, where it is 50 percentage points higher than in Asia and 6 times greater than in Latin America(35).

Deaths due to unsafe abortion remain close to 13% of all maternal deaths. Unsafe abortion related deaths have, however, reduced to 47 000 in 2008 from 56 000 in 2003 and 69 000 in 1990; corresponding to the decline in the overall number of maternal deaths to 358 000 in 2008 from 546 000 in 1990. Although unsafe abortions are preventable, they continue to pose undue risks to women's health and lives (36).

An estimated 21.6 million unsafe abortions took place worldwide in 2008, almost all in developing countries. Numbers of unsafe abortions have increased from 19.7 million in 2003. Although the overall unsafe abortion rate remains unchanged at about 14 unsafe abortions per 1000 women aged 15–44 years (2).

In developing countries, two in five unsafe abortions occur among women under age 25, and about one in seven women who have unsafe abortions is under 20.8 In Africa, about one-quarter of the unsafe abortions are among teenagers (ages 15 to 19), a higher proportion than in any other world region. And WHO reports that in the countries of sub-Saharan Africa unsafe abortions are responsible for as much as 50 percent of maternal deaths (32).

One recent study estimated that every year in developing countries five million women are admitted to hospital as a result of unsafe abortion. The treatment of abortion complications in hospital consumes a significant share of resources, including hospital beds, blood supply, medications, and often operating theatres, anesthesia and medical specialists. Thus, the consequences of unsafe abortion place great demands on the scarce clinical, material and financial resources of hospitals in many developing countries,(37) undoubtedly compromising other maternity and emergency services (31). Major physiological, financial and emotional costs are also incurred by the women who undergo unsafe abortion (37).

Unsafe abortion has significant negative consequences beyond its immediate effects on women's health. For example, complications from unsafe abortion may reduce women's productivity, increasing the economic burden on poor families; cause maternal deaths that leave children motherless; cause long-term health problems, such as infertility; and result in considerable costs to already struggling public health systems (13).

The indirect costs of unsafe abortion are substantial, yet more difficult to quantify. They include the loss of productivity from abortion-related morbidity and mortality on women and household members; the effect on children's health and education if their mother dies; the diversion of scarce medical resources for treatment of abortion complications; and secondary infertility, stigma, and other socio-psychological consequences (37).

2.6 Abortion in Ethiopia

Ethiopia is the second most populous country in sub-Saharan Africa with an estimated population of 74 million in 2009 (5). According to the 2005 EDHS data, the total fertility rate for Ethiopia is 5.4 births per woman. Over 50% of its population is younger than 20 years and over 50% of adults are illiterate. Sexual debut occurs on the average at the age of 20 for males and 16 for females with the median age of marriage for girls in Ethiopia at 16.1 years. About 40% of young women have their first child by 19 years and 54% of pregnancies to girls under the age of 15 are unwanted.

In 2008, 101 unintended pregnancies occurred per 1,000 women aged 15–44, and 42% of all pregnancies were unintended. This high level of unintended pregnancy is being the root cause of abortion (5).

Safe abortion services have been unavailable throughout much of Ethiopia's modern history. The 1957 penal code allowed abortions only to save the life or health of the woman. Combined with low rates of contraceptive supplies, use, and high rates of sexual violence, the restrictive law compelled many Ethiopian women to seek out the services of unskilled, back-street abortion providers.

However, in 2004, the Ethiopian Parliament passed one of Africa's most progressive abortion laws. The new penal code added indications for rape, incest, fetal abnormality, and a woman's physical or mental disabilities. The Parliament also approved abortion for minors physically or psychologically unable to care for a child. This marks a significant change for Ethiopia, where adolescents make up more than 45 percent of those seeking abortions (4).

In 2008, an estimated 382,500 induced abortions were performed in Ethiopia, for an annual rate of 23 abortions per 1,000 women aged 15–44 (5).

According to the 2000 and 2005DHS surveys, the level of unintended pregnancy in Ethiopia is high and may be increasing. The desire for smaller families is increasing, which reflects broader social and economic changes in the country: The average desired family size declined from 4.9 in 2000 to 4.0 in 2005 (15).

And although contraceptive use has increased, unmet need for contraception has remained high. Women in Addis Ababa and other urban areas are delaying marriage into their 20s, probably in response to adverse economic conditions. This delay in marriage may result in increased sexual activity among unmarried young women, raising their risk of unintended pregnancy, as well as abortion, given that childbearing outside of marriage is highly stigmatized. In DHS surveys, few unmarried women report ever having been sexually active, but data from other, smaller scale studies suggest that sexual activity among the unmarried is not uncommon (15).

The abortion rate is considerably higher than the national average in urban areas: 49 per 1,000 in Addis Ababa, the country's most urban and economically developed region, and 184 per 1,000 in the smaller urban regions of Dire Dawa and Harari. The high abortion rates in these urban areas are likely the result of many factors, including that the availability of private health care providers in these commercial centers draws women from surrounding areas (5).
Notwithstanding the new law, almost six in 10 abortions in Ethiopia are unsafe (14).

A study conducted by Goodman et al (43) on Implementation of Comprehensive Abortion Care in Ethiopia stated that in Ethiopia abortion is resulted from the deep rooted poverty, gender inequalities and lack of commitment of responsible actors to ensuring women rights to safe abortion. In addition, limited awareness of both clients and service providers on the revised 2005 Criminal Code of the Federal Democratic Republic of Ethiopia (penal code) is also one of the major obstacles that hindered women from attaining Comprehensive Abortion Care (43).

About half of all health facilities in Ethiopia provide induced abortion services. However, the proportion is much higher for public hospitals (76%) and private or nongovernmental organization (NGO) facilities (63%) than for public health centers (41%). These proportions are likely changing rapidly, as efforts are being made to expand abortion services in public facilities. Currently, private and NGO facilities provide the most induced abortions (5).

Better availability of safe abortion services can be partly achieved by significantly scaling up the delivery of medical abortion, which currently represents only a small percentage of the total abortions provided. In Ethiopia, MSIE is currently the only provider of medical abortion (Mifepristone/Misoprostol) as most health facilities do not have the staff trained or the supplies required to provide this service (7).

And According to MSIE statistics, from the total of 2008 safe abortion services, MSIE provided 80,547 which account for 78% of all safe and legal abortion services in the country (7).

2.6.1 Impact of Abortion in Ethiopia

The consequences of unsafe abortion place great demands on the scarce clinical, material and financial resources of hospitals in many developing countries, compromising other maternity and emergency services. Major physiological, financial and emotional costs are also incurred by the women who undergo unsafe abortion.

In Ethiopia, one in seven women die from pregnancy-related causes, and unsafe abortion causes more than half of the 20,000 maternal deaths that occur annually in the country (1). Abortion with sepsis, a toxic and often fatal blood condition, is the sixth-leading cause of hospital admissions for Ethiopian women and girls.

Prior studies have documented that unsafe abortion has been an important and ongoing health problem in Ethiopia. The 2005 Ethiopian Demographic and Health Survey (DHS) estimates that 673 women died of pregnancy-related causes for every 100,000 live births in the six years prior to the survey(39). In a 2001–2002 study in a major university hospital in Addis Ababa, post-abortion complications were one of the three leading causes of maternal mortality (38). According to a large-scale study in 2000 of hospitals in nine of the country's 11 regions, more than half of women treated for complications of induced abortion had gone to an untrained provider or had induced the abortion themselves(40). A bibliographic review spanning 1985–2000 indicates that many attitudinal and organizational barriers prevented women from obtaining post-abortion services without delay and that these barriers resulted in low-quality post-abortion services (41).

The work of Selamawit Negash (37) showed that, in Addis Ababa, the cost in provision of safe abortion service was estimated and the average unit cost ranges from 40.97 to 65.32 birr while the actual average unit cost treating a patient with abortion complication was 131.7 birr with-out including the patient side cost. Moreover the sensitivity analysis showed that the cost of treating complication of abortion could rise up to 323.23 birr. Thus, the health care system is spending a lot of resources for treating complications of abortion, which could be possible to prevent it through provision of safe abortion services. The average patient side cost which includes medical, non-medical and opportunity cost was found to be 535.5 birr (37).

2.7 Review of Related Literature

Even if they are not done by using data mining and DM algorithms, by using other statistical methods numerous works in abortion related researches have been conducted. Specifically to identify major factors or reasons of abortion occurrence and their association among those attributes.

A survey research was conducted to identify the impact of selected socio-demographic characteristics on induced abortion conducted by Elias Senbeto*et al (42)* by using both the classical bivariate methods and the multivariate logistic regression technique. Accordingly, the socio-demographic variables considered in the bivariate analysis were: age, place of residence, religion, occupation, marital status, educational status, contraceptive use and number of pregnancies of the responding subjects. With the exception of religion, all other variables showed significant associations with induced abortion.

And also, with the increase in age and number of pregnancies, there was a decrease in the number of mothers who had induced abortion (P <0.01 for each of the above two factors). On the other hand, as the level of education of the study subjects increased, there was an increase in the number of mothers who had abortions accordingly. In particular, among the total responding subjects, those who had a high school (or above) education were highly exposed to the risk of induced abortion with an odds ratio of 10.6

compared to illiterate women who could not read and write. Women living in urban centers were 3.5 times higher in having induced abortion as compared to those living in rural areas (P < .001). It was also observed from these bivariate analyses that subjects who ever-practiced contraception were at a higher risk of acquiring the problem of induced abortion.

The study also stated that, single women and students were 14.6 and 13.4 times higher in performing (having) induced abortions compared to married women and housewives respectively. In this bivariate analysis, religion did not show a significant association with induced abortion (42).

Two hundred fifty six women (19%) had abortions and the prevalence rates of spontaneous and induced abortion were computed as 14.3% and 4.8%, respectively. A total of 573 (42.6%) women reported to be current users of contraceptives. Among the determinant factors included in the multivariate logistic regression model, place of residence, marital status, contraceptive use, number of pregnancies and level of education attained by the women were found to be significantly and independently associated with induced abortion (P < .05 for each factor) (14).

Another research called Characteristics of women seeking abortion-related care in Addis Ababa, Ethiopia by YilmaMelkamu et al (43) Safe termination clients were younger, educated, and more often employed than clients presenting for treatment of incomplete abortion, who were more likely to be older, married, less educated, and unemployed.

Thus, in this study, the researcher has investigated how the attributes regarding the Socio- demographic characteristics and health related factors determine abortion occurrence in the selected area by applying the new computational methods of data mining technology. And this study also has examined the association of these frontier factors detailing with what degree of confidence and support they co-exist on a woman that has high probability of seeking abortion.

CHAPTER THREE

METHODS AND ALGORITHMS USED FOR KNOWLEDGE DISCOVERY

Discovery is the process of looking in a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be. In other words, the program takes the initiative in finding what the interesting patterns are, without the user thinking of the relevant questions first. In large databases, there are so many patterns that the user can never practically think of the right questions to ask. The key issue here is the richness of the patterns that can be expressed and discovered and the quality of the information delivered. This in turn determines the power and usefulness of the discovery technique.

A number of data mining algorithms have been introduced to the community that perform summarization of the data, classification of data with respect to a target attribute, deviation detection, and other forms of data characterization and interpretation. One popular summarization and pattern extraction algorithm is the association rule algorithm, which identifies correlations between items in transactional databases.

For this research, one of the data mining methods namely association rule and its algorithm Apriori were used.

The following section focuses on the description of the selected data mining method namely association rule and its algorithm; Apriori.

3.1 Association rule mining

Association rule mining refers to finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Frequent pattern here refers to pattern (set of items, sequence, etc.) that occurs frequently in a database.

The idea of mining association rules originates from the analysis of marketbasket data where rules like “A customer who buys products X1 and X2 will also buy product y with probability c%.” are found. Their direct applicability to business problems together with their inherent understandability- even for non-data mining experts- made association rules a popular mining method. Moreover it became clear that association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems (11).

An association rule is a rule which implies certain association relationships among a set of objects (such as ``occur together'' or ``one implies the other'') in a database. Given a set of transactions, where each transaction is a set of literals (called items), an **association rule** is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y (11).

Two probability measures, called support and confidence, are introduced to assess associations in the database. The support (or prevalence) of a rule is the proportion of observations that contain the item or item set of the rule. It is also known as the coverage of the rule. As defined by Witten and Frank (44), an item is an attribute value pair. The confidence is the conditional probability of B given A, P (B/A). A rule is “interesting” if the conditional probability P (B/A) is significantly different than P (B). Confidence of the rule measures the rule’s accuracy.

- Basic concepts:
  - itemset: A set of one or more items, X = {x_1, …, x_n} k-itemset: X = {x_1, …, x_k}
  - *support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
    - support of X and Y greater than user defined threshold s; i.e. support probability of s that a transaction contains X $\cup$ Y
    - An itemset X is *frequent* if X’s support is no less than a *minsup* threshold

– *Confidence:* is the probability of finding Y in a transaction with all $X_1, X_2, \ldots, X_n$.
    - confidence, *c,* conditional probability that a transaction having X also contains *Y; i.e.* conditional probability (confidence) of Y given X greater than or equal to user threshold c

Association rule mining is to find out association rules that satisfy the pre-defined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … ,Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1}⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub-problem is quite straight forward, most of the researches focus on the first sub-problem. The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation process and frequent itemsets generation process. We call those itemsets whose support exceed the support threshold as large or frequent itemsets, those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets (11).

3.1.1 How Do We Extract Association Rules from Datasets

As mentioned earlier, an association rule is an expression X_Y, where X and Y are sets of items. The meaning of such rule is quite intuitive: Given a database D of transactions- where each transaction T _ D is a set of items, X_Y express that whenever a transaction T contains X then T probably contains Y also.

The probability or rule confidence is defined as the percentage of transactions containing Y in addition to X with regard to the overall number transactions containing X.

That is, the rule confidence can be understood as the conditional probability
P(YÍ *T* / *X* Í *T*).

3.1.2Basic Principles

3.1.2.1 Formal Problem Description

As Hippet. al., (45) put it, the association rule discovery problem can be expressed mathematically as follows:

Let L= {x1. . . xn} be a set of distinct literals, called items. A set XÍ *L* with k=_X_ is called a k-itemset or simply an itemset. Let a database D be multi-set subsets of L. Each T_ D is called a transaction. We say that a transaction T _ D supports an itemset XÍ *L* if XÍ *T* holds. An association rule is an expression X_*Y* , where X, Y are itemsets and XÇY=ø holds. The fraction of transactions T supporting an itemset X with respect to database D is called the support of X, supp(X) = T_D _{XÍ *T*}_/_D_. The support of a rule X_*Y* is defined as supp (X_*Y*) = *p*(*X* È*Y*) .

The main challenge when mining association rules is the immense number of rules that theoretically must be considered. In fact the number of rules grows exponentially with | L |. Since it is neither practical nor desirable to mine such a huge set of rules, the rule sets are typically restricted by minimal thresholds for the quality measures support and confidence, **minsupp** and **minconf** respectively. This restriction allows us to split the problem into two separate parts: An itemset X is frequent if **supp**(X) ³ **min-supp**. Once, F={XÍ *L* / *X* frequent}, the set of all frequent itemsets together with their support values is known, deriving the desired association rules is straight forward: For every X _ F check the confidence of all values X/Y_*Y*,*Y* Í *X* , ø ¹ *Y* ¹ *X* and drop those that do not achieve minconf. According to its definition above, it suffices to know all support values of the subsets of X to determine the confidence of each rule. The knowledge about the

support values of all subsets of X is ensured by the downward closure property of itemset support: All subsets of a frequent itemset must also be frequent (45).

Generally speaking, the problem of discovering association rules can be divided into two steps (Nayak and Cook, n.d.):

1. Find all itemsets (sets of items appearing together in a transaction) whose support is greater than the specified threshold. Itemsets with minimum support are called frequent itemsets.

2. Generate association rules from the frequent itemsets. To do this, consider all partitioning of the itemset into rule left-hand and right-hand sides. Confidence of a candidate rule X®Y is calculated as support (XY) / support (X). All rules that meet the confidence threshold are reported as discoveries of the algorithm.

3.1.2.1.1 Traveling the Search Space

As explained above, we need to find all itemsets that satisfy min-supp. For practical applications looking at all subsets of L is doomed to failure by the huge search space.

The basic principle of most association rule algorithms is that if the parent class E' of a class E does not contain at least two frequent itemsets then E must also not contain any frequent itemset. If we encounter such a class E' on our way down the tree, then we have reached the border separating the infrequent from the frequent itemsets. We do not need to go behind this border so we prune E and all descendants of E from the search space. This procedure allows us to efficiently restrict the number of itemsets to investigate. We simply determine the support values only of those itemsets that we "visit" on our search for the border between frequent and infrequent itemsets. Today's common approaches for search employ either breadth-first search (BFS) or depth-first search (DFS). With BFS the support values of all (k-1) itemsets are determined before counting the support values of the k-itemsets. In contrast, DFS recursively descends the tree structure defined for itemsets(45).

3.1.2.1.2 Determine Itemset Supports

One common approach to determine the support value of an itemset is to directly count its occurrences in the database. Then all transactions are scanned and whenever one of the candidates is recognized as a subset of a transaction, its counter is incremented. Typically subset generation and candidate lookup is integrated and implemented on a hash tree or a similar data structure. Not all subsets of each transaction are generated but only those that are contained in the candidates or those which have a prefix in common with at least one of the candidates (45).

Another approach is to determine the support values of candidates by set intersections. A tid is a unique transaction identifier. For a single item the tidlist is the set of identifiers that correspond to the transactions containing this item. Accordingly tidlists also exist for every itemset X and are denoted by X.tidlist. The tidlist of a candidate C= X *Y* is obtained by C.tidlist=X.tidlist∩*Y*.*tidlist*. The tidlists are sorted in ascending order to allow efficient intersections. By buffering the tid lists of frequent candidates as intermediate results, we remarkably speed up the generation of the tidlists of the following candidates. Finally the actual support of a candidate is obtained by determining _C.tidlist_ (45).

3.1.3 Apriori Algorithm

When mining association rules there are mainly two problems to deal with: First of all there is the algorithmic complexity. The number of rules grows exponentially with the number of items. Fortunately today's algorithms are able to efficiently prune this immense search space based on minimal thresholds for quality measures on the rules. Second, interesting rules must be picked from the set of generated rules. This might be quite costly because the generated rule sets normally are quite large and in contrast the percentage of useful rules is typically only a very small fraction. The work concerning the second problem mainly focuses on supporting the user when browsing the rule set and the development of further useful quality measures on the rules (45).

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994(19) for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset

Apriori employs an iterative approach known as a *level-wise* search, where *k*-itemsets are used to explore (*k*+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted *L*1.Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets, which is used to find *L*3, and so on, until no more frequent *k*-itemsets can be found. The finding of each *Lk* requires one full scan of the database.(45)

This algorithm has emerged as one of the best association rule mining algorithms. It also serves as the base algorithm for most parallel algorithms. Apriori uses a complete, bottom-up search with a horizontal layout and enumerates all frequent itemsets. It is based on data passes. It identifies frequent "itemsets", subsets of items with a transaction, by performing as many data passes as specified by the user, or until there are no additional frequent itemsets to be identified.

Thus, the process or the algorithm starts by scanning all transactions in the database and computing the frequent items. Next, a set of potentially frequent candidate 2-itemsets is formed from the frequent items. Another database scan obtains their supports. The frequent 2-itemsets are retained for the next pass, and the process is repeated until all frequent itemsets have been enumerated. The algorithm has three main steps:

1. Generate candidates of length *k* from the frequent (*k* – 1) length itemsets, by a self-join on *Fk*–1. For example, for *F*2 = {*AC, AT, AW, CD, CT, CW, DW, TW*}, we get *C*3 = {*ACT,ACW, ATW, CDT, CDW, CTW*}.
2. Prune any candidate that has at least one infrequent subset. For example, *CDT* will be pruned because *DT* is not frequent.
3. Scan all transactions to obtain candidate supports. Apriori stores the candidates in a hash tree for fast support counting. In a hash tree, itemsets are stored in the

leaves; internal nodes contain hash tables (hashed by items) to direct the search for a candidate (45).

To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space. We will first describe this property, and then how it works.

3.1.3.1 Apriori property

All nonempty subsets of a frequent itemset must also be frequent. The Apriori property is based on the following observation. By definition, if an itemset *I* does not satisfy the minimum support threshold, *min sup*, then *I* is not frequent; thatis, *P*(*I*) <*min sup*. If an item *A* is added to the itemset*I*, then the resulting itemset (i.e., *I* [*A*) cannot occur more frequently than *I*. Therefore, *I* [*A* is not frequent either; thatis, *P*(*I* [*A*) <*min sup*.

During Apriori pruning phase, if there is any itemset which is infrequent, its superset should not be generated/tested. And the method is :

- Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemset. For each *k*, we construct two sets of *k* –tuples:
    - C_k *(candidate k*–tuples), those that might be frequent itemsets (support $\geq s$) based on information from the pass for *k* –1.
    - L_k = the set of truly frequent *k* –tuples.
  - Test the candidates against DB
  - Terminate when no frequent or candidate set can be generated (45)

Here the data miner uses the association Data Mining (DM) methods to derive knowledge from the MSIE abortion case data. Among the available algorithms to perform association mining tasks apriori were used in this research. The model was selected in this research due to its popularity in the recently published documents.

CHAPTER FOUR
DATA PREPARATION

Data Mining is a technology that uses various techniques to discover hidden knowledge from a data stored in large databases, data warehouses and other massive information repositories. To discover non-trivial knowledge and patterns, the database needs to undergo effective business understanding, data understanding and data preparations steps.

For this study, Hybrid DM model/process selected. Hybrid data mining process embraces six steps: understanding of the problems, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. Data understanding and preprocessing usually consumes the majority of the effort in the entire data mining process (19).

4.1 Business Understanding

Sharma and Osei-Bryson[46] argue that business understanding phase is somewhat more important than other phases like that of modeling, data preparation, data understanding, evaluation, etc. Lack of thorough discussions during Business understanding phase lead to the DM project may take a completely different direction than what was intended. Further, it creates inefficiencies with regard to time and resources as these decisions have to be dealt with in later phases (46).

Business understanding mainly concerned with determination of business objectives, assessment of the situations, and determination of data mining goals.

This study consulted domain experts and non-medical staffs to have insight into the problem domain and physical observation was conducted in the MSIE centers as well. The domain experts constitute individuals from MSIE centers who are in charge of Nursing and treating woman who seek family planning and sexual healthcare services.

4.1.1 Determination of Business Objectives

Business objectives determination requires discussion among either responsible business personnel who interact with the system in horizontal or vertical manner to finalize the business objective.

MSIE is a member of the Marie Stopes International (MSI) global partnership and a branch office of MSI in Ethiopia. MSIE was established in May 1990 following a bilateral agreement signed with the Federal Ministry of Health. MSI is a UK-based global, non-profit and non-governmental organization committed to upholding the fundamental rights of women and couples to decide freely and without coercion, the number and spacing of their children. MSI has a global network of country programs in about 40 countries in Africa, Asia, and elsewhere.

MSIE works closely with the Government of Ethiopia and other partners to contribute to the national effort of reducing maternal mortality and morbidity by increasing access to and quality of sexual and reproductive health services. MSIE currently operates 30 clinics that provide comprehensive RH services with a focus on FP and safe abortion in the country. In Addis Ababa, there are six centers, which are managed by MSIE and provide comprehensive obstetric care in addition to all ranges of FP and safe abortion care.

The primary objective of Marie Stopes International Ethiopia is to provide comprehensive sexual and reproductive health services through a network of its centres. These facilities provide a wide range of services including: general medical consultation, comprehensive family planning; pre- and post-natal care; child health checks; free condom provision; voluntary testing and counseling on HIV; and STI screening and treatment.

Regarding to abortion cases, each woman/girl, registered at MSIE clinical record at MSIE centers reception before taking examination by doctor. Then the physicians identify the pregnancy and decide whether she can undertake abortion procedure or not at MSIE based on her gestation period. At MSIE centers abortion is performed if the pregnancy is less than or equal to three months. Then she will be sent for counseling

service. At this session, she will be provided with information on the advantage and side effects of all available contraceptive methods in the center. After an abortion, patients recommended to use contraception for future use. Fortunately, in some cases though many years of experience, abortion case report may signify to identify determinant risk factors of the patients.

The MSIE abortion case report contains two general information of each woman who receives abortion service. The first category included the diagnostic results of woman in the abortion service delivery process e.g. pregnancy test result. The second category contain about socio demographic characteristics of woman and her reproductive health history with the current reason for using abortion service at MSIE centers.

Thoughtful discussions were conducted with domain experts and health background follow classmate in the area of maternal health and family planning before selecting the target dataset attributes. The discussion focused on the relevant and significant factors, from available attributes of MSIE abortion case report, which can help in identifying the risk factors of abortion. Because of these discussions, the researcher selected 7input attributes.

These attributes were selected due to their relevance or significance in identifying the risk factors of abortion. But, attributes that are not necessary for the data mining techniques to create a model are removed. Or in other words, patient name, date of exam, card no, and attributes which contain information on clinical diagnosis were discarded due to less importance for identifying risk factors of abortion and privacy issues also considered, typically for patient name. For this study, 7 attributes that were used to identify risk factors of abortion occurrence and to generate association rules (see Table 4.1).

Table 4.1 Identified List of attribute from MSIE abortion case report

List of Identified Attributes	
1	Age
2	Occupation
3	Marital status
4	Education
5	Religion
6	Previous induced abortion
7	Reasons for terminating pregnancy

4.1.2 Data Mining Goals

The principal investigator takes on the task of translating the business objective to data mining objective. The data-mining Goal 1; By using the MSIE abortion case records, To identify major risk factors that result in abortion occurrence. The data- mining goal 2; Find association rules for the factors.

4.2 Data Understanding

In order to meet the general objective of this research, collecting representative subset of MSIE abortion case data is the prerequisite. Data understanding begins with collection of initial data. The data collection process was carried out from MSIE centers which is situated in Addis Ababa, a total amount of 5361 abortion case patient's MSIE report covering from October 2010 to March 2011.

Because of the discussions with domain experts, the researcher selected 7input attributes for tasting apriori algorithm. Table 4.2has shown the7input attributes list along with their data description and data types.

Table 4.2 List of selected attributes along with their description and data types

S. No	Attribute Name	Description	Data types
1.	Age	Age of the patient in years	Numeric
2.	Occupation	Occupation of the patient	Categorical
3.	Marital status	Marital status of the patient	Categorical
4	Education	Educational status of the patient	Categorical
5.	Religion	Religion of the patient	Categorical
6.	Previous induced abortion	Previous induced abortion experience of the patient	Numeric
7.	Reasons for terminating pregnancy	the patient reasons for terminating the pregnancy	Nominal

4.2.1 Descriptive Data Summarization and Visualizations

Descriptive data summarization provides the general statistics summarization and visualization for characteristics of the data and identifies the presence of noise or outliers. Data characteristics regarding to central tendency include mean, median, mode and mid-range, while regarding to measure of data dispersion include quartiles, inter quartile range (IQR), and variance. Table 4.3displays the valid number of instance, minimum, maximum, the mean and the standard deviation of MSIE abortion case record numerical datasets.

Table 4.3 Descriptive Data summarization of Attributes

Descriptive Statistics						
S.No	**Attributes**	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
1	Age	5361	15	43	26.29	6.325
2	Previous induced abortion.	5361	o	5	.10	.411

The remaining MSIE abortion case attributes other than listed in table 4.3, namely Occupation, Marital status, Education, Religion, and Reasons for terminating pregnancy have a scale of measurement nominal.

4.3 Data Preprocessing

Today real word data are incomplete, inconsistent, noisy, redundant, missing due to their large size (gigabytes or more) and in some cases due to multiple sources. High quality data will lead to high-quality mining results and vice versa. Consequently, real world data of low quality needs preprocessing.

There are a number of data preprocessing tasks involved in this study such as data cleaning, handling outlier's data, data integration and transformation, and data reduction techniques. Data transformation, such as discretization, was applied to optimize the accuracy and efficiency of mining algorithms. Data reduction is another major task which can reduce the data size to obtain quick processing time and save memory.

The objective of data processing at this stage is two-fold; to obtain data prepared in the form required by the data mining algorithms and to expose as much information as possible for data modeling.

4.3.1. Data Cleaning

Data cleaning involves fill in missing values, smooth noisy data, identify or remove outliers etc.

4.3.1.1 Handling Missing Values

Most data encountered in practice, such as MSIE abortion case data as shown in Table 4.4, contains missing values. In the data missing values are frequently indicated by blank spaces, and sometimes unknown values placed in the field.

Missing values may occur for several reasons, such as malfunctioning measurement equipment, lack of consistency with other recorded data and thus deleted, or respondent in a survey may refuse to answer certain questions such as age, and data may not be recorded due to misunderstanding. But those missing values need to be given significant attention. For example in this MSIE abortion case data age is the most frequent missing value; what do these things mean about the presence or absence of

abortion incidence? Or does missing age have something to do with abortion incidence or it is just because of some random events? To make an informed judgment about the missing values, it would be very appropriate to consult with domain expert. Thus, in consultation with domain expert, the missing values have such significant implication on abortion incidence. Those missing values are not ordinary or random. So the missing values replaced by one of the following ways of handling missing values.

To deal with missing values, alternatives have been suggested by Larose[47] and Chakrabarti et al [48]: this are

- Ignore the missing value
- Replace the missing value manually
- Replace the missing value with a global constant to fill in the missing value
- Replace the missing value with some constant, specified by the analyst
- Replace the missing value with the field mean(for numerical variables) or the mode (for categorical variables)
- Replace the missing values with a value generated at random from the variable distribution observed.

In the data of MSIE abortion case reports the missing values of attributes are Age and Education with 60, and 32 respectively. The frequencies of missing value are shown in Table 4.4.

For this study, the missing Value were replaced globally before applying learning algorithm. This deed takes placed in SPSS version16 for each missing value with the mean for numeric attributes and the mode for nominal ones. This decision taken because of the proportions of missing values for variables is small, and likely not to have more than a small effect on the results derived from the data.

So the whole data, that is 5361records prepared in a way WEKA 3.7.5 required and were available for tasting purpose.

Table 4.4 Statistics of abortion case datasets for Missing, Mean, and Mode of attributes

Statistics					
S.No	**Attributes Name**	**N**		**Mean**	**Mode**
		Valid	**Missing**		
1	Age	5301	60	26.29	24
2	Occupation	5361			
3	Marital status	5361			
4	Education	5329	32		Secondary school
5	Religion	5361			
6	Previous induced abortion	5361			0
7	Reasons for terminating pregnancy	5361			

4.3.1.2 Handling Outliers Data

Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data. This outlier's data may arise due to typographic or measurement error. Further cause to noisy data can be faulty data collection instruments, data entry problems, data transmission problems, and inconsistencies in naming convention.

Identifying outliers is important because they may represent errors in data entry. Also, even if an outlier is a valid data point and not error they can change or affect the output. Few values are making to bend the output to a certain direction which should not be.

Certain statistical methods are sensitive to the presence of outliers and may present unstable results. One graphical method for identifying outliers for numeric variables is to examine a histogram of the variable [47] and for categorical variable one rectangle is drawn for each known value and called commonly to as a bar chart [48].

According to Chakrabarti et al [48], a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5 * IQR above the third quartile or below the first quartile. Or outlier's detection defined as [47]

1. It is lower than Q1 – 1.5(IQR)
2. It is higher than Q3 + 1.5(IQR)

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.

4.3.1.3 Data Integration and Transformation

When working on a data mining problem, it is first necessary to bring the data together into the same platform. Integrating data from different sources usually pause many challenges; these are different department may use different style of record keeping, different time periods, different conventions, different primary key and many others. Fortunately, MSIE abortion case dataset have been stored as a single Ms-Word file such that each file brought to Ms-Excel file format to create database of abortion case records. Then, from Ms-Excel datasets exported to SPSS 16 for descriptive analysis.

As per Han et al [19], Data transformation is about transforming or consolidating the data to make it appropriate for mining. Data transformation can involve the following operations:

- Smoothing; this techniques include binning, regression, and clustering which works to remove noise from the data.
- Generalization of the data; where low-level raw data are replaced by higher-level concepts. For example numeric attribute of age may be generalized to youth, middle age, and old age.
- Normalization; this operations performed on attribute to scaled the value to fall within a small specified range.

Since Apriori algorithm, which is the selected data mining algorithm, is often used in situations where attributes are nominal, the data was transformed to fit this algorithm.

Among those techniques of transformation, the attribute with numerical data type, namely Age were converted into nominal data type, since the chosen algorithm dictates the use of attributes with such data type. The age attribute was converted into six nominal categories each representing the age interval 15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39 and 40 to 45. In addition, the previous induced abortion numerical value was transformed by using its original value when data encoding takes placed. The encoding techniques handled with SPSS 16, Transform menu of recode into same variable futures; this enables to change the existing variables value with the given numeric value without creating additional variables.

4.3.1.4 Data Reduction

Data reduction techniques are applied to obtain reduced datasets and yet maintain the integrity of the original data. Strategies of data reduction include data cube aggregation; attribute subset selections, dimensionality reduction, numerosity reduction, and discretization and concept hierarchy generations [19, 48]. In this study dimensionality reduction applied.

4.3.1.4.1 Dimensionality Reduction

According to Chakrabarti et al [48] in data reduction, data encoding or transformation are applied to reduce or compress representation of the original data. Data compressions are in two kinds, lossless and lossy. Lossless are if the original data can be reconstructed from the compressed data without any loss of information we called it lossless. Whereas lossy is if we reconstruct only an approximation of the original data, then the data reduction is called lossy

The principal investigator in this research applied the lossless which does not lose any data in the compression process. In MSIE abortion case datasets the attributes of Age(15-19, 20-24, 25-29, 30-34, 35-39, 40-45), Marital status(Married, Single, Divorced Widowed, Separated), Occupation(House wife, Student, Govt. employee , Unemployed) Religion(Orthodox, Muslim, Protestant, Catholic, Others) Reasons for terminating pregnancy(Need for spacing, Not wanting a child, Partner decision, Societies disapproval, Medical reasons, Rape, Not wanting a child and Medical reasons) Education level(Illiterate, Read and write, Primary school, Junior secondary,

Secondary school University/college) have a categorical variable. Those attributes value are changed to numeric format and followed with data size reduction. The encoding techniques handled with SPSS 16, Transform menu of recode into same variable futures; this enables to change the existing variables value with the given numeric value without creating additional variables. Table 4.5 showed the detail of recoding with regard to attributes name, their old value and their new numeric value as well.

Table 4.5 Attribute Encoding New value for replacement of old value

S.No	Attributes name	Original Value	Respective New Value
1	Age	{15-19,20-24, 25-29, 30-34, 35-39, 40-45}	{1,2,3,4,5,6}
2	Marital status	{Married, Single, Divorced Widowed, Separated }	{1,2,3,4,5}
3	Occupation	{House wife, Student Govt. employee , Unemployed }	{1,2,3,4}
4	Religion	{Orthodox, Muslim, Protestant, Catholic, Others }	{1,2,3,4,5}
5	Reasons for terminating pregnancy	{Need for spacing, Not wanting a child, Partner decision, Societies disapproval, Medical reasons, Rape, Not wanting a child and Medical reasons}	{1,2,3,4,5,6,7}
6	Education level	{Illiterate, Read and write, Primary school, Junior secondary, Secondary school University/college}	{1,2,3,4,5,6}
7	Previous induced abortion	{0,1,2,3,4,5}	{0,1,2,3,4,5}

Further, the data format also was changed into attribute relation file format (arff) because Weka accepts only comma separated arff format text files.

CHAPTER FIVE

DATA MINING AND EVALUATIONS OF DISCOVERED KNOLEDGE

Once a good quality data is generated, a descriptive model is created using association rule method and its algorithm apriori. Experiment results and corresponding discussions are presented on each experiment.

5.1 Experimental Setup

The abortion case datasets was presented in a spreadsheet format. However, WEKA natively store data as an ARFF format; as a result the datasets are converted from a spreadsheet to ARFF. The ARFF file consists of a list of the instances, and attribute values for each instances separated by commas.

Mostly spreadsheet and database programs allow you to export data into a file in comma-separated value (CSV) format as list of records with commas between. Then you only load the file into text editor or word processor; add the dataset's name using the @relation tag, the attribute information using @attribute, and a data information with @data tag and save the file as raw text with .arff file format [18]. The .arff presented below depicts the sample of machine understandable format of the dataset in WEKA employed for this study.

To build the association rule model, the arff format of the selected dataset was given to WEKA 3-7-5, which supports apriori algorithm.

@relation abortion case

@attribute age {1, 2, 3, 4, 5, 6}

@attribute educ {1, 2, 3, 4, 5, 6}

@attribute marst {1, 2, 3, 4, 5}

@attributeoccupa {1, 2, 3, 4}

@attribute PIA {0, 1, 2, 3, 4, 5}

@attribute religions {1, 2, 3, 4, 5}

@attribute RIAP {1, 2, 3, 4, 5, 6, 7}

@data

4,5,1,1,1,1,1

5,3,1,2,0,1,1

3,1,2,2,0,1,4

2,1,1,1,0,1,7

2,2,1,1,0,5,7

1,4,1,2,0,1,1

4,3,2,2,0,1,6

3,5,2,2,0,2,7

3,4,2,2,0,1,5

1,5,2,2,1,2,7

1,4,2,2,1,3,6

5,5,2,2,0,1,7

3,1,4,1,0,4,5

3,5,2,2,0,2,7

2,6,1,3,0,1,7

1,4,1,2,0,3,1

4,5,2,2,0,2,7

2,6,1,3,0,2,1

3,1,1,1,0,2,5

Before the initial experiment, the researcher took and prepared 7 attributes for association rule model building purpose. The dataset include a total of 5361 records

having 7 attributes each. In this research those 5361 datasets are used for tasting purpose.

In this research, four experiments conducted with the following research design of data mining algorithms, parameters, and attribute selection.

Experiment 1

The association rule learning algorithm was applied on the whole dataset that had been cleaned and transformed into the .arff format.

Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items. But in this research, in order to generate relevant and applicable association rules, a minimum support threshold and a minimum confidence threshold set by the researcher with consulting domain experts. The information on running the algorithm on the database is presented in the table below.

Table 5.1: Run information of Apriori (5361 instances and 7 attributes)

Scheme	weka.associations.Apriori
-N(required number of rules output)	20
-T(metric type by which to rank rules)	0 (confidence)
-C (the minimum confidence of a rule)	0.9
-D (delta at which the minimum support is decreased at each iteration)	0.05
-U (upper bound for minimum support)	1.0
-M (the lower bound for the minimum support)	0.1
-S (significance of a rule at a given level)	-1.0
Relation	abortion
Instances	5361
Attributes	7 (see table 4.1 to view the list of attributes)

According to Agrawal in Stiles (nd), the support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others are referred as small itemsets. Large itemsets are used to generate the desired rules.

At the end of this experiment, several rules that satisfy the above metrics were generated. The following table depicts the size and frequency of the large item sets generated.

Table 5.2: Size and frequency of generated rules (5361 instances and 7 attributes)

Size of Generated large Itemsets	Frequency of Large itemsets
One	13
Two	24
Three	13
Four	3

As shown in the above table, there were thirteen one-item frequent itemset, twenty-four two-item frequent itemsets, thirteen three-item frequent itemsets and three four-item frequent itemsets generated.

Apriori orders rules according to their confidence and uses support as a tiebreaker. Although Apriori tries to generate ten rules, in this research the number of rules generated by the algorithm was specified to be double this default size (i.e. 20).

Best rules found

Apriori generated a number of rules that satisfy the above set minimum metrics of support and confidence. If a rule has a confidence above the minimum set confidence, then the rule holds.

For the purpose of illustration five of the 20 set of rules produced were presented below *(See appendix1 to view the complete list of best 20 set of rules generated by Apriori):*

In selecting rules for discussion, the researcher focused on the rules generated from the superset of frequent or large itemset consisting of the highest size of large itemset. And

on those domain experts were found to be interested in. The following are rules selected for discussion from experiment 1

Marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)

(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 52% support (i.e. 2756/5361) and a 96% accuracy/confidence

This is the rule with the largest item set size. The number preceding the '==>' symbol indicates the rule's support, that is the number of items covered by its antecedent. Following the rule is the number of those items for which the rule's consequent holds as well. In parenthesis is the confidence of the rule, i.e. the second number divided by the first.

. Occupa=1 PIA=0 RIAP=7 1295 ==>marst=1 1295 <conf:(1)

(If a woman occupational status = housewife and her induced abortion experience is = No and her reason for terminating the pregnancy= Not wanting a child and Medical reasons then her marital Status = married)

This rule has an 24% support (i.e. 1295/5361) and a 100% accuracy/confidence.

Apriori orders rules according to their confidence and uses support as a tiebreaker. Although Apriori tries to generate ten rules, in this research the number of rules generated by the algorithm was specified to be double this default size (i.e. 20).

Occupa=1 PIA=0 RIAP=1 1145 ==>marst=1 1145 <conf:(1)

(If a woman occupational status = housewife and her induced abortion experience is= No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 21% support (i.e. 1145/5361) and a 100% accuracy/confidence.

PIA=0 RIAP=1 1445 ==>marst=1 1418 <conf:(0.98)

(If a woman induced abortion experience is = No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 27% support (i.e. 1445/5361) and a 98% accuracy/confidence.

Educ=1 marst=1 1183 ==> PIA=0 1143 <conf:(0.97)

(If a woman Educational Status = illiterate and her marital Status = married then her induced abortion experience is = No)

This rule has an 22% support (i.e. 1445/5361) and a 97% accuracy/confidence.

The above generated rules constitute those attributes (itemsets) with large frequency throughout the dataset. The rules represent interesting regularity within the abortion case database. For example, the rule “(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)” holds that woman, who is married, housewife, and new for using induced abortion are more vulnerable for terminating the pregnancy or using induced abortion service.

The rules generated in this experiment revolve around in marital status of woman, her previous experience in induced abortion, her reason for terminating the pregnancy, her educational and occupational status. According to these rules, the other features that make up the women profile include married and housewife woman, who have not experience in using induced abortion services and with the current reason for abortion is not wanting a child, need for spacing, or medical condition is having a risk in using induced abortion.

In addition, it is easier to observe that the above rules apply to the real world and thus are meaningful. According domain experts if a woman is illiterate and housewife she may have no decision making power and vulnerable for induced abortion rather than using contraceptive methods to control unwanted pregnancy.

This rule indicates that the instances in the database are characterized by the occurrence of very high instances with previously discussed facts, Apart from

discovering surprising or hidden rules, the learning scheme also results this rule that confirm facts existing in the real world.

But those rules can be considered as trivial since anyone with the knowledge of the distribution of the respective values of these attributes can tell the possible relationship. That is in the above attributes, particular values highly dominate the value for that specific attribute. In the case of the attribute **PIA (previous induced experience)** the value 0(no experience) account for the vast majority of the instances in the database (i.e. 4945/5361). Similarly for the attribute **marital status**, the value **married** account for the lion's share of the instances in the database (i.e.3190/5361). The situation is the same for **Reason for terminating the pregnancy** where **Not wanting a child and Medical reasons** account for 3590 of the instances, and **occupational status** where there are 2761 value of **housewife***.*

Experiment 2

In the effort to discover relatively more interesting rules that underlie the abortion case dataset, continuing with the experimentation was imperative. During subsequent experimentation, leaving out the most frequent or invariable attributes that dominate the rules in the first experiment was considered to be a proper measure, since the use of these attributes resulted in more or less uninteresting or trivial rules

In this experiment, one of the attributes, namely **marital status**, were progressively removed to give chance for other attributes to be considered in the construction of the rules. Accordingly, an experiment was run over the following six attributes:, Age, Occupation, Education, Religion, Previous induced abortion and Reasons for terminating pregnancy.

The information from running the algorithm is presented below.

Table 5.3: Run information of Apriori (5361 instances and 5 attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c -1
Relation	Abortion case
Instances	5361
attributes	6

Table 5.4: Size and frequency of generated rules (5361 instances and 5 attributes)

Size of Generated large Itemsets	Frequency of Large itemsets
One	17
Two	41
Three	22
Four	1

Best rules found

Apriori was made to generate 20 rules that best satisfy the set confidence and support metrics *(see appendix 2 to view the complete list of the best 20 rules generated).* The researcher together with domain experts selected those rules that were considered to be interesting. These include:

Age=3 occupa=1 855 ==> PIA=0 835 <conf:(0.98)

(If a woman age= between 25-29 and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 16% support (i.e. 835/5361) and a 98% accuracy/confidence.

Educ=3 895 ==> PIA=0 850 <conf:(0.95)

(If a woman Educational Status =primary school then her induced abortion experience is = No)

This rule has an 16% support (i.e. 850/5361) and a 95% accuracy/confidence.

Age=3 1560 ==> PIA=0 1445 <conf:(0.93)

(If a woman age= between 25-29 and her then her induced abortion experience is = No)

This rule has an 27% support (i.e. 1445/5361) and a 98% accuracy/confidence.

Age=2 1545 ==> PIA=0 1420 <conf:(0.92)

(If a woman age= between 20-24 and her then her induced abortion experience is = No)

This rule has a 26% support (i.e. 1420/5361) and a 98% accuracy/confidence.

Educ=4 820 ==> PIA=0 755 <conf:(0.92)

(If a woman Educational Status =junior school then her induced abortion experience is = No)

This rule has an 14% support (i.e. 755/5361) and a 98% accuracy/confidence.

The above sets of rules supplement the previously discovered regularities regarding those woman age (20-24 and 25-29) and educational status (primary and junior school) who are not experienced in abortion. According to the rules generated in this experiment, woman who were exposed to abortion include woman who have the age between (20-24 and 25-29) and educational status (primary and junior school).

Despite the discovery of these interesting rules, the experimentation process was sustained in search of more and better rules. This time another attribute, namely PIA (previous induced experience) the value 0(no experience) account for the vast majority of the instances in the database (i.e. 4945/5361) was left out.

Even though the number of attributes was decreased from 6 to 5 to give chance for other attributes, there was not change in the types of rules generated.

Experiment 3

Even though the rules generated in second experiment are evaluated as meaningful and interesting, the researcher believed that more of such rules could be generated and thus preceded with another experiment. The researcher attempted to generate more rules by using binary attribute of abortion case dataset and see if there was any change.

Keeping in mind the goal of producing an improvement in the support and confidence level of generated rules, and discovering subjectively interesting rules within the problem domain, the nominal attributes (data) with multiple categories were converted to binary format then those binary attributes converted to nominal so that they can be

processed by the data mining software that handle nominal attributes only, namely the apriori algorithm. The following was the result:

The 7 nominal attributes were converted into 39 binary attributes. The converted binary representation of the 39 attributes by using the Weka knowledge explorer (see appendix 3).

The information from running the algorithm is presented below

Table 5.5: Run information of Apriori (5361 instances and 39 binary attributes**)**

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation	Abortion case
Instances	5361
attributes	39 binary attributes

Table 5.6: Size and frequency of generated rules (5361 instances and 39 binary attributes**)**

Size of Generated large Itemsets	Frequency of Large itemsets
One	15
Two	72
Three	151
Four	145
Five	72
Six	18
Seven	1

Best rules found

The Apriori algorithm generated 20 best rules on the basis of confidence and support metrics *(See Appendix 3 to view the complete list of 20 best rules generated)*. Interesting rules were selected in consultation with domain experts.

Except for few, almost all of the generated rules were found to be uninteresting and trivial. To illustrate a few rules are presented below:

age=6=0 PIA=2=0 5116 ==> PIA=3=0 5096 <conf:(1)

(If a woman age= not between 40-44 and her induced abortion experience is for two times = No) then her induced abortion experience is for three times = No)

This rule has an 95% support (i.e. 5116/5361) and a 100% accuracy/confidence.

Although not conclusive, the above rule imply that “95% of women are under the age of 40 with have not two and three times induced abortion experience“ is the characteristics of woman in the MSIE abortion case dataset. Or in another sense, if a woman age above than 40 she will not use induced abortion services with 95% support and 100% confidence. Having age above 40 can minimize the risk of the occurrence of abortion.

Experiment 4

Attempt was made to leave out less relevant attributes, and the 39 attributes were progressively reduced and the learning algorithm was made to run on the data with 28 attributes.

The information from running the algorithm is presented below.

Table 5.7: Run information of Apriori (5361 instances and 29 binary attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation	Abortion factors
Instances	5361
attributes	28 binary attributes

Table 5.8: Size and frequency of generated rules (5361 instances and 28 attributes**)**

Size of Generated large Itemsets	Frequency of Large itemsets
One	7
Two	14
Three	16
Four	7
Five	1

Best rules found

The Apriori algorithm generated 20 best rules on the basis of confidence and support metrics *(See Appendix 4 to view the complete list of 20 best rules generated)*. Interesting rules were selected in consultation with domain experts.

PIA=2=0 PIA=3=0 PIA=4=0 5281 ==> PIA=5=0 5276 <conf:(1)

(If a woman induced abortion experience for two times = No her induced abortion experience for three times = No and her induced abortion experience for four times = No then her induced abortion experience for five times = No)

This rule has an 98% support (i.e. 5271/5361) and a 98% accuracy/confidence.

Although not conclusive, the above rule imply that if a woman have not abortion experience for two, three or four times she will have not experienced for five times with 98% support and 98% confidence.

In another word the woman characteristics, in the MSIE abortion case dataset, regarding to previous abortion experience is with no experience or one previous abortion experience with satisfying the above support and confidence.

In this experiment, the findings are in line with the popular conception that such No experience are more often characteristics of woman than having experience of using induced abortion

While using the data that has been converted to the binary format for the generation of association rule, found a rule that is not conclusive, since the non-existent state of the attributes constitutes the rules. Hence, the value of the rules can only be appreciated for what they imply rather than what they state/conclude.

The researcher further attempted to generate more relevant set of rules which can give insight about the problem area of abortion risk factors. Careful attribute selection is one important step in data mining process to get a better result. To find attributes which can give more information on the specific problem area, the researcher attempted to evaluate attribute relevance to the problem of terminating pregnancy using one of Weka's facilities, attribute selection was used.

Even though the researcher used one of Weka's facilities, attribute selection to generate more relevant set of rules, there was not change in the types of rules generated.

5.2 Interpretation and discussion

From the different list of rules generated over various experiments using different set of data and attributes, a number rules with satisfactory objective measure (high support and confidence) and most importantly meeting the subjective judgment of domain experts on their interestingness and applicability were selected.

Summary of the input and output of the discovery task, and the subsequent interpretation and discussion of the discovered interesting rules is presented below.

Experiment 1

Input

The input and output for the first experiment are

Instances: 5361

Attributes: Age, Marital status, Occupation, Religion, Previous induced abortion, Reasons for terminating the pregnancy, and Education level.

Output

Marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)

(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 52% support (i.e. 2756/5361) and a 96% accuracy/confidence

Marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)

(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 52% support (i.e. 2756/5361) and a 96% accuracy/confidence

. Occupa=1 PIA=0 RIAP=7 1295 ==>marst=1 1295 <conf:(1)

(If a woman occupational status = housewife and her induced abortion experience is = No and her reason for terminating the pregnancy= Not wanting a child and Medical reasons then her marital Status = married)

This rule has an 24% support (i.e. 1295/5361) and a 100% accuracy/confidence.

Occupa=1 PIA=0 RIAP=1 1145 ==>marst=1 1145 <conf:(1)

(If a woman occupational status = housewife and her induced abortion experience is= No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 21% support (i.e. 1145/5361) and a 100% accuracy/confidence.

PIA=0 RIAP=1 1445 ==>marst=1 1418 <conf:(0.98)

(If a woman induced abortion experience is = No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 27% support (i.e. 1445/5361) and a 98% accuracy/confidence.

Educ=1 marst=1 1183 ==> PIA=0 1143 <conf:(0.97)

(If a woman Educational Status = illiterate and her marital Status = married then her induced abortion experience is = No)

This rule has an 22% support (i.e. 1445/5361) and a 97% accuracy/confidence.

The rule generated in this experiment revolve around in marital status of woman, her previous experience in induced abortion, her reason for terminating the pregnancy, her educational and occupational status. According to these rules, the other features that make up the women profile include married and housewife woman, who have not experience in using induced abortion services and with the current reason for abortion is not wanting a child, need for spacing, or medical condition is having a risk in using induced abortion.

It is easier to observe that the above rules apply to the real world and thus are meaningful. According domain experts if a woman is married, illiterate and housewife she may have no decision making power and vulnerable for abortion rather than using contraceptive methods to control unwanted pregnancy.

Experiment 2

Input

Instances: 5361

Attributes: Age, Occupation, Religion, Reasons for terminating the pregnancy, previous induced abortion and Education level.

Output

Age=3 occupa=1 855 ==> PIA=0 835 <conf:(0.98)

(If a woman age= between 25-29 and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 16% support (i.e. 835/5361) and a 98% accuracy/confidence.

Educ=3 895 ==> PIA=0 850 <conf:(0.95)

(If a woman Educational Status =primary school then her induced abortion experience is = No)

This rule has an 16% support (i.e. 850/5361) and a 95% accuracy/confidence.

Age=3 1560 ==> PIA=0 1445 <conf:(0.93)

(If a woman age= between 25-29 and her then her induced abortion experience is = No)

This rule has an 27% support (i.e. 1445/5361) and a 98% accuracy/confidence.

Age=2 1545 ==> PIA=0 1420 <conf:(0.92)

(If a woman age= between 20-24 and her then her induced abortion experience is = No)

This rule has a 26% support (i.e. 1420/5361) and a 98% accuracy/confidence.

Educ=4 820 ==> PIA=0 755 <conf:(0.92)

(If a woman Educational Status =junior school then her induced abortion experience is = No)

This rule has an 14% support (i.e. 755/5361) and a 92% accuracy/confidence.

The above sets of rules supplement the previously discovered regularities regarding those woman age (20-24 and 25-29) and educational status (primary and junior school) who are not experienced in abortion. According to the rules generated in this experiment, woman who were exposed to abortion include woman who have the age between (20-24 and 25-29) and educational status (primary and junior school) with have not experience in using abortion services before .

Experiment 3

Instances: 5361
Attributes: 39 binary attributes

Output

age=6=0 PIA=2=0 5116 ==> PIA=3=0 5096 <conf:(1)

(If a woman age= not between 40-44 and her induced abortion experience is for two times = No)then her induced abortion experience is for three times = No)

This rule has an 95% support (i.e. 5116/5361) and a 100% accuracy/confidence. Although not conclusive, the above rule imply that "95% of women are under the age of 40 with have not two and three times induced abortion experience" is the characteristics of woman in the MSIE abortion case dataset. Or in another sense, if a woman age above than 40 she will not use induced abortion services with 95% support and 100% confidence. Having age above 40 can minimize the risk of the occurrence of abortion

The rules that were found to be interesting by the domain experts include a rule where the instances in the database are characterized by the occurrence of very few instances woman age between 40 and 45. The rules implies that if a woman age more than 40 she will not use induced abortion services with 95% support and 100% confidence. And this is one of the interesting regularities in the dataset, woman between the ages of 40 and 45 are the ones who not often experienced in using induced abortion at MSIE centers.

Experiment 4
Input
Instances: 5361
Attributes: 28 binary attributes

Output

PIA=2=0 PIA=3=0 PIA=4=0 5281 ==> PIA=5=0 5276 <conf:(1)

(If a woman induced abortion experience for two times = No her induced abortion experience for three times = No and her induced abortion experience for four times = No then her induced abortion experience for five times = No)

This rule has an 98% support (i.e. 5271/5361) and a 100% accuracy/confidence.

Although not conclusive, the above rule imply that if a woman have not abortion experience for two, three or four times she will have not experienced for five times with 98% support and 100% confidence.

In another word the woman characteristics, in the MSIE abortion case dataset, regarding to previous abortion experience is with no experience or one previous abortion experience with satisfying the above support and confidence.

As mentioned by center managers and domain experts, “Each woman/girl take examination by doctor who decides whether she can undertook abortion procedure or not at MSIE based on her gestation period. Then she will be sent for counseling service. At this session she will be provided with information on the advantage and side effects of all available contraceptive methods in the center

Almost all clients who undertook abortion at MSIE centers start contraceptive use after their abortion procedure. The counselors gave information on types and benefits of contraceptives and inform clients that they can get pregnant after ten or fifteen days of their procedure. MSIE strongly encourage contraceptive as pregnancy may happen to them again.

So, it might decrease the number of woman, who have more than one experience in terminating pregnancy (have induced abortion one time or more) coming to MSIE centers for second time. That means we have 100% confidence that 5271(with 98% support) abortion case records of the total 5361 total are new for terminating the pregnancy (have no experience in induced abortion) or only one time. This was found as indication that woman previous induced abortion experience is one factor to reduce the chance of using abortion service.

The rules generated over a number of experiments constitute those attributes (itemsets) occurring in large frequency in the dataset. Despite the fact that most of the generated rules scored high in terms of the objective measures of interestingness (high support and confidence), most were found to be less interesting in the eyes of users/domain experts and the purpose of the research, which is discovering interesting rules/regularities.

In the effort to discover relatively more interesting rules that underlie the abortion case dataset, a series of experiments were conducted. During subsequent experimentation, leaving out the most frequent or invariable attributes that dominate the rules in the first experiments was considered to be a proper measure, since the use of these attributes resulted in more or less uninteresting or trivial rules.

Based on the evaluation of the domain area experts, determinant factors or characteristics of woman who are exposed to abortion were identified. Accordingly, attributes "Previous induced abortion experience: (zero time (no experience))", "Marital status: (Married)", " Occupation: (House wife)", " Reasons for terminating pregnancy: (Need for spacing, not wanting a child and Medical reasons)", "Education level: (Illiterate, Secondary school and junior school)", "Age (20-24, 25-29 and below 40)" Were identified as determinant factors the selection was done experimentally and supported by domain experts.

Based on the evaluation of the domain area experts, the rules generated above using different attributes have practical relevance to the problem of identifying risk factors of abortion and the associations between them because they did give some additional insight about the problem of why woman facing unintended pregnancy. A problem area should be first fully known before rushing to develop solutions. Therefore the result of this research can be used as input for developing programs and strategies for prevention and control of unintended pregnancy of women.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

In this section the conclusions drawn from the findings of the research and the recommendations forwarded in light of the findings and conclusions are presented.

6.1 Conclusions

The objective of this research was to extract hidden knowledge from abortion case datasets using data mining technique. The experiments carried out in this research using association mining algorithm apriori. Besides, the results were promising and encouraging especially in the eye of domain experts.

In data mining application, first the data in hand and the business problem to be solved must be analyzed and understood very well. Suitable mining techniques also play an important role for successful data mining application with data preprocessing in the present study. Mach emphasis is given for business understanding and data understanding to make sure the best possible results obtained.

In this research, the methodology employed was Hybrid Data mining process model; it involves six steps and the principal researcher rigorously passes through all the steps and iterated as needed. A total of 5361 abortion case records were used to generate association rules.

In order to identify major risk factors of induced abortion and their association, several experiments were conducted by using Data mining: Association rule mining, an exploratory data mining technique was applied to accomplish the goal of the research. To this effect, the Apriori algorithm, which is an implementation of the Association rule in the Weka software, was used.

Data evaluation and interpretation: The learning algorithm was able to generate a number of rules over a series of experiments. On account of subjective (opinions of domain experts) and objective (support and confidence) measures of interestingness, a

number of rules having practical relevance or that can increase to the current knowledge in the problem domain were identified.

Based on the evaluation of the domain area experts, determinant factors or characteristics of woman who are exposed to abortion were identified. Accordingly, attributes "Previous induced abortion experience: (zero time (no experience))", "Marital status: (Married)", " Occupation: (House wife)", " Reasons for terminating pregnancy: (Need for spacing, not wanting a child and Medical reasons)", "Education level: (Illiterate, Secondary school and junior school)", "Age (20-24, 25-29 and below 40)" Were identified as determinant factors. The selection was done experimentally and supported by domain experts.

The generated rules were found to be satisfactory from the point of view of the objective and subjective measures of interestingness. List of ten discovered interesting rules is presented below.

1. **Marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)**

(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 52% support (i.e. 2756/5361) and a 96% accuracy/confidence

2. **Marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)**

(If a woman marital Status = married and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 52% support (i.e. 2756/5361) and a 96% accuracy/confidence

3. **PIA=0 RIAP=1 1445 ==>marst=1 1418 <conf:(0.98)**

(If a woman induced abortion experience is = No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 27% support (i.e. 1445/5361) and a 98% accuracy/confidence.

4. **Occupa=1 PIA=0 RIAP=7 1295 ==>marst=1 1295 <conf:(1)**

(If a woman occupational status = housewife and her induced abortion experience is = No and her reason for terminating the pregnancy= Not wanting a child and Medical reasons then her marital Status = married)

5. **Occupa=1 PIA=0 RIAP=1 1145 ==>marst=1 1145 <conf:(1)**

(If a woman occupational status = housewife and her induced abortion experience is= No and her reason for terminating the pregnancy=Need for spacing then her marital Status = married)

This rule has an 21% support (i.e. 1145/5361) and a 100% accuracy/confidence.

6. **Educ=1 marst=1 1183 ==> PIA=0 1143 <conf:(0.97)**

(If a woman Educational Status = illiterate and her marital Status = married then her induced abortion experience is = No)

This rule has an 22% support (i.e. 1445/5361) and a 97% accuracy/confidence.

7. **Age=3 occupa=1 855 ==> PIA=0 835 <conf:(0.98)**

(If a woman age= between 25-29 and her occupational status = housewife then her induced abortion experience is = No)

This rule has an 16% support (i.e. 835/5361) and a 98% accuracy/confidence.

8. **Age=2 1545 ==> PIA=0 1420 <conf:(0.92)**

(If a woman age= between 20-24 and her then her induced abortion experience is = No)

This rule has a 26% support (i.e. 1420/5361) and a 98% accuracy/confidence.

9. **Age=6=0 PIA=2=0 5116 ==> PIA=3=0 5096 <conf:(1)**

(If a woman age= not between 40-44 and her induced abortion experience is for two times = No)then her induced abortion experience is for three times = No)

10. **PIA=2=0 PIA=3=0 PIA=4=0 5281 ==> PIA=5=0 5276 conf:(1)**

(If a woman induced abortion experience for two times = No her induced abortion experience for three times = No and her induced abortion experience for four times = No then her induced abortion experience for five times = No)

This rule has an 98% support (i.e. 5271/5361) and a 98% accuracy/confidence.

In general, the results from this study were promising to apply data mining for identifying the risk factors of induced abortion and prevention. The results were more applicable and appropriate to the problem domain since its simplicity and easily understandable rules that can be used by health care professional and policy makers as well.

In this research work, the researcher has proved that a clinical database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with abortion occurrence in Addis Ababa area. Based on the evaluation of the domain area experts, determinant factors or characteristics of woman who are exposed to abortion were identified.

Accordingly, based on MSIE abortion case database, woman with high risk to abortion incidence, were younger, married, less educated, and housewife. In addition less experienced woman in using abortion services were at high risk than experienced.

The findings of this research will conclude that, list of identified risk factors of abortion incidence or characteristics of woman who are vulnerable for induced abortion were presented below:

Accordingly, the risk factors of abortion occurrence are as follows:

a) **Woman currently married**: married women as well as age group of 20-24 and 25-29 years women constitute the major portion of induced abortion service users. So in this research being married, housewife and having the age between 20-24 and 25-29 years were identified as characteristics of woman who are in high risk in abortion incidence.

b) **Woman with no previous abortion experience**: Most of the women never use any of the family planning methods previously or less experienced woman in using abortion services were at high risk than experienced.

Almost all clients who undertook abortion at MSIE centers start contraceptive use after their abortion procedure. So, it might decrease the number of woman, who have more than one experience in terminating pregnancy (have induced abortion one time or more) coming to MSIE centers for second time. In another word, this was found as indication that woman previous induced abortion experience is one factor to reduce the chance of using abortion service.

c) **Women who are house wives:** if a woman is housewife she may have no decision making power and vulnerable for induced abortion rather than using contraceptive methods to control unwanted pregnancy. So in this research housewife woman are vulnerable or at high risk in abortion occurrence than other occupations.

d) **Woman who terminated pregnancy for need for spacing or medical reasons**: many women with children decided to have an abortion because they wanted either to postpone the next birth or due to medical reason. So woman who are using abortion as family planning method other than other methods are at high risk in unintended pregnancy and its result abortion.

e) **Woman with education level of Illiterate, Secondary school or junior school**: in this research, woman who has an education level illiterate, Secondary school or junior school are exposed in using abortion services.

6.2 Recommendations

The researcher makes the following recommendations based on the result of this study.

- The risk factors determining the incidence of induced abortions vary from place to place. Hence, it is not possible to define a single strategy that can be applied universally to deal with the problem. Noting that the demand for induced abortions is an indication of the frequency of unwanted pregnancies, the prevention of unplanned pregnancies is recognized as an important element in

the strategy for reducing the frequency of abortions. It would be inappropriate to focus too narrowly on the question of abortion as an isolated issue, but it should be viewed more widely in the context of social, economic, and political factors.

- Better documentation of the overall incidence of abortion and of its two components—clandestine or unsafe abortion and legal abortion—is essential for informing policy decisions and program design, and for monitoring the impact of existing policies and programs in Ethiopia. Data on the overall incidence of induced abortion is a crucial indicator of women's and couples' difficulties in preventing unintended pregnancies, and of their need for better contraceptive services.
- The possibility of incorporating the findings of this study in another (and operational) data mining application should be explored.
- It is indeed very important to assess the applicability of data mining techniques in identifying risk factors and their associations by using clinical datasets gathered from different hospitals and non-clinical or epidemiological datasets.

REFERENCES

1. WHO 2007 Unsafe abortion: Global and regional estimates of the incidence of unsafe abortion and associated mortality in 2003. 5th ed. Geneva, Switzerland, World Health Organization.
2. PRB, Population Reference Bureau 1997 The world's youth 1996. Washington, DC.
3. WHO 2004 Unsafe abortion: Global and regional estimates of the incidence of unsafe abortion and associated mortality in 2000. 4th ed. Geneva: World Health Organization.
4. Ministry of Health, 2006. Technical and Procedural Guidelines for Safe Abortion Services in Ethiopia.. Addis Ababa.
5. Population Reference Bureau (PRB), 2008 *World Population Data Sheet,* Washington PRB, 2008.
6. Desta Shamebo. 1994. The Butajira Rural Health Project in Ethiopia: Epidemiological Surveillance for Research and Intervention in Primary Health Care. The Ethiopian Journal of Health Development , Vol. 8, Special Issue.
7. Marie Stopes International Ethiopia. MSIE 2008 Annual Report. Ethiopia.
8. Raghavan, vijav (1998) a perspective of data mining journal of American society for information science: 49 (5)
9. ShegawA.Application of data Mining technology to predict child mortality patterns the case of Butajura rural health project.2002
10. Dons, stephen M. and Wallace, Michael W. 2000. Mining Association Rules from a pediatric primary care Decision support system. Available URL: www.amia.org/pubs/symposia/D200658.PDF
11. Wilcox AJ, Horney LF. Accuracy of spontaneous abortion recall. *American Journal of Epidemiology*, 1984, 120(5):727-733.
12. Ministry of Health/World Health Organization (MOH/WHO). 2003. Reduce Model: An advocacy tool for accelerated reduction of maternal and newborn morbidity and mortality in Ethiopia. Addis Ababa, Ethiopia.
13. WHO, UNICEF, UNFPA, and the World Bank, Trends in Maternal Mortality estimates: 1990–2008:

14. Ipas ; Facts on Unintended Pregnancy and Abortion in Ethiopia April 2010, new york
15. Susheela Singh and Tamara Fetters The Estimated Incidence of Induced Abortion in Ethiopia, 2008
16. Last, Mark and Kandel, Abraham. 2002. Automated Perceptions in Data Mining. Available URL : http://www.csee.usf.edu/~mlast/papers/perc_f1.pdf
17. Overview of Addis Ababa city solid waste management system. (2010). Addis Ababa, Ethiopia.
18. Ian H. Witten, Eibe Frank. (2005). Data Mining practical Machine Learning tools and Techniques. Second Edition, Morgan Kaufmann publishers
19. Han, Jiawei and MichelineKamber(2006), Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.
20. Liu, Bing et al (1997). Using General Impressions To Analyze Discovered Classification Rules. Department Of Information Systems And Computer Science. Available from: http://citeseer.nj.nec.com/liu97using.html
21. Iqbal Shah and Elisabeth Ahman, “Unsafe Abortion in 2008: Global and Regional Levels and Trends,” *Reproductive Health Matters* 18, no. 35 (2010).
22. David Hand, HeikkMannila and Padhraic Smyth . (2001). Principles of Data Mining. The MIT Press
23. Two Crows Corporation, 2005:Introduction to Data Mining and Knowledge Discovery, Third Editionwww.twocrows.com
24. Levin, Nissan and Zahavi, Jacob, 1999. Data Mining. Available URL: www.urbanscience.com/Data Mining.pdf
25. Richard D. De Veaux. (2009). Successful Data mining in practice. Williams College.
26. Trybula, Walter J. 1997. Data Mining and Knowledge Discovery: Annual review of Information Science and Technology (ARIST); (32): 197 – 229.
27. Larvac, Nada. 1998. Data Mining in Medicine : Selected Techniques and Applications. Available URL.:http://citeseer.nj.nec.com/lavrac98data.html

28. Prather, Jonathan C. *et. al*. 2001. Medical Data Mining : Knowledge Discovery in a clinical data available URL: http://www.amia.org/pubs/symposia/D004394.PDF
29. Wilcox AJ, Horney LF. Accuracy of spontaneous abortion recall. *American Journal of Epidemiology*, 1984, 120(5):727-733.
30. Jones EF, Forrest JD. Under-reporting of abortion in surveys of U.S. women: 1976 to 1988. *Demography*, 1992, 29(1):113-126.\
31. Udry RJ, Gaughan M, Schwingl PJ, van den Berg BJ. A medical record linkage analysis of abortion underreporting. *Family Planning Perspectives*, 1996, 28(5):228-231.
32. Population Reference Bureau, 2011, abortion facts and figures. Available URL: www.prb.org
33. Gold RB. *Abortion and women's health. A turning point for America?* New York and Washington, DC, The Alan Guttmacher Institute, 1990.
34. *Sedgh G et al., Induced abortion: incidence and trends worldwide from 1995 to 2008, Lancet, 2012, (forthcoming), and the World Health Organization.*
35. Åhman E, Shah IH, Mathers C. Mortality and morbidity due to unsafe abortion. *(Unpublished).*
36. World Health Organization, *Unsafe Abortion: Global and Regional Estimates of the Incidence of Unsafe Abortion and Associated Mortality in 2008*, 6th ed. (2011).
37. Selamawit Negash: 2008. COST ANALYSIS OF ABORTION IN ADDIS ABABA PUBLIC HOSPITALS
38. Berhan Y and Abdela A, Emergency obstetric performance with emphasis on operative delivery outcomes: does it reflect the quality of care? Ethiopian Journal of Health Development, 2004, 18(2):96–106.
39. World Health Organization (WHO), Maternal Mortality in 2005: Estimates Developed by WHO, UNICEF, UNFPA and the World Bank, Geneva: WHO, 2005.

40. Mekbib T, Gebrehiwot Y and Fantahun M, Survey of unsafe abortion in selected health facilities in Ethiopia, Ethiopian Journal of Reproductive Health, 2007, 1(1):28–43.
41. Ethiopian Society of Obstetricians and Gynecologists (ESOG), A Data Base on Abortion Literature Review, Addis Ababa, Ethiopia: ESOG, 2000.
42. Elias Senbeto et al,Prevalence and associated risk factors of Induced Abortion in northwest Ethiopia, [*Ethiop.J.Health Dev.* 2005;19(1) 37-44]
43. WHO Library,2008: global and regional estimates of the incidence of unsafe abortion and associated mortality in 2008. -- Library Cataloguing-in-Publication Data Unsafe abortion 6th ed.
44. Witten, Ian H. and Frank, Eibe. 2000. Practical Machine Learning Tools and Techniques with Java Implementations. USA: Academic Press.
45. Hipp, Jochen and et. al., 2000. Algorithms for Association Rule Mining: A General Survey and Comparison. Available at: http://www.cs.sfu.ca/coursecentral/884/G2/2002-3/references/high00.pdf
46. Sumana Sharma, Kweku-MuataOsei-Bryson. (2008). Framework for formal implementation of the business understanding phase of data mining projects; Virginia Commonwealth University, United States, Elsevier Ltd
47. Daniel T. Larose. (2005). Discovering Knowledge in Data; An introduction to Data Mining. A John Wiley & Sons, Inc. Publication
48. Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf HartmutGüting, Jaiwei Han, Xia Jiang, MichelineKamber, Sam S. Lightstone, Thomas P. Nadeau, Richard E. Neapolitan, Dorian Pyle, MamdouhRefaat, Markus Schneider, Toby J. Teorey, Ian H. Witten. (2009). Data Mining Know it All. Morgan Kaufmann Publishers

Appendix 1

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: abortion_simbolic-weka.filters.unsupervised.attribute.Remove-R8

Instances: 5361

Attributes: 7

agenew

educ

marst

occupa

PIA

religions

RIAP

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.2 (1072 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 13

Size of set of large itemsetsL(2): 24

Size of set of large itemsetsL(3): 13

Size of set of large itemsetsL(4): 3

Best rules found:

1. occupa=1 RIAP=7 1340 ==>marst=1 1340 <conf:(1)> lift:(1.49) lev:(0.08) [442] conv:(442.67)

2. occupa=1 PIA=0 RIAP=7 1295 ==>marst=1 1295 <conf:(1)> lift:(1.49) lev:(0.08) [427] conv:(427.8)

3. occupa=1 religions=1 1281 ==>marst=1 1281 <conf:(1)> lift:(1.49) lev:(0.08) [423] conv:(423.18)

4. occupa=1 PIA=0 religions=1 1235 ==>marst=1 1235 <conf:(1)> lift:(1.49) lev:(0.08) [407] conv:(407.98)

5. occupa=1 RIAP=1 1196 ==>marst=1 1196 <conf:(1)> lift:(1.49) lev:(0.07) [395] conv:(395.1)

6. occupa=1 PIA=0 RIAP=1 1145 ==>marst=1 1145 <conf:(1)> lift:(1.49) lev:(0.07) [378] conv:(378.25)

7. occupa=1 2761 ==>marst=1 2756 <conf:(1)> lift:(1.49) lev:(0.17) [907] conv:(152.02)

8. occupa=1 PIA=0 2660 ==>marst=1 2655 <conf:(1)> lift:(1.49) lev:(0.16) [873] conv:(146.45)

9. PIA=0 RIAP=1 1445 ==>marst=1 1418 <conf:(0.98)> lift:(1.47) lev:(0.08) [450] conv:(17.05)

10. RIAP=1 1511 ==>marst=1 1474 <conf:(0.98)> lift:(1.46) lev:(0.09) [462] conv:(13.14)

11. occupa=1 RIAP=7 1340 ==> PIA=0 1295 <conf:(0.97)> lift:(1.05) lev:(0.01) [58] conv:(2.26)

12. marst=1 occupa=1 RIAP=7 1340 ==> PIA=0 1295 <conf:(0.97)> lift:(1.05) lev:(0.01) [58] conv:(2.26)

13. occupa=1 RIAP=7 1340 ==>marst=1 PIA=0 1295 <conf:(0.97)> lift:(1.5) lev:(0.08) [432] conv:(10.39)

14. educ=1 marst=1 1183 ==> PIA=0 1143 <conf:(0.97)> lift:(1.05) lev:(0.01) [51] conv:(2.24)

15. occupa=1 religions=1 1281 ==> PIA=0 1235 <conf:(0.96)> lift:(1.05) lev:(0.01) [53] conv:(2.11)

16. marst=1 occupa=1 religions=1 1281 ==> PIA=0 1235 <conf:(0.96)> lift:(1.05) lev:(0.01) [53] conv:(2.11)

17. occupa=1 religions=1 1281 ==>marst=1 PIA=0 1235 <conf:(0.96)> lift:(1.5) lev:(0.08) [410] conv:(9.72)

18. occupa=1 2761 ==> PIA=0 2660 <conf:(0.96)> lift:(1.04) lev:(0.02) [113] conv:(2.1)

19. marst=1 occupa=1 2756 ==> PIA=0 2655 <conf:(0.96)> lift:(1.04) lev:(0.02) [112] conv:(2.1)

20. marst=1 religions=1 1743 ==> PIA=0 1677 <conf:(0.96)> lift:(1.04) lev:(0.01) [69] conv:(2.02)

Appendix 2

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: abortion_simbolic-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R3

Instances: 5361

Attributes: 6

agenew

educ

occupa

PIA

religions

RIAP

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (536 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 17

Size of set of large itemsetsL(2): 41

Size of set of large itemsetsL(3): 24

Size of set of large itemsetsL(4): 1

Best rules found:

1. agenew=3 occupa=1 855 ==> PIA=0 835 <conf:(0.98)> lift:(1.06) lev:(0.01) [46] conv:(3.16)

2. educ=1 occupa=1 985 ==> PIA=0 955 <conf:(0.97)> lift:(1.05) lev:(0.01) [46] conv:(2.47)

3. occupa=1 religions=2 610 ==> PIA=0 590 <conf:(0.97)> lift:(1.05) lev:(0.01) [27] conv:(2.25)

4. occupa=1 RIAP=7 1340 ==> PIA=0 1295 <conf:(0.97)> lift:(1.05) lev:(0.01) [58] conv:(2.26)

5. occupa=1 religions=1 1281 ==> PIA=0 1235 <conf:(0.96)> lift:(1.05) lev:(0.01) [53] conv:(2.11)

6. occupa=1 2761 ==> PIA=0 2660 <conf:(0.96)> lift:(1.04) lev:(0.02) [113] conv:(2.1)

7. occupa=1 RIAP=1 1196 ==> PIA=0 1145 <conf:(0.96)> lift:(1.04) lev:(0.01) [41] conv:(1.78)

8. RIAP=1 1511 ==> PIA=0 1445 <conf:(0.96)> lift:(1.04) lev:(0.01) [51] conv:(1.75)

9. occupa=1 religions=1 RIAP=1 656 ==> PIA=0 625 <conf:(0.95)> lift:(1.03) lev:(0) [19] conv:(1.59)

10. agenew=2 occupa=1 910 ==> PIA=0 865 <conf:(0.95)> lift:(1.03) lev:(0) [25] conv:(1.54)

11. educ=3 895 ==> PIA=0 850 <conf:(0.95)> lift:(1.03) lev:(0) [24] conv:(1.51)

12. agenew=2 RIAP=1 580 ==> PIA=0 550 <conf:(0.95)> lift:(1.03) lev:(0) [15] conv:(1.45)

13. educ=5 occupa=1 576 ==> PIA=0 545 <conf:(0.95)> lift:(1.03) lev:(0) [13] conv:(1.4)

14. religions=1 RIAP=1 831 ==> PIA=0 785 <conf:(0.94)> lift:(1.02) lev:(0) [18] conv:(1.37)

15. agenew=3 1560 ==> PIA=0 1445 <conf:(0.93)> lift:(1) lev:(0) [6] conv:(1.04)

16. religions=2 1145 ==> PIA=0 1060 <conf:(0.93)> lift:(1) lev:(0) [3] conv:(1.03)

17. religions=3 950 ==> PIA=0 875 <conf:(0.92)> lift:(1) lev:(0) [-1] conv:(0.97)

18. educ=4 820 ==> PIA=0 755 <conf:(0.92)> lift:(1) lev:(0) [-1] conv:(0.96)

19. religions=1 2586 ==> PIA=0 2380 <conf:(0.92)> lift:(1) lev:(0) [-5] conv:(0.97)

20. agenew=2 1545 ==> PIA=0 1420 <conf:(0.92)> lift:(1) lev:(0) [-5] conv:(0.95)

Appendix 3

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: abortion_simbolic-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 5361

Attributes: 39

agenew=1

agenew=2

agenew=3

agenew=4

agenew=5

agenew=6

educ=1

educ=2

educ=3

educ=4

educ=5

educ=6

marst=1

marst=2

marst=3

marst=4

marst=5

occupa=1

occupa=2

occupa=3

occupa=4

PIA=0

PIA=1

PIA=2

PIA=3

PIA=4

PIA=5

religions=1

religions=2

religions=3

religions=4

religions=5

RIAP=1

RIAP=2

RIAP=3

RIAP=4

RIAP=5

RIAP=6

RIAP=7

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.95 (5093 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 15

Size of set of large itemsetsL(2): 72

Size of set of large itemsetsL(3): 151

Size of set of large itemsetsL(4): 145

Size of set of large itemsetsL(5): 72

Size of set of large itemsetsL(6): 18

Size of set of large itemsetsL(7): 1

Best rules found:

1. RIAP=2=0 5196 ==> PIA=4=0 5196 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.85)

2. marst=3=0 5195 ==> PIA=4=0 5195 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.85)

3. marst=3=0 5195 ==> PIA=5=0 5195 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.85)

4. marst=3=0 PIA=5=0 5195 ==> PIA=4=0 5195 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.85)

5. marst=3=0 PIA=4=0 5195 ==> PIA=5=0 5195 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.85)

6. marst=3=0 5195 ==> PIA=4=0 PIA=5=0 5195 <conf:(1)> lift:(1) lev:(0) [9] conv:(9.69)

7. PIA=5=0 RIAP=2=0 5191 ==> PIA=4=0 5191 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.84)

8. PIA=3=0 RIAP=2=0 5171 ==> PIA=4=0 5171 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

9. marst=3=0 PIA=3=0 5170 ==> PIA=4=0 5170 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

10. marst=3=0 PIA=3=0 5170 ==> PIA=5=0 5170 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

11. marst=3=0 PIA=3=0 PIA=5=0 5170 ==> PIA=4=0 5170 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

12. marst=3=0 PIA=3=0 PIA=4=0 5170 ==> PIA=5=0 5170 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

13. marst=3=0 PIA=3=0 5170 ==> PIA=4=0 PIA=5=0 5170 <conf:(1)> lift:(1) lev:(0) [9] conv:(9.64)

14. marst=5=0 RIAP=2=0 5166 ==> PIA=4=0 5166 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

15. PIA=3=0 PIA=5=0 RIAP=2=0 5166 ==> PIA=4=0 5166 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

16. marst=3=0 marst=5=0 5165 ==> PIA=4=0 5165 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

17. marst=3=0 marst=5=0 5165 ==> PIA=5=0 5165 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

18. marst=3=0 marst=5=0 PIA=5=0 5165 ==> PIA=4=0 5165 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

19. marst=3=0 marst=5=0 PIA=4=0 5165 ==> PIA=5=0 5165 <conf:(1)> lift:(1) lev:(0) [4] conv:(4.82)

20. marst=3=0 marst=5=0 5165 ==> PIA=4=0 PIA=5=0 5165 <conf:(1)> lift:(1) lev:(0) [9] conv:(9.63)

Appendix 4

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: abortion_simbolic-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R13-17

Instances: 5361

Attributes: 28

agenew=1

agenew=2

agenew=3

agenew=4

agenew=5

agenew=6

educ=1

educ=2

educ=3

educ=4

educ=5

educ=6

occupa=1

occupa=2

occupa=3

occupa=4

PIA=0

PIA=1

PIA=2

PIA=3

PIA=4

PIA=5

religions=1

religions=2

religions=3

religions=4

religions=5

RIAP

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.95 (5093 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 7

Size of set of large itemsetsL(2): 14

Size of set of large itemsetsL(3): 16

Size of set of large itemsetsL(4): 7

Size of set of large itemsetsL(5): 1

Best rules found:

1. PIA=5=0 5356 ==> PIA=4=0 5351 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
2. PIA=4=0 5356 ==> PIA=5=0 5351 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
3. PIA=3=0 5336 ==> PIA=4=0 5331 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
4. PIA=3=0 5336 ==> PIA=5=0 5331 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
5. PIA=3=0 PIA=5=0 5331 ==> PIA=4=0 5326 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
6. PIA=3=0 PIA=4=0 5331 ==> PIA=5=0 5326 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
7. PIA=2=0 5311 ==> PIA=4=0 5306 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
8. PIA=2=0 5311 ==> PIA=5=0 5306 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.83)
9. PIA=2=0 PIA=5=0 5306 ==> PIA=4=0 5301 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
10. PIA=2=0 PIA=4=0 5306 ==> PIA=5=0 5301 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
11. PIA=2=0 PIA=3=0 5286 ==> PIA=4=0 5281 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
12. PIA=2=0 PIA=3=0 5286 ==> PIA=5=0 5281 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
13. PIA=2=0 PIA=3=0 PIA=5=0 5281 ==> PIA=4=0 5276 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
14. PIA=2=0 PIA=3=0 PIA=4=0 5281 ==> PIA=5=0 5276 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.82)
15. educ=6=0 5181 ==> PIA=4=0 5176 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.81)
16. educ=6=0 5181 ==> PIA=5=0 5176 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.81)

17. educ=6=0 PIA=5=0 5176 ==> PIA=4=0 5171 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.8)

18. educ=6=0 PIA=4=0 5176 ==> PIA=5=0 5171 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.8)

19. agenew=6=0 5166 ==> PIA=4=0 5161 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.8)

20. agenew=6=0 5166 ==> PIA=5=0 5161 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.8)