ADDIS ABABA UNIVERSITY SCHOOL OF INFORMATION SCIENCE GRADUATE PROGRAM

PREDICTING FERTILITY RATE IN ETHIOPIA USING DATA MINING TECHNIQUES

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science.

BY YOSEF FEKADU

OCTOBER, 2016

ADDIS ABABA UNIVERSITY
SCHOOL OF INFORMATION SCIENCE

PREDICTING FERTILITY RATE IN ETHIOPIA USING DATA MINING TECHNIQUES

BY YOSEF FEKADU

Name and Signature of Members of the Examining Board

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

Declaration

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

_____ _____

Signature Date

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

_____

Dr. Wondwossen Mulugeta

CERTIFICATE

This is to certify that Yosef Fekadu has worked on “Predicting Fertility Rate in Ethiopia. Using data mining techniques.” under my supervision. This work is original in nature and it is suitable for submission in the partial fulfillment of the requirement for the Degree of Master in Information Science.

_____

Dr. Wondwossen Mulugeta

ACKNOWLEDGEMENTS

First, I would like to a very much grateful thank to my advisor Dr. Wondwossen Mulugeta and examiners Dr. Tibebe and Dr. Million for their constructive comments and overall guidance.

Secondly, I would also like to thank Mr. Fasil Tessema, and all the coordinators and data managers in all the HDSS research centers for allowing me to carry out this research using the required data from the HDSS databases.

Table of Contents

Table of Figures

Table of tables

ABSTRACT

Introduction: Fertility rates are at a very high levels in Africa and some Arabic countries, followed next by the countries of Central and South America. Some of the social factors that can influence fertility rates are: race, level of education, religion, use of contraceptive methods, abortion, impact of immigration, etc. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases.

Objective: The main objective of this study is to apply data mining to predict fertility rate in Ethiopia, particularly for four research centers named as Arbaminch DSS, Dabat DSS, Gilgel Gibe DSS and Kilite Awelaelo DSS. This can greatly support for policy makers, planners, and healthcare providers working on the control of fertility rate in Ethiopia.

Methods and Material: The methodology used for this research was a hybrid six-step Cios Knowledge Discovery Process. The required data was collected from the data warehouse built for this purpose that stores data from four different research centers for the period of 2007 - 2015. The researcher used two popular data mining algorithms (C4 J48 Decision Trees and Naïve Bayes Classifier) to develop the predictive model using a larger dataset (68,033 cases). The researcher also used a 10-fold cross validation and 90% split test mode for data mining methods of the two predictive models for performance comparison purposes.

Results: The results indicated that the decision tree (J48 algorithm) is the best predictor with pruned parameter of the tree of 10-fold cross-validation mode; it has 76.4% accuracy on the holdout dataset (this predictive accuracy is better than any reported in the literature), Naïve Bayes Classifier came out to be the second with supervised discretization has 69% accuracy.

Conclusion: The results from this study confirmed the application of data mining for predicting fertility rate in Ethiopia. In the future, more classification studies by using a possible large amount of HDSS dataset with epidemiological information and employing other classification algorithms, tools and techniques could yield better results.

CHAPTER ONE
Introduction

1.1. Background

1.1.1. Fertility Rate

Population change can occur in one of the three major processes: Fertility, Mortality and migration. One of the ways to get the rate at which women have children is the Total Fertility Rate (TFR) or Fertility Rate in Short (Oden, and Lior, 2005).

A Fertility Rate is a measure of the average number of children a woman will have during her childbearing years. The fertility patterns are different between countries and over time. Many factors such as education and mortality can affect fertility rates. Most of these factors are difficult to measure because they involve subjectivity and some of them may not apply across cultures. This makes it especially difficult to find variables that can be used to predict future fertility rates (Oden, and Lior, 2005).

Fertility rates are still at very high levels in Africa and some Arabic countries, followed next by the countries of Central and South America. Lower rates are found in Europe and other industrialized countries like Canada and Japan. Based on the data from World Bank (for 2013), Singapore and Portugal had 1.2 and 1.3 fertility rates respectively, which were the lowest and Mali and Niger had 6.8 and 7.6 fertility rate respectively, which were the highest. Ethiopia had a fertility rate of 4.6 (World Bank, 2013)

Prediction of the population rate can enable us to forecast the future size and composition of populations. This can be used for many purposes, including predicting demand for food, water, education, medical services, labor markets, and future impact on environment.

Fertility is a key driver of the size and composition of the population. Fertility decline has been a primary determinant of population aging and projected levels of fertility have important implications on the age structure of future populations, including the pace of population aging.

1.1.2. Data Mining

The steady growth of computers and information technology helped the availability of data on different location with various formats. The abundance of data, together with the need for a powerful data analysis tools in many counties has been described as data rich but information poor society.

This fast growth and tremendous amount of data, collected and stored in large and numerous databases need a powerful tool to elicit useful information. The tool helps to get benefit from the collected data, by identifying relevant and useful information. Data mining is one of the solutions to analyze huge amount of data and turn such data into useful information and knowledge (Han and Kamber, 2006).

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns (Han and Kamber, 2006).

Data mining is the drawing out of hidden predictive data from huge databases. It entails utilizing data mining software, human creativity, and sound methodology to discover relationships, dependencies, patterns and anomalies. It incorporates various technical approaches including clustering, studying classification rules, data summarization, locating dependency networks, detecting anomalies, and evaluating change (Selam, 2011)

1.2. Statement of the problem

As data became more abundant, however, limitations in computational capabilities prevented the practical application of mathematical models. At present, not only are data available for analysis but computational resources are capable of supporting a variety of sophisticated methods. Consequently, data mining tools are now being used for health related data.

Previous investigations (Gams, 2007) were attempted to show in the areas of Fertility rate have proved the applicability of DM as well as using simple statistical method. Helen, (2011) has proved the applicability of DM technology in predicting fertility rate. The researcher has noted that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with females in childbearing age in rural communities.

The problems of previous research efforts regarding to the fertility rate were conducted not only related to the small proportion of dataset used, but also the data analysis was conducted by using simple statistical techniques (such as logistic regression and verification), applying simple algorithm, and/or lack of standard DM tools. The researcher believed there are at least 10 years differences of the data collected from the database.

Furthermore, the classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real world data and also more robust to being used by less expert users. And also because of the use of computers for closed loop business data storage and generation there now exists large quantities of data that is available to users. Likewise, the fact that computer hardware has dramatically increased ability in storing and processing the data makes some of the most powerful data mining techniques feasible today. (Alex Berson, 2013)

In this study therefore an attempt is made to demonstrate the application of different DM techniques for constructing a model for predicting fertility rate. To this end, the study attempts to explore, investigate, and answer the following main research questions.

- What are the major determinant attributes that contributes to the pattern of fertility rate?
- Which DM techniques are more appropriate to predict the pattern of fertility rate in Ethiopia?
- Is there any pattern that can be extracted from HDSS data for the prediction of fertility rate?

1.3. Objective of the study

The general and specific objectives of the research are described below.

1.3.1. General Objective

The general objective of this study is to design a predictive model using data mining techniques that can help predict fertility rate in Ethiopia.

1.3.2. Specific Objectives

To achieve the general objective, the following specific objectives are attempted in this study

- To understand the problem domain by reviewing related literatures and documents.
- To develop a data warehouse that can handle data flow from different health and demographic surveillance Systems (HDSS, henceforth) research centers in different formats.
- To clean HDSS (Health and Demographic Surveillance System) dataset by applying preprocessing tasks like cleaning, transformation and attribute selection.
- To build a model using data mining tool on cleaned HDSS data, which helps to apply data mining techniques in identifying fertility rate patterns
- To evaluate the performance of the model with domain experts and using test datasets.

1.4. Research methodology

Methodology refers to a documented approach which is used to perform activities in a manner which is coherent, consistent, accountable and repeatable. Methodology is a process that mainly consists of intellectual activities. In this section the researcher discussed about the dataset used for this research and the methodology applied, a hybrid model of Cios six-step methodology.

1.4.1. Research Design

For this study a hybrid six-step Cios KDP model is used to achieve the goal of building predictive model using data mining technique. This model was chosen on the reason that it contains all the advantages of well-known and used methodology called CRISP-DM and provides a more general, research oriented description. It also has more detailed feedback mechanisms that is helpful for

achieving the research objective. This methodology is tools independent and combines both aspects of the academic and industrial models.

Based on the hybrid model of Cios six-step methodology (see figure 1.1), the required tasks methods are identified in order to predict the fertility rate in Ethiopia. These are: Problem domain understanding, Data understanding, Data preparation, Data mining, Evaluation of the discovered knowledge and Knowledge Discovery usage.

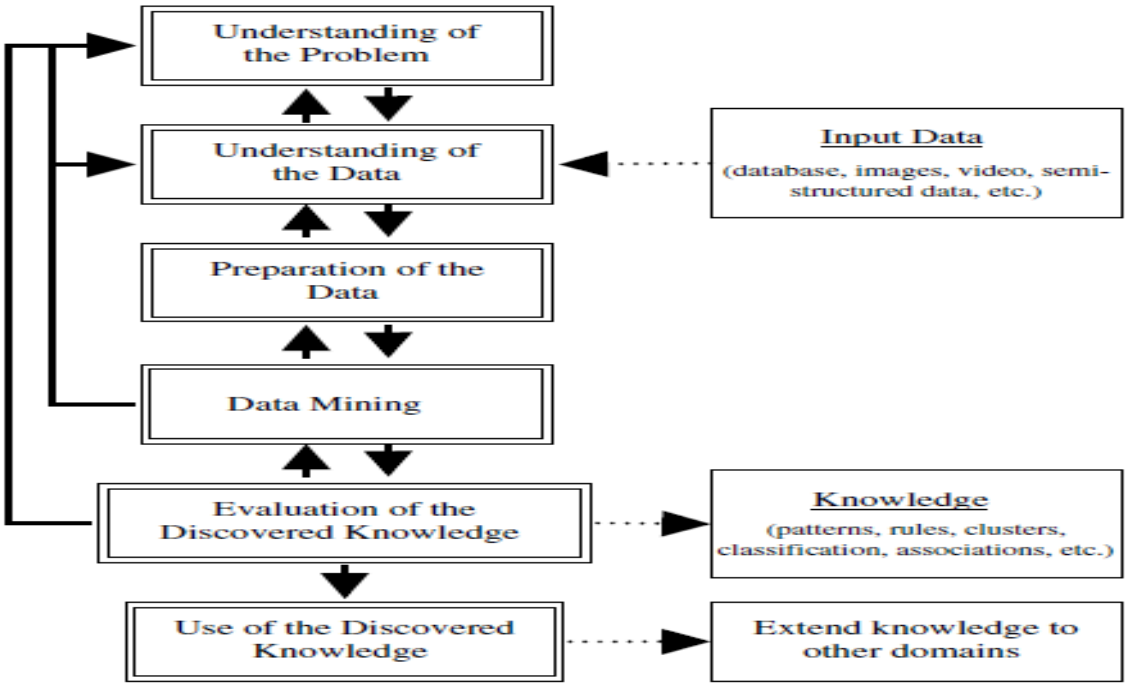


Figure 1.1. Hybrid model of Cios six-step methodology

1.4.2. Problem domain understanding

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. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed. (Cios, Pedrycz, Swiniarski, and Kurgan, 2007)

1.4.3. Data Understanding

This step includes collecting sample data and deciding which data, including format and size, are needed. Background knowledge can be used to guide these efforts. 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. (Cios, Pedrycz, Swiniarski, and Kurgan, 2007).

1.4.3.1.Dataset

There are six research centers that are working on Health and Demographic Surveillance System (HDSS) in Ethiopia. These sites are: Arba Minch HDSS – Run by Arba Minch University, Butajira HDSS – run by Addis Ababa University, Dabat HDSS – run by University of Gondar, Gilgel Gibe HDSS – run by Jimma University, Kersa HDSS – run by Haramaya University, and Kilite Awelaeelo HDSS- Run by Mekelle University.

- Gilgel Gibe Field Research Center (GGRC) Health and Demographic Surveillance System is located surrounding the Gilgel Gibe Hydroelectric dam, within four districts of Jimma Zone, Oromia Region, Southwest Ethiopia. Its global position is between latitudes 07.4253 and 07.5558oN and longitudes 037.1153 and 037.2033o E with agro climatic zone of midland. The center comprised of 11 kebeles (smallest administrative structure in Ethiopia) of which 3 are small towns. By the end of 2012, the population of the center was 57,914 of which 18,124 (31.3%) were urban and 39,790 (68.7%) rural. Children under five years of age constituted 16.6% whereas women in the reproductive age (15-49 years) were 22.8% of the population.
- The Dabat Research Center (DRC) Health and Demographic Surveillance System is located in Dabat district, Gondar. This site was initially selected purposively as a surveillance site for its unique three climatic conditions, namely Dega (high land and cold), Woina dega (mid land and temperate) and Kolla (low land and hot). The center comprised of 10 kebeles. Out of which 7 kebeles were rural and the remaining 3, urban. The first survey conducted in 2008 revealed that there were a total of 45,640 people living in the study area. Out of the total population in the study area, 35,894 (78.6 %) were from the rural area. There were 9,526 households with an average household size of 4.79 persons. About 49 % (22,378) were male and 51 % (23,262) were female, with a male to female ratio of 1:1.04.

- The Kiltie Awlaeelo HDSS includes 10 kebeles (districts) selected from Eastern zone considering agro climatic, rural/urban and other several other factors to assure representations. Nine of the study districts are rural and only one is from urban. The site is located 802km North of Addis Ababa, the capital of Ethiopia. The surveillance was started in 2009, with a baseline population of 65, 848 (urban 87.2% and 13.7% from rural) living in 14,454 households. The population comprises mainly of orthodox Christians (93.4%) and others like Moslems (1.6), catholic and protestant (0.08%). The population distribution shows it is quite youthful. 84% of the total population is of less than 45 year of age. Only 5.2% of the population is aged 65 and above years old.
- Butajirra Research and Health Program (BRHP) is located in one of the most densely populated parts of Ethiopia, Meskan, Mareko and Silti districts. The districts are part of the Southern Nations Nationalities and Peoples' Region (SNNPR). The estimated size of the Districts is 797 km2, of which Butajira town covers approximately 9 km2. The districts lie at an average of 2100 m above sea level ranging from 1750 to 3400 meters above sea level, from 1750 m in lowlands to 3400 m in mountainous areas, which are sparsely inhabited. The original DSS population in 1987 was around 28 000 and grew over 23 years to about 70 000 individuals.
- Kersa HDSS, The field site has 12 Kebeles, ten are rural and two are urban (Kersa and Weter towns). By their attitude, 2 are highland two are low land the remaining 8 are mid land. All the 12 study Kebeles have road access during the dry season. According to the 2007 census, the district has a total population of 172,626; out of which, 6.87 are urban dwellers. With an estimated area of 463.75 square kilometers

In general, about 657,000 individuals are followed in all the sites and about 80,000 households are under surveillance. But for this thesis the researcher can only manage to use the data from four HDSS research centers (gilgel gibe HDSS, Kilite Awelaelo HDSS, Arba Minch HDSS and Dabat HDSS). The reason for failing to use all the data from all the research centers is that two of the centers, namely kersa research center and butajira research centers had some problem on their data and are unwilling to give their data before they solved the problems.

The total individuals in the data warehouse from the four HDSS sites is 318,112. But after taking only females between the age of 14 and 49 the total number of records eligible for this thesis reduced to 68,033.

1.4.4. Data Preparation

This step concerns 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 (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in step 1 (Cios, Pedrycz, Swiniarski, and Kurgan, 2007)

1.4.5. Data Mining

Here the data miner uses various DM methods to derive knowledge from preprocessed data (Cios, Pedrycz, Swiniarski, and Kurgan, 2007).

1.4.6. Evaluation of the discovered knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. (Cios, Pedrycz, Swiniarski, and Kurgan, 2007)

1.4.7. Knowledge Discovery usage

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed (Cios, Pedrycz, Swiniarski, and Kurgan, 2007)

1.5. Scope and limitation of the study

This research is conducted based on the data obtained from four different centers that are operating on Health and demographic surveillance System in Ethiopia. All the four HDSS research centers are also member of INDEPTH Network, the international HDSS network. One of the goal of this study is to create a data warehouse that contains an integrated data the research centers and to mine data that holds records of fertility and to come up with a model that can predict fertility rate in Ethiopia.

Predictive modeling of data mining task is performed for the extraction of unknown knowledge and interesting patterns from different variables that have relationship with fertility.

There are a number of limitations for this research: one is that the researcher's intention at the start of the research was to create a warehouse for all the six research centers and to perform the data mining to all the data with the aim of widening the data representation. But in the process the researcher learned that two of the six research centers were in transition from one database management system to another due to failure on handling the required data. Due to this reason they were not willing to share their data before the amendment is finalized.

The other limitation the researcher observed during the data preprocessing stage was that, the HDSS data doesn't include all basic variables related to fertility rate. The likes of Housing condition of the family (how the house is built, does the house has separate bedrooms, does the house has electric light…. etc.), whether the woman under the study is using any type of contraceptive methods, any type of clinical records related to fertility, etc.

Actually in the last couple of years these kind of information were started to be collected by some of the research centers but the data were not matured enough to be used for this study.

1.6. Significance of the study

The main aim of this study is to provide a predictive model on fertility rate in Ethiopia. This is so because fertility analysis is the central importance in demographic analysis as births are a vital component of the population growth. A study in fertility also provides important information about women's reproductive behavior and attitudes. The findings in this study will

contribute to policy makers, program managers and all users understanding of changes in the fertility level.

Though demographic and surveillance data is collected and updated in different research centers in different universities in the country, little is done on analyzing the fertility data and discovering the knowledge in the dataset if demographic data. Even in these little attempts, the data analyses were conducted by using simple statistical methods. The absence of significant attempt that has been made so far to carry out research in this area using data mining technique rationalizes the importance of this research. Data mining provides automated pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods.

More over most of fertility rates of different countries mentioned in the literature review are using clinical data to determine fertility rate but in this study the researcher uses only demographic data to predict the fertility rate.

CHAPTER TWO

Literature review

In this chapter the history of data mining, its application, how different it's from similar fields such as AI and statistics, data mining processes, techniques and methodologies are discussed.

2.1. Overview of data mining

Technology now allows us to capture and store vast quantities of data. The amount of data in the world, in our lives, seems to be increasing and there's no end in sight (Witten and Frank, 2005). Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amount of data. We are overwhelmed by data - medical data, demographic data, financial data and marketing data (Jiawie, Han and Kamber, 2006). It has been estimated that the amount of data stored in the world's databases doubles every 20 months (Witten and Frank, 2005).

To undertake large data analysis project, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems (Hand, Heikki and Smyth, 2001).

Witten and Frank (2005) states that lack of data is no longer a problem at the current stage. However, the inability to generate useful information from data is the problem. As the volume of data increases inexorably, the proportion of people understands decreases, alarmingly. Lying in hidden data, in all these data potentially useful information, i.e. rarely made explicit or taken advantage of.

According to Kumar et al (2008) the health care environment is generally perceived as being 'rich in information' yet having 'knowledge poor'. There is a wealth of data available within the health care systems. Baylis (1999) states health care generates large amounts of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce large amounts of clinical data. This data is a strategic resource for health care institutions.

Finding useful patterns in data has been given a variety of names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing (Fayyad, 1996). Many People treat data mining as a synonym for another popularly used term Knowledge Discovery and Database (henceforth KDD). Others view data mining as an essential step in the process of KDD, the term data mining is becoming more popular than the term KDD (Han & Kamber, 2006).

KDD is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex datasets. Data mining is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction (Oden and Lior, 2005).

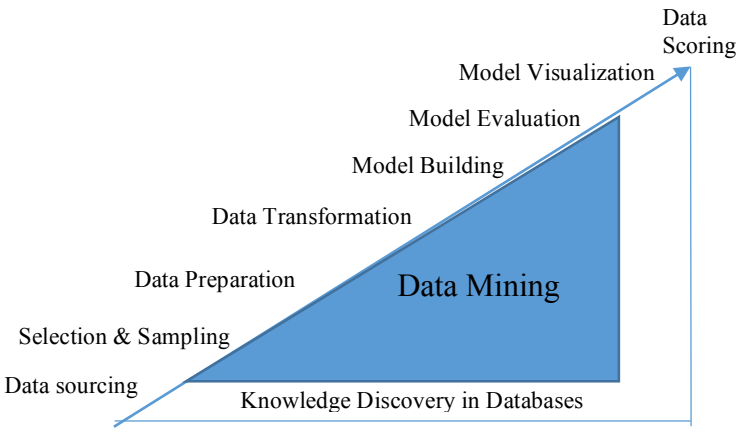


FIGURE 2-1 The relationship between data mining and Knowledge Discovery

One of the aims of data mining can be seen as the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful. This relationships and summaries derived through data mining are often referred to as models or patterns. A model is a high-level description, summarizing a large collection of data and describing its important features. The structure of the model is a global summary of a dataset. In

contrast to the global nature of models, local patterns make statements only about restricted regions of the space panned by the variables (Hand, 2001).

The patterns discovered during data mining must be meaningful in that they must lead to an economic benefit. Researchers often strive to discover the patterns that govern how physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations. (Witten & Frank, 2005).

Data mining can be defined in several ways, which differ primarily in their focus on different aspects of data mining. One of the earliest definitions of data mining is:

The non-trivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley, 1991).

To avoid confusion and mix-up its necessary here to distinguish data mining from the previous activities of statistical modeling and knowledge discovery

- **Statistical Modeling**: the use of parametric statistical algorithms to group or predict an outcome or event, based on predicator variables.
- **Data Mining**: the use of machine learning algorithms to find faint patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit in some form.
- **Knowledge discovery**: the entire process of data access, data exploration, data preparation, modeling, model deployment, and model monitoring.

2.2. Data Mining and Knowledge Discovery in a database

KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, Artificial Intelligence, knowledge acquisition for expert systems, data visualizations, and high-performance computing (Cios, Pedrycz, Swiniarski, and Kurgan, 2007). The main goal here is extracting high-level knowledge from low-level data in the context of large datasets. the data mining, as a component of KDD, uses combined techniques from machine learning, pattern recognition, and statistics to find patterns.

KDD focuses on the overall process of knowledge discovery from data including how the data are stored and accessed, how algorithms can be scaled to massive datasets still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported (Sciences Applied, 2010).

A driving force behind KDD is the database field. Indeed, the problem of effective data manipulation when data cannot fit in the main memory is of fundamental importance of KDD. Database techniques for gaining efficient data access, grouping and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger datasets. A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps to set the stage for KDD. (Sciences Applied, 2010).

Fayyad (1996) defined data mining as a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumerations of patterns over the data. According to this definition data mining is the step that is concerned with the actual extraction of knowledge from data. To emphasize the necessity that data mining algorithms need to process large amounts of data, the desired patterns has to be found under acceptable computational efficiency limitations.

Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (Fayyad, 1996). Based on this definition, data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. This means that the objective of the data mining exercise plays no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as "secondary" data analysis.

2.3. Data Mining and Data Warehouse

In predictive models, the values or classes as researchers are predicting are called the response, dependent or target variables. The values used to make the prediction are called the predictor or independent variables. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. On the other hand, descriptive techniques are sometimes referred to as unsupervised learning because there is no already known result to guide the algorithms (Two Cows Corporation, 2005).

The relevance of the field of databases to KDD is obvious from the name. Databases provide the necessary infrastructure to store, access, and manipulate the raw data. With parallel and distributed database management systems, they provide the essential layers to insulate the analysis for the extensive details of how the data is stored and retrieved. The focus here is only on the aspects of database research relevant to the DM step. A strongly related term is online analytical processing (henceforth OLAP), which mainly concerns providing new ways of manipulating and analyzing data using multidimensional methods. This has been primarily driven by the need to overcome limitations posed by SQL and relational DBMS schemes for storing and accessing data (Sumathi, Sivanadam, 2006).

Supporting operations from the DM perspective has an emerging research area in the database community. In the DM step itself, new approaches for functional dependency analysis and efficient methods for finding association rules directly from databases have emerged and are starting to appear as products. In addition, classical database techniques for query optimization and new object-oriented databases make the task of searching for patterns in databases much more reasonable (Sumathi, and Sivanadam, 2006).

Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involve data cleaning, data integration and data transformation and can be viewed as an important preprocessing step for DM. Hence, the data warehouse has become an increasingly

important platform for data analysis and OLAP and will provide an effective platform for DM (Berry, Linoff, 1997). Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today's competitive, fast-evolving world. Many people feel that with competition mounting in every industry, data warehousing is the latest must-have marketing weapon-a way to retain customers by learning more about their needs.

Moreover, data warehouses have been defined in many ways, making it difficult to formulate a rigorous definition. Loosely speaking, a data warehouse refers to a database that is maintained separately from an organization's operational databases. Data warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis (Han and Kamber, 2006).

According to Immon (1997), a data warehouse is "an integrated collection of data about a collection of subjects (units), which is not volatile in time and can support decision taken by the management". The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

In sum, a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making (Han and Kamber, 2006).

2.4. Data Mining Process

In the previous section the researcher had explained the difference between DM and KDD. Before presenting the state of the art of the existing processes, we mention that in the scientific literature there is a lot of confusion between the terms "process" and "methodology". A process is represented by a sequence of steps executed in order to produce a certain result. A methodology is

defined as an instance of a process, by specifying the tasks that should be executed, the inputs, the outputs and the way the tasks should be executed. In brief, a process gives the user the tasks that should be executed and a methodology tells the user also "how to" perform those tasks (Kaur and Krishan, 2006).

As its explained well before, DM is one among the most important steps in the KDD process. Most writers considered it even as the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the dataset (Kaur and Krishan, 2006).

Nowadays, with the explosion of information, DM has become one of the top ten emerging technologies that will change the world (Piatetsky-Shapiro and Frawley, 1991). There are two basic styles of DM: hypothesis testing and KD (Knowledge Discovery). Hypothesis testing is a top-down approach that is used when a confirmation or a rejection of an already defined hypothesis is needed. The other style is KD (relevant for this research study). It is a bottom-up approach and it is used when we want to find something that we do not know searching available data. It can be directed or undirected. There is no target field in undirected knowledge discovery. Instead, what the researcher wants from a computer is to recognize the schemes within the data that are of some importance (Glasnik, 2008).

DM is a rather complicated process that has to be planned very carefully in order to be successful. It has to be organized within one of the proposed rigorous procedures (Berry and Linoff, 1997). According to Sumathi and Saarenvitra (2006), “once a data warehouse has been developed, the DM process falls into four basic steps: data selection, data transformation, DM, and result interpretation”.

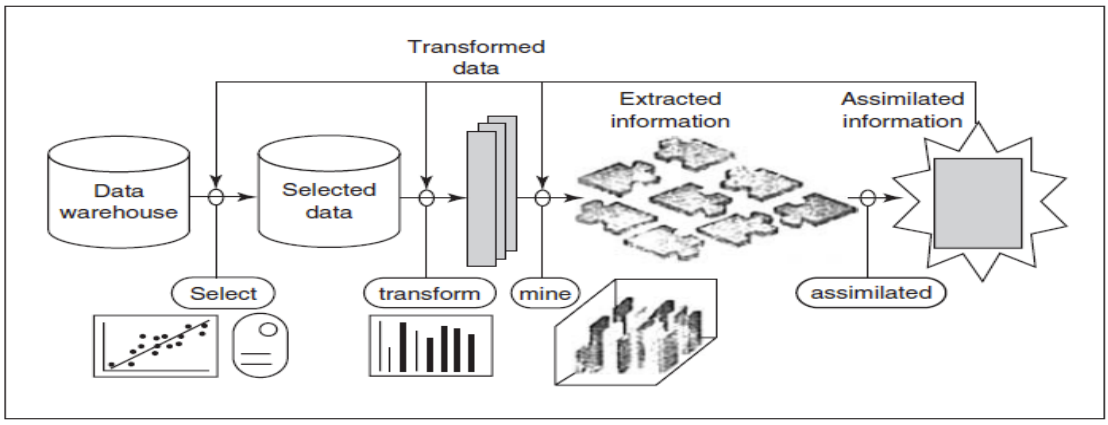


Figure 2.2: Data Mining Process (Cisos, 2007)

The DM process is iterative and interactive. The process is iterative, which means that sometimes it may be necessary to repeat the previous steps. The problem with this process, as with all the existing processes for DM, is the lack of user guidance.

2.4.1. Identifying the Target Dataset (Selection)

The first step in the DM process is to select the target data. The selected data types may be organized along multiple tables: during data selection, the user might need to perform table joints. Furthermore, even after selecting the desired database tables, mining the contents of the entire table is not always necessary for identifying useful information. Under certain conditions and for certain types of DM operations (such as when creating a classification or regression model), it is usually a less expensive operation to sample the appropriate table, which might have been created by joining other tables, and then mine only the sample.

2.4.2. Preparing the Data for Analysis (Transformation)

After selecting the desired database tables and identifying the data to be mined, the user typically needs to perform certain transformations on the data. Three considerations dictate which transformation to use: the task (mailing-list creation, for example), the DM operations (such as predictive modeling), and the DM technique (such as decision trees) involved. Transformation methods include organizing data in desired ways (organization of individual client data by household), and converting one type of data to another (changing numeric values into nominal

ones, so, that they can be processed by a decision tree). Another transformation type, the definition of new attributes (derived attributes), involves applying mathematical or logical operators on the values of one or more data base attributes-for example, by defining the ratio of two attributes (Theeuwen, 2001).

2.4.3. Building and Testing the Model (Data Mining)

The user consequently mines the transformed data using one or more techniques to extract the desired type of information. For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulations manager might need to first use clustering to segment the subscriber database, and then apply rule induction to automatically create a classification model for each desired cluster. The problem of overfitting is another issue that deserves a due consideration at this step. According to Mitchell (1997), "over fitting is an attempt to create overly complex DM model that fits noise in the training data or unrepresentative features of the particular training data that decreases the generalization accuracy of the model over other unseen instances".

2.4.4. Evaluating the Model (Result Interpretation)

The user must finally analyze the mined information according to his decision-support task and goals. Such analysis identifies the best of the information. For example, if a classification model has been developed, during result interpretation, the DM application will test the model's robustness, using established error-estimation methods such as cross validation. During this step, the user must also determine how best to present the selected mining-operation results to the decision maker, who will apply them in taking specific actions. For example, the user might decide that the best way to present the classification model is logically in the form of if-then rules.

Three observations emerge from this four-step process (Sumathi and Sivanadam, 2006):

- Mining is only one step in the overall process.
- The process is not linear but involves a variety of feedback loops.
- Visualization plays an important role in the various steps.

2.5. Data Mining Techniques

According to Berry and Linoff (1997), having an in depth knowledge and understanding of different data mining techniques is indispensable for the following reasons:

- In order to make use of and take the advantage of a specific technique, it is important to know the details of each technique.
- To find out the best applicable technique for the problem at hand.
- To know the advantages and disadvantages of a technique.

It is evident that no one technique is applicably suited to all data mining problems. Determining the best technique that fits to the specific data mining problem and familiarizing with the available techniques is extremely essential. According to Thearling (2003), the most commonly used data mining techniques are: Decision tree, neural networks, genetic algorithms, nearest neighbor method and rule induction.

As Levin and Zahavi (1999) stated, data mining techniques can be categorized into two major application groups: **Predictive modeling** and **descriptive modeling**. In each of these applications, data mining differs in the approach taken to solve problems. Each application is usually geared in solving a particular type of problem. That is, a specific algorithm is favored over others depending what the problem posed by the data miner.

In predictive modeling tasks, one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models such as classification, regression and other AI-based models. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results.

On the other hand, descriptive models belong to the realm of unsupervised learning; it is called unsupervised learning since there are no already known results to guide the algorithm. Such models interrogate the database to identify patterns and relationships in the data. Clustering, segmentation

and visualization methods, among others, belong to this family of descriptive models (NaZ & Jacob, 1999).

As Han and Kamber (2001) stated, in this unsupervised learning users may sometimes have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus, it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations of applications.

2.5.1. Predictive modeling

Prediction is arguably the strongest goal of data mining. The aim is to build a model that can permit the value of one variable to be predicted from the known values of other variables. Classification and regression are two good examples of data analysis that can be used to extract models describing important data classes or to predict future data trends (Han JaK & Micheline, 2001).

Classification problems, as stated by Deogan (2011), aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. However, the quality of the discovered knowledge is heavily dependent on the algorithms used to analyze the data.

Classification methods create classes by examining already classified cases and inductively finding the pattern (or rule) typical to each class. Data mining uses machine–learning methods such as decision trees, neural networks and Bayes to classify objects based on a dependent variable.

According to Han and Kamber (2001) data classification is a two-step process. In the first step, a model is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step the model is used for

classification. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set is evaluated.

As Rea (2002) stated, once classes are defined the system should infer rules that govern the classification and therefore should be able to find the description of each class. The writer further argued that the descriptions should only refer to the predicting attributes of the training set so that the positive examples should satisfy the description and none of the negative. A rule is said to be correct if its description covers all the positive examples and none of the negative examples of a class. Basically, in classification tasks, the system, given a case or tuple with certain known attribute values should be able to predict to which class this case belongs to. Although the choice of techniques suitable for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are the Decision trees, Bayes and Neural networks (Plate, 1997). The decision trees and the neural networks are discussed below.

2.5.1.1. Decision tree

One of the most commonly used data mining techniques for classification tasks are decision trees. Decision trees are simple knowledge representation and they classify examples to a finite number of classes. In decision tree induction, the nodes of the tree are labeled with attribute names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generates decision tree from a given set of attribute-value tuples. Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that can be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest, and C5.0 (Two Crows Corporation, 1999).

A decision tree is constructed by repeatedly causing a tree construction algorithm in each generated node of the tree (Larvac & Nada, 1998). The classification is performed separately for each leaf, which represents a conjunction of attribute valued in a rule (Last, Maimon, Oded, & Kandel

Abraham, 2002).

Structurally, decision trees consist of two types of nodes; non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features of the diagnosed case. Terminal nodes, on the other hand generate a decision (Rudolfer, Georgios, Peers, & Ian, 2002).

A typical decision tree is shown in Figure 2.2 below.

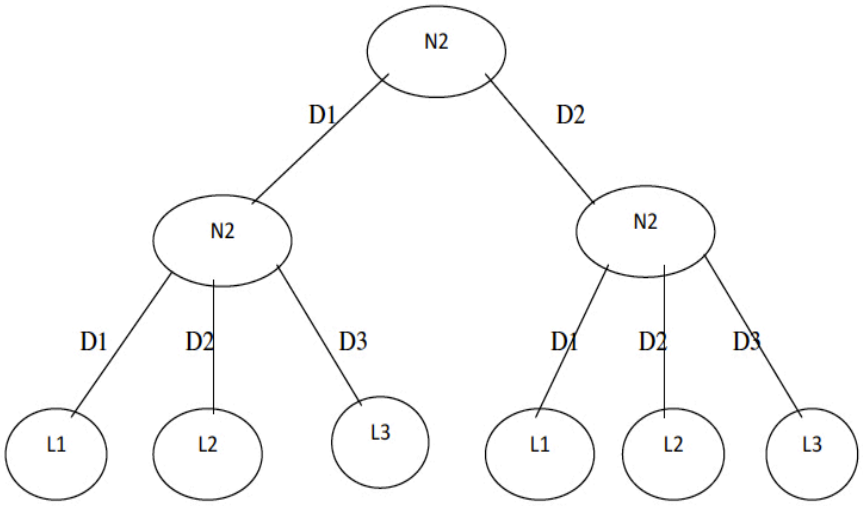


Figure 2.3: A Decision Tree with Decision (Ni) and Leaf (Li) nodes, and decisions (Di)

The very advantage of decision trees is that they can handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets (Two Crows Corporation, 1999).

The commonly agreed drawbacks for which the decision tress are criticized is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In addition, all splits are made sequentially, so each split is dependent on its predecessor (Two Crows Corporation, 1999).

2.5.1.2. Neural networks

The leading models in the AI based knowledge discovery are Neural Networks (henceforth NN) models. NN is a biologically inspired model which tries to mimic the performance of the network or neurons, or nerve cells, in the human brain. Expressed mathematically, a NN model is made up of a collection of processing units (neurons, nodes), connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. A typical NN contains several input nodes connected to one or more output nodes, through an intermediate set of hidden nodes (Frohlich, 1999).

The structure of neural network is very similar to the structure of the neurons in the human brain. All of the processing of a neural network is carried out by this set of neurons or units. Each neuron is a separate communication device, doing its own relatively simple job. A unit's function is simply to receive input from other units and, as a function of the inputs it receives, to compute an output value, which it sends to other units. The system is inherently parallel in that many units can carry out their computations at the same time (Frohlich, 1999).

According to Frohlich (1999), in neural networks, neurons are grouped in layers, often classified as input, hidden, and output layers. Inputs layer is a processing element that receives the input to the neural network and hidden layers are processing elements between a neural network's input layer and its output layer. On the other hand, output layer is the processing element that produces neural network's output.

There could be a number of input, hidden and output neurons in each corresponding layer. For example, in the following neural network (figure 2.3), there are three inputs, three hidden and three output neurons. In fact, the network consists of one input, hidden and output layer.

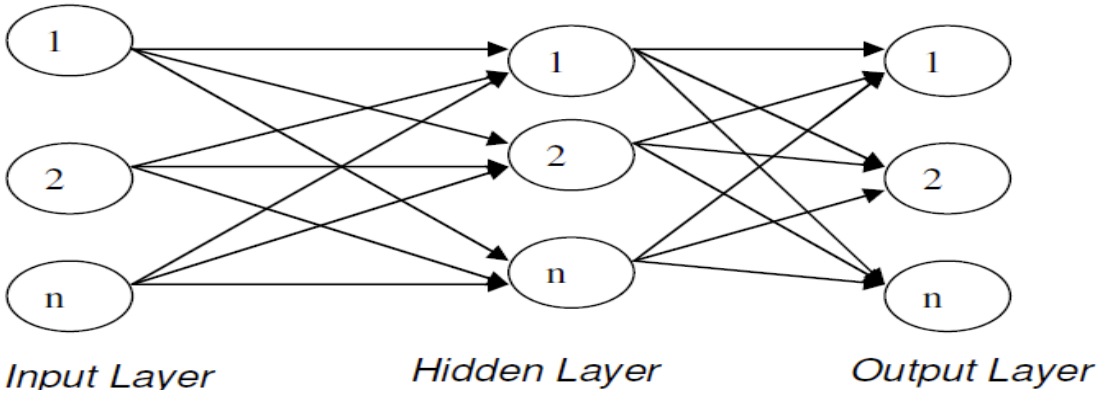


Figure 2.4: A simple neural network

NN have become of particular interest in data mining because they offer a means for efficiently modeling large and complex problems in which there are hundreds of independent variables that have many interactions (Frohlich, 1999).

2.5.2. Descriptive modeling

Fayyad et. al. (2010) states that Description involves using some variables or fields in the database and focuses on finding human-interpretable patterns describing the data. The well-known descriptive modeling in data modeling are clustering (or segmentation) algorithm and association rule discovery.

2.5.2.1. **Clustering**

Han and Kamber (2001) stated the process of grouping a set of physical or abstract objects into class of similar objects is called clustering. Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It is mapping a data item into one of several clusters which are not pre-specified but are determined from the data. Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures.

Two Crows Corporation (2011) mentioned that the goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. The categories

(clusters) can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories.

According to Han and Kamber (2001), each cluster that is formed can be viewed as a class of objects from which rules can be derived. As Rea (2013) points out there are a number of approaches for mining clusters. They are in general categorized in to partitioning and hierarchical clustering.

The partitioning clustering algorithms group data into un-nested and non-overlapping groups that usually optimize a clustering (Arzucan, 2004). Levin and Zahavi (1999) argues that perhaps the most common of all automatic partitioning clustering is the K-means algorithm, which assigns observations to one of K classes to minimize the within-cluster-sums-of-squares. Also worth mentioning are the Judgmental-based or manual segmentation methods which are still very popular in direct marketing applications to carve up a customers' list into homogeneous segments.

Hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering (Arzucan, 2004). In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on criteria. The merging can be done by using single link, complete link, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are split into number of clusters based on certain criteria, and this is called as top down approach.

2.5.2.2. Association rule discovery

According to Feyen (2011) discovery is the process of looking in a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be. In other words, the program takes the initiative in finding what the interesting patterns are, without the user thinking of the relevant questions first. In large databases, there are so many patterns that the user can never practically think of the right questions to ask. The key issue here is the richness of the patterns that can be expressed and discovered and the quality of the information delivered. This in turn

determines the power and usefulness of the discovery technique.

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. The rules are given in the form: if item A is part of an event, then X% of the time item B is also part of the event. The rules are written as A_B, where A is called the antecedent or left-hand side (LHS), and B is called the consequent or right hand side (RHS). More formally, association rules are of the form A_B, that is, (A1, - - - , Am _ B1, - - - - - Bn) Where Ai (for i =1, - - - m} and Bj (for j = 1, - - - - n } are attribute-value pairs.

The associations rule A_B is interpreted as database tuples that satisfy the condition in A are also likely to satisfy the condition in B.

Support and confidence are the probability measures, introduced to assess associations in the database. The support (or prevalence) of a rule is the proportion of observations that contain the item or item set of the rule. It is also known as the coverage of the rule. As stated by Witten and Frank (Witten & Eibe, 2000), an item is an attribute value pair. The confidence is the conditional probability of B given A, P (B/A). A rule is "interesting" if the conditional probability P (B/A) is significantly different than P (B). Confidence of the rule measures the rule's accuracy.

Association algorithms find these rules by doing the equivalent of sorting the data while counting occurrences so that they can calculate confidence and support. The efficiency with which they can do this is one of the differentiators among algorithms. One should be able to evaluate rules using different techniques especially because of the combinatorial explosion that results in enormous number of rules (Witten & Eibe, 2000).

As written by Han and Kamber (2001), association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Rules that satisfy both a minimum support threshold and a minimum confidence threshold are strong rules. A good example of the use of associations is the analysis of the claim forms submitted by patients to a medical insurance company. By defining the set of items to be the collection of all medical procedures that can be performed on a patient and the records to correspond to each claim form, the application

can find, using the association function, relationships among medical procedures that are often performed together (WHO, 2007).

In specific situations the association discovery components can help one to:

- Manage existing customers. Determine response propensities by segmenting customers on purchase patterns and attributes.
- Use knowledge of customer segment attributes to recommend items or actions that might appeal to each segment.
- Acquire new customers. Analyze purchase pattern and attribute data from an outside source to develop customer segmentation models. Then "acquire" new customers whose characteristics resemble those of your best customers by offering them targeted products and services.
- Detect patterns of potentially harmful behavior. Detect patterns of events or behavior that can help identify the potential of bioterrorist attacks and infrastructure intrusions.
- Spot fraud, waste and abuse. Detect patterns of fraudulent and abusive behavior so you can take steps to prevent future occurrences.
- Improve Web site navigation. Make it easier for people to make Web-based purchases by enhancing site navigation and how items are presented. Medical diagnosis/research.
- Identify telltale symptoms to aid in effective diagnosis.

2.6. Data Mining Methodologies

The ultimate goal of the KD Process (henceforth KDP) model is to achieve overall integration of the entire process with industrial standards. Another important objective is to provide interoperability and compatibility between the different software systems and platforms used throughout the process. Integrated and interoperable models would serve the end user in automating, or more realistically semi-automating, work with KD systems. The efforts to establish a KDP model were initiated in academia.

Although, the models usually emphasize independence from specific applications and tools, they

can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. this discussion is restricted to those models that have been popularized in the literature and have been used in real KD projects which is proper to the researcher's business domain.

2.6.1. Knowledge Discovery in Database (KDD)

As mentioned earlier, KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular DM method. While the DM step is characterized by the extraction of patterns hidden in the data, the whole KDD process is broader and includes all the processing (data selection, preprocessing and transformation) that is needed for this to occur, making it possible to evaluate and interpret the results that were obtained after the DM techniques were used. The KDD process is a set of continuous activities that include five steps: Data Selection, Preprocessing, Formatting, Data Mining and Interpretation (Cios, K. et. al. (2000)).

Moreover, the process starts by understanding the application's domain and the targets that must be reached. Then, a selection can be drawn from these data so that, one may work with the data that are of interest. The pre-processing step is the one in which missing or inconsistent data are analyzed and treated. During the formatting step data are prepared. So, DM can be used as, for instance, to map categorical data among numerical data or using methods to reduce dimensions in the data (Fayyad, et. al. (1996a)). According to Silver (2008), "preprocessing and formatting may take up to 80% of the time needed for the whole process".

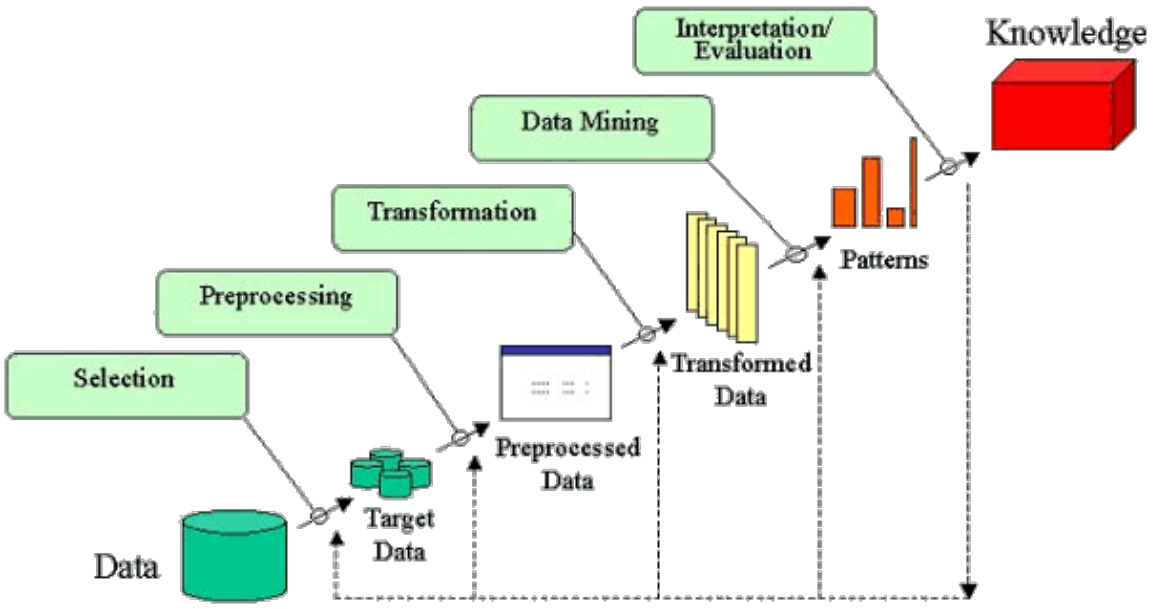


Figure 2.5: KDD Process model (Fayyad, (1996)

The main advantage of the knowledge discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge. KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular DM method. The KDD process is a set of continuous activities that include five steps: Data Selection, Pre-processing, Formatting, DM, and Interpretation. The main purpose with the KDD process is to obtain knowledge hidden in data that may be useful for decision-making, by using methods, algorithms, and techniques from different scientific areas. According to Tan et. al. (2008), these include "Statistics, Artificial Intelligence, Machine Learning, and Pattern Recognition".

The Fayyad, et. al. (1996a) KDD process consists of the following five steps:

- Select a target data set: The initial step is based on data needed for the DM process may be obtained from many different and heterogeneous data sources.
- Data preprocessing: In this step the data to be used by the process may have incorrect or missing data. There may be abnormal data from multiple sources involving different data types and metrics.
- Data transformation: Attributes and instances are added and/or eliminated from the target data. Data from different sources must be converted into a common format for processing.

- Data mining: A best model for representing the data is created by applying one or more DM algorithms.
- Interpretation/evaluation: The final step the researcher examines the output from step 4 to determine if what has been discovered is both useful and interesting.

Another important step not contained in the KDD process is goal identification. The focus of this step is on understanding the domain being considered for KD.

2.6.2. CRoss Industry Standard Process for Data Mining (CRISP-DM)

The goal of the project was to define and validate an industry- and tool-neutral DM process model that which would make the development of large as well as small DM projects faster, cheaper, more reliable and more manageable.

CRISP-DM is a process model because it is the "de facto standard" for developing DM and KD projects. In addition, CRISP-DM is the most used methodology for developing DM projects. Analyzing the problems of DM and KD projects, a group of prominent enterprises developing DM projects, proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM or independent so it can be used with any DM tool and it can be applied to solve any DM problem. CRISP-DM defines the phases to be carried out in a DM project as well as defines for each phase the tasks and the deliverables for each task (Julio and Adem, 2009; Julio and Adem, 2007)

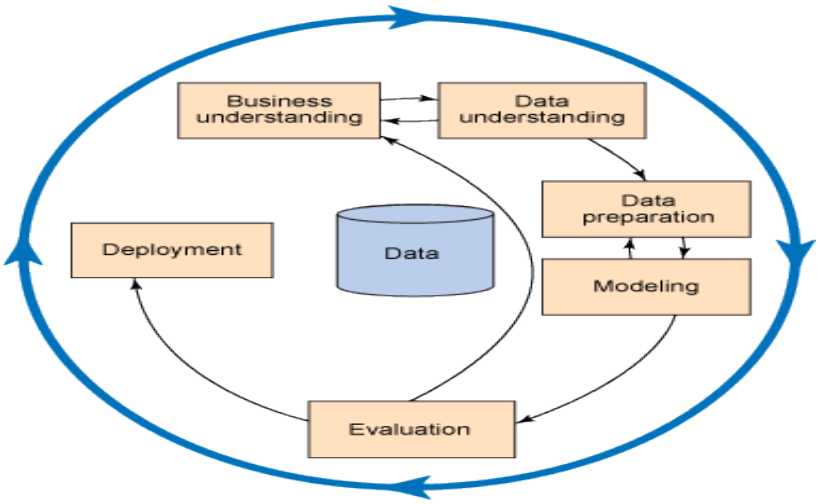


Figure 2.6: CRISP-DM Process model (IBM SPSS Modeler CRISP-DM Guide)

- Business understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- Data understanding: This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- Data preparation: The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type.
- Evaluation: Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives.
- Deployment: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.6.3. SEMMA

The methodology and approach that SAS Institute proposes is referred to as SEMMA, for Sample, Explore, Modify, Model, and Assess. Beginning with a statistically representative sample of data, users can apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and affirm the model's accuracy (Fayyad, et. al., 1996).

- Sample: The first step is to extract a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly.
- Explore: This phase involves searching speculatively for unanticipated trends and anomalies so as to gain understanding and ideas.
- Modify: The insights that are gained from the exploration phase enable knowledge workers to group the most productive subsets and clusters of data together for further analysis and exploration.
- Model: This process involves searching automatically for a variable combination that reliably predicts a desired outcome.
- Assess: During this evaluation process, assessment of the results gained from modeling provides indications as to which results should be conveyed to senior management, how to model new questions that have been raised by the previous results and thus, proceed back to the exploration phase.

2.6.4. Hybrid Model

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model developed by Cios et. al. It was developed based on the CRISP-DM model by adopting it to academic research (Cios, 2000; Julio and Adem, 2009).

A hybrid of the above-mentioned approaches may be considered in determining a suitable goal for DM. All KDD model process models emphasize the iterative nature of the process that a DM application is conducted. Typically, goals are selected, an experiment is conducted, based on

results at each stage, a step is revisited or moves to the next step (Julio and Adem, 2009).

A description of the six steps depicted in figure 2.7 as follows:

- Understanding of the 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.
- Understanding of the data: This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts.
- Preparation of the data: This step concerns deciding which data will be used as input for DM methods in the subsequent step.
- Data mining: Here the data miner uses various DM methods to derive knowledge from preprocessed data.
- Evaluation of the discovered knowledge: Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.
- Use of the discovered knowledge: This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains.

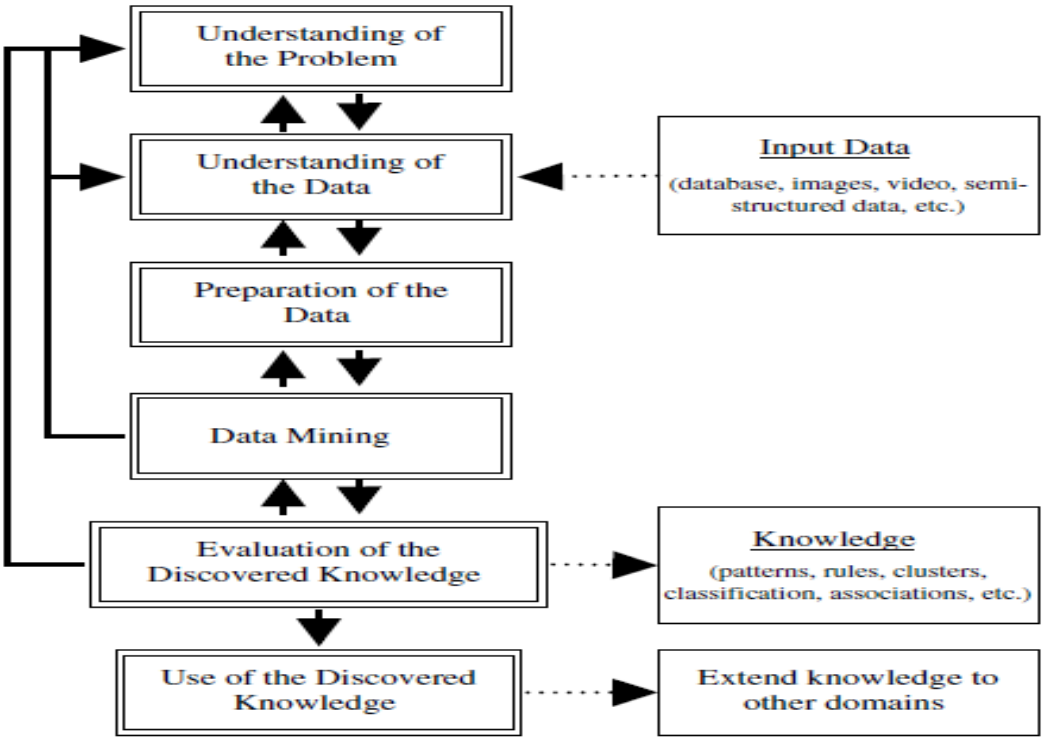


Figure 2.7: Hybrid Process model

In this study the investigator selected the hybrid process model to predict the pattern of fertility rate in Ethiopia, since the model emphasizes the iterative aspects of the process, drawing from the experience of users of previous models. It identifies and describes several explicit feedback loops.

2.7. Overview of fertility rate

Populations change through three major processes: fertility, mortality, and migration. A useful way to express the rate at which women have children is the Total Fertility Rate (TFR). TFR is the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given set of age-specific fertility rates. If the average woman has approximately 2 children in her lifetime, this is just enough to maintain the population (Gams, 2007).

A fertility rate is a measure of the average number of children a woman will have during her childbearing years. The fertility patterns are different between countries and over time. Many different factors can affect fertility rates. Many of these factors are difficult to measure because

they involve subjectivity and some of them may not apply across cultures. This makes it especially difficult to find variables that can be used to predict future fertility rates. (Norville, Gomez & Brown, 2005)

During the peak years of the baby boom (in the late 1950's), the fertility rate was 3.91 in Canada and 3.77 in the U.S. By the 1970's it had fallen below 2 in both countries. While the U.S. rate went back up to 2.08, Canada's rate kept falling to 1.52 in 1999. Such differences in fertility exist even in countries as similar as Canada and the U.S. Many developed countries in Europe are also below the replacement level of 2.1, but the U.S. is still near this rate. This is a huge gap, which raises the question: "Why this big difference?"

Fertility rates are still at very high levels in Africa and some Arabic countries, followed next by the countries of Central and South America. Lower rates are found in Europe and other industrialized countries like Canada and Japan. Characteristics that influence changes in fertility rates are related either with the economic situation or with social patterns. Some of the social factors that can influence fertility rates are: race, level of education, religion, use of contraceptive methods, abortion, impact of immigration, children as a source of labor (on family farms), children as support for couples at older ages, costs of raising children, female labor force participation, government programs to encourage or discourage childbearing, postponement of marriage, age of first birth and divorce rates (Norville, Gomez & Brown, 2005).

2.8. Review of demographic trends

There has been a shift in behavior in many societies. This can be seen in many factors such as: postponement of marriage, increasing age of first birth, increasing divorce rates, lower marriage rates, more births outside marriage, an increasing number of women in the labor force, greater levels of education for women, a decreasing need for children to support elderly parents, a shift from rural to urban societies and government programs to encourage or discourage having children. Together with these factors, general mortality rates have declined, leading to improvements in life expectancy which continue in most countries. Also, many advances in medical technologies are being realized including improvements in birth control methods and progress in the cure or successful treatment of many diseases.

Demographic changes have long been recognized as a critical component of economic development, although there is considerable debate about whether population growth helps or hinders development, and whether governments can substantially alter population growth rates. With regard to the first of these debates, Malthusian arguments that population growth induces food supply constraints have been viewed with skepticism by most economists, given the possibility of population-induced technological change in both agriculture (Boserup 1965) and industry (Henderson 2010), with the latter capable of financing food imports. More recently, David Bloom and colleagues (Bloom et al. 2007; Bloom and Williamson 1998) have argued that the change in age structures matters more than population growth per se, with the transition from low to high age dependency ratios inducing higher savings rates and greater investments in education. In successful Asian countries, this demographic "window of opportunity"—chiefly induced by lower fertility rates—explains 25-40 percent of the region's miraculous economic growth. High population growth has also been associated with increased poverty (Eastwood and Lipton 2004) and increased malnutrition (Headey 2011). However, what is not clear is the extent to which direct population policies substantially influence fertility rates. Economists view fertility rates as a demand-led factor heavily influenced by income levels, livelihoods, and education factors. Efforts to directly influence fertility rates (through contraception, family planning or more draconian measures) may therefore be less important than these indirect factors. Indeed, even in the Chinese context, the effectiveness of the country's one-child policy is heavily debated among demographers.

In Ethiopia these debates are highly relevant. Although the country is undergoing rapid economic growth (albeit from a low base), Ethiopia has a long history of Malthusian population dynamics (Pankhurst 1985).

In the middle ages, rapid population growth and the resulting deforestation of the densely populated highlands contributed to the collapse of the agricultural bases of several early empires. In more recent times Ethiopia has sadly become notorious for some of the worse famines of the 20th century. And although there are large tracts of mostly uninhabited land in Ethiopia's lowland peripheries, the population-dense highlands face shrinking farm sizes and major problems of deforestation and soil degradation (Ringheim, Teller, and Sines 2009; Yusuf et al. 2005).

These problems are clearly related to population growth rates that are too high relative to the local resource base and to traditional farming practices. Indeed, the total population is estimated to be growing at 2.6 percent per year, chiefly as a result of high fertility in rural areas (CSA 2008). Moreover, the worldwide Demographic and Health Surveys (2010) data suggest that Ethiopia not only has one of the world's highest fertility rates (at 5.4 children in 2005), the country also has the world's largest rural-urban fertility differential: the projection in 2005 was that an average rural Ethiopian woman was expected to give birth to 6 children in her lifetime, relative to just 2.4 children in the country's urban areas.

2.9. Related works

As the researcher reviews a number of literatures, the closest to this thesis is an attempt made by Be'ewnetu Tekabe on "Predicting pattern of under-five mortality in Ethiopia using data mining technology: the case of Butajira rural health program" June 2012. In this study he tried to use different data mining tools, such as naïve bayes and J48 to predict under-five children mortality in butajira HDSS study area. The methodology used for this research was a hybrid six-step Cios Knowledge Discovery Process. The required data was collected from butajira rural health program project.

Another attempt was made by matjaz Gams and Jana Crivec on "Demographic analysis of Fertility using Data Mining Tools". In this study they were trying to use data mining techniques to discover which attributes have the highest impact on country fertility rates.

Predicting Low Birth weight using data mining techniques on Ethiopia demographic and health survey datasets" by Biset Desalegn, June 2011.

"Predicting the occurrence of measles outbreak in Ethiopia using data mining technology" by Selam Assmamaw, July 2011.

CHAPTER THREE

Methods for Model building

As the researcher already expressed in the methodology section in chapter one, the focus of this thesis is on realizing predictive models from knowledge of classification. The aim is to model connections between input, or predictor variables, and the outcome or prediction using observed classification. Hence, it is important to explain the classification implementation for model building and experiments were carried out in the DM process, which also involve DM tool selection and algorithms were used for modeling. The DM algorithms used in this research to predict the pattern of fertility rate in Ethiopia based on an integrated dataset from four HDSS (Health and Demographic Surveillance System) sites were J48, decision tree algorithm and Naïve Bayes classifier.

There are a number of machine intelligent tools, techniques and algorithms that are available in the market but at the same time not all tools, techniques and algorithms are the best for all problems in the dataset. Different data sets will produce different results based on the algorithms used. In this thesis the researcher tested some algorithms based on decision trees, rule based classification and Naïve Bayes classifier probability. The researcher's aim was to find the best tool, techniques, and algorithms to predict the pattern of fertility in Ethiopia that based on the available data in the HDSS dataset.

3.1. The WEKA Tool

WEKA was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems-and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset. WEKA is a collection of machine learning algorithms for solving real-world DM problems. It contains 41 different algorithms for classification and numeric prediction (Ian and Eibe, 2005).

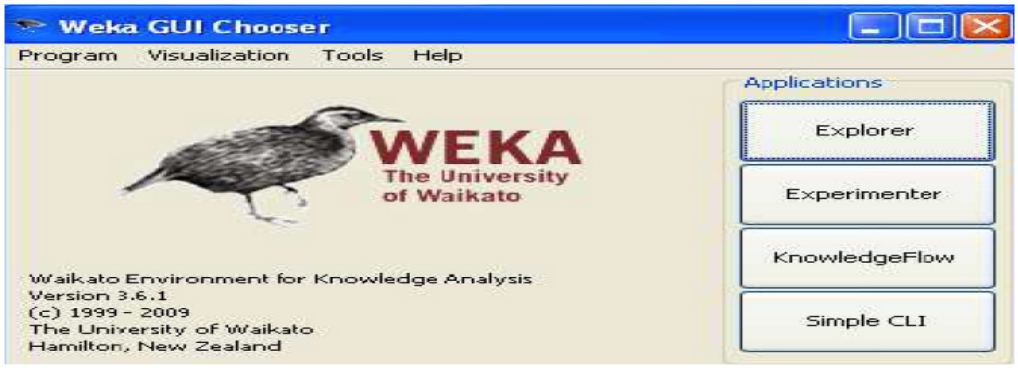


Figure 3.1: WEKA GUI application main window

A number of DM methods were implemented and experimented in the WEKA software. Some of them were based on decision trees like the J48 decision tree, some are rule-based like PART and decision tables, and some of them are based on probability and regression, like the Naïve Bayes algorithms were implemented.

One way of using WEKA is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances. A third is to apply several different learners and compare their performance in order to choose one for prediction. The learning methods are called classifiers, and in the interactive WEKA interface you select the one you want from a menu lists. Many classifiers have tunable parameters, which you access through a property sheet or object editor. A common evaluation module is used to measure the performance of all classifiers (Ian and Eibe, 2005).

Implementations of actual learning schemes are the most valuable resource that WEKA provides. But tools for preprocessing the data, called filters, come a close second. Like classifiers, you select filters from a menu and tailor them to your requirements (Amir and Shahram, 2011).

The investigator showed how different filters can be used, list the filtering algorithms, and describe their parameters in the prediction of the pattern of fertility rate in Ethiopia, particularly in four different HDSS areas. The data is often presented in dbase, SPSS, Epi-info, spreadsheet or database. However, WEKA's native data storage method is attribute relation file format (henceforth ARFF) format. You can easily convert from a spreadsheet to ARFF. The bulk of an ARFF file consists of a list of the instances, and the attribute values for each instance are separated by commas. Most spreadsheet and database programs allow you to export data into a file in

comma-separated value (henceforth CSV) format as a list of records with commas between items. Having done this, you need only load the file into a text editor or word processor; add the dataset's name using the @relation tag, the attribute information using @attribute, and a @data line; and save the file as raw text. However, you don't actually have to go through these steps to create the ARFF file yourself, because the explorer can read CSV spreadsheet files directly (Ian and Eibe, 2005).

3.2. Decision Tree Classifiers

Han and Kamber (2006) defined decision tree as a flowchart like tree structure, where each internal node (non-leaf 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.

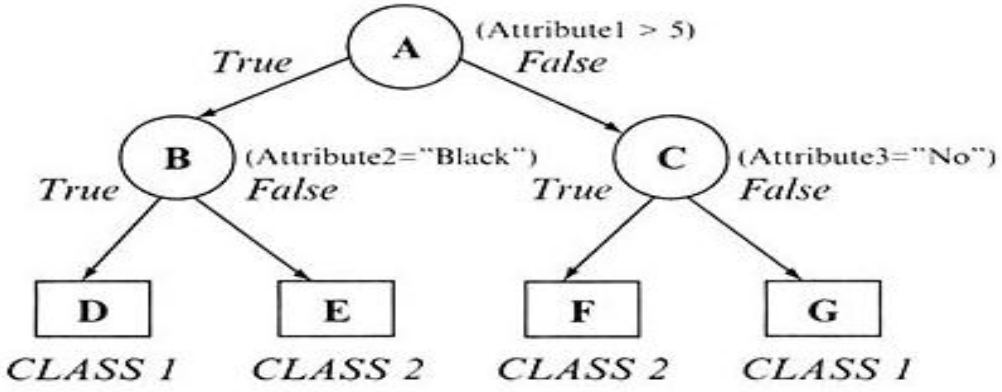


Figure 3.2: A simple Decision Tree

Decision tree is powerful in its functions and a very popular tool for classification and making prediction. It is constructed top-down: specific instances are put in sets, and as the tree grows, smaller subsets which are mainly leaf and branches are gradually divided. Each leaf represents the classifying outcomes of the decision tree, and each branch from a leaf node corresponds to the possible values/criterions for this attribute. Decision trees can classify variables according to a certain rule or classify data based on some data characteristics (Chang, 2007).

Decision tree is one of the easier data structure to understand DM. Rules from the training dataset are first extracted to form the decision tree which is then used for classification of the testing dataset. A decision tree is necessarily a tree with an arbitrary degree that classifies instances. They

are a powerful tool for classification and predication but require extensive computation. Creating the tree based on the training set takes time although making decisions once the tree is made is not time consuming. Classification tree algorithms may be divided into two groups: one whose result is a binary tree and other that yields non-binary trees (also called multiway) splits (Gorge, 2004).

In decision trees, the leaf node represents the complete classification of a given instance of the attribute and the decision node specifies the test that is conducted to produce the leaf node. Thus, with a decision tree, the sub tree that is created after any node is necessarily the outcome of the test that was conducted. In addition, a decision tree is used to classify a certain instance from the root of the tree till the leaf node which provides the outcome of that instance. A major issue in using decision tree is to find out how deep the tree should grow and when it should stop. Usually, if all the attributes are different and lead to the same outcome, the decision tree might not be the most effective in making decision and, at the same time, the size of the tree will be large (Lewis, 2002).

There are a number of algorithms that are based on decision trees. The investigator compared results of different decision tree based tools, techniques and algorithms to evaluate each for a given dataset. The researcher hopes to determine the decision tree or algorithm that provides better accuracy for the particular dataset. Some of the most common and effective types of algorithms based on decision trees are C4.5, PART and CART (Gorge, 2004).

The C4.5 algorithm is a part of the multiway split decision tree. C4.5 yields a binary split if the selected attribute/variable is numerical, but if there are other variables representing the attributes it will result in a categorical split. That is, the node will be split into C nodes where C is the number of categories for that attribute (Gorge, 2004). The J48 decision tree in WEKA is based on the C4.5 decision tree algorithm.

The presence of nodes containing insignificant attributes results in increased depth, which detracts from the effectiveness of decision trees. Moreover, decision tree algorithms not only choose the best splitting attribute for a node, but also decide what values or how many branches to assign to that node. Poorly designed decision tree algorithms that assign random values often cause an

ineffective rendering of the decision tree technique (Gorge, 2004).

The C4.5 algorithm: utilizes the same basic inductive tree creation approach as ID3, but extends its capabilities to classification of continuous data by grouping together discrete values of an attribute into subsets or ranges. Another advantage of C4.5 is that it can predict values for data with missing attributes based on knowledge of the relevant domains (Lewis, 2002).

As mentioned above, the researcher implemented C4.5 learners, each based on J48, but with the possibility of using one of three alternative models in the leaves to construct the best fitted model:

- J48-Linear: each leaf may contain a classifier that uses linear regression functions to approximate class membership (the so-called Classification via Regression classifier in WEKA (Witten, et. al., 1999).
- J48-IB1: a leaf may contain a simple nearest neighbor classifier (Cover and Hart, 1997) using one neighbor (i.e., IB 1, in the terminology of (Aha, et. al., 1991).
- J48-Bayes: a leaf may contain a Naïve Bayes Classifier (Langley, et. al., 1992) that uses a normal distribution assumption for the continuous attributes (John and Langley, 1995).

The learning algorithm was presented with a set of examples relevant to the problem domain classification task done at the HDSS dataset in order to predict the pattern of fertility rate. The aim of the learning method is to produce a tree that correctly classifies all examples in a subset of the training set. All other examples in the training set are then classified using the tree. If the tree gives the correct answer for all of these examples then it is correct for the entire training set, and the iterative process terminates. If not, a selection of the incorrectly classified examples is added to the initial subset and the process starts again (Quinlan, 1993)

Moreover, the researcher implemented a divide-and-conquer strategy used to construct the proposed decision tree (Quinlan, 1986). The choice of the test option mode to partition the training set was crucial for the complexity of the inducted tree. The test is to select an attribute for the root tree and subsequent sub-trees with the parameter of test option of pruned or unpruned situation. The information gain measure was used to select the test attribute at each node in the tree. Such a

measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that simple (but not necessarily the simplest) tree is found. So, after the above proposed procedures the following explanations were used for the HDSS data warehouse to formulate the best attributes selection by using entropy and information gain techniques.

In addition to select the test attribute (i.e., the best attribute for splitting), the entropy and information gain need to be calculated for each attribute. The C4.5 algorithm adopts an information-based method that relies on two assumptions (Velickov and Solomatine, 2000). If set S represent the training set and x, y and z are number of examples of classes X, Y and Z respectively, then the assumptions are used to formulate the best attributes selection that can be applied on the WEKA tool:

- Any correct decision tree for S will classify examples in the same proportion as their representation in S. Thus, an arbitrary example belongs to class X, Y and Z with probability:

  $\frac{X}{(X+Y+Z)}, \frac{Y}{(X+Y+Z)}$ or $\frac{Z}{(X+Y+Z)}$ respectively and

- When a decision tree is used to classify an example (training set), it returns a class. A decision tree can thus be regarded as a source of a message X, Y, or Z, with the expected information needed to generate this message given by:

$$I(X,Y,Z) = -\frac{X}{(X+Y+Z)}\log_2\left(\frac{X}{(X+Y+Z)}\right) - \frac{Y}{(X+Y+X)}\log_2\left(\frac{Y}{(X+Y+Z)}\right) - \frac{Z}{(X+Y+X)}\log_2\left(\frac{XZ}{(X+Y+Z)}\right)$$

From these assumptions, the expected information required for the decision tree with attribute A as its root is given by:

$$E(A) = \sum \frac{Xi + Yi + Zi}{X + Y + Z} . I(Xi + Yi + Zi)$$

Where xi, yi, and zi are number of examples for the classes of X, Y and Z respectively with value Ai of the attribute A. The summation gives the total expected information for attribute A. The information gained by branch the tree on A is:

$$GAIN(A) = I(X, Y, Z) - E(A)$$

At each non-leaf node of the decision tree, the gain of each untested attribute is determined. This gain in turn depends on the value of xi, yi, and zi for each value Ai of the attribute A. Every example is examined to determine its class and its value of A. Thus, the total computational requirement per iteration is proportional to the product of size of the training set, the number of attributes, and the number of non-leaf nodes in the decision tree. The training stage of the algorithm results is a classifier in a form of decision tree, which can be used to classify an unseen set of testing samples.

The attribute with the highest information gain is considered as the most discriminating attribute of the set under consideration. So, an attribute that yields maximum information gain will be chosen for data set partitioning. Then, a node is created and labeled with the chosen attribute, branches are formed for each value of the attribute, and the samples are partitioned accordingly. The same criteria were applied to each split sample on this research project. The iterative divide and conquer process executes until no further split was required (Emmanuel, 2007).

Furthermore, a set of classification rules can be extracted from the decision tree by tracing the path from the root to each leaf (corresponding class). This set of rules can be consequently plugged into propitiate knowledge-based system (Velickov and Solomatine, 2000). So, the researcher computed the C4.5 algorithm using J48 method in order to get the best fitted model that can appropriate to predict the pattern of fertility in Ethiopia, specifically for the HDSS sites and also the investigator tried to generate rules from the J48 decision tree with comparing to the PART rules by the parameter of accuracy measures.

3.3. Neural Networks

Neural networks are network structures consisting of a number of nodes connected through

directional links. Each node represents a processing unit, and the links between nodes specify the causal relationship between connected nodes (Kantardzic, 2003).

Neural networks can be used for many purposes, notably descriptive and predictive data mining. They were originally developed in the field of machine learning to try to imitate the neurophysiology of the human brain through the combination of simple computational elements (neurons) in a highly interconnected system (Giudici, 2003).

Neural networks became popular in the 1980s because of a convergence of several factors. First, computing power was readily available, especially in the business community where data was available. Second, analysts became more comfortable with neural networks by realizing that they are closely related to known statistical methods. Third, there was relevant data since operational systems in most companies had already been automated. Fourth, useful applications became more important than the holy grails of artificial intelligence. Building tools to help people superseded the goal of building artificial people. Because of their proven utility, neural networks are, and will continue to be, popular tools for data mining (Berry and Linoff, 2004).

According to Two Crows Corporation (2005), a neural network (Figure 3.3) starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the 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 (e.g., the weight from node 1 to 3 is W13 — see Figure 3.3), adds them together, applies a function (called the activation or squashing function) to them, and passes the output to the node(s) in the next layer. For example, the value passed to node 4 is:

Activation function applied to ([W14 * value of node 1] + [W24 * value of node 2])

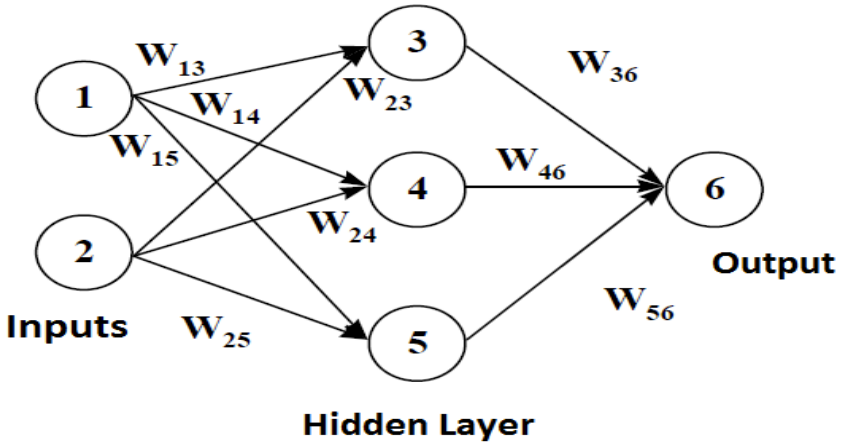


Figure 3.3: A Neural Network Architecture

Each node may be viewed as a predictor variable (nodes 1 and 2 in this example) 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. In fact, if there is a linear activation function but no hidden layer, neural nets are equivalent to a linear regression; and with certain non-linear activation functions, neural nets are equivalent to logistic regression.

The connection weights (W's) are the unknown parameters which are estimated by a training method. Originally, the most common training method was backpropagation; newer methods include conjugate gradient, quasi-Newton, Levenberg-Marquardt, and genetic algorithms. Each training method has a set of parameters that control various aspects of training such as avoiding local optima or adjusting the speed of conversion (Two Crows Corporation, 2005).

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, you may have to experiment with these parameters.

3.3.1. Multilayer Perception

As Giudici (2003) stated that the multilayer perceptron is a neural network which is the most used architecture for predictive data mining. It is a feed forward network with possibly several hidden layers, one input layer and one output layer, totally interconnected. It can be considered as a highly

nonlinear generalization of the linear regression model when the output variables are quantitative, or of the logistic regression model when the output variables are qualitative.

The network is feed forward if the processing propagates from the input side to the output side unanimously, without any loops or feedbacks. In a layered representation of the feed forward neural network, there are no links between nodes in the same layer; outputs of nodes in a specific layer are always connected as inputs to nodes in succeeding layers. This representation is preferred because of its modularity, i.e., nodes in the same layer have the same functionality or generate the same level of abstraction about input vectors (Kantardzic, 2003).

3.4. Naïve Bayes Classifiers

Naïve Bayes classifiers method is based on probabilistic knowledge. This method goes by the name Naïve Bayes, because it's based on Bayes' rule and "naively" assumes independence- it is only valid to multiply probabilities when the events are independent (Gorge, 2004). Thus, the Naïve Bayes rule outputs probabilities for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data. In Naïve Bayes classifier, the probability of the attributes are calculated based on normal distribution's mean, standard deviation, weighted sum, and precision. So, the investigator tried to show the experiments on Naïve Bayes algorithms in order to get the best fitted model for the classification as well as prediction of the HDSS dataset.

Naïve Bayes gives a simple approach, with clear semantics, to representing, using, and learning probabilistic knowledge. Impressive results can be achieved using it. It has often been shown that Naïve Bayes rivals, and indeed outperforms, more sophisticated classifiers on many datasets. The moral is, always try the simple things first. Repeatedly in machine learning people have eventually, after an extended struggle, obtained good results using sophisticated learning methods only to discover years later that simple methods such as IR and Naïve Bayes do just as well-or even better. The Naïve Bayes classification reads a set of examples from the training set and uses the Bayes theorem to estimate the probabilities of all classifications. For each instance, the classification with the highest probability is chosen as the prediction class (Ian and Eibe, 2005). Most of the Bayesian

classifiers utilize model that gives the probability of the data conditioned on the hypothesized model: P (XH, p), known as likelihood function (Velickov and Solomatine, 2000).

Moreover, Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Studies comparing classification algorithms have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases (Gorge, 2004). In this case the researcher tried to show in comparable performance by implementing several experiments with J48 decision tree algorithm and Naïve Bayes classifier on HDSS data warehouse.

As mentioned before, Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve" (Fayyad, et. al., 1996a).

In this research study, the researcher made experiments based upon the Bayes approach defines the classification problem in terms of probabilities that formulated by the underneath proof. More specifically, the three main concepts required are conditional probability, Bayes Theorem, and the Bayes decision rule. The conditional probability P (A\B), which is used to define independent events (Michael and Murray, 2006 and Ian and Eibe, 2005), is defined by

$$P(A\backslash B) = \frac{P(AUB)}{P(B)},$$

Where P (A\B) is the probability that event A happens, given that B is observed. Similarly,

$$P(B\backslash A) = \frac{P(AUB)}{P(A)},$$

Where P (B\A) is the probability that event B happens, given that A is observed. It then follows (by substitution) that

$$P(A\ n\ B) = P(A)P(B\backslash A).$$

Although, the premise of Bayes Theorem starts with an initial degree of belief that an event was occur, and then with new information this degree of belief can be "updated" (Michael, W. B. and Murray, B., 2006). These two degrees are represented, respectively, by the prior probability P (A\B) and the posterior probability P (B\A), which are related by

$$P(A\backslash B) = \frac{P(A)P(B\backslash A)}{P(B)}.$$

Finally, the Bayes decision rule states that based on the posterior probabilities, it is possible to assign an element x to a class with the largest probability. In particular, for the classifying problem, the conditional probabilities described above can be defined as follows: let x be a data sample (vector of features) and ωi one of the possible classes (Michael, W. B. and Murray, B., 2006). Then, P (x\ ωi) is the prior probability, because it can be obtained based on prior knowledge (i.e., the distribution constructed from training data). Given class i, it specifies the probability of finding x within this class. Similarly, P (ωi\x) is the posterior probability, because it is computed based on posterior knowledge. Given sample x, it specifies the probability that x belongs to class j. For a given x, if then x belongs to class 1, otherwise to class 2.

$$P(\omega 1|\chi) > P(\omega 2|\chi),$$

Where

$$P(\omega 1|\chi) = \frac{P(\chi|\omega i)\mathrm{P}(\omega i)}{P(\chi)},$$

The denominator term of equation is the overall probability of x in all the classes. For a given x, one must compute the posterior probabilities for all ωi classes, then assign x to the class that yields the maximum posterior probability (Aha, et. al., 1991).

Generally speaking, the Bayesian methodology for classification as well as prediction of the pattern of fertility rate in Ethiopia, particularly for the HDSS data warehouse follows these five steps (John and Langley, 1995) :

1. Collect data, and estimate parameters such as mean and covariance for each class (for the parametric approach the researcher assumed that all the probability density functions have a Gaussian behavior).

2. Choose a set of features.
3. Choose a model and derive a decision rule with these parameters.
4. Train the classifier and apply the decision rule by using a discriminant function (a way to represent a pattern classifier), and apply it to a test data set to classify each sample.
5. Evaluate the decision rule. Measure the accuracy /error rate in order to improve the choice of features and the overall design of the classifier.

3.5. Performance Evaluation for Predictive Model

Throughout this section the investigator had tacitly assumed that the goal of the performance evaluation was to maximize the success rate of the predictive model for HDSS dataset. Predictive models are evaluated in terms of correctness, often referred to as performance, and applicability. The performance measures are almost always geared towards the evaluation of an instance of a model type, and are almost always realization method independent. Applicability measures also contain measures that apply to the model type itself, pertaining to the need of models to be evaluated in terms of their context (Vinterbo, 1999).

Once a predictive model is developed using the fertility rate HDSS dataset, the model should be checked as to how it will perform for the future data which, it has not seen during the model building process.

3.5.1. 10-Fold Cross Validation

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold.

According to Olson and Delen (2008) 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation. In fact, in data mining community, for methods-comparison studies with relatively smaller datasets, k-fold type of experimentation methods is recommended. In essence, the main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample.

The cross validation estimate of the overall accuracy of a model is calculated by simply averaging the 10 individual accuracy measures

$$CVA = \frac{1}{10}\sum_{1}^{10} Ai$$

Where CVA stands for cross validation accuracy and A is the accuracy measure (e.g., it-rate, sensitivity, specificity, etc.) of each folds.

There are three steps to perform 10-Fold Cross Validation (Olson and Delen, 2008)):

Step 1: The complete dataset is randomly divided into 10 disjoint subsets (i.e., folds) with each containing approximately the same number of records. Sampling is stratified by the class labels to ensure that the proportional representation of the classes is roughly the same as those in the original dataset.

Step 2: For each fold, a classifier is constructed using all records except the ones in the current fold. Then the classifier is tested on the current fold to obtain a cross-validation estimate of its error rate. The result is recorded.

Step 3: After repeating the step 2 for all 10 folds, the ten cross validation estimates are averaged to provide the aggregated classification accuracy estimate of each model type.

3.5.2. Confusion Matrix

The confusion matrix is a useful tool for analyzing how well the researcher's classifier can recognize tuples of different classes. The following procedures and rules were implemented to confirm the model performance evaluation for the results of the predicted model of the fertility rate in Ethiopia, particularly for HDSS areas. Given M classes; a confusion matrix is a table of at least size M by M. An entry, CMi,j in the first M rows and M columns indicates the number of tuples of class i that were labeled by the classifier as class j. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry CM1, 1 to entry CMm,m, with the rest of the entries being close to zero (Selam, A., 2011).

In building a classification model, the confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models (Amir and Shahram, 2011).

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 3.1: Confusion Matrix

As shown in table 3.1, a confusion matrix table of size two by two, the following measures can be calculated to measure predicted pattern of the fertility rate for HDSS dataset's accuracy of the model, True Positive Rate, False Positive Rate, Accuracy, Precision, Recall, F-measure and ROC Curve.

The True Positive Rate of a classifier is expected by dividing the correctly classified positives by the total positive count.

$$True\ Positive\ Rate = \frac{TP}{TP+FN}$$

The True Negative Rate of a classifier is estimated by dividing the incorrectly classified negatives by the total negatives count.

$$True\ Negative\ Rate = \frac{TN}{TN+FP}$$

The Accuracy of a classifier is projected by dividing the total correctly classified positives and negatives instances by the total number of samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is calculated by dividing correctly classified instances by the total number of correctly and incorrectly classified samples.

$$Precision = \frac{TP}{TP+FP}$$

F-Measure is calculated as the harmonic mean of recall and precision.

$$F - Measure = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$$

3.5.3. Area Under the ROC Curve

A Receiver Operating Characteristics (ROC) curve is a technique for visualizing, organizing and selecting classifiers based on their performance. In essence, it is another performance evaluation technique for classification models and also useful tool for comparing two or more classification models. ROC curves have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers. The use of ROC analysis has been extended into visualizing and analyzing the behavior of diagnostic systems. Recently, the medical decision making community has developed an extensive literature on the use of ROC curves as one of the primary methods for diagnostic testing (Olson and Delen, 2008).

As Han and Kamber (2006) described ROC curve shows the trade-off between the true positive rate or sensitivity (proportion of positive tuples that are correctly identified) and the false positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model. That is, given a two-class problem, it allows us to visualize the trade-off between the rate at which the model can accurately recognize ‘yes’ cases versus the rate at which it mistakenly identifies ‘no’ cases as ‘yes’ for different “portions” of the test set. Any increase in the true positive rate occurs at the cost of an increase in the false-positive rate. The area under the ROC curve is a measure of the accuracy of the model.

In order to plot an ROC curve for a given classification model, true positive (TP) rate is plotted on the Y axis and false positive (FP) rate is plotted on the X axis (see Figure 3.4). As Han and Kamber

(2006) described the process of drawing Roc curve we start at the bottom left-hand corner (where the true positive rate and false-positive rate are both 0), we check the actual class label of the tuple at the top of the list. If we have a true positive (that is, a positive tuple that was correctly classified), then on the ROC curve, we move up and plot a point. If, instead, the tuple really belongs to the 'no' class, we have a false positive. On the ROC curve, we move right and plot a point. This process is repeated for each of the test tuples, each time moving up on the curve for a true positive or toward the right for a false positive.

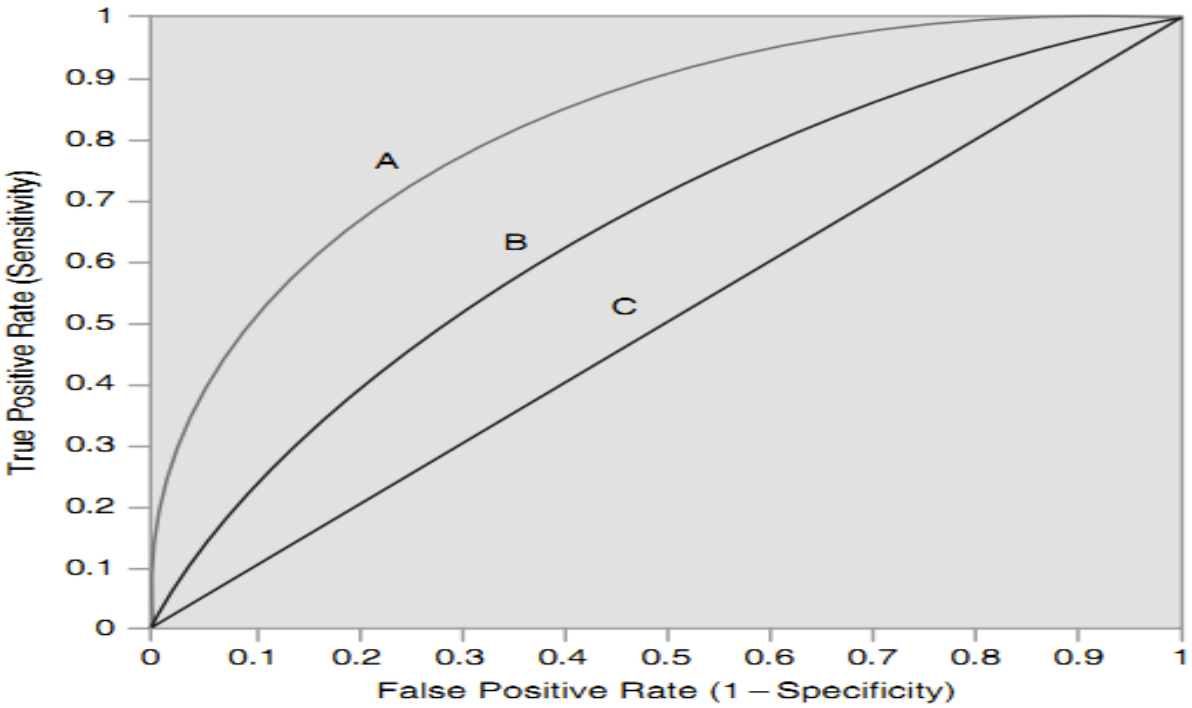


Figure 3.4: A Sample ROC curves

As Olson and Delen (2008) noted to compare classifiers or to judge the fitness of a single classifier one may want to reduce the ROC measures to a single scalar value representing the expected performance. A common method to perform such task is to calculate the area under the ROC curve, Area under the Curve (AUC) is a portion of the area of the unit square, and its value will always be between 0 and 1.0. A perfect accuracy gets a value of 1.0. The diagonal line y = x represents the strategy of randomly guessing a class. For example, if a classifier randomly guesses the positive class half the time (much like flipping a coin), it can be expected to get half the positives and half the negatives correct; this yields the point (0:5; 0:5) in ROC space, which in turn translates into area under the ROC curve value of 0.5. No classifier that has any classification power should have

an AUC less than 0.5.

For example, In Figure 3.5 classification performances of three classifiers (A, B and C) are shown in a single ROC graph. Since the AUC is the commonly used metric for performance comparison of prediction models, one can easily tell that the best performing classifier (out of the three that is being compared to each other) is A, followed by B. The classifier C is not shoving any predictive power; staying at the same level as random chance.

ROC Area	Performance
0.9 – 1.0	Excellent (A)
0.8 – 0.9	Good (B)
0.7 – 0.8	Fair (C)
06 – 0.7	Poor (D)
0.5 – 0.6	Fail (F)

Table 3.2: Performance measure of ROC Area

To summarize, three algorithms namely J48 classifier, Multilayer Perception and Naïve Bayes were used for model building and 10-fold cross validation, predictive accuracy, TP rate, TN rate, precision, and F-Measure, are six measures used for the evaluation of classification and prediction methods while predictive accuracy, TP rate, TN rate and ROC Area were used to compare the models.

CHAPTER FOUR

Business understanding and data preprocessing

One of the phases in the knowledge discovery process is understanding the business domain. Without a keen understanding of the business domain, no matter what tools used or how good techniques followed, may not provide useful result (Helen, 2003). Having an in-depth knowledge in the business domain enables data analysts clearly set the objectives and attempts to be made to attain the defined goals.

This chapter provides interesting features for business understanding and data preprocessing of the HDSS dataset emphasizing its ability to accurately measure the fertility rate. The current state of HDSS dataset was summarized and the contribution of HDSS dataset to our understanding of fertility rate in Ethiopia, particularly for HDSS areas was discussed briefly.

Domain experts are consulted to have insight into the problem domain. The domain experts constitute three individuals from Jimma University public health college, Haramaya University Public Health college and Mekelle University Public Health collage. On the basis of the insight gained from discussion with domain experts and review of relevant documents, a clear understanding of the data is achieved.

4.1. Problem domain and business understanding

The six networked HDSS sites of Butajira, Dabat, Gilgel Gibe, Kersa, Arbaminch and Kilte Awlaelo, respectively, have been run by Addis Ababa, Gondar, Jimma, Haramaya, Mekelle and Arbaminch Universities.

- Butajira Rural Health Program (BRHP) was established in 1987, covers 10 (1 urban and 9 rural) randomly selected Kebeles from the former Meskan and Mareko Woreda which constituted 82 rural and 4 urban Kebeles in the Woreda which are now divided into ‘Meskan’, ‘Mareko’ and “Silti” Woredas.
- Dabat Research Center (DRC)was established in 1996, covers 10 randomly selected Kebeles (seven rural and three urban) from 32 Kebeles in Dabat district.

- Gilgel Gibe Field Research Center (GGFRC) was established in September 2005, comprised of 10 Kebeles (eight rural and two urban) bordering the hydroelectric dam within about 10 km radius found in four Woredas (Sekoru, Tiro Afeta, Omo Nada and Kersa).
- Kersa Demographic Surveillance and Health Research Center (KDS-HRC) was established in September 2007, covers 12 Kebeles (two urban and ten rural) from 38 randomly selected Kebeles.
- Kilte Awlaelo Demographic and Health Development Program (KA-DHDP) was established in 2009, consists of 10 (9 rural and 1 urban) randomly selected Kebeles from 39 Kebeles in Wukro and Atsiwmberta Districts.
- Arba Minch Demographic and Health Development Program (AM-DHDP) was established in 2009, comprised of 9 Kebeles (one small town and 8 rural Kebeles) that were randomly selected from 29 Kebeles in the Woreda.

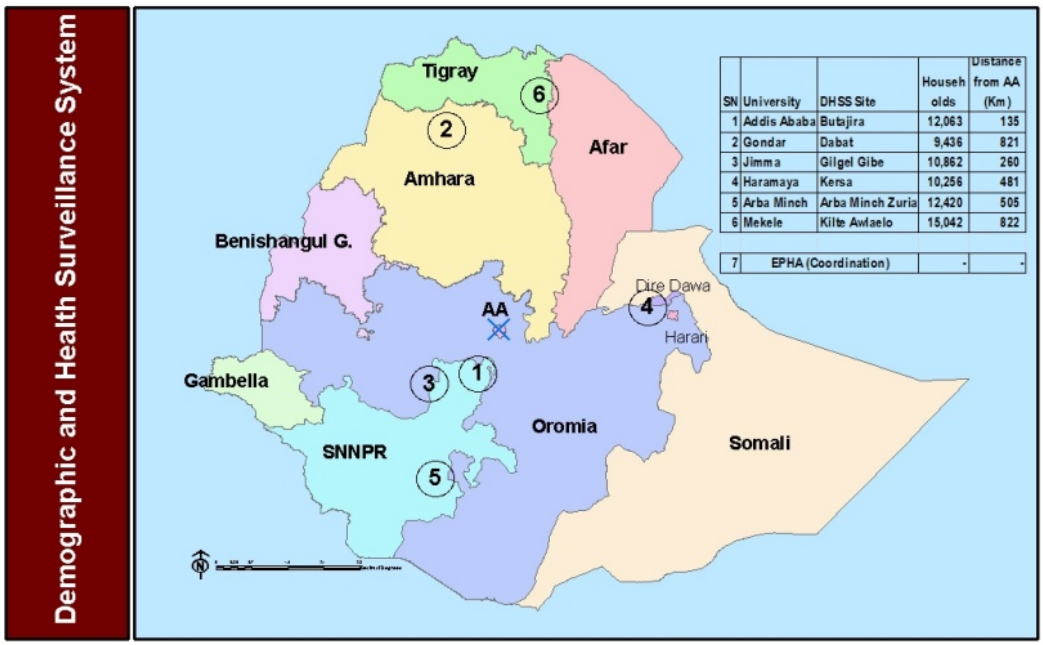


Fig 4.1. Locations of HDSS sites

Populations change through three major processes: fertility, mortality, and migration. A useful way to express the rate at which women have children is the Total Fertility Rate (TFR). TFR is the

average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given set of age-specific fertility rates (WIKI-1, Christenson, McDevitt, and Stanecki, 2004). If the average woman has approximately 2 children in her lifetime, this is just enough to maintain the population (Gams, 2007).

A fertility rate is a measure of the average number of children a woman will have during her childbearing years. The fertility patterns are different between countries and over time. Many different factors can affect fertility rates. Many of these factors are difficult to measure because they involve subjectivity and some of them may not apply across cultures. This makes it especially difficult to find variables that can be used to predict future fertility rates.

The World Fertility Report 2013: Fertility at the Extremes, the fifth in a series, adopts a particular focus on countries where fertility levels are high (more than 3.2 children per woman) and countries where fertility levels are low (2.0 children per woman or less). In 1990-1995, around the time of the ICPD, 105 countries had high fertility as opposed to just 66 countries in 2005-2010 (the period of the most recent fertility estimates). Several countries, such as Iran, the United Arab Emirates and Viet Nam, experienced rapid fertility declines over this time period, moving from high fertility to low fertility over the span of a single generation. High-fertility countries are increasingly concentrated in sub-Saharan Africa while low fertility countries have moved from being predominantly European to include countries from Asia and Latin America and the Caribbean. (world fertility report, 2013)

Fertility rates are still at very high levels in Africa and some Arabic countries, followed next by the countries of Central and South America. Lower rates are found in Europe and other industrialized countries like Canada and Japan.

4.1.1. Workflow in the HDSS Areas

HDSS is a set of field and computing operations to handle the longitudinal follow-up of well-defined entities or primary subjects (individuals, households, and residential units) and all related demographic and health outcomes within a clearly circumscribed geographic area. Unlike a cohort study, HDSS follows up the entire population of such a geographic area. DSS is an intensive study technique that produces data with substantial advantages over other data (Berhane and Byass, 2003,

IDRC, 2002).

An initial census enumerates and registers the entire population of a well-defined geographic area, the DSA, and after that regular visits are made to each registered location within the DSA in order to record demographic and health-related events that have taken place since the previous visit and to update the status of all entities registered at the location. The DSS study population is typically defined as those people who are resident within the DSA, and it follows that there are only two ways to be admitted to the DSS study population; through birth or in-migration to the DSA. Likewise, there are two ways to exit the DSS study population; through death or outmigration from the DSA. Defined in this way a DSS is similar to a population register; the primary differences are

1. a DSS usually monitors a comparatively small population intensely and
2. a DSS is an active data collection system that invests considerable effort to track down and visit each member of the study population several times during each year rather than waiting for events and status updates to occur when individuals contact "the system" for some other reason, as is often the case in a population register (IDRC, 2002).

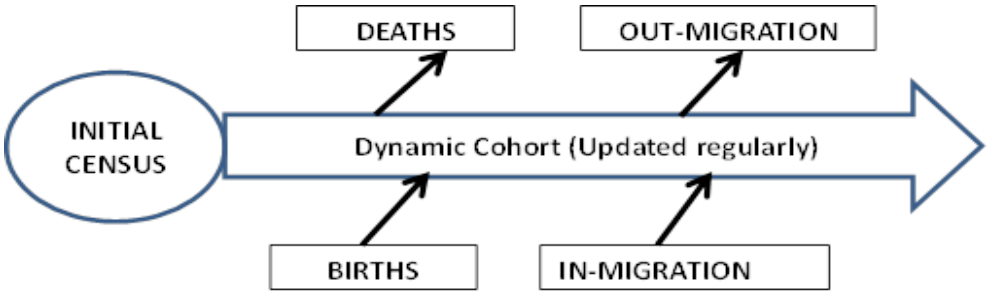


Fig 4.2. Health and Demographic surveillance system model (INDEPTH network)

The demographic surveillance methodology grew out of the need for accurate information describing the "at risk" (denominator) population living in rural areas in the developing world where vital registration systems either do not exist, or when they exist do not function well enough to provide this information. The primary advantage of most DSSs is that they are the only producers of accurate individual-level, community-based data in the remote rural areas of the Developing World where they are typically (and purposefully) situated (International Development Research Centre., 2002).

Beyond being the only sources of high quality data in these areas, DSSs produce prospective, fully linked individual, household and community-level data that describe reasonably large whole populations and often include a rich set of prospectively monitored attributes that make possible nuanced longitudinal analyses of fertility at several levels.

By its very nature DSS is longitudinal, and most DSSs collect data over the course of many years thereby describing the history of their study populations. This provides the ability to measure and describe trends in fertility or to control for trends in order to isolate and study other factors contributing to the level of fertility. Because most DSSs visit each registered location several times per year, the temporal resolution and accuracy of the data are both high allowing high resolution trends to be calculated and controlled for; for example, seasonal changes in the risk of dying-important in areas where malaria is a significant cause of death (International Development Research Centre., 2002).

HDSS tracks the presence of individuals in a defined study area. These individuals can enter and leave the study area in a small set of well-defined ways (for example, entering through birth or in-migration and leaving through death or out-migration). The International Network for the continuous Demographic Evaluation of Populations and Their Health in developing countries (henceforth INDEPTH) reference model uses events to record the ways individuals enter (or return to) and leave the study area over time (International Development Research Centre., 2002).

When a DSS tracks episodes, the concept of the "time resolution" of this tracking is very important. Below a certain time threshold, movements into or out of a particular place are not recorded. DSSs are concerned not only with the physical location or residence of individuals but also with their membership in social groups (such as households) and their relationships with other individuals (such as marital unions or parenthood). Many DSSs also need to reconstruct genealogies and to record isolated events, such as pregnancy outcomes or births and deaths external to the study area (International Development Research Centre., 2002).

4.2. Data Understanding

4.2.1. Fertility rate Based on HDSS Dataset

During the past 30 years, HDSSs have been established in a number of field research sites in various parts of the developing world where routine vital-registration systems were poorly developed or nonexistent. Although these systems may have been developed differently in terms of their initial rationale, they are all required to track a limited and common set of key variables determining population dynamics and demographic trends. DSSs have similar approaches to defining key variables and their relationships and to developing systems for collection, storage, and analysis of these data (International Development Research Centre., 2002).

4.2.2. Data Collection

It is a bare fact that the concept of data mining doesn't exist without data. There is some real benefit if the data is already part of a data warehouse. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. Furthermore, many of the problems of data consolidation have already been addressed and maintenance procedures have been put in place. Setting up a large data warehouse that consolidates data from multiple sources, resolves data integrity problems.

The primary data source for this research consumption is the four HDSS research centers in the country, namely: The Gilgel Gibe HDSS research center, the Killiteawelaelo HDSS research center, The Arbaminch HDSS research center and the Dabat HDSS research center. The data collected from these research center were primarily in different format. Some of them were in .csv format others where in .dbf, and others were in access database.

Taking the advantages of using data warehouse for data mining into consideration and considering the variety of the data sources used for this research, the researcher tried to build a data warehouse. This data warehouse contains an integrated data from the different HDSS research centers and from different data sources.

The database management tool used to build the data warehouse was mysql server and mirth connect was used to apply different integration rules and to transform data from the sources to the main data warehouse.

4.2.3. Data Source Description

As previously stated, the data was obtained from four HDSS research centers database. The HDSS is a network of research centers run by Jimma University, Mekelle University, Arbaminch University, and Gondar University. The research is financially supported by CDC through Ethiopian Public Health Association. The original DSS population differs in different research centers but currently each research center has more than 60,000 individuals to follow on average.

As mentioned earlier in chapter one, each HDSS registers births, deaths, marriages, new households, out- and in-migrations, and internal moves (migration within HDSS kebeles). Household and environmental variables were measured during the censuses. The study bases are now well established and is being used for other more focused studies on essential health problems of the country, using qualitative, as well as quantitative, research methods. So far, research on under 5 child mortality, childhood respiratory illnesses, other infectious diseases, reproductive health, and mental health has been conducted using the study-base infrastructure. Mostly the data collected and managed by the HDSS dataset are cleaned; as a result, these datasets were less likely to contain missing values (International Development Research Centre., 2002).

4.2.4. Data Quality Assurance

Data quality assurance mechanisms have been instituted at several points to ensure the integrity of the data. The most critical of these is field supervision. Field supervisors perform the immediate supervision of data collection procedures on a daily basis (Berhane, et. al., 2004).

Their tasks include checking of each completed data form and visiting randomly selected households each month on a weekly-distributed timetable. The research assistants perform the next level of supervision. They are responsible for the overall supervision of the data flow from the household level to the computer system. Research assistants also perform data checking at the field level in randomly selected households. Researchers also work in the field to provide on-site technical assistance and guidance as well as checking data quality. More recently, with the advent and easy availability of the Global Positioning System (henceforth GPS), mapping exercises at the household level has been carried out (IDRC, 2002).

4.3. Preprocessing

4.3.1. Data Preparation

Data preparation is the most important phases of the data analysis activity which involves the construction of the final data set (data that will be fed into the modeling tool) from the initial raw data. Data preparation generates a dataset smaller than the original one, which can significantly improve the efficiency of data mining. This task includes: attribute selection, filling the missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision to settle inconsistency.

4.3.1.1. Attribute Selection

Deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types (Shearer, 2000).

Therefore, in this thesis the attributes are selected with the help of domain expert and extensive literature review. Because taking all the variables in the database we have, feed them to the data mining tool and find those which are the best predictors may not be work very well. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models (Two Crows Corporations, 2005).

Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts in order to simplify the task of modeling. The data warehouse developed by the researcher from the data sources contains many attributes, and to decide on the relevant attributes for this study the researcher has discussed with domain expert in the area. As described in table 4.1, the following attributes are selected from the 9 year updated data: Birth number, Residence, Region, Mother's education, mother’s occupation, marital status, educational status, and relationship to the head of the house and others. The final selected attributes were prepared and preprocessed as stated in the following section, before developing the models.

No	Attributes	Description	Values	Data Type	Missing and Unknown
1	ethnicity	Ethnicity of the individual	Different ethnicity types depends of the area they are living in	nominal	0
2	Educ_status	The education status of the individual	Educational status in interval (1-3, 4-8, 9-12, higher education)	nominal	310
3	religion	Religion of the individual	Orthodox, Muslim, protestant, catholic, other beliefs	nominal	0
4	occupation	The occupation of the individual	Housewife, farmer, government worker, teacher etc.	nominal	662
5	rltn_head	The relationship of the individual with the head of the house he/she lives in	Head,spouse, child, borther/sister,parent,relative,not relative,etc	nominal	0
6	birth_num	The number of births the woman gives so far	numeric	numeric	0
7	marital_status	The marital status of the woman	Married, unmarried, divorced.	nominal	0
8	setting	Where the woman lives (rural or urban)	Rural, urban	nominal	0

Table 4.1. List of independent attributes

4.3.1.2. Statistical Summary of the Attributes (features)

Here the selected attributes used for model building are statistically described in details. This statistical summary of the attributes is helpful for understanding of the data set for experimentation.

Total number of population

Table 4.2 below shows the total number of men and women living in each site's study area.

site	gender F	M	Total
arbaminch	44,347	44,797	89,144
dabat	31,408	29,285	60,693
gilgelgibe	44,672	42,738	87,410
kilteawelaelo	42,152	39,318	81,470
Total	162,579	156,138	318,717

Table 4.2. Total number of Male and Female

Total number of fertile women per site

This is the total number of women whose age is between 14 – 49 that are currently living in the study area and are at their childbearing age.

SITE	Freq.	Percent	Cum.
arbaminch	19,416	28.54	28.54
dabat	13,431	19.74	48.28
gilgelgibe	17,221	25.31	73.59
kilteawelaelo	17,965	26.41	100.00
Total	68,033	100.00	

Table 4.3. Number of female at the age of child bearing

No. of Births per site

The table below shows the total number of live births registered in each site. We can divide this birth pattern as high middle or low based on the number of children a woman gave.

BIRTH_NUM	SITE arbaminch	dabat	gilgelgib	kilteawel	Total
0	13,099	8,653	9,763	12,589	44,104
1	3,912	2,862	2,368	3,611	12,753
2	1,974	1,693	1,954	1,627	7,248
3	384	153	1,805	134	2,476
4	41	36	1,072	4	1,153
5	6	27	235	0	268
6	0	6	19	0	25
7	0	0	5	0	5
8	0	1	0	0	1
Total	19,416	13,431	17,221	17,965	68,033

Table 4.4. Number of births per site

Place of Residence

Place of residence is nominal attributes; the possible values of this attribute are rural and urban. The nominal value of this attribute is described in table 4.5. As we can see the (Modal value) majority of respondents reside in rural areas.

site	Setting		
	Rural	Urban	Total
arbaminch	**16,003**	**3,413**	**19,416**
dabat	**9,610**	**3,821**	**13,431**
gilgelgibe	**9,967**	**7,254**	**17,221**
kilteawelaelo	**15,166**	**2,799**	**17,965**
Total	**50,746**	**17,287**	**68,033**

Table 4.5. Place of residence of women in each site

Religion

This nominal attribute has six distinct values (orthodox, catholic, protestant, Muslim, tradition and others. The detail summary of this attribute is described on table 4.6.

site	religion					
	Orthodox	Islam	Catholic	Protestan	Other	Total
Arbaminch	**6,403**	**42**	**43**	**12,928**	**0**	**19,416**
Dabat	**13,041**	**387**	**2**	**1**	**0**	**13,431**
GilgelGibe	**4,257**	**12,315**	**499**	**147**	**3**	**17,221**
Kilteawelaelo	**17,658**	**291**	**11**	**4**	**1**	**17,965**
Total	**41,359**	**13,035**	**555**	**13,080**	**4**	**68,033**

Table 4.6. Religion type and number in each site

Mother's education

Mother’s education is indirectly related to the number of births she gave. Mothers Education is nominal attribute that contains eight distinct values (illiterate, Read and write, grade 1 to 3, grade 4 to 6, grade 7 to 8, grade 9 to 10, grade 11 to 12 and above 12 Primary). The most frequent value for Educational Level of the women is Illiterate as shown in table 4.8.

	site				
educ_status	Arbaminch	Dabat	GilgelGib	Kilteawel	Total
Illitrate	**10,635**	**5,964**	**8,332**	**6,247**	**31,178**
Read & Write	**3,240**	**1,023**	**1,014**	**847**	**6,124**
1–3	**3,604**	**2,137**	**3,969**	**3,772**	**13,482**
4–6	**1,211**	**1,699**	**954**	**3,723**	**7,587**
7–8	**477**	**801**	**393**	**1,715**	**3,386**
9–10	**101**	**1,075**	**213**	**1,075**	**2,464**
11–12	**4**	**405**	**86**	**279**	**774**
12+	**144**	**327**	**2,260**	**307**	**3,038**
Total	**19,416**	**13,431**	**17,221**	**17,965**	**68,033**

Table 4.7. Mother's educational background

Marital Status

Table 4.8 shows the distribution of women by marital status. The marital status attributes is nominal. This attribute contains three distinct values married refers to both legal or formal marriage, unmarried, divorced. We can see from the table that the most frequent value for this attribute is unmarried.

	marital_status			
site	Divorsed	Married	UNMarried	Total
Arbaminch	**0**	**9,775**	**9,641**	**19,416**
Dabat	**56**	**6,032**	**7,343**	**13,431**
GilgelGibe	**57**	**9,510**	**7,654**	**17,221**
Kilteawelaelo	**148**	**7,791**	**10,026**	**17,965**
Total	**261**	**33,108**	**34,664**	**68,033**

Table 4.8. Number of married, divorced and unmarried women per site

Occupation of the mother

This nominal attributes shows the work the mother do for living. This attribute somehow shows the income of the mother.

occupation	site				Total
	Arbaminch	Dabat	GilgelGib	Kilteawel	
Farmer	1,541	296	1,881	1,174	4,892
Student	584	214	600	302	1,700
House Wife	1,889	329	1,940	671	4,829
Merchant	149	274	93	141	657
Gov_employee	89	235	243	295	862
nongov_employee	249	11	147	774	1,181
Daily_Laborer	7,503	8,887	5,165	5,375	26,930
Unemployed	2,851	2,307	2,363	5,881	13,402
Other	1,692	878	633	715	3,918
UNKNOWN	2,869	0	4,156	2,637	9,662
Total	19,416	13,431	17,221	17,965	68,033

Table 4.9. Occupation of the mother

Relationship-To-the-Head

The following table shows the relationship of the woman to the head of the house she is living in. this variable has 8 values. Head refers to the head of the house, Spouse refers to the wife or husband of the head of the house, child, Sibling refers to the sister or brother of the head of the house, parent, grand parent, other relative and non-relative.

RLTN_HEAD	site				Total
	Arbaminch	Dabat	GilgelGib	Kilteawel	
Head	1,067	1,028	1,584	2,076	5,755
Spouse	9,118	5,884	8,012	9,344	32,358
Child	7,211	5,211	6,063	5,825	24,310
Sibling	406	152	240	220	1,018
Parent	281	7	30	11	329
Grand_Parent	79	484	762	349	1,674
Other_Relative	620	266	530	140	1,556
Non_Relative	634	399	0	0	1,033
Total	19,416	13,431	17,221	17,965	68,033

Table 4.10. Relationship of a woman to the head of the house

4.3.2. Handling Missing Values

For many real-world applications of data mining, even when there are huge amounts of data, the subset of cases with complete data may be relatively small. A number of problems are faced while bringing the data into proper format. Missing data is the most common problem that comes up during the data analysis process. Missing values minimizing the accuracy of classification and

rules generated by the selected data mining algorithm. Missing values lead to the difficulty of extracting useful information from that data set. Solving the problem of missing data is of a high priority in the field of data mining and knowledge discovery. Handling missing values by appropriate methods does not affect the quality of the data. In this thesis the two widely used methods are applied. One is avoiding the missing data and other is data Imputation (Kantardzic, 2003).

Avoiding the missing data is not time consuming and same time it is very easy to follow. But there are many drawbacks associated with this method. Deleting records may result in losing some information. If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something.

Data imputation is another method of handling missing values. By using this method, we try to fill missing values in the records and attributes. This method is quite useful because by following this method we can make sure we have all the information from responders. There are different approaches suggested for Data imputation in handling attributes with missing values in the data set. One approach among them is the use of attributes mean to fill in the missing values when the attribute type is numerical (Han and Kamber, 2006; Kantardzic, 2003).

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, the researcher has to continue to encounter missing values in fields, especially in databases with a large number of fields. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, the researcher considered carefully about how to handle the thorny issue of missing data (Larose, 2005)

Having efficient methods to fill up missing values extends the applicability in terms of accuracy for many DM methods. The accuracy of the tool is increased and with a larger training set better rules and decision trees can be developed which contributes towards better classification of the data to predict the pattern of fertility in Ethiopia, particularly in HDSS areas.

After ignoring attributes that have no DM value, the remaining attributes were checked for missing

values, inconsistencies and other interpretable observations. The data collected had a small number of variables/attributes with missing values.

A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the data. Further, it seems like a waste to omit the information in all the other fields, just because one field value is missing. Replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables) (Larose, 2005). Therefore, in this research study the investigator tried to handle the missing values by replacing missing value with the field mean, since they are numerical attributes. Table 4.12 summarizes attributes and percentage (%) of missing values associated with each other.

Research Center	Attribute Name	No. of missing values	% of missing values	Mean value of missing values
Dabat	Occupation	662	4.93	7
Dabat	Educ_status	200	1.5	4
Gilgel Gibe	Educ_status	110	0.82	4

Table 4.12. Missing values and their percentage

As a result, the missing values of the dataset were handled in accordance with the above suggestion. The missing value of Occupation and Educ_status attributes were filled by their mean values since they are numeric value type.

4.3.3. Data transformation and reduction

The data may also need to be transformed into forms appropriate for mining. The process of data transformation might include smoothing (e.g. using bin means to replace data errors), Normalization, where the attribute data are scaled so as to fall within a small specified range (scaling the data inside a fixed range), and Attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process (Chakrabarti, et al, 2009).

In data transformation, the data were transformed or consolidated into forms appropriate for

mining. Data transformation can involve the following:

- Smoothing: this works to remove noise from the data. Such techniques include binning, regression, and clustering.
- Generalization of the data: where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.
- Normalization: where the attribute data are scaled so as to fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0. Attribute construction (or feature construction): where new attributes are constructed and added from the given set of attributes to help the mining process.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results (Han and Kamber, 2006). Data transformation is necessary for two purposes to fix problems with the data such as missing values and categorical variables that take on too many values, and to bring information to the surface by creating new variables to represent trends and other ratios and combinations.

In addition, the following data transformation and reformatting operation had employed in order to create new attributes from the existing ones and to reformat the original values of some attributes in the dataset selected for analysis.

A. Creating the Class attribute: it is not included in the original dataset instead Birth_num (total number of births per woman) as an attribute. This derived attribute is a dependent variable can help to classify individuals in to different groups. This classification would help to predict rate of fertility for a woman at her childbearing age. Therefore, the class attribute has derived from the birth_num (Number of births per woman)
B. Creating the Age attribute: in the original database AGE is not included as an attribute in the HDSS databases of the sites, instead the researcher used Date_of_birth as an attribute. However, using AGE as an input (independent) variable can help to categorize the individuals in to different age groups. This categorization would help to identify fertility

patterns in women with different age groups. Therefore, the age attribute has derived from the date_of_birth (Date of Birth).

In this research "mother age" attribute was discretized to reduce the unlike values of the attributes in order to obtain Knowledge (patterns), and to make the data set suitable for mining tools. Then this attribute is later discretized in to seven bins. Table 4.13 shows the discretized labels of mother's age attribute.

RECODE of age_latest	Freq.	Percent	Cum.
14–19	20,297	29.83	29.83
20–24	11,208	16.47	46.31
25–29	9,313	13.69	60.00
30–34	8,076	11.87	71.87
35–39	7,890	11.60	83.47
40–44	5,506	8.09	91.56
45–50	5,743	8.44	100.00
Total	68,033	100.00	

Table 4.13. Age group and frequency of women in that age group

Therefore, the researcher discretized some attributes by converting numeric attributes to nominal: specify which attributes, number of bins, whether to optimize the number of bins, and output binary attributes in order to get the best-fitted model. Table 4.14 provides summary of the original attributes and derived attributes with their values.

No.	Original Attribute	Derived Attribute	values
1	Date_of_birth	Age	numeric
2	Birth_num	TFR	nominal

Table 4.14. Original attributes and derived attributes with their value type

4.3.4. Data Preparation for Weka software

Weka needs the data set to be prepared in some Weka understandable formats. The researcher first has transformed the original different formats into mysql data warehouse. Then the data from mysql has exported to CSV format. Then preprocessing activities are performed in Stata and the file is saved into Weka acceptable comma separated values (CSV) or comma delimited file format. Weka native data format is known as the ARFF (Attribute Relation File Format). It is basically a CSV (comma separated value) format with some extra headers to specify what type each attribute is (numerical, binary, nominal). The CSV file format is converted into ARFF by using Weka

mining software, to take advantage of easier data manipulation and also compatible interaction with Weka software. During scan of the preprocessed data some basic statistics summary will be produced for each attributes. For categorical attributes, the frequency for each attribute value is shown. Moreover; the data that was converted into ARFF format is passed through important steps of data preprocessing mentioned in the previous section.

4.3.5. Setting the class attribute

In decision tree classification technique, which is supervised learning, predefined classes are required in order to train and build classification models. The setting of predefined class is done intentionally because the employed technique for this study is decision tree classification. In order to classify records into different classes the target attribute selected in this research is TFR. By discussing the matter with domain experts and with the coordinators of the HDSS project, the researcher come up with three different values for TFR - low TFR, normal TFR and high TFR. Therefore, the attribute is the dependent attribute while the rest of the variables are the independent attributes for this particular study.

4.3.6. Data type conversion

Before proceeding into building the models, some attributes and class attributes that were in numeric type were converted into nominal using Weka's attribute type converter in order to enable Weka's implementation of decision tree classifier and rule induction algorithm.

4.4. Model Building

According to the selected methodology of CRISP_DM the next step following data preparation is model building. The major activities model building includes selection of modeling technique, generating test design, building model and model assessment.

4.4.1. Selection of modeling technique

The initial step in model building is selection of specific modeling technique. The selection of modeling technique is based on the objective formulated. Since the purpose of this research is to develop a predictive model for fertility rate, classification algorithms has been used for building

the model. The analyses were performed using WEKA environment. Inside the Weka system, there exist many classification algorithms which can be classified into two types; rule induction and decision-tree algorithms.

Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IF-THEN rules. Meanwhile, decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class outputs (Witten and Frank, 2005). Decision trees are one of the most widely used and practical forms of machine learning and data mining. Decision tree models are built by a process that is known as recursive partitioning. The classification algorithms used in this study were J48 and PART.

The J48 algorithm is the Weka implementation of the C4 top-down decision tree learner proposed by J. Ross Quinlan. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. The C4 decision tree algorithm is used in this research. This algorithm is implemented altering parameters such as confidence factor, pruning and unpruning, changing the generalized and binary split decision tree classification options as shown in figure. 4.3. It is important to understand the variety of options available when using this algorithm, as it can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in some cases, each choice may require some consideration (Witten and Frank 2005; Han and Kamber, 2006).

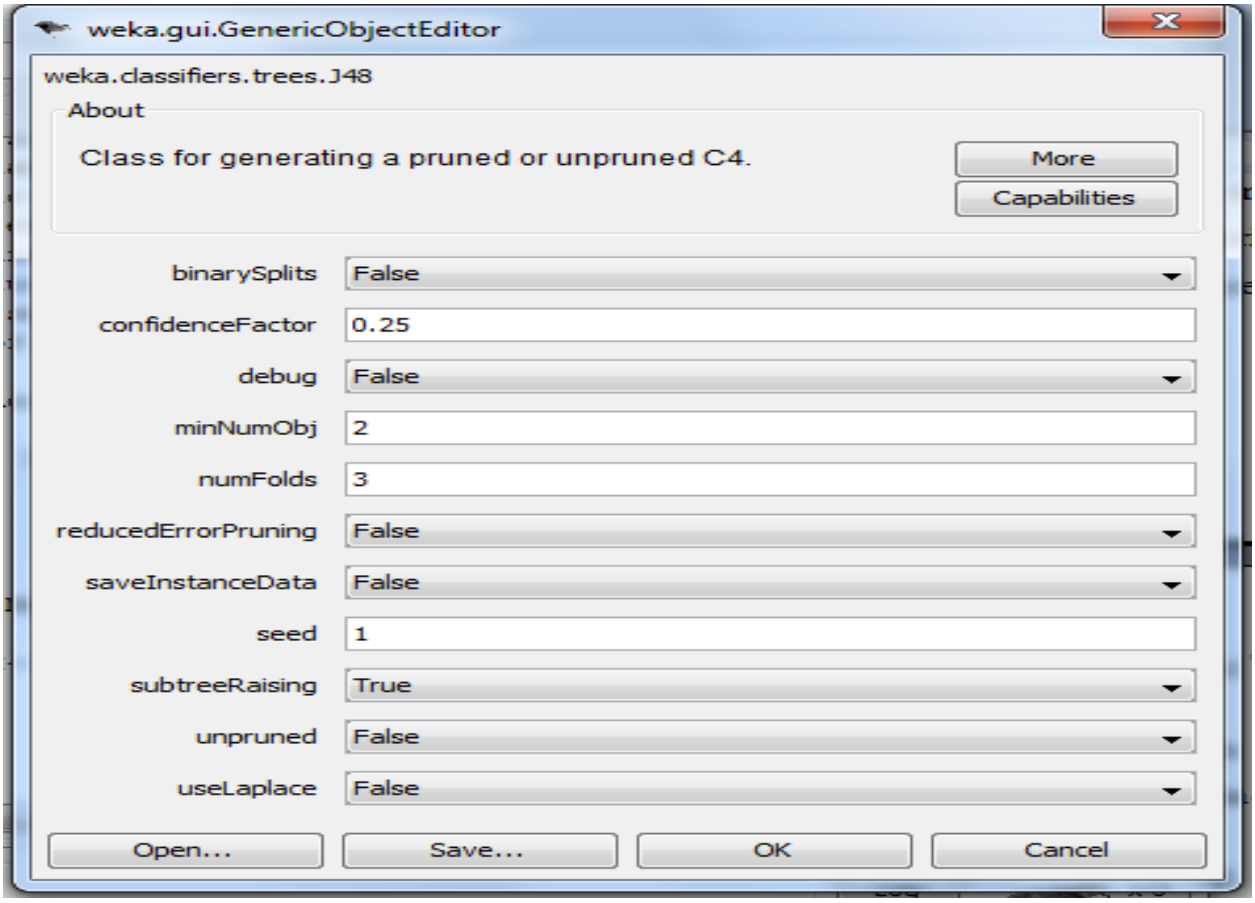


Figure 4.3: J48 Classifier Parameters Window in Weka software

The J48 algorithm gives several options as shown in figure 4.3 in related to tree pruning. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over-fitting. The basic algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular habit of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy (Witten and Frank 2005).

J48 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf -- basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards

toward the root. The second type of pruning used in J48 is termed subtreeraising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtreeraising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtreeraising can be somewhat computationally complex (Witten and Frank 2005).

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over-fitting. This approach is known as reduced error pruning. Though the method is straight-forward, it also reduces the overall amount of data available for training the model. For particularly small datasets, it may be advisable to avoid using reduced error pruning (Witten and Frank 2005).

Other error rate methods statistically analyze the training data and estimate the amount of error inherent in it. The mathematics is somewhat complex, but this approach seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires a confidence threshold or Confidence Factor, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the training data is expected to conform fairly closely to the data you'd like to test the model on, this figure can be lowered. The reverse is true if the model performs poorly on new data; try decreasing the rate in order to produce a more pruned (i.e., more generalized) tree (Witten and Frank 2005).

There are several other options that determine the specificity of the model. The minimum number of instances per leaf is one powerful option (MinNumObj). This allows us to dictate the lowest number of instances that can constitute a leaf. The higher the number, the more general the tree, lowering the number will produce more specific trees, as the leaves become more granular. The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option

effectively treats the data as a nominal value. Turning this encourages more generalized trees. There is also an option available for using Laplace smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities (Witten and Frank 2005).

The most basic parameter is the tree pruning option. Depending on how the training and test data have been defined that the performance of an unpruned tree may superficially appear better than a pruned one. As described above, this can be a result of over fitting. It is important to experiment with models by intelligently adjusting these parameters. Often, only repeated experiments and familiarity with the data will tease out the best set of options (Witten and Frank 2005).

4.4.2. Generation of test design

Numerous measures are used for rule evaluation in machine learning and knowledge discovery. In classification rule induction, the most frequently used measure is classification accuracy. Other standard measures include precision and recall, sensitivity and specificity.

Prior to building a model, a procedure needs to be defined to test the model's quality and validity. In supervised data mining tasks like classification, it is common to use error rates as quality measures for data mining models. Therefore, the test design specifies that the dataset should be separated into training and test set, the model is built on the training set and its quality estimated on the test set. The process of building predictive models requires a well-defined training and validation protocol in order to ensure the most accurate and robust prediction (Two Crows Corporations, 2005).

WEKA mining software has the facility to extract a random sample and then test the accuracy of the classifier on disjoint collection of cases. For this study 10 -fold cross validation has been used. To evaluate the robustness of the classifier, the normal methodology is to perform cross validation on the classifier. 10-fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier (Witten and Frank, 2005).

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

This chapter is devoted to discuss on the models to be built and experiments carried out together with their analysis. The experiments were run on a larger dataset in order to address the main objectives of the research study with respect to the minimum data set that consists of 9 attributes. This will help in understanding the different stages that were used in various DM algorithms.

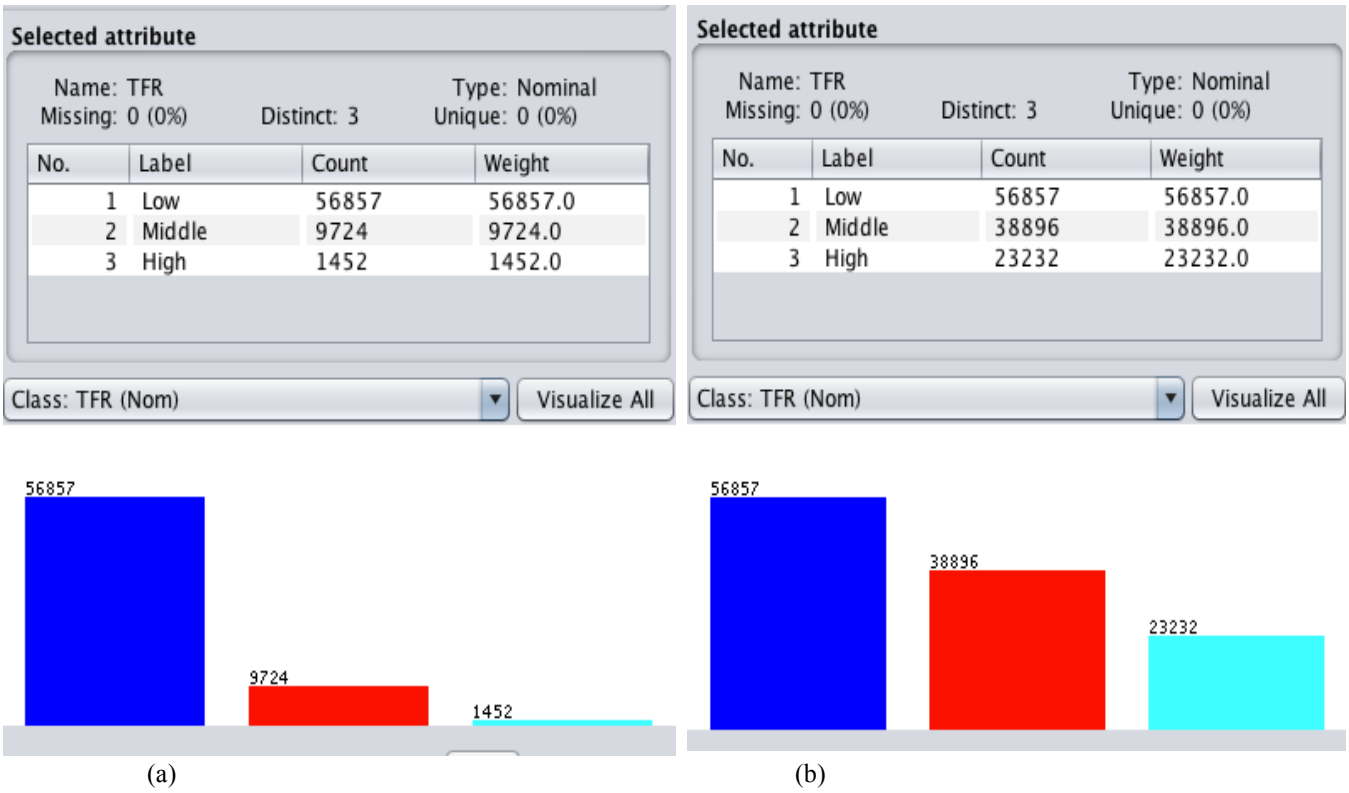
In this study an attempt was made to design a model that enables to predict the pattern of Fertility in Ethiopia. To this end, classification algorithms such as J48 decision tree and Naïve Bayes classifiers were selected and experimented on HDSS data warehouse was consulted to extract the dataset required for training and testing the models created by the classifiers. For creating predictive model, a total size of 68,033 records were used for training and testing. The validations were done using 10-fold cross validation and 90% split test option.

5.1. Dataset Preparation

The data collected for this research project came from different HDSS research centers in different format, CSV, dbase, SPSS, etc. format. Then the researcher develops a data warehouse and transform each dataset into the warehouse. The dataset initially had more than 80 attributes and 318,112 records but after preprocessing stage, it was reduced to 9 attributes and 68,033 records for building the predictive model for fertility rate. Preprocessing was computed on STATA software that was extracted from HDSS data warehouse. Then after, the preprocessed dataset converted to Comma Separated Values (.csv) and then Attribute Relation File Format (.arff) which was compatible with WEKA software for model building.

A dataset of HDSS was imbalanced if the classification categories are not approximately equally represented (Nitesh et. al., 2007). Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large. In the case of HDSS fertility data, the class variable status has a higher imbalance. Therefore, the researcher used Synthetic Minority Oversampling Technique

(henceforth SMOTE) automatic operation by filter where minority classes are over sampled by generating synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases (Selam, 2011).



(a) (b)

Figure 5.1: Side by side view of the class variable: (a) Original data; (b) Balanced data using SMOTE.

Figure 5.1 shows a side by side review of the class attribute status after SMOTE operation applied to the minority classes. Originally there were 56,857 records in the majority class and only 9724 and 1452 records in the minority classes but after applying SMOTE the records of the minority classes increased into 38,896 and 23232.

5.2. Model Building

The researcher took 9 attributes: 8 features and 1 class attribute for building predictive model. The selection of attribute was made using subjective judgment of the investigator, reviewing literature and discussion with public health professionals from different research centers in different universities. To build the predictive model, the. arff and/or .csv format of the selected dataset was given to WEKA. The researcher used J48 decision tree algorithms and Naïve Bayes classifiers.

5.2.1. Building Classification Model using WEKA Software

Classification is learning a function that maps (classifies) a data item in to one of several predefined classes. In relation to this, Chapman, et al. (2000) noted that "when learning classification rules, the system had to find the rules that predict the class-label, which is the dependent or predicted attribute's value, from the independent or predicting attributes' value". Accordingly, it was the concern of this phase to generate classification rules that were assigning the correct class label to previously unseen and unlabeled women. In this phase, the main issue to be discussed was to build a classification model. Due to the fact that well-grown decision trees were as comparably useful as other classifiers, the type of classification model selected to be built was decision tree (Selam, 2011). The other reason for selecting decision tree model is, compared to other classification model types like Neural Network classification, decision tree has a significant advantage because it can be built manually-and so, is easily explained (Theeuwen, et. al., 2001).

Apart from this, decision tree operations are completely interactive, iterative and they benefit from powerful visualization features. To carry out this phase, the well-known software, WEKA 3-8-0 had been attempted to use. Therefore, further discussion in this investigation was done by making use of the models found from the selected WEKA's algorithms runs.

5.3. J48 Decision Tree Model Building Using WEKA Software

The J48 decision tree C4 algorithm builds decision trees from a set of predefined training dataset using the concept of information entropy and attribute ordering. It uses the fact that each attribute of the data was used to make a decision by splitting the data into smaller subsets. As decision tree is a classifier, any previously unseen record with the required degree of attributes was fed into the tree. At each node, it will be sent either left or right to some test. Finally, it will reach a leaf node and be given the label associated with that leaf. At this junction, the researcher interested to generating rules of assigning the women at child bearing age of HDSS members to the class they belong.

The classification C4 algorithms were implemented in WEKA 3-8-0 to build a classification model in such a way that testing the model would be possible after training it. For most of the experiments

carried out in this phase, the experiments were 10-fold cross validation and Percentage (%) split test options were used, the total record was partitioned in to two, the training and test datasets. These two datasets were found from the final data set by using a stratified sampling technique where the different classes, found in the classification, were considered as strata for 10-fold cross validation.

The reason behind applying stratified sampling technique was, partitioning the total record where the contribution of each class in the resulting datasets was proportional. To avoid the problem of over-fitting the researcher tried to experiment by pruning the tree, 10% of the total record was selected as a test sub-data set and the remaining 90% as a training sub-data set.

5.3.1. The WEKA Decision Tree Experiment and Analysis

WEKA 3-8-0 supports many types of classification algorithms. Among the classification algorithms that WEKA 3-8-0 supports, the J48 C4 algorithms was used with different input parameters as well as different types of related classifiers. J48 algorithm is WEKA's implementation of the C4 decision tree learner. The corresponding algorithm used to extract rules from the decision trees is J48 or PART.

By making use of WEKA 3-8-0 a total of 15 experiments were carried out, where 4 of the experiments were for constructing decision trees with 10-fold cross validation, 5 were different values of percentage split test, 4 were for constructing decision trees with 85% split test and the remaining 2 experiments were for Naïve Bayes classifier with or without supervised discretization respectively. In relation to this, J48 was the algorithm used to construct the decision trees in the 13 experiments. The extraction of the corresponding rules, in the remaining 13 experiments, from the decision trees was managed using J48 or PART.

To display the run parameters and the outputs of the respective experiments, three tables (Table 5.1, Table 5.2 and Table 5.3) were used. As displayed in all tables, the different experiments were carried out by using all the 9 attributes of the records with different schemes were applied in the experiment and two different test modes (ways of feeding records to the algorithms). Analysis of the J48 decision tree predictive model were made in terms of detailed accuracy, precision, recall,

F-measure and ROC curve of the classifier based on a confusion matrix of each predictive model resulted of different classes (Low, Medium and High classes in this research thesis). The experiment number from 1 up to 4 were applied on the 10-fold cross validation test mode, 5 experiments (i.e. from 5 up to 9) were relied on different value of percentage split test mode and the remaining 4 experiments (i.e. from 10 up to 13) were relied on 85% split test mode. The experiments for J48 decision tree classification models are listed under beneath in Table 5.1.

Experiment 1	pruned J48 decision tree with default confidence factor (i.e. 0.25), with 10-fold cross validation test mode.
Experiment 2	pruned J48 decision tree with default confidence factor (i.e. 0.50), with 10-fold cross validation test mode.
Experiment 3	Unpruned J48 decision tree with default confidence factor (i.e. 0.25) and with 10-fold cross validation test mode
Experiment 4	unpruned J48 decision tree with confidence factor (i.e. 0.50) and with 10-fold cross validation test mode
Experiment 5	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 66% split test mode
Experiment 6	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 70% split test mode
Experiment 7	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 80% split test mode
Experiment 8	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 85% split test mode
Experiment 9	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 90% split test mode
Experiment 10	Unpruned J48 decision tree with default confidence factor (i.e. 0.25) and 85% split test mode
Experiment 11	Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 85% split test mode
Experiment 12	Unpruned J48 decision tree with confidence factor 0.50 and 85% split test mode
Experiment 13	Pruned J48 decision tree with confidence factor 0.50 and 85% split test mode

Table 5.1. List of experiments conducted using J48 decision tree

These experiments were analyzed to compare them in terms of different performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC

curve. The models were also compared with regard to the pattern or KD of the predictive model. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. The sensitivity is the measure of the ability of a prediction model to select instances of a certain class from a data set. The specificity corresponds to the true negative rate which is commonly used in two class problems (Emamu, A., 2011).

As an outcome of different combinations of these scheme and J48 algorithm was used in the experiments. These combinations were: J48 decision learner algorithm and J48 and/or PART decision rule extractor algorithm.

The two J48 decision tree test modes were:

- T1 - Inputting all the records with a 10-fold cross-validation test mode, and
- T2 - Inputting all the records Percentage (%) split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model.

Furthermore, the result obtained from these experiments was summarized in table 5.2 with respective performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC curve.

S. No	**Comparing parameters**	**Experiments' No.**			
		1	**2**	**3**	**4**
1	Testing Mode	T1	**T1**	T1	T1
2	Pruning	No	**Yes**	No	Yes
3	Confidence Factor	0.25	**0.25**	0.50	0.50
4	Size of Tree	4233	**2192**	**4233**	3092
5	No. of Leaves	3596	**1839**	**3596**	2608
6	Time Taken (sec.)	0.33	**0.4**	**0.34**	0.4
7	Precision	0.775	**0.776**	**0.775**	0.773
8	F-Measure	0.766	**0.767**	**0.766**	0.765
9	Sensitivity	0.763	**0.764**	0.763	0.763
10	Specificity	0.882	**0.881**	0.882	0.881
11	ROC area	0.903	**0.898**	0.903	0.901
12	Accuracy (%)	76.332	**76.4**	76.332	76.2

Table 5.2: Input parameters and the resulting J48 Decision Trees' with 10-fold CV test mode.

As can be observed from Table 5.2, except with the number of leaves and the size of tree, they all have a comparable result in all the other measures. Experiment 2 has significantly smaller number

of leaves and size of tree in comparison with the others. Therefore, experiment 2 outperforms than the other experiments in performance for HDSS datasets.

The data collected, preprocessed and analyzed using classification (J48 decision tree) was presented in the below Table 5.3. The researcher tried to classify with different values of percentage (%) split test parameters of trained and tested data to look the performance of the system. The following are some samples of the experiments

Experiment No.	Split Test Mode in %	Accuracy in %
5	66	75.79
6	70	75.82
7	80	75.85
8	85	75.969
9	90	75.761

Table 5.3: Input parameters and resulting J48 DT with different percentage split test mode.

As it can be observed from the above table, the 85% split test of data for training is better than the other percentages split test options. The selected percentage has 75.97% correctly classified instances. Percentage split test parameter of 80% training set has also 75.85 % correctly classified instances, but with relatively low precision, recall, F-measure and ROC curve. Moreover, the J48 decision tree model produced was from the table 5.3 has experiment 2 which has 85% split test mode which is train a model and then supply the unseen remaining part of the record for testing the performance of the model and its accuracy level was 75.97%. The 85% test option mode shows a better performance than other experiments.

The percentage split test option was used to partition the dataset into training and testing data and this parameter was set to 85, 85% for training and 15% for testing. The result of this learning scheme was summarized and presented in Table 5.3. Moreover, the result obtained from these experiments is summarized in table 5.4 with respective performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC curve.

S. No	**Comparing parameters**	**Experiments' No.**			
		10	**11**	**12**	**13**
1	Testing Mode	T2	**T2**	T2	T2
2	Pruning	No	**Yes**	No	Yes
3	Confidence Factor	0.25	**0.25**	0.50	0.50

4	Size of Tree	4233	**2192**	4233	3092
5	No. of Leaves	3596	**1839**	3596	2608
6	Time Taken (sec.)	0.5	**1.66**	.44	0.9
7	Precision	0.776	**0.770**	0.775	0.773
8	F-Measure	0.767	**0.763**	0.766	0.765
9	Sensitivity	0.763	**0.760**	0.763	0.763
10	Specificity	0.763	**.760**		0.880
11	Accuracy (%)	76.345	**75.969**	76.345	76.333

Table 5.4: J48 Decision Trees' with 90-percentage split test mode parameters.

As can be observed from this Table 5.4, experiments 11 has comparatively better than the other experiments with extracted accuracies. This is because, the researcher used to build J48 decision tree with default confidence factor (i.e. 0.25) and 85% split test mode and also the pruned parameter of the classifier.

The model has accuracy of 76.4% using 10-fold cross-validation and 76% accuracy using 85% split test options. Moreover, the model has almost similar true positive rate and false positive rate for both 10-fold cross validation and 85% split test.

The best J48 decision tree model of the classification generated from experiment 2 of the 10-fold cross-validation mode. The model shows a better performance evaluation than other models. The 10-fold cross-validation model also scored a better performance than 85% split test. Therefore, the test options mode used to build the decision tree for experiment 2 10-fold cross-validation mode options which is J48 pruned decision tree with default confidence factor (i.e. 0.25), are statistically significant in splitting the decision tree. Furthermore, suggestions gathered from the domain experts from different HDSS research centers in different universities and literatures indicated that these attributes have a great role in the prediction tasks.

5.3.2. Confusion Matrix for J48 Decision Tree Model

A confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models. Moreover, the overall predictive accuracy on unseen instances it is often helpful to see a breakdown of the classifier's performance.

=== Confusion Matrix ===

```
   a     b     c   <-- classified as
45370 10535   952 |    a = Low
 5632 25770  7494 |    b = Middle
  364  3565 19303 |    c = High
```

The entries in the confusion matrix have the following meaning:

- 45370 is the number of **correct** predictions that an instance is **High**,
- 10535 is the number of **incorrect** predictions that an instance is **Middle**,
- 952 is the number of **incorrect** predictions that an instance is **Low**,
- 5632 is the number of **incorrect** predictions that an instance is **High**,
- 25770 is the number of **correct** predictions that an instance is **Middle**,
- 7494 is the number of **incorrect** predictions that an instance is L**ow**,
- 364 is the number of **incorrect** predictions that an instance is **High**,
- 3565 is the number of **incorrect** predictions that an instance is **Middle**,
- 19303 is the number of **correct** predictions that an instance is **Low**.

5.3.3. ROC Analysis for J48 Decision Tree Model

ROC curves are similar to lift charts in that they provide a means of comparison between individual models and determine thresholds which yield a high proportion of positive hits. In the below figure the horizontal axis of an ROC graph measures the false positive rate as a percentage. The vertical axis shows the true positive rate. The top left hand corner is the optimal location in an ROC curve, indicating high TP (true-positive) rate versus low FP (false-positive) rate.

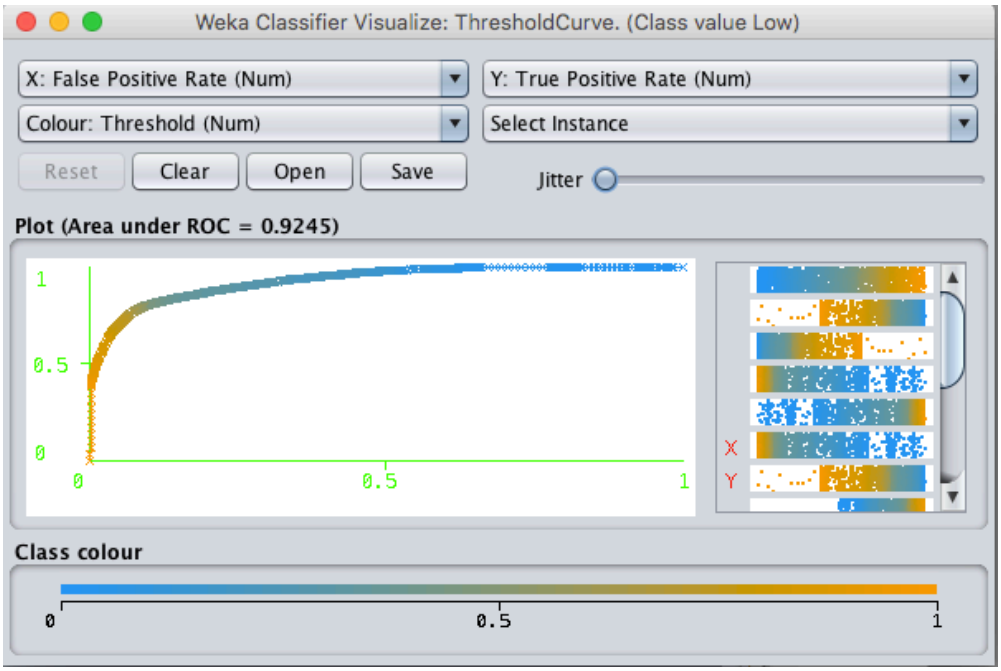


Figure 5.2: ROC curve of the J48 decision tree model

The true positive (TP) rate and false positive (FP) rate values of different classifiers on the same test set are often represented diagrammatically by a ROC Graph. The abbreviation ROC analysis stands for 'Receiver Operating Characteristics Graph', which reflects its original uses in signal processing applications. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. Class value Low gives the ROC accuracy of 92.45%. The larger the area under the ROC curve (AUC), the higher the likelihood that an actual positive case and also the better the model. It will be assigned a higher probability of being positive than an actual negative case. The AUC for the model is 0.9245 which is closer to 1, that is AUC measure is especially useful for data sets with unbalanced target distribution (one target class dominates the other).

Figure 5.4 shows that, the tree view of the predictive model built J48 algorithm with 10-fold cross validation test mode options which is J48 pruned decision tree with default confidence factor (i.e. 0.25) using 9 attributes. For clear understanding of the tree, the run information for the predictive model is annexed at Annex I.

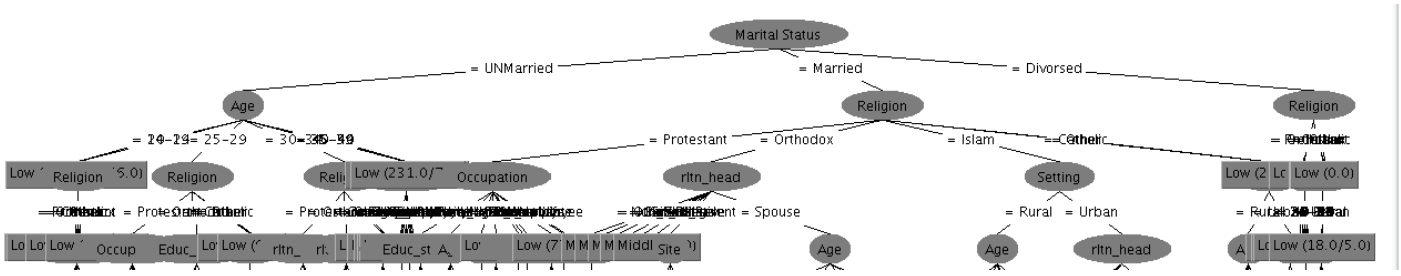


Figure 5.3 Partial tree view of predictive model using 10-fold cross validation mode

5.4. Naïve Bayes Classifier Model Building using WEKA Software

It is method of classification that does not use rules, a decision tree or any other explicit representation of the classifier. Rather, it uses the branch of Mathematics known as probability theory to find the most likely of the possible classifications. The Naïve Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which the researcher used to calculate the probability of each of the possible classifications in turn. Having done this the researcher chooses the classification with the largest value. Taking into account the nature of the underlying probability model, the Naïve Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect (Berhane and Byass, 2003).

In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data. However, various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains.

The Naïve Bayes (Han and Kamber, 2006) classifier provides a simple approach, with clear semantics, to representing and learning probabilistic knowledge. It is termed naïve because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process. Two experiments were done with the Naïve Bayes algorithm using different combination and inputting all records with an inputting 10-fold cross validation test with and without supervised

discretization parameter, which is train a model and then supply the unseen remaining part of the record for testing the performance of the model. Table 5.5 summarizes the results with respective performance matrices values, accuracies, discretization parameter, time taken in sec. in the execution, precision, F-Measure, sensitivity, specificity and ROC curve.

- Experiment 14: T3-Naïve Bayes classifiers with 10-fold cross-validation mode without supervised discretization.
- Experiment 15: T3-Naïve Bayes classifiers with 10-fold cross-validation mode with supervised discretization.

No.	Computing Parameters	Experiment No	
		13	14
1	Testing Mode	T3	T3
2	Discretization	No	Yes
3	Time Taken (Sec.)	0.04	0.21
4	Mean absolute error	0.2227	0.2227
5	Precision	0.714	0.714
6	F-Measure	0.694	0.694
7	Recall	0.691	0.691
8	ROC Area	0.861	0.861
9	Accuracy (%)	69.088	69.088

Table 5.5: Summary of Naïve Bayes Experiment Results

It is quite imperative to see the results generated by the Naïve Bayes by changing the parameters of using with discretization or without discretization or the display mode in old and new formats. Except there is a difference in the time taken to build the model; all the parameters have the same learning capacity for the datasets trained and tested. This necessitates comparing the result of J48 and Naïve Bayes.

As can be observed from the above table 5.5, the model scored in Naïve Bayes was 69.088% accuracy and 69.1% true positive rate using 10-fold cross-validation test mode without or with supervised discretization and the model.

5.4.1. ROC Analysis for Naïve Bayes Classifiers

ROC analysis is performed by drawing curves in two dimensional spaces, with axes defined by the True Positive rate and False Positive rate, or equivalently, by using terms of sensitivity and specificity. That is, the y-axis represents Sensitivity = True Positive rate, while the x-axis

represents 1−Specificity = False Positive rate. The AUC for the pattern of fertility rate records generated from the Naïve Bayes Classifier is presented in the below figure 5.5. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. Class value Low gives the ROC accuracy of 89.6%.

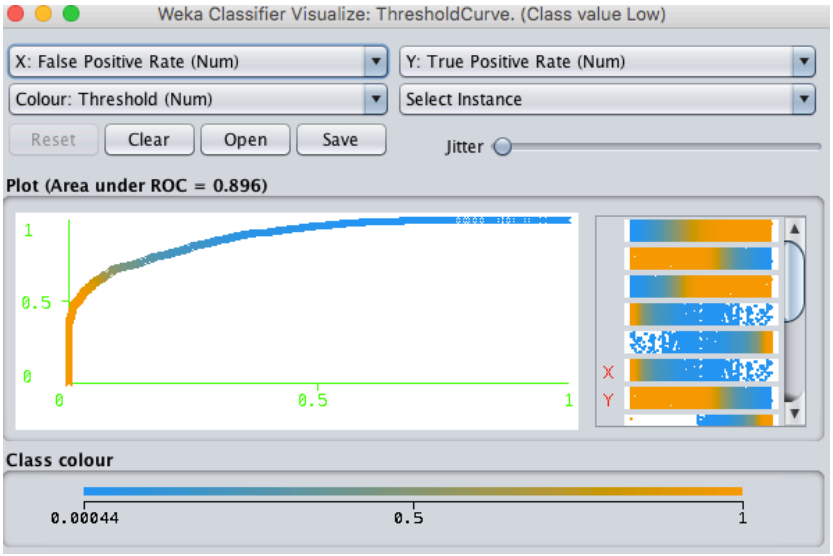


Figure 5.5: ROC curve from the Naïve Bayes Classifier

In the above figure the horizontal axis of an ROC graph measures the false positive rate as a percentage. The vertical axis shows the true positive rate. The top left hand corner is the optimal location in an ROC curve, indicating high TP (true-positive) rate versus low FP (false-positive) rate. ROC plots allow for visual comparison of several models (classifiers). For each model, the researcher calculated its sensitivity and specificity, and draws it as a point on the ROC graph.

5.5. Comparison of J48 and Naïve Bayes models

Comparison of the two models is made in terms of the general model accuracy, detailed accuracy by class such as the precision, ROC Area, recall and the rules generated for interpretation. The following table 5.6 gives the relative comparison between the two models.

Performance Testing	**J48 Decision tree**	**Naïve Bayes**
Accuracy (%)	76.4	69.088
Av. Precision	0.776	0.714
Av. Recall	0.764	0.691

Av. True Positive (TP) Rate	0.760	0.691
Av. False Positive (FP) Rate	0.121	0.148
Av. ROC Area	0.901	0.861

Table 5.6 comparison of J48 and Naïve Bayes models

Table 5.6 shows there is a relative better model prediction in the case of J48 in correctly identifying the dataset. The ROC Area for Naïve Bayes indicates 0.86 lower when compared with the ROC Area under J48 which accounts 0.901. This signifies the number of correctly classified datasets are higher in the model built by J48 than the Naïve Bayes. The overall model accuracy of J48 (76.4%) shows it has better prediction. The relative better performance of J48 algorithm can be attributed to the nature of the data such as the handled missing values; the data consistency etc. Naive Bayes has a better prediction if the attributes are conditionally independent to each other. For the given data under study J48 has shown better accuracy and the rules generated by this model are used for interpretation. It is worth mentioning however, Naïve Bayes is also a candidate to be used for predicting even though its performance is relatively low.

And also, a comparison of the performance evaluation in the table 5.5 between J48 decision tree algorithm and Naïve Bayes classifier are illustrated in figure 5.6.

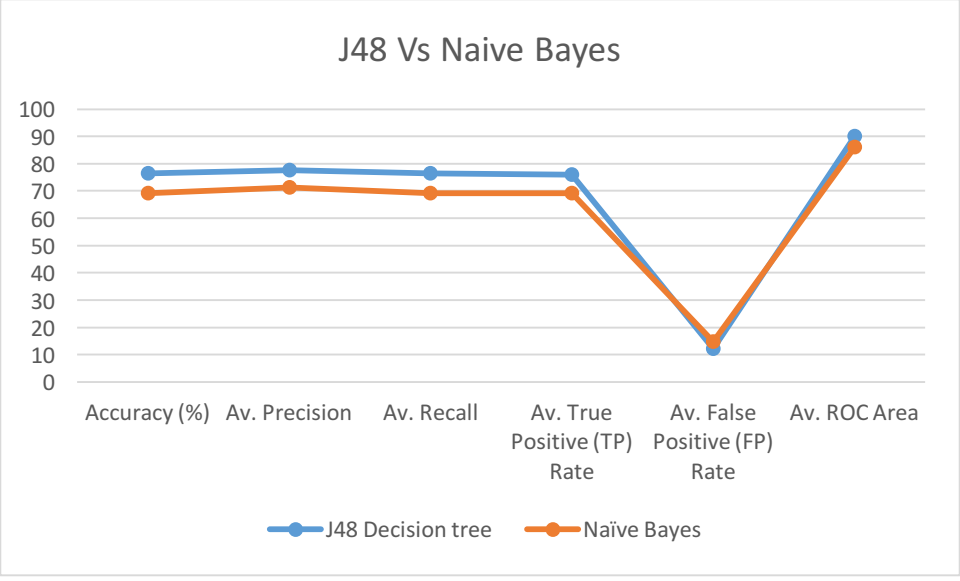


Figure 5.6: Bar Graph Visualization of Performance comparison of J48 Decision Tree and Naïve Bayes classifier with 90% split test mode.

One of the purposes of this study was to compare the J48 decision tree algorithm and Naïve Bayes

classifier DM model and to select the one, which performs the best. Accordingly, each experiment carried out in this research had employed both J48 decision tree and Naïve Bayes classifier. In all experiments the same datasets were used. The output of these experiments indicated that the classification task of records using the fertility rate dataset from HDSS research centers.

In Figure 5.6 the researcher visualized the line graph of the performance evaluation obtained for the different tools. The highest accuracy is found by the J48 decision tree method. Thus, it is considered also the base case. All the J48 decision tree algorithm tools tested have performed much better than the Naïve Bayes classifier method.

The result scores of the Naïve Bayes classifier for time taken to execute the model have better than the J48 decision tree model. However, the overall result scores of the J48 decision tree model higher than that of the Naïve Bayes classifier model. In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, Time taken for execution, Precision, recall and ROC Area). The results were achieved using inputting 10-fold cross-validation. In comparison to the above studies, the researcher found that the predictive model achieved a classification accuracy of 0.764 with a precision of 0.776, time taken to execute in sec. has 1.08, Recall is 0.764, and ROC Area of 0.901.

5.6. Generating Rules

From the decision tree developed in the aforementioned experiments, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node (Bao, 2003).

This produces rules that are unambiguous in that it doesn't matter in what order they are executed. Decision tree and decision rule solutions offer a level of interpretability that is unique to symbolic models. The solutions may be directly inspected to understand the decision surfaces that exist in the data (Apte and Weiss, 1997).

PART is a separate-and-conquer rule learner proposed by (Witten and Frank 2005). The algorithm generates sets of rules called ‘decision lists’ which are ordered set of rules. PART builds a partial C4.5 decision tree in each iteration and converts the "best" leaf into a rule. The parameters mentioned in J48 decision tree algorithm are also applicable here.

When the researcher compared the performance measure as well as the results obtained from both decision tree algorithms i.e. J48 and PART models are nearly equal performance. In terms of accuracy, execution time, AUC, sensitivity and specificity; PART is slightly better than J48. Therefore, the researcher selected the PART algorithm for generating better rules. The following are some of the rules extracted from the PART listed below and some of the rules supposed to be interesting and are selected by domain experts as well as from the literatures, are presented as follows:

1. Marital Status = UNMarried, | Age = 14-19: Low (19974.0/56.0)
2. Marital Status = UNMarried, | Age = 20-24 | | Religion = Protestant: Low (2031.0/269.0)
3. Marital Status = UNMarried | Religion = Orthodox | Occupation = Unemployed: Low (3161.0/273.0)
4. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Child: Low (1222.0/136.0)
5. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Parent: Low (0.0)
6. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Non_Relative: Low (110.0/21.0)
7. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Other_Relative: Low (85.0/4.0)
8. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Sibling: Low (39.0/1.0)
9. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Grand_Parent: Low (94.0/1.0)
10. Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Head: Low (0.0)

The first rule shows that if the individual marital status is Unmarried and at the age of between 14 and 19 then the birth rate is low. The third raw shows that if religion of the woman is protestant and she is married and her education status is only read and write and age is between 30 and 34 then the fertility rate is medium. Rule 20 shows that if site is gilgel gibe and the woman is wife and setting is rural and age is between 30 and 34 and religion is Islam and education status of the woman is only read and write then the probability is high.

There were also some interesting outputs: those whose age is b/n 30-34 and educ_status is 4-6 and their Religion is Catholic and who are living in rural area have high TFR. But those whose age is b/n 30-34 and educ_status is 4-6 and their Religion is Protestant and who are living in rural area have Middle TFR

It can be revealed from the above rules that most important variables for building model to classify the pattern of fertility rate in Ethiopia were site, relation to head, the setting, age and education status. Therefore, these attributes play a significant role in classifying records at the higher level of the tree which indicates their statistical significance than other variables occupation.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

Machine Learning algorithms are improving as the number of DM tools, techniques and algorithms increase. Healthcare data is a good test bed for DM. A great deal of data in health care is still being gathered and organized using pen and paper. Indeed, the data contains and reflects activities and facts about the organization. But, the data's hidden value, the potential to predict health trends, has largely gone unexploited. The increase in data volume causes great difficulties in extracting useful information and knowledge for decision support. It is to bridge this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as DM or KDD has emerged in recent years.

The application of DM technology has increasingly become very popular and proved to be relevant for many sectors such as health related sectors. Particularly, in the public health, DM technology has been applied for predicting the pattern of fertility rate for effective and efficient predictive model, determinants and patterns that contributes to the patterns of fertility rate.

This research has tried to assess the application of DM technology to predict the pattern of fertility rate in Ethiopia, for developing a classification model. Such a classification model could enable the public health department in each university in the HDSS project as well as for the governmental and non-governmental organizations to implement predictive model in Ethiopia.

This investigation, conducted according to the hybrid KDP model, was carried-out in six major parts namely: business understanding, data understanding, data preparation, model building, evaluation, and use of the discovered knowledge. However, since a DM task is an iterative process, these steps were not followed strictly.

Before applying the DM model the investigator developed a data warehouse to handle the data collected from four different data sources and the data required for this study is extracted from the

data warehouse. A data set with 118,985 total HDSS members' records was used to develop a classification model. Since, this research was intended to fill a gap left by a related research; some valuable experiences of the previous research were used.

In the classification phase, the J48 decision tree algorithm and Naïve Bayes classifier, which is WEKA's implementation of the C4 algorithms, were used. From the study, rltn_head, occupation, religion, educ_status, site, setting and age attributes were the significant DM and Public Health values.

In order to select a classification model that can classify the HDSS members, the models were built by employing the J48 decision tree algorithms. In the decision tree selection process, more emphasis was given to important attributes to be used, the number of records considered and the size of the tree and the corresponding number of rules extracted from the tree.

From the experiments done using a WEKA version 3-8-0, it was observed that, for a given number of attributes, as the number of records used to develop a decision tree increases the corresponding number of rules generated will possibly increase. Due to this observation, not to get a minimum number of rules, among all the models developed for comparison, the models developed from the 118,985 records 10-fold cross-validation option and attribute selection were given due attention. Accordingly, the better J48 decision tree with the corresponding extracted rules was selected as a working model to classify members into their corresponding classes. As a result, the classification accuracy of the selected J48 decision tree seems convincing than the Naïve Bayes classifier. That is, among the 118,985 data inputted to the model learner with a 10-fold cross-validation, 76.4%.

The suggestions and opinions given by domain experts in the entire investigation were observed and found to be very important in the model development process, particularly, in the classification phase. The overall predictive model building process made by employing the J48 decision tree algorithm and Naïve Bayes classifier demonstrated that DM is a method that should be considered to predict the pattern of fertility rate in Ethiopia, particularly for HDSS area.

6.2. Recommendations

This investigation has been conducted mainly for an academic purpose. However, it revealed the

potential applicability of DM technology to classify pattern of fertility rate in the HDSS data warehouse. Moreover, it is the researcher's belief that the contribution of this research work could be a good experience for a competitive study in public health as well as information science sectors in the future.

Apart from this, it is the researcher's faith that the findings of the research would encourage public health sector to work on the application of DM technology to minimize fertility rate, and as a result gain a competitive advantage based on demographic, socio-economical, parental, environmental, and epidemiological factors alone.

Therefore, the researcher strongly recommends the following:

- In this research encouraging results were obtained, further investigation should be done by integrating the numerous fertility related data sources, such as ……...
- Further extensive experiments should be required by using large amounts of dataset and applying different classification techniques.
- There is a need to undertake different DM research investigations based on clinical datasets from different health facilities.
- Further study is recommended to the problem domain specifically and fertility in general that apply those unused DM models, tools and algorithms.

References

- **Aha, D. et. al. (1991)**. Instance-Based Learning Algorithms. Machine Learning 6(1), Washington DC, USA.
- **Amir, F. and Shahram, J. (2011)**. An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network. International Journal of Advanced Science and Technology. Vol. 34, Shiraz University, Shiraz, Iran.
- Apte, C. and Weiss, S. M. (1997). Data Mining with Decision Trees and Decision Rules. T. J. Watson Research Center IBM Research Division York town Heights, NY 10598, New York, USA.
- **Arzucan** Ozg¨ur. (2004). Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization .Turkish: Bo˘gazin University.
- **Berhane, Y. and Byass, P. (2003)**. Butajira DSS Ethiopia, Department of Community Health, Faculty of Medicine AAU and Department of Public Health and Clinical Medicine Umea University, INDEPTH Monograph Volume I Part C.
- **Berhane Y., Fitaw Y., Worku A., (2004)**. Impact of child mortality and fertility preferences on fertility status in rural Ethiopia. East Africa Medical Journal 81(6):300-6. Addis Ababa, Ethiopia.
- **Berry J.A. Michael and Linoff S. Gordon (2004)**. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Second Edition. Wiley Publishing, Inc., Indianapolis, Indiana
- **Berry** MaL, G. (1997). Data mining techniques: For marketing, sales and customer support. New York. John Wiley and Sons, Inc.
- **Bloom**, D.E., and J.G. Williamson. (1998). "Demographic Transitions and Economic Miracles in Emerging Asia." World Bank Economic Review 12 (3): 419–455.
- **Bloom**, D.E., D. Canning, G. Fink, and J. Finlay. (2007). Realizing the demographic dividend: Is Africa any different? Program on the Global Demography of Aging, Working Paper 23. Cambridge, Mass, U.S.A.: Harvard University.
- **Boserup**, E. (1965). The conditions of agricultural growth: The economics of agrarian change under population pressure. Chicago: Aldine. London: Allen & Unwin.

- **Central** Statistical Agency (CSA) and ORC Macro. (2001). Ethiopia Demographic and Health Survey 2000. Addis Ababa, Ethiopia: ; Calverton, Md., U.S.A: Central Statistical Agency; ORC Macro.
- **Central** Statistical Agency (CSA) and ORC Macro. (2006). Ethiopia demographic and health survey 2005. Addis Ababa, Ethiopia; Calverton, Md., U.S.A: Central Statistical Agency; ORC Macro.
- **Central** Statistical Agency (CSA). (2008). Summary and statistical report of the 2007 population and housing census. Addis Ababa.
- **Chang, C. L. (2007)**. A study of applying DM to early intervention for developmentally-delayed children. Expert Systems with Applications. Huwei, Taiwan.
- **Chakrabarti .S ,Earl C., Eibe F., Ralf H.G., Jaiwei H. , Xia J., Micheline K., Sam S. L.,Thomas P. ,Richard E. ,Dorian P., Mamdouh R.,Markus S.,Toby J. and Witten H. (2009)**. Data mining know it all. Morgan Kaufmann Publishers 30 Corporate Drive, Suite 400 Burlington, Unite State
- **Cios**, Pedrycz, Swiniarski, and Kurgan (2000). A knowledge discovery approach to diagnosing myocardial perfusion: IEEE Engineering in Medicine and Biology Magazine, New York, USA.
- **Cios**, K. and Kurgan, L. (2005). Trends in data mining and knowledge discovery. Springer Verlag, London, UK.
- **Cisos,** Wilord Pedrycz, roman W. Swiniarski, and Lukasz A. Kurgan. (2007). Data Mining: a knowledge discovery approach
- **Claire** Norville Rocio Gomez Robert L. Brown "Some Causes of Fertility Rates Movements"< https://uwaterloo.ca/waterloo-research-institute-in-insurance-securities-and-quantitative finance/sites/ca.waterloo-research-institute-in-insurance-securities-and-quantitative finance/files/uploads/files/03-02.pdf> accessed February 15, 2016
- **Colin Shearer. (2000)**. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Ware Housing Volume 5, Number 4.
- **Cover, T. M. and Hart, R. E. (1997)**. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theoo'IT-13(I), London, UK.
- **David**, Hand, Heikki Mannila, and Padhraic Smyth. (2001). *Principles of Data Mining.* The MIT Press, Cambridge, Massachusetts, London England

- **David L. Olson and Dursun, D. (2008)**. Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg
- **Demographic** and Health Survey (DHS). (2010). Measure DHS Demographic and Health Survey Stat-Compiler, Sponsored by USAID. <http://www.statcompiler.com/>. Accessed February 10, 2016.
- **Deogan**. (2001). Data Mining: research Trends, Challenges, and Applications [database on the Internet]. <http://citeseer.nj.nec.com/deogun97data.html>.[Access date January 6,2016]
- **Eastwood**, R., and M. Lipton. (2004). "The impact of changes in human fertility on poverty." Journal of Development Studies 36 (1): 1–30.
- **Emmanuel, N. O. (2007)**. Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. Turks & Caicos Islands Community College.
- Han, J. and Kamber, M. (2006). Data Mining: concepts and Techniques. 2nd ed. Morgan kufman Publishers, San Francisco, USA.
- **Fayyad (1996a)**. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39, 11, 27-34. New York, USA.
- **Fayyad** U, Piatetsky-shapiro, G. and Smyth, Padharic. (1996). From Data Mining to Knowledge Discovery in Databases. [database on the Internet].[Access date January 20, 2016]
- **George, A. (2004)**. Application of Data Mining in Medical Applications. MSc. Thesis, University of Waterloo, Ontario, Canada.
- **Giudici, P. (2003)**. Applied Data Mining Statistical Methods for Business and Industry. John Wiley & Sons Ltd, Chichester , England
- **Han, J. and Kamber, M. (2006)**. Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco
- **Han** JaK, Micheline, (2001). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.
- **Headey**, D. (2011). "Turning Economic growth into nutrition-sensitive growth." Paper presented at 2020 Conference: Leveraging Agriculture for Improving Nutrition and Health, New Delhi, India, February 10-12.
- **Helen T. (2003)**. Application of data mining technology to identify significant patterns in

census or survey data. Masters Thesis Addis Ababa University, Addis Ababa,Ethiopia.

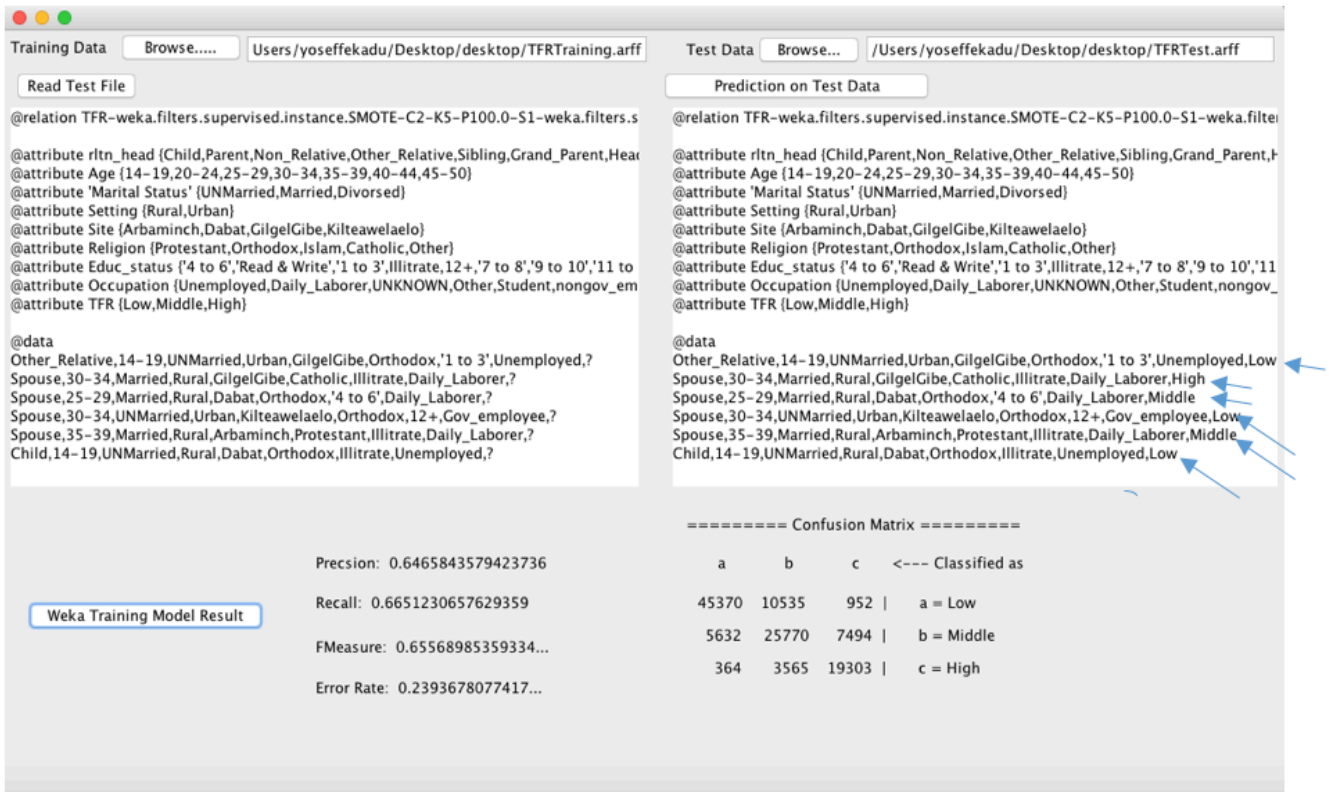
- **Henderson**, J.V. (2010). "Cities and Development." Journal of Regional Science 50 (1): 515– 540.
- **Ian, H.W. and Eibe, F. (2005)**. Data Mining Practical Machine Learning Tools and Techniques. 2nd ed., University of Waikato, New Zealand.
- **International Development Research Centre. (2002)**. Population and Health in Developing Countries. INDEPTH Network 2002, Vol. 1, Ottawa, ON, Canada K1G 3H9. Available URL: http://www.idrc.ca.
- **Jiawie, Han and Micheline Kamber. (2006)**. Data mining Concept and Techniques. 2nd Ed. Morgan Kaufmann Publishers, San Francisco
- **John, G. H. and Langley, P. (1995)**. Estimating continuous distributions in Bayesian classifiers. Proc. of 11th Conference on Uncertainty in Artificial Intelligence 338-345. Montreal, Canada.
- **Julio**, P. and Adem, K. (2009). Data Mining and Knowledge Discovery in Real Life Applications, I-Tech pub., Vienna, Austria.
- **Julio**, P. and Adem, K. (2007). Data Mining and Knowledge Discovery in Real Life Applications, Printed in Croatia.
- **Kantardzic, M. (2003)**. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New Jersey
- **Langley (1992)**. An Analysis of Bayesian Classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence 223-228. AAAI Press/MIT Press, Cambridge/Menlo Park. Montreal, Canada.
- **Larose, D. T. (2005)**. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.
- **Larvac**, Nada. (1998). Data Mining in Medicine: Selected Techniques and Applications.
- Available URL.: <http://citeseer.nj.nec.com/lavrac98data.pdf>.
- **Last** M, Maimon, oded, and Kandel Abraham (2002). Knowledge Discovery in Mortality Records:An info-fuzzy Approach. Available URL:<<http://www.csee.usf.edu/softec/med_dm3.pdf>>.
- **Levin** NaZ, Jacob (1999). Data Mining. Available URL:www.urbanscience.com/Data Mining.pdf.

- **Lewis, R. (2002)**. An introduction to classification and regression tree (CART) analysis, in Annual Meeting of the Society for Academic Emergency Medicine. San Francisco, CA, USA.
- **Mehmed Kantardzic (2003)**. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New Jersey.
- **M. Christenson, McDevitt, T., and Stanecki, K. (2004)**. Global Population Profile: 2002. International Population Reports. Health Studies Branch, International Programs Center, Washington Plaza II, Room 313A U.S. Census Bureau, Washington, DC 20233-8860.
- **M. Gams (2007).** Osnovna demografska gibanja (Basic Demographic Dynamics). In J. Malač ič , M. Gams (eds.), Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century. Ljubljana: "Jožef Stefan" Institute, pp. 35-37.
- **Michael, W. B. and Murray, B. (2006)**. Lecture Notes in Data Mining. University Of Tennessee, USA.
- **Plate** T. (1997) Visualizing the function computed by a Feed forward Neural Network. Available URL: http://pws.prserv.net/tap/papers/nc2000.pdf.
- **Quinlan, J. R. (1986)**. Induction of Decision Trees: Machine Learning 1(1):81-106. Edinburgh University Press.
- **Quinlan, J. R. (1993)**. C4.5; Programs for Machine Learning. Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, USA. His web site is URL: http://www.rulequest.com
- **Rea** A. (2002). Data Mining An introduction Student Notes. <<http :// www pccqub acuk /tec /courses/datamining/stu_notes/dm_book_1html>>.[Accessed febraury 10,2016].
- **Richard Pankhurst (1985).** History of Ethiopian Towns (Athiopistische Forschungen)
- **Ringheim**, K., C. Teller, and E. Sines. (2009). Ethiopia at a crossroads: Demography, gender, and development. Population Reference Bureau Policy Brief. <http://www.prb.org/pdf09/ethiopia.pdf >. Accessed January 8, 2016
- **Selam, A. (2011)**. Predicting the Occurrence of Measles Outbreak in Ethiopia Using DM Technology. MSc. Thesis, Addis Ababa University, Ethiopia.
- **Sumathi and Sivanadam (2006)**. Introduction to Data Mining and its applications
- **Thearling** K. (2003). An introduction to data mining.Available at: <<http:// www3. shore. net/~kht/text/dmwhite.pdf>>.

- **Trybula** WJ. (1997) Data Mining and Knowledge Discovery. Annual Review of Information Science and Technology (ARIST) (32) : 197 - 229.
- **Two Crows Corporation (2005)**. Introduction to Data Mining and Knowledge Discovery. Third Edition. Two Crows Corporation. 10500 Falls Road, Potomac, USA.
- **Uddin**, J. (2008). Child Mortality in a Developing Country: A Statistical Analysis. Journal of Applied Quantitative Methods, Sylhet, Bangladesh.
- **Velickov, S. and Solomatine, D. (2000)**. Predictive Data Mining: Practical Example, Moscow, Russia.
- **Vinterbo, S. A. (1999)**. Predictive Models in Medicine: Some Methods for Construction and Adaptation. Norwegian University of Science and Technology, Oslo, Norway.
- **WIKI-1** http://en.wikipedia.org/wiki/Failed_States_Index
- **WHO**, (2007). Regional training workshop on blood donor recruitment: pre and post Donation counseling Available at: <http://www.who.int/countries/eth/news/2008/blood_donor recruitment/en/index.html> [Access date February 08,2016]
- **Witten, (1999)**. Data Mining. Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, USA.
- **World Fertility Report (2013)** - Fertility at the Extremes
- **Yusuf**, M., A. Mekonnen, M. Kassie, and J. Pender. (2005). "Cost of land degradation in Ethiopia: A critical review of past studies." Unpublished manuscript.

Annex I

A J48 10-fold cross validation Prediction interface



Annex II

A J48 cross validation DT Generated for HDSS Dataset

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: TFR-weka.filters.supervised.instance.SMOTE-C2-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C2-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-
weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.instance.Randomize-S42
Instances: 118985
Attributes: 9
rltn_head, Age, Marital Status, Setting, Site, Religion, Educ_status, Occupation, TFR
Test mode: 10-fold cross-validation

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

J48 pruned tree

Marital Status = UNMarried, | Age = 14-19: Low (19974.0/56.0)
Marital Status = UNMarried, | Age = 20-24 | | Religion = Protestant: Low (2031.0/269.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Unemployed: Low (3161.0/273.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Child: Low (1222.0/136.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Parent: Low (0.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Non_Relative: Low (110.0/21.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Other_Relative: Low (85.0/4.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Sibling: Low (39.0/1.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Grand_Parent: Low (94.0/1.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Head:
Low (0.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Spouse | Setting = Rural | Site = Arbaminch: Low (2.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =
Spouse | Setting = Rural | Site = Dabat: Middle (277.0/94.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head =

Spouse | Setting = Rural | Site = GilgelGibe: Middle (0.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Spouse | Setting = Rural | Site = Kilteawelaelo: Low (9.0/1.0)
Marital Status = UNMarried | Religion = Orthodox | Occupation = Daily_Laborer | rltn_head = Spouse | Setting = Urban: Low (38.0/1.0)

Number of Leaves : 1839

Size of the tree : 2192

Time taken to build model: 1.62 seconds

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

Correctly Classified Instances	90443	76.0121 %
Incorrectly Classified Instances	28542	23.9879 %
Kappa statistic	0.6244	
Mean absolute error	0.2187	
Root mean squared error	0.3331	
Relative absolute error	52.3553 %	
Root relative squared error	72.8873 %	
Total Number of Instances	118985	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.798	0.097	0.883	0.798	0.838	0.707	0.925	0.918	Low
	0.663	0.176	0.646	0.663	0.654	0.483	0.832	0.675	Middle
	0.831	0.088	0.696	0.831	0.757	0.696	0.946	0.761	High
Weighted Avg.	0.760	0.121	0.769	0.760	0.762	0.632	0.898	0.808	

=== Confusion Matrix ===

a	b	c	<-- classified as
45370	10535	952	a = Low
5632	25770	7494	b = Middle
364	3565	19303	c = High