ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TO PREDICT THE LIKELIHOOD OF CONTRACEPTIVE METHOD USE AMONG WOMEN AGED 15-49

BY

ABRAHAM GEBREGIORGIS

JUNE, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TO PREDICT THE LIKELIHOOD OF CONTRACEPTIVE METHOD USE AMONG WOMEN AGED 15-49

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

ABRAHAM GEBREGIORGIS

JUNE, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TO PREDICT THE LIKELIHOOD OF CONTRACEPTIVE METHOD USE AMONG WOMEN AGED 15-49

BY

ABRAHAM GEBREGIORGIS

JUNE, 2012

Name and signature of Members of Examining Board

Name	Title	Signature	Date
Dr. Million Meshesah	Advisor	_____	_____
Dr. Wubegzier Mekonnen	Advisor	_____	_____
Dr. Alemayehu Mekonnen	int. examiner	_____	_____
Dr. Rahel Bekele	Ext.examiner	_____	_____

Declaration

I declare that this thesis is my original work and has not been presented for a degree in any university.

_____

Abraham Gebregiorgis

June, 2012

This thesis has been submitted for examination with my approval as university advisor.

_____

Million Meshesha(phD)

_____

Wubegzier Mekonnen (phD)

ACKNOWLEDGEMENT

First I would like to praise the glory of God for giving me strength and hope not only to carry on my study but also for keeping me walk in the bad times.

I am truly thankful to Dr. Million Meshesha, my advisor for his endless support, effort and encouragement. It is really a pleasure to work with him.

I also would like to thank Dr. Wubegzier Mekonnen, my advisor for his support and guidance.

I sincerely appreciate for the good communication I had with Meseret Ayano.

My sincere gratitude goes to my friends for the discussion I had with them and good comments we interchange.

My family and old friend's thank you for being the spice of my life.

TABLE OF CONTENTS

LIST OF FIGURES

Pages

LIST OF TABLES

Pages

ACRONYMS /ABBREVIATIONS

CM	Contraceptive Method
CPR	Contraceptive Prevalence Rate
CSA	Central Statistics Agency
DM	Data Mining
EDHS	Ethiopia Demographic Health Survey
EMOH	Ethiopian Ministry of Health
FGAE	Family Guidance Association of Ethiopia
FMOH	Federal Ministry of Health
FP	Family Planning
HIV/AIDS	Human Immunodeficiency Virus /Acquired immunodeficiency Syndrome
KDD	Knowledge Discovery in Database
PHC	Primary Health Care
RH	Reproductive Health
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

In Ethiopia a gap between knowledge and use of contraceptive method is observed from many studies. According to the 2005 Ethiopian Demographic Health survey report the knowledge about any modern method among women is 86%, Contraceptive Acceptance Rate is 50.1% whereas the Contraceptive Prevalence Rate is 13.9%. Therefore the main objective of this study is to predict the likelihood of contraceptive method use among women aged 15-49 based on demographic, socio-economic, geographic, reproductive history and knowledge factors using data mining classification techniques in order to tackle the barrier factors of contraceptive method use and increase the prevalence of contraceptive method use of those women with low likelihood of practicing family planning by working on the factors.

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

Experimental results show that J48 decision tree performs better than Naïve Bayes. From this model 253 rules are generated. One important rule detected was; women who do not know any contraceptive method have no any chance of using contraceptive method. But having knowledge of contraceptive method could not be a guarantee in order to use contraception. Other factors such as Partner occupation, partner education level, wife's education level, FP message, wealth index, Visit by FP worker were found to be most determinant factors as well. It is therefore recommended all concerned parties to strengthen the promotion of contraceptive method knowledge, improve both partner and wife education level. FP message and Visit by FP workers are also important in increasing CM use.

KEYWORD: Data Mining, Family Planning, Contraceptive Method, Decision Tree, Naïve Bayes

CHAPTER ONE

INTRODUCTION

1.1. Background

Ethiopia is one of the most populous countries in the world with 73.8 million people, of which 84% of them resides in rural areas and the rest lives in urban areas.

Ethiopia is one of the least developed countries with the Per-capita income of 102 USD. The economy of the country is mostly depends on agriculture. The distribution of female and male is 49.5 and 50.5 respectively. The population is also ethnically diversified .There are 80 ethnic groups with different cultures and religions. Orthodox, Muslim, protestant constitutes 43.5%, 33.9%, 18.6% respectively and the remaining is attributed by other types of religions. The average age of Ethiopia is 17 years and 60% are under the age 20 which could be said the population is young. From the total population women age 15-49 covers 23.4%.The population density is 67.9/km2 (1, 2).

Even though the fertility rate is decreasing from time to time it is still low which in average each women gives birth to 5.4 children. There are different reasons that contribute for the high rate of fertility .One of the primary reason is low contraceptive use. According to EDHS the contraceptive acceptance rate among Ethiopian Women (that is the desire to use any contraceptive method) is 50.1% where as the contraceptive prevalence rate of women is (the actual use of contraceptive method) 10.3%. The choice of contraceptive method is also different .There are few types of contraceptive method used by most women. Injectables hold the highest rate with 9.9% followed by pill 3.1%, with female sterilization, IUCD, implants, condom, LAM each constituting 0.2% (1, 3).

A gap between knowledge and use of contraceptive method is observed from many studies. The Studies conducted so far have used traditional statistical analysis which formulates a hypothesis and test the validity on the data set. While data mining is applied on large dataset and its intention is not to test a hypothesis rather it tries to discover if there are hidden patterns and relationships from a large amount of data with no prior assumption. It further predicts the

likelihood of a certain phenomenon from a previous collected and trained data. This study then aims to reach each woman by classifying them as actual or non actual users of contraceptive method based on previously collected and trained data. This helps to adjust the counseling and teaching methods to raise awareness and tackle the factors that hinder them not to use.

1.1.1. Family Planning Program

Family planning is a mechanism for limiting the size of the family and spacing the pregnancy through the use of either traditional or modern contraceptive method voluntarily. It allows parents to arrange and gain their desired number of children, to space and limit their births. Contraceptive method is one way which enables to attain the goal of family planning .Other services like treatment of infertility is also included under the service of family planning (4, 5).

Spacing or limiting the number of pregnancy has an impact on the wellbeing of the mother, the child and on the economy of the family and the country as a whole. Maternal mortality, infant and child mortality, unwanted pregnancy, abortion and post abortion complications can be minimized by an effective use of contraceptive method. Use of methods like condom can as well prevent the transmission of HIV/AIDS and other sexually transmitted diseases (4, 5).

Different strategies which incorporate family planning were implemented. In 1970's the government of Ethiopia has implemented the PHC which is one of the four layer of the health care sector delivery system established in 1998.The PHC is formulated to deliver services related to maternal and child care, immunization, family planning, nutritional health and micronutrient supplements to 2500 people in order to make an equitable distribution of health service. But this has limitation in achieving the desired goal. To alleviate the shortcoming in the implementation of the PHC the health extension program has come into existence. HEP is a community based program creating awareness and providing health service at the lowest level of the community. Two females trained for one year are recruited and deployed to each kebele to work on prevention of diseases and promotion of health service. The services provided in the health extension program are categorized in to four major areas; diseases prevention and control; family health, hygiene, environmental sanitation; and health education and communication (5, 6).

In 1993 population policy of Ethiopia was then formulated to narrow the gap between economic development and population growth rate aimed at reducing fertility through utilization of family planning services and promoting socio-economic development to reduce the number of fertility which was 7.7 per each woman at that time to 4 and raise contraceptive prevalence rate to 44 percent in the year 2015(7, 8). The population of Ethiopia was increasing rapidly due to un planned pregnancy and the low awareness of family planning (9). In fact one of the main targets of family planning is to limit the growth of population so that the population of the country grows in line with growth of its economy and the available resources.

So many reasons lead for the formulation of the 1993 population policy of Ethiopia. It is shown in many countries that without limiting the growth of population it is difficult to achieve the millennium development goal, which strives to improve the living standard and life with respect to the universal strategy. It is also evident that the population growth has outpaced the economic growth. It is difficult to get developed if the population growth is not controlled. Being an agrarian country, with large population size and poor resource utilization and management the problem has worsened (8, 10).

The 2005-2014 reproductive health framework has adopted by the ministers of health of the different countries worldwide through WHO with collaboration of its partners. The framework was adopted to incorporate family planning in the reproductive health service and make it national agenda due to the accreditation of its importance. Family planning has get recognition because it plays a key role in achieving the health and development targets of the MDGs 4, 5, and 6 (4).

Ethiopian Ministry of Health (EMOH), USAID (Unites States Agency for International Development), UNFPA (United Nations Population Fund) and other organizations has met to discuss about the issue of improving the 1993 population policy and create a RH framework in early 2004. The RH framework was intended to include family planning, safe mother hood, post abortion care, adolescent RH, and HIV/AIDS. The RH strategy was primary developed with the purpose of three main priorities. The first priority is to assist governments in their effort to achieve the Millennium development goals which intends to make a sustainable development and eliminate poverty. In the 2006 RH strategy, of the 8 MDGs improving maternal health,

promoting gender equality and fighting against HIV/AIDS was the core targets. In addition these helps to address the variables of RH domains such as fertility, gender, age at first birth, contraceptive prevalence, method choice, traditional practices, literacy and other factors with direct implications for the health of men and women throughout their lifecycles. The second is the need to respond to the socioeconomic and demographic realities of Ethiopia today. The third is to build on the notable advances realized in the health sector over the past decade (2, 11).

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

Family planning enhances efforts to improve family health by limiting the number of family and spacing the pregnancy. However, traditional beliefs, religious barriers and lack of male involvement, low level of woman education with less empowerment, lack of knowledge, less income and other barriers have weakened family planning interventions. It is also confirmed that there is high unmet need for family planning in Ethiopia. When a woman needs to limit her family size or space the pregnancy but do not use any method to do this it is called unmet need (4).

1.1.2. Data Mining

The world is producing an incredible amount of data every day due to the progress in computer power and storage capacity, the increase of transaction and the availability of tools for capturing these huge data. However, the ability to analyze these data did not matured to pace with the increased production of data. For extracting relevant information rapidly and making decision from the massive, inconsistence and noise data it is essential to make analysis. Today, a number of software tools like SPSS, WEKA, and SAS etc are employed to help analysts to organize their data, generate overviews and explore the information space in order to extract potentially useful information (13, 14).

It is attracting an interest and becoming a necessity to understand large, complex, information-rich data sets in the different fields of business, healthcare, hospitals, pharmaceuticals and other health domains due to the fact that customer and transactional data are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including machine learning techniques, for discovering knowledge from data is called data mining (12).

Data mining is an iterative and discovery driven process which it doesn't put a priori assumption about what will constitute the outcome. It is the search for new, valuable and non-trivial information from a large dataset (10).

The two major task of data mining are prediction and description. In the case of description the task is to identify patterns or relations in the data. It describes and examines the property of the data as it is .Association; clustering and summarization are some of the methods used in creating descriptive model. whereas in the prediction there is a predetermined class label and in addition to finding different kinds of structures and relations in the data, further it derives rules and models that enable prediction and decision making in new situations. The techniques employed here have the capability to foretell about the probability of happening a certain phenomena or future values based on previously collected and trained data (14). Based on the analysis of a set of training data (data objects whose class label is known) a prediction models are constructed. The derived model may be represented in various forms, such as if-then rules, decision trees, mathematical formulae, or neural networks (13). Some of the techniques applied in prediction model formulation are Naïve Bayes, support vector machine, decision tree, neural network etc.

Data mining has emerged from different disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization and information retrieval. Data mining is an analysis tool which may uncover an important and previously unknown and overlooked data pattern. This uncovered hidden knowledge contributes for business, scientific and medical strategy. The observed gap between data and information has made to look for an improved tool in order to mine important knowledge from a large amount of data (13)

1.2. Statement of the Problem

Greater use of family planning services can contribute directly to the Millennium Development Goals (MDGs) to reduce child mortality and improve maternal health; family planning helps to reduce the number of high-risk pregnancies that result in high levels of maternal and child illness and death (15).

High fertility and rapid population growth have an impact on the overall socio-economic development of a country in general and maternal and child health in particular. Maternal and child mortality are two of the major health problems challenging healthcare organizations, especially in developing countries. The majority of maternal deaths are the direct result of complications encountered during pregnancy and arising from unsafe terminations (16, 17, 18).

Ethiopia is among countries with low contraceptive prevalence rate. Because of the low prevalence level of contraceptive use and higher total fertility rate in developing countries including Ethiopia, unwanted pregnancy and births are increasing which affect the maternal and child health (19). Many of the unwanted pregnancies end with abortion. Unsafe abortion remains a significant cause of maternal morbidity and mortality in many of the developing countries (20).According to the World Health Report (21) unwanted and unintended pregnancy are the most common cause of maternal mortality in developing countries.

Women are believed to have high awareness of contraceptive method but researches (3, 6, 9, 10) showed that the use of any method is still low compared with the knowledge they have about any method. According to EDHS and other studies the gap between the knowledge and use of contraceptive method is high. The knowledge about any modern method among women is 86%, Contraceptive Acceptance Rate is 50.1% whereas the Contraceptive Prevalence Rate is 13.9% (3).

Some of the Studies done in Ethiopia so far was used simple statistical techniques such as regression on a limited set of data to assess the factors which contribute to the low utilization of contraceptive method. The analysis made by using such traditional methods focus on problems with a manageable number of variables and cases they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in dataset. Studies like that of EDHS on

the other hand collected large amount of data in a population based and they employed The Census and Survey Processing System (CSPro) for data entry, editing, tabulation, and dissemination.

On the other hand data mining handle large amounts of data and not simply perform data or information retrieval (finding aggregate values, perform deductive query) and its intention is not to test a hypothesis rather it tries to discover if there are hidden patterns and relationships from a large amount of data with. It is not concerned with those areas of statistics involving how best to collect the data in the first place so as to answer a specific question, such as experimental design and survey design. Data mining essentially assumes that the data have already been collected, and is concerned with how to discover patterns. The data sets often encountered in data mining are huge by statistical standards. The well-established statistical tools may fail under such circumstances. Data mining also enables to generate knowledge and predict the likelihood of a certain phenomena from a previous or historical data using techniques like decision tree, Naive Bayesian method and neural network, etc.

This study is one possible way of showing the application of data mining in health care data. The EDHS data is used for this purpose. The main reason of this is because there is no organization which has captured much data related to respondent's background. Collecting relevant data that best describes the population with different variables is important during service provision in order to make analysis in the future and create a prediction model. For instance, it is possible to identify or predict whether a woman continues or discontinues using contraceptive method if relevant data are recorded. But the case is different in the health care providers of Ethiopia .As to the best knowledge of the researcher health care providers do not even ask why they quit and there is no way of knowing why they might stop coming.

It is therefore the aim of this study is to apply data mining classification techniques to discover patterns (if there are common characteristics) in the contraceptive method users and non users. So that based on the identified possible factors that determine the utilization of contraceptives and the identified target groups (women with low likelihood of utilizing contraceptive) program managers can design programs and help for proper implementation with regard to increasing utilization of family planning.

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

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

1.3. Objective of Study

1.3.1. General Objective

The main objective of this study is to predict the likelihood of contraceptive method use among women aged 15-49 in Ethiopia in order to tackle the barrier factors of contraceptive method and increase CPR.

1.3.2. Specific Objectives

To achieve the general objective the following specific objectives are attempted.

- To identify the factors of contraceptive method use.
- To investigate if there are hidden patterns in the dataset.
- To develop a classification model using data mining techniques that classifies women according to their likelihood of utilization of contraceptive method.
- To assess the applicability of data mining techniques in health care data.

1.4. Scope and Limitation of the study

The scope of this study was restricted to classification based socio economic, demographic, geographic, reproductive history and knowledge factors. Data mining classification techniques such as decision tree and Naïve Bayes are the algorithms used to develop the classification model.

The scope of this research was also restricted to the data collected from the CSA. The DHS data which was collected in 2005 was employed for the study. The age groups included were age is 15 -49.it only includes women, men are out of this scope.

It also covers all regions and the two city administrations (Addis Ababa and Dire Dawa) of Ethiopia.

The unavailability of related literature was one of the limitations encountered to undertake the study. Another limitation of the study was the researcher could not use the DHS 2011 data because it was not accessible and released in the time the researcher has asked for permission.

1.5. Research Methodology of the study

1.5.1. Research Design

In order to obtain an optimal and desired outcome the researcher has followed KDD process model. Different data mining process models are available; CRISP, SEMMA and KDD are the most common data mining process models which are vastly used. CRISP is the de facto industry standard while KDD is most preferable for the academic purpose. KDD is selected for three main reasons. First KDD is best suited for academic purpose; second KDD reduces the skill required for knowledge discovery to the non-experts .Third KDD is independent from any tool and technique so one can use any technique during the study.

The steps of the KDD process are data understanding and data selection, preprocessing, transformation, data mining, evaluation and interpretation of the mined data, using the detected patterns and the useful knowledge discovered to attain our objective.

The steps of KDD are shown in figure 1.1

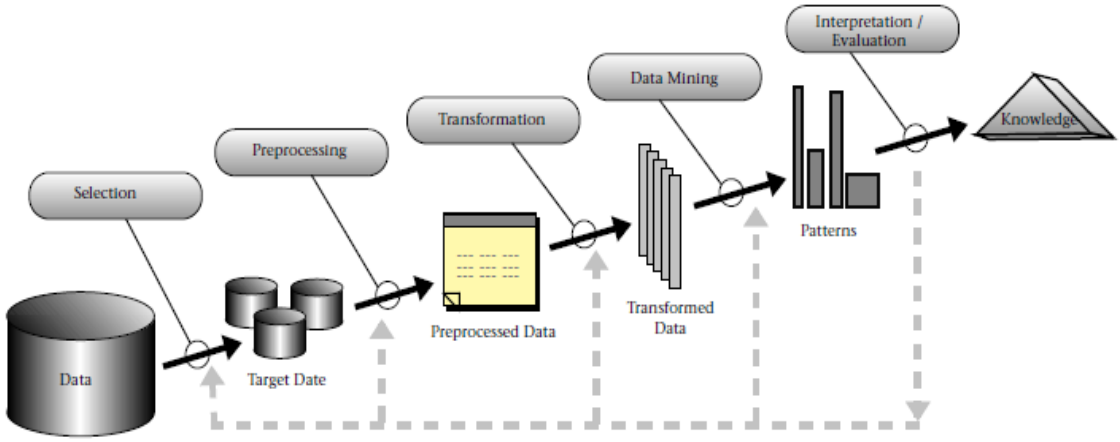

Figure 1.1KDD process model steps

1.5.1.1. Data Understanding and Selection

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with data, to identify data quality problems, to get first insight and detect appropriate subsets of the data being collected.. The researcher used data collected by CSA. In this study the 2005 Ethiopia Demographic Health survey data is used as a source of data. The EDHS collects data related to fertility and family planning behavior, child mortality, adult and maternal mortality, children's nutritional status, the utilization of maternal and child health services, knowledge of HIV/AIDS and prevalence of HIV/AIDS and anemia at the national level . As part of the world DHS program the EDHS was conducted for the first time in 2000. The survey was conducted by MOH and CSA with the cooperation of NGO's. The survey covers 9 regions(Tigray , Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, SNNP, Gambela, Harari) and 2 city administrations (Addis Ababa and Dire Dawa).

One part of the questionnaire is the Women's Questionnaire which has information related to Household and respondent characteristics , Fertility levels and preferences , Knowledge and use of family planning , Childhood mortality , Maternity care ,Childhood illness, treatment, and preventative actions, Anemia levels among women and children , Breastfeeding practices ,

Nutritional status of women and young children , Malaria prevention and treatment , Marriage and sexual activity ,Awareness and behavior regarding AIDS and STIs ,Harmful traditional practices , Maternal mortality from all women age 15-49 years.

Out of the14,070 woman age 15-49 respondents 63.4% of them are married ,25%,6.6% and 4% are never married, divorced and widowed respectively .with respect to the place they live 82.2% of them lives in Rural areas and the rest lives in urban areas. Most of the women are illiterate only 22.2% and 10.5 % has primary and secondary level of education. According to their religion Orthodox, Muslim and protestant holds 49.2%, 28.5%, 18.9% respectively. Oromiya, Amhara, SNNP, Tigray holds the majority of the population with 35.6%, 24.7%, 21.3% and 6.5% respectively (3). To understand the nature of the data, descriptive statistics is used. Based on the appropriateness of the problem domain , document analysis, literatures reviewed in chapter two section 2.3 and 2.4 and based on the information obtained from family planning service providers attributes which are relevant to the study are chosen.

1.5.1.2. Data Preprocessing

One of the most important task in doing research related to data mining is preparing the data in a way suitable for the data mining tools and techniques. The data may have missing attribute values, contain only aggregate value , hold errors or outliers or they can be inconsistent(contain discrepancies in names and codes) therefore this step targets on data preprocessing in order to obtain consistent and clean data. In order to achieve higher efficiency, accuracy, the highest value of data mining and detect knowledge data should be clean and consistent. During this step missing values are filled with either mean or mode value according their data type .Therefore it is checked if there are any noises and redundancy but the data was clean form such problems. Outlier is also checked using the box plot but there was no any outlier in the numeric attributes.

1.5.1.3. Transformation

Once the data is cleaned it was transferred into a data mining capable format. Data might be found in a number of different formats therefore this needs data integration which is detecting and resolving data value conflicts. For the same real world entity different sources might represent it in a different possible representation. During transformation data reduction is also performed either manually or automatically. It obtains reduced representation in volume but

gives the same or similar result of analysis. Different techniques are available for reduction. The most commonly used are lossless and loss full selection. A representative selection can be used to draw conclusions to the entire data.

To make the data more suitable for data mining tasks encoding of some attribute is performed. Discritization is applied to the attributes with continuous value (age and number of living children).

1.5.1.4. Data Mining Techniques

This phase is engaged on searching for patterns of interest in a particular representational form, depending on the DM objective (classification, clustering, association or prediction). The output of this step is a detected pattern.

It is where the techniques of data mining are applied to create a model. Since the goal of the study is to classify actual and non actual contraceptive method users, two techniques of classification are selected (decision tree algorithm and Naïve Bayes classification algorithm). In chapter two it has discussed the reason for choosing these techniques.

Decision tree and Naïve Bayes algorithms are selected based on the literatures reviewed in chapter two. Different experiments are then performed on the unbalanced and balanced data set in WEKA 3.6.3 machine learning software using 10-fold, percentage split test modes with their default and adjusted parameter. Best first and gain ration attribute evaluation methods are also adopted to examine if there is any improvement in the performance of the hired algorithm.

1.5.1.5. Interpretation/Evaluation of the Discovered Knowledge

After mining the required pattern the interpretation and evaluation of the mined patterns is accomplished. The interpretation is concerned with whether the detected pattern is interesting or not. It verifies whether it has knowledge or not. Its responsibility is to represent the result in an appropriate way.

The performance of the algorithms adopted in the study are measured and evaluated based on their accuracy, recall and precision. Rules were then generated from the preferred model. Detecting interesting rules and interpreting these are carried out.

The number of leaves and size of tree is also considered in order to compare the experiments performed using the J48 decision tree algorithm.

1.5.2. Tools

The data preparation is done using SPSS, Excel and WEKA. After preparing the data in a form which is suitable for data mining techniques then it was analyzed using the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a collection of machine learning algorithms implemented in java. It provides techniques which helps to achieve data mining tasks. A graphical interface is also available that enable to load data, preprocess, apply the data mining techniques and visualize the model as a rule or graphically. Applications written using the WEKA class libraries can be run on any computer with a Web browsing capability; this allows to apply the machine learning techniques on our data regardless of the computer platform.

1.6. Ethical Consideration and Result Dissemination

A research is basically conducted to alleviate problems faced by the society. When an investigator conducts a research it might involve the society either directly or indirectly. Especially in the health related issues much of the data required to be gathered has a relation with the privacy of the respondents or the people who are considered in the investigation. It needs great care in order not to violate their privacy and protects them from any harmful deeds which might happen during the investigation. Making consent with the respondent or organization involved in the study is also necessary before rushing into the task. In this thesis the 2005EDHS is used as the source of data and a paper from Addis Ababa University School of Information Science has received to acquire it and for any other cooperation's and then a contract form is filled in CSA. Agreements such as not to work less than the minimum expected work and not to handover the data to third party and other agreements are made. Finally Ethical clearance is obtained from AAU School of Public Health for the approval of the proposal.

In order to disseminate the result of this study ;The final outcome will be given to the School of Public Health and School of Information Science .A copy of it will be send to the CSA. The softcopy will also be available on the e-resource of AAU website. It will be also attempted to publish it in different journals.

1.7. Significance of the Study

Data mining has shown an encouraging result by discovering hidden knowledge in a different health care data. Studies indicate about the gap between knowledge and use of contraceptive method. These studies have showed the correlation or association of the different factors of socio economic, geographic, reproductive history, knowledge etc with the use of contraceptive methods. It is observed from these researches that having contraceptive method knowledge did not guarantee that they will use it. There are other barriers which hinder them not to use any contraceptive method. Focusing on data and background characteristics of respondent is important in order to identify the driven factors of contraceptive method use.

In the EDHS there are women who are current users and woman currently non users. Those women do have characteristics which make them to differ and there might be factors which highly contribute for this difference. It is through these deviations they are grouped into some class. Identification of those women unlikely to use contraceptive method in priori might help organizations which are concerned with family planning related issues to target on those women who have low likelihood of using contraceptive method and on the factors which influences them not to use. A rule is generated which will be applied for the purpose of predicting about other women with such characteristics. This helps organizations in adjusting the counseling and teaching methods of the target group. The finding of this study can also help for policy making, monitoring and evaluating the activities for the government and different concerned agencies such as family planning service providers. It is hoped that this study could contribute to the improvement of family planning services in the country through appropriate service delivery approaches and strategies.

It further helps in resource allocation once we know priori about the woman's probability of using any of the contraceptive method .It is known that data mining is discovery driven that is it

lets the techniques to discover the most determinant factors which contribute for the level of utilizing any method. Organizations such as government hospitals, NGO (Family Guidance association, Marie stop) captures few background variables which are different in each organization. Knowing clients' background and analyzing and further predicting based on the historical data is very important in identifying target groups, determinant factors; in indicating how to work with the target groups and to observe the trends easily if there are any changes. The case is different in Ethiopia. Most Organizations do not focus on data for decision making or any other purposes**.**

1.8. Organization of Thesis

The thesis is organized into six chapters. The first chapter introduces the research by defining the problem that initiates the study, significance and scope of the research. The general and specific objectives of the study and the methodology are also included in this chapter.

The second chapter of the thesis discusses the literatures reviewed. Literatures related to data mining, determinants of family planning practice, application of data mining in health care and application of data mining techniques in contraceptive method data are reviewed.

Techniques and methods used for creating the classification model; Decision tree and Naïve Bayes algorithms the performance measurement used to compare the experiment conducted are discussed in chapter three.

The fourth chapter of the thesis focuses on data understanding and selection, the different methods of preprocessing and transformation performed are discussed in detail.

In chapter five the various parameters adopted in the experiment are also discussed in this section of the thesis. In addition this chapter includes a presentation and interpretation of the rules discovered after different experiments were conducted.

The sixth chapter of the thesis presents the concluding remarks and recommendations that are forwarded on the basis of the outcomes of the experiment.

CHAPTER TWO

2. LITERATURE REVIEW

Due to the huge amount of data and the need to convert this into an important information and knowledge data mining has come into existence. Further this outcome helps in fraud detection, market analysis, customer type classification, diseases diagnosis etc. (13). The advancement of machine learning technologies and algorithms, progress in database technologies, data storage capabilities and parallel computing has contributed for the emergence of data mining. Computers, data and information have become the basis for decision making in many enterprises and we have transferred into information age. Large amount of information about customers, employee, products and transactions has gathered and it needs analysis for decision making and prediction for future (23).

2.1. Data mining Technology

Data mining is an iterative and discovery driven process which doesn't put a priori assumption about what will constitute the outcome. It is the search for new, valuable and non-trivial information from a large dataset (12).

Data mining has emerged from different disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization and information retrieval. Data mining is an analysis tool which may uncover an important and previously unknown and overlooked data pattern. This uncovered hidden knowledge contributes for business, scientific and medical strategy. The observed gap between data and information has made to look for an improved tool in order to mine important knowledge from a large amount of data (13).

The two major task of data mining are prediction and description. In the case of description the task is to identify patterns or relations in the data. It describes and examines the property of the data as it is .Association, clustering and summarization are some of the methods used in creating descriptive model. whereas in the prediction there is a predetermined class label and in addition to finding different kinds of structures and relations in the data, further it derives rules and

models that enable prediction and decision making in new situations. The techniques employed here have the capability to foretell about the probability a certain phenomenon or future values based on previously collected and trained data (23). Based on the analysis of a set of training data (data objects whose class label is known) a prediction models are constructed. The derived model may be represented in various forms, such as if-then rules, decision trees, mathematical formulae, or neural networks (13). The techniques applied in prediction model formulation are Naïve Bayes , support vector machine, decision tree, neural network etc.

A typical data mining can have the following components

- Database, data warehouse or other information repository which are the sources of data for discovering hidden patterns and knowledge.
- Database or data warehouse server which provides users request of data
- Knowledge base, this is responsible for guidance of searching, evaluation
- Data mining engine is where the different functions that used for analysis such as clustering, classification, and association tasks are found in.
- Pattern evaluation module is a component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It may access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be
- Graphical user interface is a module that communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

A data analysis system that does not handle large amounts of data or a system that can only perform data or information retrieval (finding aggregate values, perform deductive query) does not have to be considered as a data mining system. Data mining is a tool for discovering a novel, understandable and important pattern from large datasets by analyzing data from different perspectives (13, 24).

Different data repository are available which data mining is performed on; thus include, relational data base , data warehouse ,transactional database , flat files, advanced database system, and the world wide web, etc. some of these are discussed here below.

Data warehouse: A data warehouse is a collection of information gathered from different sources and put in a centrally located storage area and it is integrated, time variant and subject oriented data repository. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing (13, 24).

Relational database: A relational database is a collection of tables consisting of rows and columns. The columns represent the attributes which contain information of a single type and the row represents some objects or instances. Each record is identified by a unique identification called key. The intersection of a row and column which is the smallest piece of information and retrieved by query language in a table is called field. Data mining can be applied to further search for trends and data patterns on the relational database (13).

Transactional Databases: Transactional database is a database containing records of customer activities. In transactional database often each record represents a single transaction which in most cases containing a unique transaction id, customer account, sales activity, date of activity, items purchased. From this database it would be possible to answer questions like which items sold together or goes together. This question cannot be answered by a regular data retrieval system but it is possible for data mining to answer such questions by identifying frequent item set using association rule (13, 25).

2.2. Data Mining Tasks

The task of data mining could be developing of a descriptive of predictive model.

A model is a high-level description, summarizing a large collection of data and describing its important features. Often a model is global in the sense that it applies to all points in the measurement space. Data mining model building is data-driven. There is no any prior assumption we put, rather the techniques of data mining reveals if there are any relations with in the data set. The existence of relationship does not mean there is casual relationship (example:

someone with a yellow finger has a lung cancer, this doesn't mean yellow finger is the cause of lung cancer but someone who smoke cigarette develops a yellow finger). A model is good if it is general in other term if the characteristics of the generated data are not different to the real world data (26). In general there are two types of model these are:

2.2.1. Descriptive Model

Descriptive model creates a concise and convenient representation of a given data set, which is called descriptive model. Descriptive modeling helps us to look into the important aspect of the data. Association is one typical examples of descriptive model.

Association rule are a form of unsupervised learning. The aim of the association rule is to detect if there are common relationships between objects. Association rule may be applied to categorical or numerical data. The rules are generated based on the concept of support (the percentage of the occurrence of an item) and confidence indicates the strength of the rule. The two steps involving in generating association rule are ; first the frequently occurring items which satisfies the minimum threshold (support) are generated then using the these a rule that satisfies the minimum confidence are generated(19, 27).

Clustering is unsupervised machine learning technique that classifies the data set into unlabeled category and putting those with similar characteristics in one group. Clustering is different from classification and prediction. Unlike predictive and classification it categorizes the dataset into subgroups that are not previously defined.

Clustering examines the nature of the population whether the data falls into distinct groups, with members within each group being similar to other members in that group but different from members of other groups. The outcome of grouping can have a different form. It can be exclusive grouping (one instances belongs to only one group), may overlap (belongs to more than one group), or it may belong to a different group with a probability of .The two common algorithms applied for clustering are hierarchical and partition algorithm. (26, 27, 28).

2.2.2. Predictive model

This type of a model allows not only describing the characteristics of a data set but also enables to tell about the future data values of the population from which the data were draw . The goal of predictive model is to estimate a function from the training data set that can predict a value y given input variables X_i. The predicted variable is called response variable and the input or dependent variables are called explanatory variable. Depending on the data nature of the predicted variable the task of predictive model can be classification (categorical) or regression (real-valued).Predictive model is similar to classification the difference is in case of predictive model the value of predicted variable lies in the future in other word there is no predefined group. For example, predicting the price of a certain item after three months. Whereas in classification there is a target categorical variable which input variable are mapped. Diagnosing whether a particular disease is present or not and determining whether a credit card is fraudulent are examples of classification model type (26, 29).

Classification is a supervised learning i.e. there are pre specified target variables. During classification there is a mapping from some input variable to a categorical variable. The input variables of from training data set are used to build a model that classifies new data into class labels. The most commonly used techniques of classification are Artificial Neural Network, Decision tree, and Naïve Bayes etc. (30).

Different algorithm are applied to achieve the tasks of data mining, common algorithms used for classification are; Artificial neural network, decision tree and Naïve Bayes.

Artificial neural network works by learning the decision boundary surface. It doesn't follow any statistical distribution it is simply generated as a function of human brain. Neural network has nodes arranged in layers. The structure of the layer is different according to the complexity of the problem domain. But in general there are three types of layer; input, hidden and output layer. Weights (initial inputs) are given to the nodes of input layer processed and compared with the actual value and errors are and again feed to the system. This repeats until a minimum error is scored (25, 26, 31).

Decision tree is another type classification algorithm which is one of the oldest techniques used for classification. Decision tree works by dividing variables recursively so that the divided part has greatest gap which majority of the data partitioned belongs to one class. One variable believed to classify the data set into two distinct groups is chosen the classified variables are divided further into some group this goes until a leaf node is encountered. A decision tree is a tree like structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision tree are easy to understand and explain. They handle both continuous and discrete variables. Their prediction speed is fast. One disadvantage of the basic form of tree is that it is monothetic, each node is split on just one variable. Sometimes, in real problems, the class variable changes most rapidly with a combination of input variables. For example, in a classification problem involving two input variables, it might be that one class is characterized by having low values on both variables while the other has high values on both variables The true purpose of a classification tree is to classify the data into distinct groups or branches that create the strongest separation in the values of the dependent variable. Classification trees are very good at identifying segments with a desired behavior such as response or activation. This identification can be quite useful when a company is trying to understand what is driving market behavior. It also has an advantage over regression in its ability to detect nonlinear relationships. This can be very useful in identifying interactions for inputs into other modeling techniques Different algorithms are available in decision tree, CART(Classification and Regression Trees,), Chi-Square Automatic Interaction Detection (CHAW), CLS, ID3, C4.5, CHAID are some of them(25, 26, 27, 31).

Bayesian classification is unsupervised classification technique which is based on the theory of Bayes. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved (13)

2.3. Levels and Determinants of Contraceptive Method Use in Ethiopia

Contraceptive use has increased in many parts of the world .It has increased by 9% globally from 54% to 63% in the year between 1990 and 2007. And the improvement is from 17% to 28% in Africa, 57% to 67% in Asia, and 62% to 72% in Latin America and the Caribbean (4).

In Ethiopia the current use of any contraceptive method among married woman has increased from 4.8 to 14.7% between the year 1990 and 2005. The significance increase has brought by the increase of modern methods (increased from 2.9 to 13.9.Use of modern contraceptive methods has more than doubled from 6.3% of currently married women in the 2000 EDHS to 13.9% in the 2005 EDHS. The most commonly used modern method is Injectables (10 percent), followed by the pill (3 percent) (3).The percentage increase in the use of method is different among regions with the highest increase in Oromiya region and the least is in Somalia .But in Afar region it has lowered by 19% .knowledge about any of the contraceptive method among woman has also increased by 39%, 62% was scored at the time of 1990 but in 2005 it has reached 86%. Pill and Injectables were the most known contraceptive methods by women (3, 33).

CSA (3) also revealed that contraceptive use increases with educational attainment. Women with good level of education have shown better use of contraceptive method. In contrast woman with low education scored minimal use of modern contraceptive method. It also indicated women do not begin to use contraception until they have had at least one child. Contraceptive use also varies significantly according to geographical area where a woman resides has made a difference. Those woman found in the urban area are more likely to use contraceptive method than those woman residing in rural areas. Even if the progress shown in the survey of DHS is higher in the rural area than the urban the prevalence is still very low. This pattern is also reflected in use of contraception by regions. Use of modern contraceptive methods differs significantly with 45 percent in Addis Ababa and 3 percent in Somalia region.

A number of factors contribute for the use of contraceptive method among woman .The choice of contraceptive method is also different because of so many variables. As the knowledge about any method increases, and when parents get their second child; the use of contraceptive method tend to increase. Empowering woman makes a difference in the use and choice of method. Woman at

the age 15-19 have high level of desire using contraceptive method but it declines after the age of 30. Husband's profession, religion also influences woman whether to use or not. Fear of side effects and lack of knowledge holds back woman not to be part of family planning. The level of education, place of residence(urban and rural), accessibility of family planning service, marital status, ethnicity, discussion between couples and other factors are stated in the EDHS and other related studies to have some kind of association with the contraceptive method use (3, 34).

In a population based study of family planning utilization in Mojo town by Abebe and Negatu (35) using cross section method had examined the associations between selected explanatory variables and the main study variable (FP service utilization). In their study they have identified five variables which predicts the likelihood of FP service utilization these are total number of living children, literacy status, women's' approval of FP, spousal communication and exposure to media. The study showed that total number of children and family planning service utilization are strongly related. If women have more children who are living with them, the possibility of using family planning methods for limiting is expected to be high, and if the number of children desired by women is perceived to be 'not enough', they may use family planning methods for spacing purpose. The educational status of women was also found to be significant predictors of family planning services use. Compared to women with no education, literate women have more probability of using contraceptive method. Women who approved family planning services use were also more likely to use family planning than women who disapproved use of family planning. Media exposure is another important determinant of family planning service use. Women who are exposed to any one of the three media, namely, radio, television, or newspapers have higher family planning service use compared to women who have had no media exposure at all.

In a survey of the current level of knowledge, attitudes and practices related to family planning in the Amhara, Oromiya, SNNPR and Tigray regions which constitutes more than 86% of Ethiopian population taken by path finder in 2004 have employed a stratified multistage sampling design .The study reveal that Current use of any method of contraception among all women of reproductive age was highest in Oromiya (24%), followed by Tigray and Amhara with 20% each, and SNNPR at 17%. However, the levels of use for modern methods were lower (21 % in Oromiya, 17% in Tigray, 18% in Amhara, and 14% in SNNPR). The use of contraception

was considerably higher among currently married women in all the four regions, compared to all women of reproductive age. 27% of currently married women in Tigray, 31% in Amhara, 32% in Oromiya, and 23% in SNNPR reported use of any method of contraception. The most widely used and known type of contraceptive method are Temporary methods (pills and Injectables) in both urban and rural areas. The study also identifies the intention of using contraceptive method in the future among the non users and found that in is high in both rural and urban area with 59% and 65% respectively. The main reason respondent's mentioned for not using were inconvenience to obtain and shortage of supplies and demand for more children (36).

A study conducted by Wubegzier M (37) to measure contraception use and unmet need and to identify the determinants among 5746 married women of Butajira used crude and adjusted odds ratio of logistic regression model. The study revealed that contraceptive prevalence rate among married women was is 25.4%, unmet need of contraception is 52.4%. out of this 74.8% was attributed by unmet need for spacing and the remaining was for limiting . About 99% of women in Butajira district knew at least one method of contraception even though some family planning methods were better known than others. Dipo-Provera and Pills were known by more than 97% of married women .The least known modern contraceptive method was foam/jelly. Women who are rural residence are found to have less probability of using contraception compared with urban women. Married women who attend school are more likely to use contraceptive method than those who are at low level of education or who don't attend school. Use of contraception also increases with the increase of their husbands' education level. Discussion with partner was also important for practicing family planning. Even though the desire for children reduces with increasing the number of living children it did not significantly declined.

2.4. Data Mining Application in HealthCare Data

In their study of machine learning classification techniques performance in detection breast cancer Aruna S et.al (38) has applied Naïve Bayes , RBF neural network, decision tree and support vector machine Gaussian RBF kernel on the a breast cancer data set to identify the best predictor and they measure the performance with respect to accuracy, sensitivity and specificity .they employed WEKA software for creating the model.

Yue H et.al (39) has applied data mining techniques (C4.5, B1 and naïve byes) to a diabetic patient database. Their intention was to identify significant factors influencing diabetes control; Predicting individuals in the population with poor diabetes control status based on physiological and examination factors with the intention is to improve the quality of treatment by automating the handling of routine situations, particularly blood glucose control, and ensuring a better quality of life by providing support from expert . Data integration has done to merge separated sources available and data transformation (discretization) is performed in order to make the techniques work on discrete variables. Identification and removal of irrelevant and redundant information in other word feature selection is done to improve the efficiency of the data mining techniques.

Shegaw A. (40) has applied neural network and decision tree models on data set containing 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.

2.5. Related works

Lemaire V et.al.(41) has used Naïve Bayes technique on Indonesia data set to predict the contraceptive method choice (no contraceptive method, short-term contraceptive method or long-term contraceptive method) using the explanatory variables; age, education, husband's education, number of children ever born, religion, working or not, husband's occupation, standard of living index, FP message) .By selecting variables which are considered as possible targets for policies experimented to increase the probability of one class. i.e. to increase the probability of using a short-term contraceptive and the probability of using a long-term contraceptive .

Applying the method to increase the probability of using a long-term contraceptive showed that the most significant lever variable is the education level. Out of 1473 instances, 577 instances were already at a high education level. Out of the remaining 895 instances, 99 were predicted to

switch from no contraceptive to a long term contraceptive if the education level was changed from whatever value (low or middle) to a high value, and 30 instances were predicted to switch from short term contraceptive to long term contraceptive with the same change in education level. Media exposure could not make significant impact (only 2 instances changed to long term contraceptive, by changing the media exposure to good media exposure). Applying the method to increase the probability of using a short term contraceptive, 157 instances were predicted to switch from no contraceptive to short term contraceptive with a higher education, and 18 with change to good media exposure. (41).

Pejić M et.al. (42) has also used a decision tree technique and CHAID algorithm to explain the choice of contraceptive method among woman of Indonesia. The purpose of their study was to categorize or classify woman into non users, short term users and long term users based on some variables such as; wife's age, husband profession, wife's education level, media exposure, number of child ever born, religion, life index, etc. they have used the dataset of Indonesia and applies data mining techniques to determine if there are common characteristics among Indonesian woman in choosing contraceptive method. After applying this model the target is to show how to raise awareness among Indonesian woman on contraceptive method use. The technique applied by the investigators is Decision tree which is a powerful and popular tool for classification and prediction. It is also appealing because it can be understood easily, as it can be graphically presented as tree as well as in the form of rules (in English or in SQL).

Husband's profession was detected as first determinant factors the reason they put for this is because the majority of the population are Islamic and in this case husband's domination is large. Another important thing observed is the number of ever born children. After the second child has born a significant change comes and it prevails long term usage. The third important variable is woman's education level and it shows a significant change. When a woman's education increases non usage declines and the long term usage also increases. Woman's age has also an influence in contraceptive method use. When woman's age reaches 32 they intend to use long term methods.

Since the purpose of this study is to predict the likelihood of contraceptive method use with a predefined class (woman who use method and do not use) the task employed is classification. In

the next chapter the techniques which help to achieve the goal of classification task will be discussed in detail. Decision tree is preferable because of it is easy to understand, interpret and it can be graphically presented as tree as well as in the form of rules and gives us a good measure of accuracy. Bayesian classifier is also have a comparable performance with decision tree and selected neural network classifiers. Bayesian classifiers have also showed a high accuracy and speed when applied to large databases.

Different variable supposed to influence the use of contraceptive method has observed from the different literatures reviewed. Comparing the EDHS with other studies possible determinant variables are viewed and fed to the appropriate techniques in order to get real and significant factors which differentiates contraceptive method users and non users.

As to the knowledge of the researcher there is no any research applied data mining techniques to a contraceptive method data set for any purpose in Ethiopia.

CHAPTER THREE

3. METHODS AND TECHNIQUES

Since the task of data mining selected achieve the objective of this study is classification, different data mining classification techniques are reviewed in chapter two. Based on these discussions decision tree and Naïve Bayes are chosen. A detail of these algorithms and the performance measurement used to compare these algorithms are discussed in the following section.

3.1. Decision Tree

Decision trees follow a top-down approach in order to search a solution for a problem. Decision tree are inductive, that is, they generalizes from the observed training data. There are different algorithms which generates a decision tree and rule. Decision trees handle both categorical and continuous response variables. If the target variable is categorical the decision tree is called classification tree but if it is continuous it will be regression tree.

In the top-down approach attributes are selected from the training sample in the root node and a branch from the root node is created for each value of the attribute. Recursively this branching is applied in the child node until all samples at the node belongs to one class. Selection of an attribute is most important in generating a better decision tree. ID3 and its successor C4.5 select their attribute based on entropy and information gain. To determine which attributes comes at the top of the tree the entropy and information gain is calculated recursively. The principle is to calculate the information gain of each candidate attributes and the attribute with the highest information gain is set as the top node of the tree. It is first necessary to understand the concept of entropy, which is the disorderness with the target classifier. If the randomness is high then entropy is high and if the entropy is zero then there is a complete uniformity with respect to the classifier.

The attribute with the highest information gain(less entropy) minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or “impurity” in

these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found (13).

ID3 and its new version of C4.5 use univariate split and follow greedy search method that is based on growing and pruning decision tree structure (i.e., no backtracking). Decision trees that use univarate splits have a simple representational form, making it relatively easy for the user to understand the generated model; at the same time, they represent a restriction on the expressiveness of the model. In general, any restriction on a particular tree representation can significantly restrict the functional form and thus the approximation power of the model (27).

Entropy is calculated as follows (43).

$$Entropy(s) = -p_+ log_2 p_+ - p_- log_2 p_- \qquad 3.1$$

Where p_+: the proportion of positive examples in collection S

P-: the proportion of negative examples in collection S

The next step is to calculate the information gain of each candidate attribute and is calculated as follows

$$IG(S, A) = entropy(s) - \sum_{vE} Value(A) - \left(\frac{s_v}{s}\right) entropy(s_v) \qquad 3.2$$

The information Gain(IG) of attribute A in collection S where Values (A) is the set of possible values for attribute A and S_v is the subset of S for which attribute A has the value v. We can calculate for each possible attribute its expected entropy. This is the degree to which the entropy would change if branch on this attribute. You add the entropies of the two children, weighted by the proportion of examples from the parent node that ended up at that child then, that is, by minimizing the number of tests allowed for classification. A typical decision tree has a structure as shown below.

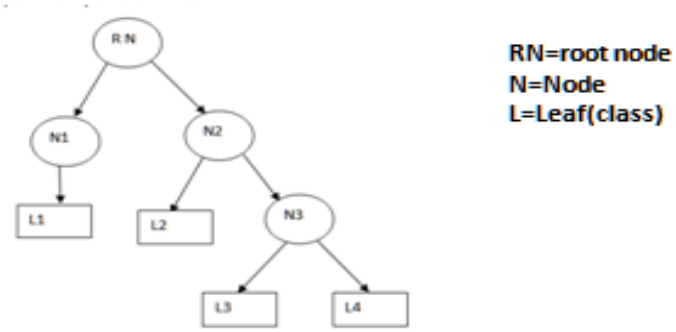

Figure 3.1: An example of simple Decision Tree Structure

The upper most node is where classification starts and is called root node. Some algorithms splits a node into two branches which is called binary tree whereas others split it into more than two branches each node this tree is called multiway tree. The splitting of node continues until a leaf node is reached. There are two nodes in decision tree. Leaf node represents a class and decision node indicates a node that needs further test until a leaf is encountered. N represents a node to be splitted and L represents leaf node (class).

A decision tree might grow large while there are ways we can reduce the size of tree with a good predictive accuracy. The size of the tree can be adjusted either while it is being generated which is called pre pruning or it can be reduced after the tree is generated known as post-pruning. The main goal of pruning is to generate a tree with fewer branches and good accuracy. Decision tree algorithms also are capable of deciding what values or how many branches to be assigned to the selected node. Poorly designed decision tree algorithms that assign random values often cause an ineffective representation of the decision tree technique (27).

There are a number of algorithms used to generate a decision tree. CART, CHIAD, ID3, C4.5 are some of the commonly applied algorithms of decision tree. Since in this study WEKA version 3.6.4 is used the WEKA implementation of C4.5 that is J48 is hired to train the dataset and create a decision tree.

The basic algorithm for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered. The following algorithm shows the generating of a decision tree from a training tuples of data partition D (13).

***Input**:*

- *Data partition, D, which is a set of training tuples and their associated class labels;*
- *attribute list, the set of candidate attributes;*
- *Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.*

***Output**: A decision tree*

Method:

1. *create a node N;*
2. ***if** tuples in D are all of the same class, C **then***
3. ***return** N as a leaf node labeled with the class C;*
4. ***if** attribute_list is empty **then***
5. ***return** N as a leaf node labeled with the majority class in D;*
6. *apply Attribute_selection_method(D, attribute list) to find the "best" splitting criterion;*
7. *label node N with splitting criterion;*
8. ***if** splitting_attribute is discrete-valued and multiway splits allowed **then***
9. *attribute list ← attribute_list – splitting_attribute; // remove splitting attribute*
10. ***for each** outcome j of splitting_criterion*
11. *let D_j be the set of data tuples in D satisfying outcome j;*
12. ***if** D_j is empty **then***
13. *attach a leaf labeled with the majority class in D to node N;*
14. ***else** attach the node returned by Generate decision tree(Dj , attribute list) to node N;*
    endfor
15. ***return** N*

Selecting the splitting criterion or optimal attribute at each node involves choosing the attribute that maximizes information gain. In other words, the optimal attribute minimizes the information needed in the resulting sub tree to classify the data set. Information gain is also referred to as entropy reduction, where entropy denotes the measure of randomness in a data collection. After selecting the lowest entropy attribute, branches representing the attribute values or choices are added to the node. Since the number of branches depends on the cardinality of the attribute domain, domains with a high cardinality can adversely affect the performance of a DT. Much like the ideal stopping point quandary, accuracy is often compromised by limiting the available choices to enhance performance (13).

J48 employs two pruning methods; the first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf, basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub tree raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub tree raising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub tree raising can be somewhat computationally complex.

3.1.1. Constructing Rules from a Decision Tree

If-then rules can be generated from a decision tree. In comparison with a decision tree, the IF-THEN rules may be easier for humans to understand, particularly if the decision tree is very large. For each path form the root node to a leaf node one rule is extracted. Each splitting criterion along a given path is logically ANDed to form the rule antecedent (“IF” part). The leaf node holds the class prediction, forming the rule consequent (“THEN” part). A disjunction (logical OR) is implied between each of the extracted rules.

Because the rules are extracted directly from the tree, they are mutually exclusive and exhaustive. By mutually exclusive, this means that we cannot have rule conflicts here because no two rules will be triggered for the same record. (We have one rule per leaf, and any record can map to only one leaf.) By exhaustive, there is one rule for each possible attribute value

combination, so that this set of rules does not require a default rule. Therefore, the order of the rules does not matter (they are unordered). Since we end up with one rule per leaf, the set of extracted rules is not much simpler than the corresponding decision tree. But in some cases the generated rules might be difficult to understand and interpret than a tree .this happens if the rule extracted becomes large with repeated and irrelevant attributes. In this situation it is necessary to prune the generated rules. For a given rule antecedent, any condition that does not improve the estimated accuracy of the rule can be removed, thereby generalizing the rule. C4.5 extracts rules from an un pruned tree, and then prunes the rules using a pessimistic approach similar to its tree pruning method. The training tuple and their associated class labels are used to estimate rule accuracy. However, because this would result in an optimistic estimate, alternatively, the estimate is adjusted to compensate for the bias, resulting in a pessimistic estimate. In addition, any rule that does not contribute to the overall accuracy of the entire rule set can also be pruned. Other problems arise during rule pruning, however, as the rules will no longer be mutually exclusive and exhaustive. For conflict resolution, C4.5 adopts a class-based ordering scheme. It groups all rules for a single class together, and then determines a ranking of these class rule sets. Within a rule set, the rules are not ordered. C4.5 orders the class rule sets so as to minimize the number of false-positive errors (i.e., where a rule predicts a class, C, but the actual class is not C). The class rule set with the least number of false positives is examined first. Once pruning is complete, a final check is done to remove any duplicates. When choosing a default class, C4.5 does not choose the majority class, because this class will likely have many rules for its tuples. Instead, it selects the class that contains the most training tuples that were not covered by any rule (13).

3.2. Bayesian Classification

Bayesian classification is based on Bayes' theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This assumption is called class

conditional independence. It is made to simplify the computations involved .When the assumption holds true, and then the Naïve Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables. Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification (13).

3.2.1. Naïve Bayes Classification

Naïve Bayes classification algorithm is one classification technique that does not use any rule like that of decision tree. The foundation of Naïve Bayes is the probability theory. The straightforward of calculating probability is to look for the frequent event and classify the unseen instance to the frequent occurring event. Using more complex probability types is better prediction of the unseen event. To obtain the prior probability we divide the frequency of most frequent event by total number of instance. The probability of an event occurring if we know that an attribute has a particular value (or that several variables have particular values) is called the conditional probability. This is also called posterior probability since it calculates the probability after it gains information where as priori probability calculates before it obtains information. Naïve Bayes is capable of integrating both priori and postpriori probability in one (43).

The Naïve Bayesian classifier, or simple Bayesian classifier, works as follows (13).

Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (X1, X2,…,Xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2, … , An.

Suppose that there are m classes, C1, C2… Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if

$$P(Ci|X) > P(Cj|X) for 1 \leq j \leq m; j \neq i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis

$$p(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \qquad 3.3$$

As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) =…= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|,where |Ci,D| is the number of training tuples of class Ci in D.

Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X|Ci) = \prod_{k=1}^{n} P(Xk|Ci) \qquad 3.4$$

$$= P(X1|Ci)\ X\ P(X2|Ci)\ X…X\ P(Xn|Ci)$$

the probabilities P(X1|Ci), P(X2|Ci), …, P(Xn|Ci) can easily be estimated from the training tuples. Xk refers to the value of attribute Ak for tuple X.

To compute P(X|Ci), we consider the following:

(a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having the value Xk for Ak, divided by |Ci,D|, the number of tuples of class Ci in D.

(b) If Ak is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad 3.5$$

$$P(Xk|Ci) = g(Xk, \mu ci, \sigma ci) \qquad 3.6$$

μCi and σCi are the mean (i.e., average) and standard deviation, respectively, of the values of attribute Ak for training tuples of class Ci.

In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci.

The classifier predicts that the class label of tuple X is the class Ci if and only if

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj) \; for 1 \leq j \leq m; j \neq i \qquad 3.7$$

When using the Naıve Bayes method to classify a series of unseen instances the most efficient way to start is calculating all the prior probabilities and also all the conditional probabilities involving one attribute, though not all of them may be required for classifying any particular instance. Even though Naïve Bayes algorithm has a good performance it has some shortcomings. The most obvious one being that it relies on all attributes being categorical. In practice, many data sets have a combination of categorical and continuous attributes, or even only continuous attributes. This problem can be resolved by converting the continuous attributes to categorical. A second problem is that estimating probabilities by relative frequencies can give a poor estimate if the number of instances with a given attribute/value combination is small (43).

The other drawback of Naïve Bayes is overfitting which results in some cases in excessively large rule sets and/or rules with very low predictive power for previously unseen data is a problem in Naïve Bayes. A classification algorithm is said to overfit to the training data if the model is strongly dependent on particularly on a certain feature of the dataset. The effect of

overfitting is the prediction works well for the data used to train the model but it is poor when it comes to new instances. Realistically, overfitting will always occur to a greater or lesser extent simply because the training set does not contain all possible instances. It only becomes a problem when the classification accuracy on unseen instances is reduced to the large extent. Overfitting commonly occurs even though the sensitivity of algorithms towards it is different. The use of different mechanism is necessary in order to avoid overfitting (43).

3.3. Performance measurement

One performance measurement used to compare classification algorithms is confusion matrix. Confusion matrix is an n by n table that contains information related to how many cases are correctly and incorrectly classified. Figure 1.2 shows a simple 2 by 2 confusion matrix.

		predicted	
		Class1	Class2
actual	Class1	TP	FN
	Class2	FP	TN

Figure 3.2: A simple 2X2 confusion matrix

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier in other word referred to as the overall recognition rate of the classifier i.e. how well the classifier recognizes tuples of the various classes.

An algorithm with high accuracy may not be acceptable because the classifier could be labeling only one class (i.e. the percentage of accuracy might be contributed by the majority of the class if the number of cases with each class is unbalanced). To tackle this trick we have to access how well the classifier can recognize the positive tuples(i.e. sensitivity) and how well it can recognize the negative tuples (i.e. specificity) A precision is also used to access the percentage of tuples labeled as positive(non user of CM) that actually are positive tuples(non user of CM method). These measures are defined as

TP rate (sensitivity), TN rate(specificity), precision and accuracy are calculated as follows

$$Sensitivity = \frac{TP}{TP + FN} \quad (3.8)$$

$$specificity = \frac{TN}{TN + FP} \quad (3.9)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.10)$$

$$precision = \frac{TP}{TP + FP} \quad (3.11)$$

CHAPTER FOUR

4. DATA UNDERSTANDING AND PREPROCESSING

since the objective of this study is to classify women according to their likelihood of practicing family planning in to actual and non actual users of contraceptive method. In order to mine useful and interesting patterns among features of the data set the desired data was selected and understood then all the necessary preprocessing and transformation is performed.

4.1. Data Understanding and Data Selection

Since the goal of this study was to understand the factors that hinder woman from using any method of contraception with respect to their socio-economic, demographic, knowledge, reproduction and other related factors , the researcher has asked if there are such data in the organizations which work in contraceptive related issues. Family guidance association and Marie stops are two of the main organizations engaged in contraceptive related service. But the necessary data could not be found in either of them. The EDHS data was found to be more relevant source. During this stage data should also be checked for completeness, redundancy, missing values attribute values, etc.

4.1.1. Contraception

Family planning refers to use of modern contraceptives or natural techniques to limit or space pregnancies. Modern methods of contraception include the pill, female and male sterilization, IUD, Injectables , implants, male and female condom, diaphragm, and emergency contraception. Traditional methods include periodic abstinence, withdrawal and folk methods (44).

Family planning is also called Contraception which can be defined as any means to prevent pregnancy (45).

Different methods of contraception are used to prevent pregnancy; these are abstinence, hormonal methods, barrier methods that Prevent sperm from moving up the upper female genital tract, devices placed into the womb, natural methods, female tubal ligation and male vasectomy. Condom is the only one currently known that protects against HIV transmission in addition to

using pregnancy. Any person or couple that is sexually active and wants to avoid sexually transmitted infections may decide to use condoms.

There are several factors that determine people either to use or not to use contraceptive method. The main reason people use contraception (family planning) is to prevent a pregnancy i.e. if heterosexual couples are sexually active and they don't want children currently do not want children at all, want to space their birth or do not want to add any children.

Other factors that lead people to use contraception includes

- Greater enjoyment of sex because they are less worried about pregnancy and/or STIs.
- Economic considerations – contraception is usually more affordable than providing for a child
- They may be benefited from a better quality of life by having a smaller number of children
- A person may also wish to pursue personal goals before having a child

On the other hand some people (especially young people) don't use contraception, even though they don't want to become pregnant. So many reasons are stated for this, some of these are

- Young people might be embarrassing due to the Social norms about sex
- Religion may forbid the use of contraception
- The Social norms about sex can also make it embarrassing for people to talk about contraception with their sexual partner and may let them not to use
- The partner of the person who wants to use might not approve of contraception
- The role of women in the society may create an influence on producing a child
- People, especially women and young people, are sometimes coerced into having sexual relations.
- The side effects of contraception whether it is real or a myths is also a factor which make women not to use any method.
- The living standard of people may create any awareness where to get contraception.
- People may not know a method which is acceptable for them.

- People (especially men) are often reluctant to use barrier methods, such as condoms, because of the perception that they decrease sexual pleasure.
- Lack of knowledge about being at risk of pregnancy or STI may not let them to use contraception (45).

In addition to the above reasons studies reviewed in chapter two revealed that several other variables determine whether a woman uses or not any method of contraception.

4.1.2. Demographic Health Survey

The central statistics agency is an organization that conducts and handles statistical issues of the government of the Federal Democratic Republic of Ethiopia. Since its establishment in 1960 it has been and is involved in socio-economic and demographic data collection, processing, evaluation and dissemination that are used for the country's socio-economic development and planning, monitoring and policy formulation. These tasks are performed through running National Integrated Household and Enterprise survey program (NIHESP), undertaking ad-hoc surveys, conducting census, and compilation of secondary data from administrative records. The Demographic and health survey is of the surveys included in the program among others (46).

The demographic and health surveys (DHS) are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition. Generally there are two main types of DHS surveys. these are the standard DHS surveys, which have large sample sized (usually between 5,000 and 3,000 households) and typically are conducted about every 5 years, to allow comparisons over time and the second type of survey is the Interim DHS surveys, which focus on the collection of information on key performance monitoring indicators but may not include data for all impact evaluation measures (such as mortality rates). These surveys are conducted between rounds of DHS surveys and have shorter questionnaires than the standard DHS surveys (44).

The standard DHS survey holds information for the following topics, among others:

- Anemia-prevalence of anemia, iron supplementation
- Child health- vaccinations, childhood illness, newborn care

- Domestics violence-prevalence of domestic violence and consequences of violence
- Education- literacy , attendance, highest level achieved
- Environment health-water, sanitation, cooking fuel
- Family planning- knowledge and use of contraceptives
- Female genital cutting- prevalence and attitude of female genital cutting
- Fertility and fertility preference-total fertility rate, desired family size, marriage and sexual activity
- Gender/domestic violence-history of domestic violence, frequency and consequences of violence
- HIV/AIDS knowledge, attitudes , and behavior- knowledge of HIV prevention, misconceptions , stigma , higher-risk sexual behavior , previous HIV testing
- HIV prevalence- prevalence of HIV by demographic and behavioral characteristics
- Household and respondent characteristics- electricity , housing quality , possessions, education and school attendance, age , sex, employment.
- Infant and child mortality –infant and child mortality rates
- Malaria-ownership and use of mosquito nets, prevalence and treatment of fever, indoor residual spraying for mosquitoes.
- Maternal health – antenatal , delivery and postnatal care
- Maternal mortality – maternal mortality ratio.
- Nutrition- child feeding practice, vitamin supplementation, anthropometry, anemia, salt iodization.
- Tobacco use- tobacco use, exposure to second –hand smoke
- Unmet need for family planning
- Wealth –division of households into 5 wealth quintiles to show relationship between wealth , population and health indicators
- Women's empowerment- gender attitudes, women decision making power, education and employment of men vs women

4.1.3. DHS and Family Planning

One of the DHS major topics is family planning which contains information of knowledge and use of contraceptive methods, both modern and traditional. It collects information about the

source of contraceptive method and whether the required information to make an informed choice is provided. Information related to the discussion with health care providers about family planning and the exposure to family planning and coverage of social marketing programs are covered. The following indicators are recognized to have relation to family planning in the DHS (44).

- Knowledge of contraceptive methods(women and men)
- Ever use of contraception(women and men)
- Current use of contraceptive by background characteristics
- Number of children at first use
- Knowledge of fertile period
- Contraceptive effect of breast feeding
- Timing of sterilization
- Source of supply for modern contraceptive method
- Contraceptive discontinuation rates
- Future use of contraception
- Reason for not using contraception
- Preferred method of contraception for future use
- Heard family planning on radio and television or in a newspaper or magazine
- Acceptability of media messages on family planning
- Contact of non-users with family planning providers
- Was family planning use mainly the woman's /the husband/partner's or a joint decision
- Informed choice of family planning methods
- Unmet need for family planning

The 2005 EDHS data related to a woman age 15-49 is the target of this study. The women's questioner has 10 sections these are

- Respondent's background
- Reproduction
- Contraception
- Pregnancy, delivery, postnatal care and nutrition

- Immunization, health, and women's nutrition
- Marriage and sexual activity
- Fertility preference
- Husband’s background and woman’s work
- HIV/AIDS and other sexually transmitted infections
- Harmful traditional practices
- Maternal mortality

Attributes that are not necessary for the data mining techniques to create a model are removed. Based on the appropriateness to the domain problem, data understanding and literatures reviewed in chapter two section 2.3 and 2.4 the attributes that are considered to be relevant are chosen. Therefore the following attributes are selected

Respondent’s age, region, type of place, religion, marital status, education level, partner’s education level, occupation, partner’s occupation, number of living children, partner’s age, partner approves FP, knowledge of FP, heard FP news paper last month, heard FP on radio last month, heard FP on television last month, visited by FP worker during last 12 m, discuss FP with partner, wealth index. The three attributes (heard FP on radio, heard FP on TV and heard FP on news paper) are combined and a new attribute named FP message is created. If a woman has at least heard family planning in either of the three medias in the last month before the interview then it is considered she has a FP message. The number of attribute which were 17 at first becomes 15 after the merging is done.

These variables have a relationship with the use of contraceptive method for example; Place of residence one characteristic that determines access to services and exposure to information pertaining to reproductive health and other aspects of life. Education is also an important factor influencing an individual’s attitude and outlook on various aspects of life. Acquiring knowledge about family planning is an important step towards gaining access to and using a suitable contraceptive method in a timely and effective manner. Individuals who have adequate information about the available methods of contraception are better able to make choices about planning their families. Exposure to family planning messages widens the horizon of understanding on issues related to contraceptive use and helps in the realization of its importance

in achieving desired family size. Additionally, it contributes to the enhancement of the health of both children and mothers. Measuring the extent of exposure to such information helps program managers and planners to effectively target population subgroups for information, education, and communication (IEC) activities. Religious adherents also vary widely in their views on birth control. This can be true even between different branches of one faith.

Table 4.1 : Selected Variables and Percentage of Missing Attribute Value

No	Attribute name	Data type	Description	Missing value
1	Current age respondent	Numeric	Current age of a woman	0%
2	Num of living children	Numeric	women's child who are currently alive	0%
3	Region	Categorical	Region they belong	0%
4	Type of place of residence	Categorical	Place of residence Rural or urban	0%
5	Highest educational level	Categorical	Education level they attain	0%
6	Religion	Categorical	Religion of the women	4(0%)
7	Knowledge of any method	Categorical	Knowledge about any method of contraceptive method	0%
8	Heard FP on radio	Categorical	Whether a woman has heard FP on radio or not in the last month before the interview.	4(0%)
9	Heard FP on Tv	Categorical	Whether a woman has heard FP on TV or not in the last month before the interview.	3(0%)
10	Heard FP on news paper	Categorical	Whether a woman has heard FP on news paper in the last month before the interview.	4(0%)
11	Visited by FP worker last 12m	Categorical	Whether a woman is visited by FP workers in the last 12 m before the interview.	8(0%)

No	Attribute name	Data type	Description	Missing value
12	Current marital status	Categorical	Marriage status (married, divorced, single)	0%
13	Partner's education level	Categorical	Education status of a partner	3911(27.8%)
14	Partner's occupation	Categorical	Job of a partner	3921(27.9%)
15	Respondent's occupation	Categorical	Occupation of a woman	25
16	Wealth index	Categorical	Living standard	0%
17	Current use by method type	Categorical	Method they use currently (modern, folk, traditional or no use)	

The demographic characteristics of the variables is shown in the following table

Table 4.2: Descriptive statistics of the selected variable

No	Demographic characteristics	Number of instance	Number of instance in %	No	Demographic characteristics	Number of instance	Number of instance in %
1	**Region**			**5**	**Respondent age**		
	1(Tigray)	1257	8.9%		1(15-19)	3252	23.1%
	2(Afar)	789	5.6%		2(20-24)	2617	18.6%
	3(Amhara)	1943	13.8%		3(25-29)	2557	18.2%
	4(Oromiya)	2230	15.8%		4(30-34)	1754	12.5%
	5(Somali)	669	4.8%		5(35-39)	1629	11.6%
	6(Ben-Gumz)	846	6.0%		6(40-44)	1181	8.4%
	7(SNNP)	2087	14.8%		7(45-49)	1080	7.7%
	12(Gambela)	729	5.2%	**6**	**Knowledge of any method**		
	13(Harari)	844	6.0%		0(no knowledge)	2494	17.7%

No	Demographic characteristics	Number of instance	Number of instance in %	No	Demographic characteristics	Number of instance	Number of instance in %
	14(Addis Abeba)	1869	13.3%		1(has knowledge)	11576	82.3%
	15(Dire Dawa)	807	5.7%	**7**	**Marital status**		
2	**Type of place**				0(never married)	3830	27.2%
	1(Rural)	9647	68.6%		1(living together)	1596	11.3%
	2(Urban)	4423	31.4%		2(not living together	3830	27.2%
3	**Education level**			**8**	**Partners' occupation**		
	0(No education)	8454	60.1%		0 (not working)	95	.7%
	1(Primary)	2966	21.1%		1(agric-employee)	7314	52.0%
	2(Secondary)	2292	16.3%		2(non agric-employee)	2740	19.5%
	3(Higher)	358	2.5%	**9**	**Respondent occupation**		
4	**Religion**				0(not working)	9121	64.8%
	1(Orthodox)	6809			1(agric-employee)	1822	12.9%
	2(Catholic)	143			2(non agric-employee)	3080	21.9%
	3(Protestant)	2301	16.4%	**10**	**Wealth Index**		
	4(Moslem)	4522	32.1%		1(poor)	4818	34.2%
	5(Traditional)	172	1.2%		2(medium)	2051	14.6%
	6(Other)	119	.8%		3(rich)	7201	51.2%
11	**Visited by FP worker**			**14**	**FP message**		
	0(No)	13194	93.8%		0(not exposed to FP message)	8900	63.3%
	1(Yes)	868	6.2%		1(exposed to FP message)	5163	36.7%
12	**Partners' education level**			**15**	**Current use of method**		
	0(No education)	5946	42.3%		0 (No)	12377	88.0%
	1(Primary)	2287	16.3%		1(Yes)	1693	12.0%
	2(Secondary)	1559	11.1%				
	3(Higher)	367	2.6%				
	Don't know	46	.3%				
13	**Number of living children**						
	0(no child)	4924	35.0%				
	1(1 or 2 child)	3473	24.7%				
	2 (more than 3)	5673	40.3%				

4.2. Data Preprocessing

The main reason we need preprocessing is that because our source of data might contain incomplete information or data with noise, invalid or outlier. In order to have a good classification, prediction in general to achieve the goal of data mining we have to preprocess our data.

Handling missing values

Missing data are data that are without value. There might be different reasons why these cells remain vacant. Missing data can affect the output of the model. Different options are available for handling of missing value; the simplest one is to leave those records with missing value. But this might lead to a biased subset of dataset. We can use other options to overcome the above problem (28).

- Filling manually the missing values, but this is tedious and is not feasible if the dataset is large.
- Filling all missing values with some global constant, this is simple but might create confusion while we apply data mining techniques.
- Filling the missing values with mean value (for numerical) or mode (for categorical) attributes.
- Filling missing values with the attribute mean for all samples belonging to the same class as the given tuple.
- Filling missing values with the most probable value using different technique.

Attributes with “don’t know” is treated as missing value in addition to the values originally identified as missing. After the data is exported to excel numerical variables are filled with the mean value and categorical are filled with the most frequent value (mode). Attributes with more large percentage of missing value are excluded from the list. Since partner’s age, discussion FP with partner, partner approve FP have such a case they are excluded from the dataset.

The following table shows the attributes that have missing values and the substituted new value of the missing attribute values.

Table 4.3 : Replaced Value of an Attributes Missing Value

No	Attribute	Replaced value
1	Religion	Orthodox
2	Visited by FP	No
3	FP message	No
4	Partners education level	No education
5	Partners occupation	Agric-employee
6	Respondent occupation	Not working

Identifying outliers

Outliers are extreme values observed in the dataset that are found close to the limits of the data range or go against the trend of the remaining data. 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. Sometimes two-dimensional scatter plots can help to reveal outliers in more than one variable. It is also possible to use z-score standardization or inter quartile range which is numerical method to identify and correct outliers (29).

Z-score is a type of numerical method applied to identify outliers within a dataset. The concept of this method is that values found much farther than 3 standard deviations from the mean of the attribute value are outliers. In general outliers have either greater than 3 or less than -3 standard deviation. One drawback of this method is both mean and standard deviation are sensitive to outliers, that is, if an outlier is added to the dataset the mean and standard deviation are affected. It is recommended not to use measures that are themselves affected by outliers (22).

To overcome the above problem another statistical method that are less sensitive to outliers are devised. The inter quartile range is such a method. The quartiles of a data set divide the data set into four parts, each containing 25% of the data.

The first quartile (Q1) is the 25th percentile.

The second quartile (Q2) is the 50th percentile, that is, the median.

The third quartile (Q3) is the 75th percentile.

The inter quartile range (IQR) is calculated as IQR = Q3 − Q1

It is a measure of variability that is much more robust than the standard deviation and may be interpreted to represent the spread of the middle 50% of the data. According to the IQR a data value is an outlier if:

a. It is located 1.5(IQR) or more below Q1, or

b. It is located 1.5(IQR) or more above Q3

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.

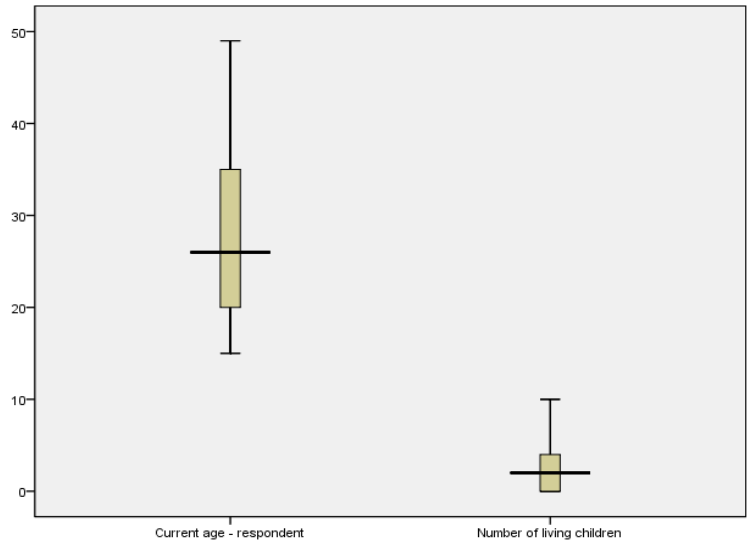

Figure 4.1 Box Plot for Numeric Attributes to Identify Outliers

The box plot consists of the median, the quar-tilesQ1 andQ3, and the smallest and largest observations. Typically, the ends of the box are at the quartiles, so that the box length is the

interquartile range, IQR. The median is marked by a line within the box. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

4.3. Data transformation

Transformation refers to making suitable the available data or converting to some form of representation for further processing or task. Variables may have ranges that vary greatly from each other. For some data mining algorithms, such differences in the ranges will lead to a tendency for the variable with greater range to have too much influence on the results. It is necessary to normalize numerical variables, to standardize the scale of effect each variable has on the results.

Discretization converts those continuous values into discrete values i.e. the original attribute values are categorized into a certain interval labeled with a new representation .The attribute number of living child is one attribute which discretization was performed on. The attribute has 20 values in the original format but after discretization was performed the number of values becomes 3. The age attribute was available in two formats; discrete (grouped age) and continuous value. Therefore the grouped age attribute is selected.

Table 4.4: A Discretized age Attribute

Age value	Represented value
15-19	1
20-24	2
25-29	3
30-34	4
35-39	5
40-44	6
45-49	7

The following table shows the number of living children attribute after disccretization is applied

Table 4.5: A Discretized num of living children attribute

Attribute value	New representation
0	0
1-2	1
3-20	2

The numeric values (such as age and number of living children) are converted to nominal. The value of partner's occupation, respondent's occupation , marital status, knowledge of any method , current use of method also recoded to a certain label. The initial selected attributes with their original value i.e. before some transformation is applied is given in appendix A. The domain of each attribute after recoding is shown below.

1. Region :{1:Tigray, 2:Afar, 3: Amhara, 4:oromina, 5:Somalia ,6:Ben-Gumuz,7:SNNP ,12:Gambela, 13: Harari, 14: Addis Ababa ,15: Dire Dawa}
2. Type of Place :{1: rural ,2: urban }
3. Religion:{1:Orthodox, 2:catholoic ,3:Protestant ,4:Muslim ,5:Traditional ,6:other}
4. VisitedbyFPworker:{0:No, 1:Yes}
5. Exposure to FP message: {0: No ,1 :Yes}
6. Education: {0:no education ,1:Primary, 2: secondary, 3:higher}
7. Number of living children :{0: no children,1: 1 or 2 children,2: three or more}
8. Knowledge of FP :{0: knows no FP method,1: knows a FP method}
9. Marital status: {0-never married, 1:living together (married , living together), 2-not living together (widowed, divorced, not living together)}
10. Partners' Education: {0:no education ,1:Primary, 2: secondary, 3:higher}
11. Partners' Occupation: {0: did not work, 1: agric-employee,2: non agric-employee}
12. Respondent's Occupation :{0: not working, 1: agric-employee,2: non agric-employee}
13. Wealth index: {1:poor,2:medium, 3:rich}
14. Respondents' Age: {1:15-19,2:20-24, 3:25-29, 4:30-34, 5:35-39,6:40-44,7:45-49}
15. currentCmuse: {No, Yes}

Format

The data was originally stored in SPSS format so in order to make it work in WEKA it has exported to excel. In the excel work sheet Missing values were replaced with the most frequent value (mode) if the data type is nominal and if the data type is numeric it is filled with mean value. Attributes that was numeric also are merged into a certain group (nominal) .The next step was to save it in .csv (comma delimited format) which is a format where commas are placed between values in adjacent columns. The database was then opened in Word, header information added. That is, the @ symbol was placed in front of the relation name and each attribute. The @DATA symbol was placed before the data. Finally, prior to saving the file extension was changed to .arff.

If the variable or attribute is nominal a list of possible values contained in a brace is required otherwise generic words like 'real' to mean continuous are used. By default the last attribute in the list of the attributes name in the dataset designates the target class. However, it is also possible to choose any attribute name as a target class no matter its position in the list while running the program. A ample of data set prepared for WEKA is shown below

Figure 4.2: A sample of data set prepared for WEKA

CHAPTER FIVE

5. EXPERIMENTATION

In order to validate and compare the classification performance of the techniques the 10 fold cross validation and percentage split are used. Both methods are tested with default value and by changing the default value it was checked if a better performance is scored. The data set is re-sampled since the number of records of the target variable with the value (non-user of contraceptive method) were much larger than the number of contraceptive method user , which is 12, 377 and 1, 693 respectively. Unbalanced data may create a bias during classification. SMOTE was used repeatedly to increase the number of contraceptive method user records. Finally 12, 377 and 6, 772 number of records for non contraceptive method users and contraceptive method users respectively are used for training and testing. The performance of the techniques on both the unbalanced and balanced data was then measured.

5.1. Experiment Set Up

Different experiments was conducted to investigate

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

5.2. Classification Model Building Using Decision Tree (J48 algorithm)

One of the classification techniques applied for building the classification model in this thesis is the J48 algorithm. J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Initially the classification model is built with the default

parameter values of the J48 algorithm. Table 5.1displays the default parameters with their values for the J48 decision tree algorithm.

Table 5.1: Default parameters with their values for the J48 decision tree algorithm.

Parameter	Description	Default value
Confidence factor	The confidence factor used for pruning (smaller values incur more pruning)	.25
minNumobj	The minimum number of instances per leaf	2
Unpruned	Whether pruning is performed or not	False

Experiment I

The first experimentation is performed with the J48 default parameters. The default 10-fold cross validation test option is employed for training and testing the classification model. In 10-fold cross-validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or "folds," 1, 2, 3, 10, each approximately equal size. Training and testing is performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier; the classifier of the second iteration is trained on folds 1, 3, 4, …, 10 and tested on the second fold; and so on.

Using these J48 default parameters the classification model is developed. This experiment has produced a decision tree that has 119 numbers of leaves and 155 tree size.

Table 5.2 Confusion Matrix Output of experiment I

Actual	**Predicted**		**Recall**
	CM non user	CM user	
CM non user	12008	369	97.01%(TP rate)
CM user	1139	554	32.72% (TN rate))
Precision	91.3%	60.02%	

The overall classification accuracy of was good. From the total of 14,070 instances 89.28 % are correctly classified; only 10.71% are misclassified. As table 5.2 depicts from the total 12, 377 non contraceptive method users 97.01 % are correctly classified where as 2.09% are incorrectly classified as contraceptive method users. It can also be observed that from the total 1, 693 which are actually contraceptive method users only 32.72 % are classified as contraceptive method user and the remaining 67.28% are grouped as non users which should be categorized as contraceptive method users.

Experiment II

When the number of leaves is large it makes difficult to traverse through all the nodes of the tree in order to come out with valid rule sets. Therefore, to make ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves. However increasing the minimum number of objects that should appear in the leaf nodes has its own disadvantage as some of the instances may be classified incorrectly. In other words, records in a given leaf could be in different class and there could be attributes that could further split the records in the same node into disjoint classes.

With this objective, the minNumObj (minimum number of instances in a leaf) parameter is tried with 5,10,15,20 and 25 and a better accuracy with minimum number of leaves and small size of tree has produced when the minNumObj is set to 20 and 25. The number of leaves which was 119 has reduced to 58.The outcome of the experiment with minNumobj set to 20 is shown Table 5.3

Table 5.3 Confusion Matrix Output of experiment II

Actual	Predicted		Recall
	CM non user	CM user	
CM non user	12035	342	97.24%(TP rate)
CM user	1164	529	31.25% (TN rate)
Precision	91.12%	60.73%	

It is shown in the above table 5.3 that from the total of 14,070 records 89.30% of them are correctly classified only 10.70% of the instances are misclassified .It is also indicated that out of the total 12, 377 instances which was non contraceptive method users 97.24% of them are identified correctly and of the total of 1, 693 actual contraceptive method users 31.25% are predicted correctly and 68.75% of them are classified as non contraceptive method users which supposed to be in the category of contraceptive method user.

When the minNumobj is set to 20 not only the size of tree and number of leaves are reduced but also the accuracy is improved slightly from 89.28% to 89.30%. But in most cases when the minNumobj increases the accuracy decreases.

Experiment III

This experiment is performed, by changing the default testing option (the 10-fold cross validation) to the percentage split. In this learning scheme a percentage split is used to partition the dataset into training and testing data. The purpose of using this parameter was to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieved a better classification accuracy than the first experimentation. First this experiment has run with the default value of the percentage split (66%). It has also tried with 70% and 80%t of split. The outcome of classification with 66% percentage split is presented in table 5.4 as follows.

Table 5.4 Confusion Matrix Output of experiment III

Actual	**Predicted**		**Recall**
	CM non user	CM user	
CM non user	4074	166	96.08%(TP rate)
CM user	344	200	36.76%(TN rate)
Precision	92.21%	54.64%	

This experiment has given a comparative classification accuracy compared to the above experiment. From the total records 66% which is set for testing 89.34% of them are correctly classified the rest 10.66% records are misclassified. Again from the total 4, 090 non contraceptive method users 96.08% are identified as non contraceptive method users correctly. The remaining are predicted incorrectly. From the total of 544 actual contraceptive method users 36.76% of the instance are depicted as actual users while 63.24% of them are misclassified.

When the percentage of training is increased from 66% to 80% the overall accuracy was increased to 90.09%. This shows increasing the percentage of training data set increases the accuracy.

Experiment IV

In this experimentation the effect of pruned and un pruned decision tree was observed using the default parameters of J48 Algorithm. In experiment I the outcome of the pruned decision tree tested with the default values of J48 is presented .But observe the effect of pruning the parameter un pruned was set to True during this experimentation.

The number of leaves and size of tree produced was 119 and 155 respectively when the parameter un pruned was set to false (i.e. pruned decision tree). However, the number of leaves and the size of tree increased to 2, 073 and 2, 529 respectively when un pruned value is set to true. Table 5.5 shows the confusion matrix of un pruned decision tree

Table 5.5 Confusion Matrix Output of experiment VIII

Actual	**Predicted**		**Recall**
	CM non user	CM user	
CM non user	11704	673	94.56%(TP rate)
CM user	1157	536	31.66%(TN rate)
Precision	91.00%	44.33%	

The overall accuracy of this was is 86.99% which is less than the accuracy scored in pruned decision tree (experiment I). The unpruned tree not only produced a large number of leaves but it also scored a low accuracy compared to the pruned tree performed in experiment I.

Experiment V

The fifth experiment is performed by first balancing the original data set using SMOTE (synthetic Minority over sampling technique). A dataset is imbalanced if the classification categories are not approximately equally represented. Many real-world data are characterized by unbalanced data and this may affect the classification accuracy.

Table 5.6 Confusion Matrix Output of experiment V

Actual	**Predicted**		**Recall**
	CM non user	CM user	
CM non user	10806	1571	87.31%(TP rate)
CM user	1714	5058	74.69%(TN rate)
Precision	86.3%	76.3%	

When SMOTE is applied to the original data set the number of instances which were 14,070 has increased to 19, 149. In order to make the classification categories equally represented which was at first 12,377(non contraceptive method users) and 1, 693(contraceptive method user) SMOTE is applied and the number of records becomes 12,377(non contraceptive method users)

and 6, 772(contraceptive method users). Then an experiment was conducted on this dataset with the J48 default values. The overall accuracy becomes 85.60% when the experiment is conducted with default values of J48 but the number of leaves and size of tree has increased. In this experiment we can observe the effect of balancing a data set. Since the number of contraceptive method users are few compared to the non users the accuracy of classifying contraceptive method user was very low but when the data was made balanced using SMOTE its accuracy has increased even if the overall accuracy has lowered.

The other issue is the number of leaves and size of tree. The number of leaves and size of tree has increased to 868 and 1, 137 respectively.

To examine if the size of tree and number of leaves can be reduced the minNumobj is set to 20 and tested. The number of leaves and size of tree lowered to 253 and 319 which were 868 and 1, 137 respectively but its accuracy is lowered to 82.85%.

Table 5.6 depict that from the total 12,377 which are non contraceptive method user 87.31% of them are identified correctly as non users and 13.69% are classified as contraceptive method users which supposed to be classified as non users. Again from the total 6, 672 actual contraceptive method users 74.69% are correctly classified as actual users but the remaining 25.31% are classified as non user which should be classified as user of the method.

Experiment VI

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

When the Bestfirst method of attribute selection is applied to the 14, 070 records only four attributes are selected including the class attribute these are partner's education level, knowledge of any method, partner's occupation and the fourth attribute is the class attribute current use of contraceptive method.

This decision tree has produced only 1 leave and a size of tree 1.The main reason behind this is the un equivalent distribution of the class attribute .The number of instances of non contraceptive method user is 12, 377 where as the number of contraceptive method user is 1,693.

Since the default method of attribute selection best first selects only 4 attributes before SMOTE is applied (in un balanced data set) and 6 attributes were selected including the class variable after SMOTE is applied (i.e. the data set is made equivalent) But this selection gives low accuracy. The attribute selection was also performed by another type of attribute selection method, that is, gain ratio attribute evaluation. This method ranks attributes according to their significance. Therefore the first 10 attributes are selected according to their rank by the researcher. These are:

Knowledge of any method, Partner occupation, Partner's education level, Current marital status, Wealth index, Type of place, FP message, Number of living children, Religion, Education level.

The outcome of this performed with J48 default parameters on a balanced data set is shown in table 5.7

Table 5.7: Confusion Matrix Output of experiment VI

Actual	**Predicted**		**Recall**
	CM non use	CM user	
CM non user	10953	1424	88.49%(TP rate)
CM user	2025	4747	70.01%(TN rate)
precision	84.40%	76.92%	

Compared to the same type of experiment conducted on all attribute (experiment I) The overall accuracy is reduced to 82.00% despite the reduction of accuracy the number of leaves and size of tree are comparatively reduced.

5.3. Classification Model Building Using Naïve Bayes Algorithm

The other classification technique applied in this study is Naïve Bayes algorithm. Two experiments are conducted using the default value of 10-fold cross validation and the percentage split.

Experiment VII

The first experiment is performed using 10-fold cross validation and the outcome of this is presented in table 5.8. The overall accuracy of this test has lowered from what has scored in the previous experiment of decision tree. Out of the 14, 070 total records 85.40% are classified correctly.

Table 5.8: Confusion Matrix Output of experiment VII

Actual	**Predicted**		**Recall**
	CM non use	CM user	
CM non user	11107	1270	89.73%(TP rate)
CM user	783	910	53.75%(TN rate)
Precision	93.41%	41.74%	

When the accuracy of individual class is observed 89.73% of non contraceptive method user are correctly classified as non contraceptive method user and the remaining 10.27% are wrongly categorized as contraceptive method. Out of 1, 693 actual contraceptive method user 53.75% are correctly identified as actual users where as 46.25% are classified as non contraceptive method user which should not be. Compared to the same type of experiment in decision tree this one gives a better accuracy for the classification of actual contraceptive method user. Because this class has few numbers of instances its respective (corresponding) accuracy has been very low

Experiment VIII

The second type of experiment is performed using the percentage split of the default 66%. Again to investigate if the accuracy could be improved by increasing the percent of training data set to it has tries with 70% and 80% of split. Only a slight improvement is observed with the 80% of split .The outcome of the Naïve Bayes with a split of 66% is shown in table 5.9. Out of the total instances set for testing 85.87% are correctly classified.

Table 5.9: Confusion Matrix Output of experiment VIII

Actual	**Predicted**		**Recall**
	CM non use	CM user	
CM non user	3805	435	89.75%(TP rate)
CM user	243	301	55.33%(TN rate)
Precision	93.99%	40.89%	

The corresponding class accuracy is also improved. 89.75% of non contraceptive method users are correctly classified as non contraceptive method users and 55.33% of actual contraceptive method users are classified as actual users correctly.

5.4. Evaluation and Interpretation

Two of the classification techniques employed in the model building (J48 decision tree algorithm and Naïve Bayes classification algorithm) are applied on both the original un balanced data set and balanced data set with their default parameter and with new adjusted parameter values .The experiments conducted for classification model building using J48 decision tree and Naïve Bayes algorithm and their number of leaves, size of tree, accuracy, TP, TN and precision is presented in table 5.10 as follows.

Table 5.10: summery of experimental result of J48 decision tree and Naïve Bayes

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

	testmode=10-fold dataset=unbalanced attribute=all						
VIII	weka.classifiers.bayes.NaiveBayes testmode=split 66% dataset=unbalanced attribute=all	-	-	85.87%	0.897	0.553	93.99

Key: Exp-experiment**,** NL-number of leaves, ST-size of Tree,C-confidence level, M-minimum number of instance per leaf, U-unpruned, TP-true positive, TN-true negative

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

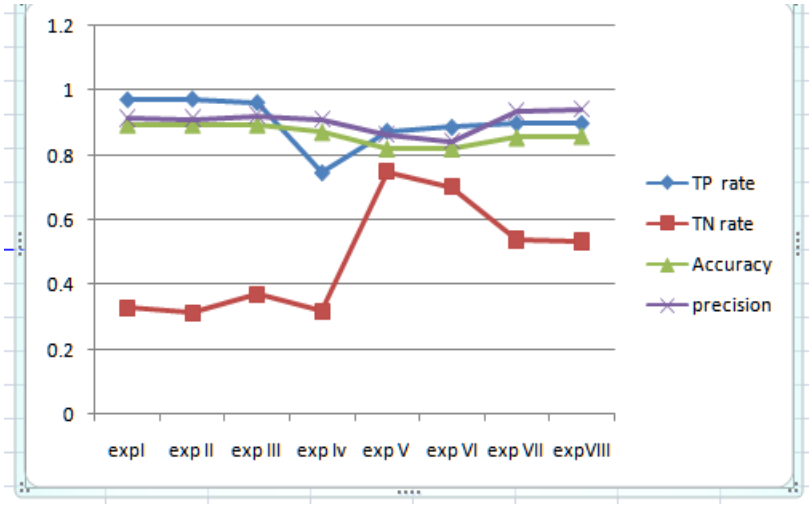

Figure 5.1: comparison of the Experiments conducted

Form the above graph we can see that the true negative is low for all experiments except for experiment V and VI. The reason behind this is that in these two experiments the data set is balanced. Even though experiments I, II and III have higher accuracy than the others but there is a gap in classifying TP and TN (they have low True negative value). In figure 5.1 it shown that there is no much difference in the precision of all the experiment. Therefore experiment V and

VI are selected and compared their accuracy, number of leaves and size of tree. Experiment V has an accuracy of 82.85%, 253 number of leaves and 319 size of tree. Experiment VII has an accuracy of 82.00%, 208 numbers of leaves and 296 size of tree. Even though experiment VI has lower number of leaves and almost equal accuracy with experiment IV, its true negative value is low compared to experiment V. Therefore experiment V is selected as a working model.

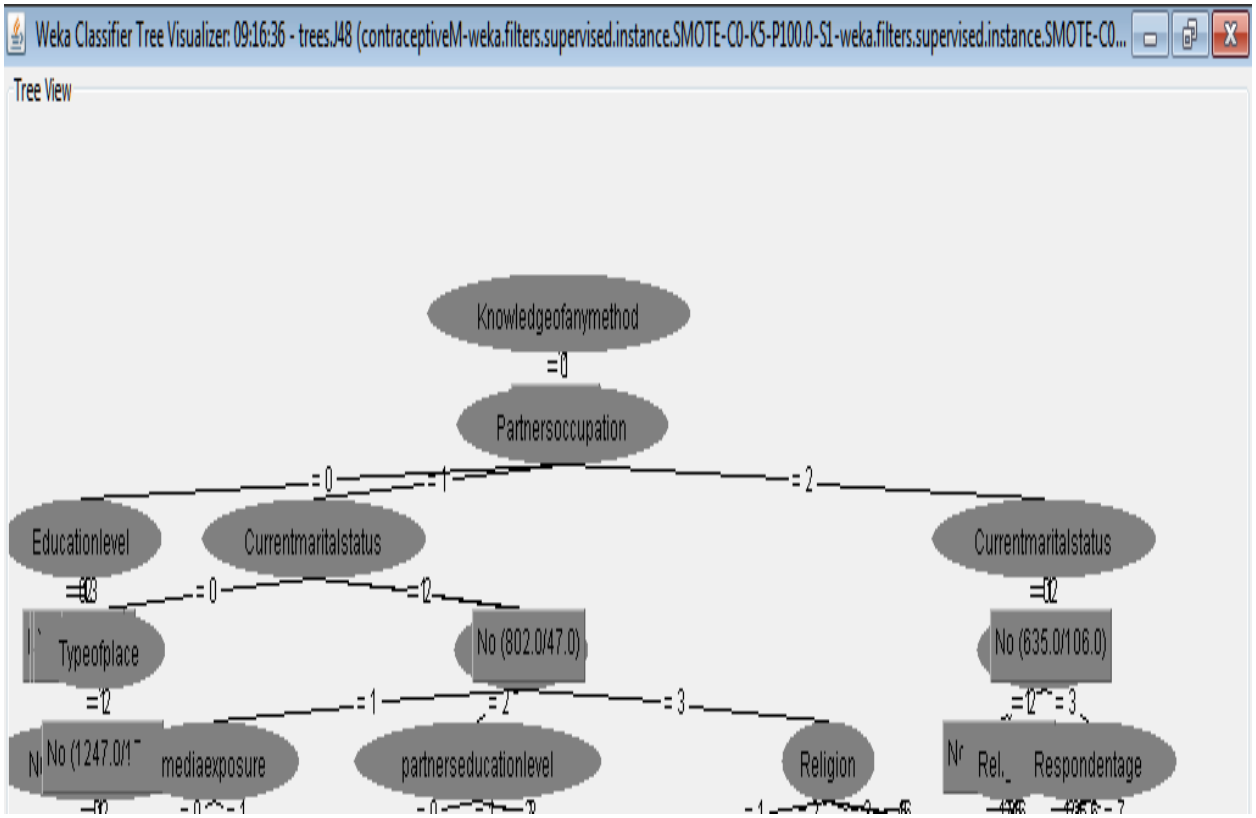

Figure 5.2: A Preview of A Tree Structure Generated From the Selected Model

In this study the best performer is experiment V. From this model a set of rules are extracted simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node.

The following are some of the rules extracted from the decision tree. Therefore those which cover more cases and have better accuracy are chosen. The following rules indicate the possible conditions in which a woman could be classified in each of the non contraceptive method user and contraceptive method user classes.

Rule1.

If Knowledge of any method = No **Then** contraceptive method use: No (2494.0)

Rue 2.

If Knowledge of any method = Yes and partner occupation agric-employee and Current marital status = Never married and Type of place = Rural and Num of living children = no child: **Then** contraceptive method use: No (2175.0/346.0)

Rule 3.

If Knowledge of any method = Yes and partner occupation=agric-employee and Current marital status = Never married and Type of place = Urban: **Then** contraceptive method use: No 1247.0/17.0)

Rule 4.

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

Rule 5.

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

Rule 6

If Knowledge of any method = Yes and partner occupation=Non agric-employee and Current marital status = Living together and Wealth index=rich and Respondent age= 30-34 **Then** contraceptive method use: Yes (702.0/121.0)

Rule 7

if Knowledge of any method = Yes and Partners occupation = Non employee and Education level=No education then contraceptive method use= No (39.0/4.0)

Rule 8

if Knowledge of any method = Yes and Partners occupation = Non employee and Education level=Primary then contraceptive method use= No (18.0/4.0)

rule 9

if Knowledge of any method = Yes and Partners occupation = Non employee and Education level=Secondary then contraceptive method use: Yes (28.0/9.0)

Rule 10

if Knowledge of any method = yes and partner occupation =Agric-employee and current marital status= married and wealth index= poor and FP message=no and region=1(tigray)and Visited by FP =no Then contraceptive method use: No (435.0/115.0)

Rule 11

if Knowledge of any method = yes and partner occupation =Agric-employee and current marital status= married and wealth index= poor and FP message=no and region=1(tigray)and VisitedbyFp = Yes Then contraceptive method use : yes (77.0/29.0)

Rule 12

if Knowledge of any method = yes and partner occupation =Agric-employee and current marital status= married and wealth index= poor and FP message=Yes and region=1(tigray)and partners education level = No education Then contraceptive method use: No (46.0/7.0)

Rule 13

if Knowledge of any method = yes and partner occupation =Agric-employee and current marital status= married and wealth index= poor and FP message=Yes and region=1(tigray)and partners education level = Primary Then contraceptive method use: Yes (45.0/11.0)

Discussion on classification model's generated rules

From the generated rules it is observed that most determinant factors are Knowledge of any method, Partner occupation, Partner's education level, Current marital status, Wealth index, Type of place, FP message, Number of living children, Religion, Education level.

A woman who does not know any contraceptive method has a zero chance of using any method. Another important observation was that knowledge about a family planning could not be the only reason which makes women to use any method of contraception .Even though a woman knows a method there are other factors which make them either to use or not to use. Knowledge of any method, Partner occupation, Current marital status, Partner's education level, Wealth index, Type of place, FP message, Number of living children, Religion, Respondent age are other factors that determine them either to practice family planning or not. For example

A rural or urban women who are not married, have no child and their partner is agriculture employee are less likely to use contraceptive method. It was also observed that women aged 24-34 who are rich and whose partner are a non agric-employee and living together have a high probability of using contraception.

Some of the interesting rules show that how education is so important ; even though women whose husband are non employee are less likely to use any contraceptive method when their education level is improved they have a chance of using contraceptive method. Poor women have low likelihood of utilizing any contraceptive method; this is may be because poor women have low exposure to FP message and have low education compared with the women who are wealthy.

Another interesting rule detected was Visit by Family planning workers and partners education level are important factors; even though women are poor and have no any family planning message they are likely to use contraceptive method if they are visited by FP workers or their partner have at least a primary education

In general when women' s education level ,partner's education level increases they are more likely to use contraceptive method and having a visit by FP workers is also important in order to increase contraceptive method utilization.

The study conducted in Indonesia has used CHAID algorithm on a small number of cases(1473) and nine variables(women's age, education, husband's education , number of children ever born, religion , employment , husband's profession, index of living standard, exposure to media, method of contraception used (no usage, long term method, short term method). In this study 15 attributes are used and the numbers of cases are 14070. The result showed that the most important variable in case of women's choice of contraceptive methods is a husband's profession. It is the same in case of this study husbands' profession is the most determinant variable excluding the variable knowledge about any contraceptive which is in fact not included in the Indonesian dataset. The classification class were short term, long term and no use but in this study the short term and long term are merged and named contraceptive method user and the second class is non user. Another important variable detected was the number of ever born children. After the second child has born a significant change comes and it prevails long term usage. The third important variable is woman's education level and it shows a significant change. When a woman's education increases non usage declines and the long term usage also increases. Woman's age has also an influence in contraceptive method use. When woman's age reaches 32 they intend to use long term methods.

CHAPTER SIX

6. CONCLUSION AND RECOMMENDATION

6.1. Conclusion

Data mining has applied in different health care data to explore important patterns that has overlooked by traditional statistical methods. The purpose of this study is to apply data mining in classifying contraceptive method users and non users among women of reproductive age it aims to identify those actual users and non users in order to trace the hindrance factors and raise awareness of those women with low likelihood of practicing family planning by working on the factors. The data collected by Central Statistics Agency (CSA) is used as a source for this study. Once the data is brought from CSA the KDD process model is followed to construct the classification model. Data understanding and selection, data preprocessing such as handling of missing value, transformation are performed. The other steps followed are data mining task and algorithm selection, interpretation and evaluation.

Different experiments are conducted using decision tree J48 algorithm and Naïve Bayes classification algorithm with their default value and adjusted values. In addition to this techniques like SMOTE is used in order to balance the number of records of each class. The other issue of conducting the experiment is selecting relevant attributes using WEKA built in function. The best first and gain ratio attribute evaluation are used to select the most determinant variables.

The effect of attribute selection is also examined before and after SMOTE is applied using best first attribute selection and gain ratio attribute evaluation method.

An experiment using the Naïve Bayes classification algorithm is also tried to construct a classification model. But the test made using 10- fold and percentage split did score lower accuracy than the J48 algorithm.

By comparing the overall accuracy, respective class accuracy, precision, number of leaves and size of tree experiment V is selected as a working model. Finally rules were generated from the selected model. This model has generated 253 rules and scored an overall accuracy of 82.85%. This model has scored a true positive (classifying non user of contraceptive method) 87.3% and a true negative (classification of contraceptive method user) of 74.7% and a precision of 86.3%.

In this research data mining techniques have revealed an important socio-economic, demographic, geographic, reproductive history and knowledge factors associated with contraceptive method use.

The variables *Knowledge of any method, Partner occupation, Partner's education level, Current marital status, Wealth index, Type of place, FP message, Number of living children, Religion, Education level* were found to be the most determinant attributes of contraceptive method users and non users. By integrating with the knowledge based system the generated rules it is possible to identify those women who are less likely of using contraceptive method and to provide advice on how raise the use of contraceptive method.

6.2. Recommendation

Data mining has applied in health care data to support decision making. In this study two data mining classification techniques are applied on contraceptive method data and a good performance was scored in both techniques. The researcher recommends the following points based on the outcome of the research.

- The techniques employed in this study were decision tree and Naïve Bayes algorithm. Even though an encouraging result was obtained, using other types of techniques with a different parameter might perform better therefore it is recommended other researchers to test with other types of techniques like Artificial neural network, support vector machine etc. it is also possible to classify them according to their choice of CM.
- The researcher believes that other important features which can make this study more interesting were not included in the family planning service providers. Data mining can

have tremendous potential and benefits if health care providers able to capture, store, prepare and mine data .Therefore recording important variable such as partners' occupation, partners' education level and education level etc. while providing service might help for decision making regarding to classify those contraceptive method users and non user based on the various determining factors.

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

REFERENCE

1. EMOH. Family Planning Policy Service. Addis Ababa, Ethiopia: Federal Democratic Republic of Ethiopia, ministry of Health; 2010.
2. EMOH. National Reproductive Health Strategy 2006 –2015. Addis Ababa, Ethiopia: Federal Democratic Republic of Ethiopia, ministry of Health; 2006.
3. CSA. Ethiopia Demographic and Health Survey. Ababa, Ethiopia: CSA; 2005.
4. WHO. Family planning Fact sheet No 351[internet]. [Cited April 2011 l]. Available from: www.who.int/mediacentre/factsheets/fs35/en/index.html
5. EMOH. Family Planning Extension Package. Ethiopia: Ababa, Ethiopia. Federal Democratic Republic of Ethiopia, ministry of Health; 2003.
6. Habtamu A. The health extension program of Ethiopia: summery of concepts, progress, achievement and challenge. Addis Ababa, Ethiopia. WHO Ethiopia country office; 2007.
7. National population policy. [internet][Cited April 2011). Available from: http://www.allbusiness.com/society-social/families-children-family/16272042-1.html
8. Onsembe J. Ethiopia Situation Analysis on Population, Reproductive Health and Gender. Addis Ababa, Ethiopia: UNFPA Country Technical services Team; 2005.
9. Ethiopian Society Population of Studies. Levels, trends and determinants life time and desired fertility in Ethiopia: findings from EDHS 2005.Addis Ababa, Ethiopia; 2008.
10. Aynalem A. population policy and projection. www.*ethiodemographyandhealth.org/*
11. USAID. Application of the Allocate Model in Ethiopia, Policy Project; 2005.
12. Mehmed K. Data Mining: Concept, Model, Methods, and Algorithms.Jhon Wiley &Sons; 2003.
13. Han J, Camber M. Data Mining Concepts and Techniques. Morgan Kaufmann publisher; 2000.
14. Howard Hamilton. Knowledge discovery in database [internet]. [cited April 2011]. Available from: http://www2.cs.uregina.ca/~dbd/cs831/i.
15. USAID. Achieving the MDGs. The contribution of family planning.2009
16. Gaym, A. A review of maternal mortality at Jimma hospital, South-Western Ethiopia. Ethiopian Journal of Health Development;2000

17. Merrick, T.W. Population and poverty: new views on an old controversy. International.2002
18. Family Planning Perspectives.Population Reports. Why Family Planning Matters. [internet]. [Cited June 2011]. Available from: http:// www.Aol.com/astanar/smith.html.
19. The Millennium Development Goals Report. 2009
20. WHO. Unsafe Abortion: Global and Regional Estimates of the Incidence of Unsafe Abortion and Associated Mortality in 2000. Geneva, Switzerland: WHO; 2004.
21. WHO. World Health Report:Make Every Mother and Child Count. Geneva, Switzerland: WHO;2005
22. Fayyad U, Piatetsky G, Smyth P. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence
23. Collier K, Carey B, Grusy E, Marjaniemi C. A Perspective on Data Mining. Northern Arizona University; 1998.
24. Jackson J. Data Mining: A Conceptual Overview.communication of the association system;2002
25. Rud O. Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management.USA: John Wiley & Sons, Inc; 2001.
26. Hand D, Mannila H, Smyth P. Principles of Data Mining. USA: The MIT Press; 2001.
27. Larose D .Discovering Knowledge in Data an Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc; 2005.
28. Chakrabarti S, Cox E, Frank E, Guting R, Han J, Jiang X, et al . Data Mining Know It All. USA: Morgan Kaufmann; 2009.
29. Berry M, Murray B. Lecture Notes in Data Mining.USA: World Scientific Publishing Co. Pte. Ltd; 2006.
30. Guo Y, Grossman R. High performance data mining Scaling Algorithms, Applications and Systems.2002. USA: Kulwer Academic publisher; 1999.
31. Pyle D. Data Preparation for Data Mining. USA: Morgan Kaufmann Publishers, In; 1999.
32. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery, Third Edition.

33. Macro International Inc. Trends in Demographic and Reproductive Health Indicators in Ethiopia: Further analysis of the 2000 and 2005 Demographic and Health Surveys Data. Calverton, Maryland USA; 2007.
34. Getu D. Examination of The Levels and Determinant Factors of Fertility and Contraceptive Use in Northwest Ethiopia: With Special Reference to the Application of the Bongaarts' Model. Addis Ababa, Ethiopia; 2009
35. Abebe G, Nigatu R . Family planning service utilization in Mojo town, Ethiopia: A population based study. Journal of Geography and Regional Planning, June 2011 .Available from: http://www.academicjournals.org/JGRP
36. Birhan research & development consultancy. Knowledge, Attitudes, and Practices in Family Planning Results of a September 2004 Survey In Amhara, Oromia, SNNPR and Tigray Regions of Ethiopia;2005.
37. Wubegzier M.Determinants of low family planning use and high unmet need in Butajira District, South Central Ethiopia. Mekonnen and Worku Reproductive Health 2011, 8:37.
38. S. Aruna, SP. Rajagopalan, L.V. Nadakishore. Knowledge Based Analysis of Various Statistical Tools In Detecting Breast Cancer. CCSEA 2011, CS & IT 02.india;2011
39. Hung Y, McCullagh P, Black N, Harper R. Feature Selection and Classification Model Construction on Type 2 Diabetic Patient's Data.
40. Shegaw A.Application of data Mining technology to predict child mortality patterns the case of Butajura rural health project.2002
41. Lemaire V, Hue C, Bernier O, Vincent L, Carine H and Olivier B. Correlation Analysis in Classifiers Results For the Contraceptive Method Choice Data Set. France.
42. Pejić M Bach and Ćosić D .Data mining usage in health care management: literature survey. Medicinski Glasnik, Volume 5, Number 1, 2008.
43. Bramer M. Principles of Data Mining. London: British Library Cataloguing in Publication Data ; 2007.
44. Measure DHS. Demographic and health survey [internet] .USA: Measure DHS;2011[cited march 2012]. Available from: http://www.measuredhs.com/
45. IPPF. Family planning. [internet].London :;2012.[cited march 2012].Available **From:** http://www.ippf.org/en/Where/et.htm

46. CSA. Function of the CSA [internet].Addis Ababa:CSA;2009.[cited 2012 march].Available From : http://www.csa.gov.et/

Appendices

Appendix A. Initial selected attributes with their initial domain

1. Region :{1:Tigray, 2:Afar, 3: Amhara, 4:oromina, 5:Somalia ,6:Ben-Gumuz,7:SNNP ,12:Gambela, 13: Harari, 14: Addis Ababa ,15: Dire Dawa}
2. Type of Place :{1: rural ,2: urban }
3. Religion:{1:Orthodox, 2:catholoic ,3:Protestant ,4:Muslim ,5:Traditional ,6:other}
4. VisitedbyFPworker:{0:no, 1:yes}
5. Heard FP on tv:{ 0: no ,1 :yes }
6. Heard FP on radio: {0: no ,1 :yes }
7. Heard FP on news paper: {0: no ,1 :yes}
8. Education: {0:no education ,1-primary, 2-secondray ,3-higher}
9. Number of living children as:{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}
10. Knowledge of FP as:{0: knows no FP method,1: knows folkorik, 2-knows traditional,3-knows modern FP method}
11. Marital status: {0-never married, 1:married,2-widowed,3-divorced,4-not living together}
12. Partner's Education:{ 0:no education ,1-primary, 2-secondray ,3-higher }
13. Partner's Occupation: {0: did not work, 1: agric-employee,2: non agric-employee
14. Respondent's Occupation :{0: not working, 1: agric-employee,2:non agric-employee}
15. Wealth index: {1:poorer,2:poor,3:medium, 4:rich, 5:richer}
16. Responden't Age: {1:15-19,2:20-24, 3:25-29, 4:30-34, 5:35-39,6:40-44,7:45-49}
17. currentCMuse {No,Yes}

Appendix B Outcome of decision tree of experiment V

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 20

Relation contraceptiveM-weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1

Instances: 19149

Attributes: 15

Region

Typeofplace

Educationlevel

Religion

VisitedbyFp

partnerseducationlevel

Respondentage

Numoflivingchildren

Knowledgeofanymethod

Currentmaritalstatus

Partnersoccupation

Respondentoccupation

Wealthindex

FPmessage

currentuseofmethod

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Knowledgeofanymethod = 0: No (2494.0)

Knowledgeofanymethod = 1

| Partnersoccupation = 0

| | Educationlevel = 0: No (39.0/4.0)

| | Educationlevel = 1: No (18.0/4.0)

| | Educationlevel = 2: Yes (28.0/9.0)

| | Educationlevel = 3: No (2.0)

| Partnersoccupation = 1

| | Currentmaritalstatus = 0

| | | Typeofplace = 1

| | | | Numoflivingchildren = 0: No (2175.0/346.0)

| | | | Numoflivingchildren = 1

| | | | | Respondentage = 1: No (4.0/1.0)

| | | | | Respondentage = 2: Yes (63.0/12.0)

| | | | | Respondentage = 3: Yes (31.0/12.0)

| | | | | Respondentage = 4: No (9.0/3.0)

| | | | | Respondentage = 5: No (14.0/1.0)

| | | | | Respondentage = 6: No (3.0/1.0)

| | | | | Respondentage = 7: No (2.0)

| | | | Numoflivingchildren = 2: No (2.0)

| | | Typeofplace = 2: No (1247.0/17.0)

| | Currentmaritalstatus = 1

| | | Wealthindex = 1

| | | | FPmessage = 0

| | | | | Region = 1

| | | | | | VisitedbyFp = 0: No (435.0/115.0)

| | | | | | VisitedbyFp = 1: Yes (77.0/29.0)

| | | | | Region = 2: No (184.0/1.0)

| | | | | Region = 3

| | | | | | partnerseducationlevel = 0

| | | | | | | Religion = 1

| | | | | | | | Respondentage = 1: No (27.0/7.0)

| | | | | | | | Respondentage = 2: No (54.0/21.0)

| | | | | | | | Respondentage = 3: No (54.0/10.0)

| | | | | | | | Respondentage = 4: No (50.0/18.0)

| | | | | | | | Respondentage = 5

| | | | | | | | | Respondentoccupation = 0: No (45.0/14.0)

| | | | | | | | | Respondentoccupation = 1: Yes (30.0/5.0)

| | | | | | | | | Respondentoccupation = 2: No (3.0)

| | | | | | | | Respondentage = 6: No (24.0)

| | | | | | | | Respondentage = 7: No (23.0/1.0)

| | | | | | | Religion = 2: No (0.0)

| | | | | | | Religion = 3: Yes (6.0)

| | | | | | | Religion = 4: No (88.0/25.0)

| | | | | | | Religion = 5: No (2.0)

| | | | | | | Religion = 6: No (0.0)

| | | | | | partnerseducationlevel = 1: No (39.0/8.0)

| | | | | | partnerseducationlevel = 2: Yes (11.0/2.0)

| | | | | | partnerseducationlevel = 3: No (0.0)

| | | | | Region = 4: No (516.0/66.0)

| | | | | Region = 5: No (67.0)

| | | | | Region = 6: No (81.0/12.0)

| | | | | Region = 7

| | | | | | Educationlevel = 0: No (283.0/32.0)

| | | | | | Educationlevel = 1: No (25.0/1.0)

| | | | | | Educationlevel = 2: Yes (8.0/1.0)

| | | | | | Educationlevel = 3: No (0.0)

| | | | | Region = 12: No (65.0/11.0)

| | | | | Region = 13: No (44.0)

| | | | | Region = 14: No (2.0)

| | | | | Region = 15: No (88.0/27.0)

| | | | FPmessage = 1

| | | | | Region = 1

| | | | | | partnerseducationlevel = 0: No (46.0/7.0)

| | | | | | partnerseducationlevel = 1: Yes (45.0/11.0)

| | | | | | partnerseducationlevel = 2: Yes (4.0/1.0)

| | | | | | partnerseducationlevel = 3: No (0.0)

| | | | | Region = 2: No (20.0/1.0)

| | | | | Region = 3: Yes (110.0/37.0)

| | | | | Region = 4

| | | | | | Educationlevel = 0: No (117.0/32.0)

| | | | | | Educationlevel = 1: Yes (25.0/9.0)

| | | | | | Educationlevel = 2: Yes (1.0)

| | | | | | Educationlevel = 3: No (0.0)

| | | | | Region = 5: No (6.0)

| | | | | Region = 6: No (9.0/3.0)

| | | | | Region = 7: No (19.0/2.0)

| | | | | Region = 12: No (4.0/1.0)

| | | | | Region = 13: No (8.0)

| | | | | Region = 14: No (0.0)

| | | | | Region = 15: No (6.0)

| | | Wealthindex = 2

| | | | partnerseducationlevel = 0

| | | | | Region = 1

| | | | | | Numoflivingchildren = 0: No (7.0)

| | | | | | Numoflivingchildren = 1: No (20.0/1.0)

| | | | | | Numoflivingchildren = 2

| | | | | | | Respondentoccupation = 0

| | | | | | | | Respondentage = 1: Yes (0.0)

| | | | | | | | Respondentage = 2: No (1.0)

| | | | | | | | Respondentage = 3: Yes (21.0/5.0)

| | | | | | | | Respondentage = 4: No (7.0/1.0)

| | | | | | | | Respondentage = 5: Yes (34.0/6.0)

| | | | | | | | Respondentage = 6: Yes (22.0/5.0)

| | | | | | | | Respondentage = 7: No (8.0/3.0)

| | | | | | | Respondentoccupation = 1: No (37.0/12.0)

| | | | | | | Respondentoccupation = 2: No (5.0/1.0)

| | | | | Region = 2: No (12.0)

| | | | | Region = 3

| | | | | | Respondentage = 1: No (25.0/5.0)

| | | | | | Respondentage = 2: Yes (79.0/31.0)

| | | | | | Respondentage = 3: No (57.0/27.0)

| | | | | | Respondentage = 4: No (31.0/6.0)

| | | | | | Respondentage = 5: No (26.0/3.0)

| | | | | | Respondentage = 6: No (43.0/20.0)

| | | | | | Respondentage = 7: No (34.0/9.0)

| | | | | Region = 4: No (177.0/43.0)

| | | | | Region = 5: No (6.0)

| | | | | Region = 6

| | | | | | Respondentage = 1: Yes (3.0/1.0)

| | | | | | Respondentage = 2: No (14.0/1.0)

| | | | | | Respondentage = 3: No (15.0/3.0)

| | | | | | Respondentage = 4: Yes (24.0/10.0)

| | | | | | Respondentage = 5: Yes (27.0/9.0)

| | | | | | Respondentage = 6: No (16.0/5.0)

| | | | | | Respondentage = 7: No (4.0)

| | | | | Region = 7: No (132.0/14.0)

| | | | | Region = 12: No (9.0)

| | | | | Region = 13: No (32.0/1.0)

| | | | | Region = 14: No (3.0/1.0)

| | | | | Region = 15: Yes (34.0/14.0)

| | | | partnerseducationlevel = 1

| | | | | Respondentage = 1

| | | | | | Numoflivingchildren = 0: Yes (30.0/13.0)

| | | | | | Numoflivingchildren = 1: No (26.0/10.0)

| | | | | | Numoflivingchildren = 2: No (1.0)

| | | | | Respondentage = 2

| | | | | | Educationlevel = 0

| | | | | | | FPmessage = 0

| | | | | | | | Region = 1: No (5.0/1.0)

| | | | | | | | Region = 2: No (0.0)

| | | | | | | | Region = 3: No (3.0)

| | | | | | | | Region = 4: Yes (21.0/6.0)

| | | | | | | | Region = 5: No (0.0)

| | | | | | | | Region = 6: No (3.0)

| | | | | | | | Region = 7: No (20.0/8.0)

| | | | | | | | Region = 12: No (1.0)

| | | | | | | | Region = 13: No (1.0)

| | | | | | | | Region = 14: No (0.0)

| | | | | | | | Region = 15: Yes (1.0)

| | | | | | | FPmessage = 1: Yes (57.0/10.0)

| | | | | | Educationlevel = 1: No (20.0/4.0)

| | | | | | Educationlevel = 2: No (1.0)

| | | | | | Educationlevel = 3: Yes (0.0)

| | | | | Respondentage = 3

| | | | | | Respondentoccupation = 0: Yes (184.0/60.0)

| | | | | | Respondentoccupation = 1: No (20.0/2.0)

| | | | | | Respondentoccupation = 2: No (14.0/3.0)

| | | | | Respondentage = 4: No (56.0/5.0)

| | | | | Respondentage = 5: No (47.0/18.0)

| | | | | Respondentage = 6: No (28.0/9.0)

| | | | | Respondentage = 7: No (11.0)

| | | | partnerseducationlevel = 2

| | | | | Respondentage = 1: No (8.0/3.0)

| | | | | Respondentage = 2: No (20.0/8.0)

| | | | | Respondentage = 3: No (9.0/1.0)

| | | | | Respondentage = 4: No (13.0/2.0)

| | | | | Respondentage = 5: Yes (27.0/4.0)

| | | | | Respondentage = 6: No (1.0)

| | | | | Respondentage = 7: No (0.0)

| | | | partnerseducationlevel = 3: No (1.0)

| | | Wealthindex = 3

| | | | Religion = 1

| | | | | Numoflivingchildren = 0

| | | | | | partnerseducationlevel = 0: No (32.0/4.0)

| | | | | | partnerseducationlevel = 1: No (35.0/11.0)

| | | | | | partnerseducationlevel = 2: Yes (14.0/4.0)

| | | | | | partnerseducationlevel = 3: Yes (2.0)

| | | | | Numoflivingchildren = 1

| | | | | | Respondentage = 1

| | | | | | | Educationlevel = 0: No (22.0/7.0)

| | | | | | | Educationlevel = 1: Yes (21.0/5.0)

| | | | | | | Educationlevel = 2: Yes (0.0)

| | | | | | | Educationlevel = 3: Yes (0.0)

| | | | | | Respondentage = 2: Yes (148.0/54.0)

| | | | | | Respondentage = 3

| | | | | | | FPmessage = 0: No (38.0/3.0)

| | | | | | | FPmessage = 1: Yes (28.0/12.0)

| | | | | | Respondentage = 4: No (9.0/1.0)

| | | | | | Respondentage = 5: No (7.0)

| | | | | | Respondentage = 6: No (10.0)

| | | | | | Respondentage = 7: No (8.0)

| | | | | Numoflivingchildren = 2

| | | | | | Respondentage = 1: Yes (1.0)

| | | | | | Respondentage = 2: Yes (61.0/16.0)

| | | | | | Respondentage = 3

| | | | | | | partnerseducationlevel = 0: No (63.0/21.0)

| | | | | | | partnerseducationlevel = 1: Yes (91.0/24.0)

| | | | | | | partnerseducationlevel = 2: Yes (61.0/11.0)

| | | | | | | partnerseducationlevel = 3: Yes (0.0)

| | | | | | Respondentage = 4

| | | | | | | partnerseducationlevel = 0: Yes (137.0/44.0)

| | | | | | | partnerseducationlevel = 1: No (29.0/6.0)

| | | | | | | partnerseducationlevel = 2: Yes (9.0/4.0)

| | | | | | | partnerseducationlevel = 3: Yes (0.0)

| | | | | | Respondentage = 5

| | | | | | | Region = 1: Yes (39.0/6.0)

| | | | | | | Region = 2: Yes (1.0)

| | | | | | | Region = 3: Yes (132.0/34.0)

| | | | | | | Region = 4: No (28.0/12.0)

| | | | | | | Region = 5: Yes (0.0)

| | | | | | | Region = 6: Yes (4.0/1.0)

| | | | | | | Region = 7: No (13.0/2.0)

| | | | | | | Region = 12: Yes (10.0/2.0)

| | | | | | | Region = 13: Yes (1.0)

| | | | | | | Region = 14: Yes (0.0)

| | | | | | | Region = 15: Yes (0.0)

| | | | | | Respondentage = 6

| | | | | | | VisitedbyFp = 0: No (112.0/48.0)

| | | | | | | VisitedbyFp = 1: Yes (27.0/4.0)

| | | | | | Respondentage = 7: No (81.0/19.0)

| | | | Religion = 2: No (17.0/6.0)

| | | | Religion = 3

| | | | | Region = 1: Yes (0.0)

| | | | | Region = 2: Yes (0.0)

| | | | | Region = 3: Yes (11.0)

| | | | | Region = 4

| | | | | | Educationlevel = 0

| | | | | | | partnerseducationlevel = 0: Yes (50.0/11.0)

| | | | | | | partnerseducationlevel = 1: No (27.0/4.0)

| | | | | | | partnerseducationlevel = 2: Yes (11.0/4.0)

| | | | | | | partnerseducationlevel = 3: Yes (0.0)

| | | | | | Educationlevel = 1: Yes (100.0/14.0)

| | | | | | Educationlevel = 2: No (5.0/1.0)

| | | | | | Educationlevel = 3: Yes (0.0)

| | | | | Region = 5: Yes (0.0)

| | | | | Region = 6: No (22.0/8.0)

| | | | | Region = 7

| | | | | | FPmessage = 0

| | | | | | | Numoflivingchildren = 0: No (5.0)

| | | | | | | Numoflivingchildren = 1

| | | | | | | | Educationlevel = 0: No (25.0/2.0)

| | | | | | | | Educationlevel = 1: Yes (34.0/15.0)

| | | | | | | | Educationlevel = 2: Yes (1.0)

| | | | | | | | Educationlevel = 3: No (0.0)

| | | | | | Numoflivingchildren = 2

| | | | | | | | Respondentage = 1: Yes (0.0)

| | | | | | | | Respondentage = 2: No (7.0/2.0)

| | | | | | | | Respondentage = 3

| | | | | | | | | partnerseducationlevel = 0: Yes (26.0/5.0)

| | | | | | | | | partnerseducationlevel = 1: Yes (44.0/14.0)

| | | | | | | | | partnerseducationlevel = 2: No (6.0/1.0)

| | | | | | | | | partnerseducationlevel = 3: Yes (0.0)

| | | | | | | | Respondentage = 4: Yes (58.0/20.0)

| | | | | | | | Respondentage = 5: No (24.0/4.0)

| | | | | | | | Respondentage = 6: Yes (26.0/11.0)

| | | | | | | | Respondentage = 7: No (21.0/7.0)

| | | | | | | FPmessage = 1: No (62.0/13.0)

| | | | | | Region = 12: Yes (34.0/14.0)

| | | | | | Region = 13: Yes (0.0)

| | | | | | Region = 14: No (1.0)

| | | | | | Region = 15: Yes (0.0)

| | | | | Religion = 4

| | | | | | Educationlevel = 0: No (440.0/124.0)

| | | | | | Educationlevel = 1

| | | | | | FPmessage = 0: No (74.0/21.0)

| | | | | | FPmessage = 1

| | | | | | | partnerseducationlevel = 0: Yes (37.0/9.0)

| | | | | | | partnerseducationlevel = 1: Yes (37.0/13.0)

| | | | | | | partnerseducationlevel = 2: No (8.0)

| | | | | | | partnerseducationlevel = 3: Yes (0.0)

| | | | | Educationlevel = 2: Yes (9.0/4.0)

| | | | | Educationlevel = 3: No (0.0)

| | | | Religion = 5: No (14.0/1.0)

| | | | Religion = 6: No (11.0/1.0)

| | Currentmaritalstatus = 2: No (802.0/47.0)

| Partnersoccupation = 2

| | Currentmaritalstatus = 0: Yes (1.0)

| | Currentmaritalstatus = 1

| | | Wealthindex = 1: No (123.0/26.0)

| | | Wealthindex = 2

| | | | Religion = 1

| | | | | Numoflivingchildren = 0: No (7.0/1.0)

| | | | | Numoflivingchildren = 1: Yes (21.0/10.0)

| | | | | Numoflivingchildren = 2: Yes (40.0/14.0)

| | | | Religion = 2: No (1.0)

| | | | Religion = 3: No (10.0)

| | | | Religion = 4: No (34.0/16.0)

| | | | Religion = 5: No (0.0)

| | | | Religion = 6: No (0.0)

| | | Wealthindex = 3

| | | | Respondentage = 1: Yes (221.0/55.0)

| | | | Respondentage = 2: Yes (1102.0/170.0)

| | | | Respondentage = 3: Yes (1305.0/200.0)

| | | | Respondentage = 4: Yes (702.0/121.0)

| | | | Respondentage = 5

| | | | | Numoflivingchildren = 0: No (19.0/3.0)

| | | | | Numoflivingchildren = 1: Yes (118.0/29.0)

| | | | | Numoflivingchildren = 2

| | | | | | FPmessage = 0

| | | | | | | Religion = 1: Yes (42.0/7.0)

| | | | | | | Religion = 2: Yes (0.0)

| | | | | | | Religion = 3: Yes (9.0/4.0)

| | | | | | | Religion = 4: No (21.0/4.0)

| | | | | | | Religion = 5: Yes (0.0)

| | | | | | | Religion = 6: Yes (0.0)

| | | | | | FPmessage = 1: Yes (418.0/49.0)

| | | | Respondentage = 6

| | | | | Numoflivingchildren = 0: No (10.0)

| | | | | Numoflivingchildren = 1: No (43.0/15.0)

| | | | | Numoflivingchildren = 2: Yes (251.0/68.0)

| | | | Respondentage = 7

| | | | | Region = 1: No (7.0/1.0)

| | | | | Region = 2: No (6.0)

| | | | | Region = 3: No (12.0/3.0)

| | | | | Region = 4: No (18.0/6.0)

| | | | | Region = 5: No (0.0)

| | | | | Region = 6: No (2.0)

| | | | | Region = 7: No (13.0/1.0)

| | | | | Region = 12: No (0.0)

| | | | | Region = 13: Yes (25.0/11.0)

| | | | | Region = 14

| | | | | | Educationlevel = 0: Yes (47.0/13.0)

| | | | | | Educationlevel = 1: No (18.0/4.0)

| | | | | | Educationlevel = 2: Yes (26.0/10.0)

| | | | | | Educationlevel = 3: No (13.0/5.0)

| | | | | Region = 15: No (30.0/5.0)

| | Currentmaritalstatus = 2: No (635.0/106.0)

Number of Leaves : 253

Size of the tree : 319

Time taken to build model: 25.25 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	15864	82.8451 %
Incorrectly Classified Instances	3285	17.1549 %

Kappa statistic 0.6229

Mean absolute error 0.2501

Root mean squared error 0.3583

Relative absolute error 54.7039 %

Root relative squared error 74.9424 %

Total Number of Instances 19149

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.873	0.253	0.863	0.873	0.868	0.881	No
	0.747	0.127	0.763	0.747	0.755	0.881	Yes
Weighted Avg.	0.828	0.208	0.828	0.828	0.828	0.881	

=== Confusion Matrix ===

a	b	<-- classified as
10806	1571	a = No
1714	5058	b = Yes