ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

DEVELOPING A PREDICTIVE MODEL FOR FERTILITY PREFERENCE OF WOMEN OF REPRODUCTIVE AGE USING DATA MINING TECHNIQUES

By
TARIKU DEBELA

A thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

JANUARY, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

DEVELOPING A PREDICTIVE MODEL FOR FERTILITY PREFERENCE OF WOMEN OF REPRODUCTIVE AGE USING DATA MINING TECHNIQUES

By

TARIKU DEBELA

A thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

Advisors:
Ato Getachew Jemaneh
Melesse Tamiru (PhD)

JANUARY, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

DEVELOPING A PREDICTIVE MODEL FOR FERTILITY PREFERENCE OF WOMEN OF REPRODUCTIVE AGE USING DATA MINING TECHNIQUES

By

TARIKU DEBELA

JANUARY, 2013

A thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

Name and Signature of Members of the Examining Board

Name	Title	Date	Signature
_____	**Chairman**	_____	_____
_____	**Advisor**	_____	_____
_____	**Advisor**	_____	_____
_____	**Examiner**	_____	_____

DECLARATION

I declare that this thesis is my original work and has not been submitted as a partial requirement for a degree in any university

Name: _____

Signature: _____

Date: _____

This Thesis has been submitted for examination with my approval as a university advisor

Getachew Jemaneh (Ato)	Melesse Tamiru (Dr.)
Signature: _____	Signature: _____
Date: _____	Date: _____

DEDICATION

I would like to dedicate this thesis to my parents, brothers and sisters for all their love and support

ACKNOWLEDGEMENTS

First of all, I would like to be grateful to my almighty God for his unreserved provision during this research work.

I am honestly thankful to my advisors, Ato Getachew Jemaneh and Dr. Melesse Tamiru for their advice, guidance, constructive and on time comments while undertaking this study.

I also forward my sincere gratitude to CSA staff for identifying the experimental dataset for this study and their general comments on the nature of the dataset.

I specially appreciate my brothers who encouraged and supported me in accomplishing this thesis.

Lastly, I would like to thank Berhanu Atnafu, my classmate, for his friendly and cooperative approach while doing group project works, assignments and general academic discussions in the last two years.

TABLE OF CONTENTS

LIST OF FIGURES

Title **page**

LIST OF TABLES

Title **Page**

ACRONYMS

ANN	Artificial Neural Network
APPGPDRH	All Party Parliamentary Group on Population Development and Reproductive Health
ARFF	Attribute Relation File Format
ASCII	American Standard Code for Information Interchange
AUC	Area Under the Curve
CAD	Coronary Artery Disease
CDW	Clinical Data Warehouse
CSA	Central Statistical Agency
DHS	Demographic and Health Survey
DM	Data Mining
DSS	Demographic Surveillance System
EDHS	Ethiopian Demographic and Health Survey
GDP	Gross Domestic Product
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome
ID3	Iterative Dichotomiser
KDD	Knowledge Discovery in Database
KDP	Knowledge Discovery Process
MDGs	Millennium Development Goals
MICS	Multiple Indicator Cluster Survey
MLP	Multilayer Perceptron
RDBMS	Relational Database Management System
ROC	Receiver Operating Characteristics
SEERS	Surveillance Epidemiology and End Results
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-sampling Technique
SNNP	Southern Nations Nationalities and Peoples
SPSS	Statistical Package for the Social Sciences
SRH	Sexual and Reproductive Health
SVM	Support Vector Machine
TFR	Total Fertility Rate
TN	True Negative
TP	True Positive
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

Background: Fertility is one of the major factors that determine the overall size, distribution and/or structure of a population. High fertility in developing countries(particularly, in the poorest of those countries) poses detrimental consequences like a high fraction of women experiencing pregnancies of order five and above and a greater likelihood of short inter-pregnancy intervals. These are threat to the health of mothers and their children. At a macro-level, high fertility also contributes to high population growth which in turn results in slow economic growth, environmental degradation and unemployment, among others. Assessing fertility preference helps identify the proportion of women who demand for children and those who intend to limit childbearing. This aids in developing and implementing appropriate intervention programs for the purpose of achieving reductions in fertility levels necessary to slow population growth.

Objective: To explore the possibility of applying data mining techniques in developing a model that can predict fertility preferences of women of reproductive age from EDHS2011 women's survey dataset collect by CSA.

Methodology: For this study, a six-step hybrid knowledge discovery process model was adopted. Through the steps, a dataset containing 15 attributes and 16515 records of women was constructed for building models.

Results: Three data mining classification algorithms, J48, Naïve Byes and neural Network (Multilayer Perceptron), were tested using 10-fold-cross-validation. The classifiers were implemented on the dataset with all and selected features. Several experiments were constructed and the accuracy achieved on selected feature subset was 75.92%, 77.34%, 78.03% for Naïve Bayes, Multilayer Perceptron and J48, respectively.

Conclusion: In this study, feature selection generally improved prediction performance of the classifiers. J48 model with accuracy of 78.03% was found to be relatively better predictor of fertility preference of women. This research study did indicate that data mining can be applied to women's dataset to identify determinants of fertility preference and classify women according to their childbearing preferences. Age, number of living children, education, child death experience, marital status, sex of child and region are found to be the most important factors that determine fertility preference of women.

CHAPTER ONE

INTRODUCTION

1.1 Human Population Dynamics

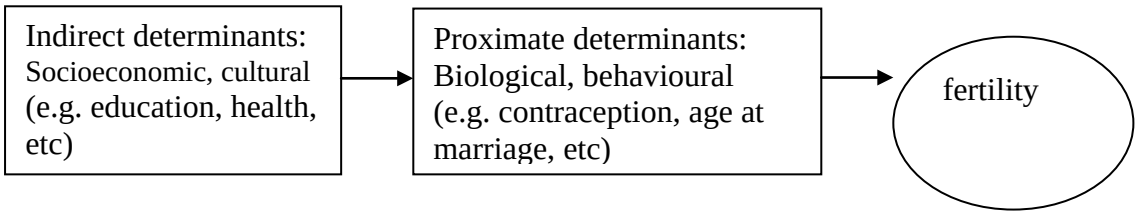
A human population is a group of people living in the same geographic area. Human populations undergo three distinct phases of their life cycle: growth, stability, and decline. Generally, the study of factors that affect growth, stability, and decline of populations is referred to as population dynamics. Human population dynamics aids in tracking factors related to changes in population. Predicting population changes is important because these changes can affect economic, social, and environmental systems. An increase in human population can, for example, impact the quality of natural resources like biodiversity, air, land, and water [1].

Populations can change through three fundamental processes, fertility, mortality, and migration, which determine the overall size, distribution and/or structure of a population and have consequences on one another [2]. Fertility involves the number of children that women have, and it is a function of a woman's fecundity (her physiological ability to conceive and bear children and of social, cultural, economic, and health factors that influence reproductive choices in the country in consideration [1].

Fertility level of a nation is indicated by the total fertility rate (TFR) which is defined as the number of children a woman would have by the end of her child bearing years if she were to pass through those years bearing children at the observed age-specific fertility rates. The basic factors influencing a country's total fertility rate include relationship status (the fraction of women who are married or in a relationship that exposes them to the possibility of becoming pregnant); use of contraception; the fraction of women who are infecund —for example, because they are breastfeeding a child; and the prevalence of induced abortion [3].

1.2 Determinants of Fertility

Factors that affect fertility are basically categorized into proximate (direct) and socioeconomic and environmental (indirect or background) variables. The indirect determinants include the social, cultural, economic, institutional, psychological, health, and environmental variables, and the proximate determinants consist of all biological and behavioural factors through which the background variables must operate to affect fertility. The proximate determinants have direct influence on fertility, whereas, socioeconomic variables can affect fertility only indirectly by modifying the proximate determinants [4].



Bongaarts, et.al.[5] noted that the proximate variables improve understanding of the operation of the socioeconomic determinants in the study of fertility. In general, a socioeconomic variable can have negative fertility effects through one set of proximate variables and positive effects through another set. The overall net effect of a socioeconomic variable on fertility can therefore be positive, negative, or insignificant depending on the relative contributions of the positive and negative effects of the proximate determinant.

The following is a brief description of some of Bongaarts' proximate determinants:

Proportion of women married or in sexual unions - This variable measures the degree to which women of reproductive age are exposed to the risk of conceiving.

Frequency of intercourse - This determinant directly affects the probability of conceiving among ovulating women.

Postpartum abstinence - Prolonged abstinence from sexual relations while a newborn is breastfeeding.

Lactational amenorrhea - Following a pregnancy a woman remains unable to conceive until the normal pattern of ovulation and menstruation is restored.

Contraception - Any practice undertaken deliberately to reduce the risk of conception

Induced abortion - This includes any practice that deliberately interrupts the normal course of gestation.

Spontaneous intrauterine mortality - A proportion of all conceptions fail to end in a live birth because some pregnancies spontaneously terminate prematurely in a miscarriage or stillbirth.

Pathological sterility - A number of diseases, especially gonorrhea, can cause primary or secondary sterility. Primary sterility results in childlessness because a sterilizing disease is contracted before a first birth. Secondary sterility results in an inability to bear additional children, sometimes very early in the childbearing years, and is due to the onset of disease among women who already have borne offspring.

1.2 Consequences of High Fertility

Most of sub-Saharan Africa countries, Ethiopia being one of them, are characterized by high fertility. High fertility – defined as a total fertility rate (TFR) of five or more births per woman over her reproductive age – poses detrimental consequences like health risks for children and their mothers, less human capital investment, slow economic growth, and threats to the environment. The high-fertility countries lag in many development indicators, as reflected for example in their rate of progress toward achievement of the Millennium Development Goals (MDGs). High total fertility has micro-and-macro-level demographic consequences; high incidence of births of order five and above, a high fraction of women experiencing pregnancies of order five and above, and a greater likelihood of short inter-pregnancy intervals are micro-level consequences whereas, at the macro-level, rapid population growth rate with the concomitant rapid growth in the size of successive birth cohorts is considered the main demographic feature [6].

Mahy [7] analysed Demographic and Health Survey (DHS) data for the risks of death during four birth intervals, neonatal, infant, early childhood and under-five, and found that children from higher-order births (due to high fertility) are known to be at greater risk of dying during infancy and early childhood. Short inter-pregnancy intervals – a characteristic of high-fertility counties- also substantially contributes to child mortality, low birth weight, preterm birth and small size for gestational age. [8] Maternal mortality is a fundamental measure of a country's overall health and development status. This is also more likely at higher pregnancy orders. Women with five or more pregnancies have a significantly higher risk of dying due to

maternal causes. Women at pregnancy orders five and six suffer higher mortality than women of fewer pregnancies, i.e., since pregnancy is an absolute requirement for maternal mortality fewer pregnancies lower the lifetime risk for mothers [9].

The effect of high fertility is also reflected in human capital investment, for example, in formal schooling of children. Based on quantity-quality trade-off, parents consciously decide to have fewer children in order to invest more per child for schooling. If there is high fertility or population growth in a country, or if a child has many siblings, the chance of that child going to school is reduced, that is, large and growing child cohorts exert downward pressure on schooling expenditure per child [10].

Fertility and economic growth are also interrelated. In general, there is a negative correlation between fertility and economic growth. Even so, this simple correlation cannot be regarded as revealing the true causal relationship between fertility and economic growth. Fertility has a negative impact on productive output, reflecting expenditure on child-rearing rather than production of goods (income generation). When the effect of the overall level of fertility and population growth rate is considered - rather than growth rates of different age-strata of the population- drop in fertility raises productive output in the long run [11].

Policy settings that support growth are primarily the key drivers of economic growth, while population size and structure play an important secondary role in facilitating or hindering economic growth. Coale and Hoover [12] analysed the relationship between population growth and economic development for India in the 1950s ─ characterized at the time by low GDP growth, low industrialization, and heavy reliance on subsistence agriculture (typical of Ethiopia today) ─ and concluded that population growth might adversely affect the prospects for economic development because of the population‘s increasing size and its structure with high and rising child dependency ratios. They argued that the combined effect of these two factors divert national resources away from investment in expanding production and increasing the capital/labor ratio ― to meet the growing needs for schools, health, housing, and other infrastructure needed to avoid compromising the future population‘s wellbeing and productivity. Similarly at the household level, they divert resources away from saving for productive investment, to meet current consumption needs.

The effect of fertility and other demographic factors on the natural environment is determined by institutional factors such as land-tenure regulations and agricultural practices and by consumption patterns. This effect varies significantly across regions and even between localities. Even so, High fertility with the concomitant high population growth rate has an impact on soil erosion, soil nutrient depletion, land use pattern, quality and amount of water and the quality of air. The effect of high fertility on natural resources is more noticeable in Africa. Migration within the region is being exacerbated by water scarcity, drought, and land degradation [13].

1.4 Fertility Trends in Ethiopia

Fertility is one of the elements in population dynamics that has significant contribution towards changing population size and structure over time. The sum of age-specific fertility rates (known as the total fertility rate, or TFR) is a summary measure of the level of fertility. It is the number of children a woman would have by the end of her childbearing years if she were to pass through those years bearing children at the current observed age-specific rates. In Ethiopia, fertility has been at its highest levels at the end of the twentieth century. Total fertility rate increased between the 1970s and early 1990s from about 5.2 children per woman in 1970 to 6.4 in 1990. Since then, however, it has been moderately declining, mostly in urban areas. The three-year rate preceding the EDHS survey declined from 6.4 children per woman in 1990 to 5.5 children per woman in 2000. According to the 2005 EDHS, TFR was 5.4 children per woman in 2005. It can be noticed that the last 15 years, since 1990, TFR declined by only one child per woman [14].

Currently, Ethiopia is at its earliest stage of fertility transition and has intermediate levels of fertility (a TFR of about 5). The 2011 TFR estimate (4.8) shows a decline in TFR from the estimates reported in the 2005 EDHS (5.4) and the 2000 EDHS (5.5). The decline in fertility in the last five years is due to a decrease in fertility in rural areas; among rural women the TFR decreased from 6.0 children in the 2005 EDHS to the current level of 5.5. Although the growth rate appears to have begun a downward trend from mid 1990s, the speed of the decline is very slow and even by 2020 the rate of growth of the population is unlikely to be any lower than 1.3% per year [15]. The youthful age structure, the outcome of high fertility levels, contributes to a continuing future rapid population growth.

1.5 Fertility Preferences

Recently, considerable attention has been paid in the demographic literature to subjective preferences, intentions, ideals, and expectations toward fertility. Fertility preference studies have looked at peoples' preferences using different terminologies and definitions. Desired family size, ideal number of children, fertility preference, desire for additional children and fertility intentions are some of the measures that have been used to describe and/or estimate the number of children that people actually want to have. Although they may happen to be quite similar, they are normally not identical. Desired family size refers to the number of children an individual would have liked to have in his/her whole life irrespective of the number he/she already has. Responses to questions about desire for additional children also referred to as fertility or reproductive intention is considered as fertility preference in the demographic and health surveys (DHS). Fertility intention is often referred to as fertility expectation. Ideal family size describes the existence of a societal norm regarding family size, while expected size describes a personal norm [16].

The predictive value of these measures on actual fertility is that it is usually based on the desires of women respondents, whereas studies have shown that fertility intentions of their husbands or partners do matter and has a great influence on actual fertility outcome. Despite slight difference among them, these preference measures continue to be very relevant because of their importance in the estimation of actual fertility [16].

People's preferences for fertility have a predictive value for fertility and might tell how many children they would eventually have. Research in fertility preferences, particularly in developing countries, is reconsidered because of its relationship with and important bearing on the complex family building processes. The subject of parental attitudes and aspirations in relation to household fertility decision-making have gained importance in recent fertility researches since these seem to be related to the future course of fertility in a society. Fertility preferences/desires and intentions are central in theoretical and empirical approaches to studying childbearing behavior. Fertility declines when childbearing becomes a subject of conscious choice, that is, when having children becomes a subject about which it is possible to have preferences. Measuring fertility intentions and determining the extent to which they predict fertility behavior is also important for population policy and the implementation of family planning programs [16-17].

The measurement of fertility preference is essential for understanding the dynamics of fertility change. This measurement is important for forecasting medium-term changes in fertility and to estimate assess the prevalence of unwanted births, and thus the prevalence of an unmet need for family planning services [18**]**.

Factors affecting fertility preference are divided into two general areas of variables: Institution/Policy variables and Socio-economic variables. Institution is identified as organised and regulated systems that operate in the basic areas of societal life such as government institutions which organize laws, policies and state power, religious institutions, border institutions which organize the distribution and allowance of goods and services, and institutions that transmit knowledge such as schools and universities. Policy is identified as the social policies put into place by the government institution. Other institution/policy variables include the removal of social security nets such as government-guaranteed employment and assigned housing. Socio-economic variables are identified as demographic factors related to the socio- economic status of an individual such as level of education, income, housing, age, and residence identity. Institution/policy and socio-economic variables are highly inter-connected that one variable is not exclusive of another. It is due to the interconnected nature of the variables that there is no one single factor that is impacting fertility preference; instead, it is a multitude of interconnected factors that are affecting fertility preferences [19].

The concept of fertility preferences theoretically captures the extent to which human agency or intentional behavior affects the reproduction process. Fertility preference measures can broadly be defined as measures that seek to capture some dimension of an individual's attitude or motivation to influence fertility outcomes. Fertility preference data are routinely collected in demographic studies and are used for various purposes. Typical uses of information on fertility preference include estimating completed fertility for couples and extending cohort fertility in aggregate-level forecasts [20].

In developing countries, information on fertility preferences of men and women is usually used to:

- ❖ estimate the demand for fertility control,

- assess the level of unmet need for contraception
- estimate trends in ideal family size,
- assess the prevailing need for contraception,
- assess the extent of unwanted and mistimed pregnancies, and
- as an indicator of future fertility trends.

Research in fertility preferences, particularly in developing countries, has received considerable attention in recent years because of its relationship with and important bearing on the complex family building processes. The subject of parental attitudes and aspirations in relation to household fertility decision-making have gained importance in recent fertility researches since these seem to be related to the future course of fertility in a society. The thrust of most of the research has been on the desired or preferred family size as a substitute variable for eventual fertility or completed family size. If a respondent's stated fertility preferences are related in some way to her eventual fertility, then information on fertility preference should have a predictive value in forecasting the future course of fertility, and under voluntary control the desired number of children will be an increasingly important determinant to fertility [21].

1.6 Population Policy of Ethiopia

Population dynamics, including changes in population growth rates, age structures and distributions of people, are closely linked to national developmental challenges and their solutions. It is noted that much attention should be given to population dynamics in order to solve national and global development challenges. Stabilizing the growth of human population is a goal that must be achieved if nations are to preserve their options for the future and improve the factors that contribute to their sustainability. By the year 2050, world population is projected to be 9.3 billion. Practically all growth is estimated to occur in developing countries since fertility is high in these countries. This is especially observed in the poorest of those countries which already produce virtually all of the world’s human numbers. In these countries, rapid population growth is more likely to cause or exacerbate challenges such as climate change and global warming, fragile and failed states, migration and refugee crises, food and water insecurity, poverty, disease, debt, illiteracy and add pressure on the economy, basic health and social services and the environment if left unchecked [22].

Population policy is a deliberately constructed or modified institutional arrangements and/or specific programs through which governments influence, directly or indirectly, demographic change [23]. Ethiopia has formulated a population policy in 1993 to bring population growth rates in line with other policy targets and achieve socio economic advancement. Reducing the total fertility rate to approximately 4.0 children per woman and reducing maternal mortality rates, infant mortality and childhood morbidity and mortal by the year 2015 are among the specific objectives of the policy [24].

Ethiopia is one of the 189 countries committed to achieving the Millennium Development Goals (MDGs) by 2015. A lack of access to Sexual and Reproductive Health (SRH) information and services leading to high fertility and subsequent population growth, particularly in the poorest countries, continues to pose significant challenges to development and the attainment of the MDGs. And high levels of fertility and population growth make it far more difficult for families and societies to overcome poverty [25].

The Millennium Development Goals (MDGs) are a set of eight important principles and goals ranging from poverty reduction to universal primary education. The time-bound goals are as follows [24]:

1. Eradicate Extreme poverty and hunger.
2. Achieve universal primary education.
3. Promote gender equality and empower women.
4. Reduce child mortality.
5. Improve maternal health.
6. Combat HIV/AIDS, Tuberculosis, maternal and other diseases.
7. Ensure environmental sustainability.
8. Develop a global partnership for development.

Achieving these goals would be major challenge for Ethiopia due mainly to the effects of uncontrolled population growth; hence, the need for population policy.

1.7 Research Methodology

In this study, to develop a model that can predict fertility preference of women of reproductive age using data mining methods based on their demographic and socioeconomic characteristics a hybrid Knowledge Discovery Process (KDP) model was adopted. Hybrid Knowledge Discovery Process (KDP) model is research-oriented and originated from academic and industrial models combining aspects of both. According to Cios et al. [26], a typical hybrid KDP model comprises sex steps: business/domain understanding, data understanding, data preparation, data mining, evaluating the discovered knowledge and use of the discovered knowledge.

The data source for this study was EDHS2011 survey data acquired from CSA of Ethiopia. The 2011 Ethiopia Demographic and Health Survey (EDHS) was conducted by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health [27]. After identifying women's dataset which contains 16511 records, features related to fertility preference were selected with the help of domain expert. The dataset was prepared for mining purpose by cleaning, discretization, and reduction of the data. Three data mining methods, namely Decision Tree, NaiveBayes and Neural Network, algorithms were used to develop models and the performance of each was evaluated based classification accuracy, TP rate, TN rate, F-measure and Area Under the ROC curve.

Various books, journals articles and papers from the Internet have been assessed and the effects of high fertility, fertility preference, basic concepts of data mining and its applications in different fields have been discussed briefly.

Tools

Choice of appropriate tools that aid in achieving the set objectives in any research is an essential step in research design. Accordingly, tools that aid in accomplishing the data mining goal of this study were selected. The tools used are Weka 3.6.3 machine learning software, SPSS version 16.0 and Microsoft Excel.

Weka Machine Learning Software

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software, developed at the University of Waikato, New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It is platform independent, that is, it runs on the major operating systems like Linux, Windows, and Macintosh. It is also open source and free software [28].

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre-processing tools. It includes virtually all the popular algorithms. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning and a wide variety of learning algorithms. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand [28].

SPSS

SPSS (Statistical Package for the Social Sciences) is used for manipulating, analyzing, and presenting data that provides a powerful statistical-analysis and data-management system using descriptive menus and simple point-and-click interface. In this study, it was used for some exploratory data analysis and for exporting the dataset into Microsoft Excel.

1.8 Statement of the Problem

Fertility is one of the factors that determine the size, structure and/or distribution of population. It can be defined as the reproductive performance of an individual, a couple or a group of population. Fertility can be measured by crude birth rate or total fertility rate and it contributes to population growth rate. The latter is an indicator of fertility level of a country [29].

High fertility is an issue for developing countries like Ethiopia. It has micro-and-macro-level demographic consequences. micro-level effects include, among others, high incidence of births of order five and above, a high fraction of women experiencing pregnancies of order five and above, and short inter-pregnancy intervals; at the macro-level, rapid population

growth rate with the concomitant rapid growth in the size of successive birth cohorts is considered the main demographic consequence. High maternal, infant and child mortality rates and low weight at birth are related to the micro-level consequences of high fertility [6].

High fertility of a country leads to rapid population growth. The continuation of rapid population growth presents serious challenges to the future development of a country Ethiopia, though declining, is characterized by high fertility. With the assumption of moderate fertility decline, the population of Ethiopia is projected to be 130 million by 2030. As human numbers increase in Ethiopia, the population carrying capacity of the environment decreases. A high population growth rate induces increased demand for resources and the rate at which these resources are exploited. In Ethiopia, where technology has not kept pace with the demands for greater productivity, environmentally harmful and economically counterproductive methods of exploiting land and associated resources are common practices to meet immediate needs. Such practices result in harsh climatic conditions and declining soil quality at alarming rate [30].

Fertility preference defines the demand for children and also indicates the motivation to limit fertility (particularly within marriage) deliberately, the latter being a major precondition for fertility decline. Evidences from researches prove that responses to questions on fertility preference are expressive and meaningful for measuring reproductive choices, for assessing the motivation for fertility regulation, and for analyzing future prospects of fertility change. Furthermore, fertility preference measures have been used to describe and/or estimate the number of children that people actually want to have [31]. Pritchett [32] argued that women's fertility choices are the primary determinants of actual fertility, and he added that policies that improve objective conditions for women are the most important voluntary and sustainable way to achieve the reductions in fertility necessary to slow population growth.

Currently in Ethiopia, population growth has been recognized as one of the main challenges to poverty reduction. The effect of population growth is reflected in various conditions, like food insecurity, population pressure on the land, low incomes in rural areas, and youth unemployment in urban areas. Alleviating these and other related conditions demands harmonizing population growth with socio-economic development. Along this line, much attention is being given to demographic factors in formulating multisectoral development

strategies to end poverty and to reduce the total fertility rate, for instance, to 4 lifetime births by 2015 [33].

Identifying critical demographic and socioeconomic factors that determine women's fertility preference is an essential step in the process of fertility regulation. Women's survey dataset at Ethiopian Central Statistical Agency can be more effectively exploited by applying data mining techniques to develop a model that can predict fertility preferences of women of reproductive age so that appropriate intervention programs could be designed and implemented for the purpose of achieving reductions in fertility level necessary to slow population growth. Regarding this, as far as the knowledge of the researcher is concerned, no researches have been conducted locally on fertility preference using data mining techniques.

The purpose of this study, therefore, is to apply data mining classification methods to women's dataset for extracting hidden patterns from the data and developing a model that could predict fertility preferences of women of reproductive age. The dataset is one of the datasets collected by Central Statistical Agency (CSA) of Ethiopia under Ethiopian Demographic and Health Survey (EDHS).

1.9 Research Objectives

1.9.1 General Objective

The general objective of this study is to explore the possibility of applying data mining techniques in developing a model that can predict fertility preferences of women of reproductive age from EDHS dataset. The model could aid in designing and implementing appropriate intervention programs for the purpose of achieving reductions in fertility level necessary to slow population growth.

1.9.2 Specific Objectives

- Conduct a review of literature on some data mining concepts and methods in general, and the application of the methods in healthcare in particular.

- Select and preprocess the dataset required for the data mining problem from the database of EDHS acquired from CSA.

- Identify appropriate data mining algorithms and software appropriate for classification task, which is the core of the research.
- Prepare the data for model building which includes data encoding, accounting for missing values, and converting data formats.
- Build data mining models using the selected algorithms and select the best model based on model performance measures.
- Report results and make recommendations.

1.10 Scope and Limitation of the Study

Beckman [34] argued that, within marriage, men's influence on fertility decisions is so strong that it can not be ignored. Fertility preferences of their husbands or partners do matter and has a great influence on actual fertility outcome. Even so, this study covers the analysis of fertility preference of women only using data mining methods based on dataset acquired from CSA.

Studies on fertility in general and fertility preference in particular using data mining technology are almost non-existent. This could be considered as the major limitation of this study.

1.11 Significance of the Study

A typical use of information on fertility preference is for assessing future fertility trends. A comprehensive analysis of levels of fertility preference of couples assists the government as well as reproductive health programmers in designing appropriate and/or fortifying existing intervention programmes with the aim of regulating fertility so that slow population growth could be achieved. The outcome of this study will also add to the body of knowledge on fertility preferences. The output of this study may also be used as a complementary approach to statistical methods in analysing survey data and a launch pad for future studies of fertility preference using data mining technology.

1.12 Dissemination of Results

This thesis could be disseminated through one or all of the following ways:

- Giving away hardcopies for concerned organizations, specifically CSA. This way, the research could be used for supporting decision making or as a basis for further research in the area.
- Publishing in different journals.
- Presentation on different conferences/workshops.

1.13 Thesis Organization

This thesis consists of five chapters. In the first chapter, factors that determine fertility, trends and current level of fertility in Ethiopia, fertility preference and related concepts are briefly discussed. Problem statement, research objectives, research design, scope and significance of the study are also presented in this chapter.

Chapter two is review of literature. A few relevant documents are scanned to present a brief account of data mining or knowledge discovery from database, the KDD processes and the role of data mining in different application areas including healthcare. Some data mining methods/techniques and what tasks data mining can accomplish through the methods are also outlined here.

The third chapter deals with data pre-processing and a description of algorithms and performance measures employed for model building and evaluation. Of course, identified objectives for collecting the dataset, aligned with the objectives of data mining task for this study, along with the description of the dataset are presented in this chapter. These are business and data underrating steps blended here.
The fourth chapter is about the experimentation. In this chapter, different experiments ran using the selected algorithms and their corresponding interpretations together with the performance of the developed models are discussed.

In the final chapter, conclusions about general discussions, experimentation results, and performance of the data mining methods employed are drawn. Finally, recommendations were forwarded based on the findings.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Data Mining

Data mining is relatively a new field which lies at the interface of statistics, database technology, pattern recognition, machine learning, artificial intelligence, data visualization, and expert systems. It is usually not simple to set clear boundaries between each of these disciplines [35]. This interplay of data mining with theses disciplines is evidenced by the fact that data mining involves an integration of techniques from database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis [36].

Different authors define data mining differently. The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably, but in realty, data mining is the core step in the KDD process. Han et al. [36] defined data mining as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Witten et al.[28] defined data mining as the process of extracting implicit, previously unknown and potentially useful information from data. Hand et al. [35] defined data mining as the analysis of observational data set to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Fayyad et al. [37] defined KDD as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

According to Hand et al. [35], data mining is usually performed on data that have already been collected for some purpose other than the data mining analysis. That is, the objectives of the data mining exercise do not help in designing the data collection strategy. That is why the third definition refers to ‘observational data’ rather than ‘experimental data’. Hence, data mining is often referred to as secondary data analysis. The knowledge mined through a data mining exercise is often referred to as model or pattern. This knowledge can be represented

using linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

Different situations call for the need for data mining. For example, sophistication in data collection tools and database technologies has resulted in an avalanche of data. Huge data are collected and stored in different formats and various data repositories until it is beyond human ability for comprehension. In the absence of powerful data analysis tools, this has been described as data rich but information poor situation, where data stored in large data repositories become ‘data tombs’ or data archives without being exploited by converting the data into information. As a result, in such situation, decision makers are usually obliged to make important decisions based on their intuition, rather than based on the information-rich data stored in data repositories simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. Another situation where the need for powerful data analysis tools arises is for expert system technologies, which usually rely on users or domain experts to manually input knowledge into knowledge bases. Such procedure is prone to biases, errors, extremely time-consuming and costly. Data mining applications provide tools for performing data analysis to uncover hidden but important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [36].

Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. This enables organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. Data mining technology delivers two key business intelligence benefits: it enables enterprises, regardless of industry or size, in the context of defined business objectives, to automatically explore, visualize and understand their data, and to identify patterns, relationships and dependencies that influence business outcomes; it enables relationships uncovered and identified through the data mining process to be expressed as business rules, or predictive models. These outputs can be communicated in traditional reporting formats like presentations, briefs, electronic information sharing, to guide business planning and strategy. The outputs can also be expressed as programming code, can be deployed or integrated into business operating systems to generate predictions of future outcomes, based on newly generated data, with higher accuracy and certainty [38].

Larose [39] put it as the following regarding companies which apply (or do not) data mining to their data repositories:

> *Data mining empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies that do not apply these techniques are in danger of falling behind and losing market share, because their competitors are using data mining and are thereby gaining the competitive edge. Discovering Knowledge in Data, the step-by-step hands-on solutions of real-world business problems using widely available data mining techniques applied to real-world data sets will appeal to managers, and others who need to keep abreast of the latest methods for enhancing return on investment.*

Data mining technology can be applied in various application areas in different fields. Different industries have been exploiting the advantage of this relatively new technology. Some of these (with few specific areas of application) include banking; for predicting levels of bad loans and fraudulent credit card use, predicting credit card spending by new customers, predicting which kinds of customers will best respond to new loan offers, manufacturing and production; for predicting machinery failures, finding key factors that control optimization of manufacturing capacity, insurance; forecasting claim amounts and medical coverage costs, predicting which customers will buy new policies, Health care; correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes, learning how to provide proper treatments elements that affect medical coverage, airlines; capturing data on where customers are flying and the ultimate destination of passengers who change carriers in hub cities so that airlines can identify popular locations that they do not service, checking the feasibility of adding routes to capture lost business, marketing; classifying customer demographics that can be used to predict which customers will respond to a mailing or buy a particular product [38].

Data mining tools and techniques can be applied to, at least in principle, any kind of data repository and transient data (- data streams). The data for mining purpose could be obtained from relational databases, data warehouses, transactional databases, advanced database systems, flat files, data streams, and the World Wide Web. Advanced database systems include object-relational databases and specific application-oriented databases, like spatial databases, time-series databases, text databases, and multimedia databases [36].

This application of data mining to different data repositories makes it a fast-growing discipline. Moreover, the advancement of database technologies, data warehouses, availability and increased access to Web and intranet data, the development of off-the-shelf software suites and the huge growth in computing power and high capacity storage devices are among developments in information technology that played crucial role in the emergence of the field of data mining and knowledge discovery [39].

2.1.1 Data Mining Tasks

The tasks of data mining are very diverse and distinct based on the type of patterns to be mined from a database. Data mining applications employ different kinds of data mining methods and techniques to find different kinds of patterns and/or knowledge. Generally, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables/fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data mining tasks into one of two categories [40]:

- Predictive data mining tasks
  perform inference on the current data (pre-classified data set) in order to make predictions about unseen data, that is, produce the model of the system described by the given data set (training set) or,
- Descriptive data mining tasks
  characterize the general properties of the data in the database, that is, produce new, nontrivial information based on the available data set.

In prediction, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, and other similar tasks. On the other hand, in description, the gaol is to gain an understanding of the analysed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data mining g applications can vary considerably. The goals of prediction and description are achieved by using data mining methods/techniques, briefly described next, for the following data mining tasks [40]:

- Classification – discovery of a predictive learning function that classifies a data item into one several predefined classes.

- Regression – discovery of a predictive learning function which maps a data item to a real-valued prediction variable.
- Clustering – a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
- Summarization – an additional descriptive task that involves methods for finding a compact description for a set of data.
- Dependency modelling – finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
- Change and deviation detection– discovering the most significant changes in a data set.
- Frequent patterns/associations – discovering frequent patterns, patterns that occur frequently in data. It is also discovering togetherness of data objects. Such togetherness is termed as association rule.

2.1.2 Basic Data Mining Methods

Machine learning methods are commonly categorized as either supervised or unsupervised learning methods. Supervised approaches require both the input (predictors) variables and the output (response) variable, whereas, unsupervised approaches rely solely upon the input (explanatory) variables [39]. The following are a few of data mining methods with the corresponding tasks.

2.1.2.1 Classification

Classification is the most commonly applied data mining task which employs a set of pre-classified examples with target categorical variable (training set) to develop a model or function that can classify examples whose class labels are unknown (test set). The data classification process involves learning and classification steps. In the learning step, the training data set is analysed by classification algorithm. In classification step, test data set is used to estimate the accuracy of the model built in the learning step. If the accuracy is acceptable, the model can be applied to new data tuples [41]. For example, any of the following classification methods can be employed to classify a breast cancer as malign or benign.

Examples of classification methods:

- Classification by decision tree induction
- Bayesian classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification based on association

2.1.2.2 Clustering

Clustering is grouping of a set of objects (whose classes are unknown) into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters, that is, the objects are clustered so that the intra-class similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. The objects are assigned to their respective class and their common features in the cluster are summarized to form the class description. Clustering is a type unsupervised learning that can be employed for the purpose of identifying classes of objects in a data set prior to classification task so that cost of classification can be minimized. For instance, cluster analysis can be applied to categorize genes with similar functionality [41].

Examples of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

2.1.2.3 Prediction

Unlike classification, prediction is used to predict the value of a dependent continues variable, rather than a categorical label. Hence, regression and numeric prediction are synonymously used. Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable [41].

Examples of regression methods

- Linear Regression
- Multivariate Linear Regression

- Nonlinear Regression
- Multivariate Nonlinear Regression

2.1.2.4 Association

Association is usually used to find frequent item sets among large data sets. It is the process of finding attributes which 'go together', mostly in transactional data sets. Association is also known as affinity analysis or market basket analysis. Association rules are of the form 'If antecedent, then consequent', together with a measure of the support and confidence associated with the rule. For instance, association rule methods can be applied to determine proportion of cases in which a new drug will exhibit dangerous side effects [39].

Examples of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

2.1.3 Knowledge Discovery Process Models

Knowledge Discovery in Database, otherwise, Data Dining is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process, involving the use of algorithms that explore the data, develop the model and discover previously unknown patterns. The KDP process models define a sequence of steps, with eventual feedback loops, that should be followed to discover knowledge, for example, patterns in data. Such process models help organizations and data mining practitioners better understand the KDP and provide a roadmap to follow while planning and executing a data mining project. The model may be used for understanding phenomena from the data, analysis and prediction [42]. There are different KDD process models and two of them are briefly discussed below.

According to Fayyad et al, [37], the knowledge discovery process is interactive and iterative (moving back to previous steps may be required) at each step and consists six steps. The scene for the KDD process commences with developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint. The first step is creating a target dataset which involves selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be

performed. Data cleaning and pre-processing comes next and involves removing noise, if appropriate, collecting the necessary information to model or account for noise, and deciding on strategies for handling missing data fields. Once the data are pre-processed, the next step is data transformation where data are transformed or consolidated into forms appropriate for mining, for example, by performing summary or aggregation operations; this step leads to the data mining phase which involves searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. Interpreting the mined patterns, the fifth step, entails visualization of the extracted patterns and models or visualization of the data given the extracted models visualization of the extracted patterns and models or visualization of the data given the extracted models, and finally, knowledge representation involves using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge [37].

Fayyad et al. [37] also indicated that the KDD process implicitly entails two other important steps, namely matching the goals of the KDD process set at the outset to a particular data-mining method like summarization, classification, regression, clustering, and so on, and exploratory analysis and model selection which involves choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. Model selection can be achieved through deciding which models and parameters might be appropriate and matching a particular data-mining method with the overall criteria of the KDD process. The KDD process is depicted in Figure 2.1.

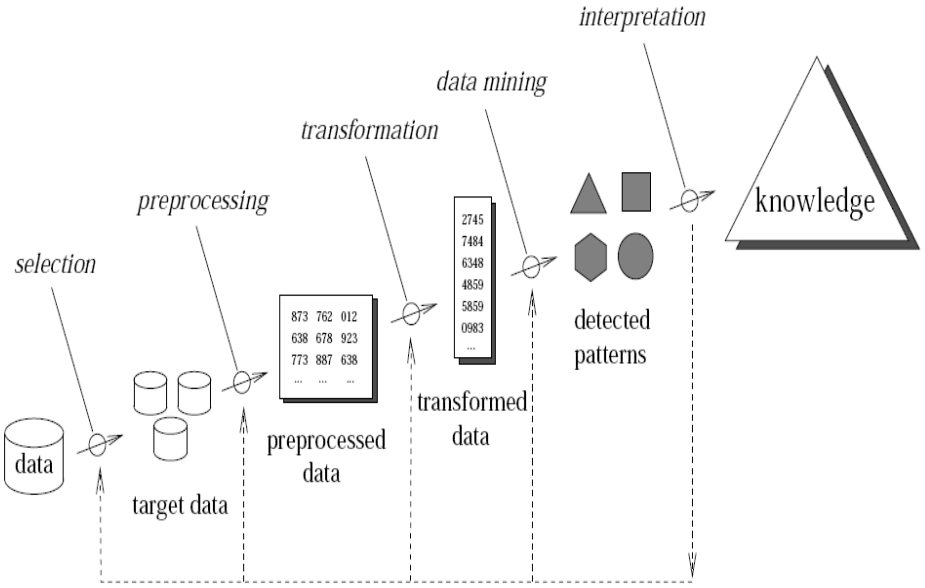


Figure 2.1 The KDD Process Model Steps

Another KDD process model is a hybrid model which is research-oriented and originated from academic and industrial models combining aspects of both. According to Cios et al. [26], a typical hybrid KDD Process model comprises sex steps. The first step is understanding the business or problem domain. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology, translating business project goals into DM goals, and the initial selection of DM tools to be used later in the process. The next step is understanding the available data. This step includes collecting sample data and deciding which data, including format and size, will be needed. Knowledge of domain experts is essential for data understanding. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals. In the third step, data should be prepared for mining. This step deals with deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms to reduce dimensionality, by derivation of new attributes, and by summarization of data. The end results are data that meet the specific input requirements for

the DM tools selected in Step one. \the forth step is data mining where data miner uses various DM methods to derive knowledge from pre-processed data. Evaluation of the discovered knowledge comes net and includes understanding the results, checking whether the discovered knowledge is novel and interesting, and interpretation of the results by domain experts. The final step is use of the discovered knowledge. This final step consists of planning where and how to use the discovered knowledge. The hybrid KDP model was adopted for this study.

2.2 Challenges of Data Mining in Health Care

Heterogeneity of Health Data

Health data usually is fragmented and distributed between hospitals, insurance companies and government departments or large portion of potentially relevant health information may not be stored electronically. Even when electronic data is available, it is usually scattered in several small databases through different clinics, hospitals, and laboratories. This data can also be in many different formats (e.g. text, image, and video) and is collected from various sources, such as patient records, doctor's comments, and laboratory test results. Medical data may also be collected from various images, interviews with the patient, and physician's notes and interpretations. All these data may have bearing on the diagnosis, prognosis, and treatment of the patient. Most medical procedures employ imaging as a preferred diagnostic tool, hence, the need to develop methods for efficient mining in databases of images, which are more difficult than mining in purely numerical/text databases [43].

Other challenges that impede extensive use of data mining in healthcare include the issue of data ownership, fear of lawsuits, and privacy concerns. Cios and Moore[44] explained these issues as below:

Data ownership: The question of who or which institution owns patient data is unsettled in most cases. Particularly, it is unclear whether the patients, the physicians, the laboratories, or the insurance companies own the data collected from patients. This situation usually leads to lawsuits in case of breach of health data.

Fear of lawsuits: In medical communities there is a fear of malpractice and other costly lawsuits that add to the challenges of applying data mining in healthcare. Potential lawsuits

that may be triggered by discovering anomalies in patient medical histories leave medical professionals unwilling to share patient data with researchers.

Privacy issues: this involves protecting patient privacy and doctor-patient confidentiality which adds another sets of challenges to data mining in health care. Administrators and researchers should pay strict attention to privacy and security when transferring, storing, or mining patient data. In many cases, patient records need to be anonymous, i.e., patient identities need to be removed at the time of information collection), anonymized, that is, patient identities are removed after the data is collected, or de-identified, i.e., patient identities are encrypted and can be restored under certain institutional policies.

The issue of **up-to-date clinical data** is still another challenge. Medical databases are updated constantly by adding new results from lab tests and other medical equipment for patients. Subsequently, this needs data mining techniques that can incrementally update the discovered knowledge, but such techniques currently are under research. Moreover, missing/incomplete data exacerbate the problem because clinical databases often lack some data required for analysis or discovery. Some data elements are not collected due to omission, irrelevance, excess risk or inapplicability in a specific clinical context. This becomes a constraint to the data mining process, when, for example, a complete set of data elements is required for some learning methods. Besides, even though some data mining methods can handle missing values, the data that was not collected may have independent information value and should not be ignored [45**]**.

2.3 The Need for Healthcare Data Warehouses

Inmon [46] defined data warehouse as subject-oriented, integrated, time-variant and non-volatile data in support of management decisions.

Subject-oriented means hat all relevant data about a subject is gathered and stored as a single set in a useful format. Information is presented according to specific subjects or areas of interest like customer, vendor, product, and sales while the integrative characteristic of a data warehouse refers to data being integrated from multiple heterogeneous sources such as relational databases, flat file and online transaction records and stored in a globally acceptable format with consistent naming conventions, measurements, encoding structures, and physical attributes even when the underlying operational systems store the data

differently. A data warehouse is time-variant in a sense that it stores historical data and covers a much longer time horizon than any other data repository (several years to decades); the time element is included implicitly or explicitly in every key structure in a data warehouse and a non-volatile characteristic of a data warehouse implies that data warehouse contains read-only data, which are updated in planned periodic cycles, not frequently, so once the data is stored in a data warehouse it is not easily changed each time an operational process is executed. Information is consistent regardless of when the warehouse is accessed [46].

Data warehouse is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. Data mining potential can be enhanced if the appropriate data are collected and stored in a data warehouse. A data warehouse is a relational database management system (RDBMS) designed specifically to meet the needs of transaction processing systems. It can be loosely defined as any centralized data repository which can be queried for business benefit. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats, as well as integrating data throughout an enterprise. Regardless of location, format, or communication requirements, data warehousing makes it is possible to incorporate additional or expert information [47].

Most applications of data mining in clinical and administrative decision support systems require homogeneous and centralized data repositories. These data repositories can be created by integrating medical data from various sources, presenting in common formats and storing separately from operational databases. This can be achieved through data warehouse technologies. Centralized data repositories enable effective application of data mining methods and techniques to healthcare data even though it may be initially costly to build and maintain such data warehouses. Data warehouse in clinical care systems is referred to as Clinical data Warehouse (CDW). It is a place where healthcare providers can gain access to clinical data gathered in patient care process. A Clinical Data Warehouse (CDW) can facilitate efficient storage, enhance timely analysis and increases the quality of real time decision making processes. Clinical data warehouses in combination with data mining can help healthcare systems in providing improved clinical care, better administration of health, and in undertaking enhanced quality medical research [48**].**

2.4 Application of Data Mining in Health Care

Data mining can be applied in different application areas in healthcare. The major application areas include the evaluation of treatment effectiveness, management of healthcare, customer relationship management, detection of fraud and abuse, early detection and/or prevention of diseases, Policy-making in public health, and early detection and management of pandemic diseases and public health policy formulation.

Treatment effectiveness: Data mining models can aid in evaluating the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For instance, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Successful standardized treatments for specific diseases can also be identified using data mining. Other data mining applications related to treatments include associating the various side-effects of treatment, identifying common symptoms to aid diagnosis, determining the most effective drug compounds for treating patients who respond differently from other patients to certain drugs, and determining proactive steps that can reduce the risk of drug side-effects [49].

Healthcare management*:* Data mining applications can be employed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. The goal here is to effectively manage resource allocation by identifying high risk areas and predicting the need for and usage of various resources. For example, a key problem in healthcare is measuring the flow of patients through hospitals and other healthcare facilities. If the inpatient length of stay can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced. Data mining also can be used to identify and understand high-cost patients [49].

Customer relationship management: data mining applications can be developed in the healthcare industry to determine the preferences, usage patterns, and current and future needs of individuals to improve their level of satisfaction. These applications also can be used to

predict other products that a healthcare customer is likely to purchase, whether a patient is likely to comply with prescribed treatment or whether preventive care is likely to produce a significant reduction in future utilization. Similarly, pharmaceutical companies can also benefit from data mining applications. By tracking which physicians prescribe which drugs and for what purposes, pharmaceutical companies can decide whom to target, identify what is the least expensive or most effective treatment plan for an ailment, identify physicians whose practices are suited to specific clinical trials, and map the course of an epidemic to support pharmaceutical salespersons, physicians, and patients [50].

Fraud and Anomaly Detection: Data mining shows promise in aiding in prevention of health care-fraud. Data mining applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others. For instance, these applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. Another key area where data mining based fraud detection is useful is the detection and prediction of faults in medical devices [51].

2.4.1 Some Case Studies

In this section, some literature on the application of data mining methods to disease prognosis and diagnosis are reviewed and the performance of the different data mining algorithms used is indicated.

Medical prognosis is a field in medicine that deals with the science of estimating the complication and recurrence of disease to predict the survival of patient or group of patients, that is, medical prognosis involves prediction modelling to estimate different parameters related to patient health. Prognosis is important because the type and intensity of the medications are based on it. Thus, these estimates can help design treatments as per the outcomes of diagnoses. Survival analysis a field in medical prognosis that deals with the application of various methods to estimate the survival of a particular patient suffering from a disease over a particular time period. For instance, in cancer diagnosis, there are three predictive foci of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability [52].

Delen et al. [53] used SEERS (Surveillance Epidemiology and End Results) dataset, with 17 variables and 202932 records, to predict breast cancer survivability. They used three data mining algorithms, namely, Artificial Neural Network (ANN), logistic regression, and decision tree to develop different models. In their experiment, with 10-fold cross validation on the test set, the classification accuracy of ANN was found to be 91.2% with a sensitivity of 94.5% and a specificity of 87.5%, and that of logistic regression was 98.2% with sensitivity of 90.2% and a specificity of 87.9%. C5 was found to be the best model with a classification accuracy of 93.6%, sensitivity of 96% and specificity of 90.7%.

Breast cancer diagnosis is another medical application that poses a great challenge to health professionals. Breast cancer is known to be the leading cause of death of women in the world, particularly in developed countries. Early detection is the most effective way to reduce breast cancer deaths. But **e**arly detection requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The goal of breast cancer diagnostic prediction is to assign patients to either a benign group that is noncancerous or a malignant group that is cancerous [52].

Sudhir et al. [54] applied Artificial Neural Network (ANN) to World Breast Cancer dataset, with 11 variables and 683 instances, to classify breast cancer patients into benign (non-cancerous) and malign (cancerous) groups. With twenty experiments, they came up with a 97% classification accuracy of Support Vector Machine (SVM) as compared to manual (physician's) detection of breast cancer which is 85%. This model with this high accuracy rate can aid physician's decision in avoiding biopsy. They recommended the training of the model with more instances, and that SVM can be utilized for the diagnosis of other types of diseases.

Cardiovascular diseases are among the deadly diseases in the world. Coronary Artery Disease (CAD) is the most common fatal heart disease that needs early diagnosis and detection based on its prevalence. The most reliable method for CAD diagnosis is angiography, but it is costly, time-consuming, and hazardous. Other popular non-invasive methods which have been proved to be effective include analysis and mining of patients' medical information [55].

Habibi et al. [55] used SMO, Naïve Bayes, and ensemble methods in Weka to develop a model that predicts CAD from a dataset with 38 features. The classification accuracies of the methods, respectively, are 86.95%, 87.22% and 88.52%. That is, the highest accuracy achieved (88.52 %.) is after feature selection (16 features were used) by the ensemble method. This proved the claim that feature selection and ensemble methods improve classification/prediction accuracy. The researchers also used association rule mining techniques that generated rules whit high confidences.

They claimed that they have achieved better classification accuracy than the previous studies on predicting CAD. They suggested the addition of new features from other sources like lab or echo data to achieve even better accuracy in predicting CAD.

Asha et al. [56] applied association rule mining method, Apriori, to tuberculosis patients dataset and generated interesting rules with high support and confidences, which aid physicians in identifying hidden symptoms associated with one another in diagnosing tuberculosis.

In summary, these studies indicate that data mining technology has high potential in clinical health care. Particularly, data mining classification methods are very effective in the early diagnosis and detection of chronic diseases. Prognosis and diagnosis of theses diseases effectively aids physicians in decision making, for instance, in decision to avoid surgical biopsy which are very dangerous for the patient.

2.5 Related Work

Gams et al. [57] used decision tree algorithms (J48) for demographic analysis of 147 countries described by 95 basic attributes to see their effect on fertility rate. Most of the attributes for the study were obtained from the Internet with the assumption that they might have impact on fertility rate. They grouped the attributes into biological, economical, social, cultural, anthropological and psychological factors. The data mining goal for this study was to classify countries according to high (TFR>2) and low (TFR<=2), and low (< 2), middle (2-3) and high (>3) total fertility rates.

In the study, several models were developed based on all and selected attributes for all countries, developed, and developing countries. The researchers also used social, economical

and educational attributes separately to classify counties according to high-and-low and high, middle and low total fertility rates. With this, they achieved different classification accuracy for each category. For instance, the classification accuracy of J48, with all features, in classifying countries according to high and low total fertility rate was 80.3%. The researchers claimed that this result is consistent with practically all literature in the demographic field which attribute, for instance, death of newborns to social and economical status of mothers who need to have several children to compensate for those dead, fewer children and lower newborn mortality to higher education of mothers, and low percentage of stillborns to costs of child life-support.

The researchers concluded that they have discovered major features that determine fertility and suggested that, since they are not experts in fertility studies or demography, verification of their conclusions by an expert and further analyses of interesting new patterns on fertility as matter of further research. They also recommended analysis of fertility related to individuals using data mining methods; their analysis is related to various counties.

Nayab [69] used multiple regression analysis to assess factors that affect fertility preference of women of reproductive age in Punjab province (Pakistan). The researcher applied purposive proportional stratified sampling technique to select 246 currently married women in the age group of 15-49 years. In the study, participant observation and structured open-ended questionnaire were used for collecting data on sex of child, number of living children, age, education, number of living sons, number of living daughters, **w**omen's work participation, economic status of household, urban exposure, and inter-spousal communication.

According to the study, son preference was found to be a major reason for women to continue child bearing, that is, only when the desired number of sons, rather than the overall family size, was achieved did women consider stopping child bearing. The researcher also claimed that lack of communication between spouses led women to desire for more children, attributing this to women who were not sure about the fertility preferences of their husbands could not use any fertility regulating method without their consent. Economic back ground of the respondents did not seem to affect the fertility preferences and behaviour of the respondents. On the other hand, the educational attainment of women, increased age at marriage, and their urban exposure were found to have a lowering effect on both their fertility

preferences and fertility behaviour, as compared to those who had never been to school, were married younger, and were lifetime rural residents, respectively.

Ibisomi [70] applied multinomial logistic regression model to analyse trends in fertility preferences of currently married women in Nigeria. The researcher used three years (1990, 1999 and 20030) Nigerian Demographic and Health Survey dataset. The explanatory variables used in the study were age, residence, region and number of living children and the response variable was fertility preference. The study indicated variations in percentage of women who want no more children and of those declared infecund across regions. Percentage of respondents who want another child generally declines as number of surviving children increases in all the years, and the situation was the same for those who want another child after two years except when number of surviving children is less than two. The result of the study also showed that women tend to limit childbearing as number of surviving children increases. Urban-rural analysis showed that respondents who want to have another child within two years and those declared infecund were more in the rural while there were more respondents who want no more children in the urban area.

Demographic researches indicate that demographic and socio-economic variables influence fertility preference of men and women. In the above studies, the authors analyzed factors that affect total fertility rate of countries worldwide (using data mining technique) and statistical methods were also employed for the analysis of fertility preferences. In this study, the objective was to analyze fertility preference of women of reproductive age at a country level using data mining methods.

CHAPTER THREE

DATA PREPROCESSING AND ALGORITHMS USED

In this study, data mining classification methods were employed for the analysis of fertility preference of women of reproductive age. The objective was to develop prediction models using Decision Tree, Neural Network and Bayesian Classifiers. Understanding the domain/business area where data mining is to be applied and the data to be used for mining purpose are the necessary steps in data mining procedure to determine the objective of data mining from the business perspective. Data pre-processing is also an essential step in the process for preparing dataset that is appropriate for mining. The objective was to develop prediction models using Decision Tree, Neural Network and Bayesian Classifiers. In this section, domain and data understanding, data pre-processing and the algorithms used to build the models together with matrices used for performance measures and comparison are discussed in brief.

3.1 Data Source

The data source for this study is EDHS2011 survey data acquired from CSA of Ethiopia. The 2011 Ethiopia Demographic and Health Survey (EDHS) was conducted by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. This is the third Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. The survey interviewed a nationally representative population in about 18,500 households. Out of these households, a nationally representative sample of 16,515 women of age 15–49 and 14,110 men of age 15–59 were interviewed. This represents a response rate of 95% for women. The sample design for the 2011 EDHS provides estimates at the national (total, urban, and rural) and regional levels. The data were collected on key indicators relating to family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment, and knowledge of HIV/AIDS. The primary objectives of the 2011 EDHS are to provide up-to-date information for planning, policy formulation, monitoring, and evaluation of population

and health programmes in the country. The survey excluded institutional living arrangements (e.g., army barracks, hospitals, police camps, and boarding schools) [27].

The sample for the 2011 EDHS was designed to provide population and health indicators at the national and regional levels. The sample was designed in a way that allowed for specific indicators, like contraceptive use, to be calculated for each of Ethiopia's eleven geographic/administrative regions; namely, nine regional states (Tigray, Affar, Amhara, Oromia, Somali, Benishangul-Gumuz, SNNP, Gambela and Harari) and two city administrations (Addis Ababa and Dire Dawa)[27].

All women of age 15-49 who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were eligible to be interviewed. Standard questioners were used for data collection. These questionnaires were adapted from model survey instruments developed for the MEASURE DHS project and the UNICEF Multiple Indicator Cluster Survey (MICS) to reflect the population and health issues relevant to Ethiopia. The dataset used for this study contains 16515 records of all women of reproductive age (women who are 15-49 years old) [27].

3.1.1 Data on Fertility Preferences

Information on fertility preferences is utilized for different purposes. A typical use of such information is that it aids in understanding the potential demand for fertility control in a given population. The goal of generating information on fertility preferences is to assess the potential demand for family planning services for the purposes of spacing or limiting future childbearing. To achieve this goal, the objective of collecting data on fertility preferences is to know the proportion of women of reproductive age who want more children or who want to limit childbearing, among others. In EDHS2011 survey, to elicit information on fertility preferences, several questions were asked of women (pregnant or not) on whether they want to have another child or not.

3.1.2 Defining Data Mining Objectives

The goal of data mining in this study is to state project objectives translated to data mining objective. Thus, based on the objectives of collecting data on fertility preferences of women of reproductive age, the data mining objectives are set as follows:

- Given women's demographic variables, classify them into those who want more children and those who want no more children. These are the main fertility preference indicators used in this study.
- Identify and select attributes that are assumed to be determinants of fertility preference of a woman.
- Identify patterns from the dataset.

The first objective is indicative of the data mining problem to address. Hence, Decision tree, Naïve Bayes and Neural Network were selected based on this objective.

3.2 Selecting the Target Dataset

EDHS2011 survey data at CSA consists of eight datasets. These datasets were not self-explanatory for the researcher and hence, an expert advice was sought in order to identify women's dataset. Thus, with the help of domain expert at CSA, the dataset was identified and from this dataset, fifteen attributes assumed to be determinants of fertility preference of women were selected. This dataset contains 16515 records, each representing individual women.

3.3 Description of Attributes

The selected women's data set consists of 16515 records/cases. Each record represents an individual woman for which data are collected on background characteristics (like age, education, media exposure), education, employment history, marriage, fertility preference, and others. The dataset is in SPSS format and the variables are not explicitly presented, but rather simply coded using a certain format. For instance, V013 represents age of women in the SPSS file. The variables are recoded using the variable description document accompanying the data set. Out of several variables, fifteen of them which are related to fertility preference were selected. The variables provide socio-demographic information for each woman.

The selected variables are age, region, type of place of residence, educational status, religion, wealth index, number of living children, number of son died, number of daughters died, frequency of watching television, marriage status, knowledge of methods, whether a woman is working, sex of child and fertility preference. These attributes together with their description are presented in Table 3.1.

Table 3.1 Selected attributes with descriptions

No.	Attribute	Description	Type
1	Age	Current age of woman	numeric
2	Region	De facto region of residence	nominal
3	Residence	De facto type of place of residence	nominal
4	Education	Highest education level attended	nominal
5	Religion	Religion to which a woman is devouted	nominal
6	Number of living children	Total number of living children a woman has	numeric
7	Number of sons died	Total number of sons who have died	numeric
8	Number of daughter died	Total number of daughters who have died.	numeric
9	Marriage status	Whether the respondent has ever been married	nominal
10	Knows method	Knowledge of contraceptive method	nominal
11	Wealth index	Wealth status of a woman	nominal
12	Respondent currently working	Whether or not a respondent is employed	Nominal
13	Frequency of watching television	Media exposure	Nominal
14	Sex	Sex of child	Nominal
15	Fertility preference	Whether a woman wants more children or wants no more children	Nominal

3.4 Some Exploratory Data Descriptions

Some of the primary reasons for performing exploratory data analysis is to investigate the variables, look at histograms of the numeric variables, examine the distributions of the categorical variables, and explore the relationships among sets of variables.

The objective of exploratory data description here is to look at the distribution of response variable and its relationships with some predictor variables and to explain these relationships. These relationships measure the strength of dependency between attributes.

After selecting attributes (by consulting domain expert) assumed to be relevant to the target variable, the final dataset, which consists of 15 variables (14 predictor(independent) variables and 1 dependent variable) and 16515 records, was constructed. Table 3.2 shows summary of predictor variables.

Table 3.2 Predictor variables with number of distinct values

Categorical variables	Number of unique values
Age	7
Region	11
Residence	2
Education	4
Religion	6
Wealth Index	5
Frequency of watching television	3
Marriage status	2
Knows method	2
Working	2
Sex of child	2

Continuous variables		Mean	S.D.	Range
Number of living children	13	2.309	2.44	0-12
Number of sons died	7	2.49	0.638	0-6
Number of daughters died	8	0.2	0.568	0-9

The dependent variable is a binary categorical variable with two categories: 0 and 1, 0 denoting 'want no more children' and 1 denoting 'want more children'. The distribution of the dependent variable is shown in Table 3.3.

Table 3.3 Distribution of dependent variable

Category	Count	Percentage
1 (Want more children)	10971	66.0
0 (Want no more children)	5544	34.0
Total	16515	100.0

The above table shows the proportion of women who want to limit childbearing and those who demand more children. As can be seen from the tables, 5544 (34%) of the woman want no more children and 10971 (66%) of them want more children, indicating unbalanced proportion of the classes of the target variable.

Cross-tabulations of variables quantifies the relationship between the variables. Table 3.4 shows the cross-tabulation of a woman's age and the dependent variable quantifying the relationship between a woman's age and her fertility choice.

Table 3.4 Cross-tabulation of Age with the dependent variable

Count				
		Class		
		0	1	Total
Age	15-19	553	3282	3835
	20-24	416	2606	3022
	25-29	840	2345	3185
	30-34	859	1241	2100
	35-39	1022	936	1958
	40-44	923	391	1314
	45-49	931	170	1101
Total		5544	10971	16515

The counts in the 0 column add up to the total number of woman who want to limit childbearing (5544) and the 1 column add up to the total number of woman who tend to have more children (10971). From Table 3.3, it is clear that older women tend to limit childbearing while younger women tend to have more children.

3.5 Attribute-Relation File Format (ARFF)

The dataset was exported to Microsoft Excel from SPSS. Then, this file in Excel format was converted to CSV (Comma Separated Values) file format which is plain text file format. This is used to create Attribute Relation File Format (ARFF) which is Weka's standard file format.

Attribute Relation File Format (ARFF) file is an ASCII text file that describes a list of instances sharing a set of attributes. It is WEKA's standard method of representing the instances and attributes found in data sets. ARFF files have two distinct sections. The first section is the header information, which is followed by the data information. The keyword *relation* indicates the name for the file, which is followed by a block defining each attribute in the data set (the columns in the data) and their types The header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types [28].

The ARFF header section of the file contains the relation declaration and attributes declarations. The relation declaration begins with '@' and the 'relation' statement and the attribute declaration begins with '@' and the 'attribute' statement. Sample of the ARFF format of the dataset is presented in Table 3.5.

Table 3.5 ARFF file format of the dataset

@relation	Fertilitypreference.arff
@attribute	Age {15-19,20-24,25-29,30-34,35-39,40-44,45-49}
@attribute	Region {1,2,3,4,5,6,7,8,9,10,11}
@attribute	Residence {1,2}
@attribute	Education {0,1,2,3}
@attribute	Religion {1,2,3,4,5,6}
@attribute	WealthIndex {1,2,3,4,5}
@dataa	
30-34,3,2,0,1,3,1,0,0,3,?,4,1,1,3,1,1	
15-19,3,2,1,1,1,?,0,0,0,0,?,0,1,1,?,1	
40-44,3,2,0,1,1,2,4,0,7,?,1,3,0,1,0,0	
25-29,3,2,1,1,1,2,0,0,2,0,?,0,0,1,?,0	
20-24,3,2,0,1,1,2,0,0,2,?,1,0,1,3,0,1	
30-34,3,2,0,1,3,1,1,0,6,?,4,3,1,3,1,0	
…	

3.6 Data Cleaning

Real-world data are usually incomplete, noisy, and inconsistent due to various reasons. Data cleaning involves handling missing values, smooth out noisy data, identifying outliers, and correcting inconsistencies and redundancy in the data.

3.6.1. Handling Missing Values

The presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset as a training sample. Rates of less than 1% missing data are generally considered trivial and 1-5% manageable. However, 5-15% requires sophisticated methods to handle and more than 15% may severely impact any kind of interpretation. Several methods have been suggested in the literature for treating missing data. Some of such methods include filling in the missing value manually (may not be feasible and time-consuming), replacing all missing attribute values by the same constant, using the most probable value to fill in the missing values and replacing a missing value with the attribute's mean or mode value (for numeric or nominal attributes, respectively). This last approach is the most commonly used method of handling missing values in a dataset [58].

In EDHS surveys data, missing values are due to partial or incomplete reporting of information and inconsistent responses to different questions in the survey. Missing values in the dataset for this research are left blank, and these are replaced with '?' using Microsoft Excel. All attributes with missing values among the selected attributes are nominal attributes. Using Weka's tool for replacing missing values, attribute's modal value is used for replacing the missing values for the nominal attributes shown in Table 3.6.

Table 3.6 Attributes with missing values

No.	Attribute	Count of missing values
1	Frequency of watching television	15
2	Respondent working	20
3	Knows method	7
4	Sex	5619
5	Marriage status	10204
6	Religion	8

3.7 Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node --nonleaf node-- denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node -or terminal node holds a class label. The topmost node in a tree is the root node [36]. Decision tree is one of the most widely used and practical methods for inductive inference. Decision Tree learning is a method of approximating discrete-valued target functions in which a learned function is represented by a decision tree. Learned trees can also be re-represented as sets of 'If-then' rules to improve human readability. Decision tree learning methods are among the most popular inductive inference algorithms and have been successfully applied to various tasks ranging from learning to diagnose medical cases to learning to assess credit risk of loan applications [59].

Decision trees classify instances by sorting them down the tree from the root to some leaf node which classifies a given instance at that node. At each node, a test of some attribute of an instance is specified and each branch descending from that node corresponds to one of the possible values for the attribute. An instance is classified by starting at the root node of the tree testing the attribute specified at this node, then moving down the tree branch corresponding to the value of the attribute at the next node. This process is recursive until tree leaf is reached. That is, Decision Trees are built from nodes, branches and leaves that indicate the variables, conditions, and outcomes, respectively. The most predictive variable is placed at the top node of the tree. The operation of decision trees is based on the ID3 or C4.5 algorithms [60].

In general, decision trees represent a disjunction of conjunction of constraints on the attribute values of instances, that is, each path from the root to a leaf indicates a conjunction of attribute tests and the built tree is a collection of these disjunctions [60]. A decision tree structure looks like the following (Figure 3.1) where A, B and C are attributes and C1, C2, C3 and C4 are classes of outcomes of tests on attributes at each node.

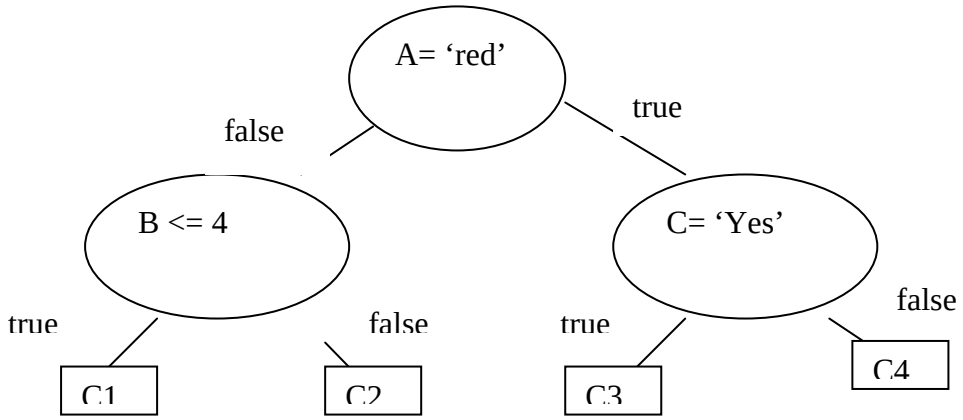


Figure 3.1 A Decision Tree structure

3.7.1 J48 Algorithm

.Decision tree J48 is an implementation of C4.5 algorithm for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's ID3 (Iterative Dichotomiser) algorithm. J48 builds decision trees from a set of labelled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. It examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, and training data with missing attribute values. Furthermore, it provides an option for pruning trees after creation [61].

J48 uses a measure known as information gain based on the concept of entropy – a measure commonly used in information theory– to select among the candidate attributes at each step while growing the tree. The goal is to select an attribute that is most useful for classifying instances at a given node. Entropy measures the impurity of an arbitrary collection of instances. That is, it is a measure of the homogeneity of the instances. Given a collection S containing k possible values with probabilities p1, p2, …, pk, the entropy of S is: [62].

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

Where P_i is the proportion of S belonging to class i.

Given entropy as a measure of the impurity of a collection of instances, a measure of the effectiveness of an attribute in classifying the instances can be defined. This measure is known as information gain. It measures the expected reduction in entropy caused by partitioning the instances according to the attribute. Thus, the information gain, Gain(S, A) of an attribute A relative to a collection of instances S is defined as:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where Values(A) is the set of all possible values for attribute A, and $\mathbf{S_v}$ is the subset of S for which attribute A has value v [62].

3.7.1.1 Avoiding Overfitting the Data

When building a decision tree, one may encounter many of the branches of the tree reflecting anomalies in the training data due to noise or outliers. That is, the algorithm produces trees that over fit the training instances. In overfitting the data, the accuracy of a model on training instances increases as the tree is grown, but its accuracy over independent test instances first increases and then begin to decline after some number of nodes is reached[36, 62]

Tree pruning methods, which typically use statistical measures to remove the least reliable branches, address this problem of overfitting. There are mainly two approaches to tree pruning in decision tree learning: prepruning and post pruning. In prepruning approach, a tree is pruned by halting its construction early, for instance, by halting a further splitting or partitioning of the subset of training tuples at a given node, while it is being grown so that the node becomes a leaf. The other approach is postpruning and is more commonly used due to the difficulty in the first approach to estimate precisely when to stop growing the tree. In

postpruning, subtrees are removed from a fully grown tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf with the most frequent class among the subtree being replaced [62].

3.8 Bayesian Classifier

Bayesian classifiers are statistical classifiers that can predict class membership probabilities, for instance, the probability that a given tuple belongs to a particular class.

Bayesian classification is based on Bayes' theorem. Bayes' theorem is named after Thomas Bayes, who did early work in probability and decision theory during the 18th century. If X is a data tuple, according to Bayes, X is considered to be "evidence", and it is described by measurements made on a set of n attributes. Let H be some hypothesis, such that the data tuple X belongs to a specified class C. For classification problems, the objective is to determine P(H|X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X. In other words, it is to determine the probability that tuple X belongs to class C, provided that the attribute description of X is known. P(H|X) is the posterior probability, or a posteriori probability, of H conditioned on X [36].
According to Bayes' theorem, the probability P(H|X) can be expressed in terms of probabilities P(H), P(X|H), and P(X) as:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)},$$

and these probabilities may be estimated from the given data[36].

3.8.1 Naive Bayesian Classifier

Naive 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 computation involved and, in this sense is considered as 'naïve'. Studies comparing classification algorithms have found that 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 [36].

The naive Bayesian classifier works as follows:

Let T be a training set of samples, each with their class labels. There are k classes, C1, C2, . . . , Ck. Each sample is represented by an n-dimensional vector, X = {x1, x2, . . . , x_n}, depicting n measured values of the n attributes, A1, A2, . . . ,An, respectively. Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class Ci if and only if

$$P(Ci|X) > P(Cj|X) \text{ for } 1 \leq j \leq m, j \neq i$$

Thus, we find the class that maximizes P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem,

$$P(Ci|X) = \frac{P(X|Ci)\,P(Ci)}{P(X)}$$

As P(X) is the same for all classes, only P(X|Ci)P(Ci) need to be maximized. If the class a priori probabilities, P(Ci), are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = . . . = P(Ck), and we would therefore maximize P(X|Ci), otherwise we maximize P(X|Ci)P(Ci). Note that the class a priori probabilities may be estimated by P (Ci) = freq (Ci, T)/|T|.

Given data sets with many attributes, it would be computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci) P(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 sample. Mathematically this means that

$$P(\boldsymbol{X}\text{j}Ci) = \prod_{k=1}^{n} P(xk\text{j}Ci$$

The probabilities P(x1|Ci), P(x2|Ci), . . . , P(xn|Ci) can easily be estimated from the training set. Recall that here xk refers to the value of attribute Ak for sample X performance measures. If Ak is categorical, then P(xk|Ci) is the number of samples of class Ci in T having the value xk for attribute Ak, divided by freq(Ci, T), the number of sample of class Ci in T. If Ak is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean μ and standard deviation s, defined by:

$$g(x, \mu, \sigma) = 1/ (2\pi\sigma)^{1/2}\, e^{-((x-\mu)2/2\sigma2}\,,$$

$$\text{so that} \qquad P(xk\text{j}Ci) = g(xk, \mu Ci\,, \sigma Ci\,).$$

μ_{Ci} and σ_{Ci} , which are the mean and standard deviation of values of attribute Ak , should be calculated for training samples of class Ci.

And finally, 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 X is Ci if and only if it is the class that maximizes P(X|Ci)P(Ci) [36].

3.9 Artificial Neural Networks

Artificial neural networks simulate human thinking and learn from examples. ANNs consist of nodes called neurons and weighted links between the neurons. Each neuron processes incoming information and may propagate information forward if warranted by its activation function. Artificial Neural Networks are massively interconnected networks in parallel of simple elements (usually adaptive), with heuristic organization, which try to interact with the objects of the real world in the same way that the biological nervous system does [63].

Artificial Neural Networks are used in many important engineering and scientific applications like signal enhancement, noise cancellation, pattern classification, prediction. Besides, they are used in many commercial products, such as modems, image processing and recognition systems, speech recognition and bio-medical instrumentation [64].

Neural networks offer a means for efficiently modelling large and complex problems in which there may be hundreds of predictor variables that have many interactions. Neural nets can be used in classification problems where the output is a categorical variable or for regressions in which the output variable is continuous [65].

A neural network is made of three layers: input layer, hidden and output layers and each layer consists of nodes. In input layer each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer or to an output layer. The output layer consists of one or more response variables. After the input layer, each node takes in a set of inputs, multiplies them by a connection weight Wxy (for example, in figure 3.2, the weight from node 1 to 3 is W13), adds them together, applies an activation or squashing function to

them, and passes the output to the node(s) in the next layer. For example, the value passed from node 4 to node 6 is given by [65]:

Activation function applied to ([W_{14} * value of node 1] + [W_{24} * value of node 2])

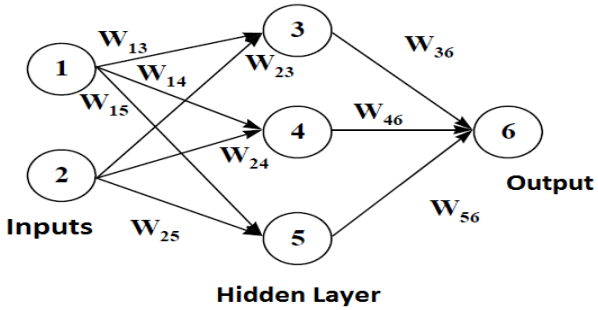


Figure 3.2 A Neural Network Architecture

Each node in neural nets can be viewed as a predictor variable (nodes 1 and 2 above) or as a combination of predictor variables (nodes 3 through 6). Node 6 is a non-linear combination of the values of nodes 1 and 2, because of the activation function on the summed values at the hidden nodes. Neural nets with a linear activation function but no hidden layer are equivalent to a linear regression and those with certain non-linear activation functions are equivalent to logistic regression. The architecture (or topology) of a neural network is the number of nodes and hidden layers and how they are connected. In designing a neural network, either the user or the software must choose the number of hidden nodes and hidden layers, the activation function, and limits on the weights. While there are some general guidelines, the user should experiment with these parameters. The feed-forward backpropagation network is among the most common types of neural networks [65].

3.9.1 Multilayer Perceptron

A Multilayer Perceptron is the most known and most frequently used type of neural network. On most occasions, the signals are transmitted within the network in one direction: from input to output. That is, there is no loop; the output of each neuron does not affect the neuron itself. This architecture is called feed forward neural network. In the MLP structure, the neurons are grouped into layers. The first and last layers are called input and output layers respectively because they represent inputs and outputs of the overall network. The remaining layers are known as hidden layers. Typically, a multilayer Perceptron consists of a set of sensory units or source nodes that constitute the input layer, one or more hidden layers of computational nodes and an output layer of computational nodes. The input signal propagates through the network in a forward direction on a layer-by-layer basis. Multilayer Perceptrons

have been applied successfully to solve difficult and diverse problems by training them in a supervized manner [66].

3.10 Measures of Performance Evaluation

3.10.1 K-fold-Cross-Validation

K-fold cross-validation is most commonly used when comparing the predictive accuracy of two or more methods in order to minimize the bias associated with the random sampling of the training and holdout data samples. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," *D*1, *D*2, … , *Dk*, each of approximately equal size. Training and testing is performed *k* times. In iteration *i*, partition *Di* is reserved as the test set, and the remaining partitions are collectively used to train the model [36]. The cross-validation estimate of the overall accuracy is calculated from the average of the k individual accuracy measures and is given as:

$$CVA = \sum_{i=1}^{k} A_i$$

where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each fold [53].

3.10.1.1 10-Fold-Cross-Validation

According to Witten et al. [28], extensive tests on several datasets using different learning techniques have shown that 10 is about the right number of folds to get the best estimate of error, and this claim is also backed up by some theoretical evidence. Although these claims are by no means conclusive, and debate continues to linger in machine learning and data mining circles about what is the best scheme for evaluation, 10-fold cross-validation has become the standard method in practical terms. Tests have also shown that the use of stratification improves results slightly. Thus, the standard evaluation technique in situations where only limited data is available is stratified 10-fold-cross-validation.

In 10-fold cross-validation, the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original dataset (stratified). Each

fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates. 10 seem to be an optimal number of folds that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process [53].

10-fold cross-validation is implemented as follows:

- The entire dataset is randomly divided into 10 disjoint subsets (folds), with each fold containing approximately the same number of records. The sampling is stratified by the class labels to ensure that the subset class proportions are roughly the same as those in the whole dataset.

- For each fold, a classifier is constructed using the nine of the 10 folds and tested on the tenth one to obtain a cross-validation estimate of its error rate.

- The 10 cross-validation estimates are then averaged to provide an estimate for the classier accuracy constructed from all the data.

3.10.2 Confusion Matrix

Given m classes, a confusion matrix is a table of at least size m by m. Confusion matrix is a useful tool for analyzing how well a classifier can recognize instances of different classes. A binary classification model classifies each instance into one of two classes, for instance, positive or negative class. This results in four possible classifications for each instance: true positive, true negative, false positive, or false negative. A confusion matrix for two classes is shown in Table 3.7. A confusion matrix juxtaposes the observed classifications for a phenomenon (rows) with predicted classifications of a model (columns). The classifications that lie along the major diagonal of the table are the correct classifications, that is, true positives and true negatives. The other cells represent model misclassifications. It is common to call true positive *hits*, true negatives *correct rejections*, false positive *false alarms*, and false negative *misses*. A number of model performance metrics can be derived from a confusion matrix, the most common metric being classifier accuracy defined by the following formula [36]:

Accuracy = TP + TN/TP + TN + FP + FN

Other performance metrics include precision and recall defined as follows:

Precision = TP/TP + FP

Recall = TP/ TP + FN

Table 3.7 A two-class confusion matrix

		Predicted	
Observed		C1	C2
	C1	True positive	False negative
	C2	False positive	True negative

3.10.3 Area Under the ROC Curve

ROC curves are two-dimensional graphs that usually depict the performance and performance trade-offs of a classification model. Receiver Operating Characteristics (ROC) analysis provides tools to select possibly optimal models and to discard suboptimal ones. Receiver operating Characteristics (ROC) curves usually convey the same information as confusion matrix in a much more intuitive and robust way. The machine learning community most often uses the ROC AUC (Area Under ROC) statistic for model comparison. This measure can be interpreted as the probability that when we randomly pick one positive and one negative example, the classifier will assign a higher score to the positive example than to the negative. Two performance metrics, True Positive Rate (TPR) and False Positive Rate (FPR), are used to construct ROC curves. ROC curves are constructed by plotting the true positive rate against the false positive rate [57].

True Positive Rate (TPR) = TP/TP + FN = Recall

False Positive Rate (FPR) = FP/FP + TN

A ROC curves of two classification models is depicted in figure 3.3. The plot shows a diagonal line where for every true positive of such a model, it is just as likely to encounter a false positive. Thus, the closer the ROC curve of a model is to the diagonal line, the less accurate the model. If the model is really good, initially it is more likely to encounter true

positives in moving down the ranked list. Thus, the curve would move steeply up from zero. Later, on encountering fewer and fewer true positives, and more and more false positives, the curve tends to flatten and becomes more horizontal [36].

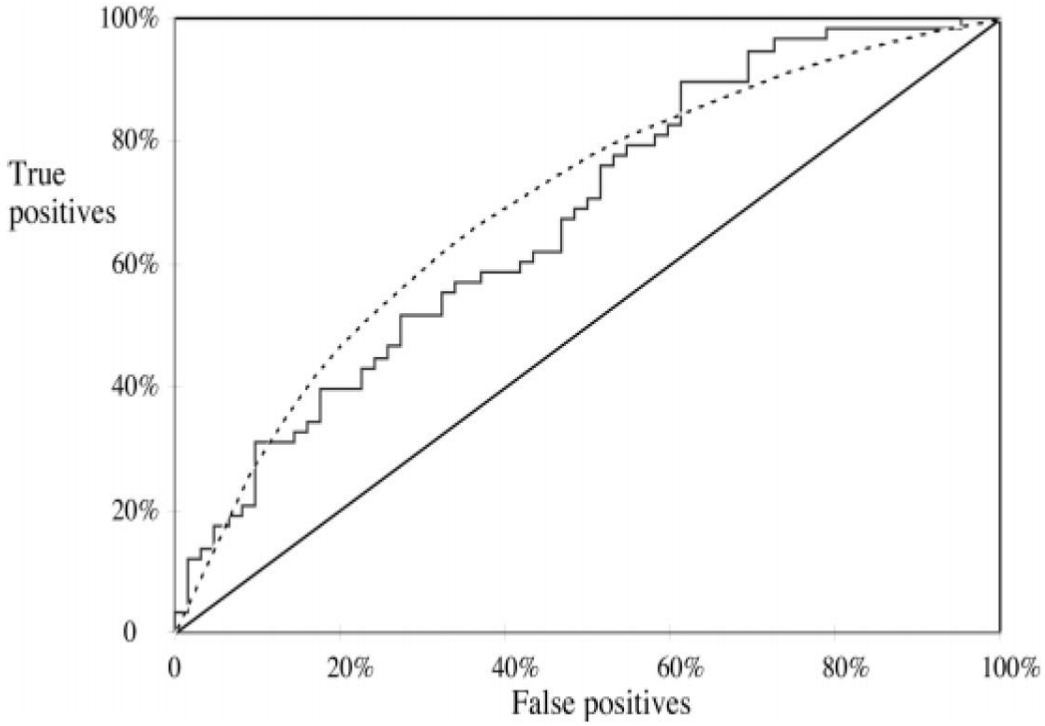


Figure 3.3 The ROC curves of two classification models

The following are observations regarding AUC (Area Under ROC):

1. It has value between [0, 1].
2. A random classifier has an AUC ~0.5.
3. The higher the value of AUC the better the distinguishing capability of the Classifier [36].

CHAPTER FOUR

EXPERIMENTATION

The main objective of this study was to predict fertility preference of women of reproductive age. Hence, data mining classification methods were chosen to develop predictive models. Different experiments were done using three algorithms: J48 Decision Tree, Naïve Bayes and Multilayer Perceptron in Weka 3.6.3.

4.1 Experimental Design

Different experiments were constructed for each classifier using the entire dataset consisting of 16515 instances. A stratified 10-fold cross-validation was used to estimate the performance of each classifier. This performance estimation approach has been proved to be statistically good enough in evaluating the performance of data mining classifier algorithms. Overall classification accuracy, TP ate, TN rate, precision, recall, and ROC area were used to evaluate and compare the performance of the models. These measures were driven from the confusion matrix of the models.
The experiments were conducted for each classifier based on:

1. All attributes.
2. Selected attributes and.
3. additionally for J48 Decision Tree,
    a. Pruning with all attributes.
    b. Pruning with selected attributes.

The above scenarios were considered to investigate the effects of feature selection on the classification accuracy of each classifier and the time taken to build models by each classifier. Moreover, it is to see the effect of tree pruning method on classification accuracy of J48 and the tree size generated by this algorithm.

4.2 Feature Selection

To proceed with the experiments, feature subset selection is performed. Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy. In general, feature selection techniques can be categorized into two: filter methods and wrapper methods. Filter methods operate independently of the learning algorithms while wrapper methods take into account the learning algorithms to be used [67]. To select feature subsets, Weka's entropy based information gain attribute evaluator algorithm with ranker search strategy was used.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 15 Class):
        Information Gain Ranking Filter

Ranked attributes:
 0.162382   1 Age
 0.145263   8 NumofLvgchldrn
 0.029369   4 Education
 0.027642  10 Numofsonsdied
 0.025659  11 Married
 0.025373   9 Numofdaughtersdied
 0.01774    7 SexofChild
 0.015921   2 Region
 0.005598  12 FrqcyofwatchgTV
 0.004005   6 WealthIndex
 0.002378   5 Religion
 0.002327   3 Residence
 0.001178  14 Working
 0.000168  13 Knowsmethod

Selected attributes: 1,8,4,10,11,9,7,2,12,6,5,3,14,13 : 14
```

Figure 4.1 Summary of ranked attributes

As can be seen from figure 4.1, eight out of fourteen attributes with relatively better information gain are Age, NumofLvngchldrn, Education, Numofsonsied, Married, Numofdaughtersdied, SexofChild and Region.

4.3 Model Building Using J48 Classifier

Experiment1

In this experiment**,** the performance of J48 classifier in predicting fertility preference is evaluated. Two models were built (both unpruned) in two scenarios on all 14 attributes and the selected 8 attributes.

In the first scenario, it took 2.42 seconds to build a model with tree size 691 and 513 leaves. It took 1.89 seconds to build the second model and the model generated tree size of 244 and 167 leaves. The model is faster and relatively less complex than the first model. Figure 4.2 shows sample run information for the first scenario.

```
Number of Leaves  :     513

Size of the tree :      691


Time taken to build model: 2.42 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12848               77.7959 %
Incorrectly Classified Instances      3667               22.2041 %
Kappa statistic                          0.4743
Mean absolute error                      0.313
Root mean squared error                  0.4131
Relative absolute error                 70.1699 %
Root relative squared error             87.4694 %
Total Number of Instances            16515

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.887     0.438      0.8       0.887     0.841      0.767    1
                 0.562     0.113      0.715     0.562     0.63       0.767    0
Weighted Avg.    0.778     0.329      0.772     0.778     0.77       0.767

=== Confusion Matrix ===

    a    b   <-- classified as
 9731 1240 |    a = 1
 2427 3117 |    b = 0
```

Figure 4.2 Sample output for unpruned J48 classifier with all attributes

Table 4.1 Performance measures for experiment 1

Model	TP Rate	TN Rate	Accuracy	Precision	F-Measure	ROC area
J48 unpruned with all attributes	0.778	0.562	77.80%	0.772	0.77	0.767
J48 unpruned with selected attributes	O.779	0.532	77.86%	0.773	0.768	0.781

Confusion matrix in figure 4.2 indicates that the first model built with J48 (unpruned) with all attributes correctly classified 12848 (77.80%) instances and the number of incorrect classification is 3667.That is, the error rate for this model is 22.20%. This overall accuracy rate can be deemed as good but looking at the precision and recall of the model can be more interesting. For instance, the precision for woman who are labelled as wanting more children is 80% and the recall is approximately 89% (Figure 4.2). This means that out of the total number of women who are labelled as wanting more children, 80% of them are actually found to be those who want more children and 89% of women who actually want more children actualize their stated preference for children. This shows that these levels of precision and recall are non-trivial and may be very useful in the application domain. The average precision and recall of the model (Table 4.1), respectively, are 77.2% and 77.8%.

The TP rate of 77.8% indicates that the model correctly identified out of 10971 women, 9731 of them who want more children and 1240 of them were incorrectly classified as though they want no more children while they actually have the desire to have more children. The TN rate of 56.2% shows that the model correctly classified 3117 woman out of 5544 women as those who want no more children and 2447 of them were incorrectly classified as those who want more children while actually they don't. There is high disparity between TP rate and TN rate indicating the bias of the model toward the majority class due to the fact that the dataset is unbalanced.

The goal of the second scenario in this experiment is to inspect the effect of reducing attributes on the performance of a model. The second model built with J48 (unpruned) with the selected 8 attributes correctly classified 12858 (77.86%) instances with an error rate of 22,14% (3657 instances misclassified, not shown here), indicating a slight increase in classification accuracy over the first model built with all attributes. The tree size and number of leaves were also reduced. Other measures like recall and ROC area have also slightly increased (see Table 4.1). This shows selecting relevant attributes for model building improves the performance of the model and decision tree size.

Experiment 2

The aim of this experiment is to investigate:

a. The effect of tree pruning on the performance of J48 classifier with all attributes.
b. The effect of tree pruning on the performance of J48 classifier with the selected 8 attributes.

Tree pruning (post-pruning) is implemented on a fully induced decision tree, and examines the tree to remove statistically insignificant nodes. Working from the bottom up, the probability (or relative frequency) of sibling leaf nodes will be compared, and any overwhelming dominance of a certain leaf node will result a pruning of that node in one of several ways. The error estimate of each child node is calculated and used to derive the total error of the parent node. The parent node is then pruned according to the relative frequencies of the child nodes, and this replacement node's error is compared with that of the old parent node, which was influenced by the child nodes' error. This comparison will dictate whether or not pruning is advantageous at a given node [68].

The parameter used to prune tree (post-pruning) is labelled by Weka as a confidence factor. For this experiment, J48 classifier was tested with confidence factors ranging from 0.24 to 0.1 by a decrement of 0.2. To assess the effect of tree pruning on the performance of J48 classifier, a series of trials were made, and the classification accuracy for each trial recorded. The goal was to get a model with a better overall classification accuracy and the least complex decision tree that is easy to interpret by domain users. The number of minimum instances per node (minNumObj) was kept at 2 (the default) and the number of folds for the testing set (numFolds) was changed and held at 10 during pruning. Classification accuracies of models from J48 pruned at different confidence factors are shown in Table 4.2.

Table 4.2 Classification accuracy of J48 pruned for different confidence factors

Model									
J48 pruned with all attributes	Accuracy(%)	77.76	77.81	77.89	78.07	78.21	78.37	78.26	78.25
	Con. factor	0.24	0.22	0.2	0.18	0.16	0.14	0.12	0.1
J48 pruned with selected attributes	Accuracy(%)	77.94	77.98	78.03	77.95	77.86	77.82	77.74	77.66
	Con. factor	0.24	0.22	0.2	0.18	0.16	0.14	0.12	0.1

As shown in Table 4.2, performance of the classifier with all attributes increased as the confidence factor increased up to 0.14 at a peak of 78.37% accuracy, after which the accuracy began to degrade. This is the accuracy with minimum error rate (21.63%). For J48 classifier with the selected attributes, the maximum accuracy rate was reached at a confidence factor of 0.2 and then it began to decline afterwards. This is also the accuracy with minimum error rate. But when decision tree complexity is considered, the least complex decision trees are reached at a confidence factor 0.1 for both scenarios. Confusion matrices and performance measures for J48 classifier models found at confidence factors 0.14 and 0.2 are, respectively, shown in Table 4.3 and Table 4.4. Figure 4.3 shows sample output for pruned J48 classifier with all attributes.

```
Number of Leaves  :     126

Size of the tree :      170


Time taken to build model: 2.51 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12923               78.2501 %
Incorrectly Classified Instances      3592               21.7499 %
Kappa statistic                          0.4828
Mean absolute error                      0.3207
Root mean squared error                  0.4039
Relative absolute error                 71.9089 %
Root relative squared error             85.5392 %
Total Number of Instances            16515

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.44       0.801     0.895     0.845      0.785    1
                 0.56      0.105      0.729     0.56      0.634      0.785    0
Weighted Avg.    0.783     0.327      0.777     0.783     0.774      0.785

=== Confusion Matrix ===

    a    b   <-- classified as
 9816 1155 |    a = 1
 2437 3107 |    b = 0
```

Figure 4.3 Sample output for pruned J48 classifier with all attributes

In this experiment, J48 classifier algorithm run on a full training set with 14 attributes took 2.01 seconds to build the model and the model generated relatively smaller and less complex tree with a size of 170 and 126 leaves. With the selected 8 attributes, a model built with J48 classifier generated smaller and less complex tree (relatively) with a size of 117 and 85 leaves and it was built in 1.65 seconds.

Table 4.3 Confusion matrices for J48 classifier (Experiment 2)

Model	Confusion Matrix		
J48 pruned with all attributes	1 (predicted)	0 (predicted)	Observed
	9841	1130	1
	2443	3101	0
J48 pruned with selected attributes	1 (predicted)	0 (predicted)	Observed
	9920	1051	1
	2586	2958	0

Table 4.4 Performance measures of J48 classifier (experiment 2)

Model	Accuracy	TP rate	TN rate	Precision	F-measure	ROC area
J48 pruned with all attributes	78.37%	0.784	0.559	0.778	0.775	0.785
J48 pruned with selected attributes	78.03%	0.78	0.534	0.775	0.769	0.783

As shown in Table 4.3, the tree-pruned J48 classifier model built with all attributes correctly classified 12942 (78.37%) instances and 3573 (21.63%) of the instances were classified incorrectly. The accuracy rate of this model has increased by 0.57% as compared to the unpruned model with all attributes in experiment 1. The time taken to build the model was 1.65 second.

Out of 10971 women who are labelled as those who want more children, the model correctly identified 9841 of them and the remaining 1130 were incorrectly identified as those who want no more children. With this, the model achieved TP rate of 0.784 (see Table 4.4). With a TN rate of 0.559, the model correctly identified 3101 out of 5544 women who want no more children and the remaining 2443 were incorrectly classified as those who want more children while they actually do. This shows that the model is good at identifying women who want more children than those who want no more children.

The average precision score of 77.8% indicates that the model is successful in identifying relevant values for each class. The ROC area 0.785 shows the goodness of the model at discriminating between true positives and false positives, and the F-Measure score of 77.5% implies that the precision and recall of the model are significantly balanced (TP rate = Recall, see table 4.4) Generally, tree size and number of leaves significantly decreased as confidence factor decreased

4.4 Model Building Using Naïve Bayes Classifier

Experiment 3

The purpose of this experiment was to evaluate the performance of Naïve Bayes classifier algorithm by considering two cases: building two models using the dataset with all attributes and the selected attributes sub set. The algorithm was run using 14 attributes and the selected attribute subset containing 8 attributes. The goal is to investigate whether attribute reduction improves or degrades the performance of the model built using the full set of attributes.

In the first case, the algorithm was run using the entire dataset with 14 attributes and it took 0.13 second to build the model. For the second case, the algorithm was run using the selected attributes and the model was constructed in 0.09 seconds.

Table 4.5 Confusion matrices for Experiment 3

Model			
Naïve Bayes with all attributes	1 (predicted)	0 (predicted)	Observed
	9167	1804	1
	2326	3218	0
Naïve Bayes with selected attributes	1 (predicted)	0 (predicted)	Observed
	9402	1569	1
	2408	3136	0

Table 4.5 shows that the first model built using 14 attributes correctly classified 12385 (74.99%) instances and 4130 (25.1%) instances were incorrectly classified. This accuracy rate was achieved after descretizing the numeric attributes. The second model built using selected 8 attributes correctly classified 12538 (75.92%) instances while 3977 (24.08%) instances were classified incorrectly. Naïve Bayes classifier works with the assumption that predictor variables are independent given the class. The improved accuracy of 75.92% by the second model may be attributed to the elimination of redundant attributes.

Table 4.6 Performance measures for Experiment 3

Model	Accuracy	TP rate	TN rate	Precision	F-measure	ROC area
Naïve Bayes with all attributes	74.99%	0.75	0580	0.745	0.747	0.784
Naïve Bayes with selected attributes	75.92%	0.759	0.566	0.753	0.754	0.79

Table 4.6 shows a slight increase in all performance measures but TN rate reflecting the effect of attribute reduction on model performance

4.5 Model Building Using Neural Network

In this experiment, the performance of Neural Network in predicting fertility preference was evaluated. Two scenarios were considered to do the experiment: building model using multilayer Perceptron with the entire 14 attributes and repeating this using only the selected attributes to see the effect of attribute reduction on the performance of the classifier.

Experiment 4

WeKa's Multilayer Perceptron algorithm has several parameters which can influence its performance. The hiddenLayers parameter, for instance, is used to set the number of hidden layers (if available in the network). The predefined value for hidden layer is a, which is the average of number of predictor variables and number of class values. Other important parameters include learning rate and momentum. Firstly, this experiment was done on the default parameter values for all parameter, and then Multilayer Perceptron was tested by varying learning rate and momentum.

Table 4.7 Confusion matrices for Experiment 4

Model			
Multilayer Perceptron with all attributes	1 (predicted)	0 (predicted)	Observed
	9195	1776	1
	2348	3196	0
Multilayer Perceptron with selected attributes	1 (predicted)	0 (predicted)	Observed
	9746	1225	1
	2546	2998	0

Multilayer Perceptron model built using all attributes correctly classified 12391 (75.03 %) instances and 4124 (24.97%) instances were incorrectly classified. The second model with reduced attributes incorrectly classified 3771 (22.84%) while 12744 instances were correctly classified with an accuracy rate of 77.17 %, indicating an increment of the accuracy by more than 2 % due to attribute reduction. The time taken to build the first model using all attributes was 1660.89 seconds while the second model generated from Multilayer Perceptron run on the selected 8 attributes was built in 522.71 seconds.

Table 4.8 Performance measures for experiment 4

Model	Accuracy	TP rate	TN rate	Precision	F-measure	ROC area
Multilayer Perceptron with all attributes	75.03 %	0.75	0.577	0.745	0.747	0.781
Multilayer Perceptron with selected attributes	77.17 %	0.772	0.547	0.765	0.763	0.811

Table 4.8 shows performance results for models built with all and reduced attributes. Multilayer Perceptron model built using reduced attributes improved accuracy, average TP rate, precision, f-measure and ROC area.

In this experiment, Multilayer Perceptron was also tested by varying the learning rate and momentum parameters using the selected attributes. Results of Multilayer Perceptron are presented in Table 4.9, where LR is learning rate, M is momentum and NHL is number of hidden layers.

Table 4.9 Performance measures of MLP for Variable Parameters

Model	LR	M	NHL	TP rate	FP rate	Precision	Accuracy (%)
Multilayer Perceptron with selected attributes	0.2	0.7	a	0.773	0.338	0.767	77.34 %
	0.3	0.7	a	0.766	0.351	0.759	76.60%
	0.5	0.7	a	0.77	0.33	0.764	77.03%
	0.6	0.9	a	0.645	0.312	0.703	64.52%

The first three values for learning rate and momentum are obtained by trial while the 0.6 and 0.9 learning rate- and-momentum combination was used as suggested in [64]. As shown in Table 49, this latter combination resulted in the worst performance but it took the least time to build the model indicating the fact that higher learning rates make Multilayer Perceptron algorithm faster. The first model (LR =0.2, M = 0.7) performed better with better accuracy

and precision. This shows that farther tuning of the parameters including the number of hidden layers might improve the performance of the algorithm.

4.6 Discussion

To develop models that can predict fertility preference of women, J48, Naïve Bayes and Multilayer Perceptron algorithms were used. Five experiments were done based on all features and selected 8 features of the dataset**.** The experiments were designed to investigate whether tree pruning improves or degrades the performance of J48 classifier, to evaluate the effect of attribute reduction on the performance of the three algorithms and the effect of parameter tuning on the performance of MLP and finally, to compare the performance of the algorithms in predicting fertility preference of women of reproductive age.

4.6.1 Effects of J48 Classifier Decision Tree Pruning

In experiment 2, a series of trials were made to assess the effect of tree pruning on the performance of J48 classifier. Eight different confidence factors were used while tree pruning for the first model built using the data set with all attributes and then, the same set of confidence factors were also used for tree pruning on the second model built using the selected attributes. In general, decision tree pruning improved the classification accuracies of both models, but the increase in accuracy, which is < 1%, does not seem significant. However, decision tree pruning significantly decreased tree size and number of leaves generated which are important for easy interpretation.

4.6.2 Effects of Attribute Selection

For each classifier, two experiments were done to develop two models. The first model for all classifiers was built using the dataset containing all attributes, and then, the second model for all classifiers was constructed using only the selected attributes. The aim was to observe the effect of attribute reduction on the accuracy of the models. Figure 4.3 shows the accuracy of the classifiers before and after attribute reduction.

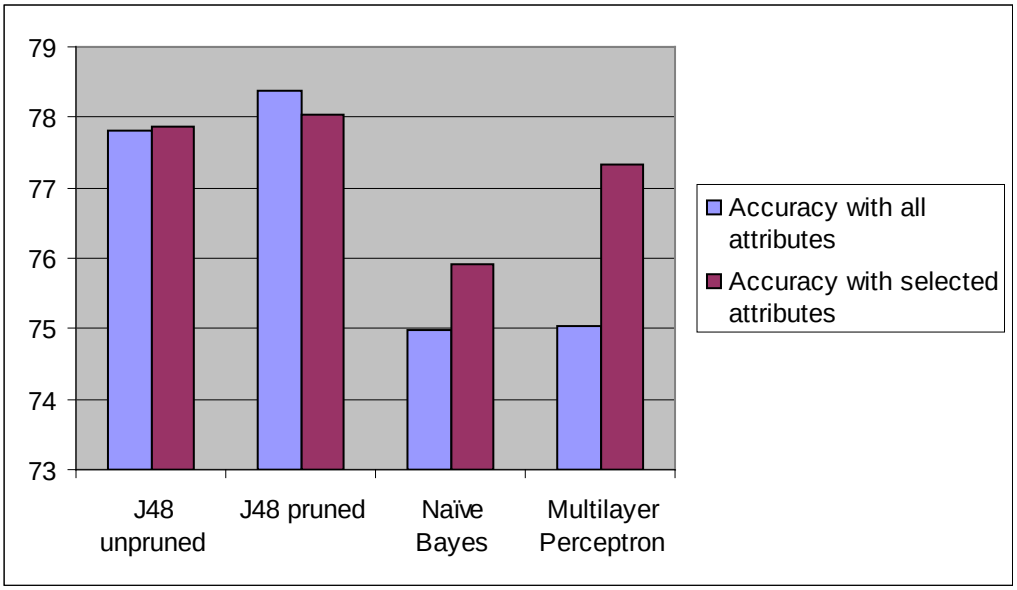


Figure 4.4 Effects of attribute selection on classification accuracy

As shown in Figure 4.3, for most of the classifiers, the overall accuracy slightly increased after attribute reduction. After attribute reduction, the accuracy for J48 model (unpruned) increased from 77.80% to 77.86%, and for the pruned J48classifier model, it increased from 77.94% to 78.03%. For Naïve Bayes classifier model, the accuracy increased from 74.99% to 75.92% after attribute reduction. This shows accuracies for J48 (unpruned) and Naïve Bayes models increased only by less than 1%. A significant change in accuracy was observed for Multilayer Perceptron model. It increased from 75.03% to 77.17%, with a significant increment of 2.14%. Attribute reduction also significantly decreased the time taken to build the models. In general, attribute reduction improved the classification accuracy of the models given the attribute selection algorithm used. The algorithm used for selecting attributes was (InfoGainAttributeEval + Ranker in Weka) which works heuristically while selecting attributes.

4.7 Model Comparison

In all experiments, a base model for each classifier was constructed using the entire dataset with all attributes and then, a second model was developed using the same dataset with the selected attributes. For all models, a 10-fold-cross-validation test was implemented. The models were compared using performance evaluation metrics like accuracy, TP rate, TN rate, F-measure and ROC area. The time take by each classifier to build the models was also considered. Results for each model is presented in Table 4.10.

Table 4.10 Performance summary for each model

Model	Accuracy	TP Rate	TN Rate	F-Measure	ROC Area	Time (sec)
J48 unpruned with all attributes	77.80%	0.778	0.562	0.77	0.767	2.56
J48 unpruned with selected attributes	77.86%	O.779	0.532	0.768	0.781	2.17
J48 pruned with all attributes	78.37%	0.784	0.559	0.775	0.785	2.0
J48 pruned with selected attributes	78.03%	0.78	0.534	0.769	0.783	1.65
Naïve Bayes with all attributes	74.99%	0.75	0580	0.747	0.784	0.13
Naïve Bayes with selected attributes	75.92%	0.759	0.566	0.754	0.790	0.09
MLP with all attributes	75.03 %	0.75	0.577	0.747	0.781	1660.89
MLP with selected attributes	77.34%	0.773	0.547	0.765	0.807	513.49

As shown in Table 4.10, J48 classifier (pruned and unpruned) applied to the dataset with all attributes and selected attributes achieved classification accuracies ranging from 77.80% to 78.37%. With the highest classification accuracy of 78.37%, it is the best classifier in predicting fertility preference. Multilayer Perceptron applied to the dataset with selected attributes was found to be the second achiever with an accuracy rate of 77.34%. Naïve Bayes implemented on the dataset with reduced attributes and MLP with all attributes performed nearly equally with accuracies, respectively, of 75.92% and 75.03 %. And finally, Naïve Bayes applied to the dataset with all attributes achieved the least classification accuracy of 74.99%.

Regarding TP rate and TN rate, all classifiers yielded similar results in all cases but the big difference observed between these measures is due to the fact that the dataset is unbalanced.

In terms of F-Measure, all classifiers achieved high average results indicating that the precision and recall of the models are higher. That is, the models are good at retrieving more relevant instances than irrelevant for each class, where most of the relevant instances to a given class are actually retrieved. ROC area was also used for measuring the performance of the models. Regarding this, all classifiers yielded high and similar results. As shown in Table 4.10, ROC area for algorithm built using the dataset with selected attributes slightly improves over the ROC area for the same algorithm built using the dataset with all attributes. Overall, the values for ROC area for all models indicate the goodness of the models in identifying true positives.

The classifiers were also compared based on the time required for building the models. Accordingly, Naïve Bayes classifier was found to be the fastest while J48 was a bit slower, and the time taken to run Multilayer Perceptron was significantly so high. It was observed that all the classifiers performed faster when run on the dataset with reduced attributes than with all attributes. This indicates that attribute reduction is very important in reducing execution time for algorithms, particularly for algorithms like Multilayer Perceptron which take longer time while constructing model.

In summary, several models were built using the dataset with all and selected attributes to investigate the effect of attribute reduction on the classification accuracy and execution time of the model. Besides, the effect of decision tree pruning on the performance of J48 classifier was investigated. The classifiers were tested on 10-fold-cross validation. In terms of TP rate, TN rate and F-Measure, all of the classifiers performed nearly equally while MLP slightly improved ROC area when run on selected attributes. Based on classification accuracy, J48 classifier performed the best with a classification accuracy of 78.37% (unpruned) and 78.03% (pruned), respectively. Multilayer Perceptron achieved an accuracy of 77.34%. Naïve Bayes classifier performed the least with accuracy of 75.92%. On tuning a few parameters, Multilayer Perceptron tended to improve prediction performance. This indicates that tuning of more parameters may improve further the performance of the classifier.

With the dataset with all and reduced attributes, J48 relatively performed the best in predicting fertility preference of women. J48 classifier also provides decision rules that domain users can easily understand and interpret.

4.8 Some Specific Rules

In this study, J48 classifier has achieved relatively the highest classification accuracy as compared to Naïve Bayes and Multilayer Perceptron. Accordingly, J48 classifier model was selected as the best model that can predict fertility preference of women. This model generated 60 rules and the following rules that contain most instances of the dataset were extracted.

Rule 1
If NumofLvgchldrn <= 1 and Married = 0 (No) then 1(want more child)/7183.0/931.0/

Rule 1 implies that women who have no or one child and who are unmarried tend to desire for more children.

Rule 2
If NumofLvgchldrn <= 1 and Married = 1(Yes) and Age = 15-19 then 1(want more child) (160.0/38.0)

Rule 3
If NumofLvgchldrn <= 1 and Married = 1(Yes) and Age =20-24 then 1(want more child) (215.0/47.0)

Rule 4
If NumofLvgchldrn <= 1 and Married = 1(Yes) and Age =25-29 then 1(want more child) (183.0/53.0)

Rule 5

If NumofLvgchldrn <= 1 and Married = 1(Yes) and Age =30-34 1(want more child) (78.0/23.0)

Rule 6
If NumofLvgchldrn <= 1 and Married = 1(Yes) and Age =30-34 then 1(want more child) (79.0/30.0)

According to rules 2 to 6, younger women who are in marriage and have no or one living child have no intention to limit child bearing.

Rule 7

If NumofLvgchldrn>4 and Age = 40-44 then 0(want no more child) (772.0/203.0)

Rule 8
If NumofLvgchldrn>4 and Age =45-49 then 0(want no more child) (705.0/83.0)

The implication of rules 7 and 8 is that women intend to limit childbearing because of old age or have more than four living children.

Rule 9

If Age = 35-39 and Education = 0 and Region = 1 and Region = 3 Region = 4 then 0(want no more child) (300/103)

From Rule 9 and other rules (not listed here) regarding women's educational status and their desire for children, uneducated or less educated women are more likely to want to limit childbearing than the more educated ones. This may be due to the fact that they already have more children than the educated ones.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Fertility is one of the factors that determine the overall size, distribution and structure of human population. It is a reproductive performance of an individual, a couple or a group of population. Crude birth rate, age-specific fertility rate and total fertility rate are among the measures used to estimate fertility. Total fertility rate is an indicator of fertility level of a country. It is a measure of the number of children a woman would have by the end of her child bearing years if she were to pass through those years bearing children at the observed age-specific fertility rates.

Demographic research has shown that socio-economic and cultural factors influence fertility through biological and behavioural determinants. The socio-economic determinants, which are deemed as indirect determinants, include social, cultural, economic, institutional, psychological, health, and environmental factors and biological and behavioural determinants, the direct determinants of fertility, include proportion of women married or in sexual unions, frequency of intercourse, and postpartum abstinence, among others**.** The proximate determinants have direct influence on fertility, whereas, socioeconomic variables can affect fertility only indirectly by modifying the proximate determinants.

High fertility results in high incidence of births of order five and above, a high fraction of women experiencing pregnancies of order five and above, and a greater likelihood of short inter-pregnancy intervals. These experiences cause harmful consequences like health risks for children and their mothers. High fertility also contributes to high population growth rate which causes population explosion. Rapid population growth negatively influences human capital investment, economic growth, and the environment. This effect of high fertility is reflected, for instance, in formal schooling of children where children in high fertility countries have less chance of going to school, depletion of natural resources and migration

Fertility preference is a conscious choice for child bearing. People's preferences for fertility have a predictive value for fertility and might be indicative of how many children they would eventually have. This estimation of actual fertility can be achieved through fertility preference measures which used to capture some dimension of an individual's attitude or motivation to influence fertility outcomes. Moreover, fertility preference defines the demand for children and indicates the motivation to limit fertility deliberately. Attitude on fertility preference are expressive and meaningful for measuring reproductive choices, for assessing the motivation for fertility regulation, and for analyzing future prospects of fertility change. The measurement of fertility preference is also essential for understanding the dynamics of fertility change, to make forecasts about medium-term changes in fertility, and to measure the prevalence of unwanted births, and thus the prevalence of an unmet need for family planning services so that different intervention programs could be designed and implemented.

Socio-economic factors that affect fertility preferences include social and economic status of an individual such as level of education, income, housing, age, and residence identity. Institutional factors like legal and religious institutions also influence people's fertility preference.

Data mining technologies have many applications in various industries. Business organization like banking, telecommunications, manufacturing industries, airlines, and health insurance companies have been exploiting data mining technologies for gaining competitive advantage in different business application areas. In banking, for instance, data mining can be applied for predicting levels of bad loans and fraudulent credit card use and credit card spending by new customers. Likewise, in manufacturing companies, data mining technologies are used for predicting machinery failures and for finding key factors that control optimization of manufacturing capacity. Capturing data on where customers are flying and the ultimate destination of passengers who change carriers in hub cities so that airlines can identify popular locations that they do not service and checking the feasibility of adding routes to capture lost business are some of the applications of data mining in airlines.

Data mining is also gaining importance in healthcare systems and institutions. Data mining can be applied in different application areas in healthcare. The major application areas include the evaluation of treatment effectiveness, management of healthcare, customer relationship management, detection of fraud and abuse, early detection and/or prevention of

diseases, policy-making in public health, and early detection and management of pandemic diseases and public health policy formulation.

The effectiveness of medical treatments can be evaluated using data mining models. A typical analysis of treatment effectiveness is comparing the outcomes of patient groups treated with different drug regimens for the same disease or condition to determine which treatments work best and are most cost-effective. Data mining models are also found to be more effective in the diagnosis and prognosis of chronic diseases like breast cancer, heart diseases, and diabetes.

However, some real issues constrain the application of data mining in healthcare. One of such issues is the heterogeneity of healthcare data. Healthcare data usually are fragmented and distributed between hospitals, insurance companies and government departments or may not be available in electronic format. To solve these problems, data warehousing technologies can be implemented for integrating distributed databases into a central data repository to create efficient and effective data access environment. The question of data ownership, fear of lawsuits and privacy issues also pose similar challenges to the application of data mining in health care. Medical professionals usually are unwilling to share patient data with researchers for fear of potential lawsuits that may be triggered by discovering anomalies in patient medical histories. Regarding privacy issues, administrators and researchers are required to pay strict attention to privacy and security when transferring, storing, or mining patient data. This is to keep patient privacy and doctor-patient confidentiality. While such practices are inevitable, they still hinder the application of data mining in healthcare. The need for incremental data mining techniques is another challenge in medical data mining due to the fact that medical data are usually up-to-date since new data are constantly being added to the existing databases from various medical sources. Data mining model built from such databases become obsolete since the model does not include new data.

In this paper, the main objective was to develop a model that can predict fertility preference of women of reproductive age. Three classification algorithms in Weka, J48, Naïve Bayes and Multilayer Perceptron, were implemented to build and compare different models. All classifiers were tested using ten-fold-cross-validation. For each classifier, a model was built using the dataset with all features, and then a second model was built using the dataset with reduced features and the performances of the models were compared. The goal was to

evaluate the effect of features selection on the performance of the classifiers, that is, the effect of attribute reduction on classification accuracy and execution time of the classifiers.

To assess the effect of tree pruning on the performance of J48 classifier, a series of trials were made. This classifier was tested with confidence factors (a parameter used to prune decision tree in Weka) ranging from 0.24 to 0.1 by a decrement of 0.2. The number of minimum instances per node (minNumObj) was kept at 2 and number of folds for the testing set (numFolds) was changed and held at 10 during pruning. Multilayer Perceptron was also tested by varying the learning rate and momentum parameters using the dataset with selected attributes.

Feature selection improved both classification accuracy and execution time for all classifiers. The increment in classification accuracy for J48 and Nave Bayes was less than one percent but a significant increment was observed for Multilayer Perceptron which is more than two percent. Even better classification accuracy was achieved by Multilayer Perceptron after tuning learning rate and momentum parameters, respectively, to 0.2 and 0.7. The accuracy was 77.34%. Naïve Bayes classifier performed the least with accuracy 75.92% on the dataset with reduced features.

Decision tree pruning also improved accuracy of J48 classifier. The accuracy increased from 77.76% to 78.37% with all features and it creased from 77.94% to 78.03% with selected features.

Therefore, based on classification accuracy, J48 model is found to be relatively the best model in predicting fertility preference of women. Furthermore, J48 model provides decision rules that can easily be understood and interpreted.

It was learned that age, number of living children, education, child death experience, marital status, sex of child and region are the most important factors that determine fertility preference of women.

5.2 Recommendations

This research paper has provided an initial insight into the potential applicability of data mining methods in predicting fertility preference of women based on some demographic and socioeconomic characteristics. With this, the main objective of this study was achieved. Based on the findings of this study, it is recommended that the following issues need to be addressed in future studies:

1. The main objective of this study was to develop a model that can predict fertility preference of women. Yet, it has been claimed in the literature that fertility preference of men is also equally or more important in the process of family building. That is, the process of family formation, particularly within marriage, is dependent on the consensus between women and their husbands (partners). Accordingly, identification of relevant features from men's dataset and merging these into women's dataset should be considered in future studies.
2. In this study, all the classifies improved accuracy and other performance measures after implementing a filter type feature selection method. However, this feature selection method does not involve the learning algorithm to be used while selecting features. Thus, in future study, applying other feature selection methods, like wrapper, that involve the learning algorithm while selecting features might result in more improvement in the performance of the classifiers.

3. Testing other algorithms like logistic regression and support vector machine(SVM) and comparing the results with the results of this study is also important.

4. The dataset used in this study is imbalanced. It is commonly known that classifiers trained on imbalanced datasets create bias in prediction and overfitting in the test set. While decision tree pruning and feature selection methods used for this study might have addressed this problem to some extent, it is important that other most direct techniques, like Synthetic minority over-sampling technique (SMOTE), should be applied to balance the dataset and see if this could improve the prediction accuracy of the classifiers.

5. While this initial work didn’t guarantee a complete software solution for classifying women according their childbearing preferences, it did, however, prove that it is possible to apply data mining in this area and hopefully leads the way to further research and a working implementation.

References

[1] Human Population Dynamics. http://www.learner.org/envsci/guide.pdf. Accessed 8 Oct 2012

[2] Mekonnen W. and Worku A. 2011. Determinants of Fertility in Rural Ethiopia: The Case of Butajira Demographic Surveillance System (DSS). http://www.biomedcentral.com/1471-2458/11/782. Accessed 8 Oct 2012

[3] Onsembe J.O. 2006. ETHIOPIA: Situation Analysis on Population, Reproductive Health and Gender. Addis Ababa

[4] Davis K. and Blake J. 1956. Social Structure and Fertility: An Analytic Framework. *In Economic Development and Cultural Change, Vol. 4, No. 4: 211-23*

[5] Bongaarts J. Frank O. Lesthaeghe R. 1984. The Proximate Determinants of Fertility in Sub-Saharan Africa. *In Population and Development Review, Vol. 10, No. 3*

[6] Casterline J.B. 2010. Determinants and Consequences of High Fertility: A Synopsis of the Evidence. Washington: World Bank

[7] Mahy M. 2003. Childhood Mortality in the Developing World: A Review of Evidence From the Demographic and Health Surveys, DHS Comparative Report 4. Calverton, MD: ORC Macro

[8] Rutstein S. O. 2008. Further Evidence of the Effects of Preceding Birth Intervals on Neonatal, Infant, and Under-Five-Years Mortality and Nutritional Status in Developing Countries: Evidence From the Demographic and Health Surveys, DHS Working Paper 41. Calverton, MD: ORC Macro

[9] Campbell O. and Graham W. 2006. Strategies for Reducing Maternal Mortality: Getting on With What Works

[10] Kodzi I. and Kravdal O. 2010. Implications of High Fertility in Developing Countries: A Multilevel Analysis

[11] Barro R. J. 1997. Determinants of Economic Growth: A Cross-Country Empirical Study. Massachusetts: MIT Press

[12] Coale A. J. and Hoover E. M. 1958. Population Growth and Economic Development in Low Income Countries: A Case Study of India's Prospects. Princeton: Princeton University Press.

[13] Gupta M. D., Bongaarts J., Cleland J. 2011. Population, Poverty, and Sustainable Development: A Review of the Evidence, Policy Research Working Paper 5719. World Bank

[14] Teller C. H., Hailemariam A, Gebreselassie T. 2008. The Stalled Fertility Transition in Rural Ethiopia, 1990-2005: Trends and Multivariate Analysis of Socio-economic, Health Service and Contextual Factors

[15] ECA. 2002. Determinants of Fertility Decline in Africa. Addis Ababa

[16] Ali M. 2000. The Effect of Selected Socio-Demographic Characteristics on Desire for Additional Children among Couples in Bangladesh

[17] Hayford S. R. and Agadjanian V. 2012. From Desire to Behaviour: Moderating Factors in a Fertility transition. *In Demographic Research, Vol. 26, No 20, PP. 511-542.*

[18] Bushan I. 1997. Understanding Unmet Need, The Johns Hopkins School of Public Health Working paper No. 4

[19] Kennedy E. 2010. Socio-economic Factors Impacting Fertility Preferences and Fertility Behaviours in Shanghai

[20] Feyisetan B. and Casterline, J.B. 2000. Socio-Economic Status, Fertility and Contraceptive Change in Sub-Saharan Africa. *In African Population Studies, Vol. 2, No. 15, pp. 1-24*

[21] Coombs L.C. 1984. The measurement of Family Size Preference and Subsequential Fertility. *In Demography*, Vol. 11, No. 4, pp. 587-611

[22] Redding T. M. 2007. The Population Challenge: Key to Global Survival, The Population Institute working paper no. 2

[23] Demeney P. 2003. Population Policy and Concise Summary. Population Council Working Paper No. 173

[24] __ 1993. The National Population Policy of Ethiopia. Addis Ababa

[25] APPGPDRH. 2007. Return of the Population Growth Factor: Its Impact on the Millennium Development Goals: Report of Hearing

[26] Cios K. and Kurgan L. 2005. Trends in Data Mining and Knowledge discovery. *In Advanced Techniques in Knowledge Discovery and Data Mining, pp. 1–26.* London: Springer

.[27] CSA. 2011. Ethiopia Demographic and Health Survey 2011, Preliminary report. Addis Ababa

[28] Witten I. H. and Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2rd ed. Amsterdam: Morgan Kaufmann

[29] Demena M. 2005. Population and Development. Addis Ababa: Ethiopia Public Health Training Initiative

[30] MFED.2006. Ethiopia: Population Images. Population Department, Ministry of Finance and Economic Development. Addis Ababa.

[31] Mahmood N. 1992. The Desire for Additional Children among Pakistani Women: The Determinants

[32] Pritchett L. H. 1994. Desired Fertility and the Impact of Population Policies. *In population and Development Review, Vol. 20, No. 1, pp. 1-55*

[33] Ringheim K., Teller C., Sines E. 2009. Ethiopia at a Crossroads: Demography, Gender and Development

[34] Beckman L.J, Aizenberg R., Forsythe, A.B, Day T. 1983. A Theoretical Analysis of Antecedents of Young Couples' Fertility Decision and Outcomes. *In Demography,Vol. 20, No 4, pp. 519-533*

[35] Hand D., Heikki M., H., Smyth P. 2001. Principles of Data Mining. Cambridge: The MIT Press

[36] Han J. and Kamber M. 2006. Data Mining: Concepts and Techniques, 2nd ed. Amsterdam: Morgan Kaufmann

[37] Fayyad U. Piatetsky-Shapiro G., Smyth P. 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence

[38] Vikram K. and Upadhayaya N. 2011. Data Mining Tools and Techniques: A Review. *In omputer Engineering and Intelligent Systems, Vol. 2, No. 8*

[39] Larose D. T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, NJ: Wiley & Sons

[40] Fu, Y.__ Data mining: Tasks, Techniques and Applications

[41] Ramageri B. M. ___ Data Mining Techniques and Applications. *In Indian Journal of Computer Science and Engineering Vol. 1 No. 4 pp 301-305*

[42] Maimon O. __Introduction to Knowledge Discovery in Databases

[43] Ashwinkumar,U.M. 2010. Ethical and Legal Issues for Medical Data Mining

[44] Cios K., Moore and G.W. 2002. Uniqueness of Medical Data Mining: Artificial Intelligence in Medicine

[45] Hosseinkhah F. and Ashktorab H. 2009. Challenges in Data Mining on Medical Databases

[46] Inmon W. 2002. Building the Data Warehouse. 3rd ed. New York: Wiley &

Sons

[47] Ponniah P. 2002. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professional. Wiley & Sons

[48] Sahama, T. R. and Croll, P. R. 2007. A Data Warehouse Architecture for Clinical Data Warehousing. *In Conferences in Research and Practice in Information Technology, Vol. 68*

[49] Koh H.C. and Tan G. 2005. Data Mining Applications in Healthcare. *In Journal of Healthcare Information Management, Vol.19, No.2, pp. 64-72*

[50] Biafore S. 1999. Predictive Solutions Bring more Power to Decision Makers

[51] Milley A. 2000. Healthcare and Data Mining

[52] Gupta S. Kumar D. Sharma A. 2011. Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. *In Indian Journal of Computer Science and Engineering,* Vol. 2 No. 2 1

[53] Delen, D. Walker G., Kadam A. 2004. Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods

[54] Sudhir D., Ashok A., Amol P. 2006. Neural Network Aided Breast Cancer Detection and Diagnosis

[55] Alizadehsani R and Habibi J. 2012 Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features

[56] Asha T., Natarajan S. Murthy K. N. B. ___. Data Mining Techniques in the Diagnosis of Tuberculosis

[57] Gams, M and Krivec J. 2008. Demographic Analysis of Fertility Using Data Mining Tools

[58] Saar-Tsechansky M. and Provost F. 2007. Handling Missing Values when Applying Classification Models. *In Journal of Machine Learning Research, Vol. 8, PP 1217-1250*

[59] Zurada J. and Lonial S. 2005. Comparison of the Performance of Several Data Mining Methods for Bad Debt Recovery in the Healthcare Industry

[60] Quinlan J.R. 1993. C4.5: Programs for Machine Learning. California: Morgan Kaufman

[61] Aruna S.P. and Rajagopalan L.V. 2011. Emperical Comparison of Supervised Learning Algorithms in Disease Detection

[62] Mitchell D. 2010. Decision Tree Learning

[63] Choi J.P., Han T.H., Park R.W. 2009. A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis

[64] Gasca E. 2010. Artificial Neural Networks

[65] Two Crows Corporation .2005. Introduction to Data Mining and Knowledge Discovery. 3rd ed. Potomac, USA

[66] Popescu M., Perscu-Popesu L., Valentina E. Balas V. E., Mastorakis N. 2009. Multilayer Perceptron and Neural Networks

[67] Doraisamy S. 2008. A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music

[68] Drazin S. and Montag M.___. Decision Tree Analysis using Weka Machine Learning

[69] Nayab, D. E.___. Fertility Preferences and Behaviour: A Case Study of Two Villages in Punjab, Pakistan

[70] Ibisomi, L. D. G. 2007. Analysis of Fertility Dynamics in Nigeria: Exploration into Fertility Preference Implementation

APPENDIX

Rules Generated from J48 Classifier Model

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.2 -M 2
Relation: mylast-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R9-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Remove-R17-weka.filters.unsupervised.attribute.Remove-R14-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.attribute.Remove-R8-14
Instances: 16515
Attributes: 8
Age
NumofLvgchldrn
Education
Numofsonsdied
Married
Numofdaughtersdied
Region
Class
Test mode: 10-fold cross-validation

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

J48 pruned tree

NumofLvgchldrn <= 1
| Married = 0: 1 (7183.0/931.0)
| Married = 1
| | Age = 15-19: 1 (160.0/38.0)
| | Age = 20-24: 1 (215.0/47.0)
| | Age = 25-29: 1 (183.0/53.0)
| | Age = 30-34: 1 (78.0/23.0)
| | Age = 35-39: 1 (79.0/30.0)
| | Age = 40-44: 0 (47.0/13.0)
| | Age = 45-49: 0 (41.0/7.0)
NumofLvgchldrn > 1
| NumofLvgchldrn <= 4
| | Married = 0

```
|  |  |  Age = 15-19: 1 (68.0/12.0)
|  |  |  Age = 20-24: 1 (740.0/158.0)
|  |  |  Age = 25-29: 1 (1623.0/479.0)
|  |  |  Age = 30-34: 1 (925.0/351.0)
|  |  |  Age = 35-39: 1 (595.0/251.0)
|  |  |  Age = 40-44: 0 (302.0/117.0)
|  |  |  Age = 45-49: 0 (179.0/44.0)
|  |  Married = 1: 0 (761.0/200.0)
|  NumofLvgchldrn > 4
|  |  Age = 15-19: 0 (0.0)
|  |  Age = 20-24: 1 (10.0/3.0)
|  |  Age = 25-29: 1 (261.0/106.0)
|  |  Age = 30-34
|  |  |  Region = 1
|  |  |  |  NumofLvgchldrn <= 6: 1 (61.0/24.0)
|  |  |  |  NumofLvgchldrn > 6: 0 (7.0/2.0)
|  |  |  Region = 2
|  |  |  |  Married = 0: 1 (55.0/8.0)
|  |  |  |  Married = 1: 0 (3.0)
|  |  |  Region = 3: 0 (83.0/22.0)
|  |  |  Region = 4
|  |  |  |  Numofsonsdied <= 1: 0 (95.0/29.0)
|  |  |  |  Numofsonsdied > 1: 1 (6.0/2.0)
|  |  |  Region = 5
|  |  |  |  Married = 0: 1 (64.0/16.0)
|  |  |  |  Married = 1: 0 (2.0)
|  |  |  Region = 6
|  |  |  |  NumofLvgchldrn <= 5: 1 (32.0/12.0)
|  |  |  |  NumofLvgchldrn > 5: 0 (26.0/9.0)
|  |  |  Region = 7: 1 (99.0/48.0)
|  |  |  Region = 8
|  |  |  |  Education = 0: 1 (23.0/5.0)
|  |  |  |  Education = 1: 0 (8.0/1.0)
|  |  |  |  Education = 2: 1 (1.0)
|  |  |  |  Education = 3: 1 (0.0)
|  |  |  Region = 9
|  |  |  |  NumofLvgchldrn <= 5
|  |  |  |  |  Numofsonsdied <= 0: 0 (10.0/4.0)
|  |  |  |  |  Numofsonsdied > 0: 1 (6.0/1.0)
|  |  |  |  NumofLvgchldrn > 5: 0 (14.0/1.0)
|  |  |  Region = 10: 0 (7.0)
|  |  |  Region = 11: 0 (34.0/10.0)
|  |  Age = 35-39
|  |  |  Education = 0
|  |  |  |  Region = 1: 0 (103.0/49.0)
|  |  |  |  Region = 2: 1 (80.0/17.0)
|  |  |  |  Region = 3: 0 (107.0/27.0)
|  |  |  |  Region = 4: 0 (90.0/27.0)
|  |  |  |  Region = 5
|  |  |  |  |  Married = 0: 1 (72.0/21.0)
```

```
|  |  |  |  |  Married = 1: 0 (4.0/1.0)
|  |  |  |  Region = 6
|  |  |  |  |  NumofLvgchldrn <= 5
|  |  |  |  |  |  Numofdaughtersdied <= 0: 1 (17.0/5.0)
|  |  |  |  |  |  Numofdaughtersdied > 0
|  |  |  |  |  |  |  Numofsonsdied <= 2: 0 (10.0/2.0)
|  |  |  |  |  |  |  Numofsonsdied > 2: 1 (2.0)
|  |  |  |  |  NumofLvgchldrn > 5: 0 (38.0/11.0)
|  |  |  |  Region = 7: 0 (108.0/26.0)
|  |  |  |  Region = 8: 1 (24.0/4.0)
|  |  |  |  Region = 9: 0 (30.0/6.0)
|  |  |  |  Region = 10: 0 (6.0)
|  |  |  |  Region = 11: 0 (35.0/14.0)
|  |  |  Education = 1: 0 (217.0/41.0)
|  |  |  Education = 2: 1 (5.0/2.0)
|  |  |  Education = 3: 1 (4.0/2.0)
|  |  Age = 40-44: 0 (772.0/203.0)
|  |  Age = 45-49: 0 (705.0/83.0)

Number of Leaves:     60

Size of the tree:     82
```